Chapter 10 Solutions

10.1. The given model was $\mu_y = 40.5 - 2.5x$, with standard deviation $\sigma = 2.0$. **(a)** The slope is -2.5. **(b)** When x increases by 1, μ_y decreases by 2.5. (Or equivalently, if x increases by 2, μ_y decreases by 5, etc.) **(c)** When $x = 10$, $\mu_y = 40.5 - 2.5(10) = 15.5$. **(d)** Approximately 95% of observed responses would fall in the interval $\mu_y \pm 2\sigma = 15.5 \pm 2(2.0) = 15.5 \pm 4.0 = 11.5$ to 19.5.

10.2. Example 10.4 gave the estimated regression equation as $\widehat{\text{MPG}} = -7.796 + 7.874\,\text{LOGMPH}$, with $s = 0.9995$. In the text following that example, the parameter estimates were rounded to two decimal places: $\widehat{\text{MPG}} = -7.80 + 7.87\ \text{LOGMPH}$, with $s = 1.00$. **(a)** If the car travels at 35 mph, then $\text{LOGMPH} = \ln 35 \doteq 3.5553$, so we estimate $\widehat{\text{MPG}} \doteq 20.1988$ mpg (or 20.18 mpg, using the rounded estimates). **(b)** The residual is $21.0 - \widehat{\text{MPG}} \doteq 0.8012$ mpg (or 0.82, using the rounded estimates). **(c)** Because this regression line was based on speeds between about 12 and 53 mph, estimates near or outside those boundaries would not be very reliable; they grow less reliable the more we stray above 53 mph. Even the estimated mpg for 45 mph would be subject to lots of uncertainty because the points at the high end of the scatterplot exhibited more spread than those at the low end.

Note: *Some students might mistakenly use the common (base-10) logarithm instead of the natural logarithm; if they do so, they will find LOGMPH $\doteq$ 1.5441 and $\widehat{MPG} \doteq$ 4.36 mpg. Hopefully, the large residual (16.6 mpg) in part (b) would help them notice their mistake; in general, we expect residuals to fall in the range $\pm 3s$, if they follow a Normal distribution.*

10.3. Example 10.6 gives the confidence interval 7.16 to 8.58 for the slope β_1. Recall that slope is the change in y (i.e., $\widehat{\text{MPG}}$) when x (i.e., LOGMPH) changes by $+1$. **(a)** If LOGMPH increases by 1, we expect $\widehat{\text{MPG}}$ to change by β_1, so the 95% confidence interval for the change is (an increase of) 7.16 to 8.58 mpg. **(b)** If LOGMPH decreases by 1, we expect $\widehat{\text{MPG}}$ to change by $-\beta_1$, so the 95% confidence interval for the change is -7.16 to -8.58 mpg—that is, a decrease of 7.16 to 8.58 mpg. **(c)** If LOGMPH increases by 0.5, we expect $\widehat{\text{MPG}}$ to change by $0.5\beta_1$, so the 95% confidence interval for the change is (an increase of) 3.58 to 4.29 mpg.

10.4. Example 10.10 gives the 95% prediction interval $\hat{y} \pm t^* \text{SE}_{\hat{y}}$ as 17.0 to 21.0, or 19.0 ± 2.0. For df = 58, the critical value t^* is very close to 2, so $\text{SE}_{\hat{y}} \doteq 1.0$.

When $x = 40$ mph, $\text{SE}_{\hat{y}}$ would be larger because predictions are less accurate for values of x near the extremes (relative to the range of speeds used to determine the regression formula, which was approximately 12 to 53 mph).

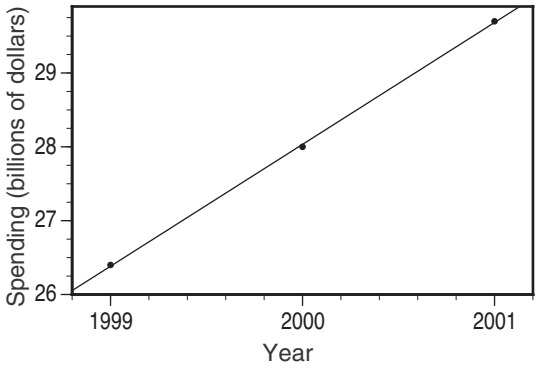
Note: *Working from the original data set, $SE_{\hat{y}}$ equals 1.01 when $x = 30$ mph, and 1.02 when $x = 40$ mph.*

10.5. **(a)** The plot suggests a linear increase. **(b)** The regression equation is $\hat{y} = -3271.9667 + 1.65x$. **(c)** The fitted values and residuals are given in the table below. Squaring the residuals and summing gives $0.001\overline{6}$, so the standard error is:

$$s = \sqrt{\frac{0.001\overline{6}}{n-2}} = \sqrt{0.001\overline{6}} \doteq 0.04082$$

(d) Given x (the year), spending comes from a $N(\mu_y, \sigma)$ distribution, where $\mu_y = \beta_0 + \beta_1 x$. The estimates of β_0, β_1, and σ are $b_0 \doteq -3271.9667$, $b_1 \doteq 1.65$, and $s \doteq 0.04082$. **(e)** We first note that $\bar{x} = 2000$ and $\sum(x_i - \bar{x})^2 = 2$, so $\text{SE}_{b_1} = s/\sqrt{2} \doteq 0.02887$. We have $\text{df} = n - 2 = 1$, so $t^* = 12.71$, and the 95% confidence interval for β_1 is $1.65 \pm t^*\text{SE}_{b_1} \doteq 1.283$ to 2.017. This gives the rate of increase of R&D spending: between 1.283 and 2.017 billion dollars per year.

Year	Spending ($billions)	Fitted values	Residuals
1999	26.4	$26.38\overline{3}$	$0.01\overline{6}$
2000	28.0	$28.0\overline{3}$	$-0.0\overline{3}$
2001	29.7	$29.68\overline{3}$	$0.01\overline{6}$



10.6. **(a)** The variables x and y are reversed: Slope gives the change in y for a change in x. **(b)** The population regression line has intercept β_0 and slope β_1 (not b_0 and b_1). **(c)** The estimate $\hat{\mu}_y = b_0 + b_1 x^*$ is more accurate when x^* is close to $\bar{x}$, so the width of the confidence interval grows with $(x^* - \bar{x})^2$.

10.7. **(a)** The parameters are β_0, β_1, and σ; b_0, b_1, and s are the *estimates* of those parameters. **(b)** H_0 should refer to β_1 (the population slope) rather than b_1 (the estimated slope). **(c)** The confidence interval will be narrower than the prediction interval because the confidence interval accounts only for the uncertainty in our estimate of the mean response, while the prediction interval must also account for the random error of an individual response.

10.8. The table below gives two sets of answers: those found with critical values from Table D, and those found with software. In each case, the margin of error is $t^*\text{SE}_{b_1} = 6.31t^*$, with $\text{df} = n - 2$.

	df	b_1	t^*	Interval	t^*	Interval
(a)	23	12.1	2.069	−0.9554 to 25.1554	2.0687	−0.9532 to 25.1532
(b)	23	6.1	2.069	−6.9554 to 19.1554	2.0687	−6.9532 to 19.1532
(c)	98	12.1	1.990*	−0.4569 to 24.6569	1.9845	−0.4220 to 24.6220

*Note that for (c), if we use Table D, we take $\text{df} = 80$.

10.9. The test statistic is $t = b_1/\text{SE}_{b_1} = b_1/6.31$, with df $= n - 2$. All three tests fail to produce significant evidence against H_0, although we are close in (a) and (c). This is consistent with the confidence intervals from the previous exercise.

	df	b_1	t	P (Table D)	P (software)
(a)	23	12.1	1.92	$0.05 < P < 0.10$	0.0677
(b)	23	6.1	0.97	$0.30 < P < 0.40$	0.3437
(c)	98	12.1	1.92	$0.05 < P < 0.10$*	0.0581

*Note that for (c), if we use Table D, we take df $= 80$.

10.10. **(a)** The plot (below, left) shows a strong linear relationship with no striking outliers. **(b)** The regression line (shown on the plot) is $\hat{y} = 1059 + 1.3930x$. **(c)** In the plot (below, right), it appears that for large x values, many residuals are negative. **(d)** A stemplot (shown) or histogram suggests a slight left skew. **(e)** To test for a relationship, we test H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$ (or equivalently, use ρ in place of β_1). **(f)** The test statistic and P-value are given in the Minitab output below: $t \doteq 15.50$, $P < 0.0001$. We have strong evidence of a non-zero slope.

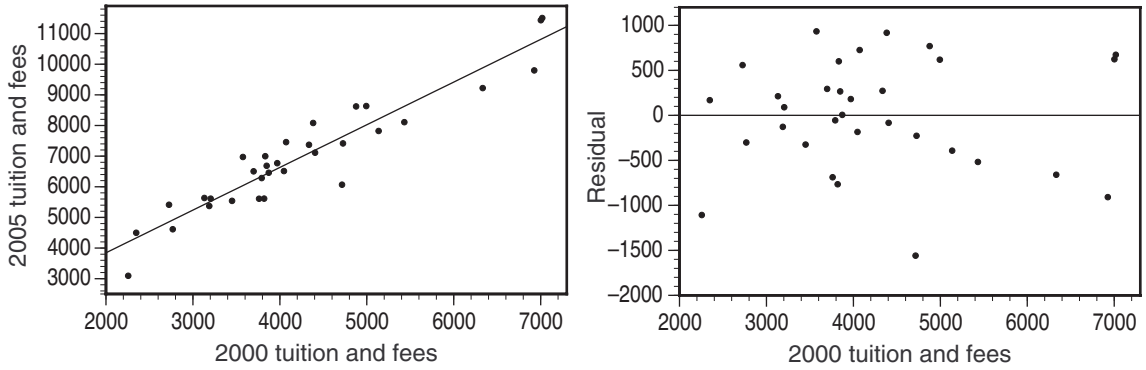
−1	5
−1	
−1	1
−0	9
−0	766
−0	5
−0	3332
−0	1100
0	0011
0	2222
0	55
0	66677
0	99

Minitab output

```
The regression equation is y2005 = 1059 + 1.39 y2000

Predictor       Coef       Stdev    t-ratio        p
Constant      1059.0       396.8       2.67    0.012
y2000        1.39301     0.08988      15.50    0.000

s = 626.6       R-sq = 88.9%     R-sq(adj) = 88.5%
```



10.11. **(a)** From the Minitab output above, we have $\text{SE}_{b_1} \doteq 0.08988$. With df $= 30$, $t^* = 2.042$, so the 95% confidence interval for β_1 is $1.3930 \pm t^*\text{SE}_{b_1} \doteq 1.2095$ to 1.5765. This slope means that a \$1 difference in tuition in 2000 changes 2005 tuition by between \$1.21 and \$1.58, so we estimate that tuition increased by 21% to 58%. **(b)** When $x = 5000$, the estimated 2005 tuition is $\hat{y} = 1059 + 1.3930(5000) \doteq \$8024$. **(c)** The Minitab output below gives the interval \$6717 to \$9331. If your software will not report this, here are the details of how to produce this interval (not for the faint of heart): We note that $s_x \doteq \$1252.17$, so $\sum(x_i - \bar{x})^2 = s_x^2(n-1) \doteq 48{,}605{,}492$. With $x^* = \$5000$, $\bar{x} \doteq \$4238.69$, $s = \$626.6$, and

$n = 32$, we have:

$$\text{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \doteq \$640$$

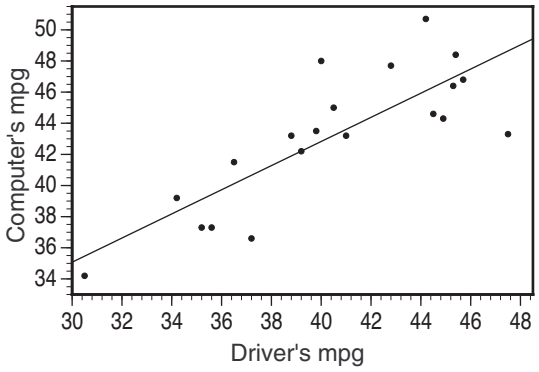
With $t^* = 2.042$, this gives the prediction interval $\$8024 \pm (2.042)(\$640) \doteq \$8024 \pm \$1307 = \$6717$ to $\$9331$.

Note: *The value "Stdev.Fit" given by Minitab (below) as 130 is* $SE_{\hat{\mu}}$. *Observe that* $SE_{\hat{y}}^2 = s^2 + SE_{\hat{\mu}}^2$*—reminiscent of the Pythagorean theorem—so that* $SE_{\hat{y}}$ *can be computed rather easily if* s *and* $SE_{\hat{\mu}}$ *are both known.*

Minitab output

```
    Fit  Stdev.Fit       95.0% C.I.           95.0% P.I.
   8024        130   (    7758,    8290)  (    6717,    9331)
```

10.12. **(a)** The scatterplot shows a fairly strong positive linear association, with no extreme outliers, so regression seems to be appropriate. **(b)** The regression equation (shown on the scatterplot) is $\hat{y} = 11.81 + 0.7754x$. **(c)** Student summaries will vary. The mpg values are certainly similar, but one notable difference is that all but three of the computer values are higher than the driver's values, and the mean computer mpg is about 2.7 mpg higher than the mean driver mpg. Additionally, the slope of the regression line is about 0.78, meaning that (on average) a 1 mpg change in the driver's value corresponds to a 0.78 mpg for the computer. The intercept, however, is about 11.8 mpg, suggesting that the computer's value is generally higher when the driver's value is small.



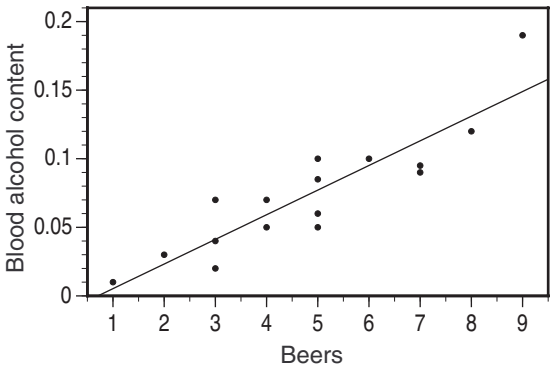
Minitab output

```
The regression equation is Computer = 11.8 + 0.775 Driver

Predictor       Coef       Stdev    t-ratio        p
Constant      11.812       5.432       2.17    0.043
Driver        0.7754      0.1335       5.81    0.000

s = 2.676       R-sq = 65.2%     R-sq(adj) = 63.3%
```

10.13. **(a)** The regression equation is $\hat{y} = -0.0127 + 0.0180x$, and $r^2 \doteq 80.0\%$. Not surprisingly, we find that BAC increases as beer consumption increases; the relationship is quite strong, with beer consumption explaining 80% of the variation in BAC. **(b)** To test H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 > 0$, we find $t = 7.48$ and $P < 0.0001$. There is very strong evidence that drinking more beers increases BAC. **(c)** The predicted mean BAC for $x = 5$ beers is 0.07712; the 90% prediction interval is 0.040 to 0.114. Steve might be safe, but cannot be sure that his BAC will be below 0.08.



Note: *For (c), we use a prediction interval (rather than a confidence interval) because we are interested in a range of values for an* individual *BAC after 5 beers, rather than the* mean *BAC. The first printing of the text asked for a* confidence *interval, and the answer given in the back of the first printing was the 90% confidence interval: 0.06808 to 0.08616.*

Minitab output

```
The regression equation is BAC = - 0.0127 + 0.0180 Beers

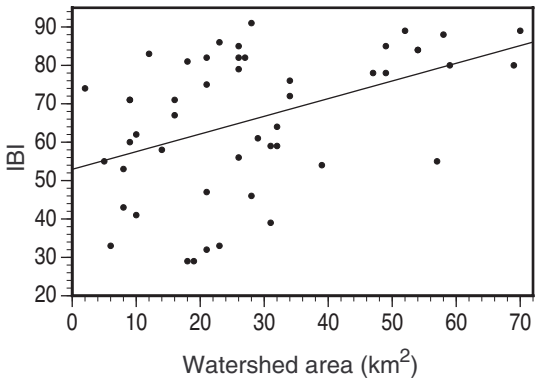
Predictor       Coef       Stdev    t-ratio        p
Constant    -0.01270     0.01264      -1.00    0.332
Beers       0.017964    0.002402       7.48    0.000

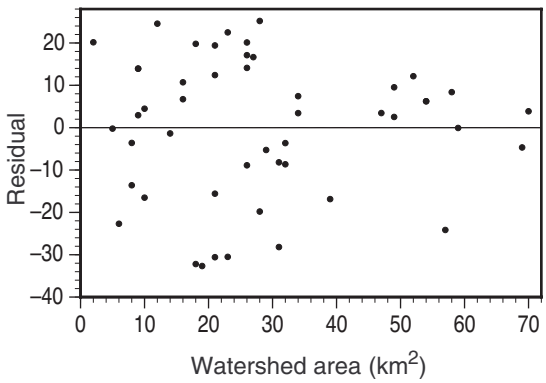
s = 0.02044     R-sq = 80.0%     R-sq(adj) = 78.6%

     Fit  Stdev.Fit       90.0% C.I.           90.0% P.I.
 0.07712    0.00513   ( 0.06808, 0.08616)  ( 0.03999, 0.11425)
```

10.14. **(a)** Stemplots are shown on the right. x (watershed area) is right-skewed; $\bar{x} \doteq 28.2857$ km^2, $s_x \doteq 17.7142$ km^2. y (IBI) is left-skewed; $\bar{y} \doteq 65.9388$, $s_y \doteq 18.2796$. **(b)** The scatterplot (next page, left) shows a weak positive association, with more scatter in y for small x. **(c)** $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, ..., 49$; ϵ_i are independent $N(0, \sigma)$ variables. **(d)** The hypotheses are H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$. **(e)** See the Minitab output on the next page. The regression equation is $\widehat{\text{IBI}} = 52.92 + 0.4602\,\text{Area}$, and the estimated standard deviation is $s \doteq 16.53$. For testing the hypotheses in (d), $t = 3.42$ and $P = 0.001$. **(f)** The residual plot (next page, right) again shows that there is more variation for small x. **(g)** As we can see from a stemplot and/or a Normal quantile plot (both on the next page), the residuals are somewhat left-skewed but otherwise seem reasonably close to Normal. **(h)** Student opinions may vary. The two apparent deviations from the model are (i) a possible change in standard deviation as x changes and (ii) possible non-Normality of error terms.

Area		IBI	
0	2	2	99
0	5688999	3	233
1	0024	3	9
1	66889	4	13
2	111133	4	67
2	66667889	5	34
3	112244	5	556899
3	9	6	0124
4		6	7
4	799	7	11124
5	244	7	56889
5	789	8	001222344
6		8	556899
6	9	9	1
7	0		





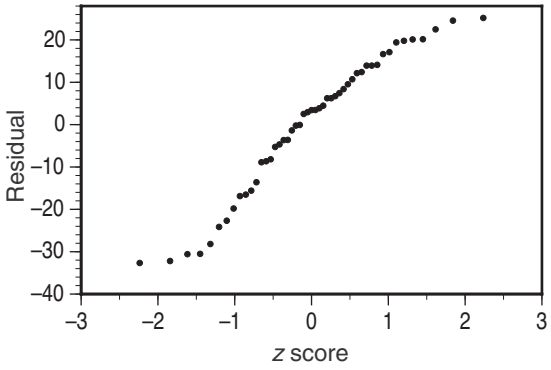
Minitab output

```
The regression equation is IBI = 52.9 + 0.460 Area

Predictor       Coef       Stdev    t-ratio        p
Constant      52.923       4.484      11.80    0.000
Area          0.4602      0.1347       3.42    0.001

s = 16.53       R-sq = 19.9%     R-sq(adj) = 18.2%
```

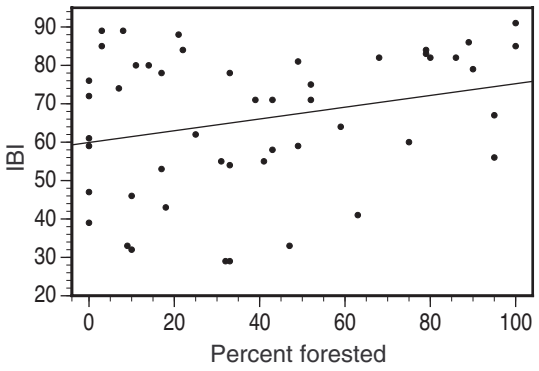
−3	2200
−2	8
−2	42
−1	9665
−1	3
−0	8885
−0	433100
0	223334
0	666789
1	022334
1	6799
2	0024
2	5

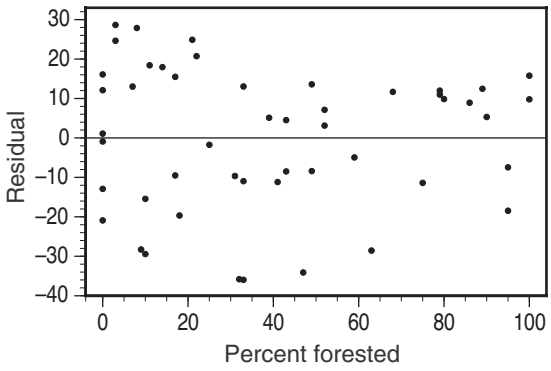


10.15. **(a)** The stemplot of percent forested is shown on the right; see the solution to the previous exercise for the stemplot of IBI. x (percent forested) is right-skewed; $\bar{x} = 39.3878\%$, $s_x = 32.2043\%$. y (IBI) is left-skewed; $\bar{y} = 65.9388$, $s_y = 18.2796$. **(b)** The scatterplot (next page, left) shows a weak positive association, with more scatter in y for small x. **(c)** $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, ..., 49$; ϵ_i are independent $N(0, \sigma)$ variables. **(d)** The hypotheses are H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$. **(e)** See the Minitab output on the next page. The regression equation is $\widehat{\text{IBI}} = 59.91 + 0.1531$ Forest, and the estimated standard deviation is $s \doteq 17.79$. For testing the hypotheses in (d), $t = 1.92$ and $P = 0.061$. **(f)** The residual plot (next page, right) shows a slight curve—the residuals seem to be (very) slightly lower in the middle and higher on the ends. **(g)** As we can see from a stemplot and/or a Normal quantile plot (both on the next page), the residuals are left-skewed. **(h)** Student opinions may vary. The three apparent deviations from the model are (i) a possible change in standard deviation as x changes, (ii) possible curvature of residuals, and (iii) possible non-Normality of error terms.

Percent forested

0	00000033789
1	0014778
2	125
3	123339
4	133799
5	229
6	38
7	599
8	069
9	055
10	00





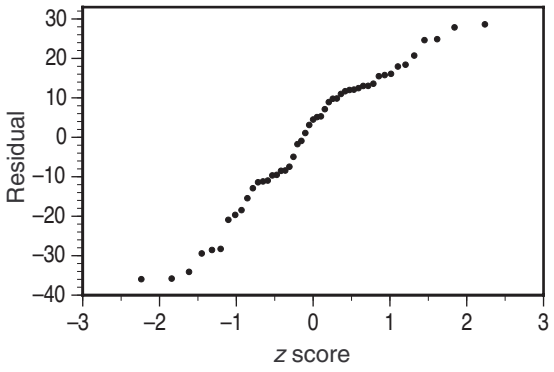
Minitab output

```
The regression equation is IBI = 59.9 + 0.153 Forest

Predictor       Coef       Stdev    t-ratio        p
Constant      59.907       4.040      14.83    0.000
Forest       0.15313     0.07972       1.92    0.061

s = 17.79       R-sq = 7.3%      R-sq(adj) = 5.3%
```

−3	55
−3	4
−2	988
−2	0
−1	985
−1	2110
−0	99887
−0	410
0	134
0	557899
1	01122333
1	55678
2	044
2	78



10.16. The first model (using watershed area to predict IBI) is preferable because the regression was significant ($P = 0.001$ vs. $P = 0.061$) and explained a higher proportion of the variation in IBI (19.9% vs. 7.3%).

10.17. The precise results of these changes depend on which observation is changed. (There are six observations which had 0% forest and two which had 100% forest.) Specifically, if we change IBI to 0 for one of the first six observations, the resulting P-value is between 0.019 (observation 6) and 0.041 (observation 3). Changing one of the last two observations changes the P-value to 0.592 (observation 48) or 0.645 (observation 49).

In general, the first change decreases P (that is, the relationship is more significant) because it accentuates the positive association. The second change weakens the association, so P increases (the relationship is less significant).

10.18. With the regression equation $\widehat{\text{IBI}} = 52.92 + 0.4602\,\text{Area}$, the predicted mean response when $x = \text{Area} = 30\ \text{km}^2$ is $\hat{\mu}_y = \widehat{\text{IBI}} \doteq 66.73$. While it is possible to find $\text{SE}_{\hat{\mu}}$ and $\text{SE}_{\hat{y}}$ using the formulas from Section 10.2, we rely on the software output shown below. ($\text{SE}_{\hat{\mu}} \doteq 2.37$, reported by Minitab as "Stdev.fit," and $\text{SE}_{\hat{y}} = \sqrt{s^2 + \text{SE}_{\hat{\mu}}^2} \doteq 16.70$, where $s \doteq 16.53$ was given in the Minitab output shown with the solution to Exercise 10.14. For $\text{df} = 47$, the appropriate critical value is $t^* = 2.0117$.) **(a)** The 95% confidence interval for μ_y is 61.95 to 71.50. **(b)** The 95% prediction interval for a future response is 33.12 to 100.34. **(c)** Among *many* streams with watershed area 30 km^2, we estimate the mean IBI to be between about 61.95 and 71.50. For an *individual* stream with watershed area 30 km^2, we expect its IBI to be between about 33.12 and 100.34. **(d)** We probably cannot reliably apply these results elsewhere; it is likely that the particular characteristics of the Ozark Highland region play some role in determining the regression coefficients.

Minitab output

```
    Fit  Stdev.Fit       95.0% C.I.           95.0% P.I.
  66.73       2.37   (   61.95,   71.50)  (   33.12,  100.34)
```

10.19. Using $\text{Area} = 10$ in the model $\widehat{\text{IBI}} = 52.92 + 0.4602\,\text{Area}$ from Exercise 10.14, $\widehat{\text{IBI}} \doteq 57.52$. Using $\text{Forest} = 25$ in the model $\widehat{\text{IBI}} = 59.91 + 0.1531\,\text{Forest}$ from Exercise 10.15, $\widehat{\text{IBI}} \doteq 63.74$. Both predictions have a lot of uncertainty; recall that r^2 was fairly small for both models. Also note that the prediction intervals (shown below) are both about 70 units wide.

Minitab output

```
- - - - - - - - - - -   IBI predicted from watershed area   - - - - - - - - - - -
    Fit  Stdev.Fit       95.0% C.I.           95.0% P.I.
  57.52       3.41   (   50.66,   64.39)  (   23.55,   91.50)
- - - - - - - - - - -   IBI predicted from percent forest   - - - - - - - - - - -
    Fit  Stdev.Fit       95.0% C.I.           95.0% P.I.
  63.74       2.79   (   58.13,   69.35)  (   27.51,   99.97)
```

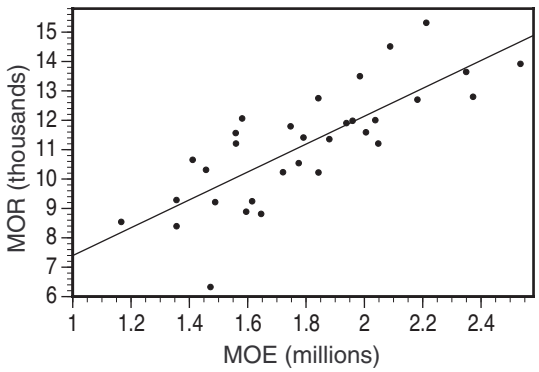
10.20. **(a)** β_0 is the population intercept, 4.6. This says that the mean overseas return is 4.6% when the U.S. return is 0%. **(b)** β_1 is the population slope, 0.67. This says that when the U.S. return changes by 1%, the mean overseas return changes by 0.67%. **(c)** The full model is $y_i = 4.6 + 0.67x_i + \epsilon_i$, where y_i and x_i are observed overseas and U.S. returns in a given year, and ϵ_i are independent $N(0, \sigma)$ variables. The residual terms ϵ_i allow for variation in overseas returns when U.S. returns remain the same.

10.21. **(a)** The stemplots (top of next page, left) are fairly symmetric. For x (MOE), $\bar{x} \doteq 1,799,180$ and $s_x \doteq 329,253$; for y (MOR), $\bar{y} \doteq 11,185$ and $s_y \doteq 1980$. **(b)** The plot (top of next page, right) shows a moderately strong, positive, linear relationship. Because we would like to predict MOR from MOE, we should put MOE on the x axis. **(c)** The model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, ..., 32$; ϵ_i are independent $N(0, \sigma)$ variables. The regression equation is $\widehat{\text{MOR}} = 2653 + 0.004742\,\text{MOE}$, $s \doteq 1238$. The slope is significantly different from 0: $t = 7.02$ ($\text{df} = 30$), $P < 0.0001$. **(d)** Assumptions appear to be met: A stemplot of the residuals shows one slightly low (not quite an outlier), but acceptable, and the plot of residuals against MOE (not shown) does not suggest any particular pattern.

MOE	
11	6
12	
13	55
14	1578
15	5589
16	14
17	2479
18	447
19	358
20	0348
21	8
22	1
23	47
24	
25	3

MOR	
6	3
7	
8	3588
9	222
10	22356
11	223455799
12	00777
13	469
14	5
15	3

Residuals	
−3	3
−2	
−2	
−1	6
−1	31110
−0	76555
−0	43221
0	00223
0	78
1	1334
1	599
2	1



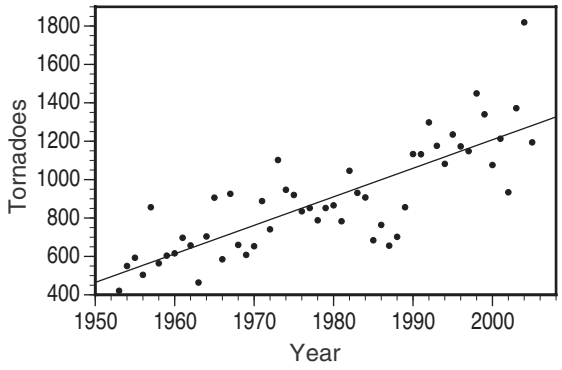
10.22. **(a)** The 95% confidence interval gives a range of values for the mean MOR of *many* pieces of wood with MOE equal to 2,000,000. The prediction interval gives a range of values for the MOR of *one* piece of wood with MOE equal to 2,000,000. **(b)** The prediction interval will include more values because the confidence interval accounts only for the uncertainty in our estimate of the mean response, while the prediction interval must also account for the random error of an individual response. **(c)** With the regression equation $\widehat{\text{MOR}} = 2653 + 0.004742\,\text{MOE}$, the predicted mean response when $x = \text{MOE} = 2{,}000{,}000$ is $\hat{\mu}_y = \widehat{\text{MOE}} \doteq 12,137$. The Minitab output below gives the two intervals, along with $\text{SE}_{\hat{\mu}}$ ("Stdev.fit").

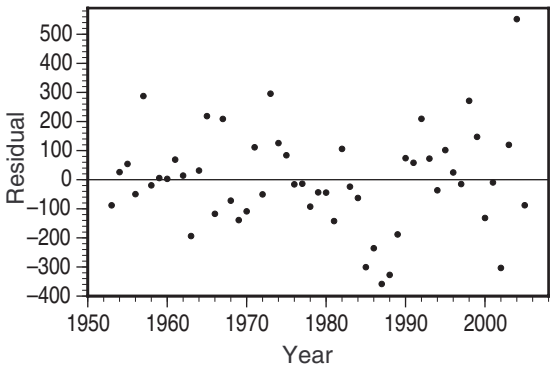
Minitab output

```
    Fit  Stdev.Fit        95.0% C.I.            95.0% P.I.
  12137        257   (   11611,   12663)  (    9554,   14720)
```

10.23. **(a)** The plot (top of next page, left) is roughly linear and increasing. **(b)** The number of tornadoes in 2004 (1819) is noticeably high. **(c)** The regression equation is $\hat{y} \doteq -28{,}516 + 14.86x$; both the slope and intercept are significantly different from 0. In the Minitab output that follows, we see $\text{SE}_{b_1} \doteq 1.527$. With $t^* = 2.0076$ for df $= 51$, the confidence interval for β_1 is $b_1 \pm t^*\text{SE}_{b_1} = 14.86 \pm 3.07 \doteq 11.79$ to 17.93 tornadoes per year. **(d)** Apart from the large residual for 2004, there are no striking features in the plot (top of next page, right). **(e)** Based on a stemplot (right), the 2004 residual is an outlier; the other residuals appear to be roughly Normal.

−3	5200
−2	3
−1	9843310
−0	9887654443211110
0	001223556778
1	001124
2	001789
3	
4	
5	5





Minitab output

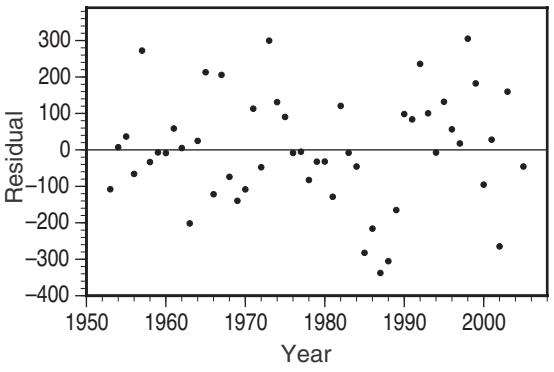
```
The regression equation is Tornado = -28516 + 14.9 Year

Predictor       Coef       Stdev    t-ratio        p
Constant      -28516        3022      -9.44    0.000
Year          14.862       1.527       9.73    0.000

s = 170.0       R-sq = 65.0%     R-sq(adj) = 64.3%
```

10.24. Refer to the scatterplot shown in the previous solution. With the 2004 tornado count removed, the regression equation is $\hat{y} \doteq -26{,}162 + 13.67x$. The estimated rate of increase (that is, the slope) is lower than that found in the previous solution by about 1 tornado per year. With $\text{SE}_{b_1} \doteq 1.396$ and $t^* = 2.0086$ for df $= 50$, the confidence interval for β_1 is $b_1 \pm t^*\text{SE}_{b_1} = 13.67 \pm 2.80 \doteq 10.87$ to 16.47 tornadoes per year. In examining the residuals, neither the stemplot nor the plot of residuals versus year (both below) suggest any deviations from the assumptions of our regression model.

−3	30
−2	86
−2	10
−1	6
−1	32200
−0	9876
−0	444333000000
0	001223
0	55899
1	01233
1	58
2	013
2	79
3	0



Minitab output

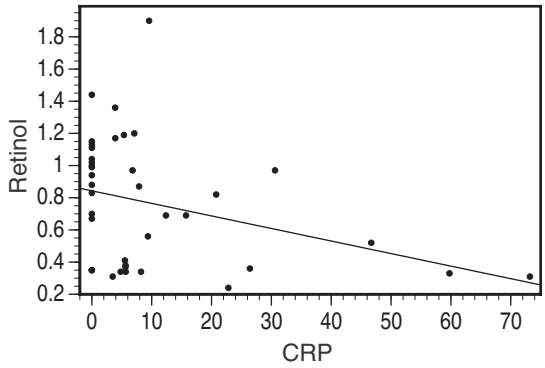
```
The regression equation is Tornado = - 26162 + 13.7 Year

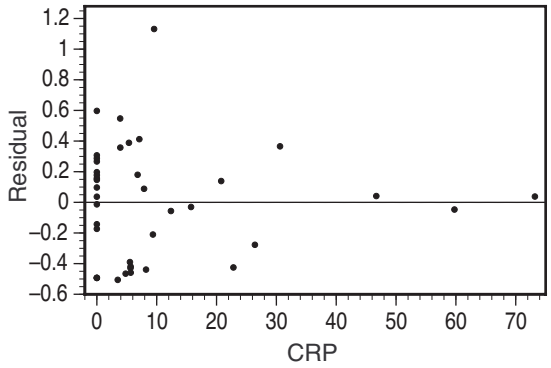
Predictor       Coef       Stdev    t-ratio        p
Constant      -26162        2763      -9.47    0.000
Year          13.666       1.396       9.79    0.000

s = 151.5       R-sq = 65.7%     R-sq(adj) = 65.0%
```

10.25. (a) x (CRP) is sharply right-skewed with high outliers; $\bar{x} = 10.0323$ and $s_x = 16.5632$. y (retinol) is slightly right-skewed; $\bar{y} = 0.7648$ and $s_y = 0.3949$. **(b)** No; no assumption is made about the distribution of x values. Note that this does not mean that we do not care about the distribution of the x values; the outliers cause trouble, as we see in (d). **(c)** The regression equation is $\widehat{\text{Retinol}} = 0.8430 - 0.007800\ \text{CRP}$, $s = 0.3781$. With $\alpha = 0.05$, the slope is significantly different from 0: $t = -2.13$ (df $= 38$), $P = 0.039$. **(d)** The high outliers in CRP are influential, as we can see from the small residuals on the right end of the plot below. Additionally, a stemplot or quantile plot of the residuals (not shown) shows that the distribution is right-skewed rather than Normal.

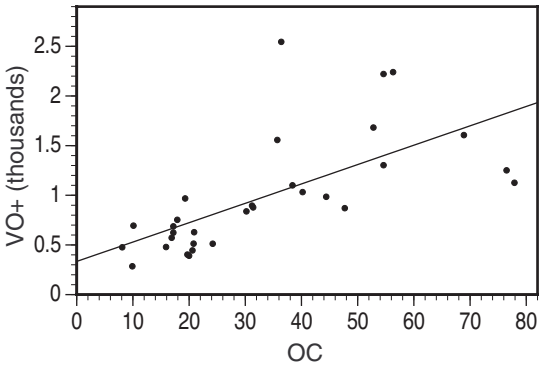
CRP		Retinol
0	0000000000000003334	2333333333333
0	55555677899	455
1	2	6667
1	5	88889999
2	02	00011111
2	6	23
3	0	4
3		1
4		9
4	6	
5		
5	9	
6		
6		
7	3	





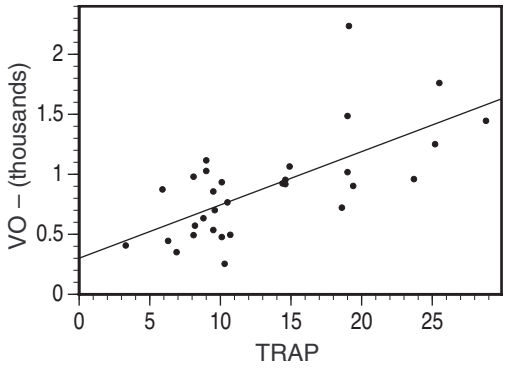
10.26. (a) Both distributions are right-skewed. OC has $\bar{x} = 33.4161$ and $s_x = 19.6097$; VO+ has $\bar{y} = 985.8065$ and $s_y = 579.8581$. **(b)** Put OC on the x axis because we hope to use it as the explanatory variable. We see a positive association, but one point is an outlier (it is far above the pattern of the rest of the points) and there appears to be more scatter about the line for large values of OC. **(c)** The regression equation is $\hat{y} = 334.0 + 19.505x$, $s = 443.3$. The slope is significantly different from 0: $t = 4.73$, $P < 0.0005$. The residuals appear to be somewhat right-skewed, and the unusual point noted in (b) corresponds to a high outlier in the distribution of residuals.

OC		VO+		Residuals	
0	89	0	23	−0	7
1	0	0	4444555	−0	5
1	5677799	0	66667	−0	33322222
2	00004	0	888899	−0	11000000000
2		1	011	0	0001
3	011	1	23	0	23
3	568	1	5	0	5
4	04	1	66	0	
4	7	1		0	88
5	244	2		1	
5	6	2	22	1	
6		2	5	1	5
6	8				
7					
7	67				



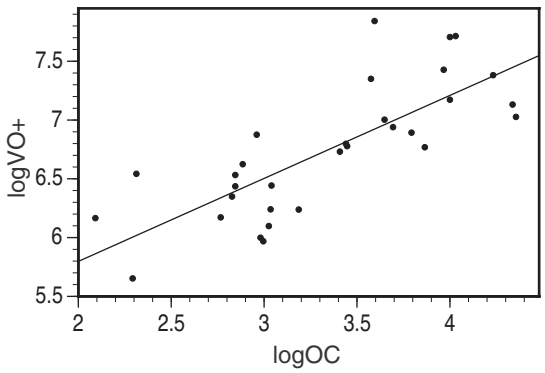
10.27. **(a)** Both distributions are right-skewed (TRAP more than VO–). TRAP has $\bar{x} = 13.2484$ and $s_x = 6.5282$; VO– has $\bar{y} = 889.1935$ and $s_y = 427.6161$. **(b)** Put TRAP on the x axis because we hope to use it as the explanatory variable. We see a positive association, but one point is an outlier (it is far above the pattern of the rest of the points). **(c)** The regression equation is $\hat{y} = 300.9 + 44.406x$, $s = 319.7$, The slope is significantly different from 0: $t = 4.97$, $P < 0.0005$. The unusual point noted in (b) corresponds to a high outlier among the residuals; otherwise the distribution of residuals seems reasonably Normal.

TRAP		VO–		Residuals	
0	3	0	23	−0	54
0	5	0	4444455	−0	32222
0	66	0	6777	−0	1111110000000
0	888899999	0	889999999	0	0111
1	00000	1	0001	0	33333
1		1	2	0	4
1	4444	1	44	0	
1		1	7	0	
1	89999	1		1	0
2		2			
2	3	2	2		
2	55				
2					
2	8				



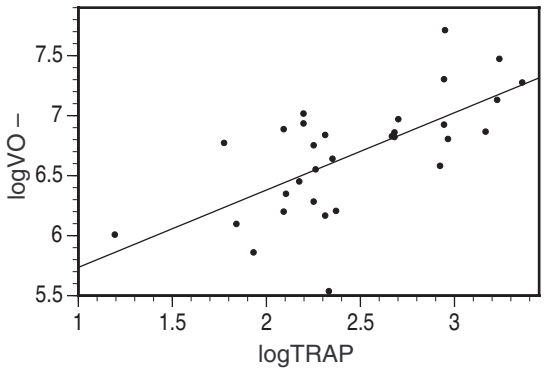
10.28. After taking (natural) logarithms, both distributions are considerably less skewed; logOC is irregular, while logVO+ is quite symmetric. logOC has $\bar{x} = 3.3379$ and $s_x = 0.6085$; logVO+ has $\bar{y} = 6.7419$ and $s_y = 0.5554$. (If common logarithms are used, multiply these results by about 2.3026.) The scatterplot shows a positive association, stronger than that seen in Exercise 10.26. The regression equation is $\hat{y} = 4.3852 + 0.7060x$, $s = 0.3580$. The slope is significantly different from 0: $t = 6.57$, $P < 0.0005$. The distribution of the residuals appears to be much more Normal than in Exercise 10.26.

logOC		logVO+		Residuals	
2	0	5	6	−0	5444
2	23	5	99	−0	33332
2		6	011	−0	110000000
2	7	6	223	0	0001
2	8888999	6	4455	0	223
3	0001	6	67777	0	44445
3		6	889	0	
3	44455	7	0011	0	9
3	667	7	33		
3	89	7	4		
4	000	7	77		
4	233	7	8		

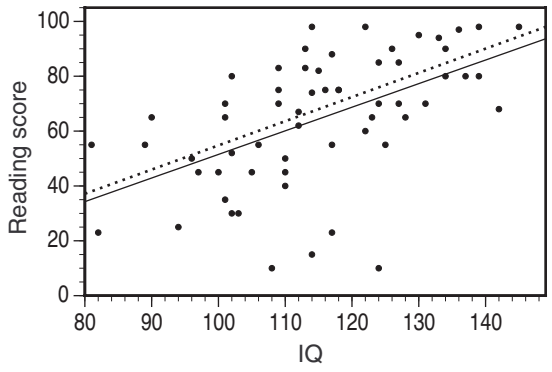


10.29. After taking (natural) logarithms, both distributions are somewhat irregular, but slightly more symmetric. logTRAP has $\bar{x} = 2.4674$ and $s_x = 0.4979$; logVO– has $\bar{y} = 6.6815$ and $s_y = 0.4832$. (If common logarithms are used, multiply these results by about 2.3026.) The scatterplot shows a positive association; the outlier visible in Exercise 10.27 has moved closer to the other points. The regression equation is $\hat{y} = 5.0910 + 0.6446x$, $s = 0.3674$. The slope is significantly different from 0: $t = 4.78$, $P < 0.0005$. The distribution of the residuals appears to be reasonably Normal, apart from a low outlier.

logTRAP		logVO–		Residuals	
1	1	5	5	−1	0
1		5		−0	
1		5	8	−0	
1	7	6	001	−0	444
1	89	6	2223	−0	3222
2	001111	6	455	−0	110000
2	22233333	6	677	0	00000011
2		6	8888888999	0	2223
2	6667	7	01	0	4455
2	99999	7	23	0	7
3	1	7	4		
3	223	7	7		

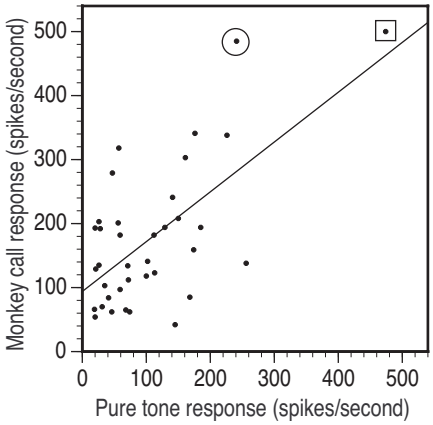


10.30. **(a)** With all 60 points, the regression equation is $\hat{y} = -34.55 + 0.8605x$, $s \doteq 20.17$. (This is the solid line in the scatterplot on the right.) The slope is significantly different from 0: $t = 4.82$, $P < 0.0005$. **(b)** Without the four points from the bottom of the scatterplot, the regression equation is $\hat{y} = -33.40 + 0.8818x$, $s \doteq 15.18$. (This is the dashed line in the scatterplot.) The slope is again significantly different from 0: $t = 6.57$, $P < 0.0005$. With the outliers removed, the line changes slightly; the most significant change is the decrease in the estimated standard deviation s. This correspondingly makes t larger (i.e., b_1 is *more* significantly different from 0) and makes the regression line more useful for prediction (r^2 increases from 28.9% to 44.4%). Of course, we should not arbitrarily remove data points; more investigation is needed to determine why these students' reading scores were so much lower than we would expect based on their IQs.



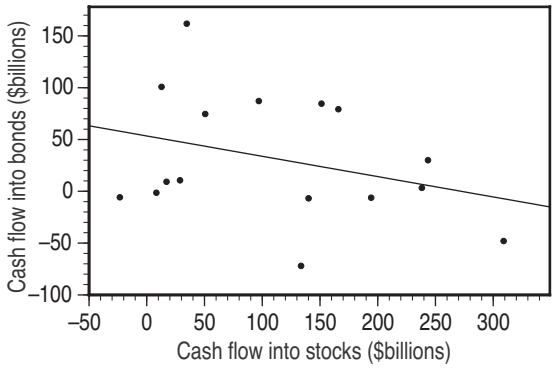
10.31. **(a)** Both variables are right-skewed. For pure tones, $\bar{x} \doteq 106.20$ and $s \doteq 91.76$ spikes/second, and for monkey calls, $\bar{y} = 176.57$ and $s_y = 111.85$ spikes/second. **(b)** There is a moderate positive association; the third point (circled) has the largest residual; the first point (marked with a square) is an outlier for tone response. **(c)** With all 37 points, $\widehat{\text{CALL}} = 93.9 + 0.778$ TONE and $s = 87.30$; the test of $\beta_1 = 0$ gives $t = 4.91$, $P < 0.0001$. **(d)** Without the first point, $\hat{y} = 101 + 0.693x$, $s = 88.14$, $t = 3.18$. Without the third point, $\hat{y} = 98.4 + 0.679x$, $s = 80.69$, $t = 4.49$. With neither, $\hat{y} = 116 + 0.466x$, $s = 79.46$, $t = 2.21$. The line changes a bit, but always has a slope significantly different from 0.

Tone		Call		Residual	
0	122222233444	0	4	−1	65
0	55556777	0	566667889	−1	3
1	0011244	1	011223334	−0	8876555
1	566778	1	5889999	−0	44444331100
2	24	2	0004	0	012334
2	5	2	7	0	667888
3		3	0134	1	14
3		3		1	7
4		4		2	0
4	7	4	8		
		5	0		



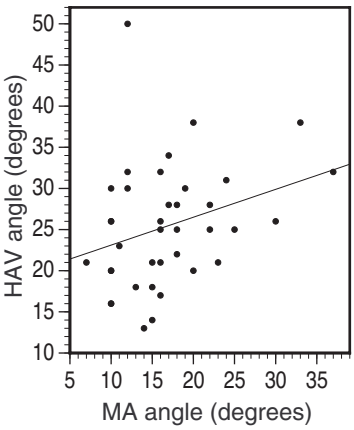
10.32. The model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$; ϵ_i are independent $N(0, \sigma)$ variables. **(a)** β_0 represents the fixed costs. **(b)** β_1 represents how costs change as the number of students changes. This should be positive because more students mean more expenses. **(c)** The error term (ϵ_i) allows for variation among equal-sized schools.

10.33. **(a)** The scatterplot shows a weak negative association; the regression equation is $\widehat{\text{Bonds}} = 53.41 - 0.1962$ Stocks with $s \doteq 59.88$. **(b)** For testing H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$, we have $t = -1.27$ (df = 14) and $P = 0.226$. The slope is *not* significantly different from 0. **(c)** The scatterplot shows a lot of variation, so s is large and t is small.



10.34. **(a)** The t statistic for testing H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$ is $t = b_1/\text{SE}_{b_1} = 0.76/0.44 \doteq 1.73$ with df = 80. This has $P = 0.0880$, so we do not reject H_0. **(b)** For the one-sided alternative $\beta_1 > 0$, we would have $P = 0.0440$, so we could reject H_0 at the 5% significance level.

10.35. See also the solutions to Exercises 2.20 and 2.68.



(a) MA angle is the explanatory variable, so it should be on the horizontal axis of the scatterplot. (This scatterplot has the same scale on both axes because both variables are measured in degrees.) **(b)** The scatterplot shows a moderate-to-weak positive linear association, with one clear outlier (the patient with HAV angle 50°). **(c)** The model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, \ldots, 38$; ϵ_i are independent $N(0, \sigma)$ variables. **(d)** Because doctors expect there to be a positive association, we use a one-sided alternative: H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 > 0$. **(e)** For the estimated slope $b_1 \doteq 0.3388$, we have $t = 1.90$ (df $= 36$) and $P = 0.033$ (half of Minitab's two-sided P-value); this is significant evidence (at $\alpha = 0.05$) to support the doctors' belief.

Minitab output

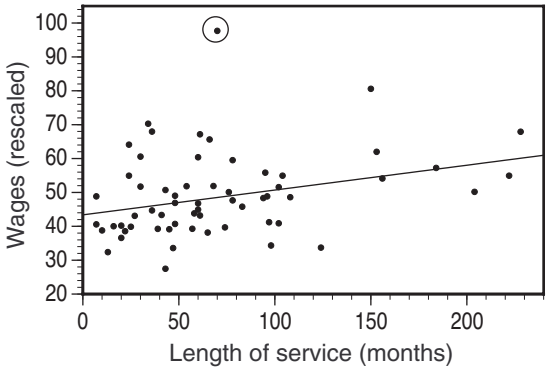
```
The regression equation is HAV = 19.7 + 0.339 MA

Predictor       Coef       Stdev    t-ratio        p
Constant      19.723       3.217       6.13    0.000
MA            0.3388      0.1782       1.90    0.065

s = 7.224       R-sq = 9.1%      R-sq(adj) = 6.6%
```

10.36. Software (Minitab output above) reports $b_1 \doteq 0.3388$ and $\text{SE}_{b_1} \doteq 0.1782$. For a t distribution with df $= 36$, $t^* \doteq 2.0281$ for a 95% confidence interval, so the interval is -0.0226 to 0.7002. The slope was significantly different from 0 using a one-sided alternative, but this interval tells us that it could be 0 (or even slightly negative); we would not reject $\beta_1 = 0$ in favor of a two-sided alternative.

10.37. **(a)** Aside from the one high point (70 months of service, and wages 97.6801), there is a moderate positive association—fairly clear but with quite a bit of scatter. **(b)** The regression equation is $\widehat{\text{WAGES}} = 43.383 + 0.07325$ LOS, with $s \doteq 10.21$ (Minitab output below). The slope is significantly different from 0: $t = 2.85$ (df $= 57$), $P = 0.006$. **(c)** Wages rise an average of 0.07325 wage units per week of service. **(d)** We have $b_1 \doteq 0.07325$ and $\text{SE}_{b_1} \doteq 0.02571$. For a t distribution with df $= 57$, $t^* \doteq 2.0025$ for a 95% confidence interval, so the interval is 0.0218 to 0.1247.



Minitab output

```
The regression equation is wages = 43.4 + 0.0733 los

Predictor       Coef       Stdev    t-ratio        p
Constant      43.383       2.248      19.30    0.000
los          0.07325     0.02571       2.85    0.006

s = 10.21       R-sq = 12.5%     R-sq(adj) = 10.9%
```

10.38. The table below summarizes the regression results with the outlier excluded, and those with all points. **(a)** The intercept and slope estimates change very little, but the estimate of σ increases from 10.21 to 11.98. **(b)** With the outlier, the t statistic decreases (because s has increased), and the P-value increases slightly—although it is still significant at the 5% level. **(c)** The interval width $2t^*\text{SE}_{b_1}$ increases from 0.1030 to 0.1207—roughly the same factor by which s increased. (Because the degrees of freedom change from 57 to 58, t^* decreases from 2.0025 to 2.0017, but the change in s has a much greater impact.)

	b_0	b_1	s	t	P	Interval width
Outlier excluded	43.383	0.07325	10.21	2.85	0.006	0.1030
All points	44.213	0.07310	11.98	2.42	0.018	0.1207

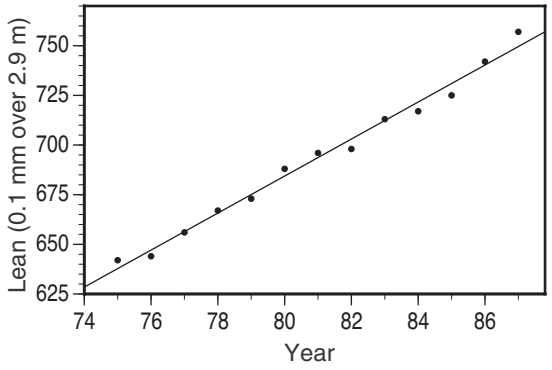
Minitab output

```
The regression equation is wages = 44.2 + 0.0731 los

Predictor       Coef       Stdev    t-ratio        p
Constant      44.213       2.628      16.82    0.000
los          0.07310     0.03015       2.42    0.018

s = 11.98       R-sq = 9.2%      R-sq(adj) = 7.6%
```

10.39. **(a)** The trend appears to be quite linear. **(b)** The regression equation is $\widehat{\text{Lean}} = -61.12 + 9.3187\ \text{Year}$ with $s \doteq 4.181$. The regression explains $r^2 = 98.8\%$ of the variation in lean. **(c)** The rate we seek is the slope. For df $= 11$ and 99% confidence, $t^* = 3.1058$, so the interval is $9.3187 \pm (3.1058)(0.3099) = 8.3562$ to 10.2812 tenths of a millimeter/year.



Minitab output

```
The regression equation is Lean = -61.1 + 9.32 Year

Predictor       Coef       Stdev    t-ratio        p
Constant     -61.12       25.13      -2.43    0.033
Year         9.3187      0.3099      30.07    0.000

s = 4.181       R-sq = 98.8%     R-sq(adj) = 98.7%
```

10.40. **(a)** $\hat{y} = -61.12 + 9.3187(18) \doteq 107$, for a prediction of 2.9107 m. **(b)** This is an example of extrapolation—trying to make a prediction outside the range of given x-values. Minitab reports that a 95% prediction interval for $\hat{y}$ when $x^* = 18$ is about 62.6 to 150.7. The width of the interval is an indication of how unreliable the prediction is.

Note: *Minitab's "Stdev.Fit" value of 19.56 is $SE_{\hat{\mu}}$, so $SE_{\hat{y}} \doteq \sqrt{s^2 + SE_{\hat{\mu}}^2} \doteq 20.00$, which agrees with the margin for the prediction interval: $t^*SE_{\hat{y}} = (2.201)(20.00) \doteq 44.02$.*

Minitab output

```
    Fit  Stdev.Fit       95.0% C.I.           95.0% P.I.
 106.62      19.56   (   63.56,  149.68)  (   62.58,  150.65) XX
 XX denotes a row with very extreme X values
```

10.41. (a) Use $x = 109$ (the number of years after 1900). **(b)** $\hat{y} = -61.12 + 9.3187(109) \doteq 955$, for a prediction of 2.9955 m. **(c)** A prediction interval is appropriate because we are interested in one future observation, not the mean of all future observations; in this situation, it does not make sense to talk of more than one future observation. In the output below, note that Minitab warns us of the risk of extrapolation.

Minitab output

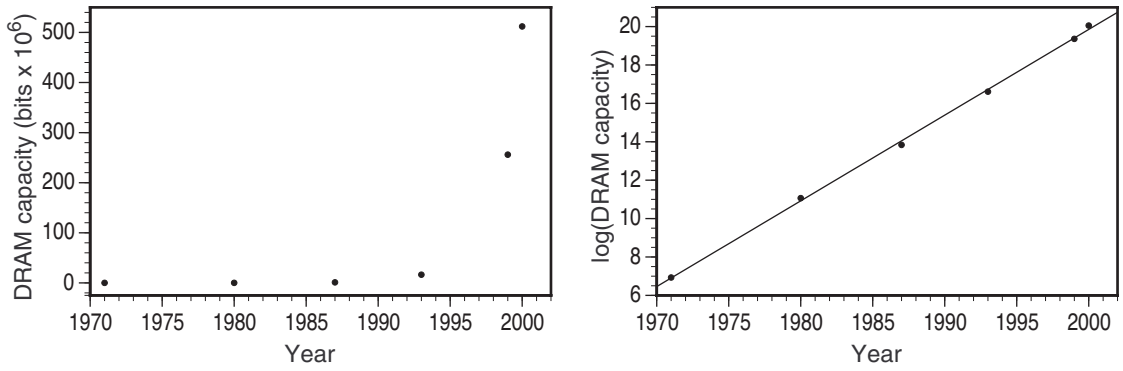
```
    Fit  Stdev.Fit       95.0% C.I.           95.0% P.I.
954.62       8.75   (  935.34,  973.89)  (  933.26,  975.97) XX
XX denotes a row with very extreme X values
```

10.42. A negative association makes sense here: If the price of beer is above average, fewer students can afford to drink, while more drinking happens when beer is cheaper.

Note: *The fact that the correlation is relatively small indicates that the price of beer is not a crucial factor in determining the prevalence of binge-drinking. In particular, a straight-line relationship with the cost of beer only explains about* $r^2 \doteq 13\%$ *of the variation in binge-drinking rates.*

10.43. To test H_0: $\rho = 0$ vs. H_a: $\rho \neq 0$, we compute $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \doteq -4.16$. Comparing this to a t distribution with df $= 116$, we find $P < 0.0001$, so we conclude the correlation is different from 0.

10.44. (a) Scatterplot below, left. **(b)** Scatterplot below, right. **(c)** The regression equation is $\hat{y} = -872.93 + 0.4464x$ with $s \doteq 0.1739$. For 95% confidence with df $= 4$, $t^* = 2.7765$, so with $b_1 \doteq 0.4464$ and $\text{SE}_{b_1} \doteq 0.006856$, the confidence interval is 0.4274 to 0.4654.



Minitab output

```
The regression equation is logBits = - 873 + 0.446 year

Predictor       Coef       Stdev    t-ratio        p
Constant     -872.93       13.63     -64.03    0.000
year        0.446390    0.006856      65.11    0.000

s = 0.1739      R-sq = 99.9%     R-sq(adj) = 99.9%
```

10.45. Recall that testing H_0: $\rho = 0$ vs. H_a: $\rho \neq 0$ is the same as testing H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$. In the solution to Exercise 10.33, we had $t = -1.27$ (df $= 14$) and $P = 0.226$, so we cannot reject H_0.

10.46. **(a)** With $r = -0.19$ and $n = 713$, we have $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \doteq -5.16$. **(b)** Comparing to a t distribution with df $= 711$ (or anything reasonably close), the P-value is less than 0.0001, so we conclude that $\rho \neq 0$.

10.47. Because DFT $=$ DFM $+$ DFE and SST $=$ SSM $+$ SSE, we can find the missing degrees of freedom (DF) and sum of squares (SS) entries by subtraction: df $=$ DFE $= 28$ and SSE $= 10152.4$. The missing entry in the mean square (MS) column is MSE $=$ SSE/DFE $\doteq 362.6$. (We can also compute MSE $=$ MSM/F $\doteq 362.7$—the same answer, up to rounding.)

10.48. $s = \sqrt{\text{MSE}} \doteq 19.0416$ and $r^2 = \frac{\text{SSM}}{\text{SST}} = \frac{3445.9}{13598.3} \doteq 0.2534$.

10.49. As $s_x = \sqrt{\frac{1}{29}\sum(x_i - \bar{x})^2} = 16.45\%$, we have $\sqrt{\sum(x_i - \bar{x})^2} = s_x\sqrt{29} \doteq 88.5860\%$, so:

$$\text{SE}_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} \doteq \frac{19.0420}{88.5860} \doteq 0.2150$$

Alternatively, note that we have $F = 9.50$ and $b_1 = 0.663$. Because $t^2 = F$, we know that $t = 3.0822$ (take the positive square root, because $t = b_1/\text{SE}_{b_1}$, and b_1 is positive). Then $\text{SE}_{b_1} = b_1/t = 0.2151$—the same answer, up to rounding. (Note that with this approach, we do not need to know that $s_x = 16.45\%$.)

With df $= 28$, $t^* = 2.0484$ for 95% confidence, so the 95% confidence interval is $0.663 \pm 0.4403 = 0.2227$ to 1.1033.

10.50. **(a)** With $\bar{x} \doteq 80.9$, $s_x \doteq 17.2$, $\bar{y} \doteq 43.5$, $s_y \doteq 20.3$, and $r \doteq 0.68$, we find:

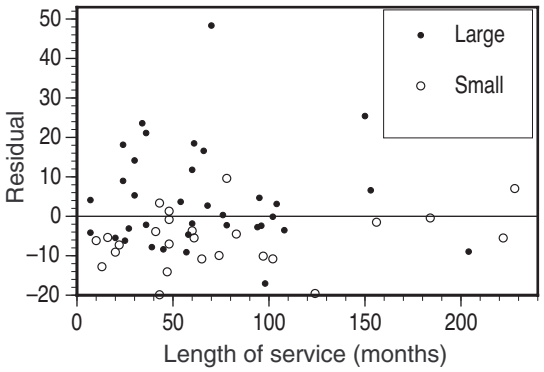
$$b_1 = (0.68)\left(\frac{20.3}{17.2}\right) \doteq 0.8026$$

$$b_0 = 43.5 - (0.8026)(80.9) \doteq -21.4270$$

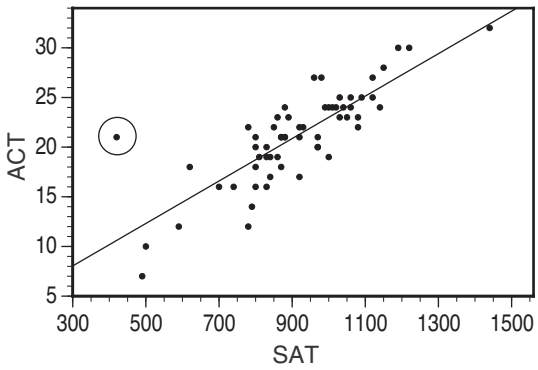
(Answers may vary slightly due to rounding.) The regression equation is therefore $\widehat{\text{GHP}} = -21.4270 + 0.8026\,\text{FVC}$. **(b)** Testing $\beta_1 = 0$ is equivalent to testing $\rho = 0$, so the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \doteq 6.43$ (df $= 48$), for which $P < 0.0005$. The slope (correlation) is significantly different from 0.

10.51. Use the formula $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with $r = 0.5$. For $n = 20$, $t = 2.45$ with df $= 18$, for which the two sided P-value is $P = 0.0248$. For $n = 10$, $t = 1.63$ with df $= 8$, for which the two sided P-value is $P = 0.1411$. With the larger sample size, r should be a better estimate of ρ, so we are less likely to get $r = 0.5$ unless ρ is really not 0.

10.52. Most of the small banks have negative residuals, while most large-bank residuals are positive. This means that, generally, wages at large banks are higher, and small bank wages are lower, than we would predict from the regression.



10.53. **(a)** Not surprisingly, there is a positive association between scores. The 47th pair of scores (circled) is an outlier—the ACT score (21) is higher than one would expect for the SAT score (420). Since this SAT score is so low, this point may be influential. No other points fall outside the pattern. **(b)** The regression equation is $\hat{y} = 1.626 + 0.02137x$. The slope is significantly different from 0: $t = 10.78$ (df $= 58$) for which $P < 0.0005$. **(c)** $r = 0.8167$.



Minitab output

```
The regression equation is ACT = 1.63 + 0.0214 SAT

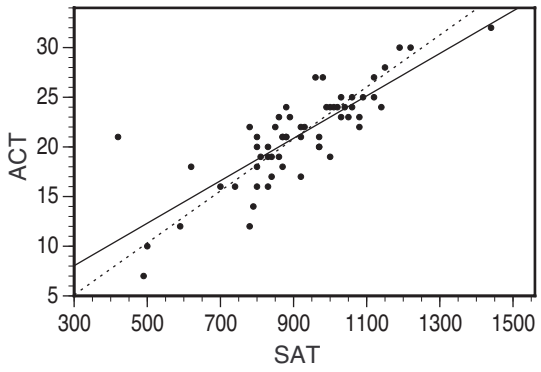
Predictor       Coef       Stdev    t-ratio        p
Constant       1.626       1.844       0.88    0.382
SAT         0.021374    0.001983      10.78    0.000

s = 2.744       R-sq = 66.7%     R-sq(adj) = 66.1%
```

10.54. **(a)** The means are identical (21.133). **(b)** For the observed ACT scores, $s_y = 4.714$; for the fitted values, $s_{\hat{y}} = 3.850$. **(c)** For $z = 1$, the SAT score is $\bar{x} + s_x = 912.7 + 180.1 = 1092.8$. The predicted ACT score is $\hat{y} \doteq 25$ (Minitab reports 24.983), which gives a standard score of about 1 (using the standard deviation of the *predicted* ACT scores. **(d)** For $z = -1$, the SAT score is $\bar{x} - s_x = 912.7 - 180.1 = 732.6$. The predicted ACT score is $\hat{y} \doteq 17.3$ (Minitab reports 17.285), which gives a standard score of about -1. **(e)** It appears that the standard score of the predicted value is the same as the explanatory variable's standard score. (See note below.)

Notes: **(a)** This will always be true because $\sum_i \hat{y}_i = \sum_i (b_0 + b_1 x_i) = n\,b_0 + b_1 \sum_i x_i = n(\bar{y} - b_1\bar{x}) + b_1 n\,\bar{x} = n\,\bar{y}$. **(b)** The standard deviation of the predicted values will be $s_{\hat{y}} = |r| s_y$; in this case, $s_{\hat{y}} = (0.8167)(4.714)$. To see this, observe that the variance of the predicted values is $\frac{1}{n-1}\sum_i (\hat{y}_i - \bar{y})^2 = \frac{1}{n-1}\sum_i (b_1 x_i - b_1\bar{x})^2 = b_1^2 s_x^2 = r^2 s_y^2$. **(e)** For a given standard score z, note that $\hat{y} = b_0 + b_1(\bar{x} + z\,s_x) = \bar{y} - b_1\bar{x} + b_1\bar{x} + b_1 z\,s_x = \bar{y} + z\,r\,s_y$. If $r > 0$, the standard score for $\hat{y}$ equals z; if $r < 0$, the standard score is $-z$.

10.55. (a) For SAT: $\bar{x} = 912.\overline{6}$ and $s_x = 180.1117$. For ACT: $\bar{y} = 21.1\overline{3}$ and $s_y = 4.7137$. Therefore, the slope is $a_1 \doteq 0.02617$ and the intercept is $a_0 \doteq -2.7522$. **(b)** The new line is dashed. **(c)** For example, the first prediction is $-2.7522 + (0.02617)(1000) \doteq 23.42$. Up to rounding error, the mean and standard deviation of the predicted scores are the same as those of the ACT scores: $\bar{y} = 21.1\overline{3}$ and $s_y = 4.7137$.



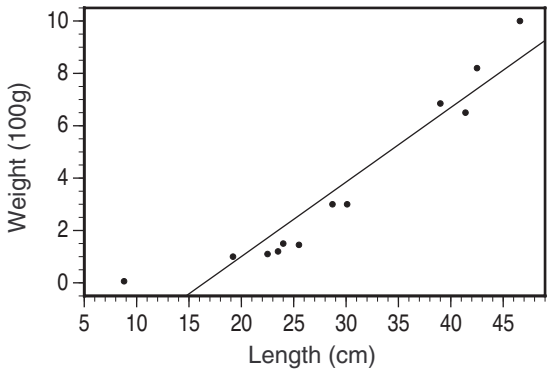
Note: *The usual least-squares line minimizes the total squared vertical distance from the points to the line. If* instead *we seek to minimize the total of* $\sum_i |h_i v_i|$*, where* h_i *is the horizontal distance and* v_i *is the vertical distance, we obtain the line* $\hat{y} = a_0 + a_1 x$*—except that we must choose the sign of* a_1 *to be the same as the sign of* r*. (It would hardly be the "best line" if we had a positive slope with a negative association.) If* $r = 0$*, either sign will do.*

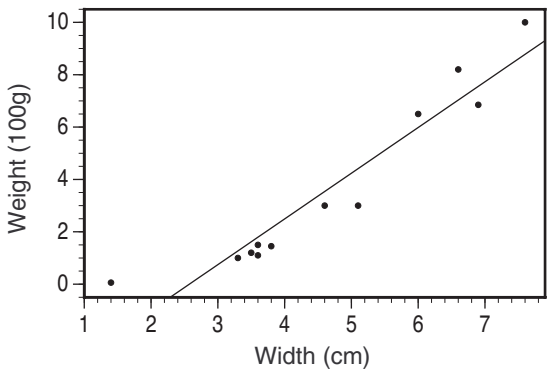
10.56. (a) The regression equations are:

$\widehat{\text{WEIGHT}} = -468.91 + 28.462$ LENGTH with $s \doteq 109.4$ and $r^2 \doteq 0.902$

$\widehat{\text{WEIGHT}} = -449.44 + 174.63$ WIDTH with $s \doteq 107.9$ and $r^2 \doteq 0.905$

(b) Both scatterplots suggest that the relationships are curved rather than linear. (Points to the left and right lie above the line; those in the middle are generally below the line.)





Minitab output

```
- - - - - - - - - - - -   MODEL 1: LENGTH & WEIGHT   - - - - - - - - - - - -
The regression equation is weight = -469 + 28.5 length

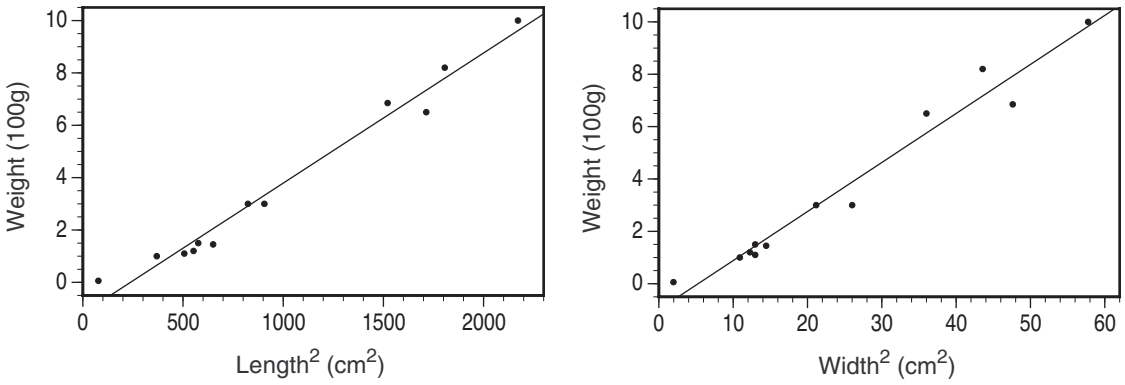
Predictor       Coef       Stdev    t-ratio        p
Constant     -468.91       92.55      -5.07    0.000
length        28.462       2.967       9.59    0.000

s = 109.4       R-sq = 90.2%     R-sq(adj) = 89.2%
- - - - - - - - - - - -   MODEL 2: WIDTH & WEIGHT   - - - - - - - - - - - -
The regression equation is weight = -449 + 175 width

Predictor       Coef       Stdev    t-ratio        p
Constant     -449.44       89.27      -5.03    0.000
width         174.63       17.93       9.74    0.000

s = 107.9       R-sq = 90.5%     R-sq(adj) = 89.5%
```

10.57. (a) For squared length: $\widehat{\text{Weight}} = -117.99 + 0.4970$ SQLEN, $s \doteq 52.76$, $r^2 = 0.977$. **(b)** For squared width: $\widehat{\text{Weight}} = -98.99 + 18.732$ SQWID, $s \doteq 65.24$, $r^2 = 0.965$. Both scatterplots look more linear.



Minitab output

```
- - - - - - - - -  MODEL 1: SQUARED LENGTH & WEIGHT  - - - - - - - - -
The regression equation is weight = -118 + 0.497 sqlen

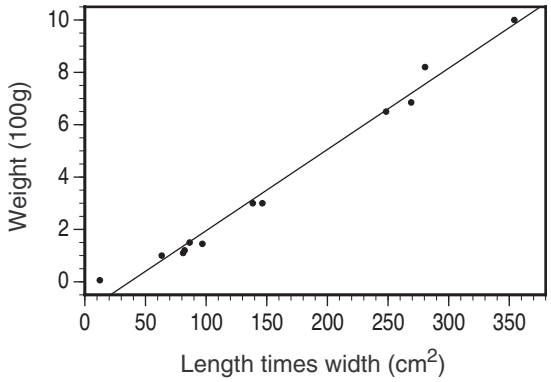
Predictor       Coef       Stdev    t-ratio        p
Constant     -117.99       27.88      -4.23    0.002
sqlen        0.49701     0.02400       20.71    0.000

s = 52.76       R-sq = 97.7%     R-sq(adj) = 97.5%
- - - - - - - - -  MODEL 2: SQUARED LENGTH & WEIGHT  - - - - - - - - -
The regression equation is weight = -99.0 + 18.7 sqwid

Predictor       Coef       Stdev    t-ratio        p
Constant      -98.99       33.67      -2.94    0.015
sqwid         18.732       1.126      16.64    0.000

s = 65.24       R-sq = 96.5%     R-sq(adj) = 96.2%
```

10.58. (a) The regression line is $\widehat{\text{WEIGHT}} = -115.10 + 3.1019(\text{LENGTH})(\text{WIDTH})$, $s \doteq 41.69$, $r^2 = 0.986$. **(b)** As measured by r^2, this last model is (by a slim margin) the best. (However, this scatterplot again gives some suggestion of curvature, indicating that some other model might do better still.)



Minitab output

```
The regression equation is weight = -115 + 3.10 lenwid

Predictor       Coef       Stdev    t-ratio        p
Constant     -115.10       21.87      -5.26    0.000
lenwid        3.1019      0.1179       26.32    0.000

s = 41.69       R-sq = 98.6%     R-sq(adj) = 98.4%
```

10.59. The table on the right shows the correlations and the corresponding test statistics. The first two results agree with the results of (respectively) Exercises 10.14 and 10.15.

	r	t	P
IBI/area	0.4459	3.42	0.0013
IBI/forest	0.2698	1.92	0.0608
area/forest	−0.2571	−1.82	0.0745

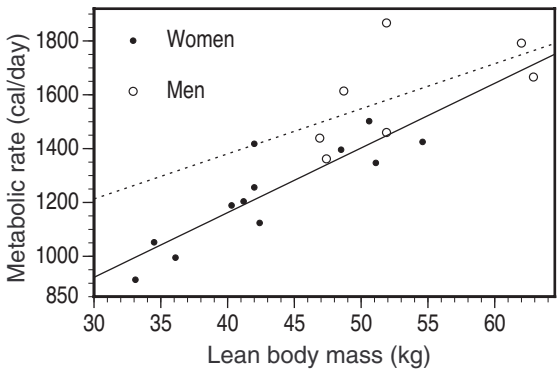
10.60. The correlation was significant for vegetables, fruit, and meat, and nearly significant for eggs. All the significant correlations are negative, meaning (for example) that children with *high* neophobia tend to eat these foods *less* frequently.

	r	t	P
Vegetables	−0.27	−6.65	0.0000
Fruit	−0.16	−3.84	0.0001
Meat	−0.15	−3.60	0.0004
Eggs	−0.08	−1.90	0.0576
Sweet/fatty snacks	0.04	0.95	0.3430
Starchy staples	−0.02	−0.47	0.6355

10.61. Quickness and creativity are both (significantly) positively correlated with all GRE scores, so creative people are *not* penalized (contradicting the critics), while quick workers do better (as some have suggested). Depth is positively associated with verbal scores (refuting the opinion that deep thinkers are penalized). The only other significant correlations are those for conscientious workers, who apparently tend to score *lower* on all parts of the GRE. Depending on what "conscientiousness" measures, this might simply be the flip side of the positive correlation with quickness; perhaps "conscientious" means (in part) that these people work more slowly.

	Analytical	Quantitative	Verbal
Conscientiousness	−0.17**	−0.14**	−0.12*
Rationality	−0.06	−0.03	−0.08
Ingenuity	−0.06	−0.08	−0.02
Quickness	0.21***	0.15**	0.26***
Creativity	0.24***	0.26***	0.29***
Depth	0.06	0.08	0.15**

10.62. See also the solution to Exercise 2.13. **(a)** The association is linear and positive; the women's points show a stronger association. As a group, males typically have larger values for both variables. **(b)** The women's regression line (the solid line in the graph) is $\hat{y} = 201.2 + 24.026x$, with $s \doteq 95.08$ and $r^2 = 0.768$. The men's line (the dashed line) is $\hat{y} = 710.5 + 16.75x$, with $s \doteq 167.1$ and $r^2 = 0.351$. The women's slope is significantly different from 0 ($t = 5.76$, df $= 10$, $P < 0.0005$), but the men's is not ($t = 1.64$, df $= 5$, $P = 0.161$). These test results, and the values of s and r^2, confirm the observation that the women's association is stronger—however, see the solution to the next exercise.



10.63. (a) These intervals (in the table below) overlap quite a bit. **(b)** These quantities can be computed from the data, but it is somewhat simpler to recall that they can be found from the sample standard deviations $s_{x,\mathrm{w}}$ and $s_{x,\mathrm{m}}$:

$$s_{x,\mathrm{w}}\sqrt{11} \doteq 6.8684\sqrt{11} \doteq 22.78 \quad \text{and} \quad s_{x,\mathrm{m}}\sqrt{6} \doteq 6.6885\sqrt{6} \doteq 16.38$$

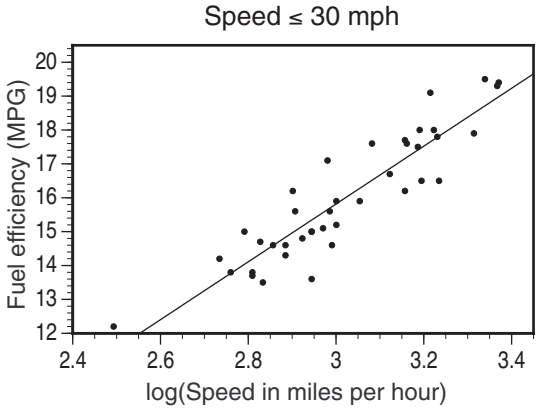
The women's SE_{b_1} is smaller in part because it is divided by a large number. **(c)** In order to reduce SE_{b_1} for men, we should choose our new sample to include men with a wider variety of lean body masses. (Note that just taking a larger sample will reduce SE_{b_1}; it is reduced even *more* if we choose subjects who will increase $s_{x,\mathrm{m}}$.)

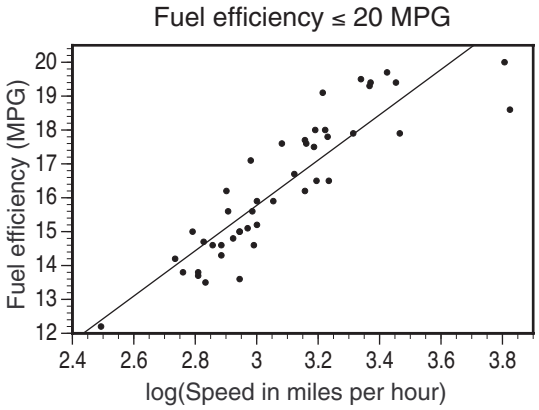
	b_1	SE_{b_1}	df	t^*	Interval
Women	24.026	4.174	10	2.2281	14.7257 to 33.3263
Men	16.75	10.20	5	2.5706	−9.4699 to 42.9699

10.64. Scatterplots, and portions of the Minitab outputs, are shown below. The equations are:

For all points, $\widehat{\mathrm{MPG}} = -7.796 + 7.8742\ \mathrm{LOGMPH}$
For speed $\leq$ 30 mph, $\widehat{\mathrm{MPG}} = -9.786 + 8.5343\ \mathrm{LOGMPH}$
For fuel efficiency $\leq$ 20 mpg, $\widehat{\mathrm{MPG}} = -4.282 + 6.6854\ \mathrm{LOGMPH}$

Students might make a number of observations about the effects of the restrictions; for example, the estimated coefficients (and their standard errors) change quite a bit.





Minitab output

```
- - - - - - - - - - - - - - - - - - -  All points  - - - - - - - - - - - - - - - - - - -
Predictor       Coef       Stdev    t-ratio        p
Constant      -7.796       1.155      -6.75    0.000
logMPH         7.8742      0.3541      22.24    0.000

s = 0.9995      R-sq = 89.5%     R-sq(adj) = 89.3%
- - - - - - - - - - - - - - - - -  Speed 30 mph or less  - - - - - - - - - - - - - - - - -
Predictor       Coef       Stdev    t-ratio        p
Constant      -9.786       1.862      -5.26    0.000
logMPH         8.5343      0.6154      13.87    0.000

s = 0.7600      R-sq = 83.5%     R-sq(adj) = 83.1%
- - - - - - - - - - - - - - -  Fuel efficiency 20 mpg or less  - - - - - - - - - - - - - -
Predictor       Coef       Stdev    t-ratio        p
Constant      -4.282       1.647      -2.60    0.013
logMPH         6.6854      0.5323      12.56    0.000

s = 0.9462      R-sq = 78.6%     R-sq(adj) = 78.1%
```