Andrew Plum

**ECON 453**

Fall 2023

Problem Set 1 – 40 points

Submit by end of day Monday September 25th

Please download the gretl session file “PS1 Session” from Canvas. This is a dataset that contains information at the state level regarding economic outcomes, demographic information, and public health issues related to the pandemic. Note: I have already deleted Washington DC from this dataset, so you do not need to worry about it influencing our estimates (you’re welcome). Please let me know if you have any questions about the dataset.

You will submit one document (Word or PDF) for this problem set. Please copy/paste the relevant regression results or graphs into your document, then add your discussions.

1. Let’s begin by looking at the relationship between the state’s economy heading into the pandemic and the severity of the pandemic in that state. The **CovidDeathRateTotal** variable is a measure of cumulative deaths through the end of last week (per 100,000 people).
   1. (4 points) Run a simple linear regression using the COVID Death Rate Total(deaths per 100,000 people) as the dependent variable and GDP per capita (2019, representing the economy before the pandemic) as the regressor.
      1. Report/copy the regression results.

A screenshot of a computer

Description automatically generated

* + 1. What is your predicted COVID death rate for Idaho, and what is the residual for Idaho? What does the residual tell us?

Predicted COVID death rate for Idaho = 518.056 – 3.24791(40.10519132) = 387.79794806

Residual for Idaho: -78.80

The residual here tells us the difference between the predicted value and the observed value for Idaho.

* + 1. Interpret the coefficient (numerically), discuss the statistical significance and the overall fit of our model. Overall, what is this telling us and does this result match your expectations?

The coefficient is saying for every $1000 increase of GDP per capita in a particular state, the model predicts the deaths from COVID per every 100,000 people goes down roughly 3.25 people. The regressor GDP per capita from 2019 is statistically significant although the R^2 for the model is low at 0.147156. This result was not what I was expecting because I believe places with higher GDP per capita tend to be more densely populated and disease spreads better in more densely populated places. An explanation why the regression equation says this, might be because states with more resources per person can then dedicate those resources to saving their people’s lives. The low R^2 suggests that there might be other factors that explain why the death rate of COVID goes down as GDP per capita goes up.

* 1. (4 points) Is there a non-linear relationship between GDP per capita and the COVID death rate in a state? Add a quadratic term to your model.
     1. Report/copy your regression results.

A screenshot of a computer

Description automatically generated

* + 1. How much does your model predict the COVID death rate will differ between a state that had a GDP per capita of 50,000 and one that had GDP per capita of $60,000?

Predicted COVID death rate when GDP per capita is $50,000 = 1108.74 – 25.1744(50) + 0.196610(50)^2 =

= 1108.74 – 1258.72 + 491.525 = 341.545

Predicted COVID death rate when GDP per capita is $60,000 = 1108.74 – 25.1744(60) + 0.196610(60)^2 =

= 1108.74 – 1510.464 + 707.796 = 306.072

COVID death rate difference between states

with GDP per capita’s of $60,000 and $50,000 = 306.072 – 341.545 = -35.473

* + 1. Discuss what this model tells us about the relationship between a state’s economy and the severity of the pandemic, whether this makes sense to you, and whether it was a good idea to use a non-linear model here.

Because the model with the quadratic term has a higher R^2 and adjusted R^2 than the model with only the linear term, this indicates that the relationship between the state’s economy and the severity of the pandemic is better represented by a quadratic equation rather than a linear equation because there is some non-linearity to the relationship. Like the linear model, this quadratic model is saying that the COVID death rate decreases as GDP per capita increases, however this happens at a diminishing rate because the squared term is positive. I think it would be good to use the non-linear model here because the adjusted R^2 is higher for the quadratic model than for the linear model. However, the quadratic model’s regressors are less statistically significant which might be problematic.

* 1. (5 points) Let’s account for regional variation in our model. Start with the model from question 1a (linear model between GDP and Death rate total), then add a series of dummy variables that allows us to test if the death rate varies by region.
     1. Report/copy your regression results.

A screenshot of a computer

Description automatically generated

I didn’t know if I supposed to leave a certain region out so I chose to leave out the West so the coefficients of the region variables would be positive.

* + 1. What does your model predict the COVID Death rate is for a state in the Northeast that had a GDP per capita of 50,000 in 2019?

COVID death rate of state in the Northeast @ GDP per capita of $50,000 in 2019 =

= 414.177 – 2.26084(50) + 41.1619(1) = 342.2969

* + 1. Summarize what we learn about the differences across regions in terms of the death rate.

The South has the highest COVID death rate, and the West has the lowest COVID death rate. Also, the South has statistical significance when the West is left out of the model.

* + 1. What has happened to the estimate of the relationship between GDP per capita and the COVID Death Rate (as compared to the model in 1a)? Explain what this tells us.

The effect the coefficient of GDP per capita has on the COVID death rate in this model has decreased when compared to the same coefficient in the model for 1a. There is also less statistical significance for the regressor here than there is in the previous model. What this means is that some of the variation in COVID death rates that was attributed to GDP per capita likely can be attributed with how the regions vary.

* + 1. Based on your results, what should we do to improve/clean up this model measuring the relationship between COVID death rates and regions?

One thing that I noticed in this new model was that the difference between R^2 and adjusted R^2 was lower. This indicates some of the regional variables aren’t adding meaningful value to the model. The variables Northeast and Midwest could be omitted from the model which would lower the p-values of the other variables which would mean the model is better.

* 1. (5 points) Create your own model explaining/predicting the COVID Death Rate across states. Your model should use the death rate as the dependent variable and should include some control for regional differences. Beyond that, use your intuition, heart, and imagination to create a model that includes at least two more regressors.
     1. Explain your reasoning for constructing the model in this way. What are your hypotheses (what are you expecting to find)?

The variables I chose to include in my model are variables I that I thought might impact the COVID death rate. The variables I chose were the dummy variable for if the state was in the West, if the governor of the state is a Republican, percent vaccinated against COVID in the state, the obesity rate of the state, and poverty rate of the state. I chose the West over the South as my regional dummy variable because although both stood out as regions, I know the South has the highest obesity and poverty rates in the country, so I wanted to avoid collinearity with those variables. I expect negative coefficients for if the state is in the West, percent vaccinated against COVID and positive coefficients for the obesity rate of the state and the poverty rate of the state. I don’t know what to expect for if the governor is a Republican although I lean towards saying it would be positive.

* + 1. Report/copy your regression results.

A screenshot of a computer

Description automatically generated

* + 1. Summarize the findings from your model. You should address significance, explanatory power, the sign and magnitude of coefficients, and whether the results matched your expectations.

West, Percent vaccinated, and the governor being a republican all have negative coefficients. Obesity and poverty all have positive coefficients. I was wrong about republican governors having a negative coefficient. I can see this making sense though because states that have democrats as governors tend to be more densely populated which probably translates into a higher COVID death rate. Another thing that I kind of found surprising was that the poverty variable has very high statistical significance while the other variables except for West have very low statical significance. The regressors with low statistical significance should be taken with a grain of salt. Percent vaccinated also had a very tiny coefficient. The coefficient of poverty was also higher than I thought it would be. I thought it would have around the same magnitude as the coefficient for obesity. The percentages for obesity are higher than the percentages for poverty in the data set so this probably affects the difference in the coefficients between obesity and poverty. The low statistical significance of many of the regressors seems to be problematic. I would have thought they would have been higher for 5 variables.

* + 1. What should be done to improve your model going forward?

To improve the model, we’ll want lower p-values for the regressors in it. Because there are high p-values in our model, there might be issues of collinearity between variables. I could see if removing any of the variables would solve the issue. But I have a feeling that obesity and poverty might be correlated so I think removing poverty to check and see if the statical significance of obesity goes up would be a good thing to do. Because percent vaccinated had a surprisingly low magnitude, I would want to pay attention to if changing the variables drastically affects it although that might not happen because most of the US population is vaccinated today, and I don’t know if it varies by state as much as it used. I would also want to see if changing the region greatly changes the overall results. I could restrict the sample by region to see how that affects it.

1. Let’s take a look at the role of policy in state economic outcomes during the pandemic. The variable **SahoDays** lists the number of days during the early part of the pandemic (March, April, and May of 2020) that the state had a stay-at-home order in place.
   1. (3 points) Run a regression using **Unemployment2020** as the dependent variable and **SahoDays** as the independent variable.
      1. Report/copy your regression results.

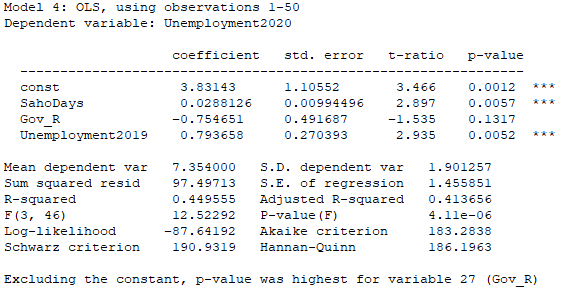
A screenshot of a computer

Description automatically generated

* + 1. Provide a numeric interpretation of the estimated coefficient. Do the results match your expectations? Explain.

The linear regression is saying that for every day there was a stay-at-home order at the beginning of the pandemic, the 2020 unemployment rate in the state increased by 0.0421957 percent. The p-value is very low for Saho Days, and there is a decently high R^2 for the regressor as well. I would have thought that the stay-at-home orders for each day would have had a greater effect on unemployment because people are staying at home, which is less time spent engaging with the economy. I know it adds up, but I still thought it would be higher.

* 1. (4 points) Next, add two additional regressors to your model from part a: (1) a dummy variable to account for the party of the governor in each state, and (2) the unemployment rate in 2019.
     1. Report/copy your regression results.



* + 1. Summarize what we learn from the coefficients about the role of the governor and prior unemployment in the level unemployment during 2020. Do these results match your expectations?

We can’t be super confident about the state having a Republican governor having an effect in the 2020 unemployment because the p-value for this regressor was high and not under the 0.05 statistically significant cut off. If a relationship does exist then if there was Republican governor in office for the state, there was 0.754651 percent less unemployment than if there was a Democratic governor in office. This is likely attributed to Republican policy being against stay-at-home orders to keep the economy open. Also, for every percent of unemployment in the state during 2019, there was 0.793658 percent unemployment in the state that persisted in 2020. This makes sense because unemployment typically doesn’t change very drastically over the course of a year, and it can take some people longer than a year to find a job.

* + 1. What has happened to your coefficient on the state policy variable (**SahoDays**)? How much would an additional month (30 days) of stay-at-home orders affect unemployment in this model, and how does that compare to your model from part a? Which estimate should we trust more? Explain.

The coefficient for Saho Days has decreased because some of the unemployment that was attributed to it before when it was the only regressor in the model is now being attributed to the new regressors in the new model. We should expect to see 0.864378 percent (0.0288126 x 30 = 0.864378) more unemployment if there is a stay-at-home order for 30 days. This differs from the model in part a because there the coefficient was higher in value at 0.0421957. Based off the model from part a, we would expect to see 1.265871 percent (0.0421957 x 30 = 1.265871) more unemployment if there is a stay-at-home order for 30 days. I would say we should trust this new model with the extra regressors rather than the model from part a because the new model has a higher adjusted R^2 and Saho Days is still very statistically significant.

* 1. (5 points) For this one, you will run two regressions. Each one will use the 2020 unemployment rate as the dependent variable, and two variables, the number of days with a stay-at-home order and the 2019 unemployment rate, as regressors.
     1. For the first regression, restrict the sample to states with a Republican Governor. Report/copy your results.

Regression with 27 States with Republican governors:

A white text with black numbers

Description automatically generated

* + 1. For the second regression, restrict the sample to states with a Democratic Governor. Report/copy your results.

Regression with 23 states with Democratic governors:

A white text with black numbers

Description automatically generated

* + 1. Summarize what we learn by comparing the coefficients, significance, and measures of explanatory power for your two specifications. Do these results seem reasonable?

The stay-at-home orders had more of an impact in increasing unemployment for states with Republican governors than for states with Democratic governors. This might be because the states with Republican governors tend to be more rural so stay at home orders might have more of impact because people are more distanced from other people making it more difficult to engage in the economy. Unemployment from 2019 also persisted more in states with Democratic governors. Another thing to note is none of the regressors are statistically significant and the adjusted R^2 was low for the regression focusing on states with Democratic governors whereas the regressors are statistically significant and the adjusted R^2 was high for the regression focusing on states with Republican governors. I looked at the data of the states, and there might be potential outliers, which might be ruining the statistical significance and explanatory power of the regression of states with Democratic governors. The two states with highest unemployment were Hawaii and Nevada, and this was likely attributed to the reduction in tourism during the pandemic.

* 1. (3 points) Create a scatterplot that examines the relationship between the number of stay-at-home order days and the 2020 unemployment rate across states.
     1. Copy the plot into your document



* + 1. Discuss what you notice in the plot in terms of the relationship and the nature of each of the variables. What do you notice from the plot that might be affecting our results and our understanding of the impact of policy on economic outcomes?

There are two or more groupings of states on the plot. There is a big gap of empty space near the beginning of the plot from around 1 to 23 days. In the second grouping of states from about 24 to 75 days there might be a problem with heteroskedasticity. My guess is that most of the states with Republican governors are grouped on or closer to the left-hand side of the plot whereas the states with Democratic governors are more grouped near the right-hand side of the plot and not grouped as close together. The 1 or 2 potential outliers in the data are states with Democratic governors. This is why the states with Republican governors have better regression results in terms of statistical significance of the regressors and explanatory power than the states with Democratic governors. If you were to remove the 1 or 2 outliers, I think the p-values and the adjusted R^2 would improve for the regression that has the states with Democratic governors. Because of the natural groupings in the data, I could also create categorical variables for the number of stay-at-home days a state had in 2020 to see if that improves the model.

1. (7 points) Freedom! Using the information in the dataset, create a set of two regressions that examine relationships present between the variables provided. These regressions should be related to each other. There are many ways this could be accomplished – you could run the same regression with two subsamples, you could run two regressions with the same dependent variable and different sets of regressors, you could run the first regression with a linear relationship between x and y and then run a second where you add a quadratic term, you could run one looking at the total COVID death rate as the dependent variable and compare to one using the 2020 COVID death rate as the dependent variable, and so on. The only limit to what you can do is your imagination (and the restrictions that n>k and there is no perfect collinearity).
   * 1. Explain what the overall concept of your analysis is. Why are you including the variables you are including, what do you expect to find, etc. Explain both specifications that you will run and (if it is not obvious) why the two are related.

For the first and second regression the dependent variable will be the percentage of votes Biden received for a state in the 2020 election. The first regression will have Unemployment 2020 and the COVID death rate for 2020 as the regressors. The second regression will have GDP growth rate for 2020 and the stay-at-home days for 2020 as the regressors. The two are regressions are related because the percentage the state voted for Biden in the 2020 election is the dependent variable. Both regressions have an independent variable that relates to the economy and another variable that relates to COVID.

* + 1. Report/copy your results for the first model. Summarize what we learn in terms of significance/magnitude/explanatory power.

A screenshot of a computer

Description automatically generated

In this regression, there likely is a relationship between the stay-at-home days in 2020 for a state and the percent the state voted for Biden in the 2020 election because the p-value for this is very low; however, this is not the case with the GDP growth rate for the state in 2020 and the percent the state voted for Biden in the 2020 election because the p-value was quite high above 0.05 for this relationship. Another thing to note, is the R^2 is moderately high here, and there is not much of a difference between the R^2 and the adjusted R^2 here.

* + 1. Report/copy your results for the second model. Summarize what we learn in terms of significance/magnitude/explanatory power.

A white paper with black text

Description automatically generated

In this regression, there likely is a relationship between the unemployment rate in 2020 for a state and the percent the state voted for Biden in the 2020 election because the p-value for this is low; however, this is not the case with the COVID Death Rate for the state in 2020 and the percent the state voted for Biden in the 2020 election because the p-value was above 0.05 for this relationship. The p-value for unemployment of the state in 2020 in this regression is higher than the p-value for the number of stay-at-home days for the state in the previous regression. Another thing to note, is the R^2 is not that high here and it is lower than the prior regression, but more importantly there is more of a difference between the R^2 and the adjusted R^2 here than the previous regression. This may be because of issues of collinearity.

* + 1. Explain what we learn overall from your analysis. Did the results match your expectations? What might be improved going forward?

The first regression indicates that there likely is a relationship between the number of stay-at-home days for a state in 2020 and the percent the state voted for Biden in the 2020 election. The second regression indicates that there is also likely a relationship with how the unemployment rate for a state in 2020 affected the percent of the vote that Biden received in the 2020 election. Although, this relationship is not as likely as the former one mentioned because the p-value for unemployment of a state in 2020 in the second regression was higher than the p-value for the number of stay-at-home days for the state in 2020 in the first regression which indicates it has less statistical significance. To further explore the relationships, I ran a third regression with all the regressors from the previous two regressions:

A screenshot of a computer

Description automatically generated

In this third regression, the number of stay-at-home days was by far the most statistically significant of the four regressors. Another thing to note is the R^2 in this regression is just slightly higher than the R^2 in the first regression, and more importantly the difference in R^2 and adjusted R^2 in this regression is greater than the first regression. This may be because of issues of collinearity. We can conclude that of the relationships we tested for, the strongest one was the one between the number of stay-at-home days for a state in 2020 and how that correlated with the percent the state voted for Biden in the 2020 election. The best regression model to use of the three regressions would probably be the first regression, although we may want to consider dropping the GDP growth rate for the state 2020 if the R^2 and adjusted R^2 aren’t much lower in a new regression without it than when it was in the regression. To improve from there, we would want to test for other regressors and keep them in the model if their relationships are statistically significant and if they are meaningful to the explanatory power of the model.