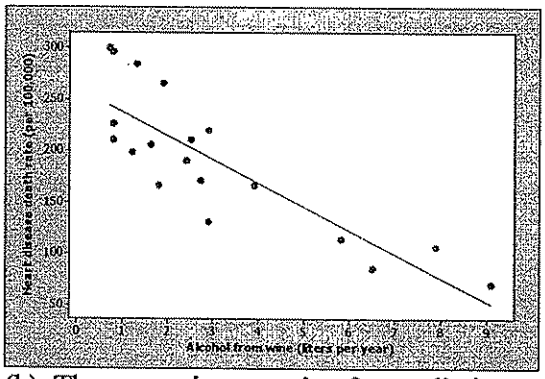
Review for Ch. 3 Test pp251-255

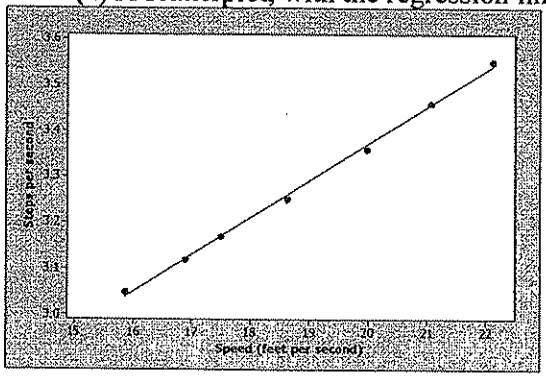
3.79 (a) A scatterplot, with the regression line, is shown below. There is a negative association between alcohol consumption and heart disease.



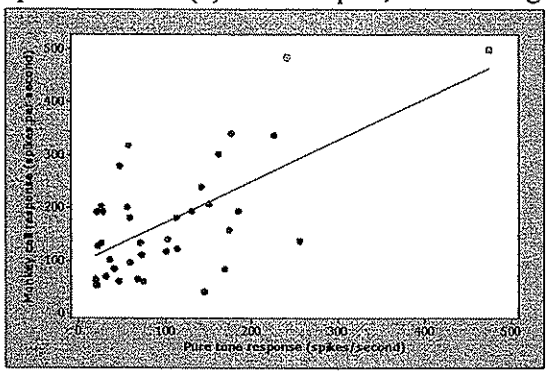
(b) The regression equation for predicting y = heart disease death rate from x = alcohol consumption is $\hat{y} \doteq 260.56 - 22.969x$. The slope provides an estimate for the average decrease (slope is negative) in the heart disease death rate for a one liter increase in wine consumption. Thus, for every extra liter of alcohol consumed, the heart disease death rate decreases on average by about 23 per 100,000. The intercept provides an estimate for the average death rate (261 per 100,000) when no wine is consumed. (c) The correlation is $r = -0.843$, which indicates a strong negative association between wine consumption and heart disease death rate. $r^2 = 0.71$, so 71% of the variation in death rate is accounted for by the linear relationship with wine consumption. (d) The predicted heart disease death rate is $\hat{y} \doteq 260.56 - 22.969 \times 4 \doteq 168.68$. (e) No. Positive r indicates that the least-squares line must have positive slope, negative r indicates that it must have negative slope. The direction of the association and the slope of the least-squares line must always have the same sign. Recall $b = r\left(\frac{s_y}{s_x}\right)$ and the standard deviations are always nonnegative.

3.80 (a) The point at the far left of the plot (Alaska) and the point at the extreme right (Florida) are unusual. Alaska may be an outlier because its cold temperatures discourage older residents from remaining in the state. Florida is unusual because many individuals choose to retire there. (b) The linear association is positive, but very weak. (c) The outliers tend to suggest a stronger linear trend than the other points and will be influential on the correlation. Thus, the correlation with the outliers is $r = 0.267$, and the correlation without the outliers is $r = 0.067$.

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



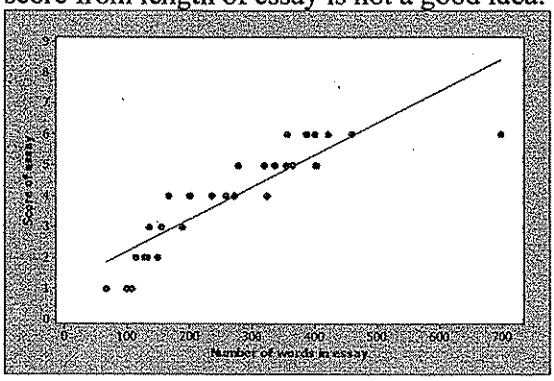
3.83 (a) One possible measure of the difference is the mean response: 106.2 spikes/second for pure tones and 176.6 spikes/second for monkey calls—an average of an additional 70.4 spikes/second. (b) A scatterplot, with the regression line $\hat{y} = 93.9 + 0.778x$, is shown below.

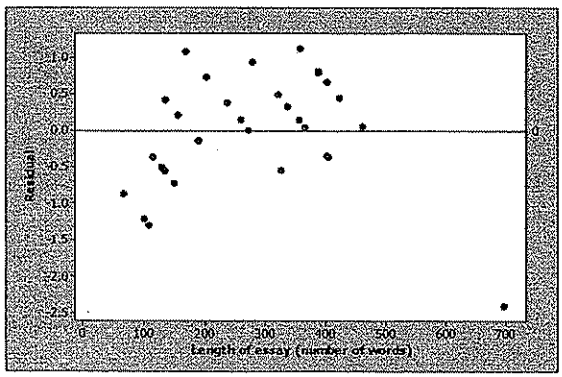


The third point (pure tone 241, call 485 spikes/second) has the largest residual; it is circled. The first point (474 and 500 spikes/second) is an outlier in the x direction; it is marked with a square. (c) The correlation drops only slightly (from 0.6386 to 0.6101) when the third point is removed; it drops more drastically (to 0.4793) without the first point. (d) Without the first point, the regression line is $\hat{y} = 101 + 0.693x$; without the third point, it is $\hat{y} = 98.4 + 0.679x$.

3.85 The slope is $b = 0.5\left(\frac{2.7}{2.5}\right) = 0.54$. The regression line, shown below, for predicting $y =$ husband's height from $x =$ wife's height is $\hat{y} = 33.67 + 0.54x$.

3.86 *Who?* The individuals are the essays provided by students on the new SAT writing test. *What?* The variables are the word count (length of essay) and score. Both variables are quantitative and take on integer values. *Why?* The data were collected to investigate the relationship between length of the essay and score. *When, where, how, and by whom?* The data were collected after the first administration of the new SAT writing test in March, 2005. Dr. Perelman may have obtained the data from the Educational Testing Service or from colleagues who scored the essays. *Graphs:* The scatterplot below, with the regression line included, shows a relationship between length of the essay and score, but the relationship appears to be nonlinear. The residual plot also shows a clear pattern, so using the least-squares regression line to predict score from length of essay is not a good idea.





Numerical summaries: The correlation between word count and score is 0.881. The least squares regression line for predicting $y =$ score from $x =$ word count is $\hat{y} = 1.1728 + 0.0104$. This line accounts for about 77.5% of the variation in score. *Interpretation:* Even though the scatterplot shows a moderately strong positive association between length of the essay and score, we do not want to jump to conclusions about the nature of this relationship. Better students tend to give more thorough explanations so there could be another reason why the longer essays tend to get high scores. In fact, a careful look at the scatterplot reveals considerably more variation in the length of the essays for students who received a score of 4, 5, or 6. If Dr. Perelman's made his second conclusion about being right over 90% of the time by rounding the correlation coefficient from 0.88 to 0.9, then he made a serious mistake with his interpretation of the correlation coefficient. If scores were assigned by simply sorting the word counts from smallest to largest, the error rate would be much larger than 10%.