Norwegian University
of Life Sciences

Master’s Thesis 2016 30 ECTS
The Department of Ecology and Natural Resource Management (INA)

Cost Analysis and Cost Estimation Model for 1-10 MW Small-Scale Hydropower Projects in Norway

Torfinn Belbo
MSc. Renewable Energy / MSc. Fornybar energi

PREFACE

This thesis marks the end of six years as a student at the Norwegian University of Life Sciences. I have no words for how much I have appreciated the time I have spent here, and how much I have learned and grown in life and academics.

I would first like to thank all friends that I have made, my fellow students in Mannskoret Over Rævne and to the whole "Ås organism" for making my time at this university a truly unforgettable period of my life.

The analysis in this thesis was made possible by the Norwegian Water Resources and Energy Directorate (NVE). They provided me access to their electronic archive at their office in Oslo. I would like to thank the section head Fredrik Arnesen for granting the access, and Seming Skau who has contributed with valuable knowledge.

Professor Torjus Folsland Bolkesjø has been my main supervisor. I would like to thank Professor Bolkesjø for valuable feedback and inputs during the course of this process. I also owe gratitude to my co-supervisor, Ph.D candidate Marko Viiding, who helped leading me to the topic of the thesis and who gave me valuable feedback along the way. Thank you to Daniel Juddson Lohmann, my dear friend, for the thorough proofreading.

Finally, to my fiancée Andrea, to my family, and to my friends. For help, support, encouragement and for all your patience this semester. Thank you, so very much.

Any errors found in the thesis are my sole responsibility.

Ås 13.5.2016 ______________________________

ABSTRACT

Small-scale hydropower has been one of the frontiers in the development of new renewable electricity generation in Norway from the turn of the 21st century. Further, small-scale hydropower will be one of the key technologies to utilize in order to fulfill the political objective to increase the production of renewable electricity toward 2020. A better understanding of factors affecting the development of such schemes would therefore be valuable, but has received little attention in academic research.

This thesis conducts a quantitative analysis of key external cost drivers for small-scale hydropower projects (1-10 MW) commissioned in the past ten years, and attempts to model a cost estimation tool to aid in assessment of new projects that are ready to be deployed.

The results showed that investment costs are consistently underestimated in license application budgets. The median difference for total costs in nominal values was found to be 49.1%, and 1.12 NOK/kWh (estimated annual production). The analysis further documents that:

1) Specific total investment costs (in NOK/kWh) increased in real values during the past ten years, with an estimated average growth rate for all projects at 3.7 pp per year (with 2005 as the base year).
2) Total investment costs increase with longer construction periods. An average increase of 37.8% per year in real values was estimated when all projects in the dataset were taken into consideration.
3) Specific investment costs differ between geographical regions. Projects in Western Norway tend to have relatively low specific costs with an observed average of 3.79 NOK/kWh in real values. Projects in Northern Norway tend to have a relatively high specific investment cost, with an observed average of 4.7 NOK/kWh in real values.
4) Projects developed by the owners of the water resource were found to have lower reported costs than projects developed by professional project development companies. Projects in the non-professional group had an observed average specific cost of 3.95 NOK/kWh in real values, while the projects in the professional group had an average specific cost of 4.71 NOK/kWh in real values.

Two cost estimation models for investment costs of small-scale hydropower projects were developed, with use of multiple linear regression. The first was developed for predicting total investment costs, and achieved a mean absolute error rate of 18.0%,. The second was developed for predicting partial costs, and achieved an error rate of 15.6%.

This thesis contributes to the literature by documenting the scales of underbudgeting in license applications, and by analyzing estimated effects of the selected cost driving factors. The cost estimation models developed may prove useful in comparing future small hydropower projects with respect to their investment costs. They can be used to produce independent cost predictions, and complement license application budget estimates for increased accuracy and indication of the cost uncertainty for each project.

If applied for analyzing future project development, these findings may be useful for the decision makers and for the hydropower industry.

SAMMENDRAG

Småskala vannkraft har vært en av driverne for utvikling av ny fornybar kraftproduksjon i Norge siden årtusenskiftet. Framover vil også småkraft være en av driverne for å oppfylle de politiske målene om økt fornybar kraftproduksjon fram mot 2020. En bedre forståelse av faktorer som påvirker utviklingen av ny småkraft er derfor fordelaktig, men har mottatt lite oppmerksomhet i forskningen.

I denne masteroppgaven utføres en kvantitativ analyse av eksterne kostnadsdrivere for småkraftprosjekter (1-10) satt i drift i løpet av de siste ti årene. Et kostnadsestimeringsverktøy blir utviklet med hensyn på å kunne bidra til å evaluere nye småkraftprosjekter som har mottatt konsesjon.

Resultatene i denne analysen viser at investeringskostnadene regelmessig underbudsjetteres i budsjettene i konsesjonssøknadene. Median differanse mellom budsjetterte og innrapporterte totale kostnader var 49,1% og 1,12 kr/kWh (estimert årlig produksjon). Analysen dokumenterer videre at:

Spesifikk investeringskostnad (i kr/kWh) hadde en realøkning i løpet av de siste ti årene, med en estimert gjennomsnittlig rate for alle prosjekter på 3,7 p.p. per år (med 2005 som basisår)

Totale investeringskostnader økte ved lengre byggeperiode. En gjennomsnittlig økning tilsvarende 37,8% per år i faste priser ble estimert når alle prosjekter i datasettet var inkludert.

Spesifikk investeringskostnad varierer mellom geografiske regioner. Prosjekter på Vestlandet tenderer til å ha relativt lave spesifikke investeringskostnader, med et observert gjennomsnitt på 3,79 kr/kWh. Prosjekter i Nord-Norge tenderer til å ha relativt høye spesifikke investeringskostnader, med et observert gjennomsnitt på 4,7 kr/kWh i faste priser.

Prosjekter utviklet av grunneiere hadde lavere innrapporterte kostnader enn prosjekter utviklet av profesjonelle aktører. Prosjekter utviklet av grunneierne hadde en observert gjennomsnittlig spesifikk kostnad på 3,95 kr/kWh i faste kroner, mens prosjekter utviklet av profesjonelle aktører hadde en observert gjennomsnittlig spesifikk kostnad på 4,71 kr/kWh i faste kroner.

To kostnadsestimeringsmodeller for investeringskostnader for småkraftprosjekter ble utviklet ved hjelp av lineær regresjonsanalyse. Den første modellen ble utviklet for estimering av totale utbyggingskostnader, og hadde en gjennomsnittlig absolutt feilrate på 18.0%. Den andre modellen ble utviklet for estimering av totale delkostnader, med en feilrate på 15.6%.

Denne masteroppgaven bidrar til forskningslitteraturen ved å dokumentere omfanget av underbudsjettering i konsesjonssøknadene, og ved å analysere effektene av de utvalgte eksterne kostnadsdrivende faktorene. Kostnadsestimeringsmodellene være nyttige for å sammenligne framtidige småkraftprosjekter med hensyn på utbyggingskostnad. De kan brukes som et uavhengig verktøy for å estimere utbyggingskostnader, og komplementere budsjettestimatene for økt nøyaktighet og en indikasjon på usikkerheten av utbyggingskostnaden for enkeltprosjekter.

Disse funnene kan være nyttige for beslutningstakere og vannkraftsektoren for å analysere framtidig utvikling av småkraft i Norge.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of definitions and translations

English terms used in this thesis. Hydropower terminology mainly in accordance with ESHA (2004), Norwegian terms in accordance with NVE (2010b).

English	**Norsk**
Absorption capacity *See maximum discharge*	Slukeevne
Compensation flow *Minimum flow of water required to pass the dam/intake and run free in the water course.*	Minstevannføring
Connection fee *In this context, a fee paid to the grid owner (transmission or distribution network operator) to cover parts of the investment costs in cases when the capacity of the existing grid must be upgraded in order to allow for the new power production*	Anleggsbidrag
Energy *In electricity terms, work performed by electrical energy, measured in kWh or MWh.* 1 Wh = 1 J/s × 3600 s = 3600 J	Energi
Fifth percentile water flow *The average water flow rate level in which the water flow rate is below in five percent of the time for a given period (year or season) based on a given dataset of hydrological measurements*	Fem persentil vannføring
Generator *Device transforming mechanical energy from the turbine to electrical energy*	Generator
Head, gross *Difference between intake MASL and MASL at center of turbine (Pelton turbines) or downstream water level (Francis turbines).*	Brutto fallhøyde
High-head power plants *Hydro power plant with gross head higher than 300 m, as defined by NVE (2015c).*	Høytrykksanlegg
Installed capacity *Production capacity of the generator(s) of the hydropower plant, in kW or MW*	Installert effekt
Intake *Construction where the water is led from the river to the waterway*	Inntak
MASL. – Meters Above Sea Level	Meter over havet (moh.)
Maximum discharge *Maximum discharge of water in turbine(s), measured in* m^3/s	Slukeevne
Micro hydropower plants *Hydropower plants with an installed capacity of less than 100 kW (NVE 2010b).*	Mikrokraftverk
Mini hydropower plants *Hydropower plants with an installed capacity of between 100 and 1000 kW (NVE 2010b).*	Minikraftverk
Nominal values, prices or costs *Values not adjusted for cost inflation. Also termed as current prices.*	Nominelle verdier, priser eller kostnader

English	**Norsk**
Ordinary low water flow *An observed average minimum water flow rate based on a given dataset of hydrological measurements (this measure has a more precise calculation procedure used by NVE)*	Alminnelig lavvannføring
Penstock *Pressure pipe leading the water from the intake to the turbine, made from ductile or cast iron, steel, glass-fiber, plastic and/or concrete*	(Trykk-) rør
Power station/house *The power station building in which the turbine, generator and control equipment is installed, and sometimes the transformer.*	Kraftstasjon
Power *Capacity to perform work, measured in watt (which equals joules per second)*	Kraft/effekt
Real values, prices or costs *Values adjusted for inflation or a specific cost index. Also termed as constant prices.*	Reelle verdier, priser eller kostnader
Run-of-river scheme *Hydropower plants with no water magazine, often with a low head.*	Elvekraftverk
Shaft/pressure shaft *In this context, structure in rock to lead the water from the intake to the turbine, constructed by drilling in rock. Distinguished from tunnels by having a smaller cross-section and constructed only by drilling.*	(Trykk-) sjakt
Small hydropower plants *Hydropower plants with an installed capacity of between 100 and 1000 kW (NVE 2010b).*	Småkraftverk
Specific (investment) cost *Investment cost per estimated average annual production, measured in NOK/kWh (NVE 2015c).*	Spesifikk (investerings-) kostnad
Tailrace *Canal or tunnel which leads the water from the turbine(s) back to the main stream (tail)*	Utløpskanal/-tunnel
Transformer *Device changing the voltage of the electrical current to desired level.*	Transformator
Tunnel *In this context, a structure in rock to lead the water from the intake to the turbine, constructed by blasting or drilling in rock. Sometimes also used in tailrace to lead the water from the power station back to the river.*	Tunnel
Turbine *Mechanical device transforming kinetic energy from the water to mechanical energy*	Turbin
Waterway *Collective term for structures leading the water from the intake/headwater to the tail water, including penstocks, tunnels, shafts and canals.*	Driftsvannvei/vannvei

1 INTRODUCTION

Since the turn of the 21st century the Norwegian power sector has experienced a new era in the development of hydropower. In the period from 1990 to 2015, 4000 MW of new capacity has been installed in about 900 hydropower plants, the majority of which has been developed within the last ten years (OED 2016). Small-scale hydropower projects have been developed in large numbers in all regions of the country.

Norway is a hydropower nation, with around 96% of its electricity production generated hydropower (SSB 2016b), and this production volume is set to increase . As of 2012, Norway has enacted an electricity certificate scheme with Sweden, aiming to increase the total annual production of renewable electricity by 28.4 TWh by the end of 2021. As of January 1st, 2016, Norway and Sweden are half-way to the production target, with an increased annual production of nearly 14 TWh (NVE & Energimyndigheten 2016). The new production is expected to come mainly from small, mini and micro-scale hydropower and wind power, as these technologies are relatively mature and have the lowest costs (NVE 2015c). By the end of 2015, The Norwegian Water Resources and Energy Directorate (NVE) had approved new projects comprising 16.3 TWh annual production of hydropower (NVE 2015a) and almost 20 TWh annual production of wind power (NVE 2015b). The limited available volume of electricity certificates will not allow all of these projects to be developed and receive certificates, and such projects will most likely not be economically viable without the income from electricity certificates. This master thesis explores the question of which of these small hydropower projects (SHP, 1-10 MW) are more likely to be implemented, and the factors that affect this selection.

The motivation behind this thesis is to make use of the past ten years of experience from small-scale hydropower project development in the future planning and decision making. **Table 1** shows the number, capacity and estimated average annual production of small hydropower plants that are in operation, license applications filed to NVE and being processed, and projects that have received license, but have not yet been built. It shows that there are more small hydropower production projects in the pipeline than there are in operation. Based on the political goals of increasing the power production, and the fact that hydropower remains one of the most cost-efficient renewable power production solutions, a high number of new small-scale hydropower projects may be developed. In spite of this there will be a selection among the projects available. This raises questions of which lessons can be learned for the future project developments from the past ones, and which future projects will most likely have a low cost per production unit.

Table 1: Status for small-scale hydropower projects in Norway.

	No. of plants	Capacity [MW]	Prod. [TWh/yr]
Installed	614	2067	8.5
License applications	331	1105	3.3
License received, not built	335	1151	3.6

Installed capacity as of 1.1.2015(NVE 2015e). License applications under process by NVE as of 29.11.2015 (NVE 2015a). Projects that have received license, but are not yet built, as of November 2015 (NVE 2015d).

No published journal articles have been found on this topic in Norway. Stokke (2014) completed a master thesis on deviations between budgeted and actual investment costs for small hydropower projects, identifying a trend of optimistic budgeting by project developers. The main issues leading to higher costs were related to project planning, in particular related to soil mechanics at dam/intake and in the waterway and time duration of the construction period. Haga and Espegren (2013) conducted a similar study for NVE with more quantitative data and reached similar conclusions.

1.1 Approach and structure

This thesis expands the work from the aforementioned studies. It is based upon the same type of data as in Haga and Espegren (2013), but with more recent data and a more extensive record. The thesis has two main objectives. The first is to conduct a numeric analysis of deviations between budgeted costs and investment costs, and to identify and quantify factors that affect investment costs. The research questions to be answered in the first part are:

> *Is there a significant trend of higher reported costs than budgeted in the license application?*
>
> *Do the following four factors have a significant effect on investment costs? (Construction year, construction time, geography and ownership of the hydropower project).*

The second objective is to develop a cost-estimation tool for small-scale hydropower projects (1-10 MW). The research question to be answered here is:

> *Can an investment cost estimation model be developed that has higher accuracy than budget estimates from license applications?*

The tool is built upon experience-based data, and is meant to be applied to projects that have received the required license, but have not yet been built. The tool should provide an estimate of investment cost for the projects, along with an estimate of its uncertainty.

The goal for the cost estimation tool is for it to be accurate enough to suffice as a bench-marking tool for license owners, investors, decision makers, and the public. It could be used together with the budget-estimates of each hydropower project as a measure of uncertainty of the project costs, and as an indication of how different/similar the project is to the projects upon which the cost estimation model has been developed.

It can also be used as a cost-ranking tool to compare several hydropower projects, ultimately providing useful information to the public on which projects are likely to be the most cost-efficient.

The structure of the thesis is as follows: Following this introduction, Chapter 2 provides a background for the current state of the Norwegian small hydropower sector. Chapter 3 presents an in-depth literature review of former studies relevant for the analysis in this thesis. Chapter 4 presents the theory and methodology on which the thesis is based. It introduces basic information on small-scale hydropower in Norway, including cost elements of hydropower projects, and presents the relevant methodology for statistical analysis used in this thesis, collection and handling of the data used in the analysis, and finally a detailed documentation of how the data was analyzed. Chapter five presents some main characteristics of the portfolio of projects in the collected dataset which gives a context for the results of the analysis. In chapter six, the main results are presented according to the structure of the research questions. In chapter seven the results are discussed with regards to previous literature, interpretation, internal and external validity of the findings. Chapter eight gives a conclusion of the work done in this thesis.

2 Motivation: Status for small hydro in Norway

The following section will present recent developments in the Norwegian small-scale hydropower market that are important in the context of developing a cost estimation tool. It will also explore the motivation behind developing a cost estimation tool for small-scale hydropower projects in Norway in greater detail.

2.1 Recent changes in the small hydro sector

2.1.1 Electricity price development

One recent development within the hydropower sector is increased hesitation on the part of small hydro license-holding owners to invest in the realization of their hydropower projects. This stems from a decrease in electricity prices in Norway over the past five years, and low expected future electricity prices, which make investments in new electricity generation less attractive.

To shed light upon the falling investment rate and profitability in the market, it is valuable to observe electricity price development. **Figure 1** shows the development of the electricity price from 2001. Although there is no clear trend in the price development over the whole period, the past five years show a gradual decrease in prices, which has disincentivized investors

It may also be useful to look at how the market views the future electricity price. Financial contracts for future electricity deliverances are traded for up to five years ahead in time. Such contracts are the best estimate of future electricity prices, according to the knowledge existing in the market. **Figure 2** provides a snapshot of market expectations for the next four years. The market estimates prices around 200 NOK/MWh - more than a third lower than the average price between 2005 and 2010, and about one fourth lower than the average price between 2011 and today. This is another indication of a market in decline, which makes investors more pessimistic.

2.1.2 Change in ownership

A second development related to investments is a growing interest in small hydro from foreign capital funds. The German investment fund Aquila Capital recently acquired Småkraft AS and Norsk grønnkraft AS, some of the largest owners of small hydro plants. These acquisitions made the company the largest player in the Norwegian small hydro sector. Scottish-owned SL Capital Partners LLP recently acquired Nordic Power, the owner of 13 small hydro plants.

Norwegian hydropower companies are now pulling out of small hydro because the profitability is lower than their required rate of return. Many of the foreign companies now looking to invest in Norway have high equity and are interested in secure, long-term investments with a lower required rate of return.

This combination of circumstances may lead to increased sales from Norwegian small hydro owners to foreign companies. There may also be a shift from new investments being driven by local Norwegian owners seeking profitable investments using their own and local capital, to investments being driven by foreign companies, demanding lower rates of return.

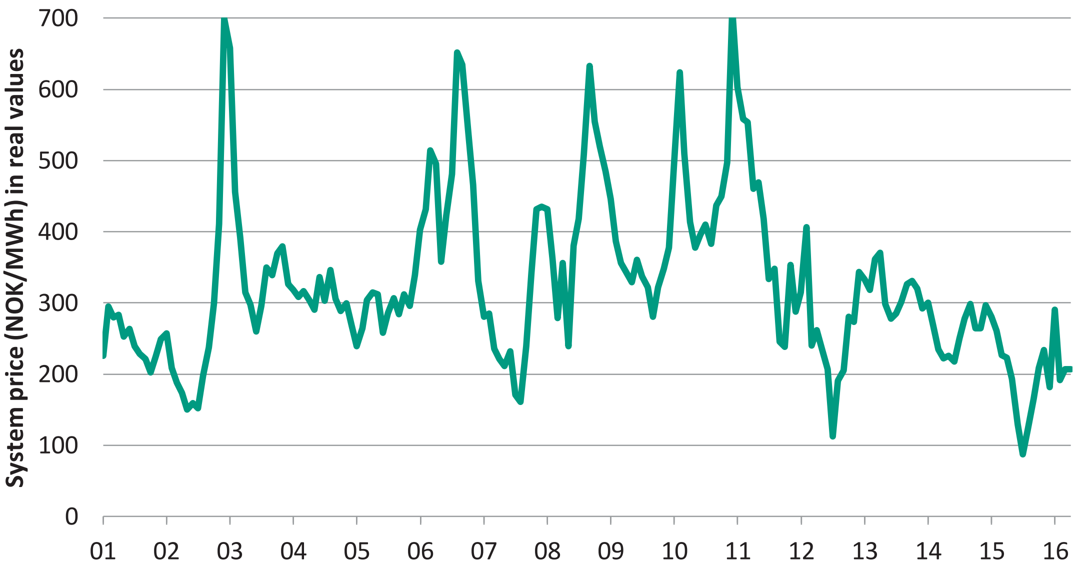

Figure 1: System price for Nordpool day-ahead market 2001-2016, adjusted for inflation. Real prices as of April 2016 (Nord Pool 2016; SSB 2016a).

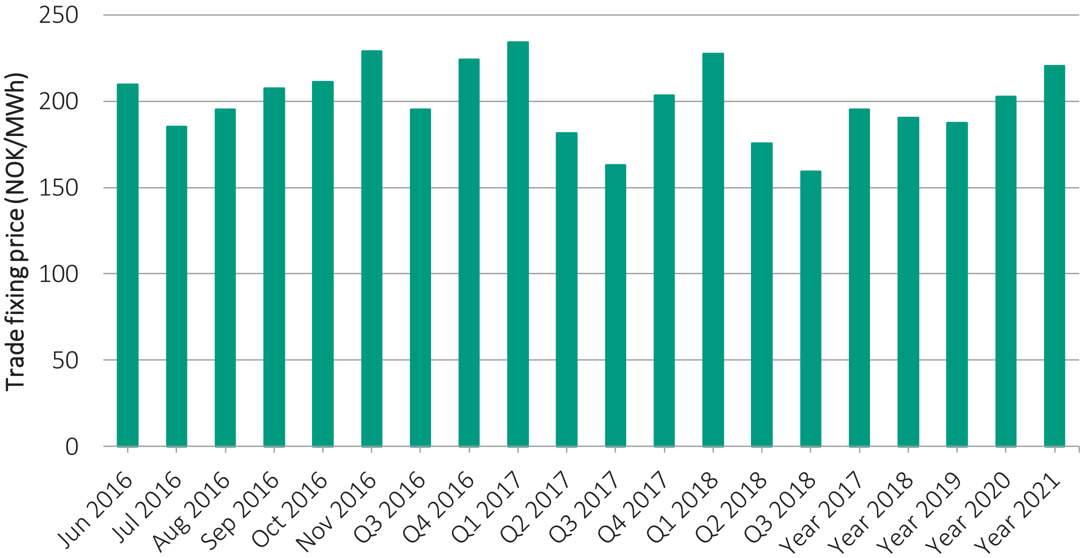

Figure 2: Futures trading prices from Nasdaq OMX, snapshot from trades made on Wednesday May 11th 2016 . Each column shows trade fixing prices for contracts months, quarters and years ahead (NasdaqOMX 2016)

2.1.3 Recent developments of small-scale hydropower costs

There are also factors on the cost-side that lead to lower profitability for new small hydro projects. Head of Småkraftforeninga Knut Olav Tveit points to two framework conditions lowering the potential profitability for investors (Aspen 2014). First, many of the pending projects are in areas with grids that lack the capacity for their power. Within the current framework the owners must pay their share for grid investments caused by their initiative, increasing the investment cost. Second, NVE has increased their demands for minimum flow of water in rivers with a power plant is in operation. This decreases the load duration (the number of hours during the year with enough river discharge for the power plant to produce electricity) for the plant, ultimately lowering profitability. Thus, to a certain extent these two conditions contribute to a lower rate of return.

2.2 MOTIVATION FOR DEVELOPING A COST ESTIMATION TOOL FOR SCREENING OF PROJECTS

Currently, there are no easily accessible cost ranking tools for small hydropower projects in Norway that have received licensure. The public has access to recent license applications at NVE's webpages(NVE 2015a), which provide budgeted investment costs and estimated average yearly production. Several factors make it difficult to compare the attractiveness across all projects, including:

- Varying application dates (license applications dating back to 2005), prevent the costs of most of the applications from factually representing the actual costs of today.
- the quality of the budgeting will vary due to the fact that some applications are carried out by owners with limited experience, some carried out by companies with experience from other small hydropower projects. This is discussed later in this chapter.

It would be ideal to have a cost ranking to provide accurate and aligned estimates of investment costs for the projects that have received license.

There does exist a more general tool to assess water resources, their potential and estimated investment costs for possible small hydropower projects. NVE carried out a resource assessment for small hydropower projects in Norway in 2004 (NVE 2004). This assessment was done with use of digital spatial analysis methods to calculate the potential of small-scale hydropower projects with a production capacity between 50 and 10 000 KW. It identified possible projects with a specific investment cost lower than 3 and costs between 3 and 5 NOK/kWh. The result revealed a possible potential of 9467 schemes. 4128 schemes were discovered having estimated costs lower than 3 NOK/kWh and an estimated average yearly production of 18 TWh. 5339 schemes were discovered having estimated costs between 3 and 5 NOK/kWh, and an estimated average yearly production of 7 TWh. In addition to this resource assessment meant to identify projects smaller than 1000 kW, an assessment was conducted for possible hydropower projects larger than 1000 kW in the 1980s with the "Samlet plan". Under this assessment, possible projects were identified and evaluated in more detail, with regards to production- and economic potential, environmental considerations, etc. The results from the "Samlet plan"-report was regarded by the report of NVE (2004) as being more accurate than their purely quantitative approach, so the results from "Samlet plan" are also included in the report. They report 7 TWh from the "Samlet plan" as having lower costs than 3 NOK/kWh. In total, the report showed a potential of 25 TWh of new small-scale hydropower projects with an investment cost lower than 3 NOK/kWh, with around 5 TWh of it having potential of being commissioned within a ten-year period.

The purpose of NVE's resource assessments was to help the public to identify possible projects, to be a basis for developing local energy plans, for local authorities to make land use plans, and for the central authorities to get an overview of the overall potential of water resources in Norway. The analysis does not take ownership arrangements into consideration, nor does it incorporate environmental considerations. It was not intended to evaluate specific projects, and therefore has limited value if the aim is to compare different license applications and their attractiveness.

NVE's resource assessments give an overview of potential projects, but are not linked to projects that are actually in the pipeline of the license process. The license applicant may choose a different river span than suggested by the spatial analysis, may have come up with more detailed hydrological data from assessments, and may also need to change the specifications of the hydropower project according to demands from the license authority itself, NVE. Thus, this tool cannot be used to evaluate the possible potential of a specific licensed project.

The tool used by project owners for estimating investment costs for their projects, is NVE's cost base for small hydropower plant (2010a). It provides cost data for small hydropower projects of high accuracy, and is updated each year with an index for the various components. Although it has high accuracy, the guide is too detailed to be used in an overall screening of hydropower projects akin the one this thesis is meant to carry out. Still, the guide will be useful to validate the model created in this thesis.

.

3 Literature review

In this section, relevant literature is presented to put the analysis in a context. The first section describes in greater detail the two studies found on cost analysis of Norwegian small-scale hydropower. The second section presents an in-depth literature review of cost estimation modeling.

3.1 Research on budgeted versus reported costs for SHP in Norway

Stokke (2014) assessed 24 small hydropower projects and used a survey to collect data from the owners on deviations between budgeted investment costs and actual costs. Stokke identified a trend of optimistic budgeting from the owners in their license applications. The report shows that 23 out of 24 hydropower plants in his analysis had higher actual costs than budgeted costs. The various components of the projects had errors of different magnitude. The owners in his study reported that the largest errors were related to planning and administrative costs, intake/dam and electro- mechanics.

In an internal report for NVE, Haga and Espegren (2013) did an analysis of the deviation between budgeted costs and reported costs for 74 small-scale hydropower plants. They used data from hydropower plants commissioned between 2008 and 2013, with use of budgeted costs in license applications and actual costs reported from the plant owners. They found that 83% of the hydropower projects in their study had higher actual investment costs per kWh average production than budgeted in the license applications. For 60% of the power plants, the investment costs ended up more than 0.5 NOK/kWh over budgeted investment cost, and more than 1.5 NOK/kWh over the budgeted cost for 20% of the plants. As seen in **Table 2**, their analysis indicates that there is a trend of underbudgeting in the early-phase project planning, with the highest deviations in costs for the intake.

Table 2: Deviations between budgeted and actual costs for small hydropower projects in Haga and Espegren (2013). n = 58.

Expected deviation	Median difference	Mean difference
Intake	44.1%	36.9%
Waterway	22.7%	13.6%
Power station	1.8%	-1.4%
Total	16.6%	13.6%

They found that the median deviation between budgeted total costs and actual total costs was 0.806 NOK/kWh, with a standard deviation of 0.858 NOK/kWh for the 74 projects in the dataset. These are important findings that suggest that more efforts should be put into accurate budgeting.

From in-depth interviews with some of the power plant owners, they found that the projects in many cases had:

- Lower annual production than planned (which itself leads to a higher cost per kWh estimated annual production),
- longer planning- and construction period than planned
- unpredictable entrepreneurial costs for intake and waterway due to
  - insufficient knowledge about geological and geotechnical parameters on the project site in the early phase planning
  - changes in the intake and waterway detail-plans

These three factors may or may not be representative for all power plants, but are common and therefore important to take into consideration.

3.2 Cost estimation modeling

Cost estimation is defined by GAO (2009 in Preface, p. i) as

> *"the summation of individual cost elements, using established methods and valid data, to estimate the future costs (…), based on what is known today."*

Cost estimation is carried out in any phase of a project, for example in pre-feasibility, feasibility, detail planning, or in tenders. The accuracy of cost estimation estimates increases with the level of detail of the project plan. For the scope of this thesis, the focus is on cost estimation on a pre-feasibility or feasibility-level of detail.

In this section, published academic literature dealing with cost-estimation of infrastructure projects will be reviewed (NVE has its own early-phase cost estimation methodology which will be presented in the next chapter). The section includes cost estimation methods for hydropower projects, as well as methods from other industries.

Cost estimation models reviewed here make use of experience-based, quantifiable data in some form, which can be utilized for predicting costs for new projects. Such models may be based upon various methods, such as statistical regression, fuzzy logic, artificial neural networks, case-based reasoning, factor and pattern time series analysis, genetic algorithm and particle swarm optimization (Cavazzini et al. 2016; Elfaki et al. 2014; Gordon 1983; Kim et al. 2004; Kim et al. 2012; Smith & Mason 1997; Trost & Oberlender 2003; Tuhtan 2007; Wang et al. 2012; Ökmen & Öztaş 2010).

There does not seem to be a broad consensus in the literature which model is superior. Some studies find that networks can have higher precision than other methods when little guidance is given in constructing the model (Gunduz & Sahin 2015; Kim et al. 2004; Smith & Mason 1997). Multiple regression models may perform better when they are well-defined, when the model developer has knowledge of the underlying relationship between variables, and when the relationship between the cost predictor variables and the cost response has a functional form without discontinuities (Smith & Mason 1997).

Kim et al. (2004) compared the performance of three different types of cost estimation methods. The methods tested were multiple regression, neural network and case-based reasoning. The three methods were applied to prediction of construction costs for residential buildings. They report a mean absolute error rate (abbreviated to MAER henceforth) for each model type. The error is calculated as the mean absolute deviation between the predicted values and the reported costs, divided by the reported costs. The multiple regression model had a MAER of 6.95%, the best neural network model gave an MAER of 2.97% and the case-based reasoning model gave a MAER of 4.81%.

Smith and Mason (1997) carried out a comparison between multiple regression and neural network models for cost estimation on simulated and real data. The simulated data was created using a third order function and by adding noise. Here they tested the neural network model performance against three different regression models, namely a first-order model, a second order model, and a model fitted using the same functional form as from which the data was generated. In this experiment, the second and third ordered models outperformed the neural network mode. The neural network model performed better than the first-order term regression model.

In the real data sample problem, the two methods were tested on prediction of costs for pressure vessels for chemical production, based on 20 cases. When tested on the real data, a first order regression model was fitted with three predictor variables, on a data subset of 16 observations (where four observations had been excluded due to extreme values). The performance of the multiple regression model and the artificial neural network models was tested using a leave-one-out cross-validation procedure. The performance test showed a significantly better performance of the neural network on all test parameters. While the neural network model had a MAER of 10.72%, the regression model had a MAER of 30.39%.

The authors commented on the choice to not explore interactions and second order models in the multiple regression model selection for the real dataset: In this case, the authors had no a priori knowledge of the true relationship between costs and physical features. There are a large number of possible sub-models which can be tested once interaction terms, second or third order terms and transformations are included as possible predictor variables. The authors argue that constructing and testing such complex regression models based on random selection of higher order terms and/or data transformation defeats the purpose of cost estimation models because they should be simplistic and require little insightful knowledge of the physical features and their interactions. However, with the computation capacity of modern statistical computer software, a model developer should be able to develop complex models which can be reduced by stepwise regression methods, requiring little computational time (to a certain limit).

3.2.1 *A closer look at former cost estimation studies for hydropower projects*

There have been a number of articles published concerning cost estimation and reducing uncertainty of cost estimates, which are of relevance for this thesis. Research with a more broad approach to cost estimation, levelized cost of energy (LCOE), uncertainty and sensitivity analysis for small-scale hydropower projects, include Merrow and Schroeder (1991), Bacon and Besant-Jones (1998), Jenssen et al. (2000), Kaldellis et al. (2005), and Kaldellis (2007). An extensive volume of research has been carried out on cost estimation methods for partial costs for hydropower projects, which will be explained in greater detail in the following section.

3.2.2 *Partial cost correlations*

The earliest study found considering cost estimation for hydropower was Gordon and Penman (1979). They established a cost estimation model that has been the basis of the majority of the subsequent research on cost estimation modeling. Based on analyses of 64 estimates of projects that were to be installed at existing dams, they developed cost equations, called "correlations" in the literature, for hydropower plants up to 5 MW. According to Cavazzini et al. (2016), this was the first study which established a correlation between the cost of electro-mechanical equipment, power, and hydrological head. The correlation equations have the following basic form:

$$C = aP^bH^c,$$

where C is the electro-mechanical equipment cost, P is the power capacity and H is the net head. The a, b and c are coefficients found using statistical regression on a dataset of hydropower projects.

Gordon later published several other studies using the same methodology; Gordon (1981) did a similar study on hydropower hydro power station costs between five and 1000 MW with heads between 10 – 300 m. Gordon (1983) (as cited in Singal et al. 2010 p. 117) developed a methodology for early-phase estimation of project costs for hydropower projects. The methodology in the latter paper developed was based on a statistical analysis of data from 170 projects. The estimation model had head and capacity as the main input parameters, was calibrated for large hydropower projects with medium- to high hydrological heads, and had a measured estimation accuracy of $\pm 40 - 50\%$. Gordon and Noel (1986) developed a methodology for estimating minimum costs for new small-scale hydropower plants, based on analysis of cost data from 141 projects. (It was not possible to access the full-texts of any of the above-cited publications by Gordon. Information about the publication was retrieved from other papers citing these publications, and from the abstracts of the publications, where they were available. It was still worthwhile to mention them.)

Singal et. al. published a series of papers on cost estimation of small hydropower schemes in India, relying on the same basic methodology of Gordon. Singal and Saini (2007) developed a cost equation for small-size, low head run-of-river projects with hydrological head between six and 15 meters and installed capacity of one to ten megawatts, with an accuracy of ±12%. Singal and Saini (2008) developed cost-estimation equations for small, low-head dam-toe hydropower plants based on the number of turbines, a hydrological head of 3-20 meters, and a capacity of 1-5 MW. Singal et al. (2010) developed a set of cost estimation equations for projects with heads in the range of 3-20 m and capacity between 1-5 MW with the use of statistical analysis. Their model validation showed an accuracy of ±11%. Mishra et al. (2012) developed a cost estimation equation for electromechanical equipment based on hydrological head and installed capacity. This is based on a log-log-transformed least squares regression analysis. They reported a prediction accuracy of ±10%.

Ogayar and Vidal (2009) developed a cost-estimation model for electro-mechanical equipment for small hydropower plants in Spain based on the methodology of Gordon and Penman (1979). They developed individual models for the three main turbine types: Pelton, Francis and Kaplan. They also did a comparison with a list of studies using the same model approach (Anagnostopoulos & Papantonis 2007; Kaldellis et al. 2005; Kaldellis 2007; Montanari 2003; Sheldon 1981; Willer 1991). The cost estimation model of Ogayar and Vidal (2009) had an error range between 19.52% and -9.50% for the cases in their study, and their model performed better than the cost equations proposed by the papers that had been reviewed.

Aggidis et al. (2010) had a similar model approach as Ogayar and Vidal (2009) and made cost-estimation equations for turbines and electro-mechanical components in small-scale hydropower schemes in the UK. The input variables in Aggidis et al. (2010) were: hydrological head, discharge, turbine type, installed capacity and partial costs. They report prediction accuracy of the equations down to ±10%, and up to ±25% for different turbine types, and ±25% accuracy for electro-mechanical equipment.

Zhang et al. (2012) developed a similar set of cost estimation equations for total project costs and electro-mechanical costs for different turbine types in the US. Their cost-equations were also based on the methodology of Gordon and Penman (1979). They reported a very low accuracy of the total project cost equation due to a small sample size. The regression results for the electro-mechanical costs had also a lower accuracy than that of many other studies.

Cavazzini et al. (2016 p. 749) attempt to develop the cost correlation methodology further by adding turbine discharge as a third cost determining variable in the cost equation. The model is estimated using a Particle Swarm Optimization method. They present a thorough literature review of formerly developed cost equations, where many of the above-mentioned authors are cited. Their model performed with mean errors below 10% for electro-mechanical equipment for Pelton and Francis turbines and below 20% for Kaplan turbines. Their model outperformed the other studies reviewed in the paper, with lower mean errors.

3.2.3 Cost-estimation for small hydropower model using linear regression and artificial neural networks

Gunduz and Sahin (2015) developed and tested two cost estimation models for small hydropower projects based on a subset of physical features of the projects. They built a multiple regression model and compared it to a model based on the neural network method. The physical feature variables they used as initial input variables were: Project cost, installed capacity, average discharge (of river), project design discharge (turbine discharge), project design head, length of tunnel, length of channel, length of transmission line, diameter of penstock, length of penstock, five year occurrence flood discharge, hundred year occurrence flood discharge, and catchment area of basin. The dataset contained 54 projects, and the model performance was validated on a selection of five projects.

The full multiple linear regression model in this study is a first-order model, with all of the above-mentioned variables, with no interaction terms, squared terms or transformations of variables. They conducted a backwards stepwise selection where coefficients with high p-values were omitted, step by step. Their final model was:

$$\begin{aligned} Cost = \hat{\beta}_0 + \hat{\beta}_1 \times Turbine\ discharge \\ + \hat{\beta}_2 \times Gross\ head + \hat{\beta}_3 \\ \times Tunnel\ length + \hat{\beta}_4 \\ \times Transmission\ line\ length \\ + \hat{\beta}_5 \\ \times 100\ years\ flood\ discharge \end{aligned}$$

The model validation was done by calculating the mean absolute prediction error for the five validation sample projects, based on the formula above. The best subset regression model gave a mean absolute prediction error rate of 9.94% for the validation samples, while the best artificial neural network model gave a mean absolute prediction error rate of 5.04%.

4 Theory and methods

This chapter begins by introducing of some of the key characteristics of small-scale hydropower and the setting in which the analysis was carried out. In the succeeding sections the methodology for data collection, data handling and data analysis is described.

4.1 Regulatory, technical and economic characteristics of small-scale hydropower in Norway

In this section, relevant concepts for small-scale hydropower (SHP) in Norway will be introduced to give the reader a notion of the framework in which the analysis in this thesis was conducted. The section includes a brief explanation of the legislative framework for small-scale hydro, central components of small-scale hydropower plants and typical cost characteristics of small-scale hydropower plants.

In order to build and establish small-scale hydropower schemes in Norway, one is required to apply for a license, according to Section 8 of the Act relating to river systems and groundwater [Act No. 82 of November 24th 2000: the Water Resources Act]. The main criterion for receiving a license for such measures is given in section 25, which states that

> *"A licence may be granted only if the benefits of the measure outweigh the harm and nuisances to public and private interests affected in the river system or catchment area" (OED & NVE 2007) Section 25*

NVE is in charge for the administrative procedure of granting licenses for small-scale hydropower schemes. The license application process requires applicants to submit a detailed plan and budget for the hydropower project (NVE 2016c). One of the assessment criterions is the economy of the project (which NVE considers in each case according to section 25 of the water resource act).

SHP plants usually have little or no water magazine capacity. The schemes comprise of the following main components:

- Dam and/or intake, in some cases canal
- Waterway: penstock, tunnels and/or shafts
- Power station building with turbine, generator, transformer (sometimes in a separate building structure) and control equipment
- Tailrace leading the water from the turbine back to the stream
- Connection line to the nearest local or regional transmission network
- Roads to the power station, and usually to the intake, and sometimes along the penstock.

For more details on physical features of SHP plants, see ESHA (2004) or NVE (2010b).

The typical cost shares of components will vary according to the characteristics of the hydropower plant. As shown in **Table 3**, the waterways can be a major cost driver, followed mechanical equipment and electro-technical equipment. According to a recent publication from NVE (2015c), SHP with a hydrological head lower than 300 meters usually have an investment cost distribution of 59% related to general civil works (which includes waterway, dam, intake, power station building and access roads), 24% related to mechanical equipment and 17% related to electro-technical equipment. High-head hydropower plants, including SHP, have a higher share of costs related to civil works of 69%, 13% related to mechanical equipment and 18% related to electro-technical equipment.

Table 3: Cost components for small-scale hydropower plants (1-10 MW), from (NVE 2010b)

Cost component	Cost share
Access roads to power station and intake	1 - 5%
Dams and intakes	5 - 10%
Waterways (penstocks, tunnels, shafts, and/or canals)	10 - 50%
Mechanical equipment: Turbines, turbine control, valves, etc.	20 - 30%
Electro-technical equipment: Generators, control- and device installations, transformers	15 - 25%
Power station buildings	2 - 5%
Power line connections	5 - 15%
Administrative work, contracts and planning, detail project planning, construction management	7 - 10%
Water rights ("fallrettighet"), miscellaneous costs	2 - 5%

NVE has developed a cost basis for hydropower projects. It was first released in 1982, was initially revised in 1987, and has been revised every five years since 1990. In 2010 the first cost basis was released for small-scale hydropower, projects with installed capacity below 10 MW (NVE 2016b).

The cost basis is intended to provide hydropower license applicants and project developers proper cost estimates for early-phase planning of hydropower projects. NVE also use it as a cost reference and estimation tool when assessing the economy and budget estimates in license applications.

The cost base contains unit cost estimates for parts and materials for all components of small hydropower schemes, as well as "expert" advice and tips for minimizing costs. It has a range of unit cost equations for various components of hydropower projects, including (NVE 2012; NVE 2016b):

- unit price for dam length based on dam height for various types of dams,
- total price for the intake based on discharge (m^3/s),
- power station (building) costs based on discharge and hydrological head for power stations in the open and underground,
- unit cost per meter canal for rock blasting and in soil based on maximum discharge,
- unit cost per meter tunnel drilling/blasing and shaft drilling based on tunnel cross-section area and shaft diameter,
- generator, transformer and control equipment costs based on active effect (kW) capacity,
- unit costs per meter power line based on mast type, turbine costs per installed effect (NOK/kW) based on discharge capacity differentiated for net hydrological head and turbine type (Pelton, Francis, Kaplan and others),
- unit cost per meter for different penstock types based on penstock diameter and pressure class,
- hatch price based on hatch area differentiated per hatch type,
- cost per installed effect (NOK/kW) for different turbine types and hydrological head for complete set of electro-mechanical equipment based on maximum discharge.

4.2 DATA COLLECTION

The investment cost analysis and cost estimation tool development in this thesis is based upon data material collected from NVEs electronic archive. Four sources of data were used, and the content and considerations for each source are given below.

4.2.1 Budgets and project plans

The data for budgeted costs and other features of the plans for the SHP-plants was reviewed and collected from four types of documents in NVE's electronic archive. These include: NVE's internal license application database, license applications, revisions of the license applications and the final decision documents from NVE.

The dataset was built upon an extract from NVEs license database, which contains main data from license applications. Data used from this source were: name of the project, license ID, county, installed capacity, estimated average annual production, budgeted cost, specific cost, date of budget cost. This was accessed by email correspondence with Erlend Støle Hansen November 4th 2015.

In the initial data collection process, detailed project data from 82 of the total 153 projects were collected from the license documents. Projects had often undergone changes during the license application process. In some cases these changes had not been updated in the license database. For these cases, the budgeted costs, annual production estimate, installed capacity and cost date were cross-checked against the license database, and revised if the database record was not the latest revision.

In cases where NVE had detected large deviations between the budget estimate in the license application and their own estimate, the latter was used.

For the rest of 71 projects, the data was cross-checked and revised if the cost date or other data were missing.

4.2.2 The form for commissioning

Once a hydropower project has been commissioned, the owners are required to submit a form to NVE in which they state the date of commissioning, and details about the physical properties of the scheme, as well as investment costs. This was the primary source on which the cost analysis and cost estimation models were based.

The following data has been collected from the forms: Date of commissioning; installed capacity; annual average production estimate; intake-, waterway-, power station- and total costs; gross head; dam dimensions; waterway type, properties and dimensions; turbine types and -properties.

In 30 out of the 153 forms, the partial costs had not been reported. The accuracy of the reported costs range from no decimals to two decimals (MNOK).

In some cases, the dam/intake, turbine type dimensions were not reported. In such cases, the detail plan for the project was consulted to find these values. This plan the plan must then be accepted by NVE in all SHP projects prior to construction start-up.

4.2.3 Construction start-up date

The reported costs were to be transformed from nominal prices to real prices via index regulation. The construction start-up date has been used as a temporal cost correction point for each project. NVE has a record of these dates in an internal hydropower database. The construction start-up date was missing for 58 out of 153 SHP-plants. For these cases, the date was supplemented by accessing electricity certificate ("green" certificate) applications, in which the owners are required to document the construction start-up date. The construction start-up date could not be found for two projects out of the 153 in the record.

4.2.4 NVEs cost index for hydropower projects

The index regulation of budgeted costs and reported costs was based on the official cost index for hydropower projects which is released annually by NVE (2016a). This is adapted to small-scale hydropower plants, and high-pressure hydropower plants (> 300 m gross head), with the average partial cost shares for each type according to the cost report from NVE (2015c). The cost index values can be found in **Table 14** in Appendix 1: Cost index for small-scale hydropower plants.

4.3 Data handling

The project data was recorded in a single data table in Excel. Once the data collection was finished, the data was loaded into the open source statistics program R. All recoding of variables, as well as variable algebra was done using R. The packages used here include the core stats package (R Core Team 2015) and the coding grammar package `dplyr` (Wickham & Francois 2015). All R coding has been documented in Appendix 5: R code .

All plots produced in R were made using the `ggplot2`-package (Wickham 2009).

4.4 Brief introduction of the statistical methods

The analysis in this master thesis relies upon location tests and multiple linear regression. A brief introduction to the methods used will be presented in the following section.

4.4.1 Two-sample tests

Student's t-Tests (t-test) and Mann-Whitney-Wilcoxon tests (Wilcoxon tests), and variants of these were both used for:

- Two-sample location tests on difference between two independent samples, such as the prediction model estimates and the budget estimates of investment costs
- Paired sample tests on difference between two dependent samples, such as the budget estimates and the reported costs

The t-test is in general valid only under the assumption that the population is normally distributed around the mean (for example the mean difference between budget estimates and actual costs). When dealing with a sample from the population, this is assumed to follow a t-distribution, which is dependent upon the sample size. If the distribution of the sample does not follow the t-distribution, the test is not valid. In order to test for this validity, the Shapiro-Wilks test was used, which is reported to be the most reliable test for normality (Razali & Wah 2011). If the p-value for this test is below the α level of significance, then the null-hypothesis that the observed data is drawn from a normally distributed population is rejected.

The assumption of normality was frequently violated in the tests conducted in this thesis. As a solution to this, non-parametric methods can be more precise and effective when the underlying assumptions are not satisfied for methods based on normal theory (Hollander et al. 2014). The term non-parametric refers to the fact these methods are not relying on assumptions of underlying probability distributions for the population. Therefore, Wilkoxon tests were utilized alongside the t-tests for all location tests. The underlying assumption for the unpaired Wilcoxon test is that the distributions of the two samples are have identical probability functions. For the paired test the assumption is that the distribution of the differences is symmetric.

All the above-mentioned tests were performed as two-tailed hypothesis tests in this thesis, which entails that the null-hypothesis is that the expected values (mean or median) for two independent samples are equal, the difference between two dependent samples equals zero, the mean of the estimated linear regression coefficient equals zero. The alternative hypothesis is that the expected values are not equal, and that the expected difference or estimated coefficient is not zero.

The null hypothesis was rejected if the test statistic had a larger value than the critical value for rejecting the null-hypothesis for the given level of significance. For more details of these tests, see Løvås (2010 Ch. 8.) and Hollander et al. (2014 Ch. 3-5)

A formal set-up of the hypothesis tests is stated below. This is an example of a t-test for paired samples:

$H_0: \mu_d = 0$ vs. $H_1: \mu_d \neq 0$

Where H_0 is the null-hypothesis that the population difference μ_d equals zero, and the H_1 is that the difference is not zero. H_0 is rejected if the test statistic $|T| > t_{\alpha/2}$. The level of significance for all hypothesis tests in this thesis is set to $\alpha = 0.05$.

The t-Test was conducted using the `t.test`-command, the Shapiro-Wilks test using the `shapiro.test`-command and the Wilcoxon test using the `wilcox.test`, all in the `stats` core package (R Core Team 2015).

T-tests are also used frequently in the thesis for the significance of regression model coefficients. Here, the underlying assumption is that the residuals, i.e the difference between the observed sample values and the fitted values, are normally distributed around zero.

4.4.2 Analysis of variance and Kruskal Wallis Rank Sum Test

For instances with more than two groups, tests for analysis of variances (ANOVA) were conducted. This method was also frequently used for assessing linear models and their predictor variables..

The one-way ANOVA tests whether the variance within each group is similar to the variance across all groups. The one-way ANOVA relies upon the assumptions that the within-group standard deviation is equal, and that the observations within each group are normally distributed.

The null-hypothesis is that the expected value is equal for all groups, the alternative hypothesis is that at least one of the group have different expected value compared to the others, stated as:

$H_0: \mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$

H_1: At least one μ differs from the rest

In case the observations within each group are not normally distributed, the Kruskal-Wallis' test can serve as an alternative. This was done in one instance to test for differences between reported investment costs for different counties and regions.

These tests were conducted using the `aov` command and the `kruskal.test`-command in the `stats` core package in R (R Core Team 2015).

4.4.3 Simple linear regression

Linear regression was frequently used as analytical method in this analysis.

The relationship between individual and external factors, namely construction year, construction time, counties and regions, and investment costs was analyzed using simple linear regression.

Linear regression relies upon the method of minimizing the squared error between the estimated values of the fitted regression line and the observed values.

Linear regression, including simple and multivariate regression, relies upon the following assumptions(Hyndman & Athanasopoulos 2013b; Mendenhall & Sincich 2003):

- Linearity between the predictor variable(s) and the response variable
- The random errors are independent
- The random errors are uncorrelated with each other and with the predictor variables
- The random errors have mean zero, are normally distributed, and with a constant variance

Violations of these assumptions may lead to unreliable test results for the model performance, as well as misspecified models and poor prediction/explanation power.

The calculations behind the model metrics will not be presented here, as this is thoroughly explained in several textbooks (Hyndman & Athanasopoulos 2013a; Johnson & Wichern 2007; Løvås 2010; Mendenhall & Sincich 2003).

All regression analyses were done using the lm-command in the stats core package. In all regression analyses, the predictor coefficients ($\hat{\beta}$), their p-values, the residual standard error (s), the multiple R^2, the global F-test and its p-value, were evaluated for each model.

4.4.4 Assessing linear model assumptions

Violations of assumptions of the linear models were assessed by plotting diagnostics plots, testing for normality of residuals using the Shapiro-Wilks test, the Global Validation of Linear Model Assumptions test and by assessing multicollinearity among the predictor variables. These assessments will be introduced briefly.

Diagnostics plots are commonly used for model evaluation. The diagnostics plots used in this analysis includes six separate plots,:

1) Residuals versus fitted values for detecting observations with high residuals, as well as patterns of the residuals, such as heteroscedasticity.
2) Standardized residuals versus theoretical quantiles to assess normality of the residuals,
3) Square root of standardized (absolute) residuals versus fitted values to identify patterns in the residuals.
4) Cook's distance per observation number to identify observations with a large cook's distance.
5) Residuals versus leverage to identify observations with a large influence on the fitted model, i.e., with high leverage and high residuals.
6) Cook's distance versus leverage to identify observations which have a high impact on the fitted model.

In addition to the diagnostics plot, the models were assessed using the Global Validation of Linear Model Assumptions (GVLMA) methodology presented in Peña and Slate (2012). The GVLMA-test evaluates four test parameters, namely;

1) skewness of the residuals of the fitted model,
2) of deviations from kurtosis of the assumed normal distribution of the residuals,
3) whether the link function of the model is misspecified, i.e. lack of linearity between predictor variables and the response variable,
4) heteroscedastic and/or dependent residuals.

Finally, a global test statistics is used to detect whether at least one of the four test-statistics suggests that the linear model assumptions are violated in the current model.

In addition to this model assumption test, a Shapiro-Wilks normality test was applied to the residuals, although this and the GVLMA should yield the same results.

To test for multicollinearity among predictor variables, variance inflation factors were calculated for each model. The variance inflation factor for a given predictor variable increases with correlation between this variable and the remaining predictor variables. According to Mendenhall and Sincich (2003), VIF-values of ten or higher are signs of highly correlated predictor variables.

The diagnostics plots were produced with the `autoplot`-command in the `ggfortify`-package (Horikoshi & Tang 2015).The GVLMA was carried out by the `gvlma`-command from a package with the same name (Peña & Slate 2014). The Shapiro-Wilks-test was conducted using the command `shapiro.test` in the `stats` core package (R Core Team 2015).

If the linear model assumptions were not met by a fitted model, two measures were carried out. In case of heteroscedasticity of residuals, a logarithmic transformation of the response variable was carried out, as this allows for increasing variance with increasing values of the (non-transformed) response variable. (Mendenhall & Sincich 2003 Ch. 7). Consequently, the log-transformed model was tested with regards to violation of model assumptions. The second measure was to delete observations with relatively large residuals. In such cases, it was documented how many observations were deleted in the resulting model.

4.4.5 *Second order multiple regression models*

Second-order (and third-order) models with both quantitative and qualitative variables were tested during the model-fitting procedure in this analysis. These can be challenging to interpret, and therefore a brief introduction has been given. An example model is:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ + \beta_4 x_1 x_2 + \beta_5 x_2^2 + \beta_6 x_2 x_3$$

Where x_1 is a dummy variable $x_1 = \begin{cases} 1 & if\ true \\ 0 & if\ not\ true \end{cases}$, x_2 and x_3 are quantitative variables. $\beta_4 x_1 x_2$ is an interaction term between the dummy variable x_1 and the quantitative variable x_2, meaning that if $x_1 = 0$, the term is also 0: $\beta_4 \times 0 \times x_2 = 0$. β_4 can thus be interpreted as the slope of x_2 given that $x_1 = 1$ (is true). The term $\beta_5 x_2^2$ is a quadratic term which determines the shape of the curve for x_2, for all terms including x_2, when all other x_k are held constant. A positive β_5 gives a parabolic curve for x_2, opening upwards; a negative β_5 gives a parabolic curve for x_2, opening downwards. The term $\beta_6 x_2 x_3$ is an interaction term between two quantitative predictor variables, which is applicable when the relationship between $E(y)$ and x_2 is dependent of the value of x_3, and vice versa. This implies that, when considering a simplified subset model $E(y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_6 x_2 x_3$, that the slope for x_2 is $(\beta_2 + \beta_6 x_1)$ when holding x_3 fixed and the slope for x_3 is $(\beta_3 + \beta_6 x_2)$ when holding x_2 fixed. Thus, the slopes for x_2 changes depending on x_3, and vice versa.

In some cases, the response variable and/or predictor variables were transformed. Transformation of the response variable is described in the previous sub-section. Transformation of predictor variables may also increase prediction performance of the linear regression model.

4.4.6 *Cross validation of models*

The main method of cross-validating the multiple regression models in this paper is by using leave-one-out analysis, a method also referred to as the jackknife. One of the measures using this techniques is called prediction sum of squares (PRESS), suggested first by Allen (1974) for model evaluation. The PRESS is calculated by

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i,-i}\right)^2$$

Where $\hat{y}_{i,-i}$ is the predicted value of y_i when model is fitted excluding observation i, and n is the number of observations in the dataset.

When comparing models with unequal size of the dataset, the standardized measure root mean square error of prediction (RMSEP) may be more appropriate:

$$RMSEP = \sqrt{\frac{PRESS}{n}}$$

The PRESS and RMSEP measures were the main performance indicator used in model evaluation of the cost estimation models. Cross-validation using k-fold cross validation may be a better performance test for the true model performance upon new data, but due to the limited size of the dataset, the leave-one-out procedure was chosen.

4.4.7 Predictive versus explanatory modeling

Multiple linear regression analysis can have two different goals: explaining and predicting. Shmueli (2010 p. 290) states that

> *"Explanatory modeling and predictive modeling reflect the process of using data and statistical (or data mining) methods for explaining or predicting, respectively"*

In the first instance, the goal is to explore the causal relationship between variables and to explain the various predictor variables' effect on the response variable. Regression analysis is a useful tool to test hypotheses on whether and to what extent y is affected by variables $x_1, x_2, \dots, x_k$, based on a pattern suggested by a theoretical framework. On the other hand, regression analysis can also utilized for building prediction models on how the response of y can be predicted by a set of variables. This distinction also has implications for model building. While the "true" model is sought in explanatory modeling, the model with best predictive properties is sought in predictive modeling, and these approaches may often end up with different model results (ibid).

The cross-validation method is, for example, less used for explanatory modeling because it may cost statistical power (ibid p. 297) (which is defined as "the probability of rejecting the null-hypothesis when it is false"(Everitt & Skrondal 2010 p. 334)). Predictive modeling often involves more exploration of possible variable transformations and dimension reduction (such as principle component analysis and - regression) than for explanatory modeling. When transforming variables, this may lead to a better fit of the model and reduce sampling variance, but these measures can make the interpretation of each coefficient's effect on the response more difficult (Shmueli 2010 p. 297).

In explanatory modeling,

> *"variable choice is based on the role of the construct in the theoretical causal structure and on the operationalization itself."* (ibid p. 297),

On the other hand, in predictive modeling

> *"there is no need to delve into the exact role of each variable in terms of an underlying causal structure. Instead, criteria for choosing predictors are quality of the association between the predictors and the response, data quality, and availability of the predictors at the time of prediction, known as ex-ante availability"* (ibid p. 298)

Another important aspect is the role of multicollinearity. In explanatory modeling, the presence of multicollinearity between two or more variables will make it difficult to explain each individual coefficient's effect on the response. In prediction however, the focus lies on the response variable and reduction of prediction error. For this purpose, multicollinearity may not be a problem. According to Spyros et al. (1998 p. 288, as cited in Shmueli 2010 p. 299);

> *"multicollinearity is not a problem unless either (i) the individual regression coefficients are of interest, or (ii) attempts are made to isolate the contribution of one explanatory variable to Y, without the influence of the other explanatory variables. Multicollinearity will not affect the ability of the model to predict."*

There is also an important and distinct difference between explanatory power and predictive power. Explanatory power is measured by R^2 and F-statistics. Predictive power is measured by the selected prediction model's performance on new data. Predictive power is not necessarily inferred from explanatory power, and, in the words of Shmueli (2010 p. 300):

> *"While predictive power can be assessed for both explanatory and predictive models, explanatory power is not typically possible to assess for predictive models*

> *because of the lack of* $\mathcal{F}$ *(the underlying, causal function of* $\Upsilon = \mathcal{F}(\mathrm{X})$*, ed.rem.) and an underlying causal structure. Measures such as* R^2 *and* F *would indicate the level of association, but not causation."*

This fact has an important effect on the set-up of the statistical analysis, and the assumptions made. While this thesis deals with both explanatory models and predictive models, the goals and procedure of the analysis will differ for the two approaches. The first part of the analysis deals with explanatory modeling to find indices for difference in investment costs due to economical, temporal and geographical factors. The latter part deals with predictive modeling, where the goal is to build a statistical prediction model for future projects. Following section 4.4.7, this had implications for how the analysis was conducted in the two parts.

4.5 Documentation of the data analysis

4.5.1 Test for difference between budgeted and reported costs

In this analysis, tests were conducted on the observed difference between the budgeted costs and the reported costs. The budget estimates and the reported costs reported for each projects are assumed to be paired/dependent in these tests, allowing for a comparison of costs within each project, instead of a comparison of the means/medians of all budget estimates and reported costs in the sample. The tests were thus conducted as dependent two-tailed tests for paired samples of whether or not there is a significant difference.

The tests were conducted on both values in millions of Norwegian kroner (MNOK), and on values in percent of budgeted cost. The difference in MNOK is computed as:

$$\mu_{di} = x_{ri} - x_{bi} \ ,$$

where u_{di} is the difference d in MNOK for observation pair i, x_{ri} is the reported cost r for observation i, and x_{bi} is the budgeted cost b for observation i.

The relative difference in percent was computed as:

$$u_{dpi} = \frac{x_{ir} - x_{ib}}{x_{ib}} \times 100\% \ ,$$

where μ_{dpi} is the relative difference d in percent p for observation pair i.

This standardization is relevant because the size and absolute cost of the projects varies from tens to hundreds of MNOK. A difference of one million NOK is less critical in a 70 MNOK project compared to a 30 MNOK project. It is therefore interesting to analyze deviations in both MNOK and percent of budgeted costs.

It is also relevant to test whether the investment costs per production unit have changed, and to analyze whether this data record follows the same pattern of difference in investment costs per production unit as found in Haga and Espegren (2013). Therefore, tests were performed on the difference between budgeted costs per production volume (GWh) estimated in the license application and reported costs per production volume estimated after completion of the projects.

As described earlier, the owners of the power plants report the investment cost of their projects in both partial costs and as a total cost. The total cost is often higher than the sum of the partial costs (as seen in **Figure 3** and **Figure 4**). It is therefore also relevant to look at the difference between the sum of partial costs and the budgeted total costs. The possible sources for these deviations are discussed in section 7.1.

In total, 12 tests were conducted on the differences between the budgeted and reported costs

4.5.2 Tests for the effect of construction year on reported costs

Tests were conducted to observe the difference in investment costs for projects undertaken in different years, in order to analyze the cost development over time. In order to evaluate projects across different scales, two different standardization measures were made. In the first, changes in specific costs were tested, which refers to investment cost per kWh estimated average annual production. The second standardization calculation measured costs per installed production capacity, namely cost per watt.

Specific costs per annual production is the standardization measure used by NVE to compare costs of hydropower projects of different scales. It is useful to evaluate different projects based upon their expected income from power production. When interpreting the development of investment cost per production unit over time, any change can be due to changes in both costs and production volume. In order to isolate costs for the infrastructure of projects across different scales, cost per installed production capacity was also calculated and tested.

The data was first explored graphically. The relationship between construction year and investment costs was then analyzed using simple linear regression with construction year as the predictor variable, and investment cost as the response variable.

Four models were tested: Nominal and real costs per estimated annual production, and nominal and real costs per installed capacity. By testing cost development in both nominal and real values, it was possible to measure whether there was still a temporal effect on investment costs after the investment costs were adjusted for the cost increase for cost components.

Before running the models, the predictor variable was recoded from construction date to construction year, and furthermore, the construction year was set to one for the first construction year recorded in the dataset (2005), two for 2006, and continuing. The accuracy of this recoding measure is discussed in section 7.1.

The predictor variable year is set to numeric, starting at one in 2005 (the first year of construction in the dataset) and increasing by one per year. This facilitates interpretation of the coefficient.

4.5.3 Test for the relationship between construction time and costs

Tests were conducted to measure whether there was a significant trend difference between specific investment costs of SHP plants with differing construction period lengths.

The construction time was calculated as the time difference between construction start date and date of operation.

The data was plotted and examined graphically. The relationship between construction time and total investment costs was analyzed further using simple linear regression.

A non-transformed linear model was fitted at first, but inspection of the diagnostics plot indicated violation of the statistical assumptions of the model (in particular, normality of the residuals). In the second attempt to fit the model, the response variable was log-transformed. Observations with extreme values were removed if needed to meet the model assumptions.

4.5.4 Test for the relationship between geography and investment costs

Tests were conducted to measure whether there was a significant difference between specific investment costs of SHP plants in different regions of Norway for the dataset.

The observed specific investment costs (in real prices) were plotted per county in a box-plot, and grouped by geographical region.

The difference in costs between counties was tested with use of one-way ANOVA and linear regression.

4.5.5 Test for the relationship between license holder type and investment cost

Tests were conducted to measure whether there was a significant difference between specific investment costs of SHP plants and project developers having multiple projects in their portfolio and project developers having only one or a few projects in their portfolio.

The observed specific investment costs (in real prices) were plotted per license holder type in a box-plot.

The difference in costs between holder groups was tested with use of independent two-sample tests. In the t-test, the variance within each sample was assumed to be non-equal.

The classification of owners is given in **Table 15** in Appendix 2: License owner classification

4.5.6 Prediction models for investment costs estimation

In this analysis, the goal was to build a model as accurate as possible using multiple linear regression. As described in section 4.4, the model building strategy for predictive modeling sometimes differs from that of explanatory modeling. We know from previous literature that physical features such as size, structure types, material types and equipment types of small-scale hydropower plants are associated with investment cost. The question is which physical parameters have the highest correlation, the lowest variance, and the best predictability on the investment cost, given the sample of data available for this analysis.

In order to develop the best possible subset model for prediction, all reported physical features of the plants available in the dataset were tested in the full model, and followed by a stepwise selection of predictor variables.

The variables included were: installed capacity, gross head, maximum discharge, dam height, dam length, penstock length, penstock diameter, penstock types, tunnel length, tunnel cross section area, shaft length, shaft cross section area, turbine types, county, construction year, year of operation and construction time, as well as dummy variables for waterway type (penstock, tunnel and shaft). These are all variables that have been used as predictor variables in the studies mentioned in the literature review or in the cost base from NVE (2012), with the exception of county, license owner, year of operation, year of construction and construction time.

The variables are in different scales, ranging from decimals to thousands. Scaling of variables may be useful for easier interpretation of the model coefficients, and to reduce multicollinearity between interaction terms or second order terms. Scaling does not affect the accuracy of the fitting process. Multicollinearity may cause rounding errors in calculation of the model estimates (Mendenhall & Sincich 2003 Ch. 7), but this is neglected in this analysis. As mentioned in section 4.4.7, multicollinearity will not affect the prediction performance of the model (Spyros et al. 1998, as cited in Shmueli 2010).

Attempts were made to build models using scaled variables, but they were abandoned due to the fact that such models are less practical in use when applied to a new dataset.

In the first attempts to build a subset model, only first order terms were included, and no transformations of the variables were made. With no transformation, the model failed to satisfy the model assumptions. Therefore, the response variable was log-transformed.

Attempts were also made to develop a second order model with interaction terms on both quantitative and qualitative variables and squared terms. These included logarithmic and squared terms of all physical sizes, and interaction terms between: installed capacity and gross head; dam height and dam length; penstock diameter and penstock length; penstock types, penstock length and penstock diameter; tunnel binary variable, tunnel length and tunnel cross section area; shaft binary variable, shaft cross section area, and shaft length, construction time and construction year.

At first, the model selection relied upon two-way stepwise selection based on AIC, using the stepAIC-function in the `MASS` package (Venables & Ripley 2002). However, when cross-checked against manual model selection based on PRESS evaluation, the `stepAIC` overfit the model, leading to higher PRESS-values than the manual selection, despite the findings in the literature that AIC is the asymptomatic equivalence to PRESS in model selection. AIC tend to have a bias towards overfitting when the sample size is low. According to (Burnham & Anderson 2004) the AICc measure should be used in model selection where n/K is smaller than about 40, where K is the number of predictor variables in the full model. This unfortunately came to the author's knowledge at a late stage in the model selection process.

There exist packages for unsupervised model selection based on AICc, such as the `MuMLn` and the `AICcmodavg`, but due to time constraints, these were not utilized.

The model performances were evaluated with regards to PRESS, using the `press`-command in the `DAAG` package, and for every performance test violation of model assumptions were also tested.

If the test indicated that the assumptions were not acceptable, the model was revised. Deleted observations were documented.

Multicollinearity among the predictor variables was also assessed, using the `vif`-command in the `car` package (Fox & Weisberg 2011).

Several attempts were made using the stepAIC-function, and thereafter using manual backward selection based on PRESS/RMSEP. The number of trials was not counted, but would probably sum up to 200-400 or more. The subset-models reported here is thus a result of both subjective assessments of the model variables, mixed with the pure quantitative approach using stepAIC.

Two models are reported in this chapter. The first is based on total investment costs, and the second is based on sum of partial costs. In both models, projects with tunnels have been excluded due to the fact that such projects tend to have significantly higher costs than those without tunnels, and this lead to lower model accuracy when those were included. The total costs, as reported from the owners, may include external costs having little to do with the projects themselves (for example grid connection fees). The sum of reported partial costs are likely to be more directly correlated to the physical properties of the hydropower projects. It is still relevant to observe the extent to which these factors affect the accuracy of the models, and therefore both have been included.

It should be noted, that in the model selection, factors such as license holder class and estimated annual production in some cases added prediction power to the model. In spite of this, these factors were left out. It is difficult to interpret the effect of license ownership on future projects. There may be multiple factors leading to significantly lower costs for non-professional license owners. It may be due to non-reported ower efforts in the construction period. It may also be the case that non-professional owners keep their water resource and develop it at their own hands because the water resource is of a high quality and easy to exploit. In such cases, the difference between non-professionally owned projects and the professionally owned projects in part is due to the quality of the water resource. The annual production estimate introduces another source of uncertainty when used for prediction of future projects.

Once the best subset model was selected, a leave-one-out cross-validation was carried out using the `cv.lm` command in the DAAG package (Maindonald & Braun 2015). This cross-validation package is designed for K-fold cross validation, but if the number of folds is set to the number of observations, it is equivalent to a leave-one out cross validation. This command returns the leave-one-out prediction for each observation, which was used as a performance indicator for the model's prediction performance compared to the budget estimates.

The model performance was sensitive to deletion and adding of observations for both models.

The difference between the cost estimates and the reported costs was calculated as the mean absolute error rate (MAER) between estimates and real costs, according to the formula below

$$MAER = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|C_{ri} - C_{pi}\right|}{C_{ri}}\right) \times 100\%,$$

Where C_{ri} is the reported r cost C for observation i, the C_{pi} is the cross-validated prediction p cost for observation i, for n number of observations. This error measurement is equivalent to the one used in Gunduz and Sahin (2015), Kim et al. (2004), Smith and Mason (1997).

The performance of the best subset models compared to the budget estimates was plotted and analyzed graphically and tested numerically using two-sample tests for difference.

The model predictions, the real budget estimates, and the real reported costs were all plotted in the same graph, with the observations sorted by increasing reported costs. This was done to provide a visual representation of the model results. A second graph was produced where the real reported costs were set to zero, and the real budget estimates, and the model predictions were plotted as deviations from the true costs. This was done to enable a closer graphical analysis and comparison of the residuals of the estimates.

A chi-squared test was also conducted to determine whether there was a significant pattern of simultaneous over- and underestimation for both budget- and model estimates.

5 THE DATASET

The data collection resulted in a dataset of 153 cases with budgeted and reported total costs for SHP projects. The variables that have been used in this analysis are listed in **Table 4**. Some variables do not have a full record due to missing information from the forms from which the data was collected.

Table 4: Selection of variables recorded in the dataset.

Full variable name	Coded variable name	n
License registration number	Kdb_ID	153
Power plant name	Title	153
License status date	Main_Status_date	153
License holder	Holder	153
County	County	153
Planned effect	Effect	153
Planned annual average production	Production	153
Budgeted cost	Est_Cost	153
Budgeted specific cost	Spec_Cost	153
Date of budgeted cost	Cost_Date	153
Reported date of operation	Date_Operation_R	153
Reported Installed capacity	Max_Effect_R	153
Reported annual production estimate	Ann_Prod_Est_R	153
Reported Dam and intake costs	Inlet_Cost_R	123
Reported waterway costs	Penstock_Costs_R	123
Reported power station costs	PP_Costs_R	123
Reported total costs	Total_Costs_R	153
Reported gross head	Gross_Head_R	153
Reported dam height 1	Dam_1_Height_R	151
Reported dam length 1	Dam_1_Length_R	151
Reported dam height 2	Dam_2_Height_R	12
Reported dam length 2	Dam_2_Length_R	12
1st penstock section length	Penstock1_ Length_R	152
1st penstock section diameter	Penstock1_Dia_R	151
1st penstock type	Penstock1_Type_R	151
2nd penstock section length	Penstock2_ Length_R	71
2nd penstock section diameter	Penstock2_Dia_R	71
2nd penstock type	Penstock2_Type_R	69
Reported tunnel length	Tunnel_Length_R	23
Reported tunnel cross section area	Tunnel_Cross_Sect_R	23
Reported shaft length	Shaft_Length_R	13
Reported shaft cross section area	Shaft_Cross_Sect_R	13
Reported turbine 1 type	Turbine1_Type_R	147
Reported turbine 1 power capacity	Turbine1_Effect_R	152
Reported turbine 1 maximum discharge	Turbine1_Abs_Cap_ R	152
Reported turbine 2 type	Turbine2_Type_R	31
Reported turbine 2 power capacity	Turbine2_Effect_R	32
Reported turbine 2 maximum discharge	Turbine2_Abs_Cap_R	32
Construction start date	Construction_Date	151

Figure 3 and **Figure 4** show reported investment costs and reported specific costs for all observations available in this dataset, in real costs adjusted to the 2015 index.

What these summary statistics tell us:

- There are some large deviations between sum of partial costs and the total costs for several projects. In some cases this is most likely because the owners have to pay a grid fee ("anleggsbidrag") to the distribution system operator for improving the electricity line in the area. In other cases it may be that the owners have excluded miscellaneous and administrative costs from the partial costs. This is uncertain, and a challenging source of error.
- Projects with tunnels have in most cases substantially higher total costs than projects with other waterway types, as seen in **Figure 3**. At the same time, projects with tunnels and shafts are fairly equally distributed in the range of specific costs seen in **Figure 4**, which means that projects with tunnels and shafts often have higher annual production.

- There are relatively few observations of projects with tunnels and shafts.

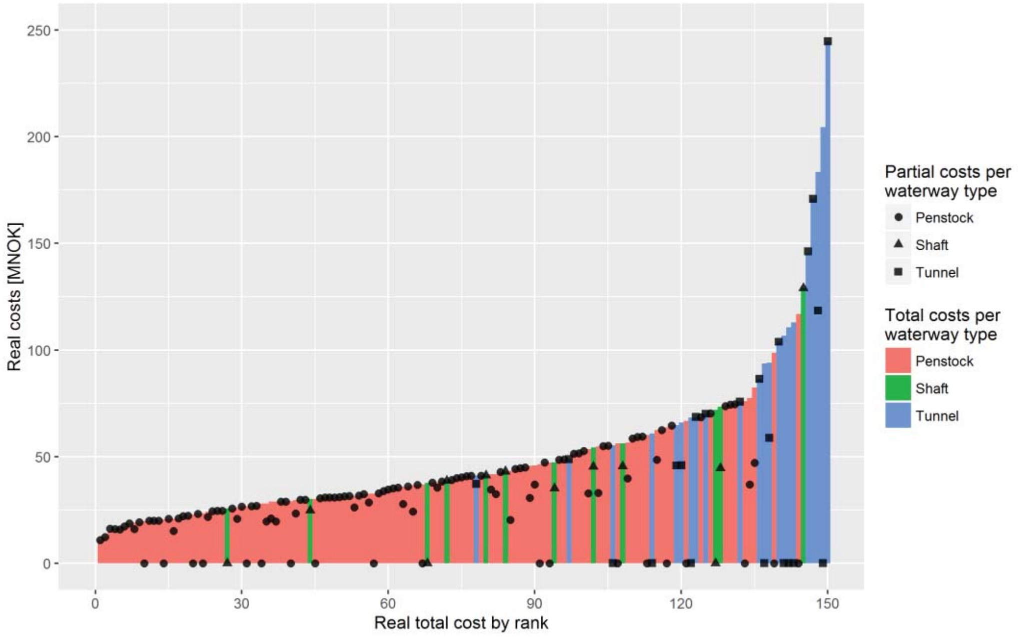

Figure 3: Real total investment costs in the dataset, stacked by investment cost. Colored bars show total costs by their main waterway type. The points show sum of partial costs by waterway type. Partial costs set to zero when not reported.

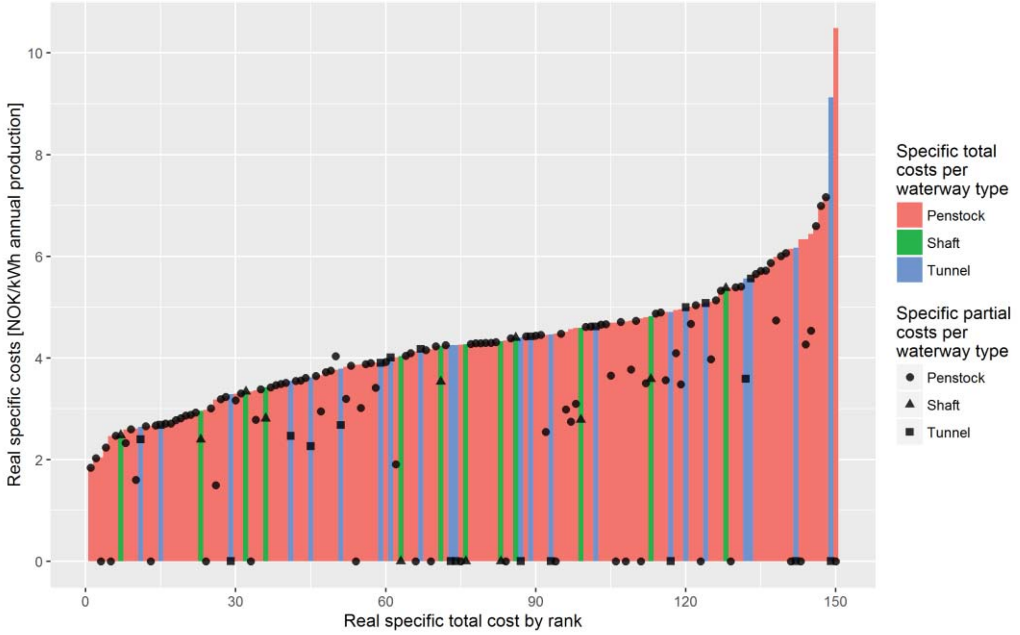

Figure 4: Real specific total investment costs in the dataset, stacked by specific total cost. Colored bars show specific costs by their main waterway type. The points show specific sum of partial costs by waterway type. Partial costs set to zero when not reported.

Table 5 summarizes the investment cost data for the total costs and sum of partial costs, grouped by waterway type.

Table 5: Real investment costs grouped by waterway type.

Cost components	Water-way Type	Min	1st Quant.	Median	Mean	3rd Quant.	Max	n (ex NAs)	NAs
Real Total Costs	Penstock	10.83	26.75	34.28	38.66	47.20	116.79	116	2
Real Total Costs	Shaft	25.60	36.55	45.14	53.29	60.14	129	12	0
Real Total Costs	Tunnel	40.95	66.56	90.01	103.47	112.33	244.77	22	1
Real Total Costs	Total	10.83	28.87	39.56	49.34	59.02	244.77	150	3
Real Partial Costs	Penstock	10.83	23.36	31.74	34.85	40.95	74.57	97	21
Real Partial Costs	Shaft	24.73	38.70	43.04	49.66	45.23	129	9	3
Real Partial Costs	Tunnel	37.16	51.28	72.95	94.40	114.85	244.77	14	9
Real Partial Costs	Total	10.83	25.47	35.33	42.91	47.52	244.77	120	33
Rel. Intake Cost	Penstock	2.0%	7.5%	10.6%	12.7%	16.7%	37.5%	98	20
Rel. Intake Cost	Shaft	3.2%	6.2%	7.8%	8.0%	8.9%	16.7%	9	3
Rel. Intake Cost	Tunnel	4.6%	6.3%	9.1%	11.6%	12.5%	28.8%	15	8
Rel. Intake Cost	Total	2.0%	7.2%	10.3%	12.2%	15.8%	37.5%	122	31
Rel. Waterway Cost	Penstock	13.8%	28.3%	34.0%	35.8%	41.6%	62.5%	98	20
Rel. Waterway Cost	Shaft	25.4%	35.2%	38.9%	42.3%	46.8%	60.8%	9	3
Rel. Waterway Cost	Tunnel	16.6%	31.7%	38.3%	39.3%	46.7%	58.2%	16	7
Rel. Waterway Cost	Total	13.8%	29.6%	35.7%	36.7%	42.5%	62.5%	123	30
Rel. Station Costs	Penstock	7.3%	37.8%	45.1%	45.9%	55.0%	70.9%	98	20
Rel. Station Costs	Shaft	24.6%	31.4%	37.9%	36.5%	39.3%	46.8%	9	3
Rel. Station Costs	Tunnel	25.7%	29.6%	37.7%	41.1%	53.4%	72.0%	16	7
Rel. Station Costs	Total	7.3%	36.0%	43.0%	44.6%	54.6%	72.0%	123	30
Rel. Partial Costs	Penstock	47.0%	100%	100%	94.4%	100%	107.1%	98	20
Rel. Partial Costs	Shaft	60.7%	80.8%	83.4%	86.8%	100%	100%	9	3
Rel. Partial Costs	Tunnel	62.5%	80.7%	100%	90.6%	100%	100.8%	15	8
Rel. Partial Costs	Total	47.0%	91.4%	100%	93.4%	100%	107.1%	122	31

All relative costs are based on nominal costs, and are calculated as percent of total costs

Figure 5 shows the average share of partial costs for the SHP projects in this dataset and provides a graphical summary of cost component characteristics of the dataset.

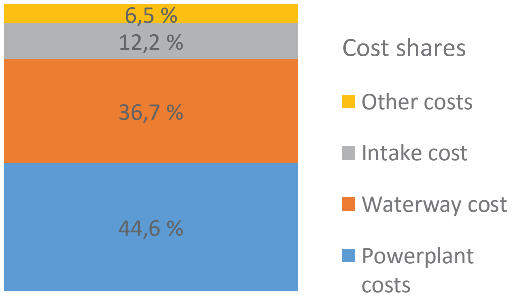

Figure 5: Average shares of partial costs per total costs for SHP projects in the dataset. 'Other costs' indicates the difference between sum of partial costs and the total reported costs. Differences were calculated based on nominal costs.

6 Results

6.1 Budgeted vs. actual, reported costs

The results from these tests, given in **Table 6**, show a significant difference between budgeted and reported costs in nearly every case. The table shows test results from twelve different comparisons, each tested with two different methods, namely the two-sample, two-tailed paired t-test and Wilcoxon signed rank test with continuity correction. All valid tests show a significantly larger reported cost than budgeted cost. The Student's t-Test of difference between the sum of real partial costs and the real budgeted costs is not significant, but the Shapiro-Wilks test on normality of the differences indicates the normality assumption is violated. When standardized, the relative difference is significantly larger than zero for the same comparison, and this test is valid with respect to the normality assumption.

Table 6: Statistical tests for difference between budgeted and reported costs. One-sample t-Test and Wilcoxon ranked sum test. Output in bold for t-tests indicate valid test (i.e. the assumption of normality is satisfied according to the Shapiro-Wilks test), otherwise the Wilcoxon test output is in bold.

Type of difference tested		Two-sample paired T-test (two-sided)					Two-sample paired Wilcoxon test (two-sided)			
	n	Mean diff.	Lower 95% CI	Upper 95% CI	p-value	Shapiro p-value	Median diff.	Lower 95% CI	Upper 95% CI	p-value
Difference between nominal total and budgeted costs	153	14 MNOK	11.2 MNOK	16.9 MNOK	<2E-16	7.00E-14	**11.6 MNOK**	**9.6 MNOK**	**13.7 MNOK**	**< 2.2E-16**
Relative difference between nominal total and budgeted costs	153	53.8%	45.8%	61.8%	<2E-16	1.00E-08	**49.1%**	**42.2%**	**56.6%**	**< 2.2E-16**
Difference between real total and budgeted costs	150	8.07 MNOK	5.11 MNOK	11.0 MNOK	3.00E-07	2.00E-13	**6.76 MNOK**	**4.87 MNOK**	**8.96 MNOK**	**1.00E-11**
Relative difference between real total and budgeted costs	150	24.2%	18.7%	29.6%	4.00E-15	4.00E-04	**21.6%**	**16.4%**	**27.1%**	**4.00E-14**
Difference between nominal sum of partial and budgeted costs	123	8.69 MNOK	5.94 MNOK	11.4 MNOK	6.00E-09	7.00E-10	**7.15 MNOK**	**5.4 MNOK**	**9.2 MNOK**	**2.00E-12**
Relative difference between nominal sum of partial and budgeted costs	123	38.8%	30.5%	47.1%	1.00E-15	4.00E-04	**35.7%**	**27.3%**	**44.2%**	**1.00E-14**
Difference between real sum of partial and budgeted costs	121	2.08 MNOK	-1.06 MNOK	5.23 MNOK	0.2	3.00E-10	**2.55 MNOK**	**0.518 MNOK**	**4.65 MNOK**	**0.01**
Relative difference between real sum of partial and budgeted costs	121	**12.8%**	**6.51%**	**19.0%**	**1.00E-04**	**0.2**	**11.20%**	**4.83%**	**17.9%**	**5.00E-04**
Difference between nominal specific total and specific budgeted costs	153	1.2 NOK	1.02 NOK	1.39 NOK	3.00E-09	3.00E-10	**1.12 NOK**	**0.969 NOK**	**1.29 NOK**	**<2E-16**
Difference between real specific total costs and specific budgeted costs	150	0.734 NOK	0.536 NOK	0.932 NOK	1.00E-11	3.00E-06	**0.699 NOK**	**0.514 NOK**	**0.869 NOK**	**1.00E-11**
Difference between nominal specific sum of partial and budgeted costs	123	**0.822 NOK**	**0.647 NOK**	**0.996 NOK**	**7.00E-16**	**0.6**	**0.803 NOK**	**0.626 NOK**	**0.987 NOK**	**1.00E-13**
Difference between real specific sum of partial and budget costs	121	**0.308 NOK**	**0.0992 NOK**	**0.517 NOK**	**0.004**	**0.8**	**0.317 NOK**	**0.113 NOK**	**0.537 NOK**	**4.00E-03**

Nominal costs = not index adjusted, real costs = index regulated to 2015, Specific costs = NOK/kWh estimated average annual production, n = sample size, CI = confidence interval of the estimates, Shapiro p-value = p-value of Shapiro-Wilks-test on normality of differences between observation pairs

6.2 Four selected external cost-driving factors

6.2.1 Construction year and costs

Figure 6 presents the cost development of the sample of small hydropower projects in this analysis over time. The figures are in specific costs. The figure shows boxplots for the four different cost measures : nominal specific partial costs; nominal specific total costs; real partial costs; and real total costs. By graphic inspection, there appears to be a trend of increasing costs, although projects in 2014 have lower costs than the previous two to three years.

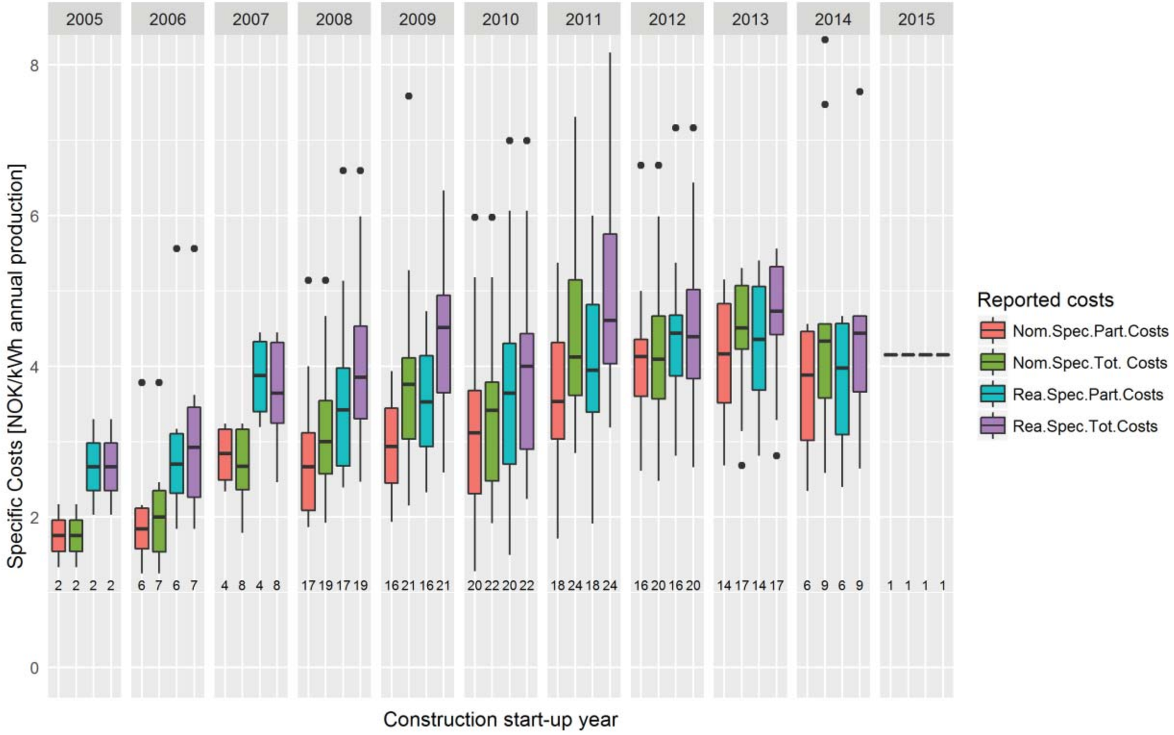

Figure 6: Specific investment costs in real values for power plants in this analysis. Four different series are shown per year: nominal specific partial costs, nominal specific total costs, real partial costs, and real total costs. Three observations were omitted due to extreme values (Specific cost higher than 8 NOK/KWh). Same number of observation per series as in **Table 6**. The colored boxes show the range of the interquartile (from the lower 25% quartile to the upper 75% quartile), the black line within the boxes is the median, and the whiskers extend up/down 1.5 times the range of the inter quartile, dots are cases outside of this range. Number of cases per group below each boxplot.

The trend of increased costs, in both nominal and real costs, is explored further with use of linear regression. **Table 7** shows the R output for fitted regression models with construction year as the predictor variable.

Both models on specific costs have significant p-values at $\alpha = 0.05$ in the t-statistics for the intercepts and $\hat{\beta}$s, and F-statistics for the models.

As expected, the Multiple R^2 is higher in the model with nominal costs ($R^2 = 0.301$) than on the model with real costs ($R^2 = 0.0822$). The $\hat{\beta}$-coefficient for construction start year is higher for Model 1 with nominal costs ($\hat{\beta} = 0.251$) than Model 2 with real costs ($\hat{\beta} = 0.130$). The difference between the coefficients is significant with a p-value of < 2E-16 when tested with a two-tailed independent two-sample t-test.

In order to satisfy the assumptions of normality of residuals, three and four observations were omitted in model 1 and model 2 respectively. In the resulting models, the Shapiro-Wilks test for normality in the residuals gave p-values of 0.1 and 0.09 respectively, and graphical assessment of the diagnostics plots indicated no severe violations of the linear model assumptions.

The slope in Model 1 corresponds to an average annual increase of 10.85 pp (percentage points) in nominal costs per production unit from 2005 to 2015 (with 2005 as base year = 100%). The slope in Model 2 corresponds to an average annual increase of 3.74 pp in real costs per production unit from 2005 to 2015 (with 2005 as base year = 100%). In comparison, the cost index increases by 5.2 pp on average when 2005 is set as base year =100%.

When fitting Model 3 and 4, the response variable had to be log-transformed in order to satisfy the linear model assumptions. Five were deleted in Model 3, with nominal costs per MW ≥ 20. Six observations were deleted in Model 4, with real costs per MW ≥ 25.

Year model 3 has a significant F-statistic, with a p-value of 0.0005. Year model 4 does not yield significance for the global F-statistic, with a p-value of 0.26. This indicates the index adjustment of the costs cancels out the cost increase, and Year model 4 can explain no excessive growth in costs per installed capacity based on the dataset.

In Model 3 with nominal costs, the $\hat{\beta}$-coefficient for construction year is significant, with a coefficient of 0.05 per year. The multiple R^2 is 0.15. All linear model assumptions were met for both models, assessed by visual inspection of diagnostics plots and by the Global Validation of Linear Model Assumptions tests.

The slope in Model 3 corresponds to a cost increase of 5.87% per year with 2005 as base year (= 100%). Since the function here is non-linear, the increase in percentage points per year is not constant.

Table 7: Four regression models on investment costs and year of construction.

Year model 1: Nominal specific total costs.	Estimate	Std. Error	Pr(>|t|)
Intercept	2.06	0.209	< 2e-16
Construction Year	0.251	0.0317	6.4E-13

Construction year 2005 = 1. Residual standard error: 0.860 on 145 degrees of freedom. R^2 = 0.301. Global F-test: p-value: 6.35e-13. Original n = 150, 3 observations deleted due to extreme values, ≥ 6.

Year model 2: Real specific. total costs.	Estimate	Std. Error	Pr(>|t|)
Intercept	3.35	0.239	< 2e-16
Construction Year	0.130	0.0362	4.5E-04

Construction year 2005 = 1. Residual standard error: 0.982 on 144 degrees of freedom. R^2 = 0.0822. Global F-test: p-value: 4.51E-04. Original n = 150, 4 observations deleted due to extreme values, ≥ 6.99.

Year model 3: Nominal costs per kW.	Estimate	Std. Error	Pr(>|t|)
Intercept	2.10	0.0668	< 2e-16
Construction Year	0.0503	0.0102	2.1E-06

Log-transformed response variable. Construction year 2005 = 1. Residual standard error: 0.274 on 143 degrees of freedom. R^2 = 0.146. Global F-test p-value:2.07e-06. Original n = 150, 5 observations deleted due to extreme values, ≥ 20 MNOK/MW.

Year model 4: Real costs per kW.	Estimate	Std. Error	Pr(>|t|)
Intercept	2.49	0.0655	<2e-16
Construction Year	0.0113	0.00993	0.257

Log-transformed response variable Construction year 2005 = 1. Residual standard error: 0.264 on 142 degrees of freedom. R^2 = 0.00905. Global F-test: p-value: 0.257. Original n = 149, 6 observations deleted due to extreme values, ≥ 25 MNOK/MW

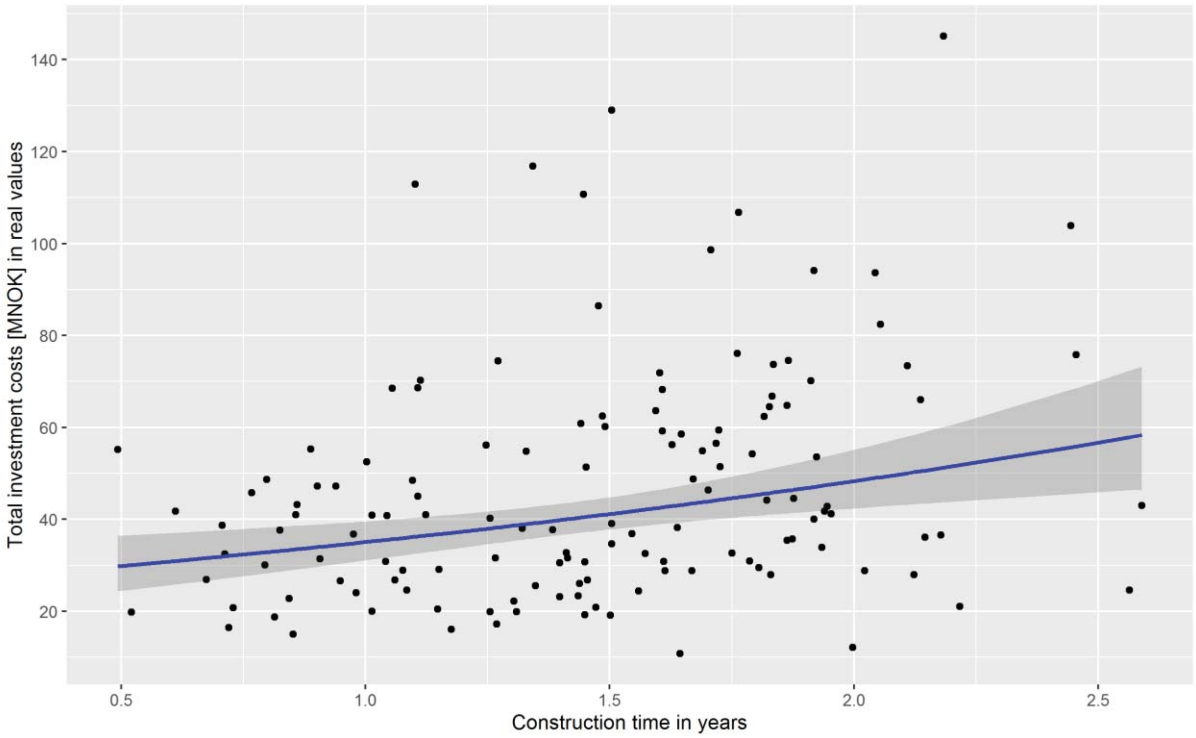

Figure 7: Total costs in real values and construction time with fitted linear regression line in blue, with 95% confidence interval for the expected prediction values in the shaded ribbon. 10 observations omitted: 6 > 150 MNOK investment cost, and 5 > 3 years construction time.

6.2.2 Construction time and costs

Construction time may affect project costs. A linear regression model was fitted to explore this effect. As seen in the scatter plot in **Figure 7**, there is a weak trend of increasing costs with increased construction time.

A non-transformed linear model was fitted at first, but when tested using the Global Validation of Linear Model Assumptions test, and the Shapiro-Wilks test on normality of residuals the tests indicated violation of the statistical assumptions of the model. Inspection of the diagnostics plot gave the same impression. In the second attempt to fit a model, the response variable was log-transformed. A Shapiro-Wilks test on normality of residuals (p-value of 0.3), and a visual inspection of the diagnostics plot indicated that the statistical assumptions of the regression model with log-transformed response variable were satisfied. The GVLMA-test indicated model assumptions were satisfied. The R output is shown in **Table 8**.

The fitted regression model indicates a very weak, but still significant correlation between construction time and reported investment costs, with $\hat{\beta}_{construction\ time} = 0.32$ and $R^2 = 0.079$. The coefficient for construction time translates into an estimated cost growth rate per year of 37.8%. The weak correlation indicates the model lacks predictor variables.

Table 8: Regression models on construction time (in years) as predictor for real total costs.

Construction time model.	Estimate	Std. Error	Pr(>|t|)
Intercept	3.24	0.145	< 2e-16
Construction time	0.321	0.0945	9.2E-04

Response variable log-transformed Construction time in years. Residual standard error: 0.499 on 134 degrees of freedom. R^2 = 0.079. Global F-test: p-value: 9.20E-04. Original n = 146, 10 observations deleted due to extreme values: 6 observations ≥ 150 MNOK total investment costs, 4 observations ≥ 3 years construction time

6.2.3 Geography and costs

Costs of installing small hydropower plants vary due to geographical features. This can be due to socio-economic factors, as well as purely physical factors. **Figure 8** shows specific investment cost in real values grouped by county.

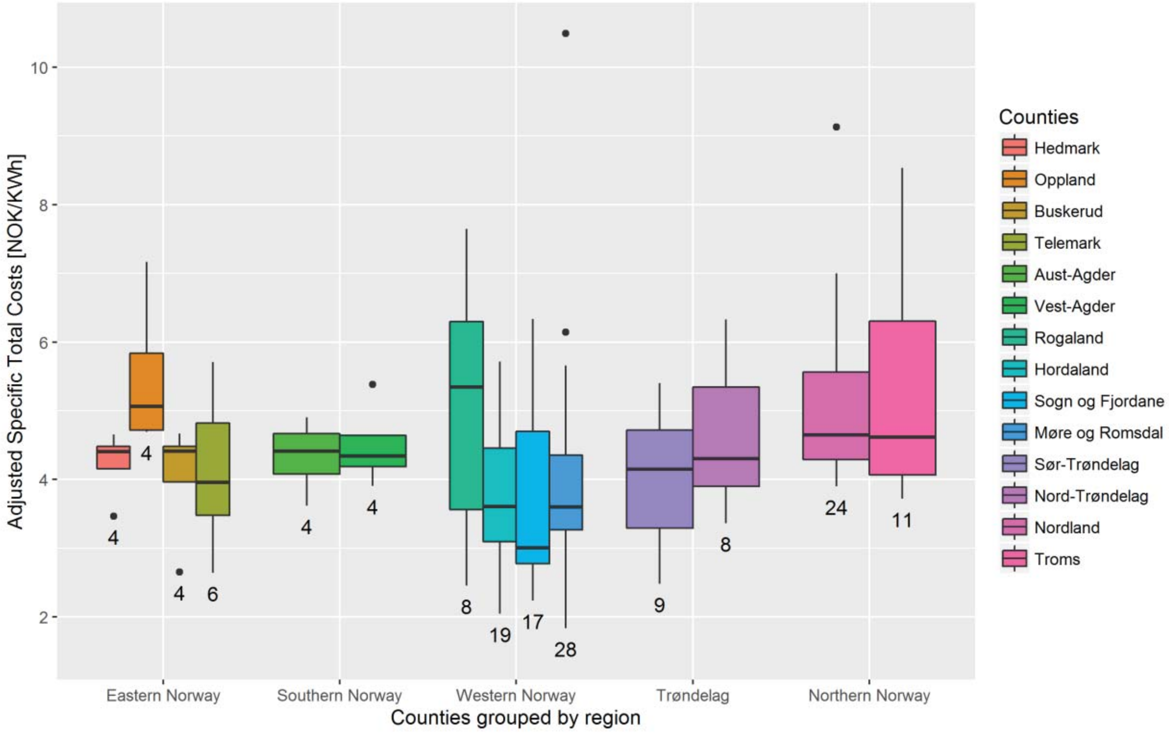

Figure 8: Boxplot with specific investment costs in real values for SHP projects per county, grouped by region. The colored boxes show the range of the interquartile (from the lower 25% quartile to the upper 75% quartile), the black line within the boxes is the group median, and the whiskers extend up/down 1.5 times the range of the inter quartile, dots are cases outside of this range. Number of cases per group is indicated below each boxplot.

To check whether there is a significant difference between costs among different counties and regions, Kruskal Wallis Rank Sum Tests and one-way ANOVA tests were carried out. Two test-rounds were conducted: one with no removal of observations, and a second with removal of observations with total specific investment costs equal to or higher than 6.5 NOK/kWh (nine observations, with an original n of 150).

For counties in the first test-round, the Kruskal-Wallis test on counties gave a p-value of 0.004. The ANOVA gave a p-value of 0.0073, but the Shapiro-Wilks test on normality of residuals indicated that this assumption was violated with a p-value of less than 0.0001.

For regions in the first test round, the Kruskal-Wallis test indicates a significant difference. The test gave a p-value below 0.0001. The ANOVA gave a p-value of 0.0013, but the Shapiro-Wilks test on normality of residuals indicated that this assumption was violated with a p-value below 0.0001.

In the second round two linear models were fitted, one with counties as predictor and one with regions as predictor, both with specific total costs as response. In addition to deleting observations with specific costs equal to or higher than six, the group with the lowest mean cost (in this case the County of Sogn og Fjordane and the region of Western Norway) was set as the first factor, ie. the base response for the model. In this case, all effect sizes relate to the group with the lowest mean, such that the effect sizes, and the p-values of each factor relate to the lowest group mean. The differences among the other factors may not be significant.

When linear models for regions and counties were fitted with the default factor, (in this case Hedmark and Eastern Norway, which both have means close to the mean investment cost (4.14 NOK/kWh) across all projects), the effect sizes were not significant.

For counties in the second round, the regression model gave a significant p-value for the global F-test, with a value of 0.0802, residual standard error of 0.985 and a multiple R^2 of 0.166. Nearly half of the counties had significant effect sizes, including: Intercept (Sogn og Fjordane) with a mean of 3.71, Oppland + 0.52, Nord-Trøndelag + 0.93, Nordland + 1, and Troms +0.968. All values were given in NOK/kWh. All model assumptions were satisfied according to the GVLMA-test, the Shapiro-Wilks test on normality of residuals, and inspection of the diagnostics plot.

For regions in the second round, the global F-test of the regression model was significant with a p-value of 0.0007, residual standard error of 0.972 and a multiple R^2 equal to 0.131. The intercept (Western Norway) had a mean of 3.79 and the region of Northern Norway had a mean of + 0.912, both significant. The other three regions did not have significant effect sizes, but had all p-values below 0.1, with the region of Trøndelag + 0.518, Eastern Norway + 0.478, Southern Norway + 0.623. All model assumptions were satisfied according to the GVLMA-test, the Shapiro-Wilks test on normality of residuals, and inspection of the diagnostics plot.

There are relatively few observations in the dataset from projects in Southern Norway, Trøndelag and Eastern Norway, and most counties here have group means close to the total mean. This means more data would be required in order to test whether significant differences may occur.

6.2.4 License holder and cost

It is relevant to investigate whether or not the companies that have specialized in developing small-scale hydropower projects are able to realize their projects at lower costs than companies organized by land owners. The boxplot in **Figure 9** shows investment costs for the holders classified as professional developers, versus the "non-professional" project developers, i.e. property owners who own the license and the power plant themselves. The figure indicates, quite surprisingly, that the professional developers have higher investment costs per kWh than the non-professionals.

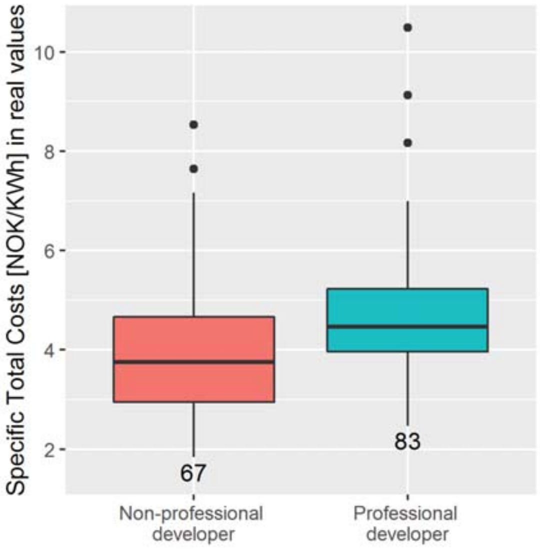

Figure 9: Boxplot with specific total investment costs in real values, by owner type. Number of cases for each group summarized at bottom of graph. The colored boxes show the range of the interquartile (from the lower 25% quartile to the upper 75% quartile), the black line within the boxes is the group median, and the whiskers extend up/down 1.5 times the range of the inter quartile.

The mean of the non-professional group is 3.95 NOK/kWh while the professional group averaged 4.71 NOK/kWh, giving a difference of 0.76 NOK/kWh. The median of the non-professional group is 3.75 NOK/kWh and of the professional group 4.46 NOK/kWh, giving a difference of 0.714 NOK/kWh. To check for a significant difference between the expected values (mean and median) of the two groups, two independent two-sample, two-tailed tests were carried out. The t-test gave a 95% confidence interval of 0.336 to 1.20 NOK/kWh of the difference between the professional-group and the non-professional group, with a significant p-value of 0.0006. The Wilcoxon rank sum test gave an estimated difference of 0.787 NOK/kWh, with a 95% confidence interval of 0.408 to 1.158 MNOK/kWh, and a p-value of 0.0001. Both tests suggest there is a statistically significant difference between the expected values for two groups.

The difference is also significant for both tests when applied to total investment costs per MW installed capacity.

6.3 Prediction models for investment costs

An extensive model selection procedure was carried out to find an appropriate subset model for investment cost prediction.

The full model for partial costs, including all first order terms and selected second order terms and interaction terms, yielded a multiple R^2 of 0.889, a residual standard error of 0.217 (on logarithmic scale), and a significant global F-statistic, with a p-value below 0.0001. The PRESS and RMSEP were both infinite. The model coefficients for the full model will not be given here.

6.3.1 Prediction model 1: Total costs

The final subset-model had the function given in **Equation 1**, and its metrics are given in **Table 9**. It is a model with seven predictor variables. Installed capacity has one first order term and a squared term, with positive and negative coefficients respectively, giving a downward-facing parabolic curve when isolated against the log-transformed response variable. The dam height term and the construction time terms are first-order terms which indicate a linear relationship between the log-transformed response variable and these predictor variables (when all other terms are kept constant). The weighted average penstock diameter term is squared, with no first-order term included. This can be justified by the fact that the cross-section area of the penstock may be a more applicable predictor than just its diameter. It was tested whether the true area of the inner penstock cross section gave a different model result by replacing the square term by an exact area term $\left(\frac{penstock\ diameter}{2}\right)^2 \times \pi$, but this gave no change in the model output, so they can be regarded as equivalent in this context. The log-transformed waterway length term can be interpreted as displaying a linear relationship between log-transformed costs and log-transformed waterway length when all other variables are kept constant. The dummy variable of shaft can be interpreted as an increase in average costs from when the SHP plant has shaft as part of the waterway.

$$\begin{aligned}\log\big(total\ costs\ (real)\big) &= \hat{\beta}_0 + \hat{\beta}_1 \times Installed\ capacity\ [MW] + \hat{\beta}_2 \times (Installed\ capacity\ [MW])^2 \\ &+ \hat{\beta}_3 \times Dam\ heigth\ [m] + \hat{\beta}_4 \times \log(Waterway\ Length\ [m]) + \hat{\beta}_5 \\ &\times (Weighted\ average\ penstock\ diameter\ [m])^2 + \hat{\beta}_6 \\ &\times Construction\ year\ [years, 2005\ =\ 1] + \hat{\beta}_7 \times Construction\ time\ [years] \\ &+ \hat{\beta}_8 \times Shaft(TRUE)\end{aligned} \qquad \textbf{(1)}$$

Table 9: Prediction model 1. Model data for best subset model for predicting (log-transformed) total costs.

Predictors					Data range	
	$\hat{\beta}$ *no.*	$\hat{\beta}$-estimate	$\hat{\beta}$-SE	p-value	Min	Max
Intercept	$\hat{\beta}_0$	1.03E+00	3.68E-01	5.90E-03		
Installed capacity [MW]	$\hat{\beta}_1$	4.55E-01	9.26E-02	3.10E-06	1.20	5.60
(Installed capacity [MW])^2	$\hat{\beta}_2$	-3.51E-02	1.32E-02	9.00E-03	1.20	5.60
Dam height [m]	$\hat{\beta}_3$	1.13E-02	6.03E-03	6.46E-02	0.0	32.0
(Weighted average penstock diameter [m])^2	$\hat{\beta}_4$	1.50E-01	3.05E-02	3.10E-06	0.20	2.10
log(Waterway Length [m])	$\hat{\beta}_5$	1.39E-01	5.16E-02	8.20E-03	170	4948
Construction year [years, 2005 = 1]	$\hat{\beta}_6$	1.77E-02	9.47E-03	6.47E-02	2005	2015
Construction time [years]	$\hat{\beta}_7$	1.42E-01	3.00E-02	6.70E-06	0.49	4.56
Shaft(TRUE)	$\hat{\beta}_8$	1.41E-01	6.72E-02	3.78E-02	0	1

Response variable log-transformed using natural logarithm.
Residual standard error = 0.213, 111 degrees of freedom, $R^2 = 0.779$, adjusted $R^2 = 0.763$,
F-statistic 48.9on 8 and 111 DF . p-value < 2E-16. PRESS = 5.93, RMSEP = 0.222.
Four observations were omitted due to extreme residuals.

The evaluation of model assumption for the selected model yielded a p-value of 0.9 for the Shapiro-Wilks test on normality of residuals. All four directional tests in the GVLMA test reported that the model assumptions were acceptable, as well as the global test. The test gave p-values above 0.5, for all five tests. High VIF-values were detected, 42.08 and 40.75 on installed capacity and squared installed capacity respectively, and otherwise were below 1.7.

A diagnostics plot was plotted and assessed, given in **Figure 10**. It shows a few influential observations, namely 128 with a relatively high Cook's distance and residual error, 122 with relatively high leverage and cook's distance and 117 with a relatively high Cook's distance. The diagnostics plot was found to be acceptable with respect to normality of residuals and absence of patterns in the residuals vs. fitted and scale-location plots.

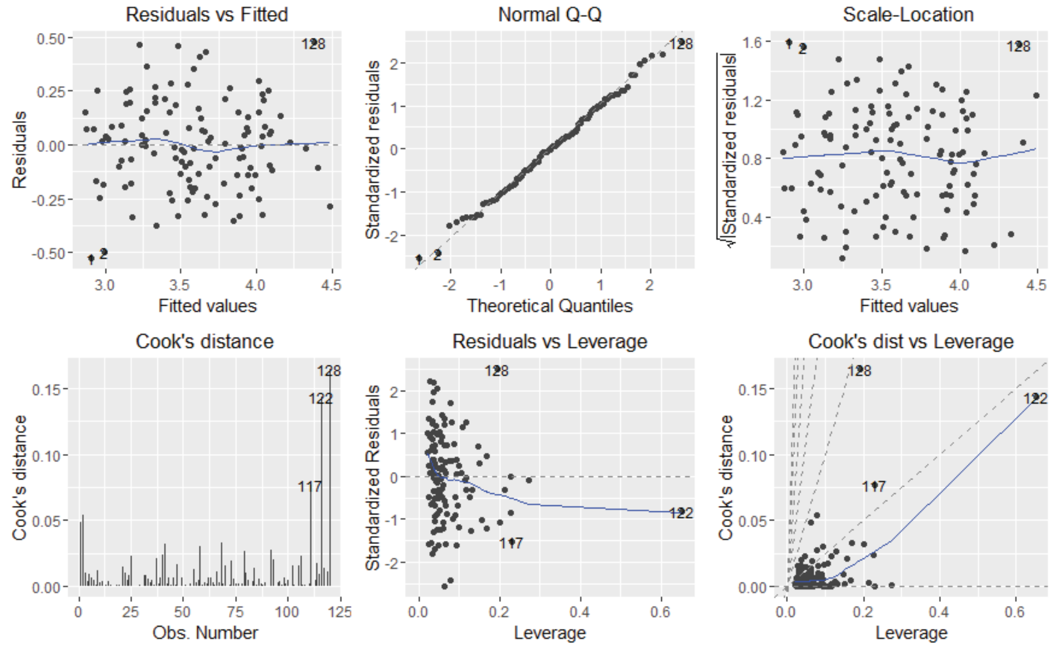

Figure 10: Six diagnostics plots for Prediction model 1 for total investment costs.

The results from the leave-one-out cross-validation are presented graphically in **Figure 11**. This shows the total investment costs as predicted by the linear model, with confidence and prediction intervals for the estimates, along with the budgeted (real) costs, reported (real) costs, and the leave-one-out cross-validation estimate. The figure shows that neither the model nor the budgets are able to give precise estimates of the actual investment costs. The confidence and prediction intervals for each observation are based on the linear model estimates, and are therefore conservative. If the intervals were based on the cross-validated estimates they would have been slightly wider (due to a higher residual standard error).

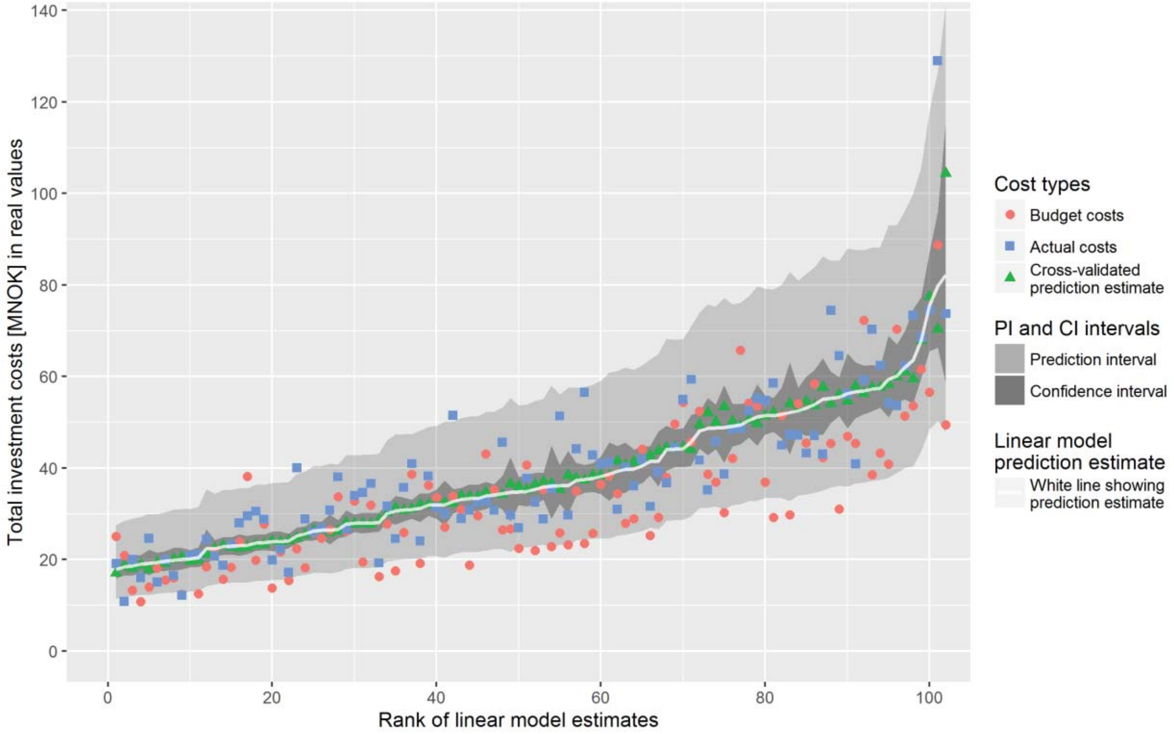

Figure 11: Prediction model 1 estimates with confidence and prediction intervals of total investment costs compared to budgeted costs, actual costs, and leave-one-out cross-validation estimates. The CI and PI are not based on the cross-validated prediction, and are thus conservative.

In order to more closely inspect the deviations from the reported costs, a plot for the deviations between actual costs, modeled costs and budgeted costs is given in **Figure 12**. The figure indicates no systematic pattern or correlation between the budget cost estimates and model cost estimates. It shows that the cross-validation estimates and the linear model predictions are mostly consistent, with the exception of one substantial deviation of approximately 20%.

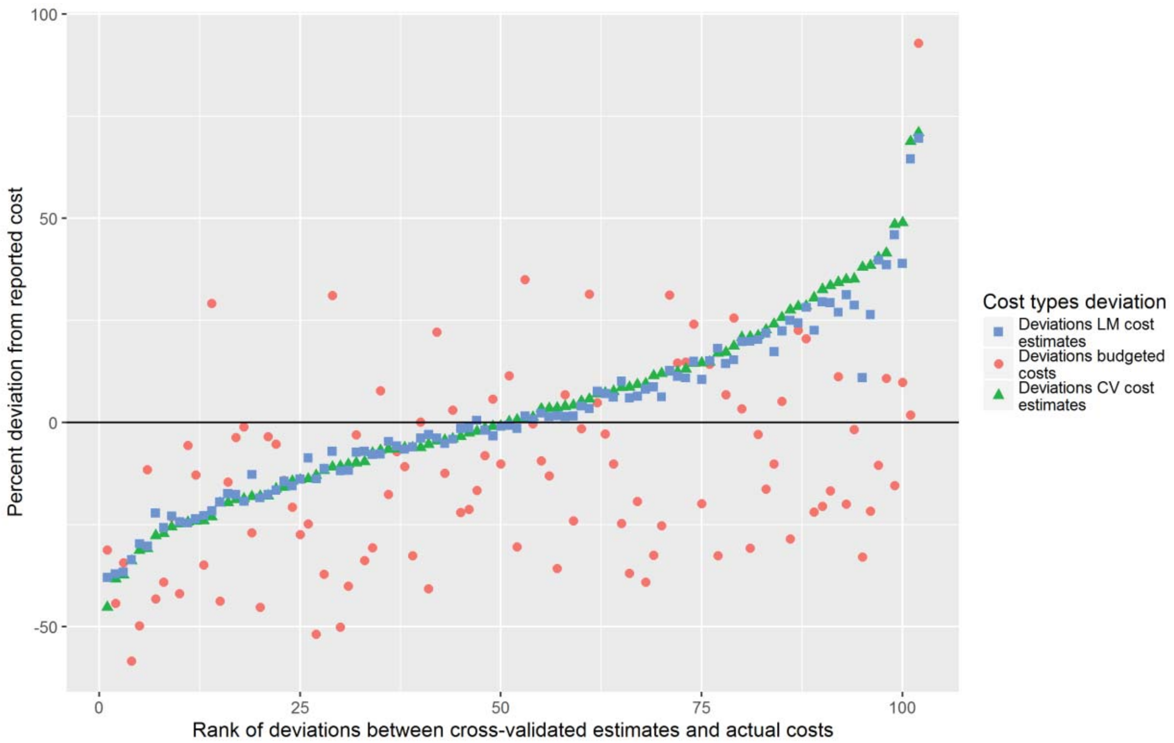

Figure 12: Plot for relative deviations in Prediction model 1 estimates and budgeted costs as percentage of reported costs, sorted by deviation size of the cross-validated linear model prediction estimates. Negative values show underestimated costs (estimated costs lower than actual costs), positive values show overestimated costs (estimated costs higher than actual costs).

Further analysis of model performance is given in section 6.3.3. The model results for each observation is given in **Table 16** in Appendix 3: Prediction model 1 dataset

6.3.2 Prediction model 2: Sum of partial costs

The final subset-model had the function given in **Equation 2**, and its metrics are given in **Table 10**. It is a model with six predictor variables. Installed capacity has one first order term and a squared term, with positive and negative coefficients respectively, giving a downward-facing parabolic curve when isolated against the log-transformed response variable. The weighted average penstock diameter term is squared, with no first-order term included, the justification for this choice is found in the previous model description in section 6.3.1. The waterway length variable has one first order term and one squared term. These two terms have also a downward-facing parabolic curve when isolated. For the time variables the terms indicate a positive linear function between the variables and the log-transformed cost response term. This model was fitted using a substantially smaller dataset than the previous model, with only 90 observations. The record of reported partial costs is smaller than that of total costs, and the range of the model, i.e. range of each predictor variable, was reduced by removing outliers in an attempt to develop a model more precise for the majority of the cases.

$$\begin{aligned}\log(\mathit{sum\ of\ (real)\ partial\ costs}) \\ = \hat{\beta}_0 + \hat{\beta}_1 \times \mathit{Installed\ capacity}\ [MW] + \hat{\beta}_2 \times (\mathit{Installed\ capacity}\ [MW])^2 \\ + \hat{\beta}_3 \times (\mathit{Weighted\ average\ penstock\ diameter}\ [m])^2 + \hat{\beta}_4 \\ \times \mathit{Waterway\ Length}\ [m] + \hat{\beta}_5 \times (\mathit{Waterway\ Length}\ [m])^2 + \hat{\beta}_6 \\ \times \mathit{Construction\ year}\ [\mathit{years}, 2005\ =\ 1] + \hat{\beta}_7 \times \mathit{Construction\ time}\ [\mathit{years}]\end{aligned} \qquad \mathbf{(2)}$$

Table 10: Prediction model 2. Model data for best subset model for predicting (log-transformed) sum of partial costs.

Predictors	$\hat{\beta}$ *no.*	$\hat{\beta}$-estimate	$\hat{\beta}$ SE	p-value	Data range Min	Max
Intercept	$\hat{\beta}_0$	1.31E+00	2.15E-01	3.80E-08		
Installed capacity [MW]	$\hat{\beta}_1$	5.47E-01	8.84E-02	2.30E-08	1.20	5.60
(Installed capacity [MW])^2	$\hat{\beta}_2$	-4.80E-02	1.28E-02	0.00031	1.20	5.60
(Weighted average penstock diameter [m])^2	$\hat{\beta}_3$	1.88E-01	3.27E-02	1.60E-07	0.20	2.10
Waterway Length [m]	$\hat{\beta}_4$	5.98E-04	1.89E-04	0.00224	170	2730
(Waterway Length [m])^2	$\hat{\beta}_5$	-1.50E-07	5.76E-08	0.01095	170	2730
Construction year [years, 2005 = 1]	$\hat{\beta}_6$	2.56E-02	9.88E-03	0.01142	2005	2015
Construction time [years]	$\hat{\beta}_7$	1.23E-01	2.67E-02	1.40E-05	0.49	4.56

Response variable log-transformed using natural logarithm.
Residual standard error = $e^{0.18} = 1.2$ MNOK, 82 degrees of freedom (DF), $R^2 = 0.833$, adjusted $R^2 = 0.819$,
F-statistic: 58.5 on 8 and 82 DF . p-value < 2E-16. PRESS = 3.17, RMSEP = $e^{0.188} = 1.21$ MNOK.
Mean absolute error rate of CV prediction = 15.6%.
13 observations were omitted due to extreme residuals and/or high values of Cook's distance or Leverage.

The evaluation of model assumption for the selected model yielded a p-value of 0.1 for the Shapiro-Wilks test on normality of residuals. All four directional tests in the GVLMA test reported that the model assumptions were acceptable, as was the global test. The test gave p-values above 0.15, for all five tests. High VIF-values of 40.57 and 40.11 were detected for installed capacity and squared installed capacity respectively, and VIF-values of 32.00 and 29.48 were detected for waterway length and waterway length squared respectively, otherwise below 1.7.

The diagnostics plot in **Figure 13** shows no single influential observations, and no apparent pattern of the residuals, and was found to be acceptable.

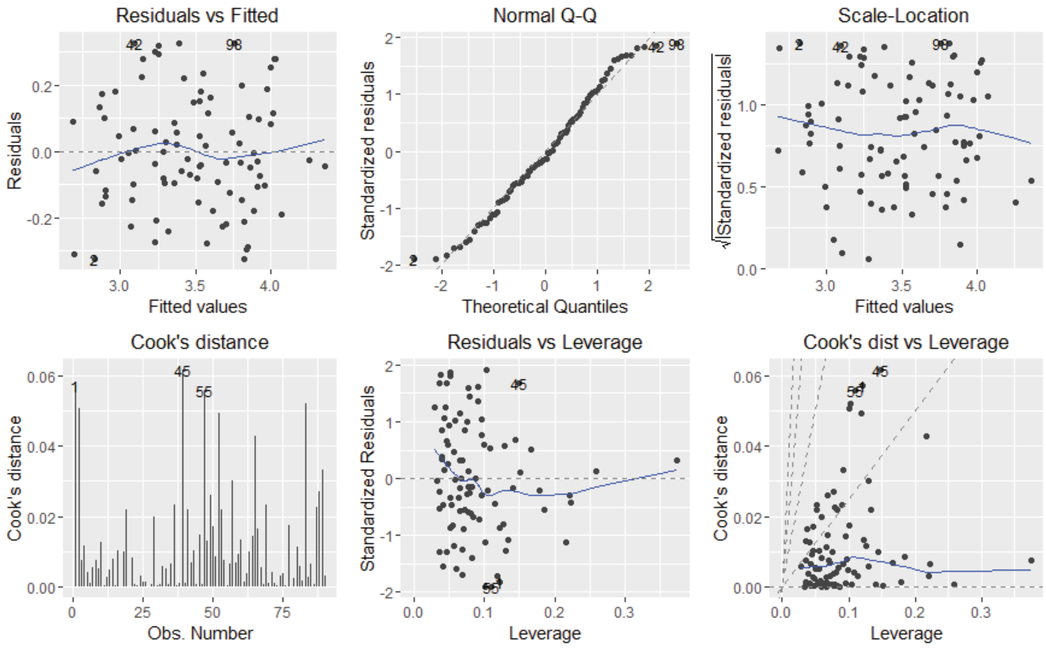

Figure 13: Six diagnostics plots for Prediction model 2 on partial costs.

The results from the leave-one-out cross-validation are displayed graphically in **Figure 14**. This shows the partial investment costs (intake, waterway and power station costs) as predicted by the linear model, with confidence and prediction intervals for the estimates, along with the budgeted (real) costs, reported (real) costs, and the leave-one-out cross-validation estimate. The figure confirms a slightly better performance than Prediction model 1 with respect to the reported costs.

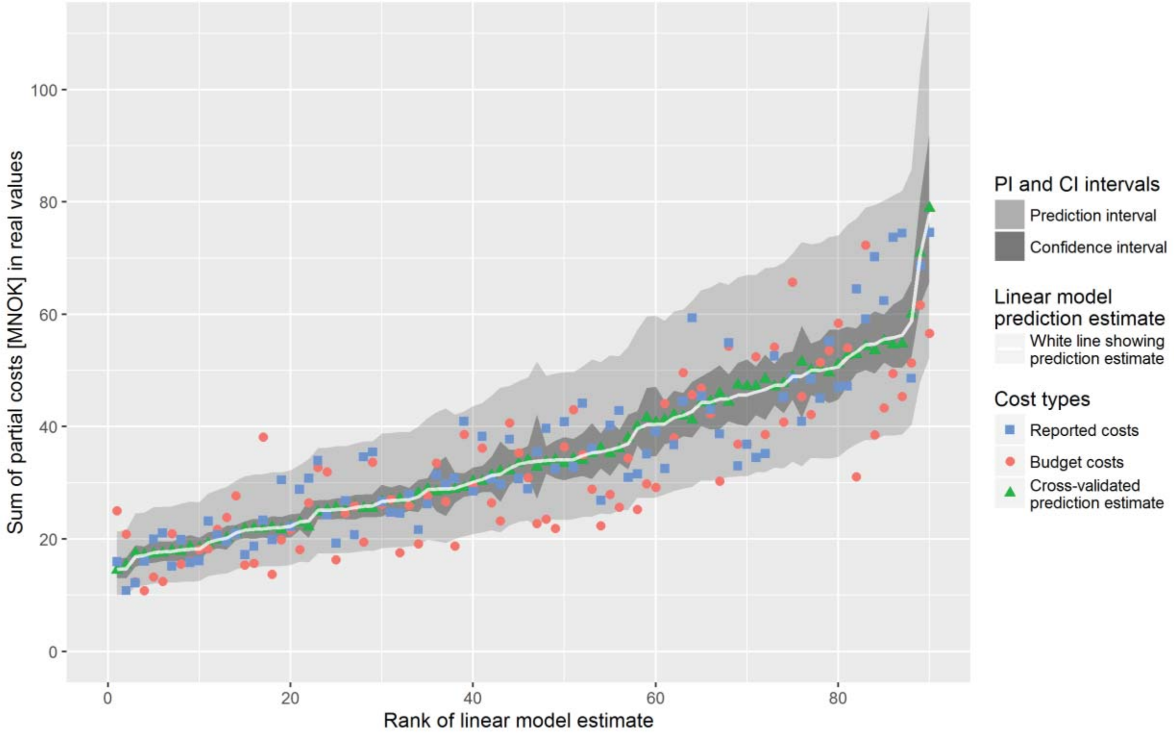

Figure 14: Prediction model 2 estimates with confidence and prediction intervals of partial costs compared to budgeted costs, actual costs, and leave-one-out cross-validation estimate. Predictions for the same dataset as the linear model was estimated from. The CI and PI are not based on the cross-validated prediction, and are thus conservative.

In order to more closely inspect the deviations from the reported costs, a plot for the deviations between actual costs, modeled costs and budgeted costs is given in **Figure 15**. The figure indicates no systematic pattern or correlation between the budget cost estimates and model cost estimates. It shows that the cross-validation estimates and the linear model predictions are mostly consistent, with the exception of one substantial deviation of approximately 20%.

The model results for each observation is given in **Table 17** in Appendix 4: Prediction model 2 dataset

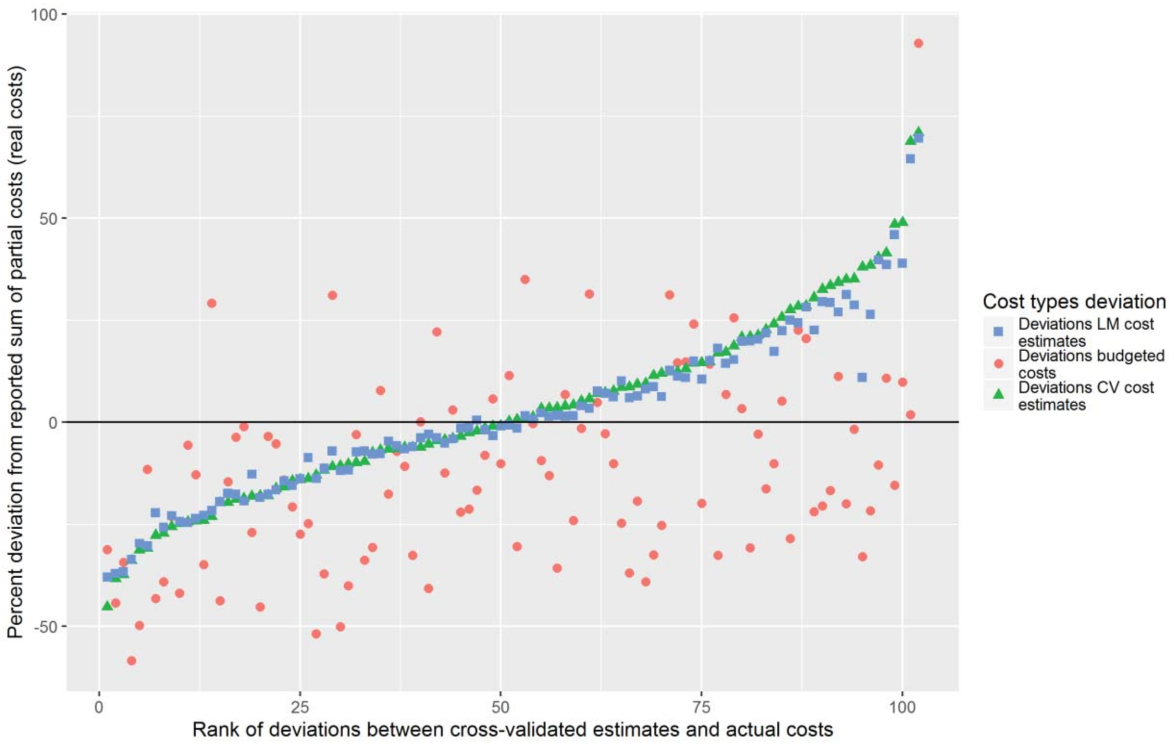

Figure 15: Plot for relative deviations in Prediction model 2 estimates and budgeted costs as percentage of reported costs, sorted by deviation size of the cross-validated linear model prediction estimates. Negative values are underestimated costs (estimated costs lower than actual costs), positive values are overestimated costs (estimated costs higher than actual costs)

6.3.3 Performance test of the two linear models compared to budget estimates

The prediction performance for this model was tested using two-sample tests. The results are given in **Table 11**. The results show that, although both prediction models on average have smaller percent deviations in both relative and absolute percent, the test cannot conclude that the models have significantly better prediction performance on average than the budget estimates. It also shows that when the model is tailored to a more homogenous subset of the data, and only the partial costs are used as a benchmark, the prediction model 2 still cannot perform significantly better than the budget estimates.

Tests were performed, as an extension from the graphical analysis of the standardized deviations of the models in **Figure 12** and **Figure 15**, to identify any presence of system patterns of devation in the estimates among the budget estimates and the model estimates. **Table 12** and **Table 13** give cross-tabulations of the frequency of under- and overestimation of costs in the budget estimates and estimates of the two prediction models. Chi-square tests were performed performed to determine whether or not there was an equal distribution of simultaneous under- and overestimation of the two estimates. The results from the chi-square tests for both models confirm that the distribution is not equal. A closer look at both tables show that the prediction models have a relatively even distribution of under- and overestimation (which is also required by the linear model assumptions of normality of residuals), and the budget estimates show a clear trend of underestimation of total costs, and a weak trend of underestimation compared to the sum of partial costs. More compelling in the analysis are the diagonals of the tables: Both show that the estimates tend to “agree”, in the sense that they are both under- or overestimating simultaneously. For model 1 on total costs, the model estimates and the budget estimates are both under or over in 62.7% of the cases, and for model 2 in 66.7% of the cases. This may be a sign that both estimation methods lack the some underlying information from which the estimates were based.

Table 11: Two-sample tests for difference between performance of Prediction model 1 and 2 compared to the budget cost estimates.

Diff to actual costs: predicted costs using leave-one-out vs. budgeted costs		Two-sample, unpaired Welch-test (two-sided, unequal variances)					Two-sample, unpaired Wilcoxon test (two-sided)			
	n	Mean diff CV-pred	Mean diff Budget.	Lower 95% CI	Upper 95% CI	p-value	Diff. in location	Lower 95% CI	Upper 95% CI	p-value
Model 1: Absolute difference in percent	102	18.0%	21.7%	-7.79 pp	0.47 pp	0.08	-3.4 pp	-7.27 pp	0.261 pp	0.07
Model 1: Difference in percent	102	2.84%	-11.9%	8.3 pp	21.2 pp	1E-05	14.9 pp	8.63 pp	21.3 pp	8E-06
Model 2: Absolute difference in percent	90	15.6%	20.7%	-9.32 pp	-1.02 pp	0.01	-3.49 pp.	-7.24 pp.	0.26 pp	0.07
Model 2: Difference in percent	90	1.90%	-4.41%	-0.44 pp	13.1 pp	0.07	7.58 pp	1.07 pp	14.3 pp	0.03

Differences CV-pred = Cross-validated prediction - reported costs. Differences budget = Budgeted costs – reported costs.
Differences between the two samples given as: CV prediction deviation– budget cost deviation. pp = percent points

Table 12: Cross-table for trends in budget and model estimates, Prediction model 1.

		CV Model estimates		
Budget estimates		Low	High	Row Total
Low	N	43	31	74
	Chi-square contribution	1.25	1.20	
	N / Row Total	0.58	0.42	0.73
	N / Col Total	0.86	0.60	
	N / Table Total	0.42	0.30	
High	N	7	21	28
	Chi-square contribution	3.30	3.17	
	N / Row Total	0.25	0.75	0.28
	N / Col Total	0.14	0.40	
	N / Table Total	0.07	0.21	
Column Total	N	50	52	102
	N / Row Total	0.49	0.51	

Pearson's Chi-squared test:
Chi^2 = 8.91, degrees of freedom (DF). = 1, p-value = 2.84E-03
Pearson's Chi-squared test with Yates' continuity correction:
Chi^2 = 7.63, DF = 1, p-value = 5.73E-03

Table 13: Cross-table for trends in budget and model estimates, Prediction model 2.

		CV Model estimates		
Budget estimates		Low	High	Row Total
Low	N	32	21	53
	Chi-square contribution	2.56	2.14	
	N / Row Total	0.60	0.40	0.59
	N / Col Total	0.78	0.43	
	N / Table Total	0.36	0.23	
High	N	9	28	37
	Chi-square contribution	3.66	3.06	
	N / Row Total	0.24	0.76	0.41
	N / Col Total	0.22	0.57	
	N / Table Total	0.10	0.31	
Column Total	N	41	49	90
	N / Row Total	0.46	0.54	

Pearson's Chi-squared test:
Chi^2 = 11.4, DF. = 1, p-value = 7.27E-04
Pearson's Chi-squared test with Yates' continuity correction:
Chi^2 = 10, DF. = 1, p-value = 1.56E-03

7 DISCUSSION

This paper has been devoted to two main analyses: The first is an explanatory analysis of external project factors that may affect the cost of small-scale hydropower plants, and the second consists of predictive modeling of investment costs of small-scale hydropower projects. In this chapter, strengths and weaknesses of the analysis, and the results will be discussed with respect to the findings in these analyses compared to previous research. The chapter starts off with a section on limitations to the analysis.

7.1 LIMITATIONS OF THE DATA AND THE ANALYSIS

In this section, the main limitations of the data collection and statistical analysis are stated and discussed. Several factors in this analysis' weaknesses may be due to the data quality, and some to the assumptions made in the modeling process.

In statistical analysis, one of the assumptions is that the sample data is randomly selected from a larger population. The regulation requiring SHP owners to submit commissioning reports came into place in 2007. This introduced a time bias, as projects commissioned before this year were not included in this data record. In addition, not all owners submitted commissioning forms. This is the case for 51 projects, more than half of which dated back before 2009. No measures were taken to overcome this bias.

There are some concerns to be considered regarding the data quality from the commissioning reports themselves. The license holders are not required to submit third party documentation for most of the features in the report (which include investment costs, intake and waterway dimensions); only for the performance characteristics of the turbine and the generator are these required. This information may therefore in many cases be inaccurate. Additionally, the owners are required to submit the form within one month after the power plant has been commissioned. With such a short timeframe, there is no guarantee that the owner has finalized the economic assessment of the project, which means the reported costs may be partly estimates, and thus inaccurate.

For several of the projects reported in the dataset, the total costs and the sum of partial costs differed. It is not possible to know with certainty which costs have been added to the total costs when these two numbers differ. In some cases it might be the cost of a connection fee, while. In other cases, administrative costs and other miscellaneous costs not directly related to the inlet, penstock and power plant might have been added only in the "total costs" figure.

No measures were done to overcome accuracy issues related to reported investment costs. In Haga and Espegren (2013) they adjusted the partial costs by splitting the difference between the sum of partial costs and total costs evenly on the three cost components. Consequently, the sum of adjusted partial costs and total were thus equal in their analysis. However, as the modeling process showed, there was more noise in the reported total costs than in the sum of partial costs.

The adjustments made to prices according to the cost index were based on average numbers. Each hydropower plant had different shares of partial costs, each of which has had a slightly different price development during the last ten years. The most precise way to adjust the costs would therefore be to break down both budgeted costs and reported costs into each category of the index and adjust accordingly. This would, however, require a more detailed assessment of each project, which was not prioritized within the timeframe of this research. The costs in real values reported in this thesis should therefore be treated as an approximation of what the actual budgeted and reported costs would have been in 2015 currency if adjusted precisely.

The cost index adjustment was based on the reported construction start-up date. This may have not have yielded accurate cost adjustments, as the main cost-driving construction activity may have occurred at a later point in time/later year than the reported start-up date. For projects with a construction period spanning over several years, the costs may also have increased within the construction period. Despite these possible shortcomings, there were few other options for more accurate index adjustments of the costs with the available data. This method of index adjustment was also in line with the method used in Haga and Espegren (2013), which was based on expert advice by NVE.

The cost index is updated annually, and no interpolation was done to increase the time resolution of the index. Therefore, the construction date variable was recoded to construction year and adjusted to the index accordingly. Due to this fact, some level of accuracy may have been lost in the construction year variable. On the other hand, this loss of accuracy was likely negligible when compared to the accuracy of the other variables, due to factors including the above-mentioned inaccuracies found in the rest of the dataset.

The construction time variable used in this analysis was calculated as the time difference between the construction start-up date and the date of commissioning. The time lag between when the construction work itself was actually finished and the reported date of operation may vary in the dataset, such that the length of the construction period as calculated here was was likely exaggerated compared to the actual construction period. In order to account for this, one would need to obtain more detailed data from each power plant owner, and that was outside of the scope of this thesis.

The extent to which the above-mentioned inaccuracies have added noise and/or introduced a systematic bias to the dataset is (to the author's knowledge) impossible to measure using the data available in this thesis. This could be tested later if quality-assured data were collected and tested against this sample.

7.2 Budgeted versus reported costs

The observed tendency of underbudgeting in the license applications supports the findings in Stokke (2014) and Haga and Espegren (2013). Haga and Espegren found a mean deviation of 0.806 NOK/kWh and a median deviation of 0,821 NOK/kWh. In this analysis, a mean deviation of 0.734 NOK/kWh and median deviation of 0.699 NOK/kWh was found (see **Table 6**). Haga and Espegren's findings lie within the 95% confidence intervals for both mean (t-test) and the median (Wilcox test). The two findings are thus consistent in statistical sense.

An increase in total costs from the budget estimate in the license application is not necessarily a problem if the rise in costs is due to changes in plans leading to a higher annual production. However, the tests for deviation between estimated annual production in the license application and the updated estimate reported after commissioning show no significant trend in any direction, but have a mean and median decrease in estimated annual production of 209 kWh and 230 kWh respectively. This indicates the increased specific costs per kWh in the majority of the cases is not due to decrease in the production estimates, but due to an actual increase in costs.

There is a significant median increase in the nominal total costs of 49.1%, and a median increase in nominal specific total costs of 1.12 NOK/kWh, which signifies an unambiguous trend of underbudgeting in the license applications. As Stokke (2014) and

Haga and Espegren (2013) suggest, this may be due to several unforeseeable factors, such as changes in plans, a lengthy license process, a lack of forecasting of cost increase from rising prices, and inadequate planning of the construction period. Still, if the budgeted investment cost in the license applications is to be regarded as a useful evaluation criterion for the owners themselves, NVE, other decision makers, possible investors, and the public in general, more precise budget estimates would be advantageous, or even a requirement.

7.3 Cost-drivers

7.3.1 *Construction year*

The average increase in nominal investment costs over time were found to be 10.9 pp (percentage points, 2005 = 100%) per year in total investment cost per estimated annual production, and 5.9% per year in total investment costs per installed capacity. The average increase in real investment costs over time was found to be 3.7 pp per year for total investment costs per estimated annual production, while the fitted model for real investment costs per installed capacity indicated no significant increase in costs. In comparison, the cost index for small hydropower had an average increase of 5.2 pp and 4.3% during the same period, while Prediction model 1 and 2 had effect sizes equivalent to 1.79% and 2.59% cost growth per year respectively (when all other variables were kept at dataset average). It should be noted that both prediction models were estimated on a subset of the data from which the simple linear models for construction year were estimated.

On the other hand, if the physical dimensions of the hydropower plants have grown (in particular waterway length) due to less accessible water resources, this would explain part of the difference between the isolated effects in the prediction models compared to effect sizes of the simple linear model. When assessing the data, there is a trend of increasing waterway lengths over time.

The results from the second model indicate hydropower projects have become more expensive per kWh estimated production, also when the costs are adjusted for the growth in the SHP cost index.

No significant cost increase was detected for real costs per production capacity. There is still, however, an average deviation of 1.6% between the growth in costs per MW in nominal values and the growth in the SHP cost index. This deviation cannot be regarded as significant in a statistical sense, but may still indicate an actual cost increase.

These two results from cost development per MWh and per MW combined suggest a trend of decreasing production volume per investment cost over time, but also a possible cost increase exceeding the inflation in the SHP cost index. The fact that the construction year term provided a significant contribution as a predictor in the cost estimation models for investment costs supports the second argument that costs have in fact increased.

NVE (2015c) mention that the hydropower projects with the best water resources and lowest specific investment costs generally are those first developed. NVE claim license applications display a trend of increasing specific costs for hydropower projects over time (Hansen, E. Personal communication 06.05.2016). Graphical inspection of the data from budgeted specific investment costs supports this trend.

Another possible explanation of less production volume per investment costs is that the requirements for mitigation of environmental impacts have become stricter over time. NVE claim that stricter environmental requirements have not been documented during the past ten years (ibid.), although the requirements for environmental assessments have become stricter.

NVE do state the compensation flow has increased. The compensation flow requirement has in general increased from the ordinary low water flow up to the 5th percentile during the past ten to fifteen years, indicating that less of the water flow can be utilized for production (ibid).

In conclusion, the observed increase of investment costs per annual production supports a possible explanation that the most accessible and easily exploitable water resources have been developed, and that over time the potential projects left are the less cost efficient ones.

7.3.2 Construction time

Total investment costs were found to increase with construction time, with an observed effect of 37.8% per year, when no other predictor variables were considered. These results should, however, be treated with caution. While the model and its predictor variable did display a significant contribution to explaining investment costs, the unexplained variance is still very high. This fitted regression model is thus not useful for prediction of increase in costs per length of construction period.

When more predictors were included into the regression, the observed effect was much lower. In Prediction model 1 and 2, the isolated effect of construction time corresponded to an increase in total costs by 15.3% and 13.1% per year respectively. It should be noted that both prediction models have been trained on a subset of the data from which the simple linear model was fitted. In these subsets, projects with tunnels are excluded, and such projects have both higher costs, and somewhat longer construction periods. This increases the effect size of construction time on investment costs when projects with tunnels are included.

In any case, the detection of construction time as a significant prediction variable supports the findings in Stokke (2014) and Haga and Espegren (2013), where both studies point to prolonged construction periods as one of the main causes of budget overruns.

7.3.3 Geography

The results from this analysis indicate a significant difference between total investment costs per annual production across counties and regions in Norway. Western Norway, and in particular Sogn og Fjordane stand out having the lowest specific costs. Northern Norway, and both the counties Nordland and Troms have a relatively high number of observations, and group means that are higher than the total mean. This indicates a significantly higher specific cost in this region and for both counties compared to the rest of the regions and counties. For the rest of the counties, finding a clear pattern is more difficult, and no significant difference has been found in specific costs among those counties and regions.

Lower specific costs for projects in Western Norway could be expected, as this region has a relatively high level of precipitation combined with a topography with a large relief. It is the region in Norway with the highest potential for small-scale hydropower according to NVE (2004). For Northern Norway, the observed higher specific costs may then be more surprising. Nordland and Troms are both counties with high levels of precipitation and mountainous areas, and some of the highest potential for small-scale hydropower according to NVE (2004), ranking among the top five together with Hordaland, Sogn og Fjordane and Møre og Romsdal. As the region of Northern Norway is less densely populated, the transmission and distribution networks are also more thinly dispersed This translates into increased costs due to longer distances to existing power grid and other infrastructure, and other factors related to distances, population density and economic activity (higher transport costs, less contractor expertise available, etc.). This would need to be analyzed further in order to give a plausible and concise explanation.

7.3.4 License owners and costs

The results indicate specific investment costs are higher for professional developers than non-professional project developers. There may be several explanations for this, but two will be elaborated upon here:

Renewable energy projects such as small-scale hydropower are capital intensive. The production infrastructure has a high initial capital requirement, while the costs for operation and maintenance are low, and the infrastructure has a very long lifetime.

Access to necessary investment capital at a sufficiently low cost may therefore be a concern for private license owners. The distribution of ownership of the small-scale hydropower projects might therefore be a result of a "natural selection", where the owners of the most cost-efficient projects have been able to raise the necessary capital themselves for developing their projects, while owners of less cost-efficient projects have sold their water rights to professional developer companies with more available investment capital. Given that an owner or a group of owners can raise the necessary capital with low costs at its own hands, it will be more attractive for them not to sell the water rights and keep all income themselves.

For the professional project developers, access to new and attractive projects is restricted by the ownership of the water rights. Their portfolio of possible investment may therefore restricted to projects made available only by the owners willing to sell their water rights.

When it comes to the project development process itself, it would seem likely that professional developers with experience from developing previous projects would be able to achieve lower costs. However, this could be outweighed by the lower quality/accessibility of the water resource if private, non-professional owners in general have access to the most cost-efficient projects and develop those themselves.

Another likely explanation that may contribute to the cost differences is that private owners may not report the cost of their work efforts in the project development process. In many cases, these owners are farmers that can do some of the construction work themselves. They are also likely to spend a considerable time on administrative work. The "true" costs of privately owned projects may therefore not have been reported. Professional developers, on the other hand, would naturally account for all costs from their labor.

The above-mentioned arguments are not supported by literature and should be regarded as new questions to be answered in future studies rather than reliable explanations.

7.4 PREDICTION MODEL FOR INVESTMENT COSTS

These results indicate the regression method for cost estimation can be applied to Norwegian small-scale hydropower projects, and that it can be used to estimate investment costs for future projects. In this section the performance of the models in this thesis is compared to performance of models in previous studies, along with some considerations of the generalization of the prediction models, and practical use of them.

7.4.1 Model performance compared to other methods

Prediction model 1 had a mean absolute error rate (MAER) of 18% compared to the reported total costs in real values, while Prediction model 2 on partial costs had a MAER of 15.6%. As recalled from the literature review, the regression model in Kim et al. (2004) for residential building costs had a MAER of 6.95%. Smith and Mason (1997) had a MAER of 30.4% for their regression model for costs of pressure vessels for chemical production, while Gunduz and Sahin (2015) reported an absolute error rate of 9.94% for their regression model on hydropower plant costs.

The papers presenting cost estimation equations in Section 3.2.2. or the most part do not give mean error rates, but rather error rate ranges. To allow for comparison, the prediction error ranges for the prediction models in this thesis are presented here. Prediction model 1 has an absolute error rate from of 0.0283% to 70.9%, and a range of relative error rate from -45.4% to 70.9%. Prediction model 2 has an absolute error rate from 0.0006% to 43.9%, and a range of relative error rate between -30.7% and 43.9%.

The studies referred to in the literature are listed below, with their ranges:

- Gordon (1983) (in Singal et al. 2010) reported a measured estimation accuracy of $\pm 40 - 50\%$ in an early-phase estimation model for project costs.
- Singal and Saini (2007) achieved an accuracy of $\pm 12\%$ in their cost equation for small-size, low head run-of-river projects.
- Singal et al. (2010) achieved an accuracy of $\pm 11\%$ in their cost estimation equations for small hydropower projects cost.
- Ogayar and Vidal (2009) reported an error range between-9.50% and 19.52% for the cases in their study.
- Aggidis et al. (2010) reported error rates down to $\pm 10\%$, and up to $\pm 25\%$ for different turbine types, and $\pm 25\%$ accuracy for electro-mechanical equipment in their cost-estimation equations.
- Cavazzini et al. (2016) reported mean errors below 10% for electro-mechanical equipment for Pelton and Francis turbines and below 20% for Kaplan turbines. The latter paper reported higher accuracy compared to several of the above-mentioned models.

Many of the aforementioned cost equations consider only parts of the hydropower projects, for example turbine costs or hydropower station costs. With this approach, fewer variables contribute to the estimate cost. With this isolation of variables, it is easier to achieve increased accuracy.

The motivation behind developing two prediction models was in part to compare how well the two models would perform with slightly different data, and whether different predictor variables would add to the prediction performance with the two response variables. It is not altogether surprising that mainly the same predictor variables were the best contributors in both models, as none of the variability which lies in the difference between the sum of partial costs and the total costs can be said to be attributed to any of the variables collected inn this dataset. While Prediction model 1 cannot be said to add value as an accurate model, it may still have its virtues when compared to Prediction model 2. Although Prediction model 2 is the obvious choice when attempting to estimate investment costs for new projects as accurate as possible, Prediction model 1 can give information about how uncertain the estimates are. In other words, since the lack of predictor variables is apparent, how much did total investment costs vary for other projects similar the one analyzed?

7.4.2 Loss of accuracy through the model generalization

Small hydropower projects in Norway are all “tailor-made” to fit the specific site characteristics. Translated into model terms, this means that hydropower projects have a high number of changing variables, all of which contribute to the total investment cost to varying extents. It is therefore challenging to take all the possible characteristics/variables of hydropower projects into consideration when developing a regression model from a limited dataset. It would be preferable to include as many variables as possible, in order to project the full extent of a new project as accurately as possible when predicting the investment cost. However, as observed in the model fitting process, the multiple linear regression method became less accurate when all variables were included. On the other hand, the fewer variables included, the more generalized the model becomes. The models developed here will predict the same cost for one project with a 2 m dam, and one with a 100 m wide dam (when all other variables are kept the same), even though the project costs should clearly differ. Thus, the model predictions for new projects must be interpreted and used with caution.

The main scarcity when estimating such models is the amount of available data. With a dataset containing ten times more observations, many of the variables that were left out in this study could possibly contribute to higher prediction accuracy. This is, however, a challenge for the majority of such modeling problems.

One of the main motivations to develop a prediction model for new hydropower projects was to have an independent project assessment tool for estimating investment costs for new projects. The budgets in the license applications clearly have their shortcomings, judging by the observed inaccuracy. The goal was therefore to develop a model which could take experience from previous project developments into consideration and incorporate the cost deviations which are unforeseeable at the planning stage, and project onto new projects.

7.4.3 Selection of modeling method.

The majority of the literature found on cost estimation of small-scale hydropower projects had a different methodological approach than multiple linear regression. If a different and more advanced method had been used, this might have yielded a more precise prediction model. On the other hand, the advantage of multiple linear regression is that the methodology is well-documented. The simplicity of the model makes it transparent and easy to interpret. In order to achieve higher prediction accuracy, more sophisticated modeling methods could be considered for future research.

7.4.4 Practical use of the prediction models and limitations

For practical use, these prediction models could be valuable for comparing multiple small-hydropower projects. The results indicate that although both prediction models developed in this study had lower mean absolute error rates on average, they were still not significantly better than the budget estimates when tested.

When comparing multiple projects, it is advisable to standardize the cost estimates by the estimated average annual production. Together with the budget estimates (when these are adjusted to real costs), it could be used as a second estimate (or third, when you take into consideration the budget assessment done by NVE in the license application process). It could also be used as an estimate of both expected costs and the uncertainty of the cost estimate. Here, uncertainty is meant both in terms of internal model uncertainty (confidence and/or prediction interval), and “external” uncertainty as compared to the budgeted cost. A large deviation between the model prediction estimate and the budgeted cost for a specific project would indicate that this project differs in some parameters compared to the sample of projects upon which the prediction model was estimated. This can be useful when screening of multiple projects, for example in the case of a possible investor looking into investment projects or portfolios.

When utilized for prediction of costs for new projects, the prediction models perform well only when the data is within the range from which the prediction models were calculated. Extrapolation is burdened with high uncertainty. It is therefore likely that the two prediction models developed here cannot be utilized for all new projects. Projects with tunnels are in any case out of scope for the models reported here.

The time variables included in the prediction models will introduce more uncertainty, but were regarded as non-negligible. The effect of construction year on an investment ahead in time will be uncertain. From a more technical perspective, it may also introduce higher confidence and prediction intervals for the cost estimate when the input variables are outside the ranges of the training dataset. The variable should be kept, as is it signifies that there has been a trend of increasing costs (in spite of the costs having been adjusted for inflation and general growth in contractor costs) which should also somehow be accounted for in cost estimation for future projects. In the practical use of the model for future projects, two choices can be made: One is to set the construction year value to 2015 (11) in all new projects for which the prediction model is utilized, and use a prognosis for the future cost index development to scale up the investment cost for a future year. The second choice may be to plug in the actual estimated construction year, and in addition use a prognosis for future cost index development, which will give a higher estimate of the investment cost.

The construction time variable must also be based on estimation of construction time for future projects. As reported in previous studies (Haga & Espegren 2013; Stokke 2014), the actual construction period is often lengthier than estimated in the license applications. This research tested whether project size had any effect on construction time, but no trends were found. The best estimate here may therefore be the mean construction time for the reported projects.

7.5 Implications of the findings and future work

The findings in this thesis can serve as a basis to improve the budget estimation process for license applicants, and the assessment of the budgets made by NVE. If the budgeted costs are to serve as a reasonable cost estimate, the rather low accuracy found in this study should call for improvements. NVE can use this for further adjustment of their cost basis for small-scale hydropower projects, which is frequently used as a basis for these budgets. The budget estimates would become more accurate if they also included an estimate of inflation in costs, which was also suggested by Stokke (2014) and Haga and Espegren (2013)

The cost estimation tool developed in this thesis clearly has its shortcomings, but may be developed further to a more accurate tool, by:

- Collecting more accurate accounting figures from existing hydropower projects: Excluding highly unpredictable numbers, such as grid connection fee, expropriation fees, etc. will give more comparable data, and the model estimation is likely to be more accurate.
- Different model development approach: Using more sophisticated modeling tools, such as particle swarm optimization (see Cavazzini et al. 2016).
- Incorporating spatial analysis, adding variables such as distance to existing roads and closest power grid connection point, geological characteristics of intake and waterway areas and topography (as also suggested in Stokke 2014)

The results in Section 6.3.3 show an apparent correlation of simultaneous under- and over-estimation of investment costs by both budget costs and estimated costs. This pattern may indicate that both cost estimation methods lack information and have a common bias. For further development of a cost estimation model, this could be investigated more in detail.

8 Conclusion

This thesis conducts numeric analyses of costs for small-scale hydropower in Norway. It shows that budgets in license applications have consistently low accuracy in their estimates of total investment costs. The thesis documents four external cost drivers for SHP projects, namely:

1) Specific costs have increased significantly more than the general cost inflation for SHP projects throughout the past ten years
2) Investment costs increase significantly with longer construction periods
3) SHP projects in Northern Norway have significantly higher costs than SHP projects in Western Norway, although both regions have among the best potential for SHP production
4) Non-professional developers of small-scale hydropower projects achieve significantly lower specific investment costs than professional development companies.

The thesis also sets out to develop a cost estimation tool for SHP projects. Two models were estimated based on multiple linear regression. One achieved a moderate accuracy in estimation of partial costs of SHP projects.

The results from the analysis call for improvements of the budgets in the license applications, if these should serve as a reasonable estimate of the total investment cost of SHP projects. The analysis of the four cost driving factors for SHP project costs is the first of its kind in the literature, and these results may be useful information for investors and project developers looking into new development objects.

The cost estimation tools may serve as a valuable independent tool for SHP project assessment. Together with index adjusted budget estimates, they can aid in the cost ranking of multiple projects for investors and project developers who are seeking new ventures.

REFERENCES

Aggidis, G. A., Luchinskaya, E., Rothschild, R. & Howard, D. C. (2010). The costs of small-scale hydro power production: Impact on the development of existing potential. *Renewable Energy*, 35 (12): 2632-2638.

Allen, D. M. (1974). The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction. *Technometrics*, 16 (1): 125-127.

Anagnostopoulos, J. S. & Papantonis, D. E. (2007). Optimal sizing of a run-of-river small hydropower plant. *Energy Conversion and Management*, 48 (10): 2663-2670.

Aspen, J. G. (2014). Interessa for småkraftverk har tørka ut. *Nationen*, pp. 4-5.

Bacon, R. W. & Besant-Jones, J. E. (1998). Estimating construction costs and schedules - Experience with power generation projects in developing countries. *Energy Policy*, 26 (4): 317-333.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference - understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33 (2): 261-304.

Cavazzini, G., Santolin, A., Pavesi, G. & Ardizzon, G. (2016). Accurate estimation model for small and micro hydropower plants costs in hybrid energy systems modelling. *Energy*, 103: 746-757.

Elfaki, A. O., Alatawi, S. & Abushandi, E. (2014). Using Intelligent Techniques in Construction Project Cost Estimation: 10-Year Survey. *Advances in Civil Engineering*, 2014.

ESHA. (2004). Guide on How to Develop a Small Hydropower Plant European Small Hydropower Association.

Everitt, B. S. & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. 4th ed. Cambridge, UK: Cambridge University Press.

Fox, J. & Weisberg, S. (2011). *An R Companion to Applied Regression*. Thousand Oaks CA: Sage.

GAO. (2009). GAO Cost Estimating and Assessment Guide, Best Practices for Developing and Managing Capital Program Costs, GAO-09-3SP: United States Government Accountabity Office. 420 pp.

Gordon, J. L. & Penman, A. C. (1979). Quick estimating techniques for small hydro potential. *International Water Power and Dam Construction*, 31 (9): 46-51.

Gordon, J. L. (1981). Estimating hydro station costs. *International Water Power and Dam Construction*, 33: 31-3.

Gordon, J. L. (1983). Hydropower costs estimates. *International Water Power and Dam Construction*, 35 (11): 30-37.

Gordon, J. L. & Noel, R. C. R. (1986). The economic limits of small and low-head hydro. *International Water Power and Dam Construction*, 38 (4): 23-26.

Gunduz, M. & Sahin, H. B. (2015). An early cost estimation model for hydroelectric power plant projects using neural networks and multiple regression analysis. *Journal of Civil Engineering and Management*, 21 (4): 470-477.

Haga, S. T. & Espegren, H. M. (2013). *Forholdet mellom budsjetterte og faktiske kostnader*: Norges vassdrags- og energidirektorat. Unpublished manuscript.

Hollander, M., Wolfe, D. A. & Chicken, E. (2014). *Nonparametric Statistical Methods*. 3rd ed. Hoboken, New Jersey, USA: John Wiley & Sons, Inc. 819 pp.

Horikoshi, M. & Tang, Y. (2015). *ggfortify: Data Visualization Tools for Statistical Analysis Results*, R package version 0.1.0.

Hyndman, R. J. & Athanasopoulos, G. (2013a). *Forecasting: principles and practice*. Melbourne, Australia: OTexts. Available at: https://www.otexts.org/fpp/5/3 (accessed: 11.05.2016).

Hyndman, R. J. & Athanasopoulos, G. (2013b). *Forecasting: principles and practice*. Section 5.1 Introduction to multiple linear regression. Melbourne, Australia: OTexts. Available at: https://www.otexts.org/fpp/5/1 (accessed: 12.4.2016).

Jenssen, L., Mauring, K. & Gjermundsen, T. (2000). Economic Risk and Sensitivity Analysis for Small-scale Hydropower Projects. *Technical Report, International Energy Agency*.

Johnson, R. A. & Wichern, D. W. (2007). *Applied multivariate statistical analysis*. 6th ed. Upper Saddle River, New Jersey, USA: Pearson Education, Inc. 773 pp.

Kaldellis, J. K., Vlachou, D. S. & Korbakis, G. (2005). Techno-economic evaluation of small hydro power plants in Greece: a complete sensitivity analysis. *Energy Policy*, 33 (15): 1969-1985.

Kaldellis, J. K. (2007). The contribution of small hydro power stations to the electricity generation in Greece: Technical and economic considerations. *Energy Policy*, 35 (4): 2187-2196.

Kim, G. H., An, S. H. & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment*, 39 (10): 1235-1242.

Kim, M., Lee, S., Woo, S. & Shin, D. H. (2012). Approximate cost estimating model for river facility construction based on case-based reasoning with genetic algorithms. *Ksce Journal of Civil Engineering*, 16 (3): 283-292.

Løvås, G. G. (2010). *Statistikk for universiteter og høgskoler*. 2 ed. Oslo: Universitetsforlaget. 489 pp.

Maindonald, J. H. & Braun, W. J. (2015). *DAAG: Data Analysis and Graphics Data and Functions*, R package version 1.22.

Mendenhall, W. & Sincich, T. (2003). *A Second Course in Statistics: Regression Analysis*. 6th ed. Upper Saddle River, New Jersey, USA: Pearson Education, Inc. 880 pp.

Merrow, E. W. & Schroeder, B. R. (1991). *Understanding the costs and schedule of hydroelectric projects*. AACE Transactions. 35th Annual Meeting.

Mishra, S., Singal, S. K. & Khatod, D. K. (2012). Costing of a Small Hydropower Projects. *International Journal of Engineering and Technology*, 4 (3): 239-242.

Montanari, R. (2003). Criteria for the economic planning of a low power hydroelectric plant. *Renewable Energy*, 28 (13): 2129-2145.

NasdaqOMX. (2016). Market prices Nordic electricity. Available at: http://www.nasdaqomx.com/commodities/market-prices (accessed: 11.5.2016).

Nord Pool. (2016). Elspot prices. Available at: http://www.nordpoolspot.com/Market-data1/Elspot/Area-Prices/ALL1/Hourly/?view=table (accessed: 6.12.2015).

NVE. (2004). Beregning av potensial for små kraftverk i Norge. *Rapport*: Norges vassdrags- og energidirektorat. 29 pp.

NVE. (2010a). Kostnadsgrunnlag for små vannkraftanlegg (opp til 10.000 kW): Norges vassdrags- og energidirektorat (NVE). 84 pp.

NVE. (2010b). Veileder i planlegging, bygging og drift av små kraftverk. *Veileder nr. 1/2010*. Oslo: Norges vassdrags- og energidirektorat. 137 pp.

NVE. (2012). Cost base for small-scale hydropower plants (< 10 000 kW): Norwegian Water Resources and Energy Directorate (NVE). 85 pp.

NVE. (2015a). *Konsesjonssaker Vannkraft*: Norges vassdrags- og energidirektorat. Available at: http://www.nve.no/no/Konsesjoner/Konsesjonssaker/Vannkraft/ (accessed: 1.12.2015).

NVE. (2015b). *Konsesjonssaker Vindkraft*: Norges vassdrags- og energidirektorat. Available at: http://www.nve.no/no/Konsesjoner/Konsesjonssaker/Vindkraft/ (accessed: 27.11.2015).

NVE. (2015c). Kostnader i energisektoren: Kraft, varme og effektivisering. *Rapport*: Norges vassdrags- og energidirektorat.

NVE. (2015d). *Licence applications dataset*.

NVE. (2015e). *Vannkraft*: Norges vassdrags- og energidirektorat. Available at: http://www.nve.no/no/Energi1/Fornybar-energi/Vannkraft/ (accessed: 27.11.2015).

NVE. (2016a). *Kostnadar i energisektoren*: Norges vassdrags- og energidirektorat. Available at: https://www.nve.no/energiforsyning-og-konsesjon/energiforsyningsdata/kostnadar-i-energisektoren/ (accessed: 22.4.2016).

NVE. (2016b). Kostnadsgrunnlag for små vannkraftanlegg (opp til 10.000 kW). *Rapport nr 40-2016*: Norges vassdrags- og energidirektorat. 89 pp.

NVE. (2016c). *Trinn 2 - Søknaden*: Norges vassdrags- og energidirektorat. Available at: https://www.nve.no/energiforsyning-og-konsesjon/vannkraft/sma-kraftverk/saksgang-for-sma-kraftverk/trinn-2-soknaden/ (accessed: 22.4.2016).

NVE & Energimyndigheten. (2016). *Elsertifikater: Kvartalsrapport nr. 4 2015*: Norges vassdrags- og energidirektorat, and Svenska Energimyndigheten.

OED & NVE. (2007). *Act No. 82 of 24 November 2000 relating to river systems and groundwater (Water Resources Act) NOTE: Unofficial translation - for information only.*: Norwegian Ministry of Petroleum and Energy. 22 pp.

OED. (2016). *Kraft til endring. Energipolitikken mot 2030 (Meld. St. 25 (2015–2016))*: The Royal Norwegian Ministry of Petroleum and Energy (OED). 229 pp.

Ogayar, B. & Vidal, P. G. (2009). Cost determination of the electro-mechanical equipment of a small hydro-power plant. *Renewable Energy*, 34 (1): 6-13.

Peña, E. A. & Slate, E. H. (2012). Global validation of linear model assumptions. *Journal of the American Statistical Association*.

Peña, E. A. & Slate, E. H. (2014). *gvlma: Global Validation of Linear Models Assumptions*, R package version 1.0.0.2.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*, Version 3.2.2. Vienna, Austria.

Razali, N. M. & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2 (1): 21-33.

Sheldon, L. H. (1981). Cost analysis of hydraulic turbines. *International Water Power and Dam Construction*.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, 25 (3): 289-310.

Singal, S. K. & Saini, R. P. (2007). *Analytical approach for cost estimation of low head small hydro power schemes*. International Conference on Small Hydropower-Hydro Sri Lanka. 24 pp.

Singal, S. K. & Saini, R. P. (2008). Cost analysis of low-head dam-toe small hydropower plants based on number of generating units. *Energy for Sustainable Development*, 12 (3): 55-60.

Singal, S. K., Saini, R. P. & Raghuvanshi, C. S. (2010). Analysis for cost estimation of low head run-of-river small hydropower schemes. *Energy for Sustainable Development*, 14 (2): 117-126.

Smith, A. E. & Mason, A. K. (1997). Cost estimation predictive modeling: Regression versus neural network. *The Engineering Economist*, 42 (2): 137-161.

Spyros, G. M., Wheelwright, S. C. & Hyndman, R. J. (1998). *Forecasting: Methods and Applications*. 3rd ed. Hoboken, New Jersey, USA: John Wiley & Sons, Inc. 642 pp.

SSB. (2016a). *Tabell: 03013: Konsumprisindeks*. Statistikkbanken: Statistisk sentralbyrå. Available at: https://www.ssb.no/statistikkbanken/.

SSB. (2016b). *Tabell: 08583: Elektrisitetsbalanse (MWh)*. Statistikkbanken: Statistisk sentralbyrå. Available at: https://www.ssb.no/statistikkbanken/ (accessed: 28.4.2016).

Stokke, T. (2014). *Småkraftverk: En analyse av avvik mellom budsjetterte og faktiske investeringskostnader*. M.Sc. thesis. Ås: Norges miljø- og biovitenskapelige universitet, Institutt for naturforvaltning. 51 pp.

Trost, S. M. & Oberlender, G. D. (2003). Predicting accuracy of early cost estimates using factor analysis and multivariate regression. *Journal of Construction Engineering and Management*, 129 (2): 198-204.

Tuhtan, J. A. (2007). *Cost Optimization of Small Hydropower*. M.Sc. thesis. Stuttgart: Universität Stuttgart, Institut für Wasserbau. 104 pp.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer. 498 pp.

Wang, W.-C., Wang, S.-H., Tsui, Y.-K. & Hsu, C.-H. (2012). A factor-based probabilistic cost model to support bid-price estimation. *Expert Systems with Applications*, 39 (5): 5358-5366.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*: Springer-Verlag New York.

Wickham, H. & Francois, R. (2015). *dplyr: A Grammar of Data Manipulation*, R package version 0.4.3.

Willer, D. C. (1991). Powerhouse and small hydropower project cost estimated. In Johns, S. G. (ed.) *Hydropower engineering handbook*. New York: McGraw-Hill.

Zhang, Q. F., Smith, B. & Zhang, W. (2012). Small hydropower cost reference model. *Oak Ridge National Laboratory. US Department Of Energy*.

Ökmen, Ö. & Öztaş, A. (2010). Construction cost analysis under uncertainty with correlated cost risk analysis model. *Construction Management and Economics*, 28 (2): 203-212.

Appendices

Appendix 1: Cost index for small-scale hydropower plants

Table 14: Cost index per 1st January from 1997 to 2015 for hydropower plant components from NVE (2016a). The aggregated costs are adapted from NVE (2015c p. 49), based on cost shares per component for the two plant types.

Year per. 1.1.	Mechanical equipment	Electro-technical equipment	General Civil works (incl. penstock)	Tunnel works	Dam works	Total Civil works	High Head (small) Plants aggregated	Small-scale Hydro Plants aggregated
2015	1.46	1.36	2.13	1.89	1.79	1.94	1.8	1.72
2014	1.4	1.32	2.08	1.84	1.75	1.89	1.76	1.68
2013	1.37	1.3	2.04	1.78	1.71	1.84	1.72	1.64
2012	1.34	1.27	1.99	1.74	1.67	1.8	1.68	1.6
2011	1.31	1.24	1.9	1.67	1.58	1.72	1.61	1.54
2010	1.25	1.19	1.8	1.63	1.51	1.65	1.54	1.47
2009	1.21	1.15	1.71	1.6	1.5	1.6	1.5	1.43
2008	1.12	1.08	1.6	1.51	1.42	1.51	1.41	1.34
2007	1.06	1.02	1.48	1.42	1.31	1.4	1.31	1.26
2006	0.96	0.96	1.38	1.33	1.24	1.32	1.23	1.17
2005	0.93	0.91	1.36	1.29	1.19	1.28	1.19	1.13
2004	0.91	0.81	1.21	1.25	1.13	1.2	1.11	1.06
2003	0.89	0.81	1.18	1.22	1.1	1.17	1.09	1.04
2002	0.87	0.82	1.14	1.19	1.07	1.13	1.06	1.02
2001	0.88	0.8	1.1	1.16	1.05	1.1	1.04	1
2000	1.12	1.06	1.08	1.07	1.04	1.06	1.07	1.08
1999	1.025	1.04	1.05	1.04	1.02	1.03	1.03	1.03
1998	1.012	1.02	1.02	1.02	1.01	1.015	1.015	1.02
1997	1	1	1	1	1	1	1	1

APPENDIX 2: LICENSE OWNER CLASSIFICATION

Table 15: Company classification into professional and non-professional owners

Kdb_ID	Title	Holder	Company Classification
4579	Færåsen kraftverk	AGDER ENERGI VANNKRAFT AS	pro
5215	Akslandselva kraftverk	AKSLANDSELVA KRAFTVERK AS	non-pro
5295	Bentsjord kraftverk	BEKK OG STRØM AS	Pro
4854	Bele Kraftverk	BELE KRAFT AS	pro
4474	Berdalselva kraftverk	BERDALSKRAFT AS	non-pro
5394	Måren kraftverk	BKK PRODUKSJON AS	pro
4940	Dalelva kraftverk	DALELVO KRAFT AS	non-pro
5168	Dversetelva kraftverk	DRAGEFOSSEN KRAFTANLEGG AS	non-pro
4211	Dvergsdalsdalen kraftverk	DVERGSDALSDALEN KRAFTVERK AS	non-pro
4108	Lya kraftverk	E-CO Vannkraft as	pro
4714	Eitro kraftverk	EITRO KRAFTVERK AS	non-pro
4904	Embla kraftverk	EMBLA KRAFT AS	non-pro
5046	Follsjå kraftverk	FOLLSJÅ KRAFT AS	pro
5639	Fossbråten kraftverk	FOSSBRÅTEN KRAFTVERK AS	non-pro
4212	Frammarsvik kraftverk	FRAMMARSVIK KRAFT AS	non-pro
4770	Føssa kraftverk	FØSSA KRAFTVERK AS	non-pro
3987	Gautvella småkraftverk	GAUTVELLA KRAFTVERK AS	non-pro
4406	Gjesdal kraftverk	GJESDAL KRAFT AS	non-pro
4095	Gjetingsdalen kraftverk	GJETINGSDALEN KRAFT AS	non-pro
4514	Gryta kraftverk	GRYTA KRAFT AS	non-pro
4670	Grønlielva kraftverk	GRØNLIELVA KRAFTVERK AS	non-pro
5492	Folkedal kraftverk -opprusting og utviding	HARDANGER ENERGI AS	pro
5199	Bordalsbekken kraftverk	HAUGALAND KRAFT AS	pro
4620	Øvre Forsland kraftverk	HELGELAND KRAFT AS	pro
2712	Laksen kraftverk	HELGELAND KRAFT AS	pro
4425	Kangsliåga kraftverk	HELGELAND SMÅKRAFT AS	pro
4837	Holdal kraftverk	HOLDALBEKK KRAFT AS	non-pro
4948	Hopselva kraftverk	HOPSELVA KRAFTVERK AS	pro
4291	Hovland kraftverk	HOVLAND KRAFT AS	non-pro
4571	Hynna Kraftverk	HYNNA KRAFT AS	non-pro
5247	Fossan kraftverk	HÅLOGALAND KRAFT AS	non-pro
5070	Håra Kraftverk	HÅRA KRAFT AS	non-pro
5096	Sundli kraftverk, reg. Strømtjønna	Jamtåsbekken vasskraftlag DA v/Røhme	non-pro
4088	Julfoss kraftverk	JULFOSS KRAFT AS	non-pro
5326	Kaldåna småkraftverk	KALDÅNA KRAFT AS	non-pro
4363	Veslefallet kraftverk	KIÆR MYKLEBY Anders Kiær	non-pro
2969	Landakraft (tidl. Landa kraftverk)	KLØVKRAFT AS	non-pro
5377	Kulu kraftverk	KULU KRAFTVERK AS	non-pro
5535	Kverninga kraftverk	KVERNINGA KRAFTVERK AS	non-pro
3028	Kvernstad kraftverk	KVERNSTAD KRAFT AS	non-pro
4834	Kvitno kraftverk	KVITNO KRAFT AS	pro

Kdb_ID	Title	Holder	Company Classification
5324	Kvitvella Electrisitetsverk i Kvitvella-fossen	KVITVELLA ELECTRISITETSVERK AS	non-pro
4805	Kviven kraftverk	KVIVEN KRAFT AS	non-pro
4842	Kylland småkraftverk	KYLLAND KRAFT AS	non-pro
4637	Kysinga kraftverk	KYSINGA KRAFT AS	non-pro
5140	Litj-Hena kraftverk	LITJ-HENA KRAFTVERK AS	pro
4288	Fjellet kraftverk, Godal kraftverk med reg. og overføring	LØVENSKIOLD FOSSUM KRAFT	pro
6015	Åmot kraftverk	LØVENSKIOLD-FOSSUM KRAFT AS	pro
5272	Middøla kraftverk	MIDDØLA KRAFT AS	pro
5028	Misfjord kraftverk	MISFJORD KRAFT AS	non-pro
4900	Muoidejohka kraftverk	MUOIDEJOHKA KRAFT AS	pro
5191	Mygland kraftverk	MYGLAND KRAFT AS	pro
4703	Nape kraftverk	NAPE KRAFT AS	non-pro
4468	Nedre Neset kraftverk	NESET KRAFT AS	pro
4470	Østre Neset kraftverk	NESET KRAFT AS	pro
4469	Vestre Neset kraftverk	NESET KRAFT AS	pro
5016	Fjelna kraftverk	NORDIC POWER AS	pro
4848	Saltdalelva kraftverk	NORDIC POWER AS	pro
4998	Ellenelva kraftverk	NORDIC POWER AS	pro
4502	Røstad kraftverk	NORDIC POWER AS	pro
5025	Bergselva kraftverk	NORDIC POWER AS	pro
4348	Lappvikelva (Tidl. Lappvik) kraftverk	NORDIC POWER AS	pro
4381	Mortensdalelva kraftverk	NORDIC POWER AS	pro
4293	Storå kraftverk	NORDIC POWER AS	pro
4340	Glesåa kraftverk	NORDRE LØSSET AS	non-pro
2283	Forsanvatn kraftverk	NORD-SALTEN KRAFT AS	pro
6106	Kvemma kraftverk	NORSK GRØNNKRAFT AS	pro
4308	Leirelva kraftverk	NORSK GRØNNKRAFT AS	pro
5550	Kvassteinåga kraftverk	NORSK GRØNNKRAFT AS	pro
4309	Tverråga kraftverk	NORSK GRØNNKRAFT AS	pro
1784	Havdal kraftverk	NORSK GRØNNKRAFT AS	pro
5094	Skravlåga kraftverk	NORSK GRØNNKRAFT AS	pro
5092	Kinnforsen kraftverk	NORSK GRØNNKRAFT AS	pro
4490	Urdsdalselva kraftverk	NORSK GRØNNKRAFT AS	pro
5833	Tossevikelva kraftverk	NORSK GRØNNKRAFT AS	pro
4393	Imsland småkraftverk	NORSK KRAFT HOLDING AS	pro
4540	Nydalselva småkraftverk	NYDALSELVA KRAFT AS	non-pro
5485	Nyvikelva kraftverk	NYVIKELVA KRAFT DA	pro
4535	Herje kraftverk	RAUMA ENERGI AS	non-pro
5643	Ravnåga kraftverk	RAVNÅGA KRAFTVERK AS	non-pro
4200	Rendalselva Kraftverk (Tidl. Kraftutbygging Rendalselva)	RENDALSELVA KRAFTVERK AS	non-pro
4394	Ringdal kraftverk	RINGDAL KRAFTVERK AS	non-pro
3826	Vassfossen kraftverk	RISDAL ENERGI AS	non-pro
5171	Rodal kraftverk	RODAL KRAFT AS	non-pro
5379	Ryddøla kraftverk	RYDDØLA KRAFTVERK AS	non-pro
3137	Saksenvik småkraftverk	SAKSENVIK KRAFT AS	pro

Kdb_ID	Title	Holder	Company Classification
2932	Mølnelva minikraftverk	SANDVIK ENERGI AS	non-pro
4754	Seimsdal(Tidl. Øyni) kraftverk	SEIMSDAL KRAFT AS	non-pro
5623	Kraftverk i Gaudalselva	SETERKRAFT AS	pro
2511	Reinskar kraftverk,regulering av Storvatn	SJØFOSSEN ENERGI AS	pro
4014	Steinåga kraftverk	SJØFOSSEN ENERGI AS	pro
5229	Liarelva Kraftverk	SKL PRODUKSJON AS	pro
4737	Skripelandsfossen kraftverk	SKRIPELANDSFOSSEN KRAFT AS	non-pro
4347	Rasdalen kraftverk	SMÅKRAFT AS	pro
4505	Søberg kraftverk	SMÅKRAFT AS	pro
4201	Furegardene (Tidl. Sagelvi) kraftverk	SMÅKRAFT AS	pro
4600	Dokkelva kraftverk	SMÅKRAFT AS	pro
5570	Reinåga kraftverk	SMÅKRAFT AS	pro
4203	Knutfoss kraftverk	SMÅKRAFT AS	non-pro
4321	Usma kraftverk	SMÅKRAFT AS	pro
4396	Skarelva kraftverk	SMÅKRAFT AS	pro
2548	Kveaså kraftverk (Tidl. Kveasåni småkraftverk)	SMÅKRAFT AS	pro
4598	Kanndalen kraftverk	SMÅKRAFT AS	pro
4497	Stublielva småkraftverk	SMÅKRAFT AS	pro
4601	Sagelvi kraftverk	SMÅKRAFT AS	pro
4158	Vågaåna småkraftverk	SMÅKRAFT AS	pro
4427	Bruvollelva kraftverk	SMÅKRAFT AS	pro
4215	Tveitaskar (Tidl. Tveitaskarelva) Kraftverk	SMÅKRAFT AS	pro
4536	Tjøtaelva kraftverk	SMÅKRAFT AS	pro
4390	Bergstø kraftverk	SMÅKRAFT AS	pro
4525	Langdalselva kraftverk	SMÅKRAFT AS	pro
5754	Kaldsåni kraftverk	SMÅKRAFT AS	pro
4341	Blådalselva kraftverk	SMÅKRAFT AS	pro
5473	Tyttebærelva kraftverk	SMÅKRAFT AS	pro
4451	Stokkelandsåna kraftverk	SMÅKRAFT AS	pro
5431	Valåi/Vålåe kraftverk	SMÅKRAFT AS	pro
4323	Eidsetelva kraftverk	SMÅKRAFT AS	pro
4480	Tua småkraftverk	SMÅKRAFT AS	pro
4541	Torvikelva kraftverk	SMÅKRAFT AS	pro
5026	Snefjellå kraftverk	SNEFJELLÅKRAFT AS	non-pro
4508	Fykanvannet kraftverk	STATKRAFT ENERGI AS	pro
4008	Rødberg kraftverk	STATKRAFT SF	pro
4388	Steindøla kraftverk	STEINDØLA KRAFT AS	pro
4752	Innhavet kraftverk	STORVATNET KRAFT AS	non-pro
4384	Furset kraftverk - Planendring ved gjenoppbygging	Stranda Energiverk AS	pro
5220	Rødset kraftverk	Stranda Energiverk AS	pro
4866	Strandos kraftverk	STRANDOS KRAFT AS	non-pro
4877	Stølsdalselva kraftverk	STØLSDALSELVA KRAFTVERK AS	non-pro
4400	Sværen kraftverk	SVÆREN KRAFT AS	non-pro
4669	Syrifossen kraftverk	SYRIKRAFT AS	non-pro
4629	Dyrkorn kraftverk	TAFJORD KRAFTPRODUKSJON AS	pro

Kdb_ID	Title	Holder	Company Classification
5699	Simavika kraftverk	TROMSØ KOMMUNE	non-pro
4636	Litlebø kraftverk	TRYGGESTAD KRAFT AS	non-pro
4130	Urke kraftverk	TUSSA ENERGI AS	pro
5788	Viddal kraftverk	TUSSA ENERGI AS	pro
5029	Standal kraftverk	TUSSA ENERGI AS	pro
4851	Draura kraftverk	TUSSA ENERGI AS	pro
4273	Dalegjerdet kraftverk	Tussa Energi AS	pro
4653	Skår småkraftverk	TUSSA ENERGI AS	pro
4573	Tveitelva kraftverk	TVEITELVA KRAFTVERK AS	non-pro
4634	Usma kraftverk	USMA KRAFT AS	non-pro
4160	Veka kraftverk	VEKA KRAFT AS	non-pro
4760	Venna kraftverk	VENNA KRAFT AS	non-pro
5048	Vikaelva kraftverk	VIKAELVA KRAFTVERK AS	non-pro
5184	Voldsetelva kraftverk	VOLDSETELVA KRAFTVERK AS	non-pro
4101	Vågen kraftverk	VÅGEN KRAFT AS	non-pro
4270	Eldrevatn kraftverk	Østfold Energi AS	pro
4791	Hanestadnea kraftverk	ØVERGAARD ENERGI AS	non-pro
4477	Øvstedal minikraftverk	ØVSTEDAL KRAFT AS	non-pro
4374	Øyadalen kraftverk	ØYADALEN KRAFTVERK AS	non-pro
4555	Ågskar småkraftverk	ÅGSKARKRAFT AS	non-pro
5339	Åkraelva kraftverk	ÅKRAELVA KRAFTVERK AS	non-pro
4886	Åselva småkraftverk	ÅSELVA KRAFT AS	non-pro

Appendix 3: Prediction model 1 dataset

Table 16: Prediction model 1 output. Reported costs and predictor variables omitted due to confidentiality.

Nr	Budg. Costs [MNOK]	CV Pred. [MNOK]	LM Pred [MNOK]	LM CI Lower [MNOK]	LM CI Upper [MNOK]	LM PI Lower [MNOK]	LM PI Upper [MNOK]
1	25.05	17.83	17.92	15.89	20.21	11.33	28.35
2	16.95	17.88	18.03	16.15	20.13	11.43	28.45
3	13.23	18.45	18.57	16.38	21.05	11.72	29.41
4	20.88	19.25	18.57	16.63	20.74	11.77	29.30
5	13.95	19.29	19.83	17.08	23.02	12.43	31.63
6	10.81	19.32	19.14	17.35	21.12	12.17	30.12
7	15.51	19.64	19.66	17.52	22.05	12.44	31.05
8	18.14	19.81	19.49	17.48	21.73	12.35	30.73
9	20.91	20.04	20.09	18.18	22.19	12.76	31.62
10	12.50	20.05	20.11	18.01	22.46	12.75	31.74
11	15.98	20.55	20.15	17.66	23.00	12.70	31.98
12	12.40	20.81	19.97	17.67	22.57	12.62	31.61
13	21.65	21.40	21.44	19.76	23.28	13.67	33.63
14	23.89	22.49	22.77	20.48	25.31	14.45	35.89
15	38.13	22.58	22.84	20.87	24.99	14.54	35.87
16	16.23	22.66	22.40	19.63	25.55	14.11	35.54
17	18.38	22.96	23.20	19.37	27.78	14.39	37.41
18	21.76	23.04	22.96	21.16	24.91	14.64	36.00
19	15.68	23.20	23.00	21.07	25.12	14.65	36.12
20	18.27	23.27	23.27	21.39	25.32	14.83	36.51
21	19.87	23.28	23.50	21.65	25.49	14.98	36.85
22	27.73	23.33	23.47	21.81	25.25	14.98	36.75
23	21.62	23.69	23.62	21.46	25.99	15.02	37.14
24	13.76	23.78	23.59	21.49	25.89	15.01	37.08
25	15.40	24.51	24.31	22.75	25.98	15.54	38.03
26	22.31	24.65	24.98	23.21	26.89	15.95	39.12
27	17.55	25.54	25.61	23.01	28.50	16.24	40.38
28	18.16	25.62	25.68	24.05	27.42	16.42	40.17
29	25.94	25.93	25.93	23.81	28.24	16.52	40.69
30	33.68	25.97	26.28	24.30	28.43	16.77	41.20
31	24.64	26.34	26.36	24.53	28.32	16.83	41.26
32	11.23	26.47	26.26	21.26	32.44	16.09	42.89
33	26.44	26.92	27.23	23.96	30.94	17.18	43.15
34	19.49	27.21	27.44	25.24	29.85	17.49	43.06
35	31.92	27.38	27.64	25.52	29.93	17.63	43.33
36	32.76	27.63	27.78	25.82	29.90	17.74	43.51
37	26.21	28.08	27.99	25.13	31.17	17.75	44.14
38	16.29	28.71	28.35	26.22	30.67	18.09	44.44
39	27.70	29.95	29.98	28.04	32.06	19.17	46.91
40	17.55	30.41	30.03	26.99	33.42	19.05	47.35
41	25.95	30.60	30.79	28.18	33.65	19.61	48.36
42	38.66	30.92	31.14	29.04	33.39	19.90	48.74
43	36.23	31.52	31.72	29.27	34.36	20.23	49.73
44	33.54	31.53	31.53	28.41	34.98	20.01	49.67
45	33.81	32.09	32.67	29.98	35.59	20.81	51.27
46	26.70	32.21	31.67	25.85	38.79	19.46	51.52
47	26.47	32.21	32.88	29.51	36.63	20.85	51.85
48	19.11	32.54	32.02	28.90	35.48	20.33	50.43
49	27.57	33.09	32.53	29.50	35.88	20.68	51.19
50	27.05	33.09	32.69	27.87	38.34	20.42	52.32
51	30.89	33.18	33.04	30.59	35.68	21.08	51.77
52	43.04	33.51	33.47	30.49	36.74	21.30	52.61
53	29.53	33.58	33.54	30.68	36.67	21.36	52.68
54	18.76	33.76	33.62	30.65	36.89	21.39	52.84
55	35.30	33.99	33.91	31.62	36.36	21.66	53.07
56	23.21	34.25	33.35	27.53	40.40	20.59	54.02
57	40.68	34.84	34.99	31.56	38.80	22.22	55.12
58	25.80	35.01	36.0 6	31.88	40.79	22.78	57.09
59	22.40	35.47	35.04	31.94	38.44	22.29	55.07
60	35.36	35.53	35.25	32.35	38.41	22.46	55.33
61	21.92	35.73	35.55	32.15	39.31	22.58	55.97
62	34.99	35.96	36.35	32.84	40.24	23.08	57.24

Nr	Budg. Costs [MNOK]	CV Pred. [MNOK]	LM Pred [MNOK]	LM CI Lower [MNOK]	LM CI Upper [MNOK]	LM PI Lower [MNOK]	LM PI Upper [MNOK]
63	33.14	36.48	36.36	33.58	39.36	23.19	56.99
64	25.69	36.71	36.92	33.87	40.25	23.52	57.95
65	22.79	36.73	36.38	28.87	45.85	22.08	59.94
66	23.52	36.86	37.44	34.40	40.74	23.86	58.74
67	31.07	37.97	36.41	30.61	43.30	22.64	58.56
68	38.22	38.04	38.49	32.44	45.67	23.95	61.85
69	31.96	38.16	37.81	35.04	40.80	24.13	59.24
70	28.92	38.22	39.03	35.49	42.92	24.82	61.37
71	28.92	38.44	38.32	34.60	42.43	24.33	60.34
72	36.45	38.91	39.03	34.91	43.63	24.73	61.60
73	34.40	39.67	38.72	33.74	44.44	24.36	61.55
74	27.95	40.01	40.03	34.68	46.21	25.14	63.74
75	29.21	40.15	40.03	34.49	46.46	25.10	63.85
76	44.11	40.38	40.45	36.54	44.78	25.69	63.69
77	25.27	42.04	41.63	38.34	45.20	26.54	65.29
78	54.35	42.78	43.34	39.20	47.92	27.53	68.23
79	49.61	43.30	43.36	39.34	47.79	27.56	68.21
80	45.66	44.04	45.12	39.77	51.18	28.48	71.48
81	37.97	44.81	44.20	39.32	49.68	27.97	69.85
82	39.18	46.88	46.25	42.44	50.41	29.47	72.60
83	38.63	47.26	46.06	40.40	52.50	29.03	73.06
84	52.41	47.37	47.14	43.23	51.40	30.03	74.00
85	36.93	48.23	48.09	43.31	53.38	30.52	75.77
86	37.95	48.30	48.80	43.49	54.77	30.89	77.10
87	42.15	48.45	48.45	43.90	53.46	30.79	76.24
88	65.71	49.31	49.29	44.92	54.08	31.36	77.46
89	54.15	49.72	49.81	45.92	54.03	31.76	78.11
90	29.21	50.08	50.30	46.70	54.19	32.11	78.79
91	36.91	50.17	50.33	46.16	54.88	32.07	79.01
92	29.80	50.75	50.01	40.94	61.10	30.77	81.29
93	53.55	50.84	51.24	44.70	58.74	32.25	81.42
94	51.43	51.48	51.03	45.60	57.11	32.33	80.57
95	46.31	52.04	51.85	47.56	56.52	33.03	81.38
96	30.29	52.18	49.65	41.45	59.47	30.79	80.07
97	54.06	52.58	52.32	47.51	57.61	33.26	82.28
98	58.39	52.92	52.33	45.67	59.97	32.94	83.15
99	45.36	53.13	54.16	48.74	60.18	34.37	85.36
100	31.06	53.41	53.73	49.69	58.09	34.28	84.20
101	45.51	54.02	53.04	46.73	60.21	33.47	84.05
102	72.29	54.97	55.16	50.14	60.69	35.08	86.75
103	56.78	55.10	56.67	49.07	65.44	35.58	90.25
104	38.51	55.23	55.80	50.93	61.13	35.51	87.67
105	43.30	55.56	55.77	51.56	60.31	35.58	87.41
106	56.65	56.01	57.07	51.15	63.66	36.18	90.02
107	42.32	56.24	54.34	46.36	63.69	33.96	86.95
108	45.34	56.34	55.21	49.39	61.73	34.98	87.14
109	46.88	56.56	56.52	48.55	65.78	35.40	90.23
110	34.28	58.08	58.24	53.92	62.92	37.17	91.27
111	70.32	58.30	57.99	51.95	64.74	36.76	91.50
112	51.34	59.54	59.75	52.82	67.59	37.74	94.59
113	40.84	59.87	59.14	50.56	69.17	36.98	94.57
114	53.58	61.93	63.49	53.60	75.21	39.53	101.97
115	61.61	65.58	65.85	57.55	75.36	41.46	104.59
116	88.76	67.84	76.59	63.20	92.82	47.27	124.07
117	56.60	72.23	72.48	62.60	83.92	45.48	115.52
118	93.37	79.39	72.73	62.62	84.46	45.59	116.03
119	56.32	91.10	84.89	68.76	104.82	52.00	138.60
120	49.47	96.88	81.08	56.72	115.89	45.91	143.18

Appendix 4: Prediction model 2 dataset

Table 17: Prediction model 2 output. Reported costs and predictor variables omitted due to confidentiality.

Nr.	Budg.Costs	LM Pred	CV Pred	LM CI lower	LM CI upper	LM PI lower	LM PI upper
	[MNOK]	[MNOK].	[MNOK].	[MNOK]	[MNOK]	[MNOK]	[MNOK]
1	25.05	14.48	14.31	12.79	16.39	9.73	21.55
2	20.88	14.63	15.26	12.82	16.70	9.81	21.83
3	10.81	17.03	17.10	15.50	18.73	11.54	25.14
4	13.23	17.52	17.27	15.57	19.71	11.80	26.02
5	12.40	16.69	17.31	14.79	18.84	11.23	24.82
6	12.50	17.57	17.31	15.84	19.49	11.88	26.00
7	15.51	18.03	17.84	16.00	20.32	12.13	26.80
8	20.91	17.71	17.92	16.01	19.59	11.98	26.18
9	18.14	18.25	18.47	16.34	20.40	12.31	27.06
10	15.98	18.18	18.56	15.89	20.81	12.18	27.16
11	18.27	19.43	19.21	17.71	21.31	13.17	28.66
12	21.76	20.04	20.01	18.46	21.76	13.62	29.50
13	23.89	20.07	20.11	18.00	22.38	13.55	29.73
14	27.73	21.05	21.06	19.65	22.56	14.34	30.91
15	19.87	22.11	21.71	20.27	24.12	15.01	32.58
16	38.13	21.81	21.75	20.22	23.52	14.84	32.05
17	15.40	21.72	21.91	20.22	23.32	14.79	31.90
18	15.68	21.83	22.03	19.97	23.87	14.81	32.19
19	13.76	21.95	22.10	19.92	24.18	14.86	32.41
20	21.62	22.21	22.21	20.14	24.50	15.04	32.81
21	26.44	23.45	22.35	20.26	27.14	15.64	35.16
22	18.16	23.18	23.03	21.73	24.73	15.80	34.00
23	32.76	25.23	24.95	23.48	27.12	17.18	37.06
24	31.92	25.18	25.22	23.42	27.08	17.14	36.99
25	24.64	25.48	25.43	23.60	27.53	17.33	37.47
26	19.49	25.86	25.49	23.82	28.08	17.57	38.06
27	33.68	25.98	25.52	23.80	28.36	17.63	38.28
28	16.29	25.46	25.77	23.60	27.48	17.32	37.43
29	25.94	25.71	26.05	23.45	28.18	17.43	37.92
30	17.55	26.84	27.03	24.24	29.72	18.15	39.69
31	26.21	27.02	27.06	24.12	30.27	18.21	40.08
32	25.95	27.10	27.06	25.03	29.33	18.42	39.86
33	27.05	27.14	27.29	24.81	29.68	18.41	40.01
34	26.70	28.24	27.75	23.33	34.19	18.50	43.12
35	19.11	27.82	28.46	24.97	31.01	18.78	41.21
36	33.54	29.08	28.64	24.93	33.92	19.34	43.72
37	18.76	29.21	29.09	26.46	32.24	19.77	43.16
38	27.70	28.99	29.10	26.99	31.14	19.74	42.58
39	38.66	29.88	29.51	27.75	32.17	20.34	43.91
40	29.53	30.32	30.43	27.81	33.06	20.58	44.67
41	36.23	30.75	30.44	28.39	33.31	20.90	45.24
42	26.47	31.16	31.19	28.06	34.60	21.06	46.11
43	23.21	31.00	31.38	25.96	37.03	20.43	47.06
44	40.68	32.97	32.61	29.72	36.58	22.28	48.78
45	22.79	33.92	33.01	26.91	42.74	21.78	52.81
46	36.45	34.31	33.68	30.52	38.58	23.11	50.95
47	35.30	33.64	33.77	31.12	36.35	22.87	49.46
48	23.52	34.16	33.96	31.70	36.82	23.25	50.21
49	34.99	35.00	34.18	31.21	39.25	23.59	51.93
50	30.89	34.08	34.39	31.22	37.19	23.13	50.21
51	21.92	34.28	34.43	30.92	38.01	23.18	50.71
52	28.92	35.22	35.07	30.40	40.81	23.48	52.82
53	27.95	36.21	35.57	31.36	41.81	24.17	54.24
54	25.69	36.53	36.21	33.51	39.81	24.80	53.81
55	22.40	35.83	36.49	32.67	39.30	24.29	52.86
56	34.40	37.26	38.36	32.41	42.83	24.91	55.72
57	45.66	41.74	40.09	36.98	47.11	28.08	62.06
58	25.27	39.80	40.25	36.68	43.18	27.04	58.57
59	29.21	40.52	40.84	34.52	47.56	26.88	61.07
60	29.80	39.98	41.09	34.14	46.81	26.55	60.20
61	37.97	40.73	41.15	36.33	45.66	27.45	60.43
62	49.61	41.54	41.34	37.72	45.75	28.13	61.34

Nr.	Budg.Costs	LM Pred	CV Pred	LM CI lower	LM CI upper	LM PI lower	LM PI upper
	[MNOK]	[MNOK].	[MNOK].	[MNOK]	[MNOK]	[MNOK]	[MNOK]
63	44.11	40.88	41.59	36.98	45.19	27.66	60.42
64	46.88	42.99	42.80	38.78	47.66	29.06	63.59
65	42.32	43.91	44.02	38.83	49.65	29.52	65.32
66	54.35	45.00	44.34	40.75	49.70	30.45	66.50
67	40.84	45.85	45.91	40.87	51.43	30.89	68.04
68	36.91	44.45	46.11	39.23	50.36	29.86	66.16
69	54.15	46.45	46.18	42.84	50.37	31.57	68.35
70	30.29	45.50	46.58	39.78	52.05	30.47	67.94
71	36.93	45.44	46.88	39.64	52.07	30.41	67.89
72	52.41	46.02	46.99	41.72	50.77	31.15	67.99
73	53.55	48.67	47.79	42.55	55.67	32.60	72.67
74	65.71	47.85	47.80	43.72	52.37	32.45	70.55
75	38.63	46.55	48.37	40.83	53.07	31.21	69.43
76	51.43	48.52	48.83	43.69	53.88	32.79	71.80
77	58.39	48.75	48.96	43.02	55.24	32.75	72.56
78	45.34	47.28	49.19	39.69	56.33	31.18	71.69
79	42.15	49.81	49.91	45.16	54.94	33.72	73.58
80	54.06	51.61	52.00	46.41	57.38	34.86	76.40
81	31.06	52.53	52.05	48.60	56.78	35.72	77.24
82	49.47	53.63	52.32	48.46	59.36	36.27	79.30
83	72.29	52.81	52.47	48.41	57.61	35.84	77.81
84	38.51	53.71	52.50	48.34	59.67	36.29	79.49
85	45.36	55.05	53.43	49.15	61.66	37.11	81.66
86	43.30	54.24	53.89	50.14	58.68	36.88	79.77
87	70.32	54.19	56.42	48.90	60.06	36.64	80.15
88	51.34	56.36	57.24	50.20	63.29	37.97	83.67
89	61.61	68.22	68.17	59.28	78.52	45.60	102.08
90	56.60	75.94	76.32	63.63	90.62	50.04	115.23

Appendix 5: R code

All relevant coding used for data handling, data analysis and plotting is given below.

```
# Intropart: Loading packages, read file, read columns correctly, make adjustments

# Packages used in this script
library(MASS)
library(car)
library(DAAG)
library(gvlma)
library(leaps)
library(bootstrap)
library(bootStepAIC)
library(ggplot2)
library(cowplot)
library(dplyr)
library(boot)
library(reshape2)
library(ggfortify)
library(gmodels)

# Saving default graphical parameters
def.par <- par(no.readonly = TRUE) # save default, for resetting...

phi <- 1.61803399
graphics.off()
wi=9/2.51; he=9/2.51; windows(width = wi, height = he);
wi=14/2.51; he=(14/phi)/2.51; windows(width = wi, height = he);
wi=7.38/2.51; he=(7.38/phi)/2.51; windows(width = wi, height = he);
wi=23/2.51; he=(23/phi)/2.51; windows(width = wi, height = he);
wi=18/2.51; he=(19/phi)/2.51; windows(width = wi, height = he);
wi=23/2.51; he=(23/3*2)/2.51; windows(width = wi, height = he);

setwd("C:/Users/tobe/Desktop/Master_lokal")

#############################################################################
4.3.                                  Data                                  handling
##########################################################################

# Read power plant data
Rselection  <-  read.csv("Reported_costs3.csv",  na.strings=c("NULL",  "-9999",  "IO",
"00.01.1900", ""), header=TRUE)
head(Rselection)

# Set selected cells with no input to 0, not NA ---------
Rselection$Dam1_Height_R[is.na(Rselection$Dam1_Height_R)] <- 0
Rselection$Dam1_Length_R [is.na(Rselection$Dam1_Length_R )] <- 0
Rselection$Dam2_Height_R [is.na(Rselection$Dam2_Height_R )] <- 0
Rselection$Dam2_Length_R [is.na(Rselection$Dam2_Length_R )] <- 0
Rselection$Penstock1_Length_R [is.na(Rselection$Penstock1_Length_R )] <- 0
Rselection$Penstock1_Dia_R [is.na(Rselection$Penstock1_Dia_R )] <- 0
Rselection$Penstock2_Length_R [is.na(Rselection$Penstock2_Length_R )] <- 0
Rselection$Penstock2_Dia_R [is.na(Rselection$Penstock2_Dia_R )] <- 0
Rselection$Tunnel_Length_R [is.na(Rselection$Tunnel_Length_R )] <- 0
Rselection$Tunnel_Cross_Sect_R [is.na(Rselection$Tunnel_Cross_Sect_R )] <- 0
Rselection$Shaft_Length_R [is.na(Rselection$Shaft_Length_R )] <- 0
Rselection$Shaft_Cross_Sect_R [is.na(Rselection$Shaft_Cross_Sect_R )] <- 0
Rselection$Turbine2_Effect_R [is.na(Rselection$Turbine2_Effect_R )] <- 0
Rselection$Turbine2_Abs_Cap_R [is.na(Rselection$Turbine2_Abs_Cap_R )] <- 0
Rselection$Generator1_Cap_R [is.na(Rselection$Generator1_Cap_R )] <- 0
Rselection$Generator2_Cap_R [is.na(Rselection$Generator2_Cap_R )] <- 0

# Penstock type ---------
Rselection$Penstock1_Type_R <- as.character.factor(Rselection$Penstock1_Type_R)
Rselection$Penstock2_Type_R <- as.character.factor(Rselection$Penstock2_Type_R)

Rselection$Penstock1_Type_R [Rselection$Penstock1_Type_R == "Duktil"] <- "Duktile"
Rselection$Penstock2_Type_R [Rselection$Penstock2_Type_R == "Duktil"] <- "Duktile"
Rselection$Penstock1_Type_R [Rselection$Penstock1_Type_R == "Duktilt"] <- "Duktile"
Rselection$Penstock2_Type_R [Rselection$Penstock2_Type_R == "Duktilt"] <- "Duktile"
Rselection$Penstock1_Type_R [Rselection$Penstock1_Type_R == "Duktilt"] <- "Duktile"
Rselection$Penstock2_Type_R [Rselection$Penstock2_Type_R == "Duktilt"] <- "Duktile"
Rselection$Penstock1_Type_R [Rselection$Penstock1_Type_R == "DSJ"] <- "Duktile"
```

```
Rselection$Penstock2_Type_R [Rselection$Penstock2_Type_R == "DSJ"] <- "Duktile"
Rselection$Penstock1_Type_R [Rselection$Penstock1_Type_R == "GJS"] <- "Duktile"
Rselection$Penstock2_Type_R [Rselection$Penstock2_Type_R == "GJS"] <- "Duktile"
Rselection$Penstock1_Type_R [Rselection$Penstock1_Type_R == "STJ"] <- "Duktile"
Rselection$Penstock2_Type_R [Rselection$Penstock2_Type_R == "STJ"] <- "Duktile"
Rselection$Penstock1_Type_R [Rselection$Penstock1_Type_R == "K10"] <- "Duktile"
Rselection$Penstock2_Type_R [Rselection$Penstock2_Type_R == "K10"] <- "Duktile"
Rselection$Penstock1_Type_R [Rselection$Penstock1_Type_R == "K9"] <- "Duktile"
Rselection$Penstock2_Type_R [Rselection$Penstock2_Type_R == "K9"] <- "Duktile"
Rselection$Penstock1_Type_R [Rselection$Penstock1_Type_R == "GUP"] <- "GRP"
Rselection$Penstock2_Type_R [Rselection$Penstock2_Type_R == "GUP"] <- "GRP"

attach(Rselection)
Rselection$Penstock_Types_R <- "NA"
for (t in 1:nrow(Rselection)) {
  if (isTRUE(Rselection$Penstock1_Type_R[t] == "Duktile" &
Rselection$Penstock2_Type_R[t] == "GRP")) Rselection$Penstock_Types_R[t] <-
"GRP_Duc"
  else
    if (isTRUE(Rselection$Penstock1_Type_R[t] == "GRP" &
Rselection$Penstock2_Type_R[t] == "Duktile")) Rselection$Penstock_Types_R[t] <-
"GRP_Duc"
    else
      if (isTRUE(Rselection$Penstock1_Type_R[t] == "GRP/duktil"))
Rselection$Penstock_Types_R[t] <- "GRP_Duc"
      else
        if (isTRUE(Rselection$Penstock2_Type_R[t] == "GRP/duktil"))
Rselection$Penstock_Types_R[t] <- "GRP_Duc"
        else
          if (isTRUE(Rselection$Penstock1_Type_R[t] == "GRP" &
Rselection$Penstock2_Type_R[t] == "PE")) Rselection$Penstock_Types_R[t] <- "GRP_PE"
          else
            if (isTRUE(Rselection$Penstock1_Type_R[t] == "PE" &
Rselection$Penstock2_Type_R[t] == "GRP")) Rselection$Penstock_Types_R[t] <- "GRP_PE"
            else
              if (isTRUE(Rselection$Penstock1_Type_R[t] == "PE" &
Rselection$Penstock2_Type_R[t] == "Duktile")) Rselection$Penstock_Types_R[t] <-
"PE_Duk"
              else
                if (isTRUE(Rselection$Penstock1_Type_R[t] == "Duktile" &
Rselection$Penstock2_Type_R[t] == "PE")) Rselection$Penstock_Types_R[t] <- "PE_Duk"
                else
                  if (isTRUE(Rselection$Penstock1_Type_R[t] == "Stål" &
Rselection$Penstock2_Type_R[t] == "Duktile")) Rselection$Penstock_Types_R[t] <-
"St_Duk"
                  else
                    if (isTRUE(Rselection$Penstock1_Type_R[t] == "Duktile" &
Rselection$Penstock2_Type_R[t] == "Stål")) Rselection$Penstock_Types_R[t] <-
"St_Duk"
                    if (isTRUE(Rselection$Penstock1_Type_R[t] == "GRP" &
Rselection$Penstock2_Type_R[t] == "Stål")) Rselection$Penstock_Types_R[t] <-
"GRP_St"
                    else
                      if (isTRUE(Rselection$Penstock1_Type_R[t] == "GJS/GRP"))
Rselection$Penstock_Types_R[t] <- "GRP_Duc"
                      else
                        if (isTRUE(Rselection$Penstock2_Type_R[t] == "GJS/GRP"))
Rselection$Penstock_Types_R[t] <- "GRP_Duc"
                        else
                          if (isTRUE(Rselection$Penstock_Types_R[t] == "NA"))
Rselection$Penstock_Types_R[t] <- Rselection$Penstock1_Type_R[t]
                          if (isTRUE(Rselection$Penstock_Types_R[t] == ""))
Rselection$Penstock_Types_R[t] <- "NA"
}
detach(Rselection)
Rselection$Penstock_Types_R <- as.factor(Rselection$Penstock_Types_R)

# Turbines -----
Rselection$Turbine1_Type_R <- as.character.factor(Rselection$Turbine1_Type_R)
Rselection$Turbine2_Type_R <- as.character.factor(Rselection$Turbine2_Type_R)

attach(Rselection)
Rselection$Turbine_Types_R <- "NA"
for (t in 1:nrow(Rselection)){
  if (isTRUE(Rselection$Turbine1_Type_R[t] == "Francis" &
Rselection$Turbine2_Type_R[t] == "Francis"))
    Rselection$Turbine_Types_R[t] <- "2xFrancis"
  else
    if (isTRUE(Rselection$Turbine1_Type_R[t] == "Pelton" &
Rselection$Turbine2_Type_R[t] == "Pelton"))
      Rselection$Turbine_Types_R[t] <- "2xPelton"
    else
      if (isTRUE(Rselection$Turbine1_Type_R[t] == "Francis" &
Rselection$Turbine2_Type_R[t] == "Pelton"))
        Rselection$Turbine_Types_R[t] <- "FrancisPelton"
      else
        if (isTRUE(Rselection$Turbine1_Type_R[t] == "Pelton" &
Rselection$Turbine2_Type_R[t] == "Francis"))
          Rselection$Turbine_Types_R[t] <- "FrancisPelton"
        else
          if (isTRUE(Rselection$Turbine_Types_R[t] == "NA"))
Rselection$Turbine_Types_R[t] <- Rselection$Turbine1_Type_R[t]
          else
            if (isTRUE(Rselection$Turbine1_Type_R[t] == ""))
Rselection$Turbine_Types_R[t] <- "NA"
}
Rselection$Turbine_Types_R <- as.factor(Rselection$Turbine_Types_R)
detach(Rselection)

Rselection$Turbine_Types_R2 <- "NA"
for (t in 1:nrow(Rselection)){
  if (isTRUE(Rselection$Turbine1_Type_R[t] == "Francis"))
    Rselection$Turbine_Types_R2[t] <- "Francis"
  else
    if (isTRUE(Rselection$Turbine1_Type_R[t] == "Pelton"))
      Rselection$Turbine_Types_R2[t] <- "Pelton"
    else
      if (isTRUE(Rselection$Turbine1_Type_R[t] != "Francis" |
Rselection$Turbine1_Type_R[t] != "Pelton" | Rselection$Turbine1_Type_R[t] == ""))
        Rselection$Turbine_Types_R2[t] <- "Other"
      if (is.na(Rselection$Turbine1_Type_R[t]))
        Rselection$Turbine_Types_R2[t] <- "Other"
}
Rselection$Turbine_Types_R2 <- as.factor(Rselection$Turbine_Types_R2)

# exclude variables ----
exclude <- names(Rselection) %in% c("Main_status", "Municipality",
"Power_Plant_Built", "Penstock_Cover", "Kommentar")
Rselection <- Rselection[!exclude]

# Correct reading of data ---------
Rselection$Cost_Date <- as.Date(Rselection$Cost_Date, format = "%d.%m.%Y")
Rselection$Main_status_date <- as.Date(Rselection$Main_status_date, format =
"%d.%m.%Y")
Rselection$Penstock_Dia <- as.numeric(Rselection$Penstock_Dia)
Rselection$Penstock_Length <- as.numeric(Rselection$Penstock_Length)
Rselection$Tunnel_Length <- as.numeric(Rselection$Tunnel_Length)
Rselection$Shaft_Dia <- as.numeric(Rselection$Shaft_Dia)
Rselection$Shaft_Length <- as.numeric(Rselection$Shaft_Length)
Rselection$Earth_Cable <- as.numeric(Rselection$Earth_Cable)
Rselection$Sea_Cable <- as.numeric(Rselection$Sea_Cable)
Rselection$Dam_Length <- as.numeric(Rselection$Dam_Length)
Rselection$Pwr_Station_Base <- as.numeric(Rselection$Pwr_Station_Base)
Rselection$Road_Length <- as.numeric(Rselection$Road_Length)
Rselection$No_Turbines <- as.numeric(Rselection$No_Turbines)
Rselection$VannKV_Yr <- as.numeric(Rselection$VannKV_Yr)
Rselection$Date_Operation_R <- as.Date(Rselection$Date_Operation_R, format =
"%d.%m.%Y")
Rselection$VannKV_Date <- as.Date(Rselection$VannKV_Date, format = "%d.%m.%Y")
Rselection$Byggestart <- as.Date(Rselection$Byggestart, format = "%d.%m.%Y")
Rselection$Byggestart_supplert <- as.Date(Rselection$Byggestart_supplert, format =
"%d.%m.%Y")
Rselection$Shaft_Length_R <- as.numeric(Rselection$Shaft_Length_R)
Rselection$Tunnel_Length_R <- as.numeric(Rselection$Tunnel_Length_R)

# Aggregation of som of the data
Rselection$Water_Way_Length <- Rselection$Penstock_Length + Rselection$Tunnel_Length
+ Rselection$Shaft_Length
Rselection$Connector_Length <- Rselection$Earth_Cable + Rselection$Sea_Cable +
Rselection$Pwr_Line
Rselection$Water_Way_Length_R <- Rselection$Penstock1_Length_R +
Rselection$Penstock2_Length_R + Rselection$Tunnel_Length_R +
Rselection$Shaft_Length_R
```

```
Rselection$Total_Dam_Length_R       <-       Rselection$Dam1_Length_R            +
Rselection$Dam2_Length_R
Rselection$Abs_Cap_R      <-      Rselection$Turbine1_Abs_Cap_R                  +
Rselection$Turbine2_Abs_Cap_R
Rselection$Penstock_Lengths_R                                                   <-
Rselection$Penstock1_Length_R+Rselection$Penstock2_Length_R
Rselection$Sum_Partial_Costs_R         <-         Rselection$Inlet_cost_R        +
Rselection$Penstock_Cost_R + Rselection$PP_cost_R
Rselection$Sum_Partial_Costs_R[Rselection$Sum_Partial_Costs_R == 0] <- NA

Rselection$Total_Costs3_R <- as.numeric(NA)
for(t in 1:nrow(Rselection)){
  if(is.na(Rselection$Total_Cost2_R[t]))
    Rselection$Total_Costs3_R[t] <- Rselection$Total_Costs_R[t]
  else
    Rselection$Total_Costs3_R[t]       <-       Rselection$Total_Cost2_R[t]       +
Rselection$Total_Costs_R[t]
}

Rselection <- mutate(Rselection,
                     Unadj_Spec_Tot_Cost_R = Total_Costs3_R/Ann_Prod_Est_R)
Rselection       <-       mutate(Rselection,       Unadj_Spec_Partial_Costs       =
Sum_Partial_Costs_R/Ann_Prod_Est_R)

# Create a single column with construction date
Rselection$Construction_Date <- as.Date("01.01.1900", format = "%d.%m.%Y")
for (t in 1:nrow(Rselection)){
  if (isTRUE(Rselection$Byggestart[t] > 0))
    Rselection$Construction_Date[t] <- Rselection$Byggestart[t]
  else
    Rselection$Construction_Date[t] <- Rselection$Byggestart_supplert[t]
}

Rselection$Construction_Time <- NA
for(t in 1:nrow(Rselection)){
  if(isTRUE((Rselection$Date_Operation_R[t] - Rselection$Construction_Date[t]) > 0))
    Rselection$Construction_Time[t]  <-  as.numeric(Rselection$Date_Operation_R[t]  -
Rselection$Construction_Date[t])
  else
    Rselection$Construction_Time[t] <- as.numeric(NA)
}
Rselection <- mutate(Rselection, Construction_Time_Yr = Construction_Time/365)

Years <- seq(from = 1900, to = 2016, by = 1)
Rselection$Construction_Year <- as.numeric(NA)
for(n in 1:length(Years)){
  for(i in 1:nrow(Rselection)) {
    if(isTRUE(Rselection$Construction_Date[i] >= paste(Years[n], "-01-01", sep="") &
Rselection$Construction_Date[i] <= paste(Years[n], "-12-31", sep="")))
      Rselection$Construction_Year[i] <- as.numeric(Years[n])
  }
}
Rselection$Construction_Year_fac <- as.factor(Rselection$Construction_Year)

Rselection   <-   mutate(Rselection,   Construction_Year0   =   Construction_Year-
min(Construction_Year, na.rm=TRUE)+1)

# Creating dummy variables for tunnell and shaft
Rselection$Waterway_Type <- "NA"
for(t in 1:nrow(Rselection)){
  if(isTRUE(Rselection$Tunnel_Length_R[t] > 0))
    Rselection$Waterway_Type[t] <- "Tunnel"
  else
    if(isTRUE(Rselection$Shaft_Length_R[t] > 0))
      Rselection$Waterway_Type[t] <- "Shaft"
    else
      if(isTRUE(Rselection$Penstock1_Length_R[t] > 0))
        Rselection$Waterway_Type[t] <- "Penstock"
}
Rselection$Waterway_Type <- as.factor(Rselection$Waterway_Type)

Rselection$Penstock_Bin <- as.logical("FALSE")
Rselection$Tunnel_Bin <- as.logical("FALSE")
Rselection$Shaft_Bin <- as.logical("FALSE")
for(t in 1:nrow(Rselection)){
  if(isTRUE(Rselection$Tunnel_Length_R[t] > 0))
    Rselection$Tunnel_Bin[t] <- as.logical("TRUE")
```

```
}

for(t in 1:nrow(Rselection)){
  if(isTRUE(Rselection$Shaft_Length_R[t] > 0))
    Rselection$Shaft_Bin[t] <- as.logical("TRUE")
}

for(t in 1:nrow(Rselection)){
  if(isTRUE(Rselection$Penstock1_Length_R[t] > 0))
    Rselection$Penstock_Bin[t] <- as.logical("TRUE")
}

Rselection$Penstock_Types_sub <- NA
for (t in 1:nrow(Rselection)) {
  if(isTRUE(Rselection$Penstock_Types_R[t] == "GRP"))
    Rselection$Penstock_Types_sub[t] <- "GRP"
  else
    if(isTRUE(Rselection$Penstock_Types_R[t] == "St_Duk"))
      Rselection$Penstock_Types_sub[t] <- "St_Duk"
    else
      if(isTRUE(Rselection$Penstock_Types_R[t] == "Duktile"))
        Rselection$Penstock_Types_sub[t] <- "Ductile"
      else
        if(isTRUE(Rselection$Penstock_Types_R[t] == "PE_Duk"))
          Rselection$Penstock_Types_sub[t] <- "PE_Duk"
        else
          if(isTRUE(Rselection$Penstock_Types_R[t]       ==       "GRP_Duc"       |
Rselection$Penstock_Types_R[t] == "GRP_PE" | Rselection$Penstock_Types_R[t] == "PE"
| Rselection$Penstock_Types_R[t] == "NA"))
            Rselection$Penstock_Types_sub[t] <- "Other"
          else
            if (is.na(Rselection$Penstock_Types_R[t]))
              Rselection$Penstock_Types_sub[t] <- "Other"
}
Rselection$Penstock_Types_sub <- as.factor(Rselection$Penstock_Types_sub)

# Adding missing start-up date to reported date from the record in Vannkraftdatabasen
for(t in 1:nrow(Rselection)){
  if(is.na(Rselection$Date_Operation_R[t]))
    Rselection$Date_Operation_R[t] <- Rselection$VannKV_Date[t]
}

# Converting start-up date to start-up year:
Years <- seq(from = 1900, to = 2016, by = 1)
Rselection$Operation_Year <- as.numeric(NA)
for(n in 1:length(Years)){
  for(i in 1:nrow(Rselection)) {
    if(isTRUE(Rselection$Date_Operation_R[i] >= paste(Years[n], "-01-01", sep="") &
Rselection$Date_Operation_R[i] <= paste(Years[n], "-12-31", sep="")))
      Rselection$Operation_Year[i] <- as.numeric(Years[n])
  }
}
Rselection$Operation_Year_fac <- as.factor(Rselection$Operation_Year)

# Recode year variable to start from 1 in 2005
Rselection    <-    mutate(Rselection,    Operation_Year0    =    Operation_Year-
min(Operation_Year, na.rm=TRUE)+1)

# Adjust investment cost for date of cost--------
# Read cost index data
IndexTable <- read.csv("Index_Table_Rev.csv", header = T)

# Adjust budgeted investment cost for date of cost--------
Rselection$Adj_Cost <- NA
for(n in 1:nrow(IndexTable)){
  for(i in 1:nrow(Rselection)) {
    if(isTRUE(Rselection$Cost_Date[i]   >=   paste(2016-n,   "-01-01",   sep="")   &
Rselection$Cost_Date[i]     <=     paste(2016-n,     "-12-31",     sep="")     &
Rselection$Gross_Head_R[i] < 300))
      Rselection$Adj_Cost[i]                                                   <-
Rselection$Est_Cost[i]/IndexTable$Small_Hydro_Plants[n]*IndexTable$Small_Hydro_Plan
ts[1]
    else
      if(isTRUE(Rselection$Cost_Date[i]   >=   paste(2016-n,   "-01-01",   sep="")   &
Rselection$Cost_Date[i]     <=     paste(2016-n,     "-12-31",     sep="")     &
Rselection$Gross_Head_R[i] >= 300))
```

```
        Rselection$Adj_Cost[i] <-
Rselection$Est_Cost[i]/IndexTable$High_Head_Plants[n]*IndexTable$High_Head_Plants[1
]
  }
}
# Specific budgeted total cost in real values
Rselection$Adj_Spec_Cost <- Rselection$Adj_Cost/Rselection$Production

# Index regulating total reported costs
Rselection$Adj_Tot_Cost_R <- as.numeric(NA)
for(n in 1:nrow(IndexTable)){
  for(i in 1:nrow(Rselection)) {
    if(isTRUE(Rselection$Construction_Date[i] >= paste(2016-n, "-01-01", sep="") &
Rselection$Construction_Date[i] <= paste(2016-n, "-12-31", sep="") &
Rselection$Gross_Head_R[i] < 300))
      Rselection$Adj_Tot_Cost_R[i] <-
Rselection$Total_Costs_R[i]/IndexTable$Small_Hydro_Plants[n]*IndexTable$Small_Hydro
_Plants[1]
    else
      if(isTRUE(Rselection$Construction_Date[i] >= paste(2016-n, "-01-01", sep="") &
Rselection$Construction_Date[i] <= paste(2016-n, "-12-31", sep="") &
Rselection$Gross_Head_R[i] >= 300))
        Rselection$Adj_Tot_Cost_R[i] <-
Rselection$Total_Costs_R[i]/IndexTable$High_Head_Plants[n]*IndexTable$High_Head_Pla
nts[1]
  }
}
# Specific total reported costs in real values
Rselection$Adj_Spec_Tot_Cost_R <-
Rselection$Adj_Tot_Cost_R/Rselection$Ann_Prod_Est_R

# Index regulating total reported costs 2, with extra column with misc., unknown
extra costs
Rselection$Adj_Tot_Cost2_R <- as.numeric(NA)
for(n in 1:nrow(IndexTable)){
  for(i in 1:nrow(Rselection)) {
    if(isTRUE(Rselection$Construction_Date[i] >= paste(2016-n, "-01-01", sep="") &
Rselection$Construction_Date[i] <= paste(2016-n, "-12-31", sep="") &
Rselection$Gross_Head_R[i] < 300))
      Rselection$Adj_Tot_Cost2_R[i] <-
Rselection$Total_Costs3_R[i]/IndexTable$Small_Hydro_Plants[n]*IndexTable$Small_Hydr
o_Plants[1]
    else
      if(isTRUE(Rselection$Construction_Date[i] >= paste(2016-n, "-01-01", sep="") &
Rselection$Construction_Date[i] <= paste(2016-n, "-12-31", sep="") &
Rselection$Gross_Head_R[i] >= 300))
        Rselection$Adj_Tot_Cost2_R[i] <-
Rselection$Total_Costs3_R[i]/IndexTable$High_Head_Plants[n]*IndexTable$High_Head_Pl
ants[1]
  }
}
# Specific total reported costs 2 in real values
Rselection$Spec_Adj_Tot_Cost2_R <-
Rselection$Adj_Tot_Cost2_R/Rselection$Ann_Prod_Est_R

# Index regulating reported partial costs
Rselection$Adj_Partial_Costs_R <- as.numeric(NA)
for(n in 1:nrow(IndexTable)){
  for(i in 1:nrow(Rselection)) {
    if(isTRUE(Rselection$Construction_Date[i] >= paste(2016-n, "-01-01", sep="") &
Rselection$Construction_Date[i] <= paste(2016-n, "-12-31", sep="") &
Rselection$Gross_Head_R[i] < 300))
      Rselection$Adj_Partial_Costs_R[i] <-
Rselection$Sum_Partial_Costs_R[i]/IndexTable$Small_Hydro_Plants[n]*IndexTable$Small
_Hydro_Plants[1]
    else
      if(isTRUE(Rselection$Construction_Date[i] >= paste(2016-n, "-01-01", sep="") &
Rselection$Construction_Date[i] <= paste(2016-n, "-12-31", sep="") &
Rselection$Gross_Head_R[i] >= 300))
        Rselection$Adj_Partial_Costs_R[i] <-
Rselection$Sum_Partial_Costs_R[i]/IndexTable$High_Head_Plants[n]*IndexTable$High_He
ad_Plants[1]
  }
}
Rselection$Adj_Partial_Costs_R[(Rselection$Adj_Partial_Costs_R==0)] <- NA
Rselection$Adj_Spec_Part_Cost_R <-
Rselection$Adj_Partial_Costs_R/Rselection$Ann_Prod_Est_R
```

```
# Costs per installed capacity for reported total costs in nominal and real values
Rselection <- mutate(Rselection,
                     Nom_Cost_per_MW = Total_Costs3_R/Max_Effect_R,
                     Adj_Cost_per_MW = Adj_Tot_Cost2_R/Max_Effect_R)

# Create a new variable with average penstock diameter weighted by penstock length -
---
Rselection$Penstock1_Dia_R[Rselection$Penstock1_Dia_R==0] <- NA
Rselection$Adj_Penstock_Dia <- as.numeric(NA)
for (t in 1:nrow(Rselection)){
  if(isTRUE(Rselection$Penstock2_Dia_R[t] > 0))
    Rselection$Adj_Penstock_Dia[t]                                              <-
((Rselection$Penstock1_Dia_R[t]*Rselection$Penstock1_Length_R[t]+Rselection$Penstoc
k2_Dia_R[t]*Rselection$Penstock2_Length_R[t])/(Rselection$Penstock1_Length_R[t]+Rse
lection$Penstock2_Length_R[t]))
  else
    Rselection$Adj_Penstock_Dia[t] <- Rselection$Penstock1_Dia_R[t]
}

# Creating an aggregated variable with average dam height weighted by dam length for
power plants with two dams.
Rselection$Adj_Dam_Height <- as.numeric(NA)
for (t in 1:nrow(Rselection)){
  if(isTRUE(Rselection$Dam2_Height_R[t] > 0))
    Rselection$Adj_Dam_Height[t]                                                <-
((Rselection$Dam1_Height_R[t]*Rselection$Dam1_Length_R[t]+Rselection$Dam2_Height_R[
t]*Rselection$Dam2_Length_R[t])/(Rselection$Dam1_Length_R[t]+Rselection$Dam2_Length
_R[t]))
  else
    Rselection$Adj_Dam_Height[t] <- Rselection$Dam1_Height_R[t]
}

# Grouping licence holders into professional and non-professional project deveopers.
Classified by
SHP_owner_classification <- read.csv(file = "SHP_owner_classification.csv")

# merge two data frames by ID and Country
Rselection <- merge(Rselection,SHP_owner_classification,by="Kdb_ID")
Rselection <- arrange(Rselection, desc(Date_Operation_R))
select(Rselection, Kdb_ID, VannKVnavn, Holder, Company_class)

# Grouping counties into regions
Rselection$Region <- NA
Rselection$County <- as.character(Rselection$County)
for (t in 1:nrow(Rselection)){
  if  (Rselection$County[t]  ==  "Troms"  |  Rselection$County[t]  ==  "Nordland"  |
Rselection$County[t] == "Finnmark")
    Rselection$Region[t] <- "Northern Norway"
  else
    if  (Rselection$County[t]  ==  "Nord-Trøndelag"  |  Rselection$County[t]  ==  "Sør-
Trøndelag")
      Rselection$Region[t] <- "Trøndelag"
    else
      if  (Rselection$County[t]  ==  "Hordaland"  |  Rselection$County[t]  ==  "Møre  og
Romsdal"  |  Rselection$County[t]  ==  "Rogaland"  |  Rselection$County[t]  ==  "Sogn  og
Fjordane")
        Rselection$Region[t] <- "Western Norway"
      else
        if  (Rselection$County[t]  ==  "Aust-Agder"  |  Rselection$County[t]  ==  "Vest-
Agder")
          Rselection$Region[t] <- "Southern Norway"
        else
          if (Rselection$County[t] == "Telemark" | Rselection$County[t] == "Buskerud"
|   Rselection$County[t]   ==   "Hedmark"   |   Rselection$County[t]   ==   "Oppland"   |
Rselection$County[t]    ==    "Akershus"    |    Rselection$County[t]    ==    "Oslo"    |
Rselection$County[t] == "Vestfold" | Rselection$County[t] == "Østfold")
            Rselection$Region[t] <- "Eastern Norway"
}
Rselection$Region <- as.factor(Rselection$Region)
Rselection$County <- as.factor(Rselection$County)

# Sorting county factor by county number
Rselection$County                                                                  =
factor(Rselection$County,levels(Rselection$County)[c(13,7,6,11,5,10,4,9,14,1,12,2,8
,3)])
```

```
Rselection$County = factor(Rselection$County,levels(Rselection$County)[seq(14,1,-
1)])
Rselection$Region =
factor(Rselection$Region,levels(Rselection$Region)[c(2,4,5,3,1)])
Rselection$Region = factor(Rselection$Region,levels(Rselection$Region)[seq(5,1,-1)])

#Write all data into a new table
write.csv(Rselection, file = "Full_dataset_2016_05_09.csv")

#Summary of the dataset
summary(Rselection)

##############################################################################
5. The dataset
##############################################################################

# Table with Summarized costs
Rselection$Adj_Partial_Costs_R[(Rselection$Adj_Partial_Costs_R==0)] <- NA
Rselection$Adj_Tot_Cost_R[(Rselection$Adj_Tot_Cost_R==0)] <- NA
Rselection$Inlet_cost_R[(Rselection$Inlet_cost_R==0)] <- NA
Rselection$Penstock_Cost_R[(Rselection$Penstock_Cost_R==0)] <- NA
Rselection$PP_cost_R[(Rselection$PP_cost_R==0)] <- NA

# Real total costs grouped
Adj_Tot_Cost_R_by_waterway <- summarise(group_by(Rselection, Waterway_Type), Min =
min(Adj_Tot_Cost_R, na.rm=TRUE), Q1=quantile(Adj_Tot_Cost_R, probs=0.25,
na.rm=TRUE), Q2=quantile (Adj_Tot_Cost_R, probs=0.50, na.rm=TRUE), Mean =
mean(Adj_Tot_Cost_R, na.rm=TRUE), Med=median(Adj_Tot_Cost_R, na.rm=TRUE),
Q3=quantile(Adj_Tot_Cost_R, probs=0.75, na.rm=TRUE), Max = max(Adj_Tot_Cost_R,
na.rm=TRUE), n = n()-sum(is.na(Adj_Tot_Cost_R)), NAs = sum(is.na(Adj_Tot_Cost_R)))
write.csv(Adj_Tot_Cost_R_by_waterway, file = "Adj_Tot_Cost_R_by_waterway2.csv")

# Real total costs ungrouped
Adj_Tot_Cost_R <- summarise(Rselection, Min = min(Adj_Tot_Cost_R, na.rm=TRUE),
Q1=quantile(Adj_Tot_Cost_R, probs=0.25, na.rm=TRUE), Q2=quantile (Adj_Tot_Cost_R,
probs=0.50, na.rm=TRUE), Mean = mean(Adj_Tot_Cost_R, na.rm=TRUE),
Med=median(Adj_Tot_Cost_R, na.rm=TRUE), Q3=quantile(Adj_Tot_Cost_R, probs=0.75,
na.rm=TRUE), Max = max(Adj_Tot_Cost_R, na.rm=TRUE), n = n()-
sum(is.na(Adj_Tot_Cost_R)), NAs = sum(is.na(Adj_Tot_Cost_R)))
write.csv(Adj_Tot_Cost_R, file = "Adj_Tot_Cost_R2.csv")

# Real sum of partial costs grouped
Adj_Partial_Costs_R_by_waterway <- summarise(group_by(Rselection, Waterway_Type),
Min = min(Adj_Partial_Costs_R, na.rm=TRUE), Q1=quantile(Adj_Partial_Costs_R,
probs=0.25, na.rm=TRUE), Q2=quantile (Adj_Partial_Costs_R, probs=0.50, na.rm=TRUE),
Mean = mean(Adj_Partial_Costs_R, na.rm=TRUE), Med=median(Adj_Partial_Costs_R,
na.rm=TRUE), Q3=quantile(Adj_Partial_Costs_R, probs=0.75, na.rm=TRUE), Max =
max(Adj_Partial_Costs_R, na.rm=TRUE), n = n()-sum(is.na(Adj_Partial_Costs_R)), NAs =
sum(is.na(Adj_Partial_Costs_R)))
write.csv(Adj_Partial_Costs_R_by_waterway, file =
"Adj_Partial_Costs_R_by_waterway.csv")

# Real partial costs ungrouped
Adj_Partial_Costs_R <- summarise(Rselection, Min = min(Adj_Partial_Costs_R,
na.rm=TRUE), Q1=quantile(Adj_Partial_Costs_R, probs=0.25, na.rm=TRUE), Q2=quantile
(Adj_Partial_Costs_R, probs=0.50, na.rm=TRUE), Mean = mean(Adj_Partial_Costs_R,
na.rm=TRUE), Med=median(Adj_Partial_Costs_R, na.rm=TRUE),
Q3=quantile(Adj_Partial_Costs_R, probs=0.75, na.rm=TRUE), Max =
max(Adj_Partial_Costs_R, na.rm=TRUE), n = n()-sum(is.na(Adj_Partial_Costs_R)), NAs =
sum(is.na(Adj_Partial_Costs_R)))
write.csv(Adj_Partial_Costs_R, file = "Adj_Partial_Costs_R.csv")

# Relative dam and inlet costs by waterway type
Rel_Inlet_cost_R_by_waterway <- summarise(group_by(Rselection, Waterway_Type), Min =
min(Inlet_cost_R/Total_Costs_R*100, na.rm=TRUE),
Q1=quantile(Inlet_cost_R/Total_Costs_R*100, probs=0.25, na.rm=TRUE), Q2=quantile
(Inlet_cost_R/Total_Costs_R*100, probs=0.50, na.rm=TRUE), Mean =
mean(Inlet_cost_R/Total_Costs_R*100, na.rm=TRUE),
Med=median(Inlet_cost_R/Total_Costs_R*100, na.rm=TRUE),
Q3=quantile(Inlet_cost_R/Total_Costs_R*100, probs=0.75, na.rm=TRUE), Max =
max(Inlet_cost_R/Total_Costs_R*100, na.rm=TRUE), n = n()-
sum(is.na(Inlet_cost_R/Total_Costs_R*100)), NAs =
sum(is.na(Inlet_cost_R/Total_Costs_R*100)))
write.csv(Rel_Inlet_cost_R_by_waterway, file = "Rel_Inlet_cost_R_by_waterway2.csv")

# Relative intake costs ungrouped
```

```
Rel_Inlet_cost_R <- summarise(Rselection, Min = min(Inlet_cost_R/Total_Costs_R*100,
na.rm=TRUE), Q1=quantile(Inlet_cost_R/Total_Costs_R*100, probs=0.25, na.rm=TRUE),
Q2=quantile (Inlet_cost_R/Total_Costs_R*100, probs=0.50, na.rm=TRUE), Mean =
mean(Inlet_cost_R/Total_Costs_R*100, na.rm=TRUE),
Med=median(Inlet_cost_R/Total_Costs_R*100, na.rm=TRUE),
Q3=quantile(Inlet_cost_R/Total_Costs_R*100, probs=0.75, na.rm=TRUE), Max =
max(Inlet_cost_R/Total_Costs_R*100, na.rm=TRUE), n = n()-sum(is.na(Inlet_cost_R)),
NAs = sum(is.na(Inlet_cost_R)))
write.csv(Rel_Inlet_cost_R, file = "Rel_Inlet_cost_R2.csv")

#Relative waterway costs by waterway type
Rel_Penstock_Cost_R_by_waterway <- summarise(group_by(Rselection, Waterway_Type),
Min = min(Penstock_Cost_R/Total_Costs_R*100, na.rm=TRUE),
Q1=quantile(Penstock_Cost_R/Total_Costs_R*100, probs=0.25, na.rm=TRUE), Q2=quantile
(Penstock_Cost_R/Total_Costs_R*100, probs=0.50, na.rm=TRUE), Mean =
mean(Penstock_Cost_R/Total_Costs_R*100, na.rm=TRUE),
Med=median(Penstock_Cost_R/Total_Costs_R*100, na.rm=TRUE),
Q3=quantile(Penstock_Cost_R/Total_Costs_R*100, probs=0.75, na.rm=TRUE), Max =
max(Penstock_Cost_R/Total_Costs_R*100, na.rm=TRUE), n = n()-
sum(is.na(Penstock_Cost_R)), NAs = sum(is.na(Penstock_Cost_R)))
write.csv(Rel_Penstock_Cost_R_by_waterway, file =
"Rel_Penstock_Cost_R_by_waterway2.csv")

#Relative waterway costs ungrouped
Rel_Penstock_Cost_R <- summarise(Rselection, Min =
min(Penstock_Cost_R/Total_Costs_R*100, na.rm=TRUE),
Q1=quantile(Penstock_Cost_R/Total_Costs_R*100, probs=0.25, na.rm=TRUE), Q2=quantile
(Penstock_Cost_R/Total_Costs_R*100, probs=0.50, na.rm=TRUE), Mean =
mean(Penstock_Cost_R/Total_Costs_R*100, na.rm=TRUE),
Med=median(Penstock_Cost_R/Total_Costs_R*100, na.rm=TRUE),
Q3=quantile(Penstock_Cost_R/Total_Costs_R*100, probs=0.75, na.rm=TRUE), Max =
max(Penstock_Cost_R/Total_Costs_R*100, na.rm=TRUE), n = n()-
sum(is.na(Penstock_Cost_R)), NAs = sum(is.na(Penstock_Cost_R)))
write.csv(Rel_Penstock_Cost_R, file = "Rel_Penstock_Cost_R2.csv")

# Relative powerplant costs by waterway type
Rel_PP_cost_R_by_waterway <- summarise(group_by(Rselection, Waterway_Type), Min =
min(PP_cost_R/Total_Costs_R*100, na.rm=TRUE),
Q1=quantile(PP_cost_R/Total_Costs_R*100, probs=0.25, na.rm=TRUE), Q2=quantile
(PP_cost_R/Total_Costs_R*100, probs=0.50, na.rm=TRUE), Mean =
mean(PP_cost_R/Total_Costs_R*100, na.rm=TRUE),
Med=median(PP_cost_R/Total_Costs_R*100, na.rm=TRUE),
Q3=quantile(PP_cost_R/Total_Costs_R*100, probs=0.75, na.rm=TRUE), Max =
max(PP_cost_R/Total_Costs_R*100, na.rm=TRUE), n = n()-sum(is.na(PP_cost_R)), NAs =
sum(is.na(PP_cost_R)))
write.csv(Rel_PP_cost_R_by_waterway, file = "Rel_PP_cost_R_by_waterway2.csv")

# Relative PP-costs ungrouped
Rel_PP_cost_R <- summarise(Rselection, Min = min(PP_cost_R/Total_Costs_R*100,
na.rm=TRUE), Q1=quantile(PP_cost_R/Total_Costs_R*100, probs=0.25, na.rm=TRUE),
Q2=quantile (PP_cost_R/Total_Costs_R*100, probs=0.50, na.rm=TRUE), Mean =
mean(PP_cost_R/Total_Costs_R*100, na.rm=TRUE),
Med=median(PP_cost_R/Total_Costs_R*100, na.rm=TRUE),
Q3=quantile(PP_cost_R/Total_Costs_R*100, probs=0.75, na.rm=TRUE), Max =
max(PP_cost_R/Total_Costs_R*100, na.rm=TRUE), n = n()-sum(is.na(PP_cost_R)), NAs =
sum(is.na(PP_cost_R)))
write.csv(Rel_PP_cost_R, file = "Rel_PP_cost_R2.csv")

#Relative sum of partial costs costs by waterway type
Rel_sum_partial_costs_R_by_waterway <- summarise(group_by(Rselection,
Waterway_Type), Min = min(Sum_Partial_Costs_R/Total_Costs_R*100, na.rm=TRUE),
Q1=quantile(Sum_Partial_Costs_R/Total_Costs_R*100, probs=0.25, na.rm=TRUE),
Q2=quantile (Sum_Partial_Costs_R/Total_Costs_R*100, probs=0.50, na.rm=TRUE), Mean =
mean(Sum_Partial_Costs_R/Total_Costs_R*100, na.rm=TRUE),
Med=median(Sum_Partial_Costs_R/Total_Costs_R*100, na.rm=TRUE),
Q3=quantile(Sum_Partial_Costs_R/Total_Costs_R*100, probs=0.75, na.rm=TRUE), Max =
max(Sum_Partial_Costs_R/Total_Costs_R*100, na.rm=TRUE), n = n()-
sum(is.na(Sum_Partial_Costs_R)), NAs = sum(is.na(Sum_Partial_Costs_R)))
write.csv(Rel_sum_partial_costs_R_by_waterway, file =
"Rel_sum_partial_costs_R_by_waterway2.csv")

#Relative sum of partial costs ungrouped
Rel_Partial_Costs <- summarise(Rselection, Min =
min(Sum_Partial_Costs_R/Total_Costs_R*100, na.rm=TRUE),
Q1=quantile(Sum_Partial_Costs_R/Total_Costs_R*100, probs=0.25, na.rm=TRUE),
Q2=quantile (Sum_Partial_Costs_R/Total_Costs_R*100, probs=0.50, na.rm=TRUE), Mean =
mean(Sum_Partial_Costs_R/Total_Costs_R*100, na.rm=TRUE),
Med=median(Sum_Partial_Costs_R/Total_Costs_R*100, na.rm=TRUE),
Q3=quantile(Sum_Partial_Costs_R/Total_Costs_R*100, probs=0.75, na.rm=TRUE), Max =
max(Sum_Partial_Costs_R/Total_Costs_R*100, na.rm=TRUE), n = n()-
sum(is.na(Sum_Partial_Costs_R)), NAs = sum(is.na(Sum_Partial_Costs_R)))
write.csv(Rel_Partial_Costs, file = "Rel_Partial_Costs2.csv")

################################################################################
6.1 Two-sample tests on difference between budgeted and actual costs
############################################################################

Relative_Costs <- data.frame(Rselection$Est_Cost, Rselection$Total_Costs3_R)
Relative_Costs <- setNames(Relative_Costs, c("Est_Cost","Total_Costs_R"))
Relative_Costs$Spec_Total_Costs_R <-
(Rselection$Total_Costs3_R/Rselection$Ann_Prod_Est_R)
Relative_Costs$Spec_Total_Costs <-Rselection$Spec_Cost2
Relative_Costs$Rel_Est_Cost <- Rselection$Est_Cost/Rselection$Est_Cost*100
Relative_Costs$Rel_Spec_Est_Cost <- Rselection$Spec_Cost2/Rselection$Spec_Cost2*100
Relative_Costs$Rel_Total_Costs_R <-
Rselection$Total_Costs3_R/Rselection$Est_Cost*100
Relative_Costs$Rel_Spec_Total_Costs_R <-
(Rselection$Total_Costs3_R/Rselection$Ann_Prod_Est_R)/Rselection$Spec_Cost2*100
Relative_Costs$Rel_Adj_Cost <- Rselection$Adj_Cost/Rselection$Adj_Cost*100
Relative_Costs$Rel_Spec_Adj_Cost <-
Rselection$Adj_Spec_Cost/Rselection$Adj_Spec_Cost*100
Relative_Costs$Rel_Adj_Tot_Cost_R <-
Rselection$Adj_Tot_Cost2_R/Rselection$Adj_Cost*100
Relative_Costs$Rel_Spec_Adj_Tot_Cost_R <-
Rselection$Spec_Adj_Tot_Cost2_R/Rselection$Adj_Spec_Cost*100
Relative_Costs$Rel_Partial_Costs_R <-
Rselection$Sum_Partial_Costs_R/Rselection$Est_Cost*100
Relative_Costs$Rel_Spec_Partial_Costs_R <-
Rselection$Unadj_Spec_Partial_Costs/Rselection$Adj_Spec_Cost*100
Relative_Costs$Rel_Adj_Partial_Costs_R <-
Rselection$Adj_Partial_Costs_R/Rselection$Adj_Cost*100
Relative_Costs$Rel_Spec_Adj_Partial_Costs_R <-
Rselection$Adj_Spec_Part_Cost_R/Rselection$Adj_Spec_Cost*100
Relative_Costs$Sum_Partial_Costs_R <- Rselection$Sum_Partial_Costs_R
Relative_Costs$Unadj_Spec_Partial_Costs <- Rselection$Unadj_Spec_Partial_Costs
Relative_Costs$Adj_Tot_Cost_R <- Rselection$Adj_Tot_Cost2_R
Relative_Costs$Adj_Spec_Tot_Cost_R <- Rselection$Spec_Adj_Tot_Cost2_R
Relative_Costs$Adj_Cost <- Rselection$Adj_Cost
Relative_Costs$Adj_Spec_Cost <- Rselection$Adj_Spec_Cost
Relative_Costs$Adj_Partial_Costs_R <- Rselection$Adj_Partial_Costs_R
Relative_Costs$Adj_Spec_Partial_Costs_R <- Rselection$Adj_Spec_Part_Cost_R

# Uncorrected budgeted vs actual costs
attach(Relative_Costs)
t.test(Total_Costs_R, Est_Cost, paired=TRUE, alternative="two.sided", var.equal =
TRUE)
shapiro.test((Total_Costs_R-Est_Cost))
var(Total_Costs_R)
var(Est_Cost)
wilcox.test(Total_Costs_R, Est_Cost,paired=TRUE, alternative="two.sided",
conf.int=TRUE)

t.test(Rel_Total_Costs_R, Rel_Est_Cost, paired=TRUE, alternative="two.sided",
var.equal = TRUE)
shapiro.test((Rel_Total_Costs_R-Rel_Est_Cost))
var(Rel_Total_Costs_R)
var(Rel_Est_Cost)
wilcox.test(Rel_Total_Costs_R, Rel_Est_Cost,paired=TRUE, alternative="two.sided",
conf.int=TRUE)

# Inflation/index adjusted budgeted vs. actual costs
t.test(Adj_Tot_Cost_R, Adj_Cost,paired=TRUE, alternative="two.sided", var.equal =
TRUE)
shapiro.test((Adj_Tot_Cost_R-Adj_Cost))
var(Adj_Tot_Cost_R, na.rm = TRUE)
var(Adj_Cost, na.rm = TRUE)
wilcox.test(Adj_Tot_Cost_R, Adj_Cost,paired=TRUE, alternative="two.sided",
conf.int=TRUE)

t.test(Rel_Adj_Tot_Cost_R, Rel_Adj_Cost,paired=TRUE, alternative="two.sided",
var.equal = TRUE)
shapiro.test((Rel_Adj_Tot_Cost_R-Rel_Adj_Cost))
var(Rel_Adj_Tot_Cost_R, na.rm = TRUE)
var(Rel_Adj_Cost, na.rm = TRUE)
```

```
wilcox.test(Rel_Adj_Tot_Cost_R, Rel_Adj_Cost,paired=TRUE, alternative="two.sided",
conf.int=TRUE)

# Partial costs
t.test(Sum_Partial_Costs_R, Est_Cost, paired=TRUE, alternative="two.sided",
var.equal = TRUE)
shapiro.test((Sum_Partial_Costs_R-Est_Cost))
var(Sum_Partial_Costs_R, na.rm = TRUE)
var(Est_Cost, na.rm = TRUE)
wilcox.test(Sum_Partial_Costs_R, Est_Cost,paired=TRUE, alternative="two.sided",
conf.int=TRUE)

t.test(Rel_Partial_Costs_R, Rel_Est_Cost, paired=TRUE, alternative="two.sided",
var.equal = TRUE)
shapiro.test((Rel_Partial_Costs_R-Rel_Est_Cost))
var(Rel_Partial_Costs_R, na.rm = TRUE)
var(Rel_Est_Cost, na.rm = TRUE)
wilcox.test(Rel_Partial_Costs_R, Rel_Est_Cost,paired=TRUE, alternative="two.sided",
conf.int=TRUE)

t.test(Adj_Partial_Costs_R, Adj_Cost, paired=TRUE, alternative="two.sided",
var.equal = TRUE)
shapiro.test((Adj_Partial_Costs_R-Adj_Cost))
var(Adj_Partial_Costs_R, na.rm = TRUE)
var(Adj_Cost, na.rm = TRUE)
wilcox.test(Adj_Partial_Costs_R, Adj_Cost,paired=TRUE, alternative="two.sided",
conf.int=TRUE)

t.test(Rel_Adj_Partial_Costs_R, Rel_Adj_Cost, paired=TRUE, alternative="two.sided",
var.equal = TRUE)
shapiro.test((Rel_Adj_Partial_Costs_R-Rel_Adj_Cost))
var(Rel_Adj_Partial_Costs_R, na.rm = TRUE)
var(Rel_Adj_Cost, na.rm = TRUE)
wilcox.test(Rel_Adj_Partial_Costs_R, Rel_Adj_Cost,paired=TRUE,
alternative="two.sided", conf.int=TRUE)

# Specific costs
t.test(Spec_Total_Costs_R, Spec_Total_Costs, paired=TRUE, alternative="two.sided",
var.equal = TRUE)
shapiro.test((Spec_Total_Costs_R-Spec_Total_Costs))
var(Spec_Total_Costs_R, na.rm = TRUE)
var(Spec_Total_Costs, na.rm = TRUE)
wilcox.test(Spec_Total_Costs_R, Spec_Total_Costs,paired=TRUE,
alternative="two.sided", conf.int=TRUE)

t.test(Adj_Spec_Tot_Cost_R, Adj_Spec_Cost, paired=TRUE, alternative="two.sided",
var.equal = TRUE)
shapiro.test((Adj_Spec_Tot_Cost_R-Adj_Spec_Cost))
var(Adj_Spec_Tot_Cost_R, na.rm = TRUE)
var(Adj_Spec_Cost, na.rm = TRUE)
wilcox.test(Adj_Spec_Tot_Cost_R, Adj_Spec_Cost,paired=TRUE, alternative="two.sided",
conf.int=TRUE)

t.test(Unadj_Spec_Partial_Costs, Spec_Total_Costs, paired=TRUE,
alternative="two.sided", var.equal = TRUE)
shapiro.test((Unadj_Spec_Partial_Costs-Spec_Total_Costs))
var(Unadj_Spec_Partial_Costs, na.rm = TRUE)
var(Spec_Total_Costs, na.rm = TRUE)
wilcox.test(Unadj_Spec_Partial_Costs, Spec_Total_Costs,paired=TRUE,
alternative="two.sided", conf.int=TRUE)

t.test(Adj_Spec_Partial_Costs_R, Adj_Spec_Cost, paired=TRUE,
alternative="two.sided", var.equal = TRUE)
shapiro.test((Adj_Spec_Partial_Costs_R-Adj_Spec_Cost))
var(Adj_Spec_Partial_Costs_R, na.rm = TRUE)
var(Adj_Spec_Cost, na.rm = TRUE)
wilcox.test(Adj_Spec_Partial_Costs_R, Adj_Spec_Cost,paired=TRUE,
alternative="two.sided", conf.int=TRUE)
detach(Relative_Costs)

attach(Rselection)
t.test(Ann_Prod_Est_R, Production, paired=TRUE, alternative="two.sided", var.equal =
TRUE)
shapiro.test((Ann_Prod_Est_R-Production))
range((Ann_Prod_Est_R-Production))
summary((Ann_Prod_Est_R-Production))
hist((Ann_Prod_Est_R-Production))
```

```
var(Ann_Prod_Est_R, na.rm = TRUE)
var(Production, na.rm = TRUE)
wilcox.test(Ann_Prod_Est_R, Production,paired=TRUE, alternative="two.sided",
conf.int=TRUE)
detach(Rselection)

ggplot(Rselection, aes(x=Prod_Diff)) + theme_grey() +
    geom_histogram(binwidth=0.5,
                   colour="black", fill="dark green", alpha = 0.5)

###############################################################################
6.2.1 Year of construction
###########################################################################

Rselection_years <- filter(Rselection, Construction_Year > 0)

give.n <- function(x){
  return(c(y = 1.1, label = length(x)))
}

Rselection_long <- select(Rselection_years, Kdb_ID, Unadj_Spec_Partial_Costs,
Unadj_Spec_Tot_Cost_R, Adj_Spec_Part_Cost_R, Spec_Adj_Tot_Cost2_R,
Construction_Year_fac)
Rselection_long <- melt(Rselection_long,
id.vars=c("Kdb_ID","Construction_Year_fac"), na.rm=TRUE)

boxplot_years_spec <- ggplot(Rselection_long, aes(x=variable, y=value,
fill=variable))+geom_boxplot()+facet_grid(.~Construction_Year_fac) +
labs(x="Years",y="Specific Costs [NOK/kWh annual production]") + theme_grey() +
theme(axis.ticks = element_blank(), axis.text.x = element_blank()) +
scale_fill_discrete(name="Reported costs", breaks=c("Unadj_Spec_Partial_Costs",
"Unadj_Spec_Tot_Cost_R", "Adj_Spec_Part_Cost_R", "Spec_Adj_Tot_Cost2_R"),
labels=c("Spec.Part.Costs", "Spec.Tot. Costs", "Adj.Spec.Part.Costs",
"Adj.Spec.Tot.Costs")) + coord_cartesian(ylim = c(0, 8)) + stat_summary(fun.data =
give.n, geom = "text", size = 2.5)

dev.set(5)
boxplot_years_spec
ggsave("boxplot_years_spec_costs_03-05-16.png")

#Rselection_years0 <- filter(Rselection, Unadj_Spec_Tot_Cost_R > 0,
Construction_Year0 > 0)
Rselection_years <- filter(Rselection, Unadj_Spec_Tot_Cost_R > 0
                           # , Unadj_Spec_Tot_Cost_R <6
                           , Construction_Year0 > 0)

years.lm <- lm(Unadj_Spec_Tot_Cost_R ~ Construction_Year0, data = Rselection_years2)
summary(years.lm)
anova(years.lm)
confint(years.lm)

Rselection_years2 <- filter(Rselection, Total_Costs3_R < 150)
Rselection_years2 <- arrange(Rselection_years2, Total_Costs3_R)
select(Rselection_years2, Total_Costs3_R)
years2.lm <- lm(log(Total_Costs3_R) ~ Construction_Year0, data = Rselection_years2)
summary(years2.lm)
gvmodel <- gvlma(years2.lm)
summary(gvmodel)
anova(years2.lm)

shapiro.test(years.lm$residuals)
layout(matrix(c(1,2,3,4),2,2))
plot(years.lm)
par(def.par)

Costs_per_MW <- filter(Rselection, Construction_Year0 > 0
                       #, Unadj_Spec_Tot_Cost_R <6
)
Costs_per_MW$Nom_Cost_per_MW <-
Costs_per_MW$Total_Costs3_R/Costs_per_MW$Max_Effect_R
Costs_per_MW$Adj_Cost_per_MW <-
Costs_per_MW$Adj_Tot_Cost2_R/Costs_per_MW$Max_Effect_R
Costs_per_MW <- arrange(Costs_per_MW, Adj_Cost_per_MW)
Costs_per_MW <- Costs_per_MW[(Costs_per_MW$Nom_Cost_per_MW < 20),]
Costs_per_MW$Nom_Cost_per_MW <-
Costs_per_MW$Total_Costs3_R/Costs_per_MW$Max_Effect_R
```

```
Costs_per_MW$Adj_Cost_per_MW <-
Costs_per_MW$Adj_Tot_Cost2_R/Costs_per_MW$Max_Effect_R
Costs_per_MW <- arrange(Costs_per_MW, Adj_Cost_per_MW)
Costs_per_MW2 <- filter(Costs_per_MW
                        , Tunnel_Bin == FALSE
                        #, Adj_Cost_per_MW < 15
                        , Adj_Cost_per_MW > 0)
Costs_per_MW2$Adj_Cost_per_MW

Costs_per_MW <- arrange(Costs_per_MW, Max_Effect_R)
select(Costs_per_MW, Max_Effect_R, Adj_Tot_Cost2_R, Adj_Cost_per_MW)

years_MW.lm <- lm(log(Nom_Cost_per_MW) ~ log(Construction_Year0), data =
Costs_per_MW)
gvmodel <- gvlma(years_MW.lm)
summary(gvmodel)

plot(Costs_per_MW$Nom_Cost_per_MW ~ Costs_per_MW$Construction_Year0)

#summary(years_MW.lm)
#anova(years_MW.lm)

layout(matrix(c(1,2,3,4),2,2))
plot(years_MW.lm)
shapiro.test(years_MW.lm$residuals)
par(def.par)

years_MW_adj.lm <- lm(Adj_Cost_per_MW ~ Construction_Year0, data = Costs_per_MW2)
gvmodel <- gvlma(years_MW_adj.lm)
summary(gvmodel)
#anova(years_MW_adj.lm)
layout(matrix(c(1,2,3,4),2,2))
plot(years_MW_adj.lm)
shapiro.test(years_MW_adj.lm$residuals)
par(def.par)

Rselection_years2 <- filter(Rselection,Spec_Adj_Tot_Cost2_R < 6.99,
Spec_Adj_Tot_Cost2_R > 0
                            , Construction_Year0 > 0
)
years_adj.lm <- lm(Spec_Adj_Tot_Cost2_R ~ Construction_Year0, data =
Rselection_years2)
summary(years_adj.lm)
anova(years_adj.lm)
confint(years_adj.lm)

x_1 <- coef(summary(years.lm))[2, "Estimate"]
x_2 <- coef(summary(years_adj.lm))[2, "Estimate"]
s_1 <- coef(summary(years.lm))[2, "Std. Error"]
s_2 <- coef(summary(years_adj.lm))[2, "Std. Error"]
n_1 <- length(years.lm$fitted.values)
n_2 <- length(years_adj.lm$fitted.values)

s_x1x2 <- sqrt(s_1^2/n_1 + s_2^2/n_2)
t <- (x_1-x_2)/s_x1x2
df <- ((s_1^2/n_1)+(s_2^2/n_2))^2/
  (
    ((s_1^2/n_1)^2/(n_1-1)) + ((s_2^2/n_2)^2/(n_2-1))
  )

abs(qt(0.05/2, df))

2*pt(-abs(t),df=df)

# s_x1x2 <- sqrt(coef(summary(years.lm))[2, "Std.
Error"]^2/(length(years.lm$fitted.values))+coef(summary(years_adj.lm))[2, "Std.
Error"]^2/(length(years_adj.lm$fitted.values)))

# t <- (coef(summary(years.lm))[2, "Estimate"]-coef(summary(years_adj.lm))[2,
"Estimate"])/s_x1x2

years_adj2.lm <- lm(Spec_Adj_Tot_Cost2_R ~ Construction_Year0 + Construction_Time_Yr,
data = Rselection_years)
summary(years_adj2.lm)
anova(years_adj2.lm)

layout(matrix(c(1,2,3,4),2,2))
```

```
plot(years_adj.lm)
shapiro.test(years_adj.lm$residuals)
par(def.par)

###############################################################################
6.2.2 Construction time
###########################################################################

Rselection_Construction_Time <- Rselection
# Rselection_Construction_Time <- mutate(Rselecion_Construction_Time,
log10Adj_Tot_Cost_R = log10(Adj_Tot_Cost_R))
Rselection_Construction_Time0 <- filter(Rselection, Adj_Tot_Cost2_R > 0,
Construction_Time_Yr > 0)
Rselection_Construction_Time <- filter(Rselection
                                       , Adj_Tot_Cost2_R > 0
                                       , Adj_Tot_Cost2_R < 150
                                       , Construction_Time_Yr < 3
                                       , Construction_Time_Yr > 0)
#
# construction_time.lm <- lm(Adj_Tot_Cost2_R ~ Construction_Time_Yr, data =
Rselection)
# summary(construction_time.lm)
# anova(construction_time.lm)
# shapiro.test(construction_time.lm$residuals)
# layout(matrix(c(1,2,3,4),2,2))
# plot(construction_time.lm)
# par(def.par)

construction_time.lm_log <- lm(log(Adj_Tot_Cost2_R) ~ Construction_Time_Yr, data =
Rselection_Construction_Time)
summary(construction_time.lm_log)
anova(construction_time.lm_log)
gvmodel <- gvlma(construction_time.lm_log)
summary(gvmodel)
shapiro.test(construction_time.lm_log$residuals)

layout(matrix(c(1,2,3,4),2,2))
plot(construction_time.lm_log)
par(def.par)

Predictions_Test <- predict(construction_time.lm_log, newdata =
Rselection_Construction_Time, interval = "confidence")
Rselection_Construction_Time$preds <- exp(Predictions_Test[,1])
Rselection_Construction_Time$lowerCI <- exp(Predictions_Test[,2])
Rselection_Construction_Time$upperCI <- exp(Predictions_Test[,3])

construction_time_scatter <- ggplot(Rselection_Construction_Time,
aes(x=Construction_Time_Yr, y=Adj_Tot_Cost2_R)) +
  geom_point() +
  theme_grey() +
  labs(x="Construction time in years",y="Total investment costs [MNOK] in real
values") +
  geom_line(aes(x=Construction_Time_Yr, y=preds), colour = "blue", size = 1)+
  geom_ribbon(aes(ymin=lowerCI, ymax=upperCI), alpha=0.2) +
  scale_y_continuous(breaks=seq(0, 140, 20))

dev.set(5)
construction_time_scatter
ggsave("construction_time_scatter_07-05-15.png")

###############################################################################
6.2.3 Geography
###########################################################################

give.n <- function(x){
  return(c(y = (min(x)-0.3), label = length(x)))
}

boxplot_geography <- ggplot(Rselection, aes(factor(Region), y=Spec_Adj_Tot_Cost2_R,
fill=factor(County)))+
  geom_boxplot() +
  labs(x="Counties grouped by region",y="Adjusted Specific Total Costs [NOK/KWh]") +
  theme_grey() +
  stat_summary(fun.data = give.n, geom = "text", fun.y = median,
position=position_dodge(width=0.75)) +
  guides(fill=guide_legend(title = "Counties")) +
  scale_y_continuous(breaks=seq(0, 10, 2))
```

```
dev.set(5)
boxplot_geography
ggsave("boxplot_geography_08-05-16.png")

Rselection_region <- filter(Rselection, Spec_Adj_Tot_Cost2_R > 0)
summarize(Rselection_region,                              mean(Spec_Adj_Tot_Cost2_R),
median(Spec_Adj_Tot_Cost2_R))

krusk_county <- kruskal.test(Spec_Adj_Tot_Cost2_R ~ County, data = Rselection_region)
aov_County <- aov(Spec_Adj_Tot_Cost2_R ~ County, data=Rselection_region)
krusk_county
summary(aov_County)
shapiro.test(aov_County$residuals)

krusk_region <- kruskal.test(Spec_Adj_Tot_Cost2_R ~ Region, data = Rselection_region)
aov_region <- aov(Spec_Adj_Tot_Cost2_R ~ Region, data=Rselection_region)
krusk_region
summary(aov_region)
shapiro.test(aov_region$residuals)

Rselection_region <- filter(Rselection, Spec_Adj_Tot_Cost2_R <6.5)
Rselection_region$Region                                                            =
factor(Rselection_region$Region,levels(Rselection_region$Region)[c(3,4,1,2,5)])
Rselection_region$County                                                            =
factor(Rselection_region$County,levels(Rselection_region$County)[c(9,1,2,3,4,5,6,7,
8,10,11,12,13,14)])

krusk_county <- kruskal.test(Spec_Adj_Tot_Cost2_R ~ County, data = Rselection_region)
aov_County <- aov(Spec_Adj_Tot_Cost2_R ~ County, data=Rselection_region)
krusk_county
summary(aov_County)
shapiro.test(aov_County$residuals)

krusk_region <- kruskal.test(Spec_Adj_Tot_Cost2_R ~ Region, data = Rselection_region)
aov_region <- aov(Spec_Adj_Tot_Cost2_R ~ Region, data=Rselection_region)
krusk_region
summary(aov_region)
shapiro.test(aov_region$residuals)

lm_county <- lm(Spec_Adj_Tot_Cost2_R ~ County, data = Rselection_region)
aov(lm_county)
gvmodel <- gvlma(lm_county)
summary(gvmodel)
shapiro.test(lm_county$residuals)

layout(matrix(c(1,2,3,4),2,2))
plot(lm_county)
par(def.par)

lm_region <- lm(Spec_Adj_Tot_Cost2_R ~ Region, data = Rselection_region)
aov(lm_region)
gvmodel <- gvlma(lm_region)
summary(gvmodel)
shapiro.test(lm_region$residuals)

layout(matrix(c(1,2,3,4),2,2))
plot(lm_region)
par(def.par)

###############################################################################
6.2.4 LICENSE                  HOLDER                   AND                   COST
###########################################################################

# Boxplot
boxplot_holder_spec        <-        ggplot(Rselection,        aes(x=Company_class,
y=Spec_Adj_Tot_Cost2_R))+geom_boxplot(aes(fill=Company_class))    +    labs(x="SHPP
owner/developer classification",y="Specific Total Costs [NOK/KWh] in real values") +
guides(fill=FALSE) + theme_grey() +  stat_summary(fun.data = give.n, geom = "text")
+ scale_x_discrete(breaks=c("pro", "non-pro"), labels=c("Professional \ndeveloper",
"Non-professional \ndeveloper")) + theme(axis.title.x = element_blank())

dev.set(2)
boxplot_holder_spec # se plottet
ggsave("boxplot_boxplot_holder_spec_small_8-5-16.png")

t.test(Rselection$Spec_Adj_Tot_Cost2_R ~ Rselection$Company_class, var.equal=FALSE)
```

```
wilcox.test(Rselection$Spec_Adj_Tot_Cost2_R ~ Rselection$Company_class,
alternative="two.sided", conf.int=TRUE, paired = FALSE)
shapiro.test(Rselection$Spec_Adj_Tot_Cost2_R[Rselection$Company_class == "non-pro"])
shapiro.test(Rselection$Spec_Adj_Tot_Cost2_R[Rselection$Company_class == "pro"])
med_non_pro <- median(Rselection$Spec_Adj_Tot_Cost2_R[Rselection$Company_class ==
"non-pro"], na.rm=TRUE)
med_pro <- median(Rselection$Spec_Adj_Tot_Cost2_R[Rselection$Company_class ==
"pro"], na.rm=TRUE)
med_pro - med_non_pro

t.test(Rselection$Adj_Cost_per_MW ~ Rselection$Company_class, var.equal=FALSE)
wilcox.test(Rselection$Adj_Cost_per_MW ~ Rselection$Company_class,
alternative="two.sided", conf.int=TRUE, paired = FALSE)
shapiro.test(Rselection$Adj_Cost_per_MW[Rselection$Company_class == "non-pro"])
shapiro.test(Rselection$Adj_Cost_per_MW[Rselection$Company_class == "pro"])

###############################################################################
6.3.1 Prediction Model 1 total costs
###########################################################################

Rselection5 <- Rselection[(Rselection$Tunnel_Bin==FALSE),]
include <- names(Rselection5) %in% c("Kdb_ID","Adj_Cost", "Adj_Partial_Costs_R",
"Adj_Tot_Cost_R", "Total_Costs_R", "Max_Effect_R", "Adj_Penstock_Dia",
"Construction_Year0","Construction_Time_Yr", "Water_Way_Length_R", "Adj_Dam_Height",
"Total_Dam_Length_R", "Shaft_Bin")
Rselection5 <- Rselection5[include]
Rselection5 <- arrange(Rselection5, Adj_Tot_Cost_R)
Rselection5 <- Rselection5[-c(127,126,70, 41),]
Subset3 <- na.omit(Rselection5)
summary(Subset3)
Subset3[(Subset3$Adj_Dam_Height ==0),]

# full.model6_log
full.model6_log <- lm(log(Adj_Tot_Cost_R) ~
                      Max_Effect_R
                      +I(Max_Effect_R^2)
                      +Adj_Dam_Height
                      +I(Adj_Penstock_Dia^2)
                      +log(Water_Way_Length_R)
                      +Construction_Year0
                      +Construction_Time_Yr
                      +Shaft_Bin
                      , data = Rselection5)
summary(full.model6_log)
press(full.model6_log)
sqrt(press(full.model6_log)/length(full.model6_log$fitted.values))

par(mfrow = c(1, 2))
diagnostics_plot_partial_costs_incl_time <-
  autoplot(full.model6_log, which = 1:6, ncol = 3, label.size = 3)+ theme_grey()
dev.set(5)
diagnostics_plot_partial_costs_incl_time

vif(full.model6_log) # variance inflation factors
gvmodel <- gvlma(full.model6_log)
summary(gvmodel)
shapiro.test(full.model6_log$residuals)

# Plot with the model
predicted <- predict(full.model6_log, Subset3, se.fit=TRUE, interval = "prediction")
predicted_ci <- predict(full.model6_log, Subset3, se.fit=TRUE, interval =
"confidence")
full.model6_log_cv <- cv.lm(data=Subset3, form.lm=full.model6_log, m= nrow(Subset3),
plotit = F)
full.model6_log_cv$pred_lower <- predicted$fit[,2]
full.model6_log_cv$pred_upper <- predicted$fit[,3]
full.model6_log_cv$lm_predict <- predicted$fit[,1]
full.model6_log_cv$ci_lower <- predicted_ci$fit[,2]
full.model6_log_cv$ci_upper <- predicted_ci$fit[,3]

full.model6_log_cv_trans <- mutate(full.model6_log_cv,
                                   True_costs = exp(log(Adj_Tot_Cost_R)),
                                   Budget_Costs = Adj_Cost,
                                   LM_Prediction = exp(lm_predict),
                                   CV_Prediction = exp(cvpred),
                                   True_costs_Rel = True_costs/True_costs*100,
                                   LM_Prediction_Rel = (LM_Prediction -
True_costs)/True_costs*100,
                                   CV_Prediction_Rel = (CV_Prediction -
True_costs)/True_costs*100,
                                   Budget_costs_rel = (Adj_Cost -
True_costs)/True_costs*100,
                                   LM_Prediction_Abs_Perc_Diff = abs(LM_Prediction -
True_costs)/True_costs*100,
                                   CV_Prediction_Abs_Perc_Diff = abs(CV_Prediction -
True_costs)/True_costs*100,
                                   Budget_costs_Abs_Perc_Diff = abs(Adj_Cost -
True_costs)/True_costs*100,
                                   Pred_diff = LM_Prediction - True_costs,
                                   Pred_diff_Rel = (LM_Prediction -
True_costs)/True_costs*100,
                                   Budg_diff = Adj_Cost - True_costs,
                                   Budg_diff_Rel = (Adj_Cost -
True_costs)/True_costs*100,
                                   PI_upper = exp(pred_upper),
                                   PI_lower = exp(pred_lower),
                                   CI_upper = exp(ci_upper),
                                   CI_lower = exp(ci_lower)
)

summarize(full.model6_log_cv_trans, min(CV_Prediction_Abs_Perc_Diff),
max(CV_Prediction_Abs_Perc_Diff), min(CV_Prediction_Rel), max(CV_Prediction_Rel))

full.model6_log_cv_trans <- arrange(full.model6_log_cv_trans, LM_Prediction)
full.model6_log_cv_trans$Index <-
seq.int(along.with=full.model6_log_cv_trans$LM_Prediction)

CV_mod2_investment_cost_scatter <- ggplot(full.model6_log_cv_trans, aes(x=Index)) +
  geom_ribbon(aes(ymin=PI_lower, ymax=PI_upper, alpha="PI_lower")) +
  geom_ribbon(aes(ymin=CI_lower, ymax=CI_upper, alpha="CI_lower")) +
  scale_alpha_manual(name = "PI and CI intervals", breaks = c("PI_lower",
"CI_lower"), labels = c("Prediction interval", "Confidence interval"),
values=c(0.4,0.2))+
  geom_point(aes(y = CV_Prediction, color="CV_Prediction", shape = "CV_Prediction"),
size =2)+
  geom_point(aes(y=Budget_Costs, color="Budget_Costs", shape = "Budget_Costs"), size
=2) +
  geom_point(aes(y=True_costs, color="True_costs", shape = "True_costs"), size =2)+
  theme_grey() +
  labs(x="Rank of linear model estimates",y="Total investment costs [MNOK] in real
values") +
  scale_colour_discrete("Cost types", breaks = c("Budget_Costs", "True_costs",
"CV_Prediction"), labels = c("Budget costs", "Actual costs", "Cross-validated
\nprediction estimate"))+
  scale_shape_discrete("Cost types", breaks = c("Budget_Costs", "True_costs",
"CV_Prediction"), labels = c("Budget costs", "Actual costs", "Cross-validated
\nprediction estimate"))+
  geom_line(aes(y=LM_Prediction, linetype="LM_Prediction"), size = 1, color="white",
alpha = 0.8) +
  scale_linetype_manual(name="Linear model \nprediction estimate",
breaks=c("LM_Prediction"), labels =c("White line showing \nprediction estimate"),
values = c("solid"))+
  coord_cartesian(ylim = c(0, 135), xlim = c(0,100)) +
  scale_y_continuous(breaks=seq(0, 140, 20)) +
  scale_x_continuous(breaks=seq(0, 120, 20))

dev.set(5)
CV_mod2_investment_cost_scatter
ggsave("LM1_Total_Cost_cost_scatter_7-5-16.png")

full.model6_log_cv_long2 <- select(full.model6_log_cv_trans, Kdb_ID, Index,
Budget_costs_rel, CV_Prediction_Rel, LM_Prediction_Rel)
full.model6_log_cv_long2 <- arrange(full.model6_log_cv_long2, CV_Prediction_Rel)
full.model6_log_cv_long2$Index <-
seq.int(along.with=full.model6_log_cv_long2$CV_Prediction_Rel)
full.model6_log_cv_long2 <- melt(full.model6_log_cv_long2, id.vars=c("Kdb_ID",
"Index"), na.rm=TRUE)

CV_mod1_standardized_cost_scatter <- ggplot(full.model6_log_cv_long2, aes(x=Index,
y=value)) +
  geom_point(aes(colour=variable, shape = variable, fill = variable), size =2)+
  theme_grey() +
  scale_fill_discrete(name ="Cost types deviation", breaks = c("LM_Prediction_Rel",
"Budget_costs_rel", "CV_Prediction_Rel"), labels = c("Deviations LM cost
\nestimates", "Deviations budgeted \ncosts", "Deviations CV cost \nestimates"))+
  scale_colour_discrete(name ="Cost types deviation", breaks = c("LM_Prediction_Rel",
"Budget_costs_rel", "CV_Prediction_Rel"), labels = c("Deviations LM cost
\nestimates", "Deviations budgeted \ncosts", "Deviations CV cost \nestimates"))+
  scale_shape_discrete(name ="Cost types deviation", breaks = c("LM_Prediction_Rel",
"Budget_costs_rel", "CV_Prediction_Rel"), labels = c("Deviations LM cost
\nestimates", "Deviations budgeted \ncosts", "Deviations CV cost \nestimates"))+
  labs(x="Rank of deviations between cross-validated estimates and actual
costs",y="Percent deviation from reported cost") +
  geom_hline(yintercept=0)

dev.set(5)
CV_mod1_standardized_cost_scatter
ggsave("CV_standardized_cost_scatter_1-5-16-2.png")

attach(full.model6_log_cv_trans)

t.test(CV_Prediction_Abs_Perc_Diff, Budget_costs_Abs_Perc_Diff,
alternative="two.sided", paired = FALSE, var.equal = FALSE) #
shapiro.test(CV_Prediction_Abs_Perc_Diff)
shapiro.test(Budget_costs_Abs_Perc_Diff)
wilcox.test(CV_Prediction_Abs_Perc_Diff, Budget_costs_Abs_Perc_Diff,
alternative="two.sided", conf.int=TRUE, paired = FALSE) #
#  t-test
t.test(CV_Prediction_Rel, Budg_diff_Rel, alternative="two.sided", paired = FALSE,
var.equal = FALSE) #
shapiro.test(CV_Prediction_Rel)
shapiro.test(Budg_diff_Rel)
wilcox.test(CV_Prediction_Rel, Budg_diff_Rel, alternative="two.sided",
conf.int=TRUE, paired = FALSE) #

t.test(CV_Prediction_Abs_Perc_Diff, Budget_costs_Abs_Perc_Diff,
alternative="two.sided", paired = TRUE, var.equal = FALSE) #
shapiro.test(CV_Prediction_Abs_Perc_Diff)
shapiro.test(Budget_costs_Abs_Perc_Diff)
var(CV_Prediction_Abs_Perc_Diff)
var(Budget_costs_Abs_Perc_Diff)
wilcox.test(CV_Prediction_Abs_Perc_Diff, Budget_costs_Abs_Perc_Diff,
alternative="two.sided", conf.int=TRUE, paired = TRUE) #
#  t-test
t.test(CV_Prediction_Rel, Budg_diff_Rel, alternative="two.sided", paired = TRUE,
var.equal = FALSE) #
shapiro.test(CV_Prediction_Rel)
shapiro.test(Budg_diff_Rel)
var(CV_Prediction_Rel)
var(Budg_diff_Rel)
wilcox.test(CV_Prediction_Rel, Budg_diff_Rel, alternative="two.sided",
conf.int=TRUE, paired = TRUE) #
detach(full.model6_log_cv_trans)

full.model6_log_cv_trans$budg_high <- "TRUE"
full.model6_log_cv_trans$budg_low <- "TRUE"
full.model6_log_cv_trans$budg_eq <- "TRUE"
full.model6_log_cv_trans$lm_high <- "TRUE"
full.model6_log_cv_trans$lm_low <- "TRUE"
full.model6_log_cv_trans$lm_eq <- "TRUE"
full.model6_log_cv_trans$eq_sign_lm_bud_diff <- "TRUE"
for(t in 1:nrow(full.model6_log_cv_trans)) {
  ifelse(isTRUE(full.model6_log_cv_trans$Budget_Costs[t] >
full.model6_log_cv_trans$True_costs[t]), full.model6_log_cv_trans$budg_high[t] <-
"TRUE", full.model6_log_cv_trans$budg_high[t] <-"FALSE")
  ifelse(isTRUE(full.model6_log_cv_trans$Budget_Costs[t] <
full.model6_log_cv_trans$True_costs[t]), full.model6_log_cv_trans$budg_low[t] <-
"TRUE", full.model6_log_cv_trans$budg_low[t] <- "FALSE")
  ifelse(isTRUE(full.model6_log_cv_trans$Budget_Costs[t] ==
full.model6_log_cv_trans$True_costs[t]), full.model6_log_cv_trans$budg_eq[t] <-
"TRUE", full.model6_log_cv_trans$budg_eq[t] <- "FALSE")
  ifelse(isTRUE(full.model6_log_cv_trans$CV_Prediction[t] >
full.model6_log_cv_trans$True_costs[t]), full.model6_log_cv_trans$lm_high[t] <-
"TRUE", full.model6_log_cv_trans$lm_high[t] <- "FALSE")
  ifelse(isTRUE(full.model6_log_cv_trans$CV_Prediction[t] <
full.model6_log_cv_trans$True_costs[t]), full.model6_log_cv_trans$lm_low[t] <-
"TRUE", full.model6_log_cv_trans$lm_low[t] <- "FALSE")
```

```
  ifelse(isTRUE(full.model6_log_cv_trans$CV_Prediction[t]                    ==
full.model6_log_cv_trans$True_costs[t]),    full.model6_log_cv_trans$lm_eq[t]    <-
"TRUE", full.model6_log_cv_trans$lm_eq[t] <- "FALSE")

  ifelse(isTRUE(
    (full.model6_log_cv_trans$CV_Prediction[t]                                  >
full.model6_log_cv_trans$True_costs[t]
     &full.model6_log_cv_trans$Budget_Costs[t]                                  >
full.model6_log_cv_trans$True_costs[t])
    |(full.model6_log_cv_trans$CV_Prediction[t]                                 <
full.model6_log_cv_trans$True_costs[t]
      &full.model6_log_cv_trans$Budget_Costs[t]                                 <
full.model6_log_cv_trans$True_costs[t]))
    ,full.model6_log_cv_trans$eq_sign_lm_bud_diff[t] <- "TRUE"
    , full.model6_log_cv_trans$eq_sign_lm_bud_diff[t] <- "FALSE")
}
full.model6_log_cv_trans$budg_high <- as.factor(full.model6_log_cv_trans$budg_high)
full.model6_log_cv_trans$budg_low <- as.factor(full.model6_log_cv_trans$budg_low)
full.model6_log_cv_trans$budg_eq <- as.factor(full.model6_log_cv_trans$budg_eq)
full.model6_log_cv_trans$lm_high <- as.factor(full.model6_log_cv_trans$lm_high)
full.model6_log_cv_trans$lm_low <- as.factor(full.model6_log_cv_trans$lm_low)
full.model6_log_cv_trans$lm_eq <- as.factor(full.model6_log_cv_trans$lm_eq)
full.model6_log_cv_trans$eq_sign_lm_bud_diff                                   <-
as.factor(full.model6_log_cv_trans$eq_sign_lm_bud_diff)

full.model6_log_cv_trans$eq_sign_lm_bud_diff                                   <-
as.factor(full.model6_log_cv_trans$eq_sign_lm_bud_diff)

summary(full.model6_log_cv_trans$eq_sign_lm_bud_diff)

# 2-Way Cross Tabulation
Cross_table_costs         <-         CrossTable(full.model6_log_cv_trans$budg_high,
full.model6_log_cv_trans$lm_low, chisq=TRUE)
chisq.test(full.model6_log_cv_trans$budg_high, full.model6_log_cv_trans$lm_low)

###############################################################################
6.3.2          Prediction          Model          2          partial          costs
###########################################################################

Rselection6 <- Rselection[(Rselection$Tunnel_Length_R==0),]
include  <-  names(Rselection6)  %in%  c("Kdb_ID","Adj_Cost",  "Adj_Partial_Costs_R",
"Adj_Tot_Cost_R",      "Total_Costs_R",      "Max_Effect_R",      "Adj_Penstock_Dia",
"Construction_Year0", "Company_class", "Construction_Time", "Construction_Time_Yr",
"Shaft_Bin", "Water_Way_Length_R")
Rselection6 <- Rselection6[include]
Rselection6 <- arrange(Rselection6, Adj_Partial_Costs_R)
Rselection6 <- Rselection6[-c(106, 65, 16, 11, 72,96, 29, 77, 81, 92, 53, 91, 27),]
Subset6 <- na.omit(Rselection6)
summary(Subset6)

full.model2_log <- lm(log(Adj_Partial_Costs_R) ~
                        Max_Effect_R
                      + I(Max_Effect_R^2)
                      + I(Adj_Penstock_Dia^2)
                      + Water_Way_Length_R
                      + I(Water_Way_Length_R^2)
                      + Construction_Year0
                      + Construction_Time_Yr
                      , data = Subset6)

#  Evaluating model by plots and tests
gvmodel <- gvlma(full.model2_log)
summary(gvmodel)
shapiro.test(full.model2_log$residuals)
vif(full.model2_log) # variance inflation factors
press(full.model2_log)
sqrt(press(full.model2_log)/length(full.model2_log$fitted.values))

par(mfrow = c(1, 2))
diagnostics_plot_partial_costs_incl_time <- autoplot(full.model2_log, which = 1:6,
ncol = 3, label.size = 3)+ theme_grey()
dev.set(5)
diagnostics_plot_partial_costs_incl_time # se plottet

predicted <- predict(full.model2_log, Subset6, se.fit=TRUE,  interval = "prediction")
predicted_ci  <-  predict(full.model2_log,  Subset6,  se.fit=TRUE,    interval  =
"confidence")
```

```
Model2_partial_costs_cv <- cv.lm(data=Subset6, form.lm=full.model2_log, m=
nrow(Subset6), plotit = F)
Model2_partial_costs_cv$pred_lower <- predicted$fit[,2]
Model2_partial_costs_cv$pred_upper <- predicted$fit[,3]
Model2_partial_costs_cv$lm_predict <- predicted$fit[,1]
Model2_partial_costs_cv$ci_lower <- predicted_ci$fit[,2]
Model2_partial_costs_cv$ci_upper <- predicted_ci$fit[,3]

Model2_partial_costs_cv_trans <-
  mutate(Model2_partial_costs_cv,
         True_costs = exp(log(Adj_Partial_Costs_R)),
         LM_Prediction = exp(lm_predict),
         CV_Prediction = exp(cvpred),
         True_costs_Rel = True_costs/True_costs*100,
         LM_Prediction_Rel = (LM_Prediction - True_costs)/True_costs*100,
         CV_Prediction_Rel = (CV_Prediction - True_costs)/True_costs*100,
         Budget_costs_rel = (Adj_Cost - True_costs)/True_costs*100,
         LM_Prediction_Abs_Perc_Diff = abs(LM_Prediction -
True_costs)/True_costs*100,
         CV_Prediction_Abs_Perc_Diff = abs(CV_Prediction -
True_costs)/True_costs*100,
         Budget_costs_Abs_Perc_Diff = abs(Adj_Cost - True_costs)/True_costs*100,
         Pred_diff = LM_Prediction - True_costs,
         Budg_diff = Adj_Cost - True_costs,
         Budget_Costs = Adj_Cost,
         PI_upper = exp(pred_upper),
         PI_lower = exp(pred_lower),
         CI_upper = exp(ci_upper),
         CI_lower = exp(ci_lower)
)

summarize(Model2_partial_costs_cv_trans, min(CV_Prediction_Abs_Perc_Diff),
max(CV_Prediction_Abs_Perc_Diff), min(CV_Prediction_Rel), max(CV_Prediction_Rel))

mean(Model2_partial_costs_cv_trans$CV_Prediction_Abs_Perc_Diff)
t.test(Model2_partial_costs_cv_trans$CV_Prediction_Abs_Perc_Diff, alternative
="two.sided")
shapiro.test(Model2_partial_costs_cv_trans$CV_Prediction_Abs_Perc_Diff)
sd(Model2_partial_costs_cv_trans$CV_Prediction_Abs_Perc_Diff)
mad(Model2_partial_costs_cv_trans$CV_Prediction_Abs_Perc_Diff)
wilcox.test(Model2_partial_costs_cv_trans$CV_Prediction_Abs_Perc_Diff, alternative
="two.sided", conf.int = TRUE)
median(Model2_partial_costs_cv_trans$CV_Prediction_Abs_Perc_Diff)
summary(Model2_partial_costs_cv_trans$CV_Prediction_Abs_Perc_Diff)

Model2_partial_costs_cv_trans <- arrange(Model2_partial_costs_cv_trans,
LM_Prediction)
Model2_partial_costs_cv_trans$Index <-
seq.int(along.with=Model2_partial_costs_cv_trans$LM_Prediction)

CV_mod2_partial_cost_scatter <- ggplot(Model2_partial_costs_cv_trans, aes(x=Index))
+
  geom_ribbon(aes(ymin=PI_lower, ymax=PI_upper, alpha="PI_lower")) +
  geom_ribbon(aes(ymin=CI_lower, ymax=CI_upper, alpha="CI_lower")) +
  scale_alpha_manual(name = "PI and CI intervals", breaks = c("PI_lower",
"CI_lower"), labels = c("Prediction interval", "Confidence interval"),
values=c(0.4,0.2))+
  geom_point(aes(y = CV_Prediction, color="CV_Prediction", shape = "CV_Prediction"),
size =2)+
  geom_point(aes(y=Budget_Costs, color="Budget_Costs", shape = "Budget_Costs"), size
=2) +
  geom_point(aes(y=True_costs, color="True_costs", shape = "True_costs"), size =2)+
  theme_grey() +
  labs(x="Rank of linear model estimate",y="Sum of partial costs [MNOK] in real
values") +
  scale_colour_discrete("Cost types", breaks = c("True_costs", "Budget_Costs",
"CV_Prediction"), labels = c("Reported costs", "Budget costs", "Cross-validated
\nprediction estimate"))+
  scale_shape_discrete("Cost types", breaks = c("True_costs", "Budget_Costs",
"CV_Prediction"), labels = c("Reported costs", "Budget costs", "Cross-validated
\nprediction estimate"))+
  geom_line(aes(y=LM_Prediction, linetype="LM_Prediction"), size = 1, color="white",
alpha = 0.8) +
  scale_linetype_manual(name="Linear model \nprediction estimate",
breaks=c("LM_Prediction"), labels =c("White line showing \nprediction estimate"),
values = c("solid")) +
coord_cartesian(ylim = c(0, 110), xlim = c(0,90)) +
  scale_y_continuous(breaks=seq(0, 120, 20)) +
  scale_x_continuous(breaks=seq(0, 100, 20))

dev.set(5) #om jeg tror den første ramma passer for figuren.
CV_mod2_partial_cost_scatter # se plottet
ggsave("LM2_Partial_Cost_cost_scatter_7-5-16.png")

Model2_partial_costs_cv_long2 <- select(Model2_partial_costs_cv_trans, Kdb_ID,
Index, Budget_costs_rel, CV_Prediction_Rel, LM_Prediction_Rel)
Model2_partial_costs_cv_long2 <- arrange(full.model6_log_cv_long2,
CV_Prediction_Rel)
Model2_partial_costs_cv_long2$Index <-
seq.int(along.with=full.model6_log_cv_long2$CV_Prediction_Rel)

Model2_partial_costs_cv_long2 <- melt(Model2_partial_costs_cv_long2,
id.vars=c("Kdb_ID", "Index"), na.rm=TRUE)

CV_mod2_standardized_cost_scatter <- ggplot(Model2_partial_costs_cv_long2,
aes(x=Index, y=value)) +
  geom_point(aes(colour=variable, shape = variable, fill = variable), size =2)+
  theme_grey() +
  scale_fill_discrete(name ="Cost types deviation", breaks = c("LM_Prediction_Rel",
"Budget_costs_rel", "CV_Prediction_Rel"), labels = c("Deviations LM cost
\nestimates", "Deviations budgeted \ncosts", "Deviations CV cost \nestimates"))+
  scale_colour_discrete(name ="Cost types deviation", breaks = c("LM_Prediction_Rel",
"Budget_costs_rel", "CV_Prediction_Rel"), labels = c("Deviations LM cost
\nestimates", "Deviations budgeted \ncosts", "Deviations CV cost \nestimates"))+
  scale_shape_discrete(name ="Cost types deviation", breaks = c("LM_Prediction_Rel",
"Budget_costs_rel", "CV_Prediction_Rel"), labels = c("Deviations LM cost
\nestimates", "Deviations budgeted \ncosts", "Deviations CV cost \nestimates"))+
  labs(x="Rank of deviations between cross-validated estimates and actual
costs",y="Percent deviation from reported sum of partial costs (real costs)") +
  geom_hline(yintercept=0)

dev.set(5) #om jeg tror den første ramma passer for figuren.
CV_mod2_standardized_cost_scatter # se plottet
ggsave("CV_standardized_cost_scatter_2-5-16.png")

summarize(Model2_partial_costs_cv_trans, RMSEP_CV_Pred = sum(sqrt((CV_Prediction-
True_costs)^2))/nrow(Model2_partial_costs_cv_trans), RMSE_LM_Pred =
sum(sqrt((LM_Prediction-True_costs)^2))/nrow(Model2_partial_costs_cv_trans),
RMSE_Budget = sum(sqrt((Budget_Costs-
True_costs)^2))/nrow(Model2_partial_costs_cv_trans), Abs_Diff_CV_Pred =
sum(CV_Prediction_Abs_Perc_Diff)/nrow(Model2_partial_costs_cv_trans),
Abs_Diff_LM_Pred =
sum(LM_Prediction_Abs_Perc_Diff)/nrow(Model2_partial_costs_cv_trans),Abs_Diff_Budge
t = sum(Budget_costs_Abs_Perc_Diff)/nrow(Model2_partial_costs_cv_trans))

sqrt((Model2_partial_costs_cv_trans$True_costs_Rel-
Model2_partial_costs_cv_trans$CV_Prediction_Rel)^2)

attach(Model2_partial_costs_cv_trans)
t.test(CV_Prediction_Abs_Perc_Diff, Budget_costs_Abs_Perc_Diff,
alternative="two.sided", paired = FALSE, var.equal = FALSE)
shapiro.test(CV_Prediction_Abs_Perc_Diff)
shapiro.test(Budget_costs_Abs_Perc_Diff)
wilcox.test(CV_Prediction_Abs_Perc_Diff, Budget_costs_Abs_Perc_Diff,
alternative="two.sided", conf.int=TRUE, paired = FALSE)
t.test(CV_Prediction_Rel, Budg_diff_Rel, alternative="two.sided", paired = FALSE,
var.equal = FALSE) #
shapiro.test(CV_Prediction_Rel)
shapiro.test(Budg_diff_Rel)
wilcox.test(CV_Prediction_Rel, Budg_diff_Rel, alternative="two.sided",
conf.int=TRUE, paired = FALSE) #
detach(Model2_partial_costs_cv_trans)

Model2_partial_costs_cv_trans$budg_high <- "TRUE"
Model2_partial_costs_cv_trans$budg_low <- "TRUE"
Model2_partial_costs_cv_trans$budg_eq <- "TRUE"
Model2_partial_costs_cv_trans$lm_high <- "TRUE"
Model2_partial_costs_cv_trans$lm_low <- "TRUE"
Model2_partial_costs_cv_trans$lm_eq <- "TRUE"
for(t in 1:nrow(Model2_partial_costs_cv_trans)) {
  ifelse(isTRUE(Model2_partial_costs_cv_trans$Budget_Costs[t] >
Model2_partial_costs_cv_trans$True_costs[t]),
Model2_partial_costs_cv_trans$budg_high[t] <- "TRUE",
Model2_partial_costs_cv_trans$budg_high[t] <-"FALSE")
```

```
  ifelse(isTRUE(Model2_partial_costs_cv_trans$Budget_Costs[t] <
Model2_partial_costs_cv_trans$True_costs[t]),
Model2_partial_costs_cv_trans$budg_low[t] <- "TRUE",
Model2_partial_costs_cv_trans$budg_low[t] <- "FALSE")
  ifelse(isTRUE(Model2_partial_costs_cv_trans$Budget_Costs[t] ==
Model2_partial_costs_cv_trans$True_costs[t]),
Model2_partial_costs_cv_trans$budg_eq[t] <- "TRUE",
Model2_partial_costs_cv_trans$budg_eq[t] <- "FALSE")
  ifelse(isTRUE(Model2_partial_costs_cv_trans$CV_Prediction[t] >
Model2_partial_costs_cv_trans$True_costs[t]),
Model2_partial_costs_cv_trans$lm_high[t] <- "TRUE",
Model2_partial_costs_cv_trans$lm_high[t] <- "FALSE")
  ifelse(isTRUE(Model2_partial_costs_cv_trans$CV_Prediction[t] <
Model2_partial_costs_cv_trans$True_costs[t]),
Model2_partial_costs_cv_trans$lm_low[t] <- "TRUE",
Model2_partial_costs_cv_trans$lm_low[t] <- "FALSE")
  ifelse(isTRUE(Model2_partial_costs_cv_trans$CV_Prediction[t] ==
Model2_partial_costs_cv_trans$True_costs[t]),
Model2_partial_costs_cv_trans$lm_eq[t] <- "TRUE",
Model2_partial_costs_cv_trans$lm_eq[t] <- "FALSE")
  ifelse(isTRUE(
    (Model2_partial_costs_cv_trans$CV_Prediction[t] >
Model2_partial_costs_cv_trans$True_costs[t]
     &Model2_partial_costs_cv_trans$Budget_Costs[t] >
Model2_partial_costs_cv_trans$True_costs[t])
    |(Model2_partial_costs_cv_trans$CV_Prediction[t] <
Model2_partial_costs_cv_trans$True_costs[t]
      &Model2_partial_costs_cv_trans$Budget_Costs[t] <
Model2_partial_costs_cv_trans$True_costs[t]))
    ,Model2_partial_costs_cv_trans$eq_sign_lm_bud_diff[t] <- "TRUE"
    , Model2_partial_costs_cv_trans$eq_sign_lm_bud_diff[t] <- "FALSE")
}
Model2_partial_costs_cv_trans$budg_high <-
as.factor(Model2_partial_costs_cv_trans$budg_high)
Model2_partial_costs_cv_trans$budg_low <-
as.factor(Model2_partial_costs_cv_trans$budg_low)
Model2_partial_costs_cv_trans$budg_eq <-
as.factor(Model2_partial_costs_cv_trans$budg_eq)
Model2_partial_costs_cv_trans$lm_high <-
as.factor(Model2_partial_costs_cv_trans$lm_high)
Model2_partial_costs_cv_trans$lm_low <-
as.factor(Model2_partial_costs_cv_trans$lm_low)
Model2_partial_costs_cv_trans$lm_eq <-
as.factor(Model2_partial_costs_cv_trans$lm_eq)
Model2_partial_costs_cv_trans$eq_sign_lm_bud_diff <-
as.factor(Model2_partial_costs_cv_trans$eq_sign_lm_bud_diff)

# 2-Way Cross Tabulation
Cross_table_costs <- CrossTable(Model2_partial_costs_cv_trans$budg_high,
Model2_partial_costs_cv_trans$lm_high, chisq=TRUE)
chisq.test(Model2_partial_costs_cv_trans$budg_high,
Model2_partial_costs_cv_trans$lm_high)
```

Norges miljø- og biovitenskapelig universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway