Statistical Analysis of University Application Data in R

Jakub Pucilowski

```
file_path <- "/Users/jakubpucilowski/Desktop/Dataset-8.csv"
Dataset.resit <- read_csv(file_path)
```

```
## Rows: 570 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): City, TuitionFees, AvgSalary, Apps
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# it is key to identify
# what "BLUE" represents, first of all it highlights that the OLS estimators are
#linear functions of the select data set so they are a linear combinations of all
#the dependent variables within the data set. Therefore this means that it's
#regression coefficients would be shown as linear functions of the values in the
#dependent variable. Secondly it is considered unbiased when it gives the true
#value of the parameter it is estimating. Ensuring that there is no systematic
#error in the estimation. Lastly, the "best" in this statement correlates to the
#variance being the smallest among all estimators, so the OLS estimator has the
#smallest possible variance around the true parameter, as a result making it the
#most precise. We know this from the Gauss-Markov Theorem which claims,in a
#classical linear regression model the OLS estimators are "BLUE". The claims are;
#that the relationship between the dependent and independent variables is linear,
#across the observations there are no errors that are correlated with each other,
#the errors have a constant variance and that there's no perfect linear
#relationship between each of the independent variables. In conclusion the
#statement holds that under certain conditions the OLS estimators are "BLUE"
#which in turn makes them the best more favorable option for estimating the
#parameters of a linear model.
```

```
avg_salary_mean <- mean(Dataset.resit$AvgSalary, na.rm = TRUE)
avg_salary_min  <- min(Dataset.resit$AvgSalary, na.rm = TRUE)
avg_salary_max  <- max(Dataset.resit$AvgSalary, na.rm = TRUE)
num_observations <- sum(!is.na(Dataset.resit$AvgSalary))
summary_table <- data.frame(
  Statistic = c("Mean", "Minimum", "Maximum", "Number of Observations"),
  Value = c(avg_salary_mean, avg_salary_min, avg_salary_max, num_observations)
)
cat("Summary of Statistics for Variable AvgSalary\n\n")
```

```
## Summary of Statistics for Variable AvgSalary
```

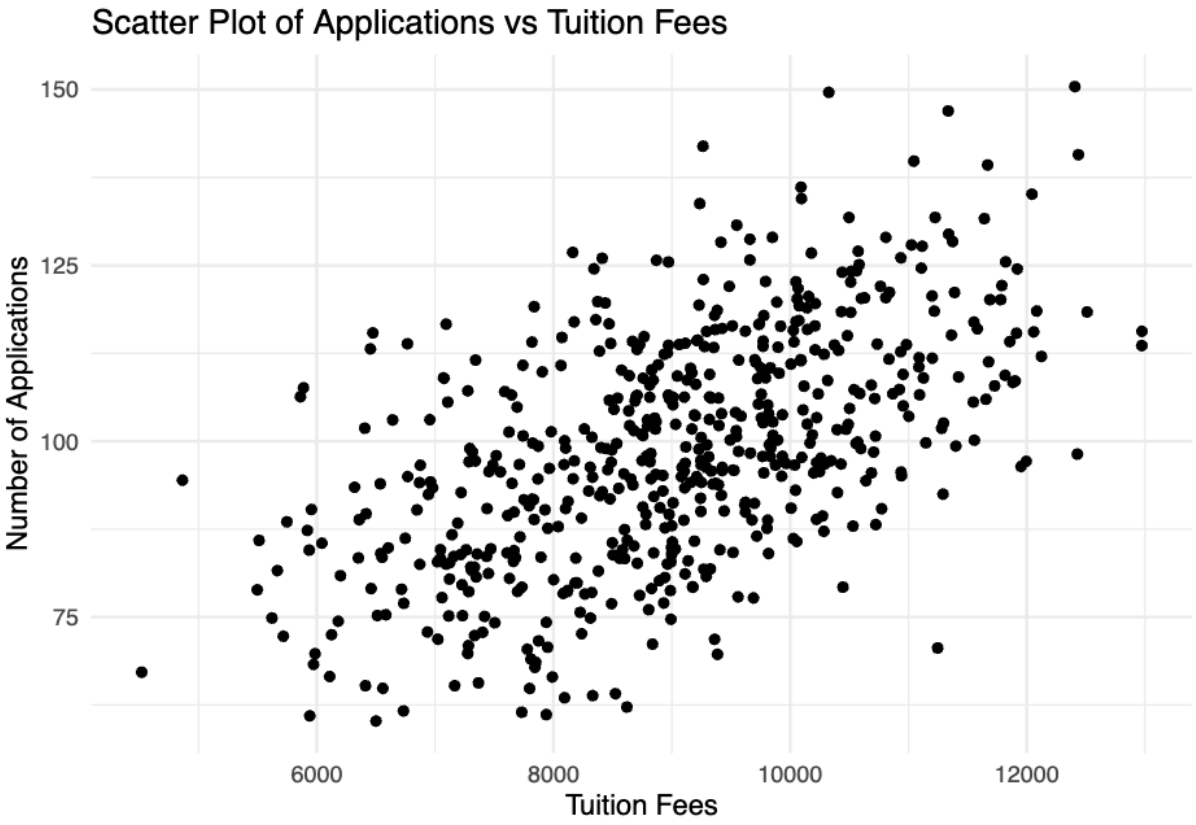
```r
print(summary_table)
```

```
##                Statistic    Value
## 1                   Mean 37051.72
## 2                Minimum 18998.55
## 3                Maximum 52882.44
## 4 Number of Observations   570.00
```

```r
total_universities <- nrow(Dataset.resit)
universities_in_city <- sum(Dataset.resit$City == 1)
share_in_city <- (universities_in_city / total_universities) * 100
city_share_table <- data.frame(
  Location = c("In the City", "Not in the City"),
  Number_of_Universities = c(universities_in_city, total_universities -
                               universities_in_city),
  Share_of_Universities = c(share_in_city, 100 - share_in_city)
)
print(city_share_table)
```

```
##          Location Number_of_Universities Share_of_Universities
## 1     In the City                   153              26.84211
## 2 Not in the City                   417              73.15789
```

```r
library(ggplot2)
ggplot(Dataset.resit, aes(x = TuitionFees, y = Apps)) +
  geom_point() +
  labs(title = "Scatter Plot of Applications vs Tuition Fees",
       x = "Tuition Fees",
       y = "Number of Applications") +
  theme_minimal()
```




```
#Overall when observing the Scatter plot graph generated there is a
#positive trend and relationship between rising tuition fees and the
#number of applications
```

```
library(stargazer)
model1 <- lm(Apps ~ TuitionFees, data = Dataset.resit)
model2 <- lm(Apps ~ TuitionFees + City, data = Dataset.resit)
model3 <- lm(Apps ~ TuitionFees + City + AvgSalary, data = Dataset.resit)
stargazer_output <- stargazer(model1, model2, model3,
                              type = "text",
                              dep.var.labels = "Apps",
                              covariate.labels = c("Tuition Fees", "City",
                                                   "Avg Salary"),
                              omit.stat = c("f", "ser", "adj.rsq"),
                              digits = 2,
                              align = TRUE,
                              column.sep.width = "1pt",
                              star.cutoffs = NA,
                              model.numbers = FALSE,
                              header = FALSE,
                              add.lines = list(
                                c("R-squared",
                                  round(summary(model1)$r.squared, 2),
                                  round(summary(model2)$r.squared, 2),
                                  round(summary(model3)$r.squared, 2)),
                                c("Sample Size",
```

```
                                    nobs(model1),
                                    nobs(model2),
                                    nobs(model3))
                                ),
                                single.row = FALSE
)
```

```
##
## =====================================
##                  Dependent variable:
##               ------------------------
##                          Apps
## -------------------------------------
## Tuition Fees   0.01      0.01     -0.09
##               (0.0004) (0.0003) (0.02)
##
## City                     19.82     19.96
##                          (1.00)   (0.97)
##
## Avg Salary                         0.02
##                                   (0.004)
##
## Constant       42.40     38.22     12.52
##                (3.49)    (2.69)   (5.14)
##
## -------------------------------------
## R-squared       0.32      0.6       0.62
## Sample Size     570       570       570
## Observations    570       570       570
## R2              0.32      0.60      0.62
## =====================================
## Note:                               NA
```

```
cat(stargazer_output)
```

```
##  ====================================                Dependent variable:                ---------
```

```
#In model one the R2 value is 0.32 which explains 32% of the variance in
#applications when only using tuition fees, this value is low implying that
#tuition fees alone aren't a reliable prediction of the number of applications.
#When looking at model two the R2 value is 0.60 which explains 60% of the
#variance in applications if both tuition fees and whether the institution
#is in a city or not is including, this is an improvement from model one,
#indicating that the addition of the city variable significantly increases
#the model's presentation.
#When considering model three's R value of 0.62 it explains 62% of the variance
#in applications by including tuition fees, city and average salary variables.
#This model has the highest R2 value implying that including average salary
#only slightly improves the model compared to model two.
#In closing, the worst fitting model for the data is model one due to its lowest
#R2 value out of all three models being 0.32 implying it has the highest amount
#of variance.
```

```r
#First to work out how much applications will decrease due to increased tuition
#fees based off model three it is key to identify it's coefficient, which is
#-0.09. From this we know that for every additional £1 the number of applications
#decreases by 0.09. So when in application to this question of an increase in
#tuition fees by £1000 we would get a decrease by 90.

coefficient_tuition_fees <- -0.09
change_in_tuition_fees <- 1000
change_in_applications <- coefficient_tuition_fees * change_in_tuition_fees
cat("The expected change in the number of applications is:",
    change_in_applications, "\n")
```

```
## The expected change in the number of applications is: -90
```

```r
coef_model2 <- coef(summary(model2))["TuitionFees", ]
coef_model3 <- coef(summary(model3))["TuitionFees", ]
t_value_model2 <- coef_model2["Estimate"] / coef_model2["Std. Error"]
t_value_model3 <- coef_model3["Estimate"] / coef_model3["Std. Error"]
cat("Model 2 - TuitionFees Coefficient:", coef_model2["Estimate"], "\n")
```

```
## Model 2 - TuitionFees Coefficient: 0.00613948
```

```r
cat("Model 2 - TuitionFees Standard Error:", coef_model2["Std. Error"], "\n")
```

```
## Model 2 - TuitionFees Standard Error: 0.0002936471
```

```r
cat("Model 2 - TuitionFees t-value:", t_value_model2, "\n")
```

```
## Model 2 - TuitionFees t-value: 20.90768
```

```r
cat("Model 2 - Significant at 5% level:", abs(t_value_model2) > 1.96, "\n\n")
```

```
## Model 2 - Significant at 5% level: TRUE
```

```r
cat("Model 3 - TuitionFees Coefficient:", coef_model3["Estimate"], "\n")
```

```
## Model 3 - TuitionFees Coefficient: -0.09242147
```

```r
cat("Model 3 - TuitionFees Standard Error:", coef_model3["Std. Error"], "\n")
```

```
## Model 3 - TuitionFees Standard Error: 0.01694986
```

```r
cat("Model 3 - TuitionFees t-value:", t_value_model3, "\n")
```

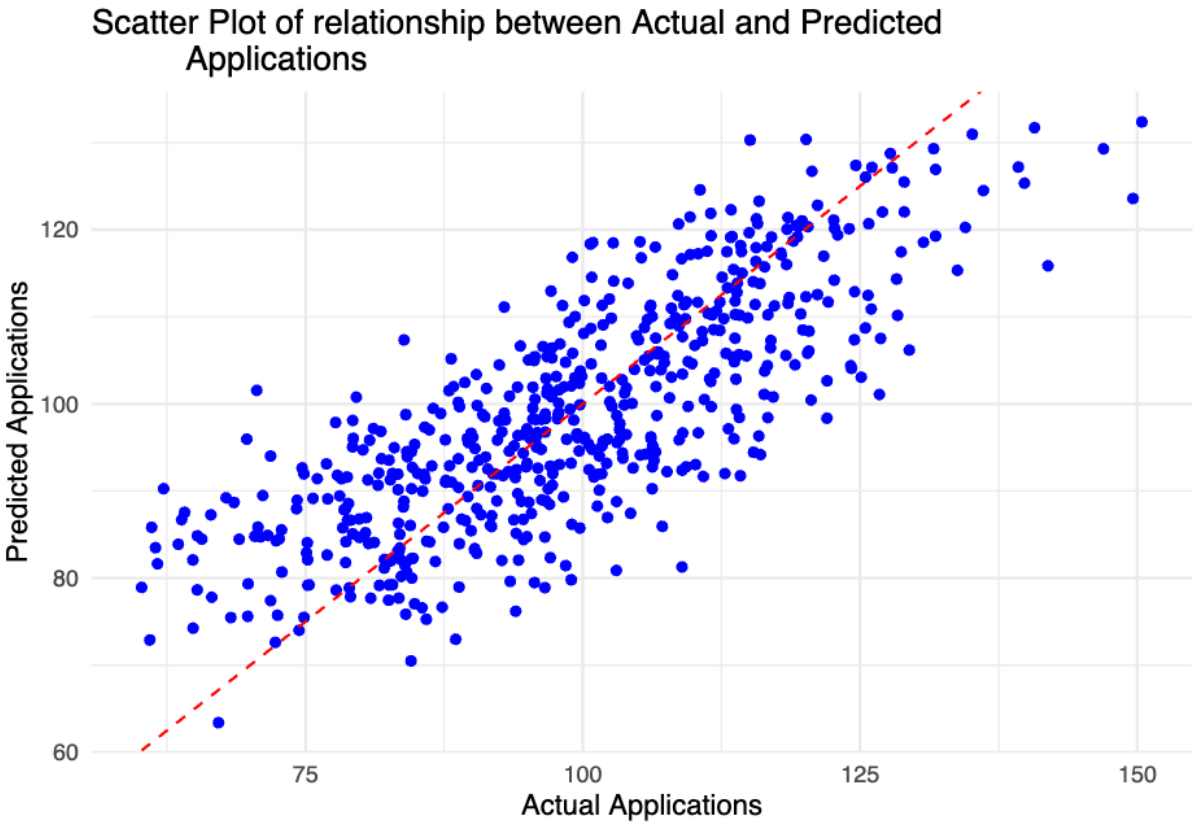
```
## Model 3 - TuitionFees t-value: -5.45264
```

```r
cat("Model 3 - Significant at 5% level:", abs(t_value_model3) > 1.96, "\n")
```

```
## Model 3 - Significant at 5% level: TRUE
```

```r
#The coefficient for tuition fees in model 2 is 0.01 when rounded to 2 decimals
#after the comma, meaning that in model 2 a £1 increase relates to an increase
#of 0.01 in the the amount of applications.
#When looking at model 3 the coefficient is -0.09 for tuition fees when rounded
#to 2 decimals points after the comma. Meaning that a £1 increase in tuition fees
#relates to a decrease of 0.09 in the number of applications.
#Now when interpreting whether the coefficient statistically is significant at
#the 5% level in each of the models, from my output both seem to be "TRUE" when
#tested for significant at the 5% level indicating that in both models the
#coefficient is statistically significant.
#When comparing the coefficients in the 2 models we can see it changes from a
#positive in model 2, to a negative in model 3. Suggesting that the influence of
#tuition fees on applications is reversed when average graduate salary is
#included in the model, suggesting that average salary interacts with the effect
#of tuition fees. In model 2 only the tuition fees and city are predictors, the
#coefficient produced, implies that when holding city constant the increase in
#tuition fees also related to more applications.
#While in model 3 we add average salary as a predictor, and the change to a
#negative coefficient in model 3 shows that after accounting for average salary
#the influence of tuition fees is negative, which in conclusion shows that the
#higher the tuition fees the fewer the applications. This shows the true
#relationship between tuition fees and applications while in model 2 the positive
#relationship could be due to variable bias.
```

```r
library(ggplot2)
predicted_apps <- predict(model3, Dataset.resit)
ggplot(Dataset.resit, aes(x = Apps, y = predicted_apps)) +
  geom_point(color = "blue") +  # Scatter plot points
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  # 45-degree reference line
  labs(x = "Actual Applications", y = "Predicted Applications") +
  ggtitle("Scatter Plot of relationship between Actual and Predicted
          Applications") +
  theme_minimal()
```




```r
head(predicted_apps)
```

```
##         1         2         3         4         5         6
## 100.44010  85.85034 122.05304  82.84641  91.19990 109.89803
```

```r
#When interpreting this plot, if the model perfectly predicted the number of
#applications then all points would be on the line of reference which is the Red
#45 degree line. when points on the plot are not on the line it suggests
#discrepancies between the values of actual and predicted applications, with
#points above the line implying overestimation while points under the line imply
#underestimation. As most points lie close to the reference line it implies that
#model 3 produces a strong prediction of the number of applications, nevertheless
#it is key to observe any outstanding deviation from the reference line as it
#could show areas that the model requires improvement on, such as considering
#more variables in the data set.
```

#Appendix of my R script

```r
library(ggplot2)
library(stargazer)
library(readr)
file_path <- "/Users/jakubpucilowski/Desktop/Dataset-8.csv"
Dataset.resit <- read_csv(file_path)
avg_salary_mean <- mean(Dataset.resit$AvgSalary, na.rm = TRUE)
avg_salary_min  <- min(Dataset.resit$AvgSalary, na.rm = TRUE)
```

```r
avg_salary_max  <- max(Dataset.resit$AvgSalary, na.rm = TRUE)
num_observations <- sum(!is.na(Dataset.resit$AvgSalary))
summary_table <- data.frame(
  Statistic = c("Mean", "Minimum", "Maximum", "Number of Observations"),
  Value = c(avg_salary_mean, avg_salary_min, avg_salary_max, num_observations)
)
cat("Summary of Statistics for Variable AvgSalary\n\n")
print(summary_table)
total_universities <- nrow(Dataset.resit)
universities_in_city <- sum(Dataset.resit$City == 1)
share_in_city <- (universities_in_city / total_universities) * 100
city_share_table <- data.frame(
  Location = c("In the City", "Not in the City"),
  Number_of_Universities = c(universities_in_city, total_universities -
                               universities_in_city),
  Share_of_Universities = c(share_in_city, 100 - share_in_city)
)
print(city_share_table)
library(ggplot2)
ggplot(Dataset.resit, aes(x = TuitionFees, y = Apps)) +
  geom_point() +
  labs(title = "Scatter Plot of Applications vs Tuition Fees",
       x = "Tuition Fees",
       y = "Number of Applications") +
  theme_minimal()
library(stargazer)
model1 <- lm(Apps ~ TuitionFees, data = Dataset.resit)
model2 <- lm(Apps ~ TuitionFees + City, data = Dataset.resit)
model3 <- lm(Apps ~ TuitionFees + City + AvgSalary, data = Dataset.resit)
stargazer_output <- stargazer(model1, model2, model3,
                              type = "text",
                              dep.var.labels = "Apps",
                              covariate.labels = c("Tuition Fees", "City",
                                                   "Avg Salary"),
                              omit.stat = c("f", "ser", "adj.rsq"),
                              digits = 2,
                              align = TRUE,
                              column.sep.width = "1pt",
                              star.cutoffs = NA,
                              model.numbers = FALSE,
                              header = FALSE,
                              add.lines = list(
                                c("R-squared",
                                  round(summary(model1)$r.squared, 2),
                                  round(summary(model2)$r.squared, 2),
                                  round(summary(model3)$r.squared, 2)),
                                c("Sample Size",
                                  nobs(model1),
                                  nobs(model2),
                                  nobs(model3))
                              ),
                              single.row = FALSE
)
```

```r
cat(stargazer_output)
coefficient_tuition_fees <- -0.09
change_in_tuition_fees <- 1000
change_in_applications <- coefficient_tuition_fees * change_in_tuition_fees
cat("The expected change in the number of applications is:",
    change_in_applications, "\n")
coef_model2 <- coef(summary(model2))["TuitionFees", ]
coef_model3 <- coef(summary(model3))["TuitionFees", ]
t_value_model2 <- coef_model2["Estimate"] / coef_model2["Std. Error"]
t_value_model3 <- coef_model3["Estimate"] / coef_model3["Std. Error"]
cat("Model 2 - TuitionFees Coefficient:", coef_model2["Estimate"], "\n")
cat("Model 2 - TuitionFees Standard Error:", coef_model2["Std. Error"], "\n")
cat("Model 2 - TuitionFees t-value:", t_value_model2, "\n")
cat("Model 2 - Significant at 5% level:", abs(t_value_model2) > 1.96, "\n\n")

cat("Model 3 - TuitionFees Coefficient:", coef_model3["Estimate"], "\n")
cat("Model 3 - TuitionFees Standard Error:", coef_model3["Std. Error"], "\n")
cat("Model 3 - TuitionFees t-value:", t_value_model3, "\n")
cat("Model 3 - Significant at 5% level:", abs(t_value_model3) > 1.96, "\n")
library(ggplot2)
predicted_apps <- predict(model3, Dataset.resit)
ggplot(Dataset.resit, aes(x = Apps, y = predicted_apps)) +
  geom_point(color = "blue") +  # Scatter plot points
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  # 45-degree reference line
  labs(x = "Actual Applications", y = "Predicted Applications") +
  ggtitle("Scatter Plot of relationship between Actual and
          Predicted Applications") +
  theme_minimal()
head(predicted_apps)
```