Advanced Econometrics

Professor: Julie Shi

Timekeeper: Rui Zhou

Fall 2023

Contents

1 The Linear Regression Model

The multiple linear regression model is used to study the relationship between a dependent variable and one or more independent variables. The generic form of the linear regression model is

$$\begin{aligned} y &= f(x_1, x_2, \cdots, x_K) + \varepsilon \\ &= x_1\beta_1 + x_2\beta_2 + \cdots + x_K\beta_K + \varepsilon \end{aligned}$$

where y is the dependent or explained variable and $x_1, \cdots, x_K$ are the independent or explanatory variables. One's theory will specify $f(x_1, x_2, \cdots, x_K)$. This function is commonly called the **population regression equation** of y on $x_1, \cdots, x_K$. In this setting, y is the *regressand* and $x_k (k = 1, \cdots, K)$ are the *regressors* or covariates. The underlying theory will specify the dependent and independent variables in the model, though it is not always obvious which is appropriately defined as each of these. The term ε is a random disturbance, so named because it "disturbs" an otherwise stable relationship. The disturbance arises for several reasons, primarily because we cannot hope to capture every influence on an economic variable in a model, no matter how elaborate.

We assume that each observation in a sample $(y_i, x_{i1}, x_{i2}, \cdots, x_{iK})$ for $i = 1, \cdots, n$, is generated by an underlying process described by

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$$

The observed value of y_i is the sum of two parts, a deterministic part and the random part ε_i. Our objective is to estimate the unknown parameters of the model, use the data to study the validity of the theoretical propositions, and perhaps use the model to predict the variable y. How we proceed from here depends crucially on what we assume about the stochastic process that has led to our observations of the data in hand.

1.1 Assumptions of Linear Regression Model

1.1.1 Linearity

Let the column vector $\mathbf{x}_k$ be the n observations on the variable x_k, for $k = 1, \cdots, K$, and assemble these data in a $n \times K$ data matrix X. In most contexts, the first column of X is assumed to be a column of 1s so that β_1 is the constant term in the model. Let y be the n observations, $y_1, \cdots, y_n$, and let ε be the column vector containing the n disturbances. The linear model as it applies to all n observations can now be written as:

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \cdots \mathbf{x}_K\beta_K + \varepsilon$$

Or in the matrix form of Assumption 1:

$$y = X\beta + \varepsilon$$

The linearity assumption is not so narrow as it might first appear. In the regression context, linearity refers to the manner in which the parameters and the disturbance enter the equation, not necessarily to the relationship among the variables.

1.1.2 Full Rank

Assumption 2 is that there are no exact linear relationships among the variables.

$$X \text{ is an } n \times K \text{ matrix with rank } K.$$

Hence, X has full column rank; the columns of X are linearly independent and there are at least K observations. This assumption is known as an **identification condition**. If there are fewer than K observations, then X cannot have full rank. Hence, we make the (redundant) assumption that n is at least as large as K.
In a two-variable linear model with a constant term, the full rank assumption means that there must be variation in the regressor x. If there is no variation in x, then all our observations will lie on a vertical line. This situation does not invalidate the other assumptions of the model; presumably, it is a flaw in the data set. The possibility that this suggests is that we *could have* drawn a sample in which there was variation in x, but in this instance, we did not. Thus, the model still applies, but we cannot learn about it from the data set in hand. Another interpretation is, in this case the invariant variable is collinear with the constant term.

Technically speaking, the OLS estimator for coefficients, $b = (X'X)^{-1}X'y$, requires the inverse of $(X'X)^{-1}$ exists. Since $\operatorname{rank}\left((X'X)^{-1}\right) = \operatorname{rank}(X)$, so naturally we require X be full column rank.

1.1.3 (Zero) Condition Mean

Assumption 3 is that, the disturbance is assumed to have conditional expected value zero at every observation, which we write as

$$\mathrm{E}\left[\varepsilon_i | X\right] = 0$$

For the full set of observations, we write Assumption 3 as

$$\mathrm{E}\left[\varepsilon | X\right] = \begin{bmatrix} \mathrm{E}\left[\varepsilon_1 | X\right] \\ \mathrm{E}\left[\varepsilon_2 | X\right] \\ \vdots \\ \mathrm{E}\left[\varepsilon_n | X\right] \end{bmatrix} = \mathbf{0}$$

There is a subtle point worth noting. The left-hand side states, in principle, that the mean of each ε_i conditioned on *all observations* $\mathbf{x}_i$ is zero. This conditional mean assumption states, in words, that no observations on X convey information about the

expected value of the disturbance. It is conceivable—for example, in a time-series setting—that although $\mathbf{x}_i$ might provide no information about $\mathrm{E}\left[\varepsilon_i|\cdot\right]$, $\mathbf{x}_j$ at some other observation, such as in the next time period, might. Our assumption at this point is that there is no information about $\mathrm{E}\left[\varepsilon_i|\cdot\right]$ contained in any observation $\mathbf{x}_j$. We will also assume that the disturbances convey no information about each other. That is, $\mathrm{E}\left[\varepsilon_i|\varepsilon_1, \cdots, \varepsilon_{i-1}, \varepsilon_{i+1}, \cdots, \varepsilon_n\right] = 0$. In sum, at this point, we have assumed that the disturbances are purely random draws from some population.
The zero conditional mean implies that the unconditional mean is also zero, since

$$\mathrm{E}\left[\varepsilon_i\right] = \mathrm{E}_X\left[\mathrm{E}\left[\varepsilon_i|X\right]\right] = \mathrm{E}_X\left[0\right] = 0$$

This assumption also means there is no correlation between regressors and disturbances:

$$\mathrm{Cov}\left[\mathrm{E}\left[\varepsilon_i|X\right], X\right] = \mathrm{Cov}\left[\varepsilon_i, X\right] = 0$$

Notice that the converse is not true: $\mathrm{E}\left[\varepsilon_i\right] = 0$ cannot imply $\mathrm{E}\left[\varepsilon_i|X\right] = 0$.

Later we will see, zero conditional mean assumption is critical for unbiasedness of OLS estimators.

In most cases, the zero overall mean assumption is not restrictive. The mean could have been something else. But, if the original model does not contain a constant term, then assuming $\mathrm{E}\left[\varepsilon_i\right] = 0$ could be substantive. This suggests that there is a potential problem in models without constant terms. As a general rule, regression models should not be specified without constant terms unless this is specifically dictated by the underlying theory. Arguably, if we have reason to specify that the mean of the disturbance is something other than zero, we should build it into the systematic part of the regression, leaving in the disturbance only the unknown part of ε. Assumption 3 also implies that

$$\mathrm{E}\left[y|X\right] = X\beta$$

Assumptions 1 and 3 comprise the *linear regression model.* The regression of y on X is the conditional mean, $\mathrm{E}\left[y|X\right]$, so that without assumption 3, $X\beta$ is not the conditional mean function.

The remaining assumptions will more completely specify the characteristics of the disturbances in the model and state the conditions under which the sample observations on $\mathbf{x}$ are obtained.

1.1.4 Spherical Disturbances

Assumption 4 concerns the variances and covariances of the disturbances:

$$\begin{cases} \mathrm{Var}\left[\varepsilon_i | X\right] = \sigma^2, \forall i = 1, \cdots, n \\ \mathrm{Cov}\left[\varepsilon_i, \varepsilon_j\right] = 0, \forall i \neq j \end{cases}$$

Constant variance is labeled homoskedasticity. Uncorrelatedness across observations is labeled generically nonautocorrelation. Disturbances that meet the assumptions of homoskedasticity and nonautocorrelation are sometimes called spherical disturbances. The two assumptions imply that

$$\mathrm{E}\left[\varepsilon\varepsilon' | X\right] = \begin{bmatrix} \mathrm{E}\left[\varepsilon_1\varepsilon_1' | X\right] & \mathrm{E}\left[\varepsilon_1\varepsilon_1' | X\right] & \cdots & \mathrm{E}\left[\varepsilon_1\varepsilon_1' | X\right] \\ \mathrm{E}\left[\varepsilon_2\varepsilon_1' | X\right] & \mathrm{E}\left[\varepsilon_2\varepsilon_1' | X\right] & \cdots & \mathrm{E}\left[\varepsilon_2\varepsilon_n' | X\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}\left[\varepsilon_n\varepsilon_1' | X\right] & \mathrm{E}\left[\varepsilon_n\varepsilon_1' | X\right] & \cdots & \mathrm{E}\left[\varepsilon_n\varepsilon_n' | X\right] \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

$$\iff \mathrm{Var}\left[\mathrm{E}\left[\varepsilon | X\right]\right] = \sigma^2 \mathbf{I}$$

Homoskedasticity guarantees the efficiency of OLS estimators, as we will see in Gauss-Markov theorem.

By using the variance decomposition formula, we find

$$\mathrm{Var}\left[\varepsilon\right] = \mathrm{E}\left[\mathrm{Var}\left[\varepsilon | X\right]\right] + \mathrm{Var}\left[\mathrm{E}\left[\varepsilon | X\right]\right] = \sigma^2 \mathbf{I}$$

Once again, we should emphasize that this assumption describes the information about the variances and covariances among the disturbances that is provided by the independent variables. For the present, we assume that there is none. We will also drop this assumption later when we enrich the regression model. We are also assuming that the disturbances themselves provide no information about the variances and covariances. Although a minor issue at this point, it will become crucial in our treatment of time-series applications.

1.2 OLS Estimation

1.2.1 Derivation of Coefficients Estimation

The unknown parameters of the stochastic relationship $y_i = x_i'\beta + \varepsilon_i$ are the objects of estimation. It is necessary to distinguish between population quantities, such as β and ε_i, and sample estimates of them, denoted as b and e_i. The population regression is

$$\mathrm{E}\left[y_i | x_i\right] = x_i'\beta$$

whereas our estimate of $\mathrm{E}\left[y_i | x_i\right]$ is denoted:

$$\hat{y}_i = x_i'b$$

The disturbance associated with the i-th data point is

$$\varepsilon_i = y_i - x_i'\beta$$

For any value of b, we shall estimate ε_i with the *residual* defined as

$$e_i = y_i - x_i b$$

From the definitions,

$$y_i = x_i'\beta + \varepsilon_i = x_i'b + e_i$$

The population quantity β is a vector of unknown parameters of the probability distribution of y_i whose values we hope to estimate with our sample data $(y_i, x_i), i = 1, 2, \cdots, n$. This is a problem of statistical inference. However, it is instructive to begin by considering the purely algebraic problem of choosing a vector b so that the fitted line $x_i'b$ is close to the data points. The measure of closeness (or, deviations) constitutes a fitting criterion. Among them, the one used most frequently is **least squares**.

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

$$\sum_{i=1}^{n} e_{i0}^2 = \sum_{i=1}^{n} (y_i - x_i'b_0)^2$$

where b_0 denotes the choice for the coefficient vector. In matrix terms, minimizing the sum of squares requires us to choose b_0 to:

$$\begin{aligned} \min_{b_0} S(b_0) = e_0'e_0 &= (y - Xb_0)'(y - Xb_0) \\ &= y'y - b_0'X'y - y'Xb_0 + b_0'X'Xb_0 \\ &= y'y - 2y'Xb_0 + b_0'X'Xb_0 \end{aligned}$$

Notice the third equality holds because both $b_0'X'y$ and $y'Xb_0$ are numbers, or 1×1 matrix.

The necessary condition for a minimum is:

$$\frac{\partial S(b_0)}{\partial b_0} = -2X'y + 2X'Xb_0 = \mathbf{0}$$

Let b be the solution. Then we find that b satisfies the **least squares normal equation**:

$$X'Xb = X'y$$

The normal equation lays the foundation for coefficient estimation and is therefore very important.

If the inverse of $X'X$ exists, which follows from the full column rank assumption of X, then the solution is:

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

From this solution to minimize the sum of squares,

$$\frac{\partial^2 S(b_0)}{\partial b_0 \partial b_0'} = 2X'X$$

which must be a positive definite matrix. To see this, let $q = c'X'Xc$ for some arbitrary nonzero vector c. Then

$$q = c'X'Xc = (Xc)'(Xc) = ||Xc||^2$$

Since X is of full column rank, then Xc cannot be $\mathbf{0}$ for any nonzero c. Thus, q must be positive for any nonzero vector c. Therefore, we can safely conclude that if X has full column rank, then the least squares solution b is unique and minimizes the sum of squared residuals.

Remarks: Our deduction after getting b aims to prove that b does minimize the sum of squared residuals, and b is unique. This is because in the preceding parts, although we have obtained a local optimizer, we do not know whether b is a maximizer and a minimizer, and whether b is unique.

Here we summarize our results in this part:

OLS Estimator and Normal Equation Suppose we hope to estimate $y = X\beta + \varepsilon$, with criterion of minimizing sum of squared residuals. The normal equation is given by

$$X'Xb = X'y$$

Under assumption of X's full column rank, it follows that

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

1.2.2 Projector and Residual Maker

The vector of least squares residuals is

$$\begin{aligned} e &= y - Xb \\ &= y - X(X'X)^{-1}X'y \\ &= \{I - X(X'X)^{-1}X'\}y \\ &:= My \end{aligned}$$

The $n \times n$ matrix M defined as $M = I - X(X'X)^{-1}X'$ is fundamental in regression analysis. In view of $e = My$, we can interpret M as a matrix that produces the vector of least squares residuals in the regression of y on X when it premultiplies any vector y. Thus, M is called the ”residual maker”. The residual maker is perpendicular to column space of X:

$$MX = \mathbf{0}$$

One way to interpret this result is that, if X is regressed on X, a perfect fit will result and the residuals will naturally be $\mathbf{0}$.

Moreover, we can check that M is symmetric and idempotent, and M is then non-negative definite. The properties of M are quite useful in following theories on least squares.

Coincidentally, our definition of estimated linear model means that y is partitioned into two meaningful parts: the fitted values $\hat{y} = Xb$ and the residual e. Since $MX = 0$, the two parts are orthogonal.

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

The matrix P is a **projection matrix**. It is the matrix formed from X such that when a vector y is premultiplied by P, the result is the fitted values in the least squares regression of y on X. This is also the *projection* of the vector y onto the column space of X. From the projection meaning in P, it is quick to verify that

$$PX = X$$

In meanings of vectors spaces, space of regressors and space of residuals are orthogonal to each other:

$$PM = MP = 0$$

The residual maker and projector sum up to construct the original space:

$$P + M = I$$

Same as M, P is symmetric and idempotent, and P is thus non-negative definite. The transformation of P and M are crucial in regression analysis.

By far, we can see the partition of y is quite meaningful:

$$y = Py + My = \text{Projection} + \text{Residual}$$

Graphically speaking, the column space of X is perpendicular to the residuals'. In mathematical forms, we can also see the Pythagorean theorem at work in the sum of squares:

$$\begin{aligned} y'y &= y'P'Py + y'M'My \\ &= \hat{y}'\hat{y} + e'e \end{aligned}$$

In manipulating equations involving least squares results, the following equivalent expressions for the sum of squares residuals are often useful:

$$\begin{aligned} e'e &= y'M'My = y'My = y'e = e'y \\ e'e &= y'y - b'X'Xb = y'y - b'X'y = y'y - y'Xb \end{aligned}$$

Here we summarize our results in this part:

Projector and Residual Maker If we regress y on X, we define the projector P and residual maker M as

$$\begin{cases} P = X(X'X)^{-1}X' \\ M = \mathbf{I} - X(X'X)^{-1}X' \end{cases}$$

Both P and M are symmetric, idempotent and thus non-negative definite. Moreover, Py will return the fitted values in regression of y on X, and My gives the residuals. P and M correspond to two orthogonal spaces of variation of y. And

$$PX = X, MX = \mathbf{0}$$

1.2.3 Partitioned Regression

1.2.3.1 FWL Theorem Suppose that the regression involves two sets of variables X_1 and X_2. Thus,

$$y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

In order to obtain the algebraic solutions for b_1 and b_2, we start from normal equation in partitioned form:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}$$

A solution can be obtained using the partitioned inverse matrix. Alternatively, the two partitioned equations can be manipulated directly to solve for b_2. We first solve for b_1:

$$b_1 = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2b_2 = (X_1'X_1)^{-1}X_1'(y - X_2b_2)$$

This solution says that b_1 is the set of coefficients in the regression of y on X_1, *minus* a correction vector. However, if $X_1'X_2 = \mathbf{0}$, i.e., X_1 and X_2 are orthogonal, then b_1 can be simplified to: $b_1 = (X_1'X_1)^{-1}X_1'y$. The general result is given in the following theorem.

Orthogonal Partitioned Regression In the multiple linear least squares regression of y on two sets of variables, X_1 and X_2, if the two sets of variables are orthogonal, then the separate coefficient vectors can be obtained by separate regressions of y on X_1 alone and y on X_2 alone.

In most cases, the two sets of variables X_1 and X_2 are not orthogonal, then the solution for b_1 and b_2 would be much more involved than just simple regressions. The more general solution is given by the following theorem.

Frisch-Waugh-Lovell Theorem In the linear least squares regression of vector y on two sets of variables, X_1 and X_2, the subvector b_2 is the set of coefficients obtained when the residuals from a regression of y on X_1 alone are regressed on the set of residuals obtained when each column of X_2 is regressed on X_1.

Proof From the two sub-equations in the partitioned normal equation, we obtain b_1 from one equation and plug in expression of b_1 into another equation. After collecting terms, the solution for b_2 is:

$$\begin{aligned} b_2 &= \{X_2(I - X_1(X_1'X_1)^{-1}X_1')X_2\}^{-1}\{X_2'(I - X_1(X_1'X_1)^{-1}X_1')y\} \\ &= (X_2'M_1X_2)^{-1}(X_2'M_1y) \end{aligned}$$

where M_1 is the residual maker for regression on X_1. Notice that M_1X_2 is a matrix of residuals; each column of M_1X_2 is a vector of residuals in the regression of the corresponding column of X_2 on the variables in X_1.

By exploiting the fact that M_1 is symmetric and idempotent, the result can be rewritten as:

$$b_2 = (X_2^{*'}X_2^*)^{-1}X_2^{*'}y^*$$

where $X_2^* = M_1X_2$, and $y^* = M_1y$.

This process described by FWL theorem is commonly called *partialling out* or *netting out* the effect of X_1. For this reason, the coefficients in a multiple regression are often called the *partial regression coefficients.*

It is not hard to compute the expectation and variance of b_2, and verify that b_2 is unbiased.

$$\begin{aligned}
\mathrm{E}\left[b_2|X\right] &= \mathrm{E}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1y\right)|X\right] = \mathrm{E}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1\left(X_1\beta_1+X_2\beta_2+\varepsilon\right)\right)|X\right] \\
&= \mathrm{E}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1X_1\beta_1+X_2'M_1X_2\beta_2+X_2'M_1\varepsilon\right)|X\right] \\
&= \mathrm{E}\left[0+\beta_2+\left(X_2'M_1X_2\right)^{-1}X_2'M_1\varepsilon|X\right] \\
&= \beta_2 \\
\mathrm{Var}\left[b_2|X\right] &= \mathrm{Var}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1y\right)|X\right] = \mathrm{Var}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1\left(X_1\beta_1+X_2\beta_2+\varepsilon\right)\right)|X\right] \\
&= \mathrm{Var}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1X_1\beta_1+X_2'M_1X_2\beta_2+X_2'M_1\varepsilon\right)|X\right] \\
&= \mathrm{Var}\left[\left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1\varepsilon\right)|X\right] \\
&= \left(X_2'M_1X_2\right)^{-1}X_2'M_1\left(\mathrm{Var}\left[\varepsilon|X\right]\right)M_1'X_2\left(X_2'M_1X_2\right)^{-1} \\
&= \sigma^2\left(X_2'M_1X_2\right)^{-1}
\end{aligned}$$

An application of FWL theorem is to compute a single coefficient.

Individual Regression Coefficients The coefficient on z in a multiple regression of y on $W = [X, z]$ is computed as $c = (z'Mz)^{-1}(z'My) = (z^{*'}z^*)^{-1}z^{*'}y^*$, where z^* and y^* are the residual vectors from least squares regressions of z and y on X; $z^* = Mz$ and $y^* = My$ where M is residual maker for X: $M = \mathbf{I} - X(X'X)^{-1}X'$.

Besides, consider the case in which X_1 is $\mathbf{1}$, namely a constant term which is a column of 1s in the first column of X. For each vectors regressed on $\mathbf{1}$, the resulted residuals are deviations from the sample mean. The interesting discovery is summarized as in the corollary.

Regression with a Constant Term The slopes in a multiple regression that contains a constant term are obtained by transforming the data to deviations from their means and then regressing the variable y in deviation form on the explanatory variables, also in deviation form.

From the two theorems in partitioned regression and the corollary of regression with a constant term, we formalize our integrated ideas into the following theorem.

Orthogonal Regression If the multiple regression of y on X contains a constant term and the variables in the regression are uncorrelated, then the multiple regression slopes are the same as the slopes in the individual simple regressions of y on a constant and each variable in turn.

The use of multiple regression involves a conceptual experiment that we might not be able to carry out in practice, the *ceteris paribus.* FWL theorem rationalizes all the thought experiment with *ceteris paribus*, even if the sample contains no such pair of observations.

1.2.3.2 Partial Correlation Coefficient From the perspective of *ceteris paribus*, to see the correlation between two variables, we should first purge out effects of other variables, and use a **partial correlation coefficient** as a measure. Follow the preceding notations, the partial correlation coefficient is defined as

$$r_{yz}^{*2} = \frac{(z_*' y_*)^2}{(z_*' z_*)(y_*' y_*)}$$

Note that there is a convenient shortcut.

Relationship of Partial Correlation Coefficient and t Statistic Once the multiple regression is computed, the t ratio for testing the hypothesis that the coefficient of z equals zero can be used to compute

$$r_{yz}^{*2} = \frac{t_z^2}{t_z^2 + \text{degrees of freedom}}$$

where the degrees of freedom is equal to $n - (K+1)$.

Proof By definition, the squared t ratio for z's coefficient is

$$\begin{aligned} t_z^2 &= \frac{c^2}{s^2 (W'W)^{-1}_{K+1,K+1}} \\ &= \frac{c^2}{\left(\dfrac{u'u}{n-(K+1)}\right)(W'W)^{-1}_{K+1,K+1}} \end{aligned}$$

where $(W'W)_{K+1,K+1}$ is the $(K+1)$, namely last, diagonal element of $(W'W)^{-1}$. To see it specifically, we introduce the following useful partitioned inverse formula:

Partitioned Inverse Formula For the general 2×2 partitioned matrix, one form of the partitioned inverse is

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1}(\mathbf{I} + A_{12}F_2A_{21}A_{11}^{-1}) & -A_{11}^{-1}A_{12}F_2 \\ -F_2A_{21}A_{11}^{-1} & F_2 \end{bmatrix}$$

where

$$F_2 = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$$

The upper left block could also be written as

$$F_1 = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$$

Specifically, finding the inverse of $X'X$ is the most common in regression analysis, where $X = [X_1, X_2]$.

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} = \begin{bmatrix} (X_1'M_2X_1)^{-1} & -(X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1} \\ -(X_2'M_1X_2)^{-1}X_2'X_1(X_1'M_1X_1)^{-1} & (X_2'M_1X_2)^{-1} \end{bmatrix}$$

where $M_1 = \mathbf{I} - X_1(X_1'X_1)^{-1}X_1', M_2 = \mathbf{I} - X_2(X_2'X_2)^{-1}X_2'$. The off-diagonal elements are symmetric.

By partitioned inverse formula, the last lower right element of the matrix equals $(z'Mz)^{-1} = (z_*'z_*)^{-1}$. From FWL theorem, we also have that $c = (z_*'z_*)^{-1}z_*'y_* = \dfrac{(z_*'y_*)}{(z_*'z_*)}$. For convenience, let $df = n - (K+1)$. Then,

$$t_z^2 = \frac{\dfrac{(z_*'y_*)^2}{(z_*'z_*)^2}}{\dfrac{u'u}{df}\cdot(z_*'z_*)^{-1}} = \frac{(z_*'y_*)^2\cdot df}{(u'u)(z_*'z_*)}$$

$$\Longrightarrow \frac{t_z^2}{t_z^2+df} = \frac{\dfrac{(z_*'y_*)^2\cdot df}{(u'u)(z_*'z_*)}}{\dfrac{(z_*'y_*)^2\cdot df}{(u'u)(z_*'z_*)}+df} = \frac{\dfrac{(z_*'y_*)^2}{(u'u)(z_*'z_*)}}{\dfrac{(z_*'y_*)^2}{(u'u)(z_*'z_*)}+1} = \frac{(z_*'y_*)^2}{(z_*'y_*)^2+(u'u)(z_*'z_*)}$$

Divide numerator and denominator by $(z_*'z_*)(y_*'y_*)$ to move closer to r_{yz}^{*2}:

$$\begin{aligned} \frac{t_z^2}{t_z^2+df} &= \frac{(z_*'y_*)^2}{(z_*'y_*)^2+(u'u)(z_*'z_*)} \\ &= \frac{\dfrac{(z_*'y_*)^2}{(z_*'z_*)(y_*'y_*)}}{\dfrac{(z_*'y_*)^2}{(z_*'z_*)(y_*'y_*)}+\dfrac{(u'u)(z_*'z_*)}{(z_*'z_*)(y_*'y_*)}} \\ &= \frac{r_{yz}^{*2}}{r_{yz}^{*2}+\dfrac{u'u}{y_*'y_*}} \end{aligned}$$

By the theorem of change in the sum of squares when a variable is added to a regression,

$$u'u = e'e - c^2(z_*'z_*) = y_*'y_* - \frac{(z_*'y_*)^2}{z_*'z_*}$$

Inserting this into our previous result to get:

$$\begin{aligned}
\frac{t_z^2}{t_z^2+df} &= \frac{r_{yz}^{*2}}{r_{yz}^{*2}+\dfrac{u'u}{y_*'y_*}} \\
&= \frac{r_{yz}^{*2}}{r_{yz}^{*2}+\dfrac{y_*'y_*-\dfrac{(z_*'y_*)^2}{z_*'z_*}}{y_*'y_*}} \\
&= \frac{r_{yz}^{*2}}{r_{yz}^{*2}+(1-r_{yz}^{*2})} \\
&= r_{yz}^{*2}
\end{aligned}$$

1.2.4 Goodness of Fit

Variation of the dependent variable is defined in terms of deviations from its mean. The *total variation* in y is the sum of squared deviations:

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

In terms of the regression analysis, we may write the full set of observations as

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

Notice that here normal equation plays an important role. The first observation comes from the assumption that $E[\varepsilon|X]=0$, that is, the regressors are uncorrelated with the disturbances,

$$\begin{aligned}
& X'Xb = X'y \\
\Longleftrightarrow\ & X'(y-Xb) = 0 \\
\Longleftrightarrow\ & X'e = 0
\end{aligned}$$

Therefore, if the regression contains a constant term, the residuals will add up to be 0. From this, the mean of the predicted values of y_i will equal the mean of the actual values, that is,

$$\bar{\hat{y}} = \bar{y}$$

From this result we can immediately derive:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i = (x_i - \bar{x})'b + e_i$$

The second observation is that, as defined before, projector P and residual maker M are perpendicular to each other, and y can be partitioned into the two orthogonal parts:

$$y = \hat{y} + e = Py + My$$

So are the variations in each part (where M_0 is the residual maker for $\mathbf{1}$).

$$\begin{aligned}
&y = Xb + e \\
\Longrightarrow &M_0 y = M_0 X b + M_0 e \\
\Longrightarrow &y' M_0 y = b' X' M_0 X b + e' e \\
\Longrightarrow &SST = SSR + SSE
\end{aligned}$$

In the deduction we have made good use of the fact that $X'e = e'X = 0$. We can now obtain a measure of how well the regression line fits:

$$R^2 = \frac{SSR}{SST} = \frac{b'X'M_0Xb}{y'M_0y} = 1 - \frac{e'e}{y'M_0y}$$

As we have shown, R^2 must be between 0 and 1, and it measures the proportion of the total variation in y that is accounted for by variation in the regressors. It equals zero if the regression is a horizontal line, that is, if all the elements of b except the constant term are zero. In this case, the predicted values of y are always $\bar{y}$, so deviations of x from its mean do not translate into different predictions for y. As such, x has no explanatory power. The other extreme, $R^2 = 1$, occurs if the values of x and y all lie in the same hyperplane (on a straight line for a two variable regression) so that the residuals are all zero. If all the values of y_i lie on a vertical line, then R^2 has no meaning and cannot be computed.

Regression analysis is often used for forecasting. In this case, we are interested in how well the regression model predicts movements in the dependent variable. With this in mind, an equivalent way to compute R^2 is also useful. First

$$\begin{cases}
\hat{y} = Xb \\
y = \hat{y} + e \\
X'e = 0 \\
\bar{e} = 0 \Longrightarrow M_0 e = 0
\end{cases} \qquad \Longrightarrow \hat{y}' M_0 \hat{y} = \hat{y}' M_0 y$$

$$\begin{aligned}
R^2 &= \frac{SSR}{SST} = \frac{b'X'M_0Xb}{y'M_0y} \\
\Longrightarrow \qquad &= \frac{\hat{y}'M_0\hat{y}}{y'M_0y} = \frac{\hat{y}'M_0\hat{y}}{y'M_0y} \cdot \frac{\hat{y}'M_0y}{\hat{y}'M_0\hat{y}} \\
&= \frac{\left[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})\right]^2}{\left[\sum_{i=1}^n (y_i - \bar{y})^2\right]\left[\sum_{i=1}^n (\hat{y}_i - \bar{y})^2\right]}
\end{aligned}$$

which is the squared correlation between the observed values of y and the predicted values.

There are some problems with the use of R^2 in analyzing goodness of fit. The first concerns the number of degrees of freedom used up in estimating the parameters. R^2 will never decrease when another variable is added to a regression equation. First, we will see change in the sum of squares will never increase when adding a new variable into a regression.

Change in the Sum of Squares When a Variable is Added to a Regression If $e'e$ is the sum of squared residuals when y is regressed on X and $u'u$ is the sum of squared residuals when y is regressed on X and z, then

$$u'u = e'e - c^2(z_*'z_*) \leq e'e$$

where c is the coefficient on z in the long regression of y on $[X, z]$ and $z_* = Mz$ is the vector of residuals when z is regressed on X.

Proof In the long regression of y on X and z, the vector of residuals is $u = y - Xd - zc$. Note that unless $X'z = \mathbf{0}$, d will not equal $b = (X'X)^{-1}X'y$. Moreover, unless $c = \mathbf{0}$, u will not equal $e = y - Xb$. From FWL theorem, $c = (z_*'z_*)^{-1}(z_*'y_*)$. From the equation in partitioned regression, we know

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

Inserting this expression for d in that for u gives

$$u = y - Xd - zc = y - Xb - X(X'X)^{-1}X'zc = e - Mzc = e - z_*c$$

Then,

$$u'u = e'e + c^2(z_*'z_*) - 2c(z_*'e)$$

Notice that $e = My = y_*$, and $z_*'e = z_*'y_* = c(z_*'z_*)$. Inserting this result in $u'u$ immediately above gives the result in the theorem.

The result can be extended to the case when there is a set of variables added to a regression.

Change in the Sum of Squares When a Set of Variables is Added to a Regression If $e'e$ is the sum of squares when y is regressed on X_1, and $u'u$ is the sum of squared residuals when y is regressed on $[X_1, X_2]$, then

$$u'u = e'e - b_2'X_2'M_1X_2b_2$$

Proof First in the setup, we estimate the following two equations:

$$\begin{aligned} y =& X_1 b + e \\ y =& X_1 b_1 + X_2 b_2 + u \end{aligned}$$

From FWL theorem, we know that

$$b_2 = (X_2' M_1 X_2)^{-1} (X_2' M_1 y)$$

From partitioned regression formula, we can solve for b_1 from b_2:

$$\begin{aligned} b_1 &= (X_1' X_1)^{-1} X_1' (y - X_2 b_2) \\ &= (X_1' X_1)^{-1} X_1 y - (X_1' X_1)^{-1} X_1' X_2 b_2 \\ &= b - (X_1' X_1)^{-1} X_1' X_2 b_2 \end{aligned}$$

Therefore, we can link the residuals of long regression and residuals of the short regression:

$$\begin{aligned} u &= y - X_1 b_1 - X_2 b_2 \\ &= y - X_1 (b - (X_1' X_1)^{-1} X_1' X_2 b_2) - X_2 b_2 \\ &= (y - X_1 b) + X_1 (X_1' X_1)^{-1} X_1' X_2 b_2 - X_2 b_2 \\ &= e + X_1 (X_1' X_1)^{-1} X_1' X_2 b_2 - X_2 b_2 \\ &= e - M_1 X_2 b_2 \\ \Longrightarrow u' u &= e' e - e' M_1 X_2 b_2 - b_2' X_2' M_1 e + b_2' X_2' M_1 X_2 b_2 \\ &= e' e - 2 b_2' X_2' M_1 e + b_2' X_2' M_1 X_2 b_2 \end{aligned}$$

To simplify the result, we should notice

$$\begin{cases} e = M_1 y \\ b_2 = (X_2' M_1 X_2)^{-1} (X_2' M_1 y) \end{cases} \Longrightarrow b_2' X_2' M_1 e = (y' M_1 X_2') (X_2' M_1 X_2)^{-1} (X_2' M_1 y)$$

$$\begin{aligned} b_2' X_2' M_1 X_2 b_2 &= (y' M_1 X_2') (X_2' M_1 X_2)^{-1} (X_2' M_1 X_2) (X_2' M_1 X_2)^{-1} (X_2' M_1 y) \\ &= (y' M_1 X_2') (X_2' M_1 X_2)^{-1} (X_2' M_1 y) \end{aligned}$$

Therefore, we find that $b_2' X_2' M_1 e = b_2' X_2' M_1 X_2 b_2$, which can be used to simplify the result. However, since $b_2' X_2' M_1 X_2 b_2$ is more meaningful, we keep this term to get:

$$u' u = e' e - b_2' X_2' M_1 X_2 b_2$$

Based on the results of change in sum of squares, we are straightforward to see change in R^2 when a variable is added to a regression.

Change in R^2 When a Variable Is Added to a Regression Let R^2_{Xz} be the coefficient of determination in the regression of y on X and an additional variable z, let R^2_X be the same for the regression of y on X alone, and let r^{*2}_{yz} be the partial correlation between y and z, controlling for X. Then

$$R^2_{Xz} = R^2_X + (1 - R^2_X)r^{*2}_{yz}$$

Proof Following the same notations, we have shown

$$u'u = e'e - c^2(z'_* z_*)$$

By definition, $r^{*2}_{yz} = \frac{(z'_* y_*)^2}{(z'_* z_*)(y'_* y_*)}$; so we have

$$u'u = e'e - c^2(z'_* z_*) = e'e(1 - r^{*2}_{yz})$$

Now divide through both sides of the equality by $y'M_0 y$. Notice that $1 - R^2_X = \dfrac{u'u}{y'M_0 y}$ and $1 - R^2_{Xz} = \dfrac{e'e}{y'M_0 y}$. Rearrange the result produces the equation illustrated above.

Thus, the R^2 in the longer regression cannot be smaller. It is tempting to exploit this result by just adding variables to the model; R^2 will continue to rise to its limit of 1. Therefore, we introduce the **adjusted** R^2 (for degrees of freedom), which incorporates a penalty for adding superfluous variables:

$$\bar{R}^2 = 1 - \left(\frac{e'e}{n - K}\right) \Big/ \left(\frac{y'M_0 y}{n - 1}\right)$$

For computational purposes, the connection between R^2 and $\bar{R}^2$ is

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

The adjusted R^2 may decline when a variable is added to the set of independent variables. Indeed, $\bar{R}^2$ may even be negative. To consider an admittedly extreme case, suppose that x and y have a sample correlation of zero. Then the adjusted R^2 will equal $-\dfrac{1}{n-2}$. (Thus, the name "adjusted R-squared" is a bit misleading—$\bar{R}^2$ is not actually computed as the square of any quantity.) Whether $\bar{R}^2$ rises or falls depends on whether the contribution of the new variable to the fit of the regression more than offsets the correction for the loss of an additional degree of freedom. The general result (the proof of which is left as an exercise) is as follows.

Change in $\bar{R}^2$ When a Variable is Added to a Regression In a multiple regression, $\bar{R}^2$ will fall (rise) when the variable x is deleted from the regression if the square to the t ratio associated with this variable is greater (less) than 1.

Proof Follow the preceding notations. Suppose at first we have the regression on X, and then we consider adding a new variable z (or equivalently think of the reverse procedure). We construct the difference of $\bar{R}^2$ from the definition of R^2:

$$\begin{cases} R^2_{Xz} = 1 - \frac{u'u}{y'M_0y} \\ R^2_X = 1 - \frac{e'e}{y'M_0y} \end{cases}, \quad \begin{cases} \bar{R}^2_{Xz} = 1 - \frac{n-1}{n-(K+1)}(1 - R^2_{Xz}) \\ \bar{R}^2_X = 1 - \frac{n-1}{n-K}(1 - R^2_X) \end{cases}$$

Since the t ratio is rather complicated, we construct t by proof:

$$\begin{aligned} &\bar{R}^2_{Xz} - \bar{R}^2_X = \frac{n-1}{n-K}(1 - R^2_X) - \frac{n-1}{n-(K+1)}(1 - R^2_{Xz}) \geq 0 \\ \Longleftarrow &\frac{n-1}{n-K} \cdot \frac{e'e}{y'M_0y} \geq \frac{n-1}{n-(K+1)} \cdot \frac{u'u}{y'M_0y} \\ \Longleftarrow &(n-K-1) \cdot e'e \geq (n-K) \cdot u'u \\ \Longleftarrow &(n-K-1) \cdot \left(u'u + c^2(z_*'z_*)\right) \geq (n-K) \cdot u'u \\ \Longleftarrow &(n-K-1) \cdot c^2(z_*'z_*) \geq u'u \\ \Longleftarrow &\frac{c^2}{\dfrac{u'u}{n-K-1} \cdot (z_*'z_*)^{-1}} \geq 1 \\ \Longleftarrow &\left(\frac{c}{\sqrt{\frac{u'u}{n-K-1} \cdot (z_*'z_*)^{-1}}}\right)^2 \geq 1 \end{aligned}$$

In fact, t ratio for z is

$$t_z = \frac{c}{\sqrt{\frac{u'u}{n-K-1} \cdot (z_*'z_*)^{-1}}}$$

Therefore, we conclude that the adjusted R^2 would increase after adding a new variable z if and only if $t_z^2 \geq 1$.

A second difficulty with R^2 concerns the constant term in the model. The proof tat $0 \leq R^2 \leq 1$ requires X to contain a column of 1s. Or more generally speaking, the orthogonal decomposition of variance of y stands upon the fact that a constant term should be included in the model. If NOT, then $M_0e \neq e$ and $e'M_0X \neq \mathbf{0}$, and the term $2e'M_0Xb$ in $y'M_0y = (M_0Xb + M_0e)'(M_0Xb + M_0e)$ in the expansion would not drop out. Consequently, when we compute R^2 as

$$R^2 = 1 - \frac{e'e}{y'M_0y}$$

The result would be unpredictable, that is, R^2 is no longer meaningful. It may even be negative. However, even if you consider an alternative computation,

$$R^2 = \frac{b'X'M_0y}{y'M_0y}$$

This is equally problematic. Again, this calculation will differ from the one obtained with the constant term included; this time, R^2 may be larger than 1. Some computer packages bypass these difficulties by reporting a third "R^2", the squared sample correlation between the actual values of y and the fitted values from the regression. However, even though this time the determinant coefficient falls within 0 and 1, its meaning might be deceptive; it is no longer the proportion of variance explained by the model.