Lab 3

STAT 108

1/26/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- 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: Gift aid at Elmhurst College

GITHUB LINK: https://github.com/theeho/lab03

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:

```
data(elmhurst)
view(elmhurst)
```

The `elmhurst` dataset contains the following variables:

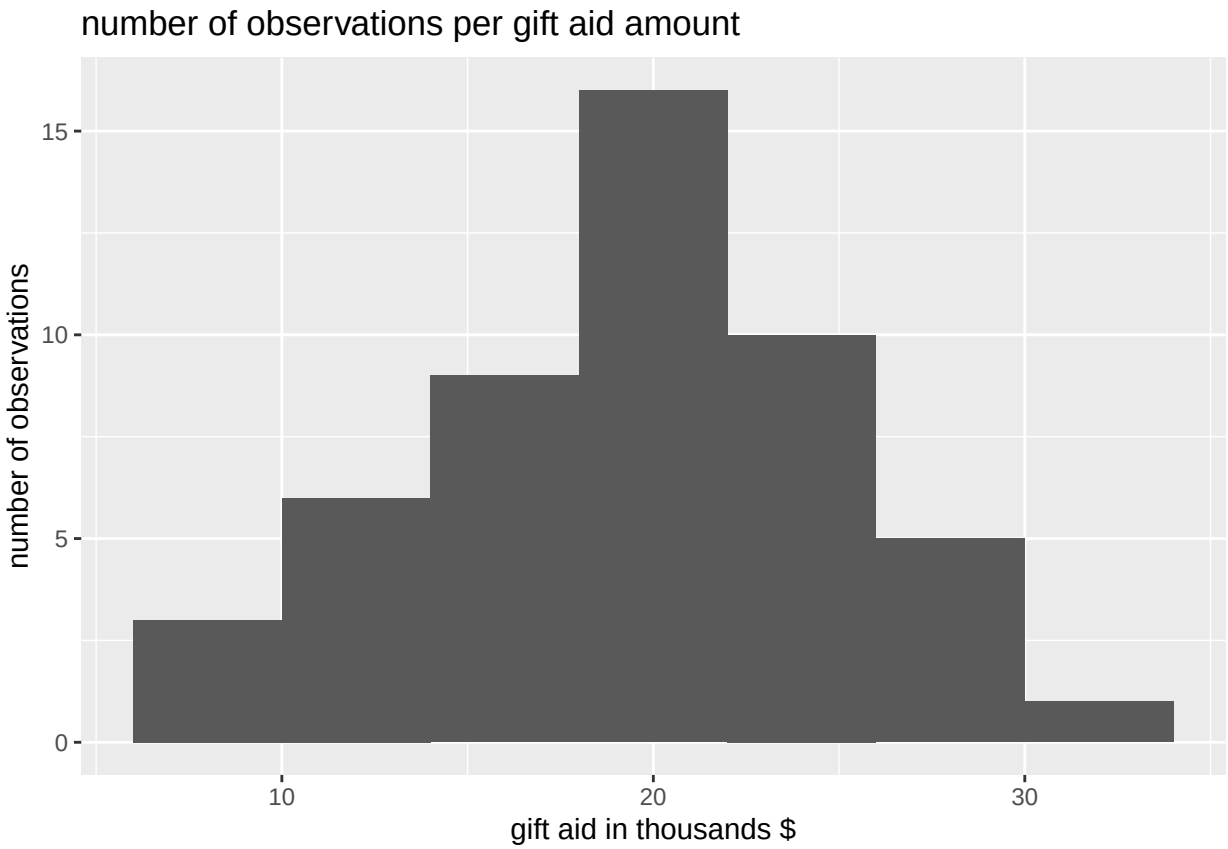
`family_income`	Family income of the student
`gift_aid`	Gift aid, in ($ thousands)
`price_paid`	Price paid by the student (= tuition - gift_aid)

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) +
  geom_histogram(mapping = aes(x = gift_aid), binwidth = 4) +
  labs(x = "gift aid in thousands $",
       y = "number of observations",
       title = "number of observations per gift aid amount")
```



The distribution of gift aid appears to follow the shape of a bell curve. Although it does not appear to be symmetrical. There are a few outliers above 30k USD in gift aid.

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(min = min(gift_aid),
            q1 = quantile(gift_aid, .25),
            median = median(gift_aid),
            q3 = quantile(gift_aid, .75),
            max = max(gift_aid),
            iqr = IQR(gift_aid),
            mean = mean(gift_aid),
            std_dev = sd(gift_aid)
            )
```

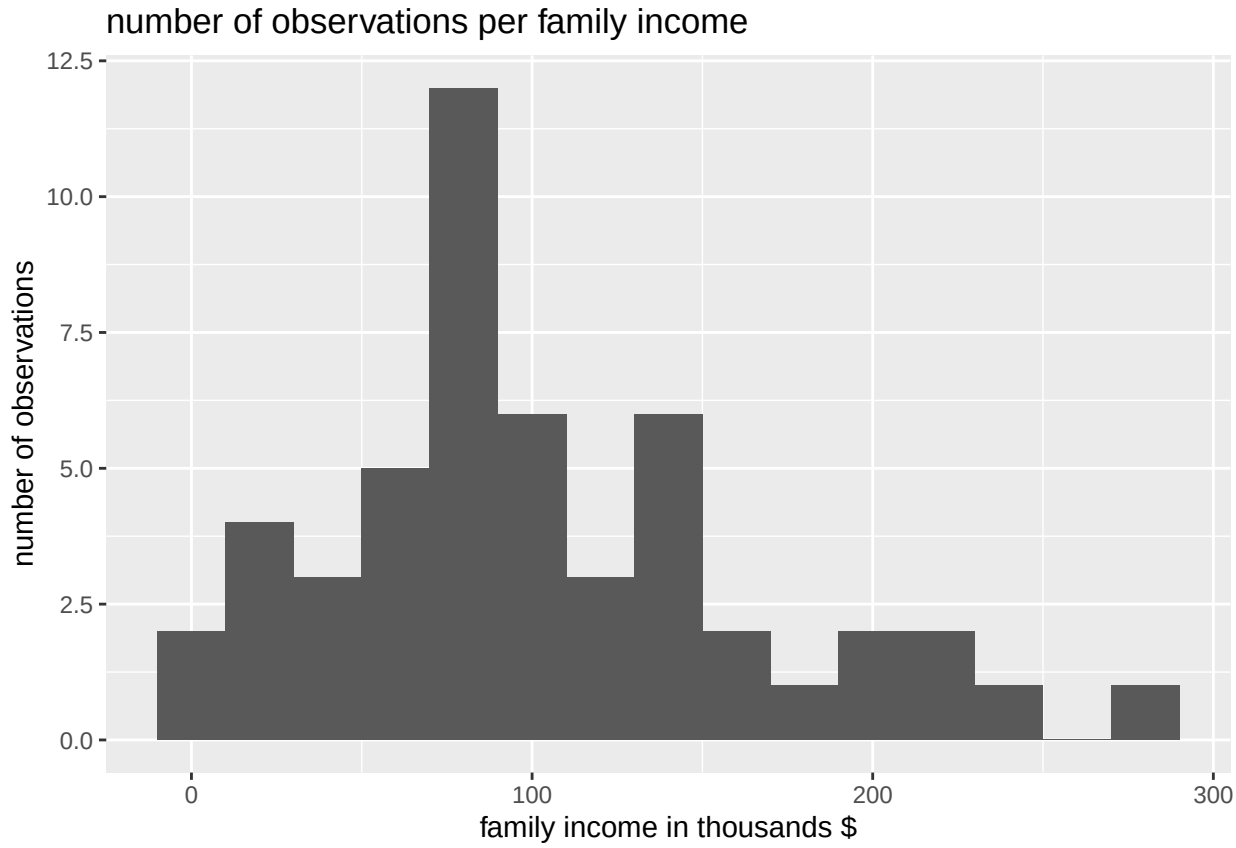
```
## # A tibble: 1 x 8
##     min    q1 median    q3   max   iqr  mean std_dev
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1     7  16.2   20.5  23.5  32.7  7.26  19.9    5.46
```

The miniumum gift aid is 7,000 USD while the maximum is 32,720 USD. These where the percise values of the 2 outliers we observed in the histogram. The middle 50% of the gift aid is between 16,250 USD and

23.515 USD. The mean is roughly centered in the data as it is very close to the median. Based on this summary, I would say the distribution is reasonably balanced and not too far spread with an IQR of 7,625 USD and a STD of 5,460 USD.

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) +
  geom_histogram(mapping = aes(x = family_income), binwidth = 20) +
  labs(x = "family income in thousands $",
       y = "number of observations",
       title = "number of observations per family income")
```



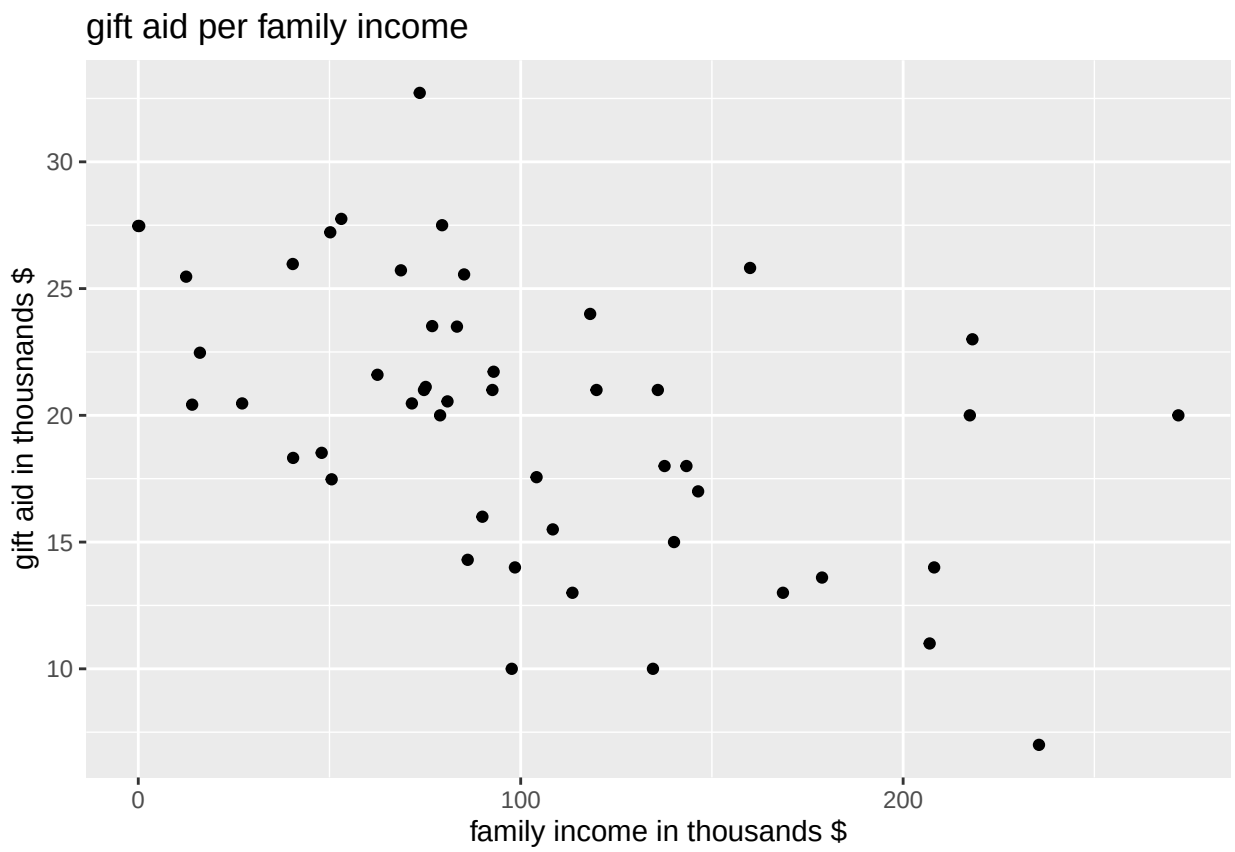

```
elmhurst %>%
  summarise(min = min(family_income),
            q1 = quantile(family_income, .25),
            median = median(family_income),
            q3 = quantile(family_income, .75),
            max = max(family_income),
            iqr = IQR(family_income),
            mean = mean(family_income),
            std_dev = sd(family_income)
            )
```

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

The majority of the incomes based on the plot seem to be around 75k USD and 100k USD. This is reflected with an 88k USD median and how 50% of the incomes are between 65k USD and 137k USD. However, there is a fair bit of income outliers in the higher range. The max is 271k USD and the 75% quantile is at 137k USD. Based on this, I would say that the data is pretty spread out with a very high STD and IQR of 63k USD and 73K USD repspectively.

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(x=family_income, y=gift_aid)) + geom_point()+
  labs(x = "family income in thousands $",
       y = "gift aid in thousnands $",
       title = "gift aid per family income ")
```



Based on this scatter plot, there appears to be some relationship between gift aid and family income. Although there are a few strange outliers such as the highest family income recieving about 20k USD in gift aid. The relationship seems to be that an increase in family income is associated with a decrease in gift aid. Althougfh this relationship seems to be less apparent as the family income increases past 150k USD.

Simple Linear Regression

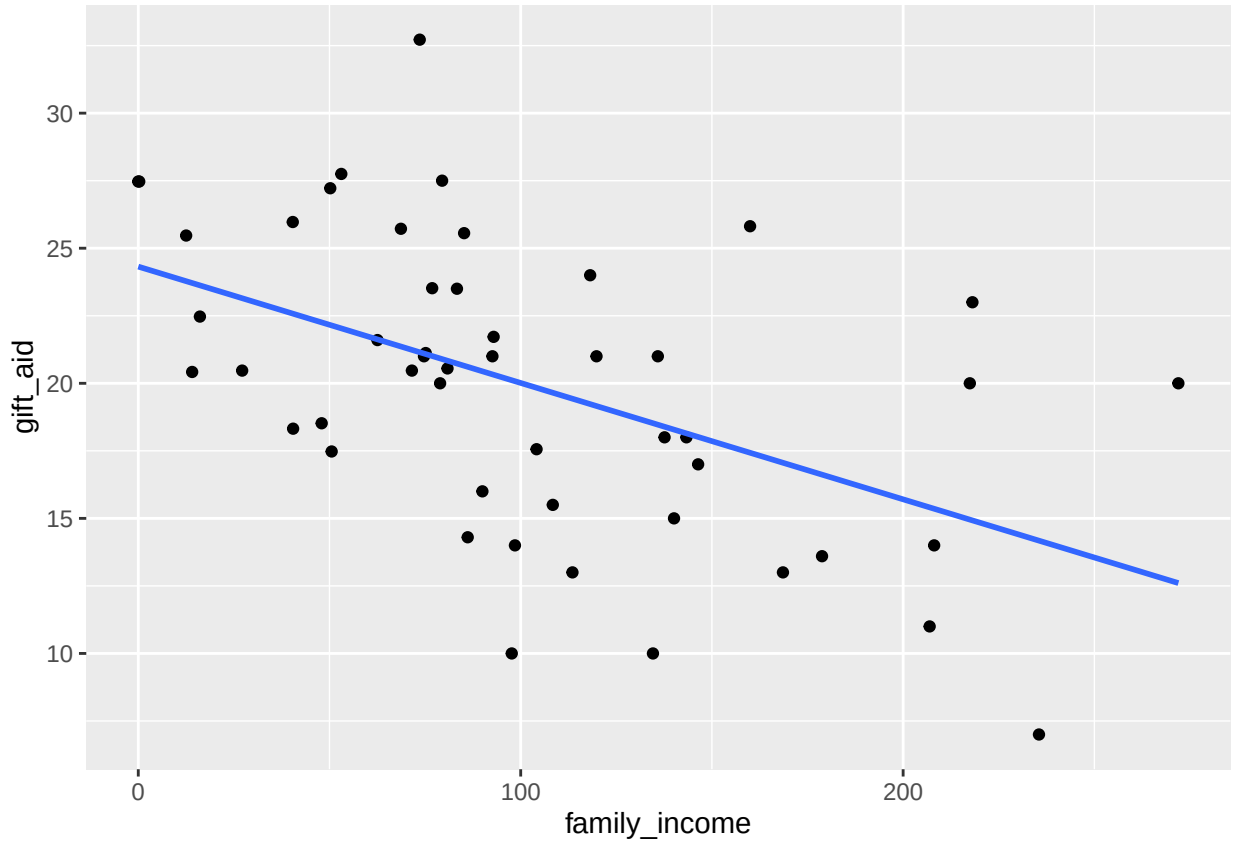
5. Use the `lm` function to fit a simple linear regression model using `family_income` to explain variation in `gift_aid`. Complete the code below to assign your model a name, and use the `tidy` and `kable` functions to neatly display the model output. *Replace X and Y with the appropriate variable names.*

```
linear_Model <- lm(gift_aid ~ family_income, data = elmhurst)
tidy(linear_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

```
ggplot(data = elmhurst, aes(x = family_income, y = gift_aid)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



6. Interpret the slope in the context of the problem.

Every 1 USD increase in family income is ascociated with a .043 USD decrease in gifted aid.

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.

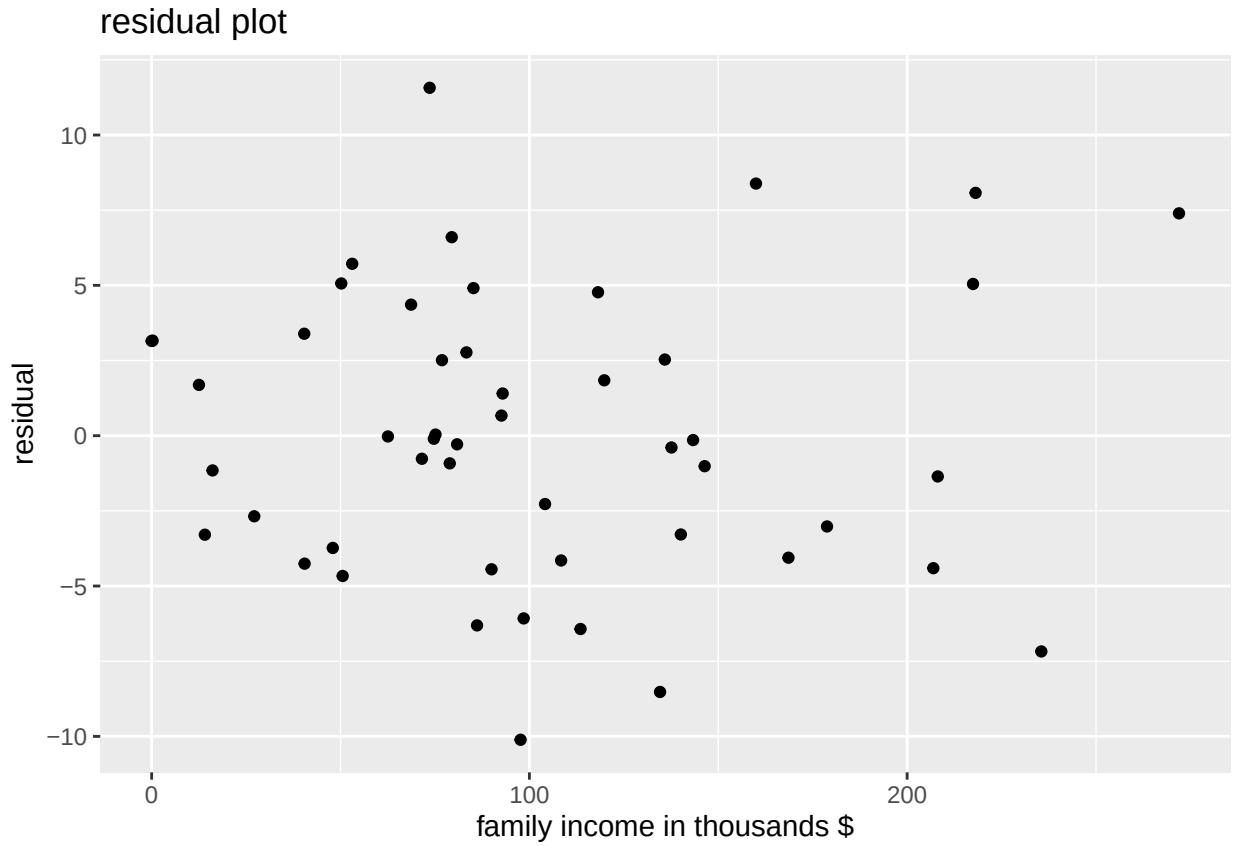
Let's begin by calculating the residuals and adding them to the dataset. Fill in the model name in the code below to add residuals to the original dataset using the `resid()` and `mutate()` functions.

```
residual_model<- elmhurst %>%
  mutate(resid = residuals(linear_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 = residual_model, aes(x = family_income, y = resid)) +
  geom_point() +
  labs(x = "family income in thousands $",
       y = "residual",
       title = "residual plot")
```



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.

It appears that lower values of family income have much less residual variance than higher values of family income. However, there does not appear to be a discernible shape of the residuals. It appears to be random.

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.*

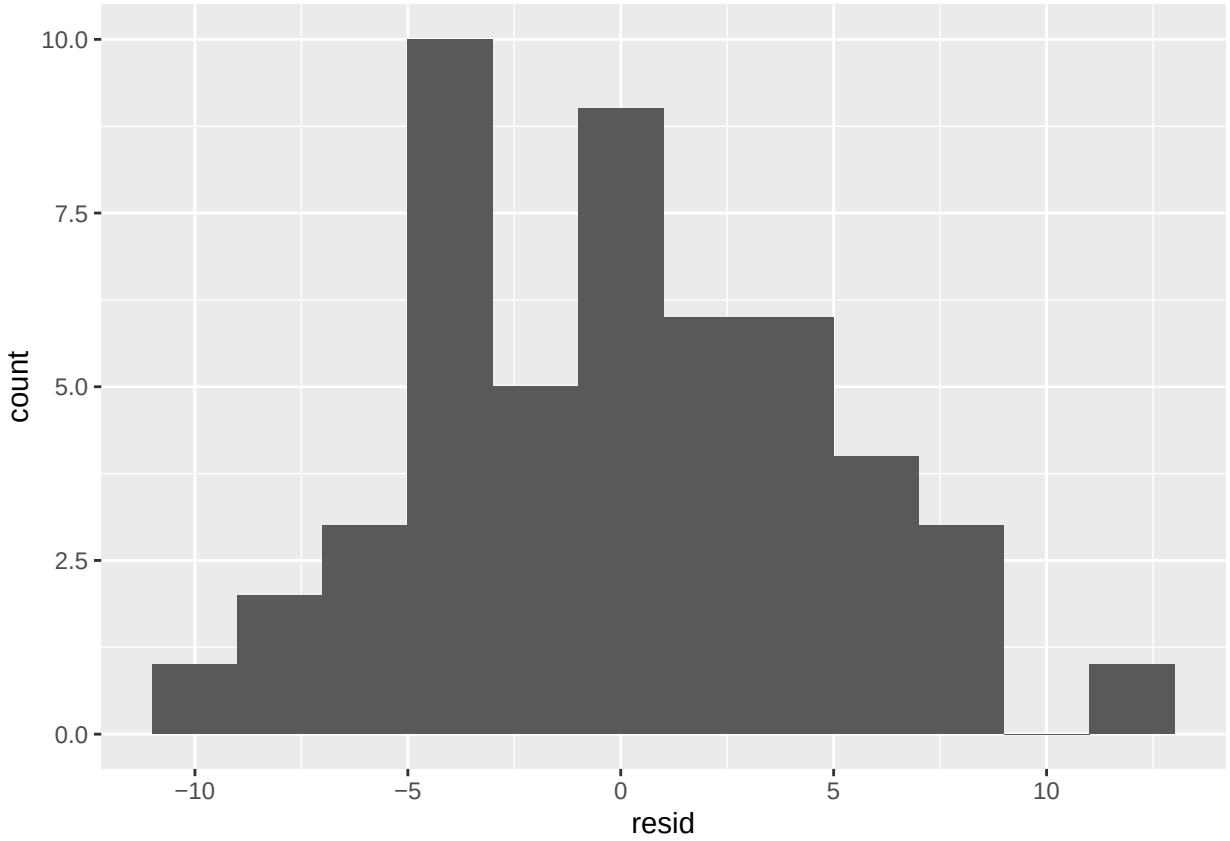
There does not appear to be a constant variance for the residuals. Residual values of family income less than 100k USD seem to have a constant variance. However, larger income families seem to have a much higher error. Thus the constant variance assumption does not seem to be satisfied.

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.

    Based on the plot below, we do see some resemblence of a normal distribution. However, it does not appear to be completely symetric. If this was a normal distribution we would expect a constant increase in count as the residual approached 0. However, There seems to be a dip after -5. There is also an outlier at around residual = 10.

```
ggplot(data = residual_model) +
  geom_histogram(mapping = aes(x = resid), binwidth = 2)
```



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.

Based on the description of the dataset "This dataset contains information about 50 randomly selected students from the 2011 freshmen class at Elmhurst College." It can be assumed that this is a random sample.

Using the Model

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

```
summary(linear_Model)
```

```
##
## Call:
## lm(formula = gift_aid ~ family_income, data = elmhurst)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1128  -3.6234  -0.2161   3.1587  11.5707
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   24.31933    1.29145  18.831  < 2e-16 ***
## family_income -0.04307    0.01081  -3.985 0.000229 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.783 on 48 degrees of freedom
## Multiple R-squared:  0.2486, Adjusted R-squared:  0.2329
## F-statistic: 15.88 on 1 and 48 DF,  p-value: 0.0002289
```

The interpretation of the R-Squared is that about 25% of the variance in financial aid is ascociated with family income.

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.

```
predict(linear_Model, newdata = data.frame(family_income = 90))
```

```
##        1
## 20.44288
```

Based on the model, she should expect 20,449 USD in 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.

Our model does not contain any data greater than 272k USD. Therefor our model contains no evidence to support a prediction of a family income of 310,000 USD. Because of this, it would not be wise to use this model to calculate the predicted aid for this student.

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.