

LAB 3: Notes

In this lab we will analyze the backpain data. This is a simulated dataset patterned after the back pain example from the course. The variables in the dataset are:

Doctor – Doctor ID number.

Cost – Cost of treatment in dollars.

Logcost – Logarithm of cost.

Actlim – Number of days in the past six months in which activity has been limited.

Undrstnd – Did the patient understand the doctor's advice (1=yes, 0=no)

Age – Age of the patient.

Educ – Education of the patient (0 means ≤ 12 years, 1 means 13-16 years, 2 means ≥ 16 years education).

Thoraic – Whether the back pain was cervical/thoraic or other (1=yes, 0=no).

Pracstyl – Practice style of the doctor (0=low, 1=medium, 2=high frequency of prescription of medicines and hospitalization for treatment of back pain).

- 1) Analyze the response Undrstnd using SAS Proc GENMOD. For now we will ignore the fact that we have multiple observations per doctor.
 - a) Declare Educ and Pracstyl to be categorical predictors.
 - b) Fit a model with Age, Educ and Pracstyle as predictors
 - c) Provide an interpretation of the coefficient for age.

The age coefficient is -0.0223 , and the model is a logit model so it is a log odds. Exponentiating gives $\exp(-0.0223) = .97795$ or about .98. So the odds of understanding decrease by about 2% with each increasing year of age. In terms of decade of age, $\exp(10 \cdot -0.0223) = \exp(-0.223) = .8001$ or about .80. So with each increase in a decade of age, the odds of understanding decrease by about 20%.

- 2) Analyze the response Actlim using SAS Proc GENMOD. Follow the steps as in question 1)

For actlim we declare a Poisson distribution as a start. That generates an age coefficient of 0.0075 with a p -value of approximately 0. The default is a log link. So we exponentiate the coefficient to get $\exp(0.0075) = 1.0075$. The interpretation is then that associated with each increase in year of age is a .75% increase in activity limitation days. Or, associated with a decade increase in age is $\exp(0.075) = 1.078$ or about an 8% increase in activity limitation days.

However we need to be wary of the Poisson assumption that the mean equals

*the variance. Calculating means and variances of the data by, say, levels of education shows the data are highly overdispersed (variance > mean). Using the PSCALE option on the MODEL statement estimates a constant of proportionality (variance = ϕ *mean, with $\phi = 7.25$) and the p-value for age becomes approximately 0.08.*

- 3) Repeat questions 1) and 2) but now add an extra line below the model statement: repeated subject=doctor / type=exch;
What is the purpose of this command? What differences do you notice?

This command accommodates the repeated measures on the doctor by using a robust variance estimate. However, the first part of the analysis (“Analysis of initial parameter estimates”) is nothing more than the naïve analysis assuming no correlation).