Multiple Linear Regression Model Derivations*

Alexandros Theloudis†

November 7, 2015

1 The model

The multiple linear regression model can be written in the population as:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

where:

- y is the independent variable
- β_0 is the constant
- x_j, $j = \{1, 2, \ldots, k\}$ is the independent variable
- β_j is the coefficient on x_j
- u is the error term

Advantages of controlling for more variables:

- Zero conditional mean assumption more reasonable
  - Closer to estimating causal/ceteris paribus effects (everything else equal)
- More general functional form
- Better prediction of y / better fit of the model

1.1 Assumptions

The Multiple Linear Regression Model is usually thought of in the context of the following assumptions:

1. *MLR.1 - Linearity*
   The model is linear in parameters. Notice that

   $$y = \beta_0 + \beta_1 \ln x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

   is also linear in parameters. For $z = \ln x_1$ the above model can be rewritten as:

   $$y = \beta_0 + \beta_1 z + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

   which is clearly linear.

*Prepared for the ECON2007 Practical Lecture on 18 October 2013. It draws material from Magne Mogstad's lecture notes entitled "Multiple Regression: Estimation".

†Alexandros Theloudis: Research Student, Department of Economics, University College London, Gower Street, WC1E 6BT London (email: alexandros.theloudis.10@ucl.ac.uk).

2. *MLR.2 - Random sample*
   The sample is a random draw from the population. The data in the sample are $\{(x_{i1}, \ldots, x_{ik}, y_i) : i = 1, \ldots, n\}$, where $\{x_{i1}, \ldots, x_{ik}, y_i\}$ are *i.i.d.* (independent and identically distributed).

3. *MLR.3 - Full rank or no perfect collinearity*
   $x_j \neq c$ and there is no exact linear relationship among any x_j in the population, *i.e.* x_j cannot be written as $\sum_{-j} \alpha_{-j} x_{-j}$, where α_{-j} are constants and $-j = 1, \ldots, j-1, j+1, \ldots, k$.

4. *MLR.4 - Zero conditional mean*
   Conditional on $x_1, \ldots, x_k$ the mean of u is 0, *i.e.* $\mathbb{E}[u|x_1, \ldots, x_k] = 0$. Notice that this assumption also implies that $\mathbb{E}[u] = 0$ and $\mathbb{E}[x_j u] = 0$, $\forall j$. (*Can you prove this?*)

5. *MLR.5 - Homoscedasticity*
   The variance of u is constant and independent of x, *i.e.* $\mathbb{E}[u^2|x_1, \ldots, x_k] = \sigma^2$.

For the remaining part of this note, we will assume for simplicity that the model can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

(*i.e.* with two independent variables only). All the results derived herein can be easily extended to accommodate the general case of k independent variables.

2 The OLS estimator

We want to estimate the following MLR

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i. \tag{1}$$

Let $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ be the OLS estimators for β_0, β_1, and β_2 respectively. The following expressions are important:

$$\begin{aligned}
y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{u}_i \\
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \\
\hat{u}_i &= y_i - \hat{y}_i
\end{aligned}$$

Notice that these expressions resemble the ones derived for fitted values and residuals in the context of the Simple Linear Regression Model. How can we interpret $\hat{\beta}_1$ in this context? $\hat{\beta}_1$ measures the *ceteris-paribus* change in y given an one unit change in x_1; put differently, it measures the change in y given an one unit change in x_1, holding x_2 fixed. Obviously, $\hat{\beta}_2$ has an analogous interpretation.

How do we actually obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ in the context of the Multiple Linear Regression Model? There are three equivalent ways: the minimization of the *sum of squared residuals*, the *partialling-out* method, and, finally, a method which makes use of a number of *moment conditions.*

2.1 Sum of squared residuals

We can obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ by minimizing the sum of squared residuals

$$Q(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} \right)^2$$

We derive the first order conditions of $Q(\hat{\boldsymbol{\beta}})$ with respect to $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ and we set them equal to 0:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} \right) = 0$$

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_1} &= -2\sum_{i=1}^{n} x_{i1}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) = 0 \\
\frac{\partial Q}{\partial \beta_2} &= -2\sum_{i=1}^{n} x_{i2}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) = 0
\end{aligned}
$$

These are three equations with three unknowns; solving them simultaneously we get $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ (you do not have to remember the following formulae):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_{i2}-\bar{x}_2)^2 \sum_{i=1}^n (y_i-\bar{y})(x_{i1}-\bar{x}_1) - \sum_{i=1}^n (y_i-\bar{y})(x_{i2}-\bar{x}_2) \sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2 \sum_{i=1}^n (x_{i2}-\bar{x}_2)^2 - \left(\sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)\right)^2}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2 \sum_{i=1}^n (y_i-\bar{y})(x_{i2}-\bar{x}_2) - \sum_{i=1}^n (y_i-\bar{y})(x_{i1}-\bar{x}_1) \sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2 \sum_{i=1}^n (x_{i2}-\bar{x}_2)^2 - \left(\sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)\right)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

where $\bar{x}_j = n^{-1}\sum_{i=1}^n x_{ij}$ and $\bar{y} = n^{-1}\sum_{i=1}^n y_i$.

2.2 The partialling-out method

A more intuitive way to obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ is the following. First, we estimate a Simple Linear Regression of x_1 on x_2 (and any other independent variables in the context of a general k-variable model):

$$x_{i1} = \alpha_0 + \alpha_1 x_{i2} + r_{i1}$$

where r_{i1} is an error term. We will use the Simple Linear Regression tools to get the OLS estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ which will then allow us to construct the residual:

$$\hat{r}_{i1} = x_{i1} - \hat{\alpha}_0 - \hat{\alpha}_1 x_{i2}$$

How can we interpret residual $\hat{r}_{i1}$? It is the variation in x_{i1} that is left after removing the variation in x_{i2}.

Properties of residual $\hat{r}_{i1}$
The usual properties apply to $\hat{r}_{i1}$:

1. Residuals sum to zero: $\sum_{i=1}^n \hat{r}_{i1} = 0$
2. Residuals are orthogonal to regressors: $\sum_{i=1}^n \hat{r}_{i1} x_{i2} = 0$
3. Sum of products between residual and dependent variable equals sum of squared residuals: $\sum_{i=1}^n \hat{r}_{i1} x_{i1} = \sum_{i=1}^n \hat{r}_{i1}^2$

   Proof:

$$
\begin{aligned}
\sum_{i=1}^{n} \hat{r}_{i1} x_{i1} &= \sum_{i=1}^{n} \hat{r}_{i1}\left(x_{i1} - \hat{\alpha}_0 - \hat{\alpha}_1 x_{i2} + \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2}\right) \\
&= \sum_{i=1}^{n} \hat{r}_{i1}\left(\hat{r}_{i1} + \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2}\right) \\
&= \sum_{i=1}^{n} \left(\hat{r}_{i1}^2 + \hat{\alpha}_0 \hat{r}_{i1} + \hat{\alpha}_1 x_{i2} \hat{r}_{i1}\right) \\
&= \sum_{i=1}^{n} \hat{r}_{i1}^2 + \hat{\alpha}_0 \underbrace{\sum_{i=1}^{n} \hat{r}_{i1}}_{=0} + \hat{\alpha}_1 \underbrace{\sum_{i=1}^{n} \hat{r}_{i1} x_{i2}}_{=0} = \sum_{i=1}^{n} \hat{r}_{i1}^2
\end{aligned}
$$

After having obtained $\hat{r}_{i1}$, we regress y_i on $\hat{r}_{i1}$ and a constant:

$$y_i = \theta_0 + \theta_1 \hat{r}_{i1} + v_i$$

This is nothing but the Simple Linear Regression Model again. The OLS estimate of the slope coefficient is

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n \left(\hat{r}_{i1} - \bar{\hat{r}}_1\right) y_i}{\sum_{i=1}^n \left(\hat{r}_{i1} - \bar{\hat{r}}_1\right)^2};$$

and as $\bar{\hat{r}}_1 = 0$, we can rewrite this as:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2}. \tag{2}$$

How can we interpret this estimated coefficient? $\hat{\theta}_1$ measures the change in y which is due to an one-unit change in x_1 after having x_2 partialled out; put differently, $\hat{\theta}_1$ measures the change in y which is due to an one-unit change in x_1 holding x_2 fixed. But that is exactly what $\hat{\beta}_1$ is measuring too (return to the discussion at the beginning of this section).

Formally, replacing $\hat{r}_{i1}$ in (**??**) with an analytical expression consisting of x_{i1} and x_{i2} only (obtained from the regression in the first stage), one can see that $\hat{\theta}_1$ is exactly the same as $\hat{\beta}_1$ from the minimization of squared residuals. In the next section we will prove that $\hat{\theta}_1$ (or $\hat{\beta}_1$; used interchangeably hereafter) is an unbiased estimate of the unknown β_1 in the population.

Going back to (**??**), how can we obtain $\hat{\beta}_2$? Using the same two-step approach one can show that

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n \hat{r}_{i2} y_i}{\sum_{i=1}^n \hat{r}_{i2}^2}.$$

More generally, in a model with k independent variables

$$\hat{\beta}_j = \frac{\sum_{i=1}^n \hat{r}_{ij} y_i}{\sum_{i=1}^n \hat{r}_{ij}^2}, \qquad j = 1, 2, \ldots, k$$

where $\hat{r}_{ij}$ is the OLS residuals from a regression of x_j on the other explanatory variables and a constant.

Finally, the estimated constant is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

where $\bar{x}_j = n^{-1} \sum_{i=1}^n x_{ij}$ and $\bar{y} = n^{-1} \sum_{i=1}^n y_i$. In the general case with k variables, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \cdots - \hat{\beta}_k \bar{x}_k$.

2.3 Moment conditions

As in the context of the Simple Linear Regression, one can use the sample counterparts of the $2+1$ moment conditions (or $k+1$ in the case of a general k-variable model) that follow from *MLR.4*:

$$\begin{aligned} \mathbb{E}[u] &= 0 \\ \mathbb{E}[x_j u] &= 0 \qquad j = 1, 2 \end{aligned}$$

The derivation is then straightforward. *Can you show it?*

3 Unbiasedness

Using the partialling-out method we showed that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2}$$

We will now show analytically that $\hat{\beta}_1$ is an unbiased estimator of the true β_1 in the population:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&\overset{MLR.1}{=} \frac{\sum_{i=1}^n \hat{r}_{i1} \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i\right)}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&= \frac{\sum_{i=1}^n \hat{r}_{i1}\beta_0}{\sum_{i=1}^n \hat{r}_{i1}^2} + \frac{\sum_{i=1}^n \hat{r}_{i1}\beta_1 x_{i1}}{\sum_{i=1}^n \hat{r}_{i1}^2} + \frac{\sum_{i=1}^n \hat{r}_{i1}\beta_2 x_{i2}}{\sum_{i=1}^n \hat{r}_{i1}^2} + \frac{\sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&= \beta_0 \underbrace{\frac{\sum_{i=1}^n \hat{r}_{i1}}{\sum_{i=1}^n \hat{r}_{i1}^2}}_{=0} + \beta_1 \underbrace{\frac{\sum_{i=1}^n \hat{r}_{i1} x_{i1}}{\sum_{i=1}^n \hat{r}_{i1}^2}}_{=1} + \beta_2 \underbrace{\frac{\sum_{i=1}^n \hat{r}_{i1} x_{i2}}{\sum_{i=1}^n \hat{r}_{i1}^2}}_{=0} + \frac{\sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&= \beta_1 + \frac{\sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2}
\end{aligned}$$

Taking expectations conditional on x_1 and x_2 we get:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1 | (x_{i1}, x_{i2}) \forall i] &= \beta_1 + \mathbb{E}[\frac{\sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2} | (x_{i1}, x_{i2}) \forall i] \\
&\overset{MLR.2}{=} \beta_1 + \mathbb{E}[\frac{\sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2} | x_1, x_2] \\
&\overset{*}{=} \beta_1 + \frac{\sum_{i=1}^n \hat{r}_{i1} \mathbb{E}[u_i | x_1, x_2]}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&\overset{MLR.4}{=} \beta_1
\end{aligned}$$

where $*$ implies that as we are conditioning on x_1 and x_2, $\hat{r}_{i1}$ is no longer random and can therefore exit the expectations operator. Unbiasedness ($\mathbb{E}[\hat{\beta}_1] = \beta_1$) now follows from the Law of Iterated Expectations. Similarly, it can be shown that $\hat{\beta}_2$ is also an unbiased estimator for β_2.

To prove unbiasedness of $\hat{\beta}_0$, one has to notice that:

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \\
&= \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \\
&= \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x}_1 + (\beta_2 - \hat{\beta}_2)\bar{x}_2
\end{aligned}$$

Then, unbiasedness follows from the unbiasedness of $\hat{\beta}_1$ and $\hat{\beta}_2$.

4 Variance of OLS estimator

Recall that

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2}.$$

We derive the variance as follows:

$$\begin{aligned}
Var\left(\hat{\beta}_1 | (x_{i1}, x_{i2}) \forall i\right) &= Var\left((\sum_{i=1}^n \hat{r}_{i1}^2)^{-1} \sum_{i=1}^n \hat{r}_{i1} u_i | (x_{i1}, x_{i2}) \forall i \right) \\
&\overset{MLR.2}{=} Var\left((\sum_{i=1}^n \hat{r}_{i1}^2)^{-1} \sum_{i=1}^n \hat{r}_{i1} u_i | x_1, x_2 \right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{*}{=} \quad (\sum_{i=1}^{n} \hat{r}_{i1}^2)^{-2} Var\left(\sum_{i=1}^{n} \hat{r}_{i1} u_i | x_1, x_2\right) \\
&\overset{*\&**}{=} \quad (\sum_{i=1}^{n} \hat{r}_{i1}^2)^{-2} \sum_{i=1}^{n} \hat{r}_{i1}^2 Var\left(u_i | x_1, x_2\right) \\
&\overset{MLR.5}{=} \quad \sigma^2 / \sum_{i=1}^{n} \hat{r}_{i1}^2 \\
&= \quad \sigma^2 / (SST_1 (1 - R_1^2))
\end{aligned}
$$

where $*$ implies that as we are conditioning on x_1 and x_2, $\hat{r}_{i1}$ no longer varies and can therefore exit the variance operator; $**$ implies that the variance of the sum is equal to the sum of the variances only if the covariances are 0. To see this last point, think about the simple case with $n = 2$. In this case:

$$
\begin{aligned}
Var\left(\sum_{i=1}^{2} \hat{r}_{i1} u_i | x_1, x_2\right) &= Var\left(\hat{r}_{11} u_1 + \hat{r}_{21} u_2 | x_1, x_2\right) \\
&= Var\left(\hat{r}_{11} u_1 | x_1, x_2\right) + Var\left(\hat{r}_{21} u_2 | x_1, x_2\right) + Cov\left(\hat{r}_{11} u_1, \hat{r}_{21} u_2 | x_1, x_2\right) \\
&= \sum_{i=2}^{2} Var\left(\hat{r}_{i1} u_i | x_1, x_2\right) + \hat{r}_{11} \hat{r}_{21} Cov\left(u_1, u_2 | x_1, x_2\right) \\
&= \sum_{i=2}^{2} Var\left(\hat{r}_{i1} u_i | x_1, x_2\right)
\end{aligned}
$$

as the last covariance is 0 because of *MLR.2*. Finally, notice that $\sum_{i=1}^{n} \hat{r}_{i1}^2 = SST_1(1 - R_1^2)$ because $R_j^2 = 1 - \frac{\sum_{i=1}^{n} \hat{r}_{i1}^2}{\sum_{i=1}^{n} (x_{i1} - \overline{x}_1)^2} \equiv 1 - \frac{\sum_{i=1}^{n} \hat{r}_{i1}^2}{SST_1}$.

5 Final remarks

There are many unbiased estimators of β_j, $j = \{0, 1, 2, \ldots, k\}$. A theorem usually referred to as the Gauss-Markov theorem states that under assumptions *MLR.1* through *MLR.5*, $\hat{\beta}_j$ is the best linear unbiased estimator (BLUE) of β_j, $j = \{0, 1, 2, \ldots, k\}$ because:

- **B**est = smallest variance
- **L**inear in parameters
- **U**nbiased: $\mathbb{E}[\hat{\beta}_j] = \beta_j$
- **E**stimator: $\hat{\beta}_j = function(data)$