Exam 2023

PROBLEM 1 - Education and church membership

We use the dataset given in the description of the exam.
We will investigate a model of the form:

$$Church = \alpha + \beta_1 \cdot H3020 + \beta_2 \cdot H4058 + \beta_3 \cdot H5020 + \beta_4 \cdot H8090 + \epsilon ,$$

Where Church is the church membership per municipality, α is an intercept, the β 's are estimated parameters, and ϵ is the residual.

H3020 is a vocational education related to food
H4058 is a short technical education, requiring a high school diploma
H5020 is a medium length pedagodigal education
H8090 is a PhD in a health related area

Question 1.1
The R of the model is

We start by running the SAS-script with the data included in the Exam Set. We now have the data in the exam2023 and can call proc reg or proc GLM

```
proc glm data=exam2023;
model church = H3020 H4058 H5020 H8090;
run;
```

we get

The GLM Procedure

Dependent Variable: church

R-Square	Coeff Var	Root MSE	church Mean
0.995127	7.301293	3072.134	42076.57

Using R we run the script with the data and we can then use

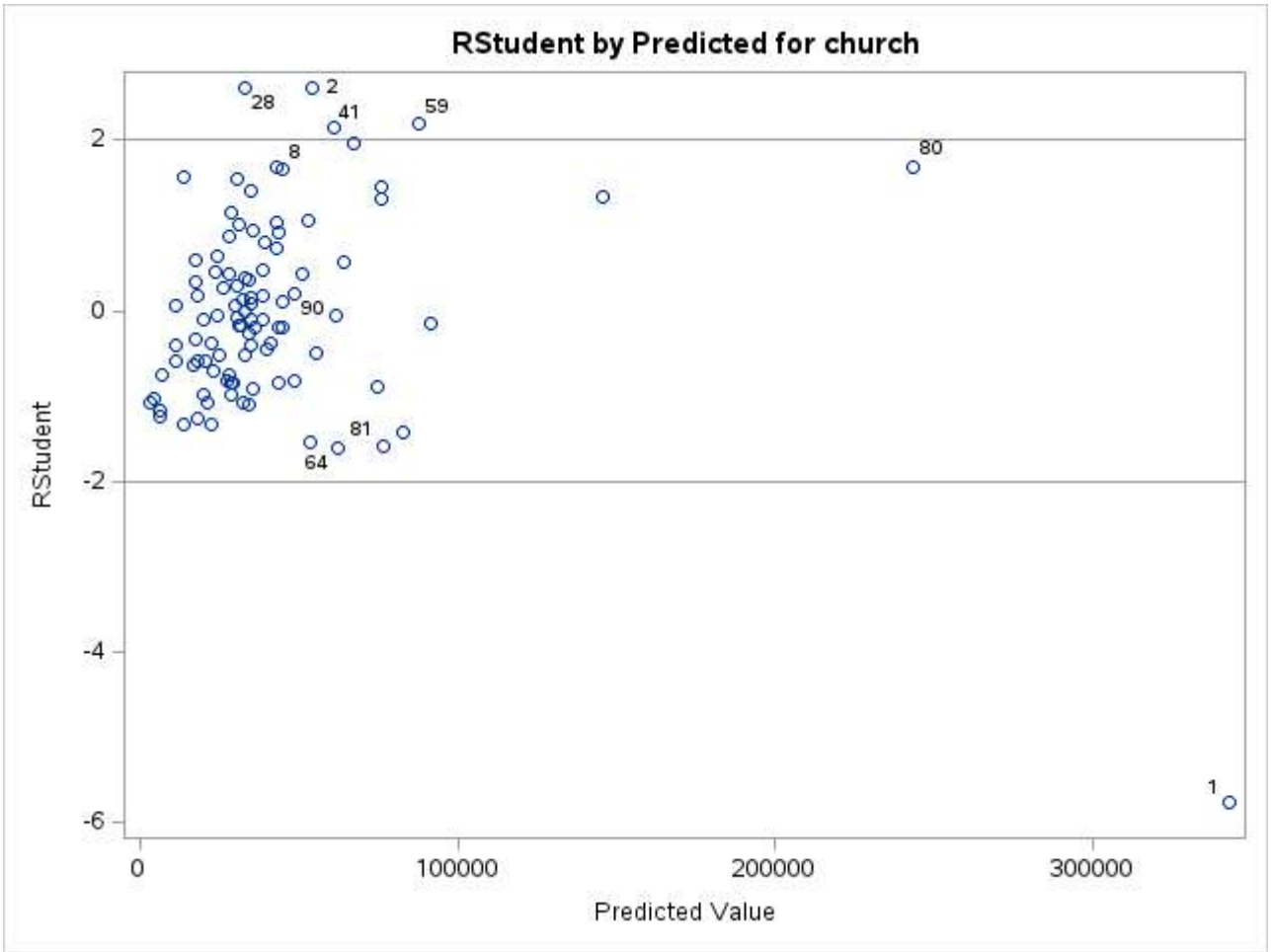
```
lmfit = lm(church ~ H3020 + H4058 + H5020 + H8090, data = exam2023)
# Q 1.1
summary(lmfit)
```

```
Residual standard error: 3072 on 93 degrees of freedom
Multiple R-squared:  0.9951,   Adjusted R-squared:  0.9949
F-statistic:  4748 on 4 and 93 DF,  p-value: < 2.2e-16
```

Question 1.2
The following number of observations have an absolute Rstudent value larger than 2

We modify the code above using proc reg instead

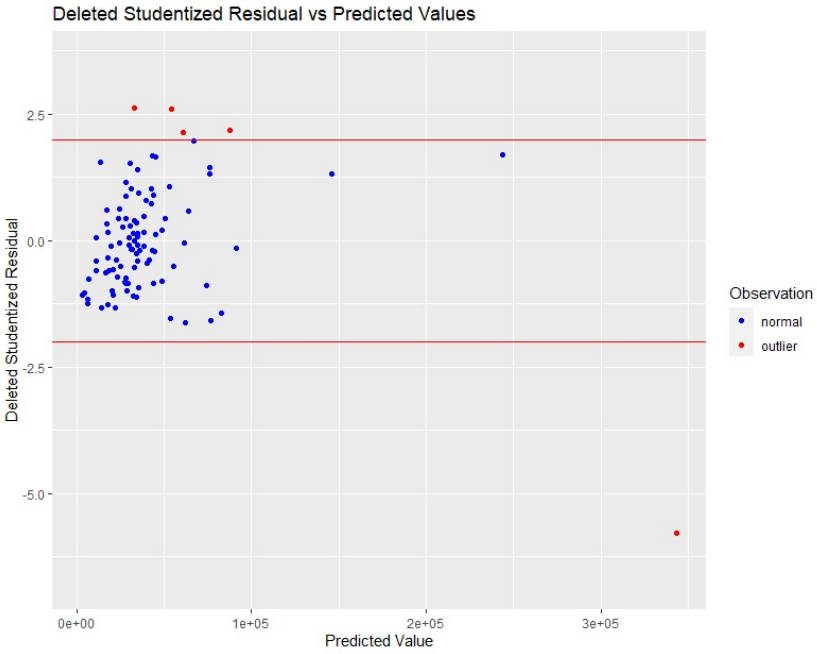
```
proc reg data=exam2023 plots(label)=all;
model church = H3020 H4058 H5020 H8090 / influence;
run;
```



We count 5 obervations

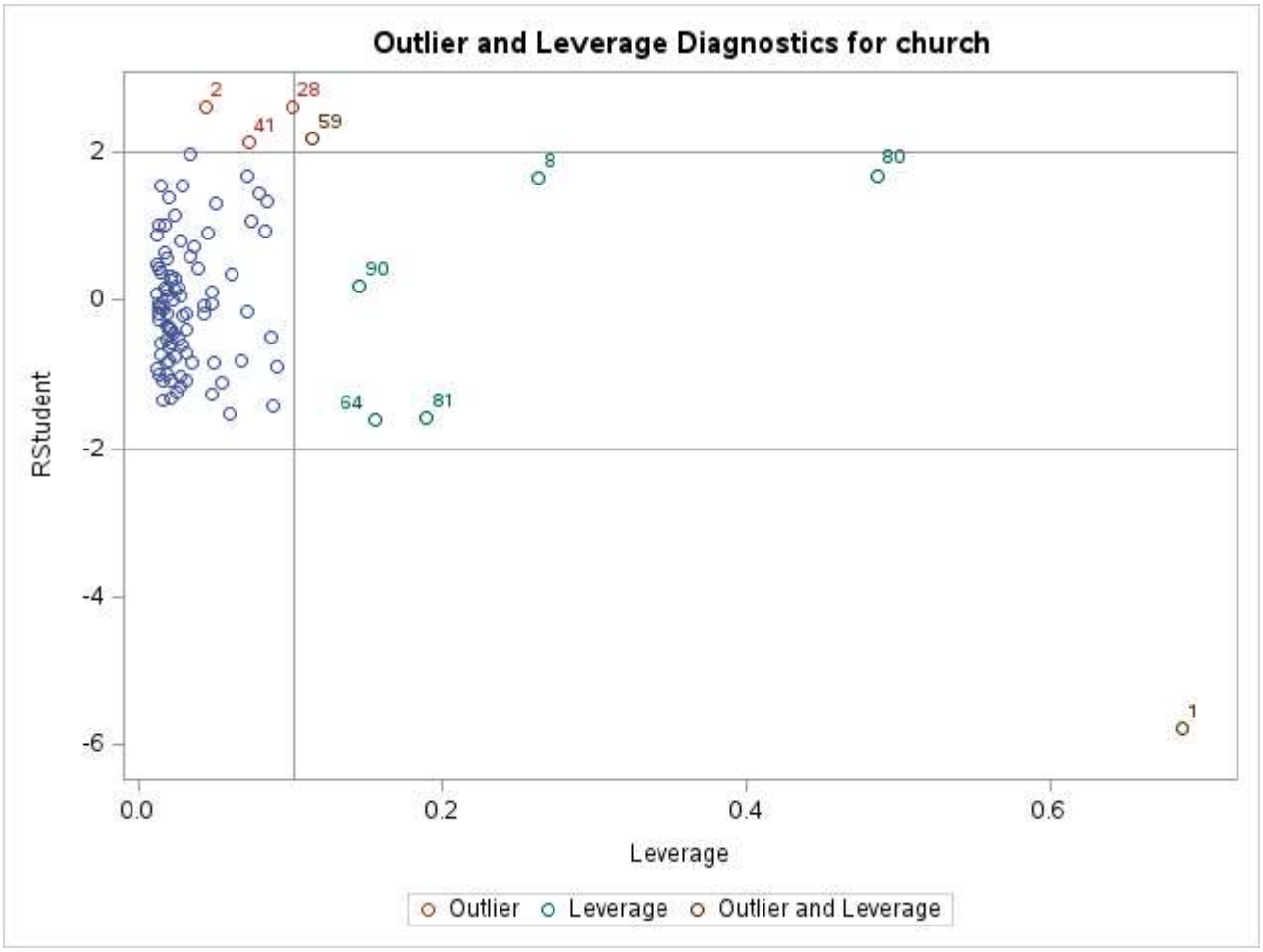
# Plotted to resemble the outputs from SAS:
Diag = olsrr::ols_plot_diagnostics(lmfit,print_plot = FALSE)
ggpubr::ggarrange(Diag$plot_1,Diag$plot_2,Diag$plot_3,
Diag$plot_4,pred,Diag$plot_6, Diag$plot_9,
Diag$plot_7,Diag$plot_8, ncol=3, nrow=3)

Q1.2
Diag$plot_2



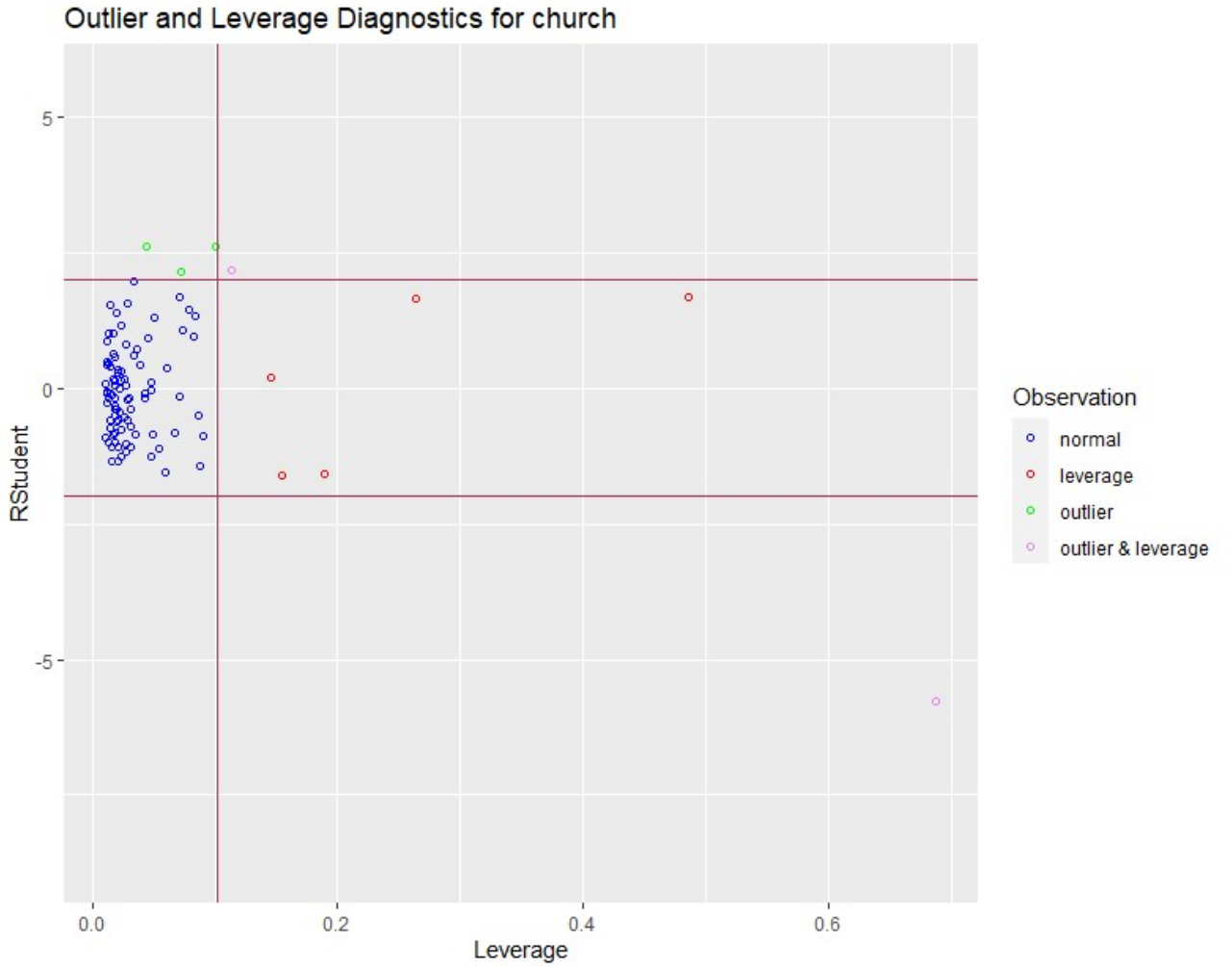
Question 1.3

The number of influential observations in the model, i.e., observations with both anabsolute RSTUDENT larger than 2 and a leverage larger than 0.102 are:



We count 2

Q1.3
Diag$plot_3



We count 2

Question 1.4

What observation – if removed – would change the estimated parameter vector the most(measured by an appropriate confidence region)

We have from the book p 204-205

Cook's D

A confidence region for the parameter θ is all the vectors θ^*, which satisfy

$$\frac{1}{p\hat{\sigma}^2}(\hat{\theta}-\theta^*)^T\mathbf{x}^T\mathbf{x}(\hat{\theta}-\theta^*) \leq F(p, n-p)_{1-\alpha}.$$

We use the left hand side as a measure of the distance between the parameter vector and $\hat{\theta}$. We let $\hat{\theta}(i)$ be the estimate, which corresponds to the deletion of the i'th observation

$$\mathbf{y}(i) = (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n)^T$$

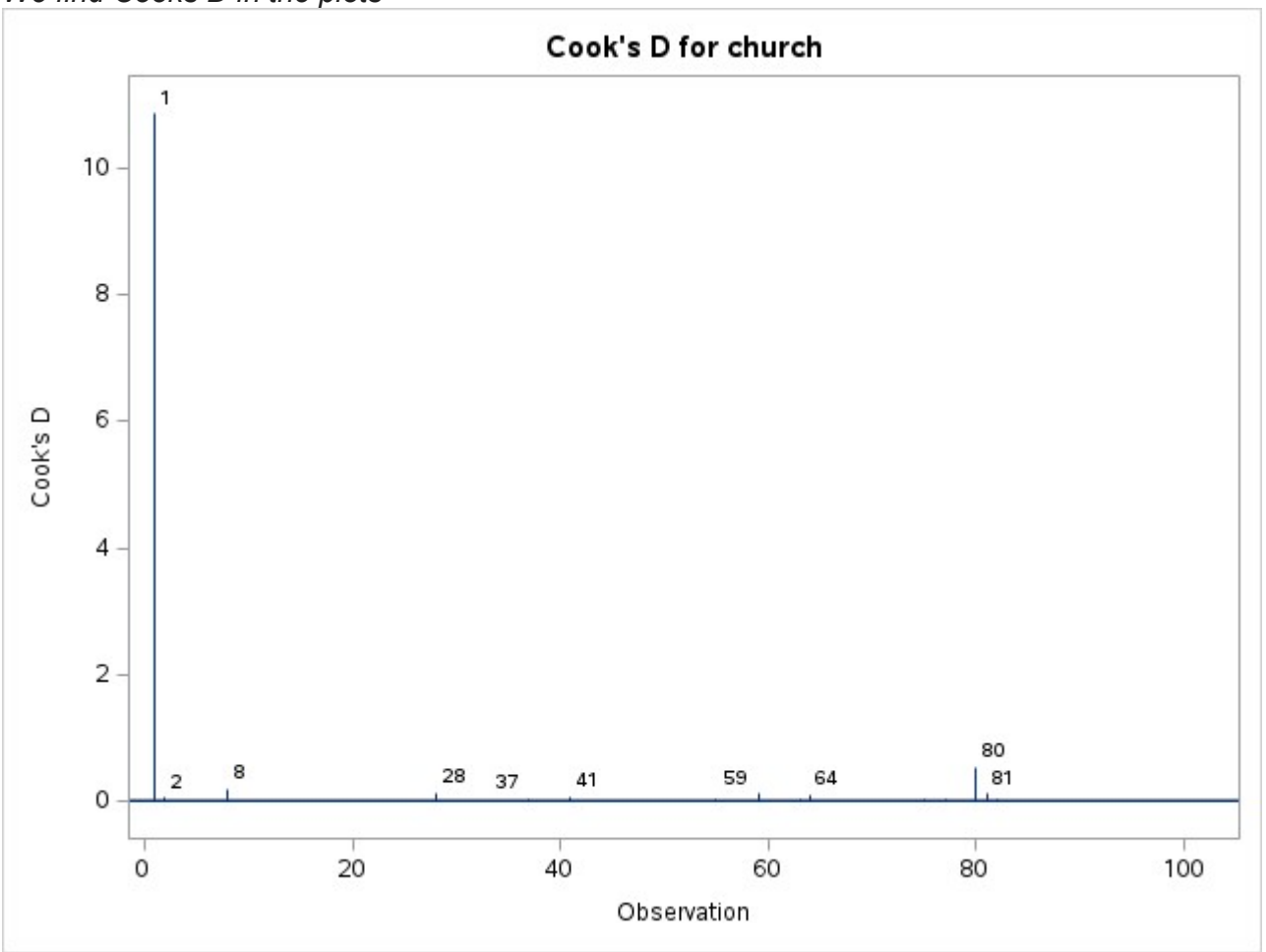
and therefore have

$$\hat{\theta}(i) = [\mathbf{x}(i)^T\mathbf{x}(i)]^{-1}\mathbf{x}(i)^T\mathbf{y}(i).$$

Cook's D then equals

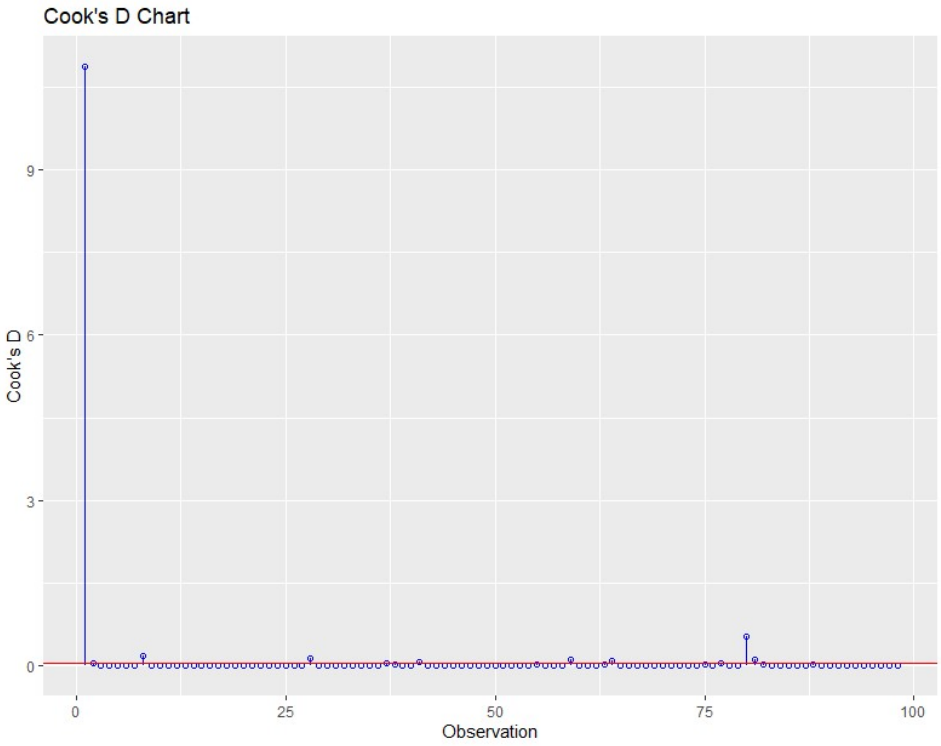
$$\frac{1}{p\hat{\sigma}^2}(\hat{\theta} - \hat{\theta}(i))^T\mathbf{x}^T\mathbf{x}(\hat{\theta} - \hat{\theta}(i)).$$

We find Cooks D in the plots



Observation 1

Q1.4
Diag$plot_6


Cook's D Chart
Cook's D
9
6
3
0
0
25
50
75
100
Observation

Question 1.5
We now consider if there is multicollinearity present in the model:

We again modify the code

```
proc reg data=exam2023 plots(label)=all;
model church = H3020 H4058 H5020 H8090 / influence vif tol;
run;
```

And get

The REG Procedure
Model: MODEL1
Dependent Variable: church

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	3224.84206	545.75487	5.91	<.0001	.	0
H3020	1	22.58819	2.47020	9.14	<.0001	0.03727	26.83336
H4058	1	19.00561	1.92033	9.90	<.0001	0.07301	13.69691
H5020	1	3.64654	0.86273	4.23	<.0001	0.01784	56.06065
H8090	1	33.53105	3.39951	9.86	<.0001	0.11621	8.60528

```
# Q1.5
# Calculate VIF for each variable
vif_values <- car::vif(lmfit)
# Create a data frame with variable names and VIF values
independent_variables <- names(coefficients(lmfit)[-1]) # Exclude the intercept
vif_df <- data.frame(Variable = independent_variables, VIF = vif_values)
# Calculate Tolerance as the reciprocal of VIF
vif_df$Tolerance <- 1 / vif_df$VIF
# Print the VIF and Tolerance values
print(vif_df)
```

```
      Variable       VIF  Tolerance
H3020    H3020 26.833364 0.03726704
H4058    H4058 13.696907 0.07300918
H5020    H5020 56.060649 0.01783782
H8090    H8090  8.605281 0.1162077
```

Yes, there are several independent variables with a VIF smaller than 10 and a Tolerance larger than0.1
No, all independent variables have a VIF smaller than 10 and a Tolerance larger than 0.1
No, all the parameters are significantly different from zero, as shown by the t-tests. We thus havecertain estimates and no signs of multicollinearity.
No, since at least one variable has a Tolerance larger than 0.1 and a VIF smaller than 10
Don't know
Yes, there are several independent variables with a VIF larger than 10 and a Tolerance less than 0.1

Question 1.6
When performing a backwards elimination, the first variable to be eliminated has an F-value of:

We can use the table from the previous question:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	3224.84206	545.75487	5.91	<.0001	.	0
H3020	1	22.58819	2.47020	9.14	<.0001	0.03727	26.83336
H4058	1	19.00561	1.92033	9.90	<.0001	0.07301	13.69691
H5020	1	3.64654	0.86273	4.23	<.0001	0.01784	56.06065
H8090	1	33.53105	3.39951	9.86	<.0001	0.11621	8.60528

# Q1.6
summary(lmfit)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3224.8421   545.7549   5.909 5.62e-08 ***
H3020         22.5882     2.4702   9.144 1.32e-14 ***
H4058         19.0056     1.9203   9.897 3.38e-16 ***
H5020          3.6465     0.8627   4.227 5.53e-05 ***
H8090         33.5310     3.3995   9.863 3.98e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Although we only have t-values here, we have from the book

Remark 2.25

The test given in Theorem 2.23 is of course equivalent to the ordinary F-test presented in Theorem 2.21 for $k - r = 1$. This may be established directly using the fact that the square of a t-distributed random variable is F-distributed, mnemonically written

$$[t(f)]^2 = F(1, f).$$

The advantage by using Theorem 2.23 is that we with obvious modifications may test one-sided hypothesis like $H_0: \theta_{i_0} \leq c$ against $H_1: \theta_{i_0} > c$. This is not possible with the F-test.

*The first one to be eliminated has the smallest t-value (and thus smallest F-value). We find in the table 4.23, which we square 4.23*4.23 = 17.8929*

Question 1.7

Considering only the four independent variables in the model. When performing a forward selection, the first variable to be included is:

We modify the code again

```
proc reg data=exam2023 plots(label)=all;
model church = H3020 H4058 H5020 H8090 / selection=forward;
run;
```

and get

Forward Selection: Step 1
Variable H5020 Entered: R-Square = 0.9829 and C(p) = 231.4741

```
# Q1.7
Fwd = olsrr::ols_step_forward_p(lmfit,penter=0.500)
Fwd
```

```
Selection Summary
--------------------------------------------------------------------------------
        Variable                    Adj.
Step    Entered     R-Square    R-Square      C(p)          AIC          RMSE
--------------------------------------------------------------------------------
   1    H5020          0.9829      0.9828    231.4741    1975.6479    5656.6942
   2    H4058          0.9897      0.9894    105.2928    1928.5897    4427.2488
   3    H8090          0.9907      0.9905     86.6175    1919.7408    4211.0760
   4    H3020          0.9951      0.9949      5.0000    1858.8849    3072.1336
--------------------------------------------------------------------------------
```

*Note: "*When all effects are variables (that is, effects have one degree of freedom and no hierarchy), using ADJRSQ, AIC, AICC, BIC, CP, RSQUARE, or SBC as the selection criterion for forward selection produces the same sequence of additions*"*

Question 1.8

Irregardless of the previous question we now consider two models:

$$\text{M1: } Church = \alpha + \beta_1 \cdot H3020 + \beta_2 \cdot H4058 + \beta_3 \cdot H5020 + \beta_4 \cdot H8090 + \epsilon,$$

$$\text{M2: } Church = \tilde{\alpha} + \tilde{\beta}_1 \cdot H3020 + \tilde{\epsilon},$$

The usual test statistic for testing for a simpler model is:

We find in the book page 129

Test statistic for $H_0: E(\boldsymbol{Y}) \in H$ against $H_1: E(\boldsymbol{Y}) \in M \backslash H$:

$$\frac{\|p_M(Y) - p_H(Y)\|^2/(k-r)}{\|Y - p_M(Y)\|^2/(n-k)} = \frac{(SS_{res}(Hyp) - SS_{res}(Mod))/(DF_{res}(Hyp) - DF_{res}(Mod))}{SS_{res}(Mod)/DF_{res}(Mod)}$$

Figure 2.4

We use proc glm to fit the two models

```
title 'M1';
proc glm data=exam2023;
model church = H3020 H4058 H5020 H8090;
run;
title 'M2';
proc glm data=exam2023;
model church = H3020;
run;
```

M1

The GLM Procedure

Dependent Variable: church

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	179243679045	44810919761	4747.92	<.0001
Error	93	877734460.55	9438004.9522		
Corrected Total	97	180121413506			

M2

The GLM Procedure

Dependent Variable: church

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	164660712988	164660712988	1022.43	<.0001
Error	96	15460700518	161048963.73		
Corrected Total	97	180121413506			

```
# Q1.8
lmreduced = lm(church ~ H3020, data = exam2023)

M1SS_error <- sum((exam2023$church - predict(lmfit))^2)
M1SS_model <- sum((predict(lmfit) - mean(exam2023$church))^2)
M2SS_error <- sum((exam2023$church - predict(lmreduced))^2)
M2SS_model <- sum((predict(lmreduced) - mean(exam2023$church))^2)

M1SS_error
M1SS_model
M2SS_error
M2SS_model
```

```
> M1SS_error
[1] 877734461
> M1SS_model
[1] 179243679045
> M2SS_error
[1] 15460700518
> M2SS_model
[1] 164660712988
```

((15460700518 - 877734461)/(96 - 93)) / (877734461 / 93) = 515.04409

PROBLEM 2 - Education and municipal compensation

We still use the dataset from the exam description.

We will now look at the relationship between the number of people with a long tertiary (i.e. Masters orequivalent) education in each municipality and whether a given municipality is a net donor or recipient inthe municipal tax equilisation scheme (kommunal udligning).

We will investigate this by means of a Linear Discriminant Analysis, where we have the class variable
gives -R for Recieves and G for Gives –
and the educational variables:
H70 H7020 H7025 H7030 H7035 H7039 H7059H7075 H7080 H7090 H7095

H70 Total (note that the total does not equal the sum of the educations below, as some categories havebeen omitted)
H7020 Educational
H7025 Humanities and Theological
H7030 Artistic
H7035 Science
H7039 Social Science
H7059 Technical Science
H7075 Food, Bio- and Laboratory Technology
H7080 Agriculture, Nature and Environment
H7090 Health Science
H7095 Police and Defence, etc.

We will use equal priors and losses.

Question 2.1
The number of resubstitution misclassifications in the model are:

We use proc discrim

```
proc discrim data=exam2023;
class gives;
var H70 H7020 H7025 H7030 H7035 H7039 H7059 H7075 H7080 H7090 H7095;
run;
```

and find in the output

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.EXAM2023
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into gives			
From gives	**G**	**R**	**Total**
G	7 87.50	1 12.50	8 100.00
R	0 0.00	90 100.00	90 100.00
Total	7 7.14	91 92.86	98 100.00
Priors	0.5	0.5	

We count one misclassification

```
# Q2.1 ================================
# Prior probabilities for the classes
# SAS by default has equal prior probabilities, so we need equal prior in R to receive the same results
different_classes <- unique(exam2023$gives)
num_classes <- length(different_classes)
prior <- rep(1/num_classes, num_classes)
#linear discriminant analysis
# Define the classes (mfr)
exam2023$class <- as.factor(exam2023$gives)
# Define the variables for the analysis
variables <- c("H70", "H7020", "H7025", "H7030", "H7035", "H7039", "H7059", "H7075", "H7080", "H7090", "H7095")
# Perform Linear Discriminant Analysis
z <- MASS::lda(class ~ ., data = exam2023[, c("class", variables)],prior=prior)
Class_Level_Information = data.frame("Frequency" = z$counts,"Proportion"=z$counts/z$N,"Prior"=z$prior)
print("Class Level Information:")
Class_Level_Information

n <- nrow(exam2023)
Classes <- nlevels(exam2023$gives)
paste0("DF Within Classes = ",n-Classes)
paste0("DF Between Classes = ",Classes-1)
zpred <- predict(z)
#Confusion Matrix:
print("Confusion Matrix:")
xtabs(~exam2023$gives+zpred$class)
```

```
zpred$class
exam2023$gives  G  R
```

G	7	1
R	0	90

Question 2.2

We now test whether the variables H7020 H7025 H7030 can be omitted from the model.
Using the usual test statistic we get a p-value in the range of:

We use

Theorem 5.21

The critical region for testing the hypothesis that the last $p-q$ variables do not contribute to the discrimination between the populations π_1 and π_2, i.e. the hypothesis that $\Delta^2_{(2|1)} = 0$ against all alternatives is

$$\left\{x_{11}, \cdots, x_{2n_2} \;\middle|\; \frac{n_1+n_2-p-1}{p-q} \frac{d^2-d_1^2}{(n_1+n_2)(n_1+n_2-2)/(n_1n_2)+d_1^2} > F(p-q, n_1+n_2-p-1)_{1-\alpha}\right\}$$

Here d^2 and d_1^2 are the observed values of D^2 and D_1^2.

We thus need to find Mahalanobis' distance for the full and reduced model
We modify our code

```
title 'Full model';
proc discrim data=exam2023;
class gives;
var H70 H7020 H7025 H7030 H7035 H7039 H7059 H7075 H7080 H7090 H7095;
run;
title 'Reduced model';
proc discrim data=exam2023;
class gives;
var H70 H7035 H7039 H7059 H7075 H7080 H7090 H7095;
run;
```

Full model

The DISCRIM Procedure

Generalized Squared Distance to gives

From gives	G	R
G	0	32.13320
R	32.13320	0

Reduced model

The DISCRIM Procedure

Generalized Squared Distance to gives		
From gives	G	R
G	0	24.91181
R	24.91181	0

We further find

Number of Observations Read	98
Number of Observations Used	98

We collect
p = 11
q = 8
n1 = 8
n2 = 90
D_full = 32.13320
D_reduced = 24.91181
DF1 = p-q = 3
DF2 = n1 + n2 –p -1 = 86

We now insert into the theorem

FVAL = ((n1+n2-p-1)/(p-q))((D_full – D_reduced) / (((n1+n2)(n1 + n2 -2)) / (n1*n2) + D_reduced) = 5.4508*

And test in F(3, 86)

```
data test;
Q2_3 = 1- cdf('F',5.4508,3,86);
run;
```

0.001769613

[0.001; 0.005[

```
# Q2.2 ================================
pcov <- Rfast::pooled.cov(as.matrix(exam2023[,variables]),exam2023$class)
Means <- as.matrix(z$means)
invCov <- solve(pcov)
# Extract unique levels from exam2023$gives
unique_levels <- levels(exam2023$gives)
num_col <- length(unique(exam2023$gives))
# Create an empty matrix to store the Mahalanobis distances with the equal priors
maha <- matrix(c(rep(0, num_col^2)), ncol = num_col)
```

```
# Define the names for rows and columns (assuming unique_levels contains the names)
rownames(maha) <- unique_levels
colnames(maha) <- unique_levels
for (i in 1:num_col) {
  for (j in 1:num_col) {
    mu <- Means[i, ] - Means[j, ]
    maha[i, j] <- mu %*% invCov %*% mu
  }
}
# squared Mahalanobis distances between the 6 mfr types.
# maha : Assuming equal priors
print("Generalized Squared Distance to species (equal priors):")
maha


# Define the variables for the analysis
variables_reduced <- c("H70", "H7035", "H7039", "H7059", "H7075", "H7080", "H7090",
"H7095")
# Perform Linear Discriminant Analysis
z_reduced <- MASS::lda(class ~ ., data = exam2023[, c("class",
variables_reduced)],prior=prior)
pcov_reduced <-
Rfast::pooled.cov(as.matrix(exam2023[,variables_reduced]),exam2023$class)
Means_reduced <- as.matrix(z_reduced$means)
invCov <- solve(pcov_reduced)
# Extract unique levels from breakfast$mfr
unique_levels <- levels(exam2023$gives)
num_col <- length(unique(exam2023$gives))
# Create an empty matrix to store the Mahalanobis distances with the equal priors
maha_reduced <- matrix(c(rep(0, num_col^2)), ncol = num_col)
# Define the names for rows and columns (assuming unique_levels contains the names)
rownames(maha_reduced) <- unique_levels
colnames(maha_reduced) <- unique_levels
for (i in 1:num_col) {
  for (j in 1:num_col) {
    mu <- Means_reduced[i, ] - Means_reduced[j, ]
    maha_reduced[i, j] <- mu %*% invCov %*% mu
  }
}
# squared Mahalanobis distances between the 6 mfr types.
# maha : Assuming equal priors
print("Generalized Squared Distance to species (equal priors):")
maha_reduced

p <- 11
q <- 8
n1 <- 8
n2 <- 90
D_full <- maha[1,2]
D_reduced <- maha_reduced[1,2]
```

```
DF1 <- p-q
DF2 <- (n1+n2-p-1)
FVAL <- ((n1+n2-p-1)/(p-q)) * ( (D_full - D_reduced) / ( ( (n1+n2)*(n1 + n2 -2) ) / (n1*n2) + D_reduced))
pval = pf(FVAL, DF1, DF2, lower.tail = FALSE)
pval
```

```
0.001769614
```

Question 2.3

We now consider the full model and change the priors to be 0.1 for the 'G' class and 0.9 for the 'R' class. The squared generalized distance from the mean of group G to itself is :

Can be solved without SAS or R. We use

Definition 5.15

Assuming that the hypothesis $H_0 : \Sigma_1 = \cdots = \Sigma_k$ is true, we define *the squared generalized distance from* $\widehat{\mu}_j$ *to population* π_i as

$$D_i^2(\widehat{\mu}_j) = \begin{cases} \left(\widehat{\mu}_i - \widehat{\mu}_j\right)^T \widehat{\Sigma}^{-1} \left(\widehat{\mu}_i - \widehat{\mu}_j\right) & \text{if the priors are equal} \\ \left(\widehat{\mu}_i - \widehat{\mu}_j\right)^T \widehat{\Sigma}^{-1} \left(\widehat{\mu}_i - \widehat{\mu}_j\right) - 2\log p_i & \text{if the priors are not all equal} \end{cases}$$

If the hypothesis is *not* true, we define *the squared generalized distance from* $\widehat{\mu}_j$ *to population* π_i as

$$D_i^2(\widehat{\mu}_j) = \begin{cases} \left(\widehat{\mu}_i - \widehat{\mu}_j\right)^T \widehat{\Sigma}_i^{-1} \left(\widehat{\mu}_i - \widehat{\mu}_j\right) & \text{if the priors are equal} \\ \left(\widehat{\mu}_i - \widehat{\mu}_j\right)^T \widehat{\Sigma}_i^{-1} \left(\widehat{\mu}_i - \widehat{\mu}_j\right) + \log\det\widehat{\Sigma}_i - 2\log p_i & \text{if the priors are not all equal} \end{cases}$$

Since we use LDA, we assume that the dispersion for both groups are equal.
*We thus only need to calculate -2*log(P_G), where P_G is the prior fro group G.*

*-2*log(P_G) = -2*log(0.1) = 4.6052*

We confirm using SAS

The DISCRIM Procedure

Generalized Squared Distance to gives

From gives	G	R
G	4.60517	32.34392

Generalized Squared Distance to gives		
From gives	G	R
R	36.73837	0.21072

And using R

```
different_classes <- unique(exam2023$gives)
num_classes <- length(different_classes)
prior <- c(0.9, 0.1)
#linear discriminant analysis
# Define the classes (mfr)
exam2023$class <- as.factor(exam2023$gives)
# Define the variables for the analysis
variables <- c("H70", "H7020", "H7025", "H7030", "H7035", "H7039", "H7059", "H7075", "H7080", "H7090",
"H7095")
# Perform Linear Discriminant Analysis
z <- MASS::lda(class ~ ., data = exam2023[, c("class", variables)],prior=prior)
Class_Level_Information = data.frame("Frequency" = z$counts,"Proportion"=z$counts/z$N,"Prior"=z$prior)
print("Class Level Information:")
Class_Level_Information

pcov <- Rfast::pooled.cov(as.matrix(exam2023[,variables]),exam2023$class)
Means <- as.matrix(z$means)
invCov <- solve(pcov)
# Extract unique levels from exam2023$gives
unique_levels <- levels(exam2023$gives)
num_col <- length(unique(exam2023$gives))
# Create an empty matrix to store the Mahalanobis distances with the unequal priors
maha <- matrix(c(rep(0, num_col^2)), ncol = num_col)
# Define the names for rows and columns (assuming unique_levels contains the names)
rownames(maha) <- unique_levels
colnames(maha) <- unique_levels
for (i in 1:num_col) {
  for (j in 1:num_col) {
    mu <- Means[i, ] - Means[j, ]
    maha[i, j] <- mu %*% invCov %*% mu -2*log(prior[i])
  }
}
# squared Mahalanobis distances between the 6 mfr types.
# maha : Assuming equal priors
print("Generalized Squared Distance to species (equal priors):")
maha
```

```
          [,1]     [,2]
[1,]  0.210721 32.34392
[2,] 36.738367  4.60517
```

PROBLEM 3 - Education and correlation

We still consider the dataset from the exam description.
We will now investigate the correlation between the variables: H7035 H7095

Question 3.1

The fraction of variance of H7035 that can be explained by H7095 is:

We have from page 29

and the *squared coefficient of correlation represents the reduction in variance. i.e. the fraction of Y's variance, which can be explained by X,* since

$$\rho^2 = \frac{\mathrm{V}(Y) - \mathrm{V}(Y|X = x)}{\mathrm{V}(Y)}.$$

We thus need to find the squared correlation

We use the following script

```
proc corr data=exam2023;
var H7035 H7095;
run;
```

The CORR Procedure

2 Variables:	H7035 H7095

Pearson Correlation Coefficients, N = 98
Prob > |r| under H0: Rho=0

	H7035	**H7095**
H7035	1.00000	0.69681 <.0001
H7095	0.69681 <.0001	1.00000

We take the correlation and square it
0.69681*0.69681 = 0.4855

# Q3.1
cor.test(exam2023$H7035, exam2023$H7095)

Question 3.2
The usual test statistic for the correlation between H7035 and H7095 being different from zero is:

We use

Theorem 1.37

Let $R = R_{ij|m+1\ldots p}$ be the empirical partial correlation coefficient between Z_i and Z_j conditioned on (or: for given) $Z_{m+1,\ldots,Z_p}$. It is assumed to be computed from the unbiased estimates of the variance-covariance matrix and from n observations. Then

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2-(p-m)} \sim \mathrm{t}(n-2-(p-m)),$$

if $\rho_{ij|m+1,\ldots,p} = 0$.

Proof omitted

Remark 1.38

The number $(p-m)$ is the number of variables which are fixed (conditioned upon). The degrees of freedom are therefore equal to the number of observations minus 2 minus the number of fixed variables. *The theorem is also valid if $p-m=0$ i.e. if we have the case of an unconditional correlation coefficient.*

We have
(p-m) = 0
R = 0.69681
R^2 = 0.4855
n =98

and insert

$$\frac{0.69681}{\sqrt{1-0.69681^2}}\sqrt{98-2} = 9.5187$$

Question 3.3
The 90% confidence interval for the correlation between H7035 and H7095 is:

We modify our SAS-script

```
proc corr data=exam2023 fisher(alpha=0.1);
var H7035 H7095;
run;
```

The CORR Procedure

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	90% Confidence Limits		p Value for H0:Rho=0
H7035	H7095	98	0.69681	0.86107	0.00359	0.69495	0.597157	0.772393	<.0001

[0.60; 0.77]

```
# Q3.3

cor.test(exam2023$H7035, exam2023$H7095, conf.level=0.9)

        Pearson's product-moment correlation

data:  exam2023$H7035 and exam2023$H7095
t = 9.5186, df = 96, p-value = 1.61e-15
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.5994632 0.7738384
sample estimates:
      cor
0.6968072
```

PROBLEM 4 - Education and sport

We will now investigate the relation between long tertiary education and locomotive sports, by means of aCanonical Correlation analysis, and use the dataset from the exam description.

We study the relation between the education variables
H7020 H7025 H7030 H7035 H7039 H7059 H7075 H7080 H7090 H7095
and the sports variables
Cycling CanoeKayak Gliding Hanggliding HorseRiding Sailing Skiing SurfRafting

Question 4.1

The number of significant canonical correlations (at level 5%) are:

We use the following SAS script

```
Proc cancorr data=exam2023;
var H7020 H7025 H7030 H7035 H7039 H7059 H7075 H7080 H7090 H7095;
with Cycling CanoeKayak Gliding Hanggliding HorseRiding Sailing Skiing
SurfRafting;
run;
```

and find in the output

The CANCORR Procedure

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.930450	0.918138	0.013632	0.865738	6.4481	4.2576	0.6566	0.6566	0.01543133	6.00	80	515.96	<.0001
2	0.828596	0.802072	0.031824	0.686572	2.1905	1.6930	0.2231	0.8797	0.11493437	3.44	63	462.3	<.0001
3	0.576387	0.428491	0.067803	0.332222	0.4975	0.0684	0.0507	0.9304	0.36670089	1.92	48	407.54	0.0004
4	0.547943	.	0.071050	0.300242	0.4291	0.2867	0.0437	0.9741	0.54913575	1.54	35	351.58	0.0296

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
5	0.352969	0.242350	0.088885	0.124587	0.1423	0.0720	0.0145	0.9885	0.78475040	0.88	24	294.25	0.6269
6	0.256312	0.147744	0.094864	0.065696	0.0703	0.0311	0.0072	0.9957	0.89643441	0.63	15	235.05	0.8462
7	0.194217	.	0.097705	0.037720	0.0392	0.0363	0.0040	0.9997	0.95946733	0.45	8	172	0.8897
8	0.054063	-.105380	0.101238	0.002923	0.0029		0.0003	1.0000	0.99707723	0.09	3	87	0.9681

4 are significant

In R

```
varY <- c("H7020", "H7025", "H7030", "H7035", "H7039", "H7059", "H7075", "H7080",
"H7090", "H7095")
varX <- c("Cycling", "CanoeKayak", "Gliding", "Hanggliding", "HorseRiding", "Sailing", "Skiing",
"SurfRafting")

Exx = as.matrix(cor(exam2023[, c(varX)]))
Eyx = as.matrix(cor(exam2023[, c(varY)],exam2023[, c(varX)]))
Exy = as.matrix(cor(exam2023[, c(varX)],exam2023[, c(varY)]))
Eyy = as.matrix(cor(exam2023[, c(varY)]))

invExx = solve(Exx)
invEyy = solve(Eyy)

#canonical correlation:
Cancorr = geigen::geigen(Eyx%*%invExx%*%Exy,Eyy,symmetric = TRUE)
values = sort(Cancorr$values,decreasing = TRUE)

#E is the residual variation after having predicted Y by means of X
H = Eyx%*%invExx%*%Exy
E = Eyy - Eyx%*%invExx%*%Exy
#epsilon <- 1e-4  # A small positive value
#H <- H + epsilon * diag(dim(Eyx)[1])

eigen(H)

invE = solve(E)
```

```
Ev <- eigen(invE%*%H)
var = Ev$values

# Eigenvalues, Proportion and Cumulative proportion of Variance:
varPC <- var/sum(var)
cumu = c(1:length(var))
for (i in 1:length(var)){
  cumu[i] = sum(varPC[1:i])
}

results <- data.frame("CanCor" = sqrt(values),"Squared CanCor" =
values,"eigenvalues"=var,"proportion"=varPC,
                      "cumulative" = cumu)
results <- results[1:8,]

n = 98
p = length(Exx[,1])
q = length(Eyy[,1])

HypTest <- CCP::p.asym(results$CanCor,n,p,q,tstat="Wilks")
print("Table with information about the Canonical Correlations similar to the output in SAS:")
results_all <- data.frame(results,HypTest)
results_all
```

```
     CanCor Squared.CanCor eigenvalues   proportion cumulative    id       stat     approx df1      df2      p.value
1 0.9304504    0.865737883 6.448117349 0.6566329407  0.6566329 Wilks 0.01543133 6.00018681  80 515.9626 0.000000e+00
2 0.8285963    0.686571888 2.190524273 0.2230682721  0.8797012 Wilks 0.11493437 3.43664216  63 462.3045 1.731948e-14
3 0.5763868    0.332221791 0.497503193 0.0506623821  0.9303636 Wilks 0.36670089 1.92017196  48 407.5367 4.144694e-04
4 0.5479430    0.300241514 0.429064485 0.0436930439  0.9740566 Wilks 0.54913575 1.53834835  35 351.5793 2.959301e-02
5 0.3529687    0.124586924 0.142317869 0.0144926954  0.9885493 Wilks 0.78475040 0.88215828  24 294.2510 6.269456e-01
6 0.2563118    0.065695747 0.070315153 0.0071604227  0.9957098 Wilks 0.89643441 0.63305038  15 235.0490 8.462314e-01
7 0.1942168    0.037720153 0.039198735 0.0039917358  0.9997015 Wilks 0.95946733 0.44943595   8 172.0000 8.896726e-01
8 0.0540626    0.002922765 0.002931333 0.0002985072  1.0000000 Wilks 0.99707723 0.08500865   3  87.0000 9.680601e-01
```

4 are significant

Question 4.2

Based on the first set of canonical variables, the least liked sport by people with a long education is:

Question 4.3

Based on the Canonical Correlation Analysis, the two educations that favor Gliding themost are

We use the same script for both

4.2: HorseRiding

4.3: H7035 and H7090

The CANCORR Procedure

Canonical Structure

Correlations Between the VAR Variables and Their Canonical Variables

	V1	V2	V3	V4	V5	V6	V7	V8
H7020	0.8984	0.2585	0.1014	0.1028	-0.1376	-0.0891	0.1988	0.0477
H7025	0.9216	0.2186	0.1275	0.1066	-0.1043	-0.0568	0.2076	0.0537
H7030	0.9343	0.1653	0.1584	0.1161	-0.1152	0.0185	0.1787	0.0846
H7035	0.8791	0.3144	0.0877	0.1557	-0.0967	-0.1132	0.2199	0.0405
H7039	0.8949	0.1410	0.1893	0.1818	-0.1170	-0.0331	0.2563	0.1105
H7059	0.8757	0.1489	0.1054	0.2701	-0.1447	-0.0207	0.2506	0.0488
H7075	0.9219	-0.0115	0.1176	0.2254	-0.1134	-0.0203	0.2240	0.0836
H7080	0.9095	-0.0464	0.1995	0.1707	-0.1559	-0.0199	0.2087	-0.0246
H7090	0.8013	0.3447	0.1581	0.1855	-0.0964	-0.1895	0.2380	0.0780
H7095	0.6164	0.0301	0.2334	0.0866	0.1946	0.2233	0.3408	-0.0424

Correlations Between the WITH Variables and Their Canonical Variables

	W1	W2	W3	W4	W5	W6	W7	W8
Cycling	0.8485	0.1337	0.2683	0.1288	0.0553	0.1160	-0.1940	-0.3457
CanoeKayak	0.5452	-0.0170	0.3869	0.6197	-0.2783	0.1379	-0.2576	-0.0761
Gliding	0.1453	0.9734	-0.1572	0.0343	0.0422	0.0355	-0.0464	-0.0168
Hanggliding	0.7528	0.0446	0.1071	0.1092	-0.0431	0.4032	0.4920	0.0380
HorseRiding	-0.1856	0.2335	0.6913	0.2133	0.3240	0.1520	-0.1601	-0.4837
Sailing	0.6568	-0.0220	0.6617	0.2614	-0.1729	-0.0851	-0.1526	0.0396
Skiing	0.5040	-0.1684	-0.1047	0.6129	0.4649	-0.3159	0.0926	0.0811
SurfRafting	0.3434	-0.0335	0.3440	0.1199	0.2331	0.4819	-0.5782	0.3569

Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables

	W1	W2	W3	W4	W5	W6	W7	W8
H7020	0.8359	0.2142	0.0584	0.0563	-0.0486	-0.0228	0.0386	0.0026
H7025	0.8575	0.1811	0.0735	0.0584	-0.0368	-0.0146	0.0403	0.0029
H7030	0.8693	0.1369	0.0913	0.0636	-0.0407	0.0047	0.0347	0.0046
H7035	0.8180	0.2605	0.0505	0.0853	-0.0341	-0.0290	0.0427	0.0022
H7039	0.8327	0.1168	0.1091	0.0996	-0.0413	-0.0085	0.0498	0.0060
H7059	0.8148	0.1234	0.0607	0.1480	-0.0511	-0.0053	0.0487	0.0026
H7075	0.8578	-0.0095	0.0678	0.1235	-0.0400	-0.0052	0.0435	0.0045
H7080	0.8463	-0.0384	0.1150	0.0935	-0.0550	-0.0051	0.0405	-0.0013
H7090	0.7455	0.2856	0.0911	0.1016	-0.0340	-0.0486	0.0462	0.0042
H7095	0.5735	0.0249	0.1345	0.0474	0.0687	0.0572	0.0662	-0.0023

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables

	V1	V2	V3	V4	V5	V6	V7	V8
Cycling	0.7895	0.1108	0.1546	0.0706	0.0195	0.0297	-0.0377	-0.0187
CanoeKayak	0.5073	-0.0141	0.2230	0.3396	-0.0982	0.0354	-0.0500	-0.0041
Gliding	0.1352	0.8066	-0.0906	0.0188	0.0149	0.0091	-0.0090	-0.0009

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables								
	V1	V2	V3	V4	V5	V6	V7	V8
Hanggliding	0.7004	0.0370	0.0617	0.0598	-0.0152	0.1034	0.0956	0.0021
HorseRiding	-0.1727	0.1935	0.3984	0.1169	0.1143	0.0390	-0.0311	-0.0262
Sailing	0.6111	-0.0182	0.3814	0.1432	-0.0610	-0.0218	-0.0296	0.0021
Skiing	0.4689	-0.1396	-0.0604	0.3358	0.1641	-0.0810	0.0180	0.0044
SurfRafting	0.3196	-0.0278	0.1982	0.0657	0.0823	0.1235	-0.1123	0.0193

```
# Correlations between the education variables (Eyy) and their Canonical variables
CorPanel = Eyy%*%Cancorr$vectors

#corresponding standardized canonical coefficients / correlations:
Coefficients = data.frame("Variables" = varY,
                  "CoefV1" = round(Cancorr$vectors[,10],2),
                  "CoefV2" = round(Cancorr$vectors[,9],2),
                  "CorrV1" = round(CorPanel[,10],2),
                  "CorrV2" = round(CorPanel[,9],2),
                  "Diff.between.Sq.Corr." = round(CorPanel[,10],1)^2-round(CorPanel[,9],1)^2)

print("Standardized Canonical Coefficients for the Panel Variables and Correlations Between the Panel
Variables and Their Canonical Variables:")
Coefficients

#for sports
Cancorr2 = geigen::geigen(Exy%*%invEyy%*%Eyx,Exx,symmetric = TRUE)

# Correlations between the Phys/Chem variables (Exx) and their Canonical variables
CorPhysChem = Exx%*%Cancorr2$vectors

#corresponding standardized canonical coefficients / correlations:
Coefficients2 = data.frame("Variables" = varX,
                   "CoefW1" = round(Cancorr2$vectors[,8],2),
                   "CoefW2" = round(Cancorr2$vectors[,7],2),
                   "CorrW1" = round(CorPhysChem[,8],2),
                   "CorrW2" = round(CorPhysChem[,7],2),
                   "Diff.between.Sq.Corr." = round(CorPhysChem[,8],1)^2-round(CorPhysChem[,7],1)^2)
print("Standardized Canonical Coefficients for the Phys/Chem Variables and Correlations Between the
Phys/Chem Variables and Their Canonical Variables")
Coefficients2
```

```
      Variables CoefV1 CoefV2 CorrV1 CorrV2 Diff.between.Sq.Corr.
H7020     H7020   1.73   1.17  -0.90   0.26                  0.72
H7025     H7025  -4.75 -11.64  -0.92   0.22                  0.77
H7030     H7030  -1.36   4.75  -0.93   0.17                  0.77
H7035     H7035   1.83   9.07  -0.88   0.31                  0.72
```

H7039	H7039	4.43	2.81	-0.89	0.14	0.80
H7059	H7059	0.16	-1.09	-0.88	0.15	0.80
H7075	H7075	-2.16	-3.50	-0.92	-0.01	0.81
H7080	H7080	0.37	0.41	-0.91	-0.05	0.81
H7090	H7090	-1.21	-1.90	-0.80	0.34	0.55
H7095	H7095	0.00	0.13	-0.62	0.03	0.36

	Variables	CoefW1	CoefW2	CorrW1	CorrW2	Diff.between.Sq.Corr.
Cycling	Cycling	-0.83	-0.26	-0.85	0.13	0.63
CanoeKayak	CanoeKayak	0.30	-0.08	-0.55	-0.02	0.25
Gliding	Gliding	-0.01	1.02	-0.15	0.97	-0.99
Hanggliding	Hanggliding	-0.14	0.03	-0.75	0.04	0.64
HorseRiding	HorseRiding	0.45	0.13	0.19	0.23	0.00
Sailing	Sailing	-0.29	0.27	-0.66	-0.02	0.49
Skiing	Skiing	-0.13	-0.07	-0.50	-0.17	0.21
SurfRafting	SurfRafting	-0.05	-0.06	-0.34	-0.03	0.09

PROBLEM 5 - Sports across municipalities

For this problem, you must rely on the embedded tables and figures when answering the questions!
We will continue our investigations of the sports variables by means of a factor analysis.
Cycling CanoeKayak Gliding Hanggliding HorseRiding Sailing Skiing SurfRafting

Question 5.1

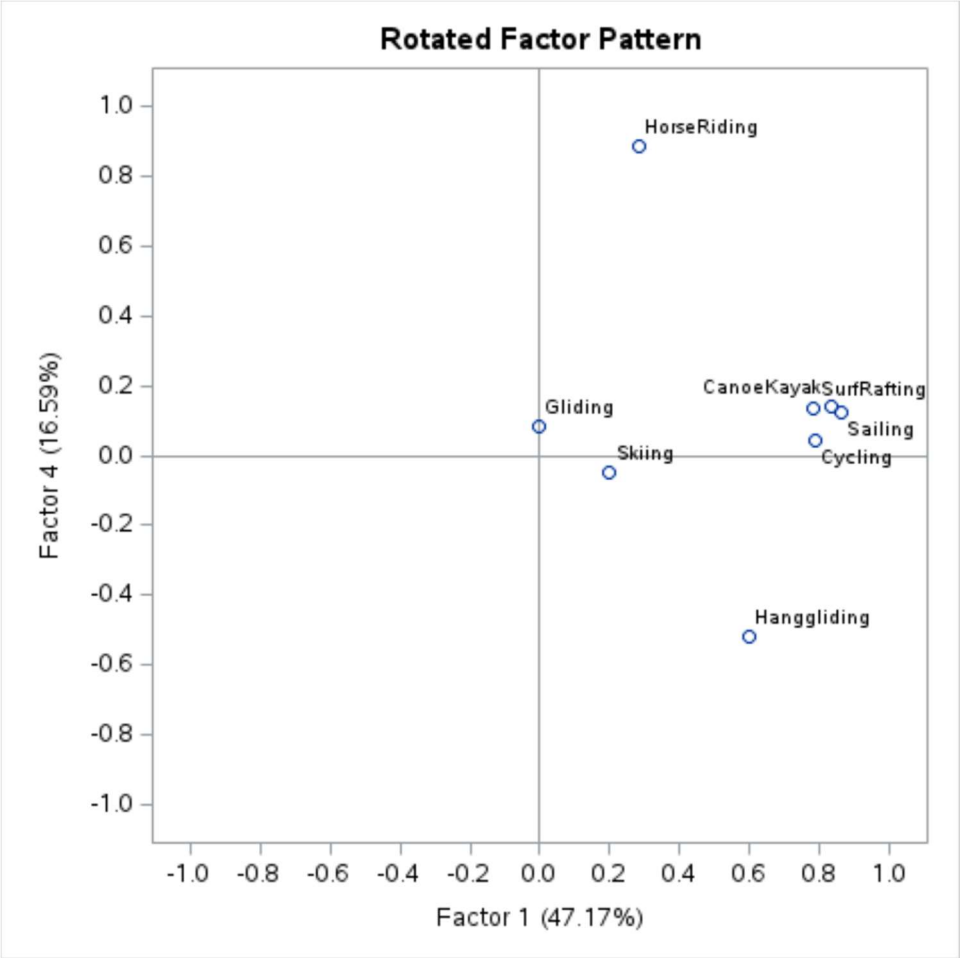
How many components must we include to account for 90% of the variance

Eigenvalue	Difference	Proportion	Cumulative
1	3.63522464	2.40142303	0.4544	0.4544
2	1.23380161	0.15264929	0.1542	0.6086
3	1.08115232	0.35973429	0.1351	0.7438
4	0.72141803	0.12339588	0.0902	0.8339
5	0.59802215	0.20522564	0.0748	0.9087
6	0.39279651	0.18852167	0.0491	0.9578
7	0.20427484	0.07096492	0.0255	0.9833
8	0.13330992	0.0167	1.0000

Five components/factors

Question 5.2

We now perform a factor analysis with 4 factors.
The 4th VARIMAX rotated factor, can be interpreted as



Mean of Cycling CanoeKayak Sailing SurfRafting, and a contrast between HorseRiding andHanggliding

Contrast between Cycling CanoeKayak Sailing SurfRafting and Gliding Skiing

Mean of all variables

Don't know

Mean of Cycling CanoeKayak Sailing SurfRafting

Contrast between HorseRiding and Hanggliding

PROBLEM 6

We now leave the exciting world of education and sports and consider something else entirely.

We consider independent random variables $Y_i \sim N(\mu_i, \sigma^2)$, organized in a two-way layout with expected values as presented in the table below.

	columns	
rows	$E(Y_1) = \mu + \alpha + \beta$	$E(Y_2) = \mu - \alpha + \beta$
	$\mathrm{E}(Y_3) = \mu - \alpha - \beta$	$E(Y_4) = \mu - \alpha + \beta$

In the sequel, you may find the following expressions useful

$$\begin{bmatrix} 4 & -2 & 2 \\ -2 & 4 & 0 \\ 2 & 0 & 4 \end{bmatrix}^{-1} = \frac{1}{8}\begin{bmatrix} 4 & 2 & -2 \\ 2 & 3 & -1 \\ -2 & -1 & 3 \end{bmatrix}$$

$$\frac{1}{8}\begin{bmatrix} 4 & 2 & -2 \\ 2 & 3 & -1 \\ -2 & -1 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 2 & 0 & 2 & 0 \\ 2 & -1 & 0 & -1 \\ 0 & 1 & -2 & 1 \end{bmatrix}$$

Question 6.1

The ordinary least squares estimator $\hat{\beta}$ of β is of the form $aY_1 + bY_2 + cY_3 + dY_4$, where $\begin{bmatrix} a & b & c & d \end{bmatrix}$ is:

We have the least squares estimate given in the problem. We can read directly from that
*¼ * [0 1 -2 1]*

Question 6.2

The variance of $\hat{\beta}$ is:

We have

Theorem 2.3

Let $\mathbf{x}$ and $\boldsymbol{\theta}$ be given as in the preceding section and let $\boldsymbol{Y} \sim \mathrm{N}_n(\mathbf{x}\,\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is given by $\mathbf{x}\hat{\boldsymbol{\theta}}$ being the projection (with respect to $\boldsymbol{\Sigma}$) onto M, $\hat{\boldsymbol{\theta}}$ is a solution to the so-called *normal equation(s)*

$$(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\hat{\boldsymbol{\theta}} = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}y.$$

If $\mathbf{x}$ has full rank k, then

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y},$$

and being a linear combination of normally distributed variables $\hat{\boldsymbol{\theta}}$ is also normally distributed with parameters

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{\theta} \\ \mathrm{D}(\hat{\boldsymbol{\theta}}) &= \sigma^2(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}. \end{aligned}$$

It is especially noted that $\hat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta}$.

Since we have independent observations with variance σ^2, we get

$$D\left(\begin{bmatrix}\hat{\mu}\\ \hat{\alpha}\\ \hat{\beta}\end{bmatrix}\right) = \sigma^2\frac{1}{8}\begin{bmatrix}4 & 2 & -2\\ 2 & 3 & -1\\ -2 & -1 & 3\end{bmatrix},$$

.

Question 6.3

The correlation between $\hat{\mu}$ and $\hat{\beta}$ is:

We have that

Theorem 2.3

Let $\mathbf{x}$ and $\boldsymbol{\theta}$ be given as in the preceding section and let $\boldsymbol{Y} \sim \mathrm{N}_n(\mathbf{x}\,\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is given by $\mathbf{x}\hat{\boldsymbol{\theta}}$ being the projection (with respect to $\boldsymbol{\Sigma}$) onto M, $\hat{\boldsymbol{\theta}}$ is a solution to the so-called *normal equation(s)*

$$(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\hat{\boldsymbol{\theta}} = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}y.$$

If $\mathbf{x}$ has full rank k, then

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y},$$

and being a linear combination of normally distributed variables $\hat{\boldsymbol{\theta}}$ is also normally distributed with parameters

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{\theta} \\ \mathrm{D}(\hat{\boldsymbol{\theta}}) &= \sigma^2(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}. \end{aligned}$$

It is especially noted that $\hat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta}$.

Since we have independent observations with variance σ^2, we get

$$D\left(\begin{bmatrix}\hat{\mu}\\ \hat{\alpha}\\ \hat{\beta}\end{bmatrix}\right) = \sigma^2\frac{1}{8}\begin{bmatrix}4 & 2 & -2\\ 2 & 3 & -1\\ -2 & -1 & 3\end{bmatrix},$$

Since this is not a correlation matrix we need to normalize it
We use from page 8

$$\rho_{ij} = \mathrm{Corr}(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{V}(X_i)\mathrm{V}(X_j)}}.$$

We insert, giving us the correlation $corr(\hat{\mu}, \hat{\beta})$

$$Corr(\hat{\mu},\hat{\beta}) = \frac{-\frac{1}{4}\sigma^2}{\sqrt{\frac{4}{8}\sigma^2}\sqrt{\frac{3}{8}\sigma^2}} = -\frac{\frac{-1}{4}}{\sqrt{\frac{12}{64}}} = -\frac{4}{\sqrt{48}} = -\frac{4}{\sqrt{3\cdot 16}} = -\frac{1}{\sqrt{3}},$$

$$\frac{-2\sigma^2}{\sqrt{4\sigma^2\cdot 3\sigma^2}} = \frac{-2}{\sqrt{4\cdot 3}} = \frac{-1}{\sqrt{3}}$$

Question 6.4

We want to make a change to the experimental design, such that both \alpha and \beta can be uniquely estimated and such that their estimates becomes uncorrelated. This can beachieved with the following layout

The tedious way is to find $D(\hat{\theta}) = \sigma^2(x^T\Sigma^{-1}x)^{-1}.$ *for all of them*

Alternatively we can find the one where we have the same combinations with opposite sign – i.e. a balanced design.

	columns	
rows	$E(Y_1) = \mu + \alpha + \beta$	$E(Y_2) = \mu - \alpha + \beta$
	$E(Y_3) = \mu - \alpha - \beta$	$E(Y_4) = \mu + \alpha - \beta$

However, let us try them all. Here using matlab
We use this order

	columns	
rows	$E(Y_1) = \mu + \alpha + \beta$	$E(Y_2) = \mu - \alpha + \beta$
	$\mathrm{E}(Y_3) = \mu - \alpha - \beta$	$E(Y_4) = \mu + \alpha - \beta$

	columns	
rows	$E(Y_1) = \mu + \alpha + \beta$	$E(Y_2) = \mu - \alpha + \beta$
	$\mathrm{E}(Y_3) = \mu + \alpha + \beta$	$E(Y_4) = \mu - \alpha + \beta$

	columns	
rows	$E(Y_1) = \mu - \alpha - \beta$	$E(Y_2) = \mu - \alpha + \beta$
	$\mathrm{E}(Y_3) = \mu - \alpha - \beta$	$E(Y_4) = \mu - \alpha + \beta$

	columns	
rows	$E(Y_1) = \mu - \alpha + \beta$	$E(Y_2) = \mu - \alpha + \beta$
	$\mathrm{E}(Y_3) = \mu - \alpha + \beta$	$E(Y_4) = \mu - \alpha + \beta$

	columns	
rows	$E(Y_1) = \mu + \alpha + \beta$	$E(Y_2) = \mu + \alpha + \beta$
	$\mathrm{E}(Y_3) = \mu - \alpha - \beta$	$E(Y_4) = \mu - \alpha + \beta$

First one is the answer
The second is uncorrelated BUT it is not invertable and we thus cannot uniquely estimate the parameters
The third one is uncorrelated BUT it is not invertable and we thus cannot uniquely estimate the parameters
The fourth one is both correlated and it is not invertable and we thus cannot uniquely estimate the parameters
The fifth one is correlated

```
x = [1  1 1
     1  -1 1
     1 -1 -1
     1 1  -1]

x'*x
inv(x'*x)

x = [1  1 1
     1  -1 1
     1 1 1
     1 -1  1]

x'*x
inv(x'*x)

x = [1  -1 -1
     1  -1 1
```

```
     1 -1 -1
     1 -1  1]

x'*x
inv(x'*x)

x = [1  -1 1
     1  -1 1
     1 -1 1
     1 -1  1]

x'*x
inv(x'*x)

x = [1  1 1
     1  1 1
     1 -1 -1
     1 -1  1]

x'*x
inv(x'*x)
```

PROBLEM 7

We consider the random normal variable

$$Z = \begin{bmatrix} Y \\ X \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ X_1 \\ X_2 \end{bmatrix}, \quad E(Z) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad D(Z) = \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Question 7.1

For $D(Z)$ to be regular, ρ must be in the range

From page 438

Here we note, that an $n \times n$ matrix A is said to be ***regular*** if the corresponding linear transformation is bijective. This is equivalent with the existence of an ***inverse matrix***, i.e. a matrix A^{-1}, which satisfies

$$A\,A^{-1} = A^{-1}A = I$$

From page 442

We find as a special case that **A** is regular if and only if det **A** is not equal to 0.

Det(D(Z)) = -(p^2 – 1)^3

That is fulfilled for

$$-1 < \rho < 1$$

Question 7.2

The first squared canonical correlation between Y and X is given by:

Theorem 6.13

Let the situation be as in the previous theorem. Then we have

$$(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \varrho_r^2\Sigma_{yy})a_r = 0$$
$$\det(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \varrho_r^2\Sigma_{yy}) = 0$$

respectively

$$(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \varrho_r^2\Sigma_{xx})b_r = 0$$
$$\det(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \varrho_r^2\Sigma_{xx}) = 0$$

We can use e.g. matlab:

```
solve(det([p^2 p^3; p p^2]*inv([1 p; p 1])*[p^2 p; p^3 p^2] - r^2 * [1 p; p 1]) == 0,r)
solve(det([p^2 p; p^3 p^2]*inv([1 p; p 1])*[p^2 p^3; p p^2] - r^2 * [1 p; p 1]) == 0,r)

0
 0
 p
-p

Which we need to square and thus get p^2
```

Question 7.3

The variable Y_1 explains the following fraction of the variance in Y_2

We have from page 29

and the *squared coefficient of correlation represents the reduction in variance. i.e. the fraction of Y's variance, which can be explained by X,* since

$$\rho^2 = \frac{\mathrm{V}(Y) - \mathrm{V}(Y|X = x)}{\mathrm{V}(Y)}.$$

We find the correlation from

$$D(Z) = \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

And square it p^2

Question 7.4

The conditional mean $E\left(\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \mid \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$ is:

We use

Theorem 1.27

If X_2 is regularly distributed, i.e. if Σ_{22} has full rank, then the distribution of X_1 conditioned on $X_2 = x_2$ is again a normal distribution, and the following holds

$$\begin{aligned} \mathrm{E}(X_1|X_2 = x_2) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \mathrm{D}(X_1|X_2 = x_2) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

If Σ_{22} does not have full rank then the conditional distribution is still normal and Σ_{22}^{-1} in the above equations should be substituted by a generalised inverse Σ_{22}^{-}.

We insert

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} \rho^2 & \rho^3 \\ \rho & \rho^2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_1 - 0 \\ x_2 - 0 \end{bmatrix} = \begin{bmatrix} x_1\left(\frac{\rho^2}{\rho^2 - 1} - \frac{\rho^4}{\rho^2 - 1}\right) \\ x_1\left(\frac{\rho^2}{\rho^2 - 1} - \frac{\rho^3}{\rho^2 - 1}\right) \end{bmatrix} = \begin{bmatrix} x_1\rho^2 \\ x_1\rho \end{bmatrix}$$

```
simplify([p^2 p^3; p p^2]*inv([1 p; p 1])*[x1;x2])
```

*p^2*x1*
*p*x1*

Question 7.5

The conditional dispersion $D\left(\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \mid \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$ is:

We use

Theorem 1.27

If X_2 is regularly distributed, i.e. if Σ_{22} has full rank, then the distribution of X_1 conditioned on $X_2 = x_2$ is again a normal distribution, and the following holds

$$\begin{aligned} \mathrm{E}(X_1|X_2 = x_2) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \mathrm{D}(X_1|X_2 = x_2) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

If Σ_{22} does not have full rank then the conditional distribution is still normal and Σ_{22}^{-1} in the above equations should be substituted by a generalised inverse Σ_{22}^{-}.

We insert

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} - \begin{bmatrix} \rho^2 & \rho^3 \\ \rho & \rho^2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} \rho^2 & \rho \\ \rho^3 & \rho^2 \end{bmatrix} = \begin{bmatrix} 1-\rho^4 & \rho-\rho^3 \\ \rho-\rho^3 & 1-\rho^2 \end{bmatrix}$$

```
simplify([1 p;p 1] - [p^2 p^3; p p^2]*inv([1 p; p 1])*[p^2 p; p^3 p^2])
```

[1 - p^4, - p^3 + p]
[- p^3 + p, 1 - p^2]

Question 7.6

The squared partial correlation $\rho^2_{Y_1Y_2|X_1X_2}$ is given by:

We use

Theorem 1.27

If X_2 is regularly distributed, i.e. if Σ_{22} has full rank, then the distribution of X_1 conditioned on $X_2 = x_2$ is again a normal distribution, and the following holds

$$\begin{aligned} \mathrm{E}(X_1|X_2 = x_2) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \mathrm{D}(X_1|X_2 = x_2) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

If Σ_{22} does not have full rank then the conditional distribution is still normal and Σ_{22}^{-1} in the above equations should be substituted by a generalised inverse Σ_{22}^{-}.

We use from the previous question
[1 - p^4, - p^3 + p]
[- p^3 + p, 1 - p^2]

And combine with this from page 8

$$\rho_{ij} = \mathrm{Corr}(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{V}(X_i)\mathrm{V}(X_j)}}.$$

$$\frac{\rho - \rho^3}{\sqrt{1-\rho^4}\sqrt{1-\rho^2}} = \frac{\rho}{\rho^2 + 1}$$

```
simplify( ((p-p^3)/(sqrt(1-p^4)*sqrt(1-p^2)))^2)
```
p^2/(p^2 + 1)

Question 7.7

The squared multiple correlation $\rho^2_{Y_1|X_1X_2}$ is

Theorem 1.42

We consider the situation above. Let σ_i be the i'th column in Σ_{xy}, i.e. σ_i^T is the i'th row in Σ_{yx}. Further, let σ_{ii} denote the i'th diagonal element, i.e. the variance of Y_i

Then

$$\rho_{y_i|x} = \frac{\sqrt{\sigma_i^T \Sigma_{xx}^{-1} \sigma_i}}{\sqrt{\sigma_{ii}}}.$$

If we let

$$\Sigma_i = \begin{bmatrix} \sigma_{ii} & \sigma_i^T \\ \sigma_i & \Sigma_{xx} \end{bmatrix},$$

then

$$1 - \rho_{y_i|x}^2 = \frac{\det \Sigma_i}{\sigma_{ii} \det \Sigma_{xx}} = \frac{V(Y_i|X)}{V(Y_i)},$$

$$1 - \frac{det\left(\begin{bmatrix} 1 & \rho^2 & \rho^3 \\ \rho^2 & 1 & \rho \\ \rho^3 & \rho & 1 \end{bmatrix}\right)}{\det\left(\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)} = \rho^4$$