Linear Regression

Nicholas Kortessis

Contents

Linear Regression

The functional idea with linear regression is that the character of an individual can be partly predicted with information about another character of that same individual. Finding this information is called "regression" and linear regression is a special case where one character (called a response) can be predicted by a linear function of some other character (called a predictor).

Typically, we write the predictor with symbol X and we write the response with the symbol Y. And when we pay attention to a single individual i, the predictor for individual i is X_i and the response for individual i is Y_i.

A typical example is that you want to might think that there is information about the length of different parts of a flower might be related to each other. R has data on the length of different parts of flowers for different species of plants in the genus *Iris*. This data were oridianlly collected by Edgar Anderson and published in 1935. If you are interested in the original paper, the citation is

- Anderson, Edgar. 1935. The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2–5.

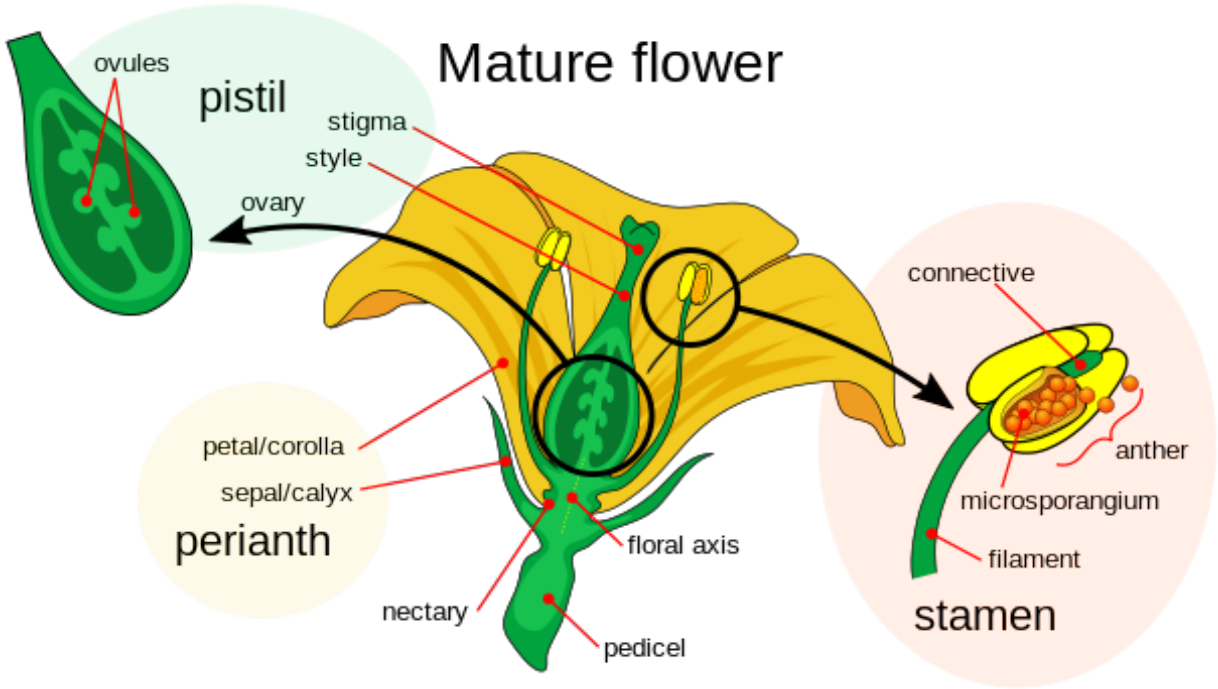
```
data("iris")
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

This data set has 50 plants of three species and measurements of sepal and petal width and height for each individual. (Remember, petals are the colorful leaf type structures on flowers and sepals are typically green leaf type structures at the very base of a flower.)



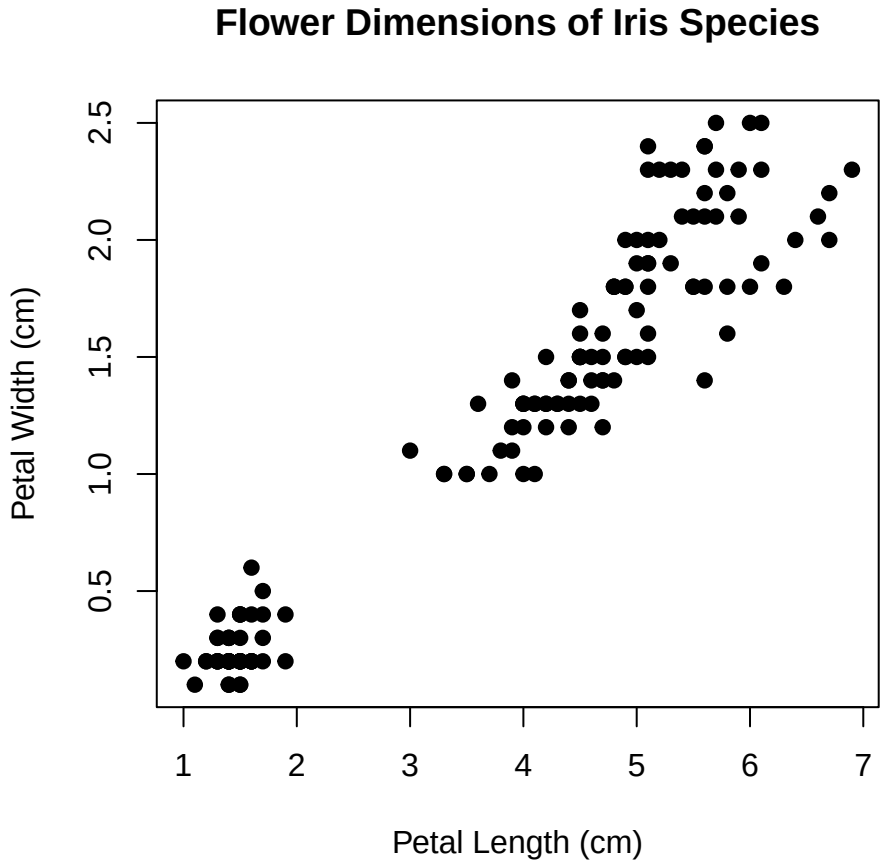
The characteristics of individual 4 in the data set are

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 4          4.6         3.1          1.5         0.2  setosa
```

This is a natural dataset to ask some questions about the **relationship** between multiple characteristics in individuals.

1. Do plants with longer petals have wider petals?
2. If so, how much does wider are petals for a given length?
3. How much can one understand about petal width from length?
4. If I know petal length, what should I expect petal width to be?

Each of these questions can be answered with regression. To make things simple, let's start with just a plot that helps us think about all these questions. It's just a plot of the petal length and width of each individual. Each point represents an individual.



It definitely looks like longer petals are also wider (and vice versa). It also looks like we can say that the increase in petal width with petal length is approximately linear, meaning we can draw a straight line through it.

We can write this mathematically. Let Y_i be the petal width of individual i and let X_i be the petal length of individual i. We are going to create a *probability model* to describe the relationship between Y and X as follows

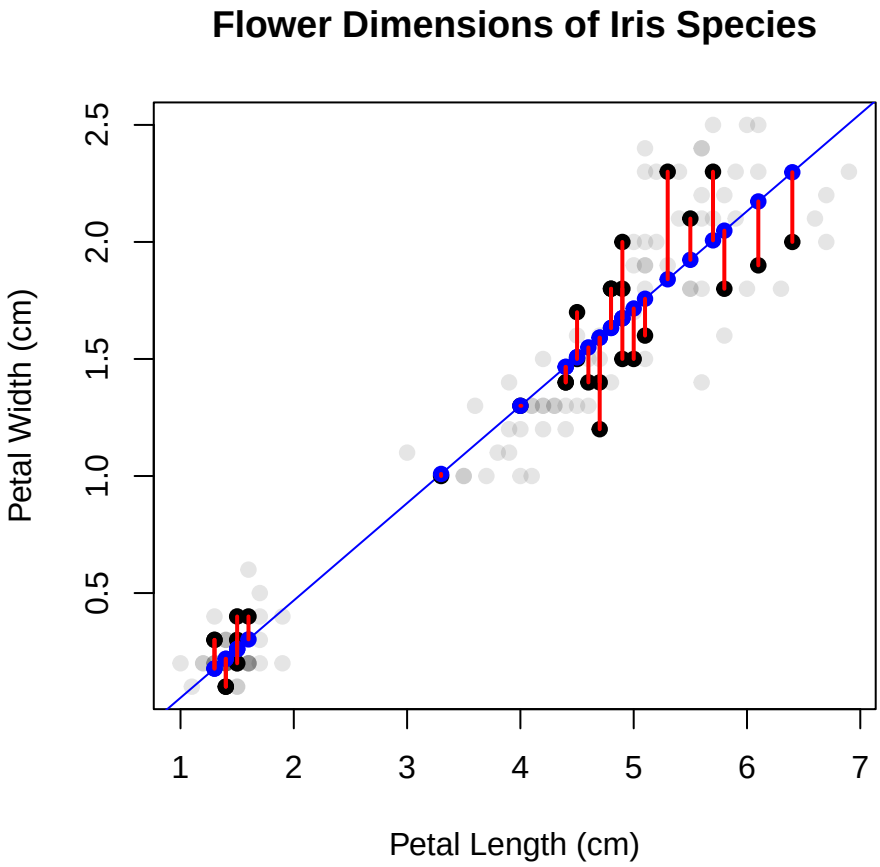
$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{a line}} + \underbrace{\epsilon_i}_{\text{residual}}$$

This model has two components: a line and a residual.

The **line** represents, well, any line that we could draw through the data points. The line has two parameters to describe it, an intercept β_0 and a slope, β_1. The intercept represents the y-value of the line when $X = 0$, and the slope represents how much the y-value of the line changes with one unit increase in the value of X.

The **residual** represents all that is leftover about the value of Y_i that is NOT described by the line. In this way, the residuals are very much like the errors in a linear model.

The figure below shows the line and the residual for a couple of points in the *Iris* dataset.

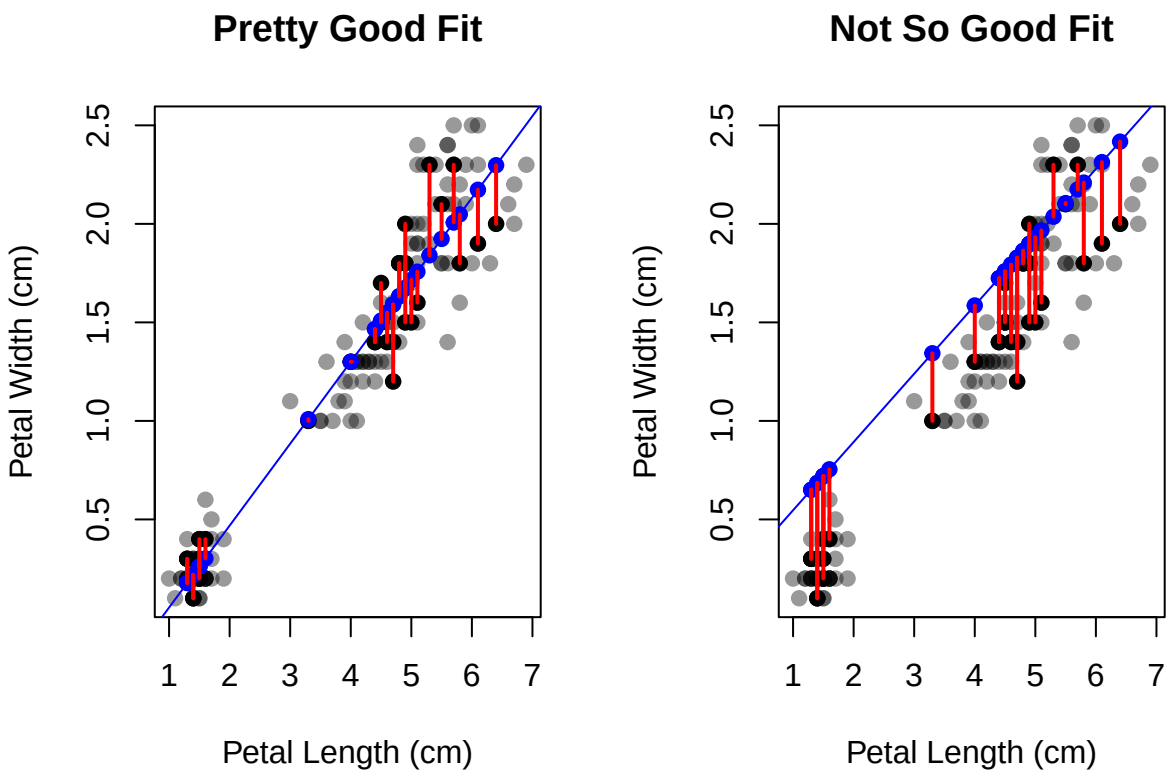


In this figure, the full line is given in blue, and the particular values of the line for specific individuals is given by blue dots. The residuals are given by the red lines that move from the blue line to the actual data points.

This line is actually what is considered to be the ***best fit line***. But how does one determine a line that fits the best?

Finding the best fit line

To find a best fit line, one needs to first describe a method for describing fit. To illustrate fit and lack of fit, look at the figure below with two different lines.



Visually, you can see that the panel on the right doesn't fit as well. You might also notice that this is because the blue points given by the line are consistently further away from the actual data points. The distance between the blue points given by the line and the actual data values is exactly the residuals, given by the red lines. Mathematically, we just ask for the difference between the data points and the line, which is

$$Y_i - (\beta_0 + \beta_1 X_i) = \beta_0 + \beta_1 X_i + \epsilon_i - (\beta_0 + \beta_1 X_i) = \epsilon_i,$$

which is just the residuals!

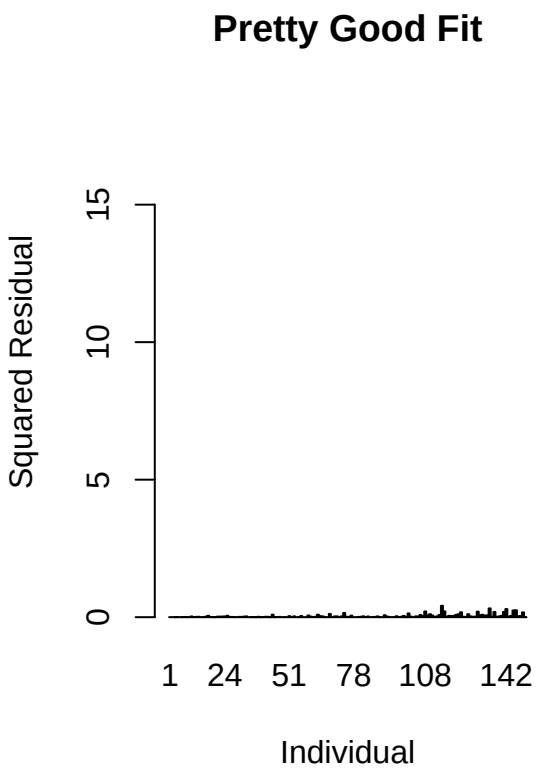
Poor fit is then given by larger distances of residuals. The more the residual distance of the points from any given line, the worse they fit.

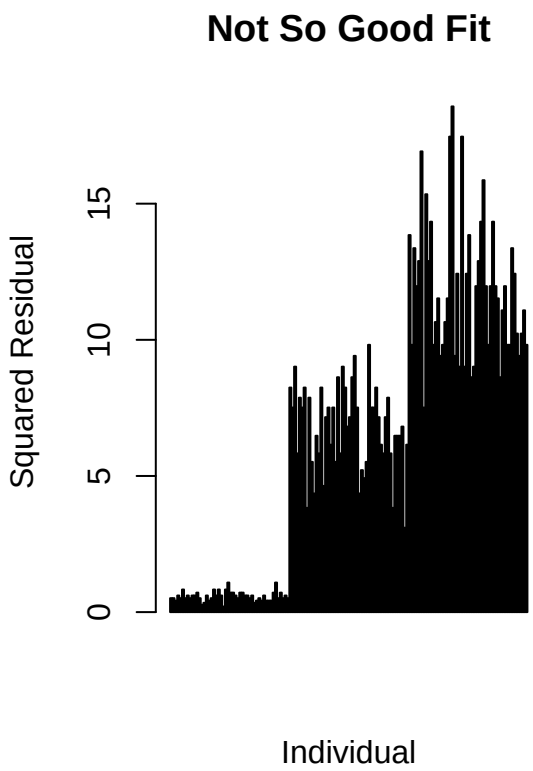
We have many times before measured how much individuals differ from some specified value whenever we are calculating variability. We measure variability using squares. Here, we can measure how much the line differs from the data points by looking at **squared residuals,** ϵ^2.

If we calculate the squared residuals for each data point in each of these panels, we get the following

```
par(mfcol = c(1,2))
barplot(mdl$residuals^2,
     xlab = 'Individual', ylab = 'Squared Residual',
     main = 'Pretty Good Fit',
     ylim = c(0, max(residuals^2)))

barplot(residuals^2, xlab = 'Individual', ylab = 'Squared Residual',
        main = 'Not So Good Fit')
```





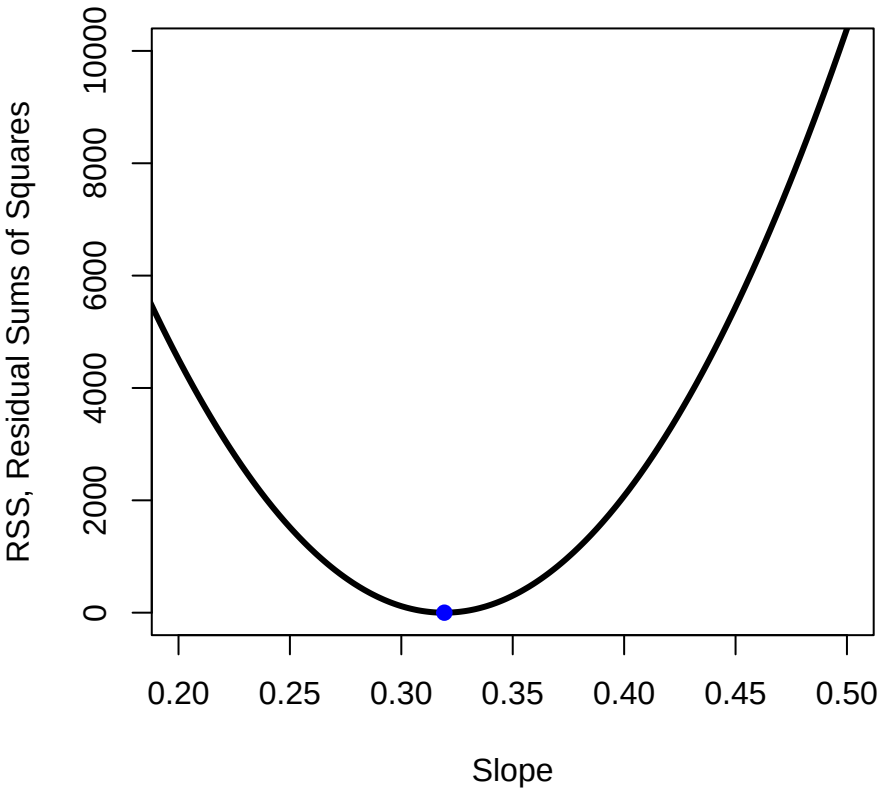
Because the first line fits so good, the squared residuals are barely perceptible on the graph. And for the poor fitting line, the squared residuals look much larger. One way to evaluate lack of fit is by looking at the total of these squared residuals

$$\text{Residual Sums of Squares} = RSS = \sum_{i=1}^{n} \epsilon^2.$$

Lines fit worse when this residual sums of squares is higher and lines fit better when the residual sums of squares is smaller. How does one find the best fit line then? You simply find the line with the smallest residual sums of squares.

What this means in practice is to find the values of β_0 and β_1 that leads to the smallest value of RSS. This can be done using the tools of calculus, but here is a way to think about how this might work.

Let's pick $\beta_0 = 0$ because this says that flower petals have zero width when their flower petals have zero length. Seems pretty reasonable. Now we can scan across a bunch of values of β_1, the slope and calculate RSS each time. When we do that, we get the following plot.



The blue points is the minimum value of the residual sums of squares. The slope at this point is $\hat{\beta}_1 = 0.32$. We give this parameter a "hat" to signify that it is an **estimate** of the slope. This estimate is found by **minimizing residual sums of squares.** This particular estimate is therefore known as the **least squares estimate** (LSE) or sometimes known as **ordinary least squares** (OLS) estimation.

Luckily, we don't need to do this process every time. Using the tools of calculus, we have the following equation for the LSE of the slope and the intercept.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

What this says is that you can estimate the slope using the data points and then use that to estimate the intercept. (Remember from calculus that if you want to find the minimum of a function $f(x)$, you find the x-values where the slope of the function is zero, $df(x)/dx = 0$, and the curvature of the function is positive, $d^2f(x)/dx^2 > 0$. You can see from the figure that the slope of the line at the blue point is zero and the curvature is positive).

If we use this data on flower petals and widths in the equations for the least squares estimates, we find that the least squares estimates for the slope and intercept are

$$\hat{\beta}_1 = 0.416$$

and

$$\hat{\beta}_0 = -0.363.$$

These are subtly different than the method used above because you actually need to find the **combination of** slopes and intercepts that **together minimize RSS**. We didn't exactly do that because we just picked a value of the intercept of zero and then found the slope that minimized RSS.

Interpreting the slope estimator, $\hat{\beta}_1$

The equations for the slopes estimator looks intimidating, but it has a very nice interpretation, if you just give yourself a chance to figure out what is going on.

To see what is going, first consider the numerator. The denominator is always positive, so it doesn't matter as much at this point.

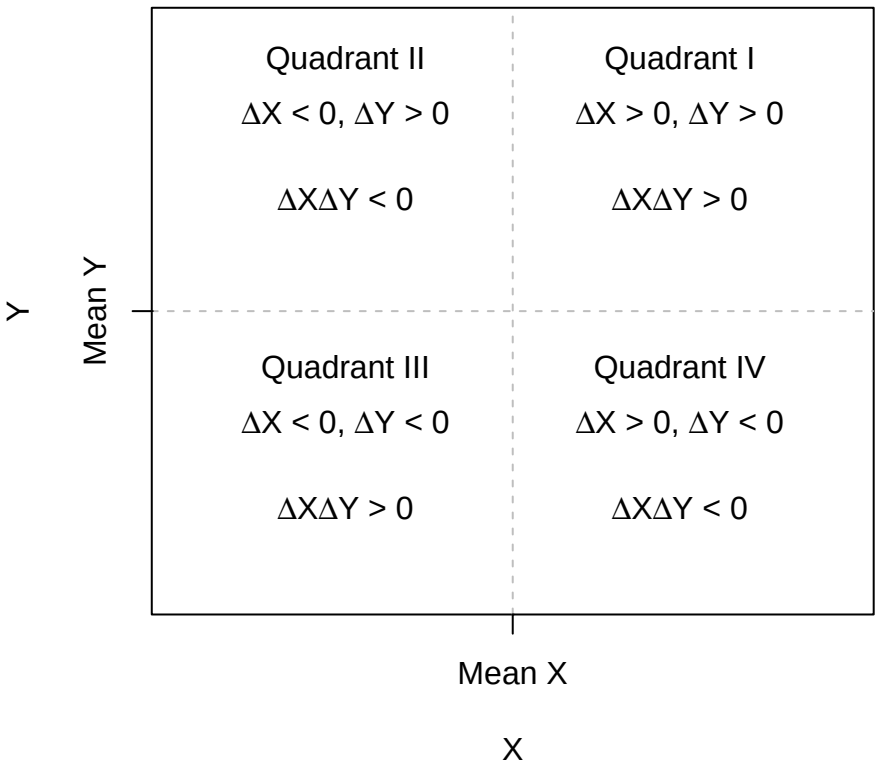
Here is the numerator of $\hat{\beta}_1$:

$$\sum_{i=1}^{n} \overbrace{(X_i - \bar{X})}^{\Delta X} \overbrace{(Y_i - \bar{Y})}^{\Delta Y}.$$

I've labelled these as ΔX and ΔY. These values are how much a data point's x and y values differ from the average. When the point is above average, it's delta value will be positive. When the point is below average, the delta value will be negative. Importantly, this is a product of two delta values, and so we have the following possibilities for each point:

	Below Average X ($\Delta X < 0$)	Above Average X ($\Delta X > 0$)
Above Average Y ($\Delta Y > 0$)	$\Delta X \Delta Y < 0$ (quadrant II)	$\Delta X \Delta Y > 0$ (quadrant I)
Below Average Y ($\Delta Y < 0$)	$\Delta X \Delta Y > 0$ (quadrant III)	$\Delta X \Delta Y < 0$ (quadrant IV)

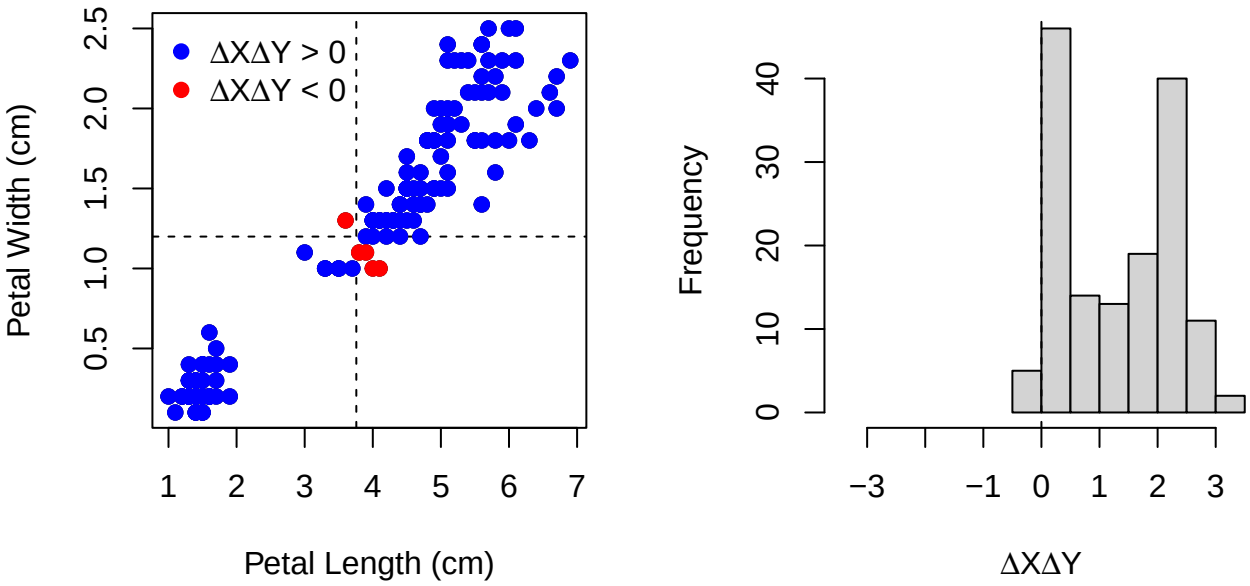
The graph below shows the quadrants and what the sign of $(X_i - \bar{X})(Y_i - \bar{Y}) = \Delta X_i \Delta Y_i$ is.



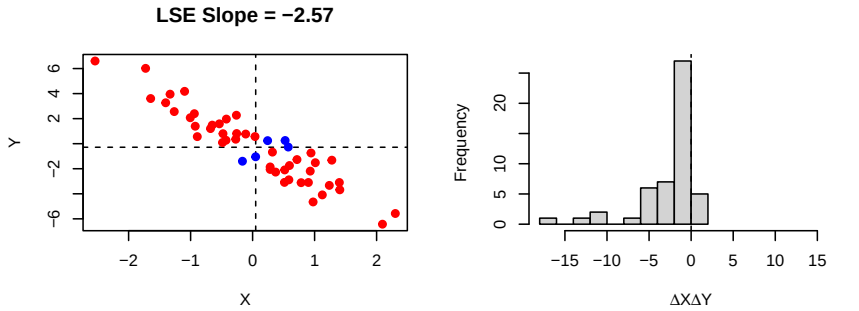
For data that show a positive relationship, many values will lie in quadrants I and III, which both have positive values of the product of ΔX and ΔY. The more data points in these quadrants, the more evidence that accumulates for a positive relationship.

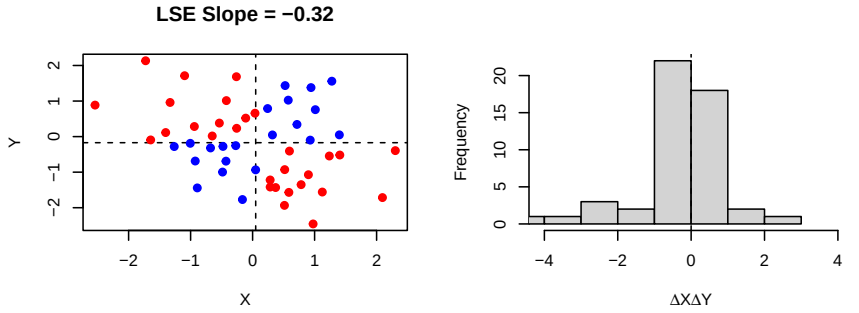
The same argument applies when there are many data points in quadrants II and IV. In that case, most of the data points with have ΔX and ΔY values that are opposite in sign and so will show up as evidence of a negative relationship.

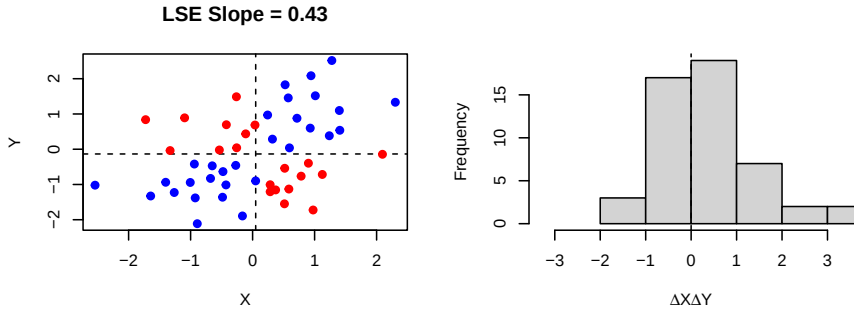
For the *Iris* dataset, it is clear that the relationship is positive. This is because most of the datapoints lie in quadrants I and III. The figure below colors the data points depending on the sign of $\Delta X \Delta Y$. Those with positive sign are in blue and those with negative sign are colored red. You can see that the vast majority are blue, indicating a lot of evidence of a positive relationship. The histogram on the right also shows the distribution of the $\Delta X \Delta Y$ values for all the data points in the data set. The LSE of the slope uses the average of these values.

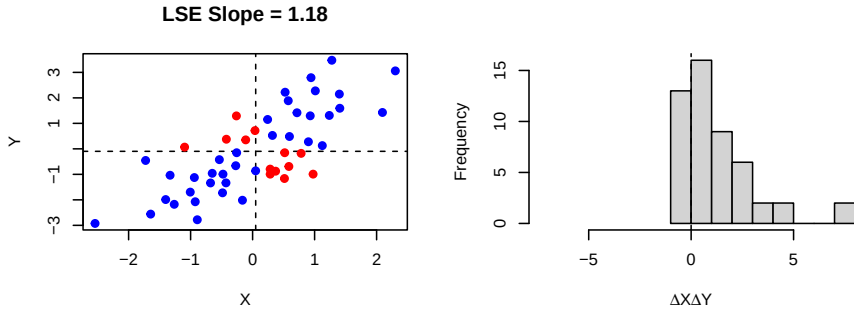


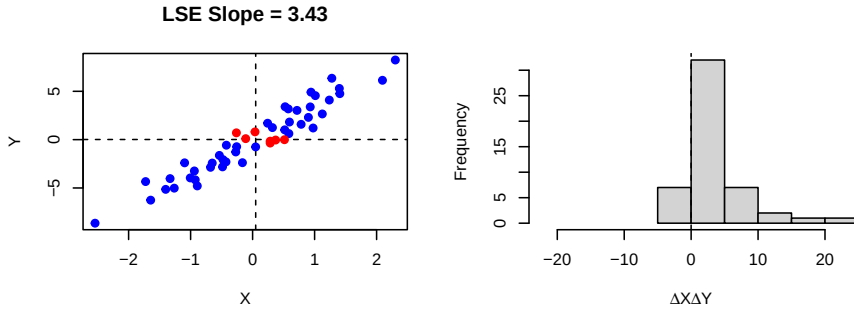
Here is the same kind of figures for other kinds of data with different relationships.


LSE Slope = −2.57
Y
X
Frequency
ΔXΔY


LSE Slope = −0.32
Y
X
Frequency
ΔXΔY


LSE Slope = 0.43
Y
X
Frequency
ΔXΔY

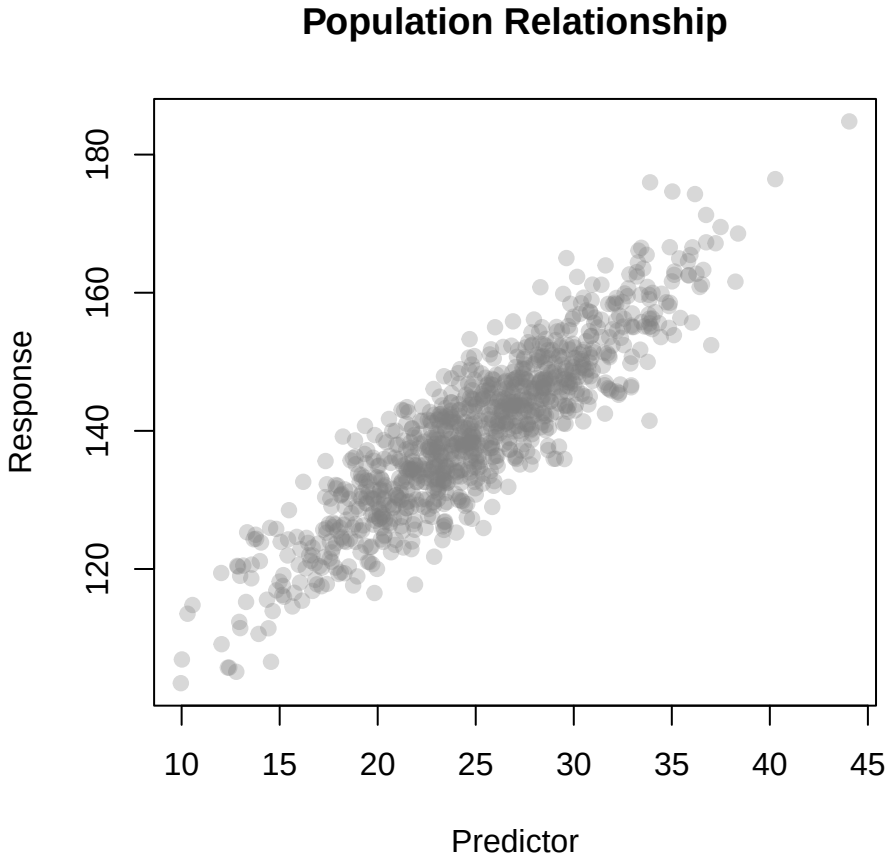

LSE Slope = 1.18
Y
X
Frequency
ΔXΔY


LSE Slope = 3.43
Y
X
Frequency
ΔXΔY

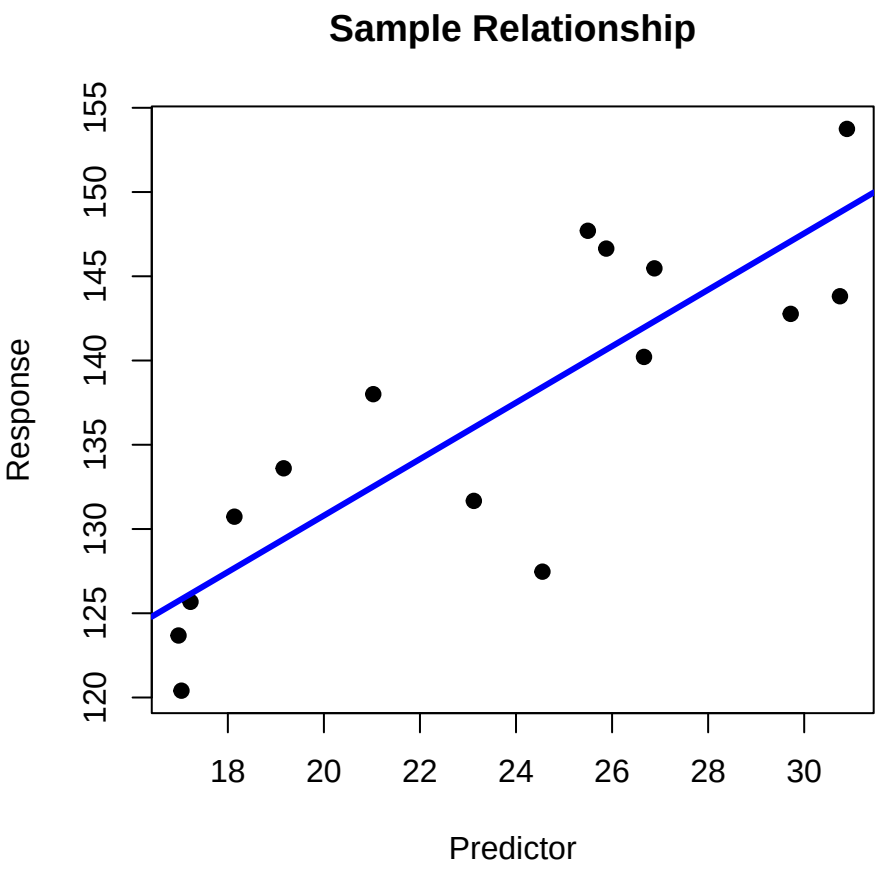
Including Uncertainty in Estimates

Just like with other forms of estimation, the parameters of a linear model are come with some uncertainty. It turns out that the sampling distributions for the slope and the intercept both follow t-distributions for simple linear regression. But an important point is that the mean and intercept estimates are not independent of each other. Larger estimates of the slope require different estimates of the intercept. This means that uncertainty in the slope and intercept estimates cannot be considered independently. It's better to think about their integration in making lines.

To capture this idea, imagine that we have a population with an actual, true relationship between a predictor and response such as that in the figure below. Here, the true relationship is given by a line with an intercept of 90 and a slope of 2.



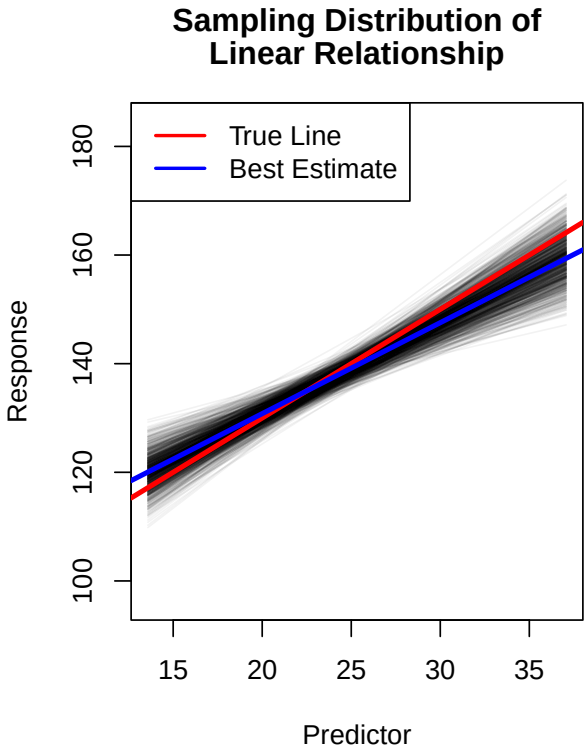
Now imagine that that we sample this population and make a best fit line, as in the figure below.

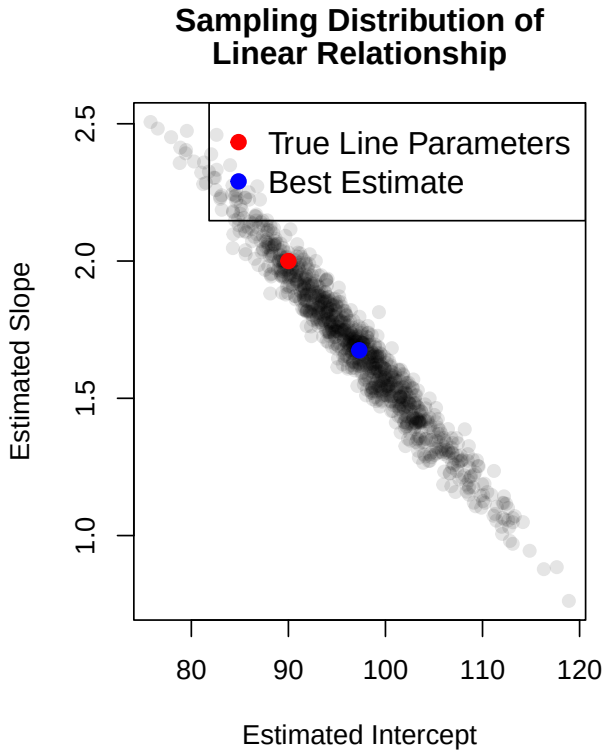


Now imagine that we resample this population many times and every time we resample the population, we fit a new line to the new sample. Because the samples will differ a little bit each time, the lines will be slightly different from one another for each sample. The figure below shows the lines from 1000 different samples of the population. Each line is a different sample. The blue line is the best fit line from the first sample and the red line is the actual fit.

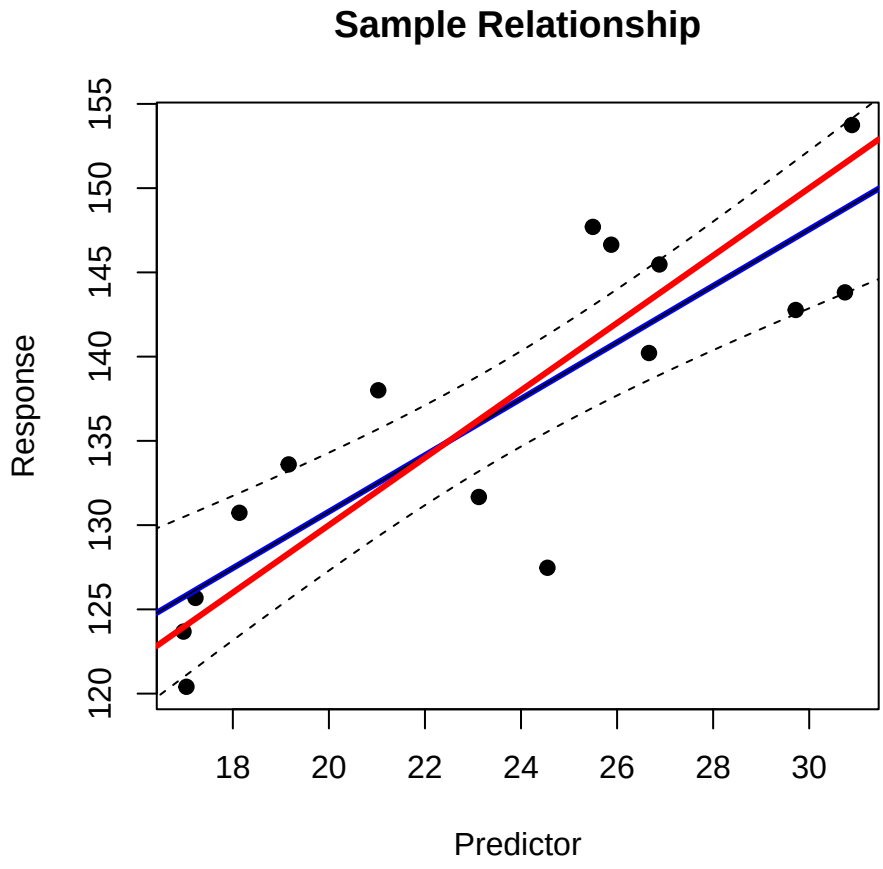
What this shows is a confidence interval of the relationship between the predictor and response. Confidence intervals don't apply just to the means, but to any properties that you might be interested in about a population.

We can show this sampling distribution in line space (the left panel below) or in the 2-dimensional space of the two parameters that describe the line: the intercept and slope (the right panel below). You can see that higher intercepts are associated with smaller slopes and vice versa.

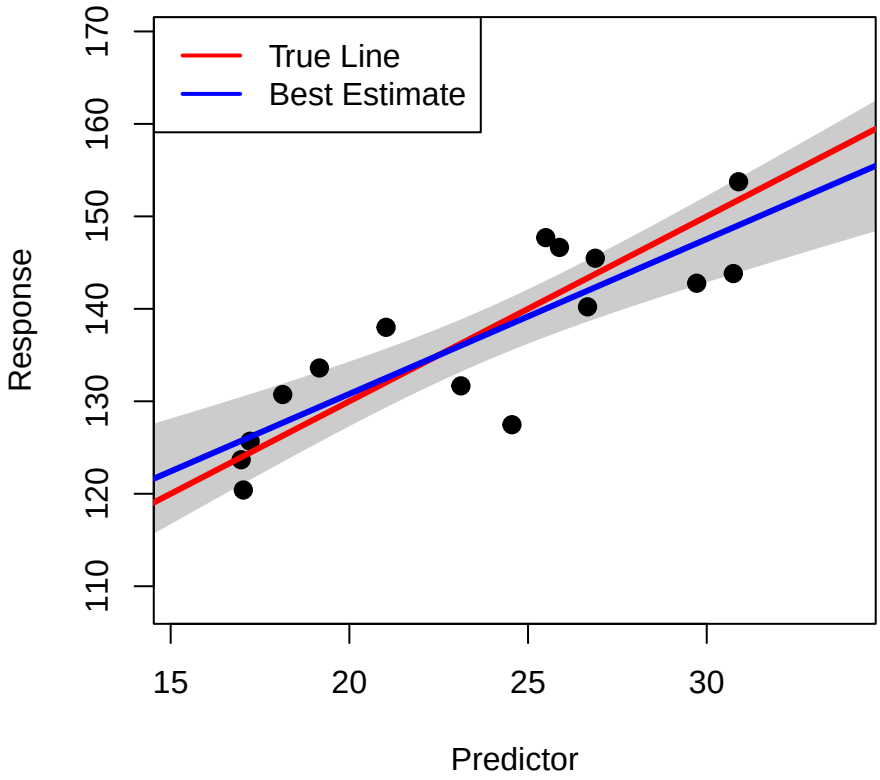




We can show 95% of the lines that come from resampling with a 95% confidence interval of the line. Here is what that looks like on the first sample of data.



Another way to show this confidence interval is with a shaded region that reflects where 95% of the lines fall.



Statistical programs such as R will provide the best least squares estimate as well as the standard errors of the coefficients, among other things. Here is what R's model output looks like.

```
##
## Call:
## lm(formula = Response ~ Predictor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9520  -4.3340  -0.4677   4.4577   7.6967
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  97.3025     7.1105  13.684 4.26e-09 ***
## Predictor     1.6748     0.2955   5.668 7.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.572 on 13 degrees of freedom
## Multiple R-squared:  0.7119, Adjusted R-squared:  0.6898
## F-statistic: 32.13 on 1 and 13 DF,  p-value: 7.691e-05
```

You can see the estimates of the parameters as well as the standard errors of the estimates. Remember that the standard errors are the standard deviation of the sampling distribution of each parameter. Moreover, R also provides a t-value under the hypothesis that each estimate is zero. Last, it provides a p-value associated with each t-value.

Checking Model Assumptions

For linear regression to be valid, a number of assumptions need to be met. To see them, let's look again at the description of the model. Simple linear regression models have the following form:

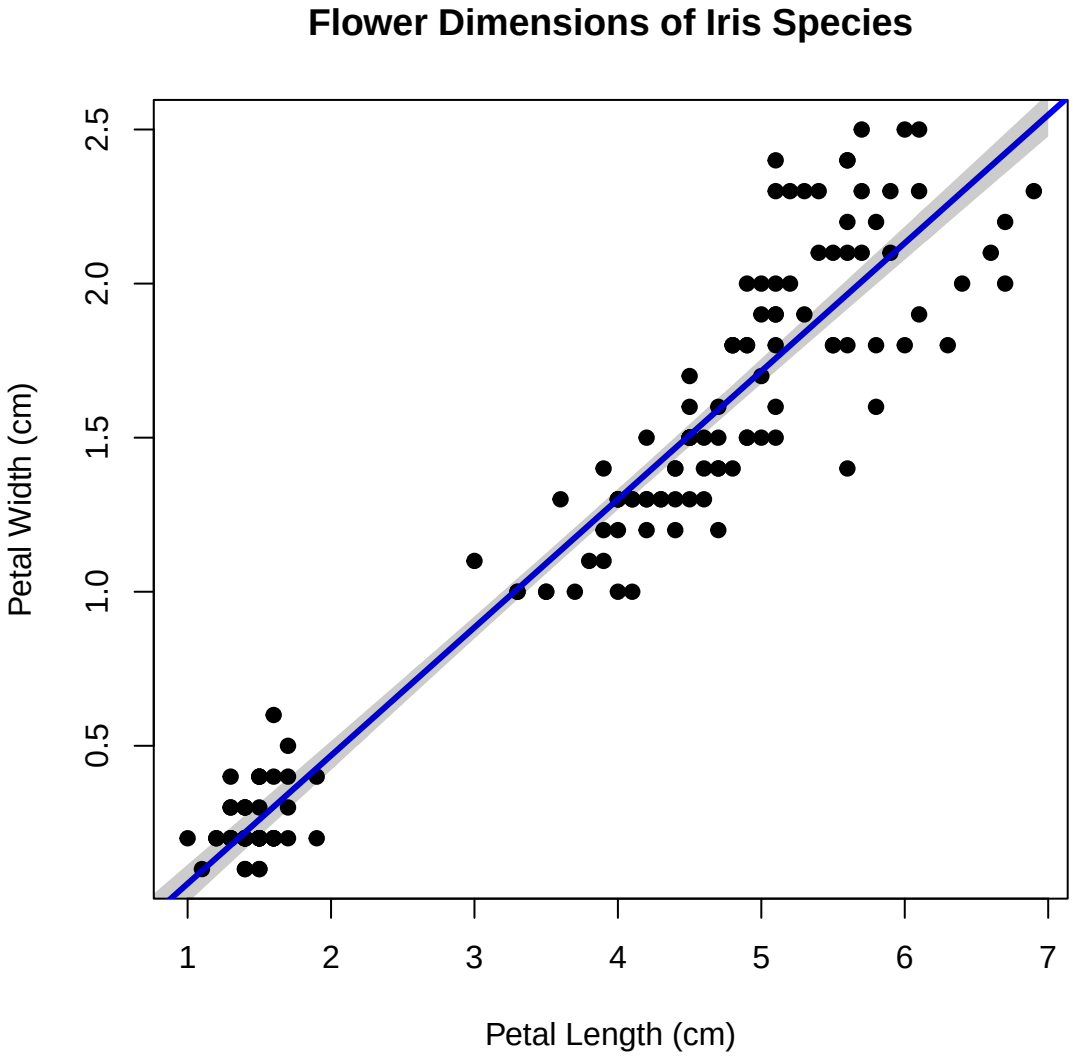
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

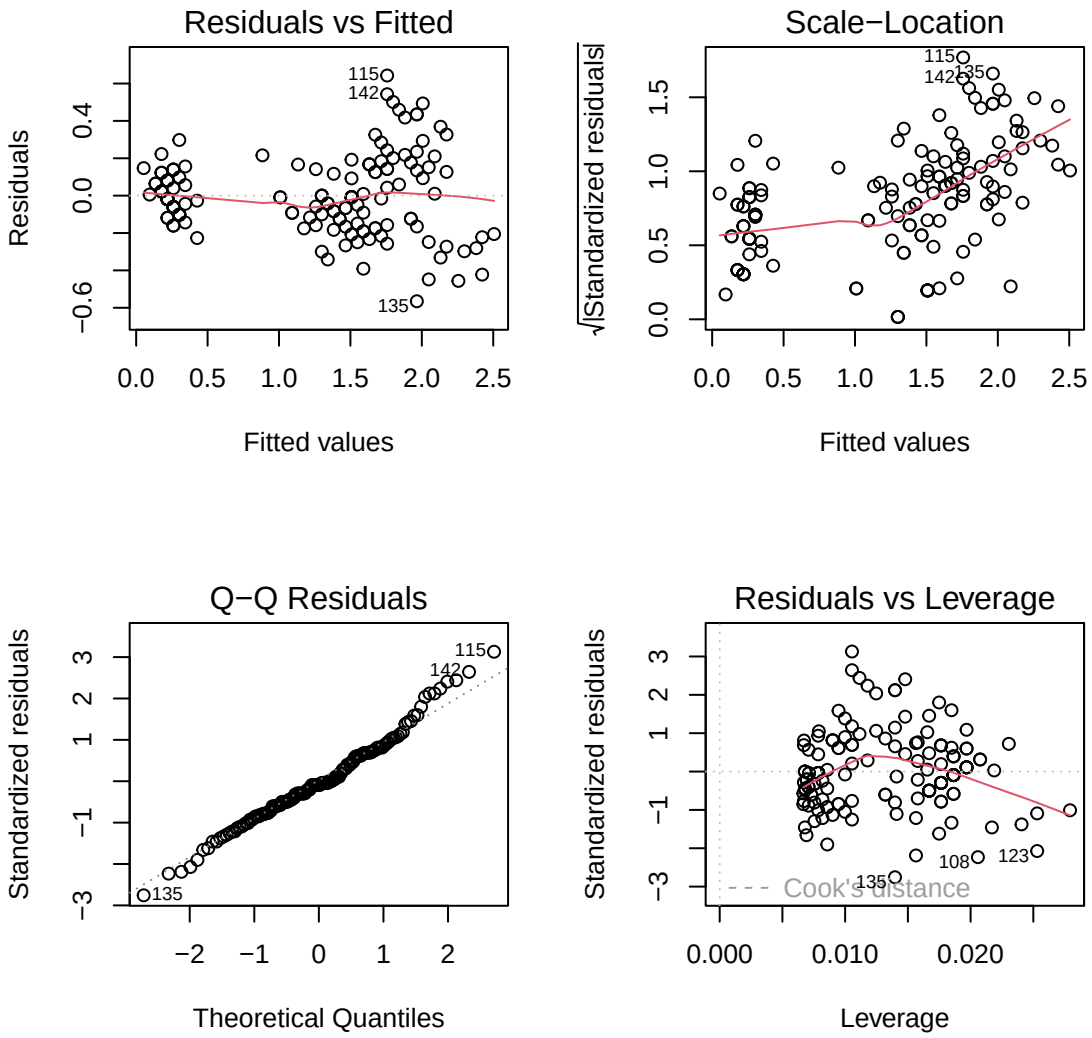
where

$$\epsilon_i \sim N(0, \sigma^2).$$

are the residuals. This means that the residuals are normally distributed and have a mean of zero and a variance that is constant for all values of X. Thus, the most important feature of model fit comes from inspecting the residuals.

Remember that the residuals are the difference between the observed values of Y and the predicted values of Y from the model. We can inspect the residuals directly and look for zero mean, constant variance, and normality using residual diagnostic plots. Here is the Iris data and the fit and the regression diagnostics for the model.





The topleft regression diagnostic plot shows how the residuals are distributed for the predictors. The main feature here is to look for patterns in the average of the residuals, shown by the red line. If you see consistent patterning, this is a sign that the average of the residuals is NOT zero for some values of the predictors, a clear violation of the zero average residual assumption.

The topright regression diagnostic plot shows the magnitude of residuals plotted against the fitted values. Here, the key thing to look for is if the size of the residuals changes with the fitted values. If it does, this violates the assumption of constant variance of residuals.

The bottomleft plot shows a Q-Q plot of the residuals. This plot helps us evaluate whether the residuals are normally distributed. If the residuals are normally distributed, the points should fall along the diagonal line. If they do not, this is a sign that the residuals are not normally distributed. As with our discussion of Q-Q plots before and checking for normality, no population is ever truly normally distributed. Here we are looking to see if there are obvious deviations from normality as indicated by a curve or an S-shape.

These plots each identify consistency of the model with the assumptions inherent in the mathematical model $\epsilon_i \sim N(0, \sigma^2)$.

The bottomright plot shows the magnitude of the residuals plotted against the leverage of each point. Leverage is a metric that details how much influence each data point has on the fitted line. Points with high leverage are those that are far away from the mean of the predictor variable. These points can have a large influence on the fitted line and can be influential points. Whether they are influential points or not is determined by the Cook's distance, which is a measure of how much the fitted line changes when the point is removed.

Points with high leverage and high Cook's distance are influential points. This plot helps identify any such points.

If you find a point with high leverage and high Cook's distance, you should ask whether there is something special about this point. If it is highly influential, you can remove it, refit the model, and evaluate whether your conclusions depend on the inclusion of this point.

In this example with the Iris example, there are no obvious problems with the regression model.

Evaluating Model Fit

The model fit can be evaluated using the R^2 statistic. This statistic is the proportion of variance in the response variable that is explained by the predictor variable. The R^2 statistic tells us whether the regression line explains a large or small amount of the variation in the repsonse variable. R^2 values always are between 0 and 1. When it is small, it means that the points do not fall close to the line, indicating that other factors beside the predictor are needed to describe any given point. When R^2 is large, it means that the points fall close to the line, indicating that the predictor variable explains a lot of the variation in the response variable.

The R^2 statistic is calculated as the proportion of the total sum of squares that is explained by the regression model. The total sum of squares is the sum of the squared differences between each data point and the mean of the response variable. The regression sum of squares is the sum of the squared differences between each data point and the fitted line. The residual sum of squares is the sum of the squared differences between each data point and the fitted line.

For the Iris example, the R^2 statistic is 0.93, indicating that the regression line explains 93% of the variation in petal width. This is a large amount of variation and indicates that the predictor variable, petal length, is a good predictor of petal width. You can see this visually because most points fall very close to the line. In essence, if I tell you the petal length of a flower, you can predict the petal width with a high degree of precision.

Making Predictions

The lines are not everything. The line only represents the mean of the response for a given value of the predictor. To see this, we can rewrite the generic linear regression model as follows:

$$Y_i \sim \text{Normal}(\mu = \beta_0 + \beta_1 X_i, \sigma^2).$$

In this writing, it says that every point Y_i is normally distributed with a mean of $\beta_0 + \beta_1 X_i$ and a variance of σ^2. This means that the points are not all on the line, but rather they are distributed around the line.

We can use this information to make predictions. For example, if we have an Iris petal length of $X_i = 5$ cm, this model describes the probability distribution for the petal width as

$$Y_i \sim \text{Normal}(\mu = -0.363 + 0.416 \times 5, \sigma^2).$$

This is incomplete because we need a measure of variability around the mean, σ^2. This variability is also estimated in linear regression. In R, this value is estimated as a standard deviation (i.e. σ) and is referred to as the "residual standard error". You can see it in the summary model output below.

```
##
## Call:
## lm(formula = Petal.Width ~ Petal.Length, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.56515 -0.12358 -0.01898  0.13288  0.64272
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
## Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2065 on 148 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266
## F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

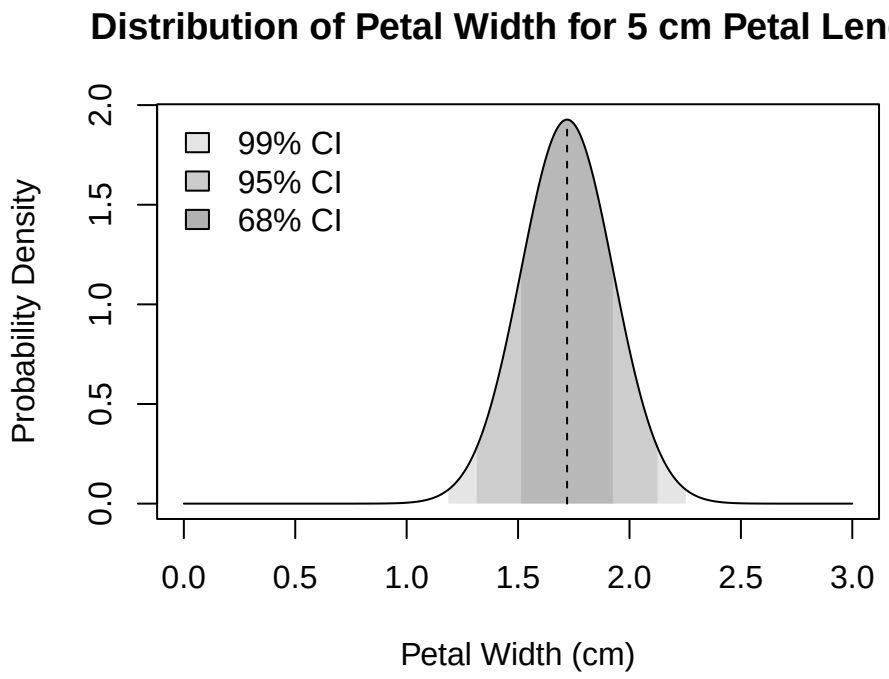
This summary says that the residual standard error is 0.207, meaning that we can write the full probability distribution for petal width as

$$Y_i \sim \text{Normal}(\mu = -0.363 + 0.416 \times 5, \sigma^2 = 0.207^2).$$

Doing all the algebra gives us

$$Y_i \sim \text{Normal}(\mu = 1.72, \sigma^2 = 0.0428).$$

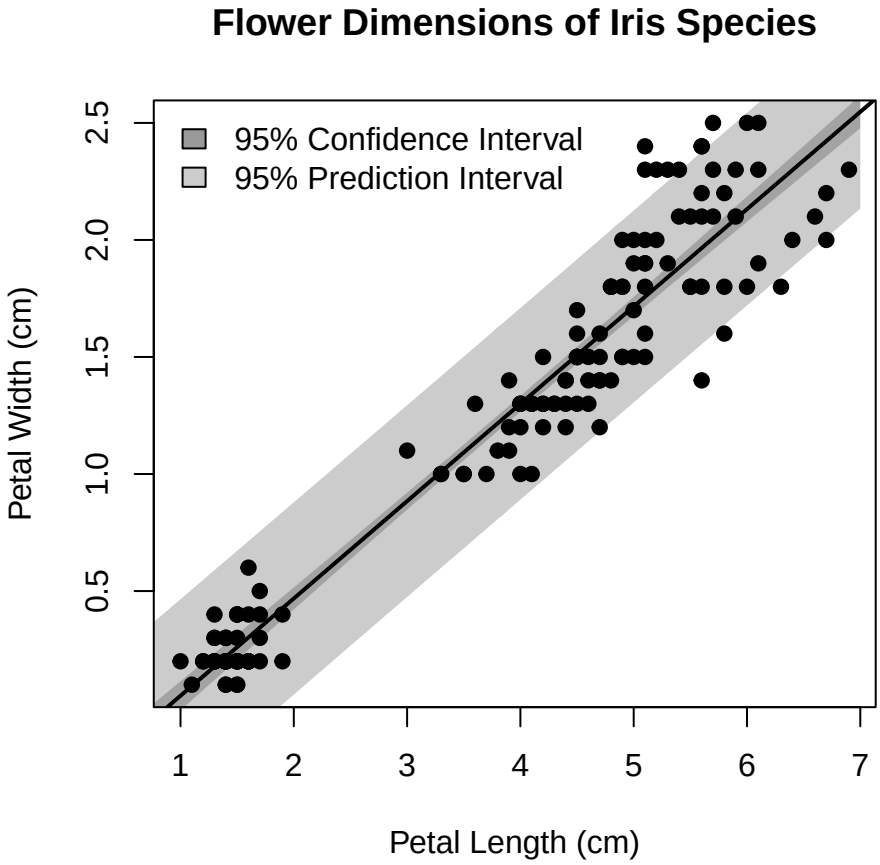
Visually this looks like the plot below



We can now make probabilistic predictions about the nature of flower petal widths given flower petal lengths. This model says that 68% of flowers with a length of 5cm have a width in the range of 1.72 +/- 0.207 cm, or between 1.51 and 1.93 cm. This also means that 95% of flowers with a length of 5cm have a width in the range of about 1.72 +/- 2 standard deviation of the mean, which is between 1.31 and 2.13 cm.

We can visually show this on the original scatterplot by making *prediction intervals* which show where 95% of the data points lie under the best fit model.

The figure below shows the iris data set, the 95% confidence interval showing where 95% of best fit lines are, and the prediction interval showing where 95% of the data points are. The prediction interval is wider than the confidence interval because it includes the variability of the data points around the line.


Flower Dimensions of Iris Species
95% Confidence Interval
95% Prediction Interval
Petal Width (cm)
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7
Petal Length (cm)