Problem Set 5 Answers

Economics 3125

Fall 2013

Claire S.H. Lim

Attach printouts of your Stata do file and log file for computer exercises.

1. The data in 401K.DTA are a subset of data analyzed by Papke (1995) to study the relationship between participation in a 401(k) pension plan and the generosity of the plan. The variable *prate* is the percentage of eligible workers with an active account; this is the variable we would like to explain. The measure of generosity is the plan match rate, *mrate*. This variable gives the average amount the firm contributes to each worker's plan for each \$1 contribution by the worker. For example, if $mrate = 0.50$, then a \$1 contribution by the worker is matched by a \$0.50 contribution by the firm.

   (a) Find the average participation rate and the average match rate in the sample of plans.

   Answer. The average *prate* is about 87.36 and the average *mrate* is about 0.732.

   (b) Now, estimate the simple regression equation

   $$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate$$

   and report the results along with the sample size and R-squared.

   Answer. The estimated equation is: $\widehat{prate} = 83.08 + 5.86 \cdot mrate$, $n = 1,534$, $R^2 = 0.075$.

   (c) Interpret the intercept in your equation. Interpret the coefficient on *mrate*.

   Answer. The intercept implies that, even if $mrate = 0$, the predicted participation rate is 83.08 percent. The coefficient on *mrate* implies that a one-dollar increase in the match rate is estimated to increase *prate* by 5.86 percentage points. This assumes, of course, that this change in *prate* is possible (if, say, *prate* is already at 98%, this interpretation makes no sense).

   (d) Find the predicted *prate* when $mrate = 3.5$. Is this a reasonable prediction? Explain what is happening here.

   Answer. If we plug $mrate = 3.5$ into the equation, we get $\widehat{prate} = 83.07 + 5.86 \cdot 3.5 = 103.59$. This is impossible, as we can have at most a 100 percent participation rate. This illustrates that,

especially when dependent variables are bounded, a simple regression model can give strange predictions for extreme values of the independent variable. (In the sample of 1,534 firms, only 34 have $mrate \geq 3.5$.)

(e) How much of the variation in *prate* is explained by *mrate*? Is this a lot in your opinion?

Answer. Looking at the R^2, $mrate$ explains about 7.5% of the variation in $prate$. This is not much, and suggests that many other factors influence 401(k) plan participation rates

2. Using data from 1988 for houses sold in Andover, Massachusetts, from Kiel and McClain (1995), the following equation relates housing price (*price*) to the distance from a recently built garbage incinerator (*dist*):

$$\widehat{\log(price)} = 9.40 + 0.312\log(dist)$$

$$n = 135, R^2 = 0.162$$

(a) Interpret the coefficient on log(*dist*). Is the sign of this estimate what you expect it to be?

Answer. Yes. If living closer to an incinerator depresses housing prices, then being farther away increases housing prices.

(b) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of *price* with respect to *dist*? (Think about the city's decision on where to put the incinerator.)

Answer. If the city chose to locate the incinerator in an area away from more expensive neighborhoods, then $log(dist)$ is positively correlated with housing quality. This would violate assumption SLR.4 (the zero conditional mean assumption, $E(u|x) = 0$), and thus OLS estimation would be biased.

(c) What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

Answer. Size of the house, number of bathrooms, size of the lot, age of the home, and quality of the neighborhood (including school quality), are just a handful of factors. As mentioned in part (b), these could certainly be correlated with *dist* [and $log(dist)$].

3. This problem uses the data in MEAP93.DTA. (The dataset is described in Wooldridge, Example 2.12.) We want to explore the relationship between the math pass rate (*math10*) and spending per student

(expend).

(a) Do you think each additional dollar spent has the same effect on the pass rate, or does a diminishing effect seem more appropriate? Explain.

Answer. It seems plausible that another dollar of spending has a larger effect for low-spending schools than for high-spending schools. At low-spending schools, more money can go toward purchasing more books, computers, and for hiring better qualified teachers. At high levels of spending, we would expend little, if any, effect because the high-spending schools already have high-quality teachers, nice facilities, plenty of books, and so on.

(b) In the population model

$$math10 = \beta_0 + \beta_1 \log(expend) + u$$

argue that $\beta_1/10$ is the percentage point change in *math10* given a 10% increase in *expend*.

Answer. If we take changes, as usual, we obtain: $\Delta math10 = \beta_1 \Delta log(expend) \approx (\beta_1/100)(\%\Delta expend)$. Therefore if $\%\Delta expend = 10$, we get, $\Delta math10 = \beta_1/10$.

(c) Use the data in MEAP93.DTA to estimate the model from part (b). Report the estimated equation in the usual way, including the sample size and *R*-squared.

Answer. The regression results are: $\widehat{math10} = -69.34 + 11.16 \cdot log(expend)$, $n = 408$, $R^2 = 0.0297$.

(d) How big is the estimated spending effect? Namely, if spending increases by 10%, what is the estimated percentage point increase in *math10*?

Answer. If expend increases by 10 percent, $\widehat{math10}$ increases by about 1.1 percentage points. This is not a huge effect, but it is not trivial for low-spending schools, where a 10 percent increase in spending might be a fairly small dollar amount.

(e) One might worry that regression analysis can produce fitted values for *math10* that are greater than 100. Why is this not much of a worry in this data set?

Answer. In this data set, the largest value of *math10* is 66.7, which is not especially close to 100. In fact, the largest fitted value is only about 30.2.

4. In many local governments one important source of revenue is property taxes. In many places (including Tompkins County) real estate is assessed by the government at full market value. The assessments

then determine the taxes that are due from each property. One question that arises is whether the assessors do an adequate job in determining values of real estate. One way to check this is to compare assessments of recently sold houses with their assessed values. This comparison can be made formal with the regression model

$$price_i = \beta_0 + \beta_1 assess_i + u_i$$

where $price_i$ is the sale price of a house in dollars and $assess_i$ is the assessed value of the house in dollars. Using a random sample of 88 houses, the following estimates were obtained (standard errors are in parentheses):

$$\widehat{price}_i = \underset{(16.27)}{-14.47} + \underset{(0.049)}{0.976} assess_i \qquad R^2 = 0.82$$

(a) What values of the intercept and slope parameters are consistent with assessors correctly valuing homes relative to market prices?

Answer. If assessors correctly value houses relative to market prices, then changes in assessment values and sale prices should match exactly and there should be no systematic under- or over-valuation of houses. This implies $\beta_0 = 0$ and $\beta_1 = 1$.

(b) Test the null hypothesis that the intercept parameter is consistent with assessors correctly valuing homes. Use a t-test with significance level 5%.

Answer. The null hypothesis is $H_0 : \beta_0 = 0$ and the alternative hypothesis is $H_A : \beta_0 \neq 0$. Since the sample is large, we can assume the test statistic follows the standard normal distribution regardless of whether the errors u are normally distributed. With $\alpha = 5\%$, the critical value for this test is $c_{0.025} = 1.96$. The test statistic is

$$t = \frac{\hat{\beta}_0 - 0}{se(\hat{\beta}_0)} = \frac{-14.47}{16.27} = -0.89$$

Because $|t| \leq 1.96$, we fail to reject the null hypothesis.

(c) Compute the p-value for a test of the null hypothesis that the slope parameter is consistent with assessors correctly valuing homes. At what significance levels can you reject this null hypothesis?

Answer. The test is of the null hypothesis $H_0 : \beta_1 = 1$ against the alternative $H_A : \beta_1 \neq 1$. As in part (b), the test statistic approximately follows the standard normal distribution. The test

statistic is

$$t = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)} = \frac{0.976 - 1}{0.049} = -0.49$$

The p-value for this test is $P(|Z| > |-0.49|) = 0.62$. We would reject the null hypothesis only when $\alpha > 62\%$.

(d) How much of the variation in price is explained by the variation in assessments?

Answer. The R^2 tells us that the regression explains 82% of the variation in sale prices.

(e) What is the expected market price predicted by the fitted model for a house that is assessed at \$100,000?

Answer. For $assess_i = 100000$, the estimated market price in dollars is $\widehat{price}_i = -14.47 + 0.976(100000) = 97600$.

(f) Based on your answers, is there evidence that assessors are systematically over-valuing or under-valuing houses? Explain carefully.

Answer. The estimates $\hat{\beta}_0 < 0$ and $\hat{\beta}_1 < 1$ suggest that assessors tend to over-value houses. But the hypotheses $\beta_0 = 0$ and $\beta_1 = 1$ stand up well to the data when tested separately, so there is little evidence that assessors are systematically wrong.

(g) Can you think of any additional variables that should be included in the model? Does the omission of those variables affect the validity of assumption SLR.4? How would you have to reinterpret the estimates if you thought that assumption SLR.4 was violated?

Answer. The effects of attributes that are observed by the assessor, such as square footage, number of bedrooms, and neighborhood quality, will be reflected in the variable *assess* and therefore are not really omitted variables. Macroeconomic conditions and unobserved (by the assessor) house attributes will also affect sale prices, but assumption SLR.4 is only violated if these factors are correlated with assessed values. If unobserved attributes were correlated with assessed values, then sale prices would systematically differ from assessed values for some range of *assess*, and we would expect assessors to use this information when updating assessments. That process might remove the correlation between *assess* and the omitted variables, so assumption SLR.4 may be valid in this context.

Do and Log Files

```
DO FILE

*Econ 3125, Statistics and Applied Econometrics
*Program Name: problem_set5.do

set more off
capture log close
local path "C:/Econ 3125/Problem Sets/Problem Set 5"
log using "‘path’/problem_set5.log", replace

/*Question 1*/
.
. use "‘path’\401k.dta"
.
. /*Part (a)*/
.
. sum prate mrate
Variable | Obs Mean Std. Dev. Min Max
---------+--------------------------------------------------------
prate    | 1534 87.36291 16.71654 3 100
mrate    | 1534 .7315124 .7795393 .01 4.91
.
. /*Part (b)*/
.
. reg prate mrate
Source   | SS       df     MS                          Number of obs = 1534
---------+------------------------------               F( 1, 1532) = 123.68
Model    | 32001.7271 1    32001.7271                   Prob > F = 0.0000
```

```
Residual | 396383.812 1532 258.73617                    R-squared = 0.0747
-------------+------------------------------           Adj R-squared = 0.0741
Total    | 428385.539 1533 279.442622                   Root MSE = 16.085
------------------------------------------------------------------------------
prate    | Coef.    Std. Err.  t    P>|t|  [95% Conf. Interval]
-------------+----------------------------------------------------------------
mrate    | 5.861079 .5270107 11.12  0.000   4.82734 6.894818
_cons    | 83.07546 .5632844 147.48 0.000  81.97057 84.18035
------------------------------------------------------------------------------
.

/*Question 3*/
use "‘path’\meap93.dta"

/*Part (c)*/
reg math10 lexpend

log close
exit

LOG FILE

log:  C:/Econ 3125/Problem Sets/Problem Set 5/problem_set5.log
.
. /*Question 3*/
.
. use "‘path’\meap93.dta"
.
. /*Part (c)*/
.
```

```
. reg math10 lexpend
Source   | SS        df     MS                        Number of obs = 408
-------------+------------------------------          F( 1, 406) = 12.41
Model    | 1329.42517 1   1329.42517                  Prob > F = 0.0005
Residual | 43487.7553 406 107.112698                  R-squared = 0.0297
-------------+------------------------------          Adj R-squared = 0.0273
Total    | 44817.1805 407 110.115923                  Root MSE = 10.35
------------------------------------------------------------------------------
math10   | Coef.    Std. Err. t    P>|t|   [95% Conf. Interval]
---------+--------------------------------------------------------------------
lexpend  | 11.16439 3.169011  3.52 0.000   4.934677 17.39411
_cons    | -69.3411 26.53013 -2.61 0.009   -121.4947 -17.18753
------------------------------------------------------------------------------
.
. log close
```