Statistical inference
MS-C1620 Exam
Department of Mathematics and Systems Analysis 1.6.2022
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.

Allowed equipment: writing tools, calculator (symbolic and graphic OK), at most A4-size cheat sheet written on one side.

P1 In each question, answer true or false. 1 point per correct answer.

(a) Sample median is the 0.5-quantile of a sample.

(b) If a distribution is symmetric around zero, it is called the normal distribution.

(c) Positive skewness indicates that the expectation of the distribution is above zero.

(d) In linear regression, correlation coefficient is the slope of the regression line.

(e) Confidence level indicates how much confidence you have in the model assumptions.

(f) Bonferroni correction is used for removing outliers from a 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 their general definitions but explain what they mean here specifically). **(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 and briefly explain two ways besides Q-Q plot for checking/testing the normality of a sample. **(2p)**

P3 Consider multiple linear regression on a sample of $n = 100$ observations of a response variable y_i and the explanatory variables $x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}$. Below are shown the linear regression model summary, variance inflation factors and the diagnostics plot (next page) 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)**

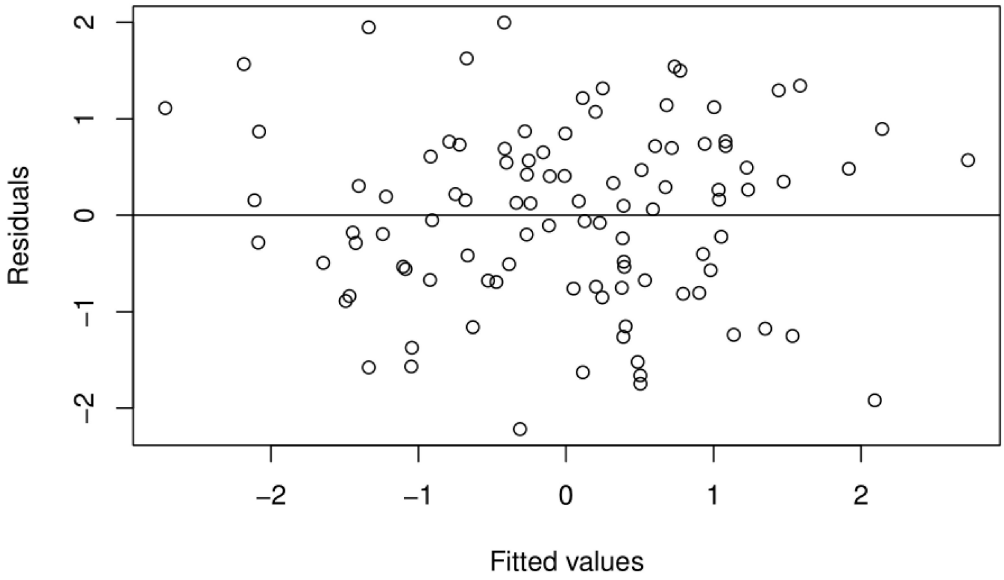
(b) Based on the variance inflation factors, can the estimated coefficients be trusted? Why or why not? Explain what it would mean if these factors are high or low. **(1p)**

(c) Give an interpretation for the estimated coefficient $\hat{\beta}_4 \approx 0.21811$. **(1p)**

(d) What does the fitted model predict for the response variable if $x_{i1} = x_{i2} = x_{i3} = 0$, $x_{i4} = 100$, and $x_{i5} = 0$? 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 Consider a dataset of $n = 7$ observations with x values

$$3.0, 3.5, 4.0, 5.0, 5.5, 6.0, 9.0$$

and y values

$$0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 2.0.$$

Let K be the **rectangular** kernel function

$$K(u) = \begin{cases} 1 & \text{if } |u| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Calculate the Nadaraya-Watson regression function at points $x = 3.5, 4.2, 7.5$ and 9.2 using K as the kernel. Give at least three decimals. If at some points it cannot be calculated, explain why. **(2p)**

(b) What happens with this dataset if we use the kernel function $K_{0.01}(u) = K(u/0.01)$? Where exactly can the regression be calculated and what are its values there? **(1p)**

(c) Calculate the regression function at $x = 4.0$ and at $x = 7.5$ using the kernel function $K_{100}(u) = K(u/100)$. Give at least three decimals. **(1p)**

(d) Explain in general terms (not just for this dataset) how bandwidth affects Nadaraya-Watson kernel regression. Consider both small and large values. **(1p)**

(e) Explain how cross-validation can be used to select bandwidth in kernel regression. **(1p)**