Data Science for Socioeconomists

Exam 1 (19.02.2024)

Ulrich Fritsche, Lisa Wegner, Victoria Hünewaldt

General remarks

- The exam consists of three parts: one multiple-choice and two essay/programming sections.
- All parts must be completed in full.
- A total of 90 points can be achieved.
- The total time of the exam is 120 minutes, the exam period is from **19.02.2024, 15:00 to 19:00.**
- Resources: RStudio via the JupyterHub (or local R/Rstudio installation), lecture notes, textbooks and the exercise materials. Usage of generative AI is allowed but has to be documented (see below).
- When you have answered all the questions, download both the Quarto document(s) and the HTML files, compress your solution(s) into a ZIP file and upload it to OpenOLAT within the deadline, together with the Declaration of Authorship you have completed. If you used generative AI add the respective declaration also to the ZIP archive and document the usage.
- Follow the guidelines in the respective tabs.
- The declaration of authorship can be found in the OpenOLAT exam room under the respective tab.
- A template for the Quarto document can be found in OpenOLAT under the respective tab.
- Help on creating a ZIP file can be found under the relevant tab.
- Instructions on how to document generative AI content can be found on the respective tab.
- Instructions on how to up- and download the Quarto file and how to create and download the respective HTML files can be found under the tab “Guide exam upload”
- You can ask comprehension questions during the exam by emailing Victoria Hünewaldt (victoria.huenewaldt@uni-hamburg.de).
- If there is any ambiguous wording in the exam that has not yet been noticed by the teaching staff, you will be notified by OLAT message.

- You are responsible for checking your OLAT messages yourself.
- If you encounter serious technical problems during the processing period that prevent you from completing the exam, you must contact the Student Administration Office immediately and provide evidence of the relevant restrictions.

We wish you a successful take-home exam!

Section 1: Multiple choice

> ! Important
>
> Please note, that multiple (max. 2) or single answers are possible.

Question 1

The challenge of the `Big Data revolution` can be described with several V's. Which term is **not** part of the challenge?

a) Volume of data
b) Veracity of data
c) Visibility of data
d) Velocity of data

Question 2

`R` has has several atomic classes of objects. Which of the the following is **not** part of the atomic classes?

a) character
b) integer
c) logical
d) factor

Question 3

You will get the code below. For which principle is this a concrete example?

```r
x <- 0:6
y <- as.numeric(x)
```

a) Function
b) Implicit coercion
c) Explicit coercion
d) Implicit coding

Question 4

Control structures in R allow you to control the flow of execution of a series of R expressions. Which of the following are examples of control structures?

a) `if` and `else`
b) `to`
c) `which`
d) `for`

Question 5

Which of the following statements is/are **not** true?

a) Lists are special types of vectors which might contain elements of different classes.
b) Lists are special types of data frames.
c) Factors are used to represent categorical data and can be unordered or ordered.
d) Unlike matrices, data frames can store different classes of objects in each column.

Question 6

Which of the following statements is/are **false**?

a) Missing values are denoted by zero.
b) NA is short for Not Available.
c) NaN is short for Not a Number.
d) NA is short for No Answer.

Question 7

What makes bad figures bad? Which of the following statements is/are true?

a) Aesthetic problems
b) Perceptual problems
c) Programming problems
d) Big data problems

Question 8

What does the `gg` in the package `ggplot2` stand for?

a) generation of graphics
b) gestalt rules of graphics
c) grammar of graphics
d) generalization of graphics

Question 9

Inference and prediction are two reasons why we might be interested in statistical learning models. Which of the following examples does **not** fit into the inference reasoning?

a) Which advertising expenses are associated with sales?
b) Which advertising expenses generate the biggest boost in sales?
c) Is this house under- or over-valued given all known fundamentals?
d) How much extra will a house be worth if it has a view of the river?

Question 10

Which of the following statements is/are true?

a) Test MSE (average squared prediction error) is U-shaped in flexibility.
b) High prediction accuracy often comes with high interpretability
c) High prediction accuracy often comes at the cost of low interpretability.
d) There is no difference between training and test MSE in the relationship to flexibility.

Question 11

The linear regression model is typically estimated using ordninary least squares. Which of the following equations are valid for this model?

a) $RSS = \sum_{i=1}^{n} |y_i - \hat{y}_i|$
b) $RSS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$
c) $RSS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^3$
d) $RSS = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$

Question 12

The standard errors of the parameters in the simple linear regression model are given by the formulas below. Which of the following statements is/are correct?

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where σ^2 stands for $Var(\varepsilon)$.

a) The variance of the estimator decreases in the number of observations n.
b) The variance of the estimator increases in the variance of x.
c) The variance of the estimator increases in the residual's variance σ^2.
d) The less information we have the more precise can the respective coefficient be estimated.

Question 13

Which of the following formulas does **not** give an expression of the "goodness-of-fit" measure R^2?

a) $R^2 = \frac{Var(\hat{y}_i)}{Var(y_i)}$
b) $R^2 = 1 - \frac{Var(\varepsilon_i)}{Var(y_i)}$
c) $R^2 = \frac{Var(y_i)}{Var(\hat{y}_i)}$
d) $R^2 = \frac{Var(\varepsilon_i)}{Var(y_i)}$

Question 14

The logistic model can be described by which set of formulas?

a) $z_i = \beta_0 + \beta_1 x_i$, $p(y_i = 1) = f(z_i) = \frac{1}{1+e^{-z_i}}$
b) $z_i = \beta_0 + \beta_1 x_i$, $p(y_i = 1) = f(z_i) = \frac{1}{(1+e^{-z_i})^2}$
c) $z_i = \beta_0 + \beta_1 x_i$, $p(y_i = 1) = f(z_i) = \frac{z_i}{1+e^{-z_i}}$
d) $z_i = \beta_0 + \beta_1 x_i$, $p(y_i = 1) = f(z_i) = \frac{1}{1-e^{-z_i}}$

Question 15

Which statement(s) is/are true?

a) As the flexibility of the statistical learning method increases, we observe a monotone increase in the training MSE and an inverse U-shape in the test MSE.
b) As the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE and an inverse U-shape in the test MSE.
c) As the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE and a U-shape in the test MSE.
d) As the flexibility of the statistical learning method increases, we observe a monotone decrease in the test MSE and a U-shape in the training MSE.

Section 2: Transfer

We've got a dataset (`income_data`) that includes data about people's monthly income measured at two points in 2023 (March and November), about their age and their gender. The dataset was created by our colleague that left the team – now we have trouble understanding the data since he left no description and stopped working. His last task was to visualize the income in regard to age. The variables look like this: Income03 for March, Income11 for November, a_2023, gender

Please use this code to load the package and the dataset `income_data`.

```r
#install.packages("remotes")
#library(remotes)
remotes::install_gitlab("BAQ6370/sozoekdsexam", host="gitlab.rrz.uni-hamburg.de",
force = TRUE)
library(sozoekdsexam)
income_data <- income_data
```

He left us with this code which produces a plot:

```r
income_data <- income_data
library(ggplot2)

p <- ggplot() +
  geom_point(data = income_data, aes(x = a_2023, y = income03)) +
  geom_point(data = income_data, aes(x = a_2023, y = income11))

print(p)
```

Your tasks:

1. How can we improve the graphic? Name 3 aspects that can be changed and justify your decision (Text)
2. Re-work the code of our colleague so that we get a graphic that visualizes income and age. Think of a way of creating income or age groups for better readability. What do the results tell us about our data? (Coding & Text)
3. How can we enhance the interpretability of the data set? Write code for three of the variables. (Coding)

As always: comment on your code and justify your decisions.

Section 3: Statistical learning exercises

For the following exercise, you have to install the package `sozoekdsexam`. Please use either library `remotes` or library `devtools` (see examples below).

```
#Either:
install.packages("remotes")
library(remotes)
remotes::install_gitlab("BAQ6370/sozoekdsexam", host="gitlab.rrz.uni-hamburg.de",
force = TRUE)

#or:

install.packages("devtools")
library(devtools)
devtools::install_git("https://gitlab.rrz.uni-hamburg.de/baq6370/sozoekdsexam.git",
force = TRUE)
```

> ! Important
>
> Please use the data set `examscores` included in the library `sozoekdsexam` in all exercises of this section. Use Quarto to code your exercise. Don't forget to explain/comment your code lines!

1. Logistic Regression

Formulate the following model in R and run a logistic regression:

$high\_mathscore = f(ParentMaritalStatus, Gender)$ or in the formula syntax of R: **high_mathscore ~ ParentMaritalStatus + Gender**

> ℹ Note
>
> To estimate the model you first need to construct "**high_mathscore**" which is a binary variable being 1 if MathScore>67 and 0 elsewise. Please drop the original variable `MathScore`from the data set for further exercises.

1. After estimating the model, interpret your results in terms of magnitude and significance. Explain the steps and motivate, why you have to take them.
2. Split your data into test and training data. Evaluate the in- & out-of-sample performance of the model using a confusion matrix and explain the metrics you have chosen. Compare and discuss the results.

2. Lasso regression

1. For the next exercise, run a logistic lasso regression on the full model (using all possible exogenous variables excluding the variable `MathScore`) and applying cross-validation. Plot the result and interpret the plot. Can you already determine the appropriate number of variables?
2. Next, find the best λ (by minimizing the mean squared error) and find a better logistic regression model than the one given in exercise 1. Write down the equation including endogenous and exogenous variables. Explain, why the Lasso regression approach can be used for model selection.
3. Evaluate the in- & out-of-sample performance of this better model using a confusion matrix and explain the metrics you have chosen.
4. Now compare the results with the results from logistic regression in exercise 1. Did you find a better performing model by applying a Lasso regression? Motivate your answer!