**POLS 6481, Spring 2021**

Instructor: Prof. Scott Basinger

Homework Assignment 5

Distributed Thursday, April 15, 2021

Due Thursday, April 22, 2021

Part I. Download the dataset SLEEP75.DTA, and open it in R. To select 580 married adults for the sample, run the following code: married <- subset(sleep, marr == 1)

The dependent variable is you will use is ***sleep*** measured as the average number of minutes per week. You will use three independent variables:

***totwrk*** is the number of minutes worked per week, from 0 (unemployed) to 6415 (nearly 107 hours);

***age*** measures the respondent’s age;

***agesq*** equals age2.

1. Estimate the following regression model using ordinary least squares:

* Are any of the estimated coefficients statistically significant?
* How much of the variation in sleep do these variables explain?
* Are you concerned about multicollinearity due to an association between *age* and *age*2? Perform an appropriate test and report your findings.

2. The mean value of ***totwrk*** in the dataset equals 2112 minutes (slightly more than 35 hours weekly). Using the model estimated in 1, calculate the predicted sleep for a (hypothetical) 25-, 35-, 45-, and 55-year-old who works the average amount; enter that number in the first row in the table atop page 2.

3. What is the estimated equation for the marginal effect of age on sleep? State it both in the abstract and substituting values from the regression model estimated in 1.

=

=

3½. At what value of age does the parabola capturing the relationship between age and sleep attain its vertex? Is this a global minimum or maximum?

4. What is the estimated equation for the standard error of the marginal effect of age on sleep? Again, state it both in the abstract and substituting values from the regression model estimated in 1.

s.e.() =

=

5. Using the vector of estimated coefficients, the variance-covariance matrix of the estimators, and your answers to 2., 3., and 4., fill in the table shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | 25 | 35 | 45 | 55 |
| Prediction *E*(|*age*) |  |  |  |  |
| Marginal effect () |  |  |  |  |
| Standard error *s.e.*() |  |  |  |  |
| *t*-statistic |  |  |  |  |

5½. Check your answers in the table using the code from lab or lecture examples. You’ll need to write equations for the predicted values (*yhat*), marginal effects (*dydx*), and standard errors of marginal effects (*sedydx*). Then, substitute values of ***age*** = 25, 35, 45 and 55, and ***age*2** = 625, 1225, 2025, and 3025, respectively, while holding ***totwrk*** equal to 2112.

6. Adapt the code presented in lab or the lecture examples to plot the predicted values () as age varies, while holding ***totwrk*** equal to 2112. (If possible within a reasonable period of time, also try to plot the prediction interval around .)

7. Adapt the code presented in lab or the lecture examples to plot the marginal effect curve () with confidence intervals.

8. Just for fun, recode the age and age-squared variables by *de-meaning* them. That is, find the average value of age, create a new variable (age – mean age), and create another new variable that is the square of (age – mean age). Check the correlation on these variables. Then, re-run the model in 1. using these new variables, and compare the results by responding to the three bullet points:

* Are any of the estimated coefficients statistically significant?
* How much of the variation in sleep do these variables explain?
* Are you concerned about multicollinearity due to an association between *age* and *age*2? Perform an appropriate test and report your findings.

If there are any differences between the models in 1. and 8., what explains the differences?

Part II. Download the dataset DISCRIM.DTA. These are ZIP code–level data on prices for various items at fast-food restaurants, along with characteristics of the ZIP code’s population. These data were used in K. Graddy (1997) “Do Fast-Food Chains Price Discriminate on the Race and Income Characteristics of an Area?” [*Journal of Business and Economic Statistics* 15: 391 – 401] Her goal was to explore whether fast-food restaurants charge higher prices in areas with a larger concentration of Black residents.

The dependent variable is you will use is ***pfries***, measured as the average price of french fries in dollars. Prices were calculated by visiting stores in four fast-food chains (*Burger King*, *Kentucky Fried Chicken*, *Roy Rogers*, and *Wendy’s*) in two states (New Jersey and Pennsylvania).

You will use two main independent variables:

***prpblck*** is the proportion of residents in a ZIP code who are Black;

***hseval*** is the median home value in a ZIP code. There are other indicators of a ZIP code’s prosperity in the dataset (median family income, proportion of residents living in poverty, etc.), but they are all highly correlated to median home values.

The dataset also includes four dummy variables you can use:

***NJ*** indicates whether the ZIP code is in New Jersey ( = 1) or Pennsylvania ( = 0).

***BK*** indicates whether the restaurants visited were *Burger King* franchises

***KFC*** indicates whether the restaurants visited were *Kentucky Fried Chicken* franchises

***RR*** indicates whether the restaurants visited were *Roy Rogers* franchises

Obviously, *Wendy’s* franchises are the omitted category.

9. Estimate the following regression model using ordinary least squares (plus any dummies you choose to add):

* Report the results in equation form, including the sample size and R-squared.
* Are any of the estimated coefficients statistically significant?
* Interpret the coefficient on *prpblck*; do you think it is substantively large?

10. Since the dependent variable and one independent variable are in dollars, a log-log model might be more appropriate. Estimate the following regression model using ordinary least squares (plus any dummies you choose to add):

* Report the results in equation form, including the sample size and R-squared.
* Are any of the estimated coefficients statistically significant?
* Interpret the coefficient on *prpblck*
* Interpret the coefficient on *log*(*income*)

11. Use the R scripts from lab and lecture examples to translate the predicted value of back into predicted values of , and then compare how well the log-log model fits compared to the level-level model.