EPRS8550

# Variance, Covariance, and Correlation

$$s_x^2 = \frac{\sum(X - \overline{X})^2}{N-1} = \frac{\sum(X - \overline{X})(X - \overline{X})}{N-1}$$

$$s_{xy} = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{N-1}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

i.e.,

$$Correlation = \frac{Covariance}{SD\ of\ x \times SD\ of\ y}$$

Table 1. Mean, Variance, Covariance and Correlation

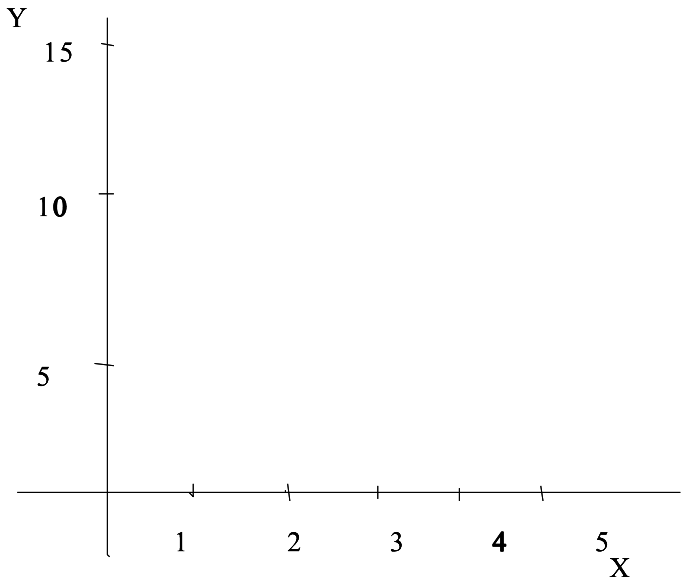
| OBS | $X$ | $X - \overline{X}$ | $(X - \overline{X})^2$ | $Y$ | $Y - \overline{Y}$ | $(Y - \overline{Y})^2$ | $(X - \overline{X})(Y - \overline{Y})$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | -2 | 4 | 3 | -4.3 | 18.49 | 8.6 |
| 2 | 1 | -2 | 4 | 5 | -2.3 | 5.29 | 4.6 |
| 3 | 1 | -2 | 4 | 6 | -1.3 | 1.69 | 2.6 |
| 4 | 1 | -2 | 4 | 9 | 1.7 | 2.89 | -3.4 |
| 5 | 2 | -1 | 1 | 4 | -3.3 | 10.89 | 3.3 |
| 6 | 2 | -1 | 1 | 6 | -1.3 | 1.69 | 1.3 |
| 7 | 2 | -1 | 1 | 7 | -0.3 | 0.09 | 0.3 |
| 8 | 2 | -1 | 1 | 10 | 2.7 | 7.29 | -2.7 |
| 9 | 3 | 0 | 0 | 4 | -3.3 | 10.89 | 0 |
| 0 | 3 | 0 | 0 | 6 | -1.3 | 1.69 | 0 |
| 11 | 3 | 0 | 0 | 8 | 0.7 | 0.49 | 0 |
| 12 | 3 | 0 | 0 | 10 | 2.7 | 7.29 | 0 |
| 13 | 4 | 1 | 1 | 5 | -2.3 | 5.29 | -2.3 |
| 14 | 4 | 1 | 1 | 7 | -0.3 | 0.09 | -0.3 |
| 15 | 4 | 1 | 1 | 9 | 1.7 | 2.89 | 1.7 |
| 16 | 4 | 1 | 1 | 12 | 4.7 | 22.09 | 4.7 |
| 17 | 5 | 2 | 4 | 6 | -1.3 | 1.69 | -2.6 |
| 18 | 5 | 2 | 4 | 7 | -0.3 | 0.09 | -0.6 |
| 19 | 5 | 2 | 4 | 10 | 2.7 | 7.29 | 5.4 |
| 20 | 5 | 2 | 4 | 12 | 4.7 | 22.09 | 9.4 |
| | | | | | | | |
| sum | 60 | 0 | 40 | 146 | 0 | 130.2 | 30 |

Calculate

Mean of X ( $\overline{X}$ )
Mean of Y ( $\overline{Y}$ )
Variance of X ( $s_x^2$ )
Standard Deviation of X ( $s_x$ )
Variance of Y ( $s_y^2$ )
Standard Deviation of Y ( $s_y$ )

Covariance of X and Y ( $S_{xy}$ )
Correlation of X and Y ( $r_{xy}$ )

Draw a scatter plot of X and Y




Draw a "best-fitting" line on the scatter plot above.

$$Y' = a + bX \quad (\text{or } \hat{Y} = b_0 + b_1 X)$$

Y' = predicted score
X = predictor
a = intercept
b = slope

$$Y' = 5.05 + .75X$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{30}{40} = .75, \quad (or \ b = r\frac{s_y}{s_x})$$

$$a = \overline{Y} - b\,\overline{X} = 7.3 - (.75)(3.0) = 5.05$$

# Table 2. Regression Analysis

| | | | | Total | | Regression | | Residual | |
|---|---|---|---|---|---|---|---|---|---|
| OBS | X | Y | Y' | $Y - \bar{Y}$ | $(Y - \bar{Y})^2$ | $Y' - \bar{Y}$ | $(Y' - \bar{Y})^2$ | $Y - Y'$ | $(Y - Y')^2$ |
| 1 | 1 | 3 | 5.8 | -4.3 | 18.49 | -1.5 | 2.25 | -2.8 | 7.84 |
| 2 | 1 | 5 | 5.8 | -2.3 | 5.29 | -1.5 | 2.25 | -0.8 | 0.64 |
| 3 | 1 | 6 | 5.8 | -1.3 | 1.69 | -1.5 | 2.25 | 0.2 | 0.04 |
| 4 | 1 | 9 | 5.8 | 1.7 | 2.89 | -1.5 | 2.25 | 3.2 | 10.24 |
| 5 | 2 | 4 | 6.55 | -3.3 | 10.89 | -0.75 | 0.5625 | -2.55 | 6.5025 |
| 6 | 2 | 6 | 6.55 | -1.3 | 1.69 | -0.75 | 0.5625 | -0.55 | 0.3025 |
| 7 | 2 | 7 | 6.55 | -0.3 | 0.09 | -0.75 | 0.5625 | 0.45 | 0.2025 |
| 8 | 2 | 10 | 6.55 | 2.7 | 7.29 | -0.75 | 0.5625 | 3.45 | 11.9025 |
| 9 | 3 | 4 | 7.3 | -3.3 | 10.89 | 0 | 0 | -3.3 | 10.89 |
| 0 | 3 | 6 | 7.3 | -1.3 | 1.69 | 0 | 0 | -1.3 | 1.69 |
| 11 | 3 | 8 | 7.3 | 0.7 | 0.49 | 0 | 0 | 0.7 | 0.49 |
| 12 | 3 | 10 | 7.3 | 2.7 | 7.29 | 0 | 0 | 2.7 | 7.29 |
| 13 | 4 | 5 | 8.05 | -2.3 | 5.29 | 0.75 | 0.5625 | -3.05 | 9.3025 |
| 14 | 4 | 7 | 8.05 | -0.3 | 0.09 | 0.75 | 0.5625 | -1.05 | 1.1025 |
| 15 | 4 | 9 | 8.05 | 1.7 | 2.89 | 0.75 | 0.5625 | 0.95 | 0.9025 |
| 16 | 4 | 12 | 8.05 | 4.7 | 22.09 | 0.75 | 0.5625 | 3.95 | 15.6025 |
| 17 | 5 | 6 | 8.8 | -1.3 | 1.69 | 1.5 | 2.25 | -2.8 | 7.84 |
| 18 | 5 | 7 | 8.8 | -0.3 | 0.09 | 1.5 | 2.25 | -1.8 | 3.24 |
| 19 | 5 | 10 | 8.8 | 2.7 | 7.29 | 1.5 | 2.25 | 1.2 | 1.44 |
| 20 | 5 | 12 | 8.8 | 4.7 | 22.09 | 1.5 | 2.25 | 3.2 | 10.24 |
| | | | | | | | | | |
| sum | 60 | 146 | 146 | 0 | 130.2 | 0 | 22.5 | 0 | 107.7 |

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

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

In English, each score Y = mean Y + regression + error

Similarly, $$\sum (Y - \bar{Y})^2 = \sum (Y' - \bar{Y})^2 + \sum (Y - Y')^2$$

$$SS_{tot} = SS_{reg} + SS_{res}$$

## Test of Significance

## ANOVA Table

| Source | SS | df | MS | F |
|---|---|---|---|---|
| regression | $\Sigma(Y' - \bar{Y})^2$ | k | $SS_{reg}/k$ | $MS_{reg}/MS_{res}$ |
| residual | $\Sigma(Y - Y')^2$ | N-k-1 | $SS_{res}/N-k-1$ | |
| total | $\Sigma(Y - \bar{Y})^2$ | N-1 | $SS_{tot}/N-1$ | |

where N = # of observation, k = # of predictors

$H_0$: X (or X's) does (do) not explain a significant amount of variation in Y.

$F_{obs} = MS_{reg}/MS_{res}$, $F_{crit} = F_{\alpha,\ k,\ N-k-1}$

Decision Rule: Reject $H_0$ if $F_{obs} \geq F_{crit}$

Complete the ANOVA table for our data.

| Source | SS | df | MS | F |
|---|---|---|---|---|
| regression | | | | |
| residual | | | | |
| total | | | | |

# Testing the Regression Coefficient

Mean Square Resiual = $MS_{res} = s^2_{y.x}$

Standard Error of Estimate = $\sqrt{MS_{res}} = s_{y.x}$

Standard Error of b = $\sqrt{\frac{MS_{res}}{\sum x^2}} = \frac{\sqrt{MSres}}{s_x \cdot \sqrt{N-1}} = s_b$

$H_0 : \beta = 0$

Test Statistic = $\frac{statistic - parameter}{SE} = t_{obs} = \frac{b}{s_b}$

$t_{crit} = t_{\alpha/2,\ N-k-1}$

Decision Rule: Reject if $|t_{obs}| \geq t_{crit}$

Test the slope for our data.

(Recall $t^2 = F$. Compare with the F ratio we obtained before.)

Confidence Interval for b

$b \pm t_{\alpha/2,\, df}(s_b)$
$.75 \pm (2.101)(.3868) = -.0627$ to $1.5627$

The slope lies within -.0627 and 1.5627 with 95% confidence. $(-.0627 \leq \beta \leq 1.5627)$

Relationship Between Correlation and Regression

In our data, $r_{xy} = .4157$ and $b = .75$

$$b = r_{xy}\frac{s_y}{s_x} = (.4157)\frac{(2.62)}{(1.45)} = .75$$

1. $r_{xy}$ and slope
2. $r^2_{xy}$ and R-square ($R^2$)

$r^2_{xy} = (.4157)^2 = .1728$ 17.28%
$R^2 = SS_{reg}/SS_{tot} = 22.5/130.2 = .1728$ 17.28%

When k = 1, $r^2_{xy} = R^2$
3.

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} = \frac{.1728/1}{(1-.1728)/18} = 3.760$$

## Factors Affecting the Precision of the Regression

1. Sample Size (N) (Larger N, _____ F)

2. Standard Error of Estimate (Larger $S_{y.x}$, _____ F)

3. Range of X (Larger range of X, _____ F)

## Assumptions

1. X is a fixed variable.

2. X is free from measurement error.

3. Regression of Y on X is linear.

4. Errors are independent of each other.

5. Errors are normally distributed at each given value of X.

6. Variabity of errors at each fixed value of X is the same across all X (homoscedasticity).