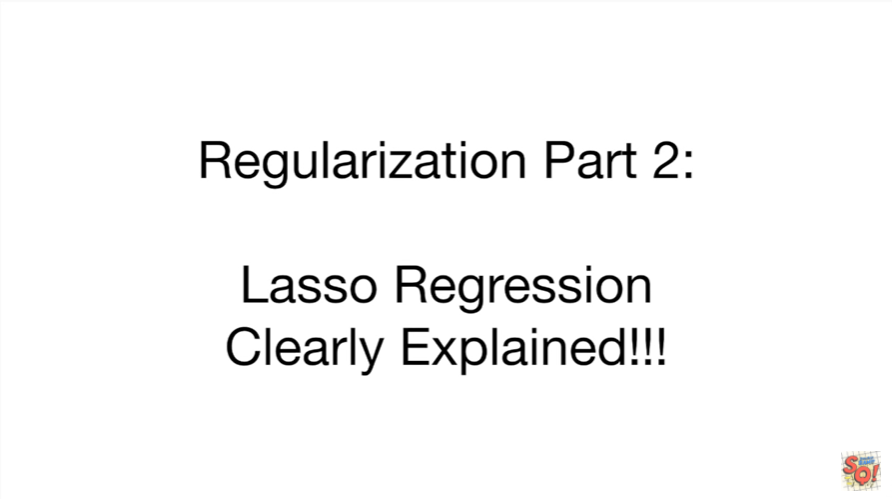
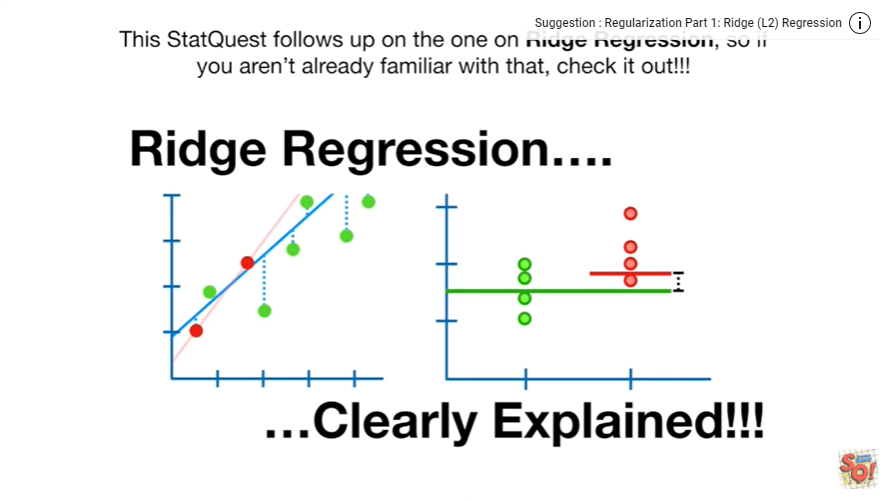
<https://www.youtube.com/watch?v=NGf0voTMlcs&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=20>



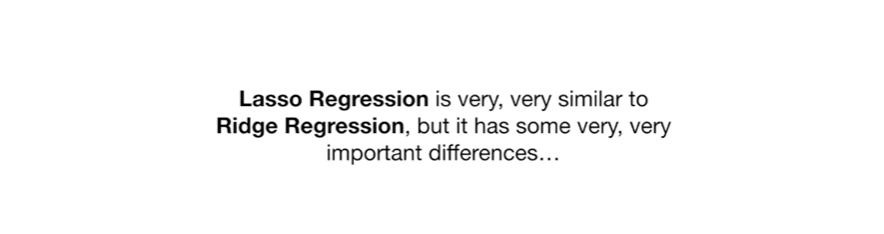
Today we're gonna do part two of our series on regularization.

We're gonna talk about lasso regression and it's going to be clearly explained.

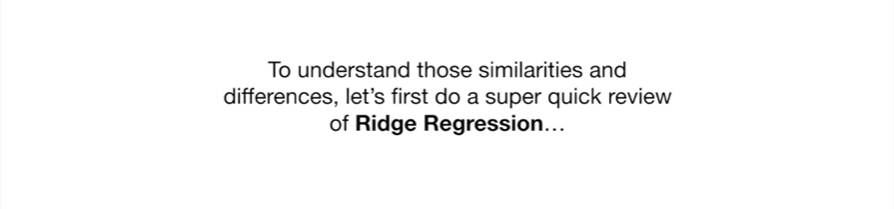


This stat quest follows up on the one on Ridge regression so if you aren't already familiar with that check it out.

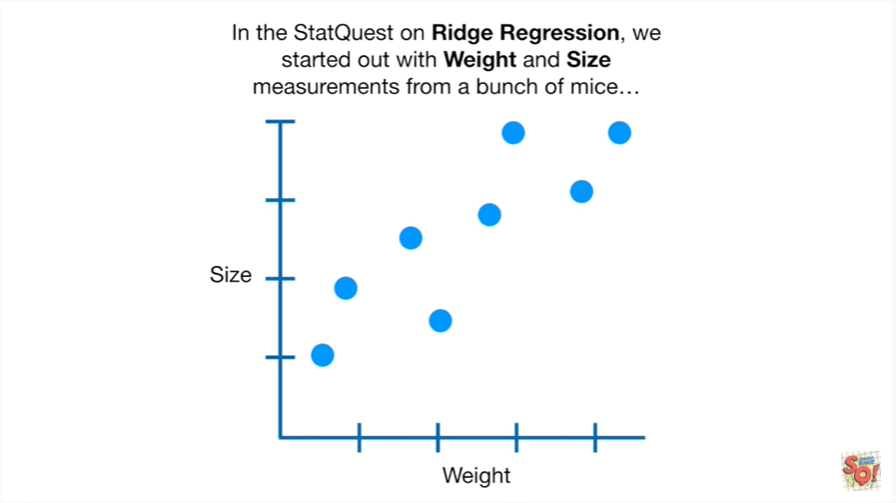
Even if you are familiar with Ridge regression you should seriously consider watching or at least skimming that stat quest because the examples in this video are based on the ones in that video last.



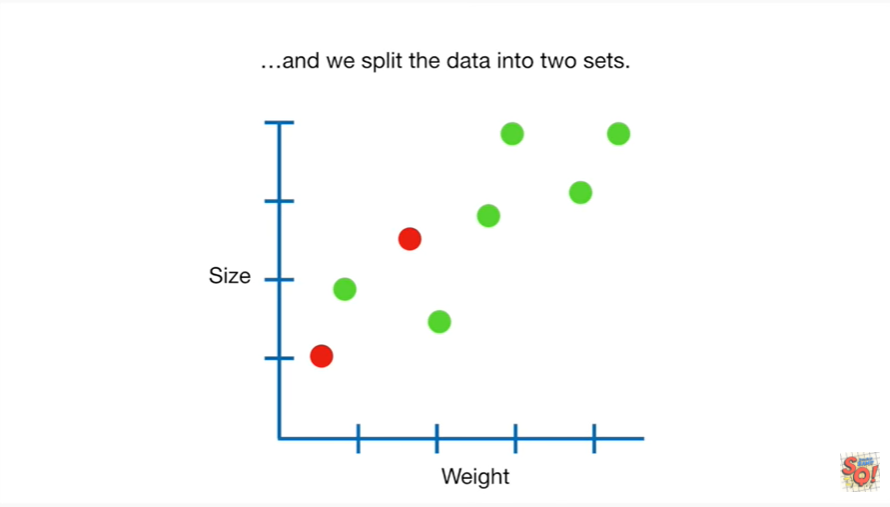
So regression is very very similar to Ridge regression but it has some very very important differences.



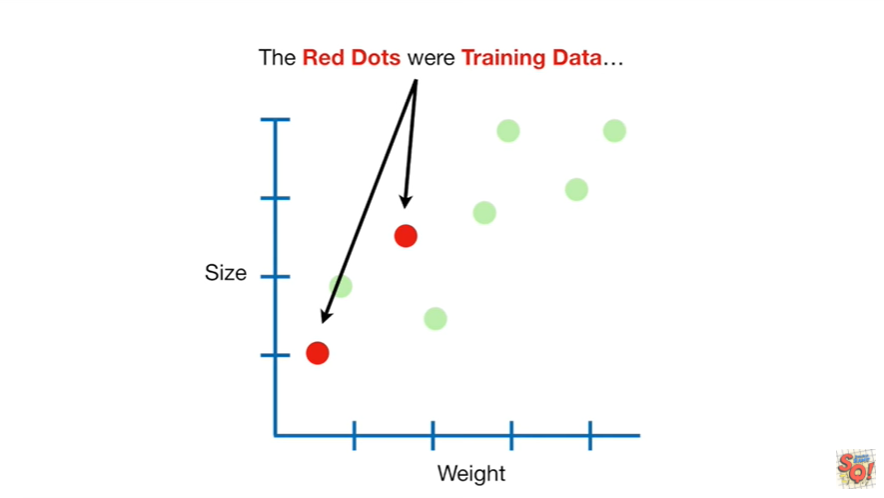
To understand those similarities and differences let's first do a super quick review of Ridge regression.



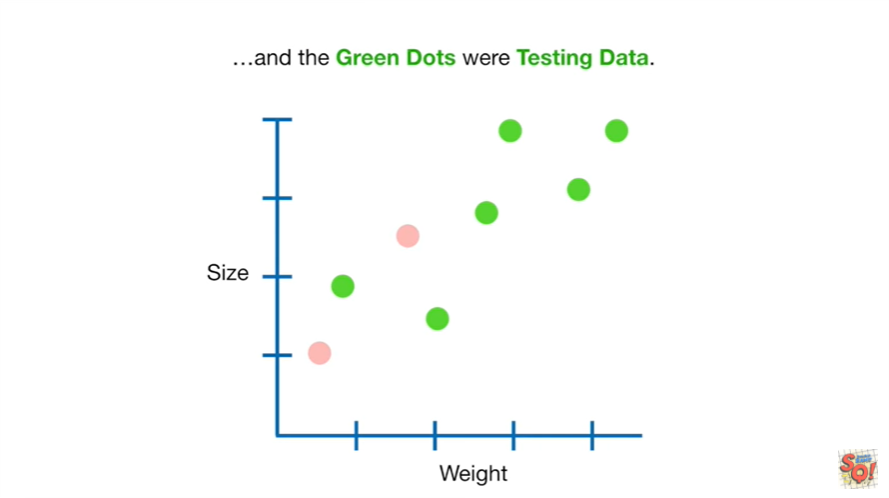
In the stat quest on Ridge regression we started out with weight and size measurements from a bunch of mice



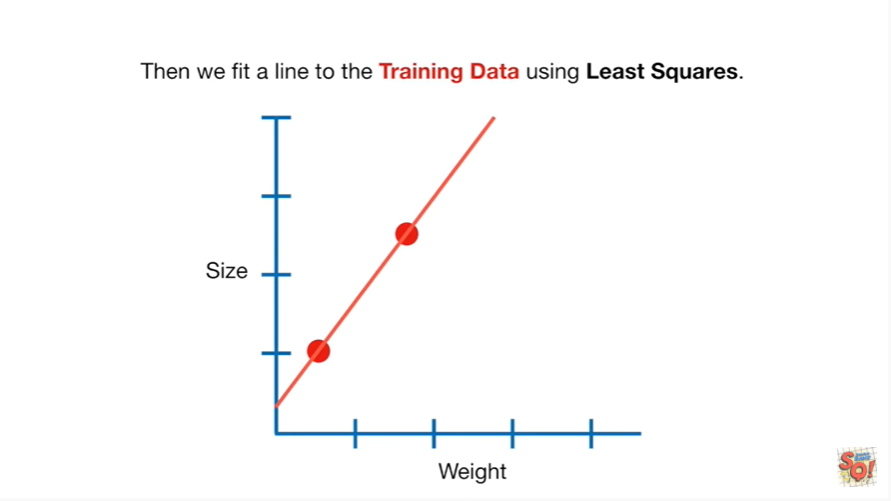
and we split the data into two sets.



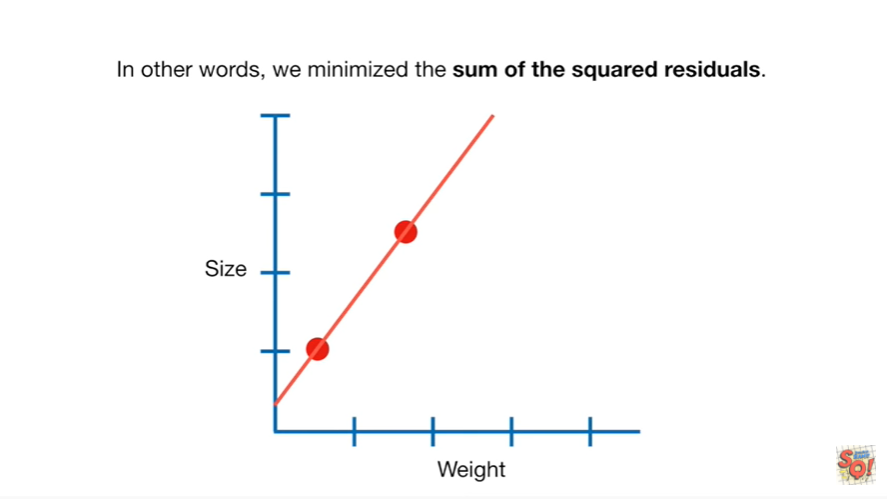
The red dots were training data



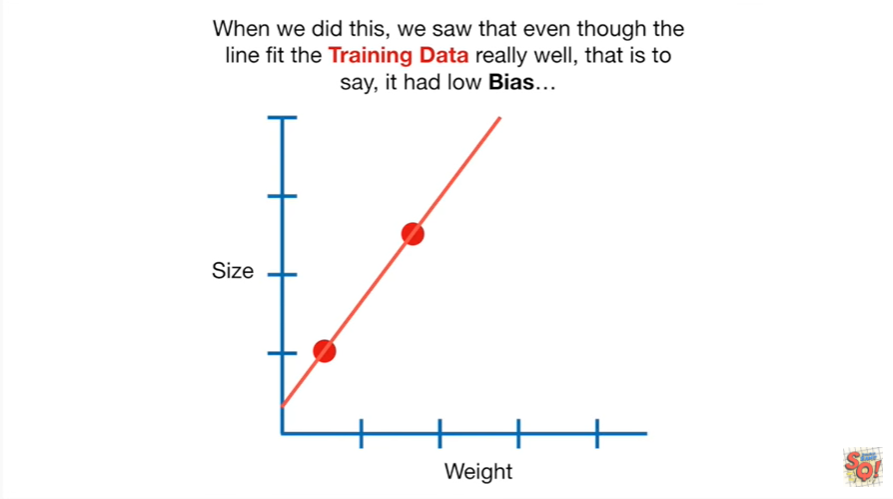
and the green dots were testing data.



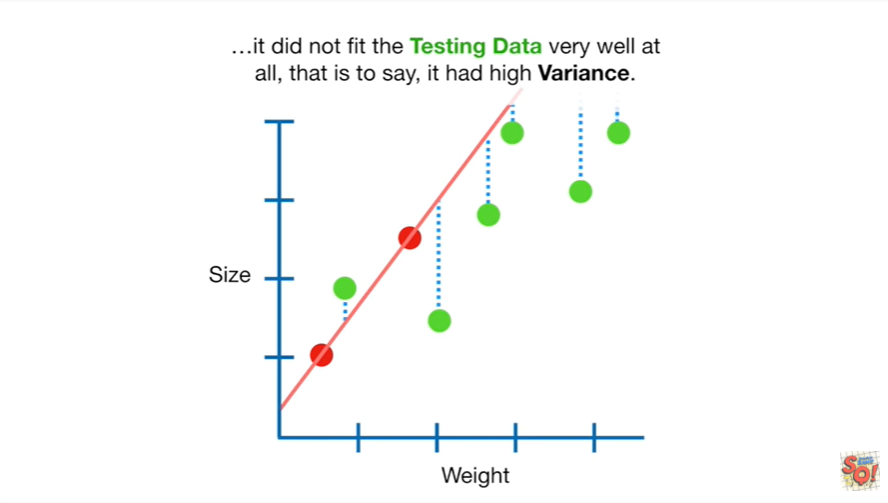
Then we fit a line to the training data using least squares.



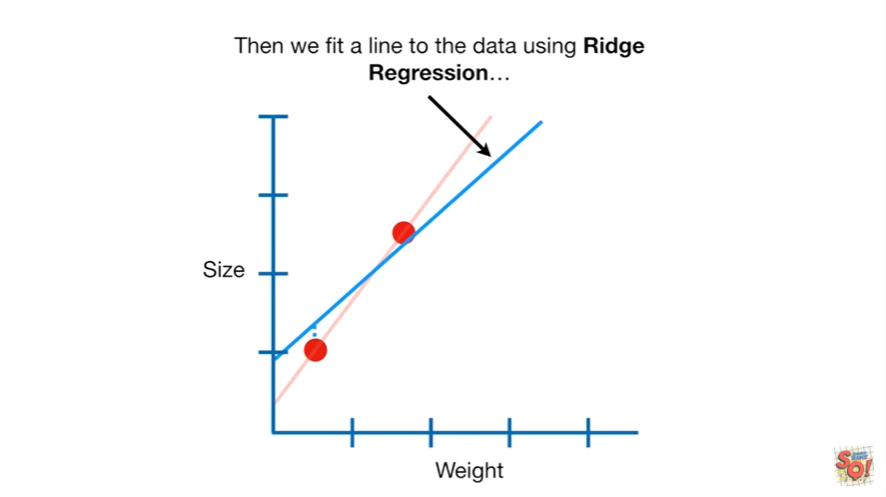
In other words we minimize the sum of the squared residuals.



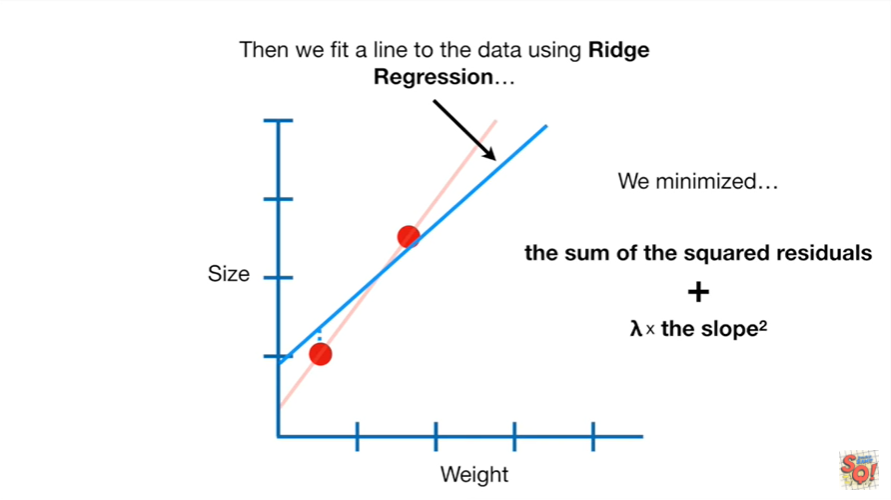
When we did this we saw that even though the line fit the training data really well that is to say it had low bias



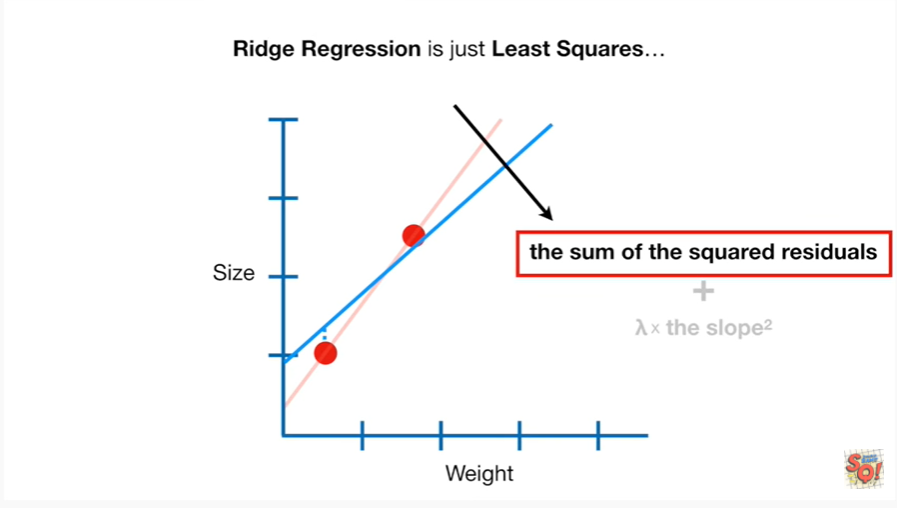
it did not fit the testing data very well at all that is to say it had high variance.



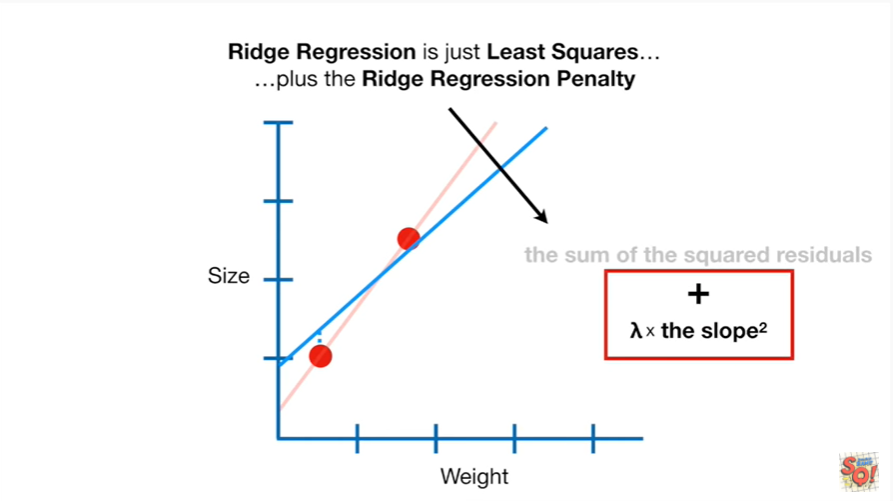
Then we fit a line to the data using Ridge regression



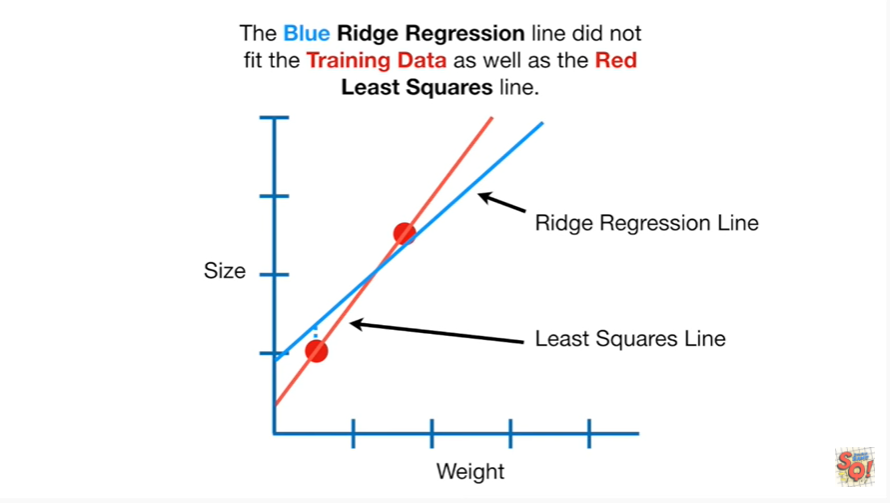
we minimized the sum of the squared residuals plus lambda times the slope squared.



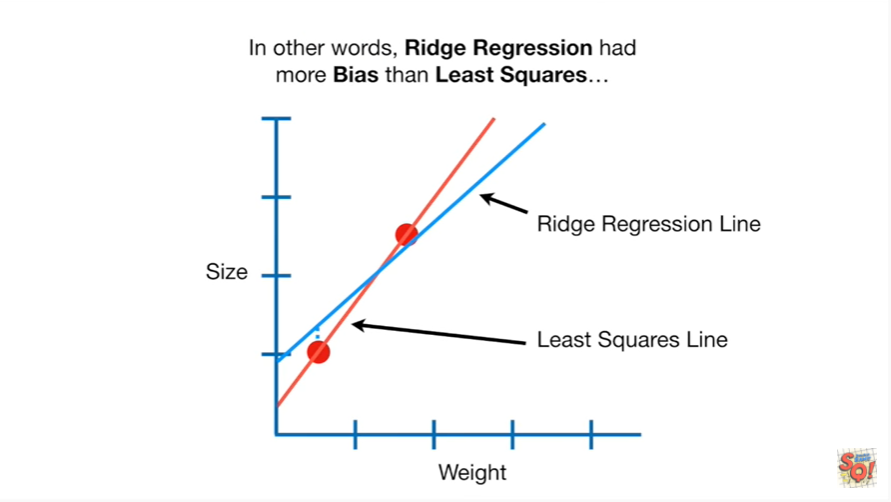
Ridge regression is just least squares



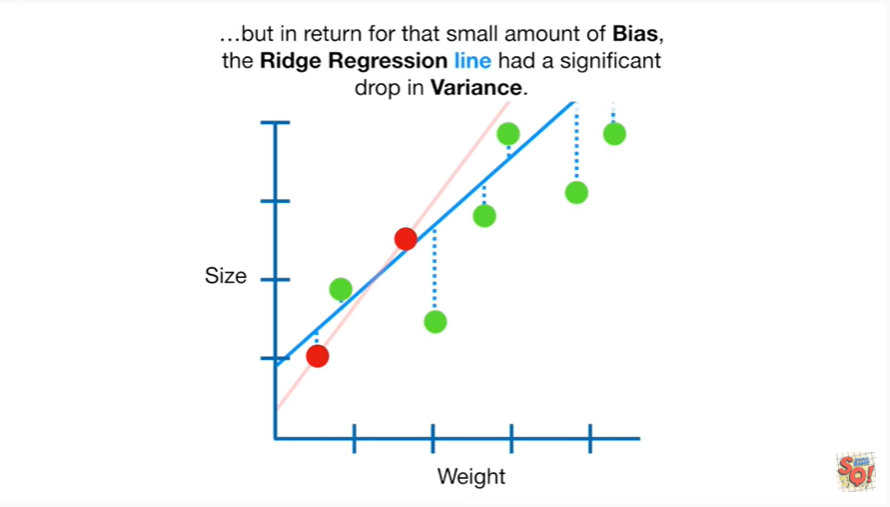
plus the ridge regression penalty.



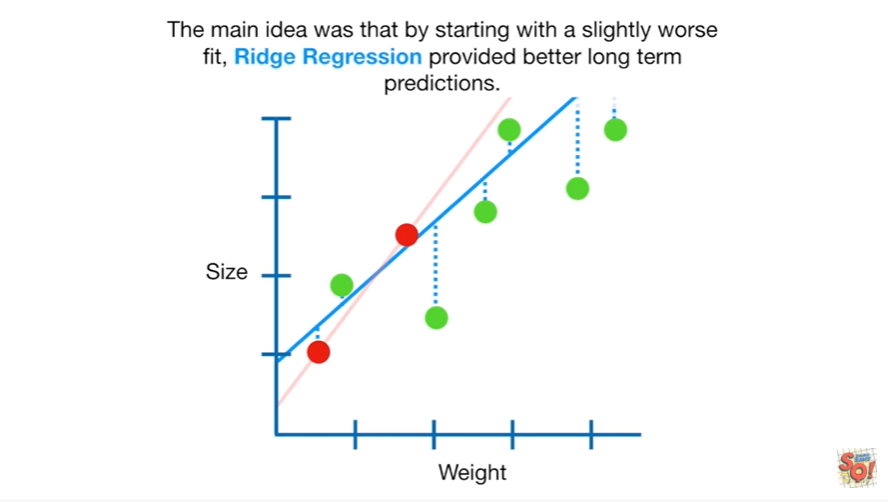
The Blue Ridge regression line did not fit the training data as well as the red least squares line.



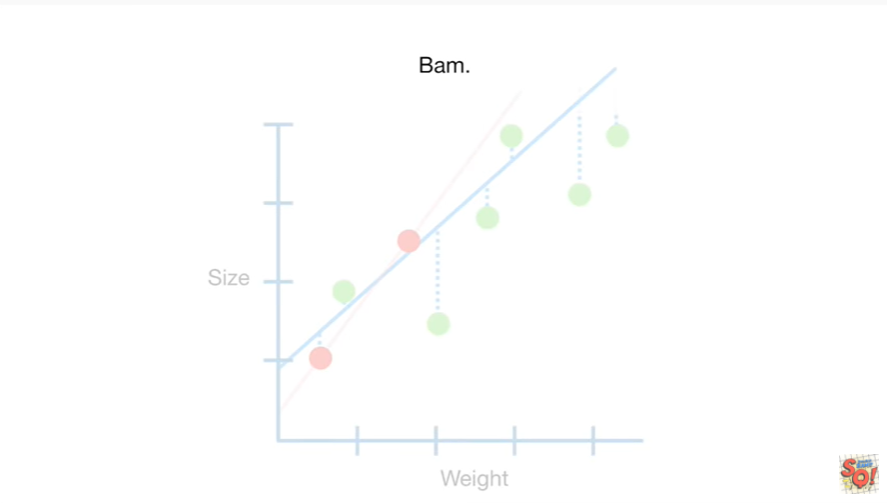
In other words Ridge regression had more bias than least squares



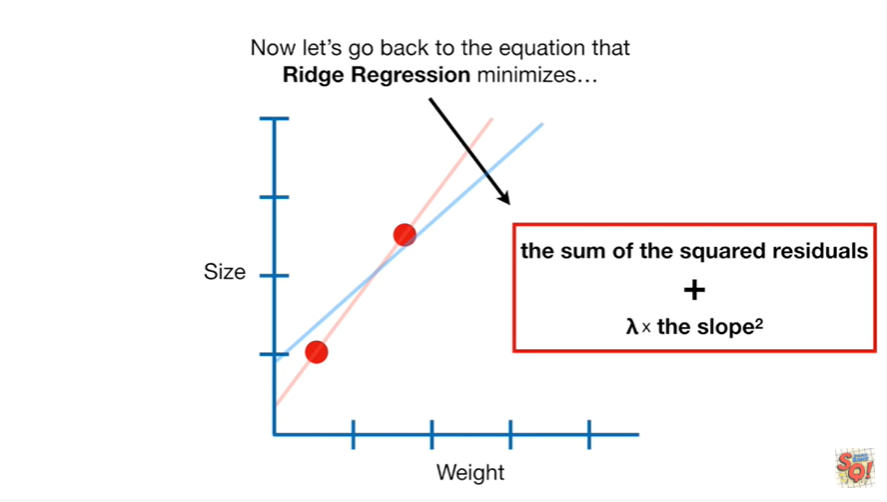
but in return for that small amount of bias the ridge regression line had a significant drop in variance.



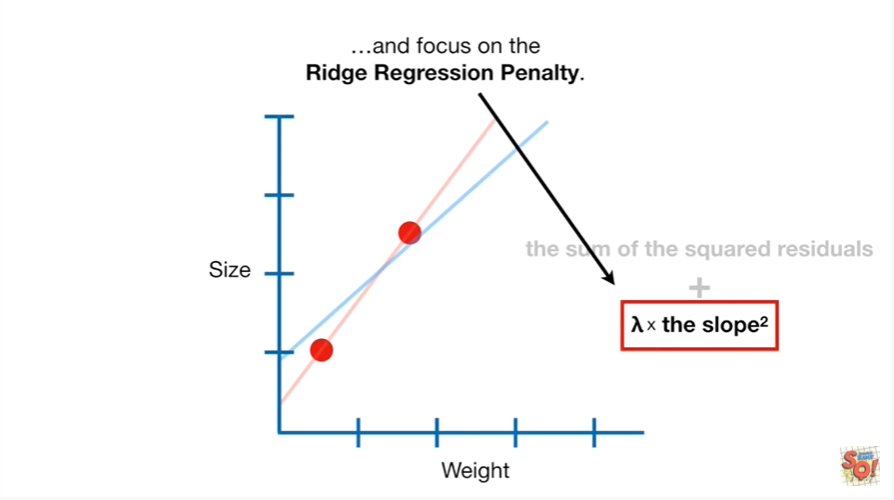
The main idea was that by starting with a slightly worse fit Richard rushman provided better long-term predictions.



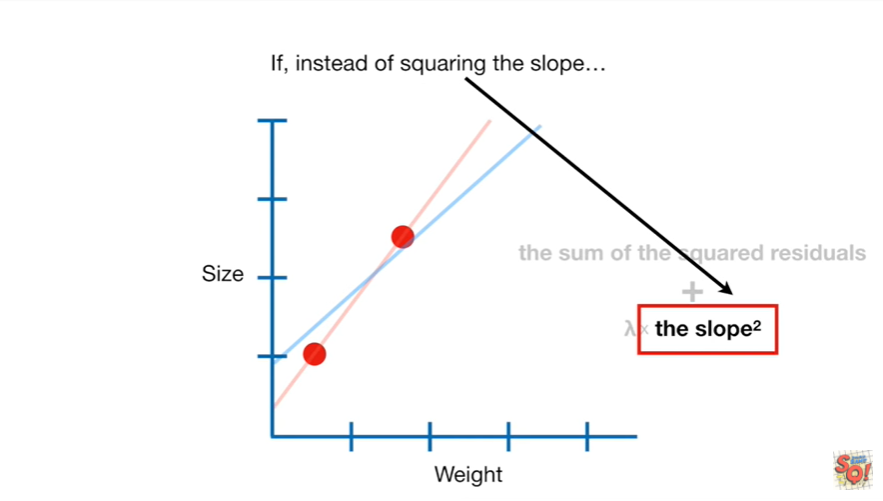
BAM.



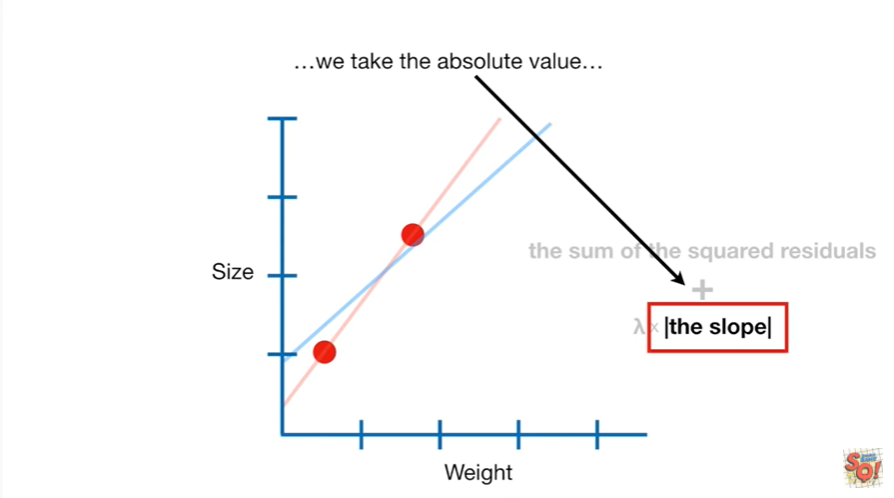
Now let's go back to the equation that Ridge regression minimizes



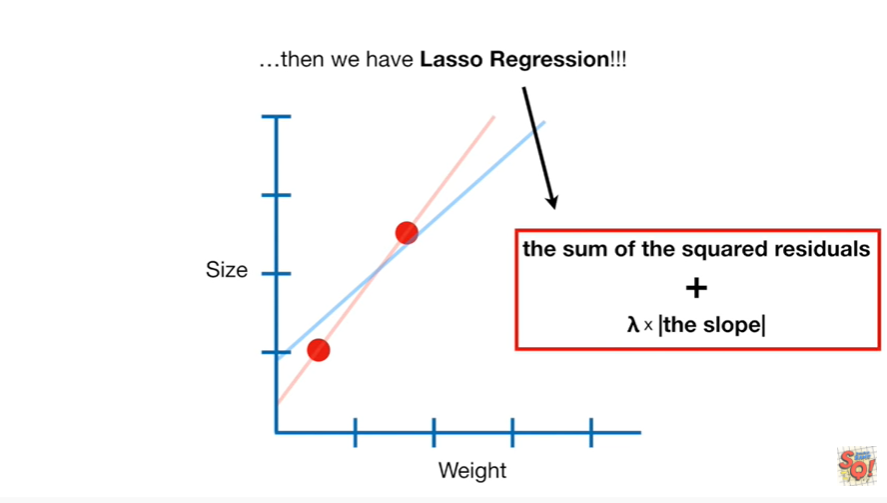
and focus on the ridge regression penalty.



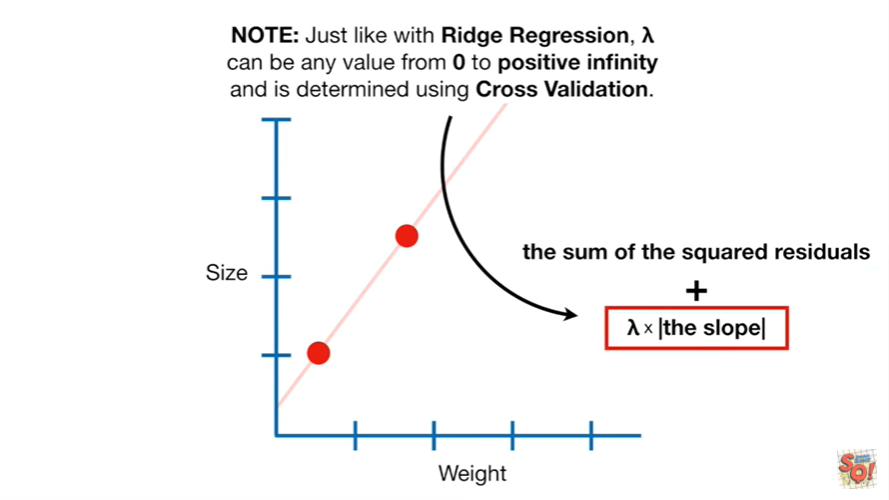
If instead of squaring the slope



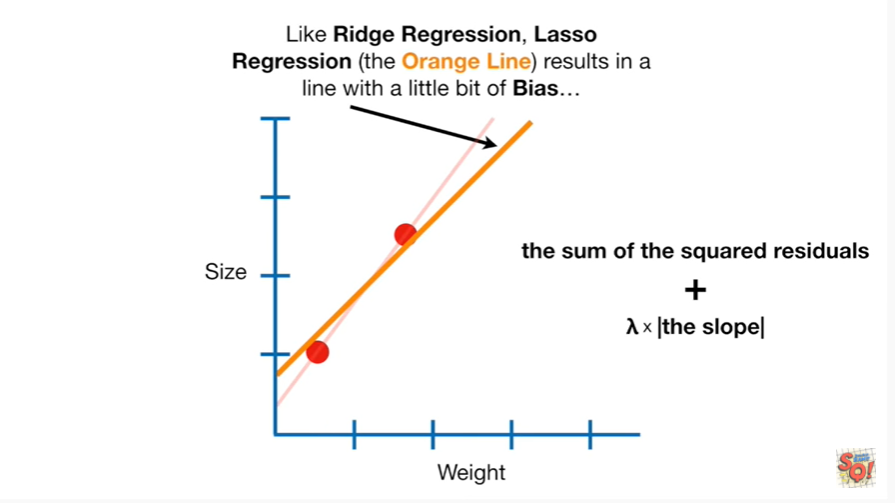
we take the absolute value



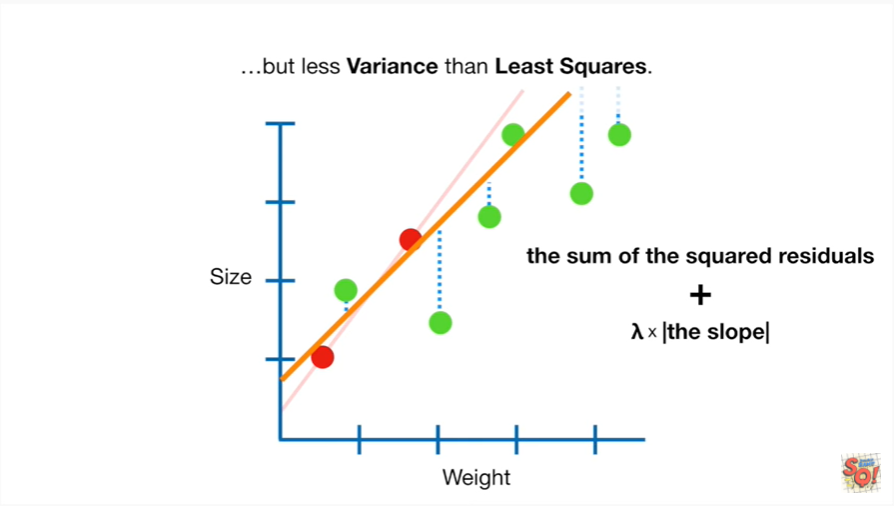
then we have lasso regression



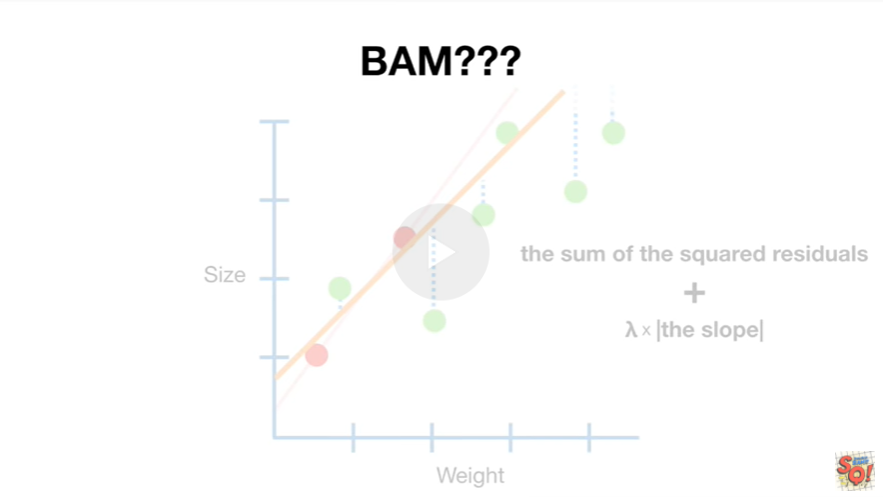
Note : just like with Ridge regression lambda can be any value from 0 to positive infinity and is determined using cross-validation.



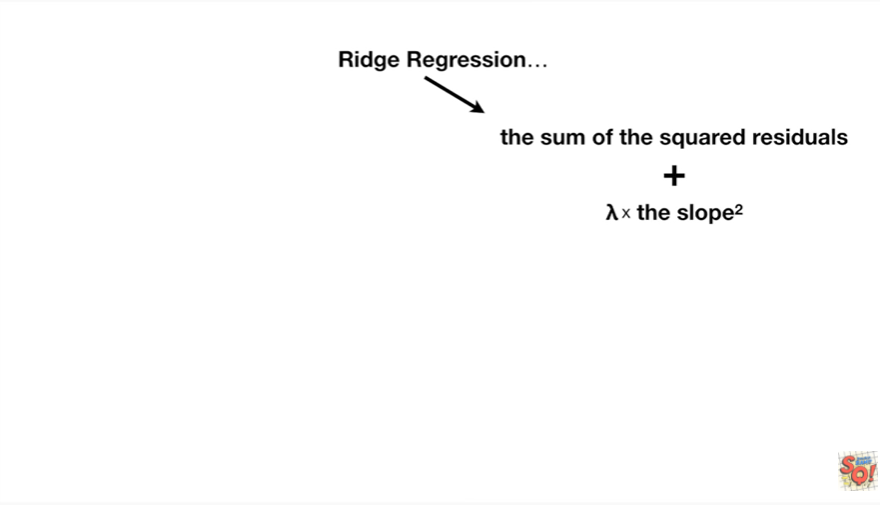
Like Ridge regression lasso regression the orange line results in a line with a little bit of bias



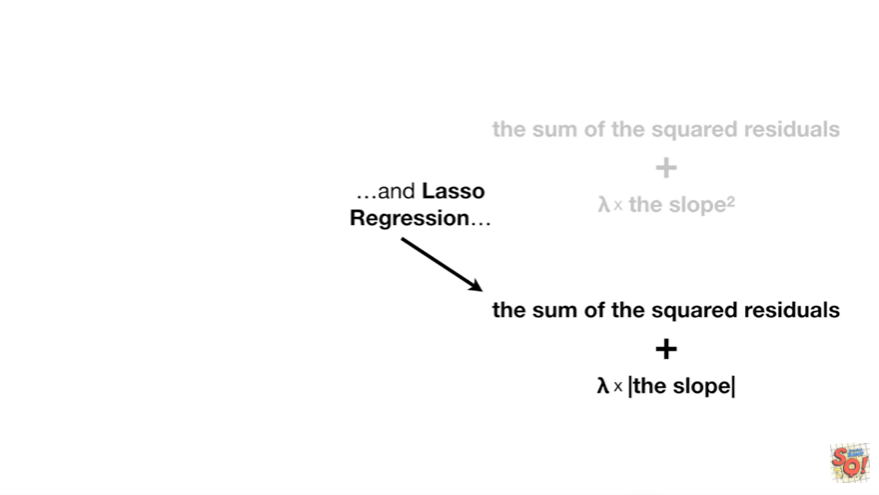
but less variance than least squares



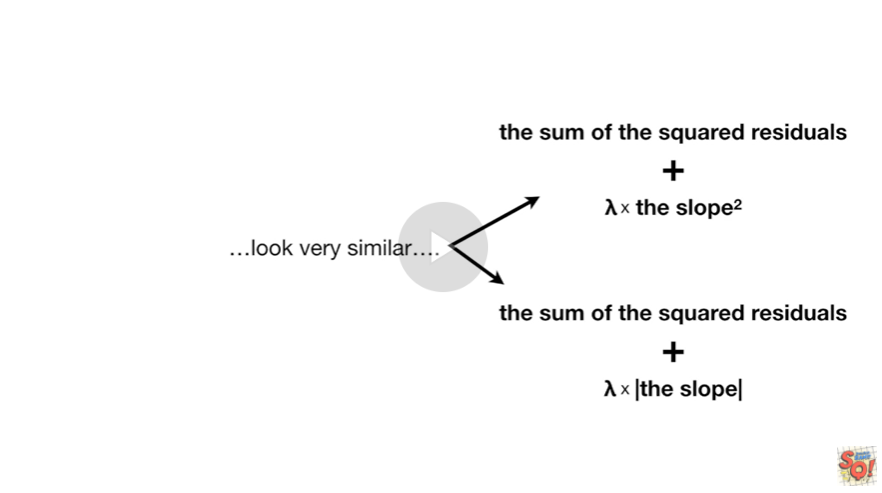
BAM ???



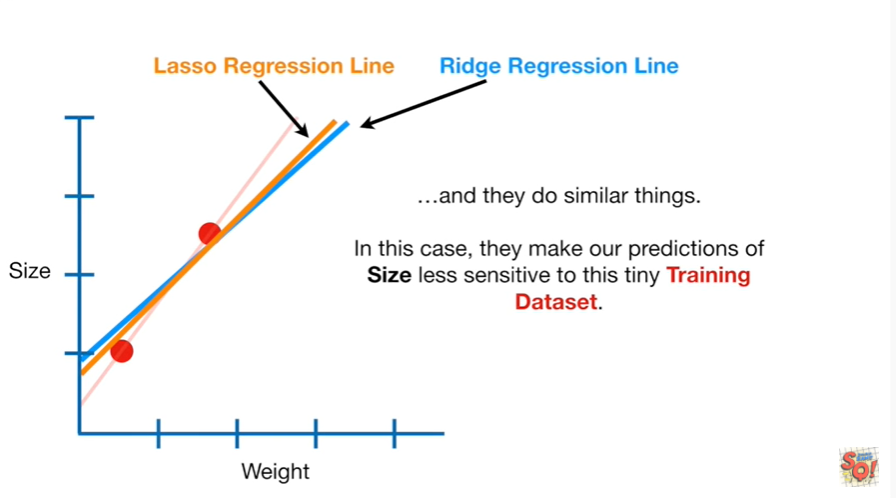
Ridge regression



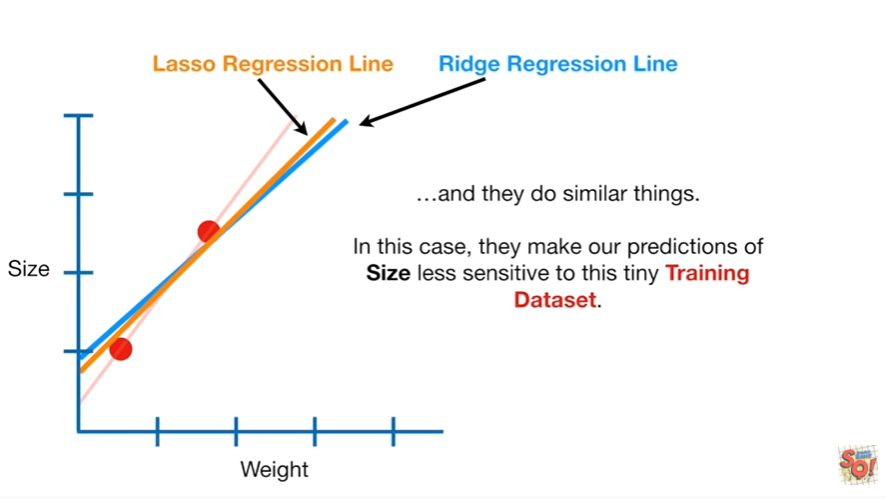
and lasso regression



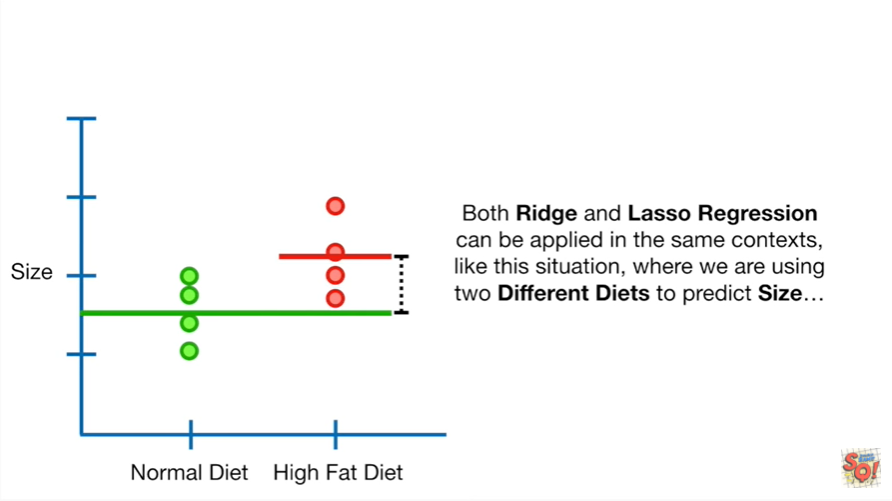
look very similar



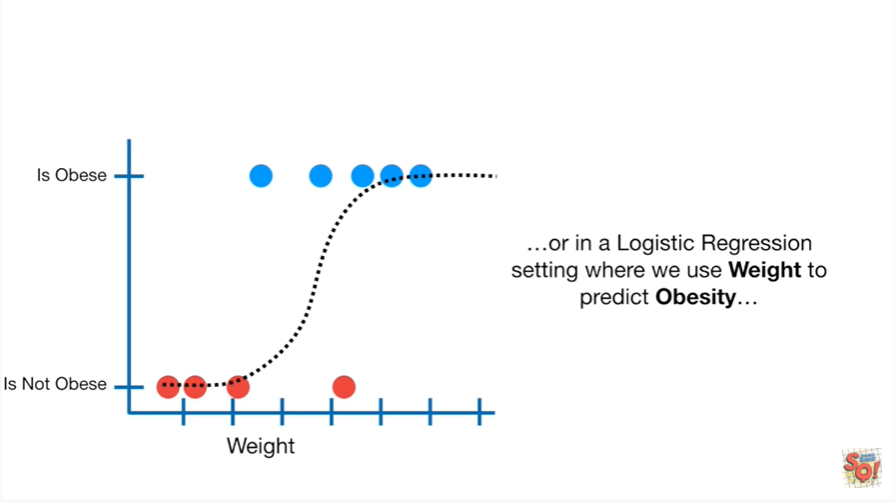
and they do similar things.



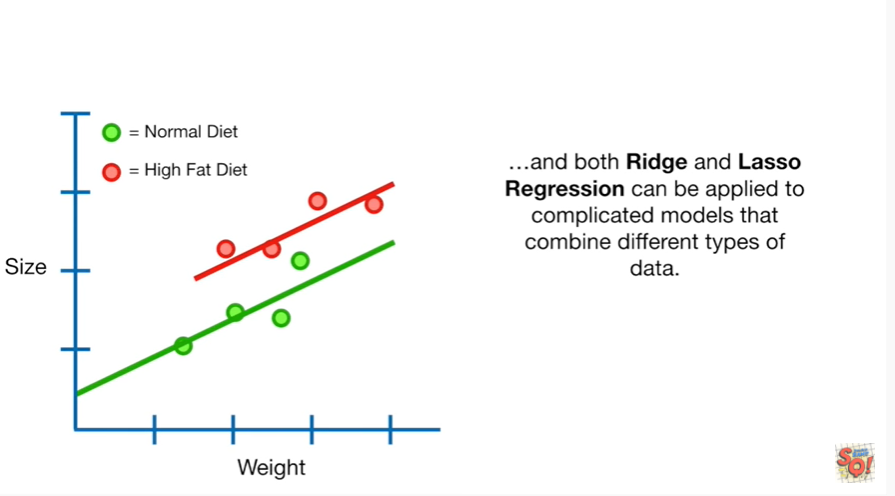
In this case they make our predictions of size less sensitive to this tiny training data set.



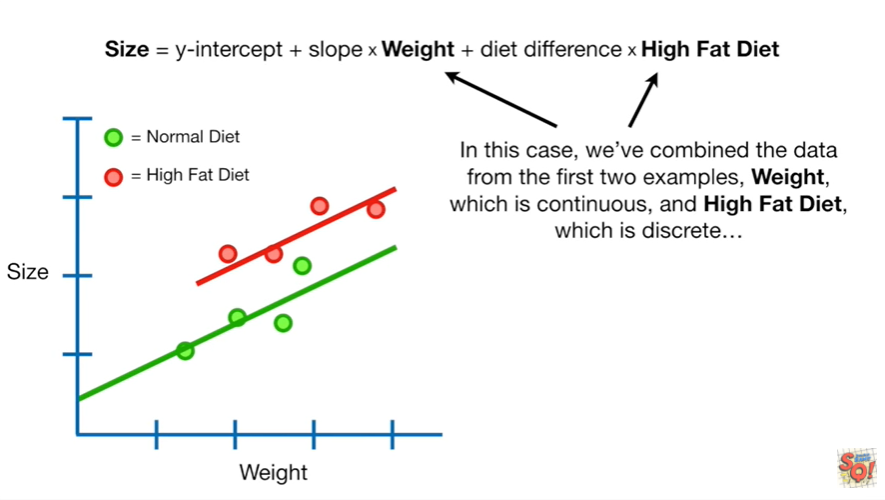
Both Ridge and lasso regression can be applied in the same context like this situation where we are using two different diets to predict size



or in a logistic regression setting where we use weight to predict obesity



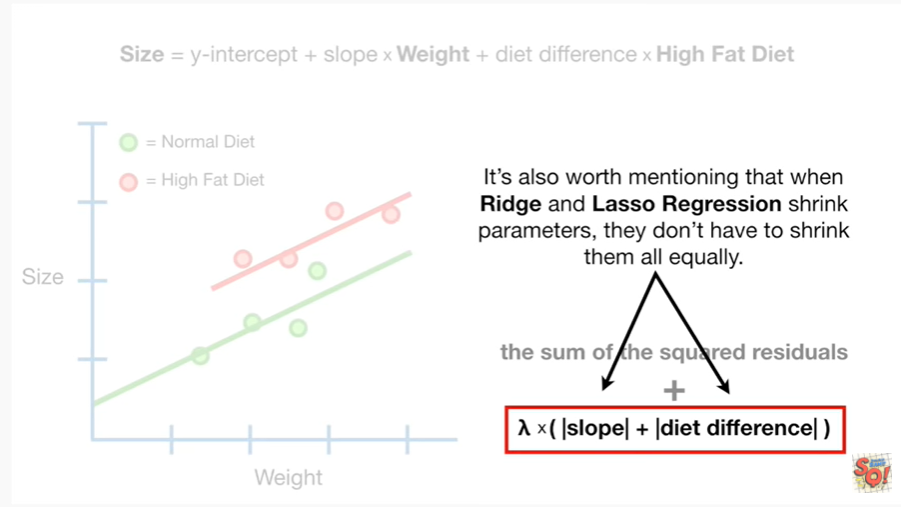
and both Ridge and lasso regression can be applied to complicated models that combine different types of data.



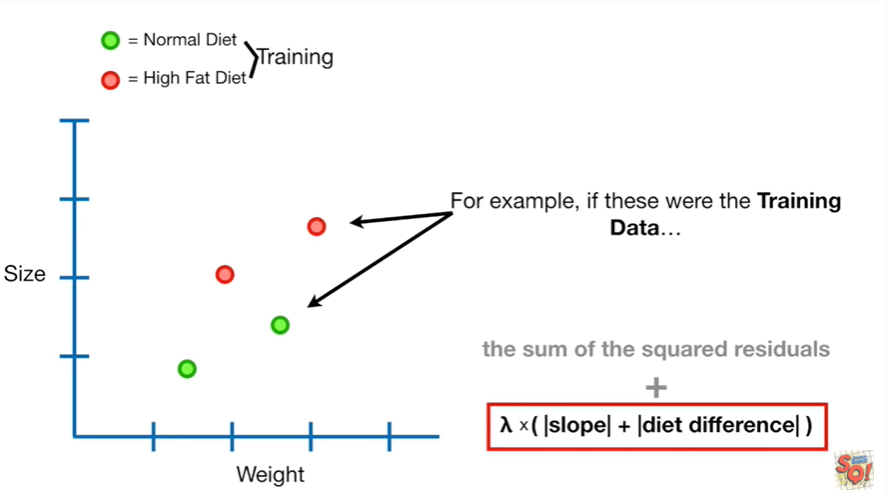
In this case we've combined the data from the first two examples weight which is continuous and high fat diet which is discrete.



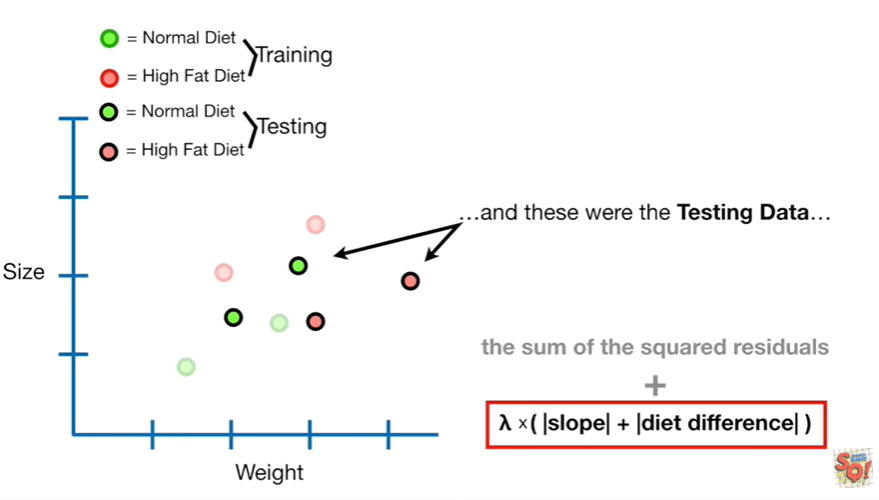
Just like the ridge regression penalty the lasso regression penalty contains all of the estimated parameters except for the y-intercept.



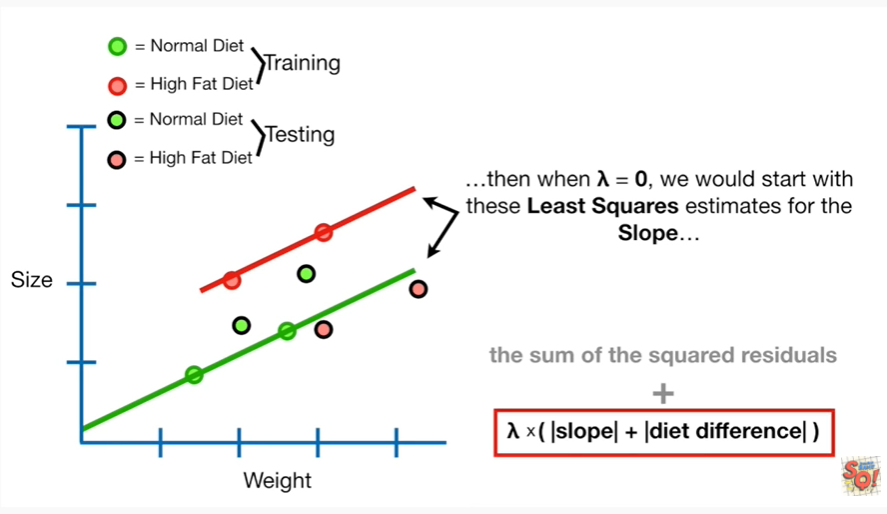
It's also worth mentioning that when Ridge and lasso regression shrink parameters they don't have to shrink them all equally.



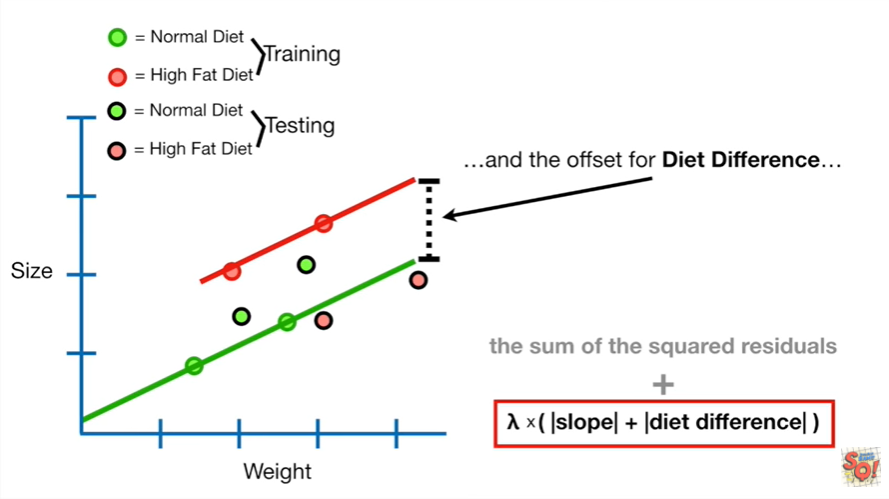
For example if these were the training data



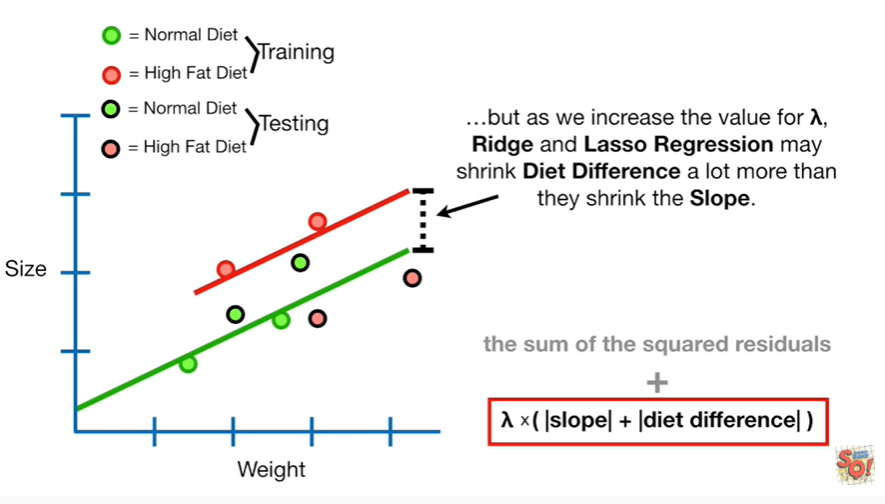
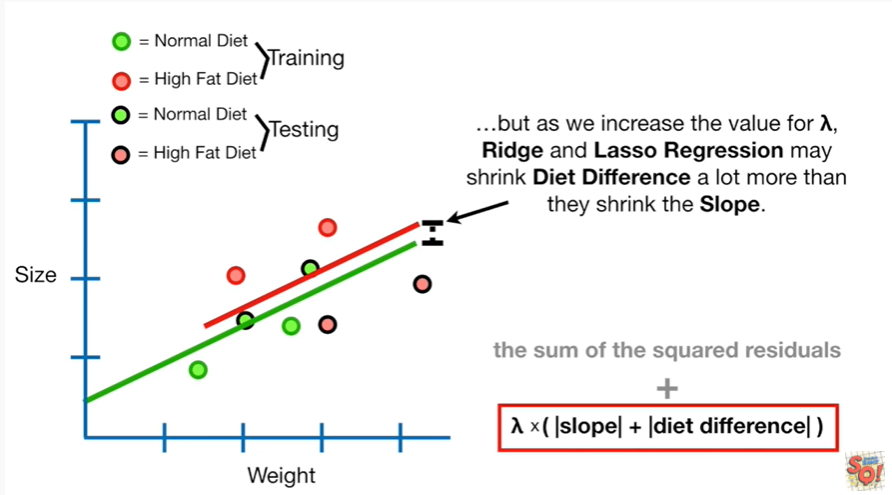
and these were the testing data



then when lambda equals zero we would start with these least squares estimates for the slope



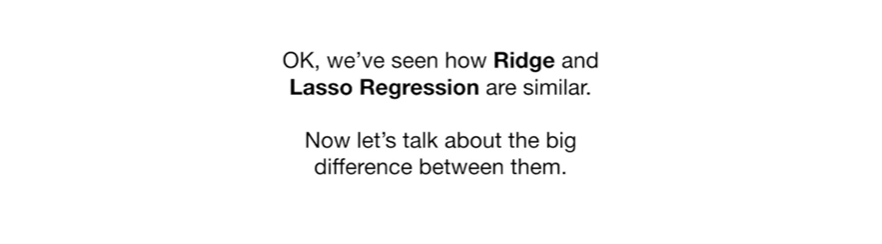
and the offset for diet difference

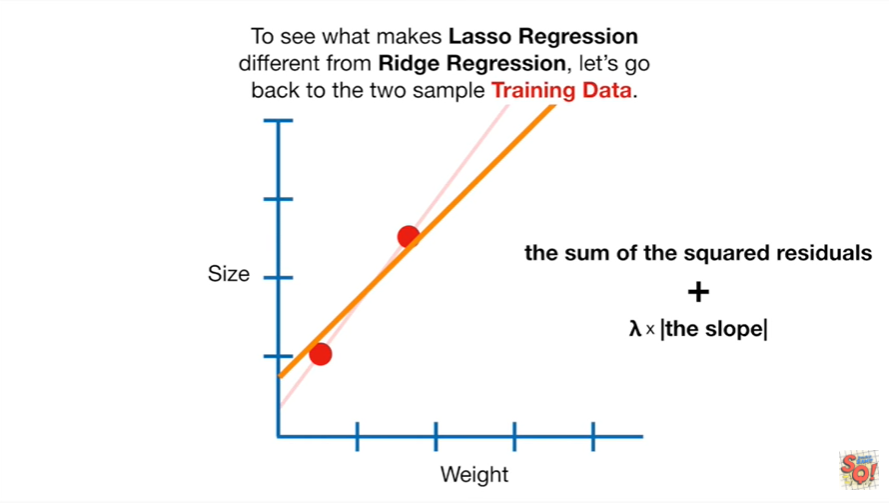
but as we increase the value for lambda, ridge and lasso regression may shrink diet difference a lot more than they shrink the slope.



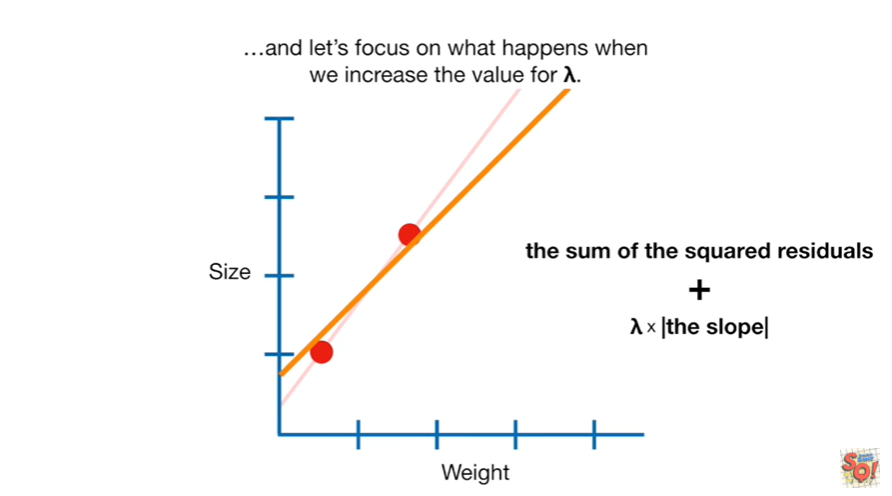
Okay we've seen how Ridge and lasso regression are similar.



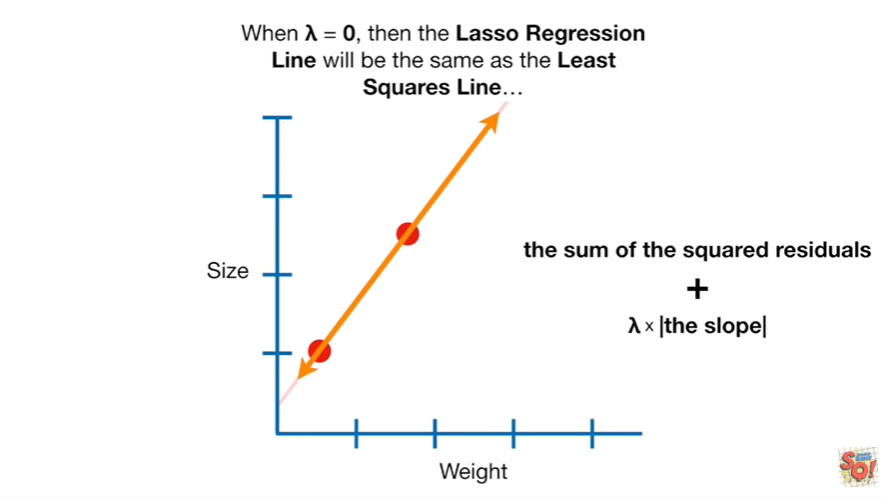
Now let's talk about the big difference between them.



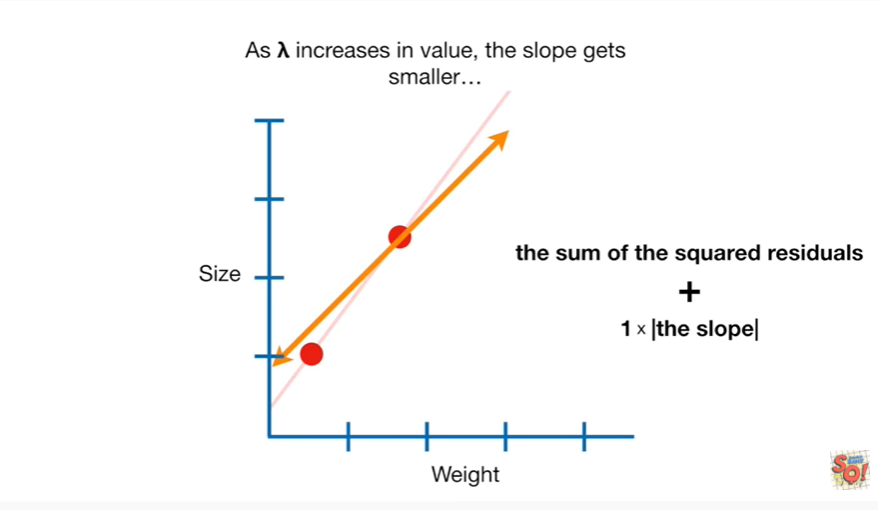
To see what makes lasso regression different from Ridge regression let's go back to the two-sample training data

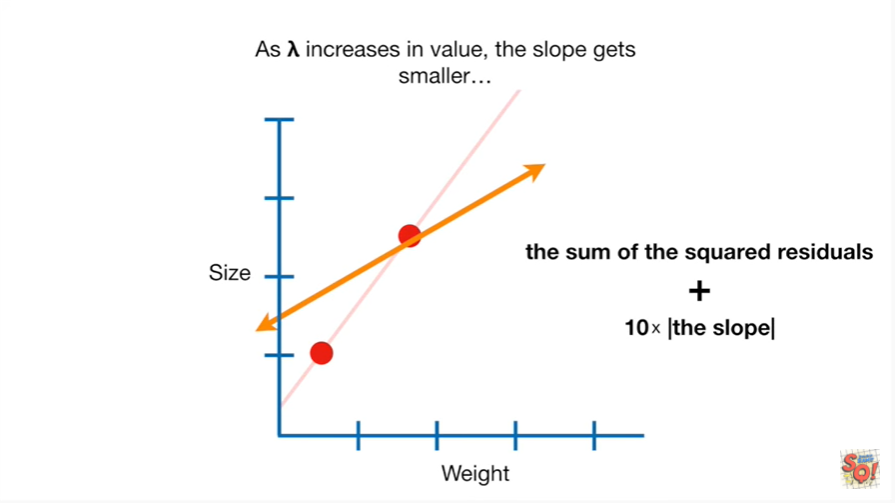
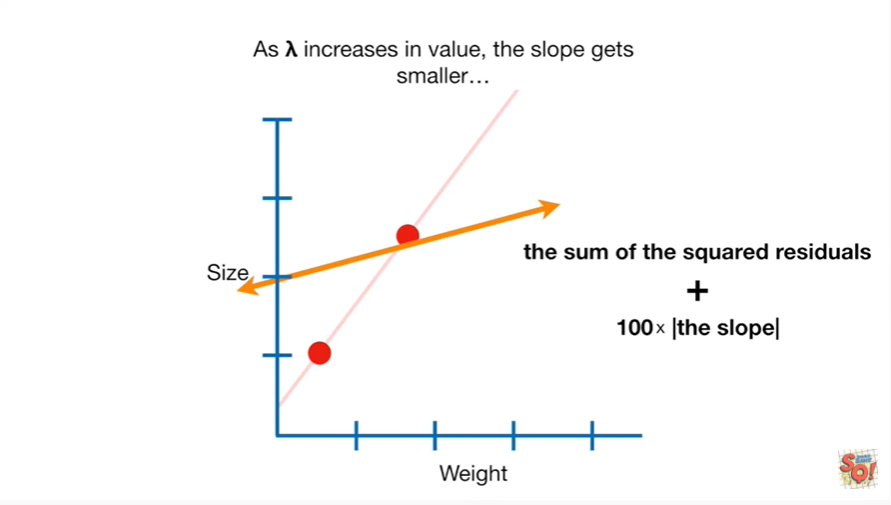


and let's focus on what happens when we increase the value for lambda.

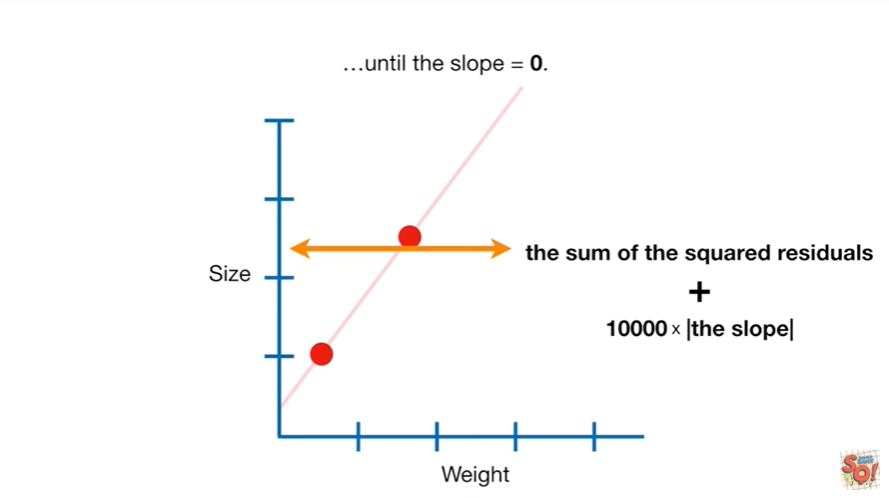


When lambda equals zero then the lasso regression line will be the same as the least squares line.

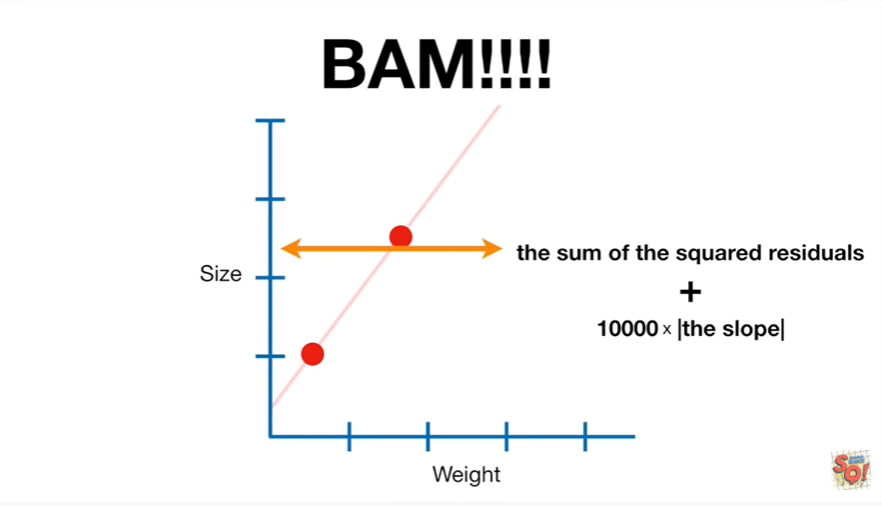


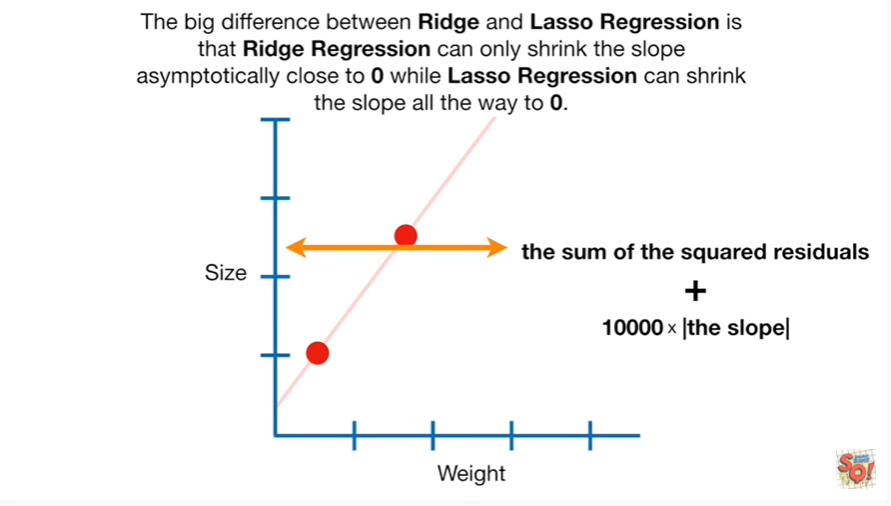
As lambda increases in value the slope gets smaller



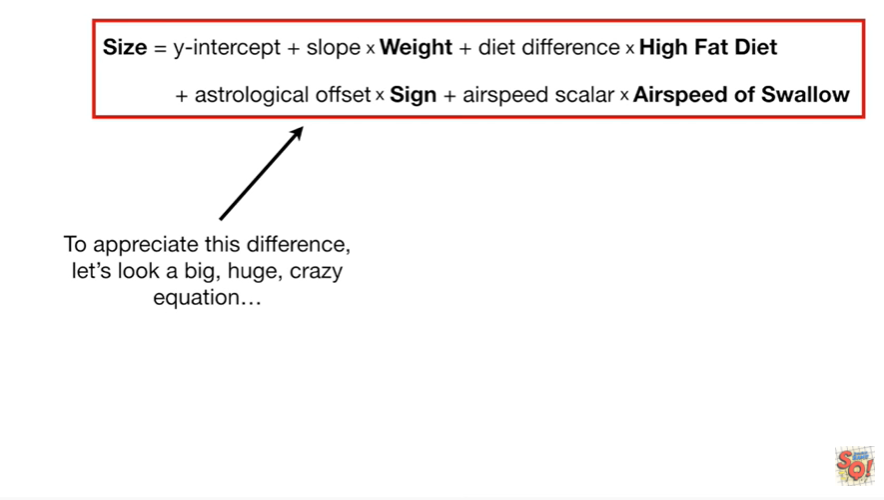
until the slope equals zero.



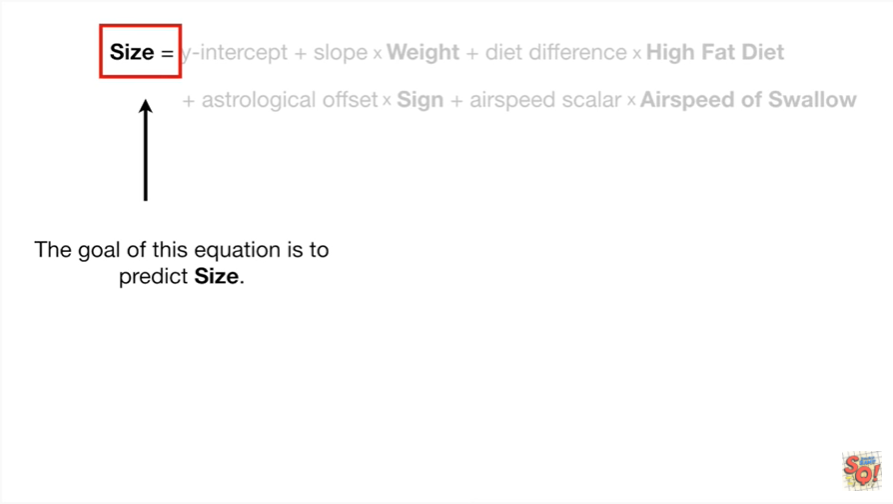
BAM !!!!



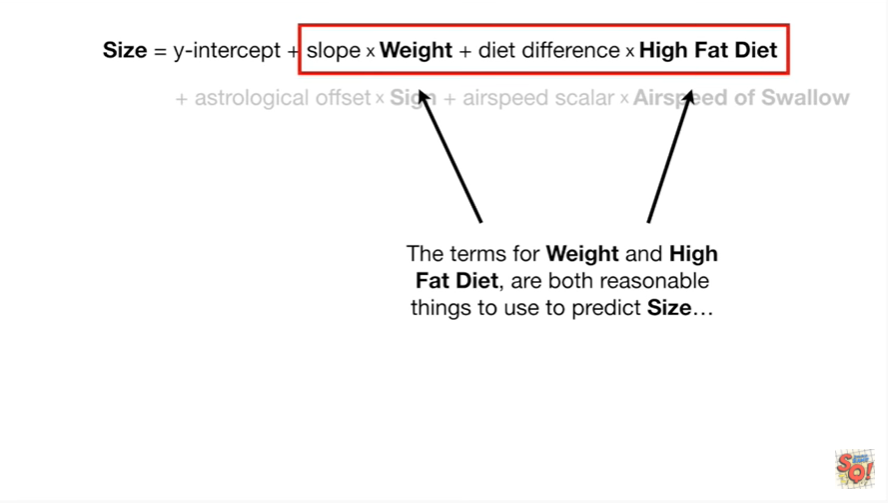
The big difference between Ridge and lasso regression is that Ridge regression can only shrink a slope asymptotically close to zero while lasso regression can shrink the slope all the way to 0.



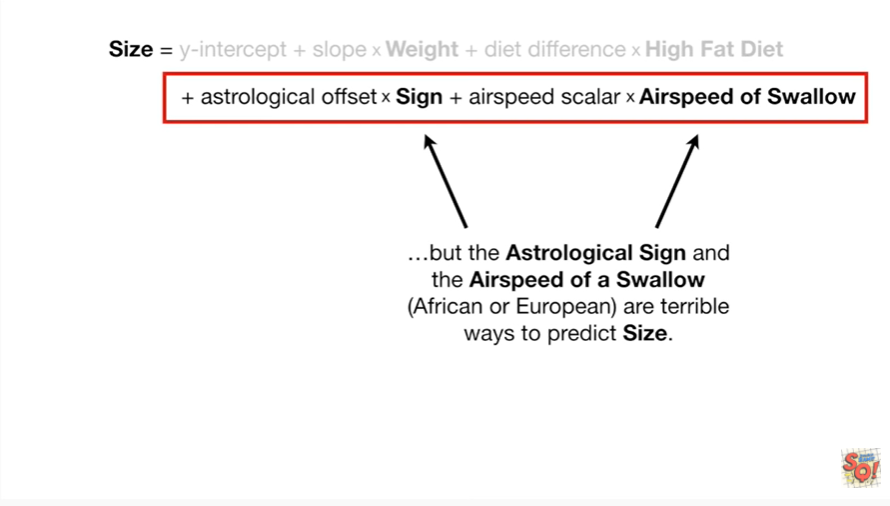
To appreciate this difference, let's look at a big huge crazy equation.



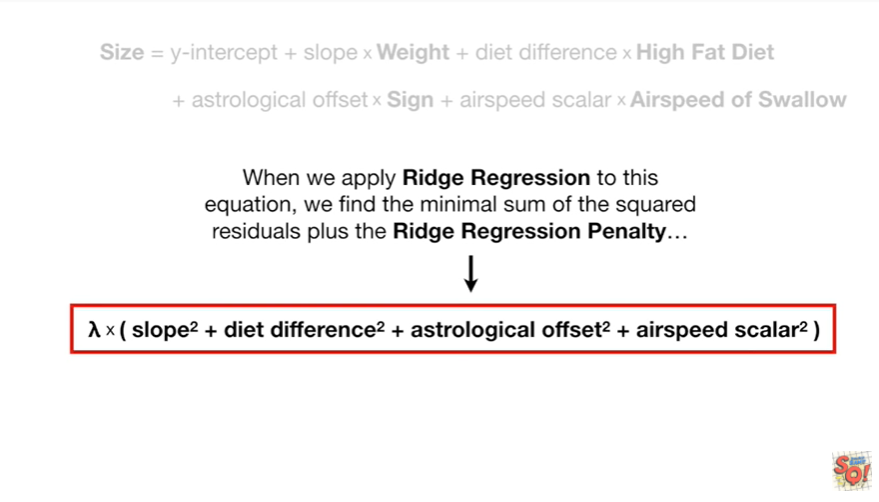
The goal of this equation is to predict size.



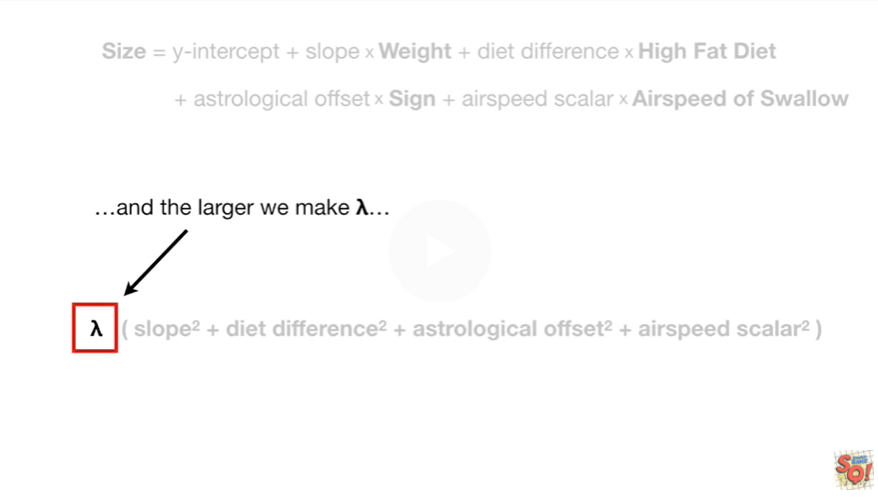
The terms for weight and high fat diet are both reasonable things to use to predict size



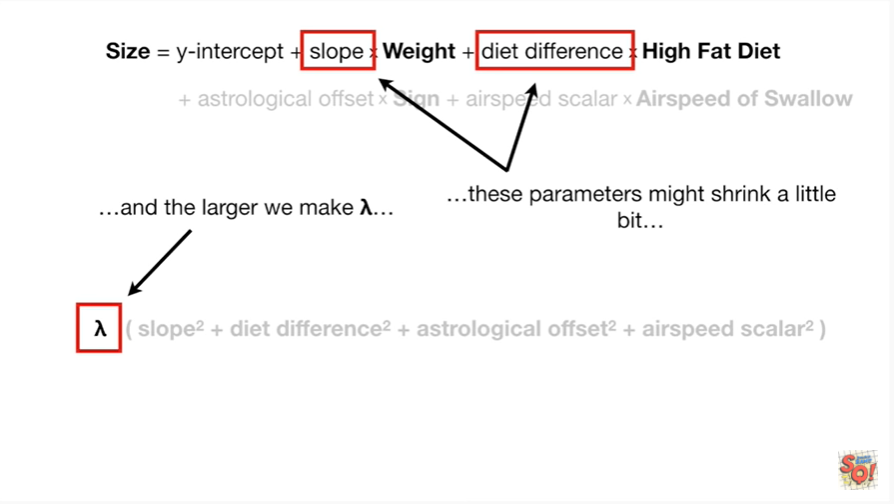
but the astrological sign and the airspeed of a swallow African or European are terrible ways to predict size.



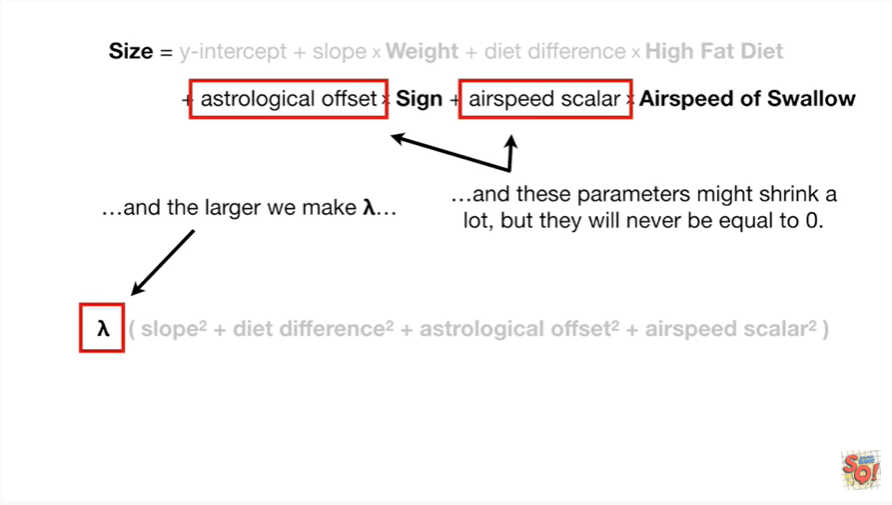
When we apply Richard Russian to this equation we find the minimal sum of the squared residuals plus the ridge regression penalty



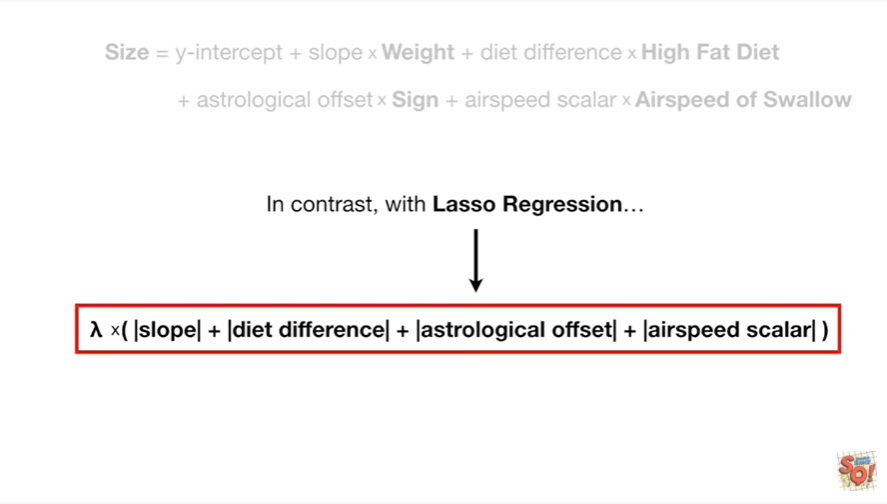
and the larger we make lambda



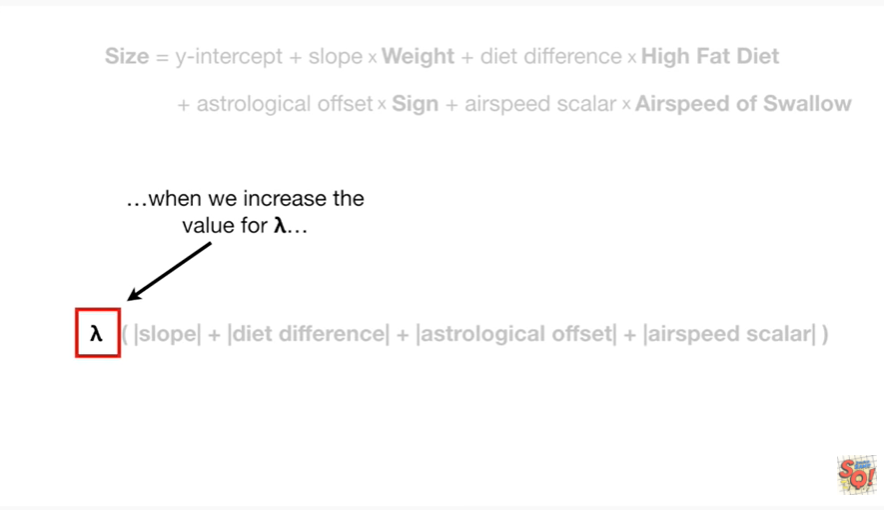
these parameters might shrink a little bit



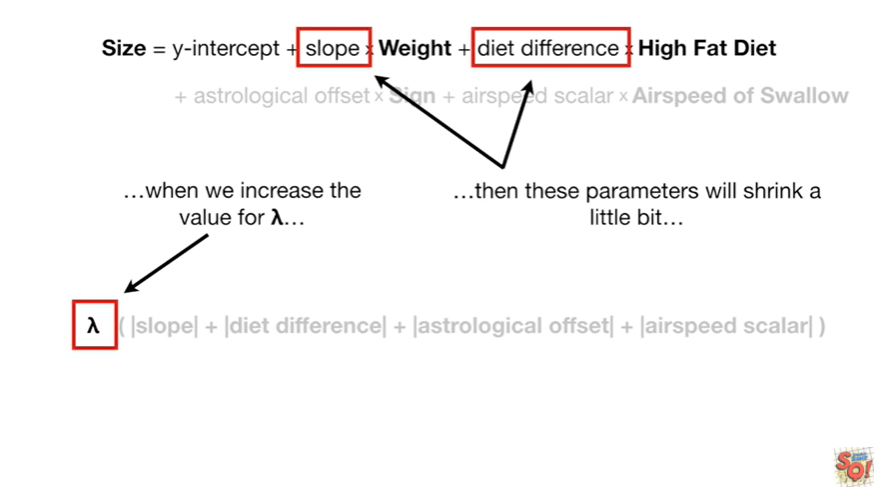
and these parameters might shrink a lot but they will never be equal to 0.



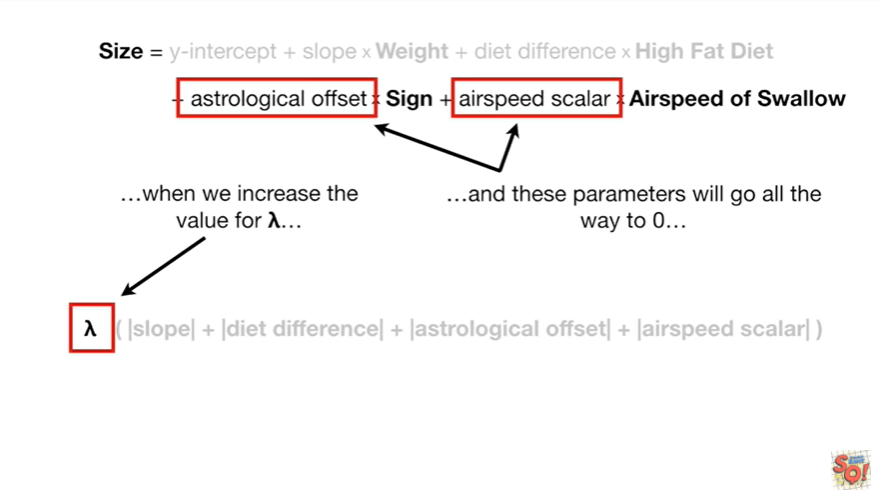
In contrast with lasso regression



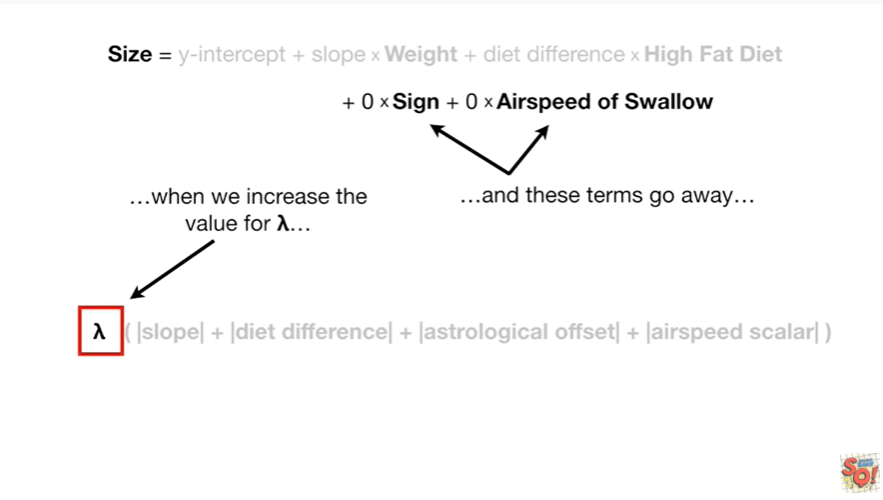
when we increase the value for lambda



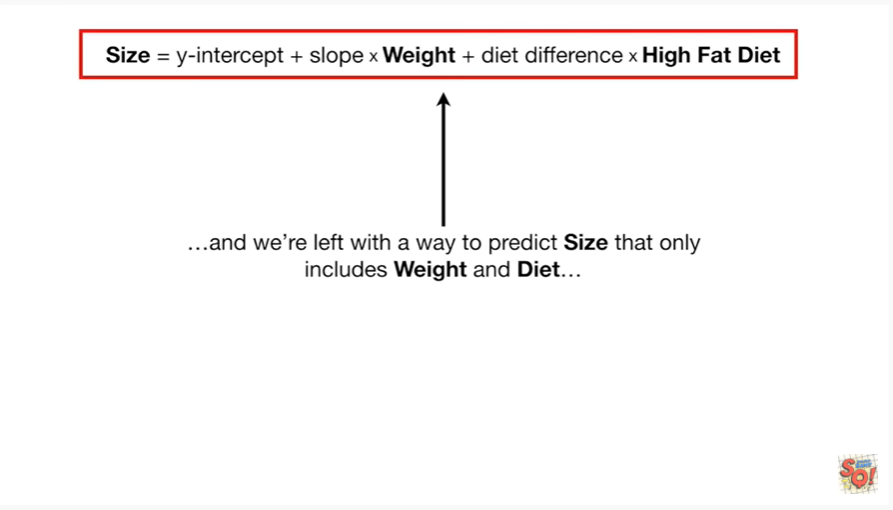
then these parameters will shrink a little bit



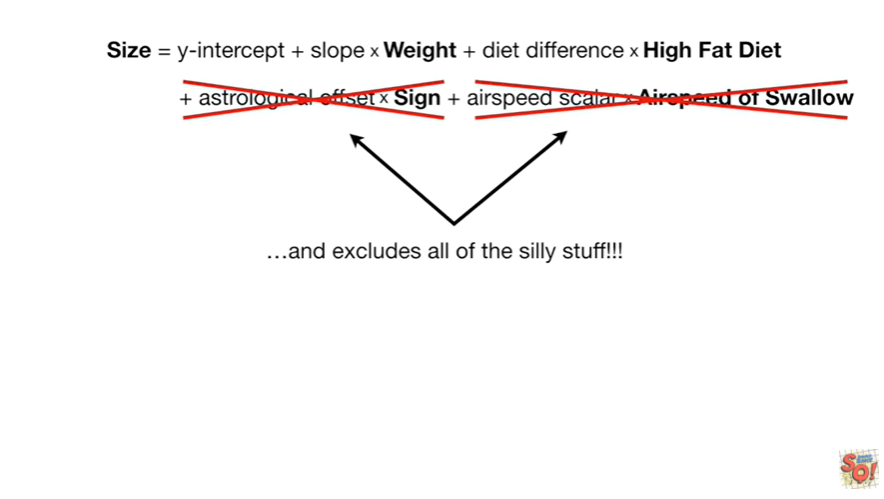
and these parameters will go all the way to 0



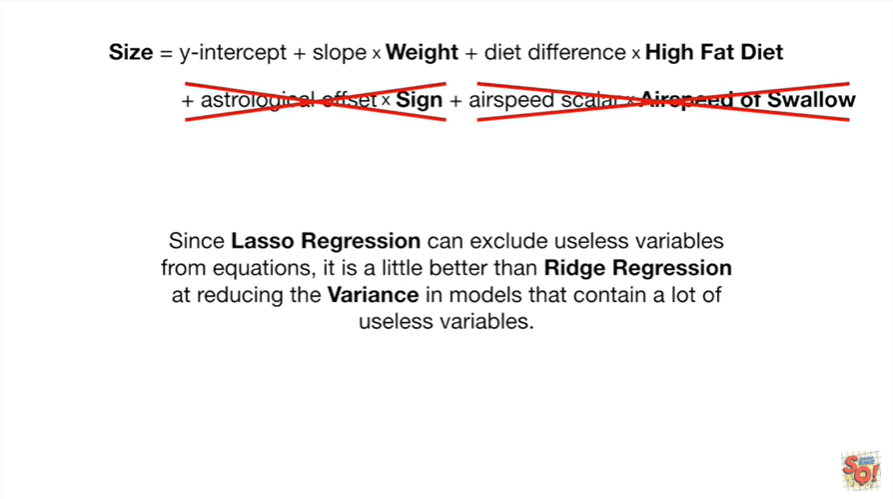
and these terms go away



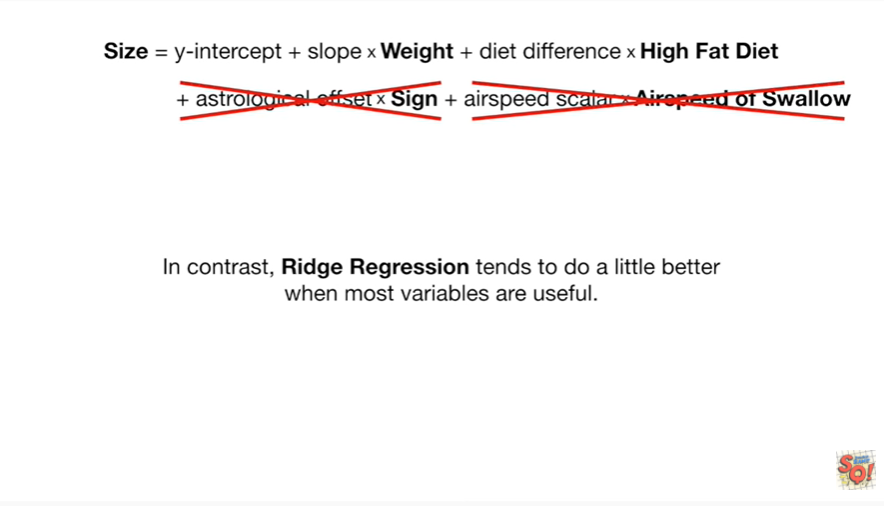
and we're left with a way to predict size that only includes weight and diet



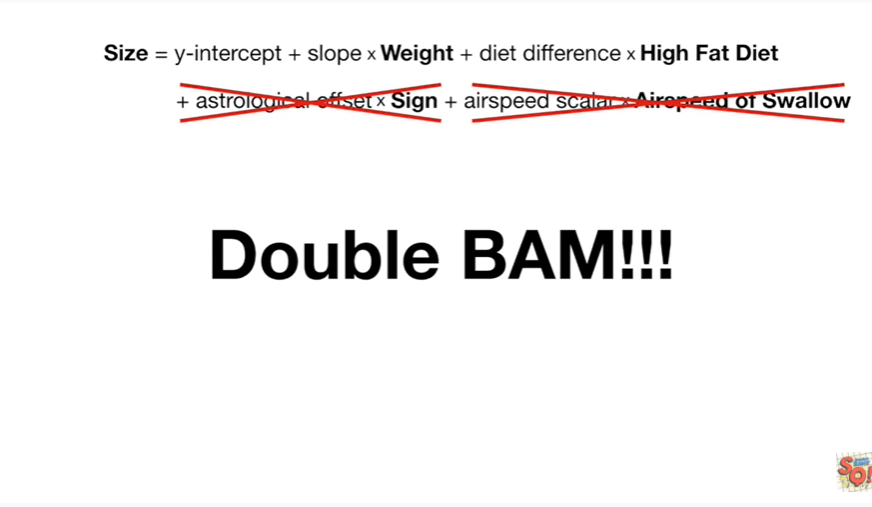
and excludes all of the silly stuff.



Since lasso regression can exclude useless variables from equations it is a little better than Ridge aggression at reducing the variance and models that contain a lot of useless variables.



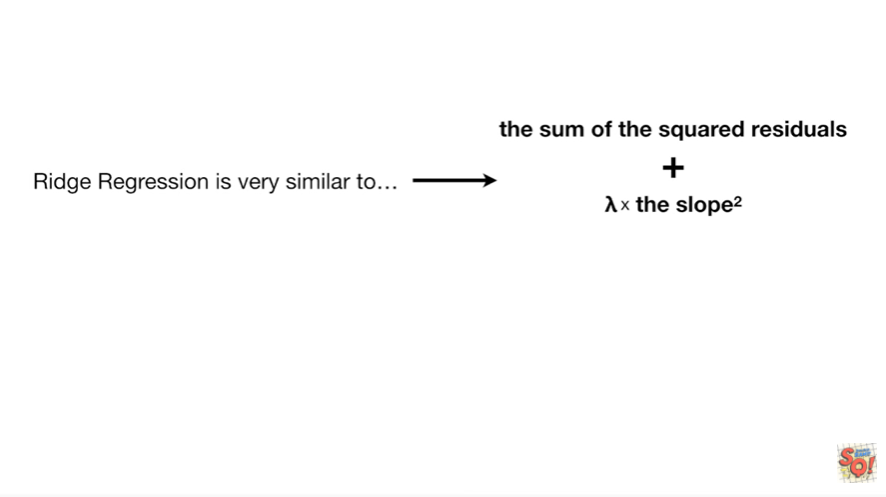
In contrast Ridge regression tends to do a little better when most variables are useful.



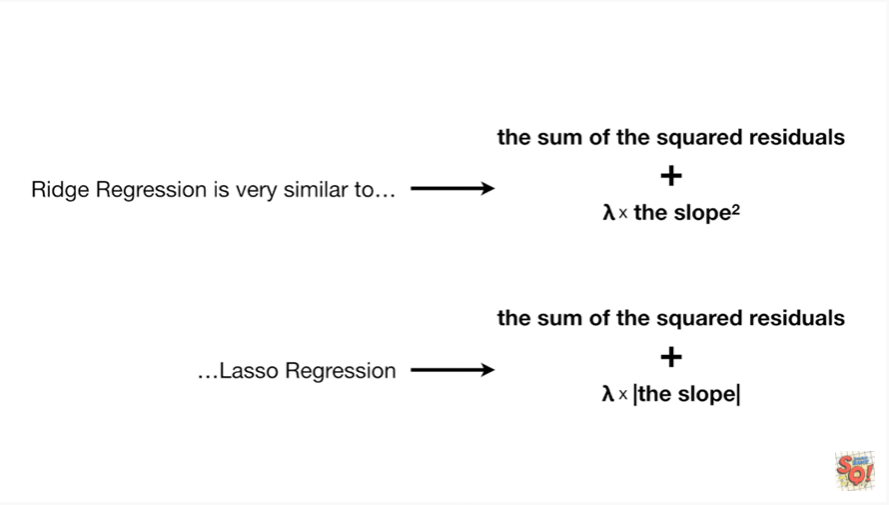
Double bam !!!



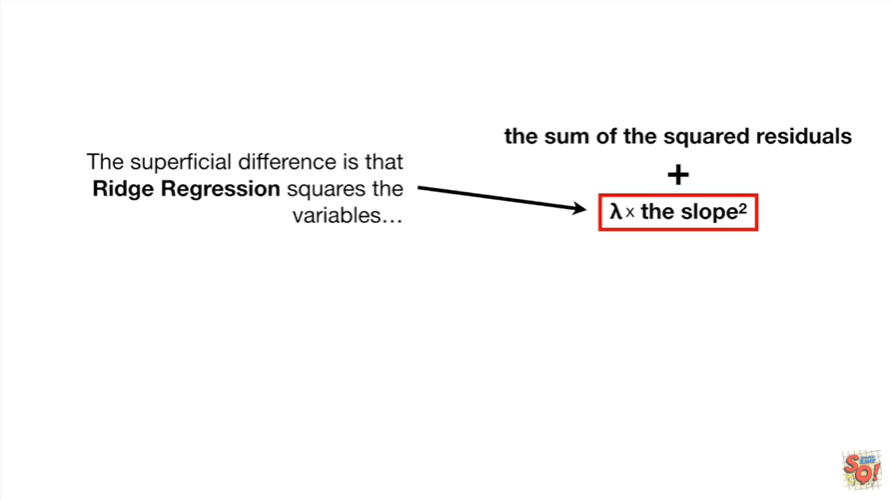
In summary :



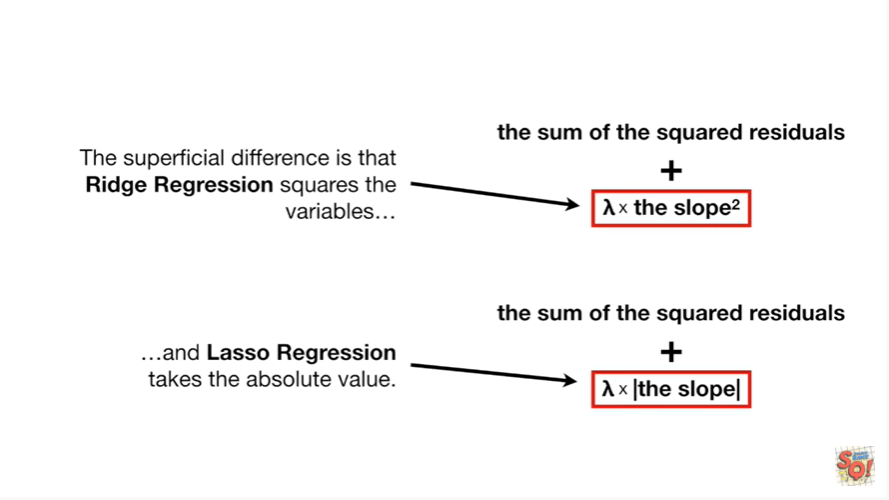
Ridge regression is very similar



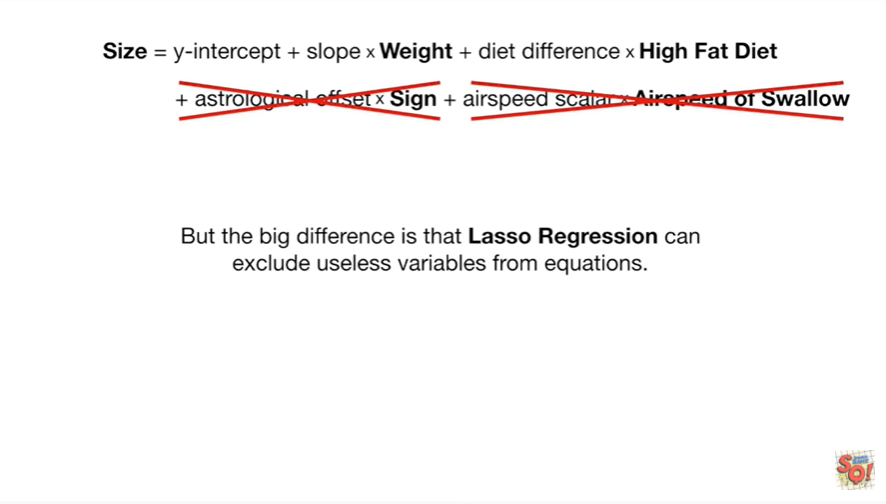
to lasso regression.



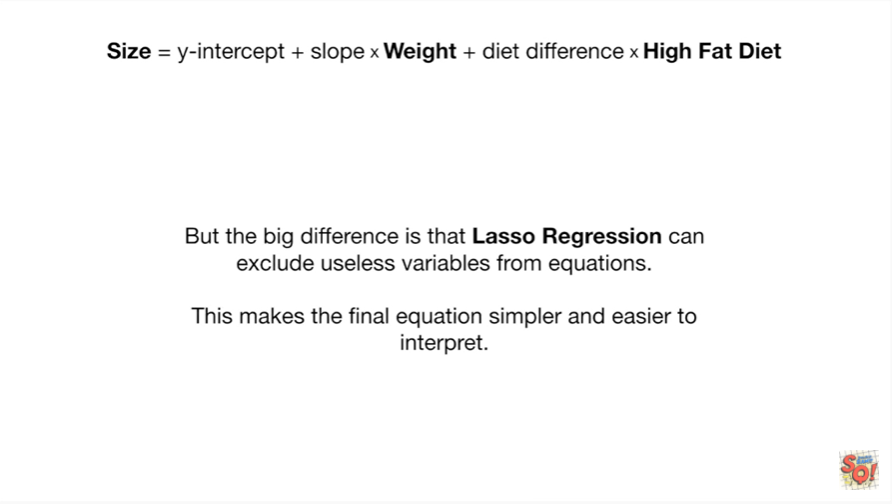
And the superficial difference is that Ridge regression squares the variables



and lasso regression takes the absolute value.



But he big difference is that Lasser aggression can exclude useless variables from equations.



This makes the final equation simpler and easier to interpret.