1. 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 variables 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).

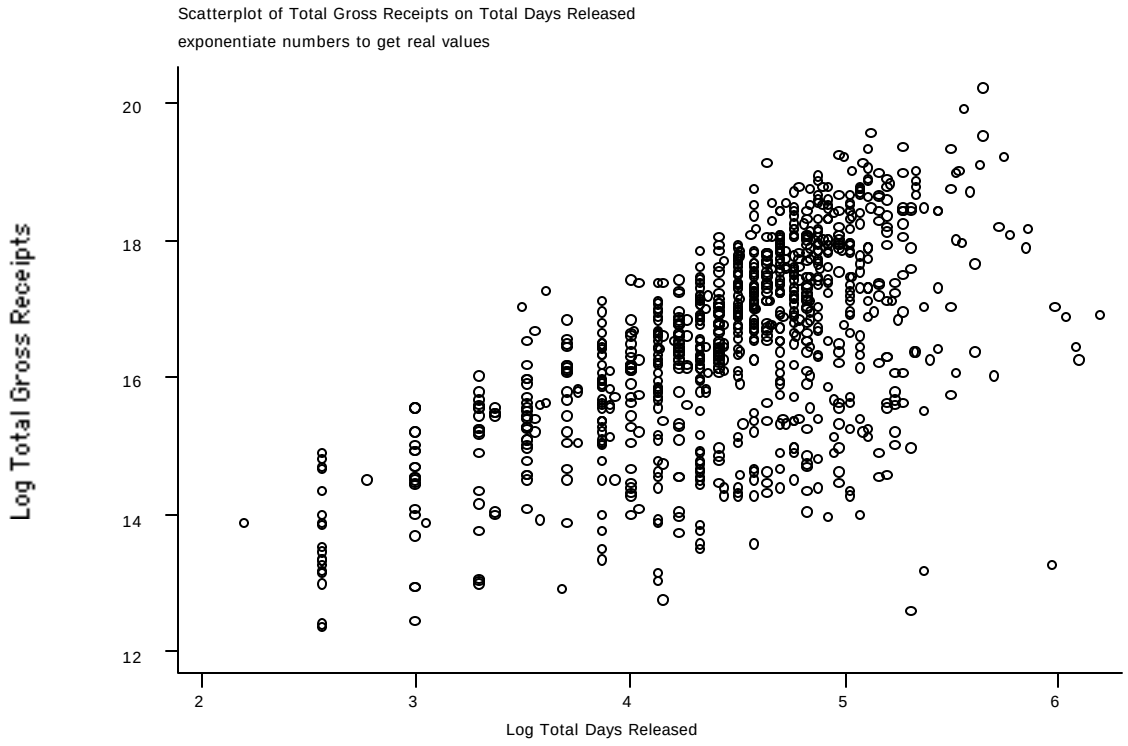
2. Introduction to Regression (10.1, 10.2)

The correlation r of days a movie has been released and the total gross is .60. The average gross receipts is about 16.5 and the standard deviation is 1.4. The average days in release is 4.5 and the standard deviation is about .64

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 =.60 which is positive relationship between the variables and it's not particularly strong or weak.



One method of using all of the information is to predict what the total gross receipts will be for a given level of days in release. 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:

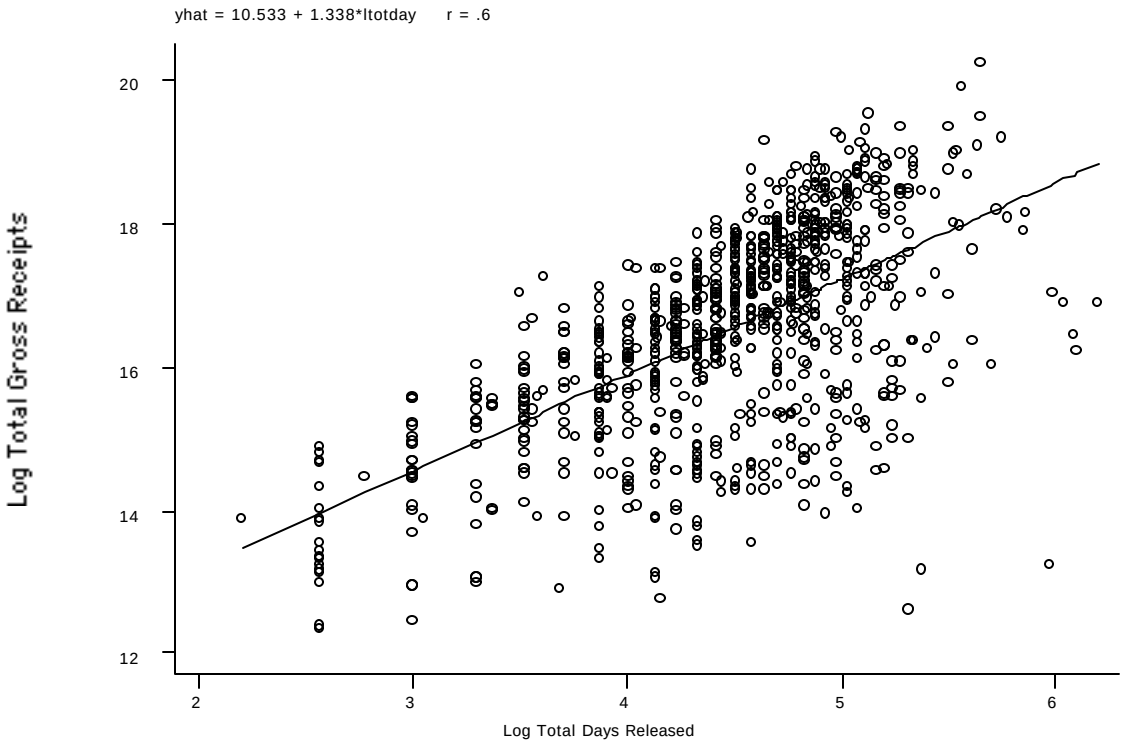
(Average days in release) + (one SD of days in release) = 4.47 + .64 = 5.14

So a movie that is one SD above the average days in release has a days in release of 5.14.

Since these two variables appear to be related (a correlation of .60 suggests some relationship) we can ask -- what is the total gross receipts for a movie that has above average (one standard deviation above average) days in release?

We can quickly answer this by taking the total gross receipt SD of 1.42, multiply by .60 to get .85 and add this result back to the average gross receipt of 16.5 to get 16.5 + .85 = 17.35

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



This is the regression line of total gross receipts on total days released. The line actually goes through the "points of averages" (page 162) that is, for each value of total days in release, the regression line is going through a best fitting line, a smoothed line, of the average of total gross receipts for that value of days in release.

From your text, the regression line of the y-variable (total gross receipts) on the x-variable (total days released) is giving an average y-value (average gross 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*(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.

3. 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 total gross receipts and total days released is .60

The average total gross receipts is 16.5 and the standard deviation is 1.42
The average total days released is 4.47 and the standard deviation is .64

A movie is chosen at random and has a total days released of 3.5. Predict the total gross receipts for that movie.

The movie is (3.5-4.47)/.64 = -1.52 standard deviations below the average total days released (the X variable) (recognize the formula? It's Z)

If we take the –1.52 standard deviations and multiply it by the correlation .60 we can relate it to the Y variable (total gross receipts) and get a -.912 standard deviations for Y. We can now give a prediction the total gross receipts by multiplying -.912 by the standard deviation for total gross receipts of 1.42:

-.912 * 1.42 = -1.295

and adding this result to the average total gross receipts (16.5 + -1.295) = 15.20 or so.

So if a movie has a release time of 3.5, given what we know about its relationship to the total gross receipts, we predict receipts of about 15.20.

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 movie is in the highest 10% of total days released.

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.

If a movie is in the top 10% of days in release, it's days in release is calculated to be 4.47 + (1.3 * .64) = 5.30

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

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

1.3* .60 = .78

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

.78 * 1.42 = 1.108

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

16.5 + 1.108 = 17.608

If your movie is in the top 10% of days released, the corresponding total gross receipts is predicted to be 17.608.

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.