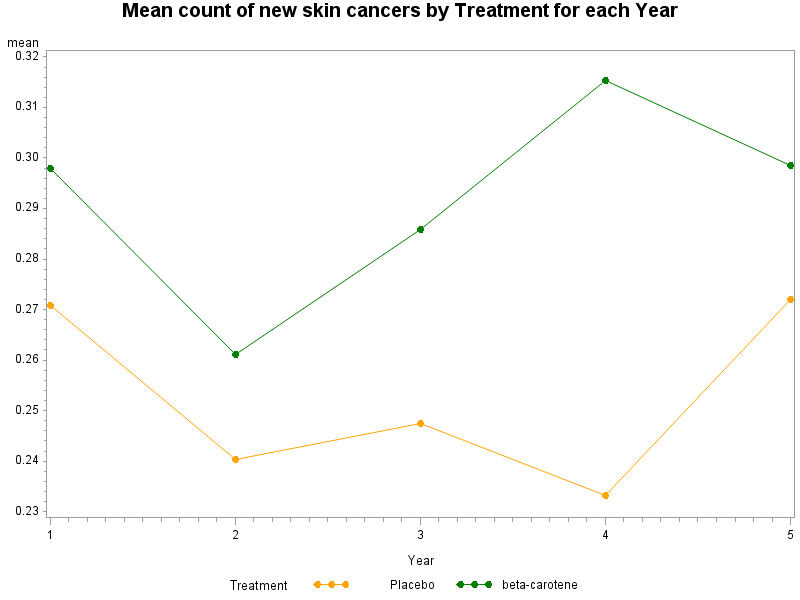
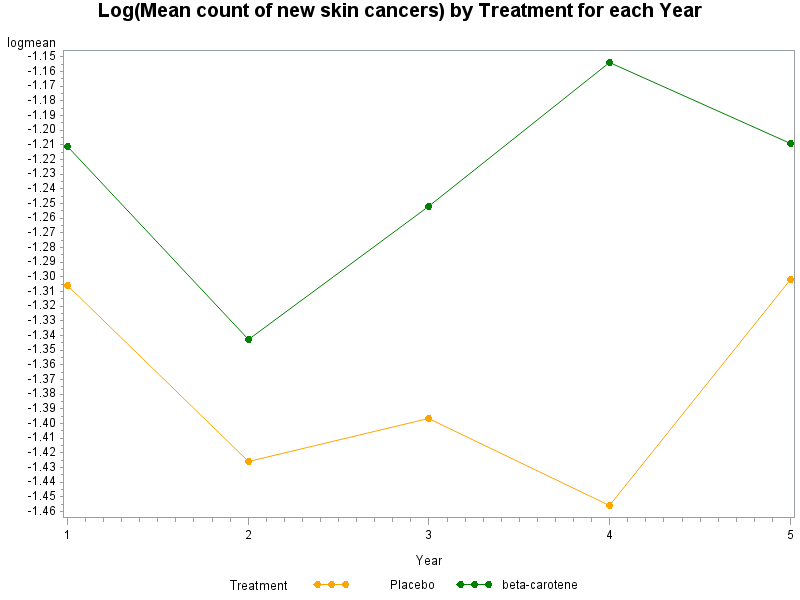
**Questions:**

**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. Please,**

1. **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.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Treatment |  | Year 1 | Year 2 | Year 3 | Year 4 | Year 5 |
| Placebo | **N** | **827** | **803** | **776** | **699** | **419** |
| Mean | 0.271 | 0.24 | 0.247 | 0.233 | 0.272 |
| **Variance** | 0.762 | 0.477 | 0.607 | 0.612 | 0.715 |
| beta-carotene | **N** | 856 | 827 | 794 | 688 | 392 |
| Mean | 0.298 | 0.261 | 0.286 | 0.315 | 0.298 |
| **Variance** | 0.647 | 0.457 | 1.117 | 1.264 | 0.803 |



During these 5 follow-up years, the count of new skin cancers, for the placebo group, drops during 2-4 years and then increases to the same level as the 1st year; for beta-carotene group, it drops at the 2nd year but then increases, at the end of the 5th year, it also reaches almost the same level as its 1st year. During the 5 years’ follow-up, the average counts of new skin placebo group are always lower than the beta-carotene group.

These data also display substantially greater variability than the mean. As a result a Poisson assumption for the variance is not appropriate for these data. Over-dispersion should be used for the variance.

Of note, there are substantial dropout, the number of observations drop about half at the end of the 5th year for both groups. It would be better to check the reasons of drop-outs (if completely random).

1. **Provide an algebraic definition for a generalized linear marginal Poisson regression model in which the only effects are for the intercept, Year (as a continuous variable) and trt\*Year. Make sure you write out the full model. Provide a table of results that you obtain from fitting the model in SAS that includes the regression coefficient estimates, empirical standard errors 95% confidence interval, and p-value for testing the hypothesis that the true parameter is zero. What do you conclude from this model about the effect of treatment?**

Note that there is no baseline counts of new skin cancers prior randomization. However, due to randomization, we assume that the baseline (Time=0) is the same for both groups and omit the main effect in the mode. We consider the following marginal model for the expected count of new skin cancers:

* Model for the mean:

(i.e., Poisson regression)

Where:

is the counts of new skin cancers for the ith subject in jth follow-up year

=1,2,3,4, 5 is the year of follow-up

if ith subject is in beta-carotene group, if ith subject is in placebo group

* Conditional variance:

(i.e., a scale parameter for overdispersion)

* Within-subject association is accounted for by assuming a common pair-wise correlation

(i.e., exchangeable or compound symmetry correlation pattern)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Analysis Of GEE Parameter Estimates | | | | | | |
| Empirical Standard Error Estimates | | | | | | |
| Parameter |  | **Estimate** | **Standard Error** | **95% Confidence Limits** | | **Pr > |Z|** |
| Intercept |  | -1.3338 | 0.0785 | -1.4876 | -1.1800 | <.0001 |
| Year |  | -0.0064 | 0.0260 | -0.0574 | 0.0446 | 0.8062 |
| Year\*Treatment | 1 | 0.0447 | 0.0333 | -0.0205 | 0.1099 | 0.1787 |
| Year\*Treatment | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |

The (empirical standard errors) estimate of trt\*year interaction, , suggests that no evidence of rejecting the null and we conclude that the treatment effect of beta-carotene is not significant in the prevention of non-melanoma skin cancer in the study population.

1. **Properly interpret each of the parameter estimates from your analysis in (ii) on the untransformed scale.**

* Estimate of intercept

represents the expected rate of new skin cancer prior randomization (Time=0). That is, the expected rate of new skin cancers at baseline for both groups is which is significant different from zero.

* Estimate of Year

For the placebo group, there is no significant reduction, the estimated rate of reduction is

which is not significant. That is, the rate decreases 0.92% per year, for the placebo group, which is not significant.

* Estimate of Year\*Treatment

represents the difference between the changes in the log expected rates per year, comparing beta-carotene group to placebo group. That is, is the ratio of rate ratio. , indicating that beta-carotene treatment doesn’t reduces the expected rates when compared to the placebo-controlled group, instead it increases the expected rate of new cancers but not statistically significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Contrast Estimate Results | | | | |
| Label | **Mean Estimate** | **Mean** | | **Pr > ChiSq** |
| **Confidence Limits** | |
| effect for treatment group | 1.0391 | 0.9751 | 1.1072 | 0.2367 |

For the beta-carotene treatment group, there is no significant reduction, instead a little increase, the estimated rate of change is:

which is not significant. That is, the expected count increases 3.9% (p=0.2367) per year for the treatment group.

* Estimate of pair-wise correlation is not very large
* Estimated scale parameter indicates substantial overdispersion

1. **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 beta-carotene treatment doesn’t effectively prevent the incidence of new skin cancers, the expected rate in beta-carotene is 1.052(p=0.2223) times the expected rate in placebo group during the study period, which is not significant. For the placebo group, there is no significant change, the rate decreases 0.92% (p=0.7378) per year. For the beta-carotene treatment group, there is also no significant change, with non-significant 4.2% (p=0.2478) increase per year.

**2. A secondary aim of the study was to evaluate risk factors for increased incidence of new skin cancers. To accomplish this goal,**

1. **Fit an appropriate generalized linear marginal model in SAS which includes Year (a linear effect), Treatment, Center, Age, Skin, Gender and Exposure as covariates. Make sure you fully write out your chosen model.**

We consider the following marginal model for the expected count of new skin cancers:

* Model for the mean:

(i.e., Poisson regression)

Where:

is the counts of new skin cancers for the ith subject in jth follow-up year

=1,2,3,4,5 is the year of follow-up

if ith subject is in beta-carotene group, if ith subject is in placebo group

is identifier number for center of enrollment for ith subject and doesn’t change with time

is the ith subject’s age in years at randomization and doesn’t change with time

is the skin type (1=burns, 0=otherwise) for ith subject at randomization and doesn’t change with time

1=Male, 0=Female for ith subject and doesn’t change with time

is the count of number of previous skin cancers prior to randomization for ith subject and doesn’t change with time

* Conditional variance:

(i.e., a scale parameter for overdispersion)

* Within-subject association is accounted for by assuming a common pair-wise correlation

(i.e., exchangeable or compound symmetry correlation pattern)

1. **Provide a table from your SAS output that includes parameter estimates, empirical standard errors and p-values.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Analysis Of GEE Parameter Estimates | | | | | | |
| Empirical Standard Error Estimates | | | | | | |
| Parameter |  | **Estimate** | **Standard Error** | **95% Confidence Limits** | | **Pr > |Z|** |
| Intercept |  | -3.7816 | 0.3327 | -4.4336 | -3.1296 | <.0001 |
| Year |  | 0.0209 | 0.0254 | -0.0289 | 0.0707 | 0.4116 |
| Treatment | beta-carotene | 0.0941 | 0.0970 | -0.0960 | 0.2842 | 0.3321 |
| Treatment | Placebo | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| center | 4 | 0.3847 | 0.1164 | 0.1565 | 0.6129 | 0.0010 |
| center | 3 | 0.6921 | 0.1525 | 0.3933 | 0.9910 | <.0001 |
| center | 2 | 0.5671 | 0.1287 | 0.3149 | 0.8193 | <.0001 |
| center | 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Age |  | 0.0149 | 0.0051 | 0.0048 | 0.0250 | 0.0039 |
| Gender | Male | 0.5242 | 0.0993 | 0.3296 | 0.7187 | <.0001 |
| Gender | Female | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Skin | burns | 0.1435 | 0.1008 | -0.0540 | 0.3410 | 0.1543 |
| Skin | Others | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Exposure |  | 0.1280 | 0.0102 | 0.1080 | 0.1480 | <.0001 |

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

There is no significant change on the incidence of new skin cancer over years, the rate ratio is 1.02=exp(0.0209), indicating that the expected rate increases 2.1% per year, which is not significant (p=0.4116). The expected rate in beta-carotene group is 1.10=exp(0.0941) times that in placebo group, which is not significant (p=0.3321). The expected rate in center 2, 3, and 4 are all higher than center 1, with rate ratio 1.76=exp(0.5671), 2.00=exp(0.6921), and 1.47=exp(0.3847), respectively. There is a significant effect of Age at randomization on the incidence, the rate ratio is 1.02=exp(0.0149), which is significant (p=0.0039). The expected rate for Male is higher than Female, with rate ratio 1.69=exp(0.5242), which is significant (p<.0001). The expected rate for patients with burns skin at randomization is higher than others, with rate ratio 1.15=exp(0.1435), but not significant (p<0.1543). There is a significant effect of exposure on the incidence, the rate ratio is 1.14=exp(0.128), which is significant (p<.0001).

1. **Fit the model from 2(i) 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.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Analysis Of GEE Parameter Estimates | | | | | | |
| Empirical Standard Error Estimates | | | | | | |
| Parameter |  | **Estimate** | **Standard Error** | **95% Confidence Limits** | | **Pr > |Z|** |
| Intercept |  | -3.7816 | 0.3327 | -4.4336 | -3.1296 | <.0001 |
| Year |  | 0.0209 | 0.0254 | -0.0289 | 0.0707 | 0.4116 |
| Treatment | beta-carotene | 0.0941 | 0.0970 | -0.0960 | 0.2842 | 0.3321 |
| Treatment | Placebo | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| center | 4 | 0.3847 | 0.1164 | 0.1565 | 0.6129 | 0.0010 |
| center | 3 | 0.6921 | 0.1525 | 0.3933 | 0.9910 | <.0001 |
| center | 2 | 0.5671 | 0.1287 | 0.3149 | 0.8193 | <.0001 |
| center | 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Age |  | 0.0149 | 0.0051 | 0.0048 | 0.0250 | 0.0039 |
| Gender | Male | 0.5242 | 0.0993 | 0.3296 | 0.7187 | <.0001 |
| Gender | Female | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Skin | burns | 0.1435 | 0.1008 | -0.0540 | 0.3410 | 0.1543 |
| Skin | Others | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Exposure |  | 0.1280 | 0.0102 | 0.1080 | 0.1480 | <.0001 |

Assuming no overdispersion, the GEE approach using empirical standard error has the exactly same parameter estimations and standard errors, as assuming overdispersion. Because the empirical standard error estimator corrects for potential overdispersion or any general misspecification of the variance.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Analysis Of GEE Parameter Estimates | | | | | | |
| Model-Based Standard Error Estimates | | | | | | |
| Parameter |  | **Estimate** | **Standard Error** | **95% Confidence Limits** | | **Pr > |Z|** |
| Intercept |  | -3.7816 | 0.2423 | -4.2565 | -3.3066 | <.0001 |
| Year |  | 0.0209 | 0.0159 | -0.0102 | 0.0520 | 0.1886 |
| Treatment | beta-carotene | 0.0941 | 0.0579 | -0.0194 | 0.2076 | 0.1043 |
| Treatment | Placebo | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| center | 4 | 0.3847 | 0.0906 | 0.2071 | 0.5623 | <.0001 |
| center | 3 | 0.6921 | 0.0956 | 0.5047 | 0.8795 | <.0001 |
| center | 2 | 0.5671 | 0.0940 | 0.3829 | 0.7513 | <.0001 |
| center | 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Age |  | 0.0149 | 0.0033 | 0.0084 | 0.0213 | <.0001 |
| Gender | Male | 0.5242 | 0.0777 | 0.3718 | 0.6765 | <.0001 |
| Gender | Female | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Skin | burns | 0.1435 | 0.0591 | 0.0277 | 0.2594 | 0.0151 |
| Skin | Others | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Exposure |  | 0.1280 | 0.0044 | 0.1193 | 0.1367 | <.0001 |
| Scale |  | 1.0000 | . | . | . | . |

Assuming no overdispersion, the GEE approach using model-based standard error has the same parameter estimations, because GEE yields consistent parameter estimation under misspecification of the within-subject association and variance. But GEE yields different and invalid standard errors, and specifically in this case, the Skin=burns turns from non-significant to significant. By ignoring the overdispersion, it underestimates the standard errors of the parameters, which leads to smaller p-values and invalid inferences.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Appendix: SAS Code

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*import dataset;

**data** skin;

infile 'C:\data\Projects\APCD High Cost\Longitudinal\skin.txt';

input id center Age Skin Gender Exposure Y Treatment Year;

**proc** **print**;**run**;

**proc** **format** ;

value treatment\_

**0**='Placebo'

**1**='beta-carotene'

;

**run**;

\*Q1.1 : Despripctive Summary;

**proc** **sort** data=skin;by treatment year;**run**;

**proc** **means** data=skin n mean var nway;

format treatment treatment\_.;

by treatment;class year;

var y;

output out=meandata(drop=\_type\_ \_freq\_) mean=mean var=var N=N;

**run**;

**proc** **transpose** data=meandata out=outmean;

by treatment;

id year;

var N mean var ;

**proc** **print** data=outmean;

**run**;

\* Plot the untransformed means by treatment gorup;

**proc** **gplot** data=meandata ;

symbol1 color=orange interpol=join value=dot;

symbol2 color=green interpol=join value=dot;

plot mean\*year=treatment ;

title 'Mean count of new skin cancers by Treatment for each Year';

**run**;

\* After log transform;

**data** meandata;

set meandata;

logmean=log(mean);

**run**;

**proc** **gplot** data=meandata;

symbol1 color=orange interpol=join value=dot;

symbol2 color=green interpol=join value=dot;

plot logmean\*year=treatment;

title 'Log(Mean count of new skin cancers) by Treatment for each Year';

**run**;

\* transpose from wide format to long format;

**data** skin;

set skin;

ct=year;

**proc** **sort**;by descending treatment;

**run**;

**proc** **genmod** data=skin order=data;

class id treatment ct;

model y=year treatment\*year/dist=poisson link=log scale=pearson;

repeated subject=id/withinsubject=ct type=exch modelSE;

estimate 'effect for treatment group' year **1** treatment\*year **1** ;

**run**;

\*Q2:evaluate risk factors for increased incidence of new skin cancers;

**proc** **format**;

value skin\_

**1**=burns

**0**=Others

;

**run**;

**proc** **format**;

value gender\_

**0**=Female

**1**=Male

;

**run**;

**proc** **format**;

value Treatment\_

**1**=Beta-carotene

**0**=Placebo

;

**run**;

**proc** **sort**;by descending treatment descending skin descending gender descending center;

**run**;

**proc** **genmod** data=skin order=data;

format skin skin\_.;format gender gender\_.;format treatment treatment\_.;

class treatment id ct center skin gender;

model y= Year treatment center age gender skin exposure/dist=poisson link=log scale=pearson ;

repeated subject=id/withinsubject=ct type=exch modelSE;

**run**;

\*no overdispersion;

**proc** **genmod** data=skin order=data;

format skin skin\_.;format gender gender\_.;format treatment treatment\_.;

class treatment id ct center skin gender;

model y= Year treatment center age gender skin exposure/dist=poisson link=log scale=**1**;

repeated subject=id/withinsubject=ct type=exch modelSE;

**run**;