TutorTube: Simple Linear Regression in StatCrunch

Spring 2020

Introduction

Hello! In this video I will be going through the steps involved in solving a typical linear regression problem using StatCrunch. Given a data set, we will draw a scatter diagram and then find the correlation coefficient, the critical value for r, and the equation of the regression line. Then, using our regression line, we will find the residual for a given value.

So, let's say you have a problem like this:

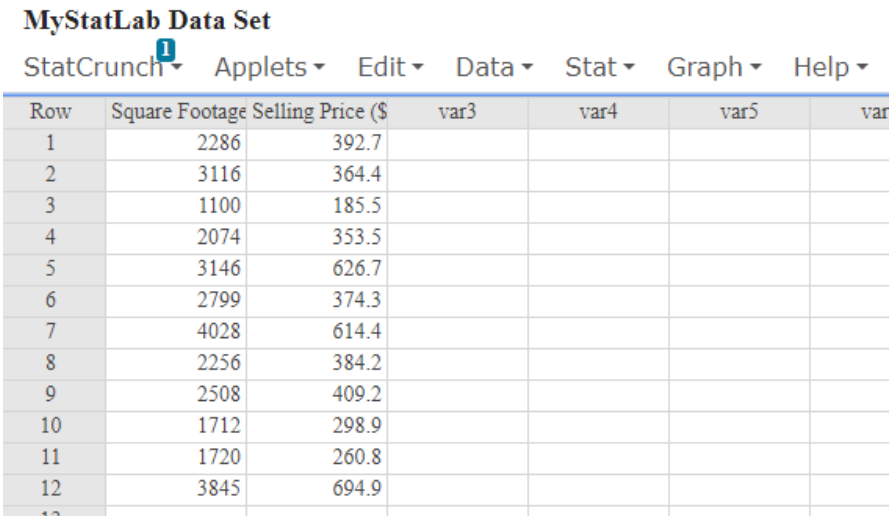
"The accompanying data represent the square footage and selling price (in thousands of dollars) for a random sample of homes for sale in a certain region."

Square Footage, x	Selling Price ($000s), y
2286	392.7
3116	364.4
1100	185.5
2074	353.5
3146	626.7
2799	374.3
4028	614.4
2256	384.2
2508	409.2
1712	298.9
1720	260.8
3845	694.9

Figure 1: Problem Data (Pearson, 2020)

Scatter Diagram

Part a) says "Draw a scatter diagram of the data." The first thing we do is open up our data in StatCrunch.



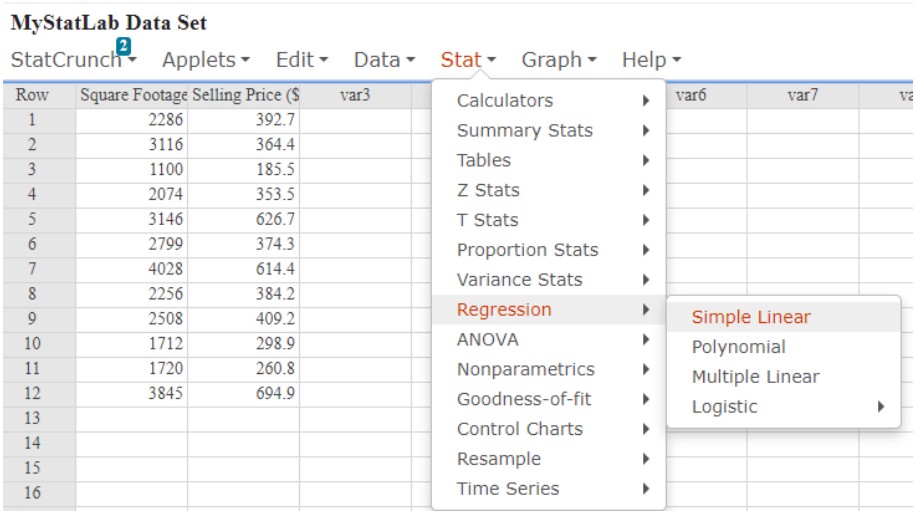
MyStatLab Data Set

StatCrunch Applets Edit Data Stat Graph Help

Row	Square Footage	Selling Price ($	var3	var4	var5	var
1	2286	392.7				
2	3116	364.4				
3	1100	185.5				
4	2074	353.5				
5	3146	626.7				
6	2799	374.3				
7	4028	614.4				
8	2256	384.2				
9	2508	409.2				
10	1712	298.9				
11	1720	260.8				
12	3845	694.9				

Figure 2: Data in StatCrunch

We can make a scatter plot using the "Regression" function under "Stat -> Regression -> Simple Linear".



MyStatLab Data Set

StatCrunch Applets Edit Data Stat Graph Help

Row	Square Footage	Selling Price ($	var3	var6	var7
1	2286	392.7			
2	3116	364.4			
3	1100	185.5			
4	2074	353.5			
5	3146	626.7			
6	2799	374.3			
7	4028	614.4			
8	2256	384.2			
9	2508	409.2			
10	1712	298.9			
11	1720	260.8			
12	3845	694.9			
13					
14					
15					
16					

Calculators
Summary Stats
Tables
Z Stats
T Stats
Proportion Stats
Variance Stats
Regression
ANOVA
Nonparametrics
Goodness-of-fit
Control Charts
Resample
Time Series

Simple Linear
Polynomial
Multiple Linear
Logistic

Figure 3: Stat > Regression > Simple Linear

Next, we select our X and Y variables. In this question, we are interested in predicting the price of a home using its square footage. This means that **Square Footage will be our explanatory or independent variable**, also known as X. Since we are trying to predict the Selling Price, **Selling Price will be our dependent or response variable**, which is our Y.

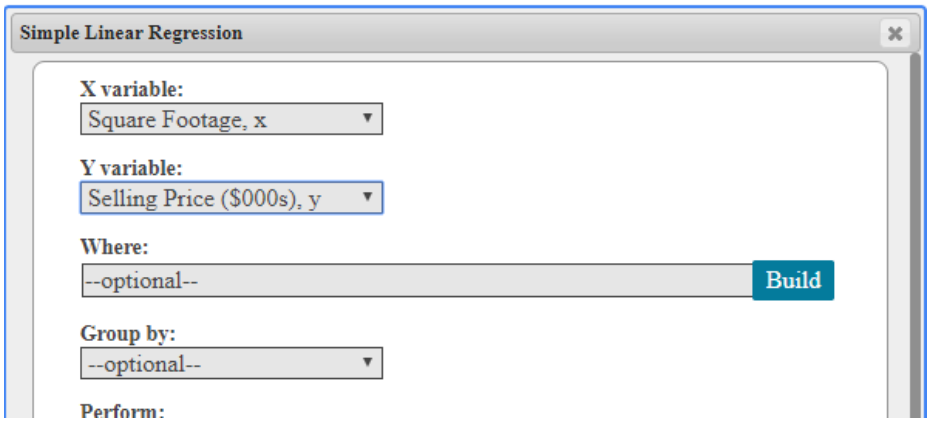


Figure 4: The Regression Window

For now, we don't have to change anything else. So just hit "Compute!"

And we have our results window:

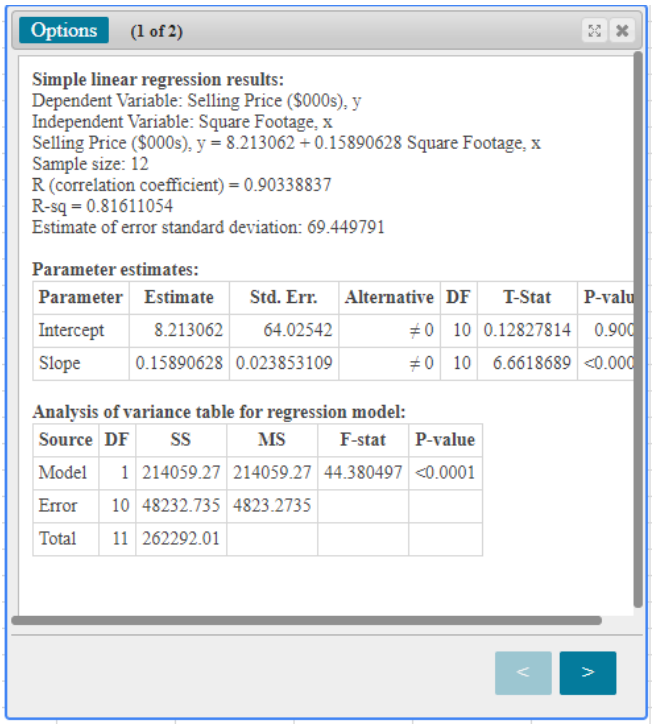


Figure 5: Regression Output Window

To find our scatter plot, click the arrow on the bottom of the results box. And there we go, we have a scatter plot with our regression line running through it.

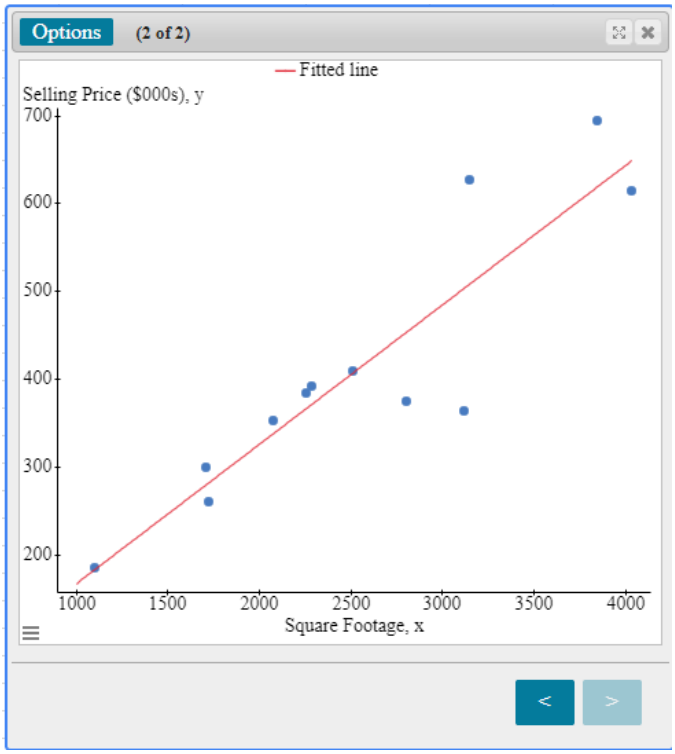


Figure 6: Scatter Plot

Correlation Coefficient, r

Next, for part b), Find the correlation coefficient, **r**. Basically, we are being asked how strongly these two variables are related.

In order to find r, we click on the left arrow to go back to the original results box. The correlation coefficient R is the 5th item down. For this data set, **r is about .903**. Be sure to pay attention to how the question wants you to round your answer.

Is There a Linear Relationship?

Next we want to know for part c): "Is there a linear relationship between these variables?" Is square footage actually a good predictor for the selling price of a house?

In order to figure this out, we need to compare the absolute value of our r to the critical value for this sample size.

First let's find the critical value. If you are using MyStatLab, you will be given a table of Critical Values for the Correlation Coefficient that looks something like this:

Critical Values for Correlation Coefficient	
n	
3	0.997
4	0.950
5	0.878
6	0.811
7	0.754
8	0.707
9	0.666
10	0.632
11	0.602
12	0.576
13	0.553
14	0.532
15	0.514
16	0.497
17	0.482
18	0.468
19	0.456
20	0.444
21	0.433
22	0.423
23	0.413
24	0.404
25	0.396
26	0.388
27	0.381

Figure 7: Critical Values or Correlation Coefficient (Pearson, 2020)

In order to find our critical value, we need to know how many observations are in our data set. In this case we had 12 observations. 12 houses. This is our "n" in this new table. The critical value associated with 12 is .576.

Next, we take the absolute value of our correlation coefficient. |r| is |.903| is just .903.

Now we compare the absolute value of r to the critical value. Since **.903 is bigger than .576**, we say that there is in fact a **linear relationship** between these variables.

The Least-Squares Regression Line

For part d), we want to find the least-squares regression line, treating square footage as the explanatory variable. Since we know that there is a linear relationship between these two variables, it makes sense to find a linear regression line for them.

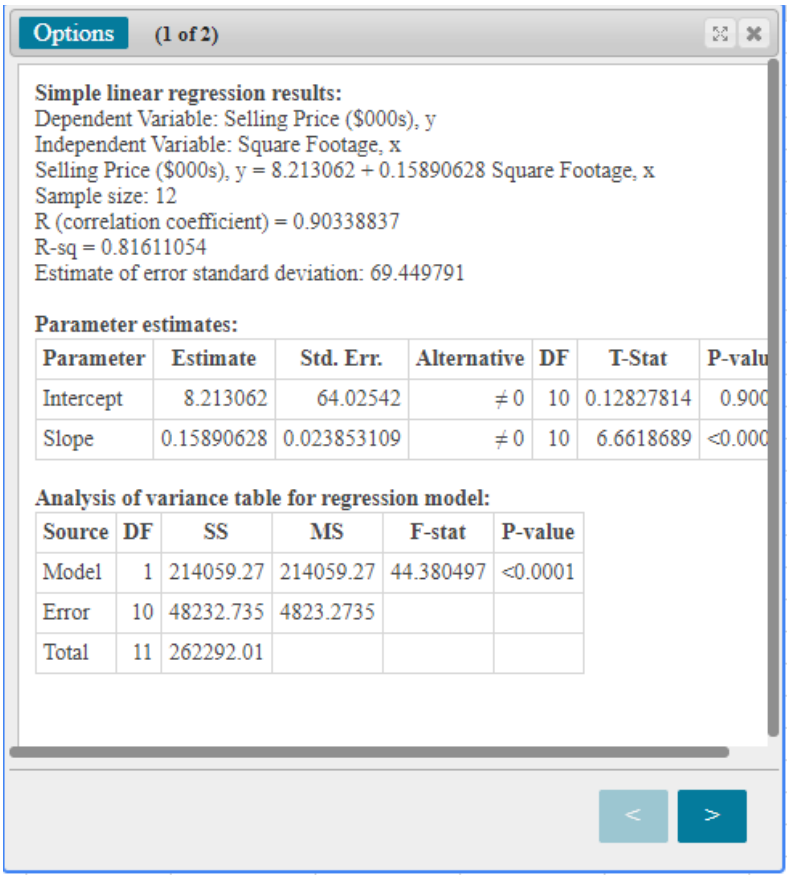


Figure 8: Output Window

This is also found in the first Results screen. It is the third item down. Something to be aware of: Some homework might want you to report the line in the form $\hat{y} = mx + b$, but StatCrunch gives you the line in the form $y = b + mx$. So be sure to keep your slope and y intercept straight when you are inputting answers.

Here for example, if we wanted the line reported in $y = mx + b$ form rounded to 3 decimals, we would write:

$$\hat{y} = .159x + 8.213$$

Interpret the Slope of the Regression Line

Next, we are asked to interpret the slope of the regression line. Many problems will ask you to also "interpret" the slope of your line. There is a format to follow when answering. The basic format for a positive slope is:

For every 1 (units of x) increase in (x-variable), we can expect a (slope) (units of y) increase in (y-variable).

So, for our example, our answer would be:

"For every 1 square foot increase in the Square Footage of a house, we can expect a .159 thousand dollar increase in the Selling Price of the house."

Predict a Value using the Linear Regression Equation

Ok now, we want to know "What is the predicted selling price of a home that is 1433 square feet?" So, in order to find the predicted value by hand, we would have to plug the value 1433 into our regression equation and solve for y. However, StatCrunch will do this for us. Go to "Options -> Edit" to get back to our Simple Linear Regression menu.

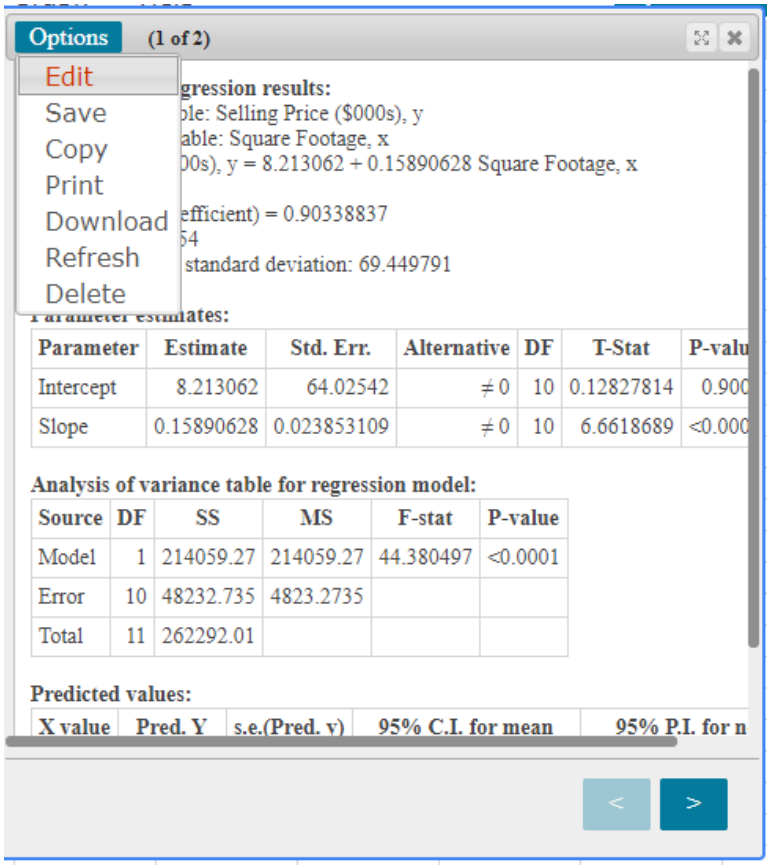


Figure 9: Use Options > Edit to return to the Regression window

And then now we find the section labeled: "Prediction of Y:" In the "X value(s):" box, enter our value of 1433.

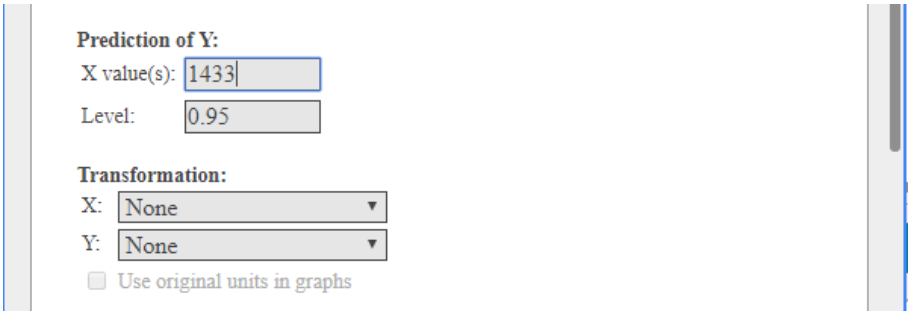
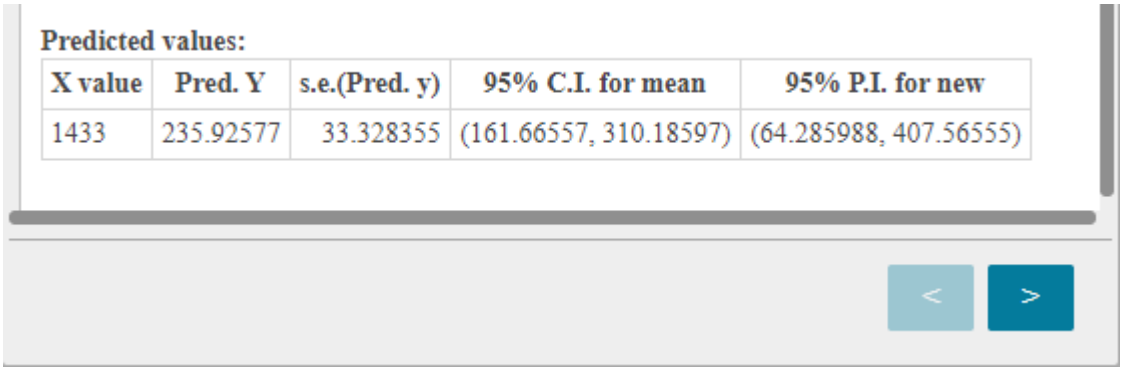


Figure 10: Enter X value(s) under Prediction of Y

Then click "Compute!" again. Now on our Results page, we have a new section at the bottom labeled "Predicted values:" The value we want for this question is listed under "Pred. Y" Which is about **235.9 (or 236)**.



Predicted values:

X value	Pred. Y	s.e.(Pred. y)	95% C.I. for mean	95% P.I. for new
1433	235.92577	33.328355	(161.66557, 310.18597)	(64.285988, 407.56555)

Figure 11: Output Window for Predicted Value

So for our question, **the predicted selling price for a house that is 1433 square feet is about 236 thousand dollars**.

Find the Residual for a Given Value

Next for our final part, "If we observe a 1433 square foot home selling for $210 thousand, what is the residual of this value?" Recall that residuals are found by subtracting the predicted value from the observed value. In other words:

Residual = Observed – Predicted

We found the predicted value in the previous question. It was 236 thousand. We know our observed value is 210 thousand. So our residual for this observation is:

210 – 236 = **-26 thousand**.

Outro

Alright, so we have answered all parts of our question. I hope you found this video helpful.

If you are a UNT student, there will be some links to other resources in the video description. Thank you for watching!

References

Pearson. (2020). MyLab: Statistics. *Pearson Higher Education Inc.*

*All calculations in this video were performed with Pearson StatCrunch 2020 software.