# Regression

Sunthud Pornprasertmanit Chulalongkorn University

## Terms

Predictor Variable, Independent Variable

Criterion Variable, Dependent Variable

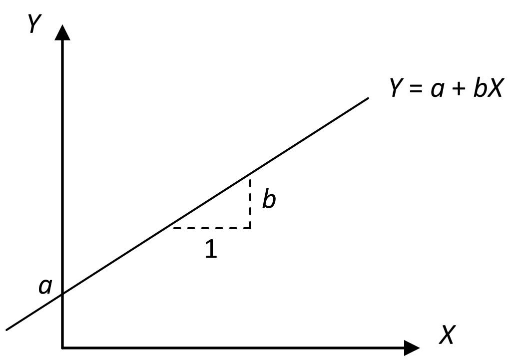
Actual Value, Predicted Value, Error of Prediction (Residual)

$$Y - Y' = e$$

## Linear Equation

Regression Line

$$Y = a + bX$$




Slope (*b*)

Y-intercept (*a*)

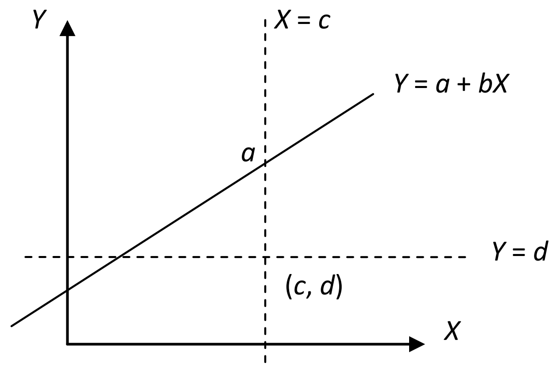
Change axis

$$Y - d = a + b(X - c)$$

Author Note

This article was written in June 2007 for teaching in Introduction to Statistics in Psychology Class, Faculty of Psychology, Chulalongkorn University

Correspondence to Sunthud Pornprasertmanit. Email: psunthud@gmail.com




## Steps of predicting value in criterion variable

### No Predictor

Which value that can predicted all value with the least error?

- Sum of Errors equal to zero
- Least Sum of Squared Errors ($SS_{error}$)

Arithmetic Mean

Example

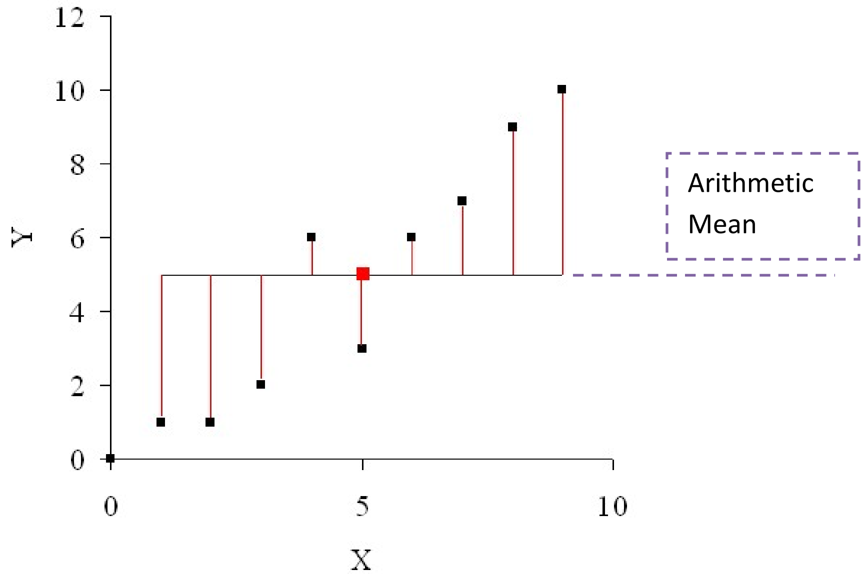
Arithmetic Mean

| ID | NCCU | Baseline Prediction | Error of Prediction | Squared error |
|---|---|---|---|---|
| **1** | 4 | 7 | -3 | 9 |
| **2** | 6 | 7 | -1 | 1 |
| **3** | 6 | 7 | -1 | 1 |
| **4** | 7 | 7 | 0 | 0 |
| **5** | 8 | 7 | +1 | 1 |
| **6** | 7 | 7 | 0 | 0 |
| **7** | 8 | 7 | +1 | 1 |
| **8** | 10 | 7 | +3 | 9 |
| **Total** | 56 | | 0 | 22 |

NCCU = Number of Credit Cards Used

Sum of Errors equal to zero.

$SS_{error} = SS_x = 22$




## One Predictor

Linear Transformation from predictor to predicted value that can predict criterion as much as possible.

$$Y' = a + bX$$

$$Y - Y' = e$$

What are the values of constants *a* and *b* that make $SS_{\text{error}}$ at least? (Supplement 1)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - Y_i')^2$$

The regression line must pass coordinate $(\bar{X}, \bar{Y})$.

$$a = \bar{Y} - b\bar{X}$$

$$b = \frac{S_{XY}}{S_X^2}$$

The more correlation between predictor and criterion, the more accuracy in prediction.

Example

Correlation Matrix

| | *1* | *2* | *3* | *4* |
|---|---|---|---|---|
| **1. NCCU** | | | | |
| **2. Family Size** | .87 | | | |
| **3. Family Income** | .83 | .67 | | |
| **4. Number of Automobiles** | .34 | .19 | .30 | |

The best predictor is family size.

Prediction Equation: $Y' = 2.87 + .97X$

Criterion ($Y$)

Predictor ($X$)

Predicted Value ($Y'$)

| ID | NCCU | Family Size | Prediction Score | Error of Prediction | Error Squared |
|---|---|---|---|---|---|
| **1** | 4 | 2 | 4.81 | -.81 | .66 |
| **2** | 6 | 2 | 4.81 | 1.19 | 1.42 |
| **3** | 6 | 4 | 6.75 | -.75 | .56 |
| **4** | 7 | 4 | 6.75 | .25 | .06 |
| **5** | 8 | 5 | 7.72 | .28 | .08 |
| **6** | 7 | 5 | 7.72 | -.72 | .52 |
| **7** | 8 | 6 | 8.69 | -.69 | .48 |
| **8** | 10 | 6 | 8.69 | 1.31 | 1.72 |
| **Total** | 56 | | | 0 | 5.50 |

Sum of Errors equal to zero.

$SS_{error}$ = 5.50

$SS_{error}$ reduces from 22 to 5.50.

For regression analysis, $SS_{error}$ from no predictor is $SS_{total}$.

The difference between $SS_{error}$ from one predictor and $SS_{total}$ is $SS_{regression}$.

$$SS_{\text{total}} = SS_{\text{regression}} + SS_{\text{error}}$$

The proportion of $SS_{regression}$ and $SS_{total}$ is coefficient of determination.

$$r^2 = \frac{SS_{\text{regression}}}{SS_{\text{total}}} = 1 - \frac{SS_{\text{error}}}{SS_{\text{total}}}$$

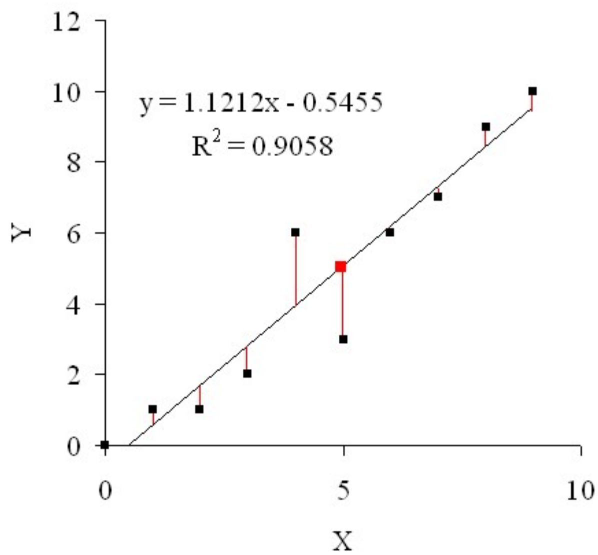
Therefore, in this example, $X$ can explain $Y$ variance equal to 16.5 (75 %).

Standard error of estimate ($S_{Y.X}$) is standard deviation of error of prediction.

$$S_{Y.X} = \sqrt{\frac{SS_{error}}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - Y_i')^2}{n}}$$

$$S_{Y.X} = S_X\sqrt{1 - r_{YX}^2}$$

When predicting value, regression analysis can provide point estimate or interval estimate (See later in confidence interval).



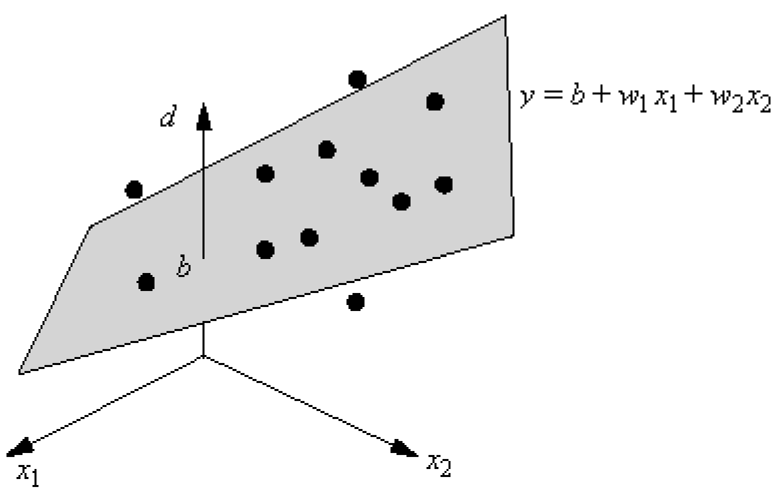

## More than One Predictor

Linear Combination from predictors to predicted value that can predict criterion as much as possible.

$$Y' = a + b_1X_1 + b_2X_2$$

$$Y - Y' = e$$

Regression Plane



What is the constants $a$, $b_1$ and $b_2$ that make $SS_{\text{error}}$ as less as possible?

The regression line must pass coordinate $(\bar{X}_1, \bar{X}_2, \bar{Y})$.

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$b_1 = \left(\frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2}\right)\frac{S_Y}{S_1}$$

$$b_2 = \left(\frac{r_{Y2} - r_{Y1}r_{12}}{1 - r_{12}^2}\right)\frac{S_Y}{S_2}$$

What do $a$, $b_1$ and $b_2$ mean?

Multicollinearity

Additional predictor should has high correlation with criterion and low correlation with other predictors, because it explain $SS_{\text{error}}$.

Partial Correlation

$$pr_{XY(Z)} = \frac{r_{XY} - r_{YZ}r_{XZ}}{\sqrt{1 - r_{YZ}^2}\sqrt{1 - r_{XZ}^2}}$$

Example

| | *1* | *2* | *3* | *4* |
|---|---|---|---|---|
| **1. NCCU** | | | | |
| **2. Family Size** | .87 | | | |
| **3. Family Income** | .83 | .67 | | |
| **4. Number of Automobiles** | .34 | .19 | .30 | |

$$pr_{31(2)} = \frac{r_{31} - r_{32}r_{12}}{\sqrt{1 - r_{32}^2}\sqrt{1 - r_{12}^2}} = \frac{.83 - (.67)(.87)}{\sqrt{1 - (.67)^2}\sqrt{1 - (.87)^2}} = .68$$

$$pr_{41(2)} = \frac{r_{41} - r_{42}r_{12}}{\sqrt{1 - r_{42}^2}\sqrt{1 - r_{12}^2}} = \frac{.34 - (.19)(.87)}{\sqrt{1 - (.19)^2}\sqrt{1 - (.87)^2}} = .36$$

The best additional predictor is Family Income.

Prediction Equation: $Y' = .482 + .63X_1 + .216X_2$

Predictor ($X_1$ and $X_2$)

Criterion ($Y$)

Predicted Value ($Y'$)

| ID | NCCU | Family Size | Family Income | Prediction Score | Error of Prediction | Error Squared |
|---|---|---|---|---|---|---|
| **1** | 4 | 2 | 14 | 4.76 | -.76 | .58 |
| **2** | 6 | 2 | 16 | 5.20 | .80 | .64 |
| **3** | 6 | 4 | 14 | 6.03 | -.03 | .00 |
| **4** | 7 | 4 | 17 | 6.68 | .32 | .10 |
| **5** | 8 | 5 | 18 | 7.53 | .47 | .22 |
| **6** | 7 | 5 | 21 | 8.18 | -1.18 | 1.39 |
| **7** | 8 | 6 | 17 | 7.95 | .05 | .00 |
| **8** | 10 | 6 | 25 | 9.67 | .33 | .11 |
| **Total** | 56 | | | | 0 | 3.04 |

Sum of Errors equal to zero.

$SS_{error}$ = 3.04

$SS_{error}$ reduces from 5.50 to 3.04

Coefficient of Multiple Determination

$SS_{regression}$ increase from 16.50 (75 %) to 18.96 (86 %)

$$R^2_{Y.12} = \frac{r^2_{Y1} + r^2_{Y2} - 2r_{Y1}r_{Y2}r_{12}}{1 - r^2_{12}}$$

The addition predictor increase explaining variance equal to 11 %.

## General Formula of Multiple Regression

$$Y' = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

What do $a$, $b_1$, $b_2$, … , $b_n$ mean?

## Recentering

$$a = \bar{Y} - b\bar{X}$$

$$Y' = a + bX$$

$$Y' = (\bar{Y} - b\bar{X}) + bX$$

1

$$Y' = \bar{Y} + b(X - \bar{X})$$

If centering at mean of predictor, what is intercept mean, in equation 1?

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$Y' = a + b_1X_1 + b_2X_2$$

$$Y' = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 + b_1X_1 + b_2X_2$$

$$Y' = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 + b_1(X_1 - \bar{X}_1) + b_2(X_2 - \bar{X}_2) \quad \Leftarrow \boxed{2}$$

$$Y' = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 + b_1(X_1 - \bar{X}_1) + b_2(X_2 - \bar{X}_2) \quad \Leftarrow \boxed{3}$$

$$Y' = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 + b_1(X_1 - \bar{X}_1) + b_2(X_2 - \bar{X}_2) \quad \Leftarrow \boxed{4}$$

What are intercept means in equation 2, 3, and 4?

## Confidence Interval of Predicted Value

For interval estimate, predicted value and confidence interval are used.

68.3 %: $Y' \pm S_{Y.X}$ or $Y' \pm S_{Y.12}$

95.4 %: $Y' \pm 2S_{Y.X}$ or $Y' \pm 2S_{Y.12}$

## Assumption in Regression Analysis

1) Independent of correlated errors
2) Errors distribute in normal distribution
3) Linearity
4) Homoscadasticity
5) Multicollinearity