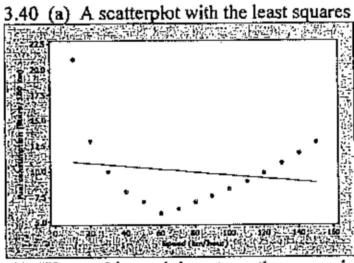
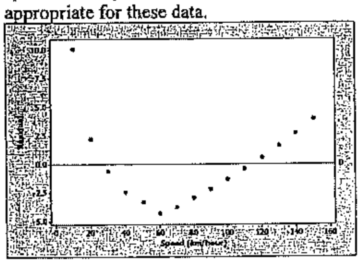
HW pp 220-222 ; pp 227-228 ; pp 230-233

3.39 (a) There is a positive, linear association between the two variables. There is more variation in the field measurements for larger laboratory measurements. The values are scattered above and below the line $y = x$ for small and moderate depths, indicating strong agreement, but the field measurements tend to be smaller than the laboratory measurements for large depths. (b) The points for the larger depths fall systematically below the line $y = x$ showing that the field measurements are too small compared to the laboratory measurements. (c) In order to minimize the sum of the squared distances from the points to the regression line, the top right part of the blue line in Figure 3.20 would need to be pulled down to go through the "middle" of the group of points that are currently below the blue line. Thus, the slope would decrease and the intercept would increase. (d) The residual plot clearly shows that the prediction errors increase for larger laboratory measurements. In other words, the variability in the field measurements increases as the laboratory measurements increase. The least squares line does not provide a great fit, especially for larger depths.

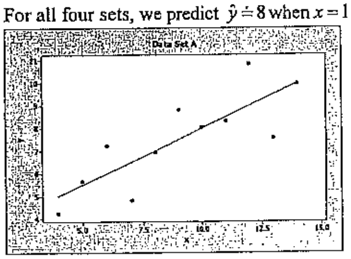
3.40 (a) A scatterplot with the least squares regression line is shown below.

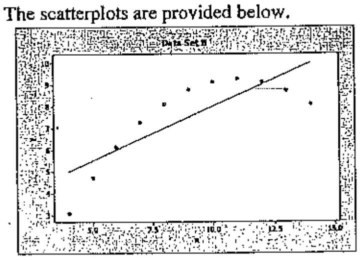


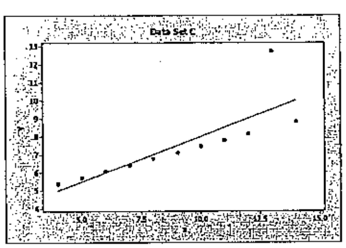
(b) We would certainly not use the regression line to predict fuel consumption. The scatterplot shows a nonlinear relationship. (c) The sum of the residuals provided is −0.01, which illustrates a slight roundoff error. (d) The residual plot indicates that the regression line underestimates fuel consumption for slow and fast speeds and overestimates fuel consumption for moderate speeds. The quadratic pattern in the residual plot indicates that the regression model is not appropriate for these data.

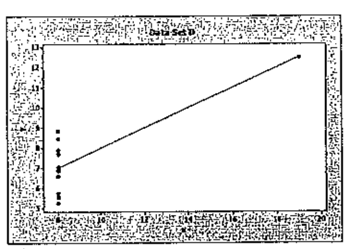


3.42 (a) The correlations are all approximately the same (To three decimal places $r_A = r_B = r_C = 0.816$ and $r_D = 0.817$), and the regression lines are all approximately $\hat{y} = 3.0 + 0.5x$. For all four sets, we predict $\hat{y} \doteq 8$ when $x = 10$. (b) The scatterplots are provided below.

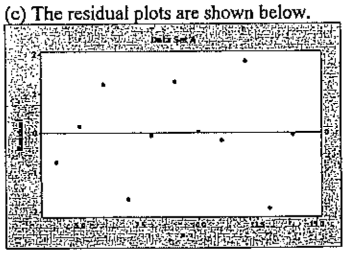


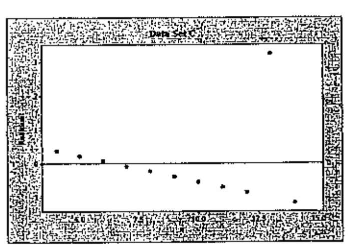


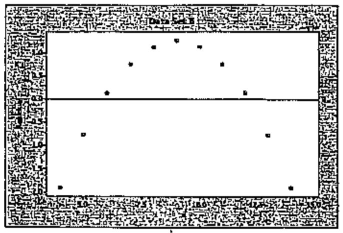


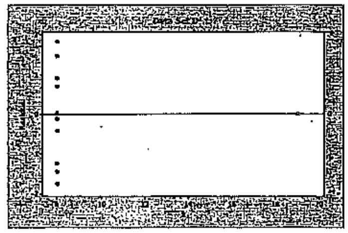


(c) The residual plots are shown below.



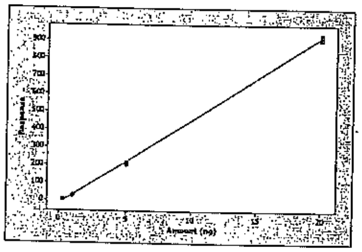






(d) The regression line should only be used for Data Set A. The variables have a moderate linear association with a fair amount of variability from the regression line and no obvious pattern in the residual plot. For Data Set B, there is an obvious nonlinear relationship which can be seen in both plots; we should fit a parabola or some other curve. For Data Set C, the point (13, 12.74) deviates from the strong linear relationship of the other points, pulling the regression line up. If a data entry error (or some other error) was made for this point, a regression line for the other points would be very useful for prediction. For Data Set D, the data point with $x = 19$ is a very influential point—the other points alone give no indication of slope for the line. The regression line is not useful in this situation with only two values of the explanatory variable x.

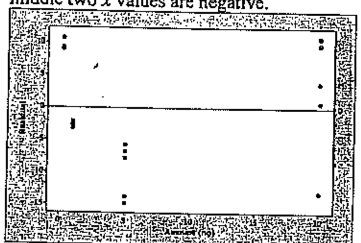
3.43 (a) The scatterplot of the data with the least-squares regression line is below.



The following table presents the measurement data:

Amount	Response
0.25	6.55
0.25	7.98
0.25	6.54
0.25	6.37
0.25	7.96
1	29.7
1	30
1	30.1
1	29.5
1	29.1
5	211
5	204
5	212
5	213
5	205
20	929
20	905
20	922
20	928
20	919

(b) The regression equation is $\hat{y} = -14.4 + 46.6x$. (c) The residual plot is below. The residuals for the extreme x-values ($x = 0.25$ and $x = 20.0$) are almost all positive; all of the residuals for the middle two x values are negative.

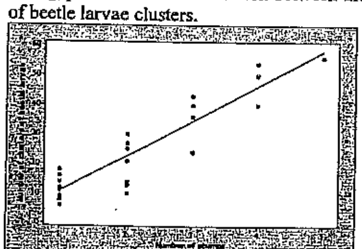


(d) $r^2 = 0.9997$; 99.97% of the variation in the response is explained by the least-squares regression with the amount of substance. This value suggests that the regression line does a great job predicting gas chromatograph readings.

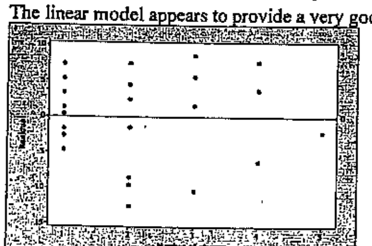
3.44 $r = \sqrt{0.16} = 0.40$ (high attendance goes with high grades, so the correlation must be positive).

3.47 (a) $r^2 = (0.596)^2 = 0.3552$. Thus, the straight-line relationship explains 35.52% of the variation in yearly changes. (b) The regression equation is $\hat{y} = 6.083 + 1.707x$. (c) The predicted change is $\hat{y} = 6.083 + 1.707 \times 1.75 = 9.0703\%$. We could have given the answer without doing calculations because the regression line must pass through $(\bar{x}, \bar{y}) = (1.75, 9.07)$.

3.48 (a) A scatterplot, with the least-squares regression line, is shown below. The plot shows a strong, positive linear association between the number of beaver-caused stumps and the number of beetle larvae clusters.



(b) The least-squares regression line is $\hat{y} = -1.29 + 11.89x$. (c) The residual plot is shown below. The linear model appears to provide a very good fit.

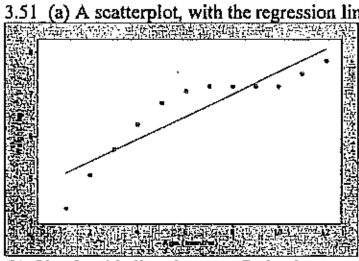


(d) About 84% of the variation in the number of beetle larvae clusters is accounted for by the linear relationship with the number of stumps.

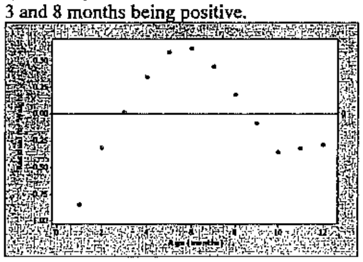
3.49 (a) The slope (1.507) says that, on average, BOD rises (falls) by 1.507 mg/L for every 1 mg/L increase (decrease) in TOC. (b) When TOC = 0 mg/L, the predicted BOD level is –55.43 mg/L. The negative value of BOD was obtained because values of TOC near zero were probably not included in the study. This is another example where the intercept does not have a practical interpretation.

3.50 (a) The least-squares line for predicting y = GPA from x = IQ has slope $b = 0.6337\left(\frac{2.1}{13.17}\right) = 0.101$ and intercept $a = 7.447 - 0.101 \times 108.9 = -3.5519$. Thus, the regression line is $\hat{y} = -3.5519 + 0.101x$. (b) $r^2 = (0.6337)^2 = 0.4016$. Thus, 40.16% of the variation in GPA is accounted for by the linear relationship with IQ. (c) The predicted GPA for this student is $\hat{y} = -3.5519 + 0.101 \times 103 = 6.8511$ and the residual is $6.8511 - 0.53 = 6.3211$.

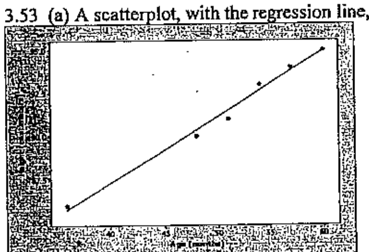
3.51 (a) A scatterplot, with the regression line, is shown below.



(b) Clearly, this line does not fit the data very well; the data show a clearly curved pattern. (c) The residuals sum to 0.01 (the result of roundoff error). The residual plot below shows a clear quadratic pattern, with the first two and the last four residuals being negative and those between 3 and 8 months being positive.



3.53 (a) A scatterplot, with the regression line, is shown below.



(b) The regression line for predicting y = height from x = age is $\hat{y} = 71.95 + 0.3833x$. (c) When x = 40 months: $\hat{y}$ = 87.28 cm. When x = 60 months: $\hat{y}$ = 94.95 cm. (d) A change of 6 cm in 12 months is 0.5 cm/month. Sarah is growing at about 0.38 cm/month; more slowly than normal.

3.55 (a) The slope of the regression line for predicting final-exam score from pre-exam totals is $b = 0.6\left(\frac{8}{30}\right) = 0.16$; for every extra point earned on the midterm, the score on the final exam increases by a mean of 0.16. The intercept of the regression line is $a = 75 - 0.16 \times 280 = 30.2$; if the student had a pre-exam total of 0 points, the predicted score on the final would be 30.2. (b) Julie's predicted final exam score is $\hat{y} = 30.2 + 0.16 \times 300 = 78.2$. (c) $r^2 = 0.36$, so only 36% of the variability in the final exam scores is accounted for by the linear relationship with pre-exam totals. About 64% of the individual variation is not accounted for by the least squares regression line, so Julie has a good reason to think this is not a good estimate.