Statistical inference
MS-C1620 Exam
Department of Mathematics and Systems Analysis 8.4.2020
Aalto University J Kohonen

Instructions: Answer in English. Write clearly and give reasons for your answers. A number only as an answer does not yield points. The exam has 4 problems, each worth 6 points.

Write your solutions by hand, clearly, on paper (or a tablet computer), and send your solutions in PDF form to the return box on the course page. Make sure that every page shows: course code, last name, first name, student number and date.

P1 Answer either TRUE or FALSE (in this problem, reasons not required). 1 point per item for correct answer, maximum amount of points obtainable is 6.

(a) Median is a measure of scatter.

(b) Descriptive statistics aims to draw conclusions about a population based on a sample.

(c) In the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the error term is usually assumed to have expected value one, $\mathbb{E}(\varepsilon_i) = 1$.

(d) If two predictors are highly correlated with each other in linear regression, this can make the coefficient estimates unstable.

(e) In hypothesis testing, the probability of a Type II error is always greater than or equal to the probability of a Type I error.

(f) In hypothesis testing, the null hypothesis is rejected when getting a p-value smaller than the significance level.

(g) LASSO can be used for variable selection.

(h) In bootstrap, the number of observations in each of the bootstrap samples is the same as the number of observations in the original sample.

P2

(a) You are testing the following hypotheses with some method of normality testing.

H_0 : The sample $x_1, \ldots, x_n$ comes from a normal distribution.

H_1 : The sample $x_1, \ldots, x_n$ does not come from a normal distribution.

Describe what it means to conduct Type I and Type II errors in this context (do not give the general definitions of Type I and II errors but instead state what they mean for this specific pair of hypotheses). **(2p)**

(b) Draw two examples of quantile-quantile (Q-Q) plots: (i) one where the sample clearly comes from a normal distribution, and (ii) one where it clearly does not. **(2p)**

(c) Name two ways besides Q-Q plot for checking/testing the normality of a sample. **(1p)**

(d) A researcher wants to model her data with Model X that makes a normality assumption. For this, she tests her data for normality and gets a p-value of 0.055 (for the hypotheses given in part a). Based on the p-value, she decides to use Model X. Can the researcher fully trust the results of the model? Explain why or why not. **(1p)**

P3 Consider multiple linear regression on a sample of $n = 100$ observations of a response variable y_i and the explanatory variables $x_{i1}, x_{i2}, x_{i3}, x_{i4}$. Below are shown the linear regression model summary, variance inflation factors and the diagnostics plot for the model fit.

(a) Describe how one can check whether the assumptions of multiple linear regression are satisfied. Are they satisfied in the current example? **(3p)**

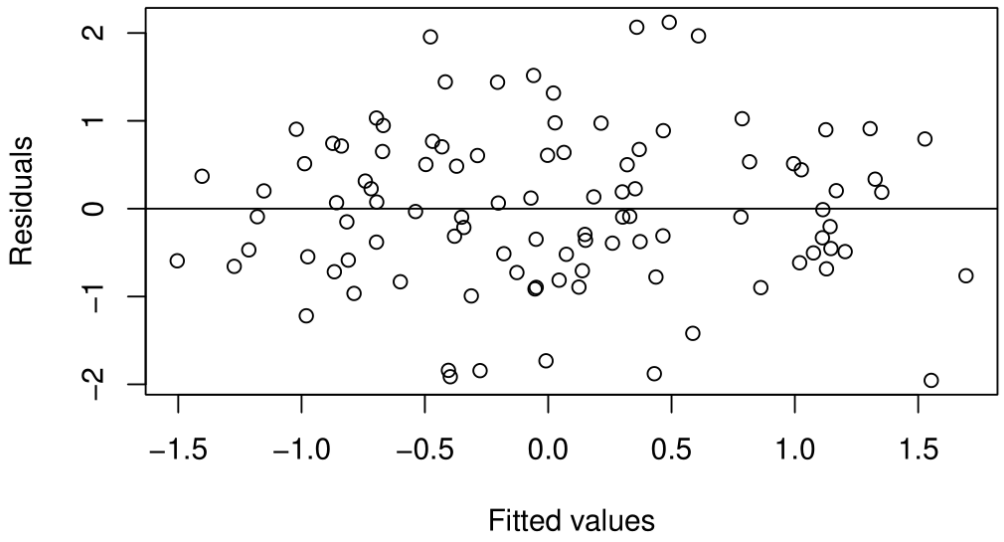
(b) Based on the variance inflation factors, can the estimated coefficients be trusted? Why or why not? Explain what it would mean if these factors are high or low. **(1p)**

(c) Give an interpretation for the estimated coefficient $\hat{\beta}_4 \approx -0.44$. **(1p)**

(d) What does the fitted model predict for the response variable if $x_{i1} = x_{i2} = x_{i3} = 0$ and $x_{i4} = 100$? Give a numerical answer and explain why or why not this prediction can be trusted. **(1p)**

```
## Call:
## lm(formula = y ~ ., data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2158 -0.6744  0.1267  0.6918  1.9987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.08357    0.09932  -0.841   0.4022
## x1          -0.59245    0.09081  -6.524 3.42e-09 ***
## x2          -0.07999    0.12061  -0.663   0.5088
## x3          -0.55658    0.21660  -2.570   0.0118 *
## x4           0.21811    0.20657   1.056   0.2937
## x5           0.72269    0.09904   7.297 9.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9499 on 94 degrees of freedom
## Multiple R-squared:  0.5637, Adjusted R-squared:  0.5405
## F-statistic: 24.29 on 5 and 94 DF,  p-value: 1.285e-15

## Variance inflation factors
##       x1       x2       x3       x4       x5
## 1.014837 1.024145 5.757164 5.738664 1.056317
```



P4

(a) How can we control the probability of Type I error in hypothesis testing? **(1p)**

(b) Why is the probability of Type II error more difficult to control in hypothesis testing than the probability of Type I error? **(1p)**

(c) When and why should you consider adjusting the significance level (for example, with the Bonferroni correction) in hypothesis testing? **(2p)**

(d) Consider testing the null hypothesis H_0 : The sample $x_1, \ldots, x_n$ comes from a normal distribution.

(i) Give an example of a statistical test for H_0 for which the probability of Type II error is 100%. **(1p)**

(ii) Give an example of a statistical test for H_0 for which the probability of Type I error is 50%. **(1p)**

Hint: In each case, you do not need to care about the other error type, and the test does not necessarily need to be a conventional statistical test (although it can be). You need to describe a procedure that makes accept/reject decisions and has the required error rate.