A. Overview

Given some 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 Ch. 7 where either could be X or Y.

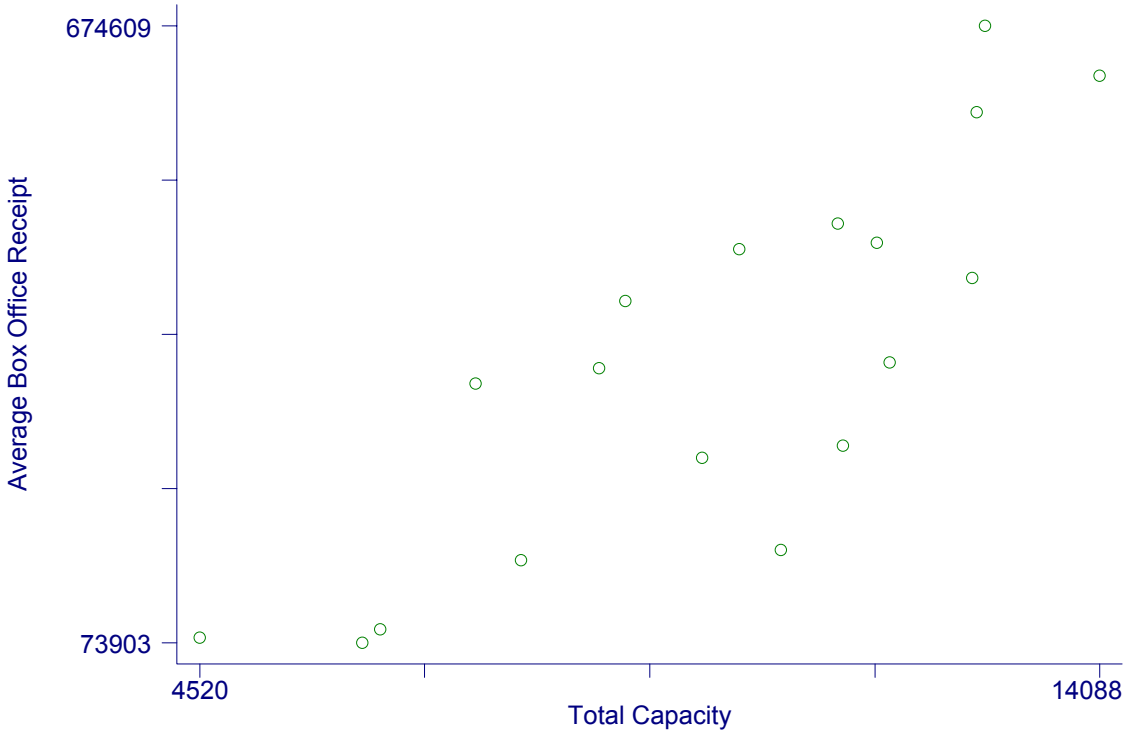
B. INTRODUCTION TO REGRESSION (Chapter 8)

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 means and SDs are below:

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



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.

C. WORKING WITH THE REGRESSION LINE AS Z SCORES

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

Notice we used a Z score above. We can do something similar to Chapter 6. Suppose a play is in the highest 10% of theater capacity and suppose theater capacity is normally distributed. Recall that if f you are in the top 10% of ANYTHING, this means you have a Z = 1.28 or you are 1.28 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, its capacity is calculated to be 9997.333 + (1.28 * 2664.093) = 13460.654

And so, its corresponding total receipts will be will be:
(1) find the product of the Standard Deviations of capacity (X) and the correlation: 1.28* .82= 1.05
(2) multiply this result by the Standard Deviation of receipts (Y): 1.05 * 187,723.80= 197,109.99
(3) add this value to the average total receipts (Y) 197,109.99 + 345,532.70 = 542, 642.69

Therefore, if your play is in the top 10% of theater capacity, the corresponding receipts is predicted to be 542,642.69

D. WORKING WITH REGRESSION IN REAL (NOT Z) UNITS

(1) It's equation is:

y = (slope*x) + intercept OR y = bx + a OR y = mx + b OR $y = b_0 + b_1x$

No straight line will pass exactly through all of the points. A fitted line comes as close as possible to all of the points simultaneously. The assumption being made here is that y is dependent on x or y is the response, dependent, or outcome variable, x is the explanatory, independent or predictor variable.

(2) Calculating the slope (b) of the line
the formula is

b = r * ((SD of Y)/(SD of X)) (see page 140-141)

The slope b measures the average observed change in Y when X changes by one unit. It is thought of as a rate of change. For our example on capacity and theater receipts:

b = .8214 * 187,723.8/2664.093 = 57.8794

Note that the correlation, r, is determining the sign of the slope b. If r were equal to 1 or negative 1, the X and Y variables would be changing at a similar rate. In this example, y is changing at a slower rate than x.

(2) Calculating the intercept: a (page 140-141)

a = (average of the y variable) - b(average of the x variable) or -233106.94=345532.70-(57.8794*9997.333)

This is the value of y when x=0. And note that the line then always crosses through the point represented by the means of x and of y. This is a check if you are calculating a regression with a hand calculator and not a computer.

Note: you must have calculated slope before you can apply this formula to calculate intercept.

(3) Put it all together in a regression equation

y = (slope*x)+ intercept so receipts = 57.8794(capacity) –233,106.94

2. Using the Regression Line

Prediction: The prediction equation for theater receipts is:
Receipts = -233106.94 + 57.8794(capacity)

Interpretation: slope tells you how much change on average to expect in Y if X is changing. Intercept tells you what Y would be if X were equal to zero (sometimes this is nonsense). Most applications of regression are interested in slope. In our example, a one seat change results in a 57.8794 dollar change in receipts. If a theater were able to add 1000 seats then we would expect a 57,879 dollar increase in receipts. For the intercept, if a theater had zero capacity, that is X=0, the model predicts that there would be a loss of 233,106.94 (negative receipts). It is nonsensical to talk about a theater having no capacity so the slope is more interesting to a researcher/business person in this particular situation.

Extrapolation (shouldn't be attempted). If someone were interested in building a new theater with 20,000 seats the regression equation predicts receipts of $924,481.06. This is an example of using a regression line to predict values outside of the range that we have. These predictions are not usually accurate and this should not be done.