University of York

Department of Health Sciences

Applied Biostatistics

Exercise: Multiple regression

Question 1

In a study of physical fitness and cardiovascular risk factors in children, blood pressure and recovery index (post exercise recovery rate, an indicator of fitness) were measured (Hoffman and Walter 1989). Multiple regression was used to look at the relationship between systolic blood pressure and recovery index, adjusted for age, race, area of residence and ponderal index (wt/ht^2). For the boys, the adjusted regression coefficient of systolic blood pressure on recovery index was given as follows: $b = -0.086$, SE $b = 0.039$, 95% CI = –0.162 to –0.010.

a) What is meant by 'multiple regression analysis'?

b) What is meant by the terms '*b*', 'SE *b*' and '95% CI'?

c) What assumptions about the variables are required for these analyses to be valid?

d) Why was the regression adjusted and what does this mean?

e) What would be the effect of adjusting for race if systolic blood pressure were related to race and recovery index were not?

f) What would be the effects of adjusting for ponderal index if blood pressure and recovery index were both related to ponderal index?

Question 2

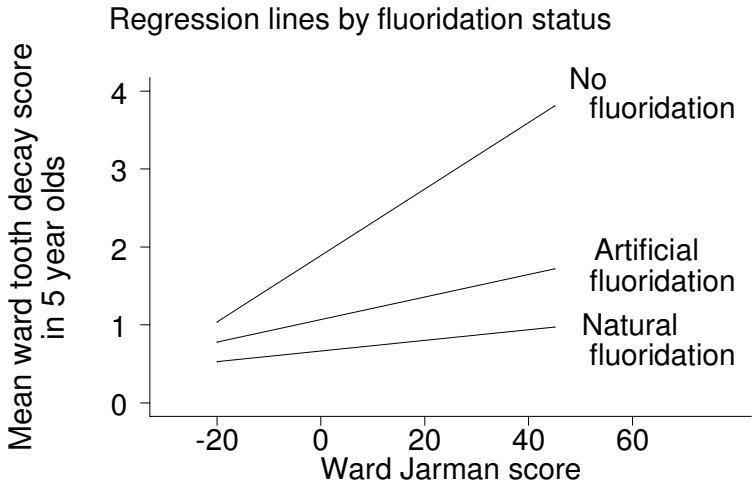
A news item in the *BMJ* (Wise 1998) reported the results of a *JAMA* (Barnes and Bero 1998) study which investigated possible bias in the publication of studies of effects of passive smoking. The *BMJ* item reported that 'a review written by authors with affiliations to the tobacco industry is 88 times more likely to conclude that passive smoking is not harmful than if the review was written by authors with no connection to the tobacco industry.' This information was taken from the *JAMA* article where the following was presented: 'In multiple logistic regression analyses . . . the only factor associated with concluding that passive smoking is not harmful was whether an author was affiliated with the tobacco industry (odds ratio, 88.4; 95 % confidence interval, 16.4 to 476; $P < 0.001$)'.

a) What is meant by 'multiple logistic regression'?

b) What is wrong with the interpretation of the odds ratio by the *BMJ* writer?

Question 3

An ecological study examined the effect of water fluoridation on tooth decay in 5 year old children using data collected at the level of the electoral ward. The electoral wards included were in three areas where the water supply was either unfluoridated, artificially fluoridated or naturally fluoridated. A multiple linear regression model was fitted with mean tooth decay in the ward as the outcome and with predictors Jarman underprivileged area score for each ward and fluoridation status (unfluoridated, artificially fluoridated or naturally fluoridated). A high Jarman score indicates an area with high deprivation. The authors reported that there was a significant interaction between the effects of Jarman score and water fluoridation on tooth decay. A graph similar to this was given (Jones *et al.*, 1997).



a) What is meant by interaction?

b) How would you interpret a statistically significant interaction here?

References

Jones, C.M., Taylor, G.O., Whittle, J.G., Evans, D., and Trotter, D.P. (1997) (Water fluoridation, tooth decay in 5 year olds, and social deprivation measured by the Jarman score: analysis of data from British dental surveys. *British Medical Journal* **315**, 514-7.

Hofman, A. and Walter, H.J. (1989) The association between physical fitness and cardiovascular disease risk factors in children in a five-year follow-up study. *International Journal of Epidemiology* **18**, 830-5.

Barnes, D.E. and Bero, L.A. (1998) Why review articles on the health effects of passive smoking reach different conclusions. *JAMA* **279**, 1566-70.

Wise, J. (1998) Links to tobacco industry influences review conclusions. *British Medical Journal* **316**, 1533.

Questions from Martin Bland and Janet Peacock: *Statistical Questions in Evidence-based Medicine*, Oxford University Press, Oxford, 2000.