Gurpinder Singh

STAT 108

1/26/2022

## Data: Gift aid at Elmhurst College

In today's lab, we will analyze the `elmhurst` dataset in the openintro package. This dataset contains information about 50 randomly selected students from the 2011 freshmen class at Elmhurst College. The data were originally sampled from a table on all 2011 freshmen at the college that was included in the article "What Students Really Pay to go to College" in *The Chronicle of Higher Education* article.

You can load the data from loading the openintro package, and then running the following command:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
library(broom)
library(modelr)
```

```
##
## Attaching package: 'modelr'
##
## The following object is masked from 'package:broom':
##
##     bootstrap
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```
data(elmhurst)
```

The `elmhurst` dataset contains the following variables:

| | |
|---|---|
| `family_income` | Family income of the student |
| `gift_aid` | Gift aid, in ($ thousands) |
| `price_paid` | Price paid by the student (= tuition - gift_aid) |

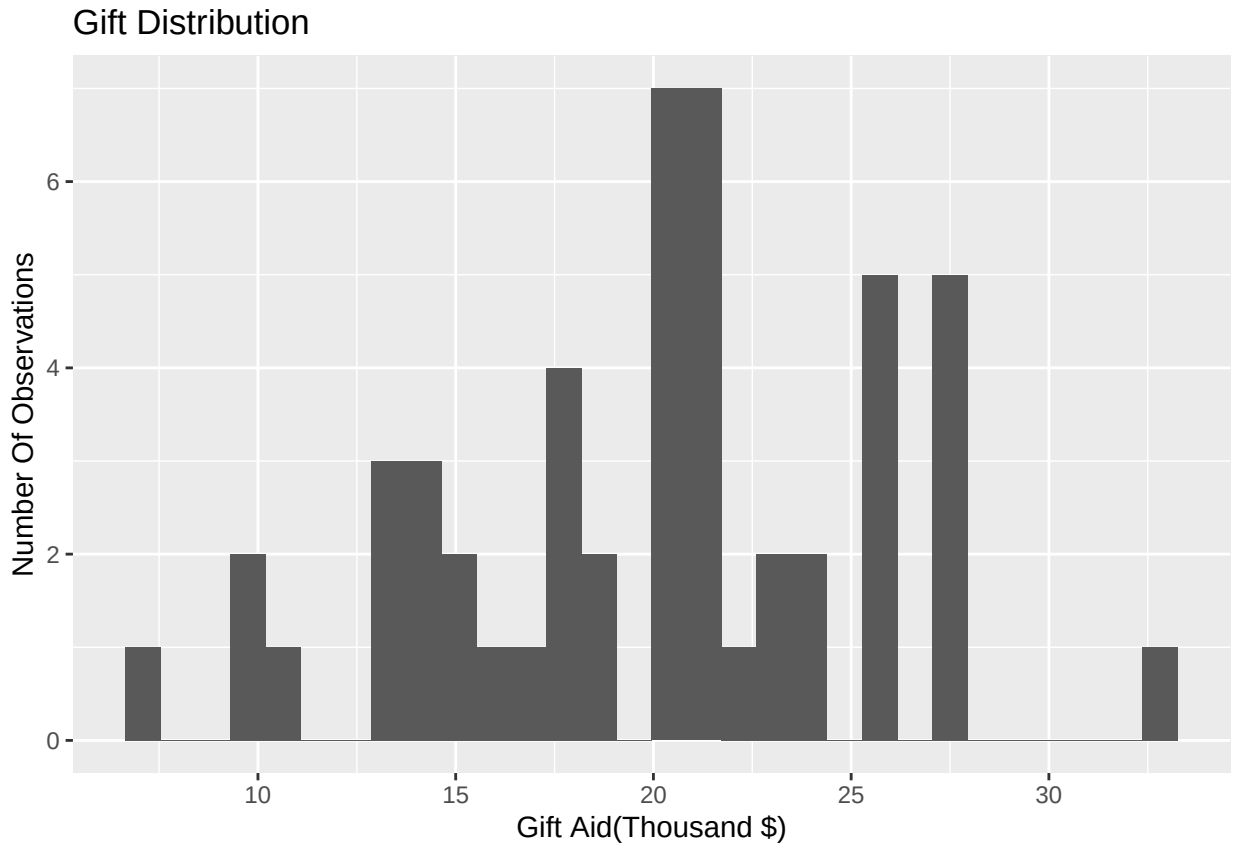
# Exercises

## Exploratory Data Analysis

1. Plot a histogram to examine the distribution of `gift_aid`. What is the approximate shape of the distribution? Also note if there are any outliers in the dataset.

```
ggplot(data = elmhurst, aes(x =gift_aid)) +
  geom_histogram() +
  labs(x ="Gift Aid(Thousand $)",
       y = "Number Of Observations",
       title = "Gift Distribution")
```

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




2. To better understand the distribution of `gift_aid`, we would like calculate measures of center and spread

of the distribution. Use the `summarise` function to calculate the appropriate measures of center (mean or median) and spread (standard deviation or IQR) based on the shape of the distribution from Exercise 1. Show the code and output, and state the measures of center and spread in your narrative. *Be sure to report your conclusions for this exercise and the remainder of the lab in dollars.*

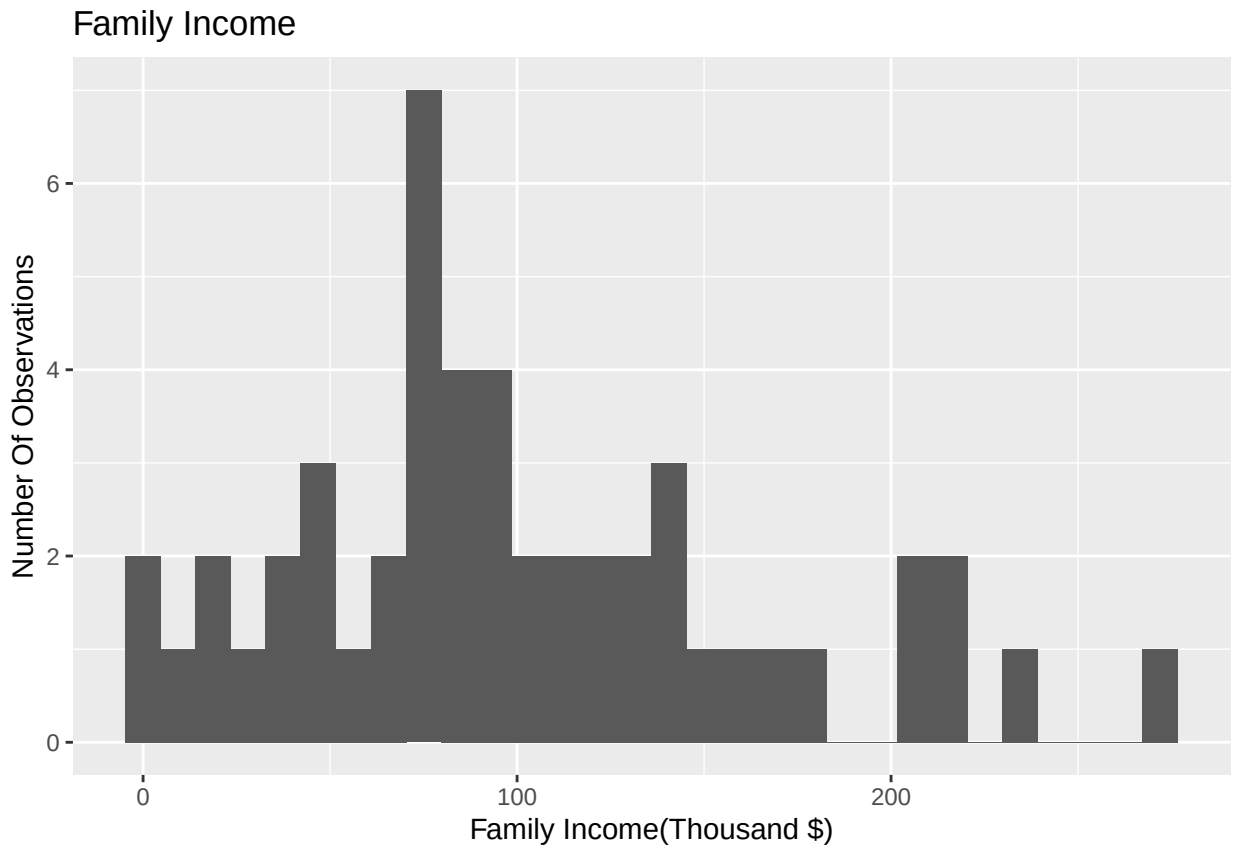
```
elmhurst %>%
  summarise(max = max(gift_aid),
            min = min(gift_aid),
            mean = mean(gift_aid),
            med = median(gift_aid),
            sd = sd(gift_aid),
            iqr = IQR(gift_aid),
  )
```

```
## # A tibble: 1 x 6
##     max   min  mean   med    sd   iqr
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  32.7     7  19.9  20.5  5.46  7.26
```

3. Plot the distribution of `family_income` and calculate the appropriate summary statistics. Describe the distribution of `family_income` (shape, center, and spread, outliers) using the plot and appropriate summary statistics.

```
ggplot(data = elmhurst, aes(x = family_income)) +
  geom_histogram() +
  labs(x = "Family Income(Thousand $)", y = "Number Of Observations", title = "Family Income")
```

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

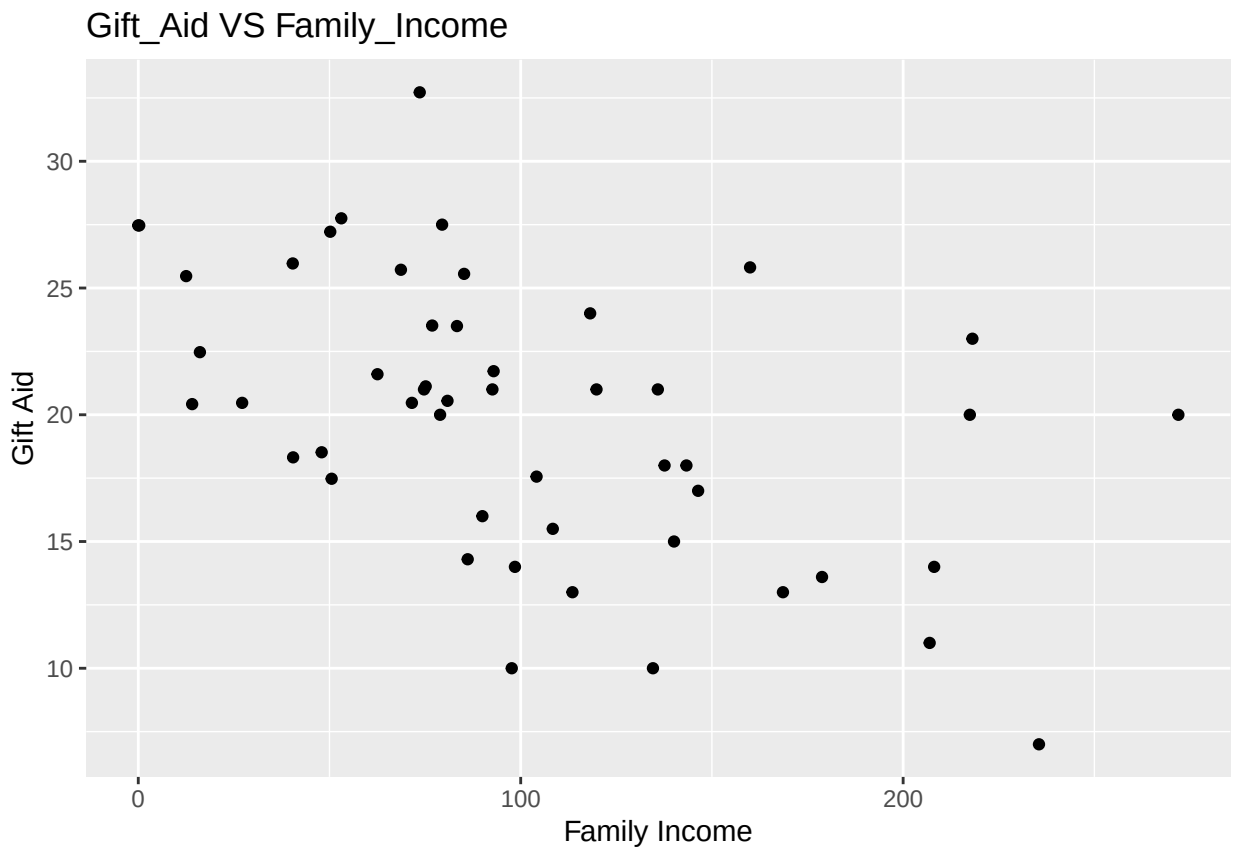



```
elmhurst %>%
  summarise(max = max(family_income),
            min = min(family_income),
            mean = mean(family_income),
            med = median(family_income),
            sd = sd(family_income),
            iqr = IQR(family_income),
  )
```

```
## # A tibble: 1 x 6
##     max   min  mean   med    sd   iqr
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  272.     0  102.  88.1  63.2  73.1
```

4. Create a scatterplot to display the relationship between `gift_aid` (response variable) and `family_income` (predictor variable). Use the scatterplot to describe the relationship between the two variables. Be sure the scatterplot includes informative axis labels and title.

```
ggplot(data = elmhurst, aes(x =family_income, y = gift_aid)) +
  geom_point() +
  labs(x ="Family Income",
       y="Gift Aid",
       title="Gift_Aid VS Family_Income")
```




## Simple Linear Regression

5. Use the `lm` function to fit a simple linear regression model using `family_income` to explain variation in `gift_aid`. Complete the code below to assign your model a name, and use the `tidy` and `kable` functions to neatly display the model output. *Replace X and Y with the appropriate variable names.*

```
reg_model <- lm(gift_aid ~ family_income, data = elmhurst)
  tidy(reg_model) %>% # output model
  kable(digits = 3) # format model output
```

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 24.319 | 1.291 | 18.831 | 0 |
| family_income | -0.043 | 0.011 | -3.985 | 0 |

6. Interpret the slope in the context of the problem. One unit increase in gift_aid results in a decrease of .043 thousand dollars of family income
7. When we fit a linear regression model, we make assumptions about the underlying relationship between the response and predictor variables. In practice, we can check that the assumptions hold by analyzing the residuals. Over the next few questions, we will examine plots of the residuals to determine if the assumptions are met.

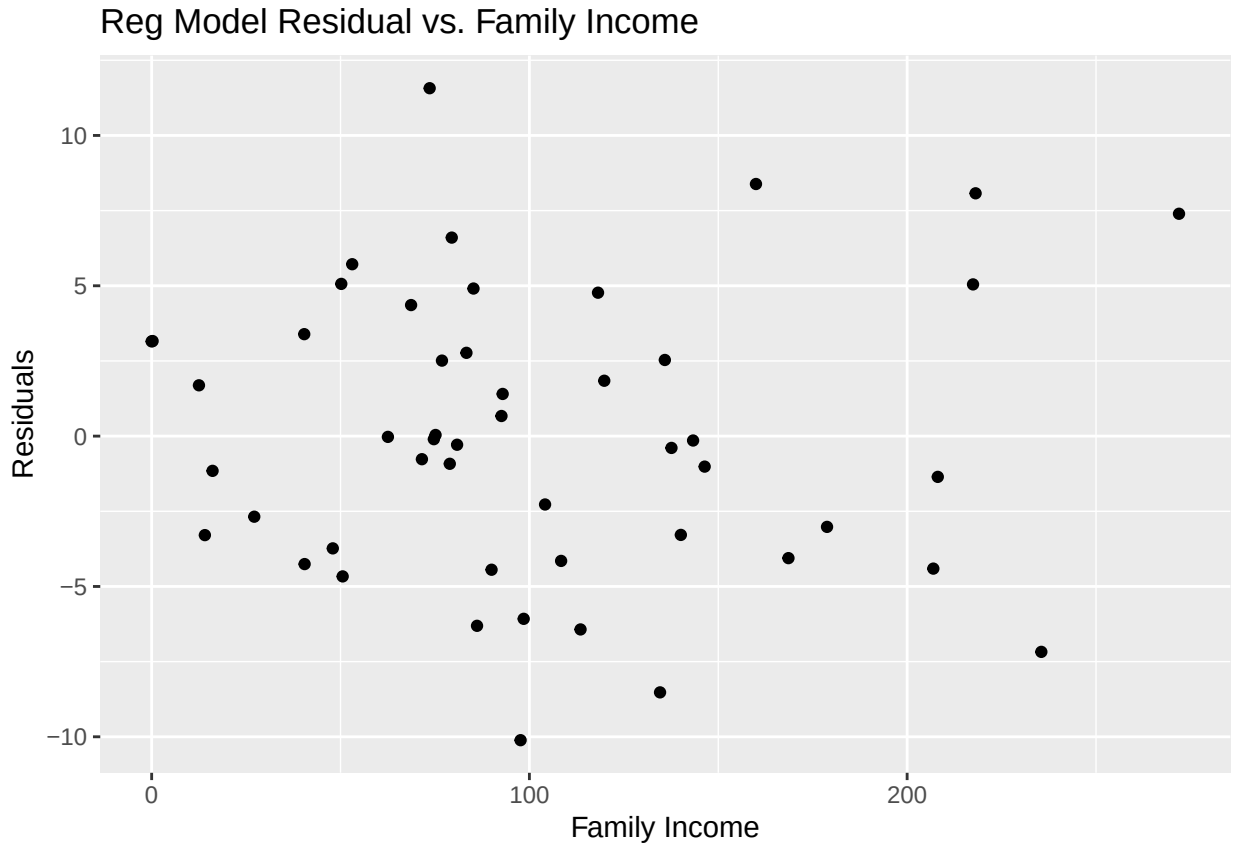
Let's begin by calculating the residuals and adding them to the dataset. Fill in the model name in the code below to add residuals to the original dataset using the `resid()` and `mutate()` functions.

```
elmhurst<- elmhurst %>%
  mutate(resid = residuals(reg_model))
```

8. One of the assumptions for regression is that there is a linear relationship between the predictor and response variables. To check this assumption, we will examine a scatterplot of the residuals versus the predictor variable.

Create a scatterplot with the predictor variable on the $x$ axis and residuals on the $y$ axis. Be sure to include an informative title and properly label the axes.

```
ggplot(data=elmhurst, aes(x =family_income, y= residuals(reg_model))) +
  geom_point() +
  labs(x="Family Income",
       y="Residuals",
       title="Reg Model Residual vs. Family Income")
```




9. Examine the plot from the previous question to assess the linearity condition.

```
- *Ideally, there would be no discernible shape in the plot. This is an indication that the linear model
- *If there is an obvious shape in the plot (e.g. a parabola), this means that the linear model does not
```

Based on this, is the linearity condition is satisfied? Briefly explain your reasoning.

The linearity condition is satisfied because there is no discernible shape in the plot. You can see that all the values are spread out appropriately

10. Recall that when we fit a regression model, we assume for any given value of $x$, the $y$ values follow the Normal distribution with mean $\beta_0 + \beta_1 x$ and variance $\sigma^2$. We will look at two sets of plots to check that this assumption holds.

    We begin by checking the constant variance assumption, i.e that the variance of $y$ is approximately equal for each value of $x$. To check this, we will use the scatterplot of the residuals versus the predictor variable $x$. Ideally, as we move from left to right, the spread of the $y$'s will be approximately equal, i.e. there is no "fan" pattern.

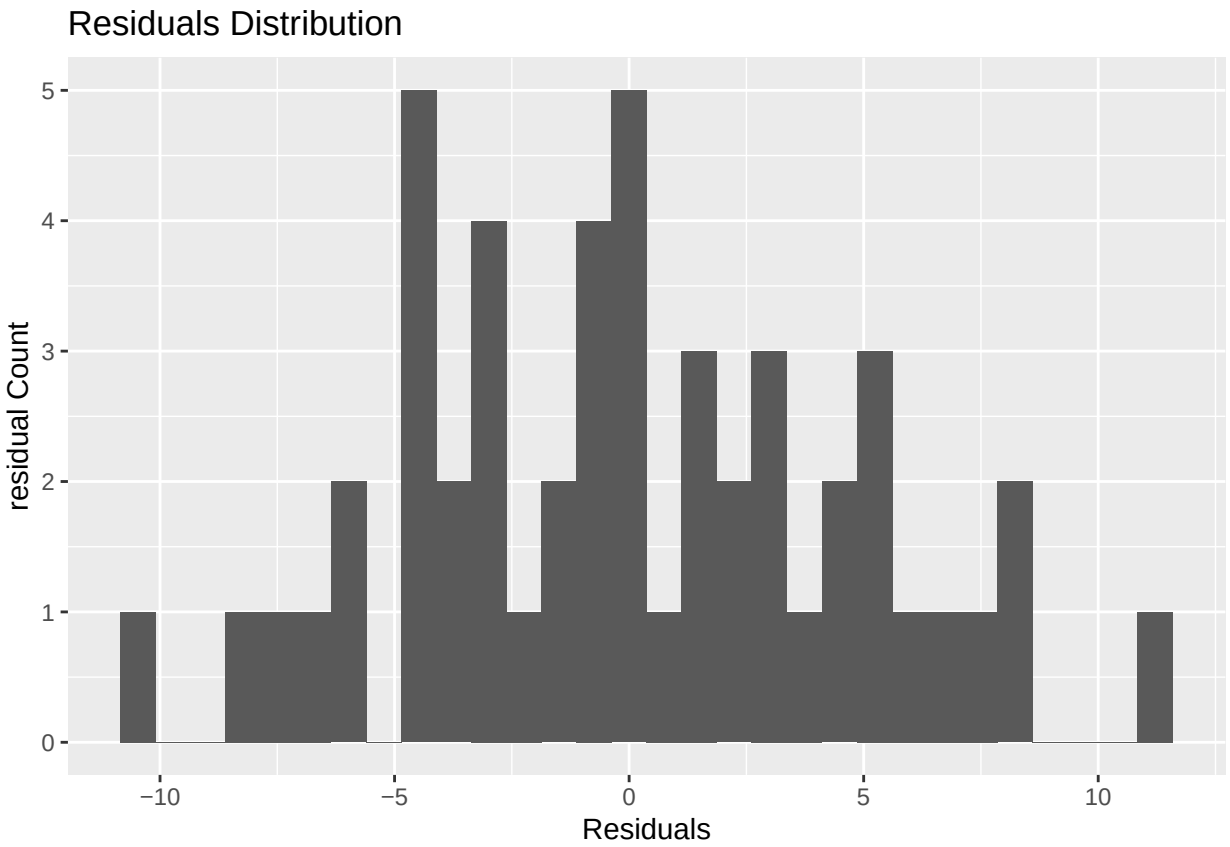
    Using the scatterplot from Exercise 8 , is the constant variance assumption satisfied? Briefly explain your reasoning. *Note: You don't need to know the value of $\sigma^2$ to answer this question.* constant variance assumption is satisfied as the variance of y is apprixmately equal for each value of x. Moving left to right the spread of y will be approximetly equal.

11. Next, we will assess with Normality assumption, i.e. that the distribution of the $y$ values is Normal at every value of $x$. In practice, it is impossible to check the distribution of $y$ at every possible value of $x$, so we can check whether the assumption is satisfied by looking at the overall distribution of the residuals. The assumption is satisfied if the distribution of residuals is approximately Normal, i.e. unimodal and symmetric.

    Make a histogram of the residuals. Based on the histogram, is the Normality assumption satisfied? Briefly explain your reasoning.

```
ggplot(data = elmhurst, aes(x = residuals(reg_model))) +
  geom_histogram() +
  labs(x = "Residuals",
       y = "residual Count",
       title = "Residuals Distribution" )
```

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




12. The final assumption is that the observations are independent, i.e. one observation does not affect another. We can typically make an assessment about this assumption using a description of the data. Do you think the independence assumption is satisfied? Briefly explain your reasoning. That is true one observation has no effect on the other. One families income should have no effect on anothers. ## Using the Model

13. Calculate $R^2$ for this model and interpret it in the context of the data.

```
rSQ <-summary(reg_model)$r.squared
```

14. Suppose a high school senior is considering Elmhurst College, and she would like to use your regression model to estimate how much gift aid she can expect to receive. Her family income is $90,000. Based on your model, about how much gift aid should she expect to receive? Show the code or calculations you use to get the prediction.

15. Another high school senior is considering Elmhurst College, and her family income is about $310,000. Do you think it would be wise to use your model calculate the predicted gift aid for this student? Briefly explain your reasoning. No because 310,000 is not included inside of the data set and would be a major anomly inside of the dataset. Making assumptions outside of the data would result in unpredictable results so the college should not. *You're done and ready to submit your work! Knit, commit, and push all remaining changes. You can use the commit message "Done with Lab 2!", and make sure you have pushed all the files to GitHub (your Git pane in RStudio should be empty) and that all documents are updated in your repo on GitHub. Then submit the assignment on Gradescope following the instructions below.*