The ANOVA identity

Chrysafis Vogiatzis

Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

Lecture 32a

ILLINOIS
ISE | Industrial & Enterprise Systems Engineering
GRAINGER COLLEGE OF ENGINEERING

Quick recap

Simple linear regression:

- Goal: get a line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.
- How? Least squares line.

Is it significant?

- Goal: check whether there is *really* a relationship.
- How? Hypothesis testing.

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

- $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$, and hence

$$T_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}/\sqrt{S_{xx}}}$$

  - where $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$;
  - and $\hat{\sigma} = \sqrt{MS_E} = \sqrt{\frac{SS_E}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$.

ILLINOIS

Quick recap

Simple linear regression:

- Goal: get a line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.
- How? Least squares line.

Is it significant?

- Goal: check whether there is *really* a relationship.
- How? Hypothesis testing.

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

- $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$, and hence

$$T_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}/\sqrt{S_{xx}}}$$

  - where $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$;
  - and $\hat{\sigma} = \sqrt{MS_E} = \sqrt{\frac{SS_E}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$.

Quick recap

Simple linear regression:

- Goal: get a line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.
- How? Least squares line.

Is it significant?

- Goal: check whether there is *really* a relationship.
- How? Hypothesis testing.

$$H_0 : \beta_1 = 0 \ \text{ vs. } \ H_1 : \beta_1 \neq 0.$$

- $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$, and hence

$$T_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}/\sqrt{S_{xx}}}$$

  - where $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$;
  - and $\hat{\sigma} = \sqrt{MS_E} = \sqrt{\frac{SS_E}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$.

ILLINOIS

The ANOVA identity

Sum of squares come in different shapes and forms..

- sum of squares of **errors**:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

- **total** sum of squares:

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

- sum of squares of **regression**:

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 .$$

The ANalysis Of VAriance (ANOVA) identity:

$$\boxed{SS_T = SS_E + SS_R}$$

ILLINOIS

The ANOVA identity

Sum of squares come in different shapes and forms..

- sum of squares of **errors**:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

- **total** sum of squares:

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

- sum of squares of **regression**:

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 .$$

The ANalysis Of VAriance (ANOVA) identity:

$$\boxed{SS_T = SS_E + SS_R}$$

The ANOVA identity

Sum of squares come in different shapes and forms..

- sum of squares of **errors**:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

- **total** sum of squares:

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

- sum of squares of **regression**:

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 .$$

The ANalysis Of VAriance (ANOVA) identity:

$$\boxed{SS_T = SS_E + SS_R}$$

ILLINOIS

The ANOVA identity

Sum of squares come in different shapes and forms..

- sum of squares of **errors**:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

- **total** sum of squares:

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

- sum of squares of **regression**:

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 .$$

The ANalysis Of VAriance (ANOVA) identity:

$$\boxed{SS_T = SS_E + SS_R}$$

ILLINOIS

The ANOVA identity

Sum of squares come in different shapes and forms..

- sum of squares of **errors**:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

- **total** sum of squares:

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

- sum of squares of **regression**:

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 .$$

The ANalysis Of VAriance (ANOVA) identity:

$$\boxed{SS_T = SS_E + SS_R}$$

ILLINOIS

Degrees of freedom

Let us take a moment and check the corresponding degrees of freedom.

- sum of squares of **errors** SS_E:

$$n - 2 \text{ degrees of freedom.}$$

- **total** sum of squares SS_T:

$$n - 1 \text{ degrees of freedom.}$$

- sum of squares of **regression** SS_R:

$$1 \text{ degree of freedom.}$$

The last follows because
$SS_T = SS_R + SS_E \implies df(SS_T) = df(SS_R) + df(SS_E)$.

Degrees of freedom

Let us take a moment and check the corresponding degrees of freedom.

- sum of squares of **errors** SS_E:

$$n - 2 \text{ degrees of freedom.}$$

- **total** sum of squares SS_T:

$$n - 1 \text{ degrees of freedom.}$$

- sum of squares of **regression** SS_R:

$$1 \text{ degree of freedom.}$$

The last follows because
$SS_T = SS_R + SS_E \implies df(SS_T) = df(SS_R) + df(SS_E)$.

Degrees of freedom

Let us take a moment and check the corresponding degrees of freedom.

- sum of squares of **errors** SS_E:

$$n - 2 \text{ degrees of freedom.}$$

- **total** sum of squares SS_T:

$$n - 1 \text{ degrees of freedom.}$$

- sum of squares of **regression** SS_R:

$$1 \text{ degree of freedom.}$$

The last follows because
$SS_T = SS_R + SS_E \implies df(SS_T) = df(SS_R) + df(SS_E)$.

ILLINOIS

Degrees of freedom

Let us take a moment and check the corresponding degrees of freedom.

- sum of squares of **errors** SS_E:

$$n - 2 \text{ degrees of freedom.}$$

- **total** sum of squares SS_T:

$$n - 1 \text{ degrees of freedom.}$$

- sum of squares of **regression** SS_R:

$$1 \text{ degree of freedom.}$$

The last follows because
$SS_T = SS_R + SS_E \implies df(SS_T) = df(SS_R) + df(SS_E)$.

Mean squares

Combining the sum of squares and the degrees of freedom, we may calculate the mean squares:

- $MS_E = \frac{SS_E}{n-2}$.
- $MS_T = \frac{SS_T}{n-1}$.
- $MS_R = \frac{SS_R}{1} = SS_R$.

ILLINOIS

The definition of R^2

Definition

R^2 is a measure of how much of the variability is accounted for by the regression model and is calculated as:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Example

Consider the following data.

x	y	$\hat{y}$	x	y	$\hat{y}$
1	7.6	7.654	6	8.74	9.019
9	10.24	9.838	7	8.99	9.292
2	7.3	7.927	8	9.93	9.565
7	8.97	9.292	1	8.47	7.654

Also note that $\overline{y} = 8.78$ *and* $SS_E = 1.629$. *What is* R^2*?*

Answer: First, for the total sum of squares: $SS_T = \sum(y_i - \overline{y})^2 = 7.2148$. Then:

$$R^2 = 1 - \frac{SS_E}{SS_T} = 1 - \frac{1.629}{7.2148} = 0.774.$$

ILLINOIS

The definition of R^2

Definition

R^2 is a measure of how much of the variability is accounted for by the regression model and is calculated as:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Example

Consider the following data.

x	y	$\hat{y}$	x	y	$\hat{y}$
1	7.6	7.654	6	8.74	9.019
9	10.24	9.838	7	8.99	9.292
2	7.3	7.927	8	9.93	9.565
7	8.97	9.292	1	8.47	7.654

Also note that $\overline{y} = 8.78$ *and* $SS_E = 1.629$. *What is* R^2*?*

Answer: First, for the total sum of squares: $SS_T = \sum (y_i - \overline{y})^2 = 7.2148$. Then:

$$R^2 = 1 - \frac{SS_E}{SS_T} = 1 - \frac{1.629}{7.2148} = 0.774.$$

ILLINOIS

The definition of R^2

Definition

R^2 is a measure of how much of the variability is accounted for by the regression model and is calculated as:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Example

Consider the following data.

x	y	$\hat{y}$	x	y	$\hat{y}$
1	7.6	7.654	6	8.74	9.019
9	10.24	9.838	7	8.99	9.292
2	7.3	7.927	8	9.93	9.565
7	8.97	9.292	1	8.47	7.654

Also note that $\overline{y} = 8.78$ *and* $SS_E = 1.629$. *What is* R^2*?*

Answer: First, for the total sum of squares: $SS_T = \sum (y_i - \overline{y})^2 = 7.2148$. Then:

$$R^2 = 1 - \frac{SS_E}{SS_T} = 1 - \frac{1.629}{7.2148} = 0.774.$$

ILLINOIS

The definition of R^2

Definition

R^2 is a measure of how much of the variability is accounted for by the regression model and is calculated as:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Example

Consider the following data.

x	y	$\hat{y}$	x	y	$\hat{y}$
1	7.6	7.654	6	8.74	9.019
9	10.24	9.838	7	8.99	9.292
2	7.3	7.927	8	9.93	9.565
7	8.97	9.292	1	8.47	7.654

Also note that $\overline{y} = 8.78$ *and* $SS_E = 1.629$. *What is* R^2*?*

Answer: First, for the total sum of squares: $SS_T = \sum (y_i - \overline{y})^2 = 7.2148$. Then:

$$R^2 = 1 - \frac{SS_E}{SS_T} = 1 - \frac{1.629}{7.2148} = 0.774.$$

The definition of R^2

Definition

R^2 is a measure of how much of the variability is accounted for by the regression model and is calculated as:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Example

Consider the following data.

x	y	$\hat{y}$	x	y	$\hat{y}$
1	7.6	7.654	6	8.74	9.019
9	10.24	9.838	7	8.99	9.292
2	7.3	7.927	8	9.93	9.565
7	8.97	9.292	1	8.47	7.654

Also note that $\overline{y} = 8.78$ *and* $SS_E = 1.629$. *What is* R^2*?*

Answer: First, for the total sum of squares: $SS_T = \sum (y_i - \overline{y})^2 = 7.2148$. Then:

$$R^2 = 1 - \frac{SS_E}{SS_T} = 1 - \frac{1.629}{7.2148} = 0.774.$$

ILLINOIS