Homework 1

Sonya Eason

Exercise 1

a) The response variable is the number of cricket chirps. The predictor variable is the temperature.

b)

$$y_i = \beta_0 \ + \ \beta_1 * x_i \ + \ \epsilon_i$$

where y_i is an observed number of chirps and x_i is the temperature at the minute when that number of chirps were observed.

c)

To conduct LLSR, we need to satisfy LINE assumptions. Specifically these are the follow assumptions in this context.

1. The mean number of cricket chirps per minute is linearly related to temperature.
2. Observations are independent. Specifically, the number of cricket chirps is not correlated with previous number of cricket chirps of future number of cricket chirps.
3. The number of cricket chirps follow a normal distribution at each temperature.
4. The variance in number of cricket chirps is equal for all temperatures.

Exercise 2

a)

The response variable is the depression score. The predictor variable is whether the patient was assigned to use an estrogen patch or not.

b)

The most obviously violated assumption is independence since we collect patient depression scores across multiple visits. We expect a person's data to be correlated with their previous visits.

In addition, the constant variance assumption is likely violated since we expect there could be more variation in postnatal depression of those who do not use the estrogen patch versus those who do use it.

Depending on what the depression score metric looks like, the normality condition could also be violated since we may have have depression scores that are more frequent on extreme ends of a metric like a lot of people that are highly depressed or not depressed. It's also hard to imagine a continuous equivalent or approximation of depression score that would look like a normal distribution.

Exercise 3

a) We have to make a statement that accounts for holding year constant because our estimate $\hat{\beta}_2$ will only speak only to the individual effect of Fast, which means if Yearnew shifts, there will be an additional effect unaccounted for.

Let's look at the winning speeds when only the fast variable changes in comparison to winning speeds when both variables change to show this.

$$\hat{y_1} = \hat{\beta}_0 + \hat{\beta}_1 * Yearnew_1 + \hat{\beta}_2 * Fast_1$$

$$\hat{y_2} = \hat{\beta}_0 + \hat{\beta}_1 * Yearnew_2 + \hat{\beta}_2 * Fast_2$$

When only the fast variable changes, which would mean $Yearnew1 = Yearnew2$, the estimated difference in winning speeds is simply $\hat{\beta}_2$ since Fast is an indicator variable.

That being said, the same can not be said about when we do not hold Yearnew constant.

Here we would also have to add the $\hat{\beta}_1 * (Yearnew_2 - Yearnew_1)$ to the $\hat{\beta}_2$ to find the difference in winning speed estimate between a fast and non-fast car without the year held constant.

As shown, in a multivariate setting, we need to hold the other variables constant to interpret the effects of a single coefficient.

b) There's no error term in Eq 1.4 because we are looking at the estimated winning speed. The estimated winning speed is like looking at the expected value and because it's an estimate, we do not need to account for error. Error is needed to account for the difference between the deterministic components and the truth. Since $\hat{y}$ is not the truth, but instead the estimated response, we do not need to account for the randomness that lies within the error terms.

Exercise 4

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(ggplot2)
library(knitr)
library(broom)
```

```
houses <- read_csv("kingCountyHouses.csv")
```

```
Rows: 21613 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (9): price, date, bedrooms, bathrooms, sqft, floors, waterfront, yr_buil...

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.
```

a)

```r
lm(price ~ sqft, houses) |>
  tidy() |>
  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

For each 100 square foot increase, houses prices in King County, Washington are expected to increase by $28,060 dollars.

```r
lm(log(price)~sqft, houses) |>
  tidy()|>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	12.218	0.006	1916.883	0
sqft	0.000	0.000	142.233	0

When square foot increases by 100, log price is expected to increase by $3.987 * 10^{-2}$.

c)

$$log(\hat{y_2}) - log(\hat{y_1}) = \hat{\beta}_0 + \hat{\beta}_1 * sqft_2 - (\hat{\beta}_0 + \hat{\beta}_1 * sqft_1)$$

$$log(\frac{\hat{y_2}}{\hat{y_1}}) = \hat{\beta}_1 * (sqft_2 - sqft_1)$$

$$\frac{\hat{y_2}}{\hat{y_1}} = e^{\hat{\beta}_1 * (sqft_2 - sqft_1)}$$

sqft increases by 100

$$\frac{\hat{y_2}}{\hat{y_1}} = e^{\hat{\beta}_1 * (100)}$$

```
exp(3.987e-04*100)
```

```
[1] 1.040675
```

The price is expected to change by a multiplicative factor of 1.040675 when sqft increases by 100.

d)

```
lm(price ~ log(sqft), houses) |>
  tidy() |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3451377.1	35169.348	-98.136	0
log(sqft)	528647.5	4650.631	113.672	0

$\hat{y_2} = \hat{\beta}_0 + \hat{\beta}_1 * log(sqft_2)$

$$\hat{y_1} = \hat{\beta}_0 + \hat{\beta}_1 * log(sqft1)$$

$$\hat{y_2} - \hat{y_1} = \hat{\beta}_0 + \hat{\beta}_1 * log(sqft2) - (\hat{\beta}_0 + \hat{\beta}_1 * log(sqft1))$$

$$\hat{y_2} - \hat{y_1} = \hat{\beta}_1 * log(\frac{sqft_2}{sqft_1})$$

square foot increases by 10%

$$\hat{y_2} - \hat{y_1} = \hat{\beta}_1 * log(\frac{1.1 * sqft_1}{sqft_1})$$

$$\hat{y_2} - \hat{y_1} = \hat{\beta}_1 * log(1.1)$$

```
528647.5 * log(1.1)
```

```
[1] 50385.49
```

The price is expected to increase by $50,385.49 when sqft increases by 10%.

Exercise 5

```
college <- read_csv("college-data.csv")
```

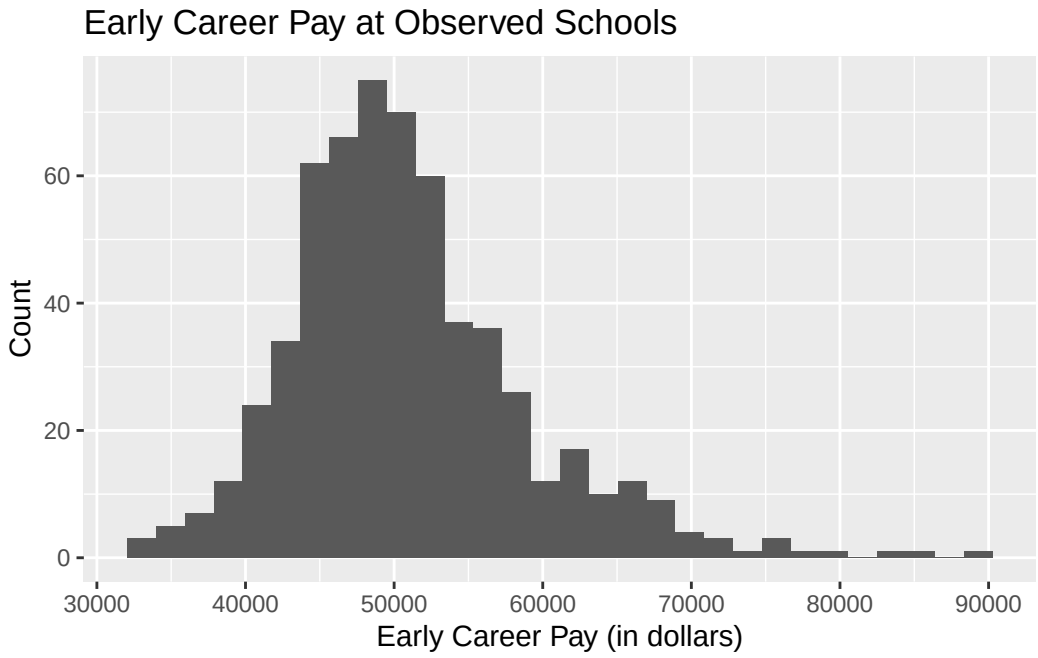
```
Rows: 593 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (5): name, state, state_code, type, degree_length
dbl (10): room_and_board, in_state_tuition, in_state_total, out_of_state_tui...

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.
```

a)

```
ggplot(college, aes(early_career_pay))+
  geom_histogram() +
  labs(x = "Early Career Pay (in dollars)",
       y = "Count",
       title = "Early Career Pay at Observed Schools")
```

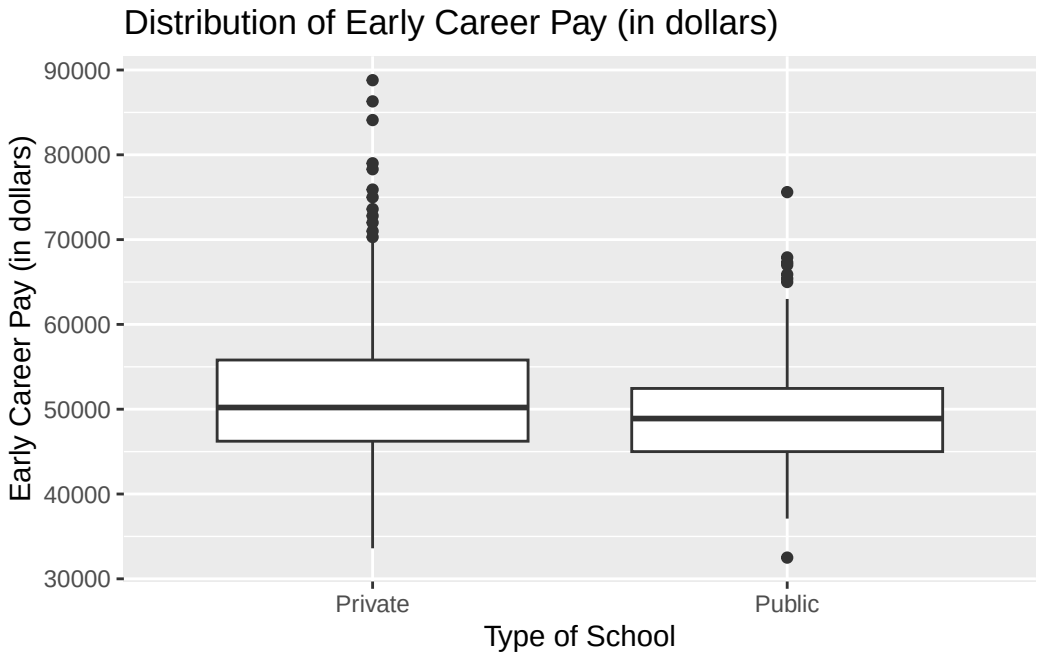
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plot appears to follow a bell curve shape like in a normal distribution. The most frequent early_career_pay values are slightly below 50k.

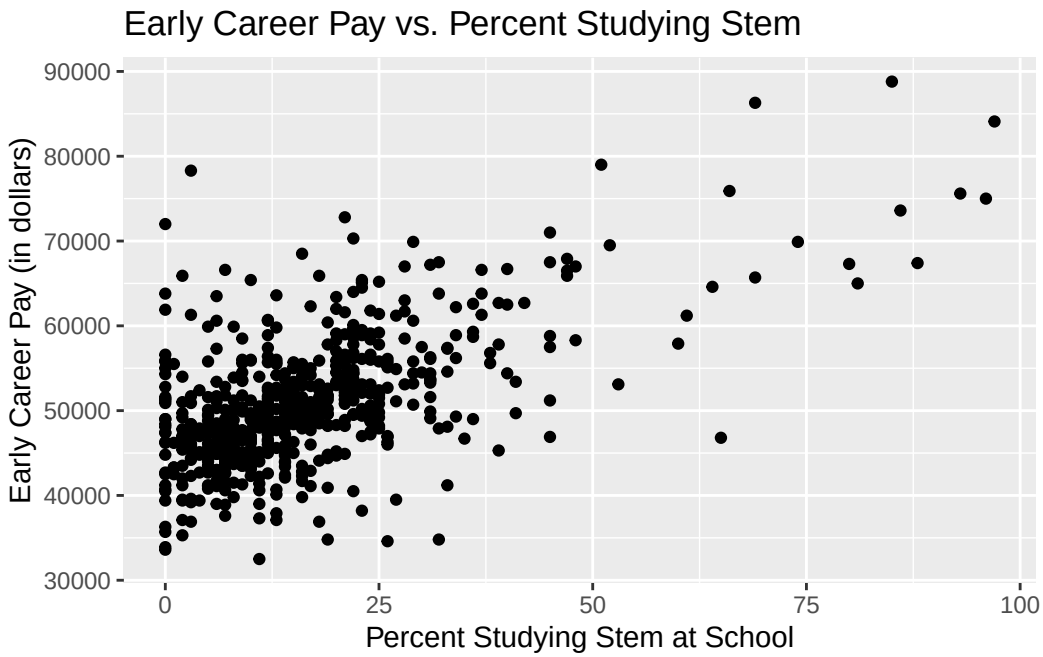
b)

```
ggplot(college, aes(type, early_career_pay)) +
geom_boxplot() +
  labs(x = "Type of School",
       y = "Early Career Pay (in dollars)",
       title = "Distribution of Early Career Pay (in dollars)")
```



It preliminarily appears that the private schools have a slightly higher early_career pay.

```
ggplot(college, aes(stem_percent, early_career_pay)) +
geom_point() +
  labs(
    x = "Percent Studying Stem at School",
    y = "Early Career Pay (in dollars)",
    title = "Early Career Pay vs. Percent Studying Stem"
  )
```



As a school's stem_percent increase, it seems their early_career pay value increases.

c)

```
lm(early_career_pay ~ out_of_state_total + type + stem_percent
   + type * stem_percent, college) |>
  tidy(conf.int = TRUE) |>
  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)

n - p - 1 degrees of freedom

```r
n = nrow(college)
#non-intercept terms
p = 4

dof = n - p - 1
```

There are 588 degrees of freedom in the estimate of the regression standard error σ.

e)

The 95% CI for amount in which intercept is higher for public institutions is between -324.813 and 2694.853.

Exercise 6

A college's institutional characteristics shed light on differences in early career pay of their attendees. In an analysis that considered out of state total, type of school (public or private), percentage of stem students, and the interaction between type of school and percent of stem students, low p-values and confidence intervals that do not incorporate 0 appear to provide evidence that percent of stem students and out of state total have a positive effect on the early career pay of school attendee. The other terms analyzed do not seem to be significant. To delve into the noted positive effects, we can say that we expect the early career pay of attendees to increase by 25 cents for each additional out of state student enrolled, holding all other variables constants, and we expect the early career pay of attendees to increase by 214 for each additional percent of stem students enrolled. To truly make the aforementioned inference on the basis of linear least squares, we need to ensure we satisfy conditions of linearity, normality, equal variance, and independence.