The association between county-level education rates and unemployment rates in California, Oregon, and Washington

Alison King
BIOST 579

Abstract

This study aims to evaluate whether county-level unemployment rates in the states of California, Washington, and Oregon vary with high school graduation and post-secondary education rates. In order to address this question, we fit two models using multiple linear regression – one without effect modification and one with state as an effect modifier. We found that the first model showed that higher college education rates slightly decreased unemployment across all states, while high school rates did not. However, we also found that state acted as a strong effect modifier in this situation, and within state, increased high school completion rates significantly decreased unemployment rates at the county-level. Within state, post-secondary education enrollment rates did not seem to affect the unemployment rate. The results of this study provide insight into important factors that may decrease unemployment within an area and provide evidence that complete and quality education can promote the wellbeing of society. This study also suggests that inter-state differences vastly change the ways in which education and unemployment interact.

Introduction

Study Objective

The primary purpose of this study is to investigate whether county-level unemployment rates in Washington, California, and Oregon differ based on the county's high school completion and college enrollment rates. The question stems from Spence's job market signaling model, which states that potential employees send a signal about their ability level to the employer by acquiring education credentials. Subsequently, the employer believes that the credential has informational value since it is difficult to obtain and is perceived to be positively correlated with greater ability. Thus, the credential theoretically enables the employer to reliably distinguish low and high ability candidates (Spence).

We also have evidence to suspect state-specific effect modification and will test to see if the association between unemployment and education level varies between the three states. It has previously been shown that one way to effectively reduce unemployment is by funding education and investing in health services (Pirim, et al.). Thus, depending on statewide funding and policy differences between the three states of interest, we suspect that the association between education and unemployment may vary by state. We also see that the diversification of industry within an area can result in lower unemployment rates, so the heterogeneity of states' industries may be another reason that the relationship differs (Izraeli & Murphy).

Data and Variables

The dataset for this study is from the 2015-2019 American Community Survey (ACS) 5-Year Estimates. It contains rates and averages based on public health data collected during this five

year span for each county in the nation. The 5-year estimates from the ACS are considered "period" estimates derived from a sample collected over a period of time, and these multiyear estimates are advantageous for this project because they include all counties regardless of size and provide increased statistical reliability of the data for less populated areas (U.S. Census Bureau). For this project, we are less concerned about the currency of the data as we are about the precision of the estimates, so we choose to use the 5-year data instead of the 1-year.

The chosen variables of interest are county and state name, percentage of adults age 25 and over with a high school diploma or equivalent, percentage of population ages 16 and up that are unemployed and looking for work, and percentage of adults age 25-44 with some post-secondary education.

Methods

Data Preprocessing

The states of interest for this study are California, Oregon, and Washington, so our data has been filtered to contain only the counties in these states. Among these states, every county is represented and there are no missing values for any of our covariates or response variables.

Descriptive Analysis

Numerical summaries of county-level rates for unemployment, high school, and college for each state are reported in the form of means and standard deviations in Table 1. Additionally, Figure 1A is a plot of all data points with multiple linear regression using high school completion rate as the predictor and unemployment rate as the response, stratified by state. Figure 1B is the same plot but using the rate of adults with some college education as the predictor.

Statistical Analysis

Primary

The primary analysis consistent with the primary objective to assess the association of high school completion and college enrollment rates with unemployment rates at the county level is:

a) Multiple linear regression for unemployment rate as the response variable with high school completion rate and college enrollment rate as predictors

Estimates are assessed for significance with Wald p-values at alpha level 0.05 and 95% confidence intervals using robust standard errors are constructed for each estimate.

Secondary

The secondary analyses address state as an effect modifier for the association found in the primary regression model and include:

a) Multiple linear regression for unemployment rate as the response variable with high school completion rate and college enrollment rate as predictors, with state and interaction terms between state and the two original covariates added to the model
b) A Rao test comparing the original regression model to the updated one that includes state and interaction terms, and Wald tests to understand how state interacts with the two original covariates

As above, model estimates are assessed by p-value at alpha level 0.05 and robust 95% confidence intervals, and the p-values of the Rao and Wald tests for state as an effect modifier are reported.

Model Checks

Since we suspect there may be multicollinearity between college enrollment and high school completion as predictors, we evaluate correlation between the two and also calculate the variance inflation factor (VIF) based on our primary analysis model. Additionally, we plot residuals from our model to ensure that we are receiving an appropriate fit.

Statistical Software

All figures, tables, and analysis are created and performed in R.

Results

Descriptive Results

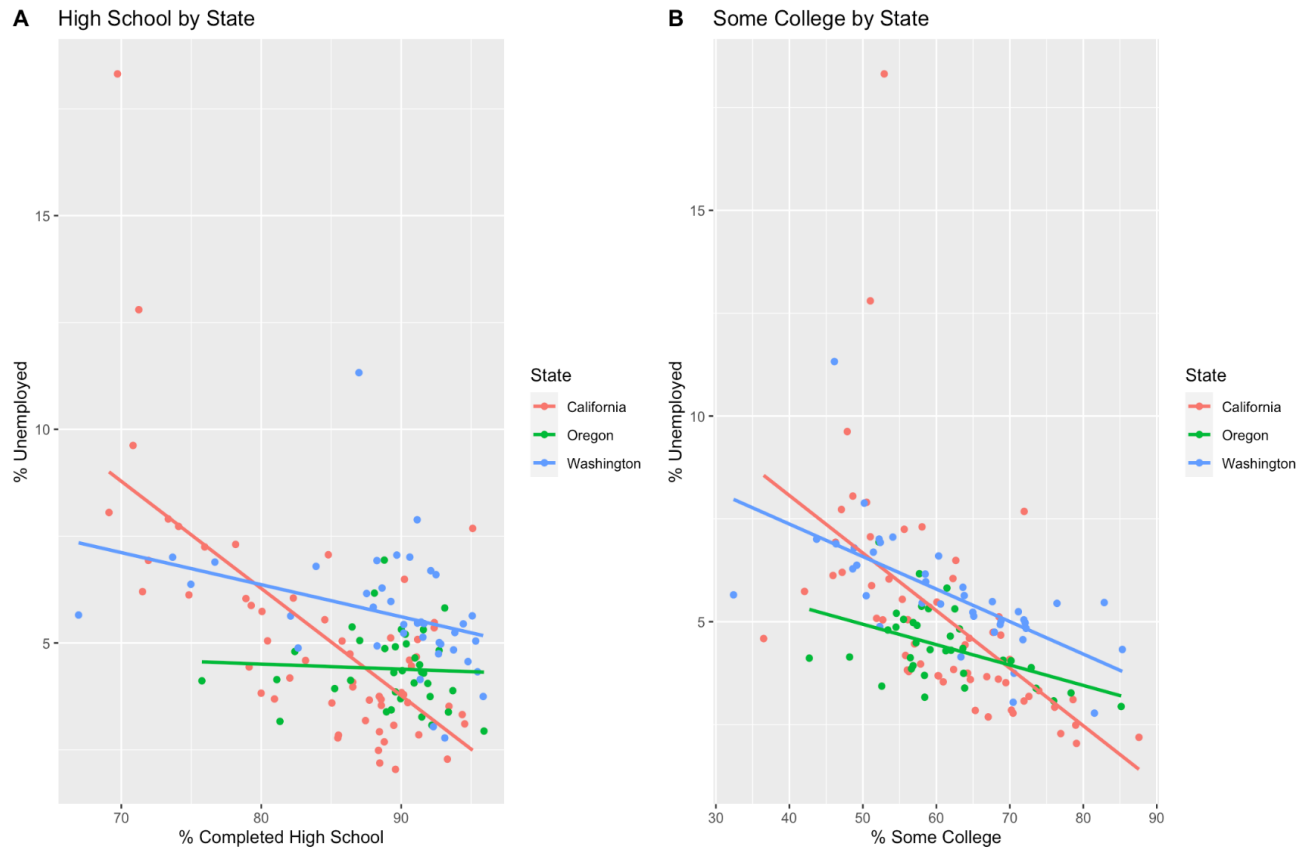
The numerical summary of the dataset is presented in Table 1 containing means and standard deviations for our two primary covariates and response variable.

Table 1: Numerical Summaries of Mean and Standard Deviation of Model Variables

State	n	Mean % High School	SD HS	Mean % College	SD C	Mean % Unemployed	SD U
California	58	84.69	7.13	61.21	10.56	5.10	2.68
Oregon	36	89.26	4.01	60.95	8.54	4.40	0.90
Washington	39	89.06	6.48	61.39	11.88	5.68	1.44

Figures 1A and1 B below demonstrate the effect modification that state has on our regression model. Here, we can visually see that the relationships for high school completion rate and some college education rate with unemployment rate within counties varies by state.

Figure 1: (A) Scatterplot of county-level high school completion rate vs. unemployment rate, stratified by state and overlaid with regression lines, (B) Scatterplot of county-level percentage of adults with some post-secondary education vs. unemployment rate, stratified by state and overlaid with regression lines.



Statistical Results

Primary Analysis

Estimates and 95% confidence intervals for our regression model for the primary analysis are presented in Table 2. We estimate that the mean county-level unemployment rate decreases by 0.082% (-0.20, 0.037) for every 1% increase in high school graduation rate, controlling for college enrollment; however, we do not find this to be significant (p=0.17). Also, the estimated mean unemployment rate decreases by 0.065% (-0.11, -0.018) for every 1% increase in college enrollment rate, controlling for high school completion, which we find to be a significant association (p=0.007). The fitted model is:

$$\widehat{Unemployment} = 16.25 - 0.082 * HighSchool - 0.065 * SomeCollege$$

where HighSchool is the county-wide high school completion rate and SomeCollege is the county-wide percentage of people with some post-secondary education.

Table 2: Estimates (95% robust confidence intervals) and p-values for regression coefficients from our primary model.

	Estimate (95% CI)	P-value
Intercept	16.25 (7.82, 24.68)	0.0002
HighSchool	-0.082 (-0.20, 0.037)	0.17

SomeCollege	-0.065 (-0.11, -0.018)	0.007

Secondary Analysis

For our secondary analysis, we include state as an additional predictor in our model as well as interaction terms between:

- State and high school completion rate, and
- State and college enrollment rate

When we fit this multiple linear regression model with state as an effect modifier, we find significant associations with unemployment rate for state, high school, and the interaction between the two. The addition of state as an effect modifier removes the significant association between college enrollment and unemployment that we see in the primary model. Estimates and confidence intervals for our regression model for the secondary analysis are presented below in Table 3. The fitted model is:

$$\widehat{Unemployment} = 25.23 - 27.00 * Oregon - 20.12 * Washington - 0.20 * HighSchool$$
$$-0.05 * SomeCollege + 0.34 * Oregon * HighSchool + 0.29 * Washington * HighSchool$$
$$-0.05 * Oregon * SomeCollege - 0.07 * Washington * SomeCollege$$

where HighSchool is the county-wide high school completion rate, SomeCollege is the county-wide percentage of people with some post-secondary education, and Oregon and Washington are binary variables equal to 1 for counties in that state. If Oregon and Washington are both equal to 0, then that county is in California.

Table 3: Estimates (95% confidence intervals) and p-values for regression coefficients from our secondary model.

	Estimate (95% CI)	P-value
Intercept	25.23 (13.39, 37.07)	<0.00005
Oregon	-27.00 (-39.72, -14.28)	0.0001
Washington	-20.12 (-33.16, -7.07)	0.003
HighSchool	-0.21 (-0.37, -0.04)	0.018
SomeCollege	-0.045 (-0.11, 0.01)	0.14
Oregon*HighSchool	0.34 (0,15, 0.52)	0.0004
Washington*HighSchool	0.29 (0.097, 0.48)	0.0035
Oregon*SomeCollege	-0.05 (-0.12, 0.023)	0.18
Washington*SomeCollege	-0.04 (-0.16, 0.019)	0.12

The Rao test for state returns a p-value less than 0.0000001, providing evidence that the model with effect modification is the better fit for our data. When we run Wald tests for the interaction term between state and high school completion, we see a significant effect (p=0.002); however, the Wald test for the interaction between state and some college was not (p=0.26). Therefore, we can conclude that within the states of Oregon, Washington, and California, the effect that high school completion rate has on county-wide unemployment rates is dependent upon the state in which the county is located.

Model Checks

Correlation between county-level high school completion rate and college enrollment rate is 0.66 and the VIF is 1.8, so we can assume that multicollinearity is not introducing bias into our model. The residual plot in Figure 2 shows that residuals appear to be randomly distributed around 0, suggesting that our multiple linear regression model is a good fit for our data.

Figure 2: A residual plot for our secondary model, showing most residuals randomly distributed around 0.

Discussion

This study found that increased education rates tended to decrease unemployment rates within a county in the states of California, Washington, and Oregon. It also found that the state that a county is in vastly affected the relationship between unemployment rates and education rates. Without adjusting for state as an effect modifier, counties that had higher percentages of adults with some post-secondary education saw lower unemployment rates overall, while the percentage of adults with a high school diploma or equivalent did not significantly affect unemployment. When we adjusted for differences between states, we saw a large difference in this relationship, with state and high school completion having significant effects on unemployment. The amount that the unemployment rate decreased with high school completion

rate also differed significantly between states. Once we adjusted for state as well, we found that college enrollment did not play a significant role in affecting unemployment rates.

This study was based on survey and census data and thus was potentially impacted by problems such as inaccuracy and misrepresentation. To mitigate this, we used robust methods in our models and tests. We also opted to use the five-year estimates from the ACS which provided more precision in our census estimates and allowed us to include smaller counties within our states. It is also important to note that there is potential spatial correlation between counties which could be affecting our results. For the purpose of this study, we decided to omit this spatial correlation; however, future studies may address the impact this could have on the relationship between education rates and unemployment.

Sources

Izraeli, O., Murphy, K. The effect of industrial diversity on state unemployment rate and per capita income. Ann Reg Sci 37, 1–14 (2003). https://doi.org/10.1007/s001680200100

Pirim, Zafer; Owings, William A.; and Kaplan, Leslie S., "The Long-Term Impact of Educational and Health Spending on Unemployment Rates" (2014). Educational Foundations & Leadership Faculty Publications. 16. https://digitalcommons.odu.edu/efl_fac_pubs/16

Spence, M. (1973). Job Market Signaling. The Quarterly Journal of Economics, 87(3), 355–374. https://doi.org/10.2307/1882010

U.S. Census Bureau, “Understanding and Using ACS Single-Year and Multiyear Estimates.” American Community Survey.
https://www.census.gov/programs-surveys/acs/library/handbooks/general.html

2021 County Health Rankings National Data.
https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation

Appendix: Figures and Tables

Table 1: Numerical summaries of mean and standard deviation of model variables.

Table 1: Numerical Summaries of Mean and Standard Deviation of Model Variables

State	n	Mean % High School	SD HS	Mean % College	SD C	Mean % Unemployed	SD U
California	58	84.69	7.13	61.21	10.56	5.10	2.68
Oregon	36	89.26	4.01	60.95	8.54	4.40	0.90
Washington	39	89.06	6.48	61.39	11.88	5.68	1.44

Table 2: Estimates (95% confidence intervals) and p-values for regression coefficients from our primary model.

	Estimate (95% CI)	P-value
Intercept	16.25 (7.82, 24.68)	0.0002
HighSchool	-0.082 (-0.20, 0.037)	0.17
SomeCollege	-0.065 (-0.11, -0.018)	0.007

Table 3: Estimates (95% confidence intervals) and p-values for regression coefficients from our secondary model.

	Estimate (95% CI)	P-value
Intercept	25.23 (13.39, 37.07)	<0.00005
Oregon	-27.00 (-39.72, -14.28)	0.0001
Washington	-20.12 (-33.16, -7.07)	0.003
HighSchool	-0.21 (-0.37, -0.04)	0.018
SomeCollege	-0.045 (-0.11, 0.01)	0.14
Oregon*HighSchool	0.34 (0,15, 0.52)	0.0004
Washington*HighSchool	0.29 (0.097, 0.48)	0.0035
Oregon*SomeCollege	-0.05 (-0.12, 0.023)	0.18
Washington*SomeCollege	-0.04 (-0.16, 0.019)	0.12

Figure 1: (A) Scatterplot of county-level high school completion rate vs. unemployment rate, stratified by state and overlaid with regression lines, (B) Scatterplot of county-level percentage of adults with some post-secondary education vs. unemployment rate, stratified by state and overlaid with regression lines.

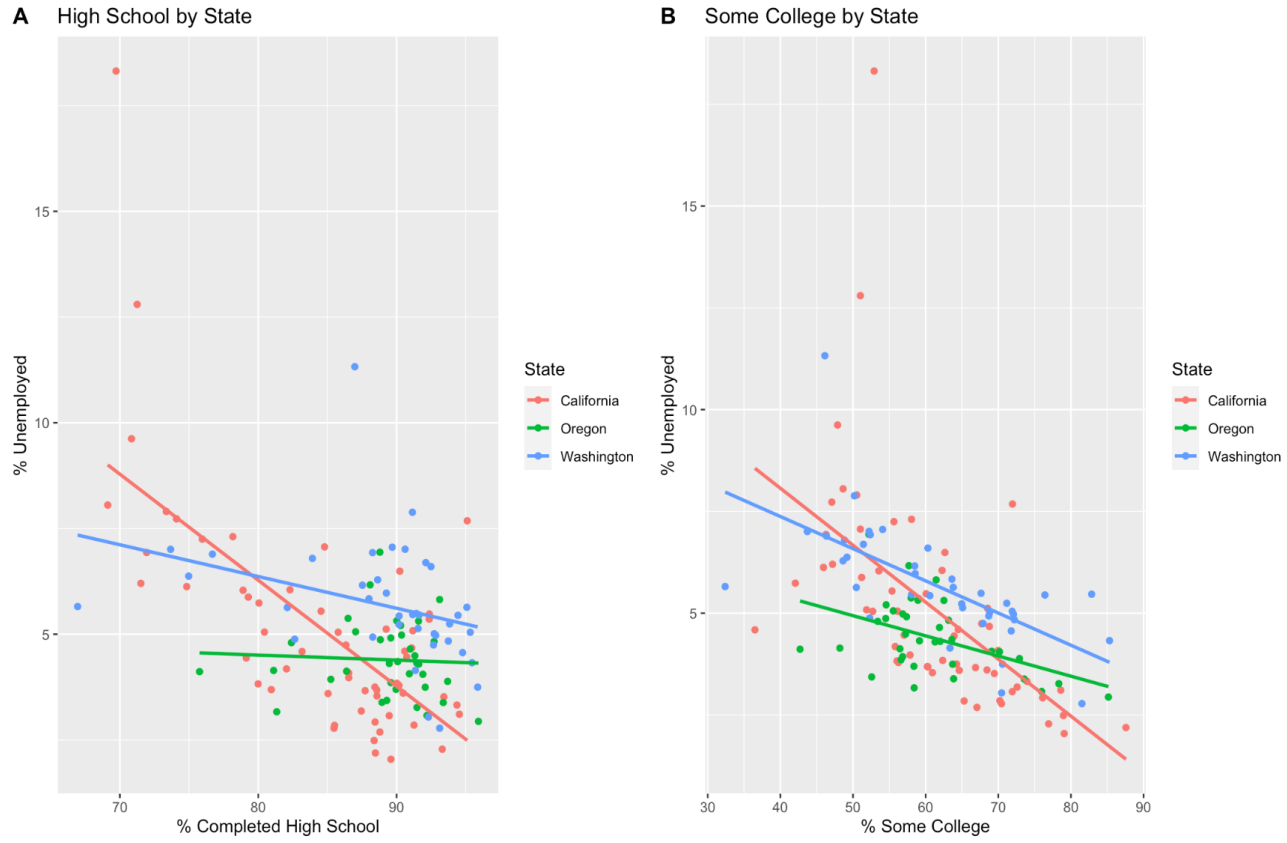
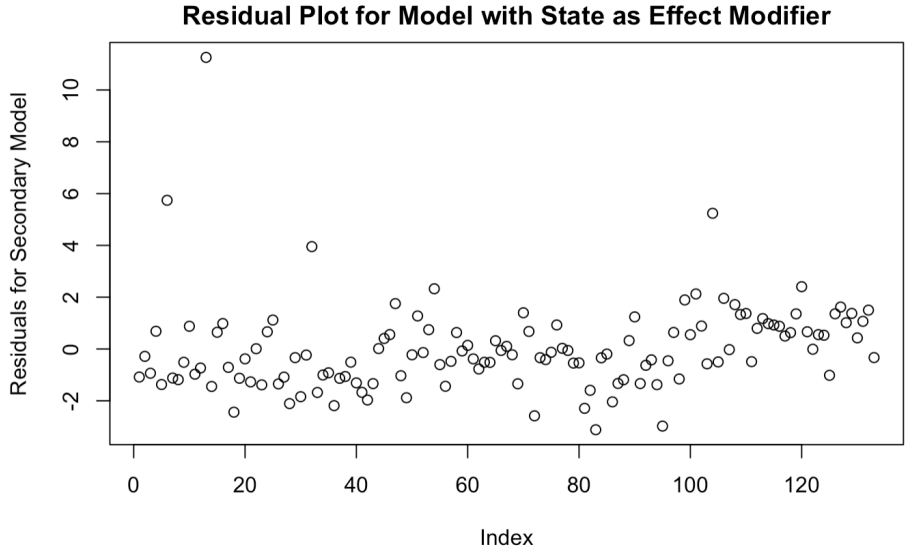


Figure 2: A residual plot for our secondary model, showing most residuals randomly distributed around 0.



Appendix: R Code

```
library(readxl)
county_data <- read_excel("~/Downloads/2021 County Health Rankings Data - v1.xlsx",
    sheet = "Ranked Measure Data", skip=1)
library(dplyr)
county_data <- county_data %>% filter(!is.na(County)) %>%
  filter(State=="California"|State=="Washington"|State=="Oregon")
library(ggplot2)

# check for multicollinearity
library(tidyverse)
county_selected <- county_data %>% select(`% Completed High School`, `% Some College`)
# correlation matrix
cor(county_selected)

## Summary Statistics for % Completed High School
county_data %>%
  group_by(State) %>%
  summarise( ave_highschool = round(mean(`% Completed High School`, na.rm=TRUE),3),
sd_value = sd(`% Completed High School`, na.rm=TRUE),
            n_obs = sum(!is.na(`% Completed High School`)),
            n_missing = sum(is.na(`% Completed High School`)),
            prop_missing = n_missing/n_obs)

## Summary Statistics for % Some College
county_data %>%
  group_by(State) %>%
  summarise( ave_college = round(mean(`% Some College`, na.rm=TRUE),3), sd_value =
sd(`% Some College`, na.rm=TRUE),
            n_obs = sum(!is.na(`% Some College`)),
            n_missing = sum(is.na(`% Some College`)),
            prop_missing = n_missing/n_obs)

## Summary Statistics for % Unemployed
county_data %>%
  group_by(State) %>%
  summarise( ave_unemployed = round(mean(`% Unemployed`, na.rm=TRUE),3), sd_value =
sd(`% Unemployed`, na.rm=TRUE),
            n_obs = sum(!is.na(`% Unemployed`)),
            n_missing = sum(is.na(`% Unemployed`)),
            prop_missing = n_missing/n_obs)

## Summary Statistics for All 3 Variables
knitr::kable(county_data %>%
  group_by(State) %>%
  summarise(n_obs = sum(!is.na(`% Completed High School`)), ave_highs = round(mean(`%
Completed High School`, na.rm=TRUE),3), sd_high = sd(`% Completed High School`,
na.rm=TRUE), ave_college = round(mean(`% Some College`, na.rm=TRUE),3), sd_college =
```

```
sd(`% Some College`, na.rm=TRUE), ave_unemployed = round(mean(`% Unemployed`,
na.rm=TRUE),3), sd_unemployed = sd(`% Unemployed`, na.rm=TRUE)
            ) , col.names = c("State", "n", "Mean % High School", "SD HS", "Mean % College", "SD
C", "Mean % Unemployed", "SD U"), digits=2, caption="Numerical Summaries of Mean and
Standard Deviation of Model Variables")

## Unemployment Histogram
ggplot(county_data, aes(`% Unemployed`)) +
  geom_histogram(binwidth = 1, alpha=0.8, fill="cornflowerblue") + labs(title="Distribution of
County-Level Unemployment Rates", y = "Count")

## Scatterplot with Linear Regression by State
plot1 <- ggplot(county_data, aes(y=`% Unemployed`, x=`% Completed High School`)) +
geom_point() + geom_smooth(method=lm, se=F) + labs(title="High School")

plot2 <- ggplot(county_data, aes(y=`% Unemployed`, x=`% Completed High School`,
color=`State`)) + geom_point() + geom_smooth(method=lm, se=F) + labs(title="High School by
State")

plot3 <- ggplot(county_data, aes(y=`% Unemployed`, x=`% Some College`)) + geom_point() +
geom_smooth(method=lm, se=F) + labs(title=" Some College")

plot4 <- ggplot(county_data, aes(y=`% Unemployed`, x=`% Some College`, color=`State`)) +
geom_point() + geom_smooth(method=lm, se=F) + labs(title="Some College by State")

library(cowplot)
plot_grid(plot1, plot3, labels = "AUTO")
plot_grid(plot2, plot4, labels = "AUTO")

# linear regression with state
library(tidyverse)
library(broom)
model1 <- lm(`% Unemployed` ~ State*(`% Completed High School` + `% Some College`),
data=county_data)

library(rigr)
model1 <- regress("mean", `% Unemployed` ~ State*(`% Completed High School` + `% Some
College`), data=county_data, robustSE=T)


# without state
model0 <- lm(`% Unemployed` ~ `% Completed High School` + `% Some College`,
data=county_data)
model0 %>% tidy() %>%
  knitr::kable(caption = "Coefficient-Level Estimates for Primary Model")

model0 <- regress("mean", `% Unemployed` ~ `% Completed High School` + `% Some
College`,  robustSE=T, data=county_data)


# with state
```

```
model1 %>% tidy() %>%
  knitr::kable(caption = "Coefficient-Level Estimates for Model")


# vif test for multicollinearity
car::vif(model0) # 1.8 is relatively low, fine to leave both predictors in the model

# plot residuals
plot(model0$residuals, ylab="Residuals for Secondary Model", main="Residual Plot for Model
with State as Effect Modifier")
plot(model1$residuals)

# LRT
my.lrt <- anova(model0, model1, test="Rao")
print(my.lrt)
```