**Machine Learning HW4**

**Section 1.22**

1. In Section 1.12.1.2, the reader was reminded that the results of a cross- validation are random, due to the random partitioning into training and test sets. Try doing several runs of the linear and k-NN code in that section, comparing results. (P.61)

setwd("Desktop/MLDatasets/")

bodyfat=read\_csv("bodyfat.csv")

#p=0.3, create train and valid datasets

n=nrow(bodyfat)

ntrain=round(0.5\*n)

trainidxs=sample(1:n,ntrain,replace=FALSE)

train=bodyfat[trainidxs,]

valid=bodyfat[-trainidxs,]

#ols, col6=age, 7=weight, 8=height

trainy=as.matrix(train[,7])

trainpredx=as.matrix(train[,c(6,8)])

lmout=lm(trainy~trainpredx)

#apply fitted model to validation dataset

validpredx=as.matrix(valid[,c(6,8)])

predy=cbind(1,validpredx)%\*%coef(lmout)

realy=valid[,7]

mean\_error=mean(as.matrix(abs(predy-realy)))

mean\_error

#cross validation: how well does ols predicts

# predvariable\_x=age&height, y=weight; mean\_error;

# p=0.3, 1st attempt=20.49054; 2nd:20.34851; 3rd:19.27954

# p=0.5, 1st:20.89973; 2nd:23.28791; 3rd:21.94913;

#k-NN, set k=25

xd=preprocessx(train[,c(6,8)],25)

kout=knnest(train[,7],xd,25)

predy\_knn=predict(kout,as.matrix(valid[,c(6,8)]),TRUE)

mean\_error\_knn=mean(as.matrix(abs(predy\_knn-realy)))

mean\_error\_knn

#cross validation: how well does k—NN predicts

# p=0.3, 1st attempt=19.07635; 2nd:20.58424; 3rd:18.39795

# p=0.5, 1st:20.10689; 2nd:22.00397; 3rd:21.04306;

**# On average, K-NN results in less absolute error than OLS does**

2. Extend (1.28: P.40) to include interaction terms for age and gender, and age2 and gender. Run the new model, and find the estimated effect of being female, for a 32-year-old person with a Master’s degree.

setwd("Desktop/MLDatasets/")

prgeng=read\_csv("prgeng.csv")

#include interaction between age gender, age^2 and gender

age2=prgeng$age^2

edu=prgeng$educ

prgeng$ms=as.integer(edu==14)

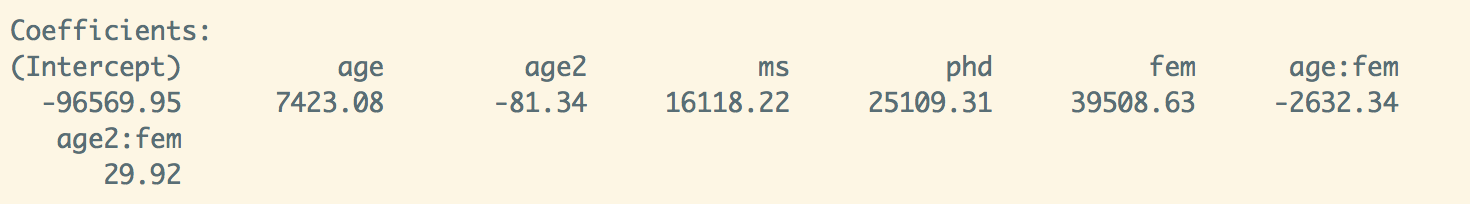
prgeng$phd=as.integer(edu==16)

prgeng$fem=prgeng$sex-1

# age:gender create interaction automatically;

# not include wkswrkd as a variable since the sencond question doesn't include it

lm\_out=lm(wageinc~age+age2+ms+phd+fem+age:fem+age2:fem,data=prgeng)



# Coefficients:

# (Intercept) age age2 ms phd fem age:fem age2:fem

# -96569.95 7423.08 -81.34 16118.22 25109.31 39508.63 2632.34 2632.34 29.92

#female, 32 years old, ms degree

dot(as.matrix(lm\_out$coefficients),cbind(1,32,32^2,1,0,1,32,32^2))

**# wageinc=59708.15**

3. Consider the bodyfat data mentioned in Section 1.2. Use lm() to form a prediction equation for density from the other variables (skipping the first three), and comment on whether use of indirect methods in this way seems feasible.

out\_bodyfat=lm(density~ siri+neck+chest+abdomen+hip+thigh+knee+ankle+biceps+forearm+wrist,bodyfat)

summary(out\_bodyfat)

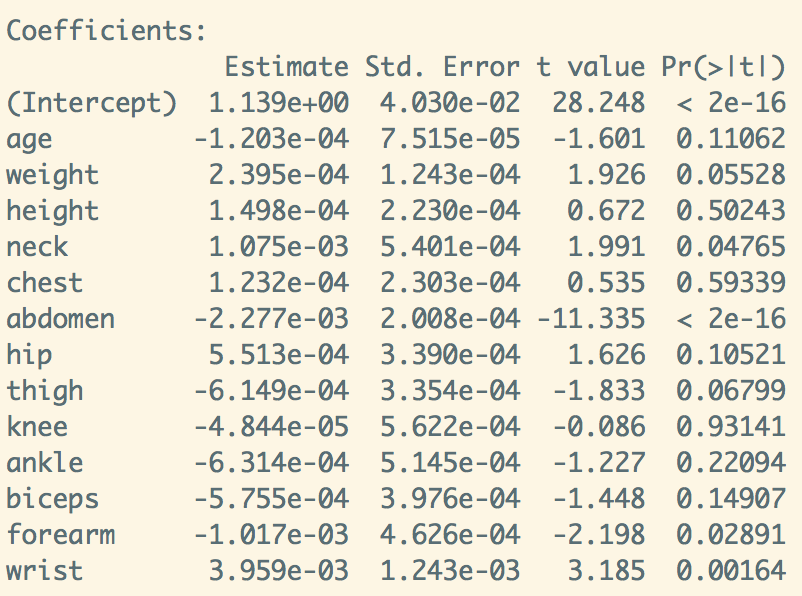
# Coefficients:

(Intercept) age weight height neck chest abdomen

1.139e+00 -1.203e-04 2.395e-04 1.498e-04 1.075e-03 1.232e-04 -2.277e-03

hip thigh knee ankle biceps forearm wrist

5.513e-04 -6.149e-04 -4.844e-05 -6.314e-04 -5.755e-04 -1.017e-03 3.959e-03



**# By checking the p-value of each variable, some variables like knee, height, chest, ankle, biceps are not significant. If we use this method, we should measure the correlation between each parameter and estimate, test the dependency or independency between each parameter, as well as the linearity.**

4. In Section 1.19.5.2, we gave this intuitive explanation:

In other words, the national mean height is a weighted average of the state means, with the weight for each state being its proportion of the national population. Replace state by gender in the following.

* (a)  Write English prose that relates the overall mean height of people and the gender-specific mean heights.

**The overall national mean height is a weighted average of national gender-specific means, with the weight for gender being its proportion of the national population.**

**To get the national gender-specific mean height, we calculated a weighted average of the state means of each gender, with the proportion for each state being its proportion of the national population.**

* (b)  Write English prose that relates the overall proportion of people taller than 70 inches to the gender-specific proportions.

**The national mean height of people taller than 70 inches is a weighted average of state means of people who are taller than 70 inches, with the for each state being its proportion of the national population.**

**Section 2.14**

1. Consider the census data in Section 1.16.1.

* (a)  Form an approximate 95% confidence interval for β6(fem) in the model (1.28).

out\_a=lm(wageinc~age+age2+wkswrkd+ms+phd+fem,data=prgeng)

summary(out\_a)

# 95% confidence interval for beta\_6=fem; estimate(fem)=-11484.49, std.e=705.3

c(-11484.49-705.30\*1.96,-11484.49+705.30\*1.96)

**# (-12866.88, -10102.10)**

* (b)  Form an approximate 95% confidence interval for the gender effect for Master’s degree holders, β6 + β7, in the model (1.28).

out\_b=lm(wageinc~age+age2+wkswrkd+ms+phd+fem+ms:fem+phd:fem,

data=prgeng)

summary(out\_b)

# beta6:estimate=-10276.797,std.e=804.498;beta7:estimate=-4157.253,

std.e=1728.329

c(-10276.797+-4157.253-1.96\*(804.498+1728.329),-10276.797+- 4157.253

+1.96\*(804.498+1728.329))

**# (-19398.39, -9469.709)**

2. The full bikeshare dataset spans 3 years’ time. Our analyses here have only used the first year. Extend the analysis in Section 2.8.5 to the full data set, adding dummy variables indicating the second and third year. Form an approximate 95% confidence interval for the difference between the coefficients of these two dummies.

day=read.csv("day.csv")

day$temp2=day$temp^2

day$clearday=as.integer(day$weathersit==1)

names(day)[15]="reg"

out\_day=lm(day$reg~temp+temp2+yr+workingday+clearday,data=day)

summary(out\_day)

# there is no second and third year, only 2011 and 2012 use variable yr

c(1716.25-56.68\*1.96, 1716.25+56.68\*1.96)

**# 95% confidence interval of year variable is (1605.157 1827.343)**

3. Suppose we are studying growth patterns in children, at k particular ages. Denote the height of the ith child in our sample data at age j by Hij, with Hi = (Hi1,...,Hik)′ denoting the data for child i. Suppose the population distribution of each Hi is k-variate normal with mean vector μ and covariance matrix Σ. Say we are interested in successive differences in heights, Dij =Hi,j+1−Hij, j=1,2,...,k−1. Define Di =(Di1,...,Dik-1)′. Explain why each Di is (k−1)-variate normal, and derive matrix expressions for the mean vector and covariance matrices.

# why each Di is (k−1)-variate normal

Proof:

Let U= (Hi2,...,Hik)′ , V= (Hi1,...,Hik-1)′ , given U~N(µU, σU^2 ), V~N(µV, σV^2 )

then, Di=U-V # note U start at Hi2 instead of Hi1

E[U-V]=E[U]-E[V]=µU-µV

Var(U-V)=Var(U)+Var(V)= σU^2+ σV^2

**Property: subtraction of two normal distributed vector is still normal, so normality satisfied**

**Mean Variance**

**(U-V)~N(µU-µV, σU^2+ σV^2-2** cov(U,V)**)**

In R, U-V mean vector= mean(µU-µV)

U-V covariance matrix= cov(U, V, use=”everything”, method=”pearson”)

4. In the simulation in Section 2.9.3, it is claimed that ρ2 = 0.50. Confirm this through derivation.

# getr2: function Estimate percentage of variation explained; package: CollapsABEL

simr2=function(n,p,nreps) {

r2s=vector(length=nreps)

for (i in 1:nreps) {

x=matrix(rnorm(n\*p),ncol=p)

y=x%\*%rep(1,p)+rnorm(n,sd=sqrt(p))

r2s[i]=getr2(x,y)

}

x

y

hist(r2s)}

getr2=function(x,y) {

smm=summary(lm(y~x))

smm$r.squared

}

**# use simr2(250,8,1000) in console, R^2 near 0.5**