


**COLLEGE OF THE CANYONS**

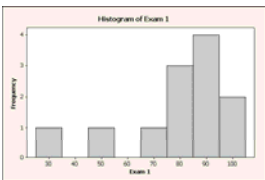
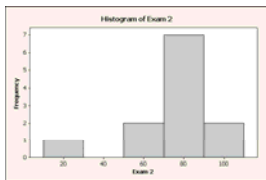
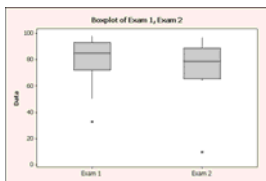


**Chapter 9 – Regression Wisdom**

### Load Today's Data

- <http://www.canyons.edu/faculty/morrowa/140/datasets>
  - Graduates
  - Inside this worksheet are four sets of data
    - Exam 1 and Exam 2 are exam scores for my math 140 in a previous semester (using a different book/exam structure)
    - Exam 1\_a and Exam 2\_a are different exam scores for math 140
    - Year and NumberGrad are graduation data for a small college
    - COCYear and COCNumber are graduation data for COC (along with source information)
- Get us Started...
  - Produce appropriate graphs and statistics for Exam 1 and Exam 2 as quantitative variables separately. Analyze the results.

### Just Checking...

**Descriptive Statistics: Exam 1, Exam 2**

Variable	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	IQR
Exam 1	79.00	19.46	378.73	33.00	72.00	85.00	93.00	98.00	21.00
Exam 2	74.17	22.79	519.42	10.00	65.50	78.50	88.50	97.00	23.00

### Quantitative Variables Together

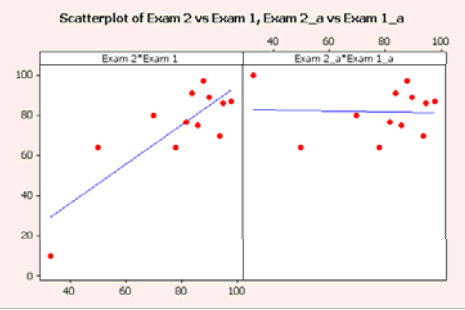
- Now create appropriate graphs to examine the relationship between
  1. Exam 1 and Exam 2
  2. Exam 1\_a and Exam 2\_a
- What do you notice?
- The teacher was concerned with the Exam 2\_a outlier, and so adjusted it to reflect the result shown in Exam 2. Is this justified? Is it fair?

### Just Checking...




- What will happen as we fit a regression line to each of these?
- Redraw the scatterplots to include the regression line.
  - Under "Multiple Plots", select "Separate Panels" and "Same Y"

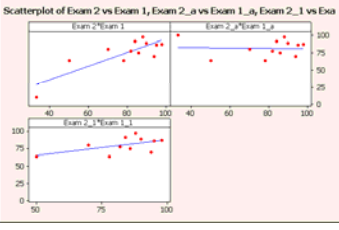
### Just Checking...



- What changes in the regression line?
- Copy the data to the neighboring column, delete the point in question from the copy, and plot all 3 datasets.

### Outliers, Leverage, and Influence

- Any outlier deserves separate inspection.
- A point with an x-value far from the rest is said to have **high leverage**.
- A point of high leverage may be **influential** if omitting it from the analysis gives a very different model.

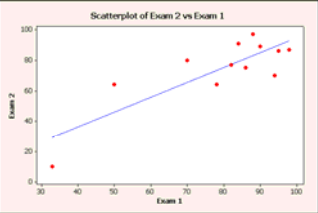
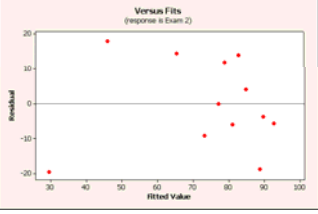


- Now continue by fitting a linear model to Exam 2 and Exam 1 (the actual data). Analyze your model.

### Regression and Model Checking

The regression equation is  $\text{Exam 2} = -2.5 + 0.970 \text{ Exam 1}$

$R\text{-Sq} = 68.6\%$

- What exactly does the equation give us?
- Interpret the slope and intercept

### Make a Prediction

- Using the data, we wish to make a prediction for students with Exam 1 scores as follows
  - Exam 1 Score: 0
  - Exam 1 Score: 100
  - Exam 1 Score: 50
- We make what's called an **extrapolation** when we make a prediction for new x's.
  - Bottom Line: We can't predict the future. (Think of the stock market.)
  - If you're going to extrapolate, don't believe that the prediction will come true.

### Predictions

- Exam 1 Score 0:
 

Obs	Fit	SE Fit	95% CI	95% PI
1	-2.48	16.83	(-39.99, 35.03)	(-50.40, 45.44)XX

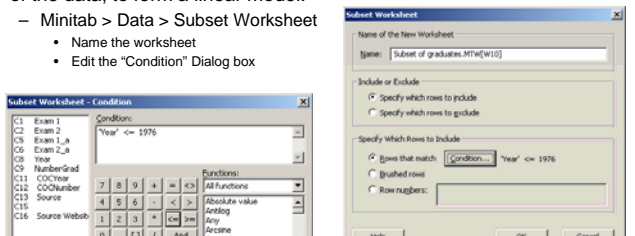
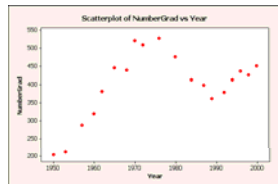
XX denotes a point that is an extreme outlier in the predictors.
- Exam 1 Score 100:
 

Obs	Fit	SE Fit	95% CI	95% PI
1	94.54	5.82	(81.57, 107.52)	(62.02, 127.07)
- Exam 1 Score 50:
 

Obs	Fit	SE Fit	95% CI	95% PI
1	46.03	7.15	(30.10, 61.96)	(12.22, 79.84)

### Potential Subsets

- Next we examine the Year NumberGrad data
- We wish to predict the number of grads in 1969
- While one line won't fit the data, we can look at a **subset**, or portion of the data, to form a linear model.
  - Minitab > Data > Subset Worksheet
    - Name the worksheet
    - Edit the "Condition" Dialog box



### Continue the Prediction...

- Note: The last procedure will give a warning because we have more than one dataset in the worksheet. For us today, this is okay.
- Performing regression (and predicting for 1969) on only the data for years less than or equal to 1976 gives the output to the right

### Regression Analysis: NumberGrad

The regression equation is  
 NumberGrad = - 27739 + 14.3 Year

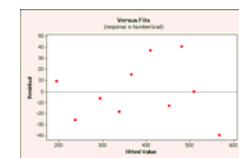
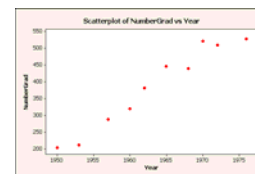
S = 27.6043 R-Sq = 95.6% R-Sq(Adj) = 95.6%

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI
1	466.05	10.71	(441.35, 490.75)

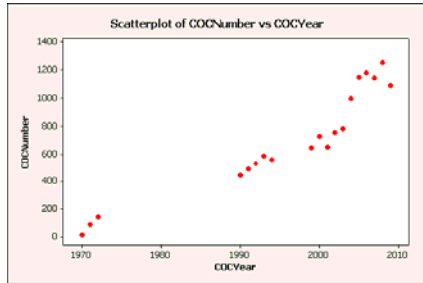
Values of Predictors for New Observations

New Obs	Year
1	1969



### Compare to COC Data

- Why do you think I was only able to find a small amount of data on the COC website?



### Lurking Variables and Rival Hypothesis

- A **rival hypothesis** is an alternate explanation for the results, different than the original investigators.
- For the following examples, provide rival hypotheses.
  1. There is a high correlation between online classes and success rates (rate of people who pass). Therefore, having more online classes will increase the college success rate.
  2. A high positive correlation was found between IQ scores and shoe size. Therefore, having larger shoes causes someone to have a higher IQ score.

### Exam 1: Chapters 1-9

- 70% Routine calculations/interpretations (like HW or class work)
- 30% Analyze a data set (like the project)
- You may bring in one 8.5"x11" sheet of notes, one side only.
- Projects are due at the beginning of class, meeting after exam. Fill out project/teammate evaluations ahead of time.
- See website for exam date –
  - Exam 1
  - You will have the entire class period to work (if you need it)
- Lecture Next Time: Review
- Class Work Next Time: Optional
  - Work on math 140 to receive a bonus class work credit. Come prepared (project, notes sheet, textbook to do review homework).

### Class Work

- To get credit, it is your responsibility to get checked off.
  1. Chapter 8 Handout
    - Rules for checking answers: No Pens in the Front!!!

### Homework

- Textbook/Routine Homework
  - Due Next Week (25% chance of collection)
    1. Read Chapter 9
    2. Pg 244-251: #1, 3, 11, 15, 19, 21, 31
- Project/Exploration Homework
  - Project #1

### Exam 1

- Exam 1, covering chapters 1-9. In one week.