Section 13.1: Using Several Variables to Predict a Response

13.1 Predicting weight

a) $\hat{y} = -121 + 3.50x_1 + 1.35x_2 = -121 + 3.50\ (66) + 1.35\ (18) = 134.3$

b) The residual is $y - \hat{y} = 115 - 134.3 = -19.3$. The actual total body weight is 19.3 pounds lower than predicted.

13.2 Does study help GPA?

a) $\hat{y} = 1.13 + 0.643x_1 + 0.0078x_2 = 1.13 + 0.643(3.5) + 0.0078(3) = 3.40$

b) For a fixed study time, the change in predicted college GPA is 0.64 (the slope) as high school GPA goes from 3.0 to 4.0.

13.3 Predicting visitor satisfaction

a) (i) $\hat{y} = 0.35 + 0.55x_1 + 0.0015x_2 = 0.35 + 0.55(4.0) + 0.0015(800) = 3.75$

(ii) $\hat{y} = 0.35 + 0.55x_1 + 0.0015x_2 = 0.35 + 0.55(2.0) + 0.0015(200) = 1.75$

b) $\hat{y} = 0.35 + 0.55x_1 + 0.0015x_2 = 0.35 + 0.55x_1 + 0.0015(500) = 0.35 + 0.55x_1 + 0.75 = 1.10 + 0.55x_1$

c) $\hat{y} = 0.35 + 0.55x_1 + 0.0015x_2 = 0.35 + 0.55x_1 + 0.0015(600) = 0.35 + 0.55x_1 + 0.9 = 1.25 + 0.55x_1$

13.4 Interpreting slopes on average monthly visitor satisfaction

a) Setting x_2 at a variety of values yields a collection of parallel lines relating $\hat{y}$ to x_1 because the model assumes that the slope for a particular explanatory variable such as x_1 is identical for all fixed values of the other explanatory variable, x_2. The slope of each parallel line is 0.55.

b) This does not imply that x_1 has a larger effect than does x_2 on y in this sample because the two variables do not use same units. We can only compare slopes if their variables use same units.

13.5 Does more education cause more crime?

a) (i) $\hat{y} = 59.1 - 0.583x_1 + 0.683x_2 = 59.1 - 0.583(70) + 0.682(0) = 18.3$

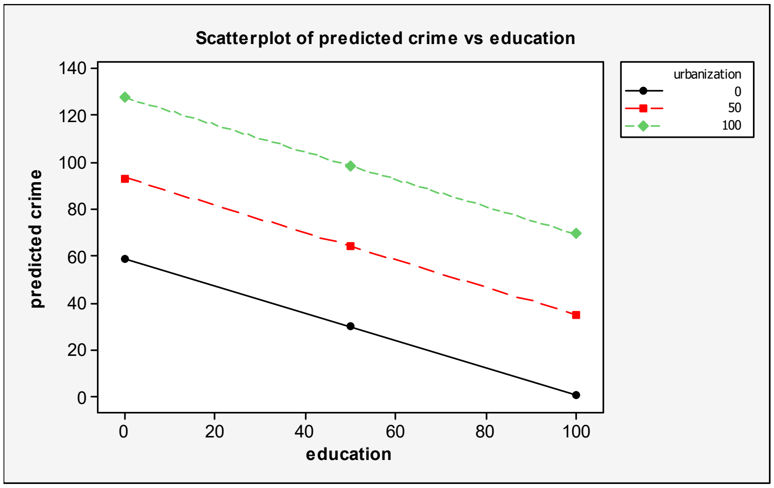
(ii) $\hat{y} = 59.1 - 0.583x_1 + 0.683x_2 = 59.12 - 0.583(80) + 0.683(0) = 12.5$

b) When we control for urbanization, crime rate changes by the slope multiplied by the change in education. When education goes up 10 (from 70 to 80), predicted crime rate changes by ten multiplied by the slope, 10(–0.5834) = –5.8.

c) (i) $\hat{y} = 59.1 - 0.583x_1 + 0.683(0) = 59.1 - 0.583x_1$

(ii) $\hat{y} = 59.1 - 0.583x_1 + 0.683(50) = 93.2 - 0.583x_1$

(iii) $\hat{y} = 59.1 - 0.583x_1 + 0.683(100) = 127.4 - 0.583x_1$

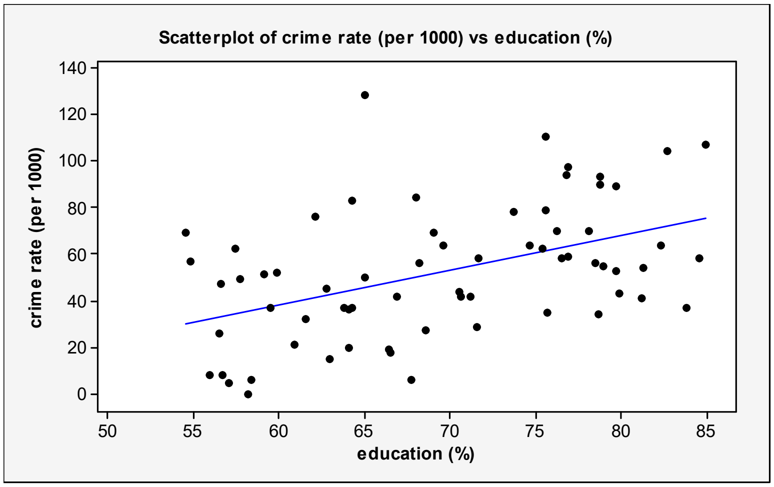


For each fixed level of urbanization, the predicted crime rate decreases by 5.8 for every 10 percentage-point increase in education

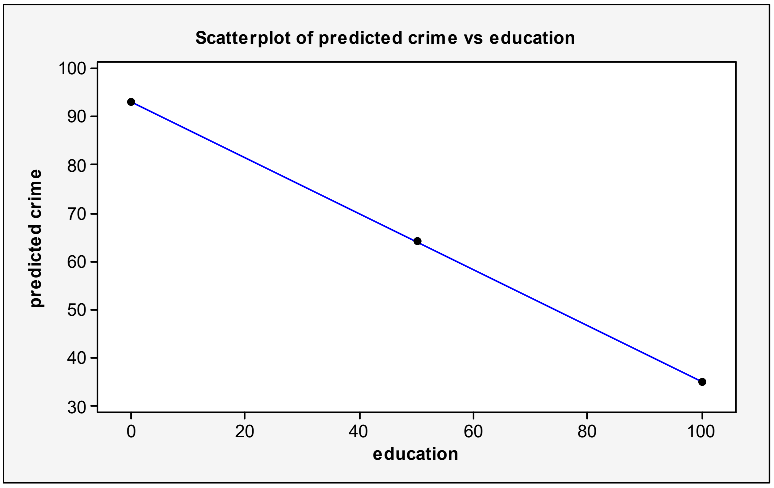
13.5 (continued)

d) The line passing through the points having urbanization = 50 has a negative slope. The line passing through all the data points has a positive slope. Simpson's paradox occurs because the association between crime rate and education is positive overall but is negative at each fixed value of urbanization. It happens because urbanization is positively associated with crime rate and with education. As urbanization increases, so do crime rate and education tend to increase, giving an overall positive association between crime rate and education.

(i)



(ii)



13.6 Crime rate and income

a) (i) $\hat{y} = 40.0 - 0.791x_1$

(ii) $\hat{y} = 104.2 - 0.791x_1$

As urbanization increases from 0 to 100%, the intercept increases from 40.0 to 104.2; the slope stays the same.

b) As income increases, predicted crime increases. This is opposite of the effect of income in the multiple regression equation.

c) (i) Ignoring urbanization, from the simple regression model with y = crime rate and x = median income, the predicted crime rate increases by 2.6 for every $1000 increase in income.

(ii) Controlling for urbanization by fitting the multiple regression model including income and urbanization, the predicted crime rate decreases by 0.8 for every $1000 increase in income.

13.7 The economics of golf

a) The regression formula for a PGA Tour golfer's earnings for 2008 is: $\hat{y} = 26,417,000 + 168,300\text{GIR} + 33,859\text{SS} - 19,784,000\text{AvePutt} - 44,725\text{Events}.$

b) The coefficient for each variable is its slope. The predicted total score will decrease by \$19,784,000 for each increase of one in the average number of putts after reaching the green, when controlling for the other variables in the model.

c) $\hat{y} = 26,417,000 + 168,300(60) + 33,859(50) - 19,784,000(1.5) - 44,725(20) = \$7,637,450$

13.8 Comparable number of bedrooms and house size effects

a) House selling price is predicted to increase by \$63 because the slope associated with a variable is the amount that the predicted value of y will increase when all other variables in the equation are held constant.

b) For a fixed house size of 2000 square feet, the predicted selling price of a two bedroom house is $\hat{y} = 60,102 + 63.0(2000) + 15,170(2) =$ \$216,442, a three bedroom house is

$\hat{y} = 60,102 + 63.0(2000) + 15,170(3) =$ \$231,612, and a four bedroom house is

$\hat{y} = 60,102 + 63.0(2000) + 15,170(4) = \$246,782.$

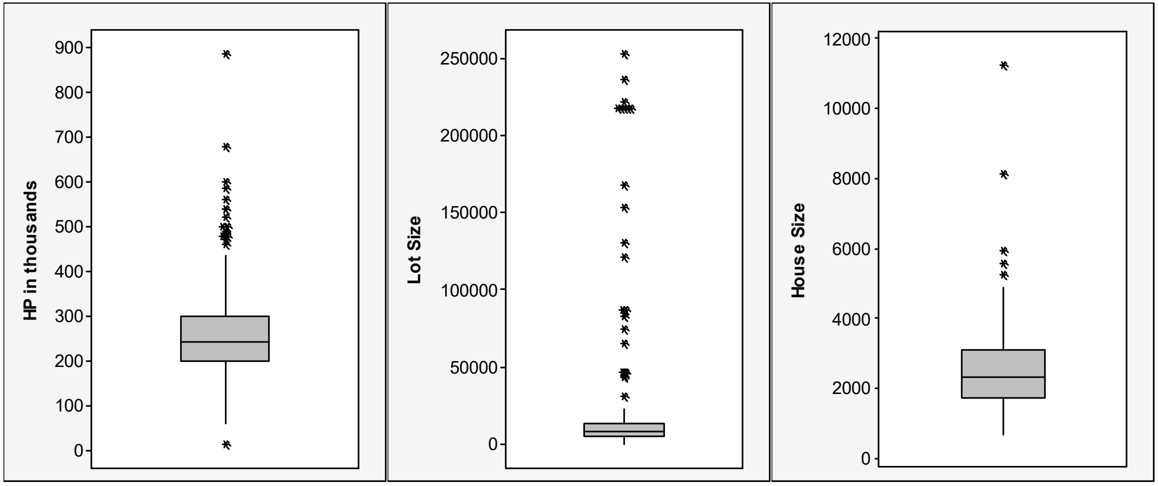
For a fixed house size, the predicted selling price will increase by \$15,170 for each additional bedroom.

13.9 Controlling has an effect

In multiple linear regression, the effect of x_1 is controlled for its relationship with the other predictor variable, x_2 . Hence, the slope of x_1 in multiple linear regression of y on x_1 and x_2 will not be the same as in simple linear regression when x_1 is the only predictor. Yes, the statement changes to "the slope of x_1 in multiple linear regression of y on x_1 and x_2, is the same as in simple linear regression when x_1 is the only predictor when x_1 and x_2 are uncorrelated," because we do not need to control x_2 if it is not associated with x_1 . Changes in x_2 will not have an impact on the effect of x_1 on y.

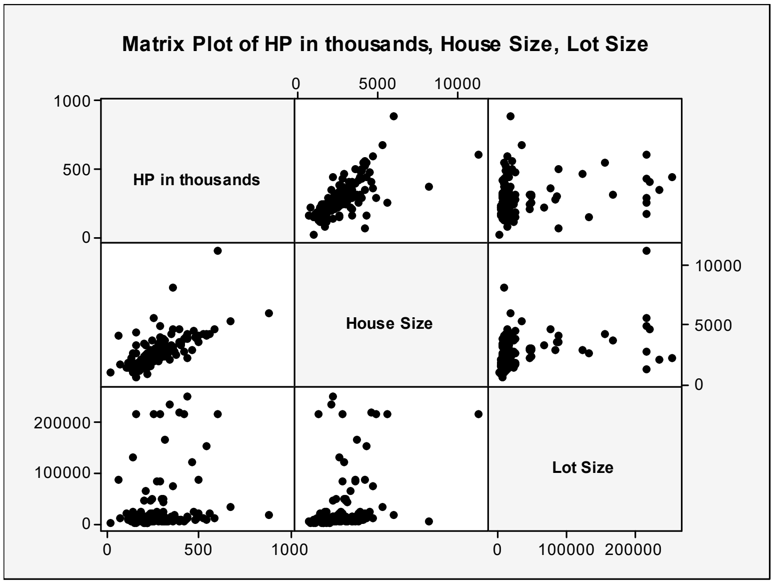
💻13.10 House selling prices

a) Box plots:



13.10 (continued)

Scatterplot matrix:



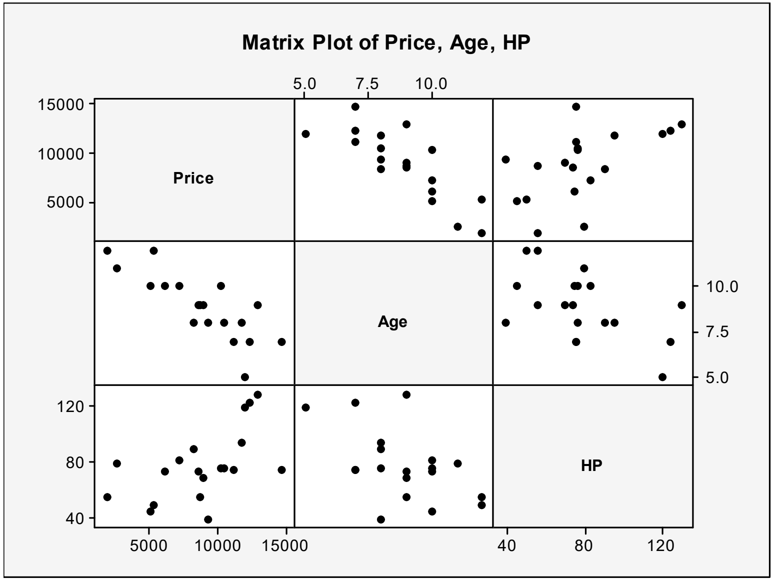
House price and house size are positively correlated. The relationship between house price and lot size is not as clear. There are a lot of large lot sizes that don't increase the house price. Many lot sizes are in the lower conventional size, but there are several large properties. It is possible that the large properties have older or farm houses in worse condition, thereby not yielding a higher price. Ignoring the large lot sizes, the variables show a positive correlation. House size and lot size have similar issues with most of the houses in the lower range.

b) From technology: price = 97,001+67.6house_size – 0.08lot_size

c) If house size remains constant, there is very little change with respect to an increase in lot size. The reason is that larger properties could be older homes in poorer condition with not as much value.

13.11 Used cars

a) The relationship between price and age is linear, negative, and strong. The relationship between price and HP is less clear; it may be linear but rather weak.



13.11 (continued)

b) From technology: Predicted Price = 19,348.7 – 1406.3Age + 25.5HP
   (i) Predicted Price = 19,348.7 – 1406.3(8) + 25.5(80) = \$10,100
   (ii) Predicted Price = 19,348.7 – 1406.3(10) + 25.5(80) = \$7,300

c) No, the predicted price difference depends on the age of the two cars because each predicted price depends on it. Only if the age of the two cars is the same does the effect of age cancel out, and the predicted price difference will be 25.5(80 – 60) = 510.

Section 13.2: Extending the Correlation and R^2 for Multiple Regression

13.12 Predicting average monthly visitor satisfaction

a) $R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = \frac{7.12}{10.09} \approx 0.71$

b) Using monthly food quality score (x_1) and the monthly number of visitors (x_2) together to predict average monthly customer satisfaction rating reduces the prediction error by 71%, relative to using $\bar{y}$ alone to predict average monthly customer satisfaction rating. The prediction is moderately better.

c) $R = \sqrt{R^2} = \sqrt{0.706} = 0.840$; there is a moderately strong association between the observed and predicted satisfaction ratings.

13.13 Predicting weight

a) Because height is more strongly correlated with weight than are age and percent body fat, it is, by itself, the best predictor of weight. Because correlations are not based on units, we can directly compare the strengths of relationships by comparing correlations.

b) One of the properties of R^2 is that it gets larger, or at worst stays the same, whenever an explanatory variable is added to the multiple regression model.

c) No, the reduction in error from using the regression equation to predict weight rather than the mean is only 1% more when adding age. It only increases from 0.66 to 0.67.

13.14 When does controlling have little effect?

Controlling for body fat and then age does not change the effect of height much because height is not strongly correlated with either body fat or age.

13.15 Price of used cars

a) $R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = \frac{222{,}102{,}253 - 69{,}534{,}753}{222{,}102{,}253} = 0.69$

b) Using both age and horsepower to predict used car price reduces the prediction error by 69%, relative to using the sample mean price $\bar{y}$.

c) 69% of the variability in used car prices can be explained by the varying age and horsepower of the cars.

13.16 Price, age, and horsepower

Because HP is correlated with price, and HP and age are correlated (older cars tend to have less HP), once age is in the model for predicting price, HP does not help in further improving the predictive power. Most of the effect of HP on price has already been captured by age.

13.17 Softball data

a) We know that the number of hits does not make much of a difference, over and above runs and errors, because its slope is so small. An increase of one hit only leads to a predicted difference 0.026 more.

b) A small increase in R^2 indicates that the predictive power doesn't increase much with the addition of this explanatory variable over and above the other explanatory variables.

13.18 Slopes, correlations, and units

a) The correlation between predicted house selling price and actual house selling price is 0.72.

13.18 (continued)

b) If selling price is measured in thousands of dollars, each *y*-value would be divided by 1000. For example, $145,000 would become 145 thousands of dollars. Each slope would also be divided by 1000 (e.g., a slope of 63.0 for house size on selling price in dollars corresponds to 0.063 in thousands of dollars for $\hat{y}$).

c) The multiple correlation would not change because it is not dependent on units.

13.19 Predicting college GPA

Technology reports that $R^2 = 25.8\%$; the multiple correlation is the square root of 0.258, which is 0.51. Using these variables together to predict college GPA reduces the prediction error by 26%, relative to using $\bar{y}$ alone to predict college GPA. There is a correlation of 0.51 between the observed college GPAs and the predicted college GPAs. Only 25.8% of the observed variability in students' college GPA can be explained by their high school GPA and study time. The remaining 74.2% of the variability in college GPA is due to other factors.

Section 13.3: Inferences Using Multiple Regression

13.20 Predicting college GPA

a) If the population slope coefficient equals 0, it means that in the population of all students, high school grade (HSG) does not predict college CPI for students having any given value for study average monthly attendance percentage (AMAP) and average daily study time (ADST). For example, for students who study 5 hours, HSG does not predict college CPI.

b) 1) Assumptions: We assume a random sample and that the model holds (each explanatory variable has a straight-line relation with μ_y controlling for the other predictors, with the same slope for all combinations of values of other predictors in model, and there is a normal distribution for *y* with the same standard deviation at each combination of values of the other predictors in the model).

2) Hypotheses: H_0: $\beta_1 = 0$; H_a: $\beta_1 \neq 0$

3) Test statistic: $t = (b_1 - 0)/se = 0.6615/0.1578 = 4.19$

4) *P*-value: The P-value is approximately 0.000.

5) Conclusion: The P-value of 0.000 gives evidence against the null hypothesis that $\beta_1 = 0$. If the null hypothesis were true, the probability would be almost 0 of getting a test statistic at least as extreme as the value observed. We have very strong evidence that high school grade (HSG) predicts college CPI, if we already know the average monthly attendance percentage (AMAP) and average daily study time (ADST). At common significance levels, such as 0.05, we reject H_0.

13.21 Study time helps CPI?

a) If the null hypothesis were true, the probability would be 0.121 of getting a test statistic at least as extreme as the value observed. It is plausible that the null hypothesis that $\beta_3 = 0$ is correct, and that study time does not predict college CPI, if we already know high school grade (HSG) and average monthly attendance percentage (AMAP).

b) The 95% confidence interval for β_3 is $b_3 \pm t_{0.025}(se) = 0.0075 \pm 1.984(0.151)$, or (–0.02, 0.04). Because 0 falls in the confidence interval, it is plausible that the slope is 0 and that average daily study time has no association with college CPI when high school grade (HSG) and average monthly attendance percentage (AMAP) are controlled.

c) No. It is likely that average daily study time (ADST) is highly correlated with high school grade (HSG) and/or average monthly attendance percentage (AMAP) and therefore it does not add much predictive power to the model to include average daily study time once HSG and AMAP are already in the model. This does not mean that study time has no association with college CPI.

13.22 Variability in college CPI

a) The residual standard deviation, 0.22, describes the typical size of the residuals and also estimates the standard deviation of *y* at fixed values of the predictors. For students with certain fixed values of high school grade (HSG), average monthly attendance percentage (AMAP) and average daily study time (ADST), college CPIs vary with a standard deviation of 0.22.

13.22 (continued)

b) Approximately 95% of college GPAs fall within about 2*s*, 0.64, of the true regression equation. When high school grade (HSG) = 3.80, average monthly attendance percentage (AMAP) = 0.90 and average daily study time (ADST) = 5.0 hours per day, college CPI is predicted to be 1.1362 + 0.6615 (3.80) + 0.2301 (.90) + 0.0075 (5) = 3.89. Thus, we would expect that approximately 95% of the college students fall between 3.89 – 0.44 = 3.45 and 3.89 + 0.44 = 4.33.

13.23 Does leg press help predict body strength?

a) 1) Assumptions: We assume a random sample and that the model holds (each explanatory variable has a straight-line relation with μ_y, with same slope for all combinations of values of other predictors in model, and there is a normal distribution for *y* with the same standard deviation at each combination of values of the other predictors in the model). Here, the 57 athletes were a convenience sample, not a random sample, so inferences are tentative.

2) Hypotheses: H$_0$: $\beta_2 = 0$; H$_a$: $\beta_2 \neq 0$

3) Test statistic: $t = (b_2 - 0)/se = 0.211/0.152 = 1.39$

4) *P*-value: The P-value is 0.17.

5) Conclusion: If the null hypothesis were true, the probability would be 0.17 of getting a test statistic at least as extreme as the value observed. It is plausible that the null hypothesis that $\beta_2 = 0$ is correct, and that the number of times an athlete can perform a 200-pound leg press does not predict upper body strength (maximum number of pounds she could bench press), if we already know the number of times she can do a 60-pound bench press.

b) The 95 confidence interval for β_2 is $b_2 \pm t_{.025}(se) = 0.2110 \pm 2.005(0.1519)$, or (–0.1, 0.5). Based on this interval, LP200 seems to have a weak impact; 0 is in the confidence interval, indicating that it is plausible that there is no association between LP200 and BP when controlling for BP60.

c) When LP200 is included in the model, the P-value of the slope associated with BP60 is 0.000, very strong evidence that the slope of BP60 is not 0.

13.24 Leg press uncorrelated with strength?

The first test analyzes the effect of LP200 at any given fixed value of BP60, whereas the second test describes the overall effect of LP200 ignoring other variables. These are different effects, so one can exist when the other does not. In this case, it is likely that LP200 and BP60 are strongly associated with one another, and the effect of LP200 is weaker once we control for BP60.

13.25 Interpret strength variability

a) The residual standard deviation estimates the standard deviation of the distribution of maxBP at given values for BP60 and LP200. This standard deviation is assumed to be the same for any combination of BP60 and LP200 values and is estimated as 7.9. The sample standard deviation of maxBP of 13.3 shows how much maxBP values vary overall, over the entire range of BP60 and LP200 values, not over just a particular pair of values.

b) Approximately 95% of BPs fall within about 2*s* = 15.8 of the true regression equation.

c) The prediction interval is an inference about where the population maxBP values fall at fixed levels of the two explanatory variables. The prediction interval indicates where a response outcome has a 95% chance of falling.

d) It would be unusual because 100 is not in the prediction interval.

13.26 Any predictive power?

a) H$_0$: $\beta_1 = \beta_2 = 0$; the null hypothesis states that neither of the two explanatory variables has an effect on the response variable *y*.

b) $F = 3.16$

c) The observed *F* statistic is 51.39 with a P-value of approximately 0.000. If the null hypothesis were true, the probability would be close to 0 of getting a test statistic at least as extreme as the value observed. This P-value gives extremely strong evidence against the null hypothesis that $\beta_1 = \beta_2 = 0$. At common significance levels, such as 0.05, we can reject H$_0$. At least one of the two explanatory variables has an effect on maxBP.

13.27 Predicting restaurant revenue

a) $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$; where μ_y is the population mean for monthly revenue, x_1 is a given level of advertising spend, x_2 is a given level of average price of its own menu items and x_3 is a given level of average price of its competitors' menu items.

b) H$_0$: $\beta_1 = 0$

c) H$_0$: $\beta_1 = \beta_2 = \beta_3 = 0$

13.28 Regression for human development

a) The 95% confidence interval for β_1 is $b_1 \pm t_{0.025}(se) = 0.1033 \pm 2.0260(0.03250)$, or (0.04, 0.17).

b) The confidence interval gives plausible values for the slope, the amount that mean human development will increase when literacy rate increases by one, when controlling for daily per capita income. For an increase of 10 units in literacy rate, we multiply each endpoint of the confidence interval by 10, giving (0.4, 1.7); this indicates that plausible increase in mean human development range from 0.4 to 1.7 when literacy rate increase by 10, controlling for daily per capita income.

13.29 Gain in human development

a) The test statistic, *F*, is 9.49, and the P-value is approximately 0.000.

b) H$_a$: At least one β parameter is not equal to 0.

c) The result in (a) indicates only that at least one of the variables is a statistically significant predictor of human development.

13.30 More predictors for selling price

a) H$_0$: $\beta_1 = \beta_2 = \beta_3 = 0$ means that house selling price is independent of size of home, number of bedrooms, and age.

b) The large *F* value and small P-value provide strong evidence that at least one of the three explanatory variables has an effect on selling price.

c) The results of the *t* tests tell us that both house size and number of bedrooms contribute to the prediction of selling price. At the 5% significance level, age does not contribute significantly when house size and number of bedrooms are included in the model.

13.31 House prices

a) H$_0$: $\beta_1 = \beta_2 = 0$; H$_a$: At least one β parameter is not equal to 0. The *F*-statistic is 108.6 and the P-value is approximately 0.000. The P-value of 0.000 gives very strong evidence against the null hypothesis that $\beta_1 = \beta_2 = 0$. It is not surprising to get such a small P-value for this test because house size and bedrooms are both statistically significant predictors of selling price.

b) The *t*-statistic is 2.85 with a P-value of 0.0025. If the null hypothesis were true, the probability would be 0.025 of getting a test statistic at least as extreme as the value observed. The P-value gives very strong evidence against the null hypothesis that $\beta_2 = 0$.

c) The 95 confidence interval for β_2 is $b_2 \pm t_{.025}(se) = 15.170 \pm 1.97(5.330)$, or (4.67, 25.67). The plausible values for the slope for number of bedrooms, when controlling for house size, range from about 4.7 to 25.7. This is more informative than the significance test because it not only tells us that the slope is likely different from 0, but it gives us a range of plausible values for the slope.

Section 13.4: Checking a Regression Model Using Residual Plots

13.32 Body weight residuals

a) These give us information about the conditional distribution.

b) The distribution of the residuals has a shape that is slightly right-skewed, although not too far from bell shaped. This suggests that the conditional distribution of *y* may be right skewed, although not too far from normal.

13.33 Strength residuals

a) The values of BP60 play a role in determining the standardized residuals against which the LP200 values are plotted.

b) The residuals are closer to 0 at lower values of LP200.

13.33 (continued)

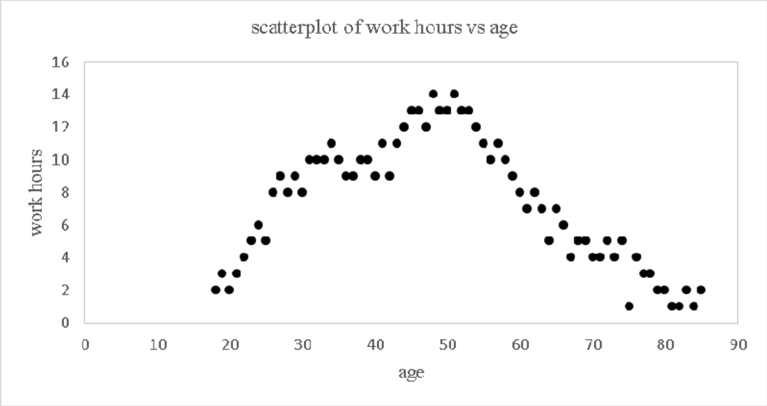
c) Without the three points with standardized residuals around –2, the data do not appear as though there is more variability at higher levels of LP200. We should be cautious in looking at residuals plots because one or two observations might prevent us from seeing the overall pattern.

13.34 More residuals for strength

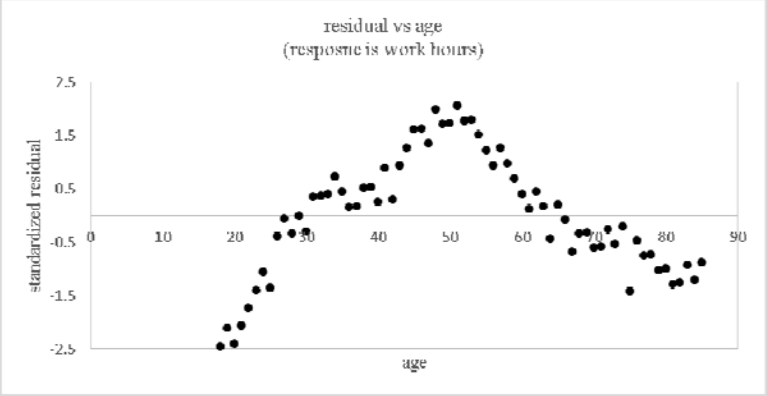
One might think that this suggests less variability at low levels and even less at high levels of BP60, but this may merely reflect fewer points in those regions. Overall, it seems OK.

13.35 Nonlinear effects of age

a) Example of possible scatterplot:

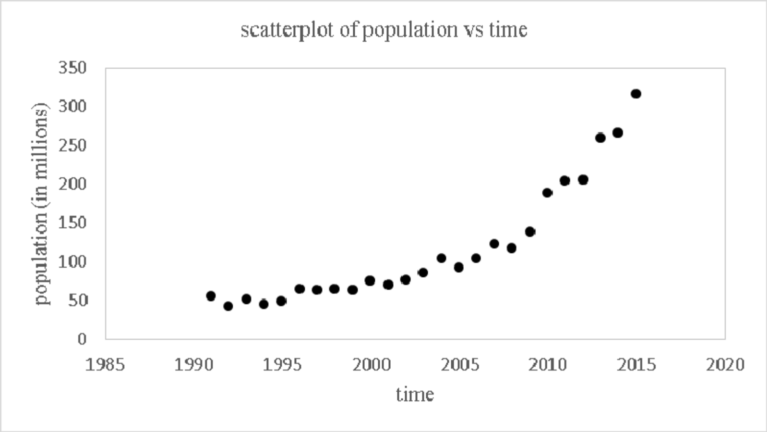


b) Example of plot of standardized residuals against the values of age:



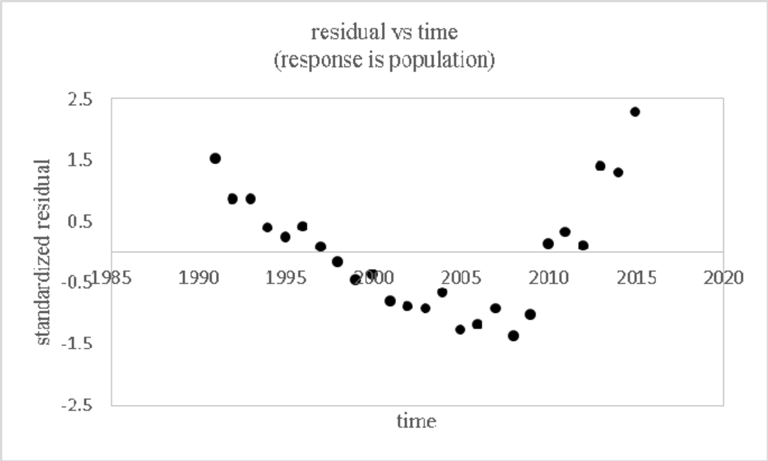
13.36 Population growth with time

a) As population generally grows exponentially with time the relationship between population and time will not be linear and hence the scatterplot will be something like the following:



13.36 (continued)

b) Example of plot of standardized residuals against the values of time:



13.37 Why inspect residuals?

The purpose of performing residual analysis is to determine if assumptions are met for tests. We standardize residuals, i.e. subtract mean and divide by standard deviation, so that irrespective of the data we can use the same threshold for outliers' detection. For instance, standardized residual shall fall between –3 and 3 for the observation to be not an outlier.

13.38 College athletes

a) The bottom left and bottom middle plots give 1RM as the response variable. They both show strong, positive associations with 1RM.

b) $\hat{y}$ = 55.01 + 0.1668LBM + 1.658REPS70; 1.66 is the amount that predicted maximum bench press changes for a one unit increase in number of repetitions, controlling for lean body mass.

c) $R^2 = 0.832$; Using these variables together to predict 1RM reduces the prediction error by 83%, relative to using $\bar{y}$ alone to predict BP.

d) The multiple correlation is 0.91. There is a strong association between the observed and predicted 1RMs.

e) $F = 7641.5/50.7 = 150.75$; The P-value is approximately 0.000. If the null hypothesis were true, the probability would be close to 0 of getting a test statistic at least as extreme as the value observed. We have very strong evidence that BP is not independent of these two predictors.

f) 1) Assumptions: We assume a random sample and that the model holds (each explanatory variable has a straight-line relation with μ_y, with same slope for all combinations of values of other predictors in model, and there is a normal distribution for y with the same standard deviation at each combination of values of the other predictors in the model). Here, the 64 athletes were a convenience sample, not a random sample, so inferences are tentative.

2) Hypotheses: H_0: $\beta_1 = 0$; H_a: $\beta_1 \neq 0$

3) Test statistic: $t = 2.22$

4) *P*-value: The P-value is 0.030.

5) Conclusion: If the null hypothesis were true, the probability would be 0.03 of getting a test statistic at least as extreme as the value observed. The P-value gives relatively strong evidence against the null hypothesis that $\beta_1 = 0$.

g) The histogram suggests that the residuals are roughly bell-shaped about 0. They fall between about –3 and +3. The shape suggests that the conditional distribution of the response variable is roughly normal.

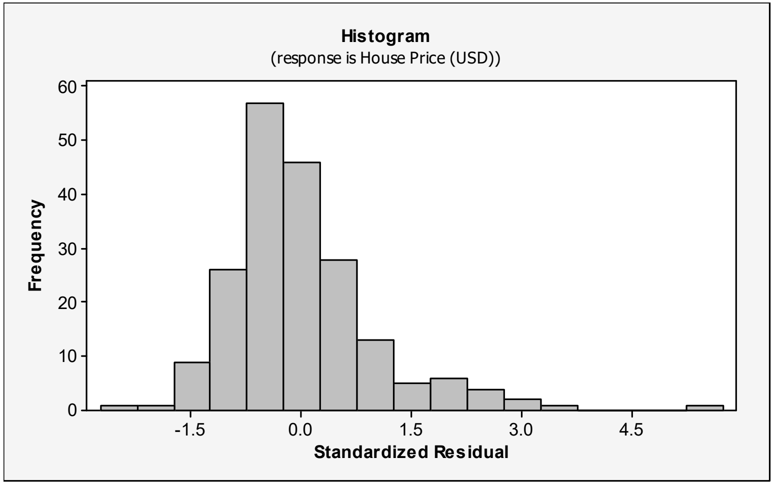
h) The plot of residuals against values of REPS70 describes the degree to which the response variable is linearly related to this particular explanatory variable. It suggests that the residuals are less variable at smaller values of REPS70 than at larger values of REPS70.

13.38 (continued)

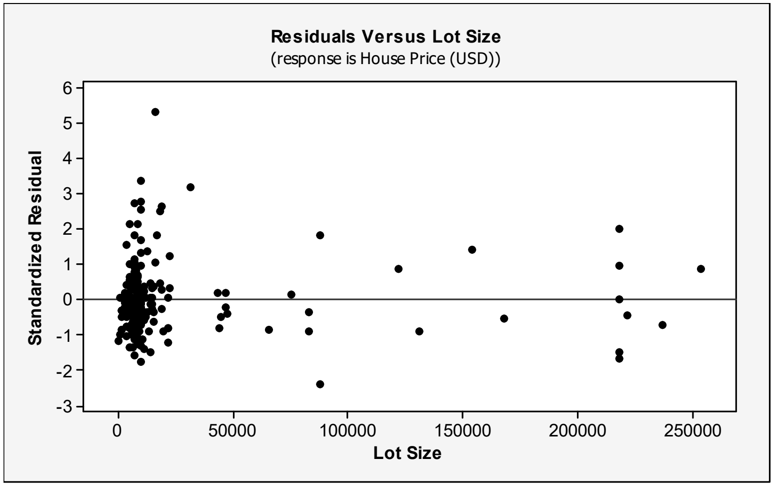
i) The individual with REPS70 around 32 and standardized residual around –3 had a 1RM value considerably lower than predicted.

13.39 House prices

a) The histogram checks the assumption that the conditional distribution of *y* is normal, at any fixed values of the explanatory variables. Although the distribution appears mostly normal, there is an outlier with a very large standardized residual greater than 6.0 that should be considered.

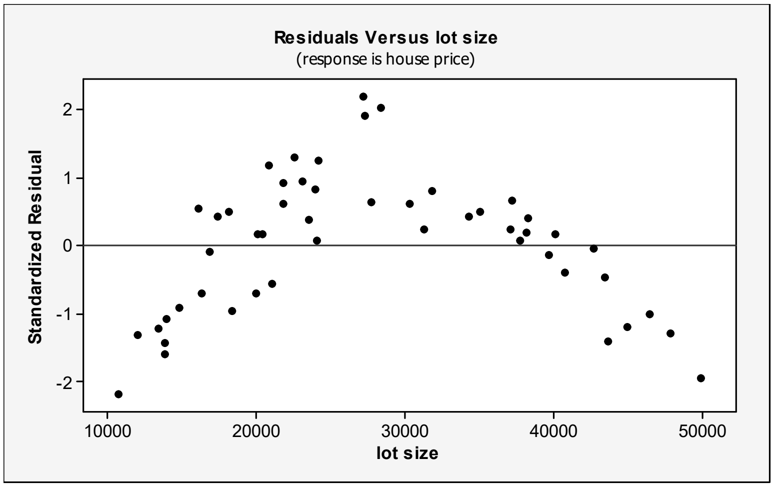


b) This plot checks the assumption that the regression equation approximates well the true relationship between the predictors and the response. The same large standardized residual (greater than 6) is evident, but there does not appear to be any discernable pattern in the residuals.



13.40 Selling prices level off

For large values of lot size, residuals would be negative and have decreasing trend. The assumption of a straight-line relationship between price and lot size, for given number of bedrooms, is violated.



Section 13.5: Regression and Categorical Predictors

13.41 U.S. and foreign used cars

a) U.S. Cars: $\hat{y} = 20,493 - 1185x_1 - 2379(1) = 18,114 - 1185x_1$

Foreign Cars: $\hat{y} = 20,493 - 1185x_1 - 2379(0) = 20,493 - 1185x_1$

b) Both prediction equations in (a) have the same slope of –1185. For a one-year increase in the age of a car, we predict that the price drops by $1,185. Since the slope is the same, this applies for both types of cars.

c) Using the equations from (a):

(i) U.S. Cars: $\hat{y} = 18,114 - 1185x_1 = 18,114 - 1185(8) = 8634$

(ii) Foreign Cars: $\hat{y} = 20,493 - 11,857x_1 = 20,493 - 1185(8) = 11,013$

The difference between them is 8634 – 11,013 = –2379, the coefficient for the indicator variable for type.

13.42 Mountain bike prices

a) For each one-pound increase in weight, the predicted price decreases by $53.75 for the same type of suspension.

b) For a given weight, the predicted price for a front end suspension bike is $643.60 less than for a full suspension bike.

13.43 Predict using house size and condition

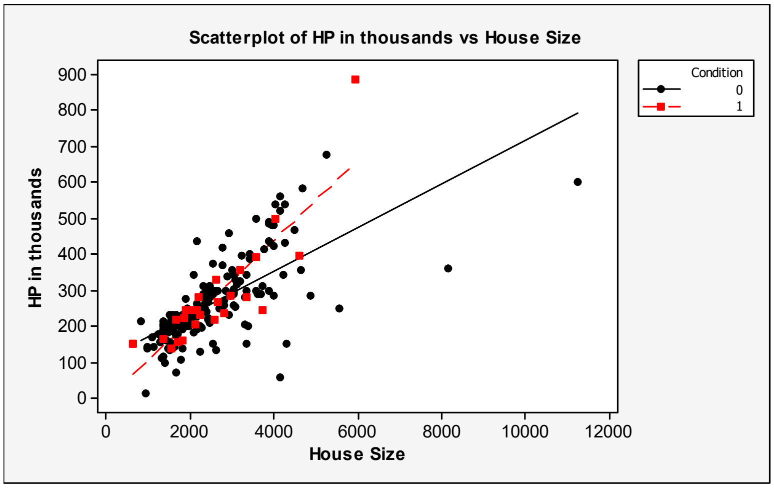
a) $\hat{y} = 96.3 + 0.0665\text{House\_Size} + 12.9\text{Condition}$

Good Condition: $\hat{y} = 96.3 + 0.0665\text{House\_Size} + 12.9(1) = 109.2 + 0.0665\text{House\_Size}$

Not Good Condition: $\hat{y} = 96.3 + 0.0665\text{House\_Size} + 12.9(0) = 96.3 + 0.0665\text{House\_Size}$

13.43 (continued)

b)



c) The difference between the predicted selling price between homes in good and not good condition, controlling for house size, is the slope for condition, 12.9.

13.44 Quality and productivity

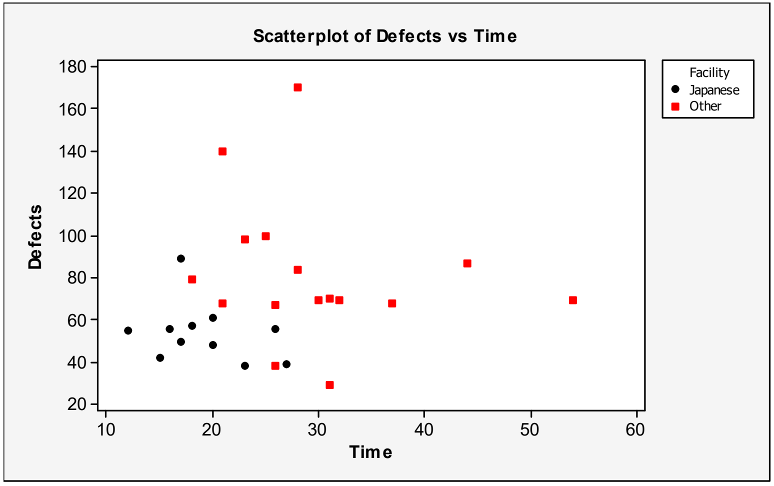
a) (i) Minimum: $\hat{y} = 61.3 + 0.35(12) = 65.5$

(ii) Maximum: $\hat{y} = 61.2 + 0.35(54) = 80.2$

b) When controlling for region, an increase of one hour leads to a decrease in predicted defects of 0.78 per 100 cars. Japanese facilities had 36 fewer predicted defects, on average, than did other facilities.

c) Simpson's paradox has occurred because the direction of the association between time and defects reversed when the variable of whether facility is Japanese was added.

d) Simpson's paradox occurred because, overall, Japanese facilities have fewer defects and take less time, whereas other facilities have more defects and take more time. When the data are looked at together, this leads to an overall positive association between defects and time.



13.45 Predicting pizza sales

a) $x_1 = 1$ if inner city; 0 if other; $x_2 = 1$ if suburbia; 0 if other.

b) Inner city: $\hat{y} =\ 6.9 + 1.2(1) + 0.5(0) = 8.1$; interstate exits: $\hat{y} =\ 6.9 + 1.2(0) + 0.5(0).$.

Interstate exit restaurants have predicted sales of \$8.10 – \$6.90 = –\$1.20, or \$1.20 less than inner city restaurants.

13.46 Houses, size and garage

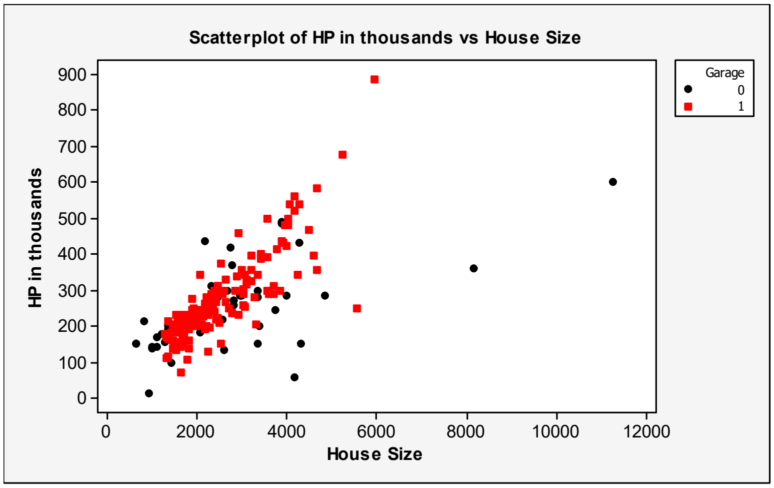
a) From technology: HP in thousands = 64.7 + 0.0674House_Size + 40.3Garage
For homes with a garage, HP in thousands = 104.96 + 0.0674House_Size.
For homes without a garage, HP in thousands = 64.66 + 0.0674House_Size.

b) The coefficient, 40.3, indicates that the predicted selling price for houses with a garage is $40,300 higher than for houses without a garage.

13.47 House size and garage interact?

a) The assumption of no interaction means that we are assuming that the slope for house size is the same for houses with and without a garage.

b)



13.48 Equal slopes for car prices?

a) interaction

b) The prediction equations do not have the same slope. For a one-year increase in the age of a used U.S. car, we predict that the price drops by $1,715. For a one-year increase in the age of a foreign car, we predict that the price drops by only $557. The effect of an increase in age differs for the two types of cars, so they need to be treated separately.

c) U.S. Cars: $\hat{y} = 23{,}417 - 1715x_1 = 23{,}417 - 1715(8) = \$9{,}697$

Foreign Cars: $\hat{y} = 15{,}536 - 557x_1 = 15{,}536 - 557(8) = \$11{,}080$

13.49 Comparing revenue

a) For one equation, we would have two explanatory variables, one for number of customers who visit the filling station and one for location (inner city = 1, and interstate exit = 1).

b) For separate equations, we would have a model with one explanatory variable, number of customers, for the inner city location, and another model with one explanatory variable, number of customers, for the interstate exit location.

Section 13.6: Modeling a Categorical Response

13.50 Income and credit cards

$$\hat{p} = \frac{e^{-3.52+0.105x}}{1+e^{-3.52+0.105x}} = \frac{e^{-3.52+0.105(35)}}{1+e^{-3.52+0.105(35)}} = \frac{1.17}{2.17} = 0.54$$

13.51 Hall of Fame induction

a) For 359 runs: $\hat{p} = \dfrac{e^{-6.7+0.0175(359)}}{1+e^{-6.7+0.0175(359)}} = \dfrac{0.6587}{1.6587} = 0.397$

For 369 runs: $\hat{p} = \dfrac{e^{-6.7+0.0175(369)}}{1+e^{-6.7+0.0175(369)}} = \dfrac{0.7847}{1.7847} = 0.440$

b) For 465 runs: $\hat{p} = \dfrac{e^{-6.7+0.0175(465)}}{1+e^{-6.7+0.0175(465)}} = \dfrac{4.2102}{5.2102} = 0.808$

For 475 runs: $\hat{p} = \dfrac{e^{-6.7+0.0175(475)}}{1+e^{-6.7+0.0175(475)}} = \dfrac{5.0153}{6.0153} = 0.834$

13.52 Cancer prediction.

a) Q1: $\dfrac{e^{-2.165+2.585(x)}}{1+e^{-2.165+2.585(x)}} = \dfrac{e^{-2.165+2.585(1)}}{1+e^{-2.165+2.585(1)}} = \dfrac{1.52}{2.52} = 0.60$

Q3: $\dfrac{e^{-2.165+2.585(x)}}{1+e^{-2.165+2.585(x)}} = \dfrac{e^{-2.165+2.585(1.85)}}{1+e^{-2.165+2.585(1.85)}} = \dfrac{13.70}{14.70} = 0.93$

b) 0.93 – 0.60 = 0.33. The probability increases by 0.33 over the middle half of the sampled radii.

13.53 Cancer prediction (continued)

b) $x = -\hat{\alpha}/\hat{\beta} = 2.165/2.585 = 0.84$

c) For females having a tumour with radius over 0.84 cm, the estimated probability is greater than 0.50.

d) For females having a tumour with radius under 0.84 cm, the estimated probability is less than 0.50.

13.54 Voting and income

a) $\dfrac{e^{-1.00+0.02(10)}}{1+e^{-1.00+0.02(10)}} = \dfrac{0.4493}{1.4493} = 0.31$

b) $\dfrac{e^{-1.00+0.02(100)}}{1+e^{-1.00+0.02(100)}} = \dfrac{2.718}{3.718} = 0.73$

The predicted probability of voting Republican increases as income increases.

13.55 Equally popular candidates

a) $x = -\hat{\alpha}/\hat{\beta} = 1.00/0.02 = 50$, or \$50,000

b) (i) Above \$50,000, the estimated probability of voting for the Republican candidate is greater than 0.50.

(ii) Below \$50,000, the estimated probability of voting for the Republican candidate is less than 0.50.

c) When x is close to the value at which $p = 0.50$, the approximate change in the predicted probability p for a one-unit increase in x, \$1,000 in this case, is $\beta/4 = 0.02/4 = 0.005$.

13.56 Many predictors of voting

a) As family income increases, people are, on average, more likely to vote Republican. As number of years of education increases, people are, on average, more likely to vote Republican. Men are more likely, on average, to vote Republican than are women.

b) (i) $\hat{p} = \dfrac{e^{-2.40+0.02x_1+0.08x_2+0.20x_3}}{1+e^{-2.40+0.02x_1+0.08x_2+0.20x_3}} = \dfrac{e^{-2.40+0.02(40)+0.08(16)+0.20(1)}}{1+e^{-2.40+0.02(40)+0.08(16)+0.20(1)}} = \dfrac{e^{-0.12}}{1+e^{-0.12}} = \dfrac{0.8869}{1.8869} = 0.47$

(ii) $\hat{p} = \dfrac{e^{-2.40+0.02x_1+0.08x_2+0.20x_3}}{1+e^{-2.40+0.02x_1+0.08x_2+0.20x_3}} = \dfrac{e^{-2.40+0.02(40)+0.08(16)+0.20(0)}}{1+e^{-2.40+0.02(40)+0.08(16)+0.20(0)}} = \dfrac{e^{-0.32}}{1+e^{-0.32}} = \dfrac{0.7261}{1.7261} = 0.42$

13.57 Graduation, gender and race

a) The response variable is whether or not the student graduated (yes or no).

b)

		Graduated		
Race	**Gender**	**Yes**	**No**	**Total**
White	Female	10,781	20,468	31,249
	Male	10,727	28,856	39,583
Black	Female	2309	10,885	13,194
	Male	2054	15,653	17,707

c) Based on these estimates, white women have the highest estimated probability of graduating. The coefficient for race is positive, indicating that 1 (white) would lead to a higher estimated probability of graduating than would 0 (black). Similarly, the coefficient for gender is positive, indicating that 1 (female) would lead to a higher estimated probability of graduating than would 0 (male).

13.58 Death penalty and race

a) Controlling for victim's race, the proportion of black defendants who received the death penalty was 15/191 = 0.079, and the proportion of white defendants who received the death penalty was 53/483 = 0.110, so the death penalty was more likely for black defendants.

b) According to this equation, the death penalty is predicted to be most likely for black defendants who had white victims. We know this because the coefficient for defendant race is negative; therefore 0 (black) would lead to a higher predicted death penalty proportion than would 1 (white). Also, the coefficient for victim's race is positive; therefore, 1 (white) would lead to a higher predicted death penalty proportion than would 0 (black).

13.59 Death penalty probabilities

a) Black defendant, white victim: $\hat{p} = \dfrac{e^{-3.596-0.868(0)+2.404(1)}}{1+e^{-3.596-0.868(0)+2.404(1)}} = \dfrac{0.304}{1.304} = 0.233$

b)

	Defendant's Race	
Victim's Race	**White**	**Black**
White	0.113	0.233
Black	0.011	0.027

Defendant's race has the same effect on the predicted probability of receiving the death penalty for both white and black victims. In both cases, black defendants are predicted to be more likely to receive the death penalty than are white defendants.

c)

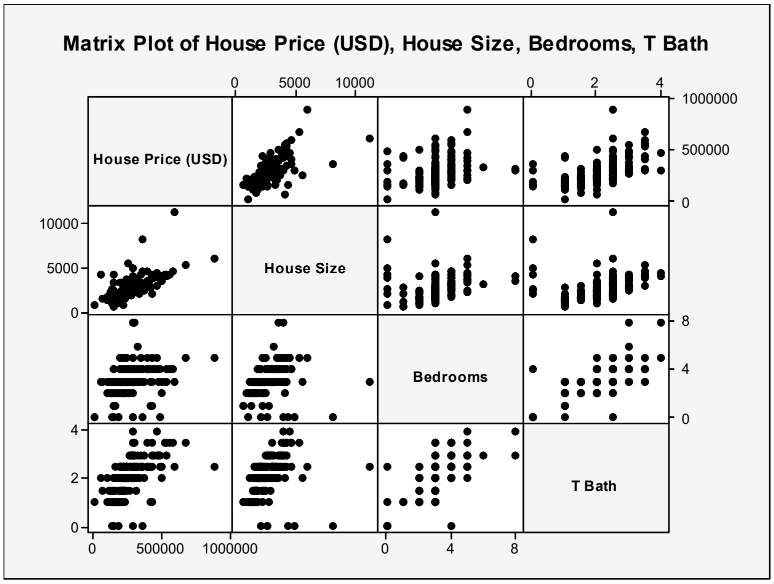
	Death Penalty		Percent
Defendant's Race	**Yes**	**No**	**Yes**
White	53	430	11.0
Black	15	176	7.9

This is an example of Simpson's paradox because the direction of association changes when we ignore a third variable. When ignoring victim's race, the predicted proportion of whites receiving the death penalty, rather than blacks, is now the higher of the two. This occurs because there are more white defendants with white victims than any other group.

Chapter Problems: Practicing the Basics

13.60 House prices

a)



The plots that pertain to selling price as a response variable are those across the top row. The highly discrete nature of x_2 and x_3 limits the number of values these variables can take on. This is reflected in the plots, particularly the plot for bedrooms by baths.

b) From technology: House Price = 39,001 + 53.2House_Size – 7885Bedrooms + 57796Bath
When number of bedrooms and number of bathrooms are fixed, an increase of one square foot in house size leads to an increase of $53.2 in predicted selling price.

c) $R^2 = \dfrac{1.60874 \times 10^{12}}{2.66887 \times 10^{12}} = 0.603$; This indicates that predictions are 60% better when using the prediction equation instead of using the sample mean $\bar{y}$ to predict y.

d) The multiple correlation, 0.78, is the square root of R^2. It is the correlation between the observed y-values and the predicted $\hat{y}$-values.

e) (i) Assumptions: multiple regression equation holds, data gathered randomly, normal distribution for y with same standard deviation at each combination of predictors.

(ii) Hypotheses: H_0: $\beta_1 = \beta_2 = \beta_3 = 0$; H_a: At least one β parameter differs from 0.

(iii) Test statistic: $F = (5.36245 \times 10^{11})/5{,}408{,}848{,}557 = 99.14$

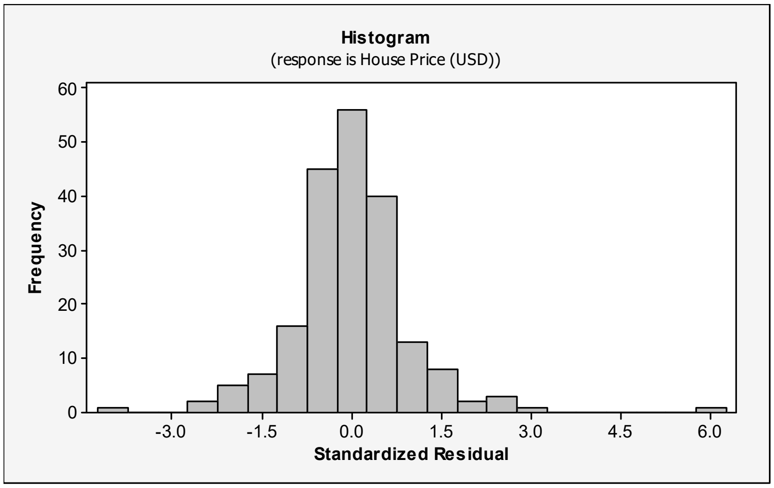
(iv) P-value: The P-value is approximately 0 for $df(3,196)$.

(v) Conclusion: If the null hypothesis were true, the probability would be close to 0 of getting a test statistic at least as extreme as the value observed. We have very strong evidence that at least one explanatory variable has an effect on y.

f) The t statistic is –1.29 with a one-sided P-value of 0.2/2 = 0.1. If the null hypothesis were true, the probability would be 0.1 of getting a test statistic at least as extreme as the value observed. At a significance level of 0.05, we cannot reject the null. It is plausible that the number of bedrooms does not have an effect on selling price. This is likely not significant because it is correlated with the other explanatory variables in this model. It might be associated with selling price on its own, but might not provide additional predictive information over and above the other explanatory variables.

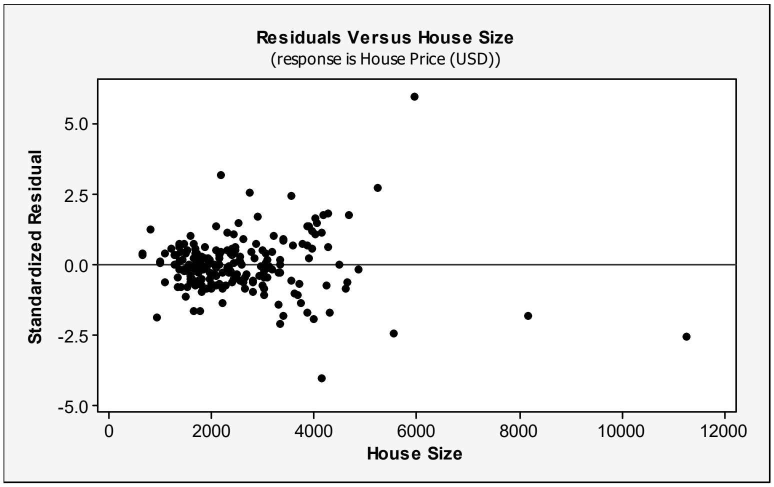
13.60 (continued)

g)



This histogram describes the shape of the conditional distribution of *y* at given values of the explanatory variables. The distribution appears fairly close to normal with two large outliers (greater than 3 in absolute value).

h)



This plot depicts the size of the residuals for the different house sizes observed in this sample. It indicates possibly greater residual variability (and hence, greater variability in selling price) as house size increases.

13.61 Predicting body strength

a) $\hat{y} = 60.6 + 1.33\text{BP\_60} + 0.21\text{LP\_200} = 60.6 + 1.33(10) + 0.21(20) = 78.1$;

The residual is $y - \hat{y} = 85 - 78.1 = 6.9$.

b) For athletes who have LP_200 = 20, the equation is $\hat{y} = 60.6 + 1.33\text{BP\_60} + 0.21(20) = 64.8 + 1.33\text{BP\_60}$; the slope of 1.33 indicates that, when controlling for LP_200, an increase of one in BP_60 leads to an increase of 1.33 in predicted BP.

c) The small difference indicates that the additional variable does not add much predictive ability.

13.62 Softball data

a) From technology, the prediction equation is: Difference = –5.00 + 0.934Hits – 1.61Errors; the slopes indicate that for each increase of one hit, the predicted difference increases by 0.93, and for each increase of one error, the predicted difference decreases by 1.61.

b) When errors = 0, the prediction equation is $\hat{y} = -5.00 + 0.934\text{Hits}$; for a predicted difference of 0, we would need 5.35 hits. ($0 = -5.00 + 0.934\text{Hits} = 0.934\text{Hits} \Rightarrow 5.00 \Rightarrow \text{Hits} = 5.00/0.934 \Rightarrow \text{Hits} = 5.35$) Thus, the team would need six or more hits so that the predicted difference is positive (if they can play error-free ball).

13.63 Violent crime

a) $\hat{y} = -270.7 + 28.334x_1 + 5.416x_2 = -270.7 + 28.334(10.2) + 5.416(92.1) = 517.1$;
The residual is 476 – 517.1 = –41.1. The violent crime rate for Massachusetts is 41.1 lower than predicted from this model.

b) (i) $\hat{y} = -270.7 + 28.334x_1 + 5.416(0) = -270.7 + 28.334x_1$

(ii) $\hat{y} = -270.7 + 28.334x_1 + 5.416(100) = 270.9 + 28.334x_1$

As percent living in urban areas increases from 0 to 100, the intercept of the regression equation increases from –270.7 to 270.9. When the second explanatory variable, poverty rate, is held constant, the increase in percent living in urban areas from 0 to 100 would result in an increase in predicted violent crime rate of 541.6.

13.64 Effect of poverty on crime

The slope with x_3 in the model represents the effect of poverty when controlling for percentage of single-parent families, as well as percent living in urban areas. The slope without x_3 in the model represents the effect of poverty when controlling only for percent living in urban areas.

13.65 Modeling fertility

a) –0.661
b) 0.443
c) 279,160
d) 155,577
e) 44.50
f) 33.4554
g) –5.21
h) 0.000

13.66 Significant fertility prediction?

a) F = 61791/1119 = 55.21; The P-value is 0.000; if the null hypothesis were true, the probability would be close to 0 of getting a test statistic at least as extreme as the value observed. We have very strong evidence that at least one of these explanatory variables predicts y better than the sample mean does.

b) The significance test would not be relevant if we were not interested in nations beyond those in the study; if this were the case, the group of nations that we studied would be a population and not a sample.

13.67 GDP, CO_2, and Internet

a) Predicted GDP $= 39{,}892 + 798\text{CO2} - 399\text{NoInternet}$

b) For a given percentage of the population not using the internet, we predict GDP to increase by \$798,000 for every metric ton increase in CO_2 emissions. Controlling for CO_2 emissions, we predict GDP to decrease by \$399,000 for every percentage point increase in the percentage of the population not using the Internet.

13.68 Education and gender in modeling income

Since the effect of education on income changes depending on gender, the explanatory variables, education and gender, are said to interact.

13.69 Horseshoe crabs and width

a) Q1: $\dfrac{e^{-12.351+0.497(24.9)}}{1+e^{-12.351+0.497(24.9)}}=\dfrac{1.0246}{2.0246}=0.51$

Q3: $\dfrac{e^{-12.351+0.497(27.7)}}{1+e^{-12.351+0.497(27.7)}}=\dfrac{4.1202}{5.1202}=0.81$

Across the middle half of crabs with respect to width, the estimated probability of having a male partner increases 0.30, from 0.51 to 0.81.

b) (i) $x=-\hat{\alpha}/\hat{\beta}=12.351/0.497=24.9$

(ii) Above a width of 24.9, the estimated probability of having a male partner nearby is greater than 0.50.

(iii) Below a width of 24.9, the estimated probability of having a male partner nearby is less than 0.50.

13.70 AIDS and AZT

a) The negative sign indicates that if an individual used AZT, the predicted probability of developing AIDS symptoms was lower.

b) black/yes: $\hat{p}=\dfrac{e^{-1.074-0.720(1)+0.056(0)}}{1+e^{-1.074-0.720(1)+0.056(0)}}=\dfrac{0.1663}{1.1663}=0.14$

black/no: $\hat{p}=\dfrac{e^{-1.074-0.720(0)+0.056(0)}}{1+e^{-1.074-0.720(0)+0.056(0)}}=\dfrac{0.3416}{1.3416}=0.26$

c) 1) Assumptions: The data were generated randomly. The response variable is binary.

2) Hypotheses: H_0: $\beta_1=0$; H_a: $\beta_1\neq 0$

3) Test statistic: $z=(b-0)/se=(-0.720-0)/0.279=-2.58$

4) *P*-value: 0.010

5) Conclusion: If the null hypothesis were true, the probability would be 0.01 of getting a test statistic at least as extreme as the value observed. We have very strong evidence against the null hypothesis that $\beta_1=0$. At significance level of 0.05, we can reject H_0.

13.71 Factors affecting first home purchase

a) We know that, other things being fixed, the predicted probability of home ownership increases with husband's earnings, wife's earnings, number of children, and home ownership because the coefficients are positive for all of these explanatory variables.

b) We know that the number of years married, given the other variables in the model, shows little evidence of an effect because the estimate divided by the standard error is small (equals –0.93).

Chapter Problems: Concepts and Investigations

13.72 Student data

The reports will be different for each student, but could include the following, along with associated graphs.
From Minitab:

```
The regression equation is
college_GPA = 2.83 + 0.203 high_sch_GPA - 0.0092 sports
Predictor         Coef  SE Coef      T      P
Constant        2.8293   0.3385   8.36  0.000
high_sch_GPA   0.20309  0.09753   2.08  0.042
sports        -0.00922  0.01161  -0.79  0.430

S = 0.341590   R-Sq = 8.8%   R-Sq(adj) = 5.6%

Analysis of Variance
Source          DF      SS      MS     F      P
Regression       2  0.6384  0.3192  2.74  0.073
Residual Error  57  6.6510  0.1167
Total           59  7.2893

Unusual Observations
Obs  high_sch_GPA  college_GPA     Fit  SE Fit  Residual St Resid
 11          2.30       2.6000  3.1580  0.1476   -0.5580  -1.81 X
 42          2.00       3.0000  3.1616  0.1351   -0.1616  -0.52 X
 50          3.00       4.0000  3.3002  0.1224    0.6998   2.19 R
 60          3.40       3.0000  3.3722  0.1345   -0.3722  -1.19 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.
```

13.73 Why regression?

The explanations will be different for each student, but should indicate that multiple regression uses *more than one* characteristic of a subject to predict some outcome.

13.74 Unemployment and GDP

a) The positive sign indicates that GDP is predicted to increase as the unemployment rate increases. One would think that GDP would decrease for increasing unemployment rate.

b) The P-value of 0.86 indicates that the coefficient is not significantly different from 0. However, this doesn’t automatically imply that it has no effect. It only means it has no effect when CO_2 and NoInternet are already included in the model.

c) R^2

d) This means that the prediction error is essentially the same whether unemployment is in the model or not. Therefore, unemployment does not help in predicting GDP when CO_2 and NoInternet are in the model.

13.75 Multiple choice: Interpret parameter

The best answer is (a).

13.76 Multiple choice: Interpret indicator

The best answer is (d).

13.77 Multiple choice: Regression effects

The best answer is (a).

13.78 True or false: *R* and R^2

a) True
b) No
c) True
d) False. R^2 is an increasing function of the number of independent variables included in the regression model.

13.79 True or false: Regression

a) True
b) True
c) False, *R*-squared cannot exceed 1.
d) False, the predicted values $\hat{y}$ cannot correlate negatively with *y*. Otherwise, the predictions would be worse than merely using $\overline{y}$ to predict *y*.

13.80 True or false: Slopes

a) True, the slope for this variable is positive in the bivariate regression equation.
b) False, a one-unit increase in x_1 corresponds to a change of 0.45 in the predicted value of *y*, only when we ignore x_2.
c) True, the slope for x_2 is 0.003 when controlling for x_1. 0.003 multiplied by 100 is 0.30.

13.81 Scores for religion

Indicator variables for a particular explanatory variable must be binary. A variable would equal 1 if an observation fell in that category, and 0 if it did not. Here, we'd have to have three variables: one for Protestant (1 = Protestant, 0 = other), one for Catholic (1 = Catholic, 0 = other), and one for Jewish (1 = Jewish, 0 = other). We would not need one for the "other" category because we would know that someone was in that category if he/she were not in the other three. Using numerical scores would treat the religion as quantitative with equidistant categories, which is not appropriate.

13.82 Lurking variable

y = math achievement score, x_1 = height, x_2 = age for a sample of children from all the different grades in a school system.

13.83 Properties of R^2

$R^2 = 1$ only when all residuals are 0 because when all regression predictions are perfect (each $y = \hat{y}$), residual SS = $\sum(y-\hat{y})^2$ = 0. When residual SS = 0, R^2 is total SS divided by total SS, which must be 1. On the other hand, $R^2 = 0$ when each $\hat{y} = \overline{y}$. In that case, the estimated slopes all equal 0, and the correlation between *y* and each explanatory variable equals 0. Under these circumstances, the residual SS would equal the total SS, and R^2 would then be 0. In practical terms, it means that R^2 is only 1 when the regression model predicts *y* perfectly, and it is only 0 when it doesn't predict *y* at all.

13.84 Why an *F* test?

When doing multiple significance tests, one may be significant merely by random variation. When there are many explanatory variables, doing the *F* test first provides protection from doing lots of *t* tests and having one of them be significant merely by random variation when, in fact, there truly are no effects in the population.

13.85 Multicollinearity

a) From Minitab:

```
Source          DF      SS      MS     F      P
Regression       2  1200.2   600.1  3.74  0.030
Residual Error  54  8674.4   160.6
Total           56  9874.6
```

The P-value of 0.030 is less than 0.05.

b) From Minitab:

```
Predictor     Coef  SE Coef     T      P
Constant    54.019    9.677  5.58  0.000
WT (lbs)    0.1251   0.1378  0.91  0.368
BF%         0.3315   0.6950  0.48  0.635
```

The P-values of 0.368 and 0.635 are both larger than 0.35.

13.86 Logistic versus linear

When $x = 0$, a little extra income is not going to make a difference; one likely can't afford a home no matter what. Similarly, when $x = 50{,}000$, a little extra income won't make much difference; one likely can afford a home no matter what. Only in the middle, when $x = 500$, is the extra income likely to "push" someone over the income level at which he or she can afford a home. In such a case, a linear regression model would not be appropriate, although a logistic regression model would.

♦♦13.87 Adjusted *R*²

10: Adjusted $R^2 = 0.500 = \{2/[10-(2+1)]\}(1-0.500) = 0.500 - 0.143 = 0.357$

100: Adjusted $R^2 = 0.500 = \{2/[100-(2+1)]\}(1-0.500) = 0.500 - 0.010 = 0.490$

1000: Adjusted $R^2 = 0.500 = \{2/[1000-(2+1)]\}(1-0.500) = 0.500 - 0.001 = 0.499$

As the sample size increases, adjusted R^2 approaches R^2.

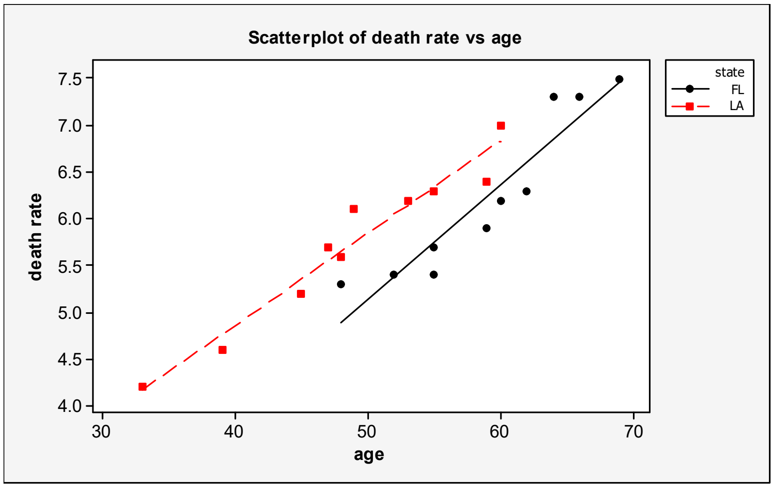
♦♦13.88 *R* can't go down

When you add a predictor, if it has no effect its coefficient is 0. Then the prediction equation is exactly the same as with the simpler model without that variable and *R* will be exactly the same as before. If having a nonzero coefficient results in better predictions overall, then *R* will increase.

♦♦13.89 Indicator for comparing two groups

If we wanted to compare two groups on a given variable, we could use regression analysis. The response variable *y* would be the same. The explanatory variable would be the two levels of the groups; one would be assigned 0 and one would be assigned 1. $\mu_1 = \mu_2$ would correspond to $\beta = 0$.

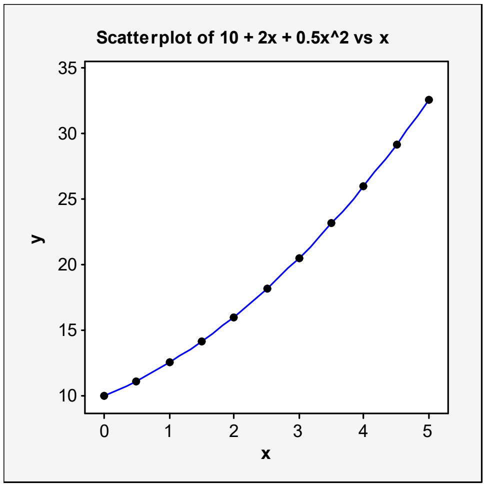
♦♦13.90 Simpson's paradox



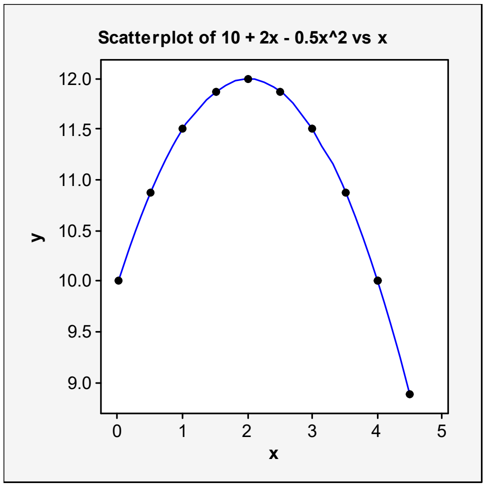
♦♦13.91 Parabolic regression

a) This is a multiple regression model with two variables. If *x*, for example, is 5, we multiply 5 by β_1 and its square, 25, by β_2.

b) (i)



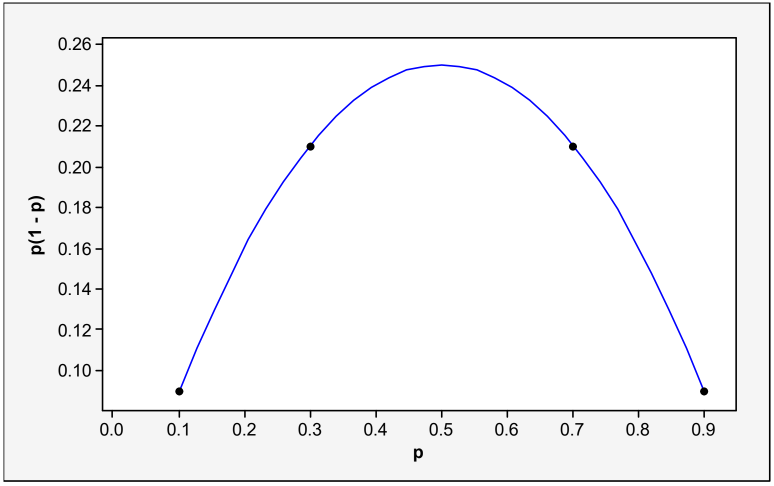
(ii)



♦♦13.92 Logistic slope

a) When *p* = 0.5, *p*(1 – *p*) = 0.5(1 – 0.5) = 0.25; 0.25 multiplied by β is the same as $\beta/4$.

b) 0.1(1 – 0.1) = 0.09, 0.3(1 – 0.3) = 0.21, 0.7(1 – 0.7) = 0.21, and 0.9(1 – 0.9) = 0.09; as *p* gets closer and closer to 1, the slope approaches 0.



♦♦13.93 When is *p* = 0.50?

$$p = \frac{e^{\alpha+\beta(-\alpha/\beta)}}{1+e^{\alpha+\beta(-\alpha/\beta)}} = \frac{e^0}{1+e^0} = \frac{1}{1+1} = \frac{1}{2} = 0.5$$

Chapter Problems: Student Activities

13.94 Class data

The responses will be different for each class.