DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

CORRELATION AND REGRESSION

I. AGENDA:
- A. Summary of hypothesis test
- B. Confidence intervals
- C. Correlation
- D. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition, Chapter 9 pages 318 to 323..

II. INFERENCE FOR REGRESSION COEFFICIENTS:
- A. The t-test for an individual regression coefficient
  1. Sample size = N
  2. Hypotheses:
     - i. Null: $H_0 \quad \beta_1 = 0$
     - ii. Alternative: $\beta_1 \neq 0$ (two-tailed)
     - iii. Or $\beta_i > 0$ or $\beta_i < 0$ (one-tailed)
  3. Expected value of coefficient estimators: $\boldsymbol{E(\hat{\beta}_1) = \beta_1}$
  4. The statistic $\boldsymbol{t_{obs}} = \dfrac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$ has a t distribution with N - 2 degrees of freedom.
  5. The standard error or standard deviation of the regression coefficient is given by

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}_{Y|X}}{\sum (X_i - \bar{X})^2}$$

  6. And $\hat{\sigma}_{Y|X}$ is $\boldsymbol{S^2 / (N - 2)}$.
- B. Example:
  1. Consider these data pertaining to cigarette consumption and four types of illness.
     - i. We will only consider two, bladder cancer and leukemia
  2. The data pertain to 44 states.

State	Cigs	Bladder	Lung	Kidney	Leukemia
AL	18.20	2.90	17.05	1.59	6.15
AZ	25.82	3.52	19.80	2.75	6.61
AR	18.24	2.99	15.98	2.02	6.94
CA	28.60	4.46	22.07	2.66	7.06
CT	31.10	5.11	22.83	3.35	7.20
DE	33.60	4.78	24.55	3.36	6.45
DC	40.46	5.60	27.27	3.13	7.08
FL	28.27	4.46	23.57	2.41	6.07
ID	20.10	3.08	13.58	2.46	6.62
IL	27.91	4.75	22.80	2.95	7.27
IN	26.18	4.09	20.30	2.81	7.00
IO	22.12	4.23	16.59	2.90	7.69
KS	21.84	2.91	16.84	2.88	7.42
KY	23.44	2.86	17.71	2.13	6.41
LA	21.58	4.65	25.45	2.30	6.71
ME	28.92	4.79	20.94	3.22	6.24
MD	25.91	5.21	26.48	2.85	6.81
MA	26.92	4.69	22.04	3.03	6.89
MI	24.96	5.27	22.72	2.97	6.91
MN	22.06	3.72	14.20	3.54	8.28
MS	16.08	3.06	15.60	1.77	6.08
MO	27.56	4.04	20.98	2.55	6.82
MT	23.75	3.95	19.50	3.43	6.90
NB	23.32	3.72	16.70	2.92	7.80
NE	42.40	6.54	23.03	2.85	6.67
NJ	28.64	5.98	25.95	3.12	7.12
NM	21.16	2.90	14.59	2.52	5.95
NY	29.14	5.30	25.02	3.10	7.23
ND	19.96	2.89	12.12	3.62	6.99
OH	26.38	4.47	21.89	2.95	7.38
OK	23.44	2.93	19.45	2.45	7.46
PE	23.78	4.89	12.11	2.75	6.83
RI	29.18	4.99	23.68	2.84	6.35
SC	18.06	3.25	17.45	2.05	5.82
SD	20.94	3.64	14.11	3.11	8.15
TE	20.08	2.94	17.60	2.18	6.59
TX	22.57	3.21	20.74	2.69	7.02
UT	14.00	3.31	12.01	2.20	6.71
VT	25.89	4.63	21.22	3.17	6.56
WA	21.17	4.04	20.34	2.78	7.48
WI	21.25	5.14	20.55	2.34	6.73
WV	22.86	4.78	15.53	3.28	7.38
WY	28.04	3.20	15.92	2.66	5.78
AK	30.34	3.46	25.88	4.32	4.90

3. Let's test the hypothesis that bladder cancer is not linearly related to cigarette consumption.
   i. That is, H_0: $\beta_{bladder,\ cig} = 0$
   ii. The alternative hypothesis is that $\beta_{bladder,\ cig} > 0$. That is, we assume either that there is no relationship or that if a relationship exists, it is

positive.

4. Here are the results from MINITAB

```
The regression equation is
Bladder = 1.09 + 0.122 Cigs

Predictor        Coef       StDev          T        P
Constant       1.0861      0.4844       2.24    0.030
Cigs          0.12182     0.01898       6.42    0.000

S = 0.6938      R-Sq = 49.5%     R-Sq(adj) = 48.3%
```

5. We see that the observed t = .12182/.01898 = 6.42.
6. The critical t for N - 2 = 42 degrees of freedom at the .001 level is about 2.576.
   i. Our observed value greatly exceeds this.
   ii. In fact, the "attained' probability of this t (or larger) under the null hypothesis is less than .000.
7. We would no doubt conclude that there is a statistically significant relationship between smoking and the incidence of bladder cancer.
   i. Whether or not this is a causal relationship cannot be determined from these data.

C. Now, is there a statistically significant relationship between smoking and leukemia?

1. Here are the results of the test of H_0 $\beta_{leukemia,\ cig} = 0$

```
The regression equation is
Leuk = 7.03 - 0.0078 Cigs

Predictor        Coef       StDev          T        P
Constant       7.0252      0.4498      15.62    0.000
Cigs         -0.00784     0.01763      -0.44    0.659

S = 0.6443      R-Sq = 0.5%      R-Sq(adj) = 0.0%
```

2. We see that the **absolute value** of the observed t (t_{obs} = -.44) is less than the critical t (one-tailed with 42 degrees of freedom) 1.282.
   i. The table shows the attained probability is .659.
3. Hence, we would accept the null hypothesis in this case.
4.

III. F TEST OF THE REGRESSION MODEL:

A. As noted last time (Class 10 on tape), we can test the regression model with an "F test."

B. Summary:
   1. We form two independent estimates of the error standard deviation, σ_ε
   2. One uses the "mean square for error"; that is, the error or residual sum of squares divided by N - 2 the degrees of freedom.
   3. The other estimator uses the "mean square for regression"; that is, the regression sum of squares divided by 1, the degree of freedom for regression.
   4. Under the null hypothesis these two are unbiased estimators of σ_ε and their ratio should be 1.0, except for sampling error.
   5. If, however, there is a linear relationship, then the expected value of the mean square for regression will be larger than the error standard deviation and the ratio will be greater than 1.

C. For this test, then, we use the F statistic, which has 1 and N - 2 degrees of freedom.
   1. For 1 and 44 - 2 = 42 degrees of freedom the critical F at the .05 level is about 4.08; at the .01 level it is about 7.31; and at the .001 level it is about 11.97.

D. Here are the observed results from the bladder cancer and cigarette consumption data.

Analysis of Variance for bladder cancer on cigarette consumption

Source	DF	SS	MS	F	P
Regression	1	19.821	19.821	41.18	0.000
Residual Error	42	20.215	0.481		
Total	43	40.036			

   1. The observed F is 19.821/.481 = 41.18.
   2. It exceeds all of the critical values and in fact, as MINITAB notes its probability under the null hypothesis is less than .000.

E. And here are the results for the leukemia data.

Source	DF	SS	MS	F	P
Regression	1	0.0821	0.0821	0.20	0.659
Residual Error	42	17.4349	0.4151		
Total	43	17.5171			

   1. As we would expect the observed F does not exceed any of the critical values and hence we accept the null hypothesis.

F. Note that the results of the two test procedures, t and F, are equivalent in the case

of simple regression.

IV. CONFIDENCE INTERVALS:

A. Repeated from Class 10 notes

B. Confidence intervals can be placed around parameter estimators in the usual way.

C. For (1 - α)% confidence intervals use a t with appropriate degrees of freedom:

$$\hat{\beta}_j \pm t_{(\alpha/2, n-2)}\ \hat{\sigma}_{\hat{\beta}_j}$$

1. The estimated β_j for j = 0, 1 is bracketed by an appropriate t times the estimated standard deviation (error) of the coefficient.
2. The $t_{(\alpha/2, N-2)}$ is chosen from the t distribution with N - 2 degrees of freedom and α/2 level of significance.

D. Examples

1. The estimated regression coefficient of bladder disease regressed on smoking is .12182 and its standard deviation is .01898.
   i. Suppose we want 99% percent confidence intervals. This suggests setting α to .01.
   ii. The appropriate (1 - .01)% t statistic with N - 2 = 42 degrees of freedom is about 2.326
   iii. Hence 99% intervals are

$$\mathbf{.1218 \pm (.01898)(2.326)}$$

   iv. The upper interval is thus .1218 + (.01898)(2.326) = .1218 + .0441 = .1659.
   v. And the lower interval is .1218 - (.01898)(2.326) = .1218 - .0441 = .0777.
   vi. Note that the intervals does not include 0. This result is equivalent to rejecting the null hypothesis at the .01 level.
   vii. Similarly, the 99.9% (α = .001) level intervals, .0729 - .1707, do not include 0, which agrees with our previous finding that the null hypothesis can be rejected at the .001 level.
2. 95 percent intervals for the estimated coefficient of leukemia on smoking, which is -.00784,

$$\mathbf{-.00784 \pm (.01763)(1.96)}$$

   i. The intervals thus go from -.0424 to .0267.
   ii. These intervals do contain 0, which fits with our previous decision

not to reject the null hypothesis at the .05 level.

V. CORRELATION AND THE CORRELATION COEFFICIENT:

A. The correlation coefficient provides another way to assess the fit of a model

B. Here are some plots of "ideal' relationships. That is, real data never follow these patterns exactly, but they sometimes approximate them.

1. Consider observations on two variables, Y (dependent) and X (independent).
2. They can be plotted on an X-Y coordinate system as we have seen.
3. Figure 1 shows and example of "positive linear correlation"

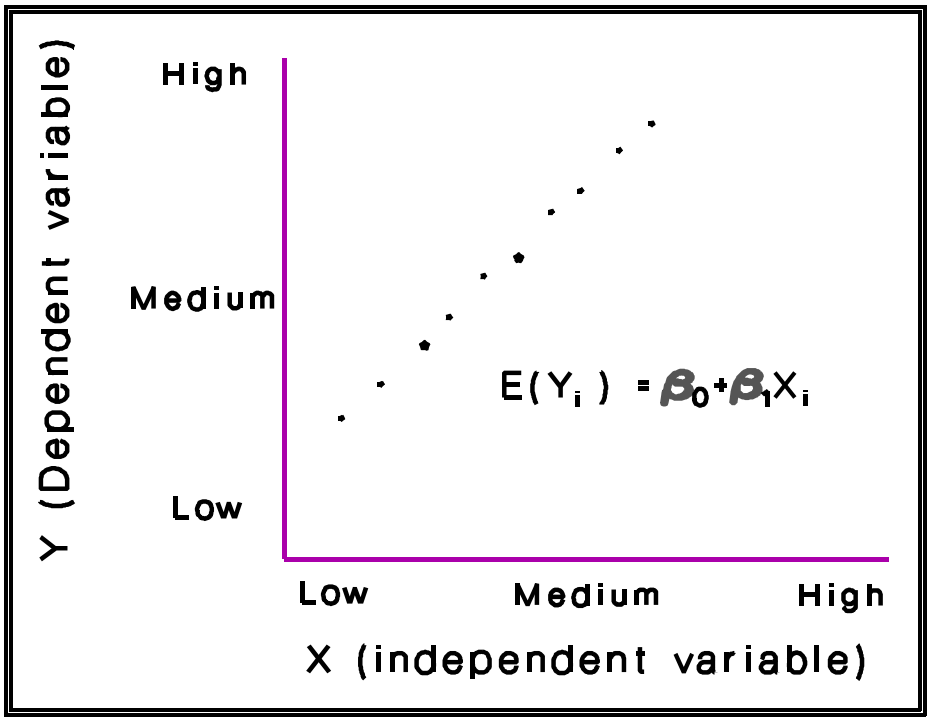


Figure 1: Example of Perfect Positive Correlation

i. Note that the data follow a pattern:

a) Large values of X are associated with large values of Y and small values of X are associated with small values of Y.

b) Moreover, all of the points lie on a straight line; the relationship is said to be **linear**.

C. The next figure illustrates perfect negative correlation

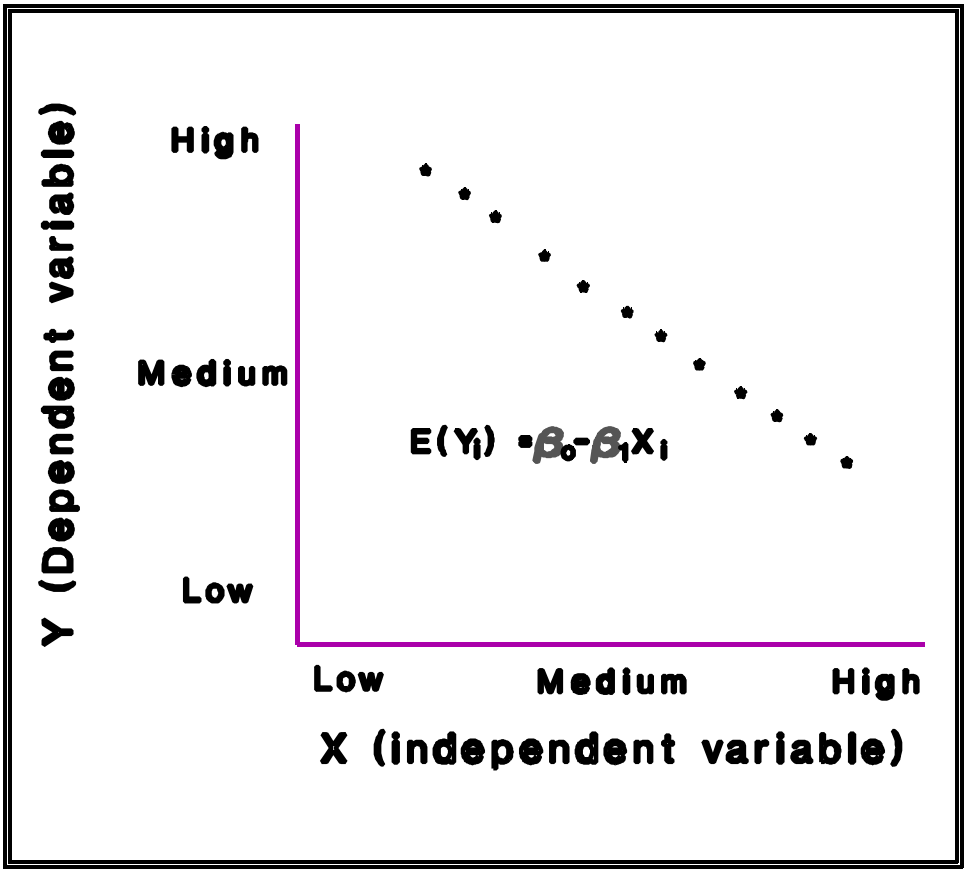


Figure 2: Example of Perfect Negative Correlation

1. The points lie on a straight line (the relationship is linear) but large values of X are associated with small values of Y and conversely small values of X are connected with large values of Y.

D. Now consider the figure on the next page (Figure 3). There the points follow the pattern of positive correlation (large values connected with large values, small values with small values), but the points do not lie on a straight line.

1. We call this sort of relationship positive but non-linear correlation.
2. Of course, a straight line will pass through or about through most of the points.

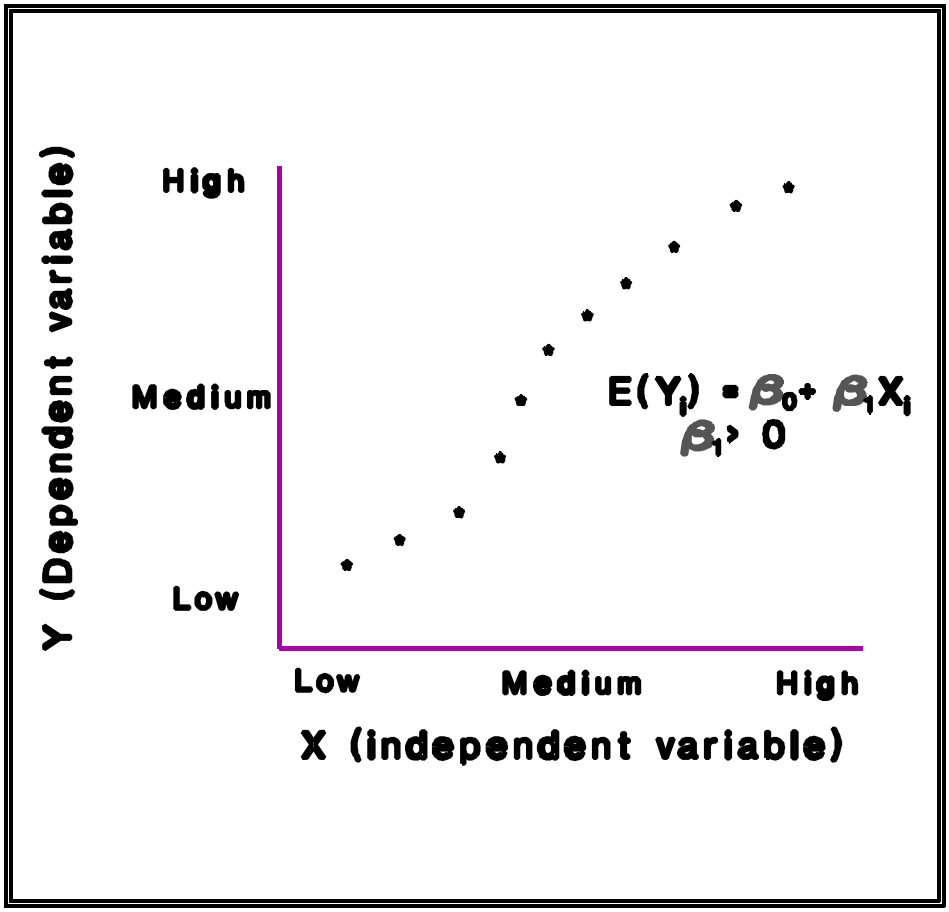


Figure 3: Nonlinear Positive Correlation

E. Correlation coefficient:

1. What is the best way to measure the "strength" and "nature" of the relationship between X and Y in these graphs? The best measure is perhaps β_1, the regression coefficient but the correlation coefficient is also used.
2. Next to R^2 the correlation coefficient is perhaps the most commonly used (and abused) statistic in applied data analysis. Since it appears so frequently and since its numerical value is easily misunderstood, we need to spend a moment or two looking at its properties.
   i. The correlation coefficient, , measures linear correlation between a pair of variables, X and Y.

VI. INTERPRETATION AND PROPERTIES:

A. The coefficient is a **bounded index**, which means that its numerical values lie between +1.0 and -1.0. In other words, it can never be larger than one in absolute value. More specifically,

1. r = 1.0 if X-Y are perfectly linearly positively correlated

2. r = -1.0 if X-Y are perfectly linearly negatively correlated
3. r = 0 if there is no linear correlation between X and Y.
4. Note, however, that r = 0 does not imply that two variables are not related or are statistically independent.

B. Symmetry: r has the same numerical value no matter whether X or Y is taken as dependent. This is not true of the regression coefficients, β_0 and β_1.

1. That is,

$$\beta_{YX} \neq \beta_{XY} \textbf{ (}\textit{in most cases}\textbf{)}$$

2. But,

$$_{YX} = \ _{XY} \textbf{ (}\textit{always}\textbf{)}$$

C. The correlation coefficient, r, is **scale-free**: that is, it is not measured in units of Y (or X). In this way it is unlike β_1, which is measured in values of Y. The measure, r, is for this reason called a standardized measure of association. Some people prefer r for this reason because they think that they can compare different variables and different populations. Later we see that this may or may not be the case.

D. Since r is the square root of R^2, it is common to square a sample or observed value of r to determine the amount of variation explained by X.

E. Further properties:

1. r will be large to the extent that (other things being equal):

   i. the variance of the errors, $\hat{}_i$, is small

   ii. the variance of X (σ_X^2) is large

   iii. the absolute value of β_1 is large

2. This last point follows from the fact that the correlation coefficient is related to β_1 as follows:

$$\beta_1 = (\sigma_Y/\sigma_X)$$

$$= \beta(\sigma_X/\sigma_Y)$$

***where*:**
σ_Y is the standard deviation of Y
σ_X is the standard deviation of X

3. These relationships hold for sample values.

4. Thus r is a function of 1) the strength of the relationship between X and Y as measured by β_1 **and** 2) the variation in X and the variation in Y.

F. The correlation coefficient by itself is not useful for these purposes:

1. Inferring causality, a point that applies to β_1 as well. See the attached diagram.
2. Judging the theoretical importance of variables. Suppose, to take a common example, that you have a dependent variable of importance, say the turnout rate. Suppose, in addition, that you have two "competing" explanations of voting as measured by two independent variables, social-economic status and level of group-party linkage. One cannot say simply that because the correlation between voting and SES is .8 whereas the correlation between turnout and party linkage is .5 that social and economic factors are more "important" explanations.
3. Comparing different populations. Suppose that you have two different populations, as for example, "underdeveloped" versus "developed" countries or black versus white. The correlation coefficient may not be a good means to compare the strength of the relationship between X and Y in the two sets of countries. Here is a pictorial explanation

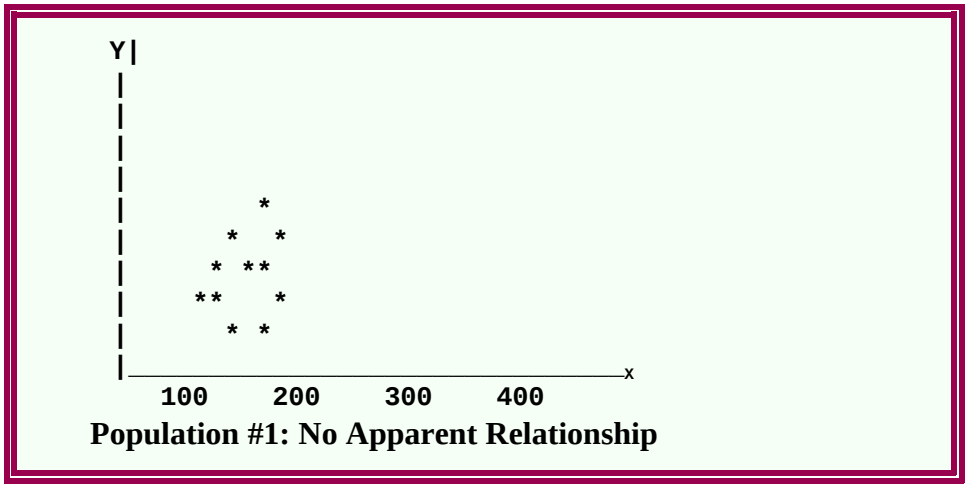
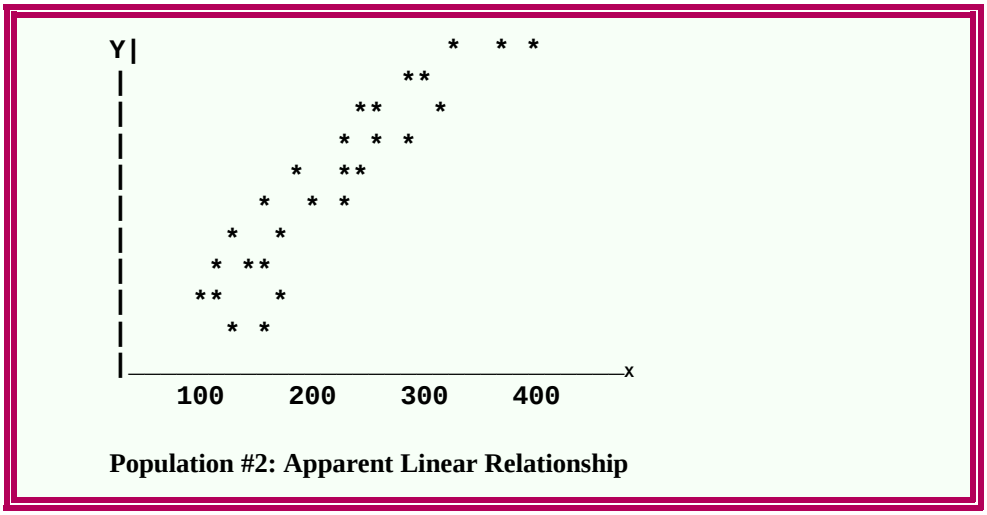


Figure 4: No Apparent Correlation

i. Note that X varies between roughly 100 and 200

4. Now consider the next figure



Population #2: Apparent Linear Relationship

Figure 5: Apparent Linear Correlation

i. Here (in this population or sample) X varies between 100 and 400.

5. These populations differ partly because there is more variation in X in the second than in the first. In a sense, X has more room to "operate" or work its effects on Y.
6. An r in the first panel (Figure 4) would be nearly 0; in the second (Figure 5) it would be close to 1.0.
7. The point is that when designing research try to maximize the variation of X. By the same token, when assessing someone else's work, ask how much the independent variable varies.
   i. Example: in studying Perot voting in the Midwest and western states, one encounters trouble using "Percent Black" as an explanatory variable because the percentages in most counties are quite low. Therefore, there is relatively little variation; X can't explain Y, because it doesn't vary enough.
8. In short, realize that r, the correlation coefficient, depends on the variation in X as well as on the strength of the relationship as measured by β_1. Thus, in general, try to measure β_1, the <u>unstandardized</u> regression parameter.

i.

G. Formal definition:

$$= \sqrt{\frac{cov(Y,X)}{var(X)var(Y)}}$$

1. The sample statistic is:

$$r = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\left[\sum_{i=1}^{N}(X_i - \bar{X})^2\right]\left[\sum_{i=1}^{N}(Y_i - \bar{Y})^2\right]}}$$

2. A computing formula is:

$$r = \frac{\sum_{1=1}^{N} X_i Y_i - \frac{(\sum_{i=1}^{N} X_i)(\sum_{i=1}^{N} Y_i)}{N}}{\sqrt{\left[\sum_{i=1}^{N} X_i^2 - \frac{(\sum_{i=1}^{N} X_i)^2}{N}\right]\left[\sum_{i=1}^{N} Y_i^2 - \frac{(\sum_{i=1}^{N} Y_i)^2}{N}\right]}}$$

3. Actually in the two variable case, the correlation coefficeint is the square root of R^2 with an appropriate sign.
   i. Thus, if you have a scatterplot and can determine the direction of the line and an R^2, you can get the corresponding value of r by taking the square root of R^2.

H. The computing formula presented above works well with a pocket calculator and a small amount of data.
   1. Use the accumulate keys to find the sums of Y and X and the crossproduct X times Y.
      i. Many calculators will also accumulate sums of squares.

I. Most of the time, however, we just use a computer program.
   1. In MINITAB look under basic statistics and then correlation.

VII. OTHER FACTORS THAT AFFECT CORRELATION AND REGRESSION:

A. Regression analysis is sensitive to extremely deviant or "leverage" values. For example consider these two plots:

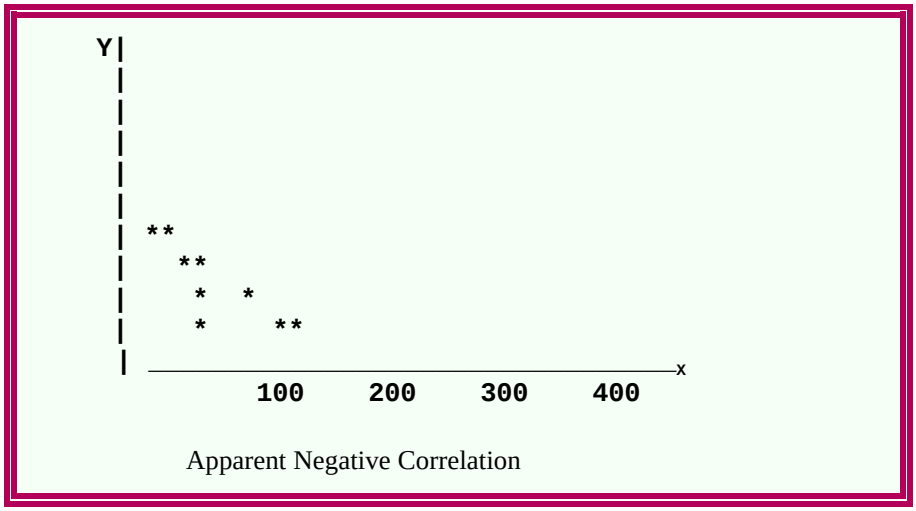


Figure 6: Apparent Negative Correlation

1. Now consider the effect of a "deviant" value.
2. In this plot the linear relationship is negative and both $\hat{\beta}_1$ and r will have negative signs. Furthermore, linear association is quite strong.
3. Now lets add a single very deviant case.

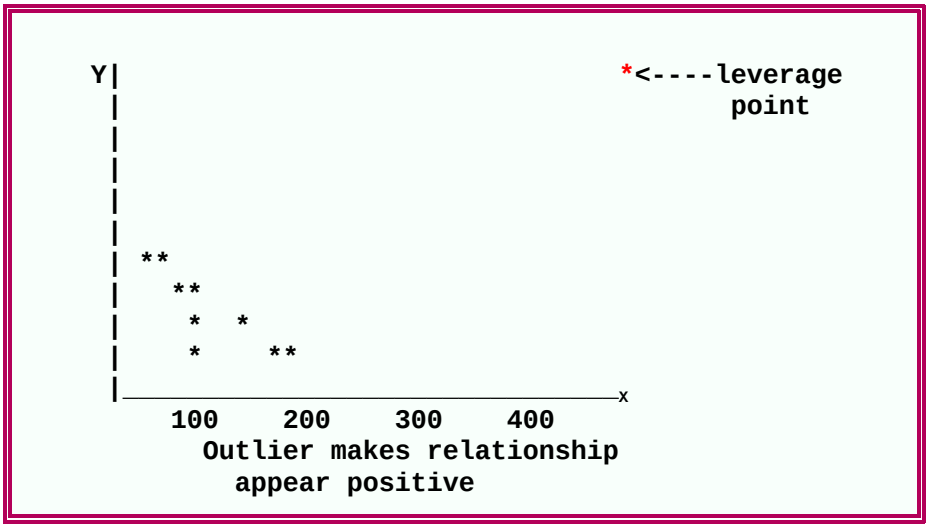


Figure 7: The Effect on an Outlier

B. Now if you were to do regression analysis, you would find that X and Y were **positively** related: that is $\hat{\beta}_1$ and r would have plus signs. In addition, the relationship might be quite strong, but it would be very misleading.

C. Moral and suggestion:
   1. Always look at a scatter plots for leverage points or cases that have large residuals. The MINITAB identifies these for you.
   2. Remove the "offending" case or cases if there are two or three and recompute the analysis.

VIII. THE MAGNITUDE OF COEFFICIENTS:

A. The correlation coefficient, r, can be thought of as a standardized regression coefficient: it measures the change in Y (in standard deviations) for a one standard deviation change in X.
   1. This feature or property supposedly allows one to compare the explanatory power variables.
   2. Suppose, for example, as we wanted to determine which variable--family background or IQ--explain most of the variation in achievement.
      i. Incidentally, this is a task undertaken in Murray and Hernstein's ***The Bell Curve***, a very controversial book.
   3. Suppose the correlation between achievement (as measured by, say, income or occupational status) and father's education or income was .4 whereas the correlation between achievement and IQ or aptitude scores was .8.
   4. It might be tempting to conclude that IQ was **twice** as "important" in explaining achievement as family background.
   5. But for reasons mentioned above and later in the semester, this conclusion could be very misleading.

B. The regression coefficient, β, is measured in units of the dependent variable. But its magnitude is affected by X's measurement scale.
   1. Example: suppose we want to know how strongly GNP per capita affects living standards as measured by percent of the population that is literate.
   2. Suppose the estimated β turns out to be .005. This might look like a small number.
      i. After all, a "one-unit change" in X (GNP per capita) only "leads" to a .005 percent increase in literacy.
      ii. But a one unit change in X is not very large; it's only $1.
      iii. So suppose we looked at a 100 unit change in X. That is, suppose we asked what effect a $100 increase in per capita income would produce.
         a) It would be 100 X .005 = .5 percent increase in literacy.
      iv. And a $1,000 increase in GNP per capita would lead to or be associated with a 5 percent increase. That's quite substantial.
   3. In other words, a country with a GNP per capita of $4,000 does not really differ from one having a GNP of $4,001. So their literacy rates wouldn't differ much, even though income does strongly affect well being.
      i. To see this compare a country with GNP per capita of $4,000 and

one with $5,000. The predicted increase in literacy would be 5 percent.

C. The lessons are:
   1. Interpret r cautiously when making claims about the strength and importance of relationships.
   2. Keep the measurement scales of X and Y in mind when thinking about the effects of the former on the latter.

IX. NEXT TIME:
   A. Multiple regression
   B. Partial regression coefficients

Go to Notes page

Go to Statistics page