UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

Instructor: **Ivo Dinov,**

Asst. Prof. of Statistics and Neurology

Teaching Assistants:

Brandi Shanata & Tiffany Head

University of California, Los Angeles, Fall 2007

http://www.stat.ucla.edu/~dinov/courses_students.html

Chapter 13

Regression & Correlation

Linear Relationships

- Analyze the relationship, if any, between variables x and y by fitting a straight line to the data
  - If a relationship exists we can use our analysis to make predictions
- Data for regression consists of (x,y) pairs for each observation
  - For example: the height and weight of individuals

Lines in 2D

(Regression and Correlation)

Vertical Lines

Horizontal Lines

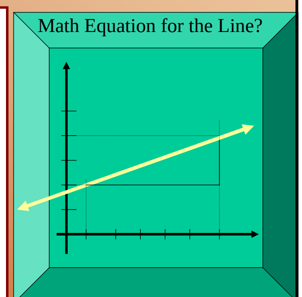
Oblique lines

Increasing/Decreasing

Slope of a line

Intercept

$Y=\alpha X + \beta$, in general.



Lines in 2D

(Regression and Correlation)

Draw the following lines:

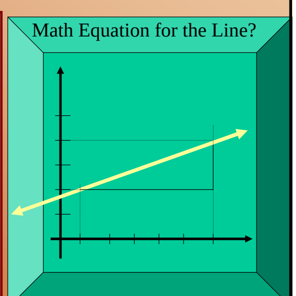
Y=2X+1

Y=-3X-5

Line through (X_1,Y_1) and (X_2,Y_2).

$(Y-Y_1)/(Y_2-Y_1)=$

$(X-X_1)/(X_2-X_1)$.



Correlation Coefficient

Correlation coefficient ($-1<=R<=1$): a measure of linear association, or clustering around a line of multivariate data.

Relationship between two variables (X, Y) can be summarized by: (μ_X, σ_X), (μ_Y, σ_Y) and the correlation coefficient, R. $R=1$, perfect positive correlation (straight line relationship), $R=0$, no correlation (random cloud scatter), $R=-1$, perfect negative correlation.

Computing $R(X,Y)$: (standardize, multiply, average)

$$R(X,Y)=\frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$X=\{x_1, x_2,\dots, x_N,\}$
$Y=\{y_1, y_2,\dots, y_N,\}$
(μ_X, σ_X), (μ_Y, σ_Y)
sample mean / SD.

Correlation Coefficient

Example:

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

Student i	Height x_i	Weight y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	167	60	6	4.67	36	21.8089	28.02
2	170	64	9	8.67	81	75.1689	78.03
3	160	57	-1	1.67	1	2.7889	-1.67
4	152	46	-9	-9.33	81	87.0489	83.97
5	157	55	-4	-0.33	16	0.1089	1.32
6	160	50	-1	-5.33	1	28.4089	5.33
Total	966	332	0	≈ 0	216	215.3334	195.0

Correlation Coefficient

Example:

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$$\mu_x = \frac{966}{6} = 161\,\text{cm}, \qquad \mu_y = \frac{332}{6} = 55\,\text{kg},$$

$$\sigma_x = \sqrt{\frac{216}{5}} = 6.573, \qquad \sigma_y = \sqrt{\frac{215.3}{5}} = 6.563,$$

$$Corr(X,Y) = R(X,Y) = 0.904$$

Correlation Coefficient - Properties

Correlation is invariant w.r.t. linear transformations of X or Y

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) =$$

$$R(aX+b, cY+d), \quad \text{since}$$

$$\left(\frac{ax_k+b-\mu_{ax+b}}{\sigma_{ax+b}}\right) = \left(\frac{ax_k+b-(a\mu_x+b)}{|a|\times\sigma_x}\right) =$$

$$\left(\frac{a(x_k-\mu)+b-b}{a\times\sigma_x}\right) = \left(\frac{x_k-\mu_x}{\sigma_x}\right)$$

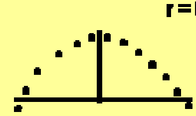
Correlation Coefficient - Properties

Correlation is Associative

$$R(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) = R(Y,X)$$

Correlation measures linear association, NOT an association in general!!! So, Corr(X,Y) could be misleading for X & Y related in a non-linear fashion.

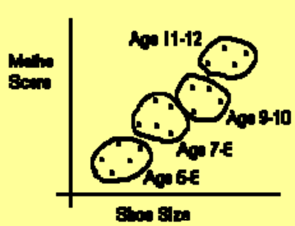
r=0



Correlation Coefficient - Properties

$$R(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) = R(Y,X)$$

1. R measures the extent of linear association between two continuous variables.
2. Association does not imply causation - both variables may be affected by a third variable – age was a confounding variable.



Linear Relationships

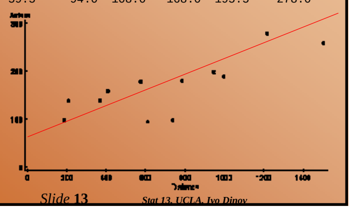
Destination	Distance	Airfare
Atlanta	*576*	*178*
Boston	370	138
Chicago	612	94
Dallas	1216	278
Detroit	409	158
Denver	1502	258
Miami	946	198
New Orleans	998	188
New York	189	98
Orlando	787	179
Pittsburgh	210	138
St. Louis	737	98

Linear Relationships

- Until now we have described data using statistics such as the sample mean

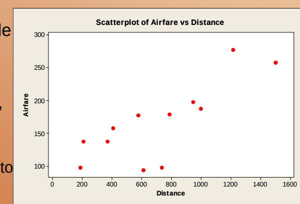
```
Descriptive Statistics: Distance, Airfare

Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
Distance  12   0    713      116    403      189    380     675    985     1502
Airfare   12   0  166.9     17.2   59.5     94.0  108.0   168.0  195.5    278.0
```

- What seems to be missing from this one sample view of the data?

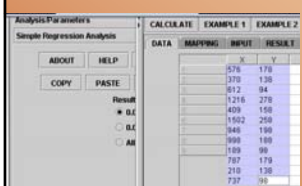


Linear Relationships

- This scatterplot gives us a view of how the dependent variable airfare (y) changes with the independent variable distance (x)
- From this data there appears to be a linear trend, but the data do not fall in an exact straight line
  - Still may be reasonable to fit a line to this data



Linear Relationships
SOCR SLR: socr.ucla.edu/htmls/SOCR_Analyses.html



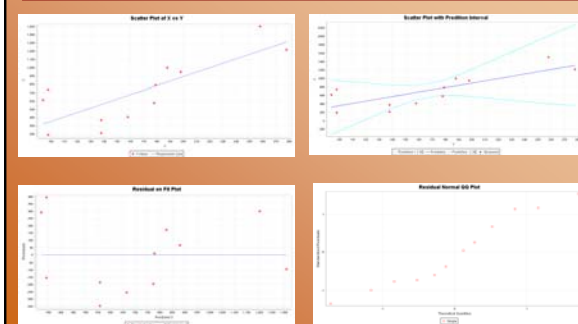
Sample Size = 12
Dependent Variable = X
Independent Variable = Y
Simple Linear Regression Results:
Mean of Y = 166.917
Mean of X = 712.667
Regression Line:
X = -186.090 + 5.384464042692271 Y
Correlation(Y, X) = .795
R-Square = .632
Intercept:

Parameter Estimate: -186.090
Standard Error: 229.137
T-Statistics: -.812
P-Value: .436

Slope:

Parameter Estimate: 5.384
Standard Error: 1.299
T-Statistics: 4.144
P-Value: .002

Linear Relationships
SOCR SLR: socr.ucla.edu/htmls/SOCR_Analyses.html



Linear Relationships

- Two Contexts for regression:

1. Y is an observed variable and X is specified by the researcher
   - Ex. Y is hair growth after 2 months, for individuals at certain dose levels of hair growth cream (X)
2. X and Y are observed variables
   - Ex. Height (Y) and weight (X) for 20 randomly selected individuals

The Fitted Regression Line

- Suppose we have n pairs (x,y)
- If a scatterplot of the data suggests a general linear trend, it would be reasonable to fit a line to the data
- The question is which is the best line?

Example Airfare (cont')

- We can see from the scatterplot that greater distance is associated with higher airfare
- In other words airports that tend to be further from Baltimore than tend to be more expensive airfare

- To decide on the best fitting line, we use the **least-squares method** to fit the least squares (regression) line

Equation of the Regression Line

- RECALL: y = *m*x + *b*
- In statistics we call this Y = $b_0 + b_1X$

where Y is the dependent variable
X is the independent variable

b_0 is the y-intercept $\boxed{\bar{y} - b_1\bar{x}}$

b_1 is the slope of the line $\boxed{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}}$

LS Estimates for the Linear Parameters

1. The least-squares line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ passes through the points (x = 0, $\hat{y}$ = ?) and (x = $\bar{x}$, $\hat{y}$ = ?). Supply the missing values.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Hands – on worksheet !

1. X={-1, 2, 3, 4}, Y={0, -1, 1, 2},

X	Y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$	
-1	0						
2	-1						
3	1						
4	2						

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Hands – on worksheet !

1. X={-1, 2, 3, 4}, Y={0, -1, 1, 2}, $\bar{x} = 2$, $\bar{y} = 0.5$

X	Y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$	
-1	0	-3	-0.5	9	0.25	1.5	
2	-1	0	-1.5	0	2.25	0	
3	1	1	0.5	1	0.25	0.5	
4	2	2	1.5	4	2.25	3	
2	0.5			14	5	5	

β_1=5/14
β_0=y^-β1*x^
β_0= 0.5-10/14

Equation of the Regression Line

- **Example**: Airfare (cont')

```
Regression Analysis: Airfare versus Distance

The regression equation is
Airfare = 83.3 + 0.117 Distance
Predictor      Coef   SE Coef      T       P
Constant      83.27     22.95   3.63   0.005
Distance    0.11738   0.02832   4.14   0.002
S = 37.8270    R-Sq = 63.2%    R-Sq(adj) = 59.5%

Analysis of Variance
Source          DF      SS      MS      F       P
Regression       1   24574   24574  17.17   0.002
Residual Error  10   14309    1431
Total           11   38883
```


Equation of the Regression Line

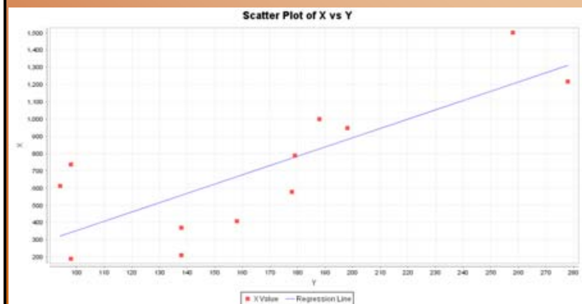
- When we write the least squares regression equation we use the following notation:

$$\hat{y} = 83.27 + 0.117x$$

- b_1 expresses the rate of change of y with respect to x
  - For every one mile increase in distance, airfare will go up by an additional 0.117 dollars.
  - We could actually describe this as for a 100 mile increase in distance airfare rises by $11.70
- b_0 expresses where the regression line will hit the y axis
  - It may or may not be interpretable, depends on the context
  - In this case does an airfare of $83.27 when distance traveled is 0 miles make sense?

Equation of the Regression Line

- NOTE: The least squares line passes through $(\bar{x}, \bar{y})$



Equation of the Regression Line

- Predict the airfare for a city that is 576 miles away. If you look at the original data set (first page), *Atlanta's* distance was 576 miles and the airfare was $178

$$\hat{y} = b_0 + b_1 x$$

= 83.27 + 0.11738(576)

= $150.88 (watch units!)

- Calculate the corresponding residual
  - HOLD that thought

    Residual = 178 – 150.88 = $27.12

Residual Standard Deviation

- The **best** straight line is the one that *minimizes the residual sums of squares*
- The residual standard deviation can be used as our description of the closeness of the data points to the regression line

$$s_{Y|X} = \sqrt{\frac{SS(resid)}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

  - how far off predictions tend to be that are made using the regression model
  - Similar idea to s (measures variability around $\bar{y}$)

    $s_{Y|X}$ (measures variability about the regression line)

Residual Standard Deviation

- Similar interpretation to ch 2.

  68% of our data falls within $\pm$ 1 $s_{Y|X}$ from the line

  95% of our data falls within $\pm$ 2 $s_{Y|X}$ from the line

- We expect most of our data to fall within $2s_{Y|X}$ from the regression line

<u>Example</u>: Airfare (cont') $s_{Y|X} = \sqrt{\frac{SS(resid)}{n-2}} = 37.83$

  - Predictions tend to be off by $37.83
  - Most of our observed values will fall within $\pm$ 2(37.83) = $75.66 from their predicted values.

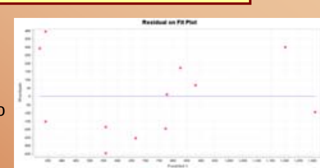
Residual Standard Deviation

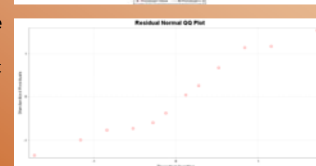
<u>Example</u>: Airfare (cont')

```
Regression Analysis: Airfare versus Distance

The regression equation is
Airfare = 83.3 + 0.117 Distance
Predictor      Coef   SE Coef      T      P
Constant      83.27     22.95   3.63  0.005
Distance    0.11738   0.02832   4.14  0.002
S = 37.8270    R-Sq = 63.2%   R-Sq(adj) = 59.5%

Analysis of Variance

Source          DF     SS     MS      F      P
Regression       1  24574  24574  17.17  0.002
Residual Error  10  14309   1431
Total           11  38883
```


Statistical Inference Concerning β_1

- How can we use statistical inference in regression?
  - Suppose we would like to investigate the relationship between X and Y
  - If X is telling us nothing about Y, what will the slope of the regression line be?
    - In other words, X is not useful for predicting Y





The Standard Error of β_1

- Before we can start with inference we need to discuss the sampling distribution of β_1
- b_1 is our estimate of β_1
  - b_1 will have some sampling error because it is an estimate based on the data
- SE_{b1} is used to describe this variability

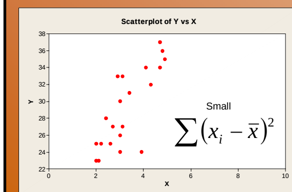
$$SE_{b_1} = \frac{s_{Y|X}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

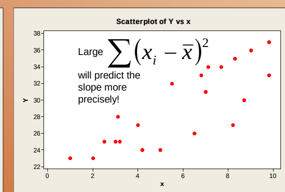
Scatter of data, less scatter about regression line = better estimate of β_1

Spread along x axis, larger = better estimate of β_1

The Standard Error of β_1

- *Two ways to make* $\sum (x_i - \bar{x})^2$ *larger:*
  - Increase n
    - more terms in the summation
  - Increase dispersion in X values
    - more spread on x axis





The Standard Error of β_1

Example: Airfare (cont')

Calculate the standard deviation of the sampling distribution of b_1 (ie. SE_{b1})

We know that $s_{Y|X} = 37.83$

And suppose $\sum (x_i - \bar{x})^2$ was given as 1,786,499

$$SE_{b_1} = \frac{s_{Y|X}}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{37.83}{\sqrt{1{,}786{,}499}} = 0.0283$$

Example:
http://socr.stat.ucla.edu/Applets.dir/RegressionApplet.html

The Standard Error of β_1

Example: Airfare (cont')

```
Regression Analysis: Airfare versus Distance

The regression equation is
Airfare = 83.3 + 0.117 Distance
Predictor     Coef  SE Coef     T      P
Constant     83.27    22.95  3.63  0.005
Distance   0.11738  0.02832  4.14  0.002
S = 37.8270   R-Sq = 63.2%   R-Sq(adj) = 59.5%

Analysis of Variance

Source          DF     SS     MS      F      P
Regression       1  24574  24574  17.17  0.002
Residual Error  10  14309   1431
Total           11  38883
```


The Standard Error of β_1

- In many studies β_1 is a clinically meaningful value (the rate of change for Y with respect to X)
  - Before we define the formula for a CI for β_1 let's remember the formula for a CI for μ

RECALL: $$\bar{y} \pm t(df)_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Where 100(1 - α) is the desired confidence

  - If we pick this apart we are really saying that a CI for μ is: the estimate of μ ± (an appropriate multiplier)x(SE)

The Standard Error of β_1

- Using similar logic:

$$b_1 \pm t(df)_{\alpha/2}\left(SE_{b_1}\right)$$

Where 100(1 - α) is the desired confidence

With df = n – 2

The Standard Error of β_1

Example: Airfare (cont')

Calculate and interpret a 95% confidence interval for the slope

$$b_1 \pm t(df)_{\alpha/2}\left(SE_{b_1}\right)$$
$$= 0.11738 \pm t(10)_{0.025}(0.02832)$$
$$= 0.11738 \pm 2.228(0.02832)$$
$$= (0.054, 0.180)$$

We are highly confident, at the 0.05 level, that the true slope of the regression of airfare on distance is between 0.054 and 0.180 \$/mi

The Standard Error of β_1

- So what does that really mean?
  - In other words, if there is a 1 mile increase in distance the airfare will go up by between \$0.054 and \$0.180.
  - Would the zero rule make sense here?

The Standard Error of β_1

```
Regression Analysis: Airfare versus Distance

The regression equation is
Airfare = 83.3 + 0.117 Distance
Predictor     Coef   SE Coef     T       P
Constant     83.27     22.95  3.63   0.005
Distance   0.11738   0.02832  4.14   0.002

S = 37.8270   R-Sq = 63.2%   R-Sq(adj) = 59.5%

Analysis of Variance
Source          DF     SS     MS      F      P
Regression       1  24574  24574  17.17  0.002
Residual Error  10  14309   1431
Total           11  38883
```


Testing the True Slope β_1

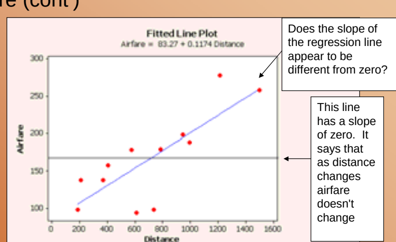
- If X is not useful for predicting Y this is like saying the true slope is zero
- In a hypothesis test our *status quo* null hypothesis would be that there is no relationship between X and Y
- #1 Hypotheses:

  H_o: $\beta_1 = 0$

  H_a: $\beta_1 \,!= 0$ or $\beta_1 > 0$ or $\beta_1 < 0$

Testing the True Slope β_1

Example: Airfare (cont')



Testing the True Slope β_1

- #2 The test statistic: $t_s = \dfrac{b_1 - 0}{SE_{b_1}}$

  with n – 2 df
- #3 P-value
  - based on the t table
  - can be directional or non-directional (multiply by 2)
  - one sided issues still apply
- #4 Conclusion (TBD)

Testing the True Slope β_1

Example: Airfare (cont')

Imagine the population of *all* cities you could fly to from Baltimore

Is the relationship we found in this sample of 12 cities strong enough to convince you that there really is a relationship for the entire population?

Testing the True Slope β_1

Test to see if distance is useful for predicting airfare in a linear model, using α = 0.05

#1 H_o: $\beta_1 = 0$

H_a: β_1 != 0

#2 $$t_s = \frac{b_1 - 0}{SE_{b_1}} = \frac{0.11738 - 0}{0.02832} = 4.145$$

#3 df = 10; 2(0.0005) < p < 2(0.005) = 0.001 < p < 0.01

Reject H_o

Testing the True Slope β_1

#4 CONCLUSION: These data provide evidence to suggest that there is a significant LINEAR relationship between distance and airfare to US cities from Baltimore, MD (0.001 < p < 0.01)

- NOTE: We're not asking if the relationship is linear
- We are already assuming that the linear relationship holds
- Why n – 2 df?
  - It takes two points to determine a straight line
  - Also n – 2 is the denominator of $s_{Y|X}$

Testing the True Slope β_1

```
Regression Analysis: Airfare versus Distance

The regression equation is
Airfare = 83.3 + 0.117 Distance
Predictor      Coef   SE Coef      T       P
Constant      83.27     22.95   3.63   0.005
Distance    0.11738   0.02832   4.14   0.002
S = 37.8270   R-Sq = 63.2%   R-Sq(adj) = 59.5%

Analysis of Variance
Source          DF     SS     MS      F      P
Regression       1  24574  24574  17.17  0.002
Residual Error  10  14309   1431
Total           11  38883
```

Be careful, p-value is for a two sided test!

Testing the True Slope β_1

Suppose we wanted to test to see if the mean airfare increases with increasing distance, using α = 0.05

What would change in our hypothesis test from before?

This means we are expecting a positive slope

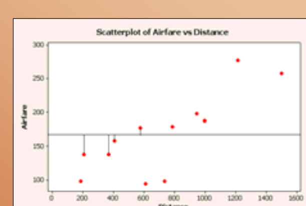
H_a: $\beta_1 > 0$

Does t_s jive with H_a? $t_s = 4.14$

0.0005 < p < 0.005

Variability in Regression

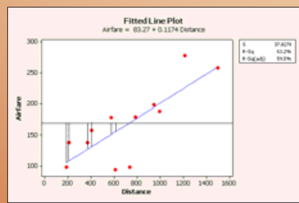
- Consider our airfare example
- The dependent variable, airfare, varies from airport to airport, regardless of distance
  - A statistical measure of the total variability in airfare is called sums of squares total



$$SS(total) = \sum (y_i - \bar{y})^2$$

Variability in Regression

- Suppose we know X_i, the distance for each airport
  - We can use a regression model to predict airfare $\hat{y} = b_0 + b_1 x$
  - Some airports have higher airfares than others and this is partly due to distance
  - The amount of variability in airfare that is explained by a linear regression with distance is called sums of squares regression



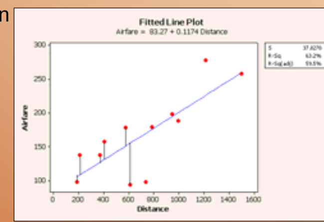
$$SS(reg) = \sum (\hat{y}_i - \bar{y})^2$$

Variability in Regression

- The amount of variation in airfare not explained by distance is called sums of squares residual
  - this is the leftover variability not explained by our regression

$$SS(resid) = \sum (y_i - \hat{y}_i)^2$$

  - NOTE: SS(total) = SS(regr) + SS(resid)
  - total variation = explained variation + unexplained variation



Variability in Regression

```
Regression Analysis: Airfare versus Distance

The regression equation is
Airfare = 83.3 + 0.117 Distance
Predictor      Coef   SE Coef      T      P
Constant      83.27     22.95   3.63  0.005
Distance    0.11738   0.02832   4.14  0.002
S = 37.8270    R-Sq = 63.2%    R-Sq(adj) = 59.5%

Analysis of Variance
Source          DF     SS     MS      F      P
Regression       1  24574  24574  17.17  0.002
Residual Error  10  14309   1431
Total           11  38883
```


Variability in Regression

- NOTE: The sums of Squares appear on minitab in the ANOVA table
  - The balance of the table is the same as we learned for ANOVA, just different formulas

Source	df	SS	MS
Regression	1	$\sum (\hat{y}_i - \bar{y})^2$	$\frac{SS(reg)}{df(reg)}$
Residual	$n^* - 2$	$\sum (y_i - \hat{y}_i)^2$	$\frac{SS(resid)}{df(resid)}$
Total	$n^* - 1$	$\sum (y_i - \bar{y})^2$	

The Coefficient of Determination

- Known as the ratio of SS(reg) to SS(total) (ratio of explained variation over total variation)
- The coefficient of determination is a measure of the strength of the linear relationship between X and Y
  - aka: "The proportion of the variability in Y that is explained by the linear regression of Y on X"
  - simply put this is a measure of the total variability of Y explained by X
- Denoted by R^2

$$R^2 = \frac{SS(reg)}{SS(total)} = 1 - \frac{SS(resid)}{SS(total)}$$

The Coefficient of Determination

- R^2 will always be:

  $0 \le R^2 \le 1$

  If there is no linear relationship between X and Y then R^2 will be close to 0

  If there is a strong linear relationship between X and Y then R^2 will be close to 1

The Coefficient of Determination

Example: Airfare (cont')

Calculate and interpret R^2

$$R^2 = \frac{SS(reg)}{SS(total)} = \frac{24574}{38883} = 0.632$$

Only 63.2% of the total variability in airfare can be explained by a linear regression with distance.

RULE OF THUMB: 81% to 100% indicates a strong linear relationship; 64% to <81% indicates is good; 49% to <64% is fair; and <49% is poor.

NOTE: R^2 close to zero does not mean that there is no relationship between X and Y, only that it is not a linear relationship.

The Coefficient of Determination

```
Regression Analysis: Airfare versus Distance

The regression equation is
Airfare = 83.3 + 0.117 Distance
Predictor      Coef  SE Coef     T      P
Constant      83.27    22.95  3.63  0.005
Distance    0.11738  0.02832  4.14  0.002
S = 37.8270   R-Sq = 63.2%   R-Sq(adj) = 59.5%

Analysis of Variance
Source          DF     SS     MS      F      P
Regression       1  24574  24574  17.17  0.002
Residual Error  10  14309   1431
Total           11  38883
```


The Coefficient of Correlation

- The correlation coefficient is also a measure of the linear relationship between X and Y

  $r = \left(\sqrt{r^2}\right) \times (\text{sign of slope})$ OR $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$

  $-1 \le r \le 1$

  If there is no linear relationship between X and Y then r will be close to 0

  If there is a strong positive linear relationship between X and Y then r will be close to +1

  If there is a strong negative linear relationship between X and Y then r will be close to -1

The Coefficient of Correlation

- **Example**: Airfare (cont')

  Calculate and interpret r

  $$r = \left(\sqrt{0.632}\right) \times (+1) = 0.795$$

  This indicates that distance and airfare have a fair positive linear relationship

  Correlation describes the tightness of the linear relationship between X and Y

  RULE OF THUMB: 0.9 to 1.0 strong linear relationship; 0.8 to <0.9 good; 0.7 to <0.8 fair; <0.7 poor

The Coefficient of Correlation

- Computer output for correlation (e.g., SOCR)

```
Correlations: Airfare, Distance

Pearson corr. of Airfare and Distance = 0.795
P-Value = 0.002
```


The Coefficient of Correlation

- If X and Y are switched the coefficient of correlation will remain unchanged.
- There is statistical inference we can make about r
  - The population correlation coefficient is ρ (rho)
  - Inference about requires a bivariate random sample – each (x, y) as having been sampled at random from a population of all (x, y) pairs
  - NOTE: Won't work when X is specified by researcher (doses)
  - It turns out that H_o: $\rho = 0$ is equivalent to H_o: $\beta_1 = 0$

Guidelines for Regression and Correlation

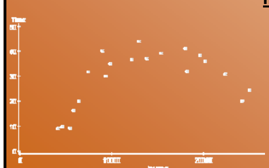
- Need to be careful interpreting correlation
  - Similar to Ch 8, an observed association between variables does not necessarily indicate causation
  - Because two variables are highly correlated does not mean that one causes the other.

Curvilinear Data

- Curvilinear data can distort regression results by:
  - a fitted line that doesn't represent the data
  - the correlation is misleadingly small
  - $s_{Y|X}$ is inflated

Curvilinear Data

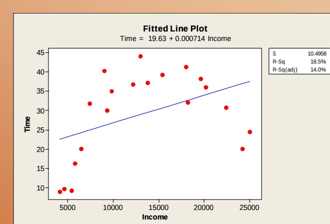
Example: For married couples with one or more offspring, a demographic study was conducted to determine the effect of the families annual income (at marriage) on time (months) between marriage and the birth of the first child.



Income	Time	Income	Time
5775	16.20	4608	9.70
9800	35.00	24210	20.00
13795	37.20	19625	38.20
4120	9.00	18000	41.25
25015	24.40	13000	44.00
12200	36.75	5400	9.20
7400	31.75	6440	20.00
9340	30.00	9000	40.20
20170	36.00	18180	32.00
22400	30.80	15385	39.20

Curvilinear Data

- Clearly a straight line model does not accurately describe what is going on with this data.
- Does this mean there is no relationship between income and time?
  - No, just that it isn't linear!

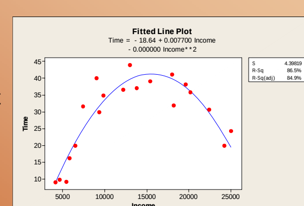


Curvilinear Data

```
Regression Analysis: Time versus Income

The regression equation is
Time = 19.6 + 0.000714 Income
Predictor       Coef     SE Coef     T       P
Constant       19.626       5.213  3.76  0.001
Income      0.0007138  0.0003528  2.02  0.058
S = 10.4958    R-Sq = 18.5%    R-Sq(adj) = 14.0%

Analysis of Variance
Source          DF       SS      MS      F      P
Regression       1    450.9   450.9   4.09  0.058
Residual Error  18   1982.9   110.2
Total           19   2433.8
```


Curvilinear Data

- Our solution would be to fit a quadratic model to address the curvature seen in the scatter plot
  - The graph shows that visually we have a good fit with a quadratic model
  - NOTE: Now that we have more than one independent variable this becomes a multiple regression problem



Curvilinear Data

```
Regression Analysis: Time versus Income, IncomeSQ

The regression equation is
Time = - 18.6 + 0.00770 Income - 0.000000 IncomeSQ
Predictor         Coef      SE Coef       T       P
Constant       -18.639        4.679   -3.98   0.001
Income       0.0077004    0.0007699   10.00   0.000
IncomeSQ   -0.00000025   0.00000003   -9.25   0.000
S = 4.39819    R-Sq = 86.5%    R-Sq(adj) = 84.9%

Analysis of Variance
Source          DF      SS       MS       F       P
Regression       2  2104.9   1052.5   54.41   0.000
Residual Error  17   328.8     19.3
Total           19  2433.8
```

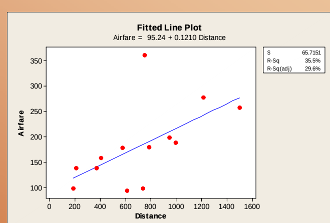

Outliers

- We know outliers as observations that are unusually large when compared to the rest of the data
- In regression analysis an outlier is a data points that is unusually far from the linear trend formed by the data
- Outliers can distort regression results by:
  - inflating $s_{Y|X}$ and reducing r
  - influencing the regression line

Outliers

Example: Airfare

Suppose we an airport to our data set which is 750 miles away with an airfare of $361



Outliers

```
Regression Analysis: Airfare versus Distance
The regression equation is
Airfare = 95.2 + 0.121 Distance
Predictor     Coef  SE Coef     T      P
Constant     95.24    39.63  2.40  0.035
Distance   0.12104  0.04919  2.46  0.032
S = 65.7151    R-Sq = 35.5%    R-Sq(adj) = 29.6%

Analysis of Variance
Source          DF     SS     MS     F      P
Regression       1  26150  26150  6.06  0.032
Residual Error  11  47503   4318
Total           12  73654
Unusual Observations
Obs  Distance  Airfare    Fit  SE Fit  Residual  St Resid
 13       750    361.0  186.0    18.3     175.0     2.77R
R denotes an observation with a large standardized residual.
```


Influential Observations

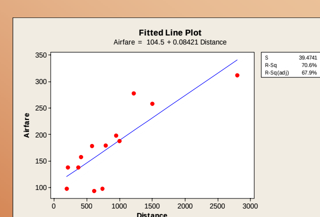
- Influential observations also affect regression results, usually in an artificially positive way
- Influential observations can distort regression results by:
  - changing fitted line
  - influences correlation

Influential Observations

Example: Airfare (cont')

Suppose we an airport to our data set which is 2800 miles away with an airfare of $312

NOTE: Not an outlier because the residual is small



Influential Observations

```
Regression Analysis: Airfare versus Distance
The regression equation is
Airfare = 105 + 0.0842 Distance
Predictor      Coef  SE Coef     T      P
Constant     104.54    18.01  5.80  0.000
Distance    0.08421  0.01638  5.14  0.000
S = 39.4741   R-Sq = 70.6%   R-Sq(adj) = 67.9%

Analysis of Variance
Source          DF     SS     MS      F      P
Regression       1  41173  41173  26.42  0.000
Residual Error  11  17140   1558
Total           12  58313

Unusual Observations
Obs  Distance  Airfare    Fit  SE Fit  Residual  St Resid
 13      2800    312.0  340.3    33.4     -28.3     -1.35 X

X denotes an observation whose X value gives it large influence.
```


Conditions for Inference

- Design conditions:
  - Random subsampling model: for each x the corresponding y is viewed as randomly chosen from the conditional population distribution of Y values
  - Bivariate random sampling model: each (x,y) pair is viewed as randomly chosen
- Conditions concerning parameters
  - $\mu_{Y|X} = \beta_0 + \beta_1 X$
  - $\sigma_{Y|X}$ does not depend on X
- Conditions concerning population distribution: the conditional distribution of Y for each fixed X is normally distributed

Conditions for Inference

- SUMMARY:
  - Same SD, for all levels of X
  - Independent observations
  - Normal distribution of Y for each fixed X
  - Random sample

Multiple Regression

- Regression can be quite complicated
- Multiple regression is an extension of simple linear regression
  - Does distance completely determine airfare?
  - Are there other factors that might influence airfare?
- There are multiple regression models that can accommodate more than one independent variable
  - These topics are covered in other statistics classes.