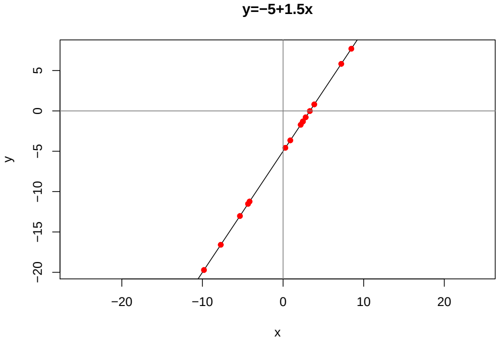
Linear Regression

Adriano Zanin Zambom [1]

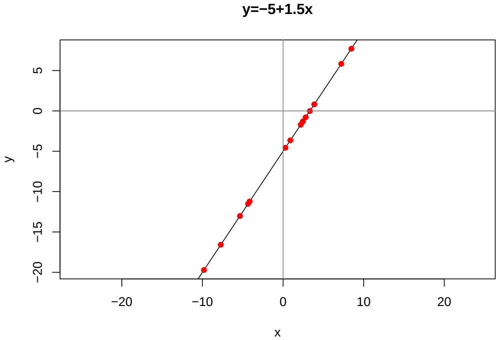
[1]Department of Mathematics
California State University Northridge

functional relationship



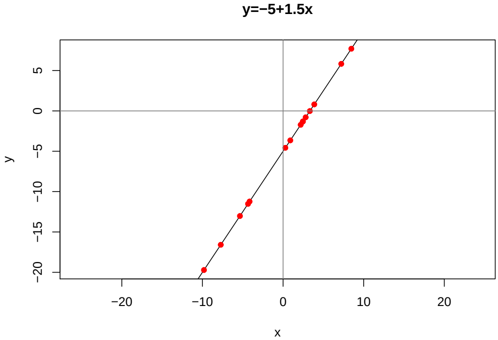
Notice that all observations fall directly on the line of the function. This is a functional relationship.

functional relationship



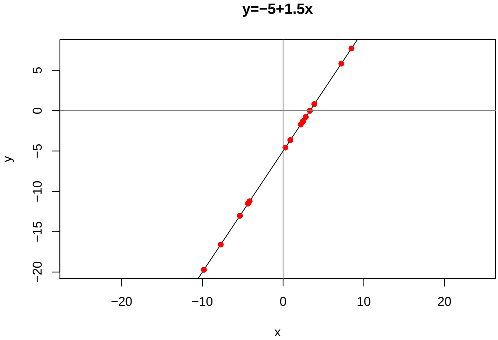
Notice that all observations fall directly on the line of the function.
This is a functional relationship.
What is the intercept?

functional relationship



Notice that all observations fall directly on the line of the function.
This is a functional relationship.
What is the intercept? -5
What is the slope?

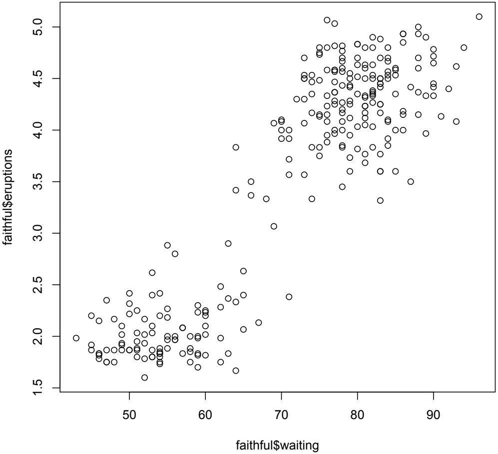
functional relationship



Notice that all observations fall directly on the line of the function.
This is a functional relationship.
What is the intercept? -5
What is the slope? 1.5

functional relationship

Real Example:
plot(faithful$waiting, faithful$eruptions)



Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i the i-th observed value of the response variable.

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i the i-th observed value of the response variable.
- X_i the i-th observed value of the predictor variable.

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i the i-th observed value of the response variable.
- X_i the i-th observed value of the predictor variable.
- β_0 and β_1 are parameters. β_0 is rerferred to as the intercept and β_1 is the slope.

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i the i-th observed value of the response variable.
- X_i the i-th observed value of the predictor variable.
- β_0 and β_1 are parameters. β_0 is rerferred to as the intercept and β_1 is the slope.
- ϵ_i is the error term.

Simple Linear Regression: Assumptions

- **Linearity**: The population regression line is straight; the relationship is linear.
- **Expected error is 0**: i.e. $E[\epsilon_i] = 0$ for all i. No observation is systematically too high or too low.
- **Constant Error**: i.e $Var[\epsilon_i] = \sigma^2$ for all i. The strength of the model is the same everywhere.
- **Uncorrelated errors**: Knowing the error of one observations gives no information about the size of any another error.

Note: Constant variance is called **homoscedasticity**.
Non-constant variance is called **heteroscedasticity**.

Simple Linear Regression: Assumptions

- More about the error term ϵ_i
- We expect the error term to be symmetric about 0
- Have bell shaped distribution

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We dont know β_0 and β_1

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We dont know β_0 and β_1
- We want to estimate them. How?

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We dont know β_0 and β_1
- We want to estimate them. How?
- Let's call the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We dont know β_0 and β_1
- We want to estimate them. How?
- Let's call the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$
- R will do it for us.
  lm(eruptions $\sim$ waiting, data $=$ faithful)
  In this example, $\hat{\beta}_0 = -1.87402$ and $\hat{\beta}_1 = 0.07563$

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We dont know β_0 and β_1
- We want to estimate them. How?
- Let's call the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$
- R will do it for us.
  lm(eruptions $\sim$ waiting, data = faithful)
  In this example, $\hat{\beta}_0 = -1.87402$ and $\hat{\beta}_1 = 0.07563$
  our estimated line is: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We dont know β_0 and β_1
- We want to estimate them. How?
- Let's call the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$
- R will do it for us.
  lm(eruptions $\sim$ waiting, data $=$ faithful)
  In this example, $\hat{\beta}_0 = -1.87402$ and $\hat{\beta}_1 = 0.07563$
  our estimated line is: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i = -1.87402 + 0.07563 X_i$

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We dont know β_0 and β_1
- We want to estimate them. How?
- Let's call the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$
- R will do it for us.
  lm(eruptions $\sim$ waiting, data $=$ faithful)
  In this example, $\hat{\beta}_0 = -1.87402$ and $\hat{\beta}_1 = 0.07563$
  our estimated line is: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i = -1.87402 + 0.07563 X_i$
- What's the predicted eruption time for a waiting time of 60min?

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We dont know β_0 and β_1
- We want to estimate them. How?
- Let's call the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$
- R will do it for us.
  lm(eruptions $\sim$ waiting, data = faithful)
  In this example, $\hat{\beta}_0 = -1.87402$ and $\hat{\beta}_1 = 0.07563$
  our estimated line is: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i = -1.87402 + 0.07563 X_i$
- What's the predicted eruption time for a waiting time of 60min? $\hat{Y} = -1.87402 + 0.07563 * 60 = 2.66378$

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- How to plot the estimated line?

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- How to plot the estimated line?
  plot(faithful$waiting, faithful$eruptions)
  fit = lm(eruptions ~ waiting, data = faithful)
  abline(fit)

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Let's look at the errors (residuals) ϵ_i

Simple Linear Regression: The Model

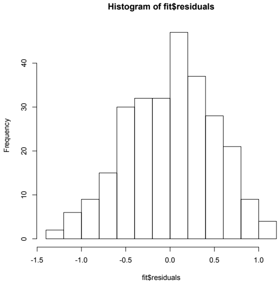
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Let's look at the errors (residuals) ϵ_i
  fit$residuals

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Let's look at the errors (residuals) ϵ_i
  fit$residuals
  hist(fit$residuals)



Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We want to test if X and Y have a significant relationship

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We want to test if X and Y have a significant relationship
- Write the hypothesis test as

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We want to test if X and Y have a significant relationship
- Write the hypothesis test as

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

- To run this test, called **hypothesis test**, we compute the **test statistic** "t" = ?

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We want to test if X and Y have a significant relationship
- Write the hypothesis test as

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

- To run this test, called **hypothesis test**, we compute the **test statistic** "t" = ?
  Associate with "t" comes a **p-value**. If the p-value is small, less than 0.05, we reject H_0

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- For the faithful example:
  fit $=$ lm(eruptions $\sim$ waiting, data $=$ faithful)
  summary(fit)

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- For the faithful example:
  fit = lm(eruptions $\sim$ waiting, data = faithful)
  summary(fit)

```
> summary(fit)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-1.29917 -0.37689  0.03508  0.34909  1.19329 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4965 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
```

Let's interpret this result

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- ```
  > summary(fit)

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

  Residuals:
       Min       1Q   Median       3Q      Max
  -1.29917 -0.37689  0.03508  0.34909  1.19329

  Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
  waiting      0.075628   0.002219   34.09   <2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 0.4965 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
  ```
- We can find $\hat{\beta}_1 = 0.075628$ and $\hat{\beta}_0 = -1.874016$
  ```

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

```
> summary(fit)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-1.29917 -0.37689  0.03508  0.34909  1.19329 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4965 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
```

- We can find $\hat{\beta}_1 = 0.075628$ and $\hat{\beta}_0 = -1.874016$
- We can see the summary of the residuals: Min, 1Q, ... Max

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- ```
> summary(fit)

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

Residuals:
     Min   1Q   Median   3Q   Max
-1.29917 -0.37689 0.03508 0.34909 1.19329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143 -11.70   <2e-16 ***
waiting   0.075628   0.002219   34.09   <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4965 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
```
- We can find $\hat{\beta}_1 = 0.075628$ and $\hat{\beta}_0 = -1.874016$
- We can see the summary of the residuals: Min, 1Q, ... Max
- The t-values and p-values ($Pr(> |t|)$) are in the last columns
```

Simple Linear Regression: The Model

- Your turn: Run a Regression with the cars dataset in R
- Plot the observations and the line
- Run the hypothesis test

Confidence intervals for regression coefficients

- A hypothesis test just tells us whether or not it is plausible that $\beta_1 = 0$

Confidence intervals for regression coefficients

- A hypothesis test just tells us whether or not it is plausible that $\beta_1 = 0$
- However, it does not give us a sense of the uncertainty of the magnitude of this association.

Confidence intervals for regression coefficients

- A hypothesis test just tells us whether or not it is plausible that $\beta_1 = 0$
- However, it does not give us a sense of the uncertainty of the magnitude of this association.
- A 95% confidence interval for β_1 is

$$\widehat{\beta}_1 \pm t^* s.e.(\widehat{\beta}_1)$$

  where t^* comes from the probability distribution and $s.e(\widehat{\beta}_1)$ is the standard error ("estimated standard deviation")

Confidence intervals for regression coefficients

- A hypothesis test just tells us whether or not it is plausible that $\beta_1 = 0$
- However, it does not give us a sense of the uncertainty of the magnitude of this association.
- A 95% confidence interval for β_1 is

$$\widehat{\beta}_1 \pm t^* s.e.(\widehat{\beta}_1)$$

where t^* comes from the probability distribution and $s.e(\widehat{\beta}_1)$ is the standard error ("estimated standard deviation")

fit = lm(eruptions ∼ waiting, data = faithful)
confint(fit)

Predicting

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Predicting a new observation
- Suppose we want to predict the eruption time for a waiting time of X = 73

Predicting

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Predicting a new observation
- Suppose we want to predict the eruption time for a waiting time of X = 73
- We can compute the predicted $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 = -1.87402 + 0.07563 * 73 = 3.64697$

Predicting

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Predicting a new observation
- Suppose we want to predict the eruption time for a waiting time of X = 73
- We can compute the predicted $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 = -1.87402 + 0.07563 * 73 = 3.64697$
- But R can give us a confidence interval for the prediction:
  new ¡- data.frame(waiting=73)
  predict(fit,new,interval="confidence")

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Another way to test the hypothesis

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

- Is to use the F-test

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Another way to test the hypothesis

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

- Is to use the F-test
  Associate with "F" comes a **p-value**. If the p-value is small, less than 0.05, we reject H_0

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- To run this test in R we use
  fit = lm(eruptions ~ waiting, data = faithful)
  anova(fit)

Simple Linear Regression: The Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- To run this test in R we use
  fit = lm(eruptions ~ waiting, data = faithful)
  anova(fit)

```
Analysis of Variance Table

Response: eruptions
           Df  Sum Sq Mean Sq F value    Pr(>F)
waiting     1 286.478 286.478  1162.1 < 2.2e-16 ***
Residuals 270  66.562   0.247
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Simple Linear Regression: The Model

- The ANOVA table

Simple Linear Regression: The Model

- The ANOVA table
- df = degrees of freedom
- SS = Sum of Squares
- Note that SST = SSR + SSE

Simple Linear Regression: The Model

- What is the link between regression and correlation?
- It can be shown that the slope $\beta_1 = \frac{sd(y)}{sd(x)} r$

Simple Linear Regression: The Model

- What is the link between regression and correlation?
- It can be shown that the slope $\beta_1 = \frac{sd(y)}{sd(x)} r$
- So we can estimate the correlation coefficient with the slope and vice versa.

Simple Linear Regression: The Model

- What is the link between regression and correlation?
- It can be shown that the slope $\beta_1 = \frac{sd(y)}{sd(x)} r$
- So we can estimate the correlation coefficient with the slope and vice versa.
- Exercise: Estimate the correlation coefficient of waiting and eruptions using the slope and the sd of each variable.

Simple Linear Regression: The Model

- How do I know if the model is good?

Simple Linear Regression: The Model

- How do I know if the model is good? There are several ways.

Simple Linear Regression: The Model

- How do I know if the model is good? There are several ways.
- R^2, the **coefficient of determination**, is used to evaluate how well a model actually fits the observed data.

Simple Linear Regression: The Model

- How do I know if the model is good? There are several ways.
- R^2, the **coefficient of determination**, is used to evaluate how well a model actually fits the observed data.
- R^2 can be interpreted as the proportion of variability among observed values of Y that is explained by the linear regression of Y on X

Simple Linear Regression: The Model

- How do I know if the model is good? There are several ways.
- R^2, the **coefficient of determination**, is used to evaluate how well a model actually fits the observed data.
- R^2 can be interpreted as the proportion of variability among observed values of Y that is explained by the linear regression of Y on X

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

Simple Linear Regression: The Model

- How do I know if the model is good? There are several ways.
- R^2, the **coefficient of determination**, is used to evaluate how well a model actually fits the observed data.
- R^2 can be interpreted as the proportion of variability among observed values of Y that is explained by the linear regression of Y on X

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

$$0 \leq R^2 \leq 1$$

- The closer to 1 R^2 is, the more variability is explained by your model (when linear)

Simple Linear Regression: The Model

- R^2 in R:
  fit = lm(eruptions ~ waiting, data = faithful)
  summary(fit)
  "Multiple R-squared: 0.8115"

Simple Linear Regression: The Model

- 3 things to never forget:

Simple Linear Regression: The Model

- 3 things to never forget:

  1. Regression models are only interpretable over the range of observed data

Simple Linear Regression: The Model

- 3 things to never forget:

1. Regression models are only interpretable over the range of observed data

2. Regression models relate to association, not causality

Simple Linear Regression: The Model

- 3 things to never forget:

  1. Regression models are only interpretable over the range of observed data

  2. Regression models relate to association, not causality

  3. If the plot of the data does not look linear, then we need to find another (non-linear) model

Simple Linear Regression: The Model

- Example

  Nonlinearity of the regression function