Brock University
Economics 3P90
Intermediate Econometrics
Ivan Medovikov

Course Review Notes

Important Notices
These notes provide a brief overview and summary of some of the topics covered in class. They are NOT a complete collection of all of the examinable material covered in this course. A significant amount of examinable material covered in class is NOT included into these Review Notes. You should not use these Review Notes as your single source for exam preparation, but review the full lecture notes in addition to these Review Notes. Review notes may contain typos.

1 Matrix Approach to Regression

1.1 Re-Writing the Model in Matrix-Vector Form

Consider a regression model given by:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i, \text{ for } i = 1, ..., n \tag{1}$$

which we can write as a system of equations:

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{21} + ... + \beta_k X_{k1} + u_1 && (2)\\ Y_2 &= \beta_1 + \beta_2 X_{22} + ... + \beta_k X_{k2} + u_2 && (3)\\ &\vdots && (4)\\ Y_n &= \beta_1 + \beta_2 X_{2n} + ... + \beta_k X_{kn} + u_n && (5)\\ & && (6) \end{aligned}$$

Letting y and u be $n \times 1$ vectors, X be an $n \times k$ matrix, and β an $k \times 1$ vector given by:

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, X = \begin{bmatrix} 1 & X_{21} & ... & X_{k1} \\ 1 & X_{22} & ... & X_{k2} \\ \vdots & & & \\ 1 & X_{2n} & ... & X_{kn} \end{bmatrix}, \text{ and } u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

we can re-write the model in a **matrix-vector form** as

$$y = X\beta + u. \tag{7}$$

1.2 Vector OLS

1.2.1 The Estimator

The Residual Sum of Squares (RSS) can then be expressed in terms of the vectors y, X and β as $\sum_{i=1}^{n} \hat{u}_i^2 = u'u = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - y'X\hat{\beta} - \hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$. To obtain the OLS estimate of the coefficients vector β, we minimize the RSS by taking its derivative with respect to β and setting it to zero. This leads to the **OLS normal equation** given by $\frac{\delta u'u}{\delta \hat{\beta}} = 0$. Solving this for $\hat{\beta}$ we get the **OLS estimator**

$$\hat{\beta} = (X'X)^{-1}(X'y). \tag{8}$$

1.2.2 Numerical and Statistical Properties

Note the following facts about the vector OLS estimator (8):

(a) Only sample data (y and X) are needed to find $\hat{\beta}$

(b) By construction, the regression line given by $\hat{y} = X\hat{\beta}$:

- (i) Passes through the point of sample means (i.e. $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2\bar{X}_2 + ... + \hat{\beta}_k\bar{X}_k$)
- (ii) Is such that $E[\hat{y}] = E[y]$
- (iii) Also, $E[\hat{u}] = 0$
- (iv) And $\hat{u}$ is uncorrelated with X's

To obtain the statistical properties of $\hat{\beta}$, we first need to make a set of assumptions, often termed **the classical linear regression model assumptions**. In particular, assume that

1. **Linearity** The true model is indeed linear and can be written as $y = X\beta + u$.
2. **Exogenous Regressors** Assume X is fixed (and is therefore independent from u).
3. **Zero-Mean Error** Assume that $E[u] = 0$.
4. **Absence of Heteroscedasticity and Serial Correlation** Assume that $E[uu'] = \sigma^2 I$, where I is a $n \times n$ identity matrix, and σ^2 is some constant.
5. **No Exact Collinearity** Assume that data matrix X has full rank.
6. **Sufficient Degrees of Freedom** Assume that we have enough data to estimate k coefficients, namely, assume that $n > k$.

Further assuming that u is Normally-distributed, we have that under the above assumptions, the vector OLS estimator $\hat{\beta}$ has the following **statistical properties**:

1. **Unbiased** $E[\hat{\beta}] = \beta$

2. **Consistent** $\hat{\beta} \to \beta$ as $n \to \infty$

3. **Normally-Distributed** $\hat{\beta}$ has a sampling distribution which is normal with mean β and variance given by
$$var(\hat{\beta}) = \sigma^2(X'X)^{-1} \tag{9}$$

4. **Efficient** Meaning that its sampling variance $\sigma^2(X'X)^{-1}$ is the smallest possible.

This is the so-called **Gauss-Markov theorem**, which tells us that under the above assumptions, the OLS estimator $\hat{\beta}$ is BLUE (Best Linear Unbiased Estimator).

Note that in practice, we also need to estimate σ^2 in order to be able to obtain an estimate of the sampling variance of the $\hat{\beta}$. We can do this using

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n-k}. \tag{10}$$

Substituting this into (9) gives us a feasible estimator for the sampling variance:

$$\hat{var}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1} = \begin{bmatrix} var(\hat{\beta}_1) & ... & cov(\hat{\beta}_1, \hat{\beta}_k) \\ ... & var(\hat{\beta}_2) & ... \\ \vdots & & \\ cov(\hat{\beta}_k, \hat{\beta}_1) & ... & var(\hat{\beta}_k) \end{bmatrix} \tag{11}$$

1.3 Generalized Least Squares

Suppose that heteroskedasticity and/or auto-correlation is present in the data. Recall that Heteroskedasticity arises when the variances of the error terms in the model are not constant, that is, when $var(u_i) = \sigma_i^2$, $i = 1, .., n$. Auto-correlation arises when at least some error terms are correlated, that is, when $cov(u_i, u_j) \neq 0$, for at least some $i \neq j$. OLS estimator is unbiased but inefficient in the presence of heteroskedasticity alone, but can be both biased and inefficient under auto-correlation. In the absence of both heteroskedasticity and auto-correlation, error co-variance matrix can be written simply as

$$E[uu'] = \sigma^2 I, \tag{12}$$

where σ^2 is some constant and I is an $n \times n$ identity matrix. In the presence of heteroskedasticity *alone*, error co-variance matrix can be written as

$$E[uu'] = \sigma^2 \Omega, \tag{13}$$

where σ^2 is some constant and Ω is an $n \times n$ diagonal matrix containing variance scaling factors. In the presence of *both* heteroskedasticity and auto-correlation, error co-variance matrix can be written as

$$E[uu'] = \sigma^2 \Omega, \tag{14}$$

where σ^2 is some constant as before and Ω is still a symmetric $n \times n$, but it is no longer diagonal (i.e. off-diagonal elements are non-zero). In this case, unbiased and efficient estimates can still be obtained using the generalized least squares (GLS) estimator given by

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}y). \tag{15}$$

1.3.1 GLS Statistical Properties

Assuming zero mean model error ($E[u] = 0$), independence of regressors X from errors u, and full rank of the data matrix X we get that the GLS estimator is unbiased and has minimum variance (efficiency) among linear estimators (Aitken's theorem). Note that lack of heteroskedasticity and auto-correlation is no longer required.

1.3.2 Feasible GLS: Weighted Least Squares

The scaling matrix Ω is typically unknown, meaning that in practice, it is replaced by an estimate $\hat{\Omega}$. In the case when only heteroskedasticity is present (and hence Ω is diagonal), an estimate $\hat{\Omega}$ can be obtained as follows:

(i) Estimate $y = X\beta + u$ using OLS, ignoring heteroskedasticity. Recall that this yields unbiased but inefficient estimates $\hat{\beta}$.

(ii) Use the estimates $\hat{\beta}$ to find the residuals $\hat{u} = y - X\hat{\beta}$.

(iii) Let $z = ln(\hat{u}^2)$, and estimate the following auxiliary regression with OLS: $z = X\delta + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon^2)$ is an error term.

(iv) Use the fitted values $\hat{z} = X\hat{\delta}$ from the auxiliary regression to estimate error variances in the original model as $\hat{\sigma}_i^2 = \exp(\hat{z}_i)$. Construct weighting matrix $\hat{\omega}$ by inserting estimates $\hat{\sigma}_i^2$ on its main diagonal.

(v) Estimate original model using feasible GLS: $\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}y)$.

1.4 Testing Linear Restrictions (the Wald's test)

One of the many advantages of the vector approach is the ability to jointly test sets of linear restrictions on the coefficients from our regression model. Suppose that we wish to test a joint hypothesis given by:

$$H_0 : A\beta = r \tag{16}$$
$$H_1 : A\beta \neq r, \tag{17}$$

where A is an $p \times k$ linear restrictions matrix representing p restrictions on the k model coefficients, and r is the $k \times 1$ vector of constants. We can test this hypothesis using the Wald's test statistic given by:

$$W = \frac{(A\hat{\beta} - r)'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - r)}{\hat{\sigma}^2}, \tag{18}$$

which follows an F distribution with p numerator and $n-k$ denominator degrees of freedom (where, again, p is the number of restrictions, k is the number of betas and n is the sample size). We reject the H_0 if $W > F_{p,n-k}$, where $F_{p,n-k}$ is the corresponding critical value at some desired level of significance. See the lecture notes for examples showing how the linear restrictions matrix A and the constants vector r are constructed.

2 Non-Linear Regression Models

In some cases, the relationship between the regressand Y and the regressors $X_{2i}, ..., X_{ki}$ may not be linear, meaning that a non-linear regression model may be more appropriate. For example, suppose that we wish to estimate a production function for the Canadian economy, which we hypothesize, is Cobb-Douglas. Letting Y_i represent the output (GDP), X_{2i} capital and X_{3i} labour, the model is given by:

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} u_i. \tag{19}$$

Note that we can easily linearise this model by taking the natural logarithm of both sides:

$$ln(Y_i) = \ln(\beta_1) + \beta_2 ln(X_{2i}) + \beta_3 ln(X_{3i}) + ln(u_i). \tag{20}$$

Letting $\tilde{Y}_i = ln(Y_i)$, $\tilde{\beta}_1 = ln(\beta_1)$, $\tilde{X}_{2i} = ln(X_{2i})$, $\tilde{X}_{3i} = ln(X_{3i})$ and $\tilde{u}_i = ln(u_i)$, we can re-write the model as:

$$\tilde{Y}_i = \tilde{\beta}_1 + \beta_2 \tilde{X}_{2i} + \beta_3 \tilde{X}_{3i} + \tilde{u}_i, \tag{21}$$

which is a linear model, and can therefore be estimated using the OLS. Letting $\dot{\tilde{\beta}}_1$, $\dot{\beta}_2$ and $\dot{\beta}_3$ represent the OLS estimates of the coefficients from the linearised model (21), we can obtain estimates of the original coefficients from the non-linear model (20) as:

$$\begin{aligned} \hat{\beta}_1 &= exp(\dot{\tilde{\beta}}_1) \\ \hat{\beta}_2 &= \dot{\beta}_2 \\ \hat{\beta}_3 &= \dot{\beta}_3. \end{aligned}$$

Models which can be linearised in a similar way are **intrinsically-linear**. Some models, however, are impossible to linearise, and we say that they are **intrinsically non-linear**.

2.1 Non-Linear Least Squares

Consider an intrinsically non-linear regression model given by:

$$Y_i = f(X_i; \beta) + u_i, \tag{22}$$

where $f(X_i; \beta)$ is some arbitrary non-linear regression function which depends on the regressor X_i and coefficients vector β (for example, we could have something like $f(X_i; \beta) = X_i^{ln(\beta)}$). While the OLS cannot be applied to model (22), we could use another least squares method, **the non-linear least squares**. Note that we can write the model residual as $\hat{u}_i = Y_i - f(X_i; \hat{\beta})$ and therefore the residual sum of squares (RSS) as $\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - f(X_i; \hat{\beta}))^2$, which can be minimized by taking the partial derivative of the RSS with respect to $\hat{\beta}$ and setting it to zero. This gives the normal equations corresponding to model (22):

$$\sum_{i=1}^{n} \frac{\delta (Y_i - f(X_i; \hat{\beta}))^2}{\delta \hat{\beta}} = 0. \tag{23}$$

We can solve the normal equation (23) for $\hat{\beta}$ and therefore obtain a non-linear least squares estimate $\hat{\beta}$. The problem is that more often than not, equation (23) may not have a closed-form solution, meaning that it may actually not be possible to solve out for $\hat{\beta}$ explicitly. Despite that, we open can find the solution numerically (using the computer), using one of the following algorithms.

2.2 Common NLLS Algorithms

2.2.1 Grid Search (Trial-and-Error)

They key idea is to try a sufficiently large number of different values for $\hat{\beta}$ hoping that by chance we find the one which solves equation (23), and is, therefore, our correct NLLS estimate. Two types of algorithms within this class are popular:

(a) **Randomized Grid Search**

   (a) Draw a random value for $\hat{\beta}$

   (b) Calculate the left-hand-side of equation (23) using this value of $\hat{\beta}$ (alternatively, calculate the RSS)

   (c) Repeat above steps a sufficiently-large number of times

   (d) Choose the value of $\hat{\beta}$ which makes the left-hand side of equation (23) as close to zero as possible (alternatively, choose the value which yields the smallest RSS)

(b) **Directed Grid Search**

   (a) Create a sufficiently-fine grid of candidate values for $\hat{\beta}$

   (b) Calculate the left-hand side of equation (23) for each of the values of $\hat{\beta}$ on the grid (alternatively, calculate RSS)

   (c) Choose the value of $\hat{\beta}$ which makes the left-hand side of (23) closes to zero (alternatively, choose the one which gives minimum RSS)

(c) **Method of Steepest Descent** The key principle behind this method is to assume that RSS is convex and differentiable, in which there exists a unique value of $\hat{\beta}$ which minimizes RSS.

   (a) Pick a starting value for $\hat{\beta}$, possibly at random

   (b) Evaluate the left-hand side of (23) at this point. If it is positive, then increasing $\hat{\beta}$ increases RSS, and vice versa.

   (c) Adjust the value of $\hat{\beta}$ accordingly (i.e. increase is derivative is negative, decrease is positive).

   (d) Repeat the above steps until you find a value of $\hat{\beta}$ such that the derivative is zero (or until $\hat{\beta}$ stops changing by much, i.e. converges).

(d) **Iterative Linearisation** Recall from your calculus course that we can approximate a continuous and differentiable function (say $g(x)$) at around some point (say x_0) using the Taylor's polynomial as:

$$g(x) = \frac{g(x_0)}{0!} + \frac{g^1(x_0)(x - x_0)}{1!} + ... + \frac{g^n(x_0)(x - x_0)^n}{n!} + R, \tag{24}$$

where $g^n(x)$ is the n'th derivative of $g(x)$ with respect to x, and R is the remainder. The larger is the value of n which we choose, the more accurate is the approximation (and the smaller is the remainder R). The nice thing about this expansion is that unlike the original function $g(x)$ on the left-hand side of (24), which may be non-linear, the approximation on the right-hand side of (24) is linear when we choose $n = 1$.

In the context of our example, this means we can approximate the non-linear regression function $f(X_i; \beta)$ around some value β_0 using a linear function as

$$f(X_i; \beta) = f(X_i; \beta_0) + f'(X_i; \beta_0)(\beta - \beta_0) + R_i, \tag{25}$$

which is linear in β! Substituting this into the non-linear model (22), we get

$$Y_i = f(X_i; \beta_0) + f'(X_i; \beta_0)(\beta - \beta_0) + R_i + u_i, \tag{26}$$

and letting $\tilde{\beta}_1 = f(X_i; \beta_0) - \beta_0 f'(X_i; \beta_0)$, $\tilde{X}_i = f'(X_i; \beta_0)$ and $\tilde{u}_i = u_i + R_i$, we can write the model as

$$Y_i = \tilde{\beta}_1 + \beta \tilde{X}_i + \tilde{u}_i, \tag{27}$$

which is a linear model and can be estimated using the OLS. The iterative linearisation algorithm is centred around the repetition of this procedure until the value of $\hat{\beta}$ converges.

(a) Generate a starting value for $\hat{\beta}_0$, possibly at random

(b) Linearise the model around this point as above, and estimate the linearised model using OLS

(c) Use the resulting estimate $\hat{\beta}$ to update the value of $\hat{\beta}_0$

(d) Repeat the above steps until $\hat{\beta}$ converges (i.e. stops changing from one repetition to another)

2.3 The Method of Maximum Likelihood

One of the most potent of the estimation methods used in econometrics is the method of maximum likelihood. Note from equation (22) that if u_i is normally-distributed with zero mean and variance σ^2, it must be that Y_i is also normally-distributed with mean $f(X_i; \beta)$ and variance σ^2. But then we can write the *likelihood* associated with observation i using the normal pdf (probability density function), or, in other words, the normal "bell-curve". This likelihood tells us the probability of observing an observations with values Y_i and X_i, which is a function of β and σ^2. Following the MLE approach, we when choose $\hat{\beta}$ and $\hat{\sigma^2}$ so that to maximize this likelihood. That is, we search for the values of the coefficients which make them the most likely, given the observed sample.

The standard normal pdf is given by

$$F(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X-\mu_x)^2}{2\sigma^2}\right), \tag{28}$$

where μ_x is the mean, and σ^2 is the variance. Then, the **likelihood of Y_i is given by**

$$F(Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - f(X_i;\beta))^2}{2\sigma^2}\right), \tag{29}$$

and the **sample likelihood** is

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} F(Y_i). \tag{30}$$

Taking the log of the sample likelihood, we get the **log-likelihood** given by

$$l(\beta, \sigma^2) = ln(\prod_{i=1}^{n} F(Y_i)) = \sum_{i=1}^{n} ln(F(Y_i)). \tag{31}$$

To find the values of $\hat{\beta}$ and $\hat{\sigma}^2$ which maximize the log-likelihood, we take the partial derivatives of (31) with respect to β and σ^2, set them to zero, and solve out for $\hat{\beta}$ and $\hat{\sigma}^2$. These are our **maximum likelihood estimates**.

Note that when $f(X_i;\beta)$ is linear (e.g. suppose that $f(X_i;\beta) = \beta_1 + \beta_2 X_i$), when MLE estimators for $\hat{\beta}_1$ and $\hat{\beta}_2$ are identical to the OLS estimators. In that sense, OLS can be thought of as being the special case of MLE, which arises when the underlying model is linear. You are welcome to verify this in your spare time (this is easier than it may sound :)).

The great thing about MLE is that under some very general assumptions, it tends to produce estimates which are **unbiased** and **efficient**.

3 Qualitative Response Models (QRMs)

So far, the dependent variable Y_i in our models was a number. There can be many situations in which it is not. For example, Y_i may be employment or home-ownership status, it can be gender, color, or political preference. A separate class of models exists in econometrics which permits the analysis of such dependent variables which are not quantitative, but rather qualitative in nature.

3.1 The Estimation Problem

Up until this point, the objective was to build and estimate the model for the regression function which gives the conditional mean of the dependent variable given some known values of the regressors: $E[Y_i|X_{2i}, ..., X_{ki}]$. When working with qualitative responses, the estimation of the conditional probability, rather than the conditional expectation, becomes the objective.

For example, suppose that we wish to study the relationship between political preference of an individual, and his/her income. Following the QRM approach, we would then attempt to build and estimate the model of the conditional probability that an individual votes, for example, liberal, given a certain level of income.

3.2 The Linear Probability Model (LPM)

Perhaps the simplest conditional probability model is the so-called Linear Probability Model, or LPM for short. For this example, suppose Y_i represents individual's home ownership status such that:

$$Y_i = \begin{cases} 1 & \text{if individual is a homeowner,} \\ 0 & \text{if not,} \end{cases} \tag{32}$$

and X_i is the corresponding income. The corresponding linear probability model is specified as

$$Y_i = \beta_1 + \beta_2 X_i + u_i. \tag{33}$$

By examining the corresponding regression function, note that in (33), $E[Y_i|X_i] = 1P(Y_i = 1|X_i) + 0P(Y_i = 0|X_i) = P(Y_i = 1|X_i) = \beta_1 + \beta_2 X_i$, meaning that **in the LPM, the conditional probability is modelled as a linear function** of the regressors. Since (33) is linear, we could estimate it using the OLS. However, note that since Y_i is either 1 or 0, for any i, then, for a given value of X_i, it must be that the error u_i is either

$$u_i = 1 - \beta_1 - \beta_2 X_i, \text{ when } Y_i = 1, \tag{34}$$

$$u_i = 0 - \beta_1 - \beta_2 X_i, \text{ when } Y_i = 0. \tag{35}$$

Clearly, u_i cannot be normally-distributed in this setting (in fact, it has a so-called Bernoulli distribution). Moreover, we can show that the variance of u_i is $\sigma^2_{u_i} = P(Y_i = 1|X_i)(1-P(Y_i = 1|X_i)$, which is clearly not constant across i, as it depends on the values of X_i. Therefore, the LPM suffers from **heteroscedasticity**! This means the OLS estimates of (33) are **unbiased**, but **inefficient**.

3.2.1 Weighted Least Squares LPM Estimation

Efficient estimates of the coefficients from (33) can be obtained despite heteroscedasticity through the use of a procedure called the **weighted least squares**. Let $P_i = P(Y_i = 1|X_i)$, and let $w_i = P_i(1 - P_i)$. Then, transform the model (33) by dividing both sides by $\sqrt{w_i}$. Let $\tilde{Y}_i = Y_i/\sqrt{w_i}$, $\tilde{X}_{1i} = 1/\sqrt{w_i}$, $\tilde{X}_i = X_i/\sqrt{w_i}$ and $\tilde{u}_i = u_i/\sqrt{w_i}$, then, then transformed model given by

$$\tilde{Y}_i = \beta_1 \tilde{X}_{1i} + \beta_2 \tilde{X}_i + \tilde{u}_i \tag{36}$$

is homoscedastic, and estimating (36) with OLS yields **unbiased** and efficient **estimates**. We call these **the weighted least squares** (or WLS) estimates.

To obtain the WLS estimates in practice, we take the following steps:

1. Estimate the model (33) using OLS, ignoring heteroscedasticity
2. Use the OLS estimates to calculate $\hat{P}_i = \hat{\beta}_1 + \beta_2 X_i$
3. Calculate the weights as $w_i = \hat{P}_i(1 - \hat{P}_i)$
4. Transform the data as $\tilde{Y}_i = Y_i/\sqrt{\hat{w}_i}$, $\tilde{X}_{1i} = 1/\sqrt{\hat{w}_i}$ and $\tilde{X}_i = X_i/\sqrt{\hat{w}_i}$
5. Estimate the transformed model $\tilde{Y}_i = \beta_1 \tilde{X}_{1i} + \beta_2 \tilde{X}_i + \tilde{u}_i$ using OLS. The resulting estimates are the WLS estimates of (33) which are **unbiased** and **efficient**.

3.2.2 The Limitations of the LPM

The key advantage of the LPM is its simplicity. However, the model is not without limitations. Firstly, notice that there's nothing to ensure that the predicted probability $\hat{P}_i$ will lie between 0 and 1. We can easily choose values for the betas and for X_i in (33) such that $\hat{P}_i$ is greater than 1 or less than 0, which, in reality, is not possible.

In addition, note that the marginal effect of the regressor X_i on P_i is $\delta P_i/\delta X_i = \beta_2$, which is constant for any value of X_i. This may be unrealistic in some applications.

3.3 The Probit Model

In many cases, the limitations of the LPM are clearly unsatisfactory. The Probit model was developed to resolve these limitations. For simplicity, assuming that we have only one regressor X_i, the Probit model is specified as:

$$P(Y_i = 1|X_i) = F(\beta_1 + \beta_2 X_i + u_i), \tag{37}$$

where $F(x)$ is the standard normal CDF (cumulative density function), which, in turn, is given by

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz, \tag{38}$$

meaning that the Probit model can be written as

$$P(Y_i = 1|X_i) = \int_{-\infty}^{\beta_1 + \beta_2 X_i + u_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \tag{39}$$

Note that by taking the inverse of F in (37) we can write the model as

$$F^{-1}(P(Y_i = 1|X_i)) = \beta_1 + \beta_2 X_i + u_i, \tag{40}$$

where the left-hand side of (40) is the so-called **normit** $I_i = F^{-1}(P(Y_i = 1|X_i))$. Therefore, in the Probit model, the normit, but not the probability, is modelled as the linear function of X_i. Since equation (40) is linear, we can estimate it using one of the least-squares methods, such as the OLS or WLS.

They key property of F is that for any value of its argument, this function will always lie between 0 and 1, **meaning that in the Probit model, the probability is always constrained to the** $(0, 1)$ **interval**. In addition, we can show that the marginal effect of the change in X_i on the conditional probability $P(Y_i = 1|X_i)$ is given by $\delta P_i/\delta X_i = \beta_2 f(\beta_1 + \beta_2 X_i)$, where f is the standard normal PDF (probability density function) given by

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{41}$$

They key thing to note is that this marginal effect depends not only on the values of the coefficients β_1 and β_2, but is also a function of X_i, meaning that **in the Probit model, the marginal effects are not constant**. The variable marginal effects and the realistic bounds imposed on the P_i are **the two key advantages of the Probit model**.

3.4 The Logit Model

While the Probit model has many favourable properties, its key **disadvantage** is that necessitates the evaluation of integrals, which can be costly when done by hand. To resolve this difficulty, the Logit model was proposed, which is considerably simpler.

For simplicity, assuming that we only have one regressor, the Logit model is specified as

$$P(Y_i = 1|X_i) = \frac{1}{1 + exp(-(\beta_1 + \beta_2 X_i + u_i))}. \tag{42}$$

The right-hand side of (42) is the so-called **logistic function**, which has a desirable property that for any values of the betas and X_i, the function is bounded between 0 and 1. This is one of the model's **key advantages**.

3.4.1 Logit Model Properties

We can show that in the Logit model,

$$P(Y_i = 0|X_i) = 1 - P(Y_i = 1|X_i) = \frac{1}{1 + exp(\beta_1 + \beta_2 X_i + u_i)}, \tag{43}$$

and that the ratio of the probabilities is given by

$$\frac{P(Y_i = 1|X_i)}{P(Y_i = 0|X_i)} = exp(\beta_1 + \beta_2 X_i + u_i). \tag{44}$$

We call this ratio the **odds ratio**. Taking the logarithm of the odds ratio, we get **the log-odds ratio** given by

$$\ln\left(\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)}\right)=\beta_1+\beta_2X_i+u_i. \tag{45}$$

Letting $Z_i=\ln\left(\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)}\right)$ be the log-odds, we can therefore write the Logit model as

$$Z_i=\beta_1+\beta_2X_i+u_i, \tag{46}$$

meaning that **in the Logit model, the log-odds is in fact a linear function of** X_i, while the probability P_i is not.

3.4.2 Estimation of the Logit Model

Since equation (46) is linear, we should in principal be able to obtain coefficient estimates using a least-squares method such as the OLS or WLS.

Note, however, that in the case that we have **individual-level data**, $P_i=1$ whenever we observe an individual i such that $Y_i=1$ (e.g. individual is a home owner, and $P_i=0$ whenever $Y_i=0$. But then, the corresponding log-odds ratio is $Z_i=ln(1/0)=ln(\infty)=\infty$ whenever $Y_i=1$, and $Z_i=ln(0/1)=ln(0)=-\infty$ whenever $Y_i=0$. From this, we have that the error term in (46), the $u_i=\infty$ whenever $Y_i=1$ and $u_i=-\infty$ whenever $Y_i=0$, meaning that **the residual sum of squares will always be infinite in the individual data case**. Since an infinite quantity does not have a minimum, it is impossible to minimize the RSS in this case, and we therefore cannot use any of the least-squares methods with individual-level data.

Alternatively, suppose that we have **grouped data**. In particular, suppose that we bin our observations into several groups where n_j is the number of cases with $Y_i=1$, and N_i is the total number of cases in bin/group j. For each of the groups, we can therefore calculate the probability $\hat{P}_j=n_j/N_j$. Note that as long as groups are non-empty and as long as $0<n_j>N_j$ it must be that $0<\hat{P}_j<1$, meaning that the corresponding log-odds $\hat{Z}_j=\hat{P}_j/(1-\hat{P}_j)$ is finite. The resulting RSS is therefore also finite, meaning that we can apply the least-squares methods.

Suppose that we estimate (46) using **OLS** and using $\hat{Z}_j$ obtained from grouped data. Under the usual assumptions, the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are **unbiased**, but, as it turns out, **inefficient**, since it can be shown that in ((46), $var(u_j)=1/(N_jP_j(1-P_j))$. Since P_j is a function X_j, so is the $var(u_j)$, meaning that it cannot be constant. **The Logit model, therefore, suffers from inheritable heteroscedasticity**.

As with the LPM, however, we can obtain **unbiased** and **efficient** Logit estimates using WLS. Assuming that we have grouped data,

1. Calculate observed probabilities $\hat{P}_j=n_j/N_j$
2. Calculate log-odds $\hat{Z}_j=\hat{P}_j/(1-\hat{P}_j)$
3. Calculate weights $\hat{w}_i=1/(N_j\hat{P}_j(1-\hat{P}_j))$

4. Use the OLS to estimate the transformed model

$$\frac{\hat{Z}_i}{\sqrt{\hat{w}_i}} = \beta_1 \frac{1}{\sqrt{\hat{w}_i}} + \beta_2 \frac{X_i}{\sqrt{\hat{w}_i}} + \frac{u_i}{\sqrt{\hat{w}_i}}. \tag{47}$$

Under the usual assumptions, **the weighted least squares estimates** of the coefficients from (47) are **unbiased** and **efficient**.

3.4.3 Using the Logit Model

Given unbiased and efficient estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, we can use the Logit model to

- **Generate probability forecasts** Given X_i, calculate predicted probability $\hat{P}(Y_i = 1|X_i) = \dfrac{1}{1+\exp(-(\hat{\beta}_1 + \hat{\beta}_2 X_i))}$
- **Test hypothesis and build confidence intervals**
- **Estimate marginal effects** If X_i is increased by 1 unit, then the resulting
  (a) **Change in log-odds** is $\delta Z_i/\delta X_i = \hat{\beta}_2$
  (b) **Change in odds ratio (absolute terms)** is $\delta(P_i/(1-P_i))/\delta X_i = \exp(\hat{\beta}_2)$
  (c) **Change in odds ratio (in % terms)** is approximately $\exp(\hat{\beta}_2) - 1$ percent
  (d) **Change in log-odds** is $\delta Z_i/\delta X_i = \hat{\beta}_2$
  (e) **Change in the probability** is $\delta P_i/\delta X_i = \hat{\beta}_2 \hat{P}_i(1-\hat{P}_i)$

4 Panel Data Models

Panel data refers to a type of data which combine both the cross-sectional and time dimensions. In general, when we have data tracking a collection of subjects across periods of time, we say we have a panel.

One of the **key advantages** of panel data over pure cross-section or time-series is that it generally leads to higher degrees of freedom (more observations) and allows the estimation of distributional effects.

A panel is said to be **balanced** is no cross-sectional units are added to or removed from the panel throughout time. It is **unbalanced** if otherwise.

A panel is said to be a **short panel** if the size of the time dimension is greater than the size of cross-sectional dimension. Otherwise, a panel is a **long panel**.

Suppose that we have a panel (Y_{it}, X_{it}), for $i = 1, ..., n$ and $t = 1, ..., T$, representing a collection of n cross-sectional units monitored for T periods of time. As usual, the objective is to model and estimate the regression function $E[Y_{it}|X_{it}]$.

4.1 Pooled OLS

Perhaps the simplest panel data model is the pooled OLS model. They key idea is to ignore the panel nature of the data, and instead treat the dataset as a cross-section of $n \times T$ observations. The pooled model is then

$$Y_j = \beta_1 + \beta_2 X_j + u_j, \text{ for } j = 1, ..., nT. \tag{48}$$

Under the usual assumptions, estimating (48) would produce **unbiased** and **efficient** estimates. The **key issue** with the pooled OLS model, however, is that the usual assumptions are likely to be violated due to the nature of the panel data.

Let α_i, for $i = 1, ..., n$, be a **nuisance parameter** which captures unobserved heterogeneity between the units in this panel. If $\alpha_i \neq 0$, then the true model is

$$Y_j = \beta_1 + \beta_2 X_j + \alpha_j + u_j, \tag{49}$$

or letting $\tilde{u}_j = \alpha_j + u_j$,

$$Y_j = \beta_1 + \beta_2 X_j + \tilde{u}_j. \tag{50}$$

If α_j is correlated with X_j, we also have that $\tilde{u}_j$ is correlated with X_j, which is a violation of CLRM assumptions, and OLS estimates will be both **biased** and **inefficient**. Moreover, if α_j is time-invariant, the model will suffer from autocorrelation (see lecture notes).

It is therefore important to account for the presence of nuisance parameter (heterogeneity) α_j.

4.2 Least Squares Dummy Variable Model (LSDVM)

The LSDV model improves upon the pooled OLS model by trying to explicitly account for the presence of heterogeneity between the units in the panel. The key idea is to include unit

dummies $D_{1,t}, D_{2,t}, ..., D_{n-1,t}$ into the model (48). Note that to avoid the **dummy variable trap**, we have at most $n-1$ unit dummies.

The LSDV model is then specified as

$$Y_{it} = \beta_1 + \sum_{i=1}^{n-1} \alpha_i D_{i,t} + \beta_2 X_{it} + u_{it}, \text{ for } i = 1, ..., n, \text{ and } t = 1, ..., T. \tag{51}$$

Note that the slope term β_2 in this model is common to all of the units $i = 1, ..., n$, but each panel unit has own unique intercept term given by $\beta_1 + \alpha_i$, for $i = 1, ..., n-1$. The dummy coefficient α_i therefore captures the difference between the regression line intercept for panel unit i and $n-1$'th unit (see lecture notes for examples).

Under the usual assumptions, and assuming that there's no other differences between panel units other than the different regression intercepts, OLS estimates of (51) are **unbiased** and **efficient**. We call such estimates the **fixed-effects estimates**.

When α_i is the same across time, we say that we have **one-way fixed effects**. When α_{it} is allowed to vary across time, we have **two-way fixed effects**.

4.2.1 Fixed-Effects Within-Group Model (FEWGM)

Just like the LSDVM, the FEWGM accounts for heterogeneity between panel units, but in another way. Instead of the dummies, FEWGM is specified using the demeaned variables. Let $\bar{Y}_i = \sum_{t=1}^{T} Y_{it}/T$ and $\bar{X}_i = \sum_{t=1}^{T} X_{it}/T$ be the means for unit i. Then, defined **demeaned variables** as

$$y_{it} = Y_{it} - \bar{Y}_i \tag{52}$$

$$x_{it} = X_{it} - \bar{X}_i. \tag{53}$$

The FE Within-Group model is then specified as

$$y_{it} = \beta_2 x_{it} + u_{it}. \tag{54}$$

Model (54) can be estimated using the OLS, which under the usual regularity assumptions, gives **unbiased** but **inefficient** estimates, which we call **within-group (or WG) estimates**. While the sampling variances can still be used for hypothesis testing in this case, pooled OLS can give estimates which are more precise (i.e. have lower sampling variance). The inefficiency of WG estimates is one of they **key disadvantages** of the FEWGM. Another is the inability to estimate heterogeneity between panel units, since, unlike the LSDV model, FEWGM does not give an estimate of α_i.

4.2.2 Random Effects Model (REM)

Similar to the LSDVM and FEWGM, the random effects model takes the potential heterogeneity between panel units into the account. Unlike in the LSDVM, however, nuisance parameter α_i is assumed to be randomly drawn from a certain distribution and is viewed as part of the model error term. The estimates of the coefficients from the REM can be obtained using an estimator called the **panel generalized least squares, or panel GLS**. The resulting estimates are **unbiased** and **efficient**, and are called **random effects estimates**.

5 Simultaneous Equations (SIMEX) Models

In our our models considered so far, we had one dependent variable Y_i, and possibly many independent variables X_{ij}, $j = 1, .., k$. They key distinction between the two is the flow of causality: we typically assume that regressors have causal relationship with the dependent variable, but themselves are determined exogenously and independently outside of the model.

In many economic systems, causality can be harder to establish. For example, market equilibrium quantities and prices can be determined jointly and simultaneously. **Simultaneous equations models** were developed to handle such systems, and are specified as systems of multiple equations which describe the joint behaviour of a set of economic variables.

Variables which are determined jointly and simultaneously within the SIMEX model are the **endogenous variables**. Variables which are treated as given, that is, independently determined outside of the model, are **exogenous variables**.

For example, suppose that we observe equilibrium market quantity Q_t and price P_t, $t = 1, ..., T$ and wish to estimate the corresponding demand and supply curves using these data. When the market is viewed as a whole, equilibrium P_t and Q_t are jointly determined by the forces of supply and demand, and we hence view them as endogenous to the model. Regressing Q_t on P_t makes little sense, since these points do not lie along a single demand or supply curve, but rather represent intersections of multiple demand and supply curves. Generally, having an endogenous variable on the right hand side will lead to **biased** and **inefficient** estimates. We therefore need to give the model additional structure.

Let Q_t^d represent equilibrium quantity demanded, and Q_t^s quantity supplied. The following is a simultaneous equations model given by three equations:

$$Q_t^d = \beta_1 + \beta_2 P_t + u_t \tag{55}$$
$$Q_t^s = \alpha_1 + \alpha_2 P_t + v_t \tag{56}$$
$$Q_t^d = Q_t^s, \tag{57}$$

where the first equation is the market demand, the second is the market supply, and the third is the market-clearing condition. We call such model the **structural model**, and its parameters β_1, β_2, α_1 and α_2 the **structural parameters**. While we cannot estimate (55) and (56) individually due to endogeneity, the model can be simplified so that to obtain estimable equation. Since $Q_t^d = Q_t^s = Q_t$, we have that

$$\beta_1 + \beta_2 P_t + u_t = \alpha_1 + \alpha_2 P_t + u_t, \text{ hence} \tag{58}$$
$$P_t = \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{u_t - v_t}{\beta_2 - \alpha_2}. \tag{59}$$

Letting $\Pi_0 = \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2}$ and $\tilde{u}_t = \frac{u_t - v_t}{\beta_2 - \alpha_2}$, we can write (58) as

$$P_t = \Pi_0 + \tilde{u}_t. \tag{60}$$

We call (60) the **reduced-form** equation for the model. In general, a reduced-form equation expresses an endogenous variable in terms of exogenous variables only. Since (60) has no endogenous variables on the right-hand side, we can estimate the equation using OLS by

regressing P_t on just on a constant, and get an unbiased and efficient estimate $\hat{\Pi}_0 = \frac{\hat{\alpha}_1 - \hat{\beta}_1}{\hat{\beta}_2 - \hat{\alpha}_2}$. Also, note that we can substitute for P_t in (55) or (56) to get a second reduced-form equation for the model, namely

$$\begin{aligned} Q_t &= \beta_1 + \beta_2 (\Pi_0 + \tilde{u}_t) + u_t) && (61) \\ &= \beta_1 + \beta_2 \Pi_0 + \beta_2 \tilde{u}_t + u_t && (62) \\ &= \Pi_1 + \bar{u}_t, && (63) \end{aligned}$$

where $\Pi_1 = \beta_1 + \beta_2 \Pi_0 = \beta_1 + \beta_2 \left(\frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} \right)$, and $\bar{u}_t = \beta_2 \tilde{u}_t + u_t$. Once again, since (63) has no endogenous variables on the right-hand side, we can estimate it with OLS by regressing Q_t on a constant, which gives unbiased and efficient estimate $\hat{\Pi}_1 = \hat{\beta}_1 + \hat{\beta}_2 \left(\frac{\hat{\alpha}_1 - \hat{\beta}_1}{\hat{\beta}_2 - \hat{\alpha}_2} \right)$. We have a total of four structural parameters in this model, but only two equations $\hat{\Pi}_1$ and $\hat{\Pi}_0$ which we could use to solve out for for $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$. Identifying all structural parameters in this model is therefore impossible, meaning that the model is **unidentified** (or under-identified).

Suppose that in addition to the two endogenous variables P_t and Q_t, we also observe consumer's income I_t and sales taxes T_t paid by the producers, which we both treat as exogenous to this model. The model now becomes:

$$\begin{aligned} Q_t^d &= \beta_1 + \beta_2 P_t + \beta_3 I_t + u_t && (64) \\ Q_t^s &= \alpha_1 + \alpha_2 P_t + \alpha_3 T_t + v_t && (65) \\ Q_t^d &= Q_t^s. && (66) \end{aligned}$$

Re-deriving the first reduced-form, we get that

$$\beta_1 + \beta_2 P_t + \beta_3 I_t + u_t = \alpha_1 + \alpha_2 P_t + \alpha_3 T_t + u_t \quad (67)$$

and solving out for P_t we get

$$\begin{aligned} P_t &= \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{-\beta_3}{\beta_2 - \alpha_2} I_t + \frac{\alpha_3}{\beta_2 - \alpha_2} T_t + \frac{u_t - v_t}{\beta_2 - \alpha_2} && (68) \\ &= \Pi_0 + \Pi_1 I_t + \Pi_2 T_t + \tilde{u}_t, && (69) \end{aligned}$$

where $\Pi_0 = \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2}$, $\Pi_1 = \frac{-\beta_3}{\beta_2 - \alpha_2}$, $\Pi_2 = \frac{\alpha_3}{\beta_2 - \alpha_2}$ and $\tilde{u}_t = \frac{u_t - v_t}{\beta_2 - \alpha_2}$. Again, since reduced form (69) has no endogenous variables, estimating it with OLS yields unbiased and efficient estimates $\hat{\Pi}_0$, $\hat{\Pi}_1$ and $\hat{\Pi}_2$. Substituting for P_t into (64) or (65) we get

$$\begin{aligned} Q_t &= \beta_1 + \beta_2 (\Pi_0 + \Pi_1 I_t + \Pi_2 T_t + \tilde{u}_t) + \beta_3 I_t + u_t && (70) \\ &= \beta_1 + \beta_2 \Pi_0 + \beta_2 \Pi_1 I_t + \beta_2 \Pi_2 T_t + \beta_2 \tilde{u}_t + \beta_3 I_t + u_t && (71) \\ &= \Pi_3 + \Pi_4 I_t + \Pi_5 T_t + \bar{v}_t, && (72) \end{aligned}$$

where $\Pi_3 = \beta_1 + \beta_2 \Pi_0 = \beta_1 + \beta_2 \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2}$, $\Pi_4 = \beta_2 \Pi_1 + \beta_3 = \beta_2 \frac{-\beta_3}{\beta_2 - \alpha_2} + \beta_3$ and $\Pi_5 = \beta_2 \Pi_2 = \beta_2 \frac{\alpha_3}{\beta_2 - \alpha_2}$. Estimating (72) with OLS gives unbiased and efficient estimates $\hat{\Pi}_3$, $\hat{\Pi}_4$ and $\hat{\Pi}_5$,

which gives a total of six equations:

$$\hat{\Pi}_0 = \frac{\hat{\alpha}_1 - \hat{\beta}_1}{\hat{\beta}_2 - \hat{\alpha}_2} \tag{73}$$

$$\hat{\Pi}_1 = \frac{-\hat{\beta}_3}{\hat{\beta}_2 - \hat{\alpha}_2} \tag{74}$$

$$\hat{\Pi}_2 = \frac{\hat{\alpha}_3}{\hat{\beta}_2 - \hat{\alpha}_2} \tag{75}$$

$$\hat{\Pi}_3 = \hat{\beta}_1 + \hat{\beta}_2 \frac{\hat{\alpha}_1 - \hat{\beta}_1}{\hat{\beta}_2 - \hat{\alpha}_2} \tag{76}$$

$$\hat{\Pi}_4 = \hat{\beta}_2 \frac{-\hat{\beta}_3}{\hat{\beta}_2 - \hat{\alpha}_2} + \hat{\beta}_3 \tag{77}$$

$$\hat{\Pi}_5 = \hat{\beta}_2 \frac{\hat{\alpha}_3}{\hat{\beta}_2 - \hat{\alpha}_2}, \tag{78}$$

which we can use to solve uniquely for all of the structural parameters in the model. We therefore say that the model is **fully identified**.

5.1 Order Condition for Identification

In order for the model to be identified, the so-called **order condition** has to hold for every equation in the SIMEX model which contains structural parameters.

Let M be the the number of endogenous variables in the model. We can then check the order condition for an equation as follows:

(i) If an equation excludes $M - 1$ variables, it is fully-identified

(ii) If an equation excludes greater or less than $M - 1$, it is not fully-identified

If a SIMEX model is fully-identified, then every equation containing structural parameters must be fully-identified[1].

[1]FYI: The converse is actually not true, meaning that the order condition is a necessary, but not a sufficient condition for identification. In order for the model to be fully-identified, another condition called the **rank condition** has to hold. This is beyond the scope of this course, and we will assume that the rank condition holds for all models considered here.