

A. Overview

Idea: given a set of data, the regression method summarizes the relationship between two variables X and Y by choosing a line that fits the points as closely as possible. Here, there is a strong sense that one of the variables (Y) **depends** on the other (X), contrast this with Chapters 8 and 9 where it was not necessary to designate one variable as Y and the other as X.

This line which summarizes the relationship can be found from just knowing the means and standard deviations of the two variable s and their correlation.

This line is called a "regression" line because the man who developed this statistical method was working with father's heights and the heights of their sons and noticed something he called "regression to the mean (or average)". Basically, he noticed that tall fathers often had sons who were average in height and short fathers often had sons who were average in height.

And that is what the regression line does -- it is going through the average of the Y-variable (the vertical axis) for a given value of the X-variable (the horizontal axis).

You could think of this line as the "best fitting" line between all of the points on the graph. Note that the line doesn't exactly pass through the points -- in fact, it misses most of them, but the average distances of the points from the line are minimized.

B. INTRODUCTION TO REGRESSION (Chapter 10.1-10.3)

Here we have some data (information) on Broadway shows, their average receipts per show and the capacity of the theater. A question we might be interested in is what is the relationship between capacity and receipts? Do bigger theaters always generate more receipts?

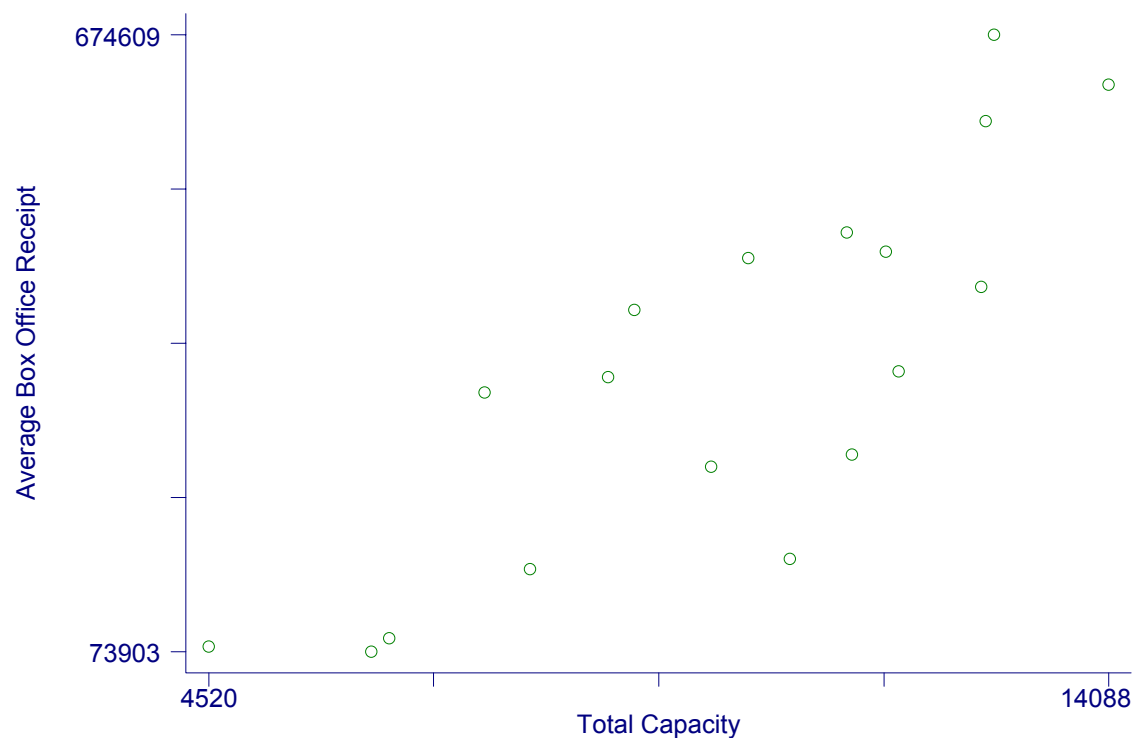
	show	receipts	capacity
1.	Angels in America	326121	7456
2.	Blood Brothers	154064	7936
3.	Cats	346723	11856
4.	Crazy for You	463377	11720
5.	Falsettos	86864	6440
6.	Fool Moon	163802	10696
7.	The Goodbye Girl	429158	12736
8.	Guys and Dolls	457087	10256
9.	Jelly's Last Jam	253951	9864
10.	Kiss of the Spider Woman	406498	9048
11.	Les Miserables	481973	11304
12.	Miss Saigon	625804	14088
13.	Phantom of the Opera	674609	12872
14.	Shakespeare for my Father	78898	4520
15.	The Sisters Rosensweig	340862	8768
16.	Someone Who'll Watch over Me	73903	6248
17.	Tommy	590334	12784
18.	The Will Rogers Follies	265561	11360

The correlation r of theater capacity and receipts is .82.

The averages and standard deviations of capacity and receipts is summarized in this table below:

Variable	Obs	Mean	Std. Dev.	Min	Max
receipts	18	345532.7	187723.8	73903	674609
capacity	18	9997.333	2664.093	4520	14088

And here is a graph of the two variables:



From the previous lecture: one of the properties of the correlation coefficient (r) is that it can be used to give us a rough idea of the relationship between two variables.

Values of r near $+1$ and -1 mean that the two variables are very closely associated. Values of r near 0 suggest that there is almost no relationship between the two. If you look the graph again, r is $=.82$ which is positive relationship between the variables and it's reasonably strong.

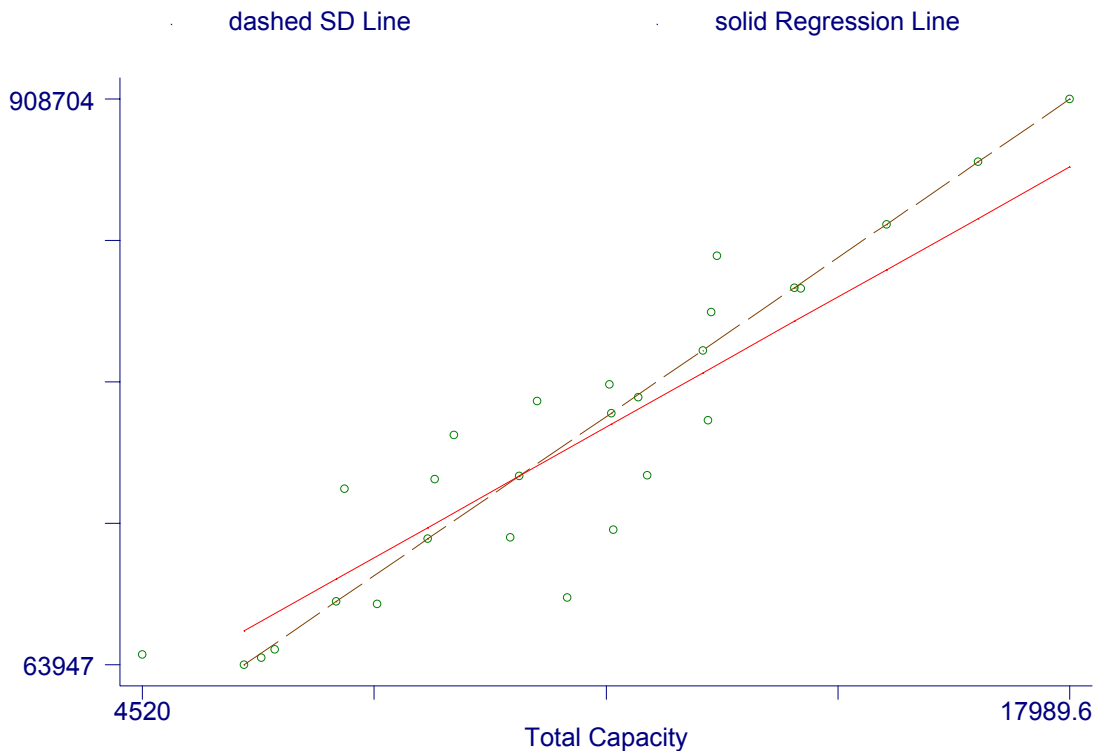
One method of using all of the information is to predict what the receipts will be for a given level of capacity. This method is called the REGRESSION METHOD. You can predict values of your y variable by knowing the averages and the standard deviations and the correlation of both X and Y :

$$(\text{Average theater capacity}) + (\text{one SD of theater capacity}) = 9997.333 + 2664.093 = 12661.426$$

So a play that is one SD above the average theater capacity has a theater capacity of about 12,661. Since these two variables (capacity and receipts) appear to be related (a correlation of $.82$ suggests some relationship) we can ask -- what are the receipts for a play that has above average (one standard deviation above average) theater capacity?

We can quickly answer this by taking the receipt SD of 187723.8, multiply by $.82$ to get 153933.52 and add this result back to the average receipt of 345532.7 to get $345532.7 + 153933.52 = 499466.22$ or near 500,000 dollars.

If you did this method for all possible theater capacity SDs (one, two, three and the numbers in between, you'd get a solid line that looks like:



and you could compare it to an SD line (which is the equivalent of a correlation = 1.0)

This is the regression line of receipts on theater capacity. The line actually goes through the "points of averages" (page 162) that is, for each value of theater capacity, the regression line is going through a best fitting line, a smoothed line, of the average receipts for a given level of theater capacity.

From your text, the regression line of the y-variable (receipts) on the x-variable (theater capacity) is giving an average y-value (average receipts) for each value of x. (page 160)

Freedman summarizes the method by stating that for each one standard deviation increase in X, there is a $r \times$ (standard deviation increase in y). The change in Y in response to X is not a one to one increase, it is not an increase of the correlation either, it's a combination of the correlation between X and Y and the standard deviation of y.

C. REGRESSION FOR INDIVIDUALS (10.3) -- PREDICTION

Chapter 10.3 has an easier way to construct this line and allows you to work on the prediction of individual values of Y given the value of X.

Back to the example above:

The correlation r of receipts and theater capacity is .82

The average receipts is 345,532.70 and the standard deviation is 187,723.80

The average theater capacity is 9997.333 and the standard deviation is 2664.093

A new play comes out and has a theater capacity of 6000. Predict the receipts for that play.

The play is $(6000 - 9997.333) / 2664.093 = -1.50$ standard deviations below the average theater capacity (the X variable) (recognize the formula? It's Z)

If we take the -1.50 standard deviations and multiply it by the correlation .82 we can relate it to the Y variable (receipts) and get a -1.23 standard deviations for Y.

We can now give a prediction for receipts by multiplying -1.23 by the standard deviation for receipts of 187,723.80:

$$-1.23 * 187,723.80 = -230,900.27$$

and adding this result to the average total receipts $(345,532.70 + -230,900.27) = \$114,632.43$ or so.

So if a new play has a capacity of 6000, given what we know about capacity's relationship to receipts, we predict (or expect) receipts of about 114,632.43

4. USING THE Z SCORE AND REGRESSION INFORMATION (10.3)

Notice we used a Z score above. We can do something similar to Chapter 5.

Suppose a play is in the highest 10% of theater capacity.

If you are in the top 10% of ANYTHING, this means you have a $Z = 1.3$ or you are 1.3 standard deviations above average (remember, a Z score could be thought of as just the number of standard deviations you are above or below the average, Z's are called "standard units").

If a play is in the top 10% of theater capacity, it's capacity is calculated to be $9997.333 + (1.3 * 2664.093) = 13460.654$

And so, its corresponding total receipts will be will be:

First find the product of the Standard Deviations of capacity (X) and the correlation:

$$1.3 * .82 = 1.066$$

Second, multiply this result by the Standard Deviation of receipts (Y):

$$1.066 * 187,723.80 = 200,113.57$$

Finally add this value to the average total receipts (Y)

$$200,113.57 + 345,532.70 = 545,646.27$$

Therefore, if your play is in the top 10% of theater capacity, the corresponding receipts is predicted to be 545,646.27

5. SUMMARY

In combination with a basic understanding of regression, the correlation coefficient can give us a sense of the amount of change that occurs between two variables. It can also be used to predict values of Y given a value of X.