EXAMINATION QUESTION PAPER - Written examination

GRA 60208
Applied Data Analytics

Department of Economics

Start date:	14.12.2018	Time 09.00
Finish date:	14.12.2018	Time 12.00

Weight:	60% of GRA 6020
Total no. of pages:	19 incl. front page
Answer sheets:	Squares, and lines
Examination support materials permitted:	All printed and handwritten support materials. BI-approved exam calculator. Simple calculator. Bilingual dictionary.

School Exam GRA6020 – Autumn 2018

The exam lasts three hours. Make sure that you do not spend too much time on a single point if you get stuck!

The first part of the exam consists of five multiple choice questions. There is just one correct answer for each question. This first part counts for 30% of the total points on the exam. For each multiple-choice question, you either get a full score or no points, and you will *not* be penalized with negative points for wrong answers.

The second part (Part 2 and Part 3) of the exam is a standard written exam and counts for 70% of the total points on the exam. Within each part, each task/sub-task carries equal marks, for example, Exercise 7 has the same weight as every sub-task in Exercise 8.

PART 1: Multiple choice questions (Counts 30%)

Please write your answer one line at the time, using clear and distinct letters. If it is impossible to make out your answer, the answer will be ignored. Any additional notes will also be ignored. You are only to write the letter of the answer. Thus, an example answer on all the multiple-choice questions would be:

1) A
2) A
3) A
4) A
5) A

where the candidate answered "A" on all the questions.

Exercise 1: A dataset has been collected to study the important relationship between U, V and W. The estimated linear regression model

$$U = \widehat{\beta}_0 + \widehat{\beta}_1 \times V + \widehat{\beta}_2 \times W + \widehat{\beta}_3 \times V \times W \tag{1}$$

is used for the analysis and the corresponding summary from R is shown below:

```
##
## Call:
## lm(formula = U ~ V + W + V:W)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1437 -0.8045 -0.1306  0.7250  2.6469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1162     0.1013   1.147    0.254
## V            -0.1277     0.1086  -1.176    0.243
## W            -0.1505     0.1038  -1.451    0.150
## V:W           0.1728     0.1149   1.503    0.136
##
## Residual standard error: 1.008 on 96 degrees of freedom
## Multiple R-squared:  0.05863,    Adjusted R-squared:  0.02921
## F-statistic: 1.993 on 3 and 96 DF,  p-value: 0.1202
```

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

A: All variables are of relative low statistical importance.

B: The p-value for the interaction term is less than 0.243.

C: The value of $\mathrm{R}^2 = 0.02921$.

D: The value of $\widehat{\beta}_0$ is equal to 0.1162.

Exercise 2: A dataset with the sales price (in US dollars) of houses has been collected; see the table below. The dataset contains 128 observations (sold houses) and include four numerical variables `Price`, `Bedrooms`, `Bathrooms` and `SqM`, where `Bedrooms` and `Bathrooms` are the number of such rooms and `SqM` is the size of the house in square meters. And two categorical variables `Brick` (`Yes`/`No`) and the type of neighborhood `Neighborhood` (`North`/`East`/`West`).

Price	Bedrooms	Bathrooms	Brick	Neighborhood	SqM
116200	3	2	No	East	166
139600	5	3	Yes	East	212
108500	3	2	No	North	198
⋮	⋮	⋮	⋮	⋮	⋮

The summary from estimating the linear regression model

$$\text{Price} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{Bathrooms} + \widehat{\beta}_2 \times \text{NeighborhoodNorth} + \widehat{\beta}_3 \times \text{NeighborhoodWest} + \widehat{\beta}_4 \times \text{SqM} \quad (2)$$

in R is shown below:

```
##
## Call:
## lm(formula = Price ~ Bathrooms + Neighborhood + SqM)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36339  -8906    927   9793  41405
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       30040.36   13300.16   2.259  0.02567 *
## Bathrooms          9613.53    3039.58   3.163  0.00197 **
## NeighborhoodNorth -9747.94    3195.14  -3.051  0.00280 **
## NeighborhoodWest  29563.61    3280.41   9.012 3.23e-15 ***
## SqM                 382.74      79.17   4.834 3.91e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14720 on 123 degrees of freedom
## Multiple R-squared:  0.7092, Adjusted R-squared:  0.6998
## F-statistic:    75 on 4 and 123 DF,  p-value: < 2.2e-16
```

Using the estimated model (2), what is the predicted price of a house of size 150 square meters, with one bathroom, in the so-called east neighborhood?

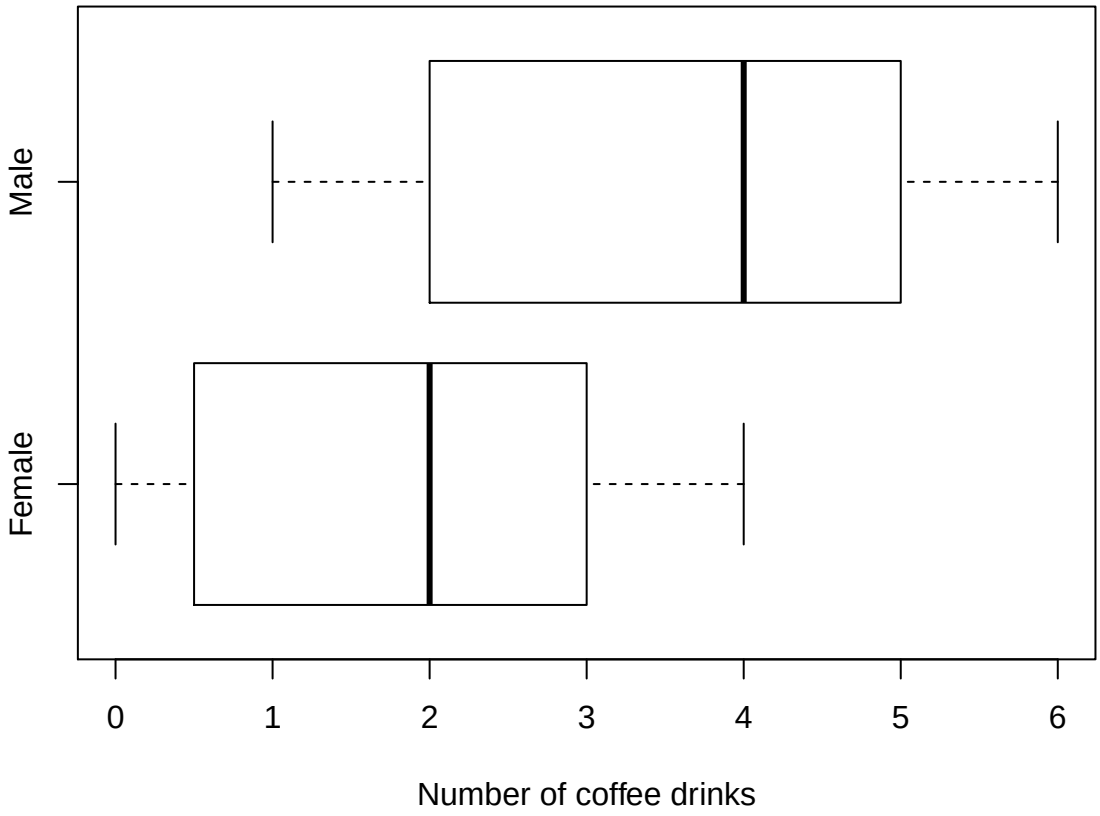
A: 39653.89

B: 97064.89

C: 87316.95

D: 126628.50

Exercise 3: In a study of daily caffeine consumption among students, the total number of coffee drinks (drinks containing caffeine) for one day was collected for each of 101 randomly chosen students (45 females and 56 males). The numbers are summarised in the box plot below:



Which of the following statements is true?

A: No women drink zero coffee drinks.

B: The number of women who drink less than two coffee drinks a day, is exactly the same as the number of those who drink two or more drinks a day.

C: The lowest number of coffee drinks for a man in this study is two.

D: At least half of the men does not drink more than four coffee drinks a day.

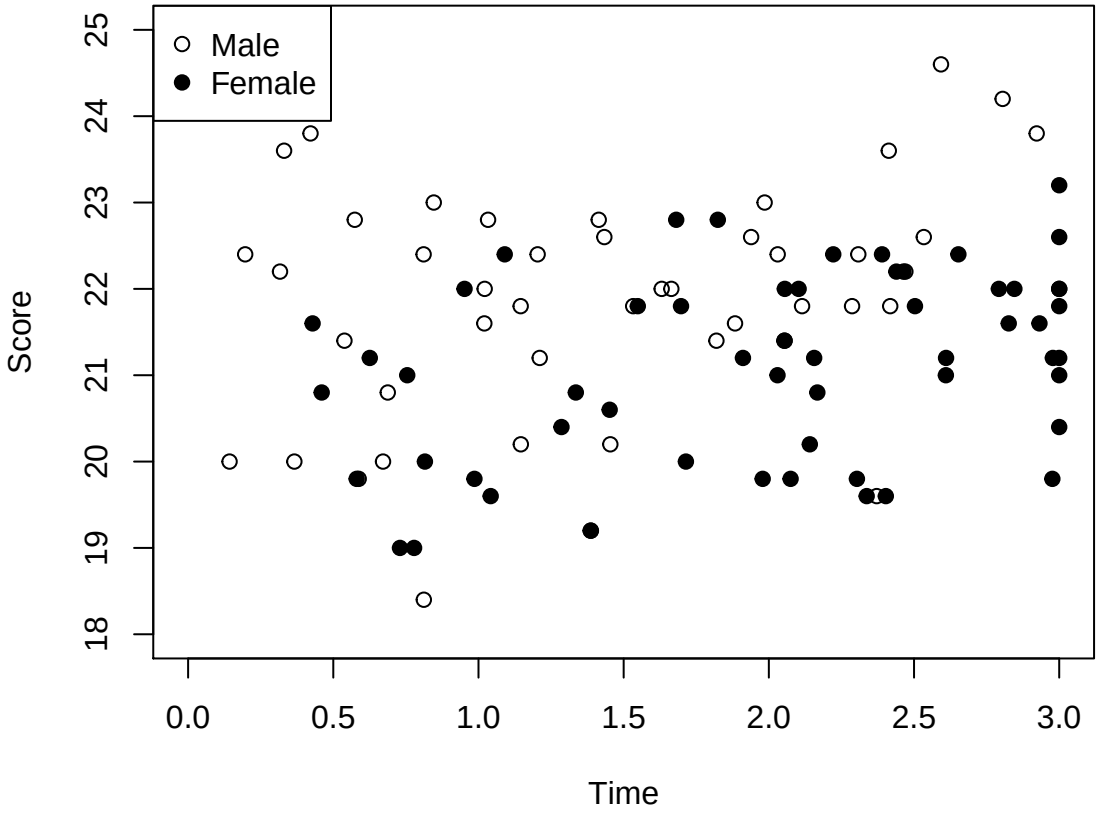
Exercise 4: Consider a study of the relationship between time used on a test and the final test score. The actual time used (`Time`), measured in hours, and the final test scores (`Score`) were collected for a random sample of 101 students. The total length of the test is 3 hours. Thus, it is not possible to spend more time on the test than this. The estimated linear regression model for investigating this relationship is given by:

$$\text{Score} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{Time} + \widehat{\beta}_2 \times \text{Female} + \widehat{\beta}_3 \times \text{Time} \times \text{Female}, \tag{3}$$

where `Female` is a dummy variable defined by

$$\text{Female}_i = \begin{cases} 0 & \text{if individual } i \text{ is a male} \\ 1 & \text{otherwise} \end{cases}.$$

The collected data is shown in the scatter plot below (note that some students got exactly the same score, but no students spent exactly the same amount of time and got exactly the same score):



Which of the following statements is true?

A: The highest score on the test belongs to a woman.

B: For a woman, the regression model (3) simplifies to $\text{Score} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{Time}$.

C: Most students using nearly all the time available were women.

D: From the scatter plot, we see that $\widehat{\beta}_0$ will be closer to 18 than 20.

Exercise 5: The estimated simple linear regression model

$$\text{Salary} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{Education} \tag{4}$$

is used with the goal of understanding the relationship between salary (`Salary`) and the number of years of higher education (`Education`). The collected dataset and the estimated model are shown in the figure below:

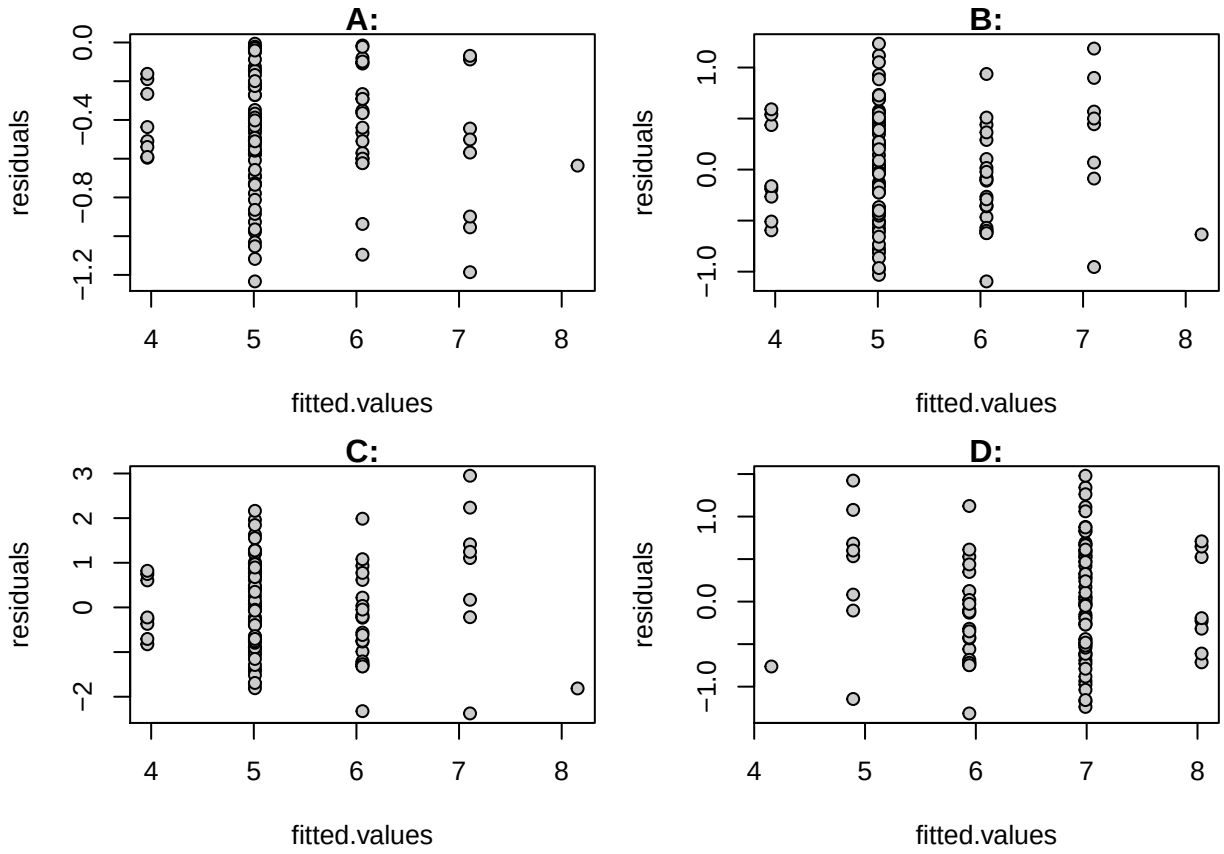
Which of the following residual plots (next page) belong to the estimated model (4) and data shown in the figure above?

A: Top left

B: Top right

C: Bottom left

D: Bottom right


A:
B:
C:
D:
residuals
fitted.values

PART 2: Mathematics (Counts 20%)

Exercise 6: Consider the following table of numbers:

Table 2: A collection of super important numbers.

i	x_i	y_i	$x_i y_i$	x_i^2
1	-1.10	-2.40	2.64	1.21
2	0.50	1.10	0.54	0.25
3	0.70	1.00	0.70	0.49
4	-0.10	0.30	-0.03	0.01

a) Use the numbers from Table 2 to compute the sample means of x_i and y_i, as given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

b) The least squares estimates for the coefficients in the straight line equation $y = \beta_0 + \beta_1 x$ are given by

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} \quad \text{and} \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Show that for the numbers in Table 2, we have that $\widehat{\beta}_0 = 0$ and that, based on the numerical values of $\bar{x}$ and $\bar{y}$, the algebraic expression for $\widehat{\beta}_1$ simplifies to (you have to give a algebraic justification of your answer):

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

c) Use the results from **b)**, and the numbers found in Table 2, to compute $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and draw the corresponding line.

PART 3: Data analysis (Counts 50%)

Exercise 7: After this year final US Open match, tennis player Serena Williams said that sexism is a problem that affects how rules are applied between male and female players. The claim is that women tend to be more penalised, i.e. receive fines more easily, than men when engaged in the same behaviour on court.

On debating that subject, the New York Times published a data table (see Table 3 below) of all fines from the Grand Slam tournaments for the past 20 years. The Grand Slam tournaments are the most important annual tennis events. The numbers in the table is based on all men and women **singles tournament**[1] matches, where the men play best-of-five sets and women best-of-three sets to win a match.

Table 3: Number of issued fines in men and women singles tournaments from the Grand Slam tournaments from 1998 to 2018 (it is possible to get more than one fine in a match).

	Men	Women
Racket Abuse	646	99
Audible Obscenity	344	140
Unsportsmanlike Conduct	287	67
Coaching	87	152
Ball Abuse	49	35
Verbal Abuse	62	16
Visible Obscenity	20	11
No Press	6	10
Time Violations	7	3
Best Effort	2	0
Default	2	0
Doubles Attire	2	1
Late for Match	1	1
First-Round Retirement	2	0
Totals	**1517**	**535**

Can anything be concluded from the data? Write at most one quarter of a page, or about 50 words. You have to justify your answer (Hint: Think, do not calculate anything).

[1] A singles tournament match involves two players competing against each other, here, two men or two women. In a match, men play up to five sets and women three, the first to win, three out of five sets for men and two of three sets for women, has won the match.

Exercise 8: You are working as an analyst in one of the country's largest online newspapers. As always, there is a discussion on how to present a new article on the frontpage. Each article presented on the frontpage is represented by an image and a corresponding article headline/image caption. The current discussion is about which combination of image and headline that is best suited to improve so-called **user engagement**. How to measure the engagement of a user is a much discussed topic. Here we will use **time-to-exit**, which is defined as the time spent by a user "engaged" in an article, i.e. the time from the user interact with (click) the article on the front page until the user leave that article (or close the browser).

For the article in question, there are two possible images to choose from, **image A** and **image B**, as shown in Figure 1 below. For illustration purposes, assume that the two images are represented by the following two frames.

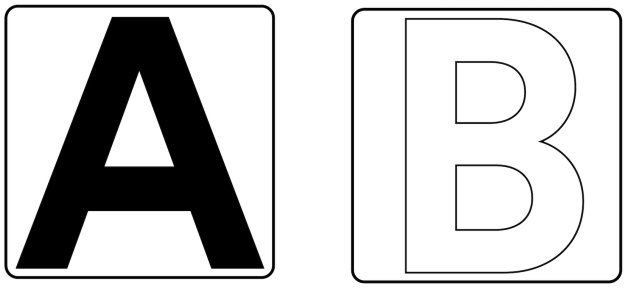


Figure 1: The two images discussed is the conservative **image A** (left) and the somewhat more revealing and scantily clad **image B** (right)

There are also two alternatives for the headline, one is more descriptive and neutral, while the other is of the sensational type. In summary, the two possible headlines to chose from are:

Headline I: The descriptive, less sensational and more neutral headline.

Headline II: The sensational *"you won't believe what happened next"* type of headline.

In order to find the optimal combination of image and headline, with respect to increasing time-to-exit, users (visitors to the site) are randomly shown one out of the four[2] potential combinations of image and headline. The corresponding time-to-exit is then measured (if they interact with the article) along with some additional explanatory variables (see complete list below).

Data: The dataset contains 678 observations (rows) and 6 different types measurements/variables (columns). Data were collected for one hour, from 15:00 to 16:00 on a Friday at the end of December in 2017. Furthermore, only users with an registered account are included in the dataset.

1. `Time`: Time-to-exit, the time, measured in seconds, from a user clicking on the article on the front page until the user exit or leave the article, or close the browser.
2. `Image`: The type of image used (`A` or `B`) to present the article.
3. `Headline`: The type of headline (`I` or `II`) used to describe the article.
4. `Gender`: The gender of the user (`Male` or `Female`) .

[2]Note that with two images and two headlines there are four different combinations and therefore four different ways of presenting the article.

5. `Age`: The age of the user.

6. `Device`: The type of device used (`Mobile`, `Tablet` or `Computer`).

a) The estimated "baseline" model (see Appendix to Part 3) is given by:

$$\text{Time} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{ImageB} + \widehat{\beta}_2 \times \text{HeadlineII} + \widehat{\beta}_3 \times \text{ImageB} \times \text{HeadlineII}.$$

Based on the corresponding R summary (given in the Appendix to Part 3), which combination of image and headline is (i) the least and (ii) the most effective? Write at most one quarter page (or about 50 words) and remember to justify your answer (Hint: Focus on the statistical and practical importance.).

b) Does the data (as described in the Appendix to Part 3) indicate that this is an article with a theme that appeals more to men or women? Write at most one quarter page (or about 50 words) and remember to justify your answer.

c) In addition to the variables included in the baseline model in **a)**, some demographic variables were also collected. The estimated full regression model is given by:

$$\text{Time} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{ImageB} + \widehat{\beta}_2 \times \text{HeadlineII} + \widehat{\beta}_3 \times \text{ImageB} \times \text{HeadlineII} + \widehat{\beta}_4 \times \text{Gender} + \widehat{\beta}_5 \times \text{Age} + \widehat{\beta}_6 \times \text{Device}.$$

Write at most one half page (or about 100 words) comparing the baseline model from **a)** to the full model. In particular, make sure to comment on changes in the main conclusions and the effect and importance of the additional variables.

d) Write at most one half page (or about 100 words) non-technical summary of the main conclusions of the analysis for your statistically challenged boss (i.e. write a summary without using statistical terminology). In particular, your boss is concerned with *gender neutrality* of the frontpage, i.e. the content (the images and articles) should generally not appeal more to one gender over the other; this is therefore something that should be addressed in the summary.

e) Write at most one half page (or about 100 words) discussing only **one** of the following issues:

1. The quality of the data and the current analysis, and also other aspects related to collecting and analysing similar types of data.

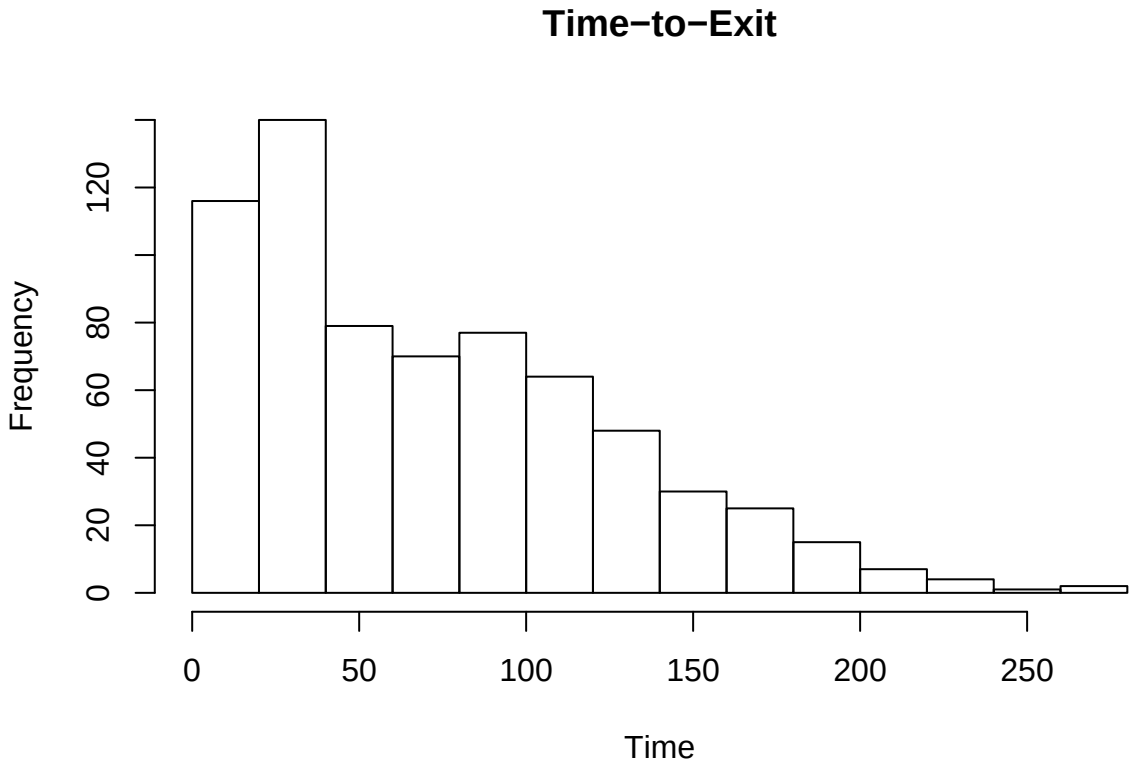
2. It is common to measure user engagement by so-called click-through rate, i.e. the number of users clicking on an article. Discuss advantages and disadvantages of measuring user engagement by click-through rate and time-to-exit (as we did here). Also, construct another way to measure user engagement and discuss your measure in relation to click-through rate and time-to-exit.

3. Discuss challenges and problems of using time-to-exit as a "global" measure for the quality of an article. (Hint: What is the challenge of using something like this to compare different types of articles).

4. Here, we tried to find the best way to combine an image and a headline for presenting a single article. Discuss potential problems of using a similar single tool/criterion for improving each article individually on the frontpage.

Appendix to Part 3

```
hist(Time)
```




```
with(data, table(Gender, Image))
```

```
##         Image
## Gender     A   B
##   Female 122  21
##   Male    64 471
```

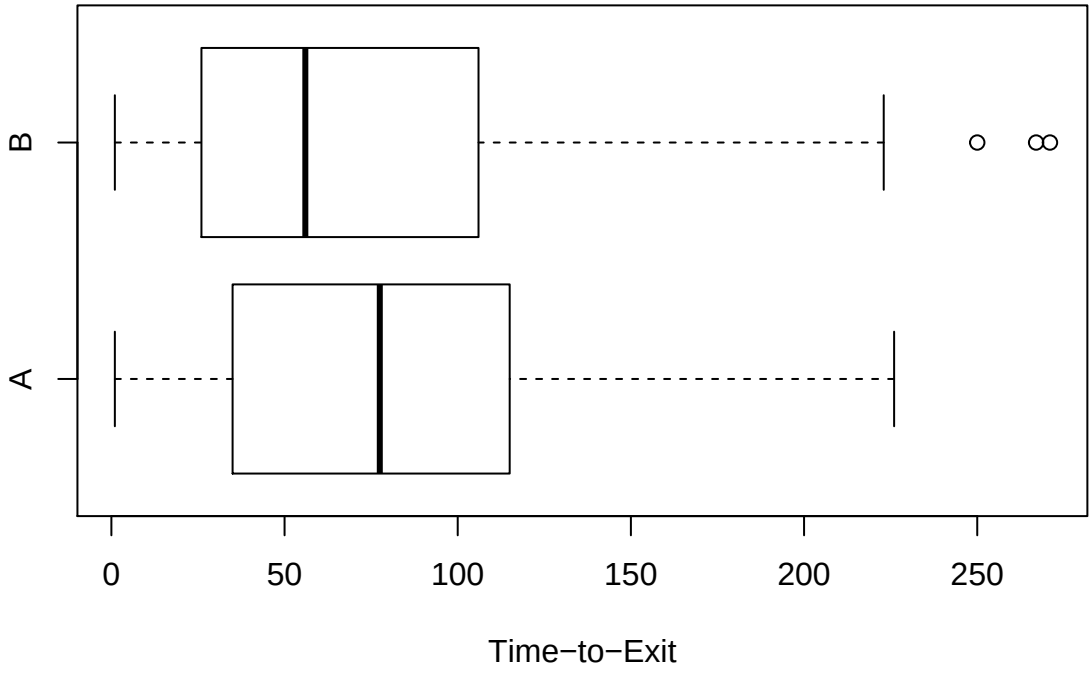
```
with(data, table(Gender, Headline))
```

```
##         Headline
## Gender     I  II
##   Female  78  65
##   Male   264 271
```

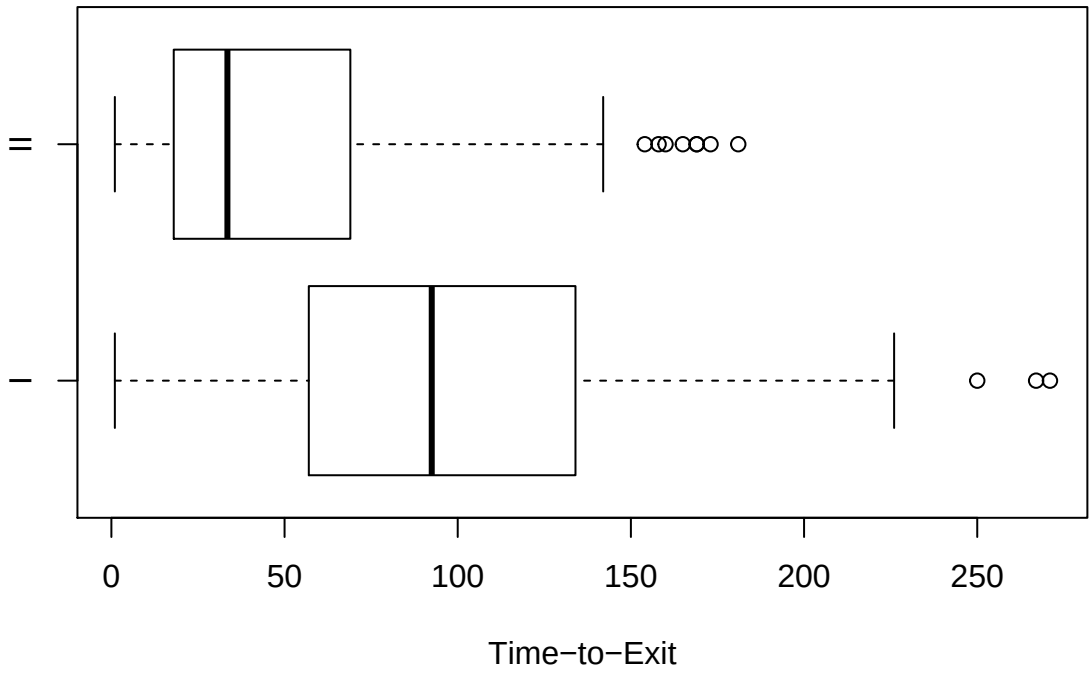
```
with(data, table(Image, Headline))
```

```
##      Headline
## Image   I  II
##     A 102  84
##     B 240 252
```

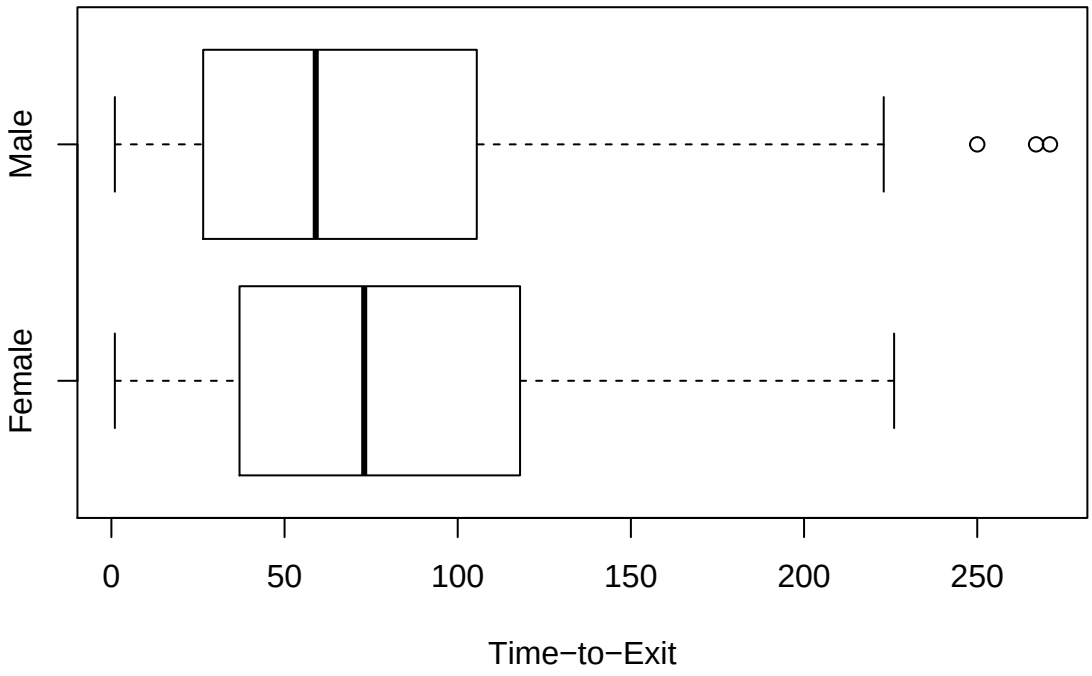
```
with(data, boxplot(Time ~ Image))
```



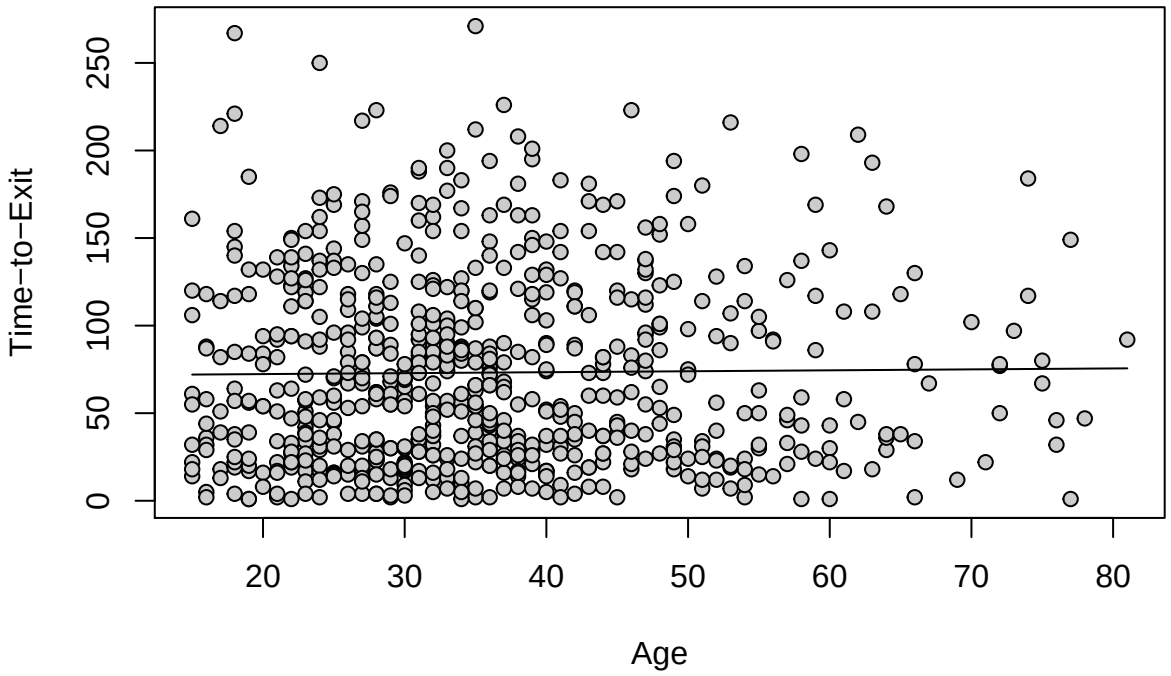

```
with(data, boxplot(Time ~ Headline))
```



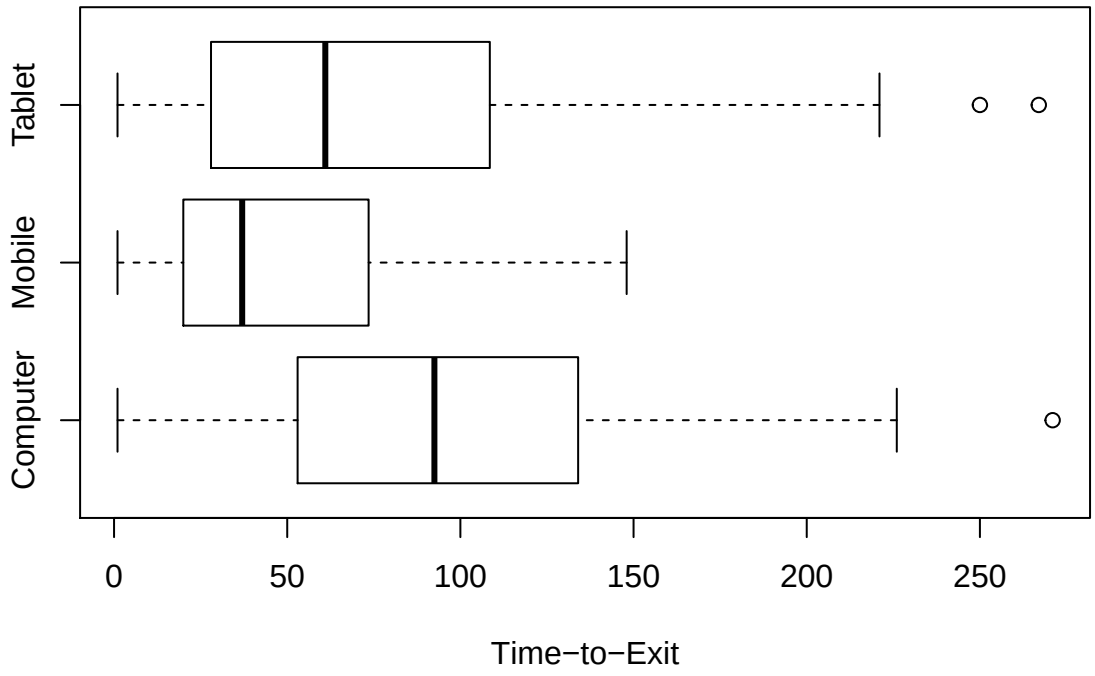
```
with(data, boxplot(Time ~ Gender))
```




```
with(data, plot(Age, Time))
with(data, lines(smooth.spline(Age, Time)))
```



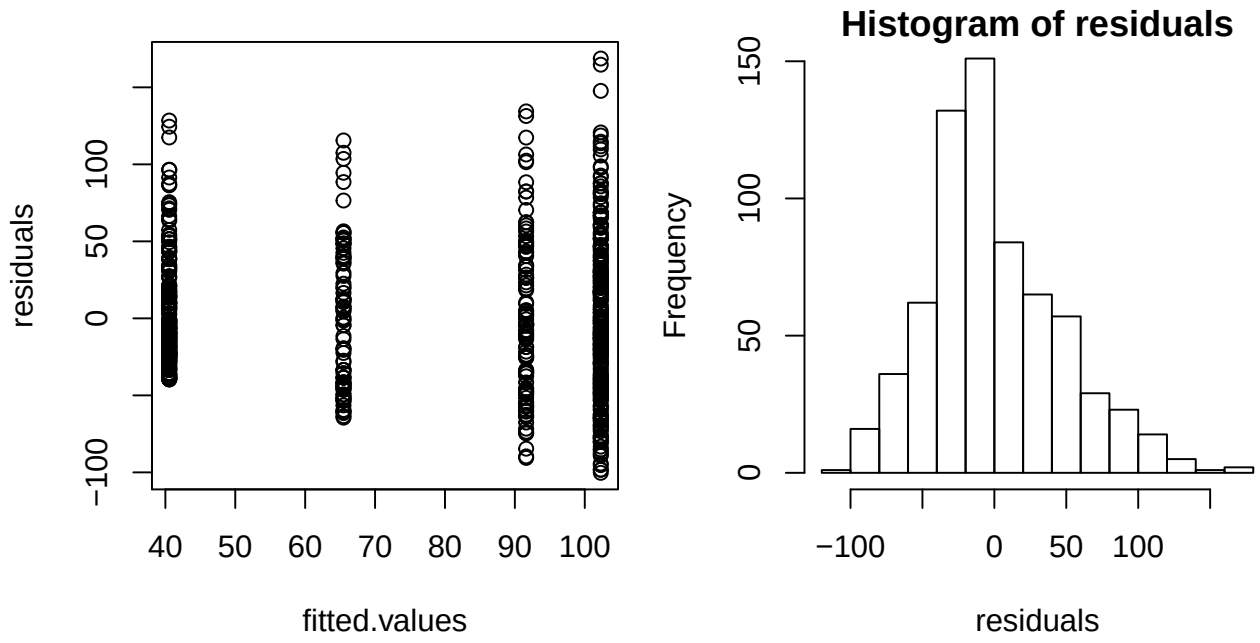
```
with(data, boxplot(Time ~ Device))
```



```r
# a) the baseline model:
baseline_lm <- with(data, lm(Time ~ Image + Headline + Image:Headline))
summary(baseline_lm)
```

```
##
## Call:
## lm(formula = Time ~ Image + Headline + Image:Headline)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.292  -30.487   -8.292   27.708  168.708
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         91.618      4.603  19.904  < 2e-16 ***
## ImageB              10.674      5.495   1.943 0.052479 .
## HeadlineII         -26.130      6.849  -3.815 0.000149 ***
## ImageB:HeadlineII  -35.611      8.031  -4.434 1.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.49 on 674 degrees of freedom
## Multiple R-squared:  0.2598, Adjusted R-squared:  0.2565
## F-statistic: 78.87 on 3 and 674 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(1, 2))
with(baseline_lm, plot(fitted.values, residuals))
with(baseline_lm, hist(residuals))
```



```r
# c) the full model:
full_lm <- with(data, lm(Time ~ Image
                              + Headline
                              + Image:Headline
                              + Device
                              + Age
                              + Gender))
summary(full_lm)
```

```
##
## Call:
## lm(formula = Time ~ Image + Headline + Image:Headline + Device +
##     Age + Gender)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.668  -29.608   -3.857   25.082  169.545
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       114.3216     6.8481  16.694  < 2e-16 ***
## ImageB              9.8837     6.0309   1.639    0.102
## HeadlineII        -29.7538     6.2302  -4.776 2.20e-06 ***
## DeviceMobile      -49.4360     4.0820 -12.111  < 2e-16 ***
## DeviceTablet      -26.7782     3.9239  -6.824 1.98e-11 ***
## Age                 0.1522     0.1239   1.228    0.220
## GenderMale         -2.7117     5.3730  -0.505    0.614
## ImageB:HeadlineII -31.2904     7.3092  -4.281 2.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.18 on 670 degrees of freedom
## Multiple R-squared:  0.3943, Adjusted R-squared:  0.388
## F-statistic: 62.31 on 7 and 670 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(1, 2))
with(full_lm, plot(fitted.values, residuals))
with(full_lm, hist(residuals))
```