the data that gave it birth than to fresh data. Thus, p-values, confidence intervals, and prediction intervals should be used cautiously.

If the data set is very large, the analyst may benefit from dividing it at random into separate model construction and validation sets. A variable selection technique can be used on the model construction set to determine a set of explanatory variables. The selected model can then be refit on the validation set, without any further exploration into suitable explanatory variables, and inferential questions can be investigated on this fit, ignoring the construction data. When the purpose of the regression analysis is prediction, it is recommended that the validation data set be about 25% of the entire set. Although the saving of the 25% of the data for validation seems reasonable for other purposes, the actual benefits are not very well understood.

12.8 SUMMARY

Model selection requires an overall strategy for analyzing the data with regression tools (see Display 9.9). After giving initial thought to a game plan for investigating the questions of interest, the preliminary analysis consists of a combination of exploration, model fitting, and model checking. Once some useful models have been identified, the answers to the questions of interest can be addressed through inferences about their parameters.

Tools for initial exploration include graphical methods (such as a matrix of scatterplots, coded scatterplots, jittered scatterplots, and interactive labeling) and correlation coefficients between various variables. Certain tricks for modeling are used extensively in the case studies—indicator variables, quadratic terms, and interaction terms. For model checking and model building, a number of other tools are suggested: residual plots, partial residual plots, informal tests of coefficients, case influence statistics, the Cp plot, the BIC, and sequential variable selection techniques. Finally, some inferential tools are presented: t-tests and confidence intervals for individual coefficients and linear combinations of coefficients, extra sums-of-squares F-tests, prediction intervals, and calibration intervals.

SAT Study

Although this example is used to demonstrate variable-selection techniques, the actual analysis is guided by the objectives, and variable selection played only a minor role. These data have been used to rank the states on their success in secondary education, but in this regard selection bias poses a serious problem. In some states, for example, a high SAT average reflects the fact that only a small proportion of students—the very best ones—took the test, rendering the self-selected sample far from representative of high school students in the state overall. One goal of the regression analysis is to establish a ranking that accounts for this selection bias. Although overcoming the limitations of a self-selected sample is impossible, it is possible to rank the states after subtracting out the effects of the different proportions of students taking the test and their different median class rankings. Accomplishing this involves fitting the regression of SAT scores on these two explanatory variables, and ranking the states according to the sizes of their residuals.

A further question is exploratory in nature: Are any other variables associated with SAT score? For example, is the amount of money spent on secondary education related to it? Variable selection techniques may be useful for sorting through various models to identify promising predictors. Whatever models are suggested, though, only the ones that include percentage of takers and median class rank should be chosen, since a question does not make sense unless it addresses the effect on SAT after these two variables associated with the selection bias are accounted for. After performing a transformation of the percentage takers and setting Alaska aside, the Cp plot selects the model *tyer* with log takers, years (studying natural science, social science, and humanities), expenditure, and median rank as explanatory variables.

Sex Discrimination Study

Model selection in this example is motivated by the need to account fully for the nondiscriminatory explanatory variables before the sex indicator variable is introduced. Initial scatterplots indicate the important explanatory variables, which permit fitting a tentative model on which residual analysis can be conducted. After transformation, the analysis consists of using a variable selection technique to identify a good set of variables for explaining beginning salary. An inferential model is formed by this set plus the sex indicator variable, making precise the objective of seeing whether sex constitutes an important explanatory factor *after* everything else is accounted for. It may be disconcerting to realize that the different variable selection routines indicate different subset models; but it is reassuring that the estimated sex effect is about the same in all cases.

12.9 EXERCISES

Conceptual Exercises

1. **State SATs.** True or false? If the coefficient of the public school percentage is not significant (p-value > 0.05) in one model, it cannot be significant in any model found by adding new variables.

2. **State SATs.** True or false? If the coefficient of income is significant (p-value < 0.05) in one model, it will also be significant in any model that is found by adding new variables.

3. **State SATs.** Why are partial residual plots useful for this particular data problem?

4. **Sex Discrimination.** Suppose that another explanatory variable is length of hair at time of hire. (a) Will the estimated sex difference in beginning salaries be greater, less, or unchanged when length of hair is included in the set of explanatory variables for adjustment? (b) Why should this variable not be used?

5. Suppose that the variance of the estimated slope in the simple regression of Y on X_1 is 10. Suppose that X_2 is added to the model, and that X_2 is uncorrelated with X_1. Will the variance of the coefficient of X_1 still be 10?

6. In a study of mortality rates in major cities, researchers collected four weather-related variables, eight socioeconomic variables, and three air-pollution variables. Their primary question concerned

the effects, if any, of air pollution on mortality, after accounting for weather and socioeconomic differences among the cities. (a) What strategy for approaching this problem should be adapted? (b) What potential difficulties are involved in applying a model-selection strategy similar to the one used in the sex-discrimination study?

7. In the Cp plots shown in Display 12.9 and Display 12.11, the model with all available explanatory variables falls on the line (with intercept 0 and slope 1), meaning that the value of Cp for this model is exactly p. Will this always be the case? Why?

8. What is the usual interpretation of the probability of an event E? How does a Bayesian interpretation differ?

9. How does the posterior function, discussed in Section 12.5, account for model-selection uncertainty?

Computational Exercises

10. A, B, and C are three explanatory variables in a multiple linear regression with $n = 28$ cases. Display 12.15 shows the residual sums of squares and degrees of freedom for all models.

DISPLAY 12.15 Data for Exercise 10

Model variables	Residual sum of squares	Degrees of freedom
None	8,100	27
A	6,240	26
B	5,980	26
C	6,760	26
AB	5,500	25
AC	5,250	25
BC	5,750	25
ABC	5,160	24

(a) Calculate the estimate of σ^2 for each model. (b) Calculate the adjusted R^2 for each model. (c) Calculate the Cp statistic for each model. (d) Calculate the BIC for each model. (e) Which model has (i) the smallest estimate of σ^2? (ii) the largest adjusted R^2? (iii) the smallest Cp statistic? (iv) the smallest BIC?

11. Using the residual sums of squares from Exercise 10, find the model indicated by forward selection. (Start with the model "None," and identify the single-variable model that has the smallest residual sum of squares. Then perform an extra-sum-of-squares F-test to see whether that variable is significant. If it is, find the two-variable model that includes the first term and has the smallest residual sum of squares. Then perform an extra-sum-of-squares F-test to see whether the additional variable is significant. Continue until no F-statistics greater than 4 remain for inclusion of another variable.)

12. Again referring to Exercise 10, calculate $\exp\{-\text{BIC} + \text{BIC}_{\min}\}$ for each model. Add these and divide each by the sum. What is the resulting posterior distribution on the models?

13. Use the computer to simulate 100 data points from a normal distribution with mean 0 and variance 1. Store the results in a column called Y. Repeat this process 10 more times, storing results in $X_1, X_2, \ldots, X_{10}$. Notice that the Y should be totally unrelated to the explanatory variables. (a) Fit the regression of Y on all 10 explanatory variables. What is R^2? (b) What model is suggested by forward selection? (c) Which model has the smallest Cp statistic? (d) Which model has the smallest BIC? (e) What danger (if any) is there in using a variable selection technique when the number of explanatory variables is a substantial proportion of the sample size?

14. Blood–Brain Barrier. Using the data in Display 11.3 (file case1102), perform the following variable-selection techniques to find a subset of the covariates—days after inoculation, tumor weight, weight loss, initial weight, and sex—for explaining log of the ratio of brain tumor antibody count to liver antibody count. (a) Cp plot (b) forward selection (c) backward elimination (d) stepwise regression.

15. Blood–Brain Barrier. Repeat Exercise 14, but include sacrifice time (treated as a factor with three levels), treatment, and the interaction of sex and treatment with the other explanatory variables.

16. Sex Discrimination. The analysis in this chapter focused on beginning salaries. Another issue is whether annual salary increases tended to be higher for males than for females. If an annual raise of $100r\%$ is received in each of N successive years of employment, the salary in 1977 is: $\text{Sal77} = \text{SalBeg} \times (1 + r)^N$. Seniority measures the number of months of employment, so the number of years of employment is $N = \text{seniority}/12$; and

$$\log(1 + r) = (12/\text{seniority}) \times (\text{Sal77}/\text{SalBeg}) = z\,(\text{say})$$

for each individual. Calculate z from beginning salary, 1977 salary, and seniority; then calculate r as $\exp(z) - 1$. Now consider r as a response variable (the average annual raise). (a) Use a two-sample t-test to see whether the distribution of raises is different for males than for females. (Is a transformation necessary?) (b) What evidence is there of a sex effect after the effect of age on average raise has been accounted for? (c) What evidence is there of a sex effect after the effects of age and beginning salary have been accounted for?

17. Pollution and Mortality. Display 12.16 shows the complete set of variables for the problem introduced in Exercise 11.23. The 15 variables for each of 60 cities are (1) mean annual precipitation (in inches); (2) percent relative humidity (annual average at 1 P.M.); (3) mean January temperature (in degrees Fahrenheit); (4) mean July temperature (in degrees Fahrenheit); (5) percentage of the population aged 65 years or over; (6) population per household; (7) median number of school years completed by persons of age 25 years or more; (8) percentage of the housing that is sound with all facilities; (9) population density (in persons per square mile of urbanized area); (10) percentage of 1960 population that is nonwhite; (11) percentage of employment in white-collar occupations; (12) percentage of households with annual income under \$3,000 in 1960; (13) relative pollution potential of hydrocarbons (HC); (14) relative pollution potential of oxides of nitrogen (NO_X); and (15) relative pollution potential of sulphur dioxide (SO_2). (See Display 11.22 for the city names.) It is desired to determine whether the pollution variables (13, 14, and 15) are associated with mortality, after the other climate and socioeconomic variables are accounted for. (*Note:* These data have problems with influential observations and with lack of independence due to spatial correlation; these problems are ignored for purposes of this exercise.)

(a) With mortality as the response, use a Cp plot and the BIC to select a good-fitting regression model involving weather and socioeconomic variables as explanatory. To the model with the lowest Cp, add the three pollution variables (transformed to their logarithms) and obtain the p-value from the extra-sum-of-squares F-test due to their addition.

(b) Repeat part (a) but use a sequential variable selection technique (forward selection, backward elimination, or stepwise regression). How does the p-value compare?

18. Suppose that a problem involves four explanatory variables (like the weather variables in Exercise 17). How many variables would be included in the corresponding saturated second-order model (Section 12.7.3)?

DISPLAY 12.16 Pollution and mortality data for 60 cities; first 6 of 60 rows

		Weather				Demographic								Pollution potentials		
SMSA	Mortality (d/100.000)	Precipitation (in/yr)	Humidity (annual av%)	January temperature (F)	July temperature (F)	Over 65(%)	Population per household	Education (years)	Housing (% sound)	Density (popn/mi^2)	Nonwhite(%)	White collar(%)	Poor(%<$3,000)	Hydrocarbons	Nitrogen oxides	Sulfur dioxide
SJCA	790.733	13	71	49	68	7.0	3.36	12.2	90.7	2702	3.0	51.9	9.7	105	32	3
WIKS	823.764	28	54	32	81	7.0	3.27	12.1	81.0	3665	7.5	51.6	13.2	4	2	1
SDCA	839.709	10	61	55	70	7.3	3.11	12.1	88.9	3033	5.9	51.0	14.0	144	66	20
LAPA	844.053	43	54	32	74	10.1	3.28	9.5	79.2	3214	2.9	43.7	12.0	11	7	32
MIMN	857.622	25	58	12	73	9.2	3.28	12.1	83.1	2095	2.0	51.9	9.8	20	11	26
DATX	860.101	35	54	46	85	7.1	3.22	11.8	79.9	1441	14.8	51.2	16.1	1	1	1

19. In the expression (Section 12.7.3) for the number of subset models with p parameters, the index j of summation refers to the number of original variables in a particular model. If there are a total of five original explanatory variables, the total number of subset models with seven parameters that have exactly three original variables is the jth term in the sum, $C_{5,3} \times C_{6,3} = 10 \times 20 = 200$. The reasoning here is that there are 10 ways to select three of the original five variables; and that, with three variables, there are three quadratic plus three product terms, or a total of six second-order terms; so, to make up a model with seven parameters, you have the constant (1) and the original variables (3), so you need $7 - 1 - 3 = 3$ second-order terms from the six available. Problem: Let the original variables be a, b, c, d, and e. Select any three of them. Then write down all 20 of the seven-parameter models involving only those three original variables.

Data Problems

20. Galapagos Islands. The data in Display 12.17 come from a 1973 study. (Data from M. P. Johnson and P. H. Raven, "Species Number and Endemism: The Galapagos Archipelago Revisited," *Science* 179 (1973): 893–5.) The number of species on an island is known to be related to the island's area. Of interest is what other variables are also related to the number of species, after island area is accounted for, and whether the answer differs for native species and nonnative species. (*Note:* Elevations for five of the islands were missing and have been replaced by estimates for purposes of this exercise.)

21. Predicting Desert Wildflower Blooms. Southwestern U.S. desert wildflower enthusiasts know that a triggering rainfall between late September and early December and regular rains through March often lead to a good wildflower show in the spring. Display 12.18 is a partial listing of a data set that might help with the prediction. It includes monthly rainfalls from September to March and the subjectively rated quality of the following spring wildflower display for each of a number of years at each of four desert locations in the southwestern United States (Upland Sonoran Desert near Tucson, the lower Colorado River Valley section of the Sonoran Desert, the Baja California region of the Sonoran Desert, and the Mojave Desert). The quality of the display was judged subjectively with ordered rating categories of poor, fair, good, great, and spectacular. The column labeled

DISPLAY 12.17 Plant species and geography of the Galapagos Islands; first 5 rows of 30

	Observed species				Distance(km)		
Island	Total	Native	Area (km^2)	Elevation (m)	From nearest island	From Santa Cruz	Area of nearest island (km^2)
Baltra	58	23	25.09	332	0.6	0.6	1.84
Bartolome	31	21	1.24	109	0.6	26.3	572.33
Caldwell	3	3	0.21	114	2.8	58.7	0.78
Champion	25	9	0.10	46	1.9	47.4	0.18
Coamano	2	1	1.05	130	1.9	1.9	903.82

DISPLAY 12.18 Monthly rainfalls (inches) for seven months and rated quality of wildflower display in the spring, for multiple years in each of four desert regions; first 5 of 122 rows

Year	Region	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Total	Rating	Score
1970	baja	0.27	0.00	0.00	0.00	0.00	0.04	0.12	0.43	Poor	0
1971	baja	1.35	0.00	0.00	0.12	0.00	0.04	0.00	1.51	Poor	0
1972	baja	0.00	0.01	0.01	0.04	0.00	0.00	0.00	0.06	Poor	0
1974	baja	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.04	Poor	0
1975	baja	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.04	Poor	0

Score is a numerical representation of the rating, with 0 for poor, 1 for fair, 2 for good, 3 for great, and 4 for spectacular. Although this is a made-up number and the suitability of regression for such a discrete response is questionable, an informal regression analysis might nevertheless be helpful for casual prediction. Analyze the data to find an equation for predicting the quality score from the monthly rainfalls. What is the predicted quality score for a season with these rainfall amounts in the Upland region: Sep: 0.45, Oct: 0.02, Nov: 0.80, Dec: 0.76, Jan: 0.17, Feb: 1.22, and Mar: 0.37? Also find a 95% prediction interval. Although the conditions are far from the ideal normality assumptions for the justification of the prediction interval, the rough prediction interval might be a useful way to clarify the precision with which the quality score can actually be predicted. (Data from Arizona-Sonora Desert Museum, "Wildflower Flourishes and Flops—a 50-Year History," www.desertmuseum.org/programs/flw_wildflwrbloom.html (July 25, 2011).)

22. Bush–Gore Ballot Controversy. Review the Palm Beach County ballot controversy description in Exercise 8.25. To estimate how much of Pat Buchanan's vote count might have been intended for Al Gore in Palm Beach County, Florida, that exercise required the fitting of a model for predicting Buchanan's count from Bush's count from all other counties in Florida (excluding Palm Beach), followed by the comparison of Buchanan's actual count in Palm Beach to a prediction interval. One might suspect that the prediction interval can be narrowed and the validity of the procedure strengthened by incorporating other relevant predictor variables. Display 12.19 shows the first few rows of a data set containing the vote counts by county in Florida for Buchanan and for four other presidential candidates in 2000, along with the total vote counts in 2000, the presidential vote counts for three presidential candidates in 1996, the vote count for Buchanan in his only other campaign in Florida—the 1996 Republican primary, the registration in Buchanan's Reform party, and the total

DISPLAY 12.19 Buchanan's 2000 presidential vote count, and predictor variables, in 67 Florida counties (first five rows)

County	Buchanan 2000	Gore 2000	Bush 2000	Nader 2000	Browne 2000	Total 2000	Clinton 1996	Dole 1996	Perot 1996	Buchanan 1996	Reform reg. 2000	Total reg. 2000
Alachua	262	47,300	34,062	3,215	658	85,235	40,144	25,303	8,072	2,151	91	120,867
Baker	73	2,392	5,610	53	17	8,072	2,273	3,684	667	73	4	12,352
Bay	248	18,850	38,637	828	171	58,486	17,020	28,290	5,922	1,816	55	92,749
Bradford	65	3,072	5,413	84	28	8,597	3,356	4,038	819	155	3	13,547
Brevard	570	97,318	115,185	4,470	643	217,616	80,416	87,980	25,249	7,927	148	283,680
...												

political party registration in the county. Analyze the data and write a statistical summary predicting the number of Buchanan votes in Palm Beach County that were not intended for him. It would be appropriate to describe any unverifiable assumptions used in applying the prediction equation for this purpose. (*Suggestion:* Find a model for predicting Buchanan's 2000 vote from other variables, excluding Palm Beach County, which is listed last in the data set. Consider a transformation of all counts.)

23. Intelligence and Class as Predictors of Future Income (Males only). In their 1994 book, *The Bell Curve: Intelligence and Class Structure in American Life*, psychologist Richard Hernstein and political scientist Charles Murray argued that a person's intelligence is a better predictor of success in life than is education and family's socioeconomic staus. The book was controversial mostly for its conclusions about intelligence and race, but also for the strength of its conclusions drawn from regression analyses on observational data using imperfect measures of intelligence and socioeconomic status. Hernstein and Murray used data from the National Longitudinal Survey of Youth (NLSY79). Display 12.20 lists the variables in the ex1223 data file on a subset of 2,584 individuals from the NLSY79 survey who were re-interviewed in 2006, who had paying jobs in 2005, and who had complete values for the listed variables, as previously described in Exercises 2.22 and 3.30. *For males only*, see whether intelligence (as measured by the ASVAB intelligence test score, *AFQT*, and its Components, *Word, Parag, Math*, and *Arith*) is a better predictor of 2005 income than education and socioeconomic status (as measured by the variables related to respondent's class and family education in 1979). Suggestion: First use exploratory techniques to decide on transformations of the response variable *Income2005*. Then, Stage 1: Use a variable selection procedure to find a subset of variables from this set: *Imagazine, Inewspaper, Ilibrary, MotherEd, FatherEd, FamilyIncome78*, and *Educ* to explain the distribution of *Income2005*. Note the percentage of variation explained by this model. Now use a variable selection procedure to find a subset starting with this same initial set of variables and the ASVAB variables. By how much is R^2 increased when the intelligence measures are included? Use an extra-sum-of-squares F-test to find a p-value for statistical significance of all the ASVAB variables in the final model (i.e., after accounting for *Educ* and the class variables). Stage 2 (reverse the order): Use a variable selection procedure to find a subset of variables from the set of all ASVAB variables. Note the percentage of variation explained by this model. Now use a variable selection procedure to find a subset from this same initial set of variables plus *Educ* and the family class, education, and 1979 income variables. By how much is R^2 increased when *Educ* and the class variables are included? Find a p-value for statistical significance of the class variables after accounting for education and ASVAB variables. Use the Stage 1 and Stage 2 results to draw conclusions about whether the class variables or the ASVAB variables are better predictors of 2005 income. (As noted in Exercise 3.30, the coded incomes greater than \$150,000 in NLSY79 were replaced in this data file by computer-simulated values to better match a true distribution of incomes.)

DISPLAY 12.20 Variables measured on 2,584 Americans who were between 14 and 22 in 1979, who took the Armed Services Vocational Aptitude Battery (AFVAB) of tests in 1981, who were available for re-interview in 2006, who had paying jobs in 2005, and who had complete records of the variables listed below

Variables Related to Family Class, Education, and Income in 1979

Imagazine	1 if any household member magazines regularly when respondent was about 14
Inewspaper	1 if any household member newspapers regularly when respondent was about 14
Ilibrary	1 if any household member had a library card when respondent was about age 14
MotherEd	Mother's years of education
FatherEd	Father's years of education
FamilyIncome78	Family's total net income in 1978

Personal Demographic Variables

Race	1 = Hispanic, 2 = Black, 3 = Non-Hispanic, Non-Black
Gender	Female or male
Educ	Years of education completed by 2006

Variables Related to ASVAB Test Scores in 1981

Science	Score on General Science component
Arith	Score on Arithmetic Reasoning component
Word	Score on Word Knowledge component
Parag	Score on Paragraph Comprehension component
Numer	Score on Numerical Operations component
Coding	Score on Coding Speed component
Auto	Score on Automotive and Shop Information component
Math	Score on Mathematics Knowledge component
Mechanic	Score on Mechanical Comprehension component
Elec	Score on Electronics Information component
AFQT	Armed Forces Qualifying Test Score (a combination of *Word, Parag, Math*, and *Arith*)

Variables Related to Life Success in 2006

Income2005	Total income from wages and salary in 2005
Esteem1	"I am a person of worth" 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree
Esteem2	"I have a number of good qualities" 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree
Esteem3	"I am inclined to feel like a failure" 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree
Esteem4	"I do things as well as others" 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree
Esteem5	"I do not have much to be proud of" 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree
Esteem6	"I take a positive attitude towards myself and others" 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree
Esteem7	"I am satisfied with myself" 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree
Esteem8	"I wish I could have more respect for myself" 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree
Esteem9	"I feel useless at times" 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree
Esteem10	"I think I am no good at all" 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree

24. Intelligence and Class as Predictors of Future Income (Females only). Repeat Exercise 23 for the subset of females.

25. Gender Differences in Wages. Display 12.21 is a partial listing of a data set with weekly earnings for 9,835 Americans surveyed in the March 2011 Current Population Survey (CPS). What evidence is there from these data that males tend to receive higher earnings than females with the same values of the other variables? By how many dollars or by what percent does the male distribution exceed the female distribution? Note that there might be an interaction between *Sex* and *Marital Status*. (Data from U.S. Bureau of Labor Statistics and U.S. Bureau of the Census: Current Population Survey, March 2011 http://www.bls.census.gov/cps_ftp.html#cpsbasic; accessed July 25, 2011.)

DISPLAY 12.21 Region of the United States (Northeast, Midwest, South, or West) where individual worked, Metropolitan Status (Metropolitan, Not Metropolitan, or Not Identified), Age (years), Sex (Male or Female), Marital Status (Married or Not Married), EdCode (corresponding roughly to increasing education categories), Education (16 categories), Job Class (Private, Federal Government, State Government, Local Government, or Private), and Weekly Earnings (in U.S. dollars) for 9,835 individuals surveyed in the March 2011 Current Population Survey; first 5 of 9,835 rows

Region	MetropolitanStatus	Age	Sex	MaritalStatus	Edcode	Education	JobClass	WeeklyEarnings
Northeast	Not Metropolitan	20	Male	Not Married	39	HighSchoolDiploma	Private	467.50
West	Metropolitan	59	Male	Married	43	BachelorsDegree	Private	1,269.00
West	Metropolitan	62	Male	Married	34	SeventhOrEighthGrade	Private	1,222.00
West	Metropolitan	39	Male	Married	39	HighSchoolDiploma	Private	276.92
South	Not Metropolitan	60	Female	Married	36	TenthGrade	Private	426.30

Answers to Conceptual Exercises

1. False. In the model with P and E, neither variable is significant (Display 12.6). But both are significant in the model *PER*.

2. False. Both income and rank are significant in the model *IR*. When T is included, however, neither is significant. (See Display 12.6.)

3. There are two reasons. (i) Percentage of takers and median rank explain so much of the variation that any additional effect of the other variables is hidden in the ordinary scatterplots. (ii) The questions of interest call for the examination of some of the variables after getting percentage of takers and median rank out of the way. The partial residual plots allow visual investigation into these questions.

4. (a) The estimated sex difference will probably be less after length of hair is accounted for. If there are differences between male and female hair lengths, then that variable is picking up the sex differences and the sex indicator variable will be less meaningful when it is included. (b) The males and females should be compared after adjustment for nondiscriminatory determinants of salary.

5. If X_2 explains some variation in Y in addition to what is explained by X_1, σ will be smaller, so the variance of the coefficient of X_1 will decrease.

6. (a) Use model selection tools to find a good-fitting model involving the weather and socioeconomic variables. Then include the pollution variables, using t- and F-statistics to judge importance. As an alternative, employ the Bayesian strategy of averaging the pollution effects over a wide range of models. (b) A key weather variable (like humidity) may not be directly related to mortality, but it may interact with air pollution variables to affect mortality. A model selection method that chooses one "best" model may leave the key variable out. (See Exercise 17.)

7. Yes. Notice what happens to the formula for Cp when the model under investigation is also the "full" model: the second term is zero, so $\mathrm{Cp} = p$.

8. The probability of E is taken to be the proportion of times when E occurs in a long run of trials. Bayesian statistics also interprets probability as a measure of belief. If M is a particular model among many, a Bayesian probability for M would be the proportion of one's total belief that is assignable to the belief that M is the correct model.

9. Estimates of the treatment effect can be made with all models involving confounding variables. The posterior function draws its conclusions based on all these estimates, with weights assigned according to the strength of the evidence supporting each model.