PPOL 503-03, PPOL 503-04, Fall 2016
Course Notes #5 Logit Models; Maximum Likelihood Estimation

I. LOGIT MODELS FOR BINARY DEPENDENT VARIABLES

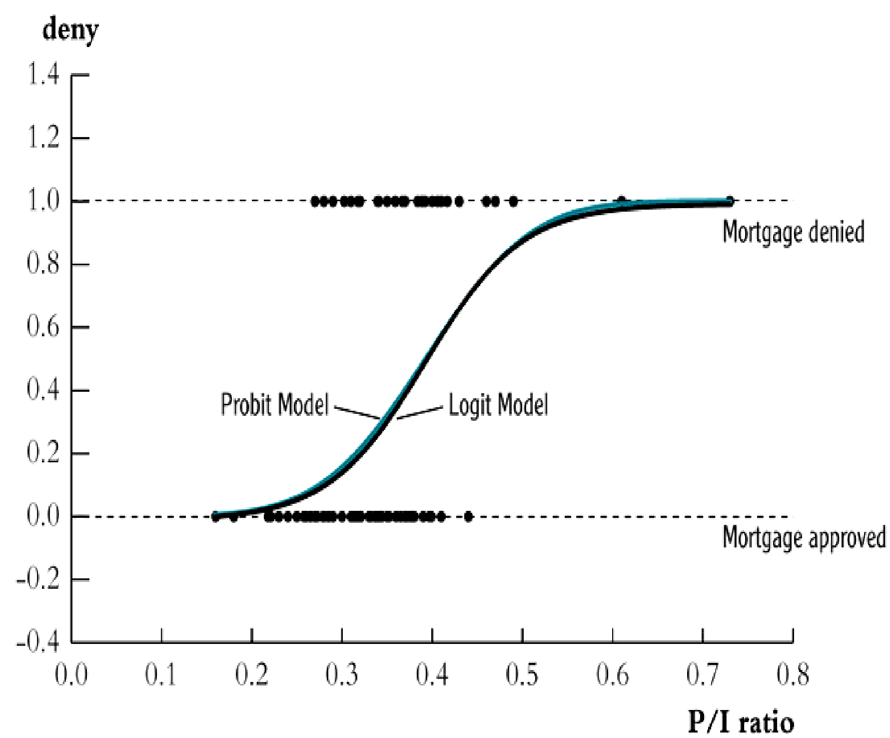
- Instead of assuming a linear relationship between the independent variables and *Y*, we can assume a functional form in which the probability of *Y* (i.e., the expected value of *Y* given **X**) approaches zero at a slower and slower rate as each *X* decreases and approaches one at a slower and slower rate as each *X* increases. (Therefore, unlike the LPM, in such a model the impact of a unit increase of *X* on the probability of *Y* is not constant.)

- The cumulative distribution function (CDF) of a random variable meets these criteria.

- The typical choice is to either use the CDF of the normal distribution (probit) or the CDF of the logistic distribution (logit). There is no definitive reason to prefer one to the other.

FIGURE 9.3 **Probit and Logit Models of the Probability of Denial, Given the P/I Ratio**

These logit and probit models produce nearly identical estimates of the probability that a mortgage application will be denied, given the payment-to-income ratio.



- **Logit Model:**
    - $P\ (Y{=}1|\mathbf{X}) = \dfrac{1}{1+e^{-(\beta_0+\beta_1 X_1+\Lambda+\beta_k X_k)}}$
- **Probit Model:**
    - $P(Y=1|\mathbf{X}) = \dfrac{1}{\sqrt{2\pi}\int_{-\infty}^{z} e^{-s^2/2} ds}$, where z = $\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$.
- Don't let these specifications scare you. Essentially, all we've done is change the functional form so that the dependent variable is restricted to predicted values between 0 and 1. You'll notice that for each of these equations, as $\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$ goes towards infinity, $P(Y{=}1|\mathbf{X})$ goes towards 1. Also, as $\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$ goes towards negative infinity, $P(Y{=}1|\mathbf{X})$ goes towards 0.

II. Maximum Likelihood Estimation

- The problem, however, is that we cannot use OLS to estimate these equations because they are not linear in their parameters. Instead, we rely on *maximum likelihood estimation* (MLE), which seems complicated but is easily executed by all statistical packages.

- MLE chooses coefficient estimates that maximize the likelihood of the sample data set being observed. The technique yields consistent and efficient estimates for large samples. MLE has asymptotic (large sample) properties: unbiased, efficient, normal distribution. Asymptotic means the estimates get better or more precise as the sample gets larger. The parameter estimates are the most precise when the value of the likelihood is maximized.

    - Assume that Y=1 N_1 times and that Y=0 N_2 times (N_1+N_2=N). Also, assume that the data are ordered so that the Y=1 observations come first.
    - The goal of MLE is to maximize the probability of the sample data set being observed. That is, MLE maximizes L = Prob(Y_1,…,Y_N). This is the same as maximizing L=Prob(Y_1) x Prob(Y_2) x … x Prob (Y_N).
    - Given the way we ordered the data above, this is the same as maximizing:
        - $L = P_1 \Lambda \ P_{N_1}(1-P_{N_1+1})\Lambda \ (1-P_N)$
    - Taking logs of both sides yields the objective of maximizing the following:
        - $\log L = \sum_{i=1}^{N_1} \log P_i + \sum_{i=N_1+1}^{N} \log(1-P_i)$
    - You then plug in the appropriate equation for the P_i's, depending on whether you are estimating a logit or probit model. MLE finds $\hat{\beta}_0$, $\hat{\beta}_1$, … $\hat{\beta}_k$ to maximize log *L* above.

- Although the likelihood function is linear in X, the probabilities are not. With logit and probit models the probabilities increase non-linearly with X. With the LPM, the probabilities increase linearly with X.

- One of the difficulties with probit and logit models is with interpreting the coefficient estimates. Given the highly non-linear nature of these models, the impact of a marginal change in X on a change in the Probability of Y is difficult to determine. Also, since the models are nonlinear, the marginal impact will vary with different levels of X. (Note that the advantage of LPM is that the coefficient estimates are easily interpreted as the change in probability of Y given a marginal change in X, and this slope is constant throughout.)

- The logit and probit coefficient estimates will tell you the sign of the impact and the standard errors will help you determine if the impacts are statistically significant. The logit and probit coefficients do not indicate the magnitude of the effects of changing an independent variable.

III. INTERPRETING RESULTS FROM LOGIT ESTIMATION

$$P_i = E(Y = 1 \mid X) = \beta_0 + \beta_1 X_1$$

Where X = payment to income ratio and Y = 1 if the mortgage application is denied

$$P_i = E(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1)}}$$

Multiply by both sides by $1 + e^{-Z}$

so $(1 + e^{-Z})\, P_i = 1$

Divide both sides by P_i and then subtract 1 to get

$e^{-Z} = 1/P_i - 1 = 1 - P_i \,/\, P_i$

since $e^{-Z} = 1/e^{Z}$ it follows that

$1/e^{Z} = 1 - P_i \,/\, P_i$

Inverting we get $e^Z = P_i / 1 - P_i$

P_i is the probability the mortgage application is denied

1 - P_i is the probability the mortgage application is not denied

Take the log of both sides yields

$\log (P_i / 1 - P_i) = Z_i = \beta_0 + \beta_1 X_1 = L_i$

L_i -- the log of the odds-ratio is linear in X and parameters.
This is the logistic distribution function. *Define* $Z_i = \beta_0 + \beta_1 X$.
Z_i ranges from $-\infty$ to $+\infty$. P_i ranges from 0 to 1. P_i is non-linearly related to Z_i.

1. The logit coefficient show how a change in an X affects the probability that the outcome will occur; logit coefficients indicate the direction but not the magnitude of effect on the probability of changing an independent variable. The most convenient way to test the significance of each individual X variables is with the Z score (coefficient divided by the standard error.)

2. Marginal effects and/or odds ratios can be calculated to demonstrate the magnitude of the effect of an independent variable on the probability of interest. The magnitude depends on the steepness of the cumulative distribution function at that point. When the probability is high or low, it is difficult to change, since it is bounded by zero or one. When the probability is moderate, the response is greatest mathematically. Marginal effect of $X_i = P(1 - P) \times \beta_i$

3. Odds Ratios: defined as the probability of an event P_i relative to the probability of the non-event 1 - P_i. If the logit coefficient for X_i is negative, the odds ratio is less than one. If the logit coefficient on X_i is positive, then the odds ratio is greater than 1. If the logit coefficient is close to zero, then the odds ratio is close to 1.

4. To calculate the odds ratio for a dummy explanatory variable such as gender = 1 if male; = 0 if female; one should exponentiate the coefficient (take the inverse log).

5. For continuous variables, the odds ratio shows how a one unit change in X (age in years) affects the probability the event occurs. Frequently, a one unit change in X is not very meaningful. Suppose one wanted to know the probability of death for a person age 80 as opposed to some one age 70, where the mean age of the sample is 70. One would first exponentiate the logit coefficient to calculate the odds ratio for age. Next, one would take the odds ratio for age and raise it to the power " t". In this case "t" is the difference in age between the sample mean and the age of interest. For this example, "t" equals 10.

IV. TESTING RESTRICTIONS AND MODEL GOODNESS OF FIT

1. F-tests do not apply for logit (probit) models. Instead, we use the *likelihood ratio statistic*, which closely corresponds to the F-statistic we used with OLS. The *LR* test compares the log-likelihood function of the unrestricted model (the model with all the variables included) to the log-likelihood function of the restricted model (the model with all – or a subset of – the variables assumed to have no effect on the dependent variable). Because MLE maximizes the log-likelihood function, dropping variables generally leads to a smaller log-likelihood. The question is whether the fall in the log-likelihood is large enough to conclude that the dropped variables are important. The LR test is a way of testing whether the difference is significant.

2. The likelihood ratio test is: $2(\text{Loglikelihood}_{UR} – \text{Loglikelihood}_{R})$ which is distributed as a chi-square. See above example. In Stata, as you did with OLS, you can use the "test" command after the logit or probit command to test different restrictions.

3. In computing the *LR* statistic for binary outcome models, it is important to know that the log-likelihood function will always be a negative number. This fact follows because Y is either zero or one and both variables inside the log function are strictly between zero and one, which means their natural logs are negative. The fact that both log-likelihood functions are negative does not change the way we compute the LR statistic, so we preserve the negative signs in computing the statistic. The multiplication by 2 is needed so the LR statistic has an approximate Chi-square distribution under the null hypothesis.

4. Goodness of Fit: In OLS an F statistic (with k degrees of freedom in the numerator and N – k – 1 degrees of freedom in the denominator) can be used to test the joint hypothesis H_0 that all the coefficients except the intercept are equal to zero. The corresponding test for the logit and probit models is the likelihood ratio

principle. This statistic is distributed as a Chi-square when the null hypothesis is true.

4. There is no statistic similar to the adjusted R^2 for either logit or probit models. Other options:
a) degree of concordance: % of cases correctly classified;
b) Pseudo $R^2 = 1 - LR_{ur} / LR_0$ where LR_{ur} is the log-likelihood function for the estimated model and LR_0 is the log-likelihood function for the model with just the intercept. If the covariates have no explanatory power then the ratio $LR_{ur} / LR_0 = 1$ and the pseudo-R^2 is zero. Usually the absolute value of LR_{ur} < the absolute value of L_0 in which case $1 - LR_{ur} / LR_0 > 0$.
Advantages: easy to compute; ranges between 0 and 1; approaches 0 as the quality of the fit diminishes and approaches 1 as model fit improves.
Disadvantages: does not penalize for increasing the number of exogenous variables.

V. EXAMPLE: A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, affect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

This data set has a binary response (outcome, dependent) variable called **admit**. There are three predictor variables: **gre**, **gpa** and **rank**. We will treat the variables **gre** and **gpa** as continuous. The variable **rank** takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest.

```
summarize gre gpa

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         gre |       400       587.7    115.5165        220        800
         gpa |       400      3.3899    .3805668       2.26          4
```

```
tab rank

       rank |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         61       15.25       15.25
          2 |        151       37.75       53.00
          3 |        121       30.25       83.25
          4 |         67       16.75      100.00
------------+-----------------------------------
      Total |        400      100.00
```

```
tab admit

      admit |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        273       68.25       68.25
          1 |        127       31.75      100.00
------------+-----------------------------------
      Total |        400      100.00
```

```
tab admit rank

           |                    rank
     admit |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         0 |        28         97         93         55 |       273
         1 |        33         54         28         12 |       127
-----------+--------------------------------------------+----------
     Total |        61        151        121         67 |       400
```

.

Using the logit model

Below we use the **logit** command to estimate a logistic regression model. The **i.** before **rank** indicates that **rank** is a factor variable (i.e., categorical variable), and that it should be included in the model as a series of indicator variables. Note that this syntax was introduced in Stata 11.

```
logit admit gre gpa i.rank

Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -229.66446
Iteration 2:   log likelihood = -229.25955
Iteration 3:   log likelihood = -229.25875
Iteration 4:   log likelihood = -229.25875

Logistic regression                               Number of obs   =        400
                                                  LR chi2(5)      =      41.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -229.25875                       Pseudo R2       =     0.0829

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   .0022644    .001094     2.07   0.038     .0001202    .0044086
         gpa |   .8040377   .3318193     2.42   0.015     .1536838    1.454392
             |
        rank |
          2  |  -.6754429   .3164897    -2.13   0.033    -1.295751   -.0551346
          3  |  -1.340204   .3453064    -3.88   0.000    -2.016992   -.6634158
          4  |  -1.551464   .4178316    -3.71   0.000    -2.370399   -.7325287
             |
        cons |  -3.989979   1.139951    -3.50   0.000    -6.224242   -1.755717
```

- In the output above, we first see the iteration log, indicating how quickly the model converged. The log likelihood (-229.25875) can be used in comparisons of nested models, but we won't show an example of that here.
- Also at the top of the output we see that all 400 observations in our data set were used in the analysis (fewer observations would have been used if any of our variables had missing values).
- The likelihood ratio chi-square of 41.46 with a p-value of 0.0001 tells us that our model as a whole fits significantly better than an empty model (i.e., a model with no predictors).
- In the table we see the coefficients, their standard errors, the z-statistic, associated p-values, and the 95% confidence interval of the coefficients. Both **gre** and **gpa** are statistically significant, as are the three indicator variables for **rank**. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

We can test for an overall effect of **rank** using the **test** command. Below we see that the overall effect of **rank** is statistically significant.

```
test 2.rank 3.rank 4.rank

 ( 1)  [admit]2.rank = 0
 ( 2)  [admit]3.rank = 0
 ( 3)  [admit]4.rank = 0

           chi2(  3) =   20.90
         Prob > chi2 =    0.0001
```

We can also test additional hypotheses about the differences in the coefficients for different levels of rank. Below we test that the coefficient for **rank**=2 is equal to the coefficient for **rank**=3. (Note that if we wanted to estimate this difference, we could do so using the **lincom** command.)

```
test 2.rank = 3.rank

 ( 1)  [admit]2.rank - [admit]3.rank = 0

           chi2(  1) =    5.51
         Prob > chi2 =    0.0190
```

You can also exponentiate the coefficients and interpret them as odds-ratios. Stata will do this computation for you if you use the **or** option, illustrated below. You could also use the **logistic** command.

logit , or

```
Logistic regression                               Number of obs   =        400
                                                  LR chi2(5)      =      41.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -229.25875                       Pseudo R2       =     0.0829

------------------------------------------------------------------------------
       admit | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf.Interval]
-------------+----------------------------------------------------------------
         gre |   1.002267   .0010965     2.07   0.038      1.00012   1.004418
         gpa |   2.234545   .7414652     2.42   0.015     1.166122   4.281877
             |
        rank |
          2  |   .5089309   .1610714    -2.13   0.033     .2736922   .9463578
          3  |   .2617923   .0903986    -3.88   0.000     .1330551   .5150889
          4  |   .2119375   .0885542    -3.71   0.000     .0934435   .4806919
------------------------------------------------------------------------------
```

GRE: OR = 1.002. The mean GRE score is 588. Suppose you want to know how the probability of being admitted to graduate school increases if one has a GRE score of 700. Take the OR and raise it to the power 112. $(1.002)^{112}$ = 1.25, which means that having a GRE score of 700 as opposed to 588 means a student is 1.25 times as likely or 25% more likely to be admitted to graduate school.

GPA: OR = 2.23. For each one unit increase in GPA, the odds of being admitted to graduate school (versus not being admitted) increase by a factor of 2.23. Alternatively stated, for each one unit increase in GPA, a student is 123% more likely to be admitted to graduate school.

Rank 2 versus Rank 1: A person who attends an undergraduate institution of rank 2 is 51% as likely (or 49% less likely) to be admitted to graduate school as someone who attended an undergraduate institution of rank 1.

Rank 3 versus Rank 1: A person who attends an undergraduate institution of rank 3 is 26% as likely (or 74% less likely) to be admitted to graduate school as someone who attended an undergraduate institution of rank 1.

Rank 4 versus Rank 1: A person who attends an undergraduate institution of rank 4 is 21% as likely (or 79% less likely) to be admitted to graduate school as someone who attended an undergraduate institution of rank 1.

Marginal Effects

Marginal effects for each individual X can be calculated as: P (1- P) x β_i

Rank 2 versus Rank 1 (the omitted group):
(.318) x (.682) x $\beta_{RANK\ 2}$ = .217 x (-.675) = -.146 or -14.6 percentage points

The probability of being admitted to graduate school is .146 or 14.6 percentage points lower for a student who attended a rank 2 undergraduate school compared to a student who attended a rank 1 institution.

Rank 3 versus Rank 1 (the omitted group):
(.318) x (.682) x $\beta_{RANK\ 3}$ = .217 x (-1.34) = -.291 or -29.1 percentage points

The probability of being admitted to graduate school is .291 or 29.1 percentage points lower for a student who attended a rank 3 undergraduate school compared to a student who attended a rank 1 institution.

You can also use predicted probabilities to help you understand the model. You can calculate predicted probabilities using the **margins** command. Below we use the **margins** command to calculate the predicted probability of admission at each level of **rank**, holding all other variables in the model at their means.

margins rank, atmeans

```
Adjusted predictions                              Number of obs   =        400
Model VCE     : OIM

Expression    : Pr(admit), predict()
at            : gre             =       587.7 (mean)
                gpa             =      3.3899 (mean)
                1.rank          =       .1525 (mean)
                2.rank          =       .3775 (mean)
                3.rank          =       .3025 (mean)
                4.rank          =       .1675 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf.Interval]
-------------+----------------------------------------------------------------
        rank |
          1  |   .5166016   .0663153     7.79   0.000     .3866261   .6465771
          2  |   .3522846   .0397848     8.85   0.000     .2743078   .4302614
          3  |    .218612   .0382506     5.72   0.000     .1436422   .2935819
          4  |   .1846684   .0486362     3.80   0.000     .0893432   .2799937
------------------------------------------------------------------------------
```

In the above output we see that the predicted probability of being accepted into a graduate program is 0.51 for the highest prestige undergraduate institutions (rank=1), and 0.18 for the lowest ranked institutions (rank=4), holding **gre** and **gpa** at their means.

Below we generate the predicted probabilities for values of **gre** from 200 to 800 in increments of 100. Because we have not specified either **atmeans** or used **at(...)** to specify values at with the other predictor variables are held, the values in the table are average predicted probabilities calculated using the sample values of the other predictor variables. For example, to calculate the average predicted probability when **gre** = 200, the predicted probability was calculated for each case, using that case's values of **rank** and **gpa**, with **gre** set to 200.

margins , at(gre=(200(100)800)) vsquish

```
Predictive margins                                Number of obs   =        400
Model VCE    : OIM

Expression   : Pr(admit), predict()
1._at        : gre             =         200
2._at        : gre             =         300
3._at        : gre             =         400
4._at        : gre             =         500
5._at        : gre             =         600
6._at        : gre             =         700
7._at        : gre             =         800

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf.
Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .1667471   .0604432     2.76   0.006     .0482807   .2852135
          2  |    .198515   .0528947     3.75   0.000     .0948434   .3021867
          3  |   .2343805   .0421354     5.56   0.000     .1517966   .3169643
          4  |   .2742515   .0296657     9.24   0.000     .2161078   .3323951
          5  |   .3178483    .022704    14.00   0.000     .2733493   .3623473
          6  |   .3646908   .0334029    10.92   0.000     .2992224   .4301592
          7  |   .4141038   .0549909     7.53   0.000     .3063237   .5218839
------------------------------------------------------------------------------
```

In the table above we can see that the mean predicted probability of being accepted is only 0.167 if one's GRE score is 200 and increases to 0.414 if one's GRE score is 800 (averaging across the sample values of **gpa** and **rank**).

Calculating Predicted Probabilities from the Logit Model

Use the coefficients from the logit model predicting the probability that an individual is admitted to graduate school. (see STATA output above)

$\text{Log}(P_i / 1 - P_i) = -3.99 + 0.0023\,(\text{GRE}) + .804\,(\text{GPA}) - .675\,(\text{RANK 2})$

$- 1.34\,(\text{RANK 3}) - 1.55\,(\text{RANK 4})$

a) Calculate the predicted probability that a person with a GE score of 800, GPA of 3.5 who attended a rank 1 institution (Georgetown) is admitted to graduate school. The rank variables drop out because rank 1 is the reference category.

$\text{Log}(P_i / 1 - P_i) = -3.99 + 0.0023\,(800) + .804\,(3.5)$

$\text{Log}(P_i / 1 - P_i) = -3.99 + 0.0023\,(800) + .804\,(3.5) = -3.99 + 1.84 + 2.81 = .66$

$e^{\text{Log}(Pi / 1 - Pi)} = e^{.66} = 1.935$

$P_i / 1 - P_i = 1.935$
$P_i = 1.935\,(1 - P_i)$
$P_i = 1.935 – 1.935\,P_i$
$2.935\,P_i = 1.935$
$P_i = 1.935 / 2.935$
$P_i = 0.659$

The probability that an individual with a score of 800 on the GRE, a 3.5 GPA who attended a rank 1 institution will be admitted to graduate school is 65.9%.

b) Calculate the predicted probability that a person with a GRE score of 550, GPA of 3.0 who attended a rank 3 institution (American University) is admitted to graduate school.

$\text{Log}(P_i / 1 - P_i) = -3.99 + 0.0023\ (550) + .804\ (3.0) - 1.34\ (1) = -.313$

$e^{\text{Log}(Pi / 1 - Pi)} = e^{-.313} = 0.73$

$P_i / 1 - P_i = .73$
$P_i = .73\ (1 - P_i)$
$P_i = .73 - .73\ P_i$
$1.73\ P_i = .73$
$P_i = .73 / 1.73$
$P_i = 0.422$
The probability that a person who has a 550 score on the GRE, a 3.0 GPA and attended a rank 3 institution is admitted to graduate school is 42.2%.