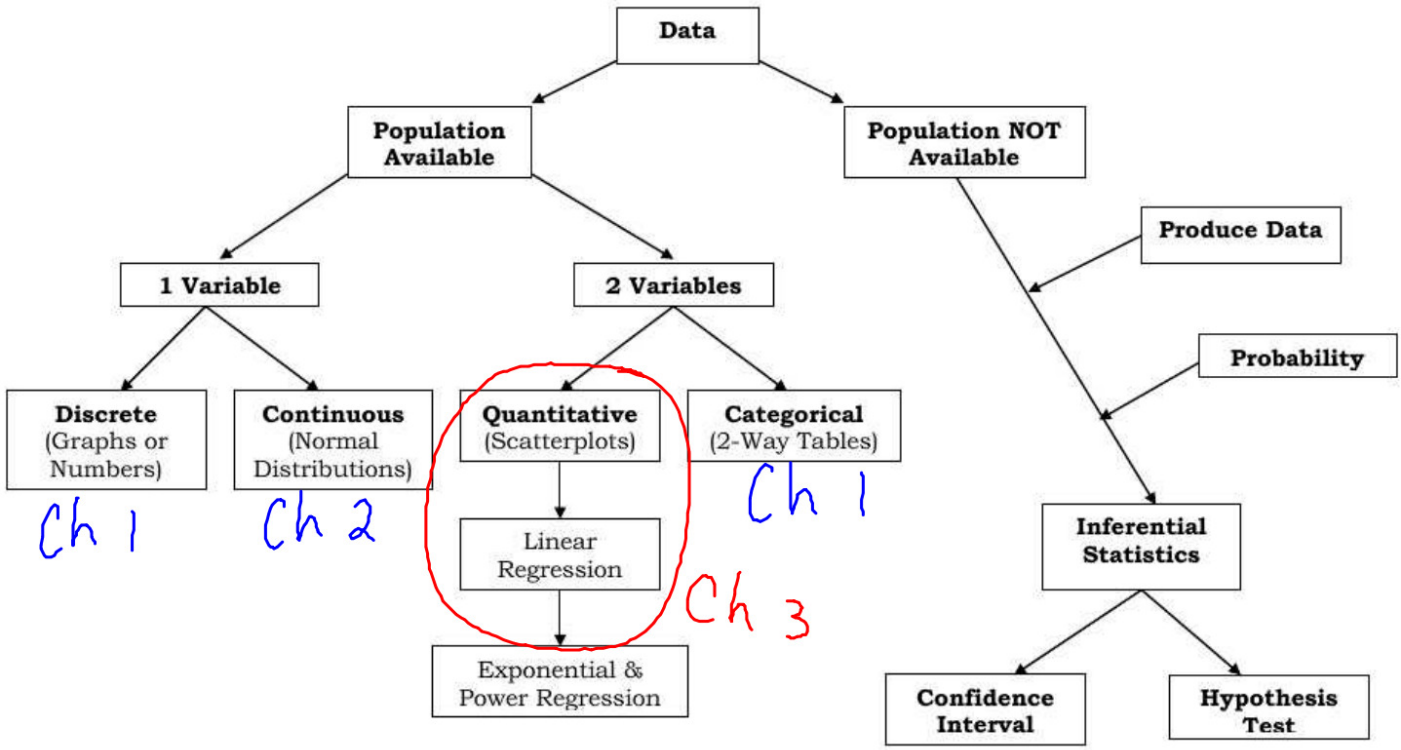
Sec 3.1

TYPES OF STATISTICS



Is there a linear relationship between 2 (x, y) quantitative variables ?

explanatory variable

response variable

★ If there is we can use x to predict y ★

Begin with Scatterplot

Response Variable

- Direction (Positive / Negative)
- Form (Linear, Curved, Clustered)
- Linear Strength (Weak, Moderate, Strong)
- Outliers?

Explanatory Variable

SAT Scores (P. 146)

Direction - Negative

Form - 2 Linear Clusters

Strength - Moderately strong

Outliers - (19, 501) and (87, 466) ?

Making Scatterplots (P. 145)

1) Determine explanatory/response variables

L_1 (x) $\longrightarrow$ L_2 (y)

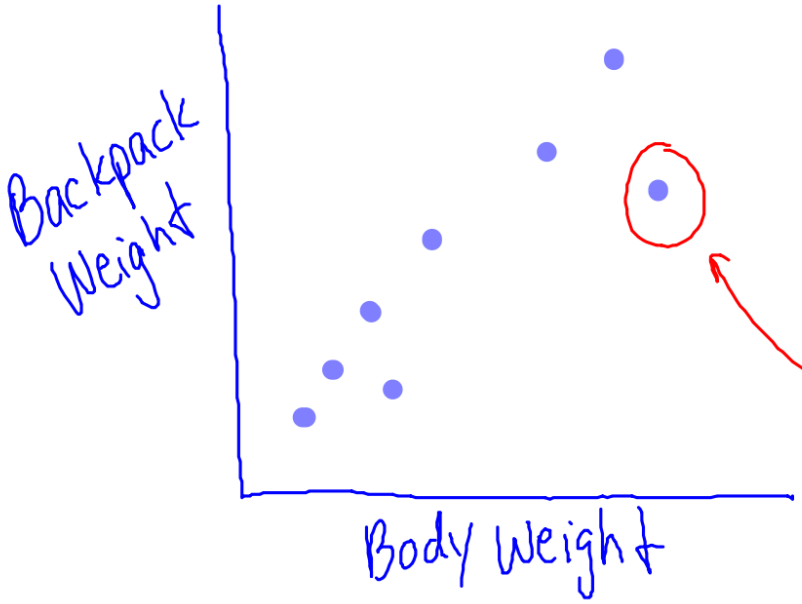
Body Weight Backpack Weight

2) Enter data into lists

3) Graph

STAT PLOT → L L L / L L L → Lists In Correct Order → ZOOM → 9

4) Sketch/Describe Graph



Direction - Positive

Form - Linear

Strength - Moderate

Outlier - 187 pound hiker

Pearson Product Moment Correlation Coefficient (r)

- Measures the direction / strength of the linear relationship between 2 quantitative variables

- Correlation does not mean causation !!

  Ex Ice cream sales → Drownings

- Uses means / standard deviations

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$\downarrow$$

$$-1 \leq r \leq 1$$

$|r|$

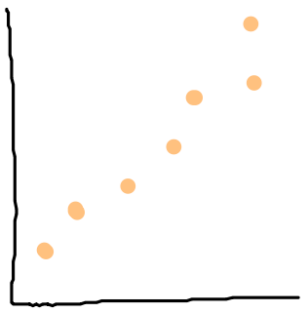
$|.90 - 1.00|$ - Strong Linear Relationship

$|.50 - .89|$ - Moderate

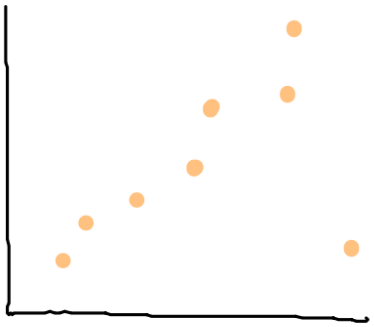
$|.25 - .49|$ - Weak

$|0 - .24|$ - None

- Correlation is not resistant



r = .90



r = .70

- Correlation is nondirectional

GPA / SAT Math = SAT Math / GPA

Calculating Correlation

1) Use Formula (P. 161, 19)

2) Use Calculator (Diagnostic "On")

   a) Scatterplot

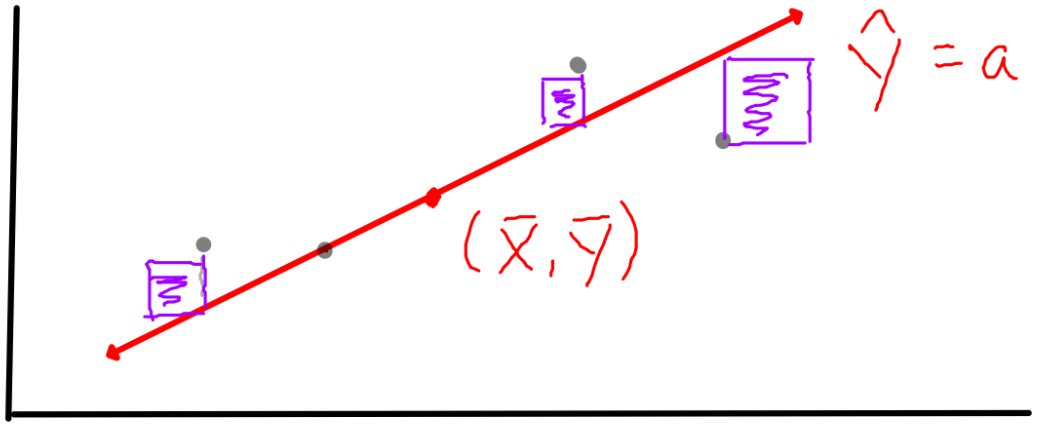
   b) STAT → CALC → LinReg (a+bx) → L_1, L_2

      $r = .79$ ($r = .95$ w/out outlier!)

Sec 3.2

Least Squares Regression Line (LSRL)

- Line ($\hat{y} = a + bx$) which minimizes the sum of the squares of the vertical distances of the observed points from the line
- LSRL is a model used to make predictions



$$\text{Slope}(b) = r\frac{s_y}{s_x}$$

$$\hat{y} = a + bx$$

$$\bar{y} = a + b\bar{x}$$

$$\text{Intercept}(a) = \bar{y} - b\bar{x}$$

Forms of Linear Equations

Algebra $y = mx + b$

Statistics $\hat{y} = a + bx$ > 2 variables

$$\hat{y} = a + b_1x_1 + b_2x_2 + \cdots b_nx_n$$ > multiple variables

- $\hat{y}$: Pulse
- a: Base Prediction
- b_1x_1: Age
- b_2x_2: Weight

Finding LSRL (Heavy Backpacks, P.145)

L_1 (x)	L_2 (y)
Body Weight	Backpack Weight

1) Using Formulas

a) Calculate Statistics

$\overline{x} = 136.13$ $\quad$ $\overline{y} = 28.63$

$S_x = 30.296$ $\quad$ $S_y = 3.462$

} 1 or 2 Variable Stats

$r = .795$ } LinReg (a + bx)

b) Calculate Slope and y-Intercept

$$b = r\frac{s_y}{s_x} = (.795)\frac{3.462}{30.296} = .0908$$

$$a = \bar{y} - b\bar{x} = 28.625 - (.0908)(136.125) = 16.2649$$

c) Write Equation In Words

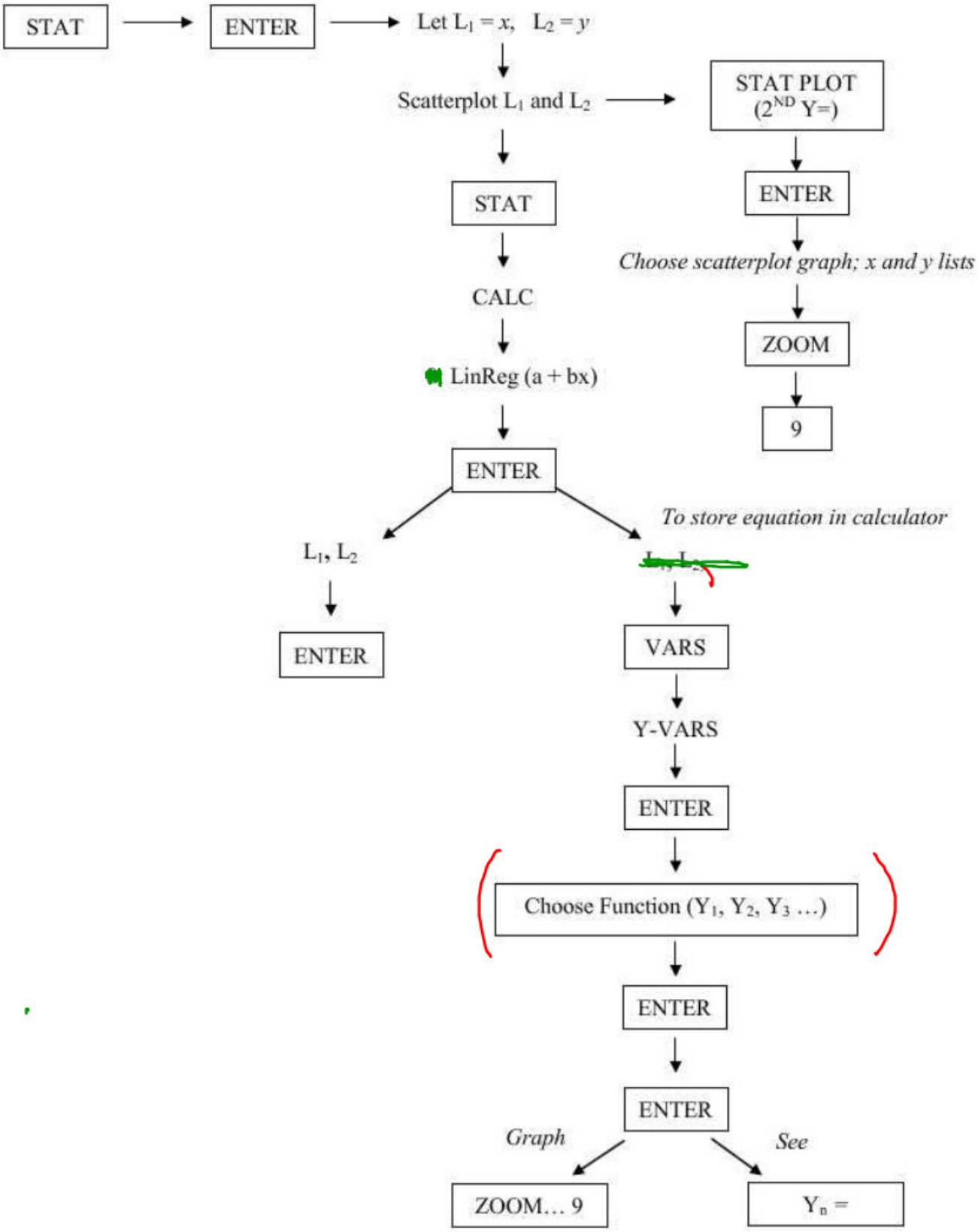
$$\hat{y} = 16.2649 + .0908x$$

$\hat{}$ Backpack Weight = 16.2649 + .0908 (Body Weight)

2) Using Calculator

- See Handout

FINDING BEST FIT LINES
(TI-83/84)



Making Predictions

1) Plug 'n Chug

$$\widehat{\text{Backpack Weight}} = 16.26 + .0907(165)$$

$$\approx 31 \text{ lbs}$$

2) Use Calculator → Need To Store Equation

- See Handout

MAKING PREDICTIONS
(TI-83/84)

Follow steps to "Finding Best Fit Lines"

↓

[VARS]

↓

Y-VARS

↓

1:Function

↓

[ENTER]

↓

Choose Function Defining Line (Y_1, Y_2, Y_3 …)

↓

[ENTER]

↓

[Y_n(*x*-value)]

↓

[ENTER]

Caution

Avoid Extrapolation ... making predictions far beyond x values

Coefficient of Determination (r^2)

- Tells what proportion/percent the variation/change in y is determined by x based on 2 linear models

LSRL

$\hat{y} = a + bx$

$\hat{y} = \bar{y}$

Understanding r-squared

Goal: To understand how r^2, the coefficient of determination, describes the **strength** of a linear model. As Rossman points out, r^2 "measures how closely the points fall to the least squares line and thus also provides an indication of how confident one can be of predictions made with the line"[1] Or to paraphrase Moore "When you report a regression, give r^2 as a measure of how successful the model is explaining the response [for a given explanatory value].[2]

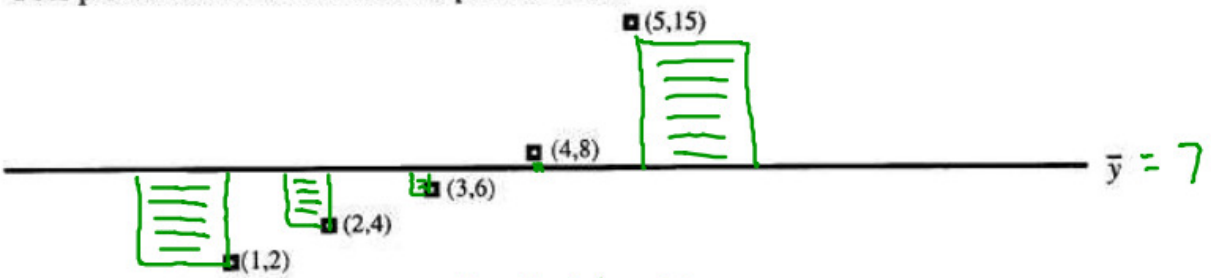
Goal of Modeling

1. To understand r^2 we need to ask the following question.

   If we did not know any modeling techniques such as regression what would the best model be?

   The answer is that, lacking any sophisticated techniques, our best model is the horizontal line containing the mean of the response values, i.e. $y = \bar{y}$.

2. Let's look at an example. Assume the data from a bivariate experiment are the points shown below.[3] On your calculator, draw a scatterplot of the data and a horizontal line representing $\bar{y}$. Your plot should look similar to the picture below.



   a. Is there error in this model? Absolutely !

   b. Draw a segment from each data point showing the error with respect to the model $y = \bar{y}$.

   c. Fill in the following chart:

		(1,2)	(2,4)	(3,6)	(4,8)	(5,15)
error with respect to the model $y = 7$	$y_i - \bar{y}$	-5	-3	-1	1	8
Square of the errors from the model $y = 7$	$(y_i - \bar{y})^2$	25	9	1	1	64

   d. Draw shaded "squares" on your plot to represent the squared values just computed (note that since the scales of the axes are not the same, the "squares" will look like rectangles). Some may overlap.

   e. $\sum (y_i - \bar{y})^2 =$ 100

SST - Total Sum of Squares About the Mean

1 *Workshop Statistics.* Allen Rossman. page 139
2 *Basic Practice of Statistics.* David S. Moore. p 127
3 Data set from Gretchen Davis at Santa Monica HS via Eric Mulfinger, Westridge School, Pasadena, CA

Understanding r-squared

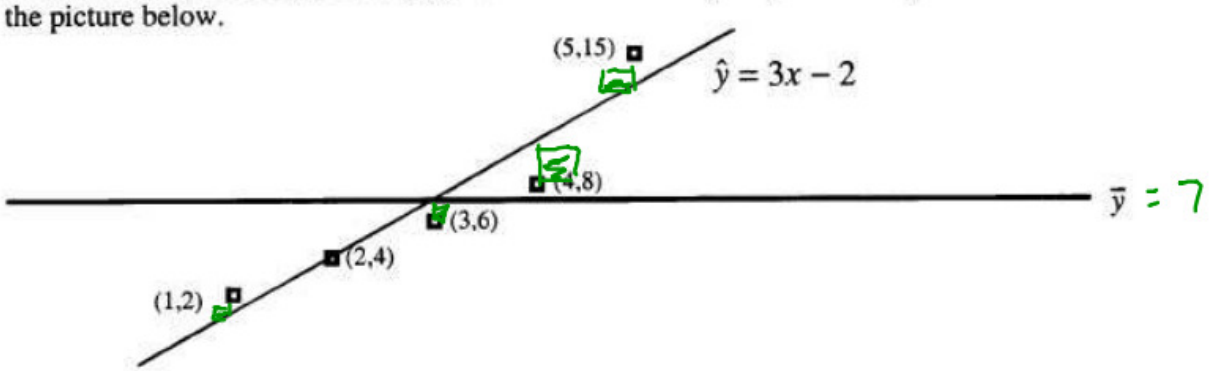
3. So what is the goal of our modeling efforts?

The goal of our modeling effort is to find a better model than the mean of the response variable.

We certainly would want the *sum of the squares of the errors from the new model* to be less than that of *the model using the mean of the response variable.*

A Better Model

1. So let's look for a better model. How about a Least Squares Regression Line (LSR line)? Using the same data as before, add the graph of the LSR line to your plot. Your plot should look like the picture below.



2. Is there still error in this new model? Comparing the two models, which appears to have lower error? yes... but not as much :)

3. Draw a segment from each data point showing the error to the model $\hat{y} = 3x - 2$.

4. What does the "hat" symbol mean on $\hat{y}$? > Predicted y

5. Fill in the following chart:

		(1,2)	(2,4)	(3,6)	(4,8)	(5,15)
error with respect to the model $\hat{y} = 3x - 2$	$y_i - \hat{y}$	1	0	-1	-2	2
Square of the errors with respect to the model $\hat{y} = 3x - 2$	$(y_i - \hat{y})^2$	1	0	1	4	4

6. Draw shaded "squares" on your plot to represent the values just computed. Does the sum of the areas of these squares seem smaller than those from the model using the response mean?

7. $\sum(y_i - \hat{y})^2 =$ 10 } SSE - Sum of Squares for Errors

Understanding r-squared

9. You have determined that $\sum(y_i - \bar{y})^2 = 100$ and $\sum(y_i - \hat{y})^2 = 10$. This information suggests what conclusion?

LSRL

Comparing Models

1. The natural next step is to find a number which gives us a sense how our new model compares with the model of the mean of the response variable. Of course we would like this number to be r^2.

$$\frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} = \frac{10}{100} = .10$$

The answer should be 0.1 . Which of the following correctly describes the proportion you just computed?

a. The proportion of how much error there is in the new model with respect to the error in the mean model.

b. The proportion of how well the new model fits the data.

2. The correct answer to the previous question was *The proportion of how much error there is the new model with respect to the error in the mean model.* Remembering that we want r^2 to show us how *well* our model measures how closely the observed values fall to the least squares line. How would we compute r^2 from the proportion of error in the model? See footnote [4] for a hint.

3. So $r^2 = 1 - \frac{1}{10} = 0.9$. Check this against the r^2 your calculator computed.

So $r^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SST - SSE}{SST} = \frac{100 - 10}{100} = .90$

In other words, you find r^2 by finding:

$$1 - \frac{\text{the sum of the squares of the error of the observed data with respect to the model}}{\text{the sum of the squares of the error of the observed data with respect to the mean of the response variable}}$$

4. Experiment with the following Geometer' Sketchpad files located on the file server in the folder APSTATS- GSP: rsqr1pt.gps & rsqr3pts.gsp. Can you explain what each square represents to your teacher?

[4] What is the correct value of r-squared? See your calculator.

MEAN 2012 SAT SCORES BY STATE

State	Participation Rate	Critical Reading	Math	Combined
Alabama	8%	538	531	1069
Alaska	54%	512	507	1019
Arizona	27%	517	525	1042
Arkansas	4%	565	566	1131
California	55%	495	512	1007
Colorado	17%	575	581	1156
Connecticut	88%	506	512	1018
Delaware	100%	456	462	918
District of Colum	83%	466	460	926
Florida	66%	492	492	984
Georgia	81%	488	489	977
Hawaii	66%	478	500	978
Idaho	20%	547	541	1088
Illinois	5%	596	615	1211
Indiana	69%	493	501	994
Iowa	3%	603	606	1209
Kansas	6%	584	594	1178
Kentucky	6%	579	575	1154
Louisiana	9%	542	536	1078
Maine	93%	470	472	942
Maryland	74%	497	502	999
Massachusetts	89%	513	530	1043
Michigan	4%	586	603	1189
Minnesota	7%	592	606	1198
Mississippi	4%	561	544	1105
Missouri	5%	589	592	1181
Montana	28%	536	536	1072
Nebraska	5%	576	585	1161
Nevada	49%	491	493	984
New Hampshire	75%	521	525	1046
New Jersey	78%	495	517	1012
New Mexico	13%	550	546	1096
New York	90%	483	500	983
North Carolina	68%	491	506	997
North Dakota	3%	588	610	1198
Ohio	19%	543	552	1095
Oklahoma	5%	568	566	1134
Oregon	57%	521	523	1044
Pennsylvania	74%	491	501	992
Rhode Island	69%	490	491	981
South Carolina	73%	481	488	969
South Dakota	3%	589	610	1199
Tennessee	10%	576	570	1146
Texas	62%	474	499	973
Utah	6%	568	566	1134
Vermont	69%	519	523	1042
Virginia	72%	510	512	1022
Washington	58%	519	530	1049
West Virginia	17%	516	502	1018
Wisconsin	4%	594	605	1199
Wyoming	5%	567	579	1146

Source: College Board

% Students Taking SAT → SAT Score

1) $r = -.88$

There is moderately strong negative linear relationship between SAT scores and the percent taking the test

2) $r^2 = .77$

77% of the variation in SAT scores (from state to state) can be explained by the percent who take it

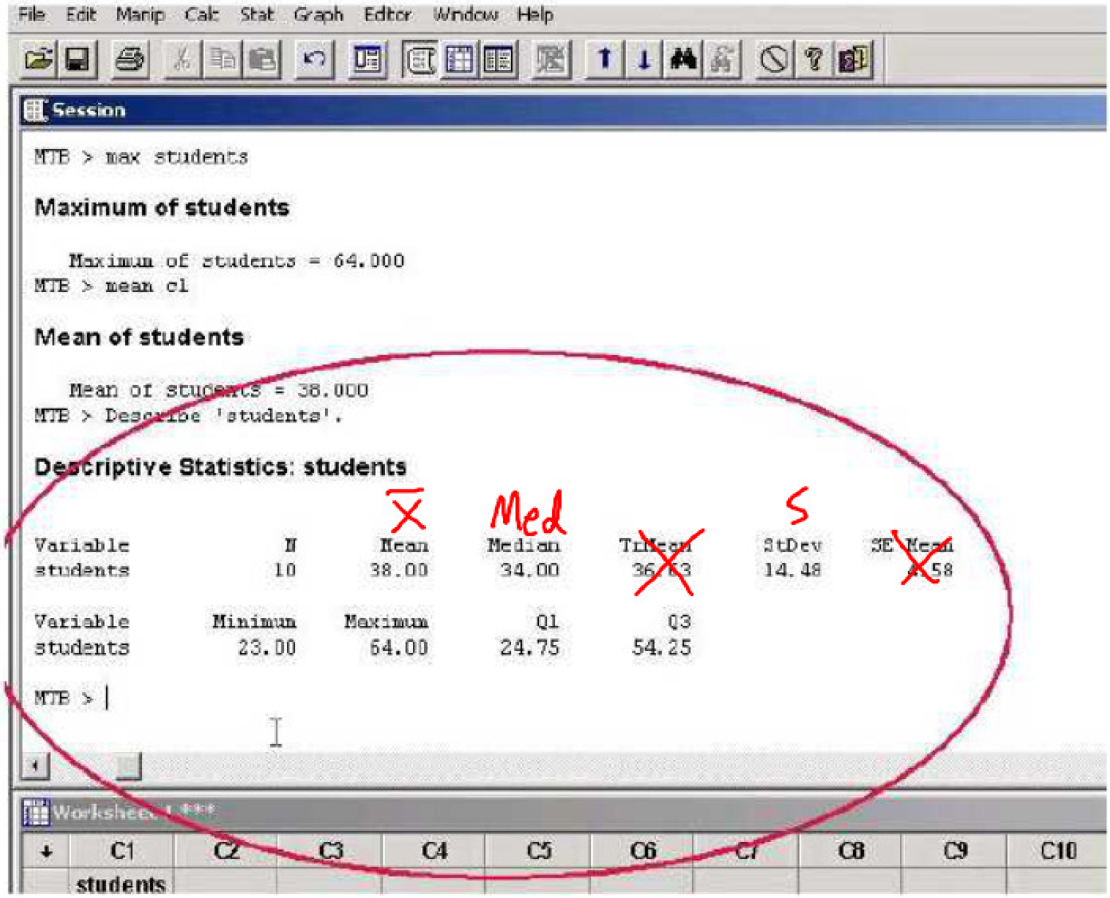
3) $\hat{y} = \bar{y} = 1068$

Regardless of % taking the SAT, predicted score will be 1068

4) $\widehat{SAT\ Score} = 1158.61 - 2.24(\%\ Taking)$

10% → 1136 90% → 957

MINITAB OUTPUT



Regression Analysis: Cost versus Income

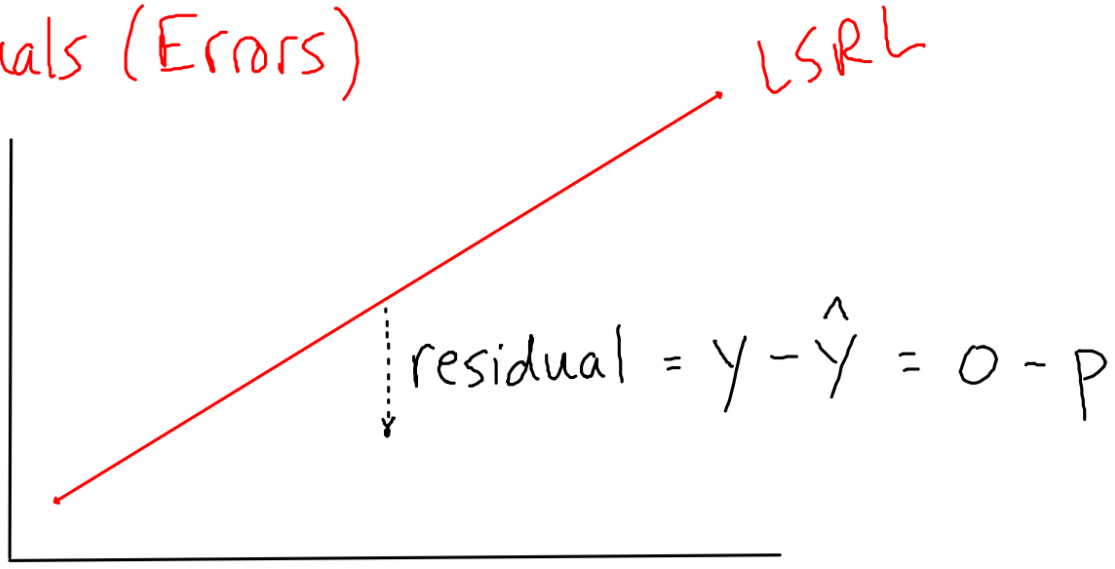
[redacted]

Predictor	Coef	SE Coef	T	P
a Constant (y-int)	438.525	3.341	131.25	0.000
b Income (slope)	0.51145	0.02326	22.00	0.000

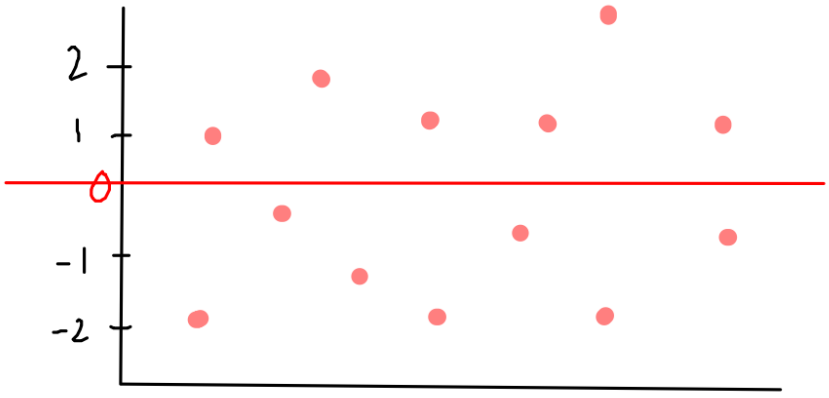
$\hat{y} = a + bx$

S = 12.2225   R-Sq = 91.0% (r^2)   R-Sq(adj) = 90.8%

Residuals (Errors)



Residual Plot (x, residuals)



Predictions Too Small

Predictions Too Big

No Pattern → Linear

Pattern → Not Linear

} Pp 176-177

Sum of all residuals = 0

$$\sum x = -.0000000000261 = 0 !$$

Mean of all residuals = 0

$$\bar{x} = -.000000000000163 = 0 !$$

Standard Deviation of Residuals (s)

Gives the approximate size of a "typical" prediction error (residual)

$$s = \sqrt{\frac{\sum (\text{residuals})^2}{n-2}}$$

Making Residual Plots (P. 164)

L_1 (x)	L_2 (y)
NEA (cal)	Fat Gain (kg)

↓

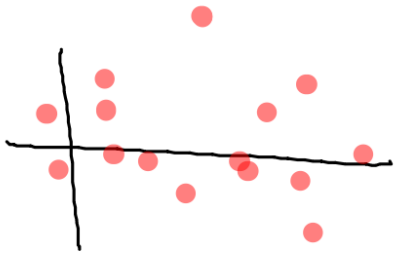
Scatterplot

↓

Calculate $r = -.77$

↓

Scatterplot (x, Residual)



Outliers

- Observations that lie outside overall pattern
- See P. 186 (Children 18 + 19)

Influential Observations

- Markedly changes slope of LSRL
- See P. 187 (Child 18 influential; Child 19 is not)

Given 2 (x,y) variables, is there a linear relationship?

Scatterplot (Form, Direction, Strength)

↓

Numerical Summaries (LSRL, r, r^2)

↓

Residual Plot

No Pattern → Linear Model?

Pattern → Exponential/Power Model (Ch 12)