Economics 475: Econometrics
Homework #1: Answers

This homework is due on January 11th.

This homework has two purposes: one is to remind (or introduce to) you the rules of mathematic expectation. The other is to have you remember how to use Stata (or whatever software you choose).

If x is discrete, the expected value of x (denoted E[x]) is: $E[x] = \Sigma(x \times f(x))$ where x is the value of a random variable and f(x) is the probability of achieving that value of the random variable. Notice, that you have been doing this for a long time already. For instance, the expected value of a fair, 6 sided die is 3.5 (one-sixth of the time you get a one, one-sixth of the time you get a two, …).

The following definitions are a summary of the rules of mathematical expectation more fully explained at: http://www.cbe.wwu.edu/krieg/Econ475/Mathematical%20Expectation.pdf

1) The expectation of a random variable is the random variable's mean. In other words $E[X] = \mu_X$.
2) The expectation of a constant is that constant. In other words E[a] = a where a is a constant.
3) I write the variance of a random number X as V[X] or Var[X]. By definition $V[X] = \sigma_X^2 = E[(X - E[X])^2]$. Remember, the variance formula is $\frac{\sum(X-\bar{X})^2}{n}$ which is really taking the average of the squared deviations of X from its mean—exactly what $E[(X - E[X])^2]$ says!
4) Like rule #3, the covariance between two random variables X and Y are typically written as Cov[X,Y] = E[(X-E[X])(Y-E[Y])].
5) The expectations operator E is a linear operator, that is if Z = X + Y then E[Z] = E[X] + E[Y].
6) Likewise E[aX] = aE[X] if a is a constant and X is a random variable.
7) Finally, E[XY] where both X and Y are random variables is equal to E[X]×E[Y] only if X and Y are independent (that is Cov[X,Y] = 0).

1. Here are some paired observations of two random variables:

Observation	X	Y
1	5	3
2	6	5
3	9	12
4	1	-4
5	2	0

Find E[X], E[Y], V[X], V[Y], E[XY], and Cov[XY] and describe in words what each of these are.

E[X] = (5+6+9+1+2)/5 = 4.6.
E[Y] = (3+5+12-4+0)/5 = 3.2
E[X] and E[Y] are the mathematical expectations of X and Y or, more commonly known as the expected values of the random variables X and Y, or more simply put their (sample) means.

$V[X] = E[X^2] - (E[X])^2$ (we didn't use this above—can you prove that this is equal to the above expression for V[X])? $E[X^2] = (5^2 + 6^2 + 9^2 + 1^2 + 2^2)/5 = 29.4$. $(E[X])^2 = 4.6^2 = 21.16$ so $V[X] = 29.4 - 21.16 = 8.24$. Likewise, V[Y] = 28.56. Both are measures of the variance which is most easily understood as the average of the squared deviations of observations from their means.
$E[XY] = (5\times3+6\times5+9\times12+1\times-4+2\times0)/5 = 29.8$ E[XY] is the average of the product of X and Y (note if X and Y are negatively correlated $E[XY]<0$ and vice versa).
The Covariance between X and Y is defined as $E[XY] - E[X]\times E[Y]$}. In this case the covariance is $29.8 - 4.6\times3.2 = 15.08$.
Note, if you used Excel (or similar program) to find these, you will get different answers. Why?

2. One can show that when E[X] = 0 then the variance of X is equal to $E[X^2]$. Prove this (and don't forget it—we will see it over and over again)!
$\sigma_x^2 = E[(X - E[X])]^2 = E[X^2 - 2\times X\times E[X] + (E[X])^2] = E[X^2] - 2\times E[X]\times E[X] + (E[X])^2 = E[X^2] - (E[X])^2 = E[X^2] - (0)^2 = E[X^2]$

3. Use the facts that X is a random variable with mean 4 and variance 8, Y is a random variable independent of X with mean 3 and variance 5 to answer the following questions:
a. What is E[X] and E[Y]?
4 and 3.

b. Think of a random variable $Z = 5+ 2X + Y$. What is E[Z] and V[Z]?
$V[Z] = E[Z - E[Z]]^2 = E[Z - \mu_Z]^2 = E[Z^2] - 2\mu_Z E[Z] + \mu_Z\mu_Z = E[(5+ 2X + Y)^2] - 2\times16\times16 + 16^2 = E[4X^2 + Y^2 + 20X + 10Y + 4XY +25] - 256 = 4E[X^2] + E[Y^2] + 20\times4 + 10\times3 + 4\times12 + 25 - 256 = 4E[X^2] + E[Y^2] - 73$. So, to solve this we must determine what the $E[X^2]$ and $E[Y^2]$ is. To do this, remember the equation for the variance:

$V[X] = E[X-E[X]]^2$ $=E[X^2] - 2\mu_X E[X] + \mu_X\mu_X$ $=E[X^2] - \mu_X\mu_X$ $= E[X^2] - 4\times4$ Inserting V[X] = 8 $8= E[X^2] - 4\times4$ $E[X^2] = 24$	$V[Y] = E[Y-E[Y]]^2$ $=E[Y^2] - 2\mu_Y E[Y] + \mu_Y\mu_Y$ $=E[Y^2] - \mu_Y\mu_Y$ $= E[Y^2] - 3\times3$ Inserting V[Y] = 5 $5= E[Y^2] - 3\times3$ $E[Y^2] = 14$

Thus, the variance of Z is $V[Z] = 4\times24 + 14 - 73 = 37$.

c. How would your answer to b change if there was a positive correlation between X and Y rather than X and Y being independent? Specifically, what happens to V[Z] as the correlation between X and Y becomes closer to 1? Why?
Solving this problem is the same until this step: $E[4X^2 + Y^2 + 20X + 10Y + 4XY +25] - 256$
If X and Y are uncorrelated, the $4E[XY] = 4\times E[X]\times E[Y]$. However, if the correlation is other than zero, then $4E[XY] \neq 4\times E[X]\times E[Y]$. The covariance of X and Y, by definition, is
$E[(X-E[X](Y-E[Y])]=E[XY]-\mu_Y E[X]- \mu_X E[Y]+E[X]E[Y] = E[XY] - \mu_X\mu_Y$. Thus, $Cov(X,Y) = E[XY] - \mu_X\mu_Y$. If the covariance is zero, then it must be true that $E[XY] = \mu_X\mu_Y$. If the covariance is positive, as in this case, then $E[XY] > \mu_X\mu_Y$. In our case, this means that E[XY] is greater than 12. Examining our variance equation in part b, adjusting the term E[XY] such that it is greater than 12 would indicate that the variance is greater than 37.

4. The data set "freshmen 2002 data" consists of observations of each freshmen (new students and running start students) who began college in 2002. Use this data set to answer the following questions.

a. Estimate the regression $gpa_i = B_0 + B_1 hsgpa_i + \varepsilon_i$. Does high school GPA impact first-quarter GPA? How do you interpret B_1?

I find:

```
. reg gpa hsgpa
```

Source	SS	df	MS	Number of obs	=	2,081
				F(1, 2079)	=	472.50
Model	195.606722	1	195.606722	Prob > F	=	0.0000
Residual	860.667118	2,079	.413981298	R-squared	=	0.1852
				Adj R-squared	=	0.1848
Total	1056.27384	2,080	.507823962	Root MSE	=	.64341

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsgpa	1.001795	.0460869	21.74	0.000	.9114138	1.092176
_cons	-.7431143	.1628574	-4.56	0.000	-1.062495	-.4237337

The interpretation of B_1 is that a unit change in high school GPA raises college GPA by 1.001 points.

b. I remember high school counselors telling me that the average college student earned a college GPA one point lower than their high school GPA. Using the regression in part a, test if this is true.

This happens only if $B_0 = -1$ and $B_1 = 1$. To test this, I construct a restricted model where I force these two conditions to be true. Mathematically, this is equivalent to $gpa_i - hsgpa_i + 1 = \varepsilon_i$. I compute the sum of squared residuals from this restricted model which, by definition is simply $(gpa_i - hsgpa_i)^2$. The sum of this turns out to be 1004.8. The resulting F-Test (using the RSS from part a) is $\frac{(1004.8-860.66)/2}{860.66/2079} = 174$. The $F_{c,2,2079,5\%} = 3.0001$. Thus, we reject the null hypothesis and conclude that the joint hypothesis of $B_0 = -1$ and $B_1 = 1$ is not true.

c. Estimate the regression $gpa_i = B_0 + B_1 hsgpa_i + B_2 satverb_i + B_3 satmath_i + B_4 male_i + B_5 firstgen_i + \varepsilon_i$. How do you interpret B_1? How does this differ from your answer in part a?

I find:

```
. reg gpa hsgpa satverb satmath male firstgen
```

Source	SS	df	MS	Number of obs	=	2,081
				F(5, 2075)	=	151.38
Model	282.314699	5	56.4629398	Prob > F	=	0.0000
Residual	773.959141	2,075	.372992357	R-squared	=	0.2673
				Adj R-squared	=	0.2655
Total	1056.27384	2,080	.507823962	Root MSE	=	.61073

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsgpa	.8154022	.0471343	17.30	0.000	.7229667	.9078376
satverb	.0019668	.0001961	10.03	0.000	.0015821	.0023514
satmath	.0008611	.0002145	4.01	0.000	.0004405	.0012818
male	-.1309339	.0298123	-4.39	0.000	-.189399	-.0724689
firstgen	-.0674227	.0287948	-2.34	0.019	-.1238924	-.010953
_cons	-1.566911	.178332	-8.79	0.000	-1.916639	-1.217183

We interpret B_1 differently in c than in a because of the inclusion of the other variables. Specifically, this regression tells us that a unit change in high school GPA raises college GPA by .81 points holding constant SAT scores, gender, and first generation status.

d. Does the SAT test help predict first-quarter GPA?

Because two SAT scores are included in the regression of Part C, I interpret this question to ask, "Do satverb and satmath jointly explain gpa?" To answer this, I create a restricted regression where I exclude satverb and satmath (essentially restricting their coefficients to be jointly zero), and then perform an F-test. My restricted regression is:

```
. reg gpa hsgpa male firstgen
```

Source	SS	df	MS	Number of obs	=	2,081
				F(3, 2077)	=	167.95
Model	206.21345	3	68.7378167	Prob > F	=	0.0000
Residual	850.06039	2,077	.409273178	R-squared	=	0.1952
				Adj R-squared	=	0.1941
Total	1056.27384	2,080	.507823962	Root MSE	=	.63974

gpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsgpa	.9740823	.0473937	20.55	0.000	.8811381	1.067026
male	-.0702797	.0295334	-2.38	0.017	-.1281978	-.0123617
firstgen	-.1404166	.0296835	-4.73	0.000	-.1986291	-.0822041
_cons	-.5675217	.1716375	-3.31	0.001	-.9041211	-.2309223

My resulting F-test is: $\frac{(850.06-773.95)/2}{773.95/2079} = 102$. As we saw above, the critical value of the F-statistic is 3, so we reject the null hypothesis that the coefficients on satverb and satmath are zero and conclude that SAT scores do matter.

e. I notice the coefficient on male is negative. Offer some explanations for this. Do any of your explanations violate the classical assumptions of OLS?

Cov(Male,ε) is unlikely to be zero.