# case1final

*Nathaniel Brown, Annie Tang, William Yang*

*September 13, 2017*

## Introduction

In this report, we analyze data from an international validation study to measure the effectiveness of the rat uterotrophic bioassay. The bioassay being studied attempts to measure the estrogenic effect of certain chemicals. The two chemicals used in this study have well known effects, so we would like to verify that the bioassay produces consistent results in rats that have been administered various dosages of the chemicals across two possible protocols. If the uterotrophic bioassay is an effective procedure for measuring the effects of these chemicals, then we expect to see consistent responses to various dosages across all labs and groupings.

To measure the consistency of the responses, we fit iterations of a linear mixed effects model on the provided dataset. The model is conditioned on the labs to account for lab-to-lab variability. Each iteration of the model differs slightly in transformations and conditioning of the predictors/responses until we arrive at a model that we deem most appropriate for the dataset. We then evaluate the final model to determine whether it predicts large variation in responses to the dosages between labs, or if it predicts a consistent measured response to the chemicals.

## Model-Fitting

$$y_{ij} \sim \beta_{0,i}+\beta_{1,i}x_{ij,d_1}+\beta_{2,i}x_{ij,d_2}+\beta_{3,i}x_{ij,p_B}+\beta_{4,i}x_{ij,p_C}+\beta_{5,i}x_{ij,p_D}+\beta_{6,i}x_{ij,log(w)}+\epsilon\beta_{0:5,i} \sim N(\mu_{0:5,i}, \sigma^2_{0:5,i})\epsilon \sim N(0,\sigma^2)$$

To fit the data, we used a mixed model, with fixed and random effects. Let $y_{ij}$ be the observed log(blotted uterus weight) for subject $x_{ij}$, the $j$th individual in lab $i$. $x_{ij,d_1}$ and $x_{ij,d_2}$ are the values of dose1 and dose2 for subject $x_{ij}$. $x_{ij,p_B}$, $x_{ij,p_C}$, and $x_{ij,p_D}$ are dummy variables indicating which protocol $x_{ij}$ was subjected to. $x_{ij,log(w)}$ is the log(body weight) for $x_{ij}$. Body weight and uterus weight are log-transformed to account for the right skew in the data (Fig. 1). We make the Gaussian assumption that the coefficients, $\beta$, are normally distributed according to some $\mu_i$ and $\sigma_i$. We add a random effect on all $\beta_{0:5,i}$ to account for the lab-to-lab variability in the intercepts and slopes of the blotted weight (Fig. 2). The summary statistics of the model described above can be found on Table 1.

We start at a reduced form of this model and augment it to its full form after initial analyses.

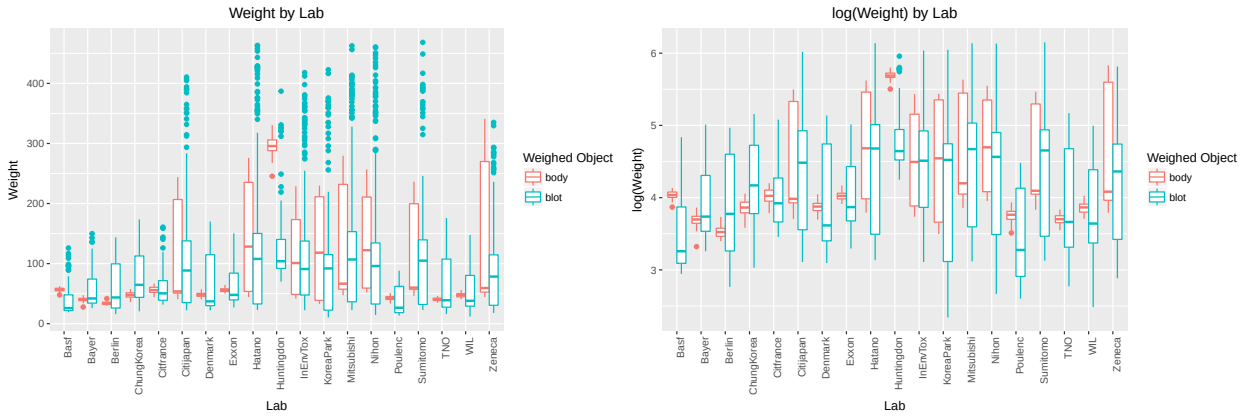



Figure 1: Log Transformations of Weights

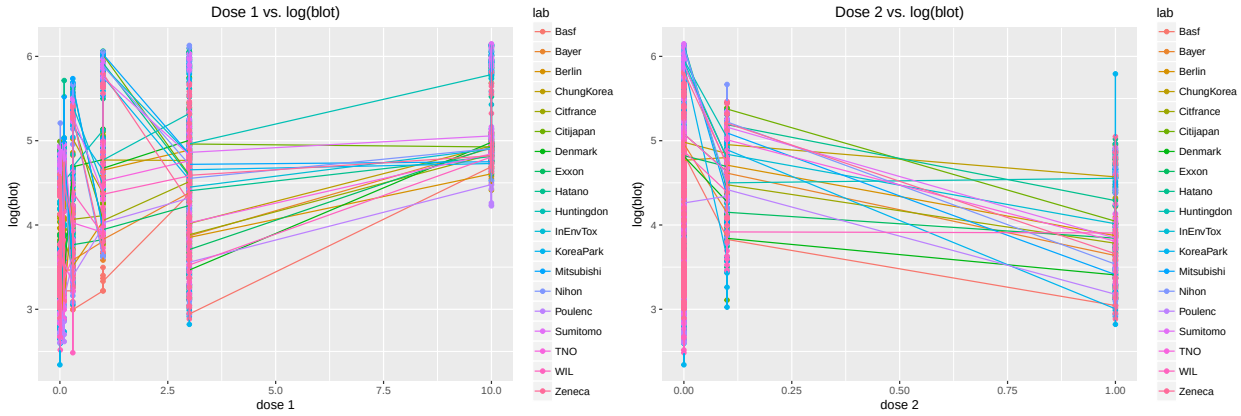


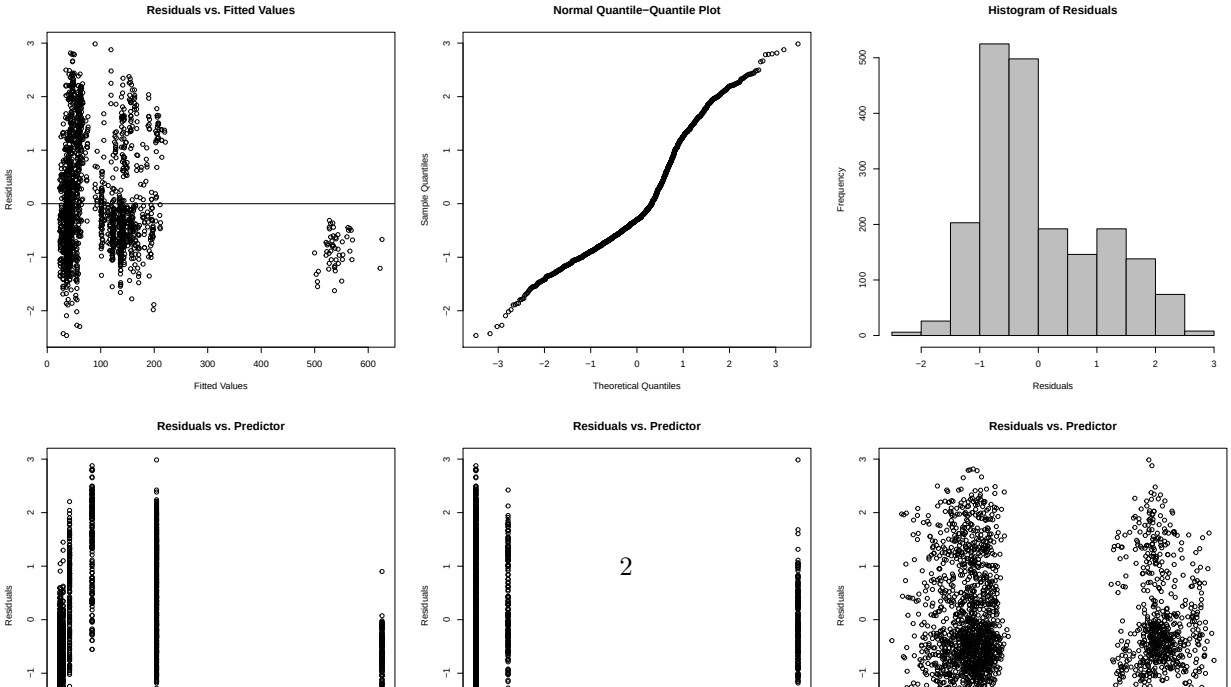

Figure 2: Lab-to-Lab Variability in Intercepts and Slopes

| Mean Fixed Eff | ects Variance Fixed | Effects |
|---|---|---|
| (Intercept) | 4.4197 | 0.1456 |
| protoB | 0.0339 | 0.0007 |
| protoC | 1.1603 | 0.0201 |
| protoD | 1.1497 | 0.0210 |
| log(body) | 0.0581 | 0.0095 |

Table 2: Summary Statistics of Initial Model

| | Mean Random Effects | Variance Random Effects |
|---|---|---|
| (Intercept) | 2.4259 | 1.1118 |
| dose1 | 0.0300 | 0.0225 |
| dose2 | 0.3259 | 0.2707 |

| | Residual Variance |
|---|---|
| residual | 0.1889 |

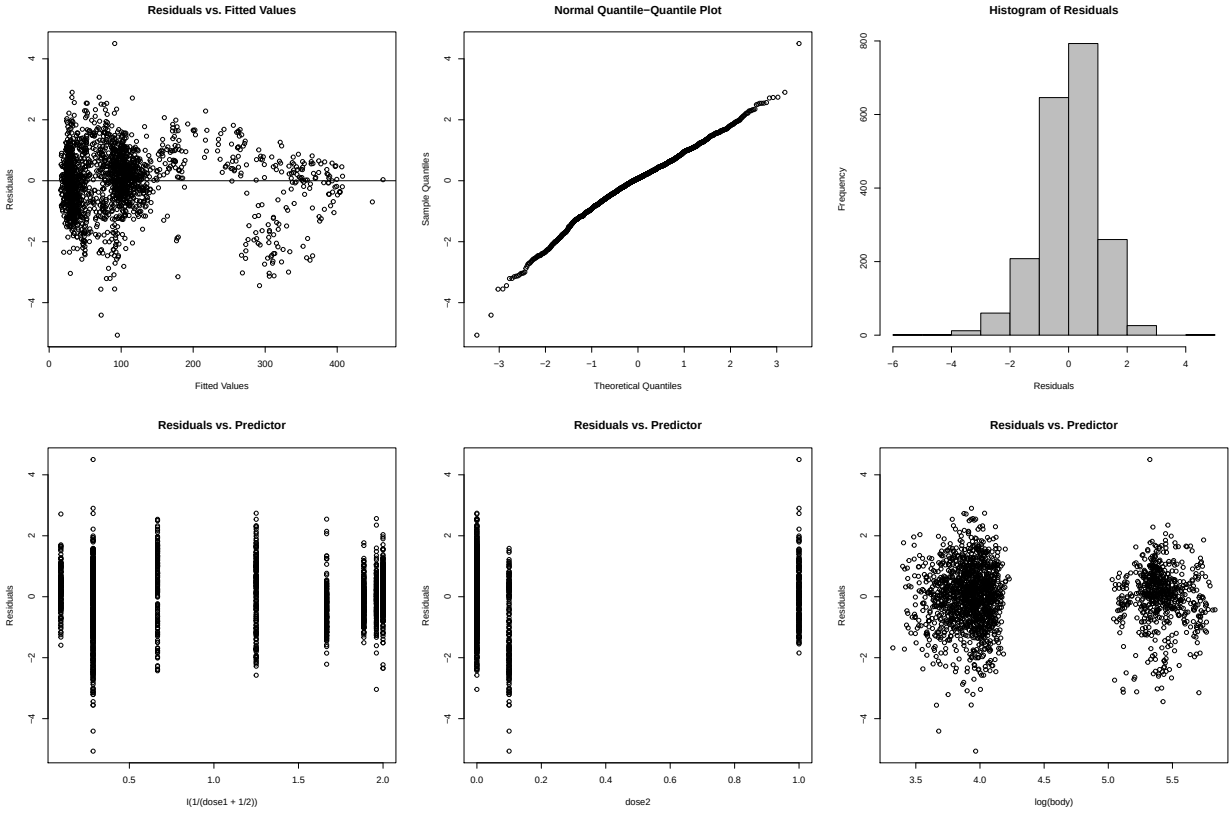


penalizes more for extreme errors, while Mean Absolute Error simply averages all of the errors. The results are in the following tables and figures:

| | Mean Fixed Effects | Variance Fixed Effects |
|---|---|---|
| (Intercept) | 2.4881 | 0.0748 |
| protoB | 0.0473 | 0.0003 |
| protoC | 0.8116 | 0.0099 |
| protoD | 0.7988 | 0.0103 |
| log(body) | 0.3217 | 0.0047 |

| | Mean Random Effects | Variance Random Effects |
|---|---|---|
| (Intercept) | 3.5663 | 1.0303 |
| I(1/(dose1 + 1/2)) | 0.4745 | 0.5324 |
| dose2 | 0.7014 | 1.0401 |

| | Residual Variance |
|---|---|
| residual | 0.081 |



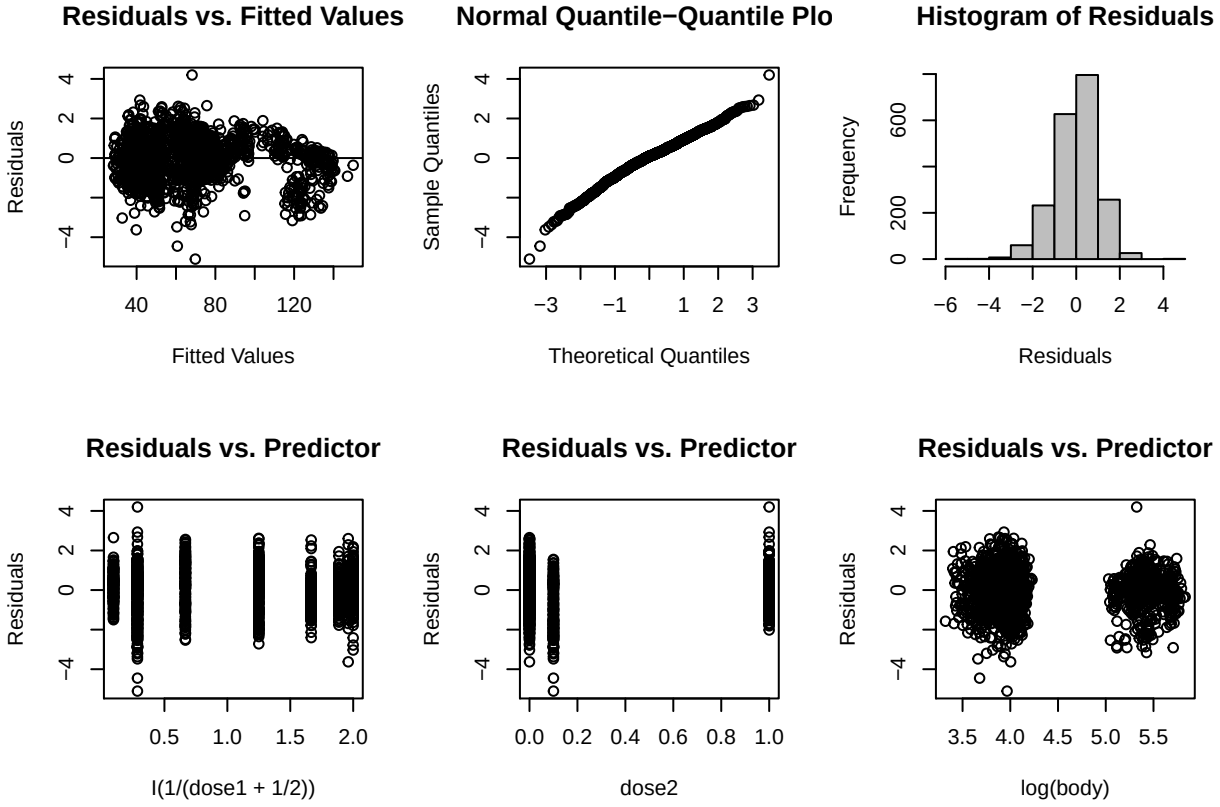

| | MAE | RMSE |
|---|---|---|
| Transformed Dose 1 | 23.3791 | 39.7445 |
| Initial | 41.6219 | 69.3082 |

**comment comment comment**. Another problem with these residual plots is the slight arching in the Residuals vs. Fitted Values. To counteract this, we added a square root transformation to the response. **well then again we might not*** The results are below:

| | Mean Fixed Effects | Variance Fixed Effects |
|---|---|---|
| (Intercept) | 1.5752 | 0.0046 |
| protoB | 0.0095 | 0.0000 |
| protoC | 0.1678 | 0.0006 |
| protoD | 0.1625 | 0.0006 |
| log(body) | 0.0964 | 0.0003 |

| | Mean Random Effects | Variance Random Effects |
|---|---|---|
| (Intercept) | 3.2542 | 0.0526 |
| I(1/(dose1 + 1/2)) | 0.5440 | 0.0323 |
| dose2 | 0.7428 | 0.0608 |

| | Residual Variance |
|---|---|
| residual | 0.005 |



| | MAE | RMSE |
|---|---|---|

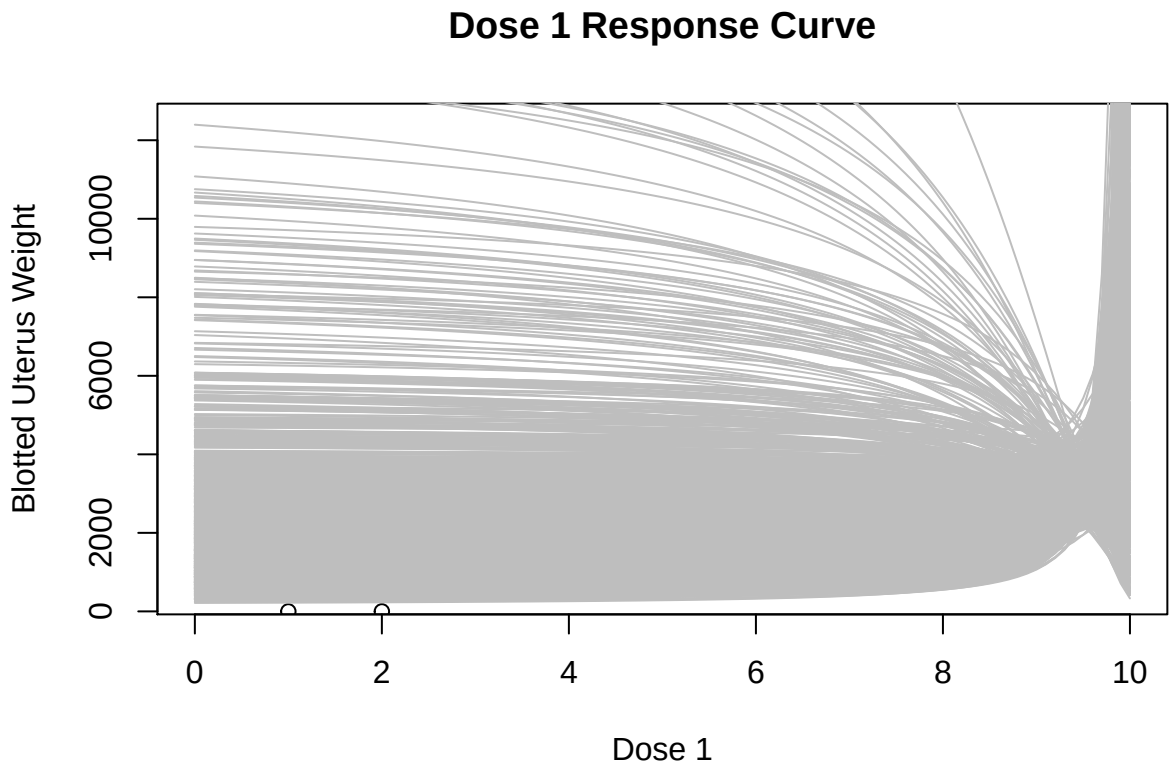
| | MAE | RMSE |
|---|---|---|
| Transformed Response and Dose 1 | 49.4019 | 84.2722 |
| Transformed Dose 1 | 23.3791 | 39.7445 |

This is our best-fitting mixed model we could fit. **maybe. it looks like sqrt log blot improves residual plot but worsens prediction**

**Results**

# INCOMPLETE

The plot below samples random dose1 effects (slopes) and lab effects (intercepts) from our final mixed model, holding all other predictors constant, and plots the resulting dose effect curve.




**Conclusion**

The variance in the dose1 effects by lab is noticeably high, so we say that the bioassay does depend on lab and thus this study fails miserably.

This explodes around 10 and it's only applicable for dose less than or equal to 10.

**Contributions**

Nathaniel Brown made the visualizations for this report. He also organized the relevant files in a Github repository for the group to access and edit. Annie Tang compiled the group work done on EDA into a .rmd and wrote the accompanying explanations for the EDA and approaches to analysis. William Yang helped pair on EDA analysis and identify approaches to handle the data. Approaches to analysis were a joint effort by all members of the group. Nathanial implemented analysis for the univariate normal and multivariate normal approaches. Implementation and analysis of the mixed effects model was a joint effort by all members of the group.