A Few Notes on Classical Linear Regression Models

David Meyer

dmm@{1-4-5.net,uoregon.edu}

Last update: April 20, 2018

1 Introduction

This note introduces the Classical Linear Regression Model (CLRM) and discusses the assumptions underlying the model. In particular, three sets of assumptions underly the CRLM:

1. Assumptions respecting the formulation of the Population Regression Equation (PRE)
2. Assumptions respecting the statistical properties of the random error term (ε) and the dependent variable (X).
3. Assumptions respecting the properties of the sample data.

The Population Regression Function (PRF), depicted in Figure 1, takes the form $f(X_i) = \mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$. Here the subscript i indexes the set of observations. For each population value X_i of X then, there is a conditional distribution of population values Y and a corresponding conditional distribution of population random errors ε such that

- $\boldsymbol{\varepsilon_i | X_i = Y_i - \mathbb{E}(Y_i|X_i) = Y_i - \beta_0 - \beta_i X_i}$
- $\boldsymbol{Y_i | X_i = \mathbb{E}(Y_i|X_i) + \varepsilon_i = \beta_0 + \beta_i X_i + \varepsilon_i}$

2 The Population Regression Equation (PRE)

Assumption A1: The first assumption we need is that the Population Regression Equation, or PRE, takes the form

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \text{\# for the population} \tag{1}$$
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \text{\# for an individual instance (unit) } i \tag{2}$$

The PRE (A1) gives the value of the regressand (dependent variable) Y for each value of the regressor (independent variable) X. The i subscripts on Y and X are used to denote individual population or sample values of the dependent variable Y and the independent variable X.

Assumption A1 implies that the PRE can be written as the sum of two parts: the PRF ($\mathbb{E}(Y_i|X_i)$) and a random error term (ε_i):

- $f(X_i) = \mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$
  where β_0 and β_1 are regression coefficients (or parameters), *the true population values of which are unknown*, and X_i is the value of the regressor X corresponding to the value Y_i of Y, and [1]

- $\varepsilon_i = Y_i - f(X_i) = Y_i - (\beta_0 + \beta_1 X_i) = Y_i - \beta_0 - \beta_1 X_i$
  where ε_i is a random error term (sometimes called a stochastic error term).
  ε_i is the difference between the observed Y_i value and the value of the population regression function for the corresponding value X_i of the regressor X.

Note that the random error terms ε_i iare unobservable because the true population values of the regression coefficients β_0 and β_1 are unknown.

The PRE (A1) incorporates three distinct sub-assumptions:

- Additive Error Term

  The error term ε_i is *additive* in the PRE, which implies that $\frac{\partial Y_i}{\partial \varepsilon_i} = 1$.

- *Linearity* in parameters/coefficients

  This assumption means that the partial derivative of Y_i with respect to each of the regression coefficients is a function only of known constants and/or the regressor X_i; it is not a function of any unknown parameters. That is, $\frac{\partial Y_i}{\beta_j} = f_j(X_i)$, for $j = 0, 1$.

- Parameter or Coefficient Constancy

  This assumption means that the regression coefficients β_0 and β_1 do not vary across observations. That is, they do not vary with the observation subscript i. This means that if $\beta_{j,i}$ is the value of the j-th regression coefficient for observation i, then $\beta_{j,i} = \beta_i$ is a constant for $\forall i$ and $j = 0, 1$.

[1] It is quite amazing that for linear models $\mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$

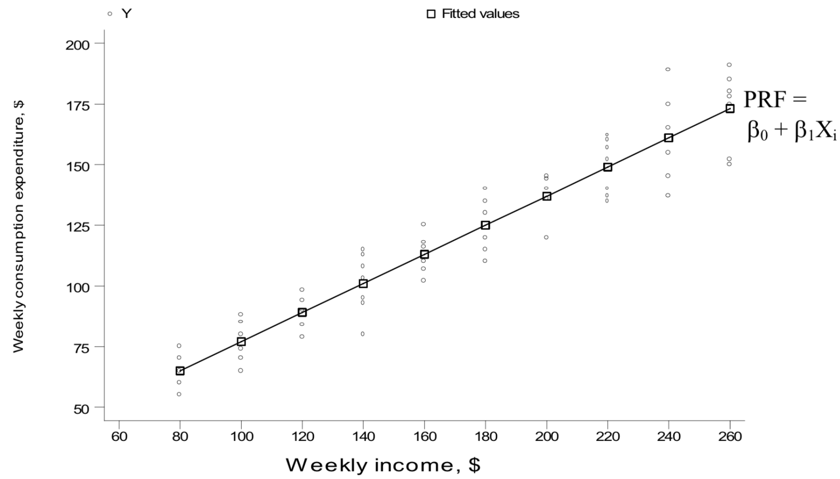

Figure 1: Example Population Regression Function: $f(X_i) = \mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$

2.1 Properties of the random error term ε_i

Assumption A2: *Zero Conditional Mean Error.* The conditional mean, or conditional expectation, of the random error terms ε_i for any given value X_i of the regressor X is equal to zero. That is

$$\mathbb{E}[\varepsilon \mid X] = 0 \text{ or } \mathbb{E}[\varepsilon_i \mid X_i] = 0 \; \forall i \tag{3}$$

This assumption is saying two things:

1. The conditional mean of the random error term ε is the same for all population values of X; it does not depend, either linearly or nonlinearly on X.
2. The common conditional population mean of ε for all values of X is zero.

There are several implications of assumption A2. The first implication is that the *unconditional mean* of the population values of the random error term ε equals zero. That is

$$\mathbb{E}[\varepsilon \mid X] = 0 \Rightarrow \mathbb{E}[\varepsilon] = 0 \tag{4}$$
$$\mathbb{E}[\varepsilon_i \mid X_i] = 0 \Rightarrow \mathbb{E}[\varepsilon_i] = 0 \tag{5}$$

This follows from the *law of iterated expectation*, which states that $\mathbb{E}[\mathbb{E}[\varepsilon \mid X]] = \mathbb{E}[\varepsilon]$. In particular, since $\mathbb{E}[\varepsilon \mid X] = 0$ by A2, we see that

$$\begin{aligned}
\mathbb{E}[\varepsilon] &= \mathbb{E}[\mathbb{E}[\varepsilon \mid X]] && \text{\# law of iterated expectation} \quad (6)\\
&= \mathbb{E}[0] && \text{\# since } \mathbb{E}[\varepsilon \mid X] = 0 \text{ (A2)} \quad (7)\\
&= 0 && \text{\# since } \mathbb{E}[c] = c \text{ for constant } c \quad (8)
\end{aligned}$$

One way to think about this is that if the conditional mean of ε for each and every population value of X equals zero, then the mean of these zero conditional means must also be zero.

Another implication of A2 is that the population values X_i of the regressor X and ε_i of the random error term ε have zero covariance. That is, the population values of X and ε are uncorrelated. That is,

$$\mathbb{E}[\varepsilon \mid X] = 0 \Rightarrow \text{Cov}(X, \varepsilon) = \mathbb{E}[X\varepsilon] = 0 \tag{9}$$
$$\mathbb{E}[\varepsilon_i \mid X_i] = 0 \Rightarrow \text{Cov}(X_i, \varepsilon_i) = \mathbb{E}[X_i\varepsilon_i] = 0 \tag{10}$$

We can see this as follows:

$$\begin{aligned}
\text{Cov}(X_i, \varepsilon_i) &= \mathbb{E}\left[[X_i - \mathbb{E}[X_i]][\varepsilon_i - \mathbb{E}[\varepsilon_i]]\right] && \text{\# definition of covariance} \quad (11)\\
&= \mathbb{E}\left[X_i - \mathbb{E}[X_i]\right]\varepsilon_i && \text{\# since } \mathbb{E}[\varepsilon_i] = 0 \text{ by A2} \quad (12)\\
&= \mathbb{E}[X_i\varepsilon_i - \mathbb{E}[X_i]\varepsilon_i] && \text{\# multiply through} \quad (13)\\
&= \mathbb{E}[X_i\varepsilon_i] - \mathbb{E}[X_i]\,\mathbb{E}[\varepsilon_i] && \text{\# since } \mathbb{E}[X_i] \text{ is a constant} \quad (14)\\
&= \mathbb{E}[X_i\varepsilon_i] && \text{\# since } \mathbb{E}[\varepsilon_i] = 0 \text{ by A2} \quad (15)\\
&= \mathbb{E}[X_i]E[\varepsilon_i] && \text{\# } \mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y] \quad (16)\\
&= 0 && \text{\# since } \mathbb{E}[\varepsilon_i] = 0 \text{ by A2} \quad (17)
\end{aligned}$$

A third implication of A2 is the conditional mean of the population Y_i values corresponding to a given value X_i of the regressor X equals the population regression function (PRF) $f(X_i) = \beta_0 + \beta_1 X_i$. This itself has a key implication, namely, that

$$\mathbb{E}(\varepsilon \mid X) = 0 \Rightarrow \mathbb{E}(Y \mid X) = f(X) = \beta_0 + \beta_1 X \text{ and} \tag{18}$$
$$\mathbb{E}(\varepsilon_i \mid X_i) = 0 \Rightarrow \mathbb{E}(Y_i \mid X_i) = f(X_i) = \beta_0 + \beta_1 X_i \qquad \forall i \tag{19}$$

The proof of this is fairly simple. Recall that by assumption $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ for $i = 1 \dots N$. Then

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i && \text{\# Linear assumption, A1} && (20)\\
\mathbb{E}[Y_i|X_i] &= \mathbb{E}[\beta_0 + \beta_1 X_i + \varepsilon_i|X_i] && \text{\# Expectation conditioned on } X_i && (21)\\
&= \mathbb{E}[\beta_0 + \beta_1 X_i|X_i] + \mathbb{E}[\varepsilon_i|X_i] && \text{\# } \mathbb{E}[X+Y|Z] = \mathbb{E}[X|Z] + \mathbb{E}[Y|Z] && (22)\\
&= \mathbb{E}[\beta_0 + \beta_1 X_i|X_i] && \text{\# } E[\varepsilon_i|X_i] = 0 \text{ by A2} && (23)\\
&= \mathbb{E}[\beta_0|X_i] + \mathbb{E}[\beta_1 X_i|X_i] && \text{\# } \mathbb{E}[X+Y|Z] = \mathbb{E}[X|Z] + \mathbb{E}[Y|Z] && (24)\\
&= \beta_0 + \mathbb{E}[\beta_1 X_i|X_i] && \text{\# } \mathbb{E}[\beta_0|X_i] = \beta_0 \text{ for constant } \beta_0 && (25)\\
&= \beta_0 + \mathbb{E}[\beta_1|X_i]\,\mathbb{E}[X_i|X_i] && \text{\# } \mathbb{E}[XY|Z] = \mathbb{E}[X|Z]\cdot\mathbb{E}[Y|Z] && (26)\\
&= \beta_0 + \beta_1\,\mathbb{E}[X_i|X_i] && \text{\# } \mathbb{E}[\beta_1|X_i] = \beta_1 \text{ for constant } \beta_1 && (27)\\
\mathbb{E}[Y_i|X_i] &= \beta_0 + \beta_1 X_i && \text{\# } \mathbb{E}[X_i|X_i] = X_i && (28)
\end{aligned}
$$

Note that in Equation 26 X and Y (β_1 and X_i) are assumed to be independent.

An amazing result really. But what we really want to do is estimate the population parameters β_0 and β_1. How do we do that? To get there we will need some more machinery.