Final Data Analysis

Yibo Yang

2019/12/7

Data preprocessing

The data set measures different variables for horses with lesions in the abdomen/digestive tract. Our goal is to model the variable respiratoryrate.

```
hrsdt = read.table("horse_data.txt", header = TRUE)
head(hrsdt)
```

```
##   respiratoryrate pulse temperature pain cellvolume totalprotein age
## 1              28    66        38.5    5         45          8.4   1
## 2              20    88        39.2    3         50         85.0   1
## 3              24    40        38.3    3         33          6.7   1
## 4              84   164        39.1    2         48          7.2   2
## 5              35   104        37.3   NA         74          7.4   1
## 6              NA    NA          NA    2         NA           NA   1
##   abdominaldistension
## 1                   4
## 2                   2
## 3                   1
## 4                   4
## 5                  NA
## 6                   2
```

There are 7 covariates available. Four of them ("pulse", "temperature", "cellvolume", "totalprotein") are numerical, and three of them ("pain", "age", "andominaldistension") are categorical.

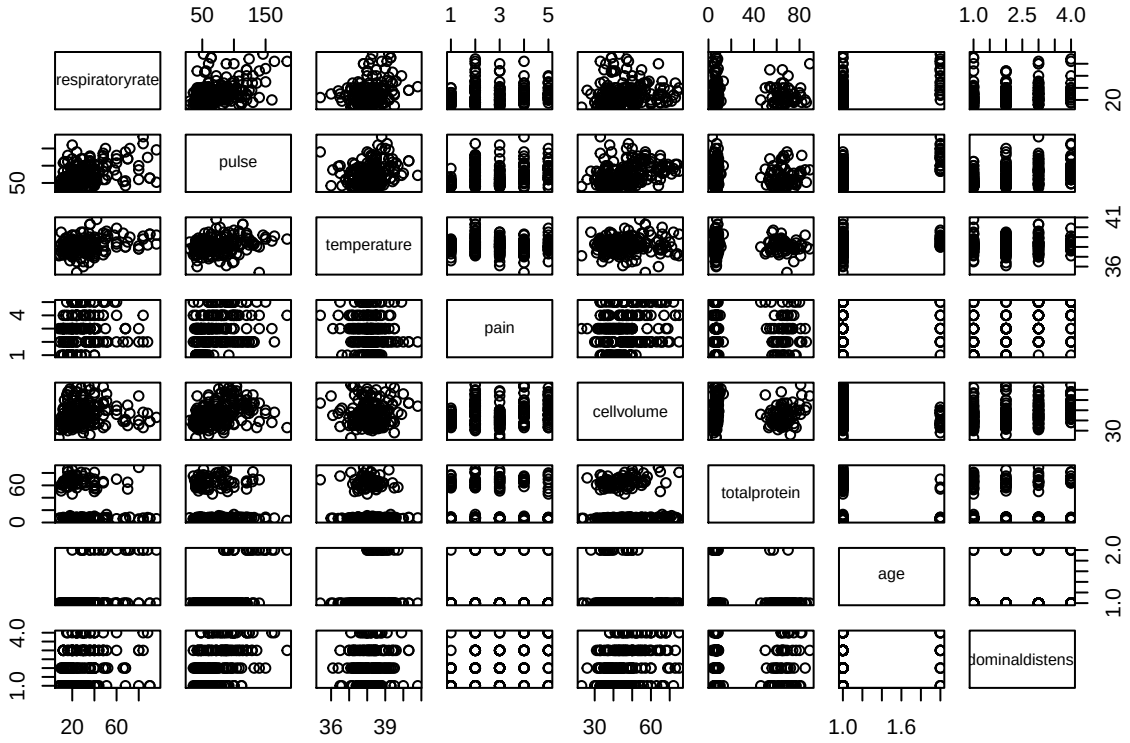
```
hrsdt$pain = as.factor(hrsdt$pain)
hrsdt$age = as.factor(hrsdt$age)
hrsdt$abdominaldistension = as.factor(hrsdt$abdominaldistension)

summary(hrsdt)
```

```
##  respiratoryrate     pulse          temperature      pain      cellvolume
##  Min.   : 8.00   Min.   : 30.00   Min.   :35.40   1   :38   Min.   :23.0
##  1st Qu.:18.50   1st Qu.: 48.00   1st Qu.:37.80   2   :59   1st Qu.:38.0
##  Median :24.50   Median : 64.00   Median :38.20   3   :67   Median :45.0
##  Mean   :30.42   Mean   : 71.91   Mean   :38.17   4   :39   Mean   :46.3
##  3rd Qu.:36.00   3rd Qu.: 88.00   3rd Qu.:38.50   5   :42   3rd Qu.:52.0
##  Max.   :96.00   Max.   :184.00   Max.   :40.80   NA's:55   Max.   :75.0
##  NA's   :58      NA's   :24       NA's   :60                NA's   :29
##   totalprotein   age     abdominaldistension
##  Min.   : 3.30   1:276   1   :76
##  1st Qu.: 6.50   2: 24   2   :65
##  Median : 7.50           3   :65
##  Mean   :24.46           4   :38
##  3rd Qu.:57.00           NA's:56
##  Max.   :89.00
##  NA's   :33
```

There are 300 observations in the data set. From the summary we know that except “age”, each variable has enties with missing values.

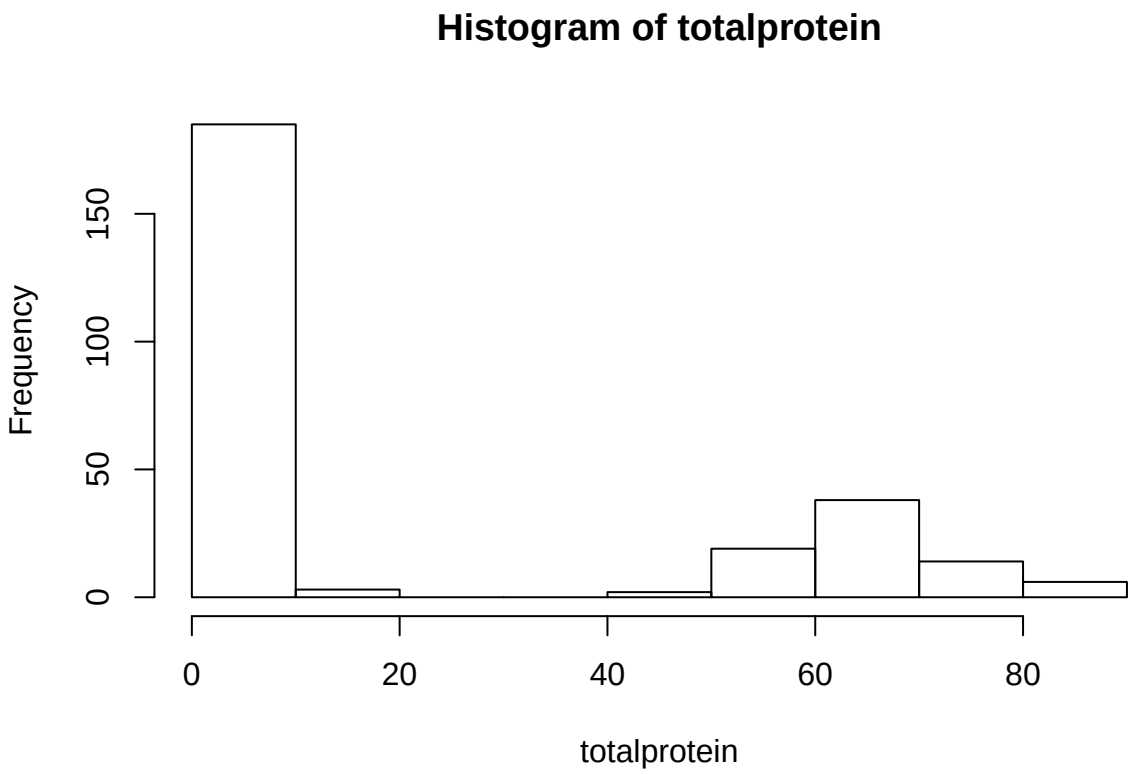
Use scatterplot matrix to take a first look of the data.

```
pairs(hrsdt)
```



We can see there are two clusters in variable “totalprotein”, plot the histogram for “totalprotein”.

```
hist(hrsdt$totalprotein, xlab = "totalprotein",main = "Histogram of totalprotein")
```

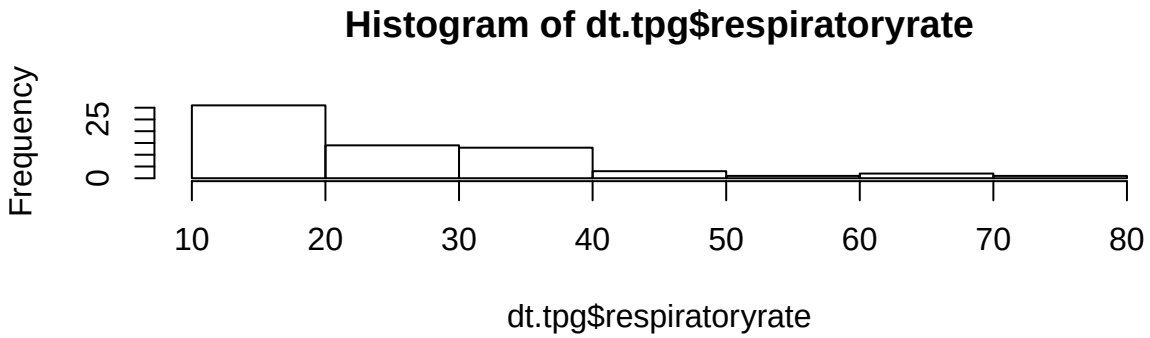


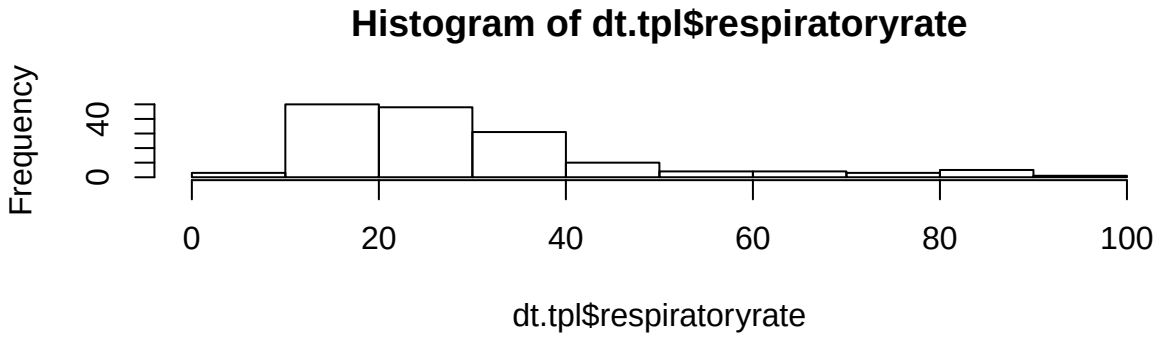
Discard observations that have missing value on "totalprotein", and seperate data set into two parts with "totalprotein" larger or smaller than 30.

```
dt.cpttp = hrsdt[!is.na(hrsdt$totalprotein),]
dt.tpg = dt.cpttp[dt.cpttp$totalprotein>30,] ## Data with totalprotein > 30
dt.tpl = dt.cpttp[dt.cpttp$totalprotein<30,] ## Data with totalprotein < 30
```

Examine the distribution of the response variable "respiratoryrate"

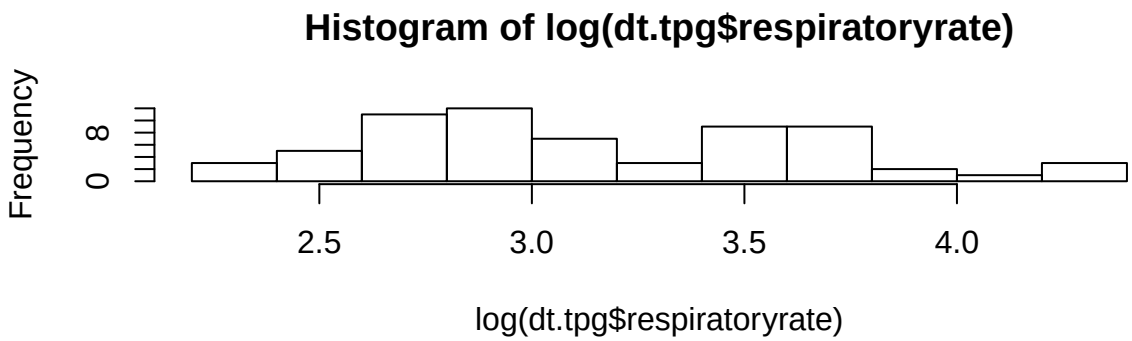
```
par(mfrow=c(2,1))
hist(dt.tpg$respiratoryrate)
hist(dt.tpl$respiratoryrate)
```

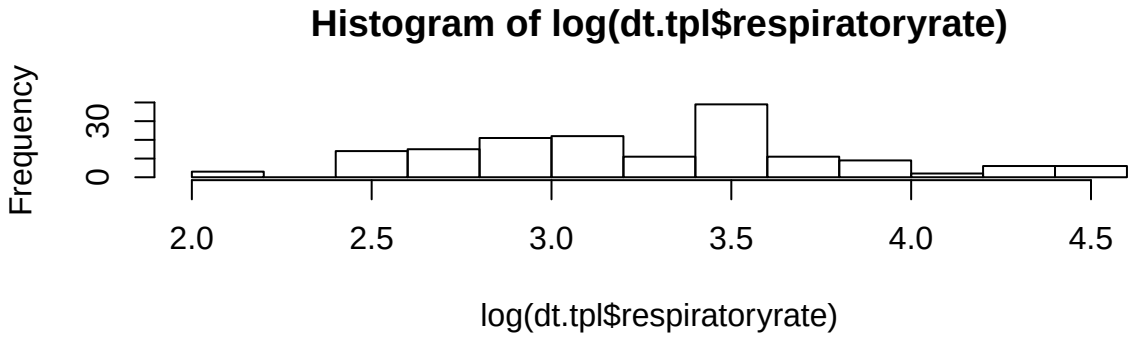




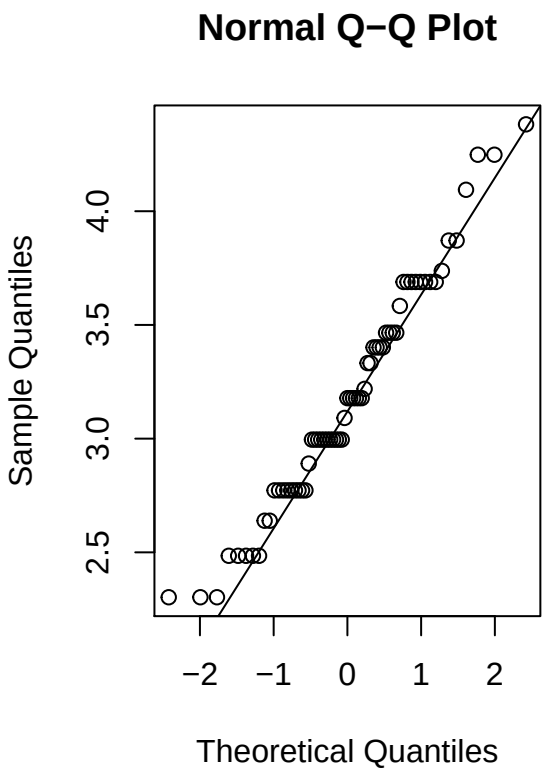
The distribution of "respiratoryrate" is skewed, take log transformation of "respiratoryrate".

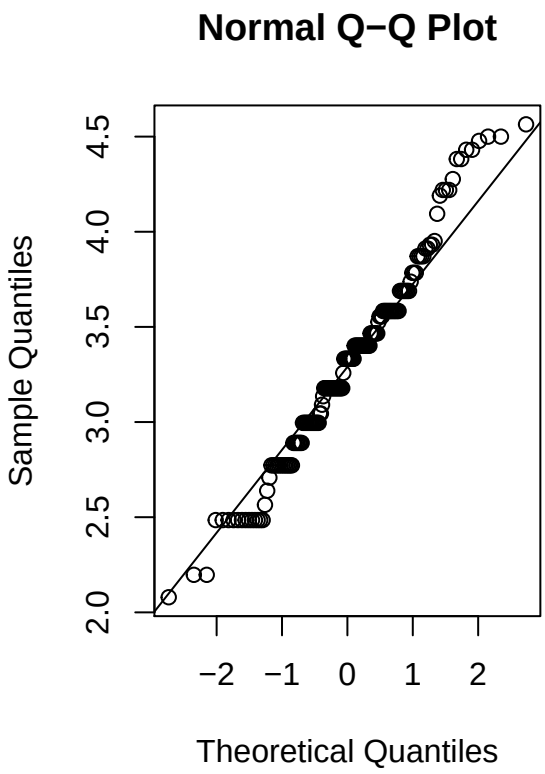
```
par(mfrow=c(2,1))
hist(log(dt.tpg$respiratoryrate))
hist(log(dt.tpl$respiratoryrate))
```






```
par(mfrow=c(1,2))
qqnorm(log(dt.tpg$respiratoryrate))
qqline(log(dt.tpg$respiratoryrate))
qqnorm(log(dt.tpl$respiratoryrate))
qqline(log(dt.tpl$respiratoryrate))
```



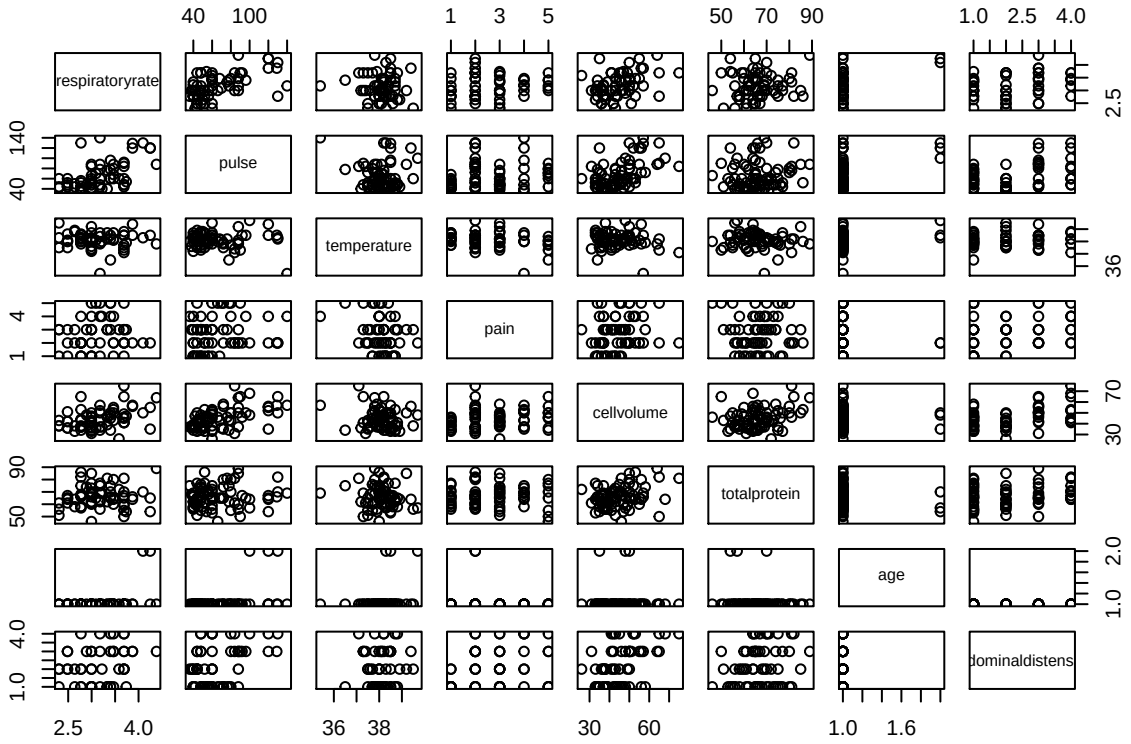


After transformation the distribution is approximately normal.

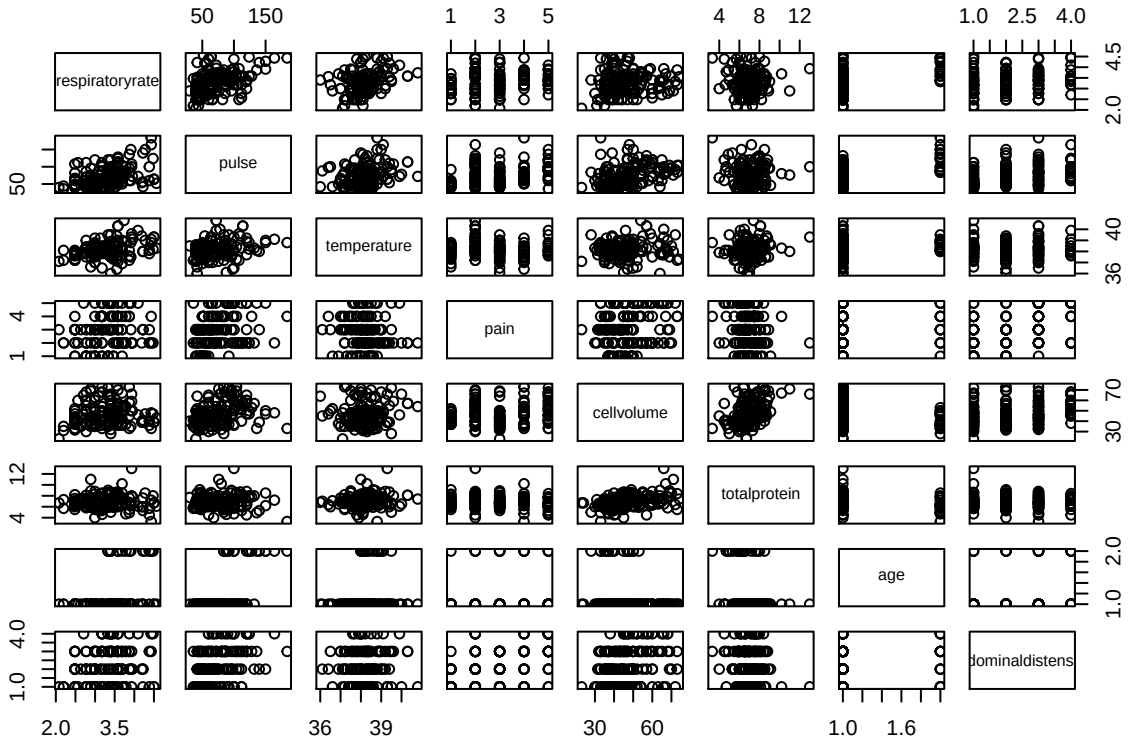
```
dt.lgtpg = dt.tpg
dt.lgtpl = dt.tpl
dt.lgtpg$respiratoryrate = log(dt.lgtpg$respiratoryrate)
dt.lgtpl$respiratoryrate = log(dt.lgtpl$respiratoryrate)
```

Scatterplot matrix and summary for two data sets.

```
pairs(dt.lgtpg)
```




```
pairs(dt.lgtpl)
```




```
summary(dt.lgtpl)
```

```
##  respiratoryrate     pulse          temperature     pain      cellvolume
##  Min.   :2.079   Min.   : 30.00   Min.   :36.0   1   :23   Min.   :23.00
##  1st Qu.:2.996   1st Qu.: 50.00   1st Qu.:37.8   2   :37   1st Qu.:39.00
##  Median :3.332   Median : 66.00   Median :38.2   3   :47   Median :45.00
##  Mean   :3.296   Mean   : 73.08   Mean   :38.2   4   :22   Mean   :46.78
##  3rd Qu.:3.584   3rd Qu.: 90.00   3rd Qu.:38.6   5   :29   3rd Qu.:52.00
##  Max.   :4.564   Max.   :184.00   Max.   :40.8   NA's:30   Max.   :74.00
##  NA's   :29      NA's   :15       NA's   :37             NA's   :3
##   totalprotein    age     abdominaldistension
##  Min.   : 3.300   1:169   1   :51
##  1st Qu.: 6.300   2: 19   2   :47
##  Median : 6.850           3   :48
##  Mean   : 6.957           4   :18
##  3rd Qu.: 7.525           NA's:24
##  Max.   :13.000
##
```

```
summary(dt.lgtpg)
```

```
##  respiratoryrate     pulse          temperature      pain
##  Min.   :2.303   Min.   : 36.00   Min.   :35.40   1   :14
##  1st Qu.:2.773   1st Qu.: 48.00   1st Qu.:37.80   2   :19
##  Median :3.178   Median : 60.00   Median :38.15   3   :17
##  Mean   :3.176   Mean   : 66.83   Mean   :38.11   4   : 9
##  3rd Qu.:3.466   3rd Qu.: 80.00   3rd Qu.:38.50   5   : 8
```

```
##  Max.   :4.382   Max.   :140.00   Max.   :39.70   NA's:12
##  NA's   :14      NA's   :2        NA's   :11
##    cellvolume     totalprotein  age    abdominaldistension
##  Min.   :26.00   Min.   :46.0   1:76   1   :24
##  1st Qu.:37.00   1st Qu.:60.0   2: 3   2   :14
##  Median :43.00   Median :65.0          3   :14
##  Mean   :44.62   Mean   :66.1          4   :10
##  3rd Qu.:50.00   3rd Qu.:70.5          NA's:17
##  Max.   :75.00   Max.   :89.0
##
```

Imputation

Discard observations with missing values in categorical variables("pain", "age", "abdominaldistension"), and impute missing values with regression iteratively.

```
dt.clgtpl = dt.lgtpl[complete.cases(dt.lgtpl[,c("pain","age","abdominaldistension")]),]
dt.clgtpg = dt.lgtpg[complete.cases(dt.lgtpg[,c("pain","age","abdominaldistension")]),]
```

Summary of data with totalprotein < 30

```
summary(dt.clgtpl)
```

```
##  respiratoryrate     pulse        temperature   pain     cellvolume
##  Min.   :2.079   Min.   : 36.0   Min.   :36.1   1:22   Min.   :23.00
##  1st Qu.:2.996   1st Qu.: 52.0   1st Qu.:37.8   2:36   1st Qu.:39.00
##  Median :3.332   Median : 66.0   Median :38.2   3:44   Median :45.00
##  Mean   :3.294   Mean   : 74.4   Mean   :38.2   4:22   Mean   :46.93
##  3rd Qu.:3.584   3rd Qu.: 92.0   3rd Qu.:38.6   5:28   3rd Qu.:53.00
##  Max.   :4.477   Max.   :184.0   Max.   :40.8          Max.   :73.00
##  NA's   :22      NA's   :11      NA's   :28            NA's   :3
##   totalprotein    age     abdominaldistension
##  Min.   : 3.300   1:138   1:43
##  1st Qu.: 6.200   2: 14   2:44
##  Median : 6.800           3:47
##  Mean   : 6.882           4:18
##  3rd Qu.: 7.500
##  Max.   :13.000
##
```

From summary we know that variables "pulse", "temperature" and "cellvolume" have missing values.

Impute data with "totalprotein" < 30

```
dt.l = dt.clgtpl
pu.na = is.na(dt.l$pulse)
te.na = is.na(dt.l$temperature)
ce.na = is.na(dt.l$cellvolume)
# Initialize missing values with mean of each variable
dt.l$pulse[pu.na] = mean(dt.l$pulse[!pu.na])
dt.l$temperature[te.na] = mean(dt.l$temperature[!te.na])
dt.l$cellvolume[ce.na] = mean(dt.l$cellvolume[!ce.na])
delta = 10
n=0
while (delta > 0.001) {
  n = n+1
  pu = predict(lm(pulse~(temperature+cellvolume+totalprotein)*(pain+age+abdominaldistension),
```

```
                  data = dt.l), dt.l[pu.na,])
  te = predict(lm(temperature~(pulse+cellvolume+totalprotein)*(pain+age+abdominaldistension),
                  data = dt.l), dt.l[te.na,])
  ce = predict(lm(cellvolume~(pulse+temperature+totalprotein)*(pain+age+abdominaldistension),
                  data = dt.l), dt.l[ce.na,])
  delta = mean(mean((dt.l$pulse[pu.na]-pu)^2)+
               mean((dt.l$temperature[te.na]-te)^2)+
               mean((dt.l$cellvolume[ce.na]-ce)^2))
  dt.l$pulse[pu.na] = pu
  dt.l$temperature[te.na] = te
  dt.l$cellvolume[ce.na] = ce
}
cat("Number of iteration is", n, "\n")
```

```
## Number of iteration is 20
```

Summary of data with "totalprotein" > 30

```
summary(dt.clgtpg)
```

```
##  respiratoryrate     pulse          temperature    pain     cellvolume
##  Min.   :2.303   Min.   : 36.00   Min.   :36.50   1:10   Min.   :26.00
##  1st Qu.:2.773   1st Qu.: 48.00   1st Qu.:37.77   2:14   1st Qu.:38.00
##  Median :3.178   Median : 60.00   Median :38.10   3:17   Median :43.00
##  Mean   :3.146   Mean   : 63.94   Mean   :38.10   4: 7   Mean   :44.35
##  3rd Qu.:3.466   3rd Qu.: 79.50   3rd Qu.:38.50   5: 7   3rd Qu.:49.50
##  Max.   :3.871   Max.   :130.00   Max.   :39.50          Max.   :75.00
##  NA's   :10      NA's   :1        NA's   :7
##   totalprotein  age    abdominaldistension
##  Min.   :46.0   1:55   1:23
##  1st Qu.:61.0   2: 0   2:13
##  Median :67.0          3:11
##  Mean   :66.8          4: 8
##  3rd Qu.:72.5
##  Max.   :85.0
##
```

From summary we know that variables "pulse" and "temperature" have missing values. Also note that after removing observations with missing value in categorical variables, there is no juvenile horse("age"=2) left in this data set, therefore "age" is not included in the following regression of this dataset. This also means that the results we get from this data set are only valid for horses with "totalprotein" > 30 and "age" $= 1$. We need more information and data to fit model for horses with "totalprotein" > 30 and "age" $= 2$".

Impute data with "totalprotein" > 30

```
dt.g = dt.clgtpg
pu.na = is.na(dt.g$pulse)
te.na = is.na(dt.g$temperature)
dt.g$pulse[pu.na] = mean(dt.g$pulse[!pu.na])
dt.g$temperature[te.na] = mean(dt.g$temperature[!te.na])
delta = 10
n=0
while (delta > 0.001) {
  n = n+1
  pu = predict(lm(pulse~(temperature+cellvolume+totalprotein)*(pain+abdominaldistension),
                  data = dt.g), dt.g[pu.na,])
```

```
  te = predict(lm(temperature~(pulse+cellvolume+totalprotein)*(pain+abdominaldistension),
                  data = dt.g), dt.g[te.na,])
  delta = mean(mean((dt.g$pulse[pu.na]-pu)^2)+
                 mean((dt.g$temperature[te.na]-te)^2))
  dt.g$pulse[pu.na] = pu
  dt.g$temperature[te.na] = te
}
cat("Number of iteration is", n, "\n")
```

```
## Number of iteration is 44
```

Discard observations with missing values in response variable "respiratoryrate".

```
dt.l = dt.l[!is.na(dt.l$respiratoryrate),]
dt.g = dt.g[!is.na(dt.g$respiratoryrate),]
summary(dt.l)
```

```
##  respiratoryrate     pulse          temperature    pain     cellvolume
##  Min.   :2.079   Min.   : 36.00   Min.   :36.40   1:19   Min.   :23.00
##  1st Qu.:2.996   1st Qu.: 50.00   1st Qu.:37.80   2:33   1st Qu.:39.00
##  Median :3.332   Median : 66.00   Median :38.13   3:37   Median :45.00
##  Mean   :3.294   Mean   : 73.51   Mean   :38.19   4:21   Mean   :46.61
##  3rd Qu.:3.584   3rd Qu.: 91.50   3rd Qu.:38.50   5:20   3rd Qu.:52.00
##  Max.   :4.477   Max.   :184.00   Max.   :40.80          Max.   :73.00
##   totalprotein    age     abdominaldistension
##  Min.   : 3.300   1:117   1:38
##  1st Qu.: 6.200   2: 13   2:37
##  Median : 6.900           3:39
##  Mean   : 6.878           4:16
##  3rd Qu.: 7.500
##  Max.   :13.000
```

```
summary(dt.g)
```

```
##  respiratoryrate     pulse          temperature   pain     cellvolume
##  Min.   :2.303   Min.   : 36.00   Min.   :36.5   1: 7   Min.   :26.00
##  1st Qu.:2.773   1st Qu.: 44.00   1st Qu.:37.9   2:13   1st Qu.:37.00
##  Median :3.178   Median : 56.00   Median :38.2   3:15   Median :43.00
##  Mean   :3.146   Mean   : 62.63   Mean   :38.2   4: 5   Mean   :43.93
##  3rd Qu.:3.466   3rd Qu.: 78.00   3rd Qu.:38.5   5: 5   3rd Qu.:48.00
##  Max.   :3.871   Max.   :130.00   Max.   :40.4          Max.   :75.00
##   totalprotein   age    abdominaldistension
##  Min.   :46.00   1:45   1:16
##  1st Qu.:61.00   2: 0   2:13
##  Median :66.00          3:11
##  Mean   :66.42          4: 5
##  3rd Qu.:72.00
##  Max.   :85.00
```

Model

Contingency table of "pain"(row) and "abdominaldistension"(column) for horse with "totalprotein" < 30 and "age" $= 1$ (adult).

```
addmargins(table(dt.l[dt.l$age==1,]$pain,dt.l[dt.l$age==1,]$abdominaldistension))
```

```
##
```

```
##          1   2   3   4 Sum
##   1     12   5   1   0  18
##   2     10  11   8   1  30
##   3      8  13  12   0  33
##   4      5   4   6   3  18
##   5      1   1   7   9  18
##   Sum   36  34  34  13 117
```

Contingency table of "pain"(row) and "abdominaldistension"(column) for horse with "totalprotein" < 30 and "age" $= 2$ (juvenile).

```
addmargins(table(dt.l[dt.l$age==2,]$pain,dt.l[dt.l$age==2,]$abdominaldistension))
```

```
##
##        1  2  3  4 Sum
##   1    0  1  0  0   1
##   2    0  0  0  3   3
##   3    1  1  2  0   4
##   4    1  0  2  0   3
##   5    0  1  1  0   2
##   Sum  2  3  5  3  13
```

Note that data are not evenly distributed among cells, and for horse with "age" = 2 (juvenile), there is not enough data to fit model for each category, therefore we fit model seperately without interacation term for this subset.

```
dt.lj = dt.l[dt.l$age==2,]
dt.la = dt.l[dt.l$age==1,]
dt.ga = dt.g

cat("The number of observations in for juvenile horse with totalprotein < 30 is", nrow(dt.lj), ".\n")
```

```
## The number of observations in for juvenile horse with totalprotein < 30 is 13 .
```

```
cat("The number of observations in for adult horse with totalprotein < 30 is", nrow(dt.la), ".\n")
```

```
## The number of observations in for adult horse with totalprotein < 30 is 117 .
```

```
cat("The number of observations in for adult horse with totalprotein > 30 is", nrow(dt.ga), ".\n")
```

```
## The number of observations in for adult horse with totalprotein > 30 is 45 .
```

As sample sizes are small, we do not hold out testing set to compare different models. Making decisions based on even smaller sample size is questionable, as small sample size leads to large variance of estimated coefficients. Also, test for outliers, infuential points, leverage points may not be valid when sample size is small, and we would only make adjustment for obvious outliers. More accuate models may need more data and information.

Model for juvenile horse("age"=2) withtotalprotein < 30

There are 13 observations in this data set.

```
md.lj1 = lm(respiratoryrate ~ pulse+temperature+cellvolume+totalprotein, data=dt.lj)
summary(md.lj1)
```

```
##
## Call:
## lm(formula = respiratoryrate ~ pulse + temperature + cellvolume +
##     totalprotein, data = dt.lj)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45697 -0.17532  0.01635  0.14848  0.56363
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.779497  11.512885   1.371    0.208
## pulse         0.014170   0.004719   3.003    0.017 *
## temperature  -0.350349   0.317940  -1.102    0.303
## cellvolume   -0.033543   0.033962  -0.988    0.352
## totalprotein  0.206624   0.156894   1.317    0.224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3396 on 8 degrees of freedom
## Multiple R-squared:  0.5499, Adjusted R-squared:  0.3249
## F-statistic: 2.444 on 4 and 8 DF,  p-value: 0.1313
```

```
md.lj2 = lm(respiratoryrate ~ pulse, data=dt.lj)
anova(md.lj2,md.lj1)
```

```
## Analysis of Variance Table
##
## Model 1: respiratoryrate ~ pulse
## Model 2: respiratoryrate ~ pulse + temperature + cellvolume + totalprotein
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     11 1.47665
## 2      8 0.92245  3    0.5542 1.6021 0.2638
```

The p-value is large, and we fail to reject the null hypothesis that two models are not different.

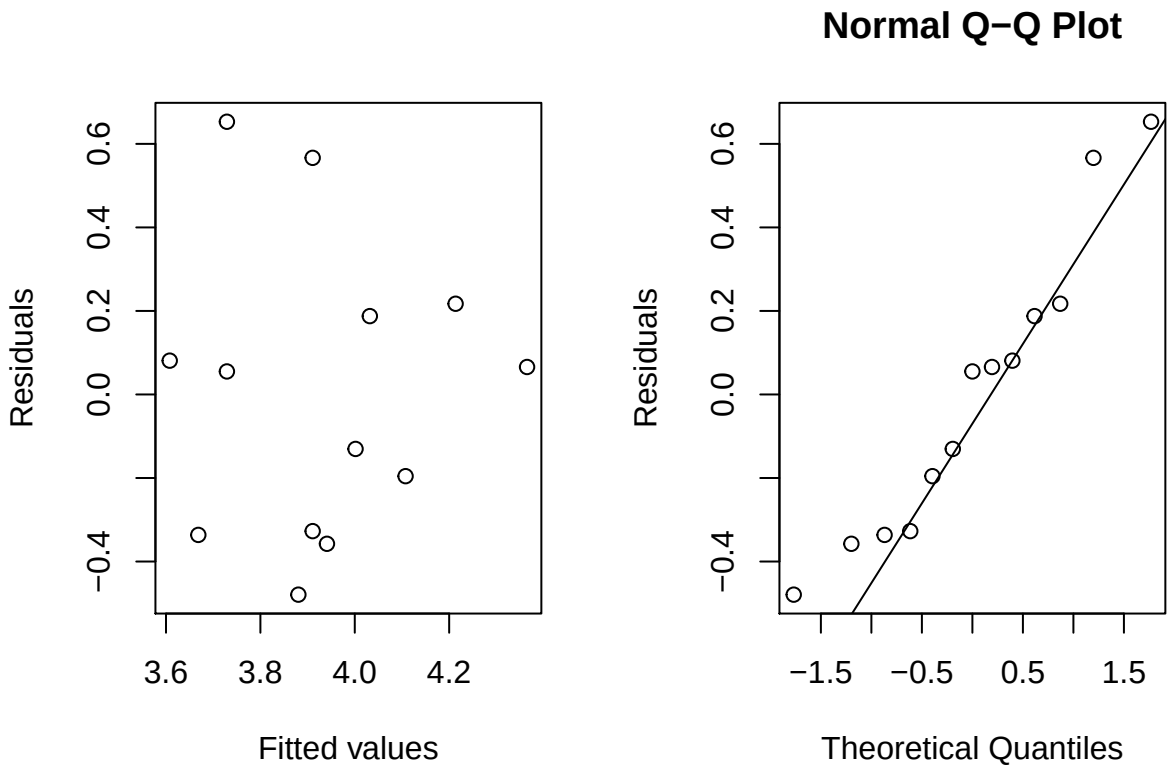
```
summary(md.lj2)
```

```
##
## Call:
## lm(formula = respiratoryrate ~ pulse, data = dt.lj)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47927 -0.32724  0.05515  0.18761  0.65298
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.971907   0.475016   6.256  6.2e-05 ***
## pulse       0.007571   0.003665   2.066   0.0632 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3664 on 11 degrees of freedom
## Multiple R-squared:  0.2795, Adjusted R-squared:  0.2141
## F-statistic: 4.268 on 1 and 11 DF,  p-value: 0.06321
```

Test for constant error and normality of errors.

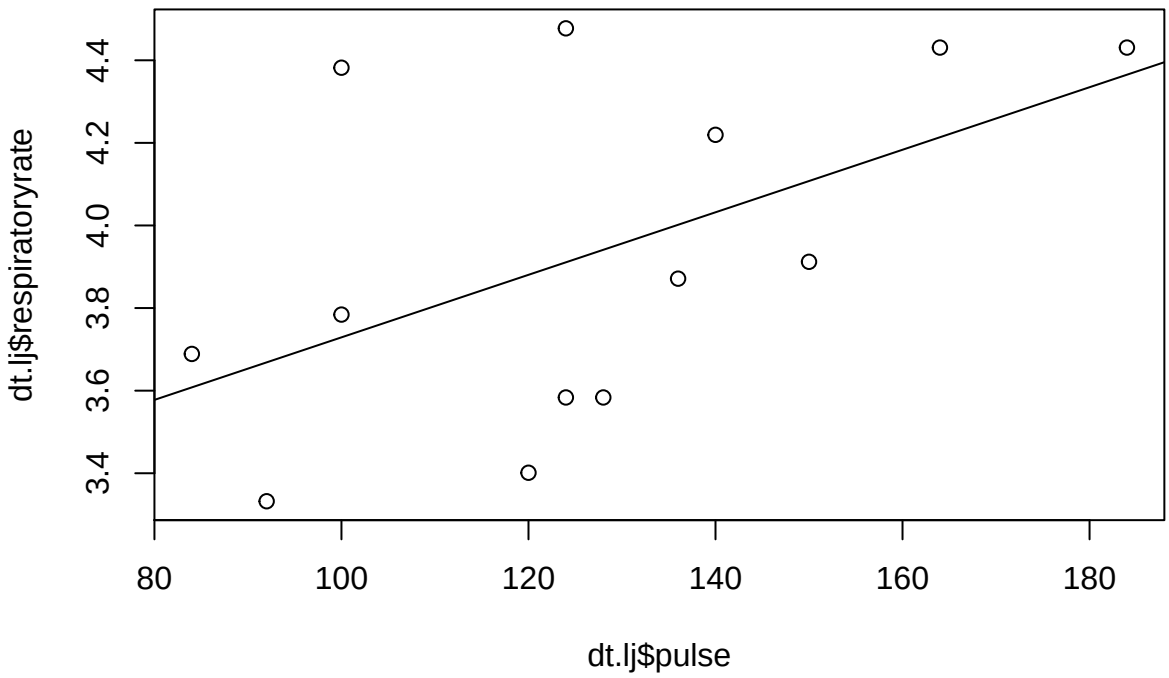
```
par(mfrow=c(1,2))
plot(md.lj2$fitted.values,md.lj2$residuals, xlab = "Fitted values", ylab = "Residuals")
```

```
qqnorm(md.lj2$residuals, ylab = "Residuals")
qqline(md.lj2$residuals)
```



From qq plot, the residuals are approximately normal. In the plot, it seems that residuals have larger variance when fitted value is small, but this may due to less data points when fitted value is large. There is not enough data to reach to a conclusion.

```
plot(dt.lj$pulse,dt.lj$respiratoryrate)
abline(md.lj2)
```



As there there are only 13 data points, we stop here and do not do further test on outliers, infuential points, leverage points etc.. More data is needed for further investigation and modeling.

Model for adult horse("age"=1) withtotalprotein < 30

There are 117 observations in this data set.

```
md.la1 = lm(respiratoryrate ~ (pulse+temperature+cellvolume+totalprotein)*(abdominaldistension+pain), da
anova(md.la1)
```

```
## Analysis of Variance Table
##
## Response: respiratoryrate
##                                  Df  Sum Sq Mean Sq F value  Pr(>F)
## pulse                             1  1.5530 1.55304  7.2672 0.00862 **
## temperature                       1  0.9793 0.97933  4.5826 0.03546 *
## cellvolume                        1  0.0624 0.06237  0.2918 0.59060
## totalprotein                      1  0.0038 0.00379  0.0177 0.89439
## abdominaldistension               3  1.2736 0.42455  1.9866 0.12297
## pain                              4  0.7873 0.19683  0.9210 0.45617
## pulse:abdominaldistension         3  0.0787 0.02624  0.1228 0.94641
## pulse:pain                        4  0.3528 0.08819  0.4127 0.79899
## temperature:abdominaldistension   3  0.5651 0.18836  0.8814 0.45459
## temperature:pain                  4  1.3857 0.34642  1.6210 0.17748
## cellvolume:abdominaldistension    3  0.7215 0.24050  1.1254 0.34411
## cellvolume:pain                   4  0.5722 0.14304  0.6693 0.61522
## totalprotein:abdominaldistension  3  0.3079 0.10263  0.4802 0.69698
```

```
## totalprotein:pain                4  0.6751 0.16877  0.7897 0.53539
## Residuals                       77 16.4553 0.21370
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
md.la2 = lm(respiratoryrate ~ pulse+temperature, data=dt.la)
anova(md.la2, md.la1)
```

```
## Analysis of Variance Table
##
## Model 1: respiratoryrate ~ pulse + temperature
## Model 2: respiratoryrate ~ (pulse + temperature + cellvolume + totalprotein) *
##     (abdominaldistension + pain)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    114 23.241
## 2     77 16.455 37     6.786 0.8582 0.6914
```

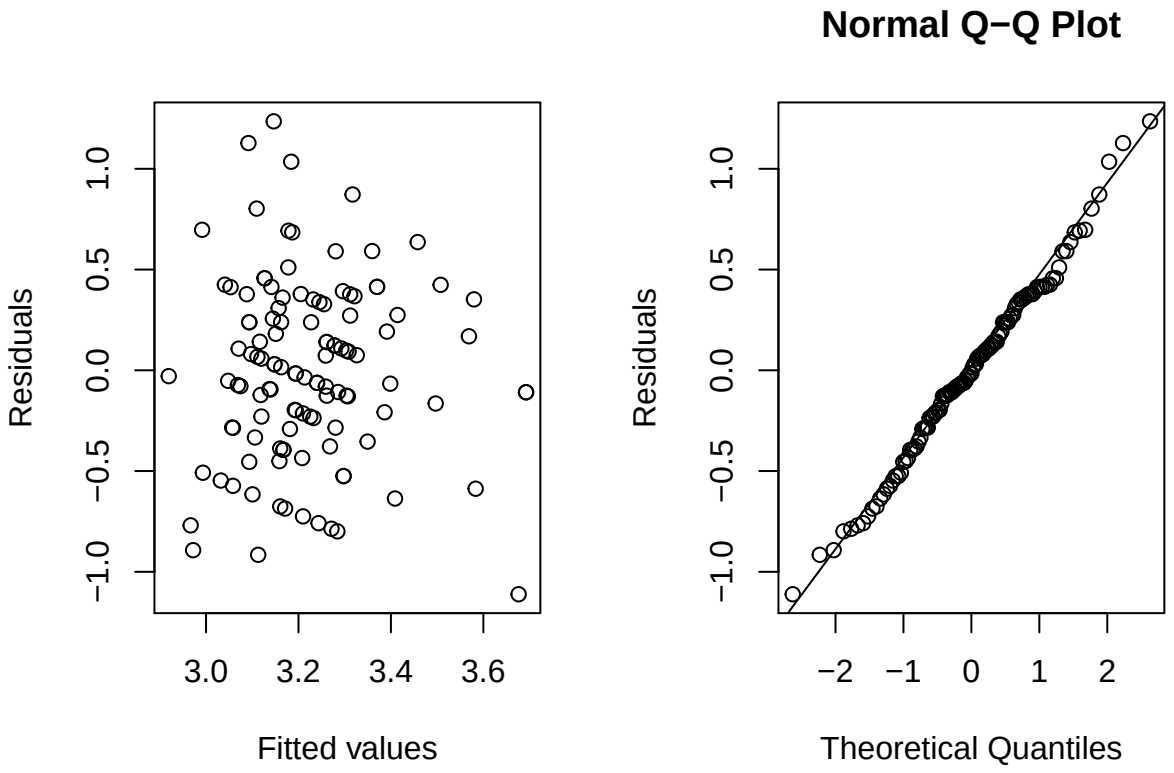
The p-value is large, and we fail to reject the null hypothesis that two models are not different.

```
summary(md.la2)
```

```
##
## Call:
## lm(formula = respiratoryrate ~ pulse + temperature, data = dt.la)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11136 -0.28450 -0.01584  0.32823  1.23558
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.773371   2.120543  -0.836   0.4047
## pulse        0.004408   0.001964   2.244   0.0268 *
## temperature  0.123157   0.056191   2.192   0.0304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4515 on 114 degrees of freedom
## Multiple R-squared:  0.09825,    Adjusted R-squared:  0.08243
## F-statistic: 6.211 on 2 and 114 DF,  p-value: 0.002753
```

Test for constant error and normality of errors.

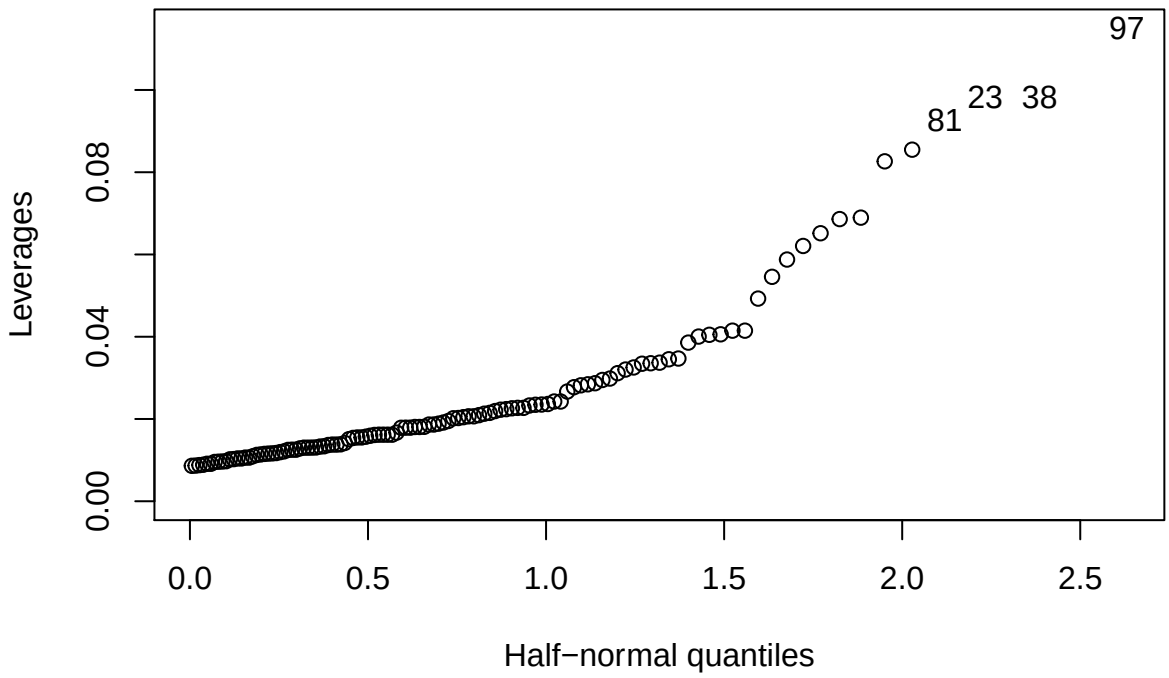
```
par(mfrow=c(1,2))
plot(md.la2$fitted.values,md.la2$residuals, xlab = "Fitted values", ylab = "Residuals")
qqnorm(md.la2$residuals, ylab = "Residuals")
qqline(md.la2$residuals)
```



From qq plot, the residuals are approximately normal. In the plot, it seems that residuals have larger variance when fitted value is small, but this may due to less data points when fitted value is large. There is not enough data to reach to a conclusion.

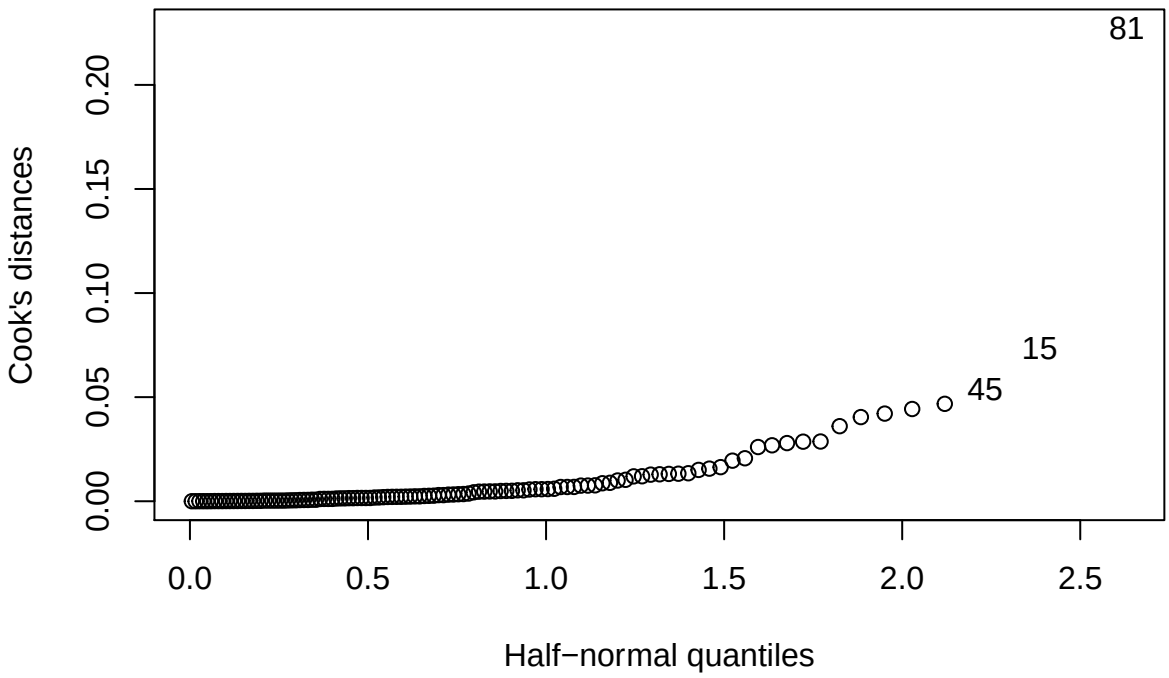
Check for leverage points.

```
hatv <- hatvalues(md.la2)
halfnorm(hatv,4,ylab="Leverages")
```



Check for influential points.

```
cook <- cooks.distance(md.la2)
halfnorm(cook, 3, ylab="Cook's distances")
```



Check for outliers.

```
#Compute Bonferroni critical value
crival = qt(.05/(nrow(dt.la)*2), md.la2$df.residual)
#Compute tudentized residuals
stures = rstudent(md.la2)
stures[which(abs(stures)>abs(crival))]
```

```
## named numeric(0)
```

Remove 97 and 81 and fit the model again

```
md.la3 = lm(respiratoryrate ~ pulse+temperature, data=dt.la[-c(81,97),])
summary(md.la3)
```

```
##
## Call:
## lm(formula = respiratoryrate ~ pulse + temperature, data = dt.la[-c(81,
##     97), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90608 -0.26167  0.02754  0.31498  1.24217
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.969553   2.267981  -1.309  0.19310
## pulse        0.005161   0.001944   2.654  0.00911 **
```

```
## temperature  0.153432   0.059992   2.558  0.01188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4419 on 112 degrees of freedom
## Multiple R-squared:  0.1276, Adjusted R-squared:  0.1121
## F-statistic: 8.193 on 2 and 112 DF,  p-value: 0.0004778
```

Model for adult horse("age"=1) withtotalprotein > 30

Contingency table of "pain"(row) and "abdominaldistension"(column) for horse with "totalprotein" > 30 and "age" = 1 (adult).

```
addmargins(table(dt.g[dt.g$age==1,]$pain,dt.g[dt.g$age==1,]$abdominaldistension))
```

```
##
##        1  2  3  4 Sum
##   1    5  2  0  0   7
##   2    3  3  5  2  13
##   3    5  6  3  1  15
##   4    1  2  1  1   5
##   5    2  0  2  1   5
##   Sum 16 13 11  5  45
```

There are 13 observations in this data set.

```
md.ga1 = lm(respiratoryrate ~ (pulse+temperature+cellvolume+totalprotein)*(abdominaldistension+pain), da
anova(md.ga1)
```

```
## Analysis of Variance Table
##
## Response: respiratoryrate
##                                  Df  Sum Sq Mean Sq F value  Pr(>F)
## pulse                             1 1.20622 1.20622  5.2868 0.05505 .
## temperature                       1 0.00112 0.00112  0.0049 0.94616
## cellvolume                        1 0.37561 0.37561  1.6463 0.24031
## totalprotein                      1 0.00420 0.00420  0.0184 0.89584
## abdominaldistension               3 0.04326 0.01442  0.0632 0.97764
## pain                              4 0.15748 0.03937  0.1726 0.94558
## pulse:abdominaldistension         3 2.05770 0.68590  3.0063 0.10417
## pulse:pain                        4 0.21787 0.05447  0.2387 0.90771
## temperature:abdominaldistension   3 0.22659 0.07553  0.3310 0.80352
## temperature:pain                  4 0.48386 0.12096  0.5302 0.71834
## cellvolume:abdominaldistension    3 0.09594 0.03198  0.1402 0.93276
## cellvolume:pain                   4 1.05150 0.26287  1.1522 0.40683
## totalprotein:abdominaldistension  3 0.43714 0.14571  0.6387 0.61376
## totalprotein:pain                 2 0.40885 0.20443  0.8960 0.45034
## Residuals                         7 1.59709 0.22816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
md.ga2 = lm(respiratoryrate ~ pulse, data = dt.ga)
anova(md.ga1, md.ga2)
```

```
## Analysis of Variance Table
##
## Model 1: respiratoryrate ~ (pulse + temperature + cellvolume + totalprotein) *
```

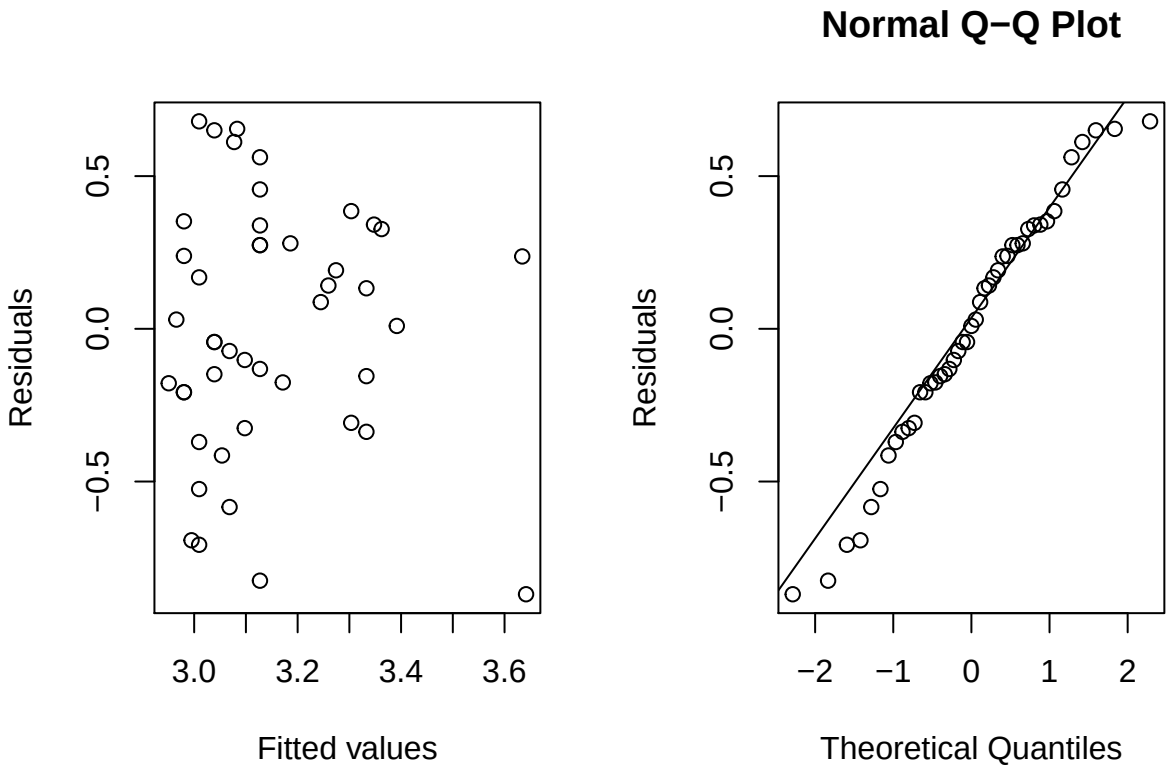
```
##     (abdominaldistension + pain)
## Model 2: respiratoryrate ~ pulse
##   Res.Df    RSS  Df Sum of Sq      F Pr(>F)
## 1      7 1.5971
## 2     43 7.1582 -36   -5.5611 0.6771 0.7934
```

```
summary(md.ga2)
```

```
##
## Call:
## lm(formula = respiratoryrate ~ pulse, data = dt.ga)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86900 -0.20763  0.00946  0.27976  0.67927
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.686275   0.181467  14.803   <2e-16 ***
## pulse       0.007349   0.002730   2.692   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.408 on 43 degrees of freedom
## Multiple R-squared:  0.1442, Adjusted R-squared:  0.1243
## F-statistic: 7.246 on 1 and 43 DF,  p-value: 0.01008
```

Test for constant error and normality of errors.

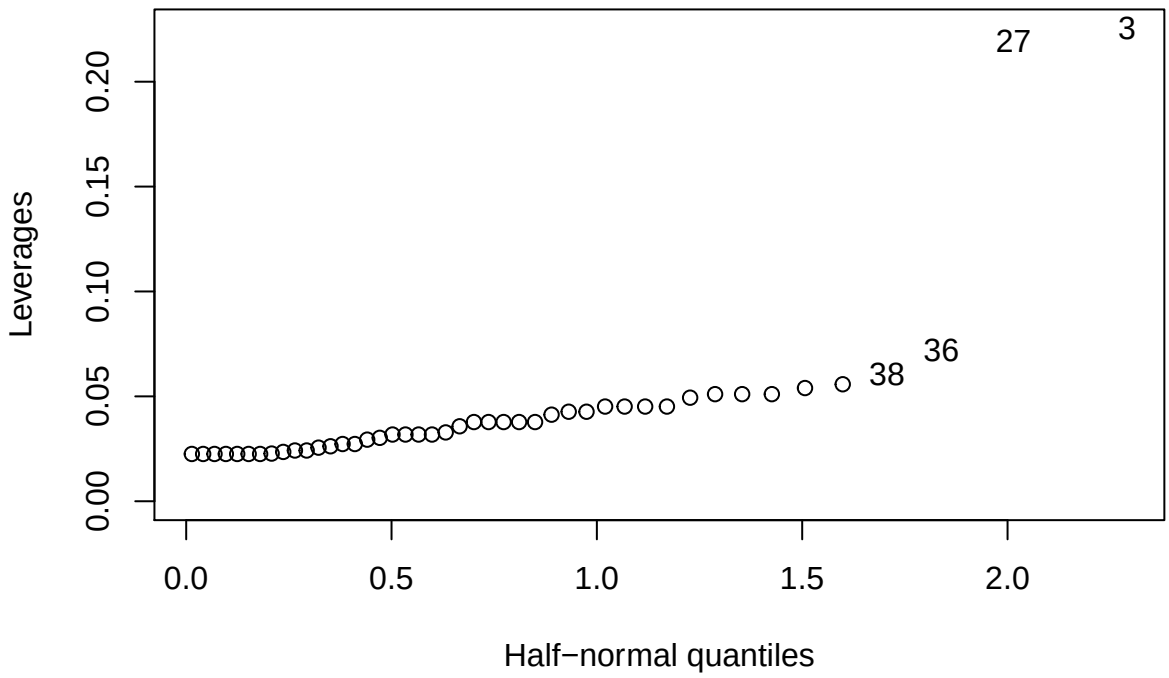
```
par(mfrow=c(1,2))
plot(md.ga2$fitted.values,md.ga2$residuals, xlab = "Fitted values", ylab = "Residuals")
qqnorm(md.ga2$residuals, ylab = "Residuals")
qqline(md.ga2$residuals)
```



From qq plot, the residuals are approximately normal. In the plot, it seems that residuals have larger variance when fitted value is small, but this may due to less data points when fitted value is large. There is not enough data to reach to a conclusion.

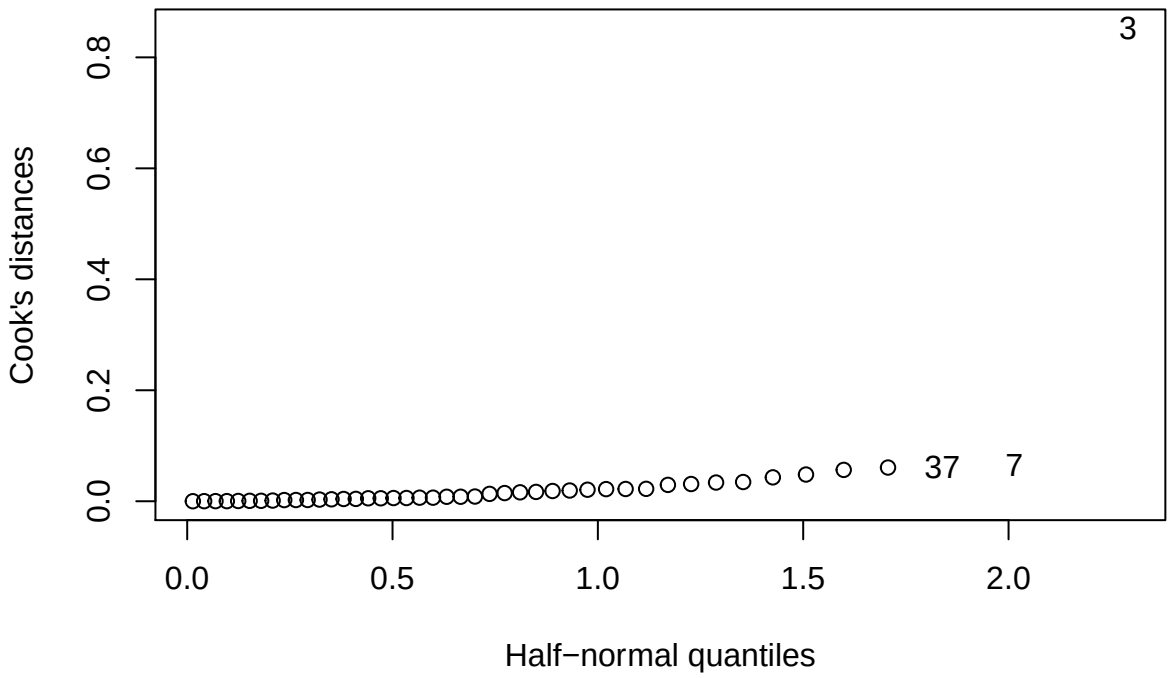
Check for leverage points.

```
hatv <- hatvalues(md.ga2)
halfnorm(hatv,4,ylab="Leverages")
```



Check for influential points.

```
cook <- cooks.distance(md.ga2)
halfnorm(cook, 3, ylab="Cook's distances")
```



Check for outliers.

```
#Compute Bonferroni critical value
crival = qt(.05/(nrow(dt.ga)*2), md.la2$df.residual)
#Compute tudentized residuals
stures = rstudent(md.la2)
stures[which(abs(stures)>abs(crival))]
```

```
## named numeric(0)
```

Remove point 3 and 27 and fit the model again

```
md.ga3 = lm(respiratoryrate ~ pulse, data = dt.ga[-c(3,27),])
summary(md.ga3)
```

```
##
## Call:
## lm(formula = respiratoryrate ~ pulse, data = dt.ga[-c(3, 27),
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84097 -0.24027 -0.01823  0.25764  0.71812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.495557   0.211363  11.807 9.01e-15 ***
## pulse       0.010800   0.003409   3.168   0.0029 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3883 on 41 degrees of freedom
## Multiple R-squared:  0.1967, Adjusted R-squared:  0.1771
## F-statistic: 10.04 on 1 and 41 DF,  p-value: 0.002898
```

Results and Discussion

We divide the horses to 4 categories with "totalprotein"<30 or "totalprotein">30 and adult("age"=1) or juvenile("age"=2). Among the 4 categories, we have little data for horses with "totalprotein">30 and "age"=2. We fit three model for other 3 categories.

We take log transformation of the response variable. The model fitted is linear model of log(respiratoryrate).

"totalprotein"<30 and "age"=2(juvenile), sample size after imputation:13

```
summary(md.lj2)$coefficients
```

```
##                Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 2.971906944 0.475015694 6.256439 6.204028e-05
## pulse       0.007571353 0.003664792 2.065971 6.321099e-02
```

"totalprotein"<30 and "age"=1(adult), sample size after imputation:117

```
summary(md.la3)$coefficients
```

```
##                 Estimate  Std. Error   t value    Pr(>|t|)
## (Intercept) -2.969553004 2.267980863 -1.309338 0.193099721
## pulse        0.005161082 0.001944453  2.654259 0.009105016
## temperature  0.153432276 0.059992193  2.557537 0.011878344
```

"totalprotein">30 and "age"=1(adult), sample size after imputation:45

```
summary(md.ga3)$coefficients
```

```
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 2.49555744 0.211362553 11.806999 9.014130e-15
## pulse       0.01080002 0.003409081  3.168013 2.897738e-03
```

We can see that log(respiratoryrate) have different linear on covariates in the three groups. More data is needed for further analyse. Especially, more data of juvenile horses is needed to discuss the difference between adult and juvenile horses.