OLS using Matrix Algebra

This version: 29 November 2018

Intermediate Econometrics / Forecasting Class Notes
Anthony Tay

The basic linear regression model can be expressed conveniently in matrix form. We present here the main OLS algebraic and finite sample results in matrix form:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_{K-1} X_{K-1,i} + \epsilon_i \ , \ i = 1, 2, ..., N.$$

We can express this relationship for every i by writing

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \cdots & X_{K-1,1} \\ 1 & X_{1,2} & X_{2,2} & \cdots & X_{K-1,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,N} & X_{2,N} & \cdots & X_{K-1,N} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{K-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

or simply

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

We will have to assume that the $(N \times K)$ matrix X has full column rank: there is no $(K \times 1)$ vector $\mathbf{c}$ such that $\mathbf{Xc} = 0$. That is, no column is perfectly correlated with any linear combination of other columns; no column is proportional to another column; no column are all zeros; the first column is the only column of constants. This assumption guarantees that the inverse of $(\mathbf{X}'\mathbf{X})$ exists.

The $\mathbf{X}$ matrix can be partitioned into columns, or rows. Emphasizing columns (which emphasizes variables):

$$\mathbf{X} = \left[\begin{array}{c:c:c:c:c} 1 & X_{1,1} & X_{2,1} & \cdots & X_{K-1,1} \\ 1 & X_{1,2} & X_{2,2} & \cdots & X_{K-1,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,N} & X_{2,N} & \cdots & X_{K-1,N} \end{array}\right] = \begin{bmatrix} \mathbf{i} & \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_{K-1} \end{bmatrix}$$

where $\mathbf{i}$ is an $(N \times 1)$ column vector of ones, and $\mathbf{X}_i$, $i = 1, 2, ..., K-1$ are $(N \times 1)$ column vectors containing observations of each of the variables. To emphasize the rows of $\mathbf{X}$, we can write:

$$\mathbf{X} = \left[\begin{array}{ccccc} 1 & X_{1,1} & X_{2,1} & \cdots & X_{K-1,1} \\ \hdashline 1 & X_{1,2} & X_{2,2} & \cdots & X_{K-1,2} \\ \hdashline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hdashline 1 & X_{1,N} & X_{2,N} & \cdots & X_{K-1,N} \end{array}\right] = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \cdots \\ \mathbf{x}_N' \end{bmatrix}$$

where $\mathbf{x}_i' = \begin{bmatrix} 1 & X_{i,1} & X_{i,2} & \cdots & X_{i,K-1} \end{bmatrix}$. This format emphasizes each observation in the regression:

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \ , \ i = 1, 2, ..., N.$$

OLS Formula and Algebraic Properties

If $\hat{\boldsymbol{\beta}}$ is some estimator for $\boldsymbol{\beta}$, then the fitted values are $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and the residuals are $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. The OLS method is to choose $\hat{\boldsymbol{\beta}}$ such that the sum of squared residuals ("SSR") is minimized. The sum of squared residuals can be calculated as

$$\begin{aligned} \text{SSR} = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

We can simplify in the last step because $\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ are scalars, and one is the transpose of the other. The transpose of a scalar is itself, thus $\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$. The OLS method is:

$$\text{OLS:} \qquad \hat{\boldsymbol{\beta}}_{OLS} = \text{argmin}_{\hat{\boldsymbol{\beta}}}\mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

The first-order conditions for this optimization problem is:

$$\frac{\partial SSR}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} = 0.$$

The solution to the FOC solves the minimization problem because

$$\frac{\partial^2 SSR}{\partial \hat{\boldsymbol{\beta}} \partial \hat{\boldsymbol{\beta}}'} = 2\mathbf{X}'\mathbf{X}$$

which is positive definite: for any $(K \times 1)$ vector $\mathbf{c} \neq 0$, we have

$$\mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c} = (\mathbf{X}\mathbf{c})'(\mathbf{X}\mathbf{c}) > 0.$$

Rewriting the FOC in terms of the residuals, we see that

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}'\hat{\boldsymbol{\epsilon}} = 0$$

We describe the condition $\mathbf{X}'\hat{\boldsymbol{\epsilon}} = 0$ by saying that $\mathbf{X}$ and $\hat{\boldsymbol{\epsilon}}$ are "orthogonal". If $\mathbf{X}$ contains the constant term, then the FOC can be written as

$$\mathbf{X}'\hat{\boldsymbol{\epsilon}} = \begin{bmatrix}\mathbf{i} & \mathbf{X}_1 & \cdots & \mathbf{X}_{K-1}\end{bmatrix}'\hat{\boldsymbol{\epsilon}} = \begin{bmatrix}\mathbf{i}' \\ \mathbf{X}_1' \\ \vdots \\ \mathbf{X}_{K-1}'\end{bmatrix}\hat{\boldsymbol{\epsilon}}.$$

which says that that (i) the residuals sum to zero, and (ii) the sample covariance between the residuals and each of the regressors is zero. In other words, OLS chooses $\hat{\boldsymbol{\beta}}$ so that (i) and (ii) hold. If $\mathbf{X}$ does not include a constant term, then the residuals do not necessarily sum to zero, and the sample covariance between the residuals and the regressors are not zero.

Solving the FOC gives:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

From this point, we will drop the 'OLS' subscript, and take for granted that $\hat{\boldsymbol{\beta}}$ is the estimator obtained using the OLS method, not any any other method (non-OLS estimators will be denoted in other ways).

Because $\hat{\boldsymbol{\beta}}$ takes the form of $\mathbf{Ay}$, each estimator in $\hat{\boldsymbol{\beta}}$ is a weighted sum of Y_i, $i = 1, 2, ..., N$. For this reason, the OLS estimator is said to be a 'linear estimator'. There is also another way of writing $\hat{\boldsymbol{\beta}}$ that is sometimes useful. Writing $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N'\end{bmatrix}$$

the OLS estimator $\hat{\boldsymbol{\beta}}$ can be expressed as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left\{ \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{bmatrix} \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix} \right\}^{-1} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} \\ &= \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i Y_i\end{aligned}$$

This form of the OLS estimator emphasizes the role that averages play in the estimation of $\boldsymbol{\beta}$, since we can write the estimator as

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i Y_i \right).$$

Fitted Values The OLS fitted values are

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.\end{aligned}$$

The matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the “hat matrix” because it “puts a hat on $\mathbf{y}$”. It is also called the projection matrix, and often denoted $\mathbf{P}$. (We won’t discuss the reason for the ‘projection’ terminology here). This matrix has some interesting and useful properties:

1. The hat matrix is symmetric:

$$\begin{aligned}\mathbf{P}' &= (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\ &= \mathbf{X}''[(\mathbf{X}'\mathbf{X})^{-1}]'\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}\end{aligned}$$

   where we use the fact that the inverse of a symmetric matrix is symmetric.

2. The hat matrix is idempotent:

$$\mathbf{PP} = \mathbf{X}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{\mathbf{I}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}$$

3. The trace of the hat matrix (the sum of its diagonal elements) is equal to K, the number

of columns in $\mathbf{X}$.

$$tr(\mathbf{P}) = tr(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = tr((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = tr(\mathbf{I}_{K\times K)}) = K.$$

4. The matrix $(\mathbf{I} - \mathbf{P})\mathbf{X} = \mathbf{0}$:

$$\begin{aligned}(\mathbf{I} - \mathbf{P})\mathbf{X} &= \mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\\ &= \mathbf{X} - \mathbf{X}\\ &= \mathbf{0}\end{aligned}$$

The matrix $\mathbf{I} - \mathbf{P}$ is often denoted by $\mathbf{M}$ because it eliMinates $\mathbf{X}$.

OLS Residuals The OLS residuals are computed as

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{P}\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{y} = \mathbf{M}\mathbf{y} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{M}\boldsymbol{\epsilon}$$

which shows nicely the relationship between the residuals and the actual noise terms. The fact that $\mathbf{M}$ is symmetric and idempotent leads to a neat formula for the sum of squared residuals:

$$\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = (\boldsymbol{\epsilon}'\mathbf{M}')(\mathbf{M}\boldsymbol{\epsilon}) = \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}.$$

The fact that $\mathbf{X}$ and $\hat{\boldsymbol{\epsilon}}$ are orthogonal means that $\hat{\mathbf{y}}$ and $\hat{\boldsymbol{\epsilon}}$ are orthogonal:

$$\hat{\mathbf{y}}'\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{\boldsymbol{\epsilon}} = 0$$

This gives a nice relationship between the sum of squares of Y_i, $\hat{Y}_i$, and $\hat{\epsilon}_i$:

$$\begin{aligned}\mathbf{y} &= \hat{\mathbf{y}} + \hat{\boldsymbol{\epsilon}}\\ \mathbf{y}'\mathbf{y} &= (\hat{\mathbf{y}} + \hat{\boldsymbol{\epsilon}})'(\hat{\mathbf{y}} + \hat{\boldsymbol{\epsilon}})\\ \mathbf{y}'\mathbf{y} &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + 2\hat{\mathbf{y}}'\hat{\boldsymbol{\epsilon}} + \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}\\ \mathbf{y}'\mathbf{y} &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}\end{aligned}$$

This is often described as “Sum of Squared Total = Sum of Squared Explained + Sum of Squared Residuals”, or SST = SSE + SSR. (Geometrically, this is nothing more than Pythagoras’ Theorem, albeit in N-dimensions.) We will use another version of this relationship (which *does* require a constant term in the regression) to develop a measure of “goodness-of-fit”.

Exercise: Show that the matrix $\mathbf{M}$ is symmetric and idempotent, with trace equal to $N - K$.

Exercise: Consider another "$\mathbf{M}$" matrix which we'll call "$\mathbf{M}_0$":

$$\mathbf{M}_0 = \mathbf{I}_{(N\times N)} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'$$

where $\mathbf{i}$ is an $(N \times 1)$ vectors of ones. Show that $\mathbf{M}_0$ is symmetric and idempotent. Suppose we pre-multipy an $(N \times 1)$ vector $\mathbf{y}$ by $\mathbf{M}_0$, i.e., take $\mathbf{M}_0\mathbf{y}$. How does this transform $\mathbf{y}$? (Ans: it substracts from each element in $\mathbf{y}$ the sample average of the elements of $\mathbf{y}$.)

Goodness-of-Fit When the constant term is included in regression, then the sample mean of the residuals is zero. Then

$$\begin{aligned}
\mathbf{y} &= \hat{\mathbf{y}} + \hat{\boldsymbol{\epsilon}} \\
\mathbf{M}_0\mathbf{y} &= \mathbf{M}_0\hat{\mathbf{y}} + \mathbf{M}_0\hat{\boldsymbol{\epsilon}} \qquad \text{subtracting means} \\
(\mathbf{M}_0\mathbf{y})'(\mathbf{M}_0\mathbf{y}) &= (\mathbf{M}_0\hat{\mathbf{y}} + \mathbf{M}_0\hat{\boldsymbol{\epsilon}})'(\mathbf{M}_0\hat{\mathbf{y}} + \mathbf{M}_0\hat{\boldsymbol{\epsilon}}) \qquad \text{Taking sum of squares} \\
\mathbf{y}'\mathbf{M}_0\mathbf{y} &= \hat{\mathbf{y}}'\mathbf{M}_0\hat{\mathbf{y}} + 2\hat{\boldsymbol{\epsilon}}'\mathbf{M}_0\hat{\mathbf{y}} + \hat{\boldsymbol{\epsilon}}'\mathbf{M}_0\hat{\boldsymbol{\epsilon}} \\
\mathbf{y}'\mathbf{M}_0\mathbf{y} &= \hat{\mathbf{y}}'\mathbf{M}_0\hat{\mathbf{y}} + \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}.
\end{aligned}$$

This says

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \overline{\hat{Y}})^2 + \sum_{i=1}^{N}\hat{\epsilon}_i^2$$

This is the "centered" version of SST = SSE + SSE. Dividing throughout by $N - 1$ (or N), we see that this is a decomposition of the sample variance of Y_i

$$\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{Y}_i - \overline{\hat{Y}})^2 + \frac{1}{N-1}\sum_{i=1}^{N}(\hat{\epsilon}_i)^2$$
$$\text{sample var}[Y_i] = \text{sample var}[\hat{Y}_i] + \text{sample var}[\epsilon_i]$$

The centered version of SST = SSE + SSR is often used to define a measure of goodness-of-fit. Dividing

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \overline{\hat{Y}})^2 + \sum_{i=1}^{N}\hat{\epsilon}_i^2$$

throughout by $\sum_{i=1}^N (Y_i - \overline{Y})^2$, we get

$$1 = \frac{\sum_{i=1}^N (\hat{Y}_i - \overline{\hat{Y}})^2}{\sum_{i=1}^N (Y_i - \overline{Y})^2} + \frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{\sum_{i=1}^N (Y_i - \overline{Y})^2}$$

The R^2 measure of Goodness-of-fit is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{\sum_{i=1}^N (Y_i - \overline{Y})^2} = 1 - \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{\mathbf{y}'\mathbf{M}_0\mathbf{y}}.$$

If we have a perfect fit, then $\sum_{i=1}^N \hat{\epsilon}_i^2 = 0$, and $R^2 = 1$. If $\hat{Y}_i = \overline{Y}$ for all $i = 1, 2, ..., N$, (which would be the case if $\hat{\beta}_0 = \hat{\beta}_1 = ... = \hat{\beta}_{K-1} = 0$) then $R^2 = 0$. All intermediate fits will result in R^2 between 0 and 1.

The R^2 gets its name from the fact that it is the square of the sample correlation coefficient between Y_i and $\hat{Y}_i$ (when the regression includes a constant). This stems from the fact that

$$\hat{\mathbf{y}}'\mathbf{M}_0\hat{\mathbf{y}} = (\mathbf{y} - \hat{\boldsymbol{\epsilon}})'\mathbf{M}_0\hat{\mathbf{y}} = \mathbf{y}'\mathbf{M}_0\hat{\mathbf{y}} = (\mathbf{M}_0\mathbf{y})'(\mathbf{M}_0\hat{\mathbf{y}})$$

Since $\hat{\boldsymbol{\epsilon}}'\mathbf{M}_0\hat{\mathbf{y}} = \hat{\boldsymbol{\epsilon}}'\hat{\mathbf{y}} = 0$. From this we can see that

$$R^2 = \frac{\hat{\mathbf{y}}'\mathbf{M}_0\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{M}_0\mathbf{y}} = \frac{\hat{\mathbf{y}}'\mathbf{M}_0\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{M}_0\mathbf{y}}\frac{\hat{\mathbf{y}}'\mathbf{M}_0\hat{\mathbf{y}}}{\hat{\mathbf{y}}'\mathbf{M}_0\hat{\mathbf{y}}} = \frac{((\mathbf{M}_0\mathbf{y})'(\mathbf{M}_0\hat{\mathbf{y}}))^2}{(\mathbf{y}'\mathbf{M}_0\mathbf{y})(\hat{\mathbf{y}}'\mathbf{M}_0\hat{\mathbf{y}})}$$

This says that

$$\begin{aligned} R^2 &= \frac{\left(\sum_{i=1}^N (Y_i - \overline{Y}_i)(\hat{Y}_i - \overline{\hat{Y}})\right)^2}{\sum_{i=1}^N (Y_i - \overline{Y}_i)^2 \sum_{i=1}^N (\hat{Y}_i - \overline{\hat{Y}})^2} \\ &= \left[\frac{\frac{1}{N-1}\sum_{i=1}^N (Y_i - \overline{Y}_i)(\hat{Y}_i - \overline{\hat{Y}})}{\sqrt{\frac{1}{N-1}\sum_{i=1}^N (Y_i - \overline{Y}_i)^2}\sqrt{\frac{1}{N-1}\sum_{i=1}^N (\hat{Y}_i - \overline{\hat{Y}})^2}}\right]^2 \end{aligned}$$

i.e., R^2 is the square of the sample correlation coefficient between Y_i and $\hat{Y}_i$.

We can still use the R^2 if a constant is not included in the regression, or if other methods of estimation are used. It's just that the R^2 can then fall below zero. (All that an R^2 less than zero means is that the fit provided by the estimated model is worse than the fit obtained via the sample mean of $\{Y_i\}_{i=1}^N$). Of course, without the constant term the R^2 no longer has the interpretation of a squared sample correlation.

(Finite Sample) Statistical Properties of the OLS Estimator

If it is the case that, conditional on all every $X_{k,i}, k = 1, 2, ..., K-1, i = 1, 2, ..., n$, that each error term is zero mean, has variance σ^2, and are uncorrelated among themselves, then we can write

$$E[\boldsymbol{\epsilon}|\mathbf{X}] = \begin{bmatrix} E[\epsilon_1|\mathbf{X}] \\ E[\epsilon_2|\mathbf{X}] \\ \vdots \\ E[\epsilon_N|\mathbf{X}] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}_{(n\times 1)}$$

and

$$E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}] = \begin{bmatrix} E[\epsilon_1^2|\mathbf{X}] & E[\epsilon_1\epsilon_2|\mathbf{X}] & \cdots & E[\epsilon_1\epsilon_n|\mathbf{X}] \\ E[\epsilon_2\epsilon_1|\mathbf{X}] & E[\epsilon_2^2|\mathbf{X}] & \cdots & E[\epsilon_2\epsilon_n|\mathbf{X}] \\ \vdots & \vdots & \ddots & \vdots \\ E[\epsilon_N\epsilon_1|\mathbf{X}] & E[\epsilon_N\epsilon_2|\mathbf{X}] & \cdots & E[\epsilon_N^2|\mathbf{X}] \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}_{(n\times n)}$$

Given these properties, we can make several statements about the OLS estimator $\hat{\boldsymbol{\beta}}$. To prove these properties, we will use

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \end{aligned}$$

(1) $\hat{\boldsymbol{\beta}}$ is unbiased

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= E[\boldsymbol{\beta}|\mathbf{X}] + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}|\mathbf{X}] \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\epsilon}|\mathbf{X}] \\ &= \boldsymbol{\beta} \end{aligned}$$

if $E[\boldsymbol{\epsilon}|\mathbf{X}] = \mathbf{0}$. It follows that $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

Note that $E[\boldsymbol{\epsilon}|\mathbf{X}] = \mathbf{0}$ is the only assumption used. No assumption regarding the variance-covariance matrix of $\boldsymbol{\epsilon}$ is used, so the unbiasedness result holds even if the errors are heteroskedastic (have different variances) or are correlated among themselves.

(2) $var[\hat{\boldsymbol\beta}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

$$\begin{aligned}
var[\hat{\boldsymbol\beta}|\mathbf{X}] &= E[(\hat{\boldsymbol\beta} - E[\hat{\boldsymbol\beta}])(\hat{\boldsymbol\beta} - E[\hat{\boldsymbol\beta}])'|\mathbf{X}] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol\epsilon\boldsymbol\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol\epsilon\boldsymbol\epsilon'|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

(3) $\hat{\boldsymbol\beta}$ has the ‘smallest’ variance among all linear unbiased estimators.

What this means is that given any other unbiased estimator of the form $\tilde{\boldsymbol\beta} = \mathbf{A}\mathbf{y}$, we have for any $(K \times 1)$ vector $\mathbf{c} \neq 0$,

$$var[\mathbf{c}'\hat{\boldsymbol\beta}] \leq var[\mathbf{c}'\tilde{\boldsymbol\beta}].$$

To show this, let $\mathbf{B} = \mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then

$$\begin{aligned}
\tilde{\boldsymbol\beta} = \mathbf{A}\mathbf{y} &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{B}]\mathbf{y} \\
&= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{B}](\mathbf{X}\boldsymbol\beta + \boldsymbol\epsilon) \\
&= \boldsymbol\beta + \mathbf{B}\mathbf{X}\boldsymbol\beta + [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{B}]\boldsymbol\epsilon
\end{aligned}$$

First we restrict our arguments to $\mathbf{A}$ such that $\mathbf{A}\mathbf{y}$ is unbiased. From the above, we see that

$$E[\mathbf{A}\mathbf{y}|\mathbf{X}] = \boldsymbol\beta + \mathbf{B}\mathbf{X}\boldsymbol\beta.$$

Unbiasedness of $\mathbf{A}\mathbf{y}$ requires that we choose $\mathbf{A}$ such that $\mathbf{B}\mathbf{X} = \mathbf{0}$. Then

$$\begin{aligned}
var[\tilde{\boldsymbol\beta}|\mathbf{X}] &= E[(\tilde{\boldsymbol\beta} - \boldsymbol\beta)(\tilde{\boldsymbol\beta} - \boldsymbol\beta)'|\mathbf{X}] \\
&= E[[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{B}]\boldsymbol\epsilon\boldsymbol\epsilon'[\mathbf{B}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]|\mathbf{X}] \\
&= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{B}\mathbf{B}'] \\
&= var[\hat{\boldsymbol\beta}] + \sigma^2\mathbf{B}\mathbf{B}'
\end{aligned}$$

It follows that for any $(K \times 1)$ vector $\mathbf{c} \neq 0$,

$$var[\mathbf{c}'\tilde{\boldsymbol\beta}] = var[\mathbf{c}'\hat{\boldsymbol\beta}] + \sigma^2\mathbf{c}'\mathbf{B}\mathbf{B}'\mathbf{c} = var[\mathbf{c}'\hat{\boldsymbol\beta}] + \sigma^2(\mathbf{B}'\mathbf{c})'\mathbf{B}'\mathbf{c} \geq var[\mathbf{c}'\hat{\boldsymbol\beta}].$$

We say that $\hat{\boldsymbol{\beta}}$ is Best among all Linear Unbiased Estimators, or "BLUE".

To obtain numerical estimates of $var[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, we require an estimator of σ^2.

(4) The expected value of the sum of squared residuals is

$$E[\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}|\mathbf{X}] = (N-K)\sigma^2.$$

We use the fact that the trace of a scalar is itself, and that the trace is a linear operator, so the expectation of a trace is the trace of the expectation:

$$\begin{aligned} E[\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}|\mathbf{X}] &= E[tr(\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}})|\mathbf{X}] \\ &= E[tr(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon})|\mathbf{X}] \\ &= E[tr[\mathbf{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}] \\ &= tr(\mathbf{M}E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}]) \\ &= \sigma^2 tr(\mathbf{M}) \\ &= (N-K)\sigma^2. \end{aligned}$$

An unbiased estimator for σ^2 is therefore

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{N-K}.$$

(5) If the error terms have a multivariate normal distribution, i.e.

$$\boldsymbol{\epsilon} \sim MN(\mathbf{0}, \sigma^2\mathbf{I}),$$

then $\hat{\boldsymbol{\beta}}$ is distributed multivariate normal:

$$\hat{\boldsymbol{\beta}} \sim MN(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

(6) Suppose we wish to test a hypothesis of the form

$$H_0 : \mathbf{r}'\boldsymbol{\beta} = r_0 \quad \text{vs} \quad H_A : \mathbf{r}'\boldsymbol{\beta} \neq r_0.$$

For instance, in the regression

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i \ , \ i = 1, 2, ..., 20,$$

we might wish to test $H_0 : \beta_1 + \beta_2 = 1$ vs $H_A : \beta_1 + \beta_2 \neq 1$. Here

$$\mathbf{r} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad r_0 = 1.$$

We can use the fact that

$$\mathbf{r}'\hat{\boldsymbol{\beta}} \sim N(\mathbf{r}'\boldsymbol{\beta}, \sigma^2 \mathbf{r}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{r}),$$

from which it follows (stated here without further elaboration) that

$$t = \frac{\mathbf{r}'\hat{\boldsymbol{\beta}} - \mathbf{r}'\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2 \mathbf{r}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{r}}} \sim t_{(N-K)}.$$

To test the hypothesis, we use the rule: reject H_0 if

$$\left| \frac{\mathbf{r}'\hat{\boldsymbol{\beta}} - r_0}{\sqrt{\hat{\sigma}^2 \mathbf{r}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{r}}} \right| > c_\alpha$$

where c_α is set so that the probability of rejecting a correct null is α (usually set at 0.01, 0.05, or 0.1). For the example presented above, we reject the hypothesis at 0.05 level of significance if

$$|t| = \left| \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{\sqrt{\hat{\sigma}^2 \mathbf{r}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{r}}} \right|$$

is greater than 2.11, which is the 0.975 percentile of the t-distribution with 17 degrees of freedom.

(7) Suppose we wish to test several hypotheses simultaneously. For example, in the regression

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i \ , \ i = 1, 2, ..., 20,$$

we might wish to test the hypotheses: $H_0 : \beta_1 + \beta_2 = 1$ and $\beta_3 = 0$ vs $H_A : \beta_1 + \beta_2 \neq 1$ or $\beta_3 \neq 0$ (or both). We can express this as testing

$$H_0 : \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{vs} \quad H_A : \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \neq \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

which we will write as $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ vs $\mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$. To test this hypothesis, we use the fact that imposing the null hypothesis as restrictions on the OLS regression will lead to a higher sum of squared residuals (OLS minimizes SSR; a restricted minimization cannot lead to a lower SSR). We then compare the restricted and unrestricted SSRs to see if imposing the restrictions leads to a substantially larger SSR, which would indicate that at least one of the hypotheses is false.

Denote the residuals from the restricted OLS estimation by $\hat{\boldsymbol{\epsilon}}_R$. We do not give the formula for the restricted least squares estimator here, but imposing restrictions on the fit is usually straghtforward. In our example, we have

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1,i} + (1-\beta_1)X_{2,i} + 0X_{3,i} + \epsilon_i \ , \\ &= \beta_0 + \beta_1(X_{1,i} - X_{2,i}) + X_{2,i} + \epsilon_i \ , \ i = 1,2,...,20. \end{aligned}$$

We would regress $Y_i - X_{2,i}$ on a constant and $(X_{1,i} - X_{2,i})$, and calculate the restricted residuals as

$$\hat{\epsilon}_i = Y_i - \hat{\beta}_0^R - \hat{\beta}_1^R(X_{1,i} - X_{2,i}) - X_{2,i} \ , \ i = 1,2,...,20$$

where the superscript R indicates restricted least squares estimators. In any case, we do not actually have to carry out the restricted least squares fit, because it can be shown (proof omitted) that:

$$\hat{\boldsymbol{\epsilon}}_R'\hat{\boldsymbol{\epsilon}}_R - \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$$

where $\hat{\boldsymbol{\beta}}$ is the *unrestricted* OLS estimator.

It turns out (again proof omitted) that if the null is true, then

$$F = \frac{(\hat{\boldsymbol{\epsilon}}_R'\hat{\boldsymbol{\epsilon}}_R - \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}})/J}{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/(N-K)} \sim F_{J,N-K}$$

where J is the number of restrictions imposed. We are testing whether $\hat{\boldsymbol{\epsilon}}_R'\hat{\boldsymbol{\epsilon}}_R - \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}$ is “too big”, so we use a one-sided (right-side) test, and reject the null if F is greater than c_α, where c_α is the 0.90, 0.95, or 0.99 percentile of the $F_{J,N-K}$ distribution.

Example We illustrate the concepts presented here with the regression

$$log(EARNINGS_i) = \beta_0 + \beta_1 S_i + \beta_2 SF_i + \beta_3 SM_i + \beta_4 WEXP_i + \beta_5 TENURE_i + \beta_6 MALE_i + \epsilon_i$$

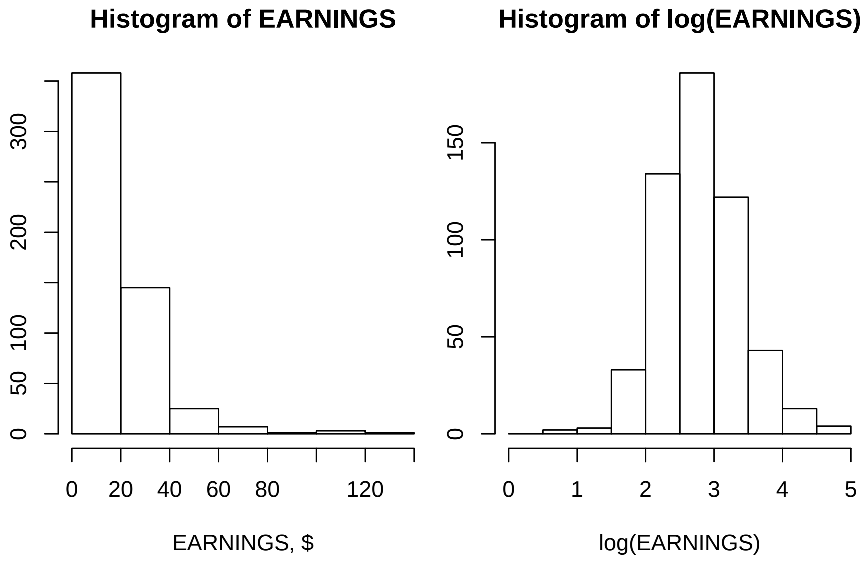
from the dataset **earnings.csv**. (Details in the r comments below)

```
# The datafile earnings.csv contains data on
# hourly earnings ($), height (inches), male (1, 0),
# s (yrs schooling), sf (yrs father's schooling), sm (yrs mother's schooling),
# tenure (yrs in current job), wexp (yrs work experience)
# from a sample of 540 people in 2002
# Data from US National Longitudinal Survey of Youth 1979, or NLYS79)
# via Dougherty textbook
earnings<-read.csv("earnings.csv")
earnings$MALE <- as.factor(earnings$MALE)
summary(earnings)
```

```
##     EARNINGS          HEIGHT      MALE          S               SF
##  Min.   :  2.13   Min.   :48.00   0:270   Min.   : 7.00   Min.   : 0.00
##  1st Qu.: 10.76   1st Qu.:64.00   1:270   1st Qu.:12.00   1st Qu.:10.00
##  Median : 16.00   Median :67.00           Median :13.00   Median :12.00
##  Mean   : 19.64   Mean   :67.67           Mean   :13.67   Mean   :11.84
##  3rd Qu.: 23.16   3rd Qu.:71.00           3rd Qu.:16.00   3rd Qu.:14.00
##  Max.   :120.19   Max.   :80.00           Max.   :20.00   Max.   :20.00
##        SM            TENURE              WEXP
##  Min.   : 0.00   Min.   : 0.01923   Min.   : 1.154
##  1st Qu.:11.00   1st Qu.: 1.93750   1st Qu.:14.596
##  Median :12.00   Median : 4.69231   Median :17.510
##  Mean   :11.58   Mean   : 7.03397   Mean   :16.900
##  3rd Qu.:12.00   3rd Qu.:10.98077   3rd Qu.:20.197
##  Max.   :20.00   Max.   :24.94231   Max.   :23.558
```

We use *log(EARNINGS)* because *EARNINGS* is very naturally skewed and non-negative.

```
par(mfrow=c(1,2))
par(mar=c(5,2,3,0.5))
hist(earnings$EARNINGS, main="Histogram of EARNINGS", xlab="EARNINGS, $",
     ylab=NULL, breaks=seq(0,140,20))
hist(log(earnings$EARNINGS), main="Histogram of log(EARNINGS)",
     xlab="log(EARNINGS)", ylab=NULL, breaks=seq(0,5,0.5))
```



The fit of regressions with dependent variables that exhibit such a high degree of skewness tends to be somewhat unsatisfactory, and the associated error terms tend to be non-normal – this is certainly the case here, since *EARNINGS* is bounded below by zero. Furthermore, using *log(EARNINGS)* means that the coefficients now have the interpretation of percent changes, e.g., we can say that the difference in hourly earnings between males and females is $100\hat{\beta}_6$ **percent**, controlling for work experiences, years of schooling, etc.

```
earnings.lm <- lm(log(EARNINGS)~S+SF+SM+WEXP+TENURE+MALE, data=earnings)
summary(earnings.lm)
```

```
##
## Call:
## lm(formula = log(EARNINGS) ~ S + SF + SM + WEXP + TENURE + MALE,
##     data = earnings)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15330 -0.29451 -0.00227  0.26747  1.85188
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.534188   0.164125   3.255  0.00121 **
## S           0.107327   0.009599  11.181  < 2e-16 ***
## SF          0.016972   0.007671   2.212  0.02736 *
## SM          0.001066   0.009513   0.112  0.91083    
## WEXP        0.021528   0.005235   4.112 4.53e-05 ***
## TENURE      0.011294   0.003443   3.280  0.00111 ** 
## MALE1       0.267785   0.042368   6.321 5.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4789 on 533 degrees of freedom
## Multiple R-squared:  0.3454, Adjusted R-squared:  0.338 
## F-statistic: 46.88 on 6 and 533 DF,  p-value: < 2.2e-16
```

The coefficient estimates are obtained from $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The corresponding standard errors are the square root of the diagonal elements of $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ where $\hat{\sigma}^2 = \frac{1}{533}\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}$, the square root of which is reported as the "Residual Standard Error":

```
sqrt(sum(earnings.lm$residuals^2)/earnings.lm$df.residual)
```

```
## [1] 0.4788524
```

The entire coefficient variance matrix is available from:

```
vcov(earnings.lm)
```

```
##               (Intercept)             S            SF            SM
## (Intercept)  2.693698e-02 -1.058149e-03  1.195743e-06 -3.110010e-04
## S           -1.058149e-03  9.214367e-05 -1.650263e-05 -1.540571e-05
## SF           1.195743e-06 -1.650263e-05  5.884838e-05 -4.092152e-05
## SM          -3.110010e-04 -1.540571e-05 -4.092152e-05  9.048808e-05
## WEXP        -5.287878e-04  1.196352e-05 -9.428842e-08 -2.546471e-06
## TENURE       4.724871e-05 -2.695823e-06  6.251336e-07  5.675369e-07
## MALE1        2.927493e-04 -2.230107e-05 -2.212113e-06 -5.506823e-06
##                      WEXP        TENURE         MALE1
## (Intercept) -5.287878e-04  4.724871e-05  2.927493e-04
## S            1.196352e-05 -2.695823e-06 -2.230107e-05
## SF          -9.428842e-08  6.251336e-07 -2.212113e-06
## SM          -2.546471e-06  5.675369e-07 -5.506823e-06
## WEXP         2.740373e-05 -6.299929e-06 -4.599204e-05
## TENURE      -6.299929e-06  1.185620e-05 -2.575173e-06
## MALE1       -4.599204e-05 -2.575173e-06  1.795007e-03
```

(the coefficient standard errors are the square root of the diagonal elements of this matrix.) The t-values are for testing (individually) if the true value of the coefficients are zero, so are equal to the coefficient estimates divided by the standard errors. There are packages and tricks for testing other individual linear hypothesis, but it is straightforward to do it directly. We do so here to illustrate the expressions given in these notes. E.g. to test $\beta_5 = \beta_6$, we can do:

```r
get_ttest_pval<-function(r,r0,mdl){
  betahat <- matrix(mdl$coefficients, ncol=1)
  tstat <- (crossprod(r,betahat) - r0) / sqrt(t(r) %*% vcov(mdl) %*% r)
  (1-pt(tstat, mdl$df.residual))*2
}
r <- matrix(c(0,0,0,0,1,-1,0), ncol=1)
r0 <- 0
print(paste("The t-test p-val is ",
            round(as.numeric(get_ttest_pval(r,r0,earnings.lm)),4),".", sep=""))
```

```
## [1] "The t-test p-val is 0.1559."
```

We do not reject the hypothesis.

The R^2 is calculated as

```r
SST <- sum((log(earnings$EARNINGS)-mean(log(earnings$EARNINGS)))^2)
SSR <- sum(earnings.lm$residuals^2)
Rsqr <- 1-SSR/SST
Rsqr
```

```
## [1] 0.3454113
```

or simply

```r
cor(log(earnings$EARNINGS), earnings.lm$fitted.values)^2
```

```
## [1] 0.3454113
```

Dividing both numerator and denominator of SSR/SST by $N-1$ shows that the R^2 is

$$1 - \frac{\text{smp.var}(\hat{\epsilon}_i)}{\text{smp.var}(Y_i)}.$$

The adjusted-R^2 is computed using unbiased versions on the variance esimators

$$Adj\text{ -}R^2 = 1 - \frac{\frac{1}{N-K}\sum_{i=1}^{N}(\hat{\epsilon}_i)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2}.$$

```
SST <- sum((log(earnings$EARNINGS)-mean(log(earnings$EARNINGS)))^2)
SSR <- sum(earnings.lm$residuals^2)
adjRsqr <- 1-(SSR/earnings.lm$df.residual)/
             (SST/(earnings.lm$df.residual+length(earnings.lm$coefficients)-1))
adjRsqr
```

```
## [1] 0.3380426
```

The adjusted-R^2 has the advantage that unlike the R^2, adding new variables to the equation might possibly reduce its value (or adding restrictions might possibly increase it.) However, it is not used often anymore as a model selection tool.

The F-statistic reported is for testing $H_0 : \beta_1 = ... = \beta_6 = 0$, i.e.,

$$F = \frac{R^2/J}{(1-R^2)/(N-K)}$$

where J is the number of restrictions (in this case, 6).

```
# Rsqr computed earlier
F = (Rsqr/(length(earnings.lm$coefficients)-1)) /
    ((1-Rsqr)/earnings.lm$df.residual)
F
```

```
## [1] 46.8753
```

To test other multiple linear hypothesis, the easiest way is to run the restricted regression, and then compute the F-statistic using

$$F = \frac{(\hat{\boldsymbol{\epsilon}}'_R\hat{\boldsymbol{\epsilon}}_R - \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}})/J}{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/(N-K)} \sim F_{J,N-K}.$$

To test that parents' years of schooling are not significant factors determining earnings, i.e., $\beta_3 = \beta_4 = 0$, we can do the following:

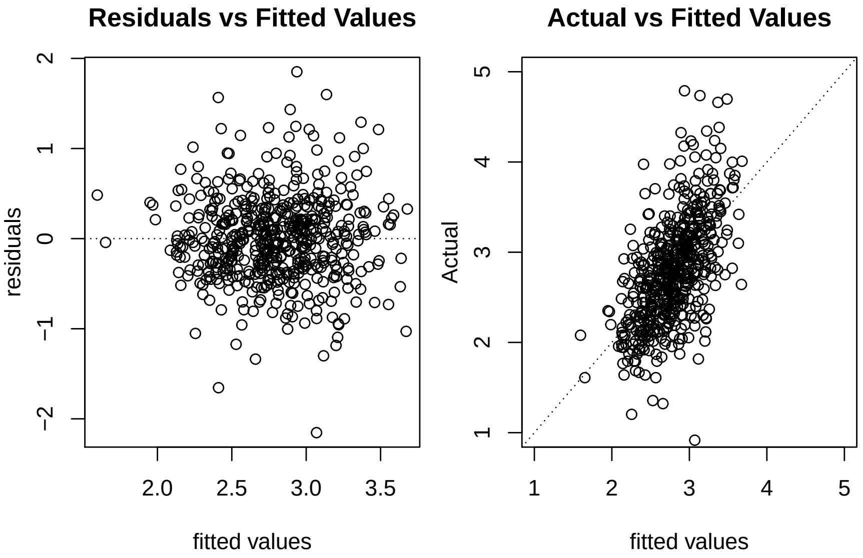
```
earnings.rlm <- lm(log(EARNINGS)~S+WEXP+TENURE+MALE, data=earnings)
SSRu <- sum(earnings.lm$residuals^2)
SSRr <- sum(earnings.rlm$residuals^2)
F = ((SSRr-SSRu)/2)/(SSRu/earnings.lm$df.residual)
# pf() gives F distribution percentiles
print(paste("The F-statistic is ", round(F,4), " with p-value ",
            round(1-pf(F,2,earnings.lm$df.residual),4),".", sep=""))
```

```
## [1] "The F-statistic is 3.7822 with p-value 0.0234."
```

We reject the hypothesis that both coefficients are zero.

We end with a discussion of visualizations of the OLS fit. It is always a good idea to plot the residuals (below left) to see if there is anything unusual. Nothing catches the eye, except one or two residuals that seem a little large in size relative to the others. We may want to see if that observation is a 'high-leverage' observation (not done here).

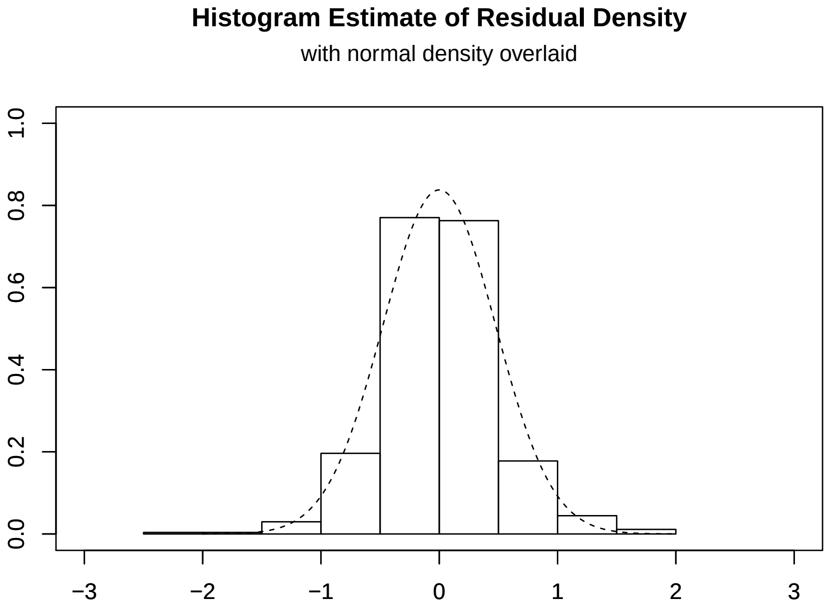
```
par(mfrow=c(1,2))
par(mar=c(5,3,3,0.8))
yhat <- earnings.lm$fitted.values
ehat <- earnings.lm$residuals
plot(x=yhat, y=ehat, main="Residuals vs Fitted Values",
     xlab="fitted values", ylab="")
title(ylab="residuals", line=2.2)
abline(h=0, lty='dotted')
plot(x=yhat, y=log(earnings$EARNINGS), main="Actual vs Fitted Values",
     xlab="fitted values", ylab="", xlim=c(1,5), ylim=c(1,5))
title(ylab="Actual", line=2.2)
abline(c(0,1), lty="dotted")
```



It is tempting to plot the actual vs the fitted values, as in the figure on the right. One would expect, if the fit is good, for the observations to lie on the 45-degree line. This plot, however, seems to lead the eye towards errors that are right-angles to the 45-degree line, rather than vertically, leading to a perception of a biased fit.

Perhaps more importantly, we might want to ask if the errors are actually normally distributed, which the t- and F-tests require.

```r
par(mar=c(2,2,4,2))
hist(earnings.lm$residuals, xlim=c(-3,3), ylim=c(0,1),
     prob=TRUE, main=NULL, ylab=NULL, xlab=NULL)
title(main="Histogram Estimate of Residual Density", line=3)
mtext("with normal density overlaid", side=3, line=1.5)
x <- seq(-2, 2, length=1000)
fx <- dnorm(x, mean=0, sd=sqrt(var(earnings.lm$residuals)))
par(new=TRUE)
plot(x, fx, type="l", lty="dashed", lwd=1, ylim=c(0,1),
     ylab="", xlab="", xlim=c(-3,3))
```



The histogram is scaled so the y-axis is probability rather than frequency, and so that the area of the histogram sums to one (so it is an actual density estimate, and we can compare it visually to a density). The overlaid normal distribution has mean zero and standard deviation equal to that of the residuals. We can calculate the skewness and kurtosis coefficients of the residuals for a better sense of ‘closeness’ to normality:

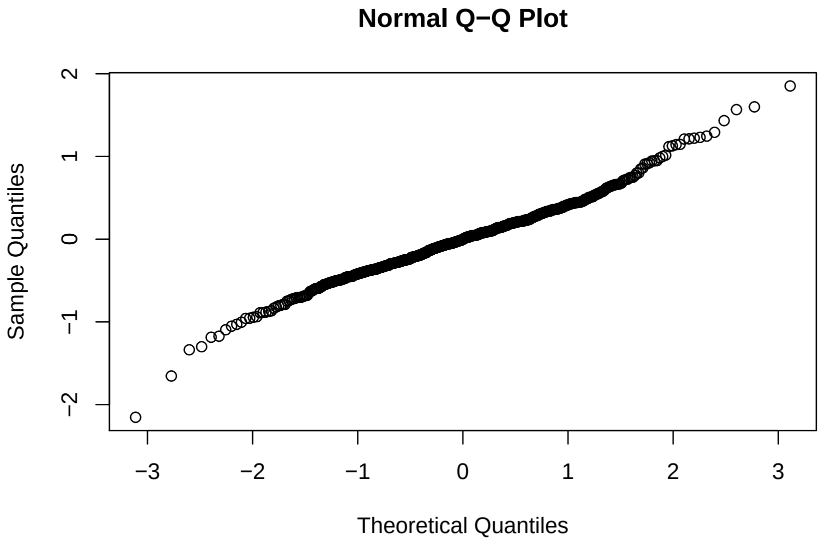
```r
N = 540
m2 = sum(earnings.lm$residuals^2)/540  # We use the fact that
m3 = sum(earnings.lm$residuals^3)/540  # OLS residuals have zero means
m4 = sum(earnings.lm$residuals^4)/540  # when calculating these moments
```

```
S = m3/(sqrt(m2)^3)
K = m4/(m2^2)
print(paste("Skewness Coefficient: ", round(S,4),
            ", Kurtosis Coefficient: ", round(K,4), ".", sep=""))
```

```
## [1] "Skewness Coefficient: 0.1146, Kurtosis Coefficient: 4.5812."
```

There appears to be a some excess kurtosis, though not much. This can also be seen from the "qqplot" of the residuals. The following plots the values of the quantiles of the residuals against the values of the corresponding quantiles of the normal distribution. If the residuals come from a normal distribution, we would expect the scatterplot to fall in a straight line:

```
qqnorm(earnings.lm$residuals)
```



As expected, the qqplot shows slightly heavier tails than what one might expect from a normal distribution. The non-normality seems very mild, nonetheless, and the t- and F-tests done here are probably reasonably accurate. In any case, it seems somewhat hopeful to expect finite sample results to hold exactly, which requires the unlikely scenario that the many assumptions made here hold exactly (e.g. homoskedasticity, normality, correct functional form, etc.). In later sections, we will loosen some of the assumptions made in this section, and use methods that are appropriate under those looser conditions.