MSIN0096 Homework 3
Due Nov 22nd, 3pm

1. (9pt) Use the data set in Q1.csv for this problem, which is the house price data used in lectures.

   (a) (2pt) Run a simple regression of area on rooms to obtain the coefficient of rooms, denoted as γ.

   (b) (2pt) Run a simple regression of price on rooms, to obtain the coefficient of rooms, denoted as $\tilde{b}_1$.

   (c) (2pt) Run a multiple regression of price on area and rooms, and obtain the coefficients of area and rooms, $\hat{b}_1$ and $\hat{b}_2$ respectively.

   (d) (1pt) Verify that $\tilde{b}_1 = \hat{b}_1 + \gamma\hat{b}_2$.

   (e) (2pt) Compare the magnitude of $\tilde{b}_1$ and $\hat{b}_1$. Explain the discrepancy.

2. (6pt) Suppose we want to estimate the effects of alcohol consumption on student's academic performance. In addition to collecting information on weekly average alcohol usage (*alcholo*) and average scores (*score*) , we also obtain attendance information (say, percentage of lectures attended, called *attend*). A standardized college entrance exam score (*entrance*) is also available.

   (a) (4pt) We estimate two following regression equations

   $$score = b_0' + b_1' alcholo + e$$

   and

   $$score = b_0 + b_1 alcholo + b_2 attend + e$$

   How would you interpret b_1' and b_1 differently?

   (b) (2pt) Should *entrance* be included as an explanatory variable? Explain.

3. (11pt) Use the data in Q3.csv to answer this question. These are postal code–level data on soda prices at fast-food restaurants, along with characteristics of the postal code population. The idea is to see whether fast-food restaurants charge higher prices on soda in areas with a larger concentration of minority population. The key variables include average soda price per unit (*psoda*), proportion of minority ethnicity population (*prpminor*), median family income (*income*), proportion of population in poverty (*prppov*), median housing value (*house*).

   (a) (2pt) Estimate a model to explain the price of soda, *psoda*, in terms of the proportion of the minority ethnicity population *prpminor*.

   $$psoda = b_0 + b_1 prpminor + e$$

   Does fast-food restaurant charge higher price on soda in areas with high minority ethnicity population concentration? Based on the regression result, can you conclude that there is price discrimination against minorities?

   (b) (2pt) Estimate another model

   $$psoda = b_0 + b_1 prpminor + b_2 income + e$$

   Does the discrimination effect become larger or smaller when you control for income? Explain why.

(c) (2pt) Now we take logarithm on *psoda* and *income* and re-estimate the model above.

$$ln(psoda) = b_0 + b_1 prpminor + b_2 ln(income) + e$$

If *prpminor* increases by 0.2, what is the estimated percentage change in psoda? Do you want to also take logarithm on *prpminor*? Explain why.

(d) (2pt) Now add the variable *prppov* to the regression in part (c). What happens to the coefficient of *prpminor*? Explain why.

(e) (3pt) Now add *house* and estimate another model,

$$ln(psoda) = b_0 + b_1 prpminor + b_2 ln(income) + b_3 propov + b_4 ln(house) + e$$

Given the results of the previous regressions, which one would you report as most reliable in determining whether the racial makeup of a postal code influences soda prices in local fast-food restaurants?

4. (9pt) Use the data in Q4.csv to answer this question. The data set includes information on education(*educ*), parents' education (*motheduc* and $fatheduc$), and ability level (*abil*, it can take negative value. The greater value, the higher ability level.) for 1,230 working men .

(a) (4pt) Estimate the regression model

$$educ = b_0 + b_1 motheduc + b_2 fatheduc + e$$

How much sample variation in educ is explained by parents' education? Please test if mother's education and father's education have the same impact.

(b) (3pt) Now estimate an equation where *abil* appears in both linear and quadratic form:

$$educ = b_0 + b_1 meduc + b_2 feduc + b_3 abil + b_4 abil^2 + e$$

Does "ability" help to explain variations in education, even after controlling for parents' education? Explain. Keep other variables unchanged, find the value of *abil* where *educ* is minimized.

(c) (2pt) Check the fraction of men in the sample have "ability" less than the value calculated in part (b). Does it make sense to include $abil^2$ in the regression? Please explain.

5. (5pt) In our house price example, we get the following regression result. All coefficients are significant.

$$price = 1441.263 + 0.0459152 age^2 - 11.83924 age$$

Suppose a family bought a brand new property and lives there for one year. How much does this property depreciate given everything unchanged? Suppose another family bought a 50-year old property and also lives there for one year. How much does this property depreciate given everything unchanged? By comparing these two numbers, what can you conclude?

6. (24pt) Use the data in Q6.cvs to answer this question, which is about 660 household's apple purchase decisions. Each household head was presented with a description of organic apples, along with prices of regular apples (*regularp*) and prices of the hypothetical organic apples (*organicp*). The variable we would like to explain, *organickg*, is the (hypothetical) kilograms of organic apples a family would like to buy. Price pairs of regular apples and organic apples were randomly assigned to each family. In addition, the data include household size (*hhsize*), household head's age (*age*), education level (*educ*), and household income (*hhincome*).

(a) (2pt) Run the regression of logarithm transformed *organickg* on logarithm transformed *organicp* and *regularp*. Report the result. Interpret the coefficients on the price variables.

(b) (1pt) Do you think the price variables together do a good job of explaining variation in *organickg*? Explain.

(c) (3pt) Run separate simple regressions of logarithm transformed *organickg* on logarithm transformed *organicp* and then logarithm transformed *organickg* on logarithm transformed *regularp*. How do the simple regression coefficients compare with the multiple regression from part (a)? Explain the difference.

(d) (2pt) Add logarithm transformed *hhincome* to the regression from part (a). Interpret the coefficient of *hhincome*. Is it as what you would expect?

(e) (2pt) Define a binary variable as $organicbuy = 1$ if $organickg > 0$ and $organicbuy = 0$ if $organickg = 0$. Estimate the linear probability model

$$organicbuy = b_0 + b_1 log(organicp) + b_2 log(regularp) + b_3 log(hhincome) + e$$

and report the results in the usual form. Carefully interpret the coefficients on the price variables.

(f) (2pt) Using the linear probability model in part (e), to predict the probability of purchasing organic apples. Compute the proportion of predicted probabilities that are either smaller than 0 or greater than 1.

(g) (2pt) Estimate the following regression using logistic regression and probit regression

$$organicbuy = b_0 + b_1 log(organicp) + b_2 log(regularp) + b_3 log(hhincome) + e$$

Compare the significance level of each variable in LPM, logistic regression and probit regression and what can you conclude?

(h) (2pt) Compute the marginal effect of *organicp* at the average using the results of logit. Compare the marginal effects of *organicp* from LPM and logit and what can you conclude?

(i) (2pt) Predict purchasing probabilities using the results of logistic regression and probit regressions. Compute the correlation coefficients of the predicted probabilities by LPM, logistics and probit. What can you conclude?

(j) (4pt) For the prediction made by logistic regression, compute the percent correctly predicted for each outcome, $organicbuy = 0$ and $organicbuy = 1$. Which outcome is best predicted by the model? (If the predicted purchasing probability is greater than 0.5, the we predict that the household will purchase. Otherwise, we predict that the household will not purchase)

7. (11pt) The data set Q7.csv includes state-level panel data on traffic fatalities for the 48 continental U.S. states) from 1980 through 2004. $totfatrte$ is the total traffic fatalities per 100,000 population in each state and year. *bac*08 indicates whether by law drivers are considered as legally intoxicated, if the blood alcohol content (BAC) is 0.08% or higher. Usually *bac*08 takes value 0 or 1. However, if the law was enacted in the middle of the year, the fraction of the year is computed for *bac*08. *perc*14_24 is the percent population aged 14 through 24. *vehicmilespc* is the number of miles driven per capita each year.

(a) (4pt) Design and estimate a regression to examine whether in general driving becomes safer over 1980-2004 disregarding the change of driving laws or drivers demographics and driving habits.

(b) (2pt) Run a regression of $totfatrte$ on $bac08$, $perc14\_24$ and $vehicmilespc$. Interpret the coefficients on $bac08$. Does strict law on driving have a negative effect on the fatality rate?

(c) (2pt) Add year fixed effect to the regression in part (b). How are the coefficient on $bac08$ different from part (b). Explain why.

(d) (3pt) Add state fixed effect to the regression in part (c), in addition to year fixed effect. Compare the coefficient on $bac08$ estimated in part (b) and (c). Which regression is the most reliable model to evaluate the impact of driving law on traffic fatalities? Explain why.

8. (7pt) The data set Q8.csv contains panel data on school districts in Michigan for the years 1992 through 1998. The response variable of interest in this question is $math4$, the percentage of fourth graders in a district receiving a passing score on a standardized math test. The key explanatory variable is $lrexpp$, which is the logarithm of real expenditures per pupil in the district. The amounts are in 1997 dollars. We want to study if increasing the school budget can improve student's math performance. Meanwhile, the data also include another 2 independent variables: $lenrol$ and $lunch$. $lenrol$ is the logarithm of total district enrollment and $lunch$ is the percentage of students in the district eligible for the school lunch program, so $lunch$ is a pretty good measure of the district-wide poverty rate. $distid$ is the school district id and $year$ is the year.

   (a) (2pt) Use OLS to regress $math4$ on $lrexpp$, $lenrol$ and $lunch$. Explain the coefficient on $lrexpp$. What can you conclude about the education expenditure and math scores?

   (b) (2pt) Add year fixed effect to the regression in part (a) and re-estimate the regression above. Compare the magnitude of coefficients on $lrexpp$, $lenrol$ and $lunch$ with the results in part (a). Explain the difference.

   (c) (3pt) Besides year fixed effect, add school district fixed effect to the regression in part (b). What can you conclude about the education expenditure and math scores now? Compare with the conclusion you get in part (a) and explain the difference. Which regression is more reliable to evaluate if education expenditure can increase math scores?

9. (4pt) Can we estimate the following model using choice data?

$$U_{ij} = k_j + \beta x_j + \gamma z_i + e_{ij}$$

   where $i = 1, ..., N$ denotes consumer and $j = 1, ..., J$ denotes brand.

   k_j is the dummy variable of alternative j. x_j is a vector of brand characteristics (e.g., price, promotion), z_i is a vector of consumer characteristics (e.g., age, gender, income). Explain why or why not.

10. (14pt) The dataset Q10.csv contains 2779 travelers' transit choices among bus, train, car and air. Variables include:

    $case$ the individual index
    alt the alternative, one of train, car, bus and air

choice one if the mode is chosen, zero otherwise
cost monetary cost
ivt in vehicle time
ovt out vehicle time
$frequency$ frequency
income income

(a) (2pt) List variables that are alternative attributes and variables that are traveler's attributes.

(b) (2pt) Estimate the following choice model

$$V_j = c_j + \alpha ivt_j + \beta ovt_j + \gamma_j income_i$$

Note that you need to use the command below to declare the choice data first.

```
transit = mlogit.data(transit, shape='long', choice='choice',
alt.var = 'alt', chid.var = 'case')
```

(c) (2pt) Based on the results in (b), as income increases, which transit mode becomes more favorable and which one becomes less favorable?

(d) (2pt) Compare the coefficients of *ivt* and *ovt*. What can you infer?

(e) (3pt) Compute the marginal effect of *income* at the average and which transit mode will be affected most if income increases?

(f) (3pt) Can you estimate the following choice model?

$$V_j = c_j + \alpha_j ivt_j + \beta_j ovt_j + \gamma_j income_i$$

If you can, report the results. If you cannot, explain why not.