Name: **ANSWER KEY**

ECON 453
In-Class Exercise 2
September 5, 2023

Please download the file "In-Class 2.gretl", a gretl "session" file. This is very similar to the dataset we worked with in class on Thursday August 31st, with a few minor adjustments. The data here come from the World Bank Development Indicators (https://data.worldbank.org/indicator) and are measured for a set of 177 countries in the year 2019. Please open the session file (**Files -> Session files -> Open session**).

1. Let's examine the relationship between agriculture and the economy in a country. Run a regression (**Model -> Ordinary Least Squares**) using GDP per capita (in 1000s) as the dependent variable and the percentage of GDP that comes from agriculture (**Ag_pct**) as the regressor.
    a. Report your estimated equation. Provide a numerical interpretation of the coefficient. Does this make sense to you?

The results should look like this:

gretl: model 1

File Edit Tests Save Graphs Analysis LaTeX

```
Model 1: OLS, using observations 1-177 (n = 173)
Missing or incomplete observations dropped: 4
Dependent variable: GDP_1000s

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       35.3979       1.83576      19.28    9.43e-045 ***
  Ag_pct      -1.32818      0.125698    -10.57    2.10e-020 ***

Mean dependent var   21.50738   S.D. dependent var   21.60501
Sum squared resid    48571.78   S.E. of regression   16.85365
R-squared            0.395012   Adjusted R-squared   0.391474
F(1, 171)            111.6504   P-value(F)           2.10e-20
Log-likelihood      -733.1207   Akaike criterion     1470.241
Schwarz criterion    1476.548   Hannan-Quinn         1472.800
```

That means my estimation equation can be written as: $\hat{y} = 35.398 - 1.328 * (Ag\_pct)$.

The coefficient is interpreted as: every increase of 1 percentage point of GDP that comes from agriculture in a country is expected to lead to a roughly $1,328 drop in GDP per capita. The negative value makes sense to me, countries where agriculture is a greater share of the economy tend to be on the lower-income end of the spectrum.

    b. Briefly discuss the significance of the Ag_pct variable, as well as the overall explanatory power of your model.

The p-value (and stars!) tell us that there is strong statistical significance to our estimated relationship. Our R^2 ***value is around 0.4. This is not the highest value I have seen but seems decent considering we only have one factor (% of GDP that comes from agriculture) that we are trying to use to explain differences across countries in GDP per capita.***

    c. What does your model predict the GDP per capita (in 1000s) should be for the U.S. (country 171)? For Ethiopia (country 54)? How far off are these predictions (find the residuals)?

There are a couple of ways to do this. One is to use the option in the regression results window in gretl (*Analysis -> Display actual, fitted, residual****). The other is to use the dataset to find the values for Ag_pct and GDP_1000s for each country. For the U.S., GDP per capita is 65.095 and the Ag % is 0.84. For Ethiopia, the GDP is 2.274 and Ag % is 33.63. We can plug the ag values in to make our predictions:***

Predicted GDP per capita for USA: $= 35.398 - (1.328 * 0.84) = 34.282$
Predicted GDP per capita for Ethiopia $= 35.3979 - (1.328 * 33.63) = -9.263$

First, we should probably note that the model has predicted a negative value of GDP per capita for Ethiopia. This is an indication that our use of a linear model might be problematic. We can also find the residuals, which tell us how far off our predictions are. For the US, the residual is (65.095-34.282) = 30.813. The actual GDP is 30,813 higher than our prediction (not great, we are almost 50% off!!!). For Ethiopia, the residual is 2.274-(-9.263) = 11.537.

2. The model in question 1 produces some unusual results. Try running 2 separate regressions, one with the bottom 50% of countries (in terms of GDP per capita) and one with the top 50% of countries. (**Sample -> Restrict, based on criterion**).
    a. Report the estimated equation, number of observations, and R^2 value *for each* of your regressions, then briefly summarize what we have learned.

The first step is to find the median value of GDP_1000s (so we know the value to use for our restriction). I did this in two steps, first – restrict so the observations that are missing Ag_pct are dropped. Then, restrict to values less than the median.

To restrict to eliminate missing Ag_pct:

gretl: restrict sample

Enter boolean condition for selecting cases:

!missing(Ag_pct)

☐ Make this restriction permanent

Help | Cancel | OK

Then, I restrict based on being below the median of GDP per capita:

gretl: restrict sample

Enter boolean condition for selecting cases:

GDP_1000s<median(GDP_1000s)

◉ add to current restriction

○ replace current restriction

Help | Cancel | OK

My results for lower-income countries (GDP_1000s<14.437):

gretl: model 5

File Edit Tests Save Graphs Analysis LaTeX

```
Model 5: OLS, using observations 1-86
Dependent variable: GDP_1000s

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const       10.5797      0.700149     15.11     1.12e-025 ***
  Ag_pct      -0.247766    0.0345839    -7.164    2.77e-010 ***

Mean dependent var    6.222076   S.D. dependent var   4.057457
Sum squared resid     868.6131   S.E. of regression   3.215685
R-squared             0.379275   Adjusted R-squared   0.371885
F(1, 84)              51.32553   P-value(F)           2.77e-10
Log-likelihood       -221.4684   Akaike criterion     446.9368
Schwarz criterion     451.8455   Hannan-Quinn         448.9123
```

And for higher-income countries (GDP_1000s>14.437):

gretl: model 6

File Edit Tests Save Graphs Analysis LaTeX

```
Model 6: OLS, using observations 1-86
Dependent variable: GDP_1000s

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       53.7095       2.87165      18.70     1.09e-031 ***
  Ag_pct      -4.91135      0.655968     -7.487    6.38e-011 ***

Mean dependent var   36.87491   S.D. dependent var   21.26398
Sum squared resid    23050.49   S.E. of regression   16.56534
R-squared            0.400247   Adjusted R-squared   0.393107
F(1, 84)             56.05774   P-value(F)           6.38e-11
Log-likelihood      -362.4458   Akaike criterion     728.8916
Schwarz criterion    733.8003   Hannan-Quinn         730.8671
```

The R^2 values are very similar between the two regressions, but what we should notice is how different our estimated coefficients are. For lower income countries, each percentage point increase in the share of GDP that comes from Agriculture is expected to decrease GDP per capita by about $248. For higher income countries, the same 1 percentage point increase is expected to decrease per capita GDP by about $4,911. This makes sense, many of the higher income countries likely have a much smaller share of GDP that comes from agriculture, so a 1 percentage point gain is a bigger deal, relatively speaking.

b. Predict the values for Ethiopia and the U.S. Are your predictions better or worse now?

For Ethiopia, we should use the lower-income country model (since we know GDP per capita is 2.274). We know from question 1 that Ethiopia has Ag_pct of 33.63, so:

Predicted GDP per capita for Ethiopia $= \mathbf{10.761 - (0.255 * 33.63) = 2.185}$

This is a much better prediction than our earlier model (residual = 2.274 – 2.185 = 0.089)
For the US, we should use the higher-income country model (since we know GDP per capita is 65.095). We know from question 1 that the US has Ag_pct of 0.84, so:

Predicted GDP per capita for USA $= \mathbf{53.7095 - (4.911 * 0.84) = 49.584}$

This is also a much better prediction than our earlier model (residual = 65.095 – 49.584 =15.511)

3. Let’s try adding a quadratic term to our equation. Create a squared version of the “Ag_pct” variable (highlight the Ag_pct variable, then choose **Add -> Squares of selected variables**). Run a regression with GDP_1000s as the dependent variable and both Ag_pct and sq_Ag_pct as regressors.
   a. Report your estimated equation. Use this equation to predict the GDP per capita for Ethiopia.

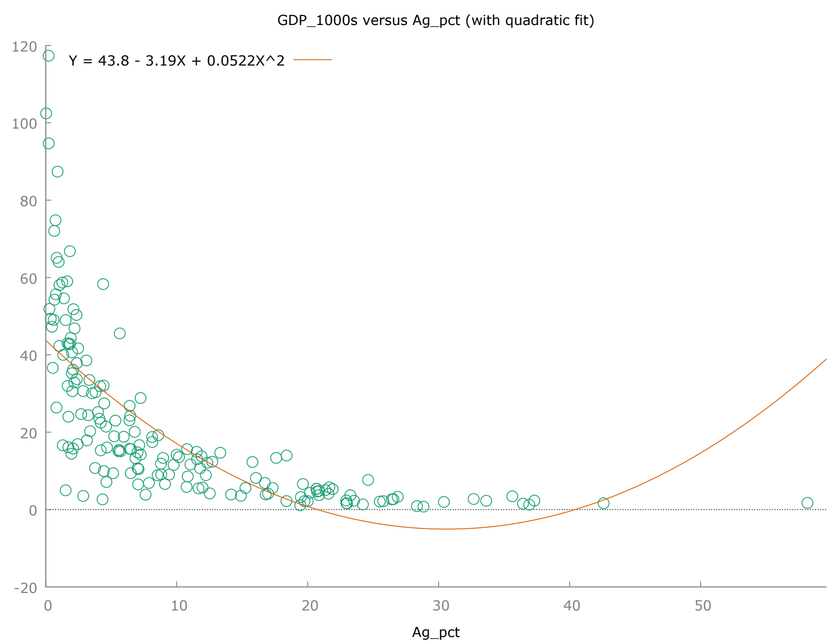
Here are the results of my regression where I have used both Ag_pct and the squared version of the variable as regressors:

```
Model 4: OLS, using observations 1-177 (n = 173)
Missing or incomplete observations dropped: 4
Dependent variable: GDP_1000s

              coefficient    std. error    t-ratio    p-value
  ---------------------------------------------------------
  const         43.7675        1.96796       22.24    3.30e-052 ***
  Ag_pct        -3.19339       0.276753     -11.54    4.07e-023 ***
  sq_Ag_pct      0.0522158     0.00711120     7.343   8.29e-012 ***

Mean dependent var   21.50738   S.D. dependent var   21.60501
Sum squared resid    36876.33   S.E. of regression   14.72819
R-squared            0.540685   Adjusted R-squared   0.535282
F(2, 170)            100.0584   P-value(F)           1.90e-29
Log-likelihood      -709.2923   Akaike criterion     1424.585
Schwarz criterion    1434.044   Hannan-Quinn         1428.422
```

We can write our estimated equation as: $\hat{y} = 43.7675 - 3.193 * (Ag_{pct}) + 0.0522 * (Ag_{pct}^{2})$.

Predicted GDP per capita for Ethiopia $= 43.7675 - (3.193 * 33.63) + (0.0522 * (33.63^2)) = -4.576$. ***Oh good, back to a negative predicted value (sarcasm).***

Our coefficient on the linear term is positive and on our squared term is negative, so we find a relationship that drops relatively quickly as Ag_pct increases, then begins to level off (and may even become positive at some point). We can visualize this by looking at a scatterplot and editing so that the trend line is quadratic:



   b. Is this version of our model an improvement over the model in question 1?

The easiest thing to look at here is the R^2 value. According to that metric, this quadratic model is an improvement over the linear model in Question 1. Our R^2 has increased from 0.395 to 0.541, which is a relatively large increase.

4. Next, let's try a different form of non-linear estimation. We are going to use the natural logarithm of GDP per capita. To create this variable, select GDP_1000s from the main gretl window, then **Add -> Logs of selected variables**. Run a regression with l_GDP_1000s as the dependent variable (this is the logged version of GDP_1000s) and Ag_pct as a regressor.
   a. Provide a numeric interpretation of the coefficient on Ag_pct.

Here are the results when I use l_GDP_1000s as my dependent variable:

```
Model 5: OLS, using observations 1-177 (n = 173)
Missing or incomplete observations dropped: 4
Dependent variable: l_GDP_1000s

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       3.49743      0.0689692     50.71    5.73e-105 ***
  Ag_pct     -0.0932607    0.00472244   -19.75    5.59e-046 ***

Mean dependent var   2.522083   S.D. dependent var   1.143539
Sum squared resid    68.55888   S.E. of regression   0.633190
R-squared            0.695187   Adjusted R-squared   0.693405
F(1, 171)            389.9998   P-value(F)           5.59e-46
Log-likelihood      -165.4121   Akaike criterion     334.8242
Schwarz criterion    341.1308   Hannan-Quinn         337.3827

Log-likelihood for GDP_1000s = -601.732
```

Since we have taken the logarithm of the y-variable, we should interpret as: every time the share of GDP coming from agriculture increases by 1 unit (1 percentage point), the GDP per capita in that country is expected to decrease by 9.32 percent.

   b. Compare the explanatory power of this model with those from the previous models and discuss briefly.

Oh man, this is the best model we have seen yet. I can barely contain my excitement. The R^2 value in this "semilog" model is 0.695, which is much better than the 0.541 from the quadratic model or the 0.395 from the linear model. Of the options we have looked at, this seems to be the clear choice in terms of modeling the relationship between the share of GDP coming from agriculture and the GDP per capita in a country.

5. Run another simple linear regression using the variables available in the dataset. Report the estimated equation, interpret the coefficient, and summarize what we learned.

I tried one in which I use life expectancy as the dependent variable and the % immunized for DPT as the regressor. My results:

```
Model 3: OLS, using observations 1-177
Dependent variable: Life_Expectancy

                  coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------------
  const            41.6686      3.32255      12.54     3.75e-026 ***
  Immunized_DPT     0.347489    0.0371210     9.361    3.86e-017 ***

Mean dependent var   72.47144   S.D. dependent var   7.474799
Sum squared resid    6552.532   S.E. of regression   6.119072
R-squared            0.333658   Adjusted R-squared   0.329850
F(1, 175)            87.62772   P-value(F)           3.86e-17
Log-likelihood      -570.7661   Akaike criterion     1145.532
Schwarz criterion    1151.884   Hannan-Quinn         1148.108
```

I learned there is a significant positive relationship between the two variables, and it seems very strong. This says, for example, that increasing the % immunized by 10 percentage points would add about 3.5 years to the life expectancy in a country.