Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Testing for Subsets of Regression Parameters

Georgia Tech

1

About This Lesson

Georgia Tech

2

Testing Overall Regression

Analysis of Variance (ANOVA) for multiple regression:

Variability Source	DF	Sum of Squares	Mean SS	F-Statistic
Regression	p	SSReg	SSReg / p	MSSReg / MSE
Residual	n-p-1	SSE	SSE / (n-p-1)	
Total	n-1	SST		

$$\mathrm{SSReg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Null hypothesis: All predictor coefficients are 0, i.e., $\mathbf{H_0}: \beta_1 = \beta_2 = \cdots = \beta_p = 0$.

Reject $\mathbf{H_0}$ if F-statistic is large (> $F_{\alpha,\ p,\ n\text{-}p\text{-}1}$ for α significance level, p and n-p-1 df).

- At least one of the coefficients is different from zero at the α significance level.

p-value = Prob($F_{p,\ n\text{-}p\text{-}1}$ > F-statistic) for F-distribution with p and n-p-1 df.

- Reject $\mathbf{H_0}$ if p-value is small.

Georgia Tech

3

Testing Subsets of Coefficients

Analysis of Variance (ANOVA):

$$\mathrm{SST}(X_1, \cdots, X_p) = \mathrm{SSReg}(X_1, \cdots, X_p) + \mathrm{SSE}(X_1, \cdots, X_p)$$

$$\mathrm{SSReg}(X_1, \cdots, X_p) = \mathrm{SSReg}(X_1) + \mathrm{SSReg}(X_2|X_1) + \mathrm{SSReg}(X_3|X_1, X_2) + \cdots + \mathrm{SSReg}(X_p|X_1, \cdots, X_{p-1})$$

$\mathbf{SSReg(X_1)}$: Sum of squares (SS) explained using only X_1

$\mathbf{SSReg(X_2|X_1)}$: ***Extra*** SS explained using X_2 in addition to X_1

$\mathbf{SSReg(X_3|X_1, X_2)}$: ***Extra*** SS explained using X_3 in addition to X_1 and X_2

$\mathbf{SSReg(X_p|X_1, \cdots, X_{p-1})}$: ***Extra*** SS explained using X_p in addition to X_1 , X_2 … X_{p-1}

Georgia Tech

4

Testing Subsets of Coefficients

- Does X_1 alone significantly aid in predicting Y?
  - SSReg(X_1) vs. SSE(X_1)
- Does the addition of X_2 significantly contribute to the prediction of Y after accounting (controlling) for the contribution of X_1?
  - SSReg(X_2 | X_1) vs. SSE(X_1, X_2)
- Does the addition of X_3 significantly contribute to the prediction of Y after accounting (controlling) for the contribution of X_1 and X_2?
  - SSReg(X_3 | X_1, X_2) vs. SSE(X_1, X_2, X_3)
- Does the addition of X_p significantly contribute to the prediction of Y after accounting (controlling) for the contribution of X_1,..., X_{p-1}?
  - SSReg(X_p | X_1, ..., X_{p-1}) vs. SSE(X_1, X_2, ..., X_p)

Georgia Tech

5

Testing Subsets of Coefficients

Partial F-test:

- Consider a full model with two sets of predictors, X_1, ..., X_p (perhaps controlling factors) and (Z_1, ..., Z_q) (perhaps additional explanatory factors):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha_1 Z_1 + \cdots + \alpha_q Z_q + \varepsilon$$

- Test whether any of the Z factors add explanatory power to the model:

$$\mathbf{H_0}: \alpha_1 = \alpha_2 = \cdots = \alpha_q = 0 \quad \textbf{vs.} \quad \mathbf{H_a}: \alpha_i \neq 0 \text{ for at least one } \alpha_i, i = 1, \cdots, q$$

$$\text{F−statistic} = \text{F}_{partial} = \frac{\text{SSReg}(Z_1, \ldots, Z_q | X_1, \ldots, X_p)/q}{\text{SSE}(Z_1, \ldots, Z_q, X_1, \ldots, X_p)/(n - p - q - 1)}$$

- Reject $\mathbf{H_0}$ if F-statistic is large (F-statistic $> \text{F}_{\alpha,\, q,\, n-p-q-1}$)
  - At least one coefficient is different from zero at the α significance level

Georgia Tech

6

Testing for Statistical Significance

- Consider a full model with the set of predictors, X_1, ..., X_p and an additional predicting variable Z:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \alpha Z + \varepsilon$$

- Test whether Z has explanatory or predictive power:

$$\mathbf{H_0}: \alpha = 0 \text{ vs } \mathbf{H_a}: \alpha \neq 0$$

$$\text{F-statistic} = \text{F}_{partial} = \frac{\text{SSReg}(Z|X_1, \dots, X_p)/1}{\text{SSE}(Z, X_1, \dots, X_p)/(n-p-2)}$$

- Reject $\mathbf{H_0}$ if F-statistic is large (F-statistic > $F_{\alpha, 1, n-p-2}$)

This is equivalent to testing for statistical significance using the t-test

Georgia Tech

7

Testing for Statistical Significance

- Consider a full model with the set of predictors, X_1, ..., X_p and an additional predicting variable Z:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \alpha Z + \varepsilon$$

- Test whether Z has explanatory or predictive power:

$$\mathbf{H_0}: \alpha = 0 \text{ vs } \mathbf{H_a}: \alpha \neq 0$$

$$\text{F-statistic} = \text{F}_{partial} = \frac{\text{SSReg}(Z|X_1, \dots, X_p)/1}{\text{SSE}(Z, X_1, \dots, X_p)/(n-p-2)}$$

- Reject $\mathbf{H_0}$ if F-statistic is large (F-statistic > $F_{\alpha, 1, n-p-2}$)

This is equivalent to testing for statistical significance using the t-test

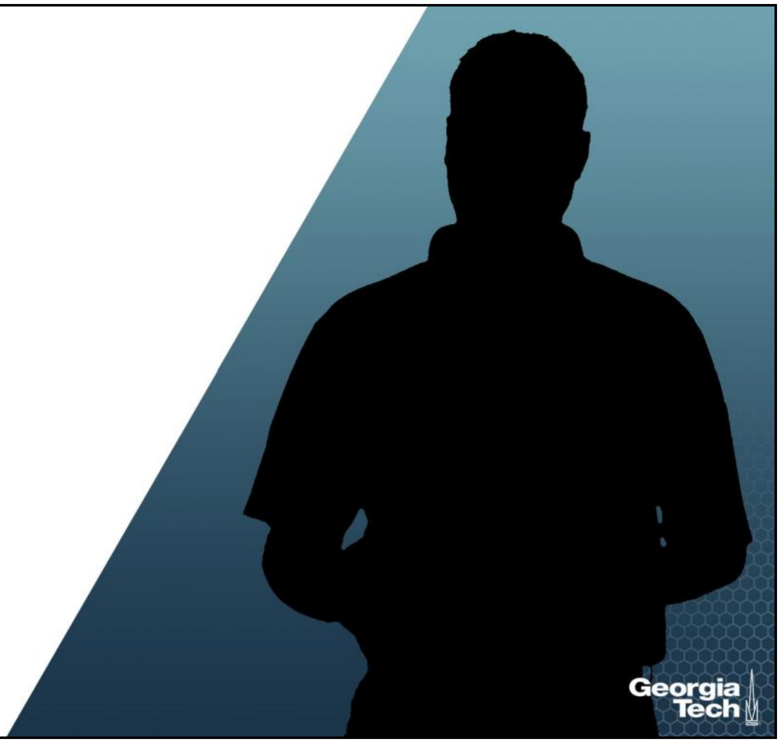
- Interpretation of the t-test for statistical significance is conditional on other predicting variables being in the model.
- The relationship between Y and X is statistically significant given all other predicting variables being in the model.

Do not perform variable selection based on the p-values of the t-tests!

Georgia Tech

8

Summary



9