Linear Models: Homework 2

Juan Andres VANEGAS JADAN

2024-2025

Answers to the questions

Question 1a

The model used is expressed below.

$$\text{exam} = \beta_0 + \beta_1 \cdot \text{homeworks} + \beta_2 \cdot \text{Biostat} + \beta_3 \cdot \text{Epi} + \beta_4 \cdot \text{Bioinf} + \epsilon$$

This multivariate model evaluates the impact of two regressors on the exam grade outcome. The first one is *homeworks*, which are the grades obtained by the students in their homework assigments, and the second one is the *specialisation*, which is expressed by a combination of dummy variables.

The estimations for the parameters are:

```
##
## Call:
## lm(formula = exam ~ homeworks + Biostat + Epi + Bioinf, data = exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.065 -12.555  -2.123   6.640  32.858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.8512    16.3896  -0.906  0.37144
## homeworks    21.6388     6.2665   3.453  0.00154 **
## Biostat       0.7154     6.2625   0.114  0.90975
## Epi           6.7541    12.8513   0.526  0.60271
## Bioinf       15.9902     8.7563   1.826  0.07689 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.19 on 33 degrees of freedom
## Multiple R-squared:  0.3425, Adjusted R-squared:  0.2628
## F-statistic: 4.298 on 4 and 33 DF,  p-value: 0.006579
```

We will analyse one by one the parameter estimation:

- (*Intercept*): the value of -14.85 tell us the baseline for Data Science students, when getting a grade of zero. This does not tell us anything about data science students, but it would be useful to compare with other specialisations using the dummy variables.

- *homeworks*: the estimation value of 21.63 tell us that for each point a student get in homeworks score, the exam grade increases by 21.64 points, when holding the specialisation dummy variables constant. Now, if we analyse the p-value of 0.00154, we can conclude that the relationship homeworks-exam is significat at a 1% level. This is a strong predictor for the exam score.
- *specialisation*:
  - *Biostat* dummy varible: the estimation value of 0.7154 tell us that Biostatistic students score 0.72 higher than Data Science students, when holding the homeworks regressor constant. However the p-value of 0.91 shows us that this statistic is not significant.
  - *Epi* dummy variable: the estimation value of 6.75 tell us that Epideomology students score 6.75 points higher than Data Science students, when holding the homeworks regressor constant. As before, a p-value of 0.6 shows us that this statistic is not significant.
  - *Bioinf*: the estimation value of of 15.99 tell us that Bio Informatic students score 15.99 points higher than Data Science students, when holding the homeworks regressor constant. While p-value of 0.07 is outside the convention of 0.05 significance, which tell us that this is not significant, it may be that Bio Informatic students perform better than Data Science students. However, given the asignation to different specialisation were not assigned randomically, we could not conclude causality by the parameters, only correlation.

If we analyse multicolinearity:

```
## homeworks   Biostat       Epi    Bioinf
##  1.029084  1.173615  1.059041  1.126714
```

We can see that all values are lower than 5, showing us that there is no prove of multicolinearity.

Question 1b

Below, the result line of *homeworks*

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 21.638830723  6.266481308  3.453107040  0.001539938
```

First, we need to obtain the degrees of freedom. We have 38 samples, so this minus the number of parameters, give as 33 as degrees of freedom. From (2.10), we know

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} \sim t_{df}. \tag{2.10}$$

so we compute the quantile at 0.975 for the t distribution at 33 degrees of freedom, and multiply by the standard error to obtain the margin error. With this we get a confidence interval of

```
## [1] "8.88957864047308 34.3880828052091"
```

This means that with 95% confidence that for each point increased in the homework grade, the exam score will increase between 8.89 and 34.39, when the specialisation dummy are held constant.

Question 1c

Since the p-value is lower than 0.05, we can refuse the null hypothesis $H_0 : \beta_1 = 0$, and makes as conclude that homeworks grades have a significant effect on the exam score.

Question 1d

If we include *simulation* score as a regressor, we obtain:

```
##
## Call:
## lm(formula = exam ~ homeworks + Biostat + Epi + Bioinf + simulation,
##     data = exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.398 -11.520  -1.627   9.116  30.054
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.1503    18.0730  -1.115  0.27318
## homeworks    21.6435     6.3126   3.429  0.00169 **
## Biostat       0.9875     6.3199   0.156  0.87682
## Epi           7.1260    12.9561   0.550  0.58613
## Bioinf       14.2550     9.1433   1.559  0.12882
## simulation    2.8089     3.8967   0.721  0.47625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.32 on 32 degrees of freedom
## Multiple R-squared:  0.353,  Adjusted R-squared:  0.2519
## F-statistic: 3.492 on 5 and 32 DF,  p-value: 0.01246
```

For this model, we see a slight increase of the parameter β_1, which correspons to the *homeworks* regressor parameter. This goes from 21.6388 to 21.6435. However, we also see an increase on the Std. Error of the the parameter, which goes from 6.26 to 6.31, and the p-value which goes from 0.0015 to 0.0017. This could be an indication that the homeworks and the simulation grades share information. In other words, this leads us to evaluate if there is multi colinearity between these regressors.

We also see that the simulation score has a positive effect of 2.81 on the exam score when other regressors remain constant. However, with a p-value of 0.48, this parameter estimate cannot be consider significant.

Question 1e

If we add the interaction effect to the model, we would get:

```
##
## Call:
## lm(formula = exam ~ homeworks + Biostat + Epi + Bioinf + simulation +
##     homeworks * simulation, data = exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.879 -11.905  -1.711  13.222  35.670
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            39.119     42.574   0.919   0.3653
```

```
## homeworks             -2.389     16.873  -0.142   0.8883
## Biostat                1.767      6.212   0.284   0.7779
## Epi                    9.391     12.778   0.735   0.4679
## Bioinf                15.529      8.996   1.726   0.0943 .
## simulation           -27.706     20.296  -1.365   0.1820
## homeworks:simulation   12.223      7.985   1.531   0.1360
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.96 on 31 degrees of freedom
## Multiple R-squared:  0.3985, Adjusted R-squared:  0.2821
## F-statistic: 3.423 on 6 and 31 DF,  p-value: 0.01041
```

Now we see that none of the parameters estimates are significant (>0.05). If we only focus on the *homeworks*$\cdot$ *simulation* effect, it gives us a value of 12.223, but with a p-value of 0.136, showing us low signifcance. Therefore, we are not able to refuse the hypthesis $H_0 : \beta_6 = 0$. In other words, we do not see evidence to include this interaction as a regressor.

Question 1f

We can check if there is multicolinearity in the last regressor by computing the VIF, with this we obtain:

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##            homeworks              Biostat                  Epi
##             7.661325             1.185787             1.075136
##               Bioinf           simulation homeworks:simulation
##             1.221084            31.152070            38.592648
```

These results show us a clear multicolinearity of the interaction regressors, the simulation grades, and homeworks grades. We should not include the interation in this model. To check if this multicolinearity happens between homeworks grades and simulation grades, we compute the VIF for the second model, which does not include the interaction between these regressors:

```
##  homeworks    Biostat        Epi     Bioinf simulation
##   1.029085   1.177817   1.060723   1.210631   1.102109
```

All values are less than five, showing us not multicolinearity between homeworks and simulations. This confirms when we saw that the interation regressor was not significant.

Question 1g

From the analysis, we can say that homeworks grades have a significant impact on the exam score. Students that perform well on their homework assigments are more likely to performe well on the exam.

Appendix with R code

Question 1a

```r
# Change the path to the data file in the following line
load(file = "~/academics/hasselt/linear-models/Data/exam.RData")

str(exam)
```

```
## 'data.frame':    38 obs. of  4 variables:
##  $ exam          : num  37 25 23 37 19 45 58 40 21 91 ...
##  $ homeworks     : num  3 2.5 2.5 2.5 2.5 2.5 3 3 1.5 3 ...
##  $ simulation    : num  1 2.5 2.5 2 1.5 1 1.5 1.5 3 3 ...
##  $ specialisation: chr  "BS" "BS" "BS" "BI" ...
```

```r
exam$Biostat <- as.numeric(exam$specialisation == "BS")
exam$DataSc <- as.numeric(exam$specialisation == "D")
exam$Bioinf <- as.numeric(exam$specialisation == "BI")
exam$Epi <- as.numeric(exam$specialisation == "E")
```

```r
m1 <- lm(exam ~ homeworks + Biostat + Epi + Bioinf, data = exam)
print(summary(m1))
```

```
##
## Call:
## lm(formula = exam ~ homeworks + Biostat + Epi + Bioinf, data = exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.065 -12.555  -2.123   6.640  32.858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.8512    16.3896  -0.906  0.37144
## homeworks    21.6388     6.2665   3.453  0.00154 **
## Biostat       0.7154     6.2625   0.114  0.90975
## Epi           6.7541    12.8513   0.526  0.60271
## Bioinf       15.9902     8.7563   1.826  0.07689 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.19 on 33 degrees of freedom
## Multiple R-squared:  0.3425, Adjusted R-squared:  0.2628
## F-statistic: 4.298 on 4 and 33 DF,  p-value: 0.006579
```

```r
vif(m1)
```

```
## homeworks   Biostat       Epi    Bioinf
##  1.029084  1.173615  1.059041  1.126714
```

Question 1b

```r
summary_m1 <- summary(m1)
homeworks_param <- summary(m1)$coefficients["homeworks", ]
print(homeworks_param)
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 21.638830723  6.266481308  3.453107040  0.001539938
```

```r
n <- nrow(exam)
m1_df <- n - (length(coef(m1)))
b1_estimation <- homeworks_param["Estimate"]
estimation_stderr <- homeworks_param["Std. Error"]
estimation_pvalue <- homeworks_param["Pr(>|t|)"]
t_value <- qt(0.975, m1_df)
margin_error <- t_value * estimation_stderr
lower_bound <- b1_estimation - margin_error
upper_bound <- b1_estimation + margin_error
print(paste(lower_bound, upper_bound))
```

```
## [1] "8.88957864047308 34.3880828052091"
```

Question 1d

```r
m2 <- lm(exam ~ homeworks + Biostat + Epi + Bioinf + simulation, data = exam)
summary(m2)
```

```
##
## Call:
## lm(formula = exam ~ homeworks + Biostat + Epi + Bioinf + simulation,
##     data = exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.398 -11.520  -1.627   9.116  30.054
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.1503    18.0730  -1.115  0.27318
## homeworks    21.6435     6.3126   3.429  0.00169 **
## Biostat       0.9875     6.3199   0.156  0.87682
## Epi           7.1260    12.9561   0.550  0.58613
## Bioinf       14.2550     9.1433   1.559  0.12882
## simulation    2.8089     3.8967   0.721  0.47625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.32 on 32 degrees of freedom
## Multiple R-squared:  0.353,  Adjusted R-squared:  0.2519
## F-statistic: 3.492 on 5 and 32 DF,  p-value: 0.01246
```

Question 1e

```r
m3 <- lm(
  exam ~ homeworks + Biostat + Epi + Bioinf + simulation + homeworks * simulation,
  data = exam
)
summary(m3)
```

```
##
## Call:
## lm(formula = exam ~ homeworks + Biostat + Epi + Bioinf + simulation +
##     homeworks * simulation, data = exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.879 -11.905  -1.711  13.222  35.670
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            39.119     42.574   0.919   0.3653
## homeworks              -2.389     16.873  -0.142   0.8883
## Biostat                 1.767      6.212   0.284   0.7779
## Epi                     9.391     12.778   0.735   0.4679
## Bioinf                 15.529      8.996   1.726   0.0943 .
## simulation            -27.706     20.296  -1.365   0.1820
## homeworks:simulation   12.223      7.985   1.531   0.1360
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.96 on 31 degrees of freedom
## Multiple R-squared:  0.3985, Adjusted R-squared:  0.2821
## F-statistic: 3.423 on 6 and 31 DF,  p-value: 0.01041
```

Question 1f

```r
vif(m3)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##            homeworks              Biostat                  Epi
##             7.661325             1.185787             1.075136
##               Bioinf           simulation homeworks:simulation
##             1.221084            31.152070            38.592648
```

```r
vif(m2)
```

```
##  homeworks    Biostat        Epi     Bioinf simulation
##   1.029085   1.177817   1.060723   1.210631   1.102109
```