hw4

Sam Mottahedi

July 22, 2017

Problem 1

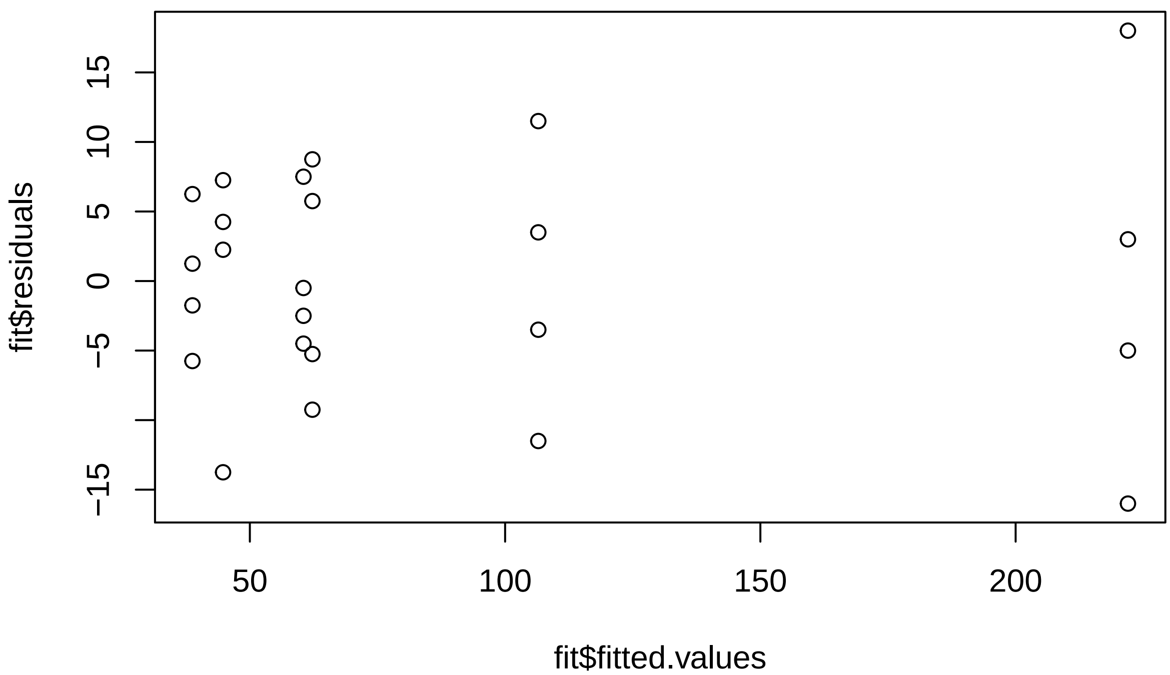
a)

```
d = read.table('CH19PR20.txt')
y = d[,1] ;A = as.factor(d[,2]); B = as.factor(d[,3])

df = data.frame(y=y, A=A, B=B)


par(mfrow=c(2,2))
fit <- lm(y ~ A + B + A * B)
```

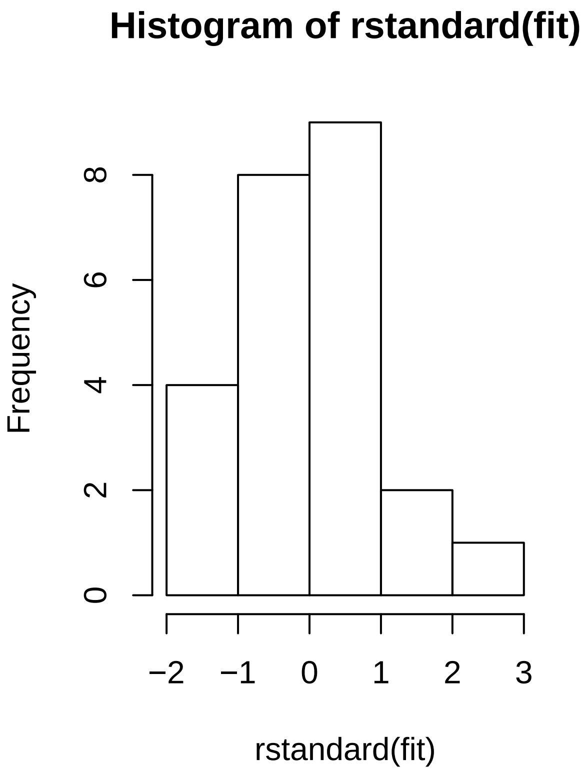
```
plot(fit$fitted.values, fit$residuals)
```

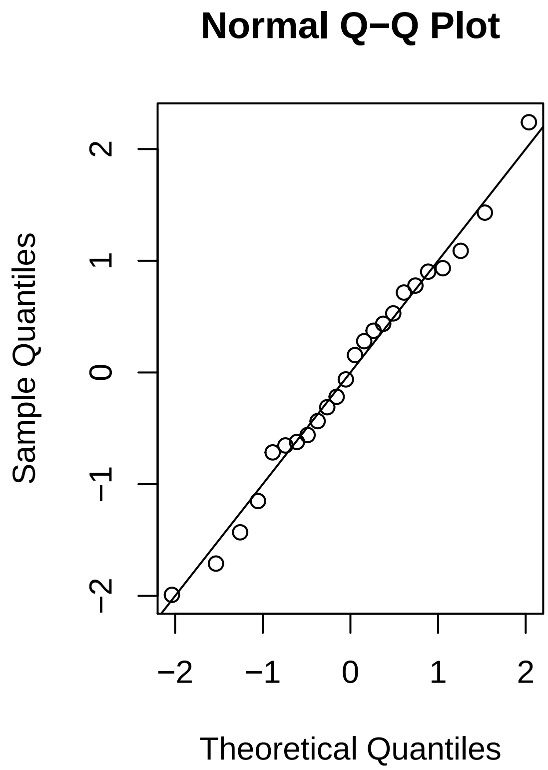


The residuals variance is not constant and increases as the prediction error increase.

b)

```
par(mfrow=c(1,2))
hist(rstandard(fit))
qqnorm(rstandard(fit))
abline(a=0, b=1)
```





The residuals are slightly skewed to the left but it's not a major departure from normality of residuals.

c)

```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## A          1  39447   39447  458.02 2.983e-14 ***
## B          2  36412   18206  211.39 3.158e-13 ***
## A:B        2  20165   10083  117.07 4.816e-11 ***
## Residuals 18   1550      86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test statistic $F = 117.07 >= 6.0129048$. the interaction term is significant.

d)

A:

The test statistic $F = 458.02 \geq 8.2854196$. the main effect for factor A is significant.

B:

The test statistic $F = 211.39 \geq 6.0129048$. the main effect for factor B is significant.

Testing is for main effect is not necessary since the test in part c showed that factor A and B interact with each other.

e)

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
df %>% group_by(A) %>% summarise(mu=mean(y)) -> means
means
```

```
## # A tibble: 2 x 2
##        A        mu
##   <fctr>     <dbl>
## 1      1 129.66667
## 2      2  48.58333
```

```r
D_hat <- means$mu[1] - means$mu[2]
s <- sqrt(2*86 / (2*4))
q = sqrt(2) * D_hat / s
q; D_hat
```

```
## [1] 24.73018

## [1] 81.08333
```

```r
qtukey(0.95, 2, 3*6)
```

```
## [1] 2.971152
```

```r
# TukeyHSD(aov(y~ A*B, df), which = 'A', conf.level = 0.95)
```

$q^* = 24.73 > 2.97$ Rejecting the null $(H_0 : D = \mu_i - \mu_{i'} = 0)$ the difference between the means of factor A is significant.

Problem 2

a)

left :

$y_{ij} = \mu_i + \epsilon_{ij}$

right:

$y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij}$

b)

In the image on the left, the treatments were randomly assigned with out considering the proximity to the wall or open walkway. In the image on the right, the each treatments are assigned in homogeneous groups and the treatment are assigned at random within each block.

c)

left:

source	SS	df	MS
BTW treatment	$SSTR = \sum n_i(\hat{(Y_{i.})} - \hat{Y}_{..}$	$r - 1 = 3$	SSTR/r-1
error	$SSE = \sum\sum(\hat{(Y_{ij})} - \hat{Y}_{i.}$	$n_T - r = 20$	SSE/ $n_T - r$
total	SSTO	$n_T - 1 = 23$	

right:

source	SS	df	MS
Blocks	$SSBL = r\sum(\hat{(Y_{i.})} - \hat{Y}_{..}$	$n_b - 1 = 5$	SSBL/$n_b - 1$
treatments	$SSTR = \sum n_b(\hat{(Y_{.j})} - \hat{Y}_{..}$	$r - 1 = 3$	SSTR/r-1
error	$SSBL.TR = \sum\sum e_{ij}^2$	$(n_b - 1)(r - 1) = 15$	SSBL.TR / $(n_b - 1)(r - 1)$
total	SSTO	$n_T - 1$	

Problem 3

```
dat <- read.table('P1.txt', header = T, colClasses = c('numeric', 'factor', 'factor'))
```

a)

```
fit <- aov(score ~teacher + method, dat )
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## teacher      9  433.4    48.2   7.716 0.000132 ***
## method       2 1295.0   647.5 103.754 1.32e-10 ***
## Residuals   18  112.3     6.2                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b)

$F^* = \frac{647.5}{6.2} = 104.4354839$

$H_0 : \tau_1 = \tau_2 = \tau_3 = 0$

H_a : not all$\tau_j$mequal zero

$F^* > 3.5545571$

rejecting the null hypothesis, the mean performance is different for the three teaching methods.

c)

```
dat %>% group_by(method) %>% summarise(mu=mean(score)) -> means
means
```

```
## # A tibble: 3 x 2
##   method    mu
##   <fctr> <dbl>
## 1      1  70.6
## 2      2  74.6
## 3      3  86.1
```

$s^s = \frac{6.2 \times 2}{10} = 1.24$

$q = 3.1599076$

$T = 2.2343921$

$Ts\{\hat{D}\} = 2.4881137$

$1.5118863 \leq \mu_2 - \mu_1 \leq 6.4881137$

$13.0118863 \leq \mu_3 - \mu_1 \leq 17.9881137$

$9.0118863 \leq \mu_3 - \mu_2 \leq 13.9881137$

we conclude that the method 3 has higher mean performance compared to method 2 and has a higher mean compared to method 1.

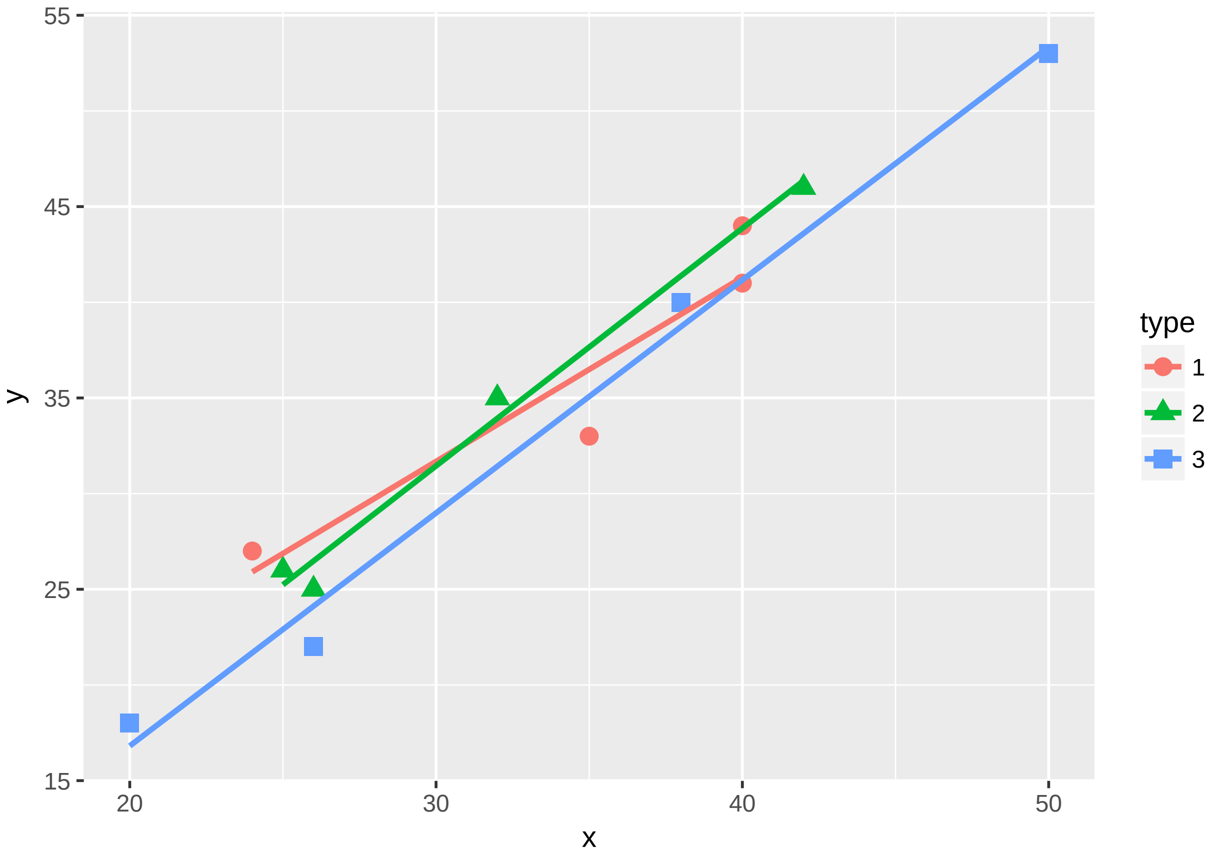
Problem 4

```
data = read.table('softdrink.dat',header=T, colClasses = c('numeric', 'numeric', 'factor'))
```

a)

```
library(ggplot2)

g <- ggplot(data, aes(x=x, y=y, shape=type, color=type)) +
    geom_point(size=2, stroke=2) +
    geom_smooth(se=FALSE, method='lm')
print(g)
```



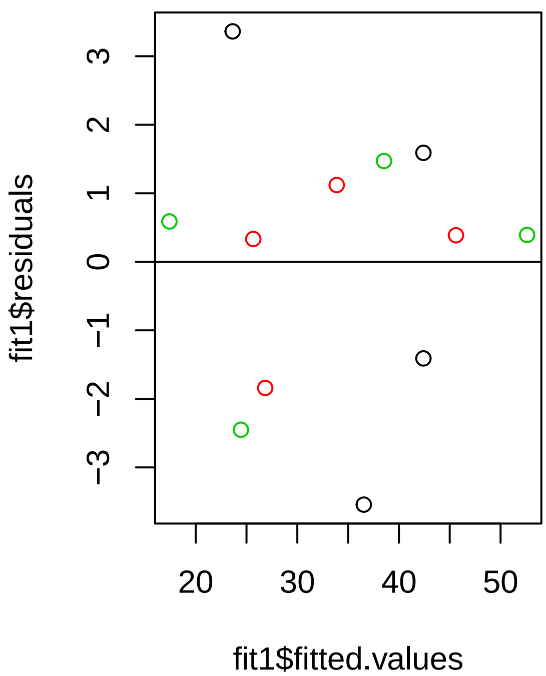
The slope for method 2 and 3 looks equal and the slope of method 1 is slightly different from the other two. Points related to method are general higher than method 3, and higher than method 1 in higher values of x.

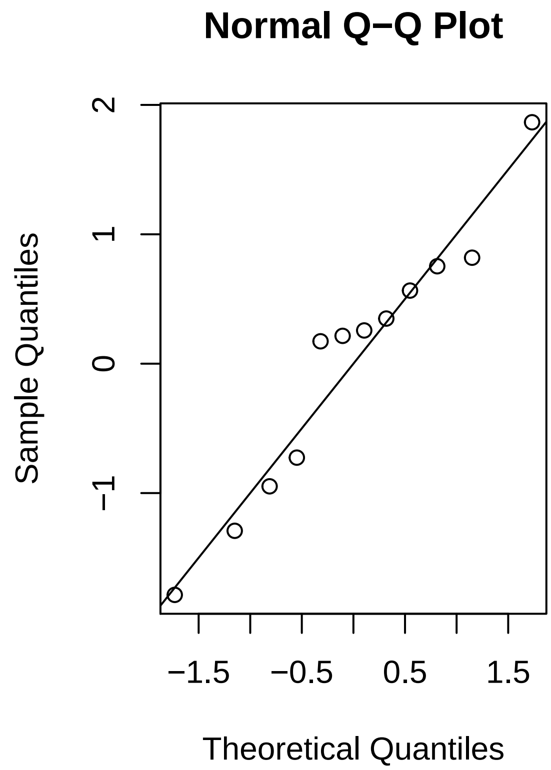
b)

```
df <- data
df$x <- df$x - mean(df$x)
fit1 <- lm(y ~ x + type, df)
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## x          1 1232.07 1232.07 234.9734 3.257e-07 ***
## type       2   11.65    5.82   1.1106    0.3753
## Residuals  8   41.95    5.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,2))
plot(fit1$fitted.values, fit1$residuals, col=df$type)
abline(a=0, b=0)
qqnorm(rstandard(fit1))
abline(a = 0, b=1)
```





The two figures shows that there no major deviation from equal variance assumption and normality of residuals assumption.

c)

```
fit2 <- lm(y ~ x, df)
# anova(fit1, fit2)
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## x          1 1232.07 1232.07 234.9734 3.257e-07 ***
## type       2   11.65    5.82   1.1106    0.3753
## Residuals  8   41.95    5.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 1232.07 1232.07  229.89 3.153e-08 ***
## Residuals 10   53.59    5.36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \tau_1 = \tau_2 = 0$

H_a : not both$\tau_1 and \tau_2$equal zero

$F^* = 1.1098927$

$F^* < F = 4.4589701$

we conclude H_0 that three truck methods do not differ in mean delivery time.

d)

```r
anova(lm(y ~ type, df))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value Pr(>F)
## type       2   26.17  13.083  0.0935 0.9116
## Residuals  9 1259.50 139.944
```

$Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}$

$H_0 : \gamma = 0$

$F^* = 232.0953516$

$F^* > F = 5.3176551$

rejecting H_0 the slope is significant.

e)

$Y_{ij} = \mu_. + \tau_i + \gamma(X_{ij} - \bar{X}_{ij}) + \beta_1 I_{ij1}(X_{ij} - \bar{X}_{ij}) + \beta_2 I_{ij2}(X_{ij} - \bar{X}_{ij}) + \epsilon_{ij}$

$H_0 : \beta_1 = \beta_2 = 0$

```r
anova(lm(y ~ x + type + x:type ,df))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## x          1 1232.07 1232.07 228.0542 5.316e-06 ***
## type       2   11.65    5.82   1.0779    0.3982
## x:type     2    9.53    4.77   0.8822    0.4615
## Residuals  6   32.42    5.40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F^* = 0.881863$

$F^* < F = 5.1432528$

concluding null, that the tree treatment lines have the same slope.

$P - value = 0.4615984$