Exercises

Section 25.1

1. **Real estate assessment** A house in the upstate New York area from which the chapter data was drawn has 2 bedrooms and 1000 square feet of living area. Using the multiple regression model found in the chapter,

$$\widehat{Price} = 20{,}986.09 - 7483.10\ Bedrooms + 93.84\ Living\ Area.$$

a) Find the price that this model estimates.
b) The house just sold for \$135,000. Find the residual corresponding to this house.
c) What does that residual say about this transaction?

2. **Chocolate** A candy maker surveyed chocolate bars available in a local supermarket and found the following least squares regression model:

$$\widehat{Calories} = 28.4 + 11.37\ Fat(g) + 2.91\ Sugar(g).$$

a) The hand-crafted chocolate she makes has 15g of fat and 20g of sugar. How many calories does the model predict for a serving?
b) In fact, a laboratory test shows that her candy has 227 calories per serving. Find the residual corresponding to this candy. (Be sure to include the units.)
c) What does that residual say about her candy?

Section 25.2

T 3. **Movie profit** What can predict how much a motion picture will make? We have data on a number of movies that includes the *USGross* (in \$), the *Budget* (\$), the *Run Time* (minutes), and the average number of *Stars* awarded by reviewers. The first several entries in the data table look like this:

Movie	USGross (\$M)	Budget (\$M)	Run Time (minutes)	Stars
White Noise	56.094360	30	101	2
Coach Carter	67.264877	45	136	3
Elektra	24.409722	65	100	2
Racing Stripes	49.772522	30	110	3
Assault on Precinct 13	20.040895	30	109	3
Are We There Yet?	82.674398	20	94	2
Alone in the Dark	5.178569	20	96	1.5
Indigo	51.100486	25	105	3.5

We want a regression model to predict *USGross*. Parts of the regression output computed in Excel look like this:

Dependent variable is USGross(\$)
R-squared = 47.4% R-squared (adjusted) = 46.0%
s = 46.41 with 120 − 4 = 116 degrees of freedom

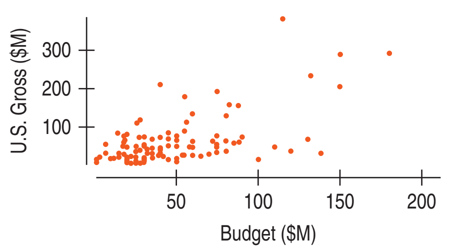
Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	−22.9898	25.70	−0.895	0.3729
Budget(\$)	1.13442	0.1297	8.75	≤0.0001
Stars	24.9724	5.884	4.24	≤0.0001
Run Time	−0.403296	0.2513	−1.60	0.1113

a) Write the multiple regression equation.
b) What is the interpretation of the coefficient of *Budget* in this regression model?

4. **Movie profit again** A middle manager at an entertainment company, upon seeing this analysis, concludes that the longer you make a movie, the less money it will make. He argues that his company's films should all be cut by 30 minutes to improve their gross. Explain the flaw in his interpretation of this model.

Section 25.3

T 5. **Movie profit once more** For the movies examined in Exercises 3 and 4, here is a scatterplot of *USGross* vs. *Budget*:

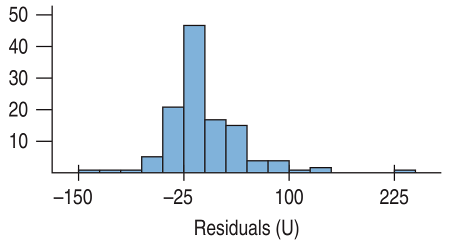


What (if anything) does this scatterplot tell us about the following Assumptions and Conditions for the regression?

a) Linearity condition
b) Equal Spread condition
c) Normality assumption

Copyright © 2014 Pearson Education, Inc.

6. Movie profit reconsidered For the movies regression, here is a histogram of the residuals. What does it tell us about these Assumptions and Conditions?



a) Linearity condition
b) Nearly Normal condition
c) Equal Spread condition

Section 25.4

T **7. Movie profit model tests** Regression output for the movies again:

a) What is the null hypothesis tested for the coefficient of *Stars* in this table?
b) What is the *t*-statistic corresponding to this test?
c) What is the P-value corresponding to this *t*-statistic?
d) Complete the hypothesis test. Do you reject the null hypothesis?

8. More movie profit tests From the regression output of Exercise 3,

a) What is the null hypothesis tested for the coefficient of *Run Time*?
b) What is the *t*-statistic corresponding to this test?
c) Why is this *t*-statistic negative?
d) What is the P-value corresponding to this *t*-statistic?
e) Complete the hypothesis test. Do you reject the null hypothesis?

Section 25.5

T **9. Interpreting R^2** In the regression model of Exercise 3,

a) What is the R^2 for this regression? What does it mean?
b) Why is the "Adjusted R Square" in the table different from the "R Square"?

T **10. Regression output interpretation** Here is another part of the regression output for the movies in Exercise 3:

Source	Sum of Squares	df	Mean Square	F-Ratio
Regression	224995	3	74998.4	34.8
Residual	249799	116	2153.44	

a) Using the values from the table, show how the value of R^2 could be computed. Don't try to do the calculation, just show what is computed.
b) What is the *F*-statistic value for this regression?
c) What null hypothesis can you test with it?
d) Would you reject that null hypothesis?

Chapter Exercises

11. Interpretations A regression performed to predict selling price of houses found the equation

$$\widehat{Price} = 169{,}328 + 35.3\,Area + 0.718\,Lotsize - 6543\,Age$$

where *Price* is in dollars, *Area* is in square feet, *Lotsize* is in square feet, and *Age* is in years. The R^2 is 92%. One of the interpretations below is correct. Which is it? Explain what's wrong with the others.

a) Each year, a house *Ages* it is worth $6543 less.
b) Every extra square foot of *Area* is associated with an additional $35.30 in average price, for houses with a given *Lotsize* and *Age*.
c) Every dollar in price means *Lotsize* increases 0.718 square feet.
d) This model fits 92% of the data points exactly.

12. More interpretations A household appliance manufacturer wants to analyze the relationship between total sales and the company's three primary means of advertising (television, magazines, and radio). All values were in millions of dollars. They found the regression equation

$$\widehat{Sales} = 250 + 6.75\,TV + 3.5\,Radio + 2.3\,Magazines.$$

One of the interpretations below is correct. Which is it? Explain what's wrong with the others.

a) If they did no advertising, their income would be $250 million.
b) Every million dollars spent on radio makes sales increase $3.5 million, all other things being equal.
c) Every million dollars spent on magazines increases TV spending $2.3 million.
d) Sales increase on average about $6.75 million for each million spent on TV, after allowing for the effects of the other kinds of advertising.

T **13. Predicting final exams** How well do exams given during the semester predict performance on the final? One class had three tests during the semester. Computer output of the regression gives

Dependent variable is Final
s = 13.46   R-Sq = 77.7%   R-Sq(adj) = 74.1%

Predictor	Coeff	SE(Coeff)	t-Ratio	P-Value
Intercept	−6.72	14.00	−0.48	0.636
Test1	0.2560	0.2274	1.13	0.274
Test2	0.3912	0.2198	1.78	0.091
Test3	0.9015	0.2086	4.32	<0.0001

Analysis of Variance

Source	DF	SS	MS	F-Ratio	P-Value
Regression	3	11961.8	3987.3	22.02	<0.0001
Error	19	3440.8	181.1		
Total	22	15402.6			

(continued)

Copyright © 2014 Pearson Education, Inc.

a) Write the equation of the regression model.
b) How much of the variation in final exam scores is accounted for by the regression model?
c) Explain in context what the coefficient of *Test3* scores means.
d) A student argues that clearly the first exam doesn't help to predict final performance. She suggests that this exam not be given at all. Does Test1 have no effect on the final exam score? Can you tell from this model? (*Hint:* Do you think test scores are related to each other?)

T **14. Scottish hill races** Hill running—races up and down hills—has a written history in Scotland dating back to the year 1040. Races are held throughout the year at different locations around Scotland. A recent compilation of information for 71 races (for which full information was available and omitting two unusual races) includes the *Distance* (miles), the *Climb* (elevation gained during the run in ft), and the *Record Time* (seconds). A regression to predict the men's records as of 2000 looks like this:

Dependent variable is Men's record
R-squared = 98.0% R-squared (adjusted) = 98.0%
s = 369.7 with 71 − 3 = 68 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-Ratio
Regression	458947098	2	229473549	1679
Residual	9293383	68	136667	

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	−521.995	78.39	−6.66	<0.0001
Distance	351.879	12.25	28.7	<0.0001
Climb	0.643396	0.0409	15.7	<0.0001

a) Write the regression equation. Give a brief report on what it says about men's record times in hill races.
b) Interpret the value of R^2 in this regression.
c) What does the coefficient of *Climb* mean in this regression?

15. Home prices Many variables have an impact on determining the price of a house. A few of these are *Size* of the house (square feet), *Lotsize,* and number of *Bathrooms.* Information for a random sample of homes for sale in the Statesboro, Georgia, area was obtained from the Internet. Regression output modeling the *Asking Price* with *Square Footage* and number of *Bathrooms* gave the following result:

Dependent Variable is Asking Price
s = 67013 R-Sq = 71.1% R-Sq (adj) = 64.6%

Predictor	Coeff	SE(Coeff)	t-Ratio	P-Value
Intercept	−152037	85619	−1.78	0.110
Baths	9530	40826	0.23	0.821
Sq ft	139.87	46.67	3.00	0.015

Analysis of Variance

Source	DF	SS	MS	F-Ratio	P-Value
Regression	2	99303550067	49651775033	11.06	0.004
Residual	9	40416679100	4490742122		
Total	11	1.39720E+11			

a) Write the regression equation.
b) How much of the variation in home asking prices is accounted for by the model?
c) Explain in context what the coefficient of *Square Footage* means.
d) The owner of a construction firm, upon seeing this model, objects because the model says that the number of bathrooms has no effect on the price of the home. He says that when *he* adds another bathroom, it increases the value. Is it true that the number of bathrooms is unrelated to house price? (*Hint:* Do you think bigger houses have more bathrooms?)

T **16. More hill races** Here is the regression for the women's records for the same Scottish hill races we considered in Exercise 14:

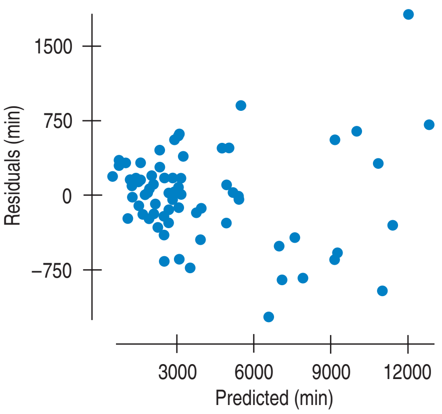
Dependent variable is Women's record
R-squared = 97.7% R-squared (adjusted) = 97.6%
s = 479.5 with 71 − 3 = 68 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-Ratio
Regression	658112727	2	329056364	1431
Residual	15634430	68	229918	

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	−554.015	101.7	−5.45	<0.0001
Distance	418.632	15.89	26.4	<0.0001
Climb	0.780568	0.0531	14.7	<0.0001

a) Compare the regression model for the women's records with that found for the men's records in Exercise 14.

Here's a scatterplot of the residuals for this regression:

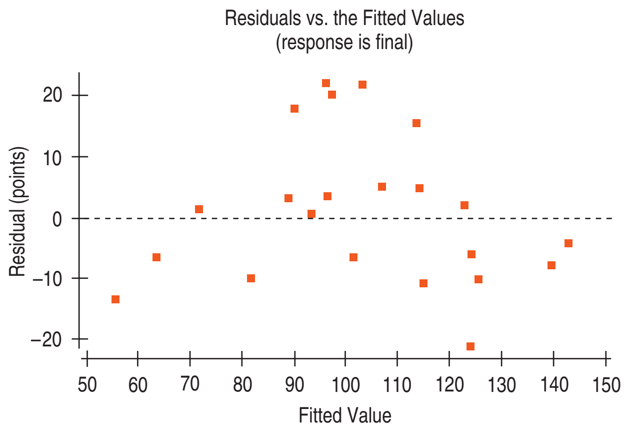


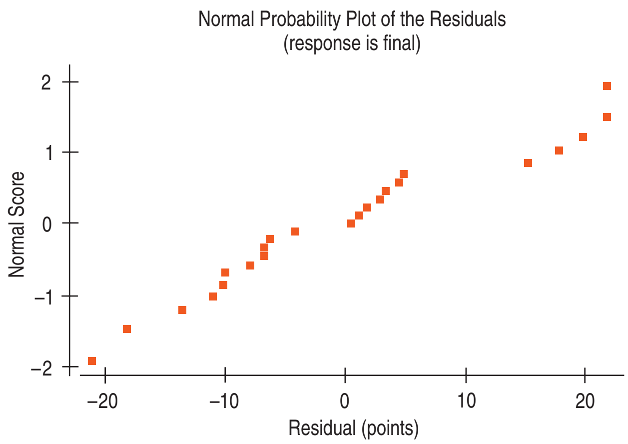
b) Discuss the residuals and what they say about the assumptions and conditions for this regression.

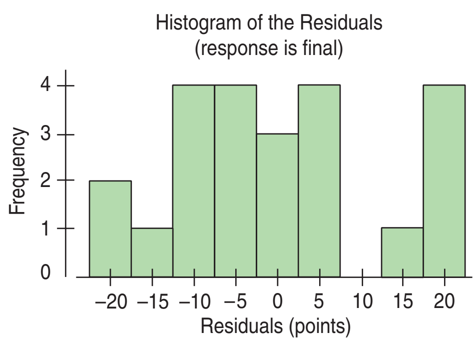
17. Predicting finals II Here are some diagnostic plots for the final exam data from Exercise 13. These were generated by a computer package and may look different from the plots generated by the packages you use. (In particular, note that the axes of the Normal probability plot are swapped relative to the plots we've made in the text. We only care about the pattern of this plot, so it shouldn't affect your interpretation.) Examine these

Copyright © 2014 Pearson Education, Inc.

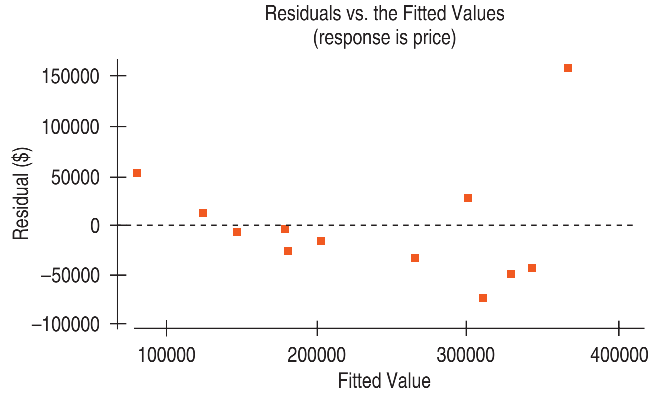
plots and discuss whether the assumptions and conditions for the multiple regression seem reasonable.

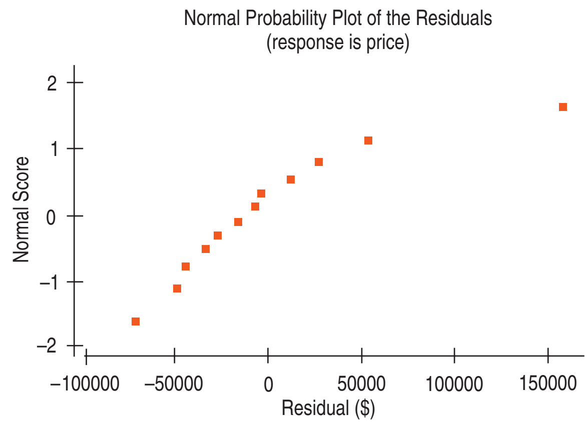


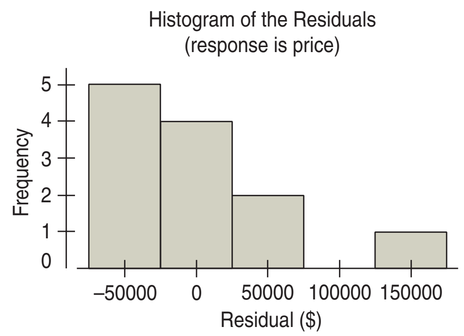




18. Home prices II Here are some diagnostic plots for the home prices data from Exercise 15. These were generated by a computer package and may look different from the plots generated by the packages you use. (In particular, note that the axes of the Normal probability plot are swapped relative to the plots we've made in the text. We only care about the pattern of this plot, so it shouldn't affect your interpretation.) Examine these plots and discuss whether the assumptions and conditions for the multiple regression seem reasonable.







19. Secretary performance The AFL-CIO has undertaken a study of 30 secretaries' yearly salaries (in thousands of dollars). The organization wants to predict salaries from several other variables.

The variables considered to be potential predictors of salary are

X1 = months of service
X2 = years of education
X3 = score on standardized test
X4 = words per minute (wpm) typing speed
X5 = ability to take dictation in words per minute

(continued)

Copyright © 2014 Pearson Education, Inc.

A multiple regression model with all five variables was run on a computer package, resulting in the following output:

Variable	Coefficient	Std. Error	t-Value
Intercept	9.788	0.377	25.960
X1	0.110	0.019	5.178
X2	0.053	0.038	1.369
X3	0.071	0.064	1.119
X4	0.004	0.307	0.013
X5	0.065	0.038	1.734

$s = 0.430$ $R^2 = 0.863$

Assume that the residual plots show no violations of the conditions for using a linear regression model.

a) What is the regression equation?
b) From this model, what is the predicted *Salary* (in thousands of dollars) of a secretary with 10 years (120 months) of experience, 9th grade education (9 years of education), a 50 on the standardized test, 60 wpm typing speed, and the ability to take 30 wpm dictation?
c) Test whether the coefficient for words per minute of typing speed (*X4*) is significantly different from zero at $\alpha = 0.05$.
d) How might this model be improved?
e) A correlation of *Age* with *Salary* finds $r = 0.682$, and the scatterplot shows a moderately strong positive linear association. However, if $X6 = Age$ is added to the multiple regression, the estimated coefficient of *Age* turns out to be $b_6 = -0.154$. Explain some possible causes for this apparent change of direction in the relationship between age and salary.

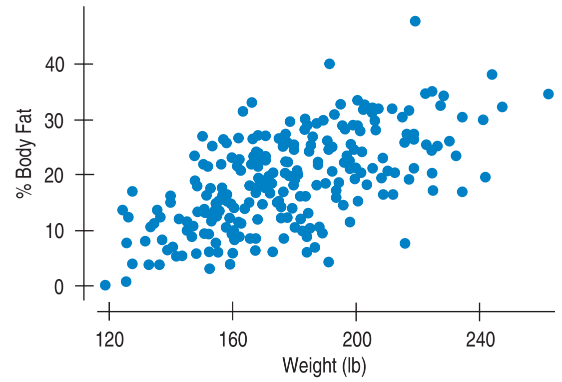
20. GPA and SATs A large section of Stat 101 was asked to fill out a survey on grade point average and SAT scores. A regression was run to find out how well Math and Verbal SAT scores could predict academic performance as measured by GPA. The regression was run on a computer package with the following output:

Response: GPA

	Coefficient	Std Error	t-Ratio	P-Value
Intercept	0.574968	0.253874	2.26	0.0249
SAT Verbal	0.001394	0.000519	2.69	0.0080
SAT Math	0.001978	0.000526	3.76	0.0002

a) What is the regression equation?
b) From this model, what is the predicted GPA of a student with an SAT Verbal score of 500 and an SAT Math score of 550?
c) What else would you want to know about this regression before writing a report about the relationship between SAT scores and grade point averages? Why would these be important to know?

T **21. Body fat, revisited** The data set on body fat contains 15 body measurements on 250 men from 22 to 81 years old. Is average *%Body Fat* related to *Weight*? Here's a scatterplot:



And here's the simple regression:

Dependent variable is Pct BF
R-squared = 38.1% R-squared (adjusted) = 37.9%
$s = 6.538$ with $250 - 2 = 248$ degrees of freedom

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	−14.6931	2.760	−5.32	<0.0001
Weight	0.18937	0.0153	12.4	<0.0001

a) Is the coefficient of *%Body Fat* on *Weight* statistically distinguishable from 0? (Perform a hypothesis test.)
b) What does the slope coefficient mean in this regression?

We saw before that the slopes of both *Waist* size and *Height* are statistically significant when entered into a multiple regression equation. What happens if we add *Weight* to that regression? Recall that we've already checked the assumptions and conditions for regression on *Waist* size and *Height* in the chapter. Here is the output from a regression on all three variables:

Dependent variable is Pct BF
R-squared = 72.5% R-squared (adjusted) = 72.2%
$s = 4.376$ with $250 - 4 = 246$ degrees of freedom

Source	Sum of Squares	df	Mean Square	F-Ratio
Regression	12418.7	3	4139.57	216
Residual	4710.11	246	19.1468	

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	−31.4830	11.54	−2.73	0.0068
Waist	2.31848	0.1820	12.7	<0.0001
Height	−0.224932	0.1583	−1.42	0.1567
Weight	−0.100572	0.0310	−3.25	0.0013

c) Interpret the slope for *Weight*. How can the coefficient for *Weight* in this model be negative when its coefficient was positive in the simple regression model?
d) What does the P-value for *Height* mean in this regression? (Perform the hypothesis test.)

Copyright © 2014 Pearson Education, Inc.

T **22. Breakfast cereals** We saw in Chapter 7 that the calorie content of a breakfast cereal is linearly associated with its sugar content. Is that the whole story? Here's the output of a regression model that regresses *Calories* for each serving on its *Protein(g), Fat(g), Fiber(g), Carbohydrate(g),* and *Sugars(g)* content.

Dependent variable is Calories
R-squared = 84.5%   R-squared (adjusted) = 83.4%
s = 7.947 with 77 − 6 = 71 degrees of freedom

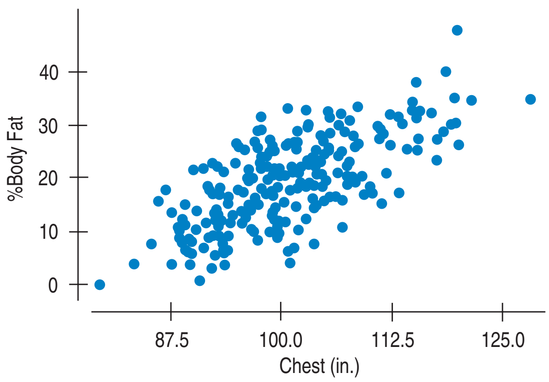
Source	Sum of Squares	df	Mean Square	F-Ratio
Regression	24367.5	5	4873.50	77.2
Residual	4484.45	71	63.1613	

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	20.2454	5.984	3.38	0.0012
Protein	5.69540	1.072	5.32	<0.0001
Fat	8.35958	1.033	8.09	<0.0001
Fiber	−1.02018	0.4835	−2.11	0.0384
Carbo	2.93570	0.2601	11.3	<0.0001
Sugars	3.31849	0.2501	13.3	<0.0001

Assuming that the conditions for multiple regression are met,

a) What is the regression equation?
b) Do you think this model would do a reasonably good job at predicting calories? Explain.
c) To check the conditions, what plots of the data might you want to examine?
d) What does the coefficient of *Fat* mean in this model?

23. Body fat again Chest size might be a good predictor of body fat. Here's a scatterplot of *%Body Fat* vs. *Chest Size*.



A regression of *%Body Fat* on *Chest Size* gives the following equation:

Dependent variable is Pct BF
R-squared = 49.1%   R-squared (adjusted) = 48.9%
s = 5.930 with 250 − 2 = 248 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	−52.7122	4.654	−11.3	<0.0001
Chest Size	0.712720	0.0461	15.5	<0.0001

a) Is the slope of *%Body Fat* on *Chest Size* statistically distinguishable from 0? (Perform a hypothesis test.)
b) What does the answer in part a mean about the relationship between *%Body Fat* and *Chest Size*?

We saw before that the slopes of both *Waist* size and *Height* are statistically significant when entered into a multiple regression equation. What happens if we add *Chest Size* to that regression? Here is the output from a regression on all three variables:

Dependent variable is Pct BF
R-squared = 72.2%   R-squared (adjusted) = 71.9%
s = 4.399 with 250 − 4 = 246 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-Ratio	P-Value
Regression	12368.9	3	4122.98	213	<0.0001
Residual	4759.87	246	19.3491		

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	2.07220	7.802	0.266	0.7908
Waist	2.19939	0.1675	13.1	<0.0001
Height	−0.561058	0.1094	−5.13	<0.0001
Chest Size	−0.233531	0.0832	−2.81	0.0054

c) Interpret the coefficient for *Chest Size.*
d) Would you consider removing any of the variables from this regression model? Why or why not?

T **24. Grades** The table below shows the five scores from an Introductory Statistics course. Find a model for predicting final exam score by trying all possible models with two predictor variables. Which model would you choose? Be sure to check the conditions for multiple regression.

Name	Final	Midterm 1	Midterm 2	Project	Homework
Timothy F.	117	82	30	10.5	61
Karen E.	183	96	68	11.3	72
Verena Z.	124	57	82	11.3	69
Jonathan A.	177	89	92	10.5	84
Elizabeth L.	169	88	86	10.6	84
Patrick M.	164	93	81	10	71
Julia E.	134	90	83	11.3	79
Thomas A.	98	83	21	11.2	51
Marshall K.	136	59	62	9.1	58
Justin E.	183	89	57	10.7	79
Alexandra E.	171	83	86	11.5	78
Christopher B.	173	95	75	8	77
Justin C.	164	81	66	10.7	66
Miguel A.	150	86	63	8	74
Brian J.	153	81	86	9.2	76
Gregory J.	149	81	87	9.2	75
Kristina G.	178	98	96	9.3	84
Timothy B.	75	50	27	10	20
Jason C.	159	91	83	10.6	71

(continued)

Copyright © 2014 Pearson Education, Inc.

Name	Final	Midterm 1	Midterm 2	Project	Home-work
Whitney E.	157	87	89	10.5	85
Alexis P.	158	90	91	11.3	68
Nicholas T.	171	95	82	10.5	68
Amandeep S.	173	91	37	10.6	54
Irena R.	165	93	81	9.3	82
Yvon T.	168	88	66	10.5	82
Sara M.	186	99	90	7.5	77
Annie P.	157	89	92	10.3	68
Benjamin S.	177	87	62	10	72
David W.`	170	92	66	11.5	78
Josef H.	78	62	43	9.1	56
Rebecca S.	191	93	87	11.2	80
Joshua D.	169	95	93	9.1	87
Ian M.	170	93	65	9.5	66
Katharine A.	172	92	98	10	77
Emily R.	168	91	95	10.7	83
Brian M.	179	92	80	11.5	82
Shad M.	148	61	58	10.5	65
Michael R.	103	55	65	10.3	51
Israel M.	144	76	88	9.2	67
Iris J.	155	63	62	7.5	67
Mark G.	141	89	66	8	72
Peter H.	138	91	42	11.5	66
Catherine R.M.	180	90	85	11.2	78
Christina M.	120	75	62	9.1	72
Enrique J.	86	75	46	10.3	72
Sarah K.	151	91	65	9.3	77
Thomas J.	149	84	70	8	70
Sonya P.	163	94	92	10.5	81
Michael B.	153	93	78	10.3	72
Wesley M.	172	91	58	10.5	66
Mark R.	165	91	61	10.5	79
Adam J.	155	89	86	9.1	62
Jared A.	181	98	92	11.2	83
Michael T.	172	96	51	9.1	83
Kathryn D.	177	95	95	10	87
Nicole M.	189	98	89	7.5	77
Wayne E.	161	89	79	9.5	44
Elizabeth S.	146	93	89	10.7	73
John R.	147	74	64	9.1	72
Valentin A.	160	97	96	9.1	80
David T.O.	159	94	90	10.6	88
Marc I.	101	81	89	9.5	62
Samuel E.	154	94	85	10.5	76
Brooke S.	183	92	90	9.5	86

T **25. Fifty states** Here is a data set on various measures of the 50 United States. The *Murder* rate is per 100,000, *HS Graduation* rate is in %, *Income* is per capita income in dollars, *Illiteracy* rate is per 1000, and *Life Expectancy* is in years. Find a regression model for *Life Expectancy* with three predictor variables by trying all four of the possible models.

a) Which model appears to do the best?
b) Would you leave all three predictors in this model?
c) Does this model mean that by changing the levels of the predictors in this equation, we could affect life expectancy in that state? Explain.
d) Be sure to check the conditions for multiple regression. What do you conclude?

State Name	Murder	HS Grad	Income	Illiteracy	Life Exp
Alabama	15.1	41.3	3624	2.1	69.05
Alaska	11.3	66.7	6315	1.5	69.31
Arizona	7.8	58.1	4530	1.8	70.55
Arkansas	10.1	39.9	3378	1.9	70.66
California	10.3	62.6	5114	1.1	71.71
Colorado	6.8	63.9	4884	0.7	72.06
Connecticut	3.1	56	5348	1.1	72.48
Delaware	6.2	54.6	4809	0.9	70.06
Florida	10.7	52.6	4815	1.3	70.66
Georgia	13.9	40.6	4091	2	68.54
Hawaii	6.2	61.9	4963	1.9	73.6
Idaho	5.3	59.5	4119	0.6	71.87
Illinois	10.3	52.6	5107	0.9	70.14
Indiana	7.1	52.9	4458	0.7	70.88
Iowa	2.3	59	4628	0.5	72.56
Kansas	4.5	59.9	4669	0.6	72.58
Kentucky	10.6	38.5	3712	1.6	70.1
Louisiana	13.2	42.2	3545	2.8	68.76
Maine	2.7	54.7	3694	0.7	70.39
Maryland	8.5	52.3	5299	0.9	70.22
Massachusetts	3.3	58.5	4755	1.1	71.83
Michigan	11.1	52.8	4751	0.9	70.63
Minnesota	2.3	57.6	4675	0.6	72.96
Mississippi	12.5	41	3098	2.4	68.09
Missouri	9.3	48.8	4254	0.8	70.69
Montana	5	59.2	4347	0.6	70.56
Nebraska	2.9	59.3	4508	0.6	72.6
Nevada	11.5	65.2	5149	0.5	69.03
New Hampshire	3.3	57.6	4281	0.7	71.23
New Jersey	5.2	52.5	5237	1.1	70.93
New Mexico	9.7	55.2	3601	2.2	70.32
New York	10.9	52.7	4903	1.4	70.55
North Carolina	11.1	38.5	3875	1.8	69.21
North Dakota	1.4	50.3	5087	0.8	72.78
Ohio	7.4	53.2	4561	0.8	70.82
Oklahoma	6.4	51.6	3983	1.1	71.42
Oregon	4.2	60	4660	0.6	72.13
Pennsylvania	6.1	50.2	4449	1	70.43
Rhode Island	2.4	46.4	4558	1.3	71.9
South Carolina	11.6	37.8	3635	2.3	67.96
South Dakota	1.7	53.3	4167	0.5	72.08

Copyright © 2014 Pearson Education, Inc.

State Name	Murder	HS Grad	Income	Illiteracy	Life Exp
Tennessee	11	41.8	3821	1.7	70.11
Texas	12.2	47.4	4188	2.2	70.9
Utah	4.5	67.3	4022	0.6	72.9
Vermont	5.5	57.1	3907	0.6	71.64
Virginia	9.5	47.8	4701	1.4	70.08
Washington	4.3	63.5	4864	0.6	71.72
West Virginia	6.7	41.6	3617	1.4	69.48
Wisconsin	3	54.5	4468	0.7	72.48
Wyoming	6.9	62.9	4566	0.6	70.29

26. Breakfast cereals again We saw in Chapter 7 that the calorie count of a breakfast cereal is linearly associated with its sugar content. Can we predict the calories of a serving from its vitamin and mineral content? Here's a multiple regression model of *Calories* per serving on its *Sodium (mg), Potassium (mg),* and *Sugars (g)*:

Dependent variable is Calories
R-squared = 38.4% R-squared (adjusted) = 35.9%
s = 15.60 with 77 − 4 = 73 degrees of freedom

Source	Sum of Squares	df	Square Mean	F-Ratio	P-Value
Regression	11091.8	3	3697.28	15.2	<0.0001
Residual	17760.1	73	243.289		

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	83.0469	5.198	16.0	<0.0001
Sodium	0.05721	0.0215	2.67	0.0094
Potass	−0.01933	0.0251	−0.769	0.4441
Sugars	2.38757	0.4066	5.87	<0.0001

Assuming that the conditions for multiple regression are met,

a) What is the regression equation?
b) Do you think this model would do a reasonably good job at predicting calories? Explain.
c) Would you consider removing any of these predictor variables from the model? Why or why not?
d) To check the conditions, what plots of the data might you want to examine?

27. Burger King 2010 revisited Recall the Burger King menu data from Chapter 7. BK's nutrition sheet lists many variables. Here's a multiple regression to predict calories for Burger King foods from *Protein* content *(g), Total Fat (g), Carbohydrate (g),* and *Sodium (mg)* per serving:

Dependent variable is Calories
R-squared = 99.8% R-squared (adjusted) = 99.8%
s = 8.51 with 111 − 5 = 106 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-Ratio
Regression	4750462	4	1187616	16394
Residual	7678.64	106	72.4400	

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	−5.826	2.568	−2.27	0.0253
Protein	3.8814	0.0991	39.1	<0.0001
Total fat	9.2080	0.0893	103	<0.0001
Carbs	3.9016	0.0457	85.3	<0.0001
Na/Serv.	1.2873	0.4172	3.09	0.0026

a) Do you think this model would do a good job of predicting calories for a new BK menu item? Why or why not?
b) The mean of *Calories* is 453.9 with a standard deviation of 234.6. Discuss what the value of *s* in the regression means about how well the model fits the data.
c) Does the R^2 value of 99.8% mean that the residuals are all actually equal to zero? How can you tell from this table?

Just Checking ANSWERS

1. 77.9% of the variation in *Maximum Wind Speed* can be accounted for by multiple regression on *Central Pressure* and *Year*.
2. In any given year, hurricanes with a *Central Pressure* that is 1 mb lower can be expected to have, on average, winds that are 0.933 kn faster.
3. First, the researcher is trying to prove his null hypothesis for this coefficient and, as we know, statistical inference won't permit that. Beyond that problem, we can't even be sure we understand the relationship of *Wind Speed* to *Year* from this analysis. For example, both *Central Pressure* and *Wind Speed* might be changing over time, but their relationship might well stay the same during any given year.

Copyright © 2014 Pearson Education, Inc.