Assignment - 2: BUAN 6312 Harikrishna Dev HXD220000

Answers

1. Use the data in APPLE to answer this question.

- 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?

The fraction of families claim they would buy ecolabeled apples are 62.42%

```
. gen ecobuy = 0

. replace ecobuy = 1 if ecolbs > 0
(412 real changes made)

. tabulate ecobuy

     ecobuy |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        248       37.58       37.58
          1 |        412       62.42      100.00
------------+-----------------------------------
      Total |        660      100.00
```

- Estimate the linear probability model below and and report the results in the usual form. Carefully interpret the coefficients on the price variables (*ecoprc* and *regprc*).

$$ecobuy = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc + \beta_4 hhsize + \beta_5 educ + \beta_6 age + u$$

We get the LRM equation as follows:

$$ecobuy = 0.4236865 + -0.8026219 \times ecoprc + 0.7192675 \times regprc + 0.0005518 \times faminc + 0.0238227 \times hhsize + 0.02$$

From the following equation, we can see that coefficients of ***ecoprc*** and ***regprc*** are ***0.803*** and ***0.719***. The p-values of these coefficients are less than 0.05, therefore they are statistically significant.

```
. reg ecobuy ecoprc regprc faminc hhsize educ age

      Source |       SS           df       MS      Number of obs   =       660
-------------+----------------------------------   F(6, 653)       =     13.43
       Model |  17.0019785         6  2.83366308   Prob > F        =    0.0000
    Residual |  137.810143       653  .211041566   R-squared       =    0.1098
-------------+----------------------------------   Adj R-squared   =    0.1016
       Total |  154.812121       659  .234919759   Root MSE        =    .45939

------------------------------------------------------------------------------
      ecobuy | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      ecoprc |  -.8026219   .1094037    -7.34   0.000    -1.017447   -.5877963
      regprc |   .7192675    .131639     5.46   0.000     .4607808    .9777543
      faminc |   .0005518   .0005295     1.04   0.298     -.000488    .0015916
      hhsize |   .0238227   .0125262     1.90   0.058    -.0007739    .0484193
        educ |   .0247849   .0083743     2.96   0.003      .008341    .0412287
         age |  -.0005008   .0012499    -0.40   0.689    -.0029551    .0019536
       _cons |   .4236865   .1649674     2.57   0.010      .099756     .747617
------------------------------------------------------------------------------
```

- 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?

We can see that we conduct a hypothesis tests on the non price variables gives us a *p_value* < *0.05*. Therefore, we can reject the null hypothesis i.e. non-price variables are jointly significant. As *t(educ)* = *2.96* is the highest t statistic value among the non price variable, we can conclude that ***education*** makes most significant effect on purchase of eco-labeled apples. This makes sense that educated customers would prefer ecolabeled apples as they would be more well equipped in understanding the benefit of the consumption of them.

```
. test faminc hhsize educ age

 ( 1)  faminc = 0
 ( 2)  hhsize = 0
 ( 3)  educ = 0
 ( 4)  age = 0

       F(  4,   653) =    4.43
            Prob > F =    0.0015
```

- 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.]

```
. gen lfaminc = ln(faminc)

. reg ecobuy ecoprc regprc lfaminc hhsize educ age

      Source |       SS           df       MS      Number of obs   =       660
-------------+----------------------------------   F(6, 653)       =     13.67
       Model |   17.278689         6  2.8797815   Prob > F        =    0.0000
    Residual |  137.533432       653  .210617813   R-squared       =    0.1116
-------------+----------------------------------   Adj R-squared   =    0.1034
       Total |  154.812121       659  .234919759   Root MSE        =    .45893

------------------------------------------------------------------------------
      ecobuy | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      ecoprc |  -.8006664   .1092981    -7.33   0.000    -1.015285   -.5860482
      regprc |    .721377   .1315196     5.48   0.000     .4631247    .9796294
     lfaminc |   .0445162   .0287239     1.55   0.122    -.0118861    .1009185
      hhsize |   .0227002    .012543     1.81   0.071    -.0019294    .0473297
        educ |    .023093   .0084508     2.73   0.006      .006499     .039687
         age |  -.0003865   .0012517    -0.31   0.758    -.0028444    .0020713
       _cons |   .3037519   .1789605     1.70   0.090    -.0476555    .6551593
------------------------------------------------------------------------------
```

We see that the *Adj-R sqr* of the second model is greater in the first model. This indicates that the second model fits better. In the second model,

2. 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).

- Regress log(uclms) on a monthly linear time trend and 11 monthly dummy variables. [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?

```
. use "C:\Users\hxd220000\Desktop\Data Sets- STATA\EZANDERS.DTA"

. regress luclms year feb mar apr may jun jul aug sep oct nov dec
```

```
      Source |       SS           df       MS      Number of obs   =       107
-------------+----------------------------------   F(12, 94)       =     14.36
       Model |  27.0363482        12  2.25302901   Prob > F        =    0.0000
    Residual |  14.7491008        94  .156905327   R-squared       =    0.6470
-------------+----------------------------------   Adj R-squared   =    0.6020
       Total |  41.7854489       106  .394202348   Root MSE        =    .39611

------------------------------------------------------------------------------
      luclms | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        year |  -.1665437   .0149503   -11.14   0.000    -.1962279   -.1368595
         feb |  -.0132261   .1867294    -0.07   0.944    -.3839816    .3575294
         mar |  -.0661643   .1867294    -0.35   0.724    -.4369198    .3045912
         apr |  -.3649279   .1867294    -1.95   0.054    -.7356834    .0058276
         may |  -.5147779   .1867294    -2.76   0.007    -.8855334   -.1440224
         jun |  -.5541234   .1867294    -2.97   0.004    -.9248789   -.1833679
         jul |  -.5191558   .1867294    -2.78   0.007    -.8899113   -.1484003
         aug |  -.3378477   .1867294    -1.81   0.074    -.7086032    .0329078
         sep |  -.7528584   .1867294    -4.03   0.000    -1.123614   -.3821029
         oct |  -.7867943   .1867294    -4.21   0.000      -1.15755  -.4160388
         nov |  -.6816665   .1867294    -3.65   0.000    -1.052422    -.310911
         dec |  -.3740492   .1926213    -1.94   0.055    -.7565034    .0084049
       _cons |   339.4264   29.66172    11.44   0.000     280.5323    398.3204
------------------------------------------------------------------------------
```

We see that coefficient of **YEAR** is *-0.1665*. This implies that the overall trend of unemployment claims decreases by *16.65%* per year. As the p-value < threshold value, we can conclude that the yearly trend is significant.

We can see that some of the monthly dummy variables are significant at a 5% level of significance, whereas some are not significant at the same threshold. This helps us understand that there is a presence of seasonal factors behind unemplyment claims.

To confirm the joint significance, we perform the Wald test on the 11 monthly dummy variables.

$$H_0 : feb - dec = 0$$

$$H_1 : feb - dec \neq 0$$

```
. test feb mar apr may jun jul aug sep oct nov dec

 ( 1)  feb = 0
 ( 2)  mar = 0
 ( 3)  apr = 0
 ( 4)  may = 0
 ( 5)  jun = 0
 ( 6)  jul = 0
 ( 7)  aug = 0
 ( 8)  sep = 0
 ( 9)  oct = 0
 (10)  nov = 0
 (11)  dec = 0

       F( 11,    94) =    4.32
            Prob > F =    0.0000
```

As the p-value < threshold, we can reject the null hypothesis. Therefore, we can conclude that the monthly dummy variables are jointly significant.

- 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?

```
. regress luclms year feb mar apr may jun jul aug sep oct nov dec ez

      Source |       SS           df       MS      Number of obs   =       107
-------------+----------------------------------   F(13, 93)       =     15.76
       Model |  28.7422487        13  2.21094221   Prob > F        =    0.0000
    Residual |  13.0432002        93  .140249465   R-squared       =    0.6879
-------------+----------------------------------   Adj R-squared   =    0.6442
       Total |  41.7854489       106  .394202348   Root MSE        =     .3745

------------------------------------------------------------------------------
      luclms | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
```

year	-.0811489	.0282722	-2.87	0.005	-.1372918	-.025006
feb	-.0132261	.1765405	-0.07	0.940	-.3638005	.3373484
mar	-.0661643	.1765405	-0.37	0.709	-.4167388	.2844101
apr	-.3649279	.1765405	-2.07	0.042	-.7155023	-.0143534
may	-.5147779	.1765405	-2.92	0.004	-.8653523	-.1642034
jun	-.5541234	.1765405	-3.14	0.002	-.9046978	-.203549
jul	-.5191558	.1765405	-2.94	0.004	-.8697303	-.1685814
aug	-.3378477	.1765405	-1.91	0.059	-.6884222	.0127267
sep	-.7528584	.1765405	-4.26	0.000	-1.103433	-.4022839
oct	-.7867943	.1765405	-4.46	0.000	-1.137369	-.4362198
nov	-.6816665	.1765405	-3.86	0.000	-1.032241	-.3310921
dec	-.3595756	.1821582	-1.97	0.051	-.7213057	.0021546
ez	-.5080266	.1456667	-3.49	0.001	-.7972917	-.2187614
_cons	170.2854	56.02201	3.04	0.003	59.03674	281.534

When ez is added to the regression, its coefficient is about −.508 (se ≈ .146). EZ decreases the unemplyment claims by:

$$100(1 - e^{-0.508}) = 39.82$$

3. Use the data in HSEINV for this exercise.

- Find the first order autocorrelation in log(invpc) and log(price) respectively. Which of the two series may have a unit root?

Answer here

- Based on your findings in part (i), estimate the equation below and report the results in standard form. Interpret the coefficient β\hat_1 and determine whether it is statistically significant.

$$log(invpc_t) = \beta_0 + \beta_1 \times \Delta log(price_t) + \beta_2 \times t + u_t$$

Answer here

- Now use Δlog(invpc_t) as the dependent variable. Re-run the equation and report the results in standard form. How do your results of the coefficient β^_1 change from part (ii)? Is the time trend still significant? Why or why not?

We must assume that around the time of EZ designation there were not other external factors that caused a shift down in the trend of log(uclms). We have controlled for a time trend and seasonality, but this may not be enough.

4. 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 $return_{t-1}$.

- To check this possibility, use the data in NYSE to estimate

$$return_t = \beta_0 + \beta_1 return_{t-1} + \beta_2 return_{t-1}^2 + u_t$$

report the results in standard form.

Answer here

- State and test the null hypothesis that E(return_t |return_(t-1)) does not depend on returnt-1. [Hint: There are two restrictions to test here.] What do you conclude?

Answer here

- Drop $return_{t-1}^2$ from the model, but add the interaction term $return_{t-1} \times return_{t-2}$. 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 $return_{t-2}$ fast.]

Answer here

- What do you conclude about predicting weekly stock returns based on past stock returns?

Answer here

5. Use the data in KIELMC for this exercise.

- The variable dist is the distance from each home to the incinerator site, in feet. Consider the model

$$log(price) = \beta_0 + \delta_0 y_8 1 + \beta_1 log(dist) + \delta_1 y_8 1 \cdot log(dist) + u.$$

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?

Answer here

- Estimate the model from part (i) and report the results in the usual form. Interpret the coefficient on $y_8 1 \cdot log(dist)$. What do you conclude?

Answer here

- What do you conclude about predicting weekly stock returns based on past stock returns?

Answer here

- Add $age, age^2, rooms, baths, log(intst), log(land), andlog(area)$ to the equation. Now, what do you conclude about the effect of the incinerator on housing values?

Answer here

- 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 here

6. 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

$$\Delta inf_t = \beta_0 + \beta_1 unem_t + e_t$$

where $\Delta inf_t = inf_t - inf_{t-1}$

- 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?

Answer here

- Suppose that et is unpredictable given all past information: $E(e_t \mid inf_(t-1), unem_(t-1), \ldots) = 0$. Explain why this makes $unem_t - 1$ a good IV candidate for $unem_t$.

Answer here

- Does $unem_t - 1$ satisfy the instrument relevance assumption? [Hint: You need to run a regression to answer this question.]

Answer here

- Estimate the expectations augmented Phillips curve by 2SLS using $unem_t - 1$ as an IV for $unem_t$. Report the results in the usual form and compare them with the OLS estimates from (i).