MATH 426: Applied Longitudinal Analysis
Homework 7, due Thursday, December 14

Please submit a PDF or .doc version of your homework to Blackboard by 11:59pm on the due date. Please type all responses. You are encouraged to use R for all calculations. Please include a commented code appendix at the end of the assignment.

Generalized Estimating Equations

The Skin Cancer Prevention Study, a randomized, double-blind, placebo-controlled clinical trial, was designed to test the **effectiveness of beta-carotene** in the prevention of non-melanoma skin cancer in high-risk subjects. A total of 1,683 subjects were randomized to either placebo or 50 mg of beta-carotene per day and were **followed for up to 5 years**. Subjects were **examined once per year** and biopsied—if a cancer was suspected—to determine the number of new cancers per year.

The outcome variable, Y, is a count of the **number of new skin cancers per year**. Selected data from the study are in the dataset skin.txt (no header). Each row of the dataset contains the following 9 variables: ID, Center, Age, Skin, Gender, Exposure, Y, Treatment, Year. These variables take values as follows:

ID: Subject identifier number
Center: Identifier number for center of enrollment [categorical]
Age: Subject's age in years at randomization
Skin: Skin type (1=burns; 0=otherwise) [evaluated at randomization and doesn't change with time]
Gender: 1=male; 0=female
Exposure: Count of number of previous skin cancers [prior to randomization] Y: Count of number of new skin cancers in the Year of follow-up
Treatment: 1=beta-carotene; 0=placebo
Year: Year of follow-up after starting randomized treatment

You may assume that the counts of new skin cancers, Y , are from one-year periods (i.e. no offset term is needed).

1. Your collaborator is interested in assessing the effect of treatment on the incidence of new skin cancers over time. As the statistician on the project, provide an analysis of the data that addresses this question.

(a) Provide a short table providing a descriptive summary of the mean count of new skin cancers by Treatment for each Year. Briefly comment on changes in incidence of new skin cancers by randomized treatment over time.

```
group: 0
  vars   n mean   sd
1    1 827 0.27 0.87
2    2 803 0.24 0.69
3    3 776 0.25 0.78
4    4 699 0.23 0.78
5    5 419 0.27 0.85
------------------------------------------------------------
group: 1
  vars   n mean   sd
1    1 856 0.30 0.80
2    2 827 0.26 0.68
```

```
3     3 794 0.29 1.06
4     4 688 0.32 1.12
5     5 392 0.30 0.90
```

There is no clear pattern in the mean count of new skin cancers over time for the group receiving the placebo, with the direction of the change in mean count switching between each year. While there is a sustained increase in cases from year 2 through 4 for the group receiving the 50 mg of beta-carotene per day, there is similarly no clear pattern as declines occur between years 1 and 2 and years 4 and 5, respectively.

(b) Fully specify a marginal Poisson regression model with an Treatment×Year (as a continuous variable) interaction and no main effect of Treatment. **You may assume an overdispersed variance however you will need to determine the association structure**. Create a table of results that includes the regression coefficient estimates, empirical standard errors, 95% confidence intervals, and p-values for testing H0 : $\beta k = 0$, k = 0, 1, 2. What do you conclude from this model about the effect of treatment?

	Estimate	Std.err	Wald	Pr(>|W|)	95% CI
(Intercept)	-1.3171	0.0786	280.62	0.0000	(-1.47, -1.16)
Year	-0.0212	0.0304	0.48	0.4871	(-0.08, 0.04)
Treatment:Year	0.0485	0.0378	1.65	0.1995	(-0.03, 0.12)

We would conclude from the model that there is no statistically significant effect of the treatment on reducing the rate of new skin cancers per year (p=0.20).

(c) Properly interpret each of the parameter estimates from your analysis in (b) on the rate ratio scale.

The rate of new skin cancers per year is exp(0.04291) = 1.05 greater for those receiving 50 mg of beta-carotene per day compared to receiving the placebo over the same time period, though the result is statistically insignificant (the p-value of 0.20 is though close to commonly accepted standards of statistical significance). Regardless of the treatment, the rate of new skin cancers per year is exp(-0.0212)=0.979 greater for subjects over the duration of the measurement period compared to baseline. The effect of the change over time, however, is not statistically significant as shown by the p-value of 0.74.

(d) In the form of the results section of a scientific abstract (3 or 4 sentences), state with justification what you conclude about the effect of beta-carotene versus placebo based on your analysis.

The Skin Cancer Prevention Study is a randomized double-blind, placebo-controlled clinical trial that was designed to test the effectiveness of beta-carotene in the prevention of non-melanoma skin cancer in high-risk subjects. We ran a specified a marginal Poisson regression model with GLM to study the effect of receiving 50 mg of beta-carotene per day over five years with measurements taking place once a year. The rate of new skin cancers per year increased at a rate 1.05 greater for those receiving 50 mg of beta-carotene per day compared to those receiving a placebo over a five-year period. However, we did not find we a statistically significant difference in the effect of the drug compared to the placebo. We also did not find a statistically significant difference in the difference for all subjects from baseline. Given that the p-value of 0.20 approaches commonly accepted standards of statistical significance, we would

not completely rule out that doctors prescribe the beta-carotene. At the very least, we would recommend further trials.

2. A secondary aim of the study was to evaluate risk factors for increased incidence of new skin cancers. (a) Fit an appropriate marginal model which includes Year (a linear effect), Treatment, Center, Age, Skin, Gender and Exposure as covariates. Make sure you fully specify your chosen model.

```
Call:
geeglm(formula = Y ~ Year + Treatment + Center + Age + Skin +
    Gender + Exposure, family = poisson(link = "log"), data = skin,
    id = ID, scale.fix = FALSE)

 Coefficients:
            Estimate  Std.err   Wald Pr(>|W|)
(Intercept) -3.83821  0.32989 135.37  < 2e-16 ***
Year         0.01648  0.02687   0.38   0.5398
Treatment    0.10706  0.10094   1.12   0.2889
Center       0.09826  0.03517   7.80   0.0052 **
Age          0.01643  0.00506  10.54   0.0012 **
Skin         0.15041  0.10883   1.91   0.1669
Gender       0.59458  0.10150  34.31  4.7e-09 ***
Exposure     0.13715  0.01096 156.46  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)      1.6  0.0824

Correlation: Structure = independenceNumber of clusters:   1683   Maximum cluster size: 5
```

(b) Provide a table that includes parameter estimates, empirical standard errors, and p-values.

	Estimate	**Std.err**	**Wald**	**Pr(>|W|)**
(Intercept)	-3.8382	0.3299	135.37	0.0000
Year	0.0165	0.0269	0.38	0.5398
Treatment	0.1071	0.1009	1.12	0.2889
Center	0.0983	0.0352	7.80	0.0052
Age	0.0164	0.0051	10.54	0.0012
Skin	0.1504	0.1088	1.91	0.1669
Gender	0.5946	0.1015	34.31	0.0000
Exposure	0.1371	0.0110	156.46	0.0000

(c) In a short paragraph (< 150 words), summarize the key findings from this model.

Holding constant the other variables in the model, the rate of new skin cancers per year is exp(0.5946) = 1.81 greater for males than that of females over the same time period. Holding constant the other variables in the model, the rate of new skin cancers per year is exp(0.15041) = 1.16 greater for those with burns than those otherwise over the same time period. Similarly, again holding constant the other variables in the model, each increase in the count of number of previous skin cancers prior to randomization increases the rate of new skin cancers over the five year period by a multiplicative factor of exp(0.4639) = 1.15. At the same time, each increase in age increases the rate of new skin cancers over

the five year period by a multiplicative factor of exp(0.0164) = 1.02. While the aforementioned effects are all statistically significant with the exception of the variable noting skin burns (p=0.1669), there is no statistically significant effect of year on the rate of new skin cancers per year as made evident by a p-value of 0.5398.

(d) Fit the model from 2(a) assuming there is no overdispersion (that is, the overdispersion parameter is equal to 1). Briefly describe the effect of this assumption on the results of the analysis.

Overdispersion occurs when actual variance is larger than assumed by the model. The variance function is dependent on the mean with Poisson and so overdisperion could be a problem. It is therefore no surprise that the standard errors for the model with no overdispersion assumed has consistently higher standard errors. Similarly, with the model assuming no overdispersion, we lose confidence in the model and it therefore makes sense that we'd have higher p-values and less statistical significance.