a4

Mushi Wang

27/11/2019

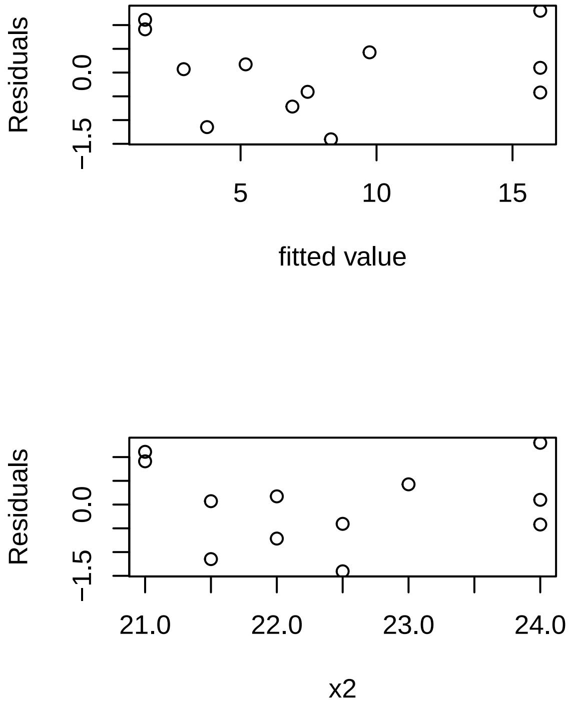
q1

```
library(readxl)
carb = read_excel("carbonation.xls")
carb
```

```
## # A tibble: 12 x 3
##        y    x1    x2
##    <dbl> <dbl> <dbl>
##  1  2.6   31    21
##  2  2.4   31    21
##  3 17.3   31.5  24
##  4 15.6   31.5  24
##  5 16.1   31.5  24
##  6  5.36  30.5  22
##  7  6.19  31.5  22
##  8 10.2   30.5  23
##  9  2.62  31    21.5
## 10  2.98  30.5  21.5
## 11  6.92  31    22.5
## 12  7.06  30.5  22.5
```

(a)

```
fit = lm(y~., carb)
par(mfrow=c(2,2))

plot(fitted(fit), residuals(fit), xlab="fitted value", ylab="Residuals")

plot(carb$x1, residuals(fit), xlab="x1", ylab="Residuals")

plot(carb$x2, residuals(fit), xlab="x2", ylab="Residuals")
```



For the plots of fitted value vs residuals and x_2 vs residuals, there is a qudratic pattern. For the plot of x_1 vs residuals, the absoluate value of residuals of $x_1 = 31$ is greater than the other x_1 values. Hence the fitted model is not adequate.

(b)

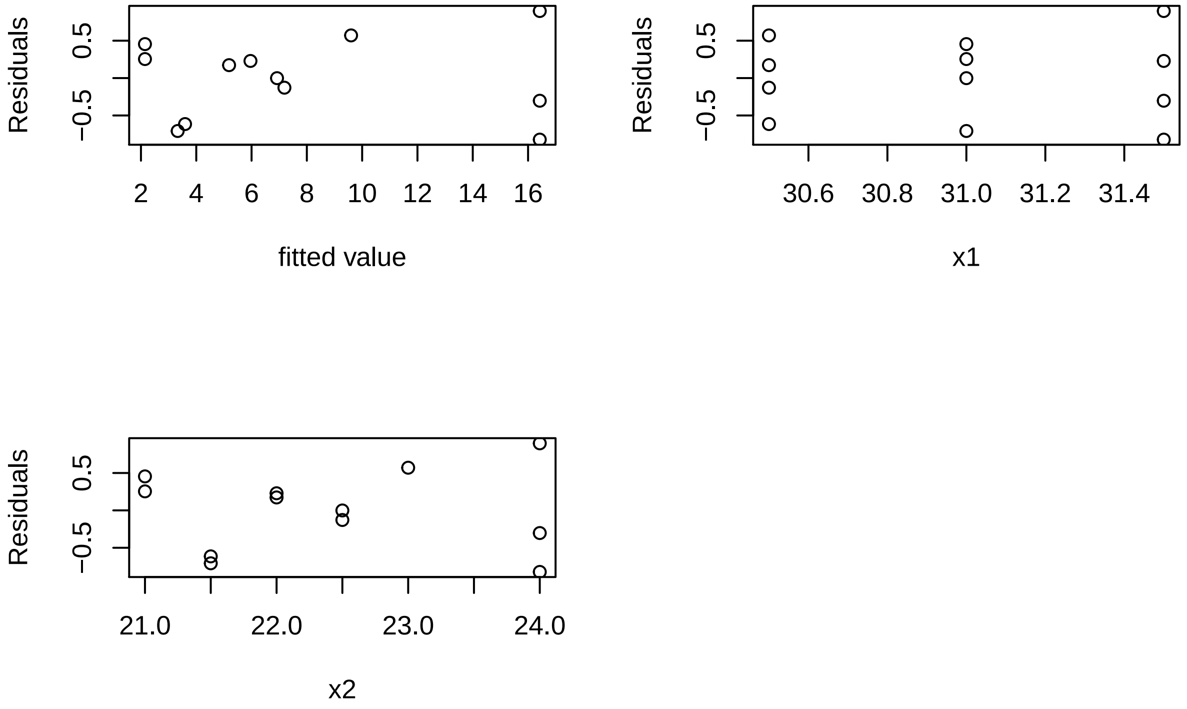
```
fit2 = lm(y~poly(x1, 2) + poly(x2, 2), carb)
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ poly(x1, 2) + poly(x2, 2), data = carb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82305 -0.38108  0.08586  0.30455  0.89695
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.9450     0.1936  41.032 1.33e-09 ***
## poly(x1, 2)1   1.0943     0.9612   1.138   0.2924
## poly(x1, 2)2   1.0682     0.9296   1.149   0.2883
## poly(x2, 2)1  16.9902     1.0532  16.132 8.55e-07 ***
## poly(x2, 2)2   2.6846     0.8239   3.258   0.0139 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6708 on 7 degrees of freedom
## Multiple R-squared:  0.9908, Adjusted R-squared:  0.9855
## F-statistic: 188.4 on 4 and 7 DF,  p-value: 3.342e-07
```

${x_2}$ and ${x_2}^2$ are significant at significant level 0.05. ${x_1}$ and ${x_1}^2$ are not significant.

(c)

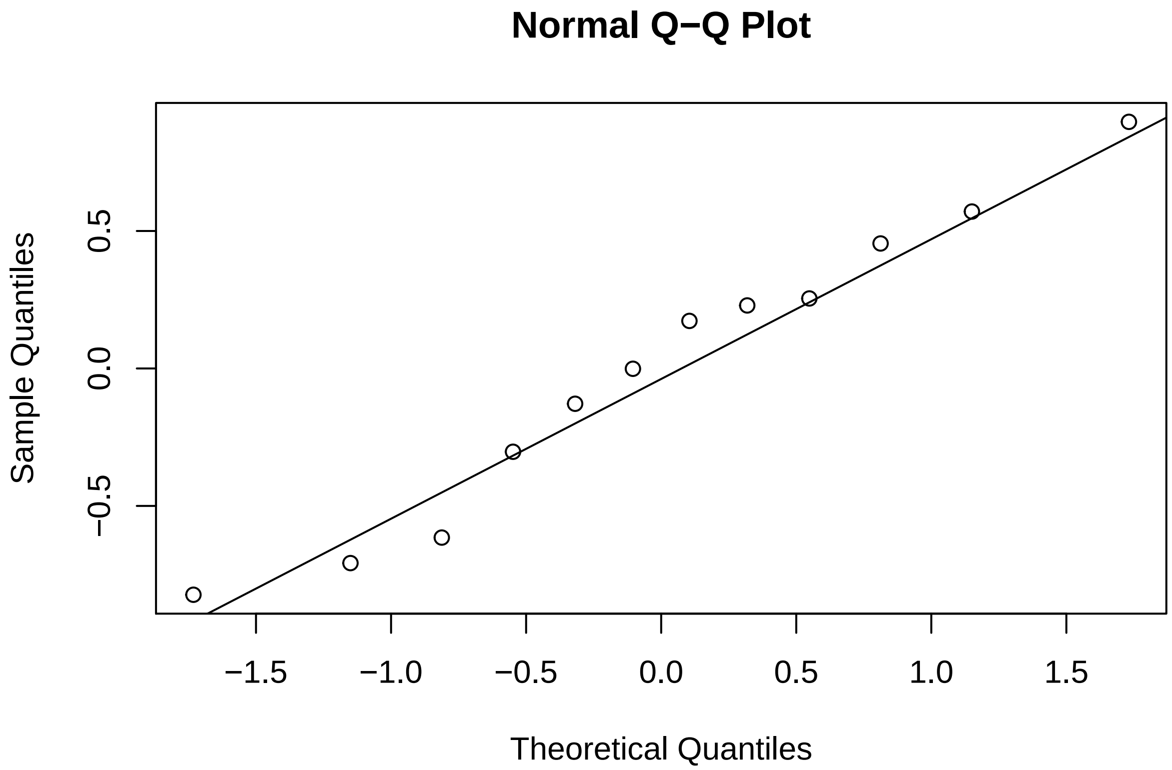
```r
par(mfrow=c(2,2))

plot(fitted(fit2), residuals(fit2), xlab="fitted value", ylab="Residuals")

plot(carb$x1, residuals(fit2), xlab="x1", ylab="Residuals")

plot(carb$x2, residuals(fit2), xlab="x2", ylab="Residuals")
```



There are no systematic patterns in any plots, and the residuals lie within a band around 0. Hence, the fitted model is adequate.

(d)

```r
qqnorm(residuals(fit2))
qqline(residuals(fit2))
```



The points in QQ plot look approxiamtely in a straight line. Hence, the residual is normally distributed.

(e)

```r
summary(fit)
```

```
## 
## Call:
## lm(formula = y ~ ., data = carb)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.4047 -0.4936  0.0860  0.5473  1.3004 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -147.4892    21.3572  -6.906 7.02e-05 ***
## x1             1.7188     0.7629   2.253   0.0508 .  
## x2             4.5570     0.2892  15.756 7.35e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9501 on 9 degrees of freedom
## Multiple R-squared:  0.9763, Adjusted R-squared:  0.971 
## F-statistic:   185 on 2 and 9 DF,  p-value: 4.896e-08
```

The adjusted R^2 of the model in part(b) is greater than the model in part(a), so we prefer the model in part(b).

##q2

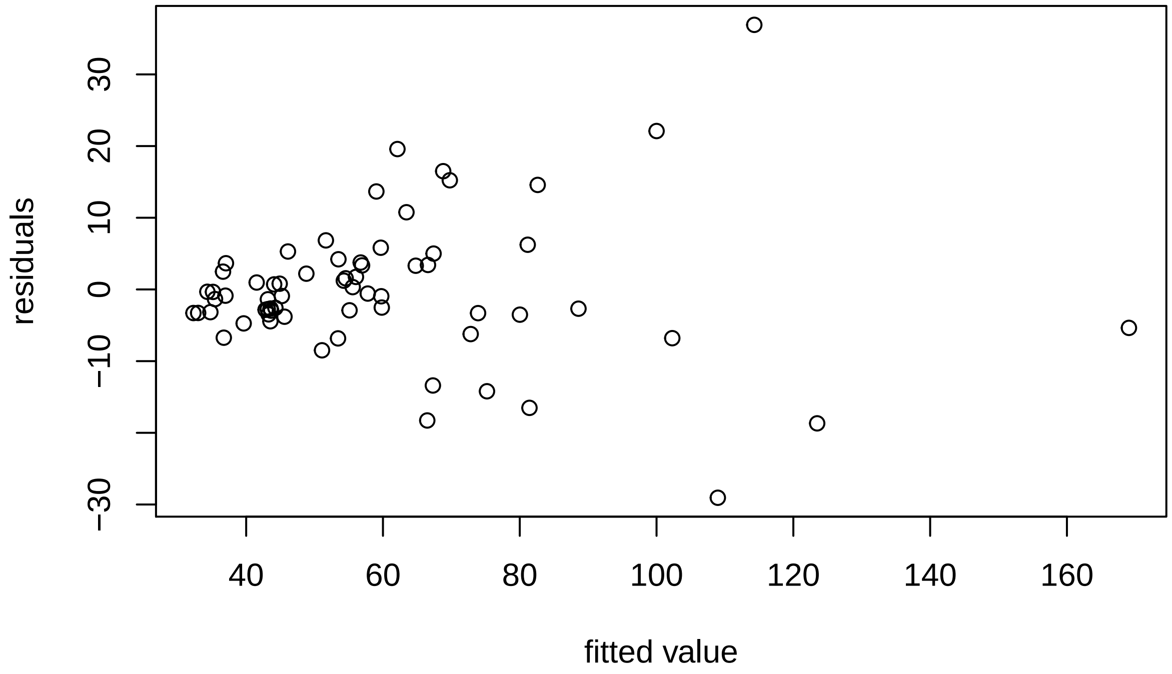
```
sal = read.table("salary.txt", header = FALSE)
colnames(sal) = c("y", "degree", "exp", "sup")
```

(a)

```
deg = factor(sal$degree)
sallm = lm(y ~ deg + exp + sup, data = sal)
summary(sallm)
```

```
##
## Call:
## lm(formula = y ~ deg + exp + sup, data = sal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.058  -3.477  -0.915   3.417  36.909
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.4714     2.8691  10.969 5.73e-16 ***
## deg2         10.8120     3.2183   3.360  0.00136 **
## deg3         22.6307     3.4846   6.494 1.81e-08 ***
## exp           1.2581     0.2273   5.535 7.23e-07 ***
## sup           1.8523     0.2276   8.137 2.86e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.14 on 60 degrees of freedom
## Multiple R-squared:  0.8633, Adjusted R-squared:  0.8542
## F-statistic: 94.76 on 4 and 60 DF,  p-value: < 2.2e-16
```

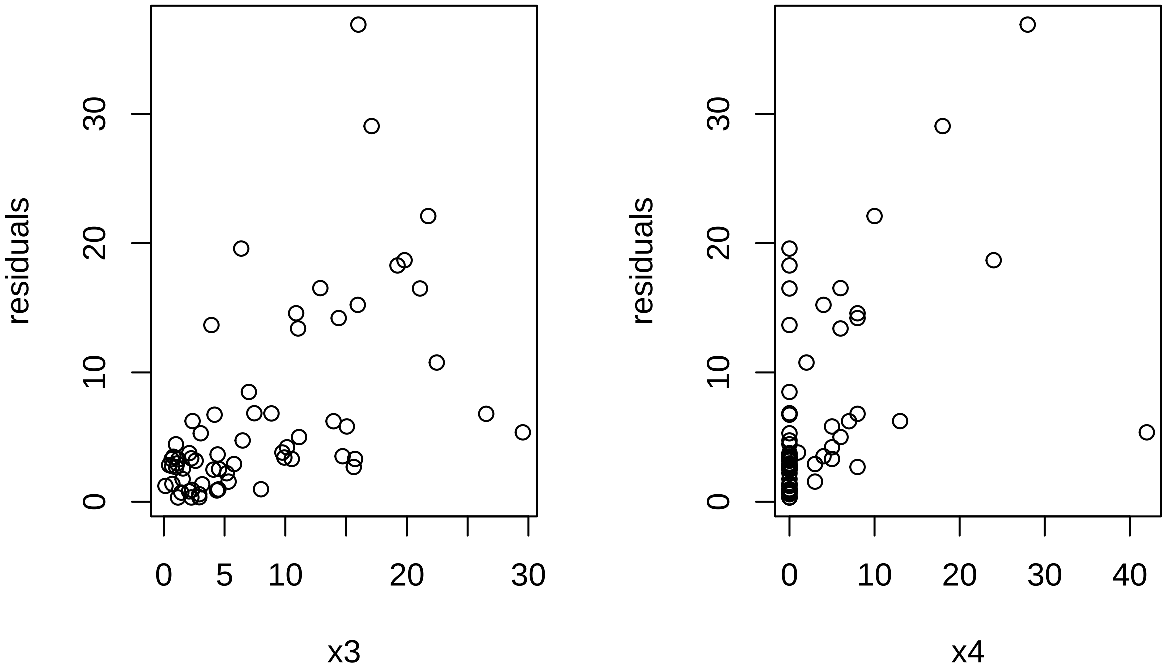
```
plot(fitted(sallm), residuals(sallm), xlab = "fitted value", ylab = "residuals")
```



The residuals are fan shaped, the variances of random errors are non-constant.

(b)

```
par(mfrow = c(1,2))
plot(sal$exp, abs(residuals(sallm)), xlab = "x3", ylab = "residuals")
plot(sal$sup, abs(residuals(sallm)), xlab = "x4", ylab = "residuals")
```



The residuals are fan shaped, the variances of random errors are non-constant.

(c)

```r
r = abs(residuals(sallm))
rsd = lm(r~sal$exp + sal$sup)
1/(fitted(rsd)^2)
```

```
##           1           2           3           4           5           6 
## 0.056292687 0.077705686 0.001532505 0.024530644 0.163067895 0.010251537 
##           7           8           9          10          11          12 
## 0.090311060 0.118907587 0.034820755 0.013179549 0.007020100 0.073711480 
##          13          14          15          16          17          18 
## 0.083172030 0.010441314 0.077705686 0.090528372 0.023024234 0.011372887 
##          19          20          21          22          23          24 
## 0.005242312 0.008537492 0.119896882 0.089023467 0.005026257 0.034323007 
##          25          26          27          28          29          30 
## 0.055657611 0.009376487 0.032596332 0.094578048 0.111415948 0.125743297 
##          31          32          33          34          35          36 
## 0.015718961 0.009792584 0.049591497 0.056938695 0.075669068 0.125033549 
##          37          38          39          40          41          42 
## 0.016351763 0.004340427 0.039776372 0.010224118 0.059753781 0.011854165 
##          43          44          45          46          47          48 
## 0.003734147 0.013888717 0.136752318 0.108216859 0.062908841 0.057595888 
##          49          50          51          52          53          54 
## 0.133575200 0.139210138 0.137564309 0.122596179 0.107368283 0.012544414 
##          55          56          57          58          59          60 
## 0.008502906 0.036729680 0.060818508 0.028140893 0.013993969 0.021181282
##          61          62          63          64          65
## 0.040548297 0.031689610 0.148355382 0.094114807 0.003540910
```

(d)

```
sal$wts = 1/(fitted(rsd)^2)
sallm2 = lm(y ~ deg + exp + sup, weights = wts, data = sal)
summary(sallm2)
```

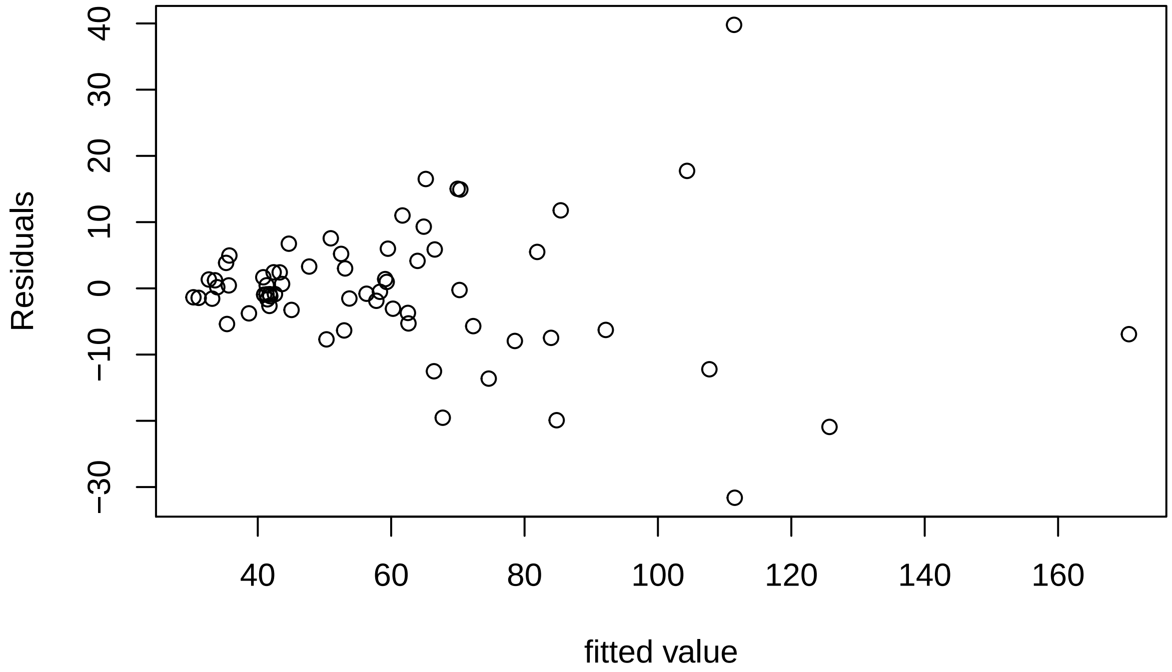
```
##
## Call:
## lm(formula = y ~ deg + exp + sup, data = sal, weights = wts)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2414 -0.7531 -0.2709  0.6915  3.3246
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.4255     1.3617  21.610  < 2e-16 ***
## deg2         10.8996     1.4918   7.307 7.50e-10 ***
## deg3         26.6849     1.6686  15.992  < 2e-16 ***
## exp           1.4253     0.2002   7.118 1.57e-09 ***
## sup           1.7239     0.3206   5.377 1.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.15 on 60 degrees of freedom
## Multiple R-squared:  0.8874, Adjusted R-squared:  0.8799
## F-statistic: 118.2 on 4 and 60 DF,  p-value: < 2.2e-16
```

These estimates are similar to the the estimates in part(a).

(e) most of the deviations are less that part(a). So the second model is more precise.

(f)

```
plot(fitted(sallm2), residuals(sallm2), xlab="fitted value", ylab="Residuals")
```



The residual plot is still fan shaped. Since the estimates from two models are relatively similar, we expect a similar residual plot.

“‘