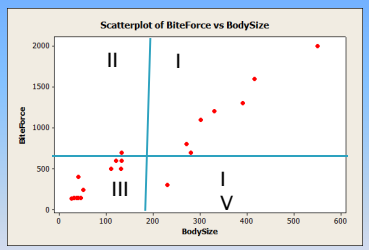
CORRELATION AND SIMPLE LINEAR REGRESSION

After studying this lecture, you should be able to:

- discuss the concept of relationship between a predictor and a dependent variable
- discuss measures of strength of linear relationships
- calculate covariance and correlation between two variables
- fit a straight line to (x,y) data
- determine if fit is reasonable
- test the significance of fitted line
- predict dependent variable for a given value of predictor

The New York Times, March 15, 2012:
"Throughout the ages, the basic crocodilian plan for dispatching prey has been simple but effective: chomp down hard and hang on tight". Scientists for a long time assumed that different snouts and teeth were related to bite strength. Greg Ericson and his team at Florida State University developed a device for measuring bite strength and measured the bite forces of 23 living crocodilians.

BodySize	BiteForce
25	140
30	145
35	142
40	148
45	144
50	240
40	400
110	500
120	600
130	500
132	600
132	700
230	300
270	800
280	700
300	1100
330	1200
390	1300
415	1600
550	2000



Look at signs of

$$P_i = (x_i - \bar{x})(y_i - \bar{y})$$

in the 4 quadrants I, II, III, IV (in the above graph):

$P_i > 0$ in Quadrants I and III
< 0 in Quadrants II and IV

This suggests using COVARIANCE as a measure of dependence.

$$\text{Covariance(X,Y)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n}$$

COVARIANCE large and >0 : strong positive relationship

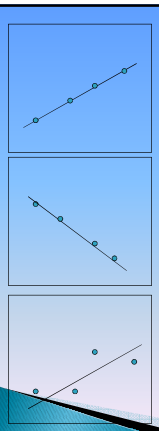
COVARIANCE large and <0 : strong negative relationship

Hope is that COVARIANCE ≈ 0 : no relationship. Unfortunately this is not true.

Also, covariance is in the units of X x Y. In the above example, unit of covariance is lbs x lbs. It is not easy to decide if a covariance is small or large. We therefore need a measure of dependence that has no units.

$$\text{Correlation(X,Y)} = r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_i (x_i - \bar{x})^2][\sum_i (y_i - \bar{y})^2]}}$$

$$= \frac{\text{cov}(x, y)}{sd(x)sd(y)}$$



r = +1

r = -1

$r^2 = 1$

$-1 \le r \le +1$

LINEAR REGRESSION MODEL

Straight Line Model $y_i = a + bx_i + e_i$

Assume errors e_i are independent and $N(0, \sigma^2)$ random variables.

THE LEAST SQUARES (LS) ESTIMATES of a and b are
obtained by minimizing

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - [a + bx_i])^2$$

The **LS Estimates of a and b** are :

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

We will illustrate fitting of a straight line for data of Example 1; all calculations are done in excel here.

	BodySize	BiteForce	X-XBAR	Y-YBAR	(X-XBAR)(Y-YBAR)	(X-XBAR)^2	(Y-YBAR)^2
	25	140	-157.7	-522.95	82469.215	24869.29	273476.703
	30	145	-152.7	-517.95	79090.965	23317.29	268272.203
	35	142	-147.7	-520.95	76944.315	21815.29	271388.903
	40	148	-142.7	-514.95	73483.365	20363.29	265173.503
	45	144	-137.7	-518.95	71459.415	18961.29	269309.103
	50	240	-132.7	-422.95	56125.465	17609.29	178886.703
	40	400	-142.7	-262.95	37522.965	20363.29	69142.7025
	110	500	-72.7	-162.95	11846.465	5285.29	26552.7025
	120	600	-62.7	-62.95	3946.965	3931.29	3962.7025
	130	500	-52.7	-162.95	8587.465	2777.29	26552.7025
	132	600	-50.7	-62.95	3191.565	2570.49	3962.7025
	132	700	-50.7	37.05	-1878.435	2570.49	1372.7025
	230	300	47.3	-362.95	-17167.535	2237.29	131732.703
	270	800	87.3	137.05	11964.465	7621.29	18782.7025
	280	700	97.3	37.05	3604.965	9467.29	1372.7025
	300	1100	117.3	437.05	51265.965	13759.29	191012.703
	330	1200	147.3	537.05	79107.465	21697.29	288422.703
	390	1300	207.3	637.05	132060.47	42973.29	405832.703
	415	1600	232.3	937.05	217676.72	53963.29	878062.703
	550	2000	367.3	1337.05	491098.47	134909.29	1787702.7
MEAN	182.7	662.95			1472400.7	451062.2	5360974.95

Example 1
(continued)

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{1472400.7}{451062.2} = 3.26$$

$$a = \bar{y} - b\bar{x} = 662.95 - 3.26 \times 182.7 = 66.56$$

Example 1 (continued)

One measure of the strength of the linear relationship is the CORRELATION COEFFICIENT

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\{\sum_{i=1}^{n}(x_i-\bar{x})^2\}\{\sum_{i=1}^{n}(y_i-\bar{y}^2\}}} = \frac{\text{Covariance}(x,y)}{sd(x).sd(y)}$$

$$= \frac{1472400.7}{1555034.776} = 0.9469$$

$$r^2 = 0.9469^2 = 0.897$$

In linear regression analysis,

R^2 = coefficient of determination, or % of variation in Y
explained by the fitted linear model (typically
expressed as a percentage)

For the data of Example 1 (BiteForce vs. BodySize)

R^2 = 89.7%

which implies that the fitted model provides a reasonable fit to that data.

PROPERTIES OF THE LEAST SQUARES ESTIMATES OF INTERCEPT AND SLOPE

$$\hat{a} \sim N(a, \sigma_{\hat{a}}^2)$$

$$\sigma_{\hat{a}}^2 = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\hat{b} \sim N(b, \sigma_{\hat{b}}^2)$$

$$\sigma_{\hat{b}}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

NOTE:
1) It follows from above that we can get confidence intervals
(or perform hypothesis tests) for the parameters a and b using
the t-tables with df = n-2, and

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

where

$$SSE = \sum_{i=1}^{n}[y_i - (\hat{a} + \hat{b}x_i)]^2$$

1) To test if X and Y are independent, we test
H_0: b = 0 vs. H_1: b≠0
by using the t-test.

2) In ANOVA approach to linear regression:
Total Sum of Squares =

$$TSS = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + b\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

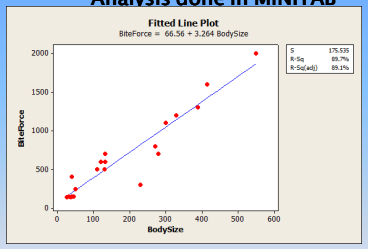
TSS = SS(Error) + SS(Regression)

df(Total) = n-1
df(Error) = n-2
df(Regression) = 1

100 r^2 = Total % of variability in data explained by the linear r

r^2 = 1 - SSE/TSS (SSE=0 then r^2 = 1)

Example 1: contd. (BiteForce as a function of BodySize)
Analysis done in MINITAB



ANOVA in Regression Analysis:

It can be shown that

$$\text{Total Sum of Squares (TSS)} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

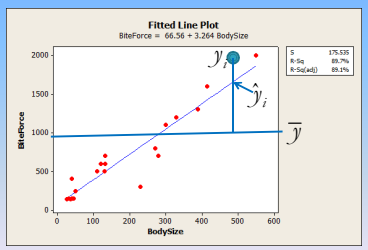
$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$= \text{SSE} + \text{SS(Regression)}$$

where $\hat{y}_i$ = fitted y-value for i-th data point $= a + bx_i$

$$y_i - \bar{y} = (y_i - \hat{y}) + (\hat{y}_i - \bar{y})$$

error regression



Example 1 : Analysis done in MINITAB

Regression Analysis: BiteForce versus BodySize

The regression equation is
BiteForce = 66.6 + 3.26 BodySize

Predictor	Coef	SE Coef	T	P
Constant	66.56	61.81	1.08	0.296
BodySize	3.2643	0.2614	12.49	0.000

S = 175.535 R-Sq = 89.7% R-Sq(adj) = 89.1%

Analysis of Variance

Source	DF	SS	MS	F	P*
Regression	1	4806352	4806352	155.99	0.000
Residual Error	18	554623	30812		
Total	19	5360975			

*Overall F-test is for H_0 : Slope (b) = 0, vs. H_1 : Slope (b) ≠ 0

P = 0.000 < .05, regression line is significant.

LS ESTIMATES USING MATRIX CALCULATIONS (Example 1)

$$y_1 = a + bx_1$$
$$y_2 = a + bx_2$$
.....
$$y_n = a + bx_n$$

In matrix notation:

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = a\begin{bmatrix} 1 \\ 1 \\ \\ 1 \end{bmatrix} + b\begin{bmatrix} x_1 \\ x_2 \\ \\ x_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix} = X\beta$$

Vector of DV values

the Design Matrix X

Vector of unknown regression parameters

$$y_1 = a + bx_1$$
$$y_2 = a + bx_2$$
.....
$$y_n = a + bx_n$$

In matrix notation:

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = a\begin{bmatrix} 1 \\ 1 \\ \\ 1 \end{bmatrix} + b\begin{bmatrix} x_1 \\ x_2 \\ \\ x_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix} = X\beta$$

$$e_1 = y_1 - (a + bx_1)$$
$$e_2 = y_2 - (a + bx_2)$$
.....
$$e_n = y_n - (a + bx_n)$$

$$\text{Error Sum of Squares} = \sum_{i=1}^{n} e_i^2 = \begin{bmatrix} e_1 & e_2 & \dots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix}$$

$$\sum_{i=1}^{n} e_i^2 = \underline{e}^T\underline{e} = (\underline{y} - X\beta)^T(\underline{y} - X\beta) =$$

$$\underline{y}^T\underline{y} - \underline{y}^T X\beta - (X\beta)^T\underline{y} + \beta^T X^T X\beta = \underline{y}^T\underline{y} - 2\beta^T X^T\underline{y} + \beta^T X^T X\beta$$

The value of $\underline{\beta}$ minimizing the Error Sum of Squares is

$\hat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{y}$ (referred to as OLS estimates)

with error variance

$\text{Var}(\hat{\beta}_i) = \hat{\sigma}^2 \times C_{ii}$

where

$\hat{\sigma}^2$ = estimated error variance = $\text{SSE}/(n-(p+1))$

p = number of predictors = 1 in simple linear regression

C_{ii} = i-th diagonal of $(X^T X)^{-1}$

(Example 1)

$$\underset{y}{\begin{bmatrix}140\\145\\...\\2000\end{bmatrix}} = \underset{\text{Intercept}}{\begin{bmatrix}1\\1\\...\\1\end{bmatrix}} + \underset{x}{\begin{bmatrix}25\\30\\...\\550\end{bmatrix}} \begin{bmatrix}a\\b\end{bmatrix}$$

$$X^T X = \begin{bmatrix}20 & 3654\\3654 & 1118648\end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix}0.124002 & -0.0004050\\-0.000405 & 0.0000022\end{bmatrix}$$

$$\hat{\underline{\beta}} = \begin{bmatrix}\hat{a}\\\hat{b}\end{bmatrix} = (X^T X)^{-1}(X^T \underline{y}) = \begin{bmatrix}66.5631\\3.2643\end{bmatrix}$$

(Example 1)

$\hat{\sigma}^2$ = estimated error variance = $\text{SSE}/(n-(p+1))$

$= 554622.7/18 = 30812.37$

$$(X^T X)^{-1} = \begin{bmatrix}0.124002 & -0.0004050\\-0.000405 & 0.0000022\end{bmatrix}$$

$\text{Var}(\hat{\beta}_0) = \text{Var}(\hat{a}) = \hat{\sigma}^2 \times C_{11} = 554622.7 \times 0.124002 = 3820.780823$

$\text{Var}(\hat{\beta}_1) = \text{Var}(\hat{b}) = \hat{\sigma}^2 \times C_{22} = 554622.7 \times 0.0000022 = 0.068310685$

$sd(\hat{a}) = \sqrt{3820.780823} = 61.81$

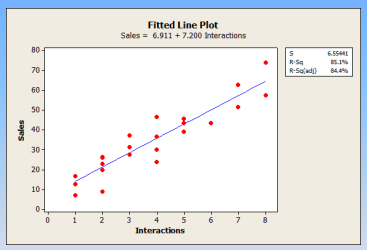
$sd(\hat{b}) = \sqrt{0.068310685} = 0.26$

(Note that these values are same as on Slide 21)

Example 2: The table contains sales data from a sample of 25 repeat customers; interactions include exposure to cross-selling opportunities and web ads. Are sales correlated to interactions, and, if so, how strong is the relationship?

Customer	Interactions	Sales	Customer	Interactions	Sales
1	1	6.9	14	3	27.5
2	3	37.3	15	3	37.3
3	2	9	16	5	39
4	5	43.4	17	6	43.5
5	2	26	18	7	51.6
6	2	22.9	19	5	45.6
7	1	12.6	20	3	31.3
8	7	62.8	21	2	19.8
9	4	23.7	22	4	30.1
10	4	36.7	23	5	43.4
11	2	26.2	24	8	73.9
12	8	57.5	25	4	46.6
13	1	16.6			

Example 2:



Example 2:

Regression Analysis: Sales versus Interactions

The regression equation is

Sales = 6.911 + 7.200 Interactions

S = 6.55441 R-Sq = 85.1% R-Sq(adj) = 84.4%

Analysis of Variance

Source	DF	SS	MS	F	P*
Regression	1	5632.42	5632.42		
Total	24	6620.50			

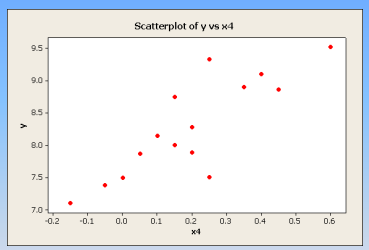
*Overall F-test is for H_0 : Slope (b) = 0, vs. H_1 : Slope (b) ≠ 0

P = 0.000 < .05, regression line is significant.

Example 3:
y = demand,
x_1 = the price (in $) of detergent FRESH offered by Enterprise Industries, x_2 = the average price (in $) of similar detergents in the market
Find a regression model for y as a function of x_1, x_2, and $x_2 - x_1$
The data set.

x1	x2	y
3.85	3.8	7.38
3.75	4	7.51
3.7	4.3	9.52
3.7	3.7	7.5
3.6	3.85	9.33
3.6	3.8	8.28
3.6	3.75	8.75
3.8	3.85	7.87
3.8	3.65	7.1
3.85	4	8
3.9	4.1	7.89
3.9	4	8.15
3.7	4.1	9.1
3.75	4.2	8.86
3.75	4.1	8.9

Scatterplot for data of Example 3



For the example on slide 1,
covariance = 0.1264857and correlation = 0.839

Regression Analysis: y versus diff

The regression equation is
y = 7.645 + 3.206 diff

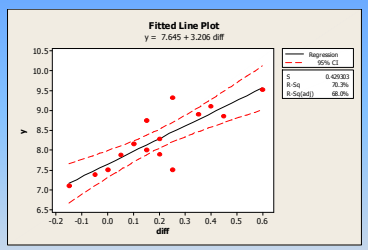
S = 0.429303 R-Sq = 70.3% R-Sq(adj) = 68.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	5.67725	5.67725	30.80	0.000
Error	13	2.39591	0.18430		
Total	14	8.07316			

Predictor	Coef	SE Coef	T	P
Constant	7.6455	0.1587	48.17	0.000
diff	3.2060	0.5776	5.55	0.000

For the data of Example 3, the fitted line is y = 7.646 + 3.206 diff



r = 0.839,
r^2 = 0.704 = percentage of variability in y explained by the line