UPPSALA UNIVERSITY
Department of Mathematics
Silvelyn Zwanzig

Regression Analysis, 5hp, 1MS555
May 25, 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

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

   (a) Is A invertible?

   (b) Show that A is a projection matrix.

   (c) Which dimension has the related subspace?

   (d) Suppose $X \sim N_6(0, I)$. Which distribution has $X^T AX$?

   (e) Suppose $X \sim N_6(0, I)$ and $b \in \mathbf{R}^6$ such that $Ab = 0$. Determine the function Ψ and the distribution of

$$\frac{b^T X}{\Psi(X^T AX)}.$$

2. Consider a simple linear regression model for independent observations

$$y_i = \beta x_i + \varepsilon_i,\ E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2,\ \ i = 1, \ldots, 3m$$

   with three different design points only: $x_i = -2$ for $i = 1, \ldots, m$, $x_i = 0$ for $i = m+1, \ldots, 2m$, and $x_i = 2$ for $i = 2m+1, \ldots, 3m$. Let

$$\overline{y}_{(1)} = \frac{1}{m}\sum_{i=1}^{m} y_i,\ \ \overline{y}_{(2)} = \frac{1}{m}\sum_{i=m+1}^{2m} y_i,\ \ \overline{y}_{(3)} = \frac{1}{m}\sum_{i=2m+1}^{3m} y_i.$$

   (a) Derive a formula for the least squares estimator basing on $\overline{y}_{(1)}, \overline{y}_{(2)}, \overline{y}_{(3)}$.

   (b) Derive a formula for the variance estimator basing on $\overline{y}_{(1)}, \overline{y}_{(2)}, \overline{y}_{(3)}$.

(c) Compare the line, which connect the points $(-2, \overline{y}_{(1)})$, $(2, \overline{y}_{(3)})$ with the line fitted by the least squares.

3. Consider the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i,\ i = 1, \ldots, n$$

where ε_i are uncorrelated, with $E\varepsilon_i = 0$ and $Var(\varepsilon_i) = (x_i + z_i)^2\sigma^2$, σ^2 is known.

(a) How is the ordinary least squares estimator for $\beta = (\beta_0, \beta_1, \beta_2)^T$ defined? Give the formulary of the estimator. Derive the formulary of the covariance matrix of the least squares estimator.

(b) How is the generalized least squares estimator for $\beta = (\beta_0, \beta_1, \beta_2)^T$ defined? Give the formulary of the estimator. Derive the formulary of the covariance matrix of the generalized least squares estimator.

(c) What means, that one estimator is better?

(d) Which estimator the better? Why?

4. Suppose you have the following data set organized in a data.frame: $t = 1960, \ldots, 2010$ is the year, $m = 1, \ldots, 12$ is the month, Temp(t,m) is the monthly average water temperature of the Fyrisån at vindbro. Further from SMHI data of monthly average amount of precipitation P(t,m) and monthly average air temperature, ATemp(t,m) in the m' th month of year t, are available.

(a) Formulate a regression model on climate change, where the water temperature is the response variable. (You can take a subset of the data.)

(b) Explain the difference between response and covariates. Which distribution assumptions are needed? Are the distribution assumptions realistic?

(c) Formulate test problems of interest. Define the respective F statistics.

(d) Give the main R commands.

5. Compare to independent regression lines with the same unknown intercept:

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

and

$$z_j = \beta_0 + t_j\gamma_1 + \varepsilon_j, j = 1, \ldots, n.$$

The errors are i.i.d. normal distributed with same variance σ^2.

(a) Formulate a joint regression model including both regression lines.

(b) Formulate the hypothesis that both lines are parallel.

(c) Formulate the model under the hypotheses.

(d) Give the formulary for the F-statistics. Determine the degrees of freedom.

6. Consider a linear model with orthogonal design and independent standard normal distributed errors:

$$y_i = \alpha \frac{1}{\sqrt{6}} + \beta x_i + \epsilon_i, \;\; x_1 = x_2 = -\frac{1}{2}, x_3 = x_4 = 0, x_5 = x_6 = \frac{1}{2}$$

(a) Give the least squares estimators of α and β.

(b) Derive a test statistic for testing $H_0 : \beta = 3$ versus $H_1 : \beta \neq 3$. (Note that, the error variance is known.)

(c) Which distribution has the test statistic under the null hypothesis?

(d) Define the p values for the test.

7. We consider the data set governmental salary "salarygov". For 495 different job classes the following variables are included: MaxSalary: maximum salary in dollars; NE: total number of employees, NW: number of woman employees, Score: score of the job based on skill level, responsibility, difficulty. The goal of the statistical study is an examination of the salary with respect to equality rights. The following R-code is carried out:

```
> head(salarygov)
                     JobClass NW NE Score MaxSalary
1               Account_clerk 52 68   258      1549
2 Account_clerk_Intermediate 26 29   269      1712
3    Account_clerk_Principal 10 13   321      2182
4       Account_clerk_Senior 16 24   273      1982
5                  Accountant  1 12   352      2555
6            Accountant_Chief  0  5   709      4060
>
> M1<-lm(MaxSalary~Score+NE+NW)
> summary(M1)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 294.6439    66.4036   4.437 1.13e-05 ***
Score         5.7446     0.1269  45.281  < 2e-16 ***
NE            3.3598     1.7271   1.945   0.0523 .
NW           -4.8626     2.3586  -2.062   0.0398 *

Residual standard error: 506 on 491 degrees of freedom
Multiple R-squared:  0.8181,    Adjusted R-squared:  0.817
F-statistic: 736.3 on 3 and 491 DF,  p-value: < 2.2e-16

> M2<-lm(MaxSalary~Score)
```

```
> summary(M2)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  295.274     62.012   4.762 2.53e-06 ***
Score          5.760      0.123  46.844  < 2e-16 ***

Residual standard error: 507.2 on 493 degrees of freedom
Multiple R-squared:  0.8165,    Adjusted R-squared:  0.8162
F-statistic:  2194 on 1 and 493 DF,  p-value: < 2.2e-16

> M3<-lm(MaxSalary~Score+NW)
> summary(M3)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 311.7123    66.0073   4.722 3.04e-06 ***
Score         5.7362     0.1272  45.113  < 2e-16 ***
NW           -0.8065     1.1056  -0.729    0.466

Residual standard error: 507.4 on 492 degrees of freedom
Multiple R-squared:  0.8167,    Adjusted R-squared:  0.816
F-statistic:  1096 on 2 and 492 DF,  p-value: < 2.2e-16
```

(a) Formulate the regression model 1 including all distribution assumptions.

(b) Which tests are carried out under model 1? Formulate the hypotheses, and the value of the respective test statistics.

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

(e) Carry out the F test comparing model M2 and M3.

(f) What is your conclusion on equality rights of governmental salary?

8. Consider the data set of Problem 7. The following R code is carried out:

```
> A<-NW/NE
> MM<-lm(MaxSalary~A+Score)
> summary(MM)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  616.7990    74.1312   8.320 8.69e-16 ***
A           -403.3350    56.2020  -7.177 2.65e-12 ***
Score          5.4203     0.1263  42.919  < 2e-16 ***

Residual standard error: 483 on 492 degrees of freedom
Multiple R-squared:  0.8339,    Adjusted R-squared:  0.8333
F-statistic:  1235 on 2 and 492 DF,  p-value: < 2.2e-16

> D<-cooks.distance(MM)
```

```
> hist(MaxSalary) # see Figure 1
> plot(1:n,D) # see Figure 1
> points(375,D[375],col=2, lwd=3)
>salarygov[375,]
                   JobClass NW NE Score MaxSalary
375 Physician_SR._Associate  0  1   809      8524

> data1<-subset(salarygov,MaxSalary<7000)
> A1<-data1$NW/data1$NE
> MM1<-lm(MaxSalary~A1+Score,data1)
> summary(MM1)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  638.4631    70.0991   9.108  < 2e-16 ***
A1          -397.7319    53.1078  -7.489 3.24e-13 ***
Score          5.3540     0.1196  44.753  < 2e-16 ***


Residual standard error: 456.4 on 491 degrees of freedom
Multiple R-squared:  0.8452,    Adjusted R-squared:  0.8446
F-statistic:  1340 on 2 and 491 DF,  p-value: < 2.2e-16
```

(a) Formulate the regression model MM including all distribution assumptions.

(b) Are the distribution assumptions fulfilled? (see Figure 2)

(c) Give the definition of an influential point and of cooks distance. What means maximal cooks distance?

(d) Is the observation 375 an outlier in the sense that it does not follow the model assumptions?

(e) What is your conclusion on governmental salary?

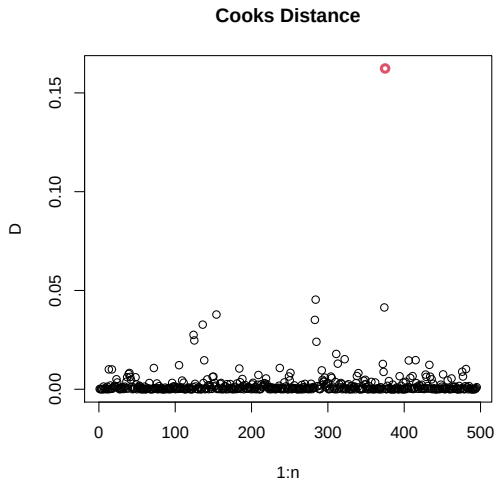

Figure 1: Illustration for Problem 8.

Figure 2: Illustration for Problem 8.