Ecn 102 - Analysis of Economic Data
University of California - Davis
John Parman

January 4, 2011

Problem Set 1 - Solutions

This problem set will be due Friday, January 21 by noon. It can be dropped off in my mailbox in the economics department mailroom. No late problem sets will be accepted. You may work in groups but everyone must turn in their own problem set. Note that this does not mean turning in multiple copies of the same problem set. Each group member must produce their own Excel results and write up their own solutions. For Excel output, please only turn in the final results and graphs you are asked for. Do not print out and turn in the complete datasets (use the 'Set Print Area' feature in Excel).

---

1. **Working With Cross-sectional Univariate Data**

   For this question, you will be working with life expectancy data for the year 2006 from the World Health Organization. The datafile (world-health.csv) can be found on Smartsite in the data folder. Note that links to definitions of the variables are provided at the bottom of the spreadsheet.

   (a) Create a histogram for female life expectancy at birth with bins having a width of 5 years. Absolute frequency should be on the vertical axis. Does the distribution of life expectancy appear to be symmetric, left-skewed or right-skewed?

   See ps1-solutions.xslx for the histogram. The distribution is left-skewed, many of the observations are bunched on the right side but there is a long left tail. We can confirm that the distribution is left-skewed by calculating the skewness. Using the SKEW() function in Excel, it turns out that the skewness is -.825. So the distribution is indeed left-skewed.

   (b) Calculate two measures of central tendency and two measures of dispersion for female life expectancy at birth. Calculate these same descriptive statistics for female healthy life expectancy at birth.

   We have several options for measures of central tendency and measures of dispersion. The descriptive statistics function in Excel will output several different measures. See ps1-solutions.xlsx for the output of descriptive statistics for female life expectancy. From this output, measures of central tendency are the mean (69.65 years), the median (73 years) and the mode (75 years). For dispersion, the descriptive statistics give us the standard deviation (11.37 years), the sample variance (129.35 years-squared) and the range (44 years). There are a variety of other measures you could have calculated in Excel to answer this question. The equivalent statistics for healthy life expectancy are shown to the right of the life expectancy statistics on the spreadsheet.

(c) Create a new variable that is defined as the difference between female life expectency and female healthy life expectency. Calculate the mean and standard deviation of this new variable. In one sentence, explain what this new variable in measuring.

> See ps1-solutions.xlsx for the calculation of this new variable and for the calculation of the various means and variances. Note that the mean of life expectancy minus healthy life expectancy is exactly equal to the mean of life expectancy minus the mean of healthy life expectancy. However, the variance of life expectancy minus healthy life expectancy is not equal to the variance of life expectancy minus the variance of healthy life expectancy.
>
> This new variable is measuring the expected number of years spent in poor health for females in each country.

(d) How does the mean of this new variable compare to the difference between the mean female life expectency and the mean female healthy life expectency? How does the variance of this new variable compare to the difference between the variance of female life expectancy and the variance of female healthy life expectancy?

> As shown by calculations in the spreadsheet, the mean of the new variable is exactly equal to the difference in the means between the means of life expectency and healthy life expectency. This is not the case for the variance. The variance of the new variable is quite different than the difference of the variances of life expectency and healthy life expectency.

(e) Let's think about generalizing your results from part (c). Suppose you have two variables x and y and you create a new variable z that is equal to $x - y$. Use the definition of the sample mean and the summation rules from class to show that:

$$\overline{z} = \overline{x} - \overline{y}. \tag{1}$$

Will the same thing be true for the variance of z? In other words, is it possible to show that:

$$\sigma_z^2 = \sigma_x^2 - \sigma_y^2 \ ? \tag{2}$$

> Let's begin by showing the $\bar{z} = \bar{x} - \bar{y}$. To do this, we can start with the formal definition of $\bar{z}$ in terms of sums and use some algebra to show that this can be rewritten in terms of $\bar{x}$ and $\bar{y}$:
>
> $$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$
>
> $$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} (x_i - y_i)$$

$$\bar{z} = \frac{1}{n}(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i)$$

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{z} = \bar{x} - \bar{y}$$

Now let's try to do this same thing for the variance of z. We'll start from the formal definition of σ_z^2 in terms of sums and see if we can rewrite it in terms of σ_x^2 and σ_y^2:

$$\sigma_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$$

$$\sigma_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - y_i - \bar{x} + \bar{y})^2$$

First, let's expand out the terms in the sum and simplify:

$$\sigma_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - x_i y_i - x_i\bar{x} + x_i\bar{y} - y_i x_i + y_i^2 + y_i\bar{x} - y_i\bar{y} - \bar{x}x_i + \bar{x}y_i + \bar{x}^2 - \bar{x}\bar{y} + \bar{y}x_i - \bar{y}y_i - \bar{y}\bar{x} + \bar{y}^2)$$

$$\sigma_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2 + y_i^2 - 2y_i\bar{y} + \bar{y}^2 - 2x_i y_i + 2x_i\bar{y} + 2\bar{x}y_i - 2\bar{x}\bar{y})$$

Now we can start rearranging terms to try to get something that looks like σ_x^2 and σ_y^2:

$$\sigma_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}((x_i - \bar{x})^2 + (y_i - \bar{y})^2 - 2x_i y_i + 2x_i\bar{y} + 2\bar{x}y_i - 2\bar{x}\bar{y})$$

$$\sigma_z^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(2x_i y_i - 2x_i\bar{y} - 2\bar{x}y_i + 2\bar{x}\bar{y})\right]$$

$$\sigma_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{1}{n-1}\sum_{i=1}^{n}(2x_i y_i - 2x_i\bar{y} - 2\bar{x}y_i + 2\bar{x}\bar{y})$$

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - \frac{1}{n-1}\sum_{i=1}^{n}(2x_i y_i - 2x_i\bar{y} - 2\bar{x}y_i + 2\bar{x}\bar{y})$$

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - \frac{1}{n-1}\left(\sum_{i=1}^{n} 2x_i y_i - \sum_{i=1}^{n} 2x_i\bar{y} - \sum_{i=1}^{n} 2\bar{x}y_i + \sum_{i=1}^{n} 2\bar{x}\bar{y}\right)$$

To simplify things further, we can use the fact that $\bar{x}$ and $\bar{y}$ are constants:

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - \frac{1}{n-1}\left(2\sum_{i=1}^{n} x_i y_i - 2\bar{y}\sum_{i=1}^{n} x_i - 2\bar{x}\sum_{i=1}^{n} y_i + 2n\bar{x}\bar{y}\right)$$

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - \frac{1}{n-1}\left(2\sum_{i=1}^{n} x_i y_i - 2n\bar{x}\bar{y} - 2n\bar{x}\bar{y} + 2n\bar{x}\bar{y}\right)$$

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - \frac{1}{n-1}\left(2\sum_{i=1}^{n} x_i y_i - 2n\bar{x}\bar{y}\right)$$

So σ_z^2 is not equal to $\sigma_x^2 - \sigma_y^2$. There are a couple of things you should notice about the final equation above. First, σ_x^2 and σ_y^2 do appear in the equation but both appear with positive signs. The reason for this is that variance is measuring how spread out the values of z are. The more variation there is in y (the larger σ_y^2 is), the more variation there will be in z whether we are subtracting or adding y in the function for z.

The second thing to notice is that we have several extra terms with both x and y in them. As we'll discuss later, these terms are accounting for the covariance between x and y. An easy intuition for why we need these extra terms is the following. Suppose that every time x_i is big, y_i is also big. Then when calculating $z_i = x_i - y_i$, a big value for x will tend to cancel out a big value for y leading to little change in the value of z. So even if x and y had large variances, the variance in z could be quite small. We need to take into account how x and y are related to each other. That is what these extra terms do.

2. **Working with Time-series Univariate Data**

For this question, use the annual GDP data for the United States from 1900 to 2000 available on Smartsite. GDP stands for gross domestic product and is a measure of the total output of the economy. The data file is in the data folder and the filename is gdp-1900-2000.csv. The variables are the following:

YEAR: the year of the observation, ranging from 1900 to 2000

NOMINAL: the nominal GDP for the United States in that year in billions of dollars

REAL: the real GDP for the United States in that year in billions of dollars

(a) Nominal GDP is measured in billions of each year's current dollars. This means that nominal GDP changes due to both changes in output and inflation. Real GDP is measured in a billions of year 2000 dollars so that differences in real GDP correspond to just differences in output. Do you think the standard deviation in

nominal GDP will be larger or smaller than the standard deviation of real GDP? Explain your answer and then calculate both standard deviations to see whether you were right.

> See ps1-solutions.xlsx for calculations of the standard deviations for nominal and real GDP. It turns out the standard deviation of real GDP is larger than the standard deviation of nominal GDP. This may seem a bit odd. Real GDP moves around because of changes in actual output. Nominal GDP moves around both because of those changes in output and because of inflation so you would expect there to likely be more variation in nominal GDP. What is happening in this case is really an issue of the units we are using. Real GDP is being measured in year 2000 dollars. For all years except 2000, this means that real GDP is measured in units that take on larger values than nominal GDP. Because standard deviation is in the units of the variable, this is leading to a larger standard deviation for real GDP. If you calculate the coefficient of variation (these calculations are shown in ps1-solutions.xlsx), a unitless measure of dispersion, you find that by that measure there is substantially more variation in nominal GDP than in real GDP.

(b) Suppose that instead of measuring nominal GDP in billions of dollars, we measured it in trillions of dollars. We could do this by dividing GDP measured in billions of dollars by 1,000:

$$GDP_{trillions} = \frac{1}{1000} GDP_{billions}.$$

Use the definition of the sample standard deviation given in class, the summation rules from class and the standard deviation for $GDP_{billions}$ you calculated in part (a) to calculate the standard deviation of $GDP_{trillions}$. (Note: You don't need to calculate $GDP_{trillions}$ to answer this question.)

> Our new variable is just the old variable multiplied by a constant. So we are considering the following situation:
>
> $$z_t = a \cdot x_t$$
>
> where z_t is our new variable, x_t is our original variable and a is a constant. The standard deviation of z is given by:
>
> $$s_z = \sqrt{\frac{1}{n-1} \sum_{t=1}^{n} (z_t - \overline{z})^2}$$
>
> First, notice that $\overline{z}$ can be rewritten in terms of x:
>
> $$\overline{z} = \frac{1}{n} \sum_{t=1}^{n} z_t$$

$$\overline{z} = \frac{1}{n}\sum_{t=1}^{n} a \cdot x_t$$

$$\overline{z} = a \cdot \frac{1}{n}\sum_{t=1}^{n} x_t$$

$$\overline{z} = a \cdot \overline{x}$$

Now we can rewrite s_z in terms of x:

$$s_z = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(a \cdot x_t - a \cdot \overline{x})^2}$$

$$s_z = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n} a^2(x_t - \overline{x})^2}$$

$$s_z = \sqrt{a^2\frac{1}{n-1}\sum_{t=1}^{n}(x_t - \overline{x})^2}$$

$$s_z = a\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_t - \overline{x})^2}$$

Notice that this expression is simply the formula for the standard deviation of x multiplied by a:

$$s_z = a \cdot s_x.$$

So when we construct a new variable that is simply equal to another variable multiplied by a constant, the standard deviation of that variable is simply the standard deviation of the original variable multiplied by the constant. So the standard deviation of GDP measured in trillions will just be $\frac{1}{1000}$ times the standard deviation of GDP measured in billions:

$$s_{GDP_{trillions}} = \frac{1}{1000} s_{GDP_{billions}}$$

$$s_{GDP_{trillions}} = \frac{1}{1000} 2482$$

$$s_{GDP_{trillions}} = 2.482$$

(c) We are often interested in the growth rate of GDP. Create a new variable that gives the annual growth rate of real GDP over the past year. This can be calculated by taking the difference between the current year's real GDP and the previous year's GDP, dividing by the previous year's real GDP and multiplying by 100.

For example, the growth rate corresponding to the 1910 observation would be calculated as:

$$g_{1910} = 100 \cdot \frac{GDP_{1910} - GDP_{1909}}{GDP_{1909}} \tag{3}$$

See ps1-solutions.xlsx for the calculation of the growth rate using the formula above.

(d) Create a line chart for the growth rate of real GDP from 1901 to 2000, with year on the horizontal axis and growth rate on the vertical axis. Does it appear that the growth rate of real GDP has become more volatile or less volatile over time?

See ps1-solutions.xlsx for the line graph. From the graph, you can see that the fluctuations in the growth rate have gotten smaller over time (although the frequency of these fluctuations does not appear to have changed much).

(e) Economists often use a mathematical shortcut for calculating growth rates. For small changes in a variable, the percent change is approximately equal to the difference in natural logs. Using this approximation, we could calculate the growth rate for the 1910 observation as:

$$g_{1910} = 100 \cdot (ln(GDP_{1910}) - ln(GDP_{1909})) \tag{4}$$

Construct a new variable that calculates the growth rate of real GDP using this approach and another variable that shows the magnitude (absolute value) of the difference between the growth rate calculated in part (c) and the growth rate calculated in this part for each year.

See ps1-solutions.xlsx for the calculation of the new variables.

(f) Create a line chart showing both the growth rate of real GDP (use either one of your growth rates) and the difference between the two measures of the growth rate for 1901 to 2000 with year on the horizontal axis. Does the log approximation look like a good approximation of the growth rate? Does the approximation work better when growth rates are large or when they are small?

See ps1-solutions.xlsx for the graph. If you look at the graph, the largest deviations between the two ways of calculating the growth rate occur when the magnitude of the growth rate is at its largest. So the approximation seems to work better for smaller growth rates.