Lab 3

Dylan Scoble

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(openintro)
```

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
data(elmhurst)
elmhurst
```

```
## # A tibble: 50 x 3
##    family_income gift_aid price_paid
##            <dbl>    <dbl>      <dbl>
##  1         92.9      21.7      14.3
##  2          0.25     27.5       8.53
##  3         53.1      27.8      14.2
##  4         50.2      27.2       8.78
##  5        138.       18        24
##  6         48.0      18.5      23.5
##  7        114.       13        23
##  8        169.       13        29
##  9        208.       14        28
## 10         12.5      25.5      16.5
## # ... with 40 more rows
```

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

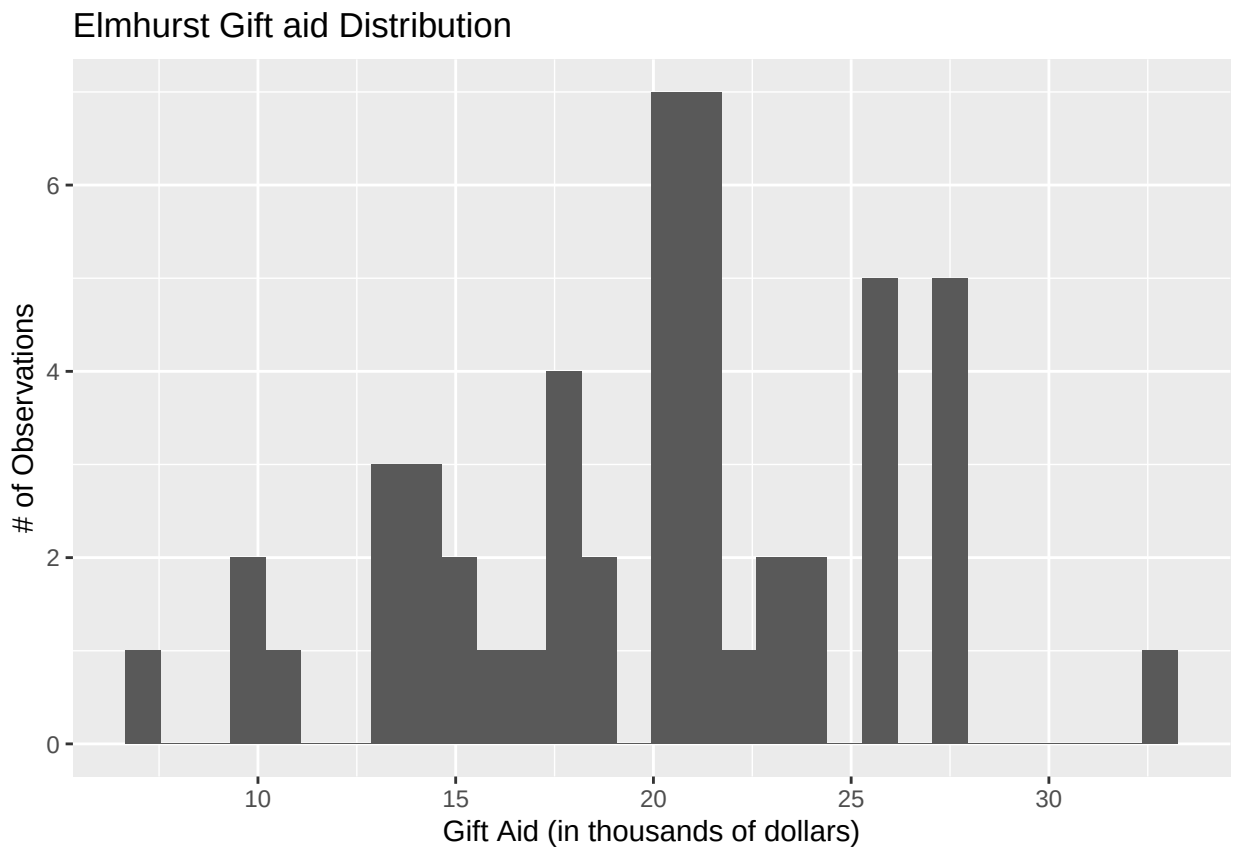
Exploratory Data Analysis

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.

The distribution has a similar shape to a normal distribution. There are no extreme outliers in this dataset.

```
ggplot(data = elmhurst, aes(x = gift_aid)) +
  geom_histogram() +
  labs(x = "Gift Aid (in thousands of dollars)",
       y = "# of Observations",
       title = "Elmhurst Gift aid 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.*

There is a center of approximately $20 and a spread of $5 - $8 in this dataset.

```
elmhurst %>%
  summarise(mean = mean(gift_aid),
            median = median(gift_aid),
            std_dev = sd(gift_aid),
            iqr = IQR(gift_aid),
  )
```

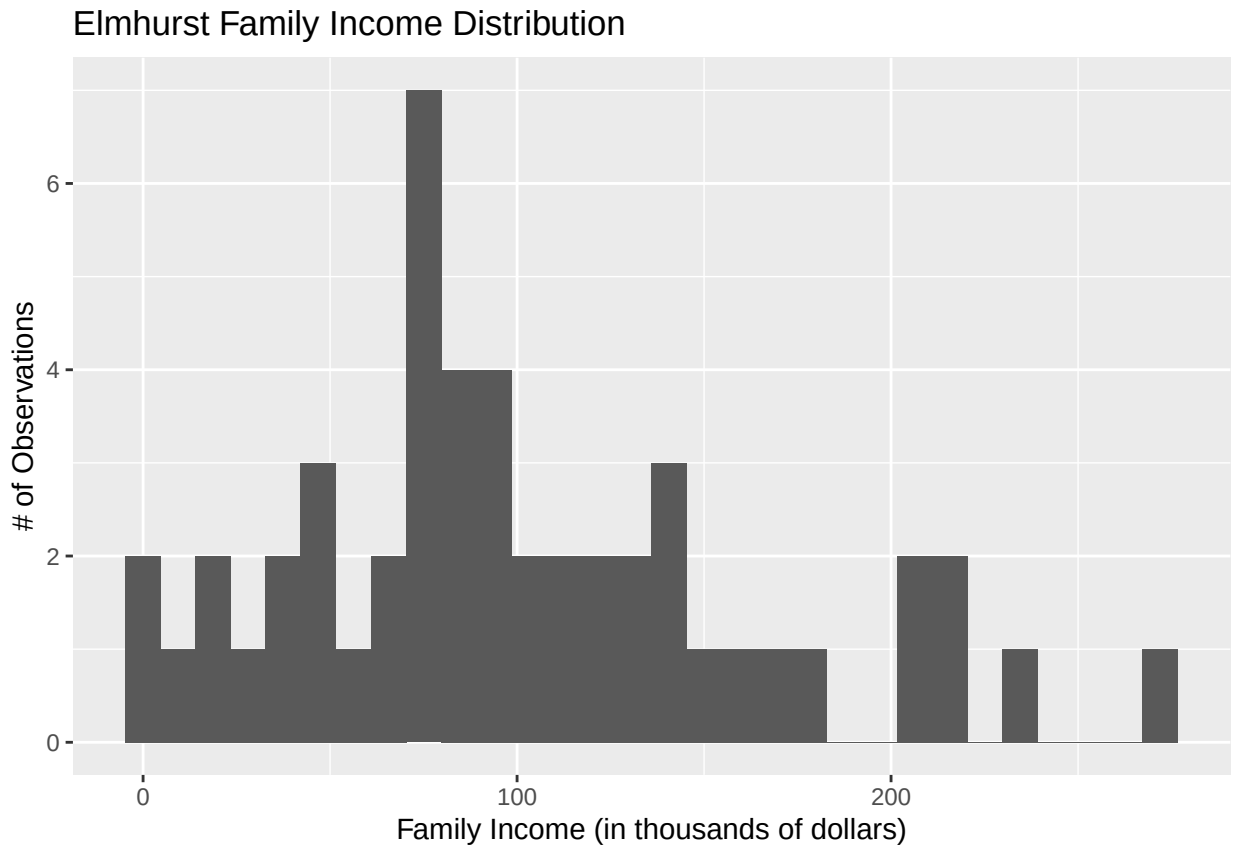
```
## # A tibble: 1 x 4
##    mean median std_dev   iqr
##   <dbl>  <dbl>   <dbl> <dbl>
## 1  19.9   20.5    5.46  7.26
```

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.

The distribution of family income is a relatively normal distribution with a center of $101.77 and a spread of $63 - $74. There is one obvious outlier that is a datapoint of a family who makes well over $200,000 in annual income.

```
ggplot(data = elmhurst, aes(x = family_income)) +
  geom_histogram() +
  labs(x = "Family Income (in thousands of dollars)",
       y = "# of Observations",
       title = "Elmhurst Family Income Distribution")
```

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




```
elmhurst %>%
  summarise(mean = mean(family_income),
            median = median(family_income),
            std_dev = sd(family_income),
            iqr = IQR(family_income),
  )
```

```
## # A tibble: 1 x 4
##    mean median std_dev   iqr
##   <dbl>  <dbl>   <dbl> <dbl>
## 1  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.

There appears to be a linear correlation between the response and the predictor variable. It would be logical to conclude that a line with a negative slope would best fit this correlation.

```
ggplot(data = elmhurst) +
  geom_point(mapping = aes(y = gift_aid, x = family_income)) +
  labs(y = "Gift Aid (in thousands of dollars)",
       x = "Family Income (in thousands of dollars)",
       title = "Elmhurst: correlation between gift aid and 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.*

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

For each $1000 increase in family income, our model predicts a decrease in gift aid of about $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.

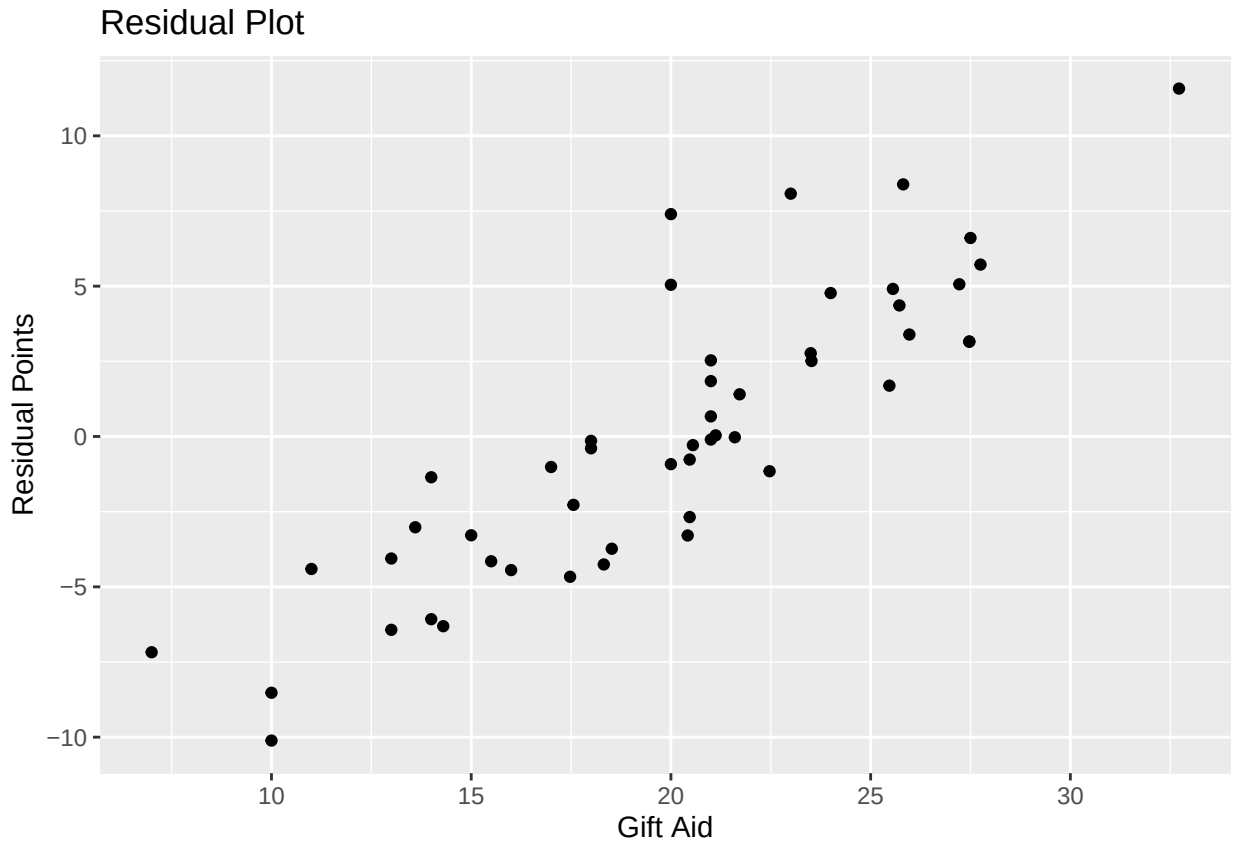
Let's begin by calculating the residuals and adding them to the dataset. Fill in the model name in the code below to add residuals to the original dataset using the `resid()` and `mutate()` functions.

```
elmhurst <- elmhurst %>%
  mutate(resid = residuals(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, mapping=aes(x=gift_aid, y=resid)) +
  geom_point() +
  labs(title="Residual Plot",
       x = "Gift Aid",
       y = "Residual Points")
```



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 no discernable shape in this plot. Therefore, the linear model adequately describes the relationship between gift aid and family income.

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.*

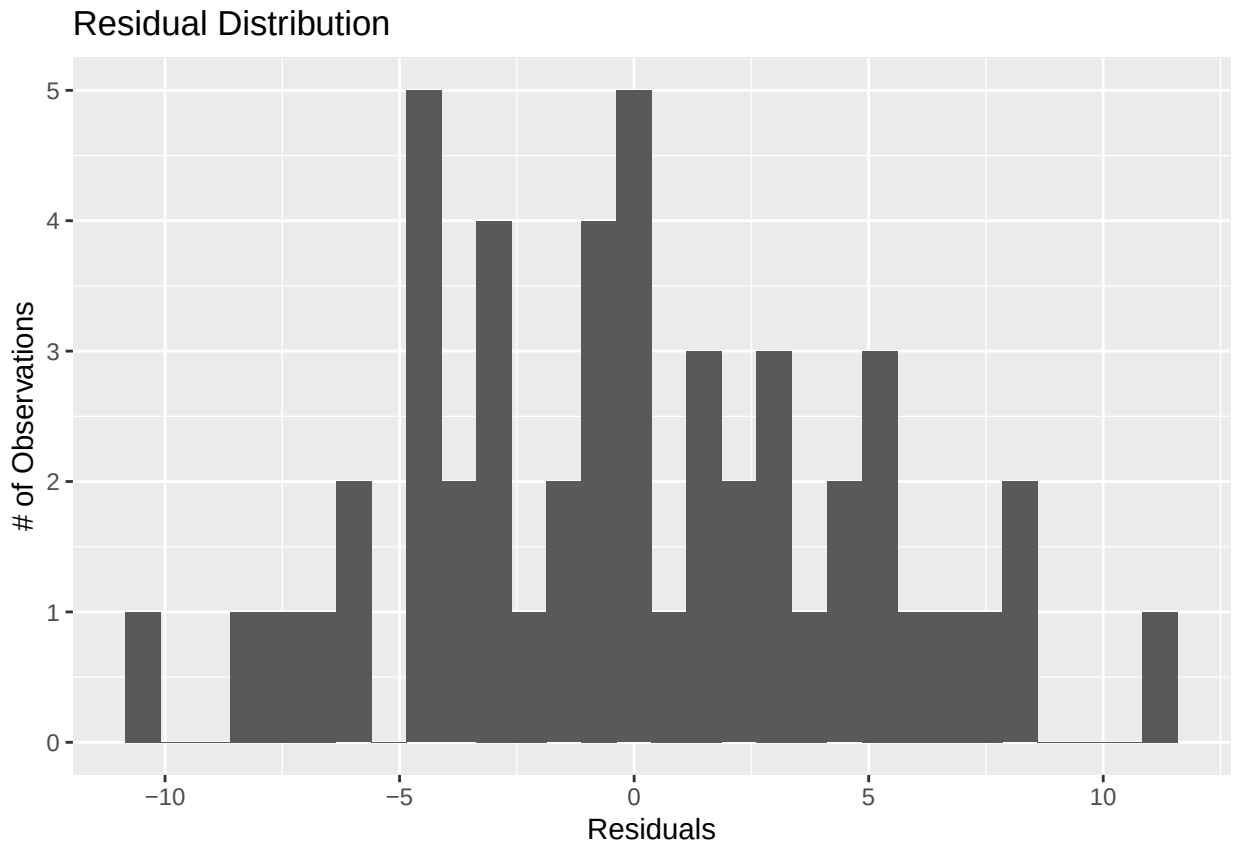
The constant variance assumption is not satisfied because the residuals scatterplot does not have a slope equal to 0. Because the slope is positive, there is a fan pattern, not a cloud pattern.

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 = resid)) +
  geom_histogram() +
  labs(x = "Residuals",
       y = "# of Observations",
       title = "Residual Distribution")
```

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



The normality assumption is satisfied because the histogram of the residuals approximately follows a normal distribution.

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.

The independence assumption is satisfied because each family in the dataset is completely different from one another. The families' income do not affect or depend on each other in any way.

Using the Model

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

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

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

An R^2 value of 0.2486 means that in this linear model, 24.86% of the data fits the model to a reasonable degree. In the context of the data and our model, 24.86% of the families in our dataset receive a reasonable amount of gift aid in relation to their 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.

Based on this model, a family with an income of $90,000 is expected to receive $20,442.88. I used the predict() function to calculate this value, shown below.

```
newdataset <-data.frame(family_income=c(90))
prediction <- predict(model, newdata=newdataset)
prediction
```

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

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.

It would not be wise to use the model because the student's family would be a major outlier in the data. The college would likely not use a model to calculate this family's aid due to their special circumstances.

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.