

Problem Set 3, ECON 30331
(Due at the start of class, Tuesday, September 22, 2009)

Dan Hungerman
Fall 2009

1. A researcher estimates a bivariate regression of the form $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ but confides to a colleague that she believes $\text{cov}(\varepsilon_i, x_i) \neq 0$ and therefore, $\hat{\beta}_1$ is a biased estimate. The colleague then asks whether one can test whether $\text{cov}(\varepsilon_i, x_i) \neq 0$. The colleague suggests that the researcher construct $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - x_i\hat{\beta}_1$ then run a regression of $\hat{\varepsilon}_i$ on x_i , that is, a regression of the form $\hat{\varepsilon}_i = \gamma_0 + x_i\gamma_1 + v_i$, then test the null $H_0: \gamma_1 = 0$ to see whether ε_i and x_i are correlated. Is this a good idea or not?

HINT: The OLS estimate of $\hat{\gamma}_1$ would be
$$\hat{\gamma}_1 = \frac{\sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

2. Consider a multivariate regression model of the form $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + u_i$. Write the 1st order conditions for the optimization problem where one is interested in minimizing the sum of squared errors (SSE) $\sum_{i=1}^n \hat{\varepsilon}_i^2$.

Suppose in a sample of 25 observations, the following facts are presented about the model above.

$$\sum_{i=1}^n x_{1i}^2 = 20 \quad \sum_{i=1}^n x_{2i}^2 = 40 \quad \sum_{i=1}^n x_{1i}x_{2i} = 0 \quad \sum_{i=1}^n x_{1i} = 0 \quad \sum_{i=1}^n x_{2i} = 0 \quad \sum_{i=1}^n y_i = 0$$

$$\sum_{i=1}^n x_{1i}y_i = 60 \quad \sum_{i=1}^n x_{2i}y_i = 80$$

Using the first order conditions (or normal equations) and these facts, provide the estimates for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. HINT: Solve for $\hat{\beta}_0$ first.

3. Consider a multivariate regression model of the form $y_i = \beta_0 + x_{1i}\beta_1 + u_i$. Suppose the R^2 from this model is R_a . True, False, or Uncertain and explain. The R^2 can never fall below R_a when additional variables are added to the model? (Think of a special case where someone adds completely irrelevant variables to the model – what will happen to the R^2 ?)
4. On the class web page is a data set law_school_1985.dta that has information about average starting salaries of graduates from 95 law schools in 1985. The data set has four other key variables
- cost average annual tuition
 - lsat average lsat of class of 1985
 - rank rank of law school according to national survey, 1=best
 - age age of law school in years

Load the data set into STATA, then construct two new variables:

```
gen lcost=ln(cost)
gen lsalary=ln(salary)
```

- a) Regress `lsalary` on four variables: `lcost`, `lsat`, `rank` and `age`. Interpret the coefficient on `lcost` – provide a numeric example of the magnitude of the coefficient on this variable?
- b) Interpret the coefficient on `rank` -- provide a numeric example of the magnitude of the coefficient on this variable?
- c) Right after the regression statement, generate a new variable that is the predicted residuals from the regression in part a) and call this variable `res1`. This can be done by using the STATA statement

```
predict res1, residual
```

Next, obtain the correlation coefficient between `res1` and two of the explanatory variables: `lsat` and `lcost`. Do the correlation coefficients between `res1` and `lcost` and `res1` and `lsat` make sense?

You can use the following statement to get the correlation coefficients

```
corr res1 lcost lsat
```

- d) Next, output the predicted values from the regression and call this variable `pred1`. This can be done with the following statement

```
predict pred1, xb
```

With the predicted values get the correlation coefficient between the actual `y` and the predicted `y`

```
corr lsalary pred1
```

Square the correlation coefficient and compare to the R^2 from the original model. What is the relationship between the correlation coefficient squared and the R^2 from the original model?

- e) Returning to the model in part a), estimate a similar model but now delete the variable “rank”. What happens to the coefficient on “lsat” when this variable is deleted? Provide an intuitive explanation for why the coefficient on “lsat” changed when “rank” was deleted.

5. On the class web page is a STATA data set called `house_price.dta`. It has data on 114 homes sold in 1998 in a small town in New England. The data set contains information on the sales price of the house (measured on thousands of dollars), the number of bedrooms, bathrooms, other rooms, square feet of living space and age of the home,

Download the data and initially estimate a regression with house prices as the outcome of interest and four covariates: age in years, # number of bedrooms, # of bath rooms, # of other rooms. Call this model 1. Interpret the coefficient on age in years and # of bedrooms and provide a numeric example.

Now, estimate a second model and add to the original regression the square feet of living space. Call this model 2. What happens to the coefficient on # of rooms, # of bedrooms and # of other rooms in this new model compared to the previous one? Why have the coefficients on these three variables changed so dramatically? Interpret the coefficient on square feet of living space.

Now estimate a third model with the same dependent variable but include only two covariates: age in years and square feet. Compare the R^2 from this model and that in Model #2. Is there much of a difference? Provide an intuitive explanation for why the difference is so small.

6. Consider a multivariate regression of the form $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + u_i$ and we know that

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{(1 - R_1^2) \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$$

Where R_1^2 is the R^2 from a regression of x_{1i} on x_{2i} . Answer the following questions.

- Suppose that x_{1i} is a linear combination of x_{2i} where $x_{1i} = a + bx_{2i}$. What is $\text{Var}(\hat{\beta}_1)$ in this case?
- Suppose instead that x_{1i} and x_{2i} are correlated but R_1^2 is very high (like 0.999). What happens to $\text{Var}(\hat{\beta}_1)$ when a highly correlated variable is added to the model?

7. Return to problem 7 on problem set 2. A pharmaceutical company is investigating the cholesterol lowering benefits of a new drug. In a sample of n subjects the company randomly assigns milligrams of active ingredients (label this as x_{1i}) and the outcome of interest, labeled as y_i , is the change in cholesterol from the start until the end of the trial. Initially, the researchers estimate a model of the form $y_i = \beta_0 + x_{1i}\beta_1 + u_i$. However, a colleague mentions that as part of the experiment, they also collected detailed data on characteristics of survey participants that predict y_i like their weight at the start of the trial, age, sex, ethnicity/race, plus other variables. The colleague asks whether one should include these covariates (label them as $x_{2i}, x_{3i}, \dots, x_{ki}$) into the basic regression?

- By estimating a model of $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \dots + x_{ki}\beta_k + u_i$, do you anticipate that the estimate on $\hat{\beta}_1$ will change?

- In a multivariate model, the estimated variance of $\hat{\beta}_1$ is given as $\hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}_\varepsilon^2}{(1 - R_1^2) \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$

What is the likely consequence of adding these additional covariates ($x_{2i}, x_{3i}, \dots, x_{ki}$) to the estimated variance of $\hat{\beta}_1$?

8. On the next page are the results from two regression models: In model (1), I regress Y on X_1 , and note that the standard error on the coefficient on X_1 is very small and the t-statistic on the coefficient on $\hat{\beta}_1$ is over 23. Note that in model (2), when I add X_2 to the model, the standard error on $\hat{\beta}_1$ increases by a factor of 3 and the t-statistic on this parameter falls to 1.39. Using the information given, provide an intuitive explanation for why the standard error increases so much on $\hat{\beta}_1$ when X_2 is added to the model.

Results for Question 8

Correlation between X_1 and X_2

```
. corr x1 x2
(obs=2489)
```

	x1	x2
x1	1.0000	
x2	0.9994	1.0000

Model 1: Regression of Y on X_1

```
. reg y x1
```

Source	SS	df	MS	Number of obs =	2489
Model	121.044173	1	121.044173	F(1, 2487) =	562.63
Residual	535.054756	2487	.215140634	Prob > F =	0.0000
Total	656.098929	2488	.263705357	R-squared =	0.1845
				Adj R-squared =	0.1842
				Root MSE =	.46383

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x1	.0765488	.0032272	23.72	0.000	.0702205	.0828771
_cons	5.059357	.0435541	116.16	0.000	4.973951	5.144763

Model 2: Regression of Y on X_1 and X_2

```
. reg y x1 x2
```

Source	SS	df	MS	Number of obs =	2489
Model	121.115811	2	60.5579054	F(2, 2486) =	281.41
Residual	534.983118	2486	.215198358	Prob > F =	0.0000
Total	656.098929	2488	.263705357	R-squared =	0.1846
				Adj R-squared =	0.1839
				Root MSE =	.46389

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x1	.1304861	.0935397	1.39	0.163	-.0529375	.3139098
x2	-.0539557	.0935159	-0.58	0.564	-.2373328	.1294213
_cons	5.059738	.0435649	116.14	0.000	4.974311	5.145166