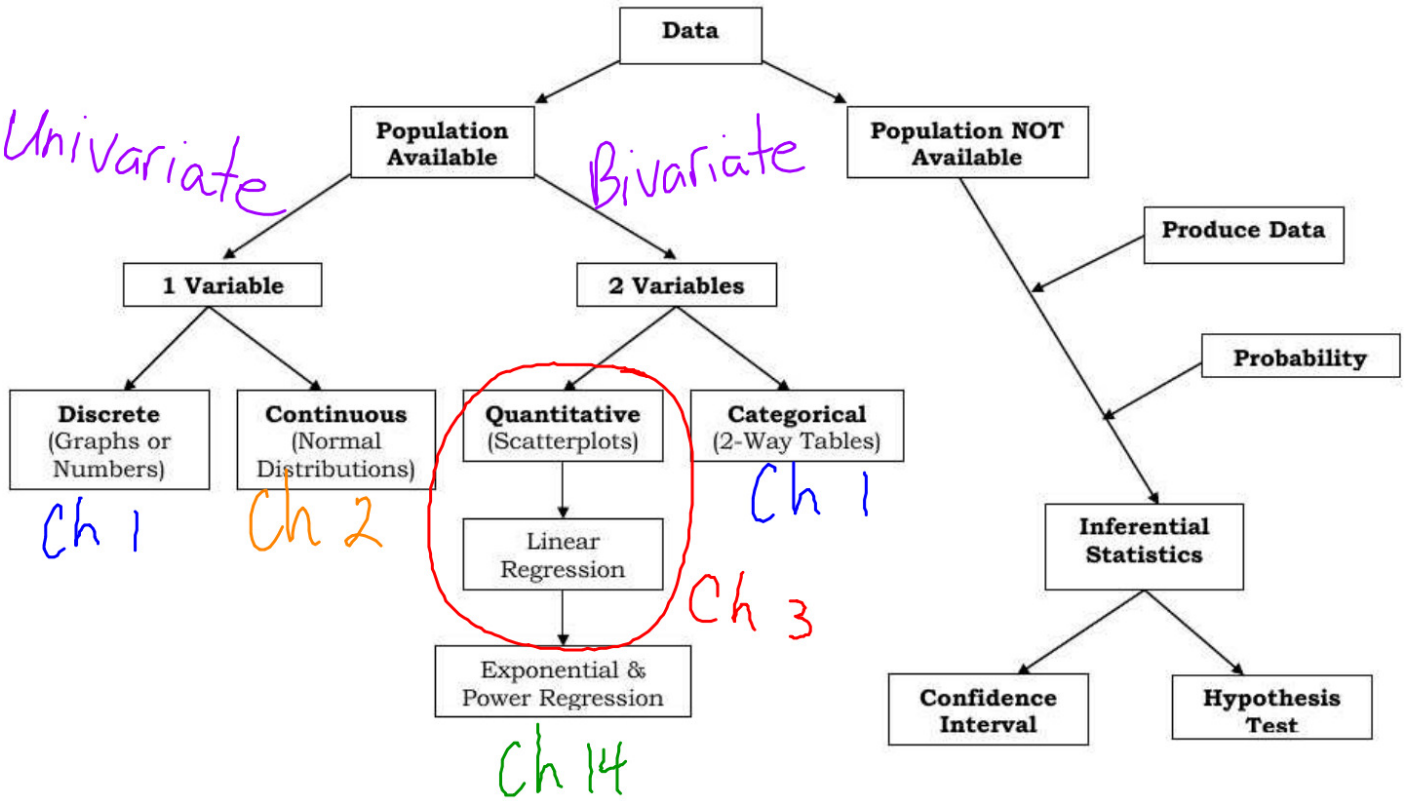
Sec 3.1

Success Criteria

- Analyze patterns in scatterplots
- Calculate and interpret correlations

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 - Clustered

Strength - Moderate

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

Making Scatterplots (P. 145)

1) Determine explanatory / response variables?

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

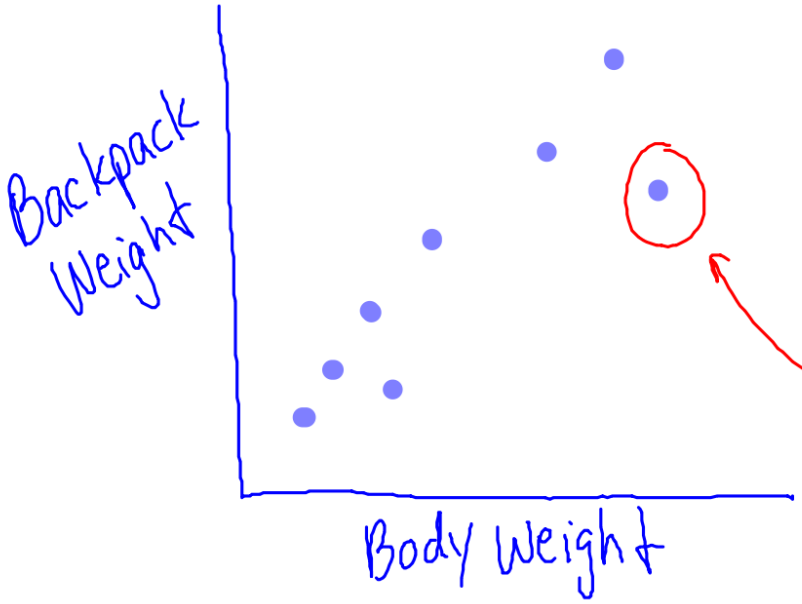
Body Weight Backpack Weight

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

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

$$\downarrow$$

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

P.151

|.90 - 1.00| - Strong Linear Relationship

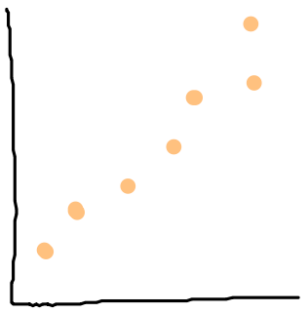
|.50 - .89| - Moderate

|.25 - .49| - Weak

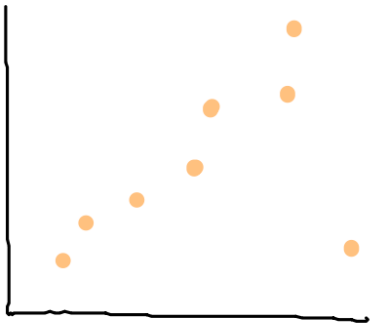
|0 - .24| - None

The stronger the correlation, the better the prediction

- 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 = .94 w/out outlier!)

Sec 3.2

Success Criteria

- Calculate Least Squares Regression Line (LSRL) using formulas and technology
- Use LSRL to make predictions

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

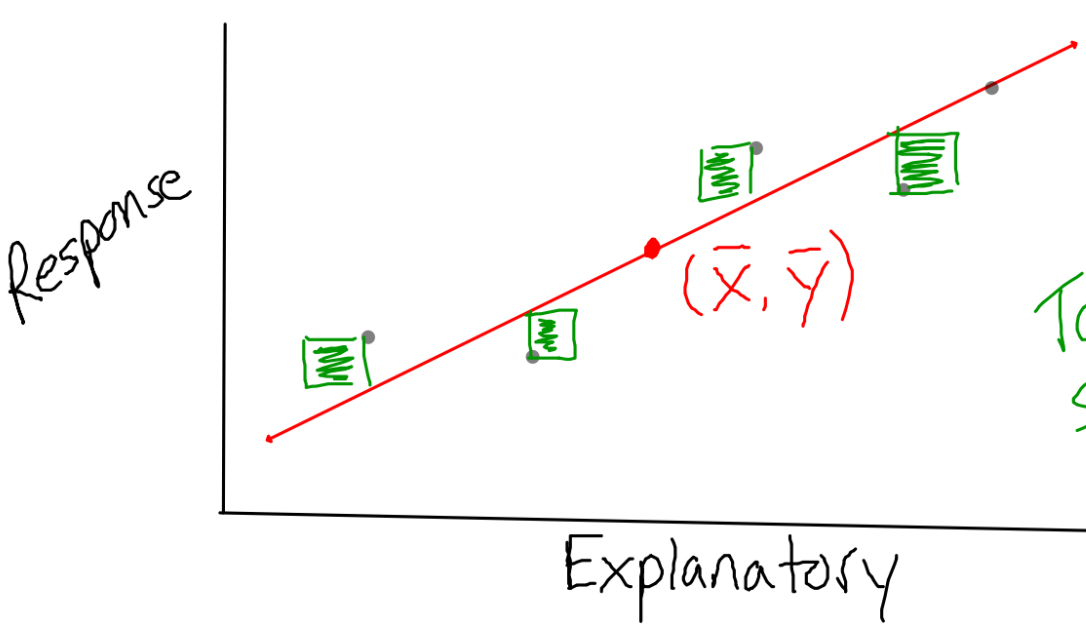
Forms of Linear Equations

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



Total sum of square (areas) small as possible

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

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

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

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

Finding LSRL (Heavy Backpacks, P.145)

$L_1(x)$ — Body Weight

$L_2(y)$ — Backpack Weight

1) Using Formulas

a) Calculate Statistics

$\overline{x} = 136.125$ $\quad$ $\overline{y} = 28.625$

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

} 1-Var stats OR 2-Var stats

$r = .7946$ } LinReg (a+bx)

b) Calculate Slope and y-Intercept

$$b = r\frac{s_y}{s_x} = (.794)\frac{3.462}{30.296} = .0908$$

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

c) Write Equation In Words

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

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

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

2) Using Calculator

- Find Correlation ($y = a + bx$)

$a = 16.2649$ $b = .0907$

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

$$\widehat{\text{Backpack Weight}} = 16.2649 + (.0907)(\text{Body Weight})$$

Making Predictions

1) Plug 'n Chug

$$\widehat{BP\ Weight} = 16.26 + .0907(Body\ Weight)$$

165

$$\widehat{BP\ Weight} = 31\ lbs$$

2) Use Calculator

a) Store Equation

LinReg(a+bx) L_1, L_2, [VARS] → Y-VARS

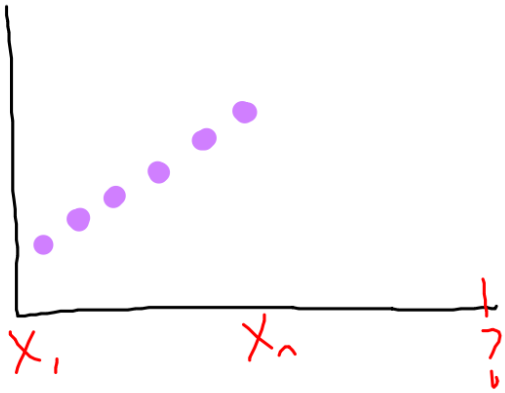
[ENTER] → [ENTER] → [ENTER]

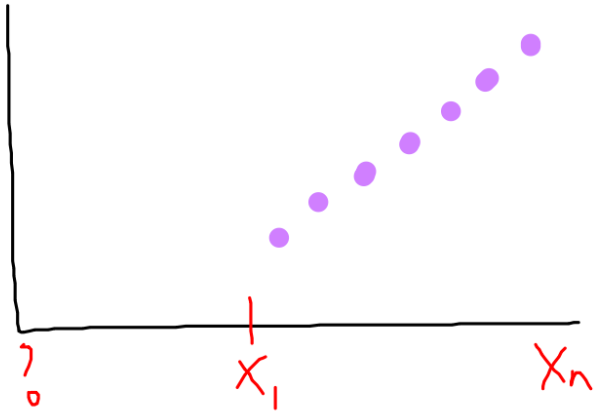
b) Make Prediction

[VARS] → Y-VARS → [ENTER] → [ENTER] → $Y_1(165)$

Caution

Avoid Extrapolation ... making predictions far beyond x values





Sec 3.2 (cont)

Success Criteria

- Interpret a coefficient of determination (r^2)
- Read a MINITAB output

UNDERSTANDING R-SQUARED

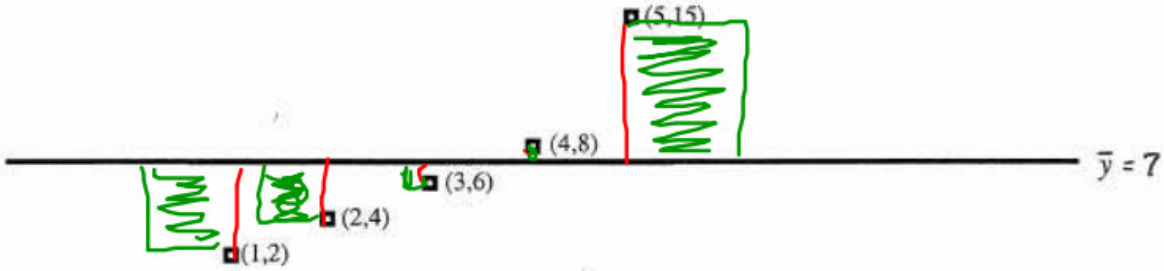
1. To understand r^2 we need to ask the following question:

   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 ($y = \bar{y}$)

2. Let's look at an example. Assume the data from a bivariate experiment are the points shown below. On your calculator, draw a scatterplot of the data and a horizontal line containing the mean of the response values ($y = 7$).

   Your plot should look similar to the picture below:



   a. Is there error in this model? Yes

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

   c. Complete the following table:

		(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 you plot to represent the squared values just computed. See Above

   e. $\sum(y_i - \bar{y})^2 =$ 100 = SST (Total Sum of Squares About the Mean)

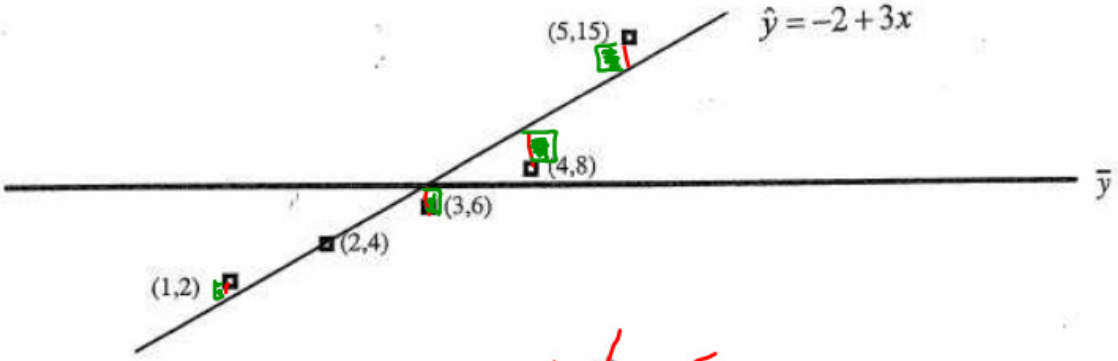
3. So what is the goal of our modeling efforts?

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

We certainly 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 (LSRL)? Using the same data, calculate the LSRL and, if possible, add the graph to your plot which should look like the picture below:



2. Is there still error in this new model? Yes

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

4. Complete the following table: See Above

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

5. Draw shaded "squares" on your plot to represent the squared values just computed. See Above Does the sum of the areas if these squares seem smaller than those from the model using the response mean? Yes!

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

Calculating r^2

$$r^2 = \frac{SST - SSE}{SST} = \frac{100 - 10}{100} = \frac{90}{100} = .90$$

✓ Check this value using your calculator.

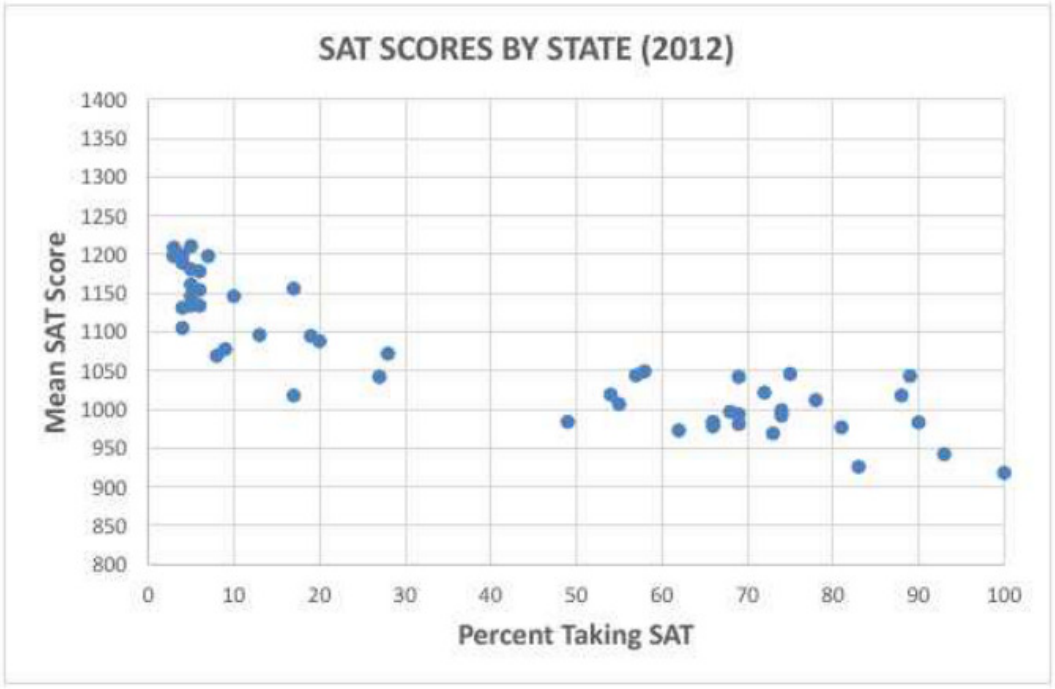
CALC → LinReg(a+bx) → $r^2 = .90$

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


SAT SCORES BY STATE (2012)
Mean SAT Score
1400
1350
1300
1250
1200
1150
1100
1050
1000
950
900
850
800
0
10
20
30
40
50
60
70
80
90
100
Percent Taking SAT

% 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) $\widehat{SAT\ Score} = \bar{y} = 1068$

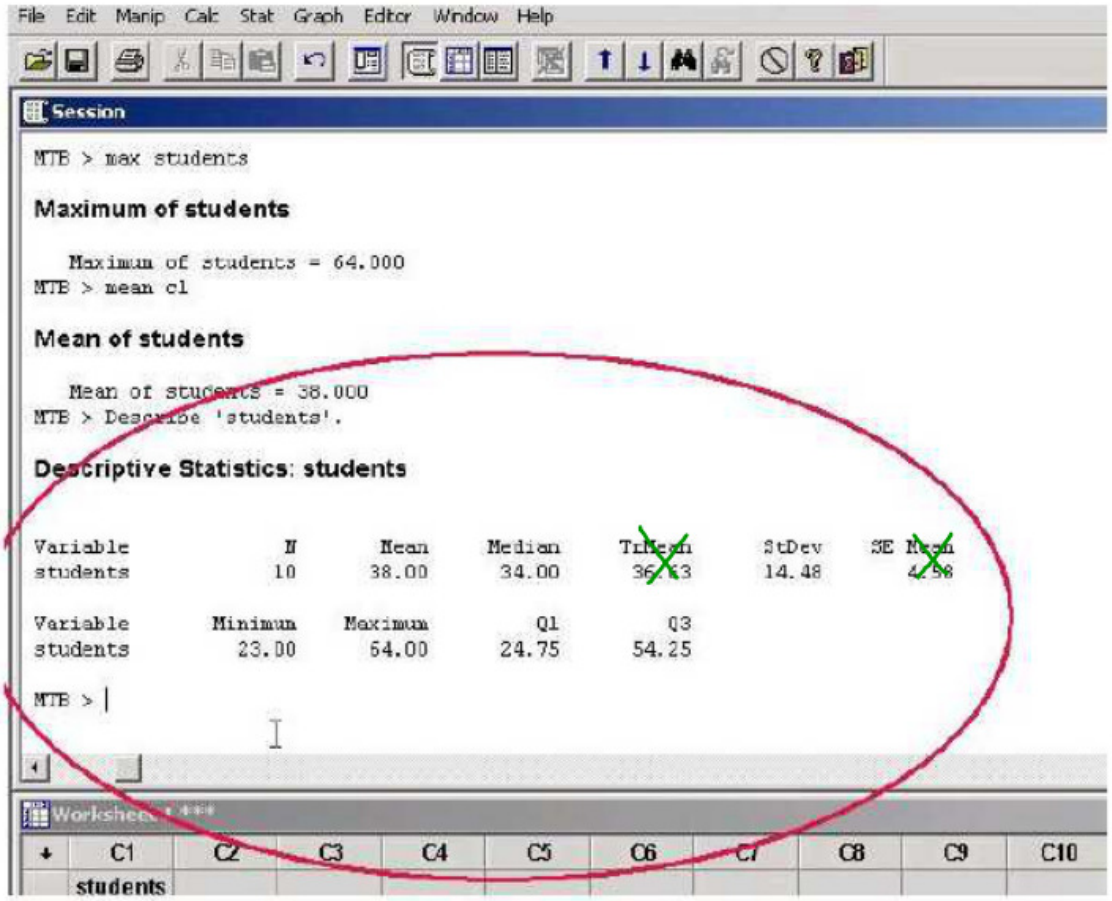
Regardless of % taking the SAT, predicted score will be 1068

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

10% → 1136 90% → 957

MINITAB OUTPUTS

Descriptive Statistics



File Edit Manip Calc Stat Graph Editor Window Help

Session

```
MTB > max students

Maximum of students

   Maximum of students = 64.000
MTB > mean c1

Mean of students

   Mean of students = 38.000
MTB > Describe 'students'.

Descriptive Statistics: students

Variable          N       Mean     Median     TrMean      StDev   SE Mean
students         10      38.00      34.00      36.13      14.48      4.58

Variable      Minimum    Maximum         Q1         Q3
students        23.00      64.00      24.75      54.25

MTB >
```

Worksheet

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
	students									

Regression Analysis: Cost versus Income

Predictor		Coef	SE Coef	T	P
Constant (y-int)	a =	438.525	3.341	131.25	0.000
Income (slope)	b =	0.51145	0.02325	22.00	0.000

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

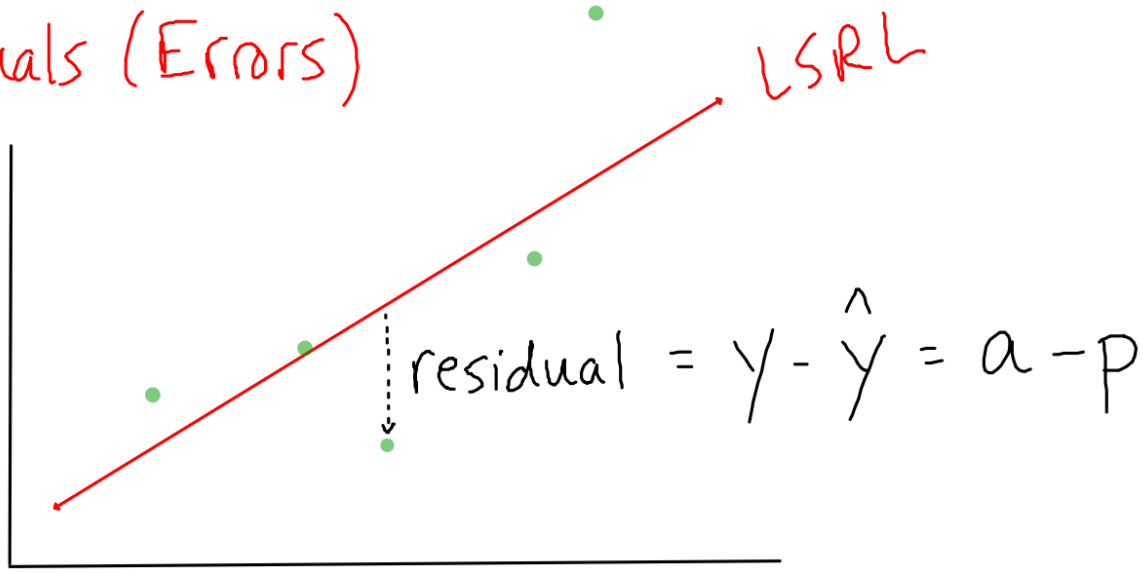
$r = ?$ | $\hat{Cost} = 438.525 + .51145(Income)$

Sec 3.2 (cont)

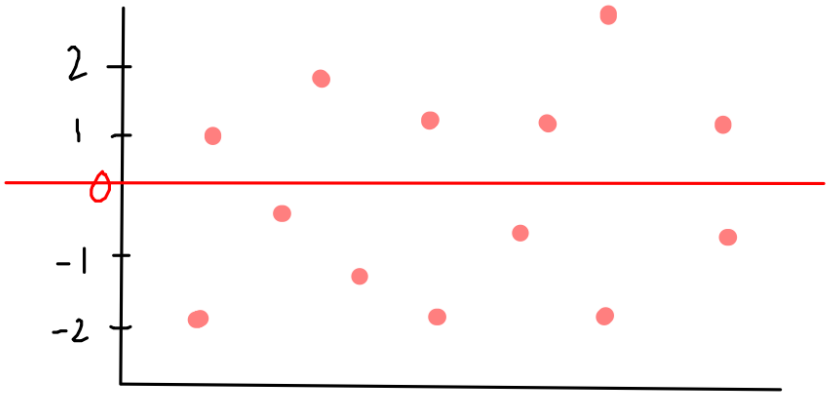
Success Criteria

- Construct and interpret a residual plot
- Determine if an outlier is influential

Residuals (Errors)



Residual Plot (x, residuals)



Predictions Too Small

Predictions Too Big

Random → Linear

Pattern → Not Linear

} Pp 176-177

Sum of all residuals = 0

Mean of all residuals = 0

Standard Deviation of Residuals (s)

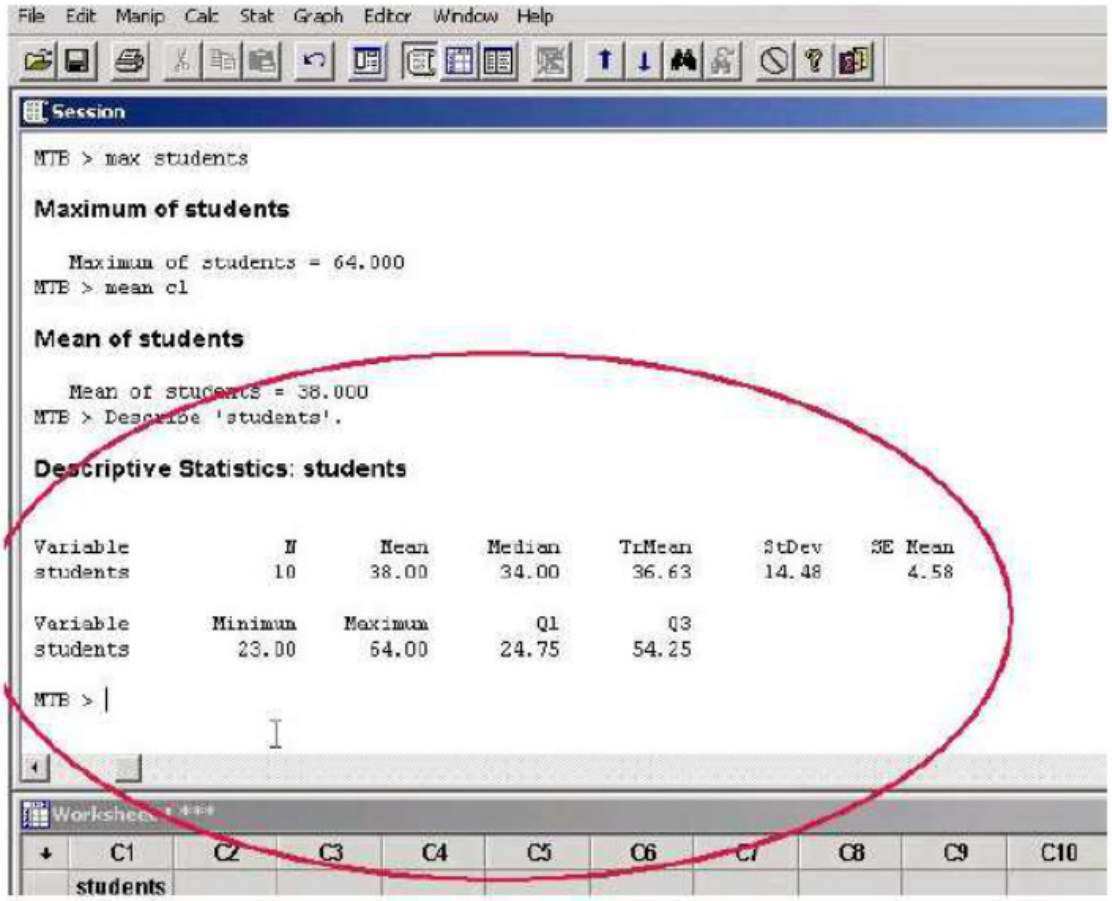
Gives the approximate size of a "typical" prediction error (residual)

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

Minitab Output ...

MINITAB OUTPUTS

Descriptive Statistics



Regression Analysis: Cost versus Income

Predictor	Coef	SE Coef	T	P
Constant	438.525	3.341	131.25	0.000
Income	0.51145	0.02325	22.00	0.000

s = 12.2225 R-Sq = 91.0% R-Sq (adj) = 90.8%

↑ standard deviation of residuals

Residual Lists

STAT → ENTER → Create "RESID" List ?

Making Residual Plots (P.145)

$L_1(x)$ Body Weight

$L_2(y)$ Backpack Weight

↓

Scatterplot

↓

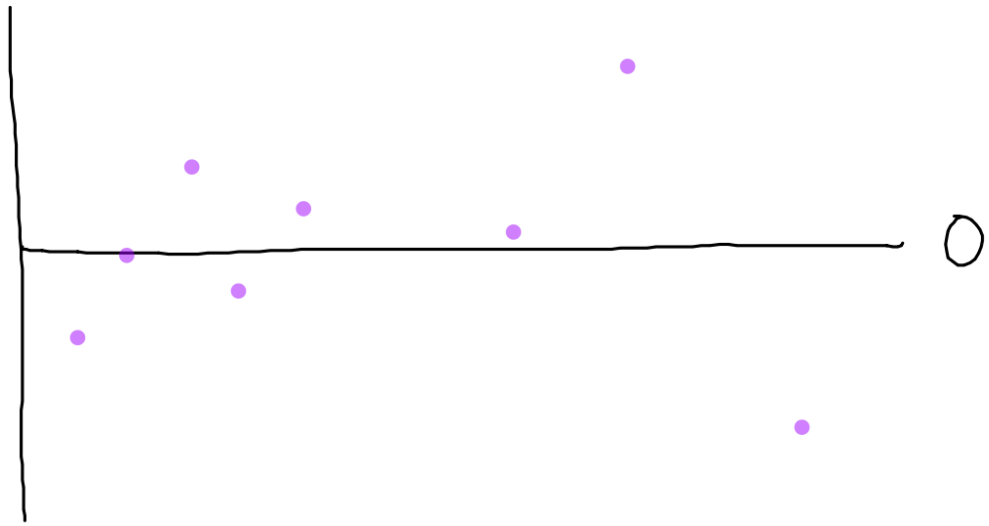
★ Calculate LSRL $\hat{Back} = 16.26 + .09(Body)$

↓

Scatterplot (X, Residuals)

L_1 RESID

Sketch Residual Plot



Let L_3 = RESID

1) Check sum of residuals = 0

$$\sum x = 0$$

2) Check mean of residuals = 0

$$\bar{x} = 0$$

} 1-Var Stats

3) Calculate standard deviation of residuals

$$s = \sqrt{\frac{\sum i(\text{residuals})^2}{n-2}} = \sqrt{\frac{30.904}{8-2}} = 2.26$$

Outliers

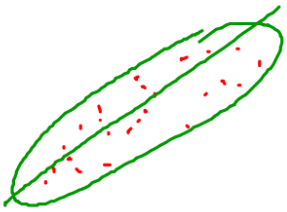
- Observations that lie outside overall pattern
- See P. 186 (Children 18 + 19)

Influential Observations

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

Review

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



Scatterplot (Form, Direction, Strength)

↓ Lin Reg

Numerical Summaries (LSRL, r^2, r)

↓

Residual Plot

No Pattern / Random

↓

Linear Model

$\hat{y} = a + bx$

Pattern / Not Random

↓

Exponential / Power Model

(ch 14)