Forum 7
Data Skills for Scientists Regression: One Numeric Predictor

2025-09-24

Introduction to Regression

Response and Predictor Variables

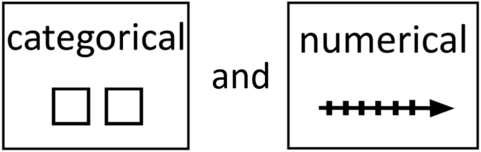
Lets start by introducing some terminology for some ideas we've already been using throughout the course.

A **response variable**, also called a dependent variable, measures an outcome of a study.

A **predictor variable**, also called an explanatory or independent variable, is used to predict, explain, or cause changes in the response variable.

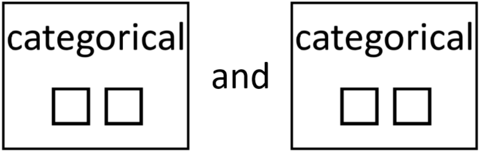
Example - Two-Sample T-Test

In a two-sample t-test, and also the Mann-Whitney U-Test, we have a numeric response variable, the outcome of interest from our study, and a categorical predictor with two categories which defines our groups.



Example - Chi-Square Test

In a Chi-square test, we have two categorical variables and we are looking for an association between them. Sometimes, it can be reasonable to consider one of these our response or outcome, and the other our predictor or explanatory variable.



categorical

□ □

and

categorical

□ □

Regression

As it turns out, all of these and more can fit into a general framework called regression.

In regression we will always have:

- one response variable, and
- some number of predictor or explanatory variables.

In this course, the response variable in regression will always be numeric and continuous, but beyond this course regression can be extended to include other kinds of response variables.

Regression

This third and final module, ‘Regression’ contains three topics, in which we explain the variation in a single numeric continuous response variable by using different kinds of predictor or explanatory variables.

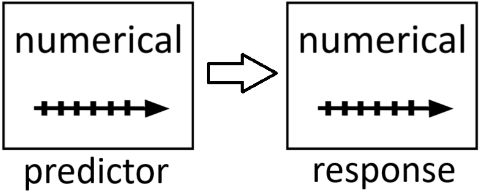
In particular, we will use:

- One Numeric Predictor;
- Multiple Numeric Predictors; and
- Categorical Predictors.

Which will each make one of the three topics in this module.

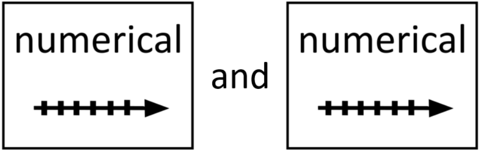
One Numeric Predictor

In this forum we will cover the first topic, using one numeric predictor to explain the variation in a numeric response variable.



Connecting the Dots

Regression of one numeric predictor follows on from the introduction to describing associations between two numeric variables in the second topic of the first module of the course, ‘Data: Summarising and Visualising’.



numerical and numerical

When describing associations however, the two numeric variables are treated interchangably. In regression, we will treat one as a predictor for the other, the response.

Before we get to that though, lets briefly revisit describing associations between two numeric variables.

Scatterplots and Correlation

In the second topic of the first module of the course, ‘Data: Summarising and Visualising’, we introduced visualising two numeric variables in a scatterplot and summarising linear associations by calculating the correlation between them.

In a scatterplot, it is traditional to plot:

- the response variable on the vertical or y-axis, and
- the predictor or explanatory variable on the horizontal or x-axis.

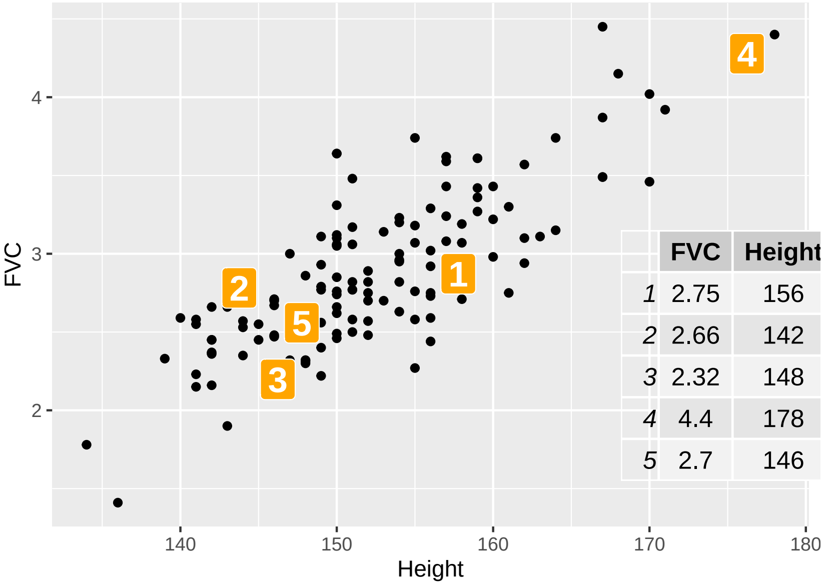
Example - FVC

Remember the FVC data set?

- **Response variable**: forced vital capacity (FVC) in litres.
- **Explanatory variable**: height in centimetres.

Let's take a look at a scatterplot of these data, and highlight the first few subjects to illustrate and recap how a scatterplot visualises data.

Example - FVC Scatterplot



	FVC	Height
1	2.75	156
2	2.66	142
3	2.32	148
4	4.4	178
5	2.7	146

Recap - Describing Associations

We described associations between two numeric variables with four characteristics:

- Direction
- Shape
- Strength
- Outliers

Direction and Shape

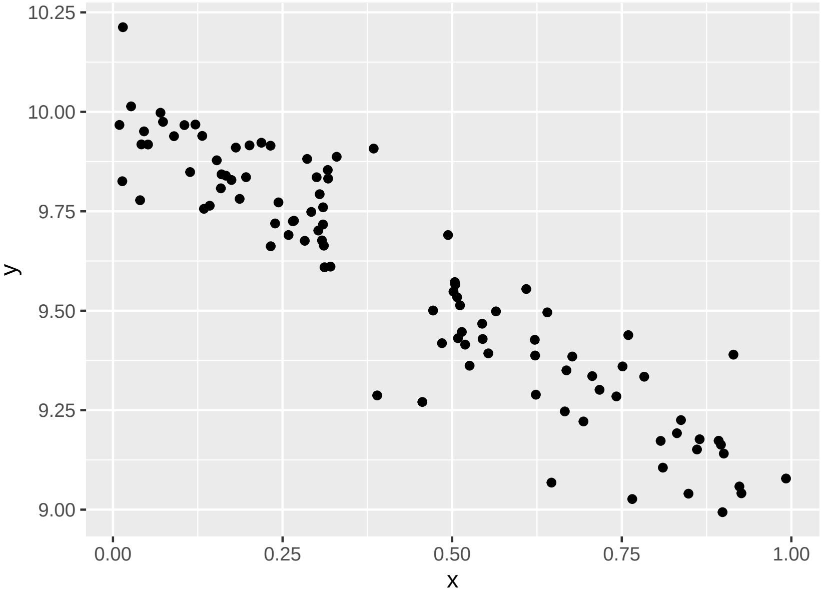
Describe the overall pattern the data follow in a scatterplot, and have direction that is either:

- Positive: as one variable increases, so does the other variable, or
- Negative: as one variable increases, the other variable decreases.

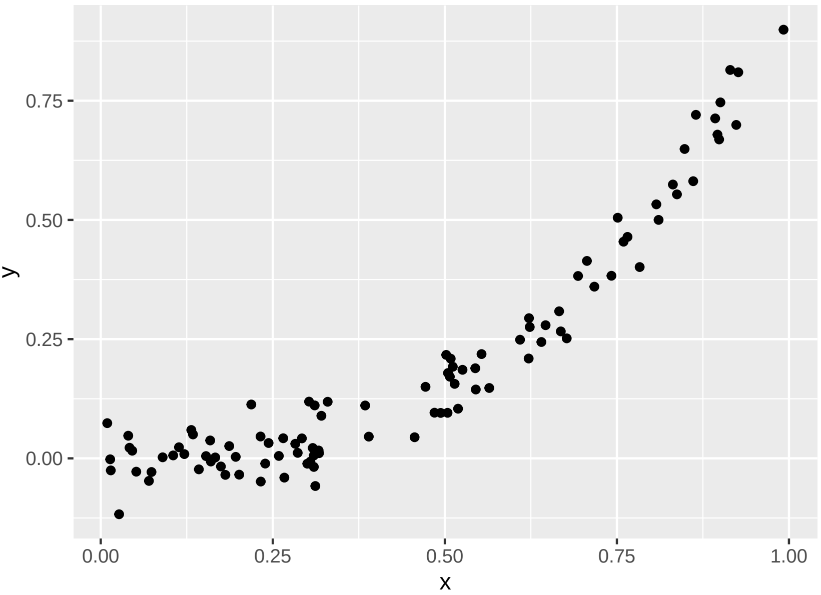
and shape that is either

- Linear: straight-line associations, or
- Non-linear: everything else.

Example - Negative and Linear Association



Example - Positive and Non-linear Association



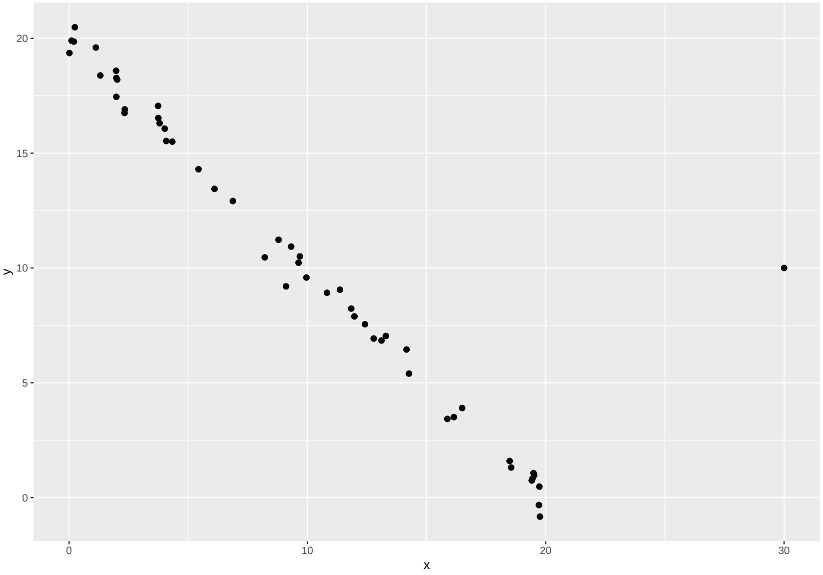
Strength and Outliers

Relate to how the data are distributed around the line that follows the overall pattern. The strength of an association describes how far away from that line the data are spread and can be where

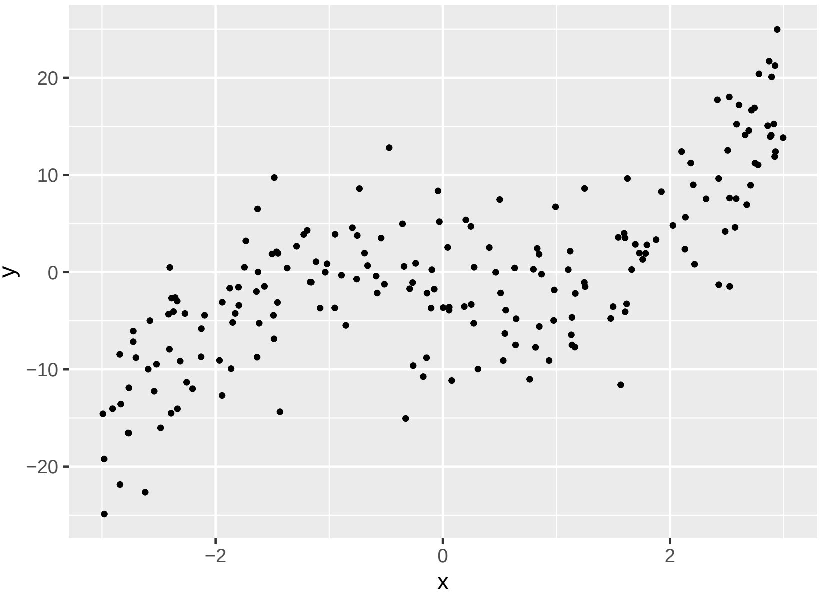
- A **strong** association is one where the data is tightly clustered around the line, and
- A **weak** association is one where the data is spread widely around the line.

and outliers describe data points that are unusually far from or contradict the overall pattern.

Example - Strong Negative Linear Association with Outlier



Example - Moderate Positive Non-linear Association



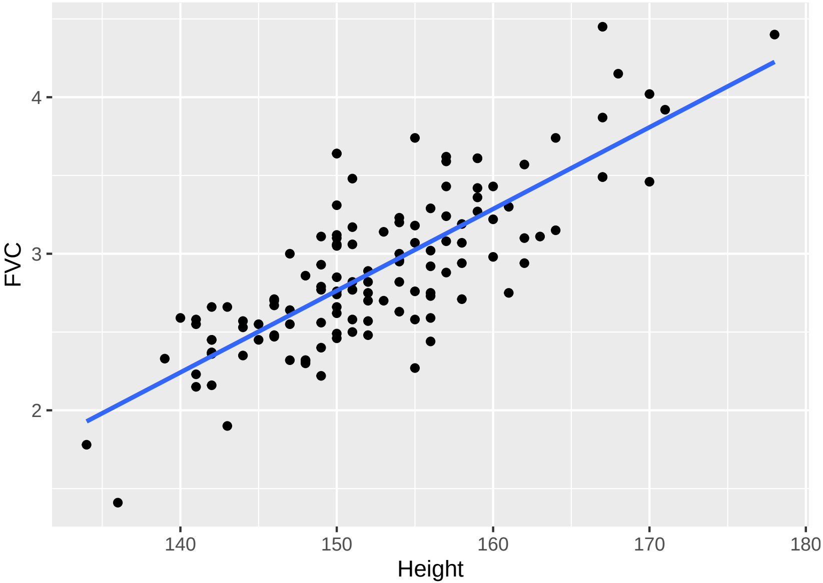
Regression line

In regression, we will try to find the ‘best’ line that describes the overall pattern of the data, and we call this the **regression line**. The regression line is also called:

- The least squares line;
- The line of best fit;
- Least-squares regression line;
- Linear regression line.

‘Least squares’ is a term that relates to what we mean by the ‘best’ line. More on this later.

Example - FVC Regression Line



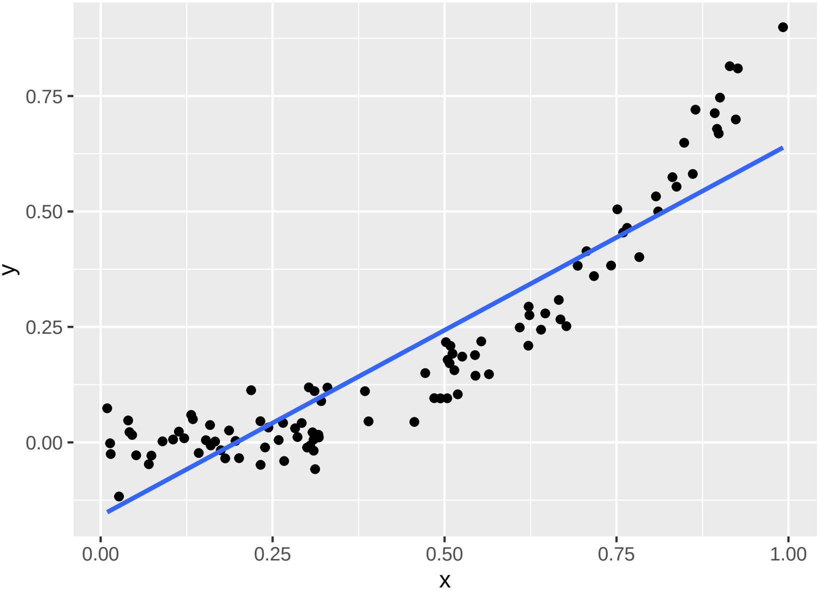
Non-linear regression

We will only try to estimate linear associations.

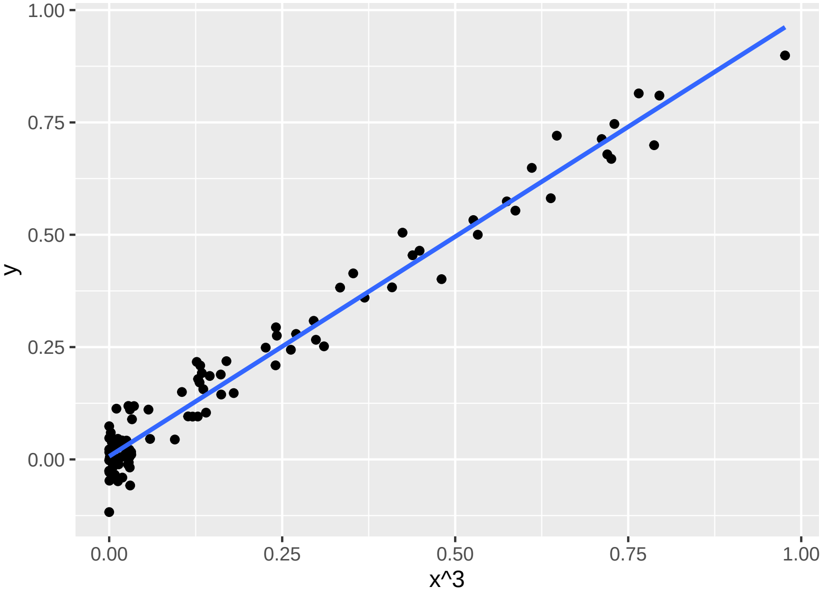
However, we can use transformations to create a linear association between transformed variables.

For example, the non-linear association between x and y shown earlier can be transformed into a linear association between x^3 and y:

Example - Association between x and y



Example - Association between x^3 and y



Topic Summary

In todays forum, we will explore regression on one numeric predictor, like in the FVC example. We'll introduce concepts in sections:

- Fitting a regression line to data,
- Inference for regression parameters,
- Checking model assumptions, and
- Making predictions.

Although we will introduce these concepts in this order,

> ℹ Note
>
> Normally you would check your model assumptions first before doing inference.

Fitting a Regression Line

Equations for Lines

You might (or might not) remember seeing the equation for a straight line written as

$$y = mx + c$$

where m is the slope and c is the intercept.

If you haven't, it could be worth revising this topic!

Equation for the Regression Line

For regression on one numeric predictor we will use the equation

$$y = b_0 + b_1 x$$

where

- x is the value of the explanatory variable.
- y is the value of the response variable for the corresponding x.
- b_1 is the **slope**, the amount by which y changes for each one-unit increase in x.
- b_0 is the **intercept**, the value of y when $x = 0$.

Interpretation of the intercept in context

If the <explanatory variable> is zero <units>, then we expect the the mean value of <response variable> to be <b_0> <units>.

Interpretation of the slope in context

If the <explanatory variable> increases by one <unit>, then we expect the mean value of the <response variable> to <increase/decrease> by <$|b_1|$> <units>.

Fitting a Regression Line

The premise of regression is not only to describe the association between two numeric variables, but to use the explanatory variable(s) x to make predictions about the response variable y.

So we want the regression line to be as close as possible to the data, specifically in the vertical (y) direction.

Fitting a regression line means to find the slope b_1 and intercept b_0 that make the vertical distances from the data to the line as small as possible.

Example - FVC

We will fit regression lines in R using the `lm()` function, and if we save the result to a variable and use the `summary()` function on it, it will summarise the result of the model fitting.

This summary contains a lot of information, and we're only interested in the values of b_0 and b_1 for now.

Example - FVC - Fit

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.0650     0.5526 -9.1658        0
Height        0.0522     0.0036 14.4264        0
```

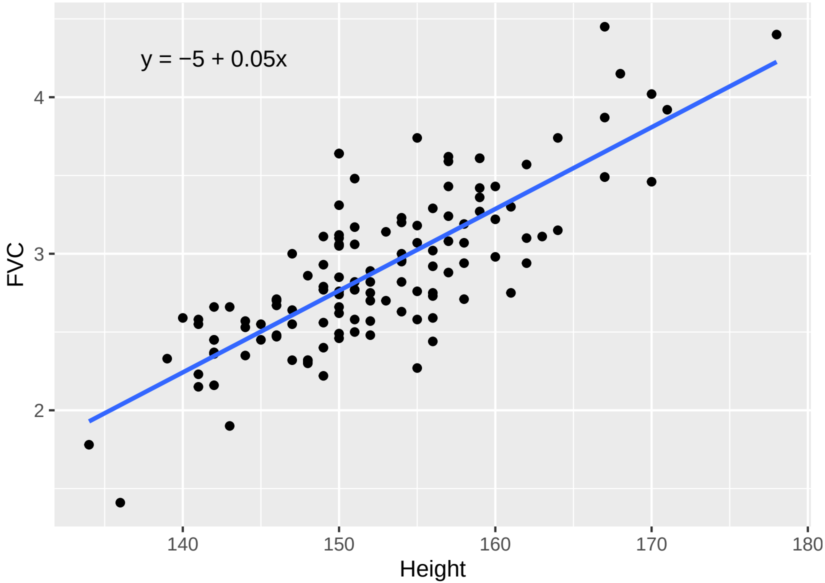
The `Estimate` column gives us our sample statistics:

- $b_0 \approx -5$ in the `(Intercept)` row, and
- $b_1 \approx 0.05$ in the `Height` row (because `Height` is our predictor).

So in this case, our fitted regression line is approximately

$$y = -5 + 0.05x$$

Example - FVC - Equation



Example - FVC - Interpretation of Slope

If the Height increases by one centimetre, then we expect the mean forced vital capacity (FVC) to increase by 0.05 litres.

We can adjust the units to make this interpretation a little nicer, for example:

- If the Height increases by 10 centimetres, then we expect the mean forced vital capacity (FVC) to increase by half a litre, or
- For every centimetre (cm) that Height increases, we expect the mean forced vital capacity (FVC) to increase by 50 millilitres (mL).

The slope being positive means that the direction of the association is positive.

Example - FVC - Interpretation of Intercept

If the Height is zero centimetres, then we expect the mean forced vital capacity (FVC) to be negative 5 litres.

What does negative volume mean? Clearly that makes no sense.

Why is that?

Interpolation vs Extrapolation

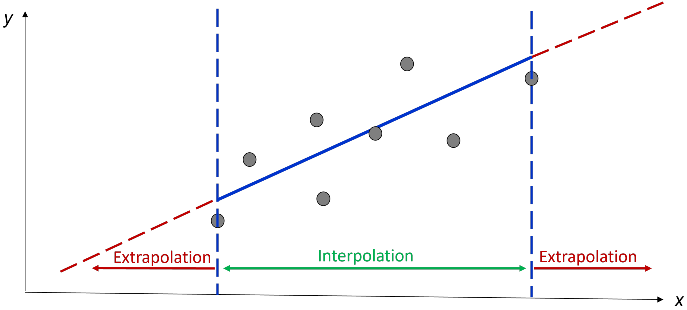
When we make predictions and interpretations within the range of our data, it is called **interpolation**.

When we extend our model to make predictions and interpretations beyond the range of our data, it is called **extrapolation**.

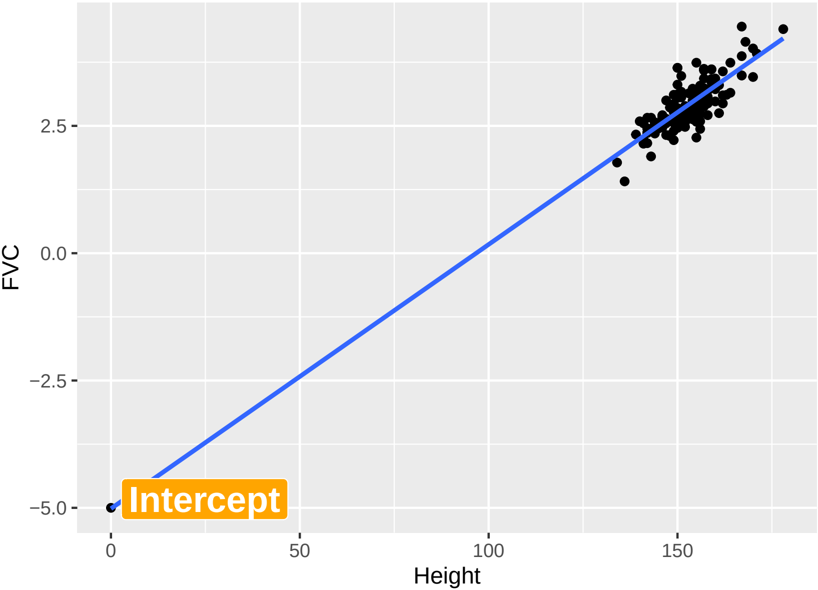
⚠ Warning

Extra care is always appropriate when extrapolating. Sometimes it can make sense, but other times the context might make it questionable or even completely unreasonable.

Interpolation vs Extrapolation



Example - FVC - Intercept



Example - FVC - Extrapolation

All the subjects of the FVC study have heights between 130cm and 180cm.

Extrapolating to much shorter individuals such as small children and infants likely doesn't make much sense as they might have quite different physiology to more fully grown people.

Extrapolating for an individual with zero height makes no sense because it is impossible for a person to have zero height! Especially while having lungs on which we could measure forced vital capacity! The physical reality of the context means that extrapolating this far is unreasonable.

So the intercept doesn't have a meaningful or useful interpretation in this context. That's ok — it doesn't need too, it allows the equation of our regression line to fit the data properly.

Correlation and Regression

Least-squares regression looks at the distances of the data points from the line only in the y direction.

As a result, the distinction between explanatory (X) and response (y) variables variables is essential in regression.

Even though correlation r ignores the distinction between x and y, there is a close connection between correlation and regression.

Coefficient of Determination

The square of the correlation, r^2, is the proportion of the variance of y that is explained by x through the regression line.

r^2 is called the **coefficient of determination**.

We can find this value in the full `summary()` of a regression fit in R.

Example - FVC - Fit

```
Call:
lm(formula = FVC ~ Height, data = fvc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75507 -0.23898 -0.00411  0.21238  0.87589

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.064961   0.552593  -9.166 1.24e-15 ***
Height       0.052194   0.003618  14.426  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3137 on 125 degrees of freedom
Multiple R-squared:  0.6248,    Adjusted R-squared:  0.6218
F-statistic: 208.1 on 1 and 125 DF,  p-value: < 2.2e-16
```

Example - Coefficient of Determination

There are two similar values for the coefficient of determination reported in this summary, `Multiple R-Squared` and `Adjusted R-Squared`, don't be too concerned about distinguishing these right now, and both are approximately 0.62 in this case anyhow.

We can interpret this value by reporting that 62% of the variance in forced vital capacity can be explained by variation in height, through the regression line we've fitted to these data.

Inference for Parameters

Population vs Sample - Parameters vs Statistics

Remember how when we are interested in a population mean μ, we use the sample mean $\bar{x}$ to estimate it since we don't know the true value of μ?

It's the same in regression. We are interested in β_1 and β_0, the slope and intercept of the population regression line

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

Inference for Slope and Intercept

Since we don't know what β_1 and β_0 are, we use our data to estimate them with the sample statistics b_1 and b_0 much like how we used $\bar{x}$ to estimate μ earlier in the course.

We can make inferences about the population parameters β_1 and β_0 in a number of ways, in particular with

- Confidence intervals, and with
- Hypothesis tests.

Confidence Intervals - Recap

Recall that to calculate a $C\%$ confidence interval for the population mean μ, we used the formula

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

Where

- $\bar{x}$ is our sample statistic estimate for μ,
- t^* is the critical value from the appropriate t-distribution, calculated to give the correct confidence level C, and
- $\frac{s}{\sqrt{n}}$ is the standard error of $\bar{x}$.

We use the same format to calculate confidence intervals for β_1 and β_0.

Confidence Intervals for Slope and Intercept

A $C\%$ confidence interval for the intercept β_0 is

$$b_0 \pm t^* \times SE_{b_0},$$

and a $C\%$ confidence interval for the slope β_1 is

$$b_1 \pm t^* \times SE_{b_1},$$

where t^* is obtained from a t-distribution with $n-2$ degrees of freedom, and the standard errors SE_{b_0} and SE_{b_1} are conveniently provided in the R output.

Example - FVC

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.0650     0.5526 -9.1658        0
Height        0.0522     0.0036 14.4264        0
```

So $SE_{b_0} = 0.55$ and $SE_{b_1} = 0.0036$.

Example - Confidence Interval for Slope

Calculate a 95% confidence interval for the slope. Interpret this interval in context.

You may assume $t^* = 1.98$ (this is calculated from a t-distribution with 125 degrees of freedom, since the FVC data set contains $n = 127$ subjects).

Example - Solution

A 95% confidence interval for the slope β_1 is given by

$$b_1 \pm t^* \times SE_{b_1},$$

and we have $b_1 = 0.0522$, $t^* = 1.98$ and $SE_{b_1} = 0.0036$. So our confidence interval is

$$0.0522 \pm 0.0072 \quad \text{or} \quad (0.0450, 0.0594)$$

rounding outwards to be conservative on the level of confidence.

We are 95% confident that for every centimetre (cm) that height increases, we expect the mean forced vital capacity (FVC) to increase between 45 and 60 millilitres (mL).

Hypothesis Testing for a Significant Association

We can also use hypothesis tests to answer specific questions about our regression model.

A common question is:

> *"Does the explantory variable help predict the response variable?"*

This question is equivalent to testing whether the slope β_1 is significantly different from zero.

Null and Alternative Hypotheses for Slope

To test if the slope is non-zero, we use the following null and alternative hypotheses:

$$H_0 : \beta_1 = 0,$$
$$H_a : \beta_1 \neq 0.$$

Test statistic

The appropriate test statistic for these hypotheses is

$$T = \frac{b_1}{SE_{b_1}},$$

where b_1 is the estimated slope, and SE_{b_1} is the standard error of the estimated slope.

If the null hypothesis is true, then the test statistic has a T-distribution with $n-2$ degrees of freedom.

The P-value is obtained using a t-distribution with $n-2$ degrees of freedom.

Conveniently, both the test statistic and p-value are already provided in the summary of a model fit:

Example - FVC

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.0650     0.5526 -9.1658        0
Height        0.0522     0.0036 14.4264        0
```

So $T = 14.4$ and the p-value is 0 (or close to zero).

Note that there is also a t-statistic and p-value for the intercept term, which are for a similar hypothesis test for the null hypothesis that the intercept is zero $H_0 : \beta_0 = 0$

Checking Assumptions

Fitted Values and Residuals

Whenever we fit a model or use it to perform inference or prediction, we need to consider what assumptions we are making.

In order to check the assumptions of a linear regression model, we use:

- Fitted values, also called predicted values or predicted response values, and
- Residuals, also called errors.

So we'll introduce these concepts first before we get to the assumptions themselves.

Fitted Values

For a data point with coordinates (x_i, y_i) on the scatterplot we use x_i in combination with our fitted regression line to predict its response variable value.

This predicted response value is the y coordinate of the point on the regression line with the same x value as the data point, x_i, and is called the **fitted value** for that data point $\hat{y}_i$.

We can do this for each data point, so the ith data point will have a fitted value:

$$\hat{y}_i = b_0 + b_1 \ x_i$$

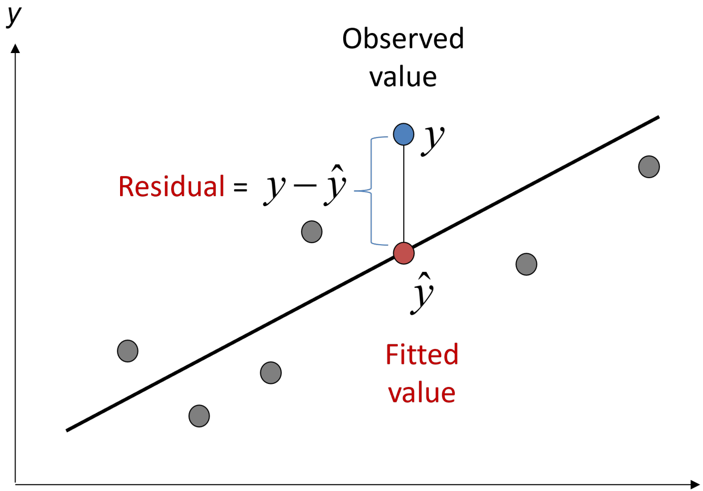
Residuals

The error in this prediction, also called the **residual** r_i, is the difference between the predicted response value $\hat{y}_i$ and and the actual response value for the data point y_i:

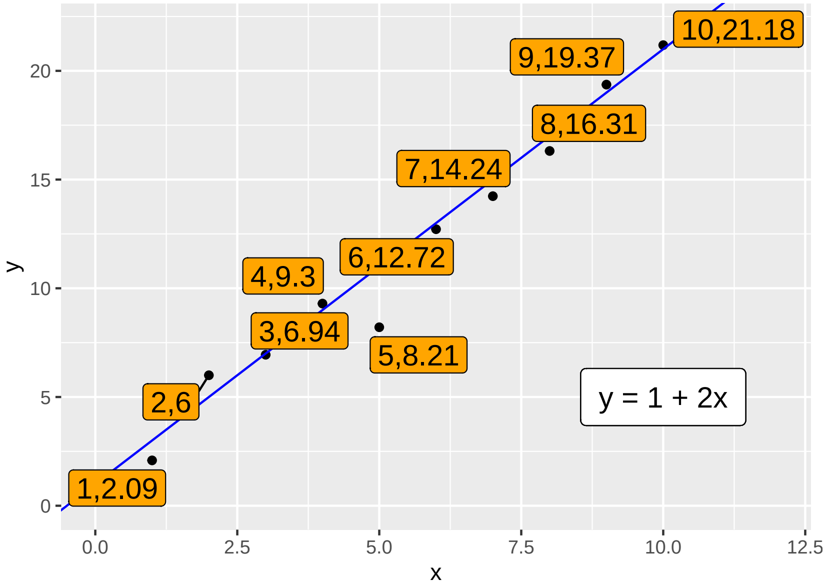
$$r_i = y_i - \hat{y}_i,$$

It can be useful to visualise this on a scatterplot in order to better understand what the fitted values and residuals are.

Fitted Values and Residuals



Example - Calculate $\hat{y}_i$ and r_i for the point with $x_i = 5$



Example - Solution

Since $x_i = 5$, we can use the equation for the regression line $y = 1 + 2x$ to calculate its fitted value

$$\hat{y}_i = 1 + 2 \times x_i = 1 + 2 \times 5 = 11$$

.

From the scatterplot, we can see that the actual response value for this point is $y_i = 8.21$, so we can calculate its residual

$$r_i = y_i - \hat{y}_i = 8.21 - 11 = -2.79$$

The residual being negative indicates the point is below the line, which matches what is shown in the scatterplot.

Assumptions of a Linear Regression Model

Now that we have the concepts of fitted values and residuals, we can look at the assumptions of a linear regression model. There are four:

- Linearity.
- Constant spread of residuals.
- Normality of the residuals.
- Independence of the residuals.

Checking Assumption - Linearity

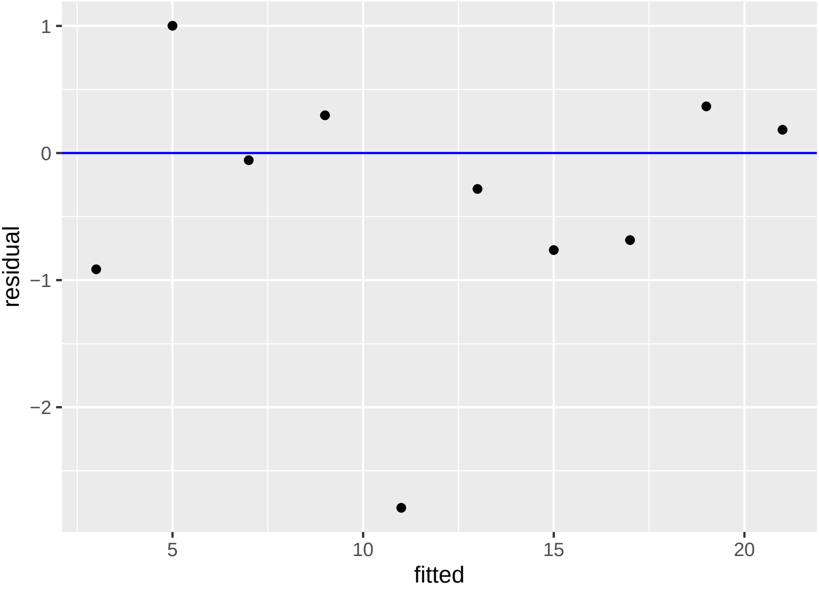
When fitting a linear regression model, we assume the overall pattern of the data has a **linear shape**.

Although this can be assessed to degree from a scatterplot of x and y, it is better to check linearity in a scatterplot of fitted values and residuals instead, as this can show deviations from linearity that might not be as obvious in a x and y scatterplot.

For the linearity assumption to be reasonable, the points in this plot should be **symmetrically scattered** above and below the horizontal zero line (corresponding to a residual of zero).

Following on from the previous example,

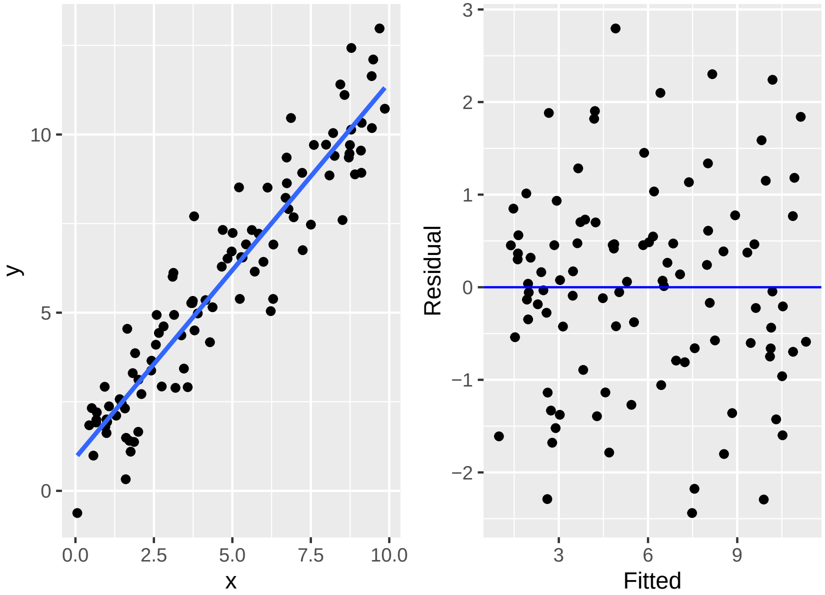
Example



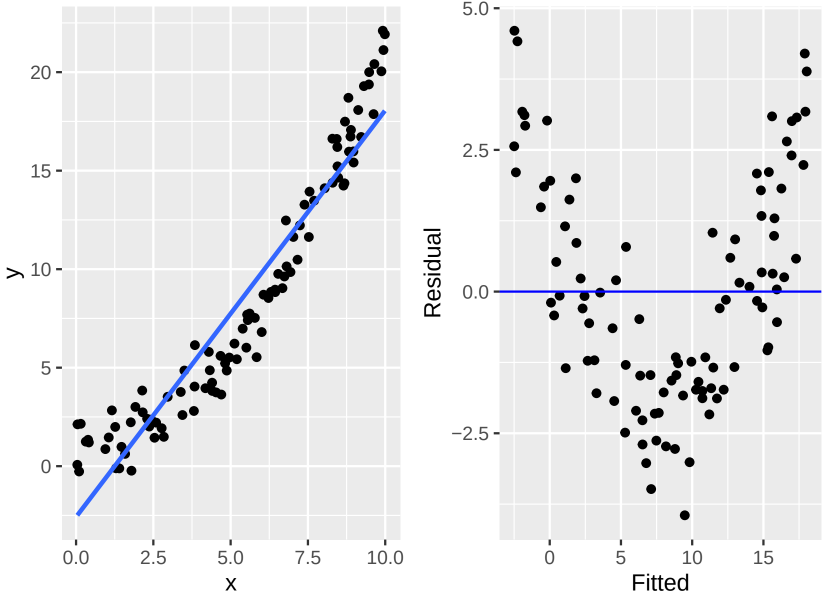
Example

In this example, we can see that the assumption of linearity is fairly reasonable, with the notable exception of one outlier, which happens to be the same one we calculated the residual for (-2.81) in the previous example.

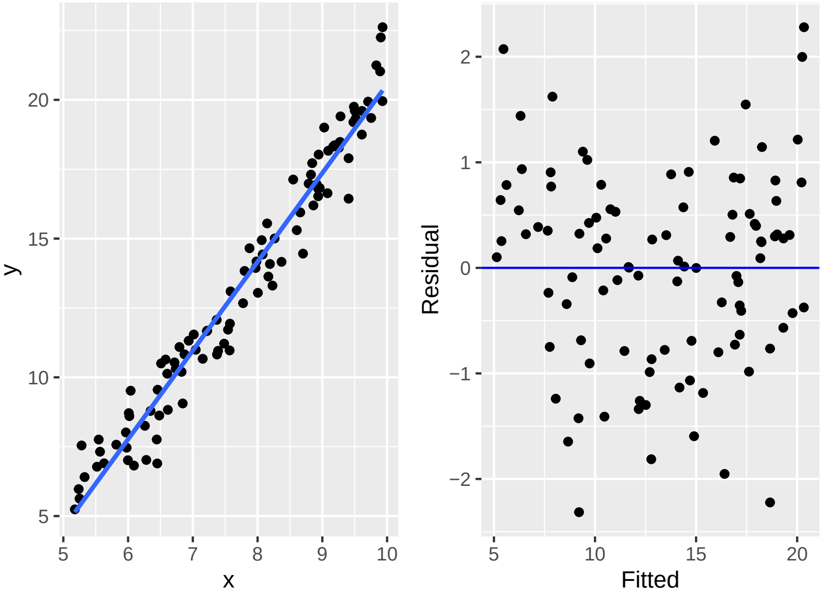
Example - Reasonable Linearity



Example - Unreasonable Linearity



Example - Less Obvious Unreasonable Linearity



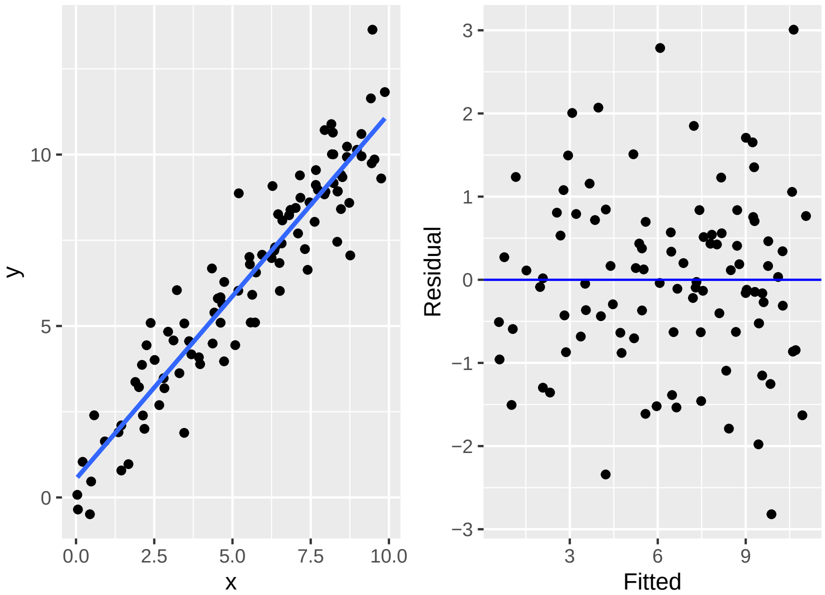
Checking Assumption - Constant Spread

When fitting a linear regression model, we assume the spread of residuals is the same along the line, i.e. for different fitted values.

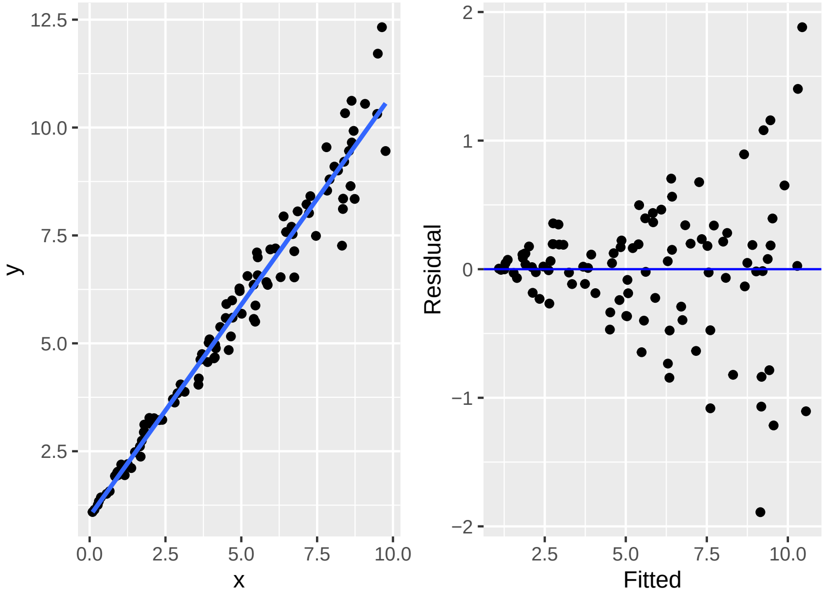
We use the same scatterplot of fitted values and residuals to check this assumption, but what we are looking for is different compared to when we check for linearity.

For the constant spread assumption to be reasonable, the points in this plot should be **roughly equally spread** in the vertical (residual) direction along the various horizontal (fitted value) positions in the plot.

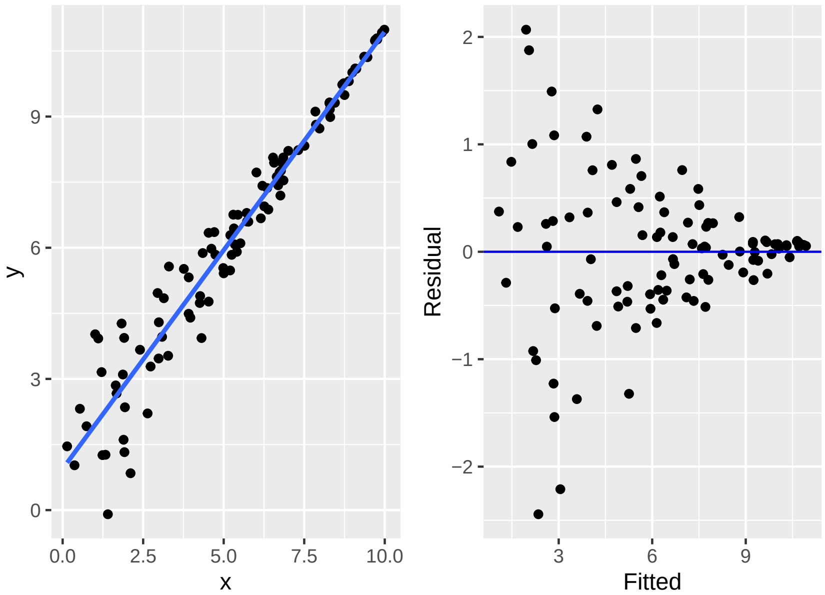
Example - Reasonable Constant Cpread



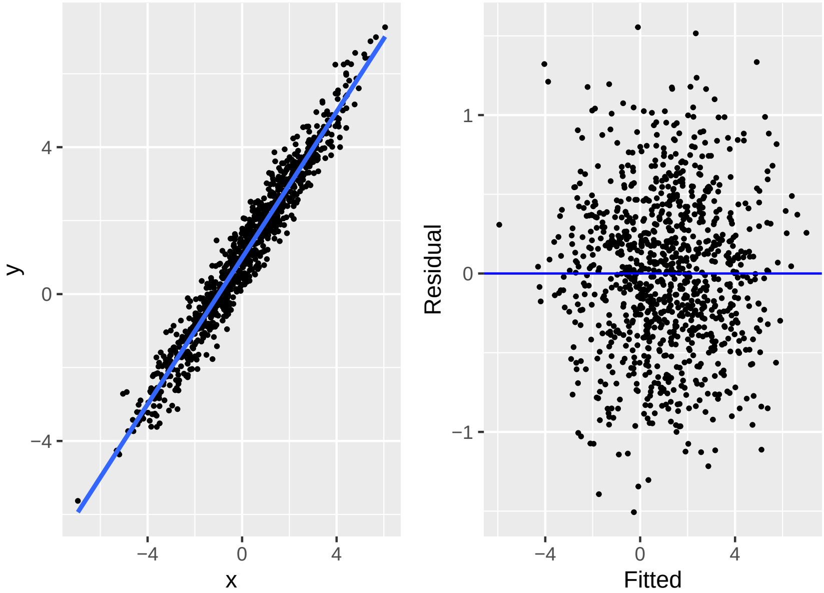
Example - Unreasonable Constant Spread



Example - Unreasonable Constant Spread



Example - Reasonable Constant Cpread



Example

This last example might seem like it has less spread on the left and right than in the centre, but it only seems like this because there are less data points on the left and right — highlighting how this can be tricky to assess.

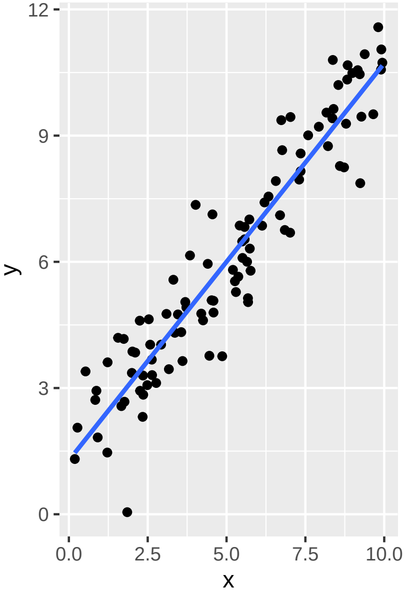
Checking Assumption - Normality

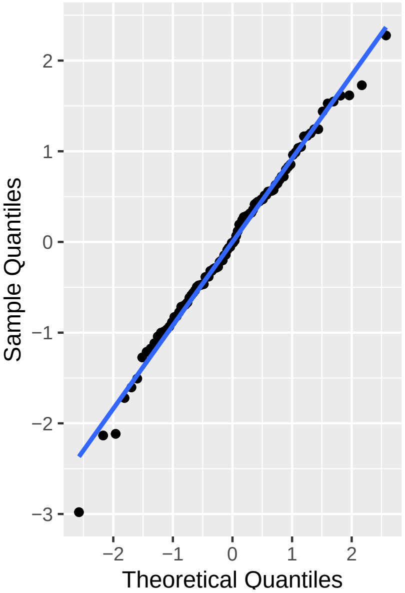
When fitting a linear regression model, we assume the residuals are normally distributed.

We use a normal quantile-quantile plot (or QQ-plot) to assess normality, as we have done throughout the course, but in this case it is a normal quantile-quantile plot of residuals, rather than of the data itself.

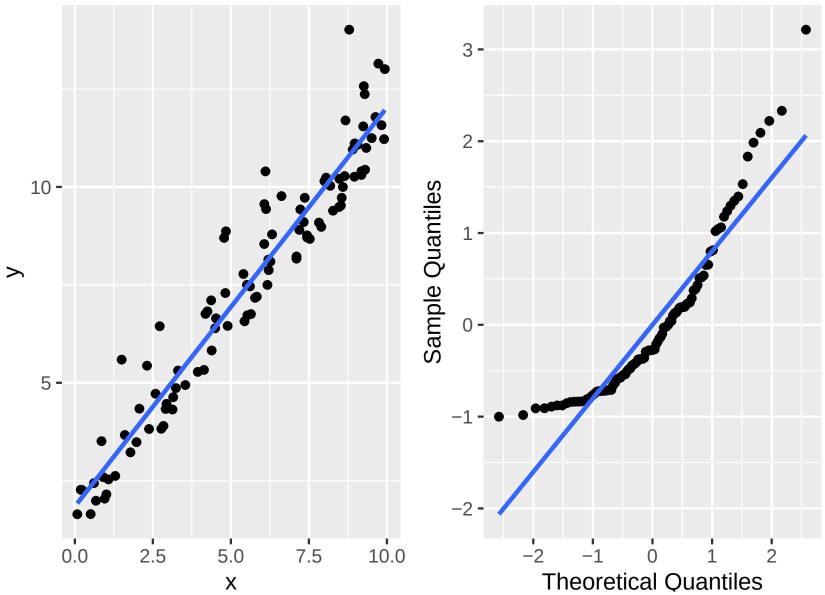
For the normality assumption to be reasonable, the points in the QQ-plot should be **roughly linear**.

Example - Reasonable Normality

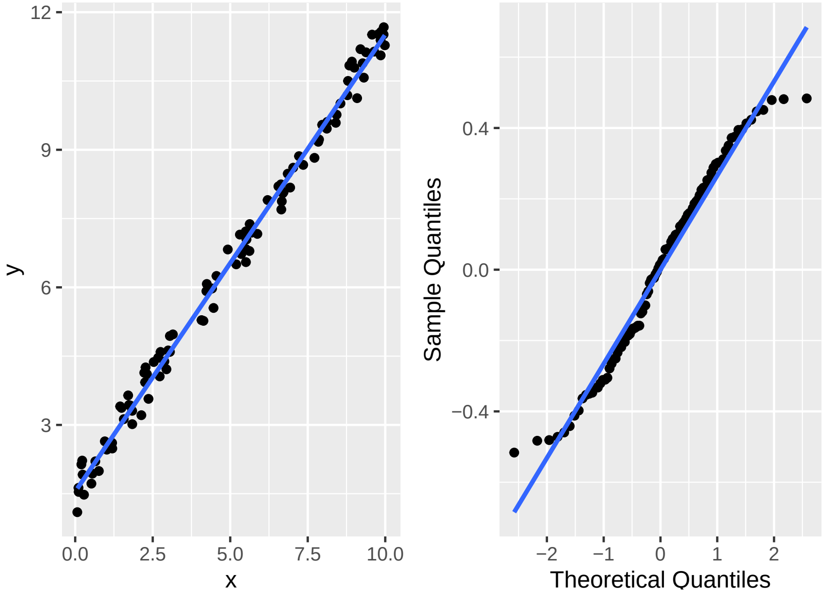




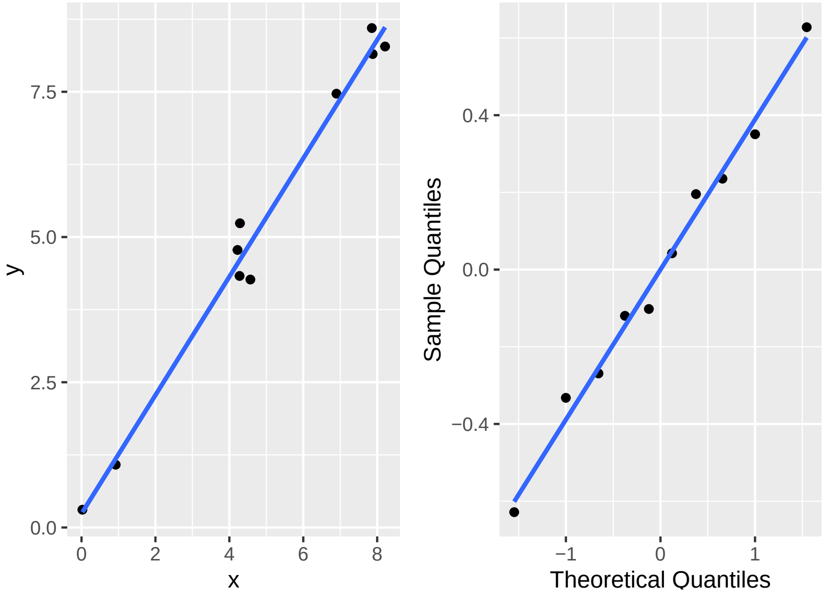
Example - Unreasonable Normality



Example - Unreasonable Normality



Example - Small Sample Size



Checking Assumption - Independence

When fitting a linear regression model, we assume that each observation is independently measured from the same population.

This means that one observation should not give us information about the other observations.

This assumption can be the trickiest to assess, as it cannot be directly checked from looking at the data, it can only be assessed by considering the experimental design.

However, looking at the data can sometimes give us clues that this assumption may be unreasonable even though the absence of such clues does not neccessarily mean the assumption is reasonable.

Examples - Unreasonable Independence

- **Clustering:** For example, if we had closely related subjects (twins, a litter of sibling puppies).
- **Spatial relationship:** For example, two neighbouring fields.
- **Temporal relationship:** For example, manufacturing process deteriorating over time.

Summary

Linearity:

- Consider the: Residual versus fitted plot.
- Reasonable if: Symmetrically scattered around zero line.
- Unreasonable if: Obvious pattern.

Constant spread:

- Consider the: Residual versus fitted plot.
- Reasonable if: Roughly equal spread left to right.
- Unreasonable if: Very different spread at different places.

Summary

Normality:

- Consider the: Normal quantile plot of residuals.
- Reasonable if: Roughly linear.
- Unreasonable if: Obvious pattern.

Independence:

- Consider the: Experimental design.
- Reasonable if: Observations do not give information about others, randomness.
- Unreasonable if: Relationship between some observations.

Making Predictions

Prediction

Now that we have our linear regression model and we've checked its assumptions we can use it to make two types of prediction:

- predict the **value** of the response variable for **an individual** with a given value of the explanatory value.
- predicting the **mean** value of the response variable for **all subjects** with a given value of the explanatory variable.

Predicting the Response for an Individual

Consider an individual with a known value of the explanatory variable x_0. If we want to estimate the individuals (unknown) response value y_0, the fitted value

$$\hat{y}_0 = b_0 + b_1 \times x_0.$$

already gives us a point-estimate for y_0. However we know that $\hat{y}_0 \neq y_0$, since our data have non-zero residuals! To take this into account, we should quantify our level of (un)certainty and calculate what is called a **prediction interval** for y_0.

Prediction Interval for an Individual

As in previous topics on confidence intervals, the form for a prediction interval for y_0 is

$$\hat{y}_0 \pm t^* SE_{\hat{y}_0},$$

where t^* is obtained from a t-distribution with $n-2$ degrees from freedom. There is a formula for the standard error $SE_{\hat{y}_0}$, but we will rely on R to calculate this for us rather than using the formula.

Example - FVC

We can use the `predict()` function on our fitted model in R to calculate a 95% prediction interval for someone with height 150cm:

```
       fit      lwr      upr
1 2.764106 2.140573 3.387638
```

Which shows us that the fitted value for a person with height $x_0 = 150$ is $\hat{y} = 2.76$, and that we would predict with 95% certainty that their FVC would be in the interval (2.14, 3.39).

- If a 12-year old boy has a height of 150 cm and a FVC of 2.3 litres, is this FVC unusually low?

Example - Solution

Since 2.3 is within our 95% prediction interval, you could argue that it is not unusually low.

However it is on the lower end, so at a lower level of certainty it might not be included in our prediction interval. In this way, the level of certainty in the prediction interval calculation can give you a measure of how unusual a value is, rather than answering the yes/no question of ‘is it unusual or not?’.

Predicting the Mean for Individuals with a given value of x

Consider all the subjects of the population that have a given value of the explanatory variable x_0.

If we knew the true linear relationship, we could calculate the mean value of the response variable for this sub-population

$$\mu_{y|x_0} = \beta_0 + \beta_1 \times x_0.$$

We use $\mu_{y|x_0}$ to denote this mean of the subpopulation of individuals with a given explanatory variable value of x_0.

Predicting the Mean for Individuals with a given value of x

We can estimate $\mu_{y|x_0}$ with the fitted value $\hat{y}_0$ as before

$$\hat{y}_0 = b_0 + b_1 \times x_0.$$

However $y = b_0 + b_1 \times x$ is only an estimate of $y = \beta_0 + \beta_1 x$, and so has some uncertainty associated with each of b_0 and b_1.

We can quantify this uncertainty by calculating a **confidence interval** for $\mu_{y|x_0}$ at a given level of confidence.

Confidence Intervals for a Mean

Our confidence interval for $\mu_{y|x_0}$

$$\hat{y}_0 \pm t^* SE_{\mu_{y|x_0}},$$

is very similar to the prediction interval for an individual's response

$$\hat{y}_0 \pm t^* SE_{\hat{y}_0},$$

but $SE_{\mu_{y|x_0}}$ is not the same as $SE_{\hat{y}_0}$. In general it will be less since we can estimate a mean more precisely than an individual value (remember the central limit theorem?).

There is a formula for $SE_{\mu_{y|x_0}}$ which is similar to (but not the same as) the formula for $SE_{\hat{y}_0}$, but we will rely on R to calculate this rather than using the formula.

Example - FVC

We can use the `predict()` function on our fitted model in R to calculate a 95% confidence interval as well, for someone with height 150cm this gives:

```
       fit      lwr      upr
1 2.764106 2.706084 2.822127
```

- Compare this to the prediction interval calculated earlier for something of the same height. How is it similar, how is it different, and would you interpret it?

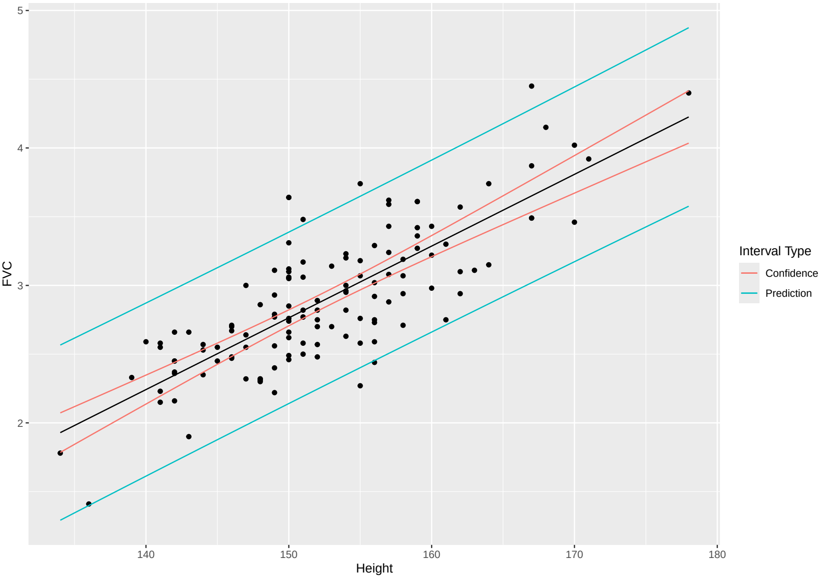
Example - Solution

We can claim that with 95% confidence, the the mean FVC of all individuals with height 150cm will be in the interval $(2.70, 2.83)$.

In contrast to the 95% prediction interval for the FVC of an individual with height 150cm $(2.14, 3.39)$, the confidence interval for the mean is much more precise — much narrower, since we can estimate the mean with much less uncertainty than we can estimate the value for an individual.

This is true in general:

Comparing PI to CI


FVC
Height
Interval Type
Confidence
Prediction