# JieTang_Lab5.R

tjj

2020-07-28

```
#Name:Jie Tang
#Course:Machine learning using R 374815
#Quarter:Summer
#Insturctor name : Michael Chang

#Quiz code part:

library(ISLR)
attach(Wage)
#Q1
#Model 1
fit1=lm(wage~age+education,data=Wage)
coef(summary(fit1))
```

```
##                              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)                 60.335975 3.24571314 18.589435  4.363715e-73
## age                          0.568694 0.05719407  9.943234  6.096970e-23
## education2. HS Grad         11.438648 2.48025404  4.611886  4.157641e-06
## education3. Some College    24.167004 2.60975659  9.260252  3.778341e-20
## education4. College Grad    39.766772 2.59031246 15.352114  2.924513e-51
## education5. Advanced Degree 64.986565 2.80837932 23.140238 3.982660e-109
```

```
#Model 2
fit2=lm(wage~poly(age,2)+education,data=Wage)
coef(summary(fit2))
```

```
##                               Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)                    85.58266   2.157936  39.659509 8.142670e-277
## poly(age, 2)1                 362.37292  35.486584  10.211547  4.350272e-24
## poly(age, 2)2                -379.43228  35.449634 -10.703419  2.911059e-26
## education2. HS Grad            10.80191   2.435240   4.435668  9.509576e-06
## education3. Some College       23.23057   2.563121   9.063394  2.229316e-19
## education4. College Grad       38.00966   2.547836  14.918411  1.337237e-48
## education5. Advanced Degree    62.76452   2.764393  22.704628 1.989941e-105
```

```
#model 3
#by poly() we can avoid long formula
fit3=lm(wage~poly(age,3)+education,data=Wage)
coef(summary(fit3))
```

```
##                               Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)                    85.60597   2.156705  39.692941 3.548046e-277
## poly(age, 3)1                 362.66754  35.466163  10.225734  3.777795e-24
## poly(age, 3)2                -379.77717  35.429337 -10.719285  2.468951e-26
## poly(age, 3)3                  74.84933  35.309477   2.119808  3.410431e-02
## education2. HS Grad            10.86075   2.433978   4.462142  8.413874e-06
## education3. Some College       23.21846   2.561633   9.063929  2.219101e-19
## education4. College Grad       37.92991   2.546628  14.894169  1.877141e-48
## education5. Advanced Degree    62.61297   2.763706  22.655439 5.196118e-105
```

```
anova(fit1,fit2,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age + education
## Model 2: wage ~ poly(age, 2) + education
## Model 3: wage ~ poly(age, 3) + education
##   Res.Df     RSS Df Sum of Sq        F Pr(>F)
## 1   2994 3867992
## 2   2993 3725395  1    142597 114.6969 <2e-16 ***
## 3   2992 3719809  1      5587   4.4936 0.0341 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Q2
#get the lowest t value
coef(summary(fit3))
```

```
##                               Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)                    85.60597   2.156705  39.692941 3.548046e-277
## poly(age, 3)1                 362.66754  35.466163  10.225734  3.777795e-24
## poly(age, 3)2                -379.77717  35.429337 -10.719285  2.468951e-26
## poly(age, 3)3                  74.84933  35.309477   2.119808  3.410431e-02
## education2. HS Grad            10.86075   2.433978   4.462142  8.413874e-06
## education3. Some College       23.21846   2.561633   9.063929  2.219101e-19
## education4. College Grad       37.92991   2.546628  14.894169  1.877141e-48
## education5. Advanced Degree    62.61297   2.763706  22.655439 5.196118e-105
```

```
#Q3
#split age group, predict result by this model
fit=lm(wage~cut(age, breaks = c(0,25,35,45,55,80)),data=Wage)
coef(summary(fit))
```

```
##                                                 Estimate Std. Error
## (Intercept)                                     76.28175   2.636361
## cut(age, breaks = c(0, 25, 35, 45, 55, 80))(25,35] 27.88222   3.057242
## cut(age, breaks = c(0, 25, 35, 45, 55, 80))(35,45] 42.78532   2.957416
## cut(age, breaks = c(0, 25, 35, 45, 55, 80))(45,55] 41.34381   2.994987
## cut(age, breaks = c(0, 25, 35, 45, 55, 80))(55,80] 40.16115   3.296416
##                                                   t value      Pr(>|t|)
## (Intercept)                                     28.934488 1.525236e-162
## cut(age, breaks = c(0, 25, 35, 45, 55, 80))(25,35]  9.120056  1.341895e-19
```

```
## cut(age, breaks = c(0, 25, 35, 45, 55, 80))(35,45] 14.467130  6.658925e-46
## cut(age, breaks = c(0, 25, 35, 45, 55, 80))(45,55] 13.804335  4.535533e-42
## cut(age, breaks = c(0, 25, 35, 45, 55, 80))(55,80] 12.183278  2.319348e-33
```

```
predict(fit, data.frame(age=35))
```

```
##       1
## 104.164
```

```
#Q4
library(splines)
#model 1
fit=lm(wage~bs(age,knots=c(25,40,60)),data=Wage)
agelims=range(age)
age.grid=seq(from=agelims[1],to=agelims[2])

pred=predict (fit ,newdata =list(age=age.grid),se=T)
predict(fit, data.frame(age=55))
```

```
##        1
## 118.2185
```

```
#model 2
fit2=lm(wage~ns(age,df=4),data=Wage)
pred2=predict (fit2 ,newdata=list(age=age.grid),se=T)
predict(fit2, data.frame(age=55))
```

```
##       1
## 118.406
```

```
#model 3
fit3=smooth.spline(age,wage,cv=TRUE)
```

```
## Warning in smooth.spline(age, wage, cv = TRUE): cross-validation with non-unique
## 'x' values seems doubtful
```

```
predict(fit3, data.frame(age=55))
```

```
## $x
##   age
## 1  55
##
## $y
##        age
## 1 118.3031
```

```
#model 4
fit4=loess(wage~age,span=.5,data=Wage)
predict(fit4, data.frame(age=55))
```

```
##       1
## 117.593
```

```
#do the comparison

#Q5
#fit a gam predict wage
library(gam)
```

```
## Loading required package: foreach
```

```
## Loaded gam 1.20
```

```
gam=gam(wage~year+s(age,5)+education,data=Wage)
coef(gam)
```

```
##                  (Intercept)                         year
##                -2340.1755578                    1.1973264
##                     s(age, 5)         education2. HS Grad
##                     0.5664159                   10.9859573
##    education3. Some College    education4. College Grad
##                    23.5449861                   38.1979357
## education5. Advanced Degree
##                    62.6007203
```

```
#Q6
predict(gam, data.frame(year = 2008, age = 48, education ="5. Advanced Degree"))
```

```
##        1
## 156.9727
```

```
#Q7
gam=gam(wage~year+s(age,3)+education,data=Wage)
predict(gam, data.frame(year = 2008, age = 48, education ="5. Advanced Degree"))
```

```
##        1
## 157.7089
```

```
# # Chapter 7 Lab: Non-linear Modeling
#
#
# # Polynomial Regression and Step Functions
#
# fit=lm(wage~poly(age,4),data=Wage)
# coef(summary(fit))
# fit2=lm(wage~poly(age,4,raw=T),data=Wage)
# coef(summary(fit2))
# fit2a=lm(wage~age+I(age^2)+I(age^3)+I(age^4),data=Wage)
# coef(fit2a)
# fit2b=lm(wage~cbind(age,age^2,age^3,age^4),data=Wage)
```

```
# agelims=range(age)
# age.grid=seq(from=agelims[1],to=agelims[2])
# preds=predict(fit,newdata=list(age=age.grid),se=TRUE)
# se.bands=cbind(preds$fit+2*preds$se.fit,preds$fit-2*preds$se.fit)
# par(mfrow=c(1,2),mar=c(4.5,4.5,1,1),oma=c(0,0,4,0))
# plot(age,wage,xlim=agelims,cex=.5,col="darkgrey")
# title("Degree-4 Polynomial",outer=T)
# lines(age.grid,preds$fit,lwd=2,col="blue")
# matlines(age.grid,se.bands,lwd=1,col="blue",lty=3)
# preds2=predict(fit2,newdata=list(age=age.grid),se=TRUE)
# max(abs(preds$fit-preds2$fit))
# fit.1=lm(wage~age,data=Wage)
# fit.2=lm(wage~poly(age,2),data=Wage)
# fit.3=lm(wage~poly(age,3),data=Wage)
# fit.4=lm(wage~poly(age,4),data=Wage)
# fit.5=lm(wage~poly(age,5),data=Wage)
# anova(fit.1,fit.2,fit.3,fit.4,fit.5)
# coef(summary(fit.5))
# (-11.983)^2
# fit.1=lm(wage~education+age,data=Wage)
# fit.2=lm(wage~education+poly(age,2),data=Wage)
# fit.3=lm(wage~education+poly(age,3),data=Wage)
# anova(fit.1,fit.2,fit.3)
# fit=glm(I(wage>250)~poly(age,4),data=Wage,family=binomial)
# preds=predict(fit,newdata=list(age=age.grid),se=T)
# pfit=exp(preds$fit)/(1+exp(preds$fit))
# se.bands.logit = cbind(preds$fit+2*preds$se.fit, preds$fit-2*preds$se.fit)
# se.bands = exp(se.bands.logit)/(1+exp(se.bands.logit))
# preds=predict(fit,newdata=list(age=age.grid),type="response",se=T)
# plot(age,I(wage>250),xlim=agelims,type="n",ylim=c(0,.2))
# points(jitter(age), I((wage>250)/5),cex=.5,pch="|",col="darkgrey")
# lines(age.grid,pfit,lwd=2, col="blue")
# matlines(age.grid,se.bands,lwd=1,col="blue",lty=3)
# table(cut(age,4))
# fit=lm(wage~cut(age,4),data=Wage)
# coef(summary(fit))
#
# # Splines
#
# library(splines)
# fit=lm(wage~bs(age,knots=c(25,40,60)),data=Wage)
# pred=predict(fit,newdata=list(age=age.grid),se=T)
# plot(age,wage,col="gray")
# lines(age.grid,pred$fit,lwd=2)
# lines(age.grid,pred$fit+2*pred$se,lty="dashed")
# lines(age.grid,pred$fit-2*pred$se,lty="dashed")
# dim(bs(age,knots=c(25,40,60)))
# dim(bs(age,df=6))
# attr(bs(age,df=6),"knots")
# fit2=lm(wage~ns(age,df=4),data=Wage)
# pred2=predict(fit2,newdata=list(age=age.grid),se=T)
# lines(age.grid, pred2$fit,col="red",lwd=2)
# plot(age,wage,xlim=agelims,cex=.5,col="darkgrey")
```

```
# title("Smoothing Spline")
# fit=smooth.spline(age,wage,df=16)
# fit2=smooth.spline(age,wage,cv=TRUE)
# fit2$df
# lines(fit,col="red",lwd=2)
# lines(fit2,col="blue",lwd=2)
# legend("topright",legend=c("16 DF","6.8 DF"),col=c("red","blue"),lty=1,lwd=2,cex=.8)
# plot(age,wage,xlim=agelims,cex=.5,col="darkgrey")
# title("Local Regression")
# fit=loess(wage~age,span=.2,data=Wage)
# fit2=loess(wage~age,span=.5,data=Wage)
# lines(age.grid,predict(fit,data.frame(age=age.grid)),col="red",lwd=2)
# lines(age.grid,predict(fit2,data.frame(age=age.grid)),col="blue",lwd=2)
# legend("topright",legend=c("Span=0.2","Span=0.5"),col=c("red","blue"),lty=1,lwd=2,cex=.8)
#
# # GAMs
#
# gam1=lm(wage~ns(year,4)+ns(age,5)+education,data=Wage)
# library(gam)
# gam.m3=gam(wage~s(year,4)+s(age,5)+education,data=Wage)
# par(mfrow=c(1,3))
# plot(gam.m3, se=TRUE,col="blue")
# plot.Gam(gam1, se=TRUE, col="red")
# gam.m1=gam(wage~s(age,5)+education,data=Wage)
# gam.m2=gam(wage~year+s(age,5)+education,data=Wage)
# anova(gam.m1,gam.m2,gam.m3,test="F")
# summary(gam.m3)
# preds=predict(gam.m2,newdata=Wage)
# gam.lo=gam(wage~s(year,df=4)+lo(age,span=0.7)+education,data=Wage)
# plot.Gam(gam.lo, se=TRUE, col="green")
# gam.lo.i=gam(wage~lo(year,age,span=0.5)+education,data=Wage)
# library(akima)
# plot(gam.lo.i)
# gam.lr=gam(I(wage>250)~year+s(age,df=5)+education,family=binomial,data=Wage)
# par(mfrow=c(1,3))
# plot(gam.lr,se=T,col="green")
# table(education,I(wage>250))
# gam.lr.s=gam(I(wage>250)~year+s(age,df=5)+education,family=binomial,data=Wage,subset=(education!="1.
# plot(gam.lr.s,se=T,col="green")
```