Linear Regression Model in Matrix Form

G.J. LI

1 The Model and Gauss-Markov Assumptions

For the linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \epsilon_i, \qquad i = 1, 2, \ldots, N. \tag{1}$$

y denotes the dependent variable and xs denote the indepedent or explanatory variables or regressors. ϵ is the random error. We use subscript i to index the observation. There are N observations in the sample, while the number of explanatory variables is K. We now stack up all the observations into vectors: $y = (y_1, y_2, \ldots, y_N)' : N \times 1$, $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_K)' : (K+1) \times 1$, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)' : N \times 1$, $\iota = (1, 1, \ldots, 1)' : N \times 1$, $x_k = (x_{1k}, x_{2k}, \ldots, x_{Nk})' : N \times 1$ (any one of the K explanatory variables) and $X = [\iota, x_1, x_2, \ldots, x_K] : N \times (K+1)$. We can then rewrite our model concisely into matrix form.

$$y = X\beta + \epsilon. \tag{2}$$

Apart from assuming y as a linear function of X, we also make the following assumptions:

Assumption 1.1. *$E(\epsilon|X) = 0$, which implies $E(\epsilon) = 0$ and ϵ_i is uncorrelated with any functions of any explanatory variable(s).*

Assumption 1.2. *The matrix X has full column rank, i.e. $rank(X) = K + 1$. In other words, there is no perfect linear relationship among the regressors.*

Assumption 1.3. *$Var(\epsilon|X) = \sigma^2 I_N$. In words, the model is homoskedastic and there is no serial correlation.*

2 Ordinary Least Squares Estimation

We now consider estimating the linear relationship between y and X, i.e. to find out the estimates of β denoted as $\hat{\beta}$. If we have estimates of β, we will at the same time have estimates of y, denoted as $\hat{y} = X\hat{\beta}$. The least squares principle is to minimize the distance between y (the actual value) and $\hat{y}$ (the model fitted value), i.e. $||y - \hat{y}|| = \sqrt{(y - \hat{y})'(y - \hat{y})}$. Since if we minimize $(y-\hat{y})'(y-\hat{y})$, we will at the same time minimize $\sqrt{(y - \hat{y})'(y - \hat{y})}$. Substituting $\hat{y} = X\hat{\beta}$, we need to solve the problem:

$$\min_{\hat{\beta}} SSR(\hat{\beta}) = (y - X\hat{\beta})'(y - X\hat{\beta}), \tag{3}$$

where SSR denotes the sum of squared residuals. We can use the first order condition to find the stationary point. Differentiating $SSR(\hat{\beta})$ with respect to $\hat{\beta}$ gives

$$\underset{(K+1)\times 1}{\frac{\partial (y - X\hat{\beta})'(y - X\hat{\beta})}{\partial \hat{\beta}}} = -2 \underset{(K+1)\times N}{X'} \underset{N\times 1}{(y - X\hat{\beta})} = 0 \tag{4}$$

Hence we have

$$X'X\hat{\beta} = X'y \tag{5}$$

Given Assumption 1.2, we know $X'X$ is positive definite and hence non-singular, PREmultiplying both sides by $(X'X)^{-1}$ yields

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y, \tag{6}$$

which is a linear function of y. So far we know $\hat{\beta}_{OLS}$ is the stationary point for $SSR(\hat{\beta})$. We can go on to check the second order condition by differentiating (4) with respect to $\hat{\beta}$ to obtain $\frac{\partial^2 SSR(\hat{\beta})}{\partial\hat{\beta}\partial\hat{\beta}'} = 2X'X$, which is a positive definite matrix. Hence we know $SSR(\hat{\beta})$ is a convex function and the stationary point $\hat{\beta}_{OLS}$ is indeed a global minimum. Since we now know $\hat{\beta}$, we know $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = P_X y$. From Lecture 1, we know P_X is the orthogonal projection matrix, which projects y orthogonally onto the column space of X. Similarly, given $\hat{\beta}$, we can have the estimated residual $\hat{\epsilon} = y - \hat{y} = (I_N - P_X)y = M_X y$, where M_x is the orthogonal projection matrix which projects y orthogonally onto the space orthogonal to the column space of X. In other words, y is now decomposed as

$$y = P_X y + M_X y \tag{7}$$

with $P_X y$ on the column space of X and $M_X y$ orthogonal to the range of X. Moreover, the following are true.

$$P_X X_1 = X_1, M_X X_1 = 0, P_X M_{X_1} P_X = P_X M_{X_1} = M_{X_1} P_X = M_{X_1} P_X M_{X_1},$$
$$M_{X_1} M_X = M_X M_{X_1} = M_X, \hat{\epsilon} = M_X y = M_X \epsilon = M_{X_1} \hat{\epsilon}, \iota' \hat{\epsilon} = 0, \tag{8}$$

where X_1 is any subset of the columns in X. For example, if $X_1 = \iota$, then $M_{X_1} = M_0 = I_N - \frac{\iota\iota'}{N}$ is the demean matrix. Pre-multiplying both sides of (7) by y' demontrates the Pythagorean theorem: $y'y = \hat{y}'\hat{y} + \hat{\epsilon}'\hat{\epsilon}$. If we pre-multiply both sides by $y'M_0$, we can have

$$y'M_0 y = y'M_0 P_X y + y'M_0 M_X y = y'P_X M_0 P_X y + y'M_X y = \hat{y}'M_0\hat{y} + \hat{\epsilon}'\hat{\epsilon}. \tag{9}$$

If we define the total sum of squares $SST = y'M_0 y$, the explained sum of squares $\hat{y}'M_0\hat{y}$ and the sum of squared residuals $SSR = \hat{\epsilon}'\hat{\epsilon}$, we have $SST = SSE + SSR$. Note that this identity only holds when there is a column of ones in X since $M_0 M_X$ will not be equal to M_X if that is not the case. We now obtain a measure of how well the regression fits the data by using the **coeffficent of determination**

$$R^2 = \frac{\hat{y}'M_0\hat{y}}{y'M_0 y} = 1 - \frac{\hat{\epsilon}'\hat{\epsilon}}{y'M_0 y}. \tag{10}$$

which is between 0 and 1 given X contains a column of ones. It measures the proportion of the total variation in y that is accounted for by variation in the regression. Note that $\frac{\hat{y}'M_0\hat{y}}{y'M_0 y} = \frac{(\hat{y}'M_0 y)^2}{\hat{y}'M_0\hat{y}y'M_0 y} = \frac{\left(\frac{\hat{y}'M_0 y}{N-1}\right)^2}{\frac{\hat{y}'M_0\hat{y}}{N-1}\frac{y'M_0 y}{N-1}} = \frac{\widehat{Cov}(y_i,\hat{y}_i)^2}{\widehat{Var}(\hat{y}_i)\widehat{Var}(y_i)}$.

In fact, R^2 is the squared sample correlation coefficient between y and $\hat{y}$.

3 Finite Sample Properties of OLS

Theorem 3.1. *Given Assumption 1.1 and 1.2, the OLS estimator $\hat{\beta}_{OLS}$ is unbiased for β.*

Proof. Substituting (2) into y on the right hand side of (6) yields $\hat{\beta}_{OLS} = (X'X)^{-1}X'(X\beta + \epsilon) = \beta + (X'X)^{-1}X'\epsilon$. Taking expectation of both sides, we have $E(\hat{\beta}_{OLS}) = \beta$. □

Theorem 3.2. *Under Assumption 1.3, $Var(\hat{\beta}_{OLS}|X) = \sigma^2(X'X)^{-1}$.*

Proof. $Var(\hat{\beta}_{OLS}|X) = Var\left((X'X)^{-1}X'\epsilon|X\right) = (X'X)^{-1}X'Var(\epsilon|X)X(X'X)^{-1} = (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}$. □

Theorem 3.3. *Under Assumption 1.1, 1.2 and 1.3, $\hat{\beta}_{OLS}$ is the best linear unbiased estimator.*

Proof. Any other linear estimator of β can be written as $\tilde{\beta} = A'y$, where A is an $N \times (K+1)$ matrix. For $\tilde{\beta}$ to be unbiased, A can consist of any functions of X. $E(\tilde{\beta}|X) = E(A'X\beta + A'\epsilon|X) = \beta$, if and only if $A'X = I_{K+1} = X'A$. Note that $Var(\tilde{\beta}|X) - Var(\hat{\beta}_{OLS}|X) = \sigma^2\left[A'A - (X'X)^{-1}\right] = \sigma^2\left[A'A - A'X(X'X)^{-1}X'A\right] = \sigma^2 A'M_X A$. Since M_X is an orthogonal projection matrix, $A'M_X A$ is positive semi-definite for any matrix $N \times (K+1)$ A. In other words, $Var(\tilde{\beta}|X)$ is greater than or equal to $Var(\hat{\beta}_{OLS}|X)$ in matrix sense. □

Theorem 3.4. *The unbiased estimator for σ^2 is*

$$\widehat{\sigma^2} = \frac{\hat{\epsilon}'\hat{\epsilon}}{N-K-1} = \frac{y'M_X y}{N-K-1}. \tag{11}$$

Proof. $E(\frac{y'M_Xy}{N-K-1}|X) = E(\frac{\epsilon'M_X\epsilon}{N-K-1}|X) = \frac{1}{N-K-1}E(tr(M_X\epsilon\epsilon'|X)) = \frac{tr(M_XE(\epsilon\epsilon'|X)}{N-K-1} = \frac{\sigma^2tr(M_X)}{N-K-1} = \sigma^2$. □

4 Statistical Inference under Normal Errors

We now add the following assumption, which allows us to obtain standard test statistics for the linear regression model.

Assumption 4.1. *The conditional distribution of ϵ on X is multivariate vector normal with mean 0 and covariance matrix $\sigma^2 I_N$, i.e. $\epsilon|X \sim N(0, \sigma^2 I_N)$.*

Theorem 4.2. *Under Assumption 4.1, $\hat{\beta}_{OLS}|X \sim N(\beta, \sigma^2(X'X)^{-1})$.*

Theorem 4.3. *Denote e_k as the kth column of I_{K+1} and β_k as the kth element in β, under Assumption 4.1, we have*

$$\frac{e_k'\hat{\beta}_{OLS} - \beta_k}{\sqrt{\sigma^2 e_k'(X'X)^{-1}e_k}} \sim N(0,1), \tag{12}$$

$$(N-K-1)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{N-K-1}, \tag{13}$$

$$\frac{e_k'\hat{\beta}_{OLS} - \beta_k}{\sqrt{\widehat{\sigma^2} e_k'(X'X)^{-1}e_k}} \sim t_{N-K-1}. \tag{14}$$

where $\widehat{\sigma^2}$ is defined in (11).

Proof. Since $\hat{\beta}_{OLS}|X \sim N(\beta, \sigma^2(X'X)^{-1})$, we have $e_k'\hat{\beta}_{OLS}|X \sim N(e_k'\beta, \sigma^2 e_k'(X'X)^{-1}e_k)$ and hence $\frac{e_k'\hat{\beta}_{OLS}-\beta_k}{\sqrt{\sigma^2 e_k'(X'X)^{-1}e_k}} \sim N(0,1)$. Note that $(N-K-1)\widehat{\sigma^2} = y'M_Xy = \epsilon'M_X\epsilon$ with $\epsilon \sim N(0, \sigma^2 I_N)$, i.e. $\frac{\epsilon}{\sigma} \sim N(0, I_N)$. Therefore, $(N-K-1)\frac{\widehat{\sigma^2}}{\sigma^2} = \frac{\epsilon'}{\sigma}M_X\frac{\epsilon}{\sigma} \sim \chi^2_{N-K-1}$. Moreover, $e_k'\hat{\beta}_{OLS} - \beta_k = e_k'(X'X)^{-1}X'\epsilon$ and

$M_X X(X'X)^{-1} e_K = 0$[1], given the results on student t distribution in Lecture 1, we have $\frac{e_k'\hat{\beta}_{OLS}-\beta_k}{\sqrt{\frac{N-K-1}{N-K-1}\frac{\widehat{\sigma^2}}{\sigma^2}\sqrt{\sigma^2 e_k'(X'X)^{-1}e_k}}} = \frac{e_k'\hat{\beta}_{OLS}-\beta_k}{\sqrt{\widehat{\sigma^2} e_k'(X'X)^{-1}e_k}} \sim t_{N-K-1}$. □

Theorem 4.4. *Under Assumption 4.1, $\hat{\beta}_{OLS}$ is the minimum variance unbiased estimator of β among any other unbiased linear or nonlinear estimator.*

Theorem 4.5. *For the following linear regression,*

$$\underset{N\times 1}{y} = X_1 \underset{K_1\times 1}{\beta_1} + X_2 \underset{K_2\times 1}{\beta_2} + \epsilon, \tag{15}$$

if $\beta_2 = 0$, then $\frac{(SSR_R - SSR_U)/K_2}{SSR_U/(N-K_1-K_2)} \sim F_{K_2,N-K_1-K_2}$, where $SSR_R = y'M_{X_1}y$, $SSR_U = y'M_X y$ and $X = [X_1, X_2]$.

Proof. Note that $M_X M_{X_1} = M_{X_1} - P_X(I - P_{X_1}) = I - P_{X_1} - P_X + P_{X1} = M_X$ and $M_X(M_{X_1} - M_X) = 0$. We can see that $SSR_R - SSR_U = y'(M_{X_1} - M_X)y = \epsilon'(M_{X_1} - M_X)\epsilon$. Since $M_{X_1} - M_X$ is an idempotent matrix with rank K_2, $\frac{SSR_R - SSR_U}{\sigma^2} \sim \chi^2_{K_2}$. Similarly $\frac{SSR_U}{\sigma^2} \sim \chi^2_{N-K_1-K_2}$. Hence $\frac{(SSR_R - SSR_U)/K_2}{SSR_U/(N-K_1-K_2)} \sim F_{K_2,N-K_1-K_2}$. □

5 Partitioned Regression

Sometimes we may want to focus on a subset of β, rather than all of its elements. We can partition X into $[X_1, X_2]$ and β into $(\beta_1, \beta_2)'$. The minimization problem in (3) now becomes

$$\min_{\hat{\beta}_1, \hat{\beta}_2} (y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2)'(y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2). \tag{16}$$

[1] In other words, $\hat{\beta}_{OLS}$ is independent of $\hat{\epsilon}$ and $\hat{\epsilon}'\hat{\epsilon}$.

If our interest is in β_2, we can concentrate out β_1 in the least squares minimization. To do this, rewrite (5) as

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}. \tag{17}$$

Solving for $\hat{\beta}_1$ in terms of $\hat{\beta}_2$ yeilds

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'(y - X_2\hat{\beta}_2). \tag{18}$$

Note that if the columns in X_1 are orthogonal to those in X_2, we can simply obtain $\hat{\beta}_1$ by regressing y on X_1. Define $M_{X_1} = I - X_1(X_1'X_1)^{-1}X_1'$, $y^* = M_{X_1}y$ and $X_2^* = M_{X_1}X_2$, where y^* and X_2^* are the estimated residuals obtained from regressing y and X_2 on X_1 respectively. Substituting (18) into (16) gives the following least squares problem.

$$\underset{\hat{\beta}_2}{min}(y^* - X_2^*\hat{\beta}_2)'(y^* - X_2^*\hat{\beta}_2). \tag{19}$$

Similar to solving (3), we can obtain the solution

$$\hat{\beta}_2 = (X_2^{*\prime}X_2^*)^{-1}X_2^{*\prime}y^* = (X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}y, \tag{20}$$

which is essentially the **Frisch-Waugh Theorem**.

Suppose β_2 is equal to 0. If we only regress y on X_1, the variance of the OLS estimator is $Var(\tilde{\beta}_1) = \sigma^2(X_1'X_1)^{-1}$. Now if we include the irrelevant variables X_2 into our regression, the estimator's variance is now

$Var(\hat{\beta}_1) = \sigma^2(X_1'M_{X_2}X_1)^{-1}$, where $M_{X_2} = I - X_2(X_2'X_2)^{-1}X_2'$.[2] Since $I - M_{X_2} = P_{X_2}$ is a positive semidefinite matrix, we know $Var(\tilde{\beta}_1)$ is no larger than $Var(\hat{\beta}_1)$ in matrix sense. If X_1 just contains one explanatory variable, $Var(\hat{\beta}_1) = \frac{\sigma^2}{SSR_{12}}$, where $SSR_{12} = SST_1(1 - R_{12}^2)$ is the sum of squared residuals obtained by regressing X_1 on X_2. In other words, if we include more irrelevant explanatory variables into our regression, which are highly correlated with the relevant variables, we will inflate the estimators' variances associated with the relevant variables, making the estimates less accurate and the t-statistics less significant.

6 More about R-Squared

Note that the objective function in (3) and (19) have the same value at the minimum. Substituting (6) into (3) and (20) into (19) gives $SSR = y'M_X y = y'M_{X_1}y - y'M_{X_1}X_2(X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}y$ and hence[3] $y'M_{X_1}y = y'M_X y + y'M_{X_1}X_2(X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}y$. Note that while $y'M_{X_1}y$ is the SSR obtained by regressing y on X_1, $y'M_X y$ is the SSR obtained by regressing y on X_1 and X_2. Since $y'M_{X_1}y \geq y'M_X y$, the R-squared from the first regression is never smaller than the one from the second regression. That is for finite sample (N is fixed), the more regressors, whether they are relevant or not, we include, the higher R-squared we can obtain. If the

[2] For $E = \begin{bmatrix} E_{11} & E_{12} \\ n\times n & n\times m \\ E_{21} & E_{22} \\ m\times n & m\times m \end{bmatrix}$, $E^{-1} = \begin{bmatrix} E_{11}^{-1} + E_{11}^{-1}E_{12}(E_{22} - E_{21}E_{11}^{-1}E_{12})^{-1}E_{21}E_{11}^{-1} & -E_{11}^{-1}E_{12}(E_{22} - E_{21}E_{11}^{-1}E_{12})^{-1} \\ -(E_{22} - E_{21}E_{11}^{-1}E_{12})^{-1}E_{21}E_{11}^{-1} & (E_{22} - E_{21}E_{11}^{-1}E_{12})^{-1} \end{bmatrix} =$
$\begin{bmatrix} (E_{11} - E_{12}E_{22}^{-1}E_{21})^{-1} & -(E_{11} - E_{12}E_{22}^{-1}E_{21})^{-1}E_{12}E_{22}^{-1} \\ -E_{22}^{-1}E_{21}(E_{11} - E_{12}E_{22}^{-1}E_{21})^{-1} & E_{22}^{-1} + E_{22}^{-1}E_{21}(E_{11} - E_{12}E_{22}^{-1}E_{21})^{-1}E_{12}E_{22}^{-1} \end{bmatrix}$.

[3] Note that $M_{X_1} - M_X = M_{X_1}X_2(X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1} = M_{X_1}P_X M_{X_1}$. We can also have $y'M_{X_1}'M_{X_1}y = y'M_{X_1}'M_X M_{X_1}y + y'M_{X_1}P_X M_{X_1}y$.

sample size is equal to the number of elements in β and $rank(X) = N$, $P_X = X(X'X)^{-1}X' = XX^{-1}(X')^{-1}X' = I$ and the regression will have perfect fit with $R^2 = 1$. The **adjusted** R^2, denoted as $\bar{R}^2$, incorporates a penalty for over-parameterization and takes the following form.

$$\bar{R}^2 = 1 - \frac{\hat{\epsilon}'\hat{\epsilon}/(N-K-1)}{y'M_0y/(N-1)} = 1 - \frac{N-1}{N-K-1}(1-R^2) \tag{21}$$

It will rise (fall) when one regressor is deleted from the regression if the t statistic associated with the variable is less (greater) than 1. (See Greene, 2011, 3.5.)

Also note that if we just regress y on X_1, analogical to (9) we can decompose the total variation $y'M_0y = y'P_{X_1}M_0P_{X_1}y + y'M_{X_1}y = y'P_{X_1}M_0P_{X_1}y + y'M_{X_1}X_2(X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}y + y'M_Xy$. This is essentially what the R function `anova()` does.[4] If $X_1 = \iota$ and X_2 is a collection of the explanatory variables, then $M_0P_{X_1} = 0$ and the R-squared is now $\frac{y'M_0X_2(X_2'M_0X_2)^{-1}X_2'M_0y}{y'M_0y}$, which is the squared sample canonical correlation between y and X_2. So another interpretation for $\hat{\beta}_{OLS}$ is the linear combination of the explanatory variables, whose squared correlation with the dependent variable is equal to the squared sample canonical correlation.[5]

[4] The result depends on the order of X_1 and X_2 in the regression.

[5] In fact, $\hat{\beta}_{OLS,slope} = a\frac{a'X_2'M_0y}{a'X_2'M_0X_2a}$, where a is the eigenvector associated with the non-zero eigenvalue for $\frac{1}{y'M_0y}(X_2'M_0X_2)^{-1}X_2'M_0yy'M_0X_2$. You do not need to memorize this for your exam.

References

[1] W.H. Greene, *Econometric Analysis*, 7th ed., Pearson Education, Chapter 2-5, 2011.

[2] J.M. Wooldridge, *Introductory Econometrics (a modern approach)*, 4th ed., South Western College, Appendix D and E, 2008.