Linear Regression

Owen G. Ward

2021-06-07

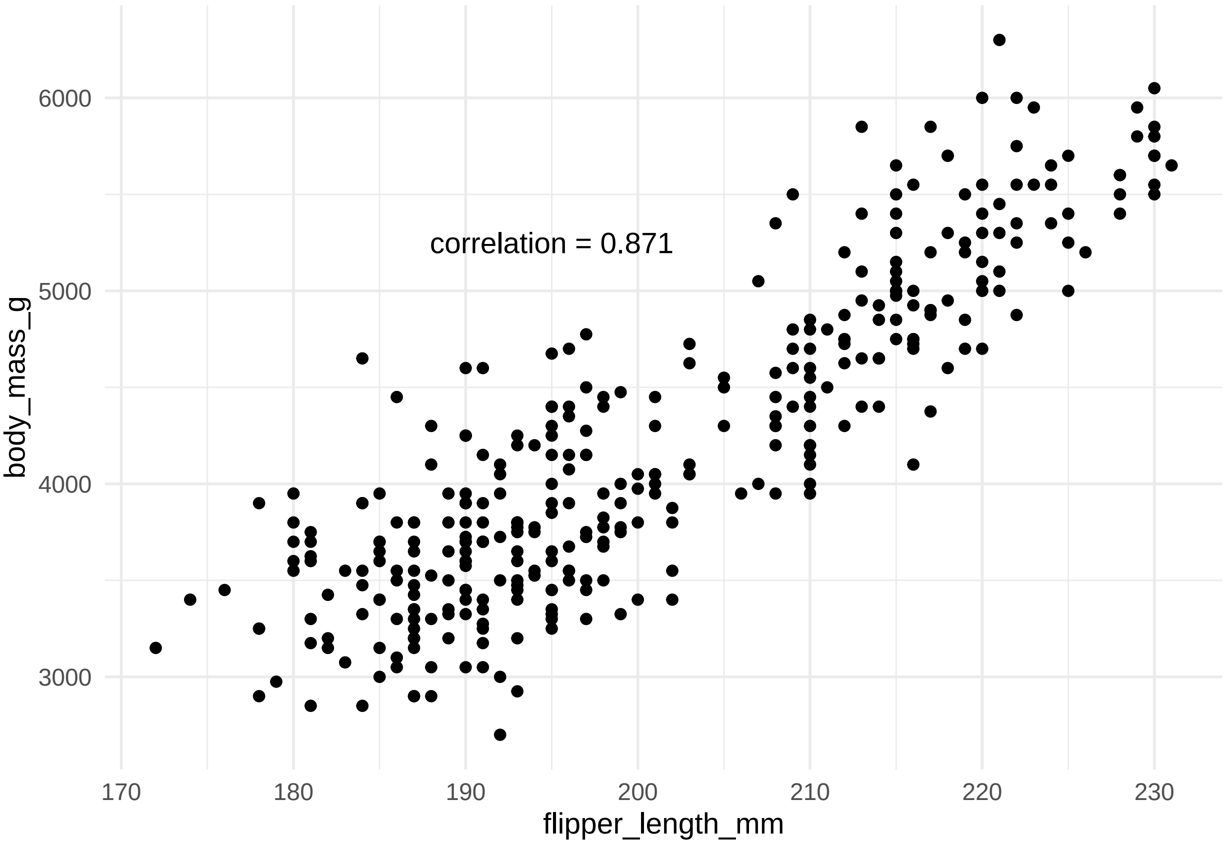
Setup

Very often we have multiple variables and we want to understand the relationship between these variables.

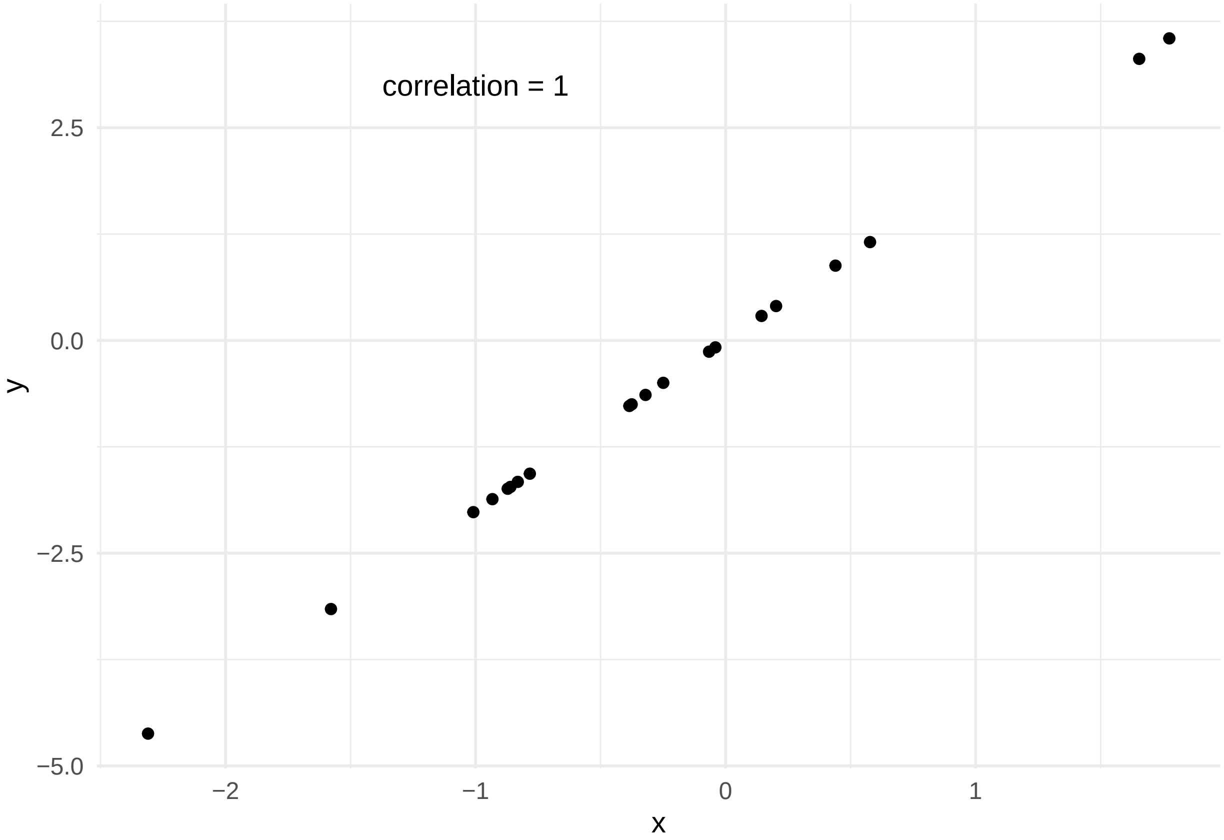
Perhaps surprisingly, a line can often describe the relationship quite well, if not exactly.

This is related to the idea of correlation we saw back in Chapter 1.

Example



If the correlation is exactly 1, then a straight line fits the data perfectly. But that is rarely ever the case with real data.



We want to fit the best line possible in the case where the data is not perfectly linear but a linear line is a good approximation.

No single line can be used for the real data above.

Have to define what we mean by a "good" fit.

To estimate a line we need to estimate the intercept and the slope.

Linear Regression Equation

Linear regression gives an equation of the form

$$y = \beta_0 + \beta_1 x + \epsilon.$$

Here β_0 is the intercept of the line, the y value of the line when $x = 0$.

β_1 is the slope of the line.

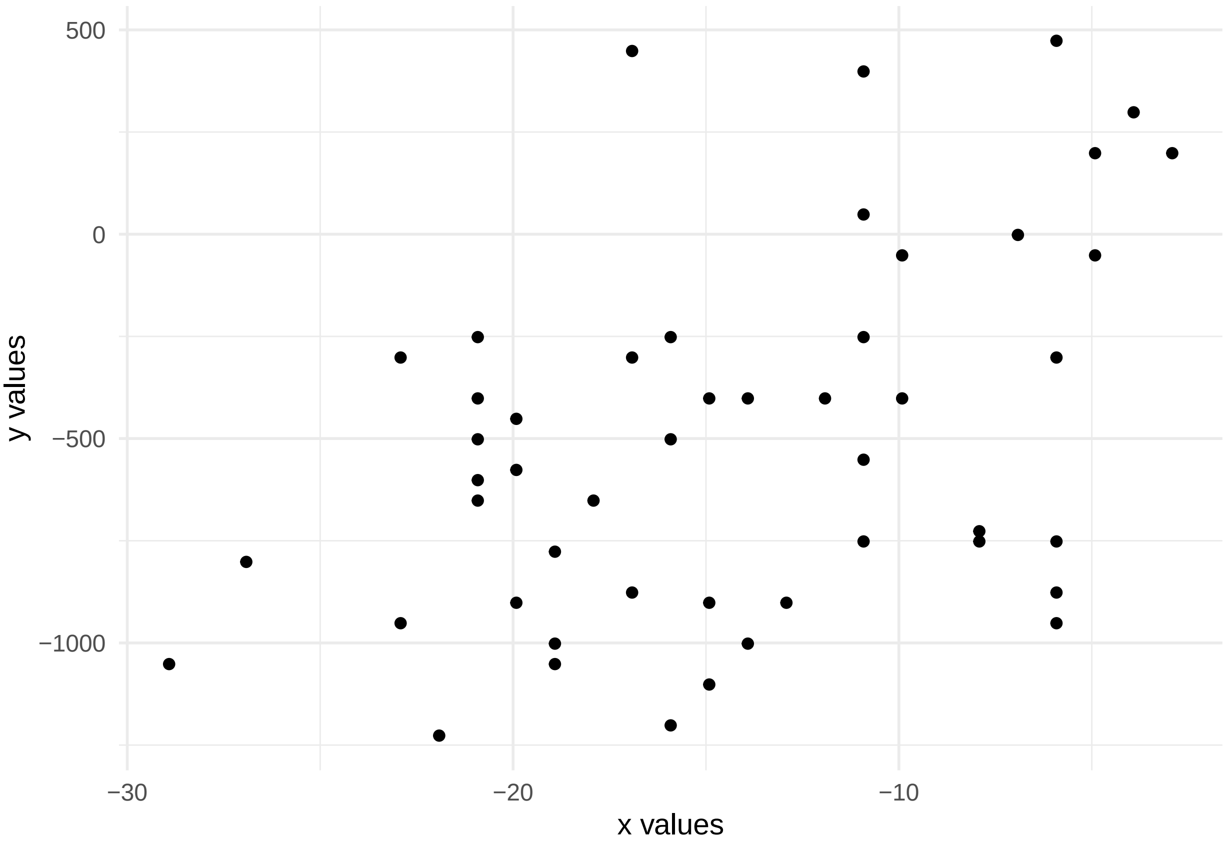
ϵ is the error of the line, the distance from the line to each point.

Linear Regression Equation

$$y = \beta_0 + \beta_1 x + \epsilon.$$

We call x the explanatory or predictor variable.

We call y the response variable.



Estimating the line

Suppose we have a line which we use to model the data, which we will call

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

What this says is that for each x_i our line gives an estimate of y_i which is the corresponding y coordinate on the line. This is called the **fitted value**

The line does not need to go through any of the actual points so

$$\hat{y}_i \neq y_i$$

We say the distance between the fitted point and the line is the **residual**

$$e_i = y_i - \hat{y}_i.$$

Parameter Estimation

We said we needed to define how to choose a “good” line.

It seems reasonable to want the residuals to be small.

We will choose to minimise the squared sum of the residuals, $Q = \sum_{i=1}^{n} e_i^2$.

Could also minimize $\sum_{i=1}^{n} |e_i|$, say, but squared sum widely used, easier to do.

Conditions

Generally require 4 conditions to be true before fitting linear regression:

- There is some linear trend in the rather, not some non linearity
- The pairs (x_i, y_i) are independent.
- The residuals are approximately normally distributed with mean 0. We run into problems if there are some residuals very far from 0.
- The variance of the residuals should not change as x changes.

We want to keep these conditions in mind every time we fit a regression.

Minimization

So want to find the values of β_0, β_1 which minimise

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} e_i^2 = \min_{\beta_0,\beta_1} Q(\beta_0, \beta_1) = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

To do this we differentiate with respect to each parameter and set th derivative equal to 0.

Estimating β_0

We differentiate with respect to β_0

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i).$$

Then solve for β_0.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Estimating β_1

We do the same process for estimating β_1.

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)$$

Then solve for β_1 when the derivative is 0.

Alternative Formulation

We have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i y_i - \bar{y}x_i)}{\sum_{i=1}^{n}(x_i^2 - \bar{x}x_i)}.$$

We can rewrite this in a more useful format.

Recall that

$$cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

and

$$\sigma_x^2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Want to rewrite $\hat{\beta}_1$ using these.

We first show that

$$\sum_{i=1}^{n}(x_i y_i - \bar{y}x_i) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Similarly, we show that

$$\sum_{i=1}^{n}(x_i^2 - \bar{x}x_i) = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Putting these together

That means that we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

If we divide both by n we get

$$\hat{\beta}_1 = \frac{\sigma_{x,y}}{\sigma_x^2},$$

the covariance divided by the sample variance of x.

Using $corr_{x,y} = \frac{\sigma_{x,y}}{\sigma_x, \sigma_y}$ we get

$$\hat{\beta}_1 = corr_{x,y} \frac{\sigma_y}{\sigma_x},$$

the product of the sample correlation and the sample standard deviation of y divided by the sample standard deviation of x.

This makes sense, higher values of correlation results in larger slopes in the fitted line.

Example

If we wanted to fit a regression to the penguin data then x is the flipper length and y is the body mass.

For this we can get the coefficients using the sample means and standard deviations $\bar{x} = 200.915$ $\bar{y} = 4201.754$ $\sigma_x = 14.062$, $\sigma_y = 801.955$ and $corr_{x,y} = 0.871$.

Properties of these estimators

After fitting this regression we can show that:

- $\sum_{i=1}^n e_i = 0$, the sum of the fitted residuals when the model includes an intercept term.
- The regression line will go through the point $(\bar{x}, \bar{y})$, if there is an intercept term.
- $\sum_{i=1}^n x_i e_i = 0$.

Unbiased

Using that $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$ can also show that the estimates for both the intercept and slope are unbiased,

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \qquad \mathbb{E}(\hat{\beta}_1) = \beta_1$$

Interpretation

We said previously that we want the the residuals to be normally distributed and most importantly, to have expected value 0.

This means that if we take the expectation of

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

that

If the fitted residuals e_i also have this property then we get that

$$\mathbb{E}(\hat{y}_i) = \beta_0 + \beta_1 x_i.$$

Suppose we want to see what the expected value of y will be at some new x point x_{new}, which is 1 unit greater than x_i?

If $x_{new} = x_i + 1$ then the expected value of y_{new} will be

$$\mathbb{E}(y_{new}) = \beta_0 + \beta_1 x_{new} = \beta_0 + \beta_1(x_i + 1)$$

This will be the expected value at y_i plus β_1,

$$\mathbb{E}(y_{new}) = \mathbb{E}(y_i) + \beta_1, \quad \text{if } x_{new} = x_i + 1.$$

Definition

The slope coefficient β_1 gives **the change in the average value of y as we increase the x value by one unit**.

Similarly, the intercept term β_0 gives the expected average value of y when $x = 0$, if the model is reasonable there (will see more next).

Example

For the penguin example we get

$$\hat{\beta}_0 = -5780, \qquad \hat{\beta}_1 = 49.686.$$

The intercept term not really interpretable, can't have a flipper with 0 length.

$\hat{\beta}_1 = 49.686$ means that as we increase the flipper length by 1mm, we expect the weight of penguins to increase by 49.69 grams, on average.

Example

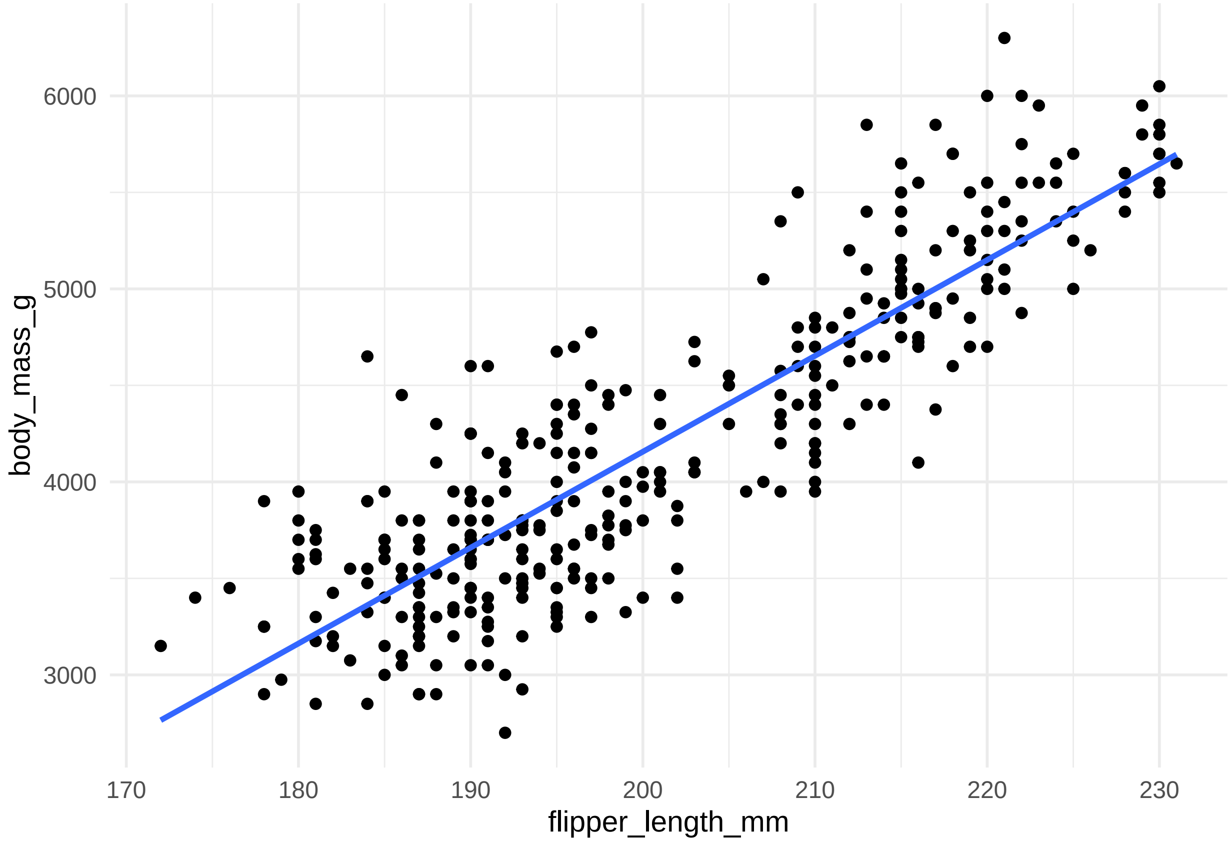
We can also use the regression coefficient estimates to get the expected mean value of the y variable at a specific x value.

The linear regression model says the expected value of y for a given x is given by

$$\mathbb{E}(y) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Using our linear regression fit, the expected body mass for a penguin with flipper length 200mm is $-5780 + 49.69(200) = 4157.1$ grams.

Fitted Regression Line



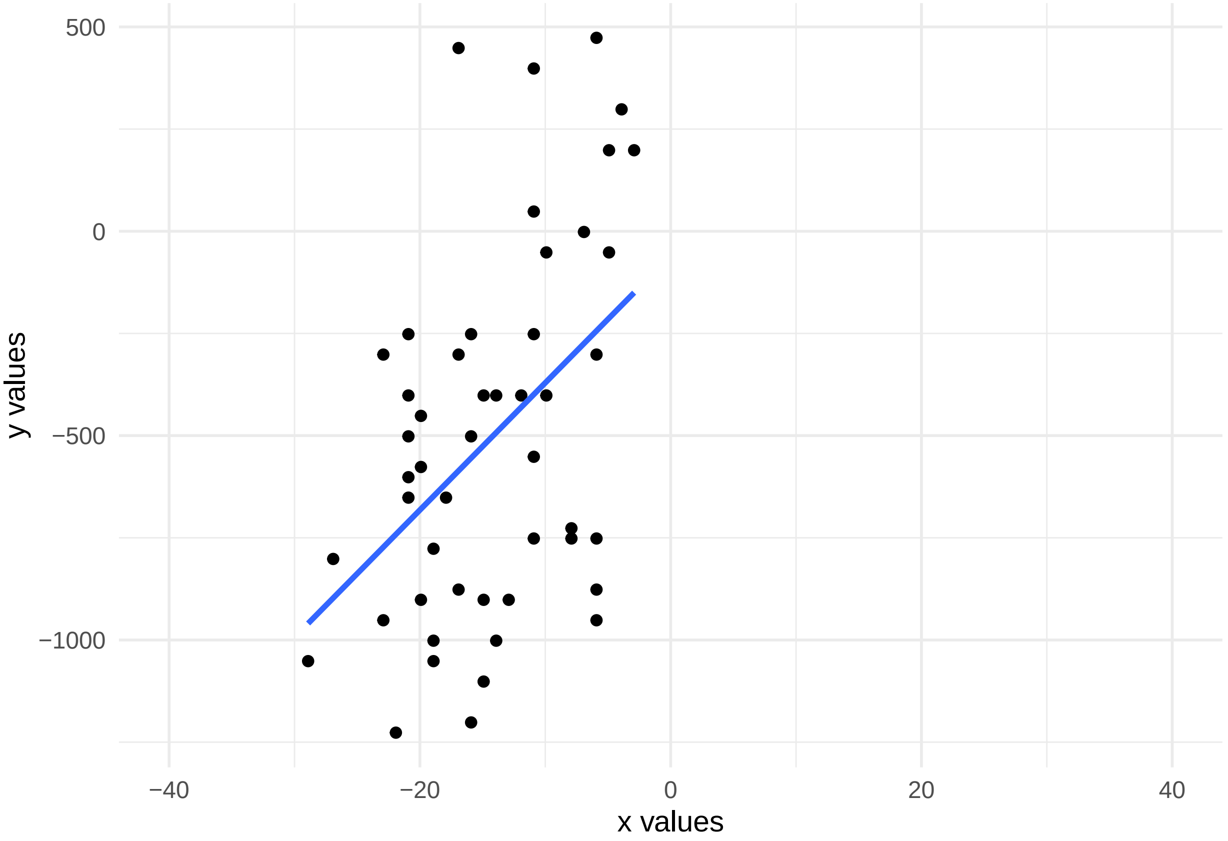
Dangers of Extrapolation

While linear regression works well, it is somewhat dependent on the data you have.

Only tells you a linear model is reasonable in the range of values of initial data. No reason to think its also valid as we move far away from those points.

Some software helps with this, wont plot the fitted line where there isn't data.

Example



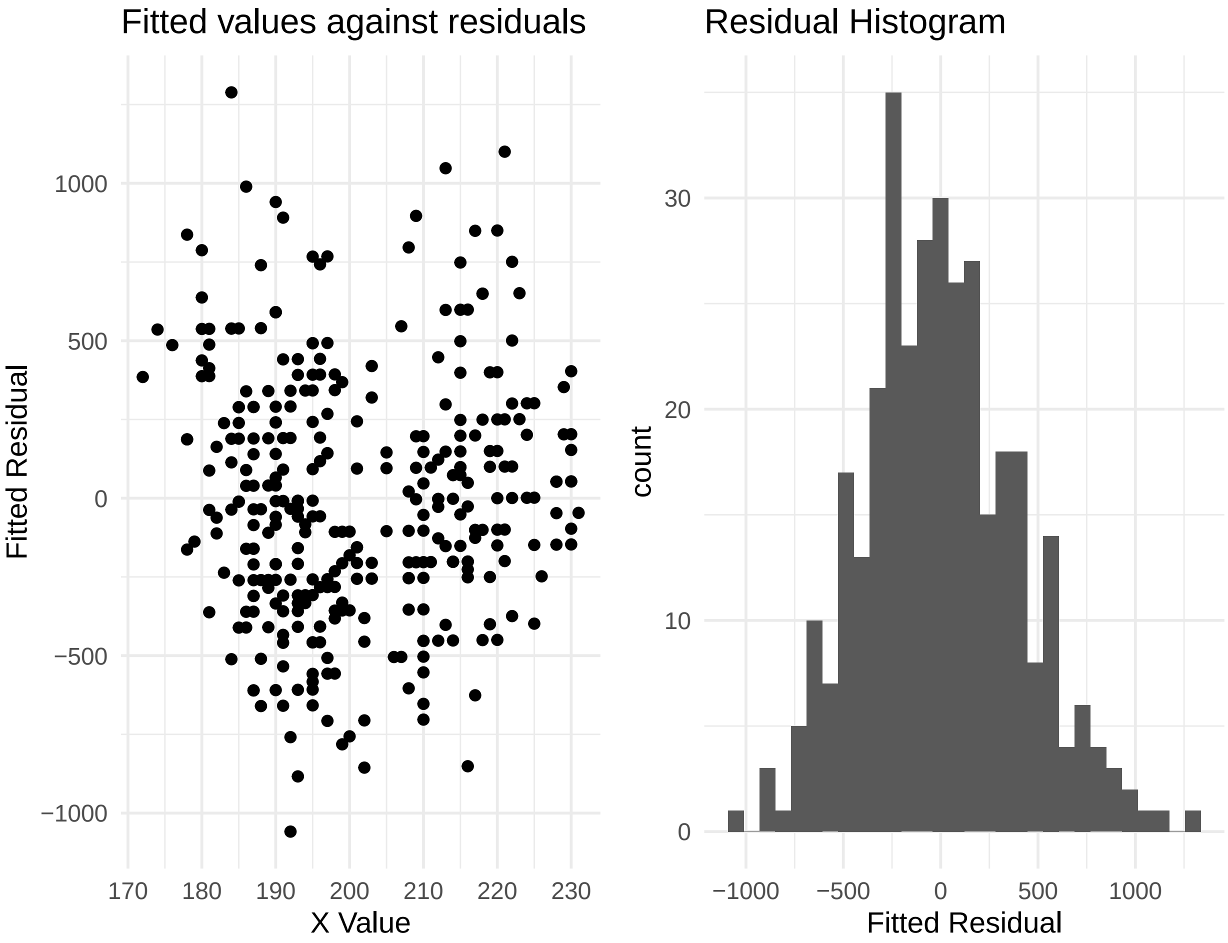
Residual analysis

What we have seen so far says that linear regression can be a really good model to describe linear relationships.

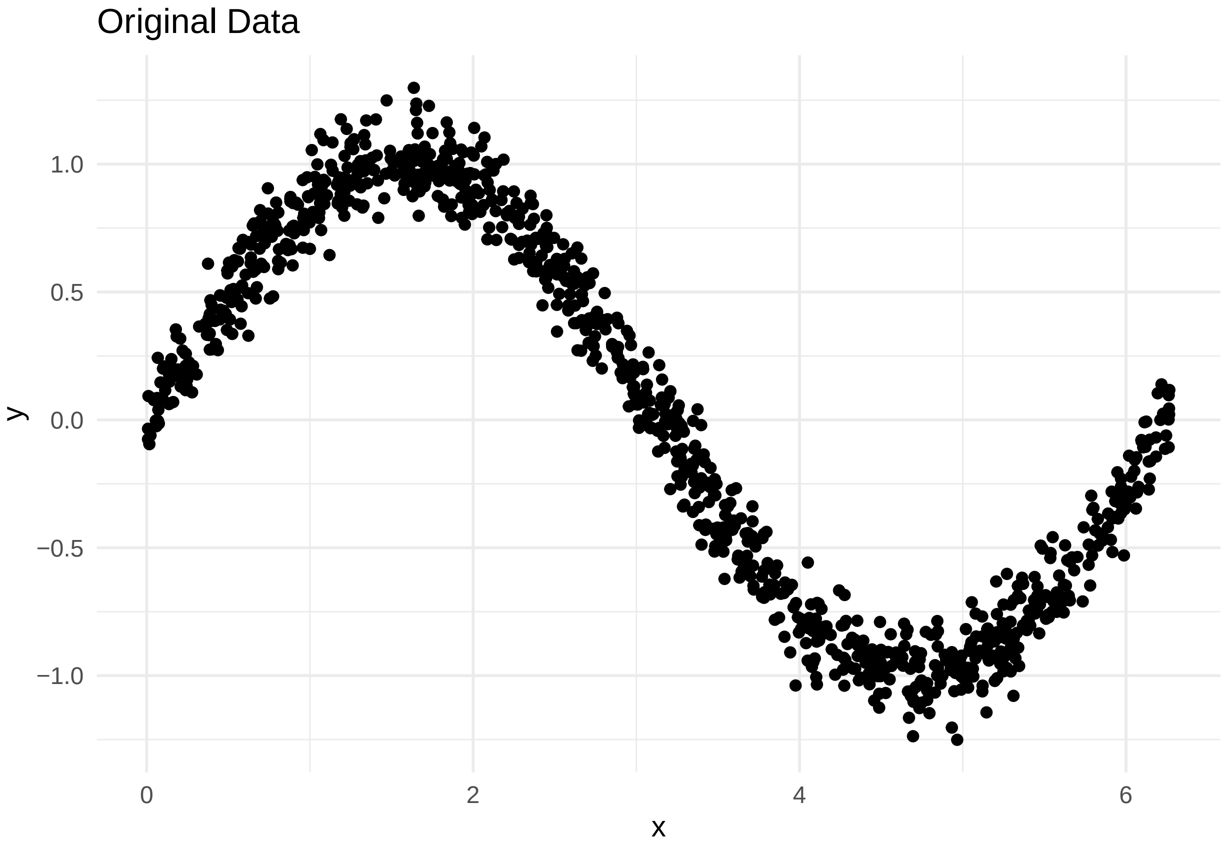
We have unbiased estimates for the coefficients, will show further properties.

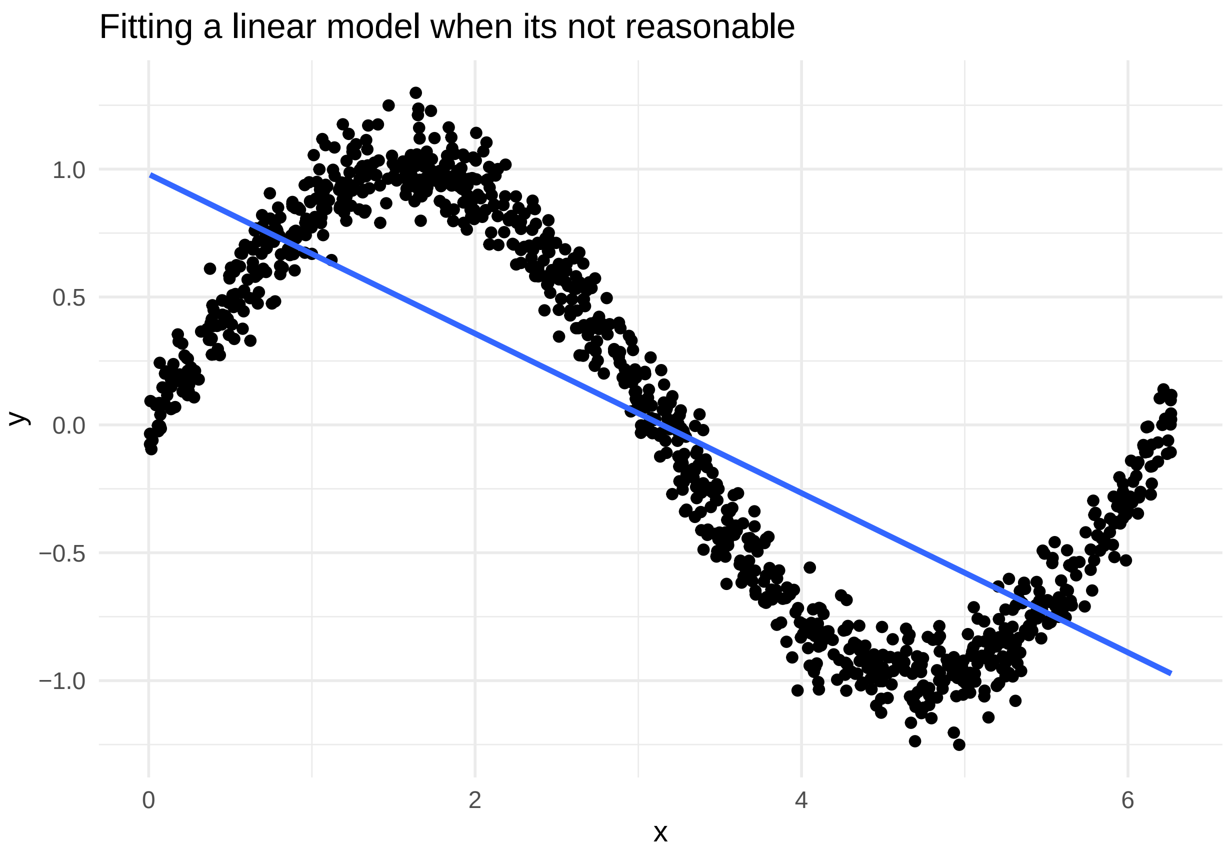
These properties rely on the residuals looking approximately normal. As such, important to confirm the residuals look reasonable.

Penguin Residuals

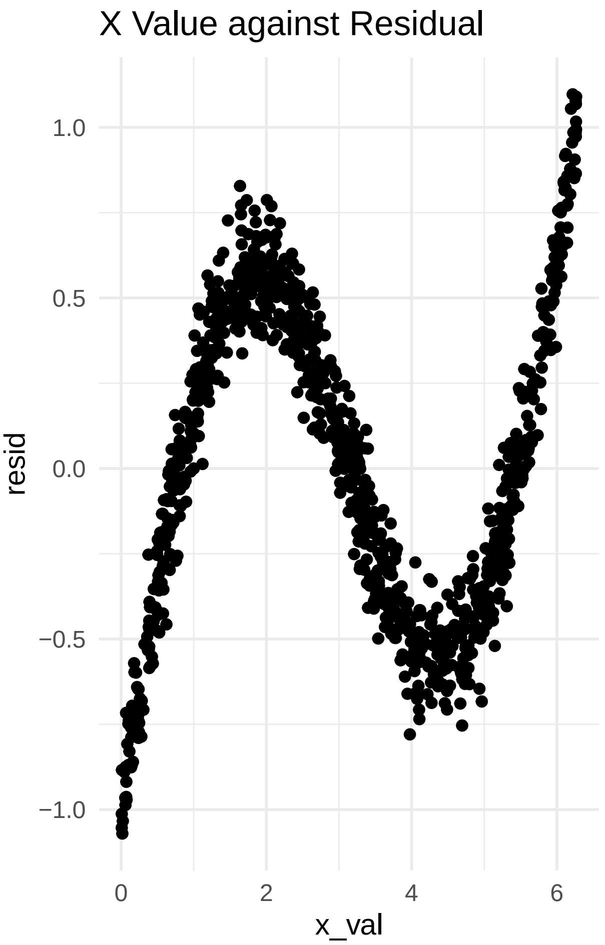


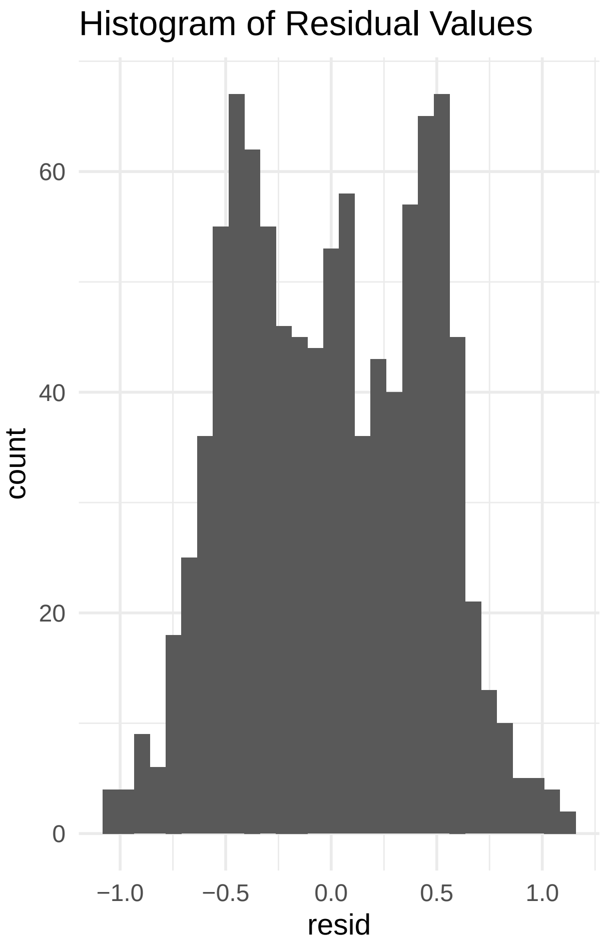
Bad Example



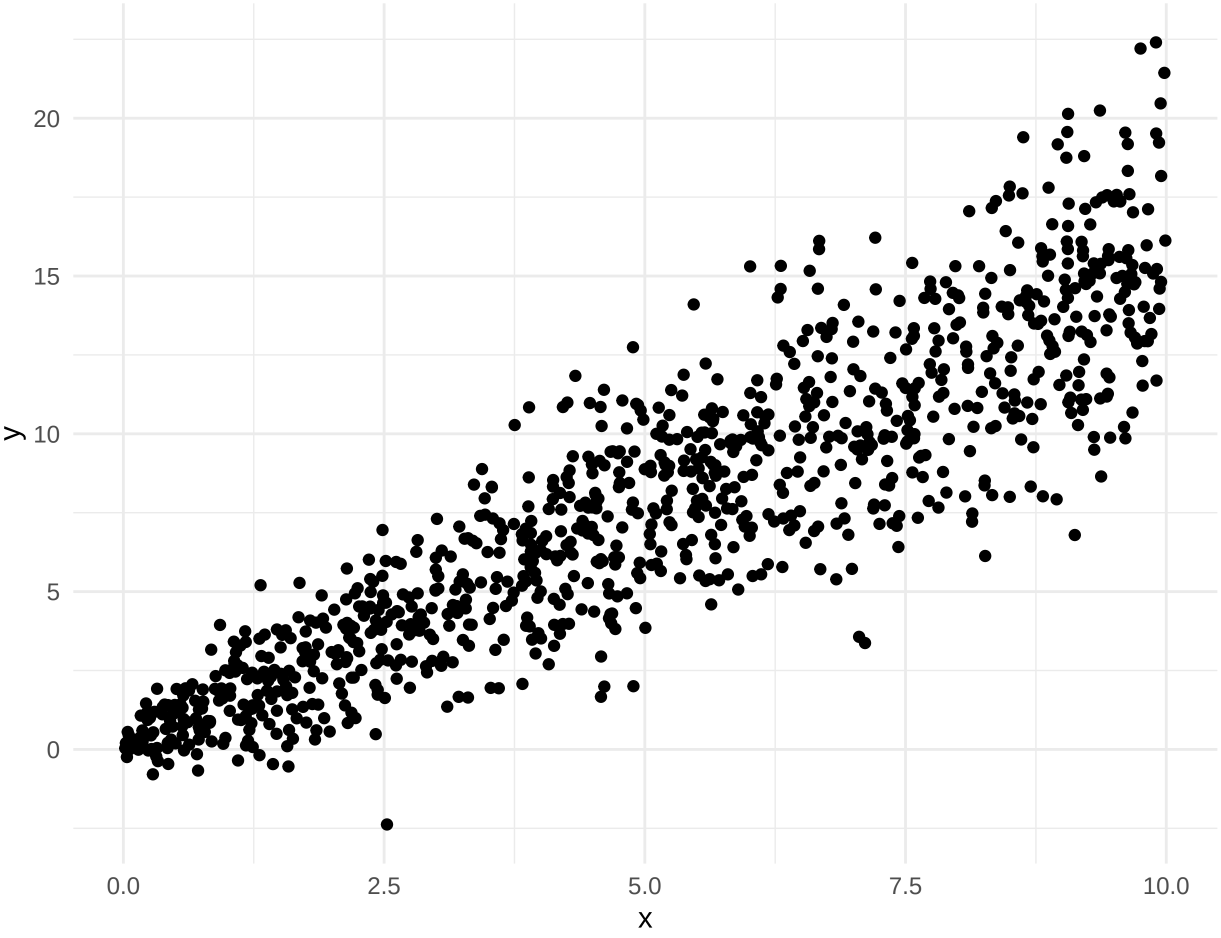

Fitting a linear model when its not reasonable
1.0
0.5
0.0
−0.5
−1.0
y
0
2
4
6
x

Residuals for an incorrect model

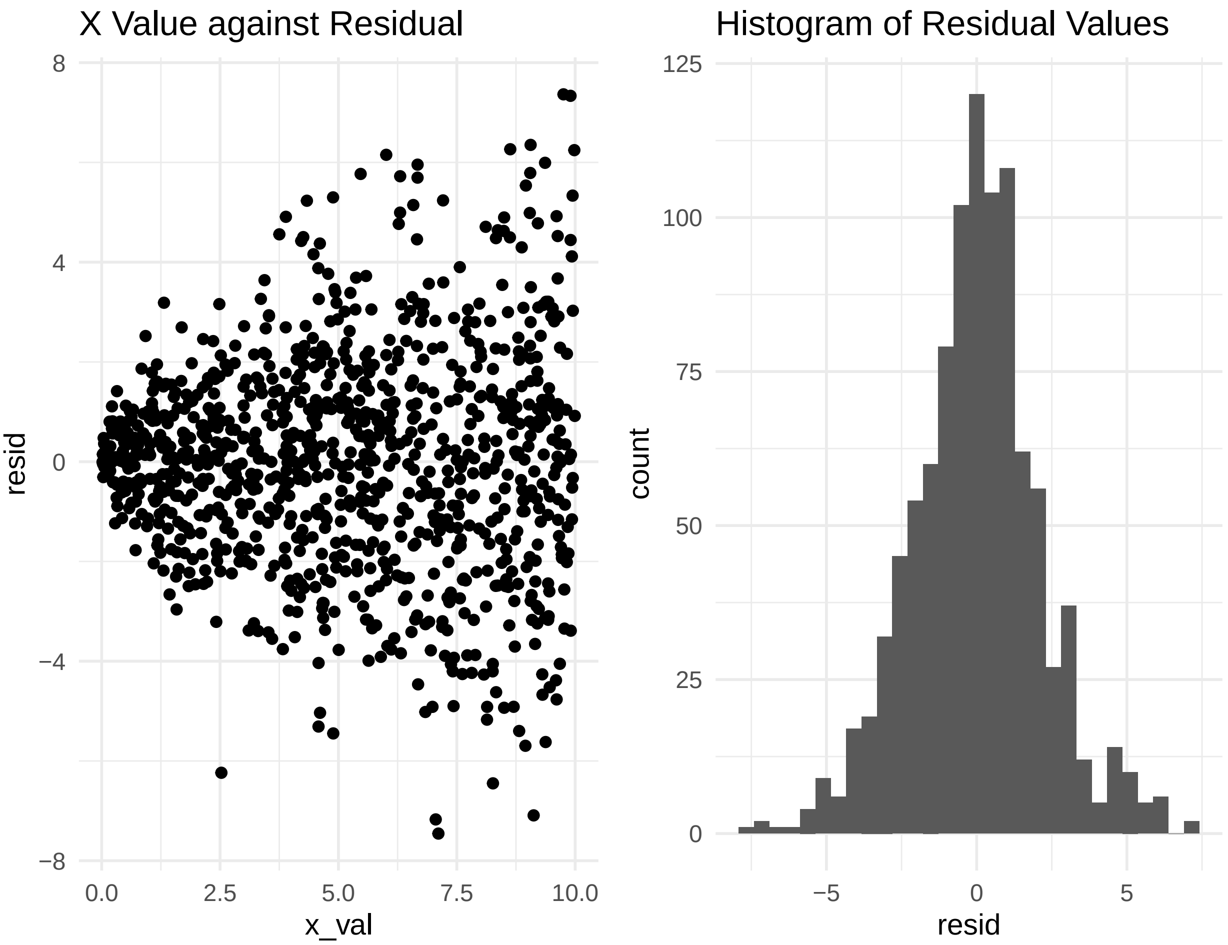




Variance of the Residuals Changing



Residuals of the Fit



Residuals

In the bad example it was pretty clear that a linear regression was incorrect from looking at the raw data. However it may not be as clear. Important to check the residuals carefully.

Remember that the interpretation of the coefficients we gave above is only valid when the model is reasonable.

The strength of a fit

Before we said that the formula for the slope, β_1, can be written in terms of the sample correlation $corr_{x,y}$. The square of this, written as R^2 is widely used to describe the strength of a linear fit.

This is a number between 0 and 1, larger values indicate a stronger relationship.

R Squared

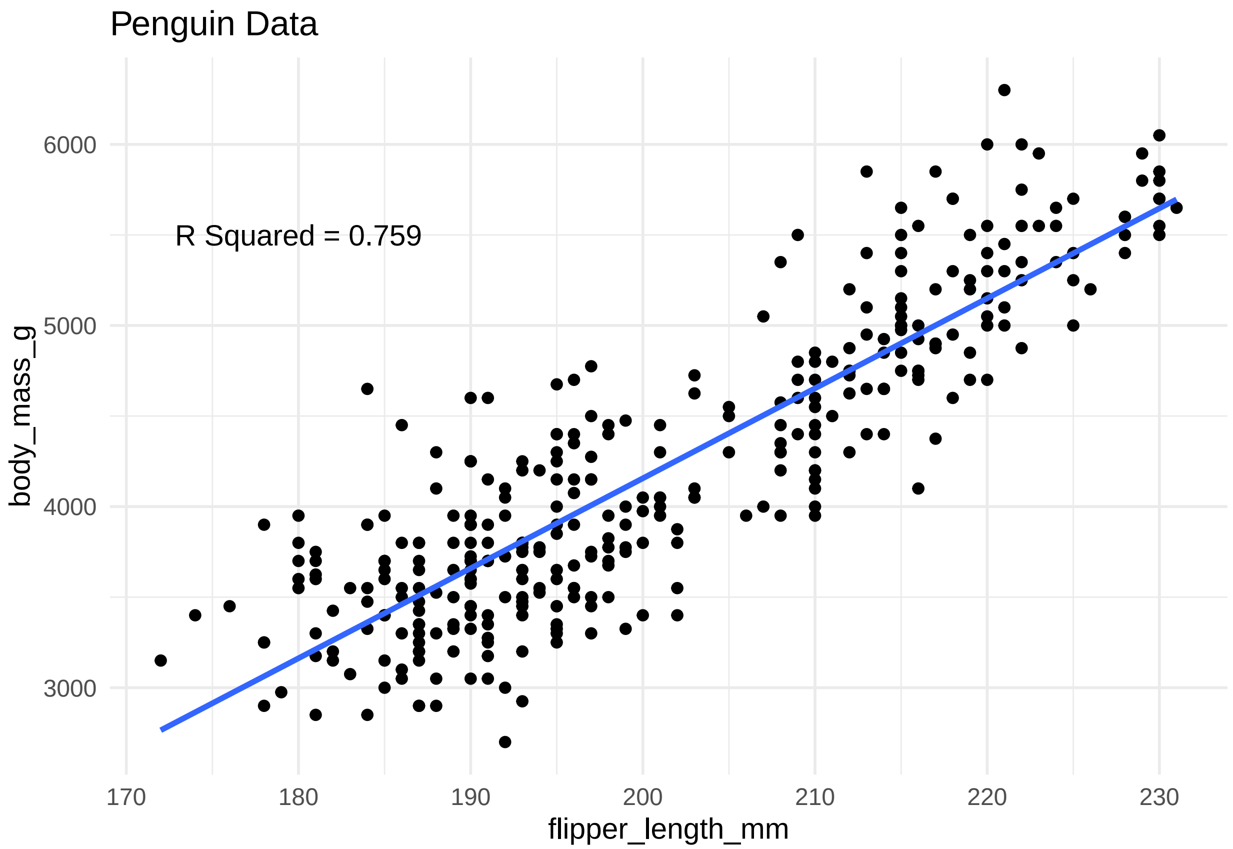
R^2 describes how much variation there is about the fitted line, relative to the total variance of the y variable.

If we fit the model and compute the residuals $e_1, \ldots, e_n$ then with s_e^2 the sample variance of the residuals, we can write

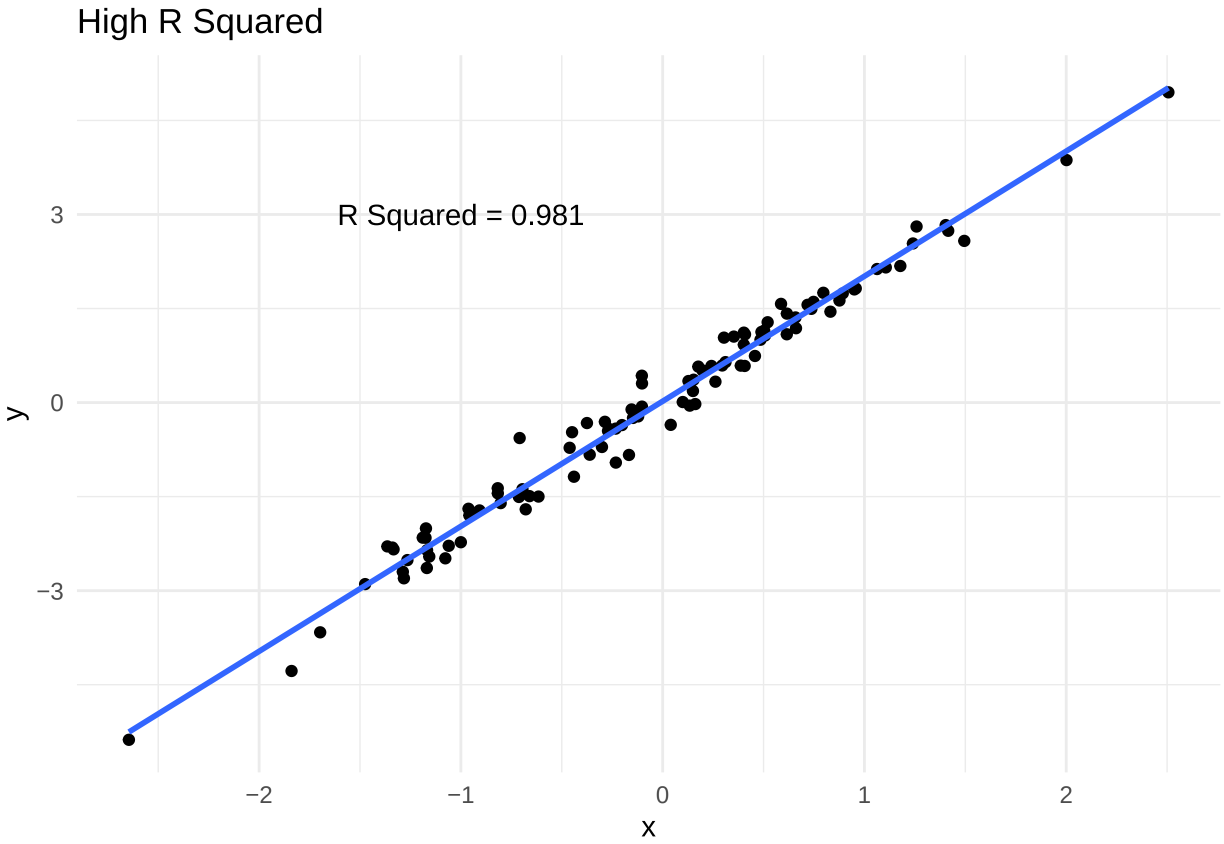
$$R^2 = \frac{s_y^2 - s_e^2}{s_y^2}.$$

As the variance of the residuals decreases, this number will be closer to 1.

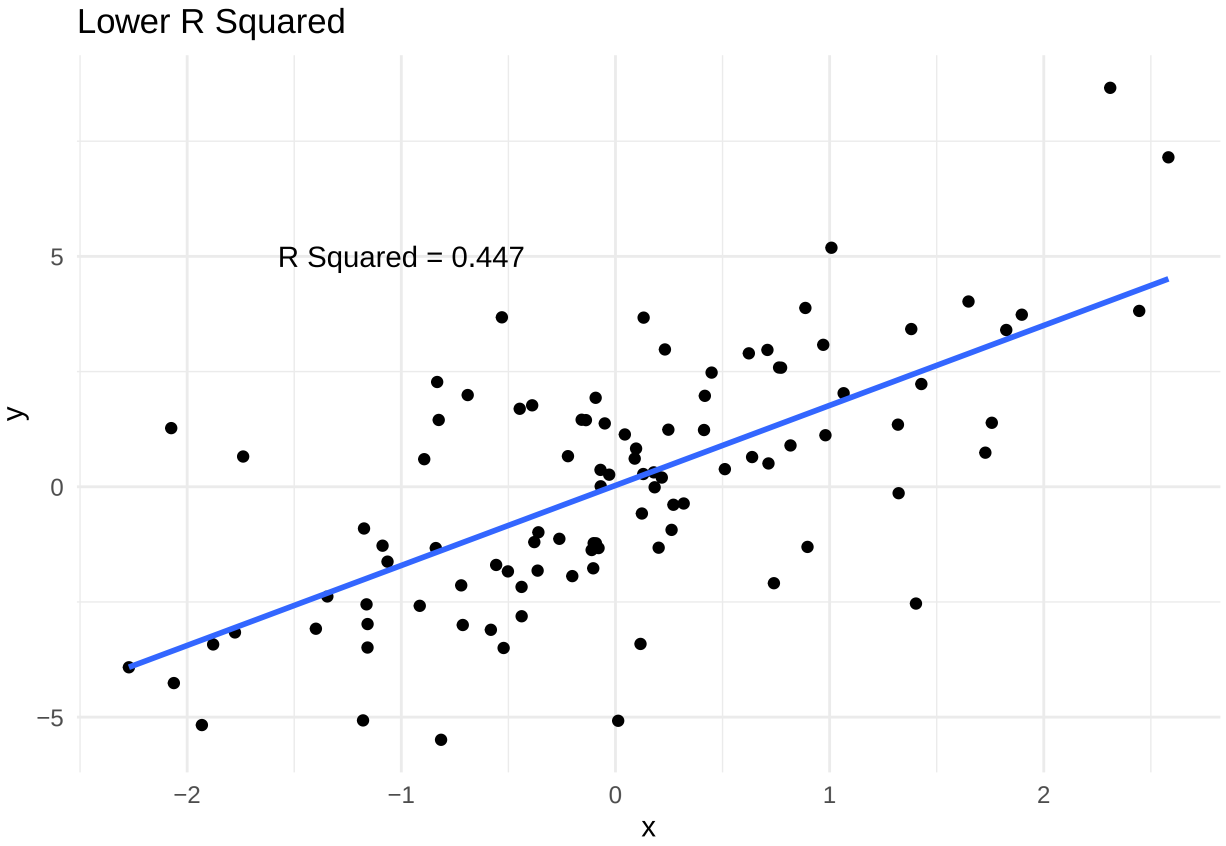
Example



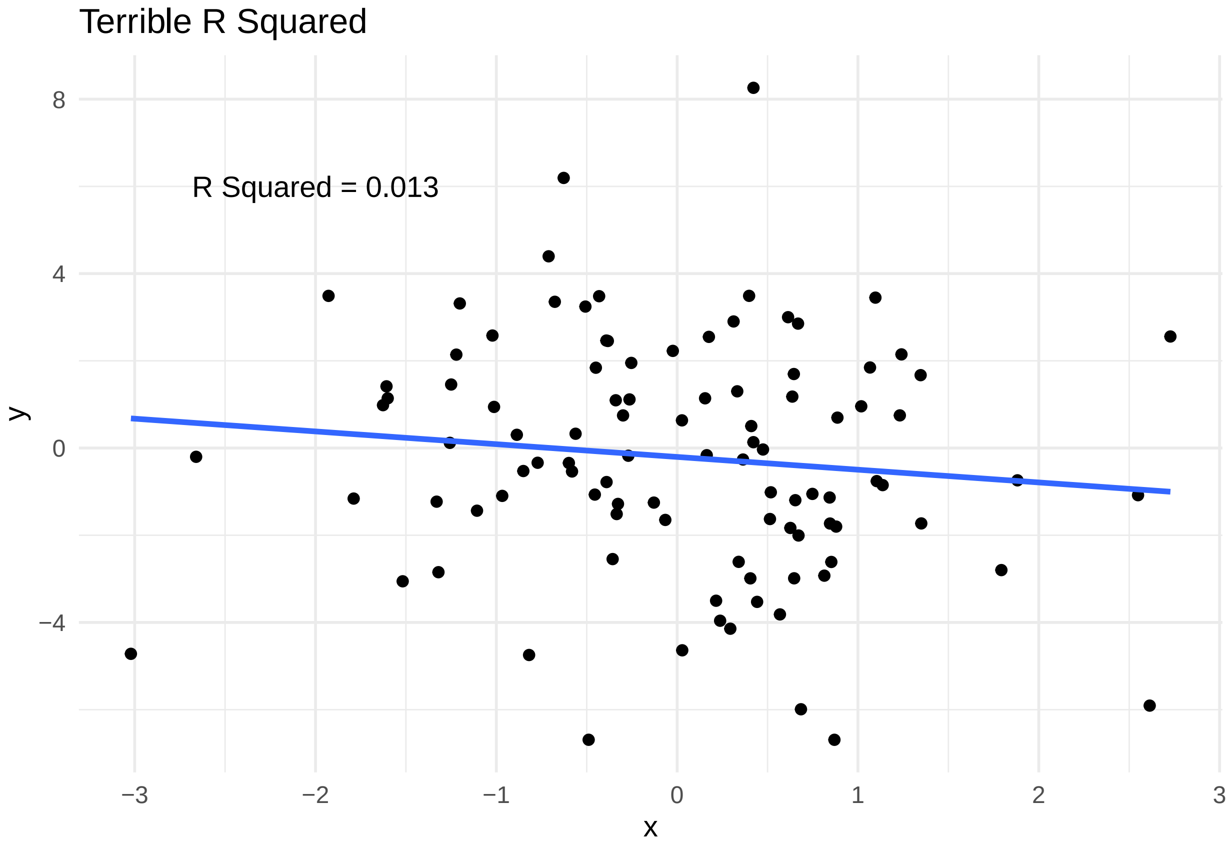
How R Squared varies



How R Squared Varies



How R Squared Varies



Example

When we fit the regression to the penguin data we had $corr_{x,y} = 0.871$ and $s_y^2 = (801.955)^2$, the sample variance of y.

We know that the $R^2 = corr_{x,y}^2 = 0.759$. We can use that to compute the variance of the residuals.

$$0.759 = \frac{(801.955)^2 - s_e^2}{(801.955)^2}.$$

We can solve this to get $s_e^2 = 154994$, so that the standard deviation of the residuals is 393.7. Compare this to the standard deviation of just the y values.

We get the same answer using `R`, where we will go through how to fit this model below.

```
sd(penguin_fit$residuals)
```

```
## [1] 393.6996
```

R Code for Regression

Fitting a linear regression is straightforward in most software packages.

As important is being able to understand the output from fitting a regression model

Thankfully it is reasonably straightforward.

Penguin Regression

```
penguin_fit <- lm(body_mass_g ~ flipper_length_mm,
                  data = penguins)
tidy(penguin_fit)
```

R Output

```
tidy_coefs <- tidy(penguin_fit)
print(tidy_coefs, width = 50)
```

```
## # A tibble: 2 x 5
##   term      estimate std.error statistic   p.value
##   <chr>        <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Interce~  -5781.     306.       -18.9 5.59e- 55
## 2 flipper_~     49.7      1.52      32.7 4.37e-107
```

```
summary(penguin_fit)$r.squared
```

```
## [1] 0.7589925
```

Confidence Intervals and Testing

The `R` output above gave estimates for the intercept term and the slope term, along with standard errors for those estimates.

This is essentially all we needed to construct confidence intervals previously, along with a z score for the chosen confidence level.

Here we would need to use a t table instead of a z table, but otherwise the procedure is identical. This would give us, for example, 95% confidence intervals for β_0 and β_1, with the same interpretation as before.

We also saw before that if we constructed a confidence interval and it didn't contain a specific value, like 0, that was equivalent to rejecting the null hypothesis of the true value being 0.

The p values given in the output above correspond to exactly that, testing whether each parameter is zero.

A slope term which could be zero would indicate there is no linear relationship present at all!

We can actually go a step further and build on this to get a confidence interval for the average value of y at a specific x value.

Even further, we can get a prediction interval for a new y at a specific x, but that is beyond the scope of this course!