Data

Data

General:

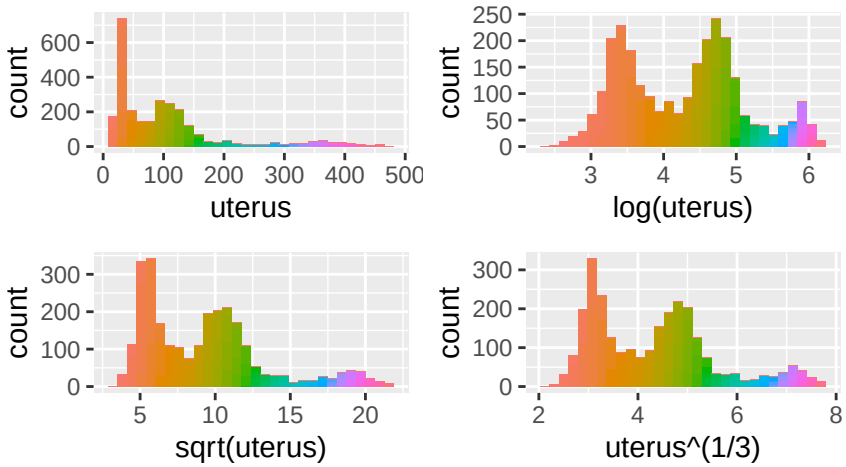
Data used in this analysis containing 2681 observations from different labs, conducting the research whether the estrogen level would effect the weight of the uterus of the rats or not. From taking a quick look to the data set, We noticed that there are 4 rows missing uterus weight values and 2 of them even missing weight values. Since these missing rows are occured rondomly,which means these rows are in different group, protocol type, and lab. Therefore, we decided to delete these four rows.

For the variables in the data set, we treat uterus, weight, EE, ZM as numeric variables and protocal, lab, group as categorical variables. Though in the research, there are only 3 kinds of dosage of ZM and 7 kinds of dosage of EE, we still treat them as numerical variables because if we treat them as categorical variables, we would lose information between different dosage. For example, 10 dose is 10 times of 2 dose. Besides the existing variables in the data set, we add a new binary variable- mature, based on the value of protocol, to indicate whether the rats are mature. If the rat is categorized as protocol A or B, it would have value 0, and if the rat is categorized as protocol C or D, it would have value 1.

EDA:

For the data set we use, we keep some colinearity issue in mind when we explore the data. Mature versus Protocol: Since mature is the new variable we created based on protocol types, there is a high colinearity between them. ZM and EE versus Group: Since groups is seperated depends on the different combination of dosage of ZM and EE, they are highly correlated with each other.

First, we plot a histogram of the response variable- uterus weight. However, the distribution is so skewed and is not normal distribution. We try square root, cube root, and log transformation, it seems that log transformation improves the distribution most. As a result, even though it is still hard to say that the distribution of log uterus weight is normal distribution, since this is the best distribution we can get, in the following analysis, we will to use log uterus weight as our response variable.



Moreover, we use plots to check the relatonship between log uterus weight with other variables. First, we take a look the difference of log uterus weight between labs. By the distribution of log uterus weight of each lab, it seems like there is some apparent difference. However, if we further seperate the data by different protocol types, in each protocol type group, each lab has similar distribution of log uterus weight. This difference we observed is caused by the fact that not all lab conduct experiment for every protocol types. Moreover, we can see apparent pattern that the data points are clustered by group when we draw the plot for log uterus weight and mean centered weight. There are four clusters in the plot. For the relation between log uterus rate and protocol, rates categorized as protocol C, D apparently have higher log uterus weight comparing to protocol A, B.

All these observations indicates that different protocol types would have different log uterus distributions. Therefore, we try to plot the relationship between log uterus weight with each variable by different protocol types. We find that when comparing log uterus weight and mean centered weight, the mature rats have a negative pattern, the immature rats do not appear to have a pattern. For the different dosage of ZM and EE, there appears to be a positive relationship between log uterus weight and ZM and a negative relationship between log uterus weight and EE. Lastly, we notice that different group of rats have different distribution of log uterus weights. However, since grouping is based on the dosage of ZM and EE and we are interested in the effect of ZM and EE, we will only include ZM and EE in our model and excluded group variable.