Bivariate Data
Further Investigations

1

OUTLIER DETECTION
QUADRATIC AND EXPONENTIAL FITTINGS

Outlier Detection

2

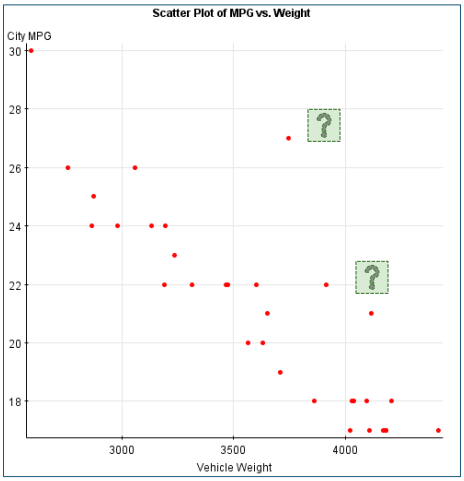
ADDRESSING THE EFFECT OF AN OUTLIER IN ANALYZING BIVARIATE DATA

3

Car MPG Data

Recall that, when visualizing the vehicle data with a scatter plot, it appeared that there may be potential outliers

- Example: What effect does an outlier have on our analysis?



The Effect on the Correlation Coefficient

4

- The linear correlation coefficient is very sensitive to outliers
  - This means that a single outlier can dramatically affect the value of the linear correlation coefficient
- Be Careful!
  - Practice caution when removing outliers
  - If a single value can affect the linear correlation coefficient so greatly, then removing that value must be done with cause
  - *Can you think of reasons why this caution is important?*

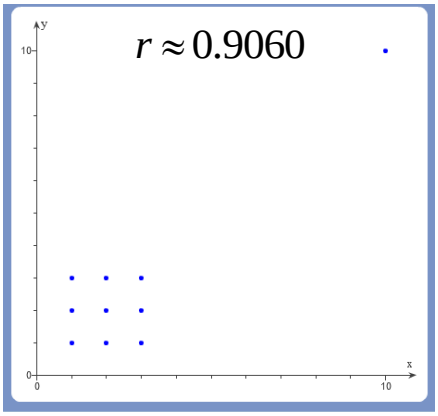
- Example:
  - Our original linear correlation coefficient was approximately
    $r \approx -0.8658$
  - With the removal of one point (3745,27), the linear correlation coefficient becomes approximately
    $r \approx -0.9266$
  - Moreover, by removing an additional two points (4115,21) and (3915,22), the linear correlation coefficient becomes approximately
    $r \approx -0.9588$

So, what does this all mean?

A Very Revealing Example

5

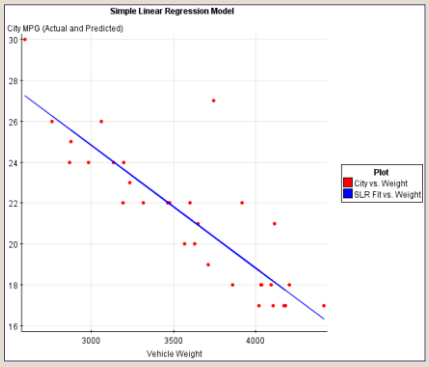
What happens if the outlier is removed in the situation below?



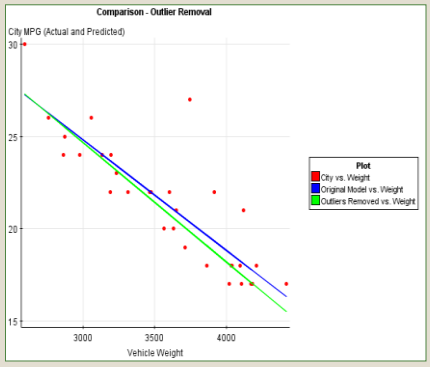
The Effect on the Regression Line

6

Recall the Previous Regression Line



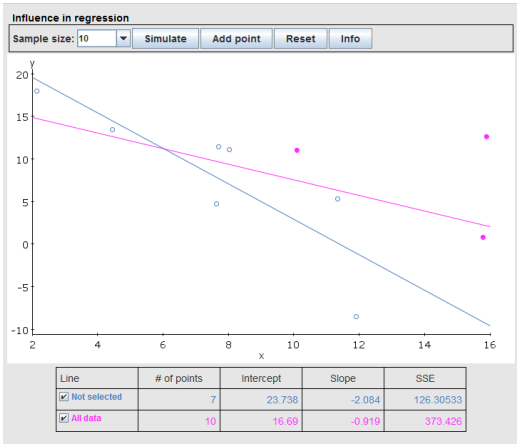
With the Outlier Removal



Another Revealing Example

7

What happens when values are removed below?



Line	# of points	Intercept	Slope	SSE
Not selected	7	23.738	-2.084	126.30533
All data	10	16.69	-0.919	373.426

Quadratic and Exponential Fittings

8

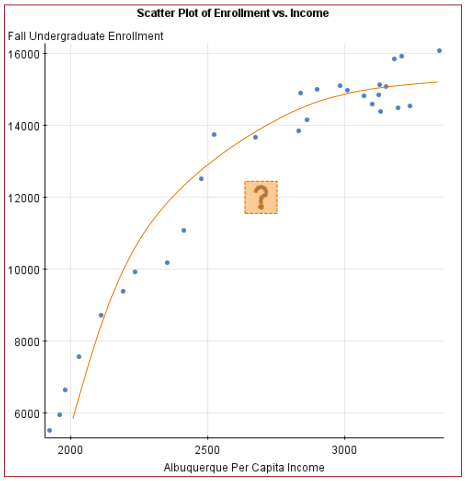
GETTING CLOSER FITTINGS AND INVESTIGATING MODELING RELATIONSHIPS

9

Enrollment Data

Recall that, when visualizing the data with a scatter plot, it appeared that there may have been an underlying quadratic relationship between the variables

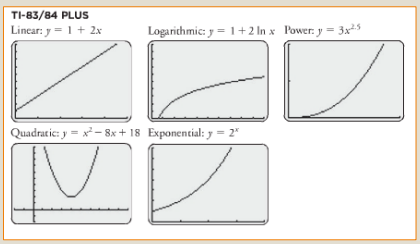
- Example: What can we do when there is a curvature apparent in the data?



Different Curves - Different Approaches

10

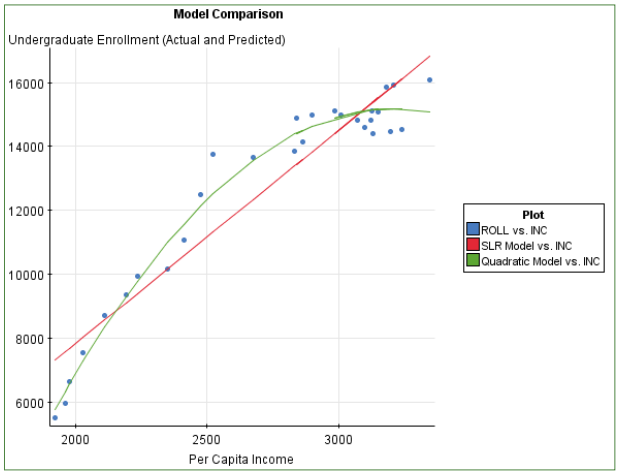
- Our previous approach considered *only* a linear relationship in bivariate data
- Nonlinear functions may fit the sample data best
  - Nonlinear mathematical models exist for model fitting
  - Some Examples:
    - Quadratic, Cubic, Logarithmic, Exponential, and Power models
  - We will focus on two models:
    - Quadratic: $\hat{y} = ax^2 + bx + c$
    - Exponential: $\hat{y} = ab^x$



11

Enrollment Data

Focusing on quadratic or exponential models to account for curvature in data, the data appears to be quadratic



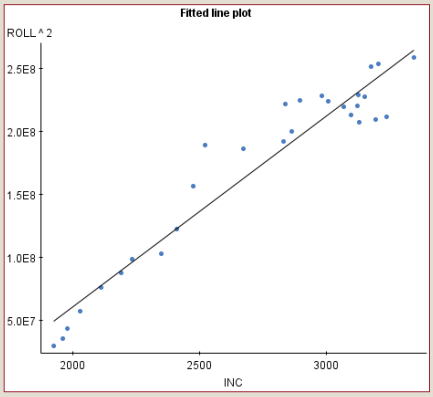
$\hat{y} = -5576.91 + 6.70x$ $\quad r^2 \approx 0.9023$

$\hat{y} = -0.01x^2 + 36.73x - 43681.52$ $\quad r^2 \approx 0.9737$

Enrollment Data (continued)

12

Is there a simpler model?



What if we square y?

- There is a slight improvement in the overall model fitting and we now have a simpler model of Per Capita Income vs. Enrollment squared
- However, this data may need to be explored further

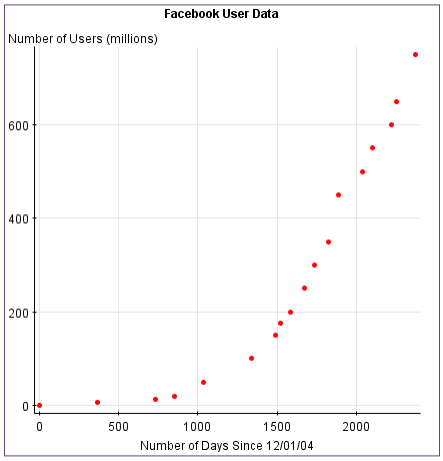
$\hat{y}^2 = -2.42E8 + 151499.33x$

$r^2 \approx 0.9270$

Another Example – Facebook Data

13

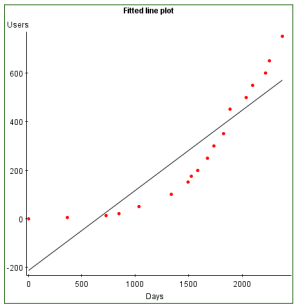
# Days Since 12/01/04	# Users (in millions)
0	1
365	5.5
730	12
851	20
1034	50
1339	100
1492	150
1523	175
1582	200
1673	250
1735	300
1826	350
1888	450
2038	500
2100	550
2222	600
2253	650
2373	750



14

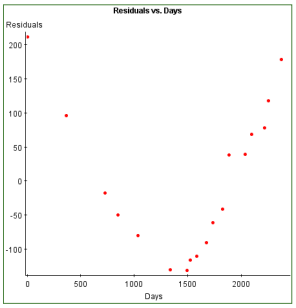
Facebook Data

Focusing on quadratic or exponential models to account for the shape, the data appears to be exponential and simple linear regression does not seem to be an appropriate fit



$$\hat{y} = -211.20 + 0.33x$$

$$r^2 \approx 0.8089$$



Facebook Data (continued)

15

- The exponential model can be used to model the data
  - It is not the only model for the data but seems appropriate
- The resulting equation is approximately:

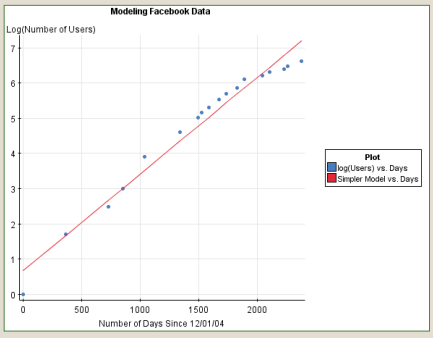
$$\hat{y} = 1.9559 \times (1.00275)^x$$

- This model is much more appropriate for prediction
- Prediction is still a main goal of modeling data
  - If the fit does not make sense for the given data, the predictions will not make sense

Facebook Data (continued)

16

Is there a simpler model?



What if we use log (y)?

- There is an improvement in the overall model fitting and we now have a simpler model of Days vs. log(Users)
- This, in some ways, verifies our choice of the exponential model

$$\log(\hat{y}) = 0.671 + 0.003x$$

$$r^2 \approx 0.9696$$

An Approach to the Modeling Cycle

General Steps for the Modeling Cycle:

1. Construct a scatter plot and verify that the pattern of the points is approximately a straight line pattern without outliers. If so, Simple Linear Regression can be used.
2. If there are outliers that are 'removable' and that affect the regression line, we may remove these points and attempt to use a linear fitting again.
3. If it appears that a different model is appropriate to account for the shape of the data, select a candidate model and assess the fit.
4. Verify the model selection by attempting to create a simple model through variable transform.
5. Use the model for prediction only if the line fits the data points well, the predictions make logical sense, and we are not predicting much beyond the scope of our available data.

Activity

In an experiment carried out by researchers, leaf litter was allowed to sit for a 20-week period in a bag in a moderately forested area. Initially, the total weight of the organic mass in the bag was 75.0 kg. Each week, the remaining amount was measured. The data is provided in a table on the next slide. Any questions?

19

Activity

Time (weeks)	Organic Mass (kg)
0	75
1	60.9
2	51.8
3	49.7
4	34.7
5	34.6
6	29.5
7	20.4
8	14.0
9	9.8
11	8.2
15	3.1
20	2.4

ANY QUESTIONS?

- Identify the role of each variable (i.e. which variable is the response variable, etc)
- Construct a scatter plot to visualize the data
  - Interpret and describe the resulting plot in terms of linear correlation, outlier detection, and shape.
- Fit a Simple Linear Regression Model for the data
  - Report the regression line and the value of the coefficient of determination.
  - Plot the line on the scatter plot you constructed and interpret the results.
  - Does this model seem appropriate for prediction? Why or why not? (*hint: try to predict the weight of the organic material at 0 weeks and, also, at 20 weeks*)
- Fit an additional mathematical model that you feel is appropriate
  - What type of model do you feel is appropriate? Why?
  - Report the coefficients for the chosen model
  - Can you interpret any of the coefficients in the model? How?
  - Plot the line on the scatter plot you previously constructed and interpret the results.
  - Does this model seem appropriate for prediction? Why?
  - Verify your model choice by performing an appropriate variable transformation and creating a 'simpler' model for the transformed data (*be sure to include a scatter plot of the transformed data, the new regression line equation, and a plot of the regression line*).