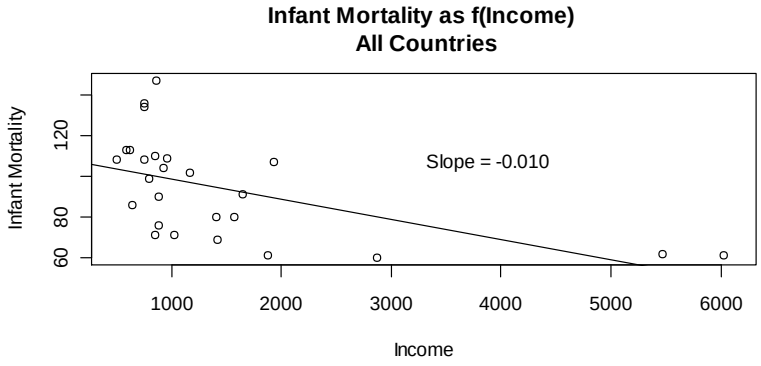
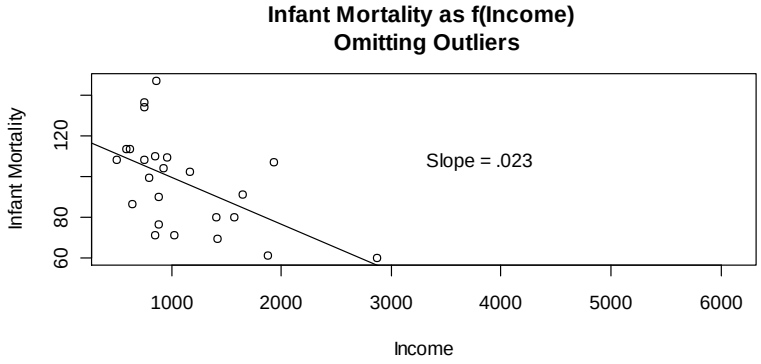
Chapter 9 - Correlation and Regression

9.1 Infant Mortality in Sub-Saharan Africa

a. & b.





c. Those two points would almost certainly draw the line toward them, which will flatten the slope. If we remove those countries we have the second graph with a steeper slope.

9.3 Significance of correlations

The minimum sample size in this example is 25, and we will use that. We would need $t = 2.069$ for a two-tailed test on $N - 2 = 23$ *df*. A little (well, maybe a lot) of algebra will show that a correlation of .396 will produce that t value.

9.5 If we put these two predictors together using methods covered in Chapter 15, the multiple correlation will be .58, which is only a small amount higher than Income alone.

9.7 I suspect that a major reason why this variable does not play a more important role is the fact that it has very little variance. The range is 3% - 7%. One cause of this may be the very high death rate among women in sub-saharan Africa. There are many fewer women giving birth at ages above 40. To quote from a United Nations report (http://www.un.org/ecosocdev/geninfo/women/women96.htm):

- Women are becoming increasingly affected by HIV. Today about 42 per cent of estimated cases are women, and the number of infected women is expected to reach 15 million by the year 2000.

- An estimated 20 million unsafe abortions are performed worldwide every year, resulting in the deaths of 70,000 women.

- Approximately 585,000 women die every year, over 1,600 every day, from causes related to pregnancy and childbirth. In sub-Saharan Africa, 1 in 13 women will die from pregnancy or childbirth related causes, compared to 1 in 3,300 women in the United States.

- Globally, 43 per cent of all women and 51 per cent of pregnant women suffer from iron-deficiency anemia.

9.9 Psychologists are very much interested in studying variables related to behavior and in finding ways to change behavior. I would guess that they would have a good deal to say about educating women in ways that would decrease infant mortality.

9.11 The relationship is decidedly curvilinear, and Pearson's *r* is a statistic on linear relationships.

9.13 Power for $n = 25$, $\rho = .20$

$d = \rho_1 = .20$

$\delta = \rho_1\sqrt{N-1} = .20\sqrt{24} = 0.98$

power $\approx$ .17

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

Regression equation: $Y = 0.0086(X) + 4.30$

If Stress score (X) = 8: $Y = 0.0086(8) + 4.30$

Predicted ln(symptoms) score is : $Y = 4.37$

9.17 Confidence interval on Y :

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

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

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

$$Y = 0.00856X + 4.30$$
$$t_{\alpha/2} = 1.983$$

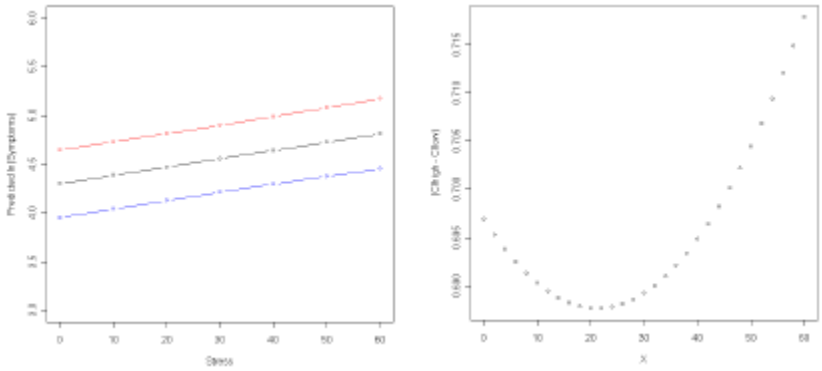
For X from 0 to 60 in steps of 10, s'$_{Y.X}$ =
0.1757 0.1741 0.1734 0.1738 0.1752 0.1776 0.1810

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

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

X = 0 10 20 30 40 50 60

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 Galton's data

a.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	23.942	2.811		8.517	.000
	midparent	.646	.041	.459	15.711	.000

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	23.942	2.811		8.517	.000
	midparent	.646	.041	.459	15.711	.000

a. Dependent Variable: child

b. Predicted height = 0.646*(Midparent) + 23.942

c. Child Means

Descriptives

child

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	392	67.12	2.247	.113	66.90	67.35
2	219	68.02	2.240	.151	67.72	68.32
3	183	68.71	2.465	.182	68.35	69.06
4	134	70.18	2.269	.196	69.79	70.57
Total	928	68.09	2.518	.083	67.93	68.25

Parent means

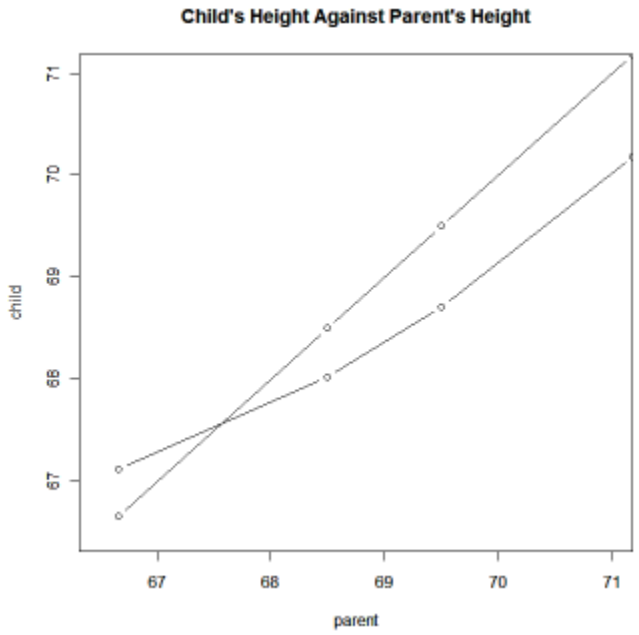
Descriptives

midparent

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	392	66.66	1.068	.054	66.56	66.77
2	219	68.50	.000	.000	68.50	68.50
3	183	69.50	.000	.000	69.50	69.50
4	134	71.18	.786	.068	71.04	71.31
Total	928	68.31	1.787	.059	68.19	68.42

d. Parents in the highest quartile have a mean of 71.18, while their children have a mean of 70.18. Those parents in the lowest quartile have a mean of 66.66, while their children have a mean of 67.14. This is what we would expect to happen.

e.



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

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

$$\delta = \rho_1\sqrt{N-1}$$
$$2.80 = .40\sqrt{N-1}$$
$$\sqrt{N-1} = 2.80/.40 = 7$$
$$N = 50$$

9.23 Katz et al. correlations with SAT scores.

a. $r_1 = .68 \quad r_1' = .829$
$r_2 = .51 \quad r_2' = .563$

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

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.

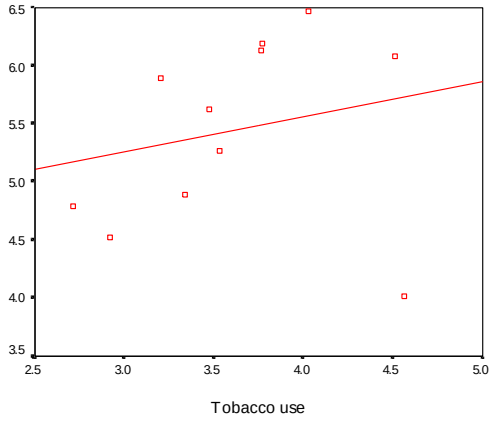
9.25 It is difficult to tell whether the significant difference between the results of the two previous problems is to be attributable to the larger sample sizes or the higher (and thus more different) values of r'. It is likely to be the former.

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.

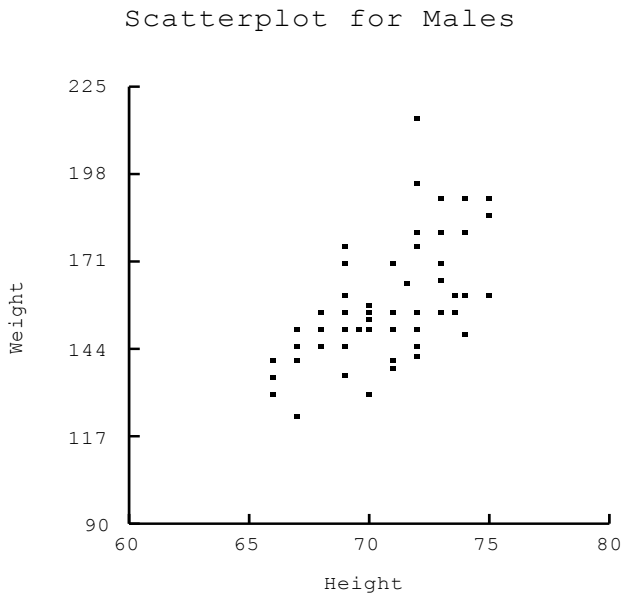


c. With Northern Ireland excluded from the data the correlation is .784, which is significant at $p = .007$.

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 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(Height) - 149.934 = 4.356*68 - 149.934 = 146.27 pounds.

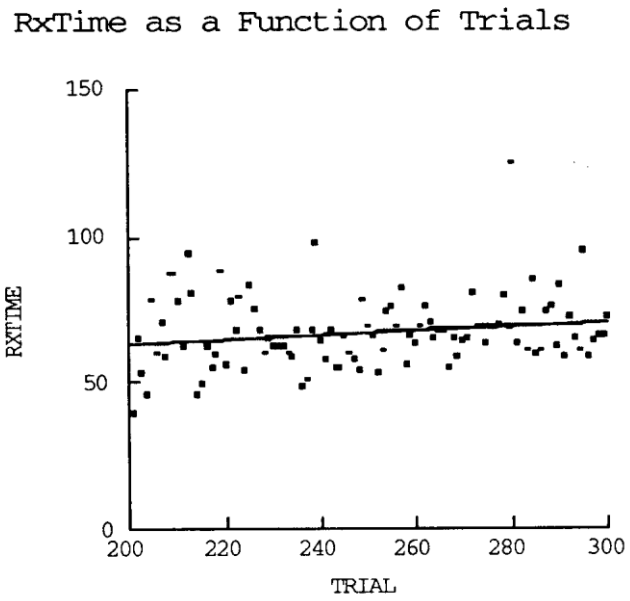
a. I weigh 146 pounds. (Well, I did two years ago.) Therefore the residual in the prediction is Y-$\hat{Y}$ = 146 - 146.27 = -0.27.

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.



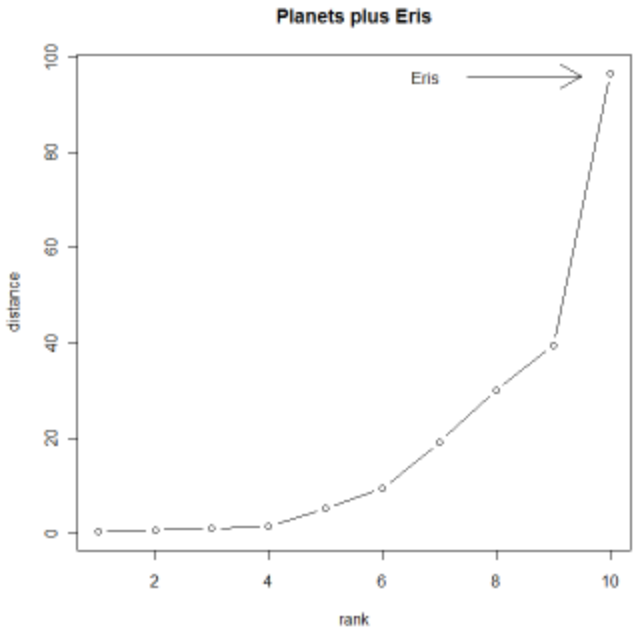
Although the regression line has a slight positive slope, the slope is not significantly different from zero. This is shown below.

```
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.

9.39 What about Eris?

Eris doesn't fit the plot as well as I would have liked. It is a bit too far away.



9.41 Comparing correlations in males and females.

$$z = \frac{r_1' - r_2'}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$

$$= \frac{.648 - .343}{\sqrt{\frac{1}{284} + \frac{1}{222}}} = \frac{.305}{\sqrt{0.0085}} = \frac{.305}{.092}$$

$$= 3.30$$

The difference between the two correlations is significant.