Final Project: The Skin Cancer Prevention Study

Minjie Fan
Instructor: Jane-Ling Wang

June 9, 2013

Abstract

In this report, we use the GEE and GLMM models to fit the skin cancer data. In both cases, we begin with a saturated model and reduce the model stepwise by Likelihood Ratio Test. The results of the GEE and GLMM models are similar, but somewhat different. After eliminating the outlying individual #37, the GEE model selects $Year$, $Gender$ and $Exposure$ as covariates. The GLMM model selects these three covariates as well. The signs of the estimates are the same for both models. In this case, similar interpretations can be applied to these two models. Note that the terms involved with Trt are all removed from these two models. This implies that there is no significant effect of beta carotene on reducing the number of new skin cancers. Model diagnostics are also conducted for the GEE and GLMM models, implying that there is no obvious lack of fit for them. Between these two models, we prefer the GLMM model to the GEE model for two reasons: the GLMM model includes subject-specific random effects; the GEE model is problematic when the dropout is missing at random. Hence, we choose the GLMM model as our final model.

1 Introduction

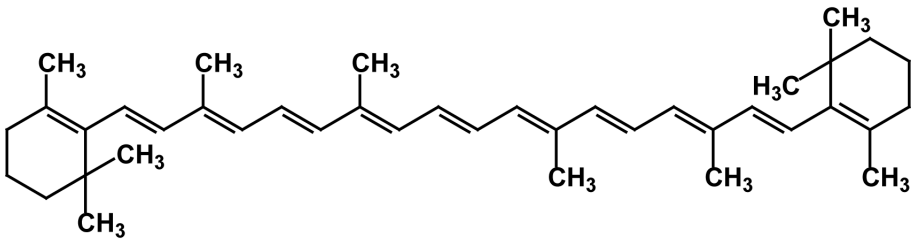

Figure 1: The molecular formula of beta carotene.

Beta carotene has been proposed as a possible preventive agent against cancer. However, current data are not sufficient to establish a direct evidence that beta carotene can decrease the risk of human cancer effectively. Greenberg et al. (1990) tested the possible cancer-preventing effects of beta carotene by randomly assigning 1805 patients who had had a

recent non-melanoma skin cancer to receive either placebo or 50mg of beta carotene per day for 5 years. Subjects were examined once a year and biopsied if a cancer was suspected to determine the number of new skin cancers occurring since the last exam. The complete data available consists of 1683 subjects with 7081 measurements in total (See Final Project requirement for detailed description of the data set). The main goal of the study is to explore the effect of beta carotene on reducing the occurrence of new skin cancers. Although the data are collected over 4 centers, we merely focus on the measurements from Center 2. This is because "center" is an external covariate, which is independent of the outcome variable and the other covariates. In spite of losing some information, the conclusion concerning the main goal will not be affected. Ultimately, the data restricted in our analysis have 315 subjects with 1248 measurements.

2 Exploratory Analysis

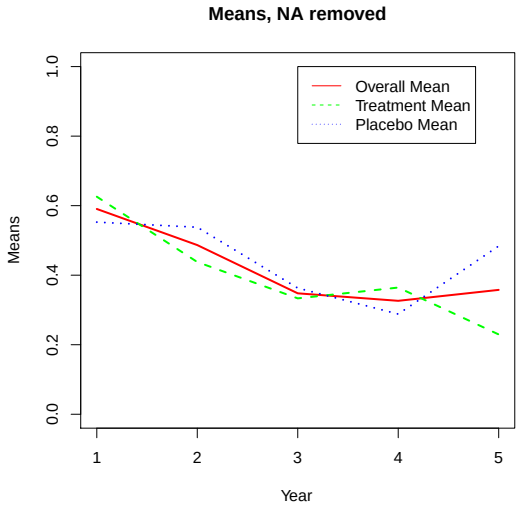

Figure 2: The group means for both treatment and placebo groups.

We randomly choose 16 patients from both treatment (i.e. beta carotene) group and placebo group, and plot their corresponding subjective trajectories (the number of new skin cancers per year versus the year of follow up), which are displayed in Fig. 3. However, it is hard to tell any difference between these two groups directly from the plot. In this case, we plot the group means for both of these two groups and the overall means as well, see Fig. 2. Note that the missing values are removed while computing the means. The three curves in the plot are entangled with each other, which might imply that there is no significant difference between the two groups. Before making any firm conclusion, however, we had better take a close look at the data by constructing appropriate statistical models. Considering the response Y is count data (discrete), we construct the models from two respects:

marginal model, i.e. generalized estimating equation (GEE) and generalized linear mixed model (GLMM).

3 Marginal Model: GEE

3.1 Methods

We begin with the saturated marginal model:

$$\begin{aligned} log(E(Y_{ij})) = \beta_1 + \beta_2 Year_{ij} + \beta_3 Trt_i + \beta_4 Trt_i \times Year_{ij} + \beta_5 Age_i + \beta_6 Skin_i \\ + \beta_7 Gender_i + \beta_8 Exposure_i + \beta_9 Trt_i \times Age_i + \beta_{10} Trt_i \times Skin_i \\ + \beta_{11} Trt_i \times Gender_i + \beta_{12} Trt_i \times Exposure_i + \beta_{13} Year_{ij}^2, \end{aligned} \tag{1}$$

where Y_{ij} is the number of new skin cancers for the i^{th} subject in the j^{th} year of follow-up, and $Year_{ij}$ denotes the year of follow up, with $Year_{ij} = j$. The variable Trt is an indicator variable for treatment group, with $Trt = 0$ if an individual was randomized to the placebo group and $Trt = 1$ if randomized to the beta carotene group. The underlying meanings of the remaining variables are just the same as in the data description. Note that the correlation structure is assumed to be unstructured temporarily. In the abovementioned saturated model, we include all the first-order terms, the interaction terms between Trt and other variables, and the quadratic term of $Year$. An omnibus test of the model coefficients of all the second-order terms except $Trt_i \times Year_{ij}$, with null hypothesis $H_0 : \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = 0$, indicates they can be eliminated from the model simultaneously (with p-value 0.5). Hence we arrive at the intermediate model:

$$\begin{aligned} log(E(Y_{ij})) = \beta_1 + \beta_2 Year_{ij} + \beta_3 Trt_i + \beta_4 Trt_i \times Year_{ij} + \beta_5 Age_i + \beta_6 Skin_i \\ + \beta_7 Gender_i + \beta_8 Exposure_i. \end{aligned} \tag{2}$$

From the estimates of its parameters, β_3, β_4 and β_5 are all insignificant. An omnibus test of β_3, β_4 and β_5, with null hypothesis $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ is conducted. The p-value is 0.51, which supports the removal of them. Thus here is our ultimate (marginal) model:

$$log(E(Y_{ij})) = \beta_1 + \beta_2 Year_{ij} + \beta_3 Skin_i + \beta_4 Gender_i + \beta_5 Exposure_i. \tag{3}$$

Note that we have supposed that the correlation pattern is unstructured. However, it is necessary to choose the optimal correlation structure among four candidates: independent, exchangeable, AR1 and unstructured. We apply the method proposed in Hardin et al. (2003). Specifically, choose a correlation structure whose sandwich estimates of the variance (of the coefficient estimates $\hat{\boldsymbol{\beta}}$) most closely approximate the naive ones. Denote the sandwich estimates as $\hat{\boldsymbol{V}}_{\text{sandwich}}$, and the naive ones as $\hat{\boldsymbol{V}}_{\text{naive}}$. The statistic used to measure the difference between these two matrices is $||\hat{\boldsymbol{V}}_{\text{sandwich}} - \hat{\boldsymbol{V}}_{\text{naive}}||_1$, where $||\cdot||_1$ denotes the entrywise

L-1 matrix norm. Typically, if the correlation structure is correctly specified, $\hat{\boldsymbol{V}}_{\text{sandwich}}$ will closely resemble $\hat{\boldsymbol{V}}_{\text{naive}}$. Hence we choose the optimal correlation structure with the smallest difference. Table 1 implies that exchangeable is the optimal correlation pattern.

Table 1: The values of the statistic $||\hat{\boldsymbol{V}}_{\text{sandwich}} - \hat{\boldsymbol{V}}_{\text{naive}}||_1$ for four different correlation structures.

Unstructured	Independent	Exchangeable	AR1
0.0244	0.0776	0.0242	0.0277

3.2 Results

Our final (marginal) model is:

$$log(E(Y_{ij})) = \beta_1 + \beta_2 Year_{ij} + \beta_3 Skin_i + \beta_4 Gender_i + \beta_5 Exposure_i, \tag{4}$$

where the correlation structure is

$$Corr(Y_{ij}, Y_{ik}) = \rho, \text{for} j \neq k,$$

and the variance of Y_{ij} is

$$Var(Y_{ij}) = \Phi E(Y_{ij}).$$

Table 2: Parameter estimates and standard errors from model (4).

Variable	Estimate	SE	P-value
Intercept	-1.49632	0.23090	9.1e-11
Year	-0.15710	0.04627	0.00069
Skin	0.28963	0.14438	0.04485
Gender	0.52267	0.20132	0.00943
Exposure	0.08373	0.00894	<2e-16
Φ	1.6	0.113	NA
ρ	0.161	0.0341	NA

Table 2 gives the estimated regression coefficients, dispersion parameter and correlation parameter. All of them are significant. Since the terms involved with Trt are eliminated from the model, the effect of beta carotene is insignificant on reducing the occurrence of new skin cancers. The expected number of new skin cancers decreases with the rate ratio of $e^{\hat{\beta}_2} = e^{-0.15710} \approx 0.855$ between two consecutive years. This might mean that human beings have a kind of natural resistance to the skin cancer. The log expected number of new skin cancers for the subject with burns skin type is larger than that for the subject without burns skin type by 0.28963. This is reasonable because typically burns skin type is at greater risk

for skin cancer. If the number of previous skin cancers is increased by 1, the log expected number of new skin cancers is also increased by 0.08373, which implies that the subject who has worse condition tends to have more new skin cancers. More interestingly, there is a distinct difference between female and male subjects. $e^{\hat{\beta}_4} = e^{0.52267} \approx 1.69$ is the rate ratio of the expected number of new skin cancers, comparing the male subject to the female one. It is more likely for males to develop skin cancers. Besides, The fact that Φ is larger 1 indicates that there exists overdispersion.

3.3 Model Diagnostic

We utilize the studentized Pearson residuals e_{ij} to do model diagnostic. In general, the studentized Pearson residuals e_{ij} is defined as $e_{ij} = \frac{Y_{ij} - g^{-1}(X_{ij}^T \hat{\beta})}{\sqrt{\phi v(\hat{\mu}_{ij})}}$. Fig. 4 indicates there is no obvious lack of fit for the model. Besides, the residual analysis can be used to detect outlying individuals. Specifically, for each individual, we calculate the mahalanobis distance between his/her observed and fitted response vector:

$$d_i = r_i^T \hat{V}^{-1} r_i,$$

where r_i is the response residual. Fig. 5 indicates that there might be an obvious outlying individual, namely the red point (individual #37), since the magnitude of its corresponding d_i is much larger than all the others. We remove this individual from the data and refit the same model again. The results are given in table 3.

Table 3: Parameter estimates and standard errors from model (4) without the individual #37.

Variable	Estimate	SE	P-value
Intercept	-1.46614	0.22983	1.8e-10
Year	-0.16287	0.04741	0.00059
Skin	0.24963	0.14163	0.07798
Gender	0.48749	0.19886	0.01423
Exposure	0.08661	0.00853	<2e-16
Φ	1.55	0.109	NA
ρ	0.135	0.0255	NA

The results with and without individual #37 are somewhat different. For example, the variable *Skin* is insignificant while eliminating the individual. Besides, Fig. 6 indicates that the trajectory of the individual deviates away from the group means for the placebo group, to which the individual belongs. Hence it is preferable to remove the individual from the data before fitting the model.

Table 4: Parameter estimates and standard errors from model (4) (excluding the variable *Skin*) without the individual #37.

Variable	Estimate	SE	P-value
Intercept	-1.30565	0.21300	8.8e-10
Year	-0.16323	0.04737	0.00057
Gender	0.46142	0.20009	0.02111
Exposure	0.09022	0.00844	<2e-16
Φ	1.55	0.109	NA
ρ	0.135	0.0258	NA

Table 4 gives the results after excluding the variable *Skin*. It can be interpreted similar as table 2 since the signs of the estimates are the same.

4 GLMM Model

4.1 Methods

We begin with the saturated model,

$$\begin{aligned} log(E(Y_{ij}|b_i)) = \beta_1 + \beta_2 Year_{ij} + \beta_3 Trt_i + \beta_4 Trt_i \times Year_{ij} + \beta_5 Age_i + \beta_6 Skin_i \\ + \beta_7 Gender_i + \beta_8 Exposure_i + b_{1i} + b_{2i} Year_{ij}. \end{aligned} \tag{5}$$

Given b_i, it is assumed that the Y_{ij} are independent and have a Poisson distribution, with $Var(Y_{ij}|b_i) = E(Y_{ij}|b_i)$. Besides, we assume that the random intercepts and slopes, b_i, have a bivariate normal distribution, with zero mean and 2×2 covariance matrix G. We fit the saturated model first, and figure out that Trt, $Trt \times Year$, Age and $Skin$ are insignificant. An omnibus test, with null hypothesis $H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$, is conducted. The resulting p-value is 0.21 indicating we fail to reject H_0. Hence we obtain an intermediate model,

$$log(E(Y_{ij}|b_i)) = \beta_1 + \beta_2 Year_{ij} + \beta_3 Gender_i + \beta_4 Exposure_i + b_{1i} + b_{2i} Year_{ij}. \tag{6}$$

The fixed effects are all significant in the intermediate model. The next step is to choose random effects. We compare the intermediate model with the following model,

$$log(E(Y_{ij}|b_i)) = \beta_1 + \beta_2 Year_{ij} + \beta_3 Gender_i + \beta_4 Exposure_i + b_i. \tag{7}$$

By Likelihood Ratio Test, which are usually more conservative since the null value is on the boundary of the parameter space, we reject the null hypothesis that there is merely a random intercept, with p-value $3.2e-07$. Thus, model (6) is our final model.

4.2 Results

Our final model is,

$$log(E(Y_{ij}|b_i)) = \beta_1 + \beta_2 Year_{ij} + \beta_3 Gender_i + \beta_4 Exposure_i + b_{1i} + b_{2i} Year_{ij}. \quad (8)$$

Table 5: Parameter estimates and standard errors from model (8)

Variable	Estimate	SE	P-value
Intercept	-1.8955	0.2196	<2e-16
Year	-0.2180	0.0493	1e-05
Gender	0.5714	0.1977	0.0039
Exposure	0.1180	0.0141	<2e-16
$g_{11} = Var(b_{1i})$	1.517	NA	NA
$g_{22} = Var(b_{2i})$	0.161	NA	NA
$g_{12} = Cov(b_{1i}, b_{2i})$	-0.342	NA	NA

Table 5 summarizes the parameter estimates and standard errors from model (8). $e^{\hat{\beta}_2} \approx 0.8041254$ is the rate ratio of the number of new skin cancers, comparing next year to current year, for a typical subject with unobserved random slope $b_{2i} = 0$. Compared with a female subject, a male subject has a larger log expected rate (by 0.5714), when the two subjects are chosen so that they have the same value for the unobserved b_{1i} and b_{2i}. Likewise, $\hat{\beta}_4$ represents the difference between the log expected rates, comparing two subjects with the former having one more previous skin cancer, when the two subjects have the same value for b_{1i} and b_{2i}. The interpretation of the fixed effects is similar to that in subsection 3.2. The change in the log expected rate is $log(E(Y_{ij+1}|b_i)) - log(E(Y_{ij}|b_i)) = \beta_2 + b_{2i}$. There is discernible heterogeneity in the subject-to-subject changes in the log expected number of new skin cancers. Specifically, approximately 95% of subjects have changes in the log expected number of new skin cancers that vary from $-0.2180 \pm 1.96\sqrt{0.161}$. b_{1i} implies that there is substantial subject-to-subject variability in terms of their expected number of new skin cancers occurring in the first follow-up year.

4.3 Model Diagnostic

We utilize Pearson's χ^2 statistic to test whether there exists overdispersion in the model (this is a really crude testing). Pearson's χ^2 is,

$$\sum_{i,j} r_{ijP}^2 = \sum_{i,j} \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}},$$

where Y_{ij} are the observations, and $\hat{\mu}_{ij}$ are the fitted values of Y_{ij}. Assume the null hypothesis H_0 is that the model fits the data. Under H_0, Pearson's χ^2 is approximately χ^2_{n-p},

where n is the number of observations, and p is the total number of parameters (= #fixed effects+#parameters contained in G). The corresponding p-value is almost 1. Hence we fail to reject the null hypothesis, which implies that the model fits the data, and thus there is no overdispersion in the model.

Fig. 7 indicates that there is no obvious outlier. This is somewhat inconsistent with the conclusion in subsection 3.3. Our explanation is that the inclusion of random effects attenuates the influence of outliers such that they are not obvious from the histograms.

5 Comparing the Marginal Model with the GLMM Model

We plot the personal fit from the marginal model and the GLMM model, for 9 randomly chosen individuals (see Fig. 8-9). It is clear that the results of the GLMM model is more favorable than those of the marginal model. Take the fit of the first individual as an example. The fitted curve from the GLMM model is much closer to the observation points, compared with that from the marginal model. Similar arguments can be applied to other individuals, such as the fifth one. There are two possible reasons to explain the abovementioned phenomenon. First, the GLMM model includes the subject-specific random effects, while the marginal model doesn't. Hence the fitted curve from the marginal model just represents the mean curve of the population to which the individual belongs. The second reason is concerned with the dropout mechanism. If the dropout is missing at random (MAR), the validity of the standard GEE approach is questionable. However, the likelihood-based methods, including the GLMM model, still work. Thus, we prefer the GLMM model to the marginal model.

6 Conclusion

In this report, we have used two models to fit the data, namely the GEE and GLMM. These two models give two similar, but different results. After eliminating the outlying individual #37, the GEE model selects $Year$, *Gender* and *Exposure* as covariates. Interestingly, the GLMM model selects the same set of covariates as the GEE model. However, the estimates of the parameters are not the same. Merely the signs of the estimates are the same. In this case, the interpretation of the estimates are similar for these two models. The terms involved with Trt are removed from both of them, which implies that beta carotene doesn't have significant effect on reducing the number of new skin cancers. Between these two models, we prefer the GLMM model to the GEE model for two reasons: the GLMM model includes the subject-specific random effects; the GEE approach is not valid in the case of MAR dropout. By Pearson's χ^2 goodness-of-fit test, we have shown that there is no overdispersion in the GLMM model.

References

[1] Garrett M. Fitzmaurice, Nan M. Laird and James H. Ware. 2011. *Applied Longitudinal Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey.

[2] Greenberg, E.R., Baron, J.A., Stukel, T.A., Stevens, M.M., Mandel, J.S., Spencer, S.K., Elias, P.M., Lowe, N., Nierenberg, D.W., Bayrd, G., Vance, J.C., Freeman, D.H., Clendenning, W.E., Kwan, T. and the Skin Cancer Prevention Study Group (1990). A clinical trial of beta carotene to prevent basal-cell and squamous-cell cancers of the skin. *New England Journal of Medicine*, **323**, 789-795.

[3] Hardin, James W. and Joseph M. Hilbe. 2003. *Generalized Estimating Equations*, Chapman & Hall/CRC Press: Boca Raton, FL.

7 Appendix

7.1 Appendix A

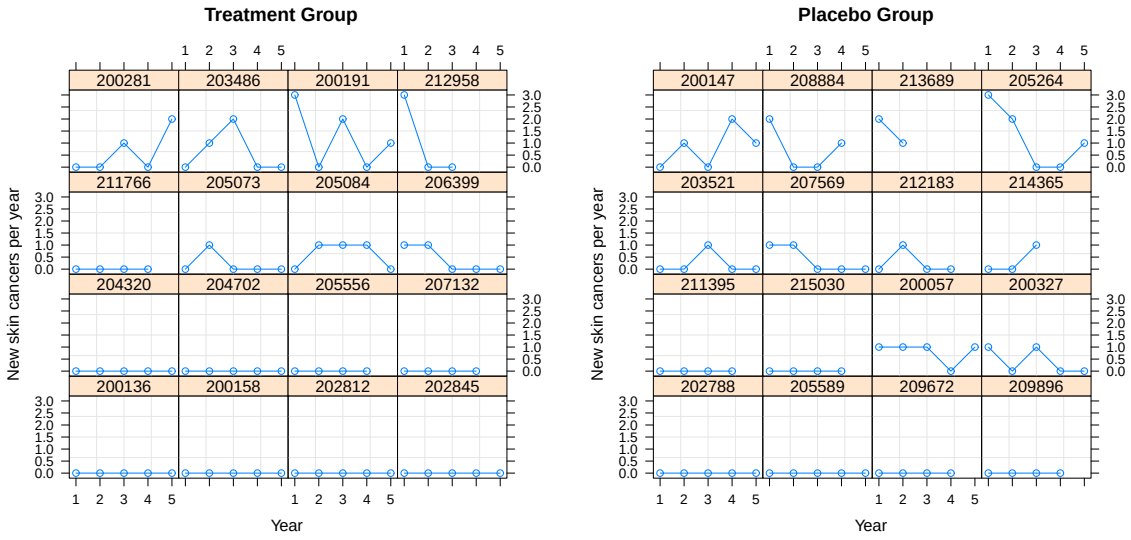

Figure 3: Overview plots for randomly chosen 16 subjects from both treatment and placebo groups.

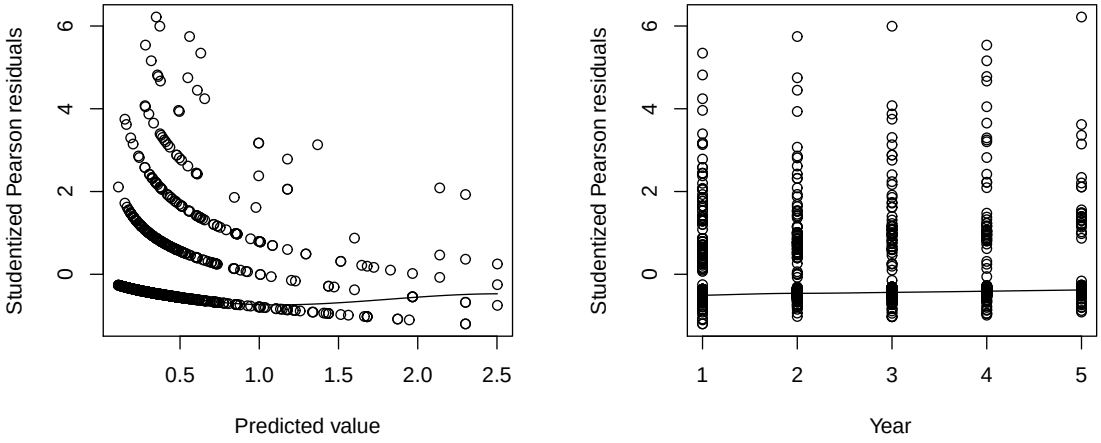

Figure 4: Residual plots (a) the studentized Pearson residuals versus predicted values (b) the studentized Pearson residuals versus years.

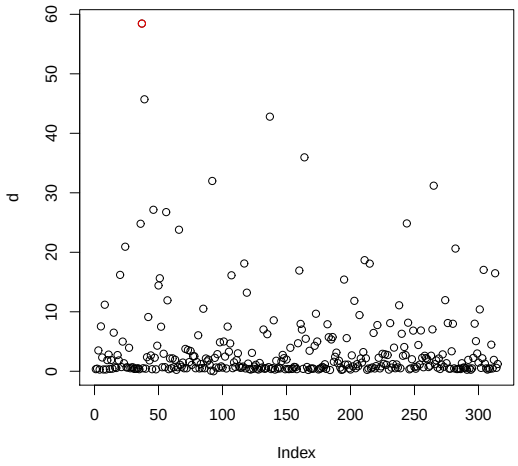

Figure 5: The plot of the mahalanobis distance d_i versus the index i.

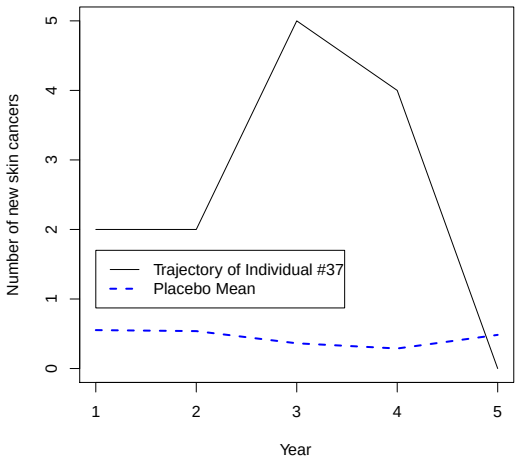

Figure 6: The plot of the trajectory of individual #37 and the group means for the placebo group.

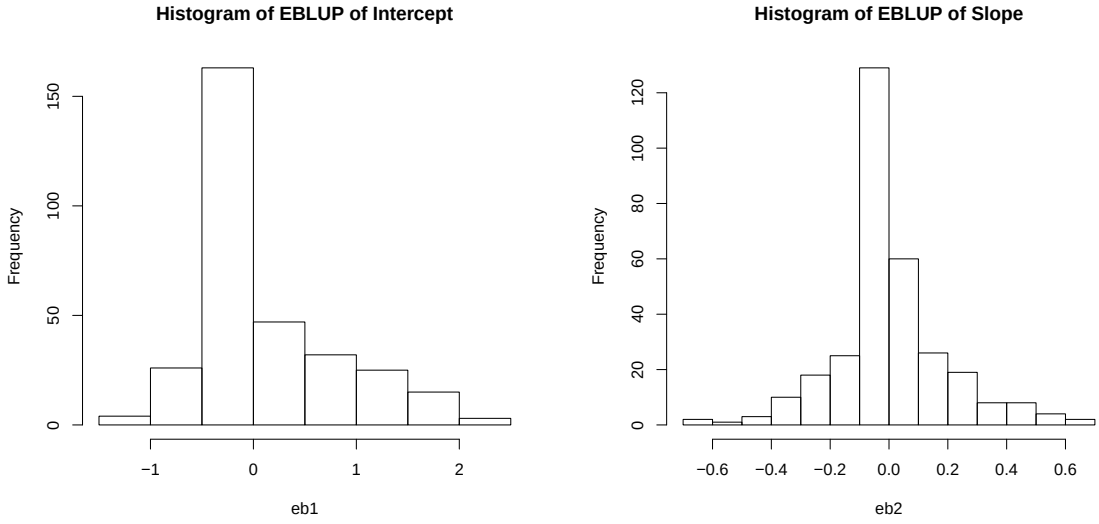

Figure 7: Histograms of empirical BLUP for (a) random intercepts and (b) random slopes.

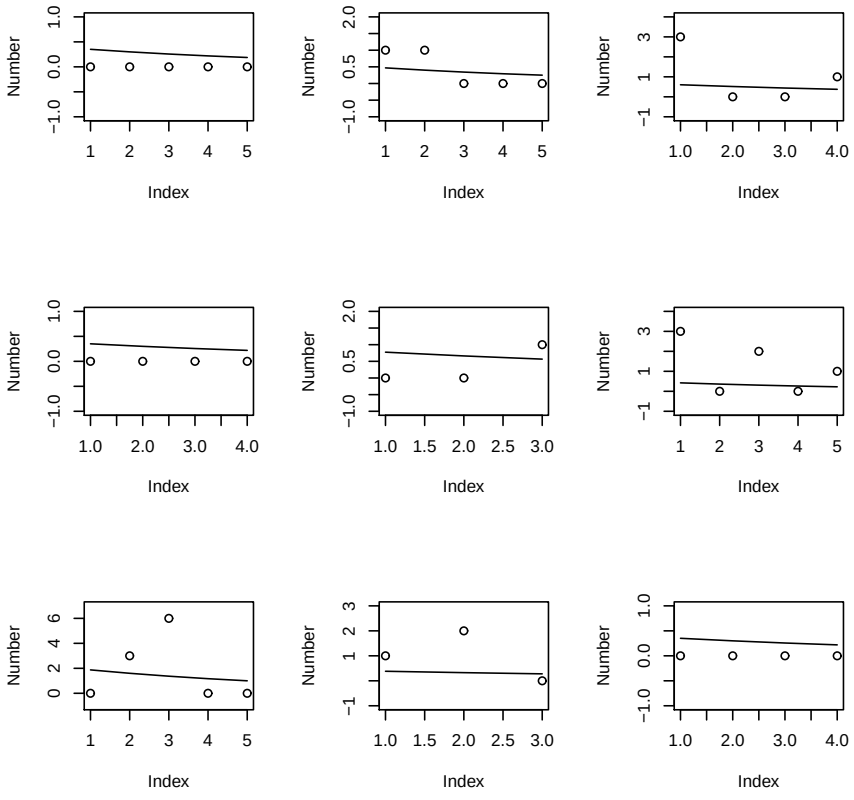

Figure 8: Subject-wise fit, using marginal model, for 9 individuals randomly chosen from the data.

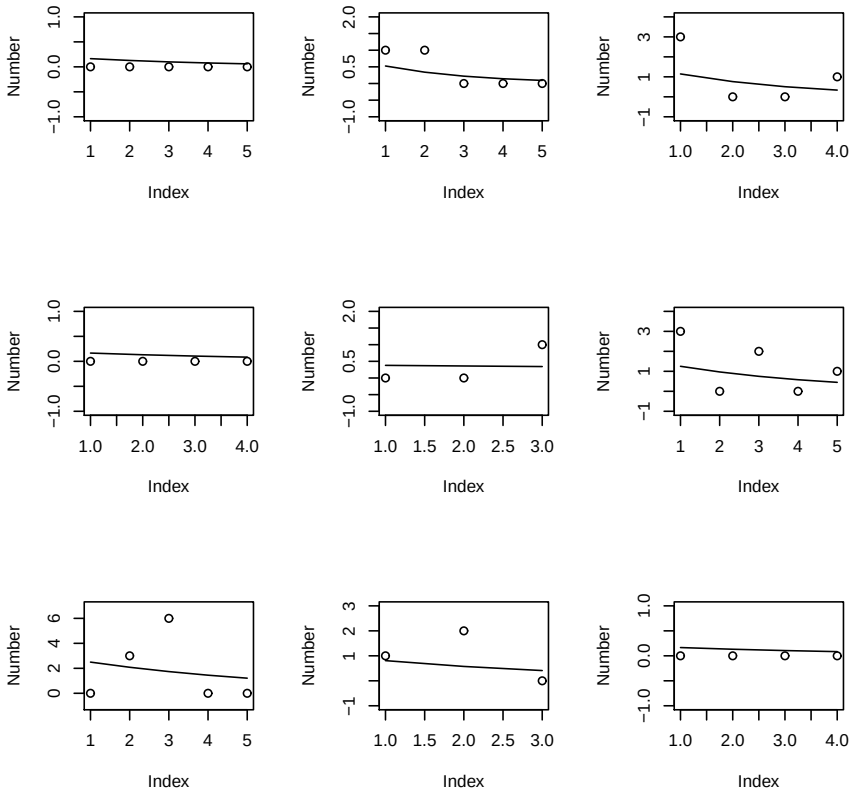

Figure 9: Subject-wise fit, using GLMM model, for 9 individuals randomly chosen from the data.

7.2 Appendix B: R Code

```
rm(list=ls())
library(lme4)
library(MASS)
library(lattice)
library(geepack)
dt<-read.table("d://skin.txt",header=FALSE)
names(dt)<-c("ID","Center","Age","Skin","Gender","Exposure","Y","Treatment","Year")
dt<-dt[dt$Center==2,]

#Exploratory Analysis
treat<-dt[dt$Treatment==1,]
placebo<-dt[dt$Treatment==0,]
treat.sample.id<-sort(sample(treat$ID,16))
placebo.sample.id<-sort(sample(placebo$ID,16))
treat.sample<-treat[treat$ID %in% treat.sample.id,]
placebo.sample<-placebo[placebo$ID %in% placebo.sample.id,]
treat.sample.gD<-groupedData(Y~Year|ID,data=treat.sample)
plot(treat.sample.gD,ylab="New skin cancers per year",main="Treatment Group")
placebo.sample.gD<-groupedData(Y~Year|ID,data=placebo.sample)
plot(placebo.sample.gD,ylab="New skin cancers per year",main="Placebo Group")

#Transform to wide format
dt.wide<-reshape(dt,v.names="Y",idvar="ID",timevar="Year",direction="wide")
Y.all<-dt.wide[,c("Y.1","Y.2","Y.3","Y.4","Y.5")]
Y.treat<-dt.wide[dt.wide$Treatment==1,c("Y.1","Y.2","Y.3","Y.4","Y.5")]
Y.placebo<-dt.wide[dt.wide$Treatment==0,c("Y.1","Y.2","Y.3","Y.4","Y.5")]
mean.all<-apply(Y.all,2,mean,na.rm=TRUE)
mean.treat<-apply(Y.treat,2,mean,na.rm=TRUE)
mean.placebo<-apply(Y.placebo,2,mean,na.rm=TRUE)
plot(mean.all,type="l",ylim=c(0,1),col="red",lwd=2,lty=1,xlab="Year",ylab="Means",
main="Means, NA removed")
lines(mean.treat,col="green",lwd=2,lty=2)
lines(mean.placebo,col="blue",lwd=2,lty=3)
legend(x=3,y=1,legend=c("Overall Mean","Treatment Mean","Placebo Mean"),col=c("red","green","blue"),
lty=c(1,2,3))

#Analysis
#Marginal model
fit.mar.sat<-geeglm(Y~Year+Treatment+I(Treatment*Year)+Age+Skin+Gender+Exposure+
I(Treatment*Age)+I(Treatment*Skin)+I(Treatment*Gender)+I(Treatment*Exposure)+I(Year^2),
id=ID,data=dt,family=poisson(link="log"),corstr="unstructured")

fit.mar<-geeglm(Y~Year+Treatment+I(Treatment*Year)+Age+Skin+Gender+Exposure,
id=ID,data=dt,family=poisson(link="log"),corstr="unstructured")

anova(fit.mar.sat,fit.mar)

fit.mar2<-geeglm(Y~Year+Skin+Gender+Exposure,
id=ID,data=dt,family=poisson(link="log"),corstr="unstructured")

anova(fit.mar,fit.mar2)
#choose fit.mar2

#compare different correlation structures
fit.mar3<-geeglm(Y~Year+Skin+Gender+Exposure,
id=ID,data=dt,family=poisson(link="log"))

fit.mar4<-geeglm(Y~Year+Skin+Gender+Exposure,
id=ID,data=dt,family=poisson(link="log"),corstr="exchangeable")

fit.mar5<-geeglm(Y~Year+Skin+Gender+Exposure,
id=ID,data=dt,family=poisson(link="log"),corstr="ar1")

var.crit <- function(mymodel) sum(abs(mymodel[["geese"]]$vbeta.naiv-mymodel[["geese"]]$vbeta))

sapply(list(fit.mar2,fit.mar3,fit.mar4,fit.mar5),var.crit)
#choose fit.mar4

#Saturated model
fit.mix<-glmer(Y~Year+Treatment+I(Treatment*Year)+Age+Skin+Gender+Exposure+(Year|ID),
data=dt,family=poisson(link="log"))
fit.mix2<-glmer(Y~Year+Gender+Exposure+(Year|ID),
data=dt,family=poisson(link="log"))
anova(fit.mix,fit.mix2)
#choose fit.mix2
```

```
76
77  fit.mix3<-glmer(Y~Year+Gender+Exposure+(1|ID),
78  data=dt,family=poisson(link="log"))
79  anova(fit.mix3,fit.mix2)
80  #include the random slope
81  #conservative
82
83  overdisp_fun <- function(model) {
84    ## number of variance parameters in
85    ##   an n-by-n variance-covariance matrix
86    vpars <- function(m) {
87      nrow(m)*(nrow(m)+1)/2
88    }
89    model.df <- sum(sapply(VarCorr(model),vpars))+length(fixef(model))
90    rdf <- nrow(model.frame(model))-model.df
91    rp <- residuals(model)
92    Pearson.chisq <- sum(rp^2)
93    prat <- Pearson.chisq/rdf
94    pval <- pchisq(Pearson.chisq, df=rdf, lower.tail=FALSE)
95    c(chisq=Pearson.chisq,ratio=prat,rdf=rdf,p=pval)
96  }
97
98  #Overdispersion
99  #dt$obs<-1:nrow(dt)
100 #fit.over<-glmer(Y~Year+Treatment+I(Treatment*Year)+Gender+Exposure+(Year|ID)+(1|obs),
101 #data=dt,family=poisson(link="log"))
102 #anova(fit.over,fit.mix2)
103 #prefer model with overdispersion
104 obj.over<-overdisp_fun(fit.mix2)
105 #no overdispersion
106
107 #Diagnostic w.r.t. marginal model
108  fit.value<-fitted(fit.mar4)
109 Y.obs<-dt$Y
110 res.raw<-Y.obs-fit.value
111 res.pear<-res.raw/sqrt(fit.value*1.6)
112 scatter.smooth(fit.value,res.pear,xlab="Predicted value",
113 ylab="Studentized Pearson residuals")
114 scatter.smooth(dt$Year,res.pear,xlab="Year",
115 ylab="Studentized Pearson residuals")
116 n<-apply((1-is.na(Y.all)),1,sum)
117 m<-length(n)
118 d<-rep(0,0)
119 p.value<-rep(0,0)
120 for (i in 1:m){
121   en<-sum(n[1:i])
122   st<-en-n[i]+1
123   C<-matrix(0.161,n[i],n[i])
124   for (j in 1:n[i]){
125     C[j,j]<-1
126   }
127   if (st!=en){
128     A<-diag(sqrt(1.6*fit.value[st:en]))
129   }
130   else{
131     A<-sqrt(1.6*fit.value[st:en])
132   }
133   V<-A%*%C%*%A
134   d[i]<-t(as.matrix(res.raw[st:en]))%*%solve(V)%*%as.matrix(res.raw[st:en])
135   p.value[i]<-pchisq(d[i],n[i],lower.tail=FALSE)
136 }
137
138 plot(d)
139 points(order(p.value)[1],max(d),col="red")
140 index.out<-order(p.value)[1]
141 en.out<-sum(n[1:index.out])
142 st.out<-en.out-n[index.out]+1
143 dt.eli<-dt[-(st.out:en.out),]
144 fit.mar6<-geeglm(Y~Year+Skin+Gender+Exposure,
145 id=ID,data=dt.eli,family=poisson(link="log"),corstr="exchangeable")
146 plot(1:5,Y.all[37,],type="l",lty=1,xlab="Year",ylab="Number of new skin cancers")
147 lines(mean.placebo,col="blue",lwd=2,lty=2)
148 legend(x=1,y=1.7,legend=c("Trajectory of Individual #37","Placebo Mean"),col=c("black","blue"),
149 lty=c(1,2),lwd=c(1,2))
150 fit.mar7<-geeglm(Y~Year+Gender+Exposure,
151 id=ID,data=dt.eli,family=poisson(link="log"),corstr="exchangeable")
152
```

```
#Diagnostic w.r.t. GLMM
EBLUP<-ranef(fit.mix2)$ID
eb1<-EBLUP[,1]
eb2<-EBLUP[,2]
hist(eb1,main="Histogram of EBLUP of Intercept")
hist(eb2,main="Histogram of EBLUP of Slope")

#Comparison between GEE and GLMM

index<-sample(1:m,9)
fit.value.rand<-fitted(fit.mix2)
par(mfrow=c(3,3))
for (i in 1:9){
  en<-sum(n[1:index[i]])
  st<-en-n[index[i]]+1
  y.min<-min(Y.all[index[i],(1:n[index[i]])])-1
  y.max<-max(Y.all[index[i],(1:n[index[i]])])+1
  plot(fit.value.rand[st:en],type="l",ylim=c(y.min,y.max),xlab="Index",
ylab="Number")
  points(1:n[index[i]],Y.all[index[i],(1:n[index[i]])],type="p")
}

dev.off()

fit.value.mar<-fitted(fit.mar4)
par(mfrow=c(3,3))
for (i in 1:9){
  en<-sum(n[1:index[i]])
  st<-en-n[index[i]]+1
  y.min<-min(Y.all[index[i],(1:n[index[i]])])-1
  y.max<-max(Y.all[index[i],(1:n[index[i]])])+1
  plot(fit.value.mar[st:en],type="l",ylim=c(y.min,y.max),xlab="Index",
ylab="Number")
  points(1:n[index[i]],Y.all[index[i],(1:n[index[i]])],type="p")
}

dev.off()
```