17.874 Lecture Notes
Part 4: Nonlinear Models

4. Non-Linear Models

4.1. Functional Form

We have discussed one important violation of the regression assumption – omitted variables. And, we have touched on problems of inefficiency introduced by heteroskedasticity and autocorrelation. This and the following subsections deal with violations of the regression assumptions (other than the omitted variables problem). The current section examines corrections for non-linearity; the next section concerns discrete dependent variables. Following that we will deal briefly with weighting, heteroskedasticity, and autocorrelation. Time permitting we will do a bit on Sensitivity.

We encounter two sorts of non-linearity in regression analysis. In some problems non-linearity occurs among the X variables but it can be handled using a linear form of non-linear functions of the X variables. In other problems non-linearity is inherent in the model: we cannot "linearize" the relationship between Y and X. The first sort of problem is sometimes called "intrinsically linear" and the second sort is "intrinsically non-linear."

Consider, first, situations where we can convert a non-linear model into a linear form. In the Seats-Votes example, the basic model involved multiplicative and exponential parameters. We converted this into a linear form by taking logarithms. There are a wide range of non-linearities that can arise; indeed, there are an infinite number of transformations of variables that one might use. Typically we do not know the appropriate function and begin with a linear relationship between Y and X as the approximation of the correct relationship. We then stipulate possible non-linear relationships and implement transformations.

Common examples of non-linearities include:

Multiplicative models, where the independent variables enter the equation multiplicatively rather than linearly (such models are linear in logarithms);

Polynomial regressions, where the independent variables are polynomials of a set of variables (common in production function analysis in economics); and

Interactions, where some independent variables magnify the effects of other independent variables (common in psychological research).

In each of these cases, we can use transformations of the independent variables to construct a linear specification with with we can estimate all of the parameters of interest.

Qualitative Independent variables and interaction effects are the simplest sort of non-linearity. They are simple to implement, but sometimes hard to interpret. Let's consider a simple case. Ansolabehere and Iyengar (1996) conducted a series of laboratory experiments involving approximately 3,000 subjects over 3 years. The experiments manipulated the content of political advertisements, nested the manipulated advertisements in the commercial breaks of videotapes of local newscasts, randomly assigned subjects to watch specific tapes, and then measured the political opinions and information of the subjects. These experiments are reported in the book *Going Negative*.

On page 190, they report the following table.

Effects of Party and Advertising Exponsure on Vote Preferences: General Election Experiments

Independent Variable	(1) Coeff. (SE)	(2) Coeff. (SE)
Constant	.100 (.029)	.102 (.029)
Advertising Effects		
Sponsor	.077 (.012)	.023 (.034)
Sponsor*Same Party	–	.119 (.054)
Sponsor*Opposite Party	–	.028 (.055)
Control Variables		
Party ID	.182 (.022)	.152 (.031)
Past Vote	.339 (.022)	.341 (.022)
Past Turnout	.115 (.030)	.113 (.029)
Gender	.115 (.029)	.114 (.030)

The dependent variable is +1 if the person stated an intention to vote Democratic after viewing the tape, -1 if the person stated an intention to vote Republican, and 0 otherwise. The variable Sponsor is the party of the candidate whose ad was shown; it equals +1 if the ad was a Democratic ad, -1 if the ad was a Republican ad, and 0 if no political ad was shown (control). Party ID is similarly coded using a trichotomy. Same Party was coded as +1 if a person was a Democrat and saw a Democratic ad or a Republican and saw a Republican ad. Opposite Party was coded as +1 if a person was a Democrat and saw a Republican ad or a Republican and saw a Democratic ad.

In the first column, the "persuasion" effect is estimated as the coefficient on the variable Sponsor. The estimate is .077. The interpretation is that exposure to an ad from a candidate increases support for that candidate by 7.7 percentage points.

The second column estimates interactions of the Sponsor variable with the Party ID variable. What is the interpretation of the set of three variables Sponsor, Sponsor*Same Party , and Sponsor*Opposite Party. The variables Same Party and Opposite Party encompass all party identifiers. When these variables equal 0, the viewer is a non-partisan. So, the coefficient on Sponsor in the second column measures the effect of seeing an ad among independent viewers. It increases support for the sponsor by only 2.3 percentage points. When Same Party equals 1, the coefficient on Sponsor is 11.9. This is not the effect of the ad among people of the same party. It is the difference between the Independents and those of the same party. To calculate the effect of the ad on people of the same party we must add .119 to .023, yielding an effect of .142, or a 14.2 percentage point gain.

Interactions such as these allow us to estimate different slopes for different groups, changes in trends, and other discrete changes in functional forms.

Another class of non-linear specifications takes the form of Polynomials. Many theories of behavior begin with a conjecture of an inherently non-linear function. For instance, a firm's production function is thought to exhibit decreasing marginal returns on investments, capital or labor. Also, risk aversion implies a concave utility function.

Barefoot empiricism sometimes leads us to non-linearity, too. Examination of data, either a scatter plot or a residual plot, may reveal a non-linear relationship, between Y and X. While we do not know what the right functional form is, we can capture the relationship in the data by including additional variables that are powers of X, such as $X^{1/2}$, X^2, X^3, as well as X. In other words, to capture the non-linear relationship, $y = g(x)$, we approximate the unknown function , $g(x)$, using a polynomial of the values of x.

One note of caution. Including powers of independent variables often leads to collinearity among the righthand side variables. Deviating the independent variables from their means before transforming them.

Example: Candidate Convergence in Congressional Elections.

An important debate in the study of congressional elections concerns how well candidates represent their districts. Two conjectures concern the extent to which candidates reflect their districts. First, are candidates responsive to districts? Are candidates in more liberal districts more liberal? Second, does competitiveness lead to closer representation of district preferences? In highly competitive districts, are the candidates more "converged"? Huntington posits greater divergence of candidates in competitive areas – a sort of clash of ideas.

Ansolabehere, Snyder, and Stewart (2001) analyze data from a survey of congressional candidates on their positions on a range of issues during the 1996 election. There are 292 districts in which both of the competing candidates filled out the survey. The authors constructed a measure of candidate ideology from these surveys, and then examine the midpoint between the two candidates (the cutpoint at which a voter would be indifferent between the candidates) and the ideological gap between the candidates (the degree of convergence). To measure electoral competitiveness the authors used the average share of vote won by the Democratic candidate in the prior two presidential elections (1988 and 1992).

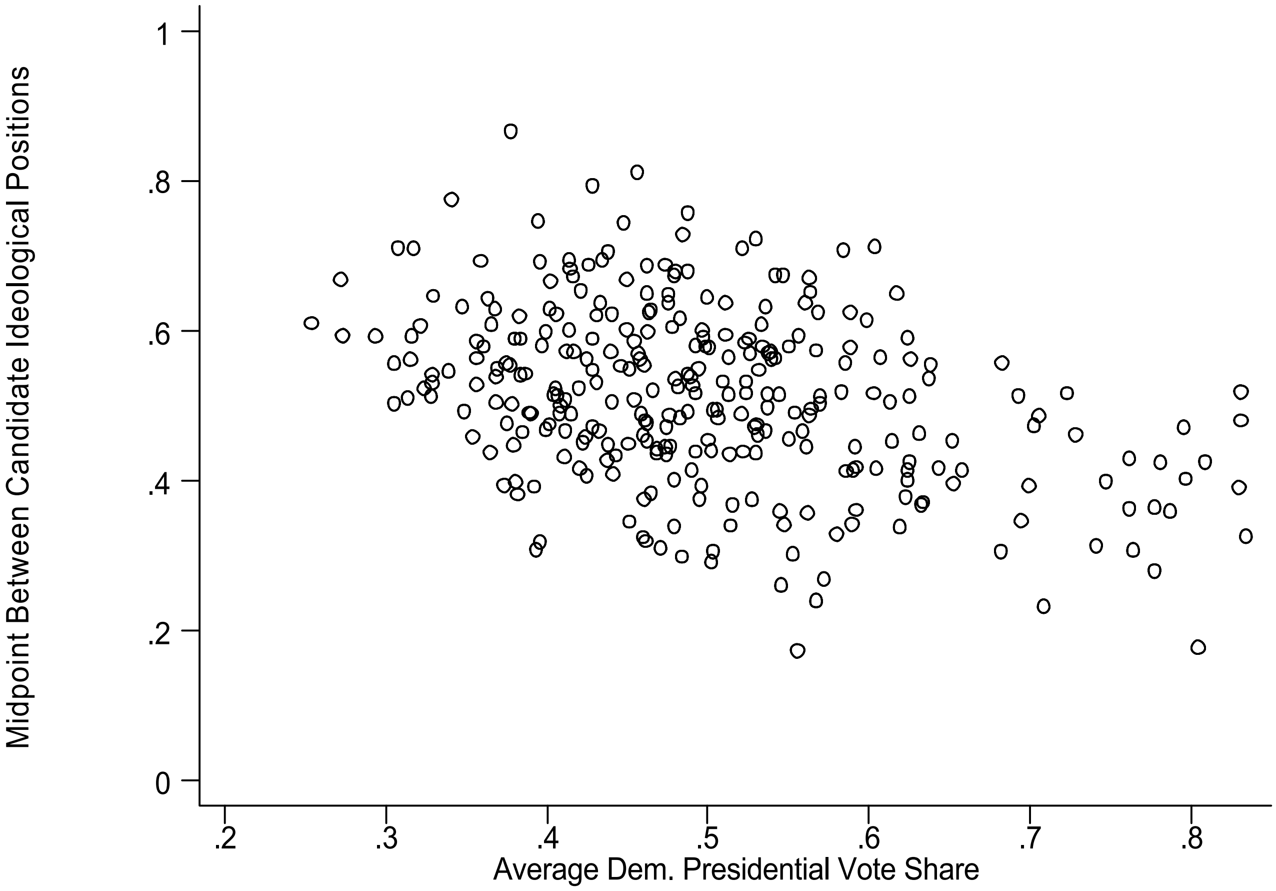
The Figures show the relationship between Democratic Presidential Vote share and, respectively, candidate midpoints and the ideological gap between competing candidates.

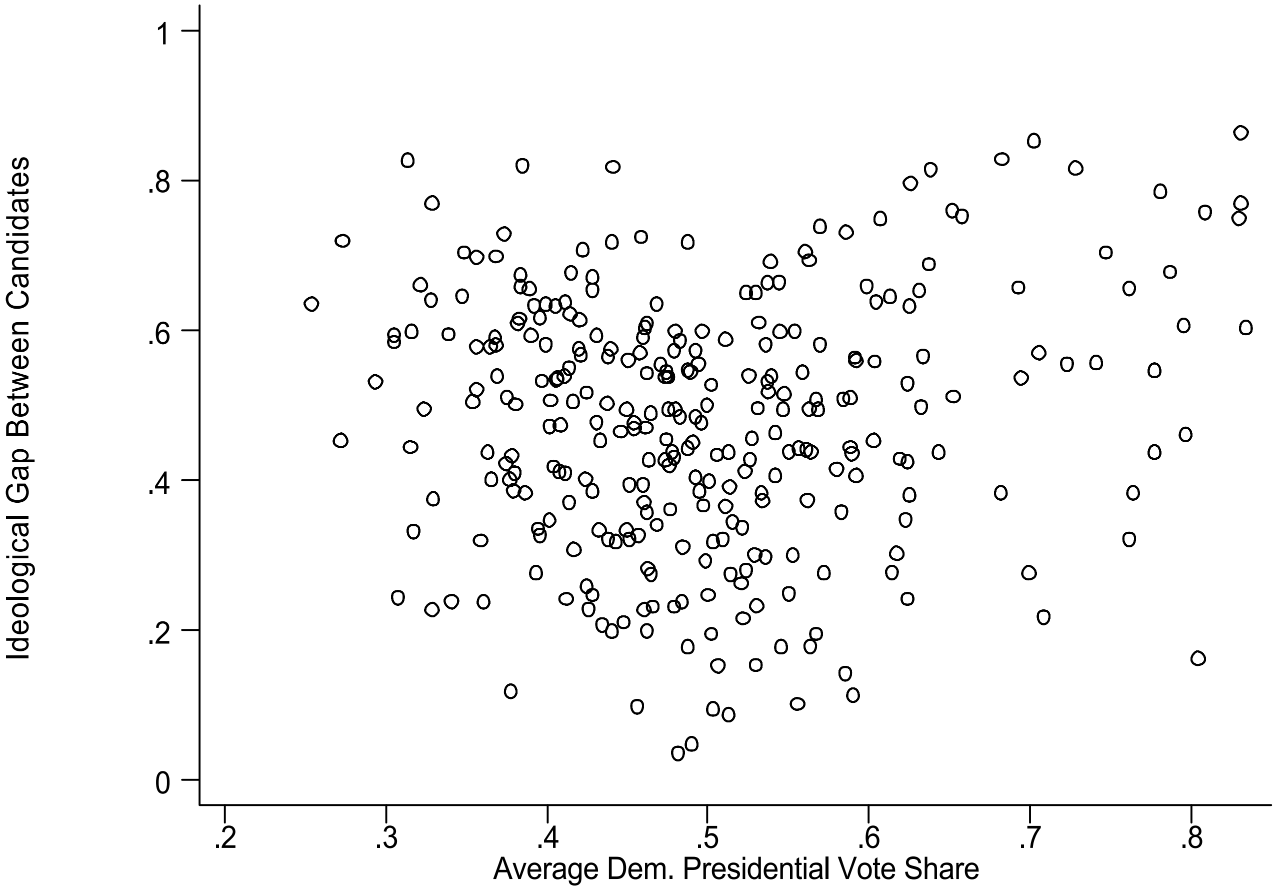
There is clear evidence of a non-linear relationship explaining the gap between the candidates

The table presents a series of regressions in which the Midpoint and the Gap are predicted using quadratic functions of the Democratic Presidential Vote plus an indicator of Open Seats, which tend to be highly competitive. We consider, separately, a specification using the value of the independent variable and its square and a specification using the value of the independent variable deviated from its mean and its square. The last column uses the absolute value of the deviation of the presidential vote from .5 as an alternative to the quadratic.

Effects of District Partisanship on Candidate Positioning
N = 292

	Dependent Variable				
	Midpoint		Gap		
Independent	(1)	(2)	(3)	(4)	(5)
Variable	Coeff.(SE)	Coeff.(SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
Democratic Presidential Vote	-.158 (.369)	—	-2.657 (.555)	—	—
Democratic Presidential Vote Squared	-.219 (.336)	—	+2.565 (.506)	—	—
Dem. Pres. Vote Mean Deviated	—	-.377 (.061)	—	-.092 (.092)	—
Dem. Pres. Vote Mean Dev. Squared	—	-.219 (.336)	—	+2.565 (.506)	—
Absolute Value of Dem. Pres. Vote, Mean Dev.	—	—	—	—	.764 (.128)
Open Seat	.007 (.027)	.007 (.027)	-.024 (.040)	-.024 (.040)	-.017 (.039)
Constant	.649 (.098)	.515 (.008)	1.127 (.147)	.440 (.012)	.406 (.015)
R^2	.156	.156	.089	.089	.110
$\sqrt{MSE}$	.109	.109	.164	.164	.162
F (p-value)	18.0 (.00)	18.0 (.00)	9.59 (.00)	9.59 (.00)	12.46 (.00)


Midpoint Between Candidate Ideological Positions
1
.8
.6
.4
.2
0
.2
.3
.4
.5
.6
.7
.8
Average Dem. Presidential Vote Share


Ideological Gap Between Candidates
0
.2
.4
.6
.8
1
.2
.3
.4
.5
.6
.7
.8
Average Dem. Presidential Vote Share

Consider, first, the regressions explaining the midpoint (columns 1 and 2). We expect that the more liberal the districts are the more to the leftward the candidates will tend. The ideology measure is oriented such that higher values are more conservative positions. Figure 1 shows a strong relationship consistent with the argument.

Column (1) presents the estimates when we naively include Presidential vote and Presidential vote squared. This is a good example of what collinearity looks like. The F-test shows that the regression is "significant" – i.e., not all of the coefficients are 0. But, neither the coefficient on Presidential Vote or Presidential Vote Squared are signficant. Tell-tale collinearity. A trick for breaking the collinearity in this case is by deviating X from its mean. Doing so, we find a significant effect on the linear coefficient, but the quadratic coefficient doesn't change. There is only really one free coefficient here, and it looks to be linear.

The coefficients in a polynomial regression measure the partial derivatives of the unknown function evaluated at the mean of X. Taylor's approximation leads us immediately to this interpretation of the coefficients for polynomial models. Recall from Calculus that Taylor's Theorem allows us to express any function as a sum of derivatives of that function evaluated at a specific point. We may choose any degree of approximation to the function by selecting a specific degree of approximation. A first-order approximation uses the first derivatives; a second order approximation uses the second derivatives; and so on. A second order approximation of an unknown fuction, then, may be expressed as:

$$y_i \approx \alpha + \beta' \mathbf{x_i} + \frac{\mathbf{1}}{\mathbf{2}} \mathbf{x_i'} \mathbf{H_0} \mathbf{x_i},$$

where

$$\mathbf{g_0} = \left[\frac{\partial \mathbf{f(x)}}{\partial \mathbf{x}} \right]_{\mathtt{x=x_0}}$$

$$\mathbf{H_0} = \left[\frac{\partial^2 \mathbf{f(x)}}{\partial \mathbf{x} \partial \mathbf{x}'} \right]_{\mathtt{x=x_0}}$$

$$\alpha = f(\mathbf{x_0}) - \mathbf{g_0'} \mathbf{x_0} + \frac{\mathbf{1}}{\mathbf{2}} \mathbf{x_0'} \mathbf{H_0} \mathbf{x_0}$$

$$\beta = \mathbf{g_0} - \mathbf{H_0} \mathbf{x_0}.$$

The coefficients on the squares and cross-product terms, then, capture the approximate second derivative. The coefficients on the linear values of x equal the gradient, adjusting for the quadrature around the point at which the data are evaluated (X_0). If we deviate all of the variables from their means first, then the coefficient on X in the polynomial regression can be interpreted straightforwardly as the gradient of the (unknown) function at the mean.

Example: Candidate Convergence in Congressional Elections, continued.

Consider the estimates in column (3). We may analyze these coefficients to ask several basic questions. What is the marginal rate of change? $\frac{\partial y}{\partial DP} = \beta_1 + 2\beta_2 DP$. Since DP ranges from .2 to .8, the rate of change in the Gap for a change in electoral competition ranges from -1.62 when DP = .2 to +1.42 when DP = .8. At what point does the rate of change equal 0 (what is the minimum)? Setting the partial equal to 0 reveals that $DP_0 = \frac{-\beta_1}{2\beta_2} = .517$, so the candidates are most converged near .5. What is the predicted value of the gap at this point? $1.127 - 2.657 * .517 + 2.565 * (.517)^2 = .438$. At the extremes, the predicted values are .64 (when DP = .8) and .70 (when DP = .2).

If we must choose among several transformations, such as logarithms, inverses, and polynomials, we typically cannot test which is most appropriate using the Wald test. Davidson and MacKinnon (1981) propose using the predicted values from one of the non-linear analysis to construct a test. Suppose we have two alternative models: $y = \mathbf{X}\beta$ and $y = \mathbf{Z}\gamma$. Consider a compound model

$$\mathbf{y} = (\mathbf{1} - \lambda)\mathbf{X}\beta + \lambda\mathbf{Z}\gamma + \epsilon$$

The J-test is implemented in two steps. First, estimate $y = \mathbf{Z}\gamma$ to get $\mathbf{Z}\hat{\gamma}$. Then, include $\mathbf{Z}\hat{\gamma}$ in the regression of $y = \mathbf{X}\beta$. The coefficient on $\mathbf{Z}\hat{\gamma}$ yields an estimate of $\hat{\lambda}$. Asymptotically, the t-ratio $\frac{\lambda\hat{-}0}{SE(\lambda}$ tests which model is appropriate.

Example: Candidate Convergence in Congressional Elections, continued.

An alternative to the quadratic specification in column (4) is an absolute value specification in column (5). Notice that the R^2 is slightly better with the absolute value. However, owing to collinearity, it is hard to test which model is more appropriate. I first regress on Gap on DP and DP^2 and $Open$ and generated predicted values, call these $GapHat$. I then regress Gap on ZG and $|DP - .5|$ and $Open$. The coefficient on $|DP - .5|$ is 1.12 with a standard error of .40, so the t-ratio is above the critical value for statistical significance. The coefficient on ZG equals -.55 with a standard error of .58. The estimated λ is not significantly different from 0; hence, we favor the absolute value specification.

4.2 Qualitative Dependent Variables.

4.2.1. Models.

Proportions. Consider a simple table.

Two Way Table

Variable	$Y = 1$	$Y = 0$
$X_1 = 1$	p_{11}	p_{10}
$X_1 = 0$	p_{01}	p_{00}

The linear effect of X_1 on Y can be modeled as the difference in means between the conditional distributions. We also allow Y to depend on other variables, say X_2. And, we can model the effects of X_1 and X_2 as additive, using i to index values of X_1 and j to index values of X_2:

$$P_{i,j} = \alpha_i + \beta_j$$

The observed proportions will follow this model plus an error term.

One relaxation of the additive effect assumption is to allow the variables to have interactive effectss on P_{ij}. Another possibility is that the effecs of X's on Y are multiplicative. A log-linear model might be specified as follows.

$$P_{ij} = A_i B_j$$

A more common model for proportions is the Logistic Regression Model. The odds that $Y = 1$ given $X_1 = i$ and $X_2 = j$ is written as $\frac{P_{ij}}{(1-P_{ij})}$. This is sometimes thought of as a measure of risk (the odds that one gets a disease or wins a lottery). The logistic model may be written as follows:

$$log\left(\frac{P_{ij}}{(1-P_{ij})}\right) = \alpha_i + \beta_j$$

This can be rewritten as:

$$P_{ij} = \frac{e^{\alpha_i+\beta_j}}{1+e^{\alpha_i+\beta_j}}$$

The table proportions themselves may be thought of as the aggregation of n individual observations on Y. The logistic regression, then, approximates an individual level Logit regression in which the dependent variable takes values 1 and 0, and the independent variables may take many values. Above we assume that they take just 2 values each. The logit function is called a "link" function, as it tells us how the additive effects are related to the dependent variables.

Latent Variables. A second way that qualitative dependent variables arise in social sciences is when there is a choice or outcome variable that depends on an underlying "latent" variable, assumed continuous. Common examples include voting, random utility models of consumer behavior, and cost-benefit calculations. The text discusses random utility models and cost benefit calculations. Let's consider voting models here.

When a legislator or individual votes, they make a decision between two alternatives, each of which is characterized by attributes. In spatial theories of voting, person i choose between two alternatives j and k. The person has a most prefered outcome, call it θ_i, and the two alternatives (parties, candidates, or proposals) represent points on the line, X_j and X_k. The voter's preferences are represented as distances from the ideal point: $u_i(X_j, \theta_i) = -(X_j - \theta_i)^2$. There may be some uncertainty about the exact distance, so $u_i(X_j, \theta_i) = -(X_j - \theta_i)^2 + \epsilon_j$. The voter chooses the alternative closer to his or her ideal point. That is, voter i chooses j over k if:

$$[-(X_j - \theta_i)^2] - [-(X_k - \theta_i)^2] + \epsilon_j - \epsilon_k > 0$$

The latent variable, Y^*, is a linear function of the set of independent variables, $\mathbf{X}$. That is,

$$\mathbf{y}^* = \mathbf{X}\beta + \epsilon.$$

However, we only observe qualitative outcomes. For example, y^* might represent how intensely someone supports or opposes a proposal, but we only observe whether they vote for the proposal or not. That is, we observe: $y = 1$ if $y^* > c$ and 0 otherwise.

$$Pr(y_i^* > c) = Pr(\mathbf{x_i'}\beta + \epsilon_\mathrm{i} > \mathbf{c}) = \mathbf{Pr}(\epsilon_\mathrm{i} > \mathbf{c} - \mathbf{x_i'}\beta)$$

If $F()$ is symmetric (as with normal and logistic), then

$$Pr(y_i^* > c) = F(\mathbf{x_i'}\beta),$$

where c is folded into $X\beta$.

Alternative Models of F. A wide variety of probability functions, F, might be employed to analyze these problems. Three are extremely common.

1. *Uniform* or *Linear Probability Model* $\epsilon \sim U(a,b)$, so $F(z) = z$.

2. *Logit* or *Logistic Probability Model*, common in psychology. $F(z) = \frac{e^z}{1+e^z}$.

3. *Probit* or *Normal Probability Model.* $\epsilon \sim N(0,1)$, so $F(z) = \Phi(z)$.

Comments: With each of these models the error variance is not constant, because the variable $Y|X$ has a Bernoulli distribution. One problem with the Linear Probability Model is that it generates out of bounds predictions; the same is true of the linear model for proportions. Generally, though, the uniform model is easier to implement and interpret, and it approximates the other functions.

4.2.2. Interpretation

Marginal Effects. The marginal effect of a change in $\mathbf{x}$ depends on the levels of other variables.

$$\frac{\partial P_i}{\partial \mathbf{x_i}} = \frac{\partial F(\mathbf{x_i'}\beta}{\partial \mathbf{x_i}}\beta$$

1. Linear Probability Model: $\frac{\partial F(z)}{\partial z} = k$, where k is the support of the uniform random variable. So

$$\frac{\partial P_i}{\partial \mathbf{x_i}} = \beta$$

2. Logit Model: $\frac{\partial F(z)}{\partial z} = \frac{e^z}{(1+e^z)^2} = \Lambda(z)(1 - \Lambda(z)$

$$\frac{\partial P_i}{\partial \mathbf{x}_\mathrm{i}} = \Lambda(\mathbf{x}_\mathrm{i}'\beta)(\mathbf{1} - \mathbf{\Lambda}(\mathbf{x}_\mathrm{i}'\beta))\beta$$

3. Probit Model: $\frac{\partial F(z)}{\partial z} = \phi(z_i)$

$$\frac{\partial P_i}{\partial \mathbf{x}_\mathrm{i}} = \phi(\mathbf{x}_\mathrm{i}'\beta)\beta$$

COMMENTS:

A general approach to interpretation is to estimate the models and then generate predicted values for each variable holding the other variables at their means.

One reason for preferring the Linear Probability Model is that the coefficients can be interpreted directly as the marginal effects of the variables on the probability that $y = 1$.

The three models are closely related to each other. If the data aren't too skewed, the coefficients can be related as follows: $\beta_U \approx .4\beta_\Phi$ and $\beta_U \approx .25\beta_\Lambda$. (See Amemiya, JEL, 1981.)

Identification. The exact parameter β is not identified. Specifically, what is estimated in these models is $\frac{\beta}{\sigma}$. Each of these probability models uses the standard distribution formula. As a result the parameter estimates depend on the variance of the underlying random variable, ϵ. The variance of ϵ will be refelected in the estimated slope. This will not affect inference, but it does affect the exact interpretation. We do not estimate the true structural parameter β, but the normalized parameter.

See handouts for comparisons.

4.3. Estimation

The probit and logit models are usually estimated using the method of Maximum Likelihood. The maximization methods do not yield closed-form solutions to the parameters of interest. So, we must approximate an exact solution using algorithms. I will review these briefly here to give you a sense of exactly what is estimated.

4.3.1. Likelihood Function

To construct the likelihood function, we consider, first, the distribution of the random variable $Y_i|\mathbf{x}_\mathrm{i}$. In the problem at hand, Y_i is a Bernoulli random variable. Denote the observed values of Y as $y_i = 1$ or 0, and note that $p_i = F(\mathbf{x}_\mathrm{i}'\beta)$. We can write the density for a given i as

$$P(Y_i = 1) = p_i^{y_i}(1-p_i)^{1-y_i} = F(\mathbf{x}_\mathrm{i}'\beta)^{\mathrm{y_i}}(\mathbf{1} - \mathbf{F}(\mathbf{x}_\mathrm{i}'\beta)^{(1-\mathrm{y_i})}.$$

We observe n observations, $i = 1, ...n$. Assuming that the observations are independent of each other, the joint density, or likelihood, is:

$$L = \Pi_{i=1}^{n} F(\mathbf{x}_\mathrm{i}'\beta)^{\mathrm{y_i}}(\mathbf{1} - \mathbf{F}(\mathbf{x}_\mathrm{i}'\beta))^{(1-\mathrm{y_i})}$$

The logarithm of the likelihood function is

$$\begin{aligned} ln(L) &= \sum_{i=1}^{n} y_i ln(F(\mathbf{x}_\mathrm{i}'\beta)) + (\mathbf{1} - \mathbf{y}_\mathrm{i})\mathbf{ln}(\mathbf{1} - \mathbf{F}(\mathbf{x}_\mathrm{i}'\beta)) && (1) \\ &= \sum_{i \in (y=1)} ln(F(\mathbf{x}_\mathrm{i}'\beta)) + \sum_{\mathrm{i} \in (\mathrm{y}=0)} \mathbf{ln}(\mathbf{1} - \mathbf{F}(\mathbf{x}_\mathrm{i}'\beta)) && (2) \end{aligned}$$

To fully specify the likelihood function, substitute the particular formula for F, e.g., $F(z) = \frac{e^z}{(1+e^z)}$ or $F(z) = \Phi(z)$.

4.3.2. Maximum Likelihood Estimation

Maximum likelihood estimates are values of β and σ such that $ln(L)$ is maximized. The gradient of the log-likelihood function with respect to β is:

$$\begin{aligned}\frac{\partial ln(L)}{\beta} &= \sum_{i=1}^{n} \frac{y_i f(\mathbf{x}_\mathrm{i}'\beta)}{(F(\mathbf{x}_\mathrm{i}'\beta)}\mathbf{x}_\mathrm{i} - \frac{(\mathbf{1}-\mathbf{y}_\mathrm{i})\mathbf{f}(\mathbf{x}_\mathrm{i}'\beta)}{(\mathbf{1}-\mathbf{F}(\mathbf{x}_\mathrm{i}'\beta))}\mathbf{x}_\mathrm{i} &\quad (3)\\ &= \sum_{i=1}^{n} \frac{f(\mathbf{x}_\mathrm{i}'\beta)}{F(\mathbf{x}_\mathrm{i}'\beta)(\mathbf{1}-\mathbf{F}(\mathbf{x}_\mathrm{i}'\beta))}(y_i - F(\mathbf{x}_\mathrm{i}'\beta))\mathbf{x}_\mathrm{i} &\quad (4)\end{aligned}$$

In the case of the Logit model, $f_i = \Lambda_i(1-\Lambda_i)$, so the first order conditions are:

$$\frac{\partial ln(L)}{\beta} = \sum_{i=1}^{n}(y_i - \Lambda(\mathbf{x}_\mathrm{i}'\hat{\beta}))\mathbf{x}_\mathrm{i} = \mathbf{0}$$

In the case of the Probit model

$$\frac{\partial ln(L)}{\beta} = \sum_{i=1}^{n} \frac{(y_i - \Phi(\mathbf{x}_\mathrm{i}'\hat{\beta}))\phi(\mathbf{x}_\mathrm{i}'\hat{\beta})}{\Phi(\mathbf{x}_\mathrm{i}'\hat{\beta})(\mathbf{1}-\mathbf{\Phi}(\mathbf{x}_\mathrm{i}'\hat{\beta}))}\mathbf{x}_\mathrm{i}$$

For neither one of these can we find an explicit solution for $\hat{\beta}$.

A solution is found using algorithms that calculate successive approximations using Taylor's Theorem. From Taylor's Theorem, a first-order approximation to the first order condition is:

$$\frac{\partial ln(L)}{\beta} \approx g(\beta^0) + H(\beta^0)(\hat{\beta} - \beta^0) = 0,$$

where β^0 is the value at which we approximate the function, $g(\beta^0)$ is the gradient at this point, $H(\beta^0)$ is the Hessian evaluated at this point. We will treat β^0 as an initial guess, usually a vector of 0's. We can rewrite this formula to have the following "updating formua," where β^1 is our "updated guess":

$$\beta^1 = \beta^0 - H(\beta^0)^{-1}g(\beta^0)$$

The updating formula can be used iteratively to arrive at a value of b^t, for the t-th iteration. The program will converge to a solution when $b^{t+1} \approx b^t$, to a chosen degree of approximation. In maximum likelihood estimation the convergence criterion used is the change in the log-likelihood between successive iterations. The value of b^t is substituted into the log-likelihood formula to calculate this value. When the percentage change in the log-likelihood is very small, the program "converges." See the handout.

4.3.3. Properties of Maximum Likelihood Estimates

Maximum likelihood estimation generally has two important properties.

First, maximum likelihood estimates are *consistent*. With large enough samples, the estimate approaches the parameter of interest, assuming that the parameter is identified. Identification is not assured. For example, in probit, we can only estimate the ratio β/σ_ϵ. Consistency for Probit is that $plimb = \beta/\sigma$.

Second, MLE's are efficient, asymptotically. One definition and two results are required to show this. The *Information Matrix* equals the expected value of the Hessian of the log-likelihood. That is

$$I[\theta] = E\left[-\frac{\partial^2 lnL(\theta)}{\partial\theta\partial\theta'}\right]$$

Consider, for example, a sample of n observations from the normal distribution. $lnL = \frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2$. The gradient is $(\frac{\partial lnL}{\partial\mu}, \frac{\partial lnL}{\partial\sigma^2})' = (\frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\mu), \frac{-n}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^n(x_i-\mu)^2)'$. The Hessian, then, is

$$H = \begin{pmatrix} -\frac{n}{\sigma^2}, -\frac{\sum_{i=1}^n(x_i-\mu)}{\sigma^4} \\ -\frac{\sum_{i=1}^n(x_i-\mu)}{\sigma^4}, \frac{n}{2\sigma^4} - \frac{\sum_{i=1}^n(x_i-\mu)^2}{\sigma^4} \end{pmatrix}$$

$$I(\theta) = E[-H] = \begin{pmatrix} -\frac{n}{\sigma^2}, 0 \\ 0, \frac{n}{2\sigma^4} \end{pmatrix}$$

Result 1: Cramer-Rao Lower Bound. Consider any estimator $\tilde{\theta}$.

$$V(\tilde{\theta}) \geq [I(\theta)]^{-1}$$

Result 2: The Asymptotic Variance-Covariance matrix of the MLE is: $I[\hat{\theta}]^{-1}$.

Consider, again, the normal example.

$$I(\theta)^{-1} = \begin{pmatrix} -\frac{n}{\sigma^2}, 0 \\ 0, \frac{n}{2\sigma^4} \end{pmatrix}^{-1} = \begin{pmatrix} -\frac{\sigma^2}{n}, 0 \\ 0, \frac{2\sigma^4}{n} \end{pmatrix}^{-1}$$

This is the Cramer-Rao lower bound, which the MLE approaches as n grows. Notice also that we can prove convergence in mean-square, and thus, consistency through examination of this matrix.

When we cannot derive closed form solutions for the variance-covariance matrix, the results above give us a handy method. Specifically, substitute the value of the maximum likelihood estimate into the Hessian and average over the values of the observations.

A further simplification in the calculation of the variance-covariance matrix arises from a useful identity: $E[\frac{\partial^2 lnL}{\partial\theta\partial\theta'}] = E[\frac{\partial lnL}{\partial\theta}\frac{\partial lnL}{\partial\theta'}]$ (see Chapter 17 of Greene). This means that we can use just the gradients (first derivatives) to compute the estimated variance-covariance matrix. Denote g_i as the gradient for any observation i. Then the estimated variance covariance matrix is:

$$I(\hat{\theta})^{-1} = [-\sum g_i g_i']^{-1}$$

[Note: Verify this with the simple normal case.]

Lastly, we should observe that the maximum likelihood estimates are approximately normally distributed. This emerges readily from consideration of the Taylor's expansion. If we approximate the function around the vector $\hat{\theta}$, we see that

$$g(\hat{\theta}) = g(\theta) + H(\theta)(\hat{\theta} - \theta) = 0$$

Hence,

$$(\hat{\theta} - \theta) = [-H(\theta)]^{-1} g(\theta),$$

which is the weighted average of a random variable (g). The usual argument then means that $\hat{\theta}$ will follow the normal distribution, even if we are dealing with a highly non-linear problem. Hence,

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, [I(\theta)]^{-1}),$$

Exercise. Least squares estimator for the linear regression model, along with $I(\theta)$ and $I(\theta)^{-1}$.

4.4. Inference

4.4.1. Confidence Intervals

As with the regression model we wish to construct confidence intervals and conduct statistical tests on single coefficients. While the values of the coefficients may be difficult to interpret immediately, statistical testing is relatively straightforward.

To conduct inferences about single parameters, consider the marginal distribution of a single parameter estimate. Because of normality, we know that the distribution of any parameter estimate, say b_k, will have mean β_k and variance equal to the kth diagonal element of $I(\theta)^{-1}$, a_{kk}. Hence,

$$Pr(|b_k - \beta_k| > z_{\alpha/2}\sqrt{a_{kk}}) < \alpha$$

Therefore, a 95 percent confidence interval for β_k is

$$b_k \pm z1.96\sqrt{a_{kk}}$$

Examples: see handouts.

4.4.1. Likelihood Ratio Tests

As with our earlier discussion of inference, we can drawm mistaken conclusions – putting too much or too little confidence in an inference – if we use the procedure for a single coefficient across many coefficients. The analogue of the Wald-test in the likelihood based models is the likelihood ratio test.

Consider two models. For simplicity assume that one of the models is a proper subset of the others – that is, the variables included in one of the models encompasses that in another model. The implicit hypothesis is that the subset of excluded variables from the more parsimonious model all have coefficients equal to 0. This amounts to a constraint on the maximum likelihood estimates. The unrestricted model will necessarily have higher likelihood. Is the difference in the likelihoods significant?

The likelihood ratio is calculated by evaluating the likelihood functions at the vector of cofficients in the two models. The ratio of the likelihood statistics is:

$$\hat{\lambda} = \frac{L(\hat{\theta_R})}{L(\hat{\theta_U})}.$$

This ratio is always less than 1. If it is much less than 1 then there is some evidence against the hypothesis.

The relevant probability distribution can be derived readily for the log of of the likelihood statistic.

$$-2ln\lambda = -2(ln(L(\hat{\theta_R})) - ln(L(\hat{\theta_U})) \sim \chi^2_J$$

where J is the number of parameters restricted by the hypothesis.

Why does this follow the χ^2? This is readily evident for the normal case. The log of the normal function yields the sum of squares. So the log of the likelihood is the difference in the sums of squares – i.e., the difference between two estimated variances.

It should be noted that this is the same information as used in the Wald test – the loss of fit. The Wald test differs primarily in that it normalizes by the fit of the unrestricted model. It can be shown that asymptotically, Wald and Likelihood ratio statistics are the same.