**ECO 5720**

**Problem Set 3**

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

(i) 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 (i.e., what fraction of observations have *ecobuy* = 1?

(ii) Estimate the linear probability model

$$ecobuy=β\_{0}+β\_{1}ecoprc+β\_{2}regprc+β\_{3}faminc+β\_{4}hhsize+β\_{5}educ+β\_{6}age+u,$$

and report the results. Noting that *ecoprc* and *regprc* denote the price (in dollars) of ecolabeled and regular apples, respectively, carefully interpret their corresponding coefficient estimates. Please note that it is important to state both the sign and the magnitude while interpreting a coefficient estimate.

(iii) How many predicted probabilities are negative? How many are greater than one?

2. Use the data in CATHOLIC to answer this question.

(i) What is the average value of mathematics scores (i.e., *math12*) for students who attended a Catholic high school (i.e., for *cathhs* = 1)? Similarly, what is the average of *math12* across students who did not attend a Catholic high school (i.e., for *cathhs* = 0)? What is the difference in average math scores between the two groups?

(ii) Run a multiple regression of *math12* on *cathhs*, *lfaminc*, *motheduc*, and *fatheduc*, and report the results. Please interpret the coefficient estimate corresponding to *cathhs* while noting its sign and magnitude.

(iii) Next, run a multiple regression of *hsgrad* (i.e., a binary indicator of high school graduation) on *cathhs*, *lfaminc*, *motheduc*, and *fatheduc*, and report the results. Interpret the coefficient estimate corresponding to *cathhs*. Please specify both the sign and the magnitude of the coefficient estimate.

(iv) In case of part (iii), how many predicted probabilities are outside the [0, 1] interval?

3. Use the data in NBASAL for this exercise.

(i) Estimate a linear regression model relating points per game to experience in the league and position (guard, forward, or center). Use centers as the base group. Report the results in the usual form.

(ii) Why do you not include all three position dummy variables in part (i)?

(iii) Holding experience fixed, does a guard score more than a center? How much more? Is the difference statistically significant?