STA 310: Homework 1

Amaris

```
# load packages
library(tidyverse)
library(tidymodels)
library(knitr)
library(patchwork)
library(dplyr)


# set default theme for ggplot2
ggplot2::theme_set(ggplot2::theme_bw())
```

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:**

   The response variable is the number of cricket chirps per minute, as this is the outcome being measured and analyzed. The predictor variable is temperature, since the researchers are interested in whether changes in temperature are associated with changes in the number of cricket chirps.

b. **Write the complete specification of the statistical model:**

Let Y_i denote the number of cricket chirps per minute observed during minute i, and let X_i denote the temperature during that same minute.

The relationship between temperature and chirping rate is modeled using simple linear regression:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n.$$

The error terms are assumed to be independent and identically distributed with a normal distribution around the regression line:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

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

- **Linearity:** The expected number of cricket chirps per minute is a linear function of temperature. This means that, on average, changes in temperature are associated with proportional changes in the chirping rate, and this relationship can be adequately described by a straight line.
- **Independence:** The observations are independent of one another. In this context, this implies that the number of chirps recorded during one minute does not influence the number of chirps recorded during any other minute.
- **Normality:** For a fixed temperature, the deviations of the observed number of chirps per minute from the regression line are normally distributed. That is, the error terms in the model follow a normal distribution centered at zero.
- **Equal Variance:** The variability in the number of cricket chirps per minute is constant across all temperatures. This means that the spread of chirp counts around the regression line does not systematically increase or decrease as temperature increase or decrease.

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:** The response variable is the depression score recorded for each patient at each visit. The predictor variables include treatment assignment (use of the estrogen patch versus no patch) and time (visit number). Because depression scores are recorded repeatedly for the same patients, each patient contributes multiple observations over time.

b. **Identify which model assumption(s) are violated. Briefly explain your choice:** The primary linear regression assumption that is violated is independence. The independence assumption requires that all observations are independent of one another. In this study, depression scores are measured repeatedly on the same patients across six visits, meaning that observations from the same patient are likely correlated. For example, a patient with a high depression score at one visit is likely to have a similar score at subsequent visits. As a result, the depression scores within a patient are not independent. Normality could also potentially be violated. The normality assumption requires that, for a given set of predictor values, the errors in depression scores are normally distributed. Depression scores are often bounded scales and may exhibit skewness, especially if many patients have very low or very high scores. As a result, the distribution of residuals may deviate from normality.

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 β_2.**

   In the Kentucky Derby case study, the coefficient β_2 in Equation (1.3) represents the effect of track condition (fast versus non-fast) after accounting for year. To see this, suppose we compare two predicted winning speeds $\hat{Y}_1$ and $\hat{Y}_2$, where $\text{Fast}_1 = 1$ and $\text{Fast}_2 = 0$. The difference in predicted winning speeds can be written as

   $$\hat{Y}_1 - \hat{Y}_2 = \beta_1(\text{Yearnew}_1 - \text{Yearnew}_2) + \beta_2(1 - 0).$$

   This difference equals β_2 only when year is held fixed, that is, when $\text{Yearnew}_1 = \text{Yearnew}_2$. Therefore, the interpretation of β_2 must explicitly state that year is held constant, or equivalently, that the comparison is made after adjusting for year.

   This clarification is necessary because year is associated with both track condition and winning speed in the data. More recent years tend to have more races run under fast conditions, and winning speeds have also increased over time. If year were not accounted for, differences in winning speed between fast and non-fast races would partly reflect differences in the years in which those races occurred rather than the effect of track condition itself. Including year in the model isolates the effect of track condition, so β_2 measures the expected change in winning speed due to fast versus non-fast conditions for a fixed year.

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

   There is no error (random variation) term ε_i in Equation (1.4) because the equation represents the fitted (estimated) values $\hat{Y}_i$ from the regression model, rather than the data-generating process for the observed responses Y_i. Equation (1.4) gives the predicted winning speed obtained by substituting the estimated regression coefficients into the linear predictor. Once the coefficients are estimated, $\hat{Y}_i$ is a deterministic function of Yearnew_i, Fast_i, and their interaction, and therefore no random error term is included.

   In contrast, the error term ε_i appears in the statistical model for the observed winning speeds Y_i to represent unexplained variability around the regression line. This random variation is reflected in the residuals $Y_i - \hat{Y}_i$ but is not part of the equation defining the fitted values themselves.

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
kingCountyHouses <- 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.**

```
price_fit <- lm(price ~ sqft, data = kingCountyHouses)
tidy(price_fit) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-43580.743	4402.690	-9.899	0
sqft	280.624	1.936	144.920	0

In Model 1, the estimated slope coefficient for `sqft` is 280.624. This means that, on average, an increase of one additional square foot in a home's size is associated with an increase of approximately $280.62 in the expected sale price of the home in King County, Washington.

Equivalently, when square footage increases by 100 square feet, the expected sale price of a house increases by approximately $28,062, holding all else constant.

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?**

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

price_fit_2 <- lm(logprice ~ sqft, data = kingCountyHouses)
tidy(price_fit_2) |>
  kable(digits = 5)
```

term	estimate	std.error	statistic	p.value
(Intercept)	12.21846	0.00637	1916.8830	0
sqft	0.00040	0.00000	142.2326	0

In Model 2, the estimated slope coefficient for `sqft` is 0.00040. This means that, on average, an increase of one additional square foot in a home's size is associated with an increase of 0.00040 in the expected value of the natural log of the sale price.

When square footage increases by 100 square feet, the expected value of `logprice` increases by approximately 0.04.

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.**

From Model 2, the fitted regression equation is

$$\mathbb{E}[\log(\text{price}) \mid \text{sqft}] = \beta_0 + \beta_1 \,\text{sqft},$$

where the estimated slope is $\hat{\beta}_1 = 0.00040$.

Consider two houses that differ in size by 100 square feet. Let $\text{sqft} = s$ for the smaller house and $\text{sqft} = s + 100$ for the larger house. The difference in expected log prices is

$$\mathbb{E}[\log(\text{price}_{s+100})] - \mathbb{E}[\log(\text{price}_s)] = \beta_1(s + 100) - \beta_1 s = 100\beta_1.$$

Using the logarithm identity $\log(a) - \log(b) = \log(a/b)$, this difference can be written as

$$\log\left(\frac{\mathbb{E}[\text{price}_{s+100}]}{\mathbb{E}[\text{price}_s]}\right) = 100\beta_1.$$

Exponentiating both sides yields

$$\frac{\mathbb{E}[\text{price}_{s+100}]}{\mathbb{E}[\text{price}_s]} = e^{100\beta_1}.$$

Substituting $\hat{\beta}_1 = 0.00040$,

$$\frac{\mathbb{E}[\text{price}_{s+100}]}{\mathbb{E}[\text{price}_s]} = e^{0.04} \approx 1.041.$$

Based on Model 2, a 100-square-foot increase in square footage is associated with an expected increase in sale price of approximately **4.1%**. (Because the response was log-transformed, this interpretation is in terms of the median price on the original scale.)

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.

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

price_fit_3 <- lm(price ~ logsqft, data = kingCountyHouses)
tidy(price_fit_3) |>
  kable(digits = 5)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3451377.1	35169.348	-98.13594	0
logsqft	528647.5	4650.631	113.67221	0

From Model 3, the fitted regression equation is

$$\mathbb{E}[\text{price} \mid \text{sqft}] = \beta_0 + \beta_1 \log(\text{sqft}),$$

where $\mathbb{E}[\cdot]$ denotes the conditional mean of the sale price and the estimated slope is $\hat{\beta}_1 = 528{,}647.5$.

A 10% increase in square footage corresponds to multiplying sqft by 1.10. Let sqft $= s$ denote the original house size and sqft $= 1.10s$ the increased size. The change in the conditional mean price is

$$\begin{aligned}\mathbb{E}[\text{price}_{1.10s}] - \mathbb{E}[\text{price}_s] &= \beta_1 \left(\log(1.10s) - \log(s)\right) \\ &= \beta_1 \log(1.10),\end{aligned}$$

using the identity $\log(a) - \log(b) = \log(a/b)$.

Substituting $\hat{\beta}_1 = 528{,}647.5$ and $\log(1.10) \approx 0.0953$,

$$528{,}647.5 \times 0.0953 \approx 50{,}350.$$

Therefore, based on Model 3, a 10% increase in square footage is associated with an increase of approximately \$50,000 in the conditional mean sale price of a house.

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:

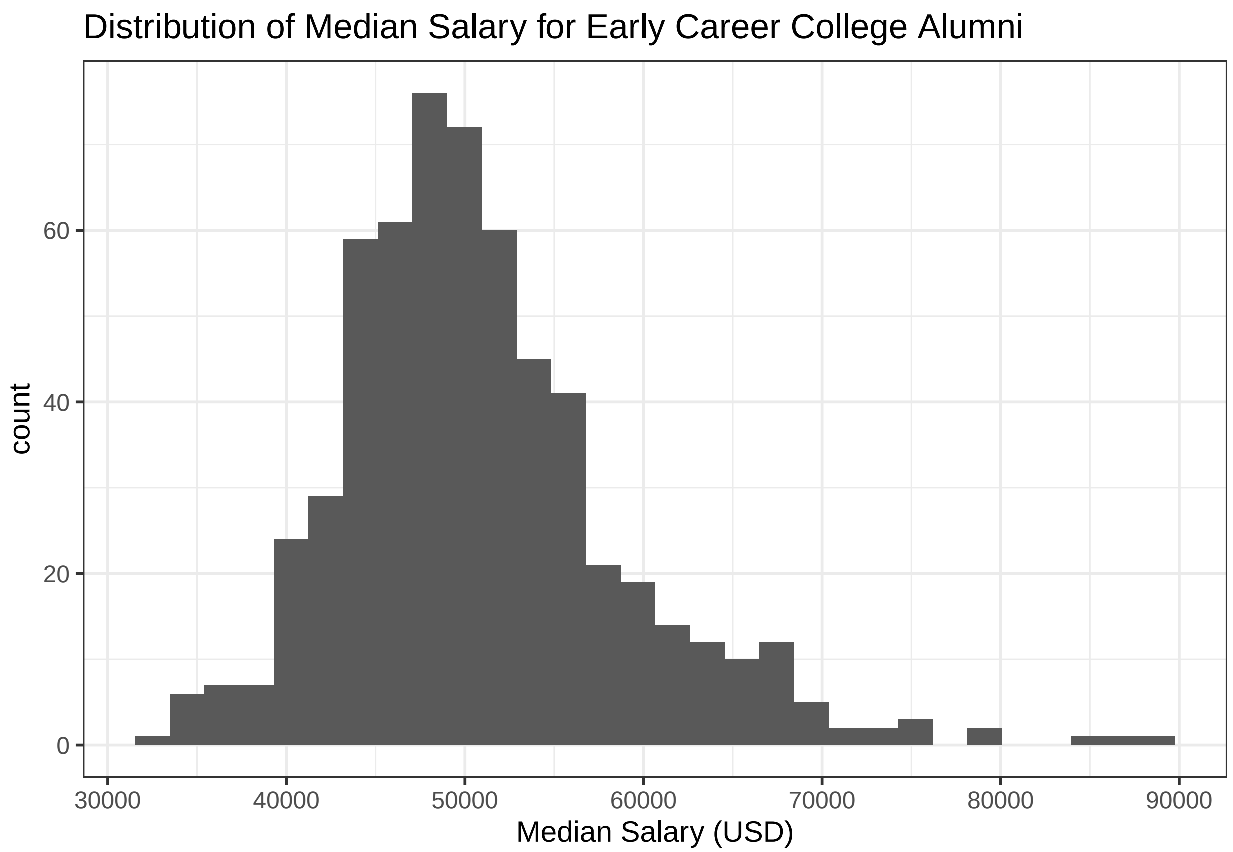
variable	class	description
name	character	Name of school
state_name	character	state name
type	character	Public or private
early_career_pay	double	Median salary for alumni with 0 - 5 years experience (in US dollars)
stem_percent	double	Percent of degrees awarded in science, technology, engineering, or 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
college_data <- read.csv("../data/college-data.csv")

ggplot(data = college_data, aes(x=early_career_pay)) +
  geom_histogram() +
  labs(
    x = "Median Salary (USD)",
    title = "Distribution of Median Salary for Early Career College Alumni"
  )
```

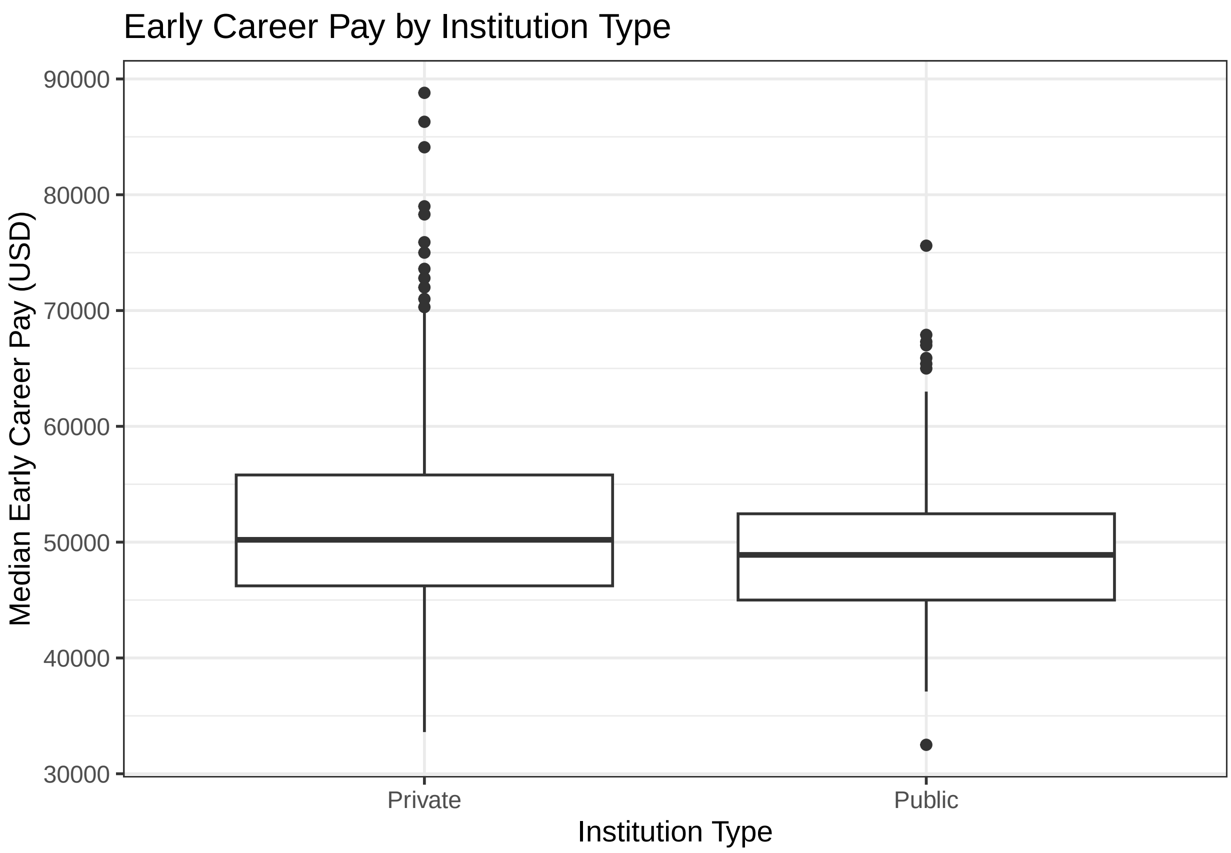
```
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```



Observations: The distribution of early career pay is unimodal and right-skewed, with most colleges having median early career salaries between approximately $40,000 and $55,000. There is a long right tail, indicating that a smaller number of colleges have substantially higher early career pay, with some values extending above $80,000.

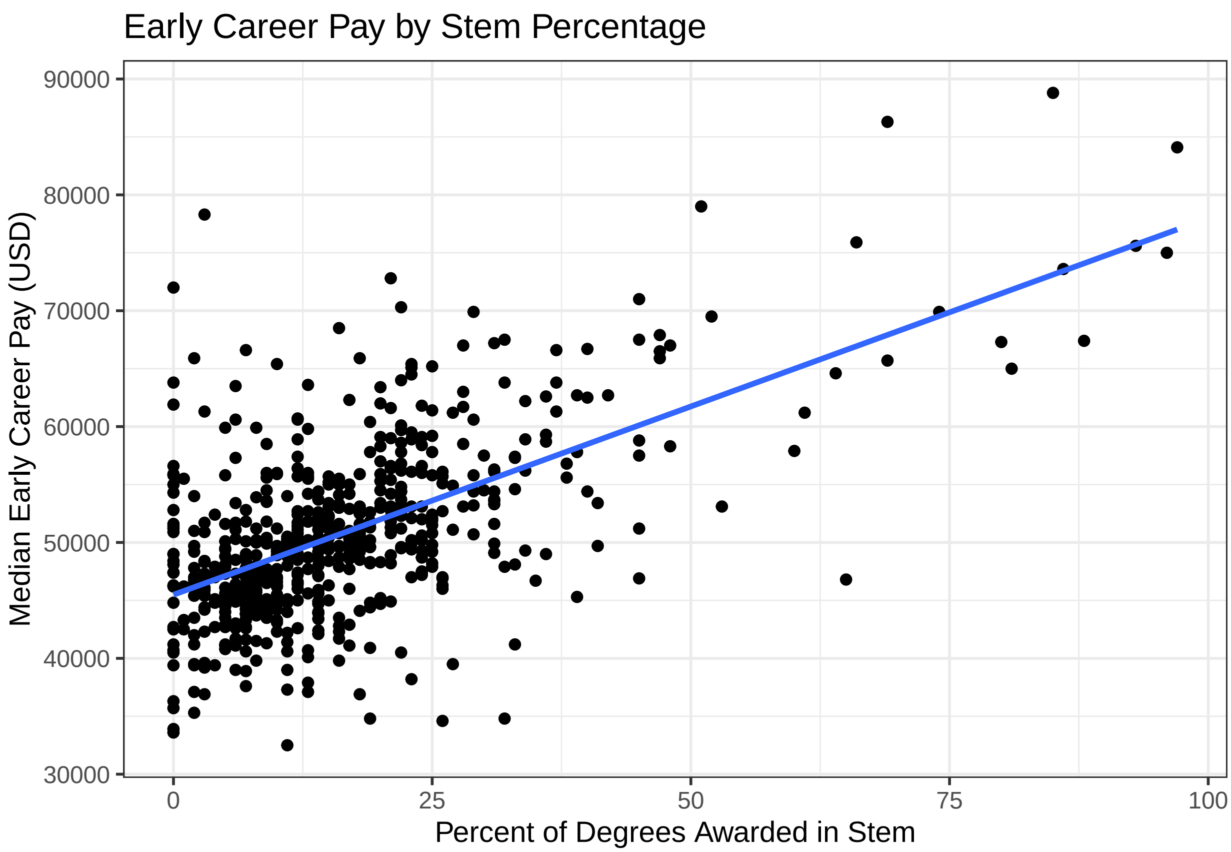
b. **Visualize the relationship between (i) `early_career_pay` and `type` and (ii) `early_career_pay` and `stem_percent`. Write an observation from each plot.****

```
ggplot(data = college_data, aes(x = type, y = early_career_pay)) +
  geom_boxplot() +
  labs(
    x = "Institution Type",
    y = "Median Early Career Pay (USD)",
    title = "Early Career Pay by Institution Type"
  )
```



```r
ggplot(data = college_data, aes(x = stem_percent, y = early_career_pay)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    x = "Percent of Degrees Awarded in Stem",
    y = "Median Early Career Pay (USD)",
    title = "Early Career Pay by Stem Percentage"
  )
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



From the boxplot, private institutions have a slightly higher median early-career pay than public institutions. The distribution for private schools also shows greater variability and more high-end outliers, indicating that while typical early-career pay is similar across institution types, private institutions are more likely to have programs associated with very high early-career salaries.

The scatterplot shows a positive linear relationship between the percentage of STEM degrees and median early-career pay. As `stem_percent` increases, the median early-career pay tends to increase, suggesting that institutions with a higher share of STEM programs generally have graduates who earn more early in their careers.

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.**

$$\begin{aligned} early\_career\_pay_i &= \beta_0 + \beta_1 \, out\_of\_state\_total_i + \beta_2 \, type && (1)\\ &+ \beta_3 \, stem\_percent_i + \beta_4 \, type * stem\_percent_i && (2)\\ &+ \epsilon_i, \quad \text{where } \epsilon_i \sim N(0, \sigma^2) && (3) \end{aligned}$$

```
early_career_pay_fit <- lm(
  early_career_pay ~ out_of_state_total + type + stem_percent + type:stem_percent,
  data = college_data
)

tidy(early_career_pay_fit, conf.int = TRUE, conf.level = 0.95) |>
  kable(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 σ?**

The dataset has 593 observations, 4 predictors, and 1 intercept. Thus, the estimate of the regression standard error has 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?**

Private institutions are the reference group. So, the difference in intercepts (public - private) is given directly by the coefficient beta Type Public. Estimate for typePublic is 1185.020 with 95% confidence interval (-324.813, 2694.853).

This means that, with 95% confidence, holding all other variables fixed and at the reference value of stem_percent, the intercept for public institutions could be anywhere from about $325 lower to $2695 higher than that of private institutions.

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).***

Early career pay varies systematically with several institutional characteristics. From the distribution of early career pay, salaries are right-skewed, with most institutions clustered around the mid-$40,000 to $55,000 range and a smaller number of schools offering substantially higher median pay. Comparing institution types, private institutions tend to have slightly higher median early career pay than public institutions, although the regression results indicate that this difference is not statistically significant at the 95% confidence level once other variables are accounted for. In contrast, the percentage of STEM degrees awarded is strongly associated with higher early career pay: both the scatterplot and the fitted regression line show a clear positive relationship, and the model estimates indicate that institutions with higher STEM representation have substantially higher median early career salaries. Finally, higher out-of-state costs are also positively associated with early career pay, suggesting that more expensive institutions tend to be associated with higher post-graduation earnings, even after controlling for institution type and STEM emphasis.

Grading

Total	**50**
Ex 1	8
Ex 2	4

Total	**50**
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.