Assignment 14

Sachin Shubham

4/24/2021

Use file Exc data.txt as a data source.

3 different schools want to compare their students mean typing speed between their classes. They have also recorded the year or grade of each student. They are not interested in the effect of Year per se, but they want to account for the effect statistically. For each of the following, answer the question, and show the output from the analyses you used to answer the question.

1) What was the LS mean typing speed for each school's class?
2) Does School have a significant effect on typing speed?
3) Does Year have a significant effect on typing speed?
4) Are the residuals reasonably normal and homoscedastic?
5) Which schools had classes with significantly different mean typing speed from which others?
6) Does this result correspond to what you would have thought from looking at the plot of the LS means and standard errors?
7) Is this design balanced or unbalanced?

Solution:

1)

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.0.5
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
library(multcompView)
```

```
## Warning: package 'multcompView' was built under R version 4.0.4
```

```
library(lsmeans)
```

```
## Warning: package 'lsmeans' was built under R version 4.0.5
```

```
## Loading required package: emmeans
```

```
## Warning: package 'emmeans' was built under R version 4.0.5
```

```
## The 'lsmeans' package is now basically a front end for 'emmeans'.
## Users are encouraged to switch the rest of the way.
## See help('transition') for more information, including how to
## convert old 'lsmeans' objects and scripts to work with 'emmeans'.
```

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
library(rcompanion)
```

```
## Warning: package 'rcompanion' was built under R version 4.0.5
```

```
##
## Attaching package: 'rcompanion'
```

```
## The following object is masked from 'package:psych':
##
##     phi
```

```
library(multcompView)
library(FSA)
```

```
## Warning: package 'FSA' was built under R version 4.0.5
```

```
## ## FSA v0.8.32. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```
##
## Attaching package: 'FSA'
```

```
## The following object is masked from 'package:car':
##
##     bootCase

## The following object is masked from 'package:psych':
##
##     headtail
```

```
setwd("D:/Statistical Methods/Assignments/Assignment 11")
#data frame
df <- read.table("Exc_data.txt",header=TRUE)
headTail(df)
```

```
##     School Year Words.per.minute
## 1        1    7               32
## 2        1    7               51
## 3        1    7               65
## 4        1    7               60
## ...    ...  ...              ...
## 45       3   10               75
## 46       3   10               51
## 47       3   10               50
## 48       3   10               63
```

```
str(df)
```

```
## 'data.frame':    48 obs. of  3 variables:
##  $ School          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Year            : int  7 7 7 7 8 8 8 8 9 9 ...
##  $ Words.per.minute: int  32 51 65 60 65 58 70 55 43 55 ...
```

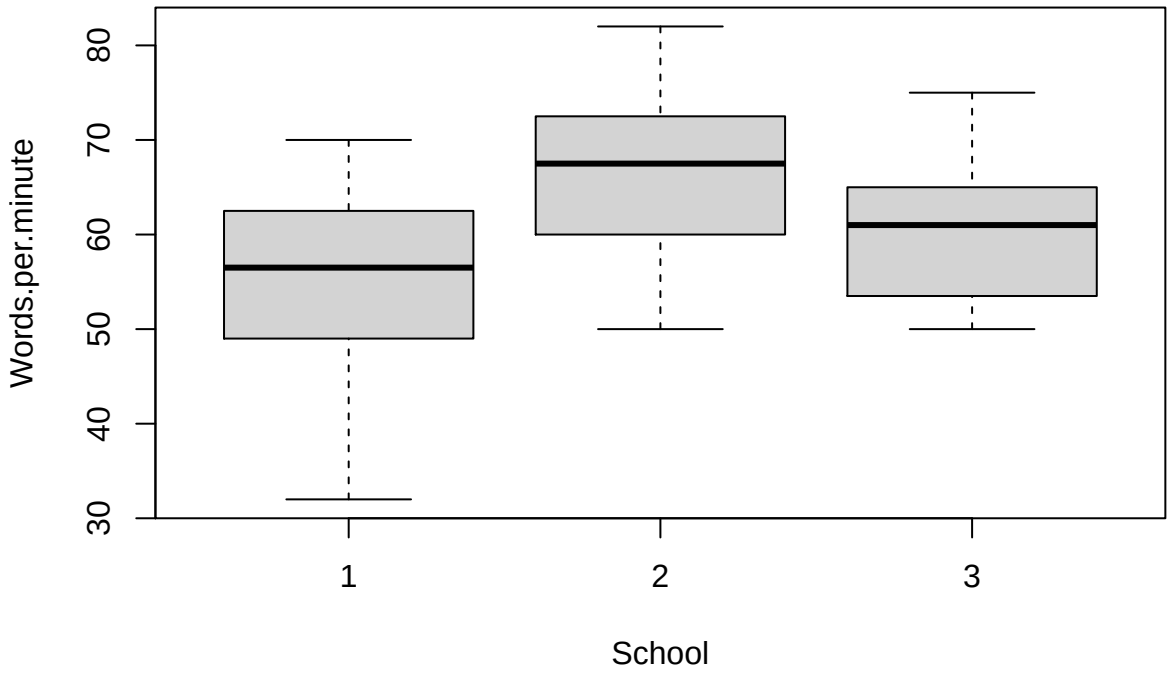
```
summary(df)
```

```
##      School       Year        Words.per.minute
##  Min.   :1   Min.   : 7.000   Min.   :32.00
##  1st Qu.:1   1st Qu.: 7.750   1st Qu.:55.00
##  Median :2   Median : 8.500   Median :60.00
##  Mean   :2   Mean   : 8.562   Mean   :60.69
##  3rd Qu.:3   3rd Qu.: 9.250   3rd Qu.:65.50
##  Max.   :3   Max.   :12.000   Max.   :82.00
```

```
#Summarize Data Frame
Summarize(Words.per.minute ~ School,data=df,digits=3)
```

```
##   School  n   mean     sd min    Q1 median    Q3 max
## 1      1 16 55.125 10.191  32 50.00   56.5 61.25  70
## 2      2 16 66.312  8.444  50 60.00   67.5 72.25  82
## 3      3 16 60.625  7.182  50 54.25   61.0 65.00  75
```

```r
#boxplot
boxplot(Words.per.minute  ~  School,data = df)
```

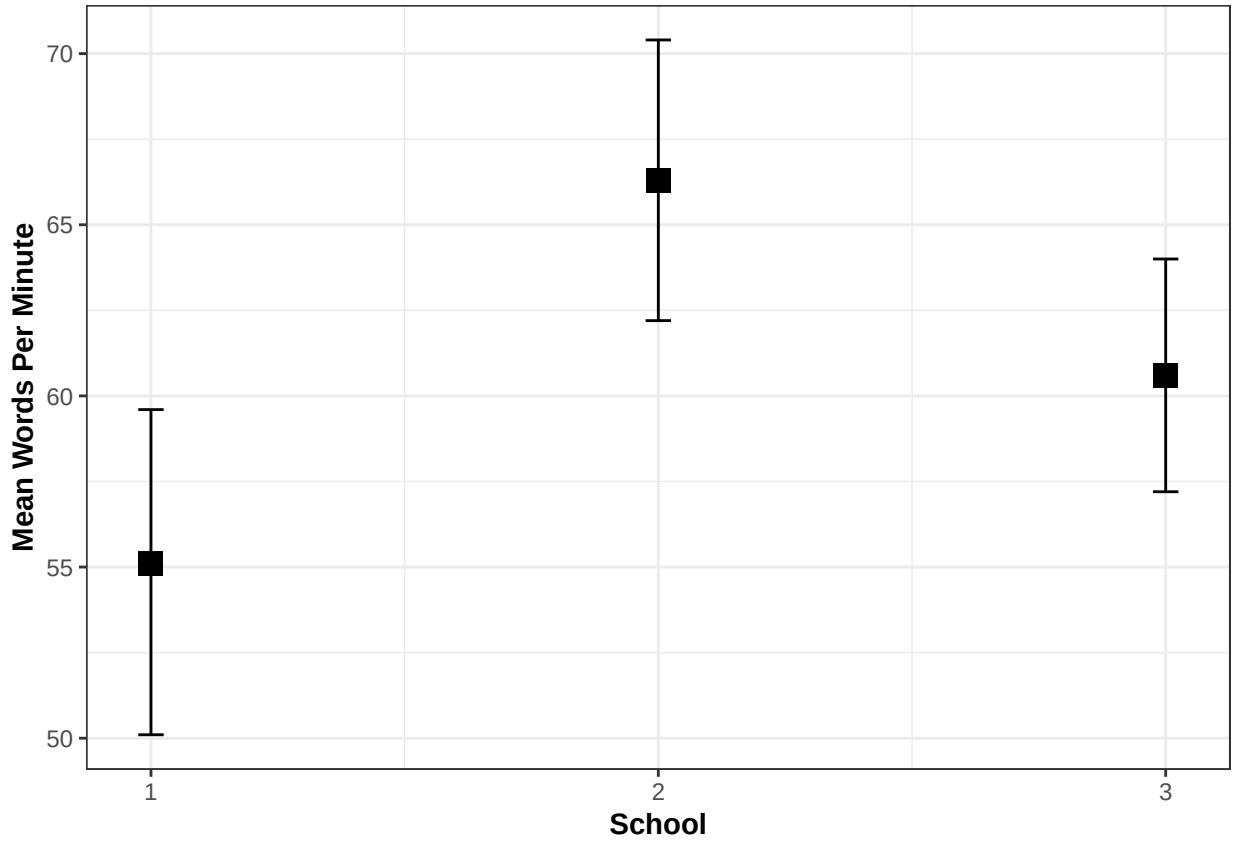



```r
#groupwiseMean
Sum = groupwiseMean(Words.per.minute ~ School,
                    data   = df,
                    conf   = 0.95,
                    digits = 3,
                    traditional = FALSE,
                    percentile  = TRUE)
print(Sum)
```

```
##   School  n Mean Conf.level Percentile.lower Percentile.upper
## 1      1 16 55.1       0.95             50.1             59.6
## 2      2 16 66.3       0.95             62.2             70.4
## 3      3 16 60.6       0.95             57.2             64.0
```

```r
#ggplot
ggplot(Sum,
       aes(x = School,
           y = Mean)) +
  geom_errorbar(aes(ymin = Percentile.lower,
                    ymax = Percentile.upper),
                width = 0.05,
```

```
              size  = 0.5) +
  geom_point(shape = 15,
             size  = 4) +
  theme_bw() +
  theme(axis.title   = element_text(face  = "bold")) +
  ylab("Mean Words Per Minute")+scale_x_continuous(breaks=c(1,2,3))
```




```
#factor
df$School = factor(df$School,levels=unique(df$School))
#model
lm.model = lm(Words.per.minute ~ School+Year,data = df)
summary(lm.model)
```

```
##
## Call:
## lm(formula = Words.per.minute ~ School + Year, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.4098  -5.7827   0.6647   5.5042  15.1664
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.4430     9.1619   5.615 1.24e-06 ***
## School2      11.2670     3.1082   3.625 0.000746 ***
```

```
## School3       5.5795     3.1082   1.795 0.079515 .
## Year          0.4238     1.0239   0.414 0.680940
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.775 on 44 degrees of freedom
## Multiple R-squared:  0.2305, Adjusted R-squared:  0.178
## F-statistic: 4.392 on 3 and 44 DF,  p-value: 0.00867
```

```
#Anova
Anova(lm.model,type = "II")
```

```
## Anova Table (Type II tests)
##
## Response: Words.per.minute
##           Sum Sq Df F value   Pr(>F)
## School    1011.7  2  6.5703 0.003186 **
## Year        13.2  1  0.1713 0.680940
## Residuals 3387.7 44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#ls mean
ls_mean = lsmeans(lm.model,~ School)
summary(ls_mean)
```

```
##  School lsmean   SE df lower.CL upper.CL
##  1        55.1 2.20 44     50.6     59.5
##  2        66.3 2.19 44     61.9     70.8
##  3        60.7 2.19 44     56.2     65.1
##
## Confidence level used: 0.95
```

```
pairs(ls_mean,adjust="tukey")
```

```
##  contrast estimate   SE df t.ratio p.value
##  1 - 2      -11.27 3.11 44 -3.625  0.0021
##  1 - 3       -5.58 3.11 44 -1.795  0.1830
##  2 - 3        5.69 3.10 44  1.833  0.1707
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

The LS mean typing speed for each school are:

School 1 : 55.1 words/min

School 2 : 66.3 words/min

School 3 : 60.7 words/min

2)

School 1 and 2 are significant with p-values 1.24e-06 and 0.000746 (< 0.05) respectively on typing speed but School 3 not significant on typing speed with p-value of 0.079515 (> 0.05)

3)

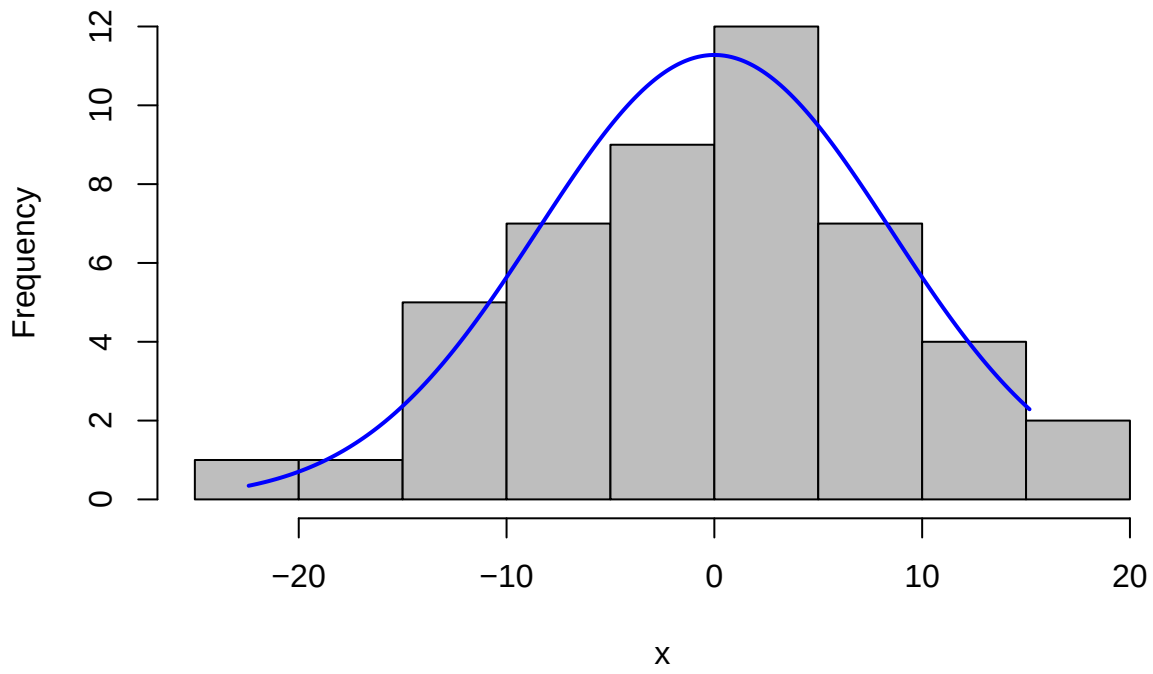
Year is not significant effect on typing speed as the p-value of Year is 0.680940 which is quite higher than 0.05.

4)

```
#checking for homoscedastic
lmtest::bptest(lm.model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm.model
## BP = 2.3218, df = 3, p-value = 0.5084
```

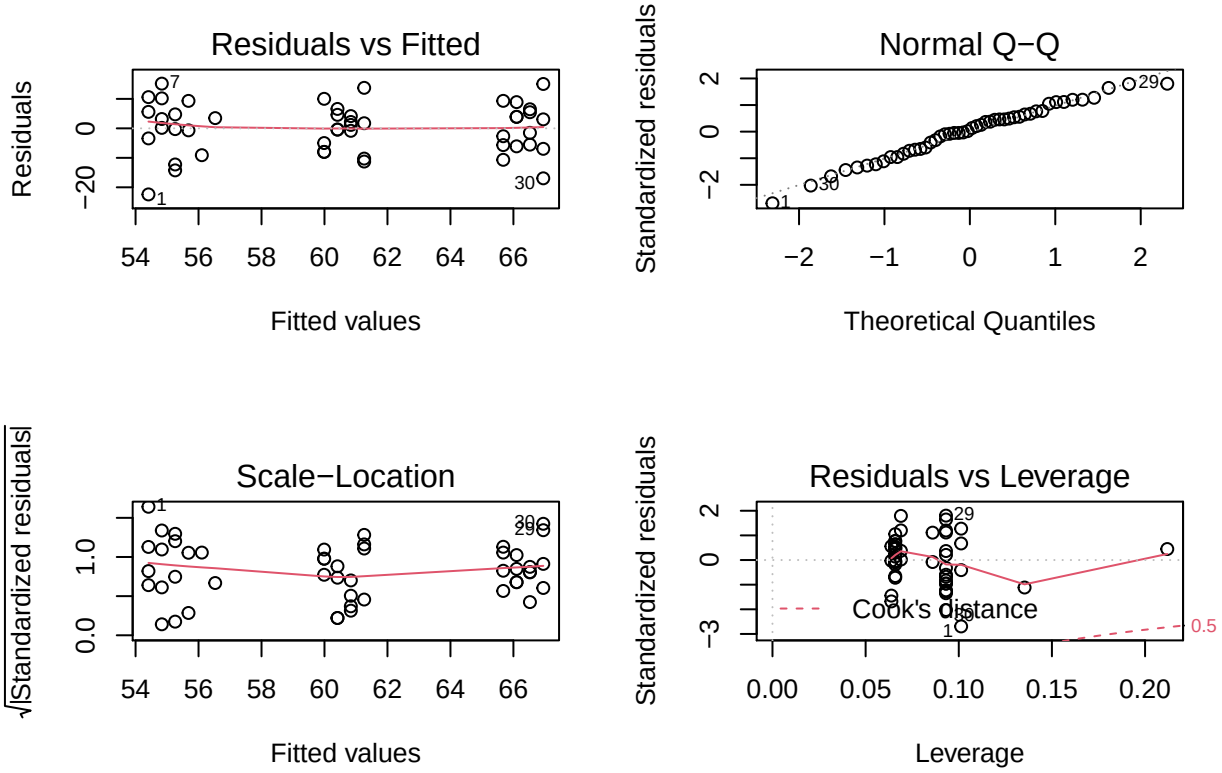
```
#plot for normality
x = residuals(lm.model)
plotNormalHistogram(x)
```




```
#checking for normality
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.98124, p-value = 0.6312
```

```
#plot model
par(mfrow=c(2, 2))
plot(lm.model)
```



With a p-value of 0.91, we fail to reject the null hypothesis and therefore infer that the residuals are homoscedastic. Also we have a much flatter line and an evenly distributed residuals in the Residuals vs Fitted plot(top-left plot)

The residuals are reasonably normal and the different school variance data appears to be equal. Looking at the Shapiro test for the residuals, the p-value is 0.6312 and greater than 0.05 so we accept the null hypothesis that the data is normally distributed

5)

```
CLD  =  multcomp::cld(ls_mean,
              alpha   = 0.05,
              Letters = letters,     ### Use lower-case letters for .group
              adjust  = "sidak")     ###  Tukey-adjusted comparisons
CLD
```

```
##  School lsmean   SE df lower.CL upper.CL .group
##  1        55.1 2.20 44     49.6     60.5  a
##  3        60.7 2.19 44     55.2     66.1  ab
##  2        66.3 2.19 44     60.9     71.8   b
##
## Confidence level used: 0.95
## Conf-level adjustment: sidak method for 3 estimates
## P value adjustment: sidak method for 3 tests
## significance level used: alpha = 0.05
```

School 1 and 2 had classes with significantly different mean typing speed from each others

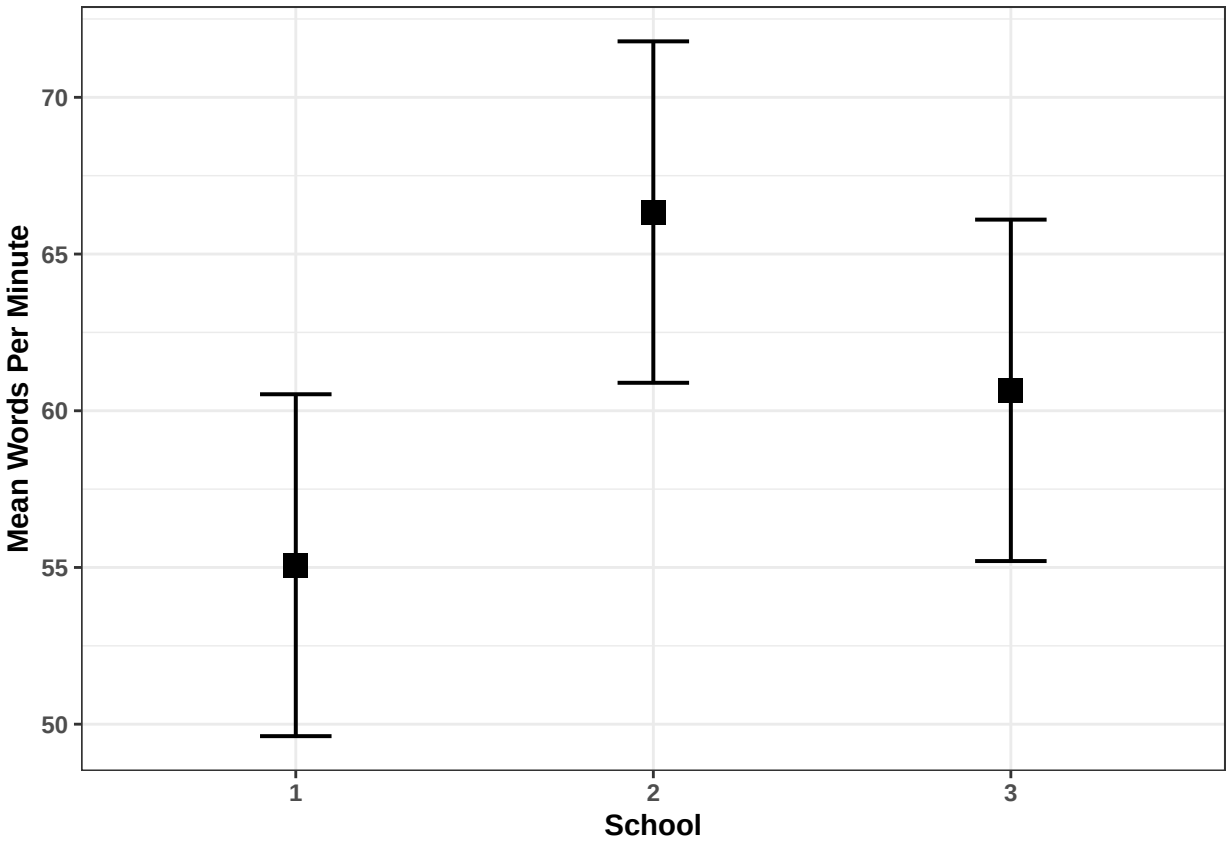
6)

```
#plot for LS Mean
ggplot(CLD,aes(x= School,y= lsmean,
           label = .group))+
  geom_point(shape  = 15,
             size   = 4) +

  geom_errorbar(aes(ymin  =  lower.CL,
                    ymax  =  upper.CL),
                width =  0.2,
                size  =  0.7) +

  theme_bw() +
  theme(axis.title   = element_text(face = "bold"),
        axis.text    = element_text(face = "bold"),
        plot.caption = element_text(hjust = 0)) +

  ylab("Mean Words Per Minute")
```



Median for School 2 is higher than School 1 and School 3.There is overlap between School 2 and School 3, also between School 1 and School 3

7)

```
#Checking design is balanced or unbalanced
VCA:::isBalanced(Words.per.minute ~ School+Year, df, na.rm = TRUE)
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                           from
##   cooks.distance.influence.merMod car
##   influence.merMod                 car
##   dfbeta.influence.merMod          car
##   dfbetas.influence.merMod         car
```

```
## [1] FALSE
```

Yes design is unbalanced.Additionally the number of data points in each category differs