

# Chapter 10-Regression

10.1 Regression equation predicting low birthweight from high-risk fertility.

$Y$  = Low Birthweight Percentage

$X$  = High-risk Fertility

$$\bar{Y} = 6.70 \quad s_Y = 0.698 \quad s_Y^2 = 0.487$$

$$\bar{X} = 46.00 \quad s_X = 6.289 \quad s_X^2 = 39.553$$

$$\text{cov}_{XY} = 2.7245$$

$$b = \frac{\text{cov}_{XY}}{s_X^2} = \frac{2.7245}{39.553} = 0.069$$

$$a = \bar{Y} - b\bar{X} = 6.70 - (0.069)(46.00) = 3.53$$

$$\hat{Y} = 0.069(X) + 3.53$$

10.3 If the high risk fertility rate jumped to 70, we would predict that the incidence of birthweight < 2500gr would go to 8.35.

$$\hat{Y} = bX + a = 0.0689X + 3.53$$

$$= 0.0689 * 70 + 3.53 = 8.35$$

This assumes that there is a causal relationship, which is plausible in some ways, but not proven.

It may be trivial to point this out, but here we have a real world situation where we can say something about changing trends in society and their possible effects.

10.5 I would be more comfortable speaking about the effects on Senegal because it is already at approximately the mean income level and we are not extrapolating for an extreme country.

This may have little to do with a statistics course in psychology, but there have been some noticeable improvements in infant mortality in Senegal, and one device that has made a difference is a warm table on which newborn infants can be placed. This may interest students who probably think of advances in medicine in terms of MRIs.

[http://transition.usaid.gov/stories/senegal/pc\\_sn\\_infant.html](http://transition.usaid.gov/stories/senegal/pc_sn_infant.html)

10.7 Prediction of Symptoms score for a Stress score of 45:

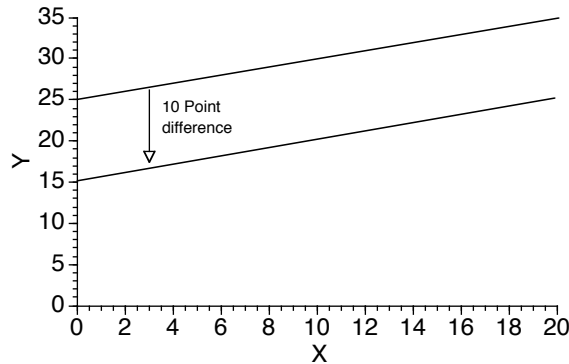
$$\text{Regression equation: } \hat{Y} = 0.7831X + 73.891$$

$$\text{If } X = 45: \quad = 0.7831 * 45 + 73.891$$

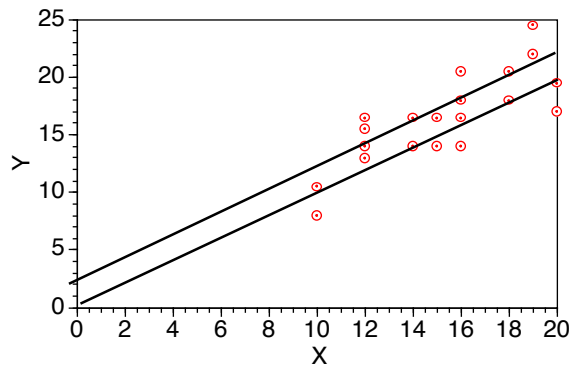
$$\text{Predicted Symptoms} \quad = 109.13$$

10.9 Subtracting 10 points from every  $X$  or  $Y$  score would not change the correlation in the slightest. The relationship between  $X$  and  $Y$  would remain the same.

10.11 Diagram to illustrate Exercise 10.10:



10.13 Adding a constant to  $Y$ :



- From this figure you can see that adding 2.5 to  $Y$  simply raised the regression line by 2.5 units.
- The correlation would be unaffected.

10.15 Predicting GPA ( $Y$ ) from ADDSC ( $X$ ):

$$b = \frac{\text{cov}_{XY}}{s_X^2} = \frac{-6.580}{154.431} = -0.0426$$

$$a = \bar{Y} - b\bar{X} = 2.456 - 0.0426 * 52.602 = 4.699$$

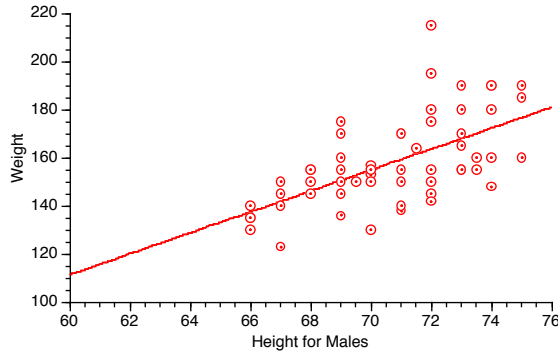
$$\hat{Y} = -0.0426X + 4.699$$

When we first collected these data I was somewhat disheartened by how well we were doing (and to some extent I still am). We can take a measure in elementary school that is quickly filled out by the teacher, and make an excellent prediction about how the student will do in high school. That

may be nice statistically, but I don't think we like to feel that children are that locked in. This is worth getting the students to think about.

10.17 The correlation dropped to  $-.478$  when I added and subtracted  $.04$  from each  $Y$  value. This drop was caused by the addition of error variance. The slope should remain the same.

10.19 Weight as a function of height for males:



The regression solution that follows is a modification of printout from SPSS.

Equation Number 1		Dependent Variable..	WEIGHT		
Variable(s) Entered on Step Number					
1..	HEIGHT				
Multiple R	.60368				
R Square	.36443				
Adjusted R Square	.35287				
Standard Error	14.99167				
Analysis of Variance					
	DF	Sum of Squares	Mean Square		
Regression	1	7087.79984	7087.79984		
Residual	55	12361.25279	224.75005		
F =	31.53637	Signif F =		.0000	
----- Variables in the Equation -----					
Variable	B	SE B	Beta	T	Sig T
HEIGHT	4.355868	.775656	.603680	5.616	.0000
(Constant)	-149.933617	54.916943		-2.730	.0085

- The intercept is given as the “constant” and is  $-149.93$ , which has no interpretable meaning with these data. The slope of  $4.356$  tells us that a one-unit increase in height is associated with a  $4.356$  increase in weight.
- The correlation is  $.60$ , telling us that for females  $36\%$  of the variability in weight is associated with variability in height.
- Both the correlation and the slope are significantly different from  $0$ , as shown by an  $F$  of  $31.54$  and a (equivalent)  $t$  of  $5.616$ .

10.21 Predicting my own weight, for which I use the equation from Exercise 10.19:

$$\hat{Y} = 4.356 * \text{height} - 149.93$$

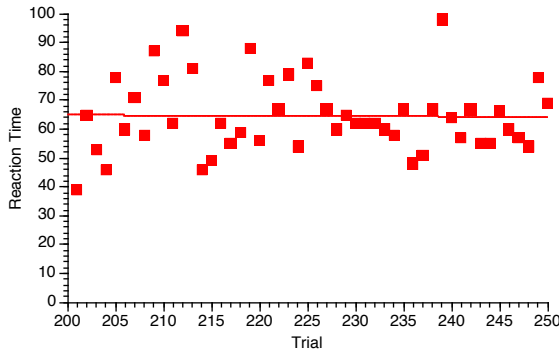
$$\hat{Y} = 4.356 * 68 - 149.93 = 146.28$$

- a) The residual is  $Y - \hat{Y} = 156 - 146.28 = 9.72$ . (Actually, I now weigh about 10 pounds less.)
- b) If the students who supplied the data gave biased responses, then, to the degree that the data are biased, the coefficients are biased and the prediction will not apply accurately to me.

10.23 Predictions for a 5'6" male and female

For the male,  $\hat{Y} = 4.356 * 66 - 149.93 = 137.57$   
 For a female,  $\hat{Y} = 2.578 * 66 - 44.859 = 125.29$   
 Difference = 12.28 pounds

10.25 Plot of Reaction Time against Trials for only the Yes/5-stimuli trials:



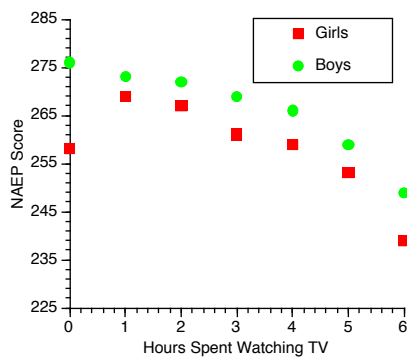
The following regression solution is a modification of SPSS printout.

Equation Number 1	Dependent Variable..	RXTIME
Variable(s) Entered on Step Number		
1..	TRIAL	
Multiple R	.01640	
R Square	.00027	
Adjusted R Square	-.02056	
Standard Error	12.76543	
Analysis of Variance		
	DF	Sum of Squares
Regression	1	2.10363
Residual	48	7821.89637
		Mean Square
		2.10363
		162.95617
F =	.01291	Signif F = .9100

----- Variables in the Equation -----					
Variable	B	SE B	Beta	T	Sig T
TRIAL	-.014214	.125100	-.016397	-.114	.9100
(Constant)	67.805186	28.267795		2.399	.0204

The slope is only -0.014, and it is not remotely significant. For this set of data we can conclude that there is not a linear trend for reaction times to change over time. From the scatterplot above we can see no hint that there is any nonlinear pattern, either.

### 10.27 The evils of television:



Regression equations:

$$\text{Boys } \hat{Y} = -4.821X + 283.61$$

$$\text{Girls } \hat{Y} = -3.460X + 268.39$$

b) The slopes are roughly equal, given the few data points we have, with a slightly greater decrease with increased time for boys. The difference in intercepts reflects the fact that the line for the girls is about 9 points below that for boys.

c) Television can not be used as an explanation for poorer scores in girls, because we see that girls score below boys even when we control for television viewing.

I have not made much of the idea of “controlling for a variable.” This would be useful to elaborate, especially to make sure that they know where to look on the graph to see this (i.e. at each value of  $X$  taken one at a time).

### 10.29 Draw a scattering of 10 data points and drop your pencil on it.

- b) As you move the pencil vertically you are changing the intercept.
- c) As you rotate the pencil you are changing the slope.
- d) You can come up with a very good line simply by rotating and raising or lowering your pencil so as to make the deviations from the lines as small as

possible. (We really minimize squared deviations, but I don't expect anyone's eyes to be good enough to do that.)

10.31 Galton's data

- a) The correlation is .459 and the regression equation is  $\hat{Y} = .646 \times \text{midparent} + 23.942$ . (Remember to weight cases by "freq".)
- b) I reran the regression requesting that SPSS save the Unstandardized prediction and residual.
- c)

**Descriptives**

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
child	1.00	392	67.1247	2.24664	.11347	66.9017	67.3478	61.70	72.20
	2.00	219	68.0196	2.24030	.15139	67.7213	68.3180	61.70	73.20
	3.00	183	68.7055	2.46458	.18219	68.3460	69.0649	63.20	73.70
	4.00	134	70.1776	2.26850	.19597	69.7900	70.5652	61.70	73.70
	Total	928	68.0885	2.51794	.08266	67.9263	68.2507	61.70	73.70
midparent	1.00	392	66.6633	1.06808	.05395	66.5572	66.7693	64.00	67.50
	2.00	219	68.5000	.00000	.00000	68.5000	68.5000	68.50	68.50
	3.00	183	69.5000	.00000	.00000	69.5000	69.5000	69.50	69.50
	4.00	134	71.1791	.78617	.06791	71.0448	71.3134	70.50	73.00
	Total	928	68.3082	1.78733	.05867	68.1930	68.4233	64.00	73.00

- d) The children in the lowest quartile slightly exceed their parents mean (67.12 vs. 66.66) and those in the highest quartile average slightly shorter than their parents (68.09 vs. 68.31).
- e) It is easiest if you force both axes to have the same range and specify that the regression line is  $\hat{Y} = 1 \times X + 0$ . (If you prefer, you can use an intercept of 0.22 to equate the means of the parents and children.)

