Linear Regression Review

Author: Alan Rosenthal

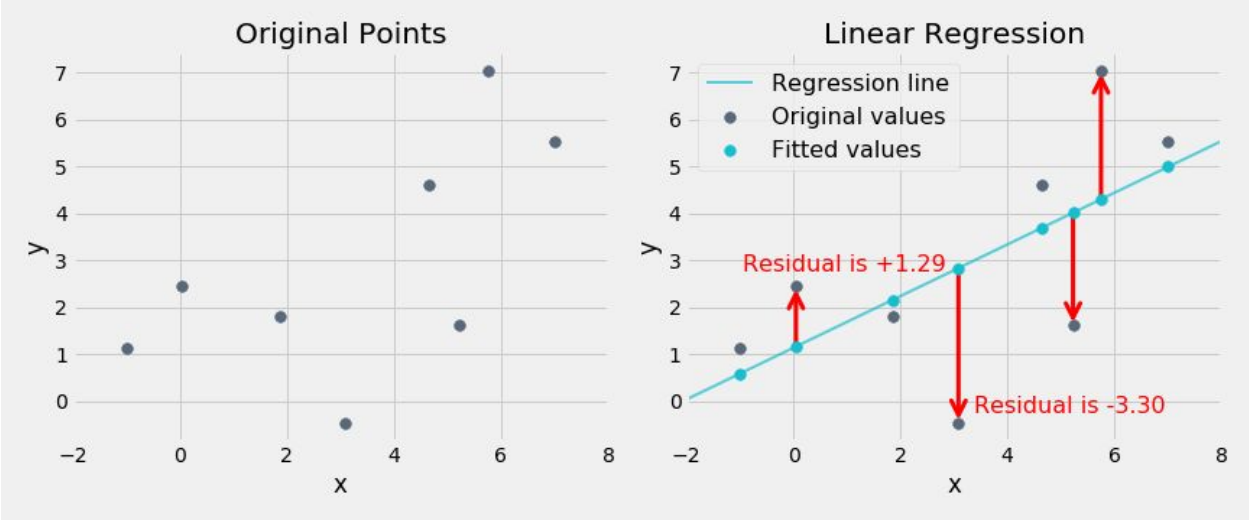
Table of Contents:

Setup and terms

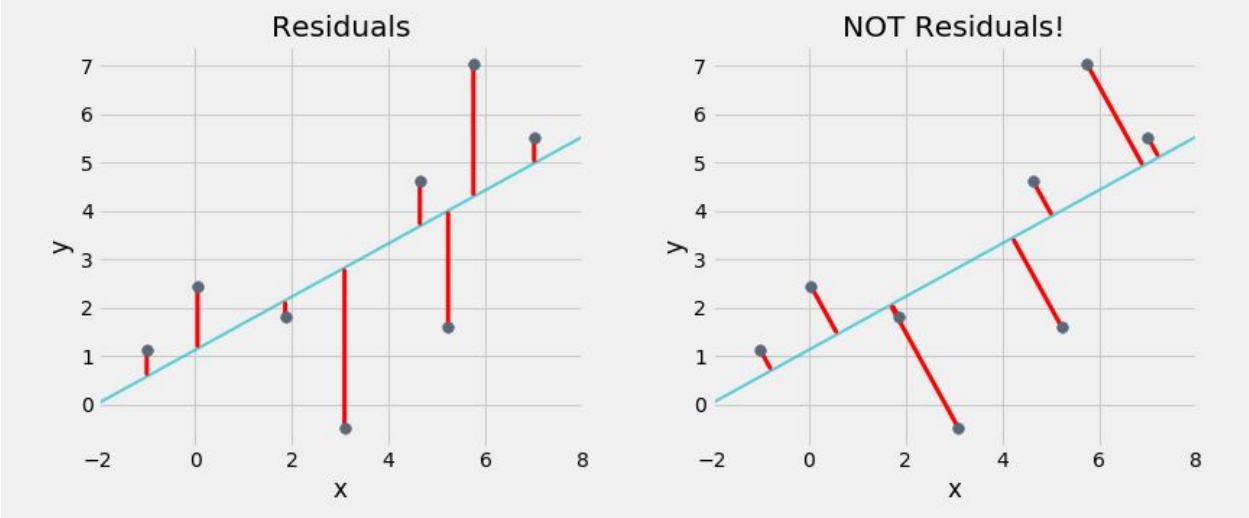
When we have two numerical variables x and y that are related to each other, we can use x to predict y. **Linear regression** is when we specifically use a line to predict. (textbook)

- The **fitted values** are the y-values of the regression line at the original x-values. (textbook)
- The **residuals** are the differences between the actual y-values and the fitted values of the points. They are the "errors" of our predictions. (textbook)

Here, the right plot shows what these terms correspond to when we do linear regression on the points in the left plot. Some of the residuals are shown with red arrows.

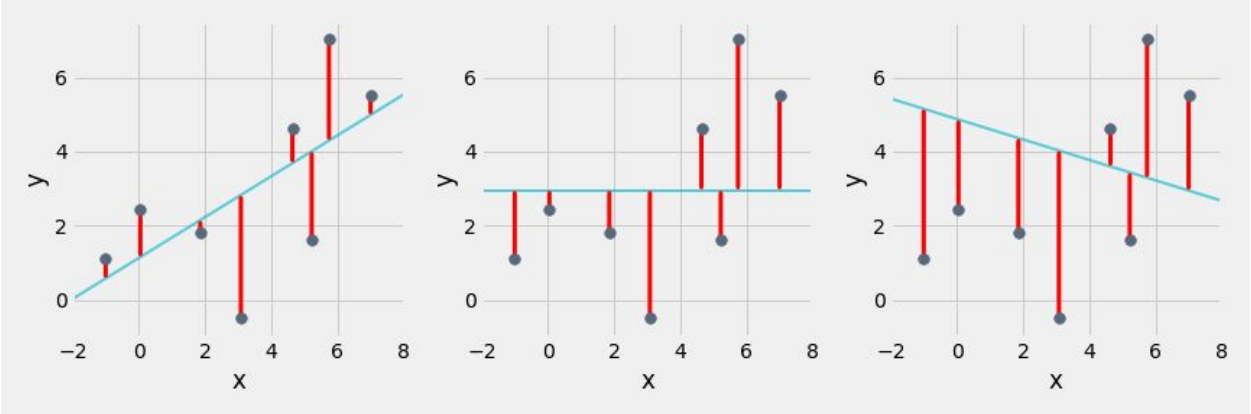


Keep in mind: the residuals are vertical differences between the points and the regression line, NOT perpendicular differences.



The defining property of the regression line

There are many lines we could use to predict, some better than others:



Since the residuals are the errors, our goal is to make them as small as possible. More precisely, the regression line minimizes the **RMSE (Root Mean Square Error)**, which is the square root of the average squared-error. This best-fit line is **unique** - every other line will have a bigger RMSE.

Why square the errors? One reason is that the errors can be positive or negative, so directly adding them wouldn't make sense because negative values would cancel positive ones - we want all errors to increase the total. Squaring the errors ensures that they are positive. (textbook)

$$\text{RMSE} = \sqrt{\frac{1}{n}\Big((\text{residual } 1)^2 + (\text{residual } 2)^2 + \dots\Big)}$$

$$\text{residual } i = y_i - \underbrace{(\textbf{slope} \cdot x_i + \textbf{intercept})}_{\text{fitted value } i}$$

We can get the regression line equation from this principle directly: we just need a way to find the magical slope and intercept that minimize the RMSE formula. This could be done in multiple ways:

1. Use the `minimize` function from the datascience module (demonstration in textbook).
2. Use the formulas in the next section for the equation of the regression line.

Both of these give the **same** line. The first way may have minute differences due to the inexact nature of the `minimize` function.

Regression line equation

First, recall the definition of the correlation coefficient:

$$r = \text{mean}(x_{su} \times y_{su})$$

The regression line equation is:

$$y = \textbf{slope} \cdot x + \textbf{intercept}$$

The **slope** and **intercept** are:

$$\textbf{slope} = r \cdot \frac{\text{SD}(y)}{\text{SD}(x)}$$
$$\textbf{intercept} = \text{mean}(y) - \textbf{slope} \cdot \text{mean}(x)$$

Where do these formulas come from?

Derivation 1: Equations

The regression line has a much more elegant form when x and y are in standard units:

$$y_{su} = r \times x_{su}$$

If the x and/or y variables are scaled by constants and shifted, the regression line gets scaled and shifted accordingly. Remember that converting to standard units is a scaling-and-shifting procedure. Therefore, if we rearrange the equation above to be in terms of x and y, instead of x and y in standard units, we will get the regression line. Here it is step-by-step:

$$y_{su} = r \times x_{su} \tag{1}$$

$$\frac{y - \text{mean}(y)}{\text{SD}(y)} = r \cdot \frac{x - \text{mean}(x)}{\text{SD}(x)} \tag{2}$$

$$y - \text{mean}(y) = \underbrace{\left(r \cdot \frac{\text{SD}(y)}{\text{SD}(x)}\right)}_{\textbf{slope}} \cdot x - \underbrace{\left(r \cdot \frac{\text{SD}(y)}{\text{SD}(x)}\right)}_{\textbf{slope}} \cdot \text{mean}(x) \tag{3}$$

$$y = \textbf{slope} \cdot x + \underbrace{\text{mean}(y) - \textbf{slope} \cdot \text{mean}(x)}_{\textbf{intercept}} \tag{4}$$

- (1) → (2) Definition of standard units

- (2) → (3) Multiply both sides by SD(y)
- (3) → (4) Add mean(y) to both sides

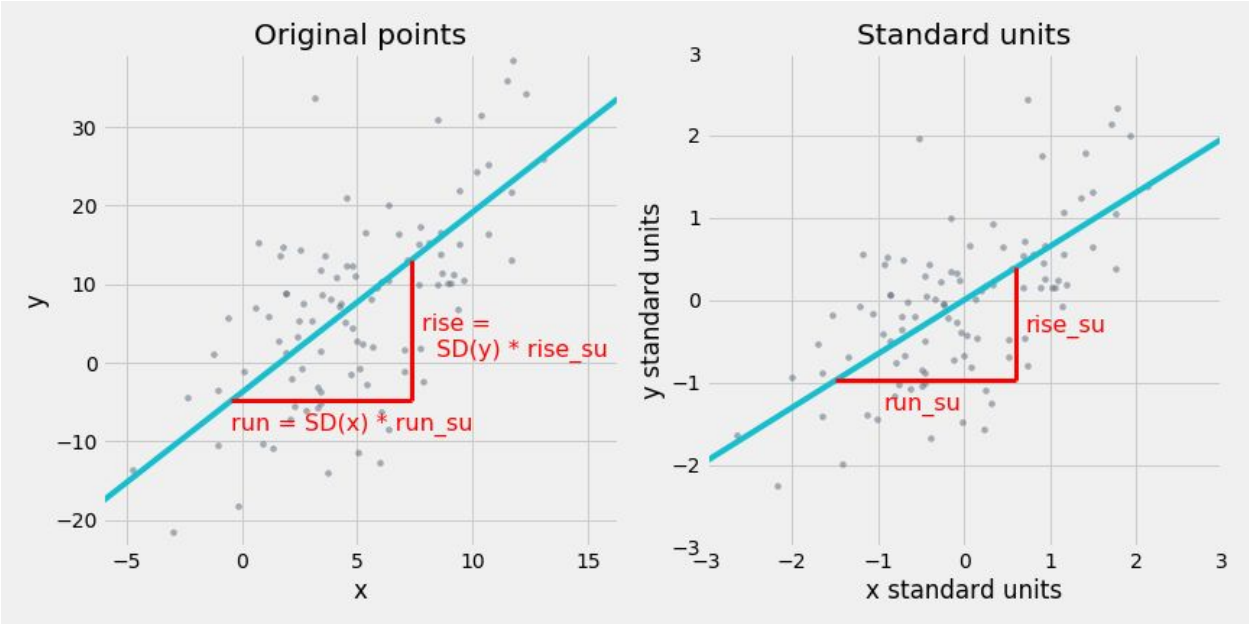
Derivation 2: Visual

We can also derive the formulas with a more “visual” approach. As before, we start with:

$$y_{su} = r \times x_{su}$$

Now, remember that the slope of a line is rise / run.

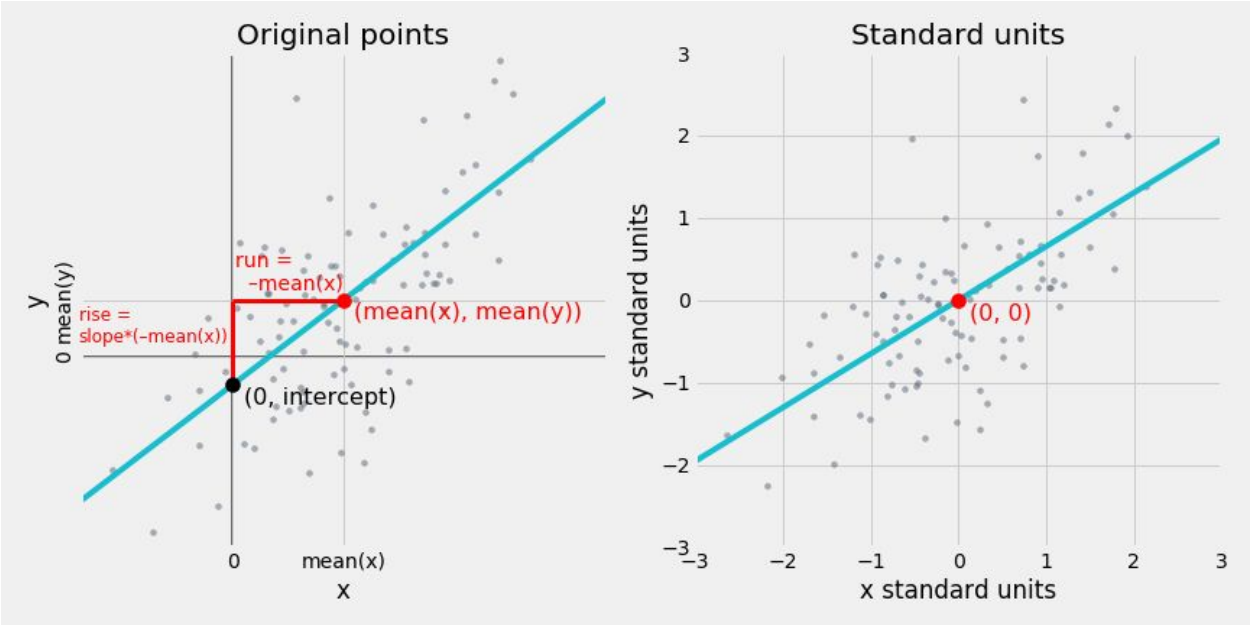
Slope:



The slope for the points in standard units is r = rise_su / run_su. We want the slope, or rise / run, of the line for the original points. We get standard units by dividing by the SD, so SD(y) units of rise correspond to one unit of rise_su, and SD(x) units of run correspond to one unit of run_su. Then:

$$\textbf{slope} = \frac{\text{rise}}{\text{run}} = \frac{\text{SD}(y)}{\text{SD}(x)} \cdot \underbrace{\frac{\text{rise}_{\text{su}}}{\text{run}_{\text{su}}}}_{r} = r \cdot \frac{\text{SD}(y)}{\text{SD}(x)}$$

Intercept:



The point (0, 0) in standard units corresponds to (mean(x), mean(y)) in original units. The intercept is the value of the regression line at x = 0. Starting at x = mean(x), we walk backwards until x = 0, for a run value of -mean(x). The corresponding rise is slope * run = slope * (-mean(x)). The y-value we end up at is the starting y_value, mean(y), plus the run:

$$\begin{aligned} \textbf{intercept} &= \text{mean}(y) + \textbf{slope} \cdot (-\text{mean}(x)) \\ &= \text{mean}(y) - \textbf{slope} \cdot \text{mean}(x) \end{aligned}$$

Regression to the mean

Roughly speaking, regression to the mean says that individuals with an extraordinary value of x tend, on average, to have a value of y that is relatively less extraordinary. (textbook)

An example would be heights in families. Super-tall parents generally have tall children, but on average not *quite as exceptionally tall* as their parents. For parents that are +3 SDs above the population mean, perhaps their kids are around +2 SDs on average.

What is the mathematical reason? Recall the equation of the regression line when x and y are in standard units:

$$y_{su} = r \times x_{su}$$

When we talk about how many SDs above or below the mean an individual is, we are literally talking about their value in standard units. Since the absolute value of the correlation coefficient r is at most 1, multiplying x_su by r shrinks it toward 0, which is the mean in standard units.

Important note: the individuals we do prediction on must come from the same distribution as the data used to fit the regression line. In general, this must be true whenever we do prediction: a model's predictions are only valid on the same underlying population that the model was fitted to.

Properties of residuals

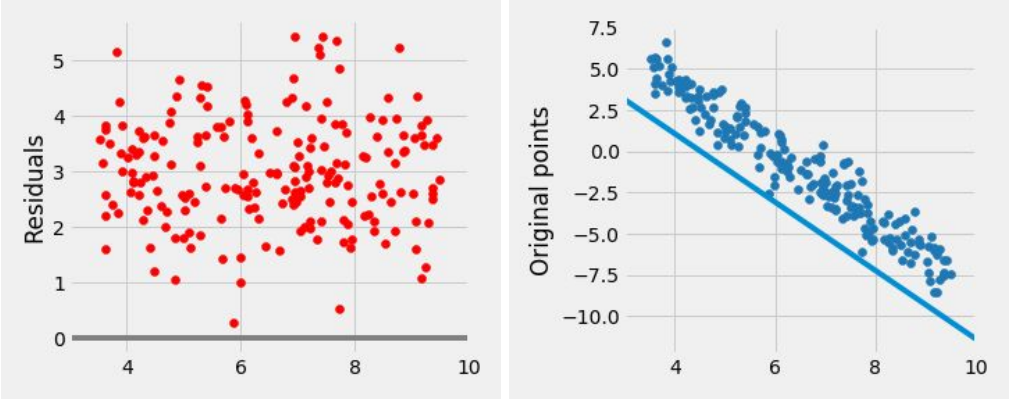
Recall that the residuals are defined as the errors that specifically the best fit line makes. They have some important properties. For more detail and derivations, see the residuals guide @3364. The highlights are repeated here:

1. Residuals have mean zero.
    - This tells us that RMSE = SD(residuals).
2. Residuals are uncorrelated with x.
3. Residuals are uncorrelated with the fitted values.
4. $\text{SD}(\text{residuals}) = \sqrt{1-r^2} \cdot \text{SD}(y)$
5. $\text{SD}(\text{fitted values}) = |r| \cdot \text{SD}(y)$

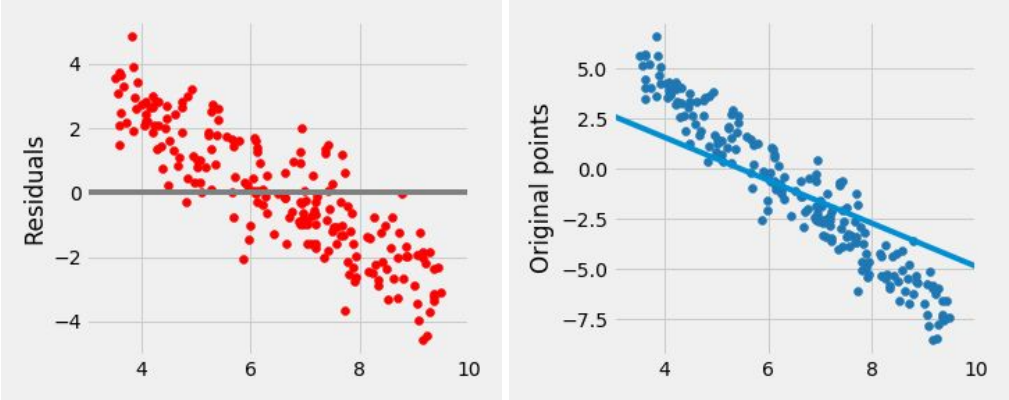
These properties are **always** true, regardless of what the data looks like.

We can understand properties 1 and 2 like so: if the property were not true, then that would mean our regression line wasn't as good as it should be.

1. If the residuals were to have nonzero mean, that would mean the regression line is drawn too low or high.



2. If the residuals had nonzero correlation with x, then the regression line has the wrong slope.



Conditions for using linear regression

Technically, we can always do linear regression on any two numerical variables - it blindly fits a line to the data. But that's not always a good idea. It works only when the data is "linear plus noise," where the noise is "nice" in certain ways.

Specifically, there should be an underlying linear trend in the data. If it helps, you can imagine that the data is generated in two steps:

1. "Underlying y-values" are set according to this linear trend.
2. Things aren't perfect, so our actual y-values are the underlying y-values with some noise added to them.

We only get to see the noisy y-values, while linear regression tries to predict the underlying y-values.

Here are conditions for the data to be well suited to linear regression:

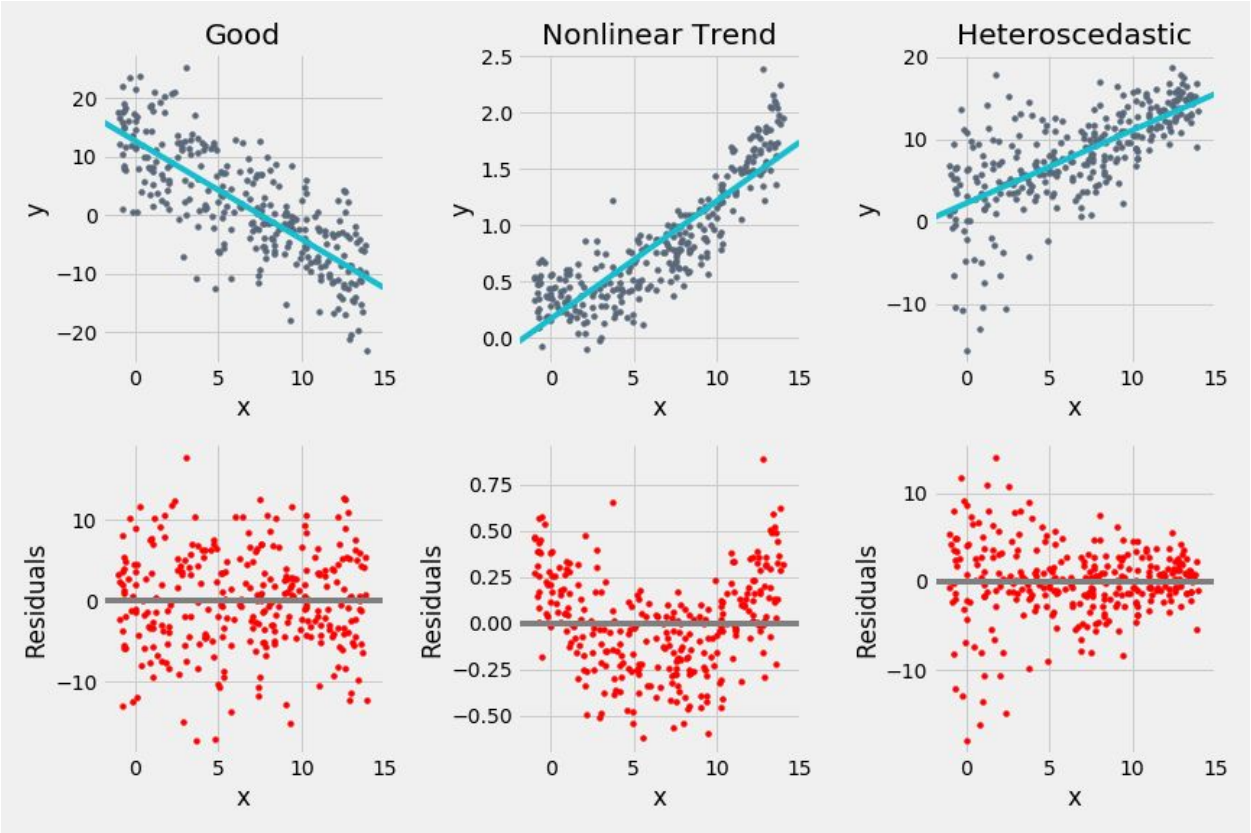
- The data follows a linear trend.
- The noise in the y-values is roughly the same across the range of x-values.

Usually, it's easier to check these conditions by looking at a **plot of the residuals**. Here are equivalent conditions, stated in terms of the residuals:

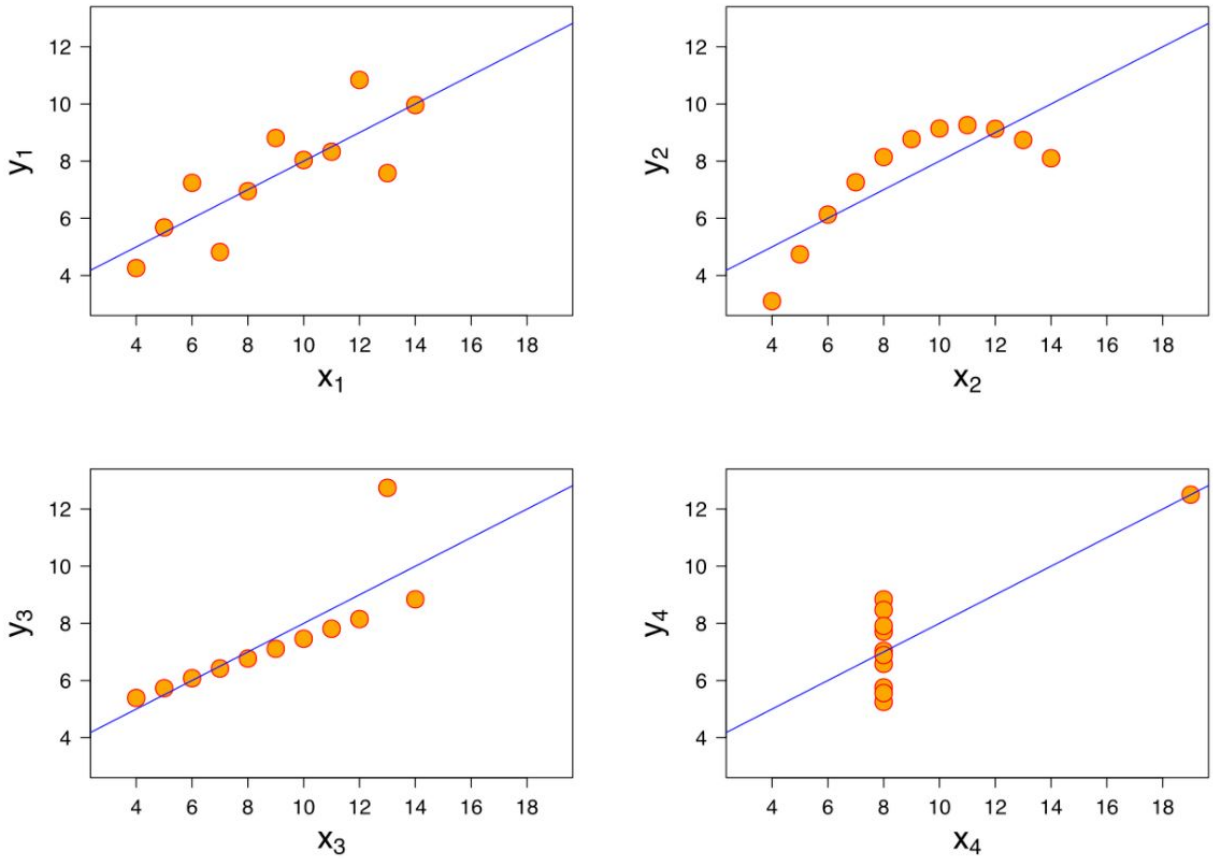
- Residuals show no pattern. They look about the same above and below the horizontal line at 0. (textbook)
- Residuals have about the same spread across the range of x-values.

- A fancy term for “same spread” is **homoscedasticity**. Uneven spread is **heteroscedasticity**. (textbook)

Here are some examples of residual plots demonstrating these characteristics. The left one is good while the other two are problematic.



Another classic example of the importance of *visually checking plots* is Anscombe’s quartet:



These four datasets, while clearly very different from one another, are identical (to several decimal places) in multiple surface-level characteristics.

- The x collections all have the same mean and SD. The y collections all do too.
- The x-y correlations are all the same.
- (implied by the other two) The linear regression lines are all the same.

Linear regression doesn't distinguish between these datasets, despite their glaring differences when plotted. The only dataset out of the four that is well suited for linear regression is the top-left.