Final Project

Benjamin Nguyen

2025-05-05

Introduction

Research Question: "How does the percentage of Pell Grant recipients at a college relate to the median earnings of students 10 years after enrollment?"

Variables:

Response: MEDIAN_EARN_10yr (renamed from MD_EARN_WNE_P10), representing the median earnings of employed graduates not enrolled in school 10 years after entry.

Explanatory: pell_grant_percentage (renamed from PCTPELL), the percentage of undergraduates receiving Pell Grants (need-based federal aid for low-income students).

Rationale: This analysis investigates whether institutions serving higher proportions of Pell Grant recipients show systemic differences in alumni earnings. The results may inform policy discussions about equitable returns on educational investments for low-income students.

Method: A linear regression model will quantify the relationship between these variables while controlling for key assumptions (linearity, homoscedasticity, and normality of residuals).

Preprocessing

Isolating the needed variables from college dataset. Also renaming the variable for readability.

```r
college_reduced <- college %>%
  select(MD_EARN_WNE_P10, PCTPELL,INSTNM)

college_reduced <- college_reduced %>%
  rename(MEDIAN_EARN_10yr = MD_EARN_WNE_P10,
         pell_grant_percentage = PCTPELL
         )
```

Visualization

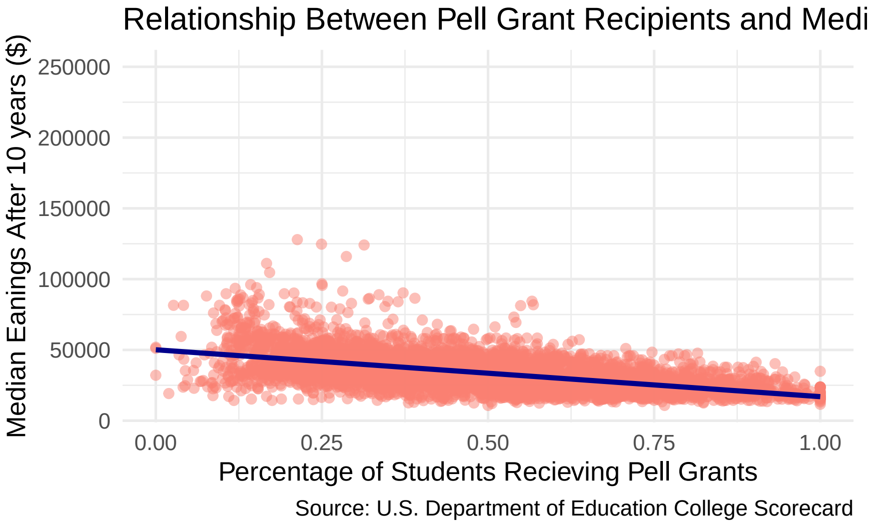
I am creating this graph to showcase the correlation between the median wage 10 years later after graduation and percentage of pell grants.

```
college_reduced %>%
  ggplot(aes(y = MEDIAN_EARN_10yr, x = pell_grant_percentage)) +
  geom_point(color = "salmon", alpha = 0.5) +
   geom_smooth(method = "lm", se = TRUE, color = "darkblue") +
  labs(
    title = "Relationship Between Pell Grant Recipients and Median Earnings",
    x = "Percentage of Students Recieving Pell Grants",
    y = "Median Eanings After 10 years ($)",
    caption = "Source: U.S. Department of Education College Scorecard"
  ) +
  theme_minimal()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’

## Warning: Removed 2243 rows containing non-finite outside the scale range
## (‘stat_smooth()‘).

## Warning: Removed 2243 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



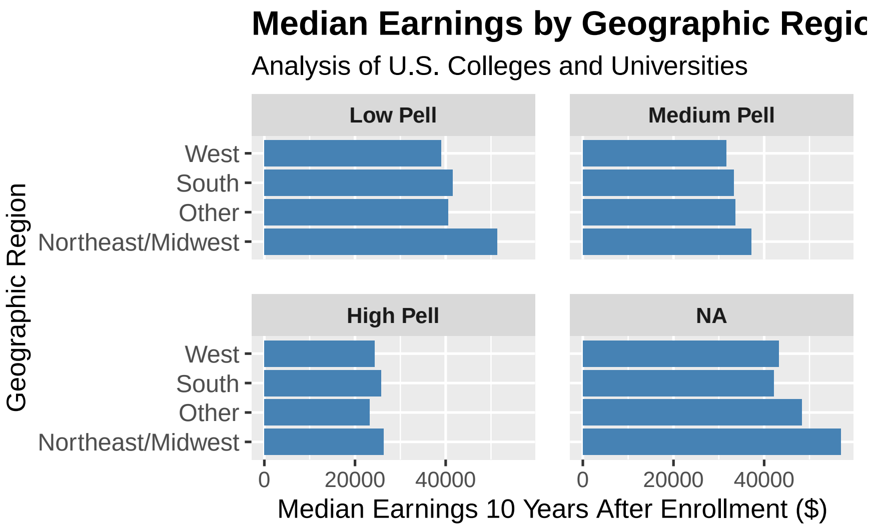
This graph shows a negative correlation between Pell grants percentage and Median Earnings after 10 years. Meaning the higher the percentage of pell grants, the lower the median wage after 10 years.

This next graph showcases the individuals variables of pell grants percentage and median income after 10 years.

```
college %>%
  left_join(college_reduced, by = "INSTNM") %>%
  mutate(region_group = case_when(
           REGION %in% c(1,2) ~ "Northeast/Midwest",
           REGION %in% c(3,4) ~ "South",
           REGION %in% c(5,6) ~ "West",
           TRUE ~ "Other")) %>%
  ggplot(aes(x = region_group, y = MEDIAN_EARN_10yr)) +
  geom_bar(stat = "summary", fun = "mean", fill = "steelblue") +
  facet_wrap(~cut(pell_grant_percentage,
                 breaks = c(0, 0.33, 0.66, 1.0),
                 labels = c("Low Pell", "Medium Pell", "High Pell"))) +
  coord_flip() +
    labs(
    title = "Median Earnings by Geographic Region and Pell Grant Prevalence",
    subtitle = "Analysis of U.S. Colleges and Universities",
    x = "Geographic Region",
    y = "Median Earnings 10 Years After Enrollment ($)",
  ) +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    axis.text.y = element_text(size = 10),
    strip.text = element_text(face = "bold"),
    panel.spacing = unit(1, "lines")
  )
```

```
## Warning in left_join(., college_reduced, by = "INSTNM"): Detected an unexpected many-to-many
## i Row 92 of ‘x‘ matches multiple rows in ‘y‘.
## i Row 92 of ‘y‘ matches multiple rows in ‘x‘.
## i If a many-to-many relationship is expected, set ‘relationship =
##   "many-to-many"‘ to silence this warning.

## Warning: Removed 2032 rows containing non-finite outside the scale range
## (‘stat_summary()‘).
```

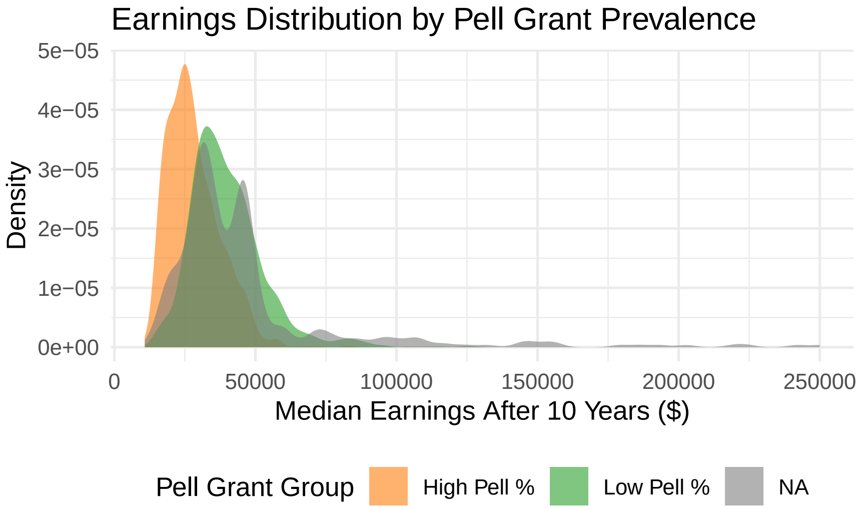


The graphs compares the median earned income after 10 years across the regions. Faceting the 4 category of pell grant percentages. Northeast/Midwest region tends to have a higher median earning after 10 years compared to the other 3 regions. This applies to all 4 categories of the pell grants.

This next graph is to help show the covariance of pell grants percentage and median income after 10 years.

```
college_reduced %>%
  mutate(pell_group = ifelse(pell_grant_percentage > median(pell_grant_percentage,
                                                              na.rm = TRUE),
                             "High Pell %", "Low Pell %")) %>%
  ggplot(aes(x = MEDIAN_EARN_10yr, fill = pell_group)) +
  geom_density(alpha = 0.6, color = NA) +
  scale_fill_manual(values = c("#ff7f0e", "#2ca02c")) +
  labs(title = "Earnings Distribution by Pell Grant Prevalence",
       x = "Median Earnings After 10 Years ($)",
       y = "Density",
       fill = "Pell Grant Group") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 1903 rows containing non-finite outside the scale range
## (‘stat_density()‘).
```



This graph shows the difference in median earnings, the difference in pell grants percentages compared to income after college. Shows that high pell % (higher than median) has a lower average of earnings after 10 years and a low pell % has a higher median earnings after 10 years.

Summary Statistics

```
summary(college$PCTPELL)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.3117  0.4622  0.4819  0.6505  1.0000     767
```

```
summary(college$MD_EARN_WNE_P10)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10800   25400   32700   35199   42000  250000    1903
```

```
var_summary1 <- college_reduced %>%
  filter(!is.na(pell_grant_percentage),
         !is.na(MEDIAN_EARN_10yr)) %>%
  summarize(
    Count = n(),
    # Pell Grant Percentage (Explanatory)
    pell_mean = mean(pell_grant_percentage),
    pell_median = median(pell_grant_percentage),
    ptell_sd = sd(pell_grant_percentage),
    pell_IQR = IQR(pell_grant_percentage),
    pell_min = min(pell_grant_percentage),
    pell_max = max(pell_grant_percentage)
  )
var_summary1
```

Count	pell_mean	pell_median	ptell_sd	pell_IQR	pell_min	pell_max
4815	0.4757343	0.4459	0.2132953	0.32625	0	1

```
var_summary2 <- college_reduced %>%
  filter(!is.na(MEDIAN_EARN_10yr))%>%
  summarise(
    # Median Earnings (Response)
    earn_mean = mean(MEDIAN_EARN_10yr),
    earn_median = median(MEDIAN_EARN_10yr),
    earn_sd = sd(MEDIAN_EARN_10yr),
    earn_IQR = IQR(MEDIAN_EARN_10yr),
    earn_min = min(MEDIAN_EARN_10yr),
    earn_max = max(MEDIAN_EARN_10yr)
  )
var_summary2
```

earn_mean	earn_median	earn_sd	earn_IQR	earn_min	earn_max
35199.4	32700	15839.07	16600	10800	250000

Data Analysis

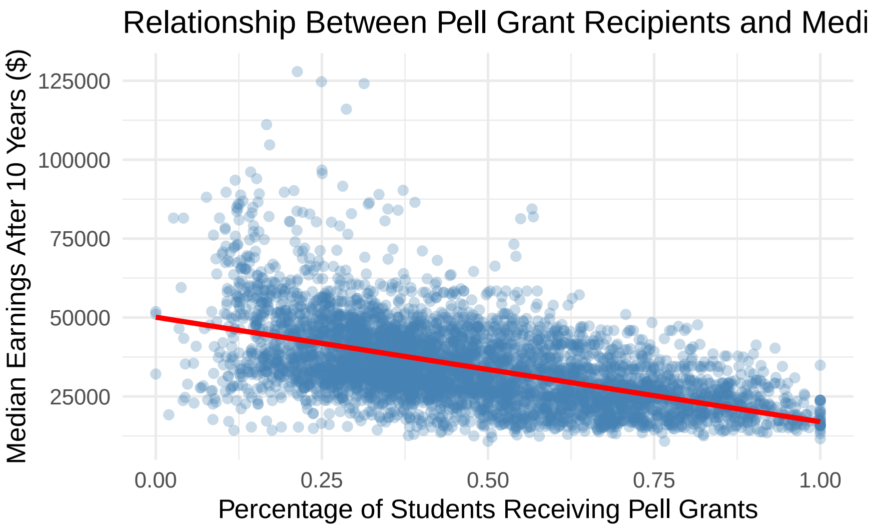
I am gonna filter out the data and take out the missing values, so that the model isn't affected.

```
model_data <- college_reduced %>%
  select(pell_grant_percentage, MEDIAN_EARN_10yr) %>%
  filter(!is.na(pell_grant_percentage), !is.na(MEDIAN_EARN_10yr))
```

Now I am going to graph the variables to see the relationship.

```
ggplot(model_data, aes(x = pell_grant_percentage, y = MEDIAN_EARN_10yr)) +
  geom_point(alpha = 0.3, color = "steelblue") +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Relationship Between Pell Grant Recipients and Median Earnings",
       x = "Percentage of Students Receiving Pell Grants",
       y = "Median Earnings After 10 Years ($)") +
  theme_minimal()
```

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



The red line represents the linear regression line of Median Earnings after 10 years on percentage of students receiving pell grants. The linear regression shows that there is a negative correlation between median earnings and pell grants received. As the percentage of Pell grants received per college goes higher, the median earnings after 10 years of graduation decreases. Meaning that Colleges who have a lower percentage of pell grants recipients tend to have higher median earnings of students after 10 years of graduation.

Now I am currently making the linear model.

```
earned_model <- lm(data = model_data, MEDIAN_EARN_10yr ~ pell_grant_percentage)

glance(earned_model) %>%
  select(r.squared, adj.r.squared, nobs) %>%
  pivot_longer(everything(),
               names_to = "Statistic",
               values_to = "Value") %>%
  mutate(
    Statistic = case_when(
      Statistic == "r.squared" ~ "R-squared",
      Statistic == "adj.r.squared" ~ "Adjusted R-squared",
      Statistic == "nobs" ~ "Number of Observations"
    ),
    Value = round(Value, 4)
  )
```

Statistic	Value
R-squared	0.2952
Adjusted R-squared	0.2950
Number of Observations	4815.0000

This model shows that the r.squared is 0.2952 meaning that 29.5% of variation in median earnings across colleges can be explained by the difference in Pell Grant recipients percentage. Pell grant percentage is a moderately strong predictor but 70.5% of earning variation is explained by other factors, such as: location, majors, institutional resources, and more.

```
tidy(earned_model)
```

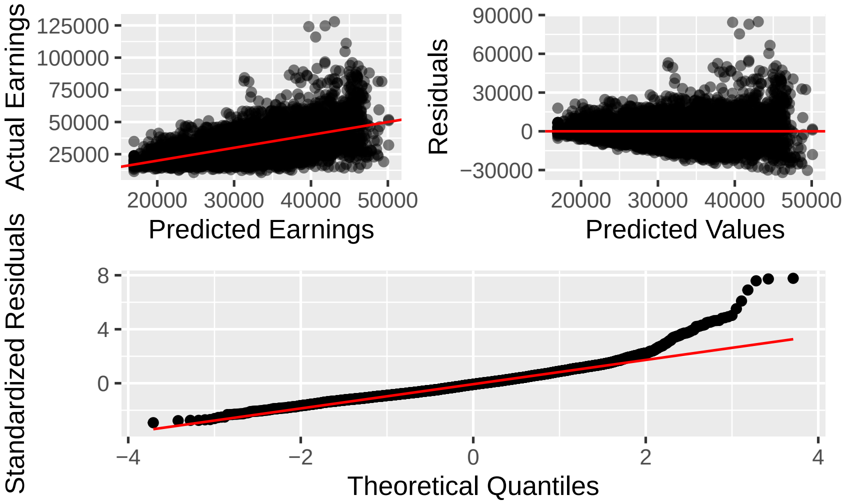
term	estimate	std.error	statistic	p.value
(Intercept)	50103.54	384.6710	130.25038	0
pell_grant_percentage	-33124.22	737.8326	-44.89395	0

```
# Observed vs Predicted
p1 <- ggplot(model_data, aes(x = predict(earned_model), y = MEDIAN_EARN_10yr)) +
  geom_point(alpha = 0.5) +
  geom_abline(slope = 1, color = "red") +
  labs(x = "Predicted Earnings", y = "Actual Earnings")

# Residuals vs Predicted
p2 <- ggplot(earned_model, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Predicted Values", y = "Residuals")
```

```r
# Q-Q Plot
p3 <- ggplot(earned_model, aes(sample = .stdresid)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(x = "Theoretical Quantiles", y = "Standardized Residuals")

# Combine plots
(p1 + p2) / p3
```



For predicted earnings and actual earnings the plot shows a consistent correlation until the prices start to increase. Showing a fan like shape.

For Residuals vs. predicted values the graph starts off good showing random points scattered across zero. As predicted values start to increase the residual grows bigger. Also like a fan shape. Showing heteroscedasticity.

For the q-q plot the graph shows that the points start to wonder as the at the higher values. Indicting more outliers at the higher values.

Conclusion

In this investigation, the data reveals a consistent and statistically significant negative relationship between the percentage of Pell Grant recipients at colleges and their graduates' median earnings ten years post-enrollment. The linear regression model estimates that with each 1% increase in Pell recipients, it corresponds to $33,124 lower median earnings ($p < 0.001$). The pattern also holds across all geographic regions. Three crucial findings support this conclusion: (1) visualizations depict a clear downward trend in earnings as Pell percentages increased, (2) density plots showed distinct earnings distributions between high and low Pell institutions, and (3) the regression model explained 29.5% of earnings variation ($R^2 = 0.295$), suggesting Pell percentage is an dominant but incomplete predictor.

While these results are statistically robust, several factors can be considered. First, the presence of an unusually large coefficient magnitude suggests that there are potential data scaling issues that should be verified. Second, residual diagnostics uncovered heteroscedasticity and non-normality, which indicate the model may overlook important nonlinear relationships or confounding variables like institutional type (public/private) or academic program mix. These limitations advise the estimated effect size should be interpreted carefully until there is further validation. Nevertheless, the findings highlight an important equity challenge in higher education outcomes that deserve deeper investigation with a collection of more comprehensive institutional data.