Analysis of Salary data by graduates from 706 universities in the United States

This program was written by Leroy Wheeler with guidance from Dr. Steve Simon on 2024-10-25. It is placed in the public domain.

Libraries

```
library(broom)
library(tidyverse)
```

Read data https://dasl.datadescription.com/datafile/earnings/

```
earnings <- read_tsv(
  file="../data/earnings.txt",
  col_types="ccnnnnnnnnc")
names(earnings) <- tolower(names(earnings))
glimpse(earnings)
```

```
Rows: 706
Columns: 11
$ school           <chr> "Princeton University", "University of Michigan-Ann A…
$ place            <chr> "Princeton, NJ", "Ann Arbor, MI", "Cambridge, MA", "H…
$ price            <dbl> 61300, 28100, 64800, 58600, 35700, 18500, 66600, 6280…
$ `price with aid` <dbl> 20600, 17300, 16500, 22400, 18200, 13400, 16000, 1900…
$ pct.need         <dbl> 59, 30, 58, 39, 51, 39, 58, 32, 27, 48, 56, 50, 37, 5…
$ `merit aided`    <dbl> NA, 16, NA, 11, 6, 24, NA, 55, 3, NA, NA, NA, 6, 2, 6…
$ earn             <dbl> 62800, 59000, 62900, 63700, 60300, 51800, 53400, 6320…
$ sat              <dbl> 1510, 1380, 1510, 1460, 1360, 1260, 1440, 1360, 1360,…
$ act              <dbl> 33, 30, 34, 33, 30, 29, 32, 31, 31, 33, 34, 34, 26, 2…
$ `sat/act`        <dbl> 1510, 1380, 1510, 1460, 1360, 1260, 1440, 1360, 1360,…
$ public           <chr> "0", "1", "0", "0", "1", "1", "1", "0", "1", "0", "0"…
```

Descriptive statistics of graduate's annual salary

```
earnings |>
  summarise(
    earn_mn=mean(earn, na.rm=TRUE),
    earn_sd=sd(earn, na.rm=TRUE),
    earn_min=min(earn, na.rm=TRUE),
    earn_max=max(earn, na.rm=TRUE),
    n_missing=sum(is.na(earn)))
```

```
# A tibble: 1 × 5
  earn_mn earn_sd earn_min earn_max n_missing
    <dbl>   <dbl>    <dbl>    <dbl>     <int>
1  45598.   6724.    28300    79700         0
```

There is a pretty large range of values for graduate's salary with no missing values.

Descriptive statistics of price paid for degree

```
earnings |>
  summarise(
    price_mn=mean(price, na.rm=TRUE),
    price_sd=sd(price, na.rm=TRUE),
    price_min=min(price, na.rm=TRUE),
    price_max=max(price, na.rm=TRUE),
    n_missing=sum(is.na(price)))
```

```
# A tibble: 1 × 5
  price_mn price_sd price_min price_max n_missing
     <dbl>    <dbl>     <dbl>     <dbl>     <int>
1   42200.   15727.     16500     70400         0
```

There is a wide range of values for amount spent per year to attend college with no missing values.

Linear model:

[Salary earned] = β0 + β1 * [Price paid for degree]

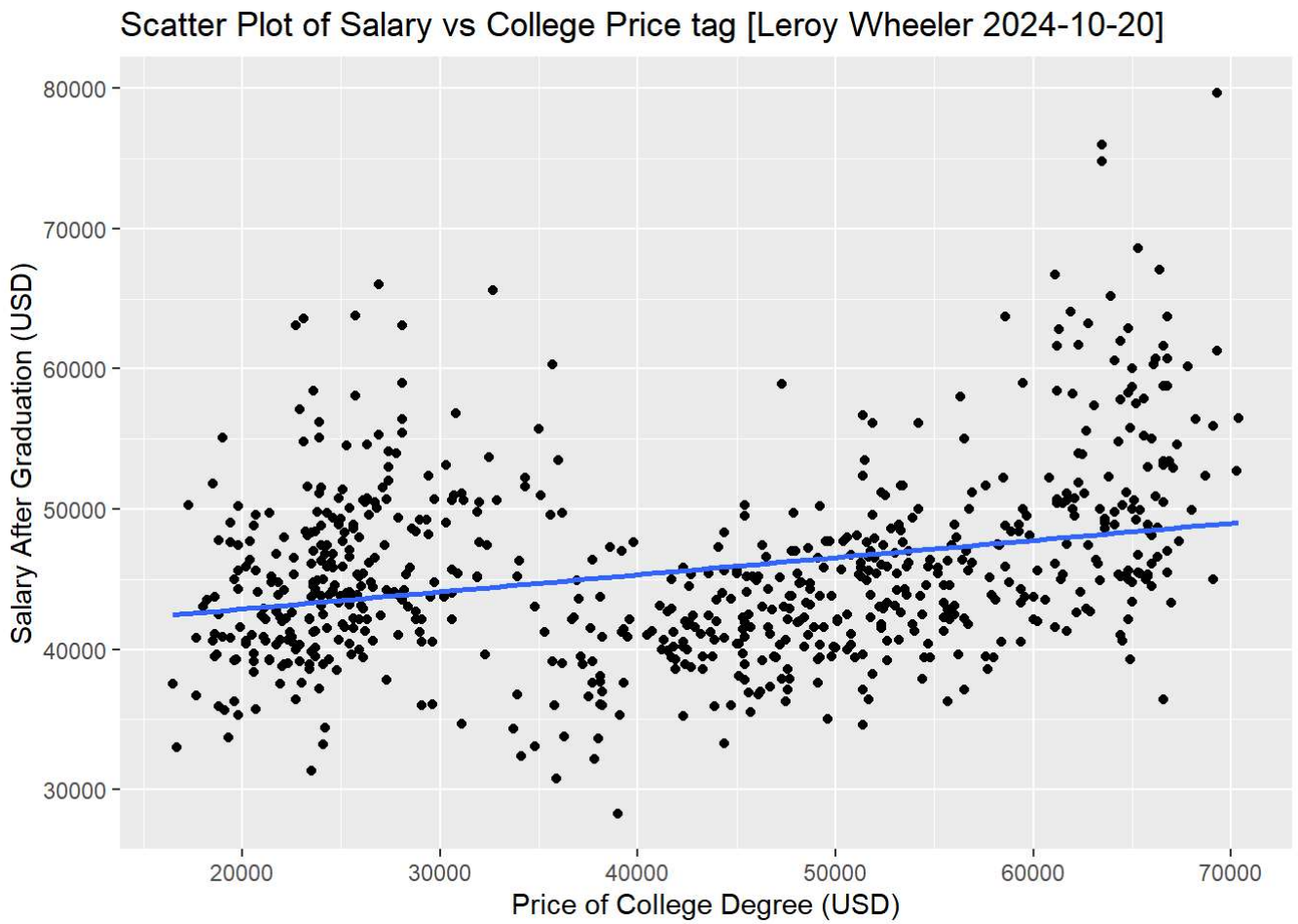
Hypothesis:

H0: β1 = 0 → No relationship between price of degree and salary earned

Ha: β1 ≠ 0 → A relationship exists between price of degree and salary earned

```
earnings|>
    ggplot(aes(price, earn)) +
      geom_point() +
      xlab("Price of College Degree (USD)") +
      ylab("Salary After Graduation (USD)") +
    geom_smooth(method="lm", se=FALSE) +
    ggtitle("Scatter Plot of Salary vs College Price tag [Leroy Wheeler 2024-10-20]")
```

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



There appears to be a small positive linear relationship between price paid for college degree and the salary earned by college graduates.

```
m1 <- lm(earn~price, data=earnings)
m1
```

```
Call:
lm(formula = earn ~ price, data = earnings)

Coefficients:
```

```
(Intercept)        price
  4.042e+04    1.227e-01
```

The linear regression model predicts that for every additional dollar spent on the college degree, the graduates with that degree would expect to earn an additional 12 cents per year in salary.

```
anova(m1)
```

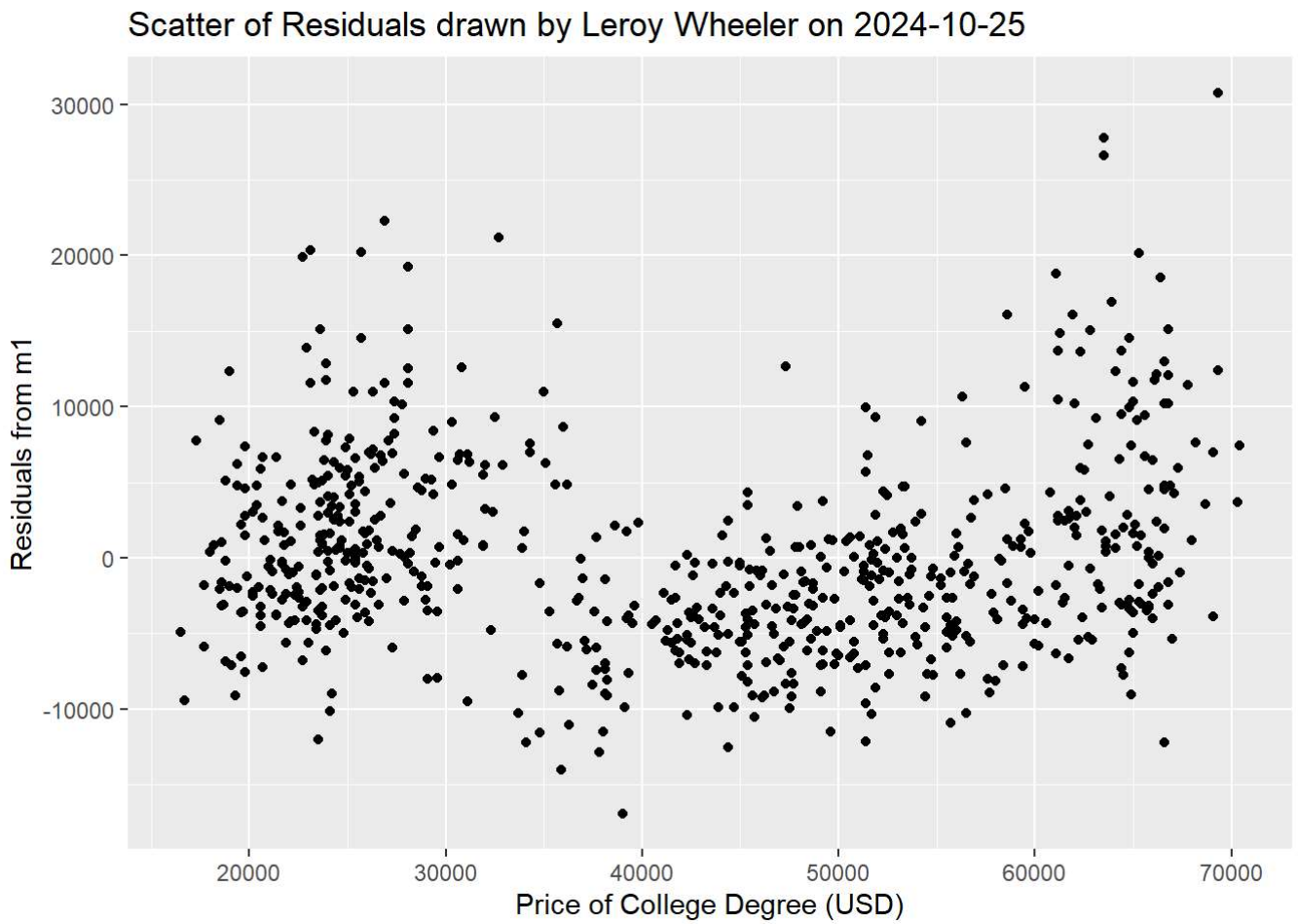
```
Analysis of Variance Table

Response: earn
           Df     Sum Sq    Mean Sq F value    Pr(>F)
price       1 2.6244e+09 2624415359  63.166 7.552e-15 ***
Residuals 704 2.9250e+10   41547620
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glance(m1)$r.squared
```

```
[1] 0.08233734
```

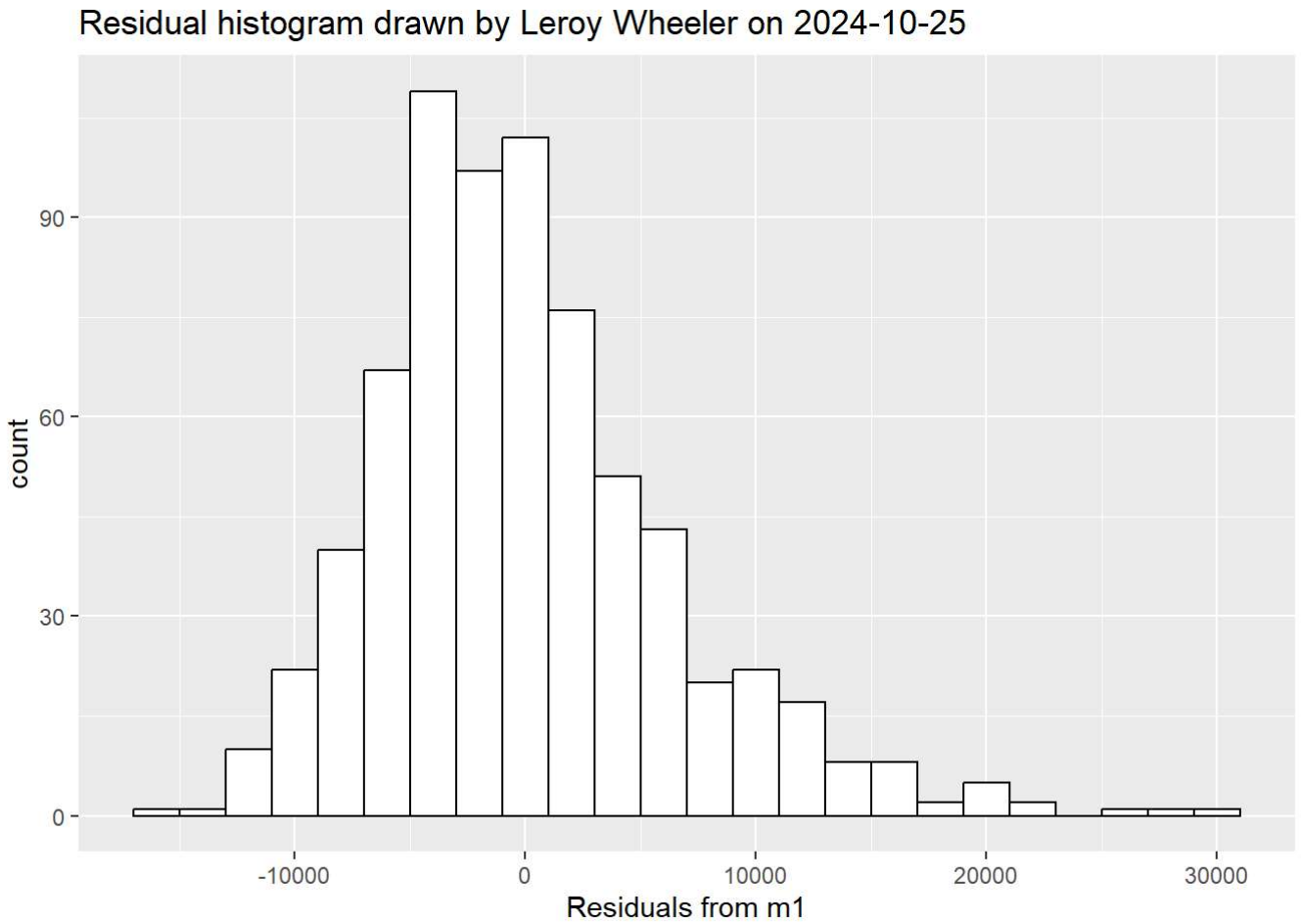
With an R squared value of 0.08, we conclude that the relationship between price paid for college degree and the salary earned upon graduation is very weak. The ANOVA table displays a large F value and a small p-value for our linear regression model, therefore we conclude that this weak relationship is likely true.

```
m1 |>
  ggplot(aes(price, .resid)) +
    geom_point() +
      ggtitle("Scatter of Residuals drawn by Leroy Wheeler on 2024-10-25") +
      xlab("Price of College Degree (USD)") +
      ylab("Residuals from m1")
```



This scatterplot displays the residual variation according to college price. It looks like there is some non-linearity in the data towards the more expensive schools.

```
m1 |>
  ggplot(aes(.resid)) +
    geom_histogram(
      binwidth=2000,
      color="black",
      fill="white") +
      ggtitle("Residual histogram drawn by Leroy Wheeler on 2024-10-25") +
      xlab("Residuals from m1")
```



A plot of the residuals demonstrates a little right skew in the residuals. This linear model may be improved by a log transformation, however with our large sample size (n = 706), we should be able to safely ignore a little departure from normal distribution.