Case 1 Take 4

Ian Hua, InHee Ho, Sonia Xu

September 25, 2017

Introduction

The following data is from an international validation study (19 labs, 6 nations) of the rat uterotrophic bioassay. The purpose of the assay was to screen chemicals for estrogenic effects--agonist (EE) or antagonist (ZM). One clear challenge from the assay was the consistency among variables and protocols throughout all data in the study. The motivation of this study is to fit the best model against the data to identify potential differences between labs, and to assess the consistency and effectiveness of the two chemicals. An original linear mixed effects model with general transformations was fit to the data; however, further exploration showed that a linear mixed effects model with kernel estimates and clustering for certain features provided better results.

Methodology

Our approach to answer our problem followed these 4 steps:

1. Data-cleaning
2. Exploratory Data Analysis
3. Fitting the data to a Linear Mixed-Effect Model
4. Interpreting the data, model, and results

Dataset

The final dataset contains 2677 observations and 6 features: lab, protocol, dose 1, dose 2, body weight, and blotted uterus weight. Three key pieces of data were removed: 4 observations due to missing data, the feature, group, due to multicollinearity, and the feature, wet, due to its variability in capturing the uterus weight. Below is a preview of the final dataset:

lab	proto	group	dose1	dose2	body	wet	blot
Berlin	A	1	0	0	31.9	21.2	17.2
Berlin	A	1	0	0	31.1	22.0	20.5
Berlin	A	1	0	0	32.9	23.5	20.1

Berlin	A	1	0	0	31.6	17.3	16.5
Berlin	A	1	0	0	33.2	18.8	15.9
Berlin	A	1	0	0	30.6	34.0	26.1

Final Model: Linear Mixed Effect Model

Transitioning from the Linear Mixed Effects Model

In our previous report, we attempted to handle the non-linearity of dose 1 and dose 2 by taking their logarithms and modeling as polynomials with a degree of 2. Overall, our previous final model fitted the data well according to the qq-plot and residual vs. fitted graph. However, using second-degree polynomials of the logarithmized values of dose 1 and dose 2 did not fully capture the uneven distribution of dose 1 and dose 2 against the response, blotted uterus weight. We can better fit the model using kernel regression and clustering. We present below our updated final model that applies kernel estimations to dose 1 and dose 2 and transforms body weight into a categorical variable via k-means clustering, and how this model better predicts our response.

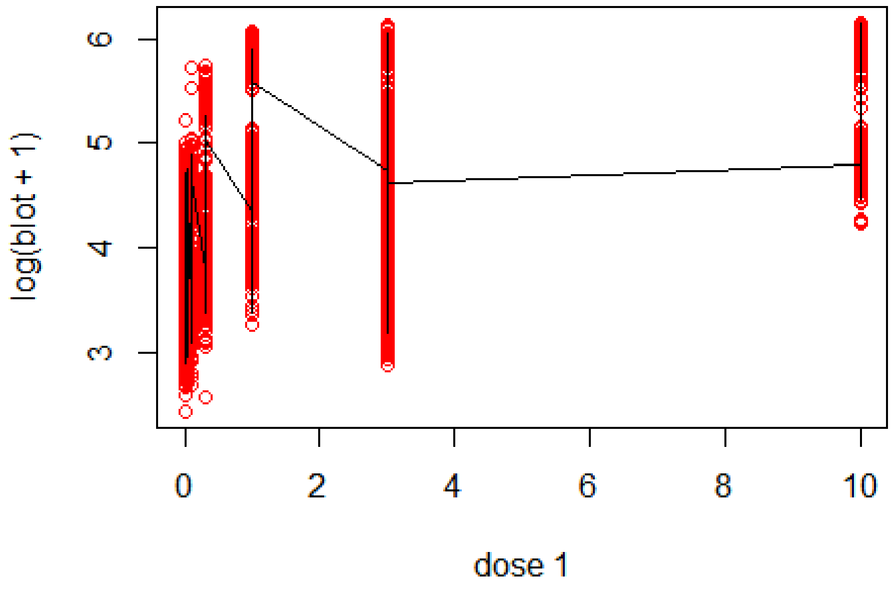
Exploratory Data Analysis

We continued from our previous report with our choices of explanatory variables. We detected significant heteroscedasticity in our response variable (blotted uterus weight) between labs, and took their logarithms for more consistent variance. Furthermore, protocols can have different effects on the response depending on the labs. We capture such difference by fitting random effects proto|lab. Finally, to capture the effects of doses of chemical 1 and 2 across all labs as well as their variation between labs, we added dose 1 and dose 2 as both fixed and random effects. However, we used different methods to better capture the distributions of dose 1, dose 2, and body weight.

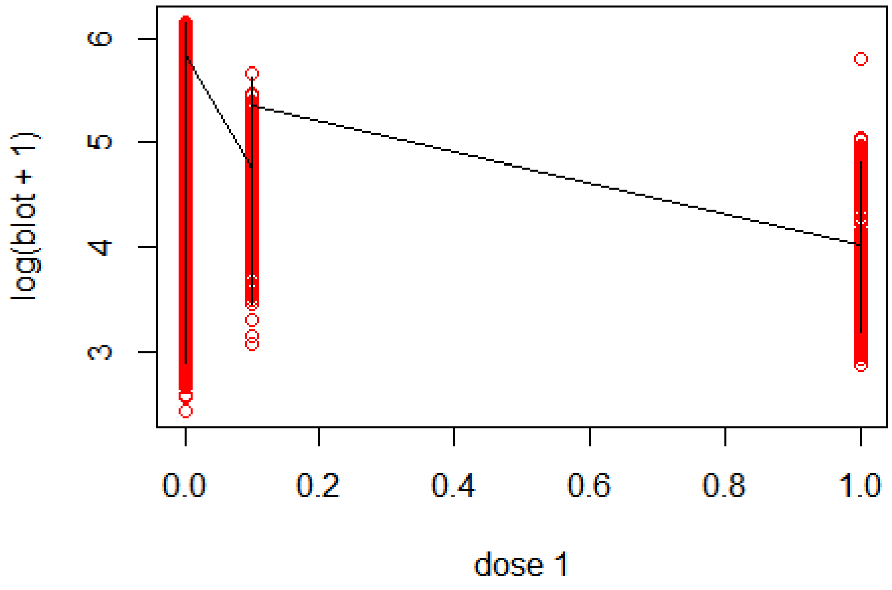
Capturing the Non-Linearity of Dose 1 and Dose 2

In order to capture the non-linearity of the distribution of dose 1 and dose 2, we decided to use kernel regression instead of modeling with second-degree polynomials of logarithmized values. We use 4 knots for dose 1 and 3 knots for dose 2.

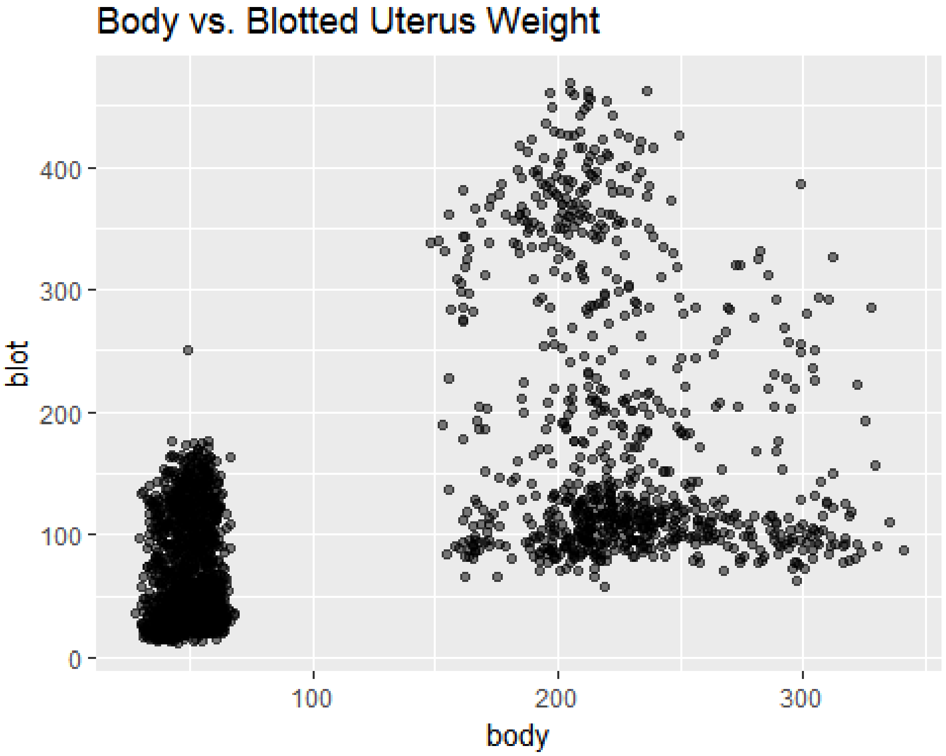
Fitted Kernels for Dose 1



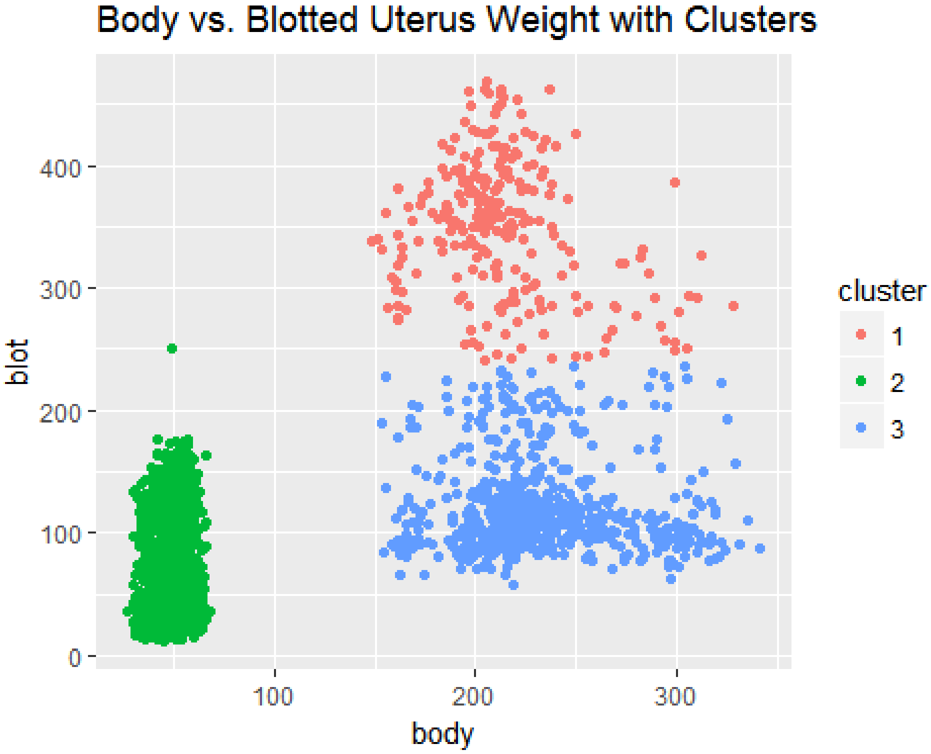
Fitted Kernels for Dose 2



Transforming Body into a Categorical Variable



Looking at the scatterplot above between body and blotted uterus weight, we observe two distinct groups of body weight. These groups are non-linear, so it would be best to distinguish them in an unsupervised fashion. K-means clustering with 3 groups resulted in the best segmentation of the data, and these groups (1,2,3) replaced the continuous variable of body weight for the final model. Below is a plot of the 3 different clusters.

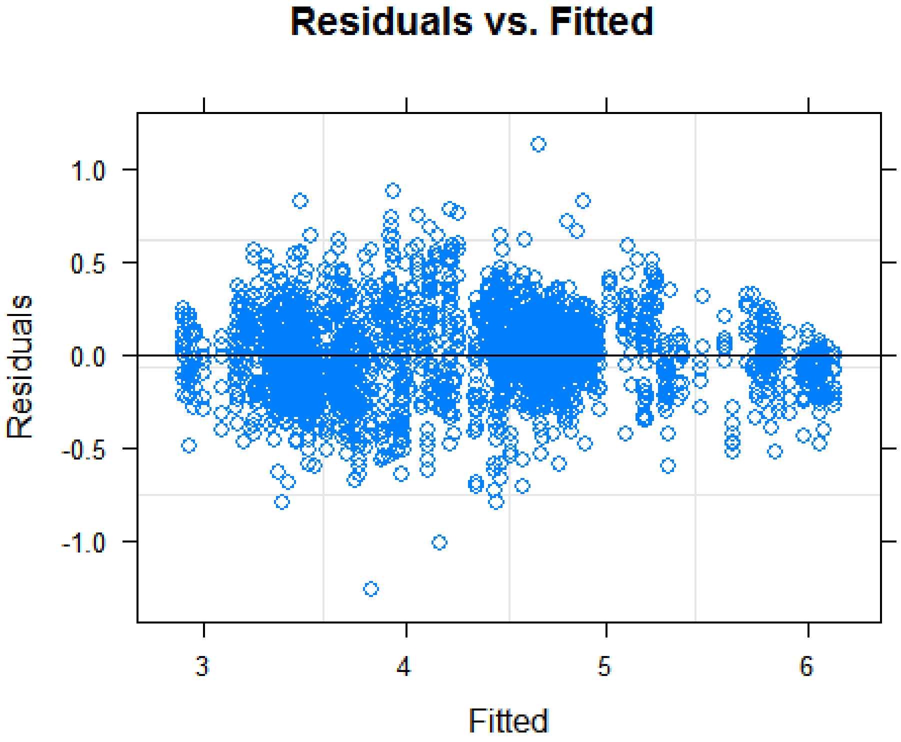


The Updated Final Model

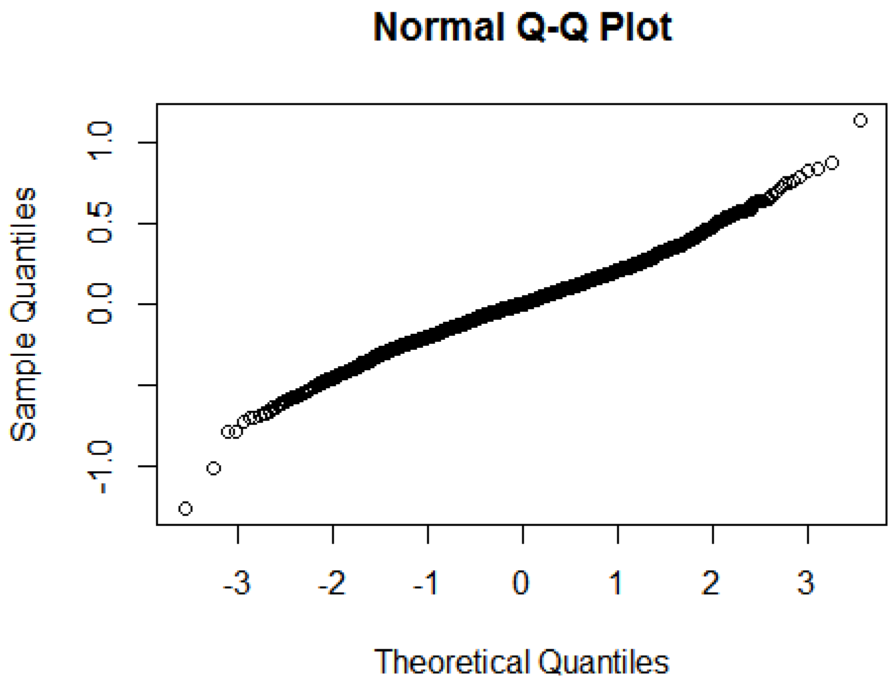
Thus, our updated final model still has the log of blotted uterus weight as the response variable; protocol, dose 1, and dose 2 as random effects between labs; and dose 1, dose 2, and body weight as fixed effects across all population. The difference between the previous and updated final model is that dose 1, dose 2, and body weight now reflect the transformations as noted above.

Checking Assumptions

The updated model satisfies the linearity checks of a linear mixed effect model.



Looking at the Residuals vs. Fitted plot, the points exhibit homoscedascity, and overall, the points are randomly and evenly distributed above and below the horizontal axis. This implies that the data should be modeled linearly.



Checking the normality assumption via the QQ-plot, the points overall follow a straight diagonal trend, which reaffirms that the points are normally distributed.

Goodness of Fit

By fitting dose 1 and dose 2 via kernel regression and clustering body into 3 segments, this updated final model produced better predictive results for the train vs. test data. The previous final model had a root mean squared error (RMSE) of 0.233 after 100 iterations. This final model had a root mean squared error of 0.222 after 100 iterations, so it is a slight improvement from our previous final model.

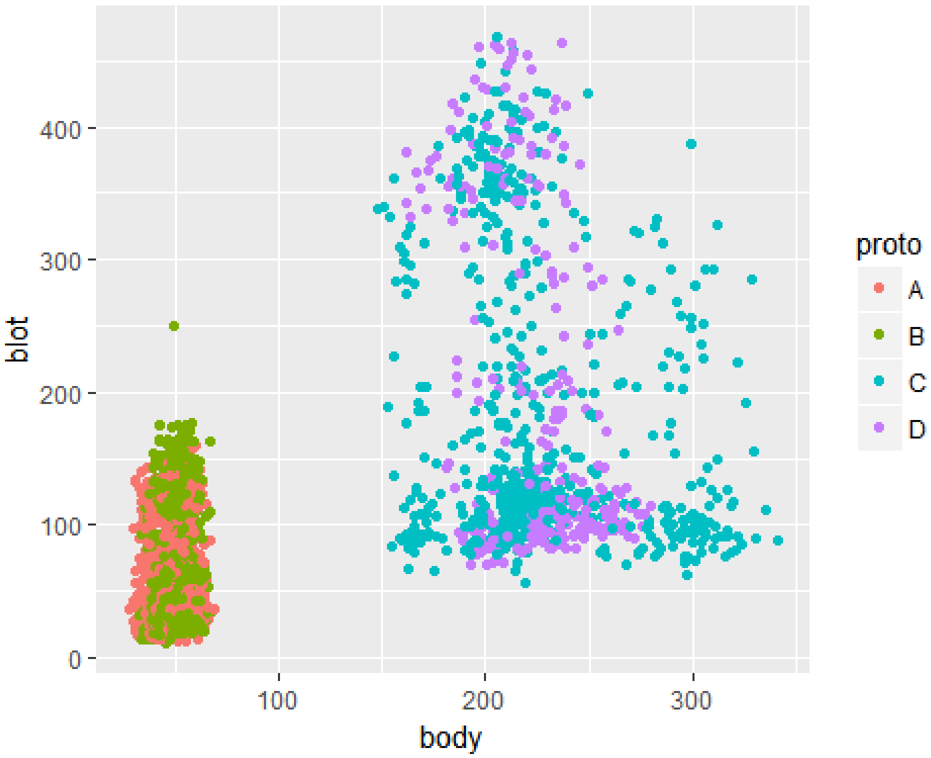
Discussion of Updated Final Model

This updated final model produced similar results to the previous final model.

Differences in the Effects of Dose 1, Dose 2, and Protocol Between Labs

High variance suggests that these labs are not comparable because it suggests that the blotted uterus weight has no distinct response across all labs after adding dose 1, dose 2, or utilizing a certain protocol. Protocol A is used as a baseline for for the random effects. Looking at the variation of the random effects of dose 1, dose 2, and protocol on lab, there exists variation in blotted uterus weight for each of these features between the different labs as denoted by the variance.

In particular, protocols C & D have a higher variance compared to protocol A. Protocol B has a slightly higher variance compared to protocol A. Thus, for example, protocols C & D may be comparable, but not protocols A & D. Intuitively, this makes sense because the Protcols A and B have lower body weights than Protcols C and D, as illustrated below:



There exists extremely high variation for dose 1, so the effect of dose 1 highly varies between labs. Although not as high, there also exists variation for dose 2, so the effect of dose 2 also varies between labs.

Differences in Body Weight Between Labs

After clustering the body weight of each rat into three categories, the body weight of group 1 (mean body weight: 349.1) > (is greater than) group 3 (mean: 122.48) > (is greater than) group 2 (mean:60.4). Using group 1 as a baseline, if the blotted uterus weight of a rat that was in group 2 increased by one unit, then we would expect a -0.696 unit decrease in the blotted uterus weight of a rat that was in group 1. If the blotted uterus weight of a rat that was in group 3 increased by one unit, then we would expect a -0.059 unit decrease in the blotted uterus weight of a rat that was in group 1.

Since all labs do not have rats with similar body weights (Lab Basf's mean body weight: 56.7 vs Lab Mitsubishi's mean body weight:144.7), these differences in body weight would affect the response, blotted uterus weight, and make the labs again, incomparable.

Recommendations

Again, differences in protocols seem to matter more than differences in labs. The random and fixed effects of dose 1 and dose 2 on blotted uterus weight vary between labs significantly. Body weight also has a slight effect on blotted uterus weight--the heavier the body weight, the heavier the blotted uterus weight. Due to the variability between labs, perhaps not all labs should be compared and rather groupings of similar labs should be investigated as there are many similar labs with negligible differences.

To avoid this problem altogether, we recommend that this assay would be best conducted in one lab, one standard body weight, and one protocol to ensure consistency in the results.

Appendix

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## log(blot + 1) ~ (1 + proto + kern_distribution(dose1, knots = 4) +
##     kern_distribution(dose2, knots = 3) | lab) + kern_distribution(dose1,
##     knots = 4) + kern_distribution(dose2, knots = 3) + bodyClust
##    Data: dat
##
## REML criterion at convergence: -58.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.5912 -0.6043 -0.0129  0.5929  5.0064
##
## Random effects:
##  Groups   Name                                 Variance Std.Dev. Corr
##  lab      (Intercept)                          0.050577 0.22489
##           protoB                               0.010348 0.10173  -0.43
##           protoC                               0.046526 0.21570  -0.88
##           protoD                               0.046489 0.21561  -0.88
##           kern_distribution(dose1, knots = 4)1 4.062746 2.01563  -0.21
##           kern_distribution(dose1, knots = 4)2 0.072460 0.26918  -0.30
##           kern_distribution(dose1, knots = 4)3 0.283399 0.53235  -0.13
##           kern_distribution(dose1, knots = 4)4 0.005585 0.07474  -0.57
##           kern_distribution(dose2, knots = 3)1 3.859439 1.96455  -0.03
##           kern_distribution(dose2, knots = 3)2 0.539903 0.73478   0.00
##           kern_distribution(dose2, knots = 3)3 0.019278 0.13885   0.15
##  Residual                                      0.051289 0.22647
##
##
##
##   0.60
##   0.63  1.00
##   0.26  0.16  0.17
##   0.12  0.42  0.42 -0.80
##   0.04 -0.08 -0.07  0.17 -0.15
##   0.25  0.75  0.74  0.02  0.40 -0.68
##   0.25  0.06  0.07 -0.30  0.35  0.02  0.01
##   0.31  0.00  0.01 -0.16  0.18  0.38 -0.30  0.64
##  -0.40 -0.20 -0.21 -0.36  0.24 -0.56  0.31  0.38 -0.32
##
## Number of obs: 2677, groups:  lab, 19
##
## Fixed effects:
##                                       Estimate Std. Error t value
## (Intercept)                            4.58083    0.03839  119.32
## kern_distribution(dose1, knots = 4)1   8.70781    0.61771   14.10
## kern_distribution(dose1, knots = 4)2 -10.22012    1.00938  -10.13
## kern_distribution(dose1, knots = 4)3 113.16792   10.25487   11.04
## kern_distribution(dose1, knots = 4)4 -14.72312    1.38620  -10.62
## kern_distribution(dose2, knots = 3)1  -1.72556    0.43767   -3.94
## kern_distribution(dose2, knots = 3)2  -4.83378    0.26064  -18.55
## bodyClust2                            -1.19219    0.05022  -23.74
## bodyClust3                            -0.01391    0.02210   -0.63
##
## Correlation of Fixed Effects:
##             (Intr) k_(1,k=4)1 k_(1,k=4)2 k_(1,k=4)3 k_(1,k=4)4 k_(2,k=3)1
## kr_(1,k=4)1 -0.169
## kr_(1,k=4)2  0.074 -0.698
## kr_(1,k=4)3 -0.070  0.660     -0.998
## kr_(1,k=4)4  0.068 -0.659      0.998     -1.000
## kr_(2,k=3)1 -0.001 -0.227      0.015     -0.001      0.001
## kr_(2,k=3)2  0.011 -0.297      0.018     -0.006      0.005      0.749
## bodyClust2  -0.412 -0.031     -0.006     -0.014      0.008      0.050
## bodyClust3  -0.515  0.004      0.044     -0.041      0.043     -0.061
##             k_(2,k=3)2 bdyCl2
## kr_(1,k=4)1
## kr_(1,k=4)2
## kr_(1,k=4)3
## kr_(1,k=4)4
## kr_(2,k=3)1
## kr_(2,k=3)2
## bodyClust2   0.075
## bodyClust3  -0.083      0.349
```

Contributions

Sonia Xu: Wrote Case 1, Take 4 (Updated Model, Discussion, etc.)

InHee Ho: Wrote Exploratory Data Analysis, Edited Case 1, Take 4

Ian Hua: Edited Case 1, Take 4