Lab 3

Aisha Lakshman

2/3/2022

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

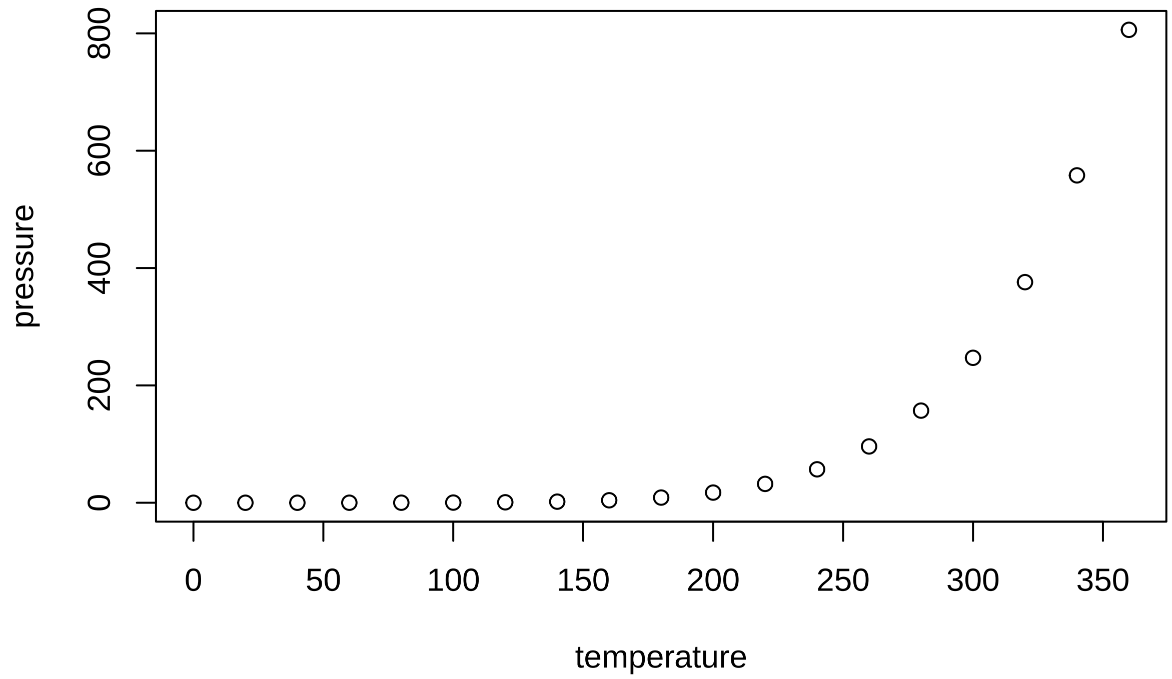
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
library(openintro)
```

```
## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata
```

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

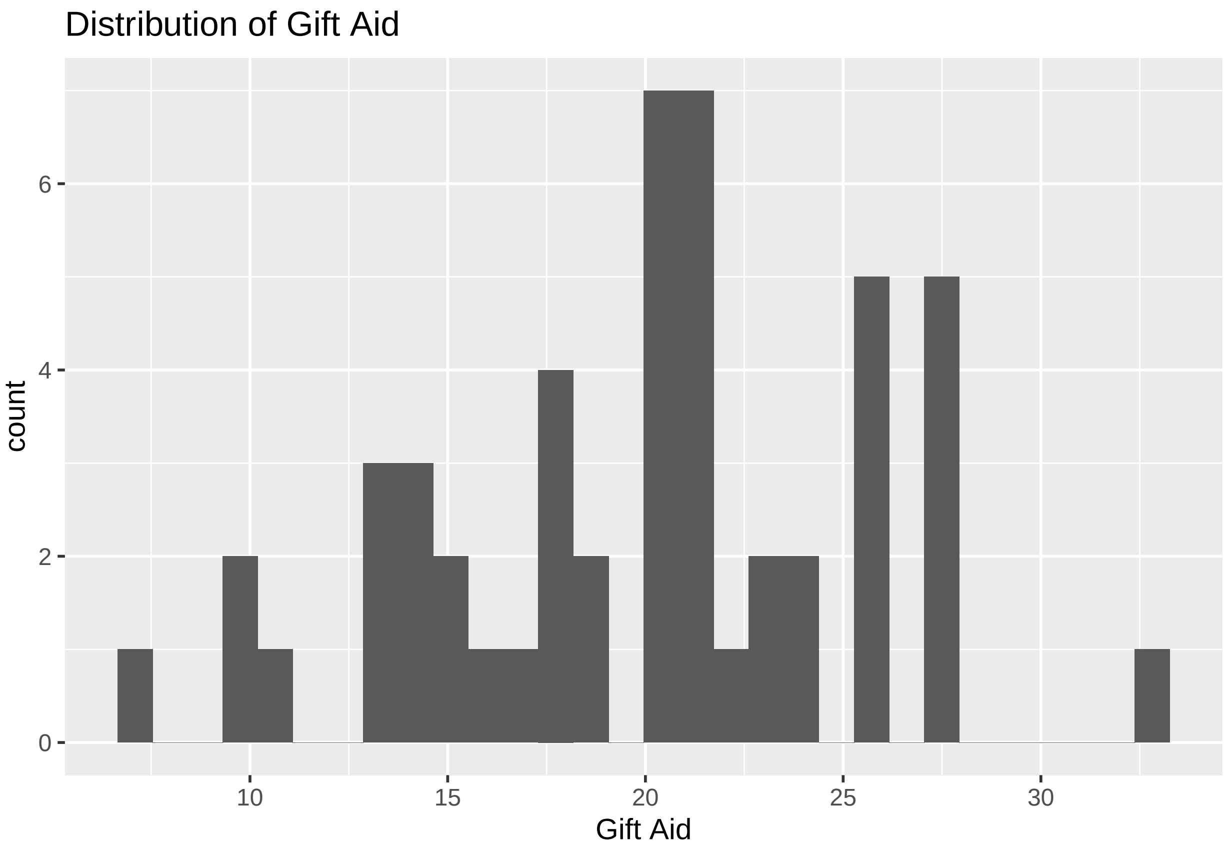
```
##
## Attaching package: 'modelr'

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

```
library(openintro)
data(elmhurst)
```

Exercise 1

```
ggplot(data = elmhurst, aes(x = gift_aid)) +
  geom_histogram() +
  labs(x = "Gift Aid", title = "Distribution of Gift Aid")
```

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



The distribution is bell-shaped, approximately following normal distribution. It looks like there are some outliers in the dataset towards the right side of the distribution, where gift aid is around $33,000.

Exercise 2

```
elmhurst %>%
  summarise(n = n(), mean = mean(gift_aid),
            sd = sd(gift_aid))
```

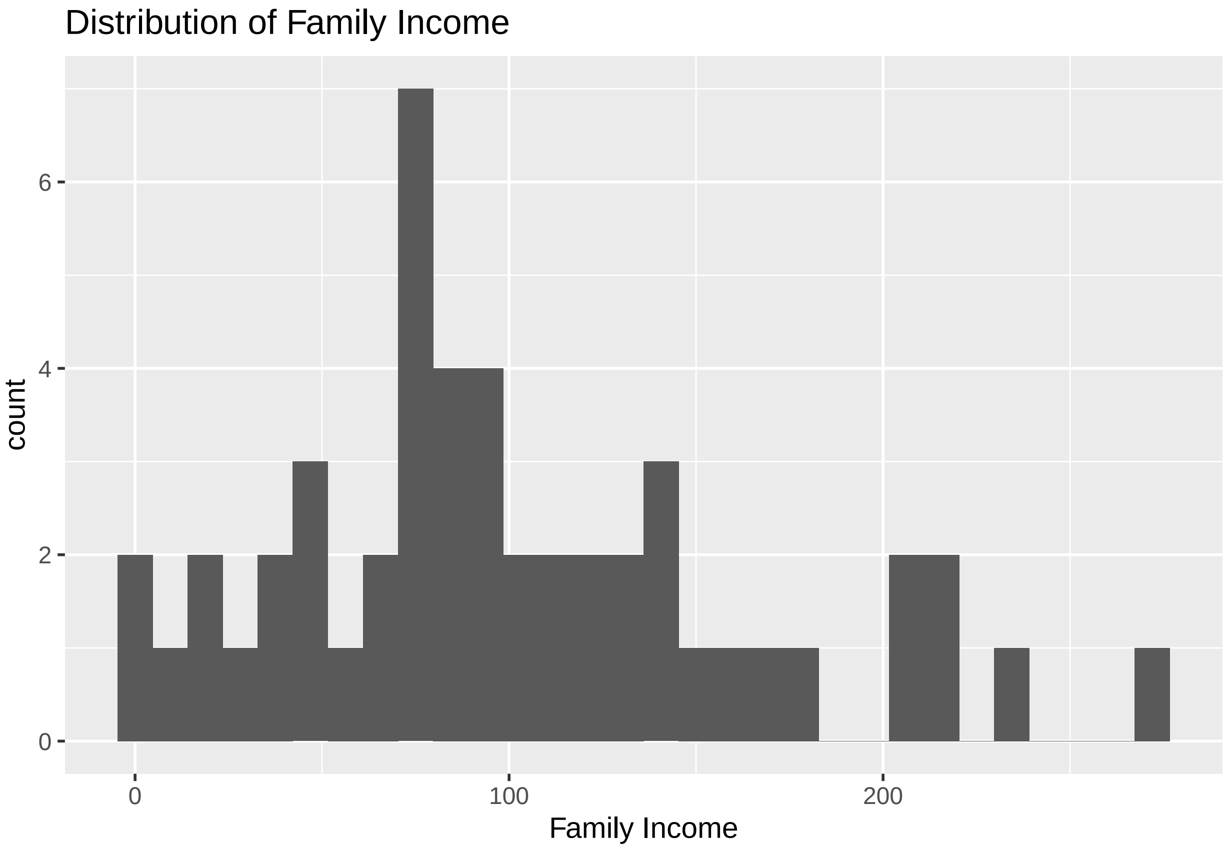
```
## # A tibble: 1 x 3
##       n  mean    sd
##   <int> <dbl> <dbl>
## 1    50  19.9  5.46
```

The mean of gift_aid is approximately 20 thousand dollars, and we can see from the distribution that the center of the spread is under gift_aid = 20 (thousand dollars). The standard deviation of gift aid is approximately 5.4 thousand dollars, which is a relatively high standard deviation under a normal distribution. This tells us that the dataset values are far from the mean, hence the outliers.

Exercise 3

```
ggplot(data = elmhurst, aes(x = family_income)) +
  geom_histogram() +
  labs(x = "Family Income", title = "Distribution of Family Income")
```

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



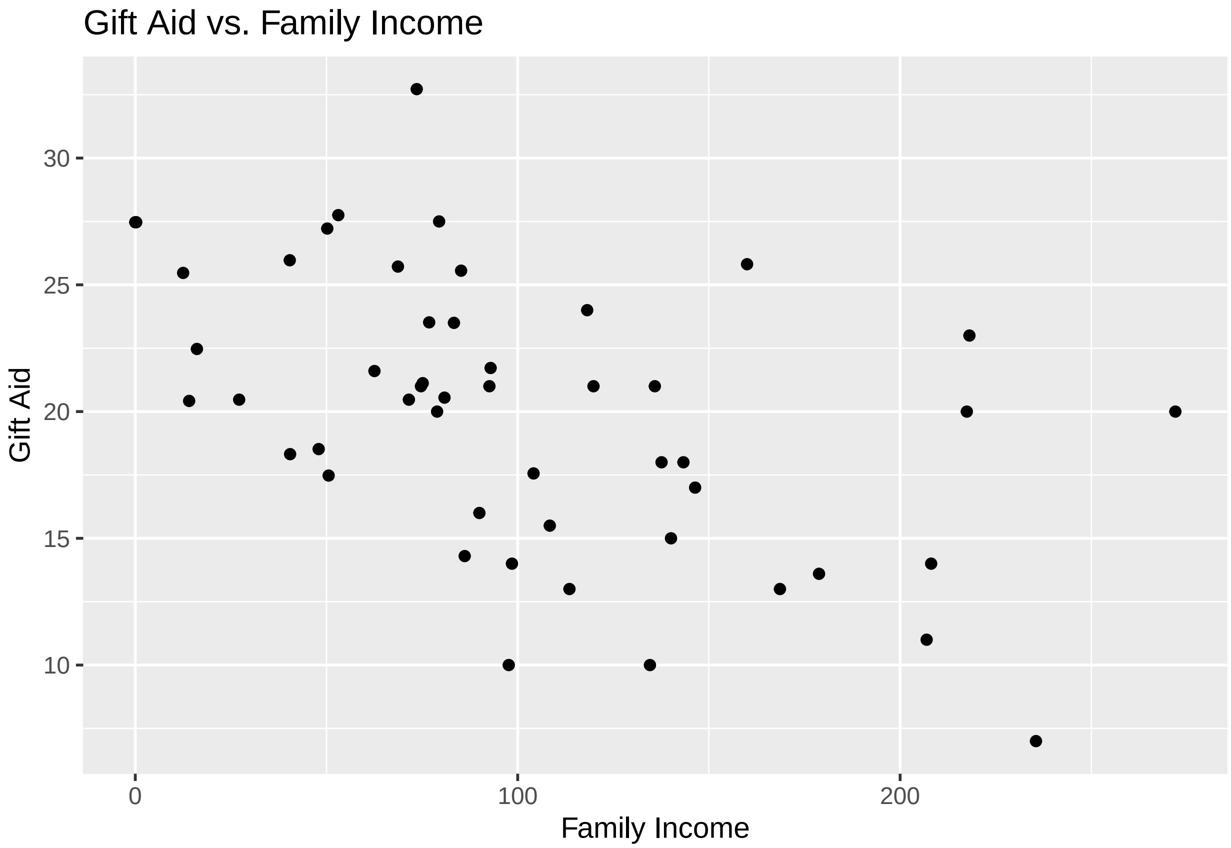
```
elmhurst %>%
  summarise(n = n(), mean = mean(family_income),
            sd = sd(family_income))
```

```
## # A tibble: 1 x 3
##       n  mean    sd
##   <int> <dbl> <dbl>
## 1    50  102.  63.2
```

The distribution is bell-shaped, approximately following a normal distribution. However, there are outliers to the right of the distribution, where family income is approximately greater than or equal to 200,000. This follows the elmhurst dataset because few entries for family income are 200,000 dollars or greater. The summary statistics for family income further explain the distribution. The mean is approximately 102,000 thousand dollars, and considering the outliers, the mean value is raised by the family incomes that are 200,000 or greater. The standard deviation is approximately 63 thousand dollars, reaffirming the outliers of family income data.

Exercise 4

```
ggplot(data = elmhurst, mapping = aes(x = family_income, y = gift_aid)) +
  geom_point() +
  labs(x = "Family Income", y = "Gift Aid", title = "Gift Aid vs. Family Income")
```



Exercise 5

```
model <- lm(gift_aid ~ family_income, data = elmhurst)
tidy(model, conf.int=TRUE) %>% kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	24.319	1.291	18.831	0	21.723	26.916
family_income	-0.043	0.011	-3.985	0	-0.065	-0.021

Exercise 6

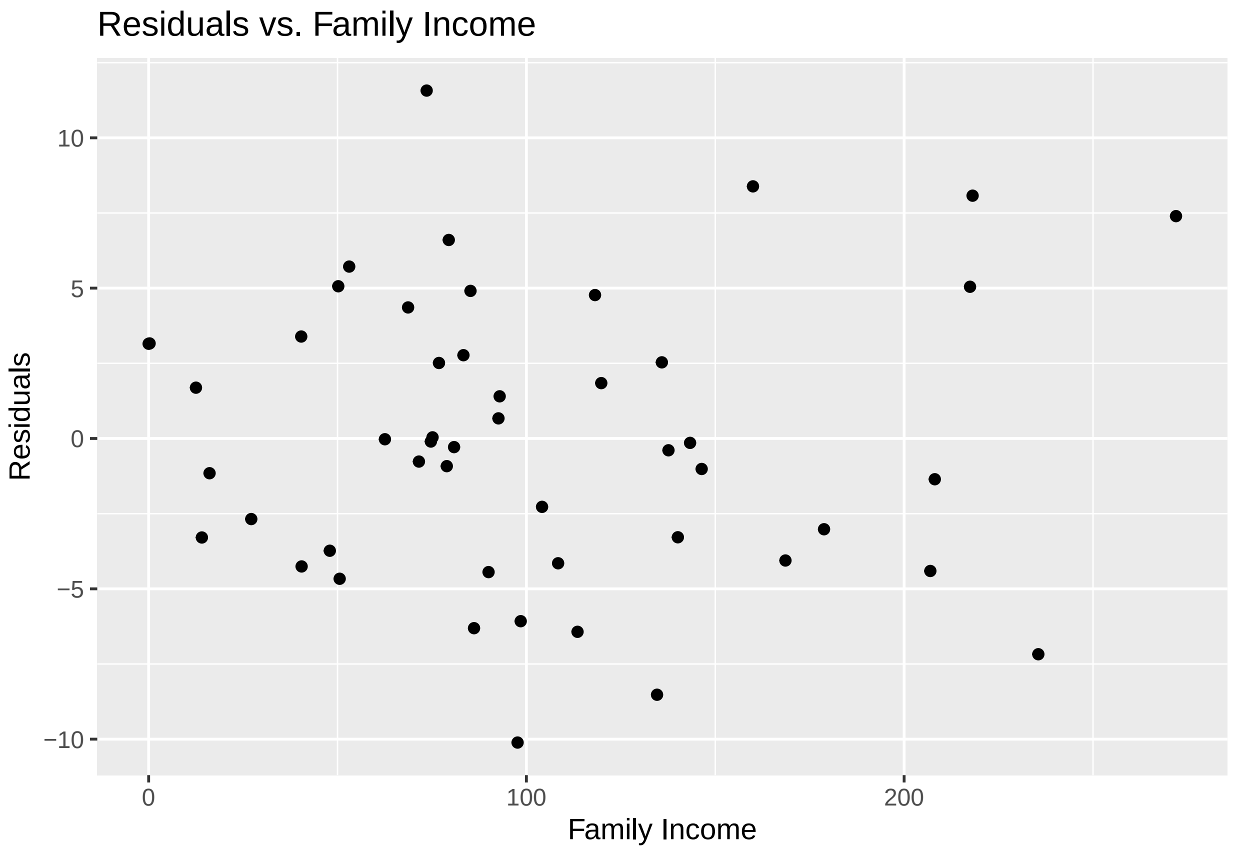
In interpreting the slope: for each additional percentage point in family income, gift aid is expected to decrease by 0.043 percentage points on average.

Exercise 7

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

Exercise 8

```
ggplot(data = elmhurst, mapping = aes(x = family_income, y = residuals(model))) +
  geom_point() +
  labs(x = "Family Income", y = "Residuals", title = "Residuals vs. Family Income")
```



Exercise 9

Based on looking at the Residuals vs. Family Income scatterplot, the linearity condition is satisfied. We can see that the points are evenly distributed, tending to cluster in the center of the plot. Additionally, the points reveal no clear pattern or shape, which means that our linear model adequately describes the relationship between gift aid and family income. Further, the absence of a discernible shape indicates that all that is left are random errors that can't be accounted for in the linear model.

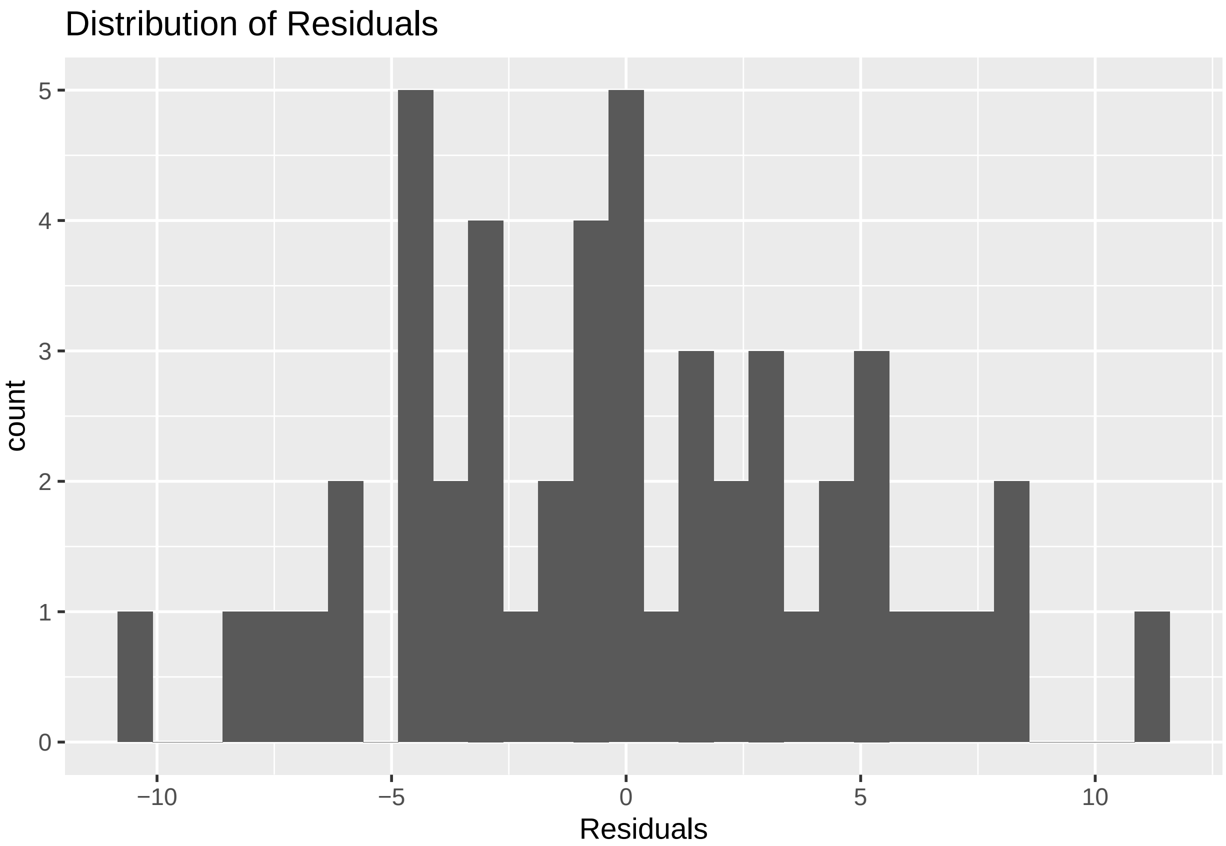
Exercise 10

By looking at the scatterplot from exercise 8, we can see an even distribution of the "y's": there is no "fan" pattern. The equal distribution of points in the scatterplot indicates that the constant variance assumption is satisfied.

Exercise 11

```
ggplot(data = elmhurst, mapping = aes(x = residuals(model))) +
  geom_histogram() +
  labs(x = "Residuals", title = "Distribution of Residuals" )
```

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



The distribution of residuals in the above histogram indicates that the Normality assumption is satisfied. The Normality assumption is satisfied because the distribution is unimodal and symmetric, approximately following a normal distribution.

Exercise 12

Based on the data description provided in the lab 3 instructions, I would say that the independence assumption is satisfied. One observation of family income does not depend on another family income, so the predictor variables are independent. However, more info about the data might yield a different conclusion. If there is a certain amount of money allocated to gift aid, gift aid might be relative to the pool of family income observations. I'm not sure if that applies here because gift aid is the response variable, but I thought it wouldn't hurt to mention.

Exercise 13

```
rsquared <- summary(model)$r.squared
rsquared
```

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

The calculation of R^2 = 0.2485582 (approximately 25%). This R^2 value indicates that 25% of the linear model's variation can be explained by variations of the inputs or predictor values (family income).

Exercise 14

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

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

According to this calculation, the 95% confidence interval is approximately (10.7, 30). Based on a family income = 90,000 dollars, I would predict that this persons gift aid will fall in the range of 10.7 thousand dollars and 30 thousand dollars.

Exercise 15

I do not think that it would be wise for this student to use my model to calculate the predicted gift aid. There are no entries in the dataset for family income that are close to a value of 310 thousand dollars, and as we have seen, observations with high family incomes are outliers in the dataset. Therefore, in the case of this student I don't think my model would accurately predict gift aid for the given predictor variable.