Course 42588 – week 10

Data analytics – linear regression 2

Today's program

- Linear regression
  - Recap on interpretation
  - Interactions and non-linearity
  - Potential problems
    - Non-linearity and non-constant variance of the error term
    - Outliers and high leverage points
    - Collinearity
  - Regression to the mean
  - Subset selection
  - Polynomial regression
  - Step functions
  - Regression splines

- 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

When to use linear regression?

- Theoretically, we need Y to be continuous and unbounded.
- Ideally, Y is normally distributed but for large samples it is not crucial.
- It may work for variables that are not themselves continuous but can be approximated by a continuous variable, or bounded variables if the observations are not close to the bounds.
- This highlights that you should always plot your Y variable to check its distribution. You should be concerned if
  - there are point masses, e.g. at 0
  - many observations are close to the bounds

Linear regression - specification

- The simplest specification is the model that is linear in all variables of interest, i.e. Y_i and each X_i are equal to the variables of interest and

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

where Y_i is the dependent variable

X_i is a vector of explanatory/independent variables

Linear regression - interpretation

- When we judge a linear regression model, we look at
  - Sign on parameters
  - Size of parameters (only comparable for same unit, otherwise use elasticities)
  - Significance (often at the 5% level but other levels are valid as well)

- Based on significance at the α level, we use the decision rule
  - If $|z^*| \leq z_{crit}(1 - \frac{\alpha}{2})$, (maybe) conclude H_0, i.e. the variable may not be important
  - If $|z^*| > z_{crit}(1 - \frac{\alpha}{2})$, conclude H_1, i.e. the variable is important

Model building

		Variables	
		Policy	**Other**
Correct sign	Significant	*Include*	*Include*
	Not significant	*Include*	*May reject*
Wrong sign	Significant	***Big Problem***	*Reject*
	Not significant	*Problem*	*Reject*

Question – discuss 1 min with neighbour

- In a t test we use the critical value at the 10%, 5% or 1% level of significance. For a normal distribution what are the critical values in a two-sided test for these three levels, respectively?

A. $The\ values\ are\ 1.5, 2.0,\ and\ 2.5$

B. $The\ values\ are\ 1.14\ 1.68,\ and\ 1.96$

C. $The\ values\ are\ 1.68, 1.96,\ and\ 2.58$

D. $The\ values\ are\ 1.1, 1.5,\ and\ 2.0$

- Answer:

Question – t test!

- Assume, that we have estimated an income paramter in a linear regression at $\hat{\beta} = 0.2$ with standard error $s = 0.09$. We would like to test if the parameter could be equal to $\beta_0 = 0$ using a t test, i.e. $z = \frac{\hat{\beta} - \beta_0}{s}$. What expression is true?

A. $\hat{\beta}$ is significantly different from 0 at the 5% level

B. $\hat{\beta}$ is significantly different from 0 at the 10% level but not at the 5% level.

C. $\hat{\beta}$ is not significantly different from 0 at the 10% level.

D. $\hat{\beta}$ is significantly different from 0.2 at the 10% level.

Answer:

Linear regression – interactions

- A linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

can easily be extended to include interactions of the x variables. Consider the case of annual income. We might hypothesise that either gender or experience has an interaction with age.

- Continuous interaction

$$Y_{income} = \beta_0 + \beta_1 X_{age} + \beta_2 X_{experience} + \beta_3 X_{age} * X_{experience} + \varepsilon_i$$

- Dummy interaction

$$Y_{income} = \beta_0 + \beta_1 X_{female} + \beta_2 X_{age} + \beta_3 X_{female} * X_{age} + \varepsilon_i$$

- What are the marginal effect of each variable in the two examples?

An example from the book

- In the example with **sales ~ TV, radio** we get the following (p. 89) when we add an interaction

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

- We see that the interaction as well as the two main effects are all three significant.

- Hierarchical principle: We keep main effects even when they are insignificant if we have added interactions.

Linear regression - transformations

- Two examples are

$$Y_{income} = \beta_0 + \beta_1 X_{age} + \beta_2 \left(X_{age}\right)^2 + \beta_3 X_{experience} + \varepsilon_i$$

$$LN(Y_{commute}) = \beta_0 + \beta_1 X_{gender} + \beta_2 LN(X_{income}) + \varepsilon_i$$

- Hierarchical principle: We keep main effects even when they are insignificant if we have added higher order effects like X squared. (not necessarily for ln())

When to use ln(x) instead of x

- Ln(x) is used when large values of x has a relatively smaller effect
  - e.g. an income change of 100,000 kr from 900,000 kr to 1,000,000 kr has a smaller effect than a change from 200,000 kr to 300,000 kr. if income is modelled using ln(), i.e. income is “logged”.

- The marginal effect in

$$y = \beta_0 + \beta_1 \ln(x_1) + \mathrm{u}$$

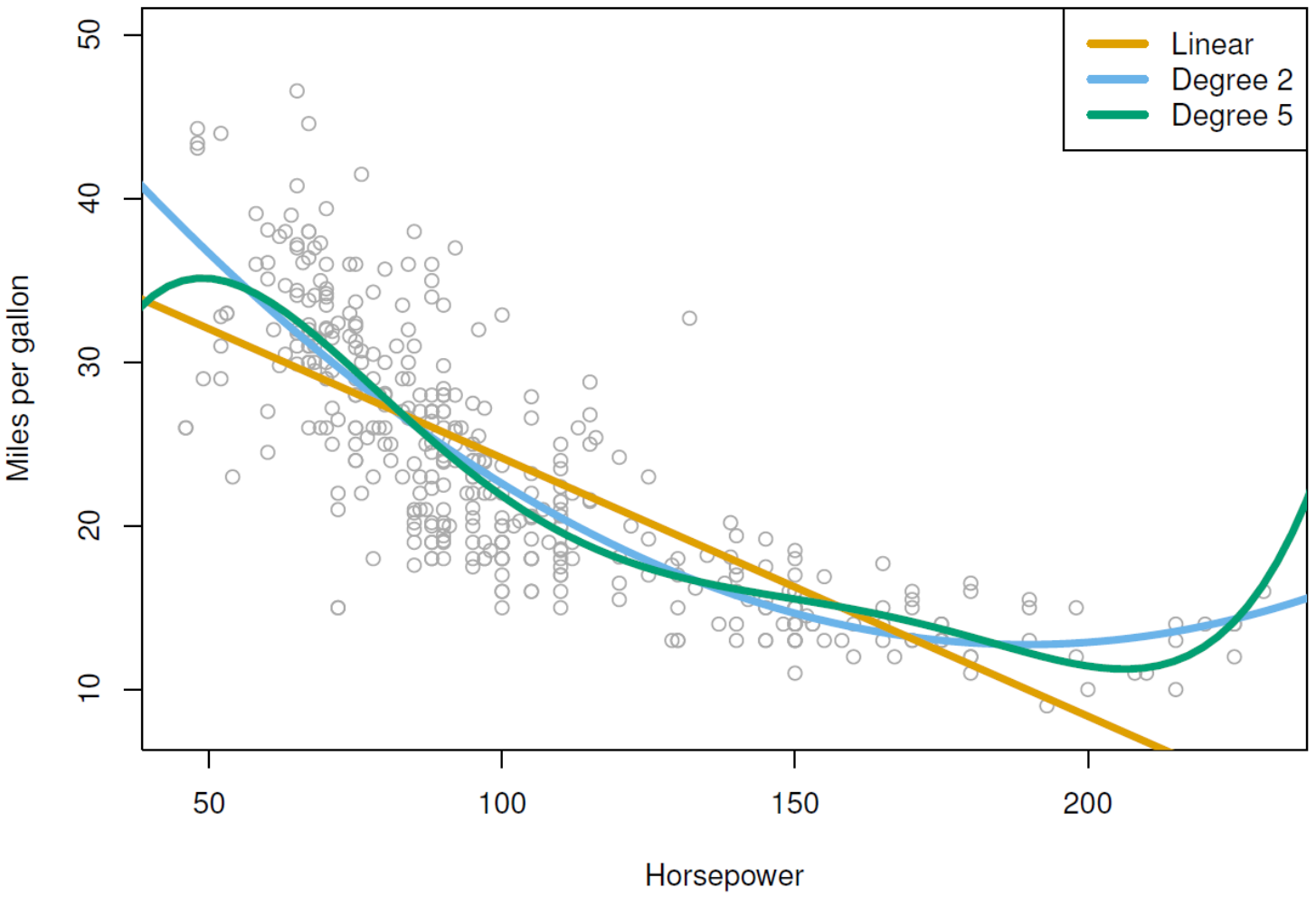
  is different from the one in the linear model. In a model with $\ln(x_1)$, the marginal effect decreases for larger values of x_1 but it never changes sign.

An example of a

- We have the following exa
  the Auto data set using ho

	Coeffic
Intercept	56.9
horsepower	−0.4
horsepower^2	0.0

- Why only use horsepower



Break

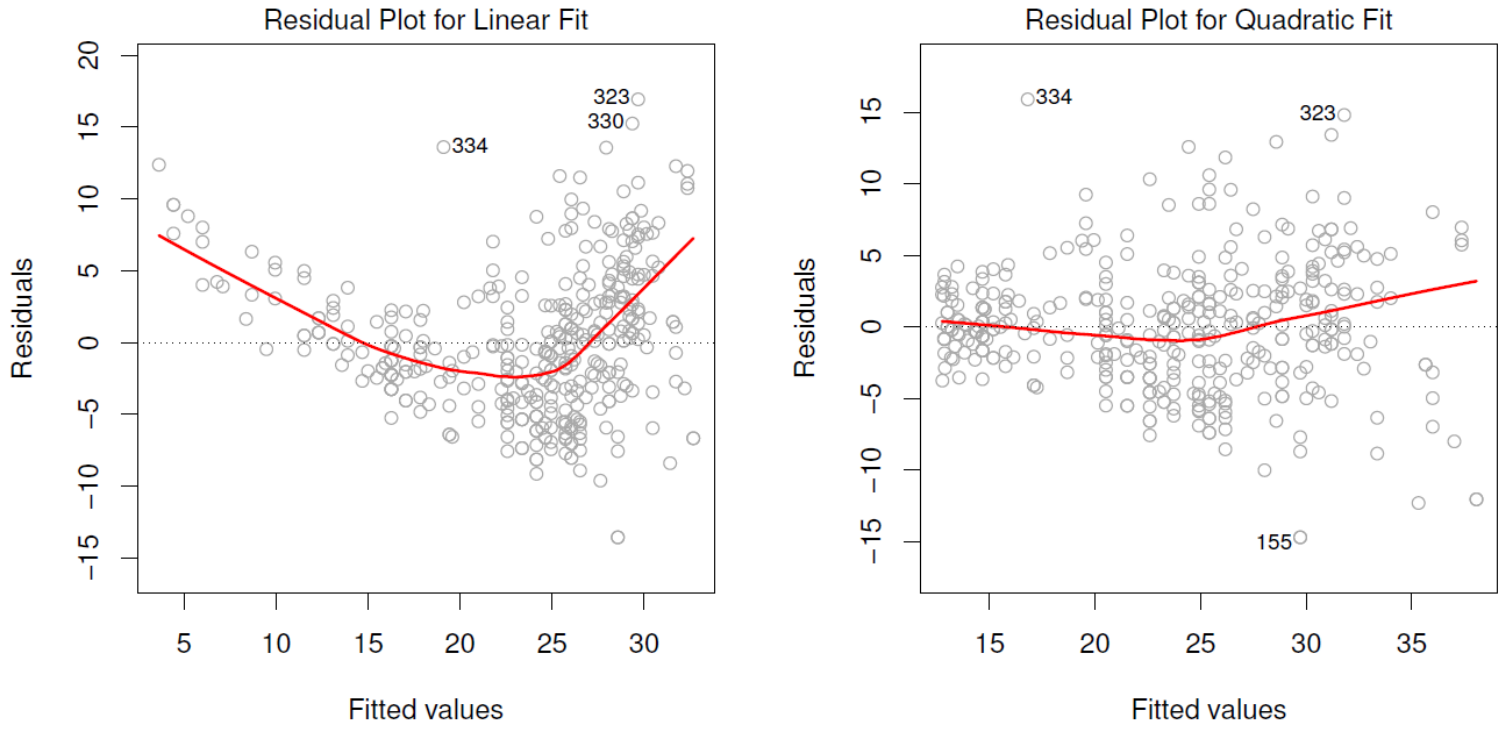
Potential problems

- Non-linearity and non-constant variance of the error term
- Correlation of error terms
- Outliers and high leverage points
- Collinearity

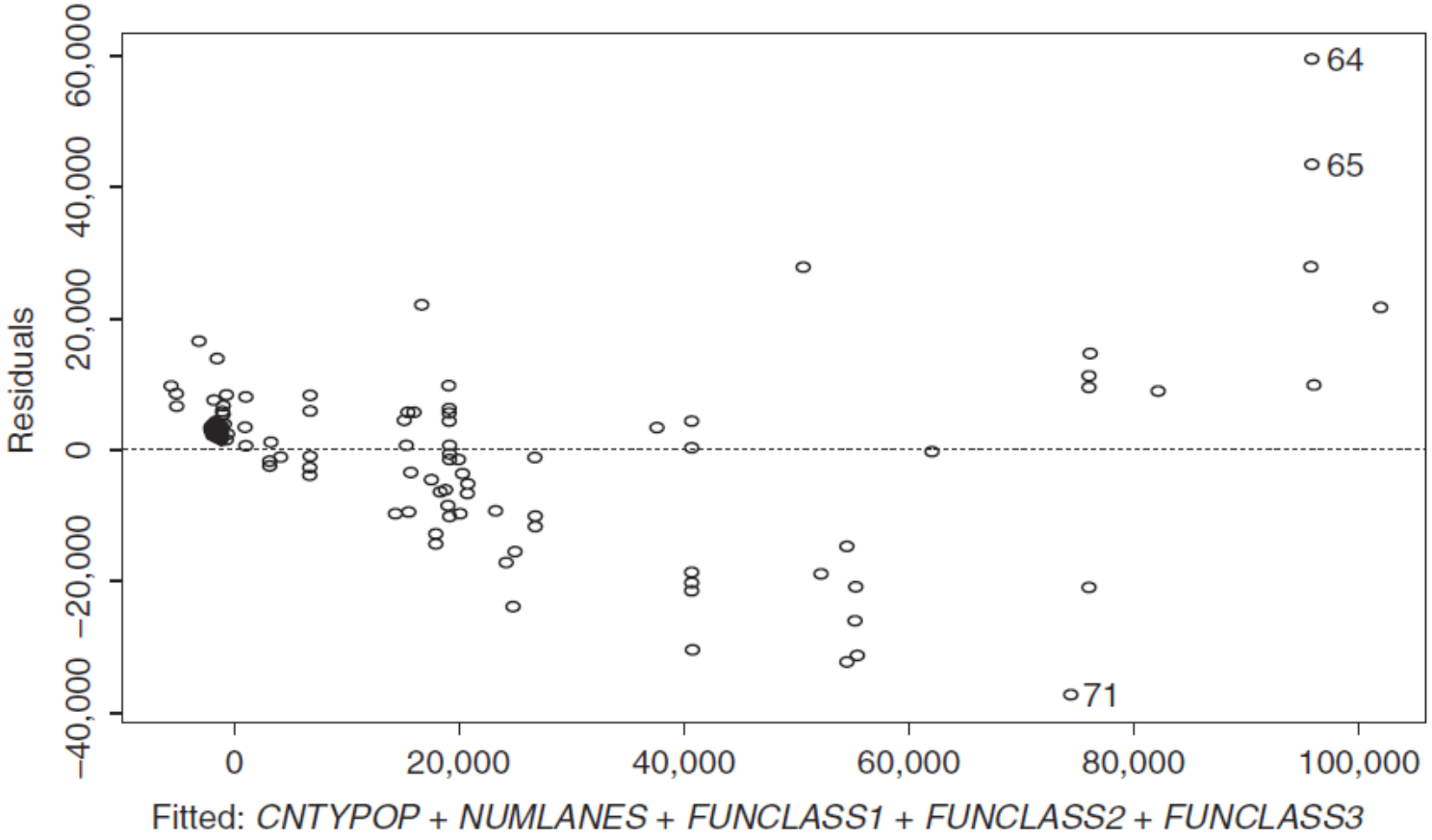
Potential problems

- Linearity is checked by plotting residuals vs. predicted or residuals vs. each explanatory variable:

- Example on Auto data from the book p. 93:



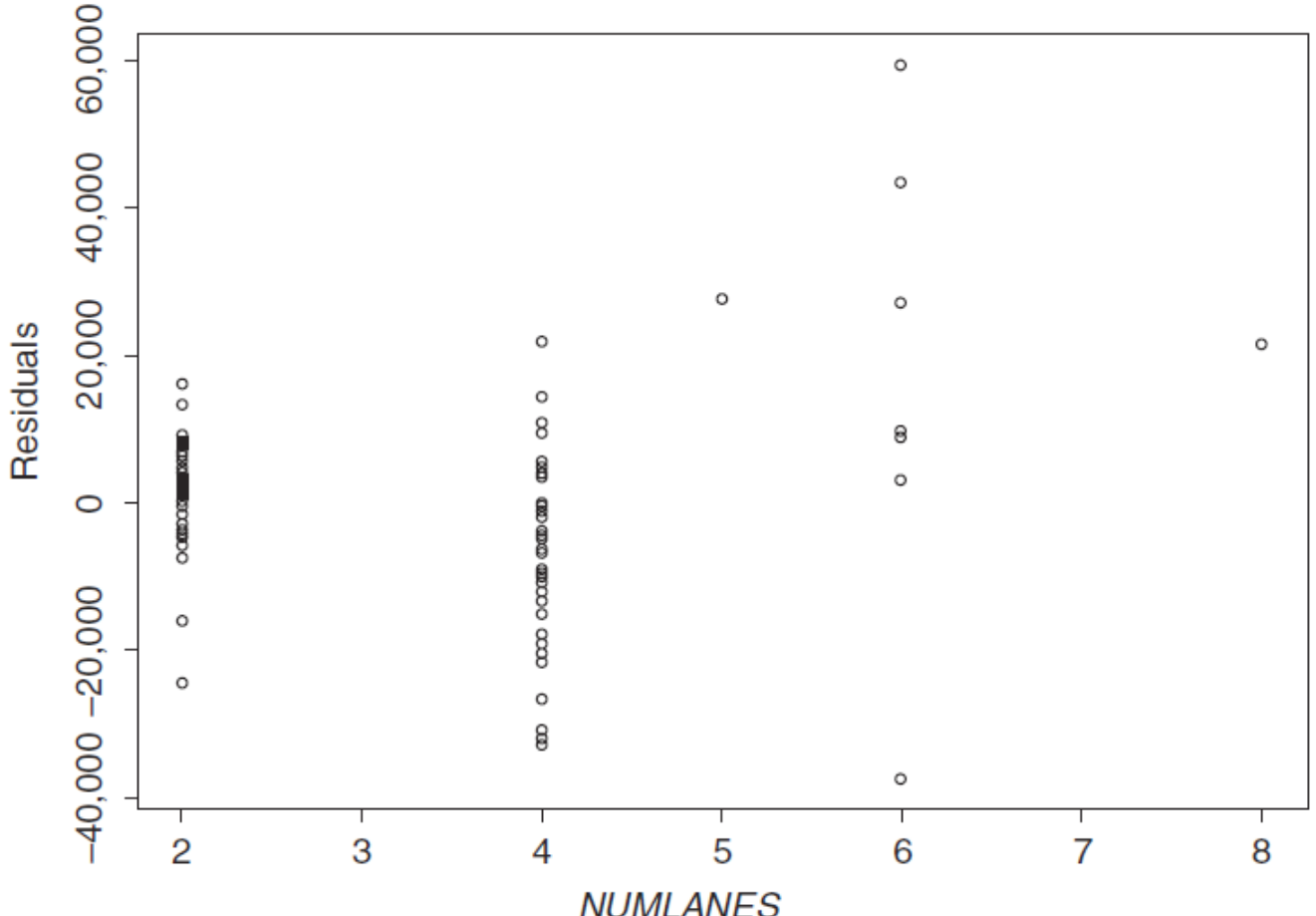
Potential problems - linearity



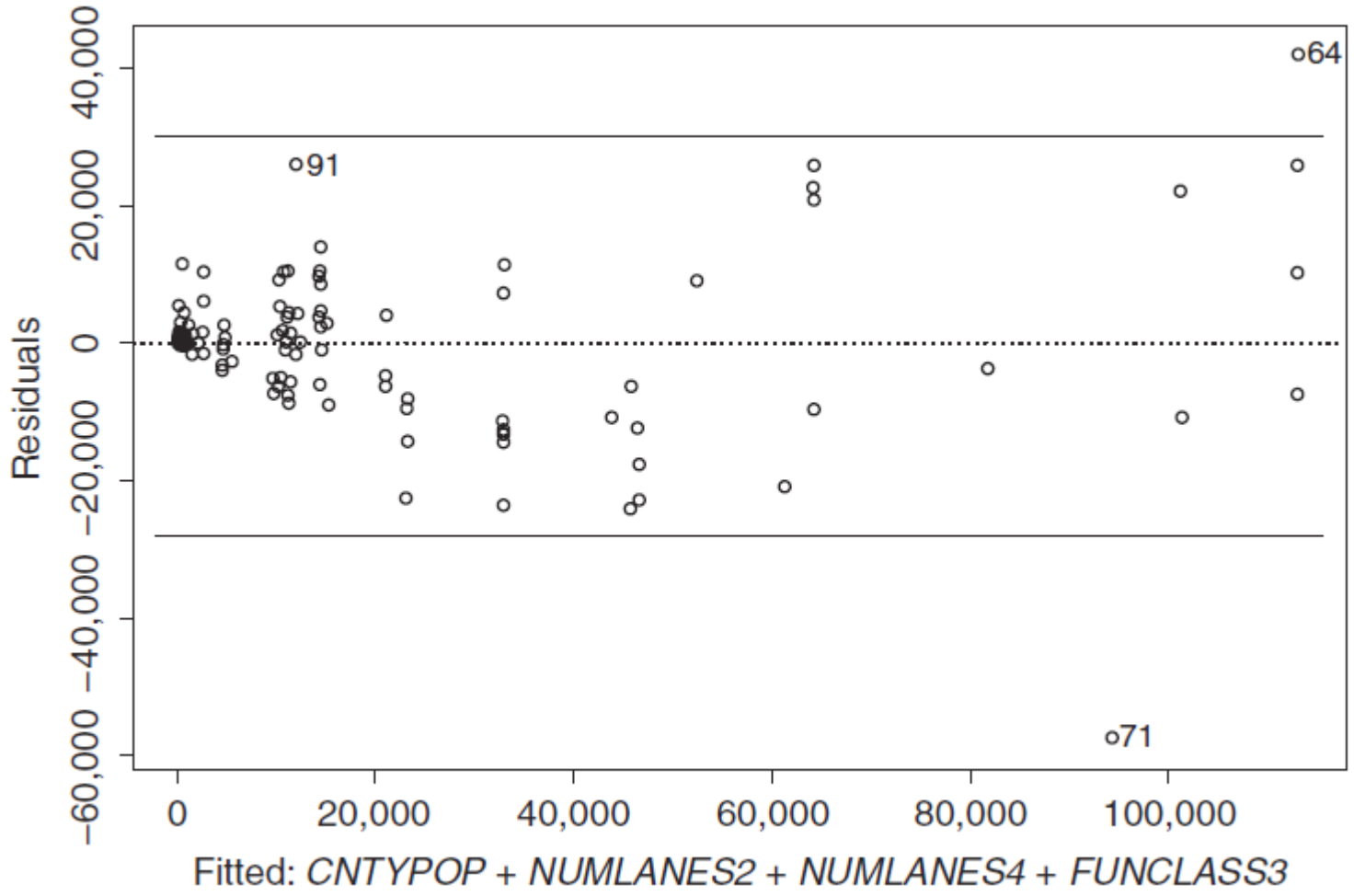
- What do you think about this plot? Discuss with your neighbour. What can you do if you think that there is a pattern?

Potential problems - linearity

- Here residuals vs. the variable NUMLANES shows problems. The residuals should be equally scattered around zero for all values of NUMLANES.



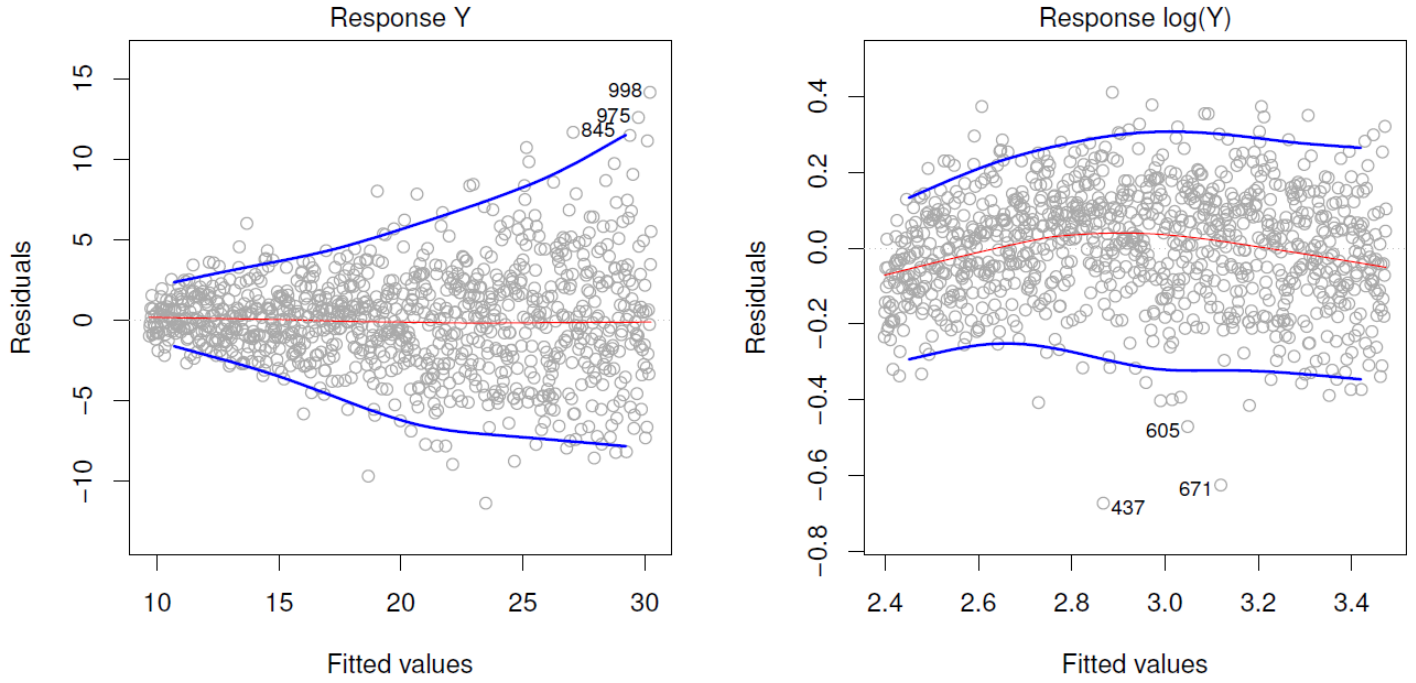
Potential problems – non-constant variance



Potential problems – non-constant variance

- Non-constant variance (also known as heteroscedasticity) can sometimes be solved by transformation of the dependent variable or one of the explanatory variables.

- Example from the book p. 96:



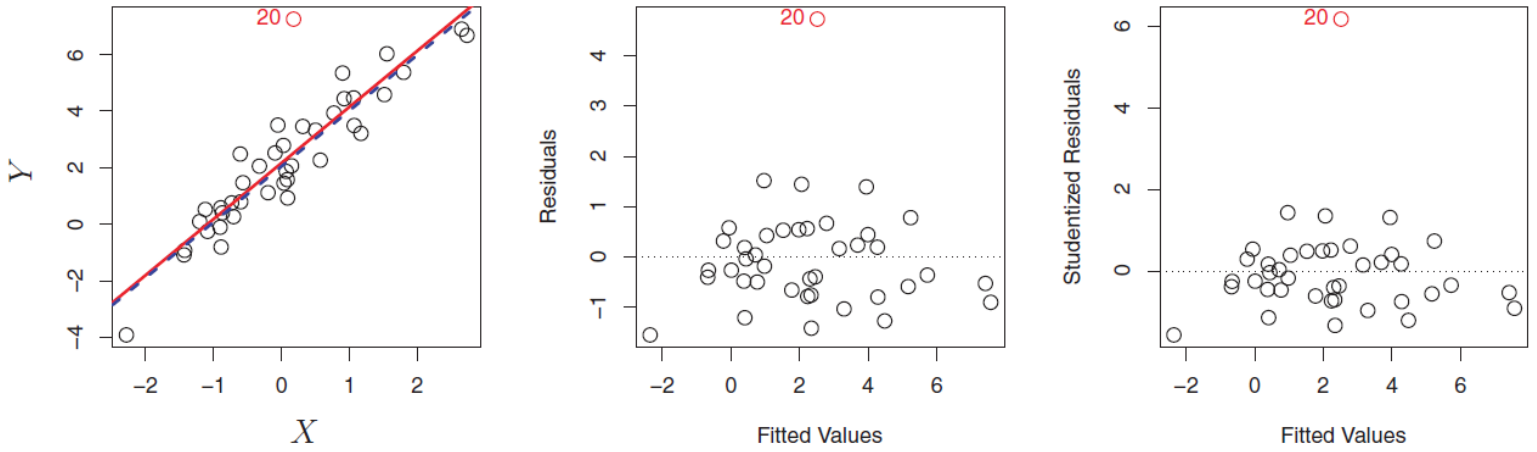
Potential problems – correlation of error terms

- Problems when observations are dependent across individuals, time, or space.
- Time series data is the most common example, e.g. stock prices. See example in book p. 95. Take a specific course if you want to work with these.
- Methods to deal with autocorrelation is beyond this course. But it is still good practise to think about the danger of correlation among error terms when you set up a model.

Outliers and high-leverage points

- The book defines outliers and high-leverage points as two different characteristics.

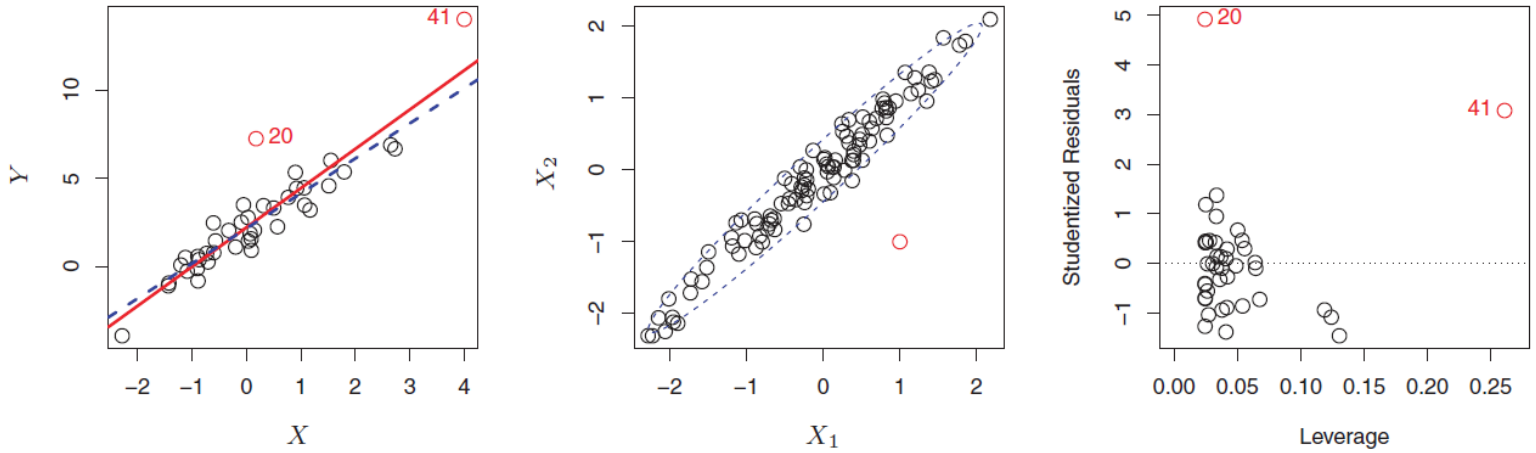
- Outliers are points that are special in the outcome dimension, i.e. based on the Y variable. As shown on p. 97 of the book:



- Outliers do not necessarily affect the regression line but could affect confidence intervals, etc. where the residual standard error (RSE) is used for calculation

Outliers and high-leverage points

- High-leverage points are points that are special in the input dimension, i.e. based on the X variables. As shown on p. 98 of the book:



- High-leverage points have potential to affect the regression line and in addition also confidence intervals, etc.

Regression Outliers

- A matrix used to find outliers is the hat matrix

$$H = X(X^T X)^{-1} X^T$$

so $\hat{Y} = HY$ and the residuals are given by $e = (I - H)Y$

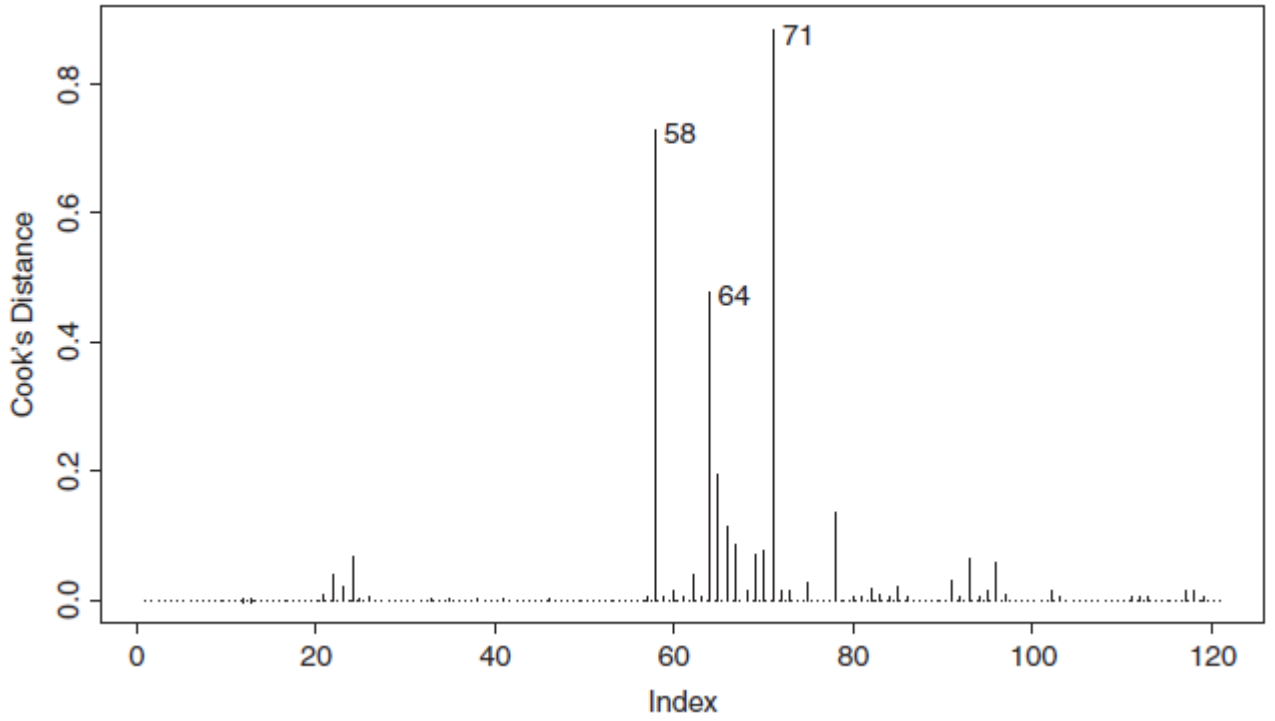
- Cook's distance measure is given by

$$D_i = \frac{{e_i}^2}{p(MSE)} \frac{h_{ii}}{(1 - h_{ii})^2}$$

where h_{ii} are the diagonal elements of H and MSE is the estimator of error term variance.

Regression Outliers – traffic example week 9

- Cook's distance measure for our data set shows three outliers. Based on individual analysis 64 and 71 are removed from the data set, while 58 is kept.



Multiple regression - multicollinearity

- Multicollinearity, i.e. strong correlation among explanatory variables. This can give problems in the estimation making it impossible to estimate separate effects for two or more variables. Examples are
  - Diesel and fuel efficiency for cars
  - Weight, horse power and motor size for cars
  - GDP (BNP), population and jobs in a zone
  - Travel time and cost between zones

- NB. Multicollinearity can lead to wrong signs on some of the coefficients, e.g. a model where you find that the number of travellers increase if travel cost increases! Then you have to change something in the model.

- Therefore always make a correlation matrix for your explanatory variables, so that you can judge if multicollinearity is a problem. Do not include non-binary nominally scaled variables in a correlcation matrix!

Multiple regression - multicollinearity

- Another indicator for multicollinearity is the variance inflation factor (VIF).
- This is the ratio of the variance of $\hat{\beta}_j$ when fitted in the full model over the variance of $\hat{\beta}_j$ when fitted in simple model with only x_j.
- A rule of thumb is that VIFs above 5 should raise some concern of multicollinearity
- What to do?
  - Remove one of the problematic variables
  - Make linear combinations of the problematic variables. The weights need to be exogenously determined.

Potential problems as OLS assumptions

- There are six assumptions that are useful for the linear regression model:
  1. Functional form
  $$Y_i = X_i\beta + \varepsilon_i$$
  2. Zero mean of disturbances
  $$E(\varepsilon_i) = 0$$
  3. Homoscedasticity of disturbances
  $$VAR(\varepsilon_i) = \sigma^2$$
  4. Nonautocorrelation of disturbances
  $$COV(\varepsilon_i, \varepsilon_j) = 0$$
  5. Uncorrelatedness of regression and disturbances
  $$COV(X_i, \varepsilon_j) = 0$$
  6. Normality of disturbances
  $$\varepsilon_i \sim N(0, \sigma^2)$$
- The conditions can be reformulated when X is stochastic. Then we think of the modelling as conditional on X, e.g. the condition $E(\varepsilon_i|X_i) = 0$ implies conditions 2 and 5.

Potential problems – checklist

- Consider whether you linear regression makes sense
- Consider whether there could be – correlation among error terms – what is you observation unit?

- Check linearity and non-constant variances using residual plots. You decide whether you use residuals or standardised residuals
- Check if any observations are outliers or high-leverage points from residual plots or leverage plots
  – Only remove observations if you have solid reasons to do so.

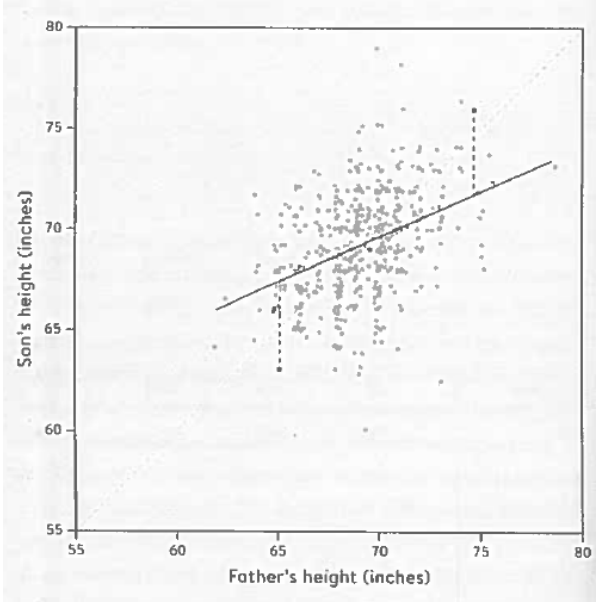
- Check if multicollinearity is an issue using either a correlation matrix or VIF scores

- Only use QQ-plots if it important for you application that error terms are normally distributed.

Break

Another potential problem - regression to the mean

- Data from 1886 (Sir Francis Galton) on the height of sons and fathers with regression.



- The marginal effect of fathers height on sons height is 0.45. What does this tell you? Discuss with your neighbour for a minute.

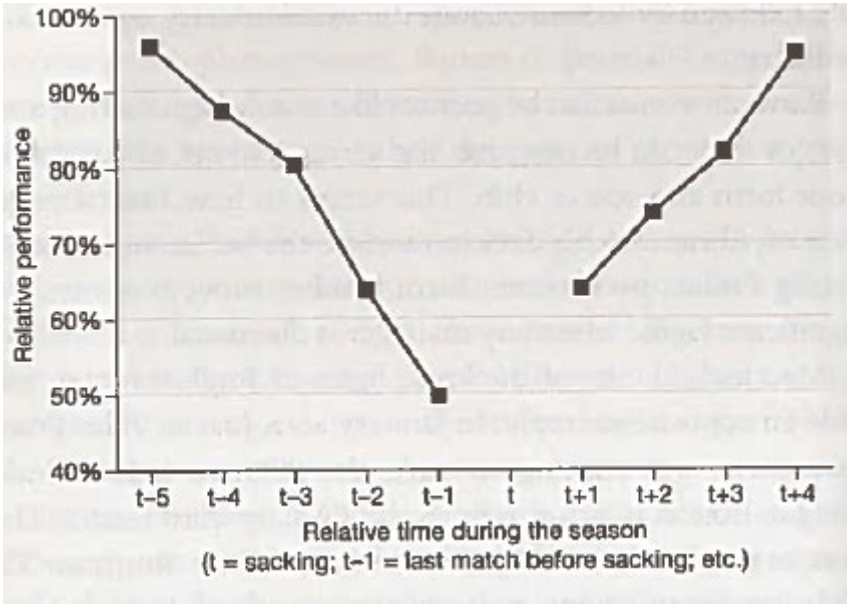
Regression to the mean – a second example

- Comparing the PISA measurements in 2003 and 2012, the math scores across countries in the two years had a correlation of -0.60.

- This of course led to dramatic conclusions about the level of math teaching in various countries.

- However, from a statistical point of view if the ratings had been purely random then we would expect at correlation of -0.71. This shows that a large part of the negative correlation can be explain as regression to the mean while only a minor part is probably due to changes in math teaching in the various countries.

Regression to the mean – a third example



- The figure shows the effect of sacking a football coach at time t (Eresdivisie, 1986-2004) for teams who lost four games in a row.

Regression to the mean

- Regression to the mean is problematic when we try to evaluate interventions using linear regression.

- If we imagine that we measure accidents on 100 Danish roads in 2010, y_{2010}. Then we choose the 10 most dangerous roads and make safety measures. In 2012, we measure again the number of accidents on the 100 roads, y_{2012}.

- The Road Directorate asks us to investigate if there has been an effect of the safety measures. So we make a regression

$$y_{2012} - y_{2010} = \beta' x + \gamma * 1(safety\ measure) + u$$

- We may get a significant effect even if the measures have no effect. This could happen if the 10 worst roads in 2010 just had a bad year in 2010 and this is the reason why they were the worst. If they have an average year in 2012 then the model will capture this as an effect of the safety measures.

Question – t test!

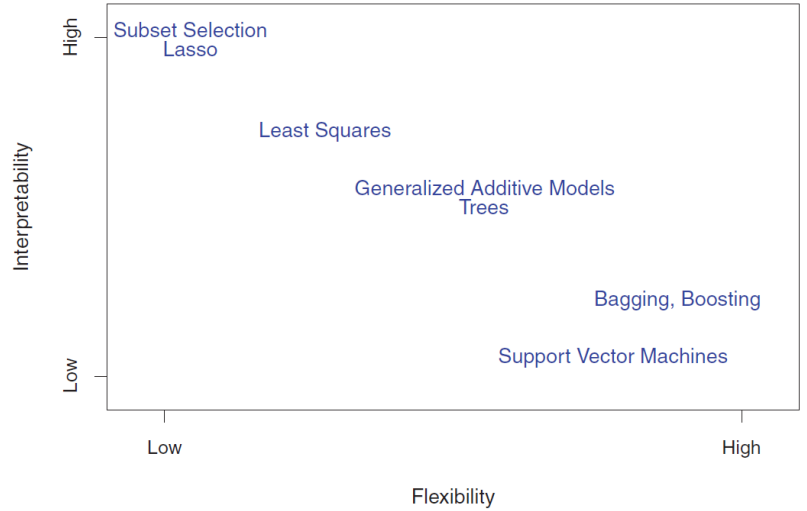
- In case we implement safety measures on the 10 most dangerous roads and estimate the linear regression $y_{2012} - y_{2010} = \beta' x + \gamma * 1(safety\ measure) + u$
  Which of the assumptions about linear regression is most problematic?

A. $E(\varepsilon_i) = 0$
B. $VAR(\varepsilon_i) = \sigma^2$
C. $COV(\varepsilon_i, \varepsilon_j) = 0$
D. $COV(X_i, \varepsilon_j) = 0$

Answer:

Model selection

- Remember that we can make models simpler or more complex
- What is best depends on the context and problem
- Next, we discuss subset selection is a little more detail
- Two arguments for subset selection
  - Prediction accuracy
  - Model interpretation



Best subset selection

- This selection method is based on looking at all possible submodels. Suppose that we have p predictors (explanatory variables).

- The procedure is as follows

1. Denote the first model M_0. This is the model with an intercept only
2. For k=1,2,…,p
   a) Estimate all $\binom{p}{k}$ models with exactly k predictors
   b) Choose the best among these models based on RSS or R^2
3. This gives $p + 1$ models. Choose the best among these using cross validation, AIC, BIC or adjusted R^2

- This is a useful method is p is small.

Forward stepwise selection

- This selection method is based on looking at a specific path of submodels. Suppose that we have p predictors (explanatory variables).

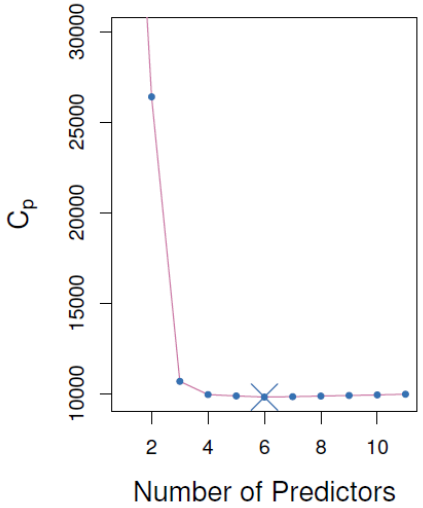
- The procedure is as follows

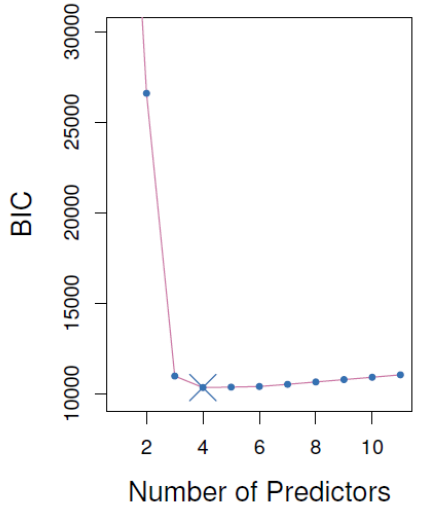
1. Denote the first model M_0. This is the model with an intercept only
2. For k=0,2,…,p-1
   a) Estimate all $p - k$ models that add one predictor to M_k
   b) Choose the best among these models based on RSS or R^2
3. This gives $p + 1$ models. Choose the best among these using cross validation, AIC, BIC or adjusted R^2

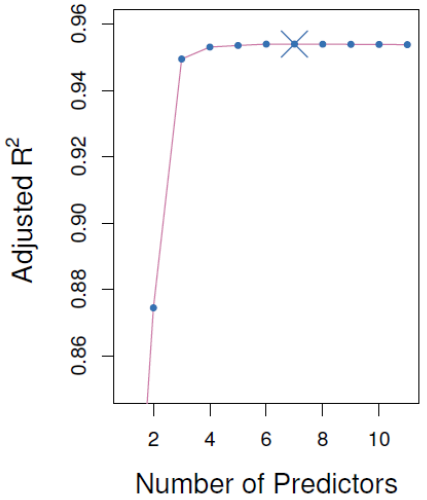
- This is a potentially useful method for any number of p.

Other selection approaches

- Backward selection – start with a
- Hybrid approaches – any combi
- All these approaches needs a m
  - adjusted $R^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$







  - Akaike Information Criterion: $\mathrm{AIC} = 2 * \mathrm{k} - 2 * \mathrm{LL}$,
  - Bayesian Information Criterion: $\mathrm{BIC} = \ln(\mathrm{n}) * \mathrm{k} - 2 * \mathrm{LL}$

where n is sample size, k number of parameters and LL the final loglikehood for the model

Moving beyond linearity

- Remember that the simplest specification is the model that is linear in all variables of interest, and

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

- In the next slides, we will discuss how to loosen the linearity assumption of a specific explanatory variable using

- Polynomial regression
- Step functions
- Regression splines

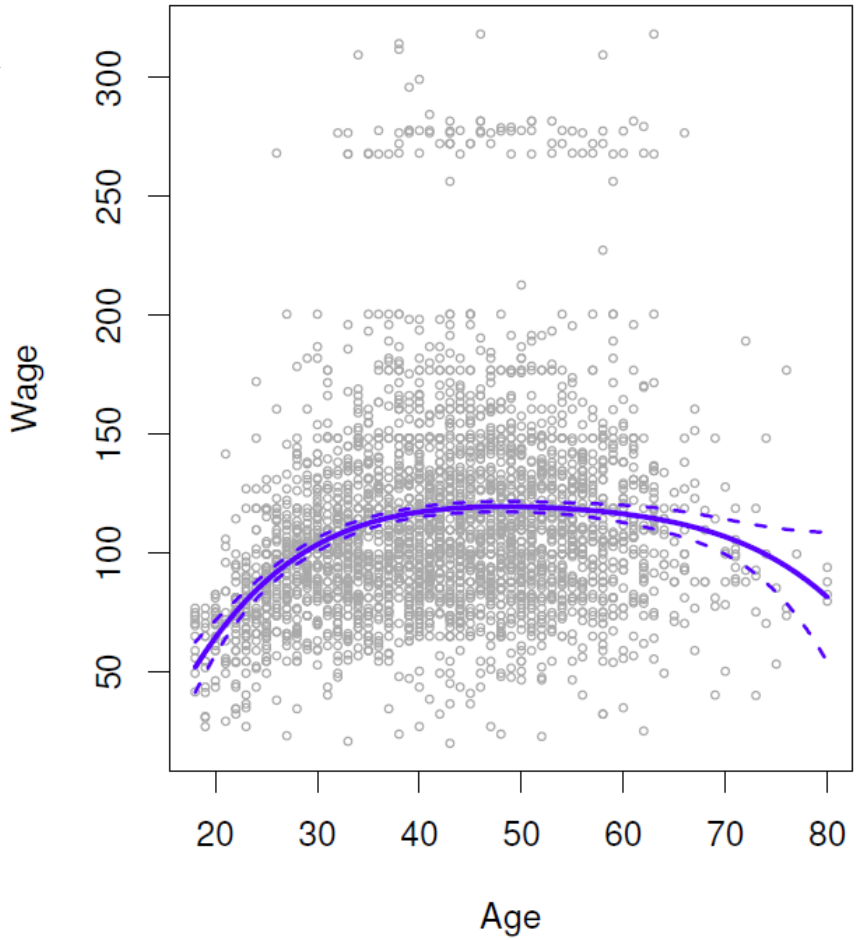
- NB. The book only covers this for a single explanatory variable in sections 7.1-7.4. However, it can easily be extended to more variables. This is sometimes referred to as Generalised Additive Models (GAMs), see section 7.7 for a brief discussion.

Polynomial regression

- Consider the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{1i}^d + \varepsilon_i$$

- If we consider wage data (p.291) modelled with age up to the fourth power we get the prediction on the right with its confidence interval.

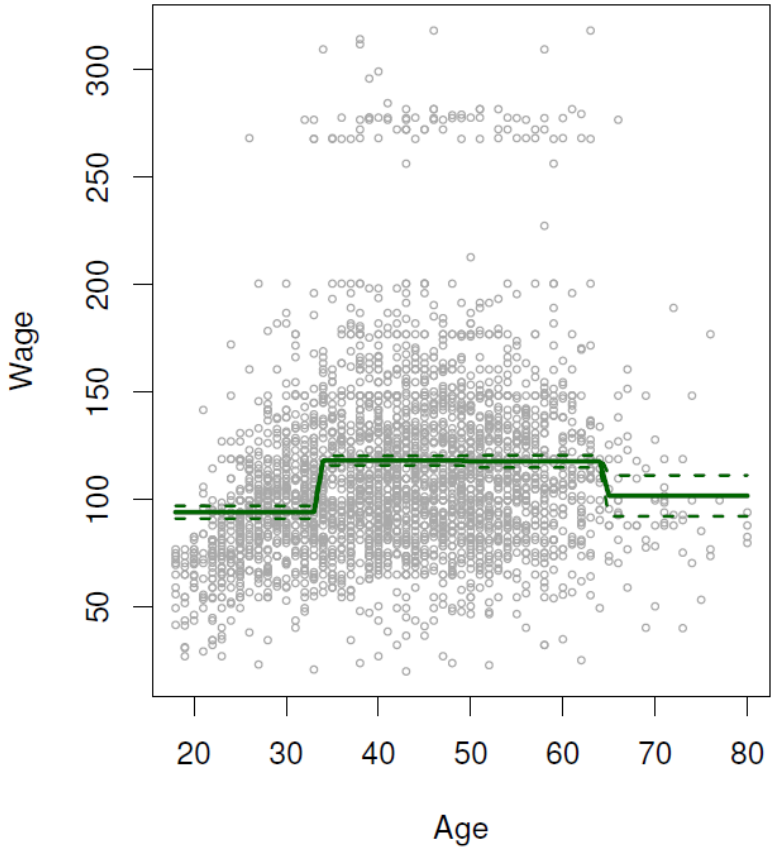


Step functions in regression

- Consider the model

$$Y_i = \beta_0 + \beta_1 C_1(X_{1i}) + \cdots + \beta_K C_K(X_{1i}) + \varepsilon_i$$

- If we consider wage data (p.293) modelled with age up to the fourth power we get the prediction on the right with its confidence interval.



Regression splines

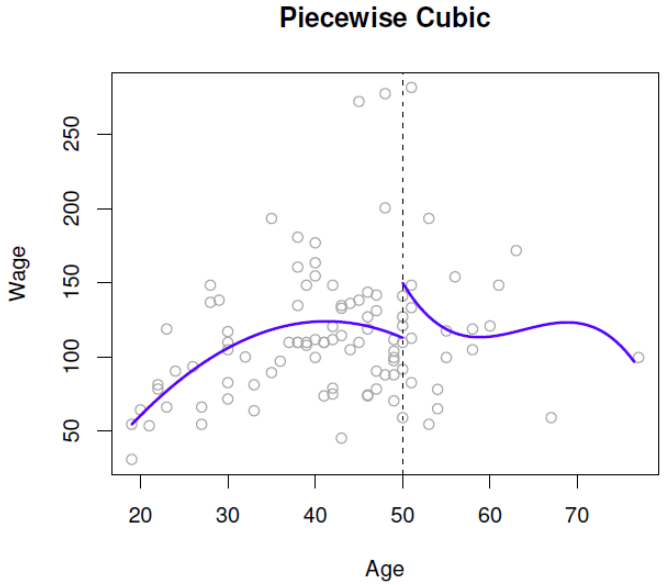
- Both polynomial regression and step functions have the format of basis functions, i.e.

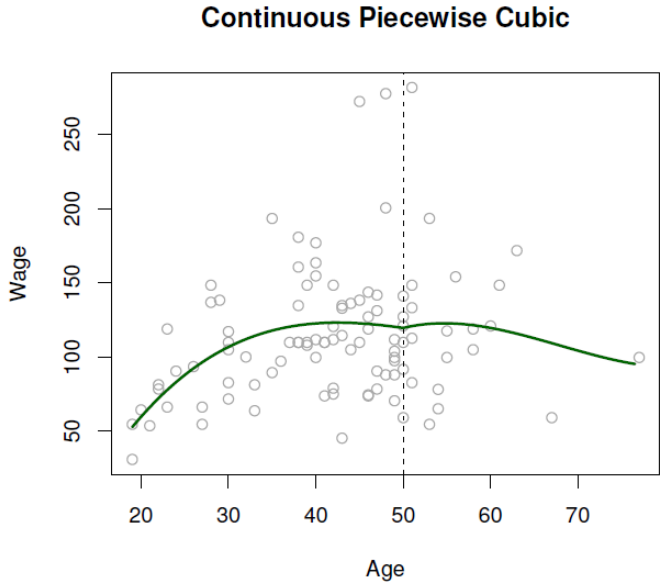
$$Y_i = \beta_0 + \beta_1 b_1(X_{1i}) + \cdots + \beta_K b_K(X_{1i}) + \varepsilon_i$$

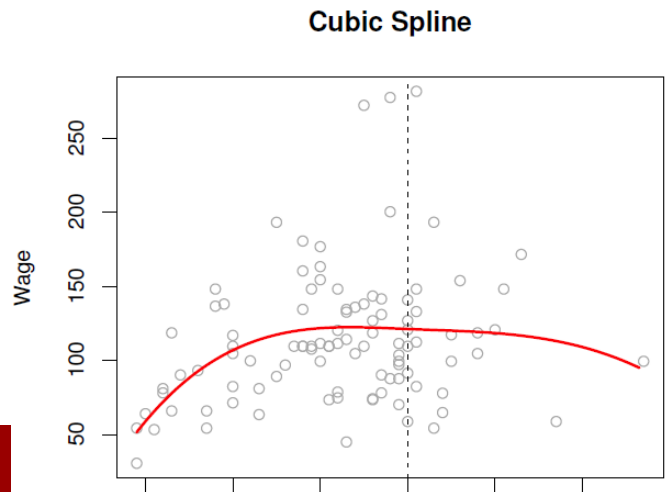
- If we combine these two ideas we can get a variety of functional forms, e.g. if we allow the functions to be polynomials.

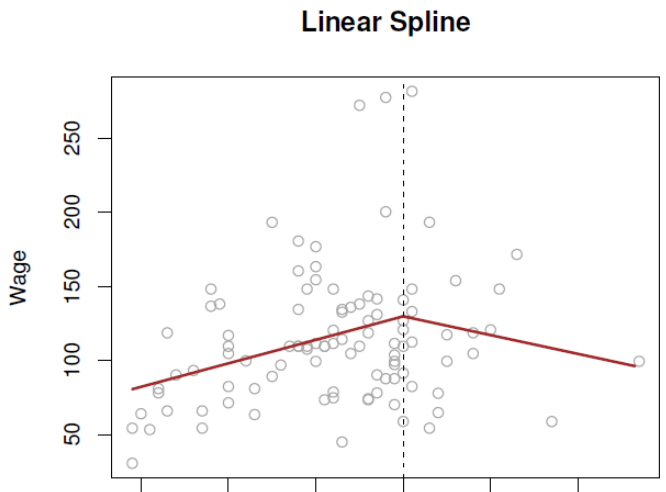
- Four examples of different basis functions on the wage data (p. 296)
  - Piecewise cubic regression
  - Continuous piecewise cubic regression
  - Cubic spline
  - Piecewise linear

Regression splines – examples





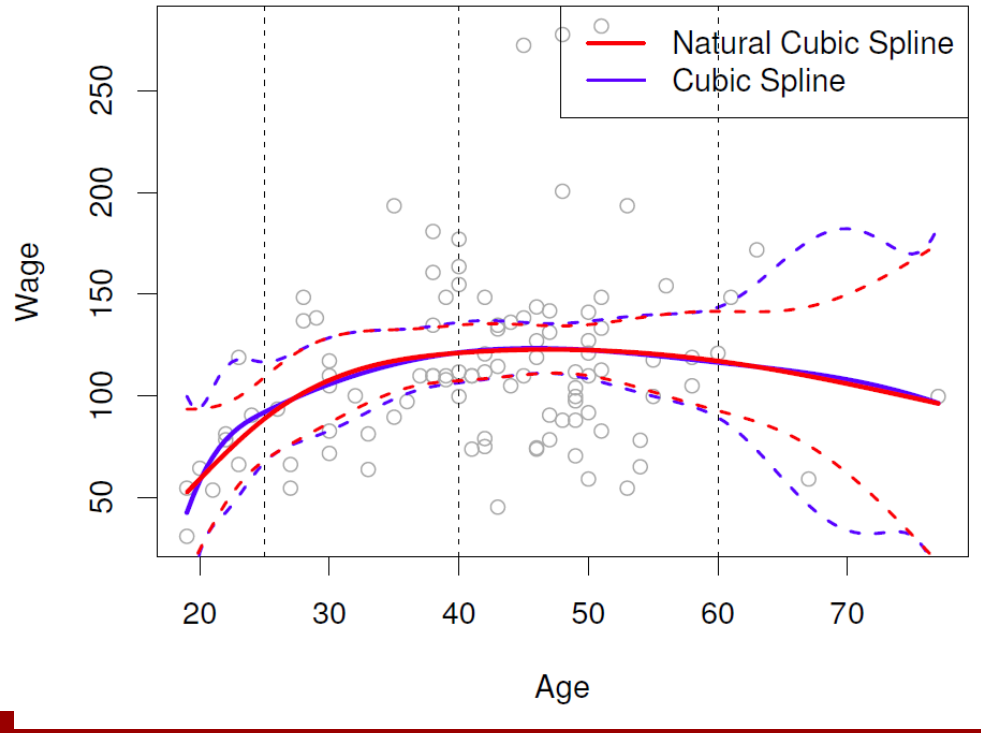




Regression splines vs. natural splines

- Splines can be data demanding as they need sufficient data within each interval. This could be problematic at the borders, which is why we have natural splines.

- These replace the two most extreme polynomials with linear splines as seen in the wage example (p.298)



Project 3

- Some groups have started working on Project 3 - part 1, i.e. the two exercises on linear regression. There will be a part 2 on classification of approx. the same size.

- All groups should continue the work on Project 3 today.

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 6 (Markus, Caroline H., Gustav, Christian) 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 3.3.3-3.4 + 6-6.1 + 7-7.4
  - (chap. 5 is a brief intro to cross-validation and bootstrapping)

- To prepare for lecture 11, you should read
  - Introduction to statistical learning chap 4-4.3 (2.2.3 is recap)

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

OLS estimation

- We can estimate the linear regression model using the ordinary least squares (OLS) method.

- The parameters, β, in the model are found as the values that minimise the function

$$Q_{min} = \sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$$

- This method gives unbiased and normal estimates under assumptions 1-6.

- Even without assumption 6 about normality, we get consistent and asymptotically normal estimates for large samples.

OLS estimation

- You can show that the estimates are given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- This formula can be used to show that

$$E(\hat{\beta}) = \beta$$

and that

$$VAR(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Checking assumptions - Exogeneity

- This assumption can be stated either as $COV(X_i, \varepsilon_j) = 0$ or $E(\varepsilon_i | X_i) = 0$.

- Exogeneity means that there is no variables affecting Y not included in the model, i.e. included in the residuals, that are related to X.

- In a annual income model

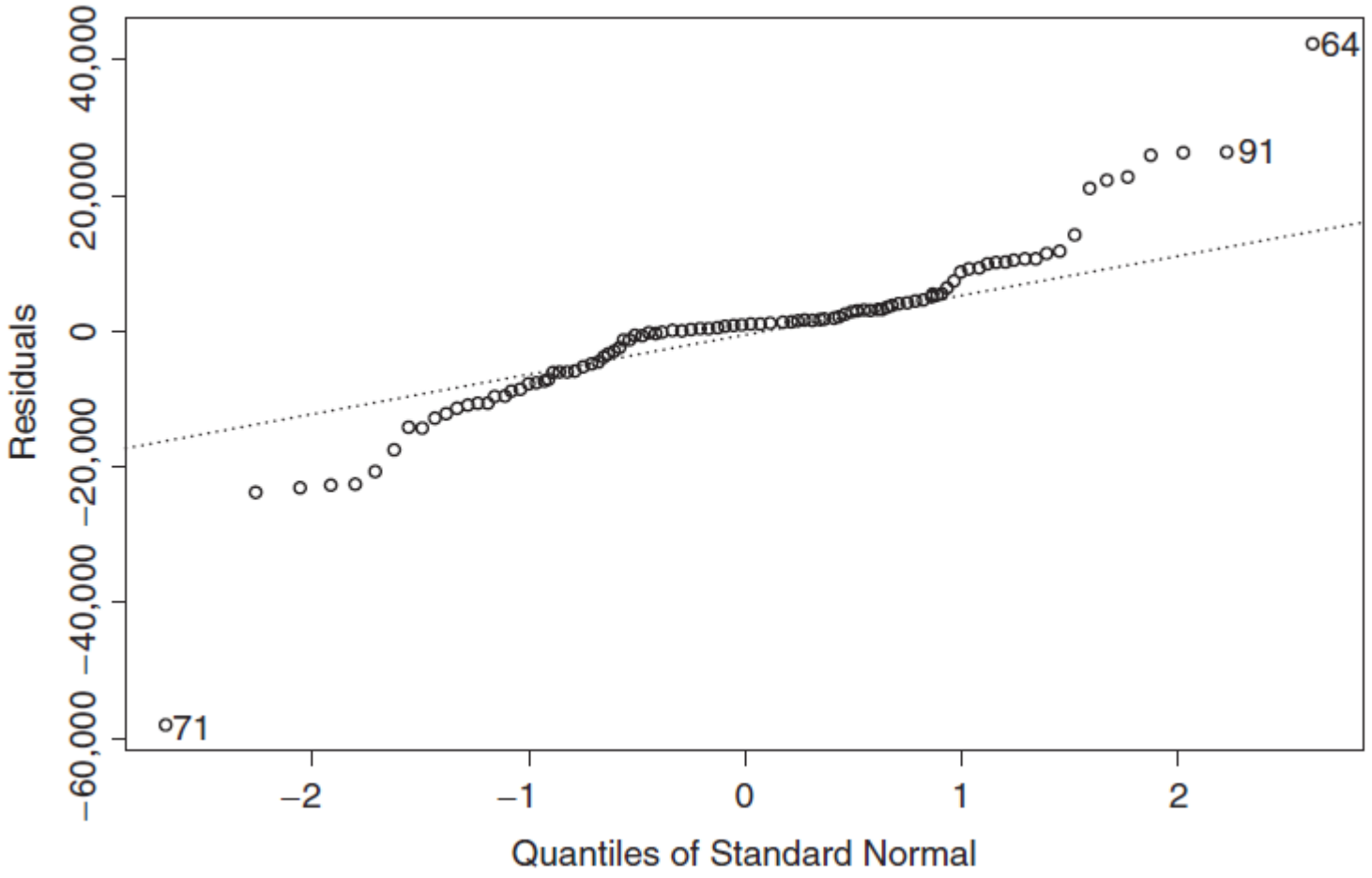
$$Y_{Income} = \beta_0 + \beta_1 X_{years\ of\ educ} + \beta_2 X_{age} + \beta_2 X_{age}^2 + \varepsilon_i$$

- There could be some aspect relevant for income that we do not have it in our data. These aspects will then be included in the residuals and they also might be correlated with years of education. Hence, $X_{years\ of\ educ}$ becomes endogenous and the parameter estimates biased.

- Methods to solve exogeneity problems are beyond this course. But be aware and discuss potential endogeneity issues. In addition, always include all relevant variables in your data.

Checking assumptions - normality

- Normality can be checked by a histogram or a Q-Q plot.

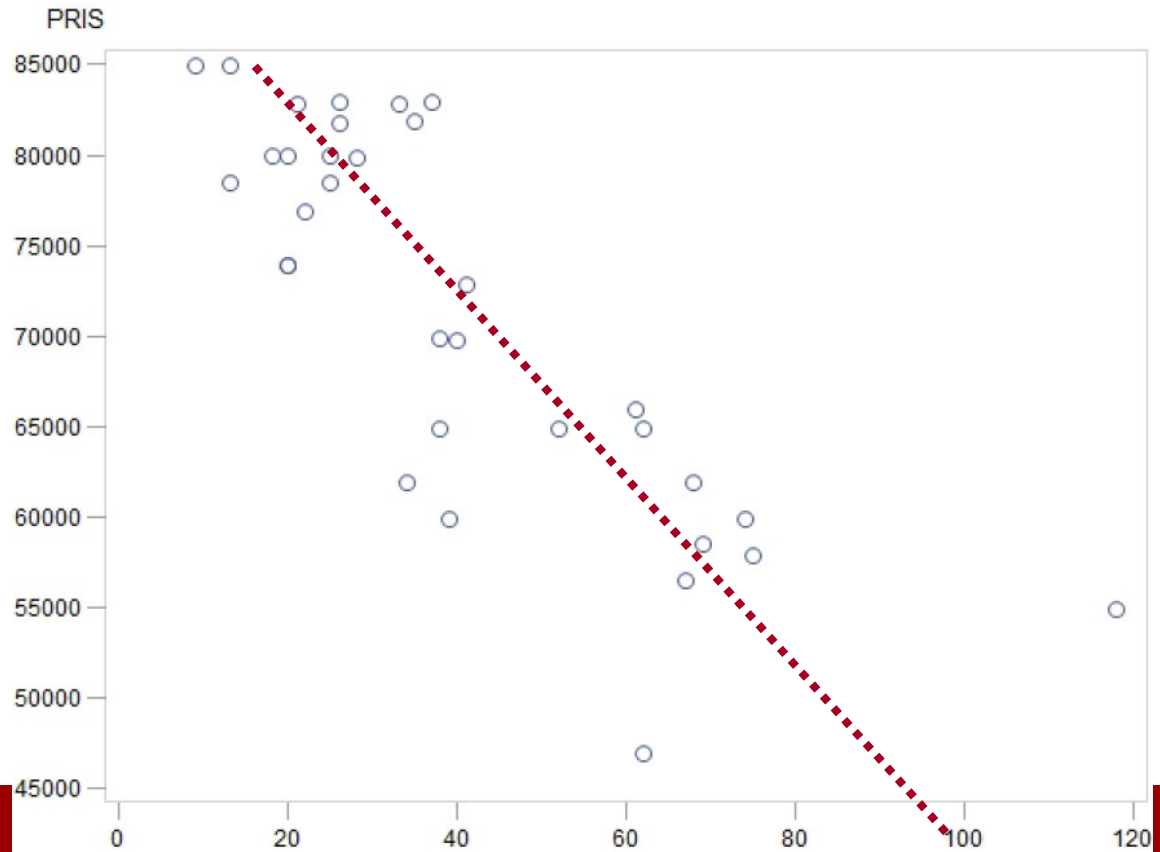


Linear regression – example

- The data set, Nissan, describes trades with 33 used Nissan Micra cars in 1987. Every row describes an observed trade. The variable Pris is the used car price, Km is the number of km driven by the car (in 1000 km), and Aar is the production year of the car (årgang).

- The price (Pris) is a continuous variable.

- A scatter plot of price on km shows that price falls approx. linearly with km.



Linear regression – Nissan Micra example

- For our Nissan Micra data, we get the following results

Number of Observations Read	33
Number of Observations Used	33

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2339579905	2339579905	58.00	<.0001
Error	31	1250518883	40339319		
Corrected Total	32	3590098788			

Root MSE	6351.32418	R-Square	0.6517
Dependent Mean	71439	Adj R-Sq	0.6404
Coeff Var	8.89051		

R^2

Adj. R^2

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	85850	2191.56767	39.17	<.0001
KM	1	-357.82396	46.98560	-7.62	<.0001

p value for test of $\hat{\beta}_0, \hat{\beta}_1$ against 0

Estimates $\hat{\beta}_0, \hat{\beta}_1$

Standard errors on $\hat{\beta}_0, \hat{\beta}_1$

T test of $\hat{\beta}_0, \hat{\beta}_1$ against 0

Linear regression – Nissan Micra example

- The model shows that *km* has a **negative significant** effect (i.e. the t test of the null hypothesis about the *km* parameter equal to zero can be rejected) on price because the t test of the *km* parameter is numerically larger than 1,96 (which is equivalent to the p value < 0.05).

- The marginal effect of km on price is $\frac{\partial y}{\partial x} = \beta_1 = -357.8\ kr/1000km$

- The prediction model is $\hat{y} = 85{,}850 - 357.8x_1 + u$

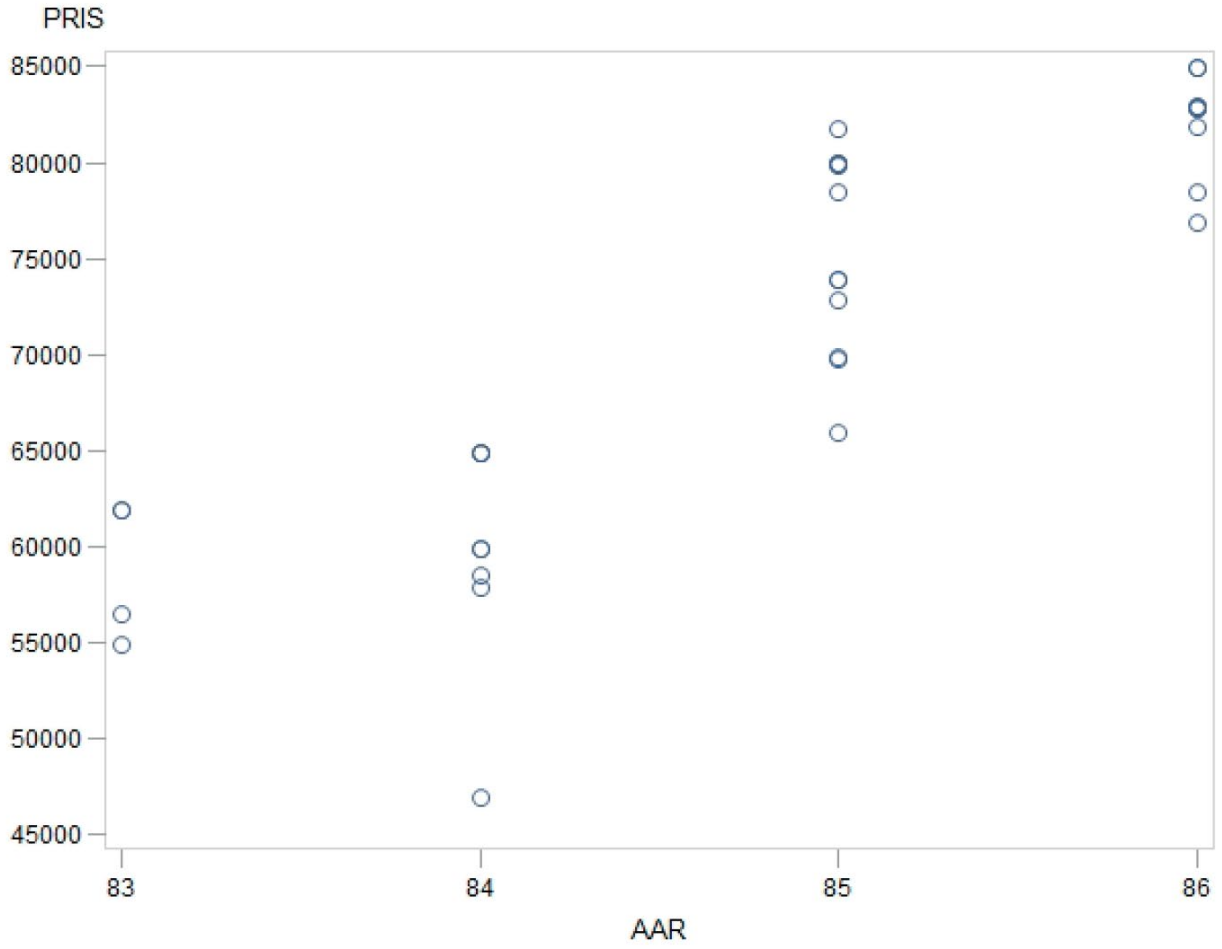
  with average predictions $\hat{y} = 85{,}850 - 357.8x_1$

  i.e. the elasticity is $\varepsilon_{Y,X} = -357.8\frac{40.3}{85{,}850-357.8*40.3} = -0.202$

  at 40.3 (measured in 1000 km).

Linear regression – example

- What is the effect of year/age on price?



Linear regression – Nissan Micra example

- Now with age as extra explanatory variable

Number of Observations Read	33
Number of Observations Used	33

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2916211485	1458105742	64.91	<.0001
Error	30	673887303	22462910		
Corrected Total	32	3590098788			

Root MSE	4739.50526	R-Square	0.8123
Dependent Mean	71439	Adj R-Sq	0.7998
Coeff Var	6.63430		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	92053	2042.94175	45.06	<.0001
KM	1	-167.98756	51.31466	-3.27	0.0027
Alder	1	-6260.33409	1235.60970	-5.07	<.0001

Linear regression – Nissan Micra example

- The model shows that both *km* and *age* have **negative significant** effects.
- The marginal effect of km on price is $\frac{\partial y}{\partial x} = \beta_1 = -168.0\ kr/1000km$
- The marginal effect of age on price is $\frac{\Delta y}{\Delta x} = \beta_1 = -6260.3\ kr/year$
- The prediction model is $\hat{y} = 92{,}053 - 168.0x_1 - 6260.3x_2 + u$

  i.e. the km elasticity is $\varepsilon_{Y,X_1} = -168.0 \frac{40.3}{92{,}053 - 168.0*40.3 - 6260.3*2.21} = -0.095$

  and the age pseudo elasticity is

  $$\varepsilon_{Y,X_2} = -6260 \frac{2.21}{92{,}053 - 168.0*40.3 - 6260.3*2.21} = -0.194$$

  at 40.3 (measured in i 1000 km) and 2.21 years.

Linear regression – Nissan Micra example

- Other functional forms

Root MSE	4201.52853	R-Square	0.8672
Dependent Mean	71439	Adj R-Sq	0.8427
Coeff Var	5.88125		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	70944	4719.72838	15.03	<.0001
KM	1	-243.18850	131.66821	-1.85	0.0757
Km_2	1	0.87702	1.04805	0.84	0.4100
d_86	1	16203	3442.55934	4.71	<.0001
d_85	1	10965	3141.07316	3.49	0.0017
d_84	1	-104.81322	2805.04282	-0.04	0.9705

Root MSE	4114.57290	R-Square	0.8632
Dependent Mean	71439	Adj R-Sq	0.8491
Coeff Var	5.75953		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	68059	3115.83545	21.84	<.0001
KM	1	-136.82099	45.60252	-3.00	0.0055
d_86	1	17174	2570.99119	6.68	<.0001
d_85	1	11568	2255.28085	5.13	<.0001

Root MSE	4179.10961	R-Square	0.8589
Dependent Mean	71439	Adj R-Sq	0.8443
Coeff Var	5.84987		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	81063	7825.70944	10.36	<.0001
ln_km	1	-5292.05744	1890.36661	-2.80	0.0090
d_86	1	17095	2705.24266	6.32	<.0001
d_85	1	12102	2221.03698	5.45	<.0001

Linear regression – Nissan Micra example

- The various models that we have tested are:

1. $pris = \beta_0 + \beta_1 km + u$
2. $pris = \beta_0 + \beta_1 km + \beta_2 alder + u$
3. $pris = \beta_0 + \beta_1 km + \beta_2 km * km + \beta_3 * d_{86} + \beta_4 * d_{85} + \beta_5 * d_{84} + u$
4. $pris = \beta_0 + \beta_1 km + \beta_3 * d_{86} + \beta_4 * d_{85} + \mathrm{u}$
5. $pris = \beta_0 + \beta_1 \ln(km) + \beta_3 * d_{86} + \beta_4 * d_{85} + u$

- The estimation results suggest that either model 4 or 5 is the best model. Which of these to prefer would depend on model control (residual plots) and non-statistical arguments, e.g. based on elasticities or theory.

- An argument could be made for model 4 based on Ocham's razor. If two competing ideas explain something similarly well, you should prefer the simpler one.