Notes for Applied Econometrics

Christopher Ferrall

Department of Economics
Queen's University

ferrallc@post.queensu.ca

February 21, 2001

These notes are based on material for Econ 351 and are intended primarily as a supplement to in-class work and assignments. Thanks to the many students who have put up with and found all the typos that no longer appear in the pages to follow. Thanks in advance to current students for putting up with the remaining typos.

I. REVIEW OF MULTIPLE LINEAR REGRESSION

I.1 Introduction

To motivate multiple regression, let's consider the demand for milk by households. To eliminate the supply curve we will assume that milk is produced with a constant Marginal Cost technology so that quantities do not affect equilibrium prices. Let m be the quantity of milk purchased (say each week) and let I be household income. Suppose the household has the utility function

$$U(m, I - pm) = m^{\alpha}(I - pm)^{1-\alpha} \tag{I.1}$$

where p is the price of milk faced by the household and U is the utility function for the household defined in terms of milk and all other goods. The Marshallian demand function in terms of log milk consumption takes the form:

$$\ln m = \ln \alpha + \ln I - \ln p_m. \tag{I.2}$$

In other words, under Cobb-Douglas utility function the amount of milk purchased by the household is log-linear in price and income, with coefficients on income and the price of milk equal to 1 in absolute value. We might specify the following regression equation:

$$\ln m = \beta_1 + \beta_2 I + \beta_3 p + u \tag{I.3}$$

The disturbance term u accounts for different *tastes* for milk across households associated with different values of α. Households with the same income and facing the same prices might have different milk consumptions because of differences in tastes: for example, one household might have young children which would push up total milk consumption.

Even with tastes differing across households, the Cobb-Douglas model implies two hypotheses about the MLRM: $\beta_2 = 1$ and $\beta_3 = -1$. These hypotheses might be interesting to test given data on milk prices, household income, and household consumption of milk.

But notice that this demand function has two variables: income and milk price. We could certainly estimate a simple linear regression model including only one variable or the other (either price or income). But since both variables can be measured much less would be left to the residual term if both were included in the analysis. Furthermore, the joint test of the Cobb-Douglas model in the paragraph above could not be implemented. The Multiple LRM is generalizes the simple LRM to allow two or more variables on the right hand side of the PRE.

I.2 Definition of the Multiple Regression Model

I.2.a Assumptions A0-A3

The multiple linear regression model (MLRM) generalizes the simple LRM:

A.0. Sample Information

The data consist of N vectors of k related variables.

$$\begin{aligned} \text{Sample} \equiv [&(X_{21}, X_{31}, \ldots, X_{k1}, Y_1), \\ &(X_{22}, X_{32}, \ldots, X_{k2}, Y_2), \\ &\vdots \\ &(X_{2N}, X_{3N}, \ldots, X_{kN}, Y_N),]. \end{aligned} \tag{I.4}$$

where X_{ji} is the value of the j^{th} variable for the i^{th} observation. For example, X_2 might be income, X_3 might be the price of milk, X_4 might be number of children, etc. Then X_{43} equals the number of children in household (or observation) 3. We begin numbering the X variables at 2 to hold a place for the constant term.

A.1. The Population Regression Equation(PRE).

The variables in X and Y are related by:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u \tag{I.5}$$

Each coefficient β_j is an unknown population parameter. Notice that the simple linear regression model is the special case of the multiple regression model, $k = 2$. Since an economic variable like Y is undoubtedly influenced by many potentially observable influences (like price, income) there is perhaps a better way to interpret any linear regression model as a special case of a larger (greater k) model. In particular, we should think of the smaller model as a difference in the observed sample information. If we do not observe all the factors that influence Y, then we have a a sample restricted in the X variables that we can observe. We do not necessarily have the *wrong* model if we only include one variable in the regression when there should be many.

The sample for a multiple regression model is written:

$$\begin{aligned}
Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \ldots + \beta_k X_{k1} + u_1 \\
Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \ldots + \beta_k X_{k2} + u_2 \\
&\vdots \\
Y_N &= \beta_1 + \beta_2 X_{2N} + \beta_3 X_{3N} + \ldots + \beta_k X_{kN} + u_N
\end{aligned} \tag{I.6}$$

A.2* The PRE written in Matrix Notation

It is much easier to work with the multiple regression model in matrix notation. To do this we have to define several vectors and matrices:

$$y \equiv \begin{vmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{vmatrix} \tag{I.7}$$

$$\beta \equiv \begin{vmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{vmatrix} \tag{I.8}$$

$$u \equiv \begin{vmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{vmatrix} \tag{I.9}$$

$$X \equiv \begin{vmatrix} 1 & X_{21} & \dots & X_{k1} \\ 1 & X_{22} & \dots & X_{k1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & X_{2N} & \dots & X_{kN} \end{vmatrix} \tag{I.10}$$

Note: The subscripts for the matrix X are the reverse of the usual row/column ordering. Why? We want to "bind" the variable name first (X_2 refers to variable number 2), but we naturally write an equation on one line (or row), forcing us to have X_2 refer to a column rather than a row.

With the notation above, all the sample information is represented by one matrix equation.

$$y = X\beta + u \tag{I.11}$$

I.2.b Interpretation of β_j in the LRM

Take expectations of both sides of (premat):

$$E[Y_i|X_2, X_3, \dots, X_k] = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k + E[u|X_2, X_3, \dots, X_k] \tag{I.12}$$

Then, using basic calculus we see

$$\frac{\partial E[Y_i|X_2, X_3, \dots, X_k]}{\partial X_j} = \beta_j \tag{I.13}$$

We can interpret the coefficient on the jth variable to be the partial derivative of the expected value of Y with respect to a change in the jth variable. For example, if Y is log-consumption of milk and X_2 is the log of household income, then β_2 is the income elasticity of milk demand.

I.3 Derivation of OLS Estimates for the Multiple Regression Model

I.3.a Definition of variables used in OLS estimation

$$\hat{\beta} \equiv \begin{vmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{vmatrix} \tag{I.14}$$

$$\hat{y} \equiv X\hat{\beta} \tag{I.15}$$

$$e \equiv y - \hat{y} = y - X\hat{\beta} \tag{I.16}$$

$$SS \equiv e'e \tag{I.17}$$

OLS estimates of β minimize the scalar function SS. In class we derive the OLS estimator for the MLRM:

$$\hat{\beta}^{OLS} \equiv (X'X)^{-1}X'y. \tag{I.18}$$

This expression is only meaningful if the $k \times k$ matrix $X'X$ is invertible. We rely on the following result which we will not prove:

> *$X'X$ is invertible if and only if The columns of X (corresponding to variables in the data set) are linearly independent of each other.*

Linear independence means that no column can be written as a linear combination of the others. You might recall that this is the same as saying that X has *full rank*. You may be tempted to *distribute* the inverse operator through the matrix multiplication:

$$(X'X)^{-1} = X^{-1}X^{-1}$$

If this is possible, then $\hat{\beta}^{OLS}$ reduces to $X^{-1}y$. But note that one can distribute the inverse operator *only if* each matrix in the multiplication has an inverse. X is a $N \times k$, and can only have an inverse if $k = N$, because only square matrices have inverses. Therefore, if $k < N$ then it is not possible to simplify (I.18) any more.

OLS in 3 Special Cases of the MLRM

Case.1. $k = 1$ (only a constant term appears on the right hand side)

Now X is simply a column of 1's and must have full rank. $X'X$ then equals the scalar N (because it equals $1 * 1 + 1 * 1 + ... + 1 * 1 = N$. The matrix X'y simply equals the sum of Y values in the sample. So the OLS estimate $\hat{\beta}_1$ reduces to $\bar{Y}$ when only a constant term is included in the regression.

Case.2. $k = 2$ (simple LRM)

This is the case studied as the Simple LRM. You should verify that:

$$(X'X) = \begin{vmatrix} N & \sum X_i \\ \sum X_i & \sum X_i^2 \end{vmatrix}. \tag{I.19}$$

Then you should use the formula for the inverse of a 2×2 matrix to get $(X'X)^{-1}$. With some manipulation it is possible to show that (I.19) gives back exactly the scalar formulas for $\hat{\beta}_1$ and $\hat{\beta}_2$ for the case $k = 2$.

Case.3. $k = N$

In this case we have as many parameters to estimate as we have observations. We have N equations in N unknowns. As long as X (which is now square $N \times N$ and $k \times k$ since $N = k$) has an inverse, then OLS will find the estimates that satisfy the equation $y = X\hat{\beta}$, which is simply $\hat{\beta} = X^{-1}y$. But we already have seen that (I.18) collapses to this expression if X is invertible. The prediction error is identically equal to 0 for each observation because we can perfectly explain each observation. If X is not invertible, then we cannot find N different values of $\hat{\beta}$. We no longer have a system of linearly independent equations because X is no longer full rank.

Using the derived formulas for the OLS estimates we get the OLS prediction and error formulas:

$$\begin{aligned} \hat{y} &= X(X'X)^{-1}X'y \qquad (I.20) \\ e &= (I - X(X'X)^{-1}X')y \\ \hat{\sigma}^2 &= e'e/(N-2) = \frac{1}{N-2}y'(I - X(X'X)^{-1}X')'(I - X(X'X)^{-1}X')y \\ &= \frac{1}{N-2}y'(I - X(X'X)^{-1}X')y \end{aligned}$$

Note that each of these expressions except the last takes the form "some matrix that involves only X" × the vector y. In other words, each is a linear function of the Y observations in the sample. The final expression is different since it includes y' and y.

Exercise: Use the formulas in (I.20) to derive **computational properties** of OLS estimates for the MLRM that are analogous to those derived for the simple LRM.

I.4 Computing and Interpreting OLS Estimates of the MLRM

Two Examples

Ex.1. Purely Numerical Example

```
. list y x1 x2 x3
y          x1         x2         x3
1. -6.545413          1          1          0
2. -1.589974          2          4   .6931472
3.  1.504102          3          9   1.098612
4. -5.820597          4         16   1.386294
5. -10.82546          5         25   1.609438
6.   -6.8942          6         36   1.791759
7. -13.92935          7         49    1.94591
8. -19.82583          8         64   2.079442
9. -17.99812          9         81   2.197225
10. -16.14051         10        100    2.302585

* Note that x2 = x1 * x1, which is okay because x2 is not LINEARLY dependent
* on x1

. regress y x1 x2 x3

Source |       SS       df       MS                  Number of obs =      10
---------+------------------------------               F(  3,     6) =   14.91
Model |   402.13325     3  134.044417              Prob > F      =  0.0035
Residual |  53.9568585     6  8.99280976               R-square      =   0.8817
---------+------------------------------               Adj R-square  =   0.8225
Total |  456.090108     9  50.6766787              Root MSE      =  2.9988
------------------------------------------------------------------------------
y |      Coef.   Std. Err.       t     P>|t|       [95 Conf. Interval]
---------+--------------------------------------------------------------------
x1 |  -18.23062   6.387585     -2.854   0.029      -33.86047   -2.600757
x2 |   .7900252   .3449069      2.291   0.062      -.0539316    1.633982
x3 |   32.77203   11.73369      2.793   0.031       4.060729    61.48333
_cons |    10.54565   5.461673      1.931   0.102       -2.81858    23.90989
------------------------------------------------------------------------------

* If Stata didn't have the  regress command, you could use
* Stata matrix commands to compute OLS estimates using (I.18)
*  See "help matrix" in Stata
*  The accum subcommand computes X'X using the variables you list
. matrix accum XpX = x1 x2 x3
. matrix list XpX, nohalf
x1          x2         x3      _cons
x1         385       3025  102.08283         55
x2        3025      25333  776.24766        385
x3   102.08283  776.24766  27.650244  15.104413
_cons          55        385  15.104413         10

*  The inv() function computes the matrix inverse
. matrix Xi = inv(XpX)
. matrix list Xi, nohalf
x1          x2          x3       _cons
x1   4.5370967  -.24149327  -8.1097843   -3.407188
x2  -.24149327   .01322843   .41656987   .18971409
x3  -8.1097843   .41656987   15.309946   5.4410995
_cons   -3.407188   .18971409   5.4410995   3.3170804

* the vecaccum subcommand treats the first listed variable as the y vector
* and computes X'y
. matrix vecaccum Xy = y x1 x2 x3
. matrix Xy = Xy'
. matrix list Xy, nohalf
y
x1  -703.48823
x2  -5634.6157
x3  -182.33711
_cons  -98.065355

. matrix beta = Xi * Xy
. matrix list beta, nohalf
y
x1  -18.230615
x2   .79002516
x3    32.77203
_cons   10.545653

* Note that the vector beta is identical to the "coef" column in the regress
* output table
*  This computes the estimated variance matrix - discussed later.
. matrix VCM = 8.99280976 * Xi
. matrix l VCM, nohalf
x1          x2          x3       _cons
x1   40.801248  -2.1717031  -72.929747  -30.640194
x2  -2.1717031   .11896076   3.7461336   1.7060627
x3  -72.929747   3.7461336   137.67943   48.930773
_cons  -30.640194   1.7060627   48.930773   29.829873
```

Ex.2. Substantive Example

Does a person's regligious background relate to their sexual behavior? Or, more to the point, was Billy Joel right when he sang "Catholic girls start much too late?". We can actually study this question "scientifically" using the NSLY used in one or more homework assignments. Here is an annotated log file for the analysis, including data cleaning, summary statistics, and estimates.

```
. keep relig fage* sex
. *  to focus on Billy's refrain, look only at young women
. drop if sex==1
(6403 observations deleted)
```

```
. *  missing values of relig coded as missing
. drop if relig<0
(20 observations deleted)

. *  religion is coded into 10 categories, will collapse to 4
. tab relig
IN WHAT|
RELIGION WAS|
R RAISED -|      Freq.     Percent        Cum.
------------+-----------------------------------
0 |        221        3.53        3.53
1 |        261        4.17        7.70
2 |       1796       28.68       36.37
3 |         95        1.52       37.89
4 |        320        5.11       43.00
5 |        523        8.35       51.35
6 |        167        2.67       54.02
7 |       2159       34.47       88.49
8 |         58        0.93       89.41
9 |        663       10.59      100.00
------------+-----------------------------------
Total |       6263      100.00

. gen none = cond(relig==0,1,0)
. gen prot = cond(relig<7&relig>0,1,0)
. gen cath = cond(relig==7,1,0)

. *  Apologize for grouping everyone else into one bundle
. gen oth = cond(relig>7,1,0)

. * Hey, why not have a recoded religion variable (4 values)?
. gen relig2 = none + 2*prot + 3*cath + 4*oth
. tab relig2
relig2|      Freq.     Percent        Cum.
------------+-----------------------------------
1 |        221        3.53        3.53
2 |       3162       50.49       54.02
3 |       2159       34.47       88.49
4 |        721       11.51      100.00
------------+-----------------------------------
Total |       6263      100.00

. * Who is raised a 2?  We need labels!!
. label define religions 1 "none" 2 "protest" 3 "catholic" 4 "non_chrst"
. label values relig2 religions
. tab relig2
relig2|      Freq.     Percent        Cum.
------------+-----------------------------------
none |        221        3.53        3.53
protest |       3162       50.49       54.02
catholic |       2159       34.47       88.49
non_chrs |        721       11.51      100.00
------------+-----------------------------------
Total |       6263      100.00
```

```
. descr fage*
3. fage83       byte   8.0g                F - AGE @FIRST SEXUAL INTERCOUR
4. fage84       byte   8.0g                F - AGE FIRST SEXUAL INTERCOURS
5. fage85       byte   8.0g                F AGE 1ST HAD SEXUAL INTERCOURS

. *  first the person is asked if they have ever had
. *  sexual intercourse. IF YES for the first time, then the person is
. *  asked age at first intercourse.  IF NO or YES in earlier year then
. *  fage is coded as negative.
. gen byte age = fage83 if fage83>0
(1259 missing values generated)

. * this adds people who said yes for first time in 1984
. replace age = fage84 if fage84>0
(2522 real changes made)

. * now 1985
. replace age = fage85 if fage85>0
(245 real changes made)

. * The variable age is now the age of first intercourse of
. * all people who have had intercourse by 1985
. tab age
age|      Freq.     Percent        Cum.
------------+----------------------------------
2 |          1        0.02        0.02
3 |          1        0.02        0.04
8 |          3        0.05        0.09
9 |          4        0.07        0.16
10 |         11        0.20        0.36
11 |         11        0.20        0.55
12 |         29        0.52        1.07
13 |         95        1.70        2.77
14 |        236        4.22        6.99
15 |        521        9.31       16.30
16 |       1020       18.23       34.54
17 |       1059       18.93       53.47
18 |       1134       20.27       73.74
19 |        620       11.08       84.82
20 |        370        6.61       91.44
21 |        231        4.13       95.57
22 |        124        2.22       97.78
23 |         62        1.11       98.89
24 |         32        0.57       99.46
25 |         18        0.32       99.79
26 |          9        0.16       99.95
27 |          3        0.05      100.00
------------+----------------------------------
Total |       5594      100.00

. regress age none cath oth
Source |       SS       df       MS                  Number of obs =    5594
---------+------------------------------               F(  3,  5590) =   31.10
Model |  460.623661     3   153.54122               Prob > F      =  0.0000
```

```
Residual |  27598.6357  5590  4.93714414              R-squared     =  0.0164
---------+------------------------------              Adj R-squared =  0.0159
Total |  28059.2594  5593   5.0168531              Root MSE      =   2.222

------------------------------------------------------------------------------
age |      Coef.   Std. Err.       t     P>|t|       [95 Conf. Interval]
---------+--------------------------------------------------------------------
none |  -.4101829    .160575     -2.554   0.011      -.7249723   -.0953935
cath |     .566983   .0659714      8.594   0.000       .4376534    .6963127
oth |   .3732208   .0980449      3.807   0.000       .1810148    .5654268
_cons |    17.2053    .0412396    417.204   0.000       17.12446    17.28615
------------------------------------------------------------------------------

. * significant t values indicate that that age
. * differs for that religion significantly compared to
. * protestant christians
. * If you ever want to see hatVar(beta_hat), use the matrix "get(VCE)" command
. matrix  eVbhat = get(VCE)
. matrix l eVbhat

symmetric eVbhat[4,4]
none       cath        oth      _cons
none  .02578433
cath   .0017007  .00435223
oth   .0017007   .0017007  .00961279
_cons  -.0017007  -.0017007  -.0017007   .0017007

. * Why don't we include dummy variable for each religion?
. regress age none cath oth prot
Source |       SS       df       MS                  Number of obs =    5594
---------+------------------------------              F(  3,  5590) =   31.10
Model |  460.623661     3   153.54122              Prob > F      =  0.0000
Residual |  27598.6357  5590  4.93714414              R-squared     =  0.0164
---------+------------------------------              Adj R-squared =  0.0159
Total |  28059.2594  5593   5.0168531              Root MSE      =   2.222

------------------------------------------------------------------------------
age |      Coef.   Std. Err.       t     P>|t|       [95 Conf. Interval]
---------+--------------------------------------------------------------------
none |  -.7834037   .1788735     -4.380   0.000      -1.134065   -.4327422
cath |    .1937622   .1027795      1.885   0.059      -.0077254    .3952499
oth |  (dropped)
prot |  -.3732208   .0980449     -3.807   0.000      -.5654268   -.1810148
_cons |   17.57853    .0889499    197.623   0.000       17.40415     17.7529
------------------------------------------------------------------------------

. * Answer:  because X'X is not invertible!!!!
```

I.5 Random Variables and Matrix Notation

To derive the statistical properties of OLS estimates, we must define the notion of a random vector and the notions of the mean and variance of a random vector.

Definitions

Def D.1. Random matrix

A random matrix is a matrix whose elements are random variables.

For example the vector of disturbance terms in (I.11) is a random vector since each element of the vector is a random variable. A constant matrix is, of course, a special type of random matrix in the same sense that a constant is a special case of a random variable that does not vary across points in the sample space.

Def D.2. Expectation of a random matrix V $(m \times r)$.

The expectation of a random matrix is the matrix of expectations of each random variable in the random vector. That is, let V be a $m \times r$ random matrix, and let v_{ij} denote the random variable in the i^{th} row and j^{th} column of V. Then

$$E\left[V\right] \equiv \begin{vmatrix} E\left[v_{11}\right] & E\left[v_{12}\right] & \dots & E\left[v_{1r}\right] \\ E\left[v_{21}\right] & E\left[v_{22}\right] & \dots & E\left[v_{2r}\right] \\ \vdots & \vdots & \vdots & vdots \\ E\left[v_{m1}\right] & E\left[v_{m2}\right] & \dots & E\left[v_{mr}\right] \end{vmatrix} \qquad (I.21)$$

Expectation is still a linear operator when defined this way. That is, if A is a $n \times m$ matrix of constants and C is a $n \times r$ vector of constants, then

$$E\left[AV + C\right] = AE\left[V\right] + C \qquad (I.22)$$

m does not have to equal r. That is, the transformation matrix A may change the dimensions of the random matrix. AV is a $n \times r$ random matrix, not a $m \times r$. The proof of this result requires you to write out what an element of the matrix $AV + C$ looks like, apply the expectation operator to it, and then see that the result is exactly the same as if you had written out $AE\left[V\right] + C$.

Def D.3. Variance of a random vector Let v be $m \times 1$.

Then $Var(v) \equiv E\left[(v - E[v])(v - E[v])'\right]$.

This generalizes the notion of covariance, which itself generalized the notion of variance. Notice that $(v - E[v])(v - E[v])'$ is a $m \times m$ matrix (when v is $m \times 1$). So $Var(v)$ is always a square matrix. If we multiply (v-E[v])(v-E[v])' out, we can see that the ith row, jth column of it contains $(v_i - E[v_i])(v_j - E[v_j])$. The expectation of this scalar random variable is by definition:

$$E[(v_i - E[v_i])(v_j - E[v_j])] = Cov(v_i, v_j) \tag{I.23}$$

Each element of $Var(v)$ is a covariance. The diagonal elements of $Var(v)$ are simply the variances of the corresponding elements of the v vector. Since $Cov(r,s) = Cov(s,r)$, the matrix $Var(v)$ is **symmetric**. Summing this up,

$$Var(v) = \begin{pmatrix} Var(v_1) & Cov(v_1,v_2) & Cov(v_1,v_3) & \ldots & Cov(v_1,v_m) \\ Cov(v_1,v_2) & Var(v_2) & Cov(v_2,v_3) & \ldots & Cov(v_2,v_m) \\ Cov(v_1,v_3) & Cov(v_2,v_3) & Var(v_3) & \ldots & Cov(v_3,v_m) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Cov(v_1,v_m) & Cov(v_2,v_m) & Cov(v_3,v_m) & \ldots & Var(v_m) \end{pmatrix} \tag{I.24}$$

Def D.4. Variance of a linear transformation

Let v be $m \times 1$ random vector and A a $n \times m$ matrix of constants. Then

$$Var[Av] = E\left[(Av - AE[v])(Av - AE[v])'\right] = E\left[A(v - E[v])(v - E[v])'A'\right] = AVar[v]A' \tag{I.25}$$

This result generalizes the simpler result that when multiplying a random variable by a constant it changes the variance by the square of the constant. In fact, if $m = n = 1$, then we simply get back the old formula $Var(av) = a^2 Var(v)$.

Here's an example. Let $v = \begin{pmatrix} v_w 1 \\ v_2 \end{pmatrix}$ with

$$Var[v] = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \tag{I.26}$$

This means that v contains two random variables that have zero covariance. (They may be statistically independent of each other as well, since independence implies zero covariance, but not necessarily). Let $A = (2 - 2)$. Then

$$w = Av = 2v_1 - 2v_2 \tag{I.27}$$

is the random variable that is simply twice the difference of the two random variables in the vector v. From the result above,

$$\begin{aligned} Var[Av] =& AVar[v]A' \\ =& \begin{pmatrix} 2 & -2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \begin{pmatrix} 2 \\ -2 \end{pmatrix} \\ =& \begin{pmatrix} 2\sigma_1^2 - 2\sigma_2^2 \end{pmatrix} \begin{pmatrix} 2 \\ -2 \end{pmatrix} \\ = 4\sigma_1^2 + 4\sigma_2^2 \end{aligned} \tag{I.28}$$

This is exactly the formula we would have used without the matrix notation, since the weighted sum of zero-covariance random variables is the sum of the variances multiplied by the square of the weights. Keeping track of variances and covariances in matrices is very convenient.

I.6 Assumptions of the LRM in Matrix Notation

Remaining Assumptions

A.2. The PRE is correct

A.3. The matrix X is fixed (non-random) across samples.

A.4. Mean 0 errors: $E[u] = 0$, where 0 is a $N \times 1$ vector containing 0s.

A.5. $Var[u] = \sigma^2 I$

In words: the variance matrix for u is diagonal (0 covariance across observations) and has equal elements on the diagonal (equal variance).

A.6. $u \sim N(0, \sigma^2 I)$

In words: The u vector is a **normal** vector, in the sense that each element of u is normally distributed.

Notice that the only change between simple and multiple regression is the change in the dimension of X. So if we had started out with matrix notation we would have been able to use these forms of the assumptions from the start. That is, **A4** above is the same as the **A4**, just using different notation. The same is true of **A3** through **A7**, the only difference being that **A5** and **A6** can stated together.

Now we can start analyzing the statistical properties of OLS estimate $\hat{\beta}^{OLS}$ and $\hat{\sigma}^2$.

I.7 Statistical Properties of OLS Estimates

Properties Under A0-A4

Prop.1. $E[\hat{\beta}] = \beta$ (OLS is unbiased)

Properties Under A0-A6

Prop.1. $Var[\hat{\beta}] = \sigma^2(X'X)^{-1}$

Prop.2. $\hat{\beta}$ is BLUE

Properties Under A0-A7

Prop.1. $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$

Prop.2. $\hat{\sigma}^2 \sim \chi^2_{N-k}$

Prop.3. $\hat{\sigma}^2$ is statistically independent of $\hat{\beta}$

I.8 Formulating and Testing Multiple Hypotheses

The null hypothesis $H_0 : \beta_2 = 0$ places one restriction on the possible values of the population parameters. But we might like to test several hypotheses together. For example,

$$\begin{aligned} H_0 : &\beta_2 = 0 \\ &\beta_3 = 0 \\ &\beta_4 = 0 \end{aligned}$$

This null hypothesis places **three simultaneous** restriction on the population parameters. One might think to test each of the restrictions with a t test and then somehow aggregate the results of those t tests into a decision about all three restrictions. But remember that a classical hypothesis test requires that α, the probability of the **Type I error**, must be controlled for. That means we must be able to determine the probability of rejecting H_0 when it is true. But each of the single t tests ignores the other restrictions, so setting α for each test separately doesn't determine the α for some combination of them. Without control over α it is not possible to determine how strongly the data support or don't support the multiple hypothesis.

We need a **single test statistic** for a multiple hypothesis. We will restrict attention to *linear hypotheses*, which are linear restrictions on the values contained in the parameter vector β. A linear restriction on β takes the form

$$R\beta = c$$

where R is a $r \times k$ matrix of constants and c is a $r \times 1$ vector of constants. For example, if $k = 5$ then the joint hypothesis above could be written:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Logical Steps of a F-tests for multiple hypotheses

Step.1. Compute RSS for the unrestricted model (without imposing H_0). Denote this RSS_{UR}.

Step.2. Impose H_0 on the OLS estimates, run the resulting regression, and compute RSS. Denote this RSS_R.

That is when minimizing the sum of squared residuals (e'e), we will force the estimates $\hat{\beta}$ to satisfy $R\hat{\beta} = c$. Adding a restriction to the estimates reduces their ability to choose values for $\hat{\beta}$ to fit the data. It must be the case that $RSS_R \geq RSS_{UR}$.

Step.3. Consider two mutually exclusive states of the world:

A. H_0 *is true*

In this world the true parameters actually satisfy the restriction being imposed on the estimated parameters in the restricted model. Imposing a true restriction should not make very much difference in the results. That is, imposing a true restriction should not change the amount of variation in Y attributed to the residual u. RSS_{UR} would not be very different from RSS_R if H_0 is true. The only reason why they would differ at all if H_0 is true is because allowing $\hat{\beta} \neq R$ allows the estimates to pick up sampling variation in the relationship between X and Y. On the other hand,

B. H_0 *is false,* H_A *is true.*

Imposing a false restriction on the estimated parameters should affect OLS abilitly to fit the Y observations. That is, if $R\beta \neq c$ then this should be would be reflected in $R\hat{\beta} \neq c$ and the RSS in the restricted and unrestricted models would be much different.

Step.4. Based on comparing these two possible states, a large difference in the restricted and unrestricted RSS would therefore be evidence against H_0.

The ratio $(RSS_R - RSS_{UR})/RSS_{UR}$ tells us the proportionate change in the amount of variation attributed to variance in u when the restriction is imposed. In fact, **under** H_0, the difference RSS_R-RSS_{UR}, which is a random variable, is distributed as a chi-squared random variable with r **degrees of freedom** (the number of linear restrictions), and it is distributed *independently* of the unrestricted RSS when H_0 is true. The intuition for this result is simply that the change in the RSS is due to explaining the sampling variation if the restriction is indeed true. And the sampling variation is independent of the model (by assumption u is drawn independently of everything else going on in the LRM).

Step.5. Because the numerator and denominator are independent under H_0, the expected value of the ratio is the ratio of expected values.

The mean of a χ^2 random variable equals its degrees of freedom. Therefore, the ratio given above would have a mean equal to $r/(N-k)$ under H_O. It makes sense to normalize the ratio so that is has a mean of 1 under H_0, which leads to **test statistics for linear hypotheses**

$$F = \frac{\frac{RSS_R - RSS_{UR}}{r}}{\frac{RSS_{UR}}{N-k}} \tag{I.29}$$

Under H_0, F follows the $F(r, N-k)$ distribution. Therefore, we should reject H_0 if $F \geq F_\alpha^\star(r, N-k)$. The alternative is simply that the restriction is not true:

$$H_A : R\beta \neq c$$

This is inherently a *two-sided* test, because we are not specifying which equations in H_A has a greater-than sign and which equations have a less-than sign.

Three Special Cases of F Tests

Case.1. $r = 1$

When the restriction matrix R contains one row ($r = 1$), then the test statistic is $F(1, N-k)$. A chi-square random variable with one degree of freedom is simply a standard normal Z variable squared. So $F(1, N-k)$ can be written $Z^2/C/(N-k)$, where C is a χ^2 with $N-k$ degrees of freedom. But this expression is exactly the square of a t variable with $N-k$ degrees of freedom. Hence a t-test is a special case of the more general F-test.

Case.2. $H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$

This particular hypothesis (with $r = k-1$ restrictions) is called the **test for overall significance**. How would you interpret this hypothesis? Under this H_0, the linear regression model would reduce to $Y = \beta_1 + u$. No X variables show up in the PRE. So this hypothesis is the hypothesis that *none of the X variables are linearly statistically related to Y*. The alternative is that at least one of the X variables is linearly related to Y, although which ones that might be is not specified.

Notice that the restricted model simply includes a constant. The OLS estimate of β_1 would in the restricted model be simply $\hat{\beta}_1 = \bar{Y}$. The RSS_R would then simply be TSS! The restricted unexplained variation is simply the total variation in Y around its sample mean. Recall that $TSS = ESS + RSS$, so RSS_R - RSS_{UR} equals $TSS - RSS_R$, which equals ESS_{UR}. So the test statistic for a test of overall significance reduces to

$$F = \frac{\frac{ESS}{k-1}}{\frac{RSS_{UR}}{N-k}} \qquad (I.30)$$

which is simply the ratio of the two MS (mean-squared) entries in the *analysis of variance table*. The F statistic reported in the Stata output table is exactly this number.¡br¿

Case.3. Test for overall significance and $k = 2$

We can go even further here. If we have a simple LRM ($k = 2$), then the test for overall significance collapses to $H_0 : \beta_2 = 0$, which we know can also be tested with the t test $\hat{\beta}/\hat{se}(\hat{\beta}_2)$. But the F version of the test is $F(1, N-k)$, which we know from example 1 is the square of a t test. And, indeed, the overall F statistic is exactly equal to the the square of the t statistic for $\hat{\beta}_2$ in a simple LRM. You should confirm this fact.

I.8.a Example of specifying a joint hypothesis test

We return to the example started earlier concerning sexual behavior and religious background.

```
. * Let's test whether expected age of first intercourse
. * differs
significantly by religious background . test none cath oth

( 1)  none = 0.0
( 2)  cath = 0.0
( 3)  oth = 0.0

F(  3,  5590) =   31.10
Prob > F =    0.0000
. * Conclusion:  We can reject the hypothesis that
. *   expected age of first intercourse is the same for
. *   all religious backgrounds
.   *
. * Let’s perform the F test directly rather than using "test"
. *  The regress above is the Unrestricted Model
. qui regress
. di "Unrestricted RSS = " _result(4) Unrestricted RSS = 27598.636
. local u =_result(4) . *  The Restricted Model imposes H0 on the estimates, which
. *  in this case means none cath oth all have zero coefficients
. *  So the Unrestricted model is
. regress age

Source |       SS       df       MS                  Number of obs =    5594
---------+------------------------------               F(  0,  5593) =       .
Model |        0.00     0           .                Prob > F      =       .
Residual |  28059.2594  5593   5.0168531               R-squared     =  0.0000
---------+------------------------------               Adj R-squared =  0.0000
Total |  28059.2594  5593   5.0168531                Root MSE      =  2.2398

------------------------------------------------------------------------------
age |      Coef.   Std. Err.       t     P>|t|       [95 Conf. Interval]
---------+--------------------------------------------------------------------
_cons |   17.42063   .0299471    581.714   0.000       17.36192    17.47934
------------------------------------------------------------------------------

. di "The Restricted RSS = "_result(4) The Restricted RSS = 28059.259
. local r = _result(4)
. di "So the F statistic for the test is " ((‘r’-‘u’)/3)/(‘u’/(5594-4)) So the F statistic
for the test is 31.099197
. * Notice that this is the F statistic reported by test, and
. * since this is a test of overall significance it is also
. * equal to the F statistic reported in the regression output
```

I.9 Prediction and Analysis of Variance in the MLRM

I.9.a Prediction

The OLS prediction for a particular Y or for the expected value of Y conditional on values of X follows the same logic as in the simple LRM. Now, however, we want to predict for a $1 \times k$ vector of X values, which would describe one observation. Let X_0 be the vector you want to predict for. Then following the earlier discussion the mean and individual predictions are the same,

$$\hat{E}[Y_0|X_0] = X_0\hat{\beta} \qquad (I.31)$$

$$\hat{Y}_0|X_0 = X_0\hat{\beta} \qquad (I.32)$$

But the variances in our predictions vary:

$$Var(\hat{E}[Y_0]) = X_0{}'Var(\hat{\beta})X_0 = \sigma^2 X_O{}'(X'X)^{-1}X_O \tag{I.33}$$

$$Var(\hat{Y}_0) = X_0{}'Var(\hat{\beta})X_0 + \sigma^2 \tag{I.34}$$

which must be estimated by

$$\hat{Var}(\hat{E}[Y_0]) = \hat{\sigma}^2 X_0{}'(X'X)^{-1}X_0 \tag{I.35}$$

$$\hat{Var}(\hat{Y}_0) = \hat{\sigma}^2(X_0{}'(X'X)^{-1}X_0X_0 + 1) \tag{I.36}$$

Notice the sharp increase in the amount of computation required to compute the estimated variance of a prediction as we move from k=2 to the general LRM. Now, to compute a confidence interval for a prediction after the estimation process itself requires a $1 \times k$ vector multiplied by a $k \times k$ matrix, the result ($1 \times k$ vector) then multiplied by a $k \times 1$. For $k = 10$, this would take hours by hand. It is very important then to learn how to use Stata's predict command and/or Stata's matrix algebra facilities.

```
.
. * Let's compute 95  confidence intervals for expected
. * age for each religion - first, what is the critical t value?
. di invt(5590,.95)
1.9603884
. * By the way:  notice that with 5590 degrees of freedom t=z
. predict agehat
. * grab estimated standard error of mean prediction
. predict ahse, stdp
. * here's one way to compute lower and upper bound
. gen  alo = agehat - ahse * invt(5590,.95)
. gen  ahi = agehat + ahse * invt(5590,.95)
. sort relig2
. by relig2: summ alo agehat ahi

-> relig2=     none
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
alo |     221    16.49089          0   16.49089   16.49089
agehat |     221    16.79512          0   16.79512   16.79512
ahi |     221    17.09935          0   17.09935   17.09935

-> relig2=  protest
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
alo |    3162    17.12446   4.17e-06   17.12446   17.12446
```

```
agehat |    3162    17.20531   4.70e-06   17.20531   17.20531
ahi |    3162    17.28615   4.73e-06   17.28615   17.28615

-> relig2= catholic
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
alo |    2159    17.67134          0   17.67134   17.67134
agehat |    2159    17.77229   3.80e-06   17.77229   17.77229
ahi |    2159    17.87323          0   17.87323   17.87323

-> relig2= non_chrs
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
alo |     721    17.40415          0   17.40415   17.40415
agehat |     721    17.57853          0   17.57853   17.57853
ahi |     721     17.7529          0    17.7529    17.7529

. *CONCLUSION:  On average, Catholic girls start .58 years to 1.38 years
. * later, 19 times out of 20.
```

I.9.b Analysis of Variance

The analysis of variance interpretation of OLS works exactly the same way in the MRLM as the Simple LRM. In particular, OLS estimates decompose the total variation in Y around the sample mean into two components: explained variation and unexplained or residual variation. So $TSS = RSS + ESS$. The only differences is that RSS, ESS, and TSS must be extended to matrix notation.

$$RSS = y'(I - X(X'X)^{-1}X')y \tag{I.37}$$

$$M \equiv \text{NxN matrix of ones} \tag{I.38}$$

$$I - \frac{1}{N}M = (I - \frac{1}{N}M)'(I - \frac{1}{N}M) \tag{I.39}$$

$$TSS = y'(I - \frac{1}{N}M)y \tag{I.40}$$

$$ESS = y'(X(X'X)^{-1}X' - \frac{1}{N}M)'(X(X'X)^{-1}X' - \frac{1}{N}M)y \tag{I.41}$$

$$R^2 \equiv \frac{ESS}{TSS} \tag{I.42}$$

I.10 Violation of MLRM Assumptions

Violations of A0-A7

A.0. Sample

If the data do not contain all the data then one cannot compute all of the OLS estimates, so violation of this assumption is fatal to estimating the full model.

A.1. PRE

There are many reasons to think about how the form of the Population Regression Equation may be wrong. Perhaps one is not sure that the PRE should be linear, log-linear, or some other transformation of the original data. Even worse, perhaps the relation between Y and X can't be linearized at all. That is one way to think about the - as extensions of the OLS model in which the relationship between Y and X is a specific one that cannot fit into the LRM framework.

A.2. PRE generates the sample

If the PRE did not generate the sample information then some other process did. So violation of these assumptions is in many respects equivalent to the violation of A1.

A.3. Mean zero error

As mentioned in the original discussion of the assumptions, A3 is really a technical assumption. If $E[u]$ is not zero then one can re-define β_1 to include the mean of u.

A.4. No Correlation between errors and exogenous variables

Recall that under A0-A4 OLS estimates are unbiased estimates of the LRM parameters. Violating Assumptions A0-A3 leads to either trivial changes (in the case of A0 and A3) or fairly esoteric statistical issues of what model seems to be generating the data. I call this esoteric because it tends to lead one away from the **economic content** of the LRM, unless one looks for other specifications that come from economic theory. However, most analyses of violations of A1 and A2 are not based on economic theory.

Violation of A4/A4* is another matter altogether. If $cov(u, X)$ is not 0 (the A4* way of seeing it), or if one cannot think of X as "exogenous" to Y (the A4 way), then one is in trouble. And this trouble has a lot to do with economic theory.

A.5. Violation of scalar variance matrix

Traditionally one would go from this point to study ways to "fix" OLS estimates when A5 and A6 do not hold. But it is important that these assumptions only enter at the point of deriving $Var(\hat{\beta})$. This means estimated standard errors and confidence intervals won't be good if A5 and/or A6 is not true. However, this is in many ways a much smaller problem than having either a non-linear model or not even having unbiased estimates. Therefore, our discussion of these assumptions is brief.

A.6. Violation of normal errors

Since OLS estimates are linear in the Y, violation of the normality assumption is really only a problem in small samples. That's because for values of N larger than 30 the Central Limit Theorem begins to kick in. That is $\hat{\beta}$ will be asymptotically normally distributed even if u is not normally distributed.

II. LIMITED DEPENDENT VARIABLES

II.1 Introduction

We have been studying the linear regression model, which in matrix form can be written:

$$Y = X\beta + u \tag{II.1}$$

This model is both quite general and fairly simple to estimate. It is general because various transformations of the data (such as taking logarithms or including squares of variables) allows (II.1) to model much more complex relationships between X and Y than its linear form might suggest. It is simple to estimate because the OLS estimates of β are linear and have good statistical properties under reasonable assumptions. Linearity also implies that OLS estimates can be computed with hundreds of variables in X and tens-of-thousands of observations.

One property of a *linear* relationship which is often annoying is that it will goes everywhere from $-\infty$ to $+\infty$ except in the case when the slope equals 0. For many types of variables this property is desirable or at least innocuous. But for many variables the domain of the relationship is inherently limited to something less than $(-\infty, +\infty)$. In general these are called *limited dependent variables*.

We already are familiar with the most restrictive case of a limited-dependent variable: a binary (0-1) random variable. However, we have been careful to make such variables the endogenous (Y) variable in a linear regression.

Types of LDV:

Typ.1. binary:

Typ.2. ordered:

Typ.3. multinomial:

Typ.4. censored:

Typ.5. truncated:

II.2 The Linear Probability Model

We will now consider how OLS performs when Y is simply Recall that we have been putting discrete variables in the X vector since the first day of class when we coded marital status as a 0 or 1 variable. However, when the discrete variable is the left-hand side variable, some difficulties arise.

There are many examples in economics when we want to explain the outcome of a binary random variable. To return to an earlier example, we might want to build a model to explain who becomes married. (Recall that our earlier analysis used marital status to explain something else.) We might want to explain who works and who doesn't, who belongs to a union and who doesn't, who drives to work and who rides the bus, who cheats on their taxes and who doesn't, which countries are net debtors and which aren't, etc. There are limitless examples outside of economics as well: which patients respond to a new treatment and which don't, which countries have capital punishment and which don't, which bridges collapse in earthquakes and which don't, etc.

Why do discrete variables on the left-hand side of (II.1) cause problems? The most obvious difficulty is that u can not be normally distributed. When u is normally distributed, Y is a continuous random variable. We haven't let this stop us before. For instance, whether you work or not is a 0/1 outcome. However, a variable that counts occurrences can be thought of as approximately continuous. But when the possible outcomes collapse to only two or three in number, and these outcomes are *qualitative* (e.g. married or not) not *quantitative*, then the normality assumption may be a problem.

Example 1.

Consider data where

$$Y = \begin{cases} 0 & \text{if person } i \text{ is unmarried} \\ 1 & \text{if person } i \text{ is married} \end{cases} \qquad (II.2)$$

Suppose we estimate the model:

$$Y = \beta_1 + \beta_2 AGE + u. \tag{$II.3$}$$

Here is some Stata output using data from the NLSY concerning marital status and age:

```
. summm
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
age |    2613    23.24914   2.990209         20         26
married |    2613    .3647149   .4814423          0          1

. regress married age
Source |       SS       df       MS                Number of obs =    2613
---------+------------------------------             F(  1,  2611) =  312.60
Model |  64.7337249     1  64.7337249             Prob > F      =  0.0000
Residual |  540.692988  2611  .207082722             R-square      =  0.1069
---------+------------------------------             Adj R-square  =  0.1066
Total |  605.426713  2612  .231786643             Root MSE      =  .45506

------------------------------------------------------------------------------
married |      Coef.   Std. Err.       t     P>|t|       [95 Conf.Interval]
---------+--------------------------------------------------------------------
age |   .0526474   .0029777     17.680   0.000       .0468085   .0584864
_cons |  -.8592924   .0697994    -12.311   0.000      -.9961602  -.7224246
------------------------------------------------------------------------------
```

Notice that AGE is positively related to marital status: older people are more likely to be married than younger people. That shouldn't be too suprising. The coefficient on age tells us more than that. It also tells at what rate the probability of being married changes with age. Now notice:

```
. predict marhat
. tab marhat age

| AGE OF R AT INTERVIEW DATE
marhat|        20         26 |     Total
-----------+----------------------+----------
.1936561 |      1198          0 |      1198
.5095406 |         0       1415 |      1415
-----------+----------------------+----------
Total|      1198       1415 |      2613
```

The predicted values from the regression take on values that Y can't possibly take on. How do we interpret a predicted marital status of 20 year-olds equal to 0.193? Are twenty year-olds 20 percent married on average? We can interpret the predicted values from a regression on a binary value as the predicted probability that Y takes on the value 1. That is, OLS estimates of (II.3) predict that the probability of 20-year-olds being married (Y=1) as 0.193. In other words:

$$\hat{P}(Y = 1|X = 20) = \hat{E}(Y|X = 20) = \hat{\beta}_1 + \hat{\beta}_2 20. \tag{$II.4$}$$

This is called, for the obvious reason, the Linear Probability Model (LPM), but that is just a fancy name for OLS when the explained variable is binary. In the LPM it is impossible to guarantee that the predicted values actually fall between 0 and 1, which probabilities must. This is a problem for the LPM model, and it arises because a linear regression model isn't well-suited for binary values.

II.3 Another Approach: Binary Response Model

Let's go back and see if we can't come up with a better model of variables such as marital status, who wins a hockey game, and so forth. In fact, let's start with the following statement: some people are more married than other people. Some people are so married that a small change in their lives wouldn't shift them to being single. Meanwhile, some people are just over the threshold (no pun intended). If not for the final beer last night they might not be married today. In the same way, some people are more single than other people.

In other words, the binary variable Y captures differences in some measured status, in this case being married or not, but underlying that status is a continuous quantity. Let $Y^\star$ be an index of how married a person is. By definition, a person is married if their value of $Y^\star$, unobserved to us, is big enough to push them into being marriage. Figure 1 illustrates the idea. The index $Y^\star$ is defined so that $Y^\star$ is negative for single persons, leading to $Y = 0$. People whose value of $Y^\star$ is greater than 0 are married, so $Y = 1$. The observed binary random variable Y is related to unobserved continuous random variable $Y^\star$ by the rule:

$$Y = \begin{cases} 0 & \text{if } Y^\star \leq 0 \\ 1 & \text{if } Y^\star > 0. \end{cases} \tag{$II.5$}$$

$Y^\star$ is a convenient construction that captures the idea that underneath the black-and-white world of binary variables often lie shades of gray. In some ways $Y^\star$ is like utility in microeconomic theory. (In fact, latent variables underneath binary choices like marriage or whether or not to ride the bus to work can be interpreted as the utility the individual places on the choice.) Utility indicates the rank of a consumption bundle to a consumer, and the magnitude of utility numbers in an economic model have no meaning except in ranking bundles. The magnitude of $Y^\star$ only matters in deciding whether the person is above or below the cut-off. We can therefore set the threshold for being married equal to 0 without loss of generality. We also can't estimate the variance of $Y^\star$, because it is an unobserved quantity, often called a **latent variable**. That is why we assume u has a fixed variance, unlike the OLS model which estimates the variance of u.

Let's now build a model for $Y^\star$. That is, let's assume that it is statistically related to other observed characteristics of the individual contained in a vector X:

$$Y^\star = X\beta + u \qquad (II.6)$$

In other words, we can think of $Y^\star$ as the variable for which the linear regression model is appropriate. For instance, the older a person is, the more likely they are to be married. In that case, the coefficient on age, say β_3, is greater than 0. Our goal, as in the regression model, is to estimate β and thereby estimate the effect of variables on the *probability* of being married. We can assume that u, and therefore $Y^\star$, is a continuous random variable. The difference between (II.3) and (II.6) is that $Y^\star$ is not observed by us, we only observe Y. We can not estimate (II.6) with OLS because we do not observe the left-hand side variable. We have built a continuous model underneath the binary random variable Y.

Equations (II.5) and (II.6) together form two-thirds of a **binary response model**. The final third describes the distribution of the disturbance term u Let the distribution of u be denoted $F(u)$. That is, $F(u)$ equals the probability that the realization of the random variable is less than or equal to u. Since Y is binary random variable, conditional on X, its whole distribution is summarized by the probability that $Y = 0$. Using (II.5) and (II.6),

$$\begin{aligned} P(Y=0|X) =& P(X\beta + u < 0|X) = P(u < -X\beta) \\ =& F(-X\beta). \end{aligned} \tag{II.7}$$

And,

$$P(Y=1|X) = 1 - P(Y=0|X) = 1 - F(-X\beta). \tag{II.8}$$

Figure 2 illustrates these formulas. The cumulative distribution function for u is graphed. The cutoff for being married is still $Y^\star = 0$, but this implies that a person is married when u is greater than $-X\beta$. The characteristics in X shift the threshold, and therefore, the probability of observing values of Y. Note the difference between this model and the Linear Probability Model (LPM). The LPM, or OLS on a binary variable, assumes implicitly that the probability is linear in $X\beta$, whereas the model we have developed assumes that the latent variable underneath Y is linear in $X\beta$.

II.4 Definition of The Logit and Probit Models

How do we estimate β in (II.6)? We have to complete our model by making an assumption about the distribution of u. We begin with the easiest assumption possible. Namely,

$$F(u) = \frac{e^u}{1+e^u}. \tag{II.9}$$

This is called the **logistic** distribution. It is straightforward to show the following properties:

Properties of the Logistic Distribution

P.1. $1 - F(u) = \frac{1}{1+e^u}$

P.2. $F(\infty) = 1$

P.3. $F(0) = 0.5$

P.4. $f(u) = \frac{e^u}{(1+e^u)^2}$

P.5. $f(u) = f(-u)$ (symmetric distribution around 0)

P.6. $E[u] = 0$

P.7. $Var[u] = \pi^{2/3}$

Since we do not observe either $Y^{\star}$ or u, we can not estimate the variance of u, so we use a distribution with fixed variance unlike the OLS model. Changing the variance of u would not affect the probability of $Y = 0$ or $Y = 1$. The density and distribution function for the logistic distribution are illustrated in Figure 3. The logistic density is similar in shape to the normal distribution, but the logistic distribution has fatter tails than the normal distribution with the same variance.

The reason why assuming u is logistic is the simplest model is because the probabilities of observing values of Y have closed-forms in terms of X and β. Assuming u is logistic, we can write the probability of $Y = 1$ and $Y = 0$ conditional on X as:

$$P(Y = 0|X; \beta) = F(-X\beta) = \frac{e^{-X\beta}}{1 + e^{-X\beta}} \qquad (II.10)$$

$$P(Y = 1|X; \beta) = 1 - F(-X\beta) = \frac{1}{1 + e^{-X\beta}}. \qquad (II.11)$$

A convenient way to express these two probabilities in one formula is

$$P(Y|X; \beta) = \left(e^{-X\beta}\right)^{1-Y} \frac{1}{1 + e^{-X\beta}}. \qquad (II.12)$$

Notice that when $Y = 1$, the first term equals 1 and when $Y = 0$ the first term equals $e^{-X\beta}$, so (II.12) captures both probabilities. Notice that in (II.12) the probability of observing the endogenous variable Y depends on both the vector of true parameters β and the value of the exogenous vector X. The notation $P(Y = 0|X; \beta)$ emphasizes that β helps determine the conditional probability.

By assuming that u follows the logistic distribution we have now completed our model for Y and X:

II.4.a Logit Binary Response Model

$$
\begin{aligned}
Y &= \begin{cases} 0 & \text{if } Y^\star < 0 \\ 1 & \text{if } Y^\star \geq 0 \end{cases} \\
Y^\star &= X\beta + u \qquad\qquad (II.13) \\
F(u) &= \frac{e^u}{1+e^u}
\end{aligned}
$$

A very similar model is the case when u is assumed to follow the standard normal distribution. That model is called the **probit** binary response model or simply a probit. In the *probit* model,

$$
F(u) = \Phi(u) = \int_{-\infty}^{u} (2\pi)^{-1/2} e^{-0.5x^2} dx. \qquad (II.14)
$$

where Φ is the standard normal (Z-variable) cumulative distribution function. The logit and probit models tend to give very similar results, but the probit model is a little more difficult to explain because the necessary probabilities do not have closed-form solutions like the logit model. (Recall that there is no closed form expression for Φ which is why we use tables of normal probability values.)

How do we interpret the coefficients in the β vector in the logit model?

How do we estimate the logit model using a sample of data? Let Y_i and X_i equal the values for the ith observation, $i = 1, 2, \ldots, N$. There are methods that make clever use of OLS estimates to estimate the logit model, so if you only had the ability to compute OLS estimates, you could estimate the logit model. However, most modern statistical programs like Stata will estimate the logit and probit models the right way. It takes more computing time, but computers are so fast these days that you might not even notice the extra time. For this reason we will not discuss methods based on OLS.

II.5 Summary

Highlights

Pt.1. When Y is a binary random variable, the linear regression modeled is re-christened the Linear Probability Model (LPM).

Pt.2. The LPM has several unattractive features, including the fact that the predicted probabilities are not guaranteed to lie between 0 and 1.

Pt.3. In the logit model, the observed binary variable Y is linked to a continuous latent (unobserved) random variable $Y^\star$ whose value determines the outcome of Y. $Y^\star$ is assumed to follow the logistic distribution. The probit model is the same, except $Y^\star$ is assumed to follow the normal distribution.

Pt.4. The logistic distribution is similar to the normal distribution, but it leads to simpler expressions for $P(Y = 0)$ and $P(Y = 1)$.

III. Intro to Maximum Likelihood Estimation

III.1 Introduction

The best method for estimating the parameters β is the method of **maximum likelihood (ML)**. ML is a non-linear estimation procedure. That is, the ML estimates $\hat{\beta}$ are *not* linear functions of Y. Like OLS, ML are chosen so as to optimize a function. As you know, OLS esimates *minimize* the sum of the residuals squared for the observations in the sample. ML estimates, on the other hand, *maximize* the likelihood function for the observations in the sample. The likelihood function is the probability of observing the values in the sample if the estimates $\hat{\beta}$ were the true parameters β. In (8.3), we have already written down the probability of observing a particular data point give the true parameters β.

Assuming that the observations in the sample are independent, the likelihood function is the **product** of the likelihood for each observation. This is exactly the same idea behind the fact that the probability of seeing two heads on two independent coin tosses equals the product of observing a head on each separate toss. Here, the probability of seeing person 1 married **and** person 2 unmarried equals the probability that person 1 is married **times** the probability that person 2 is unmarried. For a binary variable model the likelihood function is written:

$$\mathcal{L}(\hat{\beta}) = \prod_{i=1}^{N} P(Y = Y_i | X_i; \hat{\beta}). \tag{III.1}$$

The symbol $\prod$ is the product symbol just as $\sum$ is the summation symbol. We do not know the true value of β, so we do not know the true probabilities of seeing the values of Y_i. If we did know the true value of β there would be no reason to estimate the model. In the likelihood function we replace the unknown true values with estimated values. It is convenient to work with the log of $\mathcal{L}$:

$$\ln \mathcal{L}(\hat{\beta}) = \sum_{i=1}^{N} \ln P(Y = Y_i | X_i, \hat{\beta}). \tag{III.2}$$

If we maximize the log-likelihood it is the same as maximizing the likelihood, because logarithm is a monotonic transformation of the objective function. Now use the assumption that u is logistic to substitute an explicit formula for P:

$$\ln \mathcal{L}(\hat{\beta}) = \sum_{i=1}^{N} (1 - Y_i)(-X_i\hat{\beta}) - \ln(1 + e^{-X_i\hat{\beta}}). \qquad (III.3)$$

ML estimates of β are the values of $\hat{\beta}$ that maximize this objective function given the sample information (Y_i, X_i).

Example 2.

Consider the small data set below where N=7 and there is only one explanatory variable in the X matrix:

i	X	Y
-	---	---
1	66	0
2	70	1
3	69	0
4	80	1
5	68	0
6	67	1
7	72	0

We use a very small sample so that we can write out the *complete* likelihood function without using the summation sign. For this sample, the logit likelihood function for the sample would be

$$\begin{aligned}
\ln \mathcal{L}(\hat{\beta}) = &- (\hat{\beta}_1 + \hat{\beta}_2 66) - \ln(1 + e^{-(\hat{\beta}_1+\hat{\beta}_2 66})+ \\
&- \ln(1 + e^{-(\hat{\beta}_1+\hat{\beta}_2 70)}) \\
&- (\hat{\beta}_1 + \hat{\beta}_2 69) - \ln(1 + e^{-(\hat{\beta}_1+\hat{\beta}_2 69)}) \\
&- \ln(1 + e^{-(\hat{\beta}_1+\hat{\beta}_2 80)}) \\
&- (\hat{\beta}_1 + \hat{\beta}_2 68) - \ln(1 + e^{-(\hat{\beta}_1+\hat{\beta}_2 68)}) \\
&- \ln(1 + e^{-(\hat{\beta}_1+\hat{\beta}_2 67)}) \\
&- (\hat{\beta}_1 + \hat{\beta}_2 72) - \ln(1 + e^{-(\hat{\beta}_1+\hat{\beta}_2 72)}). \qquad (III.4)
\end{aligned}$$

The ML estimates are the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ that maximize this function. There are only two parameters, β_1 and β_2. If we had more data the form of the likelihood function would be the same: we would simply add the additional observations to the likelihood function.

Since our goal is to maximize $\ln \mathcal{L}$ by choosing $\hat{\beta}_1$ and $\hat{\beta}_2$, you should be tempted to take derivatives of $\ln \mathcal{L}$ with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$, set them to 0, and solve. This is indeed correct. The problem is that you won't get closed-form solutions as in with OLS estimates. That is, the ML estimates of the logit model can't be written in a formula like $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$. (Try it if you don't believe me. This is not always the case for ML. For the linear regression model, the ML estimates of β under assumption A.6 are exactly the same as the OLS estimates. See, for example, Gujarati for the derivation.) To maximize $\ln \mathcal{L}$ for the logit model requires numerical methods. We do not have time to discuss the algorithm that Stata or other programs use to find the values of $\hat{\beta}$ to maximize $\ln \mathcal{L}$. (You might have learned about Newton's method to maximize functions numerically in calculus class. Stata uses a related algorithm.)

Example 2 (Continued)

Here is Stata output for the ML logit estimates of the small data above

```
. logit y x

Iteration 0:  Log Likelihood =-4.7803567
Iteration 1:  Log Likelihood =-4.1779391
Iteration 2:  Log Likelihood =-4.1676834
Iteration 3:  Log Likelihood =-4.1676145

Logit Estimates                                         Number of obs =      7
chi2(1)       =   1.23
Prob > chi2   = 0.2683
Log Likelihood = -4.1676145                             Pseudo R2     = 0.1282

------------------------------------------------------------------------------
y |      Coef.   Std. Err.       t     P>|t|       [95 Conf. Interval]
---------+--------------------------------------------------------------------
x |   .2206555   .2349038      0.939   0.391      -.3831838    .8244948
_cons |  -15.77234   16.41216     -0.961   0.381      -57.96114    26.41647
------------------------------------------------------------------------------
```

The iterations indicate that Stata searched numerically for the ML estimates and took 4 rounds to find the ML estimates. Notice on each round the value of the likelihood falls until estimates are found for which the likelihood can not be improved. We can directly check that we and Stata agree on the form of the likelihood function, by replacing the ML estimates reported in the table into equation (III.3):

$$\begin{aligned}\ln \mathcal{L}(\hat{\beta}_{ML}) = &- (-15.77 + 0.22 \times 66) - \ln(1 + e^{-(-15.77+0.22\times 66)}) \\ &- \ln(1 + e^{-(-15.77+0.22\times 70}) \\ &- (-15.77 + 0.22 \times 69) - \ln(1 + e^{-(-15.77+0.22\times 69)}) \\ &- \ln(1 + e^{-(-15.77+0.22\times 80}) \\ &- (-15.77 + 0.22 \times 68) - \ln(1 + e^{-(-15.77+0.22\times 68)}) \\ &- \ln(1 + e^{-(-15.77+0.22\times 67}) \\ &- (-15.77 + 0.22 \times 72) - \ln(1 + e^{-(-15.77+0.22\times 72)}) \\ = &- 4.1676145. \end{aligned} \tag{$III.5$}$$

To check the last line it's easiest to use Stata:

```
. gen mylk = -(1-y)*(-15.77+0.2270*x) - ln(1+exp(-(-15.77+0.2270*x)))
. summ mylk

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
mylk |       7   -.5953735   .4052748  -1.304808  -.1419979

. display 7*(-.5953735)
-4.1676145
```

The χ^2 statistic 1.23 equals twice the difference between the final likelihood value and the likelihood value in iteration 0, $1.23 = 2 \times (-4.1676145 - (-4.7803567))$. Why this value should be reported and why it is chi-squared distributed is discussed below. The pseudo-R2 value is analogous to the R-squared derived in the linear regression model, and researchers in some disciplines rely on it as a measure of fit in logit or probit models. However, it doesn't have quite the same interpretation or definition as R-square. We will ignore it.

The rest of the table is exactly the same as in the regression output. The only difference is that the Std.Err column is based on a different formula (discussed below). But given that, the t-statistic and confidence intervals are based on the same formulas as in regression model. This means you can use the standard errors and t-ratios reported by Stata for the logit model in the same way as the regression model to test hypotheses about β and to form confidence intervals. In this case, the t-statistic on X fails to reject the null hypothesis that $\beta_2 = 0$.

Here are the results of estimating the probit model on the same data:

```
. probit y x

Probit Estimates                                        Number of obs =      7
chi2(1)       =   1.25
Prob > chi2   = 0.2633
Log Likelihood = -4.1545828                             Pseudo R2     = 0.1309

------------------------------------------------------------------------------
y |      Coef.   Std. Err.       z     P>|z|       [95 Conf. Interval]
---------+--------------------------------------------------------------------
x |   .1397627   .1418225      0.985   0.324      -.1382042    .4177296
_cons |  -9.975761     9.8966     -1.008   0.313      -29.37274    9.421219
------------------------------------------------------------------------------
```

Notice that the estimated coefficients and estimated standard errors are a bit different than those computed in the logit model. However the likelihood value, t ratios and chi-squared statistics are very similar. The difference is caused by the different shapes of the logistic and normal distribution functions. But because the two distributions are pretty similar, the logit and probit models usually give very similar results.

III.2 Statistical Properties of ML Estimates

It can be shown that the ML estimates of β, in fact any ML estimates in a very general class of statistical models, are consistent and distributed asymptotically normally. Unlike their OLS counterparts in the regression model, ML estimates are not guaranteed to be unbiased nor are they normally distributed in small samples. However, with large N, the bias goes away and the estimates become normally distributed. This means our usual t-distributions for testing hypotheses and forming confidence intervals are the same. (ML estimates also have minimum variance among all consistent estimates if the model is correctly specified, so they are also **efficient** estimates.)

The t-ratio for β_j is the familiar

$$t = \frac{\hat{\beta}_j - \beta_j}{\hat{se}(\hat{\beta}_j)},$$

where $\hat{se}(\hat{\beta}_j)$ is determined from a different formula. We will rely on Stata to compute it correctly, but some discussion of it is worthwhile. Recall that the variance-covariance matrix (VCM) for $\hat{\beta}$ contains the variances and covariances between our estimated values. The standard error of a parameter is simply the square root of the corresponding diagonal element in the VCM. Why would ML estimates vary over samples? Because the data will be different across samples and the resulting maximizing values of $\hat{\beta}$ will change. Those estimates are generated from first order conditions on a maximization problem, which involve the *first* derivatives of $\ln \mathcal{L}$. The change in $\hat{\beta}$ is therefore related to how much the first derivative of $\ln \mathcal{L}$ changes across samples. It is the *second* derivative $\ln \mathcal{L}$ which indicates the change in the second derivative. So you shouldn't be surprised that $\text{Var}(\hat{\beta})$ is related to the *Hessian* of $\ln \mathcal{L}$, that is

$$\frac{\partial^2 \ln \mathcal{L}}{\partial\hat{\beta}\partial\hat{\beta}'}.$$

In fact, it turns out that

$$\hat{\text{Var}}(\hat{\beta}) = -\frac{\partial^2 \ln \mathcal{L}}{\partial\hat{\beta}\partial\hat{\beta}'}^{-1}.$$

Why this is the right formula goes *well* beyond the scope of this class. Stata does compute this matrix, however, and reports the square-root of its diagonal elements as the estimated standard errors of $\hat{\beta}$. (By analogy, recall that OLS estimates solve the normal equation $(X'X)^{-1}\hat{\beta} = X'y$, and that $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$. The matrix $(X'X)^{-1}$ *is* the Hessian for the objective function in OLS estimation (RSS=e'e), so without realizing it, you have already seen that the inverse of a Hessian matrix determines the variance matrix for a vector of estimates.

In ML estimation, F-tests of linear restrictions can be used as in the linear regression model. Stata's **test** will work as usual after you run a logit or probit. But a better test in some ways is the *likelihood ratio test*, which relies on the chi-squared distribution. The likelihood ratio test statistic is the ln of the ratio of the likelihoods in the restricted and unrestricted model. For the test of overall significance, the chi-squared test has $k - 1$ degrees of freedom, where k is the number of estimated coefficients. That is, suppose you have the unrestricted model

$$Y^{\star} = X\beta + u \qquad (III.6)$$

and you want to test a restriction on the true parameters β, which can be written:

$$H_0: \quad f(\beta) = 0.$$

Here $f(\beta)$ is a function that states the restriction. For instance, if the hypothesis is $\beta_6 = 5$, then $f(\beta) = \beta_6 - 5$. Unlike the F-test we used earlier, the restriction does not have to be linear in β, although linear restrictions are much easier to understand and to implement.

The procedure for testing H_0 is similar in the ML case: Estimate both the restricted and unrestricted versions of the model. If the fit of the models is very different, then that gives us evidence that the restriction is not true (because if it were true the restricted and unrestricted estimates would only differ because of sampling variation in the data). In the F test we used the RSS (Residual Sum of Squares) to summarize the fit of the two models. In the logit case we don't really have a RSS, because the model generates a prediction for $Y^{\star}$ which is unobserved. Instead, the maximal value of the likelihood function itself is used to summarize the fit of the model. A larger likelihood value means the model is making the observed data more "probable." The likelihood must therefore be larger in the unrestricted model than in the restricted model. The bigger the difference the more evidence there is that the restricted model is not true and H_0 is false.

So, as usual, hypothesis testing comes down to knowing how big is big so that the probability of a type I error can be determined and controlled. This requires a test statistic that can be computed from the data and for which we know the distribution under H_0. In ML estimation, there are three test statistics commonly used. The one we will use is the likelihood ratio test. Let $\mathcal{L}_{UR}$ be the log-likelihood in the UnRestricted model and $\mathcal{L}_R$ be the log-likelihood in the Restricted model. Then the test statistic is

$$LR = -2(\mathcal{L}_R - \mathcal{L}_{UR}). \tag{III.7}$$

Notice that $\mathcal{L}_R \leq L_{UR}$, so LR must be greater than or equal to zero. (The log of a ratio is the difference in the logs, so LR is related to the likelihood ratio. Hence its name.) The larger LR is, the more the restricted model disagrees with the unrestricted one. So the key is to know what the distribution of LR is under H0. The answer:

> Under H_0, LR $\sim \chi^2_m$, where m is the number of restrictions that H_0 imposes on the unrestricted model.

To perform a likelihood-ratio test, do the following:

Step1. Specify model and H_0. Determine the number of restrictions m, and choose level of significance α.

Step2. Run both restricted and unrestricted models, record final log-likelihood values for each model.

Step3. Compute LR defined in (III.7), i.e. twice the absolute difference of the log-likelihoods.

Step4. Look up critical value of χ^2_m using the for your level of significance. (In Stata, use the **chiprob(m,LR)** function to look up the probability to the right of the calculated LR statistic and compare that to your signficance level α). Reject H_0 if LR is greater than the critical value or (equivalently) if chiprob(m,LR) $<$ α. (Note: as with the F-test, likelihood ratio tests are by their nature always *one-sided* .)

The Stata command **lrtest** will automate steps 3 and 4 for you. However, unlike

test, it will not automate the computation of the restricted and unrestricted models. So you have to run two separate logits (or probits) to compute the LR statistic.

What is the chi-square statistic reported in the output table? Stata begins iterating on the likelihood function by setting $\hat{\beta}_2 = 0$ and computing the optimal value of $\hat{\beta}_1$. So the first value of the likelihood reported in the output is the restricted model under the hypothesis that $\beta_2 = 0$. This is exactly the same hypothesis implicitly tested by the F-statistic reported in the linear regression table. In other words, it is a test of **overall significance** of the X variables. In example, the overall chi-squared statistic has m=1 degree of freedom because $k = 2$. The value 1.23 in the example is not significant at the 10 percent level, because the probability of a chi-square random variable with one degree of freedom being greater than 1.23 equals 0.26 (from the output). Therefore, we fail to reject the hypothesis that $Y^\star$ does not depend upon X. This agrees with the t statistic on $\hat{\beta}_2$.

III.3 Prediction

How do we use our ML estimates of β to make predictions? When we predict in the logit model, we are predicting the probability that Y is either 0 or 1 conditional on some vector X_0. For instance,

$$\hat{P}_{ML}(Y = 1|X_0) = \frac{1}{1 + e^{-X_0 \hat{\beta}_{ML}}}. \qquad (III.8)$$

We can use this prediction in several ways. We could say that we predict Y=1 when $\hat{P}(Y = 1|X_0) > 0.5$. In other words, one way to predict Y directly is to predict that it takes on the more likely outcome given X_0:

$$\hat{Y}|X_0 = \begin{cases} 0 & \text{if } \hat{P}(Y = 1|X_0) < 0.5 \\ 1 & \text{if } \hat{P}(Y = 1|X_0) \geq 0.5 \end{cases} \qquad (III.9)$$

When you use **predict** in Stata, it computes $\hat{P}(Y = 1|X)$ for each observation. This is analogous to the how one would interpret OLS predictions (i.e. the LPM). In the LPM, $\hat{Y}_0|X_0$ is interpreted as the predicted probability:

$$\hat{P}_{OLS}(Y = 1|X_0) = X_0 \hat{\beta}_{OLS} \qquad (III.10)$$

where $\hat{\beta}_{OLS}$ equals the OLS estimates from the LPM in (II.3).

Let's see how the predictions compare between the logit model and the OLS (LPM) model. The code below starts after running the logit command described above (so that the first predict command uses the logit model):

```
. predict yhat1
. regress y x

Source |       SS       df       MS                Number of obs =       7
---------+------------------------------              F(  1,     5) =    0.99
Model |  .282808198     1  .282808198             Prob > F      =  0.3659
Residual |  1.43147752     5  .286295503              R-square      =  0.1650
---------+------------------------------              Adj R-square  = -0.0020
Total |  1.71428571     6  .285714286             Root MSE      =  .53507

------------------------------------------------------------------------------
y |      Coef.   Std. Err.       t     P>|t|       [95 Conf. Interval]
---------+--------------------------------------------------------------------
x |   .0460385   .0463215      0.994   0.366      -.0730347    .1651118
_cons |  -2.807281   3.262017     -0.861   0.429      -11.19256       5.578
------------------------------------------------------------------------------

. predict yhat2
. list yhat1 yhat2 y

yhat1      yhat2          y
1.  .2298651   .2312634         0
2.  .4191044   .4154176         1
3.  .3665359    .369379         0
4.   .867623    .875803         1
5.  .3169634   .3233405         0
6.  .2712247   .2773019         1
7.  .5286834   .5074946         0
```

yhat1 contains $\hat{P}_{ML}(Y = 1|X)$ while yhat2 contains $\hat{P}_{OLS}(Y = 1|X)$. So ML logit says that the probability that y=1 for observation 1 is 0.23, implying that the probability that y=0 is 0.77. (You should check for yourself how Stata came up with .23.) The LPM gives the same prediction, to two decimal places. In fact, the two predictions are pretty close to each other. Notice that if we follow the two definitions of $\hat{Y}_0$, both models fail to "predict" Y correctly for observations 2,6 and 7. In both models, $Y = 0$ is more likely for $i = 2$, but the actual value was 1. Likewise for observations 6 and 7.

Also notice that $\hat{\beta}^{ML}$ and $\hat{\beta}^{OLS}$ are quite different even though the predicted probabilities are quite similar. This isn't as surprising as you might think. Recall that the logit and LPM are different statistical models for the same data. The parameters β have different interpretations in the two models even if we choose to use the same name. And notice that the t-statistic for $\hat{\beta}_2$ is similar in the two models, because $\beta_2 = 0$ in both models implies no statistical relationship between X and Y.

The potential problems with LPM do not show up in this example within the sample. Namely, the predicted probabilities all fall between 0 and 1. You might check, however, that $\hat{Y}|X < 0$ in the LPM whenever $X < 61$. In this data, X is air temperature in degrees Fahrenheit. So for reasonable temperatures, the LPM provides nonsense predictions for reasonable values of the independent variable.

III.4 Summary

Highlights

Pt.1. Estimates based on the method of Maximum Likelihood (ML) are consistent, asymptotically normally distributed, and efficient in the logit and probit models as well as many other statistical models. ML estimates are generally not unbiased in small samples.

Pt.2. In ML estimated standard errors follow different formulas than the OLS estimated standard errors, but they are used in exactly the same way to test hypotheses and to form confidence intervals about coefficients.

Pt.3. The F-test for overall significance with OLS estimates is replaced with a chi-squared test with ML estimates based on the log ratio of (maximal) likelihoods in the restricted and unrestricted models.

Pt.4. The predicted probabilities and conclusions from hypothesis tests are usually similar in the LPM, logit and probit models.