University of York

Department of Health Sciences

Applied Biostatistics

Suggested answers to exercise: Multiple regression

Question 1

a) *What is meant by 'multiple regression analysis'?* This is a statistical method used where we have a continuous outcome variable, here systolic blood pressure, and several possible predictors, here recovery index, age, race, area of residence, and ponderal index. The method estimates the relationship between the outcome and each predictor adjusting for all other predictors in the model. Here, the outcome or dependent variable is systolic blood pressure. The predictor, independent, or explanatory variables are recovery index, age, race, area of residence, and ponderal index.

b) *What is meant by the terms 'b', 'SE b' and '95% CI'?* '*b*' is the coefficient of recovery index in a multiple regression equation. It means that for two groups of subjects whose recovery index differs by one unit, and who all have the same age, race, area and ponderal index, the difference in mean systolic blood pressure will be *b* units. It is found by the method of least squares. 'SE *b*' is the standard error of the estimate of *b*. Different samples of the same size would give different estimates of *b*. SE *b* is the estimated standard deviation of the possible estimates of *b*. '95 % CI' is the 95 % confidence interval for *b*. For 95 % of possible samples, this range of values will include the value of *b* for the whole population.

c) *What assumptions about the variables are required for these analyses to be valid?* The assumptions are that the differences between the observed systolic blood pressure and the systolic blood pressure predicted by the regression equation (residuals or deviations from the regression) follow a Normal distribution and that the variance of this distribution is uniform, i.e. unrelated to recovery index, age, race, area and ponderal index. The relationships should be linear.

d) *Why was the regression adjusted and what does this mean?* Regression may be adjusted because the variable of interest, systolic blood pressure, is related to the adjusting variables, age, race, area, and ponderal index. We want to estimate the effect of recovery index on systolic pressure in children on the same age, race, area, and ponderal index.

e) *What would be the effect of adjusting for race if systolic blood pressure were related to race and recovery index were not?* If recovery index is independent of adjusting variables, the adjustment will reduce the variability of the deviations and so make the estimate of b better, in that the confidence interval will be narrower, but the estimate will not be changed.

f) *What would be the effects of adjusting for ponderal index if blood pressure and recovery index were both related to ponderal index?* If recovery index is related to the adjusting variables, the adjustment will reduce or remove the spurious relationship between recovery index and systolic blood pressure produced by both being independently related to something else. Both the coefficient and its standard error are likely to change.

Question 2

a) *What is meant by 'multiple logistic regression'?* Multiple logistic regression or logistic regression is a multifactorial statistical method used when we have a dichotomous outcome, here reporting passive smoking as harmful or not. The method allows us to estimate the relationship between this outcome and several predictor variables, here affiliation with the tobacco industry and others. It allows us to estimate the odds ratios for each predictor adjusted for all others in the model.

b) *What is wrong with the interpretation of the odds ratio by the BMJ writer?* '88 times more likely' suggests that the odds ratio has been interpreted as the increase in risk, whereas it actually represents the increase in odds. These are only similar if the condition of interest is rare. Also, the odds are 87 times greater or 88 times as great rather than 88 times more likely. This is not very important here, but would be if the odds ratio were closer to 1.0. From the *JAMA* paper, we can calculate the unadjusted odds ratio = 94. The corresponding relative risk = 7.

Question 3

a) *What is meant by interaction?* An interaction between Jarman score and fluoridation status means that the relationship between mean tooth decay and Jarman score is different in areas with different levels of fluoridation. This can be seen from the graph where the regression lines for unfluoridated, artificially fluoridated and naturally fluoridated areas are not parallel.

b) *How would you interpret a statistically significant interaction here?* We could interpret this as meaning that fluoridation appears to be more effective in areas of high deprivation. There are small differences between high and low fluoride areas when Jarman score is low but a large difference when Jarman score is high. Alternatively, we could say that tooth decay is strongly related to Jarman score in low fluoride areas but not in high fluoride areas.