Research Methods

William G. Zikmund

Bivariate Analysis - Tests of Differences

Common Bivariate Tests

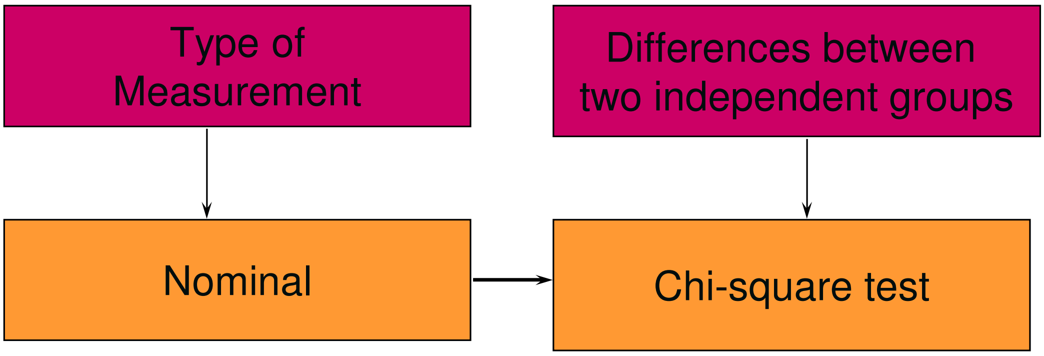
Type of Measurement	Differences between two independent groups	Differences among three or more independent groups
Interval and ratio	Independent groups: *t*-test or *Z*-test	One-way ANOVA

Common Bivariate Tests

Type of Measurement	Differences between two independent groups	Differences among three or more independent groups
Ordinal	Mann-Whitney U-test Wilcoxon test	Kruskal-Wallis test

Common Bivariate Tests

Type of Measurement	Differences between two independent groups	Differences among three or more independent groups
Nominal	*Z*-test (two proportions) Chi-square test	Chi-square test



Differences Between Groups

- Contingency Tables
- Cross-Tabulation
- Chi-Square allows testing for significant differences between groups
- “Goodness of Fit”

Chi-Square Test

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

x^2 = chi-square statistics
O_i = observed frequency in the i^{th} cell
E_i = expected frequency on the i^{th} cell

Chi-Square Test

$$E_{ij} = \frac{R_i C_j}{n}$$

R_i = total observed frequency in the i^{th} row
C_j = total observed frequency in the j^{th} column
n = sample size

Degrees of Freedom

$$(R-1)(C-1)=(2-1)(2-1)=1$$

Degrees of Freedom

$$d.f.=(R-1)(C-1)$$

Awareness of Tire Manufacturer's Brand

	Men	Women	Total
Aware	50	10	60
Unaware	15	25	40
	65	35	100

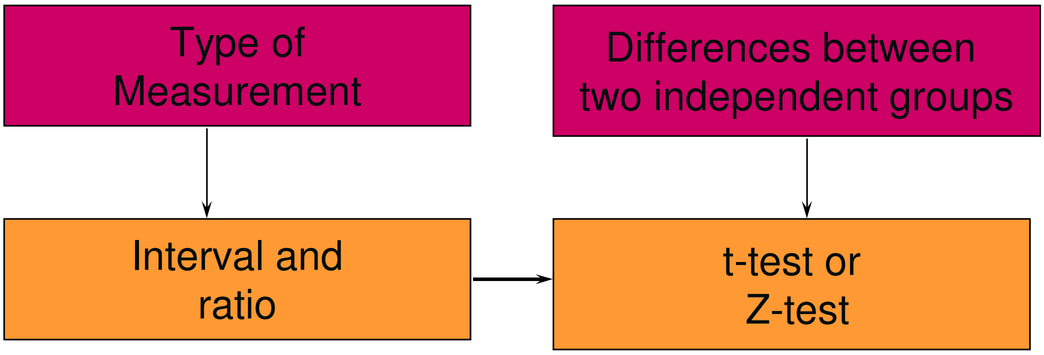
Chi-Square Test: Differences Among Groups Example

$$X^2 = \frac{(50-39)^2}{39} + \frac{(10-21)^2}{21} + \frac{(15-26)^2}{26} + \frac{(25-14)^2}{14}$$

$$\chi^2 = 3.102 + 5.762 + 4.654 + 8.643 =$$
$$\chi^2 = 22.161$$

$$d.f. = (R-1)(C-1)$$
$$d.f. = (2-1)(2-1) = 1$$

X^2=3.84 with 1 d.f.



Differences Between Groups when Comparing Means

- Ratio scaled dependent variables
- t-test
  - When groups are small
  - When population standard deviation is unknown
- z-test
  - When groups are large

Null Hypothesis About Mean Differences Between Groups

$$\mu_1 - \mu_2$$

$$OR$$

$$\mu_1 - \mu_2 = 0$$

t-Test for Difference of Means

$$t = \frac{\text{mean 1 - mean 2}}{\text{Variability of random means}}$$

t-Test for Difference of Means

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}}$$

$\bar{X}_1$ = mean for Group 1
$\bar{X}_2$ = mean for Group 2
$S_{\bar{X}_1-\bar{X}_2}$ = the pooled or combined standard error of difference between means.

t-Test for Difference of Means

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}}$$

t-Test for Difference of Means

$\bar{X}_1$ = mean for Group 1
$\bar{X}_2$ = mean for Group 2
$S_{\bar{X}_1 - \bar{X}_2}$ = the pooled or combined standard error of difference between means.

Pooled Estimate of the Standard Error

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left(\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2)}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Pooled Estimate of the Standard Error

S_1^2 = the variance of Group 1
S_2^2 = the variance of Group 2
n_1 = the sample size of Group 1
n_2 = the sample size of Group 2

Pooled Estimate of the Standard Error

t-test for the Difference of Means

$$S_{\bar{X}_1-\bar{X}_2} = \sqrt{\left(\frac{(n_1-1)S_1^2+(n_2-1)S_2^2)}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

S_1^2 = the variance of Group 1
S_2^2 = the variance of Group 2
n_1 = the sample size of Group 1
n_2 = the sample size of Group 2

Degrees of Freedom

- d.f. = n - k
- where:
  - n = $n_1 + n_2$
  - k = number of groups

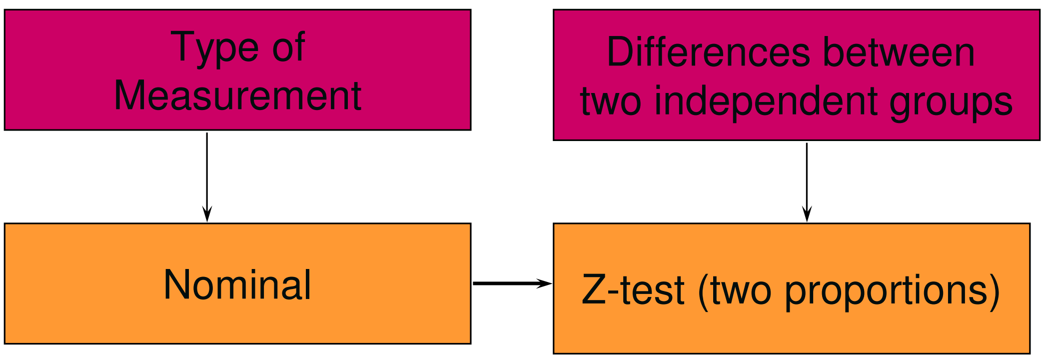
t-Test for Difference of Means Example

$$S_{\bar{X}_1-\bar{X}_2} = \sqrt{\left(\frac{(20)(2.1)^2 + (13)(2.6)^2}{33}\right)\left(\frac{1}{21}+\frac{1}{14}\right)}$$

$$= .797$$

$$t = \frac{16.5 - 12.2}{.797} \quad = \frac{4.3}{.797}$$

$$= 5.395$$



Comparing Two Groups when Comparing Proportions

- Percentage Comparisons
- Sample Proportion - P
- Population Proportion - Π

Differences Between Two Groups when Comparing Proportions

The hypothesis is:

$$H_o: \Pi_1 = \Pi_2$$

may be restated as:

$$H_o: \Pi_1 - \Pi_2 = 0$$

Z-Test for Differences of Proportions

$$H_o : \pi_1 = \pi_2$$

or

$$H_o : \pi_1 - \pi_2 = 0$$

Z-Test for Differences of Proportions

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{S_{p_1 - p_2}}$$

Z-Test for Differences of Proportions

p_1 = sample portion of successes in Group 1
p_2 = sample portion of successes in Group 2
$(\pi_1 - \pi_1)$ = hypothesized population proportion 1 minus hypothesized population proportion 1 minus
$S_{p1\text{-}p2}$ = pooled estimate of the standard errors of difference of proportions

Z-Test for Differences of Proportions

$$S_{p_1 - p_2} = \sqrt{\overline{p}\,\overline{q}\left(\frac{1}{n_1} - \frac{1}{n_2}\right)}$$

Z-Test for Differences of Proportions

$\bar{p}$ = pooled estimate of proportion of success in a sample of both groups

$\bar{q}$ = (1- $\bar{p}$) or a pooled estimate of proportion of failures in a sample of both groups

n_1= sample size for group 1

n_2= sample size for group 2

Z-Test for Differences of Proportions

$$\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Z-Test for Differences of Proportions

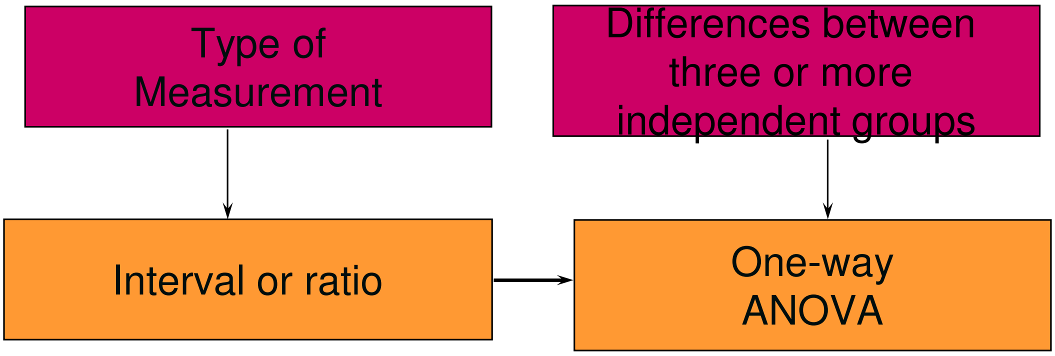
$$S_{p_1 - p_2} = \sqrt{(.375)(.625)\left(\frac{1}{100} + \frac{1}{100}\right)}$$

$$= .068$$

A Z-Test for Differences of Proportions

$$\overline{p} = \frac{(100)(.35) + (100)(.4)}{100 + 100}$$

$$= .375$$



Analysis of Variance

Hypothesis when comparing three groups

$$\mu_1 = \mu_2 = \mu_3$$

Analysis of Variance
F-Ratio

$$F = \frac{Variance - between - groups}{Variance - within - groups}$$

Analysis of Variance
Sum of Squares

$$\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_{\mathrm{within}} + \mathrm{SS}_{\mathrm{between}}$$

Analysis of Variance
Sum of SquaresTotal

$$SS_{\text{total}} = \sum_{i=1}^{n} \sum_{j=1}^{c} (X_{ij} - \bar{\bar{X}})^2$$

Analysis of Variance
Sum of Squares

X_{ij}= individual scores, i.e., the i^{th} observation or test unit in the j^{th} group

$\bar{\bar{X}}$ = grand mean

n = number of all observations or test units in a group

c = number of j^{th} groups (or columns)

Analysis of Variance
Sum of Squares Within

$$SS_{\text{within}} = \sum_{i=1}^{n} \sum_{j=1}^{c} (\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_j)^2$$

Analysis of Variance
Sum of Squares Within

X_{ij} = individual scores, i.e., the i^{th} observation or test unit in the j^{th} group

$\overline{\overline{X}}$ = grand mean

n = number of all observations or test units in a group

c = number of j^{th} groups (or columns)

Analysis of Variance
Sum of Squares Between

$$SS_{\text{between}} = \sum_{j=1}^{n} n_j (\bar{X}_j - \bar{\bar{X}})^2$$

Analysis of Variance
Sum of squares Between

X_j = individual scores, i.e., the i^{th} observation or test unit in the j^{th} group

$\bar{\bar{X}}$ = grand mean

n_j = number of all observations or test units in a group

Analysis of Variance
Mean Squares Between

$$MS_{between} = \frac{SS_{between}}{c-1}$$

Analysis of Variance
Mean Square Within

$$MS_{within} = \frac{SS_{within}}{cn-c}$$

Analysis of Variance
F-Ratio

$$F = \frac{MS_{between}}{MS_{within}}$$

A Test Market Experiment on Pricing

Sales in Units (thousands)

	Regular Price $.99	Reduced Price $.89	Cents-Off Coupon Regular Price
Test Market A, B, or C	130	145	153
Test Market D, E, or F	118	143	129
Test Market G, H, or I	87	120	96
Test Market J, K, or L	84	131	99
Mean	X_1=104.75	X_2=134.75	X_1=119.25
Grand Mean	X=119.58		

ANOVA Summary Table Source of Variation

- Between groups
- Sum of squares
  - SSbetween
- Degrees of freedom
  - c-1 where c=number of groups
- Mean squared-MSbetween
  - SSbetween/c-1

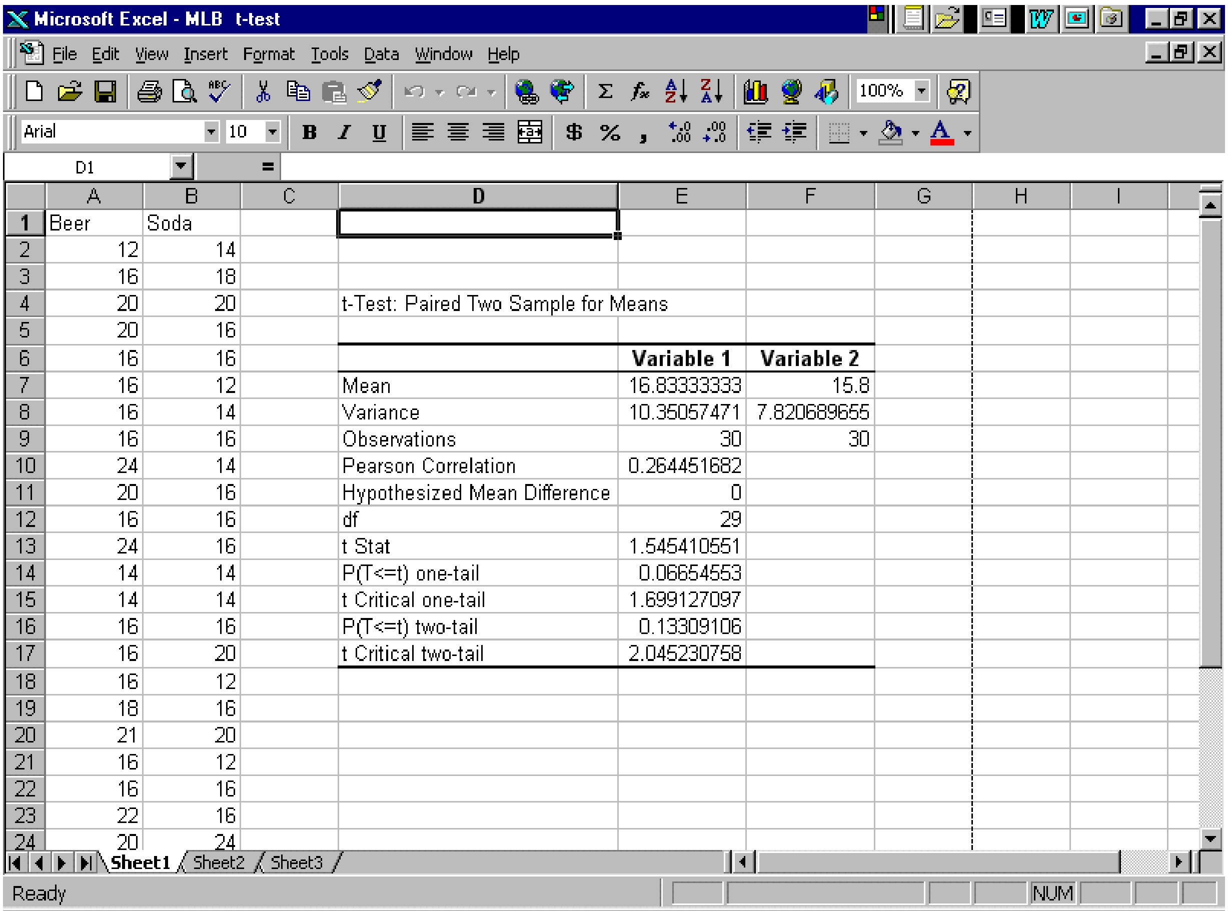
ANOVA Summary Table Source of Variation

- Within groups
- Sum of squares
  - SSwithin
- Degrees of freedom
  - cn-c where c=number of groups, n= number of observations in a group
- Mean squared-MSwithin
  - SSwithin/cn-c

ANOVA Summary Table
Source of Variation

- Total
- Sum of Squares
  - SStotal
- Degrees of Freedom
  - cn-1 where c=number of groups, n= number of observations in a group

$$F = \frac{MS_{BETWEEN}}{MS_{WITHIN}}$$



	A	B	C	D	E	F
1	Beer	Soda				
2	12	14				
3	16	18				
4	20	20		t-Test: Paired Two Sample for Means		
5	20	16				
6	16	16			**Variable 1**	**Variable 2**
7	16	12		Mean	16.83333333	15.8
8	16	14		Variance	10.35057471	7.820689655
9	16	16		Observations	30	30
10	24	14		Pearson Correlation	0.264451682	
11	20	16		Hypothesized Mean Difference	0	
12	16	16		df	29	
13	24	16		t Stat	1.545410551	
14	14	14		P(T<=t) one-tail	0.06654553	
15	14	14		t Critical one-tail	1.699127097	
16	16	16		P(T<=t) two-tail	0.13309106	
17	16	20		t Critical two-tail	2.045230758	
18	16	12				
19	18	16				
20	21	20				
21	16	12				
22	16	16				
23	22	16				
24	20	24				

Research Methods

William G. Zikmund

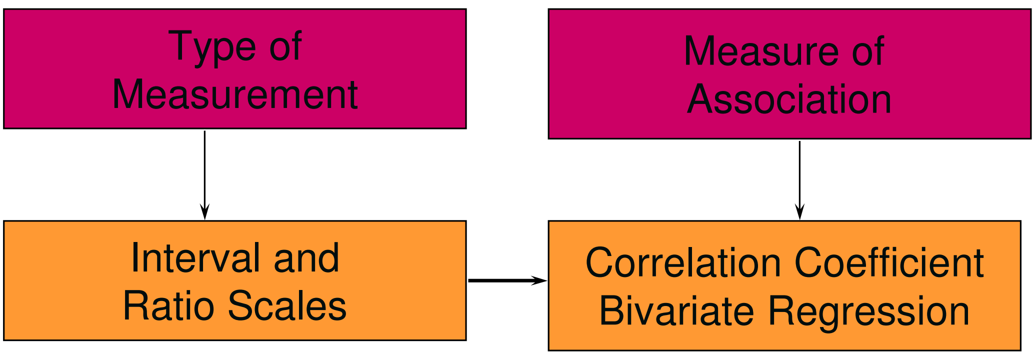
Bivariate Analysis: Measures of Associations

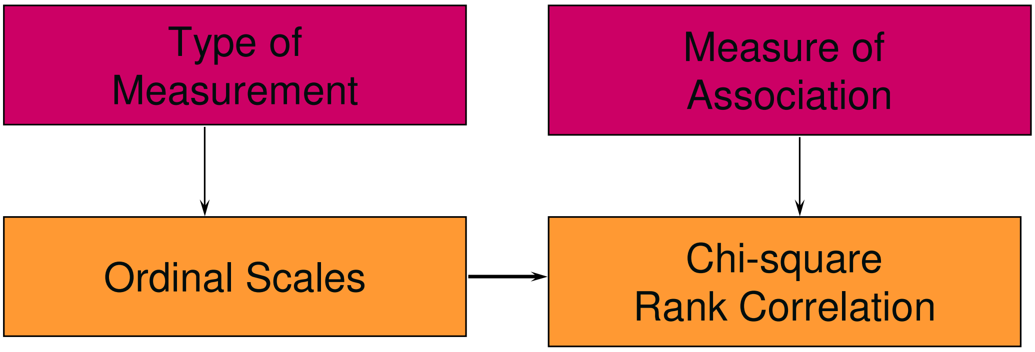
Measures of Association

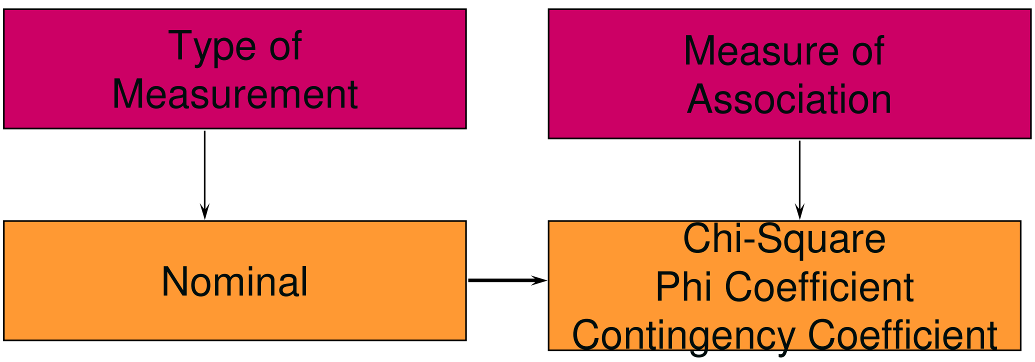
- A general term that refers to a number of bivariate statistical techniques used to measure the strength of a relationship between two variables.

Relationships Among Variables

- Correlation analysis
- Bivariate regression analysis




Type of Measurement
Measure of Association
Ordinal Scales
Chi-square
Rank Correlation


Type of Measurement
Measure of Association
Nominal
Chi-Square
Phi Coefficient
Contingency Coefficient

Correlation Coefficient

- A statistical measure of the covariation or association between two variables.
- Are dollar sales associated with advertising dollar expenditures?

The Correlation coefficient for two variables, X and Y is

$$r_{xy}$$

Correlation Coefficient

- r
- r ranges from +1 to -1
- r = +1 a perfect positive linear relationship
- r = -1 a perfect negative linear relationship
- r = 0 indicates no correlation

Simple Correlation Coefficient

$$r_{xy} = r_{yx} = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\left(Xi - \overline{X}\right)^2 \sum\left(Yi - \overline{Y}\right)^2}}$$

Simple Correlation Coefficient

$$r_{xy} = r_{yx} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

Simple Correlation Coefficient Alternative Method

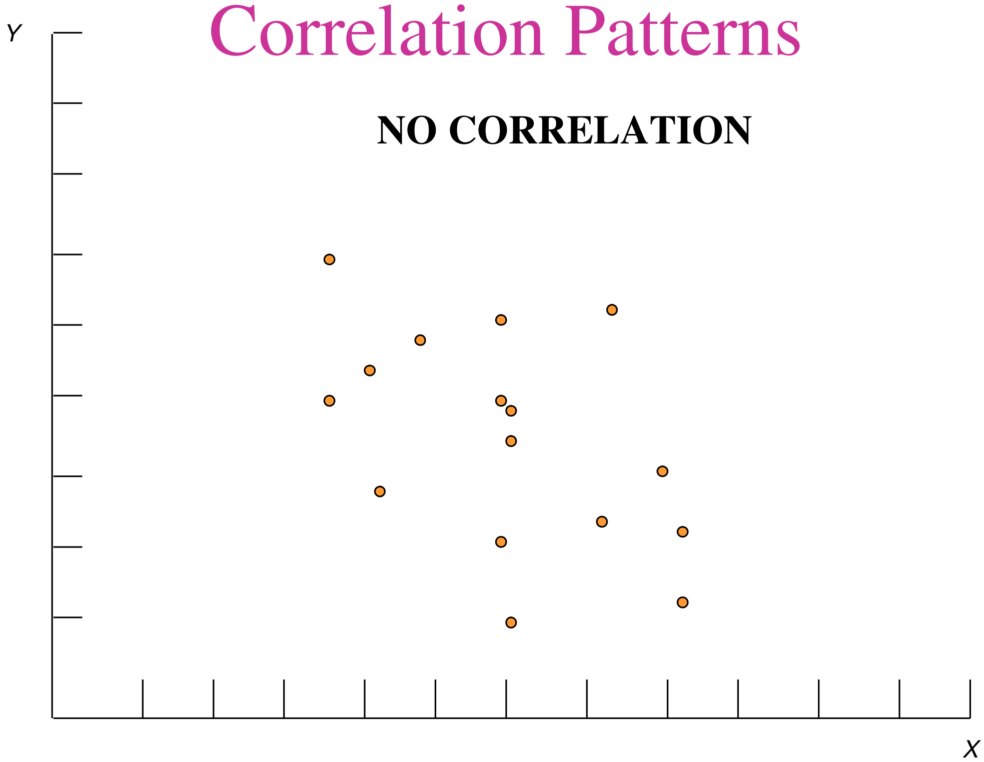
σ_x^2 = Variance of X

σ_y^2 = Variance of Y

σ_{xy}= Covariance of X and Y

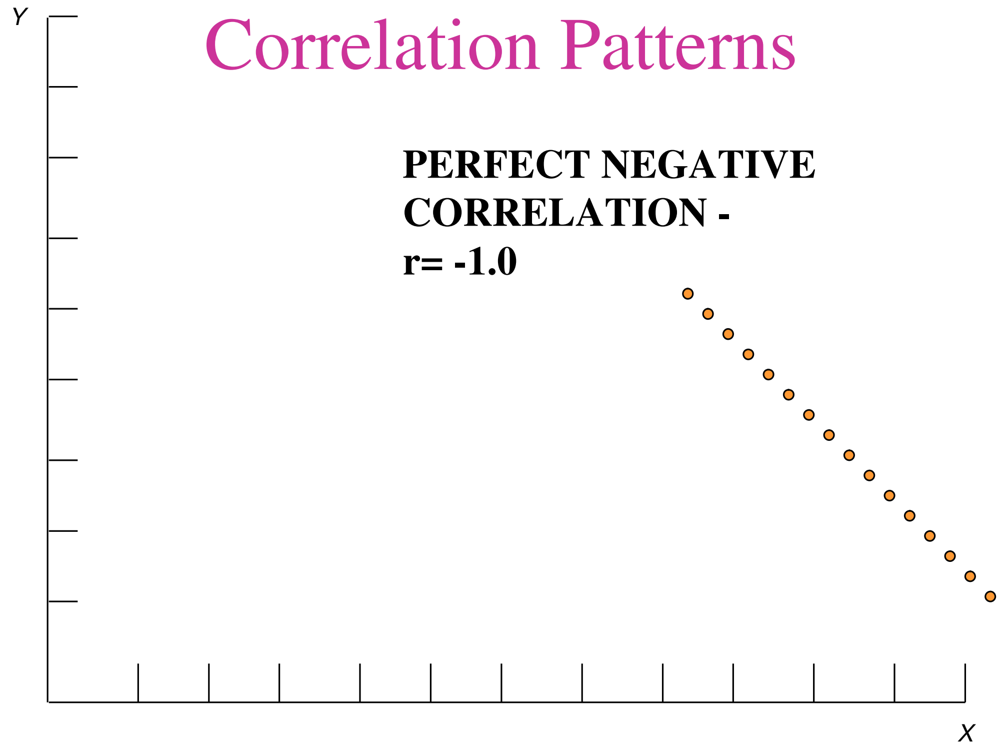
Correlation Patterns

NO CORRELATION

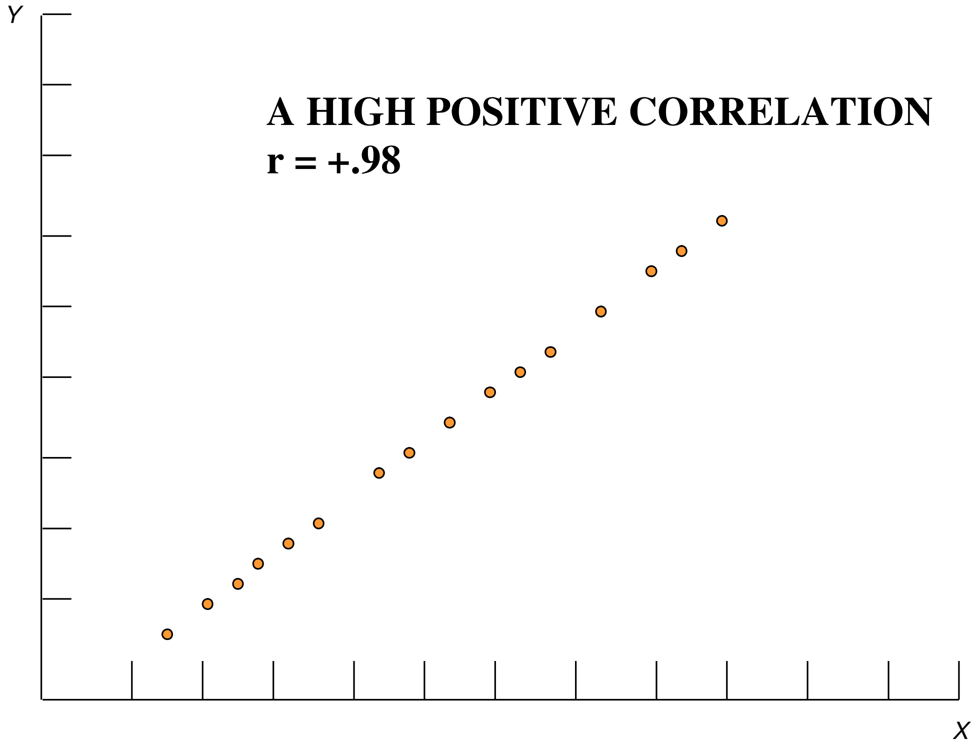


Correlation Patterns

PERFECT NEGATIVE CORRELATION - r= -1.0



Correlation Patterns



Health Economics Research Method 2003/2.

Calculation of r

$$r = \frac{-6.3389}{\sqrt{(17.837)(5.589)}}$$

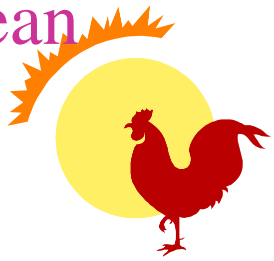
$$= \frac{-6.3389}{\sqrt{99.712}} \qquad = -.635$$

Pg 629

Coefficient of Determination

$$r^2 = \frac{Explained \ \text{variance}}{Total \ \text{Variance}}$$

Correlation Does Not Mean Causation



- High correlation
- Rooster's crow and the rising of the sun
  - Rooster does not cause the sun to rise.
- Teachers' salaries and the consumption of liquor
  - Covary because they are both influenced by a third variable

Correlation Matrix

- The standard form for reporting correlational results.

Correlation Matrix

	Var1	Var2	Var3
Var1	1.0	0.45	0.31
Var2	0.45	1.0	0.10
Var3	0.31	0.10	1.0

Walkup's First Laws of Statistics

- Law No. 1
  - Everything correlates with everything, especially when the same individual defines the variables to be correlated.
- Law No. 2
  - It won't help very much to find a good correlation between the variable you are interested in and some other variable that you don't understand any better.

Walkup's First Laws of Statistics

- Law No. 3
  - Unless you can think of a logical reason why two variables should be connected as cause and effect, it doesn't help much to find a correlation between them. In Columbus, Ohio, the mean monthly rainfall correlates very nicely with the number of letters in the names of the months!

Regression

DICTIONARY DEFINITION → **GOING OR MOVING BACKWARD**

Going back to previous conditions

- Tall men's sons

Bivariate Regression

- A measure of linear association that investigates a straight line relationship
- Useful in forecasting

Bivariate Linear Regression

- A measure of linear association that investigates a straight-line relationship
- $Y = a + bX$
- where
- Y is the dependent variable
- X is the independent variable
- a and b are two constants to be estimated

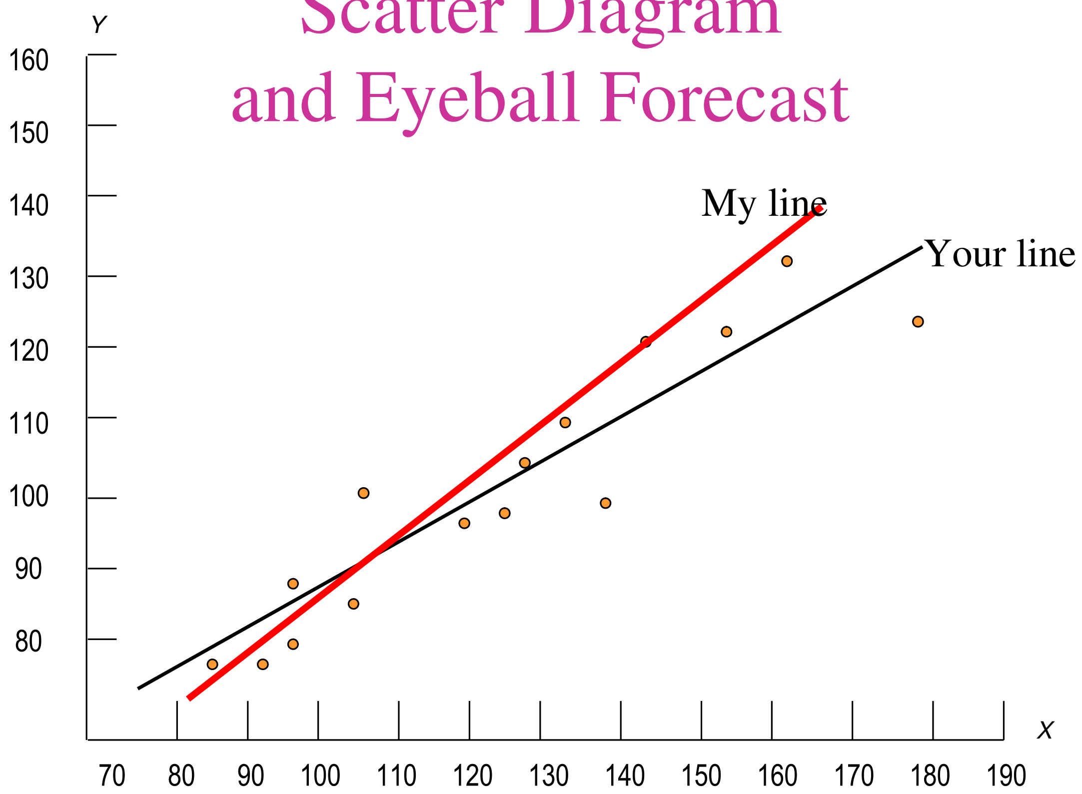
Y intercept

- a
- An intercepted segment of a line
- The point at which a regression line intercepts the Y-axis

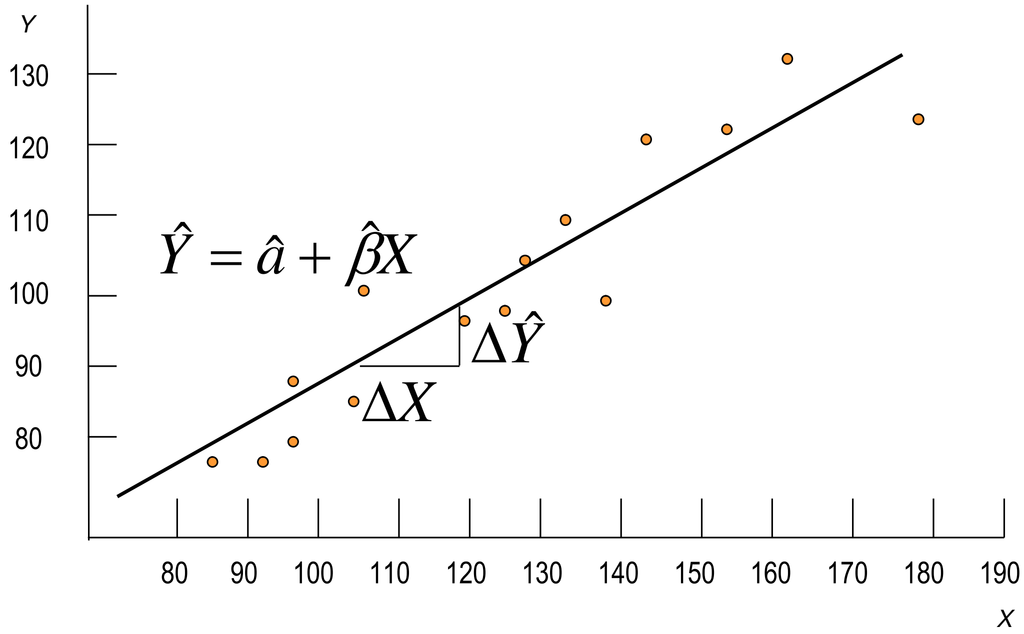
Slope

- b
- The inclination of a regression line as compared to a base line
- Rise over run
- D - notation for “a change in”

Scatter Diagram and Eyeball Forecast

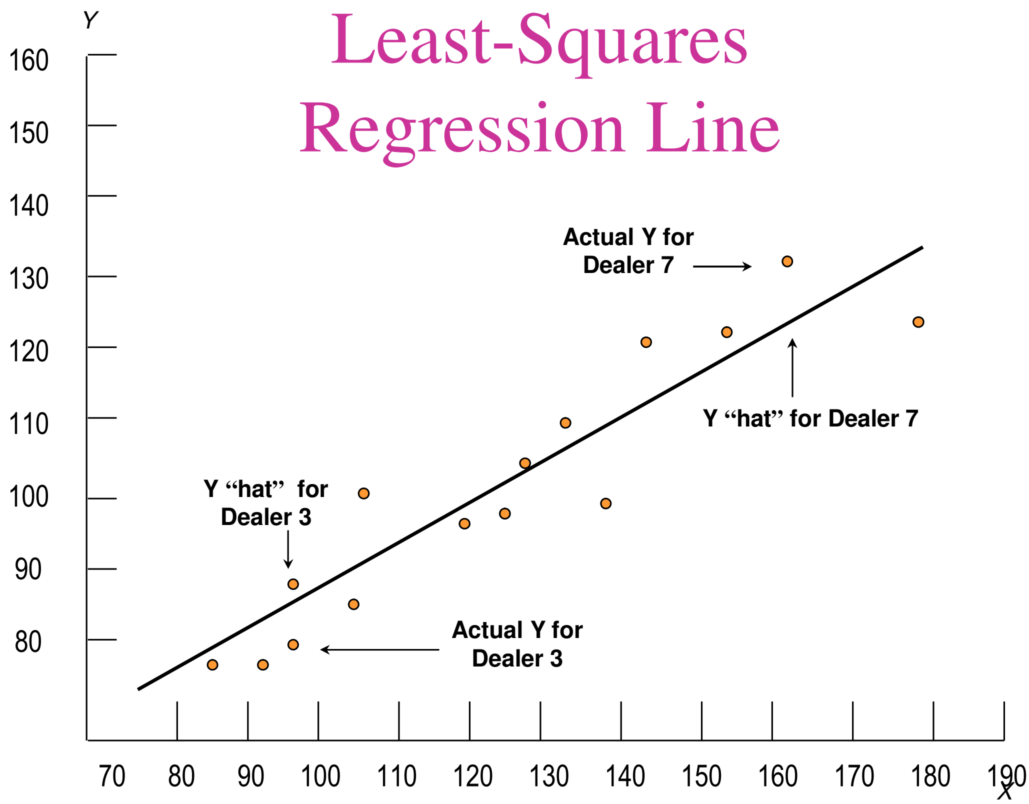


Regression Line and Slope

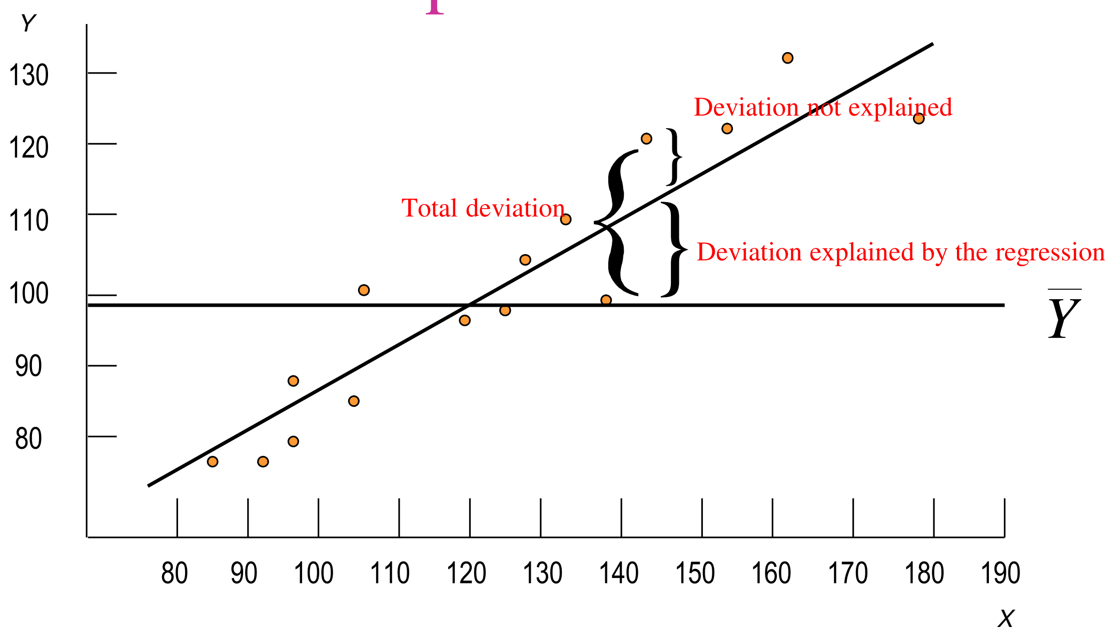


Health Economics Research Method 2003/2.

Least-Squares Regression Line



Scatter Diagram of Explained and Unexplained Variation



The Least-Square Method

- Uses the criterion of attempting to make the least amount of total error in prediction of Y from X. More technically, the procedure used in the least-squares method generates a straight line that minimizes the sum of squared deviations of the actual values from this predicted regression line.

The Least-Square Method

- A relatively simple mathematical technique that ensures that the straight line will most closely represent the relationship between X and Y.

Regression - Least-Square Method

$$\sum_{i=1}^{n} e_i^2 \text{ is minimum}$$

$$e_i = Y_i - \hat{Y}_i \quad \text{(The "residual")}$$

Y_i = actual value of the dependent variable

$\hat{Y}_i$ = estimated value of the dependent variable (Y hat)

n = number of observations

i = number of the observation

The Logic behind the Least-Squares Technique

- No straight line can completely represent every dot in the scatter diagram
- There will be a discrepancy between most of the actual scores (each dot) and the predicted score
- Uses the criterion of attempting to make the least amount of total error in prediction of Y from X

Bivariate Regression

$$\hat{a} = \overline{Y} - \hat{\beta}\overline{X}$$

Bivariate Regression

$$\hat{\beta} = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^{2}\right) - \left(\sum X\right)^{2}}$$

$\hat{\beta}$ = estimated slope of the line (the "regression coefficient")

$\hat{a}$ = estimated intercept of the *y* axis

Y = dependent variable

$\bar{Y}$ = mean of the dependent variable

X = independent variable

$\bar{X}$ = mean of the independent variable

n = number of observations

$$\hat{\beta} = \frac{15(193{,}345) - 2{,}806{,}875}{15(245{,}759) - 3{,}515{,}625}$$

$$= \frac{2{,}900{,}175 - 2{,}806{,}875}{3{,}686{,}385 - 3{,}515{,}625}$$

$$= \frac{93{,}300}{170{,}760} \quad = .54638$$

$$\hat{a} = 99.8 - .54638(125)$$

$$= 99.8 - 68.3$$

$$= 31.5$$

$$\hat{a} = 99.8 - .54638(125)$$

$$= 99.8 - 68.3$$

$$= 31.5$$

$$\begin{aligned}\hat{Y} &= 31.5 + .546(X) \\ &= 31.5 + .546(89) \\ &= 31.5 + 48.6 \\ &= 80.1\end{aligned}$$

$$\begin{aligned}\hat{Y} &= 31.5 + .546(X) \\ &= 31.5 + .546(89) \\ &= 31.5 + 48.6 \\ &= 80.1\end{aligned}$$

Dealer 7 (Actual Y value $= 129$)

$$\hat{Y}_7 = 31.5 + .546(165)$$
$$= 121.6$$

Dealer 3 (Actual Y value $= 80$)

$$\hat{Y}_3 = 31.5 + .546(95)$$
$$= 83.4$$

$$e_i = Y_9 - \hat{Y}_9$$
$$= 97 - 96.5$$
$$= 0.5$$

$$\text{Dealer 7 (Actual Y value} = 129)$$

$$\hat{Y}_7 = 31.5 + .546(165)$$
$$= 121.6$$

$$\text{Dealer 3 (Actual Y value} = 80)$$

$$\hat{Y}_3 = 31.5 + .546(95)$$
$$= 83.4$$

$$e_i = Y_9 - \hat{Y}_9$$
$$= 97 - 96.5$$
$$= 0.5$$

$$\hat{Y}_9 = 31.5 + .546(119)$$

F-Test (Regression)

- A procedure to determine whether there is more variability explained by the regression or unexplained by the regression.
- Analysis of variance summary table

Total Deviation can be Partitioned into Two Parts

- Total deviation equals
- Deviation explained by the regression plus
- Deviation unexplained by the regression

"We are always acting on what has just finished happening. It happened at least 1/30th of a second ago.We think we're in the present, but we aren't. The present we know is only a movie of the past."

Tom Wolfe in

The Electric Kool-Aid Acid Test

Partitioning the Variance

$$\left(Y_i - \overline{Y}\right) = \left(\hat{Y}_i - \overline{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

Total deviation	=	Deviation explained by the regression	+	Deviation unexplained by the regression (Residual error)

$\overline{Y}$ = Mean of the total group

$\hat{Y}$ = Value predicted with regression equation

Y_i = Actual value

$$\sum \left(Y_i - \bar{Y}\right)^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 \quad + \sum \left(Y_i - \hat{Y}_i\right)^2$$

Total variation explained	=	Explained variation	+	Unexplained variation (residual)

Sum of Squares

$$SSt \;=\; SSr \;+\; SSe$$

Coefficient of Determination r^2

- The proportion of variance in Y that is explained by X (or vice versa)
- A measure obtained by squaring the correlation coefficient; that proportion of the total variance of a variable that is accounted for by knowing the value of another variable

Coefficient of Determination r^2

$$r^2 = \frac{SSr}{SSt} = 1 - \frac{SSe}{SSt}$$

Source of Variation

- Explained by Regression
- Degrees of Freedom
  - k-1 where k= number of estimated constants (variables)
- Sum of Squares
  - SSr
- Mean Squared
  - SSr/k-1

Source of Variation

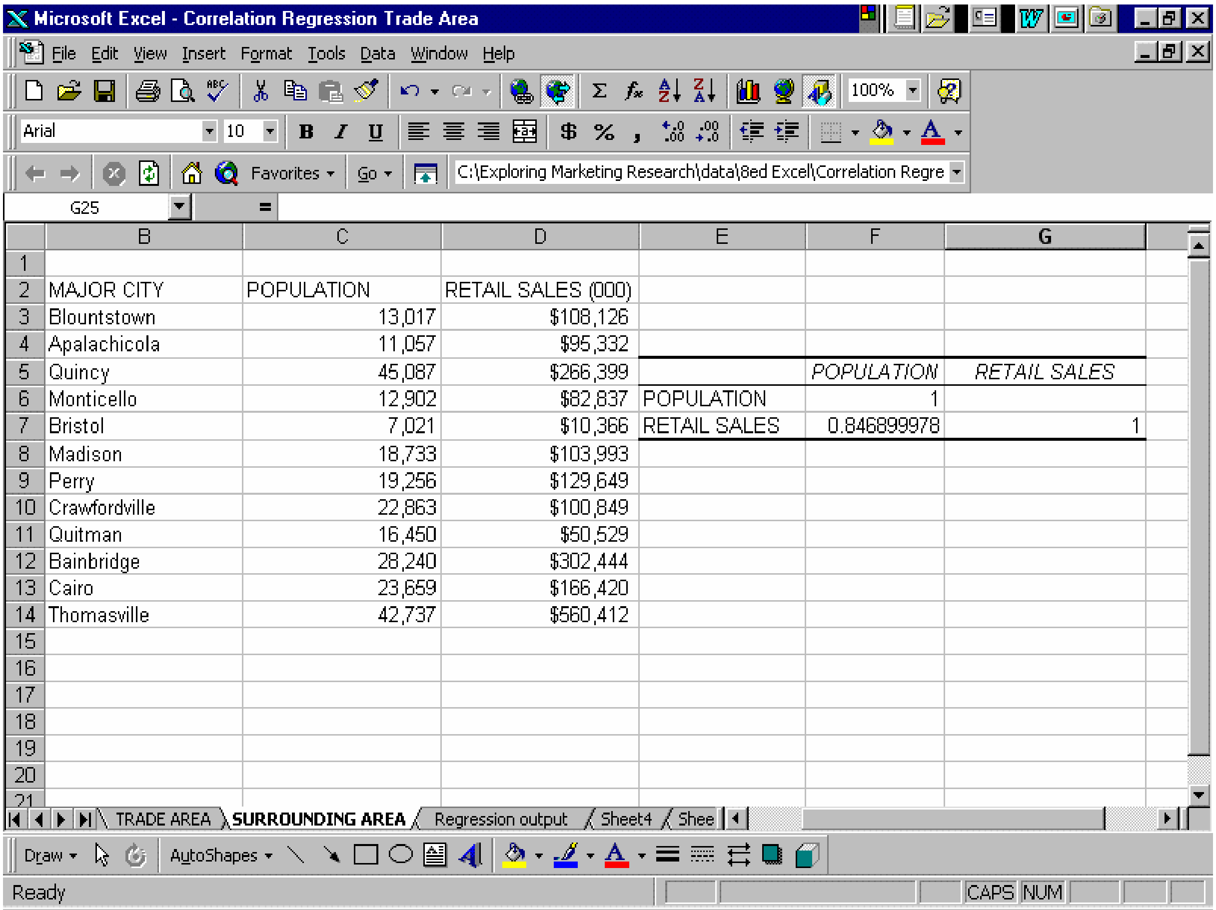
- Unexplained by Regression
- Degrees of Freedom
  - n-k where n=number of observations
- Sum of Squares
  - SSe
- Mean Squared
  - SSe/n-k

r^2 in the Example

$$r^2 = \frac{3{,}398.49}{3{,}882.4} = .875$$

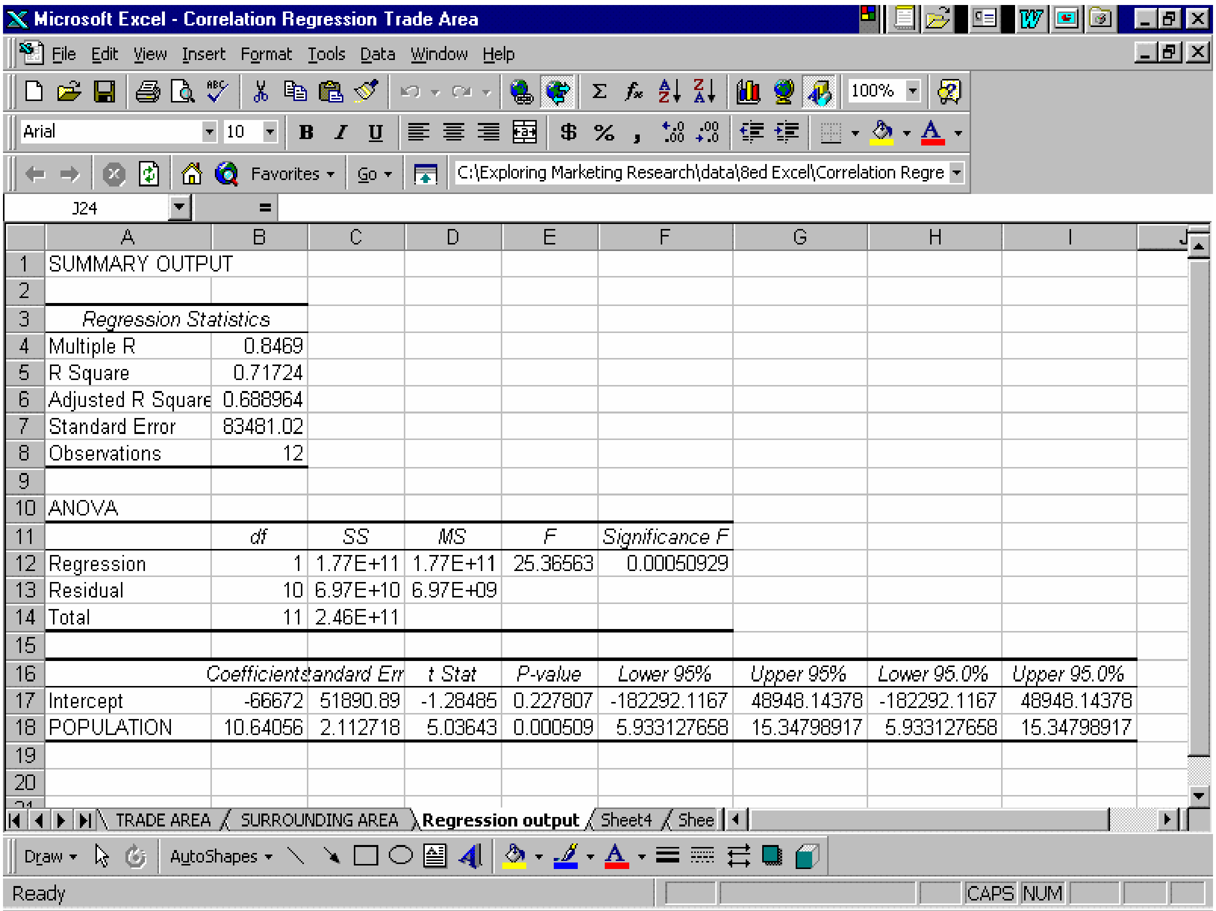
Multiple Regression

- Extension of Bivariate Regression
- Multidimensional when three or more variables are involved
- Simultaneously investigates the effect of two or more variables on a single dependent variable
- Discussed in Chapter 24



MAJOR CITY	POPULATION	RETAIL SALES (000)
Blountstown	13,017	$108,126
Apalachicola	11,057	$95,332
Quincy	45,087	$266,399
Monticello	12,902	$82,837
Bristol	7,021	$10,366
Madison	18,733	$103,993
Perry	19,256	$129,649
Crawfordville	22,863	$100,849
Quitman	16,450	$50,529
Bainbridge	28,240	$302,444
Cairo	23,659	$166,420
Thomasville	42,737	$560,412

	POPULATION	*RETAIL SALES*
POPULATION	1	
RETAIL SALES	0.846899978	1



SUMMARY OUTPUT

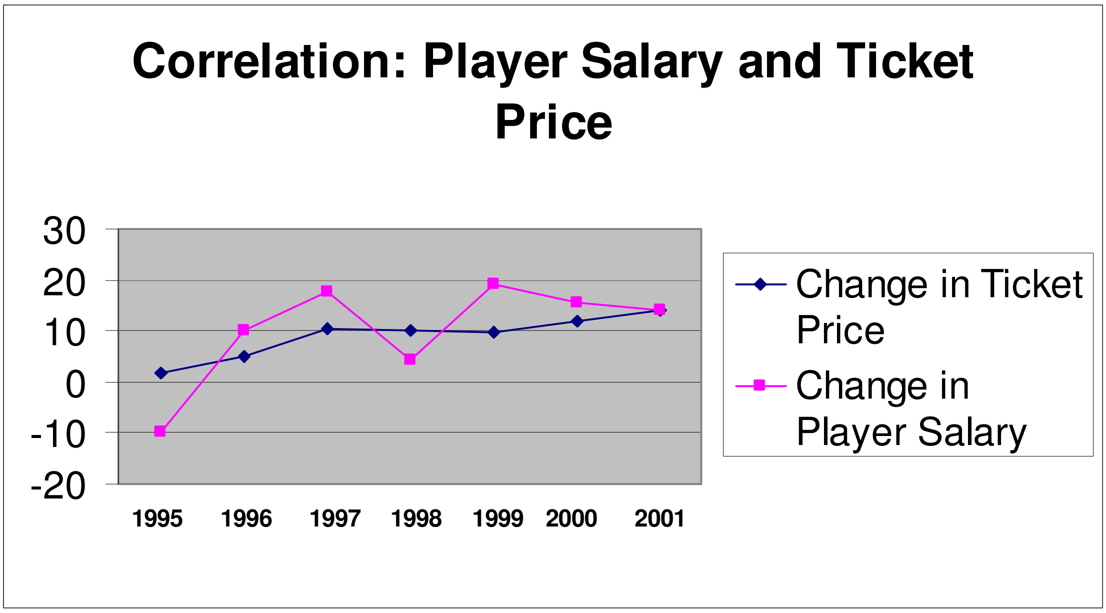
Regression Statistics	
Multiple R	0.8469
R Square	0.71724
Adjusted R Square	0.688964
Standard Error	83481.02
Observations	12

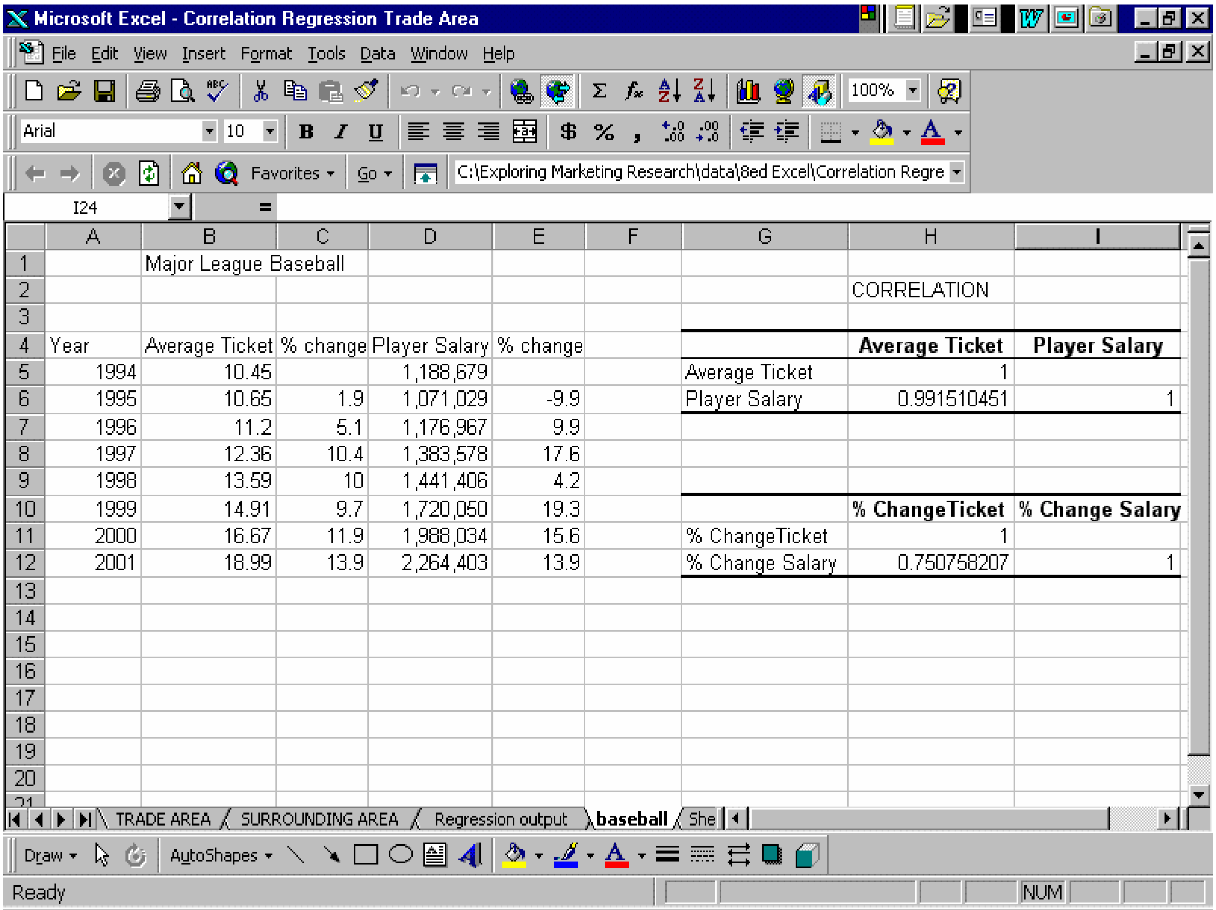
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.77E+11	1.77E+11	25.36563	0.00050929
Residual	10	6.97E+10	6.97E+09		
Total	11	2.46E+11			

	Coefficientstandard Err		*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-66672	51890.89	-1.28485	0.227807	-182292.1167	48948.14378	-182292.1167	48948.14378
POPULATION	10.64056	2.112718	5.03643	0.000509	5.933127658	15.34798917	5.933127658	15.34798917

Correlation Coefficient, r = .75





Major League Baseball

Year	Average Ticket	% change	Player Salary	% change
1994	10.45		1,188,679	
1995	10.65	1.9	1,071,029	-9.9
1996	11.2	5.1	1,176,967	9.9
1997	12.36	10.4	1,383,578	17.6
1998	13.59	10	1,441,406	4.2
1999	14.91	9.7	1,720,050	19.3
2000	16.67	11.9	1,988,034	15.6
2001	18.99	13.9	2,264,403	13.9

CORRELATION

	Average Ticket	Player Salary
Average Ticket	1	
Player Salary	0.991510451	1

	% ChangeTicket	% Change Salary
% ChangeTicket	1	
% Change Salary	0.750758207	1