Week 2: Multiple Regression Graded Assignment

Contents


```
colorize <- function(x, color) {
  if (knitr::is_latex_output()) {
    sprintf("\\textcolor{%s}{%s}", color, x)
  } else if (knitr::is_html_output()) {
    sprintf("<span style='color: %s;'>%s</span>", color,
      x)
  } else x
}
```

1 Data description and tasks to finish :

This test exercise is of an applied nature and uses data that are available in the data file TestExer2. The exercise is based on Exercise 3.14 of 'Econometric Methods with Applications in Business and Economics'.

The question of interest is whether the study results of students in Economics can be predicted from the scores on entrance tests taken before they start their studies. More precisely, you are asked to investigate whether verbal and mathematical entrance tests predict freshman grades of students in Economics. Data are available for 609 students on the following variables:

- FGPA: Freshman grade point average (scale 0-4)
- SATV: Score on SAT Verbal test (scale 0-10)
- SATM: Score on SAT Mathematics test (scale 0-10)
- FEM: Gender dummy (1 for females, 0 for males)

(A).

1. Regress FGPA on a constant and SATV. Report the coefficient of SATV and its standard error and p-value (give your answers with 3 decimals).

2. Determine a 95% confidence interval (with 3 decimals) for the effect on FGPA of an increase by 1 point in SATV.

(B).

Answer questions (**A**-1.) and (**A**-2.) also for the regression of FGPA on a constant, SATV, SATM, and FEM.

(C).

Determine the (4×4) correlation matrix of FGPA, SATV, SATM, and FEM. Use these correlations to explain the differences between the outcomes in parts (**A**) and (**B**).

(D).

1. Perform an F-test on the significance (at the 5% level) of the effect of SATV on FGPA, based on the regression in part (b) and another regression.

Note: Use the F-test in terms of SSR or R**2 and use 6 decimals in your computations. The relevant critical value is 3.9.

2. Check numerically that F = t**2.

2 Solutions :

2.1 load libraries and data :

```
library(ggplot2)
library(broom)
library(readr)
library(corrplot)
```

```
## corrplot 0.88 loaded
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
FA2 <- read_csv("TestExer2-GPA-round2.csv",)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Observation = col_double(),
##   FGPA = col_double(),
##   SATM = col_double(),
##   SATV = col_double(),
##   FEM = col_double()
## )
```

```
FA2$Observation <- as.factor(FA2$Observation)
str(FA2)
```

```
## spec_tbl_df[,5] [609 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Observation: Factor w/ 609 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ FGPA       : num [1:609] 2.52 2.33 3 2.11 2.14 ...
##  $ SATM       : num [1:609] 4 4.9 4.4 4.9 4.3 5.1 4.9 4.9 4.4 5.2 ...
##  $ SATV       : num [1:609] 4 3.1 4 3.9 4.7 4.1 4.5 4.6 5.1 4.4 ...
##  $ FEM        : num [1:609] 1 0 1 0 0 1 0 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Observation = col_double(),
##   ..   FGPA = col_double(),
##   ..   SATM = col_double(),
##   ..   SATV = col_double(),
##   ..   FEM = col_double()
##   .. )
```

2.2 Part A solution :

2.2.1 Question 1 :

```
Lin_regc <- lm(FGPA ~ SATV, data = FA2)$coef
Lin_reg <- lm(FGPA ~ SATV, data = FA2)
paste0("The coefficient rounded to 3 decimals is: ", round(Lin_regc[2], digits = 3))
```

```
## [1] "The coefficient rounded to 3 decimals is: 0.063"
```

```
paste0("The intercept rounded to 3 decimals is: ", round(Lin_regc[1], digits = 3))
```

```
## [1] "The intercept rounded to 3 decimals is: 2.442"
```

```
t <- tidy(Lin_reg, conf.int = TRUE)
paste0("The P-Value rounded to 3 decimals is: ", round(t[2,5], digits = 3))
```

```
## [1] "The P-Value rounded to 3 decimals is: 0.023"
```

```r
paste0("The Standard error rounded to 3 decimals is: ", round(t[2,3], digits = 3))
```

```
## [1] "The Standard error rounded to 3 decimals is: 0.028"
```

2.2.2 Question 2 :

```r
paste0("The 95% confidence interval for effect on FGPA with an increase by 1 point is: [", round(t[2,6]
```

```
## [1] "The 95% confidence interval for effect on FGPA with an increase by 1 point is: [0.009,0.117]"
```

2.3 Part B solution :

2.3.1 Question 1 :

```r
MLin_regc <- lm(FGPA ~ SATV + SATM + FEM, data = FA2)$coef
MLin_reg <- lm(FGPA ~ SATV + SATM + FEM, data = FA2)
paste0("The coeficients rounded to 3 decimals is: ", round(MLin_regc, digits = 3))
```

```
## [1] "The coeficients rounded to 3 decimals is: 1.557"
## [2] "The coeficients rounded to 3 decimals is: 0.014"
## [3] "The coeficients rounded to 3 decimals is: 0.173"
## [4] "The coeficients rounded to 3 decimals is: 0.2"
```

```r
summary(MLin_reg)
```

```
##
## Call:
## lm(formula = FGPA ~ SATV + SATM + FEM, data = FA2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31351 -0.29883 -0.02146  0.29419  1.09966
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.55705    0.21610   7.205 1.73e-12 ***
## SATV         0.01416    0.02793   0.507    0.612
## SATM         0.17274    0.03193   5.410 9.07e-08 ***
## FEM          0.20027    0.03738   5.358 1.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4418 on 605 degrees of freedom
## Multiple R-squared:  0.08296,    Adjusted R-squared:  0.07842
## F-statistic: 18.24 on 3 and 605 DF,  p-value: 2.411e-11
```

```
mt <- tidy(MLin_reg, conf.int = TRUE)
paste0("The P-Value rounded to 3 decimals is: ", round(mt[5], digits = 3))
```

```
## [1] "The P-Value rounded to 3 decimals is: c(0, 0.612, 0, 0)"
```

```
paste0("The Standard Errors rounded to 3 decimals is: ", round(mt[3], digits = 3))
```

```
## [1] "The Standard Errors rounded to 3 decimals is: c(0.216, 0.028, 0.032, 0.037)"
```

```
mt
```

```
## # A tibble: 4 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)   1.56      0.216      7.21  1.73e-12   1.13      1.98
## 2 SATV          0.0142    0.0279     0.507 6.12e- 1  -0.0407    0.0690
## 3 SATM          0.173     0.0319     5.41  9.07e- 8   0.110     0.235
## 4 FEM           0.200     0.0374     5.36  1.20e- 7   0.127     0.274
```

2.3.2 Question 2:

```
paste0("The 95% confidence interval for effect on FGPA with an increase by 1 point is: [", round(mt[6],
```

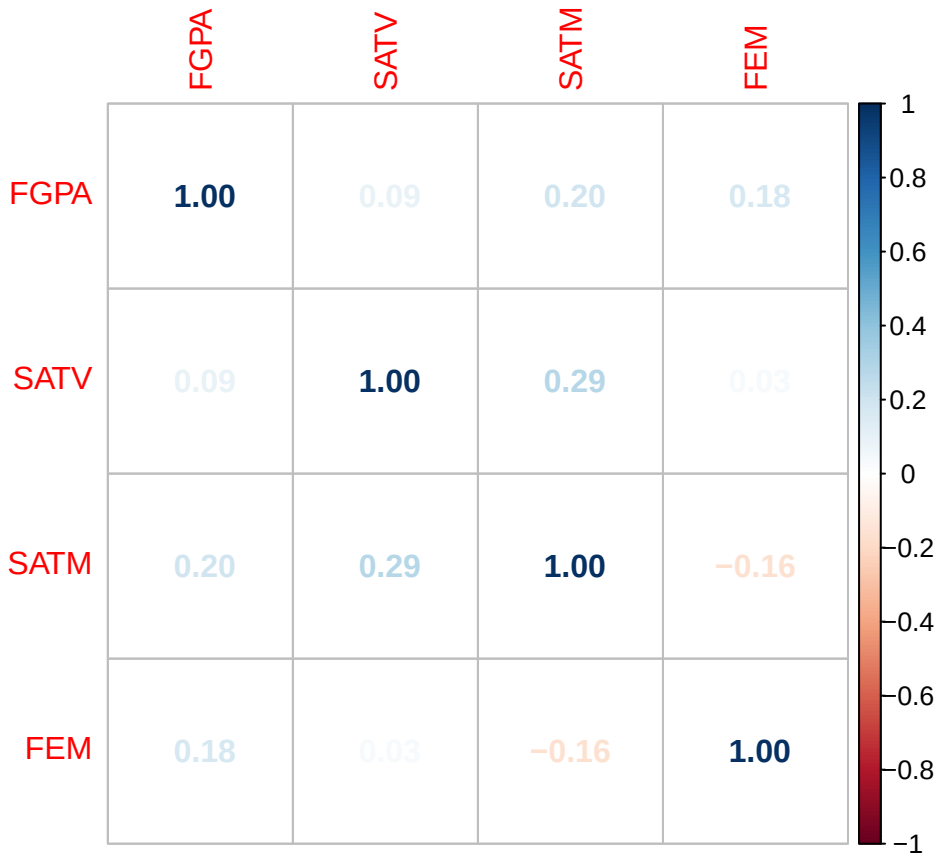
```
## [1] "The 95% confidence interval for effect on FGPA with an increase by 1 point is: [c(1.133, -0.041
```

2.4 Part C solution :

```
co_data <- FA2 %>%
  select(FGPA, SATV, SATM, FEM)
round(cor(co_data), digits = 3)
```

```
##       FGPA  SATV   SATM    FEM
## FGPA 1.000 0.092  0.195  0.176
## SATV 0.092 1.000  0.288  0.034
## SATM 0.195 0.288  1.000 -0.163
## FEM  0.176 0.034 -0.163  1.000
```

```
corrplot(cor(co_data), method = "number")
```



2.4.0.1 Interpretation ; In general, SATV has significant impact on FGPA in a linear regression model. However, since SATM and SATV are highly correlated, their influence on FGPA is reduced for only one factor according to the matrix of correlation and the multiple regression's p-values. When there was a partial dependence (Case B), we saw that SATV does not have a significant impact.

With an important correlation between SATM and FEM, and FGPA, we could only rely on those two factors in our model. The effect of SATV can be absorbed by SATM.

2.5 Part D solution :

Suing the results of Part **B** and inferences of Part **C**, we can create a new model which looks at only SATM and FEM. We can use that to determine SSR which is defined as the sum of the squared differences between the prediction for each observation and the population. We can then use the results fo this model to compare against the the full multiple regression model.

2.5.1 Question 1:

```
Mod_mod <- lm(FGPA ~ SATM + FEM, data = FA2)
ssr <- anova(Mod_mod)[3,2]
r_sq<-summary(Mod_mod)$r.squared
ssr_mr <- anova(MLin_reg)[4,2]
r_sq_mr <- summary(MLin_reg)$r.squared
F_test = (ssr-ssr_mr)/(ssr_mr/605)
```

```r
# Modified Model - SSR and R-squared
paste0("The Modified Model SSR rounded to 6 decimals is: ", round(ssr, digits = 6))
```

```
## [1] "The Modified Model SSR rounded to 6 decimals is: 118.151224"
```

```r
paste0("The Modified Model R^2 rounded to 6 decimals is: ", round(r_sq, digits = 6))
```

```
## [1] "The Modified Model R^2 rounded to 6 decimals is: 0.082575"
```

```r
# Part B Model - SSR and R-squared
paste0("The Part-B Model SSR rounded to 6 decimals is: ", round(ssr_mr, digits = 6))
```

```
## [1] "The Part-B Model SSR rounded to 6 decimals is: 118.101025"
```

```r
paste0("The Part-B Model R^2 rounded to 6 decimals is: ", round(r_sq_mr, digits = 6))
```

```
## [1] "The Part-B Model R^2 rounded to 6 decimals is: 0.082965"
```

```r
# F Statistic
paste0("The F statistic rounded to 3 decimals is: ", round(F_test, digits = 3))
```

```
## [1] "The F statistic rounded to 3 decimals is: 0.257"
```

Since the value of the F-statistic is less than the provided critical value of 3.9, we can safely conclude that the Null hypothesis H_0 is not rejected.

2.5.2 Question 2:

```r
# the t-value of SATV
t_value <- round(summary(MLin_reg)$coefficients[2,3], digits = 3)
t_value_sq <- t_value**2
identical(round(F_test, digits = 3), round(t_value_sq, 3))
```

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