.

ST 732, HOMEWORK 6, SPRING 2007

1. A study was conducted to compare two treatments for patients with bladder cancer. Each of the $n = 100$ subjects recruited into the study had recently had surgery to remove the tumor; at baseline, each was then randomized to receive either an experimental treatment, thiotepa (coded as 1), or placebo (coded as 0), which they were to take for the next 12 months. At 12 months, whether or not a new tumor had developed (0 = no, 1 = yes, the response) was recorded. Also recorded was the gender of the subject (0 = female, 1 = male).

   The data are in the file `bladder.dat` on the class web page; each data record corresponds to a single subject. The columns are (1) subject id; (2) the response, new tumor at 12 months (0 = no, 1 = yes); (3) treatment (0 = placebo, 1 = thiotepa); and (4) gender (0 = female, 1 = male).

   In parts (b)–(i) of this problem, you will build up a SAS program that does several different things through different calls to `proc genmod`. *The purpose of the problem is to illustrate to you some of the features of* `proc genmod` *that we did not discuss in class. Turn in your final program that carries out all of the required analyses and its output. Note: You should be using version 7 or higher of SAS. Remember that you need to use the* `descending` *option in your* `model` *statements so that SAS fits the model for $P(Y_j = 1)$, as discussed in the class notes.*

   Let Y_j be the new tumor response (0 or 1) for the jth subject. Define the following variables:

$$\begin{aligned} t_j &= 1 \text{ if subject } j \text{ randomized to thiotepa} \\ &= 0 \text{ if subject } j \text{ randomized to placebo} \\ g_j &= 1 \text{ if subject } j \text{ is male} \\ &= 0 \text{ if subject } j \text{ is female} \end{aligned}$$

   (a) The first goal of the study investigators was to carry out the primary analysis of the study addressing the issue of whether or not thiotepa leads to better patient outcomes in regard to new tumors than no treatment (placebo). As is standard, the analysis was to be conducted disregarding any subject covariates, so based only on comparing the randomized groups. They considered the following logistic regression model for $E(Y_j)$:

$$E(Y_j) = \frac{\exp(\beta_0 + \beta_1 t_j)}{1 + \exp(\beta_0 + \beta_1 t_j)}. \tag{1}$$

   (i) Based on model (1), write down an expression for the probability of not having a new tumor at 12 months for subjects who were randomized to thiotepa.

   (ii) Based on model (1), write down an expression for the odds ratio comparing the odds of developing a new tumor under treatment with thiotepa relative to those under placebo.

   (b) Read the data into SAS and use `proc genmod` to fit model (1). Use the variable names `treat` and `gender` for the treatment and gender indicators, respectively. Do not declare `treat` to be a `class` variable, so that you are fitting model (1) directly. Include an `estimate` statement to obtain the estimated log odds ratio and odds ratio corresponding to (ii) in part (a) above (see page 457 of the class notes). Based on the results, is there evidence to suggest that treatment with thiotepa has desirable effects on recurrence of bladder cancer tumors in this population of patients? (This addresses the primary analysis question.)

Note: In the output of the `estimate` statement, `proc genmod` presents a "Standard Error" and "Confidence Limits" for the odds ratio. Both are based on large-sample approximation. But because the odds ratio is not a *linear* function of elements of $\boldsymbol{\beta}$, these quantities are obtained by further approximations based on the basic large-sample results for estimation of $\boldsymbol{\beta}$ described in the class notes. The confidence interval is *not* found taking the estimate and adding and subtracting 1.96 times the estimated standard error; rather, it uses an approximation that is thought to be more accurate in finite samples. Thus, when making inference on the odds ratio, it is best to base it on this confidence interval. If the confidence interval does not contain 1 and has lower limit greater than 1, most practitioners would interpret this as approximate evidence suggesting that the odds ratio is greater than 1 (with "confidence" equaling the confidence level of the interval); if the upper limit is less than 1, this would mean evidence that the odds ratio is less than 1.

(c) Once they had completed the primary analysis, the study investigators wished to carry out some secondary, exploratory analyses to gain insight into the role, if any, of gender in the recurrence of bladder cancer tumors and how they respond to treatment. They considered the following logistic regression model for $E(Y_j)$:

$$E(Y_j) = \frac{\exp(\beta_0 + \beta_1 t_j + \beta_2 g_j + \beta_3 t_j g_j)}{1 + \exp(\beta_0 + \beta_1 t_j + \beta_2 g_j + \beta_3 t_j g_j)}. \qquad (2)$$

(i) Under model (2), write down an expression for the odds of female subject randomized to thiotepa developing a new tumor.

(ii) Under model (2), write down an expression for the probability that a male subject randomized to placebo develops a a new tumor.

(d) We will now gain some insight into how SAS parameterizes models and into some of the output produced by `proc genmod`. Fit model (2) using `proc genmod` two ways:

(i) Declare `treat` and `gender` to be `class` variables. Does SAS parameterize the model the same way we did in (2)?

(ii) *Do not* declare `treat` and `gender` to be `class` variables. Does SAS parameterize the model the same way we did in (2)?

Calculate *by hand* the estimate of the probability that a male subject randomized to thiotepa shows a new tumor at 12 months based on the output of each of the fits (i) and (ii). Does it matter which way the model is parameterized? Explain.

(e) Re-run your code for (d)(i), but add the options `type3` and `wald` in the `model` statement. These options given together cause `proc genmod` to construct Wald test statistics for hypotheses of the form $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ for some $\boldsymbol{L}$, where $\boldsymbol{\beta}$ is the set of parameters in the model. These will appear in the table `Wald Statistics For Type 3 Analysis` in the output. We will now see what these Wald tests are testing when the linear predictor contains an *interaction* term, as in (2), by considering the result for the `treat` effect.

Consider the model as parameterized in (2). Write down the linear predictor for each of the four possible combinations of values of t_j and g_j. Then write down the *average* of the values the linear predictor takes on when $t_j = 0, g_j = 0$ and $t_j = 0, g_j = 1$. Similarly, write down the *average* of the values the linear predictor takes on when $t_j = 1, g_j = 0$ and when $t_j = 1, g_j = 1$. Each of these averages represents the value of the linear predictor for each value of `treat`, averaged across the two levels of `gender`.

Now form the *difference* of these two averages (do the algebra to simplify). This expression is the difference in the average values of the linear predictor (averaged across levels of `gender`) for each level of `treat`. I.e., we may think of this difference as the "main effect of treatment" for the linear predictor.

Finally, write this difference in the form $\boldsymbol{L\beta}$. Re-run your code for (d)(ii), but add a `contrast` statement for testing $H_0 : \boldsymbol{L\beta} = \mathbf{0}$. Compare the output from this `contrast` to the results in your fit in (d)(i). What do you conclude that the "default" Wald statistics for `treat` and `gender` are testing in fit (d)(i)?

Note: Please see the solutions to this homework assignment for discussion of the usefulness of these hypotheses and tests for data like these.

(f) Is there evidence in the data to suggest that the way in which the probability of a subject developing a new tumor depends on the treatment they received is different depending on gender? In any parameterization of the model you choose, state the null hypothesis corresponding to this question. From any of the fits so far (state which one you choose), give the value of an appropriate test statistic and p-value for addressing this hypothesis, and state the result as a meaningful sentence.

(g) Consider the simpler model with no interaction term:

$$E(Y_j) = \frac{\exp(\beta_0 + \beta_1 t_j + \beta_2 g_j)}{1 + \exp(\beta_0 + \beta_1 t_j + \beta_2 g_j)} \qquad (3)$$

Fit model (3) using `proc genmod` two ways:

(i) Declare `treat` and `gender` to be `class` variables. Also include the options `type3` and `wald` in the `model` statement. Does SAS parameterize the model the same way we did in (3)?

(ii) Use the variables `treat` and `gender` but do not declare these to be `class variables`. Also include the options `type3` and `wald` in the `model` statement. Does SAS parameterize the model the same way we did in (3)?

What do you notice about the numerical values of the estimates from these two fits?

(h) In terms of the parameterization in model (3), write down an appropriate matrix $\boldsymbol{L}$ for a null hypothesis of the form $H_0 : \boldsymbol{L\beta} = \mathbf{0}$, where $\boldsymbol{\beta}$ is the vector of parameters in the model, that addresses the following question: "Does gender of a subject make a difference in the probability of a developing a new tumor?" In the call to `proc genmod` in part (g)(ii), use a `contrast` statement to obtain the Wald test statistic for your hypothesis. Cite the value of the statistic and the associated p-value, and compare these to the values for `gender` in the table `Wald Statistics For Type 3 Analysis` in the output of the calls in (g)(i) and (ii). What do you notice?

(i) Another way to test hypotheses of the form $H_0 : \boldsymbol{L\beta} = \mathbf{0}$ is by the "full" versus "reduced" model principle; i.e. by using a likelihood ratio test. Recall that the method used to fit generalized linear models is a maximum likelihood method; thus, this principle is applicable here. In your program, call `proc genmod` again to fit the model as in part (g)(i); however, this time, include *only* the `type3` option (leaving off the `wald` option). The result of this action is to ask `proc genmod` to compute the likelihood ratio test statistic corresponding to hypotheses of the form $H_0 : \boldsymbol{L\beta} = \mathbf{0}$ for each of the factors in the `class` statement automatically.

Compare the results of the likelihood ratio test of the issue in (h) to that obtained by Wald methods in (h). Do the qualitative conclusions agree?

2. Recall the epileptic seizure data discussed in Example 4 of Chapter 1 of the class notes. The data are from a clinical trial conducted on 59 subjects suffering from seizures who were randomized to receive a placebo (subjects 1–28) or the anti-seizure agent progabide (subjects

29–59). The numbers of seizures experienced by each subject was recorded during an 8-week baseline (pre-treatment) period and then during each of four post-treatment, consecutive 2-week periods. Here, we will focus on the *cross-section* of these data corresponding to the third 2-week period. The data are on the class web page in the file `seize3.dat`, with columns (1) subject id, (2) number of seizures during fourth 2-week period, and (3) treatment indicator (0 = placebo, 1 = progabide).

The goal of an analysis is to compare the two groups (placebo and progabide) in regard to the number of seizures experienced in the third 2-week period of the study.

(a) Let $p_j = 0$ if subject j received placebo and 1 if progabide. Consider the model for mean number of seizures given as

$$E(Y_j) = \exp(\beta_0 + \beta_1 p_j). \tag{4}$$

Fit this model parameterized exactly as in (4) (so regarding the treatment variable as numeric rather than as a classification variable) using `proc genmod`, assuming that the seizure counts follow a Poisson distribution. Based on the results, is there sufficient evidence to support the contention that the mean number of seizures during the third 2-week period is associated with whether placebo or progabide is administered? Give an appropriate test statistic based on your fit and the associated p-value.

(b) Model (4) gives the form of the expected count or number of events (seizures) in this case over a 2-week period. Thus, note that the expected *rate* of seizures per week under treatment p_j is

$$E(Y_j)/2 = \exp(\beta_0 + \beta_1 p_j)/2.$$

The *rate ratio* (for progabide relative to placebo) is thus given by the ratio of $E(Y_j)/2$ for $p_j = 1$ to $E(Y_j)/2$ for $p_j = 0$; i.e.,

$$\frac{\exp(\beta_0 + \beta_1)/2}{\exp(\beta_0)/2} = \exp(\beta_1).$$

Thus, $\exp(\beta_1)$ has interpretation as the rate ratio for progabide relative to placebo (and hence β_1 has interpretation as the log rate ratio). A rate ratio less than 1 means that progabide is effective at reducing the seizure rate relative to placebo.

It is possible to use an `estimate` statement with the `exp` option exactly as was done in Problem 1(b) to get `proc genmod` to output the estimated rate ratio, an approximate standard error, and a confidence interval for the true rate ratio. Add this to your call to `proc genmod` in (a). Based on the results, does the evidence suggest that the seizure rate is lower on progabide than on placebo?

Note: The rate ratio and the odds ratio in a logistic regression model as in Problem 1 turn out to be of the same form. Thus, all of the comments regarding approximate inference on odds ratios in the note in Problem 1(b) apply similarly to approximate inference on a rate ratio here.

3. The Skin Cancer Prevention Study was a multi-center randomized clinical trial to evaluate beta-carotene as a treatment for preventing non-melanoma skin cancer in high-risk subjects. Subjects in the trial were randomized to receive beta-carotene or placebo for five years, and each subject was to return once a year to be examined for new skin cancers. The data ($n = 1683$ subjects) are in the file `skin.dat` on the class web page, with the following columns (we won't use all of these):

Column	Variable
1	subject ID
2	center
3	age (years)
4	skin type (1 = burns, 0 = otherwise)
5	gender (0 = female, 1 = male)
6	number of previous skin cancers
7	response (number of new skin cancers at this year)
8	treatment (0 = placebo, 1 = beta-carotene)
9	year (1–5)

(a) As a preliminary step, use `proc means` (or other approach of your choice) to calculate the sample mean number of skin cancers in each treatment group at each year. Does this raw sample evidence seem to suggest any patterns over time in the mean numbers of skin cancers?

(b) Let Y_{ij} denote the number of skin cancers observed for subject i at time t_{ij} (year), $j = 1, \dots, n_i$ (you will note that not all subjects have responses observed in all 5 years). Let $\delta_i = 0$ if subject i was assigned to the placebo group and let $\delta_i = 1$ if assigned to the beta-carotene group. Write down an appropriate model for mean number of new skin cancers at time t_{ij} in terms of δ_i that has the following features:

(i) The mean number of skin cancers at time 0 (randomization; this was not actually measured) is the *same* in both groups (because the study is randomized);

(ii) Estimates of the mean number of skin cancers can never be negative;

(iii) The logarithm of the mean number of skin cancers over time potentially changes at a constant rate over time in a way that that is different in the placebo and beta-carotene groups.

Write down the vector of parameters, $\boldsymbol{\beta}$, that describes your mean model.

(c) Fit your model using `proc genmod`, allowing for the possibility that observations Y_{ij}, $j = 1, \dots, n_i$, are correlated in the same way regardless of treatment group and and exhibit the same correlation regardless of how far apart in time they are. Also allow for the possibility of overdispersion. Have your program print out the estimated correlation matrix of a data vector and calculate the standard errors for the estimates of components of $\boldsymbol{\beta}$ using the model-based estimate of covariance matrix.

(d) Informally, from the results, does there seem to be evidence of overdispersion? Also informally, does it seem necessary to take correlation into account? Explain your answers.

(e) The main question of interest was whether or not there is a difference in the pattern of change of mean number of cancers over time between beta-carotene and placebo. Write down a null hypothesis that addresses this issue, and then express it in the form $\boldsymbol{L}\boldsymbol{\beta} = \mathbf{0}$, giving the form of $\boldsymbol{L}$. Have your call to `proc genmod` provide an appropriate test statistic and p-value (you may do this however you like). Give the test statistic and p-value, and state what you conclude.

(f) In light of the results in (e), the investigators decided to do some exploratory analyses to investigate associations between mean number of cancers and other factors over time, disregarding the treatment assignments altogether. They were particularly interested in skin type. Let $s_i = 1$ if a subject had burns and 0 otherwise. Write down an appropriate model for mean number of new skin cancers at time t_{ij} in terms of s_i that has the following features:

(i) The mean number of skin cancers at time 0 may be different depending on whether subjects have burns or not.

(ii) Estimates of the mean number of skin cancers can never be negative;

(iii) The logarithm of the mean number of skin cancers over time potentially changes at a constant rate over time in a way that that is different depending on skin type.

Write down the vector of parameters, $\boldsymbol{\beta}$, that describes your mean model.

(g) Fit your model using `proc genmod` under the same assumptions on correlation and overdispersion as in (c).

(h) The main question of interest was whether or not there is a difference in the pattern of change of mean number of cancers over time between skin types. Write down a null hypothesis that addresses this issue, and then express it in the form $\boldsymbol{L\beta} = \mathbf{0}$, giving the form of $\boldsymbol{L}$. Have your call to `proc genmod` provide an appropriate test statistic and p-value (you may do this however you like). Give the test statistic and p-value, and state what you conclude.

4. A clinical trial was conducted to study the effectiveness of a treatment patients with respiratory illness. $m = 111$ patients were recruited and randomized to receive either an active treatment (coded as A) or a placebo (coded as P). At baseline (week 0), prior to administration of the assigned interventions, each subject's respiratory status was classified as "poor" (coded as 0) or "good" (coded as 1). Subjects returned to the clinic at the end of each of the next 4 weeks (coded as week = 1, 2, 3, 4), at which their respiratory status was again evaluated (recorded as 0, poor, or 1, good). Also recorded for each subject was his or her age at baseline and gender.

   The main objective of the study was to determine whether subjects receiving the active treatment show signs of improvement under treatment with the active agent. The investigators were also interested in whether the extent of improvement depended on age or gender of a subject.

   The data are in the file `respstatus.dat` on the class web page. The columns are (1) subject id; (2) drug (A,P); (3) gender (0 = female, 1 = male); (4) age at baseline (years); (5) week; (6) respiratory status (0 = poor, 1 = good).

   (a) As a preliminary step, use `proc means` (or other approach of your choice) to find the sample proportion of subjects with "good" respiratory status in each group at each week (recall that the mean of a binary random variable is the probability the variable takes on the value 1). Does this raw sample evidence seem to suggest an increase in the proportion of subjects with "good" respiratory status under active treatment relative to placebo?

   (b) Let Y_{ij} be the respiratory status for patient i at time t_{ij}, where $t_{ij} = 0, 1, 2, 3, 4$ weeks for all subjects. Let $\delta_i = 0$ if patient i was assigned to placebo and 1 if assigned to active treatment. Let $w_{ij} = 0$ if $t_{ij} = 0$ and $w_{ij} = 1$ if $t_{ij} > 0$. Consider the following model for the probability that a subject has good respiratory status at time j:

$$E(Y_{ij}) = P(Y_{ij} = 1) = \frac{\exp(\beta_0 + \beta_1 w_{ij} + \beta_2 w_{ij}\delta_i)}{1 + \exp(\beta_0 + \beta_1 w_{ij} + \beta_2 w_{ij}\delta_i)}, \tag{5}$$

   Explain what model (5) assumes about the probability of good respiratory status over the course of the study in each group. In light of the sample evidence in (a), does the model make sense? Explain.

   (c) Fit model (5) using `proc genmod` under the working assumption that pairs of respiratory status observations on the same subject exhibit the same overall correlation regardless of how close or far in time they are from one another. (You will have to define a new variable for w_{ij} in your program.)

(d) Is there evidence to suggest that, after week 0, undergoing treatment results in a change in the log odds of having good respiratory status relative to that for placebo? Express this question as a null hypothesis in terms of model (5), and obtain the relevant test statistic and p-value using `proc genmod`, allowing the possibility that the working assumption on correlation in (c) may be incorrect. What do you conclude?

(e) Using an `estimate` statement, obtain an estimate of the odds ratio comparing the odds that a patient receiving active treatment will have good respiratory status after week 1 to the odds of having good respiratory status after week 1 if s/he were to take placebo instead, and obtain a confidence interval for the odds ratio. Is there evidence to suggest that the active treatment has a positive effect on respiratory status?

(f) The investigators were also interested in whether the probability of having good respiratory status post-baseline is associated with whether a patient is male or female. Defining $g_i = 0$ if subject i is female and $g_i = 1$ if i is male, they considered the following model:

$$E(Y_{ij}) = P(Y_{ij} = 1) = \frac{\exp(\beta_0 + \beta_1 w_{ij} + \beta_2 w_{ij} g_i + \beta_3 w_{ij}\delta_i + \beta_4 w_{ij}\delta_i g_i)}{1 + \exp(\beta_0 + \beta_1 w_{ij} + \beta_2 w_{ij} g_i + \beta_3 w_{ij}\delta_i + \beta_4 w_{ij}\delta_i g_i)} \qquad (6)$$

Fit this model using `proc genmod` under the same working assumption on correlation as in (d), parameterizing the model exactly as it appears above, treating the treatment indicator and gender indicator as numeric.

(g) Under model (6), write down the log odds of good respiratory status after week 0 for a female who received active treatment and for a male who received active treatment. Use a contrast statement to obtain a test of whether the log odds is different for males and females. Based on the fit, is there evidence to suggest that the log odds is different for males and females, so that gender seems to matter in the effectiveness of treatment?