Statistics.R

015004

Tue Feb 19 11:48:24 2019

```
#################
# Load packages #
#################

library(JM)
```

```
## Warning: package 'JM' was built under R version 3.5.2
```

```
## Loading required package: MASS
```

```
## Loading required package: nlme
```

```
## Loading required package: splines
```

```
## Loading required package: survival
```

```
## Warning: package 'survival' was built under R version 3.5.2
```

```
###############
# Statistics #
###############

# Statistical Tests

## t-test
t.test(pbc2.id$serBilir[pbc2.id$drug == "D-penicil"],
       pbc2.id$serBilir[pbc2.id$drug == "placebo"])
```

```
##
##  Welch Two Sample t-test
##
## data:  pbc2.id$serBilir[pbc2.id$drug == "D-penicil"] and pbc2.id$serBilir[pbc2.id$drug == "placebo"]
## t = -1.6771, df = 265.18, p-value = 0.09469
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.8560805  0.1485513
## sample estimates:
## mean of x mean of y
##  2.794937  3.648701
```

```
## anova
aov(pbc2.id$age ~ pbc2.id$status)
```

```
## Call:
##    aov(formula = pbc2.id$age ~ pbc2.id$status)
##
## Terms:
##                 pbc2.id$status Residuals
## Sum of Squares        4783.052 30039.371
## Deg. of Freedom              2       309
##
```

```
## Residual standard error: 9.859756
## Estimated effects may be unbalanced
```

```
summary(aov(pbc2.id$age ~ pbc2.id$status))
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## pbc2.id$status   2   4783  2391.5    24.6 1.22e-10 ***
## Residuals      309  30039    97.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov(pbc2.id$age ~ pbc2.id$status + pbc2.id$sex)
```

```
## Call:
##    aov(formula = pbc2.id$age ~ pbc2.id$status + pbc2.id$sex)
##
## Terms:
##                 pbc2.id$status pbc2.id$sex Residuals
## Sum of Squares        4783.052     925.838 29113.533
## Deg. of Freedom              2           1       308
##
## Residual standard error: 9.722369
## Estimated effects may be unbalanced
```

```
summary(aov(pbc2.id$age ~ pbc2.id$status + pbc2.id$sex))
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## pbc2.id$status   2   4783  2391.5  25.301 6.71e-11 ***
## pbc2.id$sex      1    926   925.8   9.795  0.00192 **
## Residuals      308  29114    94.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## wilcoxon
wilcox.test(pbc2.id$serBilir[pbc2.id$drug == "D-penicil"],
            pbc2.id$serBilir[pbc2.id$drug == "placebo"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pbc2.id$serBilir[pbc2.id$drug == "D-penicil"] and pbc2.id$serBilir[pbc2.id$drug == "placebo"]
## W = 11951, p-value = 0.7876
## alternative hypothesis: true location shift is not equal to 0
```

```
## kruskal test
kruskal.test(list(pbc2.id$age[pbc2.id$status == "alive"],
                  pbc2.id$age[pbc2.id$status == "transplanted"],
                  pbc2.id$age[pbc2.id$status == "dead"]))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(pbc2.id$age[pbc2.id$status == "alive"], pbc2.id$age[pbc2.id$status ==     "transplanted"]
## Kruskal-Wallis chi-squared = 42.264, df = 2, p-value = 6.645e-10
```

```
kruskal.test(pbc2.id$age ~ pbc2.id$status)
```

```
##
```

```
##  Kruskal-Wallis rank sum test
##
## data:  pbc2.id$age by pbc2.id$status
## Kruskal-Wallis chi-squared = 42.264, df = 2, p-value = 6.645e-10
```

```
## chi-sqared test
tbl <- table(pbc2.id$status, pbc2.id$drug)
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 1.1822, df = 2, p-value = 0.5537
```

```
## fisher test
fisher.test(tbl)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tbl
## p-value = 0.5572
## alternative hypothesis: two.sided
```

```
## correlations
cor(pbc2.id$age, pbc2.id$serBilir)
```

```
## [1] 0.02785516
```

```
cor(pbc2.id$age, pbc2.id$serBilir, method = "spearman")
```

```
## [1] -0.01338391
```

```
cor.test(pbc2.id$age, pbc2.id$serBilir)
```

```
##
##  Pearson's product-moment correlation
##
## data:  pbc2.id$age and pbc2.id$serBilir
## t = 0.49063, df = 310, p-value = 0.624
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08344162  0.13846558
## sample estimates:
##        cor
## 0.02785516
```

```
cor.test(pbc2.id$age, pbc2.id$serBilir, method = "spearman")
```

```
## Warning in cor.test.default(pbc2.id$age, pbc2.id$serBilir, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  pbc2.id$age and pbc2.id$serBilir
## S = 5129600, p-value = 0.8138
## alternative hypothesis: true rho is not equal to 0
```

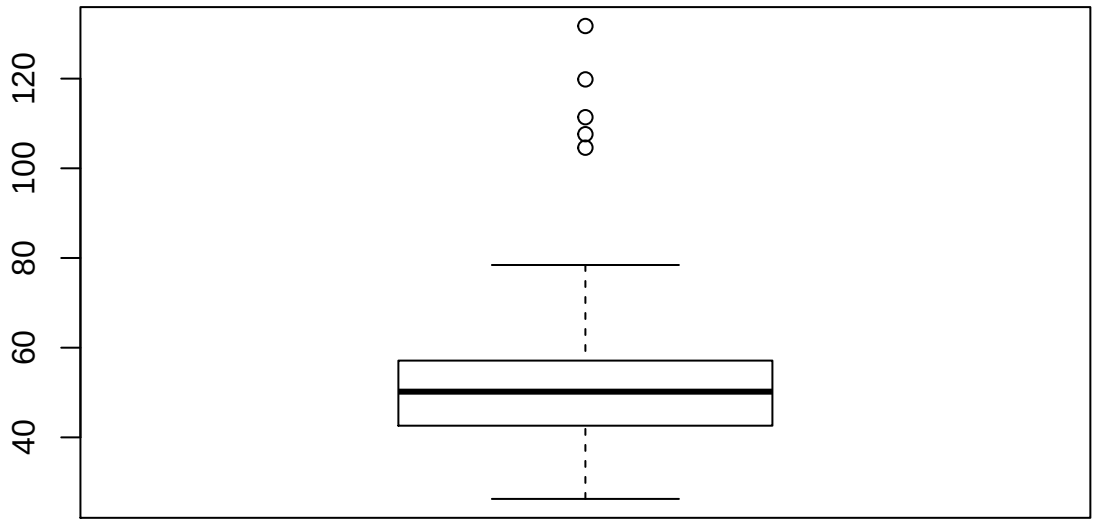
```
## sample estimates:
##         rho
## -0.01338391
```

```
# Outliers

## Create some

ind <- sample(pbc2.id$id, 5)
pbc2.id$age[pbc2.id$id %in% ind] <-
  pbc2.id$age[pbc2.id$id %in% ind] + 70

boxplot(pbc2.id$age)
```




```
head(pbc2.id[pbc2.id$age <=100, ])
```

```
##   id     years       status      drug      age    sex year ascites
## 1  1  1.095170         dead D-penicil 58.76684 female    0     Yes
## 2  2 14.152338        alive D-penicil 56.44782 female    0      No
## 3  3  2.770781         dead D-penicil 70.07447   male    0      No
## 4  4  5.270507         dead D-penicil 54.74209 female    0      No
## 5  5  4.120578 transplanted   placebo 38.10645 female    0      No
## 6  6  6.853028         dead   placebo 66.26054 female    0      No
##   hepatomegaly spiders                   edema serBilir serChol albumin
## 1          Yes     Yes edema despite diuretics     14.5     261    2.60
## 2          Yes     Yes                No edema      1.1     302    4.14
## 3           No      No      edema no diuretics      1.4     176    3.48
```

```
## 4          Yes     Yes      edema no diuretics      1.8     244    2.54
## 5          Yes     Yes                No edema      3.4     279    3.53
## 6          Yes      No                No edema      0.8     248    3.98
##   alkaline  SGOT platelets prothrombin histologic status2
## 1     1718 138.0       190        12.2          4       1
## 2     7395 113.5       221        10.6          3       0
## 3      516  96.1       151        12.0          4       1
## 4     6122  60.6       183        10.3          4       1
## 5      671 113.2       136        10.9          3       0
## 6      944  93.0        NA        11.0          3       1
```

```
## exclide them
pbc2.id <- pbc2.id[pbc2.id$age <=100, ]

# Regression Models

## linear regression
fm1 <- lm(serBilir ~ age + sex + drug, data = pbc2.id)
summary(fm1)
```

```
##
## Call:
## lm(formula = serBilir ~ age + sex + drug, data = pbc2.id)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4939 -2.4177 -1.5721  0.4211 24.2872
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.26725    1.58664   1.429   0.1540
## age            0.01967    0.02498   0.787   0.4317
## sexfemale      0.41685    0.81397   0.512   0.6089
## drugD-penicil -0.85833    0.51653  -1.662   0.0976 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.484 on 303 degrees of freedom
## Multiple R-squared:  0.01079,    Adjusted R-squared:  0.0009979
## F-statistic: 1.102 on 3 and 303 DF,  p-value: 0.3486
```

```
coef(fm1)
```

```
##   (Intercept)           age     sexfemale drugD-penicil
##    2.26724955    0.01967251    0.41685040   -0.85832651
```

```
head(fitted(fm1))
```

```
##        1        2        3        4        5        6
## 2.981865 2.936244 2.787464 2.902688 3.433750 3.987611
```

```
head(residuals(fm1))
```

```
##           1           2           3           4           5           6
## 11.51813524 -1.83624375 -1.38746391 -1.10268780 -0.03374954 -3.18761122
```

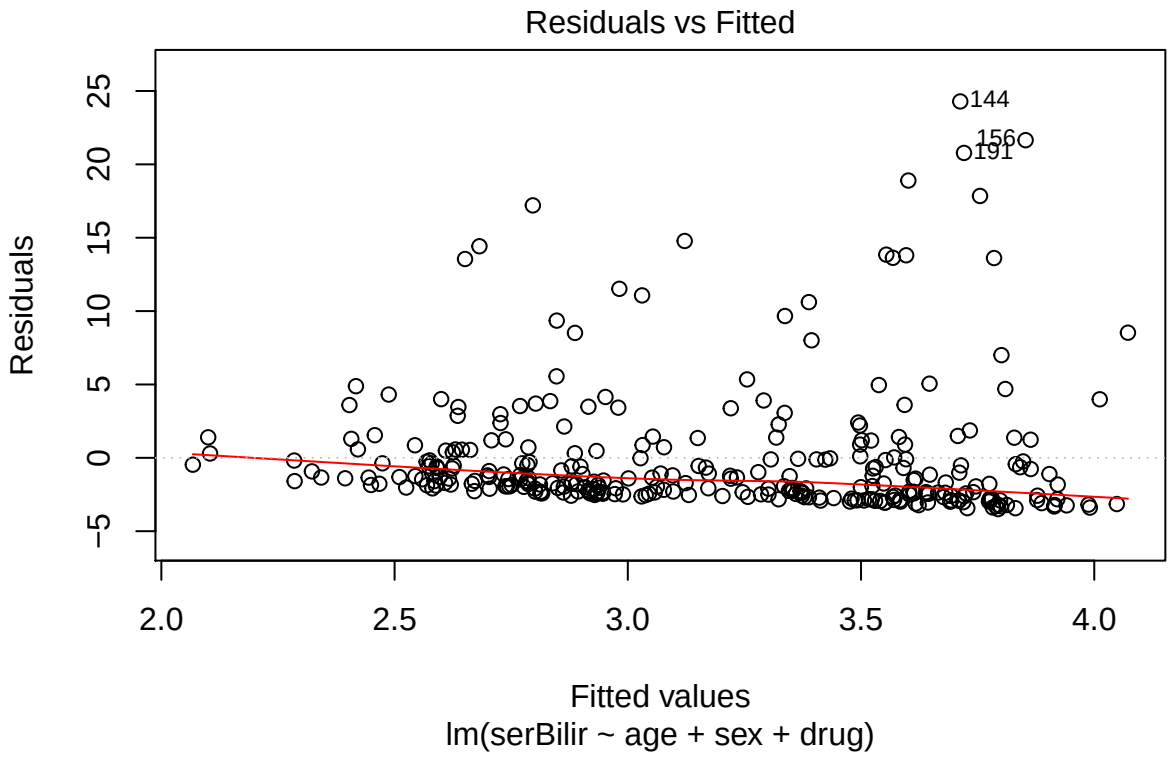
```
AIC(fm1)
```

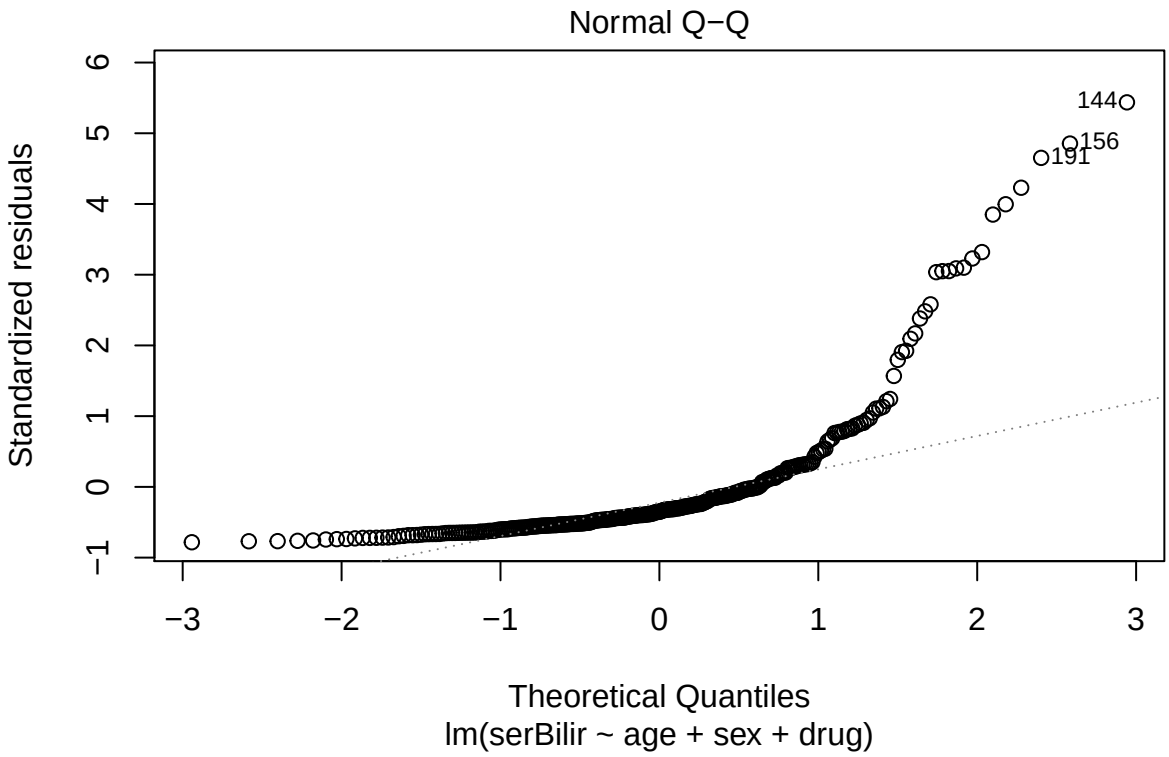
```
## [1] 1798.492
```

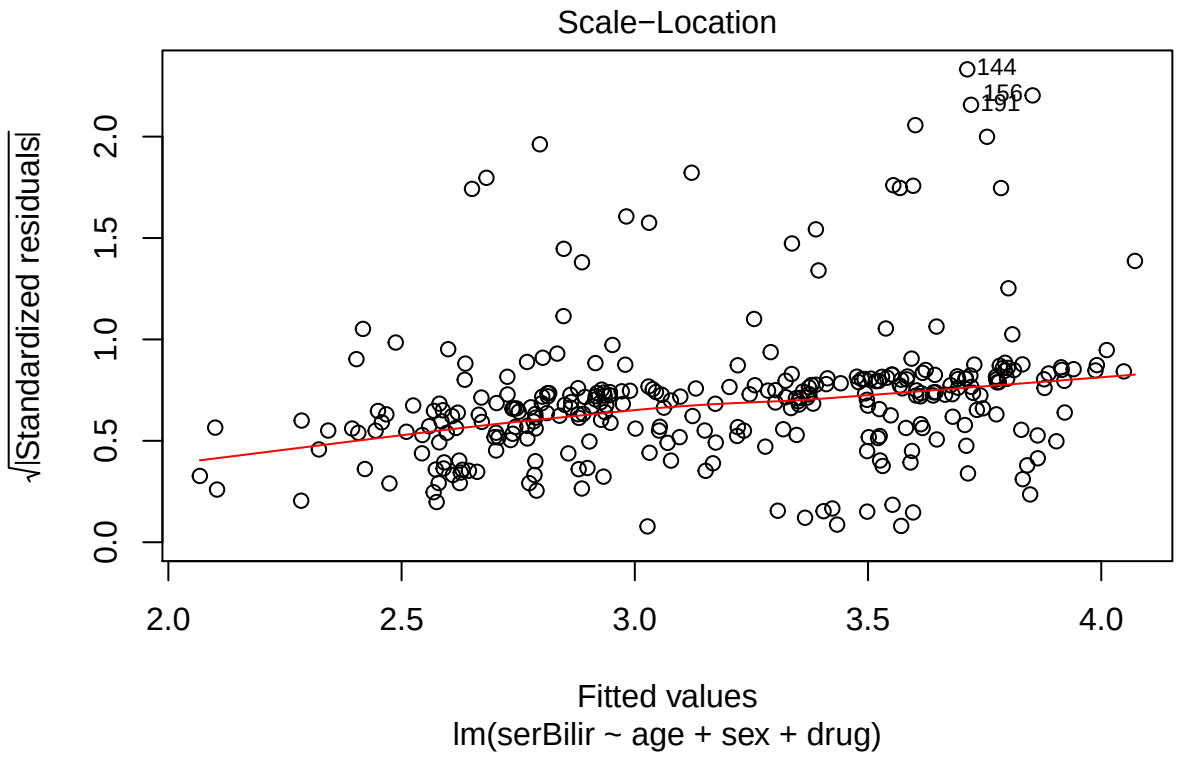
```
confint(fm1)
```

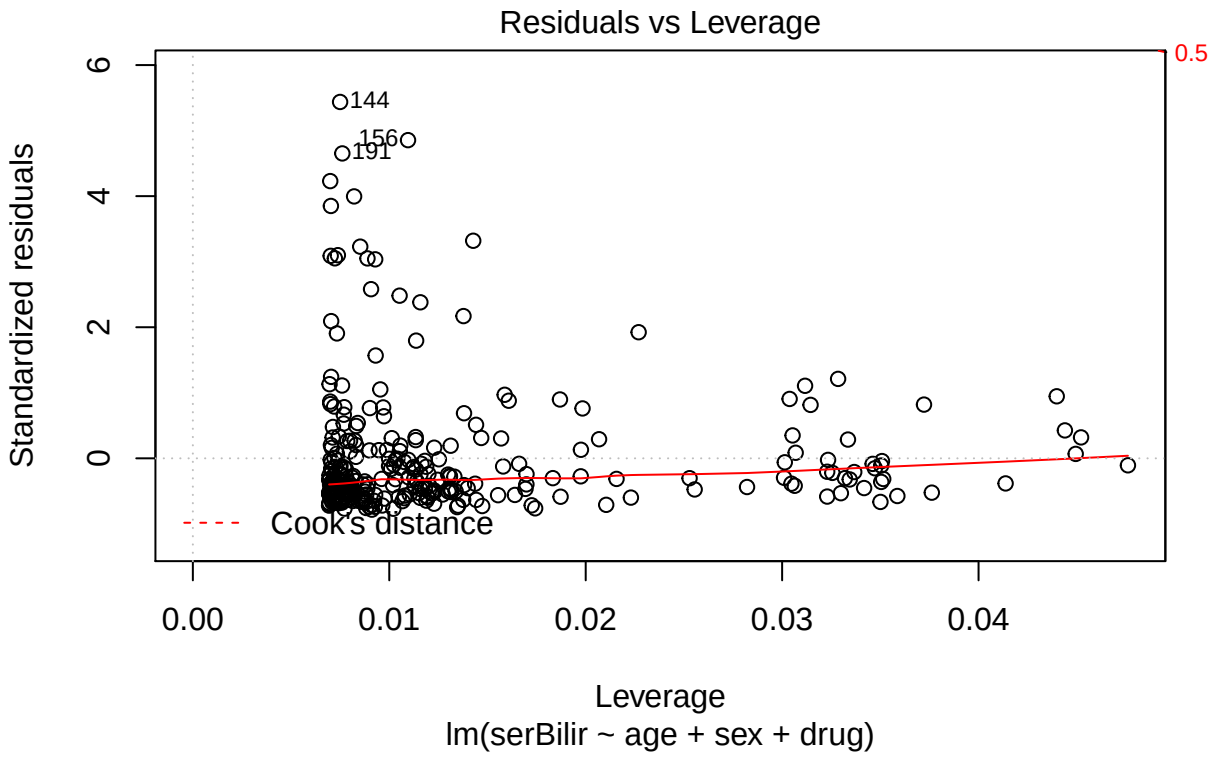
```
##                      2.5 %     97.5 %
## (Intercept)   -0.85497307 5.38947217
## age           -0.02949139 0.06883641
## sexfemale     -1.18489937 2.01860016
## drugD-penicil -1.87476717 0.15811415
```

```
# plotting the fitted model
plot(fm1)
```

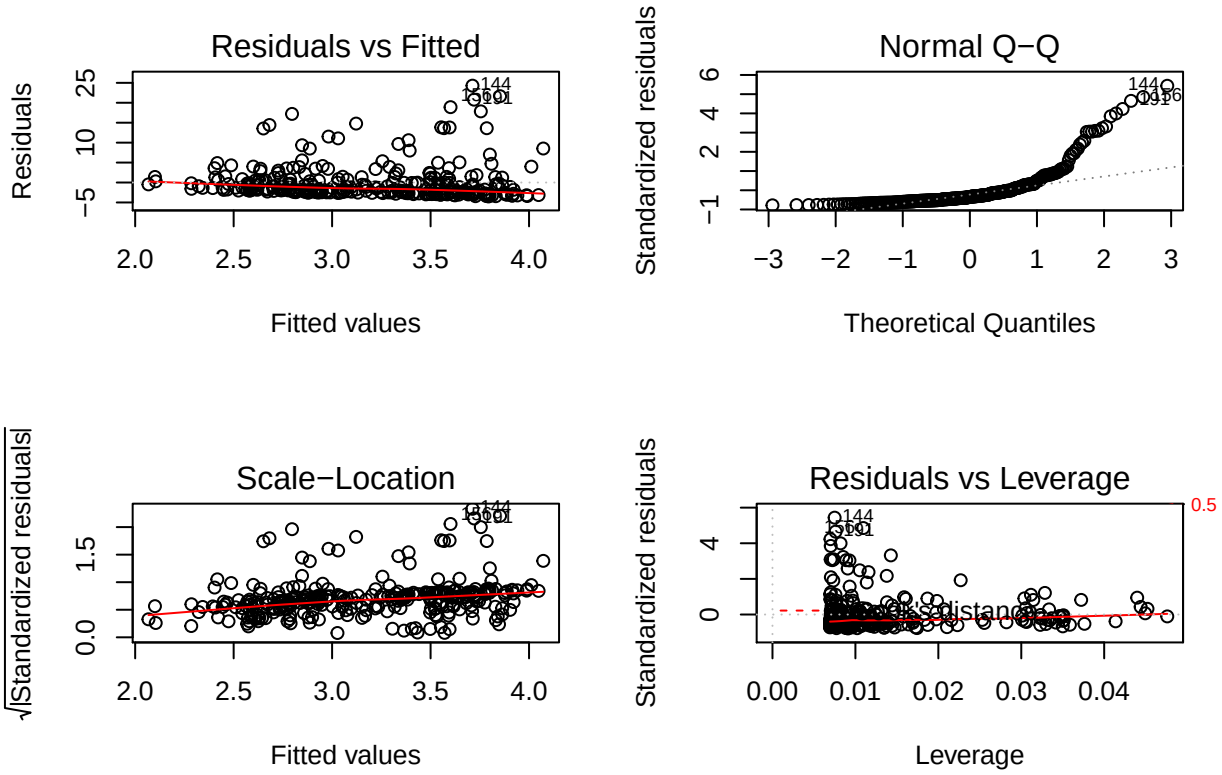



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(serBilir ~ age + sex + drug)
144
156
191


Scale−Location
144
156
191
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(serBilir ~ age + sex + drug)




```
par(mfrow = c(2, 2))
plot(fm1)
```




```
# exclude intercept
fm1b <- lm(serBilir ~ -1 + age + sex + drug, data = pbc2.id)
summary(fm1b)
```

```
##
## Call:
## lm(formula = serBilir ~ -1 + age + sex + drug, data = pbc2.id)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4939 -2.4177 -1.5721  0.4211 24.2872
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## age            0.01967    0.02498   0.787   0.4317
## sexmale        2.26725    1.58664   1.429   0.1540
## sexfemale      2.68410    1.25674   2.136   0.0335 *
## drugD-penicil -0.85833    0.51653  -1.662   0.0976 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.484 on 303 degrees of freedom
## Multiple R-squared:  0.343,  Adjusted R-squared:  0.3343
## F-statistic: 39.54 on 4 and 303 DF,  p-value: < 2.2e-16
```

```
# interaction effects
fm2 <- lm(serBilir ~ age + sex + drug + age:sex, data = pbc2.id)
```

```r
summary(fm2)
```

```
##
## Call:
## lm(formula = serBilir ~ age + sex + drug + age:sex, data = pbc2.id)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4704 -2.4150 -1.5897  0.4705 24.3000
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.39061    3.80827   0.365    0.715
## age             0.03517    0.06610   0.532    0.595
## sexfemale       1.41757    4.03398   0.351    0.726
## drugD-penicil  -0.84855    0.51877  -1.636    0.103
## age:sexfemale  -0.01811    0.07151  -0.253    0.800
##
## Residual standard error: 4.491 on 302 degrees of freedom
## Multiple R-squared:  0.011,  Adjusted R-squared:  -0.002097
## F-statistic: 0.8399 on 4 and 302 DF,  p-value: 0.5007
```

```r
fm2b <- lm(serBilir ~ age*drug + age*sex, data = pbc2.id)
summary(fm2b)
```

```
##
## Call:
## lm(formula = serBilir ~ age * drug + age * sex, data = pbc2.id)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5078 -2.4050 -1.5538  0.4251 24.2822
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.087062   4.182282   0.260    0.795
## age                 0.041076   0.074141   0.554    0.580
## drugD-penicil      -0.409108   2.536854  -0.161    0.872
## sexfemale           1.490052   4.061170   0.367    0.714
## age:drugD-penicil  -0.008786   0.049643  -0.177    0.860
## age:sexfemale      -0.019260   0.071913  -0.268    0.789
##
## Residual standard error: 4.498 on 301 degrees of freedom
## Multiple R-squared:  0.01111,    Adjusted R-squared:  -0.005322
## F-statistic: 0.676 on 5 and 301 DF,  p-value: 0.6419
```

```r
# polynomial effects
fm3 <- lm(serBilir ~ age + I(age^2) + I(age^3), data = pbc2.id)
summary(fm3)
```

```
##
## Call:
## lm(formula = serBilir ~ age + I(age^2) + I(age^3), data = pbc2.id)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -2.9899 -2.4119 -1.6783  0.2648 24.6996
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.855e+00  1.900e+01  -0.519    0.604
## age          7.359e-01  1.159e+00   0.635    0.526
## I(age^2)    -1.347e-02  2.290e-02  -0.588    0.557
## I(age^3)     8.037e-05  1.467e-04   0.548    0.584
##
## Residual standard error: 4.502 on 303 degrees of freedom
## Multiple R-squared:  0.002582,   Adjusted R-squared:  -0.007293
## F-statistic: 0.2615 on 3 and 303 DF,  p-value: 0.8531
```

```
# include smooth terms
library(splines)
fm3b <- lm(serBilir ~ ns(age, df = 3), data = pbc2.id)
summary(fm3b)
```

```
##
## Call:
## lm(formula = serBilir ~ ns(age, df = 3), data = pbc2.id)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9780 -2.4109 -1.6915  0.2714 24.7063
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.9465     1.4712   1.323    0.187
## ns(age, df = 3)1    0.4979     1.1223   0.444    0.658
## ns(age, df = 3)2    2.7753     3.4760   0.798    0.425
## ns(age, df = 3)3    0.6256     1.7536   0.357    0.722
##
## Residual standard error: 4.503 on 303 degrees of freedom
## Multiple R-squared:  0.002525,   Adjusted R-squared:  -0.007351
## F-statistic: 0.2557 on 3 and 303 DF,  p-value: 0.8573
```

```
# compare models
fm4 <- lm(serBilir ~ age, pbc2.id)
fm5 <- lm(serBilir ~ age * sex, pbc2.id)
anova(fm4, fm5)
```

```
## Analysis of Variance Table
##
## Model 1: serBilir ~ age
## Model 2: serBilir ~ age * sex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    305 6153.5
## 2    303 6144.3  2    9.1341 0.2252 0.7985
```

```
## logistic regration
gl1 <- glm(drug ~ age, data = pbc2.id, family = binomial)

gl2 <- glm(drug ~ age + sex, data = pbc2.id, family = binomial)
```

```r
summary(gl1)
```

```
##
## Call:
## glm(formula = drug ~ age, family = binomial, data = pbc2.id)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4623  -1.1648   0.9125   1.1462   1.4255
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.22847    0.56474  -2.175   0.0296 *
## age          0.02519    0.01105   2.279   0.0227 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 425.51  on 306  degrees of freedom
## Residual deviance: 420.20  on 305  degrees of freedom
## AIC: 424.2
##
## Number of Fisher Scoring iterations: 4
```

```r
summary(gl2)
```

```
##
## Call:
## glm(formula = drug ~ age + sex, family = binomial, data = pbc2.id)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.523  -1.167   0.908   1.150   1.423
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.00959    0.71524  -1.412    0.158
## age          0.02406    0.01128   2.132    0.033 *
## sexfemale   -0.18354    0.36957  -0.497    0.619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 425.51  on 306  degrees of freedom
## Residual deviance: 419.96  on 304  degrees of freedom
## AIC: 425.96
##
## Number of Fisher Scoring iterations: 4
```

```r
confint(gl1)
```

```
## Waiting for profiling to be done...
##                   2.5 %     97.5 %
```

```
## (Intercept) -2.35049430 -0.1312490
## age          0.00373369  0.0471634
```

```
anova(gl1, gl2)
```

```
## Analysis of Deviance Table
##
## Model 1: drug ~ age
## Model 2: drug ~ age + sex
##   Resid. Df Resid. Dev Df Deviance
## 1       305     420.20
## 2       304     419.96  1  0.24783
```

```
anova(gl1, gl2, test ="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: drug ~ age
## Model 2: drug ~ age + sex
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       305     420.20
## 2       304     419.96  1  0.24783   0.6186
```

```
exp(cbind(coef(gl2), confint(gl2))) # odds ratio
```

```
## Waiting for profiling to be done...
```

```
##                              2.5 %   97.5 %
## (Intercept) 0.3643684 0.08850485 1.474621
## age         1.0243486 1.00212982 1.047569
## sexfemale   0.8323212 0.39795158 1.711850
```