JAYADITYA NATH   STAT 8320 (HW4)

1. Consider the linear mixed model ;

$$\underset{\sim}{Y} \,|\, \underset{\sim}{\alpha} \sim N(X\underset{\sim}{\beta} + Z\underset{\sim}{\alpha}, \Sigma)$$

$$\underset{\sim}{\alpha} \sim N(\underset{\sim}{0}, D)$$

(a) Now, we observe that $E_Y(Y|\underset{\sim}{\alpha}) = X\underset{\sim}{\beta} + Z\underset{\sim}{\alpha}$

By law of total expectation, we know that,

$$E_{\underset{\sim}{\alpha}}[E_Y(Y|\underset{\sim}{\alpha})] = E(Y)$$

So, $E(Y) = E_{\underset{\sim}{\alpha}}(X\underset{\sim}{\beta} + Z\underset{\sim}{\alpha}) = X\underset{\sim}{\beta} + Z\,E(\underset{\sim}{\alpha}) = X\underset{\sim}{\beta} + 0 = X\underset{\sim}{\beta}$ □.

(b) Now, we observe that $V_Y(Y|\underset{\sim}{\alpha}) = \Sigma$

By law of total variance, we know that,

$$V(Y) = V_{\underset{\sim}{\alpha}}(E_Y(Y|\underset{\sim}{\alpha})) + E_{\underset{\sim}{\alpha}}(\mathrm{var}_Y(Y|\underset{\sim}{\alpha}))$$

$$= V_{\underset{\sim}{\alpha}}(X\underset{\sim}{\beta} + Z\underset{\sim}{\alpha}) + E_{\underset{\sim}{\alpha}}(\Sigma)$$

$$= ZDZ' + \Sigma \qquad [\text{as, } V_{\underset{\sim}{\alpha}}(X\underset{\sim}{\beta}) = 0]$$

■.

2. Let, $Y_i|\lambda \overset{iid}{\sim} Poi(\lambda)$; $i = 1(1)m$

The density of Y_i is:

$$f_{Y_i|\lambda}(y_i) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \quad ; \quad y_i = 0, 1, 2, \ldots$$

Also, $\lambda \sim$ Gamma (a, b)

The density of λ is:

$$f_\lambda(\lambda) = \frac{1}{\Gamma(a)\, b^a} \lambda^{a-1} e^{-\lambda/b} \quad ; \quad 0 < \lambda < \infty.$$

(a) To show that:

$$f_{Y_i}(y_i) = \int f_{Y_i|\lambda}(y_i)\cdot f(\lambda)\,d\lambda \sim \text{Negative Binomial.}$$

$$= \int_0^\infty \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\cdot\frac{1}{\Gamma(a)\cdot b^a}\cdot\lambda^{a-1}\cdot e^{-\lambda/b}\,d\lambda$$

$$= \frac{1}{\Gamma(a)}\cdot\frac{1}{b^a}\cdot\frac{1}{y_i!}\underbrace{\int_0^\infty \lambda^{(y_i+a)-1}\cdot e^{-\lambda/\left(\frac{b}{b+1}\right)}\,d\lambda}_{=\Gamma(y_i+a)\cdot\left(\frac{b}{b+1}\right)^{y_i+a}}$$

$$= \frac{(y_i+a-1)!}{(a-1)!\,y_i!}\cdot\left(\frac{b}{b+1}\right)^{y_i}\cdot\left(\frac{1}{b+1}\right)^{a}$$

$$= \binom{y_i+a-1}{\cancel{a-1}\,y_i}\cdot\left(\frac{b}{b+1}\right)^{y_i}\cdot\left(1-\frac{b}{b+1}\right)^{a} = \binom{y_i+a-1}{y_i}\cdot\left(\frac{1}{b+1}\right)^{a}\cdot\left(1-\frac{1}{b+1}\right)^{y_i}$$

So, $Y_i \sim$ Negative Binomial$\left(a, \frac{1}{b+1}\right)$

∎.

Problem 2.

Part (b)

This derivation of the negative binomial distribution is surely relevant to the notion of over-dispersed count data. Over-dispersion occurs when the variance of the count data is greater than what is expected under a Poisson distribution. This can happen due to unobserved heterogeneity, extra-Poisson variation, or clustering effects.

In this context, the Gamma distribution serves as a way to model the variability in the rate parameter λ among the different observations. By incorporating a Gamma prior for λ, we allow for this variability, which can capture over-dispersion in the data. The resulting negative binomial distribution for y_i accounts for both the Poisson variability in the counts and the additional variability in λdue to the random effect. Therefore, this derivation provides a probabilistic framework for modeling over-dispersed count data by explicitly considering the random variation in the rate parameter.

Part (c)

If we assume that λ follows a log-normal distribution instead of a Gamma distribution, the resulting marginal distribution of y_i would not still be a negative binomial distribution. Instead, it would probably be a Poisson-log-normal distribution.The Poisson-log-normal distribution arises when a Poisson-distributed random variable is multiplied by a log-normally distributed random variable.

I feel that this observation is relevant to how we perform estimation in Generalized Linear Mixed Models (GLMMs) because it highlights the importance of appropriately modeling the distribution of random effects. GLMMs allow for the incorporation of random effects to account for correlation or heterogeneity in the data, and the choice of distribution for these random effects can impact the model's performance and interpretation.

PROBLEM 3.

```r
#Load the data and reshape
data_3 = read.table("C:/Users/Jayaditya Nath/Documents/growthdata.dat")

data_3 = reshape(data_3, varying = list(2:7),
                 v.names = "Growth",
                 timevar = "Plant",
                 times = 1:6,
                 direction = "long")
# Converting 'Time' column to numeric
data_3$Time = as.numeric(as.character(data_3$V1))
data_3 = data_3[,2:5]
```

To find the initial parameter estimates, I assume that there is no random effects in the model, which in turn gives a non-linear model. Now, I try to find maximum likelihood estimates(MLE) and set them as the starting values to fit the given model. To do this, I minimize the negative log-likelihood of the given model using the *optim()* with the BFGS quasi-Newton algorithm.

```r
#Finding initial parameter values
neg_ll = function(parameters, response, time) {
  beta_1 = parameters[1]
  beta_2 = parameters[2]
  beta_3 = parameters[3]

  # Calculate the predicted values using the given function
  fit = beta_1 / (1 + exp(- (time - beta_2) / beta_3))

  # Calculate the negative log-likelihood
  neg_log_likelihood = sum(dnorm(response, mean = fit,log = TRUE))

  return(-neg_log_likelihood)
}

# Initial values for optimization
initial_params = c(200,700,1200)

# Optimize the negative log-likelihood function
optimized = optim(initial_params, neg_ll, time = data_3$Time, response = data_3$Growth,method = "BFGS")

# Extract the MLE estimates
params_init = optimized$par
```

Finally, I get the initial parameter estimates for β_1,β_2 and β_3 as 199.649757, 797.7467087 and 300.6553514.

```
library(nlme)

#Fitting the model
growth_model = nlme::nlme(Growth ~ (b1 + u) / (1 + exp(-(Time - b2) / b3)),
               data = data_3,
               fixed = b1 + b2 + b3 ~ 1,
               random = list(u ~ 1),
               groups = ~ Plant,
               start = list(fixed = c(b1 = params_init[1], b2 = params_init[2], b3 = params_init[3])),
               control=nlme::nlmeControl(pnlsTol=0.1))

summary(growth_model)
```

```
## Nonlinear mixed-effects model fit by maximum likelihood
##   Model: Growth ~ (b1 + u)/(1 + exp(-(Time - b2)/b3))
##   Data: data_3
##        AIC      BIC    logLik
##   443.3947 453.8665 -216.6974
##
## Random effects:
##  Formula: u ~ 1 | Plant
##                u Residual
## StdDev: 36.6648 7.059439
##
## Fixed effects:  b1 + b2 + b3 ~ 1
##       Value Std.Error DF  t-value p-value
## b1 199.6515  15.62142 52 12.78063       0
## b2 797.7538  15.05123 52 53.00257       0
## b3 300.6635  11.89436 52 25.27783       0
##  Correlation:
##    b1    b2
## b2 0.148
## b3 0.133 0.584
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.14299738 -0.58049127 -0.05987433  0.58532766  2.38214259
```

```
##
## Number of Observations: 60
## Number of Groups: 6
```

```r
#Testing for the presence of random effect
growth_model_rand_comp <- nlme(Growth ~ (b1) / (1 + exp(-(Time - b2) / b3)),
                    data = data_3,
                    fixed = b1 + b2 + b3 ~ 1,
                    groups = ~ Plant,
                    start = list(fixed = c(b1 = params_init[1], b2 = params_init[2], b3 = params_init[3
                    control=nlmeControl(pnlsTol=0.1))

lmtest::lrtest(growth_model,growth_model_rand_comp)
```

```
## Likelihood ratio test
##
## Model 1: Growth ~ (b1 + u)/(1 + exp(-(Time - b2)/b3))
## Model 2: Growth ~ (b1)/(1 + exp(-(Time - b2)/b3))
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -216.70
## 2  10 -216.15  5 1.0944     0.9546
```

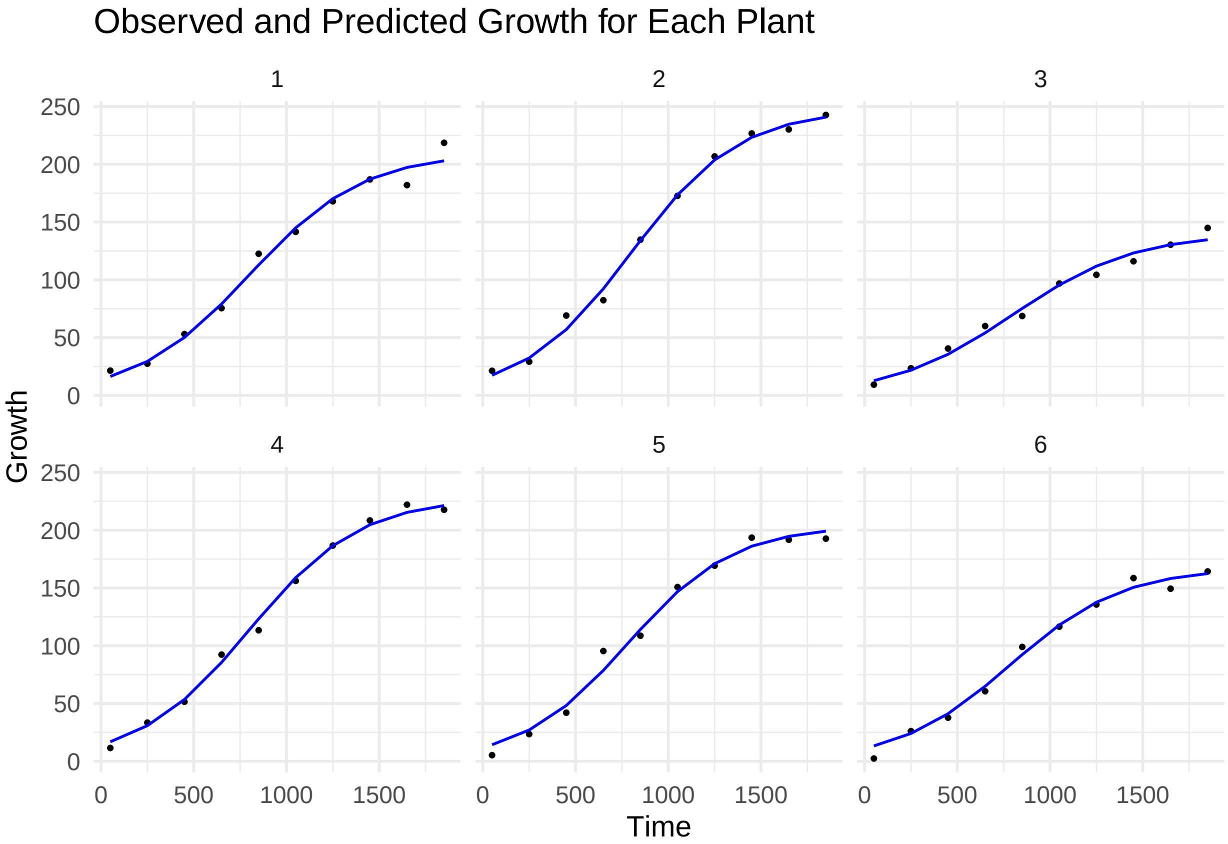
From the ANOVA table, I see that the p-value is greater than $\alpha = 0.05$, thus not allowing to reject the null hypothesis at 5% level of significance. So, I can conclude that the adding the random effects to the given model does not necessarily improve the model fit.

```r
#Testing for beta_3 = 350
growth_model_beta3_comp = nlme(Growth ~ (b1 + u) / (1 + exp(-(Time - b2) / 350)),
                          data = data_3,
                          fixed = b1 + b2 ~ 1,
                          random = list(u ~ 1),
                          groups = ~ Plant,
                          start = list(fixed = c(b1 = params_init[1], b2 = params_init[2])),control = nl


anova(growth_model, growth_model_beta3_comp)
```

```
##                         Model df      AIC      BIC    logLik   Test  L.Ratio
## growth_model                1  5 443.3947 453.8665 -216.6974
## growth_model_beta3_comp     2  4 456.1259 464.5033 -224.0630 1 vs 2 14.73118
##                         p-value
## growth_model
## growth_model_beta3_comp   1e-04
```

From the ANOVA table, I see that the p-value is less than $\alpha = 0.05$, thus allowing to reject the null hypothesis at 5% level of significance. So, I can conclude that the $\beta_3 = 350$ is not its true value.

```r
#Plotting predicted values(on the basis of the model without random effects) and observed growth
library(ggplot2)

ggplot(data_3, aes(x = Time, y = Growth)) +
  geom_point(size = 0.5) +
  geom_line(aes(y =predict(growth_model_rand_comp,newdata = data_3)), color = "blue") +
  facet_wrap(vars(Plant)) +
  labs(title = "Observed and Predicted Growth for Each Plant",
       x = "Time",
       y = "Growth")+
  theme_minimal()
```



From the plots, it can be observed that the model without the random effects provides quite a decent fit.

PROBLEM 4.

```
# Load necessary libraries
library(lme4)
library(RLRsim)
# Load the data
pred_data <- readRDS("C:/Users/Jayaditya Nath/Documents/pred_data.RDS")
```

For the given question, we are interested in knowing whether the rate of predation varies among different groups of animals, i.e, the treatment effect is significant or not. For this, we consider the following mixed effects regression model :

$Y_{ij} \sim \text{Bernoulli}(p_{ij})$

Let Y_{ij} denote the response variable indicating predation for the i-th observation in the j-th block, where , i = 1,2,. . . ,n_j and j = 1,2,. . . ,10 and n_j represents the number of observations in the j-th block.

Now,let $Y_{ij} \sim \text{Bernoulli}(p_{ij})$

$\text{logit}(p_{ij}) = \beta_0 + \alpha_i + \tau_j$

where p_{ij} s the rate of predation for the i-th observation in the j-th block.

$\text{logit}(p_{ij})$ = log-odds of predation

β_0 = fixed intercept

α_i = represents the fixed effect of the group variable for the i-th observation.

τ_j = represents the random effect of the j-th block.

$\tau_j \sim \text{Normal}(0,\sigma_\tau^2)$

```r
# Fit the mixed-effects model
model_4 <- glmer(pred ~ treat + (1 | blk), data = pred_data,family = binomial(link = "logit"))

# Check model summary
summary(model_4)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: pred ~ treat + (1 | blk)
##    Data: pred_data
##
##      AIC      BIC   logLik deviance df.resid
##     70.7     82.6    -30.4     60.7       75
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -23.2075  -0.1730   0.0976   0.2980   0.8944
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  blk    (Intercept) 11.81    3.437
## Number of obs: 80, groups:  blk, 10
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.096      1.811   2.814  0.00490 **
## treatcrabs    -3.842      1.465  -2.623  0.00871 **
## treatshrimp   -4.431      1.551  -2.856  0.00429 **
## treatboth     -5.599      1.724  -3.247  0.00117 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##             (Intr) trtcrb trtshr
## treatcrabs  -0.625
## treatshrimp -0.652  0.736
## treatboth   -0.687  0.732  0.767
```

```r
#Checking the significance of the random effects
model_4_rand = glmer(pred ~ treat + (1 | blk), data = pred_data)
exactRLRT(model_4_rand)
```

```
##
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 36.974, p-value < 2.2e-16
```

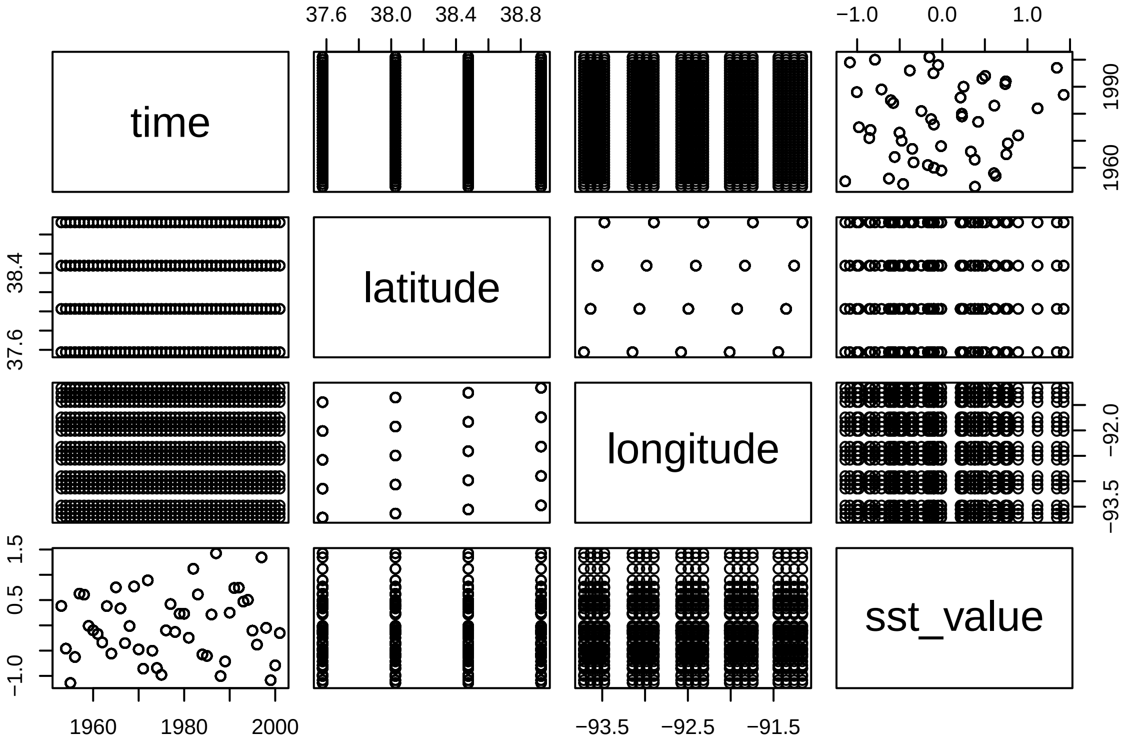
From the model summary, the group effect for the animals seem to be significant as a whole. Also, the treatments : shrimp, crab and both are significant compared to the baseline treatment effect(i.e, no treatment).

It seems from the output of the random Likelihood Ratio Test that the p-value is less than $\alpha = 0.05$, thus, we can reject the null hypothesis at 5% level of significance. So, the random effect has a significant variance different from 0.

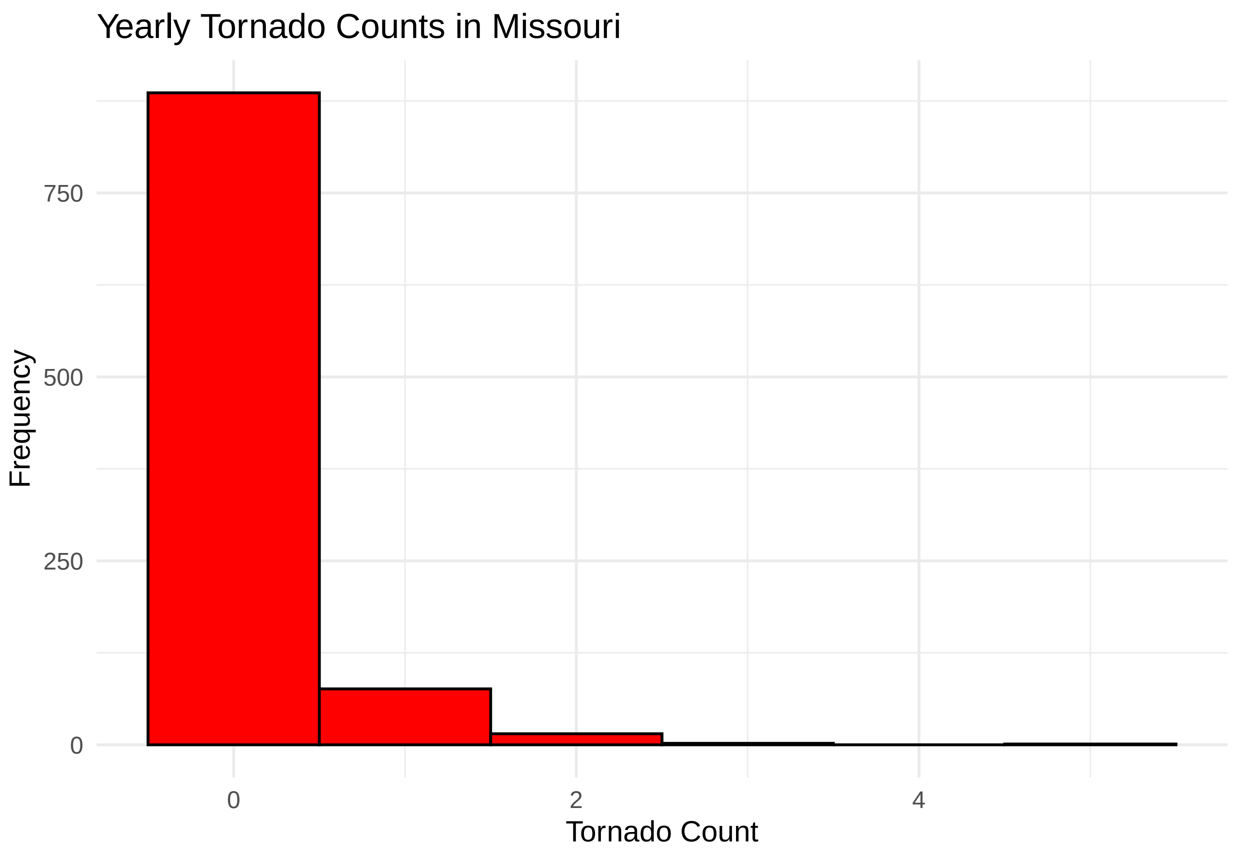
PROBLEM 5.

Introduction : The motivation of this problem is to study the relation between a set of covariates(sea surface temperature,latitude-longitude and time) and the occurence of tornadoes. We are interested to study that if there is a dependence between yearly tornado counts in the central portion of Missouri to the tropical Pacific SST.

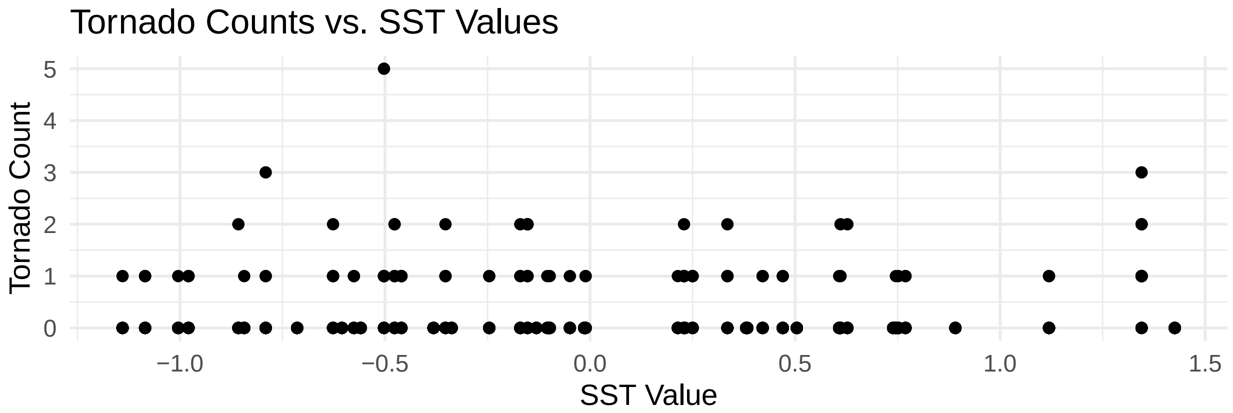
Exploratory data analysis

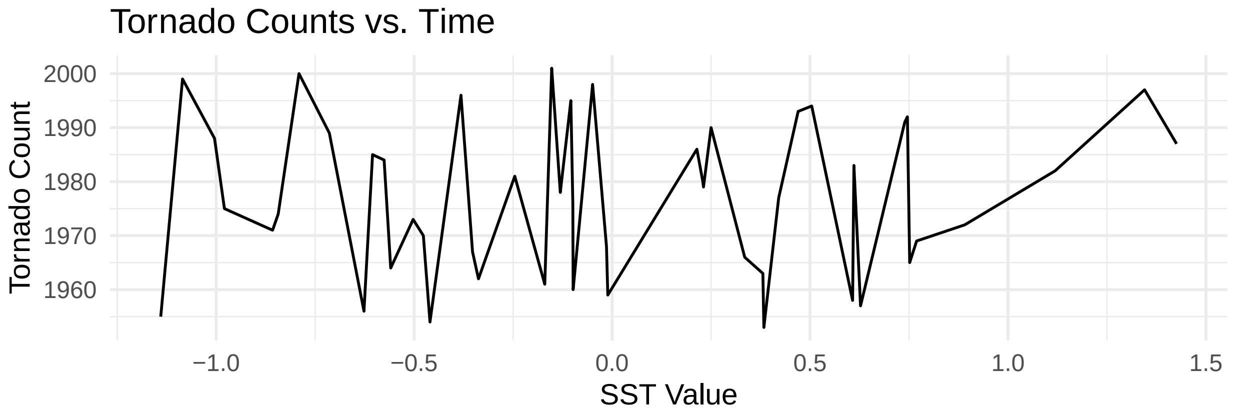


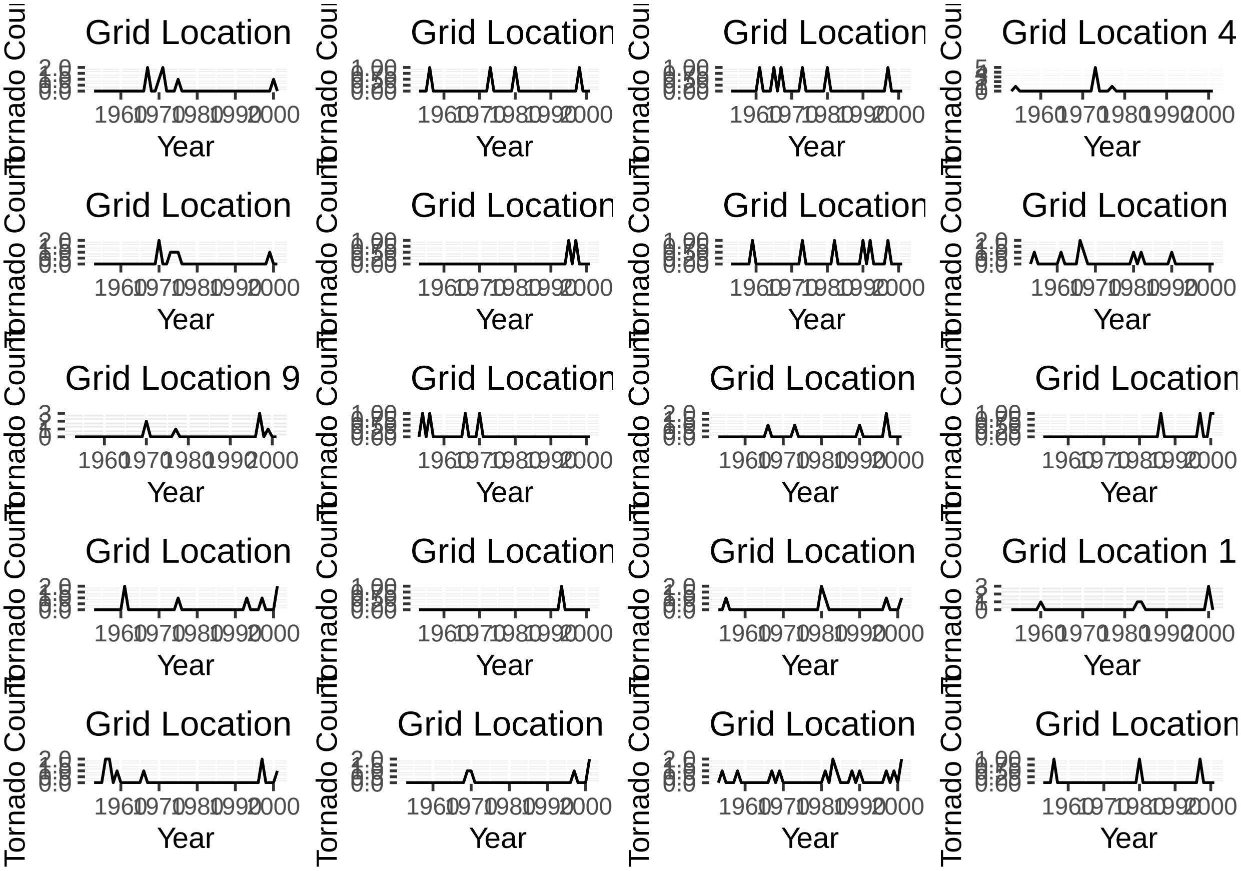
There does not seem to be any multicollinearity present among the covariates.



The above plot shows that the responses are heavily right-skewed with an excessive number of zeroes, hinting at the usage of a zero-inflated count data model.


Tornado Counts vs. SST Values
Tornado Count
0
1
2
3
4
5
−1.0
−0.5
0.0
0.5
1.0
1.5
SST Value


Tornado Counts vs. Time
Tornado Count
1960
1970
1980
1990
2000
−1.0
−0.5
0.0
0.5
1.0
1.5
SST Value



The plots are showing that there is no specific pattern in which the covariates individually influence the response variable. So, a linear model or a quadratic model can not be specified.

Discussions on models considered

On the basis of the exploratory data analysis performed, I would be using the generalized linear mixed model. Since tornado counts are non-negative integers and potentially over-dispersed, I have used a Poisson likelihood along with the log link function, which ensures that the fitted values are non-negative. For this, we have to assume that the counts of tornado occurrences in central Missouri in different years are conditionally independent given the SST values. Also, the Poisson distribution belongs to an Exponential family, which is a requirement for Generalized Linear Models.

I have considered the SST as fixed effects and considering case and time as random effects would be reasonable in some sense as repeatedly many observations have been recorded over time for each of the cases.

Initially, I have considered only fixed effects and compared them to a model having random effects and after performing ANOVA, observed that the random effects are significant. I have added a random intercept to time, a random slope to SST for each case and a random intercept to the interaction between time and case. I have used **Likelihood Ratio Test** to find that the interaction effect between time and case is also significant.

To check for over-dispersion, I plotted out the standardized quantile residuals using the *DHARMa* module and observed that the issue of over-dispersion persists. To deal with this, I have finally fitted a negative-binomial(log link) generalized linear regression mixed effects model with the same formula, considering the zero-inflated data structure also.

The model looks like :

Let Y_{ijk} represent the tornado count corresponding to the SST$_i$ at case j time point k.

So, $Y_{ij}|p_{ij} \sim NB(r, p_{ijk})$, where, r is the dispersion parameter and p_{ijk} is the probability of observing Y_{ijk} tornadoes.

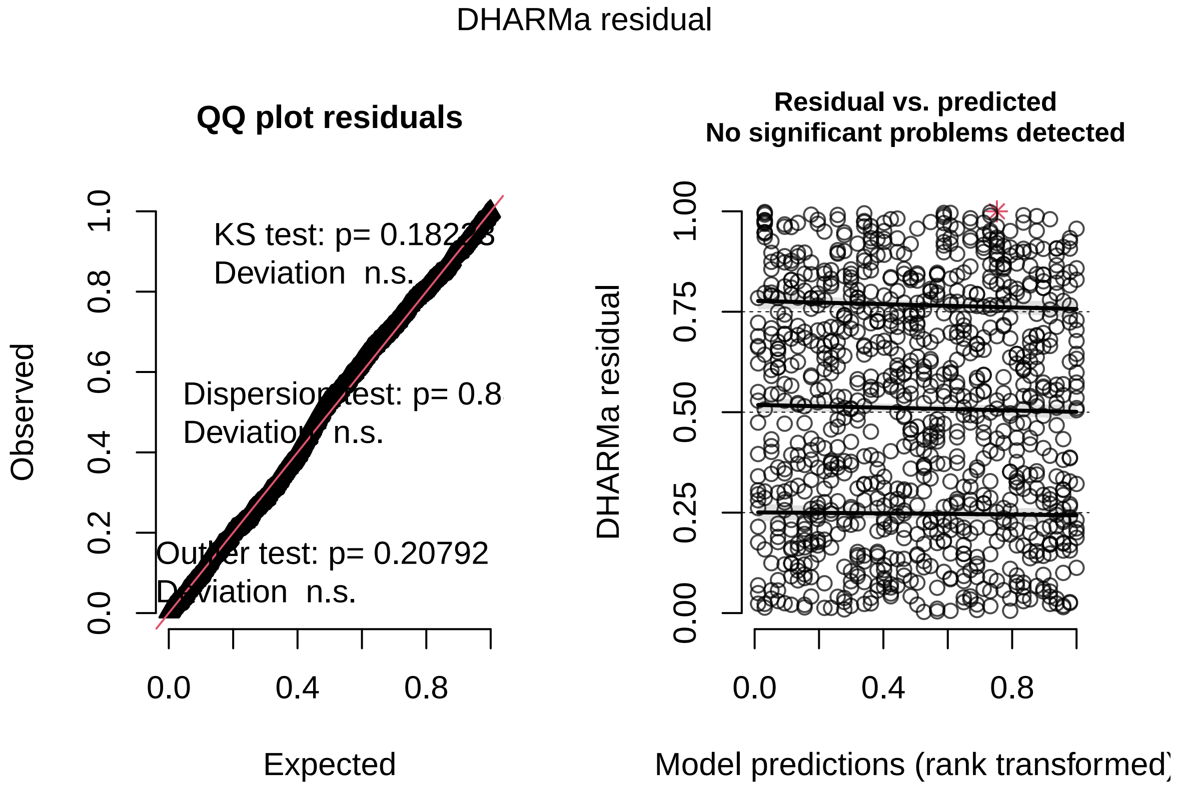
Again, $logit(p_{ijk}) = \beta_0 + \beta_1 SST_i + b_0 + b_1 SST_i + b_2$,

where, β_0 and β_1 are the fixed effects coefficients for the intercept and SST value, respectively.

b_0, b_1 and b_2 are the random effects for the intercept due to time, SST value due to case and intercept due interaction of time and case, respectively, where, b_0, b_1 and $b_2 \sim iidN(0, \sigma_\tau^2)$.

SST_i is the is ith the sea surface temperature.

Model fit



From the normal Q-Q plot, it is clear that the normality assumptions are met reasonably. The residual vs fitted plots show that there is no specific pattern in the points, which signifies the presence of linearity. Also, the points are spread optimally around the center line, signifying homoscedasticity of error variances. The plots previously have shown that the responses and thus the residuals are in no way autocorrelated. The final model also has the lowest AIC value among all other models which I have checked, thus, giving the best fit to the given data.

Inference

A very low MSE value of 0.1107043 signifies that the SST value can be used to predict the number of tornadoes occuring at a specific place at a specific time.

A p-value of 0 indicates that we reject the null hypothesis at 5% level of significance. So, we can conclude that the final model fits the data well.

Conclusion

The overall analysis shows that SST values can be considered as reasonable covariate to find out tornado counts in mid-Missouri in the presence of a few other random components as there seems to quite high dependence among them. I did not come across any limitations while performing data analysis on the specific dataset with the given methods and found the data sufficient.