Math 5305 Notes

Chapter 5

Jesse Crawford

Department of Mathematics
Tarleton State University

Outline

1. Section 5.6: Normal Theory
2. Section 5.7: The F-test
3. Chapter 5 Closing Remarks

Definition

- Suppose $U_1, \ldots, U_d$ are IID $N(0,1)$.
- Then the random variable

$$\sum_{i=1}^{d} U_i^2$$

has a χ^2 distribution with d degrees of freedom, denoted $\chi^2(d)$.

Definition

- Suppose U and V are independent, $U \sim N(0,1)$, and $V \sim \chi^2(d)$.
- Then the random variable

$$t = \frac{U}{\sqrt{V/d}}$$

has a t-distribution with d degrees of freedom, denoted $t(d)$.

Theorem

Assume a multiple regression model satisfies the assumptions from Chapter 4 and the disturbance term ϵ is normally distributed. Then

- $\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$.
- $e \perp\!\!\!\perp \hat{\beta}$
- $\|e\|^2 \sim \sigma^2 \chi^2_{n-p}$

Corollary

- *Assume a multiple regression model satisfies the assumptions from Chapter 4 and the disturbance term ϵ is normally distributed.*
- *Consider the testing problem*

$$H_0 : \beta_k = 0 \text{ vs. } H : \beta_k \neq 0.$$

- *Let*

$$t = \frac{\hat{\beta}_k}{\widehat{SE}},$$

where $\widehat{SE} = \hat{\sigma}\sqrt{(X'X)^{-1}_{kk}}$.

- *Under H_0, t has a t distribution with $n - p$ degrees of freedom.*
- *Decision Rule: Reject H_0 if $|t| > t_{\alpha/2}(n-p)$.*
- *The p-value corresponding to t is the area under a $t(n-p)$ curve outside of the range $(-t, t)$.*

Console ~/

```
> mymodel=lm(Y~X1+X2+X3)
> summary(mymodel)

Call:
lm(formula = Y ~ X1 + X2 + X3)

Residuals:
     Min       1Q   Median       3Q      Max 
-23.5493  -6.4823   0.7492   5.5936  24.9199 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 50.04082    2.94308  17.003  < 2e-16 ***
X1           4.43642    0.31869  13.921  < 2e-16 ***
X2          21.64408    3.12069   6.936  4.7e-10 ***
X3          -0.55334    0.03271 -16.919  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.463 on 96 degrees of freedom
Multiple R-squared: 0.8593,	Adjusted R-squared: 0.855 
F-statistic: 195.5 on 3 and 96 DF,  p-value: < 2.2e-16

> names(summary(mymodel))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"   
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled" 
> summary(mymodel)$r.squared
[1] 0.8593479
> 
```

Outline

1. Section 5.6: Normal Theory

2. Section 5.7: The F-test

3. Chapter 5 Closing Remarks

A Mathematics Education Example

Example

$$\begin{aligned}\text{Post-test}_i = \beta_1 &+ \beta_2\text{Pre-test}_i + \beta_3\text{HSrank}_i + \beta_4\text{MathSAT}_i + \beta_5\text{VerbSAT}_i \\ &+ \beta_6\text{NumAbsences} + \beta_7\text{GroupWork}_i + \beta_8\text{Clickers}_i \\ &+ \epsilon_i, \text{ for } i = 1, \ldots, 1000.\end{aligned}$$

Handling Dichotomous Variables

$$\text{Clickers}_i = \begin{cases} 1, & \text{if the } i\text{th subject used clickers} \\ 0, & \text{if the } i\text{th subject did not use clickers} \end{cases}$$

How would we handle a categorical variable with more than two levels?

Example

- Consider the variable high school, whose values are Abrams, Baldwin, Campbell, and Daniels.
- Variable has 4 levels.
- The first level, Abrams, is the *reference level*.
- The other three levels require "dummy variables", also called "design variables".

$$\text{Baldwin}_i = \begin{cases} 1, & \text{if the } i\text{th subject is from Baldwin High School} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Campbell}_i = \begin{cases} 1, & \text{if the } i\text{th subject is from Campbell High School} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Daniels}_i = \begin{cases} 1, & \text{if the } i\text{th subject is from Daniels High School} \\ 0, & \text{otherwise} \end{cases}$$

Full Model Incorporating High School

$$\begin{aligned}\text{Post-test}_i = \beta_1 &+ \beta_2\text{Pre-test}_i + \beta_3\text{HSrank}_i + \beta_4\text{MathSAT}_i + \beta_5\text{VerbSAT}_i \\ &+ \beta_6\text{NumAbsences} + \beta_7\text{GroupWork}_i + \beta_8\text{Clickers}_i \\ &+ \beta_9\text{Baldwin}_i + \beta_{10}\text{Campbell}_i + \beta_{11}\text{Daniels}_i \\ &+ \epsilon_i, \text{ for } i = 1, \ldots, 1000.\end{aligned}$$

- How can we test that there is not a statistically significant association between Highschool and Post-test?
-

$$\text{H}_0 : \beta_9 = \beta_{10} = \beta_{11} = 0 \text{ vs. H : At least one of } \beta_9, \beta_{10}, \beta_{11} \text{ is nonzero.}$$

- Consider a multiple regression model

$$Y = X\beta + \epsilon.$$

- Model satisfies assumptions of Chapter 4.
- $\epsilon \sim N(0, \sigma^2 I)$
- Let p_0 be an integer such that $1 \leq p_0 \leq p$, and consider the testing problem

$$\begin{aligned} &\mathrm{H}_0 : \beta_{p-p_0+1} = \cdots = \beta_p = 0 \text{ vs.} \\ &\mathrm{H} : \text{At least one of } \beta_{p-p_0+1}, \ldots, \beta_p \text{ is nonzero} \end{aligned}$$

$$\begin{aligned} &\mathrm{H}_0 : \text{The last } p_0 \text{ coefficients } \beta_j \text{ are all zero} \\ &\mathrm{H} : \text{At least one of the last } p_0 \text{ coefficients } \beta_j \text{ is nonzero} \end{aligned}$$

$$\begin{aligned}
\mathrm{H}_0 &: \text{The last } p_0 \text{ coefficients } \beta_j \text{ are all zero} \\
\mathrm{H} &: \text{At least one of the last } p_0 \text{ coefficients } \beta_j \text{ is nonzero}
\end{aligned}$$

- Let $X^{(s)}$ be the design matrix with the last p_0 columns removed ($X^{(s)}$ is $n \times (p - p_0)$).
- Under H_0, the model is

$$Y = X^{(s)}\beta^{(s)} + \epsilon^{(s)}$$

- $\beta^{(s)} \in \mathbb{R}^{p-p_0}$
- $\epsilon^{(s)} \sim N(0, \sigma^2 I)$
- Original model:

$$Y = X\beta + \epsilon$$

The F-statistic

- Fit both models
- $\hat{\beta} = (X'X)^{-1}X'Y$ and $\hat{Y} = X\hat{\beta}$.
- $\hat{\beta}^{(s)} = (X^{(s)'}X^{(s)})^{-1}X^{(s)'}Y$ and $\hat{Y}^{(s)} = X^{(s)'}\hat{\beta}^{(s)}$.
- The F-statistic for the testing problem is

$$F = \frac{(\|\hat{Y}\|^2 - \|\hat{Y}^{(s)}\|^2)/p_0}{\|e\|^2/(n-p)}.$$

- Book's definition:

$$F = \frac{(\|X\hat{\beta}\|^2 - \|X\hat{\beta}^{(s)}\|^2)/p_0}{\|e\|^2/(n-p)}.$$

Other Forms

- $$F = \frac{(\|e^{(s)}\|^2 - \|e\|^2)/p_0}{\|e\|^2/(n-p)}.$$

- $$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p_1}\right)}{\left(\frac{RSS_2}{n - p_2}\right)}.$$

The F-test

Theorem

For the testing problem described in this section, under H_0

- $U = \|X\hat{\beta}\|^2 - \|X^{(s)}\hat{\beta}^{(s)}\|^2 \sim \sigma^2\chi^2_{p_0}$
- $V = \|e\|^2 \sim \sigma^2\chi^2_{n-p}$
- $U \perp\!\!\!\perp V$
- $F \sim F(p_0, n-p)$*, that is,* F *has Fisher's* F*-distribution with* p_0 *degrees of freedom in the numerator and* $n-p$ *degrees of freedom in the denominator.*
- *Decision rule: Reject* H_0 *if* $|F| > F_\alpha(p_0, n-p)$*.*

More General F-test

Theorem

- *Consider a multiple regression model $Y = X\beta + \epsilon$ satisfying the assumptions of Chapter 4, and assume that ϵ is normally distributed.*
- *Let $V_0 \leq V \leq \mathbb{R}^p$, and consider the testing problem*

$$H_0 : \beta \in V_0 \text{ vs. } H : \beta \in V \setminus V_0.$$

- *Let e_0 and e be the residual vectors under H_0 and H, respectively.*
- *Then the test statistic*

$$F = \frac{(\|e_0\|^2 - \|e\|^2)/(\dim(V) - \dim(V_0))}{\|e\|^2/(n - \dim(V))}$$

  has an $F(\dim(V) - \dim(V_0), n - \dim(V))$ distribution.
- *We reject H_0 if $|F| > F_\alpha(\dim(V) - \dim(V_0), n - \dim(V))$.*

Example using Lab 2 Data

- Model from Lab 2

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \text{ for } i = 1, \ldots, 100.$$

- Here, $n = 100$ and $p = 4$.
- Consider the testing problem

$$\text{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. H} : \text{ At least one of } \beta_1, \beta_2, \beta_3 \text{ is nonzero}$$

- $p_0 = 3$
- $F = 195.5$

```
Console ~/
> mymodel=lm(Y~X1+X2+X3)
> summary(mymodel)

Call:
lm(formula = Y ~ X1 + X2 + X3)

Residuals:
     Min       1Q   Median       3Q      Max 
-23.5493  -6.4823   0.7492   5.5936  24.9199 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 50.04082    2.94308  17.003  < 2e-16 ***
X1           4.43642    0.31869  13.921  < 2e-16 ***
X2          21.64408    3.12069   6.936  4.7e-10 ***
X3          -0.55334    0.03271 -16.919  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.463 on 96 degrees of freedom
Multiple R-squared: 0.8593, Adjusted R-squared: 0.855 
F-statistic: 195.5 on 3 and 96 DF,  p-value: < 2.2e-16

> names(summary(mymodel))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"   
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled" 
> summary(mymodel)$r.squared
[1] 0.8593479
>
```

Interpreting t and F-tests

- Significant t and F-statistics do not prove the model is valid.
- They *assume* the model is valid.
- If the model is valid and the test statistic is significant, there is strong evidence that the null hypothesis is false.

Outline

1 Section 5.6: Normal Theory

2 Section 5.7: The F-test

3 Chapter 5 Closing Remarks

Asymptotics

Under certain regularity conditions (see endnotes for Chapter 4),

- $\hat{\beta}$ is asymptotically normal.
- Asymptotically, under the null hypothesis, the t and F-statistics have the distributions given in Chapter 5.

Other Issues

- Data snooping
- Replication
- Cross validation
- Scientific Method
- We need methods for model assessment (diagnostics).