Suggested Answers, Problem Set 3
ECON 30331

Bill Evans
Spring 2018

1. The following short program will generate the results for this question

```
log using ps3_q1.log, replace
use state_cig_data

*generate ln real prices
gen ln_r_p=ln(retail_price/cpi)

* generate ln quantities
gen ln_q=ln(packs_pc)

* get regression estimates
reg ln_q ln_r_p
predict errors, residuals

sum error

corr error ln_r_p

log close
```

The results are as follows

```
      Source |       SS       df       MS              Number of obs =    1020
-------------+------------------------------           F(  1,  1018) =  873.39
       Model |  36.0468802     1  36.0468802           Prob > F      =  0.0000
    Residual |  42.0153163  1018  .041272413           R-squared     =  0.4618
-------------+------------------------------           Adj R-squared =  0.4612
       Total |  78.0621965  1019   .07660667           Root MSE      =  .20316

------------------------------------------------------------------------------
        ln_q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ln_r_p |  -.8076941   .0273302   -29.55   0.000    -.8613241   -.7540641
       _cons |   8.834473   .1423221    62.07   0.000     8.555195    9.113751
------------------------------------------------------------------------------

. predict errors, residuals

.
. sum error

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      errors |      1020    2.92e-10    .2030564  -.7283315   .8014919

.
. corr error ln_r_p
```

```
(obs=1020)

             |   errors   ln_r_p
-------------+------------------
      errors |   1.0000
      ln_r_p |   0.0000   1.0000
```

a. $\hat{\beta}_1$ =-0.807, $\hat{\beta}_0$ = 8.83, R^2=0.46.

b. The model is of the form $\ln(q_i)=\beta_0+\ln(p_i)\beta_1+\varepsilon_i$ so $\beta_1=\dfrac{\partial\ln(q)}{\partial\ln(p)}=\dfrac{\frac{\partial q}{q}}{\frac{\partial p}{p}}=\dfrac{\%\ change\ in\ q}{\%\ change\ in\ p}=-0.81$

The parameter $\hat{\beta}_1$ is therefore an elasticity. A 10% increase in price will generate an 8.1 percent reduction in consumption.

c. Note that the sample mean of the errors is zero $\bar{\hat{\varepsilon}}=(1/n)\sum_{i=1}^{n}\hat{\varepsilon}_i$ =0

d. The sample correlation between the estimated error and x is also zero.

1. Look at the following results

```
. keep if year==1985
(969 observations deleted)

. reg packs_pc federal_tax
note: federal_tax omitted because of collinearity

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  0,    50) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |  24487.9416    50  489.758831           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  24487.9416    50  489.758831           Root MSE      =   22.13

------------------------------------------------------------------------------
    packs_pc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 federal_tax |          0  (omitted)
       _cons |   120.5275   3.098889    38.89   0.000     114.3031    126.7518
------------------------------------------------------------------------------

sum federal_tax

Variable      Obs    Mean   Std.   Dev.   Min    Max

federal_tax   51     16            0      16     16
```

The model cannot be estimated because there is no variation in the federal tax across states in 1985. Residents of CA, LA, IN, etc., all pay exactly the same tax. Look at the descriptive statistics for federal_tax – the standard deviation is zero. Recall the definition of the OLS estimate

$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. In this case x does not vary across i so $x_i = x = \bar{x}$ which means that

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$ and the model cannot be estimated.

2. If y is the teacher's evaluation and x is the average class grade, then if we estimate the model $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$, the estimated residual $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 + x_i\hat{\beta}_1$ represents the component of teaching quality that is not predicted by how easilya teacher grades. If $\hat{\varepsilon}_i$ >0 then a teacher is over-performing given how hard they grade and if $\hat{\varepsilon}_i$ <0 then they are under-performing given how hard they grade. Given this interpretation of the estimated error, the Dean could simply use $\hat{\varepsilon}_i$ as the basis for teaching awards.

3. The true model is $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ and the estimate for β_0 is $\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$. To figure out the properties of the estimate, substitute in the truth. In this case, given the true model, the truth can be characterized as $\bar{y} = \beta_0 + \bar{x}\beta_1 + \bar{\varepsilon}$ and therefore $\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = \beta_0 + \bar{x}\beta_1 + \bar{\varepsilon} - \bar{x}\hat{\beta}_1$

Taking expectation of both sides, we get

$$E\left[\hat{\beta}_0\right] = E\left[\beta_0\right] + E\left[\bar{x}\beta_1\right] + E\left[\bar{\varepsilon}\right] - E\left[\bar{x}\hat{\beta}_1\right]$$

There are four terms on the right hand side. First, not that β_1 is a constant so by definition, $E\left[\beta_0\right] = \beta_0$. Likewise, $\bar{x}$ and β_1 are constants so $E\left[\bar{x}\beta_1\right] = \bar{x}\beta_1$. Third, recall that $\bar{x}$ is fixed but $\hat{\beta}_1$ is a random variable. In class however, we demonstrated that $\hat{\beta}_1$ is an unbiased estimate of β_1 so

$$E\left[\bar{x}\hat{\beta}_1\right] = \bar{x}E\left[\hat{\beta}_1\right] = \bar{x}\beta_1$$

Finally, note that $\bar{\varepsilon} = \left(\dfrac{1}{n}\right)(\varepsilon_1 + \varepsilon_2 + ...\varepsilon_n)$ and therefore

$$E[\bar{\varepsilon}] = \left(\frac{1}{n}\right)\left(E[\varepsilon_1] + E[\varepsilon_2] + ...E[\varepsilon_n]\right)$$

By assumption, $E[\varepsilon_i]=0$ for all i, so $E[\bar{\varepsilon}] = 0$. Substituting these values into the original equation

$$E\left[\hat{\beta}_0\right] = E\left[\beta_0\right] + E\left[\bar{x}\beta_1\right] + E\left[\bar{\varepsilon}\right] - E\left[\bar{x}\hat{\beta}_1\right] = \beta_0 + \bar{x}\beta_1 - \bar{x}\beta_1 = \beta_0$$

Therefore, since $E\left[\hat{\beta}_0\right] = \beta_0$, $\hat{\beta}_0$ is an unbiased estimate.

4. We know that we can write the estimate of $\hat{\beta}_1$ as $\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} \varepsilon_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \beta_1 + \frac{\hat{\sigma}_{x\varepsilon}}{\hat{\sigma}_x^2}$

And therefore, whether $\hat{\beta}_1$ is unbiased is a function of the expected correlation between x and ε. Ask the questions, does the realization of ε convey any information about the likely value of x? In this case, one can easily argue that we would expect the popular songs are also the most downloaded songs –so we would expect $\hat{\sigma}_{x\varepsilon} > 0$ and therefore we expect $E[\hat{\beta}_1] > \beta_1$ so in this case, the estimated parameter is biased and overstates the impact of downloads on sales.

5. In this case, x is dosage and y is the change in cholesterol levels. $\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} \varepsilon_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \beta_1 + \frac{\hat{\sigma}_{x\varepsilon}}{\hat{\sigma}_x^2}$.

Ask the questions, does the realization of ε convey any information about the likely value of x? In this case, the answer is clearly no. X is determined by chance– some people assigned a high dosage, some a low dose, and some none at all. But, the important fact is that dosage is determined randomly so we expect that on average it will not be correlated with ε. In this case, $\hat{\beta}_1$ is an unbiased estimate.

6. Remember that in a regression model, we want x to be exogenous (fixed) and y to be endogenous (the choice variable). We teach in intermediate micro that BOTH prices and output are determined by the market so how can one be considered fixed? Because of the fact that prices are also determined by market, tt is unlikely you are getting the right estimate for $\hat{\beta}_1$. Let's start with the basic framework where we know that $\hat{\beta}_1 = \beta_1 + \frac{\hat{\sigma}_{x\varepsilon}}{\hat{\sigma}_x^2}$. Does the realization of ε_i reveal anything about x_i (price)? Consider an area like New York or San Francisco which has high demand $(\varepsilon_i > 0)$. Is that reflected in prices? Well, it most likely means that prices are also higher than average ($x_i > \bar{x}$). In contract, in Warren, OH where demand is low $(\varepsilon_i < 0)$ we would expect price to be lower $x_i < \bar{x}$. Therefore $\hat{\sigma}_{x\varepsilon}$ is potentially positive and hence $\hat{\beta}_1$is biased up. Other outcomes are possible. Suppose in an area with high demand $(\varepsilon_i > 0)$ firms respond and build more capacity which drives down prices. Think of the problem this way -- if you regress quantity on price, how do you know you get the demand curve and not the supply curve?

7. In this problem, we exploit the properties of natural logs, ln(ab)=ln(a)+ln(b). In the initial model, we have $\ln(Q_i)=\beta_0 + \ln(P_i)\beta_1 + \varepsilon_i$. Now, we rescale prices by multiplying all observation by the same constant, and therefore, the model would be $\ln(Q_i)=\theta_0 + \ln(P_iC)\theta_1 + \upsilon_i$. Re write this as

$\ln(Q_i)=\theta_0 + \ln(Pi)\ \theta_1 + \ln(C)\theta_1 + \upsilon_i$

Notice that $\ln(C)\theta_1$ is a constant and it can be grouped with θ_0

$\ln(Q_i)=\theta_0 + \ln(C)\theta 1 + \ln(Pi)\ \theta_1 + \upsilon_i$

The previous model will produce exactly the same estimate on the ln(price) coefficient as before, so θ_1 will equal β_1.

Another way to think about it is that the original estimate is $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(\ln(\mathrm{p}_i) - \overline{\ln(p)})(\ln(\mathrm{q}_i) - \overline{\ln(\mathrm{q})})}{\sum_{i=1}^{n}(\ln(\mathrm{p}_i) - \overline{\ln(p)})^2}$

And with the re-scaled price, the new estimate would be $\hat{\gamma}_1 = \dfrac{\sum_{i=1}^{n}(\ln(\mathrm{p}_i^*) - \overline{\ln(p^*)})(\ln(\mathrm{q}_i) - \overline{\ln(\mathrm{q})})}{\sum_{i=1}^{n}(\ln(\mathrm{p}_i^*) - \overline{\ln(p^*)})^2}$ where $p_i^* = p_i C$. Noting that ln(ab)=ln(a)+ln(b), we can write $\ln(p_i^*) = \ln(p_i) + \ln(C)$ and $\overline{\ln(p^*)} = \overline{\ln(p)} + \ln(C)$ and hence $\ln(\mathrm{p}_i^*) - \overline{\ln(p^*)} = \ln(\mathrm{p}_i) - \overline{\ln(p)}$ so $\hat{\gamma}_1 = \hat{\beta}_1$

8. This is not a very good idea. Consider the estimate for $\hat{\gamma}_1$. We can always drop one the means in the numerator so initially drop $\overline{x}$. Next, recall that in all regression models, by construction, $\bar{\hat{\varepsilon}}$ =0. Therefore, the estimate reduces to

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})x_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i x_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

We have all seen the numerator before. Look at your notes from the second day of class and look at the first order conditions of the basic minimization problem.

$$(2)\ \ \partial SSE / \partial\widehat{\beta}_1 = -2\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1\right)x_i = 0$$

Which can be reduced to read

$$\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1\right)x_i = \sum_{i=1}^{n}\hat{\varepsilon}_i x_i = 0$$

Where $\hat{\varepsilon}_i = \left(y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1\right)$. The OLS model chooses $\hat{\beta}_0$ *and* $\hat{\beta}_1$ such that x_i is, by construction, uncorrelated with $\hat{\varepsilon}_i$. Therefore, in the estimate for $\hat{\gamma}_1$, the numerator will be by construction equal to zero and in any model, $\hat{\gamma}_1$ will also, by construction, equal 0. Therefore, it does not inform us at all about whether x_i and ε_i are correlated. This is a dumb idea.

9. The OLS estimate for $\hat{\beta}_1$ in this case is $\hat{\beta}_1^* = \dfrac{\sum_{i=1}^{n}(y_i^* - \overline{y}^*)(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

Note that we can write the numerator as $\sum_{i=1}^{n} y_i^*(x_i - \overline{x})$ so the estimate would reduce to

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n} y_i^*(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Recall that $y_i^* = y_i + v_i$ *and* $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ so substitute these values into the equation above

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n} y_i^*(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(\beta_0 + x_i\beta_1 + \varepsilon_i + v_i)(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Note that by construction $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$ and $\sum_{i=1}^{n}(x_i - \overline{x})x_i = \sum_{i=1}^{n}(x_i - \overline{x})^2$

So the estimate simplifies to

$$\hat{\beta}_1^* = \beta_1 + \frac{\sum_{i=1}^{n}\varepsilon_i(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \frac{\sum_{i=1}^{n}v_i(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Taking expectations and remembering that $E[\beta_1] = \beta_1$ we get

$$E[\hat{\beta}_1^*] = \beta_1 + E\left[\frac{\sum_{i=1}^{n}\varepsilon_i(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] + E\left[\frac{\sum_{i=1}^{n}v_i(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]$$

Because we have assumed that $\mathrm{cov}(\varepsilon_i, x_i) = 0$ and $\mathrm{cov}(v_i, x_i) = 0$, both expectations on the right hand side are zero and therefore

$$E[\hat{\beta}_1^*] = \beta_1$$

Even with measurement error in y, the OLS estimate for $\hat{\beta}_1$ is still unbiased. Is this a great estimator or what?

10. There are two ways to do this – Show that R^2 equals the correlation coefficient or the reverse – show that the correlation coefficient equals R^2. Either way, you need to know the definition of the correlation coefficient.

$$\hat{\rho}(y_i,\hat{y}_i)=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})/(n-1)}{\left(\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)}\right)\left(\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2/(n-1)}\right)}$$

Notice that the (n-1)'s in both the numerator and denominator cancel so the squared correlation coefficient equals

$$\hat{\rho}(y_i,\hat{y}_i)^2=\frac{\left(\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}.$$

a) Show that the correlation coefficient equals the R^2. Notice that the denominator is (SST)(SSM). Now let's work with the numerator. Recall that we can write the numerator as $\left(\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})\right)^2=\left(\sum_{i=1}^{n}y_i(\hat{y}_i-\bar{\hat{y}})\right)^2$ and note that $y_i=\hat{y}_i+\hat{\varepsilon}_i$. Therefore,

$$\left(\sum_{i=1}^{n}y_i(\hat{y}_i-\bar{\hat{y}})\right)^2=\left(\sum_{i=1}^{n}(\hat{y}_i+\hat{\varepsilon}_i)(\hat{y}_i-\bar{\hat{y}})\right)^2=\left(\sum_{i=1}^{n}\hat{y}_i(\hat{y}_i-\bar{\hat{y}})+\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})\hat{\varepsilon}_i\right)^2$$

$$=\left(\sum_{i=1}^{n}\hat{y}_i(\hat{y}_i-\bar{\hat{y}})\right)^2=\left(\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2\right)^2=SSM^2$$

Note that we use the fact that $\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{y}_i-\bar{\hat{y}})=0$ in the first row above (you can find the derivation for this when we did the construction of the R^2. Therefore

$$\hat{\rho}(y_i,\hat{y}_i)^2=\frac{SSM^2}{SST(SSM)}=\frac{SSM}{SST}=R^2.$$

b) Show the R^2 equals the correlation coefficient. In problem set 2, we showed that we could write the R^2 as $R^2=\left(\frac{\left(\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})\right)^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}\right)$. The hint says that $\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2=\hat{\beta}_1^2\sum_{i=1}^{n}(\mathrm{x}_i-\bar{x})^2$ which means that $\sum_{i=1}^{n}(\mathrm{x}_i-\bar{x})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2/\hat{\beta}_1^2$ and $(\mathrm{x}_i-\bar{x})=(\hat{y}_i-\bar{\hat{y}})/\beta_1$ so

$$R^2 = \left(\frac{\dfrac{1}{\hat{\beta}_1^2}\left(\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\dfrac{1}{\hat{\beta}_1^2}\left(\displaystyle\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)\left(\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2\right)} \right) = \left(\frac{\left(\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\displaystyle\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)\left(\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2\right)} \right)$$