# STA 310: Homework 1

Matías Pinto

## Instructions

- Write all narrative using full sentences. Write all interpretations and conclusions in the context of the data.
- Be sure all analysis code is displayed in the rendered pdf.
- If you are fitting a model, display the model output in a neatly formatted table. (The `tidy` and `kable` functions can help!)
- If you are creating a plot, use clear and informative labels and titles.
- Render and back up your work reguarly, such as using Github.
- When you're done, we should be able to render the final version of the Rmd document to fully reproduce your pdf.
- Upload your pdf to Gradescope. Upload your Rmd, pdf (and any data) to Canvas.

## Exercises

Exercises 1 - 4 are adapted from exercises in Section 1.8 of @roback2021beyond.

### Exercise 1

Consider the following scenario:

> Researchers record the number of cricket chirps per minute and temperature during that time. They use linear regression to investigate whether the number of chirps varies with temperature.

a. Identify the response and predictor variable.

Response variable: Number of chirps per minute Predictor variable: Temperature.

b. Write the complete specification of the statistical model.

$$\text{Chirps}_i = \beta_0 + \beta_1 \times \text{temperature}_i + \epsilon$$

c. Write the assumptions for linear regression in the context of the problem.

1. Linearity: The mean of the number of chirps per minute has a linear relationship with the mean temperature.
2. Independence: The number of chirps recorded for one specific time period does not depend on the number of chirps recorded for another time period.
3. Normality: The number of chirps per minute follows a normal distribution at each level of temperature.
4. Equal variance: Variance of the number of chirps per minute is the same at all temperatures.

## Exercise 2

Consider the following scenario:

> A randomized clinical trial investigated postnatal depression and the use of an estrogen patch. Patients were randomly assigned to either use the patch or not. Depression scores were recorded on 6 different visits.

a. Identify the response and predictor variables.

Response variable: Depression scores. Predictor variable: Use of estrogen patch (categorical).

b. Identify which model assumption(s) are violated. Briefly explain your choice.

Independence is violated, because each patient has 6 depression scores recorded (for each visit), which means possible repeated values on the same patient, producing correlation and therefore data is not independent.

## Exercise 3

Use the Kentucky Derby case study in Chapter 1 of *Beyond Multiple Linear Regression.*

a. Consider Equation (1.3) in Section 1.6.3. Show why we have to be sure to say "holding year constant", "after adjusting for year", or an equivalent statement, when interpreting $\beta_2$.

In this equation, $\beta_2$ represent the **partial** change of `Y` when `Fast` increases by one unit, for horses within the same `Year`. If we do not keep the `Year` value constant (or fixed), then the interpretation of $\beta_2$ would confound the effect of `Year` with the effect of `Fast`.

b. Briefly explain why there is no error (random variation) term $\epsilon_i$ in Equation (1.4) in Section 1.6.6?

Equation 1.4 is not the statistical model, therefore there is no error term as it represents the equation produced after fitting the LLSR. It shows the expected value of $Y_i$, not the full model for the observed response variable.

## Exercise 4

The data set `kingCountyHouses.csv` in the `data` folder contains data on over 20,000 houses sold in King County, Washington (@kingcounty).

We will use the following variables:

- `price` = selling price of the house
- `sqft` = interior square footage

*See Section 1.8 of Beyond Multiple Linear Regression for the full list of variables.*

```r
library(dplyr)
Houses <- read.csv("../../data/kingCountyHouses.csv")
```

a. Fit a linear regression model with `price` as the response variable and `sqft` as the predictor variable (Model 1). Interpret the slope coefficient in terms of the expected change in price when `sqft` increases by 100.

```
model1 <- lm(price ~ sqft, data = Houses)
summary(model1)
```

```
## 
## Call:
## lm(formula = price ~ sqft, data = Houses)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1476062  -147486   -24043   106182  4362067 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -43580.743   4402.690  -9.899   <2e-16 ***
## sqft           280.624      1.936 144.920   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 261500 on 21611 degrees of freedom
## Multiple R-squared:  0.4929, Adjusted R-squared:  0.4928 
## F-statistic: 2.1e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

The model has the following equation

$$\widehat{\text{Price}} = -43,580 + 280.6 \times \widehat{\text{sqft}}$$

This means, an increase of one unit in `sqft` implies an increase in the expected house price by $280.62 on average. Now, by multiplying the square foot by 100, we get $280 \times 280.624 = 28,062.4$. Therefore, the `price` of a house is expected to increase by $28,062 for each additional 100 square feet.

b. Fit Model 2, where `logprice` (the natural log of price) is now the response variable and `sqft` is still the predictor variable. How is the `logprice` expected to change when `sqft` increases by 100?

```
Houses <- Houses |>
  mutate(logprice = log(price))

model2 <- lm(logprice ~ sqft, data = Houses)
summary(model2)
```

```
## 
## Call:
## lm(formula = logprice ~ sqft, data = Houses)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.97781 -0.28543  0.01472  0.26070  1.27628 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.222e+01  6.374e-03  1916.9   <2e-16 ***
## sqft        3.987e-04  2.803e-06   142.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3785 on 21611 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4835
## F-statistic: 2.023e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

The model has the following equation

$$\widehat{\log \text{price}} = -12.2 + 0.000399 \times \widehat{\text{sqft}}$$

Here the slope is $\beta_1 = 0.000399$, so for a 100-sqft increase we would get $0.000399 \times 100 = 0.0399$. This means that when `sqft` increases by 100, then the expected log price increases by 0.0399.

c. Recall that $log(a) - log(b) = log(\frac{a}{b})$. Use this to derive how the `price` is expected to change when `sqft` increases by 100 based on Model 2.

Let $x$ and $x + 100$ be two different `sqft` measurements. Then, according to our model, we will have that

$$\widehat{\log \text{price}_1} = -12.2 + 0.000399 \times x \tag{1}$$

$$\widehat{\log \text{price}_2} = -12.2 + 0.000399 \times (x + 100) \tag{2}$$

By doing (2) - (1), we'll get

$$\begin{aligned}
\widehat{\log(\text{price}_2)} - \widehat{\log(\text{price}_1)} &= -12.2 + 0.000399(x + 100) - (-12.2 + 0.000399x) \\
\log\left(\frac{\widehat{\text{price}_2}}{\widehat{\text{price}_1}}\right) &= 0.000399(x + 100) - 0.000399x \\
\log\left(\frac{\widehat{\text{price}_2}}{\widehat{\text{price}_1}}\right) &= 0.000399 \times 100 = 0.0399 \\
\frac{\widehat{\text{price}_2}}{\widehat{\text{price}_1}} &= e^{0.0399} \approx 1.041
\end{aligned}$$

With this, we can say that an increase in 100 `sqft` produces an increase in house price of 4.1% on average.

d. Fit Model 3, where `price` and `logsqft` (the natural log of sqft) are the response and predictor variables, respectively. How does the price expected to change when sqft increases by 10%? *As a hint, this is the same as multiplying sqft by 1.10.*

Click here for notes on interpreting model effects for log-transformed response and/or predictor variables.

```r
Houses <- Houses |>
  mutate(logsqft = log(sqft))

model3 <- lm(price ~ logsqft, data = Houses)
summary(model3)
```

```
##
## Call:
## lm(formula = price ~ logsqft, data = Houses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -606252 -170067  -33139  106342 6183772
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3451377      35169  -98.14   <2e-16 ***
## logsqft       528648       4651  113.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 290400 on 21611 degrees of freedom
## Multiple R-squared:  0.3742, Adjusted R-squared:  0.3742
## F-statistic: 1.292e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

The model has the following equation

$$\widehat{\text{price}} = -3,451,377 + 528,648 \times \widehat{\log \text{sqft}}$$

Then, an increase in 10% for `sqft` means we now get 1.10 × sqft, therefore, our equations become

$$\begin{align}
\widehat{\text{price}}_1 &= -3,451,377 + 528,648 \times \log \text{sqft} \tag{3}\\
\widehat{\text{price}}_2 &= -3,451,377 + 528,648 \times \log 1.1\text{sqft} \tag{4}
\end{align}$$

Calculating (4) - (3) gives us

$$\begin{aligned}
\widehat{\text{price}}_2 - \widehat{\text{price}}_1 &= -3,451,377 + 528,648 \times \log 1.1\text{sqft} - (-3,451,377 + 528,648 \times \log \text{sqft})\\
\widehat{\text{price}}_2 - \widehat{\text{price}}_1 &= 528,648 \times (\log 1.1\text{sqft} - \log \text{sqft})\\
\widehat{\text{price}}_2 - \widehat{\text{price}}_1 &= 528,648 \times \log \left(\frac{1.1\text{sqft}}{\text{sqft}}\right)\\
\widehat{\text{price}}_2 - \widehat{\text{price}}_1 &= 528,648 \times \log 1.10 \approx 50,400
\end{aligned}$$

Therefore, a 10% increase in `logsqft` produces an increase in house price of approximately 50,400 on average.

## Exercise 5

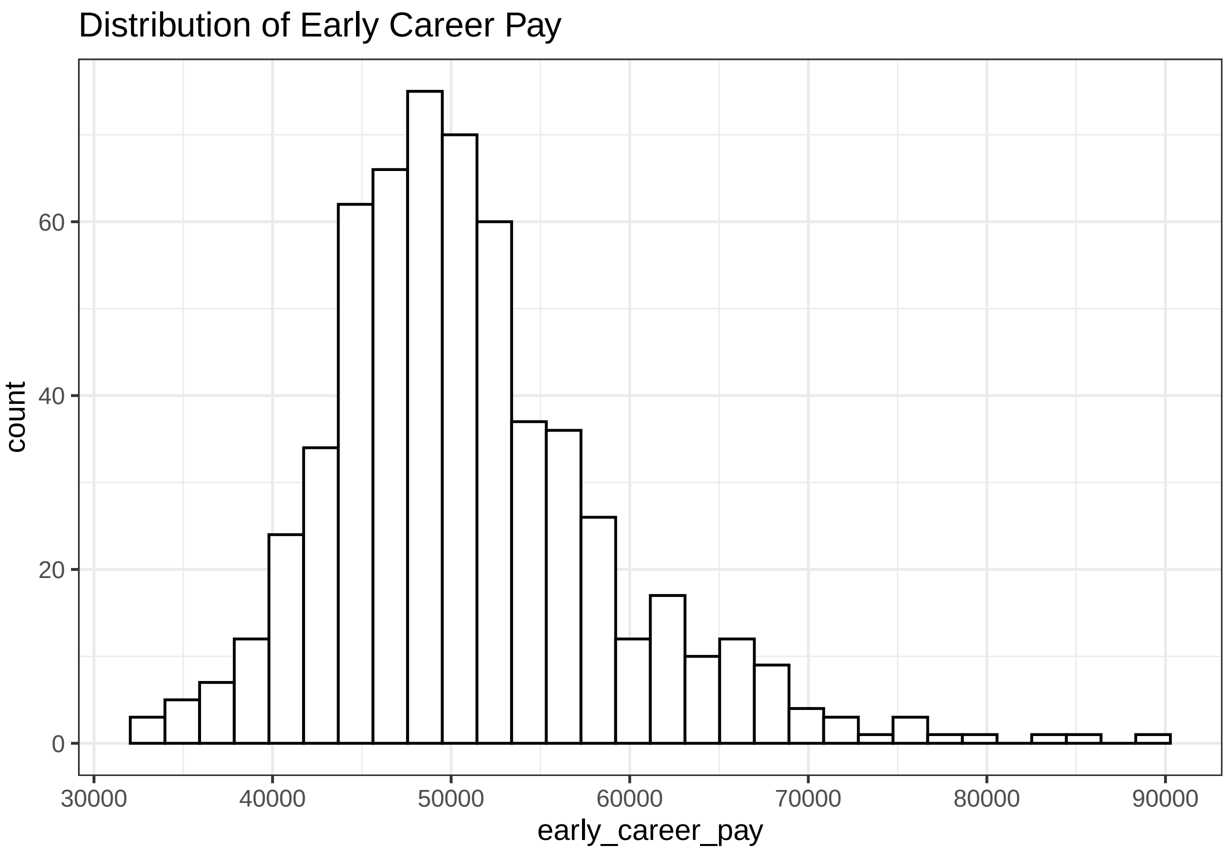
The goal of this analysis is to use characteristics of 593 colleges and universities in the United States to understand variability in the early career pay, defined as the median salary for alumni with 0 - 5 years of experience. The data was obtained from TidyTuesday College tuition, diversity, and pay, and was originaly collected from the PayScale College Salary Report.

The data set is located in `college-data.csv` in the `data` folder. We will focus on the following variables:

| variable | class | description |
|---|---|---|
| name | character | Name of school |
| state_name | character | state name |
| type | character | Public or private |
| early_career_pay | double | Median salary for alumni with 0 - 5 years experience (in US dollars) |
| stem_percent | double | Percent of degrees awarded in science, technology, engineering, or math subjects |
| out_of_state_total | double | Total cost for in-state residents in USD (sum of room & board + out of state tuition) |

a. Visualize the distribution of the response variable `early_career_pay`. Write 1 - 2 observations from the plot.

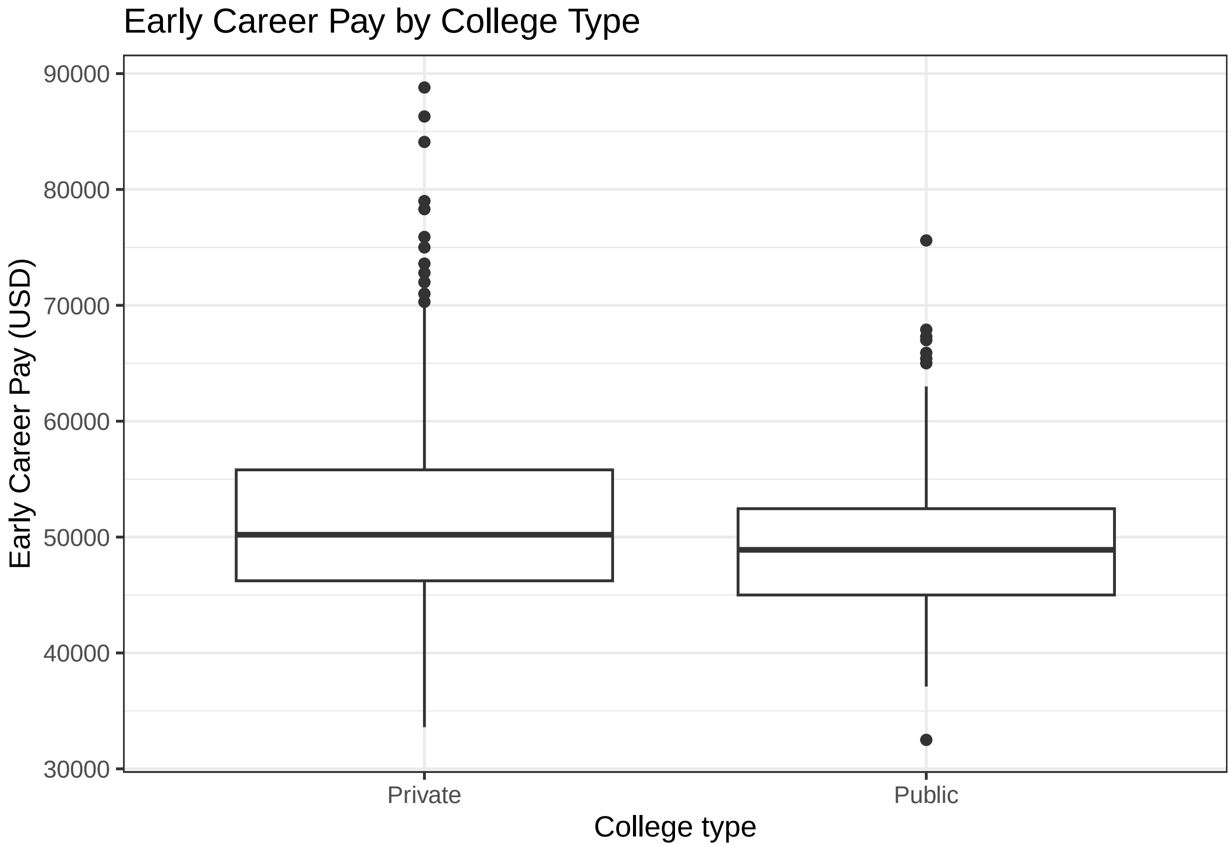
```r
library(ggplot2)
College <- read.csv("../../data/college-data.csv")

ggplot(data = College, aes(x = early_career_pay)) +
  geom_histogram(fill = "white", color = "black") +
  labs(
    title = "Distribution of Early Career Pay",
    xlab = "Early Career Pay (USD)",
    ylab = "Count"
  ) +
  theme_bw()
```



We can observe how the distribution is right-skewed, which indicates that most college have lower-to-moderate salary, while a smaller number of schools have a higher early-career pay. We can also observe that the range seems to be broad as most of the colleges have a salary between 30 to 60k USD, while a few schools fall in the 60k to 90k USD range.

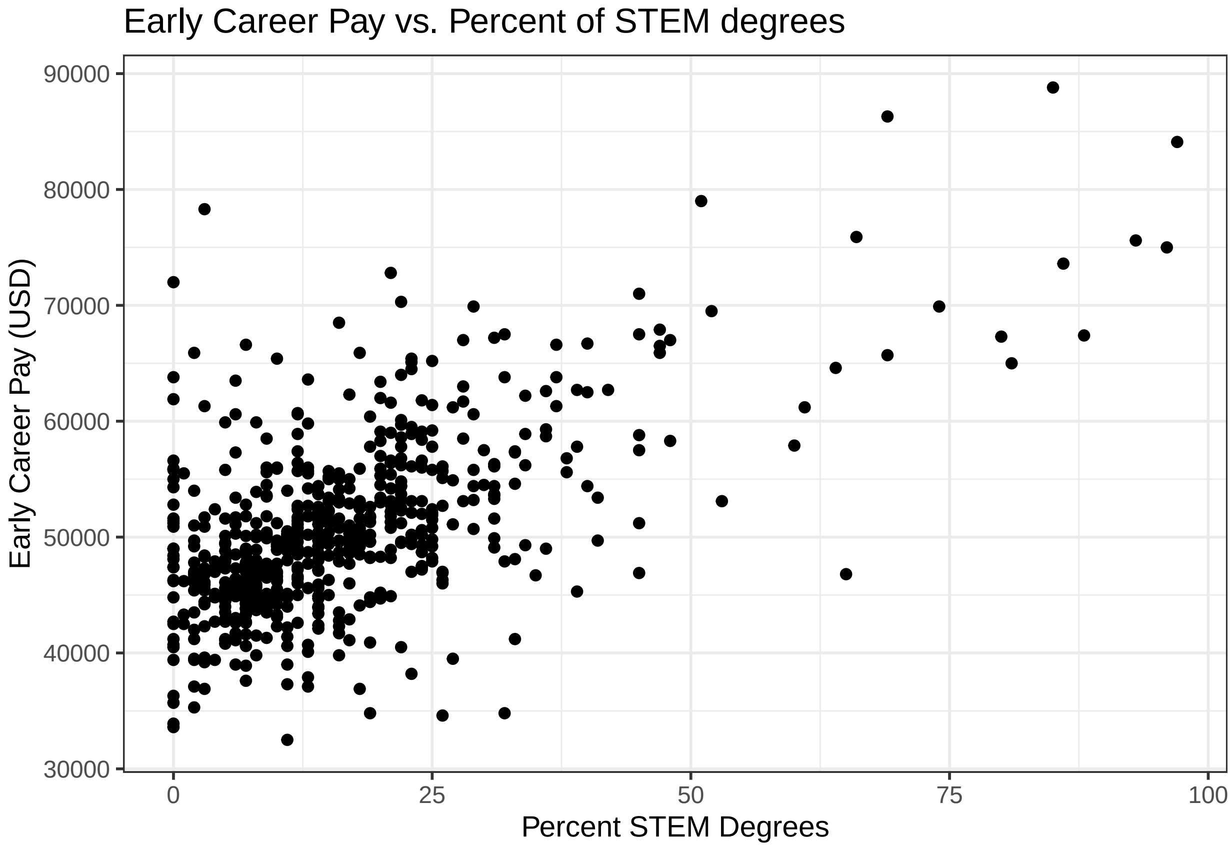
b. Visualize the relationship between (i) `early_career_pay` and `type` and (ii) `early_career_pay` and `stem_percent`. Write an observation from each plot.

```r
ggplot(College, aes(x = type, y = early_career_pay)) +
  geom_boxplot() +
  labs(
    title = "Early Career Pay by College Type",
    x = "College type",
    y = "Early Career Pay (USD)"
  ) +
  theme_bw()
```



We can observe how private college tend to have a slightly higher early career pay than public colleges, with a higher median and even higher outliers. However, other than the outliers, the boxplot shapes are similar.

```r
ggplot(College, aes(x = stem_percent, y = early_career_pay)) +
  geom_point() +
  labs(
    title = "Early Career Pay vs. Percent of STEM degrees",
    x = "Percent STEM Degrees",
    y = "Early Career Pay (USD)"
    ) +
  theme_bw()
```



There seems to be a positive correlation between the percentage of STEM degrees and early career pay, though there is a lot of variability towards lower values.

c. Below is the specification of the statistical model for this analysis. Fit the model and neatly display the results using 3 digits. Display the 95% confidence interval for the coefficients.

$$\begin{aligned} early\_career\_pay_i = \beta_0 + \beta_1\ out\_of\_state\_total_i + \beta_2\ type \quad &(5)\\ + \beta_3\ stem\_percent_i + \beta_4\ type * stem\_percent_i \quad &(6)\\ + \epsilon_i, \quad \text{where } \epsilon_i \sim N(0, \sigma^2) \quad &(7) \end{aligned}$$

```
library(tidyverse)
library(knitr)
library(broom)

model4 <- lm(early_career_pay ~
               out_of_state_total + type + stem_percent + type * stem_percent,
             data = College)

tidy_m4 <- tidy(model4, conf.int = T, conf.level = 0.95)
kable(tidy_m4, digits = 3)
```

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 36217.704 | 850.222 | 42.598 | 0.000 | 34547.862 | 37887.546 |
| out_of_state_total | 0.253 | 0.018 | 13.692 | 0.000 | 0.217 | 0.289 |
| typePublic | 1185.020 | 768.752 | 1.541 | 0.124 | -324.813 | 2694.853 |
| stem_percent | 214.306 | 19.300 | 11.104 | 0.000 | 176.402 | 252.211 |
| typePublic:stem_percent | 49.538 | 33.875 | 1.462 | 0.144 | -16.992 | 116.069 |

d. How many degrees of freedom are there in the estimate of the regression standard error $\sigma$?

Note that $d.f.s. = n - p$ where $n = 593$ is the number of observations (or colleges) while $p = 5$ is the number of parameters (including the intercept). Thus, there are 588 degrees of freedom.

e. What is the 95% confidence interval for the amount in which the intercept for public institutions differs from private institutions?

By looking at the table from (c), we can tell that the term `typePublic` has a value of 1,185 with a 95% C.I. of -324.813 and 2694.853, which means that the estimated difference between public and private institutions is 1,185 USD, and we are 95% confident that the Public institutions intercept differs from Private institutions by somewhere between -325 USD to 2,695 USD. Note that since this C.I. includes 0, it is not statistically significant at 0.05 level.

### Exercise 6

Use the analysis from the previous exercise to write a paragraph (~ 4 - 5 sentences) describing the differences in early career pay based on the institution characteristics. *The summary should be consistent with the results from the previous exercise, comprehensive, answers the primary analysis question, and tells a cohesive story (e.g., a list of interpretations will not receive full credit).*

The analysis shows that early career pay varies depending on college characteristics. For instance, private college tend to have a slightly higher median early career pay than public institutions, with this increase being more pronounced at the upper outliers. However, the difference in intercepts between public and private institutions was not statistically significant according to the Confidence Interval. We can also observe that colleges with a higher percentage of STEM degrees have higher early career pay, which correspond to the positive value of the slope for `stem_percent`. Moreover, the interaction between `type` and `stem_percent` was not statistically significant, which indicates that the effect of STEM degree count on pay is similar for both public and private institutions. In conclusion, we can deduce that cost and STEM percent contribute to differences in early career pay.

## Grading

| **Total** | **50** |
|---|---|
| Ex 1 | 8 |
| Ex 2 | 4 |
| Ex 3 | 7 |
| Ex 4 | 12 |
| Ex 5 | 12 |
| Ex 6 | 4 |
| Workflow & formatting | 3 |

The “Workflow & formatting” grade is to based on the organization of the assignment write up along with the reproducible workflow. This includes having an organized write up with neat and readable headers, code, and narrative, including properly rendered mathematical notation. It also includes having a reproducible Rmd/Quarto document that can be rendered to reproduce the submitted PDF.