University of York
Department of Health Sciences
Applied Biostatistics
Suggested answers to exercise: Correlation and Regression

Question 1

a) *What method would be used to calculate the 68g per 5 mm Hg?* 68g per 5 mm Hg is calculated using simple linear regression. This method estimates the nature of the relationship between two continuous variables, here mean 24 hour diastolic blood pressure and infant birthweight. The quantity given comes directly from the slope of the line. The method works by calculating the line of best fit through the data using the principle of least squares.

b) *What assumptions would the method require?* This method would require that birthweight followed a Normal distribution with the same variance for any given values of mean 24 hour diastolic blood pressure.

c) *What is meant by 'increase' and 'decrease' here? Do they mean that when a woman's blood pressure went down her baby's weight went up?* In this study blood pressure at 28 weeks is related to birthweight, a quantity which is measured once, at birth. This analysis is not investigating how changes in blood pressure within individuals might affect the growth of the baby. The increase in birthweight associated with a decrease in blood pressure therefore refers to the mean effect rather than an effect for an individual. The difference in mean birthweight between two groups of women whose mean blood pressure differ by 5 mm Hg would therefore be 68g, with the direction of the difference being that women with lower mean blood pressure have greater mean birthweight.

Question 2

a) *What are the interpretations of the numbers 55.9 and 0.22 in the regression equation?* The regression line shows the estimated mean ear size for given age. The line has slope 0.22, i.e. mean ear size increases by 0.22 mm for each year of age. When age is zero, the line would cross the vertical axis at 55.9. This does not mean that babies have ears of this size, because we would be extrapolating beyond the data. We cannot do this because we cannot assume that the straight line relationship will also be valid for children.

b) *Are the assumptions about the data are required for the regression analysis satisfied here?* We assume that the deviations from the regression line follow a Normal distribution and have uniform variation along the line. The second assumption looks very reasonable from the figure. The spread of the data about the line is very similar all the way along. It is difficult to tell about the Normal distribution. There appears to be some deviation from linearity, as the points tend to be below the line at each end. This may be because the older people will tend to be women, who are smaller than men.

c) *Are the conclusions justified by the data?* This is a cross-sectional not a longitudinal study, and the uncertainty in the estimate should be included. Strictly speaking, the conclusions should read 'It seems therefore that older people have bigger ears (on average by between 0.17 and 0.27 mm per year of age)'. It could be that the ears of different birth cohorts differ. After all, different birth cohorts have different mean heights at the same age. (Personally, I think that the author's interpretation is correct, but we cannot draw this conclusion from these data alone.)

Question 3

a) *What is meant by 'correlated' and 'r = 0.22'?* Two variables are correlated if when one has high values the other has high values and when one has low values the other has low values, or if when one has high values the other has low values and when one has low values the other has high values. r is the correlation coefficient which measures the strength of the linear relationship between two continuous variables. It lies between -1 and $+1$ with 0 showing no relationship. $r = 0.22$ is a positive correlation, showing that adult height tends to be greater for subjects with high birthweight but the relationship is weak and would be hard to see on a scatter diagram.

b) *What assumptions are required for the calculation of the P value?* One of the two variables must follow a Normal distribution for the P value to be valid.

c) *What can we conclude about the relationship between adult height and birth weight?* In the population which these subjects represent, adult height is related to birthweight, but the relationship is weak. Tall men tend to have been heavier at birth. We cannot conclude from these data, however, that the relationship is causal.