Chapter 5

Summarizing Bivariate Data

1

Terms

A **multivariate** data set consists of measurements or observations on each of two or more variables.

The classroom data set introduced in the slides for Chapter 1 is a multivariate data set. The data set includes observations on the variables: age, weight, height, gender, vision (correction method), and smoke (status). Age, weight and height are numerical variables while gender, vision and smoke are categorical variables.

2

Terms

A **bivariate** data set consists of measurements or observations on each of two variables.

For the rest of this chapter we will concentrate on dealing with bivariate data sets where both variables are numeric.

3

Review & Further Discussion of Scatterplots

Remember from chapter 3 that a scatterplot is a plot of pairs of observed values (both quantitative) of two different variables.

If one of the variables is considered to be a response variable (y) and the other an explanatory variable (x). The explanatory variable is usually plotted on the x axis.

4

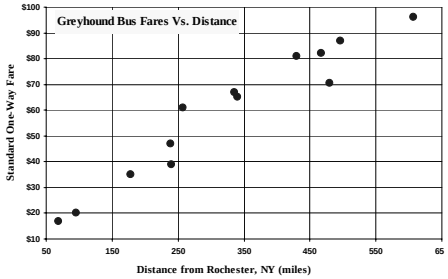
Example

The sample of one-way Greyhound bus fares (chap3) from Rochester, NY to cities less than 750 miles was taken by going to Greyhound's website. The following table gives the destination city, the distance and the one-way fare. Distance should be the x axis and the Fare should be the y axis.

Destination City	Distance	Standard One-Way Fare
Albany, NY	240	39
Baltimore, MD	430	81
Buffalo, NY	69	17
Chicago, IL	607	96
Cleveland, OH	257	61
Montreal, QU	480	70.5
New York City, NY	340	65
Ottawa, ON	467	82
Philadelphia, PA	335	67
Potsdam, NY	239	47
Syracuse, NY	95	20
Toronto, ON	178	35
Washington, DC	496	87

5

Example Scatterplot



6

Comments

The axes need not intersect at (0,0).

For each of the axes, the scale should be chosen so that the minimum and maximum values on the scale are convenient and the values to be plotted are between the two values.

Notice that for this example,

1. The x axis (distance) runs from 50 to 650 miles where the data points are between 69 and 607.
2. The y axis (fare) runs from $10 to $100 where the data points are between $17 and $96.

7

Further Comments

It is possible that two points might have the same x value with different y values. Notice that Potsdam (239) and Albany (240) come very close to having the same x value but the y values are $8 apart. Clearly, the value of y in not determined **solely** by the x value (there are factors other than distance that affect the fare.

In this example, the y value tends to increase a x increases. We say that there is a positive relationship between the variables distance and fare.

It appears that the y value (fare) could be predicted reasonably well from the x value (distance) by finding a line that is close to the points in the plot.

8

5.1: Correlation

Positive Association - Two variables are positively associated when above-average values of one tend to accompany above-average values of the other and below-average values tend similarly to occur together. (i.e., Generally speaking, the y values tend to increase as the x values increase.)

Negative Association - Two variables are negatively associated when above-average values of one accompany below-average values of the other, and vice versa. (i.e., Generally speaking, the y values tend to decrease as the x values increase.)

9

The Pearson Correlation Coefficient

A measure of the strength of the linear relationship between the two variables is called the Pearson correlation coefficient.

The Pearson sample correlation coefficient is defined by

$$r = \frac{\sum z_x z_y}{n-1} = \frac{\sum \left[\left(\frac{x-\bar{x}}{s_x} \right) \left(\frac{y-\bar{y}}{s_y} \right) \right]}{n-1}$$

10

Example Calculation

x	y	$\frac{x-\bar{x}}{s_x}$	$\frac{y-\bar{y}}{s_y}$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
240	39	-0.5214	-0.7856	0.4096
430	81	0.6357	0.8610	0.5473
69	17	-1.5627	-1.6481	2.5755
607	96	1.7135	1.4491	2.4831
257	61	-0.4178	0.0769	-0.0321
480	70.5	0.9402	0.4494	0.4225
340	65	0.0876	0.2337	0.0205
467	82	0.8610	0.9002	0.7751
335	67	0.0571	0.3121	0.0178
239	47	-0.5275	-0.4720	0.2489
95	20	-1.4044	-1.5305	2.1494
178	35	-0.8989	-0.9424	0.8472
496	87	1.0376	1.0962	1.1374
				11.6021

$\bar{x} = 325.615$

$s_x = 164.2125$

$\bar{y} = 59.0385$

$s_y = 25.506$

$$r = \frac{11.601}{13-1} = 0.9668$$

11

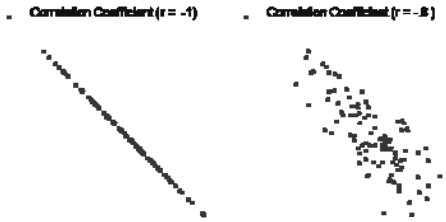
The Pearson Revisited

Using the calculation formula we have:

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} \sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}}$$

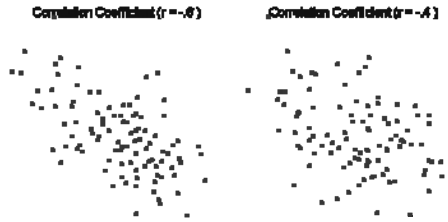
12

Some Correlation Pictures



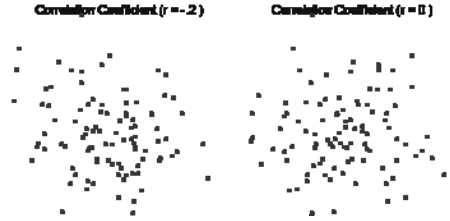
13

Some Correlation Pictures



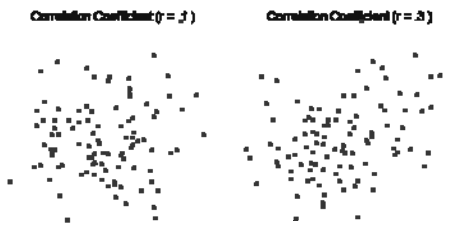
14

Some Correlation Pictures



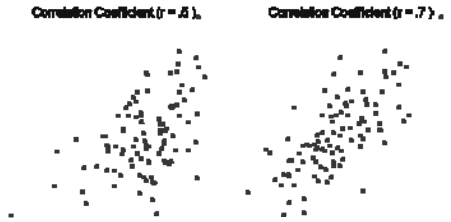
15

Some Correlation Pictures



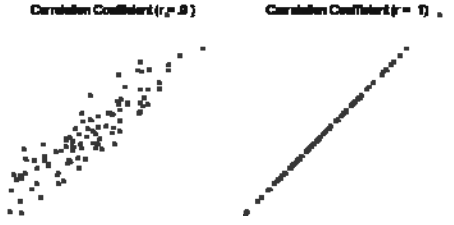
16

Some Correlation Pictures



17

Some Correlation Pictures



18

Properties of r

The value of r does not depend on the unit of measurement for each variable.

The value of r does not depend on which of the two variables is labeled x.

The value of r is between –1 and +1.

The correlation coefficient is

a) –1 only when all the points lie on a downward-sloping line, and
b) +1 only when all the points lie on an upward-sloping line.

The value of r is a measure of the extent to which x and y are linearly related.

19

An Interesting Example

Consider the following bivariate data set:

x	y
1.2	23.3
2.5	21.5
6.5	12.2
13.1	3.9
24.2	4.0
34.1	18.0
20.8	1.7
37.5	26.1

20

An Interesting Example

Computing the Pearson correlation coefficient, we find that r = 0.001

x	y	$\frac{x-\bar{x}}{s_x}$	$\frac{y-\bar{y}}{s_y}$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
1.2	23.3	-1.167	0.973	-1.136
2.5	21.5	-1.074	0.788	-0.847
6.5	12.2	-0.788	-0.168	0.133
13.1	3.9	-0.314	-1.022	0.322
24.2	4.0	0.481	-1.012	-0.487
34.1	18.0	1.191	0.428	0.510
20.8	1.7	0.237	-1.249	-0.296
37.5	26.1	1.434	1.261	1.810
				0.007
			r =	0.001

$\bar{x} = 17.488,\ s_x = 13.951,\ \bar{y} = 13.838,\ s_y = 9.721$

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) = \frac{1}{7}(0.007) = 0.001$$

21

An Interesting Example

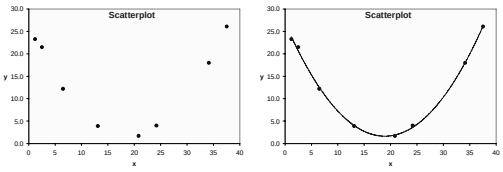
With a sample Pearson correlation coefficient, r = 0.001, one would note that there seems to be little or no linearity to the relationship between x and y.

Be careful that you do not infer that there is no relationship between x and y.

22

An Interesting Example

Note (below) that there appears to be an almost perfect quadratic relationship between x and y when the scatterplot is drawn.



23

5.2: Linear Relations

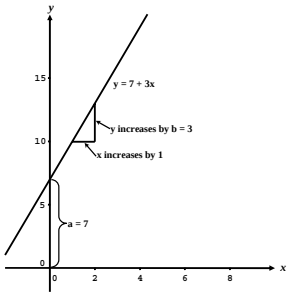
The relationship $y = a + bx$ is the equation of a straight line.

The value b, called the **slope** of the line, is the amount by which y increases when x increase by 1 unit.

The value of a, called the **intercept** (or sometimes the **vertical intercept**) of the line, is the height of the line above the value x = 0.

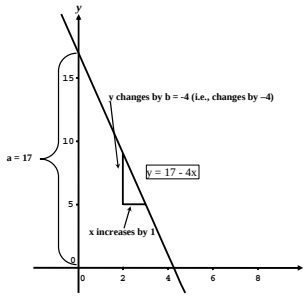
24

Example



25

Example



26

Least Squares Line

The most widely used criterion for measuring the goodness of fit of a line $y = a + bx$ to bivariate data (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ is the sum of the of the squared deviations about the line:

$$\sum [y - (a + bx)]^2$$
$$= [y_1 - (a + bx_1)]^2 + \ldots + [y_n - (a + bx_n)]^2$$

The line that gives the best fit to the data is the one that minimizes this sum; it is called the **least squares line** or **sample regression line**.

27

Coefficients a and b

The slope of the least squares line is $b = \frac{\sum\left[(x-\bar{x})(y-\bar{y})\right]}{\sum(x-\bar{x})^2}$

And the y intercept is $a = \bar{y} - b\bar{x}$

We write the equation of the least squares line as

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

where the ^ above y emphasizes that $\hat{y}$ (read as y-hat) is a prediction of y resulting from the substitution of a particular value into the equation.

28

Calculating Formula for b

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

29

Greyhound Example Continued

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
240	39	-85.615	7329.994	-20.038	1715.60
430	81	104.385	10896.148	21.962	2292.45
69	17	-256.615	65851.456	-42.038	10787.72
607	96	281.385	79177.302	36.962	10400.41
257	61	-68.615	4708.071	1.962	-134.59
480	70.5	154.385	23834.609	11.462	1769.49
340	65	14.385	206.917	5.962	85.75
467	82	141.385	19989.609	22.962	3246.41
335	67	9.385	88.071	7.962	74.72
239	47	-86.615	7502.225	-12.038	1042.72
95	20	-230.615	53183.456	-39.038	9002.87
178	35	-147.615	21790.302	-24.038	3548.45
496	87	170.385	29030.917	27.962	4764.22
4233	**768**		**323589.08**		**48596.19**

30

Calculations

From the previous slide, we have

$\sum\left[(x-\bar{x})(y-\bar{y})\right] = 48596.19$ and

$\sum(x-\bar{x})^2 = 323589.08$

So

$$b = \frac{\sum\left[(x-\bar{x})(y-\bar{y})\right]}{\sum(x-\bar{x})^2} = \frac{48596.19}{323589.08} = 0.15018$$

Also $n = 13$, $\sum x = 4233$ and $\sum y = 768$

so $\bar{x} = \frac{4233}{13} = 325.615$ and $\bar{y} = \frac{768}{13} = 59.0385$

This gives

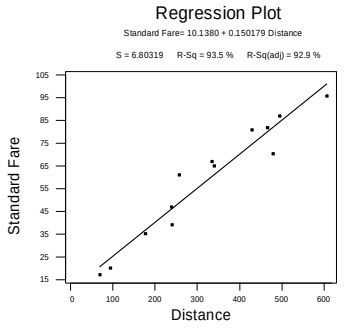
$a = \bar{y} - b\bar{x} = 59.0385 - 0.15018(325.615) = 10.138$

The regression line is $\hat{y} = 10.138 + 0.15018x$.

31

Minitab Graph

The following graph is a copy of the output from a Minitab command to graph the regression line.



32

Greyhound Example Revisited

x	y	x^2	xy
240	39	57600	9360
430	81	184900	34830
69	17	4761	1173
607	96	368449	58272
257	61	66049	15677
480	70.5	230400	33840
340	65	115600	22100
467	82	218089	38294
335	67	112225	22445
239	47	57121	11233
95	20	9025	1900
178	35	31684	6230
496	87	246016	43152
4233	**768**	**1701919**	**298506**

33

Greyhound Example Revisited

Using the calculation formula we have:

$n = 13$, $\sum x = 4233$, $\sum y = 768$

$\sum x^2 = 1701919$, and $\sum xy = 298506$

so

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{298506 - \frac{(4233)(768)}{13}}{1701919 - \frac{(4233)^2}{13}}$$

$$= \frac{48596.19}{323589.1} = 0.15018$$

As before $a = \bar{y} - b\bar{x} = 59.0385 - 0.15018(325.615) = 10.138$

and the regression line is $\hat{y} = 10.138 + 0.15018x$.

Notice that we get the same result.

34

5.3: Assessing the fit of the Line

Three Important Questions

To examine how useful or effective the line summarizing the relationship between x and y, we consider the following three questions.

1. Is a line an appropriate way to summarize the relationship between the two variables?
2. Are there any unusual aspects of the data set that we need to consider before proceeding to use the regression line to make predictions?
3. If we decide that it is reasonable to use the regression line as a basis for prediction, how accurate can we expect predictions based on the regression line to be?

35

Terminology

The **predicted** or **fitted values** result from substituting each sample x value into the equation for the least squares line. This gives

$\hat{y}_1 = a + bx_1$ =1st predicted value

$\hat{y}_2 = a + bx_2$ =2nd predicted value

...

$\hat{y}_n = a + bx_n$ =nth predicted value

The **residuals** for the least squares line are the values: $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n$

36

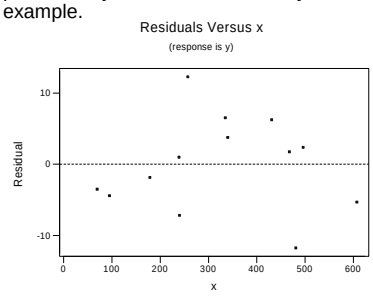
Greyhound Example Continued

x	y	Predicted value $\hat{y}=10.1+.150x$	Residual $y-\hat{y}$
240	39	46.18	-7.181
430	81	74.72	6.285
69	17	20.50	-3.500
607	96	101.30	-5.297
257	61	48.73	12.266
480	70.5	82.22	-11.724
340	65	61.20	3.801
467	82	80.27	1.728
335	67	60.45	6.552
239	47	46.03	0.969
95	20	24.41	-4.405
178	35	36.87	-1.870
496	87	84.63	2.373

37

Residual Plot

A **residual plot** is a scatter plot of the data pairs (x, residual). The following plot was produced by Minitab from the Greyhound example.

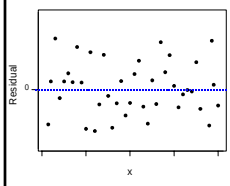


38

Residual Plot - What to look for.

Isolated points or patterns indicate potential problems.

Ideally the the points should be randomly spread out above and below zero.

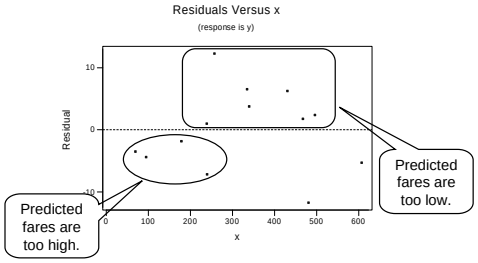


This residual plot indicates no systematic bias using the least squares line to predict the y value.
Generally this is the kind of pattern that you would like to see.

Note:
1.Values below 0 indicate over prediction
2.Values above 0 indicate under prediction.

39

The Greyhound example continued

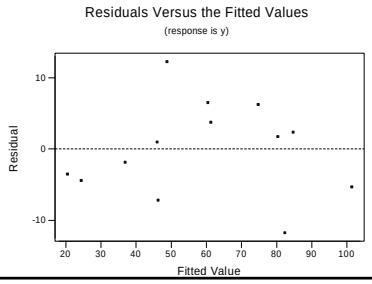


For the Greyhound example, it appears that the line systematically predicts fares that are too high for cities close to Rochester and predicts fares that are too little for most cities between 200 and 500 miles.

40

More Residual Plots

Another common type of **residual plot** is a scatter plot of the data pairs ($\hat{y}$, residual). The following plot was produced by Minitab for the Greyhound data. Notice, that this residual plot shows the same type of systematic problems with the model.



41

Definition formulae

The **total sum of squares**, denoted by **SSTo**, is defined as

$$SSTo = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2$$
$$= \sum (y - \bar{y})^2$$

The **residual sum of squares**, denoted by **SSResid**, is defined as

$$SSResid = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2$$
$$= \sum (y - \hat{y})^2$$

42

Calculational formulae

SSTo and **SSResid** are generally found as part of the standard output from most statistical packages or can be obtained using the following computational formulas:

$$SSTo = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SSResid = \sum y^2 - a\sum y - b\sum xy$$

The coefficient of determination, r^2, can be computed as $r^2 = 1 - \frac{SSResid}{SSTo}$

43

Coefficient of Determination

The **coefficient of determination**, denoted by **r^2**, gives the proportion of variation in y that can be attributed to an approximate linear relationship between x and y.

Note that the coefficient of determination is the square of the Pearson correlation coefficient.

44

Greyhound Example Revisited

$$n = 13, \sum y = 768, \sum y^2 = 53119, \sum xy = 298506$$
$$b = 0.150179 \text{ and } a = 10.1380$$

$$SSTo = \sum y^2 - \frac{(\sum y)^2}{n} = 53119 - \frac{768^2}{13} = \boxed{7747.9}$$

$$SSResid = \sum y^2 - a\sum y - b\sum xy$$
$$= 53119 - 10.1380(768) - 0.150179(298506)$$
$$= 509.117$$

45

Greyhound Example Revisited

$$r^2 = 1 - \frac{SSResid}{SSTo} = 1 - \frac{509.117}{7747.9} = 0.9348$$

.066

We can say that 93.5% of the variation in the Fare (y) can be attributed to the least squares linear relationship between distance (x) and fare.

46

More on variability

The **standard deviation about the least squares line** is denoted s_e and given by

$$s_e = \sqrt{\frac{SSResid}{n-2}}$$

s_e is interpreted as the "typical" amount by which an observation deviates from the least squares line.

47

Greyhound Example Revisited

$$s_e = \sqrt{\frac{SSResid}{n-2}} = \sqrt{\frac{509.117}{11}} = \$6.80$$

The "typical" deviation of actual fare from the prediction is $6.80.

S_e not given on Calculator, but easy to get after running regression:

$S_e = \sqrt{(\text{sum}(\text{Resid}^2)/(n-2))}$

48

Minitab output for Regression

a b

Regression Analysis: Standard Fare versus Distance

The regression equation is
Standard Fare = 10.1 + 0.150 Distance

Least squares regression line

Predictor	Coef	SE Coef	T	P
Constant	10.138	4.327	2.34	0.039
Distance	0.15018	0.01196	12.56	0.000

S = 6.803 R-Sq = 93.5% R-Sq(adj) = 92.9%

S_e r^2

Analysis of Variance

SSResid

Source	DF	SS	MS	F	P
Regression	1	7298.1	7298.1	157.68	0.000
Residual Error	11	509.1	46.3		
Total	12	7807.2			

SSTo

49

The Greyhound problem with additional data

The sample of fares and mileages from Rochester was extended to cover a total of 20 cities throughout the country. The resulting data and a scatterplot are given on the next few slides.

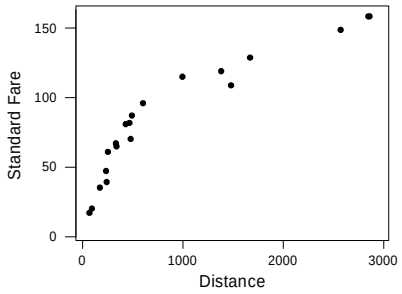
50

Extended Greyhound Fare Example

	Distance	Standard Fare
Buffalo, NY	69	17
New York City	340	65
Cleveland, OH	257	61
Baltimore, MD	430	81
Washington, DC	496	87
Atlanta, GE	998	115
Chicago, IL	607	96
San Francisco	2861	159
Seattle, WA	2848	159
Philadelphia, PA	335	67
Orlando, FL	1478	109
Phoenix, AZ	2569	149
Houston, TX	1671	129
New Orleans, LA	1381	119
Syracuse, NY	95	20
Albany, NY	240	39
Potsdam, NY	239	47
Toronto, ON	178	35
Ottawa, ON	467	82
Montreal, QU	480	70.5

51

Extended Greyhound Fare Example

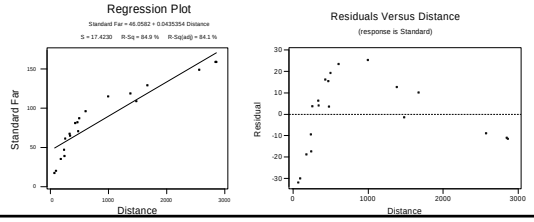


52

Extended Greyhound Fare Example

Minitab reports the correlation coefficient, r=0.921, R^2=0.849, s_e=$17.42 and the regression line Standard Fare = 46.058 + 0.043535 Distance

Notice that even though the correlation coefficient is reasonably high and 84.9 % of the variation in the Fare is explained, the linear model is not very usable.



53

5.4: Nonlinear Regression

From the previous slide we can see that the plot does not look linear, it appears to have a curved shape. We sometimes replace one or both of the variables with a transformation of that variable and then perform a linear regression on the transformed variables. This can sometimes lead to developing a useful prediction equation.

For this particular data, the shape of the curve is almost logarithmic so we might try to replace the distance with $\log_{10}$(distance) [the logarithm to the base 10) of the distance].

54

Nonlinear Regression Example

	Distance	Log_{10}(distance)	Standard Fare
Buffalo, NY	69	1.83885	17
New York City	340	2.53148	65
Cleveland, OH	257	2.40993	61
Baltimore, MD	430	2.63347	81
Washington, DC	496	2.69548	87
Atlanta, GE	998	2.99913	115
Chicago, IL	607	2.78319	96
San Francisco	2861	3.45652	159
Seattle, WA	2848	3.45454	159
Philadelphia, PA	335	2.52504	67
Orlando, FL	1478	3.16967	109
Phoenix, AZ	2569	3.40976	149
Houston, TX	1671	3.22298	129
New Orleans, LA	1381	3.14019	119
Syracuse, NY	95	1.97772	20
Albany, NY	240	2.38021	39
Potsdam, NY	239	2.37840	47
Toronto, ON	178	2.25042	35
Ottawa, ON	467	2.66932	82
Montreal, QU	480	2.68124	70.5

55

Nonlinear Regression Example

Minitab provides the following output.

Regression Analysis: Standard Fare versus Log10(Distance)

```
The regression equation is
Standard Fare = - 163 + 91.0 Log10(Distance)

Predictor         Coef      SE Coef           T        P
Constant      -163.25        10.59      -15.41    0.000
Log10(Di       91.039        3.826       23.80    0.000

S = 7.869       R-Sq = 96.9%     R-Sq(adj) = 96.7%
```

Typical Error = $7.87
Reasonably good

High r^2
96.9% of the variation attributed to the model

56

Nonlinear Regression Example

The rest of the Minitab output follows

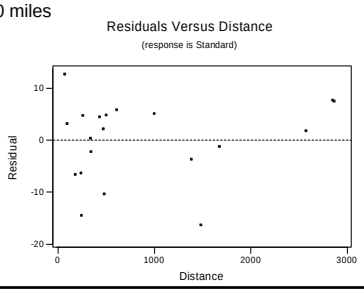
```
Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1       35068       35068    566.30    0.000
Residual Error    18        1115          62
Total             19       36183

Unusual Observations
Obs    Log10(Di    Standard         Fit      SE Fit    Residual    St Resid
 11        3.17      109.00      125.32        2.43      -16.32       -2.18R

R denotes an observation with a large standardized residual
```

The only outlier is Orlando and as you'll see from the next two slides, it is not too bad.

57

Nonlinear Regression Example

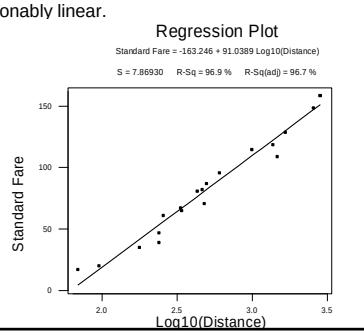
Looking at the plot of the residuals against distance, we see some problems. The model over estimates fares for middle distances (1000 to 2000 miles) and under estimates for longer distances (more than 2000 miles



58

Nonlinear Regression Example

When we look at how the prediction curve looks on a graph that has the Standard Fare and log10(Distance) axes, we see the result looks reasonably linear.

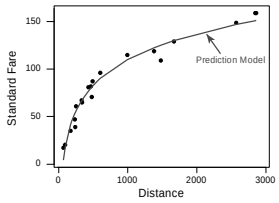


59

Nonlinear Regression Example

When we look at how the prediction curve looks on a graph that has the Standard Fare and Distance axes, we see the result appears to work fairly well.

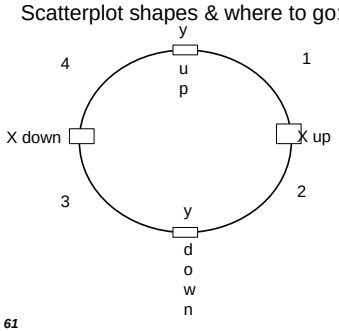
By and large, this prediction model for the fares appears to work reasonable well.



60

Transformations

Scatterplot shapes & where to go:



Power	Transform
3	$(Org)^3$
2	$(Org)^2$
1	(Org)
1/2	$(Org)^{1/2}$
1/3	$(Org)^{1/3}$
0	Log(Org)
-1	1/(Org)

61

5.5: Chapter Wrap-up

•Correlation does not imply causation.

•A correlation near zero does not imply that there is no relationship in the data.

•Least Squares line for predicting x from y is not the same as the least squares line for predicting y from x.

•Beware of extrapolation. Also be careful in interpreting the value of the slope and intercept in the least-squares lines.

•Just because we have a least squares best line, it doesn't mean we have a good prediction equation.

62

Chapter Wrap-up

•It's not enough to just look at r^2 or just look at s_e when looking at the fit of a model. Look at both & the resid plot. If r^2 & s_e look good & plot is off, you may use it, but understand there are systematic problems with your model & there **MAY** be a better model for your data.

•The linear reg model is susceptible to influential obs in the data. Be leery of the model if there is influential data.

•If the relationship between the two variables in non-linear, it is better use a non-linear model than to fit a straight line to the data. Look at the residual plot (may never get plot random). Make sure the model makes sense for the data.

63