Linear Mixed effect model-day2

Chandan Kumar Pandey

2022-10-21

Mixed effect models

possun Morphometic dataset.

```
possum <- read.csv("./data/possum.csv",header = T)
## Basic data data exploration
summary(possum)
```

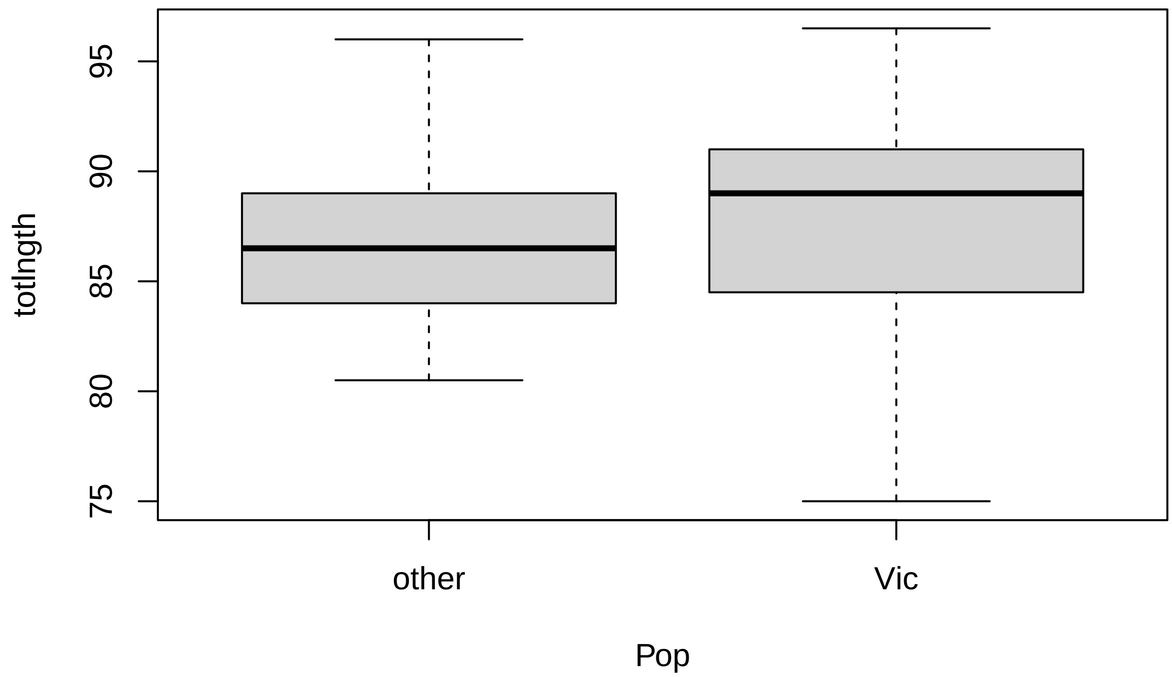
```
##       case             site           Pop                 sex
##  Min.   :  1.00   Min.   :1.000   Length:104         Length:104
##  1st Qu.: 26.75   1st Qu.:1.000   Class :character   Class :character
##  Median : 52.50   Median :3.000   Mode  :character   Mode  :character
##  Mean   : 52.50   Mean   :3.625
##  3rd Qu.: 78.25   3rd Qu.:6.000
##  Max.   :104.00   Max.   :7.000
##
##       age           hdlngth           skullw         totlngth
##  Min.   :1.000   Min.   : 82.50   Min.   :50.00   Min.   :75.00
##  1st Qu.:2.250   1st Qu.: 90.67   1st Qu.:54.98   1st Qu.:84.00
##  Median :3.000   Median : 92.80   Median :56.35   Median :88.00
##  Mean   :3.833   Mean   : 92.60   Mean   :56.88   Mean   :87.09
##  3rd Qu.:5.000   3rd Qu.: 94.72   3rd Qu.:58.10   3rd Qu.:90.00
##  Max.   :9.000   Max.   :103.10   Max.   :68.60   Max.   :96.50
##  NA's   :2
##      taill          footlgth        earconch          eye             chest
##  Min.   :32.00   Min.   :60.30   Min.   :40.30   Min.   :12.80   Min.   :22.0
##  1st Qu.:35.88   1st Qu.:64.60   1st Qu.:44.80   1st Qu.:14.40   1st Qu.:25.5
##  Median :37.00   Median :68.00   Median :46.80   Median :14.90   Median :27.0
##  Mean   :37.01   Mean   :68.46   Mean   :48.13   Mean   :15.05   Mean   :27.0
##  3rd Qu.:38.00   3rd Qu.:72.50   3rd Qu.:52.00   3rd Qu.:15.72   3rd Qu.:28.0
##  Max.   :43.00   Max.   :77.90   Max.   :56.20   Max.   :17.80   Max.   :32.0
##                  NA's   :1
##      belly
##  Min.   :25.00
##  1st Qu.:31.00
##  Median :32.50
##  Mean   :32.59
##  3rd Qu.:34.12
##  Max.   :40.00
```

```
##
```

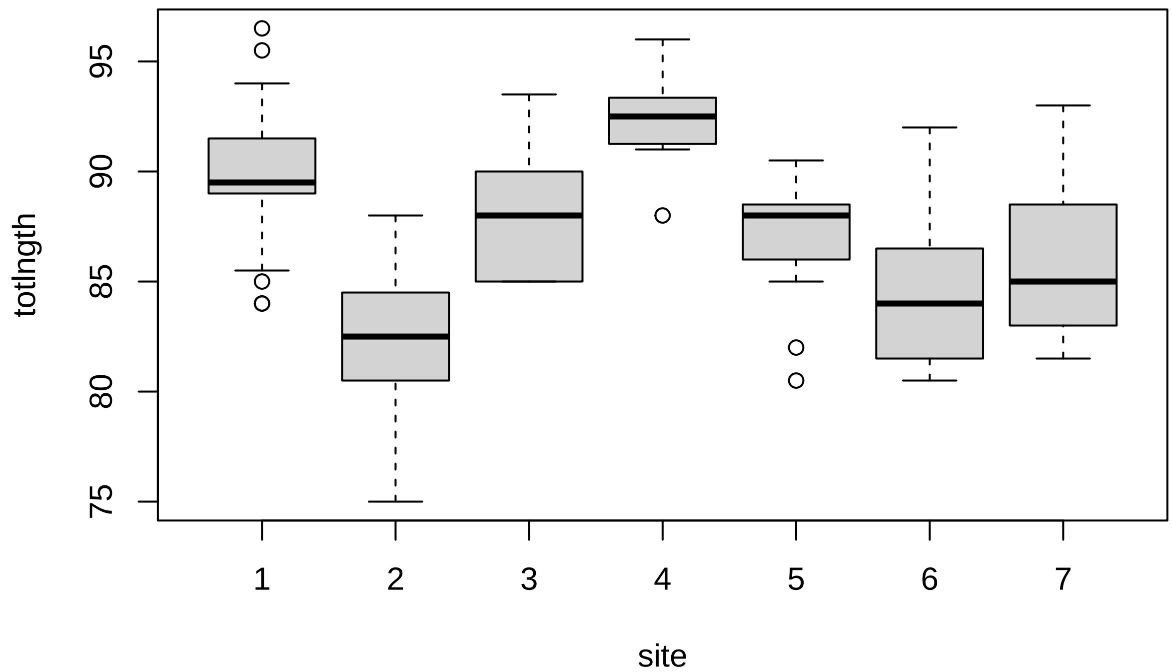
```
## convert the population,site and gender  into factors
possum$site <-as.factor(possum$site)
possum$sex <- as.factor(possum$sex)
possum$Pop <-as.factor(possum$Pop)
summary(possum)
```

```
##       case          site      Pop      sex         age           hdlngth
##  Min.   :  1.00   1:33   other:58   f:43   Min.   :1.000   Min.   : 82.50
##  1st Qu.: 26.75   2:13   Vic  :46   m:61   1st Qu.:2.250   1st Qu.: 90.67
##  Median : 52.50   3: 7                     Median :3.000   Median : 92.80
##  Mean   : 52.50   4: 7                     Mean   :3.833   Mean   : 92.60
##  3rd Qu.: 78.25   5:13                     3rd Qu.:5.000   3rd Qu.: 94.72
##  Max.   :104.00   6:13                     Max.   :9.000   Max.   :103.10
##                   7:18                     NA's   :2
##      skullw         totlngth         taill          footlgth
##  Min.   :50.00   Min.   :75.00   Min.   :32.00   Min.   :60.30
##  1st Qu.:54.98   1st Qu.:84.00   1st Qu.:35.88   1st Qu.:64.60
##  Median :56.35   Median :88.00   Median :37.00   Median :68.00
##  Mean   :56.88   Mean   :87.09   Mean   :37.01   Mean   :68.46
##  3rd Qu.:58.10   3rd Qu.:90.00   3rd Qu.:38.00   3rd Qu.:72.50
##  Max.   :68.60   Max.   :96.50   Max.   :43.00   Max.   :77.90
##                                                  NA's   :1
##     earconch          eye            chest          belly
##  Min.   :40.30   Min.   :12.80   Min.   :22.0   Min.   :25.00
##  1st Qu.:44.80   1st Qu.:14.40   1st Qu.:25.5   1st Qu.:31.00
##  Median :46.80   Median :14.90   Median :27.0   Median :32.50
##  Mean   :48.13   Mean   :15.05   Mean   :27.0   Mean   :32.59
##  3rd Qu.:52.00   3rd Qu.:15.72   3rd Qu.:28.0   3rd Qu.:34.12
##  Max.   :56.20   Max.   :17.80   Max.   :32.0   Max.   :40.00
##
```

```
## you can see the difference in the way population,site and gender columns
## box plot of lenght vs gender and pop
boxplot(totlngth~Pop, data = possum)
```




```
boxplot(totlngth~site, data = possum)
```




```
## pair plot to see if there is co - linearity
ggpairs(possum[,c(8,7,9,10)])
```

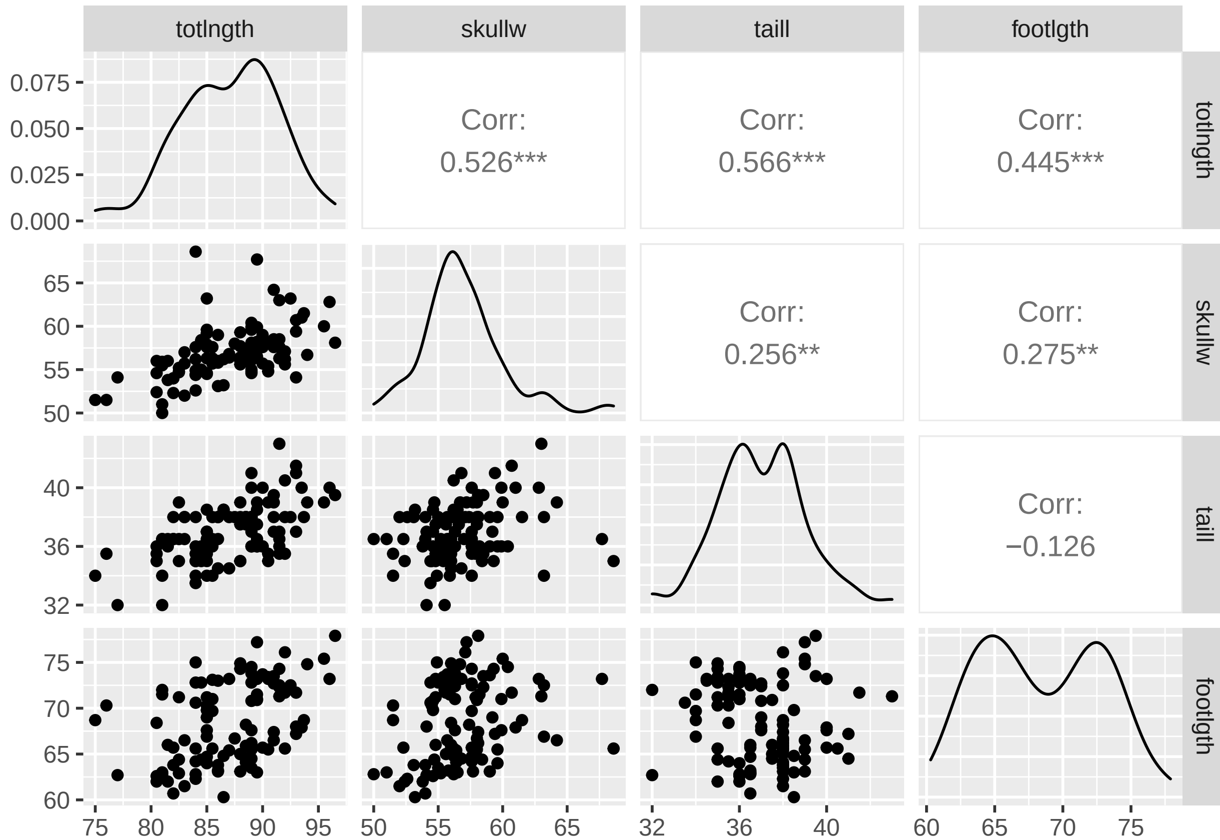
```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning: Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing non-finite values (stat_density).
```




```r
#
```

Droping the co-linear variable.

There is small amount of correlation between tail and skull, foot len vs skull. However, since we want to know of skull width can predict the possum length.

- What you take out of this graphs on the population trend of the possum

Let fit simple linear model

$$Lenght = \alpha + \beta_1 * Skullwidth + \epsilon$$

Where

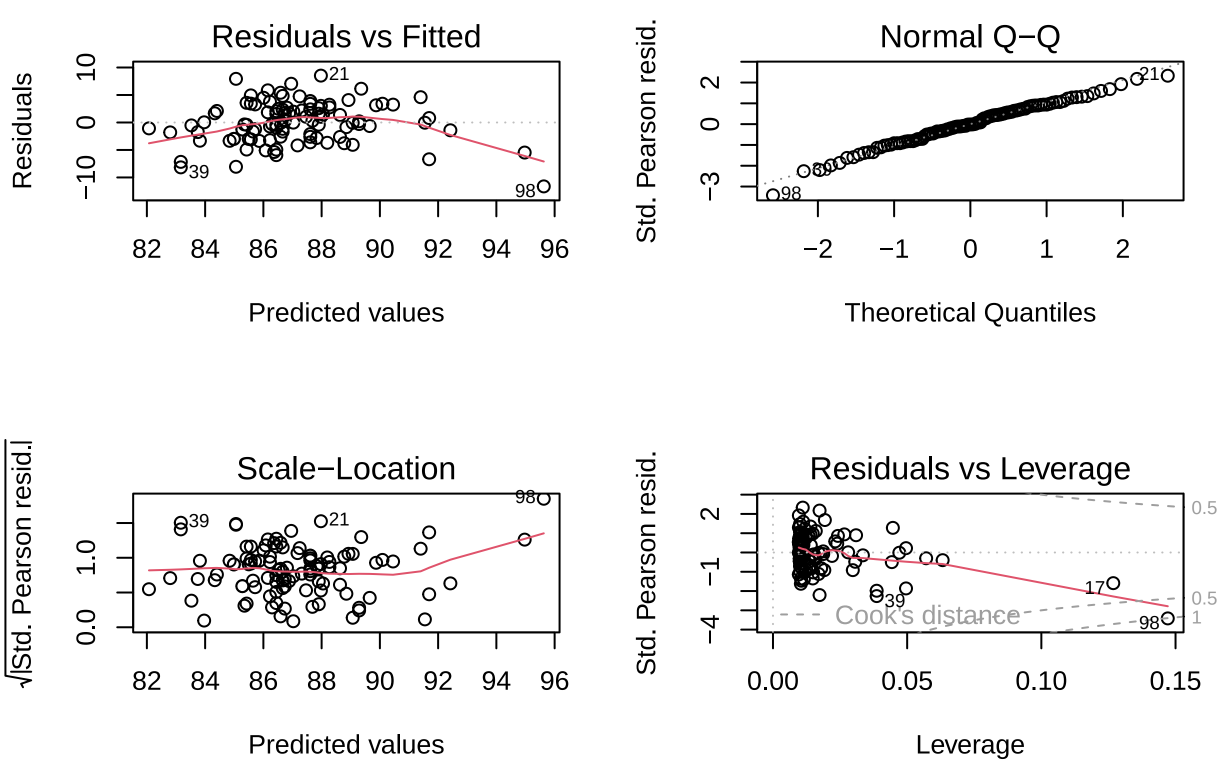
$$\epsilon \sim N(0, \sigma^2)$$

```r
mod_lm <- glm(totlngth~skullw,data = possum,family = "gaussian")
summary(mod_lm)
```

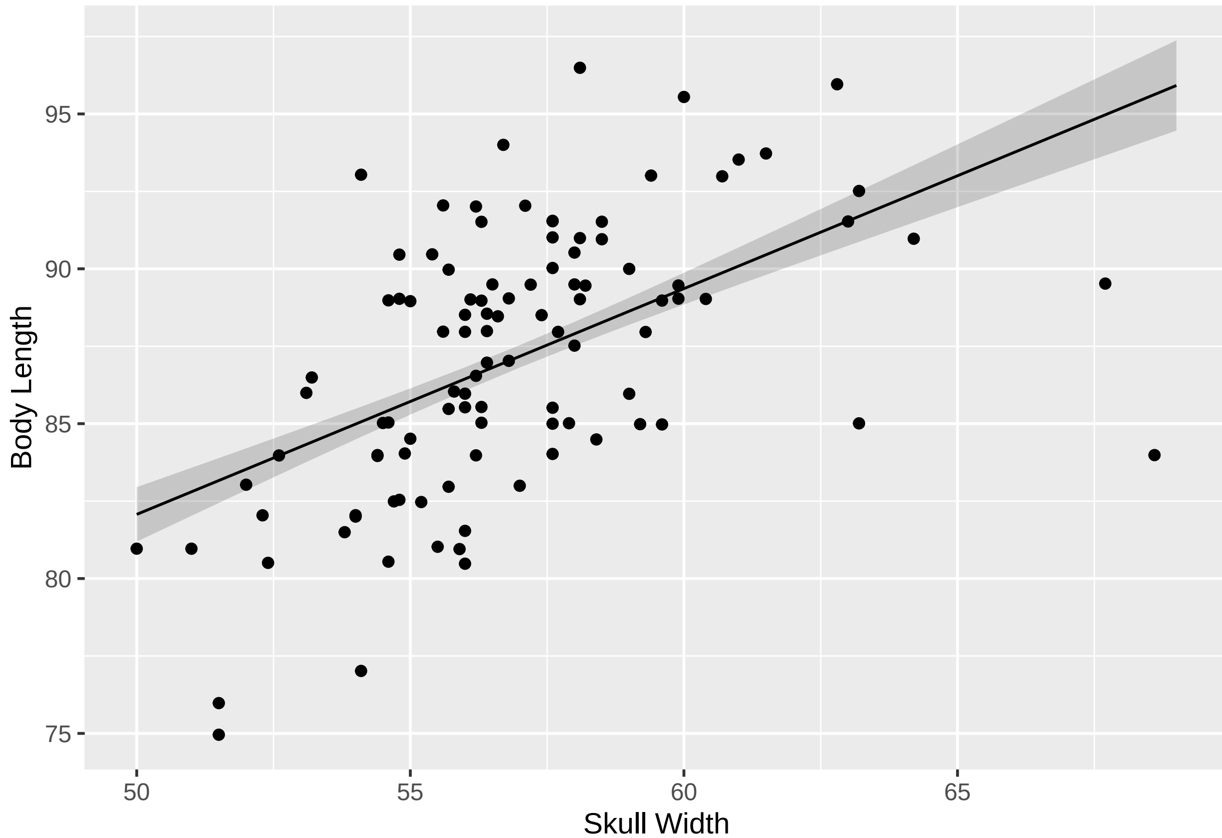
```
##
## Call:
## glm(formula = totlngth ~ skullw, family = "gaussian", data = possum)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -11.6276   -2.6156   -0.0572    2.6212    8.5250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.6305     6.6399   6.872 5.13e-10 ***
## skullw        0.7288     0.1166   6.253 9.50e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 13.56358)
##
##     Null deviance: 1913.8  on 103  degrees of freedom
## Residual deviance: 1383.5  on 102  degrees of freedom
## AIC: 570.29
##
## Number of Fisher Scoring iterations: 2
```

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




```
ggPredict(mod_lm,se = T,interactive = F)+labs(x="Skull Width", y= "Body Length")
```



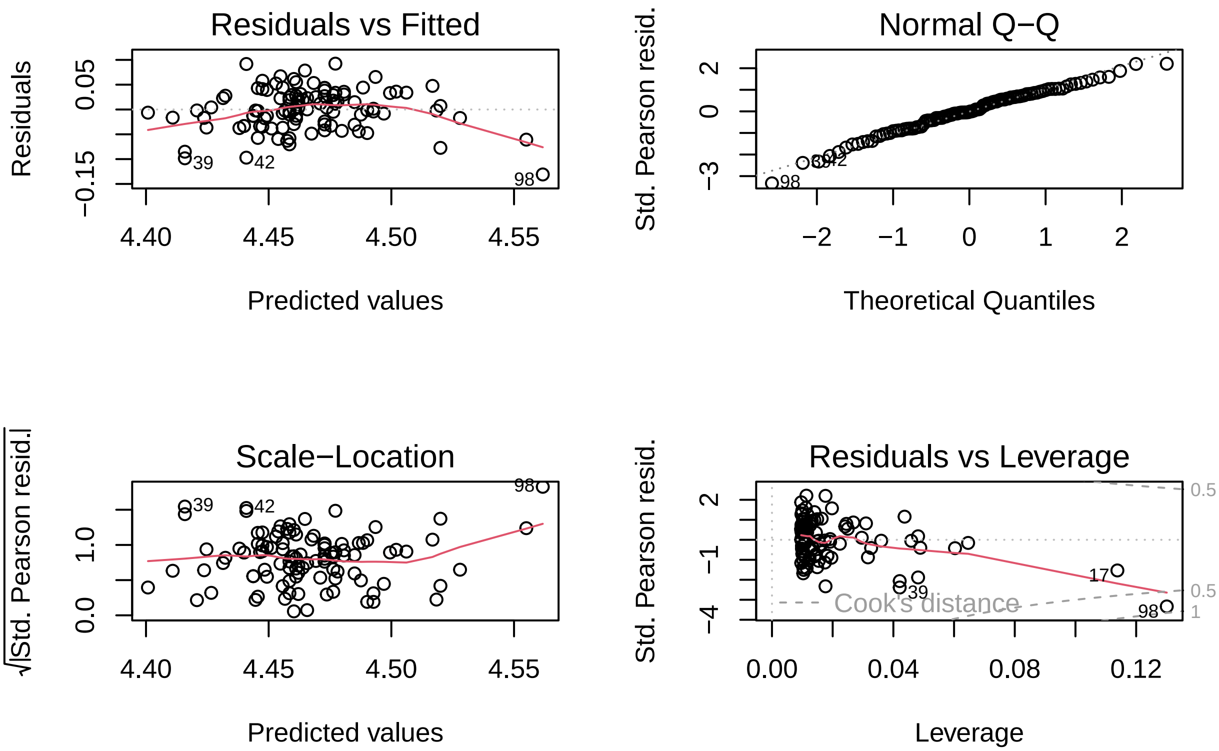
Clearly there the data is not homogeneous,

- We need to transform the variable to see if we can get out model to work

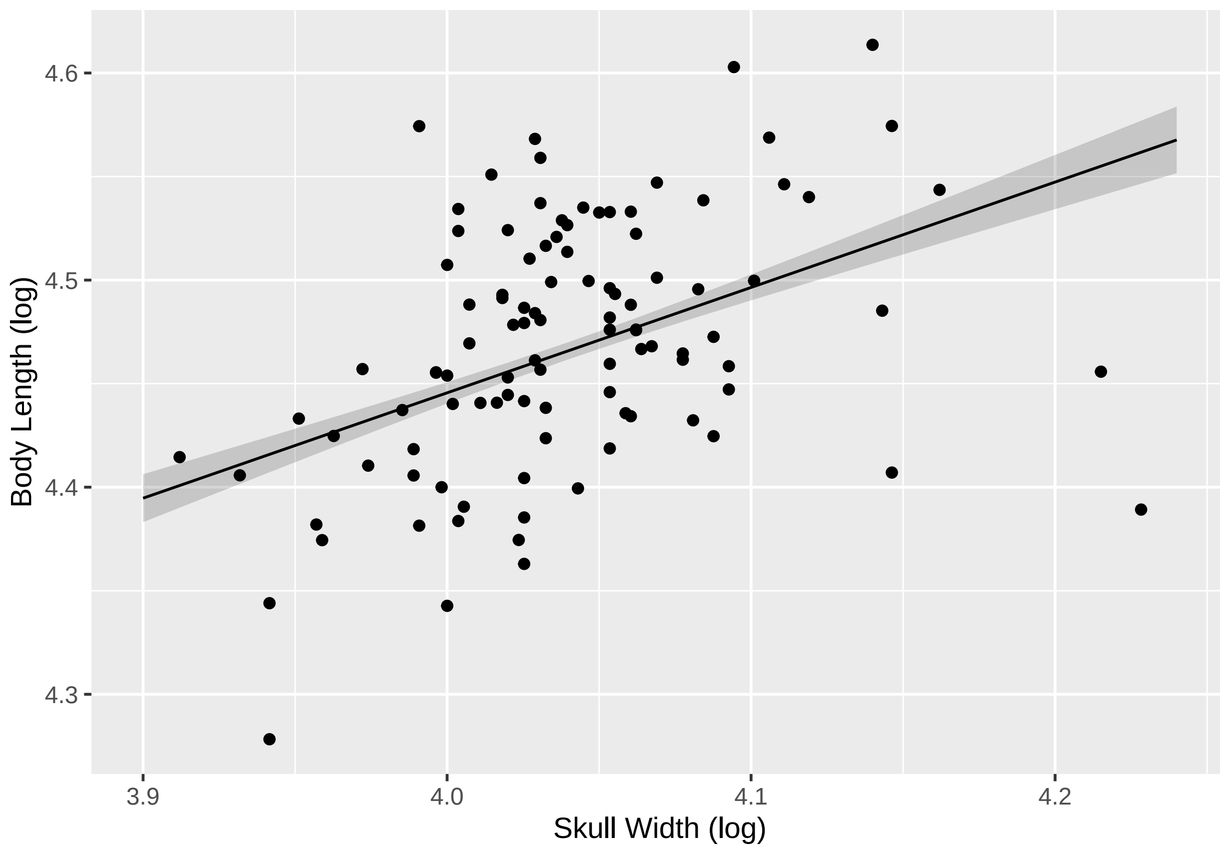
New model will be

$$\log(Length) = \alpha + \beta_1 * \log(Skullwidth) + \epsilon$$

```
possum <- possum%>%mutate(log_skullwid=log(skullw))
possum <- possum%>%mutate(log_len=log(totlngth))
mod_lm_log <- glm(log_len~log_skullwid,data = possum,family = "gaussian")
par(mfrow=c(2,2))
plot(mod_lm_log)
```




```
ggPredict(mod_lm_log,se = T,interactive = F)+labs(x="Skull Width (log)", y= "Body Length (log)")
```

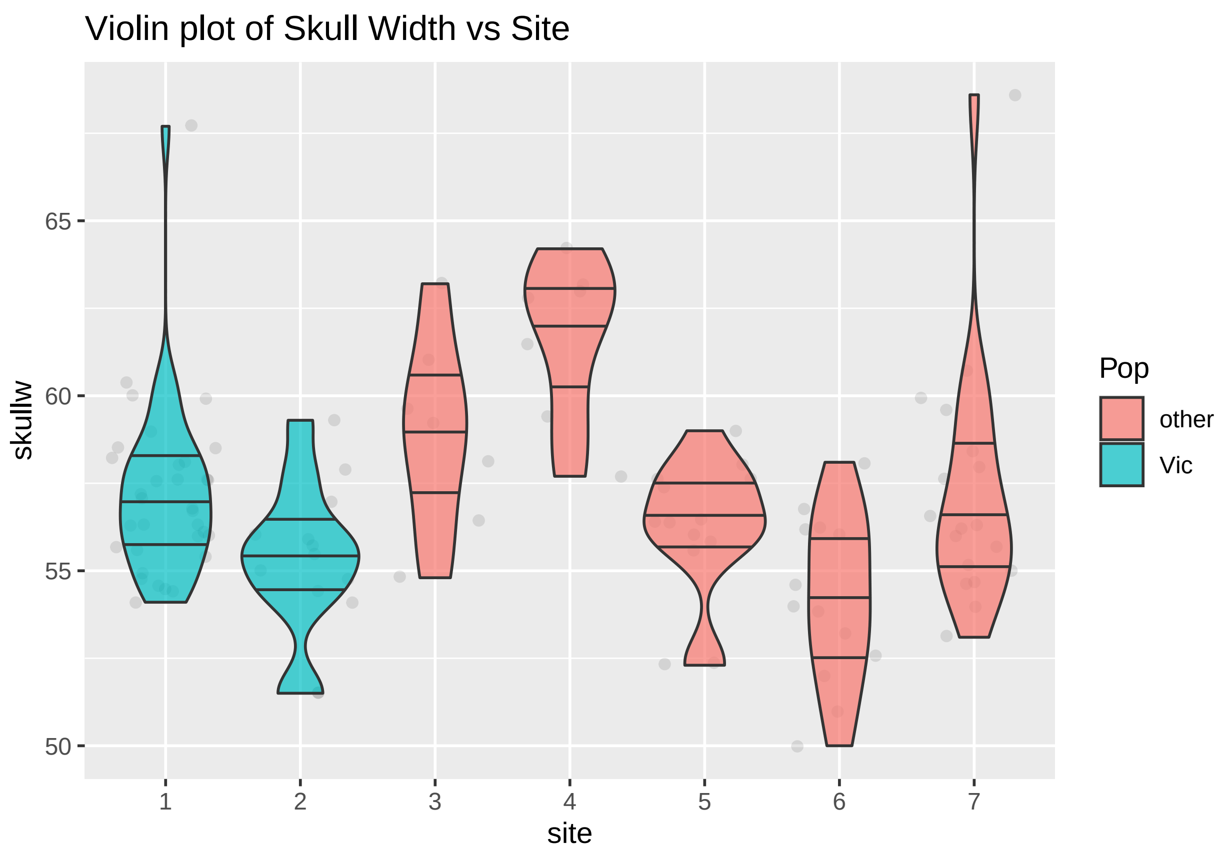


What is the assumption we are violating here for linear model of Gaussian Family ?

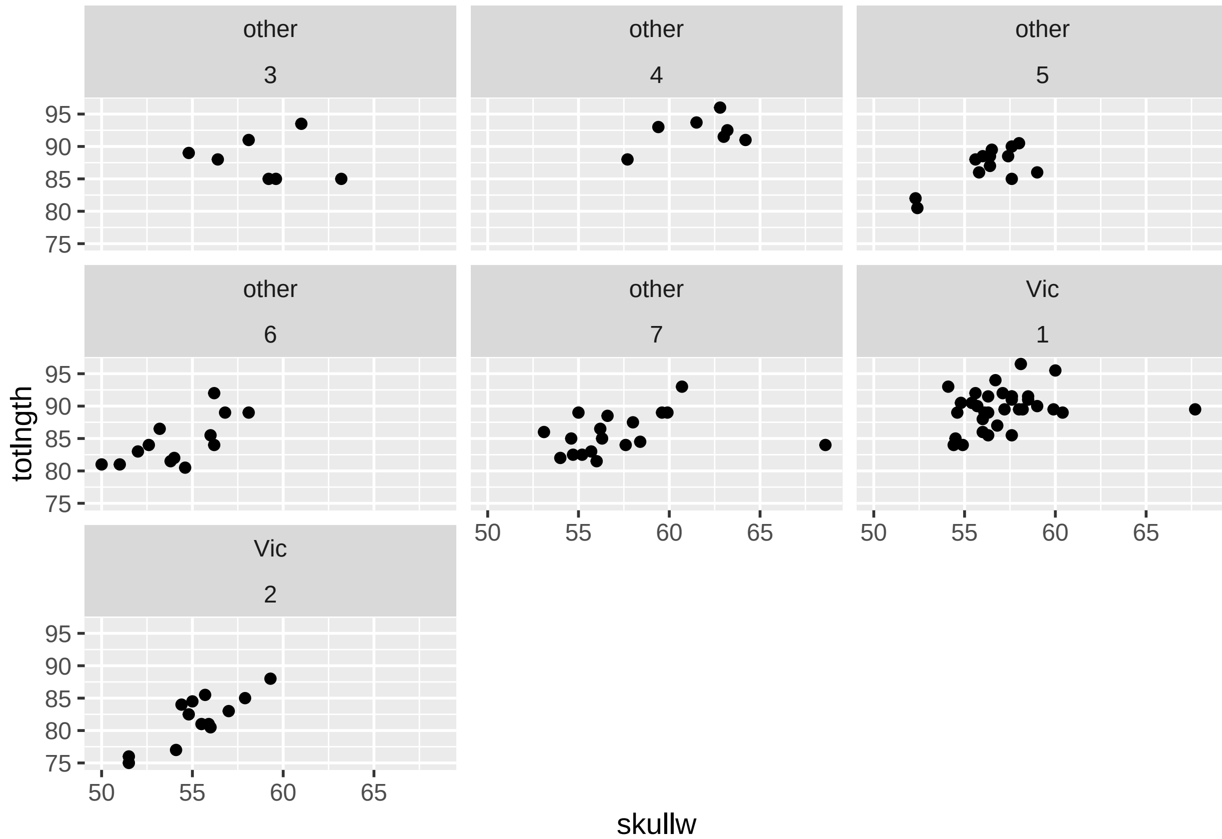
- Homoscedasticity: The variance of residual is the same for any value of X.
- what about "Independence: Observations are independent of each other."

Let us check

```
ggplot(possum, aes(x = site, y = skullw)) +
  geom_jitter(alpha = .1) +
  geom_violin(alpha = 0.7,aes(fill=Pop),draw_quantiles = c(0.25, 0.5, 0.75)) +
  ggtitle("Violin plot of Skull Width vs Site")
```




```
#scatter plot
scat_plot <- ggplot(data = possum, aes(x=skullw,y=totlngth))
scat_plot + geom_point()+facet_wrap(.~Pop+site)
```



- **the data is not independent**

As we can clearly see from the data collected the data is not independent but nested in the nature.

Moreover, the data is taken from 2 different population. In addition there are several site nested for each population.

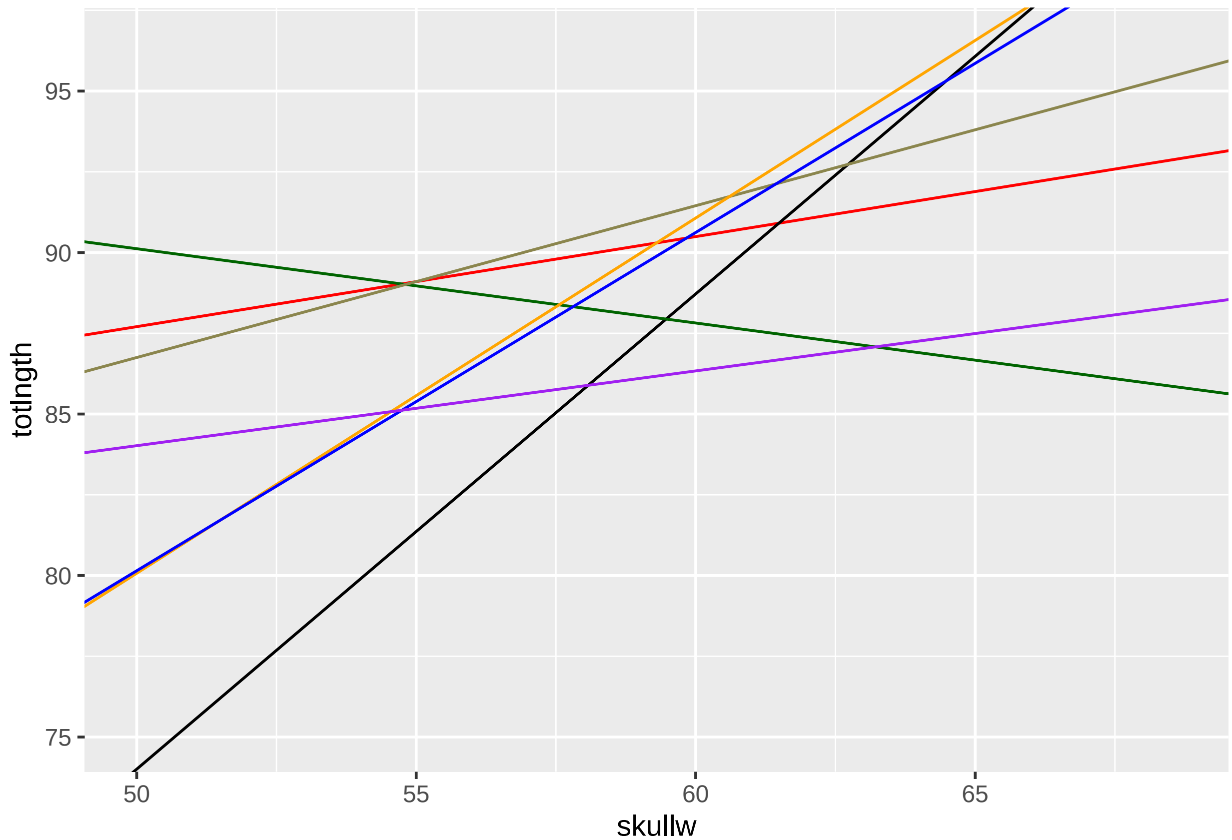
This is called nested data.

One traditional way (wrong way)

Let make model of all the site individually.

```
site<-unique(possum$site)
Beta <- vector(length = length(site))
alpha <- vector(length = length(site))
for(k in 1:length(Beta)){
  flag_mod <- lm(totlngth~skullw,data = possum[possum$site==site[k],])
  alpha[k] <- as.numeric(flag_mod$coefficients[[1]])
  Beta[k] <- as.numeric(flag_mod$coefficients[[2]])
}
##let plot regression for each site.
ggplot(possum,aes(x=skullw,y=totlngth))+geom_blank()+
  geom_abline(slope = Beta[1],intercept = alpha[1],col="red")+
  geom_abline(slope = Beta[2],intercept = alpha[2],col="black")+
```

```r
  geom_abline(slope = Beta[3],intercept = alpha[3],col="darkgreen")+
  geom_abline(slope = Beta[4],intercept = alpha[4],col="khaki4")+
  geom_abline(slope = Beta[5],intercept = alpha[5],col="orange")+
  geom_abline(slope = Beta[6],intercept = alpha[6],col="blue")+
  geom_abline(slope = Beta[7],intercept = alpha[7],col="purple")
```



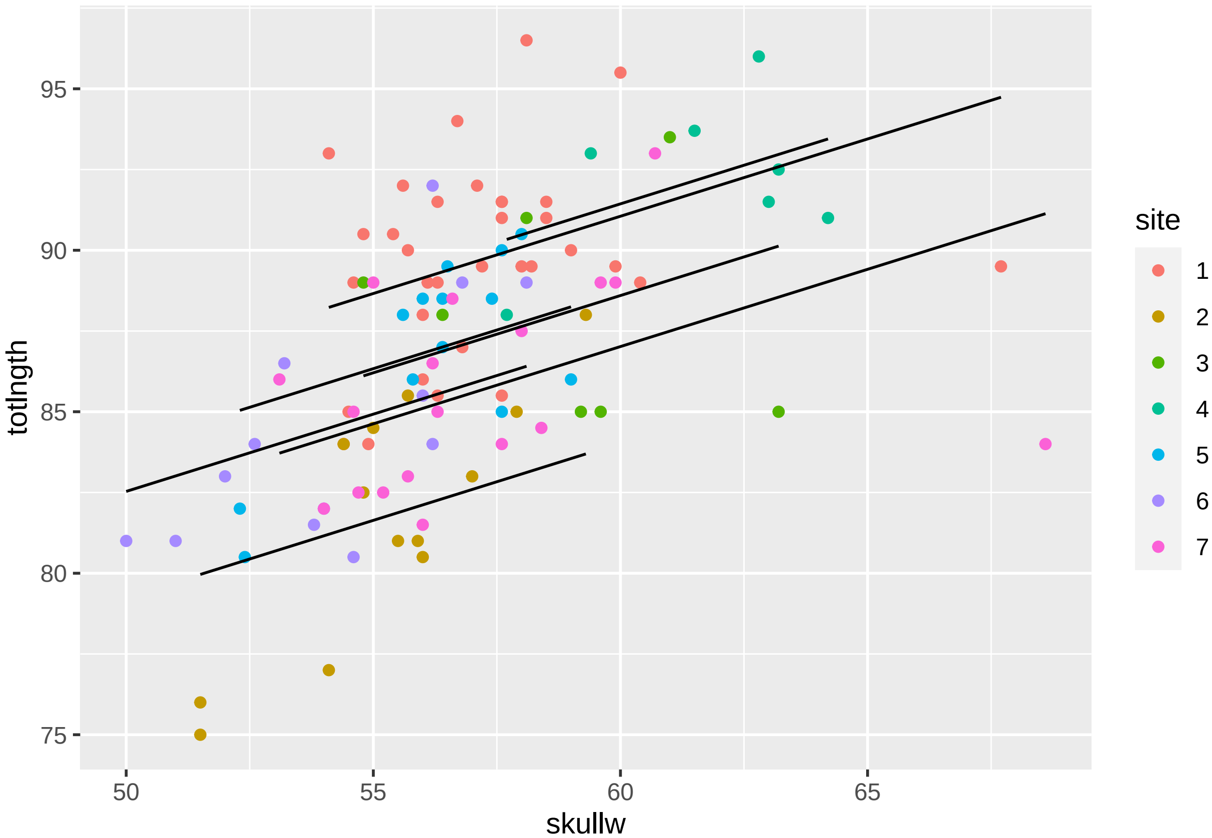
Less traditional way (method in this case)

It to account for variation in site.

```r
mod_account_site <- lm(totlngth~skullw + site ,data = possum)
AIC(mod_account_site,mod_lm)
```

```
##                  df      AIC
## mod_account_site  9 531.9851
## mod_lm            3 570.2881
```

```r
#AIC value have decrease but What is interpretation of this model ?
pred_value <- as.data.frame(predict(mod_account_site,possum,se.fit = T))
pred_value <- cbind(pred_value,possum)
ggplot(data = pred_value,aes(x=skullw,y=totlngth))+
  geom_point(aes(col=site))+
  geom_line(aes(x=skullw,y = fit,group=site))
```



Now it is clear that for some random reason the behavior of data is dependent on site.

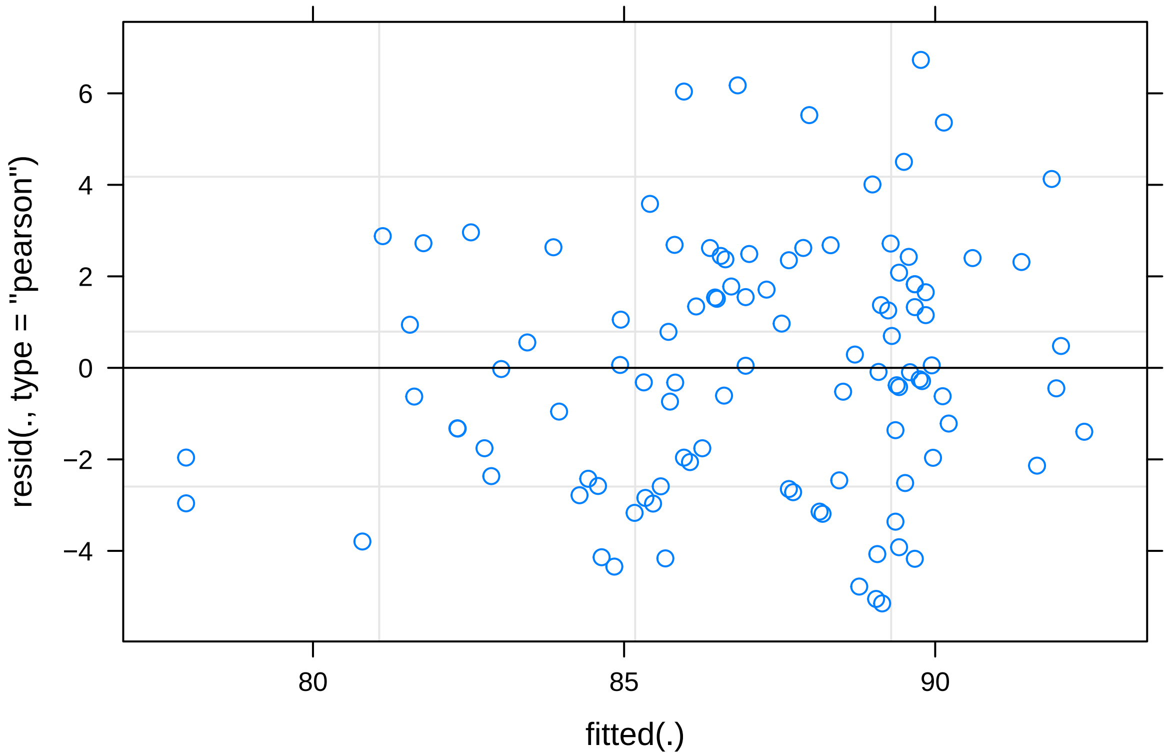
How to solve such problem

- One problem in using model $y = \alpha + \beta * X1 + \beta * X2$ is

```
model_LMM <- lmer(totlngth ~ skullw + 1 | (Pop:site),data = possum)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.998677 (tol = 0.002, component 1)
```

```
plot(model_LMM)
```




```
## clearly intercept are different.
dotplot(ranef(model_LMM, which = "Pop|site", condVar = TRUE))
```

```
## named list()
```

```
## let us see the model
print(model_LMM, corr = FALSE)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: totlngth ~ skullw + 1 | (Pop:site)
##    Data: possum
## REML criterion at convergence: 536.628
## Random effects:
##  Groups     Name        Std.Dev. Corr
##  (Pop:site) (Intercept) 28.6458
##             skullw       0.4784  -1.00
##  Residual                2.9031
## Number of obs: 104, groups:  (Pop:site), 7
## Fixed Effects:
## (Intercept)
##       88.55
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warnings
```

```r
## ##getting SE and tabular format for fixed effect.
se <- sqrt(diag(vcov(model_LMM)))
# table of estimates with 95% CI
(tab <- cbind(Est = fixef(model_LMM),
              LL = fixef(model_LMM) - 1.96 * se,
              UL = fixef(model_LMM) + 1.96 *se))
```

```
##                  Est       LL       UL
## (Intercept) 88.54983 86.91974 90.17992
```