**EC771: Econometrics, Spring 2004**

*Greene, Econometric Analysis (5th ed, 2003)*

## Chapters 2–3: Classical Linear Regression

The classical linear regression model is the single most useful tool in econometrics. Although it is often only a point of departure to more specialized methodologies, almost all empirical research will focus on the regression model as an underlying construct.

The model studies the relationship between a dependent variable and one or more independent variables, expressed as

$$y = x_1\beta_1 + x_2\beta_2 + \cdots + x_K\beta_k + \epsilon$$

where $y$ is the dependent, or endogenous variable (sometimes termed the regressand) and the $x$ variables are the independent, or exogenous variables, often termed the regressors or covariates. This of course presumes that the relationship only involves one endogenous variable, and that is in fact the setting for the classical linear regression model. If there are multiple endogenous variables in a relationship (as there would be in a demand curve, or a macro consumption function) then we must use more advanced regression techniques to deal with that endogeneity, or in economic terms simultaneity.

The term $\epsilon$ is a random disturbance, or error term; it renders the relationship between $y$ and the $x$ variables stochastic. We may rationalize the existence of a "disturbance" in several ways. It may be taken as reflecting the net effects of all the factors omitted from the specification of the regression equation: although if those factors are judged important influences, and are quantifiable, they should be included in the equation, and not omitted. A common source of stochastic variation in economic data is measurement error: that is, the relationship may appear more variable than it actually is due to the difficulty of measuring the dependent variable. (Although it is surely plausible that some of the regressors may be measured with error, that raises additional methodological issues—from an analytical standpoint, identical to those caused by simultaneity).

When we apply the regression methodology to data, we presume there is a sample of size $n$, representing observations on $y_i, x_{i1}, x_{i2}, \ldots, x_{iK}$, $i = 1, \ldots, n$. Our goal is to estimate the $K$ parameters of the regression equation, $\beta_1, \ldots, \beta_K$, as well as the error variance, $\sigma_\epsilon^2$. We may want to use the point and interval estimates of the $\beta$s to make inferences about the validity of an economic theory, or we may wish to use those estimates to generate forecasts, or predictions, from the regression equation.

*Assumptions of the classical linear regression model*

We will discuss the assumptions of the CLR model, and then consider each one in turn.

- **Linearity**: the model specifies a linear relationship between $y$ and $x_1, x_2, \ldots, x_K$.

- **Full rank**: there is no exact linear relationship among any of the independent variables of the model.

- **Strict exogeneity of $x$**: $E[\epsilon_i | x_{j1}, x_{j2}, \ldots, x_{jK}] = 0, \forall i, j$. The distribution of $\epsilon$ does not depend on past, present nor future $x$ values.

- **Spherical disturbances**: the covariance matrix of the vector $\epsilon$ is $\sigma^2 I_n$. The error terms are identically and independently distributed ($i.i.d.$).

- **Stochastic regressors**: the set of $x$ variables may include both fixed numbers and random variables, but the data generating process underlying any random $x$ is independent of that generating $\epsilon$.

- **Normally distributed errors**: for the purpose of generating interval estimates and hypothesis tests, the distribution of $\epsilon$ is assumed to be multivariate Normal.

We now turn to discussion of each assumption.

*Linearity*

Let the column vector $x_k$ be the $n$ observations on the $k^{th}$ variable. We assemble the $K$ vectors into a $n \times K$ data matrix $X$. If the regression model contains a constant term, one of the columns of $X$ is $\iota$, a vector of 1s. We may then write the multiple linear regression model as

$$y = X\beta + \epsilon$$

The vector $\beta$, of length $K$, is the primary object of regression estimation. For the model to be expressed in this linear form, it must relate $y$ to the variables in $X$ in a linear fashion with an additive error. However, that does not imply we are limited to considering models that are in a linear form: as long as they may be transformed into linear form, they may be estimated via linear regression. Consider a Cobb–Douglas function

$$y = Ax^{\beta}e^{\epsilon}$$

This model is nonlinear, but it may be transformed to linearity by taking logs. Likewise, a model relating $y$ to $1/x$ may be considered linear in $y$ and $z = 1/x$. The Cobb–Douglas form is an example of a constant elasticity model, since the slope parameter in this model is the elasticity of $y$ with respect to $x$. This transformation, in which both dependent and independent variables are replaced by their logarithms, is known as the double–log model.

The single–log model is also widely used. For instance, the growth rate model

$$y = Ae^{rt}$$

may be made stochastic by adding a term $e^{\epsilon}$. When logs are taken, this model becomes a linear relationship between $\ln y$ and $t$. The coefficient $r$ is the semi–elasticity of $y$ with respect to $t$: that is, the growth rate of $y$.

Although this sleight of hand will allow many models to be expressed in this linear form, some models cannot be written in that manner by any means. In that case, alternative estimation techniques must be employed.

*Full rank*

We assume that the data matrix $X$ is an $n \times K$ matrix with rank $K$, the number of columns (data vectors) assembled in the matrix. This rules out the situation $n < K$, implying that we must have more observations in the sample than measurements on each individual. This is not usually an issue, but can inadvertently arise when we analyze a subset of the observations.

The more common concern for the rank of $X$ is that there must be no linear dependencies among its columns: that is, no column may be expressed as linearly dependent upon the other columns. From the analytical standpoint, this is an argument that these $K$ vectors must span $K$–space, and if we are to solve $K$ equations for $K$ unknowns (the $\beta$ coefficients) those equations must not be redundant. If one or more columns of $X$ were linearly dependent, redundancies would exist. In the context of econometrics, we call the existence of a linear dependency perfect collinearity. In an intuitive sense, it indicates that a particular column of $X$ (an "explanatory variable") contains no information at the margin, since it itself may be perfectly explained by the remaining explanatory variables. A model with $K$ explanatory variables must contain $K$ distinct sources of information: both conceptually and numerically. Any situation in which there is an identity among explanatory variables, or an adding–up condition, may bring about perfect collinearity. In this case, one of the explanatory variables in the identity is clearly redundant, and not all cannot be included in the model as independent sources of variation.

*Strict exogeneity of $x$*

The disturbance is assumed to have conditional expected value of zero at every observation:

$$E[\epsilon_i|X] = 0.$$

Intuitively, this states that no observations on $X$ contain any useful information about the expected value of the disturbance for a given observation: the assumption of strict exogeneity of the $X$ variables. In the context of time series data, we may have to relax that assumption, and presume that $X$ values at some other point in time may convey information about the expected value of the disturbance at time $t$: that is, the $X$ variables are only weakly exogenous. Nevertheless, even weak exogeneity implies that current values of the $X$ variables are not informative about $\epsilon_t$.

Generally the assumption that this conditional mean is zero is not restrictive; as long as there is a constant term in the regression relationship (with a corresponding $\iota$ vector in $X$), any nonzero mean of the disturbances may be transformed to zero by adjusting the constant term. However, this implies that a constant term should generally be included in a regression model. (An exception to this rule: if a time series model is fit to first differences of the original data, it will only contain a constant term if the levels model contained a linear trend term).

*Spherical disturbances*

We assume that the distribution of $\epsilon$ is spherical: combining the two assumptions that $Var[\epsilon_i|X] = \sigma^2, \forall i$ and $Cov[\epsilon_i, \epsilon_j|X] = 0, \forall i \neq j$. These two assumptions combined imply that the disturbances are identically and independently distributed ($i.i.d.$), with a covariance matrix $\sigma^2 I$. The first assumption is that of homoskedasticity: the variance of $\epsilon$, conditioned on $X$, is identical over all observations. Violation of this assumption would imply that the errors are heteroskedastic.

The second assumption is independence of the errors, or in the context of time series data, that the errors are not serially correlated or autocorrelated. Although there is a natural connotation to this in the time series context, it may also occur in cross–sectional data: e.g., those individuals who live in the same neighborhood may have the same unusual behavioral traits. This, in fact, is addressed in Stata by the `cluster` option on many estimation commands, which allows for "neighborhood effects".

*Stochastic regressors*

In developing the regression model, it is common to assume that the $x_i$ are nonstochastic, as they would be in an experimental setting. This would simplify the assumptions above on exogeneity and the distribution of the errors, since then $X$ would be considered a matrix of fixed constants. But in social science research, we rarely work with data from the experimental setting. If we consider that some of the regressors in our model are stochastic, then our assumption concerns the nature of the data generating process than produces $x_i$. That process must be independent of the data generating process underlying the errors if the classical regression model is to be applied.

*Normally distributed errors*

It is convenient to assume that the conditional distribution of $\epsilon$ is multivariate Normal. Normality is very useful in constructing test statistics and confidence intervals from the regression model, although it is not necessary for the solution of the estimation problem.

*The least squares methodology*

In applying the least squares methodology, we note that the unknown parameters are those in the vector $\beta$ in $y = X\beta + \epsilon$. The population regression function is $E[y_i|x_i] = x_i'\beta$, whereas our estimate of that conditional mean is denoted $\widehat{y}_i = x_i'b$: that is, the vector of estimated parameters is $b$. The error, or disturbance, associated with the $i^{th}$ data point is $\epsilon_i = y_i - x_i'\beta$, while the associated regression residual is $e_i = y_i - x_i'b$. The sample is $\{y_i, x_i\}, i = 1, \ldots, n$. How might we choose a vector $b$ so that the fitted line, $x_i'b$, is close to the data points $y_i$? An obvious choice is the least squares criterion.

The least squares coefficient vector minimizes the sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i = x_i'b)^2$$

with respect to the elements of $b$. In matrix terms, we have

$$\min S = e'e = (y - Xb)'(y - Xb) = y'y - 2y'Xb + b'X'Xb$$

with first order conditions

$$\partial S / \partial b = -2X'y + 2X'Xb = 0.$$

These are the least squares normal equations,

$$X'Xb = X'y$$

with solution

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

For the solution to be a minimum, the second derivatives must form a positive definite matrix: $2X'X$. If the matrix $X$ has full rank, then this condition will be satisfied: the least squares solution will be unique, and a minimum of the sum of squared residuals. If $X$ is rank–deficient, the solution will not exist, as $X'X$ will be singular. In that instance, we cannot uniquely solve for all $K$ elements of $b$, since $X$ contains one or more linear dependencies.

Let us consider the normal equations in ordinary algebra for a "three–variable" regression problem: one in which we regress $y$ on two regressors and a constant term. Assume that the two regressors are named $T$ and $G$. Then the three normal equations are:

$$\begin{aligned} b_1 n + b_2 \sum T + b_3 \sum G &= \sum Y \\ b_1 \sum T + b_2 \sum T^2 + b_3 \sum TG &= \sum TY \\ b_1 \sum G + b_2 \sum TG + b_3 \sum G^2 &= \sum TG \end{aligned}$$

A system of three equations in three unknowns. Note that the first normal equation, if we divide through by $n$, becomes:

$$b_1 + b_2 \bar{T} + b_3 \bar{G} = \bar{Y}$$

That is, in the presence of a constant term, the regression surface passes through the multivariate point of means (which we may readily illustrate in terms of a two–variable regression).

We may solve the regression problem as two equations in two unknowns, the slopes $b_2$ and $b_3$, by demeaning each series: expressing it as deviations from its own mean. In that context, the second and third normal equations become functions of those two parameters only, since $b_1$ becomes zero for the demeaned series. Note that these equations do not depend on the raw data, but only on the sums of squares and cross products of the demeaned data. After solving these two simultaneous equations for $b_2$ and $b_3$, we may backsolve for $b_1$.

It is illustrative to consider this strategy in the two–variable or bivariate regression of $y$ on a single $x$ series. The slope parameter based on the demeaned series is

$$b_2 = \frac{\sum xy}{\sum x^2} = \frac{Cov(x, y)}{Var(x)} = r_{XY}\frac{s_y}{s_x}$$

where $r_{XY}$ is the simple correlation between $X$ and $Y$, and $s_Y$ ($s_X$) is the sample standard deviation of $Y$ $(X)$. This gives us the intercept

$$b_1 = \bar{Y} - b_2\bar{X}$$

This is known as "simple" regression, in which we have only one regressor (beyond $\iota$). In multiple regression, the slope coefficients will generally differ from those of simple regression. The simple regression coefficients are total derivatives of $Y$ with respect to $X$, whereas the multiple regression coefficients are partial derivatives. So, for instance, a solution for the aforementioned "three–variable" regression problem will yield

$$b_{yg\cdot t} = \frac{b_{yg}}{1 - r_{gt}^2} - \frac{b_{yt}b_{tg}}{1 - r_{gt}^2}$$

where $b_{yg\cdot t}$ is the multiple regression coefficient of $y$ on $g$, holding $t$ fixed, and $(r_{gt}^2)$ is the squared simple correlation between $g$ and $t$. Note that if this correlation is 1 or -1, the multiple regression cannot be computed ($X$ fails the full rank assumption). Note that

$b_{yg \cdot t}$ will differ from the simple regression coefficient $b_{yg}$ due to the second term. The second term could be zero for two reasons: (1) $t$ does not influence $y$, so $b_{yt}$ is effectively zero; or (2) $g$ and $t$ are uncorrelated, so that $b_{tg}$ is zero. In the latter case, the squared correlation coefficient is also zero, and the formula for the multiple regression coefficient becomes that of the simple regression coefficient. Conversely, as long as there is correlation among the regressors, the multiple regression coefficients will differ from simple regression coefficients, as "partialling off" the effects of the other regressors will matter. Thus, if the regressors are orthogonal, the multiple regression coefficients will equal the simple regression coefficients. Generally, we will not find orthogonal regressors in economic data except by construction: for instance, binary variables indicating that a given individual is M or F will be orthogonal to one another, since their dot product is zero.

Some results from the least squares solution of a model with a constant term:

- The least squares residuals sum to zero: the first normal equation $\bar{y} = \bar{x}'b$ implies that $\bar{e}$ is identically zero.

- The normal equations $(X'X)b - X'y = 0$ imply $-X'(y - Xb) = 0$, or $-X'e = 0$. This indicates that the residuals are constructed to be orthogonal to each column of X.

- If $\widehat{y} = Xb$ are the predicted values of the regression, $\bar{\widehat{y}} = \bar{y}$. The mean of the fitted values equals the mean of the original data (again, since $\bar{e} = 0$).

*Regression as a projection*

The vector of least squares residuals is $e = y - Xb$, or given the solution of the normal equations,

$$e = y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y = My$$

where $M$ is known as the fundamental idempotent matrix of least squares. $M$ produces the vector of LS residuals from the regression of $y$ on $X$ when it premultiplies any vector $y$. That also implies that $MX = 0$, since the residuals are constructed to be orthogonal to $X$.

Least squares partitions $y$ into two parts: that which is related to $X$, $Xb$, and that which is unrelated to $X$, $e$. Since $MX = 0$, the two parts are orthogonal. We can thus write the predicted values $\widehat{y} = y - e$ as

$$\widehat{y} = (I - M)y = X(X'X)^{-1}X'y = Py$$

where the projection matrix $P$, which is also symmetric and idempotent, generates the fitted values as the projection of $y$ on $X$, which is also the projection of $y$ into the column space of $X$. Since $M$ produces the residuals, and $P$ produces the fitted values, it follows that $M$ and $P$ are orthogonal: $PM = MP = 0$, and by simple algebra $PX = X$. We can then write $y = Py + My$, and consider the least squares problem in that context:

$$y'y = y'P'Py + y'M'My$$

or

$$y'y = \widehat{y}'\widehat{y} + e'e$$

The orthogonality of $P$ and $M$ imply that the cross–product terms that would result from that expansion are zero. Given the definition of $\widehat{y}$, we can also write

$$e'e = y'y - b'X'Xb = y'y - b'X'y.$$

*Goodness of fit and ANOVA*

Although the least squares criterion gives us a metric defining a line (surface) of best fit, the magnitude of the minimized criterion is arbitrary: the residuals are in the same units as $y$, and altering the scale of $y$ will change the magnitude of $e'e$. To judge the goodness of fit of a regression model, we consider whether the variation in the $X$ variables can explain some meaningful amount of the variation in $y$. When a constant term is included in the relationship, we are explaining the variation in $y$ about its mean. That is, the $X$ variables are not explaining why GDP is, on average, $4 trillion. Rather, we are attempting to explain the variation in GDP around that mean value. The total variation in $y$ is the sum of squared deviations:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

which is a component of the variance of $y$ (but is not that variance). Write the regression equation as

$$y = Xb + e = \hat{y} + e$$

and premultiply by

$$M^0 = [I - n^{-1}\iota\iota'],$$

the idempotent matrix which transforms variables into deviations from their own means:

$$M^0 y = M^0 X b + M^0 e$$

If we now square this equation, we find that

$$y' M^0 y = b' X' M^0 X b + e' e$$

keeping in mind that $M^0$ is idempotent, the residuals (in the presence of a constant term) have mean zero, and cross product terms vanish since $e' M^0 X = 0$. The left side of this expression is merely SST, the total sum of squares. The right side may be written as regression sum of squares, SSR, and error sum of squares, SSE. The latter is, of course, the minimized value of the least squares criterion. This identity, $SST = SSR + SSE$, corresponds to the notion that regression partitions $y$ into two orthogonal components: that explained by the regressors and that left unexplained. It is the basis for our measure of goodness of fit:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

or

$$R^2 = \frac{b'X'M^0Xb}{y'M^0y} = 1 - \frac{e'e}{y'M^0y}$$

This is known as the coefficient of determination, and in the presence of a constant term, it must lie between 0 and 1. It is, indeed, "r–squared", the squared correlation of $y$ and $\hat{y}$. In a "two–variable" regression model, since $\hat{y}$ is a linear transformation of the single $x$, it is also the squared correlation of $y$ and $x$. In a multiple regression model, $R^2$ bears no simple relationship to the simple $R^2$ measures for each regressor, but reflects the overlap in their explanation of $y$. An $R^2$ of zero indicates that the naive model $y_i = \mu + \epsilon_i$ cannot be improved with the set of regressors in use, while an $R^2$ of unity indicates that all residuals are zero.

The elements that go into the computation of $R^2$ are usually presented in computer output as the ANOVA (analysis of variance) table:

```
. reg price mpg headroom trunk weight length turn
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | **277.845193** | **6** | **46.3075321** |
| Residual | **357.220189** | **67** | **5.33164461** |
| Total | **635.065382** | **73** | **8.69952578** |

| | |
|---|---|
| Number of obs = | **74** |
| F( 6, 67) = | **8.69** |
| Prob > F = | **0.0000** |
| R-squared = | **0.4375** |
| Adj R-squared = | **0.3871** |
| Root MSE = | **2.309** |

| price | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| mpg | **-.0940633** | **.0803708** | **-1.17** | **0.246** | **-.2544842** | **.0663575** |
| headroom | **-.7312921** | **.4273691** | **-1.71** | **0.092** | **-1.584324** | **.1217401** |
| trunk | **.0982751** | **.1057208** | **0.93** | **0.356** | **-.1127446** | **.3092947** |
| weight | **.0050793** | **.0011482** | **4.42** | **0.000** | **.0027876** | **.007371** |
| length | **-.0734871** | **.0430113** | **-1.71** | **0.092** | **-.1593379** | **.0123638** |
| turn | **-.3270699** | **.1263111** | **-2.59** | **0.012** | **-.5791879** | **-.0749519** |
| _cons | **20.44725** | **6.090068** | **3.36** | **0.001** | **8.291423** | **32.60308** |

The "SS" in the table are the sums of squares in our identity above. Their mean squares are derived from their respective degrees of freedom. Although the $R^2$ is not a basis for any statistical judgment, it may be transformed into the "ANOVA $F$", which is a test of the model versus the naive model considered above: i.e. that all slopes are jointly zero. You will note that the $F$ statistic is computed from the same quantities as $R^2$, and is in fact the ratio of mean squares (MS) due to regression and error, respectively.

What are the difficulties in using $R^2$? For one thing, just as $e'e$ cannot rise when an additional regressor is added to the model, $R^2$ cannot fall. Thus a model with a large number of regressors (and the same dependent variable) will always have a higher $R^2$. To compensate for this, we often consider the adjusted $R^2$ or $\bar{R}^2$ value:

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2).$$

Note that $\bar{R}^2$ will always be less than $R^2$, and no longer has the connotation of a squared correlation coefficient; indeed, it may become negative. This measure weighs the cost of adding a regressor (the use of one more degree of freedom) against the benefit of the reduction in error sum of squares. Unless the latter is large enough to outweigh the former, $\bar{R}^2$ will indicate that a "longer" model does worse than a more parsimonious specification, even though the longer model will surely have a higher $R^2$ value. However, one cannot attach any statistical significance to movements in $\bar{R}^2$, since it can readily be shown that it will rise or fall when a single variable is added depending on that variable's $t$-statistic being greater or less than 1 in absolute value.

A second diffficulty with conventional $R^2$ relates to the constant term in the model. For $R^2$ to lie in the unit interval, the $X$ matrix must have a column $\iota$. In the absence of a constant term, $R^2$ can readily be negative (as we can easily illustrate with an improperly constrained two–variable model). Different computer programs generate different results for $R^2$ in the absence of a constant term; some omit it entirely, while others will provide an alternate formula in that context. Generally, one should not refer to $R^2$ in a model without a constant term.

One should also note that the level of $R^2$ depends upon the context. When fitting models to aggregate time–series data, $R^2$ values above 0.90 are quite common. When fitting models to individual microdata, an $R^2$ of 0.10 might be reason for rejoicing. When comparing several models, one must also be careful to ensure that the dependent variable is identical across models: not only the same observations, but in the same form. One cannot reasonably compare a levels equation with an equation fit in first differences or logarithms of the same dependent variable. It should also be recalled that $R^2$, as a squared correlation coefficient, is a measure of linear association between $y$ and $\hat{y}$. Two (or more) variables may be linked by a relationship–even in a deterministic sense–which is not linear. In that case, a measure of linear association, such as correlation, may not detect the relationship. As an example, consider $x^2 + y^2 = k$. A sample of $\{x, y\}$ points generated by that relationship will yield an $R^2$ of zero, even though the relationship is nonstochastic.