STAT 302 - Chapter 2 : Inference for Simple Linear Regression - Part 2

Harsha Perera

Key Topics

- Analysis of Variance (ANOVA)
- ANOVA Test for Simple Linear Regression
- Inference for Correlation
- Coefficient of Determination - r^2

ANOVA

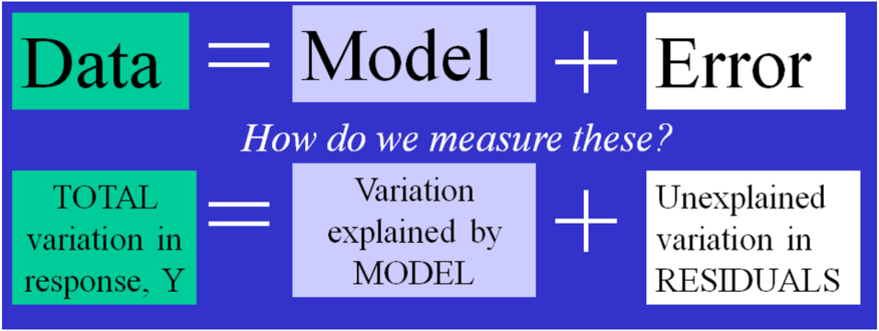
- Another way to assess the effectiveness of the model is to measure how much of the variability in the response variable is explained by the predictions based on the fitted model.

- This general technique is known as **Analysis of Variance** abbreviated as **ANOVA**.

- Even though we illustrate ANOVA in the context of the simple linear regression model, this approach could applied to any situation in which a model is used to obtain predictions.

Partitioning Variability - ANOVA

The basic idea is to partition the total variability in the response into two pieces.The **variability explained by the model** and the **variability explained by the errors** which captured in the residuals.

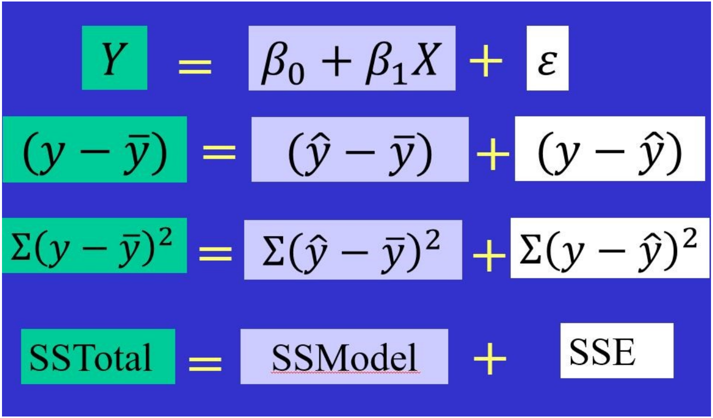


Data = Model + Error

How do we measure these?

TOTAL variation in response, Y = Variation explained by MODEL + Unexplained variation in RESIDUALS

Partitioning Variability - ANOVA



$$Y = \beta_0 + \beta_1 X + \varepsilon$$

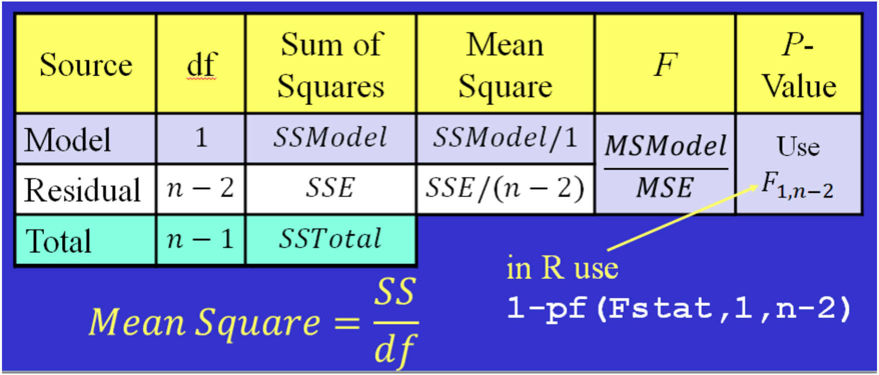
$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

$$\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2$$

SSTotal = SSModel + SSE

ANOVA Test for Simple Linear Regression

$$H_0 : \beta_1 = 0 \text{ Vs } H_a : \beta_1 \neq 0$$



Source	df	Sum of Squares	Mean Square	F	P-Value
Model	1	$SSModel$	$SSModel/1$	$\frac{MSModel}{MSE}$	Use $F_{1,n-2}$
Residual	$n-2$	SSE	$SSE/(n-2)$		
Total	$n-1$	$SSTotal$			

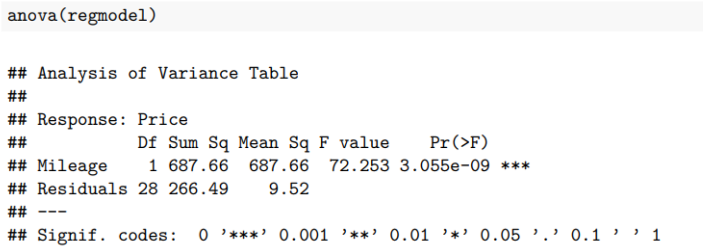
in R use

```
1-pf(Fstat,1,n-2)
```

$$Mean\ Square = \frac{SS}{df}$$

The **p-value** is obtained from the upper tail of an F-distribution with 1 and n-2 DFs.

Example 2.2



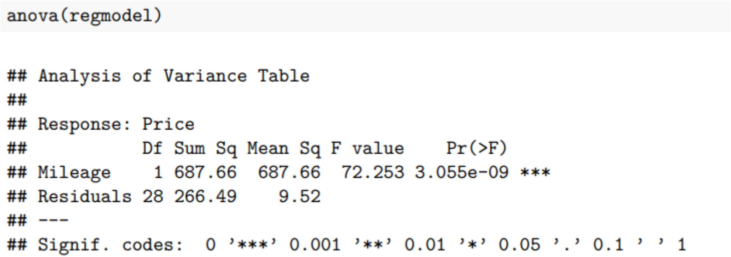
```
anova(regmodel)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Mileage    1 687.66  687.66  72.253 3.055e-09 ***
## Residuals 28 266.49    9.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $\text{MSModel} = 687.66/1 = 687.66 \quad \text{MSE} = 266.49/28 = 9.52$
- $F = 687.66/9.52 = 72.253$
- Comparing this to an F distribution with 1 numerator and 28 denominator DFs, the p-value is very close to 0.
- Thus we conclude that price and mileage have some relationship and that this model based on mileage is effective for predicting the price of used Accords.

R output of ANOVA

Analysis of Variance (ANOVA) output for the Accord Price Data :



```
anova(regmodel)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Mileage    1 687.66  687.66  72.253 3.055e-09 ***
## Residuals 28 266.49    9.52
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Examining the ”Sum Sq” column in the oputput shows the following values for this partition of variability for the Accord regression model :

SSModel $=$ 687.66   SSE$=$ 266.49   SSTotal $=$ 954.15

ANOVA Test for Simple Linear Regression

- How do we tell if the model explains a significant amount of variability or if the explained variability is due to chance alone?
- The relevant hypotheses would be the same as those for the t-test for the slope;

$$H_0 : \beta_1 = 0 \text{ Vs } H_a : \beta_1 \neq 0$$

- **Degrees of Freedom (DF)** - SSTotal has $n - 1$ DF, SSModel has just 1 DF, and SSE has $n - 2$ DF.

- SSTotal DF = SSModel DF + SSE DF

- In Accord Regression model : Total DF = 29 Model DF = 1 Error DF = 28

ANOVA Test for Simple Linear Regression

- We divide each Sum of Squares by the appropriate DF to form **Mean Squares (MS)**:

$$\text{MSModel} = \text{SSModel}/1 \quad \text{MSE} = \text{SSE}/\text{n-2}$$

- The **test statistic** is F = MSModel/ MSE where F follows an F distribution with a numerator DF 1 and a denominator DF n-2.

Inference for Correlation

- Correlation coefficient r is a number between -1 and $+1$ that measures the direction and strength of the linear relationship between two quantitative variables.

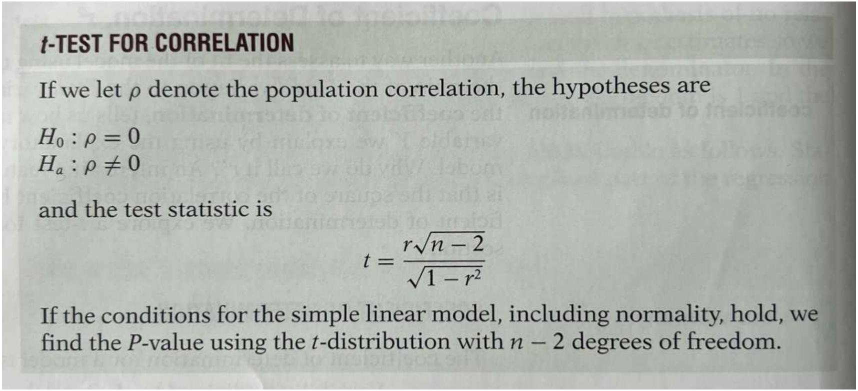
- The least square slope $\hat{\beta}_1$ is closely related to the correlation r between X and Y in a sample.

- $\hat{\beta}_1$ can be obtained from the sample correlation together with the standard deviations of X and Y as;

$$\hat{\beta}_1 = r\frac{s_Y}{s_X}$$

- Testing the null hypothesis $H_0 : \beta_1 = 0$ is the same as testing that there is no correlation between X and Y in the population from which we sampled the data.

t-Test for Correlation



t-TEST FOR CORRELATION

If we let ρ denote the population correlation, the hypotheses are

$H_0 : \rho = 0$
$H_a : \rho \neq 0$

and the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

If the conditions for the simple linear model, including normality, hold, we find the *P*-value using the *t*-distribution with $n - 2$ degrees of freedom.

- In R we can conduct the Correlation Test to obtain the p-value. Refer Example 2.4 R code and the output.

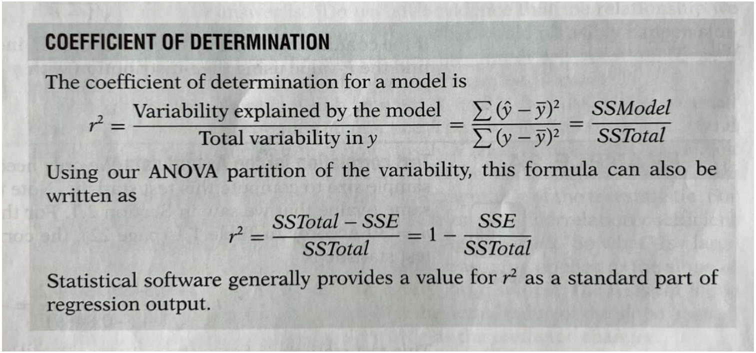
Coefficient of Determination - r^2

- Another way to assess the fit of the model using the ANOVA table is to compute the ratio of the **Model variation** to the **Total variation**.

- This statistic is known as **Coefficient of Determination**, which tells us how much variation in the response variable Y is explained by the X in the regression model.

- Why do we call it r^2 ? An interesting feature of simple linear regression is that the square of the correlation coefficient happens to be exactly the coefficient of determination.

Coefficient of Determination - r^2



COEFFICIENT OF DETERMINATION

The coefficient of determination for a model is

$$r^2 = \frac{\text{Variability explained by the model}}{\text{Total variability in } y} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{SSModel}{SSTotal}$$

Using our ANOVA partition of the variability, this formula can also be written as

$$r^2 = \frac{SSTotal - SSE}{SSTotal} = 1 - \frac{SSE}{SSTotal}$$

Statistical software generally provides a value for r^2 as a standard part of regression output.

- Value of r^2 is labeled as **"Multiple R-squared"** in the original summary(regmodel) output in R.
- Since r^2 is the fraction of the response variability that is explained by the model, we often convert the value to a percentage. In our Honda Accord example, we find that 72.1% of the variability in the prices of the Accords in this sample can be explained by the linear model based on their mileages.

Summary

- We discussed three methods to test for a significance of a linear relationship between two quantitative variables.

  1. The t-test for slope
  2. The ANOVA for regression
  3. The t-test for correlation

- These three tests are exactly equivalent in the case of simple linear regression.

Important : The t-test statistics for slope and correlation are equal (-8.50) and the F-statistic is the square of the t-statistic ($-8.50^2 = 72.25$).

While the three tests are exactly equivalent in the simple linear regression case, these tests take on different roles when we consider multiple linear regression in Chapter 3.