STA721 Final Project

Shuangjie Zhang, Xiyang Hu

12/8/2018

1. Summary

2. Introductions

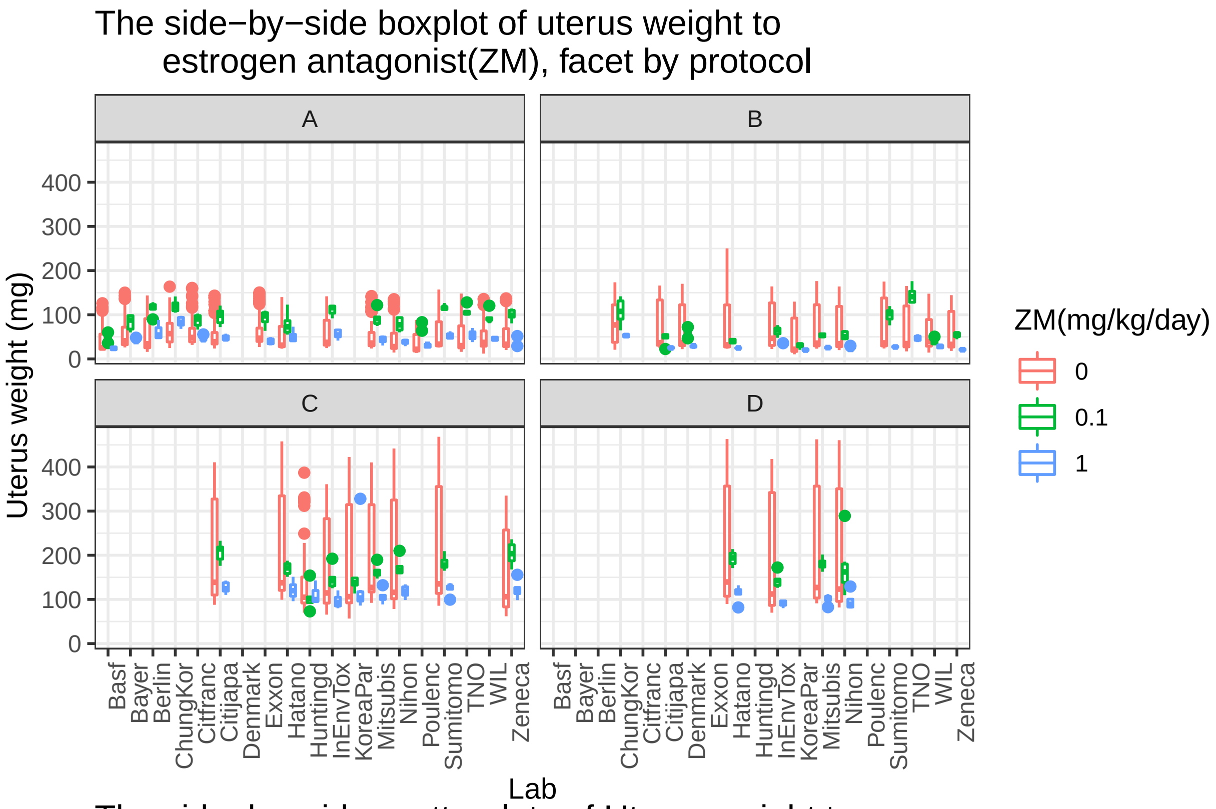
3. EDA

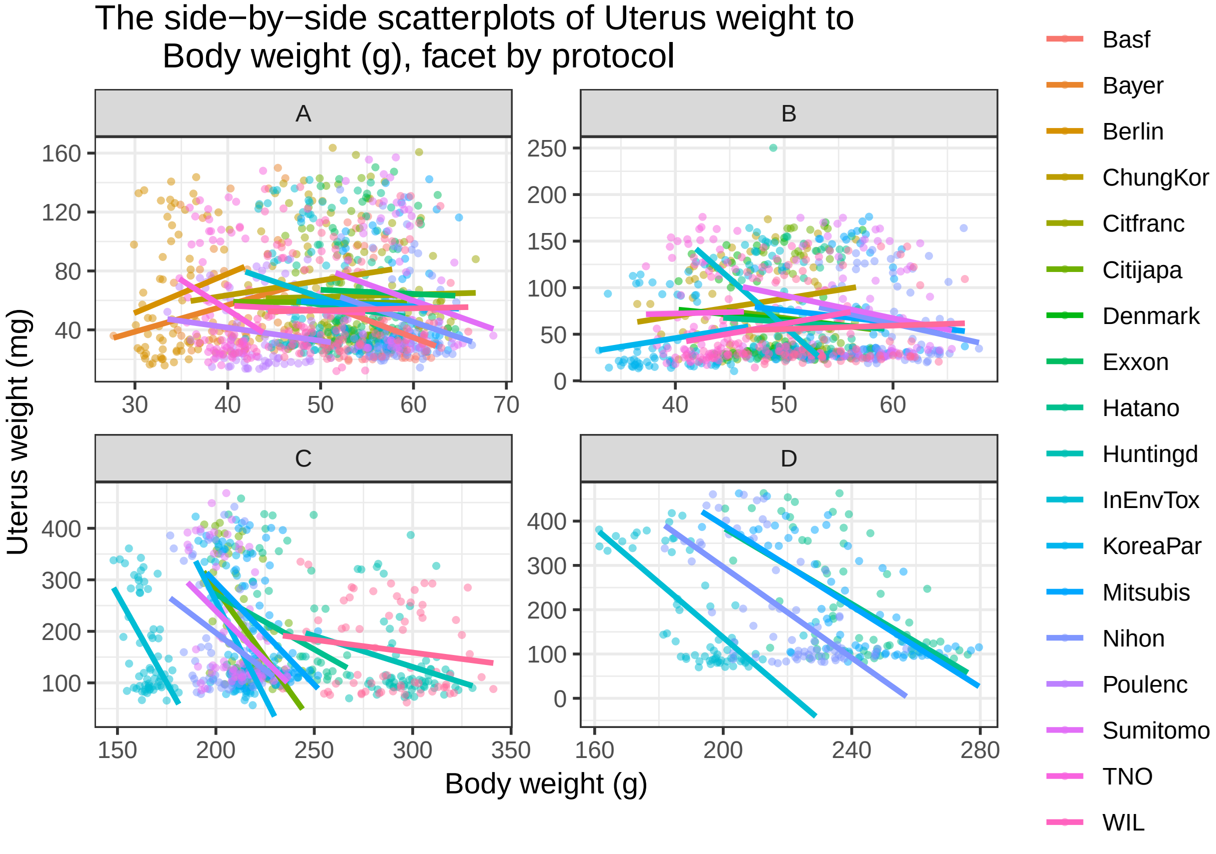
4. Model I And Result

We build a linear regression model excluding the `group` variable, because the group index varies in labs and cannot be considered as a factor. After looking into the data, we treat all variables but `uterus` and `weight` as factor. In order to use one full model to address all question, we include the interaction term of EE:protocol, ZM:protocol, EE:lab, ZM:lab. From EDA part we can find that some experiments are not done in some EE:ZM combination. So we cannot include this interaction term. Then we use `boxcox` and find that the log transformation is preferred. Therefore, the final model will be:

$$
\begin{aligned}
\log(\text{uterus}) = \beta_0 + \beta_1 \log(\text{weight}) + \beta_2\text{EE} + \beta_3\text{ZM} + \beta_4\text{lab} + \beta_5\text{protocol} \\
+\beta_6\text{EE:lab} + \beta_7\text{ZM:lab} + \beta_8\text{EE:protocol} + \beta_9\text{ZM:protocol} + \epsilon \\
\epsilon \sim N(0, \sigma^2)
\end{aligned}
$$

5. Conclusion


The side−by−side boxplot of uterus weight to estrogen antagonist(ZM), facet by protocol


The side−by−side scatterplots of Uterus weight to Body weight (g), facet by protocol

Appendix

EDA

```
bioassay_lm = bioassay[,-7]
str(bioassay_lm)
```

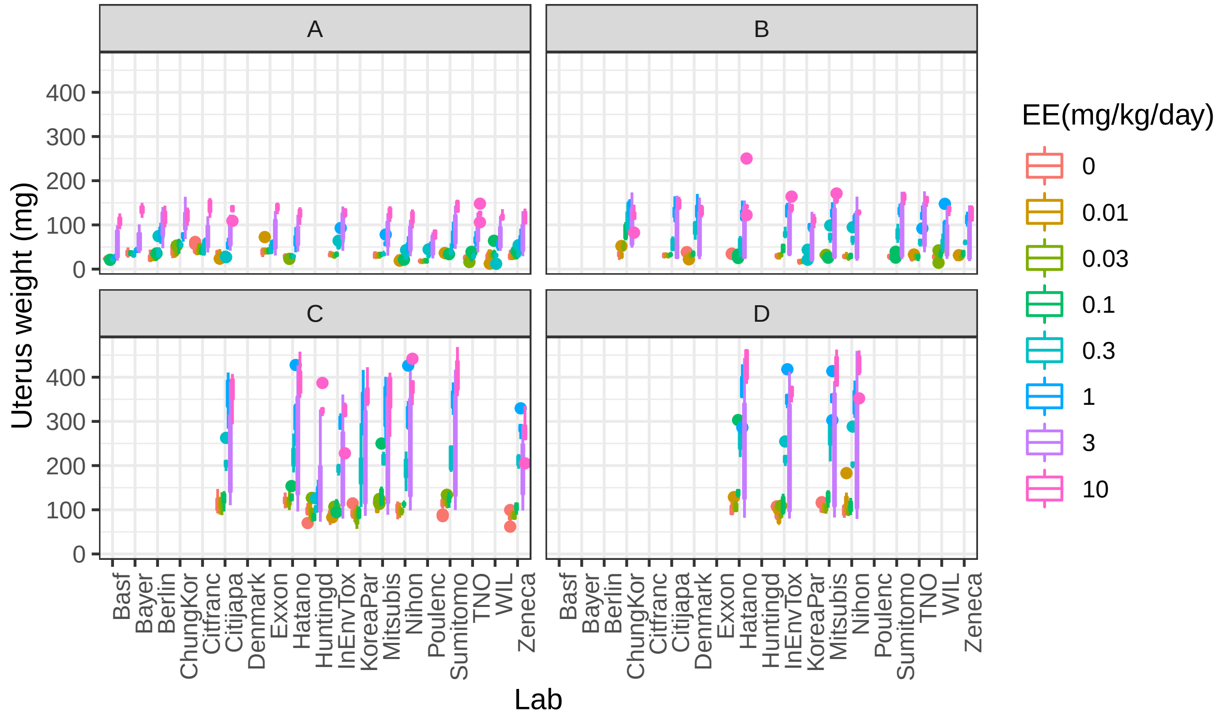
```
## 'data.frame':    2677 obs. of  6 variables:
##  $ uterus  : num  21 22 21 26 24 25 22 26 24 22 ...
##  $ weight  : num  61.9 55.9 59.1 54.8 57.5 57.6 60.3 59 59.1 61.4 ...
##  $ protocol: Factor w/ 4 levels "A","B","C","D": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EE      : Factor w/ 8 levels "0","0.01","0.03",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ ZM      : Factor w/ 3 levels "0","0.1","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ lab     : Factor w/ 19 levels "Basf","Bayer",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
table(bioassay_lm$EE, bioassay_lm$ZM)
```

```
##       
##           0 0.1   1
##   0    484   0   0
##   0.01 234   0   0
##   0.03 239   0   0
##   0.1  246   0   0
##   0.3  246   0   0
##   1    246   0   0
##   3    246 245 246
##   10   245   0   0
```

```
ggplot(data=bioassay,mapping = aes(y = uterus,x = lab,color=EE))+
  geom_boxplot()+theme_bw()+facet_wrap(~ protocol) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(x = "Lab", y="Uterus weight (mg)", title="The side-by-side boxplot of uterus weight for differen
       different dose of estrogen agonist(EE), facet by protocol", caption="", colour="EE(mg/kg/day)")
```

The side−by−side boxplot of uterus weight for different labs and different dose of estrogen agonist(EE), facet by protocol



Model Part I

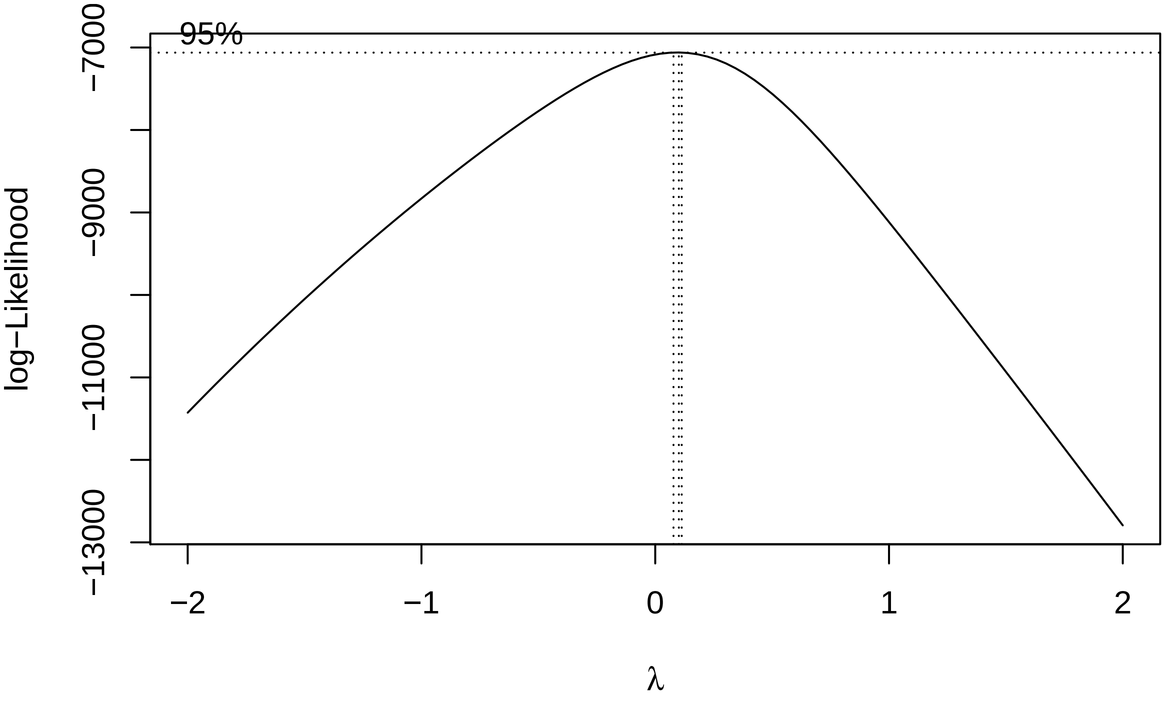
```
lm1 = lm(uterus~., data = bioassay_lm)
#summary(lm1)
step(lm1, k=log(2677))
```

```
## Start:  AIC=20175.57
## uterus ~ weight + protocol + EE + ZM + lab
##
##              Df Sum of Sq      RSS   AIC
## <none>                     4568714 20176
## - lab        18    304839  4873553 20206
## - weight      1    117187  4685901 20236
## - protocol    3    855660  5424374 20612
## - ZM          2   2030817  6599531 21144
## - EE          7   7683826 12252540 22761

##
## Call:
## lm(formula = uterus ~ weight + protocol + EE + ZM + lab, data = bioassay_lm)
##
## Coefficients:
## (Intercept)       weight    protocolB    protocolC    protocolD
##    15.82251     -0.45365      7.84315    207.53588    221.22623
##      EE0.01       EE0.03        EE0.1        EE0.3          EE1
##    -0.60177      0.26008      8.01257     47.94479    106.35605
##         EE3         EE10        ZM0.1          ZM1     labBayer
##    136.45891    150.55730    -80.51563   -127.18576      2.60266
##    labBerlin  labChungKor  labCitfranc  labCitijapa   labDenmark
##     14.84134     32.46041     26.21060     21.52689     18.95727
##     labExxon    labHatano  labHuntingd  labInEnvTox  labKoreaPar
##     23.72114     26.83352      0.09856      0.58445     -2.51500
## labMitsubis     labNihon   labPoulenc  labSumitomo       labTNO
##     24.63683     13.18893     -4.14169     28.52520     16.56429
##       labWIL    labZeneca
##     10.05022     17.93047
```

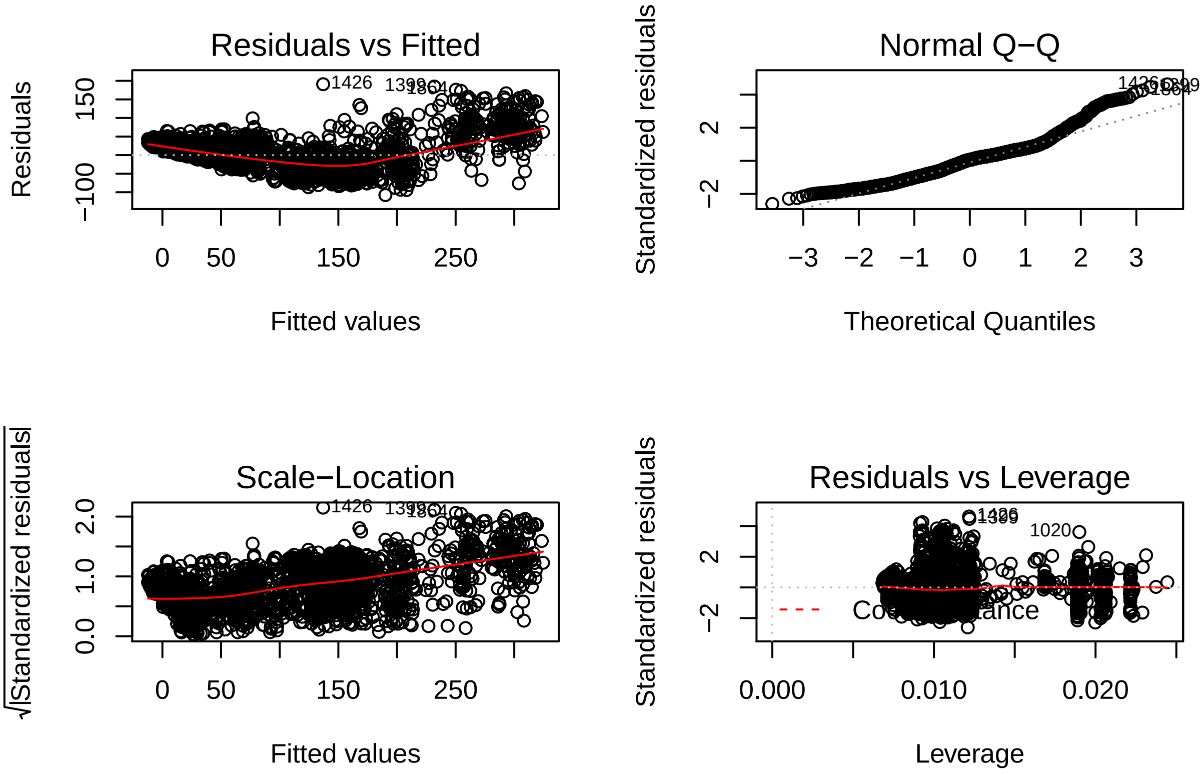
```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
box =boxcox(lm1)
```



```
lm2 = lm(formula = log(uterus) ~ log(weight) + protocol + EE + ZM + lab, data = bioassay_lm)
lm3 = lm(formula = log(uterus) ~ log(weight) + protocol + EE*lab +ZM*lab, data = bioassay_lm)
#summary(lm3)

par(mfrow=c(2,2))
plot(lm1)
```



Frequentist Random Effect Model:

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
randomeffect = lmer(log(uterus) ~ log(weight) + protocol + EE + ZM + (1+EE+ZM|lab), data = bioassay_lm)
```

```
## Warning in commonArgs(par, fn, control, environment()): maxfun < 10 *
## length(par)^2 is not recommended.

## Warning in optwrap(optimizer, devfun, getStart(start, rho$lower, rho$pp), :
## convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

## singular fit
```

```
#summary(randomeffect)
```

```
lm.full = lm(uterus~EE*lab+EE*protocol+ZM*lab+ZM*protocol+protocol+weight, data = bioassay)
#summary(lm.full)
#anova(lm.full)
```

Model Part II

Model Part III