Stats112 HW4

Viraj Vijaywargiya

2023-05-25

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nlme)
```

```
##
## Attaching package: 'nlme'
##
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
library(ggplot2)
library(mgcv)
```

```
## This is mgcv 1.8-40. For overview type 'help("mgcv-package")'.
```

```
library(readr)
library(geepack)
library(lme4)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
##
## Attaching package: 'lme4'
```

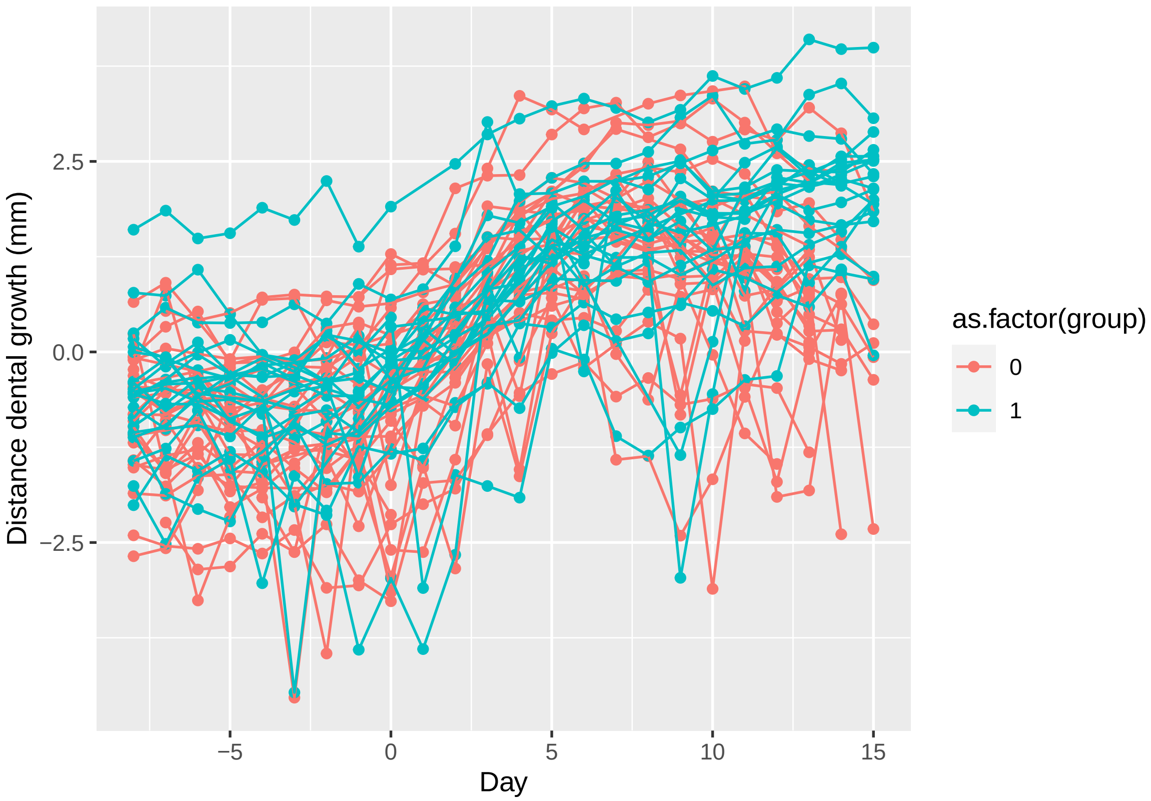
```
##
## The following object is masked from 'package:nlme':
##
##     lmList
```

1.

```
prog = read.csv("/Users/virajvijaywargiya/Downloads/progesterone.csv", header = TRUE)
```

1a)

```
prog %>%
  group_by(group) %>%
  ggplot(aes(time, PDG, group = id, color = as.factor(group))) +
  geom_point() +
  geom_line() +
  labs(x = "Day",
       y = "Distance dental growth (mm)")
```



1b)

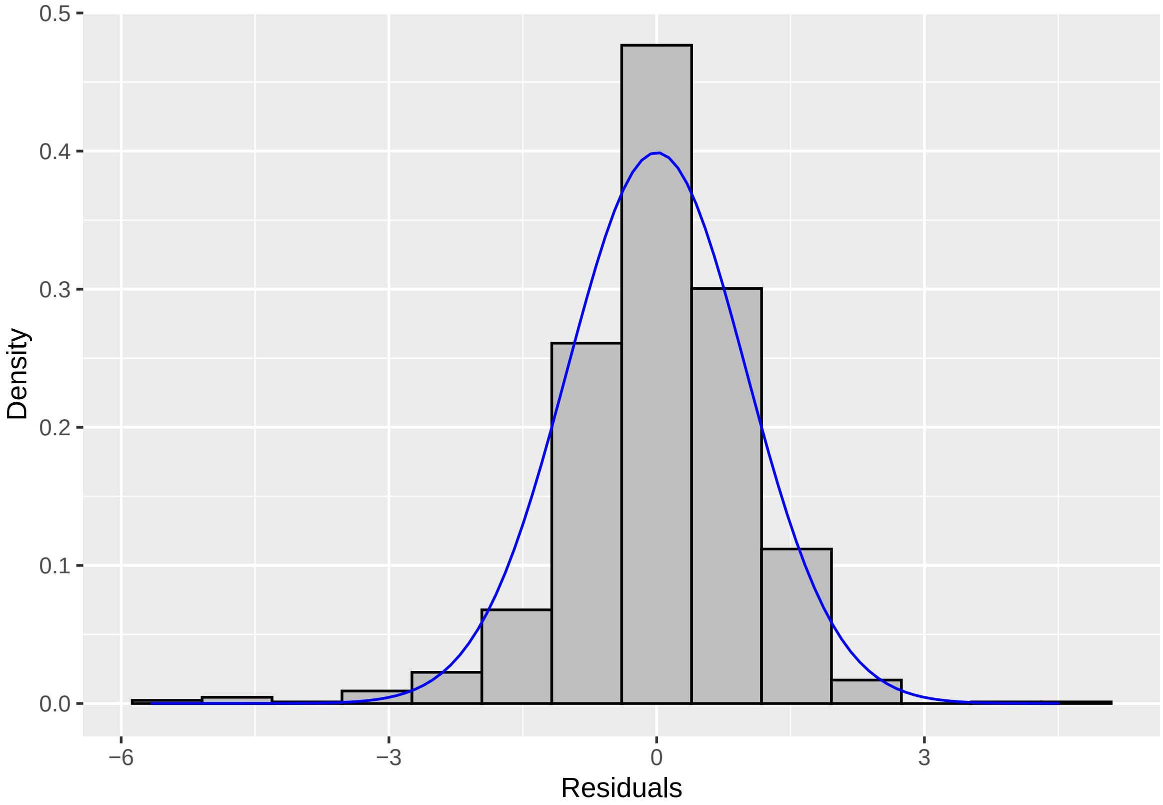
```
prog$group = as.factor(prog$group)
prog = prog %>%
  mutate(timeSqr = time^2, timeCub = time^3)

model1 = lme(PDG ~ time + group : time + timeSqr + group: timeSqr ,
                     data = prog,
                     random = ~ 1 + time + timeSqr| id,
                     method = "REML")
```

```
summary(model1)
```

```
## Linear mixed-effects model fit by REML
##   Data: prog
##        AIC      BIC    logLik
##   2623.384 2683.691 -1299.692
##
## Random effects:
##  Formula: ~1 + time + timeSqr | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev      Corr
## (Intercept) 0.875009015 (Intr) time
## time        0.048000492  0.451
## timeSqr     0.004289045 -0.526 -0.791
## Residual    0.665482856
##
## Fixed effects:  PDG ~ time + group:time + timeSqr + group:timeSqr
##                      Value  Std.Error   DF   t-value p-value
## (Intercept)     0.02653361 0.12571394 1075  0.211063  0.8329
## time            0.16131529 0.01037027 1075 15.555549  0.0000
## timeSqr        -0.00552998 0.00102273 1075 -5.407050  0.0000
## time:group1    -0.02955356 0.01482117 1075 -1.994010  0.0464
## group1:timeSqr  0.00765360 0.00137907 1075  5.549832  0.0000
##  Correlation:
##                (Intr) time   timSqr tm:gr1
## time            0.307
## timeSqr        -0.376 -0.720
## time:group1    -0.002 -0.634  0.423
## group1:timeSqr  0.012  0.452 -0.641 -0.703
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -5.26306686 -0.52146622  0.06171988  0.62227309  3.64039137
##
## Number of Observations: 1130
## Number of Groups: 51
```

1c)

```
res_population = residuals(model1, type = "response", level = 0)

Sigma_i = extract.lme.cov(model1, prog)
L_i = t(chol(Sigma_i)) #block matrix of lower triangular Cholesky factors
res_transformed <- solve(L_i) %*% res_population
tibble(r_star = res_transformed) %>%
  ggplot(aes(x = r_star)) +
  geom_histogram(aes(y = stat(density)), bins = 14, color = "black", fill = "gray") +
  geom_function(fun = dnorm, color = "blue") +
  labs(x = "Residuals", y = "Density")
```

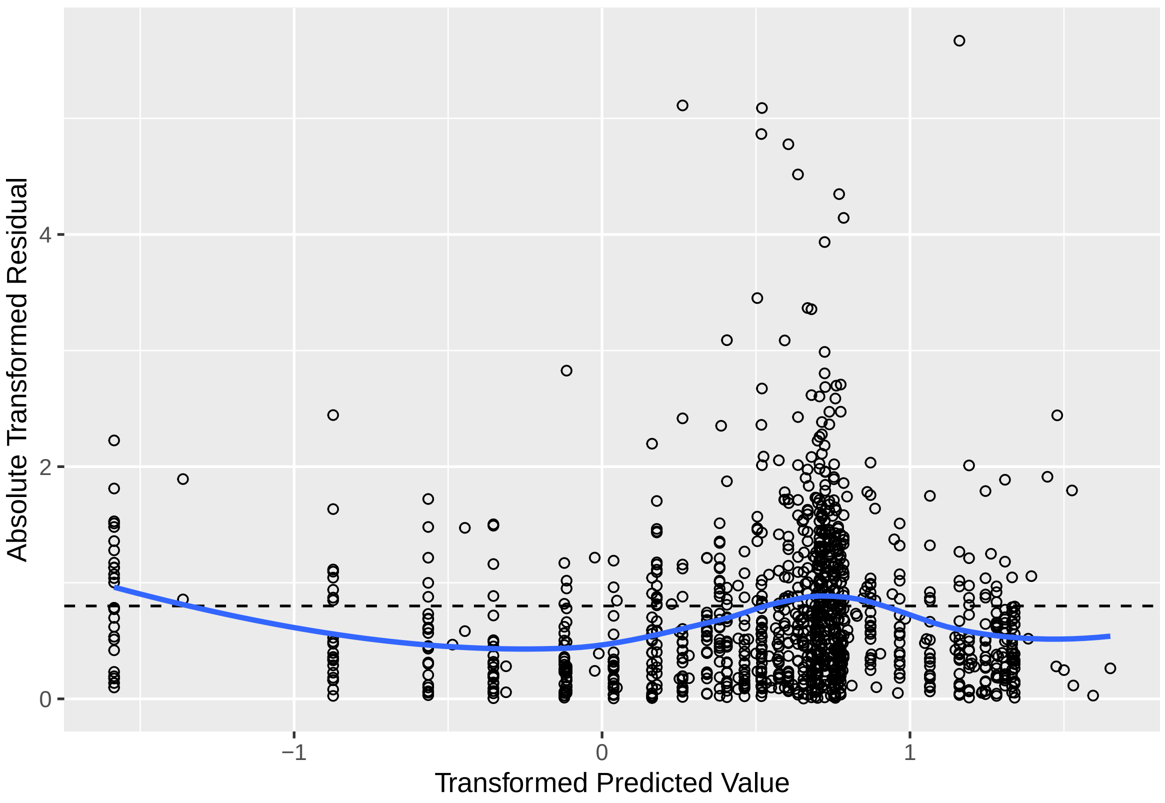


By examining the histogram of transformed residuals, we can assess their distributional characteristics, such as skewness and kurtosis, and compare them to the expected distribution (e.g., a normal distribution). Additionally, overlaying the density function (e.g., the normal distribution) on the histogram allows for visual comparison and evaluation of the fit. Therefore, transforming residuals is a valuable step in model diagnostics, helping to verify the assumptions of the statistical model, identify potential issues, and guide the need for further refinements or adjustments.

1d)

```
mu_hat = fitted(model1, level = 0)
mu_hat_transformed = solve(L_i) %*% mu_hat
abs_res_transformed = abs(res_transformed)

tibble(x = mu_hat_transformed, y = abs_res_transformed) %>%
  ggplot(aes(x = x, y = y)) +
  geom_hline(yintercept = 0.8, linetype = "dashed") +
  geom_point(shape = 1) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Transformed Predicted Value", y = "Absolute Transformed Residual")
```

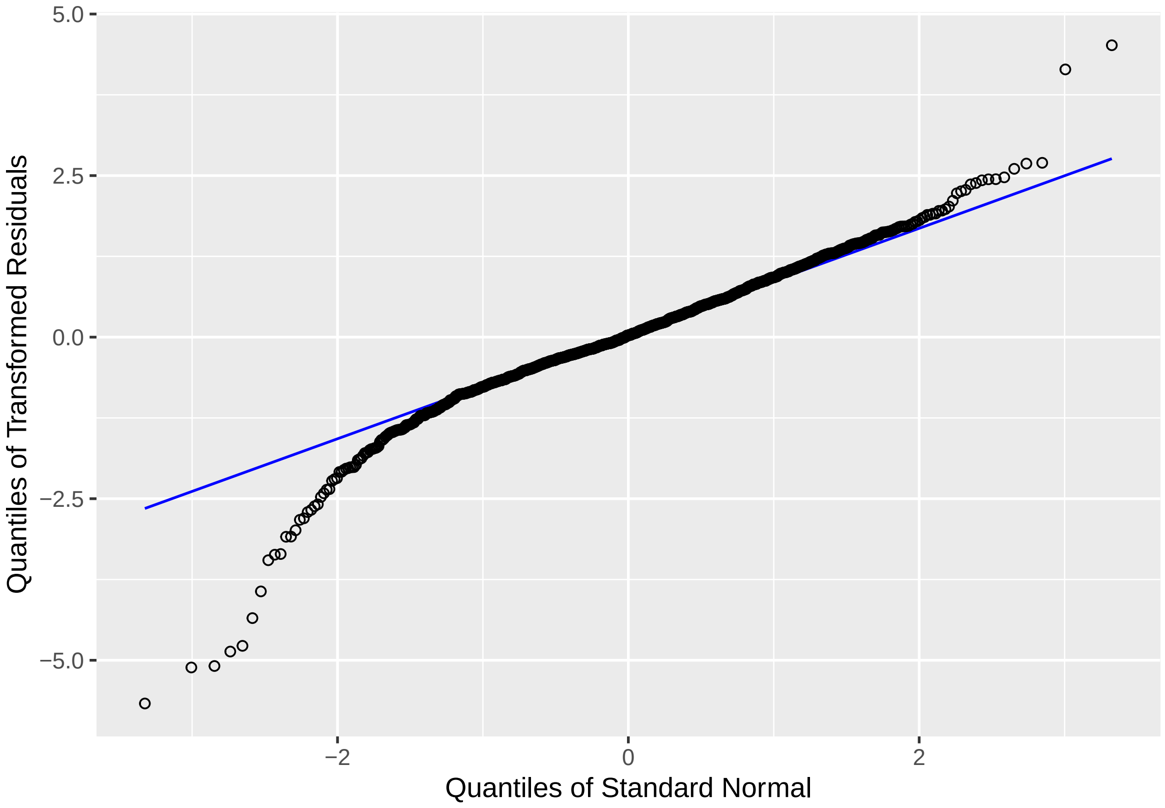
```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Smoothed line is roughly around 1, implying our systematic component is fitting the data well.

1e)

```
tibble(r_star = res_transformed) %>%
  ggplot(aes(sample = r_star)) +
  geom_qq_line(color = "blue") +
  geom_qq(shape = 1) +
  labs(x = "Quantiles of Standard Normal", y = "Quantiles of Transformed Residuals")
```



The bottom end of the Q-Q plot deviates from the straight line but the upper end is not, therefore, the distribution has a longer tail to its left and is left-skewed.

1f)

```
mahalanobis_distance = function(x){
  x <- as.matrix(x)
  t(x) %*% x
}

mahalanobis_data <- tibble(id = prog$id, r_star = res_transformed) %>%
  group_by(id) %>%
  nest() %>%
  mutate(df = map_dbl(data, ~nrow(.x)))%>%
  mutate(d = map_dbl(data, ~mahalanobis_distance(.x)))%>%
  mutate(p_value = pchisq(d, df, lower.tail= FALSE))


mahalanobis_data %>%
  arrange(p_value)
```

```
## # A tibble: 51 x 5
## # Groups:   id [51]
##       id data                 df     d  p_value
##    <int> <list>            <dbl> <dbl>    <dbl>
##  1    10 <tibble [23 x 1]>    23  98.5 2.54e-11
##  2    42 <tibble [21 x 1]>    21  82.1 3.57e- 9
##  3    43 <tibble [24 x 1]>    24  72.6 8.74e- 7
##  4    23 <tibble [23 x 1]>    23  61.0 2.71e- 5
##  5    15 <tibble [9 x 1]>      9  32.7 1.53e- 4
##  6     8 <tibble [22 x 1]>    22  47.3 1.32e- 3
##  7    48 <tibble [24 x 1]>    24  48.6 2.10e- 3
##  8    26 <tibble [24 x 1]>    24  47.2 3.16e- 3
##  9    27 <tibble [23 x 1]>    23  44.3 4.82e- 3
## 10     7 <tibble [21 x 1]>    21  31.2 7.10e- 2
## # ... with 41 more rows
```

```
sum(mahalanobis_data$p_value<0.05)
```

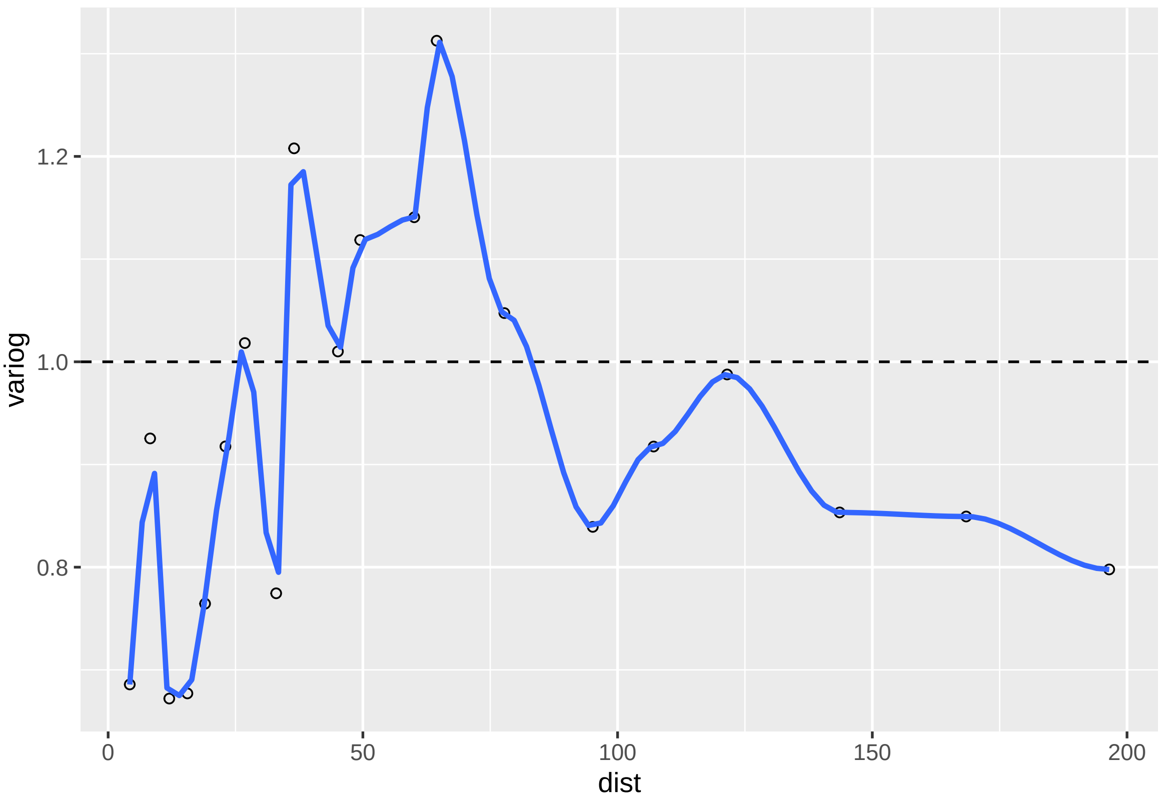
```
## [1] 9
```

9

1g)

```
Variogram(model1,
          data = prog,
          form = ~ 1 + time + timeSqr| id ,
          resType = "normalized") %>%
  as_tibble() %>%
  ggplot(aes(x = dist, y = variog)) +
  geom_hline(yintercept = 1, linetype = "dashed") +
  geom_point(shape = 1) +
  geom_smooth(method = "loess", se = FALSE, span = 0.1)
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 3.2814

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 4.9649

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 842.89
```

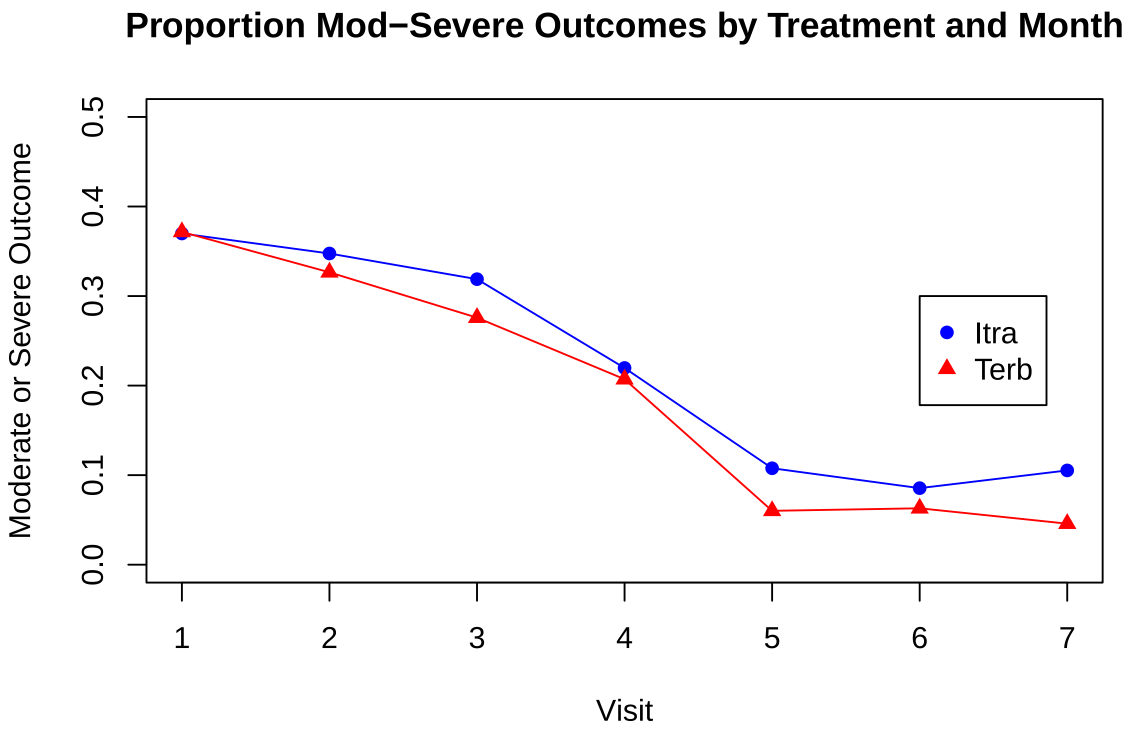


The plot of the semi-variogram fluctuates randomly around the horizontal line at 1.

2.
```
toes = read.table("/Users/virajvijaywargiya/Downloads/toenail-data.txt", header=FALSE)
names(toes) = c("ID","Y","Trt","Month","Visit")
toes$Trt = factor(toes$Trt, levels=c(0,1), labels=c("Itra","Terb"))
toes$ID = factor(toes$ID)
```

2a)

```
visits = c(1,2,3,4,5,6,7)
plot(visits, unlist(by(toes[toes$Trt=="Itra",]$Y, toes[toes$Trt=="Itra",5] , mean)), type="o",
   pch=16, col="blue",xlab="Visit", ylab="Moderate or Severe Outcome",
      main="Proportion Mod-Severe Outcomes by Treatment and Month", ylim=c(0,0.5))

points(visits, unlist(by(toes[toes$Trt=="Terb",]$Y, toes[toes$Trt=="Terb",5] , mean)), type="o",
   pch=17, col="red")

legend(6,.3,c("Itra","Terb"), col=c("blue","red"), pch=c(16,17))
```



The estimated proportion of moderate/severe infection for both the treatment groups decreases overtime (as month increases). However, it decreases slightly faster for treatment "Terb" than that for "Itra".

2b) logit(P(Y = 1)) = B0 + B1 * Month + B2 * Treatment + B3 * (Month * Treatment)

2c)

```
mod1gee= geeglm(Y ~ 1+Month*Trt , family=binomial, id=ID, corstr="exchangeable", data=toes)
summary(mod1gee)
```

```
##
## Call:
## geeglm(formula = Y ~ 1 + Month * Trt, family = binomial, data = toes,
##     id = ID, corstr = "exchangeable")
##
##  Coefficients:
##                  Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)    -0.58192  0.17206 11.439 0.000719 ***
## Month          -0.17128  0.03000 32.596 1.13e-08 ***
## TrtTerb         0.00718  0.25949  0.001 0.977924
## Month:TrtTerb -0.07773  0.05411  2.064 0.150862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##               Estimate Std.err
## (Intercept)    1.088  0.5013
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.4218  0.2119
## Number of clusters:   294  Maximum cluster size: 7
```

2d) For the existing treatment group (coded as 0): When all other variables are held constant, a unit increase in Month is associated with a change in the log odds of moderate/severe infection by the coefficient B1. This means that for each additional month, the log odds of moderate/severe infection increase by B1.

For the new treatment group (coded as 1): When all other variables are held constant, a unit increase in Month is associated with a change in the log odds of moderate/severe infection by the sum of coefficients B1 and B3. This means that for each additional month, the log odds of moderate/severe infection increase by B1 + B3. The B3 coefficient represents the difference in the effect of Month between the new treatment group and the existing treatment group. If B3 is positive, it indicates that the new treatment (Terbinafine) has a greater increase in the log odds of moderate/severe infection over time compared to the existing treatment (Itraconazole). Conversely, if B3 is negative, it suggests that the new treatment has a smaller increase in the log odds of moderate/severe infection over time compared to the existing treatment.

2e)

```
mod3gee = geeglm(Y ~ 1+Trt , family=binomial, id=ID, corstr="exchangeable", data=toes)
anova(mod1gee, mod3gee)
```

```
## Analysis of 'Wald statistic' Table
##
## Model 1 Y ~ 1 + Month * Trt
## Model 2 Y ~ 1 + Trt
##   Df   X2 P(>|Chi|)
## 1  2 63.2   1.9e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2f) The AIC (Akaike Information Criterion) and likelihood ratio test are commonly used for model selection and hypothesis testing in traditional regression models. However, they cannot be directly applied to GEE models. This is because GEE estimation involves the specification of a working correlation structure, which affects the model’s asymptotic distribution. The AIC and likelihood ratio test rely on specific assumptions about the likelihood function, which may not hold under the GEE framework. Instead, hypothesis testing in GEE models typically involves score tests or Wald tests based on robust standard errors to account for the correlation structure.

2g) logit(P(Y_ij = 1)) = B0 + b_i + B1 * Month_ij + B2 * Treatment_ij + B3 * (Month_ij * Treatment_ij)

2h)

```
mod = glmer(Y ~ 1+Month*Trt  + (1 | ID), family=binomial, data=toes, nAGQ =  5)
summary(mod)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 5) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Y ~ 1 + Month * Trt + (1 | ID)
##    Data: toes
##
##      AIC      BIC   logLik deviance df.resid
##     1270     1298     -630     1260     1903
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
##  -3.10  -0.20  -0.10  -0.01  40.64
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 13.6     3.69
## Number of obs: 1908, groups:  ID, 294
##
## Fixed effects:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.4576     0.3947   -3.69  0.00022 ***
## Month          -0.3821     0.0434   -8.81  < 2e-16 ***
## TrtTerb        -0.1298     0.5378   -0.24  0.80925
## Month:TrtTerb  -0.1336     0.0662   -2.02  0.04343 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Month  TrtTrb
## Month       -0.194
## TrtTerb     -0.665  0.220
## Mnth:TrtTrb  0.207 -0.565 -0.312
```

2i) A lower AIC value indicates a better-fitting model relative to other models being compared. To conduct a test of whether or not to include Month in the model, including both the main effect and interaction term, you can compare the AIC values of two nested models: one with Month included and one without Month.

2j) For the average or typical subject in the dataset, a unit increase in Month is associated with a change in the odds of having a moderate/severe infection. The estimated effect can be interpreted as follows: On average, for every one-unit increase in Month, the odds of a moderate/severe infection occurring increase (or decrease) by a factor of exp(B1), where B1 represents the estimated coefficient for the effect of Month in the model. This interpretation assumes all other variables in the model are held constant.

2k)

```
coef(mod)$ID[1:5,]
```

```
##   (Intercept)   Month TrtTerb Month:TrtTerb
## 1      2.0166 -0.3821 -0.1298       -0.1336
## 2      0.2892 -0.3821 -0.1298       -0.1336
## 3     -0.6847 -0.3821 -0.1298       -0.1336
## 4     -0.6766 -0.3821 -0.1298       -0.1336
## 6      1.5544 -0.3821 -0.1298       -0.1336
```

logit(P(Y_ij = 1)) = 2.0166 - 0.3821*Month_ij - 0.1298*TrtTerb_ij - 0.1336*(Month_ij * TrtTerb_ij)

2l) The GEE model focuses on the population-average treatment effect and marginal relationship between covariates and the response, whereas the GLMM model allows for individual-specific treatment effects and incorporates subject-specific random effects. The GEE model is suitable for population-level inference and marginal effects estimation, while the GLMM model is appropriate for individual-level inference and accounting for both fixed and random effects.

3.
```
skin = read.csv("/Users/virajvijaywargiya/Downloads/skin.csv")
skin$trt_num = skin$trt
skin$trt = factor(skin$trt, levels=c('0','1'),labels=c('Placebo','beta carotene'))
```

3a) log(E(Y)) = B0 + B1 * Treatment + B2 * Year + B3 * (Treatment * Year)

3b) The offset term is typically used in generalized linear models (GLMs) to account for exposure or time-at-risk when modeling rates or counts. It allows for the inclusion of an offset variable that represents the logarithm of the expected exposure or time-at-risk.

In the given scenario of preventing non-melanoma skin cancer, the outcome variable Y is already defined as the count of new skin cancers per year, which inherently accounts for the time-at-risk. Therefore, there is no need to include an offset term in the model.

By including the Year variable as a predictor in the model, we are implicitly accounting for the variation in time-at-risk across different years of follow-up. The coefficient associated with Year (B2) captures the effect of the follow-up year on the count of new skin cancers while considering the differences in exposure time.

Hence, in this specific case, the inclusion of an offset term is unnecessary since the count variable itself represents the time-at-risk, and the Year variable adequately captures the impact of time in the model.

3c)

```
gee_2 = geeglm(y ~ year + trt + year*trt,data = skin,family = poisson(link = "log"),id = id,  corst
summary(gee_2)
```

```
##
## Call:
## geeglm(formula = y ~ year + trt + year * trt, family = poisson(link = "log"),
##     data = skin, id = id, corstr = "ar1")
##
##  Coefficients:
##                       Estimate Std.err   Wald Pr(>|W|)
## (Intercept)            -1.3289  0.1234 115.93   <2e-16 ***
## year                   -0.0116  0.0329   0.12     0.73
## trtbeta carotene        0.0657  0.1644   0.16     0.69
## year:trtbeta carotene   0.0327  0.0484   0.46     0.50
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     2.62   0.377
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.545   0.111
## Number of clusters:   1683  Maximum cluster size: 5
```

3d) The coefficient of Treatment (B1) in the model represents the change in the log count of new skin cancers per year when switching from the placebo (Treatment = 0) to the treatment (Treatment = 1), while holding the year constant. In terms of interpretation, if B1 is positive and statistically significant, it suggests that the treatment (beta carotene) is associated with an increase in the log count of new skin cancers compared to the placebo. Conversely, if B1 is negative and statistically significant, it indicates that the treatment is associated with a decrease in the log count of new skin cancers compared to the placebo.

3e) The coefficient of Year (B2) in the model represents the change in the log count of new skin cancers per unit increase in the follow-up year, while holding the treatment constant. The interpretation is that if B2 is positive and statistically significant, it indicates that there is an increasing trend in the log count of new skin cancers over time, regardless of the treatment received. Conversely, if B2 is negative and statistically significant, it suggests a decreasing trend in the log count of new skin cancers over time, independent of the treatment.

3f) The coefficient of the interaction term (Treatment * Year; B3) in the model represents the additional effect on the log count of new skin cancers due to the combined influence of both treatment and follow-up year. The interpretation of this coefficient is that if B3 is positive and statistically significant, it suggests that the treatment effect on the log count of new skin cancers varies depending on the follow-up year. In other words, the impact of treatment differs across different years. If B3 is negative and statistically significant, it indicates that the treatment effect on the log count of new skin cancers also varies with the follow-up year, but in the opposite direction.

3g) log(E(Yij)) = B0 + B1 * Treatment_ij + B2 * Year_ij + B3 * (Treatment_ij * Year_ij) + b0i + b1i * Year_ij + eij

3h)

```
glmm_3 = glmer(y ~ year+trt+trt*year + (1+year | id), family=poisson, data=skin , nAGQ=0)
summary(glmm_3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 0) [glmerMod]
##  Family: poisson  ( log )
## Formula: y ~ year + trt + trt * year + (1 + year | id)
##    Data: skin
##
##      AIC      BIC   logLik deviance df.resid
##     8429     8477    -4208     8415     7074
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.535 -0.359 -0.283 -0.265  3.602
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  id     (Intercept) 2.416    1.555
##         year        0.101    0.317    -0.46
## Number of obs: 7081, groups:  id, 1683
##
## Fixed effects:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.8569     0.1066  -17.41   <2e-16 ***
## year                    -0.0365     0.0325   -1.12     0.26
## trtbeta carotene         0.0897     0.1469    0.61     0.54
## year:trtbeta carotene    0.0209     0.0447    0.47     0.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) year   trtbtc
## year        -0.775
## trtbetcartn -0.726  0.563
## yr:trtbtcrt  0.563 -0.726 -0.772
```

3i) In the random effects model, the estimated coefficient of Year (-0.0365) indicates the average rate of change in the log count of new skin cancers per unit increase in the follow-up year, while considering the random intercept and random slope. A negative coefficient suggests that, on average, there is a decreasing trend in the log count of new skin cancers over time. For every one unit increase in the follow-up year, the expected log count of new skin cancers decreases by 0.0365, holding the treatment and individual-specific variations constant.

3j) In the random effects model, the estimated coefficient of the interaction term (year:trtbeta carotene) is 0.0209. This coefficient represents the additional effect on the log count of new skin cancers due to the combined influence of both treatment (beta carotene) and follow-up year, while accounting for random intercept and random slope. A positive coefficient suggests that the treatment effect varies depending on the follow-up year. Specifically, for every one unit increase in the follow-up year, the treatment (beta carotene) is associated with a 0.0209 increase in the log count of new skin cancers, holding individual-specific variations constant.