ISM - Assignment 2

Geraldo B. Padilla F.

Questions

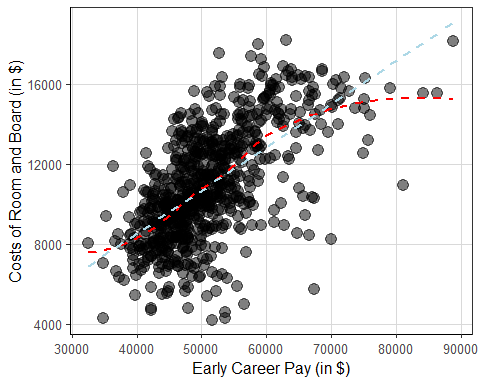
1. Based on your first assignment, are there attributes that could have been missing from the linear regression fitted in assignment 1? That is, is the assumption of all attributes being included in the model appropriate? Why or why not? If not, what other attributes may be of interest?

First, let us recall the proposed analysis. The guiding question was whether the estimated Early Career Pay (in USD) explains some variation in Room and Board costs (in USD). The underlying reasoning was that more expensive careers tend to be offered in expensive universities and cities, so students enrolling in those programs should expect high Room and board costs.

The relationship between the two variables and the linear regression model analyzed were as follow:

Figure 1:

gf\_point(room\_and\_board ~ early\_career\_pay, data = tuition, size = 4, alpha = .5) %>%  
 gf\_smooth(method = 'loess', linetype = 2, color = 'red') %>%  
 gf\_smooth(method = 'lm', linetype = 2, color = 'lightblue') %>%  
 gf\_labs(x = 'Early Career Pay (in $)',  
 y = "Costs of Room and Board (in $)")



cor(room\_and\_board ~ early\_career\_pay, data = tuition)

## [1] 0.623822

rb\_lm <- lm(room\_and\_board ~ early\_career\_pay, data = tuition)  
summary(rb\_lm)

##   
## Call:  
## lm(formula = room\_and\_board ~ early\_career\_pay, data = tuition)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8729.8 -1241.0 9.8 1441.4 6259.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.298e+02 5.171e+02 -0.251 0.802   
## early\_career\_pay 2.161e-01 9.988e-03 21.639 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2161 on 735 degrees of freedom  
## Multiple R-squared: 0.3892, Adjusted R-squared: 0.3883   
## F-statistic: 468.2 on 1 and 735 DF, p-value: < 2.2e-16

coefficients(rb\_lm)

## (Intercept) early\_career\_pay   
## -129.7551288 0.2161361

Based on these results, the relationship is claimed to be positive and approximately linear (r=0.62), with data deviation at the lower and upper extremes (the red line becomes flat at these locations). From the summary of the simple linear regression,the costs of Room and Board (RB) are, on average, -129.75 (b0) when the Early Career Pay (ECP) is equal to zero. The slope (b1) is 0.216, and can be interpreted to mean that for a one unit increase in ECP (in USD), the cost of RB increases, on average, by 0.216 dollars.

Even though the correlation index is moderately high, there is obviously room for improving the model and adding other variables to help explain the variability of the dependent variable. It is difficult for a phenomenon to be explained by a single predictor, and in this case, one could add the type of institution [Public = 0; Private = 1]. The guiding question now will be: do the estimated Early Career Pay (in USD) and the type of institution in which students are enrolled (Private or Public) explain the variation in room and board costs (in USD)?

This variable is added under the assumption that the type of institution, especially Private ones, tends to increase the overall costs of Room and Board for students.

tuition <- tuition %>%  
 mutate(type\_dummy = ifelse(type == 'Private', 1, 0))

1. Add one or more of the attributes identified in question 1 to create a multiple regression model. That is, add one or more of the attributes from question 1 while keeping the attribute from assignment 1 into the regression model. Summarize the interpretation of the regression coefficient estimates for this multiple regression model.

Based on the previous discussion, the multiple regression model is:

rb\_mlm <- lm(room\_and\_board ~ early\_career\_pay + type\_dummy, data = tuition)  
summary(rb\_mlm)

##   
## Call:  
## lm(formula = room\_and\_board ~ early\_career\_pay + type\_dummy,   
## data = tuition)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9049.4 -1198.2 104.1 1290.0 5809.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.195e+02 4.972e+02 -0.442 0.659   
## early\_career\_pay 2.040e-01 9.726e-03 20.969 < 2e-16 \*\*\*  
## type\_dummy 1.230e+03 1.570e+02 7.837 1.62e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2077 on 734 degrees of freedom  
## Multiple R-squared: 0.4363, Adjusted R-squared: 0.4348   
## F-statistic: 284.1 on 2 and 734 DF, p-value: < 2.2e-16

coefficients(rb\_mlm)

## (Intercept) early\_career\_pay type\_dummy   
## -219.5142880 0.2039532 1230.4944616

According to the model summary, the y-intercept (b0) is -219.514. Since the model includes a qualitative variable, the intercept is linked to the reference level of the indicator variable (type of institution = Public [0]). Thus, the intercept would be interpreted as follows: the average costs of Room and Board for students enrolled in Public institutions would be -219.51 when Early Career Pay equals zero. In practical terms, this intercept is hardly interpretable and perhaps a minimum-centering transformation would be useful if we were interested in a more formal analysis.

The partial slope of Early Career Pay (b1) is 0.204, not very different from the value obtained previously (b0 = 0.216), which means that for one unit change (in dollars) in the costs of Early Career Pay, Room and Board costs increase, on average and holding the membership constant (Type of institution), by 0.20 dollars. Again, In this case, the slope is small due to the metric of this variable, whose values are expressed in units of 1.000 dollar. A more practical interpretation would be that the costs of RB increase, on average, by 203,95 dollars per 1.000 units (dollars) increase in ECP.

The partial slope of Type of institution (b2) corresponds to the category of Private institutions, and its interpretation would be as follow: the costs of RB increase, on average, by 1.010,98 (b0+b1) dollars when students enroll in Private institutions. The raw value b2 (1230.49) indicates the difference in average RB costs between students enrolled in Public and Private institutions.

1. Estimate model fit indices to compare the model from assignment 1 to the multiple regression fitted in question 2. Does the model improve model fit based on that from assignment 1? Why or why not?

The r-squared and Residual Standard Error (RSE) are common indicators of model fit to the data. In this case, we have the following changes between the first and second models:

summary(rb\_lm)$r.square; summary(rb\_lm)$sigma

## [1] 0.3891538

## [1] 2160.764

summary(rb\_mlm)$r.square; summary(rb\_mlm)$sigma

## [1] 0.4363248

## [1] 2077.071

The r-squared and RSE improved. The predictors included in the multiple regression model (Early Career Pay and Type of institution) explain around 43,6% of the variability of RB costs, a slight increase compared to the previous result (38,9%). On the other hand, the sigma has decreased from 2160.76 to 2077.07, which means that the difference between the observed RB costs and the predicted ones is, on average, 2077.07 dollars.

Furthermore, since the models are nested, we can analyze the variance decomposition to support or question our results. The objective is to provide further evidence that the added predictor (Type of institution) helps to explain the variance over and above the first (simple) model.

anova(rb\_lm, rb\_mlm)

## Analysis of Variance Table  
##   
## Model 1: room\_and\_board ~ early\_career\_pay  
## Model 2: room\_and\_board ~ early\_career\_pay + type\_dummy  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 735 3431640704   
## 2 734 3166641515 1 2.65e+08 61.425 1.618e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

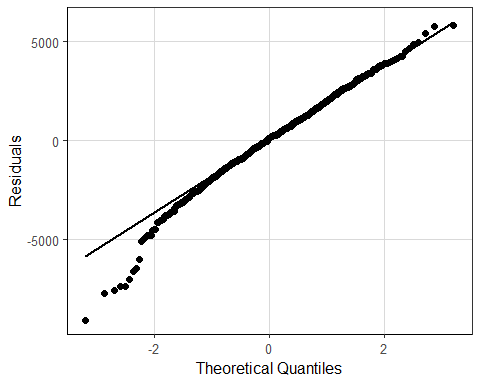
The large F-statistics (61.425) suggests that the Type of institution is related to RB costs and is a useful predictor.

1. Explore and evaluate the model assumptions regarding the residual/error term. Summarize how well the model is meeting the assumptions, citing specific statistics or visualizations to justify your conclusions. In addition, do the model assumptions seem better met compared to those from the regression in assignment 1? 2 pts

First, let us explore the assumption of normality. Based on the Q-Q plot, we could roughly assume that the residuals are normally distributed. A high proportion of the values align with the straight line within 2 standard deviations, but some problems appear at the extremes, especially if the deviation of the values at the bottom of the distribution is considered.

figure 2

tuition$resid <- resid(rb\_mlm)  
ggplot(tuition, aes(sample = resid)) +   
 stat\_qq(size = 2) +   
 stat\_qq\_line(size = 1) +  
 labs(x = 'Theoretical Quantiles', y = 'Residuals')

 The result of the Shapiro-wilk normality test is stricter: the assumption of normal distribution is not satisfied (p-value < 0.001).

shapiro.test(rb\_mlm$residuals)

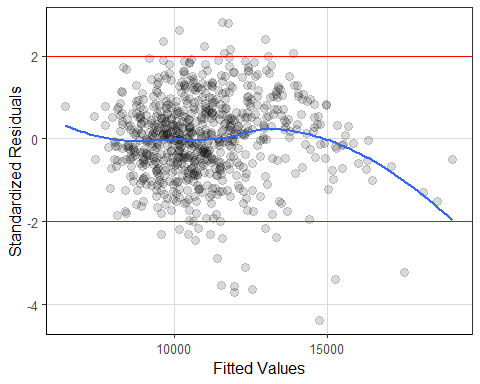
##   
## Shapiro-Wilk normality test  
##   
## data: rb\_mlm$residuals  
## W = 0.98586, p-value = 1.496e-06

However, due to the sample size (n=2,500) and the principles of the CLT theorem, we can assume that violation of this assumption will not affect the model too much.

Second, the assumption of homogeneity of variance look worst. There are several residuals falling out of the 2 standard deviations, a sign that some values could be affecting the estimation due to their large differences. Moreover, there is a decreasing pattern on the right side of the plot, which is also evidence that some residuals are correlated with the predictors.

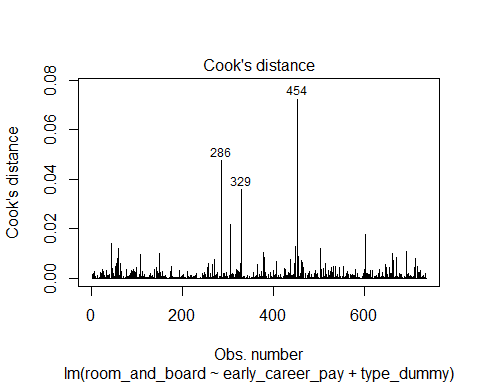
Figure 3. Scale-Location

norm\_residuals <- augment(rb\_mlm)  
gf\_point(.std.resid ~ .fitted, data = norm\_residuals, size = 3, alpha = .15) %>%  
 gf\_hline(yintercept = ~ 2, color = 'red', size = .5) %>%  
 gf\_hline(yintercept = ~ -2, color = 'red', size = .5) %>%  
 gf\_smooth(method = 'loess', size = 1) %>%  
 gf\_labs(x = 'Fitted Values',  
 y = 'Standardized Residuals')



Third, looking at the Cook’s distance plot, there are some influential values in the model that could be affecting the regression estimate, especially those related to Early Career Pay. Despite this evidence, it is difficult to decide whether or not to remove values from the data. For instance, using the rule of thumb for greater than 1 Cook’s distance, none of the values should be removed. However, if some of the values are too extreme or do not make sense compared to the rest of the values, they could be eliminated and the model should be performed again evaluating the improvement of the model.

plot(rb\_mlm, 4)



Compared to the previous model, the assumptions do not improve as much as expected. Although the fit indices (r-squared and RSE) have improved, the assumptions look really similar. Contrary to expectation, the conclusion is almost the same: the residuals are approximately normally distributed: homogeneity of variance cannot be assumed and needs to be further explored; finally, there are some leverage points, but any Cook’s distance is larger than 1, thus values should not be removed from the data set.

1. Summarize the statistical evidence surrounding the null or alternative hypotheses that are being explored for the coefficients entered into the model. Note, I did not explicitly ask you for null/alternative hypotheses, but you may want to write those down for your reference. In a few sentences, describe if the evidence provides support for or against the null hypothesis. Provide specific statistical evidence to support your justification. 2 pts

To analyze whether or not the null hypotheses can be rejected (each slope and the y-intercept equal zero in the population; there is no relationship between the predictors and the outcome), there are two useful indicators: p-values and confidence intervals.

According to the summary of the multiple linear regression, the p-value of the y-intercept is larger (0.659) than accepted at the 0.05 significance level, so most likely the value of β0 in the population is close to zero. On the other hand, the p-values of β1 (<0.001) and β2 (<0.001) indicate that there is a very low probability of observing these values in the population by chance, thus we can conclude, in this example, that they are higher than 0 (under the null hypothesis and using a sampling distribution).

summary(rb\_mlm)

##   
## Call:  
## lm(formula = room\_and\_board ~ early\_career\_pay + type\_dummy,   
## data = tuition)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9049.4 -1198.2 104.1 1290.0 5809.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.195e+02 4.972e+02 -0.442 0.659   
## early\_career\_pay 2.040e-01 9.726e-03 20.969 < 2e-16 \*\*\*  
## type\_dummy 1.230e+03 1.570e+02 7.837 1.62e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2077 on 734 degrees of freedom  
## Multiple R-squared: 0.4363, Adjusted R-squared: 0.4348   
## F-statistic: 284.1 on 2 and 734 DF, p-value: < 2.2e-16

This interpretation of the model results is supported by the 95% confidence interval limits. Only the intercept interval includes zero, so we cannot reject the null hypothesis for this estimated parameter (β0 = 0).

(CI\_b0 <- -219.5142880 + c(-1, 1) \* 1.96 \* 497.2)

## [1] -1194.0263 754.9977

(CI\_b1 <- 0.2039532 + c(-1, 1) \* 1.96 \* 0.009726)

## [1] 0.1848902 0.2230162

(CI\_b2 <- 1230.4944616 + c(-1, 1) \* 1.96 \* 157.0)

## [1] 922.7745 1538.2145

1. Create confidence intervals for the coefficients in the multiple regression model. Justify your confidence level and interpret the confidence intervals in the context of the data. What do these confidence intervals suggest about the magnitude of effect? 2 pts

The intervals were constructed in the previous question on the basis of 95% confidence. Considering the complexity of estimating social phenomena, I consider this level of confidence to be an acceptable middle ground between scientific rigor and practical flexibility.

Although not significant, the y-intercept interval can be interpreted to mean that, with 95% confidence, the average cost of Room and Board for students enrolled in Public institutions is between -$1194.0263 and $754.9977 when Early Career Pay equals zero.

For the case of partial slope b1, with 95% confidence, the unknown value of the parameter lies between 0.184848902 and 0.2230162, that is, for each unit by which Early Career Pay increases, Room and Board costs increase, on average, between 0.18 and 0.22 dollars (holding the memberships constant).

On the other hand, if we were to take 100 repeated samples and construct confidence intervals for each one, 95% of these new intervals for b2 would show that the true value of the slope lies between 922.7745 and 1538.2145. In other words, when students enroll in Private institutions (1), the average difference in Room and Board costs with students in Public institutions increases between $922.8 and $1538.2 (for a given value of ECP).

Based on the confidence intervals, we can estimate that the results are not as positive as expected. The effect of the predictors on the dependent variable raises the need to revise the model, the relationship between the variables and to look for other analysis options using other variables in the data set.

1. Based on the justification in #5 and #6, what practical implications does this result have? That is, are the statistical results that you have been describing/summarizing throughout this assignment practically relevant? Be as specific as you can in your discussion about why you believe the results are practically useful or not.

After analyzing the multiple regression model using Early Career Pay costs and Type of institution to explain the variability of Room and Board costs among students, I think that there is not much practical utility in the results.

On the positive side, both predictors were significant and are factors that increase the costs of RB. If students choose a career with higher costs during the first year they can expect to spend more money on room and board. On the other hand, those who enroll in private institutions can expect to spend more on room and board compared to public institutions, in addition to the costs of the program they study.

These statistical results represent a step towards understanding the phenomenon but are not satisfactory, especially when we consider the results of the principles of normal distribution and homogeneity. A good option to improve the model is to explore the influential and outliers and evaluate whether removing them improves the estimation.