LINEAR REGRESSION

Pearson correlation coefficient

- **Correlation coefficients** are used to measure how strong a relationship is between two variables.
- There are several types of correlation coefficient, but the most popular is Pearson's.
- **Pearson's correlation** (also called Pearson's *R*) is a **correlation coefficient** commonly used in linear regression.
- The range of the correlation coefficient is from -1 to 1.

Pearson correlation coefficient

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson correlation coefficient

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where,

- r = Pearson Coefficient
- n= number of observations
- $\sum xy$ = sum of products of the paired stocks
- $\sum x$ = sum of the x scores
- $\sum y$= sum of the y scores
- $\sum x^2$ = sum of the squared x scores
- $\sum y^2$ = sum of the squared y scores

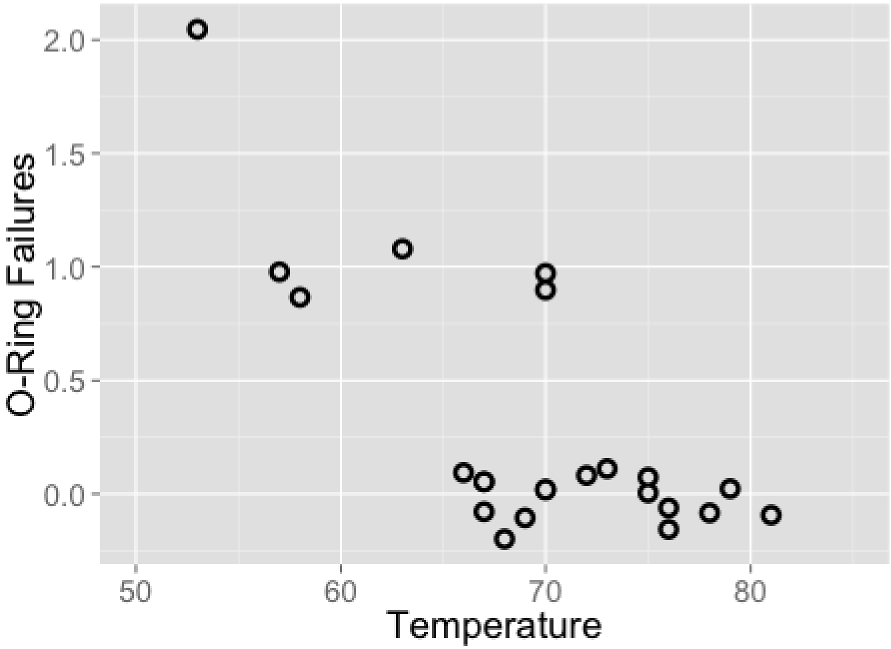
- **Find out the number of pairs of variables, which is denoted by n. Let us presume x consists of 3 variables – 6, 8, 10. Let us presume that y consists of corresponding 3 variables 12, 10, 20.**

x	y	x*y	x^2	y^2
6	12	72	36	144
8	10	80	64	100
10	20	200	100	400
24	**42**	**352**	**200**	**644**

- **Insert the values found above in the formula and solve it.**
- r = 3*352-24*42 / √(3*200-24^2)*(3*644-42^2)
  = 0.7559

What is regression?

On January 28, 1986, seven crewmembers of the United States space shuttle Challenger were killed when O-rings responsible for sealing the joints of the rocket booster failed and caused a catastrophic explosion



Regression
specifying the relationship between a single numeric **dependent variable** (the value to be predicted) and one or more numeric **independent variables** (the predictors)

Regression Problem
Predicting a real valued output

Supervised Learning
"Correct Value" for each example is given in the data

Simple Linear Regression

- Let the training data set $\mathrm{D} = \{(x_i, y_i)\}^N{}_{i=1}$

Sample No.	No. of O-ring Failures (x)	Temperature (y)
1.	2	50
2.	1	57
…	…	…
N	0	81

N – No. of Training samples
x – Input Variable
y - Target Variable
(x,y) – Training samples
x_i,y_i – i^{th} training sample

Simple Linear Regression

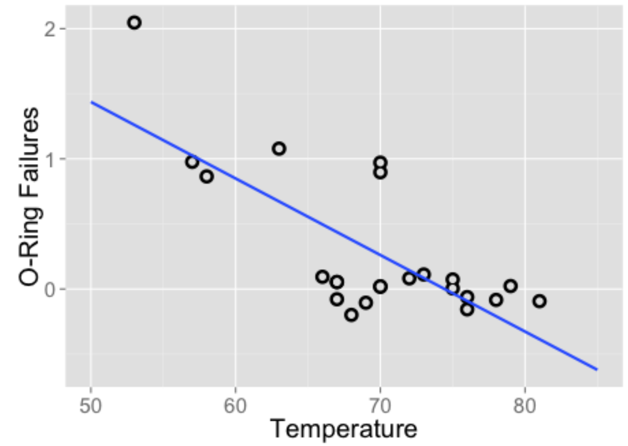
Training data

↓

Learning Algorithm

↓

Temperature → H → No. of O-ring failures



Representing the hypothesis H

H(x) = $\theta_0 + \theta_1 x$

↑

Simple Linear Regression

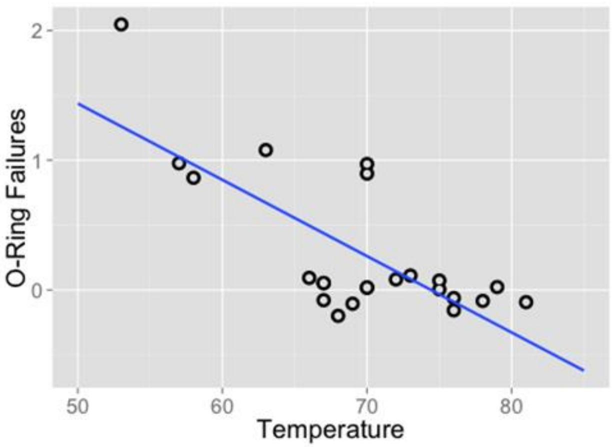
Linear Regression with a single independent variable

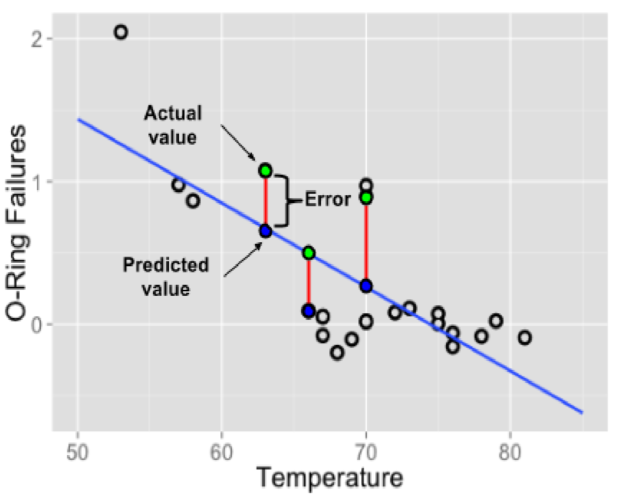
Univariate linear regression

Question to Ponder

- Can regression be used for other types of dependent variables?

Regression Analysis





Best Fit Line

Choose θ_0 and θ_1 such that H(x) is as much as close to y

θ_0 and θ_1 are called parameters

Regression analysis involves finding the optimal parameter estimates

Predicted Value　Actual Value

$$\underset{\theta_0,\ \theta_1}{Minimize} \underbrace{\sum(\underbrace{H(x) - y}_{\text{Error or Residual}})^2}_{\text{Squared Error Function}}$$

Error or Residual

Squared Error Function

Ordinary Least Square Estimation

- Using Calculus, the value of θ_1 that results in the minimum squared error is:

$$\theta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\theta_1 = \frac{Cov(x, y)}{Var(x)}$$

- The optimal value of θ_1 is:

$$\theta_0 = \bar{y} - \theta_1\bar{x}$$

Example

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
3	4			10	6

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(yi-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(xi-\bar{x})^2} = \frac{6}{10} = 0.6$$

$$\hat{b}_0 = \bar{y}\text{-}\hat{b}_1\bar{x} = 2.2$$

$$\boxed{\hat{y} = 2.2 + .6x}$$

Performance Metrics

- Evaluate the accuracy of the regression model?
  - R-Squared – Measure of squared deviation from the expected value

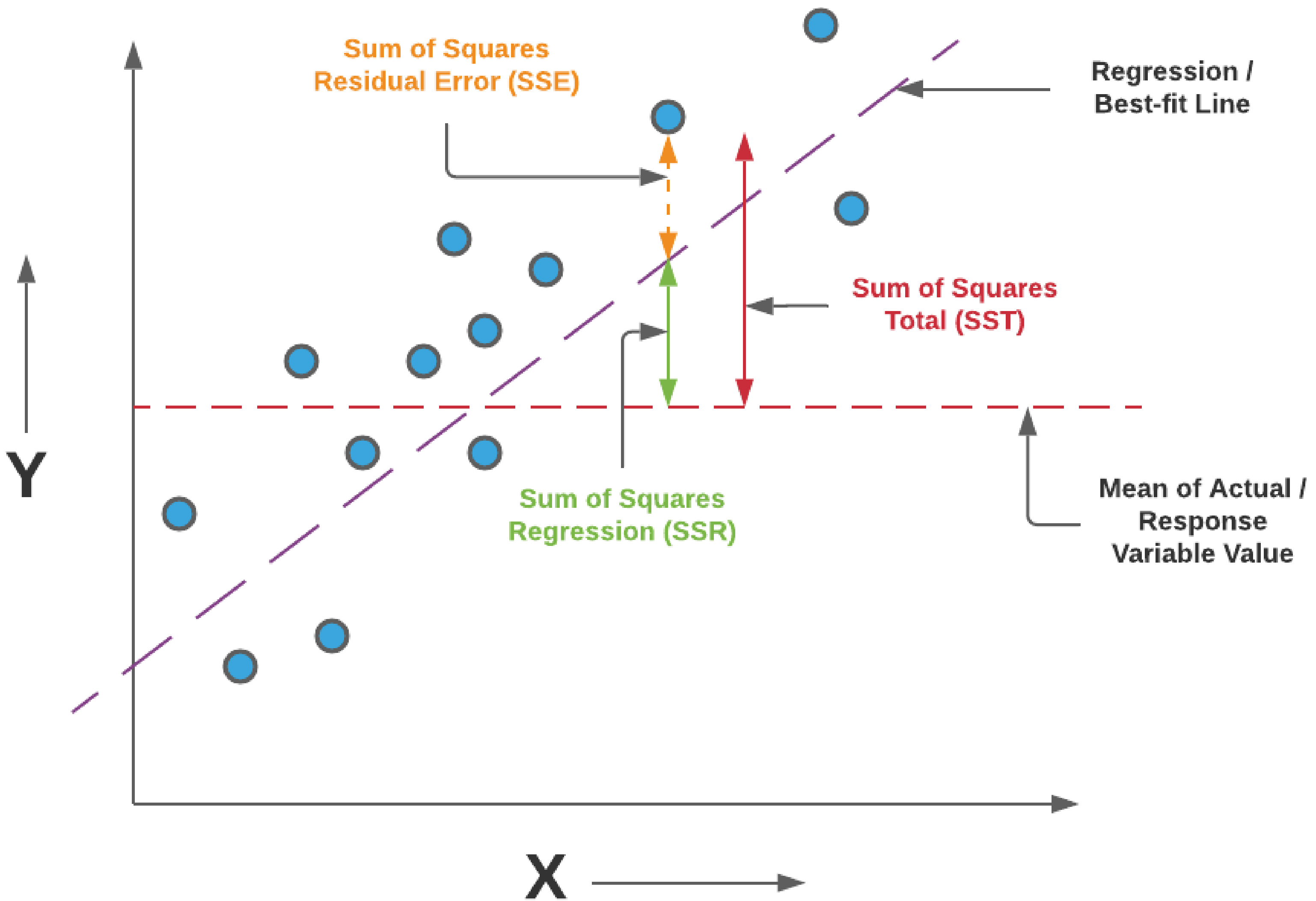
Regression Sum of Squares

$$R^2 = \frac{Explained\ Variance}{Total\ Variance} = \frac{SSR}{SST}$$

Total Sum of Squares

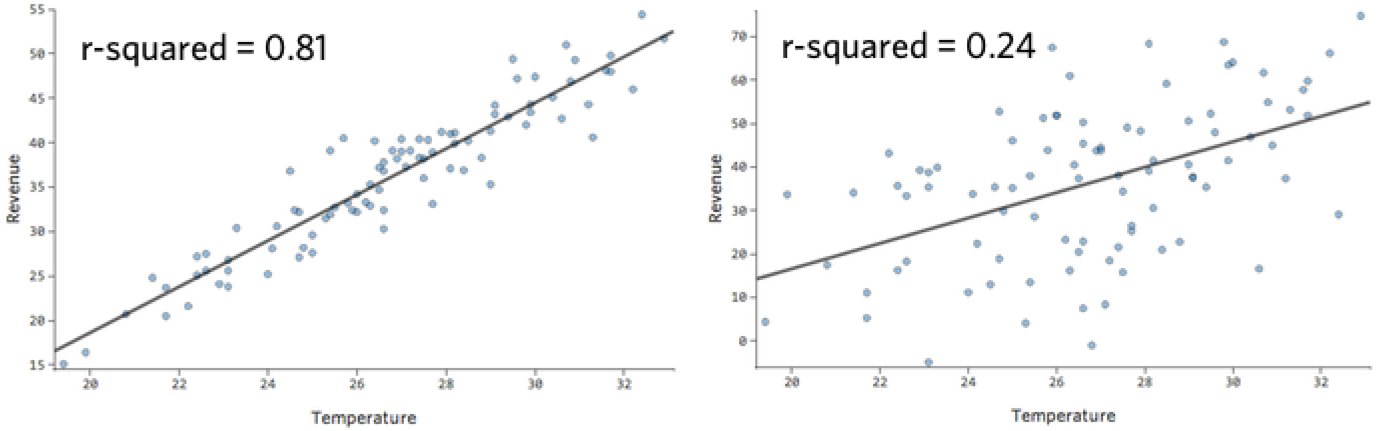
$where\ SST = \sum_{i=1}^{N}(y_i - \bar{y})^2$ and SSR= $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$

Performance Metrics



Performance Metrics

- For example, if R^2 =0.8, then 80% of variance in the data is explained by the model.



Linear Regression Model (with Normally Distributed Errors)

- In most linear regression analyses, it is common to assume that the error term is a normally distributed random variable with mean equal to zero and constant variance.

- Thus, the linear regression model is expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} + \varepsilon$$

where:

y is the outcome variable

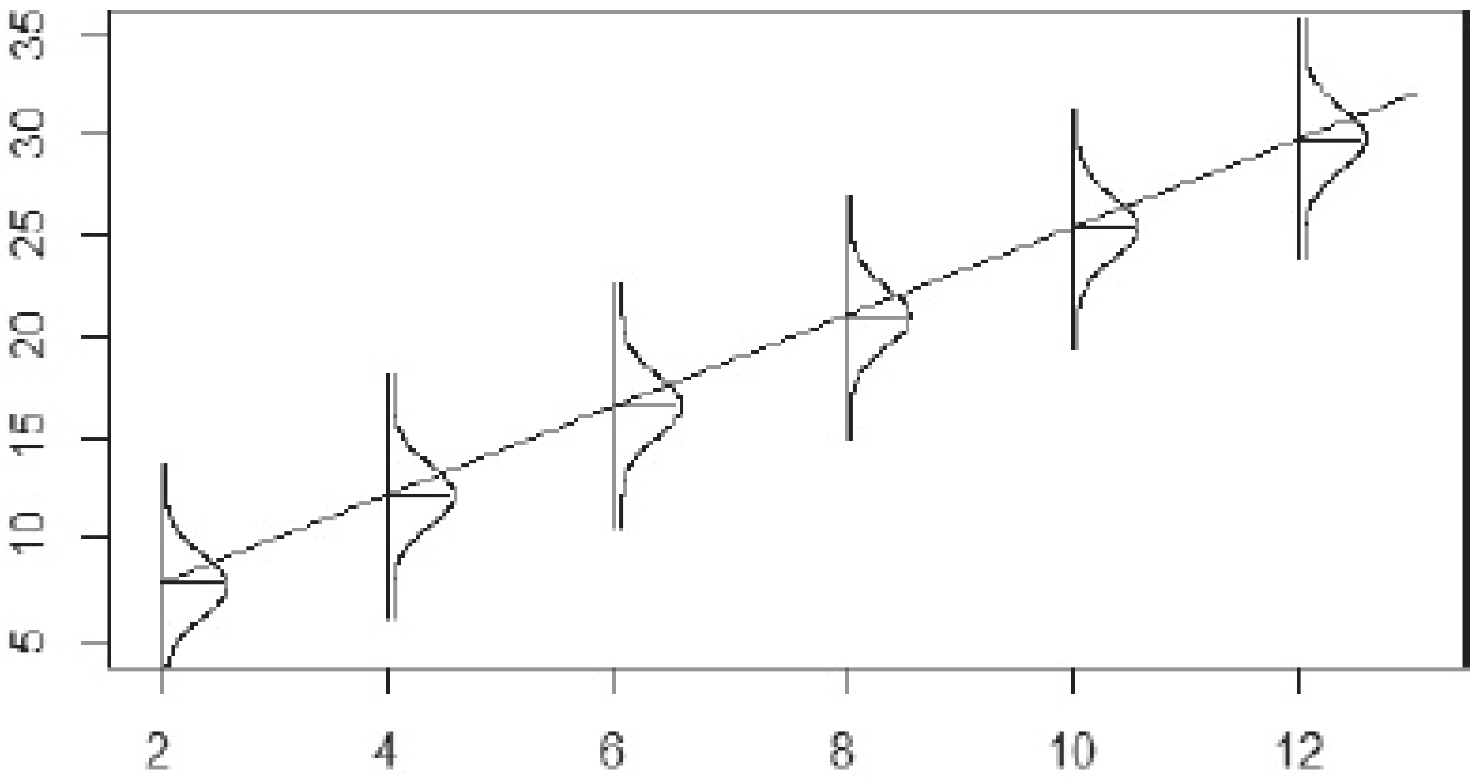
x_j are the input variables, for j = 1, 2,…, p – 1

β_0 is the value of y when each x_j equals zero

β_j is the change in y based on a unit change in x_j, for j = 1, 2,. . ., p – 1

$\varepsilon \sim N(0, \sigma^2)$ and the εs are independent of each other

Linear Regression Model (with Normally Distributed Errors)



Normal distribution about y for a given value of x

Sample data and Model

- Data: Marketing from Datarium package
- We want to predict future sales on the basis of advertising budget spent on youtube.
  - sales = b0 + b1 * youtube
- The R function lm() can be used to determine the beta coefficients of the linear model:

```
#building a linear model
model <- lm(sales~youtube,data=marketing)
model
```

```
Coefficients:
(Intercept) youtube
8.43911     0.04754
```

Model Assessment

- Before using this formula to predict future sales, you should make sure that this model is statistically significant, that is:
  - there is a statistically significant relationship between the predictor and the outcome variables
  - the model that we built fits the data in our hand very well .

Model summary

- summary(model) -outputs shows 6 components:

- **Call**
  - Shows the function call used to compute the regression model.
- **Residuals**
  - Provide a quick view of the distribution of the residuals, which by definition have a mean zero. Therefore, the median should not be far from zero, and the minimum and maximum should be roughly equal in absolute value.
- **Coefficients**
  - Shows the regression coefficients and their statistical significance.
  - Predictor variables, that are significantly associated to the outcome variable, are marked by stars.

Coefficients Significance

t-statistic and p-values:

- For a given predictor, the t-statistic (and its associated p-value) tests whether or not there is a statistically significant relationship between a given predictor and the outcome variable, that is whether or not the beta coefficient of the predictor is significantly different from zero.

- The statistical hypotheses are as follow:
  - Null hypothesis (H0): the coefficients are equal to zero (i.e., no relationship between x and y)
  - Alternative Hypothesis (Ha): the coefficients are not equal to zero (i.e., there is some relationship between x and y)

Coefficients Significance

t-statistic and p-values:

- The t-statistic measures the number of standard deviations that θ is away from 0. Thus a large t-statistic will produce a small p-value.

- Higher the t-statistic (and the lower the p-value), more significant the predictor is.

- A statistically significant coefficient indicates that there is an association between the predictor (x) and the outcome (y) variable.

Model Accuracy

- Metrics that are used to check how well the model fits our data.
  - The Residual Standard Error (RSE)
  - The R-squared (R2)
  - F-statistic

Residual standard error (RSE)

- The RSE (also known as the model sigma) is the residual variation, representing the average variation of the observations points around the fitted regression line. This is the standard deviation of residual errors.

- RSE provides an absolute measure of patterns in the data that can’t be explained by the model. When comparing two models, the model with the small RSE is a good indication that this model fits the best the data.

Residual standard error (RSE)

- Dividing the RSE by the average value of the outcome variable will give you the prediction error rate, which should be as small as possible.
- In our example, RSE = 3.91, meaning that the observed sales values deviate from the true regression line by approximately 3.9 units in average.
- Whether or not an RSE of 3.9 units is an acceptable prediction error is subjective and depends on the problem context. However, we can calculate the percentage error. In our data set, the mean value of sales is 16.827, and so the percentage error is 3.9/16.827 = 23%.

R-squared and Adjusted R-squared

- The R-squared (R2) ranges from 0 to 1 and represents the proportion of information (i.e. variation) in the data that can be explained by the model. The adjusted R-squared adjusts for the degrees of freedom.

- The R2 measures, how well the model fits the data. For a simple linear regression, R2 is the square of the Pearson correlation coefficient.

R-squared and Adjusted R-squared

- A high value of R^2 is a good indication.
- However, as the value of R^2 tends to increase when more predictors are added in the model, such as in multiple linear regression model, you should mainly consider the adjusted R-squared, which is a penalized R^2 for a higher number of predictors.
  - An (adjusted) R^2 that is close to 1 indicates that a large proportion of the variability in the outcome has been explained by the regression model.
  - A number near 0 indicates that the regression model did not explain much of the variability in the outcome.

F-Statistic

- The F-statistic gives the overall significance of the model. It assess whether at least one predictor variable has a non-zero coefficient.
- In a simple linear regression, this test is not really interesting since it just duplicates the information in given by the t-test, available in the coefficient table. In fact, the F test is identical to the square of the t test: 312.1 = (17.67)^2. This is true in any model with 1 degree of freedom.
- The F-statistic becomes more important once we start using multiple predictors as in multiple linear regression.
- A large F-statistic will corresponds to a statistically significant p-value ($p < 0.05$). In our example, the F-statistic equal 312.14 producing a p-value of 1.46e-42, which is highly significant.