The Classical Linear Regression Model

ME104: Linear Regression Analysis
Kenneth Benoit

August 14, 2012

CLRM: Basic Assumptions

1. Specification:
   - Relationship between X and Y *in the population* is linear: $\mathrm{E}(Y) = X\beta$
   - No extraneous variables in X
   - No omitted independent variables
   - Parameters (β) are *constant*
2. $\mathrm{E}(\epsilon) = 0$
3. Error terms:
   - $\mathrm{Var}(\epsilon) = \sigma^2$, or homoskedastic errors
   - $\mathrm{E}(r_{\epsilon_i,\epsilon_j}) = 0$, or no auto-correlation

CLRM: Basic Assumptions (cont.)

4. X is non-stochastic, meaning observations on independent variables are fixed in repeated samples
   - implies no *measurement error* in X
   - implies no serial correlation where a lagged value of Y would be used an independent variable
   - no *simultaneity* or *endogenous* X variables

5. $N > k$, or number of observations is greater than number of independent variables (in matrix terms: $\text{rank}(X) = k$), and no exact linear relationships exist in X

6. Normally distributed errors: $\epsilon | X \sim N(0, \sigma^2)$. Technically however this is a *convenience* rather than a strict assumption

Normally distributed errors

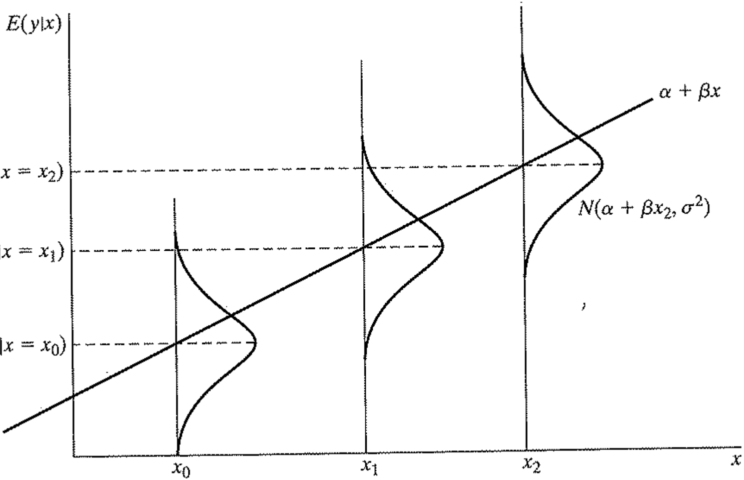


FIGURE 2.2 The Classical Regression Model.

Ordinary Least Squares (OLS)

- Objective: minimize $\sum e_i^2 = \sum(Y_i - \hat{Y}_i)^2$, where
  - $\hat{Y}_i = b_0 + b_1 X_i$
  - error $e_i = (Y_i - \hat{Y}_i)$

$$
\begin{aligned}
b &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})} \\
&= \frac{\sum X_i Y_i}{\sum X_i^2}
\end{aligned}
$$

- The intercept is: $b_0 = \bar{Y} - b_1 \bar{X}$

OLS rationale

- Formulas are very simple
- Closely related to ANOVA (sums of squares decomposition)
- Predicted Y is sample mean when $\Pr(Y|X) = \Pr(Y)$
  - In the special case where Y has no relation to X, $b_1 = 0$, then OLS fit is simply $\hat{Y} = b_0$
  - Why? Because $b_0 = \bar{Y} - b_1\bar{X}$, so $\hat{Y} = \bar{Y}$
  - Prediction is then sample mean when X is unrelated to Y
- Since OLS is then an extension of the sample mean, it has the same attractice properties (efficiency and lack of bias)
- Alternatives exist but OLS has generally the best properties when assumptions are met

OLS in matrix notation

- Formula for coefficient β:

$$\begin{aligned} Y &= X\beta + \epsilon \\ X'Y &= X'X\beta + X'\epsilon \\ X'Y &= X'X\beta + 0 \\ (X'X)^{-1}X'Y &= \beta + 0 \\ \beta &= (X'X)^{-1}X'Y \end{aligned}$$

- Formula for variance-covariance matrix: $\sigma^2(X'X)^{-1}$
  - In simple case where $y = \beta_0 + \beta_1 * x$, this gives $\sigma^2/\sum(x_i - \bar{x})^2$ for the variance of β_1
  - Note how increasing the variation in X will reduce the variance of β_1

The "hat" matrix

- The hat matrix H is defined as:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \\ X\hat{\beta} &= X(X'X)^{-1}X'y \\ \hat{y} &= Hy \end{aligned}$$

- $H = X(X'X)^{-1}X'$ is called the *hat-matrix*
- Other important quantities, such as $\hat{y}$, $\sum e_i^2$ (RSS) can be expressed as functions of H
- Corrections for heteroskedastic errors ("robust" standard errors) involve manipulating H

Three critical quantities

Y_i The observed value of dep. variable for unit i

$\bar{Y}$ The mean of the dep. variable Y

$\hat{Y}_i$ The value of outcome for unit i that is predicted from the model

Sums of squares (ANOVA)

TSS Total sum of squares $\sum(Y_i - \bar{Y})^2$

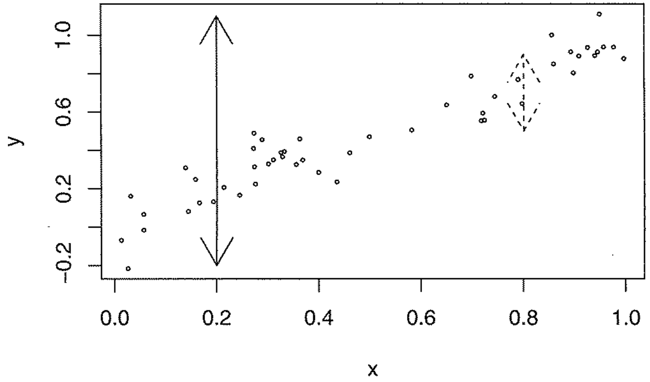
SSM Model or Regression sum of squares $\sum(\hat{Y}_i - \bar{Y})^2$

SSE Error or Residual sum of squares
$\sum e_i^2 = \sum(\hat{Y}_i - Y_i)^2$

The key to remember is that TSS = SSM + SSE

R^2

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$



- Solid arrow: variation in y when X is unknown (TSS Total Sum of Squares $\sum(y_i - \bar{y})^2$)
- Dashed arrow: variation in y when X is known (SSM Model Sum of Squares $\sum(\hat{y}_i - \bar{y})^2$)

R^2 decomposed

$$
\begin{aligned}
y &= \hat{y} + \epsilon \\
\text{Var}(y) &= \text{Var}(\hat{y}) + \text{Var}(\epsilon) + 2\text{Cov}(\hat{y}, \epsilon) \\
\text{Var}(y) &= \text{Var}(\hat{y}) + \text{Var}(\epsilon) + 0 \\
\sum (y_i - \bar{y})^2 / N &= \sum (\hat{y}_i - \bar{\hat{y}})^2 / N + \sum (e_i - \hat{e})^2 / N \\
\sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{\hat{y}})^2 + \sum (e_i - \hat{e})^2 \\
\sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{\hat{y}})^2 + \sum e_i^2 \\
TSS &= SSM + SSE \\
\frac{TSS}{TSS} &= \frac{SSM}{TSS} + \frac{SSE}{TSS} \\
1 &= R^2 + \text{unexplained variance}
\end{aligned}
$$

R^2

- A much over-used statistic: it may not be what we are interested in at all
- Interpretation: the proportion of the variation in y that is explained linearly by the independent variables

$$
\begin{aligned}
R^2 &= \frac{SSM}{TSS} \\
&= 1 - \frac{SSE}{TSS} \\
&= 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}
\end{aligned}
$$

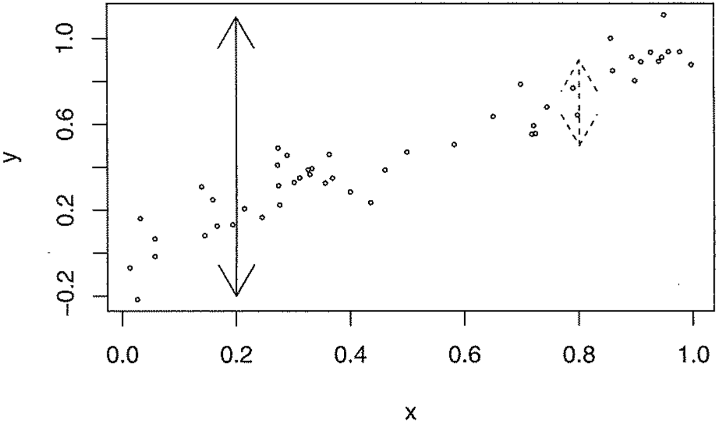
- Alternatively, R^2 is the squared correlation coefficient between y and $\hat{y}$

R^2 continued

- When a model has no intercept, it is possible for R^2 to lie outside the interval $(0, 1)$
- R^2 rises with the addition of more explanatory variables. For this reason we often report "adjusted R^2": $1 - (1 - R^2)\frac{n-1}{n-k-1}$ where k is the total number of regressors in the linear model (excluding the constant)
- Whether R^2 is *high* or not depends a lot on the overall variance in Y
- To R^2 values from different Y samples *cannot be compared*

R^2 continued

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$



- Solid arrow: variation in y when X is unknown (SSR)
- Dashed arrow: variation in y when X is known (SST)

R^2 decomposed

$$
\begin{aligned}
y &= \hat{y} + \epsilon \\
\text{Var}(y) &= \text{Var}(\hat{y}) + \text{Var}(e) + 2\text{Cov}(\hat{y}, e) \\
\text{Var}(y) &= \text{Var}(\hat{y}) + \text{Var}(e) + 0 \\
\sum (y_i - \bar{y})^2/N &= \sum (\hat{y}_i - \bar{\hat{y}})^2/N + \sum (e_i - \hat{e})^2/N \\
\sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{\hat{y}})^2 + \sum (e_i - \hat{e})^2 \\
\sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{\hat{y}})^2 + \sum e_i^2 \\
SST &= SSR + SSE \\
SST/SST &= SSR/SST + SSE/SST \\
1 &= R^2 + \text{unexplained variance}
\end{aligned}
$$

Regression "terminology"

- y is the dependent variable
  - referred to also (by Greene) as a *regressand*
- X are the independent variables
  - also known as explanatory variables
  - also known as regressors
- y is regressed on X
- The error term ϵ is sometimes called a disturbance

Some important OLS properties to understand

Applies to $y = \alpha + \beta x + \epsilon$

- If $\beta = 0$ and the only regressor is the intercept, then this is the same as regressing y on a column of ones, and hence $\alpha = \bar{y}$ — the mean of the observations
- If $\alpha = 0$ so that there is no intercept and one explanatory variable x, then $\beta = \frac{\sum xy}{\sum x^2}$
- If there is an intercept and one explanatory variable, then

$$\begin{aligned} \beta &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ &= \frac{\sum_i (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \end{aligned}$$

Some important OLS properties (cont.)

- If the observations are expressed as deviations from their means, $y* = y - \bar{y}$ and $x^* = x - \bar{x}$, then $\beta = \sum x^* y^* / \sum x^{*2}$
- The intercept can be estimated as $\bar{y} - \beta\bar{x}$. This implies that the intercept is estimated by the value that causes the sum of the OLS residuals to equal zero.
- The mean of the $\hat{y}$ values equals the mean y values – together with previous properties, implies that the OLS regression line passes through the overall mean of the data points

OLS in Stata

```
. use dail2002
(Ireland 2002 Dail Election - Candidate Spending Data)

. gen spendXinc = spend_total * incumb
(2 missing values generated)

. reg votes1st spend_total incumb minister spendXinc

      Source |       SS       df       MS              Number of obs =     462
-------------+------------------------------           F(  4,   457) =  229.05
       Model |  2.9549e+09     4   738728297           Prob > F      =  0.0000
    Residual |  1.4739e+09   457  3225201.58           R-squared     =  0.6672
-------------+------------------------------           Adj R-squared =  0.6643
       Total |  4.4288e+09   461  9607007.17           Root MSE      =  1795.9

------------------------------------------------------------------------------
    votes1st |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 spend_total |   .2033637   .0114807    17.71   0.000     .1808021    .2259252
      incumb |   5150.758   536.3686     9.60   0.000     4096.704    6204.813
    minister |   1260.001   474.9661     2.65   0.008      326.613     2193.39
   spendXinc |  -.1490399   .0274584    -5.43   0.000    -.2030003   -.0950794
       _cons |   469.3744   161.5464     2.91   0.004     151.9086    786.8402
------------------------------------------------------------------------------
```

OLS in R

```
> dail <- read.dta("dail2002.dta")
> mdl <- lm(votes1st ~ spend_total*incumb + minister, data=dail)
> summary(mdl)

Call:
lm(formula = votes1st ~ spend_total * incumb + minister, data = dail)

Residuals:
    Min      1Q  Median      3Q     Max
-5555.8  -979.2  -262.4   877.2  6816.5

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         469.37438  161.54635   2.906  0.00384 **
spend_total           0.20336    0.01148  17.713  < 2e-16 ***
incumb             5150.75818  536.36856   9.603  < 2e-16 ***
minister           1260.00137  474.96610   2.653  0.00826 **
spend_total:incumb   -0.14904    0.02746  -5.428 9.28e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1796 on 457 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.6672, Adjusted R-squared: 0.6643
F-statistic:   229 on 4 and 457 DF,  p-value: < 2.2e-16
```

Examining the sums of squares

```
> yhat <- mdl$fitted.values   # uses the lm object mdl from previous
> ybar <- mean(mdl$model[,1])
> y <- mdl$model[,1]          # can't use dail$votes1st since diff N
> SST <- sum((y-ybar)^2)
> SSR <- sum((yhat-ybar)^2)
> SSE <- sum((yhat-y)^2)
> SSE
[1] 1473917120
> sum(mdl$residuals^2)
[1] 1473917120
> (r2 <- SSR/SST)
[1] 0.6671995
> (adjr2 <- (1 - (1-r2)*(462-1)/(462-4-1)))
[1] 0.6642865
> summary(mdl)$r.squared      # note the call to summary()
[1] 0.6671995
> SSE/457
[1] 3225202
> sqrt(SSE/457)
[1] 1795.885
> summary(mdl)$sigma
[1] 1795.885
```