UPPSALA UNIVERSITY
Department of Mathematics
Silvelyn Zwanzig

Regression Analysis, 5hp, 1MS555
17 August 2022

Permitted aids: pocket calculator, one hand-written sheet of formulae (2 pages)

Time: 5 hours. For a pass (mark 3) the requirement is at least 18 points. For the mark 4, 25-31 points are necessary. For an excellent test (mark 5) the requirement is at least 32 points. Every problem is worth 5 points.

OBS: *Please explain and interpret your approach carefully. Don't try to write more than really needed, but what you write must be clear and well argued.*

1. Consider the following matrix $A = B^T(BB^T)^{-1}B$ with

$$B = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{bmatrix}.$$

   (a) Is A invertible?

   (b) Show that A is a projection matrix.

   (c) Which dimension has the related subspace?

   (d) Suppose $X \sim N_5(0, I)$. Which distribution has X^TAX?

2. Consider a simple linear regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i \ , \ i = 1, \ldots, 2m$$

   with two different design points only: $x_i = -1$ for $i = 1, \ldots, m$ and $x_i = 1$ for $i = m+1, \ldots, 2m$. Let

$$\overline{y}_{(1)} = \frac{1}{m}\sum_{i=1}^{m} y_i, \ \ \overline{y}_{(2)} = \frac{1}{m}\sum_{i=m+1}^{2m} y_i.$$

   (a) Derive a formula for the least squares estimator basing on $\overline{y}_{(1)}, \overline{y}_{(2)}$.

   (b) Calculate the covariance matrix for the least squares estimator.

(c) Compare the line, which connect the points $(-1, \overline{y}_{(1)})$ and $(1, \overline{y}_{(2)})$, with the fitted line calculated by the least squares method.

(d) Sign a picture with both lines.

3. Consider a linear regression model

$$y_i = \beta_1 x_i + \beta_2 z_i + \beta_3 t_i + \varepsilon_i, \quad i = 1, \ldots, 4m$$

with two different design points: $x_i = -1$ for $i = 1, \ldots, 2m$ and $x_i = 1$ for $i = 2m+1, \ldots, 4m$. Suppose $z_i \in \{-1, 1\}$ and $t_i \in \{-1, 1\}$.

(a) Rewrite the model in matrix form. Determine the design matrix.

(b) Give the formulary for the least squares estimator. Determine the covariance matrix of the least squares estimator.

(c) Find values for z_i and t_i such that the covariance matrix of the least squares estimator has diagonal form.

4. Consider the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, \ldots, n$$

where ε_i are dependent r.v. with $E\varepsilon_i = 0$ and $Cov(\varepsilon_i, \varepsilon_j) = \sigma^2_{i,j}$.

(a) How is the ordinary least squares estimator for $\beta = (\beta_0, \beta_1)^T$ defined? Give the formulary of the covariance matrix of the least squares estimator.

(b) How is the generalized least squares estimator for $\beta = (\beta_0, \beta_1)^T$ defined? Give the formulary of the covariance matrix.

(c) What means, that one estimator is better? Which estimator is the best? Why?

5. Let $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$. With $y := [2.9, 5.2, 9.2, 9.1, 11.2, 14.3]$ and $x = [0, 1, 2, 3, 4, 5]$.

(a) Calculate the least squares estimator (LSE) for β_1 and β_2.

(b) Calculate the LSE for β_1 and β_2, under $H : \beta_2 + \beta_1 = 0$.

(c) Suppose $\varepsilon_i \sim N(0, 1)$. Derive a test statistic for $H_0 : \beta_2 + \beta_1 = 0$.

(d) Suppose $\varepsilon_i \sim N(0, \sigma^2)$, where σ^2 is unknown. Derive a test statistic for $H_0 : \beta_2 + \beta_1 = 0$.

(e) What are the results in c) and d)? Take $\alpha = 0.05$ (Hint: The quantile of the Chi -distribution with 1 d.f. for $\alpha = 0,05$ is 3.841. The quantile of the F -distribution with 1 d.f. and 4 d.f. for $\alpha = 0,05$ is 7.71.)

(f) Sign a picture with the observations and the fitted line in the "large" model and the fitted line under the hypotheses. Interpret the test results!

6. Consider a balanced three sample problem. The variance of sample one is fourth of the variance of Sample 2 : $\sigma_1^2 = 4\sigma_2^2$. Sample 2 and Sample 3 have same variances.

   (a) Formulate the joint regression model. Introduce a dummy variable.

   (b) Estimate the variance σ_1^2 as function of only observations of Sample 1.

   (c) Estimate the variance σ_1^2 as function of all observations.

   (d) Assume normal distribution for the errors. Derive the distributions for both variance estimates.

   (e) Which estimate is better? Why?

7. Suppose you have the following data set from the tourists information on Gotland, on Öland and from Dalarna, Skåne, Stockholm, Göteborg organized in a joint data.frame containing the number of tourists this year and last year during June and July, the origin of the tourist: Sweden, Germany, Netherlands, Denmark, Norway. Furthermore you have the rate for the Swedish krona (SEK) to the EURO, Danish krone (DKK), Norwegian krone (NOK), the number of sun days, the average temperatures. In news paper it is stated, that because of the weak Swedish krona, the Swedish prefer to spend their holidays in Sweden.

   (a) Formulate a regression model for that purpose. Explain the difference between response and covariates. Which distribution assumptions are needed?

(b) Formulate the test problem for testing the statement of the news paper.

(c) Define the respective F statistics.

(d) Give the main R commands.

(e) Which covariates would you like to include in your data set.

8. Consider the following R code and R results

```
> names(UN11)
[1] "region"  "group" "fertility" "ppgdp"  "lifeExpF"  "pctUrban"
> attach(UN11)
> M1<-lm(lifeExpF~ppgdp+log(fertility)+pctUrban)
> R1<-resid(M1)
> shapiro.test(R1)
 Shapiro-Wilk normality test
data:  R1
W = 0.9596, p-value = 1.879e-05
> qq.plot(R1)
> summary(M1)
Residuals:
    Min      1Q  Median      3Q     Max
-20.028  -2.764   0.265   3.209  13.352

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.119e+01  1.838e+00  44.166  < 2e-16 ***
ppgdp           7.935e-05  2.655e-05   2.988  0.00317 **
log(fertility) -1.520e+01  1.042e+00 -14.584  < 2e-16 ***
pctUrban        6.805e-02  2.187e-02   3.111  0.00215 **

Residual standard error: 5.385 on 195 degrees of freedom
Multiple R-squared:  0.7213,    Adjusted R-squared:  0.717
F-statistic: 168.2 on 3 and 195 DF,  p-value: < 2.2e-16
> M2<-lm(lifeExpF~log(fertility))
> R2<-resid(M2)
> shapiro.test(R2)
 Shapiro-Wilk normality test
```

```
data:  R2
W = 0.9678, p-value = 0.0001571
> qq.plot(R2)
> summary(M2)
Residuals:
     Min       1Q   Median       3Q      Max
-20.3941  -3.6426   0.1302   4.1029  14.7973
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     89.3006     0.9427   94.73   <2e-16 ***
log(fertility) -18.6437     0.9293  -20.06   <2e-16 ***
---
Residual standard error: 5.818 on 197 degrees of freedom
Multiple R-squared:  0.6714,    Adjusted R-squared:  0.6697
F-statistic: 402.5 on 1 and 197 DF,  p-value: < 2.2e-16
```

(a) Give regression equation of model 1 and formulate the distribution assumptions.

(b) Are the distribution assumptions fulfilled?

(c) Which tests are carried out under model 1? Formulate the hypotheses, and the value of the respective test statistics. Can we relay on the results?

(d) Compare both models. Which model do you would recommend?

(e) How you would proceed for analysing this data set?

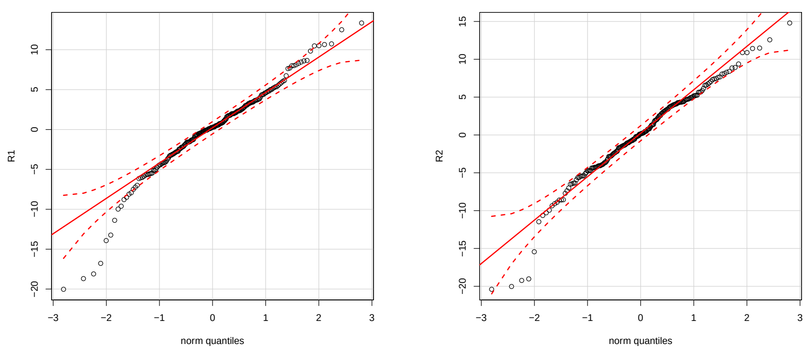

Figure 1: Problem 8