HOMEWORK 2

Question 1

The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u,$$

where sleep and totwrk (total work) are measured in minutes per week and educ and age are measured in years.

1. If adults trade off sleep for work, what is the sign of β_1?
2. What signs do you think β_2 and β_3 will have?

   Using the data in SLEEP75.RAW, the estimated equation is

   $$sleep = 3,638.25 - .148 totwrk - 11.13 educ + 2.20 age + u,$$

   $$n = 706, R^2 = .113$$

   If someone works five more hours per week, by how many minutes is sleep predicted to fall? Is this a large tradeoff?
3. Discuss the sign and magnitude of the estimated coefficient on educ.
4. Would you say totwrk, educ, and age explain much of the variation in sleep? What other factors might affect the time spent sleeping? Are these likely to be correlated with totwrk?

Question 2

In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

1. In the model

   $$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u,$$

   does it make sense to hold sleep, work, and leisure fixed, while changing study?
2. Explain why this model suffers from a multicollinearity issue.
3. How could you reformulate the model so that it does not suffer from multicollinearity?

Question 3

The following equation represents the effects of tax revenue mix on subsequent employment growth for the population of counties in the United States:

$$growth = \beta_0 + \beta_1 share_P + \beta_2 share_I + \beta_3 share_S + \text{other factors},$$

where *growth* is the percentage change in employment from 1980 to 1990, $share_P$ is the share of property taxes in total tax revenue, $share_I$ is the share of income tax revenues, and $share_S$ is the share of sales tax revenues. All of these variables are measured in 1980. The omitted share, $share_F$, includes fees and miscellaneous taxes. By definition, the four shares add up to one. Other factors would include expenditures on education, infrastructure, and so on (all measured in 1980).

1. Why must we omit one of the tax share variables from the equation?

2. Give a careful interpretation of β_1

Question 4

The following model can be used to study whether campaign expenditures affect election outcomes:

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 prtystrA + u,$$

where voteA is the percentage of the vote received by Candidate A, expendA and expendB are campaign expenditures by Candidates A and B, and prtystrA is a measure of party strength for Candidate A (the percentage of the most recent presidential vote that went to A's party).

1. What is the interpretation of β_1?

2. In terms of the parameters, state the null hypothesis that a 1% increase in A's expenditures is offset by a 1% increase in B's expenditures.

3. Estimate the given model using the data in VOTE1.DTA and report the coefficients and standard errors. Do A's expenditures affect the outcome? What about B's expenditures?

4. Are β_1, β_2, and β_3 jointly significant at a 0.1% level? Explain. State clearly the null hypothesis, the test you are using, the critical value, and the test statistic. Can you use these results to test the hypothesis in part 2?

5. Estimate a model that directly gives the t statistic for testing the hypothesis in part 2. What do you conclude? (Use a two-sided alternative.)

Question 5

Use the data in WAGE2.DTA for this exercise.

1. Estimate the model

   $$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u$$

   and report the coefficients. 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 6

Use the data in HPRICE1.DTA for this exercise.

1. Estimate the model

   $$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

   Do you reject or not the null hypothesis of homoskedastic standard errors at a 1% level? Test the hypothesis by regressing $\hat{u}^2$ on $\hat{y}$ and $\hat{y}^2$ and state clearly the test you are using, the value of the test statistic, and the critical value.

2. Re-estimate the model with heteroskedasticity-robust standard errors and compare the results.

3. Now estimate the model

   $$log(price) = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

   Do you reject or not the null hypothesis of homoskedastic standard errors at a 1% level? Test the hypothesis by regressing $\hat{u}^2$ on $\hat{y}$ and $\hat{y}^2$ and state clearly the test you are using, the value of the test statistic, and the critical value.

4. Compare the results in part 1 and part 2 of the problem. On the basis of this example, what do you conclude about how to reduce heteroskedasticity?

Question 7

The data set DRIVING.DTA includes state-level panel data (for the 48 continental U.S. states) from 1980 through 2004, for a total of 25 years. Various driving laws are indicated in the data set, including the alcohol level at which drivers are considered legally intoxicated. There are also indicators for ?per se? laws?where licenses can be revoked without a trial?and seat belt laws. Some economics and demographic variables are also included.

1. How is the variable totfatrte defined? What is the average of this variable in the years 1980, 1992, and 2004? Run a regression of totfatrte on dummy variables for the years 1981 through 2004, and describe what you find. Did driving become safer over this period? Explain.

2. Add the variables bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, and vehicmilespc to the regression from part 1. Interpret the coefficients on bac8 and bac10. Do per se laws have a negative effect on the fatality rate? What about having a primary seat belt law? (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)

3. Reestimate the model from part 2 using fixed effects (at the state level). How do the coefficients on bac08, bac10, perse, and sbprim compare with the pooled OLS estimates? Which set of estimates do you think is more reliable?

4. Suppose that vehicmilespc, the number of miles driven per capita, increases by 1,000. Using the FE estimates, what is the estimated effect on totfatrte? Be sure to interpret the estimate as if explaining to a layperson.

5. If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model then the standard errors in part (iii) are invalid. If possible, use ?cluster? robust standard errors for the fixed effects estimates. What happens to the statistical significance of the policy variables in part 3?

Question 8

Use the data in AIRFARE.RAW for this exercise. We are interested in estimating the model

$$fare_{it} = \nu_t + \beta_1 concen_{it} + \beta_2 log(dist_i) + a_i + u_{it}, \qquad t = 1, 2, 3, 4$$

where ν_t means that we allow for different year intercepts.

1. Estimate the above equation by OLS (ignoring the a_i term), being sure to include year dummies. If $\Delta concen = .10$ (the change is .10), what is the estimated percentage increase in fare?

2. Now estimate the equation using random effects. How does the estimate of β_1 change?

3. Now estimate the equation using fixed effects. How does the estimate of β_1 change relative to OLS and random effects? Why is the $log(dist_i)$ dropped from the equation?

4. Name two characteristics of a route (other than distance between stops) that are captured by a_i. Might these be correlated with $concen_{it}$?

5. Use the Hausman test to check whether a fixed effects of a random effects model is appropriate.