Course 42588 – week 9

Data analytics – linear regression 1

Today's program

- Statistical learning - prediction versus inference and the bias-variance trade-off

- Linear regression
  - Multiple linear regression
  - Example + exercise
  - Significance, hypothesis test and confidence intervals
  - Variable selection
  - GoF and adjusted R2
  - Example + exercise
  - Prediction and elasticities
  - Example + exercise

- Work on Project 3

The course plan

Week	Date	Subject/Lecture	Literature	Exercises	Teachers
1	31/1	Introduction + questions and data	AoS chap. 3	Form groups + week 1 exercise	Stefan
2	7/2	Basics on data and variables	AoS chap. 1-2 (+ OM 1)	Project 1 – start	Stefan/Guest from Genmab
3	14/2	Surveys + data types + experimental data	Paper 1 (+ OM 2-5)	Project 1 – work	Sonja / Stefan
4	21/2	Governance + causality	Paper 2 + AoS chap. 4 (+ OM 6)	Project 1 – deadline	Hjalmar / Stefan
5	28/2	More on data, e.g. real-time data, online data	Paper 3 (+ OM 7-10)	Discuss data for project 2	Guido/ Stefan
6	6/3	Visualisation	Chap. 1,5,6,7,10,23, 24,29 in Wilke + (AM 1-2)	Integrated exercises + work on project 2	Mads
7	13/3	Spatial data	Chap. 1,14 in Gimonds	Week 7 exercises + work on project 2	Mads / Guest from Niras
8	20/3	Imputation/weighting/presentation proj. 2	Paper 4	First deadline of project 2 + Week 8 exercises	Mads
9	3/4	Data analytics I	ISL ch. 3 + paper 5	Work on project 3	Stefan
10	10/4	Data analytics II	ISL ch. 6-6.1 + 7-7.4	Work on project 3	Stefan
11	17/4	Data analytics III	ISL ch. 4-4.3	Work on project 3	Stefan
12	24/4	Data analytics IV	Train 2.1-2.3 + 3.1, 3.6, 3.8-9	Work on project 3	Stefan
13	1/5	Summary and perspective	Paper 6	Project 3 – deadline	Stefan

Statistical learning

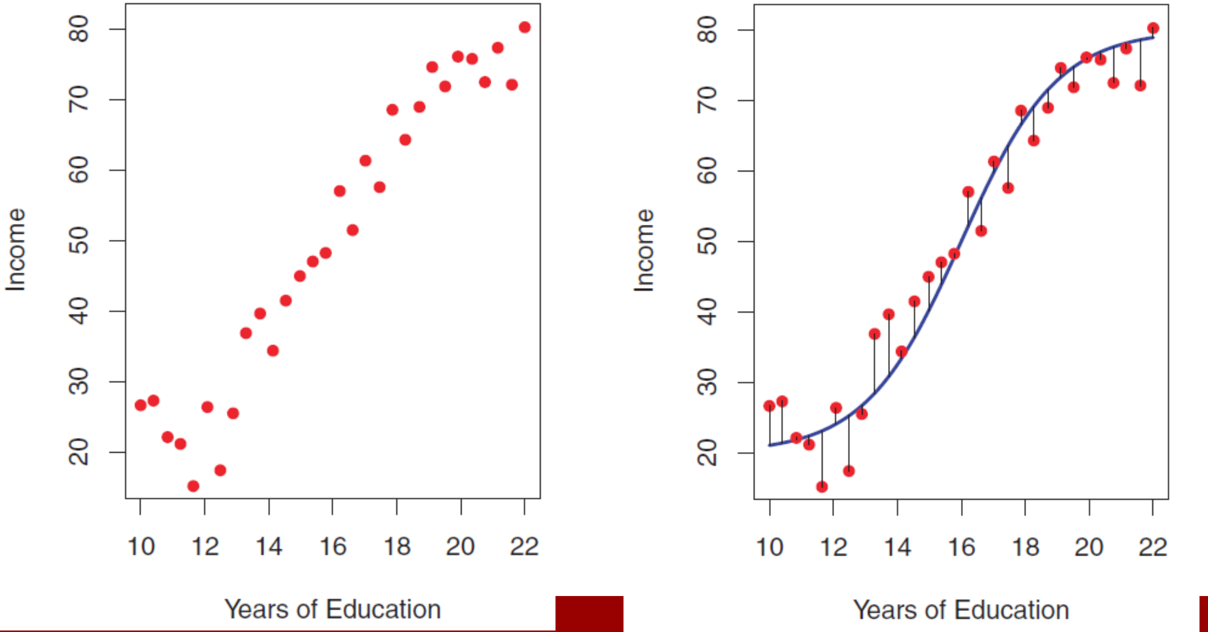
- Suppose we observe Y and p predictors $X = (X_1, \ldots, X_p)$. To analyse data, we assume

$$Y = f(X) + \epsilon$$

where ϵ is a random error term **independent** of X. Here Y is known as the dependent variable or the response variable, while X is known as predictors, explanatory variables, features

- We could be interested in
  - **Prediction**
  - **Inference**

- In either case, we need to find $f(X)$.



- Figures from Introduction to Statistical Learning by James et al.

Prediction

- The common way to predict Y (once we have estimated $f(X)$) is to use

$$\hat{Y} = \hat{f}(X)$$

- Here the accuracy (mean squared error, MSE) of $\hat{Y}$ as estimator of Y depends on two quantities
  – the reducible error, which can be lowered if we choose a better model $\hat{f}$
  – the irreducible error, which cannot be lowered by modelling since it is the uncertainty inherent in the context. It includes unmeasured variables, unmeasured variation, measurement errors.

- We have the result that

$$E\left(Y - \hat{Y}\right)^2 = \left[f(X) - \hat{f}(X)\right]^2 + Var(\epsilon)$$

Inference

- Suppose we have a model

$$Y = f(X;\beta) + \epsilon$$

- If the focus is to estimate the specific β parameters then it is inference. There can be several reasons for being interested in inference
  – To see which predictors are associated with Y
  – To find the sign and size of the β in linear models and $\frac{\partial f}{\partial X}$ in more complicated models
  – To determine whether the relationship between Y and X is approximately linear or a more complicated non-linear form is necessary
  – To find the trade-off between predictors, e.g. $\frac{\beta_1}{\beta_2}$. This could be the value of time or the value of ecological food

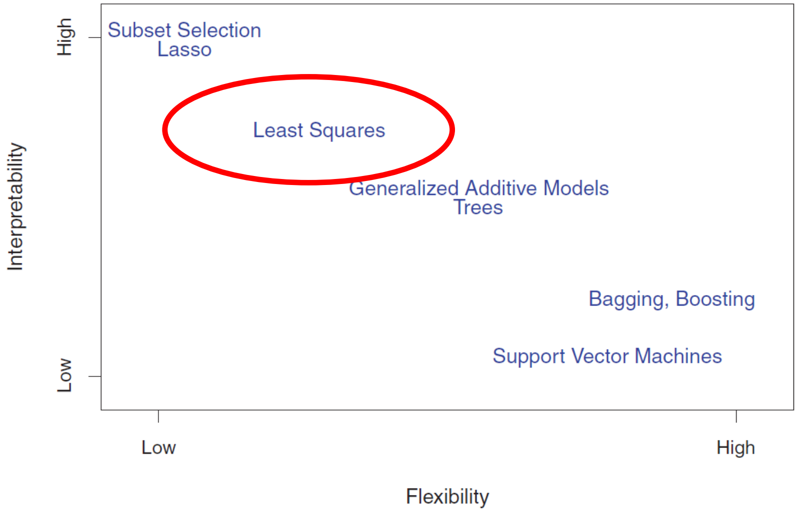
Trade-off prediction accuracy vs. interpretation

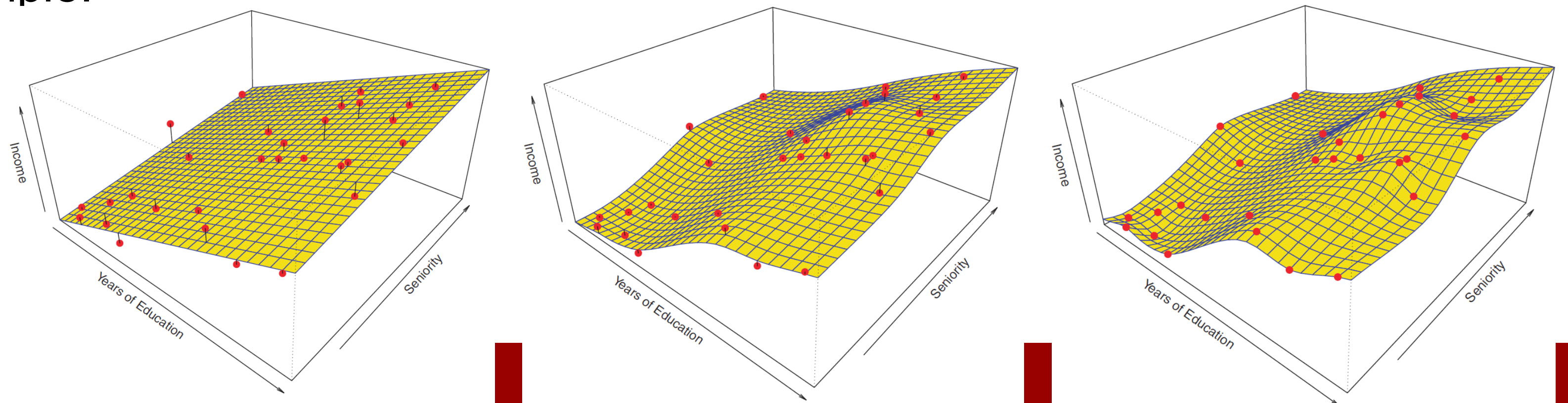
- Why would we sometimes decide to use a less accurate model?

- **Interpretation**

- Another reason for making models simpler is to avoid **overfitting**. Of course it helps using various methods, e.g. cross-validation. However, this can still overfit the model to the sample if the sample is not representative for the population of interest, e.g. a future version of the sample.

- From ISL pp. 22-24





Supervised vs. unsupervised

- We will not discuss unsupervised methods

- Supervised methods are methods were we have a dependent variable

- Supervised methods are commonly divided into
  – Regression approaches
  – Classification approaches

- Note that logistic regression is a classification approach. It has its name from the fact that we make a regression of the probabilities.

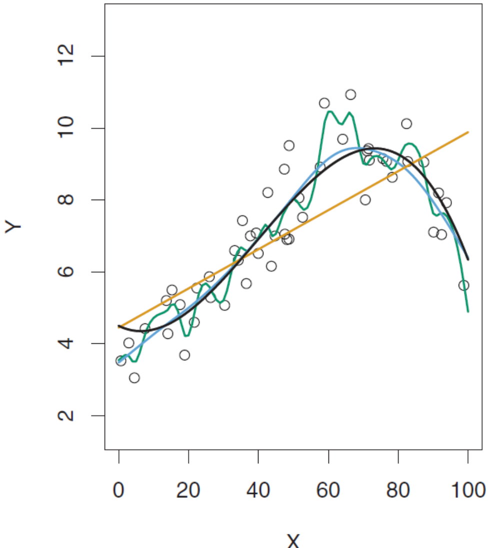
The bias-variance trade-off

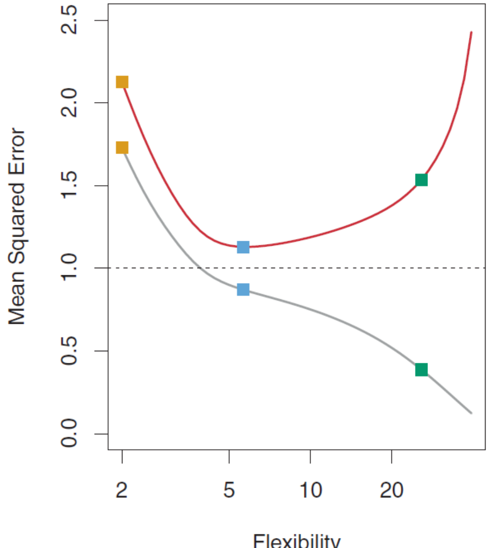
- How to measure quality of fit:
  - Measure it on the (training) sample – this may be OK if the ratio of parameters to data points is low and the model is relatively simple
  - Measure it on a test set

- A common way to measure quality of fit is using the mean square error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

- The trade-off is illustrated on p. 31



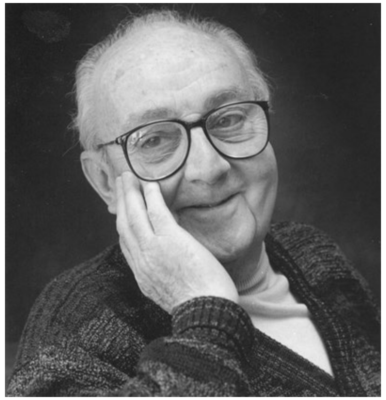


Wise words

- On modelling

“All models are wrong, but some are useful”

George Box, 1919-2013



- On uncertainty

“It is better to roughly right than precisely wrong”

John M. Keynes, 1883-1946

Linear regression - motivation

- You should all have had a introductory course to statistics where you learned about linear regression.

- The linear regression model is a standard model when we have a continuous variable and we want to analyse how this variable is affected by other variables.

- To set up a linear regression, we need to
  – Set up a frame that we think represents causal effects
  – Specify the linear relationship between the dependent variable and the explanatory variables

- ISL p. 59

” the importance of having a good understanding of linear regression before studying more complex learning methods cannot be overstated”

Question

- Find examples of continuous variables that could be dependent variables in a linear regression model
- Discuss 1 min with neighbour.

Linear regression – the model

- The linear regression model is

$$Y_i = E(Y_i|X_i) + \varepsilon_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i = X_i\beta + \varepsilon_i,$$

where Y_i is the dependent variable
X_i is a vector of explanatory/independent variables (predictors/features)

- So our linear regression is actually a model of the conditional expectation of Y_i given X_i, i.e. $E(Y_i|X_i)$.

- Some of you may only have seen a simple linear regression model

$$Y_i = E(Y_i|X_i) + \varepsilon_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

- Now we just allow for more than one explanatory variable. Everything else is more or less the same.

Linear regression – short on model building

- The linear regression model is

$$Y_i = E(Y_i|X_i) + \varepsilon_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i = X_i\beta + \varepsilon_i,$$

- So a good question is what variables to include and how to include them.
- The explanatory variables, X, are either continuous or discrete.
- For continuous variables, we just enter them as they are in data (for now).
- For discrete variables, you may enter them as they are in data (if they have a natural ordering). Or you may recode them using dummy variables.

Linear regression – recoding discrete variables

- Example:
  - Suppose you have a variable with 3 levels, e.g. 3 road types coded 0,1,2 in the variable $X_{road\ level}$.
  - Then you can choose a base level, e.g. level 0, and code the variable using two dummy variable instead, i.e. one for level 1, X_{level1}, and one for level 2, X_{level2}.
  - Suppose we try to explain traffic load then the two models are

$$Y_{load,i} = \beta_0 + \beta_1 X_{road\ level,i} + \varepsilon_i$$

$$Y_{load,i} = \beta_0 + \beta_1 X_{level1,i} + \beta_2 X_{level2,i} + \varepsilon_i$$

- What is the difference between the two models?
  - Discuss with your neighbour for 1 minute.

Example - data

- The data are annual average daily traffic (AADT) data from Minnesota. We would like to estimate the linear regression:

$$
\begin{aligned}
&AADT_i \\
&= \beta_0 + \beta_1 CNTYPOP_i + \beta_2 NUMLANES_i + \beta_3 FCTCLASS1_i + \beta_4 FCTCLASS2_i \\
&+ \beta_5 FCTCLASS3_i + \varepsilon_i
\end{aligned}
$$

Variables	Description
AADT	Annual average daily traffic on the road section
CNTYPOP	County population
NUMLANES	Number of lanes on the section
FCTCLASS1	Dummy for rural interstate
FCTCLASS2	Dummy for rural noninterstate
FCTCLASS3	Dummy for urban interstate
FCTCLASS4	Dummy for urban noninterstate

Example – OLS estimates

- The results are given below in the table

Variables	Estimate	Std. Error	T statistic	P(>|t|)
Intercept	-26234.5	4935.7	-5.3	<.0001
CNTYPOP	0.029	0.005	6.0	<.0001
NUMLANES	9953.7	1375.4	7.2	<.0001
FCTCLASS1	885.4	5830.0	0.2	0.879
FCTCLASS2	4127.6	3345.4	1.2	0.220
FCTCLASS3	35453.7	4530.7	7.8	<.0001
FCTCLASS4	0.0	-	-	-

Example - Model interpretation

- The intercept cannot be interpreted in this example. So we are not concerned with sign and significance. Always keep the intercept.
- CNTYPOP – the coefficient is significant and indicates that 1 more person in the county on average leads to 0.029 more cars on the road
- NUMLANES – the coefficient is significant and indicates that 1 more lane on average is associated with 9953.7 more cars on the road
- FCTCLASS1 and FCTCLAS2 are not significant so the data cannot tell us whether these classes can be expected to have more traffic than FCTCLASS4
- FCTCLASS3 – this class can on average be expected to have 35453.7 more cars than FCTCLASS4 and the effect is significant

Week 9 exercise – part 1

- Import the data: AADT_data

- Find the average value for the variables
  - AADT
  - CNTYPOP
  - NUMLANES
  - FCTCLASS1,…4

- Estimate the linear regression model from slides 16 and 17.

Break

Example – difference simple and multivariate

- Consider the linear regression model from before

$$AADT_i = \beta_0 + \beta_1 CNTYPOP_i + \beta_2 NUMLANES_i + \beta_3 FCTCLASS1_i + \beta_4 FCTCLASS2_i + \beta_5 FCTCLASS3_i + \varepsilon_i$$

- Also, consider a simple version with only population as explanatory variable

$$AADT_i = \beta_0{}' + \beta_1{}' CNTYPOP_i + \varepsilon_i$$

- What is the difference between an interpretation of β_1 and $\beta_1{}'$?
  – discuss for one minutes with your neighbour!

Difference continued

- In the book they have an example showing what can happen when we go from a simple to a multivariate linear regression.

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

Example on confidence intervals – part 1

- Do this exercise in groups of two!
- We will look at the height of the class based on the sample, we collected last week.
- This population has 24 individuals
- Every group draws a random sample with 5 observations:
  - Draw 5 observations from the bag
  - Note down the 5 observation on heigths
  - Put back the observations in the bag and send the bag to the next group

Variables	Estimate	Std. Error	T statistic	P(>|t|)
CNTYPOP	0.029	0.005	6.0	<.0001

T tests and p values

- In the results above, we reported t statistics and p values. What are these?

- The t statistics are the test statistics for the hypothesis that each parameter is equal to zero, i.e.

$$H_0: \beta_k = \beta_0, \qquad H_1: \beta_k \neq \beta_0$$

- The test statistic is calculated based on the formula $Z^* = \frac{\widehat{\beta_k} - \beta_0}{\hat{s}_k}$ with $\beta_0 = 0$.

- The p value is the probability that the test statistic, e.g. Z^*, could have been more extreme than the value calculated based on the data, i.e.

$$p = P(|Z^*| > Z_0)$$

Significance and confidence intervals

- Then to test significance at the α level, we calculate $Z^* = \frac{\widehat{\beta_k} - \beta_0}{\hat{s}_k} \sim_a N(0,1)$

  and compare this value to the critical values at the α level for the normal distribution, e.g. -1.96 and +1.96 at the 5% level.

NB. $|Z^*| > 1.96$ is equivalent at a p-value less than α, i.e. $\mathrm{p} < 0.05$.

- So in case the test gives a number larger than the critical values in absolute value, we can reject the null hypothesis. In this case, we say that the parameter is significant and that the related variable has a significant effect in the model.

- This corresponds to constructing $1 - \alpha$ confidence intervals around the estimates. If β_0 is not in the confidence interval, we can reject H_0. So if the hypothesis is $\beta_0 = 0$ and zero is not in the confidence interval around the estimate then we can say that the parameter is significant at the α level.

Example on confidence intervals 2 – 5 min

- Now each group has a sample of 5 observations
- Calculate the average within your sample, $\bar{X}$, in cm
- Calculate a 0.90 confidence interval where you use the known variance (σ^2=73.6 cm^2)
  – You may use that $Z_{0.95} = 1{,}64$ and
  – $[\bar{X} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\sigma^2/n}\ ;\ \bar{X} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\sigma^2/n}]$, where n is the sample size
- If you have the time feel free to calculate a 0.90 confidence interval where you estimate the variance
  – $[\bar{X} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{S^2/n}\ ;\ \bar{X} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{S^2/n}]$, where n is the sample size, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Four considerations in linear regression

- Is at least one predictor relevant?
- Do all predictors help explain Y, or is only a subset useful?
- How well does the model fit the data?
- How can we make predictions and how accurate are they?

First consideration – any predictor

- This consideration is mainly relevant when you have a large number of predictors.
- If p is large then by chance some predictors that are unrelated to Y will become significantly different from zero.
- Hence it is practice in some fields to run an F-test on the linear regression model where the null hypothesis is that all regressors are zero.
- If you need this then look it up on pp. 75-77.
- Two competing models where one is a reduced version of the other can be compared in an F test. See the book for more details.

Second consideration – variable selection

- Forward selection
  - Start with an intercept only model (M0)
  - Estimate p models with a single predictor -> keep the best model (M1)
  - Estimate p-1 models, i.e. M1 + the each of the remaining p-1 predictors -> keep the best model (M2), etc.

- Backward selection
  - Estimate a full model with all predictors
  - Remove the least significant and reestimate

- Mixed selection
  - Consider removing insignificant variables that have previously been found significant

Third consideration - Goodness-of-fit measures

- We need a measure for how well our model explains data. In general, such a measure is known as a goodness-of-fit (GoF) measure.

- We can define three different sums based on the regression

- Residual sum of squares: $\mathrm{RSS} = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$
- Total sum of squares: $\mathrm{TSS} = \sum_{i=1}^{n}(Y_i - \overline{Y_i})^2$

- These sums are used to calculate the most common goodness-of-fit measure:

$$\mathrm{R}^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Comments on R^2

- It is a measure between 0 and 1.
- It should only be compared between models explaining the same phenomenon.
- A model of 0.2 can be good in one context while a model of 0.9 can be less good in another context.
- Since R^2 will always grow when more parameters are added there is an adjusted version taking the number of parameters into account:

$$\mathrm{R}^2{}_{adjusted} = 1 - \frac{(n-1)RSS}{(n-p)TSS},$$ where n is no. obs., and p is no. parameters

Comparison of two linear regression models

- Comparison of models on the same data.

- Consider whether signs, significance and sizes make sense for the parameters in either model. Here you could also look at elasticities.

- If both models make sense, look at adjusted R^2. In general, for two models that make equal sense we statistically prefer the one with highest adjusted R^2.

- This is only useful if estimated on the same data.

Linear regression – specification briefly

- Once you have your variables, you should decide on which variables to use in your model.

- When evaluating the need for at specific variable in a model, we look at two things
  - Signs and sizes of parameters (common sense + field knowledge)
  - Whether the parameter is significant (t test or p-value)

- NB. Some would only call this statistically significant. They emphasize that real significance is a different matter.

- You can then compare models using F-tests and/or (adjusted) R-squared.

- Alternatively if you know AIC or BIC and prefer these, you are welcome to use them.

Question – discuss 1 min with neighbour

- Given the model summary below, what can you conclude?

Models	DoF	Adj. R^2	Comment
Model 0	6	0.79	Base model
Model 1	8	0.87	Model 0 + two parameters. One parameter with wrong sign
Model 2	7	0.85	Model 0 w. correct sign on new parameter

A. $Model\ 2\ is\ prefered$

B. $Model\ 1\ is\ prefered$

C. $Model\ 0\ is\ prefered$

D. $Neither\ of\ the\ models\ are\ prefered$

- Answer:

Week 9 exercise – part 2

- Estimate your own extension of the linear regression model from slides 16 and 17.
- Argue why your new more is better or argue why you prefer the previous model based on the extensions you have tested.

Break

Fourth consideration - predictions

- Three sorts for uncertainty in predictions

- $\hat{Y}$ (regression plane) is an approximation of $f(X)$ (true population regression plane). This uncertainty can be quantified by a confidence interval around $\hat{Y}$. This is a reducible error, i.e. the approximation can be improved

- The linear model is an approximation. So we have model bias. This is also part of the reducible error, i.e. we can choose a more flexible non-linear model.

- Even with the correct model there will be uncertainty due to the irreducible error. This can be captured by a prediction interval.

Prediction example

- In the book they have the Y=sales, and X=(TV, radio) example. They estimate a linear regression and predict sales for TV=$100,000 and radio=$20,000.

- This gives the prediction $\hat{Y} = 11{,}256$

- The confidence interval capturing the uncertainty in average sales across many cities is found to be

$$[10{,}985;\ 11{,}528]$$

- The prediction interval capturing the uncertainty in sales for a specific city is found to be

$$[7{,}930;\ 14{,}580]$$

- The prediction interval is always larger than the confidence interval.

Linear regression - forecasting

- In the linear regression model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i,$$

we can forecast using the estimates. After estimation, the model implies that

$$\hat{Y} = X_i \hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_p X_{pi}$$

- So if there is a change in an explanatory variable, e.g. X_{1i}, then we can calculate the expected change in Y_i by entering the new value of X_{1i} into the model.

Linear regression – forecasting example

- We have

$$E(Y_i|X_i) = \hat{Y} = X_i\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_p X_{pi}$$

- So for our example with estimates in the table. We calculate

$$\hat{Y} = E(AADT_i) = -26234.5 + 0.029 * CNTYPOP_i + \cdots + 35453.7 * FCTCLASS_i$$

- Using the formula, we can forecast traffic for a specific road, e.g.
  - CNTYPOP = 263,428
  - NUMLANES = 3.1
  - FCTCLASS3 = 1

- This gives $\hat{Y}$ = 47,707 cars on an average day

Variables	Estimate
Intercept	-26234.5
CNTYPOP	0.029
NUMLANES	9953.7
FCTCLASS1	885.4
FCTCLASS2	4127.6
FCTCLASS3	35453.7
FCTCLASS4	0.0

Question – discuss 1 min with neighbour

- What is the marginal effect, $\frac{dy}{dx}$, of x on y in the model,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \text{ ?}$$

A. $\beta_0 + \beta_1 x + \beta_2 x^2$
B. $\beta_0 + \beta_1 + 2 * \beta_2 x$
C. $\beta_1 + \beta_2 x$
D. $\beta_1 + 2 * \beta_2 x$

- Answer:

Linear regression - elasticities

- In the linear regression model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i,$$

  we can find the marginal effect of X_k on Y as:

$$\frac{\partial Y}{\partial X_k} = \beta_k$$

- This effect depends on the units of X_1 on Y. Therefore we often look at elasticities instead of marginal effects. These are given as

$$E_{Y,X_k} = \frac{\partial Y}{\partial X_k}\frac{X_k}{Y} = \beta_k \frac{X_k}{Y}$$

- The elasticity tells us the percentage change in Y caused by a percentage change in X_k

Linear regression – elasticities example

- In the linear regression model,

$$AADT_i = \beta_0 + \beta_1 CNTYPOP_i + \beta_2 NUMLANES_i + \beta_3 FCTCLASS1_i$$
$$+\beta_4 FCTCLASS2_i + \beta_5 FCTCLASS3_i + \varepsilon_i$$

we estimate the elasticity using $E_{Y,X_k} = \frac{\partial Y}{\partial X_k}\frac{X_k}{\hat{Y}} = \beta_k \frac{X_k}{\hat{Y}}$

where

$$\hat{Y} = -26234.5 + 0.029 * CNTYPOP_i + \cdots + 35453.7 * FCTCLASS_i$$

– For the observation: CNTYPOP = 263,428; NUMLANES = 3.1; FCTCLASS3 = 1, we have $\hat{Y}$ = 47,707 so the elasticity of $AADT$ wrt. $CNTYPOP$ is

$$E_{Y,X_k} = \beta_k \frac{X_k}{\hat{Y}} = 0.029 * \frac{263{,}428}{47{,}707} = 0.16$$

- So a 10 % increase in POP is expected to lead to a 1.6% increase in a road with these characteristics.

Elasticities – for discrete variables

- In the linear regression model with continuous X_k the elasticity is given by

$$E_{Y,X_k} = \frac{\partial Y}{\partial X_k}\frac{X_k}{Y} = \beta_k \frac{X_k}{Y}$$

- For discrete variables, we cannot differentiate, so instead we use a pseudo-elasticity based on

$$E_{Y,X_k} = \frac{\Delta Y}{\Delta X_k}\frac{X_k}{Y} = \beta_k \frac{X_k}{Y}$$

where $\Delta X_k = (X_k + 1) - X_k = 1$. Note that this only makes sense for binary discrete variables. For discrete variables with more values and more β's you have to do something else.

Question – discuss 1 min with neighbour

- If Y$= \beta_0 + \beta_1 \ln(X_1) + \varepsilon$, where both variables are continuous, what is then the elasticity of Y wrt. X_1?

  A. $E_{y,x} = \beta_1 * \frac{x}{y}$

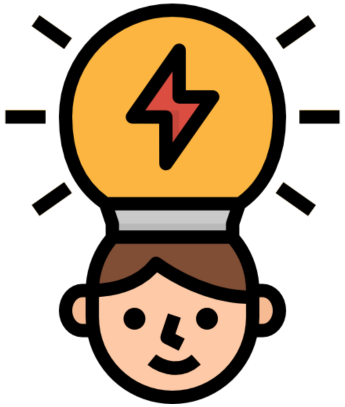
  B. $E_{y,x} = \beta_1 * \frac{1}{y}$

  C. $E_{y,x} = \beta_1 * x$

  D. $E_{y,x} = \beta_1$

- Answer:

Project 3



- The project should focuses on the AC stages in PPDAC
  - – Analyses
  - – Communication

Exam

- The oral exam will be the 28th (max 3 groups) and the 29th.
- The exam is group wise with 10-12 minutes of group presentation followed by 10-12 minutes individual discussion for each group member.
- Feel free to send me a mail if your group has a preference for either the 28th or 29th. Based on the wishes I have received by April 17th, I will make an exam plan.
- The presentation is related to project 3. The individual part is focused on project 3 but there can also be questions related to other parts of the curriculum. A curriculum list will be uploaded late April.

Feedback

- Final questions

    1. What was the most interesting you learned during the lecture?

    2. What is your most important unanswered question based on the lecture?

- Group 5 (Anders, Simen, Jonas, Alexander, Felix) should send their feedback to Stefan. Everyone else are very welcome to give feedback as well!

For next time

- Read for this week
  – Introduction to statistical learning chap 3.2-3.3.2 (2.1-2.2.2 + 3.1 is repetition)

- To prepare for lecture 10, you should read
  – Introduction to statistical learning chap 3.3.3-3.4 + 6-6.1 + 7-7.4

- Feel free to continue working on the week 9 exercise.

- Work on Project 3 (deadlines 7/5).