This homework is meant to give you additional practice with with binary response models (BRM). Binary response models are used ubiquitously throughout econometrics for causal inference as well as in machine learning for prediction so, aside from it being required for a grade, it is of your own personal interest to take this homework seriously.

To receive full credit for this homework, your D2L submission should be two files:

1. Your solutions for Problems 1 through 4 in the file "Lastname_Firstname_HW2.pdf"
2. Your R code used to solve Problems 2 and 3 in the file "Lastname_Firstname_HW2.R"
   - Please use the R script template "ECON_418-518_HW2_R_Template.R" to format your solutions for these problems.
   - You must comment each line of code you write!

This homework is worth 100 points (10%) of your grade.

Problem 1 (25 Points)

(E) Problem 17.2. Let *grad* be a dummy variable for whether a student-athlete at a large university graduates in five years. Let *hsGPA* and *SAT* be high school grade point average and SAT score, respectively. Let *study* be the number of hours spent per week in an organized study hall. Suppose that, using data on 420 student-athletes, the following logit model is obtained:

$$\widehat{\mathbb{P}}(grad = 1 \mid hsGPA, SAT, study) = \Lambda(-1.17 + 0.24 \cdot hsGPA + 0.00058 \cdot SAT + 0.073 \cdot study),$$

where

$$\Lambda(z) = \frac{\exp(z)}{1 + \exp(z)}$$

is the logit function. Holding *hsGPA* fixed at 3.0 and *SAT* fixed at 1,200, compute the estimated difference in the graduation probability for someone who spent 10 hours per week in study hall and someone who spent 5 hours per week.

Problem 2 (35 Points)

(E) Problem C17.1. Use the data in PNTSPRD for this exercise.

(i) The variable *favwin* is a binary variable that equals 1 if the team favored by the Las Vegas point spread wins. A linear probability model to estimate the probability that the favored team wins is

$$P(favwin = 1 \mid \mathit{spread}) = \beta_0 + \beta_1 \mathit{spread}.$$

Explain why, if the spread incorporates all relevant information, we expect $\beta_0 = 0.5$.

(ii) Estimate the model from part (i) by OLS. Test $H_0 : \beta_0 = 0.5$ against a two-sided alternative at the $\alpha = 0.05$ level. Use both the usual and heteroskedasticity-robust standard errors.

(iii) Is *spread* statistically significant? What is the estimated probability that the favored team wins when *spread* $= 10$?

(iv) Now, estimate a probit model for $P(\mathit{favwin} = 1 \mid \mathit{spread})$. Interpret and test the null hypothesis that the intercept is zero at the $\alpha = 0.05$ level. [Hint: Remember that $\Phi(0) = 0.5$].

(v) Use the probit model to estimate the probability that the favored team wins when *spread* $=$ 10. Compare this with the LPM estimate from part (iii).

(vi) Add the variables *favhome*, *fav25*, and *und25* to the probit model and test the joint significance of these variables using the likelihood ratio test at the $\alpha = 0.05$ level. (How many degrees of freedom are in the chi-square distribution?) Interpret this result, focusing on the question of whether the spread incorporates all observable information prior to a game.

Problem 3 (30 Points)

(E) Problem C17.2 with different Part (iv). Use the data in LOANAPP for this exercise.

(i) Estimate a probit model of *approve* on *white*. Find the estimated probability of loan approval for both whites and nonwhites. How do these compare with the linear probability estimates?

(ii) Now, add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2*, and *vr* to the probit model. Is there statistically significant evidence of discrimination against nonwhites at the $\alpha = 0.05$ level?

(iii) Estimate the model from part (ii) by logit. Compare the coefficient on *white* to the probit estimate.

(iv) Estimate the average marginal discrimination effect for probit and logit.

Problem 4 (10 Points)

Suppose MLR Assumptions 1-6 hold. Let the outcome variable $\boldsymbol{y}$ be binary or continuous. Show that when estimating the model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$ by maximum likelihood estimation (MLE) we get the same solution as that given by ordinary least squares (OLS) $\left(\widehat{\boldsymbol{\beta}} = (X'X)^{-1} X'\boldsymbol{y}\right)$ when we assume the errors are distributed according to the multivariate normal distribution conditional on X, i.e., $\boldsymbol{u} \mid X \sim \mathbb{N}(\mathbf{0}, \sigma^2 I_n)$. Some tips:

1. Since the errors are normally distributed conditional on the covariates, this implies the outcome conditional on the covariates is also normally distributed, but with a different mean. Can you show why this holds and what the conditional mean and variance is?

2. The conditional distribution of $\boldsymbol{y}$ given X is

$$f_{Y|X}(\boldsymbol{y} \mid X; \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{y} - X\boldsymbol{\beta})' (\boldsymbol{y} - X\boldsymbol{\beta})\right).$$

   This is the PDF of the multivariate normal distribution (the conditional mean of the outcome given the covariates is given in this expression). Since $\boldsymbol{y} \mid X$ is normally distributed, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2) = (\beta_0, \beta_1, \ldots, \beta_k, \sigma^2)$. So, we can now set up the log-likelihood function.

3. Now, it comes down to solving the problem

$$\underset{\boldsymbol{\beta}}{\arg\max} \left[\ln\left(f_{Y|X}\left(\boldsymbol{y} \mid X; \boldsymbol{\beta}, \sigma^2\right)\right)\right].$$

   Notice we avoid the summation operator because we are using vector/matrix notation.

   This involves taking logs of the PDF and taking the first order condition with respect to $\boldsymbol{\beta}$. Then, you solve for $\widehat{\boldsymbol{\beta}}$ in a similar fashion to how we derived the OLS solution in class. An important note is maximizing the negative of a function is the same as minimizing the function itself.

Problem 5 (5 Points Extra Credit)

Suppose y_i is a binary variable generated by the probit model

$$\begin{aligned} y_i &= \mathbb{1}(\beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + u_i > 0) \\ u_i \mid \boldsymbol{x}_i &\sim \mathbb{N}(0, 1) \end{aligned}$$

where β_0 is an unobserved intercept term and $\boldsymbol{\beta}$ is a k-dimensional vector of slope coefficients for the non-constant (observable) regressors $\boldsymbol{x}_i$.

Given a random sample $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, show that the restricted MLE of β_0 under the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ (i.e., $\beta_1 = \ldots = \beta_k = 0$) is a simple function of the sample average of the dependent variable.

Some tips:

1. Assume all slope coefficients are zero so that $y_i = \beta_0 + u_i$.
2. Use the CDF of the negative of u_i to derive the probabilities that $y_i = 1$ and $y_i = 0$. This is the standard normal CDF given by Φ.
3. Set up the log-likelihood function.
4. Take first order conditions. The derivative of $\ln\left[\Phi\left(\widehat{\beta}_0\right)\right]$ is $\frac{\phi(\widehat{\beta}_0)}{\Phi(\widehat{\beta}_0)}$ where ϕ is the PDF of the standard normal distribution.
5. Remember that Φ^{-1} (inverse CDF function) exists for the standard normal distribution.
6. Solve for $\widehat{\beta}_0$.