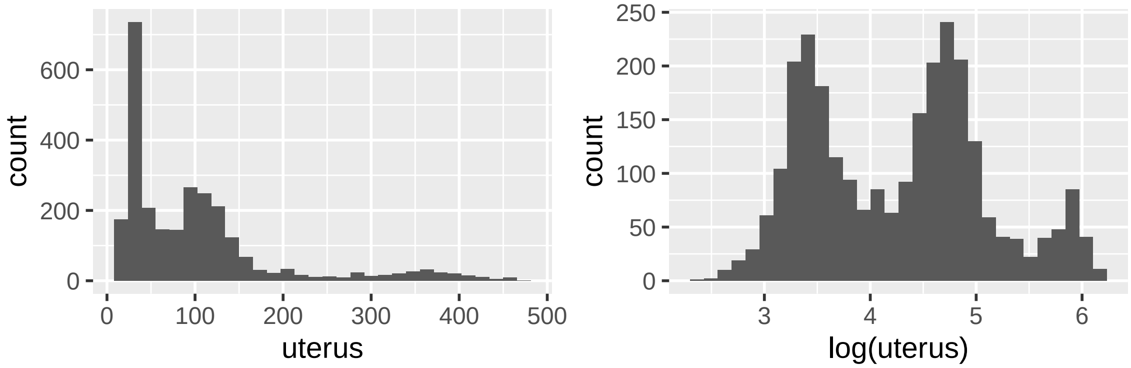
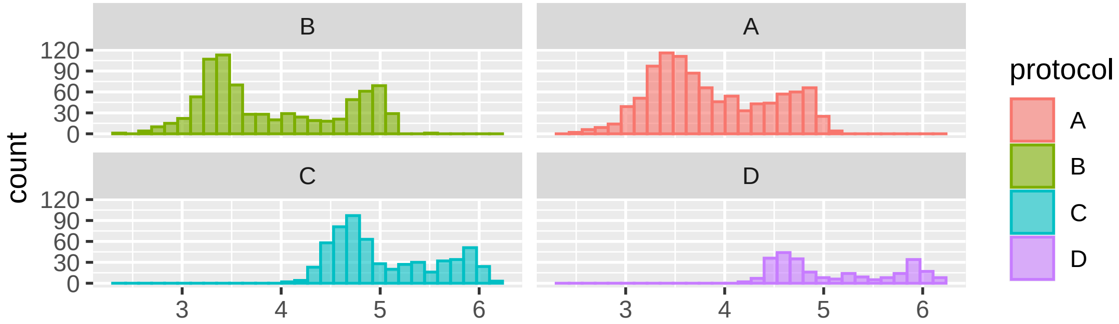
hw3

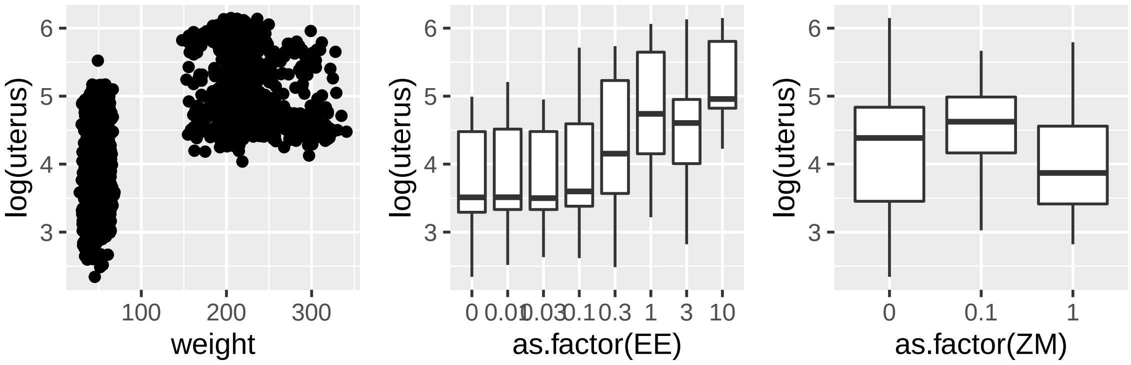
EDA



It turns out that we have a bimodal distribution for log(uterus). This is probably because of the mature status of the rats, in particular, due to protocol.



For this reason, protocol is suggested to be a random effect.



The above show the relationship of log(uterus) vs. weight, EE and ZM. We observe that weight seems to correlate to log(uterus) with 2 clear clusters, however, weight should be correlated to protocol since an immature rat tends to weight less and have lighter uterus.

The EE are positively correlated to uterus and ZM seems to have an upside-down U shape relation to uterus.

Model

For this part, I will create a basic model that answers the research question, and then create the final model by adding extra features to the model based on performance. The basic models are as below:

$$log(uterus_{ijk}) = \beta_0 + \beta_{1k}EE_{ijk} + \beta_{2k}ZM_{ijk}$$

$$\beta_0 = b_0 + b_{0j} + b_{0k}$$

$$\beta_{1k} = b_1 + b_{1k}$$

$$\beta_{2k} = b_2 + b_{2k}$$

Where i represents observation index and j represents lab and k represents protocol. b_0 is the grand mean of log(uterus), b_{0j} and b_{0k} are random intercept of lab and protocol respectively. b_1 and b_2 are fixed intercept of EE and ZM. b_{1k} and b_{2k} are the random slopes of EE and ZM with respected to protocol.

In this way, b_1 and b_2 are used for answering question 1; b_{0j} is for question 2 and b_{1k} and b_{2k} are for question 3.

Fit Model

I have tried 4 models and I will choose the best model based on BIC.

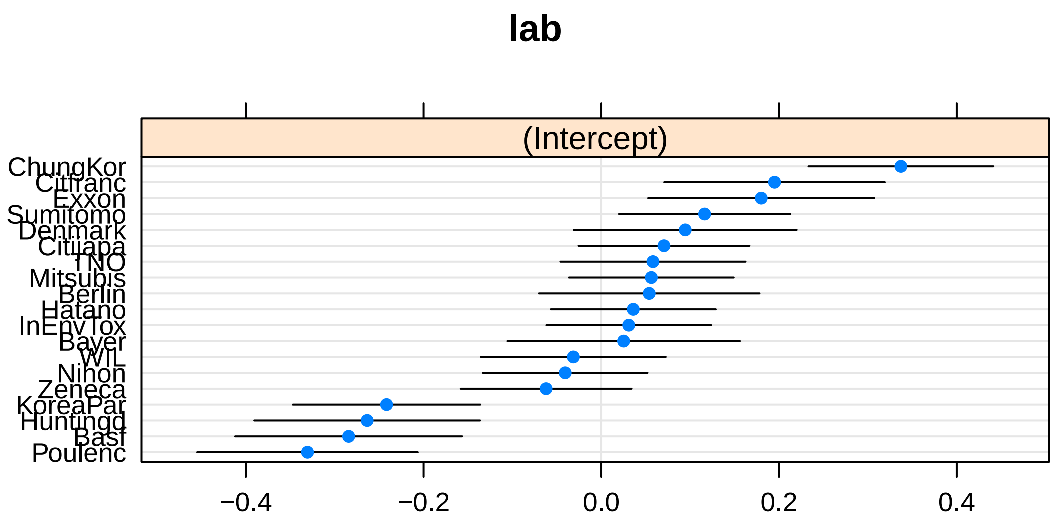
I find that the base model 'm1' (the model is specified above) has the lowest BIC at 3262 and therefore it is the best model. Analysis will be based on 'm1' later.

Table 1: Coefficient of Fixed Effects

	Estimate	Std. Error	t value
(Intercept)	4.270	0.378	11.294
EE	0.138	0.009	14.679
ZM	-0.569	0.142	-4.012

Based on the coefficient of EE and ZM, the estimate fixed intercept of EE is 0.138, which mean 1 unit increase in EE will lead to $(e^{0.138} - 1) * 100\% = 14.8$ increase in uterus; The estimate coefficient of ZM is -0.569, which mean 1 unit increase in ZM will lead to $(1 - e^{-0.569}) * 100\% = 43.3$ decrease in uterus; we confirm that uterus weight exhibit an increasing dose response trend for EE and a decreasing dose response trend for ZM. This answers question 1.

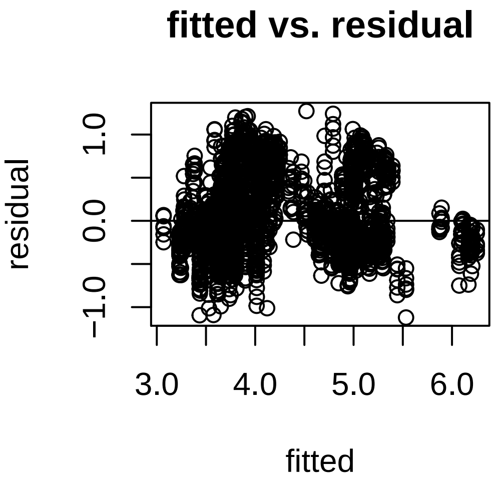
```
## $lab
```

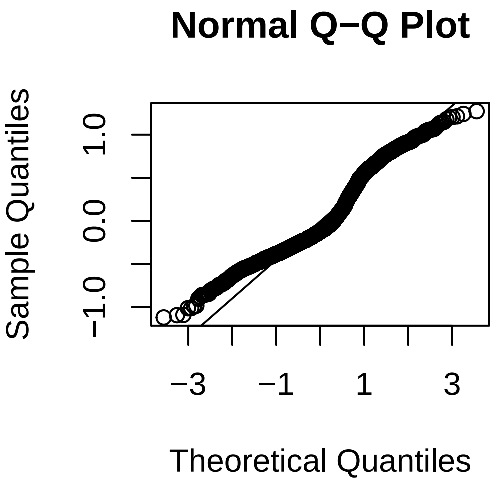


Based on the plot, the random intercept varies cross the labs. The Poulenc has the lowest uterus weight and ChungKor has the highest. This partly answers question 2.

In addition, the within-group variance is 0.596379 and the cross-group variance is 0.185207, thus the correlation between two samples from same group is $\frac{0.185207}{0.185207+0.596379} = 0.2370$ which is small.

Residual and Sensitivity Analysis





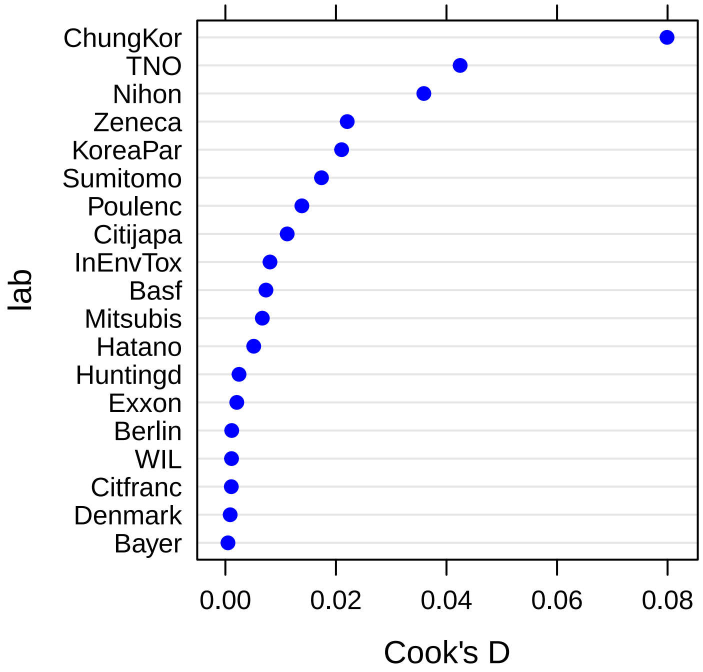
Based on the residual plots, the model exhibits violation of constant variance a bit, and it roughly follows normality.

Next, I use DFBETAS to determine the influential labs.

Table 2: dfbetas of labs

	(Intercept)	EE	ZM
Basf	-0.042	0.001	-0.034
Bayer	0.005	-0.016	-0.012
Berlin	0.006	0.001	0.042
ChungKor	0.058	-0.207	0.122
Citfranc	0.032	-0.045	-0.021
Citijapa	0.013	-0.071	-0.018
Denmark	0.017	-0.031	-0.012
Exxon	0.030	-0.049	-0.030
Hatano	0.028	0.014	0.001
Huntingd	-0.045	0.021	0.041
InEnvTox	-0.038	-0.012	-0.010
KoreaPar	-0.039	0.115	0.034
Mitsubis	0.017	-0.049	-0.102
Nihon	0.011	0.098	-0.069
Poulenc	-0.050	-0.015	0.009
Sumitomo	0.015	0.064	-0.029
TNO	0.005	0.105	0.106
WIL	-0.006	-0.012	0.017
Zeneca	-0.012	-0.076	-0.001

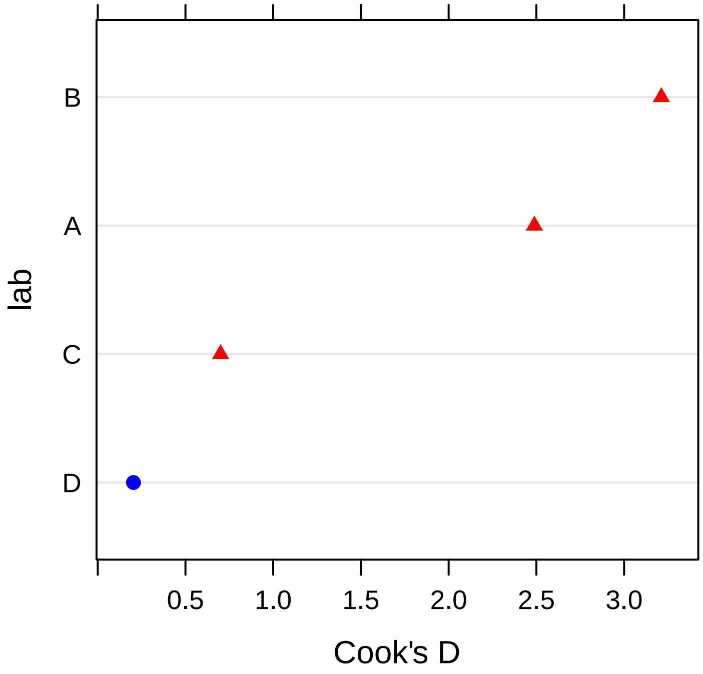
The cutoff is $\frac{2}{\sqrt{19}} = 0.459$, based on DFBETAS there is no influntial labs on the random effects.



As for the Cook's distance, lab ChungKor is the largest, but it is too small to be claimed as influential. For this reason, there is no lab that appears to be outlier. This answers question 2.

Table 3: dfbetas of protocol

	(Intercept)	EE	ZM
A	-0.594	0.404	1.115
B	-0.427	0.362	-0.624
C	0.477	-0.994	0.135
D	0.500	-0.328	-0.299



As for protocol, the cutoff is $\frac{2}{\sqrt{4}} = 1$, thus protocol A is influential in random effect of ZM and protocol C is influential in random effect of C (-0.994 is close to threshold 1). Protocol D has small DFBETAS on average and also has the smallest Cook's distance. Thus I recommend to use protocol D for detecting effects of EE and ZM. This answers question 3.

Limits

The limits of the model might be the violation of constant variance. Also I am not able to create a nested model due to the small sample size.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(MASS)
library(tidyverse)
```

```r
library(lme4)
library(knitr)
library(lattice)
library(influence.ME)
library(gridExtra)
library(hrbrthemes)
library(viridis)
library(forcats)
library(GGally)
knitr::opts_chunk$set(echo = F, warning = F, message = F, fig.align = 'center', fig.width=6, fig.height=
file = 'bioassay.txt'
data <- read.delim(file, header = TRUE, sep=' ')
summary(data)
data$protocol <- as.factor(data$protocol)
data$uterus <- as.numeric(data$uterus)
data$weight <- as.numeric(data$weight)
data$lab <- as.factor(data$lab)
summary(data)
# remove na
data <- data %>% filter(!is.na(uterus))
p1 <- ggplot(data=data, aes(x=uterus))+
  geom_histogram()

p2 <- ggplot(data=data, aes(x=log(uterus)))+
  geom_histogram()

grid.arrange(p1, p2, nrow = 1);
data %>%
  mutate(text = fct_reorder(protocol, log(uterus))) %>%
  ggplot( aes(x=log(uterus), color=protocol, fill=protocol)) +
    geom_histogram(alpha=0.6) +
    xlab("") +
    ylab("count") +
    facet_wrap(~text);
p3 <- ggplot(data, aes(x=weight, y=log(uterus)))+
  geom_point()

p4 <- ggplot(data, aes(x=as.factor(EE), y=log(uterus)))+
  geom_boxplot()

p5 <- ggplot(data, aes(x=as.factor(ZM), y=log(uterus)))+
  geom_boxplot()
grid.arrange(p3, p4,p5, nrow = 1);
m1 <- lmer(log(uterus) ~ 1 + EE + ZM + (1|lab) + (EE + ZM|protocol), data=data)
m2 <- lmer(log(uterus) ~ 1 + EE + ZM + weight + (1|lab) + (EE + ZM|protocol), data=data)
m3 <- lmer(log(uterus) ~ 1 + EE * ZM + weight + (1|lab) + (EE + ZM|protocol), data=data)
m4 <- lmer(log(uterus) ~ 1 + EE * ZM + weight + (1|lab) + (EE * ZM|protocol), data=data)
anova(m1, m2)
anova(m1, m3)
anova(m1, m4)
m1_summary = summary(m1)
kable(m1_summary$coefficients, digits = 3, caption="Coefficient of Fixed Effects")
p6 <- dotplot(ranef(m1, condVar=T))[1]
```

```r
p6
m1_summary
par(mfrow=c(1,2))
plot(fitted(m1), resid(m1), xlab = 'fitted', ylab='residual',
     main='fitted vs. residual')
abline(h=0)
qqnorm(resid(m1))
qqline(resid(m1))
m1.inf1 <- influence(m1, "lab")
kable(dfbetas(m1.inf1), digits = 3, caption="dfbetas of labs")
plot(m1.inf1, which="cook", cutoff=4/length(unique(data$lab)), sort=T, xlab="Cook's D", ylab='lab')
m1.inf2 <- influence(m1, "protocol")
kable(dfbetas(m1.inf2), digits = 3, caption="dfbetas of protocol")
plot(m1.inf2, which="cook", cutoff=4/length(unique(data$lab)), sort=T, xlab="Cook's D", ylab='lab')
```