10. ANSWERS TO PROBLEMS

1. Brain and body mass of mammals

a. The correlation coefficient r is 0.9342.

b. The scatterplot of data points does not conform to a bivariate normality pattern. Rather most of the points are clumped, and there are a few outliers.

c. *Spearman's non-parametric correlation test*
Ho: $\rho_s = 0$
Ha: $\rho_s \neq 0$
reject Ho. $r_s = 0.9535$, df=60, $P < 0.0001$. Body mass and brain size are correlated in mammals.

d. Taking the log of the data generates a scatterplot of the data that conforms reasonably well to a bivariate normal distribution.

e. The following is the equation of our regression line, using the log-transformed data (natural log with base e):

$$Y = 2.1271 + 0.7545\,X,$$

where Y is log(brain mass) and X is log(body mass). Or, you can express it as:

$$\text{Log(brain mass)} = 2.1271 + 0.7545 \text{ log(body mass)}$$

$r^2 = 0.9197$. The r^2 value represents the strength of our relationship. In other words, with the linear regression line X "explains" 92% of the variation in Y. This is a strong relationship.

f. The assumptions of linear regression are 1) that the true relationship is linear (the mean of Y for each X lies on the true regression line); 2) that the distribution of Y-values for each X is normal; and 3) that the variance in Y is the same at every X. Reasonable approaches to testing these assumptions are as follows:

Linear fit: A spline fit (lambda=10) suggests that the relationship is approximately linear. (Carrying out a "Lack Of Fit" test to the line is another way to test the assumption of linearity, and JMP IN provides an F-test of this null hypothesis. However, we haven't learned this method and its assumptions.)

Normal Y-values for each X: A plot of the residuals indicates no major departures from the assumptions. E.g., residuals are distributed above and below the line with little obvious skew or presence of outliers. To take this further, we can save the residuals and test for normality:
Ho: Residuals are normally distributed
Ha: Residuals are not normally distributed
Shapiro-Wilks W=0.9835, n=62, P=0.8104 do not reject Ho.

Equal variances in Y: A plot of the residuals gives little reason to doubt the assumption. There do not appear to be trends in the scatter of points about and below the line with different X-

values. (Substantial departures from the assumption of equal variances might show up as a non-normal distribution of residuals, but not always).

g. *Test if the slope of the regression is significantly different from zero*
Ho: $\beta = 0$
Ha: $\beta \neq 0$
$F = 687.2601$, df=1, 60, $P < 0.0001$. reject Ho. The slope differs significantly from zero. In an assignment, you should include the ANOVA table. (Note: the slope may also be tested with a *t*-test: t=26.22, df=60, P<0.0001).

h. The confidence bands measure the upper and lower bounds of 95% confidence intervals for the predicted mean *Y* value at each *X*.

i. The prediction intervals are wider than the confidence bands because they take into account the variance of the individual *Y*'s for a given *X*.

j. The species with the largest brain, after correcting for body size differences between species using regression, is the human. You can see this by examining the residual plot. Click on the point with the highest positive residual: this is point #32 (refer back to the data table, which will have row 32 selected). The water opossum, on the other hand, is the farthest data point below the line in the residual plot, in other words it has the smallest brain size after taking account of differences in body mass.

k. Other species with relatively large brains are the primates (rhesus monkey, owl monkey, baboon) and the smallest brains belong to the giant armadillo, musk shrew, pig and tenrec.

l. The method of collection of this data set is unknown, so independence would be difficult to assess here without additional information. However, we will learn of a potential complication that arises when using species values as observations, even when species are somehow "randomly" sampled. Species values are not expected to be independent because closely related species are expected to be more similar in *Y* than the average of species picked at random (and also *X*, but what matters in regression is independence of residuals in *Y*). Methods to deal with this problem take account of the phylogeny of relationships between species. Various programs to analyze species data are distributed free of charge (e.g., COMPARE at http://compare.bio.indiana.edu/)

2. Deleterious mutation rates

a. By these estimates the human/chimpanzee have the highest genomic deleterious mutation rates per generation, and *Drosophila* have the lowest.

b. *Correlation test*
Log transforming the data helps it to meet the assumption of bivariate normality reasonably well.

Ho: $\rho = 0$
Ha: $\rho \neq 0$
reject Ho. $r = 0.9625$, df=7, $P < 0.0001$. Deleterious mutation rates are correlated with generation time.

3. GAG concentration and child age

a. The assumption of linearity is not met, but a log transformation of both variables improves matters (use log(age+1) to accommodate values of age=0). A spline fit of the log-transformed data looks close to linear. The only worrisome trend is that the mean *Y* for very large *X*-values may dip a little, but this discrepancy, if real, is not large.
The residuals are fairly evenly distributed above and below the regression line. There is one worrisome trend, however: the residuals corresponding to the largest values of *X* appear to have a higher variance than the rest.

A fit of the residuals to a normal distribution yielded the following:
Ho: Residuals are normally distributed
Ha: Residuals are not normally distributed
Shapiro-Wilks W=0.9796, n=314, P=0.1640. Do not reject Ho. Our sample size is also quite large, so the parametric test should be robust.

Linear regression
Ho: $\beta = 0$
Ha: $\beta \neq 0$
reject Ho. Analysis of fit $F = 1198.16$, df = 1, 312, $P < 0.0001$. The slope is significantly different from zero. The equation of the regression line is:

$$\log(\text{GAG}) = 3.0601 - 0.5465 \log(\text{age}+1)$$

$r^2 = 0.7869$

b. For predictive purposes, to determine whether a child is atypical for his/her age, it would be useful to add the prediction intervals to the plot.

4. Survival of males and birth weight

a. No, the line does not look linear. Rather it is a concentric curve that asymptotes at about a birth weight of 5 pounds.

b. Using a spline curve (lambda=10) and eyeing survival probability, the following predictions of survival are made for different birth weights:

Birth weight	Survival
3 pounds	0.25
4 pounds	0.75
7 pounds	>0.95
10 pounds	0.95

This suggests a steep rising curve initially with increasing birth weight, a flat dome between about 6 to 9 ponds, and a decline at 10 or more pounds. Transformations would probably not work since a power relationship would not be the same for entire range of *X*.

c. Based on spline fits, the relationship between survival and gestation is similar but not identical to that between survival and birth weight. Survival seems to show a steeper drop at high birth weights than at high gestation periods, although the shapes of the two curves are similar over most of the range of data.

d. The assumption of bivariate normality does not look justified, so I used a non-parametric test, the Spearman's rank correlation
Ho: $\rho_s = 0$
Ha: $\rho_s \neq 0$
reject Ho. $r_s = 0.5241$, $P < 0.0001$. Gestation time and birth weight are correlated, but the correlation is not as strong as you might expect (the linear correlation is only $r = 0.4722$).