Department of Economics | UN 3412
Columbia University | Fall 2021

Problem Set 6
Introduction to Econometrics
Seyhan Erden and Tamrat Gashaw
for all sections

1. (50p) In recent years, public concern about "second-hand" smoke has led to smoking bans in many US workplaces. In some cases, smoking bans are determined by local ordinance that covers indoor workplaces over a certain size, sometimes with exemptions such as for bars or restaurants. In other cases, smoking bans are voluntarily adopted by individual businesses (these voluntary bans were the main type of ban during the time period of the data for this problem set).

It has been conjectured that workplace smoking bans induce smokers to quit by reducing their opportunities to smoke. In this assignment you will estimate the effect of workplace smoking bans on smoking. To do this you will use data on a sample of 10,000 US indoor workers in 1991-1993. The data set contains information on whether individuals were, or were not, subject to a workplace smoking ban, whether or not the individuals smoked, and other individual characteristics. The data are in a STATA dataset called smoking.dta (described below).

(a) (5p) Estimate the probability of smoking for
  (i) all workers (the full sample)
  (ii) workers affected by workplace smoking bans
  (iii) workers not affected by workplace smoking bans

(b) (5p) What is the difference in the probability of smoking between workers affected by a workplace smoking ban and workers not affected by a workplace smoking ban? Use a linear probability model to determine whether this difference is statistically significant.

(c) (5p) Estimate a linear probability model with smoker as the dependent variable and the following regressors: smkban, female, age, age^2, hsdrop, hsgrad, colsome, colgrad, black, and hispanic. Compare the estimated effect of a smoking ban from this regression with your answer from 1(b). Suggest a reason, based on the substance of this regression, explaining the change in the estimated effect of a smoking ban between 1(b) and 1(c).

(d) (5p) Test the hypothesis that the coefficient on smkban is zero in the population version of the regression in part (c), against the alternative that it is nonzero, at the 5% significance level.

(e) (5p) Test the hypothesis that the probability of smoking does not depend on the level of education in the regression in part (c). In words, describe the estimated relationship *between education and smoking (holding the other regressors constant).*

(f) (25p) Fill out the table below.

Estimated Effect on the Probability of Smoking of a Workplace Smoking Ban on Two Hypothetical Workers

Mr. A: male, white, non-hispanic, 20 years old, high school dropout
Ms. B: female, black, 40 years old, college graduate

	Probit Model (1)	**Logit Model (2)**	**Linear Prob. Model (3)**
Estimated coefficient on *smkban* (standard error in parentheses)			
Predicted probabilities of smoking for Mr. A:			
(i) with workplace ban			
(ii) without workplace smoking ban			
Difference, (i) – (ii)			
Predicted probabilities of smoking for Ms. B:			
(iii) with workplace ban			
(iv) without workplace smoking ban			
Difference, (iii) – (iv)			

Notes: The entry in the first row is the estimated coefficient on *smkban* in the probit model (column (1)) the logit model (column (2)), and the linear probability model (column (3)), with standard errors in parentheses; both regressions include the following control variables: *female*, *age*, *age*^{2}, *hsdrop*, *hsgrad*, *colsome*, *colgrad*, *black*, and *hispanic*. The entries in the remaining rows are predicted probabilities of smoking for the indicated hypothetical individuals, and differences in those predicted probabilities.

2. (25p) 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	2,263.5	1,965.6
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) (8p) Interpret the results. Do the coefficients have the expected signs? Why do you think age was entered both in levels and in squares?

(b) (9p) 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) (8p) 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?

3. (25p) **Maximum Likelihood Estimation Method (MLE):**

(a) (12p) Suppose $X_1, X_2, \ldots, X_n$ is a random sample from an **exponential distribution** with parameter λ. Assume that X_i's are independence and the individual pdf is given by:

$$f(x, \lambda) = \lambda e^{-\lambda x}$$

Find the MLE of this function (i.e., $\hat{\lambda}$).

(b) (13p) Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a **Poisson distribution** with parameter alpha. Assume that X_i's are independence and the individual pdf is given by:

$$f(x, \alpha) = \frac{e^{-\alpha}\alpha^x}{x!}$$

Find the MLE of this function (i.e., $\hat{\alpha}$).

<u>The following questions will not be graded, they are for you to practice and will be discussed at recitation:</u>

1. SW Exercise 11.1
2. SW Exercise 11.6
3. SW Exercise 11.7
4. SW Exercise 11.9
5. SW Empirical Exercise 11.3