

## Generalized Linear Models for Independent Data: For practice\*

---

These two problems provide practice with logistic regression and Poisson regression. If you feel you need this practice, you should complete these exercises **BEFORE** we begin covering generalized linear mixed models for correlated data in class.

Suggested solutions will be put on the class web site.

1. A school district is interested in comparing two driver education curricula, called the “Safe performace curriculum” (SPC) and the “Pre-driver licensing curriculum” (PDLC). In particular, they want to know if students who participated in SPC were less likely to have a collision than students who participated in PDLC. Three high schools offered both curricula during one year. 300 students who wanted driver education (100 from each school) were assigned to one of SPC or PDLC (roughly half and half). Gender, grade point average (GPA, 0.0 to 4.0), school (1, 2, or 3), and curriculum (SPC or PDLC) were recorded at the beginning of the driver education. One year after completion of the education, students were asked if they had had a collision (yes/no) in the past year.

Based on the results of this study, only one curriculum will be chosen by the district to be used in future years. Because of this, the district administrators are not interested in whether the curriculum worked better for certain subgroups of students (e.g., those with higher GPA, or those in school 2). In statistical terms, this means they are only interested in testing the overall curriculum effect, not in any interactions with the curriculum effect. The data can be found in `driver.dat`.

USING ONLY THE `year = 1` OBSERVATIONS,

- (a) Fit a generalized linear model on collision (yes/no) with a logit link using `PROC GENMOD`. Include a main effect for curriculum, and all main effects and interactions among the three adjusting variables: school, gender, and GPA. From this full model, reduce the model down to a final model by removing non-significant effects. Test for significance using standard likelihood ratio tests ( $T^* = (-2 \log \ell_{reduced} - (-2 \log \ell_{full}))$ ). If you have never used `PROC GENMOD` before, your code will look like this:

```
proc genmod data=dat;
  class categorical covariate list ;
  model collision = covariate list / dist = bin link = logit;
```

Verify on the output that you are modeling  $P[\textit{accident}]$  and not  $P[\textit{no accident}]$ .

- (b) What is the probability of a collision for a female with GPA of 3.0? A male with GPA of 4.0?

---

\*File: `~/Teaching/Correlated/Review/GzLMpractice.tex`; last modified August 14, 2007.

- (c) Suppose there were no interactions in your final model. What is the interpretation of  $\exp\{\beta_{GPA}\}$ ? What is the interpretation of  $\exp\{\beta_{gender}\}$ ?
  - (d) Re-fit your *final model* only using logistic regression in whatever software you normally use. (In SAS, that might be PROC LOGISTIC.) Do you get exactly the same fitted regression coefficients? Do you get exactly the same standard errors for the regression coefficients?
2. It has been hypothesized that there is a link between Sudden Infant Death Syndrome (SIDS) and environmental temperature. Between 8 January 1979 and 31 November 1985 (242 days), each day's average temperature (in degrees Celsius) and the number of SIDS deaths reported on the *following* day were recorded. The data set `sids.dat` contains these pairs of observations (count followed by temperature) which can be read as follows:

```
data dat;
  infile 'sids.dat';
  input count temp @@;
run;
```

Assume the days can be treated as independent observations.

- (a) Fit a Poisson regression to the number of SIDS deaths with temperature as the only covariate using PROC GENMOD. Your code will look like this:

```
proc genmod data=dat;
  model deaths = temp / dist = poisson link = log;
```

Is there sufficient evidence to support the hypothesized link between SIDS and temperature? What is the interpretation of  $\exp\{\beta_{temp}\}$ ?

- (b) Replace `link = log` with `link = identity` and re-fit the model. Do your results change?
- (c) Now take a log transform of the numbers of deaths and fit a simple linear regression of log deaths on temperature. Do your results change? (Because some days have 0 deaths, you will need to add a small constant to *every* number of deaths, e.g. 1/6, in order to take the log transformation.)
- (d) Write down the model equation and assumptions for each of the three models you fit. Make sure you understand the differences between these three models.