Problem set 6

Due on Thursday November 17, 2022 (by 11:59 PM EST)

Note: No credit will be given if you report only the final answers without showing formulas and calculations when appropriate. This applies to both theoretical and empirical questions. For the empirical questions, make sure to submit the R scripts and output on Latte. No credit will be given if the R output is missing.

Problem 1

A study tried to find the determinants of the increase in the number of households headed by a female. Using 1940 and 1960 historical census data, a logit model was estimated to predict whether a woman is the head of a household (living on her own) or whether she is living within another's household. The limited dependent variable takes on a value of one if the female lives on her own and is zero if she shares housing. The results for 1960 using 6,051 observations on prime-age whites and 1,294 on nonwhites were as shown in the table:

Regression	**(1) White**	**(2) Nonwhite**
Regression model	Logit	Logit
Constant	1.459 (0.685)	-2.874 (1.423)
Age	-0.275 (0.037)	0.084 (0.068)
age squared	0.00463 (0.00044)	0.00021 (0.00081)
education	-0.171 (0.026)	-0.127 (0.038)
farm status	-0.687 (0.173)	-0.498 (0.346)
South	0.376 (0.098)	-0.520 (0.180)
expected family earnings	0.0018 (0.00019)	0.0011 (0.00024)
family composition	4.123 (0.294)	2.751 (0.345)
Pseudo-R^2	0.266	0.189
Percent Correctly Predicted	*82.0*	*83.4*

where *age* is measured in years, *education* is years of schooling of the family head, *farm status* is a binary variable taking the value of one if the family head lived on a farm, *south* is a binary variable for living in a certain region of the country, *expected family earnings* was generated from a separate OLS regression to predict earnings from a set of regressors, and *family composition* refers to the number of family members under the age of 18 divided by the total number in the family.

The mean values for the variables were as shown in the table.

Variable	(1) White mean	(2) Nonwhite mean
age	46.1	42.9
age squared	2125.21	1840.41
education	12.6	10.4
farm status	0.03	0.02
south	0.3	0.5
expected family earnings	2,336.4	1,507.3
family composition	0.2	0.3

(a) Interpret the results. Do the coefficients have the expected signs? Why do you think age was entered both in levels and in squares?
(b) Calculate the difference in the predicted probability between whites and nonwhites at the sample mean values of the explanatory variables. Why do you think the study did not combine the observations and allowed for a nonwhite binary variable to enter?
(c) What would be the effect on the probability of a nonwhite woman living on her own, if *education* and *family composition* were changed from their current mean to the mean of whites, while all other variables were left unchanged at the nonwhite mean values?

Problem 2

The logit regression in Chapter 11 of your textbook reads:

$$\overline{Pr(deny=1 \mid P/Iratio, black)} = F(-4.13 + 5.37\ P/Iratio + 1.27\ black)$$

(a) Using a spreadsheet program such as Excel, plot the following logistic regression function with a single *X*, $\hat{Y}_i = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i})}}$ where $\hat{\beta}_0 = -4.13$, $\hat{\beta}_1 = 5.37$, $\hat{\beta}_2 = 1.27$.

Enter values for X_1 in the first column starting from 0 and then increment these by 0.1 until you reach 2.0. Let X_2 be 0 at first. Then enter the logistic function formula in the next column. Next allow X_2 to be 1 and calculate the new values for the logistic function in the third column. Finally produce the predicted probabilities for both blacks and whites, connecting the predicted values with a line.

(b) Using the same spreadsheet calculations, list how the probability increases for blacks and for whites as the *P*/*I* ratio increases from 0.5 to 0.6.

(c) What is the difference in the rejection probability between blacks and whites for a *P*/*I* ratio of

0.5 and for 0.9? Why is the difference smaller for the higher value here?

(d) Table 11.2 on page 368 of your textbook lists a logit regression (column 2) with further explanatory variables. Given that you can only produce simple plots in two dimensions, how would you proceed in (a) above if there were more than a single explanatory variable?

Problem 3

You have a sample size of size $n = 1$ with data $y_1 = 2, x_1 = 1$. You are interested in the value of β in the regression $Y = X\beta + u$ (Note that there is no intercept).

a. Plot the sum of squared residuals $(y_1 - bx_1)^2$ as a function of b.
b. Show that the least squares estimated of β is equal to 2.
c. Using $\lambda_{Ridge} = 1$, plot the Ridge penalty term $\lambda_{Ridge}b^2$ as a function of b.
d. Using $\lambda_{Ridge} = 1$, plot the Ridge penalized sum of squared residuals, $(y_1 - bx_1)^2 + \lambda_{Ridge}^2 b^2$.
e. Find the value of $\hat{\beta}_{Ridge}$.
f. Using $\lambda_{Ridge} = 0.5$, repeat parts (c) and (d). Also, find the new value of $\hat{\beta}_{Ridge}$.
g. Using $\lambda_{Ridge} = 3$, repeat parts (c) and (d). Also, find the new value of $\hat{\beta}_{Ridge}$.
h. Use the graphs that you produced in parts (a)-(d) for the various values of λ_{Ridge} to explain why a large value of λ_{Ridge} results in more shrinkage of the OLS estimates.

Problem 4 (empirical)

What determines whether organizations fire their leaders? It is often hard for outsiders to observe performance, but in sports, many facets of performance (particularly winning percentages) are easily observed. In this exercise you will use data on the performance and firing of NFL football coaches. The list of variables included in this data is below.

a. Run a Probit model explaining whether the coach was fired as a function of winning percentage. Graph fitted values for this model on the same graph with fitted values from a linear probability model estimated using the same variables. Explain the differences you see in the plot
b. Estimate linear probability, Probit, and Logit models of coach firings by using winning percentage, lagged winning percentage, a new coach dummy, strength of schedule, and coach tenure as independent variables. Are the coefficients across the three models substantially different? How about the t-statistics and p-values?
c. Indicated the minimum, mean, and maximum of the fitted values for each model and briefly discuss each
d. What are the correlations of the three fitted values?
e. It is kind of odd to say that the lag winning percentage affects the probability that the new coaches were fired because they weren't coaching on the years associated with the lagged winning percentages. Include an interaction for new coach dummy variable and lagged winning percentage. The effect of lagged winning percentage on the probability of being fired is now the sum of coefficients on lagged winning percentage and the interaction. Test the null hypothesis that lagged winning percentage has no effect on the probability of being fired for new coaches (meaning coaches for whom *NewCoach*=1)

Variables for Football Coach Data

Variable name	Description
FiredCoach	A dummy variable if the football coach was fired during or after the season (1=fired, 0=otherwise)
WinPct	The winning percentage of the team
LagWinPct	The winning percentage of the team in the previous year
ScheduleStrength	A measure of schedule difficulty based on records of opposing teams
NewCoach	A dummy variable indicating if the coach was new (1=new, 0=otherwise)
Tenure	The number of years the coach has coached the team