Lab 03: Simple Linear Regression

Isaac Plotkin

2/4/2022

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:

```
data(elmhurst)
```

The `elmhurst` dataset contains the following variables:

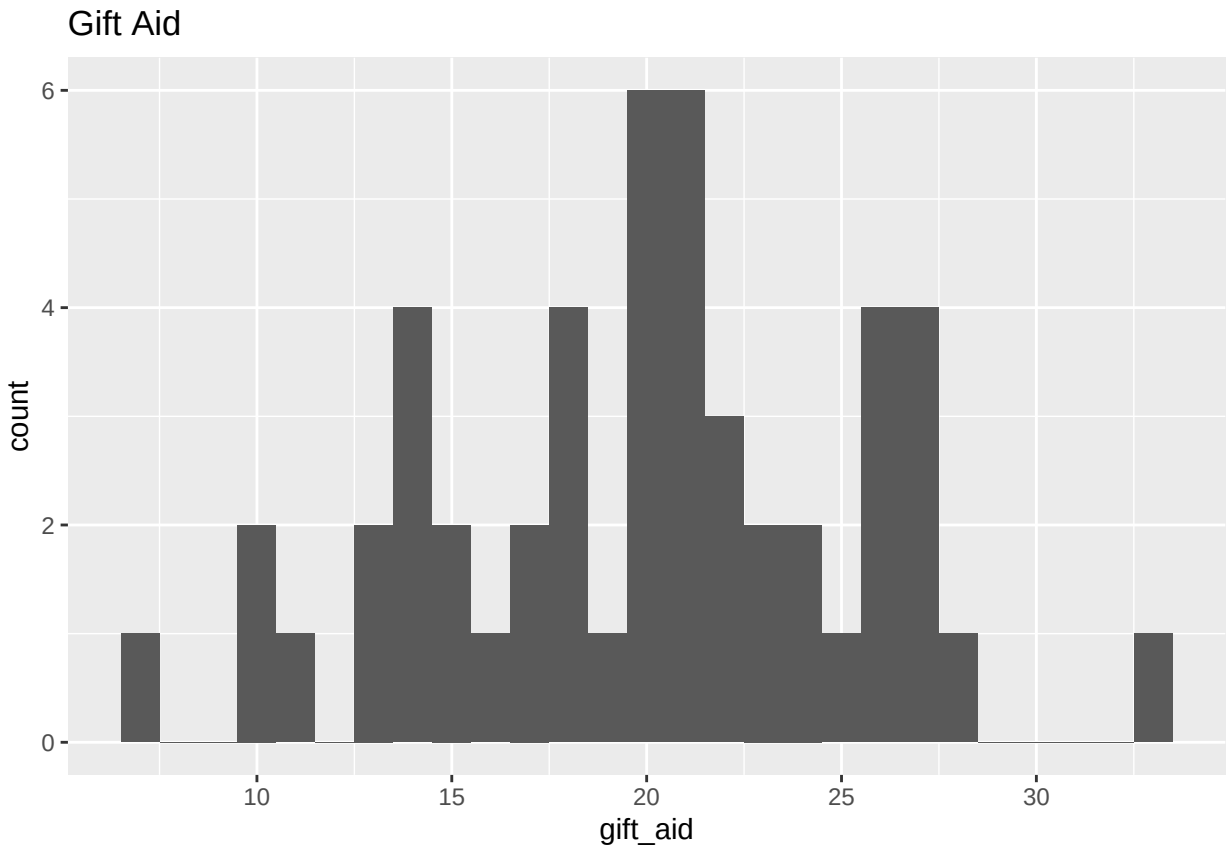
`family_income`	Family income of the student
`gift_aid`	Gift aid, in ($ thousands)
`price_paid`	Price paid by the student (= tuition - gift_aid)

Exercises

Exploratory Data Analysis

1.

```
ggplot(data = elmhurst, aes(x = gift_aid)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Gift Aid")
```



Gift_aid appears to have a normal distribution. There are 2 outliers. One person has gift_aid > $35,000 and another person has gift_aid < $5,000.

2.

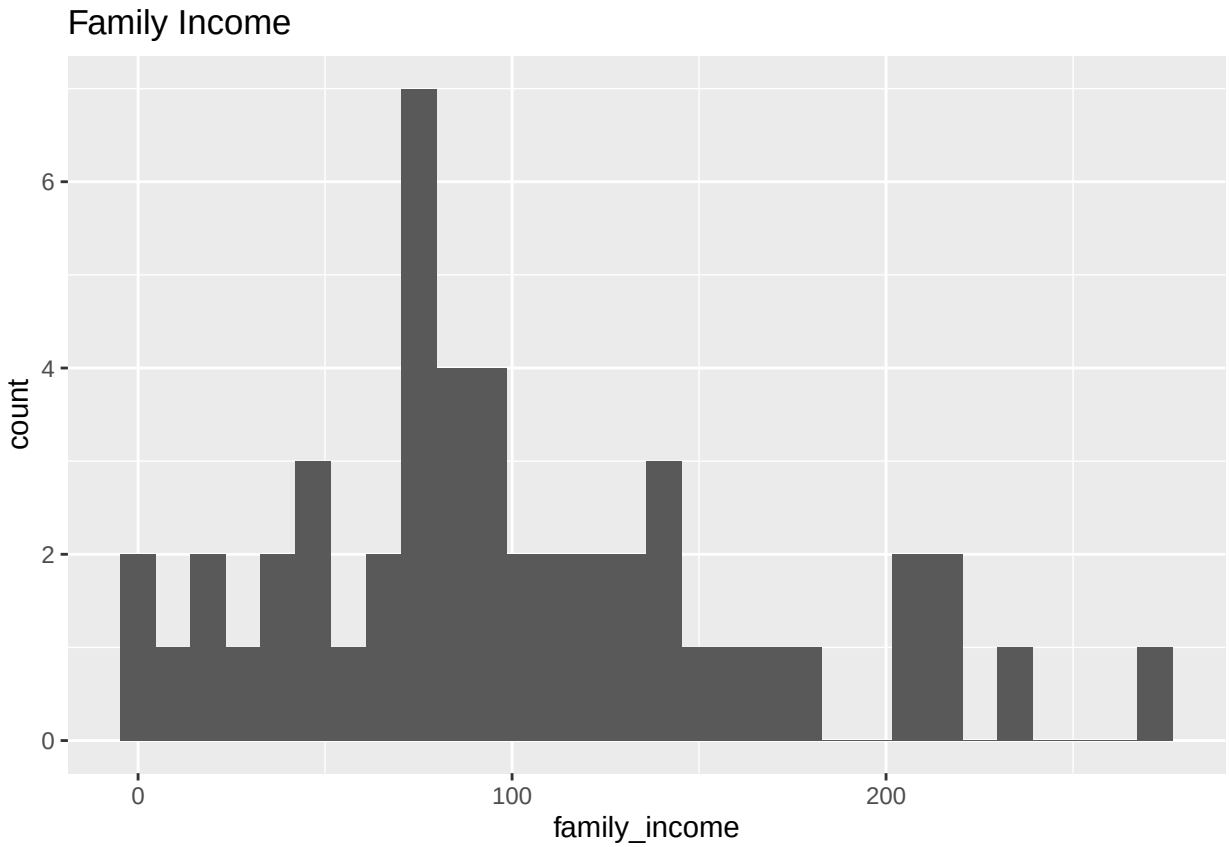
```
elmhurst %>%
  summarise(min = min(gift_aid),
            q1 = quantile(gift_aid, 0.25),
            q3 = quantile(gift_aid, 0.75),
            max = max(gift_aid),
            iqr = IQR(gift_aid),
            mean = mean(gift_aid),
            median = median(gift_aid),
            std_dev = sd(gift_aid)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

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

Gift_aid has a mean of ~$20,000 with a standard deviation of $5,460.

3.

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



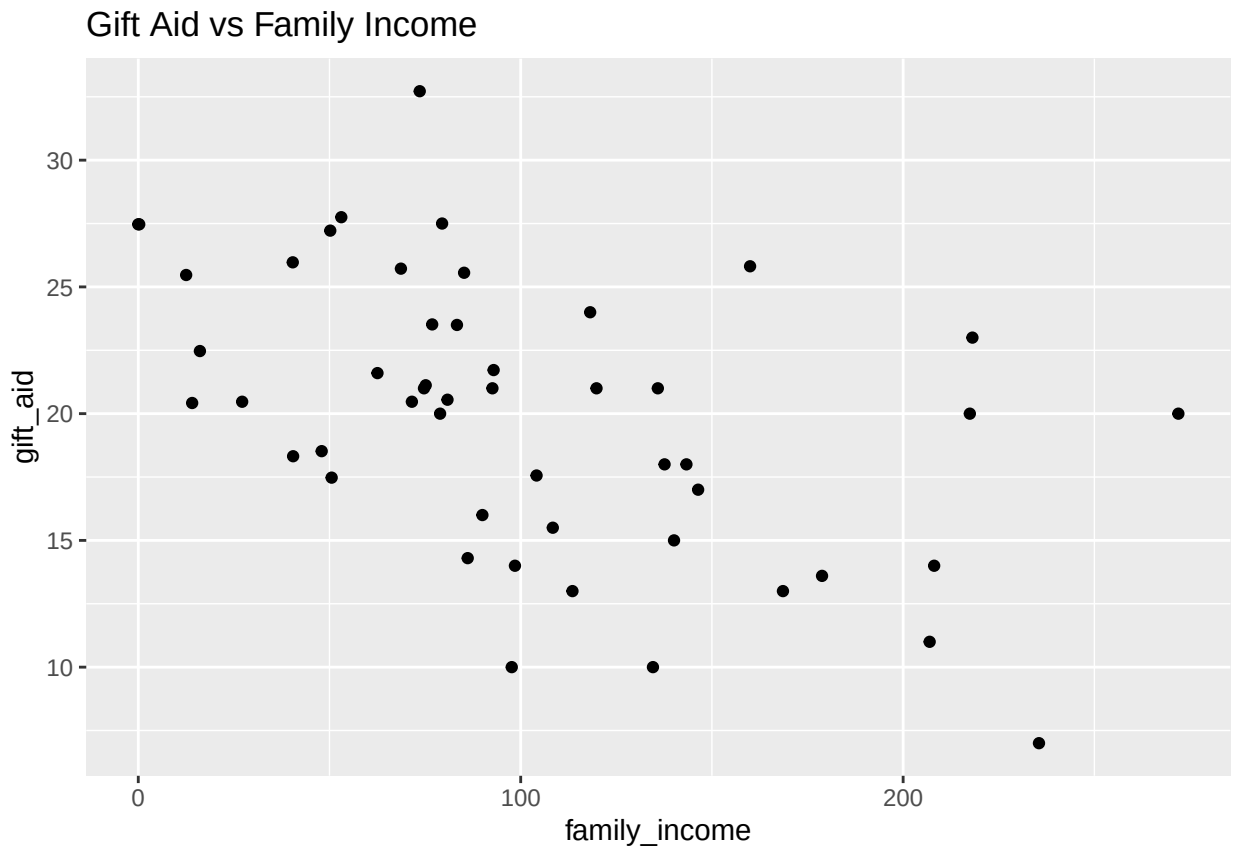

```
elmhurst %>%
  summarise(min = min(family_income),
            q1 = quantile(family_income, 0.25),
            q3 = quantile(family_income, 0.75),
            max = max(family_income),
            iqr = IQR(family_income),
            mean = mean(family_income),
            median = median(family_income),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

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

The distribution of family_income appears to be right skewed. It is centered at ~$100,000 dollars with a standard deviation of $63,206. There is one outlier that has a family income of > $250,000. There are also several people who have a family income = 0.

4.

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



There appears to be a negative correlation between family income and gift aid. As family income increases, gift aid decreases.

Simple Linear Regression

5.

```
gift_model <- lm(gift_aid ~ family_income, data = elmhurst)
tidy(gift_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

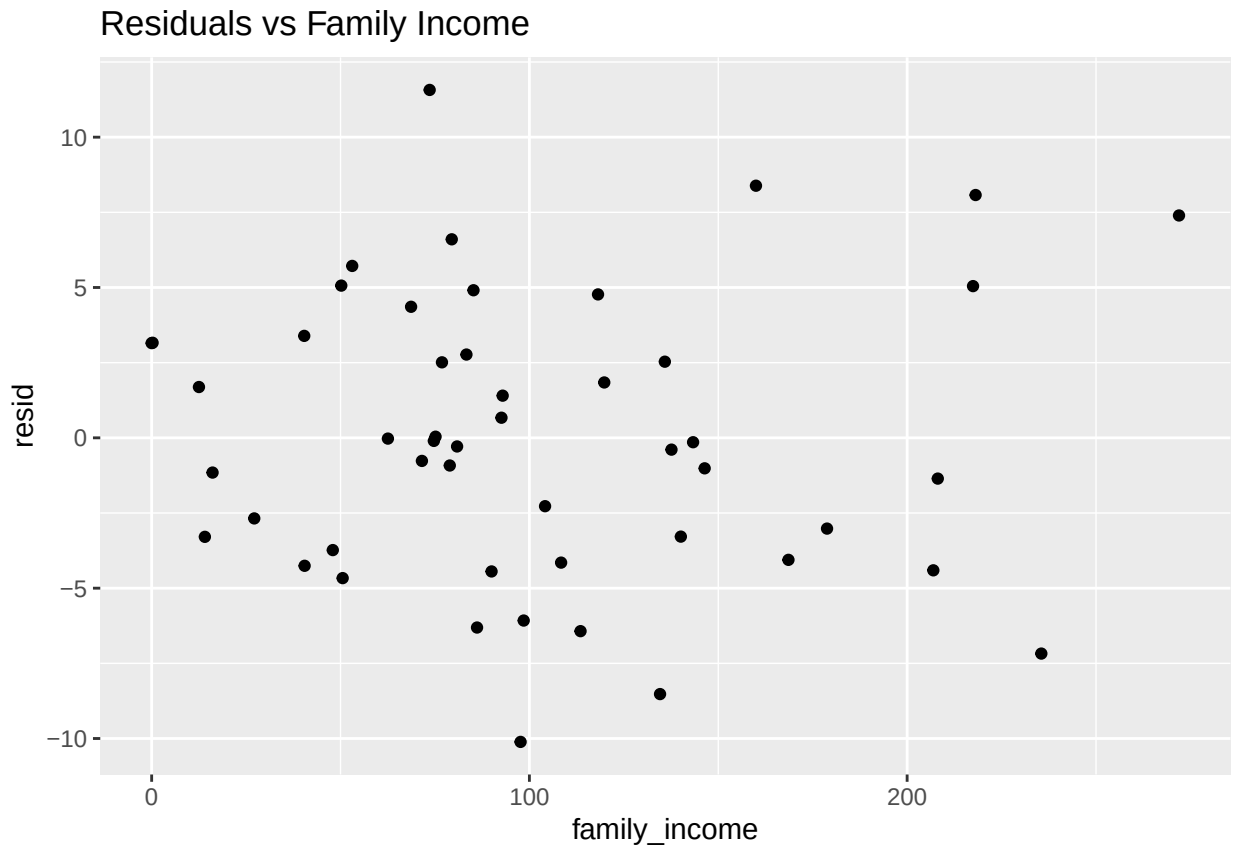
6. The slope is -0.043 for this problem. That means gift_aid goes down by $43 when family_income is increased by $1000. gift_aid = 24.319 - 0.043*family_income

7.

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

8.

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

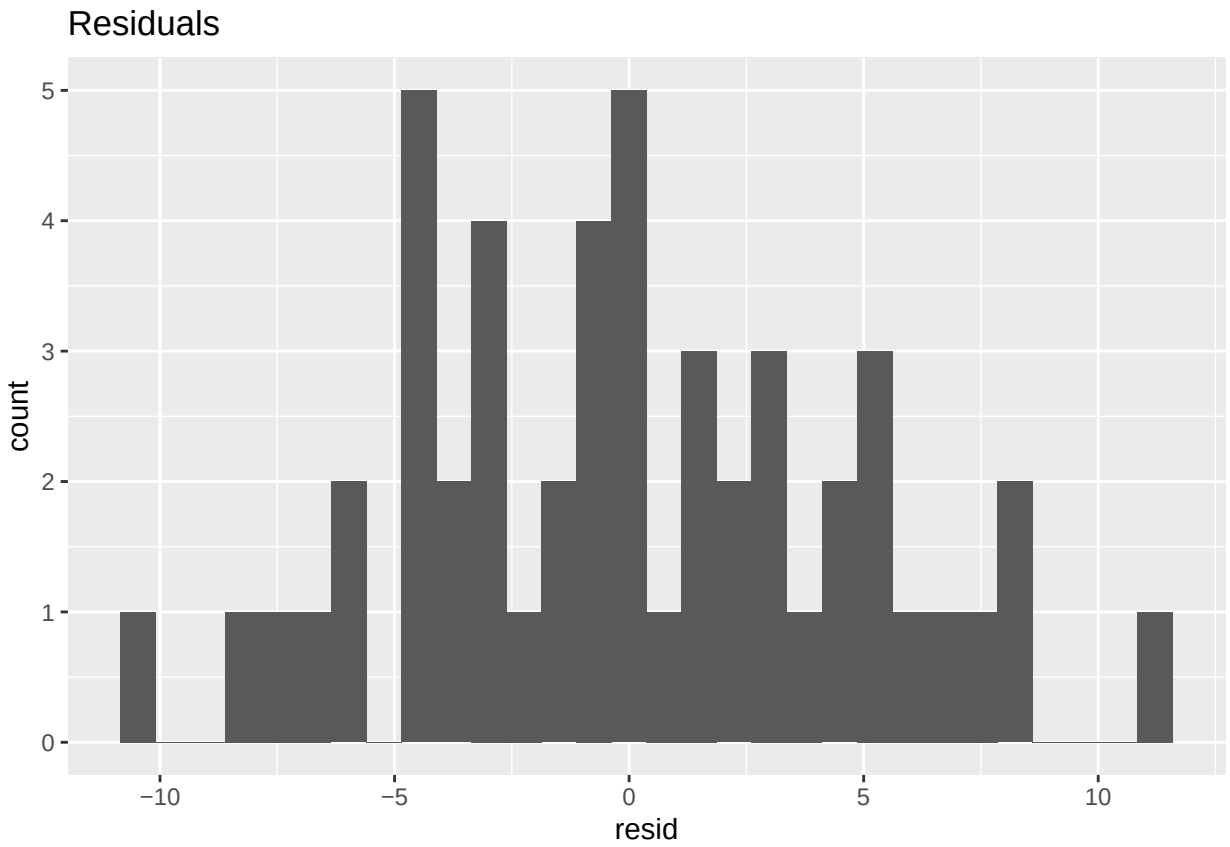


9. The linearity condition is satisfied because there is no discernible shape or patterns in the plot.

10. Yes the constant variance assumption is satisfied, The spread in the points appear constant as we move from left to right on the plot.

11.

```
ggplot(data = elmhurst, aes(x = resid)) +
  geom_histogram() +
  labs(title = "Residuals")
```



The residual histogram seems like it follows a normal distribution. Therefore, the normality assumption is satisfied.

12. The students were randomly selected so there is not a high chance that the observations are dependent on eachother. The independence assumption is satisfied.

Using the Model

13.

```
rsquare(gift_model, elmhurst)
```

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

24.86% of the variation in gift_price is explained by family_income.

14.

```
# as.numeric(predict(gift_model, data.frame(family_income = 90)))
y = 24.319 - 0.043*(90)
y
```

```
## [1] 20.449
```

She can expect to get about $20,449 in gift aid.

15. It is probably not wise to use my model because $310,000 is an outlier compared to the data used in the model. The student can use the model but they should keep in mind that the results might not be accurate.

My repository for this lab can be found here: https://github.com/iplotkin/lab-03