Statistics 301: Probability and Statistics

Simple Linear Regression (SLR)

Module 12

2018

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
- β_0: the value of the y-intercept (when $x = 0$)
- β_1: the value of the slope (the change in y due to a one unit increase in x, **not** $\frac{rise}{run}$)
- ϵ_i: the residual (error) term

The sample regression model

Is used once there are estimated values from the data:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \text{ or } \hat{y} = a + bx$$

Where:

- $\hat{y}_i$: estimate of the value of the i^{th} response (dependent) variable

- $\hat{\beta}_0$ (a): the estimate of the value of the y-intercept ($\hat{y}$ when $x = 0$)

- $\hat{\beta}_1$ (b): 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 ϵ_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 0
(2) $V(\epsilon_i) = \sigma^2_\epsilon$: 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) $\epsilon_i \sim N(0, \sigma^2_\epsilon)$: Residuals have an approximate normal distribution with mean 0 and homogeneous variance

Residuals

The *vertical distances* of each point (X_i, Y_i) from the line, are called residuals (also referred to as errors).

Residuals: ϵ_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.

Residual variation

$s^2_\epsilon = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$ 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; the scatterplot will show if there is a linear association between the explanatory (independent) variable and the response (dependent) variable
- 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.

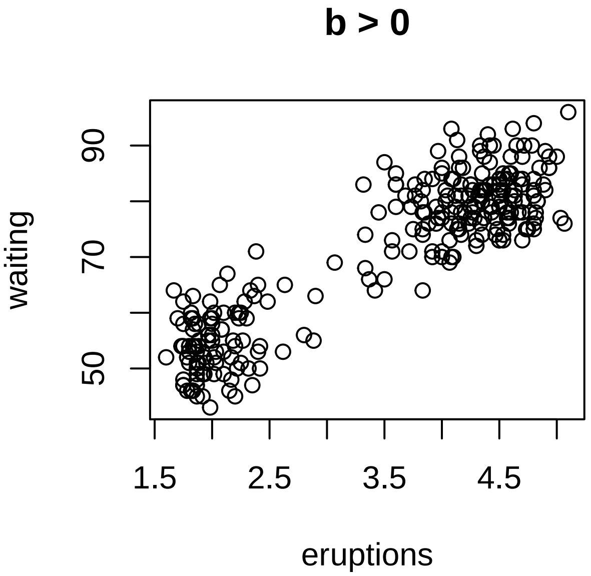
Eruptions will be the explanatory (independent) variable and waiting will be the response (dependent) variable, modelling waiting time by eruption duration; in other words, we are using the eruption time to estimate the time until the next eruption. Let x=`eruptions` and y=`waiting`.

```
  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
```

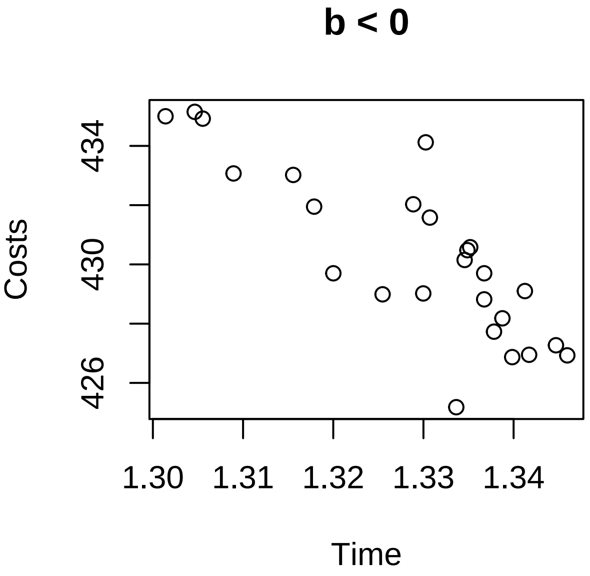
Analysis tools: scatterplot graph

This has positive slope (x increases and y increases)



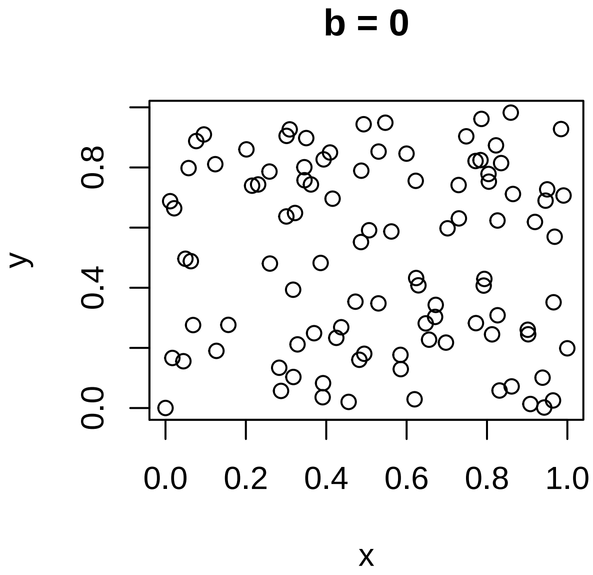
Analysis tools: scatterplot graph

This has negative slope (x increases and y decreases)



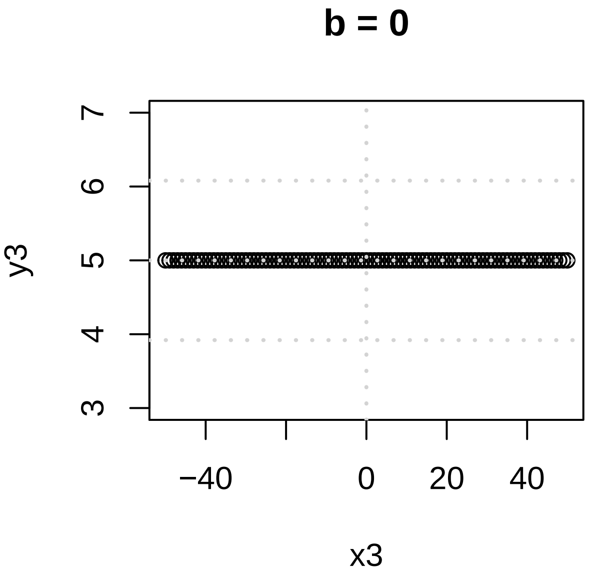
Analysis tools: scatterplot graph

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

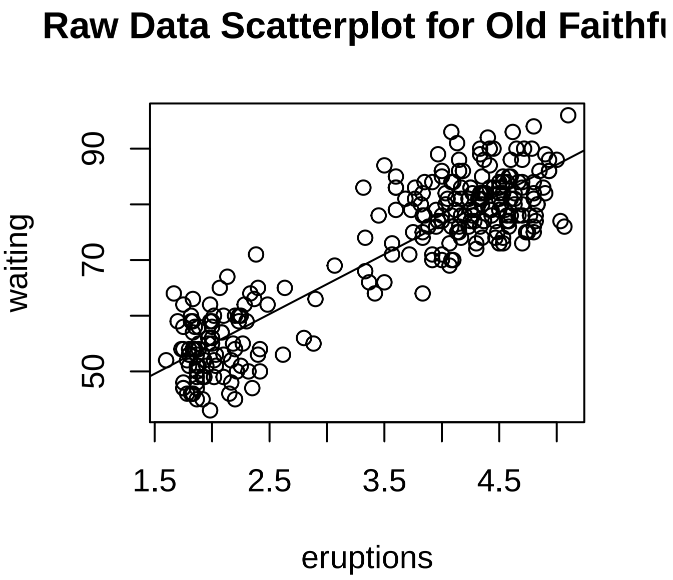


Analysis tools: scatterplot graph

This has 0 slope



Analysis tools: scatterplot graph with regression line



Slope and intercept formulas

Slope:

$$b = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{s_x^2(n-1)}$$

Intercept:

$$a = \overline{y} - b\overline{x}$$

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 ρ for the population correlation.

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

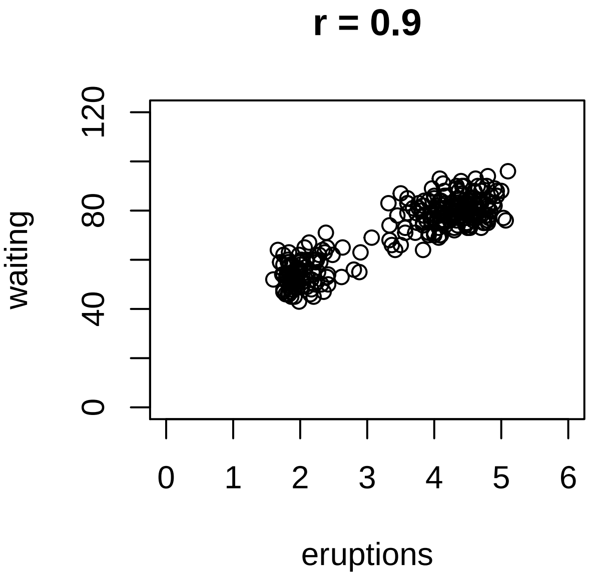
Coefficient of Determination, R^2

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

- It is the proportion (or ×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.

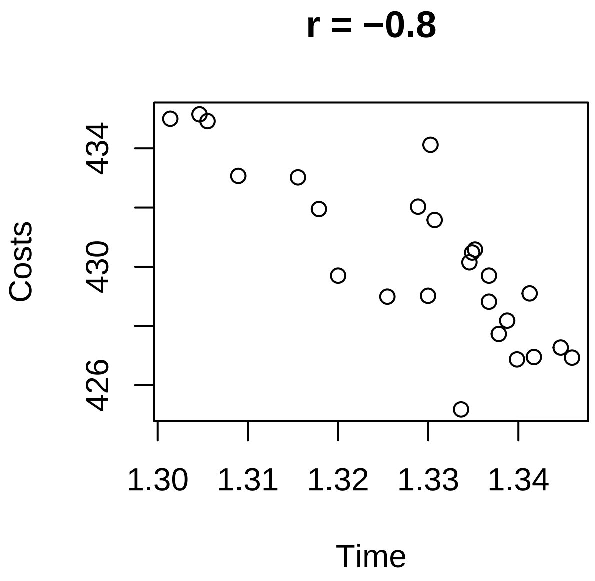
Analysis tools: scatterplot graph

Relatively strong, positive correlation



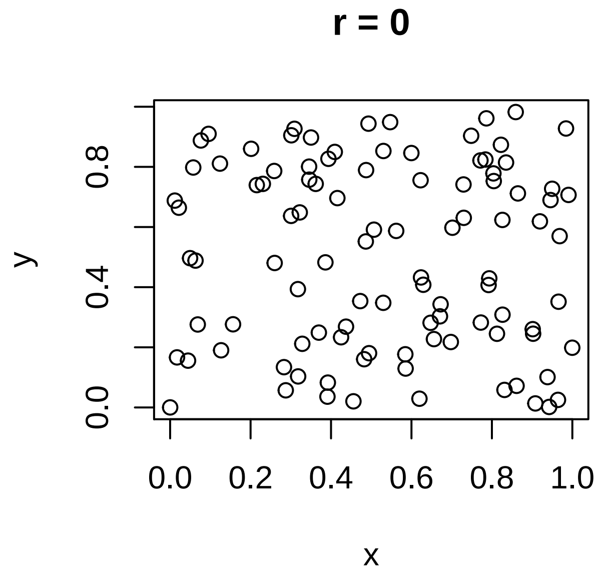
Analysis tools: scatterplot graph

Moderately strong, negative correlation



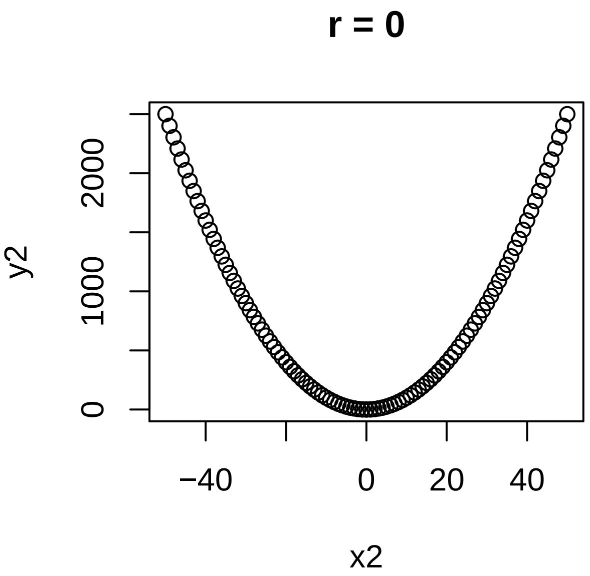
Analysis tools: scatterplot graph

No correlation



Analysis tools: scatterplot graph

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



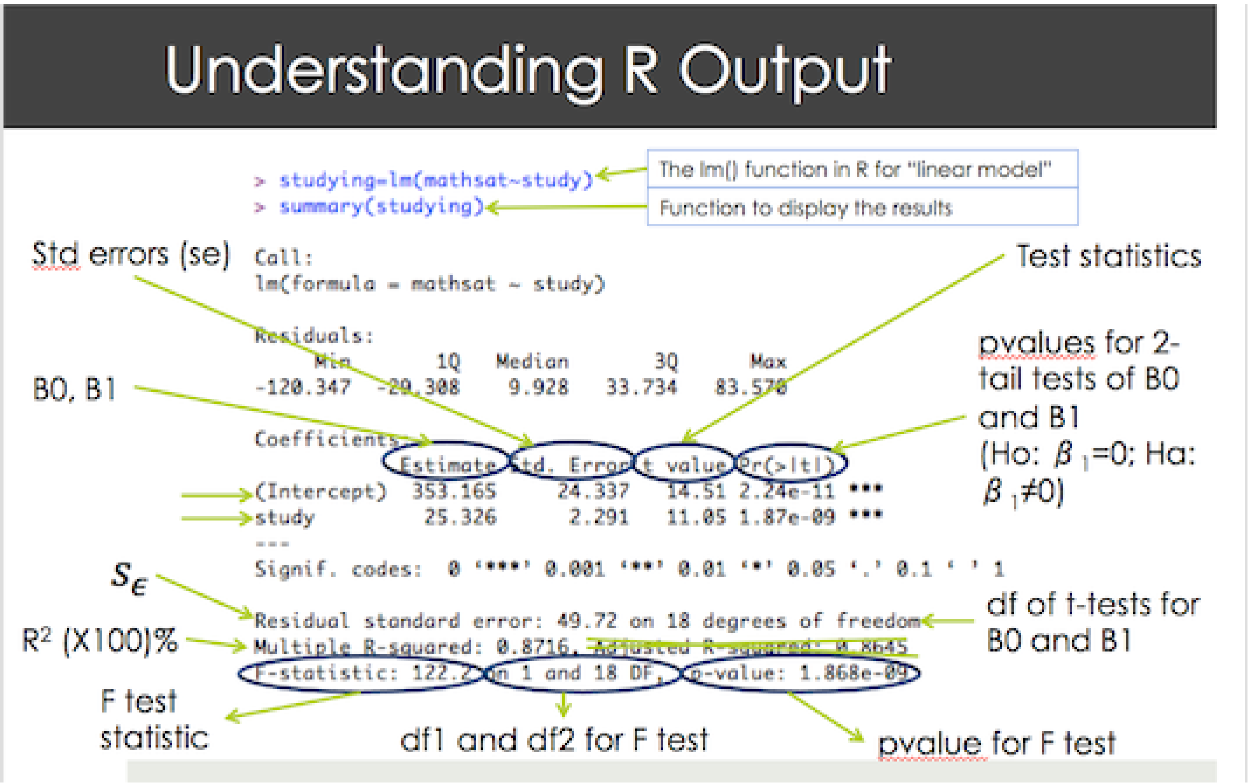
Old Faithful summary statistics

x is eruptions, y is waiting

```
             [,1]
sumxy 3787.985926
n      272.000000
xbar     3.487783
ybar    70.897059
s2x      1.302728
s2y    184.823312
sx       1.141371
sy      13.594974
```

Reading `R` output

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



Understanding R Output

```
> studying=lm(mathsat~study)
> summary(studying)

Call:
lm(formula = mathsat ~ study)

Residuals:
     Min       1Q   Median       3Q      Max 
-120.347  -29.308    9.928   33.734   83.578 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  353.165     24.337   14.51 2.24e-11 ***
study         25.326      2.291   11.05 1.87e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.72 on 18 degrees of freedom
Multiple R-squared: 0.8716, Adjusted R-squared: 0.8645 
F-statistic: 122.2 on 1 and 18 DF,  p-value: 1.868e-09
```

The lm() function in R for "linear model"
Function to display the results
Std errors (se)
Test statistics
B0, B1
pvalues for 2-tail tests of B0 and B1
(Ho: β_1=0; Ha: $\beta_1 \neq 0$)
S_ϵ
R² (X100)%
df of t-tests for B0 and B1
F test statistic
df1 and df2 for F test
pvalue for F test

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 (and usually is when discussing its results) easily

Old Faithful Output

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

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0796  -4.4831   0.2122   3.9246  15.9719 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  33.4744     1.1549   28.98   <2e-16 ***
eruptions    10.7296     0.3148   34.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.914 on 270 degrees of freedom
```

```
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

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)

$$\hat{y} = 33.47 + 10.73x$$

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

$$\hat{y}_{|x=2.283} = 33.474397 + 10.7296414 * 2.283 = 57.9701683$$

$$\hat{y}_{|x=5.1} = 33.474397 + 10.7296414 * 5.1 = 88.1955681$$

Calculating residuals, e_i

$$e_{|x=2.283} = 62 - 57.9701683 = 53.0213743$$

$$e_{|x=5.1} = 96 - 88.1955681 = 87.0213743$$

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

CIs for β_0, β_1

$$\hat{\beta}_j \pm t^{\star}(se_{\hat{\beta}_j})$$

Where $\hat{\beta}_j$ is either $\hat{\beta}_0$ (a) or $\hat{\beta}_1$ (b); same goes for the se, $t^{\star} = t_{\alpha/2,df}$ and $df = n - 2$ for both cases.

$$se_{\hat{\beta}_0} = \sqrt{s_\epsilon^2\left(\frac{1}{n} + \frac{\overline{x}^2}{s_x^2(n-1)}\right)} \qquad se_{\hat{\beta}_1} = \sqrt{\frac{s_\epsilon^2}{s_x^2(n-1)}}$$

$$s_\epsilon^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

All of these values are on the output, listed as `Std.Error` (for $se_{\hat{\beta}_0}$ and $se_{\hat{\beta}_1}$) and `Residual standard error` for s_ϵ

Hypothesis tests for the estimated slope (β_1) and intercept (β_0)

- Most often the slope b 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 (β_1 or b) and intercept (β_0 or a)

- 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 β_1, the slope

Hypotheses:

$$H_0 : \beta_1 = 0 \;\; vs. \;\; H_a : \beta_1 \neq 0$$

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 β_0, the intercept

Hypotheses:

$$H_0 : \beta_0 = 0 \;\; vs. \;\; H_a : \beta_0 \neq 0$$

Test Statistic:

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

- The $se_{\hat{\beta}_0}$ 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)

95% CI for β_0

$$\hat{\beta}_0 \pm t^{\star}(se_{\hat{\beta}_0})$$

$df = n - 2 = 272 - 2 = 270$ and $t^{\star} = t_{\alpha/2,df} = 1.969$

$$\hat{\beta}_0 \pm t^{\star}(se_{\hat{\beta}_0}) = 33.474397 \pm (1.969)(1.1548735) = 33.474397 \pm 2.273946$$

$$= 31.2004511, 35.748343$$

With 95% confidence the true y-intercept is between = 31.2004511 and 35.748343 minutes.

95% CI for β_1

$$\hat{\beta}_1 \pm t^{\star}(se_{\hat{\beta}_1})$$

$df = n - 2 = 272 - 2 = 270$ and $t^{\star} = t_{\alpha/2,df} = 1.969$

$$\hat{\beta}_1 \pm t^{\star}(se_{\hat{\beta}_1}) = 10.7296414 \pm (1.969)(0.3147534) = 10.7296414 \pm 0.6197495$$

$$= 10.1098919, 11.3493909$$

With 95% confidence the true slope is between 10.1098919 and 11.3493909 minutes.

Test for β_1

Hypotheses:

$$H_0 : \beta_1 = 0 \quad vs. \quad H_a : \beta_1 \neq 0$$

Test Statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{se_{\hat{\beta}_1}} = \frac{10.7296414 - 0}{0.3147534} = 34.0890399$$

H_0 can be rejected if $|t_{calc}| \geq t_{\alpha/2,df}$ where $df = n - 2$. $df = 272 - 2 = 270$ and $t_{\alpha/2,df} = 1.969$
Since $|34.0890399| \geq 1.969$, we reject H_0. The slope is significant (also means the relationship 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

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).

Since we are using output, r is calculated as $r = +\sqrt{R^2} = +\sqrt{0.8115} = 0.9$. There is a strong, positive linear relationship between eruptions and waiting of Old Faithful.

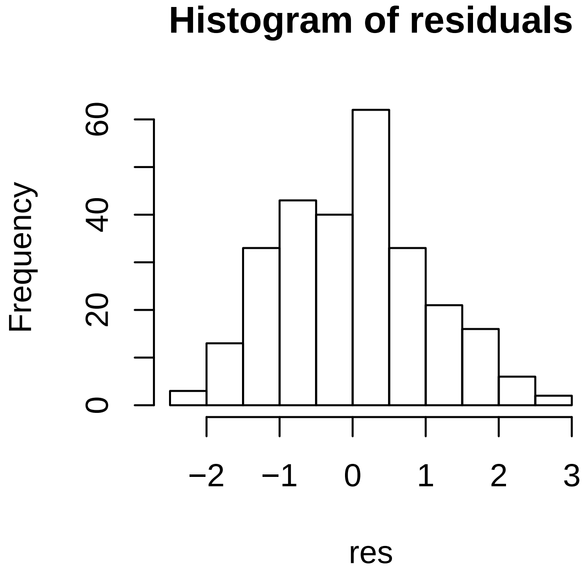
It is positive since the slope is positive (if $r > 0$ then $\beta_1 > 0$, if $r < 0$ then $\beta_1 < 0$, and vice versa)

Diagnostic plots used to check assumptions of slr

For checking assumptions, we need 3 graphs:
- Histogram of the residuals (#1,4)
- Scatterplot of residuals vs. predicted (#2)
- A normal probability plot, also called a QQ plot (#4)
- As for the 3^{rd} assumption, there is often no need to check it except in specific circumstances, so just assume it is met

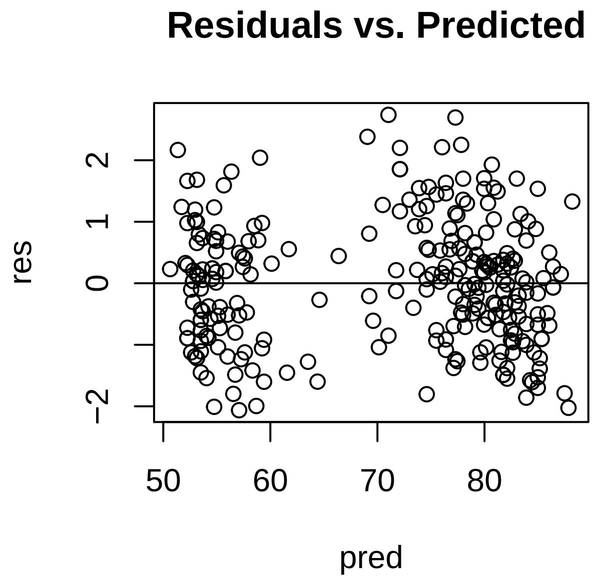
Assumption 1: $E(\epsilon_i) = 0$

Mean of the residuals is ≈ 0. For this, we look at a histogram of residuals to see if it is centered around zero (see if the histogram has the highest bar at zero)



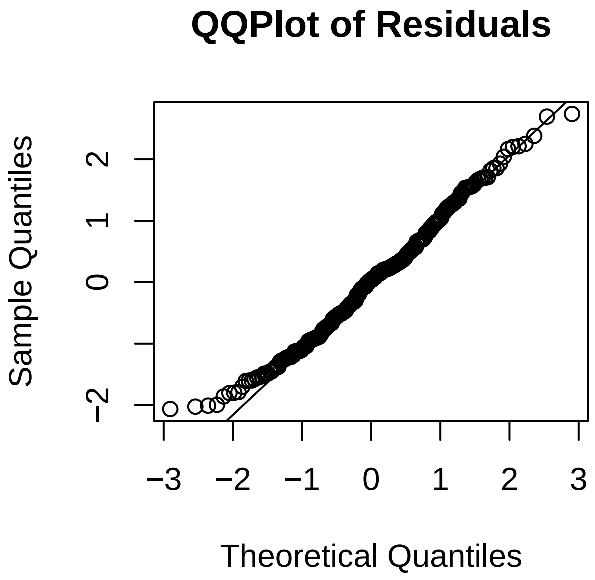
Assumption 2: $V(\epsilon_i) = \sigma^2_\epsilon$

The variance of the residuals is constant (the same) for all values of $\hat{y}$. The plot of x=predicted and y=residuals and it should have no discerable pattern (random scatter)



Assumption 4: $\epsilon_i \sim N(0, \sigma^2_\epsilon)$

Normality of residuals means that the histogram of residuals should be approximately symmetric/bell-shaped or that the QQplot (normal probability plot) shows that most points are along y=x line



Notation

- $\hat{\beta}_0$: sample intercept

- $\hat{\beta}_1$: sample slope
- $\hat{y}$: estimated value of y, called y-hat
- $\overline{x}$: mean of x values
- $\overline{y}$: mean of y values
- s_x^2, s_y^2: standard deviation of x and y values, respectively
- e_i: sample residual (estimate of ϵ_i), $e_i = y_i - \hat{y}_i$ (observed y-estimated y)
- $\hat{y} = a + bx$: regression equation
- s_ϵ^2: residual variance, variance of residuals
- s_ϵ: residual standard error, standard error of residuals