HW8

Tianqi Wu

5/1/2020

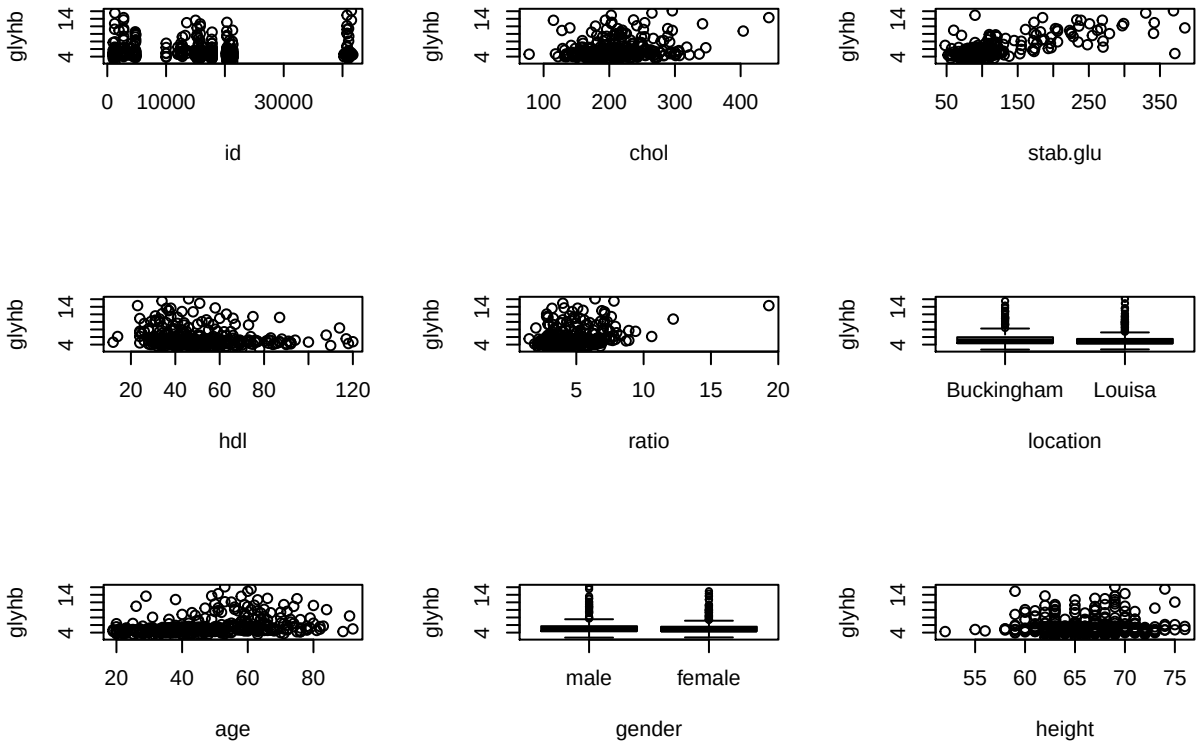
```
library(faraway)
library(rpart)
attach(diabetes)
attach(wbca)
```

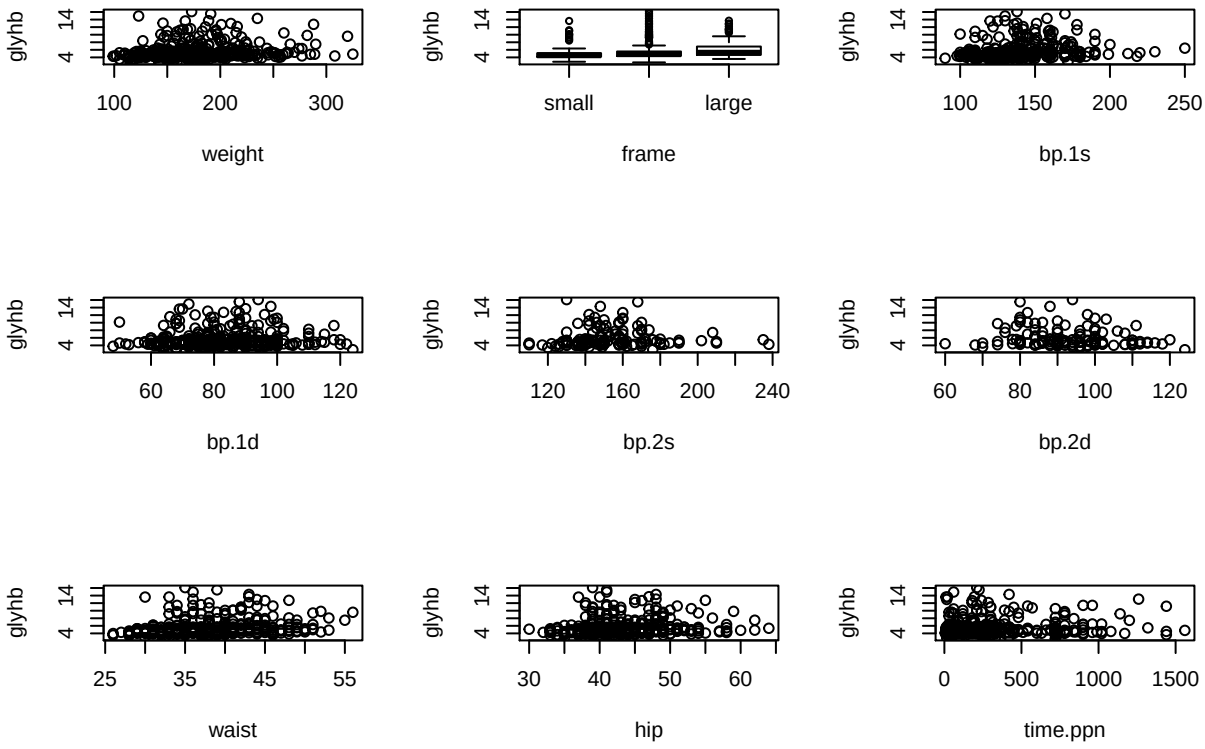
Problem 1

1(a)

From the plots, it seems that stab.glu is positively linearly related to glyhb. There is no other obvious linear trend and transformation may be needed.

```
par(mfrow = c(3, 3))
plot(glyhb~., data=diabetes)
```





1(b)

From summary, we can see that there are 262 missing values for bp.2s and bp.2d. Hence, we remove the two columns. Other variables have small amount of missing values (<15) and we just remove the rows with missing values. In total, we removed 2 columns and 37 rows. The reduced dataset has 366 observations.

```
summary(diabetes)
```

```
##        id             chol          stab.glu          hdl
##  Min.   : 1000   Min.   : 78.0   Min.   : 48.0   Min.   : 12.00
##  1st Qu.: 4792   1st Qu.:179.0   1st Qu.: 81.0   1st Qu.: 38.00
##  Median :15766   Median :204.0   Median : 89.0   Median : 46.00
##  Mean   :15978   Mean   :207.8   Mean   :106.7   Mean   : 50.45
##  3rd Qu.:20336   3rd Qu.:230.0   3rd Qu.:106.0   3rd Qu.: 59.00
##  Max.   :41756   Max.   :443.0   Max.   :385.0   Max.   :120.00
##                  NA's   :1                       NA's   :1
##      ratio            glyhb            location        age           gender
##  Min.   : 1.500   Min.   : 2.68   Buckingham:200   Min.   :19.00   male  :169
##  1st Qu.: 3.200   1st Qu.: 4.38   Louisa    :203   1st Qu.:34.00   female:234
##  Median : 4.200   Median : 4.84                    Median :45.00
##  Mean   : 4.522   Mean   : 5.59                    Mean   :46.85
##  3rd Qu.: 5.400   3rd Qu.: 5.60                    3rd Qu.:60.00
##  Max.   :19.300   Max.   :16.11                    Max.   :92.00
##  NA's   :1        NA's   :13
##      height          weight         frame         bp.1s           bp.1d
##  Min.   :52.00   Min.   : 99.0   small :104   Min.   : 90.0   Min.   : 48.00
##  1st Qu.:63.00   1st Qu.:151.0   medium:184   1st Qu.:121.2   1st Qu.: 75.00
##  Median :66.00   Median :172.5   large :103   Median :136.0   Median : 82.00
##  Mean   :66.02   Mean   :177.6   NA's  : 12   Mean   :136.9   Mean   : 83.32
##  3rd Qu.:69.00   3rd Qu.:200.0                3rd Qu.:146.8   3rd Qu.: 90.00
```

```
##  Max.   :76.00   Max.   :325.0                Max.   :250.0   Max.   :124.00
##  NA's   :5       NA's   :1                    NA's   :5       NA's   :5
##      bp.2s           bp.2d            waist           hip
##  Min.   :110.0   Min.   : 60.00   Min.   :26.0   Min.   :30.00
##  1st Qu.:138.0   1st Qu.: 84.00   1st Qu.:33.0   1st Qu.:39.00
##  Median :149.0   Median : 92.00   Median :37.0   Median :42.00
##  Mean   :152.4   Mean   : 92.52   Mean   :37.9   Mean   :43.04
##  3rd Qu.:161.0   3rd Qu.:100.00   3rd Qu.:41.0   3rd Qu.:46.00
##  Max.   :238.0   Max.   :124.00   Max.   :56.0   Max.   :64.00
##  NA's   :262     NA's   :262      NA's   :2      NA's   :2
##     time.ppn
##  Min.   :   5.0
##  1st Qu.:  90.0
##  Median : 240.0
##  Mean   : 341.2
##  3rd Qu.: 517.5
##  Max.   :1560.0
##  NA's   :3
```

```
data = subset(diabetes, select=-c(bp.2s,bp.2d))
sum(!complete.cases(data))
```

```
## [1] 37
```

```
which(is.na(data))
```

```
##  [1]  431 1237 1640 2059 2075 2080 2125 2132 2133 2208 2233 2289 2342 2343 2412
## [16] 2418 3691 3714 3823 3859 3945 4192 4484 4497 4503 4542 4544 4586 4658 4716
## [31] 4761 4766 4782 4814 4844 4850 4874 4900 5052 5247 5253 5277 5303 5455 5979
## [46] 6036 6382 6439 6559 6664 6720
```

```
data = na.omit(data)
```

1(c)

For stab.glu< 158, we have 324 obersvations and mean response is 5.028333. The largest terminal node is where the split gives largest number of observations. In this case, age<55.5 is the largest terminal node with 205 observations and the mean response is 4.598683.

```
model.1c = rpart(glyhb~., data=data)
model.1c
```

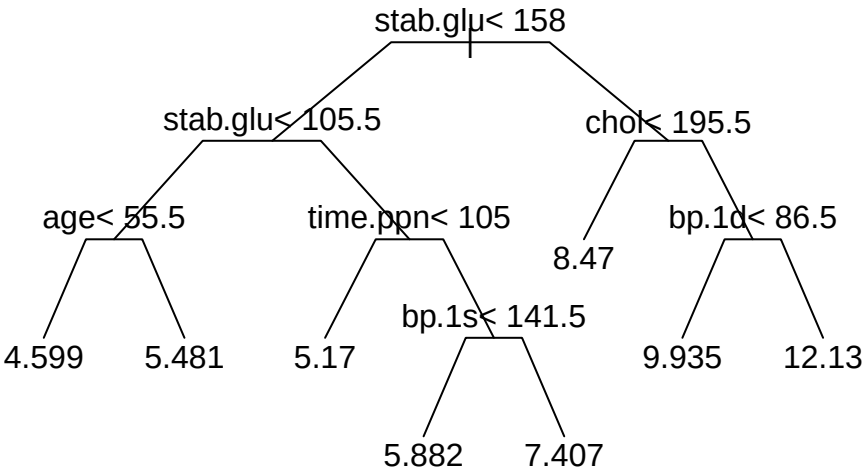
```
## n= 366
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 366 1816.85800  5.607295
##    2) stab.glu< 158 324  546.30990  5.028333
##      4) stab.glu< 105.5 268  306.23520  4.806007
##        8) age< 55.5 205  115.01170  4.598683 *
##        9) age>=55.5 63  153.73920  5.480635 *
##      5) stab.glu>=105.5 56  163.43180  6.092321
##       10) time.ppn< 105 22   27.56829  5.169546 *
##       11) time.ppn>=105 34  105.00860  6.689412
##         22) bp.1s< 141.5 16   40.96785  5.881875 *
```

```
##         23) bp.1s>=141.5 18   44.33236  7.407222 *
##    3) stab.glu>=158 42  324.14380 10.073570
##      6) chol< 195.5 17  106.69140  8.470000 *
##      7) chol>=195.5 25  144.01200 11.164000
##       14) bp.1d< 86.5 11   49.21207  9.935455 *
##       15) bp.1d>=86.5 14   65.15250 12.129290 *
```

1(d)

The most important predictor is the highest split in the tree. In this case, stab.glu is the most important predictor and it could split the observations most.

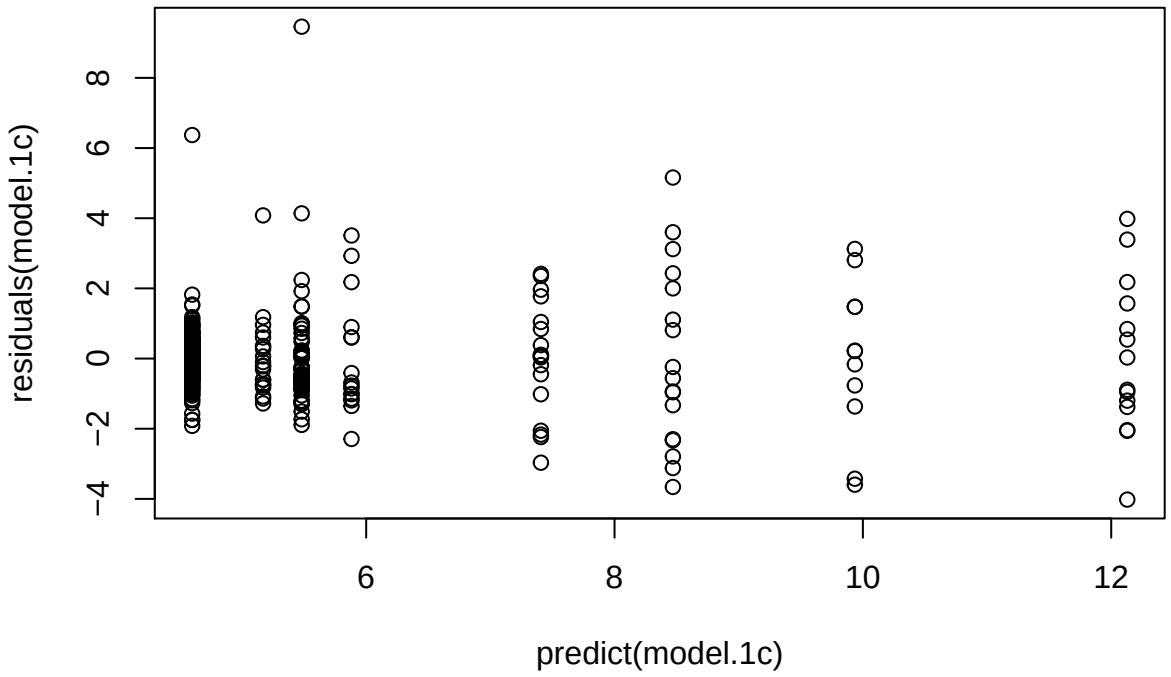
```
plot(model.1c,compress=T,uniform=T,branch=0.4,margin=.1)
text(model.1c)
```



1(e)

It seems that variance of residuals gets larger as fitted values increases. Though not obvious, we could see small trumpet pattern.

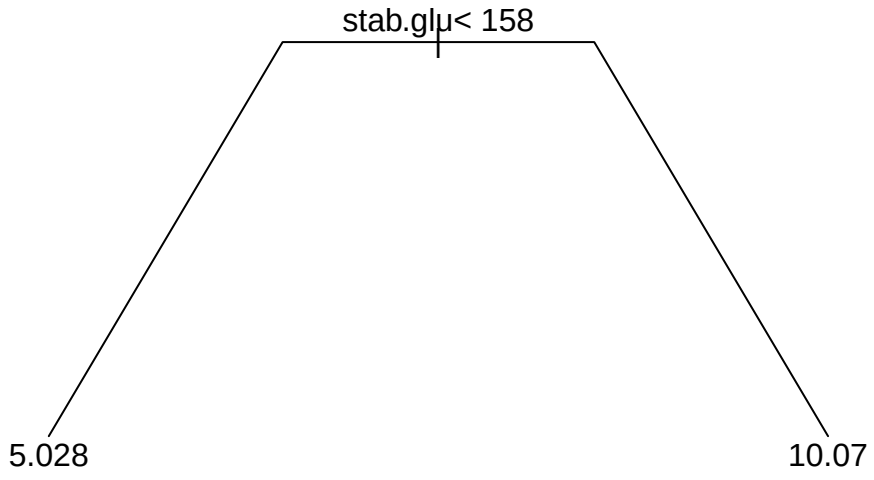
```
plot(predict(model.1c),residuals(model.1c))
```



1(f)

The smallest tree would only have one split: stab.glu< 158.

```
myCPtable = model.1c$cptable
id.min = which.min(myCPtable[,'xerror'])
my1se.err = myCPtable[id.min,'xerror'] + myCPtable[id.min, 'xstd']
id.1se = min(which(myCPtable[,'xerror'] < my1se.err))
CP.1se = (myCPtable[id.1se, 'CP'] + myCPtable[(id.1se-1), 'CP'])/2
tree.1se = prune.rpart(model.1c, CP.1se)
plot(tree.1se, compress=T,uniform=T,branch=0.4,margin=.1)
text(tree.1se)
```



Problem 2

2(a)

The residual deviance is 89.464 on 671 degrees of freedom. It can be used to determine if the models fits the data since it is like residual sum of square. The lower the residual deviance, the better the model.

```
model.2a = glm(Class~., data=wbca, family=binomial)
summary(model.2a)
```

```
##
## Call:
## glm(formula = Class ~ ., family = binomial, data = wbca)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48282  -0.01179   0.04739   0.09678   3.06425
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.16678    1.41491   7.892 2.97e-15 ***
## Adhes       -0.39681    0.13384  -2.965  0.00303 **
## BNucl       -0.41478    0.10230  -4.055 5.02e-05 ***
## Chrom       -0.56456    0.18728  -3.014  0.00257 **
## Epith       -0.06440    0.16595  -0.388  0.69795
## Mitos       -0.65713    0.36764  -1.787  0.07387 .
## NNucl       -0.28659    0.12620  -2.271  0.02315 *
## Thick       -0.62675    0.15890  -3.944 8.01e-05 ***
## UShap       -0.28011    0.25235  -1.110  0.26699
## USize        0.05718    0.23271   0.246  0.80589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 881.388  on 680  degrees of freedom
## Residual deviance:  89.464  on 671  degrees of freedom
## AIC: 109.46
##
## Number of Fisher Scoring iterations: 8
```

2(b)

According to AIC criterion, the best subset of variables would be: Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap.

```
model.2b = step(model.2a, scope=list(upper=~., lower=~1))
```

```
## Start:  AIC=109.46
## Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
##     UShap + USize
##
##         Df Deviance    AIC
## - USize  1   89.523 107.52
```

```
## - Epith  1   89.613 107.61
## - UShap  1   90.627 108.63
## <none>       89.464 109.46
## - Mitos  1   93.551 111.55
## - NNucl  1   95.204 113.20
## - Adhes  1   98.844 116.84
## - Chrom  1   99.841 117.84
## - BNucl  1  109.000 127.00
## - Thick  1  110.239 128.24
##
## Step:  AIC=107.52
## Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
##     UShap
##
##         Df Deviance    AIC
## - Epith  1   89.662 105.66
## - UShap  1   91.355 107.36
## <none>       89.523 107.52
## + USize  1   89.464 109.46
## - Mitos  1   93.552 109.55
## - NNucl  1   95.231 111.23
## - Adhes  1   99.042 115.04
## - Chrom  1  100.153 116.15
## - BNucl  1  109.064 125.06
## - Thick  1  110.465 126.47
##
## Step:  AIC=105.66
## Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap
##
##         Df Deviance    AIC
## <none>       89.662 105.66
## - UShap  1   91.884 105.88
## + Epith  1   89.523 107.52
## + USize  1   89.613 107.61
## - Mitos  1   93.714 107.71
## - NNucl  1   95.853 109.85
## - Adhes  1  100.126 114.13
## - Chrom  1  100.844 114.84
## - BNucl  1  109.762 123.76
## - Thick  1  110.632 124.63
```

2(c)

There are 11 false positives and 9 false negatives.

```
phat=model.2b$fitted.values;
mypred = (phat>0.5)
table(wbca$Class, mypred)
```

```
##    mypred
##     FALSE TRUE
##   0   227   11
##   1     9  434
```