

Introducing..... Linear Regression!

Frank Edwards

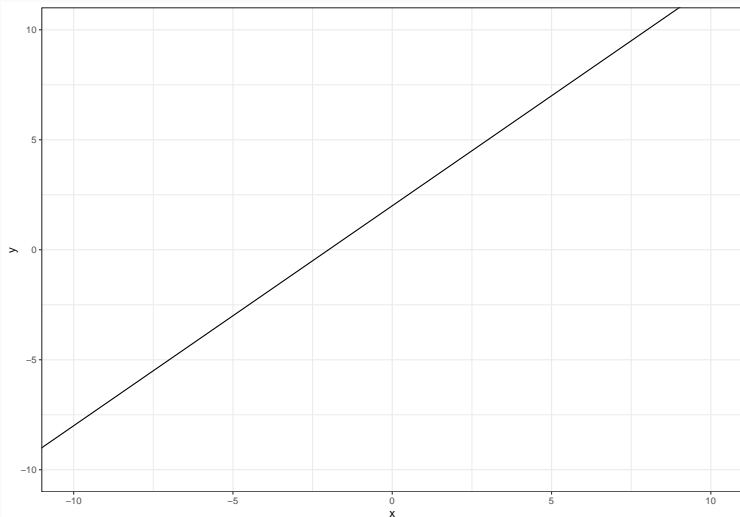
We can define a line as:

$$y = mx + b$$

Where m is the slope and b is the y-intercept.

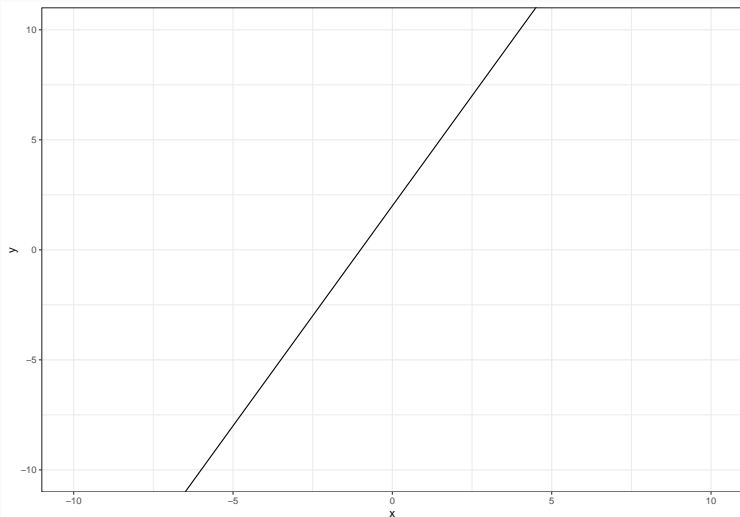
Lines: slopes

$$y = x + 2$$



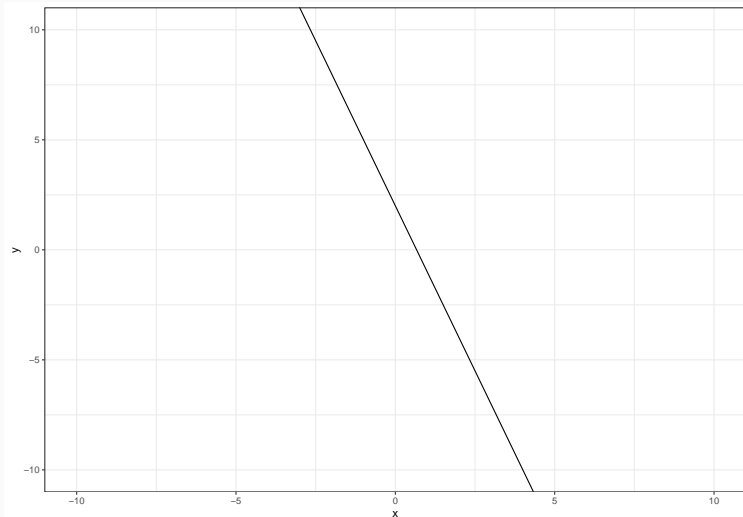
Lines: slopes

$$y = 2x + 2$$



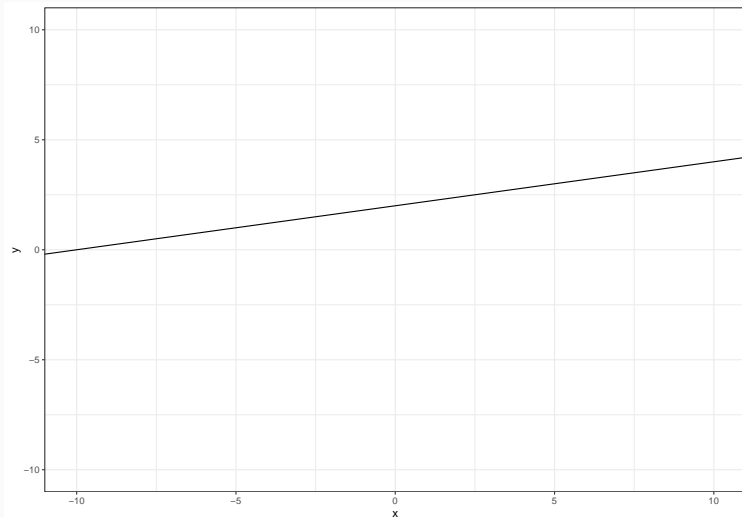
Lines: slopes

$$y = -2x + 2$$



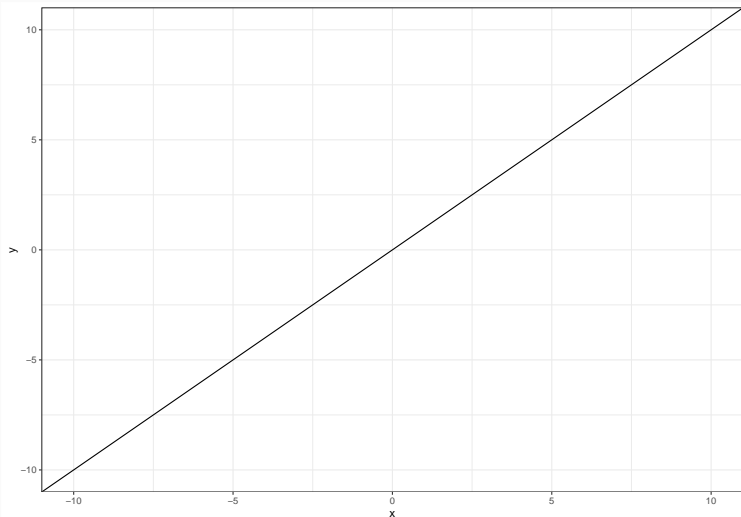
Lines: slopes

$$y = 0.2x + 2$$



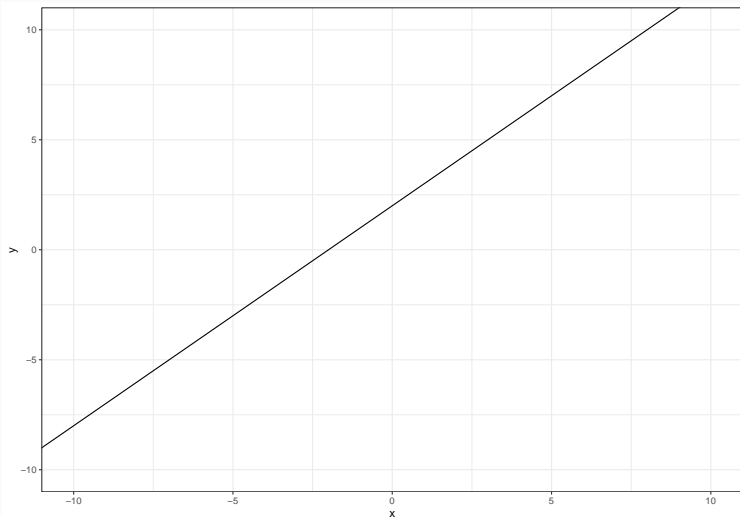
Lines: intercepts

$$y = x + 0$$



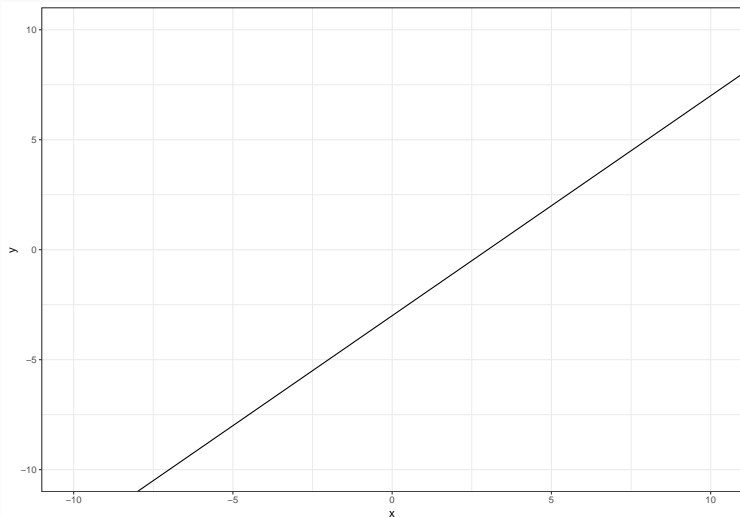
Lines: intercepts

$$y = x + 2$$



Lines: intercepts

$$y = x - 3$$



The linear regression model

We can describe the relationship between a predictor variable X and an outcome variable Y with a line:

$$y = mx + b$$

What does m describe?

The linear regression model

We can describe the relationship between a predictor variable X and an outcome variable Y with a line:

$$y = mx + b$$

What does m describe?

- The increase in y for a one-unit increase in x

What does b describe

The linear regression model

We can describe the relationship between a predictor variable X and an outcome variable Y with a line:

$$y = mx + b$$

What does m describe?

- The increase in y for a one-unit increase in x

What does b describe

- The location of y when $x = 0$

The linear regression model: expected value

We can describe the relationship between a predictor variable X and the expected value E of an outcome variable Y with the line:

$$E[Y] = \beta_0 + \beta_1 X$$

The linear regression model: expected value

We can describe the relationship between a predictor variable X and the expected value E of an outcome variable Y with the line:

$$E[Y] = \beta_0 + \beta_1 X$$

What does β_0 describe?

The linear regression model: expected value

We can describe the relationship between a predictor variable X and the expected value E of an outcome variable Y with the line:

$$E[Y] = \beta_0 + \beta_1 X$$

What does β_0 describe?

What does β_1 describe?

The error term in linear regression

We can describe the relationship between a predictor variable X and an outcome variable Y with the line:

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

Where β_0 is the y-intercept of the line, β_1 is the slope of the line, and ε is the error between the fitted line and the coordinates (X, Y)

The linear regression model

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

β_0 : The value of y when x is equal to zero

The linear regression model

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

β_0 : The value of y when x is equal to zero

β_1 : The average increase in y when x increases by one unit

The linear regression model

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

β_0 : The value of y when x is equal to zero

β_1 : The average increase in y when x increases by one unit

ε : The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of y .

The linear regression model

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

β_0 : The value of y when x is equal to zero

β_1 : The average increase in y when x increases by one unit

ε : The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of y .

The line $y = \beta_0 + \beta_1 X$ provides an expected value for y based on the values of x .

- If $\beta_0 = 2$ and $\beta_1 = 1.5$, what is the expected value of y when $x = 4$?

The linear regression model as a prediction engine

- If $\beta_0 = 2$ and $\beta_1 = 1.5$, what is the expected value of y when $x = 4$?
- When $x = 2$?

The linear regression model and prediction

We put a *hat* on variables to indicate that they are estimated from the data.

A regression line tells us to expect values \hat{Y} with the equation:

$$\hat{Y} = \beta_0 + \beta_1 X$$

and the residual, or prediction error is the difference between the observed and predicted values of Y

$$\varepsilon = Y_{obs} - \hat{Y}$$

Understanding the regression line for real data

```
## # A tibble: 10 x 2
##       x       y
##   <dbl> <dbl>
## 1 -0.311 -0.237
## 2  0.932  0.883
## 3  1.04   1.62
## 4  3.55   3.41
## 5  0.507  0.254
## 6  0.655  1.11
## 7  1.39   1.29
## 8  0.868  0.477
## 9  0.859  0.561
## 10 0.509  0.626
```

$$\beta_0 = 0.05, \beta_1 = 0.95$$

- Estimate \hat{Y} . Recall that $\hat{Y} = \beta_0 + \beta_1 X$

Understanding the regression line for real data

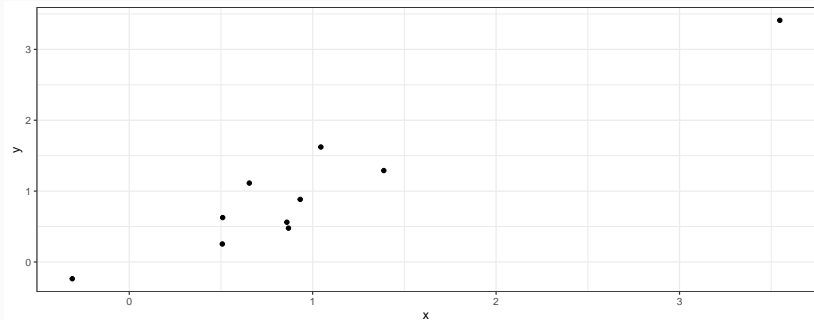
```
## # A tibble: 10 x 2
##       x       y
##   <dbl> <dbl>
## 1 -0.311 -0.237
## 2  0.932  0.883
## 3  1.04   1.62
## 4  3.55   3.41
## 5  0.507  0.254
## 6  0.655  1.11
## 7  1.39   1.29
## 8  0.868  0.477
## 9  0.859  0.561
## 10 0.509  0.626
```

$$\beta_0 = 0.05, \beta_1 = 0.95$$

- Estimate \hat{Y} . Recall that $\hat{Y} = \beta_0 + \beta_1 X$
- Estimate ε . Recall that $\varepsilon = Y - \hat{Y}$

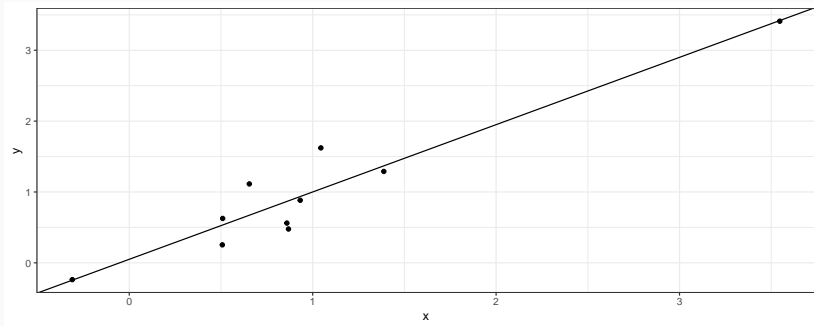
Understanding the regression line

$$\beta_0 = 0.05, \beta_1 = 0.95$$



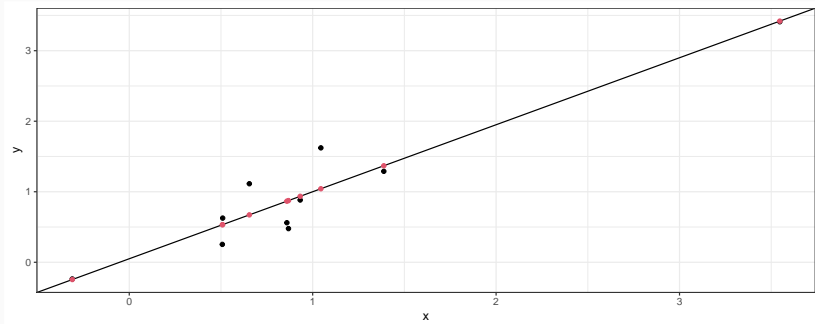
Understanding the regression line: adding the fit

$$\beta_0 = 0.05, \beta_1 = 0.95$$



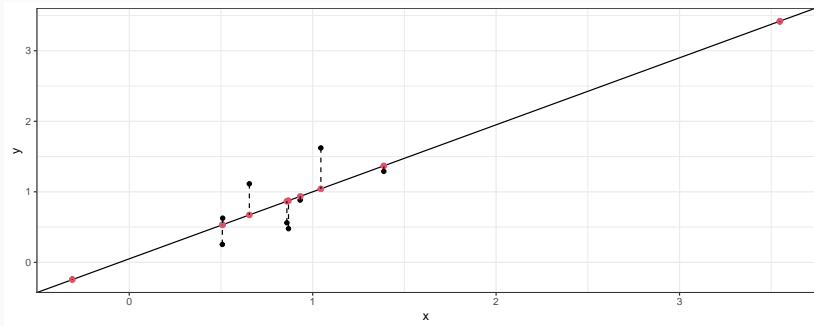
Understanding the regression line: adding \hat{y}

$$\beta_0 = 0.05, \beta_1 = 0.95$$



Understanding the regression line: adding ε

$$\beta_0 = 0.05, \beta_1 = 0.95$$



Ordinary least squares

We usually fit a linear regression using a method called *ordinary least squares*, or OLS.

Ordinary least squares

We usually fit a linear regression using a method called *ordinary least squares*, or OLS.

- This method seeks to minimize the distance between \hat{Y} and Y .

Ordinary least squares

We usually fit a linear regression using a method called *ordinary least squares*, or OLS.

- This method seeks to minimize the distance between \hat{Y} and Y .
- To do so, we minimize the sum of squared residuals (SSR)

Ordinary least squares

We usually fit a linear regression using a method called *ordinary least squares*, or OLS.

- This method seeks to minimize the distance between \hat{Y} and Y .
- To do so, we minimize the sum of squared residuals (SSR)

In other words, we solve for the values of β_0 and β_1 that result in the smallest possible value for SSR

Ordinary least squares

We usually fit a linear regression using a method called *ordinary least squares*, or OLS.

- This method seeks to minimize the distance between \hat{Y} and Y .
- To do so, we minimize the sum of squared residuals (SSR)

In other words, we solve for the values of β_0 and β_1 that result in the smallest possible value for SSR

$$\text{SSR} = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x)^2$$

We can solve for the values of β that minimize the sum of squares using linear algebra (R does this for us).

$$\beta = (X^T X)^{-1} X^T y$$

Remember that T is a matrix transpose and -1 indicates an inverse

Let's use `iris` to estimate this regression model:

$$\text{Sepal.Length} = \beta_0 + \beta_1 \text{Petal.Length} + \varepsilon$$

What does this equation say?

Estimating a model in R

We use the `lm` function to estimate linear regressions.

```
my_cool_model <- lm(Sepal.Length ~ Petal.Length,  
                    data = iris)
```

Note that `~` separates the left and right side of the model (where outcomes are on the left, predictors on the right). This is called formula notation in R

Model output

We use `summary()` to extract model output

```
summary(my_cool_model)

##
## Call:
## lm(formula = Sepal.Length ~ Petal.Length, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24675 -0.29657 -0.01515  0.27676  1.00269
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.30660     0.07839   54.94  <2e-16 ***
## Petal.Length   0.40892     0.01889   21.65  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4071 on 148 degrees of freedom
## Multiple R-squared:  0.76, Adjusted R-squared:  0.7583
## F-statistic: 468.6 on 1 and 148 DF, p-value: < 2.2e-16
```

For now `coef()` is easier to work with

```
coef(my_cool_model)
```

```
## (Intercept) Petal.Length  
##      4.3066034      0.4089223
```

What does this model tell us?

$$\cdot \beta_0 = 4.3$$

What does this model tell us?

- $\beta_0 = 4.3$
- $\beta_1 = 0.4$

What does this model tell us?

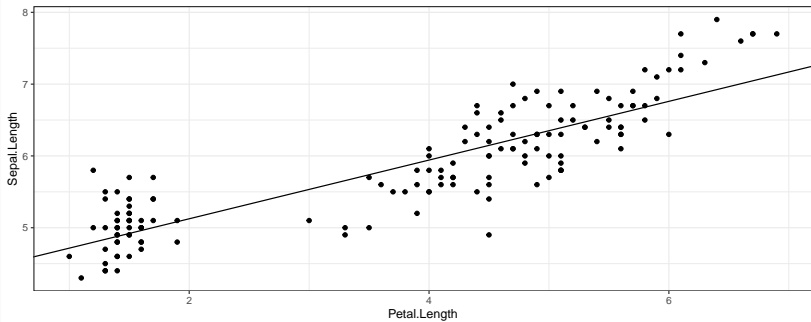
- $\beta_0 = 4.3$
- $\beta_1 = 0.4$

$$E(\text{Sepal.Length}) = \beta_0 + \beta_1 \text{Petal.Length}$$

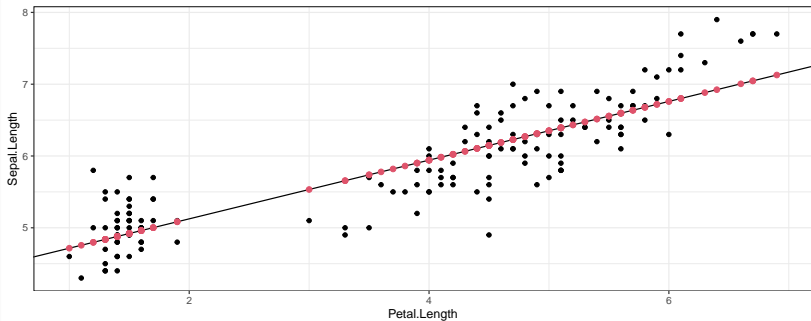
Visualized: observed data



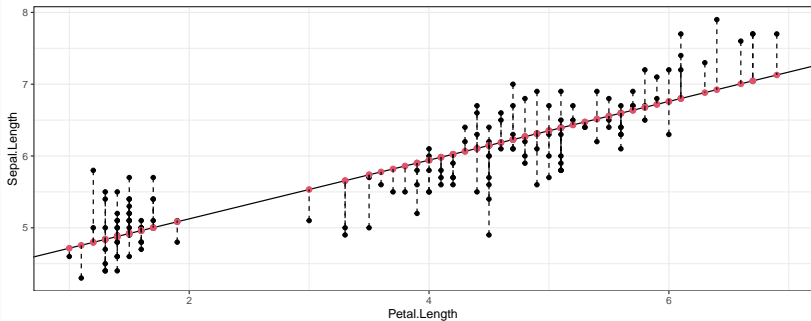
Visualized: regression line



Visualized: expected values (\hat{y})

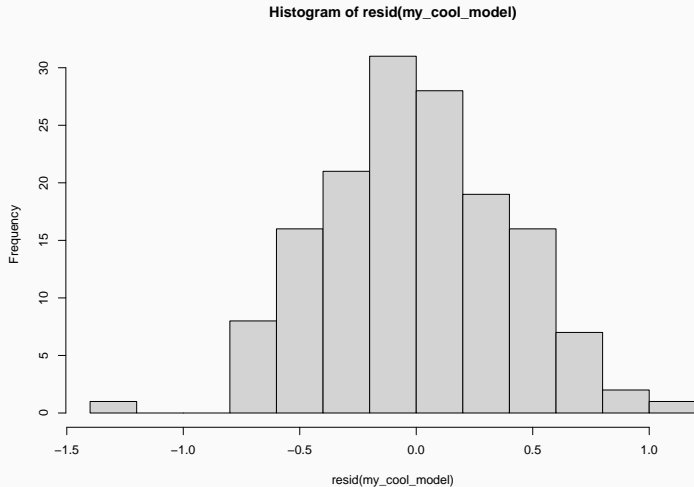


Visualized: residuals (epsilon)



Evaluating the distribution of our residuals

```
hist(resid(my_cool_model))
```



Practice with estimation,
interpretation on homework
