# Simple Linear Regression (SLR)

Statistics 251: Statistical Methods

Module 12

Updated 2020

## Simple Linear Regression (slr)

- SLR analysis explores the linear association between an explanatory (independent) variable, usually denoted as $x$, and a response (dependent) variable, usually denoted as $y$
- This type of data is called bivariate data (data with two (bi) variables)
- The point is to see if we can use a mathematical linear model to describe the association (relationship) between the two variables
- Using one known value to estimate the other value, in addition to seeing how strong the relationship is
- You are familiar with $y = mx + b$ from algebra, where $m$ is the slope and $b$ is the $y$-intercept (value of $y$ when $x = 0$), which is a mathematical linear equation, a *deterministic* equation.

## The population regression model

Notice that it is basically the same as you have seen and used before ($y = mx + b$):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where:

- $y_i$: value of the response (dependent) variable
- $\beta_0$: the value of the $y$-intercept (when $x = 0$)
- $\beta_1$: the value of the slope (the change in $y$ due to a one unit increase in $x$, **not** $\frac{rise}{run}$)
- $\epsilon_i$: the residual (error) term

## The sample regression model

Is used once there are estimated values from the data:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Where:

- $\hat{y}_i$: estimate of the value of the $i^{th}$ response (dependent) variable
- $\hat{\beta}_0$: the estimate of the value of the $y$-intercept ($\hat{y}$ when $x = 0$)
- $\hat{\beta}_1$: the estimate of the value of the slope (the change in $y$ due to a one unit increase in $x$. **Not** $\frac{rise}{run}$)
- Note that $\epsilon_i$ dropped off from the other model. This is because of the first assumption of regression, $E(\epsilon_i) = 0$: the mean of the residuals = 0.

## Assumptions of SLR

(1) $E(\epsilon_i) = 0$: the mean of the residuals is $\approx 0$
(2) $V(\epsilon_i) = \sigma_\epsilon^2$: the variance of the residuals is constant (the same) for all values of $\hat{y}$. Also called constant variance, homogeneity of variance (means same variance)
(3) $Cov(\epsilon_i, \epsilon_j) = 0$: independence of residuals
(4) $\epsilon_i \sim N(0, \sigma_\epsilon^2)$: Residuals have an approximate normal distribution with mean 0 and homogeneous variance

## Least Squares

The method most common for regression parameter estimation is called *least squares.* This method evaluates the quality of a line's fit to the data by the sum of the squared *vertical distances* of each point $(X_i, Y_i)$ from the line. Those distances are called residuals (also referred to as errors).

## Residuals

**Residuals**: $\epsilon_i$ are the population residuals and $\hat{\epsilon}_i = e_i$ are the sample residuals

$e_i = y_i - \hat{y}$. If $e_i > 0$, the model *understimated* the response and if $e_i < 0$, the model *overstimated* the response.

$s_\epsilon^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$ the average *squared* distance between each estimated $y$ and the observed value of $y$, called $MSE$, mean squared error, or residual variance (the variance of the residuals).

$s_\epsilon = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$ the average distance between each estimated $y$ and the observed value of $y$, called $RMSE$, root mean squared error, or residual standard error (the standard error of the residuals).

## Analysis tools: scatterplot graph

- First thing that is necessary is to look at a scatterplot of the two variables; it is a type of graph that you are familiar with from algebra
  - $x$ is the explanatory (independent) variable and goes along the $x$-axis
  - $y$ is the response (dependent) variable and goes along the $y$-axis
- The values of $x$ and $\hat{y}$ are an ordered pair of data, $(x, \hat{y})$ that can be graphed on the cartesian (rectangular) coordinate system
- The value of $x$ that will be given is most often one that is an observed value of $x$ so that an estimation of the residual, $e_i = \hat{y}_i - y_i$ can be calculated.

## Analysis tools: scatterplot graph

- A scatterplot of the data shows if there is a linear association between the explanatory (independent) variable and the response (dependent) variable
  - When $x$ and $y$ both increase, the slope (relationship) is positive
  - When $x$ increases while $y$ decreases, the slope (relationship) is negative
- The point of visually checking the scatterplot **before** doing the regression analysis is decide if there is at least a fair linear relationship between $x$ and $y$
  - If you do not have a linear relationship, then use of regression analysis is not recommended as the results cannot be used with the given dataset
- The regression line is also called a trend line.

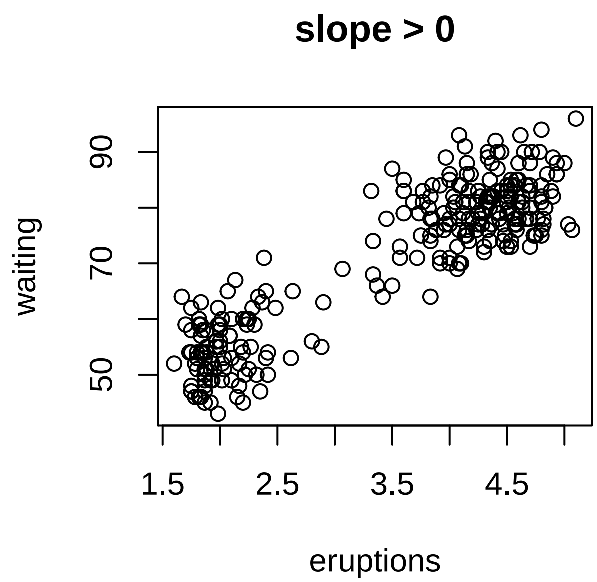
## Module example data

With the example throughout this lecture will be Old Faithful; eruptions is the duration of the eruption of Old Faithful and waiting is the interval between eruptions, both in minutes. Eruptions will be the explanatory (independent) variable and waiting will be the response (dependent) variable, modelling waiting time by eruption duration; in other words, we are using the eruption time to estimate the time until the next eruption. Let $x$=`eruptions` and $y$=`waiting`.

```
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

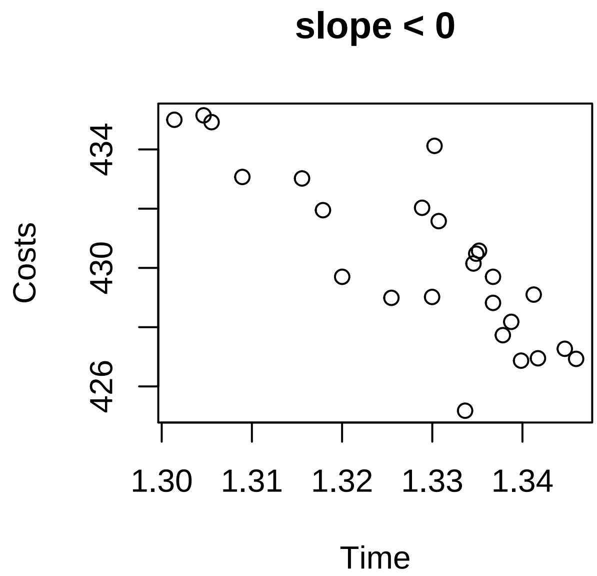
## Analysis tools: scatterplot graph

This has positive slope ($x$ increases and $y$ increases)



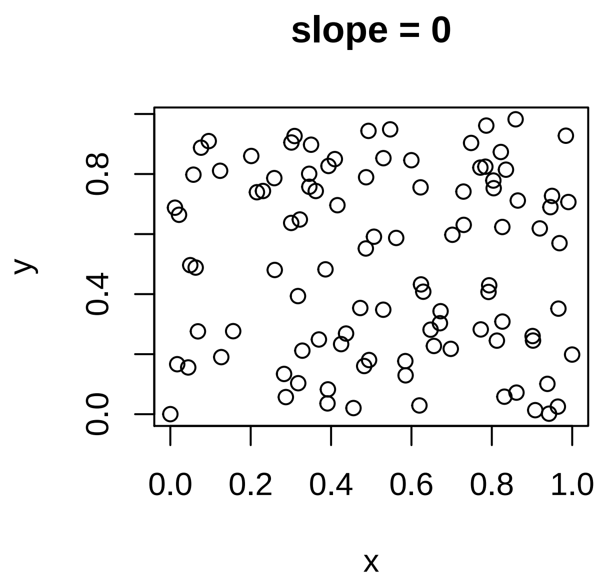

## Analysis tools: scatterplot graph

This has negative slope ($x$ increases and $y$ decreases)



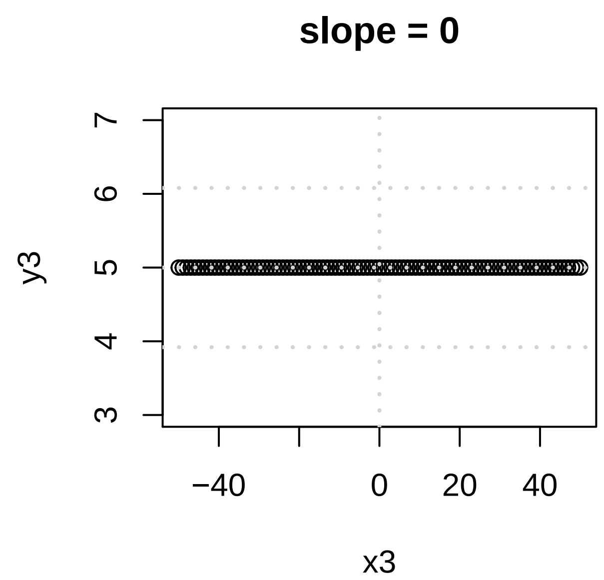

**Analysis tools: scatterplot graph**

This has 0 slope (and a lot of random scatter)




**Analysis tools: scatterplot graph**

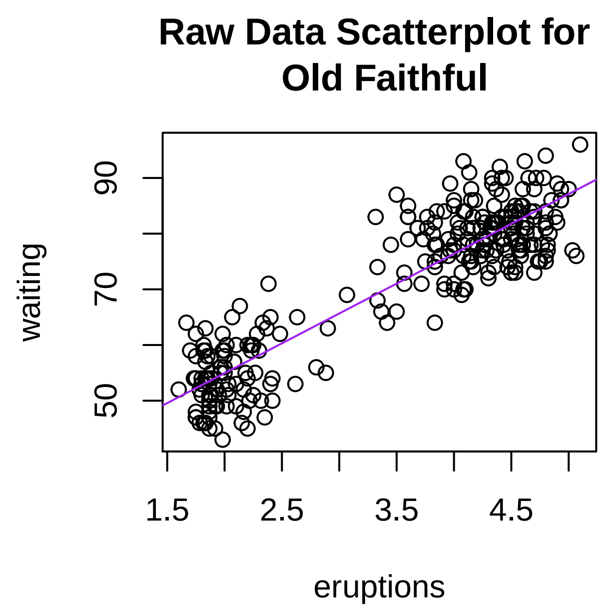
This has 0 slope




## Analysis tools: scatterplot graph with regression line

Many times in regression, we want to see what the line of the regression equation will look like on the scatterplot of the raw data. It is not strictly necessary but the point of this analysis is to explore and understand the linear relationship between two variables. If you do not have a linear relationship, then use of this analysis is not recommended as the results cannot be used with the given dataset. The regression line is also called a trend line.

## Analysis tools: scatterplot graph with regression line




## Slope and intercept formulas

**Slope**:

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{s_x^2(n-1)} = \frac{S_{XY}}{S_{XX}}$$

**Intercept**:

$$\hat{\beta}_0 = \overline{y} - b\overline{x}$$

## Old Faithful Equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \Rightarrow \hat{y} = 33.474397 + 10.729641x$$

## Using the regression equation

Use of the equation works just like you are used to; given a specified value of $x$, solve the equation for the estimated $y$ value called $\hat{y}$ (y-hat)

Find the values of $\hat{y}$ and $e_i$ for each of the following values: $(2.283, 62), (5.1, 96)$

$$\hat{y}_{|x=2.283} = 33.474397 + 10.729641 * 2.283 = 57.970168$$

$$\hat{y}_{|x=5.1} = 33.474397 + 10.729641 * 5.1 = 88.195568$$

## Using the regression equation

$$e_{|x=2.283} = 62 - 57.970168 = 53.021374$$

$$e_{|x=5.1} = 96 - 88.195568 = 87.021374$$

Since both $e_i > 0$, the model understimated the waiting times.

## Correlation

To determine the strength of the relationship between two *quantitative* variables, we use a measure called *correlation*

**Defn**: Is a calculation that measures the strength and direction (positive or negative) of the *linear* relationship between 2 *quantitative* variables, $x$ and $y$

### Correlation $\neq$ causation

It is *extremely* important to note that just because two variables have a mathematical correlation **IT DOES NOT MEAN $X$ CAUSES $Y$!!!**. To establish actual causation, repeatable experimentation must be done.

## Correlation logistics

- It is bounded between -1 and 1 ($-1 \leq r \leq 1$)
  - $r = -1$ and $r = 1$ are perfect linear relationships
  - $r = 0$ implies both no linear relationship and $x$, $y$ are independent
- $r$ makes no distinction between $x$ and $y$
- $r$ has no units of measurement
- Correlation is denoted as $r$ for sample correlation and $\rho$ for the population correlation.

$$r = \frac{1}{n-1} \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

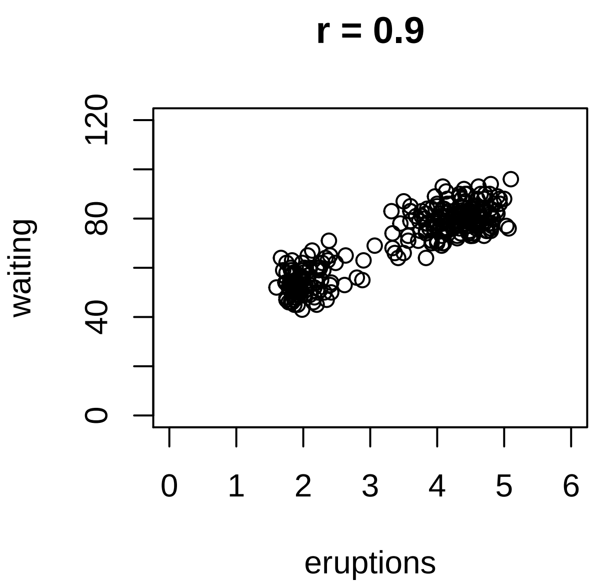
## Coefficient of Determination, $R^2$

$R^2$ is called the *coefficient of determination*:

- It is the proportion (or $\times 100\%$) of observed variation that can be explained by the relationship between $x$ and $y$
- $0 \leq R^2 \leq 1$: It is bounded between 0 (0%) and 1 (100%)
  - The closer to 1 (100%), the more variation we can explain and also the stronger the linear relationship between $x$ and $y$
    - An acceptable baseline for $R^2$ would be when $R^2 \geq 60\%$
- $R^2 = (r)^2 \therefore r = \pm\sqrt{R^2}$
  - if the slope is positive, then $r$ is positive, if the slope is negative, then $r$ is negative.

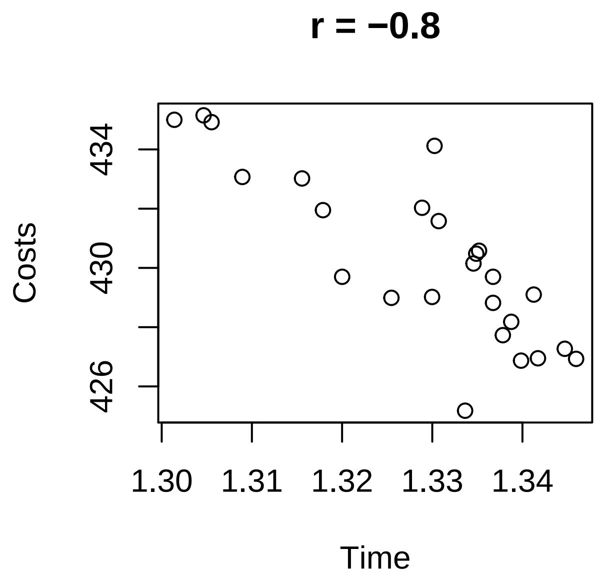
## Analysis tools: scatterplot graph

Relatively strong, positive correlation



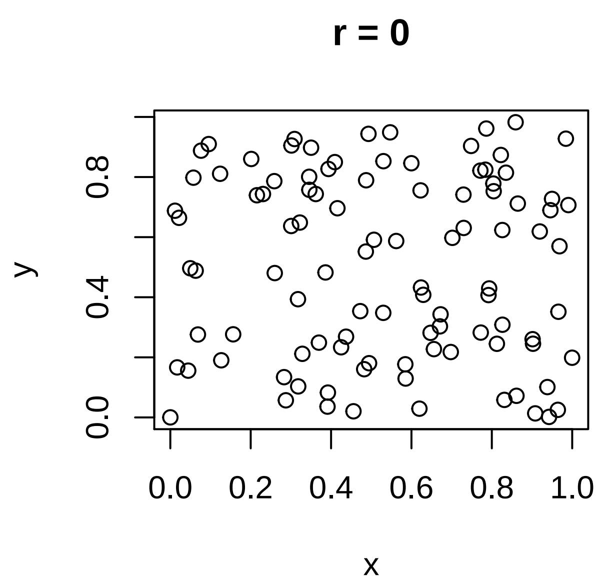

## Analysis tools: scatterplot graph

Moderately strong, negative correlation



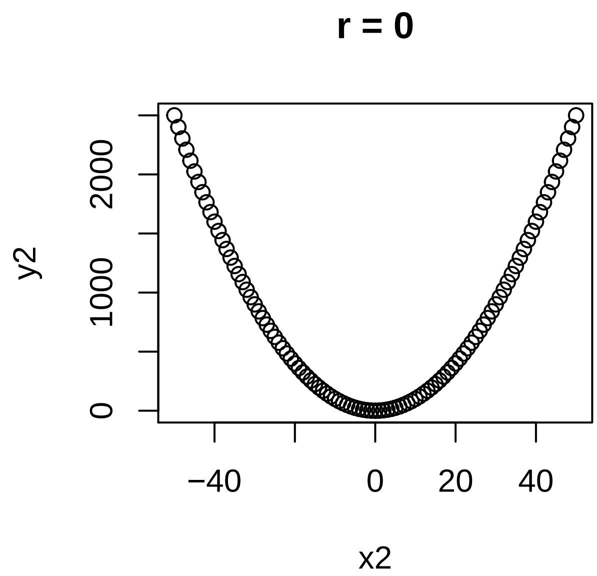

**Analysis tools: scatterplot graph**

No correlation




**Analysis tools: scatterplot graph**

No correlation but there *is* a relationship, it is not a linear relationship




### General rule-of-thumb for correlations

$r \approx 0$: no *linear* relationship
$r > |0.9|$: strong *linear* relationship
$|0.75| < r \leq |0.9|$: decent *linear* relationship
$|0.6| < r \leq |0.75|$: moderate *linear* relationship
$|0.4| < r \leq |0.6|$: fair *linear* relationship (meh)
$|0.2| < r \leq |0.4|$: weak *linear* relationship
$0 < r \leq |0.2|$: no real *linear* relationship

### Old Faithful $r$ and $R^2$

$$r = \frac{1}{n-1} \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = 0.900811$$

There is a strong, positive linear relationship between eruptions and waiting of Old Faithful.

$$R^2 = (r)^2 = (0.900811)^2 = 0.811461 \approx 81.15\%$$

We can explain approximately 81.15% of the variation in the response (waiting times) due to the linear relationship between $x$ and $y$ (which is good).

### Hypothesis tests for the estimated slope ($\beta_1$) and intercept ($\beta_0$)

- Most often the slope $\hat{\beta}_1$ is the only real test of interest

- Many times the value of $x = 0$ is not in the dataset (or the fact that mabye $x = 0$ is not possible in the population the data was sampled from). Without $x = 0$ in the dataset (or even possible at all), the intercept does not make sense in context
- Additionally, the slope is what is driving the relationship whereas the intercept just represents the value where the regression line crosses through the $y$-axis

- There are some economic datasets and many others that utilize the intercept because it make sense both mathematically and realistically.

## Hypothesis tests for the estimated slope ($\beta_1$) and intercept ($\beta_0$)

- The null hypothesis for the slope is to test if the slope is equal to zero
  - A slope of zero is a horizontal line, where any value of $x$ has the same $y$ value
- Most often of interest is whether or not it is significant, the alternative hypothesis is to see if the slope is different from zero
  - Realistically the hypothesized value could be something other than 0 if there is a need, like seeing if it has increased or decreased since the previous sample was taken and analyzed

### Test for $\beta_1$, the slope

**Hypotheses**:

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

**Test Statistic**:

$$t = \frac{\hat{\beta}_1 - \beta_1}{se_{\hat{\beta}_1}}$$

- The $se_{\hat{\beta}_1}$ and $df = n - 2$ are the same as for CIs

- Rejection criteria is the same as the $t$-tests learned in earlier modules (starting in module 9). Rejection of the null means the slope is significant; there is a significant relationship between $x$ and $y$. Not rejecting the null means there is no significant relationship between $x$ and $y$

### Test for $\beta_0$, the intercept

**Hypotheses**:

$$H_0 : \beta_0 = 0 \text{ vs. } H_a : \beta_0 \neq 0$$

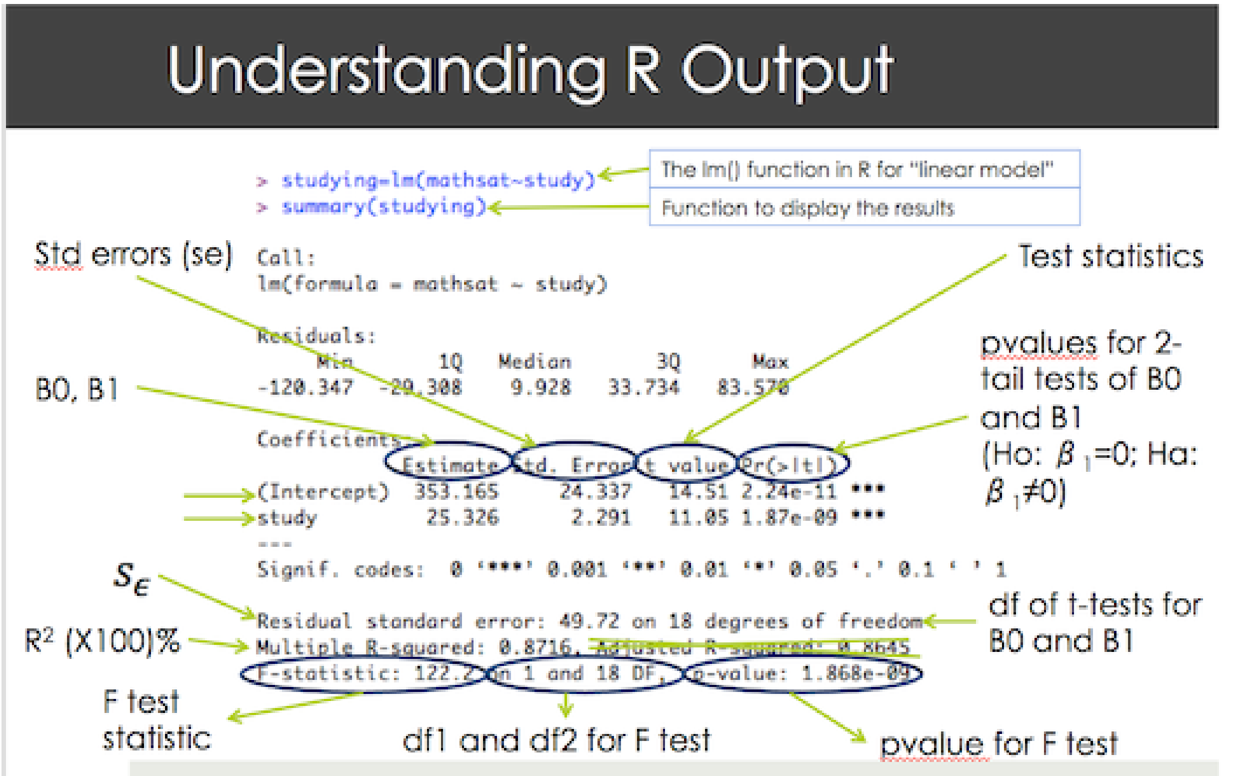
**Test Statistic**:

$$t = \frac{\hat{\beta}_0 - \beta_0}{se_{\hat{\beta}_1}}$$

- The $se_a$ and $df = n - 2$ are the same as for CIs
- Rejection criteria is the same as the $t$-tests learned in earlier modules (starting in module 9). Rejection of the null means the intercept is significant. Not rejecting the null just means the intercept is not significant (but has no impact on the significance of the slope)

### Reading `R` output

The following picture is a printout of a regression summary table from `fit=lm(y~x,data= )` and `summary(fit)`




### Notes on `R` output

`R` does not directly display correlation $r$ in the regression output but it does display the $R^2$ value (called `Multiple R-squared`)

Remember $r = \pm\sqrt{R^2}$, and use the sign of the slope to determine if $r$ is positive or negative

It is a proportion in the output but can be converted to a percent easily; it usually is when discussing its results

CIs and PIs (prediction intervals) for slope, intercept, $\hat{\mu}$, or $\hat{y}$ can be calculated but is not covered in this course

### Old Faithful Output

```
Call:
lm(formula = waiting ~ eruptions, data = faithful)

Residuals:
    Min      1Q  Median      3Q     Max
-12.080  -4.483   0.212   3.925  15.972

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   33.474      1.155    29.0   <2e-16 ***
eruptions     10.730      0.315    34.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.91 on 270 degrees of freedom
Multiple R-squared:  0.811, Adjusted R-squared:  0.811
F-statistic: 1.16e+03 on 1 and 270 DF,  p-value: <2e-16
```

## Test for $\beta_1$

**Hypotheses**:

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

$H_0$ can be rejected if $pvalue \leq \alpha$, which can be found on the output.

The procedure will require you to state the test statistic, *pvalue*, results, and conclusion.

$t = 34.09$, $pvalue = \texttt{2e-16} = 2 \times 10^{-16} = 0.0000000000000002 \approx 0 \leq \alpha(0.05)$. The null is rejected, meaning the slope is significant (also means the relationship between x and y is significant).

## $r$ and $R^2$

$R^2$ (Multiple R-squared) is 0.8115 meaning that 81.15% of the variation in the estimated response is explained by the linear relationship modeled with $x$ and $y$

$r$ could be calculated as $r = +\sqrt{R^2} = +\sqrt{0.8115} = 0.9$

It is positive since the slope is positive (if $r > 0$ then $\beta_1 > 0$, if $r < 0$ then $\beta_1 < 0$, and vice versa)

## Model assessment

The null hypothesis for the slope was rejected (slope is significant), both $r$ and $R^2$ are high, implying a strong linear relationship and 81% explainable variation in the response. This means that we have a decent model.

## Notation

- $\hat{\beta}_0$: sample intercept
- $\hat{\beta}_1$: sample slope
- $\hat{y}$: estimated value of y, called y-hat
- $e_i$: sample residual (estimate of $\epsilon_i$), $e_i = \hat{y}_i - y_i$ (estimated $y$-observed $y$)
- $y = \beta_0 + \beta_1 x + \epsilon$: population model
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$: sample regression equation
- $s_\epsilon^2$: sample residual variance, variance of residuals
- $s_\epsilon$: sample residual standard error, standard error of residuals