ECON 20: Problem Set 1

1 Statistics review: Random variables

Let Y_1, Y_2, Y_3, Y_4 represent the values of some variable in a simple random sample of a population with mean μ and standard deviation σ. Recall this means the observations are independent and identically distributed, so $E[Y_1] = E[Y_2] = E[Y_3] = E[Y_4] = \mu$; $Var[Y_1] = Var[Y_2] = Var[Y_3] = Var[Y_4] = \sigma^2$; and $Cov(Y_1, Y_2) = Cov(Y_1, Y_3) = Cov(Y_1, Y_4) = Cov(Y_2, Y_3) = Cov(Y_2, Y_4) = Cov(Y_3, Y_4) = 0$. For short, we would typically write this as $E[Y_i] = \mu$, $Var[Y_i] = \sigma^2$, and $Cov(Y_i, Y_j) = 0 \ \forall i \neq j$.

1. Show that the sample average $\bar{Y} = \frac{1}{4} \cdot (Y_1 + Y_2 + Y_3 + Y_4)$ is an unbiased estimator for μ; that is, show that $E[\bar{Y}] = \mu$.

2. Show that the weighted average $W = \frac{1}{8} \cdot Y_1 + \frac{1}{8} \cdot Y_2 + \frac{1}{4} \cdot Y_3 + \frac{1}{2} \cdot Y_4$ is an unbiased estimator for μ.

3. Calculate the variance of both $\bar{Y}$ and W.

4. Why do we prefer the sample average ($\bar{Y}$) as a way to estimate the mean even though both $\bar{Y}$ and W are unbiased?

5. When and why might be prefer some kind of weighted average? (*Hint*: not when we are using a simple random sample)

6. The square root of the variance of $\bar{Y}$ could be called the standard deviation of $\bar{Y}$ but it more commonly goes by another name. What is that name? Calculate it. Notice that it is smaller than the standard deviation of Y.

7. Let w_1, w_2, w_3, w_4 be weights (just numbers, not random variables) for a different weighted average $W = w_1Y_1 + w_2Y_2 + w_3Y_3 + w_4Y_4$. What must be true of these weights for W to be unbiased?

2 Partial derivatives

Remember how to take derivatives? If $y = x^2$, then $\frac{\partial y}{\partial x} = 2x$. If $y = \log(x)$, then $\frac{\partial y}{\partial x} = \frac{1}{x}$. The *partial* derivative, denoted with the ∂, is nothing more than a derivative which ignores all variables not involved in the derivative being taken (i.e., it treats them as constants). For example, if $y = zx^2 + z$, then $\frac{\partial y}{\partial x} = 2zx$ while $\frac{\partial y}{\partial z} = x^2 + 1$. See if you can do these problems.

1. Let $health = -2 \cdot age - 3 \cdot age^2 + 4 \cdot income$. What is $\frac{\partial health}{\partial age}$?
2. Let $health = \beta_0 + \beta_1 age \cdot \log(income) + \beta_2 Age + \beta_3 \log(income) + \epsilon$. What is $\frac{\partial health}{\partial age}$?
3. In the same equation as above, what is $\frac{\partial health}{\partial \log(income)}$?
4. Let $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ \cdot black + u$. What is $\frac{\partial \log(wage)}{\partial educ}$?
5. In the same equation as above, suppose that the variable *black* takes only two values, 0 and 1. What is $\frac{\partial \log(wage)}{\partial educ}$ when $black = 0$? And when $black = 1$?

3 Do rising health care costs reduce earnings?

Rapidly rising health care costs have been a prominent political issue for decades. Since at least the 1970's, health care costs have increased faster than inflation. At the same time, wage growth in the U.S. has been relatively slow since the 1970's. Since payment for health insurance is typically through an employer in the U.S., concerns have been raised that perhaps rising health care costs have crowded out other forms of compensation.

To analyze this question, we introduce a new concept, the variance-covariance matrix. For a list of variables, this matrix expressses the covariance between all pairs of these variables (including variables with themselves which is the variable's own variance). For example, for two variables X_1 and X_2, the matrix would be:

$$\begin{bmatrix} var(X_1) & cov(X_2, X_1) \\ cov(X_1, X_2) & var(X_2) \end{bmatrix}$$

Notice that the covariance in the right-corner is redundant since covariance is symmetric, so sometimes it is left out.

Now to the question at hand. To get preliminary evidence on whether rising health care costs may have reduced wage growth, you obtain yearly (time-series) data from the *Economic Report of the President* on growth in wages and health care costs. Their variance-covariance matrix is

$$\begin{bmatrix} 2.69322 & \\ -1.31575 & 1.72842 \end{bmatrix}$$

where the first row/column is $\%\Delta wage$ and the second row/columnn is $\%\Delta health\ spending$.

1. Provide an OLS estimate of the slope of a regression of growth in wages on growth in health care spending,

2. Provide an OLS estimate of the slope of a regression of the reverse regression: growth in health care spending on growth in wages.

3. Is the following statement true? *By estimating the reverse regression, we show that not only do rising health care costs cause wages to grow more slowly, but that slow wage growth has contributed to rising health care costs in the U.S.*

4. Do you think the question "did rising health care costs contribute to slow wage growth in the US?" is a well-posed (*ala* Dinardo 2007)? Choose yes or no and briefly justify your answer.

5. What is the R^2 of the regressions in (a) and (b)?

4 Health insurance and inflation

A common theory is that health insurance causes health care inflation, because people who have health insurance are insensitive to the price of health care. One way to investigate this would be to ask if health spending of those who have health insurance is higher than those who do not. Let Y_{i1} be the health care spending of person i if they have health insurance and Y_{i0} be the health care spending of person i if they do not. Let $T_i \in \{0, 1\}$ be a health insurance indicator, i.e. $T_i = 1$ if person i has health insurance and $T_i = 0$ otherwise.

1. In symbols, write down the estimator that compares average health care spending of those who do and do not have health insurance.

2. In symbols, what needs to be true for the estimator above to be the causal effect of health insurance on health care spending?

3. Say your answer to (ii) in words.

4. Do you think the question "Does having health insurance cause people to spend more on health care?" is well-posed? Choose yes or no and briefly explain your answer.

5 Multiple Regression Basics

Select all true answers. There may be more than one (or none).

1. The independent variable is on the one which...
   (a) Is on the right-hand side of the estimating equation.
   (b) Is also called the explanatory variable because it explains or causes the outcome.
   (c) Both
2. That the residual and the right-hand side variable are uncorrelated is ...
   (a) An untestable assumption.
   (b) Something which is always true.
3. What can be said about the estimated slope coefficient for a regression of Y on X versus the estimated slope coefficient for a regression of X on Y?
   (a) The slopes are reciprocal.
   (b) The slopes are the same sign.
   (c) The slopes are negative reciprocals of each other.
   (d) The slopes are identical.
4. Adding controls to a regression...
   (a) Never decreases the R-squared.
   (b) Means we interpret the slope "holding the controls constant."
5. The mean squared error (MSE) measures of the variance of the...
   (a) Residuals.
   (b) Dependent variable.
   (c) Independent variable.
6. The higher the R^2...
   (a) The more variation in the left-hand side variable that is "explained" (accounted for) by the right-hand side variables.
   (b) The higher is the mean-squared error.

(c) The more correlated is X with Y in a bivariate regression.

(d) The more correlated is Y with X in a bivariate regression.

(e) The more likely are Y and X to have a causal relationship.

7. Suppose you get some data on infant weights at birth and mother's smoking habits during pregnancy. You estimate a linear model $birthweight = \beta_0 + \beta_1 cigarettes + u$ via OLS (where *cigarettes* is the daily number of cigarettes the mother smoked while pregnant). You estimate that the average birth weight is 140oz and the average mother smoked 2 cigarettes per day. The sample size is 2,488 mothers. If your estimate of β_1 is -2.5, what is your estimate of β_0?

   (a) 140

   (b) 145

   (c) 135

   (d) Other

   (e) Not enough information

8. Intepret the slope, β_1, from the regression above.