Department of Economics | UN 3412
Columbia University | Fall 2021

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

1. [50 Pts] 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 in the data file. Use "describe" command in stata to see the variable definitions).

(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

```
. sum smoker

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      smoker |      10000       .2423    .4284963          0          1

. sum smoker if smkban==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      smoker |       6098    .2120367    .4087842          0          1

. sum smoker if smkban==0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      smoker |       3902    .2895951    .4536326          0          1
```

(i) probability = 24.2% overall

(ii) probability = 21.2% with smoking ban

(iii) probability = 29.0% without smoking ban

(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.

```
. reg smoker smkban, r;

Regression with robust standard errors                 Number of obs =   10000
                                                       F(  1,  9998) =   75.06
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0078
                                                       Root MSE      =  .42684

------------------------------------------------------------------------------
             |               Robust
      smoker |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smkban |  -.0775583    .008952    -8.66   0.000    -.0951061   -.0600106
       _cons |   .2895951   .0072619    39.88   0.000     .2753604    .3038298
------------------------------------------------------------------------------
```

The probability of smoking is 7.8 percentage points less if there is a smoking ban than if there is not. The *t*-statistic is -8.66 so the hypothesis that this difference is zero in population is rejected at the 1% significance level.

(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).

```
. reg smoker smkban $varlist1, r;

Regression with robust standard errors                 Number of obs =   10000
                                                       F( 10,  9989) =   68.75
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0570
                                                       Root MSE      =  .41631

------------------------------------------------------------------------------
             |               Robust
      smoker |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smkban |  -.0472399   .0089661    -5.27   0.000    -.0648153   -.0296645
      female |  -.0332569   .0085683    -3.88   0.000    -.0500525   -.0164612
         age |   .0096744   .0018954     5.10   0.000      .005959    .0133898
      age_sq |  -.0001318   .0000219    -6.02   0.000    -.0001747   -.0000889
      hsdrop |   .3227142   .0194885    16.56   0.000     .2845128    .3609156
      hsgrad |   .2327012   .0125903    18.48   0.000     .2080217    .2573807
     colsome |   .1642968   .0126248    13.01   0.000     .1395495     .189044
     colgrad |   .0447983   .0120438     3.72   0.000       .02119    .0684066
```

```
       black |  -.0275658   .0160785    -1.71   0.086    -.0590828    .0039513
    hispanic |  -.1048159   .0139748    -7.50   0.000    -.1322093   -.0774226
       _cons |  -.0141099   .0414228    -0.34   0.733    -.0953069    .0670872
------------------------------------------------------------------------------
```

After controlling for these additional variables, the estimated effect of a smoking ban is to reduce smoking by 4.7 percentage points, less than the 7.8 percentage points estimate without the control variables. This suggests that the original estimate was subject to omitted variable bias. For example, the estimated coefficients indicate that less educated individuals are more likely to smoke (condition (i) for omitted variable bias), but if less educated individuals also tend to work in places, like restaurants, that do not have smoking bans (condition (ii) for omitted variable bias), then having a smoking ban may be picking up the effect of education on smoking.

(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.

The *t*-statistic is -5.27 so the hypothesis is rejected 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).*

```
.  test hsdrop hsgrad colsome colgrad;

 ( 1)  hsdrop = 0
 ( 2)  hsgrad = 0
 ( 3)  colsome = 0
 ( 4)  colgrad = 0

       F(  4,  9989) =  140.09
            Prob > F =    0.0000
```

This requires an *F*-test because the null hypothesis is that the coefficients on *hsdrop, hsgrad, colsome,* and *colgrad* are all zero in population. The *p*-value is <.001 so the hypothesis is rejected at the 1% significance level. The less education, the larger are the coefficients, so they indicate that the probability of smoking is observed to decrease with education, holding the other regressors constant. For example, the probability of smoking is predicted to be 32.3 percentage points greater for a high school dropout than for the omitted group (those with a Master's degree or higher).

(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)			
Fraction Correctly Predicted (i.e., use $\widehat{Y}_i > 0.5$)			
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.

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

For this question, a correct answer would code Ms. B as having values of 0 for all education variables except for colgrad as the data suggests the education binary variables are mutually exclusive (no observation has a 1 for more than one variable). However, since it is ambiguous from the setup that this is how the education variables are coded, we will also accept answers that code Ms. B as having a 1 for hsgrad and a 1 for colgrad. We present the answers corresponding to both below.

Correct answer:

	Probit Model (1)	**Logit Model (2)**	**Linear Prob. Model (3)**
Estimated coefficient on *smkban* (standard error in parentheses)	-0.1586 (.0291)	-0.262 (0.049)	-.0472 (.0090)
Fraction Correctly Predicted (i.e., use $\widehat{Y}_i > 0.5$)	0.759	0.760	0.758
Predicted probabilities of smoking for Mr. A:			
(i) with workplace ban	.402	.408	.402
(ii) without workplace smoking ban	.464	.472	.449
Difference, (i) – (ii)	-.062	-0.064	-.047
Predicted probabilities of smoking for Ms. B:			
(iii) with workplace ban	.111	0.112	0.099
(iv) without workplace smoking ban	.144	0.141	0.146
Difference, (iii) – (iv)	-.033	-0.029	-0.047

Incorrect but acceptable answers for Ms. B:

	Probit Model (1)	**Logit Model (2)**	**Linear Prob. Model (3)**
Estimated coefficient on *smkban* (standard error in parentheses)	-.1586 (.0291)	-0.262 (0.049)	-.0472 (.0090)
Predicted probabilities of smoking for Ms. B:			
(iii) with workplace ban	.0.367	0.379	0.331
(iv) without workplace smoking ban	0.428	0.442	0.379
Difference, (iii) – (iv)	-0.061	-0.063	-0.047

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 (2)), 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. [25 Pts] 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?

Solution:

(a) Since these are logit estimates, the value of the coefficients cannot be interpreted easily. However, statements can be made about the direction of the relationship between the dependent variable and the regressors. There is a decrease in the probability of females of living on their own with an increase in years of education. Living on a farm also lowers the probability. These results hold both for whites and nonwhites. In addition, for whites the probability of living on her own increases up to a point with age (i.e., age < 60), but then decreases after age ≥ 60. This is the result of age entering as a level and the square of age. This relationship with regard to age is not statistically significant for nonwhites. In the south, white females are more likely to live on their own, but nonwhites are not. An increase in expected family earnings and family composition increase the probability of females living on their own.

(b) For whites, the probability is 0.90 (i.e., F(Z=2.2292)), while for nonwhites, it is 0.88 (i.e., F(Z=2.0349)). In the above approach, all coefficients are allowed to vary, whereas in a combined sample, the coefficients on the variables other than the binary race variable would have to be identical.

(c) The probability would increase to 0.81 (i.e., F(Z=1.4804)).

3. [25 Pts] Maximum Likelihood Estimation Method (MLE):

(a) **[12 Pts]** 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., $\hat{\lambda}$).

Solution

Because of independence, the likelihood function is a product of the individual pdf's:

$$L \prod_{i=1}^{n} f(x_i, \lambda) = \left(\lambda e^{-\lambda x_1}\right) \ldots \left(\lambda e^{-\lambda x_n}\right) = \lambda^n e^{-\lambda \sum x_i}$$

First take the natural logarithm of the likelihood function, which is:

$$\ln[f(x_1, x_2, \dots, x_n; \lambda)] = n[ln(\lambda)] - \lambda \sum x_i$$

Second, take the derivative of this log-likelihood function with respect to λ and set it equal to zero to solve for λ:

$$\frac{dlnf(.)}{d\lambda} = n\left(\frac{1}{\lambda}\right) - \sum x_i = 0$$

$$\lambda = {}^{n}/_{\sum x_i} = {}^{1}/_{\bar{X}}$$

(b) **[13 Pts]** Suppose $X_1, X_2, \dots, X_n$ 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., $\hat{\alpha}$).

Solution

Write the likelihood function assuming independence:

$$L\left(\prod_{i=1}^{n} f(x_i, \alpha)\right) = L\prod_{i=1}^{n} f(x_1, x_2, \dots, x_n; \alpha) = \prod_{i=1}^{n} \frac{e^{-\alpha}\alpha^{x_i}}{x_i!}$$

Re-writing the pdf function as follows:

$$L\prod_{i=1}^{n} f(x_1, x_2, \dots, x_n; \alpha) = \prod_{i=1}^{n} \frac{e^{-\alpha}\alpha^{x_i}}{x_i!} = \frac{e^{-n\alpha}\alpha^{\sum x_i}}{x_1!\, x_2! \dots x_n!}$$

To maximize the log-likelihood function and solve for the MLE $\hat{\alpha}$, first take log of the above function and set it equal to zero. Then solve for $\hat{\alpha}$ as follows.

$$lnL = -n\alpha + \ln(\alpha) . \sum x_i - \ln(x_1!\, x_2! \dots x_n!)$$

$$\frac{dlnL}{d\alpha} = -n + \frac{\sum x_i}{\alpha} = 0$$

$$\hat{\alpha} = \frac{\sum x_i}{n} = \bar{X}$$

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

1. SW Exercise 11.1

(a) The t-statistic for the coefficient on *Experience* is 0.031/0.009 = 3.44, which is significant at the 1% level.

(b) $z_{Matthew} = 0.712 + 0.031 \times 10 = 1.022; \Phi(1.022) = 0.847$

(c) $z_{Christopher} = 0.712 + 0.031 \times 0 = 0.712; \Phi(0.712) = 0.762$

(d) $z_{Jed} = 0.712 + 0.031 \times 80 = 3.192; \Phi(3.192) = 0.999$, this is unlikely to be accurate because the sample did not include anyone with more than 40 years of driving experience.

2. SW Exercise 11.6

(a) For a black applicant having a P/I ratio of 0.35, the probability that the application will be denied is $\Phi(-2.26 + 2.74 \times 0.35 + 0.71) = \Phi(-0.59) = 27.76\%$.

(b) With the P/I ratio reduced to 0.30, the probability of being denied is $\Phi(-2.26 + 2.74 \times 0.30 + 0.71) = \Phi(-0.73) = 23.27\%$. The difference in denial probabilities compared to (a) is 4.4 percentage points lower.

(c) For a white applicant having a P/I ratio of 0.35, the probability that the application will be denied is $\Phi(-2.26 + 2.74 \times 0.35) = 9.7\%$. If the P/I ratio is reduced to 0.30, the probability of being denied is $\Phi(-2.26 + 2.74 \times 0.30) = 7.5\%$. The difference in denial probabilities is 2.2 percentage points lower.

(d) From the results in parts (a)–(c), we can see that the marginal effect of the P/I ratio on the probability of mortgage denial depends on race. In the probit regression functional form, the marginal effect depends on the level of probability which in turn depends on the race of the applicant. The coefficient on *black* is statistically significant at the 1% level.

3. SW Exercise 11.7

(a) For a black applicant having a P/I ratio of 0.35, the probability that the application will be denied is

$$F(-4.13+5.37\times 0.35+1.27)=\frac{1}{1+e^{0.9805}}=27.28\%.$$

(b) With the P/I ratio reduced to 0.30, the probability of being denied is $F(-4.13+5.37\times 0.30+1.27)=\frac{1}{1+e^{1.249}}=22.29\%$. The difference in denial probabilities compared to (a) is 4.99 percentage points lower.

(c) For a white applicant having a P/I ratio of 0.35, the probability that the application will be denied is $F(-4.13+5.37\times 0.35)=\frac{1}{1+e^{2.2505}}=9.53\%$. If the P/I ratio is reduced to 0.30, the probability of being denied is $F(-4.13+5.37\times 0.30)=\frac{1}{1+e^{2.519}}=7.45\%$. The difference in denial probabilities is 2.08 percentage points lower.

(d) From the results in parts (a)–(c), we can see that the marginal effect of the P/I ratio on the probability of mortgage denial depends on race. In the logit regression functional form, the marginal effect depends on the level of probability which in turn depends on the race of the applicant. The coefficient on *black* is statistically significant at the 1% level. The logit and probit results are similar.

4. SW Exercise 11.9

(a) The coefficient on *black* is 0.084, indicating an estimated denial probability that is 8.4 percentage points higher for the black applicant.

(b) The 95% confidence interval is $0.084 \pm 1.96 \times 0.023 = [3.89\%, 12.91\%]$.

(c) The answer in (a) will be biased if there are omitted variables which are race-related and have impacts on mortgage denial. Such variables would have to be related with race and also be related with the probability of default on the mortgage (which in turn would lead to denial of the mortgage application). Standard measures of default probability (past credit history and employment variables) are included in the regressions shown in Table 9.2, so these omitted variables are unlikely to bias the answer in (a). Other variables such as education, marital status, and occupation may also be related the probability of default, and these variables are omitted from the regression in column. Adding these variables (see columns (4)–(6)) have little effect on the estimated effect of *black* on the probability of mortgage denial.

5. SW Empirical Exercise 11.3

Answers are provided to many of the questions using the linear probability models. You can also answer these questions using a probit or logit model. Answers are based on the following table:

Regressor	Dependent Variable						
	Insured	**Insured**	**Insured**	**Healthy**	**Healthy**	**Healthy**	**Any Limitation**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
selfemp	–0.128**	–0.174**	–0.210**	0.010	0.020*	0.015	–0.010
	(0.015)	(0.014)	(0.063)	(0.007)	(0.008)	(0.008)	(0.012)
age		0.010**	0.010**		0.0006	–0.002	0.003
		(0.003)	(0.003)		(0.0017)	(0.002)	(0.002)
age^2		–0.00008*	0.000*		–0.00003	0.000	0.000
		(0.00003)	(0.000)		(0.00002)	(0.000)	(0.000)
age × selfemp			0.001				
			(0.001)				
deg_ged		0.151**	0.151**			0.045*	0.061*
		(0.027)	(0.027)			(0.020)	(0.024)
deg_hs		0.254**	0.254**			0.099**	–0.012
		(0.016)	(0.016)			(0.012)	(0.012)
deg_ba		0.316**	0.316**			0.122**	–0.042**
		(0.017)	(0.017)			(0.013)	(0.014)
deg_ma		0.335**	0.335 **			0.128**	–0.078**
		(0.018)	(0.018)			(0.015)	(0.018)
deg_phd		0.366**	0.366 **			0.138**	–0.084**
		(0.026)	(0.025)			(0.018)	(0.027)
deg_oth		0.288**	0.287**			0.115**	–0.049**
		(0.020)	(0.020)			(0.014)	(0.017)
familysz		–0.017**	–0.017**			–0.001	–0.016**
		(0.003)	(0.003)			(0.002)	(0.002)
race_bl		–0.028*	–0.028*			–0.022*	–0.035**
		(0.013)	(0.013)			(0.009)	(0.010)
race_ot		–0.048*	–0.048**			–0.029	–0.046
		(0.023)	(0.023)			(0.015)	(0.016)
reg_ne		0.037**	0.037**			0.006	–0.046**
		(0.012)	(0.012)			(0.008)	(0.011)
reg_mw		0.053**	0.053**			0.012	0.008
		(0.012)	(0.012)			(0.008)	(0.011)
reg_so		0.003	0.004			0.001	–0.007
		(0.011)	(0.011)			(0.008)	(0.010)
male		–0.037**	–0.037**			0.015**	–0.005
		(0.008)	(0.008)			(0.005)	(0.007)
married		0.136**	0.136 **			0.001	–0.017**
		(0.010)	(0.010)			(0.007)	(0.009)
Intercept	0.817	0.299**	0.296**	0.927**	0.953**	0.902**	0.071
	(0.004)	(0.054)	(0.054)	(0.003)	(0.031)	(0.035)	(0.044)

Significant at the 5% * or 1% ** level.

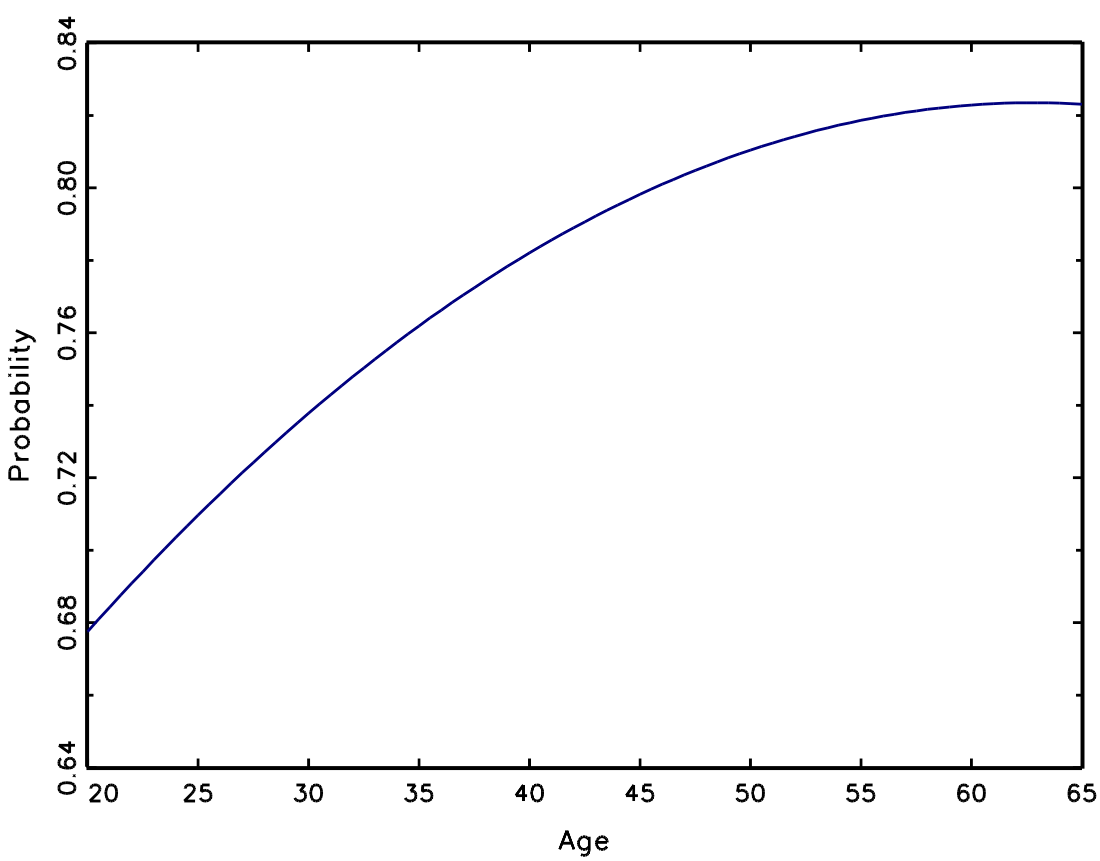
(a) Probability of being insured

	$\hat{p}$	$SE(\hat{p})$
All Workers	0.802	0.004
Self Employed	0.689	0.04
Not Self Employed	0.817	0.004

The self-employed are 12.8% less likely to have health insurance. This is a large number. It is statistically significant: from (1) in the table the difference is significant at the 1% level.

(b) From specification (2), the result is robust to adding additional control variables. Indeed, after controlling for other factors, the difference *increases* to 17.4%

(c) See specification (2). There is evidence of nonlinearity (Age^2 is significant in the regression). The plot below shows the effect of *Age* on the probability of being insured for a self-employed white married male with a BA and a family size of four from the northeast. (The profile for others will look the same, although it will be shifted up or down.) The probability if being insured increases with *Age* over the range 20–65 years.



(d) Specification (3) adds an interaction of *Age* and *selfemp*. Its coefficient is not statistically significant, and this suggests that the effect of *selfemp* does not depend on *Age*. (Note: this answer is specific to the linear probability model. In the probit model, even without an interaction, the effect of *selfemp* depends on the level of the probability of being insured, and this probability depends on *Age*.)

(e) This is investigated in specifications (4)–(7). The effect of *selfemp* on health status or "Any Limitation" is small and not and statistically significant. This result obtains when the regression controls for *Age* or for a full set of control variables.

There are potential problems with including *healthy* on the right hand side of the model because of "adverse selection" problems. It is possible that only those less healthy individuals pursue health insurance, perhaps through their employer. This causes a self-selection problem that more healthy individuals might (a) choose to be self-employed or (b) choose not to obtain health insurance. While the evidence suggests that there might not be a strong correlation between health status and self-employment, the adverse selection concerns still exist.

STATA .do file for question 12:

```
use insurance.dta, clear
gen age_sqr=age^2
gen age_selfemp=age*selfemp

reg insured selfemp, r
reg insured selfemp age age_sqr deg_ged deg_hs deg_ba deg_ma deg_phd deg_oth
familysz race_bl race_ot reg_ne reg_mw reg_so male married, r

gen age_sim=19+_n if _n<=36   // for simulated plot
gen
y_sim=_b[_cons]+_b[selfemp]+age_sim*_b[age]+(age_sim^2)*_b[age_sqr]+_b[deg_ba
]+4*_b[familysz]+_b[reg_ne]+_b[male]+_b[married]
twoway connected y_sim age_sim, xtitle("Age") ytitle("Probability") msize(0)

reg insured selfemp age age_sqr age_selfemp deg_ged deg_hs deg_ba deg_ma
deg_phd deg_oth familysz race_bl race_ot reg_ne reg_mw reg_so male married, r
reg healthy selfemp, r
reg healthy selfemp age age_sqr, r
reg healthy selfemp age age_sqr deg_ged deg_hs deg_ba deg_ma deg_phd deg_oth
familysz race_bl race_ot reg_ne reg_mw reg_so male married, r
reg anylim selfemp age age_sqr deg_ged deg_hs deg_ba deg_ma deg_phd deg_oth
familysz race_bl race_ot reg_ne reg_mw reg_so male married, r
```