Compulsory exercise 1: Group 3

TMA4268 Statistical Learning V2023

Helle Villmones Haug, Hjalmar Jacob Vinje and Sanna Baug Warholm

23 February, 2023

Problem 1

For this problem you will need to include some LaTex code. Please install latex on your computer and then consult Compulsor1.Rmd for hints how to write formulas in LaTex.

a)

b)

c)

d)

e)

Problem 2

a)

i)

The lm() function creates a variable to be estimatied which is multiplied with the binary variable of rankAssocProf and rankProf. The interpretation oof the number corresponding to the variables is that holding all other variables constant going from a AsstProf to AssocProf is assosiated with an increase in salary of 12,907.6 and going from AsstProf to Prof is assosiated with an increase in salary of 45,066.0 holding all other variables constant.

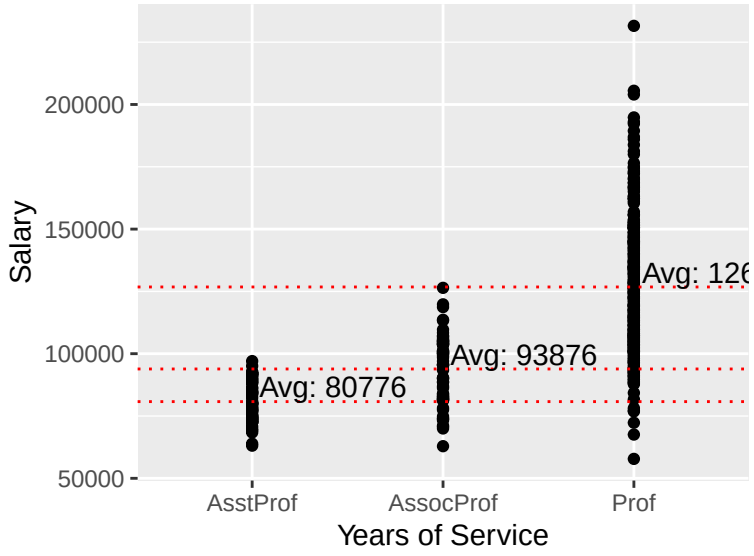
ii)

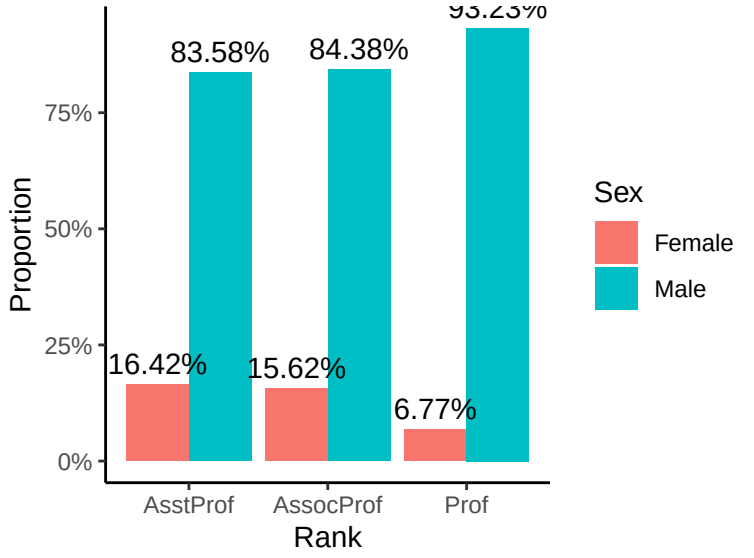
To test the whole categorical variable Rank, you can use an ANOVA test to compare the means of salary for each category of Rank.

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
rank          2 1.432e+11 7.162e+10   128.2 <2e-16 ***
Residuals   394 2.201e+11 5.586e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This shows that rank as a whole has an impact on salary at the 99% confidence level.

b)





The first graph shows that professors earn more than assistante professors and associate professors. The second shows that women make up a higher proportion of AsstProf than Prof.

This is the reason why the regression with sex as the only covariate shows that sex is statistically significant in trems of salary, but it is no longer significant when controlling for rank and the other covariates.

c)

```
model1 <- lm(salary ~ ., data = Salaries)
options(repos = c(CRAN = "https://cran.rstudio.com/"))
install.packages("ggfortify")
```

```
##
## The downloaded binary packages are in
##  /var/folders/wk/x86_p6511195p594k6qnb98h0000gn/T//RtmplJvveo/downloaded_packages
```

```
library(ggplot2)
library(ggfortify)
autoplot(model1, smooth.colour = NA)
```

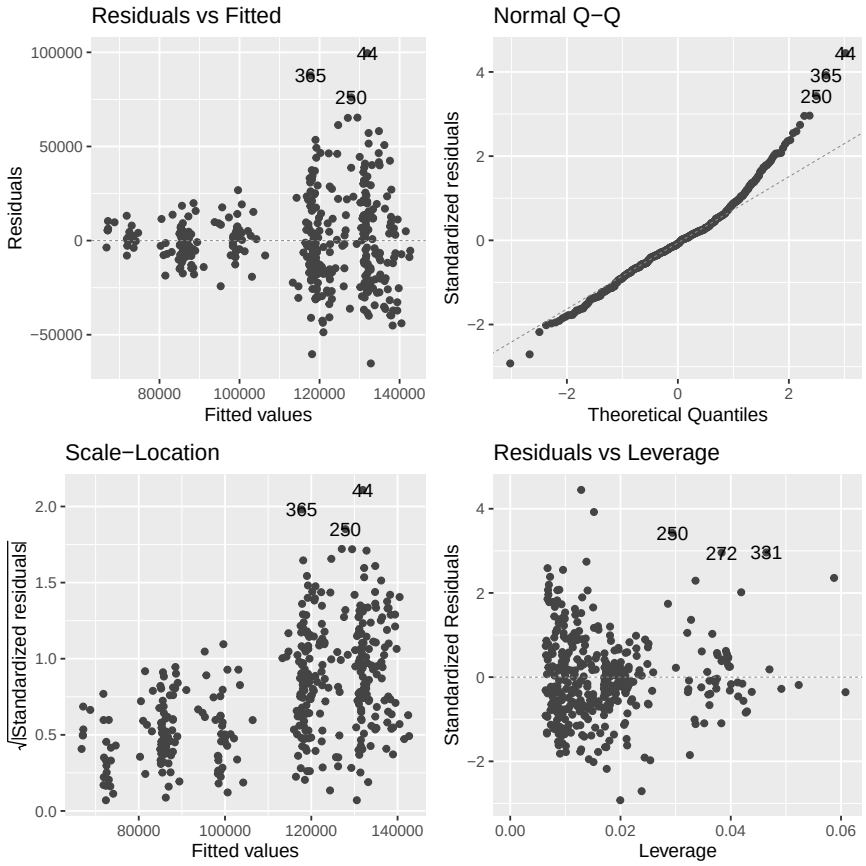


Figure 1: Diagnostic for model1

The first plot showing residuals vs fitted values shows that the data is heteroskedastic, meaning that the variance is not constant. This breaks the assumption of homoscedasticity.

ii)

```
log_salary = log(Salaries$salary)
model2 <- lm(log_salary ~ rank + discipline + yrs.since.phd + yrs.service + sex - salary, data = Salari
library(ggplot2)
autoplot(model2, smooth.colour = NA)
```

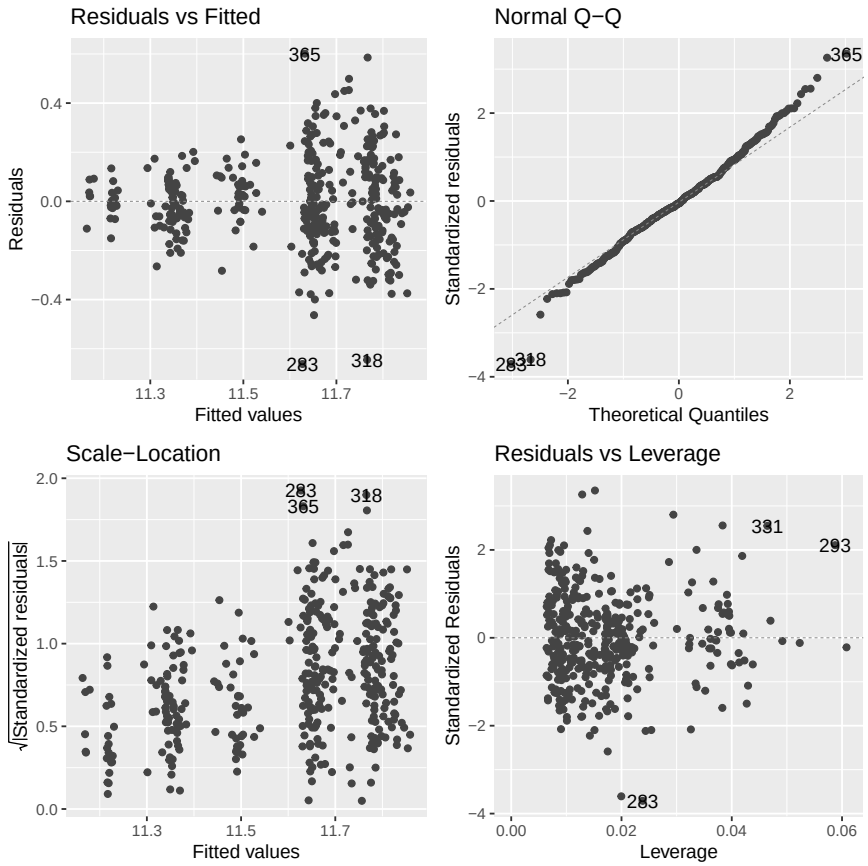


Figure 2: Diagnostic for model2

The residual variance is still heteroskedastic.

```
log_salary = log(Salaries$salary)
model2 <- lm(log_salary ~ rank + discipline + yrs.since.phd + yrs.service + sex - salary, data = Salari
library(ggplot2)
autoplot(model2, smooth.colour = NA)
```

d)

)

```
model3 <- lm(log_salary ~ rank + discipline + yrs.since.phd + yrs.service + sex + sex:yrs.since.phd -sa
summary(model3)
```

```
##
```

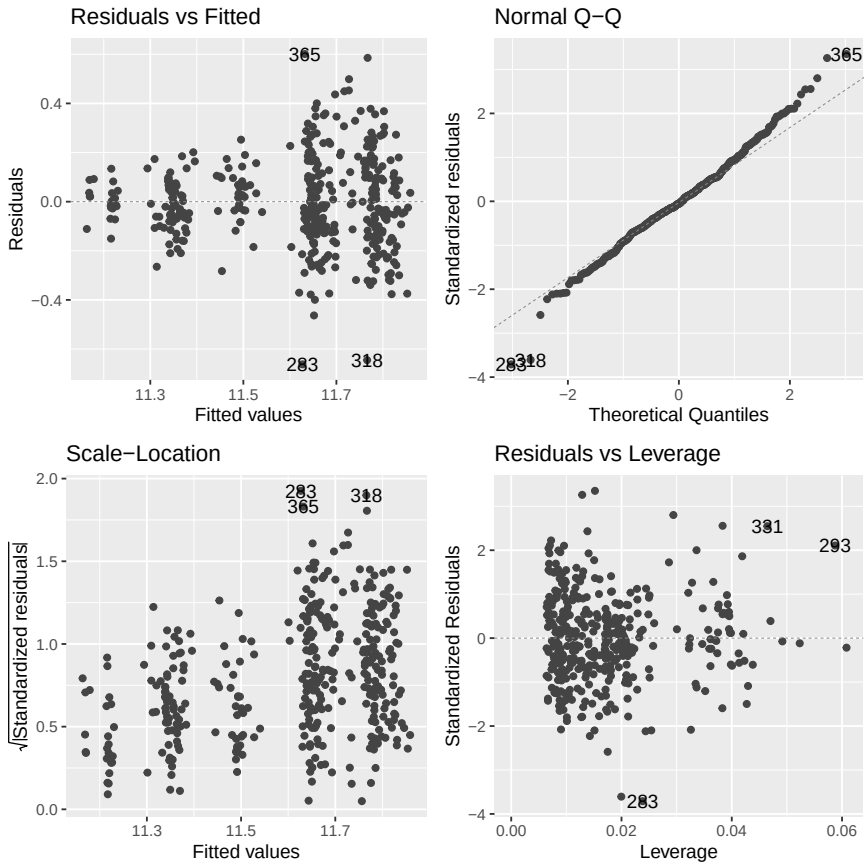


Figure 3: Diagnostic for model3

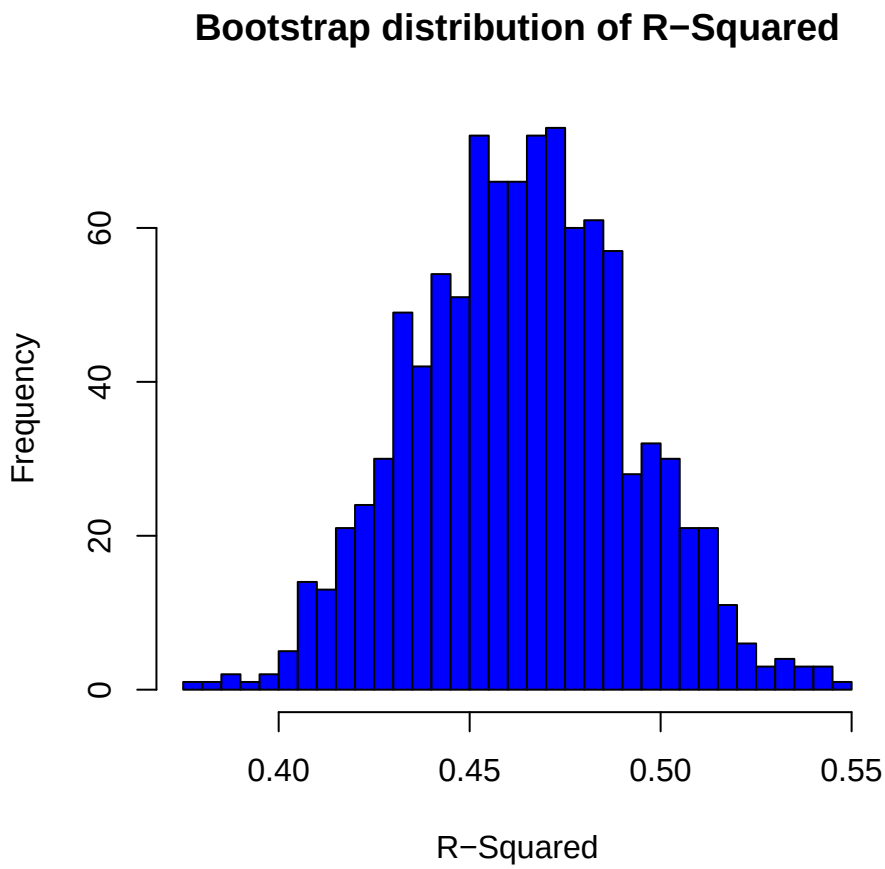
```
## Call:
## lm(formula = log_salary ~ rank + discipline + yrs.since.phd +
##     yrs.service + sex + sex:yrs.since.phd - salary, data = Salaries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66187 -0.10831 -0.00951  0.09846  0.60143
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             11.1537511  0.0591759 188.485  < 2e-16 ***
## rankAssocProf            0.1528200  0.0335575   4.554 7.05e-06 ***
## rankProf                 0.4482679  0.0344343  13.018  < 2e-16 ***
## disciplineB              0.1317818  0.0188133   7.005 1.09e-11 ***
## yrs.since.phd            0.0039500  0.0035253   1.120   0.2632
## yrs.service             -0.0038902  0.0017059  -2.280   0.0231 *
## sexMale                  0.0574914  0.0614436   0.936   0.3500
## yrs.since.phd:sexMale -0.0007049  0.0031407  -0.224   0.8225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1809 on 389 degrees of freedom
## Multiple R-squared:  0.5249, Adjusted R-squared:  0.5163
## F-statistic: 61.39 on 7 and 389 DF,  p-value: < 2.2e-16
```

ii) The interaction term is not statistically significant, so we can not conclude that Bert-Ernie is correct.

e)

```r
# i)
set.seed(4268)
getR2 <- function(data, indices)
  {fit <- lm(salary ~ rank + discipline + yrs.since.phd + yrs.service + sex,
  data = data[indices,])
summary(fit)$r.squared}
library(boot)
boot_results <- boot(data = Salaries, statistic = getR2, R = 1000, strata = Salaries$rank)
```

```r
# ii)
hist(boot_results$t, main = "Bootstrap distribution of R-Squared",
     xlab = "R-Squared", col = "blue", border = "black", breaks = 30)
```




```
# iii)
sd(boot_results$t)
```

```
## [1] 0.02803583
```

```
quant = boot_results$t[25:975]
summary(quant)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3761  0.4440  0.4644  0.4637  0.4825  0.5470
```

(iv) R-squared was calculated to 0.5249 and the bootstrap estimates 0.4546766 with a 95% confidence interval of [0.3761, 0.5470].The original R-squared lies within the 95% confidence interval.

f)

```
# i)
bert_ernie <- data.frame(rank=c("Prof", "Prof"), discipline=c("A", "B"),
                         yrs.since.phd=c(20, 20), yrs.service=c(20, 20),
```

```
                          sex=c("Male", "Male"))
preds <- predict(object=model1, newdata=bert_ernie, interval="prediction", level=0.95)
# 1. Corrected confidence to prediction
# 2. Corrected 0.975 to 0.95
preds[1, 2] > 75000
```

```
## [1] FALSE
```

He can now no longer be confident with 95% certainty that we will earn at least $75 000 at this time.

ii) The analytic expression for the lower limit of the prediction interval:

$$PI_{lower} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{1+\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0}$$

```
x_0 = c(1, 0, 1, 0, 20, 20, 1)
beta_hat = coef(model1)
alpha = 0.05
sd_hat = summary(model1)$sigma
X <- model.matrix(~rank+discipline+yrs.since.phd+yrs.service+sex, data=Salaries)
n = nrow(Salaries)
p = ncol(X)
PI_lower = t(x_0)%*%beta_hat - qt(1-alpha/2,df=n-p) * sd_hat *
  sqrt(1+t(x_0)%*%solve(t(X)%*%X)%*%x_0)
PI_lower
```

```
##          [,1]
## [1,] 72121.12
```

```
PI_lower == preds[1,2]
```

```
##      [,1]
## [1,] TRUE
```

Problem 3

Problem 4