Statistical modeling in logistics

MMS075

Lecture 7b – Ethical analysis of big data, Brief overview of course content

Acknowledgement: Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Outline

- Ethical analysis of big data

- Brief overview of course content:
  - Model type selection
  - Variable selection
  - Relevant concepts
  - Relevant plots (examples)

- Next steps

Recommended resources

Reading in ISL: Chapters 2-5 (excluding 3.5, 4.4 and 5.2), Sections 6.1 and 7.1

The videos from the Statistical Learning course, available at this link.

For the discussion of ethical analysis of big data:

- Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. (2017). Ten simple rules for responsible big data research. PLoS Comput Biol 13(3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Ethical analysis of big data

How to do analysis in an ethical way?

- Recommendations are made in the following article:

Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399

- For a full understading, read the article. Here: some highlights are given, with some own edits and additions

- Ethical aspects addressed in Assignment 3, but not in the exam

Data are people and can do harm

- Start with the assumption that all data are people until proven otherwise

- Apparently neutral data can lead to discrimination: categorization based on zip codes resulted in less access to Amazon Prime same-day delivery service for African-Americans in United States cities*

* Ingold D, Spencer S. Amazon Doesn't Consider the Race of Its Customers. Should It? Bloomberg.com, 21 April 2016. http://www.bloomberg.com/graphics/2016-amazon-same-day/. Accessed 12 June 2016

Privacy is more than a binary value

- Breaches of privacy are key means by which big data research can do harm
- Privacy is contextual* and situational** (e.g. just because something has been shared publicly does not mean any subsequent use would be unproblematic)
- Marketing based on search patterns have been perceived by some to be “creepy” or even outright breaches of privacy

* Nissenbaum H. Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press; 2009.
** Marwick AE. boyd d. Networked privacy: How teenagers negotiate context in social media. New Media & Society. 2014:1461444814543995.

Guard against data re-identification

- It is problematic to assume that data cannot be re-identified.
- When datasets thought to be anonymized are combined with other variables, it may result in unexpected re-identification
- Birthdate, gender, and zip code can identify people today*, but even seemingly harmless factors like battery usage may aid personal identification tomorrow**

*Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002;10(05):557–70.
** Michalevsky Y, Schulman A, Veerapandian GA, Boneh D, Nakibly G. Powerspy: Location tracking using mobile device power analysis.
In 24th USENIX Security Symposium (USENIX Security 15) 2015 (pp. 785–800).

Context & multiple meanings

Context affects every stage:

- data acquisition
- data cleaning
- interpretation of findings
- dissemination of the results

Context and evolution of data needs to be documented

Multiple meanings and uses:

- interpretation of those (re)using your data may differ from your own
- consider & describe potential multiple meanings (e.g. for models with confounding!)
- use clear & high quality figures*

→ Do not overstate clarity, acknowledge multiple meanings and uses
→ Document strengths & weaknesses of data and analysis

* See e.g. the following article for recommendations: Rougier, N.P., Droetboom, M., Bourne, P.E. (2014). Ten Simple Rules for Better Figures. PLoS Comput Biol. 10(9): e1003833, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4161295/

Long-term strategy

Discuss tough, ethical choices

- Make grappling with ethical questions part of standard workflow
- Ethics is often about finding a good or better, but not perfect, answer

Make code of conduct

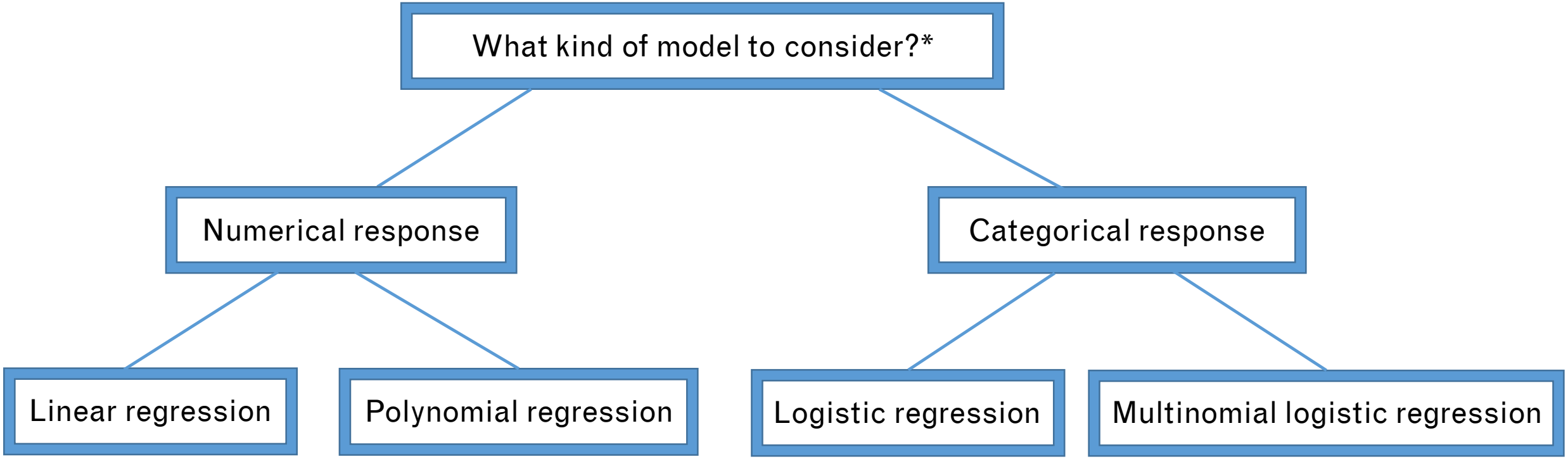
- Address issues that might be ignored until they blow up
- Make researchers and representatives of affected communities active contributors

Auditability

- Plan for & welcome audits of big data practices
- Explicit about decisions → understandability and replicability

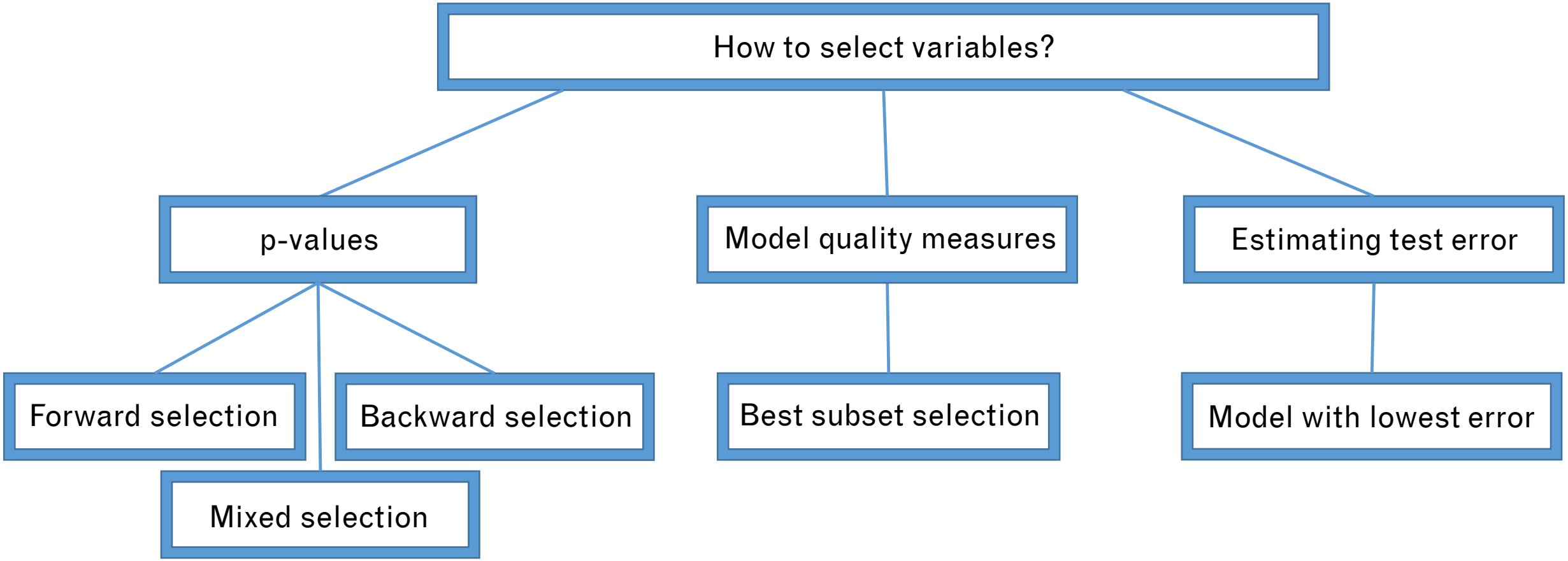
Brief overview of course content

Model type selection



* There are way more models than those shown on this slide; see ISL for many examples. This chart only shows models covered in MMS075.

Variable selection



Relevant concepts (1)

- Linear regression terminology (e.g. Least squares line, Slope, Intercept, Residuals, RSS, R^2)
- Interpretation of coefficients in regression models with one or multiple predictors, practical implications, making predictions
- Quantifying uncertainty...
  - ... of coefficients: Confidence intervals for coefficients
  - ... of predicted values: Confidence intervals, Prediction intervals
- Hypothesis testing of relationship between predictors and response:
  - Variable significance & link with confidence interval
  - Significance in one-variable and multivariable models
  - Model significance

Relevant concepts (2)

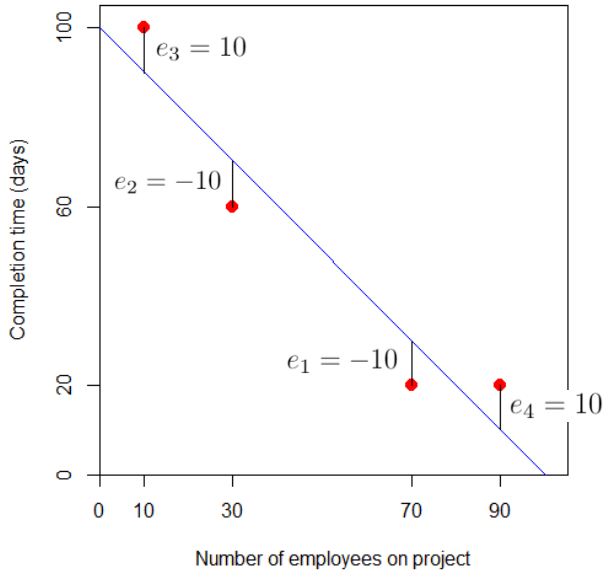
- Qualitative predictors – Dummy variables, how to define these for categorical variables with $L \geq 2$ levels
- Interaction terms – Interpretation, Hierarchical principle
- Polynomial regression: including higher degrees of predictors
- Model diagnostics:
  - Residual plots, Standardized/Studentized residuals
  - Outliers, High leverage points, Influential points
  - Non-linear relationships
  - Correlated error terms
  - Heteroscedasticity (non-constant variance of error term)
  - Collinearity, variable inflation factor

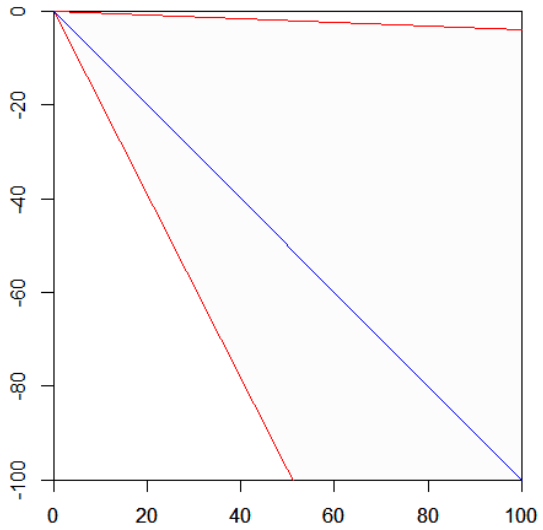
Relevant concepts (3)

- Odds, Log-odds (Logit), Interpretation of logistic regression coefficients in terms of log-odds, odds and probability
- Confounding: different effect of a predictor in one-variable and multivariable models
- Plots for regression models (examples on next slides)
- Training error and test error in regression and classification, Overfitting
- Estimating test error: Validation set, K-fold Cross-Validation, LOOCV

Geometrical interpretation of coefficients

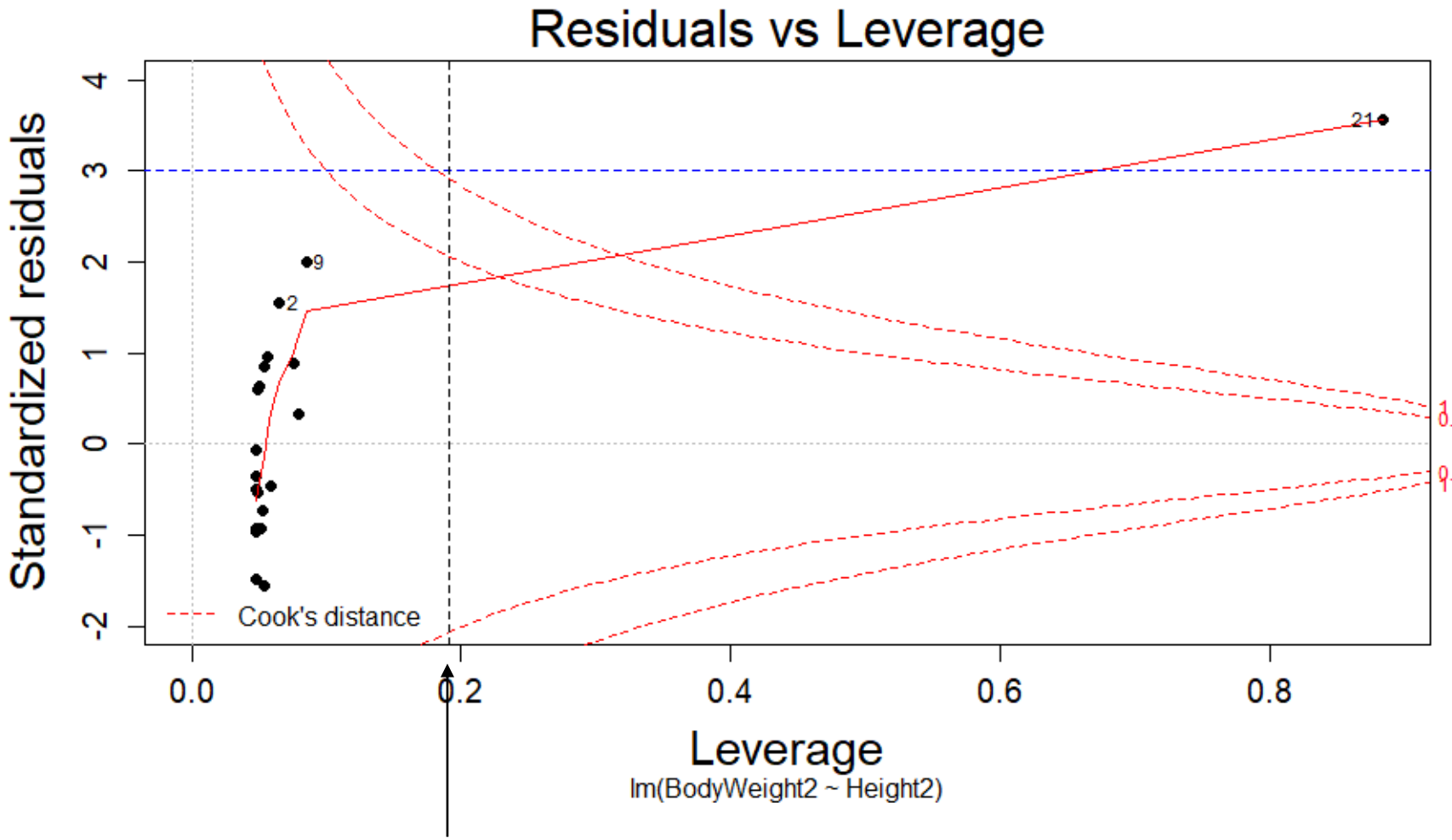
- Intercept: the value where the regression line crosses the y-axis
- Slope: average increase in Y associated with a one-unit increase in X
- Confidence interval boundaries for slope are shown below





The blue line has slope -1, like the least squares line; the red lines have slopes corresponding to the endpoints of the confidence intervals for the slope, i.e. -1±0.96.

Points with unusual and/or influential values



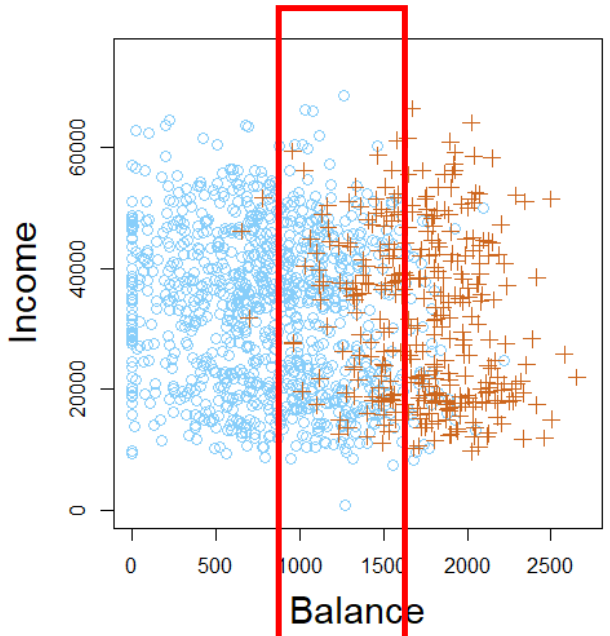
Points above this line* or below -3 are **outliers**

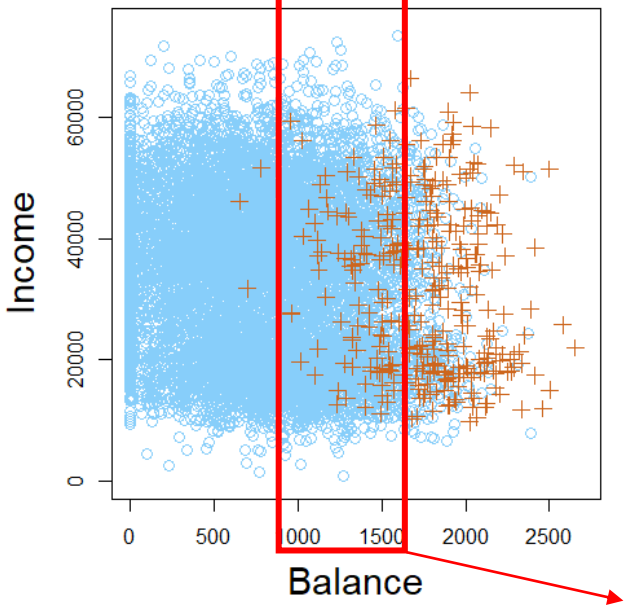
Points in the marked corners are **influential** (if corner is not shown in R → no influential points)

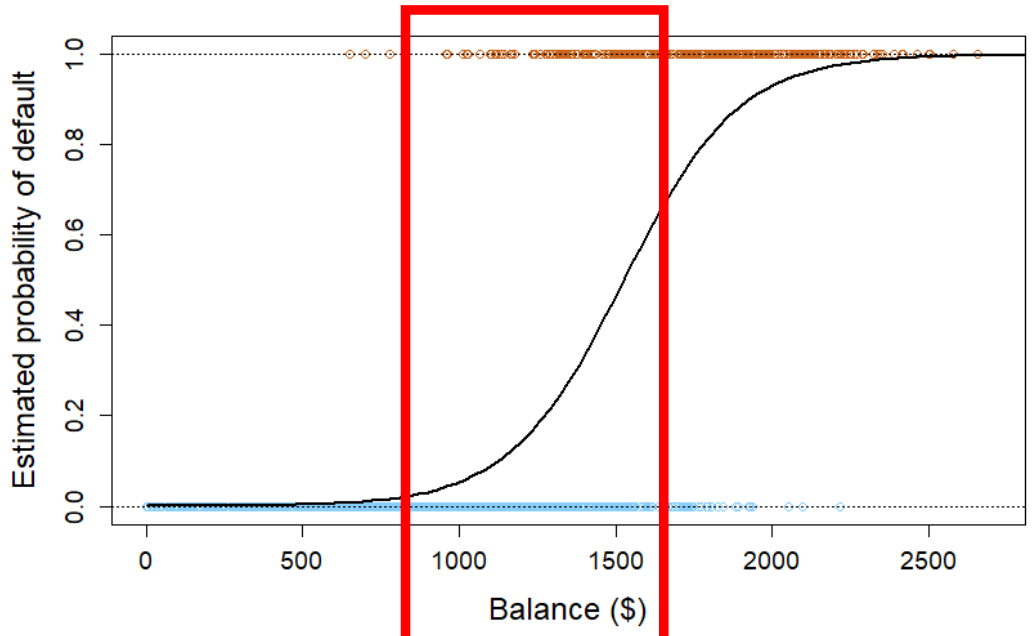
Points to the right of this line* have **high leverage**

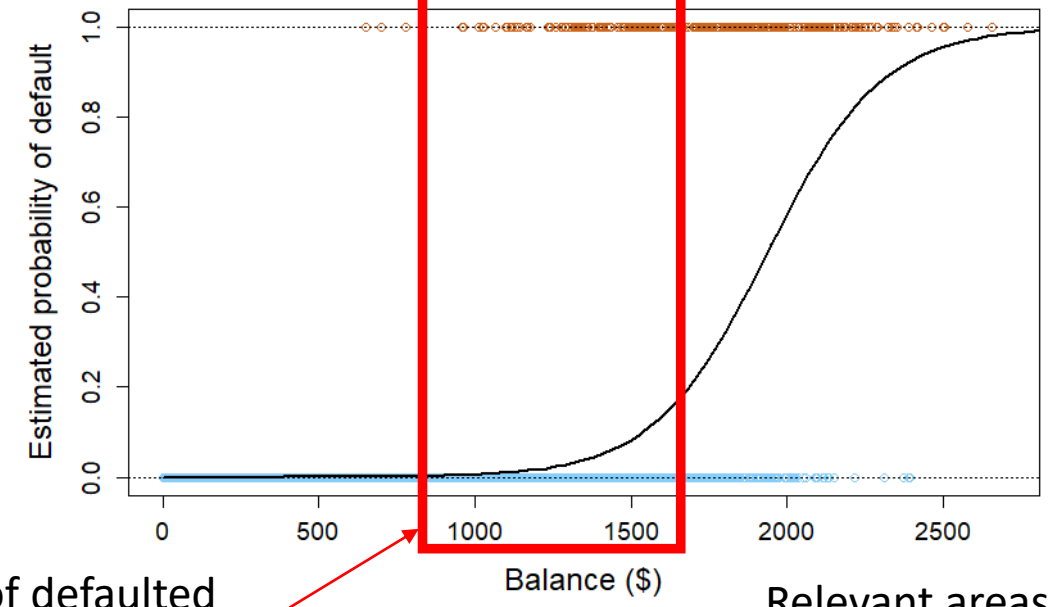
*The blue and black lines can be added manually to the R plot

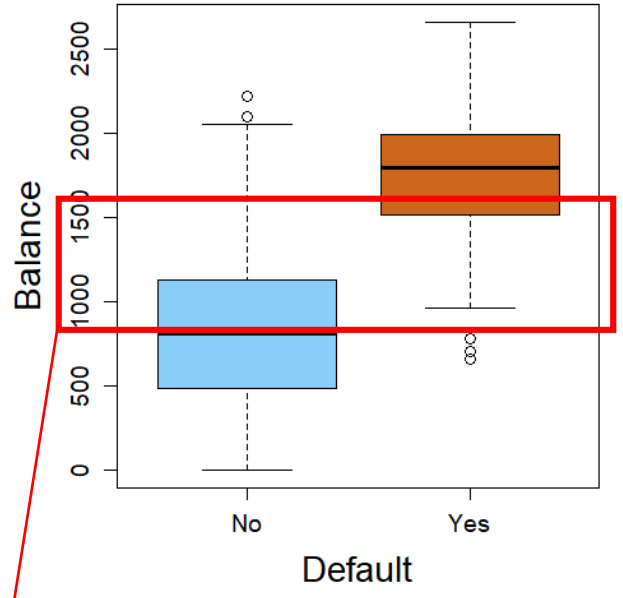
Plots relevant to logistic regression

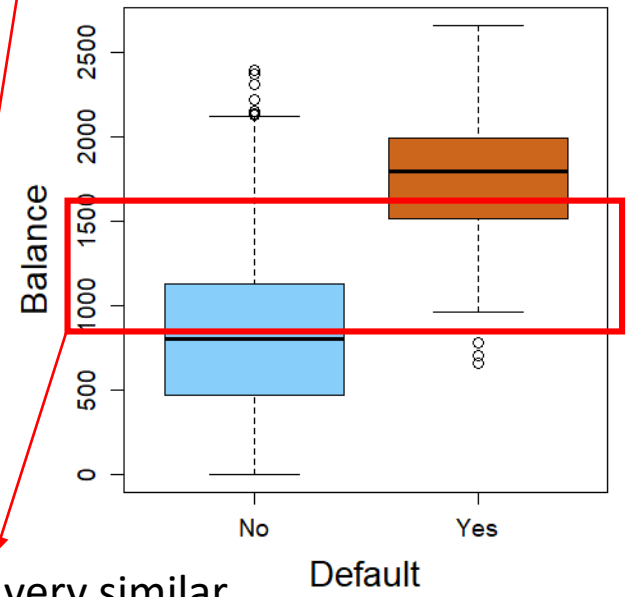








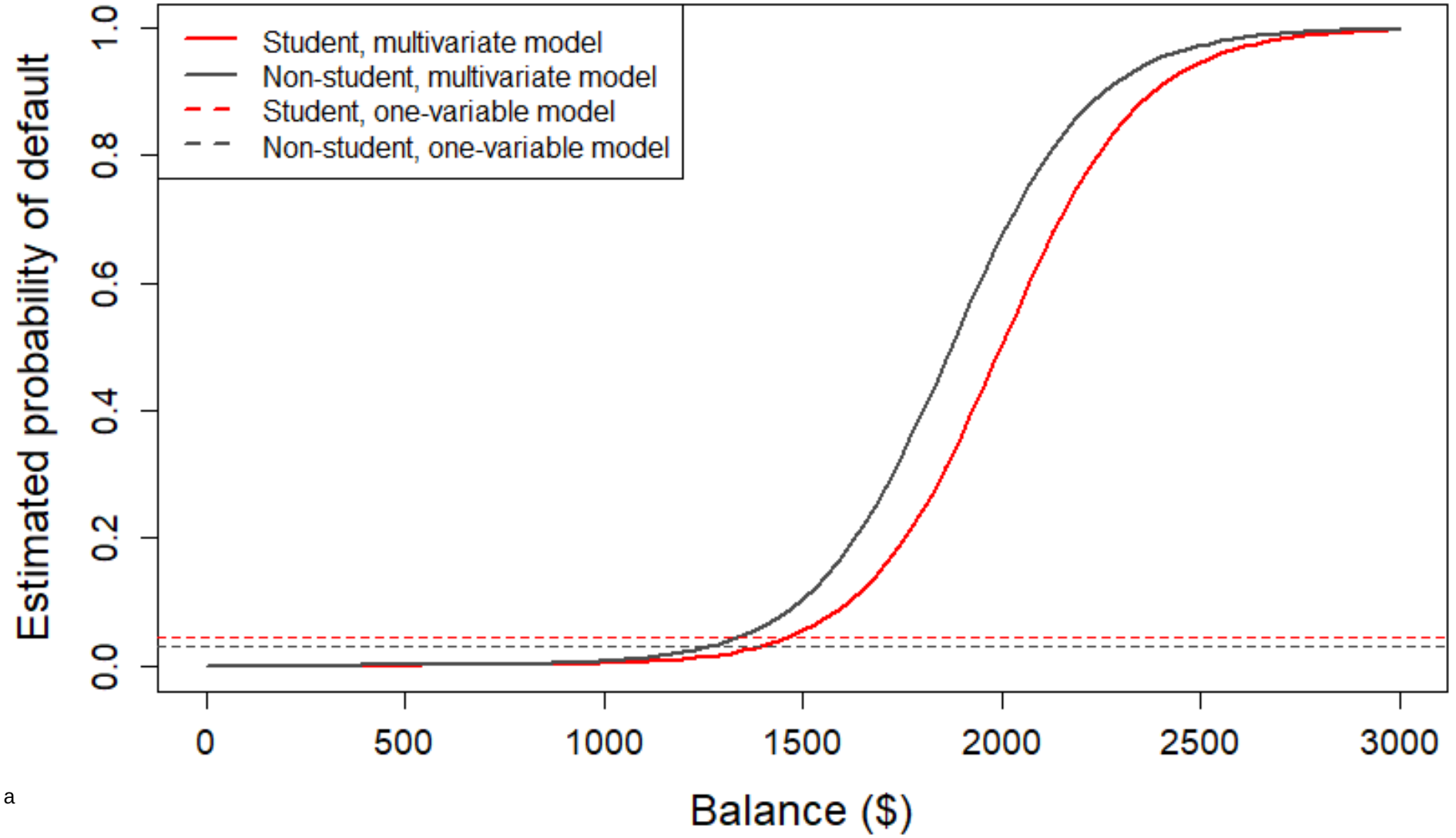




% of defaulted indicates probability

Relevant areas look very similar
→ box plots cannot indicate difference

Confounding example



Next steps

Assessment in MMS075

The description of assessment on the Student Portal:

Examination including compulsory elements

One or more project tasks (part A). Written examination (part B). The final grade is determined by the grade on the written exam.

Credit distribution

Module			Sp1	Sp2	Sp3	Sp4	Summer course	No Sp	Examination dates		
0119	Written and oral assignments, part A	5,0c Grading: UG		5,0c							
0219	Examination, part B	2,5c Grading: TH		2,5c					16 Mar 2020 am L,	10 Jun 2020 pm L,	19 Aug 2020 am L

Link between assessment and intended learning outcomes

Learning outcomes (After the course, students should be able to...)

A+E • Demonstrate an understanding of the key concepts and ideas in statistical modeling on larger datasets;

E • Describe suitable statistical methods for using on larger datasets relevant in logistics;

A+E • Choose and use appropriate statistical methods for answering a logistics related problem, and report the findings in a suitable and compelling format;

A+E • Critically evaluate statistical materials and methods and reason about their limitations;

A • Reflect on ethical aspects and considerations when collecting and analyzing larger datasets.

A: assignments, E: exam

Will the exam include ...

- **Computations** by hand or using a calculator?

No long computations.

Short computations (of 1-2 steps) are possible, formulas will be provided.

- **Choosing models** and **interpreting coefficients**?

Yes. You need to understand when the different models can be used and how the model parameters can be interpreted.

- Interpretation of **outputs from R**?

Yes. It is essential to understand summaries and plots provided by statistical software like R and find the relevant information in such outputs.

- Specification or interpretation of **commands from R**?

No. The course is about statistical modelling and not about a specific software.

Where to go from here?

- Apply knowledge in your work or further studies – potential feedback based on later experiences is greatly appreciated!
- There is a lot more to learn – see e.g. ISL or ESL, videos from the Statistical Learning course at this link
- Please fill course evaluation survey once it is sent out!
  - Your feedback is very important to me and will help future students