Lecture 8: F-Test for Nested Linear Models

Zhenke Wu
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
zhwu@jhu.edu
http://zhenkewu.com

11 February, 2016

Lecture 7 Main Points Again

Constructing F-distribution:

- $Y_i \overset{independently\ distributed}{\sim} Gaussian(\mu_i, \sigma_i^2)$
- $Z_i = \frac{Y_i - \mu_i}{\sigma_i}$; $Z_i \overset{iid}{\sim} Gaussian(0, 1)$
- Define **quadratic** forms $Q_1 = Z_1^2 + \cdots + Z_{n_1}^2$ and $Q_2 = Z_{n_1+1}^2 + \cdots + Z_{n_1+n_2}^2$
- $Q_1 \sim \chi^2_{n_1}$ with mean n_1 and variance $2n_1$
- $Q_2 \sim \chi^2_{n_2}$ with mean n_2 and variance $2n_2$
- Q_1 is **independent** of Q_2
- $F_{n_1,n_2} = \frac{Q_1/n_1}{Q_2/n_2} \sim \mathcal{F}(n_1, n_2)$ (F-distribution with n_1 and n_2 degrees of freedom; "F" for Sir R.A. Fisher)

Lecture 7 Main Points Again (continued)

- Data:
  - n observations; $p+s$ covariates
  - continuous outcome Y_i, measured with error
  - covariates: $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip}, X_{i,p+1}, \ldots, X_{i,p+s})^\top$, for $i = 1, \ldots, n$
- **Question: In light of data, can we use a simpler linear model nested within a complex one?**
- Hypothesis testing:

  (a) Null model: $\mathbf{Y} \sim \text{Gaussian}_n(\mathbf{X}_N \boldsymbol{\beta}_N, \sigma^2 \mathbf{I}_n)$
  - $\mathbf{X}_N$: design matrix $n \times (p+1)$ obtained by stacking observations X_i
  - First p (transformed) covariates and 1 intercept
  - Regression coefficients: $\boldsymbol{\beta}_N = (\beta_0, \beta_1, \ldots, \beta_p)^\top$
  - Standard deviation of measurement errors: σ

  (b) Extended model: $\mathbf{Y} \sim \text{Gaussian}_n(\mathbf{X}_E \boldsymbol{\beta}_E, \sigma^2 \mathbf{I}_n)$
  - $\mathbf{X}_E$: design matrix with intercept$+p+s$ covariates
  - $\boldsymbol{\beta}_E = (\boldsymbol{\beta}_N^\top, \beta_{p+1}, \ldots, \beta_{p+s})^\top$

  - Null model: H_0: $\beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+s} = 0$

Lecture 7 Main Points Again (continued)

Null model: H_0: $\beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+s} = 0$

Let $\boldsymbol{\beta}_{[p+]} = (\beta_{p+1}, \cdots, \beta_{p+s})^\top$

- Rationale of the F-Test
  - If H_0 is true, estimates $\widehat{\beta}_{p+1}, \cdots, \widehat{\beta}_{p+s}$ should all be close to 0
  - Reject H_0 if these estimates are sufficiently different from 0s.
  - However, not every $\widehat{\beta}_{p+j}, j = 1, \ldots, s$, should be treated the same; they have different precisions
  - Use a quadratic term to measure their **joint** differences from 0, taking account of different precisions:

$$\widehat{\boldsymbol{\beta}}_{[p+]}^\top \left(\mathrm{Var}_E[\widehat{\boldsymbol{\beta}}_{[p+]}]\right)^{-1} \widehat{\boldsymbol{\beta}}_{[p+]} \quad (1)$$

  - $\mathrm{Var}_E[\widehat{\boldsymbol{\beta}}_{[p+]}] = \sigma^2 \mathbf{A}(\mathbf{X}_E^\top \mathbf{X}_E)^{-1}\mathbf{A}^\top$, where $\mathbf{A} = [\mathbf{0}_{s\times(p+1)}, \mathbf{I}_{s\times s}]$
  - Estimate σ^2 by $\mathrm{RSS}_E/(n-p-s-1)$; RSS for "residual sum of squares"

Lecture 7 Main Points Again (continued)

- $$F = \frac{(RSS_N - RSS_E)/s}{RSS_E/(n-p-s-1)} \quad (2)$$
- $F(s, n-p-s-1)$: F-distribution with s and $n-p-s-1$ degrees of freedom
- $RSS_N = Y'(I - H_N)Y$; $H_N = X_N(X_N'X_N)^{-1}X_N$; "$H$" for **hat** matrix, or projector
- $RSS_E = Y'(I - H_E)Y$; $H_E = X_E(X_E'X_E)^{-1}X_E$
- $(RSS_N - RSS_E)/\sigma^2 \sim \chi^2_s$ and $RSS_E/\sigma^2 \sim \chi^2_{n-p-s-1}$; they are **independent**
  [Proof]:
  - Algebraic: The former is a function of $\widehat{\boldsymbol{\beta}}_E$, which is independent of RSS_E]
  - Geometric: Squared lengths of orthogonal vectors

Geometric Interpretation: Projection

- $\widehat{Y}_N = H_N Y$: fitted means under the null model
- $\widehat{Y}_E = H_E Y$: fitted means under the extended model

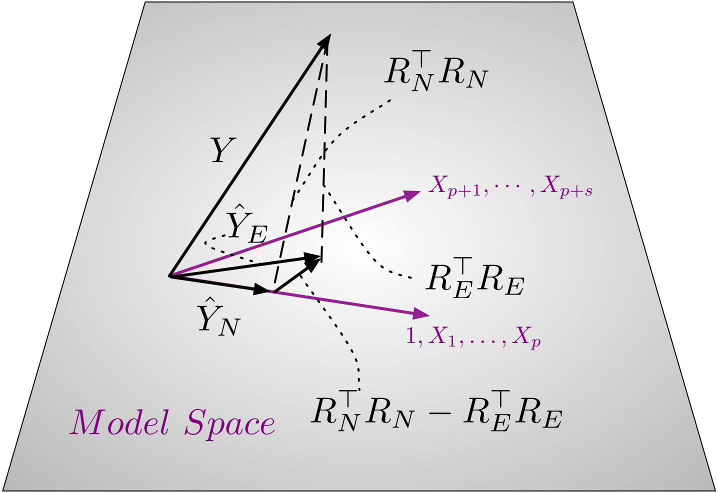

Table: ANOVA for Regression

Model	df	Resudial df	Residual Sum of Squares (RSS)	Residual **Mean** Square
Null	$p+1$	$n-p-1$	$RSS_N = R_N' R_N$	$\frac{R_N' R_N}{n-p-1} = S_N^2$
Extended	$p+s+1$	$n-p-s-1$	$RSS_E = R_E' R_E$	$\frac{R_E' R_E}{n-p-s-1} = S_E^2$
Change	s	$-s$	$(R_N' R_N - R_E' R_E)$ $= R_N' R_N - R_E' R_E$	$\frac{R_N' R_N - R_E' R_E}{s}$

- $F_{s,n-p-s-1} = \frac{(R_N' R_N - R_E' R_E)/s}{R_E' R_E/(n-p-s-1)}$
- Reject H_0 if $F > \underbrace{\mathcal{F}_{1-\alpha}(s, n-p-s-1)}_{(1-\alpha\%)\ percentile\ of\ the\ \mathcal{F}\ distribution}$, e.g., $\alpha = 0.05$

Some Quick Facts about F-distribution

Special cases of $\mathcal{F}(n_1, n_2)$

- $n_2 \to \infty$:
  - $Q_2/n_2 \xrightarrow{in\ probability} constant$
  - For a fixed n_1, $F_{n_1,n_2} \xrightarrow{in\ distribution} Q_1/n_1 \sim \chi^2_{n_1}/n_1$ as n_2 approaches infinity
  - Or equivalently $n_1 F_{n_1,\infty} \sim \chi^2_{n_1}$
- If $s = 1$:
  - The F-statistic equals $(\widehat{\beta_{p+1}}/se_{\widehat{\beta}_{p+1}})^2$ for testing the null model $H_0: \ \beta_{p+1} = 0$
  - Under H_0, it is distributed as $\mathcal{F}(1, n-p-2)$
  - Approximately distributed as $\chi^2_1/1$ when $n >> p$ (therefore 3.84 is the critical value at the 0.05 level)

F-Table

For F distribution with denominator $df_2 = 1, 2$, the 0.95 percentile increases with df_1; for $df_2 > 2$, the percentile decreases with df_1.

$df_2 \backslash df_1$	1	2	3	10	100
1	161.45	199.50	215.71	241.88	253.04
2	18.51	19.00	19.16	19.40	19.49
3	10.13	9.55	9.28	8.79	8.55
100	3.94	3.09	2.70	1.93	1.39
1000	3.85	3.00	2.61	1.84	1.26
∞	3.84	3.00	2.60	1.83	1.24

Table: 95% quantiles for F-distribution with degrees of freedom df_1 and df_2.

F-Table

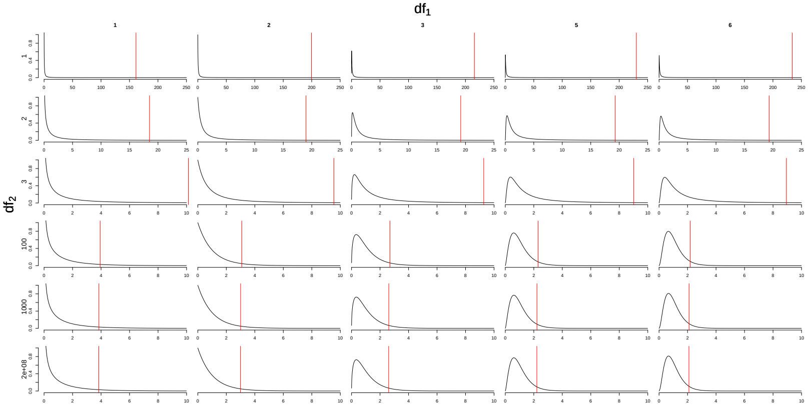

Figure: Density functions for F distributions; Red lines for 95% quantiles

Example

- Data: National Medical Expenditure Survey (NMES)
- Objective: To understand the relationship between medical expenditures and presence of a major smoking-caused disease among persons who are similar with respect to age, sex and SES
- $Y_i = \log_e(total\ \ medical\ \ expenditure_i + 1)$
- $X_{i1} = age_i - 65\ \ years$
- $X_{i2} = ♂$
- $\#\ \ of\ \ subjects : n = 4078$

Example

Table: NMES Fitted Models

Model	Design	df	Residual MS	Resid. df
A	X_1, X_2	3	1.521	4075
B	$X_1, (X_1 - (-20)^+, (X_1 - 0)^+),\ X_2$	5	1.518	4073
C	$\underbrace{[X_1, (X_1 - (-20)^+, (X_1 - 0)^+)] * X_2}_{all\ \ interactions\ \ and\ \ main\ \ effects}$	8	1.514	4070

NMES Example: Question 1

Is average log medical expenditures roughly a linear function of age?

- Compare which two models?
- Calculate Residual Sum of Squares and Residual Mean Squares.
- Calculate F-statistic; What are the degrees of freedom for its distribution under the null?
- Compare it to the critical value at the 0.05 level

NMES Example: Question 1

- H_0: Within a larger model B, model A is true (or state the scientific meaning, i.e., linearity in age).
-

$$F = \underbrace{\frac{(RSS_N - RSS_E)/\overbrace{s}^{change\ \ in\ \ df}}{\underbrace{RSS_E}_{residual\ \ sum\ \ of\ \ squares} / \underbrace{(n-p-s-1)}_{residual\ \ df}}}_{residual\ \ \textbf{mean}\ \ squares} \tag{3}$$

$$= \frac{(1.521 \times 4075 - 1.518 \times 4073)/2}{1.518} = 5.03 \tag{4}$$

- This statistic, under repeated sampling, has a $\mathcal{F}(2, 4073)$ distribution, which is approximately $\chi^2_2/2$ distributed.
- p-value: $Pr(\chi^2/2 > 5.03) = 0.0065$ by approximation or $Pr(\mathcal{F}(2, 4073) > 5.03) = 0.0066$ without approximation. The approximation is good.
- Reject linearity in age.

NMES Example: Question 2 (In-Class Exercise)

- Is the non-linear relationship of average log expenditure on age the same for ♂ and ♀? (Are there curves parallel?)
- Or equivalently, is the difference between average log medical expenditure for ♂-vs-♀ the same at all ages?

NMES Example: Question 2 (In-Class Exercise)

- H_0: Within a larger model C, model B is true (or equivalently state the scientific meaning, i.e., no interaction).
- $$F = \frac{(1.518 \times 4073 - 1.514 \times 4070)/3}{1.514} = 4.59 \qquad (5)$$
- Under repeated sampling, it is $\mathcal{F}(3, 4070)$ distributed.
- p-value $Pr(\chi^2_3/3 > 4.59) = 0.0032$ by approximation, or $Pr(\mathcal{F}(3, 4070) > 4.59) = 0.0033$ without approximation.
- Reject no-interaction assumption

Notes:

- Ingo's Notes: http://biostat.jhsph.edu/ iruczins/teaching/140.751/
- $F = \frac{n-p-s-1}{s}\left(\frac{RSS_N}{RSS_E} - 1\right) = \frac{n-p-s-1}{s}\left(\left\{\left[\frac{RSS_E/n}{RSS_N/n}\right]^{n/2}\right\}^{-2/n} - 1\right)$,
  where $\Lambda = \left[\frac{RSS_E/n}{RSS_N/n}\right]^{n/2}$ is the likelihood ratio test (LRT) statistic comparing the null versus the extended model. Because F and Λ are one-to-one, monotonically related, in this case the LRT and F-test are equivalent tests (e.g., the same p-values). However, F-statistic is preferred in practice for its nice approximations by Chi-square (e.g., when $df_2 \to \infty$) and connections to other distributions (e.g., $\mathcal{F}(1, df_2) \overset{d}{=} t^2_{df_2}$).

Next by Professor Scott Zeger:

- *Delta method* to calculate the variance of a **function** of estimates. For example, if we know the variance of **log** odds ratio (LOR) comparing two proportions, how do we obtain the variance of odds ratio (exponential of the LOR)?