Data Science for Socioeconomists

Exam 2 (20.03.2024)

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 **20.03.2024, 15:00 to 19:00.**
- Resources: RStudio via the JupyterHub (or local R/RStudio installation), lecture notes, textbooks and the exercise materials.
- 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 Independence you have completed.
- The declaration of independence can be found in the OpenOLAT exam room under the tab "Declaration form".
- A template for the quarto document can be found in OpenOLAT under the tab "Template".
- Help on creating a ZIP file can be found under the relevant tab.
- 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 Stine message.
- You are responsible for checking your Stine 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

Question 1

What is typically referred to as Moore's law?

a) The price of computer chips doubles every two years.
b) The number of transistors on a microchip doubles every two years.
c) The processor speed of computers doubles every two years.
d) The size of computer memory doubles every two years.

Answer: b) The number of transistors on a microchip doubles every two years. (Introduction, slide 17)

Question 2

A programming language consists of four basic elements. Which of the following elements are part of it?

a) Primitive constructs
b) Inline tags
c) Static semantics
d) Natural language programming

Answer: a) Primitive constructs, c) Static semantics (R Markdown, slide 7)

Question 3

The `c()` command concatenates vectors. Which of the following statements are true?

a) Within the vector all elements must be of the same class.
b) The concatenate command does not allow for implicit coercion
c) Integer vectors can be constructed in c() with a range definition.
d) Within the vector elements can be of different classes.

Answer: a) Within the vector all elements must be of the same class. c) Integer vectors can be constructed in c() with a range definition. (R Basics, slide 18)

Question 4

Which of the following commands do **not** allow for matrix combination?

a) rbind
b) nbind
c) cbind
d) mbind

Answer: b) nbind d) mbind (R Basics, slides 32/33)

Question 5

You will get the code below. Which of the statements is true?

```
f <- function(num) {
     for( i in seq_len(num)) {
       print(" Hello, world!")
     }
}
```

a) `seq_len` is a function and generates an integer vector of the form 1 up to the value of num
b) `num` is a parameter of the function f
c) `num` is a function
d) `num` is not used in the function f

Answer: a) `seq_len` is a function and generates an integer vector of the form 1 up to the value of num, `num` is a parameter of the function f (R Basics, slide 42)

Question 6

Which of the following statements is/are true?

a) We can address the variables in a dataframe by $variable\$dataframe$
b) We can address the variables in a dataframe by $dataframe\$variable$
c) We can address the variables in a dataframe by $dataframe : variable$
d) We can address the variables in a dataframe by $dataframe :: variable$

Answer: b) We can address the variables in a dataframe by <dataframe$variable> (Data and Basic transformations, slide 17)

Question 7

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

a) The pipe command is a representation of an object-oriented workflow.
b) %>% indicates the usage of a pipe in R.
c) |> indicates the usage of a pipe in R.
d) The pipe command is a representation of a functional workflow.

Answer: d) The pipe command is a representation of a functional workflow. (Data and Basic transformations, slide 29)

Question 8

The `Gestalt` rules are important factors when it comes to perception. Which of the following elements is/are **not** part of the `Gestalt` rules?

a) Proximity: Things that are spatially near to one another seem to be related.
b) Closure: Incomplete shapes are perceived as complete.
c) Discontinuity: Partially hidden objects are never completed into familiar shapes.
d) Connection: Things that are visually tied to one another seem to be related.

Answer: c) Discontinuity: Partially hidden objects are never completed into familiar shapes. (Visualisation, slide 15)

Question 9

The accuracy of $\hat{y}$ as a prediction for y depends on two quantities/elements as shown in the formula below. Which of the following the terms does/do indicate **neither element A nor B**?

$$\mathbb{E}(y-\hat{y})^2 = \mathbb{E}\left[f(x)+\varepsilon-\hat{f}(x)\right]^2 = \underbrace{\mathbb{E}\left[f(x)-\hat{f}(x)\right]^2}_{\text{A}} + \underbrace{var(\varepsilon)}_{\text{B}}$$

a) Reducible error
b) Restricted error
c) Irreducible error
d) Irrelevant error

Answer: b) Restricted error, d) Irrelevant error (Foundations, slide 7)

Question 10

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

a) Cluster analysis is an example for supervised learning.
b) For each observation of the predictor measurement(s) $x_i, i = 1, ..., n$ there is an associated response measurement y_i.
c) Linear regression is an example for unsupervised learning.
d) Unsupervised learning describes the situation in which for every observation $i = 1, ..., n$, we observe a vector of measurements x_i but no associated response y_i

Answer: a) Cluster analysis is an example for supervised learning. c) Linear regression is an example for unsupervised learning. (Foundations, slides 19)

Question 11

The expected test MSE for a given training data set can be formally decomposed into three elements as shown in the equation below. Which of the following elements is **not** part of the decomposition?

$$\mathbb{E}\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}\left(\hat{f}(x_0)\right) + \left[\text{Bias}(\hat{f}(x_0))\right]^2 + \text{Var}(\varepsilon)$$

a) The variance of $\hat{f}$.
b) The squared bias of $\hat{f}(x_0)$
c) The variance of x_0
d) The variance of the error.

Answer: c) The variance of x_0 (Foundations, slide 34)

Question 12

For the estimated linear regression model, the analytical solution can be stated as below. Which of the following statements is/are **false**?

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

a) The estimated coefficient $\hat{\beta}_1$ can be calculated as ratio of $Cov(x, y)$ and $Var(x)$.
b) The estimated coefficient $\hat{\beta}_1$ can be calculated as ratio of $Cov(x, y)$ and $\sqrt{Var(x)}$.

c) The estimated coefficient $\hat{\beta}_0$ is a function of $\bar{y}$ and $\bar{x}$.
d) The estimated coefficient $\hat{\beta}_0$ does not depend on $\hat{\beta}_1$.

Answer: b) The estimated coefficient $\hat{\beta}_1$ can be calculated as ratio of $Cov(x, y)$ and $\sqrt{Var(x)}$, d) The estimated coefficient $\hat{\beta}_0$ does not depend on $\hat{\beta}_1$. (Linear regression, slide 11)

Question 13

Which of the following principles guide our selection of estimation methods?

a) complexity
b) unbiasedness
c) efficiency
d) simplicity

Answer: b) unbiasedness, c) efficiency (Linear regression, slides 19/20)

Question 14

Which of the following is true for the linear probability model?

a) Probability can be lower than zero and higher than one.
b) The model is linear in variables but non-linear in parameters.
c) Residuals are normally distributed.
d) We get a "strange" distribution of residuals: bi-modal, not continuous.

Answer: a) Probability can be lower than zero and higher than one. d) We get a "strange" distribution of residuals: bi-modal, not continuous. (Linear regression, slide 40)

Question 15

Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event. Which formula describes Bayes' Theorem?

a) $Pr(A|B) = \frac{Pr(B|A)*Pr(A)}{Pr(B)}$
b) $Pr(A|B) = \frac{Pr(B|A)*Pr(B)}{Pr(A)}$
c) $Pr(A|B) = \frac{Pr(A|B)*Pr(A)}{Pr(B)}$
d) $Pr(A|B) = \frac{Pr(B|A)*Pr(A)}{Pr(A)}$

Answer: a) $Pr(A|B) = \frac{Pr(B|A)*Pr(A)}{Pr(B)}$ (Classifications, slide 15)

Section 2: Transfer

Please install and activate the data package `sozoekdsexam` by using the following code (make sure that library `devtools` is installed. In case it is not installed, comment out the respective lines in the code.)

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

The task: We work at Airbnb HQ and are the first people to work with a new dataset. Our main interest is the cleaning fee. Our idea is to show people creating new listings how to set their cleaning fee. Since our boss likes visuals, we created a chart - but were criticized and told that the chart needed to be easier to read.

Remark: In case `ggplot2` is not installed before activating, install it with `install.packages()` and activate thereafter.

```
library(sozoekdsexam)
library(ggplot2)

airbnbsmall <- airbnbsmall

ggplot(data = airbnbsmall, aes(x = usd_cleaning_fee)) +
  geom_point(aes(y=n_accommodates))
```

1. What is wrong with the plot? Name 3 things that could be changed in terms of readability and aesthetics, and justify your decision. (Please use complete sentences!)
2. Think about a different type of plot that would help us to look at the range of fees for each number of guests. Could it help to create groups for better readability? What do the results tell us about our data? (Please include text and code!)
3. Can we give a valid recommendation for

```
- the cleaning fee for a flat that accommodates 3 to 5 people?
- the cleaning fee for a flat that accommodates 9 to 11 people?
```

regarding the cleaning fee based on our plot? (Please use complete sentences!)

4. What other factors might affect the cleaning fee - argue for one of the variables from the data set and plot it along with the cleaning fee. Interpret the result. Note: You do not need to do a statistical analysis here to determine the best fitting variable, just choose one logically and explain your choice. (Please include text and code!)

Solutions to section 2

1. What is bad about the plot?

- axis labels
- no units are reported
- the title is missing
- chosen plot makes it hard to interpret concrete values in regard to the height of fees.
- no colors are used, everything looks the same, the eye has no visual guidance

2. Type of plot: Boxplot is recommended since we are interested in the range of fees. We should also create groups for different types of apartments, this creates a less cluttered graphic and helps with the interpretation.

```
# create categories for number of accommodates

airbnbsmall$accommodates_type <- cut(
  airbnbsmall$n_accommodates,
  breaks = seq(0, 16, by = 2),
  include.lowest = TRUE,
  labels = c(
    "0-2 persons:
    Couple",
    "3-5 persons:
    Small Family",
    "6-8 persons:
    Large Family",
    "9-11 persons:
    Group of Friends",
    "12-14 persons:
    Large Group of Friends",
    "15-17 persons:
    School Trip"
  )[1:8]  # Select the first 8 labels
)
```

```
# create boxplot for the groups
# scale the x-axis for better readability of concrete values
ggplot(airbnbsmall, aes(x=usd_cleaning_fee, y=accommodates_type,
                        fill=accommodates_type)) +
  geom_boxplot() +
  scale_x_continuous(breaks = seq(0, max(airbnbsmall$usd_cleaning_fee), by = 25))+
  labs(x = "Cleaning fee in $", y = "Type",
       title = "Cleaning fees in regard to the size of the accomodation") +
    guides(fill = "none")
```

Interpretation:

- The median value for the cleaning fee rises with a larger number of guests
- The biggest differences in cleaning fees can be seen for the group "Large Group of Friends". The interquartile range is the largest of all the groups.
- For the group "Couple" the fee has the smallest range, going from around 13$ to around 28$.
- The groups help to see a clear trend and indicate a positive relationship between the fee and the amount of guests

3. Our recommendation:

- The cleaning fee for a **flat that accommodates 3-5 people** should be around 36$.
  - It is not recommended to raise this fee higher than 50$ since this would mean that this service is pricier in your flat than in 75% of other accommodations.
  - Most people are used to paying at least 25$ for the cleaning, only 25% of the flats take a lower rate.
  - If your flat has very special features you could increase your fees up to 312$. This fee is the highest that another landlord is asking for.
- The cleaning fee for a **flat that accommodates 9 to 11 people** should be around 55 $which is the median value of fees for this group.
  - It is not recommended to raise the fee higher than 87$ since this would mean that this service is pricier in your flat than in 75% of other accommodations.
  - Most people are used to paying at least 37$ for the cleaning this kind of flat, only 25% of flats take a lower fee.
  - If your flat has very special features you could increase your fees up to 400$. This fee is the highest that another landlord is asking for.

4. Other factors that might influence the cleaning fee:
   - n_bathrooms: more bathrooms lead to more showers that need to be cleaned

- n_review_scores_rating: with a better rating one could expect a more polished flat that is pricier to clean

```
# create boxplot for the groups
# scale the x-axis for better readability of concrete values
airbnbsmall$n_bathrooms <- as.factor(airbnbsmall$n_bathrooms)
ggplot(airbnbsmall, aes(x=usd_cleaning_fee, y=n_bathrooms,
                        fill=n_bathrooms)) +
  geom_boxplot() +
    scale_x_continuous(breaks = seq(0, max(airbnbsmall$usd_cleaning_fee), by = 25))+
  labs(x = "Cleaning fee in $", y = "Number of Bathrooms",
       title = "Cleaning fees in regard to the number of bathrooms") +
    guides(fill = "none")
```

Interpretation:

- The plot helps to see a clear trend and indicates a positive relationship between the fee and the number of bathrooms.
- For flats with fewer bathrooms the fees have a smaller interquartile range and a smaller range overall.
- While the fee seems to increase slowly and relatively steady per each additional bathroom the median fee drops for 5 bathrooms and is similar to the one for 4 bathrooms while the median fee for 4.5 bathrooms is higher than both.
- The fee for flats with 6 bathrooms shows the highest median value at around 262$. 75% of the flats in this category take lower rates than 317$.

Section 3: Statistical learning exercises