Homework 10

Multiple Regression Model- Dummy dependent variables (25 points)

Due Date: Monday April 19^{th} at 11:59 pm

Instruction:

- This HW must be done in Rmarkdown!
- Please submit both the .rmd and the Microsoft word files. (Do not submit a PDF or any other image files as the TAs are going to give you feedback in your word document)
- Name your files as: HW10_groupnumber_name
- All the HW assignments are individual work. However, I highly encourage you to discuss it with your group members.
- The answer key will be uploaded on Canvas a couple of days after the due date.

Question 1 Given the following:

$$\widehat{log(bwght)} = \underset{0.22}{4.66} - \underset{0.0009}{0.004}cigs + \underset{0.0059}{0.0093}log(faminc) + \underset{0.006}{0.016}parity + \underset{0.01}{0.027}male + \underset{0.013}{0.055}white$$

$$n = 1,388, \qquad R^2 = 0.0472$$

and:

$$\widehat{log(bwght)} = \underset{0.38}{4.65} - \underset{0.0010}{0.0052}cigs + \underset{0.0085}{0.0110}log(faminc) + \underset{0.006}{0.017}parity + \underset{0.011}{0.034}male + \underset{0.015}{0.045}white - \underset{0.003}{0.003}motheduc + \underset{0.0026}{0.0032}fatheduc$$

$$n = 1,191 \qquad R^2 = 0.0493$$

The variables are defined as:
bwght = birth weight, in lbs.
cigs = average number of cigarettes the mother smoked per day during pregnancy.
parity = the birth order of this child.
faminc = annual family income.
motheduc = years of schooling for the mother.
fatheduc = years of schooling for the father.
male = Whether the child is male.
white = Whether the child is white or not.

1. In the first equation, interpret the coefficient on the variable *cigs*. In particular, what is the effect on birth weight from smoking 10 more cigarettes per day?
2. How much more is a white child predicted to weigh than a nonwhite child, holding the other factors in the first equation fixed? Is the difference statistically significant?
3. Comment on the estimated effect and statistical significance of *motheduc*.
4. From the given information, why are you unable to compute the F statistic for joint significance of *motheduc* and *fatheduc*? What would you have to do to compute the F statistic? (Hint, look at the number of observations in each model. Restricted vs Unrestricted)

Question 2 Given the following:

$$\widehat{sat} = \underset{6.29}{1,028.10} + \underset{3.83}{19.30}hsize - \underset{0.53}{2.19}hsize^2 - \underset{4.29}{45.09}female - \underset{12.71}{169.81}black + \underset{18.15}{62.31}female * black$$

$$n = 4,137 \qquad R^2 = 0.0858$$

The variables are defined as:
sat = Combine SAT score.
hsize = high school graduating class size in the hundreds.
female = Whether the student is female or not.
black = Whether a student is black or not.

1. Is there strong evidence that $hsize^2$ should be included in the model? From this equation, what is the optimal high school size?
2. Holding *hsize* fixed, what is the estimated difference in SAT score between *nonblack* females and *nonblack* males? How statistically significant is this estimated difference?
3. What is the estimated difference in SAT score between nonblack males and black males? Test the null hypothesis that there is no difference between their scores, against the alternative that there is a difference.
4. What is the estimated difference in SAT score between black females and nonblack females? What would you need to do to test whether the difference is statistically significant?

Question 3 Given the following:

$$inlf = 0.586 - 0.0034nwifeinc + 0.038educ + 0.039exper - 0.0006exper^2$$
$$- 0.016age - 0.262kidslt6 + 0.013kidsage6$$

$$n = 753, \ R^2 = 0.264$$

The variables are defined as:
inlf = (binary variable) In labor force participation by married women.
nwifeinc = Husband's earnings measured in thousands of dollars.
educ = Years of education.
exper = Years of labor market experience.
age = Age.
kidslt6 = Number of children under age 6.
kidsge6 = Number of kids between 6 - 18 years old.

Suppose that we define *outlf* to be one if the woman is out of the labor force, and zero otherwise.

1. If we regress *outlf* on all of the independent variables in the equation, what will happen to the intercept and slope estimates? (Hint: $inlf = 1 - outlf$. Plug this into the population equation $inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + ...$ and rearrange.)
2. What will happen to the standard errors on the intercept and slope estimate?
3. What will happen to the R-squared?

Computer Exercises

Question 4 Use the data in WAGE2 for this exercise.

Given the following:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure$$

$$+ \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u$$

1. Report the estimated model. Holding other factors fixed, what is the approximate difference in monthly salary between blacks and nonblacks? Is this difference statistically significant?
2. Add the variables $exper^2$ and $tenure^2$ to the equation and show that they are jointly insignificant at even the 20% level.
3. Extend the original model to allow the return to education to depend on race and test whether the return to education does depend on race.
4. Again, start with the original model, but now allow wages to differ across four groups of people: married and black, *married* and *nonblack*, *single* and *black*, and *single* and *nonblack*. What is the estimated wage differential between *married blacks* and *married nonblacks*?

Question 5 Use the data in APPLE for this exercise.
(Hint: in order to make a new dummy variable called ecobuy, you need to use a combination of mutate() and ifelse() functions.)

1. Define a binary varibale 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?
2. Estimate the linear probability of the following model.

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

   Report the results in equation form. Carefully interpret the coefficients on the price variables.
3. 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 important effect on the decision to buy ecolabeled apples? Does this make sense to you?
4. In the model from part 2, replace *faminc* with *log(faminc)*. Which model fits the data better, using *faminc* or *log(faminc)*? Interpret the coefficient on *log(faminc)*.
5. In the estimation in part 4, how many estimated probabilities are negative? How many are bigger than one? Should you be concerned?
6. For the estimation in part (iv), compute the percent correctly predicted for each outcome,$ecobuy = 0$ and $ecobuy = 1$. Which outcome is best predicted by the model?