Coefficient of determination and RMSE

At the end of the last section, we said that, in order to find the equation of the line-of-best-fit, we actually want to minimize

$$\sum (e_n)^2$$

where e_n is the residual for each data point.

Coefficient of determination

If e_n is the residual, the next thing we want to talk about is r^2, the coefficient of determination. But let's take a step back for a moment.

We've been talking about finding the regression line that best approximates the trend in the data. But we could have simply found the average of all the y-values in the set and then drawn a horizontal line through the data at that point instead.

For instance, given the data set

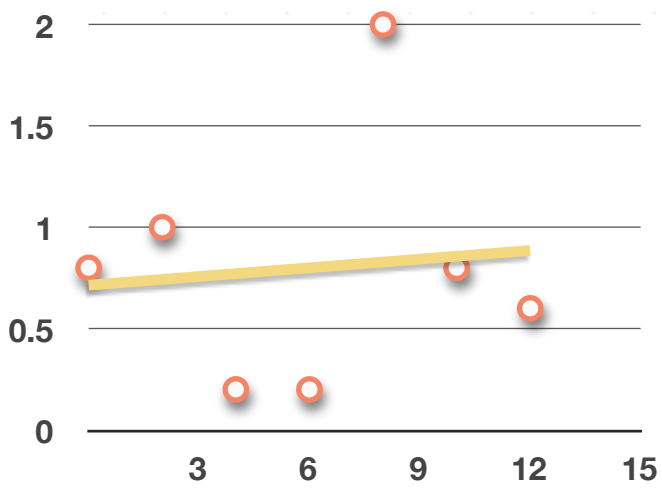
x	y
0	0.8
2	1.0
4	0.2
6	0.2
8	2.0
10	0.8
12	0.6

the mean of the y-values is

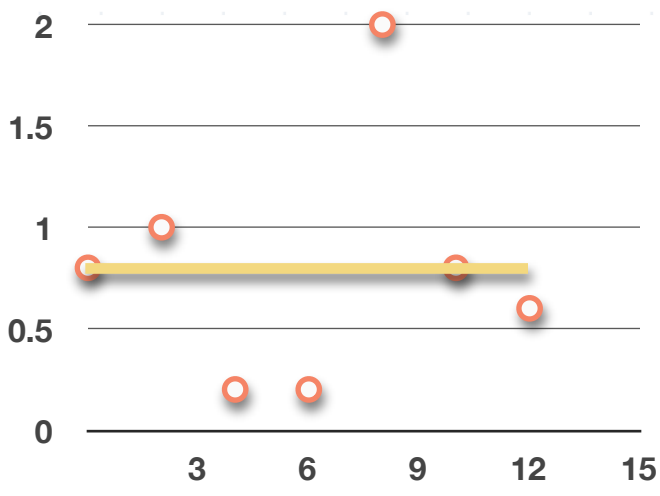
$$\bar{y} = 0.8$$

So, instead of sketching in the regression line, we could have taken a simpler path and sketched in the horizontal line at $\bar{y} = 0.8$.

Regression line:



Simple horizontal line:



The horizontal line doesn't fit the data as well as the regression line, but it's much faster to find. What we want to talk about now is how much error we eliminate by using the regression line, instead of the horizontal line.

If we eliminate a large amount of error, then we know that the regression line is a much better approximation than the simpler horizontal line. But if we only eliminate a small amount of error, then we know the horizontal line was actually a pretty good trend line for the data, and that the regression line doesn't do much better.

So how do we find the amount of error eliminated by using the regression line instead of the horizontal line? We do it with the **coefficient of determination**, r^2, which measures the percentage of error we eliminated by using least-squares regression instead of just $\bar{y}$.

We already know that the residual e of a data point is the distance between the point's actual y-value and its predicted $\hat{y}$-value from the regression line.

For the horizontal line through the data, if we were to take the residual of each data point and square it, and then add up the area inside all of those actual squares, we'd get a "sum of squares" that, in a way, measures the error of just drawing the horizontal line.

If instead we find the regression line that more accurately fits the data, and then we go through the same procedure of finding the residual for each data point compared to the regression line, and take the sum of squares again, what we'll find is that we significantly reduce the sum of squares, and therefore reduce the error. In other words, r^2 tells us how well the regression line approximates the data.

For this reason, the coefficient of determination is often written as a percent, where $100\,\%$ would describe a line that's a perfect fit to the data. The higher the value of r^2, the more data points the line passes through. If

r^2 is very small, it means the regression line doesn't pass through many of the data points.

The coefficient of determination is the square of the correlation coefficient, which is why we use r for correlation coefficient and r^2 for coefficient of determination.

Root-mean-square error

Root-mean-square error (RMSE), also called **root-mean-square deviation (RMSD)**, we can think of as the standard deviation of the residuals.

In the same way that we talked about the standard deviation of normally distributed data, and how many data points fall within one, two, or three standard deviations from the mean, we can think about RMSE as the standard deviation of the data away from the least-squares line.

Once we find the least-squares line, and we've sketched that through the data, we could draw parallel lines on either side of the least-squares line that represent standard deviations away from the regression line. If the standard deviation is very large, and these lines are far from the least-squares line, it tells us that the least-squares line doesn't fit the data very well. But if the standard deviation is very small, and these lines are close to the least-squares line, it tells us that the least-squares line does a very good job showing the trend in the data.

To find RMSE, we'll find the residual for each data point, then square it. We'll add up all of those square residuals, and then divide by n. Then we'll

take the square root of that result, and we'll get the standard deviation of the residuals.

If we continue on with the data set we've been working with,

x	y	"y-hat"	e
0	0.8	0.7143	0.0857
2	1.0	0.7429	0.2571
4	0.2	0.7715	-0.5715
6	0.2	0.8001	-0.6001
8	2.0	0.8287	1.1713
10	0.8	0.8573	-0.0573
12	0.6	0.8859	-0.2859

we want to start by squaring the residuals.

x	y	"y-hat"	e	e^2
0	0.8	0.7143	0.0857	0.0073
2	1.0	0.7429	0.2571	0.0661
4	0.2	0.7715	-0.5715	0.3266
6	0.2	0.8001	-0.6001	0.3601
8	2.0	0.8287	1.1713	1.3719
10	0.8	0.8573	-0.0573	0.0033
12	0.6	0.8859	-0.2859	0.0817

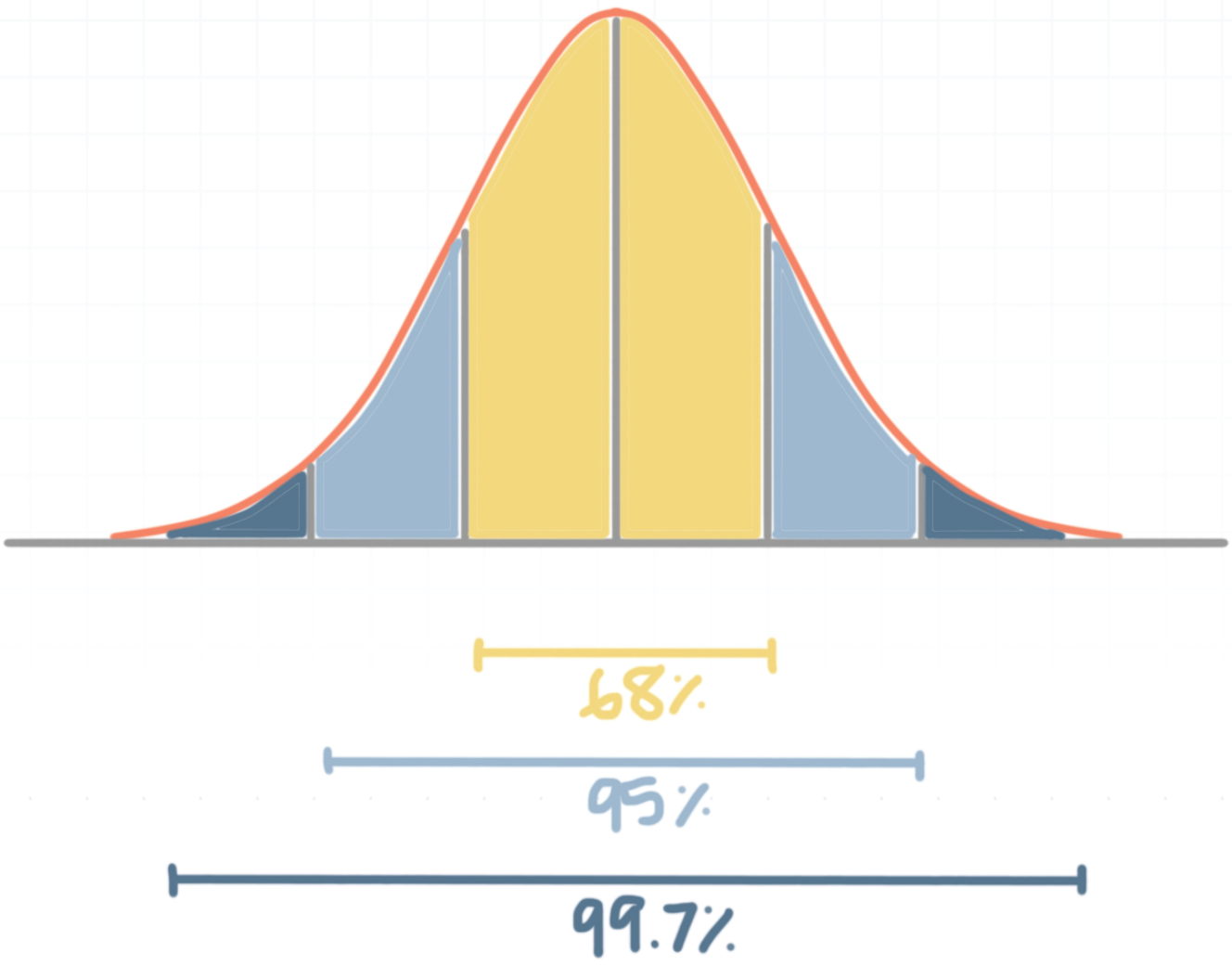
Then we sum the residuals, divide that sum by n, and take the square root of that result. Since we have $n = 7$ data points, we get

$$RMSE \approx \sqrt{\frac{0.0073 + 0.0661 + 0.3266 + 0.3601 + 1.3719 + 0.0033 + 0.0817}{7}}$$

$$RMSE \approx \sqrt{\frac{2.2170}{7}}$$

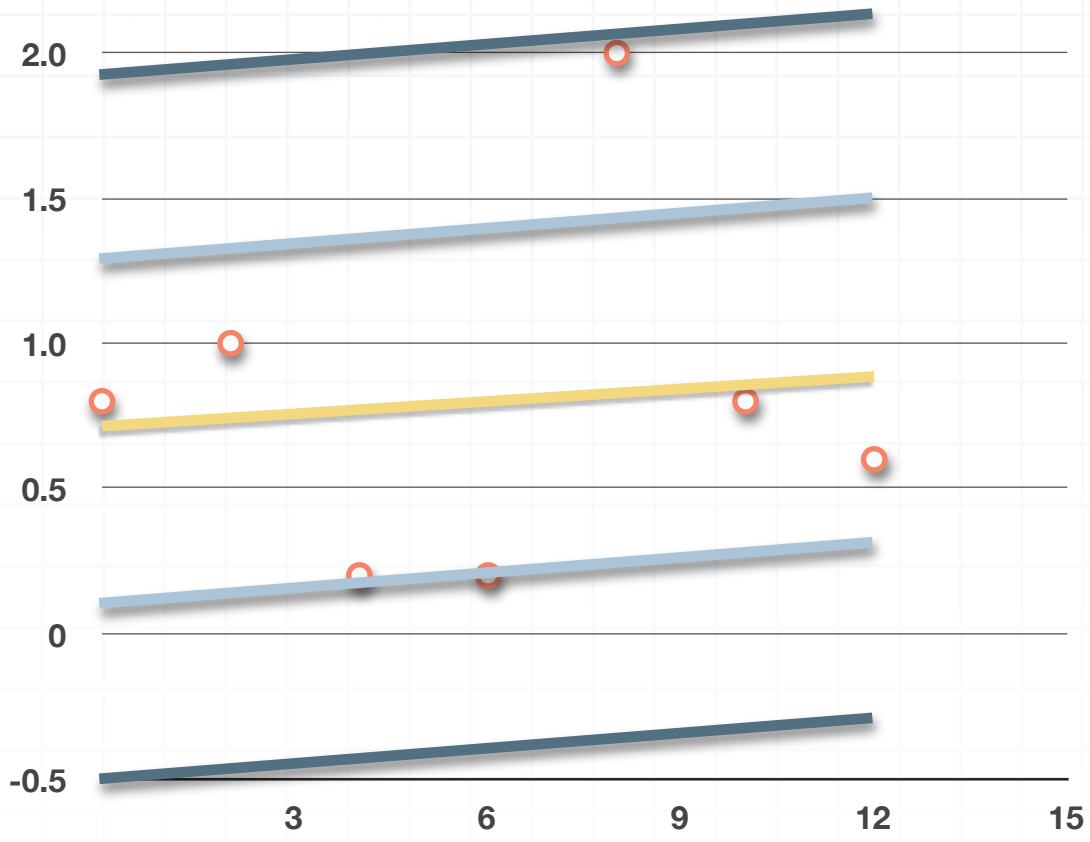
$$RMSE \approx 0.5628$$

This value is the standard deviation of the residuals, which means that, based on the normal curve from earlier in the course,



$68\,\%$ of the data points will fall within ± 0.5628 (one standard deviation) of the regression line, that $95\,\%$ of the data points will fall within $\pm 2(0.5628)$ (two standard deviations) of the regression line, and that $99\,\%$ of the data points will fall within $\pm 3(0.5628)$ (three standard deviations) of the regression line.

We could roughly sketch these standard deviation boundaries into the graph.



$68\,\%$ of the data points will fall inside the light blue lines, and $95\,\%$ of the data will fall inside the dark blue lines.

The larger the RMSE (standard deviation),

- the further apart these lines will be,
- the more scattered the data points are, and
- the weaker the correlation is in the data

The smaller the RMSE (standard deviation),

- the closer together these lines will be,
- the more tightly clustered the data points are, and
- the stronger the correlation is in the data