# CHAPTER 4
# Multiple Regression Analysis

## 4.1 Introduction

So far we have considered only one regressor $X$ besides the constant in the regression equation. Economic relationships usually include more than one regressor. For example, a demand equation for a product will usually include real price of that product in addition to real income as well as real price of a competitive product and the advertising expenditures on this product. In this case

$$Y_i = \alpha + \beta_2 X_{2i} + \beta_3 X_{3i} + .. + \beta_K X_{Ki} + u_i \quad i = 1, 2, \ldots, n \tag{4.1}$$

where $Y_i$ denotes the $i$-th observation on the dependent variable $Y$, in this case the sales of this product. $X_{ki}$ denotes the $i$-th observation on the independent variable $X_k$ for $k = 2, \ldots, K$ in this case, own price, the competitor's price and advertising expenditures. $\alpha$ is the intercept and $\beta_2, \beta_3, \ldots, \beta_K$ are the $(K-1)$ slope coefficients. The $u_i$'s satisfy the classical assumptions 1–4 given in Chapter 3. Assumption 4 is modified to include all the $X$'s appearing in the regression, i.e., every $X_k$ for $k = 2, \ldots, K$, is uncorrelated with the $u_i$'s with the property that $\sum_{i=1}^n (X_{ki} - \bar{X}_k)^2/n$ where $\bar{X}_k = \sum_{i=1}^n X_{ki}/n$ has a finite probability limit which is different from zero.

Section 4.2 derives the OLS normal equations of this multiple regression model and discovers that an additional assumption is needed for these equations to yield a unique solution.

## 4.2 Least Squares Estimation

As explained in Chapter 3, least squares minimizes the residual sum of squares where the residuals are now given by $e_i = Y_i - \widehat{\alpha} - \sum_{k=2}^K \widehat{\beta}_k X_{ki}$ and $\widehat{\alpha}$ and $\widehat{\beta}_k$ denote guesses on the regression parameters $\alpha$ and $\beta_k$, respectively. The residual sum of squares

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \widehat{\alpha} - \widehat{\beta}_2 X_{2i} - .. - \widehat{\beta}_K X_{Ki})^2$$

is minimized by the following $K$ first-order conditions:

$$\begin{aligned} &\partial(\textstyle\sum_{i=1}^n e_i^2)/\partial\widehat{\alpha} = -2\sum_{i=1}^n e_i = 0 \\ &\partial(\textstyle\sum_{i=1}^n e_i^2)/\partial\widehat{\beta}_k = -2\sum_{i=1}^n e_i X_{ki} = 0, \text{ for } k = 2, \ldots, K. \end{aligned} \tag{4.2}$$

or, equivalently

$$\begin{aligned} \textstyle\sum_{i=1}^n Y_i &= \widehat{\alpha} n + \widehat{\beta}_2 \textstyle\sum_{i=1}^n X_{2i} + .. + \widehat{\beta}_K \sum_{i=1}^n X_{ki} \\ \textstyle\sum_{i=1}^n Y_i X_{2i} &= \widehat{\alpha} \textstyle\sum_{i=1}^n X_{2i} + \widehat{\beta}_2 \sum_{i=1}^n X_{2i}^2 + .. + \widehat{\beta}_K \sum_{i=1}^n X_{2i} X_{Ki} \\ \vdots \quad & \qquad \vdots \qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ \textstyle\sum_{i=1}^n Y_i X_{Ki} &= \widehat{\alpha} \textstyle\sum_{i=1}^n X_{Ki} + \widehat{\beta}_2 \sum_{i=1}^n X_{2i} X_{Ki} + .. + \widehat{\beta}_K \sum_{i=1}^n X_{Ki}^2 \end{aligned} \tag{4.3}$$

B.H. Baltagi, *Econometrics*, Springer Texts in Business and Economics, DOI 10.1007/978-3-642-20059-5_4,

where the first equation multiplies the regression equation by the constant and sums, the second equation multiplies the regression equation by $X_2$ and sums, and the $K$-th equation multiplies the regression equation by $X_K$ and sums. $\sum_{i=1}^n u_i = 0$ and $\sum_{i=1}^n u_i X_{ki} = 0$ for $k = 2, \ldots, K$ are implicitly imposed to arrive at (4.3). Solving these $K$ equations in $K$ unknowns, we get the OLS estimators. This can be done more succinctly in matrix form, see Chapter 7. Assumptions 1–4 insure that the OLS estimator is BLUE. Assumption 5 introduces normality and as a result the OLS estimator is also (i) a maximum likelihood estimator, (ii) it is normally distributed, and (iii) it is minimum variance unbiased. Normality also allows test of hypotheses. Without the normality assumption, one has to appeal to the Central Limit Theorem and the fact that the sample is large to perform hypotheses testing.

In order to make sure we can solve for the OLS estimators in (4.3) we need to impose one further assumption on the model besides those considered in Chapter 3.

**Assumption 6:** *No perfect multicollinearity*, i.e., the explanatory variables are not perfectly correlated with each other. This assumption states that, no explanatory variable $X_k$ for $k = 2, \ldots, K$ is a perfect *linear* combination of the other $X$'s. If assumption 6 is violated, then one of the equations in (4.2) or (4.3) becomes redundant and we would have $K - 1$ linearly independent equations in $K$ unknowns. This means that we cannot solve uniquely for the OLS estimators of the $K$ coefficients.

**Example 1:** If $X_{2i} = 3X_{4i} - 2X_{5i} + X_{7i}$ for $i = 1, \ldots, n$, then multiplying this relationship by $e_i$ and summing over $i$ we get

$$\sum_{i=1}^n X_{2i}e_i = 3\sum_{i=1}^n X_{4i}e_i - 2\sum_{i=1}^n X_{5i}e_i + \sum_{i=1}^n X_{7i}e_i.$$

This means that the second OLS normal equation in (4.2) can be represented as a perfect linear combination of the fourth, fifth and seventh OLS normal equations. Knowing the latter three equations, the second equation adds no new information. Alternatively, one could substitute this relationship in the original regression equation (4.1). After some algebra, $X_2$ would be eliminated and the resulting equation becomes:

$$\begin{aligned} Y_i \;=\;& \alpha + \beta_3 X_{3i} + (3\beta_2 + \beta_4)X_{4i} + (\beta_5 - 2\beta_2)X_{5i} + \beta_6 X_{6i} + (\beta_2 + \beta_7)X_{7i} \\ & +.. + \beta_K X_{Ki} + u_i. \end{aligned} \tag{4.4}$$

Note that the coefficients of $X_{4i}$, $X_{5i}$ and $X_{7i}$ are now $(3\beta_2 + \beta_4)$, $(\beta_5 - 2\beta_2)$ and $(\beta_2 + \beta_7)$, respectively. All of which are contaminated by $\beta_2$. These linear combinations of $\beta_2$, $\beta_4$, $\beta_5$ and $\beta_7$ can be estimated from regression (4.4) which excludes $X_{2i}$. In fact, the other $X$'s, not contaminated by this perfect linear relationship, will have coefficients that are not contaminated by $\beta_2$ and hence are themselves estimable using OLS. However, $\beta_2$, $\beta_4$, $\beta_5$ and $\beta_7$ cannot be estimated separately. Perfect multicollinearity means that we cannot separate the influence on $Y$ of the independent variables that are perfectly related. Hence, assumption 6 of no perfect multicollinearity is needed to guarantee a unique solution of the OLS normal equations. Note that it applies to perfect linear relationships and does *not* apply to perfect non-linear relationships among the independent variables. In other words, one can include $X_{1i}$ and $X_{1i}^2$ like (years of experience) and (years of experience)$^2$ in an equation explaining earnings of individuals. Although, there is a perfect quadratic relationship between these independent variables, this is not a perfect *linear* relationship and therefore, does not cause perfect multicollinearity.

## 4.3 Residual Interpretation of Multiple Regression Estimates

Although we did not derive an explicit solution for the OLS estimators of the $\beta$'s, we know that they are the solutions to (4.2) or (4.3). Let us focus on one of these estimators, say $\widehat{\beta}_2$, the OLS estimator of $\beta_2$, the partial derivative of $Y_i$ with respect to $X_{2i}$. As a solution to (4.2) or (4.3), $\widehat{\beta}_2$ is a multiple regression coefficient estimate of $\beta_2$. Alternatively, we can interpret $\widehat{\beta}_2$ as a simple linear regression coefficient.

**Claim 1:** (i) Run the regression of $X_2$ on all the *other* $X$'s in (4.1), and obtain the residuals $\widehat{\nu}_2$, i.e., $X_2 = \widehat{X}_2 + \widehat{\nu}_2$. (ii) Run the simple regression of $Y$ on $\widehat{\nu}_2$, the resulting estimate of the slope coefficient is $\widehat{\beta}_2$.

The first regression essentially cleans out the effect of the other $X$'s from $X_2$, leaving the variation unique to $X_2$ in $\widehat{\nu}_2$. Claim 1 states that $\widehat{\beta}_2$ can be interpreted as a simple linear regression coefficient of $Y$ on this residual. This is in line with the partial derivative interpretation of $\beta_2$. The proof of claim 1 is given in the Appendix. Using the results of the simple regression given in (3.4) with the regressor $X_i$ replaced by the residual $\widehat{\nu}_2$, we get

$$\widehat{\beta}_2 = \textstyle\sum_{i=1}^n \widehat{\nu}_2 Y_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2 \qquad (4.5)$$

and from (3.6) we get

$$\mathrm{var}(\widehat{\beta}_2) = \sigma^2 / \textstyle\sum_{i=1}^n \widehat{\nu}_{2i}^2 \qquad (4.6)$$

An alternative interpretation of $\widehat{\beta}_2$ as a simple regression coefficient is the following:

**Claim 2:** (i) Run $Y$ on all the *other* $X$'s and get the predicted $\widetilde{Y}$ and the residuals, say $\widetilde{\omega}$. (ii) Run the simple linear regression of $\widetilde{\omega}$ on $\widehat{\nu}_2$. $\widehat{\beta}_2$ is the resulting estimate of the slope coefficient.

This regression cleans both $Y$ and $X_2$ from the effect of the other $X$'s and then regresses the cleaned out residuals of $Y$ on those of $X_2$. Once again this is in line with the partial derivative interpretation of $\beta_2$. The proof of claim 2 is simple and is given in the Appendix.

These two interpretations of $\widehat{\beta}_2$ are important in that they provide an easy way of looking at a multiple regression in the context of a simple linear regression. Also, it says that there is no need to clean the effects of one $X$ from the other $X$'s to find its unique effect on $Y$. All one has to do is to include all these $X$'s in the same multiple regression. Problem 1 verifies this result with an empirical example. This will also be proved using matrix algebra in Chapter 7.

Recall that $R^2 = 1 - RSS/TSS$ for any regression. Let $R_2^2$ be the $R^2$ for the regression of $X_2$ on all the other $X$'s, then $R_2^2 = 1 - \sum_{i=1}^n \widehat{\nu}_{2i}^2 / \sum_{i=1}^n x_{2i}^2$ where $x_{2i} = X_{2i} - \bar{X}_2$ and $\bar{X}_2 = \sum_{i=1}^n X_{2i}/n$; $TSS = \sum_{i=1}^n (X_{2i} - \bar{X}_2)^2 = \sum_{i=1}^n x_{2i}^2$ and $RSS = \sum_{i=1}^n \widehat{\nu}_{2i}^2$. Equivalently, $\sum_{i=1}^n \widehat{\nu}_{2i}^2 = \sum_{i=1}^n x_{2i}^2(1 - R_2^2)$ and the

$$\mathrm{var}(\widehat{\beta}_2) = \sigma^2 / \textstyle\sum_{i=1}^n \widehat{\nu}_{2i}^2 = \sigma^2 / \sum_{i=1}^n x_{2i}^2 (1 - R_2^2) \qquad (4.7)$$

This means that the larger $R_2^2$, the smaller is $(1 - R_2^2)$ and the larger is $\mathrm{var}(\widehat{\beta}_2)$ holding $\sigma^2$ and $\sum_{i=1}^n x_{2i}^2$ fixed. This shows the relationship between multicollinearity and the variance of the OLS estimates. High multicollinearity between $X_2$ and the other $X$'s will result in high $R_2^2$ which in turn implies high variance for $\widehat{\beta}_2$. Perfect multicollinearity is the extreme case where $R_2^2 = 1$. This in turn implies an infinite variance for $\widehat{\beta}_2$. In general, high multicollinearity among the regressors yields imprecise estimates for these highly correlated variables. The least squares regression estimates are still unbiased as long as assumptions 1 and 4 are satisfied, but these estimates are unreliable as reflected by their high variances. However, it is important to note that a low $\sigma^2$ and a high $\sum_{i=1}^{n} x_{2i}^2$ could counteract the effect of a high $R_2^2$ leading to a significant $t$-statistic for $\widehat{\beta}_2$. Maddala (2001) argues that high intercorrelation among the explanatory variables are neither necessary nor sufficient to cause the multicollinearity problem. In practice, multicollinearity is sensitive to the addition or deletion of observations. More on this in Chapter 8. Looking at high intercorrelations among the explanatory variables is useful only as a complaint. It is more important to look at the standard errors and $t$-statistics to assess the seriousness of multicollinearity.

Much has been written on possible solutions to the multicollinearity problem, see Hill and Adkins (2001) for a good summary. Credible candidates include: (i) obtaining *new and better data*, but this is rarely available; (ii) introducing *nonsample information* about the model parameters based on previous empirical research or economic theory. The problem with the latter solution is that we never truly know whether the information we introduce is good enough to reduce estimator Mean Square Error.

## 4.4 Overspecification and Underspecification of the Regression Equation

So far we have assumed that the true linear regression relationship is always correctly specified. This is likely to be violated in practice. In order to keep things simple, we consider the case where the true model is a simple regression with one regressor $X_1$.

$$\text{True model: } Y_i = \alpha + \beta_1 X_{1i} + u_i$$

with $u_i \sim \text{IID}(0, \sigma^2)$, but the estimated model is overspecified with the inclusion of an additional irrelevant variable $X_2$, i.e.,

$$\text{Estimated model: } \widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i}$$

From the previous section, it is clear that $\widehat{\beta}_1 = \sum_{i=1}^{n} \widehat{\nu}_{1i} Y_i / \sum_{i=1}^{n} \widehat{\nu}_{1i}^2$ where $\widehat{\nu}_1$ is the OLS residuals of $X_1$ on $X_2$. Substituting the true model for $Y$ we get

$$\widehat{\beta}_1 = \beta_1 \sum_{i=1}^{n} \widehat{\nu}_{1i} X_{1i} / \sum_{i=1}^{n} \widehat{\nu}_{1i}^2 + \sum_{i=1}^{n} \widehat{\nu}_{1i} u_i / \sum_{i=1}^{n} \widehat{\nu}_{1i}^2$$

since $\sum_{i=1}^{n} \widehat{\nu}_{1i} = 0$. But, $X_{1i} = \widehat{X}_{1i} + \widehat{\nu}_{1i}$ and $\sum_{i=1}^{n} \widehat{X}_{1i} \widehat{\nu}_{1i} = 0$ implying that $\sum_{i=1}^{n} \widehat{\nu}_{1i} X_{1i} = \sum_{i=1}^{n} \widehat{\nu}_{1i}^2$. Hence,

$$\widehat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \widehat{\nu}_{1i} u_i / \sum_{i=1}^{n} \widehat{\nu}_{1i}^2 \tag{4.8}$$

and $E(\widehat{\beta}_1) = \beta_1$ since $\widehat{\nu}_1$ is a linear combination of the $X$'s, and $E(X_k u) = 0$ for $k = 1, 2$. Also,

$$\text{var}(\widehat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n} \widehat{\nu}_{1i}^2 = \sigma^2 / \sum_{i=1}^{n} x_{1i}^2 (1 - R_1^2) \tag{4.9}$$

where $x_{1i} = X_{1i} - \bar{X}_1$ and $R_1^2$ is the $R^2$ of the regression of $X_1$ on $X_2$. Using the true model to estimate $\beta_1$, one would get $b_1 = \sum_{i=1}^{n} x_{1i} y_i / \sum_{i=1}^{n} x_{1i}^2$ with $E(b_1) = \beta_1$ and $\text{var}(b_1) =$

$\sigma^2/\sum_{i=1}^n x_{1i}^2$. Hence, $\text{var}(\widehat{\beta}_1) \geq \text{var}(b_1)$. Note also that in the overspecified model, the estimate for $\beta_2$ which has a true value of zero is given by

$$\widehat{\beta}_2 = \sum_{i=1}^n \widehat{\nu}_{2i}Y_i/\sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{4.10}$$

where $\widehat{\nu}_2$ is the OLS residual of $X_2$ on $X_1$. Substituting the true model for $Y$ we get

$$\widehat{\beta}_2 = \sum_{i=1}^n \widehat{\nu}_{2i}u_i/\sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{4.11}$$

since $\sum_{i=1}^n \widehat{\nu}_{2i}X_{1i} = 0$ and $\sum_{i=1}^n \widehat{\nu}_{2i} = 0$. Hence, $E(\widehat{\beta}_2) = 0$ since $\widehat{\nu}_2$ is a linear combination of the $X$'s and $E(X_k u) = 0$ for $k = 1, 2$. In summary, overspecification still yields unbiased estimates of $\beta_1$ and $\beta_2$, but the price is a higher variance.

Similarly, the true model could be a two-regressors model

$$\text{True model: } Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

where $u_i \sim \text{IID}(0, \sigma^2)$ but the estimated model is

$$\text{Estimated model: } \widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 X_{1i}$$

The estimated model omits a relevant variable $X_2$ and underspecifies the true relationship. In this case

$$\widehat{\beta}_1 = \sum_{i=1}^n x_{1i}Y_i/\sum_{i=1}^n x_{1i}^2 \tag{4.12}$$

where $x_{1i} = X_{1i} - \bar{X}_1$. Substituting the true model for $Y$ we get

$$\widehat{\beta}_1 = \beta_1 + \beta_2 \sum_{i=1}^n x_{1i}X_{2i}/\sum_{i=1}^n x_{1i}^2 + \sum_{i=1}^n x_{1i}u_i/\sum_{i=1}^n x_{1i}^2 \tag{4.13}$$

Hence, $E(\widehat{\beta}_1) = \beta_1 + \beta_2 b_{12}$ since $E(x_1 u) = 0$ with $b_{12} = \sum_{i=1}^n x_{1i}X_{2i}/\sum_{i=1}^n x_{1i}^2$. Note that $b_{12}$ is the regression slope estimate obtained by regressing $X_2$ on $X_1$ and a constant. Also, the

$$\text{var}(\widehat{\beta}_1) = E(\widehat{\beta}_1 - E(\widehat{\beta}_1))^2 = E(\sum_{i=1}^n x_{1i}u_i/\sum_{i=1}^n x_{1i}^2)^2 = \sigma^2/\sum_{i=1}^n x_{1i}^2$$

which understates the variance of the estimate of $\beta_1$ obtained from the true model, i.e., $b_1 = \sum_{i=1}^n \widehat{\nu}_{1i}Y_i/\sum_{i=1}^n \widehat{\nu}_{1i}^2$ with

$$\text{var}(b_1) = \sigma^2/\sum_{i=1}^n \widehat{\nu}_{1i}^2 = \sigma^2/\sum_{i=1}^n x_{1i}^2(1 - R_1^2) \geq \text{var}(\widehat{\beta}_1). \tag{4.14}$$

In summary, underspecification yields biased estimates of the regression coefficients and understates the variance of these estimates. This is also an example of imposing a zero restriction on $\beta_2$ when in fact it is not true. This introduces bias, because the restriction is wrong, but reduces the variance because it imposes more information even if this information may be false. We will encounter this general principle again when we discuss distributed lags in Chapter 6.

## 4.5 R-Squared Versus R-Bar-Squared

Since OLS minimizes the residual sums of squares, adding one or more variables to the regression cannot increase this residual sums of squares. After all, we are minimizing over a larger dimension parameter set and the minimum there is smaller or equal to that over a subset of the parameter space, see problem 4. Therefore, for the same dependent variable $Y$, adding more variables makes $\sum_{i=1}^n e_i^2$ non-increasing and $R^2$ non-decreasing, since $R^2 = 1 - (\sum_{i=1}^n e_i^2 / \sum_{i=1}^n y_i^2)$. Hence, a criteria of selecting a regression that "maximizes $R^2$" does not make sense, since we can add more variables to this regression and improve on this $R^2$ (or at worst leave it the same). In order to penalize the researcher for adding an extra variable, one computes

$$\bar{R}^2 = 1 - [\textstyle\sum_{i=1}^n e_i^2/(n-K)]/[\sum_{i=1}^n y_i^2/(n-1)] \tag{4.15}$$

where $\sum_{i=1}^n e_i^2$ and $\sum_{i=1}^n y_i^2$ have been adjusted by their degrees of freedom. Note that the numerator is the $s^2$ of the regression and is equal to $\sum_{i=1}^n e_i^2/(n-K)$. This differs from the $s^2$ in Chapter 3 in the degrees of freedom. Here, it is $n-K$, because we have estimated $K$ coefficients, or because (4.2) represents $K$ relationships among the residuals. Therefore knowing $(n-K)$ residuals we can deduce the other $K$ residuals from (4.2). $\sum_{i=1}^n e_i^2$ is non-increasing as we add more variables, but the degrees of freedom decrease by one with every added variable. Therefore, $s^2$ will decrease only if the effect of the $\sum_{i=1}^n e_i^2$ decrease outweighs the effect of the one degree of freedom loss on $s^2$. This is exactly the idea behind $\bar{R}^2$, i.e., penalizing each added variable by decreasing the degrees of freedom by one. Hence, this variable will increase $\bar{R}^2$ only if the reduction in $\sum_{i=1}^n e_i^2$ outweighs this loss, i.e., only if $s^2$ is decreased. Using the definition of $\bar{R}^2$, one can relate it to $R^2$ as follows:

$$(1-\bar{R}^2) = (1-R^2)[(n-1)/(n-K)] \tag{4.16}$$

## 4.6 Testing Linear Restrictions

In the simple linear regression chapter, we proved that the OLS estimates are BLUE provided assumptions 1 to 4 were satisfied. Then we imposed normality on the disturbances, assumption 5, and proved that the OLS estimators are in fact the maximum likelihood estimators. Then we derived the Cramér-Rao lower bound, and proved that these estimates are efficient. This will be done in matrix form in Chapter 7 for the multiple regression case. Under normality one can test hypotheses about the regression. Basically, any regression package will report the OLS estimates, their standard errors and the corresponding $t$-statistics for the null hypothesis that each individual coefficient is zero. These are tests of significance for each coefficient separately. But one may be interested in a joint test of significance for two or more coefficients simultaneously, or simply testing whether linear restrictions on the coefficients of the regression are satisfied. This will be developed more formally in Chapter 7. For now, all we assume is that the reader can perform regressions using his or her favorite software like EViews, Stata, SAS, TSP, SHAZAM, LIMDEP or GAUSS. The solutions to (4.2) or (4.3) result in the OLS estimates. These multiple regression coefficient estimates can be interpreted as simple regression estimates as shown in section 4.3. This allows a simple derivation of their standard errors. Now, we would like to use these regressions to test linear restrictions. The strategy followed is to impose these restrictions on the model and run the resulting restricted regression. The corresponding Restricted Residual

Sums of Squares is denoted by RRSS. Next, one runs the regression without imposing these linear restrictions to obtain the Unrestricted Residual Sums of Squares, which we denote by URSS. Finally, one forms the following $F$-statistic:

$$F = \frac{(RRSS - URSS)/\ell}{URSS/(n-K)} \sim F_{\ell,n-K} \tag{4.17}$$

where $\ell$ denotes the number of restrictions, and $n - K$ gives the degrees of freedom of the unrestricted model. The idea behind this test is intuitive. If the restrictions are true, then the RRSS should not be much different from the URSS. If RRSS is different from URSS, then we reject these restrictions. The denominator of the $F$-statistic is a consistent estimate of the unrestricted regression variance. Dividing by the latter makes the $F$-statistic invariant to units of measurement. Let us consider two examples:

**Example 2:** Testing the joint significance of two regression coefficients. For e.g., let us test the following null hypothesis $H_0; \beta_2 = \beta_3 = 0$. These are two restrictions $\beta_2 = 0$ and $\beta_3 = 0$ and they are to be tested jointly. We know how to test for $\beta_2 = 0$ alone or $\beta_3 = 0$ alone with individual $t$-tests. This is a test of joint significance of the two coefficients. Imposing this restriction, means the removal of $X_2$ and $X_3$ from the regression, i.e., running the regression of $Y$ on $X_4, \ldots, X_K$ excluding $X_2$ and $X_3$. Hence, the number of parameters to be estimated becomes $(K-2)$ and the degrees of freedom of this restricted regression are $n-(K-2)$. The unrestricted regression is the one including all the $X$'s in the model. Its degrees of freedom are $(n-K)$. The number of restrictions are 2 and this can also be inferred from the difference between the degrees of freedom of the restricted and unrestricted regressions. All the ingredients are now available for computing $F$ in (4.17) and this will be distributed as $F_{2,n-K}$.

**Example 3:** Test the equality of two regression coefficients $H_0; \beta_3 = \beta_4$ against the alternative that $H_1; \beta_3 \neq \beta_4$. Note that $H_0$ can be rewritten as $H_0; \beta_3 - \beta_4 = 0$. This can be tested using a $t$-statistic that tests whether $d = \beta_3 - \beta_4$ is equal to zero. From the unrestricted regression, we can obtain $\widehat{d} = \widehat{\beta}_3 - \widehat{\beta}_4$ with $\text{var}(\widehat{d}) = \text{var}(\widehat{\beta}_3) + \text{var}(\widehat{\beta}_4) - 2\text{cov}(\widehat{\beta}_3, \widehat{\beta}_4)$. The variance-covariance matrix of the regression coefficients can be printed out with any regression package. In section 4.3, we gave these variances and covariances a simple regression interpretation. This means that $se(\widehat{d}) = \sqrt{\text{var}(\widehat{d})}$ and the $t$-statistic is simply $t = (\widehat{d} - 0)/se(\widehat{d})$ which is distributed as $t_{n-K}$ under $H_0$. Alternatively, one can run an $F$-test with the RRSS obtained from running the following regression

$$Y_i = \alpha + \beta_2 X_{2i} + \beta_{3i}(X_{3i} + X_{4i}) + \beta_5 X_{5i} + .. + \beta_K X_{Ki} + u_i$$

with $\beta_3 = \beta_4$ substituted in for $\beta_4$. This regression has the variable $(X_{3i} + X_{4i})$ rather than $X_{3i}$ and $X_{4i}$ separately. The URSS is the regression of $Y$ on all the $X$'s in the model. The degrees of freedom of the resulting $F$-statistic are 1 and $n-K$. The numerator degree of freedom states that there is only one restriction. It will be proved in Chapter 7 that the square of the $t$-statistic is exactly equal to the $F$-statistic just derived. Both methods of testing are equivalent. The first one computes only the unrestricted regression and involves some further variance computations, while the latter involves running two regressions and computing the usual $F$-statistic.

**Example 4:** Test the joint hypothesis $H_0; \beta_3 = 1$ and $\beta_2 - 2\beta_4 = 0$. These two restrictions are usually obtained from prior information or imposed by theory. The first restriction is $\beta_3 = 1$.

The value 1 could have been any other constant. The second restriction shows that a linear combination of $\beta_2$ and $\beta_4$ is equal to zero. Substituting these restrictions in (4.1) we get

$$Y_i = \alpha + \beta_2 X_{2i} + X_{3i} + \tfrac{1}{2}\beta_2 X_{4i} + \beta_5 X_{5i} + .. + \beta_K X_{Ki} + u_i$$

which can be written as

$$Y_i - X_{3i} = \alpha + \beta_2 (X_{2i} + \tfrac{1}{2} X_{4i}) + \beta_5 X_{5i} + .. + \beta_K X_{Ki} + u_i$$

Therefore, the RRSS can be obtained by regressing $(Y - X_3)$ on $(X_2 + \frac{1}{2}X_4), X_5, \ldots, X_K$. This regression has $n - (K - 2)$ degrees of freedom. The URSS is the regression with all the $X$'s included. The resulting $F$-statistic has 2 and $n - K$ degrees of freedom.

**Example 5:** Testing constant returns to scale in a Cobb-Douglas production function. $Q = AK^\alpha L^\beta E^\gamma M^\delta e^u$ is a Cobb-Douglas production function with capital$(K)$, labor$(L)$, energy$(E)$ and material$(M)$. Constant returns to scale means that a proportional increase in the inputs produces the same proportional increase in output. Let this proportional increase be $\lambda$, then $K^* = \lambda K$, $L^* = \lambda L$, $E^* = \lambda E$ and $M^* = \lambda M$. $Q^* = \lambda^{(\alpha+\beta+\gamma+\delta)} AK^\alpha L^\beta E^\gamma M^\delta e^u = \lambda^{(\alpha+\beta+\gamma+\delta)} Q$. For this last term to be equal to $\lambda Q$, the following restriction must hold: $\alpha + \beta + \gamma + \delta = 1$. Hence, a test of constant returns to scale is equivalent to testing $H_0; \alpha + \beta + \gamma + \delta = 1$. The Cobb-Douglas production function is nonlinear in the variables, and can be linearized by taking logs of both sides, i.e.,

$$\text{log}Q = \text{log}A + \alpha \text{log}K + \beta \text{log}L + \gamma \text{log}E + \delta \text{log}M + u \tag{4.18}$$

This is a linear regression with $Y = \text{log}Q$, $X_2 = \text{log}K$, $X_3 = \text{log}L$, $X_4 = \text{log}E$ and $X_5 = \text{log}M$. Ordinary least squares is BLUE on this non-linear model as long as $u$ satisfies assumptions 1–4. Note that these disturbances entered the original Cobb-Douglas production function multiplicatively as $\exp(u_i)$. Had these disturbances entered additively as $Q = AK^\alpha L^\beta E^\gamma M^\delta + u$ then taking logs does not simplify the right hand side and one has to estimate this with non-linear least squares, see Chapter 8. Now we can test constant returns to scale as follows. The unrestricted regression is given by (4.18) and its degrees of freedom are $n - 5$. Imposing $H_0$ means substituting the linear restriction by replacing say $\beta$ by $(1 - \alpha - \gamma - \delta)$. This results after collecting terms in the following restricted regression with one less parameter

$$\text{log}(Q/L) = \text{log}A + \alpha \text{log}(K/L) + \gamma \text{log}(E/L) + \delta \text{log}(M/L) + u \tag{4.19}$$

The degrees of freedom are $n - 4$. Once again all the ingredients for the test in (4.17) are there and this statistic is distributed as $F_1, n - 5$ under the null hypothesis.

**Example 6:** Joint significance of all the slope coefficients. The null hypothesis is

$$H_0; \beta_2 = \beta_3 = .. = \beta_K = 0$$

against the alternative $H_1$; at least one $\beta_k \neq 0$ for $k = 2, \ldots, K$. Under the null, only the constant is left in the regression. Problem 3.2 showed that for a regression of $Y$ on a constant only, the least squares estimate of $\alpha$ is $\bar{Y}$. This means that the corresponding residual sum of squares is $\sum_{i=1}^n (Y_i - \bar{Y})^2$. Therefore, $RRSS$ = Total sums of squares of regression (4.1) = $\Sigma_{i=1}^n y_i^2$.

The URSS is the usual residual sums of squares $\sum_{i=1}^{n} e_i^2$ from the unrestricted regression given by (4.1). Hence, the corresponding $F$-statistic for $H_0$ is

$$F = \frac{(TSS - RSS)/(K-1)}{RSS/(n-K)} = \frac{(\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} e_i^2)/(K-1)}{\sum_{i=1}^{n} e_i^2/(n-K)} = \frac{R^2}{1-R^2} \cdot \frac{n-K}{K-1} \tag{4.20}$$

where $R^2 = 1-(\sum_{i=1}^{n} e_i^2/\sum_{i=1}^{n} y_i^2)$. This $F$-statistic has $(K-1)$ and $(n-K)$ degrees of freedom under $H_0$, and is usually reported by regression packages.

## 4.7 Dummy Variables

Many explanatory variables are qualitative in nature. For example, the head of a household could be male or female, white or non-white, employed or unemployed. In this case, one codes these variables as "$M$" for male and "$F$" for female, or change this qualitative variable into a quantitative variable called FEMALE which takes the value "0" for male and "1" for female. This obviously begs the question: "why not have a variable MALE that takes on the value 1 for male and 0 for female?" Actually, the variable MALE would be exactly 1-FEMALE. In other words, the zero and one can be thought of as a switch, which turns on when it is 1 and off when it is 0. Suppose that we are interested in the earnings of households, denoted by EARN, and MALE and FEMALE are the only explanatory variables available, then problem 10 asks the reader to verify that running OLS on the following model:

$$EARN = \alpha_M MALE + \alpha_F FEMALE + u \tag{4.21}$$

gives $\widehat{\alpha}_M$ = "average earnings of the males in the sample" and $\widehat{\alpha}_F$ = "average earnings of the females in the sample." Notice that there is no intercept in (4.21), this is because of what is known in the literature as the "dummy variable trap." Briefly stated, there will be perfect multicollinearity between MALE, FEMALE and the constant. In fact, MALE + FEMALE = 1. Some researchers may choose to include the intercept and exclude one of the sex dummy variables, say MALE, then

$$EARN = \alpha + \beta FEMALE + u \tag{4.22}$$

and the OLS estimates give $\widehat{\alpha}$ = "average earnings of males in the sample" $= \widehat{\alpha}_M$, while $\widehat{\beta} = \widehat{\alpha}_F - \widehat{\alpha}_M$ = "the difference in average earnings between females and males in the sample." Regression (4.22) is more popular when one is interested in contrasting the earnings between males and females and obtaining with one regression the markup or markdown in average earnings $(\widehat{\alpha}_F - \widehat{\alpha}_M)$ as well as the test of whether this difference is statistically different from zero. This would be simply the $t$-statistic on $\widehat{\beta}$ in (4.22). On the other hand, if one is interested in estimating the average earnings of males and females separately, then model (4.21) should be the one to consider. In this case, the $t$-test for $\widehat{\alpha}_F - \widehat{\alpha}_M = 0$ would involve further calculations not directly given from the regression in (4.21) but similar to the calculations given in Example 3.

What happens when another qualitative variable is included, to depict another classification of the individuals in the sample, say for example, race? If there are three race groups in the sample, WHITE, BLACK and HISPANIC. One could create a dummy variable for each of these classifications. For example, WHITE will take the value 1 when the individual is white and 0 when the individual is non-white. Note that the dummy variable trap does not allow the inclusion of all three categories as they sum up to 1. Also, even if the intercept is dropped, once MALE and FEMALE are included, perfect multicollinearity is still present because MALE + FEMALE = WHITE + BLACK + HISPANIC. Therefore, one category from race should be dropped. Suits (1984) argues that the researcher should use the dummy variable category omission to his or her advantage, in interpreting the results, keeping in mind the purpose of the study. For example, if one is interested in comparing earnings across the sexes holding race constant, the omission of MALE or FEMALE is natural, whereas, if one is interested in the race differential in earnings holding gender constant, one of the race variables should be omitted. Whichever variable is omitted, this becomes the base category for which the other earnings are compared. Most researchers prefer to keep an intercept, although regression packages allow for a no intercept option. In this case one should omit one category from each of the race and sex classifications. For example, if MALE and WHITE are omitted:

$$EARN = \alpha + \beta_F FEMALE + \beta_B BLACK + \beta_H HISPANIC + u \tag{4.23}$$

Assuming the error $u$ satisfies all the classical assumptions, and taking expected values of both sides of (4.23), one can see that the intercept $\alpha =$ the expected value of earnings of the omitted category which is "white males". For this category, all the other switches are off. Similarly, $\alpha + \beta_F$ is the expected value of earnings of "white females," since the FEMALE switch is on. One can conclude that $\beta_F =$ difference in the expected value of earnings between white females and white males. Similarly, one can show that $\alpha + \beta_B$ is the expected earnings of "black males" and $\alpha + \beta_F + \beta_B$ is the expected earnings of "black females." Therefore, $\beta_F$ represents the difference in expected earnings between black females and black males. In fact, problem 11 asks the reader to show that $\beta_F$ represents the difference in expected earnings between hispanic females and hispanic males. In other words, $\beta_F$ represents the differential in expected earnings between females and males holding race constant. Similarly, one can show that $\beta_B$ is the difference in expected earnings between blacks and whites holding sex constant, and $\beta_H$ is the differential in expected earnings between hispanics and whites holding sex constant. The main key to the interpretation of the dummy variable coefficients is to be able to turn on and turn off the proper switches, and write the correct expectations.

The real regression will contain other quantitative and qualitative variables, like

$$\begin{aligned} EARN = \alpha + \beta_F FEMALE + \beta_B BLACK + \beta_H HISPANIC + \gamma_1 EXP \\ + \gamma_2 EXP^2 + \gamma_3 EDUC + \gamma_4 UNION + u \end{aligned} \tag{4.24}$$

where EXP is years of job experience, EDUC is years of education, and UNION is 1 if the individual belongs to a union and 0 otherwise. EXP$^2$ is the squared value of EXP. Once again, one can interpret the coefficients of these regressions by turning on or off the proper switches. For example, $\gamma_4$ is interpreted as the expected difference in earnings between union and non-union members holding all other variables included in (4.24) constant. Halvorsen and Palmquist (1980) warn economists about the interpretation of dummy variable coefficients when the dependent variable is in logs. For example, if the earnings equation is semi-logarithmic:

$$\log(Earnings) = \alpha + \beta UNION + \gamma EDUC + u$$

then $\gamma =$ % change in earnings for one extra year of education, holding union membership constant. But, what about the returns for union membership? If we let $Y_1 = \log(\text{Earnings})$

when the individual belongs to a union, and $Y_0 = \log(\text{Earnings})$ when the individual does not belong to a union, then $g = \%$ change in earnings due to union membership $= (e^{Y_1} - e^{Y_0})/e^{Y_0}$. Equivalently, one can write that $\log(1 + g) = Y_1 - Y_0 = \beta$, or that $g = e^{\beta} - 1$. In other words, one should not hasten to conclude that $\beta$ has the same interpretation as $\gamma$. In fact, the % change in earnings due to union membership is $e^{\beta} - 1$ and not $\beta$. The error involved in using $\widehat{\beta}$ rather than $e^{\widehat{\beta}} - 1$ to estimate $g$ could be substantial, especially if $\widehat{\beta}$ is large. For example, when $\widehat{\beta} = 0.5, 0.75, 1$; $\widehat{g} = e^{\widehat{\beta}} - 1 = 0.65, 1.12, 1.72$, respectively. Kennedy (1981) notes that if $\widehat{\beta}$ is unbiased for $\beta$, $\widehat{g}$ is not necessarily unbiased for $g$. However, consistency of $\widehat{\beta}$ implies consistency for $\widehat{g}$. If one assumes log-normal distributed errors, then $E(\widehat{e^{\beta}}) = e^{\beta+0.5\text{Var}(\widehat{\beta})}$. Based on this result, Kennedy (1981) suggests estimating $g$ by $\widetilde{g} = e^{\widehat{\beta}+0.5\widehat{\text{Var}}(\widehat{\beta})}$-1, where $\widehat{\text{Var}}(\widehat{\beta})$ is a consistent estimate of $\text{Var}(\widehat{\beta})$.

Another use of dummy variables is in taking into account seasonal factors, i.e., including 3 seasonal dummy variables with the omitted season becoming the base for comparison.[1] For example:

$$Sales = \alpha + \beta_W Winter + \beta_S Spring + \beta_F Fall + \gamma_1 Price + u \tag{4.25}$$

the omitted season being the Summer season, and if (4.25) models the sales of air-conditioning units, then $\beta_F$ is the difference in expected sales between the Fall and Summer seasons, holding the price of an air-conditioning unit constant. If these were heating units one may want to change the base season for comparison.

Another use of dummy variables is for War years, where consumption is not at its normal level say due to rationing. Consider estimating the following consumption function

$$C_t = \alpha + \beta Y_t + \delta \text{WAR}_t + u_t \qquad t = 1, 2, \ldots, T \tag{4.26}$$

where $C_t$ denotes real per capita consumption, $Y_t$ denotes real per capita personal disposable income, and $WAR_t$ is a dummy variable taking the value 1 if it is a War time period and 0 otherwise. Note that the War years do not affect the slope of the consumption line with respect to income, only the intercept. The intercept is $\alpha$ in non-War years and $\alpha + \delta$ in War years. In other words, the marginal propensity out of income is the same in War and non-War years, only the level of consumption is different.

Of course, one can dummy other unusual years like periods of strike, years of natural disaster, earthquakes, floods, hurricanes, or external shocks beyond control, like the oil embargo of 1973. If this dummy includes only one year like 1973, then the dummy variable for 1973, call it $D_{73}$, takes the value 1 for 1973 and zero otherwise. Including $D_{73}$ as an extra variable in the regression has the effect of removing the 1973 observation from estimation purposes, and the resulting regression coefficients estimates are exactly the same as those obtained excluding the 1973 observation and its corresponding dummy variable. In fact, using matrix algebra in Chapter 7, we will show that the coefficient estimate of $D_{73}$ is the forecast error for 1973, using the regression that ignores the 1973 observations. In addition, the standard error of the dummy coefficient estimates is the standard error of this forecast. This is a much easier way of obtaining the forecast error and its standard error from the regression package without additional computations, see Salkever (1976). More on this in Chapter 7.

### Interaction Effects

So far the dummy variables have been used to shift the intercept of the regression keeping the slopes constant. One can also use the dummy variables to shift the slopes by letting them interact with the explanatory variables. For example, consider the following earnings equation:

$$EARN = \alpha + \alpha_F FEMALE + \beta EDUC + u \tag{4.27}$$

In this regression, only the intercept shifts from males to females. The returns to an extra year of education is simply $\beta$, which is assumed to be the same for males as well as females. But if we now introduce the interaction variable (FEMALE × EDUC), then the regression becomes:

$$EARN = \alpha + \alpha_F FEMALE + \beta EDUC + \gamma(FEMALE \times EDUC) + u \tag{4.28}$$

In this case, the returns to an extra year of education depends upon the sex of the individual. In fact, $\partial(EARN)/\partial(EDUC) = \beta + \gamma(FEMALE) = \beta$ if male, and $\beta + \gamma$ if female. Note that the interaction variable = EDUC if the individual is female and 0 if the individual is male.

Estimating (4.28) is equivalent to estimating two earnings equations, one for males and another one for females, separately. The only difference is that (4.28) imposes the same variance across the two groups, whereas separate regressions do not impose this, albeit restrictive, equality of the variances assumption. This set-up is ideal for testing the equality of slopes, equality of intercepts, or equality of both intercepts and slopes across the sexes. This can be done with the $F$-test described in (4.17). In fact, for $H_0$; equality of slopes, given different intercepts, the restricted residuals sum of squares (RRSS) is obtained from (4.27), while the unrestricted residuals sum of squares (URSS) is obtained from (4.28). Problem 12 asks the reader to set up the $F$-test for the following null hypothesis: (i) equality of slopes and intercepts, and (ii) equality of intercepts given the same slopes.

Dummy variables have many useful applications in economics. For example, several tests including the Chow (1960) test, and Utts (1982) Rainbow test described in Chapter 8, can be applied using dummy variable regressions. Additionally, they can be used in modeling splines, see Poirier (1976) and Suits, Mason and Chan (1978), and fixed effects in panel data, see Chapter 12. Finally, when the dependent variable is itself a dummy variable, the regression equation needs special treatment, see Chapter 13 on qualitative limited dependent variables.

**Empirical Example:** Table 4.1 gives the results of a regression on 595 individuals drawn from the Panel Study of Income Dynamics (PSID) in 1982. This data is provided on the Springer web site as EARN.ASC. A description of the data is given in Cornwell and Rupert (1988). In particular, log wage is regressed on years of education (ED), weeks worked (WKS), years of full-time work experience (EXP), occupation (OCC = 1, if the individual is in a blue-collar occupation), residence (SOUTH = 1, SMSA = 1, if the individual resides in the South, or in a standard metropolitan statistical area), industry (IND = 1, if the individual works in a manufacturing industry), marital status (MS = 1, if the individual is married), sex and race (FEM = 1, BLK = 1, if the individual is female or black), union coverage (UNION = 1, if the individual's wage is set by a union contract). These results show that the returns to an extra year of schooling is 5.7%, holding everything else constant. It shows that Males on the average earn more than Females. Blacks on the average earn less than Whites, and Union workers earn more than non-union workers. Individuals residing in the South earn less than those living elsewhere. Those residing in a standard metropolitan statistical area earn more on the average than those

**Table 4.1** Earnings Regression for 1982

Dependent Variable: LWAGE
Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob > F |
|---|---|---|---|---|---|
| Model | 12 | 52.48064 | 4.37339 | 41.263 | 0.0001 |
| Error | 582 | 61.68465 | 0.10599 | | |
| C Total | 594 | 114.16529 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.32556 | R-square | 0.4597 |
| Dep Mean | 6.95074 | Adj R-sq | 0.4485 |
| C.V. | 4.68377 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 5.590093 | 0.19011263 | 29.404 | 0.0001 |
| WKS | 1 | 0.003413 | 0.00267762 | 1.275 | 0.2030 |
| SOUTH | 1 | –0.058763 | 0.03090689 | –1.901 | 0.0578 |
| SMSA | 1 | 0.166191 | 0.02955099 | 5.624 | 0.0001 |
| MS | 1 | 0.095237 | 0.04892770 | 1.946 | 0.0521 |
| EXP | 1 | 0.029380 | 0.00652410 | 4.503 | 0.0001 |
| EXP2 | 1 | –0.000486 | 0.00012680 | –3.833 | 0.0001 |
| OCC | 1 | –0.161522 | 0.03690729 | –4.376 | 0.0001 |
| IND | 1 | 0.084663 | 0.02916370 | 2.903 | 0.0038 |
| UNION | 1 | 0.106278 | 0.03167547 | 3.355 | 0.0008 |
| FEM | 1 | –0.324557 | 0.06072947 | –5.344 | 0.0001 |
| BLK | 1 | –0.190422 | 0.05441180 | –3.500 | 0.0005 |
| ED | 1 | 0.057194 | 0.00659101 | 8.678 | 0.0001 |

who do not. Individuals who work in a manufacturing industry or are not blue collar workers or are married earn more on the average than those who are not. For $EXP2 = (EXP)^2$, this regression indicates a significant quadratic relationship between earnings and experience. All the variables were significant at the 5% level except for WKS, SOUTH and MS.

## Note

1. There are more sophisticated ways of seasonal adjustment than introducing seasonal dummies, see Judge et al. (1985).

## Problems

1. For the Cigarette Data given in Table 3.2. Run the following regressions:

   (a) Real per capita consumption of cigarettes on real price and real per capita income. (All variables are in log form, and all regressions in this problem include a constant).

(b) Real per capita consumption of cigarettes on real price.

(c) Real per capita income on real price.

(d) Real per capita consumption on the residuals of part (c).

(e) Residuals from part (b) on the residuals in part (c).

(f) Compare the regression slope estimates in parts (d) and (e) with the regression coefficient estimate of the real income coefficient in part (a), what do you conclude?

2. *Simple Versus Multiple Regression Coefficients.* This is based on Baltagi (1987b). Consider the multiple regression

$$Y_i = \alpha + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad i = 1, 2, \ldots, n$$

along with the following auxiliary regressions:

$$\begin{aligned} X_{2i} &= \widehat{a} + \widehat{b} X_{3i} + \widehat{\nu}_{2i} \\ X_{3i} &= \widehat{c} + \widehat{d} X_{2i} + \widehat{\nu}_{3i} \end{aligned}$$

In section 4.3, we showed that $\widehat{\beta}_2$, the OLS estimate of $\beta_2$ can be interpreted as a simple regression of $Y$ on the OLS residuals $\widehat{\nu}_2$. A similar interpretation can be given to $\widehat{\beta}_3$. Kennedy (1981, p. 416) claims that $\widehat{\beta}_2$ is not necessarily the same as $\widehat{\delta}_2$, the OLS estimate of $\delta_2$ obtained from the regression $Y$ on $\widehat{\nu}_2$, $\widehat{\nu}_3$ and a constant, $Y_i = \gamma + \delta_2 \widehat{\nu}_{2i} + \delta_3 \widehat{\nu}_{3i} + w_i$. Prove this claim by finding a relationship between the $\widehat{\beta}$'s and the $\widehat{\delta}$'s.

3. For the simple regression $Y_i = \alpha + \beta X_i + u_i$ considered in Chapter 3, show that

(a) $\widehat{\beta}_{OLS} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$ can be obtained using the residual interpretation by regressing $X$ on a constant first, getting the residuals $\widehat{\nu}$ and then regressing $Y$ on $\widehat{\nu}$.

(b) $\widehat{\alpha}_{OLS} = \bar{Y} - \widehat{\beta}_{OLS}\bar{X}$ can be obtained using the residual interpretation by regressing 1 on $X$ and obtaining the residuals $\widehat{\omega}$ and then regressing $Y$ on $\widehat{\omega}$.

(c) Check the var$(\widehat{\alpha}_{OLS})$ and var$(\widehat{\beta}_{OLS})$ in parts (a) and (b) with those obtained from the residualing interpretation.

4. *Effect of Additional Regressors on $R^2$.* This is based on Nieswiadomy (1986).

(a) Suppose that the multiple regression given in (4.1) has $K_1$ regressors in it. Denote the least squares sum of squared errors by $SSE_1$. Now add $K_2$ regressors so that the total number of regressors is $K = K_1 + K_2$. Denote the corresponding least squares sum of squared errors by $SSE_2$. Show that $SSE_2 \leq SSE_1$, and conclude that the corresponding $R$-squares satisfy $R_2^2 \geq R_1^2$.

(b) Derive the equality given in (4.16) starting from the definition of $R^2$ and $\bar{R}^2$.

(c) Show that the corresponding $\bar{R}$-squares satisfy $\bar{R}_1^2 \geq \bar{R}_2^2$ when the $F$-statistic for the joint significance of these additional $K_2$ regressors is less than or equal to one.

5. *Perfect Multicollinearity.* Let $Y$ be the output and $X_2$ = skilled labor and $X_3$ = unskilled labor in the following relationship:

$$Y_i = \alpha + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 (X_{2i} + X_{3i}) + \beta_5 X_{2i}^2 + \beta_6 X_{3i}^2 + u_i$$

What parameters are estimable by OLS?

6. Suppose that we have estimated the parameters of the multiple regression model:

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + u_t$$

by *Ordinary Least Squares* (OLS) method. Denote the estimated residuals by $(e_t, t = 1, \ldots, T)$ and the predicted values by $(\widehat{Y}_t, t = 1, \ldots, T)$.

(a) What is the $R^2$ of the regression of $e$ on a constant, $X_2$ and $X_3$?

(b) If we regress $Y$ on a constant and $\widehat{Y}$, what are the estimated intercept and slope coefficients? What is the relationship between the $R^2$ of this regression and the $R^2$ of the original regression?

(c) If we regress $Y$ on a constant and $e$, what are the estimated intercept and slope coefficients? What is the relationship between the $R^2$ of this regression and the $R^2$ of the original regression?

(d) Suppose that we add a new explanatory variable $X_4$ to the original model and re-estimate the parameters by OLS. Show that the estimated coefficient of $X_4$ and its estimated standard error will be the same as in the OLS regression of $e$ on a constant, $X_2$, $X_3$ and $X_4$.

7. Consider the Cobb-Douglas production function in example 5. How can you test for *constant returns to scale* using a $t$-statistic from the unrestricted regression given in (4.18).

8. *Testing Multiple Restrictions.* For the multiple regression given in (4.1). Set up the $F$-statistic described in (4.17) for testing

(a) $H_0; \beta_2 = \beta_4 = \beta_6$.

(b) $H_0; \beta_2 = -\beta_3$ and $\beta_5 - \beta_6 = 1$.

9. *Monte Carlo Experiments.* Hanushek and Jackson (1977, pp. 60–65) generated the following data $Y_i = 15 + 1X_{2i} + 2X_{3i} + u_i$ for $i = 1, 2, \ldots, 25$ with a fixed set of $X_{2i}$ and $X_{3i}$, and $u_i$'s that are IID $\sim N(0, 100)$. For each set of 25 $u_i$'s drawn randomly from the normal distribution, a corresponding set of 25 $Y_i$'s are created from the above equation. Then OLS is performed on the resulting data set. This can be repeated as many times as we can afford. 400 replications were performed by Hanushek and Jackson. This means that they generated 400 data sets each of size 25 and ran 400 regressions giving 400 OLS estimates of $\alpha$, $\beta_2$, $\beta_3$ and $\sigma^2$. The classical assumptions are satisfied for this model, by construction, so we expect these OLS estimators to be BLUE, MLE and efficient.

(a) Replicate the Monte Carlo experiments of Hanushek and Jackson (1977) and generate the means of the 400 estimates of the regression coefficients as well as $\sigma^2$. Are these estimates unbiased?

(b) Compute the standard deviation of these 400 estimates and call this $\widehat{\sigma}_b$. Also compute the average of the 400 standard errors of the regression estimates reported by the regression. Denote this mean by $\overline{s}_b$. Compare these two estimates of the standard deviation of the regression coefficient estimates to the true standard deviation knowing the true $\sigma^2$. What do you conclude?

(c) Plot the frequency of these regression coefficients estimates? Does it resemble its theoretical distribution.

(d) Increase the sample size form 25 to 50 and repeat the experiment. What do you observe?

10. *Female and Male Dummy Variables.*

(a) Derive the OLS estimates of $\alpha_F$ and $\alpha_M$ for $Y_i = \alpha_F F_i + \alpha_M M_i + u_i$ where $Y$ is Earnings, $F$ is FEMALE and $M$ is MALE, see (4.21). Show that $\widehat{\alpha}_F = \bar{Y}_F$, the average of the $Y_i$'s only for females, and $\widehat{\alpha}_M = \bar{Y}_M$, the average of the $Y_i$'s only for males.

(b) Suppose that the regression is $Y_i = \alpha + \beta F_i + u_i$, see (4.22). Show that $\widehat{\alpha} = \widehat{\alpha}_M$, and $\widehat{\beta} = \widehat{\alpha}_F - \widehat{\alpha}_M$.

(c) Substitute $M = 1 - F$ in (4.21) and show that $\alpha = \alpha_M$ and $\beta = \alpha_F - \alpha_M$.

(d) Verify parts (a), (b) and (c) using the earnings data underlying Table 4.1.

11. *Multiple Dummy Variables.* For equation (4.23)

$$EARN = \alpha + \beta_F FEMALE + \beta_B BLACK + \beta_H HISPANIC + u$$

Show that

(a) E(Earnings/Hispanic Female) $= \alpha + \beta_F + \beta_H$; also E(Earnings/Hispanic Male) $= \alpha + \beta_H$. Conclude that $\beta_F =$ E(Earnings/Hispanic Female) – E(Earnings/Hispanic Male).

(b) E(Earnings/Hispanic Female) – E(Earnings/White Female) = E(Earnings/Hispanic Male) – E(Earnings/White Male) $= \beta_H$.

(c) E(Earnings/Black Female) – E(Earnings/White Female) = E(Earnings/Black Male) – E(Earnings/White Male) $= \beta_B$.

12. For the earnings equation given in (4.28), how would you set up the $F$-test and what are the restricted and unrestricted regressions for testing the following hypotheses:

(a) The equality of slopes and intercepts for Males and Females.

(b) The equality of intercepts given the same slopes for Males and Females. Show that the resulting $F$-statistic is the square of a $t$-statistic from the unrestricted regression.

(c) The equality of intercepts allowing for different slopes for Males and Females. Show that the resulting $F$-statistic is the square of a $t$-statistic from the unrestricted regression.

(d) Apply your results in parts (a), (b) and (c) to the earnings data underlying Table 4.1.

13. For the earnings data regression underlying Table 4.1.

(a) Replicate the regression results given in Table 4.1.

(b) Verify that the joint significance of all slope coefficients can be obtained from (4.20).

(c) How would you test the joint restriction that expected earnings are the same for Males and Females whether Black or Non-Black holding everything else constant?

(d) How would you test the joint restriction that expected earnings are the same whether the individual is married or not and whether this individual belongs to a Union or not?

(e) From Table 4.1 what is your estimate of the % change in earnings due to Union membership? If the disturbances are assumed to be log-normal, what would be the estimate suggested by Kennedy (1981) for this % change in earnings?

(f) What is your estimate of the % change in earnings due to the individual being married?

14. *Crude Quality.* Using the data set of U.S. oil field postings on crude prices (\$/barrel), gravity (degree API) and sulphur (% sulphur) given in the CRUDES.ASC file on the Springer web site.

(a) Estimate the following multiple regression model: POIL $= \beta_1 + \beta_2$GRAVITY $+ \beta_3$ SULPHUR $+ \epsilon$.

(b) Regress GRAVITY $= \alpha_0 + \alpha_1$SULPHUR $+ \nu_t$ then compute the residuals ($\widehat{\nu}_t$). Now perform the regression

$$POIL = \gamma_1 + \gamma_2 \widehat{\nu}_t + \epsilon$$

Verify that $\widehat{\gamma}_2$ is the same as $\widehat{\beta}_2$ in part (a). What does this tell you?

(c) Regress POIL $= \phi_1 + \phi_2$SULPHUR $+ w$. Compute the residuals ($\widehat{w}$). Now regress $\widehat{w}$ on $\widehat{\nu}$ obtained from part (b), to get $\widehat{w}_t = \widehat{\delta}_1 + \widehat{\delta}_2 \widehat{\nu}_t +$ residuals. Show that $\widehat{\delta}_2 = \widehat{\beta}_2$ in part (a). Again, what does this tell you?

(d) To illustrate how additional data affects multicollinearity, show how your regression in part (a) changes when the sample is restricted to the first 25 crudes.

(e) Delete all crudes with sulphur content outside the range of 1 to 2 percent and run the multiple regression in part (a). Discuss and interpret these results.

**Table 4.2** U.S. Gasoline Data: 1950–1987

| Year | CAR | QMG<br>(1,000 Gallons) | PMG<br>($ ) | POP<br>(1,000) | RGNP<br>(Billion) | PGNP |
|---|---|---|---|---|---|---|
| 1950 | 49195212 | 40617285 | 0.272 | 152271 | 1090.4 | 26.1 |
| 1951 | 51948796 | 43896887 | 0.276 | 154878 | 1179.2 | 27.9 |
| 1952 | 53301329 | 46428148 | 0.287 | 157553 | 1226.1 | 28.3 |
| 1953 | 56313281 | 49374047 | 0.290 | 160184 | 1282.1 | 28.5 |
| 1954 | 58622547 | 51107135 | 0.291 | 163026 | 1252.1 | 29.0 |
| 1955 | 62688792 | 54333255 | 0.299 | 165931 | 1356.7 | 29.3 |
| 1956 | 65153810 | 56022406 | 0.310 | 168903 | 1383.5 | 30.3 |
| 1957 | 67124904 | 57415622 | 0.304 | 171984 | 1410.2 | 31.4 |
| 1958 | 68296594 | 59154330 | 0.305 | 174882 | 1384.7 | 32.1 |
| 1959 | 71354420 | 61596548 | 0.311 | 177830 | 1481.0 | 32.6 |
| 1960 | 73868682 | 62811854 | 0.308 | 180671 | 1517.2 | 33.2 |
| 1961 | 75958215 | 63978489 | 0.306 | 183691 | 1547.9 | 33.6 |
| 1962 | 79173329 | 62531373 | 0.304 | 186538 | 1647.9 | 34.0 |
| 1963 | 82713717 | 64779104 | 0.304 | 189242 | 1711.6 | 34.5 |
| 1964 | 86301207 | 67663848 | 0.312 | 191889 | 1806.9 | 35.0 |
| 1965 | 90360721 | 70337126 | 0.321 | 194303 | 1918.5 | 35.7 |
| 1966 | 93962030 | 73638812 | 0.332 | 196560 | 2048.9 | 36.6 |
| 1967 | 96930949 | 76139326 | 0.337 | 198712 | 2100.3 | 37.8 |
| 1968 | 101039113 | 80772657 | 0.348 | 200706 | 2195.4 | 39.4 |
| 1969 | 103562018 | 85416084 | 0.357 | 202677 | 2260.7 | 41.2 |
| 1970 | 106807629 | 88684050 | 0.364 | 205052 | 2250.7 | 43.4 |
| 1971 | 111297459 | 92194620 | 0.361 | 207661 | 2332.0 | 45.6 |
| 1972 | 117051638 | 95348904 | 0.388 | 209896 | 2465.5 | 47.5 |
| 1973 | 123811741 | 99804600 | 0.524 | 211909 | 2602.8 | 50.2 |
| 1974 | 127951254 | 100212210 | 0.572 | 213854 | 2564.2 | 55.1 |
| 1975 | 130918918 | 102327750 | 0.595 | 215973 | 2530.9 | 60.4 |
| 1976 | 136333934 | 106972740 | 0.631 | 218035 | 2680.5 | 63.5 |
| 1977 | 141523197 | 110023410 | 0.657 | 220239 | 2822.4 | 67.3 |
| 1978 | 146484336 | 113625960 | 0.678 | 222585 | 3115.2 | 72.2 |
| 1979 | 149422205 | 107831220 | 0.857 | 225055 | 3192.4 | 78.6 |
| 1980 | 153357876 | 100856070 | 1.191 | 227757 | 3187.1 | 85.7 |
| 1981 | 155907473 | 100994040 | 1.311 | 230138 | 3248.8 | 94.0 |
| 1982 | 156993694 | 100242870 | 1.222 | 232520 | 3166.0 | 100.0 |
| 1983 | 161017926 | 101515260 | 1.157 | 234799 | 3279.1 | 103.9 |
| 1984 | 163432944 | 102603690 | 1.129 | 237001 | 3489.9 | 107.9 |
| 1985 | 168743817 | 104719230 | 1.115 | 239279 | 3585.2 | 111.5 |
| 1986 | 173255850 | 107831220 | 0.857 | 241613 | 3676.5 | 114.5 |
| 1987 | 177922000 | 110467980 | 0.897 | 243915 | 3847.0 | 117.7 |

| | | |
|---|---|---|
| CAR: | Stock of Cars | POP: Population |
| RMG: | Motor Gasoline Consumption | RGNP: Real GNP in 1982 dollars |
| PMG: | Retail Price of Motor Gasoline | PGNP: GNP Deflator (1982=100) |

15. Consider the U.S. gasoline data from 1950–1987 given in Table 4.2, and obtained from the file USGAS.ASC on the Springer web site.

   (a) For the period 1950–1972 estimate models (1) and (2):

$$\begin{aligned}\log QMG &= \beta_1 + \beta_2 \log CAR + \beta_3 \log POP + \beta_4 \log RGNP \\ &\quad + \beta_5 \log PGNP + \beta_6 \log PMG + u \end{aligned} \tag{1}$$

$$\log \frac{QMG}{CAR} = \gamma_1 + \gamma_2 \log \frac{RGNP}{POP} + \gamma_3 \log \frac{CAR}{POP} + \gamma_4 \log \frac{PMG}{PGNP} + \nu \tag{2}$$

   (b) What restrictions should the $\beta$'s satisfy in model (1) in order to yield the $\gamma$'s in model (2)?

   (c) Compare the estimates and the corresponding standard errors from models (1) and (2).

   (d) Compute the simple correlations among the $X$'s in model (1). What do you observe?

   (e) Use the Chow-F test to test the parametric restrictions obtained in part (b).

   (f) Estimate equations (1) and (2) now using the full data set 1950–1987. Discuss briefly the effects on individual parameter estimates and their standard errors of the larger data set.

   (g) Using a dummy variable, test the hypothesis that gasoline demand per CAR permanently shifted downward for model (2) following the Arab Oil Embargo in 1973?

   (h) Construct a dummy variable regression that will test whether the price elasticity has changed after 1973.

16. Consider the following model for the demand for natural gas by residential sector, call it model (1):

$$\log Cons_{it} = \beta_0 + \beta_1 \log Pg_{it} + \beta_2 \log Po_{it} + \beta_3 \log Pe_{it} + \beta_4 \log HDD_{it} + \beta_5 \log PI_{it} + u_{it}$$

   where $i = 1, 2, \ldots, 6$ states and $t = 1, 2, \ldots, 23$ years. Cons is the consumption of natural gas by residential sector, $Pg$, $Po$ and $Pe$ are the prices of natural gas, distillate fuel oil, and electricity of the residential sector. $HDD$ is heating degree days and $PI$ is real per capita personal income. The data covers 6 states: NY, FL, MI, TX, UT and CA over the period 1967–1989. It is given in the NATURAL.ASC file on the Springer web site.

   (a) Estimate the above model by OLS. Call this model (1). What do the parameter estimates imply about the relationship between the fuels?

   (b) Plot actual consumption versus the predicted values. What do you observe?

   (c) Add a dummy variable for each state except California and run OLS. Call this model (2). Compute the parameter estimates and standard errors and compare to model (1). Do any of the interpretations of the price coefficients change? What is the interpretation of the New York dummy variable? What is the predicted consumption of natural gas for New York in 1989?

   (d) Test the hypothesis that the intercepts of New York and California are the same.

   (e) Test the hypothesis that **all** the states have the same intercept.

   (f) Add a dummy variable for each state and run OLS without an intercept. Call this model (3). Compare the parameter estimates and standard errors to the first two models. What is the interpretation of the coefficient of the New York dummy variable? What is the predicted consumption of natural gas for New York in 1989?

   (g) Using the regression in part (f), test the hypothesis that the intercepts of New York and California are the same.

## References

This chapter draws upon the material in Kelejian and Oates (1989) and Wallace and Silver (1988). Several econometrics books have an excellent discussion on dummy variables, see Gujarati (1978), Judge et al. (1985), Kennedy (1992), Johnston (1984) and Maddala (2001), to mention a few. Other readings referenced in this chapter include:

Baltagi, B.H. (1987a), "To Pool or Not to Pool: The Quality Bank Case," *The American Statistician*, 41: 150–152.

Baltagi, B.H. (1987b), "Simple versus Multiple Regression Coefficients," *Econometric Theory*, Problem 87.1.1, 3: 159.

Chow, G.C. (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28: 591–605.

Cornwell, C. and P. Rupert (1988), "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variables Estimators," *Journal of Applied Econometrics*, 3: 149–155.

Dufour, J.M. (1980), "Dummy Variables and Predictive Tests for Structural Change," *Economics Letters*, 6: 241–247.

Dufour, J.M. (1982), "Recursive Stability of Linear Regression Relationships," *Journal of Econometrics*, 19: 31–76.

Gujarati, D. (1970), "Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Two Linear Regressions: A Note," *The American Statistician*, 24: 18–21.

Gujarati, D. (1970), "Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Two Linear Regressions: A Generalization," *The American Statistician*, 24: 50–52.

Halvorsen, R. and R. Palmquist (1980), "The Interpretation of Dummy Variables in Semilogarithmic Equations," *American Economic Review*, 70: 474–475.

Hanushek, E.A. and J.E. Jackson (1977), *Statistical Methods for Social Scientists* (Academic Press: New York).

Hill, R. Carter and L.C. Adkins (2001), "Collinearity," Chapter 12 in B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Kennedy, P.E. (1981), "Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations," *American Economic Review*, 71: 802.

Kennedy, P.E. (1981), "The Balentine: A Graphical Aid for Econometrics," *Australian Economic Papers*, 20: 414–416.

Kennedy, P.E. (1986), "Interpreting Dummy Variables," *Review of Economics and Statistics*, 68: 174–175.

Nieswiadomy, M. (1986), "Effect of an Additional Regressor on $R^2$," *Econometric Theory*, Problem 86.3.1, 2:442.

Poirier, D. (1976), *The Econometrics of Structural Change*(North Holland: Amsterdam).

Salkever, D. (1976), "The Use of Dummy Variables to Compute Predictions, Prediction Errors, and Confidence Intervals," *Journal of Econometrics*, 4: 393–397.

Suits, D. (1984), "Dummy Variables: Mechanics vs Interpretation," *Review of Economics and Statistics*, 66: 132–139.

Suits, D.B., A. Mason and L. Chan (1978), "Spline Functions Fitted by Standard Regression Methods," *Review of Economics and Statistics*, 60: 132–139.

Utts, J. (1982), "The Rainbow Test for Lack of Fit in Regression," *Communications in Statistics-Theory and Methods*, 11: 1801–1815.

# Appendix
# Residual Interpretation of Multiple Regression Estimates

**Proof of Claim 1:** Regressing $X_2$ on all the other $X$'s yields residuals $\widehat{\nu}_2$ that satisfy the usual properties of OLS residuals similar to those in (4.2), i.e.,

$$\sum_{i=1}^n \widehat{\nu}_{2i} = 0,\ \sum_{i=1}^n \widehat{\nu}_{2i}X_{3i} = \sum_{i=1}^n \widehat{\nu}_{2i}X_{4i} = .. = \sum_{i=1}^n \widehat{\nu}_{2i}X_{Ki} = 0 \tag{A.1}$$

Note that $X_2$ is the dependent variable of this regression, and $\widehat{X}_2$ is the predicted value from this regression. The latter satisfies $\sum_{i=1}^n \widehat{\nu}_{2i}\widehat{X}_{2i} = 0$. This holds because $\widehat{X}_2$ is a linear combination of the other $X$'s, all of which satisfy (A.1). Turn now to the estimated regression equation:

$$Y_i = \widehat{\alpha} + \widehat{\beta}_2 X_{2i} + .. + \widehat{\beta}_K X_{Ki} + e_i \tag{A.2}$$

Multiply (A.2) by $X_{2i}$ and sum

$$\sum_{i=1}^n X_{2i}Y_i = \widehat{\alpha}\sum_{i=1}^n X_{2i} + \widehat{\beta}_2\sum_{i=1}^n X_{2i}^2 + .. + \widehat{\beta}_K\sum_{i=1}^n X_{2i}X_{Ki} \tag{A.3}$$

This uses the fact that $\sum_{i=1}^n X_{2i}e_i = 0$. Alternatively, (A.3) is just the second equation from (4.3). Substituting $X_{2i} = \widehat{X}_{2i} + \widehat{\nu}_{2i}$, in (A.3) one gets

$$\begin{aligned}\sum_{i=1}^n \widehat{X}_{2i}Y_i + \sum_{i=1}^n \widehat{\nu}_{2i}Y_i &= \widehat{\alpha}\sum_{i=1}^n \widehat{X}_{2i} + \widehat{\beta}_2\sum_{i=1}^n \widehat{X}_{2i}^2 + .. \\ &+\ \widehat{\beta}_K\sum_{i=1}^n \widehat{X}_{2i}X_{Ki} + \widehat{\beta}_2\sum_{i=1}^n \widehat{\nu}_{2i}^2\end{aligned} \tag{A.4}$$

using (A.1) and the fact that $\Sigma_{i=1}^n \widehat{X}_{2i}\widehat{\nu}_{2i} = 0$. Multiply (A.2) by $\widehat{X}_{2i}$ and sum, we get

$$\sum_{i=1}^n \widehat{X}_{2i}Y_i = \widehat{\alpha}\sum_{i=1}^n \widehat{X}_{2i} + \widehat{\beta}_2\sum_{i=1}^n \widehat{X}_{2i}X_{2i} + .. + \widehat{\beta}_K\sum_{i=1}^n \widehat{X}_{2i}X_{Ki} + \sum_{i=1}^n \widehat{X}_{2i}e_i \tag{A.5}$$

But $\sum_{i=1}^n \widehat{X}_2 e_i = 0$ since $\widehat{X}_2$ is a linear combination of all the other $X$'s, all of which satisfy (4.2). Also, $\sum_{i=1}^n \widehat{X}_{2i}X_{2i} = \sum_{i=1}^n \widehat{X}_{2i}^2$ since $\sum_{i=1}^n \widehat{X}_{2i}\widehat{\nu}_{2i} = 0$. Hence (A.5) reduces to

$$\sum_{i=1}^n \widehat{X}_{2i}Y_i = \widehat{\alpha}\sum_{i=1}^n \widehat{X}_{2i} + \widehat{\beta}_2\sum_{i=1}^n \widehat{X}_{2i}^2 + .. + \widehat{\beta}_K\sum_{i=1}^n \widehat{X}_{2i}X_{Ki} \tag{A.6}$$

Subtracting (A.6) from (A.4), we get

$$\sum_{i=1}^n \widehat{\nu}_{2i}Y_i = \widehat{\beta}_2\sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{A.7}$$

and $\widehat{\beta}_2$ is the slope estimate of the simple regression of $Y$ on $\widehat{\nu}_2$ as given in (4.5).

By substituting for $Y_i$ its expression from equation (4.1) in (4.5) we get

$$\widehat{\beta}_2 = \beta_2\sum_{i=1}^n X_{2i}\widehat{\nu}_{2i}/\sum_{i=1}^n \widehat{\nu}_{2i}^2 + \sum_{i=1}^n \widehat{\nu}_{2i}u_i/\sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{A.8}$$

where $\sum_{i=1}^n X_{1i}\widehat{\nu}_{2i} = 0$ and $\sum_{i=1}^n \widehat{\nu}_{2i} = 0$. But, $X_{2i} = \widehat{X}_{2i} + \widehat{\nu}_{2i}$ and $\sum_{i=1}^n \widehat{X}_{2i}\widehat{\nu}_{2i} = 0$, which implies that $\sum_{i=1}^n X_{2i}\widehat{\nu}_{2i} = \sum_{i=1}^n \widehat{\nu}_{2i}^2$ and $\widehat{\beta}_2 = \beta_2 + \sum_{i=1}^n \widehat{\nu}_{2i}u_i/\sum_{i=1}^n \widehat{\nu}_{2i}^2$. This means that $\widehat{\beta}_2$ is unbiased with

$E(\widehat{\beta}_2) = \beta_2$ since $\widehat{\nu}_2$ is a linear combination of the $X$'s and these in turn are not correlated with the $u$'s. Also,

$$\text{var}(\widehat{\beta}_2) = E(\widehat{\beta}_2 - \beta_2)^2 = E(\textstyle\sum_{i=1}^n \widehat{\nu}_{2i}u_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2)^2 = \sigma^2 / \sum_{i=1}^n \widehat{\nu}_{2i}^2$$

The same results apply for any $\widehat{\beta}_k$ for $k = 2, \ldots, K$, i.e.,

$$\widehat{\beta}_k = \textstyle\sum_{i=1}^n \widehat{\nu}_{ki}Y_i / \sum_{i=1}^n \widehat{\nu}_{ki}^2 \tag{A.9}$$

where $\widehat{\nu}_k$ is the OLS residual of $X_k$ on all the other $X$'s in the regression. Similarly,

$$\widehat{\beta}_k = \beta_k + \textstyle\sum_{i=1}^n \widehat{\nu}_{ki}u_i / \sum_{i=1}^n \widehat{\nu}_{ki}^2 \tag{A.10}$$

and $E(\widehat{\beta}_k) = \beta_k$ with $\text{var}(\widehat{\beta}_k) = \sigma^2 / \sum_{i=1}^n \widehat{\nu}_{ki}^2$ for $k = 2, \ldots, K$. Note also that

$$\begin{aligned}\text{cov}(\widehat{\beta}_2, \widehat{\beta}_k) &= E(\widehat{\beta}_2 - \beta_2)(\widehat{\beta}_k - \beta_k) = E(\textstyle\sum_{i=1}^n \widehat{\nu}_{2i}u_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2)(\sum_{i=1}^n \widehat{\nu}_{ki}u_i / \sum_{i=1}^n \widehat{\nu}_{ki}^2) \\ &= \sigma^2 \textstyle\sum_{i=1}^n \widehat{\nu}_{2i}\widehat{\nu}_{ki} / \sum_{i=1}^n \widehat{\nu}_{2i}^2 \sum_{i=1}^n \widehat{\nu}_{ki}^2\end{aligned}$$

**Proof of Claim 2:** Regressing $Y$ on all the other $X$'s yields, $Y_i = \widetilde{Y}_i + \widetilde{\omega}_i$. Substituting this expression for $Y_i$ in (4.5) one gets

$$\widehat{\beta}_2 = (\textstyle\sum_{i=1}^n \widehat{\nu}_{2i}\widetilde{Y}_i + \sum_{i=1}^n \widehat{\nu}_{2i}\widetilde{\omega}_i) / \sum_{i=1}^n \widehat{\nu}_{2i}^2 = \sum_{i=1}^n \widehat{\nu}_{2i}\widetilde{\omega}_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{A.11}$$

where the last equality follows from the fact that $\widetilde{Y}$ is a linear combination of all $X$'s excluding $X_2$, all of which satisfy (A.1). Hence $\widehat{\beta}_2$ is the estimate of the slope coefficient in the linear regression of $\widetilde{\omega}$ on $\widehat{\nu}_2$.

## Simple, Partial and Multiple Correlation Coefficients

In Chapter 3, we interpreted the square of the *simple correlation coefficient*, $r^2_{Y,X_2}$, as the proportion of the variation in $Y$ that is explained by $X_2$. Similarly, $r^2_{Y,X_k}$ is the $R$-squared of the simple regression of $Y$ on $X_k$ for $k = 2, \ldots, K$. In fact, one can compute these simple correlation coefficients and find out which $X_k$ is most correlated with $Y$, say it is $X_2$. If one is selecting regressors to include in the regression equation, $X_2$ would be the best one variable candidate. In order to determine what variable to include next, we look at *partial correlation coefficients* of the form $r_{Y,X_k.X_2}$ for $k \neq 2$. The square of this first-order partial gives the proportion of the residual variation in $Y$, not explained by $X_2$, that is explained by the addition of $X_k$. The maximum first-order partial ('first' because it has only one variable after the dot) determines the best candidate to follow $X_2$. Let us assume it is $X_3$. The first-order partial correlation coefficients can be computed from simple correlation coefficients as follows:

$$r_{Y,X_3.X_2} = \frac{r_{Y,X_3} - r_{Y,X_2}r_{X_2,X_3}}{\sqrt{1 - r^2_{Y,X_2}}\sqrt{1 - r^2_{X_2,X_3}}}$$

see Johnston (1984). Next we look at second-order partials of the form $r_{Y,X_k.X_2,X_3}$ for $k \neq 2, 3$, and so on. This method of selecting regressors is called *forward selection*. Suppose there is only $X_2, X_3$ and $X_4$ in the regression equation. In this case $(1 - r^2_{Y,X_2})$ is the proportion of the variation in $Y$, i.e., $\sum_{i=1}^n y_i^2$, that is not explained by $X_2$. Also $(1 - r^2_{Y,X_3.X_2})(1 - r^2_{Y,X_2})$ denotes the proportion of the variation in $Y$ not explained after the inclusion of both $X_2$ and $X_3$. Similarly $(1 - r^2_{Y,X_4.X_2,X_3})(1 - r^2_{Y,X_3.X_2})(1 - r^2_{Y,X_2})$ is the proportion of the variation in $Y$ unexplained after the inclusion of $X_2$, $X_3$ and $X_4$. But this is exactly $(1 - R^2)$, where $R^2$ denotes the $R$-squared of the multiple regression of $Y$ on a constant, $X_2$, $X_3$ and $X_4$. This $R^2$ is called the *multiple correlation coefficient*, and is also written as $R^2_{Y.X_2,X_3,X_4}$. Hence

$$(1 - R^2_{Y.X_2,X_3,X_4}) = (1 - r^2_{Y,X_2})(1 - r^2_{Y,X_3.X_2})(1 - r^2_{Y,X_4.X_2,X_3})$$

and similar expressions relating the multiple correlation coefficient to simple and partial correlation coefficients can be written by including say $X_3$ first then $X_4$ and $X_2$ in that order.