Chapter 11 Solutions

11.1. **(a)** The response variable is math GPA. **(b)** The number of cases is $n = 106$. **(c)** There were $p = 4$ explanatory variables. **(d)** The explanatory variables were SAT Math, SAT Verbal, class rank, and mathematics placement score.

11.2. **(a)** $\hat{y} = -1.4 + 2.6(4) - 2.3(2) = 4.4$. **(b)** No: We can compute predicted values for any values of x_1 and x_2 (although of course it helps if they are close to those in the data set). **(c)** This is determined by the coefficient of x_1: An increase of two unit in x_1 results in an increase of $(2.6)(2) = 5.2$ units in $\hat{y}$.

11.3. **(a)** The fact that the coefficients are all positive indicates that math GPA should increase when any explanatory variable increases. **(b)** With $n = 86$ cases and $p = 4$ variables, $\text{DFM} = p = 4$ and $\text{DFE} = n - p - 1 = 81$. **(c)** In the following table, each t statistic is the estimate divided by the standard error; the P-values are computed from a t distribution with $\text{df} = 81$. (The t statistic for the intercept was not required for this exercise, but is included for completeness.)

Variable	Estimate	SE	t	P
Intercept	−0.764	0.651	−1.1736	0.2440
SAT Math	0.00156	0.00074	2.1081	0.0381
SAT Verbal	0.00164	0.00076	2.1579	0.0339
HS rank	1.470	0.430	3.4186	0.0010
Bryant placement	0.889	0.402	2.2114	0.0298

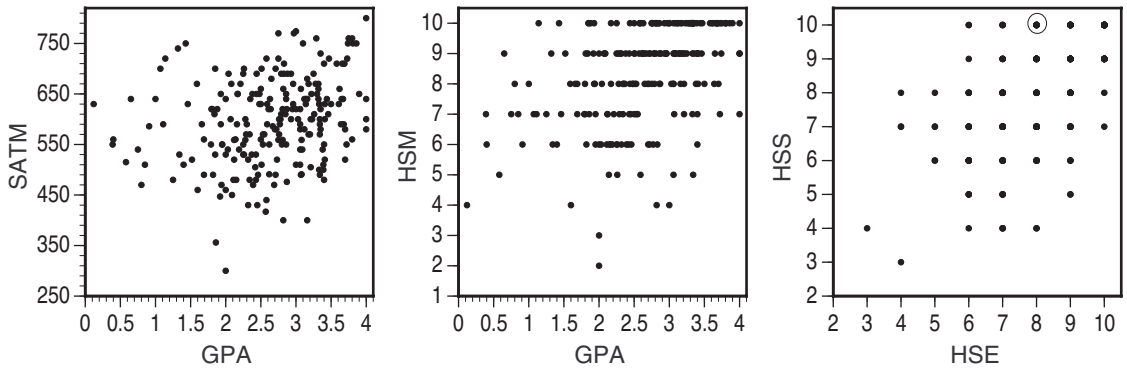
All four coefficients are significantly different from 0 (although the intercept is not).

11.4. The missing entries in the DF and SS columns can be found by noting that DFE + DFM = DFT and SSE + SSM = SST. The MS (mean square) entries are computed as SS divided by DF, and $F = \text{MSM}/\text{MSE}$. Comparison of $F = 2.5$ to an F distribution with df 5 and 60 gives $P \doteq 0.0402$, so we conclude the regression is significant. Finally, $R^2 = \dfrac{\text{SSM}}{\text{SST}} = \dfrac{175}{1015} \doteq 0.1724$.

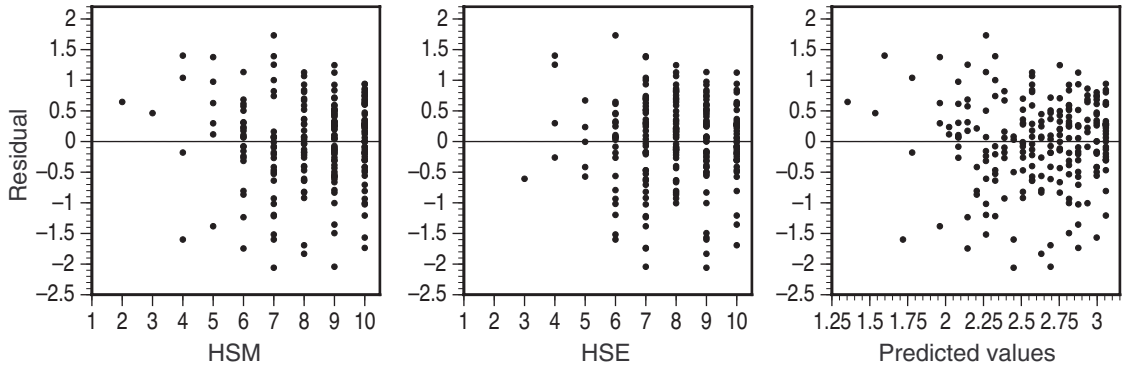
Source	DF	SS	MS	F
Model	5	175	35	2.5
Error	60	840	14	
Total	65	1015	15.62	

11.5. The correlations are found in Figure 11.3 and are summarized in the table on the right. Of the 15 possible scatterplots to be made from these six variables, three are shown below as examples. The pairs with the largest correlations are generally easy to pick out. The whole-number scale for high school grades causes point clusters in those scatterplots and makes it difficult to determine the strength of the association. For example, in the plot of HSS versus HSE below, the circled point represents 9 of the 224 students. One might guess that these three scatterplots show relationships of roughly equal strength, but because of the overlapping points, the correlations are quite different; from left to right, they are 0.2517, 0.4365, and 0.5794.

	SATM	SATV	HSM	HSS	HSE
GPA	0.2517	0.1145	0.4365	0.3294	0.2890
SATM		0.4639	0.4535	0.2405	0.1083
SATV			0.2211	0.2617	0.2437
HSM				0.5757	0.4469
HSS					0.5794



11.6. The regression equation is given in the Minitab output below. The whole-number scale for high school grades means that the predicted values also come in clusters. All but 21 students had both HSM and HSE above 5, so for all three plots, there are few residuals on the left half.



Minitab output

```
The regression equation is GPA = 0.624 + 0.183 HSM + 0.0607 HSE

Predictor       Coef       Stdev    t-ratio        p
Constant      0.6242      0.2917       2.14    0.033
HSM          0.18265     0.03196       5.72    0.000
HSE          0.06067     0.03473       1.75    0.082

s = 0.6996      R-sq = 20.2%     R-sq(adj) = 19.4%
```

11.7. 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 $t^*\text{SE}_{b_1}$, with df $= n - 3$ for (a) and (b), and df $= n - 4$ for (c) and (d).

	df	Coeff.	SE	t^*	Interval	t^*	Interval
(a)	27	10.8	2.4	2.052	5.8752 to 15.7248	2.0518	5.8756 to 15.7244
(b)	50	10.8	2.4	2.009	5.9784 to 15.6216	2.0086	5.9795 to 15.6205
(c)	26	10.8	2.4	2.056	5.8656 to 15.7344	2.0555	5.8667 to 15.7333
(d)	120	10.8	2.4	1.984	6.0384 to 15.5616	1.9799	6.0482 to 15.5518

11.8. For all four settings, the test statistic is $t = b_1/\text{SE}_{b_1} = 10.8/2.4 = 4.5$, with $\text{df} = n - 3$ for (a) and (b) and $\text{df} = n - 4$ for (c) and (d). All four P-values are very small (the first and third are 0.0001, and the other two are less than 0.00005), In all four cases, b_1 is significantly different from 0. (This is consistent with the confidence intervals from the previous exercise.)

11.9. (a) H_0 should refer to β_2 (the population coefficient) rather than b_2 (the estimated coefficient). **(b)** This sentence should refer to the *squared* multiple correlation. **(c)** A small P implies that *at least one coefficient* is different from 0.

11.10. (a) Multiple regression only assumes Normality of the error terms (residuals), not the explanatory variables. (The explanatory variables do not even need to be random variables.) **(b)** A small P-value tells us that the model is significant (useful for prediction) but does not measure its explanatory power (the accuracy of those predictions). The squared multiple correlation R^2 is a measure of explanatory power. **(c)** For example, if x_1 is significantly correlated with the response variable and x_2 is not, it might turn out that the coefficient of x_1 is *not* statistically significant, while the coefficient of x_2 *is* statistically significant.

Note: *For (c), the statement about the coefficient of x_2 is a paraphrase of the "Caution" on page 617 of the text. The statement about the coefficient of x_1 might be surprising to some students—and perhaps to some instructors—but a simple illustration confirms that it is true: Suppose that the response variable $y = ax_1 + b$ (with little or no error term), where all observed values of x_1 are positive, and the second explanatory variable is $x_2 = x_1^2$. The correlation between y and x_2 might be very large, but in a multiple regression model with x_1, the coefficient of x_2 will almost certainly not be significant.*

11.11. (a) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_7 x_{i7} + \epsilon_i$, where $i = 1, 2, \ldots, 140$, and ϵ_i are independent $N(0, \sigma)$ random variables. **(b)** The sources of variation are model ($\text{df} = p = 7$), error ($\text{df} = n - p - 1 = 132$), and total ($\text{df} = n - 1 = 139$).

11.12. (a) With $n = 73$ and $p = 5$, the degrees of freedom in the ANOVA table are $\text{DFM} = p = 5$, $\text{DFE} = n - p - 1 = 67$, and $\text{DFT} = n - 1 = 72$. With the first two degrees of freedom, we can find $\text{MSM} = \frac{\text{SSM}}{\text{DFM}} = 2.82$ and $\text{MSE} = \frac{\text{SSE}}{\text{DFE}} = 1.5$ and then compute $F = \frac{\text{MSM}}{\text{MSE}} = 1.88$. **(b)** This F statistic has df 5 and 67. **(c)** Comparing to the $F(5, 60)$ critical values in Table E, we note that $F < 1.95$, so $P > 0.10$. (Software gives 0.1095.) **(d)** This regression explains $R^2 = \frac{\text{SSM}}{\text{SST}} \doteq 12.3\%$ of the variation in the response variable.

Source	DF	SS	MS	F
Model	5	14.1	2.82	1.88
Error	67	100.5	1.5	
Total	72	114.6	$1.591\overline{6}$	

11.13. (a) We see from the table of coefficients that there were $p = 5$ explanatory variables. The F statistic has df 5 and 524, so $n - p - 1 = 524$, meaning that $n = 530$ children. **(b)** We were given $R = 0.44$, so the regression explains $R^2 \doteq 19.36\%$ of the variation in percent fat mass. **(c)** Based on the positive coefficients, predicted fat mass is higher for females, those who take in higher percents of energy at dinner, and children of parents with higher BMIs. The one negative coefficient tells us that predicted fat mass is higher for those with underreported intake (low values of EI/predicted BMR) and lower for those who overreported intake. **(d)** With $\text{df} = 524$, the appropriate critical value for a 95% confidence

interval is $t^* = 1.965$, so a 95% confidence interval for β_2 is $0.08 \pm (1.965)(0.02) \doteq 0.04$ to 0.12. Therefore, when that explanatory variable changes by 5%, percent fat mass changes by between 0.20 and 0.60.

11.14. No (or at least, not necessarily). It is possible that, although no individual coefficient is significant, the whole group (or some subset) is. Recall that the t tests "assess the significance of each predictor variable assuming that all other predictors are included in the regression equation." If one variable is removed from the model (because its t statistic is not significant), we can no longer use the other t statistics to draw conclusions about the remaining coefficients.

11.15. (a) The hypotheses and test results are:

Hypotheses	t	P	Conclusion
H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$	4.55	$P < 0.001$	Reject H_0; GPA is significant
H_0: $\beta_2 = 0$ vs. H_a: $\beta_2 \neq 0$	2.69	$P < 0.01$	Reject H_0; popularity is significant
H_0: $\beta_3 = 0$ vs. H_a: $\beta_3 \neq 0$	2.69	$P < 0.01$	Reject H_0; depression is significant

(b) $b_1 < 0$, so marijuana use decreases with increasing GPA; b_2 and b_3 are positive, so marijuana use increases with popularity and depression. **(c)** The numbers 3 and 85 are the degrees of freedom of the F statistic ($p = 3$ explanatory variables and $n - p - 1 = 85$ degrees of freedom left over after estimating the four regression coefficients). **(d)** H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ is rejected in favor of H_a: at least one β_i is nonzero. **(e)** Students may have lied (or erred) in their responses, which may make our conclusions unreliable. This risk is especially great because all four variables were measured with questionnaire responses. **(f)** We cannot assume that students are the same everywhere; for example, these conclusions might not hold for nonsuburban students or those who live in other states.

11.16. The hypotheses are essentially the same as in the previous exercise, so they are not restated here. We see from the test results that, as for marijuana, cigarette and alcohol use decreases with increasing GPA and increases with increasing popularity and depression. Cocaine use also decreases as GPA increases, but it also drops when popularity increases, and is not significantly affected by depression. The sizes of the coefficients also tell us the relative impact of the explanatory variables; for example, the effect of increasing GPA is similar for marijuana and cocaine but considerably less for cigarettes and alcohol. Similarly, the sizes of R^2 and F give some indication of the predictive usefulness of these regression formulas.

Note: *The problem of doing several significance tests deserves mention here. A procedure such as the Bonferroni method (first mentioned in Chapter 6 and discussed further in Chapter 12) would say that, as we are performing 12 t tests, we should have $P < 0.05/12 \doteq 0.0042$ before we declare a result significant at the 5% level. By this standard, considerably fewer of the individual coefficients would be significantly different from 0.*

11.17. (a) Because this coefficient is negative, U.S. subjects are *less* willing to pay more. (Specifically, for a U.S. and U.K. subject for which all other characteristics are the same, we predict that the U.K. subject would be willing to pay about 0.2304 units more than the U.S. subject.) **(b)** We have $t = -4.196$ and df ≈ 1800 (the question does not clearly state

how many explanatory variables are being used). For testing H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$, Table D tells us that $P < 0.001$. (In fact, it is much smaller than that—software gives $P \doteq 0.00003$.) **(c)** The U.K. response rate is much lower than the U.S. rate, so we might have missed a significant portion of U.K. opinion. Also, "don't know" responses are not much better than no response at all.

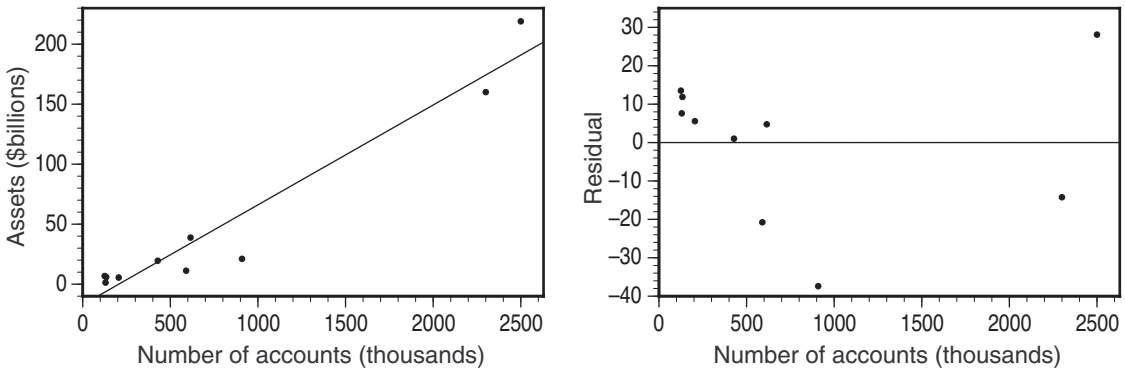
11.18. (a) No—at least, not entirely. Assuming that we had a standard regression model, we know that the fitted regression equation is $\hat{y} = b_0 + 1.02x_1 + 0.96x_2 + 0.30x_3$; the constant b_0 is not given. **(b)** No: With a multiple regression model, it can happen that some individual coefficients are not significantly different from 0 because those tests "assess the significance of each predictor variable assuming that all other predictors are included in the regression equation." **(c)** With $n = 282$ and $p = 3$, the degrees of freedom for the F statistic are 3 and 278. The P-value for this F statistic is very small; 45.64 is much larger than 5.63, the largest critical value listed for an $F(3, 200)$ distribution in Table E. **(d)** The regression explains $R^2 = 33\%$ of the variation in exercise enjoyment. **(e)** It seems likely that men and women differ in how they respond to exercise in general (and aerobic dance in particular), so we should be reluctant to extend these results to men.

11.19. (a) The regression equation is $\widehat{\text{Score}} = 3.33 + 0.82\ \text{Unfav} + 0.57\ \text{Fav}$. **(b)** Because $P < 0.01$, we reject H_0: $\beta_1 = \beta_2 = 0$ in favor of H_a: at least one of β_1 and β_2 is nonzero. **(c)** The estimates of β_0, β_1, and β_2 are all significantly different from 0 (all have $P < 0.01$). **(d)** The t statistics have $\text{df} = n - p - 1 = 152 - 2 - 1 = 149$.

11.20. (a) The regression equation is $\widehat{\text{Score}} = 3.96 + 0.86\ \text{Unfav} + 0.66\ \text{Fav}$. **(b)** Because $P < 0.01$, we reject H_0: $\beta_1 = \beta_2 = 0$ in favor of H_a: at least one of β_1 and β_2 is nonzero. **(c)** The estimates of β_0, β_1, and β_2 are all significantly different from 0 (all have $P < 0.01$). **(d)** The t statistics have $\text{df} = n - p - 1 = 162 - 2 - 1 = 159$.

11.21. All coefficients are positive, so the associations are positive, as expected. The unfavorable coefficients are larger than (and their t statistics at least as large as) those for favorable nutrients, so unfavorable nutrients do have a stronger effect.

11.22. (a) The scatterplot (below, left) suggests a positive association with a slight curve. **(b)** The regression equation is $\widehat{\text{ASSETS}} = -17.121 + 0.0832\ \text{ACCTS}$, with $s \doteq 20.19$ and $R^2 = 0.938$. The slope is significantly different from 0: $t = 10.96$ (df = 8) for which $P < 0.0005$. **(c)** The residual scatterplot (below, right) also suggests a curved relationship: The residuals are negative in the middle, and positive on the left and right ends. **(d)** The regression equation is $\widehat{\text{ASSETS}} = 7.608 - 0.00457\ \text{ACCTS} + 0.00003361\ \text{ACCTS}^2$, with $s = 12.41$ and $R^2 = 0.979$. The coefficient of ACCTS is *not* significantly different from 0 ($t = -0.19$, $P = 0.853$), but the coefficient of ACCTS^2 is ($t = 3.76$, $P = 0.007$). (The constant b_0 is also not significantly different from 0.) The plot of residuals against accounts (or squared accounts) is not shown but reveals no obvious pattern.



Minitab output

```
- - - - - - - - - - - - - - - - - - -  Linear model  - - - - - - - - - - - - - - - - - -
The regression equation is assets = -17.1 + 0.0832 accts

Predictor       Coef       Stdev    t-ratio        p
Constant     -17.121       8.778      -1.95    0.087
accts       0.083205    0.007592      10.96    0.000

s = 20.19       R-sq = 93.8%     R-sq(adj) = 93.0%
- - - - - - - - - - - - - - - - - -  Quadratic model  - - - - - - - - - - - - - - - - - -
The regression equation is assets = 7.61 - 0.0046 accts + 0.000034 acctsqd

Predictor       Coef       Stdev    t-ratio        p
Constant       7.608       8.503       0.89    0.401
accts       -0.00457     0.02378      -0.19    0.853
acctsqd   0.00003361  0.00000893       3.76    0.007

s = 12.41       R-sq = 97.9%     R-sq(adj) = 97.3%
```

11.23. With the altered quadratic term, the regression equation is:

$$\widehat{\text{ASSETS}} = -13.558 - 0.04877\ \text{ACCTS} + 0.00003361(\text{ACCTS} - 793.6)^2$$

with $s = 12.41$ and $R^2 = 0.979$. The coefficient of the new quadratic term and its standard error (and therefore its t statistics and P-value) are the same as for the old quadratic term. Also, the values of s and R^2 are unchanged. The constant term and the coefficient of ACCTS are now both significantly different from 0.

Note: *The estimates b_0 and b_1 for the new model can be obtained from those in the old model in a fairly simple way. Note that $(ACCTS - 793.6)^2 = ACCTS^2 - 1587.2ACCTS + 793.6^2$. Therefore,* new $b_0 =$ old $b_0 - 793.6^2 b_2$ *and* new $b_1 =$ old $b_1 + 1587.2 b_2$.

Minitab output

```
- - - - - - - - - - - - - -  Modified quadratic model  - - - - - - - - - - - - - -
The regression equation is assets = -13.6 + 0.0488 accts + 0.000034 accts2

Predictor       Coef       Stdev    t-ratio        p
Constant     -13.558       5.479      -2.47    0.043
accts        0.04877     0.01027       4.75    0.000
accts2    0.00003361  0.00000893       3.76    0.007

s = 12.41       R-sq = 97.9%     R-sq(adj) = 97.3%
```

11.24. Students might identify the two outliers differently. Some might choose the two points with 2300 and 2500 thousand accounts (Charles Schwab and Fidelity). If they do so, the regression results are shown below. (The quadratic regression was done using both of the models from the previous two exercises.) Without these points, the regression is less useful for prediction; the coefficients are, at best, on the borderline of significance, and R^2 is much smaller than before.

Minitab output

```
- - - - - - - - - -   Linear model without Schwab and Fidelity   - - - - - - - - -
The regression equation is assets = 2.13 + 0.0297 accts

Predictor       Coef       Stdev    t-ratio        p
Constant       2.131       5.787       0.37    0.725
accts        0.02967     0.01211       2.45    0.050

s = 9.369       R-sq = 50.0%     R-sq(adj) = 41.7%
 - - - - - - - - -   Quadratic model without Schwab and Fidelity   - - - - - - - -
The regression equation is assets = -6.74 + 0.0887 accts - 0.000063 acctsqd

Predictor       Coef       Stdev    t-ratio        p
Constant      -6.737       9.204      -0.73    0.497
accts        0.08874     0.05016       1.77    0.137
acctsqd  -0.00006251  0.00005163      -1.21    0.280

s = 9.025       R-sq = 61.4%     R-sq(adj) = 45.9%
 - - - - - - -   Modified quadratic model without Schwab and Fidelity   - - - - - - -
The regression equation is assets = 32.6 - 0.0105 accts - 0.000063 accts2

Predictor       Coef       Stdev    t-ratio        p
Constant       32.63       25.80       1.26    0.262
accts       -0.01048     0.03516      -0.30    0.778
accts2   -0.00006251  0.00005163      -1.21    0.280

s = 9.025       R-sq = 61.4%     R-sq(adj) = 45.9%
```

Other students might identify as outliers the two points that had the largest residuals in the original linear fit (Charles Schwab and E*Trade). Note, though, that these two points did *not* have the largest residuals for the quadratic fit, so this choice of outliers is perhaps not as good as the first. Without these two points, R^2 rises slightly for all models, but we see that the coefficients for the quadratic terms are not significantly different from 0—the linear model appears to be the best in this case.

Minitab output

```
- - - - - - - - - -   Linear model without Schwab and E*Trade   - - - - - - - - -
The regression equation is assets = -9.66 + 0.0721 accts

Predictor       Coef       Stdev    t-ratio        p
Constant      -9.659       4.582      -2.11    0.080
accts       0.072072    0.005166      13.95    0.000

s = 9.978       R-sq = 97.0%     R-sq(adj) = 96.5%
```

```
– – – – – – – – –  Quadratic model without Schwab and E*Trade  – – – – – – – – –
The regression equation is assets = -0.46 + 0.0345 accts + 0.000015 acctsqd

Predictor       Coef       Stdev    t-ratio        p
Constant      -0.463       6.639      -0.07    0.947
accts        0.03451     0.02219       1.56    0.181
acctsqd   0.00001533  0.00000887       1.73    0.144

s = 8.648       R-sq = 98.1%     R-sq(adj) = 97.4%
– – – – – – –  Modified quadratic model without Schwab and E*Trade  – – – – – – –
The regression equation is assets = -10.1 + 0.0588 accts + 0.000015 accts2

Predictor       Coef       Stdev    t-ratio        p
Constant     -10.120       3.980      -2.54    0.052
accts       0.058847    0.008865       6.64    0.001
accts2    0.00001533  0.00000887       1.73    0.144

s = 8.648       R-sq = 98.1%     R-sq(adj) = 97.4%
```

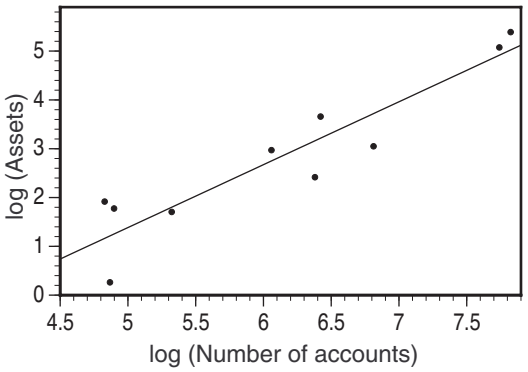
11.25. The plot of log-assets against log-accounts (below, left) appears to be reasonably linear. The regression equation is $\hat{y} = -5.058 + 1.2885x$, with $s = 0.6291$, $R^2 = 0.858$, $t = 6.96$, and $P < 0.0005$. A plot of residuals against log-accounts (below, right) suggests no particular pattern.

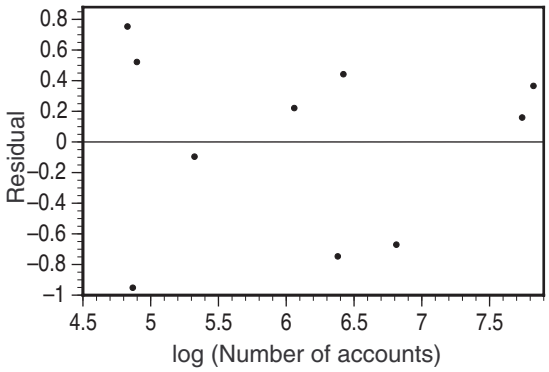
Minitab output

```
The regression equation is logasset = -5.06 + 1.29 logaccts

Predictor       Coef       Stdev    t-ratio        p
Constant      -5.058       1.150      -4.40    0.002
logaccts      1.2885      0.1852       6.96    0.000

s = 0.6291      R-sq = 85.8%     R-sq(adj) = 84.0%
```





11.26. The three regression equations and associated results are:

	$\widehat{\text{ASSETS}} =$	s	R^2	t	P
(a)	$-17.12 + 0.0832$ ACCTS	20.19	0.938	10.96	$P < 0.0005$
(b)	$-19.90 + 7.680$ MSHARE	50.54	0.609	3.53	$P = 0.008$
(c)	$-21.45 + 0.0756$ ACCTS $+ 1.158$ MSHARE	20.52	0.944	6.44	$P < 0.0005$
				0.86	$P = 0.418$

(d) Between the three models, (a) seems to be the best. Based on R^2, it is clearly better than (b) and only slightly worse than (c). The model in (c) is more complicated, and the market-share component does not make a significant contribution when the number of accounts is included in the model.

Minitab output

```
- - - - - - - - - - - - - - - -   Number of accounts   - - - - - - - - - - - - - - - -
The regression equation is assets = -17.1 + 0.0832 accts

Predictor       Coef       Stdev    t-ratio        p
Constant     -17.121       8.778      -1.95    0.087
accts       0.083205    0.007592      10.96    0.000

s = 20.19       R-sq = 93.8%     R-sq(adj) = 93.0%
- - - - - - - - - - - - - - - - - -  Market share  - - - - - - - - - - - - - - - - - -
The regression equation is assets = -19.9 + 7.68 mshare

Predictor       Coef       Stdev    t-ratio        p
Constant      -19.90       25.22      -0.79    0.453
mshare         7.680       2.177       3.53    0.008

s = 50.54       R-sq = 60.9%     R-sq(adj) = 56.0%
- - - - - - - - - - - - - - - - - - - - -  Both  - - - - - - - - - - - - - - - - - - - -
The regression equation is assets = -21.5 + 0.0756 accts + 1.16 mshare

Predictor       Coef       Stdev    t-ratio        p
Constant      -21.45       10.24      -2.09    0.074
accts        0.07559     0.01173       6.44    0.000
mshare         1.158       1.344       0.86    0.418

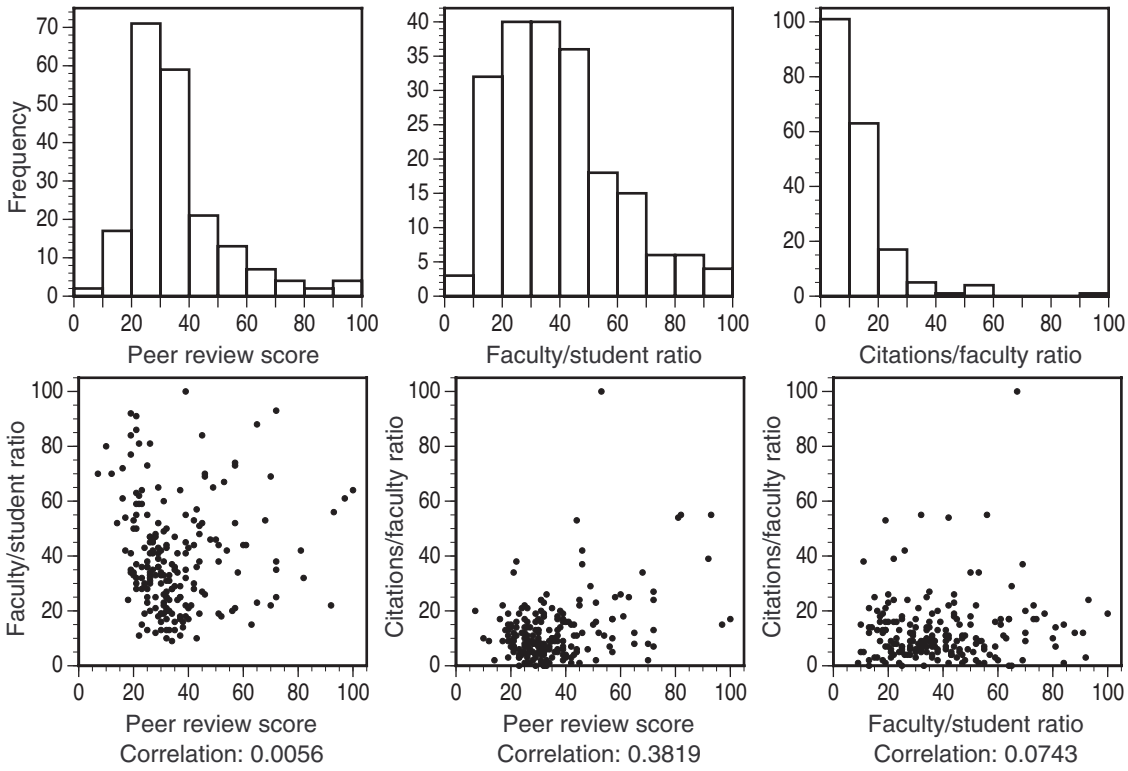
s = 20.52       R-sq = 94.4%     R-sq(adj) = 92.7%
```

11.27. (a) Histograms of the three explanatory variables are given below. All three distributions are right-skewed, especially CtoF, which has a high outlier of 100; the next largest value is 55 (see the note below). Student choices of summary statistics may vary; five-number summaries are a good choice because of the skewness, but some may also give means and standard deviations.

Variable	n	$\bar{x}$	s	Min	Q_1	M	Q_3	Max
PEER	200	35.67	16.22	7	25.0	32.0	41.5	100
FtoS	200	39.33	19.61	9	24.5	35.5	50.0	100
CtoF	197	12.57	11.84	0	5.0	10.0	16.0	100

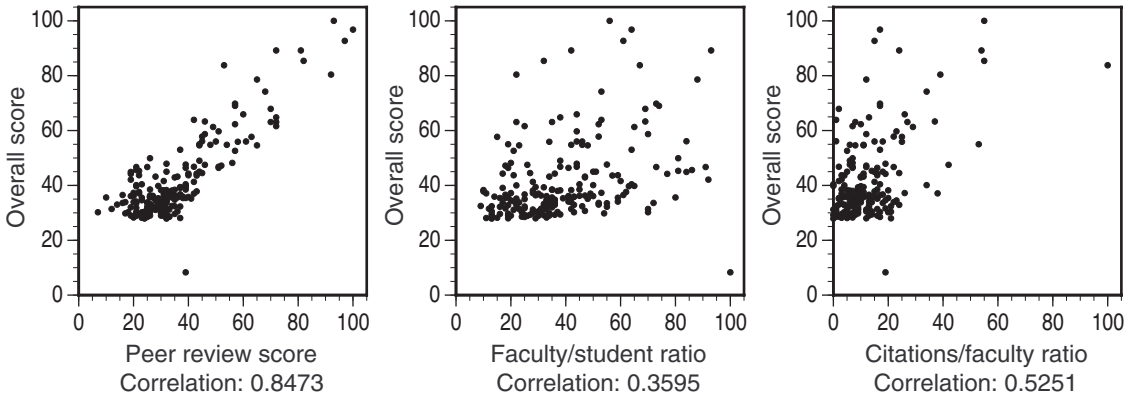
Notice especially how the skewness is apparent in the five-number summaries.
(b) Correlation coefficients are given below the scatterplots; only the correlation between PEER and CtoF is substantial. If students elect to remove the outlier 100 from the CtoF data, the correlation between PEER and CtoF rises slightly to 0.4040, while the correlation between FtoS and CtoF falls to 0.0229.

Note: *Using the 1.5 × IQR criterion for outliers, there would be* many *outliers: the 13 peer-review scores of 68 or more, the 4 FtoS ratios of 91 or more, and the 11 CtoF ratios of 34 or more. Under these circumstances—when there are so many suspected "outliers," and (except for the 100 CtoF ratio) most are fairly evenly spread out—it seems best to disregard the 1.5 × IQR criterion.*



11.28. (a) The scatterplot of peer review score looks much more linear than the other two. Apparent in all three scatterplots is a low outlier among the overall scores; that point fits fairly well with the patterns of the PEER and CtoF scatterplots but weakens the association with FtoS. (This outlier is also easily noted in a stemplot or histogram of the overall scores.) **(b)** The correlations are given below the scatterplots. If the outlier in CtoF noted in the solution to the previous exercise is removed, that correlation drops to 0.5010 (because the presence of that outlier strengthens the association). If we remove the low outlier in overall score, all three correlations increase (to 0.8611, 0.4103, and 0.5389); not surprisingly, this point has the greatest effect on the correlation with FtoS.

Note: *The fact that the scatterplots do not all suggest linear associations does not mean that a multiple regression is inappropriate. Even if the data exactly fit a multiple regression model, the pairwise scatterplots will not necessarily appear to be linear.*



11.29. **(a)** The model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, where ϵ_i are independent $N(0, \sigma)$ random variables. **(b)** The regression equation is:

$$\widehat{\text{Score}} = 3.62 + 0.690\ \text{PEER} + 0.254\ \text{FtoS} + 0.259\ \text{CtoF}$$

(c) For the confidence intervals, take $b_i \pm t^* \text{SE}_{b_i}$, with $t^* = 1.9723$ (for df $= 193$). These intervals have been added to the Minitab output below. None of the intervals contain 0 because all coefficients are significantly different from 0. **(d)** The regression explains $R^2 \doteq 88.1\%$ of the variation in overall score. The estimate of σ is $s \doteq 5.108$.

Minitab output

```
The regression equation is
Score = 3.62 + 0.690 PEER + 0.254 FtoS + 0.259 CtoF

Predictor       Coef       Stdev    t-ratio        p    95% confidence interval
Constant       3.619       1.142       3.17    0.002
PEER         0.69007     0.02432      28.38    0.000    0.6418 to 0.7383
FtoS         0.25371     0.01902      13.34    0.000    0.2160 to 0.2914
CtoF         0.25917     0.03341       7.76    0.000    0.1929 to 0.3255

s = 5.108       R-sq = 88.1%     R-sq(adj) = 87.9%
```

11.30. **(a)** Between GPA and IQ, $r = 0.634$ (straight-line regression explains $r^2 = 40.2\%$ of the variation in GPA). Between GPA and self-concept, $r = 0.542$ (straight-line regression explains $r^2 = 29.4\%$ of the variation in GPA). Since gender is categorical, the correlation between GPA and gender is not meaningful. **(b)** The model is $\text{GPA} = \beta_0 + \beta_1\,\text{IQ} + \beta_2\,\text{SC} + \epsilon_i$, where ϵ_i are independent $N(0, \sigma)$ random variables. **(c)** Regression gives the equation $\widehat{\text{GPA}} = -3.88 + 0.0772\ \text{IQ} + 0.0513\ \text{SC}$. Based on the reported value of R^2, the regression explains 47.1% of the variation in GPA. (So the inclusion of self-concept only adds about 6.9% to the variation explained by the regression.) **(d)** We test H_0: $\beta_2 = 0$ vs. H_a: $\beta_2 \neq 0$. The test statistic $t = 3.14$ (df $= 75$) has $P = 0.002$; we conclude that the coefficient of self-concept is not 0.

Minitab output

```
The regression equation is GPA = -3.88 + 0.0772 IQ + 0.0513 SelfCcpt

Predictor       Coef       Stdev    t-ratio        p
Constant      -3.882       1.472      -2.64    0.010
IQ           0.07720     0.01539       5.02    0.000
SelfCcpt     0.05125     0.01633       3.14    0.002

s = 1.547       R-sq = 47.1%     R-sq(adj) = 45.7%
```

11.31. (a) All distributions are skewed to varying degrees—GINI and CORRUPT to the right, the other three to the left. CORRUPT, DEMOCRACY, and LIFE have the most skewness. Student choices of summary statistics may vary; five-number summaries are a good choice because of the skewness, but some may also give means and standard deviations.

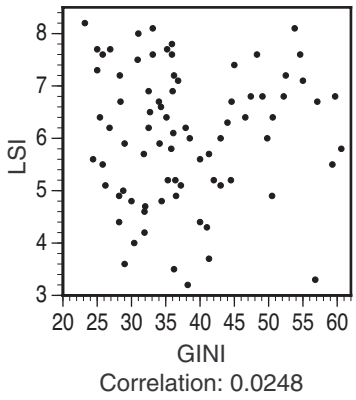
Variable	n	$\bar{x}$	s	Min	Q_1	M	Q_3	Max
LSI	80	6.0638	1.2690	3.2	5.1	6.2	7.1	8.2
GINI	80	37.9725	9.6360	23.2	30.95	35.95	44.25	60.6
CORRUPT	80	4.6950	2.4447	1.7	2.75	3.5	6.75	9.7
DEMOCRACY	80	4.1813	1.6862	0.5	3.0	4.5	5.5	6.0
LIFE	80	71.3641	9.3755	39.29	69.835	72.96	78.2	81.25

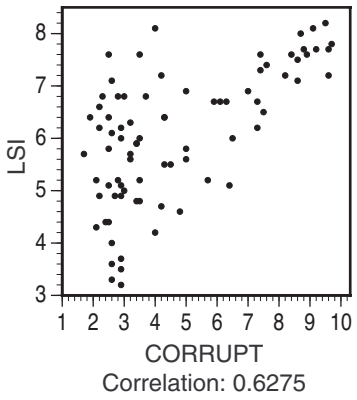
Notice especially how the skewness is apparent in the five-number summaries.

(b) Correlation coefficients are given below the scatterplots. LSI and GINI have a very small correlation (0.0248); all other pairs have moderate to large correlations. GINI is negatively correlated to the other three variables, while all other correlations are positive.

LSI

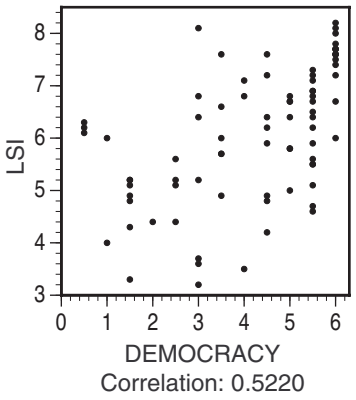
3	23
3	567
4	02344
4	6788999
5	01112222
5	5566778899
6	000122234444
6	567777888899
7	1122234
7	5666667778
8	0112

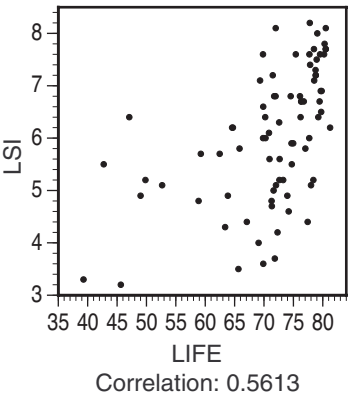




GINI

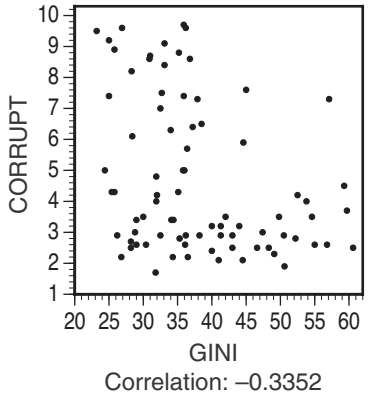
2	34
2	555556668888899
3	00011112222334444
3	55555566666667788
4	00111233444
4	567899
5	002234
5	56799
6	0

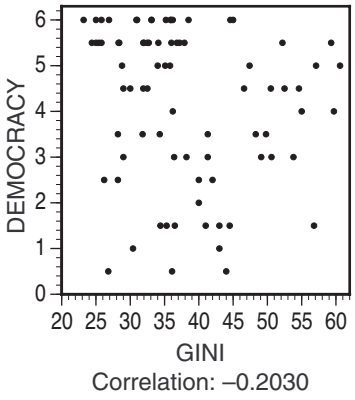




CORRUPT

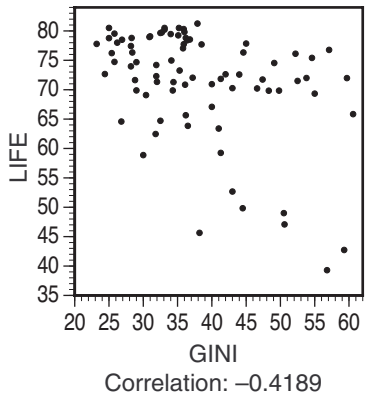
1	79
2	1122234
2	55555666667889999999
3	00222444
3	55557
4	0022333
4	58
5	000
5	79
6	134
6	5
7	03344
7	56
8	24
8	66789
9	12
9	5667

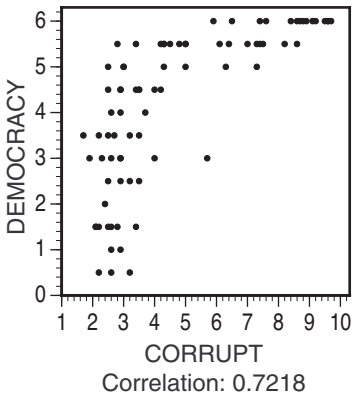




DEMOCRACY

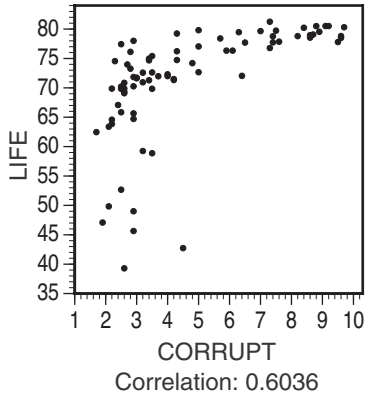
0	555
1	00
1	5555555
2	0
2	5555
3	0000000
3	555555
4	000
4	55555555
5	0000000
5	55555555555555555
6	000000000000000

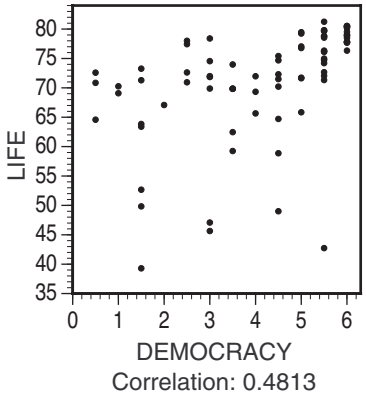




LIFE

3	9
4	2
4	5799
5	2
5	89
6	23344
6	557999999
7	000011111111222223344444
7	566666777777888888889999999
8	000001





11.32. The four regression equations are:

(**a**) $\widehat{\text{LSI}} = \quad 5.94 + 0.0033\ \text{GINI}$

(**b**) $\widehat{\text{LSI}} = -1.97 + 0.0408\ \text{GINI} + 0.0909\ \text{LIFE}$

(**c**) $\widehat{\text{LSI}} = -1.38 + 0.0398\ \text{GINI} + 0.0677\ \text{LIFE} + 0.2630\ \text{DEMOCRACY}$

(**d**) $\widehat{\text{LSI}} = -0.73 + 0.0470\ \text{GINI} + 0.0505\ \text{LIFE} + 0.0588\ \text{DEMOCRACY} + 0.2465\ \text{CORRUPT}$

Minitab output (on the next page) gives the values of R^2 for each regression and highlights non-significant P-values. We note that GINI does not contribute significantly to the first

model but is significant in every other model, and DEMOCRACY is not significant in the last model, even though it was significant in the second-to-last model. (Roughly speaking, this means that whatever information DEMOCRACY contributed to the model, CORRUPT contains that same information but contributes it more efficiently than does DEMOCRACY. Recall from the previous solution that DEMOCRACY and CORRUPT had the highest correlation.)

A full analysis of the residuals for each regression would require a total of 10 scatterplots. Shown on the next page are stemplots of the regressions for all four regressions (which show no major deviations from Normality). Also shown are two scatterplots that suggest a possible problem with the assumptions: Both the residuals from the third regression (against DEMOCRACY) and those from the fourth regression (against CORRUPT) show a hint of curvature.

Minitab output

Model 1: GINI

LSI = 5.94 + 0.0033 GINI

Predictor	Coef	Stdev	t-ratio	p
Constant	5.9398	0.5838	10.17	0.000
GINI	0.00327	0.01491	0.22	0.827

s = 1.277 R-sq = 0.1% R-sq(adj) = 0.0%

Model 2: GINI and LIFE

LSI = - 1.97 + 0.0408 GINI + 0.0909 LIFE

Predictor	Coef	Stdev	t-ratio	p
Constant	-1.972	1.265	-1.56	0.123
GINI	0.04076	0.01314	3.10	0.003
LIFE	0.09092	0.01351	6.73	0.000

s = 1.020 R-sq = 37.1% R-sq(adj) = 35.4%

Model 3: GINI, LIFE, and DEMOCRACY

LSI = - 1.38 + 0.0398 GINI + 0.0677 LIFE + 0.263 DEMOCRACY

Predictor	Coef	Stdev	t-ratio	p
Constant	-1.383	1.185	-1.17	0.247
GINI	0.03983	0.01221	3.26	0.002
LIFE	0.06774	0.01406	4.82	0.000
DEMOCRACY	0.26304	0.07208	3.65	0.000

s = 0.9468 R-sq = 46.5% R-sq(adj) = 44.3%

Model 4: All four variables

LSI = - 0.73 + 0.0470 GINI + 0.0505 LIFE + 0.0588 DEMOCRACY + 0.246 CORRUPT

Predictor	Coef	Stdev	t-ratio	p
Constant	-0.730	1.104	-0.66	0.510
GINI	0.04700	0.01139	4.13	0.000
LIFE	0.05054	0.01369	3.69	0.000
DEMOCRACY	0.05881	0.08499	0.69	0.491
CORRUPT	0.24648	0.06414	3.84	0.000

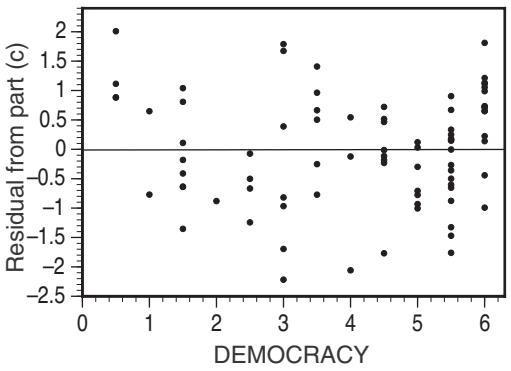
s = 0.8711 R-sq = 55.3% R-sq(adj) = 52.9%

First

−2	885
−2	430
−1	8766
−1	43222110
−0	999888865
−0	44333211100
0	0111223334
0	556666667889
1	00112344
1	5555666799
2	01

Second

−2	5
−2	
−1	99875
−1	443111100
−0	99776666655
−0	44332221110000
0	001113334444
0	5555566888999
1	000012223344
1	5
2	01

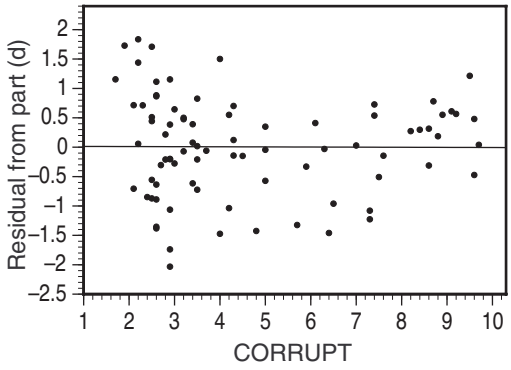


Third

−2	
−2	20
−1	776
−1	4320
−0	9998887777666665
−0	444322221111000
0	011111122334
0	55566666777788999
1	0011124
1	678
2	0

Fourth

−2	
−2	0
−1	7
−1	4443332000
−0	9888776555
−0	4333222211110000
0	0000011222333344444
0	5555556677777888
1	1124
1	5778
2	



11.33. **(a)** The coefficients, standard errors, t statistics, and P-values are given in the Minitab output shown with the solution to the previous exercise. **(b)** Student observations will vary. For example, the t statistic for the GINI coefficient grows from $t = 0.22$ ($P = 0.827$) to $t = 4.13$ ($P < 0.0005$). The DEMOCRACY t is 3.65 in the third model ($P < 0.0005$) but drops to 0.69 ($P = 0.491$) in the fourth model. **(c)** A good choice is to use GINI, LIFE, and CORRUPT: All three coefficients are significant, and $R^2 = 55.0\%$ is nearly the same as the fourth model from previous exercise. However, a scatterplot of the residuals versus CORRUPT (not shown) still looks quite a bit like the second scatterplot shown in the previous solution, suggesting a slightly curved relationship, which would violate the assumptions of our model.

Minitab output

```
LSI = - 0.74 + 0.0479 GINI + 0.0518 LIFE + 0.274 CORRUPT

Predictor       Coef       Stdev    t-ratio        p
Constant      -0.737       1.100      -0.67    0.505
GINI         0.04793     0.01127       4.25    0.000
LIFE         0.05175     0.01353       3.83    0.000
CORRUPT      0.27423     0.04988       5.50    0.000

s = 0.8681      R-sq = 55.0%     R-sq(adj) = 53.2%
```

11.34. (a) Stemplots (below) show that all four variables are right-skewed to some degree. The means and standard deviations are given in the table on the right. **(b)** Correlations and scatterplots (below) show that all six pairs of variables are positively associated. The strongest association is between VO+ and VO– and the weakest is between OC and VO–.

	Mean	SD
VO+	985.806	579.858
VO–	889.194	427.616
OC	33.416	19.610
TRAP	13.248	6.528

VO+

0	23
0	4444555
0	66667
0	888899
1	011
1	23
1	5
1	66
1	
2	
2	22
2	5

VO–

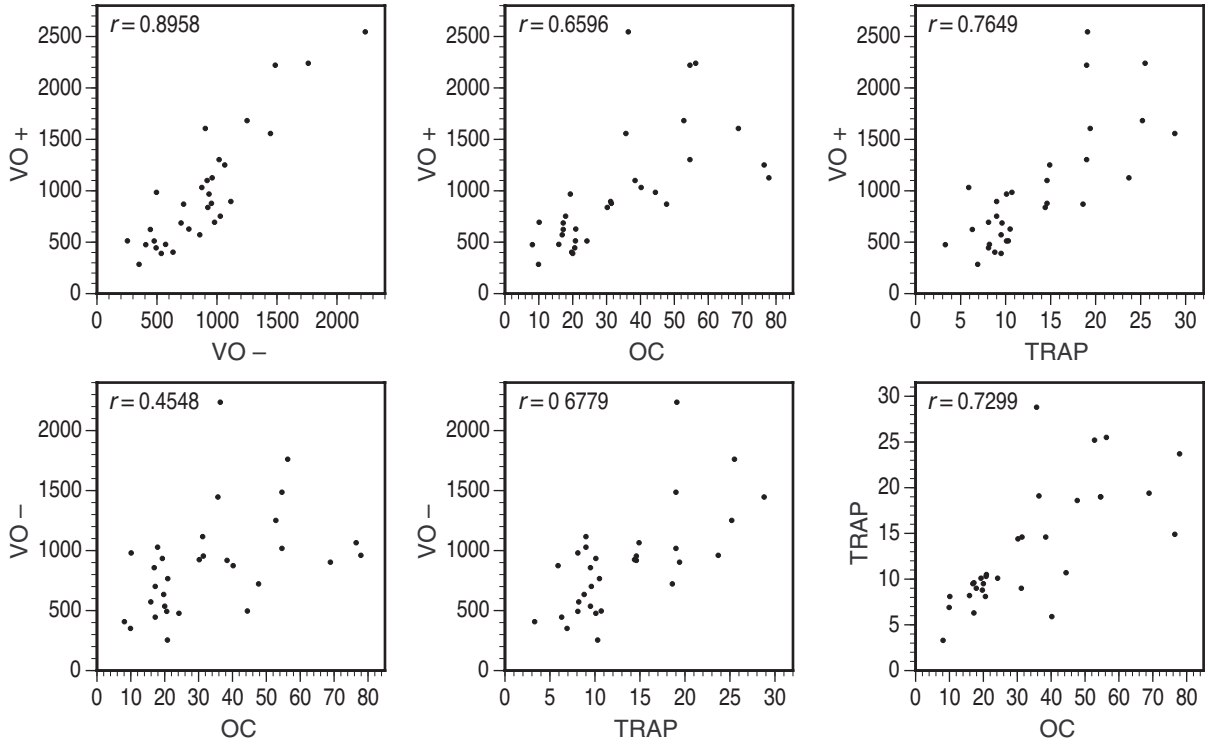
0	23
0	4444455
0	6777
0	889999999
1	0001
1	2
1	44
1	7
1	
2	
2	2

OC

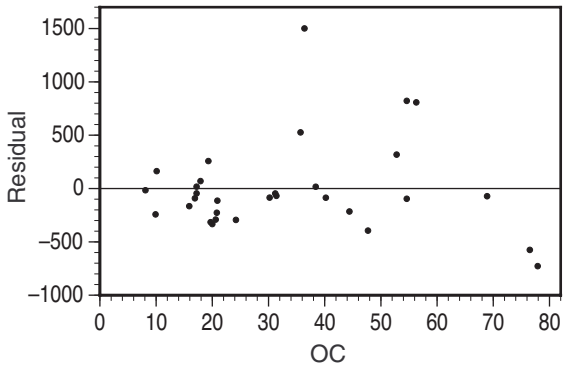
0	89
1	0
1	5677799
2	00004
2	
3	011
3	568
4	04
4	7
5	244
5	6
6	
6	8
7	
7	67

TRAP

0	3
0	5
0	66
0	888899999
1	00000
1	
1	4444
1	
1	89999
2	
2	3
2	55
2	
2	8



11.35. **(a)** See the previous solution for the scatterplot, which suggests greater variation in VO+ for large OC. The regression equation is $\widehat{\text{VOplus}} = 334.0 + 19.505\ \text{OC}$, with $s \doteq 443.3$ and $R^2 \doteq 0.435$; the test statistic for the slope is $t = 4.73$ ($P < 0.0005$), so we conclude the slope is not zero. The plot of residuals against OC suggests a slight downward curve on the right end, as well as the increasing scatter as OC increases. The residuals are also somewhat right-skewed. A stemplot and Normal quantile plot of the residuals are not shown here but could be included as part of the analysis. **(b)** The regression equation is $\widehat{\text{VOplus}} = 57.7 + 6.415\ \text{OC} + 53.87\ \text{TRAP}$, with $s \doteq 376.3$ and $R^2 \doteq 0.607$. The coefficient of OC is not significantly different from 0 ($t = 1.25$, $P = 0.221$), but the coefficient of TRAP is ($t = 3.50$, $P = 0.002$). This is consistent with the correlations found in the solution to Exercise 11.34: TRAP is more highly correlated with VO+, and is also highly correlated with OC, so it is reasonable that, if TRAP is present in the model, little additional information is gained from OC.



Minitab output

```
- - - - - - - - - - - - - - - - - -  Model 1: OC only  - - - - - - - - - - - - - - - - - -
The regression equation is VOplus = 334 + 19.5 OC

Predictor       Coef       Stdev    t-ratio        p
Constant       334.0       159.2       2.10    0.045
OC            19.505       4.127       4.73    0.000

s = 443.3       R-sq = 43.5%     R-sq(adj) = 41.6%
- - - - - - - - - - - - - - - - -  Model 2: OC and TRAP  - - - - - - - - - - - - - - - - -
The regression equation is VOplus = 58 + 6.41 OC + 53.9 TRAP

Predictor       Coef       Stdev    t-ratio        p
Constant        57.7       156.5       0.37    0.715
OC             6.415       5.125       1.25    0.221
TRAP           53.87       15.39       3.50    0.002

s = 376.3       R-sq = 60.7%     R-sq(adj) = 57.9%
```

11.36. **(a)** The model is $\text{VOplus} = \beta_0 + \beta_1\ \text{OC} + \beta_2\ \text{TRAP} + \beta_2\ \text{VOminus} + \epsilon_i$, where ϵ_i are independent $N(0, \sigma)$ random variables. **(b)** See the table on the next page for this regression equation and the significance test results. **(c)** & **(d)** The table on the next page summarizes the results for the three regressions. The estimated coefficients and P-values can change rather drastically from one model to the next. Generally, R^2 increases (sometimes only slightly) as we add more explanatory variables to the model. **(e)** The results of the regression in (b) suggest that we remove TRAP from the model. This regression equation and associated results are also in the table on the next page. Because R^2 drops only slightly for this simpler model, this is probably the best of all models we have considered to this point.

Equation: $\widehat{\text{VOplus}} =$				R^2	s
334.0	+ 19.505 OC			0.435	443.3
	SE = 4.127				
	$t = 4.73$				
	$P < 0.0005$				
57.7	+ 6.415 OC	+ 53.87 TRAP		0.607	376.3
	SE = 5.125	SE = 15.39			
	$t = 1.25$	$t = 3.50$			
	$P = 0.221$	$P = 0.002$			
−243.5	+ 8.235 OC	+ 6.61 TRAP	+ 0.975 VOminus	0.884	207.8
	SE = 2.840	SE = 10.33	SE = 0.1211		
	$t = 2.90$	$t = 0.64$	$t = 8.05$		
	$P = 0.007$	$P = 0.528$	$P < 0.0005$		
−234.1	+ 9.404 OC		+ 1.019 VOminus	0.883	205.6
	SE = 2.150		SE = 0.0986		
	$t = 4.37$		$t = 10.33$		
	$P < 0.0005$		$P < 0.0005$		

Minitab output

```
- - - - - - - - - - - -  Model 3: OC, TRAP and VOminus  - - - - - - - - - - -
The regression equation is VOplus = -243 + 8.23 OC + 6.6 TRAP + 0.975 VOminus

Predictor       Coef       Stdev    t-ratio        p
Constant     -243.49       94.22      -2.58    0.015
OC             8.235       2.840       2.90    0.007
TRAP            6.61       10.33       0.64    0.528
VOminus       0.9746      0.1211       8.05    0.000

s = 207.8       R-sq = 88.4%     R-sq(adj) = 87.2%
- - - - - - - - - - - - - -  Model 4: OC and VOminus  - - - - - - - - - - - - -
The regression equation is VOplus = -234 + 9.40 OC + 1.02 VOminus

Predictor       Coef       Stdev    t-ratio        p
Constant     -234.14       92.09      -2.54    0.017
OC             9.404       2.150       4.37    0.000
VOminus      1.01857     0.09858      10.33    0.000

s = 205.6       R-sq = 88.3%     R-sq(adj) = 87.4%
```

11.37. Stemplots (below) show that all four variables are noticeably less skewed. The means and standard deviations are given in the table on the right. Correlations and scatterplots (below) show that all six pairs of variables are positively associated. The strongest association is between LVO+ and LVO– and the weakest is between LOC and LVO–. The regression equations for these transformed variables are given in the table below, along with significance test results. Residual analysis for these regressions is not shown here.

	Mean	SD
logVO+	6.7418	0.5555
logVO–	6.6816	0.4832
logOC	3.3380	0.6085
logTRAP	2.4674	0.4978

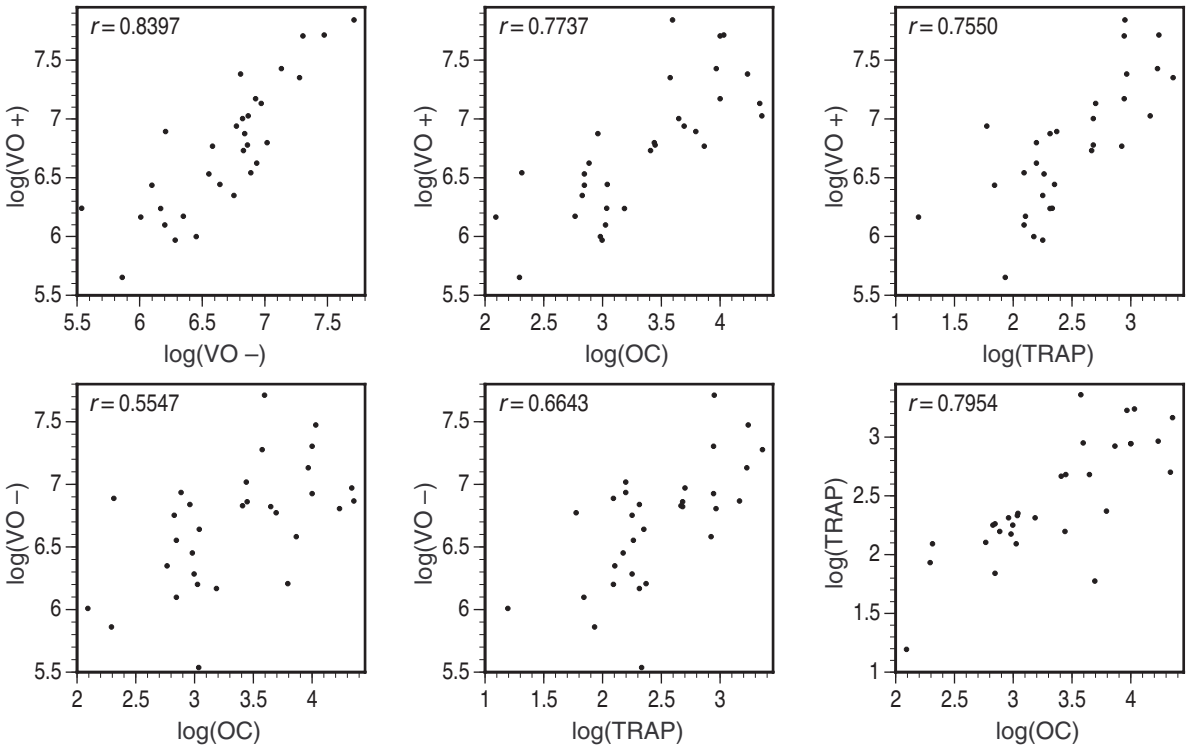
The final conclusion is the same as for the untransformed data: When we use all three explanatory variables to predict LVO+, the coefficient of LTRAP is not significantly different from 0 and we then find that the model that uses LOC and LVO– to predict LVO+ is nearly as good (in terms of R^2), making it the best of the bunch.

logVO+	
5	6
5	99
6	011
6	223
6	4455
6	67777
6	889
7	0011
7	33
7	4
7	77
7	8

logVO–	
5	5
5	
5	8
6	001
6	2223
6	455
6	677
6	8888888999
7	01
7	23
7	4
7	7

logOC	
2	0
2	23
2	
2	7
2	8888999
3	0001
3	
3	44455
3	667
3	89
4	000
4	233

logTRAP	
1	1
1	
1	
1	7
1	89
2	001111
2	22233333
2	
2	6667
2	99999
3	1
3	223



Equation: $\widehat{\text{LVOplus}} =$				R^2	s
4.3841	$+ 0.7063$ LOC			0.599	0.3580
	SE $= 0.1074$				
	$t = 6.58$				
	$P < 0.0005$				
4.2590	$+ 0.4304$ LOC	$+ 0.4240$ LTRAP		0.652	0.3394
	SE $= 0.1680$	SE $= 0.2054$			
	$t = 2.56$	$t = 2.06$			
	$P = 0.016$	$P = 0.048$			
0.8716	$+ 0.3922$ LOC	$+ 0.0275$ LTRAP	$+ 0.6725$ LVOminus	0.842	0.2326
	SE $= 0.1154$	SE $= 0.1570$	SE $= 0.1178$		
	$t = 3.40$	$t = 0.18$	$t = 5.71$		
	$P = 0.002$	$P = 0.862$	$P < 0.0005$		
0.8321	$+ 0.4061$ LOC		$+ 0.6816$ LVOminus	0.842	0.2286
	SE $= 0.0824$		SE $= 0.1038$		
	$t = 4.93$		$t = 6.57$		
	$P < 0.0005$		$P < 0.0005$		

11.38. Refer to the solution to Exercise 11.34 for the scatterplots. Note that, in this case, it really makes the most sense to use TRAP to predict VO– (because it is the appropriate biomarker), but many students might miss that detail. Both single-explanatory variable models are given in the table below. Residual analysis plots are not included. As before, we find the best model is to use OC and VO+ to predict VO–.

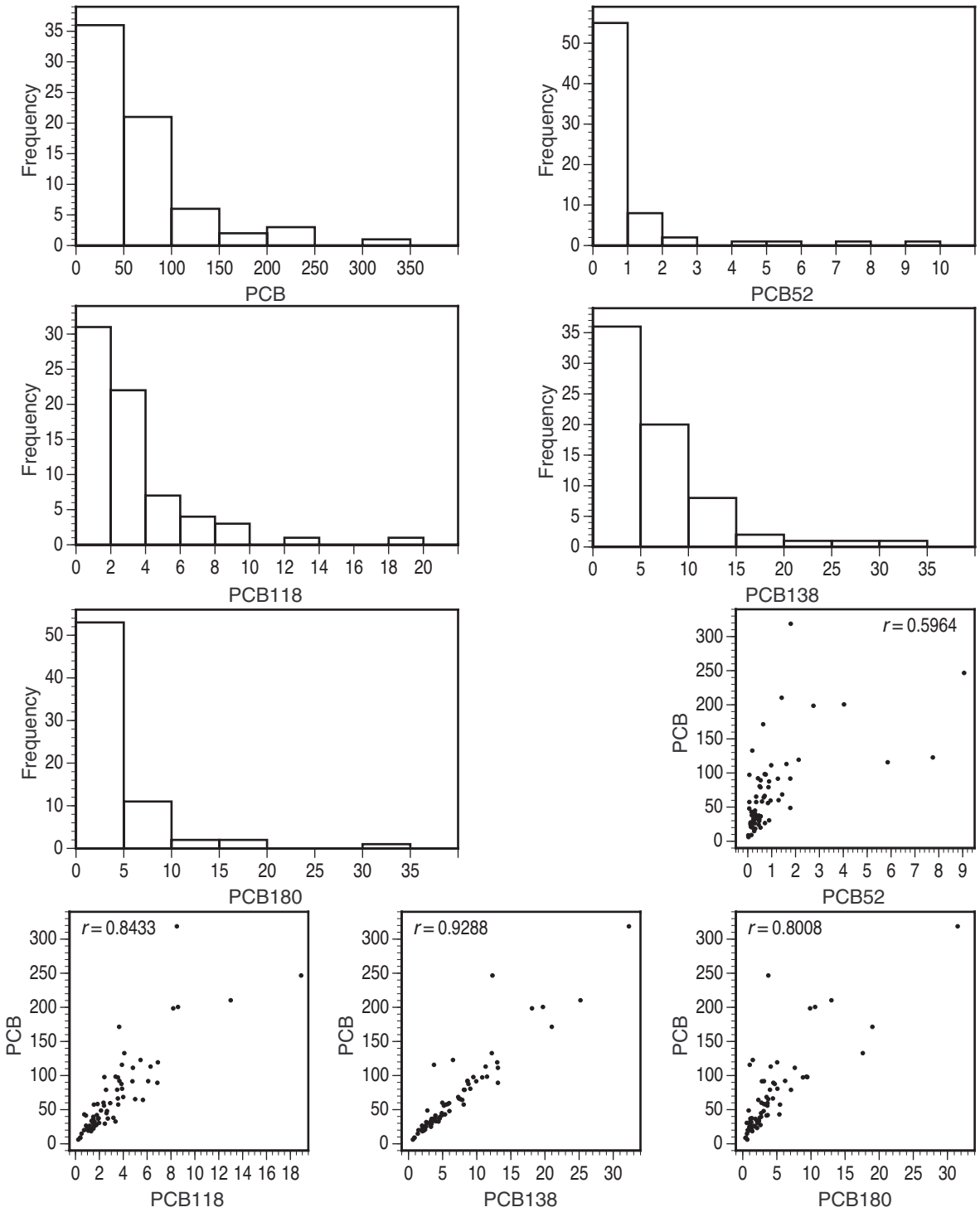
Equation: $\widehat{\text{VOminus}} =$				R^2	s
557.8	+ 9.917 OC			0.207	387.4
	SE = 3.606				
	$t = 2.75$				
	$P = 0.010$				
300.9		+ 44.41 TRAP		0.460	319.7
		SE = 8.942			
		$t = 4.97$			
		$P < 0.0005$			
309.1	− 1.868 OC	+ 48.50 TRAP		0.463	324.4
	SE = 4.418	SE = 13.27			
	$t = -0.42$	$t = 3.66$			
	$P = 0.676$	$P = 0.001$			
267.3	− 6.513 OC	+ 9.485 TRAP	+ 0.724 VOplus	0.842	179.2
	SE = 2.507	SE = 10.33	SE = 0.0900		
	$t = -2.60$	$t = 1.08$	$t = 8.05$		
	$P = 0.015$	$P = 0.290$	$P < 0.0005$		
298.0	− 5.254 OC		+ 0.778 VOplus	0.835	179.7
	SE = 2.226		SE = 0.0753		
	$t = -2.36$		$t = 10.33$		
	$P = 0.025$		$P < 0.0005$		

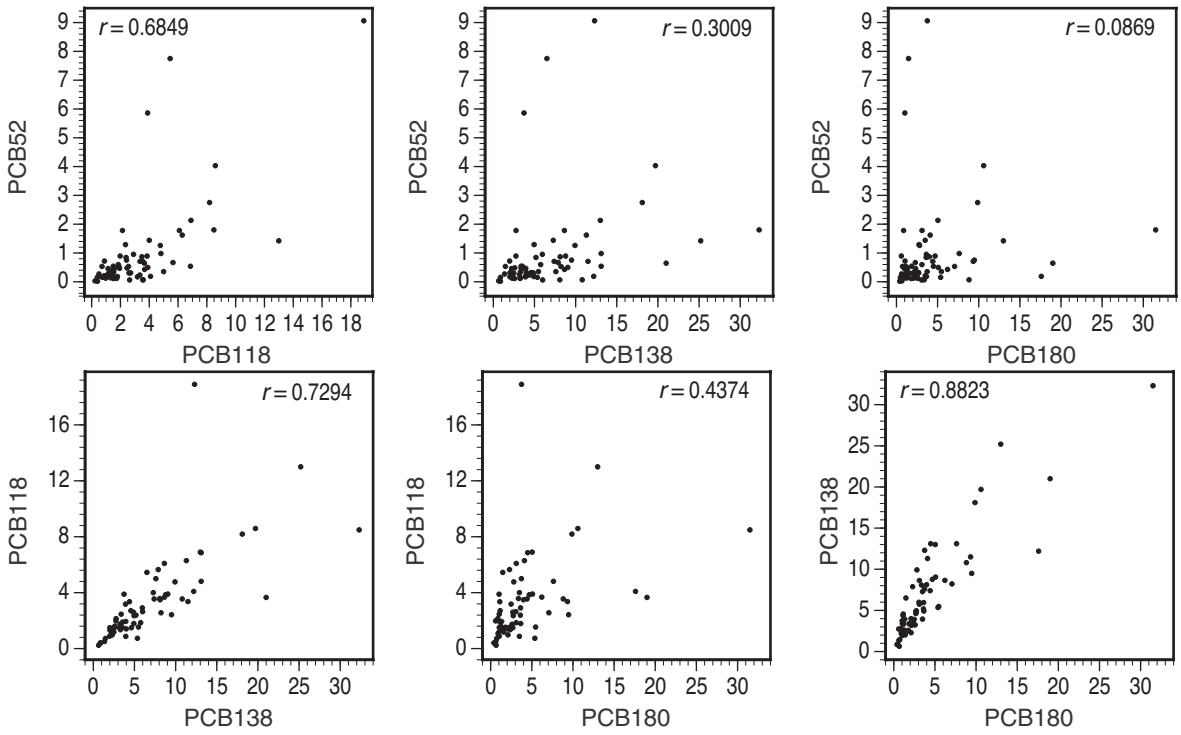
11.39. Refer to the solution to Exercise 11.36 for the scatterplots. Note that, in this case, it really makes the most sense to use LTRAP to predict LVO– (because it is the appropriate biomarker), but many students might miss that detail. Both single-explanatory variable models are given in the table below. Residual analysis plots are not included. This time, we might conclude that the best model is to predict LVO– from LVO+ alone; neither biomarker variable makes an indispensable contribution to the prediction.

Equation: $\widehat{\text{LVOminus}} =$				R^2	s
5.2110	+ 0.4406 LOC			0.308	0.4089
	SE = 0.1227				
	$t = 3.59$				
	$P = 0.001$				
5.0905		+ 0.6449 LTRAP		0.441	0.3674
		SE = 0.1347			
		$t = 4.79$			
		$P < 0.0005$			
5.0370	+ 0.0569 LOC	+ 0.5896 LTRAP		0.443	0.3732
	SE = 0.1848	SE = 0.2259			
	$t = 0.31$	$t = 2.61$			
	$P = 0.761$	$P = 0.014$			
1.5729	− 0.2932 LOC	+ 0.2447 LTRAP	+ 0.8134 LVOplus	0.748	0.2558
	SE = 0.1407	SE = 0.1662	SE = 0.1425		
	$t = -2.08$	$t = 1.47$	$t = 5.71$		
	$P = 0.047$	$P = 0.152$	$P < 0.0005$		
1.3109	− 0.1878 LOC		+ 0.8896 LVOplus	0.728	0.2611
	SE = 0.1237		SE = 0.1355		
	$t = -1.52$		$t = 6.57$		
	$P = 0.140$		$P < 0.0005$		
1.7570			+ 0.7304 LVOplus	0.705	0.2669
			SE = 0.0877		
			$t = 8.33$		
			$P < 0.0005$		

11.40. **(a)** Means and standard deviations are given in the table on the right; histograms are on the next page. All distributions are sharply right-skewed. **(b)** Scatterplots and correlations are on the following pages. All pairs of variables are positively associated, although some only weakly. In general, even when the association is strong, the plots show more variation for large values of the two variables. One correlation (PCB52/PCB180) is not significantly different from 0 ($r = 0.0869$, $t = 0.71$, $P = 0.4775$). The correlation between PCB52/PCB138 is significantly different from 0 at $\alpha = 0.05$, but not 0.01 ($r = 0.3009$, $t = 2.58$, $P = 0.0120$). All others have $P < 0.0005$.

	Mean	SD
PCB	68.4674	59.3906
PCB52	0.9580	1.5983
PCB118	3.2563	3.0191
PCB138	6.8268	5.8627
PCB180	4.1584	4.9864


Frequency
PCB
PCB52
PCB118
PCB138
PCB180
r = 0.5964
r = 0.8433
r = 0.9288
r = 0.8008



11.41. **(a)** The model is:

$$\text{PCB}_i = \beta_0 + \beta_1\,\text{PCB}_{52} + \beta_2\,\text{PCB}_{118} + \beta_3\,\text{PCB}_{138} + \beta_4\,\text{PCB}_{180} + \epsilon_i$$

where $i = 1, 2, \ldots, 69$; ϵ_i are independent $N(0, \sigma)$ random variables. **(b)** The regression equation is:

$$\widehat{\text{PCB}} = 0.937 + 11.8727\,\text{PCB}_{52} + 3.7611\,\text{PCB}_{118} + 3.8842\,\text{PCB}_{138} + 4.1823\,\text{PCB}_{180}$$

with $s = 6.382$ and $R^2 = 0.989$. All coefficients are significantly different from 0, although the constant 0.937 is not ($t = 0.76$, $P = 0.449$). That makes some sense—if none of these four congeners are present, it might be somewhat reasonable to predict that the total amount of PCB is 0. **(c)** The residuals appear to be roughly Normal, but with two outliers. There are no clear patterns when plotted against the explanatory variables (these plots are not shown).

−2	2
−1	
−1	31
−0	8776655
−0	4443333222211111000000
0	000000000001111222223333444
0	677778
1	12
1	
2	2

Minitab output

```
The regression equation is
PCB = 0.94 + 11.9 PCB52 + 3.76 PCB118 + 3.88 PCB138 + 4.18 PCB180

Predictor       Coef       Stdev    t-ratio        p
Constant       0.937       1.229       0.76    0.449
PCB52        11.8727      0.7290      16.29    0.000
PCB118        3.7611      0.6424       5.85    0.000
PCB138        3.8842      0.4978       7.80    0.000
PCB180        4.1823      0.4318       9.69    0.000

s = 6.382       R-sq = 98.9%     R-sq(adj) = 98.8%
```

11.42. **(a)** The outliers are specimen #50 (residual −22.0864) and #65 (22.5487). Because residuals are observed values minus predicted values, the negative residual (#50) is an overestimate. (The estimated PCB for this specimen is $\widehat{\text{PCB}} \doteq 144.882$, and the actual level was 122.796.) **(b)** The regression equation is:

$$\widehat{\text{PCB}} = 1.6277 + 14.4420\,\text{PCB}_{52} + 2.5996\,\text{PCB}_{118} + 4.0541\,\text{PCB}_{138} + 4.1086\,\text{PCB}_{180}$$

with $s = 4.555$ and $R^2 = 0.994$. As before, all coefficients are significantly different from 0, although the constant is barely not different ($t = 1.84$, $P = 0.071$). The residuals again appear to be roughly Normal, but two new specimens (#44 and #58) show up as outliers to replace the two we removed. There are no clear patterns when plotted against the explanatory variables (these plots are not shown).

−1	2
−1	
−0	98
−0	76
−0	5544
−0	33332222
−0	1111111100000000000
0	0000000000111111
0	22233
0	445
0	677
0	889
1	
1	
1	4

Minitab output

```
The regression equation is
PCB = 1.63 + 14.4 PCB52 + 2.60 PCB118 + 4.05 PCB138 + 4.11 PCB180

Predictor       Coef       Stdev    t-ratio        p
Constant      1.6277      0.8858       1.84    0.071
PCB52        14.4420      0.6960      20.75    0.000
PCB118        2.5996      0.5164       5.03    0.000
PCB138        4.0541      0.3752      10.80    0.000
PCB180        4.1086      0.3175      12.94    0.000

s = 4.555       R-sq = 99.4%     R-sq(adj) = 99.4%
```

11.43. **(a)** The regression equation is:

$$\widehat{\text{PCB}} = -1.018 + 12.644\,\text{PCB}_{52} + 0.3131\,\text{PCB}_{118} + 8.2546\,\text{PCB}_{138}$$

with $s = 9.945$ and $R^2 = 0.973$. Residual analysis (not shown) suggests a few areas of concern: The distribution of residuals has heavier tails than a Normal distribution, and the scatter (that is, prediction error) is greater for larger values of the predicted PCB. **(b)** The estimated coefficient of PCB118 is $b_2 \doteq 0.3131$; its P-value is 0.708. (Details in Minitab output below.) **(c)** In Exercise 11.41, $b_2 \doteq 3.7611$ and $P < 0.0005$. **(d)** This illustrates how complicated multiple regression can be: When we add PCB180 to the model, it complements PCB118, making it useful for prediction.

Minitab output

```
The regression equation is
PCB = -1.02 + 12.6 PCB52 + 0.313 PCB118 + 8.25 PCB138

Predictor       Coef       Stdev    t-ratio        p
Constant      -1.018       1.890      -0.54    0.592
PCB52         12.644       1.129      11.20    0.000
PCB118        0.3131      0.8333       0.38    0.708
PCB138        8.2546      0.3279      25.18    0.000

s = 9.945       R-sq = 97.3%     R-sq(adj) = 97.2%
```

11.44. **(a)** Because TEQ is defined as the sum TEQPCB + TEQDIOXIN + TEQFURAN, we have $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = 1$. **(b)** The error terms are all zero, so they have no scatter; therefore, $\sigma = 0$. **(c)** Except for rounding error, the regression confirms the values in (a) and (b).

Minitab output

```
The regression equation is
TEQ =0.000000 + 1.00 TEQPCB + 1.00 TEQDIOXIN + 1.00 TEQFURAN

Predictor       Coef       Stdev    t-ratio        p
Constant  0.00000032  0.00000192       0.16    0.870
TEQPCB       1.00000     0.00000 1211707.25    0.000
TEQDIOXIN    1.00000     0.00000  566800.75    0.000
TEQFURAN     1.00000     0.00001  176270.48    0.000

s = 0.000007964   R-sq = 100.0%    R-sq(adj) = 100.0%
```

11.45. The model is:

$$\text{TEQ}_i = \beta_0 + \beta_1\,\text{PCB}_{52} + \beta_2\,\text{PCB}_{118} + \beta_3\,\text{PCB}_{138} + \beta_4\,\text{PCB}_{180} + \epsilon_i$$

where $i = 1, 2, \ldots, 69$; ϵ_i are independent $N(0, \sigma)$ random variables. The regression equation is:

$$\widehat{\text{TEQ}} = 1.0600 - 0.0973\,\text{PCB}_{52} + 0.3062\,\text{PCB}_{118} + 0.1058\,\text{PCB}_{138} - 0.0039\,\text{PCB}_{180}$$

with $s = 0.9576$ and $R^2 = 0.677$. Only the constant and the PCB118 coefficient are significantly different from 0; see Minitab output below. Residuals (stemplot on the left) are slightly right-skewed and show no clear patterns when plotted with the explanatory variables (not shown).

−1	66
−1	4200
−0	987666666666555555
−0	44444333221111100
0	0000222224
0	566667788
1	23334
1	9
2	3
2	57

Minitab output

```
The regression equation is
TEQ = 1.06 - 0.097 PCB52 + 0.306 PCB118 + 0.106 PCB138 - 0.0039 PCB180

Predictor       Coef       Stdev    t-ratio        p
Constant      1.0600      0.1845       5.75    0.000
PCB52        -0.0973      0.1094      -0.89    0.377
PCB118       0.30618     0.09639       3.18    0.002
PCB138       0.10579     0.07470       1.42    0.162
PCB180      -0.00391     0.06478      -0.06    0.952

s = 0.9576      R-sq = 67.7%     R-sq(adj) = 65.7%
```

11.46. **(a)** Results will vary with software. **(b)** Different software may produce different results, but (presumably) all software will ignore those 16 specimens, which is probably not a good approach. **(c)** The summary statistics (right) and stemplots (below) are based on natural logarithms; for common logarithms, multiply mean and standard deviation by 2.3026. For LPCB126, the zero terms were replaced $\ln 0.0026 \doteq -5.9522$, which accounts for the odd appearance of its stemplot.

	Mean	SD
LPCB28	−1.3345	1.1338
LPCB52	−0.7719	1.1891
LPCB118	0.8559	0.8272
LPCB126	−4.8457	0.7656
LPCB138	1.6139	0.8046
LPCB153	1.7034	0.9012
LPCB180	0.9752	0.9276
LPCB	3.9170	0.8020
LTEQ	0.8048	0.5966

LPCB28

−5	1
−4	
−4	
−3	6
−3	
−2	88765
−2	433331111100
−1	8888776665555
−1	443222211111
−0	99998888766555
−0	3100
0	4
0	56789
1	
1	9

LPCB52

−3	85
−3	
−2	776
−2	1100
−1	98877765
−1	4443211111000
−0	8777776666665
−0	443332111100
0	22334
0	5557
1	03
1	7
2	02

LPCB118

−1	4
−0	996
−0	33110
0	0112233334444444
0	556666788899999
1	012222222333334
1	556678899
2	111
2	59

LPCB126

−5	9999999999999999
−5	
−5	
−5	222
−5	111000
−4	99999999998888
−4	7776
−4	54444
−4	3322
−4	11100
−3	999888
−3	666
−3	554

LPCB138

−0	411
0	33
0	6677788999
1	0011112222333334
1	5555566777789
2	000001111222344
2	5555589
3	024

LPCB153

−0	110
0	134
0	5578899
1	000112222222222333444
1	5677888899999
2	01222233444
2	55569
3	01244
3	57

LPCB180

−0	975
−0	443110
0	00001111222334
0	567778889999
1	0011122222233344
1	55666689
2	012223
2	589
3	4

LPCB

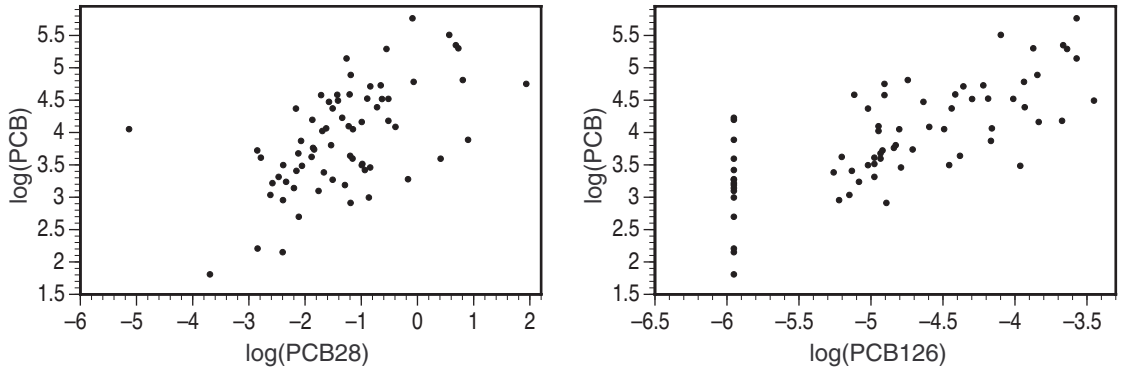
1	8
2	12
2	6999
3	0011222233444444
3	5556666777888
4	000000111233344
4	555555777788
5	1233
5	57

LTEQ

0	00000000111111
0	2223333333
0	4455555
0	66677
0	888899
1	0111111
1	22333
1	44455
1	6667777
1	888

11.47. (a) The correlations (all positive) are listed in the table below. The largest correlation is 0.956 (LPCB and LPCB138); the smallest (0.227, for LPCB28 and LPCB180) is not quite significantly different from 0 ($t = 1.91$, $P = 0.0607$) but, with 28 correlations, such a P-value could easily arise by chance, so we would not necessarily conclude that $\rho = 0$. Rather than showing all 28 scatterplots—which are all fairly linear and confirm the positive associations suggested by the correlations—we have included only two of the interesting ones: LPCB against LPCB28 and LPCB against LPCB126. The former is notable because of one outlier (specimen 39) in LPCB28; the latter stands out because of the "stack" of values in the LPCB126 data set that arose from the adjustment of the zero terms. (The outlier in LPCB28 and the stack in LPCB126 can be seen in other plots involving those variables; the two plots shown are the most appropriate for using the PCB congeners to predict LPCB, as the next exercise asks.) **(b)** All correlations are higher with the transformed data. In part, this is because these scatterplots do not exhibit the "greater scatter in the upper right" that was seen in many of the scatterplots of the original data.

	LPCB28	LPCB52	LPCB118	LPCB126	LPCB138	LPCB153	LPCB180
LPCB52	0.795						
LPCB118	0.533	0.671					
LPCB126	0.272	0.331	0.739				
LPCB138	0.387	0.540	0.890	0.792			
LPCB153	0.326	0.519	0.780	0.647	0.922		
LPCB180	0.227	0.301	0.654	0.695	0.896	0.867	
LPCB	0.570	0.701	0.906	0.729	0.956	0.905	0.829



11.48. Student results will vary with how many different models they try, and what tradeoff they consider between "good" (in terms of large R^2) and "simple" (in terms of the number of variables included in the model). The first Minitab output on the next page, produced with the `BREG` (best regression) command, gives some guidance as to likely answers; it shows the best models with one, two, three, four, five, six, and seven explanatory variables. We can see, for example, that if all variables are used, $R^2 = 0.975$, but we can achieve similar values of R^2 with fewer variables. The best regressions with two, three, and four explanatory variables are shown in the Minitab output on the next page.

Minitab output

Best subsets regression

Vars	R-sq	Adj. R-sq	C-p	s	L L P P P P P C C B B B B 2 8 1 2 3 5 8 8 2 8 6 8 3 0
1	91.4	91.3	141.4	0.23689	X
2	96.2	96.1	28.5	0.15892	X X
3	96.8	96.6	16.1	0.14696	X X X
4	97.2	97.0	8.2	0.13826	X X X X
5	97.3	97.0	8.6	0.13776	X X X X X
6	97.5	97.2	6.0	0.13389	X X X X X X
7	97.5	97.2	8.0	0.13497	X X X X X X X

Two explanatory variables

```
The regression equation is
LPCB = 2.74 + 0.175 LPCB52 + 0.813 LPCB138

Predictor       Coef       Stdev    t-ratio        p
Constant     2.74038     0.05860      46.76    0.000
LPCB52       0.17533     0.01926       9.10    0.000
LPCB138      0.81294     0.02846      28.56    0.000

s = 0.1589      R-sq = 96.2%     R-sq(adj) = 96.1%
```

Three explanatory variables

```
The regression equation is
LPCB = 2.79 + 0.0908 LPCB28 + 0.104 LPCB52 + 0.821 LPCB138

Predictor       Coef       Stdev    t-ratio        p
Constant     2.79394     0.05633      49.60    0.000
LPCB28       0.09078     0.02601       3.49    0.001
LPCB52       0.10371     0.02717       3.82    0.000
LPCB138      0.82056     0.02641      31.07    0.000

s = 0.1470      R-sq = 96.8%     R-sq(adj) = 96.6%
```

Four explanatory variables

```
The regression equation is
LPCB = 2.79 + 0.107 LPCB28 + 0.0876 LPCB52 + 0.669 LPCB138 + 0.151 LPCB153

Predictor       Coef       Stdev    t-ratio        p
Constant     2.79081     0.05300      52.65    0.000
LPCB28       0.10684     0.02503       4.27    0.000
LPCB52       0.08763     0.02610       3.36    0.001
LPCB138      0.66854     0.05538      12.07    0.000
LPCB153      0.15118     0.04921       3.07    0.003

s = 0.1383      R-sq = 97.2%     R-sq(adj) = 97.0%
```

11.49. Using Minitab's BREG (best regression) command for guidance, we see that there is little improvement in R^2 beyond models with four explanatory variables. The best models with two, three, and four variables are given in the Minitab output below.

Minitab output

Best subsets regression

Vars	R-sq	Adj. R-sq	C-p	s	L L P P P P P P C C C C C C B B B B B B 1 1 1 1 1 1 2 5 1 2 3 5 8 8 2 8 6 8 3 0
1	72.9	72.5	10.8	0.31266	X
2	76.8	76.1	2.0	0.29166	X X
3	77.6	76.6	1.6	0.28859	X X X
4	78.0	76.7	2.5	0.28816	X X X X
5	78.1	76.4	4.2	0.28981	X X X X X
6	78.2	76.1	6.1	0.29188	X X X X X X
7	78.2	75.7	8.0	0.29400	X X X X X X X

Two explanatory variables

```
The regression equation is
LTEQ = 3.96 + 0.107 LPCB28 + 0.622 LPCB126

Predictor       Coef       Stdev    t-ratio        p
Constant      3.9637      0.2275      17.42    0.000
LPCB28       0.10749     0.03242       3.32    0.001
LPCB126      0.62231     0.04801      12.96    0.000

s = 0.2917      R-sq = 76.8%     R-sq(adj) = 76.1%
```

Three explanatory variables

```
The regression equation is
LTEQ = 3.44 + 0.0777 LPCB28 + 0.114 LPCB118 + 0.543 LPCB126

Predictor       Coef       Stdev    t-ratio        p
Constant      3.4445      0.4029       8.55    0.000
LPCB28       0.07773     0.03736       2.08    0.041
LPCB118      0.11371     0.07319       1.55    0.125
LPCB126      0.54345     0.06952       7.82    0.000

s = 0.2886      R-sq = 77.6%     R-sq(adj) = 76.6%
```

Four explanatory variables

```
The regression equation is
LTEQ = 3.56 + 0.0720 LPCB28 + 0.170 LPCB118 + 0.554 LPCB126 - 0.0693 LPCB153

Predictor       Coef       Stdev    t-ratio        p
Constant      3.5568      0.4152       8.57    0.000
LPCB28       0.07199     0.03767       1.91    0.060
LPCB118      0.16973     0.08928       1.90    0.062
LPCB126      0.55374     0.07005       7.90    0.000
LPCB153     -0.06929     0.06344      -1.09    0.279

s = 0.2882      R-sq = 78.0%     R-sq(adj) = 76.7%
```

11.50. The degree of change in these elements of a regression can be readily seen by comparing the three regression results shown in the solution to Exercise 11.48; they will be even more visible if students have explored more models in their search for the best model. Student explanations might include observations of changes in particular coefficients from one model to another and perhaps might attempt to paraphrase the text's comments about *why* this happens.

11.51. In the table, two *IQR*s are given; those in parentheses are based on quartiles reported by Minitab, which computes quartiles in a slightly different way from this text's method.

	$\bar{x}$	M	s	IQR
Taste	24.53	20.95	16.26	23.9 (or 24.58)
Acetic	5.498	5.425	0.571	0.656 (or 0.713)
H2S	5.942	5.329	2.127	3.689 (or 3.766)
Lactic	1.442	1.450	0.3035	0.430 (or 0.4625)

None of the variables show striking deviations from Normality in the quantile plots (not shown). Taste and H2S are slightly right-skewed, and Acetic has two peaks. There are no outliers.

Taste

0	00
0	556
1	1234
1	55688
2	011
2	556
3	24
3	789
4	0
4	7
5	4
5	67

Acetic

4	455
4	67
4	8
5	1
5	2222333
5	444
5	677
5	888
6	0011
6	3
6	44

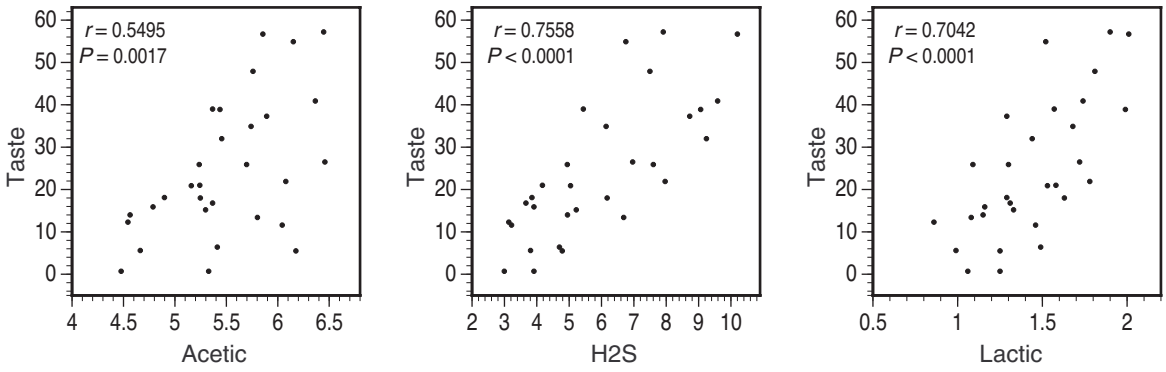
H2S

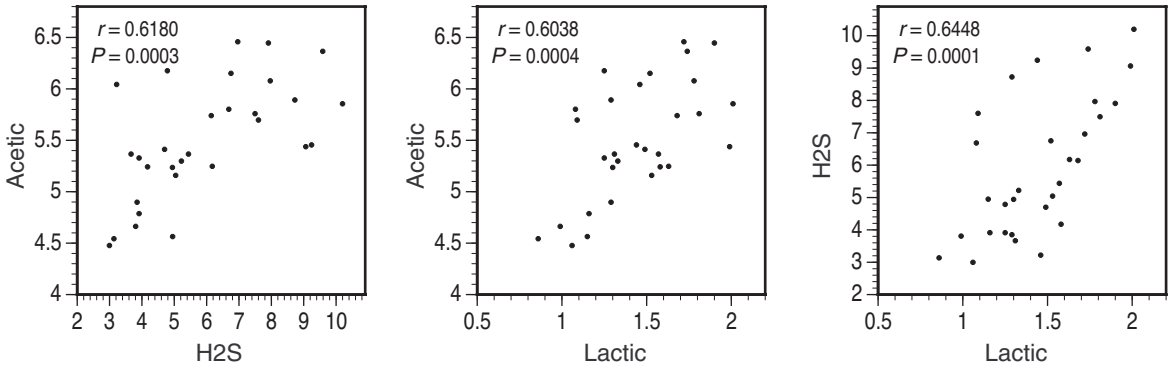
2	9
3	1268899
4	17799
5	024
6	11679
7	4699
8	7
9	025
10	1

Lactic

8	6
9	9
10	689
11	56
12	5599
13	013
14	469
15	2378
16	38
17	248
18	1
19	09
20	1

11.52. The plots show positive associations between the variables. The correlations and P-values are in the plots; all correlations are positive (as expected) and significantly different from 0. (Recall that the P-values are correct if the two variables are Normally distributed, in which case $t = r\sqrt{n-2}/\sqrt{1-r^2}$ has a $t(n-2)$ distribution if $\rho = 0$.)

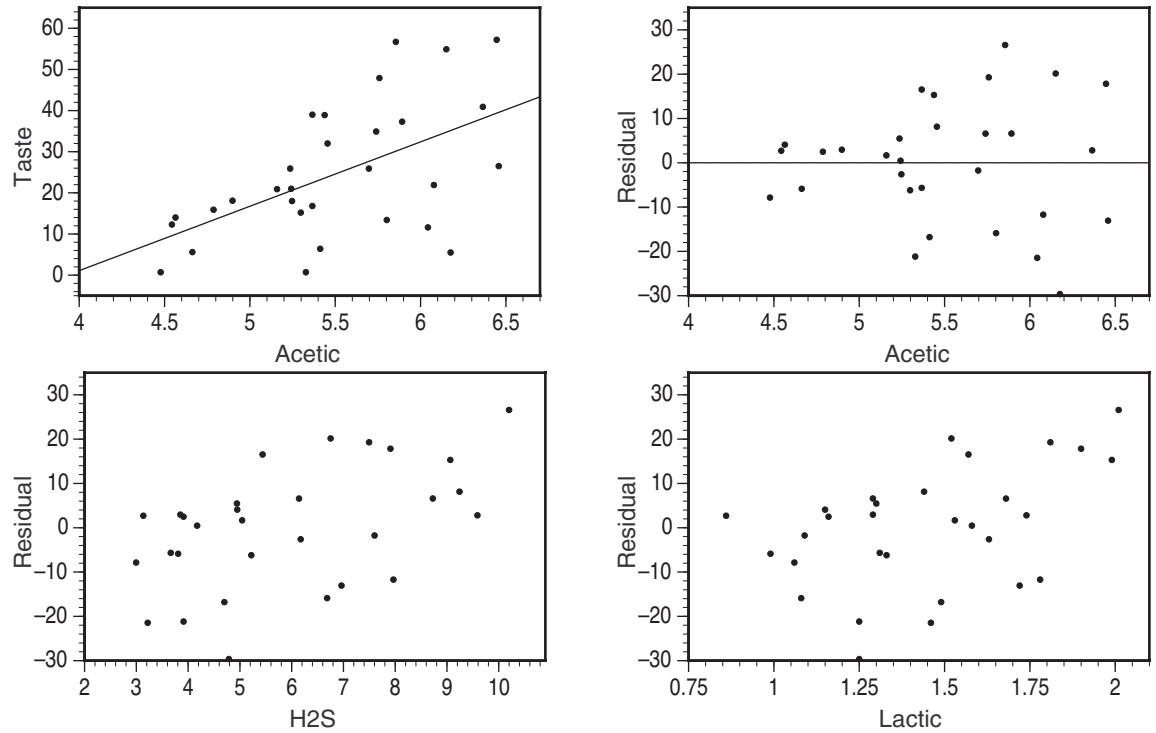




11.53. The regression equation is $\widehat{\text{Taste}} = -61.50 + 15.648$ Acetic with $s = 13.82$ and $R^2 = 0.302$. The slope is significantly different from 0 ($t = 3.48$, $P = 0.002$).

Based on a stemplot (right) and quantile plot (not shown), the residuals seem to have a Normal distribution. Scatterplots (below) reveal positive associations between residuals and both H2S and Lactic. The plot of residuals against Acetic suggests greater scatter in the residuals for large Acetic values.

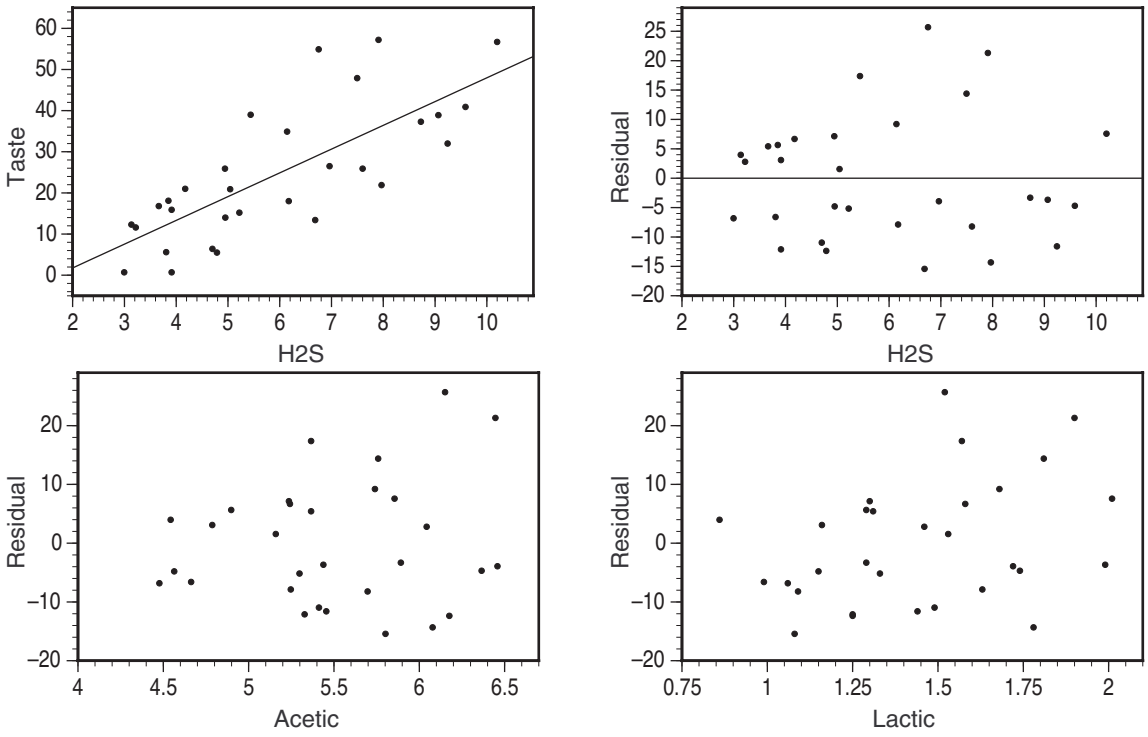
−2	9
−2	11
−1	65
−1	31
−0	7655
−0	21
0	0122224
0	5668
1	
1	5679
2	0
2	6



11.54. The regression equation is $\widehat{\text{Taste}} = -9.787 + 5.7761\ \text{H2S}$ with $s = 10.83$ and $R^2 = 0.571$. The slope is significantly different from 0 ($t = 6.11$, $P < 0.0005$).

Based on a stemplot (right) and quantile plot (not shown), the residuals may be slightly skewed, but do not differ greatly from a Normal distribution. Scatterplots (below) reveal weak positive associations between residuals and both Acetic and Lactic. The plot of residuals against H2S suggests greater scatter in the residuals for large H2S values.

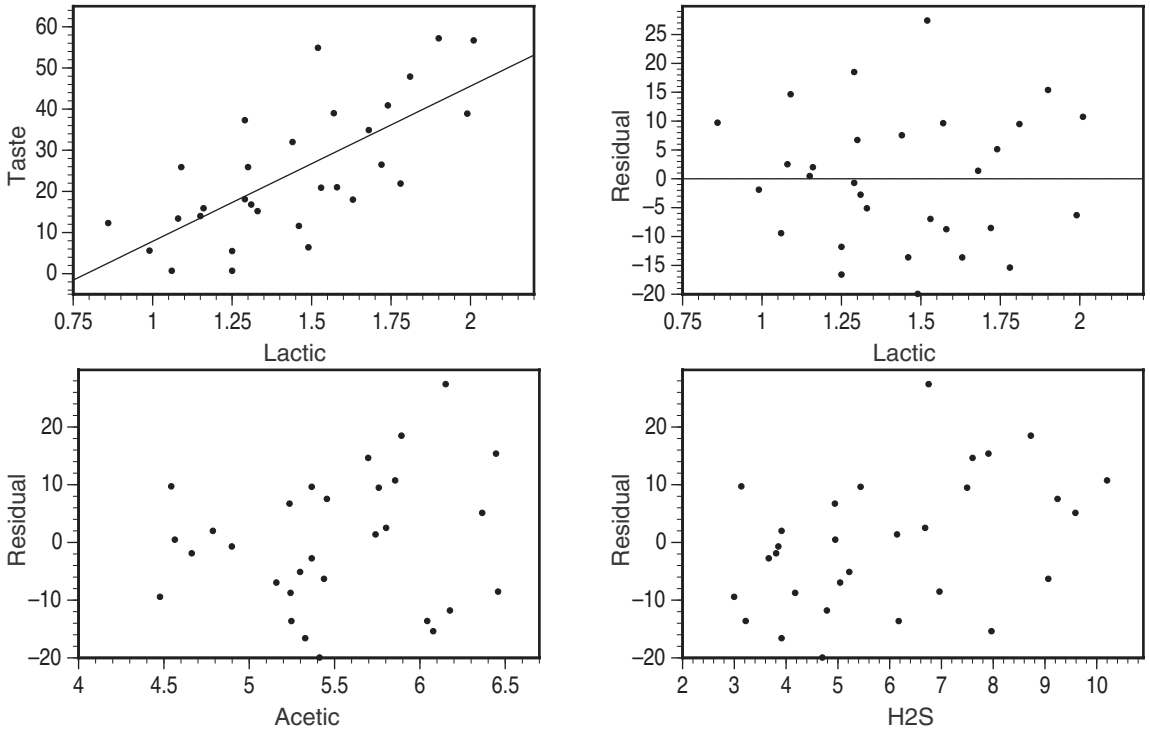
−1	5
−1	42210
−0	87665
−0	44333
0	1233
0	556779
1	4
1	7
2	1
2	5



11.55. The regression equation is $\widehat{\text{Taste}} = -29.86 + 37.720$ Lactic with $s = 11.75$ and $R^2 = 0.496$. The slope is significantly different from 0 ($t = 5.25$, $P < 0.0005$).

Based on a stemplot (right) and quantile plot (not shown), the residuals appear to be roughly Normal. Scatterplots (below) reveal no striking patterns for residuals vs. Acetic and H2S.

−1	965
−1	331
−0	988665
−0	210
0	0122
0	567999
1	04
1	58
2	
2	7



11.56. All information is in the table at the right. The intercepts differ from one model to the next because they represent different things—for example, in the first model, the intercept is the predicted value of Taste with Acetic = 0, etc.

x	$\widehat{\text{Taste}} =$	F	P	r^2	s
Acetic	$-61.50 + 15.648x$	12.11	0.002	30.2%	13.82
H2S	$-9.787 + 5.7761x$	37.29	<0.0005	57.1%	10.83
Lactic	$-29.86 + 37.720x$	27.55	<0.0005	49.6%	11.75

11.57. The regression equation is $\widehat{\text{Taste}} = -26.94 + 3.801\ \text{Acetic} + 5.146\ \text{H2S}$ with $s = 10.89$ and $R^2 = 0.582$. The t-value for the coefficient of Acetic is 0.84 ($P = 0.406$), indicating that it does not add significantly to the model when H2S is used because Acetic and H2S are correlated (in fact, $r = 0.618$ for these two variables). This model does a better job than any of the three simple linear regression models, but it is not much better than the model with H2S alone (which explained 57.1% of the variation in Taste)—as we might expect from the t-test result.

Minitab output

```
The regression equation is taste = -26.9 + 3.80 acetic + 5.15 h2s

Predictor       Coef       Stdev    t-ratio        p
Constant      -26.94       21.19      -1.27    0.215
acetic         3.801       4.505       0.84    0.406
h2s            5.146       1.209       4.26    0.000

s = 10.89       R-sq = 58.2%     R-sq(adj) = 55.1%
```

11.58. The regression equation is $\widehat{\text{Taste}} = -27.592 + 3.946\ \text{H2S} + 19.887\ \text{Lactic}$ with $s = 9.942$. The model explains 65.2% of the variation in Taste, which is higher than for the two simple linear regressions. Both coefficients are significantly different from 0 ($P = 0.002$ for H2S, and $P = 0.019$ for Lactic).

Minitab output

```
The regression equation is taste = -27.6 + 3.95 h2s + 19.9 lactic

Predictor       Coef       Stdev    t-ratio        p
Constant     -27.592       8.982      -3.07    0.005
h2s            3.946       1.136       3.47    0.002
lactic        19.887       7.959       2.50    0.019

s = 9.942       R-sq = 65.2%     R-sq(adj) = 62.6%
```

11.59. The regression equation is $\widehat{\text{Taste}} = -28.88 + 0.328\ \text{Acetic} + 3.912\ \text{H2S} + 19.671\ \text{Lactic}$ with $s = 10.13$. The model explains 65.2% of the variation in Taste (the same as for the model with only H2S and Lactic). Residuals of this regression appear to be Normally distributed and show no patterns in scatterplots with the explanatory variables. (These plots are not shown.)

The coefficient of Acetic is not significantly different from 0 ($P = 0.942$); there is no gain in adding Acetic to the model with H2S and Lactic. It appears that the best model is the H2S/Lactic model of Exercise 11.58.

Minitab output

```
The regression equation is
taste = -28.9 + 0.33 acetic + 3.91 h2s + 19.7 lactic

Predictor       Coef       Stdev    t-ratio        p
Constant      -28.88       19.74      -1.46    0.155
acetic         0.328       4.460       0.07    0.942
h2s            3.912       1.248       3.13    0.004
lactic        19.671       8.629       2.28    0.031

s = 10.13       R-sq = 65.2%     R-sq(adj) = 61.2%
```