Linear Models: Homework 2

Emanuel Gloria

2024-2025

Answers to the questions

Question 1a

The linear model “m1” suggests that only the ‘homeworks’ variable has a significant positive effect on ‘exam’ percentages, with each additional point in ‘homeworks’ increasing it by roughly 21.639. The specialization ‘Biostat’, ‘Epi’, and ‘Bioinf’ show positive coefficients, but none are statistically significant except for ‘Bioinf’, which has a borderline effect (p=0.08).

Question 1b

The output returns a confidence interval of the ‘homework’ that ranges between [8.89,34.39]. Therefore, we are 95% confident that each additional point in ‘homeworks’ increases the ‘exam’ by an amount between 8.89 and 34.39 percent, with other variables held constant. This interval does not include zero, confirming that ‘homeworks’ has a statistically significant positive effect on ‘exam’.

Question 1c

The p-value for the ‘homeworks’ coefficient is 0.00154, which is well below the common significant level of 0.05. This indicates that the effect of ‘homeworks’ on ‘exam’ is statistically significant. We can conclude that there is strong evidence that an increase in ‘homeworks’ scores is associated with an increase in ‘exam’ percentage, holding other factors constant.

Question 1d

The estimate of 21.644 in the second model suggests that, on average, each additional point in ‘homeworks’ is associated with a ~21.64 point increase in ‘exam’ percentage, holding ‘simulation’ and other specialization variables constant. This estimate is just slightly higher than in the original model.

Question 1e

The interaction effect between ‘homeworks’ and ‘simulation’ is estimated at 12.22. This means that the effect of ‘homeworks’ on ‘exam’ changes by 12.22 percent for each unit change in simulation and vis-a-vis.

The 95% confidence interval for the interaction term (homeworks:simulation) is [-4.06,28.51]. This range means that we are 95% confident that the true interaction effect lies between -4.06 and 28.51 percent. Since the interval includes 0, this sugest that the interaction effect is not statistically significant at 5% level (i.e., it is plausible that the true interaction effect could be zero).

Question 1f

Base on the model in Question 1e (m3) and the correlation between ‘homeworks’ and ‘simulation’ (0.0355), there does not appear to be a significant problem with multicollinearity in the model. The low correlation between these two variables suggest minimal linear relationship and the model’s coefficient and standard errors are stable, with no signs of multicollinearity issues such as inflated standard errors or high instability. Additionally, the confidence intervals for the coefficients are reasonably wide but not problematic, and the Adjusted R-squared indicates the model explains a modest portion of the variance without significant interference from collinearity. Therefore, multicolliearity is not a major concern in this case.

Question 1g

The analyses shows a clear and consistent positive relationship between homework and exam scores. In all models, ‘homework’ is a significant predictor of exam performance, with higher ‘homework’ scores linked to better ‘exam’ results. Specifically, for each unit increase in ‘homework’ score, ‘exam’ scores tend to increase by about 21 percent. This suggest that doing well on homework is a strong indicator of success on the exam, while other factors, including simulations, have less impact.

Appendix with R code

```
# Change the path to the data file in the following line
load(file="/Users/eman/Documents/3560_LIMO/Assignments/HW2/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" ...
```

```
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")
```

```
library(car) # To use the alias() - identify redundant variables by showiung which coeffcients are line
```

```
## Loading required package: carData
```

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

```
# Calculate the 95% confidence interval for the "homeworks" coefficient
confint(m1, "homeworks", level = 0.95)
```

```
##               2.5 %   97.5 %
## homeworks 8.889579 34.38808
```

```r
# Update the model to include the main effect of simulation
m2 <- lm(exam ~ homeworks + simulation + Biostat + Epi + Bioinf, data = exam)

# Display the summary of the updated model
summary(m2)
```

```
##
## Call:
## lm(formula = exam ~ homeworks + simulation + Biostat + Epi +
##     Bioinf, 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 **
## simulation    2.8089     3.8967   0.721  0.47625
## 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
## ---
## 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
```

```r
# Calculate the 95% confidence interval
confint(m2, level = 0.95)
```

```
##                   2.5 %   97.5 %
## (Intercept) -56.963825 16.66318
## homeworks     8.785186 34.50189
## simulation   -5.128528 10.74624
## Biostat     -11.885745 13.86069
## Epi         -19.264831 33.51680
## Bioinf       -4.369400 32.87937
```

```r
# Update the model to include the interaction effect of simulation and homework
m3 <- lm(exam ~ homeworks * simulation + Biostat + Epi + Bioinf, data = exam)

# Display the summary of the updated model
summary(m3)
```

```
##
## Call:
```

```
## lm(formula = exam ~ homeworks * simulation + Biostat + Epi +
##     Bioinf, 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
## simulation            -27.706     20.296  -1.365   0.1820
## 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 .
## 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
```

```
# Calculate the 95% confidence interval
confint(m3, level = 0.95)
```

```
##                           2.5 %    97.5 %
## (Intercept)          -47.710199 125.94848
## homeworks            -36.803168  32.02422
## simulation           -69.098719  13.68755
## Biostat              -10.902713  14.43698
## Epi                  -16.670909  35.45252
## Bioinf                -2.818035  33.87625
## homeworks:simulation  -4.061738  28.50809
```

```
# Looking at the correlation between homeworks and simulation
cor(exam[c("homeworks","simulation")])
```

```
##             homeworks simulation
## homeworks  1.00000000 0.03547538
## simulation 0.03547538 1.00000000
```

```
# Calculate for linear independence and variance inflation factor
alias(m3)
```

```
## Model :
## exam ~ homeworks * simulation + Biostat + Epi + Bioinf
```

```
vif(m3, type = "predictor")
```

```
## GVIFs computed for predictors
##                GVIF Df GVIF^(1/(2*Df)) Interacts With
## homeworks  1.156683  3        1.024556     simulation
## simulation 1.156683  3        1.024556      homeworks
## Biostat    1.185787  1        1.088938           --
## Epi        1.075136  1        1.036887           --
## Bioinf     1.221084  1        1.105027           --
##                                   Other Predictors
## homeworks                     Biostat, Epi, Bioinf
## simulation                    Biostat, Epi, Bioinf
## Biostat        homeworks, simulation, Epi, Bioinf
## Epi        homeworks, simulation, Biostat, Bioinf
## Bioinf        homeworks, simulation, Biostat, Epi
```