Stat_101_C_HW1

Junhyuk Jang

April 10, 2017

SID : 004 728 134 LEC : 2 DIS : 2B

```r
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.2.5
```

```r
# Q1
df <- read.csv("~/Desktop/UCLA_Academic/Spring 2017/STAT 101_C/HW/Heart.csv")
mean(df$Chol)
```

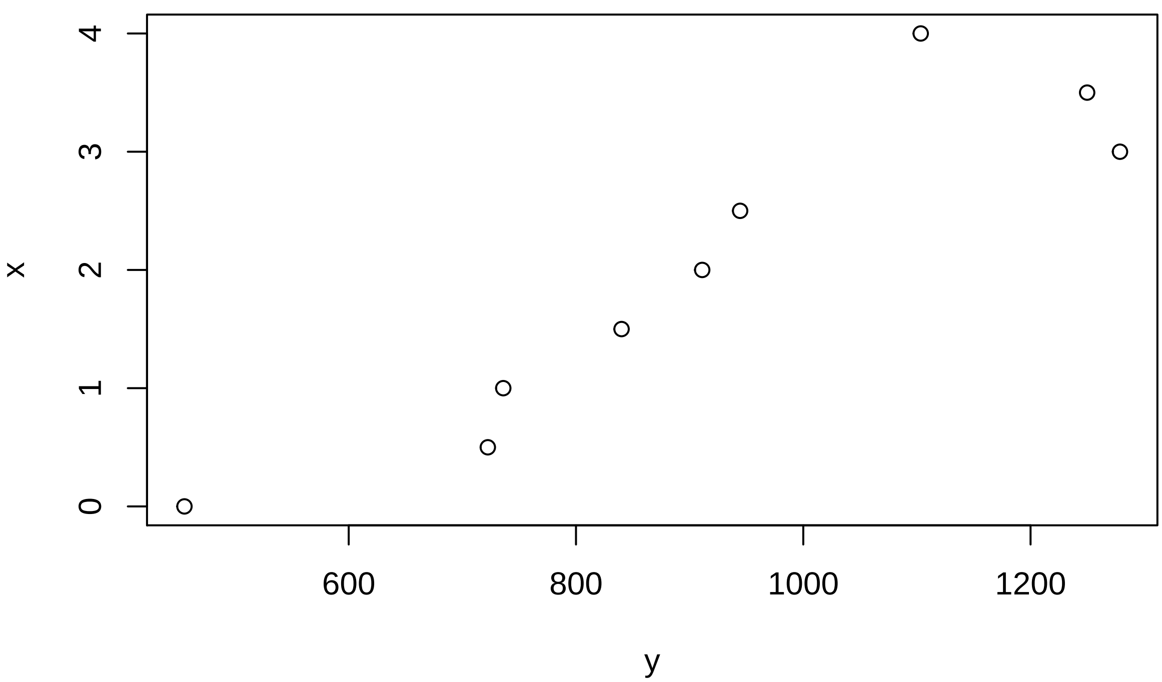
```
## [1] 246.6931
```

```r
# (a)
# prediction : (1) When MaxHR is greater than 150,and other predictors are fixed,
#                  predict the existence of heart disease.
#                (2) When Cholesterol level is greater than 250,and other
#                  predictors are fixed, predict the existence of heart disease.
#
# Inference : (1) Which predictor is the most strong effect on the
#                  heart disease?
#              (2) What is the relationship between age and heart disease

# (b)
df <- df[,-1]
df <- df[complete.cases(df),]
df$Sex = as.factor(df$Sex)
realsample = sample(seq(1,297),200,replace = F)
train = df[realsample,]
test = df[-realsample,]

logit.out = glm(AHD~.,family=binomial(link='logit'),data = train)
summary(logit.out)
```

```
##
## Call:
## glm(formula = AHD ~ ., family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8141  -0.4408  -0.1133   0.3117   2.1170
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.466302   3.941778  -0.626  0.53152
## Age                   0.003546   0.032691   0.108  0.91362
```

```
## Sex1                  1.849847   0.703203   2.631  0.00852 **
## ChestPainnonanginal -1.841063   0.604744  -3.044  0.00233 **
## ChestPainnontypical -2.012250   0.830641  -2.423  0.01541 *
## ChestPaintypical    -2.305084   0.815476  -2.827  0.00470 **
## RestBP               0.029469   0.013411   2.197  0.02800 *
## Chol                 0.005098   0.006012   0.848  0.39647
## Fbs                 -1.062689   0.753714  -1.410  0.15856
## RestECG              0.384379   0.247095   1.556  0.11981
## MaxHR               -0.026310   0.016003  -1.644  0.10017
## ExAng                1.424009   0.577570   2.466  0.01368 *
## Oldpeak              0.677030   0.293974   2.303  0.02128 *
## Slope               -0.353620   0.524702  -0.674  0.50035
## Ca                   0.954739   0.332064   2.875  0.00404 **
## Thalnormal          -1.220696   1.102029  -1.108  0.26800
## Thalreversable       0.172537   1.060844   0.163  0.87080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 276.54  on 199  degrees of freedom
## Residual deviance: 120.68  on 183  degrees of freedom
## AIC: 154.68
##
## Number of Fisher Scoring iterations: 6
```

```r
# Based on summary result, we  can say person's sex has the most strong
# effect on heart disease.

# Q2
df1 <- read.csv("~/Desktop/UCLA_Academic/Spring 2017/STAT 101_C/HW/hw1.csv")
plot(x~y,data = df1)
```



```r
# (a)
model1 <- lm(y~x,data = df1)
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -169.918  -60.665   -0.924   65.921  184.219
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   558.12      68.60   8.136 8.18e-05 ***
## x             178.78      28.82   6.204 0.000444 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 111.6 on 7 degrees of freedom
## Multiple R-squared:  0.8461, Adjusted R-squared:  0.8241
## F-statistic: 38.49 on 1 and 7 DF,  p-value: 0.0004436
```

```r
model2 <-  lm(y~poly(x,2,raw=TRUE),data = df1)
model3 <-  lm(y~poly(x,3,raw=TRUE),data = df1)
model4 <-  lm(y~poly(x,4,raw=TRUE),data = df1)
model5 <-  lm(y~poly(x,5,raw=TRUE),data = df1)

anova(model1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 479453  479453  38.488 0.0004436 ***
## Residuals  7  87201   12457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
(MSE_training_1 <- sum((model1$residuals)^2)/9)
```

```
## [1] 9688.946
```

```r
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: y
##                         Df Sum Sq Mean Sq F value   Pr(>F)
## poly(x, 2, raw = TRUE)  2 498280  249140  21.863 0.001757 **
## Residuals               6  68374   11396
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
(MSE_training_2 <- sum((model2$residuals)^2)/9)
```

```
## [1] 7597.056
```

```r
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: y
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## poly(x, 3, raw = TRUE)  3 505189  168396  13.699 0.007582 **
## Residuals             5  61465   12293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
(MSE_training_3 <- sum((model3$residuals)^2)/9)
```

```
## [1] 6829.43
```

```r
anova(model4)
```

```
## Analysis of Variance Table
##
## Response: y
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## poly(x, 4, raw = TRUE)  4 548344  137086  29.948 0.003065 **
## Residuals             4  18310    4577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
(MSE_training_4 <- sum((model4$residuals)^2)/9)
```

```
## [1] 2034.409
```

```r
anova(model5)
```

```
## Analysis of Variance Table
##
## Response: y
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## poly(x, 5, raw = TRUE)  5 548366  109673  17.991 0.01912 *
## Residuals             3  18288    6096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
(MSE_training_5 <- sum((model5$residuals)^2)/9)
```

```
## [1] 2032.052
```

```r
summary(model5)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 5, raw = TRUE), data = df1)
##
## Residuals:
##       1       2       3       4       5       6       7       8       9
##  -2.552  16.471 -39.354  33.121  26.828 -85.042  78.809 -34.170   5.888
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              457.986     77.713   5.893  0.00975 **
## poly(x, 5, raw = TRUE)1  795.440    505.187   1.575  0.21343
## poly(x, 5, raw = TRUE)2 -737.383    916.860  -0.804  0.48008
## poly(x, 5, raw = TRUE)3  293.429    618.048   0.475  0.66737
## poly(x, 5, raw = TRUE)4  -33.114    174.003  -0.190  0.86122
## poly(x, 5, raw = TRUE)5   -1.022     17.324  -0.059  0.95667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.08 on 3 degrees of freedom
## Multiple R-squared:  0.9677, Adjusted R-squared:  0.9139
## F-statistic: 17.99 on 5 and 3 DF,  p-value: 0.01912
```

```r
# (b)
# Based on only MSE value, model 5 has the smallest MSE. I decide to
# choose fifth-order polynomial model which is model 5.


# (c)
set.seed(123456)
x = seq(0,4,by=.5)
y = 500+200*x+rnorm(length(x),0,100)

predicty1 <- predict(model1,newdata = as.data.frame(x),type = "response")
predicty2 <- predict(model2,newdata = as.data.frame(x),type = "response")
predicty3 <- predict(model3,newdata = as.data.frame(x),type = "response")
predicty4 <- predict(model4,newdata = as.data.frame(x),type = "response")
predicty5 <- predict(model5,newdata = as.data.frame(x),type = "response")

(MSE_testing_1 <- sum((predicty1 - y)^2) / length(x))
```

```
## [1] 19129.35
```

```r
(MSE_testing_2 <- sum((predicty2 - y)^2) / length(x))
```

```
## [1] 21410.25
```

```r
(MSE_testing_3 <- sum((predicty3 - y)^2) / length(x))
```

```
## [1] 19643.81
```

```r
(MSE_testing_4 <- sum((predicty4 - y)^2) / length(x))
```

```
## [1] 26458.53
```

```r
(MSE_testing_5 <- sum((predicty5 - y)^2) / length(x))
```

```
## [1] 26385.96
```

```r
# (d)
# Different from the traing MSE, model1 has the smallest MSE in the
# testing data. It implies that the model 5 was overfitted. If I choose the model
# by considering low variance and low bias, I would choose model 1.
# The true model is y = 500+200*x and the model_1=558.12 + 178.78x. It seems
# the model 1 fit the a relationship the best.

# Q3
# (a)
# When we have the large sample size and small number of predictors, it would
# be better to use flexible model becasue we can predict the large number of
# parameters that are present in the model using the large number of sample size.

# (b)
# When we have the small number of observations is small, we cannot use a flexible
# statistical learning method. In this situation, use inflexible method is the
# best we can do.

#(c)
# When the relationship between the predictors and response is highly
# non-linear,we cannot use inflexible method because it is hard to fit a
# relationship. In this case, it is better to use the flexible model
# to fit a non-linear relationship for the model.

#(d)
# Simple would be better.The sentence "variance is extremely high" implies that
# observation is very far from true.The model achieved by the flexible method
# will involve all the noise.

# Q4
# (a) Classification (inference? prediction?)
#       (1) response: Whether got the leukemia or not.
#         predictors: cholesterol level,
#       a white blood cell lv,sex,a red blood cell lv
#               -> prediction
#       (2) response: Whether got the Hepatitis B Infections
#         predictors: height,weight,maxHR,Age,gender,jaundice
#               -> prediction
#       (3) response: Whether survived or not in the titanic accident.
#         predictors: passenger class, age, with family or alone
#               -> prediction
# (b) Regression     (inference? prediction?)
```

```
#       (1) stock alaysis response: the price of apple stock
#                  predictors: daily return, closing price, starting price,
#                              highest,lowest
#               -> prediction
#       (2) response: SAT score
#           predictors: mathe score, english score, nationality, Sex, Age
#               -> prediction
#       (3) response: teenagers' body weihgt
#           predictors: family income, age, sex
#               -> precition

# (c) Cluster analysis
#       (1) Divsion of the different countries into 3 groups. High GDP& democratic,
#           medium GDP&democratic,low GDP& democratic.
#           response variable is the differentiation of countries into one of
#           three catagories given above. Predictors are whether democratic
#           or not, GDP.
#          -> prediction.
#       (2) Division of  social class into "High or low or mediocre".
#           response variable is the differentiation of peoples' social class.
#           Predictors are occupations,income and educaiton.
#          -> prediction
#       (3) Division of companies into "big,medium,small".
#           reponse vaiable is the differentiation of company size.
#           Predictors are profit-making ,# of employee, #of affiliated companies.
#           -> prediction

# Q5
# (a)
# 1. Linearity of parameters. (y = x*beta + epsilon)
# 2. For all observations, the expected value of the error term is zero.
# 3. Variance of the error term is constant.
# 4. Error term is independently ditributed and not correlated.
# 5. x has no pattern with the error term.
# (b)
# If I got students SAT math score from school A,B,C,D,E and variance of each schools
# are different, I cannot get the best linear unbiased estimators becase assumption
# 3 is violated
```