Assignment 1 ISM

Geraldo B. Padilla F.

03-10-2021

**Introduction**

1. Identify a research question of interest that can use linear regression to answer the question.

Guiding question: does the estimated early career pay (in USD) explain variation in the costs of Room and board (in USD)?

1. Explore the research question you identified in #1 descriptively. In a few sentences, summarize any potential relationship, including statistical evidence (i.e., figures or statistics) to support your statements.

To start with, let us explore the two variables included in the analysis:

Table 1:

st1 <- df\_stats(~ early\_career\_pay, data = tuition, min, max, mean, median, sd, quantile(probs = c(0.25, 0.75)))  
st2 <- df\_stats(~ room\_and\_board, data = tuition, mean, min, max, median, sd, quantile(probs = c(0.25, 0.75)))  
st <- rbind(st1, st2)  
head(st)

## response min max mean median sd 25% 75%  
## 1 early\_career\_pay 32500 88800 51150.47 49700 7974.058 45900 55000  
## 2 room\_and\_board 4200 18156 10925.71 10500 2762.779 8948 12880

According to Table 1, the early career pay among universities is, on average, $51,150, with a sd of $7,974.06. On the other hand, students pay for room and board, on average, $10,925.71, with a sd of $2,762.78. The IQR of the early career pay is $9.100, and $3.932 is the IQR of room and board costs. To compare the variability between these two variables, the coefficient of variation (CV) can be calculated.

(7974.058/51150.47)\*100

## [1] 15.58941

(2762.779/10925.71)\*100

## [1] 25.28695

These results suggest that costs of room and board has more dispersion than early career pay.

On the other hand, the both variables can be analyzed using plots. The distribution of early career pay is slightly right-skewed with some outliers above $70.000.

Figure 1.1:

gf\_histogram(~early\_career\_pay, data = tuition, color = 'black', fill = 'lightblue', bins = 50) %>%  
 gf\_boxploth(65 ~ early\_career\_pay, data = tuition, fill = "gray", width = 3) %>%  
 gf\_labs(x = "Early career pay (in $)", y = 'Frequency')

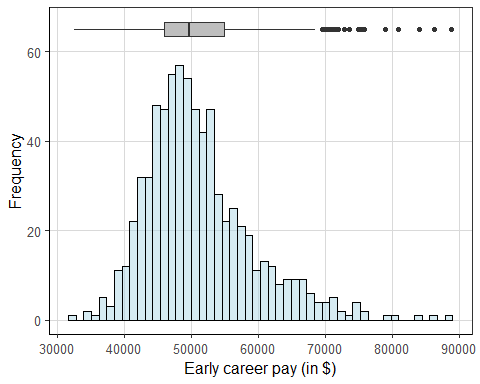
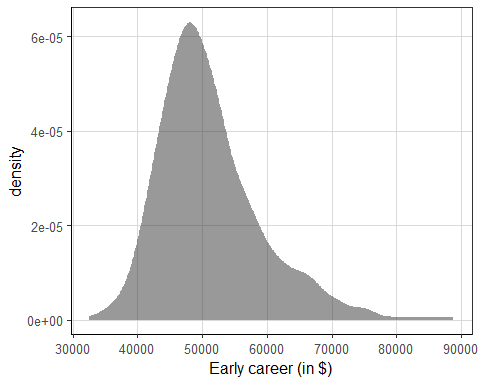


Figure 1.2

gf\_density(~ early\_career\_pay, data = tuition) %>%  
 gf\_labs(x = "Early career (in $)")



The distribution of costs of room and board seems more bell-shaped but leptokurtic.

Figure 2.1:

gf\_histogram(~room\_and\_board, data = tuition, color = 'black', fill = 'lightblue', bins = 50) %>%  
 gf\_boxploth(50 ~ room\_and\_board, data = tuition, fill = "gray", width = 3) %>%  
 gf\_labs(x = "Costs of room and board (in $)", y = 'Frequency')

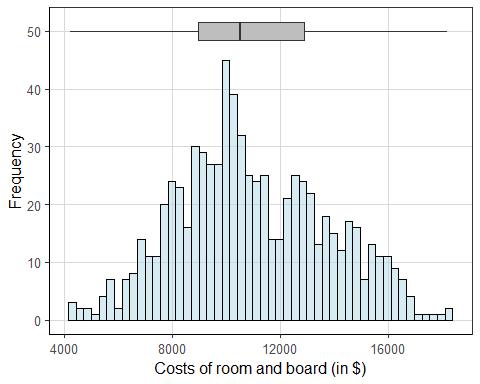
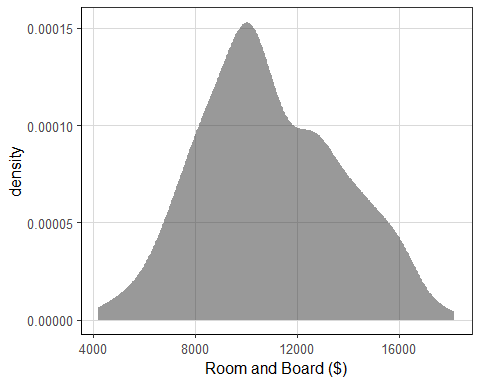


Figure 2.2

gf\_density(~ room\_and\_board, data = tuition) %>%  
 gf\_labs(x = "Room and Board ($)")



The Pearson’s correlation coefficient between the two variables is .62, therefore, the relationship between costs of room and board and early career pay is positive, strong and most possible linear.

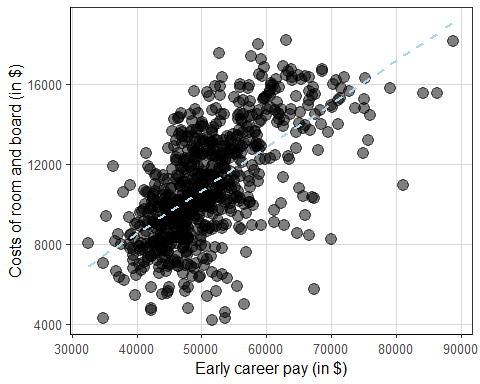
cor(room\_and\_board ~ early\_career\_pay, data = tuition)

## [1] 0.623822

The Figure 3 include a straight line across the relationship between the two variables.Although the positive correlation coefficient, the points show high variability.

Figure 3:

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



1. Fit the linear regression to answer your question from #1. Interpret the regression coefficients in the context of the problem at hand. That is, what do the coefficients mean for the attributes included in the model.

Based on the results of the simple linear regression performed, the cost of Room and Board (RB) is, on average, -129.75 (b0) when the Early career pay (ECP) equals zero. The negative sign of the y-intercept it does not make much sense because in the context of the example there is no costs below 0. However, this estimated b0 is extrapolated until the 0 value of Early career pay (x-axis), which actual minimum value is 32,500.

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

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

On the other hand, the slope (b1) is 0.216. This can be interpreted as follows: for one unit increase in Early career pay (in USD), the cost of Room and board increase, on average, 0.216 dollars. In this case, the slope is small because of the metric of the independent variable. One unit change equals 1 dollar change, and may be using 1,000 units would be better.

1. Is the model intercept as interpretable as it could be? What could be done to enhance the interpretation of the intercept? Summarize in a few sentences any theory or data elements that may help to decide how to improve the interpretation of the intercept.

The current y-intercept is clearly not much interpretable due to its negative sign. To enhance the intuitiveness of its interpretation, a linear transformation of the Early career pay can be performed. As the objective of this simple linear model is studying the relationship between the variables in the training set, a minimum-centering transformation would be the best option to analyze the data.

The mean-centering transformation is another interesting option, but there is no theoretical assumptions or any framework that support this decision. In this case, the most practical transformation is the minimum-centering because it yields a more interpretable y-intercept and keep the model as simple as possible.

1. Interpret the model estimates that show how well the model is performing. That is, what model statistics help to understand how well the model representing the outcome attribute? What do these statistics mean in the context of the problem?

There are two statistics that aid to analyze the accuracy of the model. First, the Residual Standard Error (RSE), which expresses the average difference between the observed values of Room and Board and the predicted ones by the model. In the context of this exercise, it can be state that the actual costs of Room and Board deviate from the regression line by approximately 2,161 USD, on average. However, one problem of RSE is that it does not make much sense by itself, It depends on the metric of the dependent variable but there is no a clear interpretation about if it is large or small.

summary(costs\_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

Second, the R-squared statistic. This represents the amount of variability in the costs of Room and Board accounted for by Early career pay. According to the summary of the model, 38.9% of the variability of Room and Board costs is explained by the Early career pay. Broadly speaking, it is a moderate result.

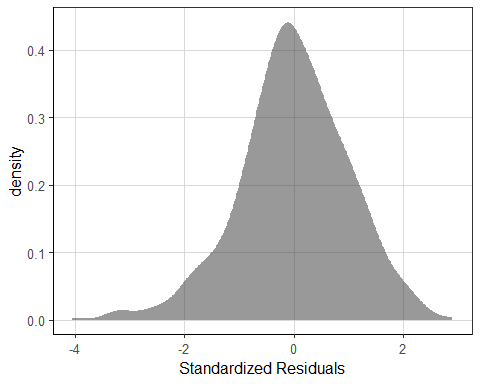
1. Explore and evaluate the model assumptions regarding the residual/error term. Summarize how well the model is meeting the assumptions, citing specific statistics or visualizations to justify your conclusions. 2 pts.

First, let us explore if the residuals () are nearly normally distributed around the regression line. Raw residuals or standardized residuals can be used. Here, the latest type will be explored because of the intuitiveness of its distribution.

Based on the density plot (Figure 4), the residuals appear quite normally distributed around the mean of zero. The distribution looks slightly left-skewed, however, the 95% of the residuals seem to fall properly within 2 standards deviations.

Figure 4

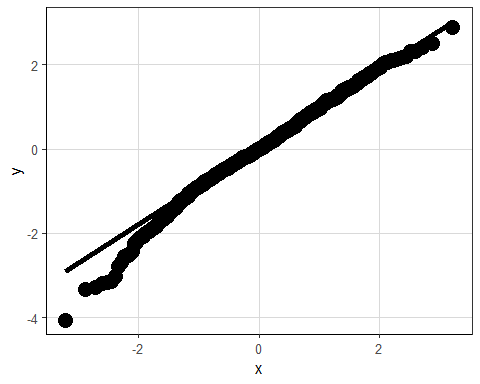
norm\_residuals <- augment(costs\_lm)  
gf\_density(~ .std.resid, data = norm\_residuals) %>%  
 gf\_labs(x = "Standardized Residuals")



This assumption can be also explore analyzing the match of the standardized residuals and theoretical percentiles. The standardized residuals closely follow the line, so normality can be assumed again.

Figure 5

ggplot(norm\_residuals, aes(sample = .std.resid)) +   
 stat\_qq(size = 5) +   
 stat\_qq\_line(size = 2)



If these evidences are not enough, the Shapiro-Wilk test could be analyzed. Running the test on residuals, the null hypothesis can be maintained at the .05 level of significance. In other words, the p-value (<0.001) of the Shapiro-Wilk test suggests that residuals follow a normal distribution.

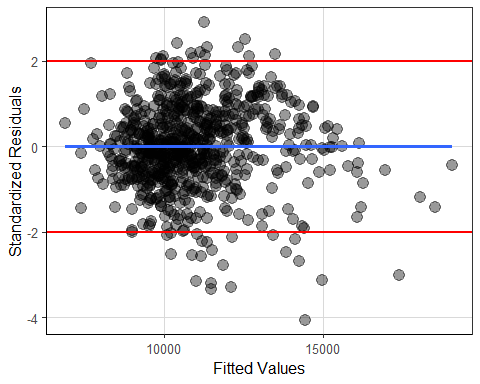
shapiro.test(costs\_lm$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: costs\_lm$residuals  
## W = 0.99064, p-value = 0.0001249

Second, the homogeneity of variance. When the regression line is laid horizontal (blue line), a high variance across the fitted values. Analyzing the standardized residuals, most of the observations falls within 2 standard deviations (red lines). The error variance look roughly equal distributed above and below the line, despite some of the values fall far away of 2 sd.

Figure 6

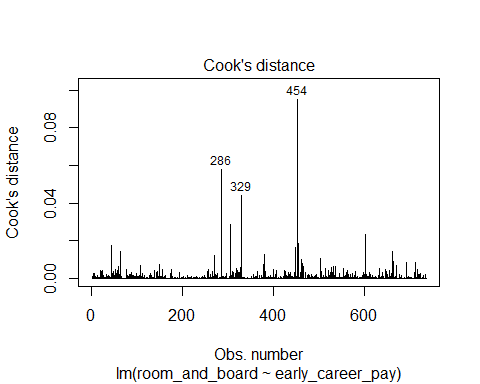
gf\_point(.std.resid ~ .fitted, data = norm\_residuals, size = 4, alpha = .4) %>%  
 gf\_hline(yintercept = ~ 2, color = 'red', size = 1) %>%  
 gf\_hline(yintercept = ~ -2, color = 'red', size = 1) %>%  
 gf\_smooth(method = 'lm', size = 1.2) %>%  
 gf\_labs(x = 'Fitted Values',  
 y = 'Standardized Residuals')



Third, influential points.Observing the plot of Cook’s distance, there are some influential values in the Early career pay that could be affecting the regression estimate. Despite this finding, it is hard to decide whether or not to eliminate this cases from the data set. For instance, using the rule of thumb for greater than 1 Cook’s distance, no one of the values should be removed.

Figure 7

plot(costs\_lm,4)

 In conclusion, the residuals are nearly normally distributed: the homogeneity of variance can be assumed but more exploration is needed using transformations; 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. Write out the null and alternative hypotheses based on your research question from #1.

More formally, the null hypothesis tend to be presented as follow:

or

These two last null hypothesis are expression of null effect between the variables.

On the other hand, the alternative hypothesis can be state as follow:

or

1. Summarize the statistical evidence surrounding the null or alternative hypotheses from question 7. In a few sentences, describe if the evidence provides support for or against the null hypothesis. Provide specific statistical evidence to support your justification.

The objective is to determine whether there is a relationship between Room and board costs and Early career pay. More specific, if the Early career pay helps to predict the variability in the Room and board costs that students have to pay. This can be supported or challenged via different alternatives. On of them is constructing Confidence Intervals for the parameters estimated.

Based on this results. with 95% of confidence, we fail to reject the null hypothesis in the case f the y-intercept but can reject it in the case of the slope. Let us analyse this in more detail.

sum\_x <- tuition %>%  
 summarise(mean\_early = mean(early\_career\_pay),  
 sum\_dev\_x\_sq = sum( (early\_career\_pay - mean\_early) ^ 2))  
se\_b0 <- sqrt(summary(costs\_lm)$sigma^2 \* ((1 / nrow(tuition)) + ( sum\_x[['mean\_early']]^2 / sum\_x[['sum\_dev\_x\_sq']]) ))  
se\_b1 <- sqrt(summary(costs\_lm)$sigma^2 / sum\_x[['sum\_dev\_x\_sq']])  
(CI\_b0 <- -129.7551288 + c(-1, 1) \* 1.96 \* 517.0658)

## [1] -1143.2041 883.6938

(CI\_b1 <- 0.2161361 + c(-1, 1) \* 1.96 \* 0.009988239)

## [1] 0.1965592 0.2357130

The C.I. of the y-intercept, constructed with 95% of confidence, is 1143.2041 - 883.6938, which means that the costs of Roam and board costs will fall, on average, between -1143.20 and 883.69 when the Early career pay equals zero. On the other hand, with 95% of confidence, one unit increase in Early career pay (one dollar increase) increase the Roam and board costs, on average, from 0.20 to 0.24. On top of that, as the first C.I. (95%) does contain zero, thus the β0 is not significantly different from 0. Conversely, as the second C.I. (95%) does not contain zero, We can conclude that the slope β1 is significantly different from zero at the .05 level.

summary(costs\_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

This interpretations are supported by the p-values. Based on the summary of the model, the p-value of the y-intercept is larger (0.802) than the accepted level of statistical significance (0.05), thus most probably the value of β0 in the population is close to zero. Moreover, the p-value of β1 (<0.001) indicates that there is a low probability of observing this value by chance under the null.

1. Based on the justification in #8, 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.

We have found that, in the current status of the model, the y-intercept is not significant different from zero. Some kind of linear or statistical transformation should be explored in order to improve this situation. In the case of the slope, evidence about statistical significance was provided, which means that Early career pay affects the costs of Room and board.

Now, in practical terms we could conclude that for one dollar increase in the early career pay, the room and board costs increase, on average, in .216. Of course, a more interpretable option would be decrease the metric of the predictor variable, for example, dividing the values by 1.000.

The results are useful, especially if the correlation coefficient and estimated slope are considered. The early career pay increase the costs of room and board, and that is something that students should take into account when make their decisions about where studying.