University Admissions R Example

Alexandra Chronopoulou

Lecture #2

Example

The director of admissions of a small college administered a newly designed entrance test to 20 students selected at random from the freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the entrance test score (X). The results of the study follow are summarized in the admissions.txt file.

Analysis

First, we need to set the working directory. In order to find the directory where the file is stored, from the top of the page, go to:

Session -> Set Working Directory -> Choose Directory

Our analysis starts by reading the data into R. We have a .txt file to read using the read.table command, and since we have no headers we add "header=FALSE". If the file contains headers, then we simply put "header=TRUE". R uses "V1" and "V2" as default values for the two columns in the file. We can rename the columns, using the code below:

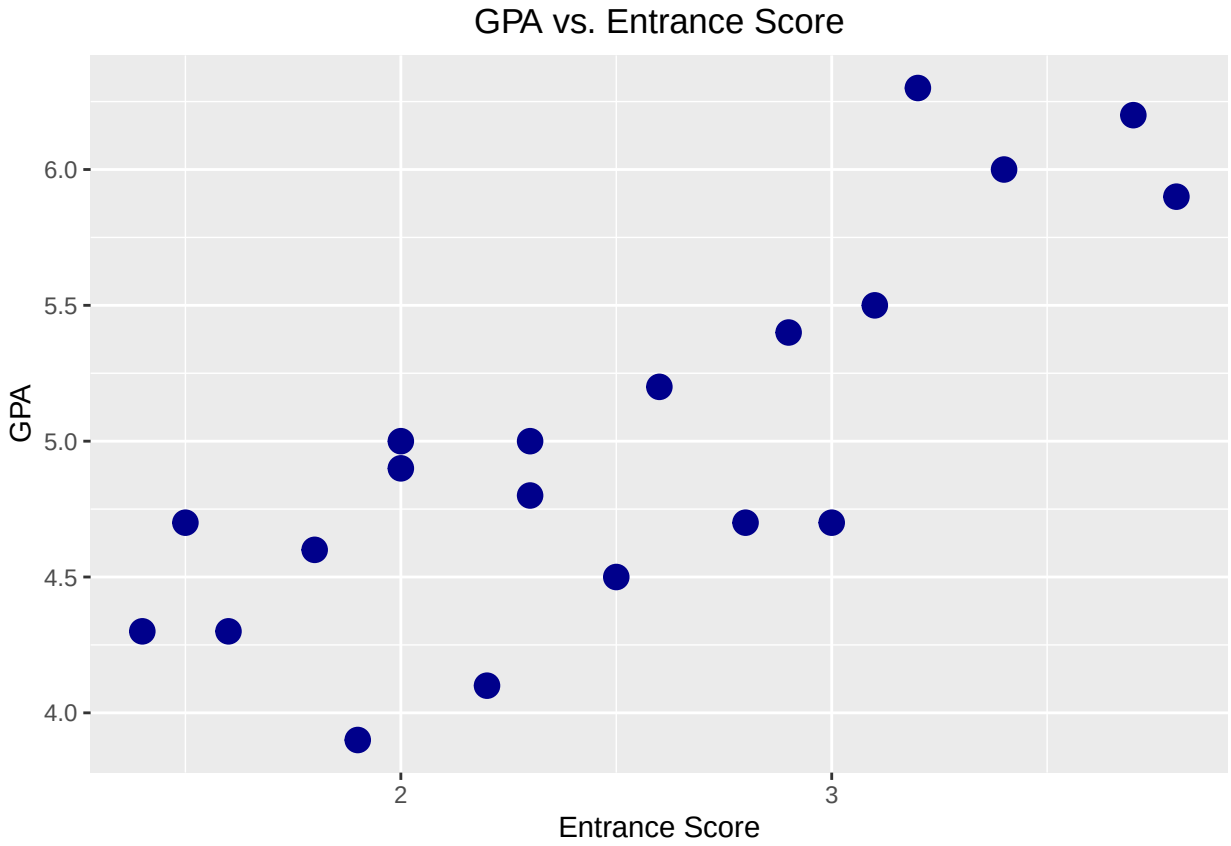
```
admissions = read.table("admissions.txt", header=FALSE)
gpa = admissions$V2
entrance_score = admissions$V1
```

We want to create a scatterplot of the dataset. The R-code and ouput are as follows:

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
scatterplot = ggplot(admissions, aes(entrance_score, gpa)) +
  geom_point(size=4, color='darkblue') +
  labs(title="GPA vs. Entrance Score",y="GPA",x="Entrance Score") +
  theme(legend.position = "none") +
  theme(plot.title = element_text(hjust = 0.5))
plot(scatterplot)
```



Based on the scatterplot, we can say that there might be a linear relation between the two variables, X and Y. Therefore, we fit a simple linear regression:

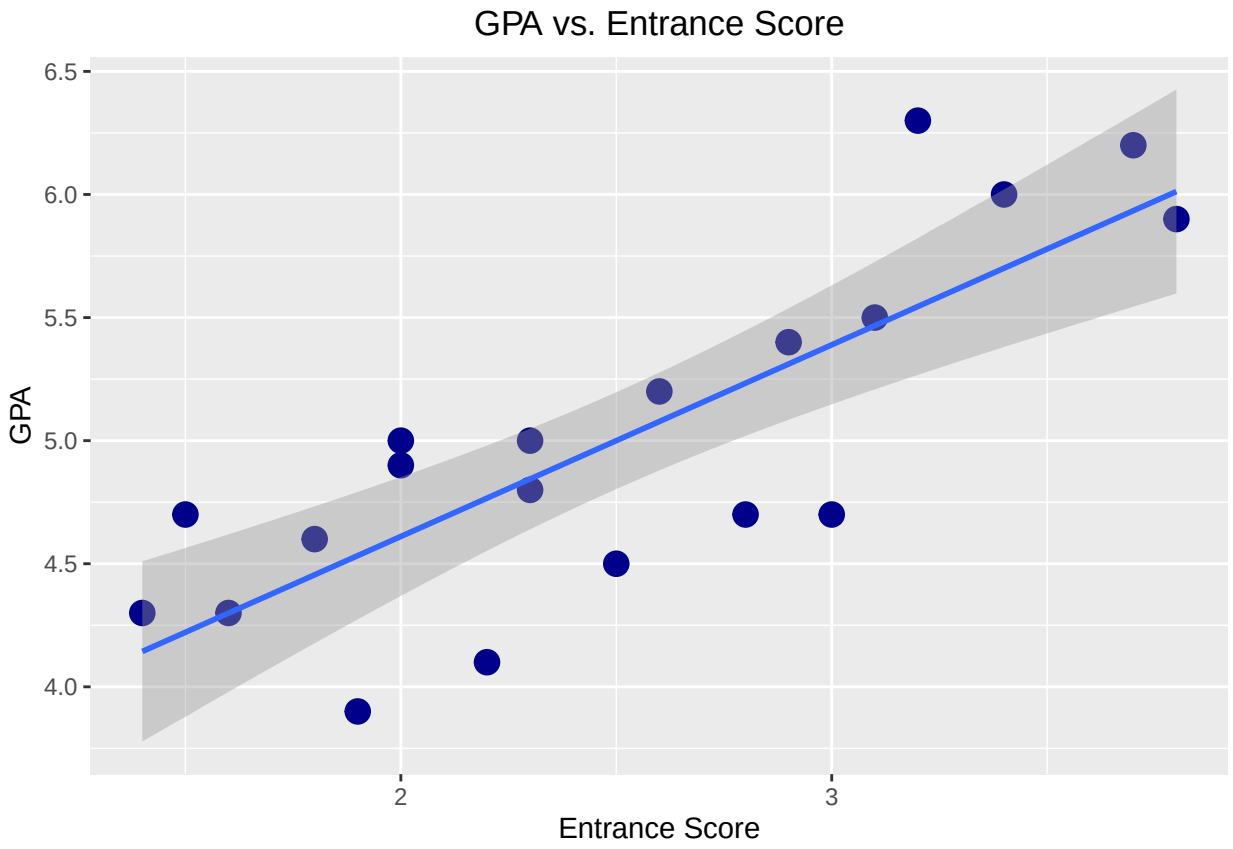
```
admissions.lm = lm(gpa~entrance_score)
summary(admissions.lm)
```

```
##
## Call:
## lm(formula = gpa ~ entrance_score)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6892 -0.2090  0.1054  0.2717  0.7551
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.0539     0.3467   8.809 6.05e-08 ***
## entrance_score   0.7785     0.1335   5.831 1.60e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4188 on 18 degrees of freedom
## Multiple R-squared:  0.6538, Adjusted R-squared:  0.6346
## F-statistic:    34 on 1 and 18 DF,  p-value: 1.597e-05
```

The fitted regression line is shown added in the scatterplot below:

```
scatterplot_reg = ggplot(admissions, aes( entrance_score, gpa)) +
  geom_point(size=4, color='darkblue') +
  labs(title="GPA vs. Entrance Score",y="GPA",x="Entrance Score") +
  theme(legend.position = "none") +
  theme(plot.title = element_text(hjust = 0.5))+
  geom_smooth(method=lm)
plot(scatterplot_reg)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



As we can see, from the R Output, the fitted values for the coefficients are:

$$\beta_0 = 3.0539, \ \beta_1 = 0.7785$$

and the line that we fitted is:

$$(\text{GPA}) = 3.0539 + 0.7785 \cdot (\text{Entrance Score})$$

In order to obtain the residuals and fitted values from R, we have

```
admissions.res = resid(admissions.lm)
admissions.pred = fitted(admissions.lm)
```