

Univariate Regression

(Part 2b – Estimation and Measures of Fit)

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w

Brooklyn College

The OLS Estimator, Predicted Values, and Residuals

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2} \quad (4.7)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (4.8)$$

The OLS predicted values \hat{Y}_i and residuals \hat{u}_i are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \dots, n \quad (4.9)$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \dots, n. \quad (4.10)$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual (\hat{u}_i) are computed from a sample of n observations of X_i and $Y_i, i = 1, \dots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

Example

For the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ find the OLS estimators $\hat{\beta}_0, \hat{\beta}_1$ given the following data sample:

Y_i	X_i
4	1
5	4
7	5
12	6

i	Y	X	$y = Y - Ybar$	$x = X - Xbar$	xy	x^2	$Y_hat = b0 + b1 * X$	u_hat
1	4	1	-3	-3	9	9	2.929	1.071
2	5	4	-2	0	0	0	7.000	-2.000
3	7	5	0	1	0	1	8.357	-1.357
4	12	6	5	2	10	4	9.714	2.286
Sum	28	16	0	0	19	14	28.000	0
Avg (Ybar Xbar)	7	4						
	My Calcs	Excel	Diff.					
n	4	4	0					
$b1 = xy/x^2 = 19/14$	1.357	1.357	0					
$b1 = covar(X,Y) / var(X)$	1.357	1.357	0					
$b0 = Ybar - b1*Xbar$	1.571	1.571	0					

Intuition: Regression Slope Coefficient vs Correlation Coefficient

Recall that the correlation coefficient is:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where $\sigma_X = \sqrt{\text{Var}(X)}$ and $\sigma_Y = \sqrt{\text{Var}(Y)}$

We've established that the slope coefficient for univariate regression is:

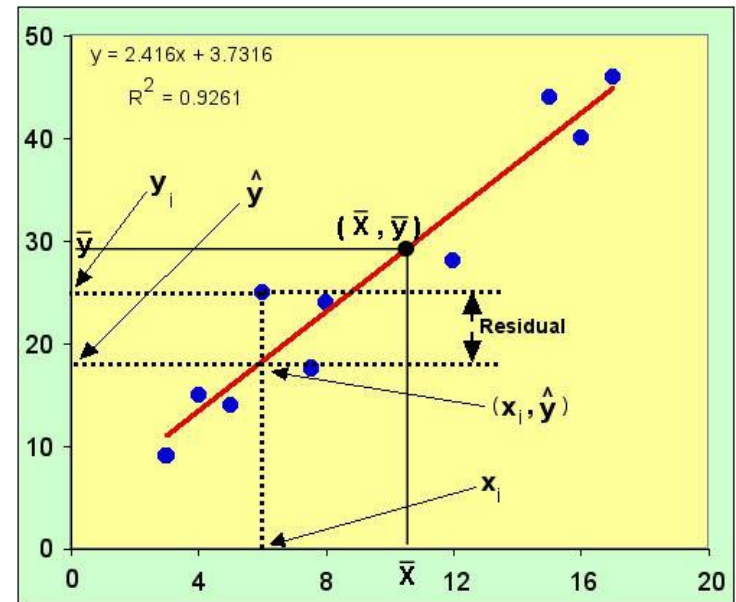
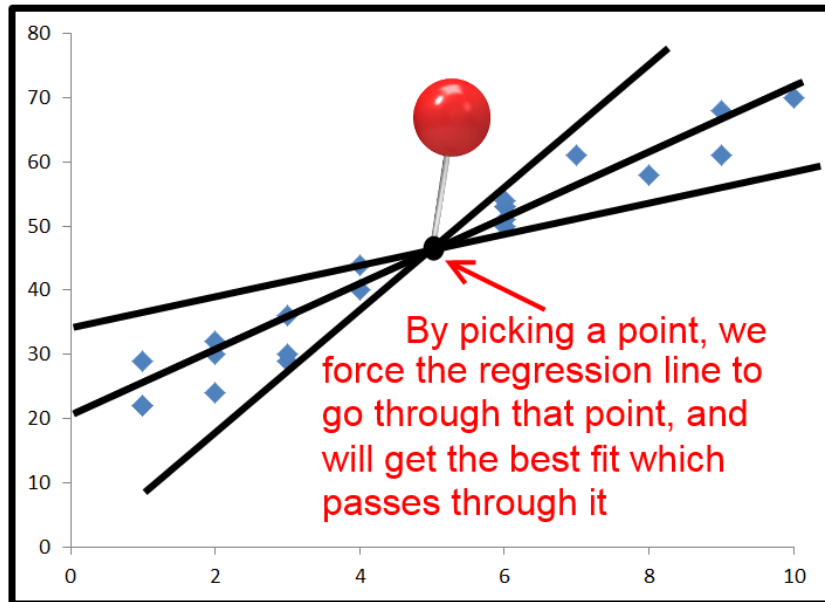
$$\hat{\beta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \equiv \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_X}$$

Which implies the following relationship between the regression slope coefficient and the correlation coefficient:

$$\hat{\beta}_1 = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}$$

Intuition: Regression line must pass through point (\bar{X}, \bar{Y})

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$



Measures of Fit (SW Section 4.3)

Two regression statistics provide complementary measures of how well the regression line “fits” or explains the data:

- The *regression R^2* (aka "coefficient of determination") measures the fraction of the variance of Y that is explained by X ; it is unitless and ranges between zero (no fit) and one (perfect fit)
- The *standard error of the regression (SER)* measures the magnitude of a typical regression residual in the units of Y .

The *regression* R^2 is the fraction of the sample variance of Y_i “explained” by the regression.

$Y_i = \hat{Y}_i + \hat{u}_i = \text{OLS prediction} + \text{OLS residual}$

→ sample var (Y) = sample var(\hat{Y}_i) + sample var(\hat{u}_i)

→ total sum of squares = “explained” SS + “residual” SS

$$\text{TSS} = \text{ESS} + \text{SSR}$$

Definition of R^2 :

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$
$$= 1 - \frac{SSR}{TSS} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

- $R^2 = 0$ means $ESS = 0$
- $R^2 = 1$ means $ESS = TSS$
- $0 \leq R^2 \leq 1$, higher R^2 means better fit
- For regression with a single X , R^2 = the square of the correlation coefficient between X and Y

The Standard Error of the Regression (SER)

The *SER* measures the spread of the distribution of u . The *SER* is (almost) the sample standard deviation of the OLS residuals:

$$\begin{aligned} SER &= \sqrt{\frac{1}{n-2} \sum_{i=1}^n (\hat{u}_i - \bar{\hat{u}})^2} \\ &= \sqrt{\frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2} \end{aligned}$$

The second equality holds because $\bar{\hat{u}} = \frac{1}{n} \sum_{i=1}^n \hat{u}_i = 0$.

$$SER = \sqrt{\frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2}$$

The *SER*:

has the units of u , which are the units of Y

measures the average “size” of the OLS residual (the average “mistake” made by the OLS regression line)

The *root mean squared error* (*RMSE*) is closely related to the *SER*:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2}$$

This measures the same thing as the *SER* – the minor difference is division by $1/n$ instead of $1/(n-2)$.

Example (continued)

For the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, find the coefficient of determination (R^2) given the following data sample:

Y_i	X_i
4	1
5	4
7	5
12	6

[illegible]

Next week – finish chapter 4

1. Probability framework for linear regression
2. The ordinary least squares (OLS) estimator and the sample regression line
3. Measures of fit of the sample regression
- 4. The least squares model assumptions**
- 5. The sampling distribution of the OLS estimator**