Notes

13 Multiple Regression (Cont'd)

MTH 3240 Environmental Statistics

Spring 2020

Notes

Objectives

Objectives:

- Interpret sums of squares, degrees of freedom, and mean squares in multiple regression (**Optional for Spring 2020**).
- Interpret the R^2 associated with a fitted multiple regression model (**Optional for Spring 2020**).
- Carry out a regression model F test (**Optional for Spring 2020**).
- Use the adjusted R^2 to compare suitabilities of models containing different sets of explanatory variables (**Optional for Spring 2020**).

Notes

Sums of Squares in Multiple Regression (Optional for Spring 2020)

- In a **multiple regression analysis**, we can split the **total variation** in the $\boldsymbol{Y_i}$ values into two parts:
  1. Variation in Y that's due to the $\boldsymbol{p}$ **variables** $\boldsymbol{X_1, X_2, \ldots, X_p}$ that are included in the model.
  2. Variation in Y due to **all other variables** *besides* $X_1, X_2, \ldots, X_p$ (which collectively produce the **random errors** in the data).

Notes

(Optional for Spring 2020)

Example

For the data on **water consumption** in 28 U.S. cities, cities' **wealth** levels and **population** sizes **explain *some*** of the **variation** in **water consumptions**, but not all of it.

Other variables that affect water consumption (e.g. temperature, precipitation, landscape characteristics, number of swimming pools, etc.) show up as **residuals** (i.e. **random error**) in a regression analysis.

The larger the contributions of these other variables to **water consumption**, the larger the residuals will be.

(Optional for Spring 2020)

- Variation in Y due to **random error** (i.e. due to **all other variables *besides*** $\boldsymbol{X_1, X_2, \ldots, X_p}$) 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*** $\boldsymbol{X_1, X_2, \ldots, X_p}$ contribute substantially more 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 the variables $\boldsymbol{X_1, X_2, \ldots, X_p}$ 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** $\bar{\boldsymbol{Y}}$. A fitted plane (when $p = 2$) with a **steeper tilt** will result in a **larger** SSR. Thus SSR will be **large** when $\boldsymbol{X_1, X_2, \ldots, X_p}$ contribute 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 $\boldsymbol{X_1, X_2, \ldots, X_p}$ ***and*** due to **other variables *besides*** $\boldsymbol{X_1, X_2, \ldots, X_p}$ (i.e. **random error**). SSTo will be large if either the $\boldsymbol{X_1, X_2, \ldots, X_p}$ **variables *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:

$$\begin{aligned}\text{Total Variation} &= \text{Variation Due to } \boldsymbol{X_1, X_2, \ldots, X_p} \\ &\quad + \text{Variation Due to All Other Variables } \textit{Besides } \boldsymbol{X_1, X_2, \ldots, X_p} \\ &= \text{Variation Due to } \boldsymbol{X_1, X_2, \ldots, X_p} \\ &\quad + \text{Variation Due to Random Error}\end{aligned}$$

Notes

(Optional for Spring 2020)

Example

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

$$\text{SSTo} = \mathbf{24.9}, \quad \text{SSR} = \mathbf{18.6}, \quad \text{and} \quad \text{SSE} = \mathbf{6.3}.$$

As expected, SSTo = SSR + SSE since

$$\mathbf{24.9} = \mathbf{18.6} + \mathbf{6.3}.$$

This shows that the majority of the variation in water consumption (**18.6** out of **24.9**) is due to cities' **wealth** levels and **population** sizes, and only a smaller portion (**6.3** out of **24.9**) due to **random error** (i.e. all other variables).

Notes

Degrees of Freedom (Optional for Spring 2020)

- As for one-factor ANOVA, each sum of squares has a corresponding **degrees of freedom**.

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

$$\begin{aligned} df \text{ for SSTo} &= n-1 \\ df \text{ for SSR} &= p \\ df \text{ for SSE} &= n-(p+1) \end{aligned}$$

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

Notes

(Optional for Spring 2020)

Example

For the water consumption data, there are $n = 28$ U.S. cities and $p = 2$ explanatory variables, the degrees of freedom are

$$df \text{ for SSTo} = \mathbf{27}, \quad df \text{ for SSR} = \mathbf{2}, \quad \text{and} \quad df \text{ for SSE} = \mathbf{25}.$$

Notes

Mean Squares (Optional for Spring 2020)

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

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

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

$$\text{MSE} = \frac{\text{SSE}}{n-(p+1)}.$$

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 water consumption of 28 U.S. cities, the **mean squares** (from software) are

$$\text{MSR} = \mathbf{9.30} \quad \text{and} \quad \text{MSE} = \mathbf{0.252}$$

and so

$$\sqrt{\text{MSE}} = \mathbf{0.501}.$$

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

R Squared (Revisited) (Optional for Spring 2020)

- Recall that the so-called $\boldsymbol{R^2}$ value measures **how well the fitted model fits the data**.
- To compute R^2, we use **sums of squares**:

Coefficient of Multiple Determination:

$$R^2 = \frac{\text{SSR}}{\text{SSTo}} = 1 - \frac{\text{SSE}}{\text{SSTo}}.$$

(Optional for Spring 2020)

- Because SSR measures variation in Y due to $X_1, X_2, \ldots, X_p$, and SSTo measures *total* variation in Y, R^2 is

$$R^2 = \frac{\text{Variation in } Y \text{ Due to } X_1, X_2, \ldots, X_p}{\text{Total Variation in } Y}.$$

This explains why $\boldsymbol{R^2}$ can be interpreted as **the proportion of variation in $\boldsymbol{Y}$ that's explained by $\boldsymbol{X_1, X_2, \ldots, X_p}$.**

Adjusted R Squared (Optional for Spring 2020)

- Our **goals** when deciding ***which* explanatory variables** to include in a model are:
  1. The model should **fit the data well** (i.e. it should explain much of the Y variation).
  2. The model should be **parsimonious** (i.e. not contain excessive numbers of explanatory variables).

(Optional for Spring 2020)

- Although R^2 is useful for summarizing how well a given model fits the data, it's **not** very useful for deciding whether a variable should be added to or removed from a model.
- The reason for this is that:
  - $\boldsymbol{R^2}$ will **always increase** if another explanatory variable is added to the model.
  - It will **always decrease** if a variable is removed from the model.

  Thus R^2 would *always* favor a model with *more* variables in it, ...

  ... and we'd end up with a **non-parsimonious** model.

Notes

(Optional for Spring 2020)

- The ***adjusted*** $\boldsymbol{R^2}$, denoted $\boldsymbol{R^2_{adj}}$, "adjusts" the R^2 value for the number of explanatory variables that are in the model.

Notes

(Optional for Spring 2020)

- Unlike R^2, the $\boldsymbol{R^2_{adj}}$ doesn't necessarily increase every time we add another variable to the model, and it doesn't necessarily decrease every time we remove a variable.

  In particular, it ***decreases*** when we add a variable that's either **unrelated to** $\boldsymbol{Y}$ altogether, or **unrelated to** $\boldsymbol{Y}$ while **controlling** for the effects of **other variables** *already* in the model.

  (The second scenario can arise when the added variable is related to Y but is **correlated** with a variable that's already in the model.)

  It ***increases*** when we remove a variable that's either **unrelated to** $\boldsymbol{Y}$ altogether, or **unrelated to** $\boldsymbol{Y}$ while **controlling** for the effects of **other variables** in the model.

Notes

(Optional for Spring 2020)

- $\boldsymbol{R^2_{adj}}$ can be used to decide whether a variable should be added or removed from a model.
  - If $\boldsymbol{R^2_{adj}}$ **increases** when we add (or remove) a variable from the model, then that variable **should** be added (or removed).
  - If $\boldsymbol{R^2_{adj}}$ **decreases** when we add (or remove) a variable from the model, then that variable **shouldn't** be added (or removed).

Notes

(Optional for Spring 2020)

- The $\boldsymbol{R^2_{\text{adj}}}$, is computed using the following:

Adjusted $\boldsymbol{R^2}$:

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-(p+1))}{\text{SSTo}/(n-1)} = 1 - \frac{\text{MSE}}{S_y^2},$$

where S_y^2 is the sample variance of $Y_1, Y_2, \ldots, Y_n$.

Notes

(Optional for Spring 2020)

Example

For the study of **water consumption** in in 28 U.S. cities, we can investigate the changes in $\boldsymbol{R^2}$ and $\boldsymbol{R^2_{\text{adj}}}$ when **wealth** is added to a model that *already* includes **population**.

For the model with **population** as the sole explanatory variable,

$$\boldsymbol{R^2 = 0.740} \quad \text{and} \quad \boldsymbol{R^2_{\text{adj}} = 0.729}.$$

When we add **wealth** to the model that *already* includes **population**, we get

$$\boldsymbol{R^2 = 0.747} \quad \text{and} \quad \boldsymbol{R^2_{\text{adj}} = 0.727}.$$

Notes

(Optional for Spring 2020)

We see that $\boldsymbol{R^2}$ increases (which is to be expected), but only marginally, and $\boldsymbol{R^2_{\text{adj}}}$ actually *decreases*.

The implication is that the improved fit that results from adding **wealth** to the model is not worth the trade-off of having an extra variable in the model.

In other words, if the goal is to find a model that *fits the data well* (explains most of the variation in **water consumption**), but is *parsimonious*, **wealth** should be left out of the model.

Notes

Regression Model $\boldsymbol{F}$ Test (Optional for Spring 2020)

- A way to **simultaneously** test for the coefficients $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_p}$ is to perform the so-called called ***regression model $\boldsymbol{F}$ test***.

  The null and alternative hypotheses are

$$\begin{aligned} H_0: &\quad \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_a: &\quad \text{The } \beta_p\text{'s aren't all 0} \end{aligned}$$

  The null hypothesis says $\boldsymbol{Y}$ *isn't* related to **any** of the variables $X_1, X_2, \ldots, X_k$. The alternative says Y *is* related to **at least one** of $X_1, X_2, \ldots, X_k$.

Notes

(Optional for Spring 2020)

- The ***regression model F test statistic*** is

F Test Statistic for the Regression Model:

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

F will be **large** when the amount Y variation that's explained by the explanatory variables is large relative to the amount of variation that's due to random error.

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 the errors ϵ_i in the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i$ follow a $\text{N}(0, \sigma)$ distribution and F is the test statistic in a **regression model F test**, then when

$$H_0: \ \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

is true,

$$F \sim F(p,\ n-(p+1)),$$

the F distribution with numerator degrees of freedom p and denominator degrees of freedom $n-(p+1)$.

Notes

(Optional for Spring 2020)

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

Notes

(Optional for Spring 2020)

Example

For the data on **water consumption** in 28 U.S. cities, 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	Sum Sq	Mean Sq	F value	P-value
Regression	2	18.6	9.30	36.9	0.000
Error	25	6.3	0.252		
Total	27	24.9			

Notes

(Optional for Spring 2020)

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

$$H_0: \quad \beta_1 = \beta_2 = 0$$
$$H_a: \quad \text{Both } \beta_k\text{'s aren't 0}$$

is $F = 36.9$ and the **p-value** is **0.0000**, indicating a **statistically significant relationship** between **water consumption** and ***at least one*** of explanatory variables **wealth** and **population**.

Notes

Notes

Notes

Notes