

Lecture 11:
F-test
STAT 632, Spring 2020

Test of all the predictors

Is there a relationship between the response variable and at least one predictor in the multiple linear regression model?

$$Y = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p + e$$

The null and alternative hypothesis for the test can be written as

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

There is no relationship between Y and the predictor variables.

$$H_A: \text{at least one } \beta_j \neq 0$$

There is a relationship between Y and at least one of the predictor variables.

Test of all the predictors

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

H_A : at least one $\beta_j \neq 0$

Test statistic:

$$F = \frac{(\text{SST} - \text{RSS})/p}{\text{RSS}/(n - p - 1)} = \frac{\text{SSreg}/p}{\text{RSS}/(n - p - 1)}$$

- ▶ $\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$ is the total sum of squares
- ▶ $\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ is the residual sum of squares
- ▶ $\text{SSreg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ is the regression sum of squares
- ▶ $\text{SST} = \text{SSreg} + \text{RSS}$
- ▶ n is the number of observations in the data set, and p is the number of predictor variables.

Test of all the predictors

Analysis of variance table:

| Source | df | Sum of Squares | Mean Square | F |
|------------|-------------|----------------|--------------------|-------------------------------|
| Regression | p | SSreg | SSreg/ p | $\frac{SSreg/p}{RSS/(n-p-1)}$ |
| Residual | $n - p - 1$ | RSS | RSS/ $(n - p - 1)$ | |
| Total | $n - 1$ | SST | | |

Most computer packages will provide some version of this table. Ronald Fisher, the creator of the table, once said it's "nothing but a convenient way of arranging arithmetic." That was in 1931, when he had to do all the calculations by hand.

Example: NY Housing Data

- ▶ Data set on housing prices from Canton, NY (scraped from Zillow.com)
- ▶ The response variable is Price (in thousands of dollars)
- ▶ The predictors
 - ▶ Beds: number of bedrooms
 - ▶ Baths: number of bathrooms
 - ▶ Size: floor area of house (in thousands of square feet)
 - ▶ Lot: size of the lot (in acres)

```
> library(Stat2Data)
```

```
> data(HousesNY)
```

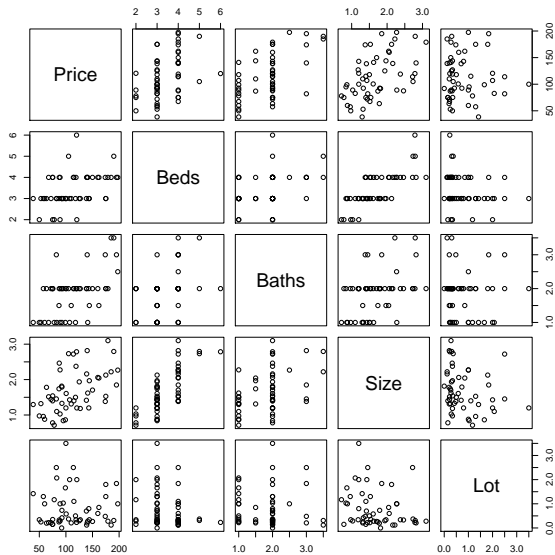
```
> dim(HousesNY)
```

```
[1] 53  5
```

```
> head(HousesNY)
```

| | Price | Beds | Baths | Size | Lot |
|---|-------|------|-------|-------|------|
| 1 | 57.6 | 3 | 2 | 0.960 | 1.30 |
| 2 | 120.0 | 6 | 2 | 2.786 | 0.23 |
| 3 | 150.0 | 4 | 2 | 1.704 | 0.27 |
| 4 | 143.0 | 3 | 2 | 1.200 | 0.80 |
| 5 | 92.5 | 3 | 1 | 1.329 | 0.42 |
| 6 | 50.0 | 2 | 1 | 0.974 | 0.34 |

```
> pairs(Price ~ Beds + Baths + Size + Lot, data=HousesNY)
```



Example

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_A : \text{at least one } \beta_j \neq 0$$

For the F-test we are comparing the full model, with all the predictors, to the null model, with no predictors.

Full model:

$$\text{Price} = \beta_0 + \beta_1 \text{Beds} + \beta_2 \text{Baths} + \beta_3 \text{Size} + \beta_4 \text{Lot} + e$$

Null (reduced) model:

$$\text{Price} = \beta_0 + e$$


```
> lm_full <- lm(Price ~ Beds + Baths + Size + Lot, data=HousesNY)
> lm_null <- lm(Price ~ 1, data=HousesNY)
```

```
> anova(lm_null, lm_full)
Analysis of Variance Table
```

Model 1: Price ~ 1

Model 2: Price ~ Beds + Baths + Size + Lot

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|--------|-------|----|-----------|--------|--------------|
| 1 | 52 | 89255 | | | | |
| 2 | 48 | 52358 | 4 | 36897 | 8.4566 | 3.01e-05 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Since the p -value < 0.001 we reject the null hypothesis that $\beta_1 = \dots = \beta_4 = 0$. Thus, we conclude, that at least one predictor is associated with Price.

The results of the F-test are also provided in the `summary()` output.

```
> summary(lm_full)
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|-----------|
| (Intercept) | 14.590 | 23.266 | 0.627 | 0.5336 |
| Beds | 2.771 | 8.730 | 0.317 | 0.7523 |
| Baths | 26.238 | 7.844 | 3.345 | 0.0016 ** |
| Size | 22.155 | 11.931 | 1.857 | 0.0695 . |
| Lot | 4.621 | 6.184 | 0.747 | 0.4585 |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.03 on 48 degrees of freedom

Multiple R-squared: 0.4134, Adjusted R-squared: 0.3645

F-statistic: 8.457 on 4 and 48 DF, p-value: 3.01e-05

Just for verification, we can also directly calculate the F-test statistic using the formula.

```
> n <- nrow(HousesNY)
> p <- 4
> rss <- sum(resid(lm_full)^2); rss
[1] 52357.9
> sst <- sum(resid(lm_null)^2); sst
[1] 89255.4
> fstat <- ((sst - rss) / p) / (rss / (n-p-1))
> fstat
[1] 8.456603
> 1 - pf(fstat, df1=p, df2=n-p-1)
[1] 3.0104e-05
```

Testing just one predictor

Can one particular predictor be dropped from the model?

To test whether the coefficient for a single predictor is 0 we can either use a t-test or F-test (the results are equivalent). In R, the t-test is less work, since the results are provided in the regression summary output.

$$H_0 : \beta_j = 0$$

$$H_A : \beta_j \neq 0$$

Test statistic:

$$T_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}; \quad \text{df} = n - p - 1$$

Example

The regression summary in a previous slide shows that Beds is not significant, and can be dropped from the model, since $t = 0.317$ with a p -value = 0.7523. Using the F-test we obtain the same result:

```
> lm_full <- lm(Price ~ Beds + Baths + Size + Lot, data=HousesNY)
> lm1 <- lm(Price ~ Baths + Size + Lot, data=HousesNY)
> anova(lm1, lm_full)
```

Analysis of Variance Table

Model 1: Price ~ Baths + Size + Lot

Model 2: Price ~ Beds + Baths + Size + Lot

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|--------|-------|----|-----------|--------|--------|
| 1 | 49 | 52468 | | | | |
| 2 | 48 | 52358 | 1 | 109.87 | 0.1007 | 0.7523 |

Notice that the p -values from the F-test and t-test are exactly the same.

Testing a subset of predictors

Suppose we want to test whether a specified subset of predictors have regression coefficients equal to 0. This is often called the **partial F-test**.

For example, using the NY housing data, the full model is given by

$$\text{Price} = \beta_0 + \beta_1 \text{Beds} + \beta_2 \text{Baths} + \beta_3 \text{Size} + \beta_4 \text{Lot} + e$$

Suppose we want to test whether the coefficients for Beds and Lot are both zero. The null and alternative hypothesis can be written as:

$$H_0: \beta_1 = \beta_4 = 0$$

$$H_A: \beta_1 \neq 0 \text{ or } \beta_4 \neq 0$$

Testing a subset of predictors

For the partial F-test we use the following test statistic:

$$F = \frac{(\text{RSS}_{\text{reduced}} - \text{RSS}_{\text{full}})/k}{\text{RSS}_{\text{full}}/(n - p - 1)}$$

- ▶ RSS_{full} is the residuals sum of squares for the model with the full set of p predictors.
- ▶ $\text{RSS}_{\text{reduced}}$ is the residuals sum of squares for the reduced model with k predictors removed.

Example

```
> lm_full <- lm(Price ~ Beds + Baths + Size + Lot, data=HousesNY)
> lm2 <- lm(Price ~ Baths + Size, data=HousesNY)
> anova(lm2, lm_full)
```

Analysis of Variance Table

Model 1: Price ~ Baths + Size

Model 2: Price ~ Beds + Baths + Size + Lot

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|--------|-------|----|-----------|--------|--------|
| 1 | 50 | 53039 | | | | |
| 2 | 48 | 52358 | 2 | 680.81 | 0.3121 | 0.7334 |

The p -value = 0.73 is large, so we do not reject the null hypothesis that $H_0 : \beta_1 = \beta_4 = 0$. So we can remove both predictors, Beds and Lot, from the model.

The R^2 for the full and reduced models are about the same, and the adjusted R^2 for the reduced model is a little higher. This agrees with the conclusion of the F-test. So the adjusted- R^2 also indicates that we can remove Beds and Lot.

```
> s1 <- summary(lm_full)
> s2 <- summary(lm2)
>
> s1$r.squared
[1] 0.4133924
> s2$r.squared
[1] 0.4057647
>
> s1$adj.r.squared
[1] 0.3645084
> s2$adj.r.squared
[1] 0.3819953
```

```
> summary(lm2)
```

```
Coefficients:
```

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|------------|
| (Intercept) | 24.641 | 15.890 | 1.551 | 0.12728 |
| Baths | 26.755 | 7.699 | 3.475 | 0.00107 ** |
| Size | 23.399 | 8.317 | 2.813 | 0.00699 ** |

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 32.57 on 50 degrees of freedom
```

```
Multiple R-squared:  0.4058, Adjusted R-squared:  0.382
```

```
F-statistic: 17.07 on 2 and 50 DF,  p-value: 2.233e-06
```

Your turn

Using the NY housing data, the full model is given by

$$\text{Price} = \beta_0 + \beta_1 \text{Beds} + \beta_2 \text{Baths} + \beta_3 \text{Size} + \beta_4 \text{Lot} + e$$

In R, conduct a partial F-test for the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \beta_4 = 0$$

$$H_A : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \beta_4 \neq 0$$

What is the p -value and your conclusion?

Summary

- ▶ Use the overall F-test to test whether there is a relationship between the response and at least one predictor in the model.
- ▶ Use the partial F-test to test whether there is a relationship between the response and a specified subset of two or more predictors in the model.
- ▶ Use a t-test to test whether there is a relationship between the response and a single predictor in the model.