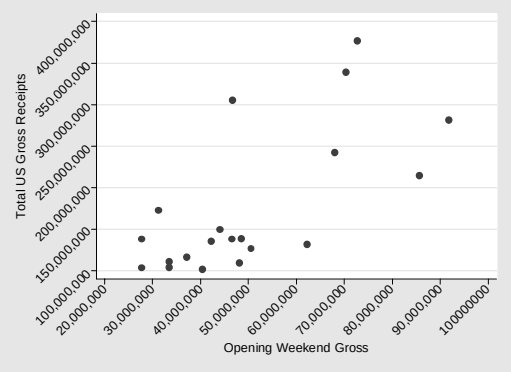
Announcements

- Unclaimed 1st & 2nd exams & homeworks (1, 2, 3, & 4) are up front
- Extra Office Hours This Week
  - Today 2:30pm-3:30pm
- Old Handouts & 1st Day Survey are up front
- One Lecture Handout Today
- Review Material Handout Today

Recall Last Time We Examined Movie Ticket Sales



r = .68

- Means and Standard Deviations are:

Variable	Obs	Mean	Std. Dev.	Min	Max
totusgross	20	177342928	85658496	102543518	377027325
weekend1	20	50361404	18775706	27557647	91774413

Working with Linear Regression Using Real Units (and not Z scores)

We can predict the total US gross receipts for this new movie using real units. We must still assume that X and Y are normal to do this.

- The line will no longer pass through the origin (0,0), so there is a slope and an **interpretable** intercept (different versions of the same line):

y = (slope*x) + intercept

also **y = bx + a**

also **y = mx + b**

also **$y = b_0 + b_1 x$**

First, Calculate the Slope of the Regression Line

- **(see page 140-141)** $b = r * \frac{SD_Y}{SD_X}$
- 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 opening weekend ticket sales and total US gross receipts:

$$b = .68 * \frac{85658496}{18775706} = 3.1023$$

- 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 faster rate (3.1023 times) than x.

Second, Calculate the Intercept of the Line

- **(see page 140-141)**
  a = (average of the y variable) – [b*(average of the x variable)]
  so
  177,342,928- (3.1023*50,361,404) = 21,106,744
- 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.

Construct The Regression Equation

- **y = (slope*x)+ intercept →**

total US gross receipts = (3.1023 *first weekend ticket sales) + 21,106,744

- Check that the equation is correct by plugging the mean of the X variable in:

total US gross receipts = (3.1023 *50,361,404) + 21,106,744

= 177,342,927.6 or 177,342,928 →

that's the mean of the Y variable

Using the Regression Equation

- The equation again:

total US gross receipts = 3.1023 (first weekend ticket sales) +21,106,744

- We can interpret the slope and intercept
  - For each additional dollar of first weekend sales (X variable) results in a **3.1023** dollar increase in total US gross receipts.
  - If first weekend sales were ZERO, we would expect the movie to earn a total of 21,106,744 in it's total run. (This result is nonsensical, which is why the Z score method is better at times)

Using the Regression Equation (cont'd)

- The equation again:

total US gross receipts = 3.1023 (first weekend ticket sales) +21,106,744

- We can predict total gross receipts from first weekend ticket sales. "National Treasure" had sales of $35,142,554
- Plug it in:

total US gross receipts = (3.1023*35,142,554) +21,106,744

=$130,129,489

INCORRECTLY USING THE Regression Equation

- The equation again:

total US gross receipts = 3.1023 (first weekend ticket sales) +21,106,744

- "Alexander" had first weekend sales of $13,687,087
- You could plug it in:

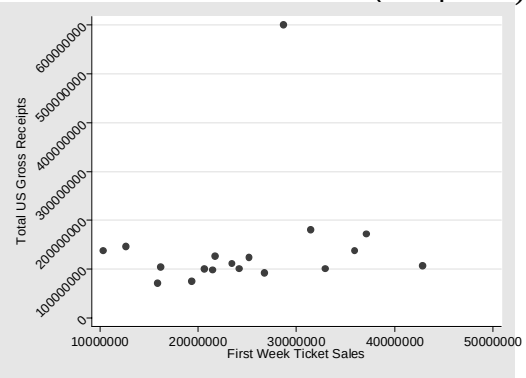
total US gross receipts = (3.1023*13687087) +21,106,744 = $63,568,194

- Problem: $13,687,087 is OUTSIDE of the range of the X values used to construct the equation, this is called EXTRAPOLATION (Chapter 9)and should not be done

Another Incorrect Use

- The movie "Seabiscuit" eventually had a total US gross receipts of $120,277,854
- Can we use the regression equation to predict it's first weekend sales?
- NO – the regression equation is built on the assumption that the X variable (first weekend) predicts the Y variable (total) and not the reverse

Outliers – Undue Influence (Chapter 9)



The extreme point is the movie "Titanic", r=.17 here if we remove it, r=.31. A single point can affect your regression line.

Residuals

Our original data with two additional columns

	title	totusgross	weekend1	predicted	residual
1.	The Lord of the Rings: The Return of the King	377,027,325	72,629,713	246,425,903	130,601,422
2.	Finding Nemo	339,714,978	70,251,710	239,048,624	100,666,354
3.	Pirates of the Caribbean	305,413,918	46,630,690	165,769,134	139,644,784
4.	The Matrix Reloaded	281,576,461	91,774,413	305,818,505	-24,242,044
5.	Bruce Almighty	242,829,261	67,953,330	231,918,360	10,910,901
6.	X2: X-Men United	214,949,694	85,558,731	286,535,595	-71,585,901
7.	Elf	173,398,518	31,113,501	117,630,158	55,768,360
8.	Terminator 3: Rise of the Machines	150,371,112	44,041,440	157,736,503	-7,365,391
9.	The Matrix Revolutions	139,270,910	48,475,154	171,491,214	-32,220,304
10.	Cheaper by the Dozen	138,614,544	27,557,647	106,598,832	32,015,712
11.	Bad Boys II	138,608,444	46,522,560	165,433,682	-26,825,238
12.	Anger Management	135,645,823	42,220,847	152,088,478	-16,442,655
13.	Hulk	132,177,234	62,128,420	213,847,741	-81,670,507
14.	2 Fast 2 Furious	127,154,901	50,472,480	177,687,519	-50,532,618
15.	S.W.A.T.	116,934,650	37,062,535	136,085,846	-19,151,196
16.	Spy Kids 3D: Game Over	111,761,982	33,417,739	124,778,596	-13,016,614
17.	Scary Movie 3	110,003,217	48,113,770	170,370,093	-60,366,876
18.	American Wedding	104,565,114	33,369,440	124,628,758	-20,063,644
19.	Daddy Day Care	104,297,061	27,623,580	106,803,376	-2,506,315
20.	Daredevil	102,543,518	40,310,419	146,161,757	-43,618,239

Residuals

•The equation again:

total US gross receipts = (3.1023*first weekend ticket sales) +21,106,744

A residual is equal to the original Y value – predicted Y value. So for example, the movie "Elf" has a predicted total gross receipts of

$117,630,158= (3.1023*31,113,501) +21,106,744

and a residual of

$173,398,518 – $117,630,158 = $55,768,360

So we "under-predicted" the total gross receipts of Elf based on information we had on it's first week ticket sales.

Why Residuals?

- The analysis of residuals is important because the determination of how well a model fits depends upon the residuals (this is advanced and outside of the scope of Stat 10). When a residual is positive (as in our worked out example) the model is underestimating the value of the Y variable, when a residual is negative, the model overestimates the value.

R^2 or R-Squared

- It's the proportion or the percentage of variation accounted for by having knowledge of the X variable when trying to predict Y.
- In our example, the r-squared is .68*.68=.46 which would be interpreted as "knowing the value of the first weekend ticket sales accounts for about 46% of the variation in total US gross receipts".

Know Your Assumptions (Chapter 9)

- Essentially, to use regression 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.