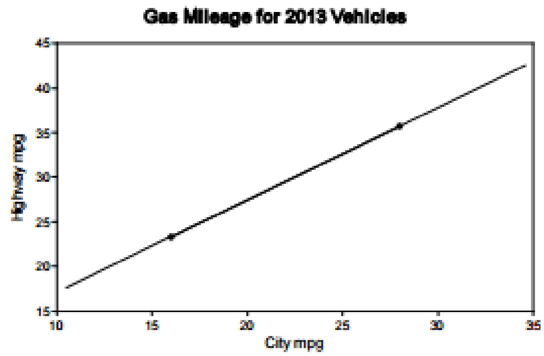
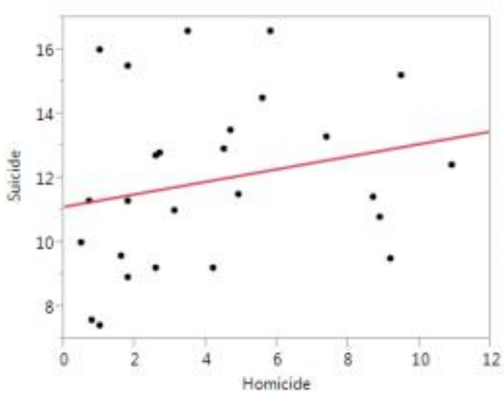
Chapter 5 – Regression



5.1 (a) The slope is 1.033. On average, highway mileage increases by 1.033 mpg for each additional 1 mpg change in city mileage. **(b)** The intercept is 6.785 mpg. This is the highway mileage for a nonexistent car that gets 0 mpg in the city. Although this interpretation is valid, such a prediction would be invalid, since 0 is outside the range of the data (this is extrapolation, which will be addressed later in the chapter). **(c)** For a car that gets 16 mpg in the city, we predict highway mileage to be 6.785 + (1.033)(16) = 23.31 mpg. For a car that gets 28 mpg in the city, we predict highway mileage to be 6.785 + (1.033)(28) = 35.71 mpg. **(d)** The regression line passes through all the points of prediction. The plot was created by drawing a line through the two points (16, 23.31) and (28, 35.71), corresponding to the city mileages and predicted highway mileages for the two cars described in (c).

5.4 (a) $\bar{x}$ = 30.280, s_x = 0.4296, $\bar{y}$ = 2.4557, s_y = 0.1579, and r = −0.8914. $b = r\frac{s_y}{s_x}$ = $(-0.8914)\frac{0.1579}{0.4296}$ = −0.3276, and $a = \bar{y} - b\bar{x}$ = 2.4557 – (−0.3276)(30.280) = 12.3754. The equation is $\widehat{Coral\ growth}$ = 12.3754 − 0.3276(Celsius Temperature). **(b)** Software agrees with these values to three decimal places, since we rounded to the fourth decimal place (where values are rounded will affect these results). **(c)** The slope is −0.3276. This means that every increase of one degree Celsius means about 0.3276 fewer mean millimeters of coral growth per year.



Linear Fit

Suicide = 11.125 + 0.1953551*Homicide

Summary of Fit

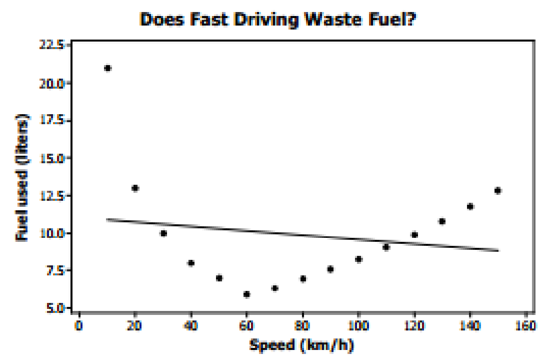
RSquare	0.052692
RSquare Adj	0.013221
Root Mean Square Error	2.657216
Mean of Response	11.95
Observations (or Sum Wgts)	26

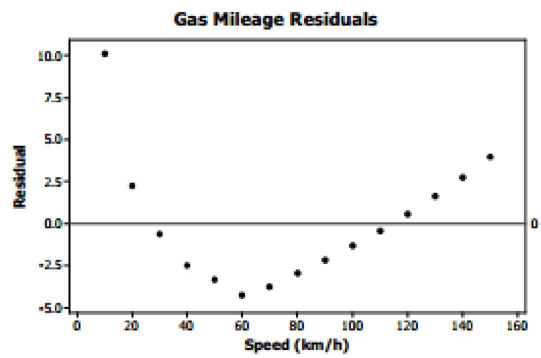
5.6 The farther r is from 0 (in either direction), the stronger the linear relationship is between two variables. In Exercise 4.30, the relationship between SRD and DMS is very strongly linear, and a regression line should enable relatively more accurate prediction than a regression line for the golfers' scores.

5.8 (a) The residuals are computed in the table using $\hat{y}$ = 12.3754 – 0.3276*x*, as computed in Exercise 5.4. **(b)** They sum to zero, except for rounding error. **(c)** From software, the correlation between *x* and $y - \hat{y}$ is 0.000025, which is zero except for rounding.

x	y	$\hat{y}$	$y - \hat{y}$
29.68	2.63	2.652	–0.022
29.87	2.58	2.590	–0.010
30.16	2.60	2.495	0.105
30.22	2.48	2.475	0.005
30.48	2.26	2.390	–0.130
30.65	2.38	2.335	0.045
30.90	2.26	2.253	0.007
		0	

5.9 (a) Plot is provided following, left. **(b)** No; the pattern is curved, so linear regression is not appropriate for prediction. **(c)** For *x* = 10, we estimate $\hat{y}$ = 11.058 – 0.01466(10) = 10.91, so the residual is 21.00 – 10.91 = 10.09. The sum of the residuals is –0.01. **(d)** The first two and last four residuals are positive, and those in the middle are negative. Plot following, right.





5.17 Possible lurking variables include the IQ and socioeconomic status of the mother, as well as the mother's other habits (drinking, diet, etc.). These variables are associated with smoking in various ways, and are also predictive of a child's IQ.

Note: *There may be an indirect cause-and-effect relationship at work here: some studies have found evidence that over time, smokers lose IQ points, perhaps due to brain damage caused by toxins from the smoke. So, perhaps smoking mothers gradually grow less smart and are less able to nurture their children's cognitive development.*

5.20 (b) 7.5. The regression line seems to pass through the point (110, 7.5).

5.21 (b) 0.2. Consider two points on the regression line—say (90,4) and (130,11). The slope of the line segment connecting these points is $\frac{11-4}{130-90}$ = 7/40 = 0.175.

5.22 (c) −3

5.23 (a) $y = 1000 + 100x$

5.24 (b) will be less than 0. As the number of packs increases, average age at death decreases. Correlation is negative, and so is the slope of the regression line.

5.25 (c) 16 cubic feet

5.26 (a) 405 cubic feet

5.27 (a) The slope of the line is positive.

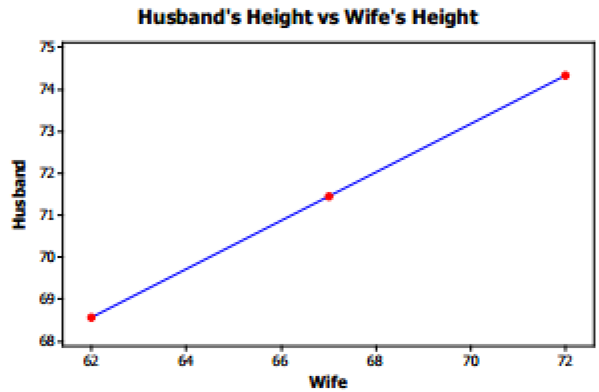
5.28 (c) prediction of gas used from degree-days will be quite accurate.

5.27 (a) $\hat{y} = 24.2 + 6.0x$

5.31 (a) Since the slope is 3721.02, the least-squares regression line says that increasing the size of a diamond by 1 carat increases its price by 3721.02 Singapore dollars, on average. **(b)** A diamond of size 0 carats would have a predicted price of 259.63 Singapore dollars. This is probably an extrapolation, since the data set on which the line was constructed almost certainly had no rings with diamonds of size 0 carats. However, if the number is meaningful (dubious), then it refers to the cost of the gold content and other materials in the ring.

5.32 (a) The regression equation is $\hat{y} = -0.126 + 0.0608x$. Each increase of one unit in social distress increases brain activity by about 0.0608 units. For $x = 2.0$, this formula gives $\hat{y} = -0.0044$. (A student who uses the more precise coefficient estimates listed under "Coef" in the Minitab output might report the predicted brain activity as −0.0045.) **(b)** This is given in the Minitab output as "R-Sq": 77.1%. The linear relationship explains 77.1% of the variation in brain activity. **(c)** Knowing that $r^2 = 0.771$, we find $r = \sqrt{r^2} = 0.878$; the sign is positive because it has the same sign as the slope coefficient.

5.34 Since we wish to regress husbands' heights on wives' heights, the women's heights will be the x-values, and the men's heights will be the y-values. **(a)** $b = r\, s_y / s_x$ = (0.5)(3.1/2.7) = 0.574, and $a = \bar{y} - b\bar{x}$ = 69.9 – (0.574)(64.3) = 32.99 inches. The regression equation is $\hat{y}$ = 32.99 + 0.574x. For every inch of a wife's height, her husband is about 0.574 inches taller. **(b)** If a wife is 67 inches tall, we predict her husband to have height $\hat{y}$ = 32.99 + (0.574)(67) = 71.448 inches. The plot, with this pair identified, is provided. **(c)** We don't expect this prediction to be very accurate because the heights of men having wives 67 inches tall varies a lot. Also, $r^2 = (0.5)^2 = 0.25$, so the linear regression explains only 25% of the variation in men's heights.



5.35 **(a)** $b = r\, s_y / s_x = (0.5)\left(\frac{8}{40}\right)$ = 0.1, and $a = \bar{y} - b\bar{x}$ = 75 – (0.1)(280) = 47. The regression equation is $\hat{y}$ = 47 + 0.1x. Each point of pre-exam total score means an additional 0.1 points on the final exam, on average. **(b)** Julie's pre-final exam total was 300, so we would predict a final exam score of $\hat{y}$ = 47 + (0.1)(300) = 77. **(c)** Julie is right; with a correlation of $r = 0.5$, $r^2 = (0.5)^2 = 0.25$, so the regression line accounts for only 25% of the variability in student final exam scores. That is, the regression line doesn't predict final exam scores very well. Julie's score could, indeed, be much higher or lower than the predicted 77. Since she is making this argument, one might guess that her score was, in fact, higher. Julie should visit the Dean.

5.36 $r = \sqrt{0.16}$ = 0.40 (high attendance goes with high grades, so the correlation must be positive).