## AP Sample Test

**AP1.** A. There aren't any points in the upper left-hand corner because the oldest child has to be older (or the same age, in the case of twins) than the youngest child. Thus all points must lie on or below the line $y = x$.

**AP2.** C. The predicted birthrate is $-0.38 \cdot 60 + 53.5 = 30.7$, so the residual is the actual birthrate of 47 minus this prediction, $47 - 30.7 = 16.3$.

**AP3.** C. Curvature in the residual plot of a linear regression is a sign of curvature in the original plot, so statement I is true. When points in the residual plot lie below the line $y = 0$, the points in the original scatterplot lie below the regression line and so the prediction is too large. Thus, statement II is true. Statement III is false because, for example, the pattern could be exponential with a high correlation.

**AP4.** A. Outliers should not be removed permanently from a data set simply because they are outliers. Further investigation is needed, as described in B, C, and D.

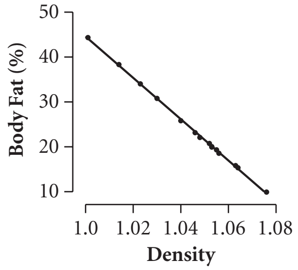
**AP5.** A. B is incorrect because the slope of the regression equation is positive, so the correlation is 0.228. C is incorrect because the value of $R^2$ doesn't give any information about linearity versus curvature. E is incorrect because it implies that each person's satisfaction tends to increase over their stay in the hospital. Instead, there may be a lurking variable of age: older people have to stay longer and they also tend to be more satisfied with their care. Or, the lurking variable might be severity of the problem. The more seriously ill a patient is, the longer they tend to have to stay, and the more grateful they are for the care they were given.

**AP6.** D. In the year 2000, $t = 50$, so $\log_{10}(\textit{population}) = 0.01 \cdot 50 + 7 = 7.5$, and thus $\textit{population} = 10^{7.5} \approx 31{,}622{,}777$ and 31,600,000 is the closest.

**AP7.** E. Choice A is a poor choice because each point represents a different Barbarian, and so does not establish trends in a particular Barbarian. Choice B is closer to an interpretation of the intercept than to the slope. Choice C might be close to correct if the $y$-intercept was near zero, but here it's far from zero. For choice D, you would have to know the scores on the two sections had equal standard deviations before making this interpretation.

**AP8.** D. Note that the correct statement E is equivalent to saying that 81% of the variation in the number of raids among Barbarians is explained by personal cleanliness.

**AP9.** **a.** **i.** From the plot, this line looks to be a good fit.

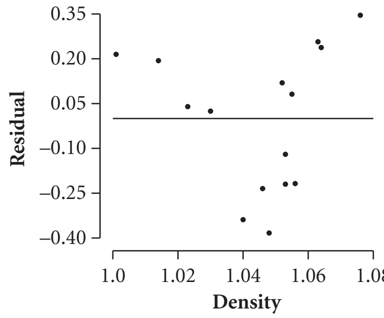



**ii.** The regression equation is $\hat{y} = 505.254 - 460.678x$ and the analysis is as follows.

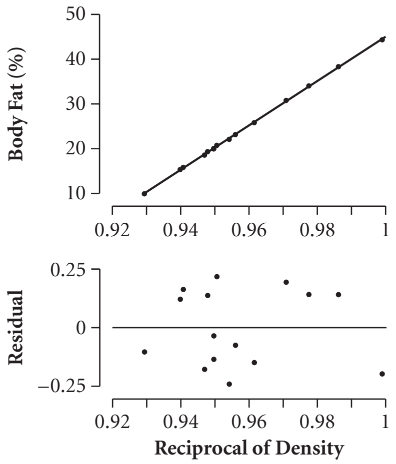
```
Dependent variable is:  % Fat
No Selector
R squared = 99.9%       R squared (adjusted) = 99.9%
s = 0.2443 with 15 - 2 = 13 degrees of freedom

           Sum of
Source     Squares      df      Mean Square   F-ratio
Regression 1206.48      1       1206.48       20223
Residual      0.775560  13         0.059658

                        s.e. of
Variable   Coefficient  Coeff   t-ratio       prob
Constant    505.254     3.386    149          ≤0.0001
Density    -460.678     3.239   -142          ≤0.0001
```

**iii.** The $r^2$ value of 0.999 seems to confirm that this model is a good fit.

**iv.** The residual plot uncovers some problems; perhaps we could do better!




**b.** **i.** As the percentage of fat increases, the body density decreases. Perhaps the positive association between the *reciprocal of density* and the *percentage of fat* would be easier to model. The pertinent plots and the regression analysis are shown on the next page.




```
Dependent variable is:  % Fat
No Selector
R squared = 100.0%      R squared (adjusted) = 100.0%
s = 0.1690 with 15 - 2 = 13 degrees of freedom

                Sum of
Source          Squares         df      Mean Square     F-ratio
Regression      1206.88         1       1206.88         42246
Residual           0.371389     13         0.028568

                                s.e. of
Variable        Coefficient     Coeff   t-ratio         prob
Constant        -450.632         2.309  -195            ≤0.0001
1/Density        495.654        -0.412   206            ≤0.0001
```

The residuals show less pattern; the plot is more like one of random scatter, suggesting that this is a better model.

The regression line has the equation

$$\%\ body\ fat = -450.63 + 495.65\left(\frac{1}{density}\right)$$

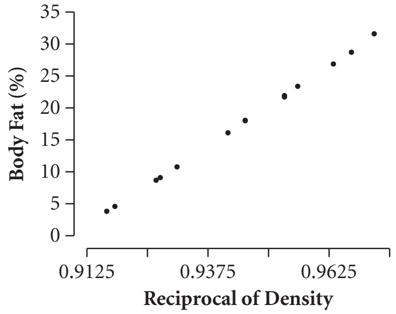
which is very close to the Siri equation.

**ii.** The correlation is close to 1 for both models, but the second proves to be a better fitting model than the first. Moral: Never use correlation as the only criterion for choosing a model.

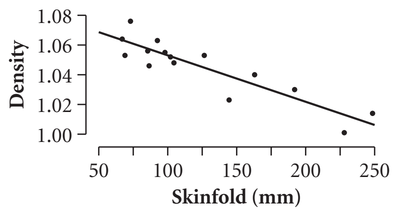
**iii.** *Percent body fat* as a function of log(*density*) works almost as well as Siri's model. The residual plot, however, has a hint of a pattern.

**AP10. a.** For women, the regression equation was $\%\ body\ fat = -450.63 + 495.65\left(\frac{1}{density}\right)$, almost identical to Siri's model. (See AP9.) The relationship is very strong and linear, with correlation almost equal to 1 and no pattern in the residual plot.

Using the same variables for men, the relationship is again extraordinarily linear (see the next scatterplot), with a correlation near 1. But this time the regression equation is $\%\ body\ fat = -453.7 + 498.97\left(\frac{1}{density}\right)$, which is similar to Siri's model but not nearly as close as the equation for women. Thus, the model fits better for women than for men.




**b.** For women, this scatterplot of *density* against *skinfold* is quite linear and does not require re-expression. Thus, a reasonable model is the regression equation, $density = 1.084 - 0.000311$ skinfold. The correlation is $-0.897$.




For men, the relationship is less strong and has some curvature (see the scatterplot); however, the linear model is an adequate one. The equation is $density = 1.105 - 0.000295\ skinfold$. The residual plot shows some heteroscedasticity.