Universiteit Utrecht

FACULTEIT SOCIALE WETENSCHAPPEN

21 January 2021

Statistics Course
Rabobank
day 1

Afdeling Methoden & Statistiek (M&S)

Five meetings

- January 21 Basics of modern statistics; Inference
- February 11 How to determine cut-off values
- March 11 Classifications
- March 25 Distributions
- April 8 …

The team from the UU

Dr. Peter Lugtig

prof. Dr. Peter van der Heijden

dr. Gerko Vink

Prof. Dr. Daniel Oberski

dr. Rebecca Kuiper

prof. dr. Stef van Buuren

Today

- Inference vs. prediction
- Inference under a design
- The role of (hypothesis) testing
- Inference using a model
  - Under- and overfitting.
  - Linear and logistic regression,
  - Nonlinear model, splines

Today

- Inference vs. prediction
- Inference under a design
- The role of (hypothesis) testing
- break
- Inference using a model
  - Under- and overfitting.
  - Linear and logistic regression,
  - Nonlinear model, splines
- *break*
- 2 cases

Inference or prediction

- “What is the inflation rate in the Netherlands”?
- “What determines variations in house prices in NL”?
- “Should Peter receive a higher credit limit?”
- ”What is the risk of Company X going bankrupt next year”?
- “Which type of households are at risk of getting behind on their mortgage”?
- “If mortgage interest rates increase to 5% next year, how many households will default on their mortgage?”

Inference or prediction

- “What is the inflation rate in the Netherlands”?
- “What determines variations in house prices”?
- “Should Peter receive a higher credit limit?”
- ”What is the risk of Company X going bankrupt next year”?
- “Which type of households are at risk of getting behind on their mortgage”?
- “If mortgage interest rates increase to 5% next year, how many households will default on their mortgage?”

Inference or prediction

- “What is the inflation rate in the Netherlands”?
- “What determines variations in house prices”?
- “Should Peter receive a higher credit limit?”
- ”What is the risk of Company X going bankrupt next year”?
- “Which type of households are at risk of getting behind on their mortgage”?
- “If mortgage interest rates increase to 5% next year, how many households will default on their mortgage?”

Inference central today

- “What is the inflation rate in the Netherlands”?
- “What determines variations in house prices”?
- “Should Peter receive a higher credit limit?”
- ”What is the risk of Company X going bankrupt next year”?
- “Which type of households are at risk of getting behind on their mortgage”?
- “If mortgage interest rates increase to 5% next year, how many households will default on their mortgage?”

Inference:
Using a small dataset to draw big conclusions

Consumer Price Inflation:

- the price of a weighted average market basket of consumer goods and services purchased by households

Market basket: not all products, but a *selection*

Weighted average: some products more important than others.

So not a random selection, but a *weighted selection*

Sampling and inference

You hardly ever have all the data

- Sample of people
  - Opinion polls to tell about voting behavior in NL
- Sample of products
  - Product prices to inform CPI in NL
- Sample of business
  - Investments, output in NL

- For good inference you need to know: sample <-> population (NL)

Inference under a design

- Ideally: You know exactly how sample was selected from population.

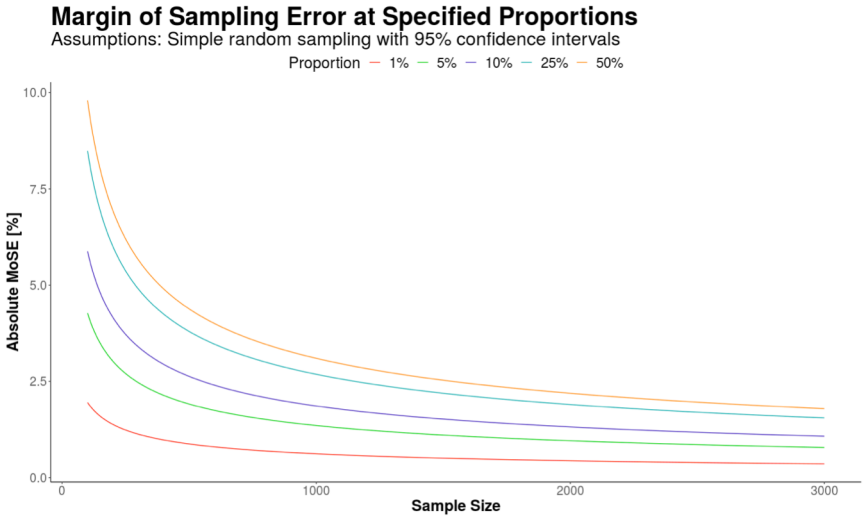
1. Simple random sample (a real lottery)
2. Stratified designs
   - Unequal selection probabilities, but still known

Design-based inference (1): Simple Random Sampling

"Real lottery"

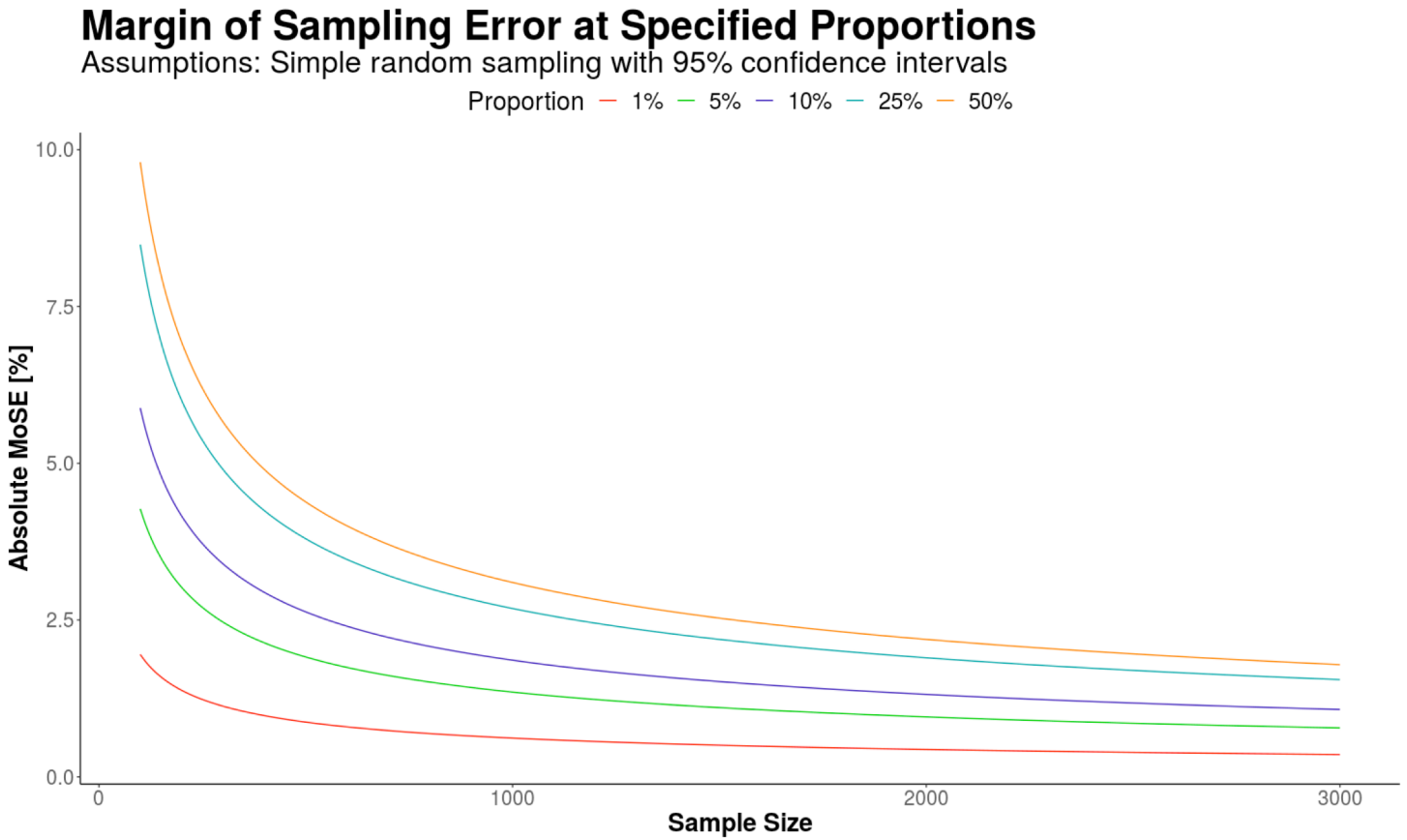
Equal chance for everyone in population to be selected

- In long run, the sample will exactly resemble the population
  - On all possible statistics!
  - Inference is easy
- In 1 sample, some uncertainty
- The larger the sample size, the lower uncertainty



- Example: a simple random sample of account holders

Design-based inference (1): Simple Random Sampling and Margin of error



Intermezzo: The idea of testing in inference

- % of unemployed in the Netherlands
  - Q4 – 2021: 2.8%
  - Q1 – 2022 : ???

- Sample 1 (n=100): 2.5%
  - Standard error (s.e.): p(1-p)/√n -> (.025*.975)/ √100 = .002 (0.2%)
  - 95% Confidence Interval: 2.5 +- 1.96*0.2. [2.1;2.9]
  - Confidence overlaps with Q4: no certain decline in employment
- Sample 2 (n=1000): 2.5%
  - Standard error: p(1-p)/√n -> (.025*.975)/ √1000 = .0008 (.08%)
  - 95% Confidence Interval: 2.5 +- 1.96*0.08 [2.34;2.66] -> decline!
- (EBB in 2020. n=45000: s.e. = .0001)

Design-based inference (2): Stratified designs

Example: CPI

- Not all products should get an equal weight!
- Products that people spend lots on, should compose a higher proportion of the basket
  - Rent, Groceries, energy, insurance
- Sample with unequal weights - stratification

- Example: sample stratified on activity of account

Design-based inference (3):

A BIG problem:

- Very often, no control over sampling design
- You just have to deal with the data you get
- Unclear how it matches the population
- Difference between sample and population?
  - Possible bias

Model-based inference (3):

- Statistical model to describe reality
  - “What is the inflation rate in the Netherlands”?
- Sometimes inference more about variation
  - “What determines variations in house prices”?

- Many, many different statistical models
- Today: Regression

Regression: overview

$$y = f(x) + \epsilon$$

Where, $f(\ )$ is the function relates the predictors x with the outcome variable y , and ϵ is the irreducible error.

Example: explain house prices (y) with floor surface (m^2).

Different goals of regression

Prediction: the goal is to obtain accurate predictions of *y*.

- Given *x* and *y*, work out *f*(*x*); use *f*(*x*) to predict values of *y*
- "engineering mindset": *A → B, focus on reaching B*

Inference: the goal is to understand the process that relates *x* and *y*.

- "Is *x* related to *y*?"
- "How is *x* related to *y*?"
- "How precise are parameters of f(x) estimated from the data?"
- "scientific mindset": *A → B, focus on " →"*

Regression: overview

$$y = f(x) + \epsilon$$

y: Observed outcome;

x: Observed predictor(s);

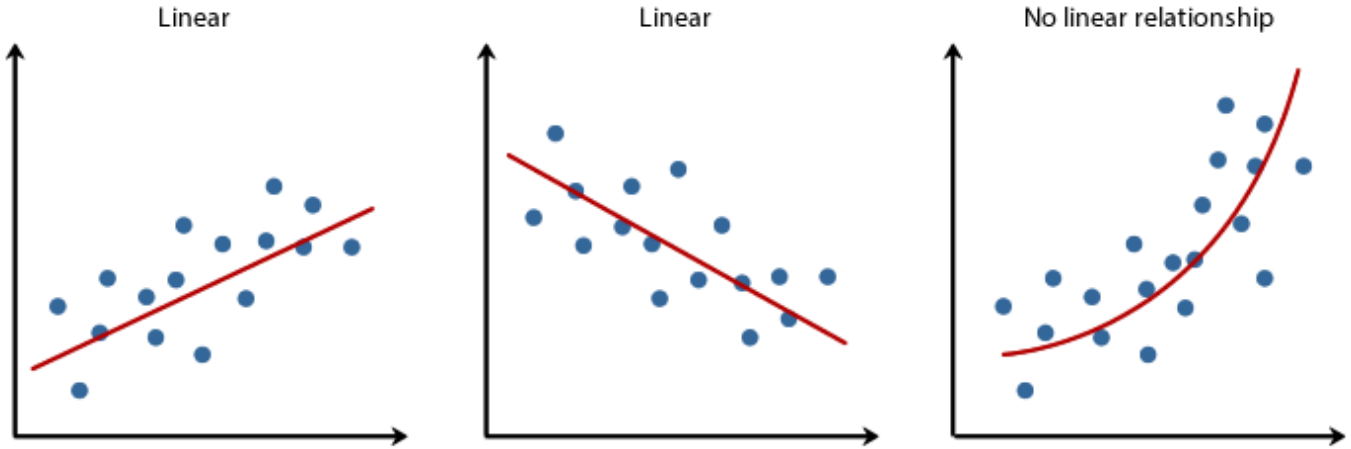
f(*x*): Prediction function, to be estimated;

ϵ: Unobserved residuals, just defined as the "irreducible error", $\epsilon = y - f(x)$

The higher the variance of the irreducible error, variance(ϵ) = σ^2, the less we can explain.

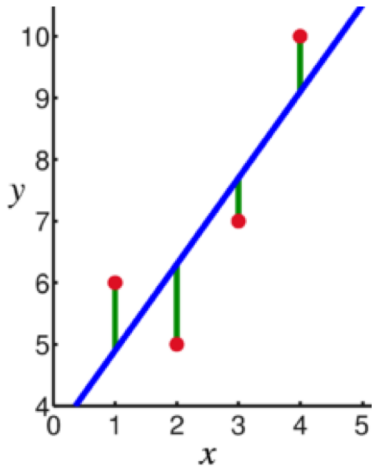
Regression: overview

- $f(x)$ can take different identities:
  - Linear regression vs non-linear regression
  - One or more predictors (x)
    - Floor surface, region, garden, etc….



Linear regression

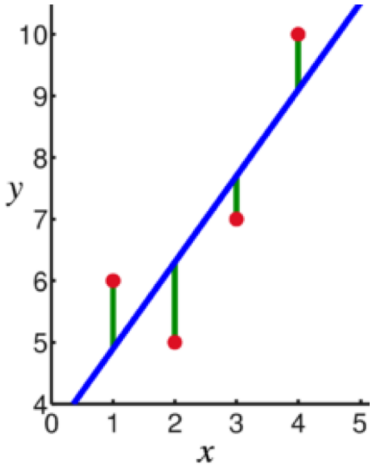
In linear regression, the mean of the outcome *y* is assumed to be a *linear combination* of the regression coefficients *(b's)*.



$$y = b_0 + b_1 x$$

In linear regression, the observations (**red**) are assumed to be the result of random deviations (**green**) from an underlying relationship (**blue**) between a outcome variable (*y*) and an predictor (*x*).

Testing in regression



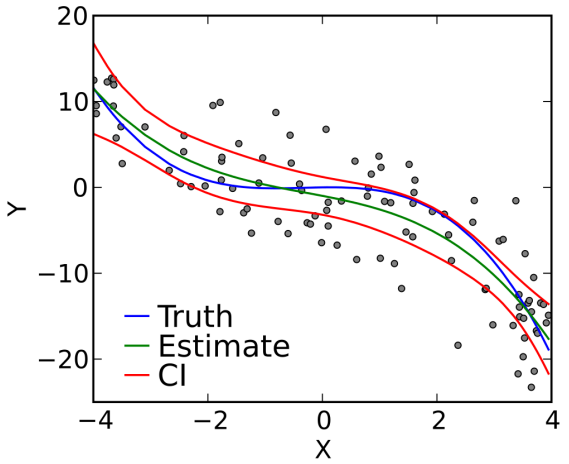
$$y = b_0 + b_1 x$$

- The blue line (b_1) is also estimated and includes uncertainty (sampling)
- We get a standard error with every coefficient
- And then test whether b_1
  - by constructing the confidence interval around B_1 (+- 1.96 * s.e.)
  - and testing whether that confidence includes 0 ($B_1 > 0$)

Break (15 min)

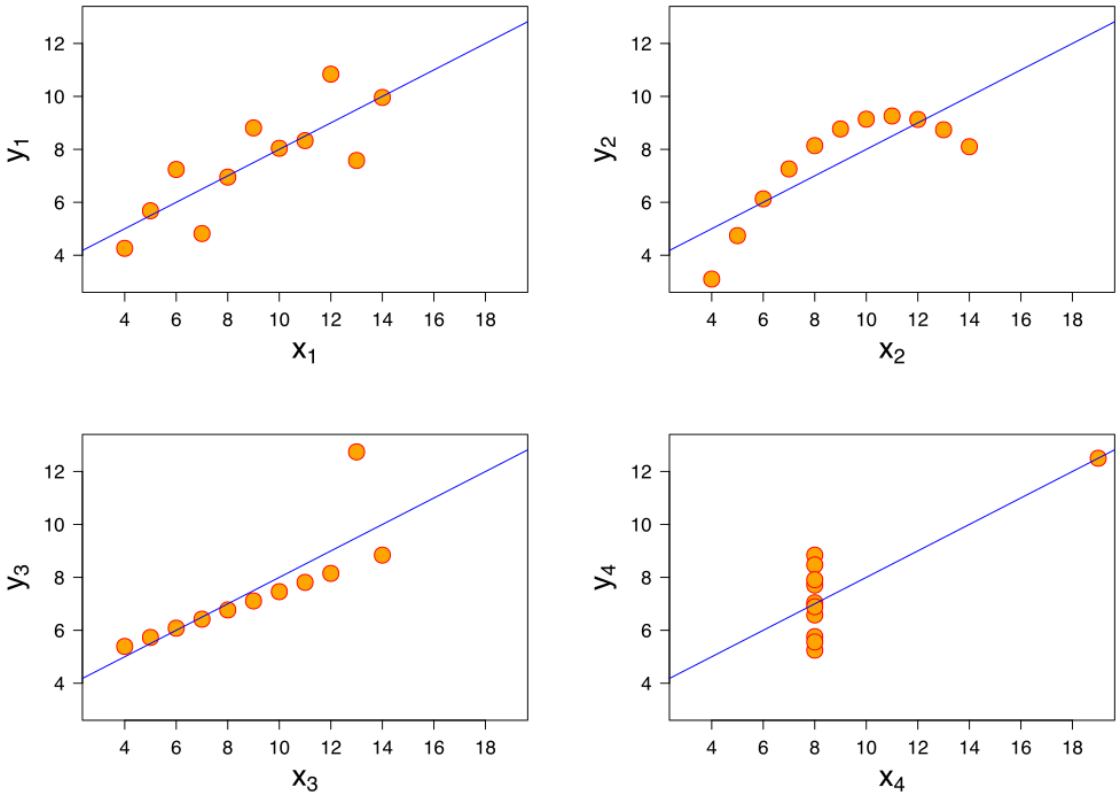
Linear regression

- $f(x)$ can take different identities in linear regression:
  - 1 or more independent variables (predictors):
    - $f(x) = b_0 + b_1 x$
    - $f(x) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$
  - Can be polynomials of different degrees:
    - *1st degree:* $f(x) = b_0 + b_1 x$
    - *2nd degree:* $f(x) = b_0 + b_1 x + b_2 x^2$
    - *nth degree:* $f(x) = b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n$



ie. Polynomial regression of 3rd degree

Issues with fitting a regression

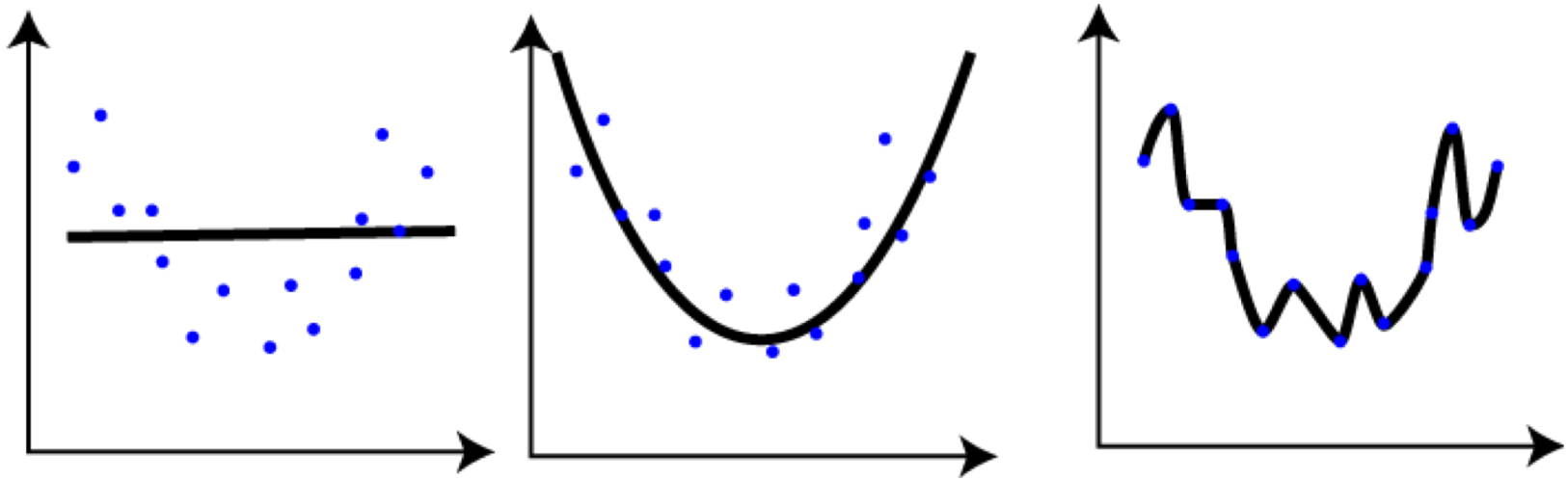


The four models produce the same fitted curves. Can you identify the following?

- *Quadratic relation.*
- *Not related.*
- *Linear relation with observed noise.*
- *Linear relation with an outlier (an extreme observation).*

Why is it so important to visually inspect the regression fitted?

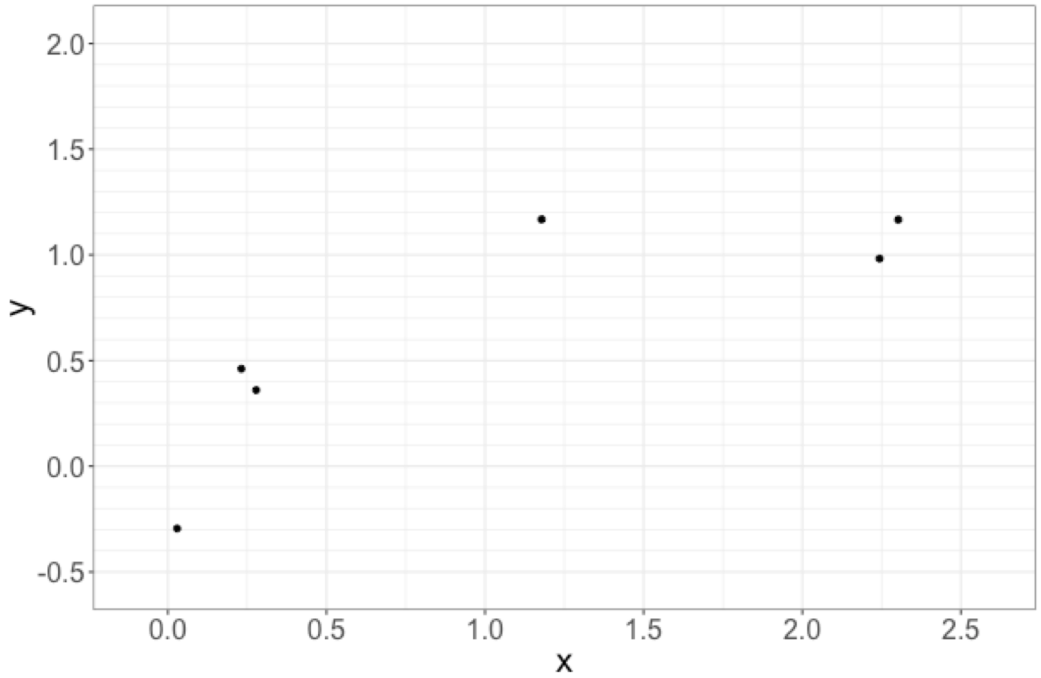
Underfitting and Overfitting



Short exercise - Linear regression

- I am going to show you a small data set, and we are going to try to estimate $f(x)$ and predict y;
- I generated this data set myself using R, so I know the true $f(x)$ and distribution of ϵ.
- For now, however assume that we have a sample

Task: predict (however you want): *y* for *x* = 0.5, 1.5, 2.*0*



Discussion of results

Linear regression:

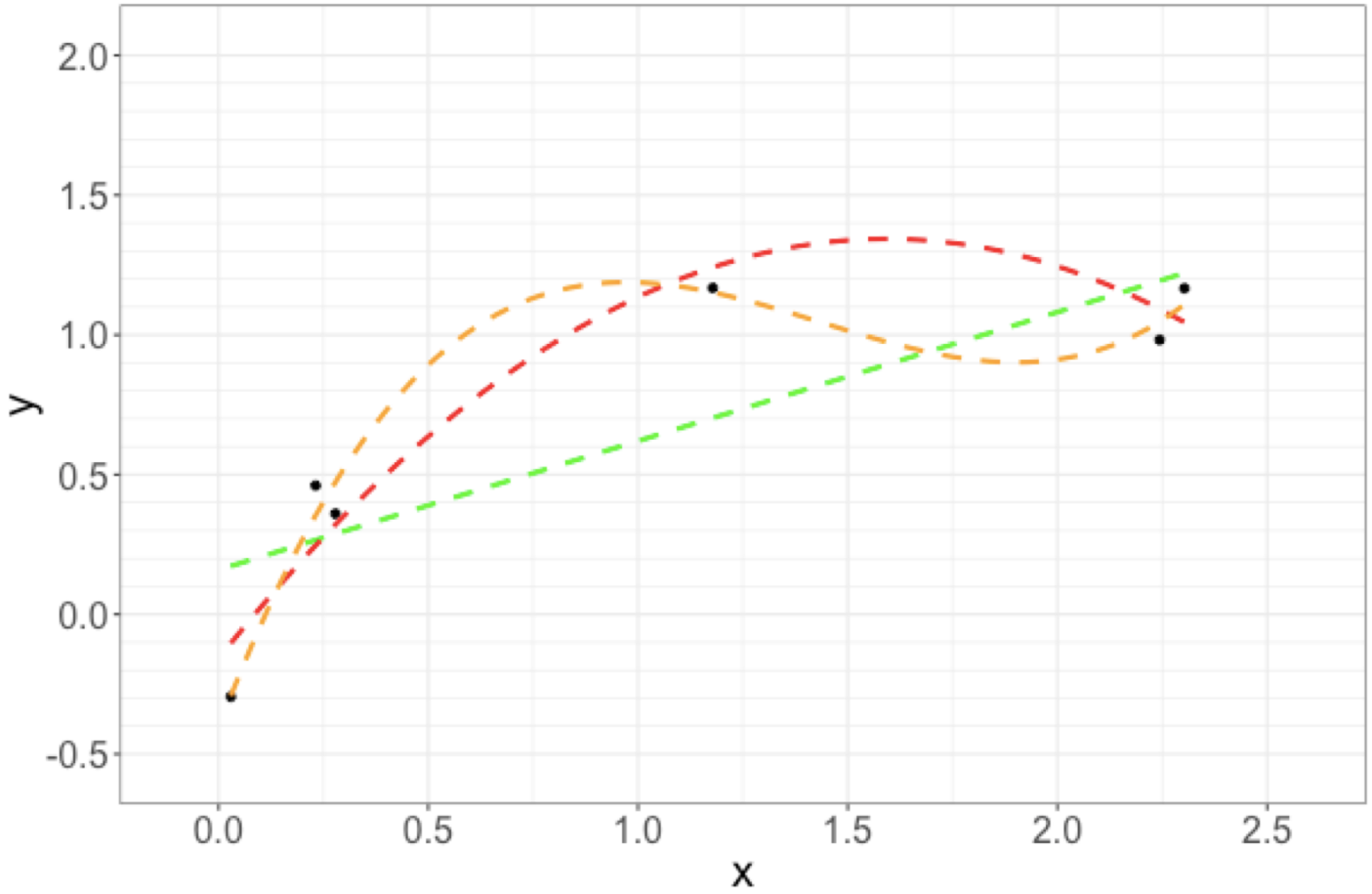
$$y_i = b_0 + b_1 x_i + \epsilon_i \qquad \text{(so, f(x)= } b_0 + b_1 \text{)}$$

Linear regression with quadratic term:

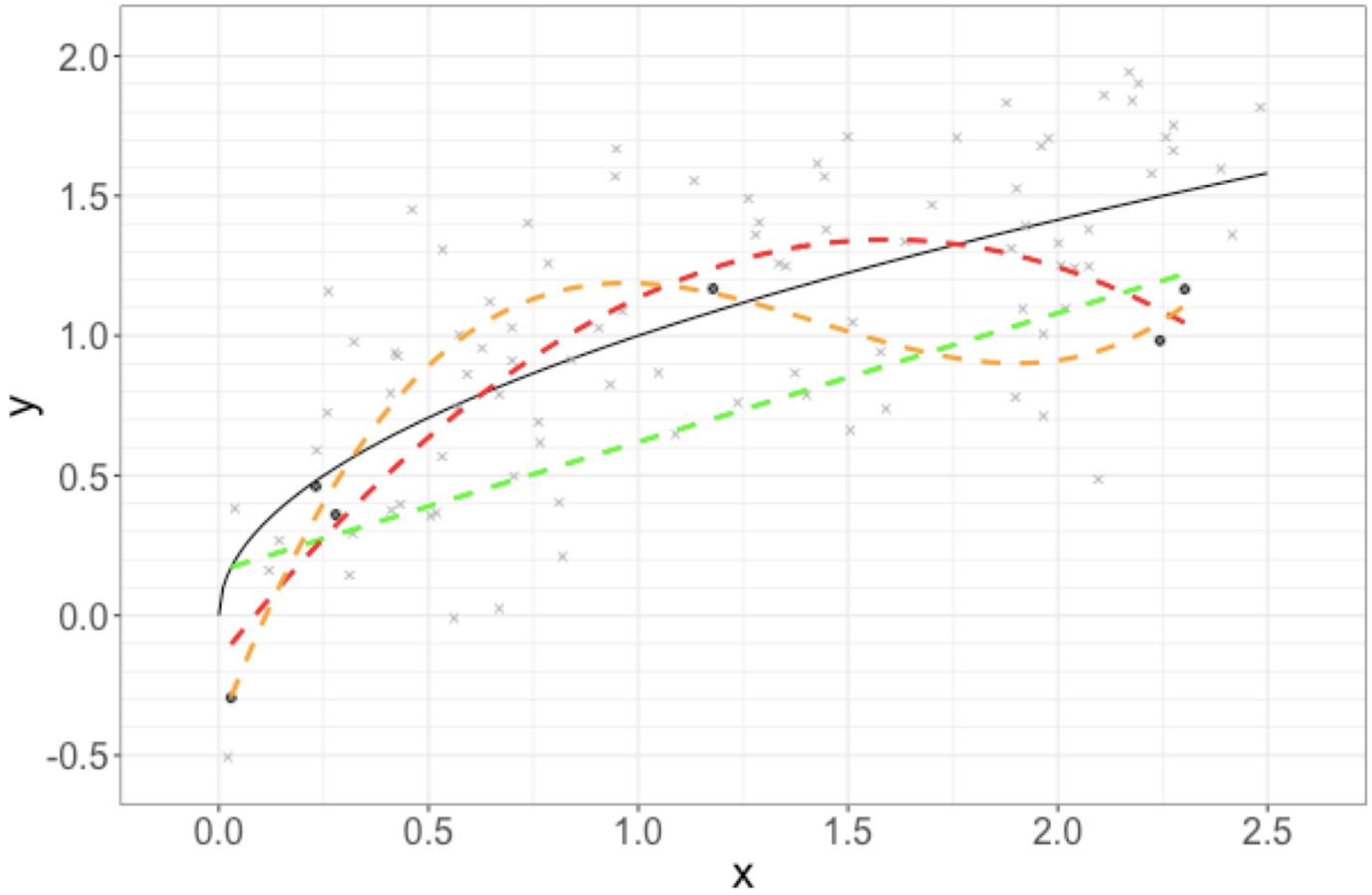
$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + \epsilon_i$$

Linear regression with quadratic and cubic terms:

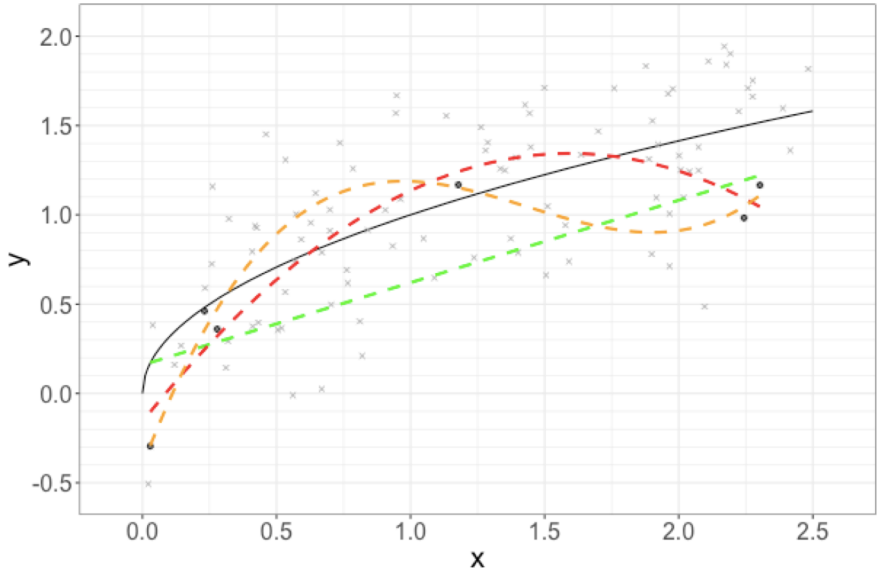
$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3 + \epsilon_i$$


2.0
1.5
1.0
0.5
0.0
-0.5
y
0.0
0.5
1.0
1.5
2.0
2.5
x

The truth (normally we don’t know this)



Discussion of results: comparing results



Model	f(0.5)	f(1.5)	f(2.0)
Eyeballing			
LR (Polynomial degree = 1)	0.40	0.85	1.10
Polynomial (degree = 2)			
Polynomial (degree = 3)			
Actual function	0.71	1.22	1.58

Evaluating model accuracy

- The predictions *y* differ from the true *y*;
- We can evaluate how much this happens “on average”.
- A systematic way of evaluating how good is the fit of a model is using the Mean Squared Error (MSE):

Mean squared error (MSE): $\text{MSE} = n^{-1} \sum_{i=1}^{n} (y - \hat{y})^2$

Discussion of results: comparing MSE results

Model	MSE
Eyeballing	
LR (Pol. degree = 1)	0.1545
Pol. degree = 2	
Pol. degree = 3	

True $f(x)$: $y = \sqrt{x} + \epsilon$

with $\epsilon \sim \text{Normal}(0; 1)$

Discussion of results: what happened?

- There were few observations, relative to the complexity of the polynomial of third degree;
- The polynomial of first degree did not offer a good representation of the nonlinear function, it **underfitted** the data
- The polynomial of second degree performed better in this example;

BUT

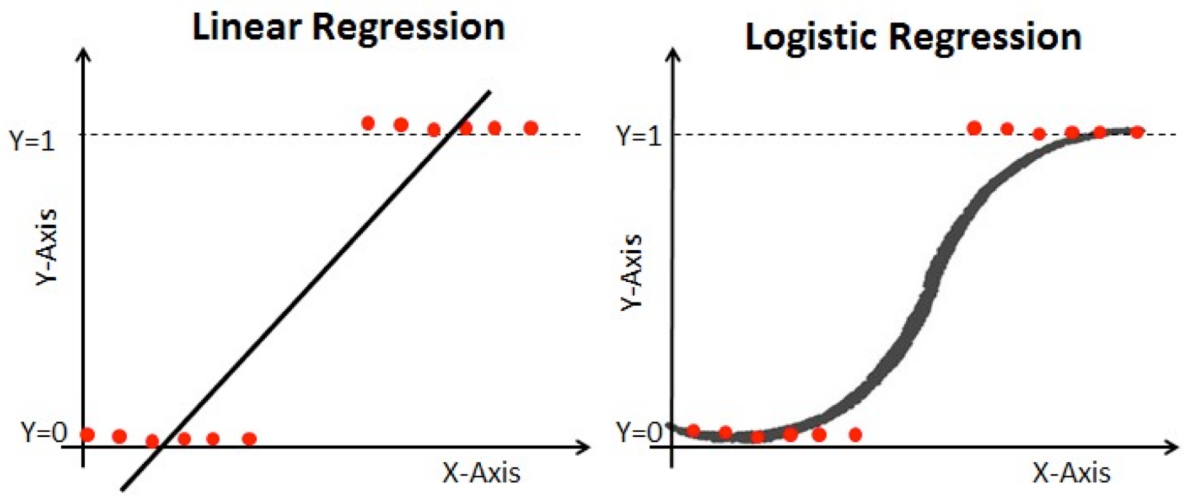
- By chance, some patterns appeared that are not in the true $f(x)$;
- The most flexible model $f(x)$ **overfitted** these patterns
- Perhaps with a larger sample the most flexible model would have performed better than the other two!

Overfitting and what to do…

- Overfitting here happens because of f(x) is too flexible
- Can also happen because of too many predictors (x_1, x_2, x_3,… x_n)
- Good practice:
  - Use predictors and complex polynomials sparingly
  - Hold out samples and cross-validation
    - Split data up in two parts
    - Use part 1 to build the model
    - And part 2 to test it

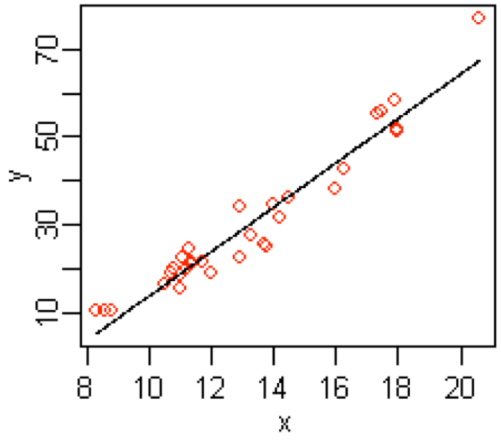
Extensions of regression models (1)

- Logistic regression:
  - Binary outcomes
  - F(x) now not linear, but logistic
  - Marginal effects to interpret results as probabilities
    - If X increases 1 unit, how how the predicted probability on Y-axis change
  - Popular in machine learning

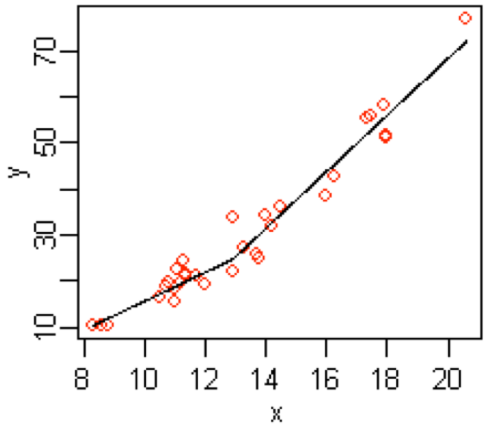


Extensions of regression models (2)

- Splines
  - Not one regression model across whole range of X
  - E.g. When X is income, age,….



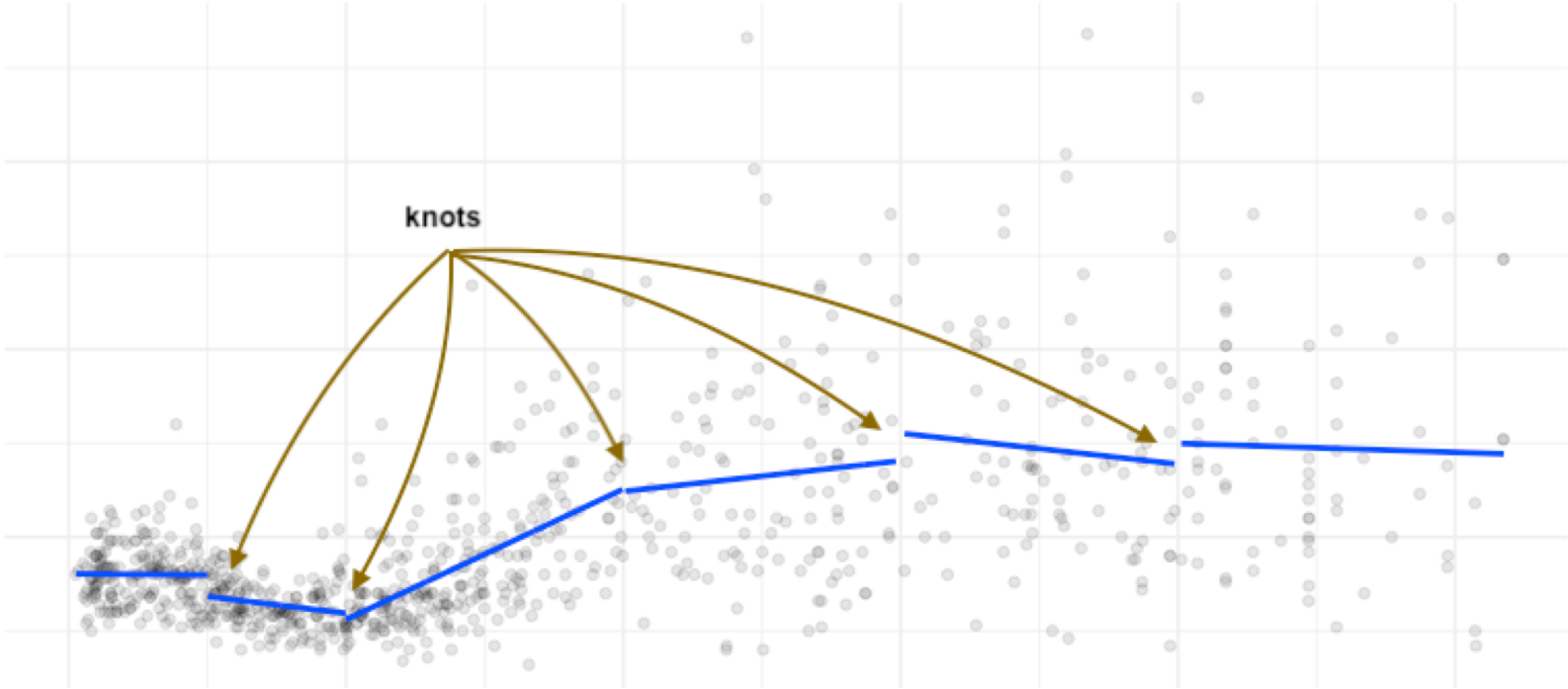
- Linear regression model



Regression splines (2)

Piecewise regression

- Example: policy changes over time on X-axis (knots)
- Regression discontinuity designs



Conclusion

- Regression can be used for inference:
  - Using a sample to understand predictors of Y
    - Which predictors are important
    - Which are not
  - Can also be used to predict actual values in Y
    - What is predicted house value (y) given values of X
    - If someone would extend house floor space 20 m^2, what would predicted Y be?
  - Underfitting is a problem, but overfitting too
    - In prediction models overfitting less of a problem (?)
- Regression is very flexible
  - It is used a lot in practice

Exercises

- 1. Introduction to JASP
- 2. Regression in JASP
- 3. Intro to R
- 4. Regression in R

Break (15 min)

Two cases

Thank you

- Questions?

- E-mail: p.lugtig@uu.nl

- G.vink@uu.nl
- P.g.m.vanderheijden@uu.nl