Statistical inference
MS-C1620
Department of Mathematics and Systems Analysis
Aalto University

Exam
19.4.2023
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.

P1 Exactly four of the following claims are true, and four are false. **Mark** each claim as TRUE or FALSE. If false, **explain** briefly (in 1–2 sentences) why it is false. (No explanation needed for true claims.) 0.5 points for each correct true/false answer, and 0.5 points for each correct explanation of a false claim (total 6 points).

(a) Bootstrapping is based on taking new samples from an already existing sample.

(b) The coefficient of determination (R^2) measures how much of the variance in the response variable is being explained by the model.

(c) If a sample comes from a normal distribution, its skewness coefficient equals zero.

(d) In hypothesis testing, high p-values are desired because they indicate high probability of correct results.

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

(f) In cross-validation, one keeps some of the data hidden when training the statistical model.

(g) Nonparametric tests generally make less assumptions about the underlying distribution than parametric tests.

(h) Positive skewness indicates that both tails of the distribution are long.

P2 Consider multiple linear regression on a sample of $n = 100$ observations of a response variable y_i and the explanatory variables $x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}$. On the next page are shown the linear regression model summary, variance inflation factors and the diagnostics plot for the model fit.

(a) Explain the diagnostic plot: what are the two axes, and what does it mean when a residual is positive or negative. Explain what one can in general see in a diagnostic plot, and interpret the current plot in this context. Also explain how residuals can be used for estimating the error variance. **(2p)**

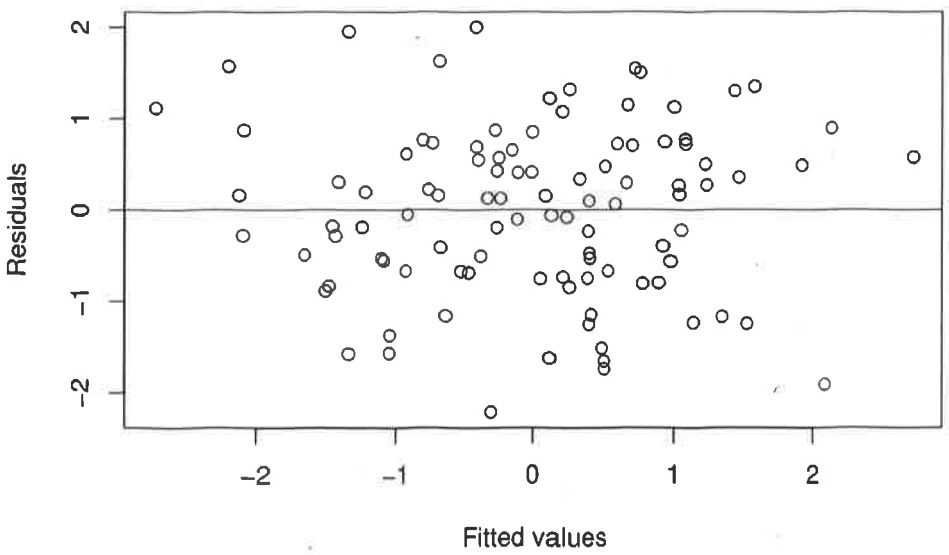
(b) Based on the variance inflation factors, can the estimated coefficients be trusted? Why or why not? Explain what it would mean if these factors are high or low. **(2p)**

(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
```



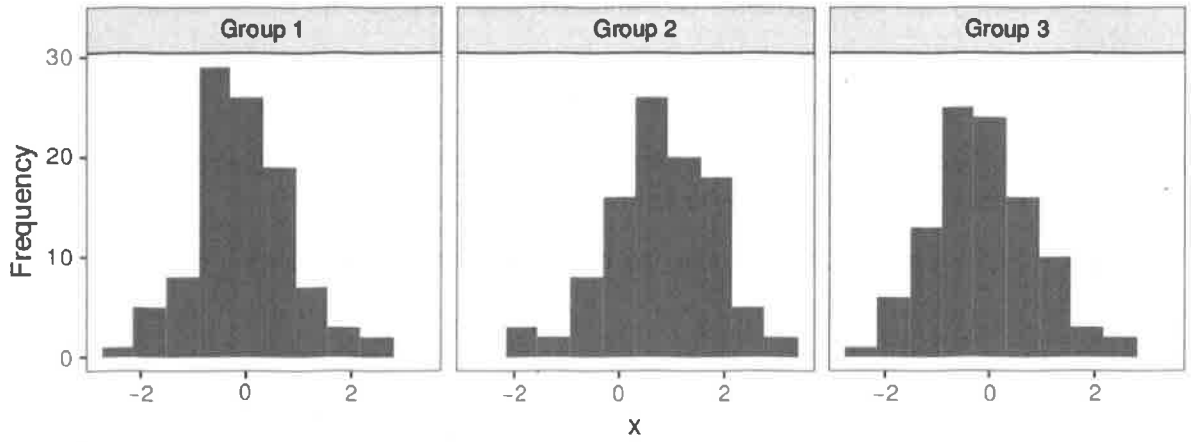
P3 Consider analysis of variance (ANOVA) on a sample of three groups with 50 observations in each of them (assume that the groups are independent and that the observations are i.i.d. within each group). The following plot shows (in this order) the histograms of the groups, the ANOVA summary, and the results of Bartlett's test.

(a) State the null hypothesis and the alternative hypothesis of ANOVA for this three-group case. **(1p)**

(b) What would you conclude based on the ANOVA results? **(1p)**

(c) Describe how one can check whether the assumptions of ANOVA are satisfied. Are they satisfied in the current example? **(2p)**

(d) The next step in the analysis would be to conduct pair-wise testing between the groups. Bonferroni correction is often used in this context. Why is this? Describe also how the Bonferroni correction is applied. **(2p)**




```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## group         2  57.83  28.913   31.61 3.57e-13 ***
## Residuals   297 271.63   0.915
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## ################################################################

##
##  Bartlett test of homogeneity of variances
##
## data:  x by group
## Bartlett's K-squared = 0.85135, df = 2, p-value = 0.6533
```

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)**