ECON 640 - PROBLEM SET 4

Instructions: *For the theory section, submit your individual answers in printed form. For the empirical section, each student must upload their own work to the* ***same GitHub repository*** *used previously (rename it to serve as your* ****Econ 640 Coding Hub****). Each problem set should be stored in a separate folder, and you do not need delete any previous homework. Include the cleaned do-file and all results (figures and tables). This problem set is due* ***at the beginning of class on **December 2nd*****.

1. Suppose y_i is a binary outcome representing whether Trump wins in state i (1 if Trump wins, 0 otherwise). Let x_i be the independent variable representing each state's employment-to-working-age population ratio, measured from 0 to 100 (e.g., 50 means 50 percent of the working-age population are employed).

   (a) First, model this relationship using the linear probability model:

   $$y_i = \beta_0 + \beta_1 x_i + u_i$$

   Derive the OLS estimators for β_0 and β_1 using the Maximum Likelihood Estimation (MLE) approach.

   (b) Next, estimate this relationship using a logit model. Derive the log-likelihood function for the logit estimator of β_0 and β_1 using MLE. Since a closed-form solution is not attainable, take the derivation to the steps where it could be solved numerically by a computer.

   (c) Suppose you have estimated $\hat{\beta}$ (a vector of the estimated parameters). Derive the marginal effects based on the estimated $\hat{\beta}$ values.

   (d) Now, consider modeling the relationship using a log-linear model:

   $$y_i = \beta_0 + \beta_1 \ln(x_i) + u_i$$

   How does the interpretation of the coefficient β_1 differ in this model compared to the linear probability and logit models?

2. Suppose the data structure now involves y_{is}, a binary outcome representing whether individual i voted for Trump in state s (1 if yes, 0 otherwise). Let x_s represent the employment-to-population ratio in state s. You want to model the following relationship using the Linear Probability Model (LPM):

   $$y_{is} = \beta_0 + \beta_1 x_s + u_{is}$$

(a) Write down the variance of the OLS estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ in matrix form.

(b) Suppose you believe there is heteroskedasticity in the data. How would you adjust the standard errors of $\hat{\beta}$ to account for heteroskedasticity? Specify the estimator for the variance of $\hat{\beta}$.

(c) If you believe there is serial correlation within each state but not across states, what estimator for the variance of $\hat{\beta}$ would you propose? Write down the estimator and explain how you would account for the within-state correlation.

(d) Finally, if you believe there is only heteroskedasticity and aim to find a BLUE estimator, what estimator should you use to feasibly estimate the parameters? Write down the estimator and outline the steps required for estimation. What is the variance of this estimator?

3. Suppose you now have panel data tracking individuals' voting behavior over time. Let y_{it} be a binary outcome representing whether individual i voted for a Republican candidate in year t (1 if yes, 0 otherwise). Let x_{it} represent individual i's employment status in year t. You aim to model the following relationship using the Linear Probability Model:

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}$$

(a) Suppose you believe an individual's voting behavior is autocorrelated over time. How would you adjust the OLS estimator's standard errors to account for this autocorrelation? Specify the estimator you would propose and write it down.

4. Empirical Task with non-simulated datasets:

(a) Use the 'sysuse auto' dataset in Stata to analyze heteroskedasticity.

- Load the dataset with the command 'bcuse auto'. Estimate a model with `price` as the dependent variable and `weight` and `mpg` (miles per gallon) as independent variables:

```
regress price weight mpg
```

After estimating the model, conduct a Breusch-Pagan test for heteroskedasticity using the command:

```
estat hettest
```

Interpret the results of the Breusch-Pagan test and discuss whether heteroskedasticity is present (learn how to utilize and interpret this test yourself).

- Re-estimate the model from (a) using White's heteroskedasticity-robust standard errors. Run the following command:

```
regress price weight mpg, robust
```

Compare the standard errors with those from the original OLS estimation. Explain why the robust standard errors may differ from the OLS standard errors.

(b) Use the 'sysuse nlsw88' dataset to estimate a model of wage with clustered standard errors.

- Load the dataset with 'bcuse nlsw88'. Estimate a model with `wage` as the dependent variable and `age` and `collgrad` (college graduate indicator) as independent variables:

```
regress wage age collgrad
```

- Now, re-estimate this model with cluster-robust standard errors by `state`. Use the command:

```
regress wage age collgrad, cluster(state)
```

Compare the cluster-robust standard errors with the original OLS standard errors. Explain why the clustered standard errors might differ, especially when considering correlation within states.

(c) Use the 'bcuse cps91' dataset to explore the interpretation of a linear-log model.

- Load the dataset with 'bcuse cps'. Estimate a linear-log model where `hrwage` is the dependent variable and `exper` (years of experience) is the independent variable. Transform `exper` using the natural log:

```
gen lnexper = ln(exper)
```

Then run the regression:

```
regress wage lnexper
```

- Interpret the coefficient on `lnexper`. Explain what it implies about the relationship between experience and wage in terms of percentage changes.
- Compare the interpretation of the `lnexper` coefficient in this model to a model with experience as a linear predictor (without logging `exper`). What does the linear-log model capture that a linear model might miss?

(d) Use the `bcuse bwght2` dataset (`http://fmwww.bc.edu/ec-p/data/wooldridge/bwght2.des`), which focuses on low birth weight, to apply a logit model. Define the variables you need based on the data's codebook.

- Estimate a logit model where `low` (an indicator for low birth weight) is the dependent variable and `smoke` (smoking during pregnancy) and `age` (mother's age) are independent variables:

```
logit low smoke age
```

- After estimating the model, calculate the marginal effects for the `smoke` variable by using:

  ```
  margins, dydx(smoke)
  ```

  Interpret the marginal effect of `smoke` on the probability of low birth weight.
- Compare the interpretation of the coefficients from the logit model to what they would represent in an OLS model. Discuss why the interpretation of the logit coefficients is not as straightforward as in the linear model.