STAT 302 - Chapter 3 : Multiple Regression - Part 5

Harsha Perera

Key Topics

- Testing subsets of predictors
- Nested F-Test

Testing subsets of predictors

- We discussed two types of tests for predictors
  1. Overall ANOVA F-test
  2. Individual t-Test

- ANOVA F-test: allows us to test the effectiveness of all the predictors in the model as a group :
  $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ vs H_1 : at least one $\beta_i \neq 0$

- t-test: allows us to test the importance of a single predictor in the model :
  $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$

- None of the methods allow us to test the contribution of subset of a predictors in the model.

- In this section we describe a general procedure to test the contribution of subset of a predictors in the model.

Nested F-Test

Allows us to test the contribution of subset of predictors in the model

- What is a **Nested Model**?
  Example: consider a **complete second order model** of two predictors

  $$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_1^2 + \beta_4 \cdot X_2^2 + \beta_5 \cdot X_1 \cdot X_2 + \in$$

- **Nested models** of the complete second order model of two predictors are
  - Interaction model:
  $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_5 \cdot X_1 \cdot X_2 + \in$
  - Second order polynomial model:
  $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_3 \cdot X_1^2 + \in$
- In this example,
  - Complete second order model is called the **Full Model**
  - Nested models are called the **Reduced Models**

Example 3:18 - House Prices : Comparing models

- **Full model**:

$$Price = \beta_0 + \beta_1 \cdot Beds + \beta_2 \cdot Baths + \beta_3 \cdot Size + \in$$

  Lets say we want to test the contribution of both Beds and Size in a single test :

  $H_0 : \beta_1 = \beta_3 = 0$ vs. H_1 : at least one $\beta_i \neq 0$

- Therefore **reduced model**:

$$Price = \beta_0 + \beta_2 \cdot Baths + \in$$

- We need to compare the effectiveness of full model (3 predictors) and the reduced model (1 predictor)
- If we reject the null hypothesis: Then we need to retain either Size or Beds in the model

Nested F-Test

We want to asses whether the amount of new variability is significant or not?

Amount of new variability $= SSModel_{full} - SSModel_{reduced}$

$$F = \frac{\frac{SSModel_{full} - SSModel_{reduced}}{\text{predictors tested}}}{\frac{SSE_{full}}{(n-k-1)}}$$

- n: total number of observations
- k: number of predictors in the full model
- P-value is computed from an F-distribution with numerator DF equal to **the number of predictors being tested** and denominator DF equal to **the error DF for the full model**
- Depending on the p-value we can decide whether to reject H_0 or not

House Prices in NYC - Nested F-Test

Full model:

$$SSModel_{full} = 23407 + 276 + 12605 = 36288$$

Reduced model:

$$SSModel_{reduced} = 27821$$

$$F = \frac{\frac{SSModel_{full} - SSModel_{reduced}}{\text{predictors tested}}}{\frac{SSE_{full}}{(n-k-1)}}$$

$$F = \frac{\frac{36288-27821}{2}}{\frac{52967}{(49)}} = 3.92$$

P-value $= 0.026 < 0.05$. We reject H_0 at 0.05 level of significance. We may want to retain either Size or Beds in the model.

Nested F-Test

An equivalent way to compute the amount of new variability explained by the predictors being tested is :

$$SSModel_{full} - SSModel_{reduced} = SSE_{reduced} - SSE_{full}$$

Therefore we can re-write the F-Test statistic as :

$$F = \frac{\frac{SSE_{reduced} - SSE_{Full}}{\text{predictors tested}}}{\frac{SSE_{full}}{(n-k-1)}}$$

House Prices in NYC - Nested F-Test :

$$F = \frac{\frac{61434-52967}{2}}{\frac{52967}{(49)}} = 3.92$$

Example 3.19 : NFL Winning Percentage : Nested F-Test

The data set contains :

- Winning percentage of each NFL team during 2016 regular season,which is out of 16 games (*WinPct*).
  $WinPct100 = WinPct \times 100$
- Number of points scored (*PointsFor*)
- Number of points allowed (*PointsAgainst*)
- Total yards the team gained (*YardsFor*)
- Number of yards for the opponents (*YardsAgainst*)
- Number of touchdowns (*TDs*)

Example 3.19 - Contd ...

Full model : Model with all five variables:

$$WinPct100 = \beta_0 + \beta_1 \cdot PointsFor + \beta_2 \cdot PointsAgainst + \beta_3 \cdot YardsFor$$

$$+ \beta_4 \cdot YardsAgainst + \beta_5 \cdot TDs + \in$$

Reduced Model : Model with PointsAgainst and YardsAgainst :

$$WinPct100 = \beta_0 + \beta_2 \cdot PointsAgainst + \beta_4 \cdot YardsAgainst + \in$$

Therefore we shoud test :

$H_0 : \beta_1 = \beta_3 = \beta_5 = 0$
$H_1 : \beta_i \neq 0$ at least one of the three

Example 3.19 - Contd ...

$$F = \frac{\frac{SSE_{redcued} - SSE_{full}}{\text{predictors tested}}}{\frac{SSE_{full}}{(n-k-1)}} = \frac{\frac{5085.9-2218.3}{3}}{\frac{2218.3}{(26)}} = \frac{\frac{2867.7}{3}}{85.37} = 11.2$$

- P-value of F statistics < 0.05
- We reject null hypothesis at 0.05 level of significance
- At least one of the *PointsFor*, *YardsFor* and *TDs* need to be included in the prediction equation