Statistics
Problem Set 1: Descriptive Statistics

Javier Tasso

1. **Cross-Sectional Data.** Download GDP per capita and CO_2 emissions per capita in 2019 from Our World in Data.

   (a) Define $x_i \overset{\text{Def}}{=} \ln(\text{GDP}_i)$. Plot the histogram. Calculate the following measures: mean, median, variance, standard deviation, first, and third quartile.

   (b) Define $y_i \overset{\text{Def}}{=} \ln(\text{CO}_{2i})$ and repeat part (a).

   (c) Focus on variables x and y. Calculate the correlation coefficient and the slope and intercept of the regression line. Make the scatterplot and plot the regression line.

   (d) When both dependent and independent variables are measured in logs, the slope of the regression line has the interpretation of an elasticity. Verify this fact by following these steps.

      i. The regression line is defined as $\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + \varepsilon_i$, where ε_i is the error.

      ii. Totally differentiate this equation both sides. You may assume ε does not change with x.

      iii. Isolate and interpret the ratio $\frac{\frac{dy}{y}}{\frac{dx}{x}}$.

2. **Time Series Data.** Download US real GDP data from FRED. Your sample is $\{\text{GDP}_t\}$ where t is each quarter in 1984Q1-2019Q4.

   (a) Plot the series.

   (b) Calculate $y_t \overset{\text{Def}}{=} \frac{\text{GDP}_t - \text{GDP}_{t-1}}{\text{GDP}_{t-1}}$. This is the growth rate of real GDP. Plot the new series and calculate its mean and standard deviation.

   (c) Calculate the autocorrelation of order h for $h = 1, 2, \ldots, 6$ and plot them. The autocorrelation of order h is defined as the correlation coefficient between y_t and y_{t-h} or $\text{AC}(h) = \frac{\text{COV}(y_t, y_{t-h})}{\text{S}_{y_t}\text{S}_{y_{t-h}}}$.

   (d) Calculate $z_t \overset{\text{Def}}{=} \ln(\text{GDP}_t) - \ln(\text{GDP}_{t-1})$. This is called the log-difference. Plot the new series and calculate its mean and standard deviation.

   (e) Verify that y_t and z_t are approximately the same following these steps.

      i. Consider $f(x) = \ln(x)$. Calculate the equation of the tangent line around $x = 1$. We call this line $g(x)$.

      ii. Intuitively argue that the ratio $\frac{\text{GDP}_t}{\text{GDP}_{t-1}}$ will be close to 1.

      iii. Verify that $f\left(\frac{\text{GDP}_t}{\text{GDP}_{t-1}}\right)$ is the log difference.

      iv. Verify that $g\left(\frac{\text{GDP}_t}{\text{GDP}_{t-1}}\right)$ is the growth rate of real GDP.

3. **Panel Data.** Download data on per capita GDP and life expectancy from Our World in Data for the years 2000-2019. Merge the data to construct a panel. Define $x_i = \ln(\text{GDP})$ to be the log of GDP and y_i to be the life expectancy in years.

(a) Make sure you have a balanced panel, that is, drop any country that has missing observations. Count the number of countries and observations.

(b) Choose one year. Make the scatterplot of x and y for that year.

(c) Choose one country. Plot the time series of x and y for that country.

We are interested in the correlation there is between x and y.

(d) (Pooling) Calculate the correlation coefficient between x and y.

(e) (Time effects) A lot of the increase over time in GDP per capita and LE is due to technological change. In recent years, we may see large values for both x and y simply because of human progress. To control for this time effect, follow these steps:

  i. For each year t calculate $\bar{x}_t$ (defined as the mean value of x_{it} that year) and $\bar{y}_t$ (defined as the mean value of y_{it} that year).

  ii. Define $x^1_{it} \overset{\text{Def}}{=} x_{it} - \bar{x}_t$ and $y^1_{it} \overset{\text{Def}}{=} y_{it} - \bar{y}_t$.

  iii. What do positive/negative values of y^1_{it} mean?

  iv. Calculate the correlation coefficient between x^1_{it} and y^1_{it}.

(f) (Individual effects) Some developed countries may have high GDP per capita and LE throughout the entire sample, while some undeveloped countries may have low values most of the time.

  i. For each country i calculate $\bar{x}_i$ (defined as the mean value of x_{it} in that country) and $\bar{y}_i$ (defined as the mean value of y_{it} in that country).

  ii. Define $x^2_{it} \overset{\text{Def}}{=} x_{it} - \bar{x}_i$ and $y^2_{it} \overset{\text{Def}}{=} y_{it} - \bar{y}_i$.

  iii. What do positive/negative values of y^2_{it} mean?

  iv. Calculate the correlation coefficient between x^2_{it} and y^2_{it}.

(g) (Individual and time effects) Now we take care of the two issues at the same time.

  i. Define $x^*_{it} \overset{\text{Def}}{=} x_{it} - \bar{x}_t - \bar{x}_i$ and y^*_{it} in a similar way.

  ii. Calculate the correlation coefficient between x^*_{it} and y^*_{it}.