STATISTICS 101

by Brandon Foltz
www.youtube.com

Video Companion Guide

PLAYLIST 14

VIDEO 1

SIMPLE LINEAR REGRESSION

STATISTICS 101

SUM OF SQUARES

$$49 + 25 + 1 + 4 + 16 + 25 = 120$$

The goal of simple linear regression is to create a linear model that minimizes the sum of squares of the residuals / error (SSE).

If our regression model is significant, it will "eat up" much of the raw SSE we had when we assumed (like this problem) that the independent variable did not even exist. The regression line will/should literally "fit" the data better. It will minimize the residuals.

The Legal Stuff...

Author's Note on the Legal Stuff

As a creator I need to put the legal language in this document. I wish I didn't have to. On a personal note, as you can imagine, I spend an immense amount of time, money, and sanity creating videos and this ancillary content. Some videos can take up to 20 hours to research, design, record, edit, and maintain. I also pay to have each video professionally captioned so you can enjoy it online and in this transcription document.

Therefore all I ask is that you respect the time, effort, and resources I put into creating this content for learners like you by keeping it to yourself for your own individual use. You should be able to print any part of the document and take notes using a tablet / stylus and software such as Drawboard PDF (Windows), PDF Viewer, PDF Expert (iPad) and others.

Transcription ReadMe

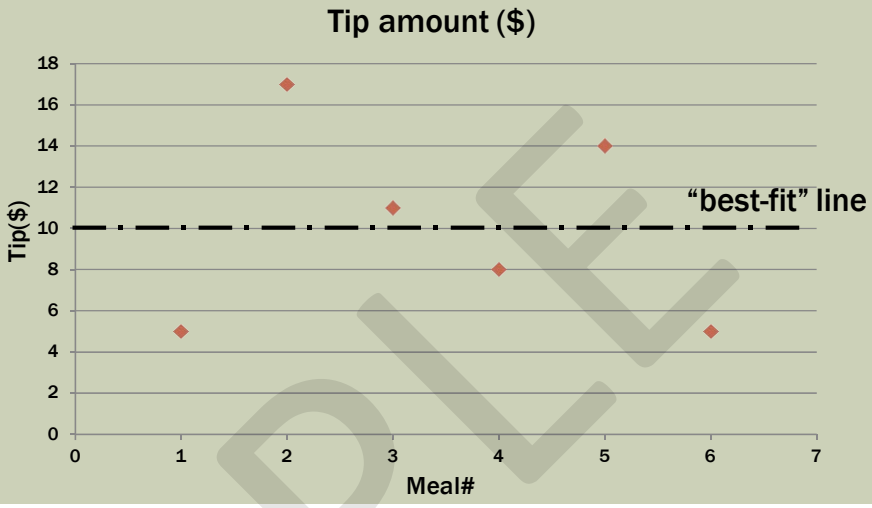
The notes pages contain the video transcript. These transcripts are professionally prepared by Rev.com and they do an outstanding job. A few things to note however. First, Rev.com transcribes everything I say **literally**. This means that my way of speaking is captured in the text. Therefore "because" is likely ['cause] and "kind of" is ['kinda] and things like that. Secondly, Rev.com does the best it can with the math notation; you can only imagine how hard it must be transcribing that notation. Therefore x_i is likely literally transcribed as "x sub i" and so on and so forth. And finally, some of my videos had sponsors that made those videos possible. I have left those slides in as a way of saying thank you. So please patronize those sponsors if possible.

Thank for your support and never stop learning. Never. - Brandon

TIPS FOR SERVICE

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

y $10



5:47
https://youtu.be/ZkjP5RJLQF4?t=347

So the first thing we're gonna do is we're going to visualize our data. As you know if you watch my other videos, I am a huge advocate of visualizing our problems, making charts, graphs, diagrams, whatever we have to do to make them visual. So the first thing we'll do is we'll make a graph of our tips. Now on the x-axis on the bottom, we have our meal number. Now that's not a variable, that's just a descriptor of what meal we're graphing. Now on the y-axis, or the vertical axis, that's where we will graph our tip amount. Let's go ahead and see what this looks like. So for meal one, with a tip of $5, so we'll go ahead and graph that at around $5. For meal two, with a tip of $17, so that goes way up there. For meal three, with a tip of $11, so that goes there. Meal four, with a tip of $8, that goes there. Meal five, that was a $14 tip. And meal number six, that was a $5 tip. So here are our data points. Remember, we're only dealing with one variable, that's the tip amount, and the meals along the bottom just describe where we're graphing each point. And the order does not matter. We could have graphed these in any order. This just happens to be the one we ended up with. Now, what's really the most you can figure out about this data? How would you predict what the tip for meal number seven would be? Is it going to be like meal number six, it's $5? Is it gonna be like meal number two, to $17? How would you come up with the best guess or estimate for the next meal using only one variable? Well, you would use its mean. So the mean for all six tips is $10. So guess what? That's the best we can do. With only one variable, the best estimate for the best prediction, for any given meal tip is $10. So go ahead and put a line at $10. So that for this model, that is our best fit line, that's

all we have. One variable, tip amount, the mean is the best predictor of any given tip amount. Now obviously if you look at this chart, our tips do not fall on the $10 line, they're scattered around it. But still, it's the mean. So that's your best estimate for the next tip for any given tip would be.

TIPS FOR SERVICE

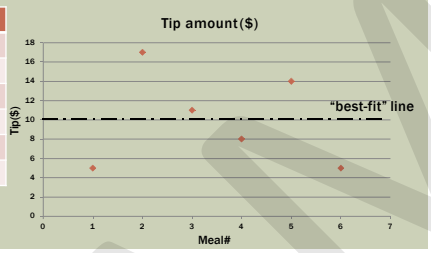
Unfortunately when you begin to look at your data, you realize you only collected data for the tip amount and not the meal amount also! So this is the best data you have.

How might you predict the tip amount for future meals using only this data?

TIPS FOR SERVICE

Meal#	Tip amount($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

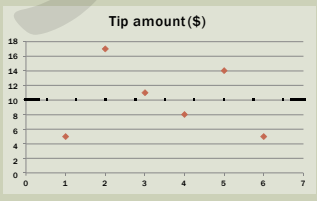
y $10



TIPS FOR SERVICE

Meal#	Tip amount($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

y $10



y $10

With only one variable, and no other information, the best prediction for the next measurement is the mean of the sample itself. The variability in the tip amounts can only be explained by the tips themselves.