Topic 3b

Analysis of variance (ANOVA) approach to regression analysis

Learning objectives

- Apply ANOVA … an (alternative) approach to testing for a linear association
- Know when to use the t-test and the F-test
- Understand and interpret regression output from software e.g. Stata

The basic idea

- Break down the variation in Y ("**total sum of squares**") into two components:
  - a component that is "due to" the change in X ("**regression sum of squares**")
  - a component that is just due to random error ("**error sum of squares**")
- If the regression sum of squares is a large component of the total sum of squares, it suggests that there is a linear association.

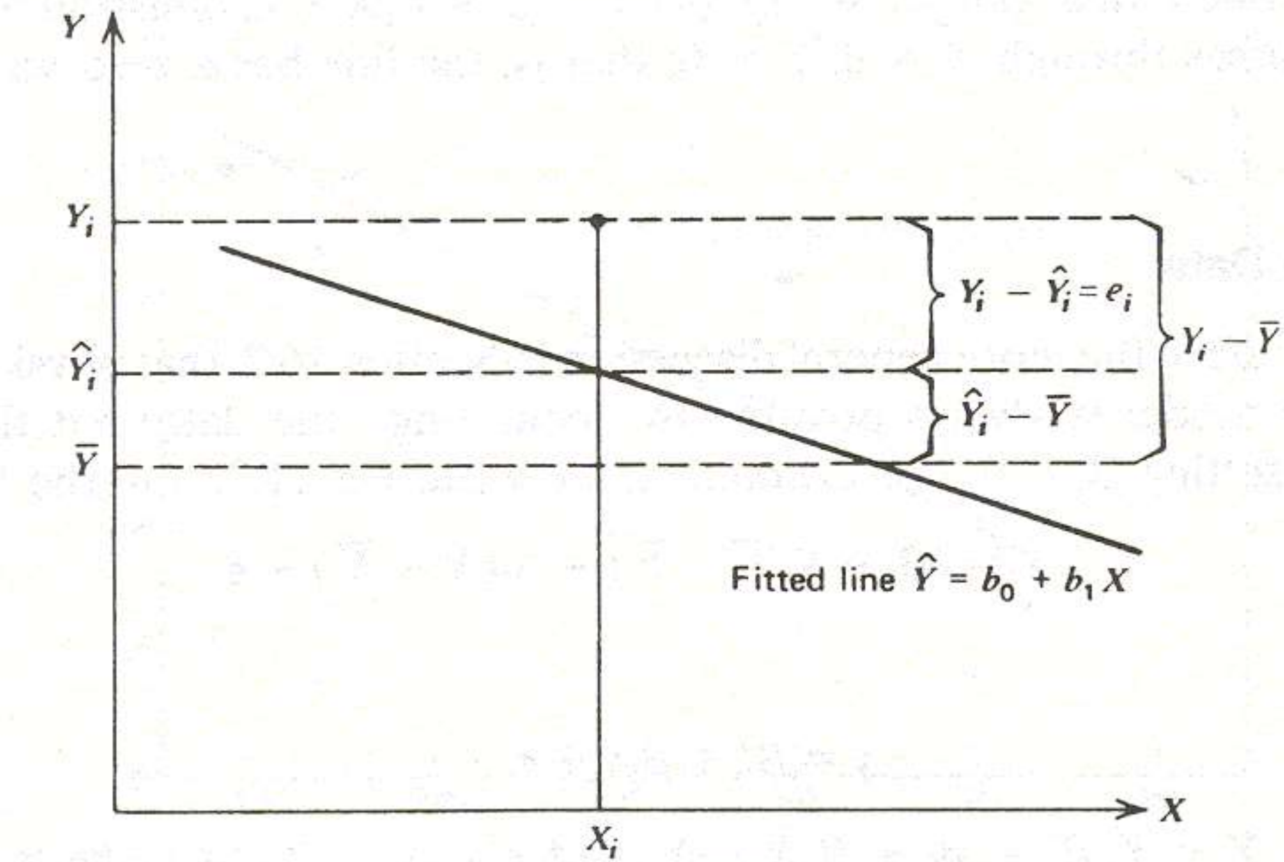


Figure 1.7. Geometrical meaning of the identity (1.3.1).

$$\boxed{\left(Y_i - \bar{Y}\right) = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)}$$

The above decomposition holds for the sum of the squared deviations, too:

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Total sum of squares (SST)

Regression sum of squares (SSR)

Error sum of squares (SSE)

SST = SSR + SSE

Breakdown of degrees of freedom

Degrees of freedom associated with SST

$$(n-1) = (k) + (n \; - k \; - 1)$$

Degrees of freedom associated with SSR

Degrees of freedom associated with SSE

Analysis of Variance (ANOVA) Table

TABLE 2.2 ANOVA Table for Linear Regression.

Source of Variation	SS	df	MS	$E\{MS\}$
Regression	$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$	k	$MSR = \frac{SSR}{k}$	$\sigma^2 + \beta_1^2 \Sigma(X_i - \bar{X})^2$
Error	$SSE = \Sigma(Y_i - \hat{Y}_i)^2$	$n\text{-}k\text{-}1$	$MSE = \frac{SSE}{n\text{-}k\text{-}1}$	σ^2
Total	$SSTO = \Sigma(Y_i - \bar{Y})^2$	$n - 1$		

Example: Mortality and Latitude

```
The regression equation is Mort = 389 - 5.98 Lat

Predictor        Coef     SE Coef         T        P
Constant       389.19       23.81     16.34    0.000
Lat           -5.9776      0.5984     -9.99    0.000

S = 19.12        R-Sq = 68.0%     R-Sq(adj) = 67.3%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			

How to find n?

- Recall the degrees of freedom?

$$(n-1) = (k) + (n-k-1)$$

Definitions of Mean Squares

We already know the **mean square error** (**MSE**) is defined as:

$$\boxed{MSE = \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-k-1} = \frac{SSE}{n-k-1}}$$

For a simple regression k=1 such that:

$$MSE = \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{SSE}{n-2}$$

Similarly, the **regression mean square** (**MSR**) is defined as:

$$\boxed{MSR = \frac{\sum(\hat{Y}_i - \ \bar{Y}_i)^2}{k} = \frac{SSR}{k}}$$

R- Squared

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Let us check from the Mortality and Latitude example!
- Latitude explains 68% of the variation in mortality. 32% remains unexplained – Has to always sum up to 100.

Adjusted-R^2

- It is adjusted based on the degrees of freedom (df)
- Relevant in multiple regression
- Adjusted R^2 can actually get smaller as additional variables are added to the model.
- As N gets bigger, the difference between R^2 and Adjusted R^2 gets smaller and smaller.

$$R^2_{adj} = 1 - (1 - R^2)\,\frac{\mathrm{n} - 1}{n - k - 1}$$

The formal F-test for slope parameter β_1

Null hypothesis H_0: $\beta_1 = 0$
Alternative hypothesis H_A: $\beta_1 \neq 0$

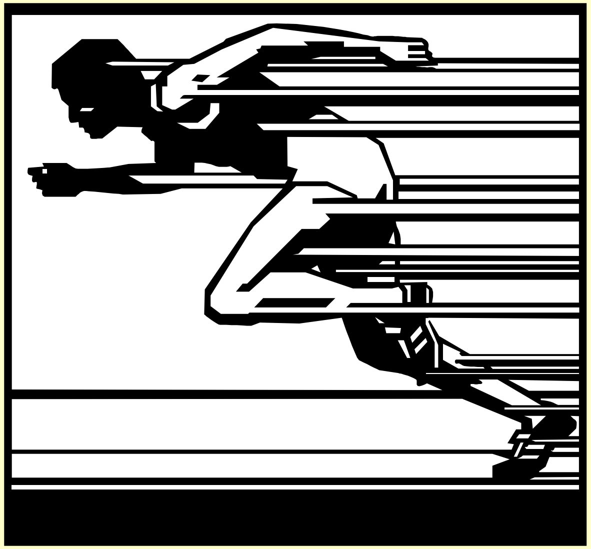
Test statistic $$F^* = \frac{MSR}{MSE}$$

P-value = What is the probability that we'd get an F* statistic as <u>large</u> as we did, if the null hypothesis is true? (<u>One-tailed</u> test!)

The P-value is determined by comparing F* to an **F distribution** with *1* **numerator degree of freedom** and *n*-*k*-*1* **denominator degrees of freedom**.

Winning times (in seconds) in Men's 200 meter Olympic sprints, 1900-1996.

Are men getting faster?



Row	Year	Men200m
1	1900	22.20
2	1904	21.60
3	1908	22.60
4	1912	21.70
5	1920	22.00
6	1924	21.60
7	1928	21.80
8	1932	21.20
9	1936	20.70
10	1948	21.10
11	1952	20.70
12	1956	20.60
13	1960	20.50
14	1964	20.30
15	1968	19.83
16	1972	20.00
17	1976	20.23
18	1980	20.19
19	1984	19.80
20	1988	19.75
21	1992	20.01
22	1996	19.32

Regression Plot
Men200m = 76.1534 - 0.0283833 Year
S = 0.298134 R-Sq = 89.9 % R-Sq(adj) = 89.4 %
22.5
21.5
20.5
19.5
Men200m
1900
1950
2000
Year

Analysis of Variance Table

DF_E = n-k-1 = 22-2 = 20

MSE = SSE/(n-2) = 1.8/20 = 0.09

MSR = SSR/1 = 15.8

```
Analysis of Variance
Source            DF      SS       MS       F        P
Regression         1    15.8     15.8    177.7    0.000
Residual Error    20     1.8     0.09
Total             21    17.6
```

DF_{TO} = n-1 = 22-1 = 21

F* = MSR/MSE = 15.796/0.089 = 177.7

P = Probability that an F(1,20) random variable is greater than 177.7 = 0.000…

For simple linear regression model, the F-test and t-test are equivalent.

Predictor	Coef	SE Coef	T	P
Constant	76.153	4.152	18.34	0.000
Year	-0.0284	0.00213	-13.33	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15.796	15.796	177.7	0.000
Residual Error	20	1.778	0.089		
Total	21	17.574			

$$\boxed{(-13.33)^2 = 177.7} \quad \left(t^*_{(n-k-1)}\right)^2 = F^*_{(1,n-k-1)}$$

Equivalence of F-test to t-test

- For a given α level, the F-test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is algebraically equivalent to the two-tailed t-test.
- Will get exactly same P-values, so…
  - If one test rejects H_0, then so will the other.
  - If one test does not reject H_0, then so will the other.

Should I use the F-test or the t-test?

- The F-test is only appropriate for testing that the slope differs from 0 ($\beta_1 \neq 0$).
- Use the t-test to test that the slope is positive ($\beta_1 > 0$) or negative ($\beta_1 < 0$).
- F-test is more useful for multiple regression model when we want to test that more than one slope parameter is 0. Test if β_1 and β_2 are jointly significant

Alternative formula for F-test

- **Null hypothesis**

H_0: $\beta_1 = \beta_2 = 0 \parallel R^2 = 0$

- **Alternative hypothesis**

H_A: $\beta_1 \neq \beta_2 \neq 0 \parallel R^2 \neq 0$

- Test statistic

$$F^* = \frac{R^2/k}{(1 - R^2)/n - k - 1}$$

- F-critical

$$Fcritical = F_{k,n-k-1}$$

k – Column, n-k-1 – Row

- When F*>F-critical,

Reject H_0

R^2 is statistically significant

When F*<F-critical,

Fail to reject H_0

R^2 is not statistically significant

P-values