# Final Project

Alyssa Garcia

2022-07-28

## Question of Interest

In recent years the number of females entering the technology field has been increasing, to explore this topic, we will be using the "College Scorecard" dataset, to determine whether there is a relationship between the percent of degrees award in Computer and Information Science and Support Services and percent of female students who completed within 4 years at an original institution, within Virginia? To examine our question, we will be using a linear model with the variables PCIP11, FEMALE_COMP_ORIG_YR4_RT, ST_FIPS, FEMALE, and INSTNM.

## Preprocessing

```r
college_reduced <- college %>%
  select(PCIP11, FEMALE_COMP_ORIG_YR4_RT, ST_FIPS, FEMALE, INSTNM)
```

To select the columns being used within this projects, we used the select function and assigned the new dataset to "college_reduced".

```r
college_reduced_2 <- college_reduced %>%
  filter(ST_FIPS == 51)
```

Since we are only interested at looking at colleges within Virginia, we used the filter function to only display Virginia colleges and assigned the dataset to "college_reduced_2".

```r
college_renamed <- college_reduced_2 %>%
  rename(
    degree_percentage_computer = PCIP11,
    female_completed_4yrs = FEMALE_COMP_ORIG_YR4_RT,
    state_code = ST_FIPS,
    demographics_female = FEMALE,
    school_name = INSTNM
  )
```

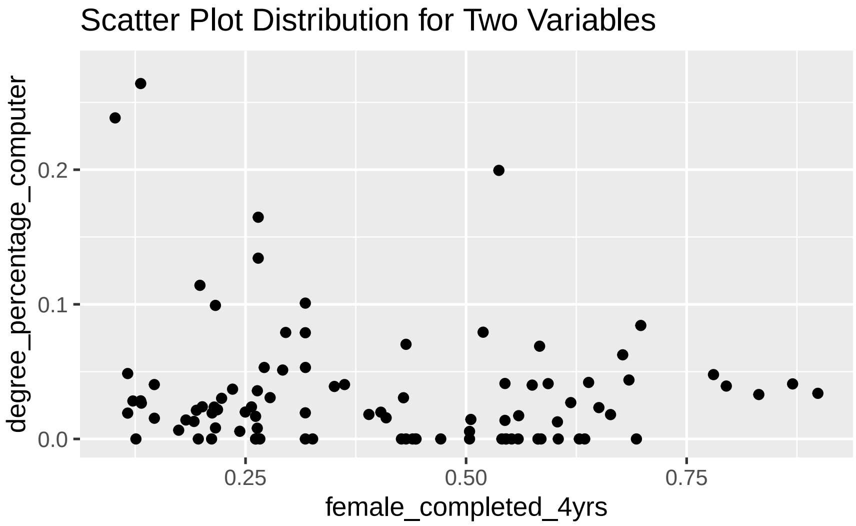
The original column names are abbreviations, to make the them more readable we used the rename function and assigned the new dataset to "college_renamed"

### Visualization

In order to visualize the covariation between our two variables, we have created a scatter plot using female_completed_4yrs and degree_percentage_computer.

```
college_renamed %>%
  ggplot() +
  geom_point(aes(x = female_completed_4yrs, y = degree_percentage_computer)) +
  labs(
    title = "Scatter Plot Distribution for Two Variables",
    x = "female_completed_4yrs",
    y = "degree_percentage_computer"
  )
```

```
## Warning: Removed 81 rows containing missing values (geom_point).
```

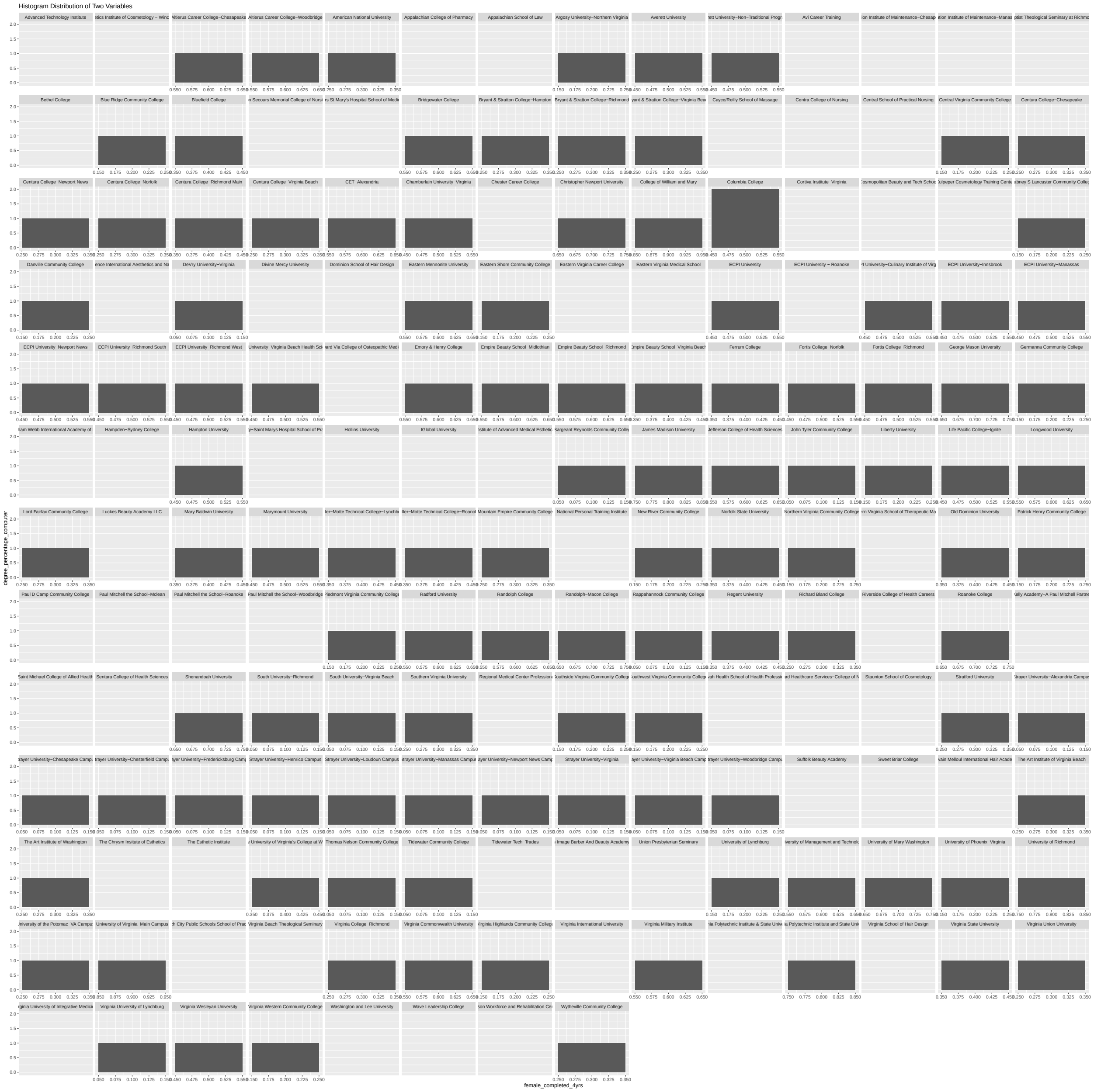


There seems to be a weak correlation between the points on this scatter plot. I seems that there is almost no pattern between the variables. Another variable that might influence these patterns could be using a different number of years taken to complete their degree.

In order to visualize the variation between our two variables, we have created a histogram using female_completed_4yrs and degree_percentage_computer.

```
college_renamed %>%
  ggplot() +
  geom_histogram(aes(x = female_completed_4yrs), bins = 10) +
  facet_wrap(~ school_name, scales = "free_x") +
  labs(
    title = "Histogram Distribution of Two Variables",
    x = "female_completed_4yrs",
    y = "degree_percentage_computer"
  )
```

```
## Warning: Removed 63 rows containing non-finite values (stat_bin).
```



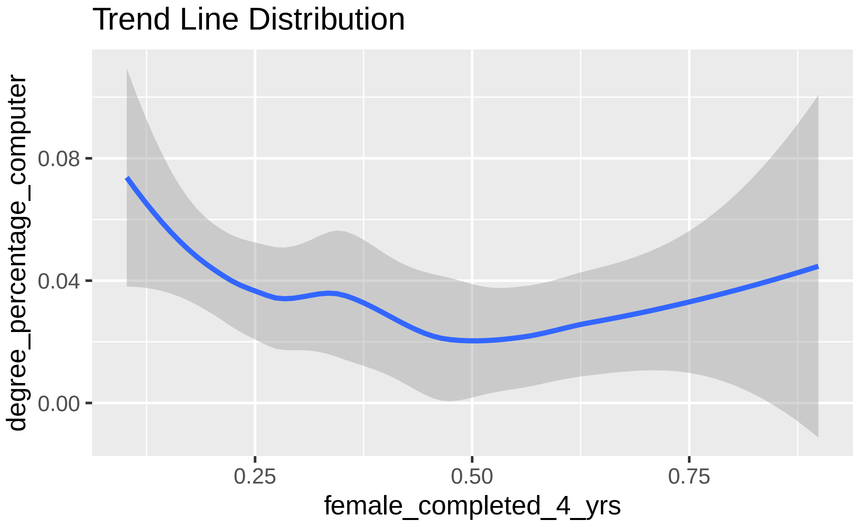
This histograms shown above mostly are mostly uniform with no outliers.

In order to visualize the covariation between our two variables, we have created a trend line using female_completed_4yrs and degree_percentage_computer.

```
college_renamed %>%
  ggplot() +
  geom_smooth(aes(x = female_completed_4yrs, y = degree_percentage_computer)) +
  labs(
    title = "Trend Line Distribution",
    x = "female_completed_4_yrs",
    y = "degree_percentage_computer"
  )
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 81 rows containing non-finite values (stat_smooth).
```



The trend line displays a negative pattern between our two variables, female_completed_4yrs and degree_percentage_computer. However, there is a slight increase towards the end of the trend line.

### Summary Statistics

```
college_statistics <- college_renamed %>%
  na.omit
```

In order to display accurate summary statistics, we have opted to remove all missing values from our dataset, which we assigned to a new data frame "college_statistics".

Now we want to look at the summary statistics of all our variables within our "college_statistics" dataframe. The first variable we will use is "degree_percentage_computer".

```
college_statistics %>%
  summarize(
    count = n(),
    mean = mean(degree_percentage_computer),
    median = median(degree_percentage_computer),
    std.dev = sd(degree_percentage_computer),
    iqr = IQR(degree_percentage_computer)
  )
```

| count | mean | median | std.dev | iqr |
|---:|---:|---:|---:|---:|
| 96 | 0.0344021 | 0.02065 | 0.0480324 | 0.04095 |

The second variable we be looking at is “demographics_female”.

```
college_statistics %>%
  summarize(
    count = n(),
    mean = mean(demographics_female),
    median = median(demographics_female),
    std.dev = sd(demographics_female),
    iqr = IQR(demographics_female)
  )
```

| count | mean | median | std.dev | iqr |
|---:|---:|---:|---:|---:|
| 96 | 0.644908 | 0.6232215 | 0.1468727 | 0.1508355 |

The third variable we be looking at is “female_completed_4yrs”.

```
college_statistics %>%
  summarize(
    count = n(),
    mean = mean(female_completed_4yrs),
    median = median(female_completed_4yrs),
    std.dev = sd(female_completed_4yrs),
    iqr = IQR(female_completed_4yrs)
  )
```

| count | mean | median | std.dev | iqr |
|---:|---:|---:|---:|---:|
| 96 | 0.3994123 | 0.3564745 | 0.2010603 | 0.3374598 |

The fourth variable we be looking at is “state_code”.Since this variable is categorical, we must use the group_by function to find the number of rows.

```
college_statistics %>%
  group_by(state_code) %>%
  summarize(
    count = n()
    )
```

| state_code | count |
|---:|---:|
| 51 | 96 |

The fifth variable we be looking at is "school_name".Since this variable is categorical, we must use the group_by function to find the number of rows. In addition to this, we also used the head() function to limit the number of rows we will see in the tibble below.

```
college_statistics %>%
  group_by(school_name) %>%
  summarize(
    count = n()) %>%
  head()
```

| school_name | count |
|:---|---:|
| Altierus Career College-Chesapeake | 1 |
| Altierus Career College-Woodbridge | 1 |
| American National University | 1 |
| Argosy University-Northern Virginia | 1 |
| Averett University | 1 |
| Averett University-Non-Traditional Programs | 1 |

## Data Analysis

Now that we have finished our exploratory data analysis, we can create our linear regression model using the lm() function.

```
college_model <- lm(degree_percentage_computer ~ female_completed_4yrs, data = college_statisti
```

To get the data frame to summarize the model's parameters, we will use the tidy() function.

```
college_model %>%
  tidy()
```

| term | estimate | std.error | statistic | p.value |
|:---|---:|---:|---:|---:|
| (Intercept) | 0.0488323 | 0.0108800 | 4.488262 | 0.0000203 |
| female_completed_4yrs | -0.0361287 | 0.0243568 | -1.483309 | 0.1413375 |

The values under the estimate column gives us the intercept and slope of our linear model.

For additional information about our model, we can obtain the R^2 parameter using the glance() function and piping it into the select() function so that it will not overflow the margin of our knitted document.

```
college_model %>%
  glance() %>%
  select(r.squared)
```

| r.squared |
|---|
| 0.0228711 |

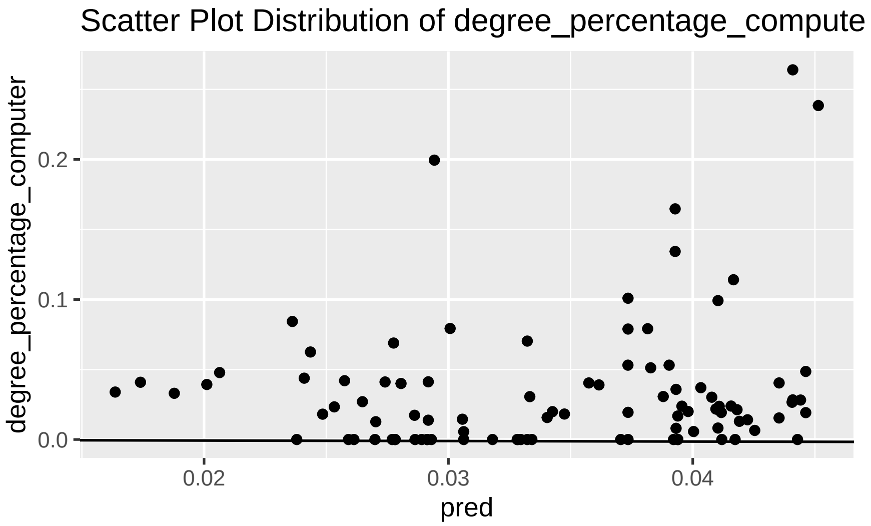
As we can see from our tibble, our R^2 is closer to 0, which means is its doing a poor job capturing the variability of the response variable.

After building our model, we want to know what it predicts and how accurate the predictions are. To do this, we must use the add_predictions() and add_residuals() function to add the model predictions and residuals to the data frame "college_df".

```
college_df <- college_statistics %>%
  add_predictions(college_model) %>%
  add_residuals(college_model)
```

Now that we have our residuals and predictions, we can start plotting these. The scatter plot below is our observed values vs. our predicted values.

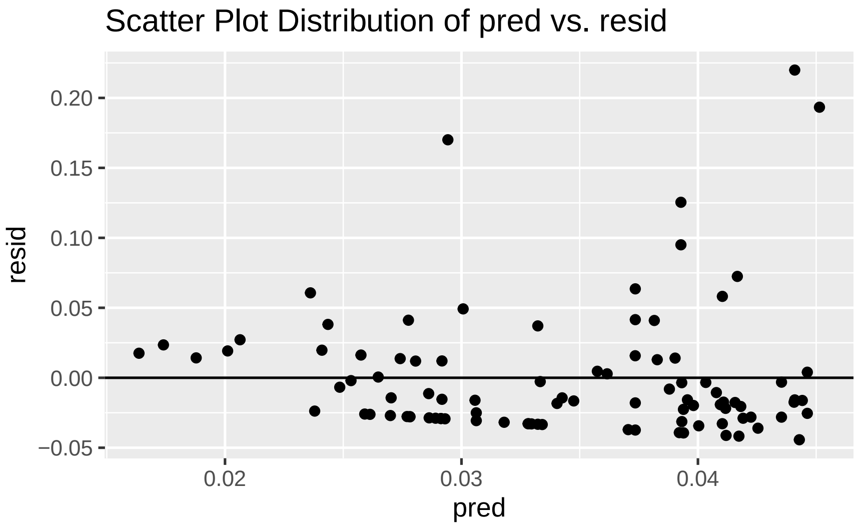
```
college_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = degree_percentage_computer)) +
  geom_abline(slope = -0.0361, intercept = 0) +
  labs(
    title = "Scatter Plot Distribution of degree_percentage_compute vs. pred",
    x = "pred",
    y = "degree_percentage_computer"
  )
```



This scatter plot does not meet the first condition, linearity, as the response and explanatory variables barely fall along the line. The scatter plot also does not meet the second condition, nearly normal residuals, as the residuals are not normally distributed. The scatter plot does not meet the third condition, constant variation, as the residuals do not fall along the line at a constant rate.

Now we can create a scatter plot of our residuals vs. predicted.

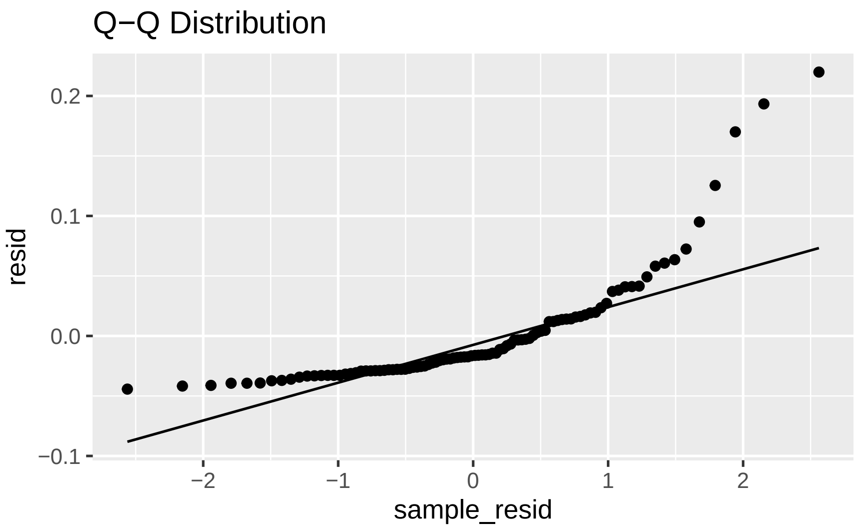
```
college_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = resid)) +
  geom_hline(yintercept = 0) +
  labs(
    title = "Scatter Plot Distribution of pred vs. resid",
    x = "pred",
    y = "resid"
  )
```



This scatter plot does meet the first condition, linearity, as the response and explanatory variables fall slightly along the line. The scatter plot does not meet the second condition, nearly normal residuals, as the residuals are not normally distributed. The scatter plot doesn't meet the third condition, constant variation, as the residuals do not fall along the line at a constant rate.

Now we can creat our Q-Q plot.

```
college_df %>%
  ggplot() +
  geom_qq(aes(sample = resid)) +
  geom_qq_line(aes(sample = resid)) +
  labs (
    title = "Q-Q Distribution",
    x = "sample_resid",
    y = "resid"
  )
```



This Q-Q plot does not meet the first condition, linearity, as the response and explanatory variables is curved and does not fall along the line. The Q-Q plot does not meet the second condition, nearly normal residuals, as the residuals are not normally distributed. The Q-Q plot doesn't meet the third condition, constant variation, as the residuals do not fall along the line at a constant rate.

## Conclusion

By reviewing all sections from our analysis, we can draw that there is no linearity, normal residuals, and constant variation between the variables "female_completed_4yrs" and "degree_percentage_computer". We can also see from our r.squared that our model was doing a poor job at capturing the variability in our response variable.

Thus, we can conclude that there is a weak to little relationship between our variables "female_completed_4yrs" and "degree_percentage_computer". This could mean that there is a small number of females within Virginia colleges that are receiving a computer degree within four years. However, there could be confounding variables, like the new average degree completion takes 5 years, that we did not look at within our model.