Classical Linear Regression (CLR) model

The $n$ observations $Y_i, X_{i1}, \ldots, X_{ik}, i = 1, \ldots, n$, on a dependent variable $Y$ and $K$ independent variables $X_1, \ldots, X_K$ satisfy

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{ik} + u_i$$

for $i = 1, \ldots, n$.

Assumption 1: $u_i, i = 1, \ldots, n$ are random variables with $E(u_i) = 0$

Assumption 2: $X_{ik}, i = 1, \ldots, n, k = 1, \ldots, K$ are deterministic, i.e. non-random, constants.

Assumption 3: (Homoskedasticity)
All $u_i$'s have the same variance, i.e. for $i = 1, \ldots, n$

$$Var(u_i) = E(u_i^2) = \sigma^2$$

Assumption 4 (No serial correlation)
The random errors $u_i$ and $u_j$ are not correlated for all $i \neq j = 1, \ldots, n$

$$Cov(u_i, u_j) = E(u_i u_j) = 0$$

For the CLR model with normal errors we make the additional assumption:

Assumption 5. The random error terms $u_i, i = 1, \ldots, n$ are random variables with a normal distribution.
Ordinary Least Squares (OLS) estimators in the simple CLR model and their sampling variance

\[
\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}
\]

(4)

\[
\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}
\]

(5)

\[
Var(\hat{\alpha}) = \frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} (X_i - \bar{X})^2} \sigma^2
\]

(6)

\[
Var(\hat{\beta}) = \frac{n \sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}
\]

(7)

\[
Cov(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sigma^2
\]

(8)

Unbiased estimator of \(\sigma^2\)

\[
s^2 = \frac{1}{n - K} \sum_{i=1}^{n} e_i^2
\]

(9)

Substitution in the sampling variances gives the estimated sampling variances. The square root of these are the standard errors \(std(\hat{\alpha}), std(\hat{\beta})\) or in the multiple CLR model \(std(\hat{\beta}_k), k = 1, \ldots, K\).

Goodness-of-fit

Predicted/computed value:

\[
\hat{Y}_i = \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_K X_{ik}
\]

(10)

OLS residual:

\[
e_i = Y_i - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} - \cdots - \hat{\beta}_K X_{ik}
\]

(11)

Properties of OLS residuals:

\[
\sum_{i=1}^{n} X_{ik} e_i = 0, \quad k = 1, \ldots, K \quad \text{(with a constant term this implies sum of OLS residuals is 0)}
\]

(12)

Decomposition of total variation:

\[
\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2
\]

(13)

Note: \(\bar{Y} = \bar{\hat{Y}}\)

Total Sum of Squares (TSS) \quad Regression Sum of Squares (ESS) \quad Error Sum of Squares (RSS)

\[
R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}
\]

(14)
Confidence interval for $\beta_k$ in CLR model with normal errors

(15) \[ \hat{\beta}_k - t_{1-\alpha, n-K} \text{std}(\hat{\beta}_k) < \beta_k < \hat{\beta}_k + t_{1-\alpha, n-K} \text{std}(\hat{\beta}_k) \]

with \( \Pr\left(T > t_{1-\alpha, n-K}\right) = \frac{1-\alpha}{2} \) and \( T \) has a \( t \) distribution with \( n-K \) degrees of freedom. \( \alpha \) is the confidence coefficient or coverage probability of the confidence interval.

Test on individual regression coefficients in CLR model with normal errors

We test

\[ H_0 : \beta_k = \beta_{k0} \]

against

\[ H_0 : \beta_k \neq \beta_{k0} \]

The test statistic (\( t \) test) is

(16) \[ T = \frac{\hat{\beta}_k - \beta_{k0}}{\text{std}(\hat{\beta}_k)} \]

We reject \( H_0 \) if \( |T| > t_{\alpha, n-K} \) with \( \Pr\left(T > t_{\alpha, n-K}\right) = \alpha \) where \( T \) has a \( t \) distribution with \( n-K \) degrees of freedom and \( \alpha \) is the significance level (and Type I error) of the test.

Test on several regression coefficients in CLR model with normal errors

We test

\[ H_0 : m \text{ regression coefficients are 0} \]

against

\[ H_1 : \text{one of them is not equal to 0} \]

The \( F \) statistic is:

(17) \[ F = \frac{ESS_0 - ESS_1}{m} \]

\[ \frac{ESS_1}{n-K} \]

with \( ESS_0, ESS_1 \) the error sum of squares of the regression model if \( H_0 \) is true and \( H_1 \) is true, respectively. We reject \( H_0 \) if \( F > f_{\alpha} \) with \( \Pr(F > f_{\alpha}) = \alpha \) where \( F \) has a \( F \) distribution with \( m \) (numerator) and \( n-K \) (denominator) degrees of freedom and \( \alpha \) is the significance level (and Type I error) of the test.
Prediction

Predictor

(18) \[ \hat{Y}_{n+1} = \hat{\alpha} + \hat{\beta} X_{n+1} \]

Prediction error

(19) \[ Y_{n+1} - \hat{Y}_{n+1} = (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) X_{n+1} + u_{n+1} \]

Prediction interval

(20) \[ \hat{Y}_{n+1} - t_{1-\alpha} s_{n+1} < Y_{n+1} < \hat{Y}_{n+1} + t_{1-\alpha} s_{n+1} \]

with \( \Pr \left( T > t_{1-\alpha} \right) = \frac{1-\alpha}{2} \) and

(21) \[ s_{n+1}^2 = s^2 \left( 1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \]