Review of Probability

ECON 4753 — University of Arkansas

Prof. Kyle Butts

1 — Summation

When working with data, it is common to have many observations. We want to have a notation that lets us work with all of those observations at once. Let's say we have some set of observations for some variable x. We can write them out as $x_1, x_2, \ldots, x_n$ where the subscript denotes which of the n observations we are referring too.

The first example of this is trying to sum up the value of a variable for all observations. We could write $x_1 + x_2 + \cdots + x_n$ to represent summing up all of the observations, but this is a lot of writing. We will use the $\sum$ notation (Σ is the greek capital letter S for "Sum"). In general, the notation will look like this:

$$\sum_{i=1}^{n} x_i \tag{1.1}$$

The first thing to notice is this subscript i. This is the 'iterator' variable and the sum notation says: Start i at 1 ($i = 1$ part) and count up by one until you reach n. The $\sum$ term says "sum up all n terms iterated by i. Last, x_i denotes what object to sum; in this case, sum the value of x for the i-th observation.

Example 1.1 (Sum of Squares)**.** Take $\sum_{i=1}^{5} i^2$. This says go from 1 to 5 and add the value of i^2.

$$\sum_{i=1}^{5} i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2$$

■

Example 1.2 (Sample Mean)**.** Say you go out to the quad and start recording people's ages. You observe the following ten people: $\{19, 20, 32, 19, 22, 40, 28, 30, 19, 21\}$. You want to calculate the sample mean of these observations which requires summing up the observations nad dividing by the number of observations (10).

We can write that as

$$\frac{1}{10}\sum_{i=1}^{10} \text{Age}_i = \frac{1}{10}\left(19+20+32+19+22+40+28+30+19+21\right) = 25$$

In general the mean is given by $\frac{1}{n}\sum_{i=1}^{n} x_i$. ■

1.1. Properties of summation

It will be useful to look at a few special cases where we know what the sum will be.

For any constant (number) c,

$$\sum_{i=1}^{n} c = n * c$$

For any constant a,

$$\sum_{i=1}^{n} ax_i = a * \sum_{i=1}^{n} x_i$$

Last, you can split up sums into parts:

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Putting them together, we have

$$\sum_{i=1}^{n} (a * x_i + b * y_i) = a * \sum_{i=1}^{n} x_i + b * \sum_{i=1}^{n} y_i$$

1.2. Applications for our class

Define $\bar{x} = 1/n \sum_{i=1}^{n} x_i$ to be our sample mean we discussed above. Let's work through the calculation of the *variance* of a variable. The variance is defined as

$$\text{var}(x) \equiv \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

We can use our rules above to simplify this a bunch

$$\begin{aligned}
\text{var}(x) &\equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2 * x_i * \bar{x} + \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2 * x_i * \bar{x} + \sum_{i=1}^{n} \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2 * \bar{x}\sum_{i=1}^{n} x_i + n * \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2 * \bar{x} * n * \bar{x} + n * \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 + n * \bar{x}^2\right),
\end{aligned}$$

where the second line follows from FOIL-ing the square, the third line comes from splitting sums, the fourth line from (i) pulling out constants and (ii) from summing the constant term $\bar{x}^2$, the fifth line from the definition of the sample mean, and the last line from simplifying terms.

Similarly, we can simplify the *covariance* between two variables:

$$\begin{aligned}
Cov(x, y) &\equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n * \bar{x} * \bar{y}\right)
\end{aligned}$$

1.3. Review Questions

1. Evaluate the following
   i. $\sum_{i=1}^{4}(i-2)$
   ii. $\sum_{i=1}^{4}(i-1)^2$
   iii. $\sum_{j=5}^{10} i$
2. Write the sample mean of the variable Height (in.) in summation notation. What is the sample mean of the following set of observations $\{68, 66, 67, 70, 65, 66\}$.

Figure 1 — Example Linear Equation for Expenditure

2 — Linear Equations

We are interested in predicting some variable y using some variable x. In general, we can write this relationship using a *function*, i.e. $y = f(x)$ for some function f. This section will focus on one of the most imporant functional forms: the linear equation.

The linear equation takes the form $y = \beta_0 + \beta_1 * x$. You may remember this from your algebra course as being $y = mx + b$ or $y = a + bx$. We use greek letters β in econometrics, but otherwise they are the same.

Here is an example. On average, we can calculate the average housing expenditure based on monthly income: Housing Expenditure $= 400 + 0.35 *$ Monthly Income. This is plotted in 1.

The first thing we can do is get predictions out for a specific income, say $\$2500$, by plugging it into our equation:

$$\text{Housing Expenditure} = 400 + 0.35 * 2500 = \$1275$$

We call these *fitted values* as you take the value of your explanatory variables and use the model fit to predict the value of the outcome variable (we will discuss this more later in the course).

Also, we can think about changing someone's income and seeing how the outcome variable changes. Say you change from x_1 to x_2, how does y change in response?

$$\begin{aligned} y_2 - y_1 &= (\beta_0 + \beta_1 * x_2) - (\beta_0 + \beta_1 * x_1) \\ &= \beta_1(x_2 - x_1). \end{aligned}$$

We can write this more succinctly as $\Delta y = \beta_1 * \Delta x$, where Δ (greek for D for "difference") takes the difference between the new and old values of the variable. Notice the slope plays an important role here. It tells us when you change x by a certain amount, then y changes by that amount *scaled* by β_1. So, if I increase x by one unit, then y changes by β_1 units.

In our housing example, if I increase my income by 500, then I change my housing expenditure by $0.35 * 500 = 175$ dollars.

Multiple variables

Say we have the following $y = \beta_0 + \beta_1 x_1 + \beta_2 + x_2$, where x_1 and x_2 are two explanatory variables. Using simlar math above, show that

$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_2$$

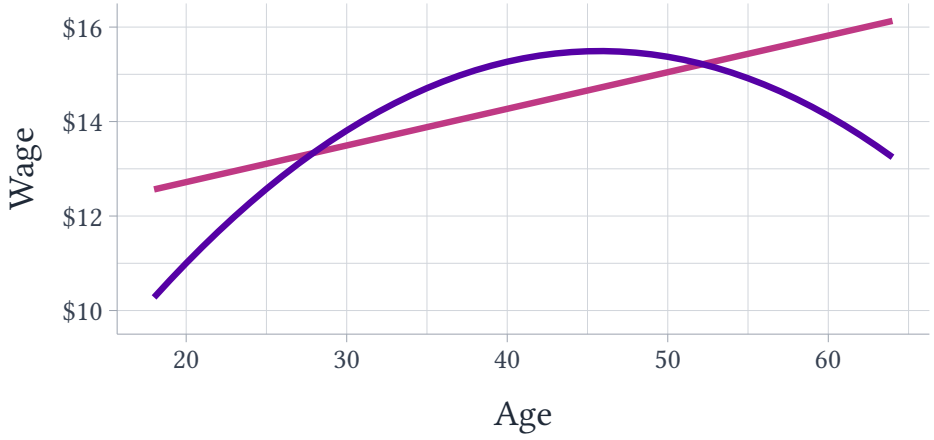
Say we want to examine how y changes when you change x_1 while holding x_2 equal (latin lesson, this is called 'ceteris palabus' or 'all else equal'). We have $\Delta x_2 = 0$, so $\Delta y = \beta_1 \Delta x_1$ just like before.

As an example, think about predicting how many dollars a person will spend on a product given (i) the price of the product p and (ii) the consumer's disposable income I. Let's say this is given by

$$Q = 120 - 9.8 * p + 0.03 * I$$

Say disposable income stays fixed at $\$900$ but the price increases from 8 to 9 dollars. How will

Figure 2 — Linear vs. Quadratic functions



the quantity demanded change?

$$\begin{aligned}\Delta Q &= -9.8 * \Delta p + 0.03 * \Delta I \\ &= -9.8 * 1 + 0.03 * 0 = -9.8\end{aligned}$$

The quantity demanded will decrease by 9.8 units.

3 — Quadratic Functions

Now say you model the relationship between x and y with a quadratic function

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

Adding polynomial terms allows the relationship between x and y to not be linear. See figure 2 for an example. The figure plots a linear relationship and quadratic relationship between age and wages. The quadratic equation better represents the facts that (i) workers are promoted more often at younger ages (quicker wage growth) and (ii) earnings peek around age 45 for workers.

The equation for wages is given by

$$\text{Wage} = -5.5 + 1 * \text{Age} - 0.01 * \text{Age}^2$$

Given this equation, how do wages change as a person ages? We could do some algebra using the method before (or using derivatives), but I'll show you the answer:

$$\Delta\text{Wage} = (1 - 2 * 0.01 * \text{Age})\, \Delta\text{Age}$$

The answer got more complicated, but that makes sense looking at the figure. Following the curve, the direction you move to stay on the line depends on where you are on the line (age). E.g. moving from 25 to 30, wages increase by $(1 - 2 * 0.01 * 25) * (30 - 25) = +\$2.50$. From 55 to 60, wages decrease by $(1 - 2 * 0.01 * 55) * (60 - 55) = -\$0.50$.

More generally, the formula is given by

$$\Delta y = \underbrace{(\beta_1 + 2 * \beta_2 * x)}_{\text{depends on starting } x} * \Delta x$$

4 — $\log$ transformations

Sometimes, we will talk about when in the course, it is beneficial to log either the outcome and/or explanatory variables. Log transformed variables change the interpretation of our linear regression lines. It takes from 'unit changes' to 'percent changes'. E.g. I could model $\log(\text{Wages})$ as a linear regression of age. My interpretation would be that increasing age by 1 year would increase wages by $\beta_1 * 100$ percent.

An example for how to interpret each combination of log and non-logged y and x variables is given in Table 1.

Table 1 — Summary of log-transformed linear equations

Model	Interpretation
$y = \beta_0 + \beta_1 x$	A 1 unit increase in x increases y by β_1 units
$\log(y) = \beta_0 + \beta_1 x$	A 1 unit increase in x increases y by $\beta_1 * 100$ percent
$y = \beta_0 + \beta_1 \log(x)$	A 1 percent increase in x increases y by β_1 units
$\log(y) = \beta_0 + \beta_1 \log(x)$	A 1 percent increase in x increases y by $\beta_1 * 100$ percent

For example, let's derive this for $y = \beta_0 + \beta_1 * \log(x)$. We have:

$$\begin{aligned} \Delta y &= \beta_1 \left(\log(x_2) - \log(x_1)\right) \\ &\approx \beta_1 * \frac{x_2 - x_1}{x_1} \\ &= \beta_1 * \%\Delta \text{ in } x \end{aligned}$$

,

where the approximation holds for small changes in Δx (think like a 1 percent increase). If you remember your high school science class, percent change is calculated (new - old / old) which is exactly what we have on the right hand side, the percent change in x.

This derivation will work the same for the other equations as summarized in Table 1. Just remember that if you have a log, then it is percent change.