2nd Assignment in Medical Biostatics II

January 13, 2025

## Authors of this assignment

1. Michael Adam, 01507402
2. Aaron Paul Osburg, 12010313

## Exercise 1

### Task

Develop a strategy to improve the linear regression model for this data set. What could be done to achieve a model fit that has less issues with residuals? First, develop a plan with several analysis strategies and summarize these ideas in written form. Second, carry out the analyses and check the resulting model(s) again. Third, describe your new model(s) appropriately. Fourth, write a summary discussing successes/failures of your strategies.

### Solution

#### Modelling Strategy

The residuals are not normally distributed as the QQ-Plot of assignment 1 showed. This could be due to a wrong or incomplete choice of predictors or a non-normal distribution of the dependent variable. If the distribution of the dependent variable is not normal, one could try to apply a so-called Box-Cox transformation, a square root or a polynomial transformation. Otherwise, or if the problems remain after applying such a transformation, additional or different predictors (such as interaction and higher order terms, splines, or fractional polynomials) could be used instead.

Potentially better predictors could be found by using AIC-based model selection algorithms (starting from a model with higher-order terms such as a fully-quadratic model) or inspecting partial residual plots of the original predictors to deduce a suitable nonlinear transformation.

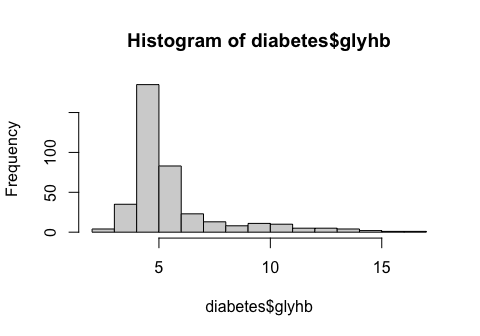
#### Analysis

As a first step, we consider the glyhb distribution:

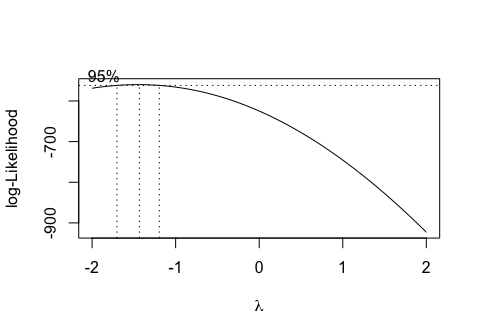
library(mfp)

## Loading required package: survival

library(splines)  
library(MASS)  
url\_diab <- "https://hbiostat.org/data/repo/diabetes.csv"  
diabetes <- read.csv(file=url\_diab)  
diabetes$BMI <- diabetes$weight / diabetes$height^2 \* 704.5   
  
hist(diabetes$glyhb)

 Clearly, the distribution of glyhb strongly deviates from a normal distribution and is highly asymmetric with a long tail on the right side. To symmetrize the distribution, we determine a Box-Cox transformation starting from the model given in assignment 1:

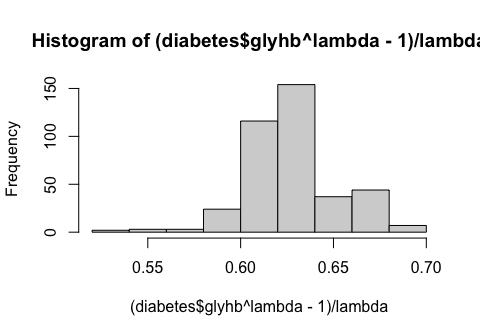
model <- lm(glyhb ~ chol + hdl + age + gender + waist + bp.1s + BMI, data=diabetes)  
b <- boxcox(model)



lambda <- b$x[which.max(b$y)]  
lambda

## [1] -1.434343

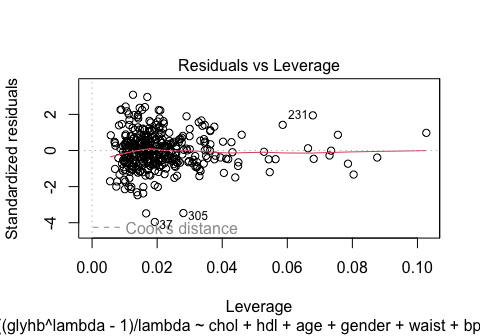
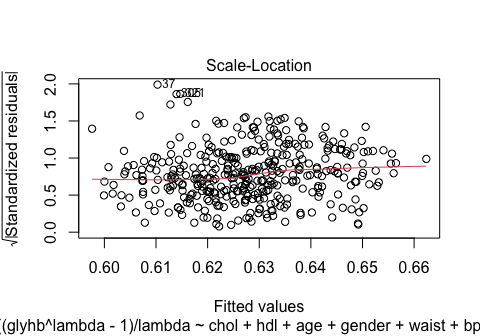
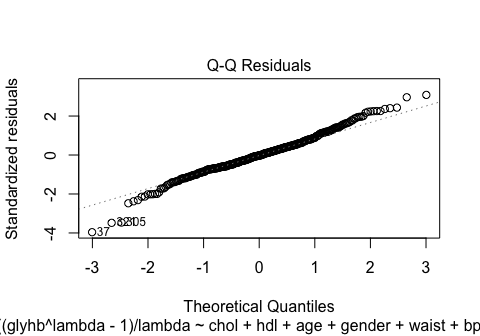
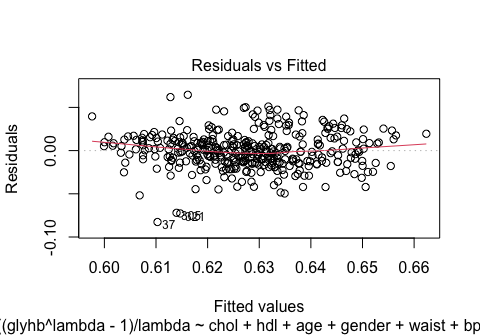
hist((diabetes$glyhb^lambda - 1)/lambda)

 The resulting distribution of the dependent variable looks more symmetric and Gaussian now, even though it still does not perfectly look like a Gaussian. To further evaluate the performance of the model, we consider the following plots:

model <- lm((glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist + bp.1s + BMI, data=diabetes)  
summary(model)

##   
## Call:  
## lm(formula = (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender +   
## waist + bp.1s + BMI, data = diabetes)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.082640 -0.012571 -0.000422 0.011270 0.064579   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.601e-01 1.064e-02 52.666 < 2e-16 \*\*\*  
## chol 9.433e-05 2.602e-05 3.626 0.000329 \*\*\*  
## hdl -2.148e-04 6.780e-05 -3.168 0.001664 \*\*   
## age 4.527e-04 7.657e-05 5.912 7.71e-09 \*\*\*  
## gendermale 2.605e-04 2.418e-03 0.108 0.914258   
## waist 1.102e-03 3.609e-04 3.053 0.002429 \*\*   
## bp.1s 2.690e-05 5.365e-05 0.501 0.616363   
## BMI -2.758e-04 3.171e-04 -0.870 0.385004   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02107 on 369 degrees of freedom  
## (26 observations deleted due to missingness)  
## Multiple R-squared: 0.2671, Adjusted R-squared: 0.2532   
## F-statistic: 19.22 on 7 and 369 DF, p-value: < 2.2e-16

plot(model)



The QQ-Plot shows that the residuals are approximately normally distributed now and deviations can only be seen in the tails. In the original models these deviations where much more pronounced. The scale-location plot shows that the tendency of the residuals to get larger for larger fitted values decreased compared to the original model from assignment 1.

To further improve the model, we apply an AIC-based model selection starting from a fully-quadratic model (except for the quadratic term of the binary predictor gender):

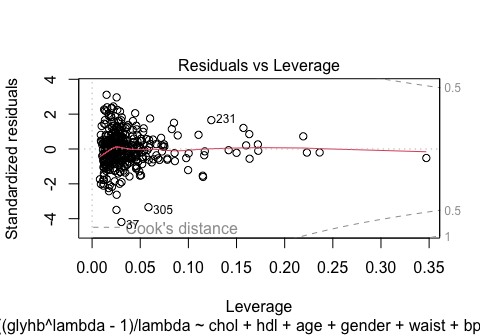
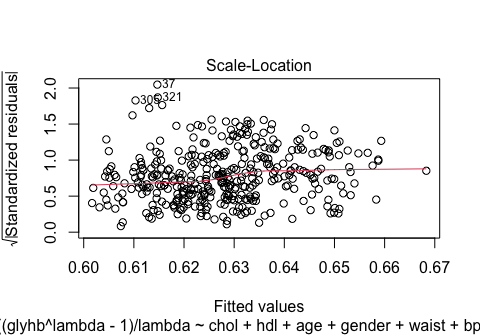
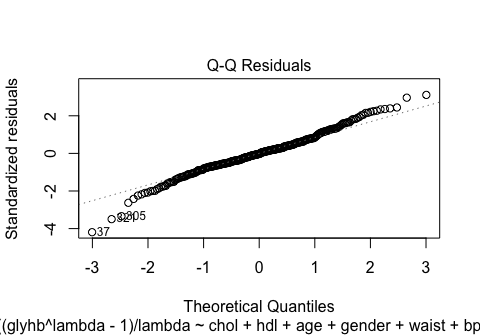
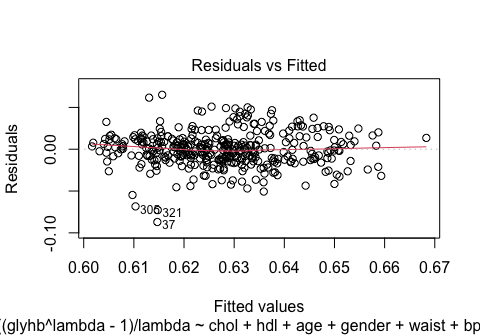
model <- lm((glyhb^lambda - 1)/lambda ~ (chol + hdl + age + gender + waist + bp.1s + BMI)^2 + I(chol^2) + I(hdl^2) + I(age^2) + I(waist^2) + I(bp.1s^2) + I(BMI^2),   
data=na.omit(diabetes))  
model <- step(model)

## Start: AIC=-1014.79  
## (glyhb^lambda - 1)/lambda ~ (chol + hdl + age + gender + waist +   
## bp.1s + BMI)^2 + I(chol^2) + I(hdl^2) + I(age^2) + I(waist^2) +   
## I(bp.1s^2) + I(BMI^2)  
##   
## Df Sum of Sq RSS AIC  
## - gender:bp.1s 1 0.00000052 0.046712 -1016.8  
## - chol:bp.1s 1 0.00000056 0.046712 -1016.8  
## - age:BMI 1 0.00000771 0.046719 -1016.8  
## - I(age^2) 1 0.00001842 0.046730 -1016.7  
## - chol:waist 1 0.00002042 0.046732 -1016.7  
## - age:waist 1 0.00002360 0.046735 -1016.7  
## - hdl:age 1 0.00006206 0.046774 -1016.6  
## - age:bp.1s 1 0.00006532 0.046777 -1016.6  
## - I(waist^2) 1 0.00009800 0.046809 -1016.5  
## - hdl:gender 1 0.00014383 0.046855 -1016.4  
## - hdl:bp.1s 1 0.00017015 0.046882 -1016.3  
## - chol:age 1 0.00018368 0.046895 -1016.3  
## - hdl:waist 1 0.00020131 0.046913 -1016.2  
## - I(bp.1s^2) 1 0.00021908 0.046931 -1016.2  
## - age:gender 1 0.00024529 0.046957 -1016.1  
## - bp.1s:BMI 1 0.00029002 0.047002 -1016.0  
## - I(chol^2) 1 0.00031957 0.047031 -1015.9  
## - waist:BMI 1 0.00038419 0.047096 -1015.7  
## - I(BMI^2) 1 0.00050210 0.047214 -1015.3  
## - chol:gender 1 0.00053913 0.047251 -1015.2  
## - hdl:BMI 1 0.00054134 0.047253 -1015.2  
## <none> 0.046711 -1014.8  
## - gender:waist 1 0.00071232 0.047424 -1014.7  
## - I(hdl^2) 1 0.00079304 0.047505 -1014.5  
## - gender:BMI 1 0.00098734 0.047699 -1014.0  
## - waist:bp.1s 1 0.00102749 0.047739 -1013.8  
## - chol:BMI 1 0.00171739 0.048429 -1011.9  
## - chol:hdl 1 0.00289454 0.049606 -1008.6  
##   
## Step: AIC=-1016.79  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(age^2) + I(waist^2) +   
## I(bp.1s^2) + I(BMI^2) + chol:hdl + chol:age + chol:gender +   
## chol:waist + chol:bp.1s + chol:BMI + hdl:age + hdl:gender +   
## hdl:waist + hdl:bp.1s + hdl:BMI + age:gender + age:waist +   
## age:bp.1s + age:BMI + gender:waist + gender:BMI + waist:bp.1s +   
## waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - chol:bp.1s 1 0.00000110 0.046713 -1018.8  
## - age:BMI 1 0.00001056 0.046723 -1018.8  
## - I(age^2) 1 0.00002027 0.046732 -1018.7  
## - chol:waist 1 0.00002199 0.046734 -1018.7  
## - age:waist 1 0.00002329 0.046735 -1018.7  
## - hdl:age 1 0.00006220 0.046774 -1018.6  
## - age:bp.1s 1 0.00006804 0.046780 -1018.6  
## - I(waist^2) 1 0.00009747 0.046809 -1018.5  
## - hdl:gender 1 0.00014393 0.046856 -1018.4  
## - hdl:bp.1s 1 0.00017571 0.046888 -1018.3  
## - chol:age 1 0.00018317 0.046895 -1018.3  
## - hdl:waist 1 0.00020080 0.046913 -1018.2  
## - I(bp.1s^2) 1 0.00026291 0.046975 -1018.0  
## - age:gender 1 0.00026982 0.046982 -1018.0  
## - bp.1s:BMI 1 0.00029805 0.047010 -1017.9  
## - I(chol^2) 1 0.00032229 0.047034 -1017.9  
## - waist:BMI 1 0.00038366 0.047096 -1017.7  
## - I(BMI^2) 1 0.00050396 0.047216 -1017.3  
## - hdl:BMI 1 0.00054758 0.047260 -1017.2  
## - chol:gender 1 0.00057208 0.047284 -1017.1  
## <none> 0.046712 -1016.8  
## - gender:waist 1 0.00071658 0.047429 -1016.7  
## - I(hdl^2) 1 0.00081791 0.047530 -1016.4  
## - gender:BMI 1 0.00098682 0.047699 -1016.0  
## - waist:bp.1s 1 0.00102804 0.047740 -1015.8  
## - chol:BMI 1 0.00184351 0.048556 -1013.5  
## - chol:hdl 1 0.00294314 0.049655 -1010.5  
##   
## Step: AIC=-1018.79  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(age^2) + I(waist^2) +   
## I(bp.1s^2) + I(BMI^2) + chol:hdl + chol:age + chol:gender +   
## chol:waist + chol:BMI + hdl:age + hdl:gender + hdl:waist +   
## hdl:bp.1s + hdl:BMI + age:gender + age:waist + age:bp.1s +   
## age:BMI + gender:waist + gender:BMI + waist:bp.1s + waist:BMI +   
## bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - age:BMI 1 0.00001232 0.046725 -1020.8  
## - chol:waist 1 0.00002109 0.046734 -1020.7  
## - I(age^2) 1 0.00002240 0.046736 -1020.7  
## - age:waist 1 0.00002252 0.046736 -1020.7  
## - hdl:age 1 0.00007981 0.046793 -1020.6  
## - age:bp.1s 1 0.00007989 0.046793 -1020.6  
## - I(waist^2) 1 0.00009643 0.046810 -1020.5  
## - hdl:gender 1 0.00014685 0.046860 -1020.4  
## - hdl:waist 1 0.00020022 0.046913 -1020.2  
## - chol:age 1 0.00021538 0.046928 -1020.2  
## - hdl:bp.1s 1 0.00024808 0.046961 -1020.1  
## - I(bp.1s^2) 1 0.00027107 0.046984 -1020.0  
## - age:gender 1 0.00028927 0.047002 -1020.0  
## - I(chol^2) 1 0.00032511 0.047038 -1019.9  
## - bp.1s:BMI 1 0.00032803 0.047041 -1019.8  
## - waist:BMI 1 0.00038313 0.047096 -1019.7  
## - I(BMI^2) 1 0.00050494 0.047218 -1019.3  
## - hdl:BMI 1 0.00056035 0.047273 -1019.2  
## - chol:gender 1 0.00057981 0.047293 -1019.1  
## <none> 0.046713 -1018.8  
## - gender:waist 1 0.00073793 0.047451 -1018.7  
## - I(hdl^2) 1 0.00081897 0.047532 -1018.4  
## - gender:BMI 1 0.00102927 0.047742 -1017.8  
## - waist:bp.1s 1 0.00103428 0.047747 -1017.8  
## - chol:BMI 1 0.00184245 0.048556 -1015.5  
## - chol:hdl 1 0.00308021 0.049793 -1012.1  
##   
## Step: AIC=-1020.75  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(age^2) + I(waist^2) +   
## I(bp.1s^2) + I(BMI^2) + chol:hdl + chol:age + chol:gender +   
## chol:waist + chol:BMI + hdl:age + hdl:gender + hdl:waist +   
## hdl:bp.1s + hdl:BMI + age:gender + age:waist + age:bp.1s +   
## gender:waist + gender:BMI + waist:bp.1s + waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - I(age^2) 1 0.00001548 0.046741 -1022.7  
## - chol:waist 1 0.00001575 0.046741 -1022.7  
## - hdl:age 1 0.00007223 0.046798 -1022.5  
## - age:bp.1s 1 0.00008370 0.046809 -1022.5  
## - I(waist^2) 1 0.00009614 0.046822 -1022.5  
## - age:waist 1 0.00012691 0.046852 -1022.4  
## - hdl:gender 1 0.00014507 0.046871 -1022.3  
## - hdl:waist 1 0.00019840 0.046924 -1022.2  
## - chol:age 1 0.00022408 0.046950 -1022.1  
## - hdl:bp.1s 1 0.00024280 0.046968 -1022.0  
## - I(bp.1s^2) 1 0.00025964 0.046985 -1022.0  
## - I(chol^2) 1 0.00031992 0.047045 -1021.8  
## - bp.1s:BMI 1 0.00033238 0.047058 -1021.8  
## - waist:BMI 1 0.00040035 0.047126 -1021.6  
## - age:gender 1 0.00046417 0.047190 -1021.4  
## - I(BMI^2) 1 0.00054077 0.047266 -1021.2  
## - hdl:BMI 1 0.00055781 0.047283 -1021.1  
## - chol:gender 1 0.00056885 0.047294 -1021.1  
## <none> 0.046725 -1020.8  
## - gender:waist 1 0.00072696 0.047452 -1020.6  
## - I(hdl^2) 1 0.00081867 0.047544 -1020.4  
## - gender:BMI 1 0.00104083 0.047766 -1019.8  
## - waist:bp.1s 1 0.00108805 0.047813 -1019.6  
## - chol:BMI 1 0.00189249 0.048618 -1017.4  
## - chol:hdl 1 0.00307211 0.049798 -1014.1  
##   
## Step: AIC=-1022.71  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(waist^2) + I(bp.1s^2) +   
## I(BMI^2) + chol:hdl + chol:age + chol:gender + chol:waist +   
## chol:BMI + hdl:age + hdl:gender + hdl:waist + hdl:bp.1s +   
## hdl:BMI + age:gender + age:waist + age:bp.1s + gender:waist +   
## gender:BMI + waist:bp.1s + waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - chol:waist 1 0.00001360 0.046755 -1024.7  
## - age:bp.1s 1 0.00006979 0.046811 -1024.5  
## - hdl:age 1 0.00007186 0.046813 -1024.5  
## - I(waist^2) 1 0.00008418 0.046825 -1024.5  
## - age:waist 1 0.00011383 0.046855 -1024.4  
## - hdl:gender 1 0.00014469 0.046886 -1024.3  
## - hdl:waist 1 0.00018685 0.046928 -1024.2  
## - hdl:bp.1s 1 0.00023212 0.046973 -1024.0  
## - I(bp.1s^2) 1 0.00026296 0.047004 -1023.9  
## - chol:age 1 0.00027150 0.047012 -1023.9  
## - I(chol^2) 1 0.00031163 0.047053 -1023.8  
## - bp.1s:BMI 1 0.00034661 0.047088 -1023.7  
## - waist:BMI 1 0.00038940 0.047130 -1023.6  
## - age:gender 1 0.00046127 0.047202 -1023.4  
## - hdl:BMI 1 0.00054780 0.047289 -1023.1  
## - chol:gender 1 0.00055809 0.047299 -1023.1  
## - I(BMI^2) 1 0.00055819 0.047299 -1023.1  
## <none> 0.046741 -1022.7  
## - gender:waist 1 0.00072719 0.047468 -1022.6  
## - I(hdl^2) 1 0.00081063 0.047552 -1022.4  
## - gender:BMI 1 0.00103385 0.047775 -1021.7  
## - waist:bp.1s 1 0.00112016 0.047861 -1021.5  
## - chol:BMI 1 0.00189850 0.048639 -1019.3  
## - chol:hdl 1 0.00306718 0.049808 -1016.1  
##   
## Step: AIC=-1024.67  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(waist^2) + I(bp.1s^2) +   
## I(BMI^2) + chol:hdl + chol:age + chol:gender + chol:BMI +   
## hdl:age + hdl:gender + hdl:waist + hdl:bp.1s + hdl:BMI +   
## age:gender + age:waist + age:bp.1s + gender:waist + gender:BMI +   
## waist:bp.1s + waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - hdl:age 1 0.0000651 0.046820 -1026.5  
## - age:bp.1s 1 0.0000661 0.046821 -1026.5  
## - I(waist^2) 1 0.0000778 0.046832 -1026.4  
## - age:waist 1 0.0001288 0.046883 -1026.3  
## - hdl:gender 1 0.0001360 0.046891 -1026.3  
## - hdl:waist 1 0.0001796 0.046934 -1026.2  
## - hdl:bp.1s 1 0.0002370 0.046991 -1026.0  
## - I(bp.1s^2) 1 0.0002690 0.047024 -1025.9  
## - I(chol^2) 1 0.0002980 0.047053 -1025.8  
## - bp.1s:BMI 1 0.0003541 0.047109 -1025.6  
## - chol:age 1 0.0003736 0.047128 -1025.6  
## - waist:BMI 1 0.0003758 0.047130 -1025.6  
## - age:gender 1 0.0005181 0.047273 -1025.2  
## - I(BMI^2) 1 0.0005447 0.047299 -1025.1  
## - chol:gender 1 0.0005515 0.047306 -1025.1  
## - hdl:BMI 1 0.0005590 0.047314 -1025.0  
## <none> 0.046755 -1024.7  
## - gender:waist 1 0.0007302 0.047485 -1024.6  
## - I(hdl^2) 1 0.0008291 0.047584 -1024.3  
## - gender:BMI 1 0.0010263 0.047781 -1023.7  
## - waist:bp.1s 1 0.0011178 0.047872 -1023.5  
## - chol:hdl 1 0.0033872 0.050142 -1017.2  
## - chol:BMI 1 0.0034064 0.050161 -1017.1  
##   
## Step: AIC=-1026.48  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(waist^2) + I(bp.1s^2) +   
## I(BMI^2) + chol:hdl + chol:age + chol:gender + chol:BMI +   
## hdl:gender + hdl:waist + hdl:bp.1s + hdl:BMI + age:gender +   
## age:waist + age:bp.1s + gender:waist + gender:BMI + waist:bp.1s +   
## waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - I(waist^2) 1 0.0000704 0.046890 -1028.3  
## - age:bp.1s 1 0.0000718 0.046891 -1028.3  
## - age:waist 1 0.0000944 0.046914 -1028.2  
## - hdl:gender 1 0.0001305 0.046950 -1028.1  
## - hdl:waist 1 0.0001366 0.046956 -1028.1  
## - hdl:bp.1s 1 0.0001723 0.046992 -1028.0  
## - I(bp.1s^2) 1 0.0002424 0.047062 -1027.8  
## - I(chol^2) 1 0.0002905 0.047110 -1027.6  
## - bp.1s:BMI 1 0.0003349 0.047155 -1027.5  
## - waist:BMI 1 0.0003571 0.047177 -1027.5  
## - chol:age 1 0.0004675 0.047287 -1027.1  
## - hdl:BMI 1 0.0004982 0.047318 -1027.0  
## - I(BMI^2) 1 0.0005443 0.047364 -1026.9  
## - chol:gender 1 0.0005516 0.047371 -1026.9  
## - age:gender 1 0.0006592 0.047479 -1026.6  
## <none> 0.046820 -1026.5  
## - gender:waist 1 0.0007027 0.047522 -1026.5  
## - gender:BMI 1 0.0009925 0.047812 -1025.6  
## - I(hdl^2) 1 0.0010071 0.047827 -1025.6  
## - waist:bp.1s 1 0.0011209 0.047941 -1025.3  
## - chol:BMI 1 0.0033629 0.050183 -1019.0  
## - chol:hdl 1 0.0034463 0.050266 -1018.8  
##   
## Step: AIC=-1028.27  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(bp.1s^2) + I(BMI^2) +   
## chol:hdl + chol:age + chol:gender + chol:BMI + hdl:gender +   
## hdl:waist + hdl:bp.1s + hdl:BMI + age:gender + age:waist +   
## age:bp.1s + gender:waist + gender:BMI + waist:bp.1s + waist:BMI +   
## bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - age:waist 1 0.0000679 0.046958 -1030.1  
## - age:bp.1s 1 0.0000977 0.046988 -1030.0  
## - hdl:waist 1 0.0000989 0.046989 -1030.0  
## - hdl:bp.1s 1 0.0001403 0.047030 -1029.9  
## - hdl:gender 1 0.0001406 0.047031 -1029.9  
## - I(bp.1s^2) 1 0.0001925 0.047083 -1029.7  
## - I(chol^2) 1 0.0002707 0.047161 -1029.5  
## - waist:BMI 1 0.0002919 0.047182 -1029.4  
## - bp.1s:BMI 1 0.0003484 0.047238 -1029.3  
## - hdl:BMI 1 0.0004551 0.047345 -1029.0  
## - chol:age 1 0.0004969 0.047387 -1028.8  
## - chol:gender 1 0.0004982 0.047388 -1028.8  
## - I(BMI^2) 1 0.0005543 0.047444 -1028.7  
## - gender:waist 1 0.0006455 0.047536 -1028.4  
## - age:gender 1 0.0006814 0.047572 -1028.3  
## <none> 0.046890 -1028.3  
## - gender:BMI 1 0.0009220 0.047812 -1027.6  
## - I(hdl^2) 1 0.0009789 0.047869 -1027.5  
## - waist:bp.1s 1 0.0011570 0.048047 -1027.0  
## - chol:BMI 1 0.0032926 0.050183 -1021.0  
## - chol:hdl 1 0.0034319 0.050322 -1020.7  
##   
## Step: AIC=-1030.08  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(bp.1s^2) + I(BMI^2) +   
## chol:hdl + chol:age + chol:gender + chol:BMI + hdl:gender +   
## hdl:waist + hdl:bp.1s + hdl:BMI + age:gender + age:bp.1s +   
## gender:waist + gender:BMI + waist:bp.1s + waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - hdl:waist 1 0.0000874 0.047045 -1031.8  
## - age:bp.1s 1 0.0000894 0.047047 -1031.8  
## - hdl:gender 1 0.0001201 0.047078 -1031.7  
## - hdl:bp.1s 1 0.0001556 0.047114 -1031.6  
## - I(bp.1s^2) 1 0.0001844 0.047142 -1031.5  
## - I(chol^2) 1 0.0002490 0.047207 -1031.4  
## - waist:BMI 1 0.0002706 0.047229 -1031.3  
## - bp.1s:BMI 1 0.0004050 0.047363 -1030.9  
## - hdl:BMI 1 0.0004390 0.047397 -1030.8  
## - chol:age 1 0.0004754 0.047433 -1030.7  
## - chol:gender 1 0.0004870 0.047445 -1030.7  
## - I(BMI^2) 1 0.0005487 0.047507 -1030.5  
## - age:gender 1 0.0006341 0.047592 -1030.2  
## - gender:waist 1 0.0006450 0.047603 -1030.2  
## <none> 0.046958 -1030.1  
## - gender:BMI 1 0.0008747 0.047833 -1029.6  
## - I(hdl^2) 1 0.0009433 0.047901 -1029.4  
## - waist:bp.1s 1 0.0015176 0.048476 -1027.8  
## - chol:BMI 1 0.0032492 0.050207 -1023.0  
## - chol:hdl 1 0.0033677 0.050326 -1022.7  
##   
## Step: AIC=-1031.82  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(bp.1s^2) + I(BMI^2) +   
## chol:hdl + chol:age + chol:gender + chol:BMI + hdl:gender +   
## hdl:bp.1s + hdl:BMI + age:gender + age:bp.1s + gender:waist +   
## gender:BMI + waist:bp.1s + waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - age:bp.1s 1 0.0000501 0.047095 -1033.7  
## - hdl:gender 1 0.0000781 0.047123 -1033.6  
## - hdl:bp.1s 1 0.0001738 0.047219 -1033.3  
## - I(bp.1s^2) 1 0.0002040 0.047249 -1033.2  
## - I(chol^2) 1 0.0002283 0.047274 -1033.2  
## - chol:age 1 0.0004168 0.047462 -1032.6  
## - hdl:BMI 1 0.0004664 0.047512 -1032.5  
## - chol:gender 1 0.0004821 0.047527 -1032.4  
## - waist:BMI 1 0.0005085 0.047554 -1032.4  
## - gender:waist 1 0.0005714 0.047617 -1032.2  
## - bp.1s:BMI 1 0.0006257 0.047671 -1032.0  
## - age:gender 1 0.0006589 0.047704 -1031.9  
## <none> 0.047045 -1031.8  
## - gender:BMI 1 0.0008078 0.047853 -1031.5  
## - I(BMI^2) 1 0.0008467 0.047892 -1031.4  
## - I(hdl^2) 1 0.0008560 0.047901 -1031.4  
## - waist:bp.1s 1 0.0018524 0.048898 -1028.6  
## - chol:BMI 1 0.0033087 0.050354 -1024.6  
## - chol:hdl 1 0.0034155 0.050461 -1024.3  
##   
## Step: AIC=-1033.68  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(bp.1s^2) + I(BMI^2) +   
## chol:hdl + chol:age + chol:gender + chol:BMI + hdl:gender +   
## hdl:bp.1s + hdl:BMI + age:gender + gender:waist + gender:BMI +   
## waist:bp.1s + waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - hdl:gender 1 0.0000687 0.047164 -1035.5  
## - I(chol^2) 1 0.0002317 0.047327 -1035.0  
## - chol:age 1 0.0003695 0.047465 -1034.6  
## - hdl:bp.1s 1 0.0003714 0.047467 -1034.6  
## - I(bp.1s^2) 1 0.0004116 0.047507 -1034.5  
## - hdl:BMI 1 0.0004603 0.047556 -1034.4  
## - chol:gender 1 0.0005212 0.047617 -1034.2  
## - waist:BMI 1 0.0005244 0.047620 -1034.2  
## - gender:waist 1 0.0005463 0.047642 -1034.1  
## - bp.1s:BMI 1 0.0006618 0.047757 -1033.8  
## <none> 0.047095 -1033.7  
## - gender:BMI 1 0.0007583 0.047854 -1033.5  
## - I(hdl^2) 1 0.0008166 0.047912 -1033.3  
## - age:gender 1 0.0008288 0.047924 -1033.3  
## - I(BMI^2) 1 0.0008751 0.047971 -1033.2  
## - waist:bp.1s 1 0.0018028 0.048898 -1030.6  
## - chol:hdl 1 0.0034103 0.050506 -1026.2  
## - chol:BMI 1 0.0034184 0.050514 -1026.2  
##   
## Step: AIC=-1035.48  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(chol^2) + I(hdl^2) + I(bp.1s^2) + I(BMI^2) +   
## chol:hdl + chol:age + chol:gender + chol:BMI + hdl:bp.1s +   
## hdl:BMI + age:gender + gender:waist + gender:BMI + waist:bp.1s +   
## waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - I(chol^2) 1 0.0002065 0.047371 -1036.9  
## - chol:age 1 0.0003223 0.047486 -1036.6  
## - hdl:bp.1s 1 0.0003711 0.047535 -1036.4  
## - hdl:BMI 1 0.0003916 0.047556 -1036.4  
## - I(bp.1s^2) 1 0.0004065 0.047571 -1036.3  
## - waist:BMI 1 0.0004645 0.047629 -1036.2  
## - chol:gender 1 0.0004731 0.047637 -1036.1  
## - gender:waist 1 0.0006077 0.047772 -1035.7  
## - bp.1s:BMI 1 0.0006788 0.047843 -1035.5  
## <none> 0.047164 -1035.5  
## - gender:BMI 1 0.0007344 0.047899 -1035.4  
## - I(hdl^2) 1 0.0007773 0.047941 -1035.3  
## - age:gender 1 0.0007829 0.047947 -1035.2  
## - I(BMI^2) 1 0.0008151 0.047979 -1035.2  
## - waist:bp.1s 1 0.0017572 0.048921 -1032.5  
## - chol:hdl 1 0.0033738 0.050538 -1028.1  
## - chol:BMI 1 0.0033857 0.050550 -1028.0  
##   
## Step: AIC=-1036.89  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(hdl^2) + I(bp.1s^2) + I(BMI^2) + chol:hdl +   
## chol:age + chol:gender + chol:BMI + hdl:bp.1s + hdl:BMI +   
## age:gender + gender:waist + gender:BMI + waist:bp.1s + waist:BMI +   
## bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - chol:age 1 0.0002505 0.047621 -1038.2  
## - chol:gender 1 0.0003111 0.047682 -1038.0  
## - hdl:bp.1s 1 0.0003232 0.047694 -1038.0  
## - hdl:BMI 1 0.0003407 0.047711 -1037.9  
## - I(bp.1s^2) 1 0.0003556 0.047726 -1037.9  
## - waist:BMI 1 0.0005283 0.047899 -1037.4  
## - I(hdl^2) 1 0.0005919 0.047962 -1037.2  
## - gender:waist 1 0.0006388 0.048009 -1037.1  
## - bp.1s:BMI 1 0.0006735 0.048044 -1037.0  
## <none> 0.047371 -1036.9  
## - age:gender 1 0.0007711 0.048142 -1036.7  
## - gender:BMI 1 0.0007812 0.048152 -1036.7  
## - I(BMI^2) 1 0.0008815 0.048252 -1036.4  
## - waist:bp.1s 1 0.0018330 0.049204 -1033.7  
## - chol:BMI 1 0.0031802 0.050551 -1030.0  
## - chol:hdl 1 0.0033774 0.050748 -1029.5  
##   
## Step: AIC=-1038.17  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(hdl^2) + I(bp.1s^2) + I(BMI^2) + chol:hdl +   
## chol:gender + chol:BMI + hdl:bp.1s + hdl:BMI + age:gender +   
## gender:waist + gender:BMI + waist:bp.1s + waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - hdl:bp.1s 1 0.0001952 0.047816 -1039.6  
## - chol:gender 1 0.0003530 0.047974 -1039.2  
## - I(bp.1s^2) 1 0.0003688 0.047990 -1039.1  
## - hdl:BMI 1 0.0004531 0.048074 -1038.9  
## - waist:BMI 1 0.0005480 0.048169 -1038.6  
## - bp.1s:BMI 1 0.0006912 0.048312 -1038.2  
## <none> 0.047621 -1038.2  
## - gender:waist 1 0.0007109 0.048332 -1038.2  
## - age:gender 1 0.0007135 0.048335 -1038.2  
## - I(hdl^2) 1 0.0007163 0.048337 -1038.1  
## - gender:BMI 1 0.0007439 0.048365 -1038.1  
## - I(BMI^2) 1 0.0009423 0.048563 -1037.5  
## - waist:bp.1s 1 0.0018952 0.049516 -1034.9  
## - chol:hdl 1 0.0033583 0.050979 -1030.9  
## - chol:BMI 1 0.0035542 0.051175 -1030.4  
##   
## Step: AIC=-1039.61  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(hdl^2) + I(bp.1s^2) + I(BMI^2) + chol:hdl +   
## chol:gender + chol:BMI + hdl:BMI + age:gender + gender:waist +   
## gender:BMI + waist:bp.1s + waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - chol:gender 1 0.0003066 0.048123 -1040.7  
## - waist:BMI 1 0.0004702 0.048286 -1040.3  
## - I(bp.1s^2) 1 0.0005423 0.048359 -1040.1  
## - age:gender 1 0.0005762 0.048393 -1040.0  
## - hdl:BMI 1 0.0006068 0.048423 -1039.9  
## - I(hdl^2) 1 0.0006599 0.048476 -1039.8  
## - bp.1s:BMI 1 0.0006773 0.048494 -1039.7  
## <none> 0.047816 -1039.6  
## - gender:waist 1 0.0007646 0.048581 -1039.5  
## - I(BMI^2) 1 0.0008559 0.048672 -1039.2  
## - gender:BMI 1 0.0008766 0.048693 -1039.1  
## - waist:bp.1s 1 0.0022798 0.050096 -1035.3  
## - chol:hdl 1 0.0033415 0.051158 -1032.4  
## - chol:BMI 1 0.0033731 0.051189 -1032.3  
##   
## Step: AIC=-1040.74  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(hdl^2) + I(bp.1s^2) + I(BMI^2) + chol:hdl +   
## chol:BMI + hdl:BMI + age:gender + gender:waist + gender:BMI +   
## waist:bp.1s + waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - hdl:BMI 1 0.0004344 0.048557 -1041.5  
## - I(bp.1s^2) 1 0.0004386 0.048562 -1041.5  
## - waist:BMI 1 0.0005488 0.048672 -1041.2  
## - age:gender 1 0.0006033 0.048726 -1041.0  
## - gender:waist 1 0.0006662 0.048789 -1040.9  
## <none> 0.048123 -1040.7  
## - I(hdl^2) 1 0.0007262 0.048849 -1040.7  
## - bp.1s:BMI 1 0.0007351 0.048858 -1040.7  
## - gender:BMI 1 0.0007522 0.048875 -1040.6  
## - I(BMI^2) 1 0.0010658 0.049189 -1039.8  
## - waist:bp.1s 1 0.0024069 0.050530 -1036.1  
## - chol:hdl 1 0.0030363 0.051159 -1034.4  
## - chol:BMI 1 0.0035060 0.051629 -1033.2  
##   
## Step: AIC=-1041.52  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(hdl^2) + I(bp.1s^2) + I(BMI^2) + chol:hdl +   
## chol:BMI + age:gender + gender:waist + gender:BMI + waist:bp.1s +   
## waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - I(bp.1s^2) 1 0.0004170 0.048974 -1042.4  
## - I(hdl^2) 1 0.0004381 0.048995 -1042.3  
## - waist:BMI 1 0.0004599 0.049017 -1042.2  
## - age:gender 1 0.0006157 0.049173 -1041.8  
## - bp.1s:BMI 1 0.0007014 0.049259 -1041.6  
## <none> 0.048557 -1041.5  
## - gender:BMI 1 0.0007990 0.049356 -1041.3  
## - gender:waist 1 0.0008290 0.049386 -1041.2  
## - I(BMI^2) 1 0.0013121 0.049869 -1039.9  
## - waist:bp.1s 1 0.0024807 0.051038 -1036.8  
## - chol:hdl 1 0.0028001 0.051357 -1035.9  
## - chol:BMI 1 0.0032317 0.051789 -1034.8  
##   
## Step: AIC=-1042.36  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(hdl^2) + I(BMI^2) + chol:hdl + chol:BMI +   
## age:gender + gender:waist + gender:BMI + waist:bp.1s + waist:BMI +   
## bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - I(hdl^2) 1 0.0004579 0.049432 -1043.1  
## - waist:BMI 1 0.0005624 0.049537 -1042.8  
## - gender:BMI 1 0.0005979 0.049572 -1042.7  
## - gender:waist 1 0.0007184 0.049693 -1042.4  
## <none> 0.048974 -1042.4  
## - bp.1s:BMI 1 0.0007287 0.049703 -1042.3  
## - age:gender 1 0.0007392 0.049713 -1042.3  
## - I(BMI^2) 1 0.0016191 0.050593 -1039.9  
## - waist:bp.1s 1 0.0024044 0.051379 -1037.8  
## - chol:hdl 1 0.0034104 0.052385 -1035.2  
## - chol:BMI 1 0.0034954 0.052470 -1035.0  
##   
## Step: AIC=-1043.09  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(BMI^2) + chol:hdl + chol:BMI + age:gender +   
## gender:waist + gender:BMI + waist:bp.1s + waist:BMI + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - waist:BMI 1 0.0004781 0.049910 -1043.8  
## - gender:BMI 1 0.0006474 0.050080 -1043.3  
## - age:gender 1 0.0006563 0.050089 -1043.3  
## - bp.1s:BMI 1 0.0006947 0.050127 -1043.2  
## - gender:waist 1 0.0007110 0.050143 -1043.2  
## <none> 0.049432 -1043.1  
## - I(BMI^2) 1 0.0014049 0.050837 -1041.3  
## - waist:bp.1s 1 0.0021366 0.051569 -1039.3  
## - chol:hdl 1 0.0030714 0.052504 -1036.9  
## - chol:BMI 1 0.0033469 0.052779 -1036.2  
##   
## Step: AIC=-1043.79  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(BMI^2) + chol:hdl + chol:BMI + age:gender +   
## gender:waist + gender:BMI + waist:bp.1s + bp.1s:BMI  
##   
## Df Sum of Sq RSS AIC  
## - bp.1s:BMI 1 0.00045175 0.050362 -1044.6  
## - age:gender 1 0.00061808 0.050528 -1044.1  
## <none> 0.049910 -1043.8  
## - gender:waist 1 0.00082846 0.050739 -1043.5  
## - gender:BMI 1 0.00098197 0.050892 -1043.1  
## - I(BMI^2) 1 0.00129484 0.051205 -1042.3  
## - waist:bp.1s 1 0.00181061 0.051721 -1040.9  
## - chol:BMI 1 0.00294400 0.052854 -1038.0  
## - chol:hdl 1 0.00310177 0.053012 -1037.6  
##   
## Step: AIC=-1044.56  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(BMI^2) + chol:hdl + chol:BMI + age:gender +   
## gender:waist + gender:BMI + waist:bp.1s  
##   
## Df Sum of Sq RSS AIC  
## - gender:waist 1 0.0006597 0.051022 -1044.8  
## - age:gender 1 0.0006945 0.051057 -1044.7  
## <none> 0.050362 -1044.6  
## - gender:BMI 1 0.0008558 0.051218 -1044.3  
## - I(BMI^2) 1 0.0013103 0.051672 -1043.1  
## - waist:bp.1s 1 0.0017553 0.052117 -1041.9  
## - chol:hdl 1 0.0033096 0.053672 -1037.9  
## - chol:BMI 1 0.0034888 0.053851 -1037.5  
##   
## Step: AIC=-1044.79  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(BMI^2) + chol:hdl + chol:BMI + age:gender +   
## gender:BMI + waist:bp.1s  
##   
## Df Sum of Sq RSS AIC  
## - gender:BMI 1 0.00021392 0.051236 -1046.2  
## <none> 0.051022 -1044.8  
## - age:gender 1 0.00086897 0.051891 -1044.5  
## - waist:bp.1s 1 0.00138472 0.052406 -1043.2  
## - I(BMI^2) 1 0.00142803 0.052450 -1043.0  
## - chol:BMI 1 0.00289993 0.053922 -1039.3  
## - chol:hdl 1 0.00299175 0.054013 -1039.0  
##   
## Step: AIC=-1046.22  
## (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist +   
## bp.1s + BMI + I(BMI^2) + chol:hdl + chol:BMI + age:gender +   
## waist:bp.1s  
##   
## Df Sum of Sq RSS AIC  
## <none> 0.051236 -1046.2  
## - age:gender 1 0.0010868 0.052322 -1045.4  
## - waist:bp.1s 1 0.0013359 0.052571 -1044.7  
## - I(BMI^2) 1 0.0025388 0.053774 -1041.6  
## - chol:hdl 1 0.0029024 0.054138 -1040.7  
## - chol:BMI 1 0.0029107 0.054146 -1040.7

#model with the terms resulting from the reduction  
model <- lm((glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender + waist + bp.1s + BMI + I(BMI^2) + chol:hdl + chol:BMI + age:bp.1s + age:gender + waist:bp.1s, data=diabetes)  
  
summary(model)

##   
## Call:  
## lm(formula = (glyhb^lambda - 1)/lambda ~ chol + hdl + age + gender +   
## waist + bp.1s + BMI + I(BMI^2) + chol:hdl + chol:BMI + age:bp.1s +   
## age:gender + waist:bp.1s, data = diabetes)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.087031 -0.011684 -0.000515 0.011699 0.065092   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.023e-01 5.491e-02 10.968 <2e-16 \*\*\*  
## chol 1.598e-04 1.422e-04 1.124 0.2619   
## hdl 8.008e-05 2.867e-04 0.279 0.7802   
## age 6.736e-04 4.144e-04 1.625 0.1050   
## gendermale 4.198e-03 7.004e-03 0.599 0.5493   
## waist -1.067e-03 1.305e-03 -0.818 0.4141   
## bp.1s -5.089e-04 3.620e-04 -1.406 0.1607   
## BMI 7.924e-04 1.465e-03 0.541 0.5888   
## I(BMI^2) -1.428e-05 1.763e-05 -0.810 0.4184   
## chol:hdl -1.229e-06 1.242e-06 -0.990 0.3229   
## chol:BMI -4.663e-07 4.144e-06 -0.113 0.9105   
## age:bp.1s -1.230e-06 2.918e-06 -0.422 0.6736   
## age:gendermale -7.240e-05 1.422e-04 -0.509 0.6111   
## waist:bp.1s 1.535e-05 8.823e-06 1.740 0.0827 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02109 on 363 degrees of freedom  
## (26 observations deleted due to missingness)  
## Multiple R-squared: 0.2777, Adjusted R-squared: 0.2518   
## F-statistic: 10.74 on 13 and 363 DF, p-value: < 2.2e-16

plot(model)

 The resulting model has the predictors

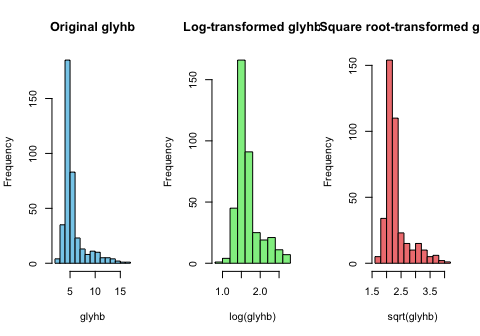
(Intercept)  
chol  
hdl  
age  
gendermale  
waist  
bp.1s  
BMI I(BMI^2) chol:hdl  
chol:BMI  
age:bp.1s  
age:gendermale waist:bp.1s

With the new model, we achieve residuals which seem to be independent of the fitted value (Residuals vs. Fitted plot) and the scale-location plot shows that the tendency of the residuals to get larger for larger fitted values decreased compared to the original model from assignment 1. Unfortunately, some of the data (the ones including NAs) had to be omitted for the model reduction. Alternatively to the Box-Cox transformation, we used sqrt- and log-transformations, but the resulting histogramms are more asymmetric:

diabetes$glyhb\_log <- log(diabetes$glyhb)

diabetes$glyhb\_sqrt <- sqrt(diabetes$glyhb)

par(mfrow = c(1, 3)) # Arrange plots in a row  
hist(diabetes$glyhb, main = "Original glyhb", xlab = "glyhb", col = "skyblue")  
hist(diabetes$glyhb\_log, main = "Log-transformed glyhb", xlab = "log(glyhb)", col = "lightgreen")  
hist(diabetes$glyhb\_sqrt, main = "Square root-transformed glyhb", xlab = "sqrt(glyhb)", col = "lightcoral")



## Exercise 2

### Task

With the diabetes data set, perform a different analysis strategy that is based on categorizing glyhb at 2 different values, and then performing two logistic regression models with the dichotomized glyhb values.

First, discuss with your partners what could be suitable values at which glyhb could be categorized for analysis. In your considerations, take into account statistical properties of the cutpoints (enough data in all categories) and clinical relevance of the cutpoints! Let’s assume your chosen cutpoints are g1 and g2.

Second, create two new binary outcome variables glyhb1 and glyhb2, where glyhb1=1 if glyhb>g1 and 0 else, or NA if glyhb is missing, and glyhb2=1 if glyhb>g2 and 0 else, or NA if glyhb is missing. Evaluate which absolute number of persons and proportions have glyhb1=1 and glyhb2=1 and glyhb1=0 and glyhb2=0 by cross-tabulating glyhb1 and glyhb2.

Third, fit two (separate) logistic regression models with glyhb1 and glyhb2 as outcome variables, using the same predictors as in Assignment 01-02. Check for nonlinear functional forms of continuous predictors. Describe the estimated associations of the predictors with the outcome variable in the model and its performance appropriately.

Report the methods and the results as you would do for a medical paper.

Compare with your linear regression model of Exercise 1 of this assignment. How do the estimated associations of the predictors with the outcome variables compare? Are the results consistent to each other? What else can be learned here?

### Methods

Commonly used threshold values for the glyhb value for healthy subjects are glyhb < 5.7%. Values between 5.7% and 6.4% are correspond to prediabetes, and values above 6.5% are are assumed to indicate diabetes (see <https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/glycated-hemoglobin>). Thus, we use the cutpoints g1=5.7% and g2=6.5%. To check the validity of this choice with respect to the statistical properties in our dataset, we check the absolute quantities of each category:

g1 <- 5.7  
g2 <- 6.5  
  
diabetes$glyhb1 <- FALSE  
diabetes$glyhb1[diabetes$glyhb > g1] <- TRUE  
diabetes$glyhb1[is.na(diabetes$glyhb)] <- NA  
  
diabetes$glyhb2 <- FALSE  
diabetes$glyhb2[diabetes$glyhb > g2] <- TRUE  
diabetes$glyhb2[is.na(diabetes$glyhb)] <- NA  
  
  
table(diabetes$glyhb1)

##   
## FALSE TRUE   
## 301 89

table(diabetes$glyhb2)

##   
## FALSE TRUE   
## 325 65

Indeed, there are enough data points in all categories. The cross-table for glyhb1 and glyhb2,

table(glyhb1=diabetes$glyhb1, glyhb2=diabetes$glyhb2)

## glyhb2  
## glyhb1 FALSE TRUE  
## FALSE 301 0  
## TRUE 24 65

confirms the expected behavior that there are no instances with glyhb2==TRUE and glyhb1==FALSE, because g2>g1. We fit two logistic regression models to the dichotomized glhyb dataset, using the same predictors as in assignment 1.

model1 <- glm(glyhb1 ~ chol + hdl + age + gender + waist + bp.1s + BMI,family=binomial(link='logit'),data=diabetes)

model2 <- glm(glyhb2 ~ chol + hdl + age + gender + waist + bp.1s + BMI,family=binomial(link='logit'),data=diabetes)

#### Results

The model coefficients are given by:

summary(model1)

##   
## Call:  
## glm(formula = glyhb1 ~ chol + hdl + age + gender + waist + bp.1s +   
## BMI, family = binomial(link = "logit"), data = diabetes)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.360866 1.478915 -5.653 1.57e-08 \*\*\*  
## chol 0.007971 0.003216 2.478 0.01321 \*   
## hdl -0.023777 0.009016 -2.637 0.00836 \*\*   
## age 0.050075 0.010099 4.958 7.11e-07 \*\*\*  
## gendermale -0.106798 0.314804 -0.339 0.73442   
## waist 0.139195 0.046699 2.981 0.00288 \*\*   
## bp.1s 0.003059 0.006338 0.483 0.62937   
## BMI -0.059106 0.040916 -1.445 0.14857   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 407.31 on 376 degrees of freedom  
## Residual deviance: 321.14 on 369 degrees of freedom  
## (26 observations deleted due to missingness)  
## AIC: 337.14  
##   
## Number of Fisher Scoring iterations: 5

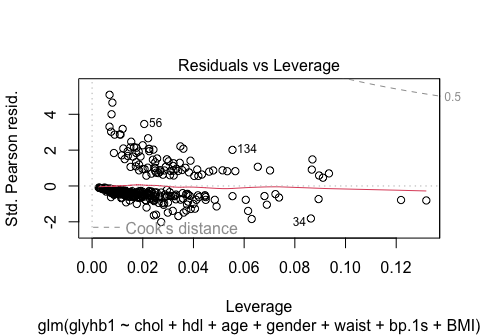
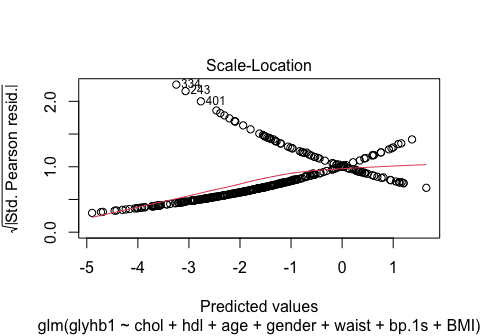
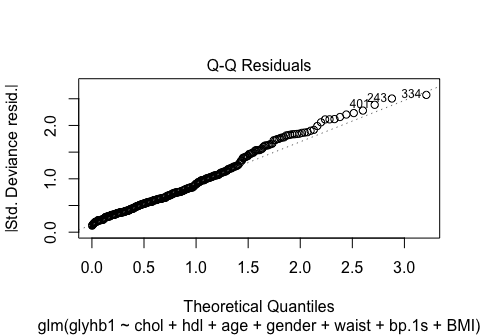
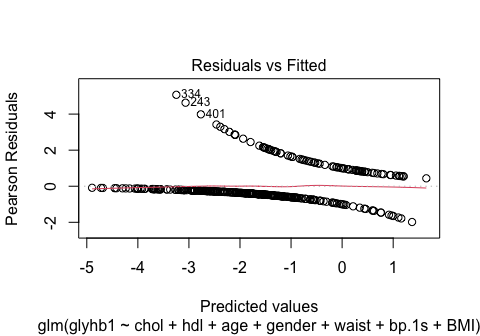
summary(model2)

##   
## Call:  
## glm(formula = glyhb2 ~ chol + hdl + age + gender + waist + bp.1s +   
## BMI, family = binomial(link = "logit"), data = diabetes)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.145063 1.666757 -5.487 4.09e-08 \*\*\*  
## chol 0.010972 0.003491 3.143 0.00167 \*\*   
## hdl -0.025155 0.010212 -2.463 0.01377 \*   
## age 0.046367 0.011218 4.133 3.58e-05 \*\*\*  
## gendermale 0.030598 0.350894 0.087 0.93051   
## waist 0.092515 0.049667 1.863 0.06250 .   
## bp.1s 0.005747 0.006881 0.835 0.40355   
## BMI -0.016083 0.043674 -0.368 0.71269   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.26 on 376 degrees of freedom  
## Residual deviance: 270.90 on 369 degrees of freedom  
## (26 observations deleted due to missingness)  
## AIC: 286.9  
##   
## Number of Fisher Scoring iterations: 5

To analyse the model performance on the datatsets we again perform an internal validation by inspecting residual plots.

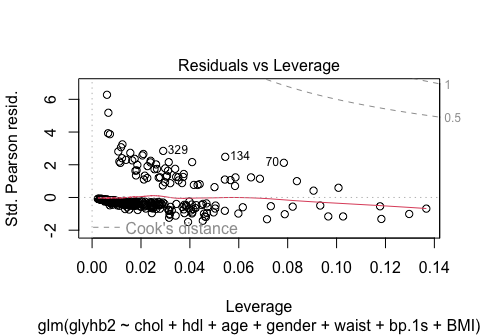
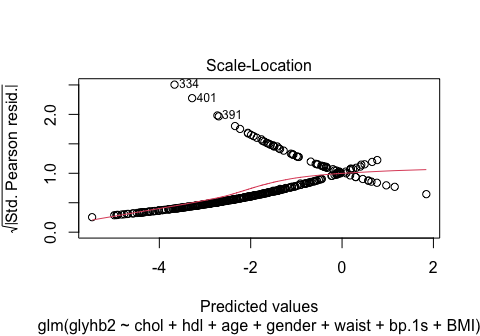
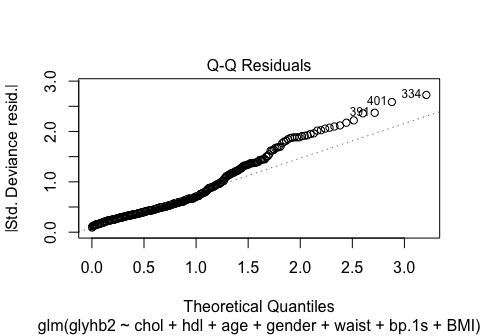
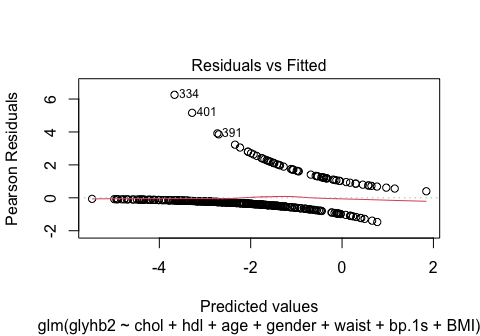
#### Residual Plots Model 1

plot(model1)



#### Residual Plots Model 2

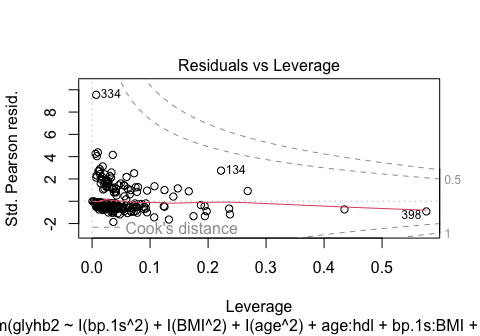
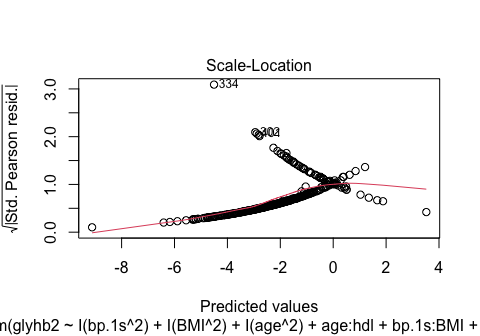
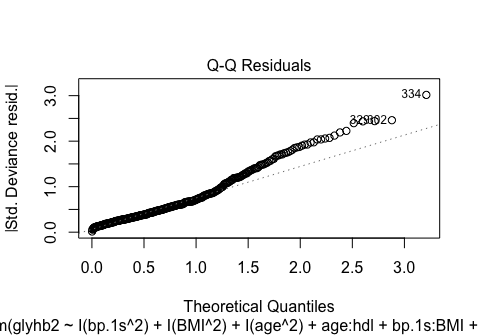
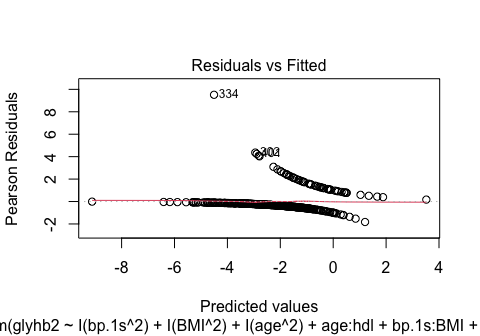
plot(model2)



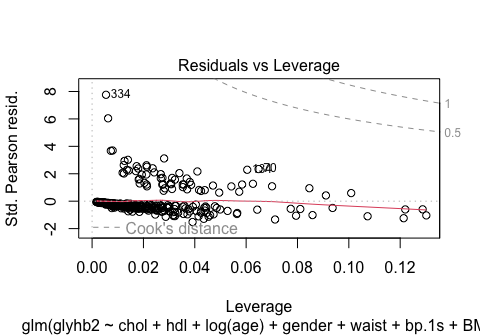
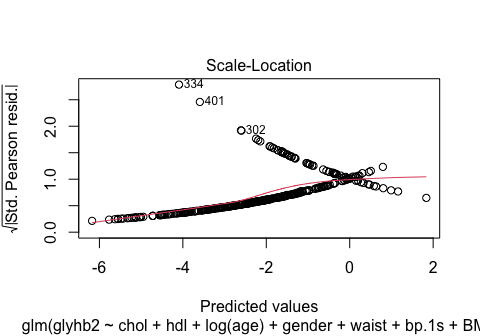
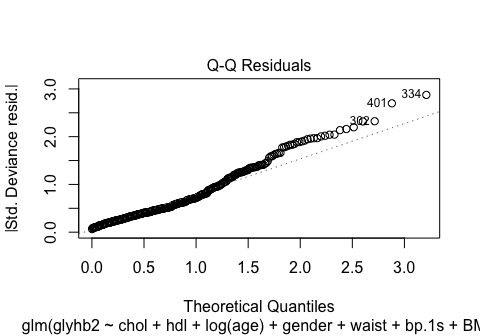
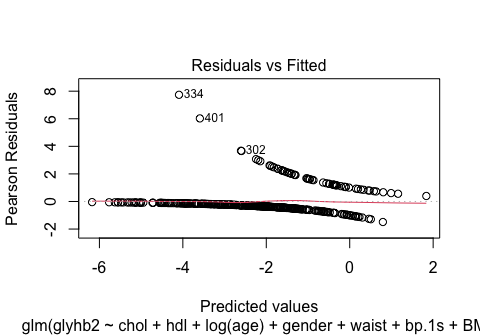
#### Anaylsis

The QQ-plot and scale-location plot for the second model (model2) do not look ideal. Therefore, we check for nonlinear functional forms of continuous predictors by using a model with interaction and quadratic terms for all predictors except for gender (assuming the integer variables such as age are considered as continuous). Utilizing an AIC-based model reduction of a fully-quadratic model we determined the following model, which seems to perform better w.r.t to the appearance of the QQ-plot.

model2 <- glm(glyhb2 ~ I(bp.1s^2) + I(BMI^2) + I(age^2) + age:hdl + bp.1s:BMI + waist:BMI + waist:bp.1s + hdl:bp.1s + hdl:waist + chol:hdl + chol:bp.1s , data=diabetes,family=binomial(link='logit'))  
plot(model2)

 The improved residual plots suggest that introducing the quadratic and mixed terms , , , , , , , , , and leads to slightly better model performance. Of course, this might still not be the optimal model since in principle other non-linear terms, e.g. log-terms or sqrt-terms, could be added, which might lead to better performance. Since age is the most significant predictor in model2, we experimented with several non-linear terms in this variable, finding that adding a simple logrithimic term produces a better QQ-plot than with AIC while keeping the model complexity lower:

modeltest <- glm(glyhb2 ~ chol + hdl + log(age) + gender + waist + bp.1s + BMI, data=diabetes,family=binomial(link='logit'))  
plot(modeltest)



summary(modeltest)

##   
## Call:  
## glm(formula = glyhb2 ~ chol + hdl + log(age) + gender + waist +   
## bp.1s + BMI, family = binomial(link = "logit"), data = diabetes)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -16.594155 2.595670 -6.393 1.63e-10 \*\*\*  
## chol 0.010689 0.003499 3.055 0.00225 \*\*   
## hdl -0.023542 0.010164 -2.316 0.02054 \*   
## log(age) 2.532441 0.601149 4.213 2.52e-05 \*\*\*  
## gendermale 0.032241 0.352635 0.091 0.92715   
## waist 0.092679 0.049818 1.860 0.06284 .   
## bp.1s 0.004976 0.006876 0.724 0.46929   
## BMI -0.014559 0.043780 -0.333 0.73948   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.26 on 376 degrees of freedom  
## Residual deviance: 268.22 on 369 degrees of freedom  
## (26 observations deleted due to missingness)  
## AIC: 284.22  
##   
## Number of Fisher Scoring iterations: 6

To describe the estimated associations of the predictors with the outcome variable in the model, we consider the model coefficients of model1 and the improved model2 and compare them to the ones of the linear model (without taking into account non-linear associations). In the original linear model, chol, hdl, age, and waist were significant predictors of the transformed glyhb whereas gender, bp.1s, and BMI were not statistically significant. Consistent to the linear model, the linear regression models highlight the key predictors chol, hdl, and age as important for predicting glyhb thresholds, but only model1 considers waist to be significant (as the linear model does). More specifically, both the linear and logisitc models consistently predict that higher cholesterol levels and older age are associated with increased odds of developing diabetes, whereas higher HDL levels reduce the odds. By introducing non-linear associations with AIC in model2 we have achieved better model fits at the cost of interpretability of the coefficients and potential overfitting. Experimentally, we found that adding a logarithimic term to model2 achieves better residual plots, the interpretation of the coefficients is straightforward and consistent to the linear model.

In conclusion, we have demonstrated that by implementing a simple transformation, the Box-Cox transformation, of the distribution of the dependent variable in the linear model the residual plots could be improved. On the other hand, the QQ-plots in the logistic model suggest that the residuals are less normally distributed than in the linear model, even if non-linear associations are taken into account.