Solutions to Worktext Exercises

Chapter 17

Visualizing Multiple Regression Analysis

Basic Learning Exercises

1. The model is logical, although some of the variables may be related to one another. Other arguments are possible if they are defended logically. It is important to understand how individual behavior differs from aggregate data (states) and to think about cause-and-effect.

Predictor	*Reasoning About Expected Sign of Estimated Coefficient*
Dropout (public high school dropout rates)	A negative sign might be expected since states with a large proportion of students who fail to complete high school are less likely to produce or retain a college-educated populace.
EdSpend% (percent of personal income spent on K-12 educ)	A positive sign might be expected since states that give students access to the educational tools needed for college success are more likely to produce and retain a college-educated populace.
Income (personal income per capita in current dollars)	A positive sign might be expected for since more affluent states would have more students with access to college preparatory education and have the ability to pay for college.
Urban (percent of population living in urban areas)	Sign could be argued to be positive (e.g., high-tech jobs in urban states attract those with more education) or negative (e.g., cities attract recent immigrants who may lack education).
Age (median age of population)	A negative sign might be anticipated because states with older populations would have more members of older generations who were less likely to graduate from college (and conversely).
FemLab (labor force participation rate among females)	A positive sign seems likely because states that attract more female workers would generally be those that offer more jobs that require college degrees.
Neast (1 if state is in the northeastern U.S., 0 otherwise)	Could be expected to be positive (e.g., the northeast has some of the top universities) or negative (e.g., the northeast has industrial jobs that require less education).
Seast (1 is state is in the southeastern U.S., 0 otherwise)	Could be expected to be positive (e.g., the south has a high growth rate with many high-tech jobs) or negative (e.g., the south historically has had more agricultural employment).
West (1 if state is in the western U.S., 0 otherwise)	Could be expected to be positive (e.g., the west is the home of high-tech silicon valley-type jobs) or negative (e.g., the west has attracted many recent immigrants who may have less education).

2. Four predictors (*Dropout, EdSpend%, Urban, and Seast*) have estimated coefficients that do not differ significantly from zero, so they can neither contradict nor confirm your expected coefficient sign. *Income* and *Neast* are the strongest predictors, followed by *Femlab* and *West*. *Age* is borderline because it is significant at $\alpha = 0.10$ but not at $\alpha = 0.05$ or $\alpha = 0.05$.

Predictor	*Expected Sign of Coefficient*	*Actual Sign of Coefficient*	*p-Value*	*Differs from Zero at*		
				α = 0.10?	*α = 0.05?*	*α = 0.01?*
Dropout	–	NA	0.657	No	No	No
EdSpend%	+	NA	0.809	No	No	No
Income	+	+	0.003	Yes	Yes	Yes
Urban	–/+	NA	0.424	No	No	No
Age	–	–	0.080	Yes	No	No

FemLab	–	+	0.026	Yes	Yes	No
Neast	–/+	+	0.004	Yes	Yes	Yes
Seast	–/+	NA	0.367	No	No	No
West	–/+	+	0.022	Yes	Yes	No

3. a) ColGrad% = 7.466 - 0.02324 Dropout - 0.1250 EdSpend% + .0006398 Income + 0.02781 Urban - 0.4403 Age + 0.2223 FemLab + 3.192 Neast + 1.041 Seast + 2.170 West. b) No, a predictor with a small coefficient (such as Income) can be significant. c) This equation might suggest to a novice that some good predictors (e.g., Income) were unimportant, or vice versa (e.g., Seast).

4. R^2 0.7705 R^2_{adj} 0.7188 F statistic 14.92 p-value 0.000
The R^2 indicates that about 77 percent of the variation in college graduation rates is accounted for by these nine predictors. However, the R^2_{adj} is noticeably lower, which suggests that some of the predictors may be redundant. The F statistic is very high, and its p-value indicates that the overall regression is significant (not due to chance) so at least some of the predictors must differ significantly from zero.

Intermediate Learning Exercises

5. At first glance, the residual histogram may seem somewhat symmetric, although there is one residual in the left tail that could be an outlier. A couple of other histogram bars differ somewhat from the superimposed bell-shaped curve.

6. a) The bars to the left of zero differ somewhat from the normal, but the bars on the right of zero are very close to the normal. The impressions are similar, except that the modal class is left of 0 in the standardized histogram. b) 48 of 50 or 96% lie within ± 2 (compared with 95% for a normal curve). c) Yes, there is one outlier in the left tail.

7. Nevada at –3.555 is an outlier (its college graduation rate is much lower than predicted by the model). Colorado at 2.198 is an unusual observation (its college graduation rate is much higher than predicted by the model).

8. Except for the outlier in the left tail (suggesting non-normality) the graph is fairly linear (suggesting normality). There are mixed indications about normality.

9. It shows the fitted value of the dependent variable (Colgrad%) plotted against the observed value of the dependent variable (Colgrad%). It shows a fairly strong correlation despite one outlier (Nevada). The multiple correlation coefficient is the same as the correlation between actual and estimated Y in a fitted regression model, and both are equal to the square root of the R^2 statistic.

10. *Income*, *Femlab*, and *Neast* have low p-values in both exhibits. *West* is significant in the regression but not in the correlation matrix, and conversely for *Urban* and *Seast*. No, a multiple regression is different than a bivariate correlation. A predictor that is significantly correlated with Y may be insignificant in a regression, and conversely.

11. 6 are significant at $\alpha = 0.05$ and 9 at $\alpha = 0.01$. This data set has many highly-correlated variables. b) Three correlations exceed 0.5000: *Income* and *Urban* (0.5534), *Income* and *Neast* (.5950), and *Dropout* and *Seast* (0.5497). Yes, these correlations suggest the possibility of multicollinearity (relationships among several predictors).

12. a) The VIFs are between 1.8 and 4.0. No VIF exceeds 10, so no single predictor is extremely related to its fellows. b) The sum of the VIFs (21.2) exceeds 10, which could cause concern.

c) There is a diffuse pattern of relationships among predictors, but removing a single predictor is not indicated as a remedy.

13. a) VIF = 1.0 would mean that the predictor was completely unrelated to its fellow predictors. b) In this regression, no VIF exceeds 5 and only 5 of the 9 exceed 2. All are below 10 (the dashed red line). c) The VIF scale is shown in log form because VIFs can become very large. d) Both the VIF plot and the correlation matrix offer clues about the presence of multicollinearity.

Advanced Learning Exercises

14. The magnitude of *Income* is quite different than the other predictors. *Income* might be rescaled (say by expressing it in thousands of dollars). Otherwise, no problems exist.

15. *Edspend%* in Montana is unusually high (7.0 compared with a mean of 4.406). *Age* in Utah is unusually low (26.2 compared with a mean of 32.82). *Femlab* in West Virginia is unusually low (44.2 compared with a mean of 58.75).

16. The interpretation should confirm your conclusions. It provides a good checklist for a novice because it checks for problems (e.g., variance inflation and data-conditioning).

17. The residuals appear to have little or no pattern, although some observers may feel there is a slight “fan-out” pattern. This suggests the absence of any violation.

18. There is little evidence of heteroskedasticity except possibly for Urban Femlab, and West (slight fan-out pattern) and Seast (possible “funnel-in”). The binary predictors yield two-valued residual plots. Since there are only two groups, it is actually easier to perform a visual check for equality of variance.

19. R^2, adjusted R^2, and multiple correlation coefficient are undefined when there is no intercept. No, it does not make sense, since there is no meaning to having these independent variables set to zero. For example, *Income* = 0 would require that there be no personal income in the state.

20. As the weaker predictors are eliminated form the model, R^2 declines only slightly up through Model 5, but then begins to fall more sharply. R2adj increases up to model 5, and then declines. The F statistic increases steadily. It would seem that the initial model contained several weak predictors whose loss did not matter much.

Model	Predictor Deleted	R^2	R^2_{adj}	F statistic
1	None	0.7705	0.7188	14.92
2	Edspend%	0.7701	0.7253	17.17
3	Dropout	0.7690	0.7305	19.97
4	Seast	0.7649	0.7322	23.32
5	Urban	0.7614	0.7343	28.08
6	Age	0.7364	0.7130	31.43
7	Neast	0.6937	0.6737	34.72
8	West	0.6436	0.6284	42.44
9	West	0.5634	0.5543	61.94

21. The VIFs are huge, and the model does not make very much logical sense *a priori*. There is positive autocorrelation (a pattern of + + + + – – – – in the residuals over time).