

Scatterplot

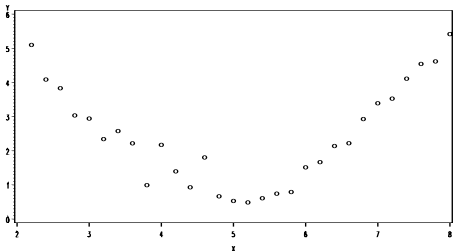
- Allows visual assessment of linear relationship
- Predictor variable (X) plotted on x-axis
- Response variable (Y) plotted on y-axis
- Regression gives “best” line through (X, Y) pairs

$$Y = \beta_0 + \beta_1 X$$

- Correlation describes “tightness” of linearity

$$-1 \leq r \leq 1$$

Example: Computing r not appropriate



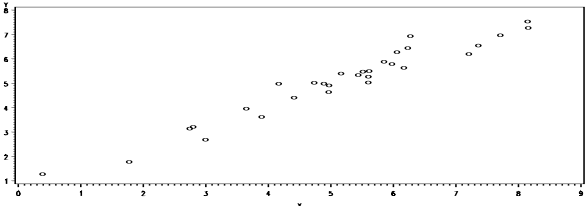
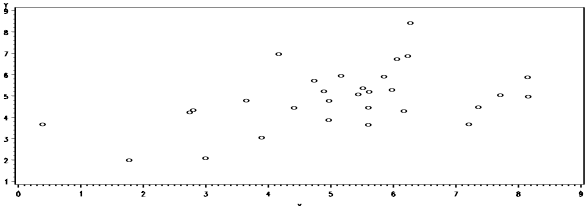
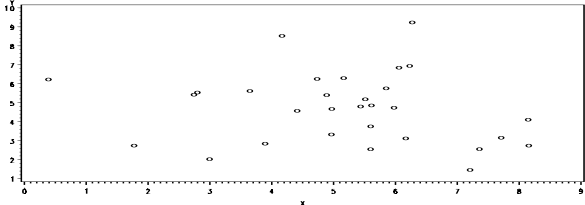
The Correlation Coefficient

in

Simple Linear Regression

Applied Regression and Other Multivariable Methods
Sections 6-1 - 6-7

Examples: $r = -0.18$, $r = 0.45$, and $r = 0.97$



The Correlation Coefficient

- Correlation coefficient is a dimensionless statistic

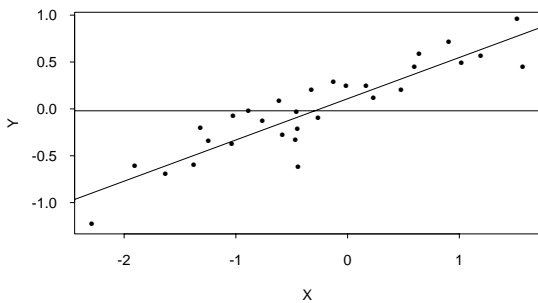
$$r = \frac{SSXY}{\sqrt{SSX * SSY}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{S_Y}{S_X} \beta_1$$

- Sign of $SSXY$ indicates direction of trend
 - $(X - \bar{X})(Y - \bar{Y})$ positive when
 - $X > \bar{X}$ and $Y > \bar{Y}$
 - $X < \bar{X}$ and $Y < \bar{Y}$
 - $(X - \bar{X})(Y - \bar{Y})$ negative when
 - $X > \bar{X}$ and $Y < \bar{Y}$
 - $X < \bar{X}$ and $Y > \bar{Y}$

- r has same sign as β_1
- Symmetry - can interchange X and Y , same r

The Correlation Coefficient

- Describes how close the data cluster about the line
- Provides an **index** of association and direction
- **Association**
 - Lack of statistical independence between X and Y
 - X provides some assistance with predicting Y
- If not using X predict Y with \bar{Y}
- If using X , predict Y with $\hat{\beta}_0 + \hat{\beta}_1 X$



6-4

Coefficient of Determination

- r^2 known as coefficient of determination. % of total variation in Y explained by regression

$$r^2 = 1 - \frac{SSE}{SSY}$$

where $SSY = \sum (Y - \bar{Y})^2$ and $SSE = \sum (Y - \hat{Y})^2$

- If perfect linear association $\rightarrow SSE = 0$
 $r = \pm 1$ and $r^2 = 100\%$
- If no linear association $\rightarrow SSE = SSY$
 $r = 0$ and $r^2 = 0\%$
- Can use r^2 to approximate reduction in pred std err

$$\frac{S_{Y|X}}{S_Y} \approx \sqrt{1 - r^2}$$

6-5

Hypothesis Test for $H_0 : \rho = 0$

- Based on bivariate normal population model
- r is sample estimate of pop parameter ρ

$$\rho = \beta_1 \frac{\sigma_Y}{\sigma_X} \rightarrow b_1 \sqrt{\frac{SSY}{SSX}} = r$$

- Use t-test to see if population linear association
- Identical to test $H_0 : \beta_1 = 0$
- The test statistic

$$T = r \sqrt{\frac{n-2}{1-r^2}}$$

under H_0 is t distributed with $n - 2$ df

6-6

General Inference on ρ

- $H_0 : \rho = \rho_0$ no longer equivalent to test of slope
 - Must construct sampling dist of r when $\rho = \rho_0$
 - This sampling dist only symmetric when $\rho = 0$
 - Will **transform** statistic to make approx normal
- $$f(r) = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \sim N \left(\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right), 1/(n-3) \right)$$
- Compute P-value through standardization and Table A-1
 - Can also construct approximate CI for ρ
 - Construct CI for $f(\rho)$, then convert back using A-5

$$L_Z = f(r) - z_{\alpha/2} \frac{1}{\sqrt{n-3}}$$

$$L_U = f(r) + z_{\alpha/2} \frac{1}{\sqrt{n-3}}$$

6-7

Example

Consider the blood toluene vs parts per million toluene data of Problem 5-15 (page 80-82). This involves $n = 60$ rats and resulted in an $r^2 = .9497$ which means $r = .9745$.

Hypothesis Test: Suppose the experimenter was interested in testing $H_0 : \rho = .95$ vs $H_A : \rho > .95$. Based on the sample

$$\begin{aligned} f(r) &= .5 \ln(1.9745/.0255) = 2.1752 \\ f(\rho_0) &= .5 \ln(1.95/.05) = 1.8318 \\ Z &= (2.1752 - 1.8318)/(1/\sqrt{60-3}) = 2.59 \end{aligned}$$

From Table A-1, the area to the **right** of 2.59 is 0.0048. For any $\alpha > .0048$, the scientist would reject H_0 in favor of the alternative.

Confidence Interval: Suppose the experimenter was interested in constructing a 99% CI for ρ . From Table A-1, $z_{.005} = 2.576$. Thus,

$$\begin{aligned} L_Z &= 2.1752 - 2.576(1/\sqrt{60-3}) = 1.8340 \\ U_Z &= 2.1752 + 2.576(1/\sqrt{60-3}) = 2.5164 \end{aligned}$$

From Table A-5, these endpoints convert to a confidence interval of (0.950, .987). If the formulas at the bottom of pg 98 are used, the interval is (.9502, .9870).

6-8

Testing the Equality of Two Correlations

- Comparing two indep populations $H_0 : \rho_1 = \rho_2$
 - Perform approximate test using normal transformation

$$\begin{aligned} Z_1 &= .5 \ln \left(\frac{1+r_1}{1-r_1} \right) & Z_2 &= .5 \ln \left(\frac{1+r_2}{1-r_2} \right) \\ Z &= \frac{Z_1 - Z_2}{\sqrt{1/(n_1-3) + 1/(n_2-3)}} \end{aligned}$$

- Comparing two variables within single population
 - Compare correlations of Y vs X_1 and Y vs X_2
 - Let r_{12} and r_{13} be the sample correlations
 - Let r_{23} be the sample correlation of X_1 vs X_2

$$Z = \frac{(r_{12} - r_{13})\sqrt{n}}{\sqrt{(1-r_{12}^2)^2 + (1-r_{13}^2)^2 - 2r_{23}^2 - (2r_{23} - r_{12}r_{13})(1-r_{12}^2 - r_{13}^2 - r_{23}^2)}}$$

6-9

SAS Procedures

- Most point estimates available in regression output
- Use corr option to print correlations
- Can also use Proc Corr procedure
 - Gives P-value for $H_0 : \rho = 0$
- Program example6-1.sas performs Chpt 6 calcs

```
data table5;
  infile 'I:\.www\datasets502\example5-1.dat';
  input indiv sbp age;

proc reg corr; model sbp=age;

proc corr; var sbp age;

data new ;
  input r r0 n alpha;
  fr=.5*log((1+r)/(1-r)); fr0=.5*log((1+r0)/(1-r0));
  z_alpha = probit(1-alpha/2); Z = (fr - fr0)*sqrt(n-3);
  L_Z = fr - z_alpha*sqrt(1/(n-3)); U_Z = fr + z_alpha*sqrt(1/(n-3));
  L_p = (exp(2*L_Z)-1)/(exp(2*L_Z)+1);
  U_p = (exp(2*U_Z)-1)/(exp(2*U_Z)+1);
  pvalue = 2*probnorm(-abs(Z));
  cards;
  .65757 .85 30 .05
;

proc print;
  var alpha pvalue L_p U_p;
run;
```

6-10

The REG Procedure

Variable	age	sbp
age	1.0000	0.6576
sbp	0.6576	1.0000

Dependent Variable: sbp

Source	DF	Sum of Squares		Mean Square	F Value	Pr > F
		Squares	Square			
Model	1	6394.02269	6394.02269	21.33	<.0001	
Error	28	8393.44398	299.76586			
Corrected Total	29	14787				

Root MSE	17.31375	R-Square	0.4324
Dependent Mean	142.53333	Adj R-Sq	0.4121
Coeff Var	12.14716		

Pearson Correlation Coefficients, N = 30

Prob > |r| under H0: Rho=0

	sbp	age
sbp	1.00000	0.65757 <.0001
age	0.65757	1.00000 <.0001

Obs	alpha	pvalue	L_p	U_p
1	0.05	0.015103	0.38960	0.82289

6-11