A. Overview

Last time we looked at theater capacity (X) and receipts (Y)

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

B. 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

C. 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.

D. Residuals

Here's the original data

	show	receipts	capacity	predic~d	residual
1.	Angels in America	326121	7456	198441.9	127679.1
2.	Blood Brothers	154064	7936	226224	-72159.98
3.	Cats	346723	11856	453111.2	-106388.2
4.	Crazy for You	463377	11720	445239.6	18137.38
5.	Falsettos	86864	6440	139636.4	-52772.39
6.	Fool Moon	163802	10696	385971.1	-222169.1
7.	The Goodbye Girl	429158	12736	504045.1	-74887.09
8.	Guys and Dolls	457087	10256	360504.2	96582.81
9.	Jelly's Last Jam	253951	9864	337815.5	-83864.47
10.	Kiss of the Spider Woman	406498	9048	290585.9	115912.1
11.	Les Miserables	481973	11304	421161.8	60811.19
12.	Miss Saigon	625804	14088	582298.1	43505.94
13.	Phantom of the Opera	674609	12872	511916.7	162692.3
14.	Shakespeare for my Father	78898	4520	28507.95	50390.05
15.	The Sisters Rosensweig	340862	8768	274379.6	66482.38
16.	Someone Who'll Watch over Me	73903	6248	128523.6	-54620.55
17.	Tommy	590334	12784	506823.3	83510.69
18.	The Will Rogers Follies	265561	11360	424403	-158842

Consider our regression equation

Receipts = -233106.94 + 57.8794(capacity)

A residual is equal to the original data value – predicted value. So for example, Angels in America has a predicted value of

-233106.94 + 57.8794(7456) = -233106.94 + 431548.81 = 198441.9

But it's actual receipts were 326,121.

The residual is 326,121-198,441.9 = 127,679.1

The analysis of residuals is important because the determination of how well a model fits depends upon the size of the residuals. When a residual is positive (as in our worked out example) the model is underestimating the value, when a residual is negative, the model overestimates the value.

E. R2 or R-squared

You don't need to know this for the final. It's an important measure, just know the verbal definition, it's the proportion or the percentage of variation accounted for by having knowledge of the X variable (capacity) when trying to predict Y (receipts).

F. Know your Assumptions

Essentially, it must be reasonable to fit a straight line to your scatter diagram of data.

You should not have outliers present in the scatter diagram.

You should not extrapolate beyond the available ranges of your data

Consider the possibility that there may other X variables that predict your Y variable better.