12 Correlation and Linear Regression (Cont'd)

MTH 3240 Environmental Statistics

Spring 2020

Notes

Objectives

Objectives:

- Interpret sums of squares, degrees of freedom, and mean squares (**Optional for Spring 2020**).
- Carry out a regression model F test for the slope in a regression model (**Optional for Spring 2020**).

Notes

Sums of Squares in Regression (Optional for Spring 2020)

- In **one-factor ANOVA**, we split the **total variation** in Y into:
  1. Variation due to the **factor** (*between-groups*).
  2. Variation due to **random error** (*within-groups*).
- In a **regression analysis**, we can split the **total variation** in Y into:
  1. Variation due to the X **variable**.
  2. Variation due to **random error**.

Notes

(Optional for Spring 2020)

- The **random error** is variation in Y that's due to **all other variables *besides*** X (or besides the *factor* in ANOVA).

Notes

(Optional for Spring 2020)

Example

Lengths of snakes explains **some** of the **variation** in their **weights**, but not all of it.

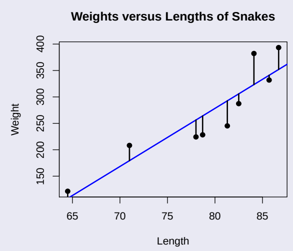
If it explained *all* of the **variation**, the points in the scatterplot would lie *exactly* on a *straight line*.

Other variables that contribute to **variation** in **weights** (e.g. metabolic rate, caloric intake, bone density, etc.) show up as **residuals** (i.e. **random error**) in the scatterplot.

(The larger their contribution is, the larger the residuals will be.)

Notes

(Optional for Spring 2020)



Notes

(Optional for Spring 2020)

Variation in **weights** of snakes is due to **two sources**:

1. **Lengths** of the snakes (the X **variable**).
2. All **other variables** *besides length* that affect weight (i.e. **random error**).

Notes

(Optional for Spring 2020)

- Variation in Y due to **random error** (i.e. due to **all other variables *besides*** X) is measured by the ***error sum of squares***, denoted **SSE** and defined as follows.

Error Sum of Squares:

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 \,.$$

SSE is just the **sum of squared residuals**. SSE will be **large** when **other variables *besides*** X contribute substantially to the variation in Y (i.e. when the variation due to **random error** is large).

Notes

(Optional for Spring 2020)

- Variation in Y due to differences in the value of the $\boldsymbol{X}$ **variable** is measured by the ***regression sum of squares***, denoted **SSR** and defined as follows.

Regression Sum of Squares:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 .$$

SSR is the **sum of squared deviations** of the **fitted values** away from the **overall mean response** $\bar{Y}$. A fitted line with a **steeper slope** will result in a **larger** SSR. Thus SSR will be **large** when the $\boldsymbol{X}$ **variable** contributes substantially to the variation in Y.

Notes

ANOVA-Like Partition of the Variation (Optional for Spring 2020)

- We measure **total variation** in Y by the ***total sum of squares***, denoted **SSTo** and defined as follows.

Total Sum of Squares:

$$\text{SSTo} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

SSTo reflects variation in Y due to $\boldsymbol{X}$ ***and*** due to **other variables *besides*** $\boldsymbol{X}$ (i.e. **random error**). SSTo will be large if either the $\boldsymbol{X}$ **variable *or* random error** contributes substantially to Y variation.

Notes

(Optional for Spring 2020)

- It can be shown that

ANOVA-Like Partition:

$$\text{SSTo} = \text{SSR} + \text{SSE}.$$

This partitions the variation in the data as:

$$\begin{aligned}\text{Total Variation} &= \text{Variation Due to } \boldsymbol{X} \\ &\quad + \text{Variation Due to All Other Variables } \textit{Besides } \boldsymbol{X} \\ &= \text{Variation Due to } \boldsymbol{X} \\ &\quad + \text{Variation Due to Random Error}\end{aligned}$$

Notes

(Optional for Spring 2020)

Example

For the snakes data, the **sums of squares** (obtained using statistical software) are

$$\text{SSTo} = \mathbf{62,255}, \quad \text{SSR} = \mathbf{51,090}, \quad \text{and} \quad \text{SSE} = \mathbf{11,165}.$$

As expected, SSTo $=$ SSR $+$ SSE since

$$\mathbf{62,255} = \mathbf{51,090} + \mathbf{11,165}.$$

This shows that the majority of the variation in weights (**51,090** out of **62,255**) is due to differences among their **lengths**, and only a small portion (**11,165** out of **62,255**) due to **random error** (i.e. all other variables).

Notes

Degrees of Freedom (Optional for Spring 2020)

- Each sum of squares has a corresponding **degrees of freedom**.

Degrees of Freedom: For linear regression, the degrees of freedom are:

$$\begin{aligned} df \text{ for SSTo} &= n-1 \\ df \text{ for SSR} &= 1 \\ df \text{ for SSE} &= n-2 \end{aligned}$$

Degrees of freedom will be used later to determine which t and F distributions **p-values** are obtained from for hypothesis tests.

Notes

Mean Squares (Optional for Spring 2020)

- A ***mean square*** is a sum of squares divided by its degrees of freedom.

Mean Squares: For linear regression, the ***mean square for regression***, ***MSR***, and ***mean squared error***, ***MSE***, are

$$\text{MSR} = \frac{\text{SSR}}{1} = \text{SSR}$$

$$\text{MSE} = \frac{\text{SSE}}{n-2}.$$

MSR and **MSE** will be used later in a so-called ***regression model*** F ***test***.

Notes

(Optional for Spring 2020)

Example

For the data on lengths and weights of $n = 9$ snakes, the **mean squares** (from software) are

$$\text{MSR} = 51,090 \quad \text{and} \quad \text{MSE} = 1,595$$

and so

$$\sqrt{\text{MSE}} = 39.9.$$

This is the size of a **typical residual**, and serves as an **estimate** of σ, the standard deviation of the N$(0, \sigma)$ distribution of the **random error** term in the regression model.

Notes

R Squared (Revisited) (Optional for Spring 2020)

- Recall that R^2 measures **how well the fitted line fits the data**.
- One way to compute R^2 is to **square the correlation** r:

$$R^2 = r^2.$$

Notes

(Optional for Spring 2020)

- It turns out that another way to compute R^2 is to use **sums of squares**:

$$R^2 = \frac{\text{SSR}}{\text{SSTo}} \qquad \left(= 1 - \frac{\text{SSE}}{\text{SSTo}}\right).$$

- Because SSR measures variation in Y due to X, and SSTo measures *total* variation in Y, R^2 is

$$R^2 = \frac{\text{Variation in } Y \text{ Due to } X}{\text{Total Variation in } Y}.$$

This explains why $\boldsymbol{R^2}$ is interpreted as **the proportion of variation in $\boldsymbol{Y}$ that's explained by $\boldsymbol{X}$**.

Notes

Regression Model F Test (Optional for Spring 2020)

- Another way to test for the slope coefficient $\boldsymbol{\beta_1}$ is to perform the so-called called ***regression model F test***.

The null and alternative hypotheses are exactly the same as for the **t test**, namely

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Notes

(Optional for Spring 2020)

- But the ***regression model F test statistic*** is

F Test Statistic for the Regression Model:

$$F = \frac{\text{MSR}}{\text{MSE}}.$$

Because **MSR** measures **variation in Y due to X** (i.e. due to the fitted line having a steep slope) and **MSE** measures **variation due to random error**, F will be **large** when the fitted line has a steep slope relative to the sizes of the residuals.

Large values of F provide evidence *against* H_0 in favor of H_a.

Notes

(Optional for Spring 2020)

Sampling Distribution of F Under H_0: If the errors ϵ_i in the regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ follow a N$(0, \sigma)$ distribution and F is the test statistic in a **regression model F test**, then when

$$H_0 : \beta_1 = 0$$

is true,

$$F \sim F(1, n-2),$$

the F distribution with numerator degrees of freedom 1 and denominator degrees of freedom $n - 2$.

Notes

(Optional for Spring 2020)

- **P-values** for the **regression model F test** are obtained from the **right tail** of the $F(1, n-2)$ distribution.

Notes

(Optional for Spring 2020)

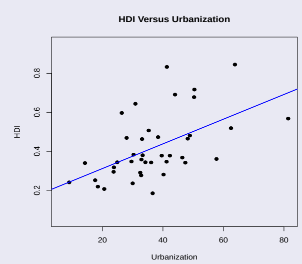
Example

For the data on the human development index (**HDI**) values and **urbanization** rates for the $n = 40$ sub-Saharan countries, statistical software reports the **sums of squares**, **degrees of freedom**, and **regression model F test** results in the following so-called ***regression ANOVA table***.

Source	DF	SS	MS	F	P-value
Regression	1	0.320	0.320	15.83	0.0003
Error	38	0.768	0.020		
Total	39	1.088			

Notes

(Optional for Spring 2020)



Notes

(Optional for Spring 2020)

From the table, the **F test statistic** for the **regression model F test** of

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \end{aligned}$$

is $F = 15.83$ and the **p-value** is **0.0003**, indicating a **statistically significant linear relationship** between the **HDI** and **urbanization** rate.

Notes

(Optional for Spring 2020)

- It can be shown that the t **test for the slope** and the **regression model** F **test** will always come to the **same conclusion**.

Notes

Notes

Notes

Notes