Bivariate Data

1

VISUALIZATION
LINEAR CORRELATION
SIMPLE LINEAR REGRESSION
RESIDUAL ANALYSIS

Visualizing Bivariate Data

2

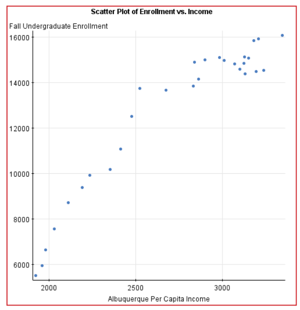
ASSESSING ASSOCIATIONS BETWEEN BIVARIATE (i.e. Paired) QUANTITATIVE DATA WITH SCATTER PLOTS

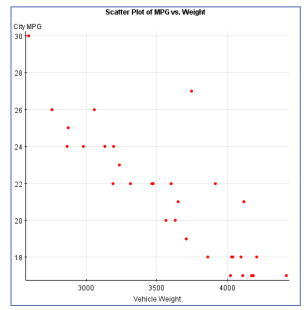
3

Scatter Plots

These plots assist in visualizing linear relationships, investigating non-linear associations, and identifying outliers in a dataset of two variables

- Example: What can we initially say about the scatter plots below?



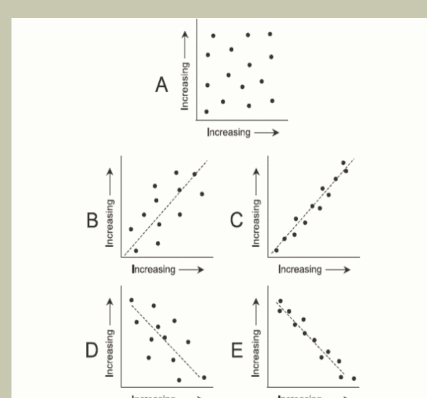


Defining Linear Correlation

(refer to p. 85-88 in text)

4

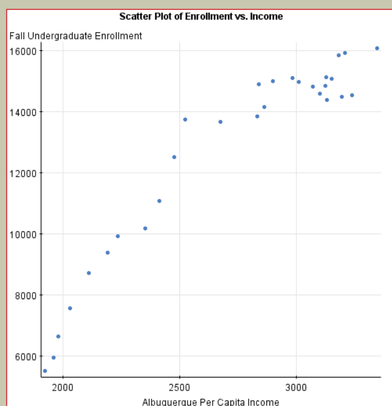
- Linear Correlation
  - Positive or Negative
  - Strong or Weak
- No Linear Correlation
  - No Correlation
  - Non-linear Relationship
- Notice: Correlation does not imply Causation (p. 89)



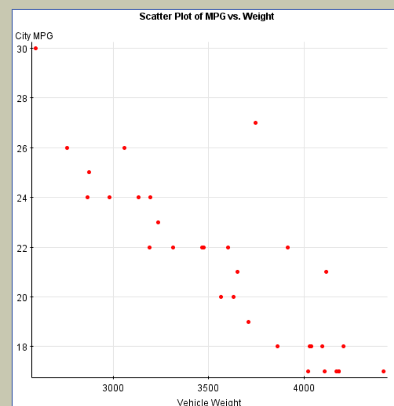
Further Explorations of Bivariate Data

5

Possible Non-Linear Relationship



Possible Outliers and/or Influential Points



Determining Linear Correlation

6

ASSESSING ASSOCIATIONS BETWEEN BIVARIATE (i.e. Paired) QUANTITATIVE DATA WITH PEARSON'S LINEAR CORRELATION COEFFICIENT (r)

7

The Linear Correlation Coefficient
(refer to p. 90-95 in text)

Properties

- Formula for Calculating r (p. 95):
  - $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{n(\sum xy)-(\sum x)*(\sum y)}{\sqrt{n(\sum x^2)-(\sum x)^2}*\sqrt{n(\sum y^2)-(\sum y)^2}}$
- Properties of r (p. 94):
  - $-1 \leq r \leq 1$
  - If all values of either variable are converted to a different scale, the value of r does not change
  - The value of r is not affected by the choice of x or y *(Can you see why this is the case?)*
  - r measures the strength of a <u>linear</u> relationship <u>only</u>!!
  - r is very sensitive to outliers in the sense that a single outlier can dramatically affect its value

Examples (continued)

8

Enrollment Data

ROLL	INC
5501	1923
5945	1961
6629	1979
7556	2030
8716	2112
9369	2192
9920	2235
10167	2351
11084	2411
12504	2475
13746	2524
13656	2674
13850	2833
14145	2863

14888	2839
14991	2898
14836	3123
14478	3195
14539	3239
14395	3129
14599	3100
14969	3008
15107	2983
14831	3069
15081	3151
15127	3127
15856	3179
15938	3207
16081	3345

City MPG Data

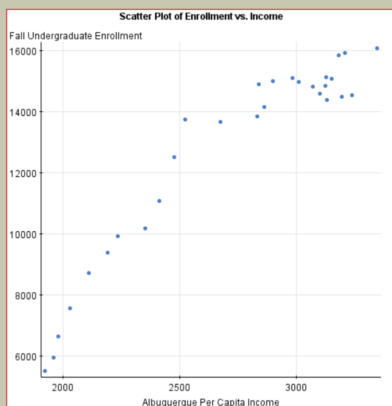
Weight	City
4035	18
3315	22
4115	21
3650	21
3565	20
4030	18
3710	19
3135	24
4105	17
4170	17
3190	22
4180	17
2760	26
3195	24
2980	24

4095	18
3860	18
4020	17
2875	25
3915	22
4205	18
4415	17
3060	26
3745	27
4180	17
3235	23
3475	22
2865	24
3600	22
2595	30
3465	22
3630	20

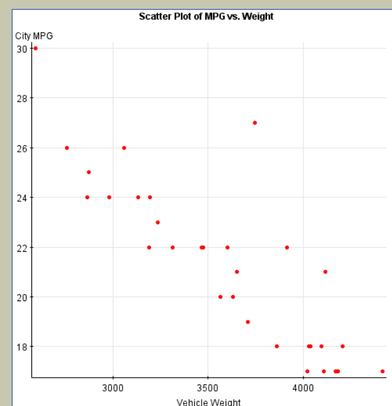
Examples (continued)

9

$r \approx 0.9499$



$r \approx -0.8658$



Simple Linear Regression

10

ASSESSING ASSOCIATIONS BETWEEN BIVARIATE (i.e. Paired) QUANTITATIVE DATA WITH SIMPLE LINEAR REGRESSION

A Line of "Best Fit"

(refer to p. 100-105)

11

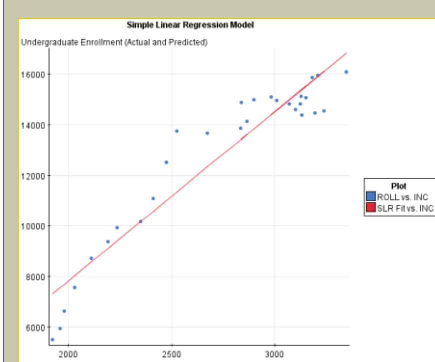
- Given a collection of paired sample data, simple linear regression attempts to algebraically describe the relationship between the two variables x and y.
- This algebraic description is often denoted as $\hat{y} = b_0 + b_1 x$
  - *Does this equation look familiar?*
  - *What do you think b_1 represents? What about b_0?*
  - *How would you describe this representation in terms of the relationship that exists between x and y?*
- The graph of the above equation is the Least Squares line
  - Also known as the "line of best fit" or the "regression line"
  - The least squares line fits the sample points best
  - The slope and intercept can be determined using the formulas on p. 103

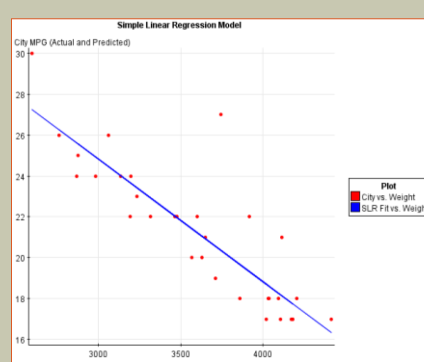
Examples (continued)

12

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

$$\hat{y} = 42.83 - 0.006x$$
$$r^2 \approx 0.7496$$





Residual Analysis

13

ASSESSING ASSOCIATIONS BETWEEN BIVARIATE (i.e. Paired) QUANTITATIVE DATA WITH SIMPLE LINEAR REGRESSION

14

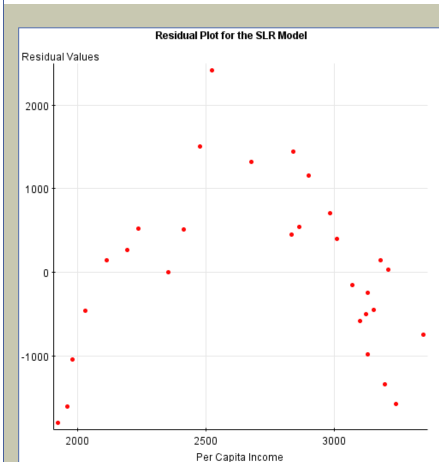
Residuals

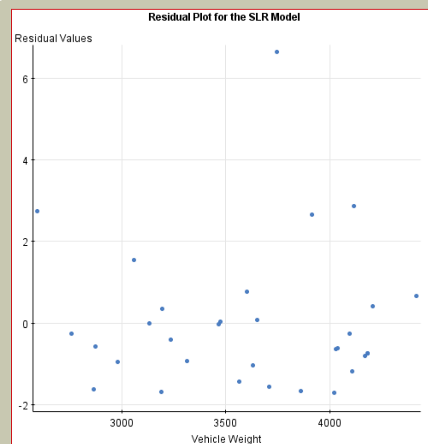
Inherent Error

- Residual – for a pair of sample x and y values, the difference between the observed sample value of y and the y-value that is predicted by using the regression equation is the residual
  - $Residual = Observed - Predicted = y - \hat{y}$
  - A residual represents, then, a type of inherent prediction error
  - *Ideally, what would we want residuals to be equal to? Would this ever be possible 0?*
- The Least Squares Property
  - The Least Squares line is a straight line that satisfies this property
  - The sum of the squares of the residuals is the smallest sum possible (*Why do you think we have to square the residuals here?*)
- Residual Plot
  - A residual plot is a scatter plot of the (x, y) values after each of the y-values has been replaced by the residual value, $y - \hat{y}$.
  - That is, a residual plot is a graph of the points $(x, y - \hat{y})$.
  - The residual plot can reveal information about the fit of the model for the given data

Examples (continued)

15





What do these plots reveal about the two fittings for the different datasets?

S. Robinson - University of Arkansas 7/4/2013

Further Investigations

16

- For the enrollment dataset, we may attempt to fit a quadratic model to account for the slight curvature
- For the city mpg dataset, there appears to be an outlier that may need evaluated more thoroughly
- Both models could be used for prediction purposes, however, when using regression equations for prediction, we must always consider the strength of the linear relationship that exists between the variables and, also, common errors such as extrapolation
  - Example:

  A simple random sample of 30 winning 5k times for competitive male runners aged 15-24 years resulted in a mean 5k time of 16.79 min. The sample linear correlation coefficient between the age of the runner and the 5k time for this sample was -0.903. The simple linear regression equation that fits this sample data was found to be $\hat{y} = 21.506 - 0.276x$ where x represents the age of the runner in years and y represents the 5k time for the runner in minutes. Can we use the regression equation above to predict the 5k time for a 65 year old competitive male runner? Why or why not?

S. Robinson - University of Arkansas 7/4/2013

A Complete Simple Linear Regression Analysis

17

- ✓ General Steps for a Complete Regression Analysis:
  - ✓ Construct a scatter plot and verify that the pattern of the points is approximately a straight line pattern without outliers
  - ✓ Assess the linear correlation between two variables of interest and create a regression equation and least squares line
  - ✓ Plot the least squares line and verify that the fitting is appropriate
  - ✓ Construct a residual plot and verify that there is no pattern (other than a straight line pattern) and also verify that the residual plot does not become thicker or thinner
  - ✓ Use a histogram to confirm that the values of the residuals have a distribution that is approximately normal
  - ✓ Consider any effects of a pattern over time

S. Robinson - University of Arkansas 7/4/2013

18

Activity (p. 114)

- Using the data in #6 on p. 114 in the text, answer #6 (a-f)
- Also, calculate the residuals for the least-squares line
- Plot the residuals and determine if there is a pattern that would suggest any further exploration

ANY QUESTIONS?

S. Robinson - University of Arkansas 7/4/2013