IIA-3 Econometrics: Supervision 7

Emre Usenmez

Lent Term 2025

Topics Covered

Faculty Qs:

Supplementary Qs: Independently pooled cross-section; panel data; difference-in-differences (DiD) estimator;

Related Reading:

Dougherty (2016), *Introduction to Econometrics*, 5^{th} ed, OUP

Chapter 10: Binary Choice and Limited Dependent Variable Models, and Maximum Likelihood Estimation

Chapter 14: Introduction to Panel Data Models

Wooldridge J M (2021) *Introductory Econometrics: A Modern Approach*, 7^{th} ed,

Section 7-5: A Binary Dependent Variable: The Linear Probability Model

Chapter 13: Pooling Cross Sections across Time: Simple Panel Data Methods

Chapter 17: Limited Dependent Variable Model and Sample Selection Corrections

Gujarati, D N and Porter, D (2009) *Basic Econometrics*, 7^{th} International ed, McGraw-Hill

Chapter 15: Qualitative Response Regression Models

Chapter 16: Panel Data Regression Models

Gujarati, D (2022) *Essentials of Econometrics*, 5^{th} ed, Sage

Chapter 6: Qualitative or Dummy Variable Regression Models

Chapter 12: Panel Data Regression Models

Stock, J H and Watson M W (2020) *Introduction to Econometrics.* 4^{th} Global ed, Pearson

Chapter 10: Regression with Panel Data

Chapter 11: Regression with a Binary Dependent Variable

Very grateful to Dr Oleg Kitov and Dr Clive Lawson for the very informative stylized answers to previous iterations of the supervision questions.

FACULTY QUESTIONS

QUESTION A:

SUPPLEMENTARY QUESTIONS

QUESTION A

(1) Explain why the following is termed a Linear Probability Model (LPM) if Y_i is a binary dependent variable (i.e. Y_i takes only the values 0 and 1):

$$Y_i = \beta_0 + \beta_1 X_i + v_i \tag{1}$$

Answer: When the dichotomous dummy variable is the dependent variable are of the form

$$Y_i = \begin{cases} 1 & \text{if some condition is satisfied} \\ 0 & \text{if some condition is not satisfied} \end{cases}$$

then $\mathbb{E}(Y_i \mid X_i)$ can be interpreted as the *conditional probability* that the event will occur given X_i, i.e. $\mathbb{P}(Y_i = 1 \mid X_i)$.

Assume $\mathbb{E}(u_i) = 0$ in order to have unbiased estimators. Then the probability of the event occuring, p_i, is assumed to be a linear function of a set of explanatory variables:

$$\begin{aligned} p_i &= 1 \times \mathbb{P}(Y = 1 \mid X_i) + 0 \times \mathbb{P}(Y_i = 0 \mid X_i) \\ &= \mathbb{P}(Y_i = 1 \mid X_i) \\ &= \mathbb{E}(Y_i \mid X_i) \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

Notice that

$$Y_i = \begin{cases} 1 & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{cases}$$

then Y_i follows the *Bernoulli probability distribution.* That is $Y_i \sim \text{Bern}(p_i)$ where $p_i = \beta_0 + \beta_1 X_i$. Accordingly,

$$\mathbb{E}(Y_i) = 0(1 - p_i) + 1p_i = p_i.$$

We can then equate

$$\begin{aligned} \mathbb{E}(Y_i \mid X_i) &= \beta + 0 + \beta_1 X_i \\ &= \mathbb{E}(Y_i) \\ &= p_i. \end{aligned}$$

This means, the conditional expectation of the LPM model can be interpreted as the conditional probability of Y_i.

In general, the expectation of Bernoulli random variable is the probability that the random variable equals 1.

Also note that if there are n independent trials, each with a probability p of success and probability $(1 - p)$ of failure, and X of these trials represent the number of successes, then X follows the *binomial distribution.* The mean of the binomial distribution is np and its variance is $np(1 - p)$.

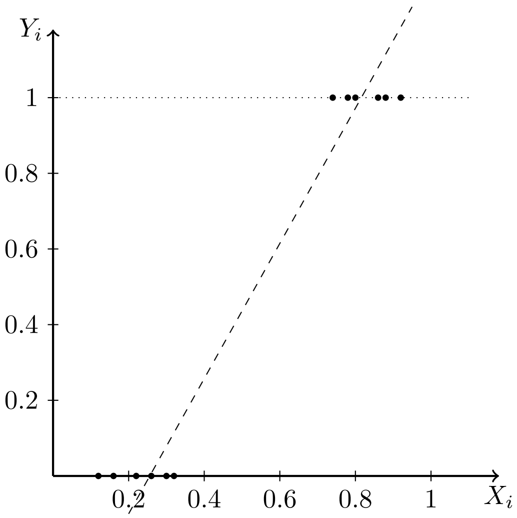
Finally, since p_i must be between 0 and 1, then we have the restriction that $0 \leq \mathbb{E}(Y_i \mid X_i) \leq 1$.

(2) Carefully outline the main problems associated with LPMs.

Answer: i) Nonfulfillment of $0 \leq \mathbb{E}(Y_i \mid X_i) \leq 1$

[-] Although Y takes a value of 0 or 1, there is no guarantee that the estimated values of Y will necessarily lie between 0 and 1. In an application, some $\hat{Y}_i$ values can turn out to be negative and some can exceed 1.

The problem emerges from the fact that OLS will fit a straight line to these points for the estimated values of β_0 nad β_1 while nothing preventing the intercept from being negative. Similarly, for high levels of X we can obtain probability higher than 1. This is a problem since a negative probability or probability higher than 1 is meaningless.



ii) Errors are non-normal and follow Bernoulli distribution or binomial probability distribution.

[-] Although OLS does not require u_i to be normally distributed to get unbiased estimates, we assume them to be the case for the purposes of statistical inference.

[-] Since v_i takes only the following two values

$$Y_i = \begin{cases} 1 & v_i = 1 - \beta_0 - \beta_1 X_{1i} \quad \text{with prob. } p_i \\ 0 & v_i = -\beta_0 - \beta_1 X_{1i} \quad \text{with prob. } 1 - p_i \end{cases}$$

it is non-normal. It instead follows a Bernoulli distribution:

$$\begin{aligned} f(Y_i) &= \begin{cases} p_i & Y_i = 1 \\ 1 - p_i & Y_i = 0 \end{cases} \\ &\equiv p^{Y_i}(1-p)^{1-Y_i} \end{aligned}$$

iii) Error term has heteroskedastic variances

[-] For Bernoulli distribution the theoretical mean is p and variance $p(1-p)$. This means the variance is a function of the mean, hence the error variance is heteroskedastic.

$$\begin{aligned} Var(v_i) &= \mathbb{E}(v_i^2) - \left(\mathbb{E}(v_i)^2\right) \\ &= \mathbb{E}(v_i^2) - 0 \\ &= \mathbb{E}((Y_i \mid X_{1i})^2) \\ &= \mathbb{P}(Y_i = 1 \mid X_{1i})(\text{value of } v_i \text{ when } Y_i = 1)^2 + \mathbb{P}(Y_i = 0 \mid X_{1i})(\text{value of } v_i \text{ when } Y_i = 0)^2 \\ &= \mathbb{P}(Y_i = 1 \mid X_{1i})(1 - (\beta_0 + \beta_1 X_{1i}))^2 + \mathbb{P}(Y_i = 0 \mid X_{1i})(0 - (\beta_0 + \beta_1 X_{1i}))^2 \\ &= p_i(1-p_i)^2 + (1-p_i)(0-p_i)^2 \qquad \text{since } \mathbb{E}(Y_i \mid X_{1i}) = p_i = \beta_0 + \beta_1 X_{1i} \end{aligned}$$

$$
\begin{aligned}
&= p_i(1 + p_i^2 - 2p_i) + (1 - p_i)p_i^2 \\
&= p_i + p_i^3 - 2p_i^2 + p_i^2 - p_i^3 \\
&= p_i - p_i^2 \\
&= p_i(1 - p_i)
\end{aligned}
$$

since p_i differs for each i, and since $Var(v_i)$ depends on p_i, the disturbance is heteroskedastic. Since $p_i = \mathbb{E}(Y_i) = \beta_0 + \beta_1 X_{1i}$, this can also be expressed as:

$$Var(v_i) = (\beta_0 + \beta_1 X_{1i})(1 - \beta_0 - \beta_1 X_{1i})$$

which varies with X_i.

iv) R^2 is not meaningful

since Y takes only two values, 0 and 1, the conventionally computed R^2 value is likely to be much lower than 1

iv) It is not logically attractive model since the marginal effects are constant

It assumes that $p_i = \mathbb{E}(Y = 1 \mid X)$ increases linearly with X. That is, the marginal effect of X remains constant throughout. This is unrealistic. In reality, we would expect that p_i is nonlinearly related to X.

---

(3) If you have not done so already, derive $Var(v)$ and use this to find a transformation that can deal with the problem of heteroskedasticity in equation (1) on page 3.

Answer: We have already derived the variance of the disturbance term.

Recall that in the presence of heteroskedasticity the OLS estimators are unbiased but inefficient. In Supervision 4, we discussed a number of ways to handle heteroskedasticity problem. Since the variance of v_i depends on X_i, one way to resolve the heteroskedasticity problem is to transform the model (1) as follows:

$$\frac{Y_i}{\sqrt{w_i}} = \beta_0 \frac{1}{\sqrt{w_i}} + \beta_1 \frac{X_i}{\sqrt{w_i}} + \frac{v_i}{\sqrt{w_i}}$$

where $\sqrt{w_i} = var(v_i) = \sqrt{\big(\mathbb{E}(Y_i \mid X_i)\big)\big(1 - \mathbb{E}(Y_i \mid X_i)\big)} = \sqrt{p_i(1 - p_i)}$. With this transformation, the transformed error term is now homoskedastic.

To see this, set $\sqrt{w_i} = \sigma_i$. then:

$$
\begin{aligned}
Var(\frac{v_i}{\sigma_i}) &= \mathbb{E}\bigg[\Big(\frac{v_i}{\sigma_i}\Big)^2\bigg] - \bigg[\mathbb{E}\Big(\frac{v_i}{\sigma_i}\Big)\bigg]^2 \\
&= \mathbb{E}\bigg[\Big(\frac{v_i}{\sigma_i}\Big)^2\bigg] \quad \text{since } \mathbb{E}\Big(\frac{v_i}{\sigma_i}\Big) = 0 \\
&= \frac{1}{\sigma_i^2}\mathbb{E}(v_i^2) \quad \text{since } \sigma_i^2 \text{ is known; thus it is a collection of constants} \\
&= \frac{1}{\sigma_i^2}\sigma_i^2 = 1
\end{aligned}
$$

which is a constant.

In practice, the true $\mathbb{E}(Y_i \mid X_i)$ is unknown, so the weights w_i of this weighted least squared regression are also unknown. To estimate the w_i, we can use the following two step-procedure:

Step 1: Run the OLS regression equation (1) despite the homoskedasticity problem and obtain $\hat{Y}_i$, which is the estimate of the true $\mathbb{E}(Y_i \mid X_i)$

Step 2: Obtain $\hat{w}_i = \hat{Y}_i(1 - \hat{Y}_i)$, which is the estimate of w_i.

Step 3: Use the estimated w_i to transform the data as above and estimate the transformed equation by OLS (i.e., by weighted least squares).

Also note that if the sample is reasonably large, we can use White's heteroskedasticity-corrected standard errors to deal with heteroskedasticity.

(4) In considering the utility from taking a paid job, a married woman considers only the effect of the wage that could be earned (X_i). Let U_i denote the utility difference between working and not working, then assuming a linear relationship we can write:

$$U_i = \alpha + \beta X_i + \varepsilon_i$$

where ε_i denotes unobserved characteristics associated with individual i, and is assumed to be a random variable that is independent and identically distributed with probability density function $f(\varepsilon_i)$.

Assuming that we observe an indicator variable Y_i, which takes the value 1 if the individual works and the utility difference exceeds zero, write down an expression for the probability that a given woman works. Assuming that the probability of working is independent across women, derive an expression for the likelihood function based on a sample of size n.

Answer: In this question we are told, using the indicator function notation, that:

$$Y_i = \mathbb{1}[U_i > 0].$$

We can estimate the probability that a given woman works in two ways:

(i) Bernoulli: The dichotomous dependent variable can be expressed as a Bernoulli distribution:

$$\begin{aligned} f(Y_i) &= \begin{cases} p_i & Y_i = 1 \\ 1 - p_i & Y_i = 0 \end{cases} \\ &\equiv p^{Y_i}(1-p)^{1-Y_i} \end{aligned}$$

Therefore, the probability of some particular sample of size n is

$$\prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} \qquad \text{for } i = 1, \dots, n \text{ and } Y_i = 0, 1$$

(ii) Likelihood Function: The likelihood function gives the joint probability density given the sample of observations. Since $Y_i \sim$ i.i.d. Bern(p), for a given value of p, the probability mass functon of Y_i is:

$$f(Y_i \; ; \; p) = p^{Y_i}(1-p)^{Y_i}.$$

The likelihood function $L(p \mid \tilde{\mathbf{Y}})$ is then given by the joint probability of observing $\tilde{\mathbf{Y}} = (Y_1, \dots, Y_n)$ denoted by $f(\tilde{\mathbf{Y}} \ ; \ p)$:

$$L(p \ ; \ \tilde{\mathbf{Y}}) = f(\tilde{\mathbf{Y}} \ ; \ p) = \prod_{i=1}^{n} f(Y_i \ ; \ p) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i}.$$

So the probability that a given woman works is expressed as:

$$L(p) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i}$$

if we take the natural log of both sides,

$$\ell(p) = \ln p \sum_{i=1}^{n} Y_i + \ln(1-p) \sum_{i=1}^{n}(1-Y_i).$$

In order to estimate the unknown parameter in such a manner that the probability of observing the given Y's is as high as possible, we apply the *maximum likelihood method*:

$$\begin{aligned}\frac{\partial \ell(p)}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} Y_i - \frac{1}{1-p}\sum_{i=1}^{n}(1-Y_i) &\overset{\text{set}}{=} 0 \\ \sum_{i=1}^{n} Y_i - p\sum_{i=1}^{n} Y_i &= p\sum_{i=1}^{n}(1-Y_i) \\ p &= \frac{1}{n}\sum_{i=1}^{n} Y_i.\end{aligned}$$

So, if say we have a sample of 17 and 2 of them wants extra work, the probability that a given woman works is $p = 2/17$. This is also the frequency of the sample and thus equivalent to the estimate of the parameter via method of moments.

However, notice that ε_i is not directly observable. This is known as *unobservable*, or **latent**, variable. In this question, U_i is the *latent variable* where

$$Y_i = \mathbb{1}[U_i > 0].$$

We can derive the response probability for Y as follows assuming ε is symmetrically distributed about zero:

$$\begin{aligned}\mathbb{P}(Y_i = 1 \mid X_i) &= \mathbb{P}(U_i > 0 \mid X_i) \\ &= \mathbb{P}(\varepsilon_i > -(\alpha + \beta X_i) \mid X_i) \\ &= 1 - F(-(\alpha + \beta X_i)) \\ &= F(\alpha + \beta X_i).\end{aligned}$$

Note that our assumption ε is symmetrically distributed about zero means that $1 - F(-s) = F(s)$ for all real numbers s.

In the binomial model, in order to estimate the nonlinear binary response models we maximized with respect to p. Here, and in general where there are explanatory variables, we can use maximum likelihood estimation to estimate nonlinear models where we maximize with respect to α and β.

Let $f(Y \mid X, \beta)$ denote the density function for a random draw Y_i from the population, conditional on $X_i = x$. The maximum likelihood estimator (MLE) of β that maximizes the log-likelihood function:

$$\underset{b}{max} \sum_{i=1}^{n} \ln f(Y_i \mid X_i, b)$$

where b is the dummy argument in the maximization problem.

In most cases $\hat{\beta}$, i.e. the MLE, is consistent and has an approximate normal distribution in large samples. This is true even though we cannot write down a formula for $\hat{\beta}$, except in very special circumstances.

For the binary response case, the conditional density is determined by two values:

$$f(1 \mid X, \beta) = \mathbb{P}(Y_i = 1 \mid X_i) = F(\beta X_i)$$
$$\text{and}$$
$$f(0 \mid X, \beta) = \mathbb{P}(Y_i = 0 \mid X_i) = 1 - F(\beta X_i)$$

This density can be written succinctly as

$$f(Y \mid X, \beta) = [F(\beta X)]^Y [1 - F(\beta X)]^{1-Y} \quad \text{for } Y = 0, 1$$

where we get $[F(\beta X)]^Y$ when $Y = 1$ and $[1 - F(\beta X)]^{1-Y}$ when $Y = 0$.

The *log-likelihood function* for observation i is a function of the parameters and the data (X_i, Y_i) and is obtained by:

$$\ell_i(\beta) = Y_i \ln[F(\beta X_i)] + (1 - Y_i) \ln[1 - F(\beta X_i)].$$

The log-likelihood for a sample size n is then obtained by summing this up across all observations:

$$L(\beta) = \sum_{i=1}^{n} \ell_i(\beta).$$

The MLE of β, denoted $\hat{\beta}$, maximizes this log-likelihood.

- If $F(\cdot)$ is standard normal cdf, then $\hat{\beta}$ is the *probit estimator*,
- If $F(\cdot)$ is standard logit cdf, then $\hat{\beta}$ is the *logit estimator*

So the maximization equation becomes:

$$\underset{\beta}{max} \sum_{i=1}^{n} \ln f(Y_i \mid X_i, \beta)$$
$$\underset{\beta}{max} \sum_{i=1}^{n} \ln \Big([F(\beta X)]^Y [1 - F(\beta X)]^{1-Y} \Big)$$
$$\underset{\beta}{max} \sum_{i=1}^{n} \Big(Y_i \ln F(\beta X_i) + (1 - Y_i) \ln[1 - F(\beta X_i)] \Big).$$

QUESTION B

(1) Using the 'labour force' data from the data set (`limdep.xls`) estimate the following LPM:

$$inlf_i = \alpha + \beta\ educ_i + \gamma\ kids_i + \varepsilon_i \tag{2}$$

Answer: The variables are:

[id:] identification number

[inlf:] =1 if in labor force, 1975

[Kids:] number of kids less than 6 years old

[educ:] years of schooling

In R:

```
laborforce_df <- read_excel("../Data/limdep.xls", sheet = "labour force")
SQB1_lm <- lm(inlf ~ Kids + educ, data = laborforce_df)
summary(SQB1_lm)
```

In STATA:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("labour force") firstrow
reg inlf Kids educ
```

```
      Source |       SS           df       MS      Number of obs   =       753
-------------+----------------------------------   F(2, 750)       =     37.34
       Model |  16.7298847         2  8.36494235   Prob > F        =    0.0000
    Residual |  167.997871       750  .223997161   R-squared       =    0.0906
-------------+----------------------------------   Adj R-squared   =    0.0881
       Total |  184.727756       752  .245648611   Root MSE        =    .47328

------------------------------------------------------------------------------
        inlf | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        Kids |  -.2241021   .0331357    -6.76   0.000    -.2891518   -.1590524
        educ |   .0463196    .007614     6.08   0.000     .0313724    .0612668
       _cons |   .0525438   .0946111     0.56   0.579    -.1331903    .2382779
------------------------------------------------------------------------------
```

which gives us:

$$\begin{array}{rlll} \widehat{inlf} = & 0.053 & +0.046\ educ & -0.224\ kids \\ t: & [0.56] & [6.08] & [-6.76] \\ se: & (0.095) & (0.008) & (0.033) \end{array}$$

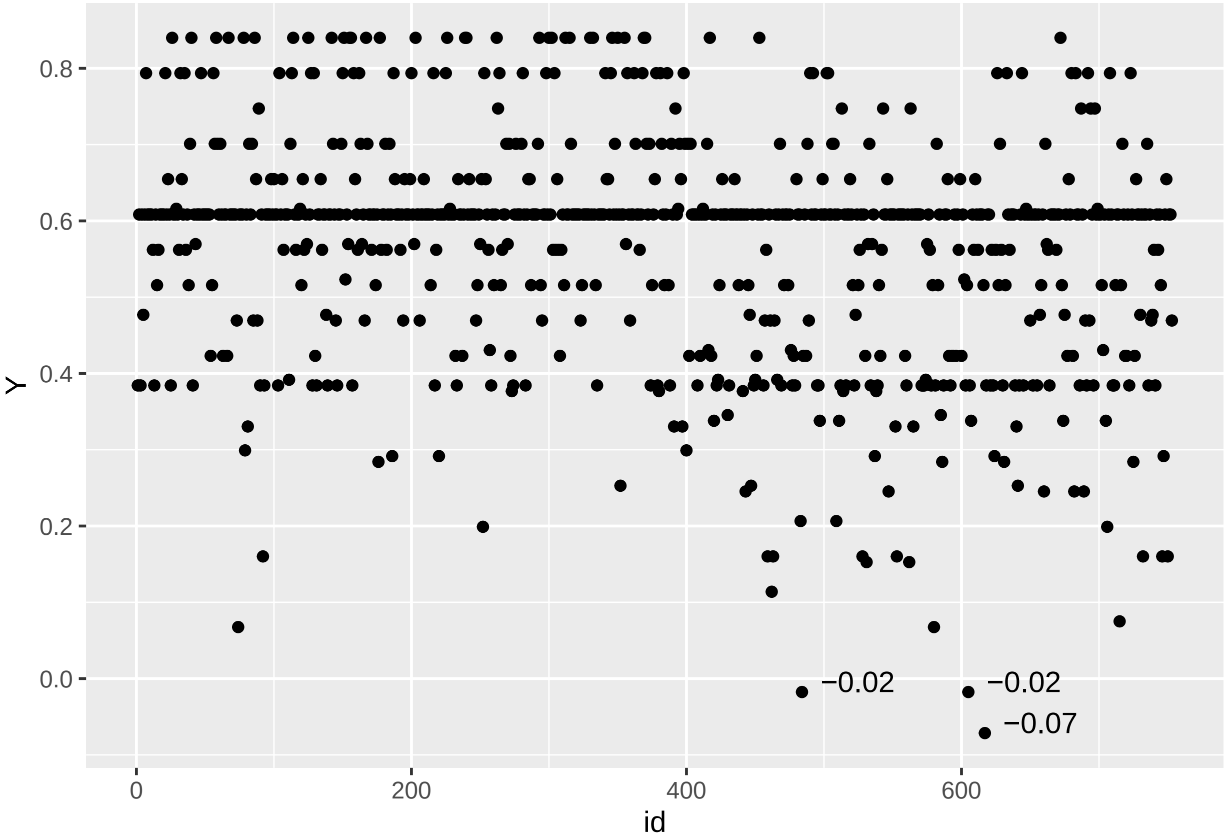
(2) Plot the fitted values and comment on the plausibility of your results.

In STATA:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("labour force") firstrow
quietly reg inlf Kids educ
predict Y
scatter Y id
```

In R:

```
Y <- predict(SQB1_lm)
ggplot(data = laborforce_df,
       aes(x=id, y=Y)) +
         geom_point() +
  geom_text(aes(label=ifelse(Y<0, round(Y,2), '')),
            hjust=-0.25, vjust=0)
```



Notice that all data points are within 0 and 1 except for three points which are negative, though very close to 0 at -0.02 and -0.07.

(3) How does the probability of being involved in the labor force change if a woman goes from having no children to having 1 child? How does this probability change if the woman has another child?

Answer: Since it is a linear probability model, it always falls by 0.224.

---

(4) If education (`educ`) can be assumed fixed at its mean value, what is the probability of being in the labor force if the woman has 3 children?

Answer: In this question we are trying to obtain the probability value for

$$\widehat{inlf} = 0.053 + 0.046\ \overline{educ} - 0.224\ kids$$

So we first need to obtain the mean value of *educ*.

In STATA:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("labour force") firstrow
sum educ
```

In R:

```
0.046*mean(laborforce_df$educ) + 0.053 - .224*3
```

```
[1] -0.0538048
```

So we are interested in the probability of being in the labor force if a woman has three children:

$$\begin{aligned}
\widehat{inlf} &= 0.053 + 0.046 \times 12.28685 - 0.224\ kids \\
&= 0.053 + 0.5651952 - 0.224\ kids \\
&= 0.6181952 - 0.224 \times 3 \\
&= -0.05380478
\end{aligned}$$

---

(5) Do your results suggest that the problems you outlined in Question A(2) are present in equation (2)?

Answer: The results do not satisfy the probability requirement $0 \leq \mathbb{E}(Y_i \mid X_i) \leq 1$ since the probability we obtained in the previous question is negative.

We can also check for heteroskedasticity using BP test first. However, since that test assumes heteroskedasticity is linear, we also perform White test.

BP test in R that gives the LM-statistic:

```r
bptest(SQB1_lm, studentize=FALSE)
# or from the `skedastic` package:
# breusch_pagan(SQB1_lm, koenker=FALSE)
```

BP test in STATA for both F-statistic and LM-statistic:

```stata
quietly cd ..
quietly import excel Data/limdep.xls, sheet("labour force") firstrow
quietly reg inlf Kids educ
hettest, rhs fstat
hettest, rhs
```

```
Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: All independent variables

H0: Constant variance

F(2, 750) =   7.57
 Prob > F = 0.0006


Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variables: All independent variables

H0: Constant variance

    chi2(2) =   3.01
Prob > chi2 = 0.2219
```

BP test manually in STATA:

```stata
quietly cd ..
quietly import excel Data/limdep.xls, sheet("labour force") firstrow
quietly reg inlf Kids educ
predict u, residuals
generate u2 = u^2
quietly regress u2 Kids educ
display e(F)
display e(r2)*e(N)
```

White test in R using the skedastic package:

```
white(SQB1_lm, interactions = TRUE)
```

In STATA:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("labour force") firstrow
quietly reg inlf Kids educ
imtest, white
```

```
White's test
H0: Homoskedasticity
Ha: Unrestricted heteroskedasticity

    chi2(5) =  45.56
Prob > chi2 = 0.0000

Cameron & Trivedi's decomposition of IM-test

--------------------------------------------------
              Source |       chi2     df         p
---------------------+----------------------------
  Heteroskedasticity |      45.56      5    0.0000
            Skewness |      75.18      2    0.0000
            Kurtosis |     250.83      1    0.0000
---------------------+----------------------------
               Total |     371.56      8    0.0000
--------------------------------------------------
```

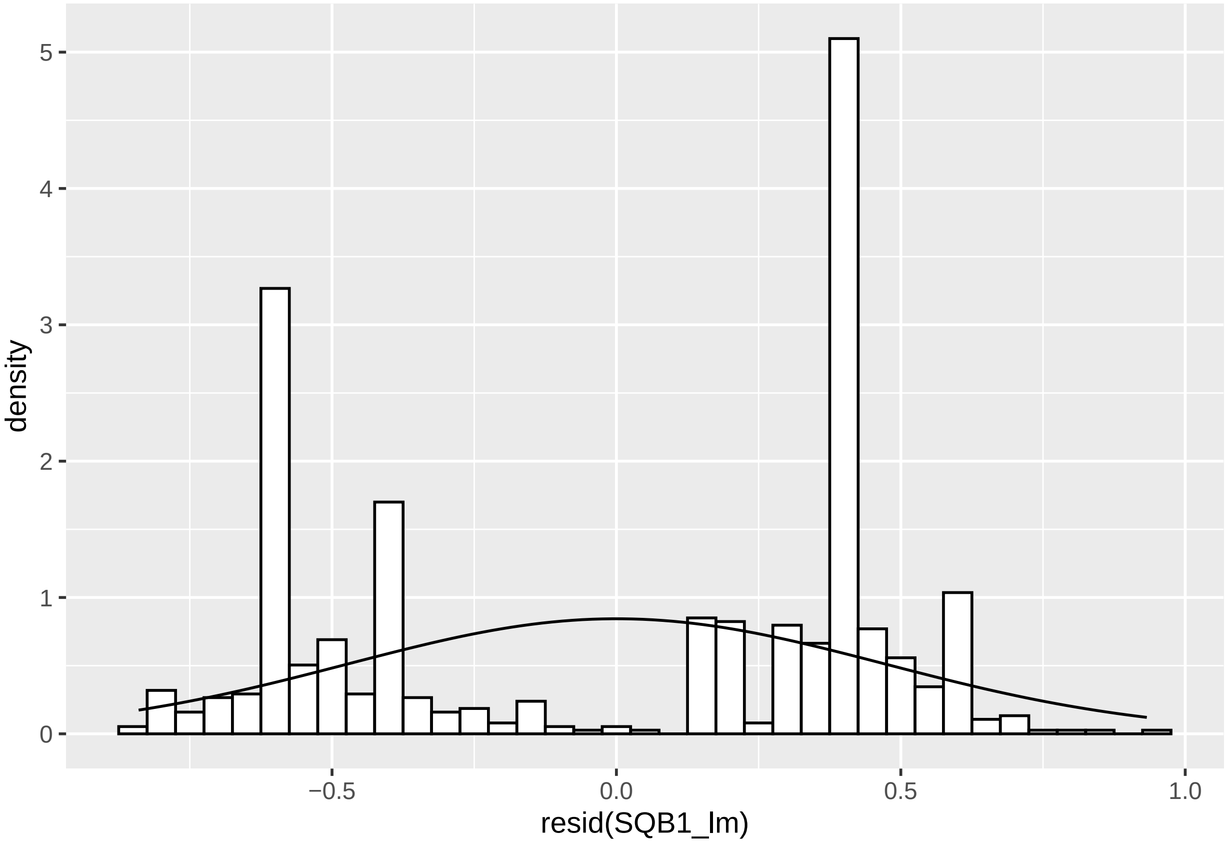
Manually in STATA:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("labour force") firstrow
quietly reg inlf Kids educ
predict u, residuals
generate u2 = u^2
generate k2 = Kids^2
generate e2 = educ^2
generate ke = Kids * educ
quietly regress u2 Kids educ k2 e2 ke
display e(N)*e(r2)
```

Thus in both BP and White tests we reject the null of homoskedasticity and conclude that presence of heteroskedasticity is very likely.

We can also check if the residuals are non-normal:

```
ggplot(laborforce_df,
       mapping = aes(x = resid(SQB1_lm))) +
         geom_histogram(
           aes(y=after_stat(density)),
           binwidth = 0.05,
           color = "black",
           fill = "white"
         ) +
  stat_function(fun = dnorm, args = list(mean = mean(resid(SQB1_lm)), sd=sd(resid(SQB1_lm))))
```



From the graph we can see that the residuals are not normally distributed. Finally, having a linear, constant marginal effects does not make much sense.

Therefore we see all three problems we outlined in Question A(2) present in equation (2).

---

(6) Estimate a Logit model of the same relationship, i.e., for

$$\mathbb{P}(inlf = 1 \mid educ, kids) \tag{3}$$

and repeat parts B(3) and B(4). Verify that the results you obtain using the ‘margins‘ command in STATA are the same as those obtained by putting different values of Kids (and the mean value of education) into logit function.

Answer: In R we build logit model by appling the ‘glm()‘ function. For the logistic regression model we specify ‘family=binomial‘.

```r
SQB6_glm <- glm(inlf ~ Kids + educ, data = laborforce_df, family = 'binomial')
summary(SQB6_glm)
```

In STATA we can do the same via:

```stata
quietly cd ..
quietly import excel Data/limdep.xls, sheet("labour force") firstrow
logit inlf Kids educ, nolog
/* nolog avoids displaying the iteration log */
```

```
Logistic regression                                     Number of obs =    753
                                                        LR chi2(2)    =  71.40
                                                        Prob > chi2   = 0.0000
Log likelihood = -479.17238                             Pseudo R2     = 0.0693

------------------------------------------------------------------------------
        inlf | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        Kids |  -1.010073   .1626127    -6.21   0.000    -1.328788   -.6913578
        educ |   .2101728    .036479     5.76   0.000     .1386754    .2816703
       _cons |  -2.053952   .4441444    -4.62   0.000    -2.924459   -1.183445
------------------------------------------------------------------------------
```

which gives us:

$$
\begin{array}{rlll}
\widehat{inlf} = & -2.054 & + 0.210\; educ & - \; -1.01\; kids \\
z: & [-4.62] & [5.76] & [-6.21] \\
se: & (0.00) & (0.00) & (0.00)
\end{array}
$$

In Question B(3) we were asked to compare the probabilities of being involved in the labor force changes as a woman going from 0 child to 1, and from 1 child to 2.

In order to obtain the marginal effects of having kids, we keep the education at its mean value and then calculate the probability using the logistic distribution function given by:

$$p_i = \mathbb{P}(Y_i = 1 \mid X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}}.$$

In R, we can do this by:

```r
# when kids=0
1/(1 + exp(-(SQB6_glm$coefficients[1] + SQB6_glm$coefficients[3]*mean(laborforce_df$educ))))
```

```
(Intercept)
   0.629112
```

```r
# when kids=1
1/(1 + exp(-(SQB6_glm$coefficients[1] + SQB6_glm$coefficients[3]*mean(laborforce_df$educ)
             + SQB6_glm$coefficients[2]*1)))
```

```
(Intercept)
    0.38186
```

```
# when kids=2
1/(1 + exp(-(SQB6_glm$coefficients[1] + SQB6_glm$coefficients[3]*mean(laborforce_df$educ)
             + SQB6_glm$coefficients[2]*2)))
```

```
(Intercept)
   0.183661
```

We can automate this process in STATA using the `margins` function with the options of `at` and `atmeans`. The `at` option calculates the marginal effects at specified values. The `atmeans` option calculates the marginal effects at mean of a dataset rather than the default behavior of calculating the average marginal effects. Finally, the `post` option causes STATA to overwrite the original regression estimates with the "margins" estimates.

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("labour force") firstrow
quietly logit inlf Kids educ
margins, at (Kids = (0 1 2 3)) atmeans noatlegend post
```

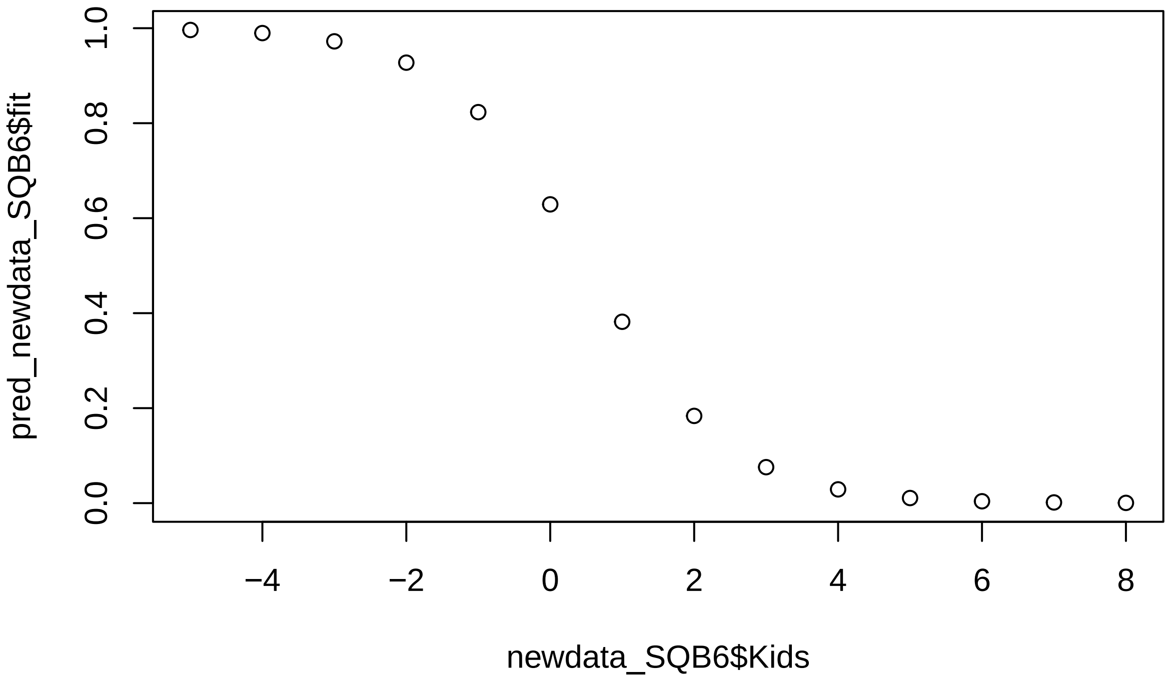
```
Adjusted predictions                                        Number of obs = 753
Model VCE: OIM

Expression: Pr(inlf), predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .6291122   .0199695    31.50   0.000     .5899727    .6682518
          2  |   .3818596    .034906    10.94   0.000     .3134451    .4502741
          3  |   .1836614   .0448032     4.10   0.000     .0958487     .271474
          4  |   .0757315   .0320468     2.36   0.018      .012921     .138542
------------------------------------------------------------------------------
```

In R, this is a more protracted process. One can use `margins()` function that replicates STATA's `margins` but it does not have the `atmeans` option. So we have to add the mean values into a new dataset first:

```
newdata_SQB6 <- expand.grid(educ = mean(laborforce_df$educ),
                            Kids = seq(0,3,1))
pred_newdata_SQB6 <- predict(object = SQB6_glm,
                             newdata = newdata_SQB6,
                             type = "response",
                             se.fit = TRUE)
mult_SQB6 <- qnorm(0.5*(1-0.95))
out_SQB6 <- cbind(pred_newdata_SQB6$fit,
                  pred_newdata_SQB6$se.fit,
                  pred_newdata_SQB6$fit + pred_newdata_SQB6$se.fit * mult_SQB6,
                  pred_newdata_SQB6$fit - pred_newdata_SQB6$se.fit * mult_SQB6)
rownames(out_SQB6) <- seq(1,4,1)
colnames(out_SQB6) <- c("margin", "Std. Err.", "lower 95% conf", "upper 95% conf")
out_SQB6
```

We can plot the margins using the `marginsplot` function in STATA or the following in R:

```
newdata_SQB6 <- expand.grid(educ = mean(laborforce_df$educ),
                            Kids = seq(-5,8,1))
pred_newdata_SQB6 <- predict(object = SQB6_glm,
                             newdata = newdata_SQB6,
                             type = "response",
                             se.fit = TRUE)
plot(newdata_SQB6$Kids,pred_newdata_SQB6$fit)
```



If we are interested in the actual marginal effects at different points rather than the differences between them, then we can use:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("labour force") firstrow
quietly logit inlf Kids educ
margins, dydx (Kids) at (Kids = (0 1 2 3)) atmeans noatlegend
```

```
Conditional marginal effects                          Number of obs = 753
Model VCE: OIM

Expression: Pr(inlf), predict()
dy/dx wrt:  Kids

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
Kids         |
         _at |
          1  |  -.2356803   .0360248    -6.54   0.000    -.3062877    -.165073
          2  |  -.2384205   .0315862    -7.55   0.000    -.3003283   -.1765126
          3  |  -.1514401   .0081041   -18.69   0.000    -.1673239   -.1355563
          4  |  -.0707013   .0163617    -4.32   0.000    -.1027696   -.0386331
------------------------------------------------------------------------------
```

In R, we can get similar marginal effects via:

```
margins(SQB6_glm, variables = "Kids", at = list(Kids=(0:3)))
```

```
Average marginal effects at specified values

glm(formula = inlf ~ Kids + educ, family = "binomial", data = laborforce_df)

 at(Kids)     Kids
        0 -0.22596
        1 -0.22799
        2 -0.15227
        3 -0.07511
```

QUESTION C

Using the 'loans' data from the data set 'limdep.xls':

(1) Regress `approve` on `white` and report your results. Interpret the coefficient on `white`. Is it statistically significant? Is it practically large? What is the probability of getting a loan if you are white?

Answer: In R:

```
loans_df <- read_excel("../Data/limdep.xls", sheet = "loans")
SQC1_lm <- lm(approve ~ white, data = loans_df)
summary(SQC1_lm)
```

in STATA:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
reg approve white
```

```
      Source |       SS           df       MS      Number of obs   =     1,989
-------------+----------------------------------   F(1, 1987)      =    102.23
       Model |  10.4743407         1  10.4743407   Prob > F        =    0.0000
    Residual |   203.59303     1,987  .102462521   R-squared       =    0.0489
-------------+----------------------------------   Adj R-squared   =    0.0485
       Total |  214.067371     1,988  .107679764   Root MSE        =     .3201

------------------------------------------------------------------------------
     approve | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       white |   .2005957     .01984    10.11   0.000     .1616864     .239505
       _cons |   .7077922   .0182393    38.81   0.000     .6720221    .7435623
------------------------------------------------------------------------------
```

The coefficient of 20% is significant with *t*-statistic of 10.11.

In order to find out the probability of getting a loan if white, we need to obtain the marginal probability:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
quietly reg approve white
margins, at (white=(1 0))
```

```
Adjusted predictions                                     Number of obs = 1,989
Model VCE: OLS

Expression: Linear prediction, predict()
1._at: white = 1
2._at: white = 0

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .9083879   .0078073   116.35   0.000     .8930766    .9236991
          2  |   .7077922   .0182393    38.81   0.000     .6720221    .7435623
------------------------------------------------------------------------------
```

or in R:

```
predict(SQC1_lm)[1]
```

Therefore, the probability of approval is about 91%.

(2) Given your answers to Question A(3) on page 5 above, compute the weighted least square estimates for part C(1) by first computing your own weights and using these to transform the relevant variables. Verify these results using a weighted least squares option in STATA. Show that in this case the results obtained are identical to those resulting from the use of the robust estimates of C(1). Generally would you expect these results to be the same?

Answer: Recall in Question A(3) we weighted the LPM to deal with heteroskedasticity as follows:

$$\frac{Y_i}{\sqrt{w_i}} = \beta_0 \frac{1}{\sqrt{w_i}} + \beta_1 \frac{X_i}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}}$$

where $\sqrt{w_i} = \sqrt{\big(\mathbb{E}(Y_i \mid X_i)\big)\big(1 - \mathbb{E}(Y_i \mid X_i)\big)} = \sqrt{p_i(1-p_i)}$.

We will therefore generate the weights based on the product of p_i and $1 - p_i$.

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
quietly reg approve white
/* using own weights */
quietly predict yhat
generate w = sqrt(yhat*(1-yhat))
generate approve_w = approve/w
generate white_w = white/w
generate cons_w = 1/w
reg approve_w white_w cons_w, noconstant

/* using weighted least squares */
vwls approve white, sd(w)

/* using robust estimates */
reg approve white, robust
```

```
      Source |       SS           df       MS      Number of obs   =     1,989
-------------+----------------------------------   F(2, 1987)      =   8698.32
       Model |  17414.1374         2  8707.06868   Prob > F        =    0.0000
    Residual |  1988.99957     1,987  1.00100633   R-squared       =    0.8975
-------------+----------------------------------   Adj R-squared   =    0.8974
       Total |  19403.1369     1,989  9.75522219   Root MSE        =    1.0005

------------------------------------------------------------------------------
   approve_w | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     white_w |   .2005957   .0268651     7.47   0.000      .147909    .2532823
      cons_w |   .7077922   .0259264    27.30   0.000     .6569465    .7586379
------------------------------------------------------------------------------


Variance-weighted least-squares regression           Number of obs   =    1,989
Goodness-of-fit chi2(1987) = 1989.00                 Model chi2(1)   =    55.81
Prob > chi2                =  0.4831                 Prob > chi2     =   0.0000
------------------------------------------------------------------------------
     approve | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       white |   .2005957   .0268516     7.47   0.000     .1479675    .2532238
       _cons |   .7077922   .0259133    27.31   0.000      .657003    .7585814
------------------------------------------------------------------------------


Linear regression                               Number of obs     =      1,989
                                                F(1, 1987)        =      55.75
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0489
                                                Root MSE          =      .3201

------------------------------------------------------------------------------
             |               Robust
     approve | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       white |   .2005957   .0268651     7.47   0.000      .147909    .2532824
       _cons |   .7077922   .0259264    27.30   0.000     .6569465     .758638
------------------------------------------------------------------------------
```

These confirm that there are very little differences in the results with t-statistic 27.3 and 7.47, respectively.

(3) Add the variables *obrat, loanprc, chist*, **and** *pubrec* **as controls. Test the hypothesis that the newly added variables are jointly significant. Is there still evidence of discrimination against non-white?**

In R:

```
summary(lm(approve ~ white + obrat + loanprc + pubrec + chist, data = loans_df))
```

In STATA:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
reg approve white obrat loanprc pubrec chist
test obrat loanprc pubrec chist
```

```
      Source |       SS           df       MS      Number of obs   =     1,989
-------------+----------------------------------   F(5, 1983)      =     70.85
       Model |  32.4439721         5  6.48879443   Prob > F        =    0.0000
    Residual |  181.623398     1,983  .091590216   R-squared       =    0.1516
-------------+----------------------------------   Adj R-squared   =    0.1494
       Total |  214.067371     1,988  .107679764   Root MSE        =    .30264

------------------------------------------------------------------------------
     approve | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
```

```
       white |   .1311088   .0193517     6.78   0.000     .0931569    .1690607
       obrat |  -.0045293   .0008432    -5.37   0.000    -.0061829   -.0028757
     loanprc |  -.1474878   .0370815    -3.98   0.000    -.2202106   -.0747649
      pubrec |  -.2466396   .0279666    -8.82   0.000    -.3014867   -.1917926
       chist |   .1339311     .01916     6.99   0.000     .0963553    .1715068
       _cons |   .9316853   .0453511    20.54   0.000     .8427446    1.020626
------------------------------------------------------------------------------


 ( 1)  obrat = 0
 ( 2)  loanprc = 0
 ( 3)  pubrec = 0
 ( 4)  chist = 0

       F(  4,  1983) =   59.97
            Prob > F =    0.0000
```

All the coefficients are individually and jointly significant. The F-stat was calculated using:

$$F = \frac{\frac{RSS_1-RSS_2}{k_2-k_1}}{\frac{RSS_2}{n-k_2}} = \frac{\frac{203.59303-181.623398}{6-2}}{\frac{181.623398}{1989-6}} = 59.97$$

We can also do the same using the robust estimates:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
reg approve white obrat loanprc pubrec chist, robust
test obrat loanprc pubrec chist
```

```
Linear regression                               Number of obs     =      1,989
                                                F(5, 1983)        =      40.16
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1516
                                                Root MSE          =     .30264

------------------------------------------------------------------------------
             |               Robust
     approve | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       white |   .1311088   .0254313     5.16   0.000     .0812339    .1809837
       obrat |  -.0045293   .0010589    -4.28   0.000     -.006606   -.0024526
     loanprc |  -.1474878   .0375634    -3.93   0.000    -.2211555     -.07382
      pubrec |  -.2466396   .0421519    -5.85   0.000    -.3293063   -.1639729
       chist |   .1339311   .0247077     5.42   0.000     .0854752    .1823869
       _cons |   .9316853   .0525287    17.74   0.000      .828668    1.034703
------------------------------------------------------------------------------


 ( 1)  obrat = 0
 ( 2)  loanprc = 0
 ( 3)  pubrec = 0
 ( 4)  chist = 0
```

```
       F(  4,  1983) =   35.13
            Prob > F =    0.0000
```

Therefore, there still seems to be significant evidence of discrimination, t-stat on white is either 6.78, or in the robust estimates, 5.16.

---

(4) Now let the effect of race interact with the variable measuring other obligations as a percent of income (*obrat*). Is the interaction term significant? Interpret your results - especially, interpret the coefficients on *white*, *obrat* and the interaction term *white.obrat*.

Answer:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
gen whiteobrat = white*obrat
reg approve white obrat loanprc pubrec chist whiteobrat
```

```
      Source |       SS           df       MS      Number of obs   =     1,989
-------------+----------------------------------   F(6, 1982)      =     61.19
       Model |  33.4569032         6  5.57615053   Prob > F        =    0.0000
    Residual |  180.610467     1,982  .091125362   R-squared       =    0.1563
-------------+----------------------------------   Adj R-squared   =    0.1537
       Total |  214.067371     1,988  .107679764   Root MSE        =    .30187

------------------------------------------------------------------------------
     approve | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       white |  -.1265832   .0796651    -1.59   0.112    -.2828194     .029653
       obrat |   -.010911   .0020907    -5.22   0.000    -.0150113   -.0068108
     loanprc |  -.1521892   .0370142    -4.11   0.000      -.22478   -.0795984
      pubrec |  -.2457575   .0278968    -8.81   0.000    -.3004676   -.1910473
       chist |   .1310461   .0191309     6.85   0.000     .0935274    .1685649
  whiteobrat |   .0075756   .0022722     3.33   0.001     .0031194    .0120317
       _cons |   1.157147   .0813592    14.22   0.000     .9975884    1.316705
------------------------------------------------------------------------------
```

Which gives us:

$$\begin{array}{rlllllll}
\widehat{approve} = 1.157 & -\,0.127\ white & -\,0.011\ obrat & -\,0.152 loanprc & -\,0.246 pubrec & +\,0.131 chist & +\,0.008 whiteobrat \\
t: [14.22] & [-1.58] & [-5.22] & [-4.11] & [-8.81] & [6.85] & [3.33] \\
se: (0.081) & (0.112) & (0.000) & (0.000) & (0.000) & (0.000) & (0.001)
\end{array}$$

or, using robust estimates:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
gen whiteobrat = white*obrat
reg approve white obrat loanprc pubrec chist whiteobrat, robust
```

```
Linear regression                               Number of obs     =      1,989
                                                F(6, 1982)        =      34.30
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1563
                                                Root MSE          =     .30187

------------------------------------------------------------------------------
             |               Robust
     approve | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       white |  -.1265832   .1039457    -1.22   0.223    -.3304375     .077271
       obrat |   -.010911   .0028964    -3.77   0.000    -.0165914   -.0052307
     loanprc |  -.1521892   .0379064    -4.01   0.000    -.2265297   -.0778487
      pubrec |  -.2457575   .0423392    -5.80   0.000    -.3287916   -.1627234
       chist |   .1310461   .0246588     5.31   0.000     .0826862    .1794061
  whiteobrat |   .0075756   .0030829     2.46   0.014     .0015296    .0136216
       _cons |   1.157147   .1058147    10.94   0.000     .9496271    1.364667
------------------------------------------------------------------------------
```

Which gives us:

$$\begin{array}{rllllllll}
\widehat{approve} = 1.157 & -\,0.127\ white & -\,0.011\ obrat & -\,0.152 loanprc & -\,0.246 pubrec & +\,0.131 chist & +\,0.008 whiteobrat \\
t: [10.94] & [-1.22] & [-3.77] & [-4.01] & [-5.80] & [5.31] & [2.46 \\
se: (0.000) & (0.223) & (0.000) & (0.000) & (0.000) & (0.000) & (0.014)
\end{array}$$

The interaction term is significant, although not at $\alpha = 1\%$ if robust estimates are considered. *white* is not significant in either, thus it is not different than 0 in explaining whether a loan application would be approved or not. An applicant gets penalized for having high ratio of other obligations as a percent of total income. As that ratio increases by one unit, the probability of loan declines by 0.01. Thus the cross product shows that although having higher *obrat* penalizes, this penalty appears lower for *white* since its coefficient is 0.008.

---

(5) Estimate the coefficient (marginal effect) of *white* if *obrat* is at its mean, using your results from Question C(4). Show that this result can be obtained by running the previous regerssion once more but this time with an amended interaction term $white \times (obrat - \overline{obrat})$, where $\overline{obrat}$ is the mean of *obrat*. Discuss your results, especially the interpretation of your coefficient on *white*.

Answer: The marginal effect of *white* if *obrat* is at its mean is going to be:

$$\beta_1\ white + (\beta_6\ whiteobrat) \times \overline{obrat}$$

In STATA:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
generate whiteobrat = white*obrat
quietly reg approve white obrat loanprc pubrec chist whiteobrat, robust
egen obratavg = mean(obrat)
display _b[white] + _b[whiteobrat]*obratavg
```

```
.11878212
```

The second part of the question is asking for us to run the regression in Question C(4) but this time with the amended interaction term $white \times (obrat - \overline{obrat})$.

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
generate whiteobrat = white*obrat
egen obratavg = mean(obrat)
gen whitrat = white * (obrat-obratavg)
reg approve white obrat loanprc pubrec chist whitrat
```

```
      Source |       SS           df       MS      Number of obs   =     1,989
-------------+----------------------------------   F(6, 1982)      =     61.19
       Model |  33.4569032         6  5.57615054   Prob > F        =    0.0000
    Residual |  180.610467     1,982  .091125362   R-squared       =    0.1563
-------------+----------------------------------   Adj R-squared   =    0.1537
       Total |  214.067371     1,988  .107679764   Root MSE        =    .30187

------------------------------------------------------------------------------
     approve | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       white |   .1187821   .0196535     6.04   0.000     .0802385    .1573257
       obrat |   -.010911   .0020907    -5.22   0.000    -.0150113   -.0068108
     loanprc |  -.1521892   .0370142    -4.11   0.000      -.22478   -.0795984
      pubrec |  -.2457575   .0278968    -8.81   0.000    -.3004676   -.1910473
       chist |   .1310461   .0191309     6.85   0.000     .0935274    .1685649
     whitrat |   .0075756   .0022722     3.33   0.001     .0031194    .0120317
       _cons |   1.157147   .0813592    14.22   0.000     .9975884    1.316705
------------------------------------------------------------------------------
```

and with robust estimators:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
generate whiteobrat = white*obrat
egen obratavg = mean(obrat)
gen whitrat = white * (obrat-obratavg)
reg approve white obrat loanprc pubrec chist whitrat, robust
```

```
Linear regression                               Number of obs     =      1,989
                                                F(6, 1982)        =      34.30
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1563
                                                Root MSE          =     .30187

------------------------------------------------------------------------------
             |               Robust
     approve | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       white |   .1187821   .0250042     4.75   0.000     .0697449    .1678193
       obrat |   -.010911   .0028964    -3.77   0.000    -.0165914   -.0052307
     loanprc |  -.1521892   .0379064    -4.01   0.000    -.2265297   -.0778487
      pubrec |  -.2457575   .0423392    -5.80   0.000    -.3287916   -.1627234
       chist |   .1310461   .0246588     5.31   0.000     .0826862    .1794061
     whitrat |   .0075756   .0030829     2.46   0.014     .0015296    .0136216
       _cons |   1.157147   .1058147    10.94   0.000     .9496271    1.364667
------------------------------------------------------------------------------
```

Thus, in all approaches the marginal effect of *white* if *obrat* is at its mean is $\beta_{white} = 0.1188$. Notice that $\overline{obrat} = 32.39$, so coefficient on *white* is the race differential when $obrat = 32.39$.

(6) Now estimate a Probit model of *approve* on *white*. Find the estimated probability of loan approval for both white and non-whites. How do these compare with the estimates from the LPM?

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
probit approve white, nolog
display normal(_b[_cons])
display normal(_b[_cons]+_b[white])
```

```
Probit regression                                       Number of obs =  1,989
                                                        LR chi2(1)    =  78.94
                                                        Prob > chi2   = 0.0000
Log likelihood = -700.87744                             Pseudo R2     = 0.0533

------------------------------------------------------------------------------
     approve | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       white |   .7839465   .0867118     9.04   0.000     .6139946    .9538985
       _cons |   .5469463    .075435     7.25   0.000     .3990964    .6947962
------------------------------------------------------------------------------

.70779219

.90838786
```

From the regression we have

$$\widehat{approve} = \Phi(0.547 + 0.784\ white) \begin{cases} \mathbb{P}(1) = \Phi(1.331) = 0.908 \\ \mathbb{P}(0) = \Phi(0.547) = 0.708 \end{cases}$$

Alternatively, we can use the `margins` command:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
quietly probit approve white
margins, at (white = (0 1)) noatlegend
```

```
Adjusted predictions                                    Number of obs = 1,989
Model VCE: OIM

Expression: Pr(approve), predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .7077922   .0259133    27.31   0.000      .657003    .7585814
          2  |   .9083879    .007036   129.10   0.000     .8945975    .9221782
------------------------------------------------------------------------------
```

In either case we see that the results are the same as the marginal effect we obtained in Question C(1) on page 19.

(7) Now add the same variables as in Question C(3) on page 21 to the probit model. Use the likelihood ratio test to assess whether the extra variables should be included in the equation. Does any statistically significant evidence of discrimination against non-whites remain?

Answer: In the first part of the question we will add all the control variables to our probit model.

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
probit approve white obrat loanprc pubrec chist, nolog
```

```
Probit regression                                      Number of obs =  1,989
                                                       LR chi2(5)    = 248.17
                                                       Prob > chi2   = 0.0000
Log likelihood = -616.25975                            Pseudo R2     = 0.1676

------------------------------------------------------------------------------
     approve | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       white |   .5291924   .0941964     5.62   0.000      .344571    .7138139
       obrat |  -.0243472   .0046577    -5.23   0.000    -.0334761   -.0152183
     loanprc |  -.9982303   .2307498    -4.33   0.000    -1.450492   -.5459691
      pubrec |  -.8080753   .1244161    -6.49   0.000    -1.051926   -.5642242
       chist |   .5686446   .0939647     6.05   0.000     .3844772    .7528119
       _cons |   2.014055   .2675501     7.53   0.000     1.489666    2.538443
------------------------------------------------------------------------------
```

With the likelihood test we are testing if the coefficients of the control variables are jointly 0. The likelihood ratio test is given by:

$$\lambda_{LR} = -2\ln\left(\frac{L^{unr}(\theta)}{L^{res}(\theta)}\right)$$

or in log-likelihoods:

$$\lambda_{LR} = -2\big(\ell^{unr}(\theta) - \ell^{res}(\theta)\big).$$

In this question $\ell^{unr}(\theta) = -616.25975$ and $\ell^{res} = -700.87744$. Therefore,

$$\lambda_{LR} = -2(-616.25975 + 700.87744) = 169.235$$

This is χ^2 distributed with 4 degrees of freedom. The probability is:

```
pchisq(169.235,4, lower.tail = FALSE)
```

```
[1] 0.0000000000000000000000000000000000152633
```

therefore we reject the null and conclude that the control variables are jointly significant. We can do the same test using the `lrtest` command:

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
quietly probit approve white
estimates store restricted
quietly probit approve white obrat loanprc pubrec chist
lrtest restricted
```

```
Likelihood-ratio test
Assumption: restricted nested within .

 LR chi2(4) = 169.24
Prob > chi2 = 0.0000
```

(8) Calculate the probability of getting a loan if you are white and non-white. Compare your results with those from Question C(1) on page 18.

```
quietly cd ..
quietly import excel Data/limdep.xls, sheet("loans") firstrow
quietly probit approve white obrat loanprc pubrec chist
margins, at (white = (1 0)) atmeans noatlegend
```

```
Adjusted predictions                                    Number of obs = 1,989
Model VCE: OIM

Expression: Pr(approve), predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .9201425   .0069723   131.97   0.000      .906477    .9338079
          2  |   .8097127   .0232573    34.82   0.000     .7641293    .8552961
------------------------------------------------------------------------------
```

Manually these marginal rates are obtained as follows:

	$\hat{\beta}_i$	$\bar{X}$	$\hat{\beta}_i \bar{X}_i$
white	0.5291924	1	0.5291924
obrat	-0.0243472	32.389	-0.788581
loanprc	-0.9982303	0.7706	-0.769236
pubrec	-.8080753	0.068879	-0.0556594
chist	0.5686446	0.8376	0.476297
constant			2.014055
			$\sum \hat{\beta}_i \bar{X} = 1.40607$
			$\Phi(1.40607) = 0.920148$

```r
-0.788581-0.769236-0.0556594+0.476297+2.014055
```

```
[1] 0.876876
```

```r
-0.0243472 * 32.389
```

```
[1] -0.788581
```

```r
pnorm(0.876876)
```

```
[1] 0.809723
```

```r
0.920148-0.8092723
```

```
[1] 0.110876
```

Similarly, if *white* $= 0$ then $\sum = 0.876876$ and $\Phi(0.876876) = 0.809723$. So now we see a difference of only 11.09%.