Final Project

Alizeh Murtaza

2023-05-08

Introduction

My question that I will be investigating is "Is There a Relationship Between Age of Entry and Highest Degree Earned?" The variables I will use for this question are:

- AGE_ENTRY
- AGEGE24
- PCT_BA
- HIGHDEG

I will use a linear model for this question.

I believe that this is an interesting question because it is relevant in the United States today. For one, I believe that there is much discrimination against people older than middle-aged. There are less oppurtunities for them and it makes living hard for them. This is just for citizens of the US. However, it can be even harder for immigrants. If someone wants to come to the US to build a better life for their family, it is possible to take over a decade for them to be allowed to immigrate. At that point, whoever wanted to immigrate may be older than the average college student. It may be harder for them to support their family or even themself if they want to obtain a college degree. I believe this question will allow us to see if there really is discrimination age-wise in college education.

Preprocessing

```
college_short <- college %>%
  select(AGE_ENTRY, AGE_ENTRY_SQ, AGEGE24,
         PCT_BA, PREDDEG, HIGHDEG, C200_4,
         C200_L4)
```

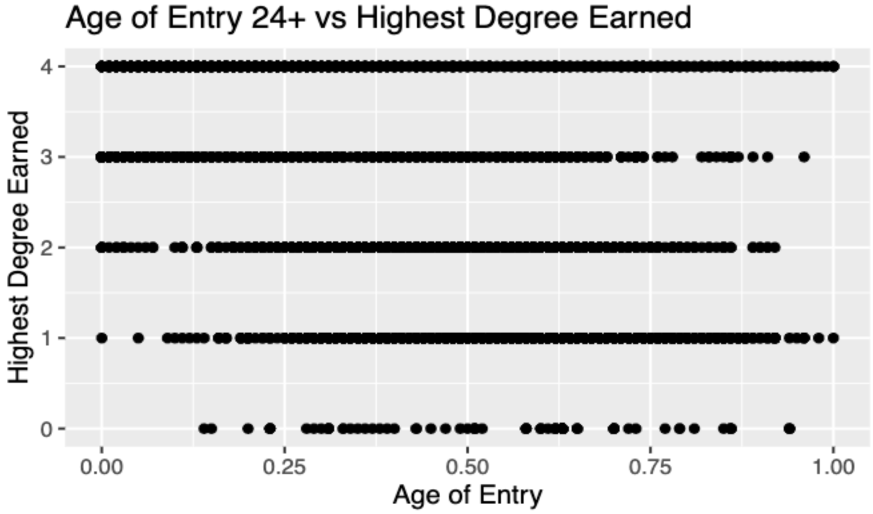
```
college_renamed <- college_short %>%
  rename(
    age_entry = AGE_ENTRY,
    age_24_plus = AGEGE24,
    pct_25_bach = PCT_BA,
    highdeg = HIGHDEG,
    comp_4 = C200_4,
    comp_l_4 = C200_L4
  )
```

Visualization

Why? I want to see the relationship between the highest degree earned and the age of the student, specifically students that are over the age of 23.

```
ggplot(data = college_renamed, mapping = aes(x = age_24_plus, y = highdeg)) +
  geom_point()+
  labs(
      title = "Age of Entry 24+ vs Highest Degree Earned",
      x = "Age of Entry",
      y = "Highest Degree Earned"
  )
```

```
## Warning: Removed 2163 rows containing missing values (geom_point).
```



Explanation:

Although this graph is not too strong in its relationship between age and highest degree earned, there is a weak, negative correlation. As the age increases, less students earn each degree. Ages 24 and above all earn Masters degree similarly, however. I believe this can be due to students who pursue a Masters degree must be highly motivated, so no matter the age of the student, they will most likely finish what they start to achieve their rigorous goal.

2.

Why?

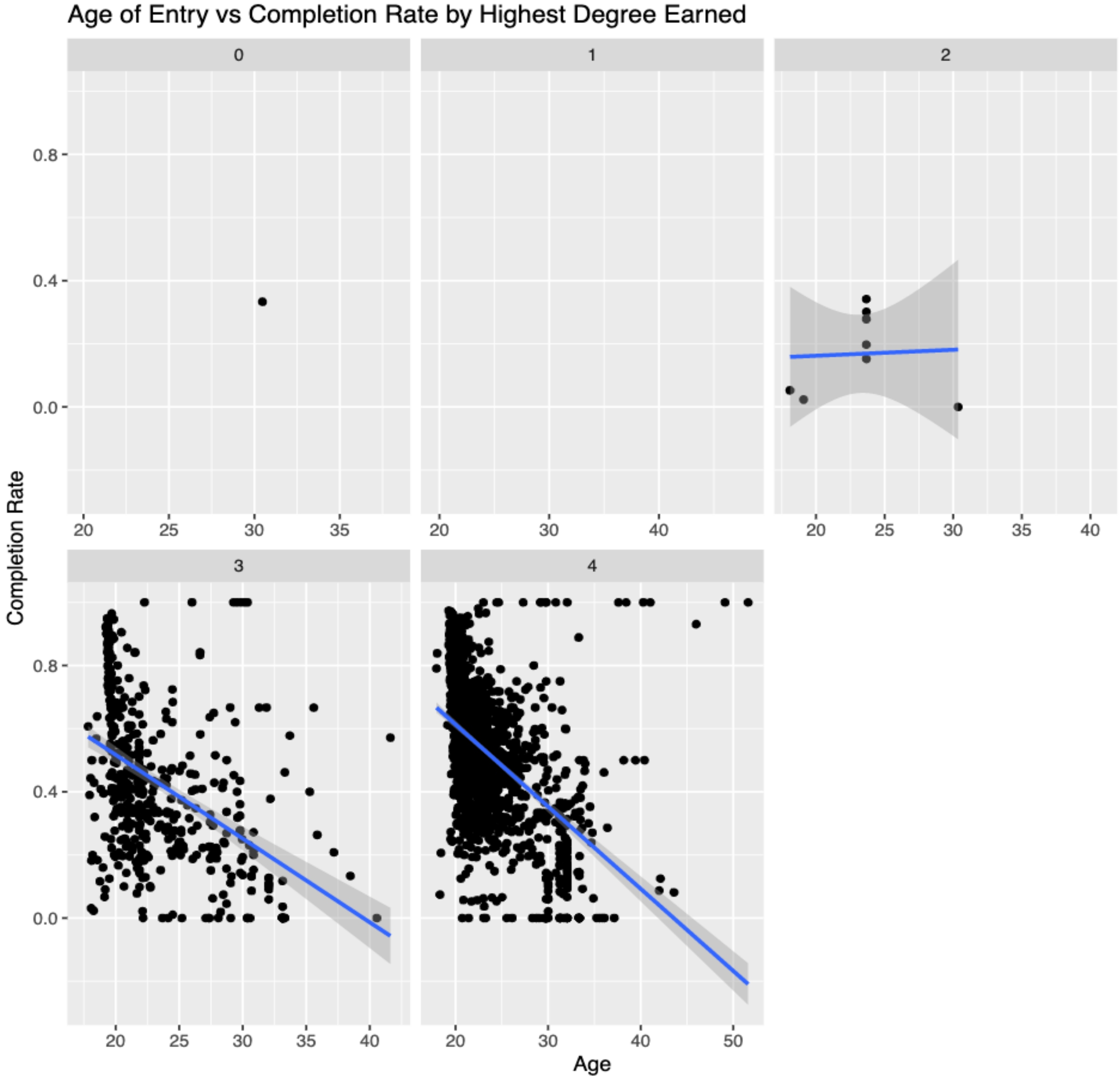
I am creating a scatter plot with the x-value of the age the student entered the university and the y-value of the completion rate. I faceted this over the highest degree each student earned. I want to see if there is a correlation between age and completion of 4 years of school and how high of a degree they earned.

```r
college_renamed %>%
  ggplot() +
    geom_point(mapping = aes(x = age_entry, y = comp_4)) +
    facet_wrap(~ highdeg, scales = "free_x")+
  geom_smooth(mapping = aes(x = age_entry, y = comp_4), method = "lm")+
    labs(
        title = "Age of Entry vs Completion Rate by Highest Degree Earned",
        x = "Age",
        y = "Completion Rate"
    )
```

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

## Warning: Removed 5036 rows containing non-finite values (stat_smooth).

## Warning: Removed 5036 rows containing missing values (geom_point).
```



Explanation:

It is shown that there are barely any or no data points in the first three plots representing Non-degree granting, certificate degree, and associate degrees. However, in the last two plots, it is shown that there are more students that earn a graduate degree if the complete 4 years of schooling, though there are also many students who only earn a bachelors degree. However, it can be seen that there is a correlation between age and the completion rate. First off, there is a significant amount of students that are arounf the ages of 20-25 than any other age, and as the age increases, the number of students seem to decrease. It is also shown that the completion rate is higher for students who are around this 20-25 year range that any studdents that are older. There seems to be a strong negative relationship between completion rate and age of entry.

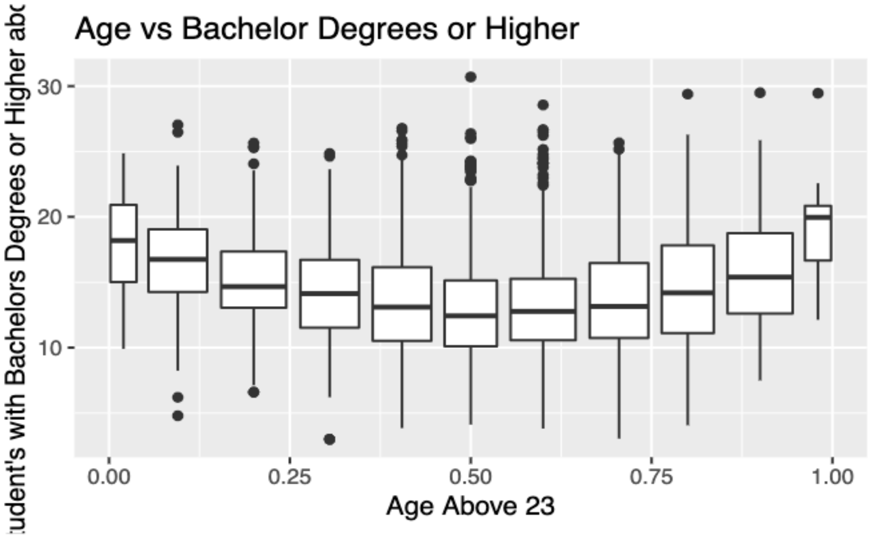
Why?

Now that I have seen the relationships between age of entry, completion rate, and highest degree earned, I want to get a little more specific. I will create a box plot to show the relationship between age of entry of students 23 and above vs the percentage of a bachelor's degree or higher.

```
ggplot(data = college_renamed, mapping = aes(x = age_24_plus, y = pct_25_bach)) +
  geom_boxplot(mapping = aes(group = cut_width(age_24_plus, 0.1))) +
  labs(
    title = "Age vs Bachelor Degrees or Higher",
    x = "Age Above 23",
    y = "Percent of student's with Bachelors Degrees or Higher above the age 25"
  )
```

```
## Warning: Removed 2163 rows containing missing values (stat_boxplot).
```

```
## Warning: Removed 39 rows containing non-finite values (stat_boxplot).
```



Explanation:

This boxplot shows that the younger and older ages in this group seem to be recieving their bachelors degree or higher. However, there are multiple outliers that recieved the degrees in the middle range of the ages. I believe this is because the younger and older students may have more motivation, while the middle-range ages may have other things to be worrying about such as supporting a family, themself, not having time, etc. While the younger and older ages may not have as many responsibilities.

Summary Statistics

```
college_renamed %>%
  group_by(highdeg) %>%
  summarize(
    count = n(),
    mean = mean(age_entry, na.rm = TRUE),
    median = median(age_entry, na.rm = TRUE),
    sd = sd(age_entry, na.rm = TRUE),
    IQR = IQR(age_entry, na.rm = TRUE),
    min = min(age_entry, na.rm = TRUE),
    max = max(age_entry, na.rm = TRUE)
  )
```

highdeg	count	mean	median	sd	IQR	min	max
0	477	28.13279	28.46959	3.857756	4.807848	19.96979	38.20000
1	2259	27.47580	27.07984	2.997906	3.482002	19.72956	48.12000
2	1513	26.01690	25.54912	2.925634	3.689392	18.10345	40.81704
3	762	24.73154	23.99054	4.686037	7.538460	17.42771	41.94805
4	2047	24.21097	22.66604	4.524652	5.333951	17.94444	51.60000

```
college_renamed %>%
  group_by(highdeg) %>%
  summarize(
    count = n()
  )
```

highdeg	count
0	477
1	2259
2	1513
3	762
4	2047

Data Analysis

```
research_model <- lm(highdeg ~ age_entry, data = college_renamed)
```

```
research_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	5.2368142	0.1001023	52.31464	0
age_entry	-0.1171196	0.0038051	-30.77940	0

```
research_model %>%
  glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.1259238	0.1257909	1.226574	947.3716	0	1	-10676.17	21358.34	21378.72	9893.492	6576	6578

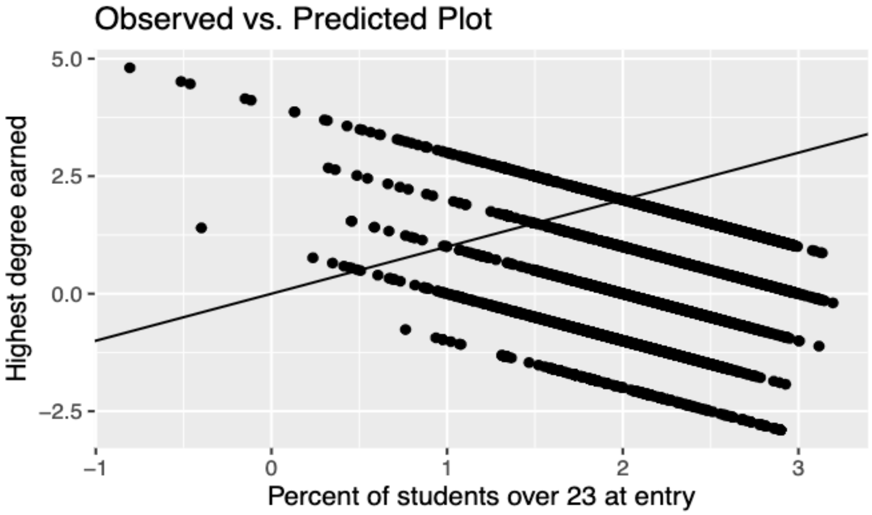
The low r^2 value tells us that this model is not very good at explaining variation because the variables are not closely correlated to each other.

```
college_df <- college_renamed %>%
  add_predictions(research_model) %>%
  add_residuals(research_model)
```

- observed vs. predicted plot

```
college_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = resid)) +
  geom_abline(slope = 1, intercept = 0) +
  labs(
      title = "Price Observed vs Predicted Graph",
      x = "Predicted",
      y = "Observed") +
    labs(x = "Percent of students over 23 at entry", y = "Highest degree earned",
       title = "Observed vs. Predicted Plot")
```

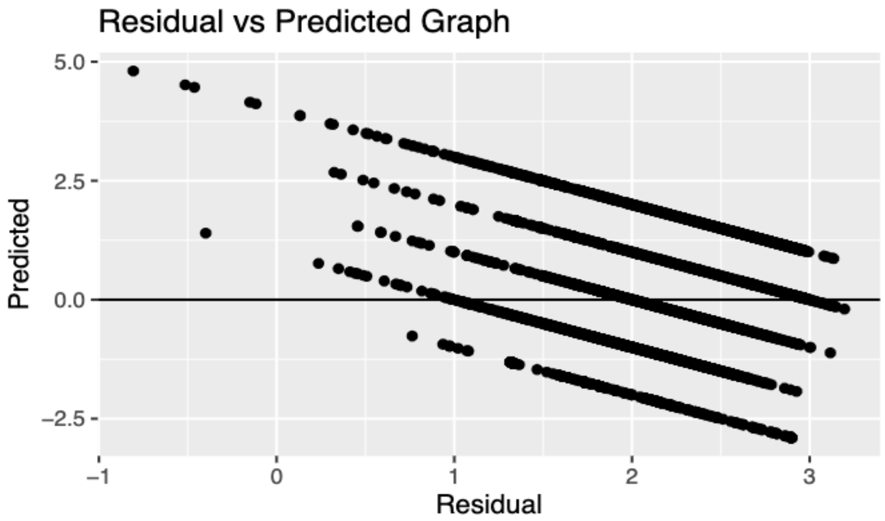
```
## Warning: Removed 480 rows containing missing values (geom_point).
```



- residual vs. predicted plot

```
ggplot(college_df) +
  geom_point(mapping = aes(x = pred, y = resid)) +
  geom_hline( yintercept = 0) +
  labs(
     title = "Residual vs Predicted Graph",
      x = "Residual",
      y = "Predicted"
  )
```

```
## Warning: Removed 480 rows containing missing values (geom_point).
```

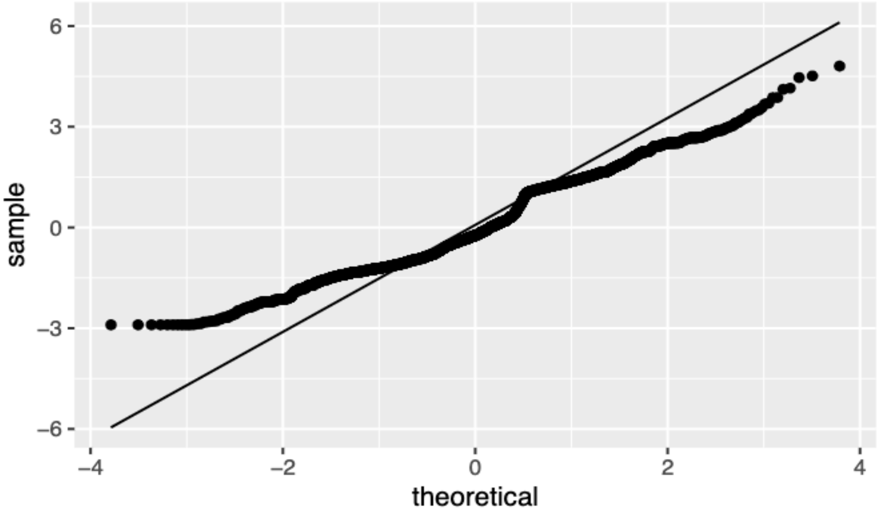


- Q-Q plot

```
college_df %>%
  ggplot() +
  geom_qq(aes(sample = resid)) +
  geom_qq_line(aes(sample = resid))
```

```
## Warning: Removed 480 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 480 rows containing non-finite values (stat_qq_line).
```



Explanation:

In the first plot, there seems to be multiple lines going the opposite of the line in which predicted values equal the observed value. This means that this does not meet the first assumption of linearity. For the second plot, since the data points do not seem to be scattered arounf the horizontal line in which the residuals are equal to 0, the assumption of constant variance is also not met. And for the third plot, the assumption for normal distribution is met as the points do in fact seem to follow the diagonal line.

Conclusion

My plots was not able to meet all of the three assumptions of linearity, however, I believe this is something we should look further into. I would answer my original question of interest by saying that I will need look into other tests in which will prove that there is a relationship between the variables of age_entry and highdeg, however, there is not a linear relationship between them. In my visualization section, it looks as if there is a relationship between my variables, however, my data analysis section show the contrary. There are most definitely confounding variables that I did not take into account when creating my models such as race, financial factors, gender, etc. I believe this is an interesting question to study in the first place because especially since George Mason University is so diverse in accepting students of all races, ethnicities, ages, etc., we should see if there is anything that can be done about any discrimination or disadvantages for students of ages other than the most common ages of students.