1. (Lecture 6) (14 points) Use the data in APPLE to answer this question.
2. (2 point) Define a binary variable as *ecobuy* = 1 if *ecolbs* > 0 and *ecobuy* = 0 if *ecolbs* = 0. In other words, *ecobuy* indicates whether, at the prices given, a family would buy any ecologically friendly apples. What fraction of families claim they would buy ecolabeled apples?
3. (3 points) Estimate the linear probability model

and report the results in the usual form. Carefully interpret the coefficients on the price variables (*ecoprc* and *regprc*).

1. (4 points) Are the nonprice variables jointly significant in the LPM? (Use the usual *F* statistic, even though it is not valid when there is heteroskedasticity.) Which explanatory variable other than the price variables seems to have the most significant effect on the decision to buy ecolabeled apples? Does this make sense to you?
2. (5 points) In the model from part (ii), replace *faminc* with log(*faminc*). Given the *R*2, which model fits the data better? How many estimated probabilities are negative? How many are bigger than one? Should you be concerned? [Hint: Use command predict y to generate fitted values.]

Answer:

1. The fraction of families claim they would buy ecolabeled apples is
2. The OLS estimates of the LPM are

If *ecoprc* increases by 1 dollar, then the probability of buying eco-labeled apples is expected to fall by about .803. If *regprc* increases by 1 dollar, the probability of buying eco-labeled apples is expected to increase by about .719.

1. The *F* statistic is 4.43, with *p*-value = .0015. Thus, based on the usual *F* test, the four non-price variables are jointly very significant. Of the four variables, *educ* appears to have the most significant effect, with *p*-value = .003. For example, a difference of four years of education implies an increase of .025(4) = .10 in the estimated probability of buying eco-labeled apples. This suggests that more highly educated people are more open to buying produce that is environmentally friendly, which is perhaps expected. Household size (*hhsize*) also has an effect. Comparing a couple with two children to one that has no children – other factors equal – the couple with two children has a .048 higher probability of buying eco-labeled apples.
2. The model with log(*faminc*) fits the data slightly better: the *R*-squared increases from .1098 to .1116. The fitted probabilities range from about .185 to 1.051, so none are negative. After sorting, we find that there are two fitted probabilities above 1, which should not be a source of concern with 660 observations.
3. (Lecture 7) (11 points) Use the data in EZANDERS for this exercise. The data are on monthly unemployment claims in Anderson Township in Indiana, from January 1980 through November 1988. In 1984, an enterprise zone (EZ) was located in Anderson (as well as other cities in Indiana). [See Papke (1994) for details.]
4. (6 points) Regress log(*uclms*) on a monthly linear time trend and 11 monthly dummy variables. **Copy and paste the stata code you use to generate the monthly linear time trend *t* here.** [Hint: Use *jan* as the base month for the monthly dummy variables.]

What was the overall trend in unemployment claims over this period? (Interpret the coefficient on the time trend.) Is there evidence of seasonality in unemployment claims?

1. (3 points) Add *ez*, a dummy variable equal to one in the months Anderson had an EZ, to the regression in part (i). Does having the enterprise zone seem to decrease unemployment claims? By how much?
2. (2 points) What assumptions do you need to make to attribute the effect in part (ii) to the creation of an EZ?

Answer:

1. (2 points for the commands, omitted here) The coefficient on the time trend in the regression of log(*uclms*) on a linear time trend and 11 monthly dummy variables is about , which implies that monthly unemployment claims fell by about 1.4% per month on average. The trend is very significant with *p*-value = .00. There is also very strong seasonality in unemployment claims, with 6 (*may*, *jun*, *jul*, *sep*, *oct*, *nov*) of the 11 monthly dummy variables having *p*-values less than .05. The *F* statistic for joint significance of the 11 monthly dummies is 3.24 and yields *p*-value of around .0009.
2. When *ez* is added to the regression, its coefficient is about . It is significant at 1% level with *p*-value = .005. It indicates that unemployment claims are estimated to fall approximate 50.8% after enterprise zone designation.
3. We must assume that around the time of EZ designation there were no other external factors that caused a shift in the trend of log(*uclms*). We have controlled for a time trend and seasonality, but this may not be enough.
4. (Lecture 8) (12 points) Use the data in HSEINV for this exercise.
5. (4 points) Find the first order autocorrelation in log(*invpc*) and log(*price*) respectively. Which of the two series may have a unit root? [Hint: Use the command correlate to find the correlation.]
6. (3 points) Based on your findings in part (i), estimate the equation

and report the results in standard form. Interpret the coefficient and determine whether it is statistically significant.

1. (5 points) Now use as the dependent variable. Re-run the equation and report the results in standard form. How do your results of the coefficient change from part (ii)? Is the time trend still significant? Why or why not?

Answer:

1. The first order autocorrelation for log(*invpc*) is about .6391. So there is no strong evidence of a unit root in it. For log(*price*), the first order autocorrelation is about .9492, which is very high. So we cannot confidently rule out a unit root in log(*price*).
2. The estimated equation is

The coefficient on implies that a one percentage point increase in the growth in price leads to a 3.88 percent increase in housing investment above its trend. The *t* statistic is 4.05 with *p*-value = 0.000, so it is very statistically significant.

1. The estimated equation is

The coefficient on has fallen substantially and is no longer significant. (*t* statistic is 1.38 with *p*-value = 0.177.) The *R*-squared is much smaller, so explains very little of the variation in . Because differencing eliminates linear time trends, it is not surprising that the estimate on the trend is very small and very statistically insignificant. (*t* statistic is 0.02 with *p*-value = 0.985.)

1. (Lecture 8) (12 points) Recall that in the example of testing Efficient Markets Hypothesis, it may be that the expected value of the return at time *t*, given past returns, is a quadratic function of *returnt-1*.
2. (2 points) To check this possibility, use the data in NYSE to estimate

;

report the results in standard form.

1. (4 points) State and test the null hypothesis that does not depend on *returnt-1*. [Hint: There are two restrictions to test here.] What do you conclude?
2. (4 points) Drop from the model, but add the interaction term . Now test the efficient markets hypothesis. [Hint: stata can create lag (or lead) variables using subscripts conveniently. For example, you can use the command gen return\_2 = return[\_n-2] to create *returnt-2* fast.]
3. (2 points) What do you conclude about predicting weekly stock returns based on past stock returns?

Answer:

1. The estimated equation is
2. The null hypothesis is . Only if both parameters are zero does not depend on *returnt‑*1. The *F* statistic is about 2.16 with *p*-value = .1161. Therefore, we cannot reject H0 even at the 10% level.
3. When we put in place of, the null can still be stated as in part (ii): no past values of *return*, or any functions of them, should help us predict *returnt*. So we regress on and and test the joint significance of the explanatory variables. The *F* statistic is about 1.80 with *p*-value = .1658. Therefore, again, we cannot reject H0 at even the 15% level.
4. Predicting *returnt* based on past returns does not appear promising. Even though the *F* statistic from part (ii) is almost significant at the 10% level, we have many observations. We cannot even explain 1% of the variation in *returnt*.
5. (Lecture 9) (11 points) Use the data in KIELMC for this exercise.
6. (3 points) The variable *dist* is the distance from each home to the incinerator site, in feet. Consider the model

If building the incinerator reduces the value of homes closer to the site, what is the sign of δ1? What does it mean if β1 > 0?

1. (3 points) Estimate the model from part (i) and report the results in the usual form. Interpret the coefficient on . What do you conclude?
2. (3 points) Add *age*, *age*2, *rooms*, *baths*, log(*intst*), log(*land*), and log(*area*) to the equation. Now, what do you conclude about the effect of the incinerator on housing values?
3. (2 points) Why is the coefficient on log(*dist*) positive and statistically significant in part (ii) but not in part (iii)? What does this say about the controls used in part (iii)?

Answer:

1. Other things equal, homes farther from the incinerator should be worth more, so . If , then the incinerator was chosen to be located – even before its construction – farther away from more expensive homes.
2. The estimated equation is

means that after the incinerator site was chosen, increasing distance from house to incinerator by 100% (double the distance) is predicted to increase house selling price by 4.8%. While it has the expected sign, it is not statistically significant (*p*-value = 0.556).

1. When we add the list of housing characteristics to the regression, the coefficient on becomes .062 (se = .050). So the estimated effect is larger – the elasticity of *price* with respect to *dist* is .062 after the incinerator site was chosen – but its *t* statistic is still only 1.24. The *p*-value for the one-sided alternative is about .108, which is close to being significant at the 10% level.
2. The fact that log(*dist*) has a much smaller coefficient and is insignificant in part (iii) indicates that the characteristics included in part (iii) largely capture the housing characteristics that are most important for determining housing prices.
3. (Lecture 10) (13 points) Use the data in PHILLIPS for this exercise. As we mentioned in Lecture 7, instead of the static Phillips curve model, we can estimate an expectations-augmented Phillips curve of the form

,

where .

1. (3 points) Estimate this equation by OLS and report the results in the usual form. In estimating this equation by OLS, we assumed that the supply shock, *et*, was uncorrelated with *unemt*. If this is false, what can be said about the OLS estimator of β1?
2. (2 points) Suppose that *et* is unpredictable given all past information: . Explain why this makes *unemt-1* a good IV candidate for *unemt*.
3. (3 points) Does *unemt-1* satisfy the instrument relevance assumption? [Hint: You need to run a regression to answer this question.]
4. (5 points) Estimate the expectations augmented Phillips curve by 2SLS using *unemt-1* as an IV for *unemt*. Report the results in the usual form and compare them with the OLS estimates from (i).

Answer:

1. The OLS estimates are

As usual, if *unemt* is correlated with *et*, OLS will be biased and inconsistent for estimating β1.

1. If , then *unemt‑*1 is uncorrelated with *et*, which means *unemt‑*1 satisfies the first requirement as an IV for *unemt*.
2. The second requirement for *unemt‑*1 to be a valid IV for *unemt* is that *unemt‑*1 must be sufficiently correlated. Regress *unemt* on *unemt-1* yields the coefficient estimate of .742 with standard error .089, so the *t* statistic is about 8.31 with *p*-value = .000. Therefore, there is a strong, positive correlation between *unemt* and *unemt*‑1, and *unemt-1* satisfy the instrument relevance assumption well.
3. The expectations-augmented Phillips curve estimated by 2SLS using *unemt-1* as an IV for *unemt* is

The IV estimate of β1 is much lower in magnitude than the OLS estimate (-.518), and also becomes not statistically different from zero. The OLS estimate had a *t* statistic of about -2.48 with *p*-value = .017, so it’s statistically different from zero at 5% level.