```
*==============================================================================
*Title: LOANAPP_monica.do
*Author: Monica Elgawly
*Date Modified: 4/17/2018
*Assignment 4 Part 2
*==============================================================================

cap log close
clear
set more off

*==============================================================================

*==============================================================================
*2. In 1975, the U.S. passed the Home Mortgage Disclosure Act. This law requires
*all mortgage lenders to release annual data on all mortgage applications. One
*goal of the act is to identify discriminatory lending--cases where banks do not
*lend to individuals based on race or gender. The data set LOANAPP.dta contains a
*random sample of loan applications from HDMA data. The variables are defined as
*follows:

** approve is a dummy variable equal to one if the loan application is approved
** black, hispan, and male are dummy variables equal to one if the borrower is
*black, hispanic or male respectively
** apr is the annual-percent rate on the loan (i.e. the interest rate)
** term measures the length of time over which the loan will be paid off
** bankruptcy is a dummy variable equal to one if the individual has ever declared
*bankruptcy
** gdlin is a variable equal to one if the borrower's credit history meets the lender's
** typical guidelines
** hh_expenditures is a measure of household expenditures as a share of annual income

*a) Estimate a linear probability model that relates loan approval rates to the
*applicant's demographic characteristics (black, hispan, male), the loan
*characteristics (apr, term), and the applicant's credit worthiness (bankruptcy,
*gdlin, hh_expenditures). Interpret the coefficients on black, hispan and male.
*Although we are interested in the relationship between demographics and loan
approval,
*why is it important to control for loan characteristics and credit worthiness?

use "/Users/monicaelgawly/Downloads/LOANAPP.dta"
reg approve black hispan male apr term bankruptcy gdlin hh_expenditures, r
*==============================================================================

*==============================================================================
*b) Find the predicted values based on the regression in part a. Summarize the
distribution
* of the predicted values. How does the average predicted approval rate compare to the
* average actual approval rate? Do you notice anything unusual about the predicted
values?
* Explain what you see.

predict ~~y2~~ approvrate
sum ~~y2~~, detail
*(where xb should be assumed)
*==============================================================================

*==============================================================================
*c) Repeat the regression in part a using only the regressors that are individually
* statistically significant at the 10% level using t-tests. Do your results change
```

```
* substantially? Are your results sensitive to dropping these controls? What might this
* tell you about OVB?
reg approve black hispan gdlin, r
*================================================================================

*================================================================================
*d) Estimate the relationship in part c using a probit model. What is the difference
*in the probability of approval between black and nonblack applicants who meet the
*borrowing guidelines (gdlin = 1)? How about the difference between hispanic and
*non-hispanic borrowers? Do the results differ much from the linear probability model?

probit approve black hispan gdlin, r

*================================================================================

*================================================================================
*e) Repeat part d using a logit model instead of a probit model. Do your results change?
logit approve black hispan gdlin, r
*================================================================================

*================================================================================
*f) Based on your work in parts a through e, comment on the disadvantages and
*advantages of using the linear probability model vs. probit/logit to estimate models
*with binary dependent variables.

*================================================================================

*================================================================================
*g) Based on your work in parts a through e, what do you conclude about discrimination
*by lenders?
```

The predicted probability of loan approval is consistently seven to eight point eight percent lower for black and Hispanic applicants. The results are consistent regardless of model used. The regressors accounting for gender, apr of the loan, length of loan term, status of bankruptcy and extent of a household's expendenitures were found to be insignificant markers in deciding whether someone receives a loan. Meeting guidelines was the predominant factor and discrimination has a significant enough liklihood that its claim should not be dismissed.

```
*================================================================================

clear
```

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.1   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
                                      College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata license expires 20 Aug 2018:
       Serial number:  301509284846
         Licensed to:  Monica Elgawly
                       Hunter College

Notes:
      1.  Unicode is supported; see help unicode_advice.

1 . log using "/Users/monicaelgawly/Downloads/metrics/do and log files/LOANAPP_monic
> a.log"
```

```
      name:  <unnamed>
       log:  /Users/monicaelgawly/Downloads/metrics/do and log files/LOANAPP_monic
> a.log
  log type:  text
 opened on:  19 Apr 2018, 01:36:19

2 . use "/Users/monicaelgawly/Downloads/LOANAPP.dta"

3 .
4 . reg approve black hispan male apr term bankruptcy gdlin hh_expenditures, r

Linear regression                               Number of obs     =      1,968
                                                F(8, 1959)        =      72.74
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3802
                                                Root MSE          =     .25867
```

approve	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
black	-.0832246	.025738	-3.23	0.001	-.1337013	-.0327479
hispan	-.0814848	.0344091	-2.37	0.018	-.1489672	-.0140024
male	.0080667	.015247	0.53	0.597	-.0218352	.0379687
apr	.0000547	.0000335	1.63	0.103	-.000011	.0001204
term	-.0000194	.0000954	-0.20	0.839	-.0002064	.0001676
bankruptcy	-.0454592	.0324919	-1.40	0.162	-.1091815	.0182632

gdlin	.6706584	.0363856	18.43	0.000	.5993	.7420169
hh_expenditures	-.0012052	.0010511	-1.15	0.252	-.0032666	.0008561
_cons	.2986692	.0539987	5.53	0.000	.1927682	.4045701

```
5 . predict approvrate
(option xb assumed; fitted values)
(15 missing values generated)

6 . sort approvrate

7 . sum approvrate, detail

                        Fitted values
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .1501102       .1257839
 5%      .244849       .1283087
10%       .84996       .1303715       Obs               1,968
25%      .931003       .1353147       Sum of Wgt.       1,968

50%     .9465646                      Mean           .8775407
                        Largest       Std. Dev.       .202182
75%     .9545963       1.021383
90%     .9628758       1.028688       Variance       .0408776
95%     .9694582       1.029386       Skewness      -2.866013
99%     .9852294       1.190825       Kurtosis       9.559944

8 . sum approve, detail

                           approve
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs               1,983
25%            1              0       Sum of Wgt.       1,983

50%            1                      Mean            .878467
                        Largest       Std. Dev.      .3268281
75%            1              1
90%            1              1       Variance       .1068166
95%            1              1       Skewness      -2.316584
99%            1              1       Kurtosis       6.366562

9 . reg approve black hispan gdlin, r

Linear regression                               Number of obs     =      1,983
                                                F(3, 1979)        =     183.37
```

Prob > F	=	0.0000
R-squared	=	0.3777
Root MSE	=	.25801

approve	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
black	-.0879547	.0253942	-3.46	0.001	-.1377569	-.0381525
hispan	-.0851574	.0340472	-2.50	0.012	-.1519296	-.0183853
gdlin	.6892772	.033956	20.30	0.000	.6226839	.7558706
_cons	.2610012	.0341644	7.64	0.000	.1939992	.3280033

```
10 . reg approve black hispan, r
```

Linear regression

Number of obs	=	1,983
F(2, 1980)	=	27.43
Prob > F	=	0.0000
R-squared	=	0.0495
Root MSE	=	.3188

approve	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
black	-.2267144	.034107	-6.65	0.000	-.2936038	-.159825
hispan	-.1429999	.0408399	-3.50	0.000	-.2230935	-.0629062
_cons	.9087657	.0070367	129.15	0.000	.8949656	.9225657

```
11 . probit approve black hispan gdlin, r

Iteration 0:   log pseudolikelihood = -733.64726
Iteration 1:   log pseudolikelihood = -497.25269
Iteration 2:   log pseudolikelihood = -496.61183
Iteration 3:   log pseudolikelihood = -496.61158
Iteration 4:   log pseudolikelihood = -496.61158
```

Probit regression

Number of obs	=	1,983
Wald chi2(3)	=	393.24
Prob > chi2	=	0.0000
Pseudo R2	=	0.3231

Log pseudolikelihood = -496.61158

approve	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
black	-.4917113	.1223809	-4.02	0.000	-.7315735	-.251849

hispan	-.5026078	.1622975	-3.10	0.002	-.820705	-.1845105
gdlin	2.211937	.1195054	18.51	0.000	1.977711	2.446163
_cons	-.5779041	.1143003	-5.06	0.000	-.8019286	-.3538795

```
12 . predict racialcomponent, pr

13 . sum racialcomponent if black == 0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
racialcomp~t	1,788	.899812	.1700964	.1399572	.9488741

```
14 . sum racialcomponent if black == 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
racialcomp~t	195	.68217	.3220565	.1423962	.8733399

```
15 . sum racialcomponent if black == 0 & gdlin == 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
racialcomp~t	1,671	.9444503	.0180236	.871062	.9488741

```
16 . sum racialcomponent if black == 1 & gdlin == 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
racialcomp~t	144	.8733399	0	.8733399	.8733399

```
17 . sum racialcomponent if hispan == 0 & gdlin == 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
racialcomp~t	1,720	.9425503	.0209267	.8733399	.9488741

```
18 . sum racialcomponent if hispan == 1 & gdlin == 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
racialcomp~t	95	.871062	0	.871062	.871062

```
19 . logit approve black hispan gdlin, r

Iteration 0:   log pseudolikelihood = -733.64726
Iteration 1:   log pseudolikelihood = -567.51525
Iteration 2:   log pseudolikelihood = -499.65477
Iteration 3:   log pseudolikelihood = -496.58832
```

```
Iteration 4:   log pseudolikelihood = -496.58333
Iteration 5:   log pseudolikelihood = -496.58333

Logistic regression                             Number of obs     =      1,983
                                                Wald chi2(3)      =     351.73
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -496.58333               Pseudo R2         =     0.3231
```

approve	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
black	-.9585783	.232843	-4.12	0.000	-1.414942	-.5022144
hispan	-.9892	.3168287	-3.12	0.002	-1.610173	-.3682271
gdlin	3.820653	.2130302	17.93	0.000	3.403121	4.238184
_cons	-.9061003	.1959751	-4.62	0.000	-1.290204	-.5219963

20 .

LOANAPP_monica.do
Monica Elgawhy
4/18/2018
Assignment 4 Part II

$\beta_1 = P(Y=1|X=0)$

(2a.) The linear probability model is the name for the mult regression model of Part II when the dependent variable is binary rather than continuous. This is true in our case as the question describes only one variable as dependent to the rest — approve.

p. 425 in textbook * The population coefficient β_1 on a regressor X is the change in probability that Y=1 associated with a unit change in X.

Black: -.083 represents a probability of those of a black race have 8.3% less of a chance of being approved for a loan.
(Refer to p. 423)

Hispanic: -.081 represents a similar outcome for those of Hispanic origin have a more difficult time getting loan approval because of the color of their skin/nationality/race/etc.

Male: This is an interesting statistic. The coefficient .0081 tells us men have a .8% advantage in consideration for loan approval/credit worthiness. Note that the p-value of .597 shows a failure to reject a null of insignificance. The data with regards to men ~~may be disregarded~~ is statistically significant.

It is important to control for loan characteristics like a high enough income because in the case of rase as Ch. 11 opens, it may not be the color of one's skin that may be correlated. It may be that white groups had a headstart in simply having higher incomes to begin with or familial connections to speed up the process — not the very color of one's skin was the reason. It is the necessity to control for correlation vs. causation. It is necessary for validity to our regression hypotheses, biases and conclusions.

(b) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + u_i$

$\widehat{approve} = \beta_0 + \beta_1 \widehat{black} + \beta_2 \widehat{hispan} + \beta_3 \widehat{male} + \beta_4 \widehat{apr} + \beta_5 \widehat{term} + \beta_6 \widehat{bankruptcy} + \beta_7 \widehat{gdlin} + \beta_8 \text{hh\_expenditures}.$

$\widehat{approve} = .30 - .083\,\widehat{black} - .081\,\widehat{hispan} + .008\,\widehat{male} + 0(\widehat{apr}) + 0(\widehat{term}) - .045\,\widehat{bankruptcy} + .671\,\widehat{gdlin} - .0012\,\text{hh\_expenditures}$

approval rates range from 12.50% to 119% utilizing the sort, sum and predict commands in Stata. Predict utilizes the data in stata to give regression data points for the number corresponding to approve for each person's characteristic apr, term, status of bankruptcy and expenditures to give an estimate as to that person's chance at loan approval in relation to all others in the set.

*The mean predicted approval rate vs actual mean approval rate

found under sum approvrate, detail | found under sum approve, detail

mean .8775407 | .878467

*There are 15 observations unaccounted for in the predicted results. Apparently this occurred because 15 people did not vouch information of being male or otherwise. Those 15 were not accounted in predicted values as they violated the need for all the other variables of the regression to be held constant.

∇. This makes sense because the decision to not disclose this info means the functional form is incorrect potentially introducing bias into the prediction (LC#19 slide#4)

c) If the regressor, the loan approval rate, is correlated with a variable that has been omitted from the analysis (the binary variables for black, hispan or gdlin) and that determines in part the dependent variable then the OLS estimator will have omitted variable bias.

Omitted variable bias occurs ~~occurs~~ when two conditions are true:

(1) When the omitted variable is correlated with the included regressor and

(2) When the omitted variable is a determinant of the dependent variable.

Ch 6 p.182 If an omitted variable is in the present situation, u_i and X_i are correlated and the conditional mean of u_i given X_i is nonzero. This correlation violates the 1st least squares assumption (OLS) and makes our estimator biased and inconsistent.

↑F ↑ more likely var are sig

In reg approve black hispan gdlin, r the biggest difference is the change in F-statistic from 72.74 to 183.37. The F-test statistic's purpose is to test the null where all explanatory variables are found insignificant, or in math, equivalent to zero. ↑F, the more difficult it is to say the explanatory variables are insignificant because it's more so the case that there's more evidence to reject the insignificance of the regressors.

p. 263

$p \leq .05$	sig
$p \leq .01$	high sig
$p > .05$	not sig

The higher the F statistic $SSR_R > SSR_U$, we are more likely to reject the null hypothesis, in other words.
∴ Our results are better once dropping the other controls/regressors. This information tells us a good thing in that even if we include a regressor, if it's found to not impact the regression, the valid/significant regressors are not impacted much, if at all by the addition. Via p-values and the use of the F-statistic we can use our data without any concern for bias or inconsistency.

d) Utilizing the probit model, we expect the coefficients provided in regression form to be something we don't interpret. Utilizing probit, we say something is more or less likely to have an impact on loan approval rate compared to the utility of the coefficients in the LPM.
The probability of approval is found utilizing the predict and sum operations in STATA under use of probit.
In doing so, we type predict [new variable], pr. (for probit) In my log I wrote predict racialcomponent, pr and then sum racialcomponent if black==0 & gdlin==1 as we're given gdlin=1 as our restriction & repeat the sum command in the circumstance that black ==1 which again is the binary (dummy) variable we change.

Refer to LC #19 Slide 3/28

* P(gdlin=1 | hispan, black=0) = 94.45% were approved of non-black sample
P(gdlin=1 | hispan, black=1) = 87.33% were approved of black sample.
Difference ⇒ There is 7.12% less liklihood for loan approval for black applicants.

* Similar findings for Hispanic applicants:
P(gdlin=1 | black, hispan=0) = 94.26% were approved if not Hispanic
P(gdlin=1 | black, hispan=1) = 87.11% were approved from Hispanic sample.
∴ There is (94.26−87.11) 7.15% less liklihood for loan approval among Hispanic applicants.

LPM predicts 8.8% less for black applicants and 8.5% less for Hispanic applicants
This difference of ~1 to 2% when it could could've been a 30% difference, as we're looking at 8.8% and 7.12% in the spectrum relative to 0 → 100%, is negligible. (Referring to reg approve black hispan gdlin, r for LPM.)

Logit linear combin of independent variables

(e) $F(z) = \dfrac{\exp(z)}{1+\exp(z)} = L(z)$

$z \Rightarrow -\infty$ num $\to 0$ den $\to 1$ $F \to 0$

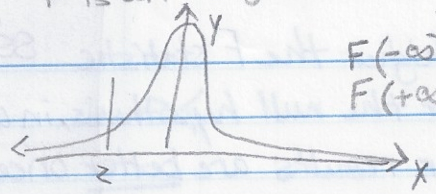
$z \to +\infty$ $F \to \dfrac{e^z}{e^z} = 1$

Logit when linear combin is zero

$= F(0) = \dfrac{e^0}{1+e^0} = \dfrac{1}{2}$

Probit

F is an integral: $F(z) = \int_{-\infty}^{z} \phi(u)\,du = \Phi(z)$



$F(-\infty) = 0$

$F(+\infty) = 1$

$\Phi(0) = \frac{1}{2}$ ✓

* helps with normal distribution assumptions

LC#20 slide #4 * $P(Y_i = 1 \mid X) = \dfrac{1}{1+e^{-(\beta_0+\beta_1 X)}}$ * $P(Y_i = 1 \mid X) = \Phi(\beta_0 + \beta_1 X)$

$\boxed{\text{Logit}}$

Black (0) $= \dfrac{1}{1+e^{-[-.906+(-.959)(0)+(-.9892)(0)+(3.821)(1)]}} = .9486$

Black (1) $= \dfrac{1}{1+e^{-[-.906+(-.959)(1)+(-.989)(0)+(3.821)(1)]}} = .876$

Our logit probabilities match our linear probability model calculated as expected!

for part (d) The mathematical, written answer to the $\boxed{\text{probit}}$ probabilities is

Black (0) $\Rightarrow \phi[-.578+(-.492)(0)+(-.5026)(0)+2.212(1)] = \phi(1.634)$

where ϕ is the cumulative standard normal distribution.

So it's $\Rightarrow .9484$

Black (1) $\Rightarrow \phi[-.578+(-.492)(1)+(-.5026)(0)+2.212(1)] = \phi(1.142)$

Where the probability according to Appendix Table 1 is .8729.

Once again the probabilities are all the same as they are expected.

Logit:

Hispanic(0) $\Rightarrow 1/1+e^{-[-.906+(-.959)(0)+(-.9892)(0)+(3.821)(1)]} = .94858$

Hispanic(1) $\Rightarrow 1/1+e^{-[-.906+(-.959)(0)+(-.9892)(1)+(3.821)(1)]} = .87278$

Probit:

Hispanic(0) $\Rightarrow \phi[-.578+(-.492)(0)+(-.5026)(0)+2.212(1)] = \phi[1.634] \Rightarrow .9484$

Hispanic(1) $\Rightarrow \phi[-.578+(-.492)(0)+(-.5026)(1)+2.212(1)] = \phi[1.1314] \Rightarrow .8708$

LC#20 Slide 4 (f) LPM, probit and logit all have the benefit of serving as a check to one another as they give accurately, consistent values to the probability of a regressor in relation to the dependent variable. LPM is easy to implement & interpret via coefficients. Yet if the functional form is incorrect we may introduce bias into the regression. Probit and Logit have straightforward calculations as long as we know the formula. Yet interpretation takes an additional step compared to LPM. Probit and Logit help account for our applied assumption of a normal distrubition (homoskedasticity) maybe more so than in LPM.

Assignment 4: Regression Models with Nonlinear Functions and Binary Dependent Variables

ECO 321

DUE: Thursday, April 19 in class

Instructions: You need to write a Stata do-file to answer these questions. Your do-file must produce a log file that contains your Stata output. You must submit your log file along with your answers to the assignment. Assignments that do not contain log files will be marked down.

1. In this question, we are interested in how education, experience, and race explain differences in wages across people. Use wage2.dta to answer this question.

a) Use the wage2.dta to estimate the following model where the return to education depends upon the amount of work experience (and visa versa):

$$log(wage_i) = \beta_0 + \beta_1 edu_i + \beta_2 exper_i + \beta_3 (educ_i \times exper_i) + u_i$$

Show that the marginal effect of another year of education equals $\hat{\beta}_1 + \hat{\beta}_3 exper$. Plug in the values of $\hat{\beta}_1$ and $\hat{\beta}_3$ to get the general form for the marginal effect of education.

b) Test whether the return to education depends on the level of experience.

c) Now allow education, experience, job tenure, marriage status, race, and geographic location to determine wages by estimating the following model:

$$log(wage_i) = \beta_0 + \beta_1 edu_i + \beta_2 exper_i + \beta_3 tenure_i + \beta_4 married_i + \beta_5 black_i + \beta_6 south_i + \beta_7 urban_i + u_i$$

Report the results in standard form. Holding other factors fixed, what is the approximate difference in monthly salary between married people and nonmarried people? Is this difference statistically significant at the 5% level?

d) Modify the model in part (c) by allowing $log(wage)$ to differ across four groups of people: married and black, married and nonblack, single and black, and single and nonblack. What is the estimated wage differential between married nonblack people and nonmarried nonblack people?

e) On a graph with education on the x-axis and $log(wage)$ on the y-axis, draw the sample regression functions for the four groups of people: (1) married and black, (2) married and nonblack, (3) single and black, and (4) single and nonblack, holding constant experience, tenure, and geographic location (i.e., south and urban). You do not need to create this graph in Stata, but you will use the

results from part (d) to draw the graph on paper.

2. In 1975, the U.S. passed the Home Mortgage Disclosure Act. This law requires all mortgage lenders to release annual data on all mortgage applications. One goal of the act is to identify discriminatory lending—cases where banks do not lend to individuals based on race or gender. The data set LOANAPP.dta contains a random sample of loan applications from HMDA data. The variables are defined as follows:

- ***approve*** **is a dummy variable equal to one if the loan application is approved**
- ***black*****,** ***hispan*****, and** ***male*** **are dummy variables equal to one if the borrower is black, hispanic, or male, respectively**
- ***apr*** **is the annual-percent rate on the loan (i.e. the interest rate)**
- ***term*** **measures the length of time over which the loan will be paid off**
- ***bankruptcy*** **is a dummy variable equal to one if the individual has ever declared bankruptcy**
- ***gdlin*** **is a variable equal to one if the borrower's credit history meets the lender's typical guidelines**
- ***hh_expenditures*** **is a measure of annual household expenditures as a share of annual income**

a) Estimate a linear probability model that relates loan approval rates to the applicant's demographic characteristics (black, hispan, male), the loan characteristics (apr, term), and the applicant's credit worthiness (bankruptcy, gdlin, hh_expenditures). Interpret the coefficients on black, hispan, and male. Although we are interested in the relationship between demographics and loan approval, why is it important to control for loan characteristics and credit worthiness?

b) Find the predicted values based on the regression in part (a). Summarize the distribution of the predicted values. How does the average predicted approval rate compare to the average actual approval rate? Do you notice anything unusual about the predicted values? Explain what you see.

c) Repeat the regression in part (a) using only the regressors that are individually statistically significant at the 10% level using t-tests. Do your results change substantially? Are your results sensitive to dropping these controls? What does might this tell you about OVB?

d) Estimate the relationship in part (c) using a probit model. What is the difference in the probability of approval between black and nonblack applicants who meet the borrowing guidelines ($gdlin = 1$)? How about the difference between hispanic and non-hispanic borrowers? Do the results differ much from the linear probability model?

e) Repeat part (d) using a logit model instead of probit model. Do the results change?

f) Based on your work in parts (a) through (e), comment on the disadvantages and advantages of using the linear probability model vs. probit/logit to estimate models with binary dependent variables.

g) on back.

g) Based on your work in parts (a) through (e), what do you conclude about discrimination by lenders?