**Chapter 10 Exercises**

**Concepts**

1. What is multicollinearity? How can a researcher ensure that her data does not have multicollinearity problems? What issues can multicollinearity pose for regressions?
2. What is a misspecification error? How can it be identified? How does it influence the results of a regression?
3. Is it always better to include a variable if the inclusion of the variable increases the *r*2 value? Why or why not?
4. What are some of the different ways to handle missing values and outliers in a multiple regression analysis? Is there one solution that works best across all cases?

**Exercises**

1. The following data reflects the test scores of a pool of applicants for a program versus whether or not they were accepted into the program.
	1. Plot the data and explain which assumption of linear regression is violated given the plot?
	2. Now use logistic regression to determine how the likelihood of being accepted into the program varies with the applicant’s test score.

|  |  |
| --- | --- |
| **Test score** | **Acceptance? 0 = No; 1 = Yes** |
| 68 | 0 |
| 79 | 1 |
| 86 | 0 |
| 90 | 1 |
| 96 | 1 |
| 68 | 0 |
| 85 | 1 |
| 71 | 0 |
| 78 | 0 |
| 73 | 0 |
| 85 | 1 |
| 77 | 0 |
| 85 | 0 |
| 91 | 0 |
| 78 | 0 |
| 82 | 1 |
| 84 | 1 |
| 94 | 1 |
| 83 | 0 |
| 68 | 0 |

1. The following data were collected from a team of computer programmers collaborating on a project. The promotion column tracks whether the programmer was promoted to a management role in the next year, while the “bugs” column tracks how many code problems each individual programmer fixed.

|  |  |
| --- | --- |
| **# Bugs fixed** | **Promoted? 0 = No; 1 = Yes** |
| 273 | 0 |
| 120 | 0 |
| 156 | 0 |
| 122 | 1 |
| 132 | 1 |
| 140 | 0 |
| 152 | 0 |
| 152 | 0 |
| 143 | 0 |
| 93 | 1 |
| 150 | 0 |
| 98 | 1 |
| 163 | 0 |
| 215 | 0 |
| 203 | 0 |
| 137 | 0 |
| 114 | 1 |
| 140 | 0 |
| 88 | 1 |
| 141 | 0 |

* 1. Use logistic regression to determine how the odds of a programmer’s promotion varies as the number of bugs she has fixed increases.
	2. Draw a graph of the results from part (a), showing how bug fixes affect promotion.
	3. What might explain this relationship between bugs fixed and promotion?
1. The expected number of primary care providers available within an hour drive of a consumer is a linear function of where the consumer lives:

 Y = 1.2 + 31.5 X1 + 22.8 X2 + 8.3 X3

where X1, X2, and X3 are dummy variables for urban, suburban, and rural areas, respectively (the remote areas category has been omitted).

* 1. What is the predicted number of primary care providers for each of the types of areas?
	2. How would you interpret the coefficient for the urban variable?
	3. What does the intercept value mean and indicate?
1. Given the following data, determine whether beta (the ‘slope’ of the logistic curve) is positive or negative. You do not need to actually find the coefficient, but justify your answer.

|  |  |
| --- | --- |
| **Y** | **X** |
| 1 | 169 |
| 0 | 127 |
| 1 | 200 |
| 0 | 150 |
| 0 | 148 |
| 1 | 192 |

1. Given the dataset in question 4, what can you say about the *R*2 if four more independent variables are added in the model? In the new model, what measure of explained variation would be more useful and why?
2. The following results were obtained from a regression on a dataset of size n = 50

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sum of squares** | **df** | **Mean square** | ***F*** |
| **Regression** | 537489 |  |  |  |
| **Residual** |  | 47 |  |  |
| **Total** | 541721 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Coefficient** ***(b)*** | **Standard error** ***(sb)*** | **VIF** |
| **Variable A** | 0.380 | 0.053 | 1.006 |
| **Variable B** | 32.2 | 0.417 | 1.006 |
| **Constant** | -298.1 | 9.893 |  |

* 1. Fill in the blanks in the ANOVA table.
	2. What is *r*2?
	3. Test the null hypothesis that *r*2 = 0 by comparing the F-statistic from the table with its critical value.
	4. What do you conclude from the variance inflation factors (VIFs)? What, if any, modification would you suggest in light of these values?
1. Researchers carry out a longitudinal study, tracking the academic performance of high school students and their eventual earnings as adults. Use the following data to answer the questions:

|  |  |  |  |
| --- | --- | --- | --- |
| **GPA** | **Standardized test score** | **Number of extra curriculars** | **Income at 30** **(in thousands of dollars)** |
| 2.4 | 1278 | 1 | 34 |
| 2.6 | 1313 | 2 | 42 |
| 2.5 | 1590 | - | 150 |
| 3.8 | 1536 | 1 | 80 |
| 3.4 | 1431 | 3 | 32 |
| 3.8 | 1405 | 2 | 48 |
| 2.8 | 1404 | 4 | 35 |
| 2.2 | 1379 | - | 40 |
| 2.7 | 1316 | 2 | 20 |
| 2.8 | 1334 | 1 | 30 |

* 1. Exclude the records with missing values and run a regression on the data. Give the r2.
	2. Replace the missing values with the mean, and run the regression and give the r2.
1. Using the SPSS Housing dataset and SPSS or Excel, answer the following:
	1. Construct a regression equation using housing price as the dependent variable and the floor area, bedrooms, bathrooms, garage, date built, whether the house is detached, and fireplace as the independent variables. Investigate any multicollinearity and outliers. Comment on the results.
	2. Attempt to improve upon the regression equation constructed in part (a), and justify the variables you include and exclude.
2. Using the Milwaukee Housing dataset and SPSS or Excel, answer the following:
	1. Construct a regression equation using the sale price of the house as the dependent variable and the lot size, the finished square feet, bedrooms, bathrooms, age, presence of full basement, attic, fireplace, air conditioning and garage as the independent variables. Investigate multicollinearity and outliers, and comment on the results.
	2. Attempt to improve upon the regression equation constructed in part (a), and justify the variables you include and exclude.
3. Using the Singapore Census dataset and SPSS or Excel:
	1. Create and describe three separate plots of the relationship between unemployment rate (dependent) and Illiteracy Rate, PctUniversityDegree, and PctRenter.
	2. Regress each independent variable individually with the dependent variable and then all together. Explain and interpret the results from each of the models and the discrepancies in the subsequent R2 observed if any.

**Chapter 10 Solutions**

* 1. The data fail to meet the assumption of homoscedasticity.



* 1. y = .205 x - 17.305
	2. y = 22.792 - 0.181 x



* 1. The negative coefficient on the variable suggests that as programmers carry out more bug fixes, they are less associated with management roles; perhaps programmers who are not doing bug fixes are doing other, more managerial tasks with their time.
	2. Remote Areas = 1.2; Urban = 32.7; Suburban = 24.0; Rural = 9.5
	3. Living in an urban area will on average increase the number of providers by 31.5.
	4. An area that is not urban, suburban, or rural will have on average 1.2 providers. In this case we might interpret this as the value for living in remote areas.
1. The slope will be positive, as the higher values are associated with 1 while the lower values are associated with 0.
2. Adding four more independent variables will increase the R2. It may be more useful to assess the model using the AIC, as this measure accounts for the trade-off between additional variables and model fit.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sum of squares** | **df** | **Mean square** | ***F*** |
| **Regression** | 537489 | 2 | 268744.5 | 2986.1 |
| **Residual** | 4232 | 47 | 90.0 |  |
| **Total** | 541721 | 49 |  |  |

* 1. r2 = 0.992
	2. The *F*-value is extreme, so that we may reject the null hypothesis and conclude that r2 is not zero.
	3. The VIFs suggest that there are no significant multicollinearity problems in the data; no adjustments are necessary.
	4. The regression is given by y = -5.220 xGPA + 0.209 xtest - 7.627 xextracurriculars - 216.537. r2 = 0.605
	5. The regression is given by y = -24.194 xGPA + 0.386 xtest - 9.637 xextracurriculars - 398.901. r2 = 0.880
	6. There are over 40 leverage values greater than the rule-of-thumb 2p/n = 14 / 499 = 0.028, with the most extreme value (0.160) being associated with a house that has 4 bathrooms yet costs 88950. However, none of the VIF values exceed 5, so that the regression suggests no multicollinearity problems. Five coefficients show up as being significant at the 0.05 level, namely garage, date built, floor area, detached, and fireplace. The bedrooms and bathrooms are not significant. The overall adjusted r2 value is 0.692.
	7. Dropping bedrooms and bathrooms as variables gives a model with an adjusted r2 of 0.691. Thus, with very little difference in r2, two of the variables can be eliminated for a more parsimonious model.
	8. There are 170 leverage values greater than the rule-of-thumb 2p/n = 20 / 1449 = 0.0138. The VIF values are low, none exceeding 5, so that there are few multicollinearity problems suggested here. Of the variables, all are shown to be significant at the 0.05 level except for the age of the lot and the presence or absence of a garage. The overall adjusted r2 value is 0.567.
	9. Dropping age and garage as variables, the adjusted r2 value is 0.566, and the model is simplified.



The illiteracy rate appears to have a positive relationship with unemployment. However, this relationship may not be linear.



The PctUniveDeg rate appears to have a negative relationship with unemployment.



PctRenter appears to have a negative relationship with unemployment. However, it is not clear how strong this relationship may be.

* 1. y = 0.069 + 0.513(IlliteracyRate) R2 = 0.596

y = 0.116 – 0.092(PctUnivDeg) R2 = 0.770

y = 0.099 – 0.63(PctRenter) R2 = 0.403

y = 0.1 + 0.291(IlliteracyRate) – 0.054(PctUnivDeg) -0.038(PctRenter) R2 = 0.802

The R2 for these regression models is reasonably high. Examining the collinearity diagnostics and correlations among the independent variables suggests there may be some concern about multicollinearity. In particular there is likely a relationship between PctUnivDeg and IlliteracyRate. We also observe a potential non-linear relationship between IlliteracyRate and UnempRate.

**Additional Chapter 10 Exercises**

**Concepts**

1. What challenges might a researcher face if he adopted a quadrat size or defined a region as a particular shape based on the default values provided by a software package? What should he consider in selecting either of these?
2. If a researcher uses nearest neighbors to detect a series of clusters in the data, are those clusters necessarily meaningful? Why or why not?
3. Can a researcher safely assume normally distributed variance among counties (which are defined politically) or census tracts (which are designed to have roughly equivalent populations)? Why or why not, in either case?
4. In specifying a spatial regression model, the measure of strength of the correlation between residuals ρ can sometimes equal zero. What does it imply about the correlation when this occurs?
5. Why might regression coefficients vary over space? Give an example of a process that might demonstrate this behaviour
6. Tobler’s first law of geography states that “Everything is related to everything else, but near things are more related than distant things”. How does geographically weighted regression capture this dynamic?

**Exercises**

1. The nests of a particular species of endangered bird have been found at the following locations:



Find the variance and the mean of the number of bird nests per cell, and use the variance-mean ratio to test the hypothesis that the nests are distributed randomly over the area (against the two-tailed hypothesis that they are not).

1. Incidents of a particular disease have been found at the following locations:



* 1. Find the variance and the mean of the number of disease incidences per cell, and use the variance-mean ratio to test the hypothesis that the incidences are distributed randomly over the area (against the two-tailed hypothesis that they are not).
	2. Aggregate the cells into square groups of 4, and repeat the analysis of part (a) at this new resolution (ignoring the obvious problems with the small sample size).
1. A regression of a health index on GDP and Gini coefficient (a measure of income disparity among the residents of a country) leaves the following residuals across a series of nations:



1. Use Moran’s *I* to determine whether there is a spatial pattern to the residuals, assuming binary connectivity for the weights.
2. For the following set of restaurants, perform an ordinary least squares regression of the average price of a main dish on the restaurant rating and age. Do the residuals show signs of autocorrelation?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X** | **Y** | **Rating** | **Age** | **Average main dish cost** |
| 11 | 13 | 3 | 1 | 24.17 |
| 6 | 9 | 2 | 13 | 14.54 |
| 6 | 4 | 1 | 2 | 12.65 |
| 9 | 8 | 3 | 4 | 24.16 |
| 10 | 9 | 4 | 1 | 24.72 |
| 8 | 7 | 2 | 1 | 16.53 |
| 12 | 3 | 3 | 5 | 26.11 |
| 10 | 6 | 2 | 1 | 7.15 |
| 13 | 7 | 5 | 1 | 25.72 |
| 7 | 15 | 2 | 1 | 16.54 |
| 5 | 11 | 4 | 1 | 19.96 |
| 7 | 10 | 2 | 58 | 46.78 |
| 13 | 14 | 2 | 2 | 9.25 |
| 5 | 6 | 1 | 5 | 15.25 |
| 7 | 12 | 1 | 3 | 10.79 |
| 13 | 6 | 1 | 1 | 7.13 |
| 11 | 6 | 3 | 1 | 16.89 |
| 6 | 6 | 4 | 6 | 22.34 |
| 6 | 8 | 3 | 42 | 39.1 |
| 9 | 13 | 4 | 1 | 28.28 |
| 10 | 10 | 2 | 2 | 18.5 |
| 9 | 2 | 4 | 1 | 20.42 |
| 10 | 12 | 3 | 1 | 20.7 |
| 8 | 13 | 2 | 1 | 12.62 |
| 9 | 9 | 4 | 2 | 26.19 |
| 7 | 9 | 3 | 2 | 19.81 |
| 7 | 7 | 1 | 31 | 25.98 |
| 9 | 11 | 3 | 1 | 22.74 |
| 9 | 15 | 3 | 1 | 26.47 |
| 11 | 8 | 4 | 1 | 25.35 |
| 12 | 6 | 2 | 1 | 16.67 |
| 12 | 7 | 3 | 1 | 21.51 |
| 14 | 11 | 1 | 1 | 19.67 |
| 11 | 11 | 2 | 15 | 23.65 |
| 8 | 8 | 3 | 2 | 15.8 |
| 16 | 9 | 1 | 1 | 23.27 |
| 5 | 12 | 4 | 1 | 26.22 |
| 17 | 8 | 4 | 1 | 21.23 |
| 11 | 14 | 5 | 33 | 45.11 |
| 13 | 9 | 2 | 1 | 18.46 |
| 4 | 8 | 2 | 1 | 13.75 |
| 11 | 9 | 4 | 2 | 25.35 |
| 8 | 12 | 5 | 2 | 26.88 |
| 13 | 12 | 1 | 3 | 10.19 |
| 8 | 11 | 2 | 1 | 21.79 |

1. Consider the following proposed relationship between a nation’s per capita GDP and its population and national spending on research. A researcher believes that the GDP of nations is dependent on the GDP of nearby nations, and wants to define a variable y\* as follows:

*y\* = y - p sigma{j=1 to n} wijyj*

Where the value of the weight wij = 1 if the two nations share a border and 0 otherwise. Find the value of p that minimizes the standard error of the estimate if y\* is regressed on population and research budget.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Country** | **GDP** | **Population (millions)** | **Research budget (billions)** | **Bordering nations** |
| 1 | 7.114 | 4.604 | 5.884 | 2,3,7,8 |
| 2 | 42.386 | 2.172 | 15.585 | 1,3,4,5 |
| 3 | 46.641 | 1.907 | 15.112 | 1,2,5,7 |
| 4 | 25.296 | 1.896 | 10.257 | 2,5,6 |
| 5 | 18.882 | 1.417 | 5.905 | 2,3,7 |
| 6 | 21.171 | 3.280 | 8.828 | 4 |
| 7 | 39.543 | 2.710 | 12.841 | 1,3,5,8 |
| 8 | 16.140 | 2.842 | 6.968 | 1,7 |

1. Use the following data and geographically weighted regression to produce a table of the regression coefficient for the density of vegetation based on rainfall at each point. Use a beta value of 0.05.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Point** | **X** | **Y** | **Rainfall (inches per year)** | **Vegetation** |
| 1 | 11.2 | 8.9 | 1568 | 2746 |
| 2 | 4.4 | 7.2 | 1827 | 2793 |
| 3 | 8.9 | 15.7 | 825 | 2728 |
| 4 | 10.1 | 5.1 | 1421 | 1929 |
| 5 | 9 | 7.6 | 615 | 1339 |
| 6 | 7.6 | 13.6 | 1073 | 2440 |
| 7 | 16.7 | 11.1 | 849 | 2232 |
| 8 | 8.4 | 10.4 | 977 | 2178 |
| 9 | 14.5 | 7.7 | 938 | 1529 |
| 10 | 5.9 | 6.2 | 1148 | 1696 |
| 11 | 7.6 | 8.3 | 1698 | 2471 |
| 12 | 8.4 | 11.4 | 1092 | 2296 |
| 13 | 11.7 | 10.3 | 1169 | 2284 |
| 14 | 2.3 | 6.1 | 1039 | 1744 |
| 15 | 8.1 | 8.2 | 1282 | 1766 |
| 16 | 9.8 | 10.8 | 1723 | 2682 |
| 17 | 13.6 | 8.2 | 1427 | 2319 |
| 18 | 7 | 7.5 | 977 | 1943 |
| 19 | 5.2 | 7.5 | 1360 | 2143 |
| 20 | 10.1 | 12.2 | 1563 | 3006 |
| 21 | 11.4 | 13.2 | 1651 | 2623 |
| 22 | 9.2 | 8.5 | 1442 | 2353 |
| 23 | 10.8 | 13.3 | 2035 | 3499 |
| 24 | 11.1 | 11.2 | 2242 | 3407 |
| 25 | 14.5 | 4.9 | 1279 | 1977 |
| 26 | 10.4 | 11.9 | 2216 | 3307 |
| 27 | 9.4 | 8.5 | 1161 | 1599 |
| 28 | 12.5 | 9.7 | 1997 | 2653 |
| 29 | 9.8 | 9.9 | 1795 | 2758 |
| 30 | 5.3 | 13.1 | 2020 | 3346 |

1. A business owner is looking to purchase a new office, and is considering the sale of other officer across a city in order to determine what prices he can expect to pay. Using the following data, perform an ordinary least squares regression of building price on lot size, the number of stories in the building, and whether the building is air conditioned.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Latitude** | **Longitude** | **Price** | **Lot Size** | **Stories** | **Air conditioning** |
| 44 | 16 | 1960 | 0.956 | 1 | 0 |
| 54 | 13 | 19982 | 7.948 | 1 | 0 |
| 51 | 9 | 11306 | 5.835 | 1 | 0 |
| 41 | 12 | 5860 | 1.563 | 1 | 0 |
| 47 | 13 | 17303 | 7.093 | 1 | 1 |
| 54 | 15 | 24418 | 9.073 | 2 | 0 |
| 53 | 6 | 16723 | 8.426 | 2 | 0 |
| 46 | 14 | 13431 | 5.507 | 1.5 | 1 |
| 45 | 11 | 3910 | 0.879 | 1.5 | 0 |
| 53 | 5 | 12028 | 5.156 | 1 | 0 |
| 32 | 6 | 26618 | 9.697 | 2 | 1 |
| 34 | 7 | 20201 | 8.092 | 1 | 1 |
| 35 | 11 | 8788 | 2.202 | 1 | 1 |
| 37 | 11 | 17242 | 6.608 | 1.5 | 0 |
| 26 | 20 | 22261 | 5.754 | 1.5 | 1 |

Give the regression equation. Do the residuals show signs of autocorrelation? Produce a graph of the residuals relative to their position in space.

1. Using geographically weighted regression with a beta value of ???, plot the spatial variation in the coefficient on plankton densities at each of the sampled locations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample** | **X** | **Y** | **Fish caught (tons)** | **Plankton density** |
| 1 | 55 | 26 | 12 | 0.639 |
| 2 | 43 | 44 | 8 | 0.189 |
| 3 | 63 | 49 | 14 | 0.470 |
| 4 | 51 | 41 | 14 | 0.563 |
| 5 | 61 | 64 | 16 | 0.439 |
| 6 | 71 | 56 | 14 | 0.399 |
| 7 | 58 | 33 | 9 | 0.294 |
| 8 | 42 | 43 | 10 | 0.289 |
| 9 | 47 | 35 | 7 | 0.137 |
| 10 | 61 | 41 | 14 | 0.563 |
| 11 | 69 | 79 | 20 | 0.484 |
| 12 | 40 | 43 | 19 | 0.815 |
| 13 | 46 | 82 | 29 | 0.730 |
| 14 | 42 | 30 | 11 | 0.513 |
| 15 | 30 | 54 | 26 | 0.953 |
| 16 | 53 | 54 | 15 | 0.477 |
| 17 | 57 | 56 | 9 | 0.161 |
| 18 | 56 | 37 | 15 | 0.665 |
| 19 | 41 | 61 | 14 | 0.356 |
| 20 | 25 | 51 | 9 | 0.201 |
| 21 | 57 | 60 | 28 | 0.944 |
| 22 | 31 | 33 | 15 | 0.756 |
| 23 | 44 | 48 | 11 | 0.294 |
| 24 | 63 | 55 | 27 | 0.996 |
| 25 | 40 | 37 | 9 | 0.264 |
| 26 | 35 | 64 | 27 | 0.853 |
| 27 | 61 | 36 | 19 | 0.960 |
| 28 | 63 | 44 | 17 | 0.699 |
| 29 | 34 | 57 | 20 | 0.675 |
| 30 | 53 | 28 | 8 | 0.270 |

**Additional Chapter 11 Solutions**

1. Variance s2 = 0.447, xmean = 0.633. VMR = 0.706. Given z = -1.120, we fail to reject the null hypothesis, as the critical values are -2.045 and +2.045.
	1. Variance s2 = 0.313, xmean = 0.528. VMR = 0.594. Given z = -1.698, we fail to reject the null hypothesis, as the critical values are approximately -1.96 and +1.96.
	2. Variance s2 = 4.611, xmean = 2.111. VMR = 2.184. Given z = 2.368, we reject the null hypothesis, as the critical value +1.96 is less than our z.
	3. I = -0.410, with *Z*-score -0.929. Thus we fail to reject the null hypothesis, where the critical values for alpha = 0.05 are -1.96 and +1.96.
2. The r2 value is .771, and the standard error is 4.069. Each of the coefficients is statistically significant, and the residuals show some evidence of a pattern:



1. The table below gives the standard errors associated with four different values of p:

|  |  |
| --- | --- |
| **P** | **Standard error** |
| 0.5 | 23.98 |
| 0.1 | 6.32 |
| 0.05 | 6.28 |
| 0.01 | 7.04 |
| 0 | 7.32 |

So we select 0.05 as the value of p, and get the following set of coefficients:

y\* = 6.871 xpop + 2.495 xresearch - 12.280

With the standard errors being 2.6 for xpop and 0.602 for xresearch, and 10.9 for the constant. All variables are significant.

|  |  |
| --- | --- |
| **Point** | **Coefficient** |
| 1 | 1.345 |
| 2 | 1.485 |
| 3 | 1.665 |
| 4 | 1.373 |
| 5 | 1.324 |
| 6 | 1.614 |
| 7 | 1.494 |
| 8 | 1.428 |
| 9 | 1.454 |
| 10 | 1.450 |
| 11 | 1.364 |
| 12 | 1.482 |
| 13 | 1.419 |
| 14 | 1.514 |
| 15 | 1.348 |
| 16 | 1.447 |
| 17 | 1.411 |
| 18 | 1.356 |
| 19 | 1.446 |
| 20 | 1.481 |
| 21 | 1.544 |
| 22 | 1.319 |
| 23 | 1.523 |
| 24 | 1.474 |
| 25 | 1.458 |
| 26 | 1.524 |
| 27 | 1.339 |
| 28 | 1.487 |
| 29 | 1.400 |
| 30 | 1.692 |

1. The regression equation on price (y), given the lot size (x1), air conditioning (x2), and number of stories (x3) is given by

 y = 2085.749 x1 + 2930.829 x2 + 2076.151 x3 - 928.369

The associated adjusted r2 value is .877, which is quite high, and the standard error of the estimate is 2622.9. Looking at the plot of residuals, however, we see some clustering in coefficients, which leads us to suspect that there may be spatial autocorrelation.



1. The graph of coefficients is given below:

