# Stats 401 Lab 12

Sanjana Gupta

3/15/2018 and 11/30/2018

# Outline

- F tests and ANOVA
- Goodness of fit
  - R squared
  - Adjusted R Squared
- Model selection
- Lab Ticket

# F Test/ANOVA Motivation

Why perform the F test?

- Want to know if additional variables are statistically significant.

How is this different from looking at the regression output?

- Regression output is testing each $b_i$ with all other $b_j$ fixed
- F test/ANOVA lets us test multiple variables for significance at once

# Hypothesis Test outline

F-Test corresponds to a **hypothesis test**.

Recall from *STATS250*, a hypothesis test has the following steps:

- Establish the Null and Alternate hypothesis ($H_0, H_a$)
- Find a test statistic (F-Statistic)
- Find the p-value
  *Prob(observing something as or more extreme than your test-stat)*
- Conclusion: reject/ fail to reject Null hypothesis

# F Test outline: Nested Models

**Establish Hypothesis**

- Let $H_0$ be the base linear model $\mathbf{Y} = \mathbb{X}\beta + \boldsymbol{\epsilon}$
- $H_a$ extends $H_0$ by adding $d$ additional variables, i.e $\mathbf{Y} = \mathbb{X}_\alpha\beta + \boldsymbol{\epsilon}$

So the Null hypothesis is that the smaller (base) model is better, whereas the alternate hypothesis is that the additional variables being considered are important and should be included in the model.

- $H_0 : \mathbf{Y} = \mathbb{X}\beta + \epsilon, \quad \text{dimension}(\mathbb{X}) = n \times q$
- $H_a : \mathbf{Y} = \mathbb{X}_a\beta_{\boldsymbol{a}} + \boldsymbol{\epsilon}, \quad \mathbb{X}_a = [\mathbb{X} \;\; \mathbb{Z}]$ where $\mathbb{Z}$ is the matrix of d additional variables

# F Test outline: Nested Models

**Get Test-Statistic**

$$f = \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)}$$

- $RSS_0$ and $RSS_a$ are the residual sum of squares for the null and alternative models
- $d$ is the difference in the degrees of freedom between the two models
- $n - q$ is the degrees of freedom in the alternative model

**Note:**

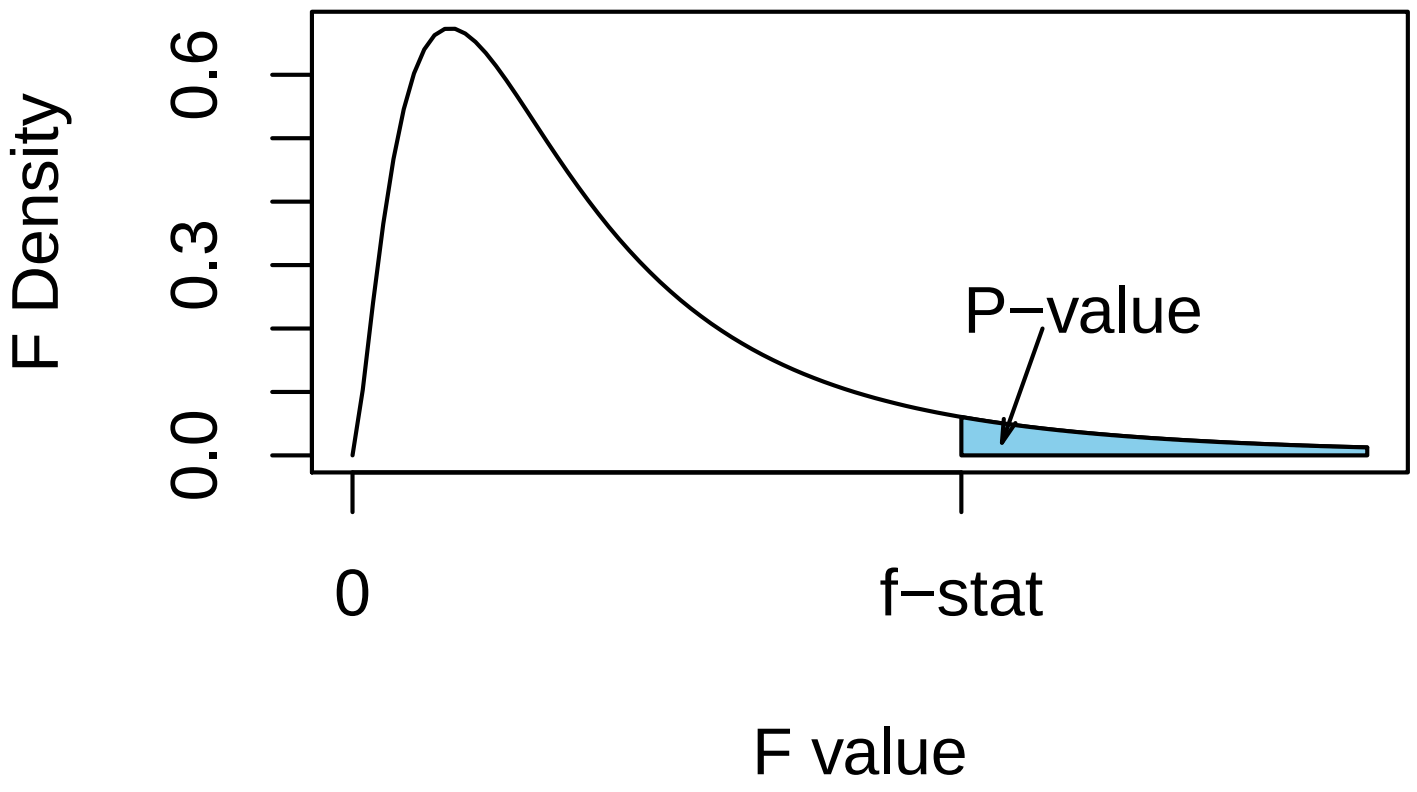
- Under $H_0$, the model generated F-statistic has an F distribution with $d$ and %n-q$ degrees of freedom.
- If $H_0$ is true, then $f \approx 1$. So large values of $f$ are evidence against the null.

# F Test outline: Nested Models

**Get p-value**

Let $F$ be a random variable that follows the *F-distribution* with degrees of freedom $d, n - q$

p-value $= \mathrm{P}(F > f)$



# F Test outline: Nested Models

**Conclusion**

- Rule: Reject Null hypothesis if the p-value $<$ significance level ($\alpha = 0.01, 0.05, 0.1$)
- Interpretation: We pick the linear model in the alternate hypothesis, which is the model containing the additional variables. Hence we conclude that the additional variables are significant.

# F Test in R: Null mean model

This is the F-Test corresponding to the `lm()` output for the regression model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

- $H_0 : \beta = 0$, i.e. all coefficients are zero
- $H_a$ : alteast one of the coefficients is non-zero

Mathematically, this corresponds to the following:

- $H_0 : Y = \beta_0 + \epsilon$
- $H_a : \mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

$$f = \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)}$$

- $q =$ Number of columns in $\mathbb{X}$
- $d =$ Number of covariates (intercept not included)
  $=$ No of columns in $\mathbb{X} - 1 = q - 1$
- $n =$ Number of observations $=$ No of rows in $\mathbb{X}$

# Example 1: Null mean model

Let us calculate the F-Statistic by hand and compare with the `lm()` output. Recall the iris data

```
data(iris)
iris <- iris[,-2] #Remove the second column of original dataset
head(iris)
```

```
##   Sepal.Length Petal.Length Petal.Width Species
## 1          5.1          1.4         0.2  setosa
## 2          4.9          1.4         0.2  setosa
## 3          4.7          1.3         0.2  setosa
## 4          4.6          1.5         0.2  setosa
## 5          5.0          1.4         0.2  setosa
## 6          5.4          1.7         0.4  setosa
```

# Example 1: Examining F-Test in R

Consider the following linear model

```
lm1 <- lm(Petal.Length~Petal.Width, data=iris)
summary(lm1)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33542 -0.30347 -0.02955  0.25776  1.39453
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08356    0.07297   14.85   <2e-16 ***
## Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.4782 on 148 degrees of freedom
```

# Example 1: Calculating the F-Statistic by hand

We know that $f = \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)}$

Let us find the relevant values:

```
RSS_a <- sum(residuals(lm1)^2)
RSS_0 <- sum((iris$Petal.Length - mean(iris$Petal.Length))**2)
#or
lm0 <- lm(Petal.Length ~ 1, data=iris) #Linear model with interc
RSS_0 <- sum(residuals(lm0)^2)

cat("RSS_0:", RSS_0, " ; RSS_a:", RSS_a, " ; n:",
    nrow(model.matrix(lm1)), " ; q:", ncol(model.matrix(lm1)))
```

```
## RSS_0: 464.3254  ; RSS_a: 33.84475  ; n: 150  ; q: 2
```

# Example 1: Calculating the F-Statistic by hand

```
## RSS_0: 464.3254  ; RSS_a: 33.84475  ; n: 150  ; q: 2
```

So,

$$
\begin{aligned}
f &= \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)} \\
&= \frac{(464.3254 - 33.84475)/1}{33.84475/148} \\
&= \frac{430.4807}{0.2286807} \\
&= 1882.453
\end{aligned}
$$

From the output, the F-Statistic is 1882.

# Example 1: Calculating the p-value by hand

p-value is

$$\text{probability}(F > f), \text{ where } F \sim F_{1,148}$$

```
pval = pf(1882.453,1,148,lower.tail = FALSE)
```

From the output, the p-value is $<$ 2.2e-16 (which holds)

**Conclusion**

Since the p-value $<$ 0.05 (and 0.01), assuming our significance level us 1%, we reject the null hypothesis and conclude that atleast one of the coefficients is non-zero. Thus, we pick the linear model with the petal width.

# Lab Activity 1: Null mean model in R

Recall the GPA dataset from Hw10 Q1

```
gpa <- read.table("https://ionides.github.io/401f18/hw/hw10/gpa.
head(gpa)
```

```
##   ID  GPA High_School ACT Year
## 1  1 0.98           61  20 1996
## 2  2 1.13           84  20 1996
## 3  3 1.25           74  19 1996
## 4  4 1.32           95  23 1996
## 5  5 1.48           77  28 1996
## 6  6 1.57           47  23 1996
```

# Lab Activity 1: Null mean model in R

We fit the following linear model

```
lm_gpa <- lm(GPA~High_School+ACT+factor(Year),data=gpa)
```

-Fit the model in R
- Write the null and alt hypothesis for the F-test performed in the `lm()` summary
- Identify the F-statistic and P-value in the output
- Compare this by manually calculating the p-value
- State your conclusion (based on this, which variables would you consider including in your linear model)

# Lab Activity 1: Null mean model in R

We fit the following linear model
`lm_gpa <- lm(GPA~High_School+ACT+factor(Year),data=gpa)`

- Fit the model in R

```
lm_gpa <- lm(GPA~High_School+ACT+factor(Year),data=gpa)
```

- Write the null and alt hypothesis for the F-test performed in the `lm()` summary
  - $H_0$ : GPA $= \beta_0(\text{constant}) + \epsilon$
  - $H_a$ : GPA $= \beta_0 + \beta_1 \text{High\_School} + \beta_2 \text{ACT} + \beta_3 \times (1997) + \beta_4 \times (1998) + \beta_5 \times (1999) + \beta_6 \times (2000) + \epsilon$
    where atleast some $\beta_i \neq 0$

# Lab Activity 1: Null mean model in R

```
summary(lm_gpa)
```

```
##
## Call:
## lm(formula = GPA ~ High_School + ACT + factor(Year), data = g
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15048 -0.28873  0.07655  0.39619  1.30415
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.217874   0.144598   8.422  < 2e-16 ***
## High_School        0.010124   0.001285   7.878 1.28e-14 ***
## ACT                0.037188   0.005951   6.248 7.21e-10 ***
## factor(Year)1997 0.083657   0.068816   1.216   0.2245
## factor(Year)1998 0.115339   0.066158   1.743   0.0817 .
## factor(Year)1999 0.080071   0.067475   1.187   0.2358
## factor(Year)2000 0.056007   0.068013   0.823   0.4105
## ---
```

# Lab Activity 1: Null mean model in R

- Compare this by manually calculating the p-value

```
RSS_a <- sum(residuals(lm_gpa)^2)
RSS_0 <- sum((gpa$GPA - mean(gpa$GPA))**2)
```

```
## RSS_0: 283.4484  ; RSS_a: 224.7415  ; n: 705  ; q: 7
```

$$f = \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)} = \frac{(283.4484 - 224.7415)/6}{224.7415/698} = \frac{9.784483}{0.3219792} = 30.38856$$

p-value: is the $P(F > f)$ where $F \sim F_{6,698}$

```
pf(30.38856,6,698,lower.tail = FALSE)
```

```
## [1] 1.795986e-32
```

# Lab Activity 1: Null mean model in R

- Conclusion: based on this, which variables would you consider including in your linear model?
  - Since the p-value is extremely small, we reject the null hypothesis and conclude that atleast some of the covariates are important
  - Looking at the summary table of `lm_gpa`, we would include High_School,ACT.
    Do you agree? Why/ why not?

# Lab Acitivity 2: Anova

For the `iris` data in Lab Activity 1, we saw that we should include `Petal.Width` while modelling `Petal.Length`. Let us evaluate whether to include `species` or not.

Let $H_0$ be the model consisting of only the `Petal.Width` and let $H_a$ be the extended model that includes `Petal.Width` as well as `species`.

- Write the Null and Alt hypothesis for this test and fit it in R
- Compute the F-statistic (by hand) for the model mentioned above. (Use the output from the fitted models above)
- Compare this with the `anova()` output
- Find the p-value and draw your conclusion

# Lab Acitivity 2: Anova (Establishing the hypothesis)

- $H_0 : Petal.Length = \beta_1 \times (Petal.Width) + \beta_0 + \epsilon$
- $H_a : Petal.Length = \beta_1 \times Petal.Width + \beta_2 \times Species_versicolor + \beta_3 \times Species_erginica + \beta_0 + \epsilon$

Fitting the models in R

```
lm_iris0 <- lm(Petal.Length~Petal.Width, data=iris)
lm_iris1 <- lm(Petal.Length~Petal.Width+Species, data=iris)
```

# Lab Acitivity 2: Anova (F-Statistic by hand)

We know that $f = \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)}$

```
RSS_0 <- sum(residuals(lm_iris0)^2)
RSS_a <- sum(residuals(lm_iris1)^2)
n <- nrow(model.matrix(lm_iris1))
q <- ncol(model.matrix(lm_iris1))
d <- ncol(model.matrix(lm_iris1)) - ncol(model.matrix(lm_iris0))
cat("RSS_0:",RSS_0,"; RSS_a:",RSS_a,"; d:",d,"; n-q:",n-q)
```

```
## RSS_0: 33.84475 ; RSS_a: 20.83344 ; d: 2 ; n-q: 146
```

Plugging this into the formula, we have

$$f = \frac{(33.84475 - 20.83344)/2}{20.83344/146} = \frac{13.01131/2}{0.1426948} = \frac{6.505655}{0.1426948} = 45.5914$$

# Lab Acitivity 2: Anova (comparing with R output)

From the previous slide,

```
## RSS_0: 33.84 ; RSS_a: 20.83 ; d: 2 ; n-q: 146
```

$$f = \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)} = \frac{(33.84-20.83)/2}{20.83/146} = \frac{13.01/2}{0.14} = \frac{6.51}{0.14} = 45.591$$

Compare with R output:

```
anova(lm_iris1)
```

```
## Analysis of Variance Table
##
## Response: Petal.Length
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## Petal.Width   1 430.48  430.48 3016.792 < 2.2e-16 ***
## Species       2  13.01    6.51   45.591 4.137e-16 ***
## Residuals   146  20.83    0.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

# Lab Acitivity 2: Anova (p-value)

p-value is

$$\text{probability}(F > f), \text{ where } F \sim F_{2,146}$$

```
pval = pf(45.591,2,146,lower.tail = FALSE); pval
```

```
## [1] 4.138018e-16
```

Compare with R output:

```
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: Petal.Length
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Petal.Width   1 430.48  430.48  1882.5 < 2.2e-16 ***
## Residuals   148  33.84    0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

# Lab Acitivity 2: Anova (conclusion)

Since our p-value is very small ($<0.01$), we reject the null hypothesis and pick the model corresponding to the alternate hypothesis. That is, we chose to include `species` in our model.

# Note on Anova

Anova checks for additional variables sequentially. That is,

- for the linear model `y~a+b`
  $H_0$ : `y~a` and $H_a$ : `y~a+b`
  i.e. checking whether to include additional variable `b`
- for the linear model `y~b+a`
  $H_0$ : `y~b` and $H_a$ : `y~b+a`
  i.e. checking whether to include additional variable `a`
- for the linear model `y~a+b+c` there will be two tests:
  - $H_0$ : `y~a` $H_a$ : `y~a+b`
    i.e. checking whether to include additional variable `b`
  - $H_0$ : `y~a+b` $H_a$ : `y~a+b+c`
    i.e. checking whether to include additional variable `c`

# F-Test and T-test

When we evaluate the importance of a single variable (i.e. when $d = 1$ in the F-test), then the F-test is equivalent to the t-test. That is, if $T \sim T_d$ and $F \sim F_{1,d}$, then $T^2$ has the same distribution as $F$. i.e. $T_d^2 \stackrel{d}{=} F_{1,d}$

Check using R

```
df=10 # Fix degrees of freedom of t distribution

for(x in c(1,5,16,25)){
  print(c(pf(x,1,df), pt(sqrt(x),df)-pt(-sqrt(x),df)))}
```

```
## [1] 0.6591069 0.6591069
## [1] 0.9506678 0.9506678
## [1] 0.9974817 0.9974817
## [1] 0.9994627 0.9994627
```

Since,

pf(x,1,df2=df) $=$ P(F<x) $=$ P($T^2$<x) $=$ P(-x $<$ T $<$ x)

$=$ P(T $<$ x) - P(T$<$ x) $=$ pf(x,df) - pf(x,df)

# Goodness of fit

This describes how well a model fits the data. We have seen the following methods:

- F tests
- R Squared
- Adj R Squared

# R-Squared Statistic

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

where RSS = Residual sum of squares $= \sum_{i=1}^{n} y_i - \hat{y}_i$

TSS = Total sum of squares $= \sum_{i=1}^{n} y_i - \bar{y}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- $R^2$ is the square of the correlation between the data and the fitted values.
- It is sometimes described as the fraction of the variation in the data explained by the linear model.
- This compares the residual sum of squares under the full model with the residual sum of squares under a model with a constant mean.

# Drawbacks of R-Squared

- Higher R-Squared is better
- Note that R-Squared always increases upon increasing the number of variables in the model. So, it will always select bigger models.
- One way to penalize $R^2$ is by using Adjusted $R^2$ instead

# Adjusted R-Squared

$$R^2_{adj} = 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)}$$

- Dividing by degrees of freedom is similar to F-test

**$R^2$ vs F-Test**

- Recall: The F test compares a full model with a model that omits specific selected explanatory variables.
- F Test can only be applied when we have nested models.
- When the models are not nested, we can compare them using $R^2_{adj}$ instead.

# Model Selection

- Model selection is the problem of selecting the best model from a group of candidate models, given data.
- Model selection techniques try to balance the *goodness of fit* and *complexity* of the candidate models.
- In general, increasing complexity (number of variables) increases the variance of the model which is not desired. Hence, we want to find the smallest model which best fits our data.
- Goodness of fit measures (such as $R^2_{adj}$, AIC, BIC) are tools to help us find the best model for our data.

# Exit ticket

- [Link between `lm()` and `anova()`] How can you get the F-test that is being done in the `lm()` output in Lab-Activity 1 using `anova()`? (*Hint: What are the null and alt linear models in the LB1 F-test? The input of the anova() function is always the bigger linear model - which is the alternate model*)
- Assume you fit the linear model `y~a+b+c`, i.e. the linear model here is $y = \beta_1 \times a + \beta_2 \times b + \beta_3 \times c + \epsilon$ where $y$ is the outcome variable and $a, b, c$ are covariates.
  - Write down the null and alt hypothesis corresponding to the F-statistic in the `lm(y~a+b+c)` output
  - Write down the null and alt hypotheses corresponding to the F-statistics in the `anova(y~a+b+c)` output