Random coefficient modeling with Prevalence of Fish species in Missouri

Report prepared by

Xiaoyu Ma

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts in Statistics

Under the direction of

Lori Thombs, Ph.D.
Department of Statistics
The University of Missouri - Columbia

May 2020

Abstract

In this paper, random coefficient model was fitted to explore the relationship between counts of several fish species and related water quality metrics including temperature, turbidity, phosphorus, nitrogen, etc. Two datasets studied in this paper were aggregated by monthly mean and monthly median respectively. Both random intercept model and random intercept-slope model were fitted to see if there were any water quality random effects. During fitting, *lme4* package in R was used to generate parameter estimates, residuals, etc. For the results, the random intercept model performed well but the random intercept-slope model came out with singular result which is the random effect variance estimates are nearly zero. After conducting the likelihood ratio test which make comparisons between models, the conclusion could be drawn that the random intercept model worked better in explaining the variances of residuals.

1 Introduction

Since 1972, the Water Quality Act Amendments (PL 92-500) has pushed the monitoring of water quality in developing the thresholds for specific contaminants in order to protect the environment (Karr 1981). Karr also stated that monitoring biological communities (i.e. fish) is also important because they are more likely to reflect the levels of water resources systems. As an environmental indicator, the simple measurement of fish species can capture the full complexity of the ecosystem (Whitfield & Elliott 2002). Fortunately, we get the data involved fish species count and water quality metrics in terms of nutrients like nitrogen, phosphorus, etc. from Missouri Department of Conservation between 2006 to 2015. Therefore, drawing inferences and making

predictions on fish counts within the State of Missouri will play an important role on protecting environment, formulating ecological policies and creating connections for people and wildlife. In order to study the species-specific effects, random intercepts and slopes would be added to the linear model. The structure of the models will be introduced in section 3.

2 Data and Summary Statistics

The original dataset was collected mainly in summer and fall between 2006 to 2015 from different conservation areas in Missouri by the Missouri Department of Conservation (MDC) which include 13528 observations and 34 features. There are 158 fish species recorded in total and criteria of more than 500 show-ups was used to select representative species in the water system which are BLUEGILL, GREEN SUNFISH, CENTRAL STONEROLLER and CREEK CHUB.

The original dataset would not be used because there are hundreds of duplicates, typos and collection errors made by people in it. The data cleaning procedure like removing wrong information and missing values was performed at the start. Besides the man-made errors in the original dataset, the other reason why aggregated datasets would be analyzed is that significant p-values would definitely show up because of the large degrees of freedom coming from large sample size in original dataset. In this case, R-square will be relatively small which indicates the poor fit of models. Two datasets were aggregated by monthly mean and monthly median respectively. After removing outliers and uninterested predictors, the mean dataset has 132 observations and median dataset has 129 observations. 8 features in both datasets are:

mon: Month of data collection.

commonname: Common name for the fish species.

count: Fish counts standardized by efforts.

temp: Temperature in Celcius.

Turb: Turbidity in Nephelometric Turbidity Units.

TP: Total Phosphorus (ug/L).

TN: Total Nitrogen (mg/L).

logc: Log of Counts.

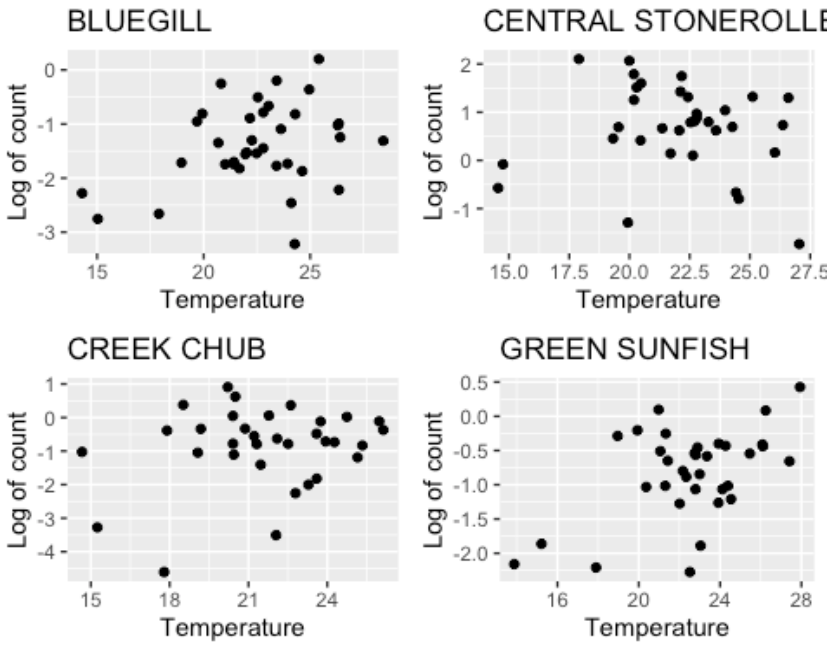


Figure 2.1.1: Summary Statistics for Mean Dataset with respect to 4 species

The counts in the dataset were standardized by the unit effort. The data collectors were using electrofishing and the operation time would be recorded during collecting data. For example, if 10 bluegills were caught for 6 minutes, the count would be

$$count = \frac{10\ bluegills}{\sqrt{360\ seconds}} = 0.5270$$

Logarithmic transformations were applied to the fish counts to make the dependent variable less skewed. Among the predictors, temperature is the main concern of researchers because it's one of the easy-to-collect predictors in reality and the linear trends would become apparent in the ordinary sense. Figure 2.1.1 and Figure 2.1.2 illustrated the summary statistics of log of counts against temperature for mean and median data accordingly. From the figures, relationships between log of counts and temperature are different with respect to species, but linear trends are clearly identified. So, we can formulate some species-specific effects in the modeling.

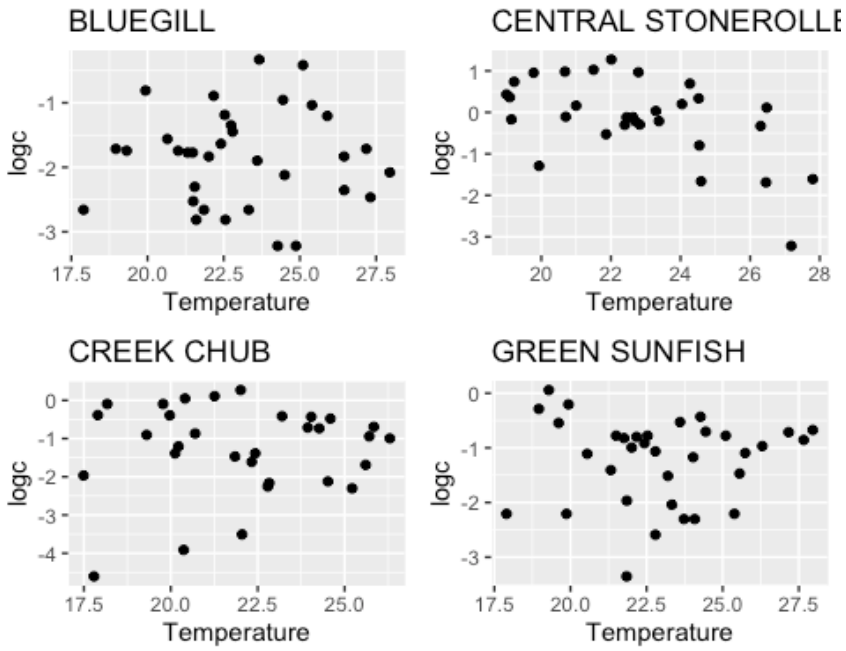


Figure 2.1.2: Summary Statistics for Median Dataset with respect to 4 species

3 Statistical Analysis

The techniques used in this part were linear regression model and random coefficient model, and the random coefficient model would be introduced in a multilevel

way. We start the analysis part from simplest form with only one covariate, temperature. This linear model is mainly for exploring the linear relationship between log of count and temperature. The form of the model is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \cdot (TEMP_i) + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2) \end{aligned} \tag{1}$$

which i is from 1 to n, n is the sample size of the dataset. β_0 is the intercept term, β_1 is the coefficient for temperature and σ^2 is the variance of error term ϵ_i.

There are total four fish species in the dataset. If we believe the differences between species will affect the level of counts, we may add a random intercept to the model. The multilevel model will be developed (Raudenbush and Bryk 2002),

$$\begin{aligned} &Level\ 1: \\ &\qquad Y_{ij} = \beta_{0j} + \beta_{1j} \cdot (TEMP_{ij}) + \epsilon_{ij} \\ &\qquad \epsilon_{ij} \sim N(0, \sigma^2) \\ &Level\ 2: \\ &\qquad \beta_{0j} = \gamma_{00} + u_{0j} \\ &\qquad u_{0j} \sim N(0, \tau_{00}^2) \end{aligned} \tag{2}$$

where $i = 1, \ldots, n_j$, n_j is the sample size of j^{th} species.

3.1 Mean Data

The linear model was fitted to the mean data and the parameter estimates are shown in Table 3.1.1. The intercept estimate is -2.1164 ($p = 0.0060$), the coefficient estimate for temperature is 0.0647 ($p = 0.0585$). In this linear model, the intercept estimate is significant and coefficient estimate of temperature is not significant but

0.0585 is a relatively small p-value. Normality assumption of residuals was also checked through checking the Q-Q plot and histogram plot of the residuals (see Appendix A.1). But the Adjusted R-squared of this linear model is low ($Adjusted\ R^2 = 0.0198$). We will make a step further to multi-level model to check if random effects were needed.

Table 3.1.1: *Linear Model Estimates of mean data*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.1164	0.7583	-2.7909	0.0060
temp	0.0647	0.0339	1.9089	0.0585

Random intercept Model

The model in the form of equation (2) was fitted by maximum likelihood method with t-test performed using Satterthwaite's method (Bates et al. 2015). For the fixed effects in the random intercept model, intercept estimate is -2.1764 ($p = 0.0028$), coefficient estimate is 0.0683 ($p = 0.0081$). Both estimates were significant if the 0.05 rule was applied, but usually people would not care about the fixed effect estimates in this multilevel model. Because only random effects can realize the subject-specific effects, fixed effects are more like baseline estimates in the model and not necessary to be significant. The random effect estimates were in Table 3.1.2. The variance covariance estimates corresponding to the parameterization in the form of equation (2) was:

$$
\begin{aligned}
&Level\ 1: \\
&\qquad Y_{ij} = \hat{\beta}_{0j} + 0.0683 \cdot (TEMP_{ij}) + \hat{\epsilon}_{ij} \\
&\qquad \hat{\epsilon}_{ij} \sim N(0, 0.6994) \\
&Level\ 2: \\
&\qquad \hat{\beta}_{0j} = -2.1764 + \hat{u}_{0j} \\
&\qquad \hat{u}_{0j} \sim N(0, 0.5482)
\end{aligned}
$$

where $\hat{\gamma}_{00} = -2.1764$, $\hat{\beta}_{1j} = 0.0683$, $\hat{\sigma}^2 = 0.6994$ and $\hat{\tau}^2_{00} = 0.5482$.

An ANOVA-like table for random effects (Kuznetsova et al. 2015) was conducted to test the null hypothesis of no random effect existed. The mechanism of this ANOVA-like table is performing likelihood ratio test on the models with or without random effects. The result is a p-value of less than 0.0001 which indicates the random intercept is needed in this model.

Table 3.1.2: *Random effects Estimates for mean data (random intercept only)*

grp	var1	var2	vcov	sdcor
commonname	(Intercept)	NA	0.5482	0.7404
Residual	NA	NA	0.6994	0.8363

Fitted lines of random intercept model are shown in Figure. The more complex model was fitted which includes all four predictors and only total nitrogen is significant ($p = 0.0356$).

Random intercept and slope Model

Once people believed that different species show different behaviors with respect to the temperature levels, we may add random slope to temperature and the model will be (Raudenbush and Bryk 2002),

$$
\begin{aligned}
&Level\ 1: \\
&\qquad Y_{ij} = \beta_{0j} + \beta_{1j} \cdot (TEMP_{ij}) + \epsilon_{ij} \\
&\qquad \epsilon_{ij} \sim N(0, \sigma^2) \\
&Level\ 2: \\
&\qquad \beta_{0j} = \gamma_{00} + u_{0j} \\
&\qquad \beta_{1j} = \gamma_{10} + u_{1j} \\
&\qquad \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00}^2 & \tau_{01}^2 \\ \tau_{10}^2 & \tau_{11}^2 \end{pmatrix} \right]
\end{aligned}
\tag{3}
$$

where τ_{11}^2 is the variance parameter for temperature random effect, $\tau_{01}^2 = \tau_{10}^2$ are the variance-covariance parameters and τ_{00}^2 is the intercept random effect. The correlations between intercept and slope were assumed here.

For the fixed effects estimate in the random intercept and slope model, the intercept estimate is -2.1435 ($p = 0.0711$) and the coefficient estimate for temperature is 0.0668 ($p = 0.0546$). Both fixed effects estimates are significant. Table 3.1.3 contains the random effect estimates and it will be shown in the form of equation (3):

$$
\begin{aligned}
&Level\ 1: \\
&\qquad Y_{ij} = \hat{\beta}_{0j} + \hat{\beta}_{1j} \cdot (TEMP_{ij}) + \hat{\epsilon}_{ij} \\
&\qquad \hat{\epsilon}_{ij} \sim N(0, 0.6908) \\
&Level\ 2: \\
&\qquad \hat{\beta}_{0j} = -2.1435 + \hat{u}_{0j} \\
&\qquad \hat{\beta}_{1j} = 0.0668 + \hat{u}_{1j} \\
&\qquad \begin{pmatrix} \hat{u}_{0j} \\ \hat{u}_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2.0739 & -0.0451 \\ -0.0451 & 0.0010 \end{pmatrix} \right]
\end{aligned}
$$

where $\hat{\gamma}_{00} = -2.1435$, $\hat{\gamma}_{10} = 0.0668$

$\hat{\sigma}^2 = 0.6908$, $\hat{\tau}_{00}^2 = 2.0739$, $\hat{\tau}_{01}^2 = \hat{\tau}_{10}^2 = -0.0451$ and $\hat{\tau}_{11}^2 = 0.0010$.

Here the problem of Singularity fitted model arose, i.e. estimated variance-covariance matrix is not full ranked or random effects variance estimates nearly zero (Bates et al. 2015). This problem often appears when the correlations are close to 1 or -1 and potential reasons would be a lack of information or too much variables in the model. The indication of singular fit are overfitting and inappropriateness of testing procedure such as Wald test or Likelihood ratio test. So, ANOVA-like table for random effects would not be performed in the random intercept and slope model. Singular fit problem also exists in the random slope models with other predictors such as turbidity, total

phosphorus and total nitrogen. So, random slope would be considered a poor fit in this dataset.

Table 3.1.3: *Random effects Estimates of mean data (random intercept and slope)*

grp	var1	var2	vcov	sdcor
commonname	(Intercept)	NA	2.0739	1.4401
commonname	temp	NA	0.0010	0.0313
commonname	(Intercept)	temp	-0.0451	-1.0000
Residual	NA	NA	0.6908	0.8311

Residual checking through Q-Q plot and histogram plot for both random intercept model and random intercept and slope model are in Appendix. A.2.

3.2 Median Data

Random intercept model

For median data, we also start with a temperature only model and model in equation (2) was used in this part. The fixed effect estimates are not significant: intercept estimate is -0.2541 ($p = 0.752$) and coefficient estimate is -0.0383 ($p = 0.244$). Table 3.2.1 contains the random effect estimates and if we show them in the form of equation (2):

$$
\begin{aligned}
&Level\ 1: \\
&\qquad Y_{ij} = \hat{\beta}_{0j} - 0.0383 \cdot (TEMP_{ij}) + \hat{\epsilon}_{ij} \\
&\qquad \hat{\epsilon}_{ij} \sim N(0, 0.8542) \\
&Level\ 2: \\
&\qquad \hat{\beta}_{0j} = 0.2541 + \hat{u}_{0j} \\
&\qquad \hat{u}_{0j} \sim N(0, 0.3462)
\end{aligned}
$$

where $\hat{\gamma}_{00} = 0.2541$, $\hat{\beta}_{1j} = -0.0383$, $\hat{\sigma}^2 = 0.8542$ and $\hat{\tau}^2_{00} = 0.3462$.

The result of ANOVA-like table for random effect is p-value less than 0.0001 which stands for random intercept is needed in the model. A model with all four predictors was also fitted but the coefficient estimates for turbidity and total phosphorus are nearly zero.

Table 3.2.1: *Random effect Estimates of median data (random intercept only)*

grp	var1	var2	vcov	sdcor
commonname	(Intercept)	NA	0.3462	0.5884
Residual	NA	NA	0.8542	0.9242

Random intercept and slope model

The model in equation (3) was used in this part. The correlations between intercept and slope were assumed here. The fixed effect estimates: intercept estimate is -0.1475 ($p = 0.925$) and coefficient estimate is -0.0431 ($p = 0.464$). Both estimates are not significant. Table 3.2.2 contains the random effect estimates and it will be shown in the form of equation (3):

$$
\begin{aligned}
&Level\ 1: \\
&\qquad Y_{ij} = \hat{\beta}_{0j} + \hat{\beta}_{1j} \cdot (TEMP_{ij}) + \hat{\epsilon}_{ij} \\
&\qquad \hat{\epsilon}_{ij} \sim N(0, 0.8049) \\
&Level\ 2: \\
&\qquad \hat{\beta}_{0j} = -0.1475 + \hat{u}_{0j} \\
&\qquad \hat{\beta}_{1j} = -0.0431 + \hat{u}_{1j} \\
&\qquad \begin{pmatrix} \hat{u}_{0j} \\ \hat{u}_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 6.6341 & -0.2236 \\ -0.2236 & 0.0075 \end{pmatrix} \right]
\end{aligned}
$$

where $\hat{\gamma}_{00} = -0.1475$, $\hat{\gamma}_{10} = -0.0431$

$\hat{\sigma}^2 = 0.8049$, $\hat{\tau}^2_{00} = 6.6341$, $\hat{\tau}^2_{01} = \hat{\tau}^2_{10} = -0.2236$ and $\hat{\tau}^2_{11} = 0.0075$.

The variance estimate for random slope is 0.0075 here which is also close to 0. The singular fit problem also existed in the random coefficient model for median data as

well as other predictors. ANOVA-like table for random effects would not be performed in this model for the same reason we talked in section 3.1.

Table 3.2.2: *Random effects Estimates of median data (random intercept and slope)*

grp	var1	var2	vcov	sdcor
commonname	(Intercept)	NA	6.6341	2.5757
commonname	temp	NA	0.0075	0.0868
commonname	(Intercept)	temp	-0.2236	-1.0000
Residual	NA	NA	0.8049	0.8971

Residual checking through Q-Q plot and histogram plot for both random intercept model and random intercept and slope model are in Appendix. B.

4 Conclusion and Ideas for Further Research

From the result in section 3.1 and 3.2, the random intercept model works well in both the mean and median datasets since the ANOVA-like table for random effect suggests significant random intercept. For better visualization, the plots of fitted lines for the random intercept model are attached in Appendix. C. For the random slope model, the singular fit problem (overfitting) always shows up no matter what predictors included. We can conclude that the random intercept model works better in terms of assigning random effects in the aggregated fish counts data analysis.

In the original dataset, there is more than 30 variables including spatial variable such as name of water system and coordinates. Spatial statistics analysis would be a topic worth exploring in the further research. Time variables such as dates of data collection are also in the original dataset. Constructing longitudinal study like including t, t^2 in the model is also one of the hot areas today.

The datasets used and R script used in this paper are attached in Appendix. D.

References

Karr, J. R. (1981). “Assessment of biotic integrity using fish communities.” Fisheries, 6, 21-27.

Whitfield, A. K. and Elliott, M. (2002). “Fishes as indicators of environmental and ecological changes within estuaries: a review of progress and some suggestion for the future.” Journal of Fish Biology, 61(Supplement A), 229-250.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

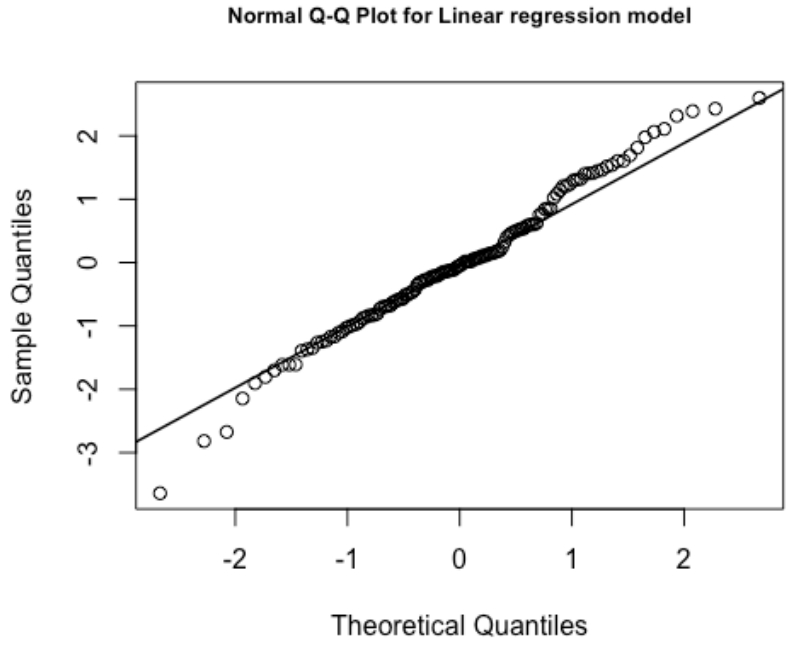
Kuznetsova, A., Brockhoff P. B., Christensen R. H. B. (2017). “lmerTest Package: Tests in Linear Mixed Effects Models.” *Journal of Statistical Software*, 82(13), 1-26. doi:10.18637/jss.v082.i13.

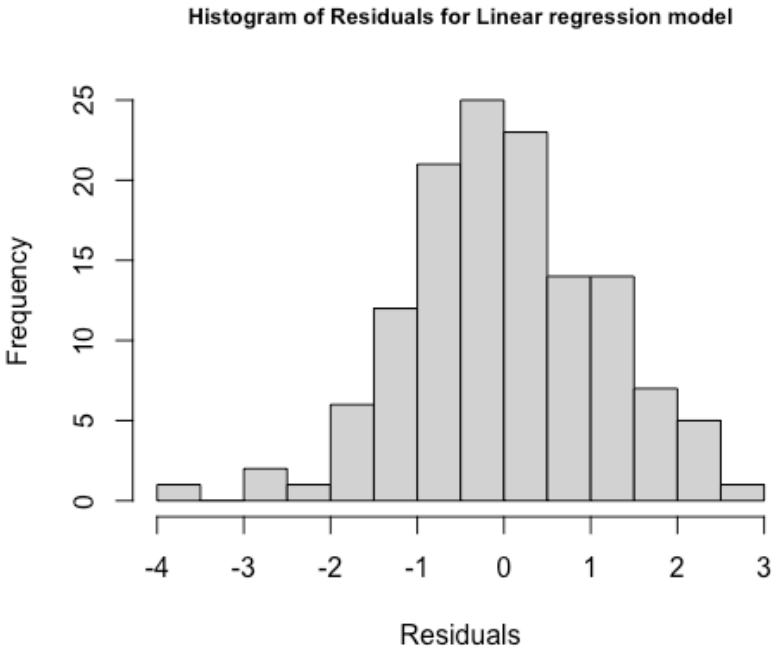
Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods 2ed*. Sage Publications, p. 75-85

Appendix

A.1

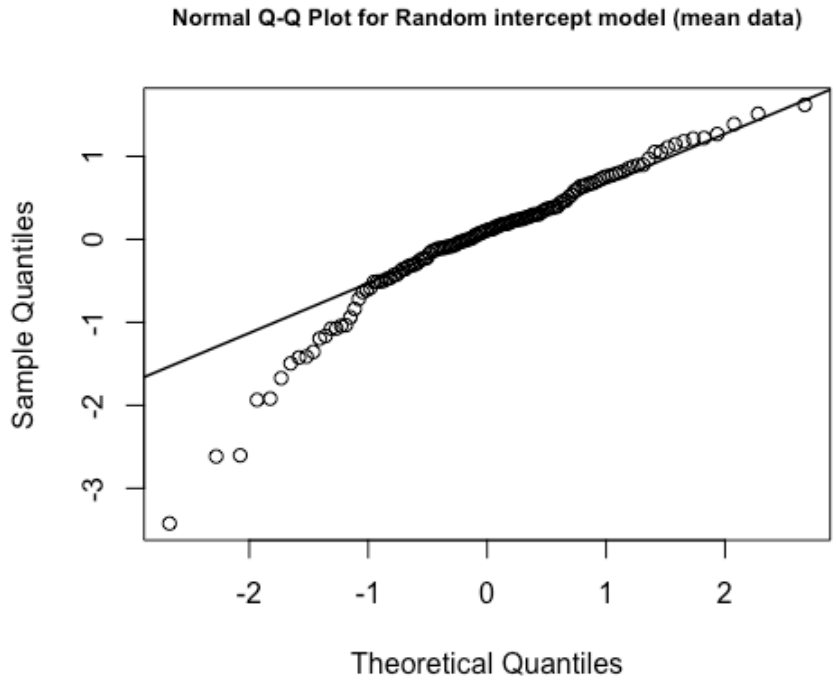
Residuals checking for linear regression model in section 3.1.

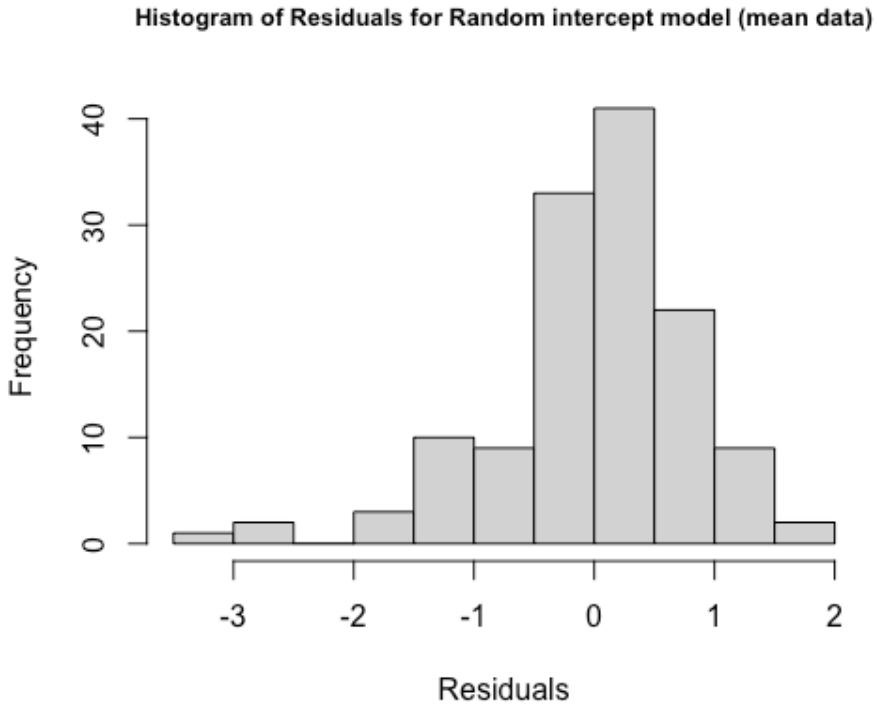




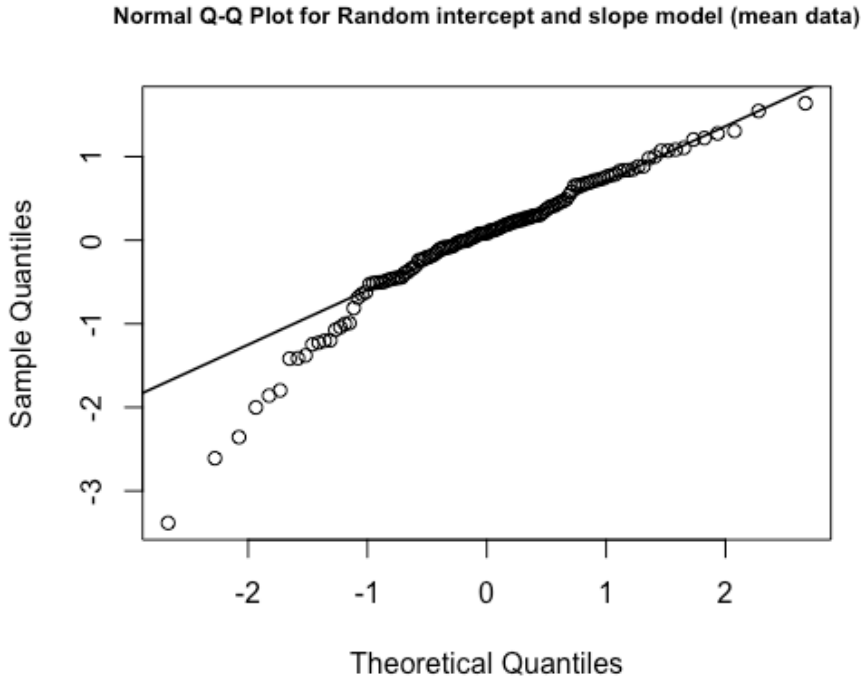
A.2

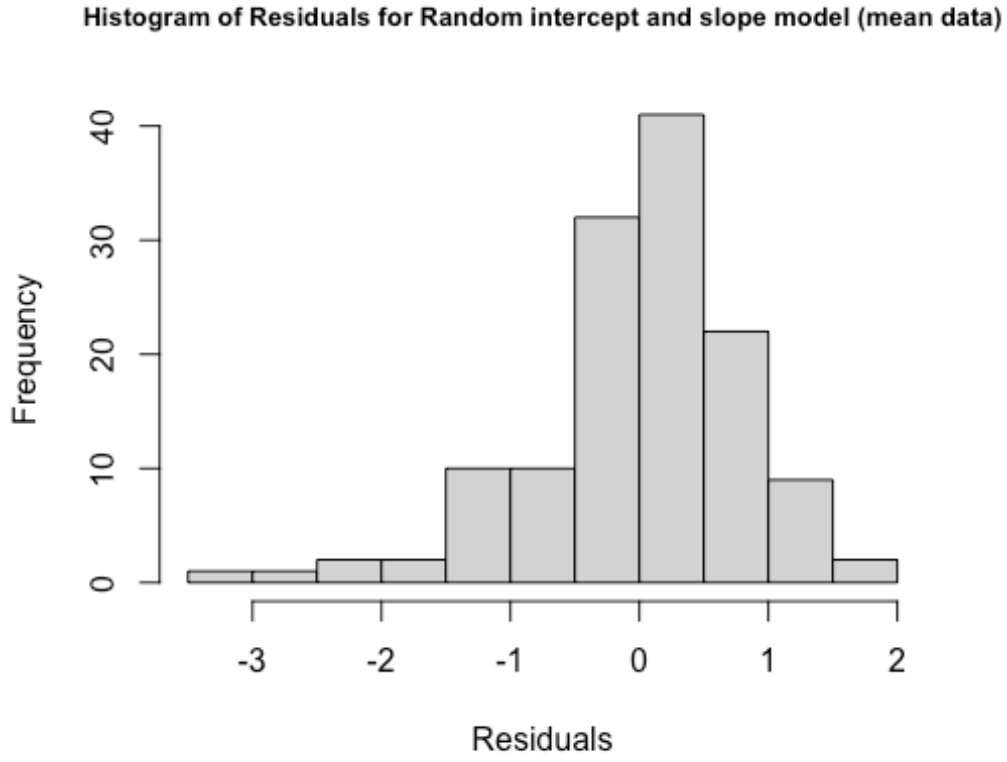
Residuals checking for the random intercept model in section 3.1.





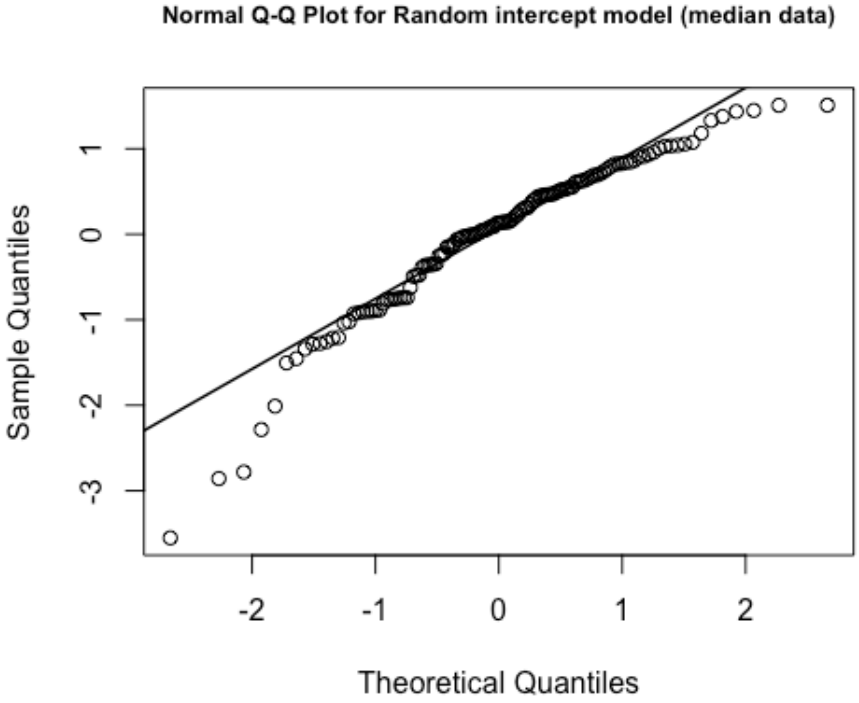
Residuals checking for the random intercept and slope model in section 3.1.

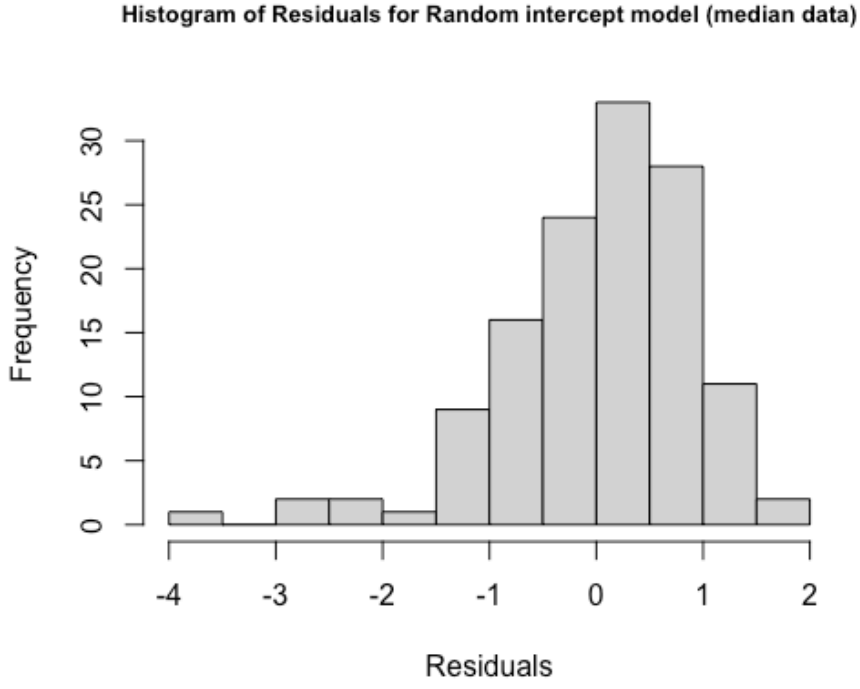




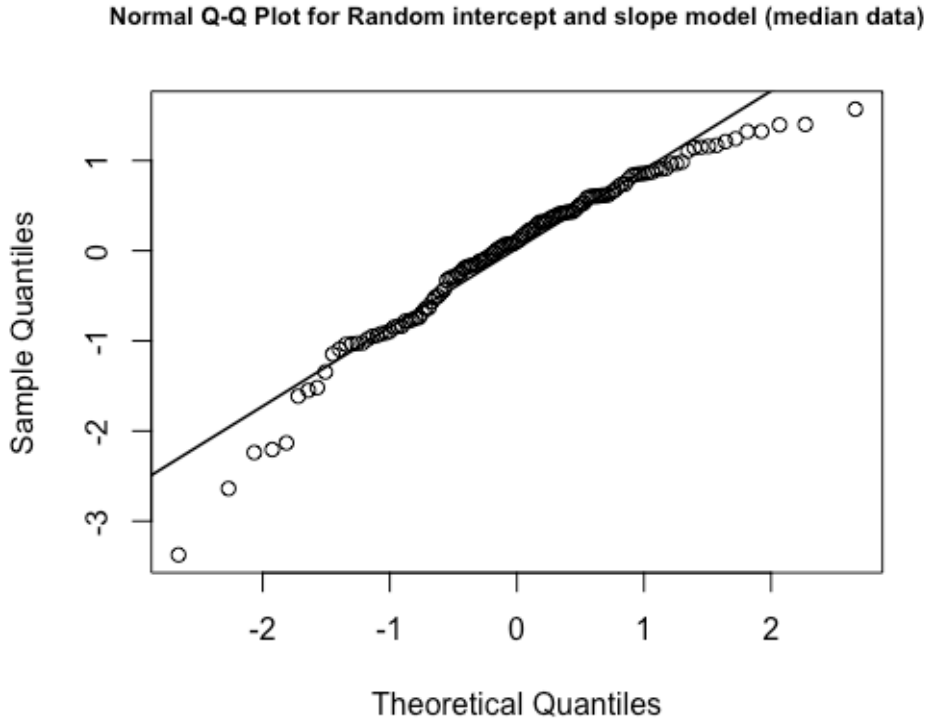
B

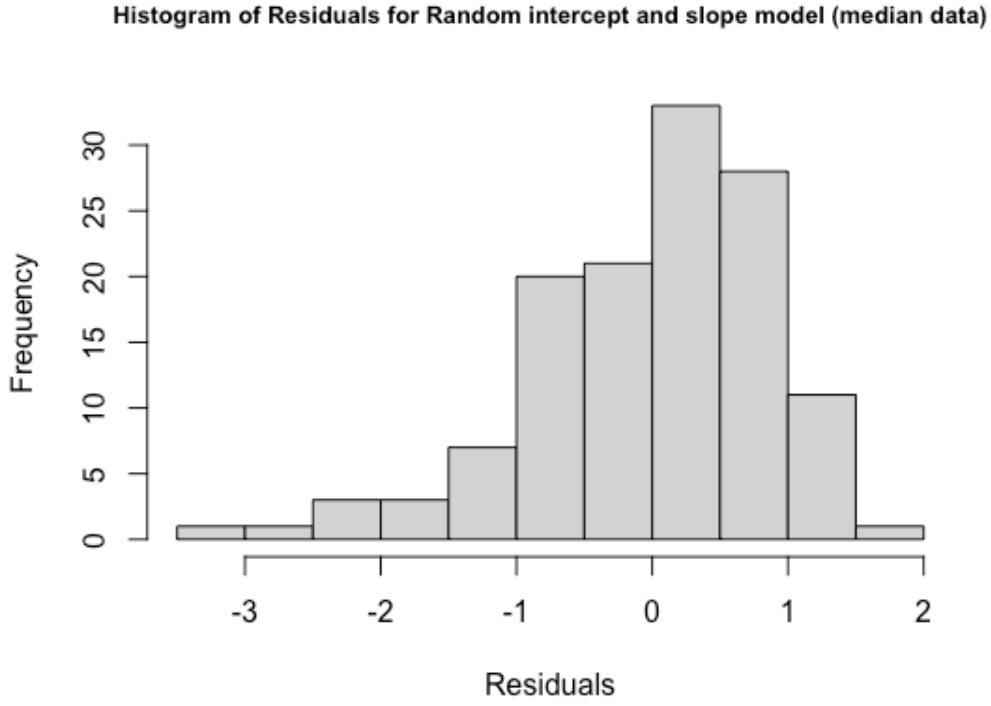
Residuals checking for the random intercept model in section 3.2.





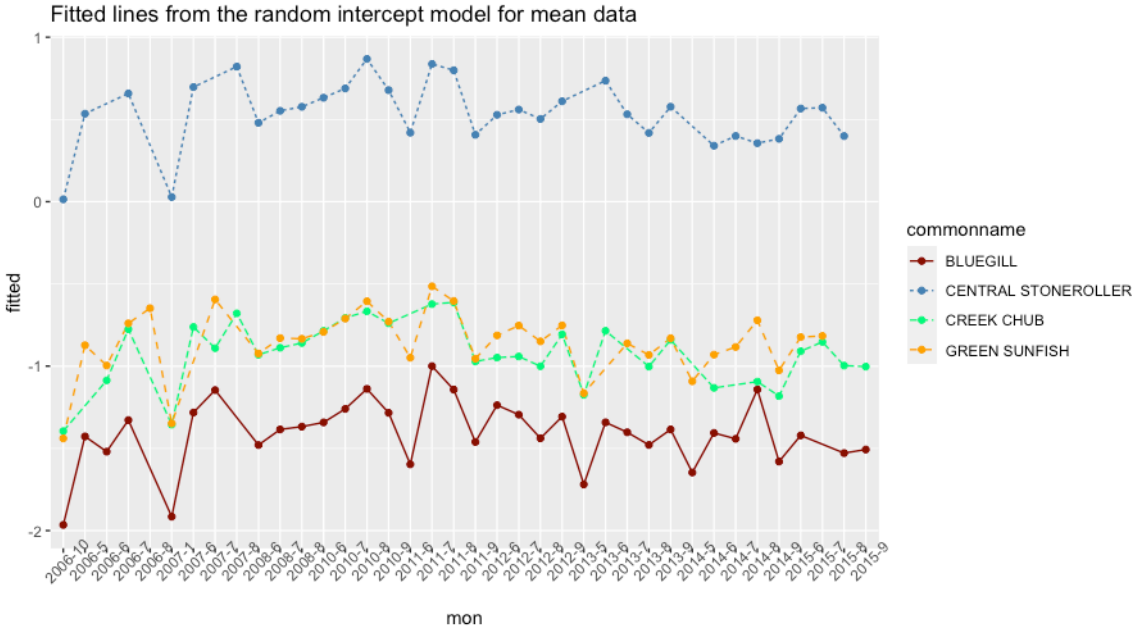
Residuals checking for the random intercept and slope model in section 3.2.

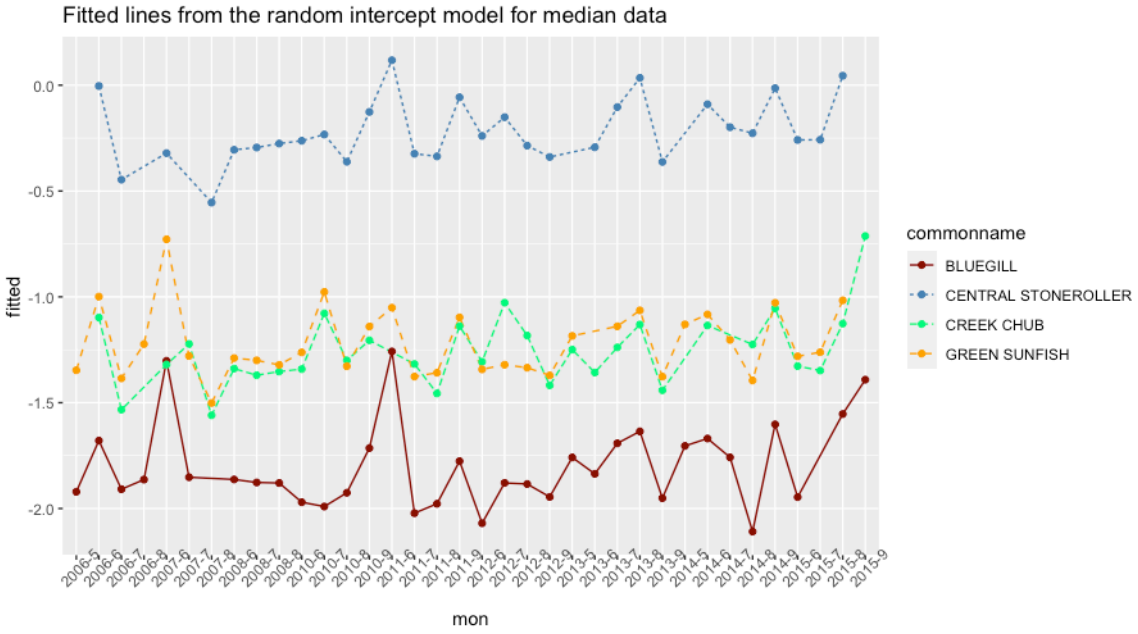




C

The fitted lines from the random intercept models mentioned in section 4.





D

The datasets and R script:

https://github.com/xiaoyuace/Stat8100-Paper