Solutions to Extra Practice Questions - Week 3

Statistics and Econometrics

Question 1

Use the data in apple.RData to answer this question.

1. Define 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 fractions of families claim they would buy eco-labeled apples?
2. Estimate the linear probability model

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

   and report the results. Carefully interpret the coefficients on the price variables.
3. Are the nonprice variables significant in the LPM? Which explanatory variable other than the price variables seems to have the most important effect on the decision to buy eco-labeled apples? Does this make sense to you?
4. In the estimation in part 2, how many estimated probabilities are negative? How many are bigger than one? Should you be concerned?

Solution

1.

```r
load("apple.RData")
data$ecobuy <- ifelse(data$ecolbs > 0, 1, 0)
mean(data$ecobuy)
```

```
## [1] 0.6242424
```

The fraction of families claim they would buy eco-labeled apples is 62.4%.

2.

```r
ecobuy.lpm <- lm(ecobuy ~ ecoprc + regprc + faminc + hhsize + educ + age, data)
r.ttest <- coeftest(ecobuy.lpm, vcov = vcovHC(ecobuy.lpm, "HC1"))
```

The OLS estimates of the LPM are

$$\widehat{ecobuy} = -\underset{(.168)}{.424} - \underset{(.106)}{.803}\,ecoprc + \underset{(.130)}{.719}\,regprc + \underset{(.001)}{.001}\,faminc + \underset{(.012)}{.024}\,hhsize + \underset{(.008)}{.025}\,educ - \underset{(.001)}{.001}\,age,$$

$n = 660, R^2 = .110$. We report robust standard errors due to the heteroskedasticity problem in LPM. If *ecoprc* increases by, say, 10 cents (.10), then the probability of buying eco-labeled apples falls by about .080. If *regprc* increases by 10 cents, the probability of buying eco-labeled apples increases by about .072.

3.

```r
linearHypothesis(ecobuy.lpm, c("faminc = 0", "hhsize = 0", "educ = 0", "age = 0"),
                 white.adjust = "hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## faminc = 0
## hhsize = 0
## educ = 0
## age = 0
##
## Model 1: restricted model
## Model 2: ecobuy ~ ecoprc + regprc + faminc + hhsize + educ + age
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F   Pr(>F)
## 1    657
## 2    653  4 4.2427 0.002133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F test, with 4 and 653 df, is 4.24, with p-value = .002. Thus, the four non-price variables are jointly very significant. Of the four variables, *educ* appears to have the most important effect. 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.

4.

```r
sum(ecobuy.lpm$fitted.values < 0)
```

```
## [1] 0
```

```r
sum(ecobuy.lpm$fitted.values > 1)
```

```
## [1] 2
```

None are negative, and there are two fitted probabilities above 1, which is not a source of concern with 660 observations.

Question 2

The variable *smokes* is a binary variable equal to one if a person smokes, and zero otherwise. We estimate a linear probability model for smokes:

$$\begin{aligned}\widehat{smokes} \;=\;& \underset{\substack{(.855)\\ [.856]}}{.656} - \underset{\substack{(.204)\\ [.207]}}{.069}\log(cigpric) + \underset{\substack{(.026)\\ [.026]}}{.012}\log(income) - \underset{\substack{(.006)\\ [.006]}}{.029}\,educ\\ & + \underset{\substack{(.006)\\ [.005]}}{.020}\,age - \underset{\substack{(.00006)\\ [.00006]}}{.00026}age^2 - \underset{\substack{(.039)\\ [.038]}}{.101}\,restaurn - \underset{\substack{(.052)\\ [.050]}}{.026}\,white,\end{aligned}$$

$n = 807, R^2 = .062$. *income* is the person's annual income; *cigpric* indicates the per pack price of cigarettes; *educ* indicates years of schooling the person received; *age* is measured in years; *restaurn* is a dummy variable which equals one if the person lives in a state with restaurant smoking restrictions; *white* equals to one if the person is Caucasian. Both the usual and heteroskedasticity-robust standard errors are reported.

1. Are there any important differences between the two sets of standard errors?

2. Holding other factors fixed, if education increases by four years, what happens to the estimated probability of smoking?
3. Interpret the coefficient on the binary variable *restaurn*.
4. A person in the data set has the following characteristics: $cigpric = 67.44, income = 6,500$, $educ = 16, age = 77, restaurn = 0, white = 0$, and $smokes = 0$. Compute the predicted probability of smoking for the observation in the data set.

Solution

1. No. For each coefficient, the usual standard errors and the robust ones are practically very similar.

2. The effect is $-.029(4) = -.116$, so the probability of smoking falls by about .116.

3. Holding other factors in the equation fixed, a person in a state with restaurant smoking restrictions has a .101 lower chance of smoking. This is similar to the effect of having four more years of education.

4. We just plug the values of the independent variables into the OLS regression line:

   $$\widehat{smokes} = .656 - .069 \cdot \log(67.44) + .012 \cdot \log(6,500) - .029(16) + .020(77) - .00026(77^2) \approx .0052.$$

   Thus, the estimated probability of smoking for this person is close to zero.