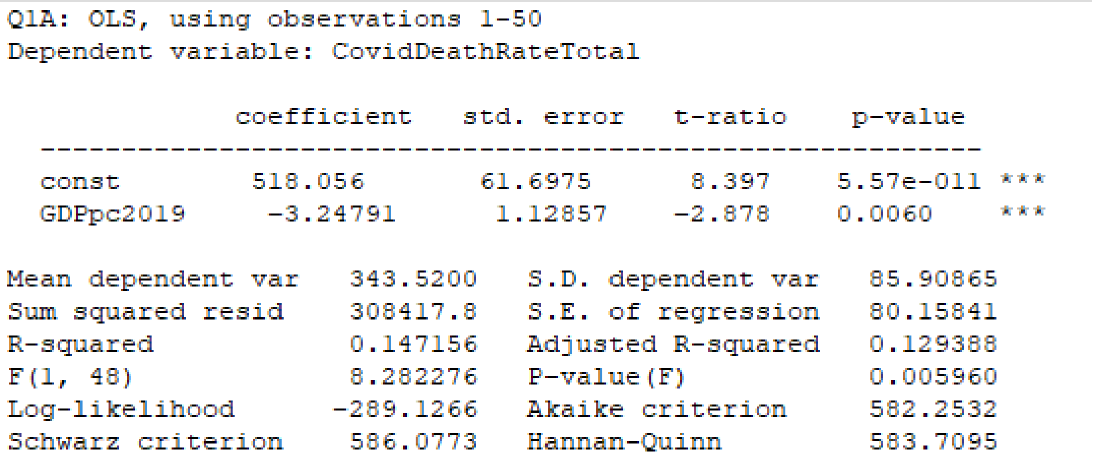
ECON 453
Fall 2023
Problem Set 1 – 40 points

ANSWER KEY

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).
   a. (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.
      i. Report/copy the regression results.



```
Q1A: OLS, using observations 1-50
Dependent variable: CovidDeathRateTotal

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        518.056      61.6975      8.397    5.57e-011 ***
  GDPpc2019     -3.24791     1.12857    -2.878    0.0060    ***

Mean dependent var   343.5200   S.D. dependent var   85.90865
Sum squared resid    308417.8   S.E. of regression   80.15841
R-squared            0.147156   Adjusted R-squared   0.129388
F(1, 48)             8.282276   P-value(F)           0.005960
Log-likelihood      -289.1266   Akaike criterion     582.2532
Schwarz criterion    586.0773   Hannan-Quinn         583.7095
```

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

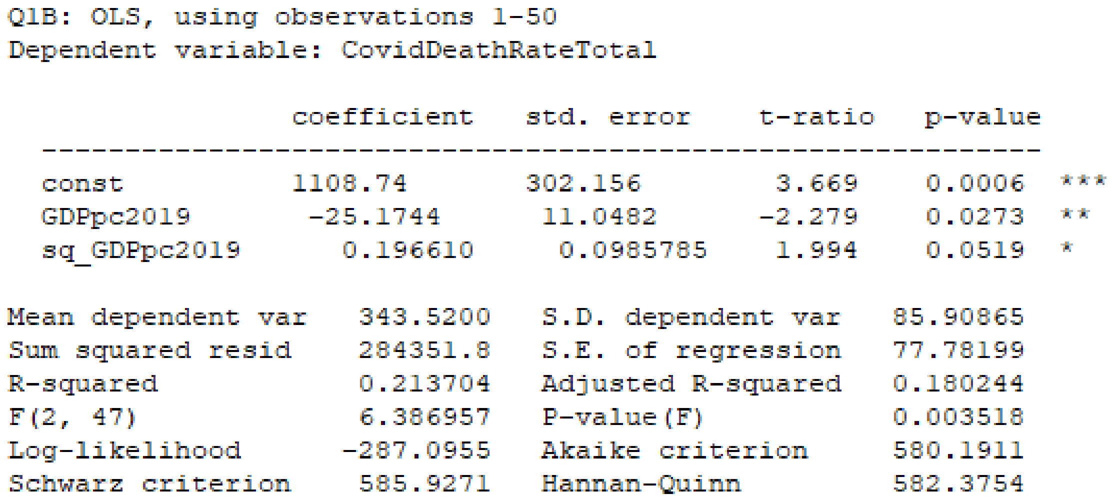
There are several ways to find this, but these easiest way is just to use the Gretl menus. From the regression results window you can select Analysis, then Display actual, fitted, residual. Here we see htat Idaho (state 12) had actual rate of 309 deaths per 100,000 in the population. Based on the GDP per capita, our model predicts the rate should have been 387.8. This means the residual (actual – predicted) is -78.8. The residual tells us that Idaho had a lower COVID death rate than we might expect based on the state's economy.

      iii. 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 tells us that increasing GDP per capita in a state by $1,000 is predicted to decrease the COVID death rate by about 3.25 people per 100,000 in the population. The result is highly statistically significant, with a p-value of 0.006. The explanatory power of our model is pretty low (R^2 of 0.147), but this is not a surprise given we only have one explanatory variable. I would say this matches my expectations; I am not surprised that states with fewer resources would be hit harder by a pandemic.

b. (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.

i. Report/copy your regression results.



```
Q1B: OLS, using observations 1-50
Dependent variable: CovidDeathRateTotal

                 coefficient   std. error    t-ratio   p-value
  ------------------------------------------------------------
  const          1108.74       302.156        3.669    0.0006   ***
  GDPpc2019       -25.1744      11.0482      -2.279    0.0273   **
  sq_GDPpc2019      0.196610     0.0985785    1.994    0.0519   *

Mean dependent var   343.5200   S.D. dependent var   85.90865
Sum squared resid    284351.8   S.E. of regression   77.78199
R-squared            0.213704   Adjusted R-squared   0.180244
F(2, 47)             6.386957   P-value(F)           0.003518
Log-likelihood      -287.0955   Akaike criterion     580.1911
Schwarz criterion    585.9271   Hannan-Quinn         582.3754
```

ii. 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?

To find this, you should be using the quadratic equation from your results to predict two times (once for GDP = 50,000, once for 60,000). Remember that the model has the variable measured in 1000s, so we need to plug in 50, not 50,000.

Prediction for 50,000 = 1108.74 - (25.1744*50) + (0.1966*50^2) = 341.52
Prediction for 60,000 = 1108.74 - (25.1744*60) + (0.1966*60^2) = 306.04

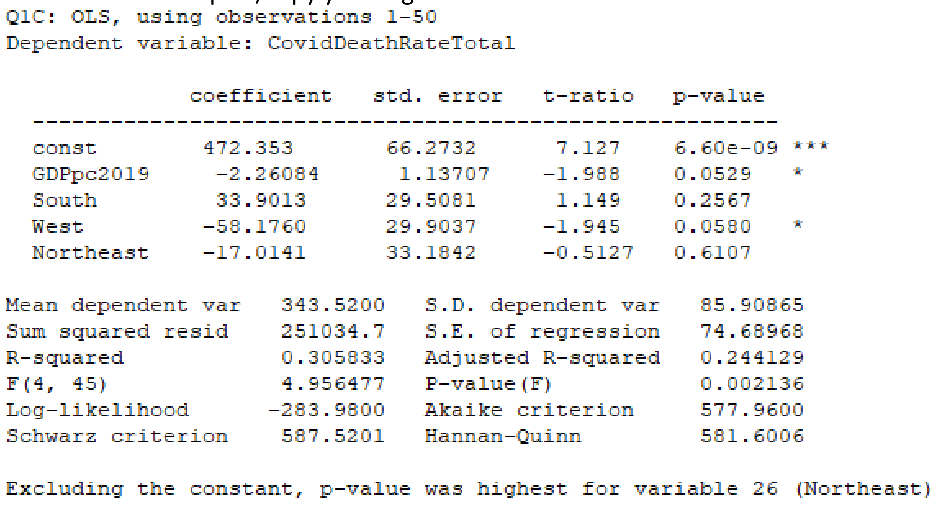
This model predicts that there will be about 35 fewer deaths per 100,000 in the population (about a 10% drop) in a state with GDP per capita of $60,000 as compared to a state with $50,000.

iii. 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.

This model tells us there is a negative relationship between GDP per capita and the COVID death rate in a state, but that the magnitude of this effect diminishes as the GDP per capita grows. In other words, there are diminishing returns to improving the economy in the state. This seems reasonable, as we know from our discussions of life expectancy, money can only improve health outcomes to a certain point. Overall, the non-linear model appears to be a good idea. Our adjusted-R^2 has improved, and each of the variables is at least weakly significant in this model.

c. (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.

i. Report/copy your regression results.



```
Q1C: OLS, using observations 1-50
Dependent variable: CovidDeathRateTotal

              coefficient   std. error   t-ratio   p-value
  ----------------------------------------------------------
  const        472.353      66.2732       7.127    6.60e-09 ***
  GDPpc2019     -2.26084     1.13707     -1.988    0.0529   *
  South         33.9013     29.5081       1.149    0.2567
  West         -58.1760     29.9037      -1.945    0.0580   *
  Northeast    -17.0141     33.1842      -0.5127   0.6107

Mean dependent var   343.5200   S.D. dependent var   85.90865
Sum squared resid    251034.7   S.E. of regression   74.68968
R-squared            0.305833   Adjusted R-squared   0.244129
F(4, 45)             4.956477   P-value(F)           0.002136
Log-likelihood      -283.9800   Akaike criterion     577.9600
Schwarz criterion    587.5201   Hannan-Quinn         581.6006

Excluding the constant, p-value was highest for variable 26 (Northeast)
```

ii. 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?

Our model predictions for this state would be:

Predicted death rate = 472.353 – (2.26084*50) – 17.0141 = 342.3

iii. Summarize what we learn about the differences across regions in terms of the death rate.

What we learn is that there do not appear to be very many significant differences across regions. My reference category is the Midwest. Neither the South nor the Northeast are found to be significantly different from the MW. The West has a lower death rate, but this is only weakly significant. If we were putting the regions in order, we would say the South has the highest death rates, followed by the Midwest, then the Northeast, and the West has the lowest death rates. Again, we are overall finding not much difference.

iv. 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 estimated coefficient dropped by about a third of its magnitude when we included the regional dummy variables. This tells us that there are likely correlations between the regions and GDP per capita (there definitely are). When we omitted the region variables, we may have been overstating the effect that the economy has on the health outcomes because the economic differences were used in place of other regional differences.

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

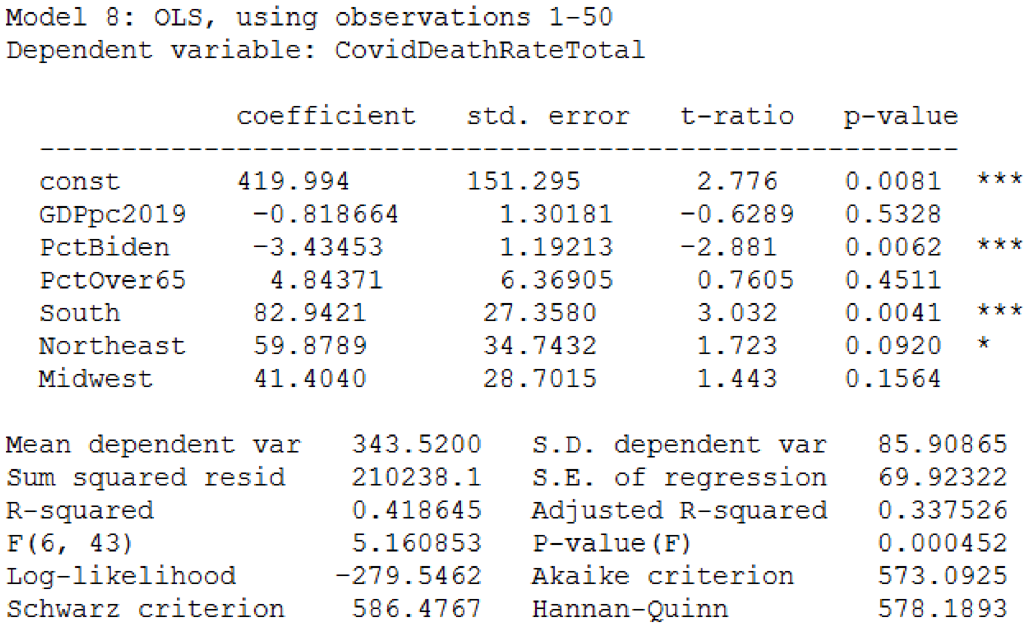
There are a lot of ways you could go with this part of the question. The most obvious thing to me is that we should probably clean up the regional dummy variables. We don’t need them all in there if they are not adding much significance. We would likely want to include either the South or the West dummy, since the biggest difference is between these two. The other thing we would want to do is try and find some more variables to add that might improve the explanatory power of our model.

d. (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.

i. Explain your reasoning for constructing the model in this way. What are your hypotheses (what are you expecting to find)?

I was pretty boring here and just tried to add some variables that I thought would improve my explanatory power. I included the percent voting for Biden and the percent over 65 as additional regressors. I kept the GDP per capita variable in there, and used the regions while omitting the West, since I know it has the lowest rates.

ii. Report/copy your regression results.



```
Model 8: OLS, using observations 1-50
Dependent variable: CovidDeathRateTotal

                coefficient   std. error   t-ratio   p-value
  -----------------------------------------------------------
  const          419.994       151.295       2.776    0.0081   ***
  GDPpc2019       -0.818664      1.30181    -0.6289   0.5328
  PctBiden        -3.43453       1.19213    -2.881    0.0062   ***
  PctOver65        4.84371       6.36905     0.7605   0.4511
  South           82.9421       27.3580      3.032    0.0041   ***
  Northeast       59.8789       34.7432      1.723    0.0920   *
  Midwest         41.4040       28.7015      1.443    0.1564

Mean dependent var   343.5200   S.D. dependent var   85.90865
Sum squared resid    210238.1   S.E. of regression   69.92322
R-squared            0.418645   Adjusted R-squared   0.337526
F(6, 43)             5.160853   P-value(F)           0.000452
Log-likelihood      -279.5462   Akaike criterion     573.0925
Schwarz criterion    586.4767   Hannan-Quinn         578.1893
```

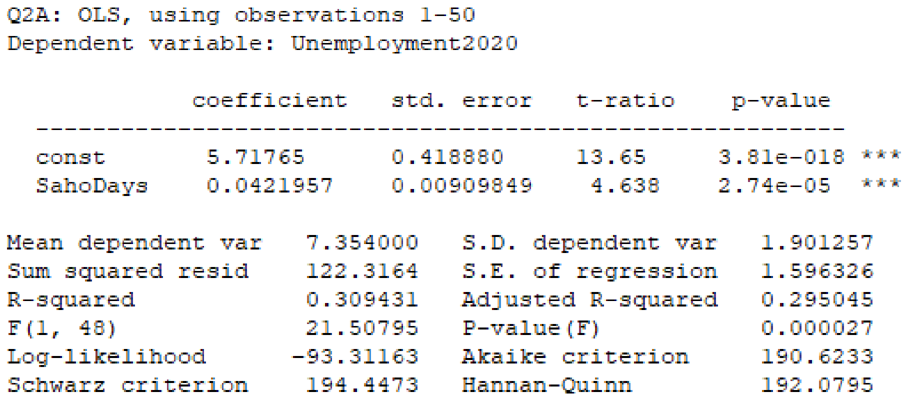
iii. Summarize the findings from your model. You should address significance, explanatory power, the sign and magnitude of coefficients, and whether the results matched your expectations.

Let's start with the good news – my explanatory power (as measured by the adjusted-R^2 value) has improved from our previous models. The Biden variable is highly significant, and the regions are showing more significance than before. On the other hand, the Over 65 variable is not doing anything for me, and the GDP per capita variable is no longer anywhere close to being statistically significant. In terms of matching my expectations, it is a mixed bag. I am not surprised that the Biden variable has a negative coefficient. I am surprised that the age variable was not significant, and that GDP lost its significance.

iv. What should be done to improve your model going forward?

First, I will probably have a nice, long cry about not doing a better job. Then I will pick myself up, dust myself off, and get to the business of improving my model. As before, an obvious first step would be to take out some of the insignificant variables. Age can go, and GDP can go at this point, though I may want to think about whether I want Biden or GDP in there (it seems I can't have both). Then I would try some other variables to see if they can add more predictive power (obesity, vaccination rates, poverty, etc.).

2. 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.
    a. (3 points) Run a regression using **Unemployment2020** as the dependent variable and **SahoDays** as the independent variable.
        i. Report/copy your regression results.

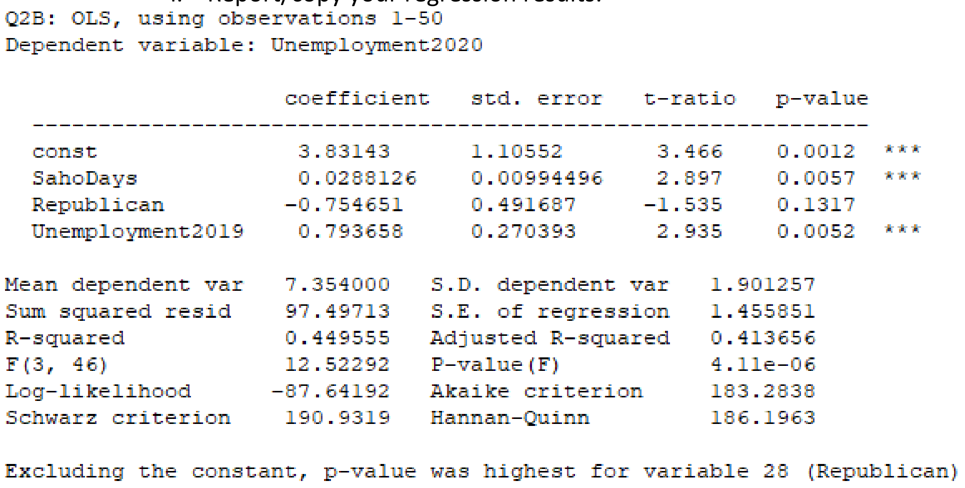


```
Q2A: OLS, using observations 1-50
Dependent variable: Unemployment2020

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       5.71765      0.418880     13.65     3.81e-018 ***
  SahoDays    0.0421957    0.00909849    4.638    2.74e-05  ***

Mean dependent var   7.354000   S.D. dependent var   1.901257
Sum squared resid    122.3164   S.E. of regression   1.596326
R-squared            0.309431   Adjusted R-squared   0.295045
F(1, 48)             21.50795   P-value(F)           0.000027
Log-likelihood      -93.31163   Akaike criterion     190.6233
Schwarz criterion    194.4473   Hannan-Quinn         192.0795
```

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

This says that every additional day with a stay-at-home order (SAHO) in place is expected to increase the unemployment rate by 0.042. This did not surprise me. More SAHO days meant more adverse impacts on the economy (to the extent they were respected/enforced). We should not take this as definitive proof, but the preliminary results are in line with what I thought.

    b. (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.
        i. Report/copy your regression results.



```
Q2B: OLS, using observations 1-50
Dependent variable: Unemployment2020

                    coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------------
  const              3.83143      1.10552       3.466    0.0012   ***
  SahoDays           0.0288126    0.00994496    2.897    0.0057   ***
  Republican        -0.754651     0.491687     -1.535    0.1317
  Unemployment2019   0.793658     0.270393      2.935    0.0052   ***

Mean dependent var   7.354000   S.D. dependent var   1.901257
Sum squared resid    97.49713   S.E. of regression   1.455851
R-squared            0.449555   Adjusted R-squared   0.413656
F(3, 46)             12.52292   P-value(F)           4.11e-06
Log-likelihood      -87.64192   Akaike criterion     183.2838
Schwarz criterion    190.9319   Hannan-Quinn         186.1963

Excluding the constant, p-value was highest for variable 28 (Republican)
```

        ii. 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?

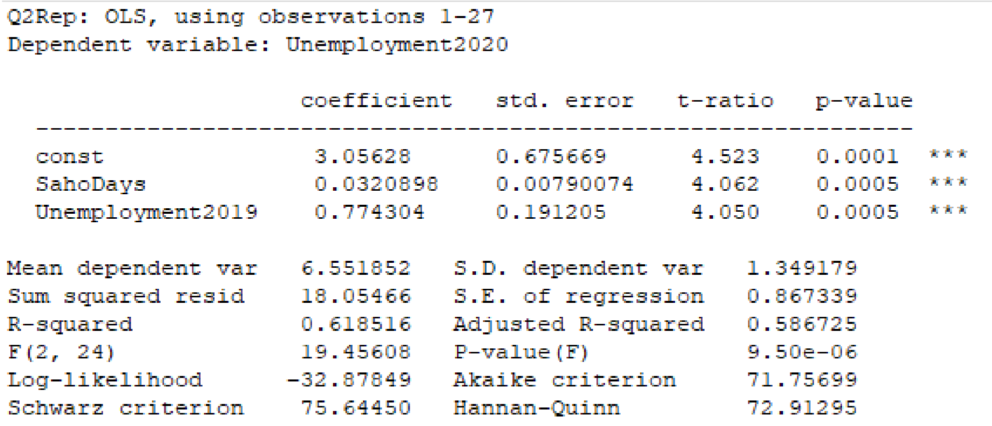
We learn that the governor party did not seem to matter for unemployment (once SAHO days was accounted for) and that states with higher unemployment in 2019 tended to have higher unemployment in 2020. The correlation between years in unemployment makes sense. This is often based on the nature of the state's economy. I am a little surprised there is no significance to the governor variable.

iii. 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 is smaller in magnitude in the model with the additional variables included. The value is 0.0422 in the model in part a and 0.0288 in the model in part b. This means 30 days of SAHO would be expected to increase unemployment by 1.266 % points if we use the first estimate and by only 0.864 if we use the second estimate. We should probably trust the second model more, since we have included an important detail, the prior year's unemployment rate. Our adjusted-R^2 is higher, and overall we have a better sense of the actual impact of the policy from this model.

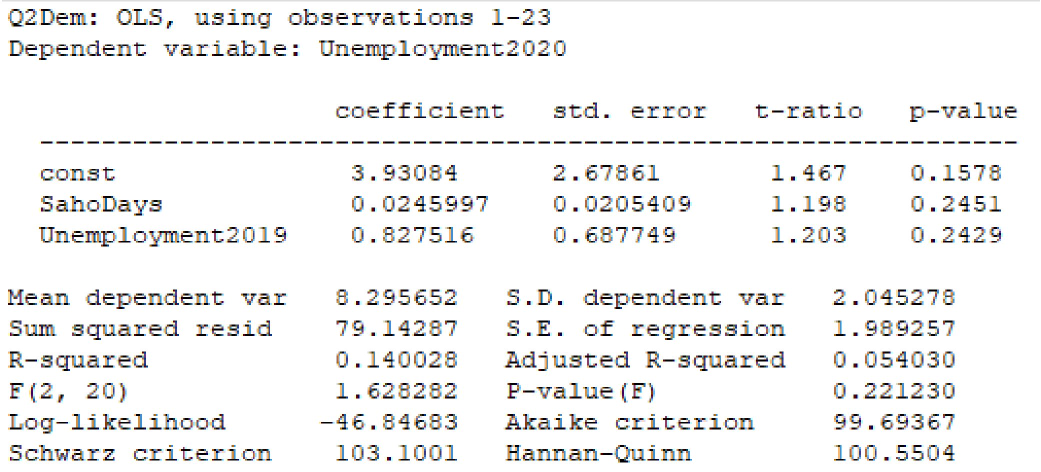
c. (5 points) For this one, you will run two regressions. Each one will use the 2020 unemployment rate as the dependent variable, and two variables, the number of days with a stay-at-home order and the 2019 unemployment rate, as regressors.

i. For the first regression, restrict the sample to states with a Republican Governor. Report/copy your results.



```
Q2Rep: OLS, using observations 1-27
Dependent variable: Unemployment2020

                     coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------------
  const               3.05628      0.675669      4.523    0.0001   ***
  SahoDays            0.0320898    0.00790074    4.062    0.0005   ***
  Unemployment2019    0.774304     0.191205      4.050    0.0005   ***

Mean dependent var   6.551852   S.D. dependent var   1.349179
Sum squared resid    18.05466   S.E. of regression   0.867339
R-squared            0.618516   Adjusted R-squared   0.586725
F(2, 24)             19.45608   P-value(F)           9.50e-06
Log-likelihood      -32.87849   Akaike criterion     71.75699
Schwarz criterion    75.64450   Hannan-Quinn         72.91295
```

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



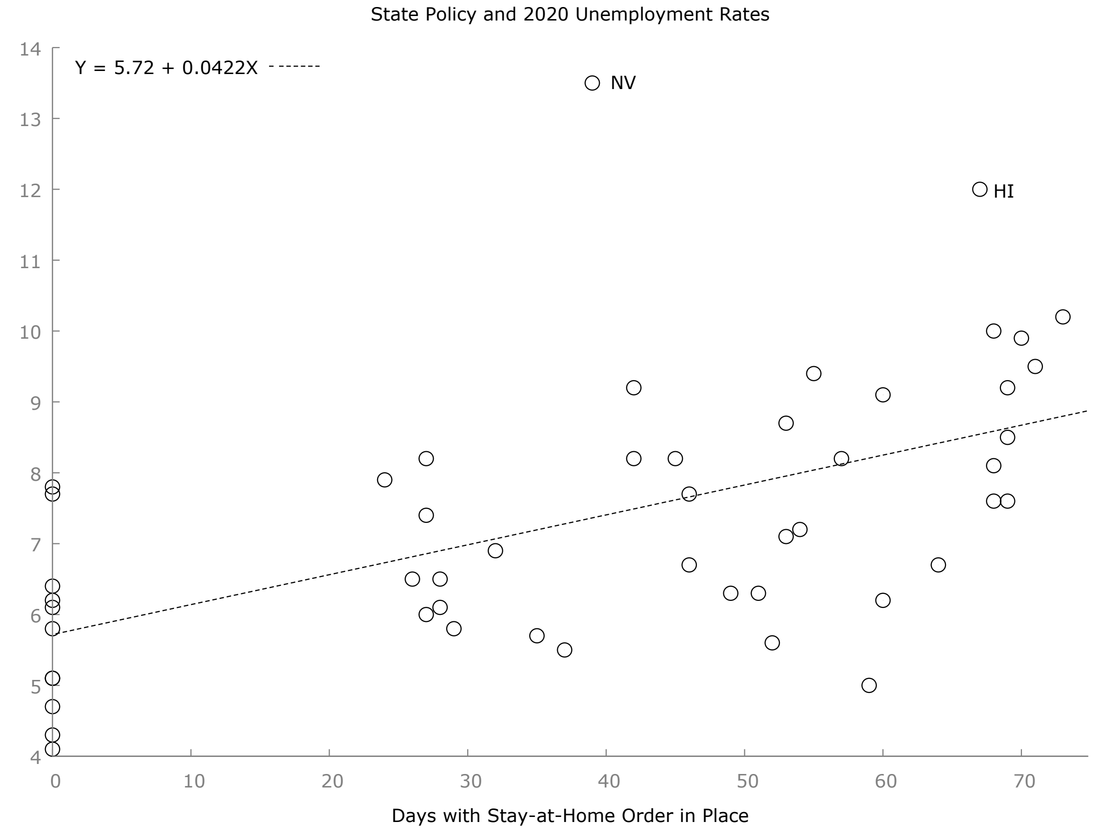
```
Q2Dem: OLS, using observations 1-23
Dependent variable: Unemployment2020

                     coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------------
  const               3.93084      2.67861       1.467    0.1578
  SahoDays            0.0245997    0.0205409     1.198    0.2451
  Unemployment2019    0.827516     0.687749      1.203    0.2429

Mean dependent var   8.295652   S.D. dependent var   2.045278
Sum squared resid    79.14287   S.E. of regression   1.989257
R-squared            0.140028   Adjusted R-squared   0.054030
F(2, 20)             1.628282   P-value(F)           0.221230
Log-likelihood      -46.84683   Akaike criterion     99.69367
Schwarz criterion    103.1001   Hannan-Quinn         100.5504
```

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

The results here are somewhat interesting. The coefficient estimates are similar in the two regressions, but there is strong significance in the Republican model and no significance in the Democratic model. As we discussed in class, this might be due to the fact that there are some outliers in the Democratic states.

If we summarize what we learn, we would conclude that the length of SAHO seemed to impact unemployment rates in states lead by a Republican, but not in states lead by a Democrat. The explanatory power of our model is pretty strong for the Republican model, and almost nothing in the Democrat model.

d. (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.
   i. Copy the plot into your document
   ii. 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?



I chose this relationship to use for the problem set because I am hoping it gives you the idea that we need to carefully study things like the effect of policy on unemployment. It is not something that we want to casually gather data for and throw something together. One thing we notice when we look at the data – there are several states that are listed as having no SAHO days. That might seem hard to believe, and might lead you to question how the SAHO days are being counted for this analysis. How do we count this if the state had a limited SAHO (over 65, for example)? We also notice that there are some outliers that might impact our results. Nevada and Hawaii had very high unemployment rates early in the pandemic because of the impact of tourism shutting down. The rise in unemployment in these states may have been impacted by a SAHO, but was likely more impacted by other factors. I also notice that there are several groupings of states (those at 0, those around 30 days, etc.). Maybe we should try to break the SAHO variable into groups instead of the number of days in each state.

3. (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).
    i. 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.
    ii. Report/copy your results for the first model. Summarize what we learn in terms of significance/magnitude/explanatory power.
    iii. Report/copy your results for the second model. Summarize what we learn in terms of significance/magnitude/explanatory power.
    iv. Explain what we learn overall from your analysis. Did the results match your expectations? What might be improved going forward?