1

INTRODUCTION TO REGRESSION ANALYSIS

1.1 INTRODUCTION

The general purpose of regression analysis is to study the relationship between one or more dependent variable(s) and one or more independent variable(s). The most basic form of a regression model is where there is one independent variable and one dependent variable. For instance, a model relating the log of wage of married women to their experience in the work force is a simple linear regression model given by $\log(\text{wage}) = \beta_0 + \beta_1 \text{exper} + \varepsilon$, where β_0 and β_1 are unknown coefficients and ε is random error. One objective here is to determine what effect (if any) the variable exper has on wage. In practice, most studies involve cases where there is more than one independent variable. As an example, we can extend the simple model relating log(wage) to exper by including the square of the experience (exper^2) in the work force, along with years of education (educ). The objective here may be to determine what effect (if any) the explanatory variables (exper, exper^2, educ) have on the response variable log(wage). The extended model can be written as

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + \varepsilon,$$

where β_0, β_1, β_2, and β_3 are the unknown coefficients that need to be estimated, and ε is random error.

An extension of the multiple regression model (with one dependent variable) is the multivariate regression model where there is more than one dependent variable. For instance, the well-known Grunfeld investment model deals with the relationship between investment (I_{it}) with the true market value of a firm (F_{it}) and the value of capital (C_{it}) (Greene, 2003). Here, i indexes the firms and t indexes time. The model is given by $I_{it} = \beta_{0i} + \beta_{1i} F_{it} + \beta_{2i} C_{it} + \varepsilon_{it}$. As before, β_{0i}, β_{1i}, and β_{2i} are unknown coefficients that need to be estimated and ε_{it} is random error. The objective here is to determine if the disturbance terms are involved in cross-equation correlation. Equation by equation ordinary least squares is used to estimate the model parameters if the disturbances are not involved in cross-equation correlations. A feasible generalized least squares method is used if there is evidence of cross-equation correlation. We will look at this model in more detail in our discussion of seemingly unrelated regression models (SUR) in Chapter 8.

Dependent variables can be continuous or discrete. In the Grunfeld investment model, the variable I_{it} is continuous. However, discrete responses are also very common. Consider an example where a credit card company solicits potential customers via mail. The response of the consumer can be classified as being equal to 1 or 0 depending on whether the consumer chooses to respond to the mail or not. Clearly, the outcome of the study (a consumer responds or not) is a discrete random variable. In this example, the response is a binary random variable. We will look at modeling discrete responses when we discuss discrete choice models in Chapter 10.

In general, a multiple regression model can be expressed as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon, \tag{1.1}$$

Applied Econometrics Using the SAS® System, by Vivek B. Ajmani

where y is the dependent variable, $\beta_0, \ldots, \beta_k$ are the $k+1$ unknown coefficients that need to be estimated, $x_1, \ldots, x_k$ are the k independent or explanatory variables, and ε is random error. Notice that the model is linear in parameters $\beta_0, \ldots, \beta_k$ and is therefore called a linear model. Linearity refers to how the parameters enter the model. For instance, the model $y = \beta_0 + \beta_1 x_1^2 + \cdots + \beta_k x_k^2 + \varepsilon$ is also a linear model. However, the exponential model $y = \beta_0 \exp(-x\beta_1)$ is a nonlinear model since the parameter β_1 enters the model in a nonlinear fashion through the exponential function.

1.1.1 Interpretation of the Parameters

One of the assumptions (to be discussed later) for the linear model is that the conditional expectation $E(\varepsilon|x_1, \ldots, x_k)$ equals zero. Under this assumption, the expectation, $E(y|x_1, \ldots, x_k)$ can be written as $E(y|x_1, \ldots, x_k) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$. That is, the regression model can be interpreted as the conditional expectation of y for given values of the explanatory variables $x_1, \ldots, x_k$. In the Grunfeld example, we could discuss the expected investment for a given firm for known values of the firm's true market value and value of its capital. The intercept term, β_0, gives the expected value of y when all the explanatory variables are set at zero. In practice, this rarely makes sense since it is very uncommon to observe values of all the explanatory variables equal to zero. Furthermore, the expected value of y under such a case will often yield impossible results. The coefficient β_k is interpreted as the expected change in y for a unit change in x_k holding all other explanatory variables constant. That is, $\partial E(y|x_1, \ldots, x_k)/\partial x_k = \beta_k$.

The requirement that all other explanatory variables be held constant when interpreting a coefficient of interest is called the *ceteris paribus condition*. The effect of x_k on the expected value of y is referred to as the marginal effect of x_k.

Economists are typically interested in *elasticities* rather than marginal effects. Elasticity is defined as the relative change in the dependent variable for a relative change in the independent variable. That is, elasticity measures the responsiveness of one variable to changes in another variable—the greater the elasticity, the greater the responsiveness.

There is a distinction between marginal effect and elasticity. As stated above, the marginal effect is simply $\partial E(y|\mathbf{x})/\partial x_k$ whereas elasticity is defined as the ratio of the percentage change in y to the percentage change in x. That is, $e = (\partial y/y)/(\partial x_k/x_k)$.

Consider calculating the elasticity of x_1 in the general regression model given by Eq. (1.1). According to the definition of elasticity, this is given by $e_{x_1} = (\partial y/\partial x_1)(x_1/y) = \beta_1(x_1/y) \neq \beta_1$. Notice that the marginal effect is constant whereas the elasticity is not. Next, consider calculating the elasticity in a log–log model given by $\log(y) = \beta_0 + \beta_1 \log(x) + \varepsilon$. In this case, elasticity of x is given by

$$\partial \log(y) = \beta_1 \partial \log(x) \Rightarrow \partial y \frac{1}{y} = \beta_1 \partial x \frac{1}{x} \Rightarrow \frac{\partial y}{\partial x}\frac{x}{y} = \beta_1.$$

The marginal effect for the log–log model is also β_1. Next, consider the semi-log model given by $y = \beta_1 + \beta_1 \log(x) + \varepsilon$. In this case, elasticity of x is given by

$$\partial y = \beta_1 \partial \log(x) \Rightarrow \partial y = \beta_1 \partial x \frac{1}{x} \Rightarrow \frac{\partial y}{\partial x}\frac{x}{y} = \beta_1 \frac{1}{y}.$$

On the other hand, the marginal effect in the semi-log model is given by $\beta_1(1/x)$.

For the semi-log model given by $\log(y) = \beta_0 + \beta_1 x + \varepsilon$, the elasticity of x is given by

$$\partial y \log(y) = \beta_1 \partial x \Rightarrow \partial y \frac{1}{y} = \beta_1 \partial x = \frac{\partial y}{\partial x}\frac{x}{y} = \beta_1 x.$$

On the other hand, the marginal effect in the semi-log model is given by $\beta_1 y$.

Most models that appear in this book have a log transformation on the dependent variable or the independent variable or both. It may be useful to clarify how the coefficients from these models are interpreted. For the semi-log model where the dependent variable has been transformed using the log transformation while the explanatory variables are in their original units, the coefficient β is interpreted as follows: For a one unit change in the explanatory variable, the dependent variable changes by $\beta \times 100\%$ holding all other explanatory variables constant.

In the semi-log model where the explanatory variable has been transformed by using the log transformation, the coefficient β is interpreted as follows: For a one unit change in the explanatory variable, the dependent variable increases (decreases) by $\beta/100$ units.

In the log–log model where both the dependent and independent variable have been transformed by using a log transformation, the coefficient β is interpreted as follows: A 1% change in the explanatory variable is associated with a β% change in the dependent variable.

1.1.2 Objectives and Assumptions in Regression Analysis

There are three main objectives in any regression analysis study. They are

a. To estimate the unknown parameters in the model.
b. To validate whether the functional form of the model is consistent with the hypothesized model that was dictated by theory.
c. To use the model to predict future values of the response variable, y.

Most regression analysis in econometrics involves objectives (a) and (b). Econometric time series analysis involves all three. There are five key assumptions that need to be checked before the regression model can be used for the purposes outlined above.

a. *Linearity:* The relationship between the dependent variable y and the independent variables $x_1, \ldots, x_k$ is linear.
b. *Full Rank:* There is no linear relationship among any of the independent variables in the model. This assumption is often violated when the model suffers from multicollinearity.
c. *Exogeneity of the Explanatory Variables:* This implies that the error term is independent of the explanatory variables. That is, $E(\varepsilon_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$. This assumption states that the underlying mechanism that generated the data is different from the mechanism that generated the errors. Chapter 4 deals with alternative methods of estimation when this assumption is violated.
d. *Random Errors:* The errors are random, uncorrelated with each other, and have constant variance. This assumption is called the homoscedasticity and nonautocorrelation assumption. Chapters 5 and 6 deal with alternative methods of estimation when this assumption is violated. That is estimation methods when the model suffers from heteroscedasticity and serial correlation.
e. *Normal Distribution:* The distribution of the random errors is normal. This assumption is used in making inference (hypothesis tests, confidence intervals) to the regression parameters but is not needed in estimating the parameters.

1.2 MATRIX FORM OF THE MULTIPLE REGRESSION MODEL

The multiple regression model in Eq. (1.1) can be expressed in matrix notation as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Here, $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is a $n \times (k+1)$ matrix containing values of explanatory variables, $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector of coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors. Note that $\mathbf{X}$ consists of a column of 1's for the intercept term $\boldsymbol{\beta}_0$. The regression analysis assumptions, in matrix notation, can be restated as follows:

a. *Linearity*: $\mathbf{y} = \boldsymbol{\beta}_0 + \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_k\beta_k + \boldsymbol{\varepsilon}$ or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
b. *Full Rank:* $\mathbf{X}$ is an $n \times (k+1)$ matrix with rank $(k+1)$.
c. *Exogeneity:* $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0} - \mathbf{X}$ is uncorrelated with $\boldsymbol{\varepsilon}$ and is generated by a process that is independent of the process that generated the disturbance.
d. *Spherical Disturbances:* $Var(\varepsilon_i|\mathbf{X}) = \sigma^2$ for all $i = 1, \ldots, n$ and $Cov(\varepsilon_i, \varepsilon_j|\mathbf{X}) = 0$ for all $i \neq j$. That is, $Var(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}$.
e. *Normality:* $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

1.3 BASIC THEORY OF LEAST SQUARES

Least squares estimation in the simple linear regression model involves finding estimators b_0 and b_1 that minimize the sums of squares $L = \Sigma_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. Taking derivatives of L with respect to β_0 and β_1 gives

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i),$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i.$$

Equating the two equations to zero and solving for β_0 and β_1 gives

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i,$$
$$\sum_{i=1}^{n} y_i x_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2.$$

These two equations are known as normal equations. There are two normal equations and two unknowns. Therefore, we can solve these to get the ordinary least squares (OLS) estimators of β_0 and β_1. The first normal equation gives the estimator of the intercept, β_0, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Substituting this in the second normal equation and solving for $\hat{\beta}_1$ gives

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

We can easily extend this to the multiple linear regression model in Eq. (1.1). In this case, least squares estimation involves finding an estimator $\mathbf{b}$ of $\boldsymbol{\beta}$ to minimize the error sums of squares $L = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Taking the derivative of L with respect to $\boldsymbol{\beta}$ yields $k+1$ normal equations with $k+1$ unknowns (including the intercept) given by

$$\partial L / \partial \boldsymbol{\beta} = -2(\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}).$$

Setting this equal to zero and solving for $\hat{\boldsymbol{\beta}}$ gives the least squares estimator of $\boldsymbol{\beta}$, $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. A computational form for $\mathbf{b}$ is given by

$$\mathbf{b} = \left(\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i^T y_i\right).$$

The estimated regression model or predicted value of $\mathbf{y}$ is therefore given by $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$. The residual vector $\mathbf{e}$ is defined as the difference between the observed and the predicted value of $\mathbf{y}$, that is, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.

The method of least squares produces unbiased estimates of $\boldsymbol{\beta}$. To see this, note that

$$\begin{aligned} E(\mathbf{b}|\mathbf{X}) &= E((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{y}|\mathbf{X}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}|\mathbf{X}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} E(\boldsymbol{\varepsilon}|\mathbf{X}) \\ &= \boldsymbol{\beta}. \end{aligned}$$

Here, we made use of the fact that $(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{X}) = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix and the assumption that $E(\boldsymbol{\varepsilon}|\mathbf{X}) = 0$.

1.3.1 Consistency of the Least Squares Estimator

First, note that a consistent estimator is an estimator that converges in probability to the parameter being estimated as the sample size increases. To say that a sequence of random variables X_n converges in probability to X implies that as $n \to \infty$ the probability that $|X_n - X| \geq \delta$ is zero for all δ (Casella and Berger, 1990). That is,

$$\lim_{n \to \infty} \Pr(|X_n - X| \geq \delta) = 0 \,\forall\, \delta.$$

Under the exogeneity assumption, the least squares estimator is a consistent estimator of $\boldsymbol{\beta}$. That is,

$$\lim_{n \to \infty} \Pr(|\mathbf{b}_n - \boldsymbol{\beta}| \geq \delta) = 0 \,\forall\, \delta.$$

To see this, let $\mathbf{x}_i$, $i = 1, \ldots, n$, be a sequence of independent observations and assume that $\mathbf{X}^T\mathbf{X}/n$ converges in probability to a positive definite matrix $\boldsymbol{\Psi}$. That is (using the probability limit notation),

$$p \lim_{n \to \infty} \frac{\mathbf{X}^T\mathbf{X}}{n} = \boldsymbol{\Psi}.$$

Note that this assumption allows the existence of the inverse of $\mathbf{X}^T\mathbf{X}$. The least squares estimator can then be written as

$$\mathbf{b} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}^T\boldsymbol{\varepsilon}}{n}\right).$$

Assuming that $\boldsymbol{\Psi}^{-1}$ exists, we have

$$p\lim \mathbf{b} = \boldsymbol{\beta} + \boldsymbol{\Psi}^{-1} p\lim\left(\frac{\mathbf{X}^T\boldsymbol{\varepsilon}}{n}\right).$$

In order to show consistency, we must show that the second term in this equation has expectation zero and a variance that converges to zero as the sample size increases. Under the exogeneity assumption, it is easy to show that $E(\mathbf{X}^T\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$ since $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$. It can also be shown that the variance of $\mathbf{X}^T\boldsymbol{\varepsilon}/n$ is

$$Var\left(\frac{\mathbf{X}^T\boldsymbol{\varepsilon}}{n}\right) = \frac{\sigma^2}{n}\boldsymbol{\Psi}.$$

Therefore, as $n \to \infty$ the variance converges to zero and thus the least squares estimator is a consistent estimator for $\boldsymbol{\beta}$ (Greene, 2003, p. 66).

Moving on to the variance–covariance matrix of $\mathbf{b}$, it can be shown that this is given by

$$Var(\mathbf{b}|\mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

To see this, note that

$$\begin{aligned} Var(\mathbf{b}|\mathbf{X}) &= Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}) \\ &= Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})|\mathbf{X}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(\boldsymbol{\varepsilon}|\mathbf{X})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned}$$

It can be shown that the least squares estimator is the best linear unbiased estimator of $\boldsymbol{\beta}$. This is based on the well-known result, called the Gauss–Markov Theorem, and implies that the least squares estimator has the smallest variance in the class of all unbiased estimators of $\boldsymbol{\beta}$ (Casella and Berger, 1990; Greene, 2003; Meyers, 1990).

An estimator of σ^2 can be obtained by considering the sums of squares of the residuals (SSE). Here, $SSE = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})$. Dividing *SSE* by its degrees of freedom, $n-k-1$ yields $\hat{\sigma}^2$. That is, the mean square error is given by $\hat{\sigma}^2 = MSE = SSE/(n-k-1)$. Therefore, an estimate of the covariance matrix of $\mathbf{b}$ is given by $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$.

Using a similar argument as the one used to show consistency of the least squares estimator, it can be shown that $\hat{\sigma}^2$ is consistent for σ^2 and that the asymptotic covariance matrix of $\mathbf{b}$ is $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$ (see Greene, 2003, p. 69 for more details). The square root of the diagonal elements of this yields the standard errors of the individual coefficient estimates.

1.3.2 Asymptotic Normality of the Least Squares Estimator

Using the properties of the least squares estimator given in Section 1.3 and the Central Limit Theorem, it can be easily shown that the least squares estimator has an asymptotic normal distribution with mean $\boldsymbol{\beta}$ and variance–covariance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. That is, $\hat{\boldsymbol{\beta}} \sim asym.N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$.

1.4 ANALYSIS OF VARIANCE

The total variability in the data set (*SST*) can be partitioned into the sums of squares for error (*SSE*) and the sums of squares for regression (*SSR*). That is, $SST = SSE + SSR$. Here,

$$SST = \mathbf{y}^T\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n},$$

$$SSE = \mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y},$$

$$SSR = \mathbf{b}^T\mathbf{X}^T\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

TABLE 1.1. Analysis of Variance Table

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Square F_0
Regression	*SSR*	k	$MSR = SSR/k$
Error	*SSE*	$n-k-1$	$MSE = SSE/(n-k-1)$
Total	*SST*	$n-1$	*MSR/MSE*

The mean square terms are simply the sums of square terms divided by their degrees of freedom. We can therefore write the analysis of variance (ANOVA) table as given in Table 1.1.

The F statistic is the ratio between the mean square for regression and the mean square for error. It tests the global hypotheses

$$\begin{aligned} H_0&: \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0, \\ H_1&: \quad \text{At least one } \beta_i \neq 0 \quad \text{for } i = 1, \ldots, k. \end{aligned}$$

The null hypothesis states that there is no relationship between the explanatory variables and the response variable. The alternative hypothesis states that at least one of the k explanatory variables has a significant effect on the response. Under the assumption that the null hypothesis is true, F_0 has an F distribution with k numerator and $n-k-1$ denominator degrees of freedom, that is, under H_0, $F_0 \sim F_{k,n-k-1}$. The p value is defined as the probability that a random variable from the F distribution with k numerator and $n-k-1$ denominator degrees of freedom exceeds the observed value of F, that is, $\Pr(F_{k,n-k-1} > F_0)$. The null hypothesis is rejected in favor of the alternative hypothesis if the p value is less than α, where α is the type I error.

1.5 THE FRISCH–WAUGH THEOREM

Often, we may be interested only in a subset of the full set of variables included in the model. Consider partitioning $\mathbf{X}$ into $\mathbf{X}_1$ and $\mathbf{X}_2$. That is, $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$. The general linear model can therefore be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$. The normal equations can be written as (Greene, 2003, pp. 26–27; Lovell, 2006)

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2 \\ \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T\mathbf{y} \\ \mathbf{X}_2^T\mathbf{y} \end{bmatrix}.$$

It can be shown that

$$\mathbf{b}_1 = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2).$$

If $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$, then $\mathbf{b}_1 = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}$. That is, if the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal, then $\mathbf{b}_1$ can be obtained by regressing $\mathbf{y}$ on $\mathbf{X}_1$. Similarly, $\mathbf{b}_2$ can be obtained by regressing $\mathbf{y}$ on $\mathbf{X}_2$. It can easily be shown that

$$\mathbf{b}_2 = (\mathbf{X}_2^T\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2^T\mathbf{M}_1\mathbf{y}),$$

where $\mathbf{M}_1 = (\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T)$ so that $\mathbf{M}_1\mathbf{y}$ is a vector of residuals from a regression of $\mathbf{y}$ on $\mathbf{X}_1$.

Note that $\mathbf{M}_1\mathbf{X}_2$ is a matrix of residuals obtained by regressing $\mathbf{X}_2$ on $\mathbf{X}_1$. The computations described here form the basis of the well-known Frisch–Waugh Theorem, which states that $\mathbf{b}_2$ can be obtained by regressing the residuals from a regression of $\mathbf{y}$ on $\mathbf{X}_1$ with the residuals obtained by regressing $\mathbf{X}_2$ on $\mathbf{X}_1$. One application of this result is in the derivation of the form of the least squares estimators in the fixed effects (LSDV) model, which will be discussed in Chapter 7.

1.6 GOODNESS OF FIT

Two commonly used goodness-of-fit statistics used are the coefficient of determination (R^2) and the adjusted coefficient of determination (R_A^2). R^2 is defined as

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

It measures the amount of variability in the response, y, that is explained by including the regressors $x_1, x_2, \ldots, x_k$ in the model. Due to the nature of its construction, we have $0 \leq R^2 \leq 1$. Although higher values (values closer to 1) are desired, a large value of R^2 does not necessarily imply that the regression model is a good one. Adding a variable to the model will always increase R^2 regardless of whether the additional variable is statistically significant or not. In other words, R^2 can be artificially inflated by overfitting the model.

To see this, consider the model $\mathbf{y} = \mathbf{X}_1\,\boldsymbol{\beta}_1 + \mathbf{X}_2\,\boldsymbol{\beta}_2 + \mathbf{U}$. Here, $\mathbf{y}$ is a $n \times 1$ vector of observations, $\mathbf{X}_1$ is the $n \times k_1$ data matrix $\boldsymbol{\beta}_1$ is a vector of k_1 coefficients, $\mathbf{X}_2$is the $n \times k_2$ data matrix with k_2 added variables, $\boldsymbol{\beta}_2$ is a vector of k_2 coefficients, and $\mathbf{U}$ is a $n \times 1$ random vector. Using the Frisch–Waugh theorem, we can show that

$$\hat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2^T\mathbf{M}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^T\mathbf{M}\mathbf{y} = \left(\mathbf{X}_{2^*}^T\mathbf{X}_{2^*}\right)^{-1}\mathbf{X}_{2^*}^T\mathbf{y}_*.$$

Here, $\mathbf{X}_{2^*} = \mathbf{M}\mathbf{X}_2$, $\mathbf{y}_* = \mathbf{M}\mathbf{y}$, and $\mathbf{M} = \mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T$. That is, $\mathbf{X}_{2^*}$ and $\mathbf{y}_*$ are residual vectors of the regression of $\mathbf{X}_2$ and $\mathbf{y}$ on $\mathbf{X}_1$. We can invoke the Frisch–Waugh theorem again to get an expression for $\hat{\boldsymbol{\beta}}_1$. That is, $\hat{\boldsymbol{\beta}}_1 = \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\left(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2\right)$. Using elementary algebra, we can simplify this expression to get $\hat{\boldsymbol{\beta}}_1 = \mathbf{b} - \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\mathbf{X}_2\hat{\boldsymbol{\beta}}_2$, where $\mathbf{b} = \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\mathbf{y}$. Next, note that $\mathbf{u} = \mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2$. We can substitute the expression of $\hat{\boldsymbol{\beta}}_1$ in this to get $\hat{\mathbf{U}} = \mathbf{u} = \mathbf{e} - \mathbf{M}\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{e} - \mathbf{X}_{2^*}\hat{\boldsymbol{\beta}}_2$. The sums of squares of error for the extra variable model is therefore given by

$$\mathbf{u}^T\mathbf{u} = \mathbf{e}^T\mathbf{e} + \hat{\boldsymbol{\beta}}_2^T\left(\mathbf{X}_{2^*}^T\mathbf{X}_{2^*}\right)\hat{\boldsymbol{\beta}}_2 - 2\hat{\boldsymbol{\beta}}_2\mathbf{X}_{2^*}^T\mathbf{e} = \mathbf{e}^T\mathbf{e} + \hat{\boldsymbol{\beta}}_2^T\left(\mathbf{X}_{2^*}^T\mathbf{X}_{2^*}\right)\hat{\boldsymbol{\beta}}_2 - 2\hat{\boldsymbol{\beta}}_2^T\mathbf{X}_{2^*}^T y_*.$$

Here, $\mathbf{e}$ is the residual $\mathbf{y} - \mathbf{X}_1\mathbf{b}$ or $\mathbf{M}\mathbf{y} = \mathbf{y}_*$. We can now, manipulate $\hat{\boldsymbol{\beta}}_2$ to get

$$\mathbf{X}_{2^*}^T\mathbf{y}_* = \left(\mathbf{X}_{2^*}^T\mathbf{X}_{2^*}\right)\hat{\boldsymbol{\beta}}_2 \text{ and}$$

$$\mathbf{u}^T\mathbf{u} = \mathbf{e}^T\mathbf{e} - \hat{\boldsymbol{\beta}}_2^T\left(\mathbf{X}_{2^*}^T\mathbf{X}_{2^*}\right)\hat{\boldsymbol{\beta}}_2 \leq \mathbf{e}^T\mathbf{e}.$$

Dividing both sides by the total sums of squares, $\mathbf{y}^T\mathbf{M}^0\mathbf{y}$, we get

$$\frac{\mathbf{u}^T\mathbf{u}}{\mathbf{y}^T\mathbf{M}^0\mathbf{y}} \leq \frac{\mathbf{e}^T\mathbf{e}}{\mathbf{y}^T\mathbf{M}^0\mathbf{y}} \Rightarrow R^2_{\mathrm{X}_1,\mathrm{X}_2} \geq R^2_{\mathrm{X}_1},$$

where $\mathbf{M}^0 = \mathbf{I} - \mathbf{i}(\mathbf{i}^T\mathbf{i})^{-1}\,\mathbf{i}^T$. See Greene (2003, p. 30) for a proof for the case when a single variable is added to an existing model.

Thus, it is possible for models to have a high R^2 yet yield poor predictions of new observations for the mean response. It is for this reason that many practitioners also use the adjusted coefficient of variation, R_A^2, which adjusts R^2 with respect to the number of explanatory variables in the model. It is defined as

$$R_A^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2).$$

In general, it will increase only when significant terms that improve the model are added to the model. On the other hand, it will decrease with the addition of nonsignificant terms to the model. Therefore, it will always be less than or equal to R^2. When the two R^2 measures differ dramatically, there is a good chance that nonsignificant terms have been added to the model.

1.7 HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

The global F test checks the hypothesis that at least one of the k regressors has a significant effect on the response. It does not indicate which explanatory variable has an effect. It is therefore essential to conduct hypothesis tests on the individual coefficients $\beta_j (j = 1, \ldots, k)$. The hypothesis statements are $H_0{:}\beta_j = 0$ and $H_1{:}\beta_j \neq 0$. The test statistic for testing this is the ratio of the least squares estimate and the standard error of the estimate. That is,

$$t_0 = \frac{b_j}{s.e.(b_j)}, \quad j = 1, \ldots, k,$$

where $s.e.(b_j)$ is the standard error associated with b_j and is defined as $s.e.(b_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$, where C_{jj} is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ corresponding to b_j. Under the assumption that the null hypothesis is true, the test statistic t_0 is distributed as a t distribution with $n-k-1$ degrees of freedom. That is, $t_0 \sim t_{n-k-1}$. The p value is defined as before. That is, $\Pr(|t_0| > t_{n-k-1})$. We reject the null hypothesis if the p value $< \alpha$, where α is the type I error. Note that this test is a marginal test since b_j depends on all the other regressors $x_i (i \neq j)$ that are in the model (see the earlier discussion on interpreting the coefficients).

Hypothesis tests are typically followed by the calculation of confidence intervals. A $100(1-\alpha)\%$ confidence interval for the regression coefficient $\beta_j (j = 1, \ldots, k)$ is given by

$$b_j - t_{\alpha/2,n-k-1} s.e.(b_j) \leq \beta_j \leq b_j + t_{\alpha/2,n-k-1} s.e.(b_j).$$

Note that these confidence intervals can also be used to conduct the hypothesis tests. In particular, if the range of values for the confidence interval includes zero, then we would fail to reject the null hypothesis.

Two other confidence intervals of interest are the confidence interval for the mean response $E(\mathbf{y}|\mathbf{x}_0)$ and the prediction interval for an observation selected from the conditional distribution $f(\mathbf{y}|\mathbf{x}_0)$, where without loss of generality $f(\bullet)$ is assumed to be normally distributed. Also note that $\mathbf{x}_0$ is the setting of the explanatory variables at which the distribution of $\mathbf{y}$ needs to be evaluated. Notice that the mean of $\mathbf{y}$ at a given value of $\mathbf{x} = \mathbf{x}_0$ is given by $E(y|\mathbf{x}_0) = \mathbf{x}_0^T\boldsymbol{\beta}$.

An unbiased estimator for the mean response is $\mathbf{x}_0^T\mathbf{b}$. That is, $E(\mathbf{x}_0^T\mathbf{b}|\mathbf{X}) = \mathbf{x}_0^T\boldsymbol{\beta}$. It can be shown that the variance of this unbiased estimator is given by $\sigma^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0$. Using the previously defined estimator for σ^2 (see Section 1.3.1), we can construct a $100(1-\alpha)\%$ confidence interval on the mean response as

$$\hat{y}(\mathbf{x}_0) - t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0} \leq \mu_{y|\mathbf{x}_0} \leq \hat{y}(\mathbf{x}_0) + t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

Using a similar method, one can easily construct a $100(1-\alpha)\%$ prediction interval for a future observation $\mathbf{x}_0$ as

$$\hat{y}(\mathbf{x}_0) - t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)} \leq y(\mathbf{x}_0) \leq \hat{y}(\mathbf{x}_0) + t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)}.$$

In both these cases, the observation vector $\mathbf{x}_0$ is defined as $\mathbf{x}_0 = (1, x_{01}, x_{02}, \ldots, x_{0k})$, where the "1" is added to account for the intercept term.

Notice that the width of the prediction interval at point $\mathbf{x}_0$ is wider than the width of the confidence interval for the mean response at $\mathbf{x}_0$. This is easy to see because the standard error used for the prediction interval is larger than the standard error used for the mean response interval. This should make intuitive sense also since it is easier to predict the mean of a distribution than it is to predict a future value from the same distribution.

1.8 SOME FURTHER NOTES

A key step in regression analysis is residual analysis to check the least squares assumptions. Violation of one or more assumptions can render the estimation and any subsequent hypothesis tests meaningless. As stated earlier, the least squares residuals can be computed as $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$. Simple residual plots can be used to check a number of assumptions. Chapter 2 shows how these plots are constructed. Here, we simply outline the different types of residual plots that can be used.

1. A plot of the residuals in time order can be used to check for the presence of autocorrelation. This plot can also be used to check for outliers.
2. A plot of the residuals versus the predicted value can be used to check the assumption of random, independently distributed errors. This plot (and the residuals versus regressors plots) can be used to check for the presence of heteroscedasticity. This plot can also be used to check for outliers and influential observations.
3. The normal probability plot of the residuals can be used to check any violations from the assumption of normally distributed random errors.