Chapter 15 - Multiple Regression

15.1 Predicting Quality of Life:

    a. All other variables held constant, a difference of +1 degree in Temperature is associated with a difference of −.01 in perceived Quality of Life. A difference of $1000 in median Income, again all other variables held constant, is associated with a +.05 difference in perceived Quality of Life. A similar interpretation applies to b_3 and b_4. Since values of 0.00 cannot reasonably occur for all predictors, the intercept has no meaningful interpretation.

    b.
$$\hat{Y} = 5.37 - .01(55) + .05(12) + .003(500) - .01(200) = 4.92$$
    c.
$$\hat{Y} = 5.37 - .01(55) + .05(12) + .003(100) - .01(200) = 3.72$$

15.3 The F values for the four regression coefficients would be as follows:

$$F_1 = \left[\frac{\beta_1}{s_{\beta_1}}\right]^2 = \left[\frac{-0.438}{0.397}\right]^2 = 1.22 \qquad F_2 = \left[\frac{\beta_2}{s_{\beta_2}}\right]^2 = \left[\frac{0.762}{0.252}\right]^2 = 9.14$$

$$F_3 = \left[\frac{\beta_3}{s_{\beta_3}}\right]^2 = \left[\frac{0.081}{0.052}\right]^2 = 2.43 \qquad F_4 = \left[\frac{\beta_4}{s_{\beta_4}}\right]^2 = \left[\frac{-0.132}{0.025}\right]^2 = 27.88$$

    I would thus delete Temperature, since it has the smallest F, and therefore the smallest semi-partial correlation with the dependent variable.

15.5   **a.** Envir has the largest semi-partial correlation with the criterion, because it has the largest value of t.

    b. The gain in prediction (from $r = .58$ to $R = .697$) which we obtain by using all the predictors is more than offset by the loss of power we sustain as p became large relative to N.

15.7 As the correlation between two variables decreases, the amount of variance in a third variable that they share decreases. Thus the higher will be the possible squared semi-partial correlation of each variable with the criterion. They each can account for more previously unexplained variation.

15.9 The tolerance column shows us that NumSup and Respon are fairly well correlated with the other predictors, whereas Yrs is nearly independent of them.

15.11 Using Y and $\hat{Y}$ from Exercise 15.10:

$$MS_{residual} = \frac{\sum(Y-\hat{Y})^2}{N-p-1}$$

$$= \frac{42.322}{15-4-1} = 4.232 \quad \text{(also calculated by BMDP in Exercise 15.4)}$$
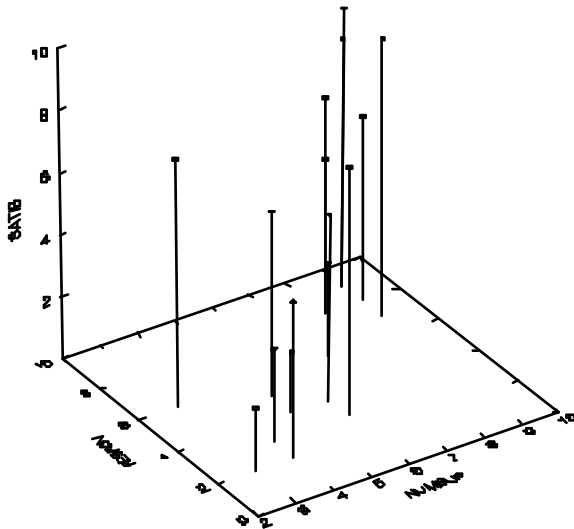
15.13 Adjusted R^2 for 15 cases in Exercise 15.12:

$$R^2_{0.1234} = .173$$

$$\text{est } R^{*2} = 1 - \frac{(1-R^2)(N-1)}{(N-p-1)} = 1 - \frac{(1-.173)(14)}{(15-4-1)} = -.158$$

Since a squared value cannot be negative, we will declare it undefined. This is all the more reasonable in light of the fact that we cannot reject $H_0:R^* = 0$.

15.15 Using the first three variables from Exercise 15.4:

a. Figure comparable to Figure 15.1:

b.

$\hat{Y} = 0.4067\text{Respon} + 0.1845\text{NumSup} + 2.3542$

The slope of the plane with respect to the Respon axis $(X_1) = .4067$
The slope of the plane with respect to the NumSup axis $(X_2) = .1845$
The plane intersects the Y axis at 2.3542

15.17 It has no meaning in that we have the data for the population of interest (the 10 districts).

15.19 It plays a major role through its correlation with the residual components of the other variables.

15.21 Within the context of a multiple-regression equation, we cannot look at one variable alone. The slope for one variable is only the slope for that variable when all other variables are held constant. The percentage of mothers not seeking care until the third trimester is correlated with a number of other variables.

15.23 Create set of data examining residuals.

15.25 Rerun of Exercise 15.24 adding PVTotal.

   b. The value of R^2 was virtually unaffected. However, the standard error of the regression coefficient for PVLoss increased from 0.105 to 0.178. Tolerance for PVLoss decreased from .981 to .345, whereas VIF increased from 1.019 to 2.900. **(c)** PVTotal should not be included in the model because it is redundant with the other variables.

15.27 Path diagram showing the relationships among the variables in the model.
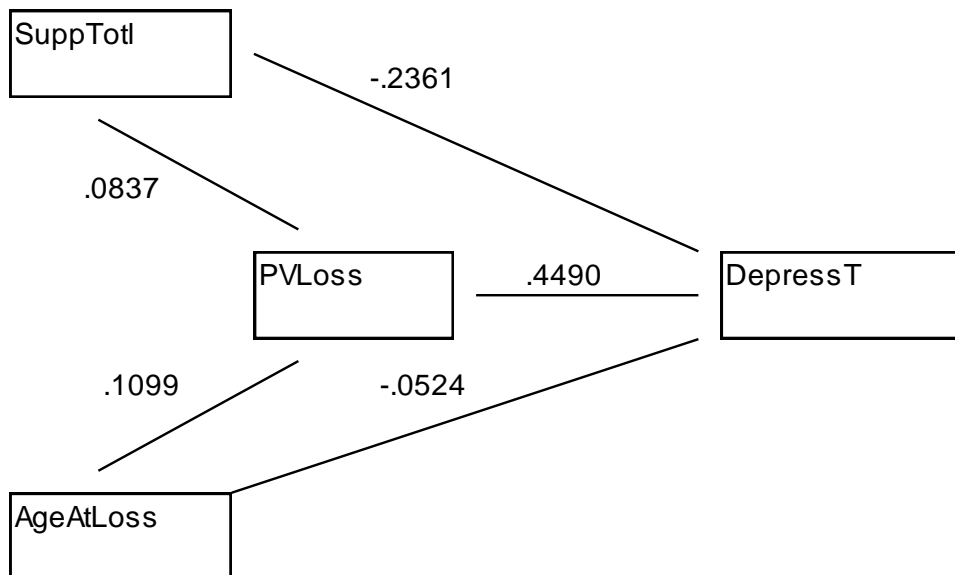
15.29 Regression diagnostics.

   Case # 104 has the largest value of Cook's D (.137) but not a very large Studentized residual ($t = -1.88$). When we delete this case the squared multiple correlation is increased slightly. More importantly, the standard error of regression and the standard error of one of the predictors (PVLoss) also decrease slightly. This case is not sufficiently extreme to have a major impact on the data.

15.31 Logistic regression using Harass.dat:

The dependent variable (Reporting) is the last variable in the data set.

I cannot provide all possible models, so I am including just the most complete. This is a less than optimal model, but it provides a good starting point. This result was given by SPSS.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	35.442	5	.000
	Block	35.442	5	.000
	Model	35.442	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	439.984	.098	.131

Classification Table[a]

			Predicted		
			REPORT		Percentage Correct
	Observed		No	Yes	
Step 1	REPORT	No	111	63	63.8
		Yes	77	92	54.4
	Overall Percentage				59.2

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	-.014	.013	1.126	1	.289	.986
	MARSTAT	-.072	.234	.095	1	.757	.930
	FEMIDEOL	.007	.015	.228	1	.633	1.007
	FREQBEH	-.046	.153	.093	1	.761	.955
	OFFENSIV	.488	.095	26.431	1	.000	1.629
	Constant	-1.732	1.430	1.467	1	.226	.177

a. Variable(s) entered on step 1: AGE, MARSTAT, FEMIDEOL, FREQBEH, OFFENSIV.

From this set of predictors we see that overall $\chi^2_{LR} = 35.44$, which is significant on 5 *df* with a *p* value of .0000 (to 4 decimal places). The only predictor that contributes significantly is the Offensiveness of the behavior, which has a Wald χ^2 of 26.43. The exponentiation of the regression coefficient yields 0.9547. This would suggest that as the offensiveness of the behavior increases, the likelihood of reporting *decreases*. That's an odd result. But remember that we have all variables in the model. If we simply predicting reporting by using Offensiveness, exp(B) = 1.65, which means that a 1 point increase in Offensiveness multiplies the odds of reporting by 1.65. Obviously we have some work to do to make sense of these data. I leave that to you.

15.33 It may well be that the frequency of the behavior is tied in with its offensiveness, which is related to the likelihood of reporting. In fact, the correlation between those two variables is .20, which is significant at *p* < .000. (I think my explanation would be more convincing if Frequency were a significant predictor when used on its own.)

15.35 BlamPer and BlamBeh are correlated at a moderate level ($r = .52$), and once we condition on BlamPer by including it in the equation, there is little left for BlamBeh to explain.

15.37 Make up an example.

15.39 This should cause them to pause. It is impossible to change one of the variables without changing the interaction in which that variable plays a role. In other words, I can't think of a sensible interpretation of "holding all other variables constant" in this situation.

15.41 Analysis of results from Feinberg and Willer (2011).

The following comes from using the program by Preacher and Leonardelli referred to in the chapter. I calculated the *t* values from the regression coefficients and their standard errors and then inserted those *t* values in the program. You can see that the mediated path is statistically significant regardless of which standard error you use for that path.

Input:	Test statistic:	p-value:
t_a 2.27	Sobel test: 2.0597124	0.03942604
t_b -4.90	Aroian test: 2.02528147	0.04283847
	Goodman test: 2.09596126	0.03608563
Reset all	Calculate	