STAT 108: Lab 3

Timothy Lanthier

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.

```
library(tidyverse)
library(knitr)
library(broom)
library(modelr)
library(openintro)
```

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

The `elmhurst` dataset contains the following variables:

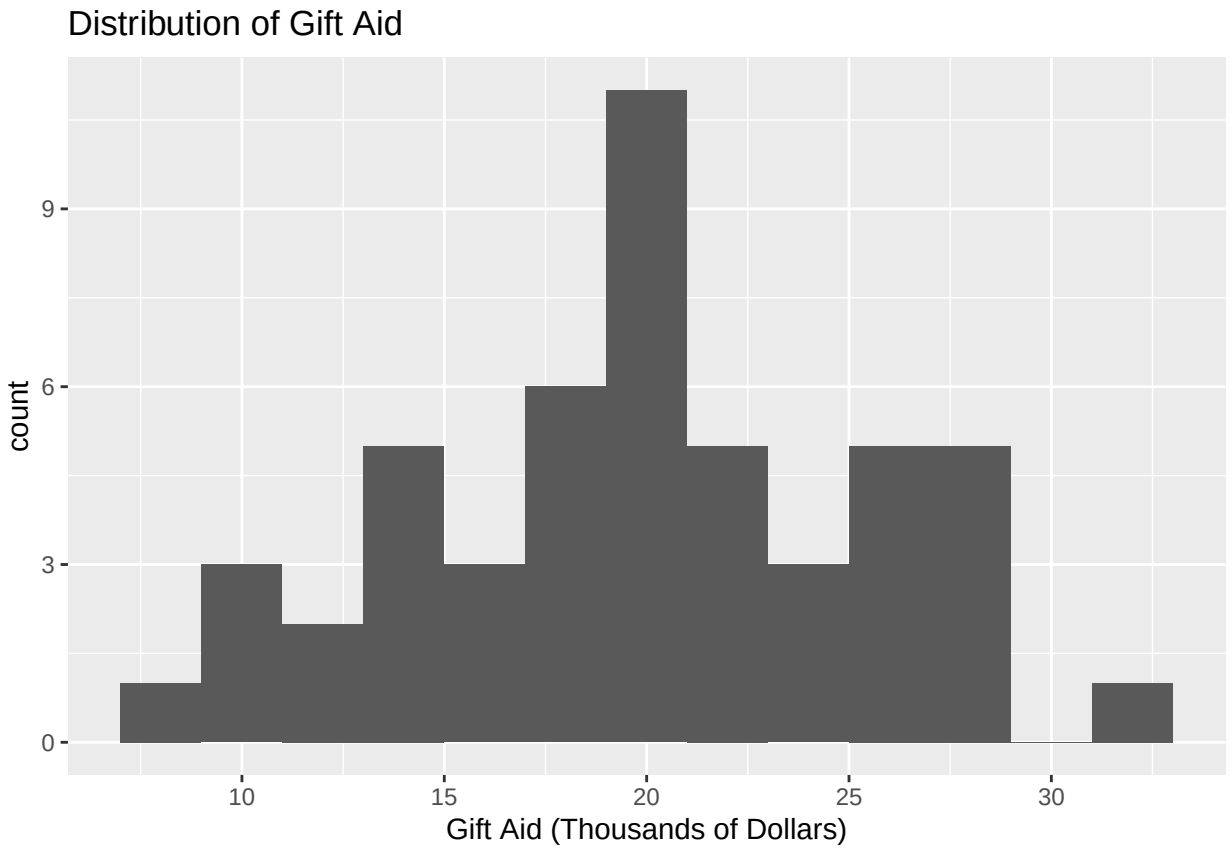
`family_income`	Family income of the student
`gift_aid`	Gift aid, in ($ thousands)
`price_paid`	Price paid by the student (= tuition - gift_aid)

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(gift_aid)) +
  geom_histogram(binwidth = 2) +
  labs(x = 'Gift Aid (Thousands of Dollars)', title = 'Distribution of Gift Aid')
```



The distribution appears to have a bell curve like shape. 'gift_aid' appears to be centered around about 20 and our distribution is fairly symmetric. We could make the argument that it is approximately normal. It looks like we have very few outliers as well. It looks like we might have an outlier at around 5 as well as one closer to 35.

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.*

```
summarise(elmhurst, mean = mean(gift_aid),
          std_dev = sd(gift_aid),
          min = min(gift_aid),
          q1 = quantile(gift_aid, 0.25),
          median = median(gift_aid),
          q3 = quantile(gift_aid, 0.75),
          max = max(gift_aid),
          IQR = q3-q1)
```

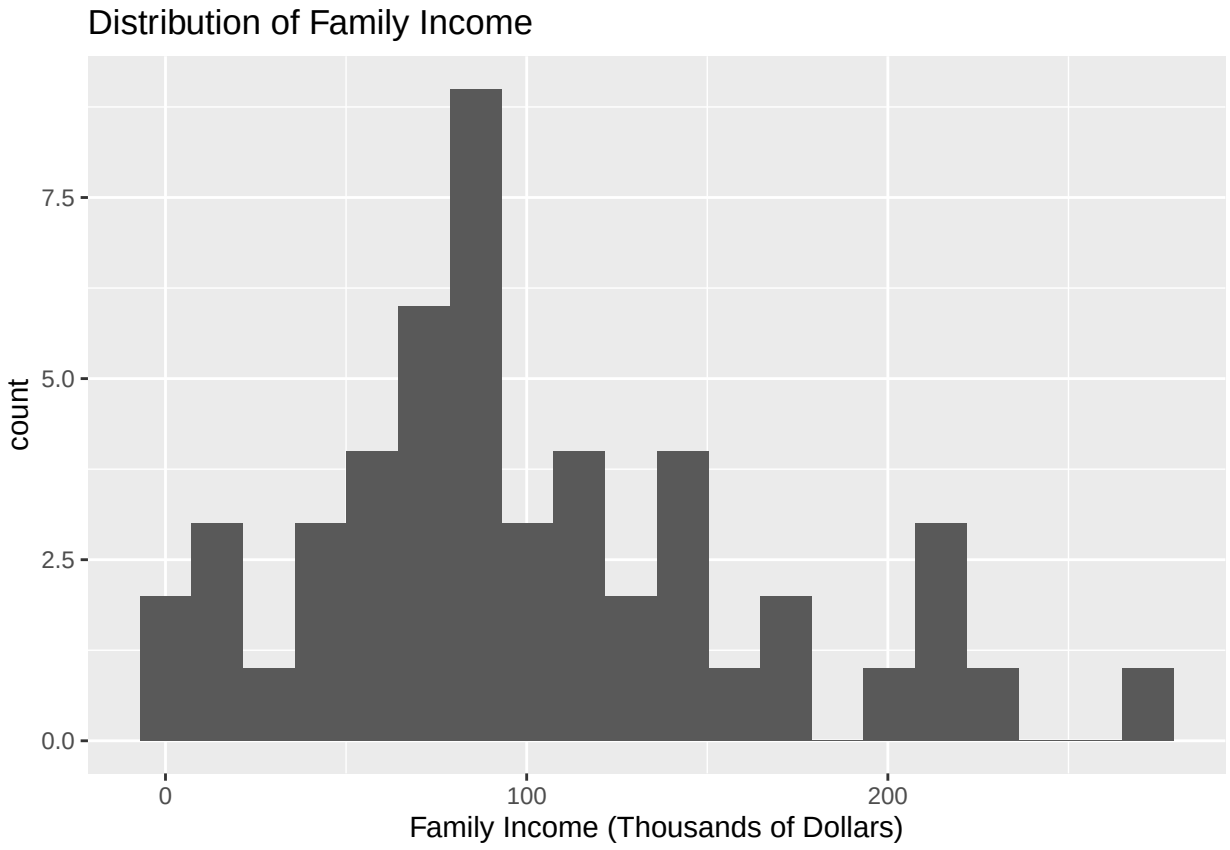
```
## # A tibble: 1 x 8
##    mean std_dev   min    q1 median    q3   max   IQR
##   <dbl>   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1  19.9    5.46     7  16.2   20.5  23.5  32.7  7.26
```

Looking at our summary statistics, we have a mean of 19.935 and a similar mean of about 20.47. So on average students appear to have received around $19,935 in aid and the typical student received around $20,470 in aid. Seeing as our mean is smaller than our median we appear to have a slight negative skew

although it seems quite insignificant. We also have a standard deviation of about $5,460 and an IQR of $7,265. Interpreting the IQR, this means that the spread in aid between the middle 50% of students is around $7,265.

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(family_income)) +
  geom_histogram(bins = 20) +
  labs(x = 'Family Income (Thousands of Dollars)',
       title = 'Distribution of Family Income')
```




```
summarise(elmhurst, mean = mean(family_income),
          std_dev = sd(family_income),
          min = min(family_income),
          q1 = quantile(family_income, 0.25),
          median = median(family_income),
          q3 = quantile(family_income, 0.75),
          max = max(family_income),
          IQR = q3-q1)
```

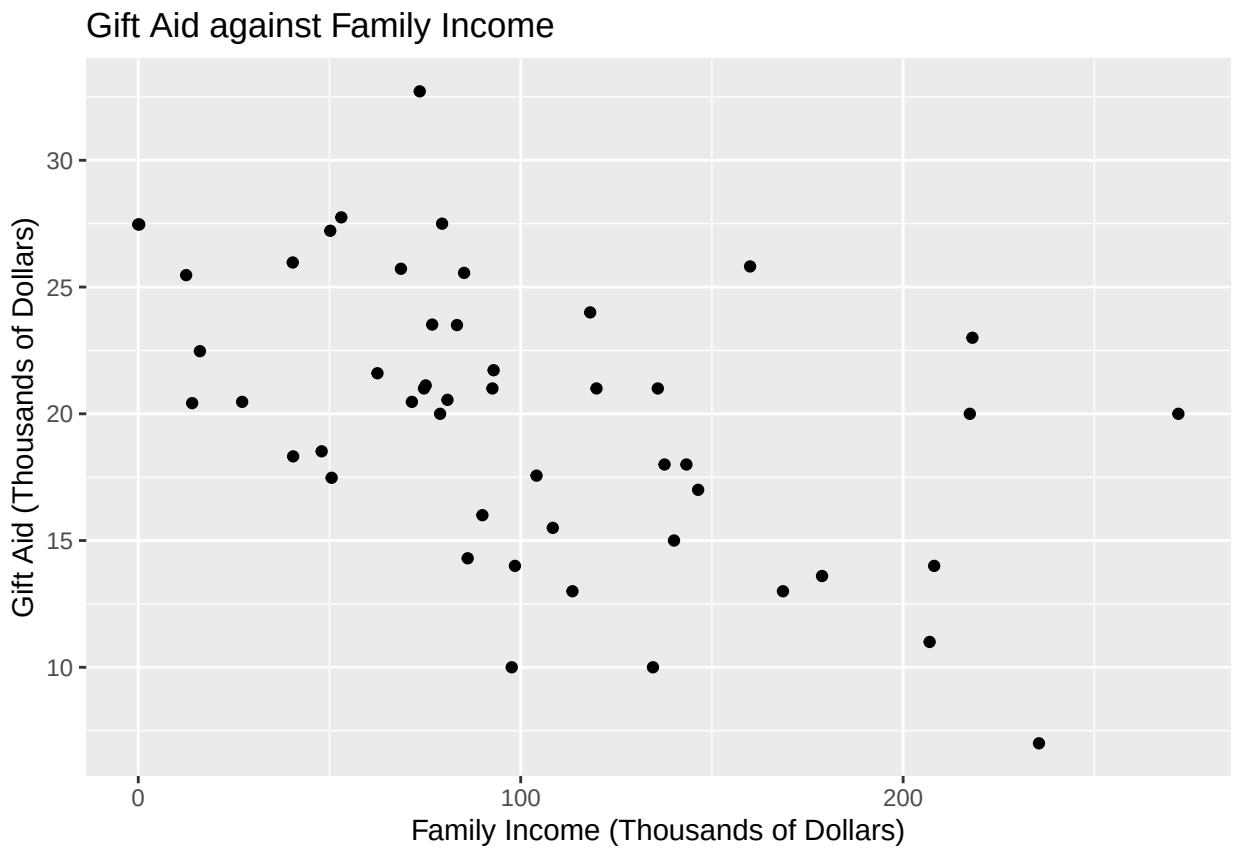
```
## # A tibble: 1 x 8
##    mean std_dev   min    q1 median    q3   max   IQR
##   <dbl>   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1  102.    63.2     0  64.1   88.1  137.  272.  73.1
```

Looking at the histogram, the distribution of family income has a somewhat heavy positive skew. This is backed up by our summary statistics as the mean family income is about $101,778 while the median is much

lower at approximately \$88,061. We also have quite a few outliers. The spread of family income is also quite large. There is an IQR of \$73,095 indicating that there is a \$73,000 income difference from the 25th and 75th percentile. We also have a standard deviation of about \$63,000 which is also very lage. The highest family income is \$271,974, which is about $\frac{271.972-101.7785}{62.306} = 2.732$ standard deviations from the mean. Looking at our histogram, it looks like we have some families with incomes around the \$240,000 range which are also likely an outliers.

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(elmhurst, aes(y = gift_aid, x= family_income)) +
  geom_point() +
  labs(x = 'Family Income (Thousands of Dollars)', y = 'Gift Aid (Thousands of Dollars)',
       title = 'Gift Aid against Family Income')
```



Looking at the scatterplot, it looks like there is a negative linear association between Gift Aid and Family Income. It appears that students with high family incomes tend to receive less aid. Meanwhile those with very small family incomes are receiving more aid. That being said, we can identify a few points that don't follow this trend. For example, there is one student who received about \$20,000 in aid but has a very high family income of around \$320,000.

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.*

```
income_model <- lm(gift_aid ~ family_income, data = elmhurst)
tidy(income_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.

Our model gives us a slope of -0.043. Since our variables both are in thousands of dollars, this means that according to our model, with an increase in family income of $1,000 we would expect that the gift aid received would be reduced by $43.

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.

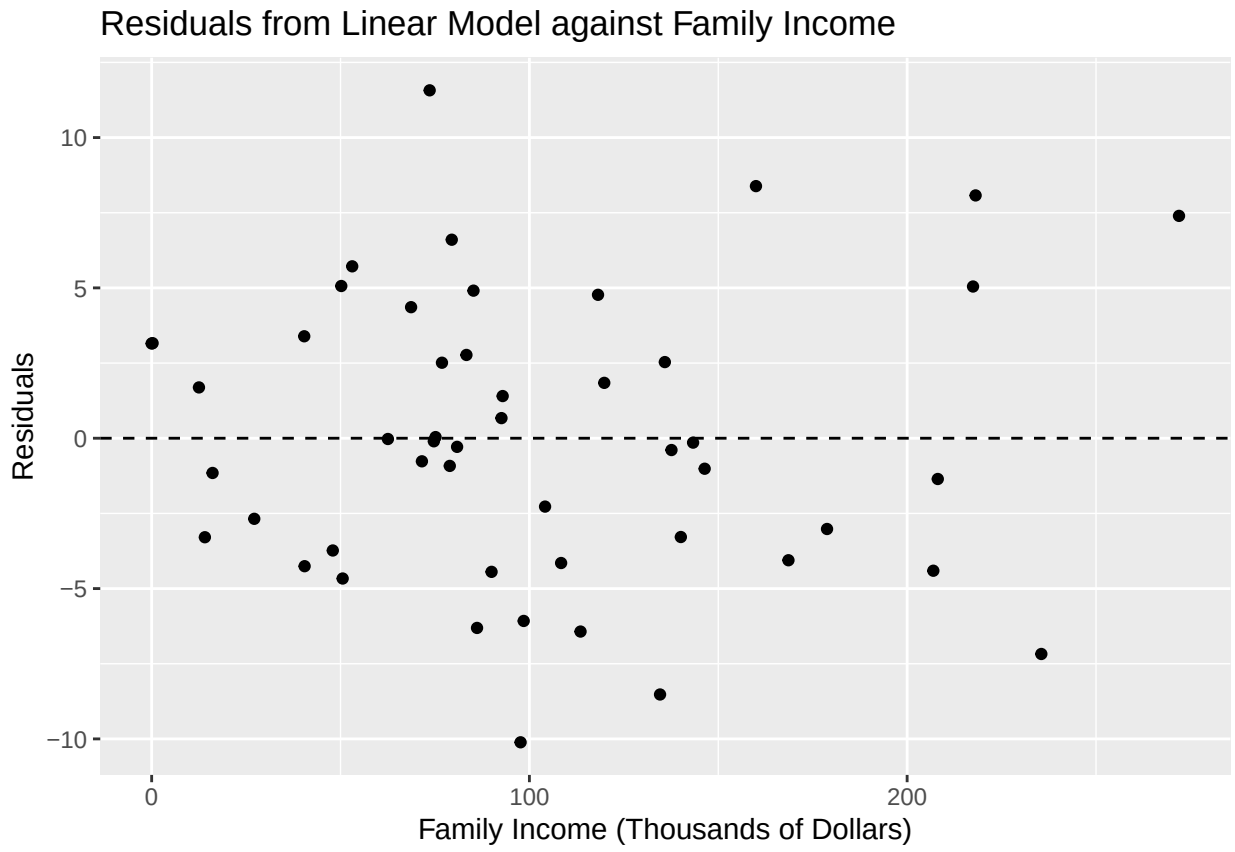
Let's begin by calculating the residuals and adding them to the dataset. Fill in the model name in the code below to add residuals to the original dataset using the `resid()` and `mutate()` functions.

```
elmhurst_resid <- elmhurst %>%
  mutate(resid = residuals(income_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(elmhurst_resid, aes(x = family_income, y = resid)) +
  geom_point() +
  labs(x = 'Family Income (Thousands of Dollars)', y = 'Residuals',
       title = 'Residuals from Linear Model against Family Income') +
  geom_hline(yintercept = 0, linetype = 'dashed')
```



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 adequately describes the relationship between the response and predictor, and all that is left is the random error that can't be accounted for in the model, i.e. other things that affect gift aid besides family income.*
   - *If there is an obvious shape in the plot (e.g. a parabola), this means that the linear model does not adequately describe the relationship between the response and predictor variables.*

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

There is not discernible shape in the plot of the residuals. Since the residuals appear to be randomly scattered around 0, then we would say that the linearity condition is satisfied. If there was some observable shape in the plot of the residuals, then it means that there is some relationship between gift aid and family income that is not described well by our linear model.

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 σ^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 σ^2 to answer this question.*

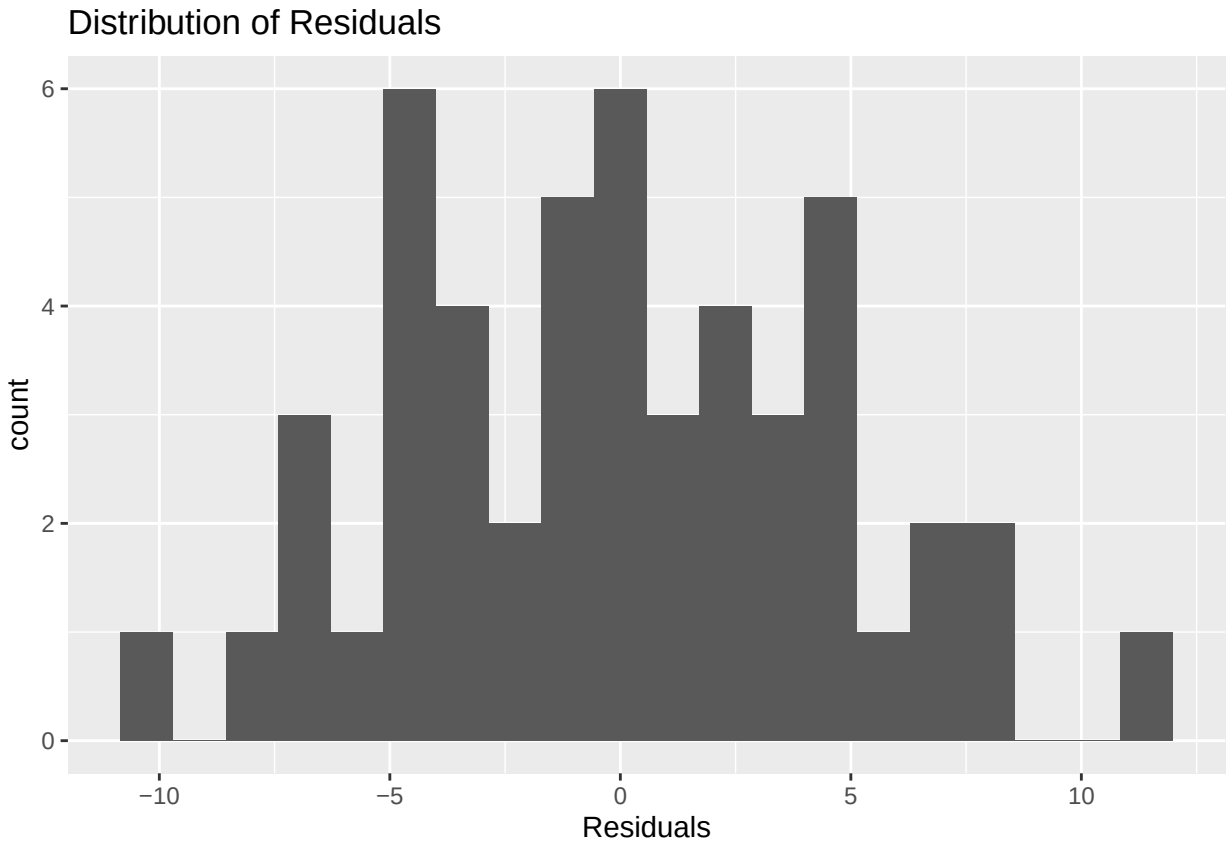
We would argue that the constant variance assumption is satisfied. Looking at the plot from exercise 8, for all levels of family income, the residuals are spread a similar amount around 0. That being said we don't

have perfectly constant variance. There are a few points around the $75,000-100,000 family income mark where the residuals are over $10,000 away from 0. That being said, these are only a few observations so we would still say the constant variance assumption is satisfied. Also, for some of the observations with family income over $200,000 we have residuals fairly far from 0. That being said, we would argue that this may be due to having such few observations for family income over $200,000. So we would still conclude that the constant variance assumption is satisfied.

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_resid, mapping = aes(resid)) +
  geom_histogram(bins = 20) +
  labs(x = 'Residuals', title = 'Distribution of Residuals')
```



It doesn't look like the normality condition is satisfied. While there is a little bit of symmetry to the distribution, the histogram is bimodal with 2 large spikes at around -5 and 0.

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.

Yes, the independence assumption is satisfied. Since the sample is a random sample from the 2011 class, then we can claim that observations are independent from one another.

Using the Model

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

```
rsquare(income_model, elmhurst)
```

```
## [1] 0.2485582
```

For our model we have the single predictor variable 'family_income' to help predict 'gift_aid'. We found an R^2 of 0.249. That means that our model using just family income explains 24.86% of the variation in gift aid.

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.

```
x0 <- data.frame(family_income = c(90))
predict.lm(income_model, x0,interval = 'prediction', conf.level = 0.95)
```

```
##        fit      lwr    upr
## 1 20.44288 10.72776 30.158
```

Based on our model, we would expect that this student would receive $20,443 of gift aid.

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. It would be unwise to use this model to predict gift aid for this student. Going back to our summary statistics for family income, in our dataset, the largest family income observed is around $270,000. Thus our model would only be appropriate for predicting gift aid for students whose family incomes lie at or below $270,000. From the dataset provided, we have no way of knowing that the linear relationship measured can be appropriately extended for families with higher incomes. Seeing as $310,000 is far above this range, it would be inappropriate to use this model to predict gift aid for this student.

Github: https://github.com/talanthier/lab-03

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.