MAT220 Lecture 19: Linear Regression

Contents

In Lecture 18, we studied paired quantitative data by plotting the observations (x_i, y_i) on a scatter diagram. We then used the sample correlation coefficient r to describe the sign and strength of a linear relationship. Correlation itself merely tells us whether the points approximately fit a straight line. But, it does not actually construct the line itself. Today we will build this line, known as the *least-squares regression line*. We will start by identifying the key features of the regression line, and then we will use it to make predictions. Finally, we will introduce the *coefficient of determination*, which measures the quality of the regression line as a predictor.

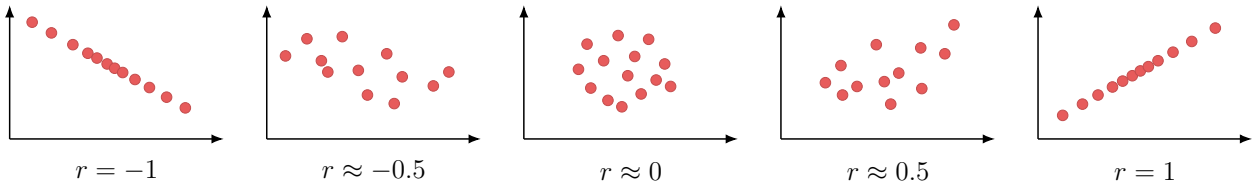
1 Linear Regression

1.1 The correlation coefficient

Suppose that we have n paired observations:

$$(x_1, y_1), \qquad (x_2, y_2), \qquad \ldots, \qquad (x_n, y_n).$$

In the previous lecture, we started to study the descriptive statistics of paired data. In particular, we looked at *linear correlation*, which described how closely the paired data followed a straight line. In the process, we introduced the *correlation coefficient*, which was a number that quantified the linear correlation between x and y. This number, which we denoted by r, ranged from -1 to 1, and the value of r dictated the strength and sign of the linear correlation present:



Algebraically, the sample correlation coefficient can be calculated using the formula:

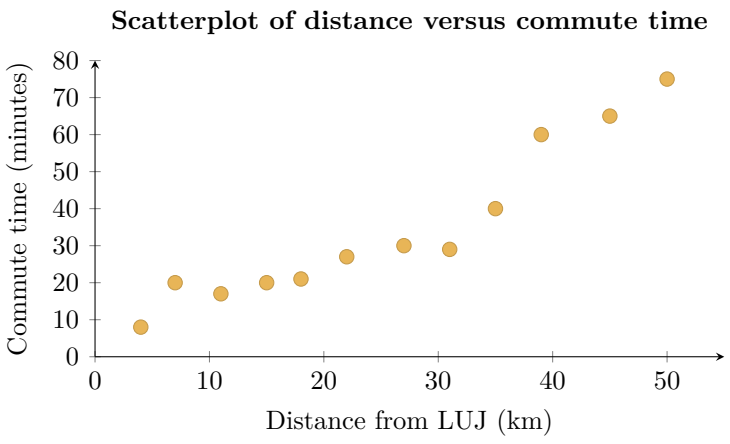
$$r = \frac{\text{cov}(x, y)}{s_x s_y},$$

where here the numerator is the covariance (a measure of the interrelatedness of the two variables) and s_x and s_y are the sample standard deviations in the x and y directions.

1.2 Example: distance and commute time

Throughout this lecture, we will use the following data set as a regular example. Suppose that twelve students record their distance from LUJ and their commute time to campus.

Student	Distance (km)	Time (min)
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75



For this data, we saw last time that

$$\bar{x} \approx 25.333, \qquad \bar{y} \approx 34.333, \qquad s_x \approx 15.011, \qquad s_y \approx 21.253, \qquad \text{and} \qquad \text{cov}(x, y) \approx 301.606.$$

Therefore,

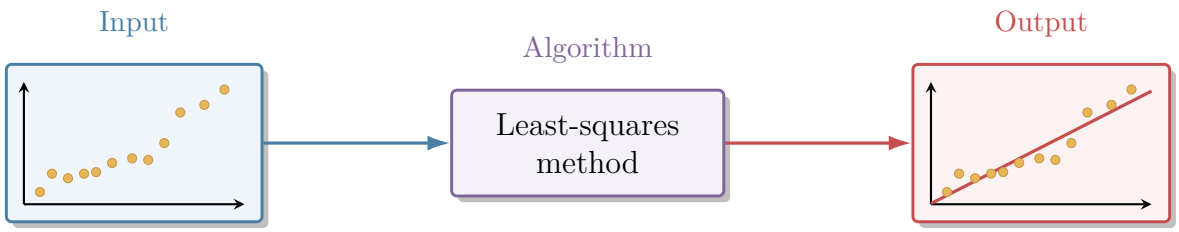
$$r = \frac{301.606}{(15.011)(21.253)} \approx 0.9454,$$

which indicates a strong, positive linear relationship between the two variables.

2 The Least-Squares Method

2.1 Regression as an algorithm

An *algorithm* is a fixed procedure that takes an input and reliably produces an output. In today's lecture, we will present an algorithm that takes paired data as an input, and produces the "line of best fit" as an output. For reasons that will be clear later, we will call this the *least-squares method*. Schematically, we have:



Recall that any non-vertical straight line can be fully described using two numbers:

- the *gradient*, which describes the rate of change of the line (i.e. the steepness), and
- the *y-intercept*, which is the value of the line when $x = 0$.

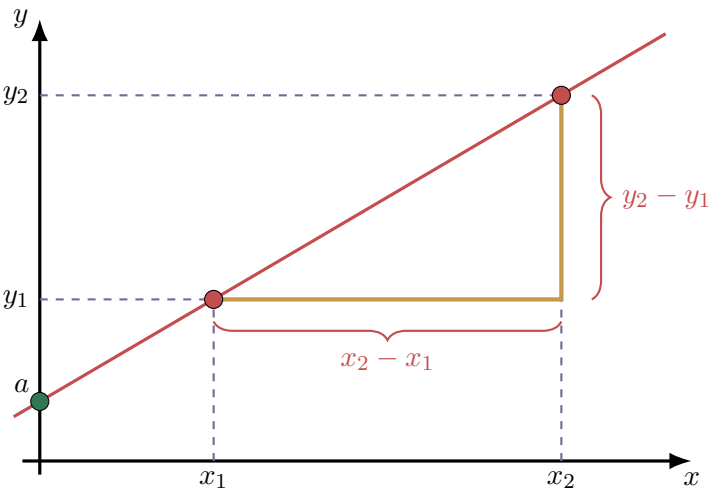
We will write our line using the notation:

$$\hat{y} = a + bx,$$

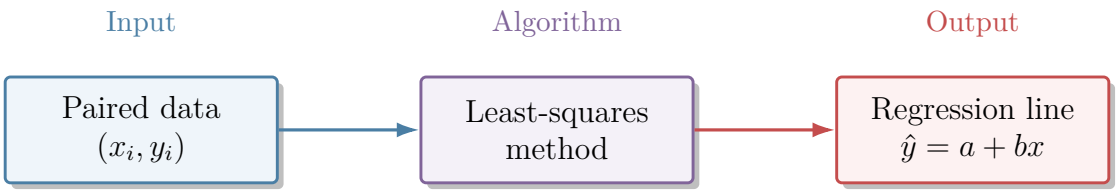
where the hat notation for the y will make sense later. Here, a is the value of the y-intercept, and b is the gradient of the line. The gradient can be calculated by picking two points on the line and using the formula:

$$b = \frac{y_2 - y_1}{x_2 - x_1},$$

that is, we take the vertical change and divide it by the horizontal change.

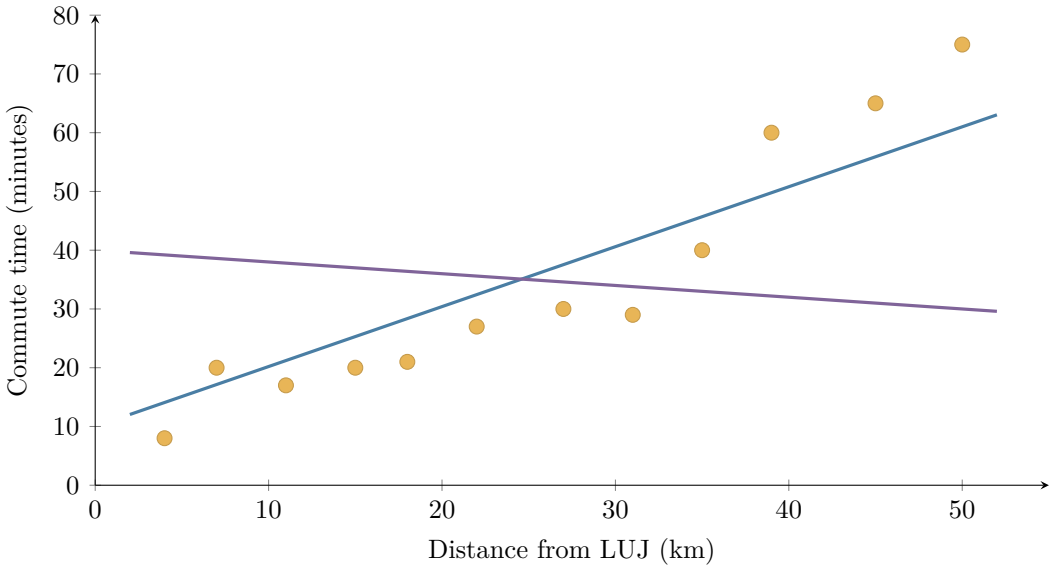


With this notation in mind, the output of the least-squares method is actually a pair of numbers a and b, so that the resulting line best fits the input data. Such a line is called a *regression line*. So, a better picture of our algorithm might be:



2.2 The least-squares criterion

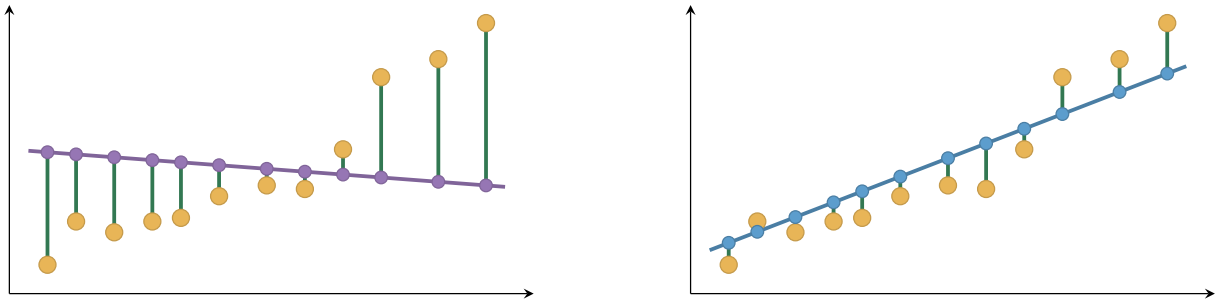
Using the data from Section 1, let's consider two lines:



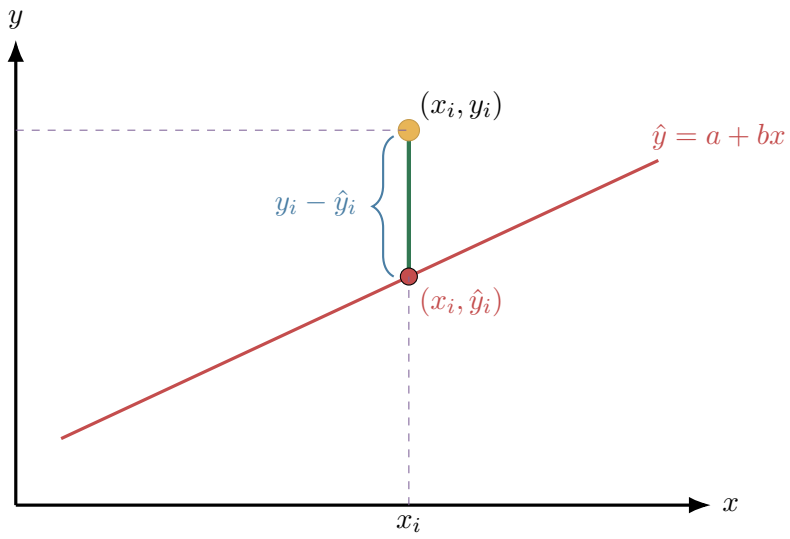
Clearly the blue line fits the data better than the purple line. But, mathematically speaking, why is this the case? In other words, what is it about the blue line that is better than the purple? One answer to this question can be found by looking at the vertical distances between the data and the regression line. As a reminder: each data point is of the form (x_i, y_i). Therefore, to describe the vertical distance from each of these points to the line, we can input the value x_i into the equation for the regression line $\hat{y}$ to find the output of the line. The value on the regression line above x_i is given by the number:

$$\hat{y}_i = a + bx_i.$$

Taken together, this defines a point $(x_i, \hat{y}_i)$ sitting on the line that is directly above x_i. Ideally, this point should be close to the observed data point (x_i, y_i). As we can see, in the case of the purple line, these vertical distances to the real data are much larger than those for the blue line:



It is in this sense that we intuitively think of the blue line as a better fit: the total of all of the lengths of these vertical lines is smaller than that of the purple line. Generally, these signed differences are called *residuals*, and they are simply the value of the difference $y_i - \hat{y}_i$:



Notice that values y_i lying above or below the line will cause $y_i - \hat{y}_i$ to have either a positive or negative sign, respectively. Therefore, if we were to naively add up all of these residuals to try and create some kind of "total error", then there is a risk that they may cancel each other out,

jeopardizing our measure of "good fit". To repair this issue, we first square all of the quantities $y_i - \hat{y}_i$ before adding them up.[1] As such, the correct measure of "total error" is actually the total *squared* error:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\big(y_i - (a + bx_i)\big)^2.$$

Different values of a and b produce different lines, and therefore different values of this sum. In the example above, the purple line's total squared error happens to be much larger than the blue line's. With this in mind, we can define the "line of best fit" as the line which minimizes this total squared error.

The least-squares criterion

The least-squares regression line $\hat{y} = a + bx$ is chosen so that

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

is as small as possible among all straight lines.

2.3 The formulas for the regression line

The exact derivation of the values of a and b requires the use of calculus. Since this is just an introductory course, we will not spend time deriving the regression line here. [2] Instead, we will merely accept that the resulting formulas for a and b are:

$$b = \frac{\text{cov}(x, y)}{s_x^2} \qquad \text{and} \qquad a = \bar{y} - b\bar{x}.$$

These two formulas allow us to write the regression line $\hat{y} = a + bx$.

Computing the regression line

Given paired data (x_i, y_i), calculate

$$b = \frac{\text{cov}(x, y)}{s_x^2} \qquad \text{and} \qquad a = \bar{y} - b\bar{x}.$$

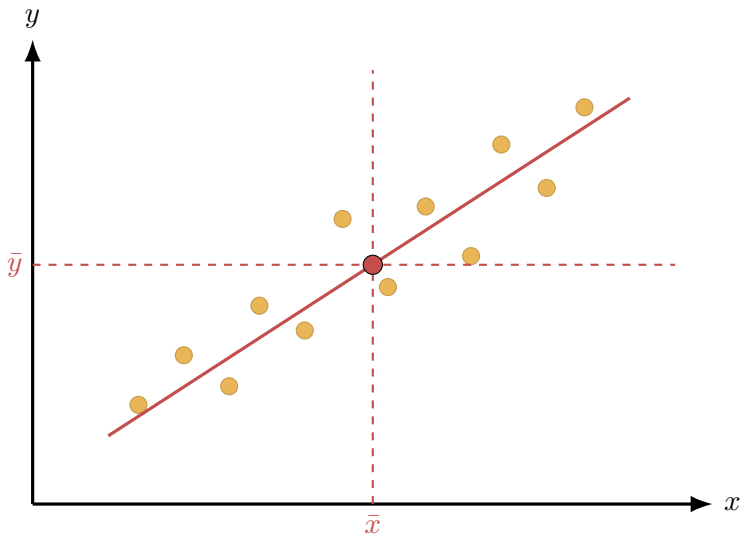
Then write the least-squares regression line as

$$\hat{y} = a + bx.$$

As a small corollary, we notice that the point $(\bar{x}, \bar{y})$ always lies on the least-squares regression line:

[1] This is the same reason that we square the quantities $y_i - \bar{y}$ in the formula for the sample variance s_y^2.

[2] For the curious: we can take the expression $\sum_{i=1}^{n}\big(y_i - (a + bx_i)\big)^2$ and differentiate it with respect to a and b. To find the minimum, we set these derivatives equal to zero and solve the pair of resulting linear equations.



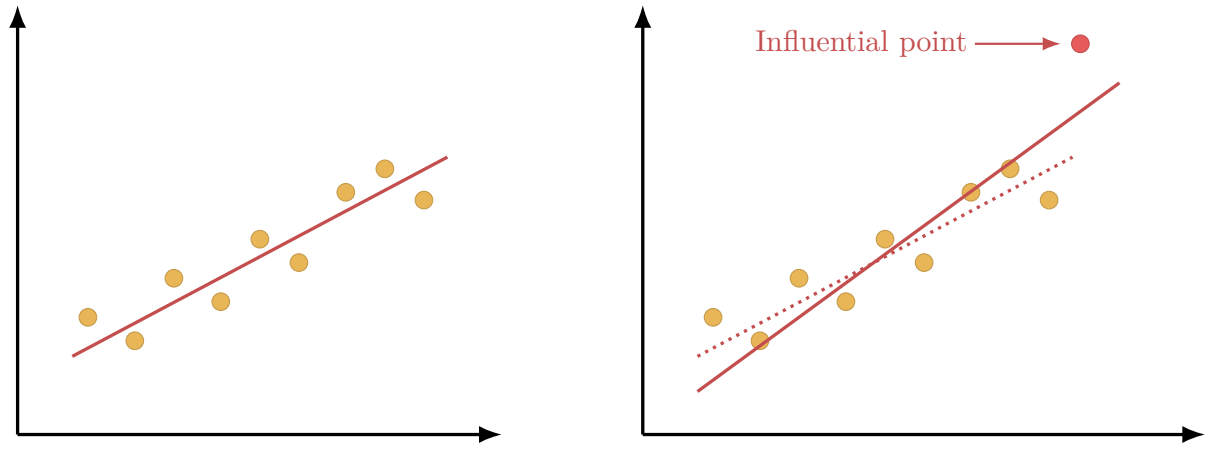
To check this, we may substitute $x = \bar{x}$ into the equation and see:

$$\hat{y} = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}.$$

Note that it doesn't matter which data we start with – it is simply a consequence of the construction of the regression line.

2.4 Influential points

A point can have a large effect on the regression line when it lies far from the main cloud of data, especially in the horizontal direction. Such an observation is called an *influential point*.



An influential point is not necessarily an issue. However, because it can radically change a regression line, the inclusion of influential points may affect any predictions that we may want to make using our regression line as a model. With this in mind, we should always be cautious when dealing with points of this type.

2.5 Example: computing the regression line

We now calculate the least-squares regression line for the commuting data of Section 1. From Lecture 18, we already know

$$\bar{x} \approx 25.333, \qquad \bar{y} \approx 34.333, \qquad s_x \approx 15.011, \qquad \operatorname{cov}(x, y) \approx 301.606.$$

We first calculate the gradient b:

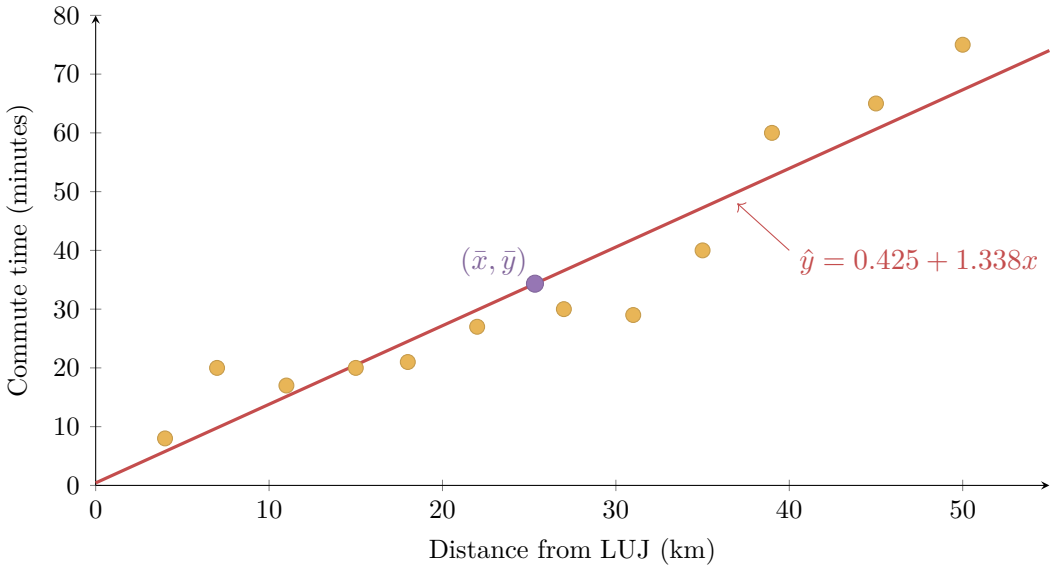
$$b = \frac{\operatorname{cov}(x, y)}{s_x^2} = \frac{301.606}{(15.011)^2} \approx 1.33849.$$

We may now use this value to calculate the y-intercept a:

$$a = \bar{y} - b\bar{x} = 34.333 - (1.33849)(25.333) \approx 0.425.$$

Putting these together, the regression line becomes:

$$\hat{y} = 0.425 + 1.338x.$$



Interpretation of the gradient

The gradient is approximately 1.338. According to the fitted linear model, each additional kilometre of distance from LUJ is associated with an increase of approximately 1.34 minutes in predicted commute time.

The intercept is approximately 0.425 minutes. In this example, the intercept is a theoretical requirement that is mainly needed to position the line correctly, and it should not automatically be given a real-world interpretation.

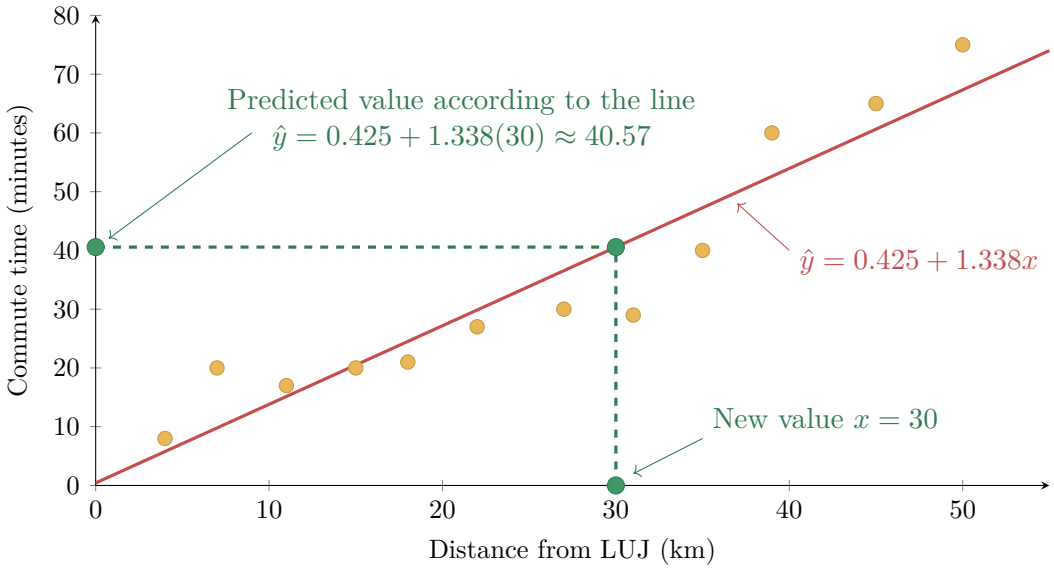
3 Using the Regression Line for Predictions

3.1 Making a prediction

Aside from visual summaries, one key benefit of a regression line is that it can act as a model that allows us to make predictions. Using the example of commute times and distance travelled, let's suppose that a student moves and their new apartment is 30 km away from campus. In this case, we can use the regression line to predict the approximate commute time by plugging the value $x = 30$ into the equation:

$$\hat{y} = 0.425 + 1.338(30) \approx 40.57.$$

Thus, the model predicts a commute time of approximately 41 minutes. Visually, our prediction corresponds to the following:

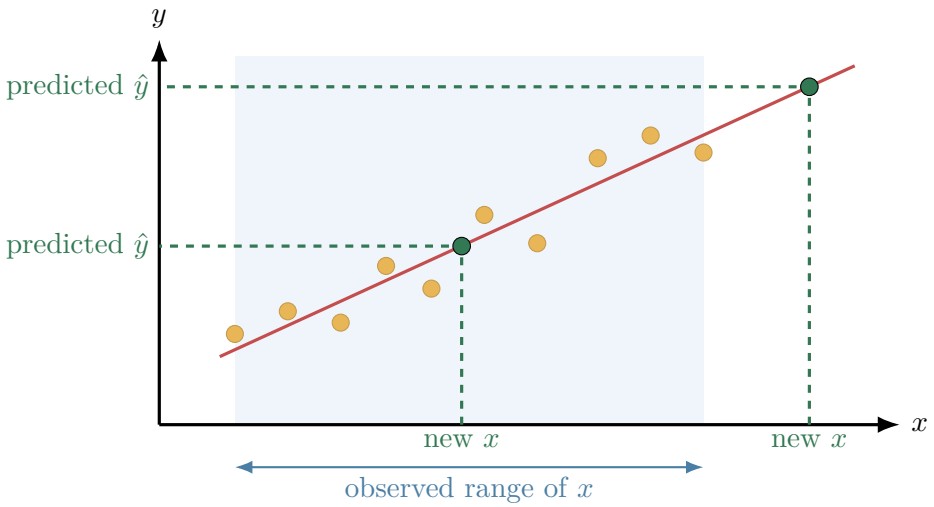


3.2 Interpolation and extrapolation

There are two fundamentally different types of prediction, known as *interpolation* and *extrapolation*.

Interpolation and Extrapolation

Interpolation makes predictions based on new values that lie within the observed range of x, whereas **extrapolation** makes predictions based on new values of x that are outside of the range of the observed values.



Typically, interpolation is usually the safer use of a regression line, because the model is being applied in a region where the data was actually observed. In contrast, extrapolation carries the risk that we use a model outside an appropriate regime. There may be other variables that affect our data at extreme values of x, and this is not accounted for in the regression line alone.

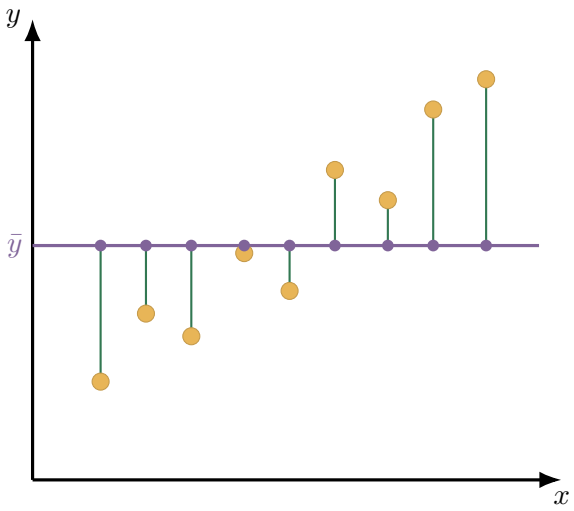
Warning about extrapolation

A linear pattern observed over one range of x may not continue outside that range. Extrapolated predictions can therefore be unreliable, even when the original correlation is strong.

3.3 The coefficient of determination

The regression line summarizes the scatter diagram and allows us to predict values of y based on new values of x. We will now introduce a number that quantifies how good this fit is. In order to do so, we first need to introduce a baseline predictor.

Suppose that we ignored the data in x completely. If we were to do this, the best general prediction for any value of y would be to use the sample mean $\bar{y}$. Of course, this wouldn't be correct most of the time, but *on average* it would be the best thing that we can do. Using the mean $\bar{y}$ to predict every value of y would be like picking a horizontal line $y = \bar{y}$:

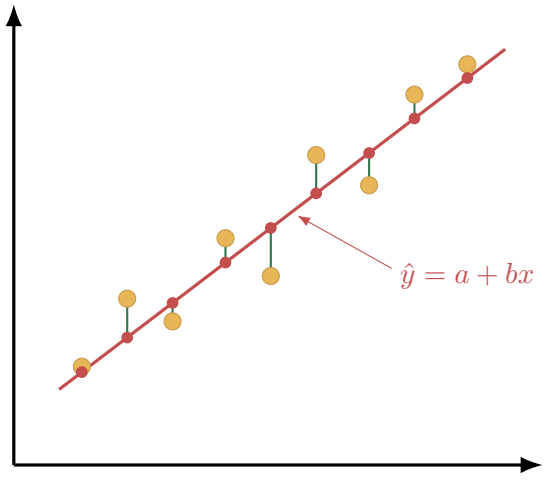


We can measure how good this prediction method is by looking at the total amount of error created between the line and the observed data points. Again, making sure to square the values first so as to avoid accidental cancellation, the total squared error coming from our naive prediction will be:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

This quantity is known as the *total squared error*, and it will serve as our baseline.

We are now going to measure how good the regression line is at predicting by comparing its error with the naive prediction. Suppose now that instead of simply guessing $\bar{y}$ for every value, we use the paired input of x_i together with the regression line. This will again give some errors:



The total squared error obtained from using the regression line as a predictor is given by the quantity

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

By construction, this is the quantity that is minimised by the least-squares criterion. Therefore, it will always be the case that

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 \geq \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

If we wanted to compare the relative size of these two numbers, we can always take their ratio:

$$\frac{\text{total squared error using regression line}}{\text{total squared error using the sample mean}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

If the regression line is a good predictor, then it should produce a small error compared to the error from using the sample mean. Therefore, we would expect the above ratio to be *small*. Conversely, if the regression line is bad as a predictor, then it would generate as much error as simply using the mean, implying that the above ratio would be close to 1. If we wanted the number to be large when the predictor is good and small when the predictor is bad, then we could subtract the number from 1:

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

This is known as the *coefficient of determination*.

Coefficient of determination

The **coefficient of determination** is

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

It measures one minus the ratio of the error produced by the regression line to the error produced by the naive guess using the sample mean. When it is defined, the number is between 0 and 1, with larger numbers implying that the regression model is a better fit.

3.4 Other forms of the coefficient of determination

We can rearrange the coefficient of determination in order to find other ways to talk about this number. One way to do that is to insert a copy of $\hat{y}_i - \hat{y}_i$ into the expression $y_i - \bar{y}$ to obtain:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

Conveniently, the regression line has the useful property that:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\Big((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\Big)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

Usually we would expect there to be cross-terms of the form $2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$, but they happen to vanish when summed over, as a consequence of the calculus used to construct the regression line. With some basic algebra, we now see:

$$\begin{aligned} 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \end{aligned}$$

We can divide both the numerator and denominator by $n-1$ to obtain:

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

On the bottom of this expression is the variance for y, i.e. the spread of the paired data in the vertical direction. On the top of the fraction we have an expression which calculates a sort-of average squared distance from the predicted values to $\bar{y}$. This is again a type of variance, but this time we are calculating the spread of the predicted values $\hat{y}_i$ around $\bar{y}$. We may therefore think of the coefficient of determination in these terms.

Alternate Expression of the Coefficient of determination

The **coefficient of determination** can be equivalently expressed as

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

Therefore, we may understand the coefficient of determination as measuring how much of the variance in y is explained by the regression line $\hat{y}$.

3.5 The coefficient of determination as r^2

Consider again the alternate form of the coefficient of determination. For convenience, we are going to insert copies of $\frac{1}{n-1}$ into the numerator and denominator:

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

Therefore, the denominator becomes s_y^2, the sample variance of y. Since the regression line is given by $\hat{y}_i = a + bx_i$ where $a = \bar{y} - b\bar{x}$, we see that:

$$\hat{y}_i - \bar{y} = a + bx_i - \bar{y} = (\bar{y} - b\bar{x}) + bx_i - \bar{y} = b(x_i - \bar{x}).$$

Therefore, the numerator of the coefficient of determination becomes:

$$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(b(x_i - \bar{x})\right)^2 = b^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = b^2 s_x^2.$$

Putting this together, we have:

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{b^2 s_x^2}{s_y^2}.$$

Using the definition of b, we see that

$$b = \frac{\mathrm{cov}(x,y)}{s_x^2} = \frac{\mathrm{cov}(x,y)}{s_x^2} \cdot \frac{s_y^2}{s_y^2} = \frac{\mathrm{cov}(x,y)}{s_x s_y} \cdot \frac{s_y^2}{s_x s_y} = r \cdot \frac{s_y}{s_x}.$$

Therefore, the coefficient of determination simplifies heavily to:

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2} = b^2 s_x^2 \div s_y^2 = \frac{\left(r\frac{s_y}{s_x}\right)^2 s_x^2}{s_y^2} = r^2\frac{s_y^2}{s_x^2}s_x^2 \div s_y^2 = r^2 s_y^2 \div s_y^2 = r^2.$$

Indeed, for simple linear regression with one explanatory variable and an intercept, the coefficient of determination is exactly the square of the sample correlation coefficient:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}.$$

This has several immediate consequences:

- the sign of r disappears when it is squared;
- positive and negative correlations of equal strength produce the same r^2;
- a strong linear correlation produces a large coefficient of determination;
- a weak linear correlation produces a small coefficient of determination.

3.6 Example: coefficient of determination for the commute data

For our commuting data, we calculated previously that $r \approx 0.9454$. Therefore,

$$r^2 \approx (0.9454)^2 \approx 0.894.$$

Thus, approximately 89.4% of the observed variation in commute times is accounted for by the linear regression on distance from LUJ in this sample. This unexplained variation may reflect other variables influencing the commute time, such as method of transport, traffic, transfers, walking time, measurement noise, and so on.