Model selection, Assumption checking and Model Evaluation

470518795

27/10/2019

```
library(MASS)
library(dplyr)
```

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

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

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

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

```
ourdata <- birthwt %>% mutate(
  race = factor(race, labels = c("white", "black", "other")), #organise race
  smoke = factor(smoke, labels = c("False", "True")), #organise smokes
  ui =factor(ui,labels=c("Yes","No")),
  low=factor(low),
  ht=factor(ht)

)
ourdata<-dplyr::select(ourdata,-c(low))
ourdata
```

Model selection

```
# 1: Run a full multiple regression
bwt_lm1=lm(ourdata$bwt ~ .,data = ourdata)
summary(bwt_lm1)
```

```
##
## Call:
## lm(formula = ourdata$bwt ~ ., data = ourdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1825.26  -435.21    55.91   473.46  1701.20 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2927.962    312.904   9.357  < 2e-16 ***
## age           -3.570      9.620  -0.371 0.711012    
## lwt            4.354      1.736   2.509 0.013007 *  
## raceblack   -488.428    149.985  -3.257 0.001349 ** 
## raceother   -355.077    114.753  -3.094 0.002290 ** 
## smokeTrue   -352.045    106.476  -3.306 0.001142 ** 
## ptl          -48.402    101.972  -0.475 0.635607    
## ht1         -592.827    202.321  -2.930 0.003830 ** 
## uiNo        -516.081    138.885  -3.716 0.000271 ***
## ftv          -14.058     46.468  -0.303 0.762598    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 650.3 on 179 degrees of freedom
## Multiple R-squared:  0.2427, Adjusted R-squared:  0.2047 
## F-statistic: 6.376 on 9 and 179 DF,  p-value: 7.891e-08
```

```r
# 2: Backward search using AIC
step_back_aic<-step(bwt_lm1,direction = "backward",trace = F)
step_back_aic
```

```
## 
## Call:
## lm(formula = ourdata$bwt ~ lwt + race + smoke + ht + ui, data = ourdata)
## 
## Coefficients:
## (Intercept)          lwt    raceblack    raceother    smokeTrue          ht1  
##    2837.264        4.242     -475.058     -348.150     -356.321     -585.193  
##        uiNo  
##    -525.524  
```

```r
summary(step_back_aic)
```

```
## 
## Call:
## lm(formula = ourdata$bwt ~ lwt + race + smoke + ht + ui, data = ourdata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1842.14  -433.19    67.09   459.21  1631.03 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2837.264    243.676  11.644  < 2e-16 ***
## lwt            4.242      1.675   2.532 0.012198 *  
## raceblack   -475.058    145.603  -3.263 0.001318 ** 
## raceother   -348.150    112.361  -3.099 0.002254 ** 
## smokeTrue   -356.321    103.444  -3.445 0.000710 ***
## ht1         -585.193    199.644  -2.931 0.003810 ** 
## uiNo        -525.524    134.675  -3.902 0.000134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.9 on 182 degrees of freedom
## Multiple R-squared:  0.2404, Adjusted R-squared:  0.2154
## F-statistic:   9.6 on 6 and 182 DF,  p-value: 3.601e-09
```

```
# 3: Run a null multiple regression model
bwt_lm2=lm(ourdata$bwt ~ 1,data = ourdata)
# 4: Forward search using AIC
step_fwd_aic=step(bwt_lm2,scope = list(lower=bwt_lm2,upper=bwt_lm1),direction = "forward",trace = F)
summary(step_fwd_aic)
```

```
##
## Call:
## lm(formula = ourdata$bwt ~ ui + race + smoke + ht + lwt, data = ourdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1842.14  -433.19    67.09   459.21  1631.03
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2837.264    243.676  11.644  < 2e-16 ***
## uiNo        -525.524    134.675  -3.902 0.000134 ***
## raceblack   -475.058    145.603  -3.263 0.001318 **
## raceother   -348.150    112.361  -3.099 0.002254 **
## smokeTrue   -356.321    103.444  -3.445 0.000710 ***
## ht1         -585.193    199.644  -2.931 0.003810 **
## lwt            4.242      1.675   2.532 0.012198 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.9 on 182 degrees of freedom
## Multiple R-squared:  0.2404, Adjusted R-squared:  0.2154
## F-statistic:   9.6 on 6 and 182 DF,  p-value: 3.601e-09
```

```
# By both forward and backwards search alogarithm with AIC, we get the same results.
```

```
# Then we try to adding or subtracting other variables, to see whether we can get a smaller aic, howeve
add1(step_fwd_aic,test="F",scope=bwt_lm1)
```

```
## Single term additions
##
## Model:
## ourdata$bwt ~ ui + race + smoke + ht + lwt
##        Df Sum of Sq      RSS    AIC F value Pr(>F)
## <none>              75937505 2452.8
## age     1    104920 75832585 2454.5  0.2504 0.6174
## ptl     1    117366 75820139 2454.5  0.2802 0.5972
## ftv     1     58307 75879197 2454.7  0.1391 0.7096
```

```
drop1(step_fwd_aic)
```

```
## Single term deletions
##
## Model:
## ourdata$bwt ~ ui + race + smoke + ht + lwt
##        Df Sum of Sq      RSS    AIC
## <none>              75937505 2452.8
## ui      1   6353218 82290723 2466.0
## race    2   6630123 82567628 2464.6
## smoke   1   4950633 80888138 2462.7
## ht      1   3584838 79522343 2459.5
## lwt     1   2674229 78611734 2457.3
```

```
# This result supports our model.
```

```
# We also can use the backward selection using p-value
drop1(bwt_lm1,test="F")
```

```
## Single term deletions
##
## Model:
## ourdata$bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv
##        Df Sum of Sq      RSS    AIC F value    Pr(>F)
## <none>              75702317 2458.2
## age     1     58238 75760555 2456.3  0.1377 0.7110122
## lwt     1   2661604 78363921 2462.7  6.2934 0.0130073 *
## race    2   6578597 82280914 2470.0  7.7776 0.0005768 ***
## smoke   1   4623219 80325536 2467.4 10.9317 0.0011423 **
## ptl     1     95285 75797602 2456.4  0.2253 0.6356065
## ht      1   3631032 79333349 2465.1  8.5857 0.0038298 **
## ui      1   5839544 81541861 2470.2 13.8077 0.0002705 ***
## ftv     1     38708 75741025 2456.3  0.0915 0.7625980
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
M2=update(bwt_lm1,. ~ .- ftv)
drop1(M2,test = "F")
```

```
## Single term deletions
##
## Model:
## ourdata$bwt ~ age + lwt + race + smoke + ptl + ht + ui
##        Df Sum of Sq      RSS    AIC F value    Pr(>F)
## <none>              75741025 2456.3
## age     1     79115 75820139 2454.5  0.1880 0.6650913
## lwt     1   2623988 78365013 2460.7  6.2360 0.0134160 *
## race    2   6552496 82293521 2468.0  7.7861 0.0005713 ***
## smoke   1   4606425 80347449 2465.5 10.9473 0.0011321 **
## ptl     1     91560 75832585 2454.5  0.2176 0.6414430
## ht      1   3592430 79333455 2463.1  8.5375 0.0039251 **
## ui      1   5817995 81559020 2468.3 13.8266 0.0002676 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
M3=update(M2,.~.-age)
drop1(M3,test="F")
```

```
## Single term deletions
##
## Model:
## ourdata$bwt ~ lwt + race + smoke + ptl + ht + ui
##        Df Sum of Sq      RSS    AIC F value    Pr(>F)    
## <none>              75820139 2454.5                      
## lwt     1   2545892 78366031 2458.7  6.0776 0.0146227 *  
## race    2   6571668 82391807 2466.2  7.8440 0.0005408 ***
## smoke   1   4530009 80350149 2463.5 10.8142 0.0012103 ** 
## ptl     1    117366 75937505 2452.8  0.2802 0.5972329    
## ht      1   3546591 79366731 2461.1  8.4665 0.0040713 ** 
## ui      1   5751122 81571261 2466.3 13.7292 0.0002804 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
M4=update(M3,.~.-ptl)
drop1(M4,test = "F")
```

```
## Single term deletions
## 
## Model:
## ourdata$bwt ~ lwt + race + smoke + ht + ui
##        Df Sum of Sq      RSS    AIC F value    Pr(>F)    
## <none>              75937505 2452.8                      
## lwt     1   2674229 78611734 2457.3  6.4093 0.0121981 *  
## race    2   6630123 82567628 2464.6  7.9452 0.0004919 ***
## smoke   1   4950633 80888138 2462.7 11.8652 0.0007099 ***
## ht      1   3584838 79522343 2459.5  8.5918 0.0038100 ** 
## ui      1   6353218 82290723 2466.0 15.2268 0.0001341 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
M4
```

```
## 
## Call:
## lm(formula = ourdata$bwt ~ lwt + race + smoke + ht + ui, data = ourdata)
## 
## Coefficients:
## (Intercept)          lwt    raceblack    raceother    smokeTrue          ht1  
##    2837.264        4.242     -475.058     -348.150     -356.321     -585.193  
##        uiNo  
##    -525.524
```

```
summary(M4)
```

```
## 
## Call:
## lm(formula = ourdata$bwt ~ lwt + race + smoke + ht + ui, data = ourdata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1842.14  -433.19    67.09   459.21  1631.03 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2837.264    243.676  11.644  < 2e-16 ***
## lwt            4.242      1.675   2.532 0.012198 *  
## raceblack   -475.058    145.603  -3.263 0.001318 ** 
## raceother   -348.150    112.361  -3.099 0.002254 **
## smokeTrue   -356.321    103.444  -3.445 0.000710 ***
## ht1         -585.193    199.644  -2.931 0.003810 **
## uiNo        -525.524    134.675  -3.902 0.000134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.9 on 182 degrees of freedom
## Multiple R-squared:  0.2404, Adjusted R-squared:  0.2154
## F-statistic:   9.6 on 6 and 182 DF,  p-value: 3.601e-09
```

```
# To sum up, we uses three methods to select model, fortunately, we get the same result. Because of thi
```

Assumption checking

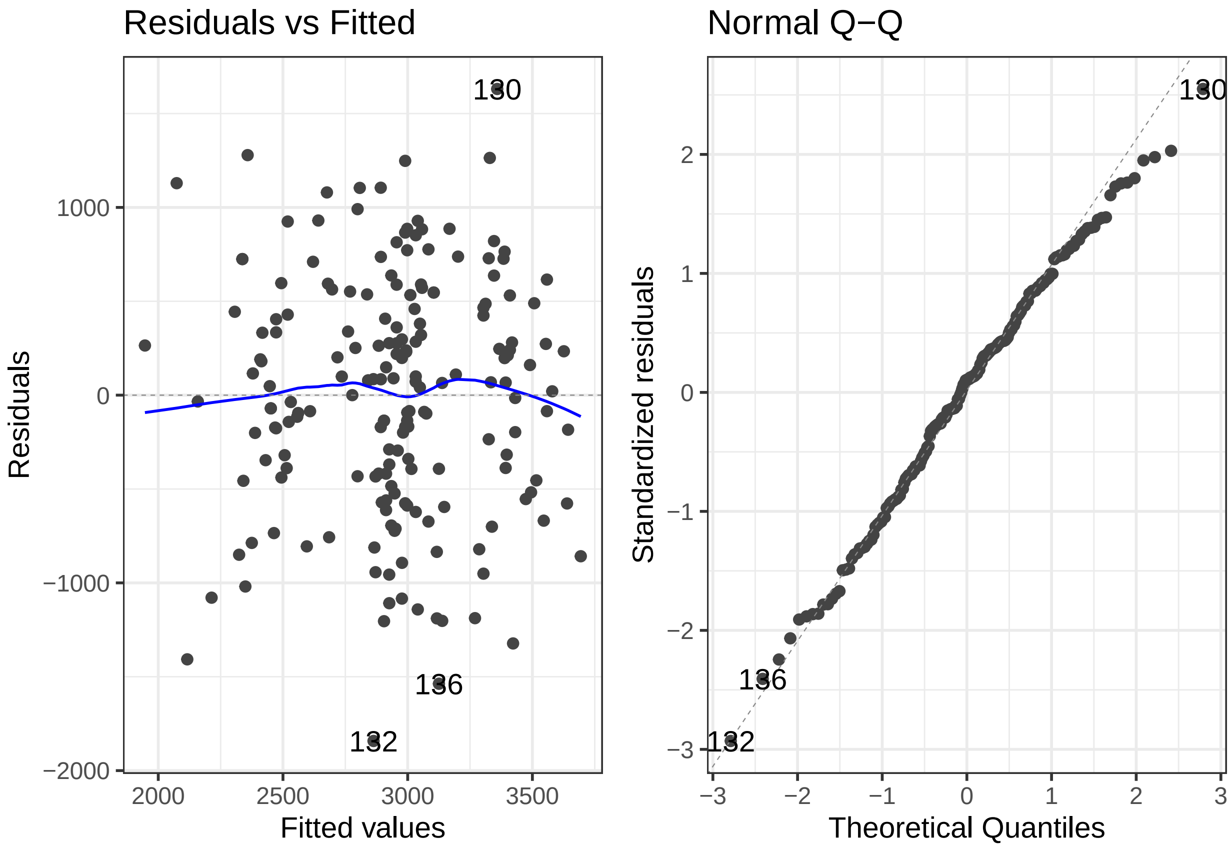
```
# 1:Linearity
## after runing the regression we check the fitted values vs residuals
library(ggfortify)
```

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

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
autoplot(M4, which=1:2)+theme_bw()
```



```
# By looking at the plot, there is no obvious pattern in the residual vs fitted values plot so it does
# Homoskedasticity: the residuals do not appear to be changing their variability over the range of the
# Normality: in the QQ plot, the points are reasonably close to the diagonal line. The top are not quit
```

```
library(caret)
```

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

```
## Loading required package: lattice
```

```
set.seed(2)
cv_full = train(data = ourdata, bwt ~., method = 'lm', trControl= trainControl(method='cv',number = 10,
cv_full
```

```
## Linear Regression
##
## 189 samples
##   8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 170, 170, 169, 169, 171, 170, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   659.2827  0.2437037  540.2351
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
cv_simplified = train(data = ourdata, bwt ~ lwt + race + smoke + ht + ui, method = 'lm', trControl= trai
cv_simplified
```

```
## Linear Regression
##
## 189 samples
##   5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 170, 170, 169, 169, 169, 170, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   657.2103  0.1975652  533.7115
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```