Introduction to Social Statistics
Handout on Correlations and Regression

Correlation Matrix

- - Correlation Coefficients - -

	CRIME	GRAD	INCOME	MANUF	TEACHSAL	UNEMP	URBAN
CRIME	1.0000						
	P= .						
GRAD	-.3200	1.0000					
	P= .024	P= .					
INCOME	.1916	.4283	1.0000				
	P= .183	P= .002	P= .				
MANUF	-.0028	-.4504	-.1819	1.0000			
	P= .985	P= .001	P= .206	P= .			
TEACHSAL	.2329	.3374	.8317	.0105	1.0000		
	P= .104	P= .017	P= .000	P= .942	P= .		
UNEMP	.5807	-.0732	.2936	.0771	.3553	1.0000	
	P= .000	P= .613	P= .039	P= .595	P= .011	P= .	
URBAN	.6059	.0133	.5910	.0515	.5753	.3992	1.0000
	P= .000	P= .927	P= .000	P= .723	P= .000	P= .004	P= .

(Coefficient / (Cases) / 2-tailed Significance) " . " is printed if a coefficient cannot be computed

Interpretation of p-values: The p-values represent the probability of obtaining a correlation coefficient equal to or greater than the computed coefficient if in fact the correlation coefficient in the population was equal to zero. Small p-values, say less than 0.05 or 0.10, indicate that the correlation coefficient is significantly different from zero.

PLOTTING REGRESSION LINES

A. SIMPLE LINEAR REGRESSION

To correctly plot a regression line, the first step is to examine the ranges of your variables. This range should be reflected in the scaling of both the dependent and independent variables.

Always include a descriptive label for the X-axis and the Y-axis, as well tick marks for the scale of the variables.

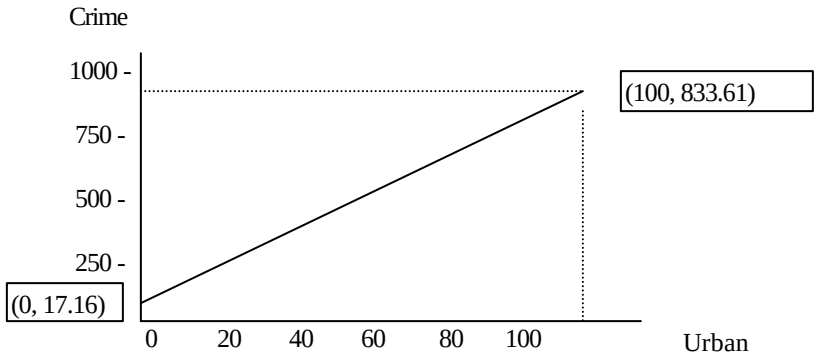
To plot a linear function of the form, $Y = a + bX$, you simply need to calculate two points (x_1, y_1) and (x_2, y_2) and draw a line between these two points. One obvious choice for a

point is the point where the line intersects the Y-axis, namely (0, a). When you compute the second point, it is a good idea to use a value of X near its maximum (or minimum).

Consider the following linear relation between violent crime rates and levels of urbanization across the 50 states:

CRIME RATE = 17.61 + 8.16 (URBAN)

The first point to is (0, 17.61), the Y-intercept. To determine my second point, I use X=100, so Y= 17.61 + (8.16) (100) = 1769.16, thus (100,833.61).



2. Linear Model with Dichotomous Variable

Consider the following SPSS output, where CRIME is the dependent variable and South is a dichotomous variable coded as 1 if the state is in the South, and 0 otherwise:

Variable	B	SE B	Beta	T	Sig T
URBAN	8.539922	1.423101	.633966	6.001	.0000
SOUTH	216.146495	68.008616	.335762	3.178	.0026
(Constant)	-68.120065	103.009107		-.661	.5116:

These results imply the following relation:

CRIME = -68.12 + 8.54(URBAN) + 216.14 (SOUTH)

To plot this function, it is necessary to find four points, as follows:

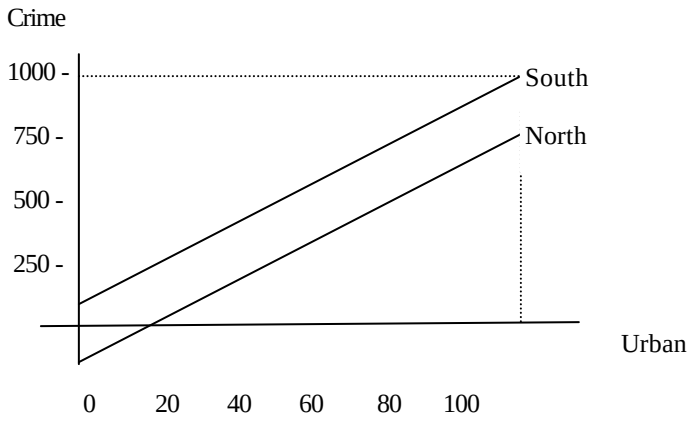
a. if South = 1 and URBAN=0:
(0, 148.2) since -68.12 + 8.54(0) + 216.14 (1) = 148.2

b. if South = 1 and URBAN=100:
(100, 1002.2) since -68.12 + 8.54(100) + 216.14(1) = 1000.2

c. if South = 0 and URBAN=0:
(0, -68.12) since -68.12 + 8.54(0) + 216.14 (0) = -68.12

d. if South = 0 and URBAN=100:
(100, 785.8) since -68.12 + 8.54(100) + 216.14(1) = 785.8

Points (a) and (b) define the line for the southern states, and points (c) and (d) define the line for the non-southern states.



3. Model with Interaction Term

Finally, Consider SPSS output a model which allows for a different relation between URBAN and CRIME in different regions:

Variable	B	SE B	Beta	T	Sig T
URBAN	7.982784	1.564557	.592607	5.102	.0000
SOUTH	2.491735	255.403147	.003871	.010	.9923
INTERACT	3.310600	3.813846	.343461	.868	.3899
(Constant)	-30.315112	112.087328		-.270	.7880

These results imply the following relation:

CRIME = -30.32 + 7.98(URBAN) + 2.49 (SOUTH) + 3.31 (SOUTH*URBAN)

To plot this function, it is necessary to find four points, as follows:

a. if South = 1 and URBAN=0:

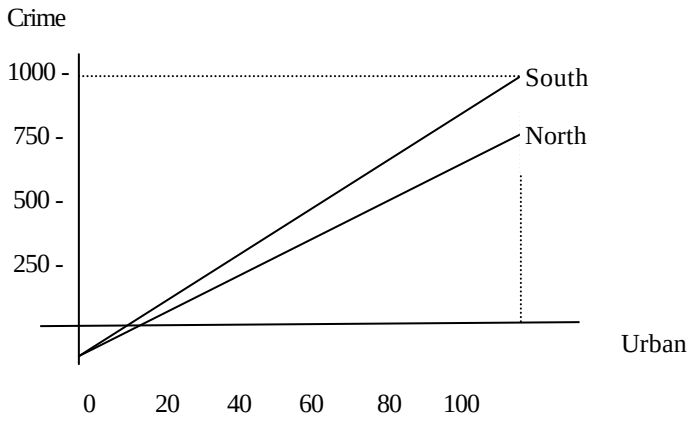
(0, -27.83) since -30.32 + 7.98(0) + 2.49(1) + 3.31(1)(0) = -27.83

b. if South = 1 and URBAN=100:
(100, 1101.17) since -30.32 + 7.98(100) + 2.49(1) +
3.31(1)(100)=1101.17

c. if South = 0 and URBAN=0:
(0, -30.32) since -30.32 + 7.98(0) + 2.49(0) +3.31 (0)(0) = -30.32

d. if South = 0 and URBAN=100:
(100, 830.81) since -30.32 + 7.98(100) + 2.49(1)
+ 3.31(0)(100)= 830.81

Points (a) and (b) define the line for the southern states, and points (c) and (d) define the line for the non-southern states.



Interpretation of Coefficients
Consider dividing the states into two sub-groups, the Southern states and the non-Southern states, and then running separate simple linear regressions for each group:

For the Southern States: $CRIME = a_s + b_s(URBAN)$
For the Northern (Non-Southern) States : $CRIME = a_n + b_n(URBAN)$

The coefficients for each group of states are available from the interaction model:

$CRIME = -30.32 + 7.98(URBAN) + 2.49\ (SOUTH) + 3.31\ (SOUTH*URBAN)$
Intercept for Southern States: $a_s = -30.32 + 2.49$
Slope for Southern States: $b_s = 7.98 + 3.31$

Intercept for Northern States: $a_n = -30.32$
Slope for Southern States: $b_n = 7.98$

The coefficient of the interaction term, 3.31, is an estimate of the difference across regions in the effect of levels of urbanization on violent crime rates. Note that although the *difference* is not significant (t=0.868, p=.389), the effects of urbanization on crime are highly significant in each region.

Introduction to Social Statistics

A. MODEL WITH INTERACTION TERMS

Consider four different models:

1. CRIME = a
2. CRIME = a + b_1(URBAN)
3. CRIME = a + b_1(URBAN) + b_2 (SOUTH)
4. CRIME = a + b_1(URBAN) + b_2 (SOUTH) + b_3 (SOUTH*URBAN)

NOTE:

A. The third model can be obtained from the fourth, by setting b_3 equal to zero. Since b_3 is a measure of the *difference* in the relation between CRIME and URBAN between the Southern and Non-Southern states, setting b_3 equal to zero implies that the regression slope is the same across region.
B. If both b_2 and b_3 are set equal to zero, then we have the second model that implies that there are no differences in crime rates across regions.
C. If all coefficients aside from the intercept (a) are set equal to zero, then we have the first model, which would simply be the mean violent crime rate for our population.

INTERPRETING COEFFICIENTS OF DETERMINATION: INHERITABILITY OF "INTELLIGENCE"

Simple Model:

Parent's "IQ" ——————→ Child's "IQ"

According to many researchers, the heritability of IQ falls somewhere between 0.40 and 0.80. In statistical terms, these are estimates of the coefficient of determination, r^2. These types of results are easily misunderstood, with potentially serious policy implications.

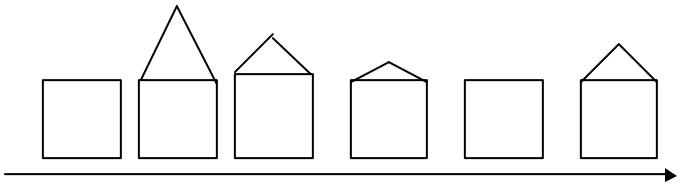
An incorrect interpretation of the coefficient of determination:

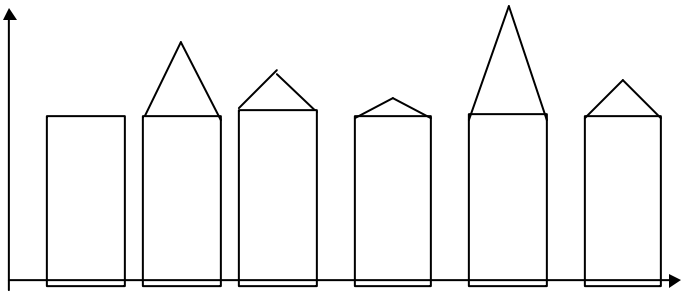
"When I -- when we -- say 60 percent heritability, it's not 60 percent of the variation. It is 60 percent of the IQ in any given person." Charles Murray quoted in "Dumb Bell" (TRB from Washington, THE NEW REPUBLIC, JAN. 2, 1995: 6).

A correct interpretation:

"Heritability ... measures the relative contribution of genes to the variation observed in a trait. It makes no more sense to talk about heritability of an individual's IQ than it does to talk about his birthrate." Richard Herrnstein and (?) Charles Murray, *The Bell Curve*, p. 106.
That is, about 60 percent of the variation in "child's IQ" is explained by genetic factor's ("Parent's IQ"). Or, more technically, the conditional distribution of Y for a given X has about a 60 percent smaller variance than the marginal distribution.

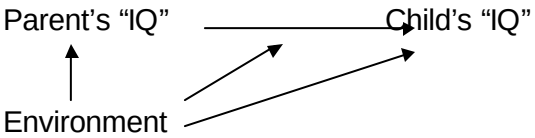
The Different Shape of Roofs:





Do these findings support a policy of "benign neglect"? Should we stop funding programs, such as Head Start, whose aim is to improve the academic (cognitive) of poor children? Should we end affirmative action in education and at the work place?

A More Complicated Model:



Phenylketonuria, Before and After Treatment

		Before Treatment Trait in Parent		After Treatment Trait in Parent	
		Yes	No	Yes	No
Trait in Child	Yes	100%	0%	0%	100%
	No	0%	100%	100%	0%