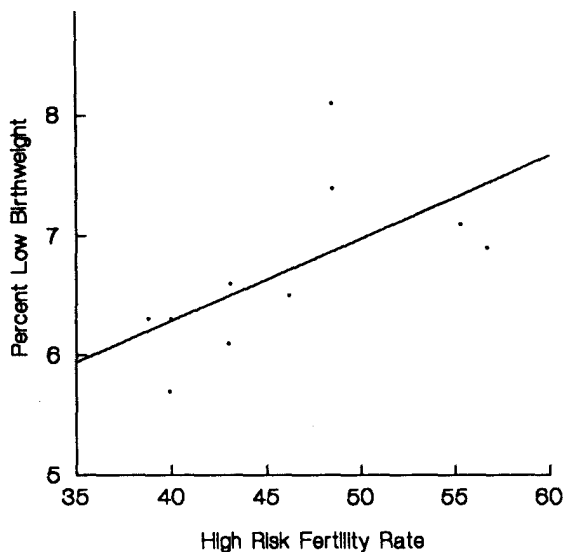


Chapter 9 - Correlation and Regression

- 9.1 Scatter diagram of percentage of LBW infants (Y) and high-risk fertility rate (X₁) in Vermont Health Planning Districts.

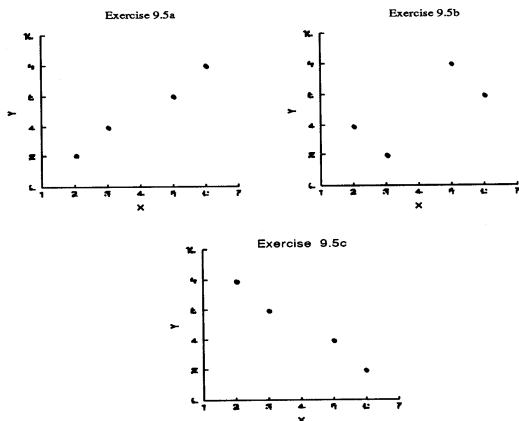


- 9.3 Correlation between percentage of LBW infants (Y) and percentage of births to unmarried mothers (X₂) in Vermont Health Planning Districts.

$$N = 10 \quad s_Y = 0.698 \quad s_{X_2} = 1.322 \quad \text{cov}_{X_2Y} = 0.3189$$

$$r = \frac{\text{COV}_{X_2Y}}{s_{X_2}s_Y} = \frac{0.3189}{(1.322)(0.698)} = .35$$

- 9.5 Three sets of data:



9.7 Correlations for the three data sets in Exercise 9.5:

a.

$$r = \frac{\text{COV}_{XY}}{s_X s_Y} \quad \text{For each set, } s_X = 1.826 \text{ and } s_Y = 2.582$$

Set 1:

$$r = \frac{\text{COV}_{XY}}{s_X s_Y} = \frac{4.67}{(1.826)(2.582)} = .99$$

Set 2:

$$r = \frac{\text{COV}_{XY}}{s_X s_Y} = \frac{3.33}{(1.826)(2.582)} = .71$$

Set 3:

$$r = \frac{\text{COV}_{XY}}{s_X s_Y} = \frac{-4.67}{(1.826)(2.582)} = -.99$$

b. Three arrangements of Y will result in the lowest possible positive correlation:

$$2 \ 8 \ 6 \ 4 \quad \text{or} \quad 6 \ 4 \ 2 \ 8 \quad \text{or} \quad 6 \ 2 \ 8 \ 4 \quad [r = .14]$$

9.9 Cerebral hemorrhage in low-birthweight infants and cognitive deficit at age 5:

a. Power calculation:

$$d = .20$$

$$\delta = d \sqrt{N-1}$$

$$= .20 \sqrt{25-1}$$

$$= 0.980 \quad \text{power} = .17$$

b. For power = .80, $\delta = 2.8$

$$\delta = d \sqrt{N-1}$$

$$2.8 = .20 \sqrt{N-1}$$

$$N = 197$$

9.11 Standard error of estimate for regression equation in Exercise 9.10:

$$N = 10 \quad r = .62 \text{ [Calculated in Exercise 9.2]}$$

$$s_{Y.X} = s_Y \sqrt{(1-r^2) \frac{N-1}{N-2}} = .698 \sqrt{(1-.62^2)(9/8)} = .580$$

9.13 If high-risk fertility rate in Exercise 9.10 jumped to 70:

$$\begin{aligned} \hat{Y} &= bX + a \\ &= .069(70) + 3.53 \\ &= 8.36 \text{ would be the predicted percentage of LBW infants born.} \end{aligned}$$

9.15 Number of symptoms predicted for a stress score of 8 using the data in Table 9.2 :

$$\text{Regression equation: } \hat{Y} = 0.0086(X) + 4.30$$

$$\text{If Stress score (X) = 8: } \hat{Y} = 0.0086(8) + 4.30$$

$$\text{Predicted ln(symptoms) score is : } \hat{Y} = 4.37$$

9.17 Confidence interval on \hat{Y}

I will calculate them for X incrementing between 0 and 60 in steps of 10

$$CI(Y) = \hat{Y} \pm t_{\alpha/2} (s'_{Y.X})$$

$$s'_{Y.X} = s_{Y.X} \sqrt{1 + \frac{1}{N} + \frac{(X_i - \bar{X})^2}{(N-1)s_X^2}} = 0.1726 \sqrt{1 + \frac{1}{107} + \frac{(X_i - \bar{X})^2}{106(156.05)}}$$

$$\hat{Y} = 0.00856X + 4.30$$

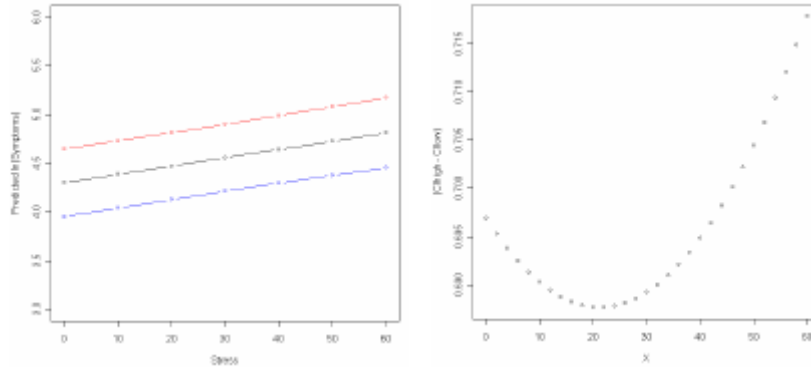
$$t_{\alpha/2} = 1.983$$

For X from 0 to 60 in steps of 10, $s'_{Y.X} =$
0.1757 0.1741 0.1734 0.1738 0.1752 0.1776 0.1810

$$CI(Y) = \hat{Y} \pm (t_{\alpha/2})(s'_{Y.X})$$

For several different values of X, calculate \hat{Y} and $s'_{Y.X}$ and plot the results.

X =	0	10	20	30	40	50	60
$\hat{Y} =$	4.300	4.386	4.471	4.557	4.642	4.728	4.814



The curvature is hard to see, but it is there, as can be seen in the graphic on the right, which plots the width of the interval as a function of X . (It's fun to play with R).

9.19 When data are standardized, the slope equals r . Therefore the slope will be less than one for all but the most trivial case, and predicted deviations from the mean will be less than actual parental deviations.

9.21 Number of subjects needed in Exercise 9.20 for power = .80:

For power = .80, $\delta = 2.80$

$$\delta = \rho_1 \sqrt{N-1}$$

$$2.80 = .40 \sqrt{30-1}$$

$$N = 50$$

9.23 Katz et al. correlations with SAT scores.

a. $r_1 = .68$ $r_1' = .829$

$r_2 = .51$ $r_2' = .563$

$$z = \frac{r_1' - r_2'}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}} = \frac{.829 - .563}{\sqrt{\frac{1}{14} + \frac{1}{25}}} = 0.797$$

The correlations are not significantly different from each other.

b. We do not have reason to argue that the relationship between performance and prior test scores is affected by whether or not the student read the passage.

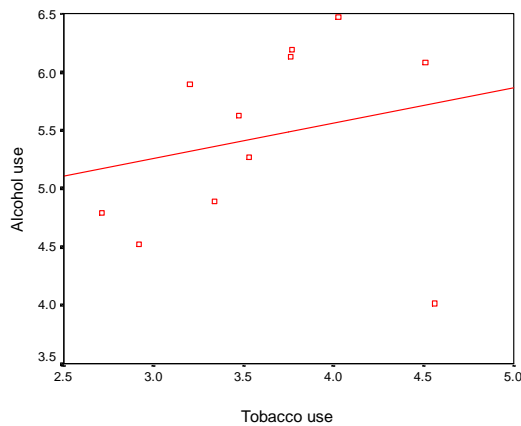
9.25 It is difficult to tell whether the significant difference between the results of the two previous problems is to be attributable to the larger sample sizes or the higher (and thus more different) values of r' . It is likely to be the former.

9.27 Moore and McCabe example of alcohol and tobacco use:

Correlations

		ALCOHOL	TOBACCO
ALCOHOL	Pearson Correlation	1.000	.224
	Sig. (2-tailed)	.	.509
	N	11	11
TOBACCO	Pearson Correlation	.224	1.000
	Sig. (2-tailed)	.509	.
	N	11	11

b. The data suggest that people from Northern Ireland actually drink relatively little.

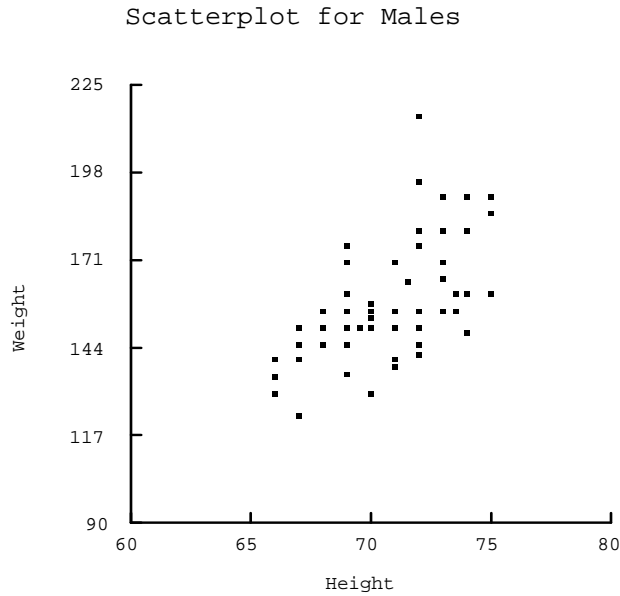


c. With Northern Ireland excluded from the data the correlation is .784, which is significant at $p = .007$.

9.29 a. The correlations range between .40 and .80.

b. The subscales are not measuring independent aspects of psychological well-being.

9.31 Relationship between height and weight for males:



The regression solution that follows was produced by SPSS and gives all relevant results.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.604 ^a	.364	.353	14.9917

a. Predictors: (Constant), HEIGHT

b. Gender = Male

ANOVA^{b,c}

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7087.800	1	7087.800	31.536	.000 ^a
	Residual	12361.253	55	224.750		
	Total	19449.053	56			

a. Predictors: (Constant), HEIGHT

b. Dependent Variable: WEIGHT

c. Gender = Male

Coefficients^{a,b}

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-149.934	54.917		-2.730	.008
	HEIGHT	4.356	.776	.604	5.616	.000

a. Dependent Variable: WEIGHT

b. Gender = Male

With a slope of 4.36, the data predict that two males who differ by one inch will also differ by approximately $4 \frac{1}{3}$ pounds. The intercept has no meaning because people are not 0 inches tall, but the fact that it is so largely negative suggests that there is some curvilinearity in this relationship for low values of Height.

Tests on the correlation and the slope are equivalent tests when we have one predictor, and these tests tell us that both are significant. Weight increases reliably with increases in height.

9.33 As a 5'8" male, my predicted weight is $\hat{Y} = 4.356(\text{Height}) - 149.934 = 4.356 \cdot 68 - 149.934 = 146.27$ pounds.

a. I weigh 146 pounds. (Well, I did two years ago.) Therefore the residual in the prediction is $Y - \hat{Y} = 146 - 146.27 = -0.27$.

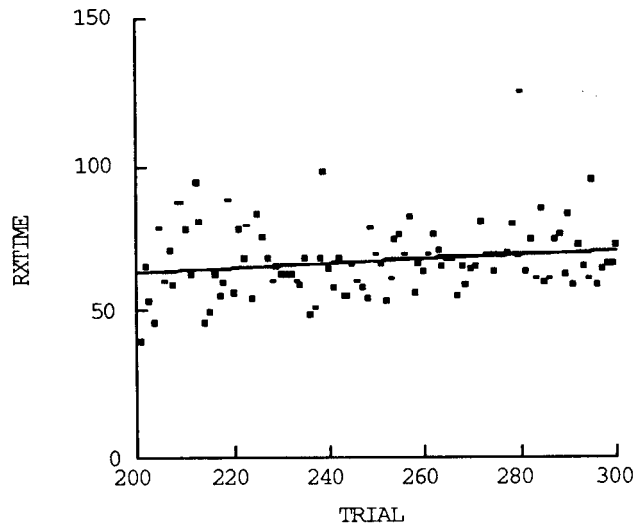
b. If the students on which this equation is based under- or over-estimated their own height or weight, the prediction for my weight will be based on invalid data and will be systematically in error.

9.35 The male would be predicted to weigh 137.562 pounds, while the female would be predicted to weigh 125.354 pounds. The predicted difference between them would be 12.712 pounds.

9.37 Independence of trials in reaction time study.

The data were plotted by "trial", where a larger trial number represents an observation later in the sequence.

RxTime as a Function of Trials



Although the regression line has a slight positive slope, the slope is not significantly different from zero. This is shown below.

DEP VAR: TRIAL N: 100 MULTIPLE R: 0.181 SQUARED MULTIPLE R: 0.033
 ADJUSTED SQUARED MULTIPLE R: 0.023 STANDARD ERROR OF ESTIMATE: 28.67506

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	221.84259	15.94843	0.00000	.14E+02	.10E-14	
RXTIME	0.42862	0.23465	0.18146	1.00000	1.82665	0.07080

ANALYSIS OF VARIANCE					
SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	2743.58452	1	2743.58452	3.33664	0.07080
RESIDUAL	80581.41548	98	822.25934		

There is not a systematic linear or cyclical trend over time, and we would probably be safe in assuming that the observations can be treated as if they were independent. Any slight dependency would not alter our results to a meaningful degree.