CSSS 508: Intro to R

3/17/06

Final Solutions

*These solutions are not the only acceptable answers. Many of the questions were open-ended and will be graded on their own rather than against these solutions*.

Download final.dat from the class website.

final<-read.table(“<http://www.stat.washington.edu/rnugent/teaching/csss508/final.dat>”)

Did we get all of the data?

dim(final)

[1] 506 14

final[1,]

crim zn indus chas nox rm age dis rad tax ptratio black lstat medv

1 0.00632 18 2.31 0 0.538 6.575 65.2 4.09 1 296 15.3 396.9 4.98 24

So we can work with the individual variables:

attach(final)

A reminder of the variables:

*crim*: per capita crime rate by town

*zn*: proportion of residential land zoned for lots over 25,000 sq.ft.

*indus*: proportion of non-retail business acres per town

*chas*: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

*nox*: nitrogen oxides concentration (parts per 10 million)

*rm*: average number of rooms per dwelling

*age*: proportion of owner-occupied units built prior to 1940

*dis*: weighted mean of distances to five Boston employment centres

*rad*: index of accessibility to radial highways

*tax*: full-value property-tax rate per $10,000

*ptratio*: pupil-teacher ratio by town

*black*: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

*lstat*: lower status of the population (percent)

*medv*: median value of owner-occupied homes in $1000

Before we answer the questions, let’s look at what we have:

summary(final)

crim zn indus chas

Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000

1st Qu.: 0.08190 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000

Median : 0.25651 Median : 0.00 Median : 9.69 Median :0.00000

Mean : 3.66439 Mean : 11.38 Mean :11.21 Mean :0.06986

3rd Qu.: 3.69599 3rd Qu.: 12.50 3rd Qu.:18.10 3rd Qu.:0.00000

Max. :88.97620 Max. :100.00 Max. :27.74 Max. :1.00000

NA's : 8.00000 NA's : 10.00 NA's : 8.00 NA's :5.00000

nox rm age dis

Min. :0.3850 Min. :3.561 Min. : 6.00 Min. : 1.130

1st Qu.:0.4490 1st Qu.:5.883 1st Qu.: 45.10 1st Qu.: 2.100

Median :0.5380 Median :6.209 Median : 77.70 Median : 3.216

Mean :0.5544 Mean :6.284 Mean : 68.73 Mean : 3.797

3rd Qu.:0.6240 3rd Qu.:6.626 3rd Qu.: 94.10 3rd Qu.: 5.212

Max. :0.8710 Max. :8.780 Max. :100.00 Max. :12.127

NA's :2.0000 NA's :6.000 NA's : 5.00 NA's : 5.000

rad tax ptratio black

Min. : 1.000 Min. :187.0 Min. :12.60 Min. : 0.32

1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.:375.33

Median : 5.000 Median :330.0 Median :19.10 Median :391.34

Mean : 9.583 Mean :409.4 Mean :18.47 Mean :356.63

3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:396.21

Max. :24.000 Max. :711.0 Max. :22.00 Max. :396.90

NA's : 7.000 NA's : 5.0 NA's : 3.00 NA's : 9.00

lstat medv

Min. : 1.730 Min. : 5.00

1st Qu.: 6.928 1st Qu.:16.80

Median :11.465 Median :21.20

Mean :12.708 Mean :22.52

3rd Qu.:17.093 3rd Qu.:25.00

Max. :37.970 Max. :50.00

NA's : 6.000 NA's : 8.00

Note that the *chas* was treated as numeric even though it’s really categorical (Yes/No).

table(chas)

chas

0 1

466 35

We have missing data in all the variables. In our analysis, we have a couple of choices:

A) We could remove all suburbs that have one or missing values.

> sub.miss<-NULL

> for(i in 1:506)

+ if(any(is.na(final[i,]))) sub.miss<-c(sub.miss,i)

> sub.miss

[1] 2 4 9 11 15 19 23 27 31 33 37 41 42 45 60 73 78 87 101

[20] 104 109 115 116 138 150 151 152 170 175 176 186 191 199 202 206 207 219 224

[39] 228 235 238 242 251 256 262 270 272 275 276 278 285 289 294 298 299 309 312

[58] 313 314 317 318 321 327 335 345 351 353 359 376 382 383 413 423 435 436 450

[77] 454 456 464 497 506

This choice would remove 81 rows (suburbs) from our data matrix.

B) We could just remove missing values while doing the analyses (na.rm = T, etc.). Some of these suburbs are only missing one value and removing them from the analysis completely (as in choice #1) can delete useful information. We’ll be using option #2 for the first few questions. When modeling in questions 5 and 6, the R procedures will remove all observations/suburbs that have one or more missing values (option #1). But we’re going to try to maximize using the information we have while we can.

*1) The majority of the zn values are zero. That is, many suburbs have no residential land zoned for lots over 25,000 sq ft. Compare the mean and variance of indus and rad for those suburbs who have no residential land zoned and those who have at least some residential land zoned.*

table(zn)

zn

0 12.5 17.5 18 20 21 22 25 28 30 33 34 35 40 45 52.5

364 10 1 1 21 4 10 10 3 6 4 3 3 7 6 2

55 60 70 75 80 82.5 85 90 95 100

2 4 3 3 15 2 2 5 4 1

We need to create a categorical variable indicating whether or not the suburb has no residential land zoned.

zn.zero<-ifelse(zn==0,1,0)

table(zn.zero)

zn.zero

0 1

132 364

Looking at the relationship between *zn* and *indus* (proportion of non-retail business acres)

boxplot(indus[zn.zero==1],indus[zn.zero==0],names=c(“Zero Res. Land Zoned”, “Some Res. Land Zoned”),main=”Proportion of Non-Retail Business Acres”)



mean(indus[zn.zero==1],na.rm=T)

13.64589

var(indus[zn.zero==1],na.rm=T)

39.9314

mean(indus[zn.zero==0],na.rm=T)

4.471641

var(indus[zn.zero==0],na.rm=T)

6.187686

The suburbs with no residential land zoned for lots over 25,000 sq ft have a much higher proportion of non-retail business acres. Large lots usually correspond to business parks, strip shopping centers, shopping malls, etc. It would make sense that areas with no zoned lots of this size would have a higher proportion of non-retail businesses. The variance of this group is also much higher, likely due to outliers.

To test if the difference in means is significant, we could use a t-test.

t.test(indus[zn.zero==1],indus[zn.zero==0])

Welch Two Sample t-test

data: indus[zn.zero == 1] and indus[zn.zero == 0]

t = 22.9887, df = 481.554, p-value = < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

8.390102 9.958394

sample estimates:

mean of x mean of y

13.645889 4.471641

We do have a significant difference between the mean of the two groups.

Now looking at the relationship between *zn* and *rad* (index of accessibility to radial highways – highways which lead out of the city)

boxplot(rad[zn.zero==1],rad[zn.zero==0],names=c(“Zero Res. Land Zoned”, “Some Res. Land Zoned”),main=”Index of Accessibility to Radial Highways”)



mean(rad[zn.zero==1],na.rm=T)

11.51811

var(rad[zn.zero==1],na.rm=T)

90.14981

mean(rad[zn.zero==0],na.rm=T)

4.484615

var(rad[zn.zero==0],na.rm=T)

3.398986

Suburbs with no residential land zoned for large lots have a higher average accessibility to radial highways. However, the variance is extremely large (because of the 130 24’s).

table(rad[zn.zero==1])

1 2 3 4 5 6 8 24

6 18 24 72 75 17 17 130

We can test if this difference in means is significant:

t.test(rad[zn.zero==1],rad[zn.zero==0])

Welch Two Sample t-test

data: rad[zn.zero == 1] and rad[zn.zero == 0]

t = 13.3576, df = 423.684, p-value = < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

5.998506 8.068475

sample estimates:

mean of x mean of y

11.518106 4.484615

Yes, we do have a significant difference in the means of the two groups.

*2) The rm variable is real-valued (because it is an average). Create a categorical variable rm2 that rounds the number of rooms to the nearest integer. What are the different values of number of rooms now and how often do each of them occur?*

First, a reminder of what *rm* looks like:

summary(rm)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

3.561 5.883 6.209 6.284 6.626 8.780 6.000

rm2<-round(rm,0)

summary(rm2)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

4.000 6.000 6.000 6.268 7.000 9.000 6.000

Different values of number of rooms:

sort(unique(rm2))

[1] 4 5 6 7 8 9

How often do each of them occur?

table(rm2)

rm2

4 5 6 7 8 9

5 37 307 124 24 3

The most frequent values are 6-room and 7-room houses.

*3) Use graphs to illustrate the differences (for at least three variables) between towns with a pupil-teacher ratio of less than 20 and towns with a pupil-teacher ratio of greater than or equal to 20. The choice of graph should be appropriate for the variable. Graphs should also be clearly labeled. Describe what you see.*

First we need to create a categorical variable indicating whether or not the suburb’s pupil-teacher ratio is greater than or equal to 20.

summary(ptratio)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

12.60 17.40 19.10 18.47 20.20 22.00 3.00

pt.20<-ifelse(ptratio>=20,1,0)

table(pt.20)

pt.20

0 1

302 201

Graphically looking at differences (a few at a time):

m<-matrix(c(1,2,3,4,5,6),ncol=2)

layout(m)

#Crime Rate: very different distributions; best looked at with histograms

hist(crim[pt.20==1],xlab=”Per Capita Crime Rate”,main=”Crime: PT Ratio >= 20”)

hist(crim[pt.20==0],xlab=”Per Capita Crime Rate”,main=”Crime: PT Ratio < 20”)

# Proportion of Residential Land zoned:

#the huge frequency of *zn* = 0 makes the distribution hard to graph. Probably more useful to look at the

#table of the two indicator variables (*zn* = 0 yes/no vs. *pt.ratio* >= 20 yes/no)

barplot(table(zn.zero,pt.20),beside=T, names=c(“Pt.Ratio >=20”, “Pt.Ratio<20”),legend.text=c(“None Zoned”, “Some Zoned”),main=”Residential Zoning vs. PT Ratio”)

#Proportion of Non-retail business acres per town: very different distributions; best looked at with histograms

hist(indus[pt.20==1],xlab="Proportion of Non-Retail Business Acres",main="Indus: Pt Ratio >=20")

hist(indus[pt.20==0],xlab="Proportion of Non-Retail Business Acres",main="Indus: Pt Ratio < 20")

#Charles River: dummy variable; again we look at the table of the two indicator variables

barplot(table(chas,pt.20),beside=T, names=c(“Pt.Ratio >=20”, “Pt.Ratio<20”),legend.text=c(“Bounds River”, “Doesn’t Bound River”),main=”Charles River vs. PT Ratio”)



The per capita crime rate appears to be much higher in towns with a greater pupil-teacher ratio. The proportion of non-retail business acres is not distributed that much differently when you look at the actual frequencies excepting a very large peak around 15-20% in the suburbs with a large pupil-teacher ratio. Differences in amount of residential land zoned are apparent in suburbs with a low pupil-teacher ratio. The differences between suburbs on the river and off the river do not appear to depend greatly on the pupil-teacher ratio.

par(mfrow=c(3,2))

#Nitrogen Oxides Concentration (parts per 10 million)

boxplot(nox[pt.20==1],nox[pt.20==0],names=c(“Pt.Ratio >=20”, “Pt.Ratio<20”),main=”NOX vs. PT Ratio”)

#Average number of rooms per dwelling

boxplot(rm[pt.20==1],rm[pt.20==0],names=c(“Pt.Ratio >=20”, “Pt.Ratio<20”),main=”Avg # of Rooms vs. PT Ratio”)

#Proportion of owner-occupied units built prior to 1940

boxplot(age[pt.20==1],age[pt.20==0],names=c(“Pt.Ratio >=20”, “Pt.Ratio<20”),main=”# of Units built pre-1940 vs. PT Ratio”)

#weighted mean of distances to five Boston employment centers

boxplot(dis[pt.20==1],dis[pt.20==0],names=c(“Pt.Ratio >=20”, “Pt.Ratio<20”),main=”Dist to Employment Centers vs. PT Ratio”)

# Index of Accessibility to radial highways: very different distributions but integer-valued

#better to use barplots but we want the ranges to be the same

rad1.freq<-rad2.freq<-rep(0,max(rad,na.rm=T)-min(rad,na.rm=T)+1)

for(i in 1:506){

if(!is.na(pt.20[i])){

if(pt.20[i]==1) rad1.freq[rad[i]]<-rad1.freq[rad[i]]+1

if(pt.20[i]==0) rad2.freq[rad[i]]<-rad2.freq[rad[i]]+1

}

}

barplot(rad1.freq,xlab=”Index of Accessibility to Radial Highways”,names=seq(1,24),col=3,main=”Rad: PT Ratio >=20”)

barplot(rad2.freq,xlab=”Index of Accessibility to Radial Highways”,names=seq(1,24),col=3,main=”Rad: PT Ratio < 20”)



In suburbs with a higher pupil-teacher ratio, the nitrogen oxides concentration appears to be higher. The average number of rooms doesn’t seem to be dependent on the pupil-teacher ratio; suburbs with a low pupil-teacher ratio perhaps have slightly larger houses. Suburbs with high pupil-teacher ratios have older houses (but note the large number of outliers) and are closet to employment centers. With regards to accessibility to radial highways, there appears to be a group of high pupil-teacher ratio suburbs with better access. Otherwise, the shape of the distributions is not that different (taking into account the frequencies).

m<-matrix(c(1,2,5,3,4,6),ncol=2)

layout(m)

# full-value property-tax rate per $10,000: very different distributions but integer-valued

#better to use barplots but we want the ranges to be the same

start.pos<-min(tax,na.rm=T)

end.pos<-max(tax,na.rm=T)

tax1.freq<-tax2.freq<-rep(0,end.pos-start.pos+1)

for(i in 1:506){

if(!is.na(pt.20[i])){

if(pt.20[i]==1) tax1.freq[tax[i]-(start.pos-1)]<-tax1.freq[tax[i]-(start.pos-1)]+1

if(pt.20[i]==0) tax2.freq[tax[i]-(start.pos-1)]<-tax2.freq[tax[i]-(start.pos-1)]+1

}

}

barplot(tax1.freq,xlab=”Property Tax”,names=seq(start.pos,end.pos),col=3,main=”Tax: PT Ratio >=20”)

barplot(tax2.freq,xlab=”Property Tax”,names=seq(start.pos,end.pos),col=3,main=”Tax: PT Ratio < 20”)

# 1000(bk-0.63)^2 : bk – proportion of blacks by town

#very different distributions best looked at with histograms

#NOTE: very small and very large values of bk both give large values of black

hist(black[pt.20==1],xlab="(Transformed) Proportion of Blacks by Town",main="Blacks: Pt Ratio >=20")

hist(black[pt.20==0],xlab="(Transformed) Proportion of Blacks by Town ",main="Blacks: Pt Ratio < 20")

# lower status of the population (percent)

boxplot(lstat[pt.20==1],lstat[pt.20==0],names=c(“Pt.Ratio >=20”, “Pt.Ratio<20”),main=”Lower Status vs. PT Ratio”)

# median value of owner-occupied homes in $1000

boxplot(medv[pt.20==1],medv[pt.20==0],names=c(“Pt.Ratio >=20”, “Pt.Ratio<20”),main=”Median House Value vs. PT Ratio”)



Aside from the large group of outlier suburbs in the high pupil-teacher ratio group, the property taxes don’t seem that dependent on pupil-teacher ratio (note frequencies).

The distributions of the transformed suburbs’ percentages of blacks do not seem dependent on pupil-teacher ratio. However, this variable could be misleading as both small and large percentages of blacks will lead to similar values of the variable. Suburbs with a higher pupil-teacher ratio have a higher percentage of lower status citizens and a lower median house value.

4) *Write a function plot.by.quartile that takes as arguments:*

1. *a continuous variable to be categorized by quartiles (var1)*
2. *a matrix of continuous variables (should not include var1)*

*Within the function, we want to do the following:*

*First, categorize var1 by quartiles, i.e. create an indicator variable with a 1, 2, 3, or 4 for each observation identifying its quartile. Then for each of the continuous variables in the matrix from the second argument (the columns), create two plots. The first is a scatterplot of var1 vs. the continuous variable with suburbs in different quartiles of var1 labeled with different colors and symbols. Use a legend to identify the symbols you used.*

*The second plot will be four boxplots of the distributions of the continuous variable for each quartile of var1, i.e. splitting the values by the categorized var1. Remember that your boxplots should all be on the same graph space so values can be easily compared across plots. Plots should be labeled/titled appropriately.*

*(So if we pass in crim and a matrix with columns zn and dis, we will have two scatterplots: one of crim vs. zn and one of crim vs. dis. Each observation in the scatterplots is labeled according to its crim quartile. Then we will have four boxplots of the zn values. The first boxplot shows the zn distribution for the suburbs in the first crim quartile and so on. Similarly, we have four boxplots of the dis values, again with the suburbs split by the crim quartiles*.)

*Also, find the mean for all four var1 quartile groups of suburbs for each continuous variable. Return the means in a matrix with 4 rows and as many columns as continuous variables in the matrix.*

*(If we pass in crim and a matrix with columns zn and dis, we return a mean.matrix of size 4 by 2.*

*The first row is the mean zn value and mean dis value for the first crim quartile group of suburbs; the second row is the mean zn value and the mean dis value for the second crim quartile group of suburbs and so on.)*

We have four main tasks in the function:

1. find the quartiles of the first argument variable
2. create scatterplots
3. create boxplots
4. find means

We want the function to be as general as possible – the function starts with several lines of code finding the dimensions of the passed-in arguments as well as the names of the variables. In order to get variable names for plot labels, we pass in var1 as a one-column matrix with the corresponding column name. We have missing values in our data, so we need to double check for NA when categorizing as well as when finding the means.

plot.by.quartile<-function(var1,cont.var.matrix){

##Finding dimensions, names, creating space, etc.

var.name<-colnames(var1)

n<-nrow(var1)

var1<-as.numeric(var1)

col.labels<-colnames(cont.var.matrix)

n.col<-ncol(cont.var.matrix)

mean.matrix<-matrix(NA,4,n.col)

##Categorizing Var1

sum.var1<-summary(var1)

var1.quart<-rep(NA,n)

for(i in 1:n){

#need to check if missing first

if(!is.na(var1[i])){

if(var1[i]<=sum.var1[2]) var1.quart[i]<-1

else if(sum.var1[2]<var1[i] & var1[i]<=sum.var1[3]) var1.quart[i]<-2

else if(sum.var1[3]<var1[i] & var1[i]<=sum.var1[5]) var1.quart[i]<-3

else var1.quart[i]<-4

}

}

par(mfrow=c(2,2))

for(i in 1:n.col){

current.var<-cont.var.matrix[,i]

##The Scatterplots

plot(var1,cont.var.matrix[,i],xlab=var.name,ylab=col.labels[i],type="n")

points(var1[var1.quart==1],current.var[var1.quart==1],col=2,pch=12)

points(var1[var1.quart==2],current.var[var1.quart==2],col=3,pch=13)

points(var1[var1.quart==3],current.var[var1.quart==3],col=4,pch=16)

points(var1[var1.quart==4],current.var[var1.quart==4],col=5,pch=18)

legend(locator(1),legend=c(“Q1”,”Q2”,”Q3”,”Q4”),col=c(2,3,4,5),pch=c(12,13,16,18),cex=.7)

title(paste(var.name,” by “,col.labels[i]))

##The Boxplots

boxplot(current.var[var1.quart==1],current.var[var1.quart==2],current.var[var1.quart==3],current.var[var1.quart==4],names=c(“Q1”,”Q2”,”Q3”,”Q4”),xlab=var.name,ylab=col.labels[i])

title(main=paste(var.name,” by “,col.labels[i]))

##The Means

mean.matrix[1,i]<-mean(current.var[var1.quart==1],na.rm=T)

mean.matrix[2,i]<-mean(current.var[var1.quart==2],na.rm=T)

mean.matrix[3,i]<-mean(current.var[var1.quart==3],na.rm=T)

mean.matrix[4,i]<-mean(current.var[var1.quart==4],na.rm=T)

}

rownames(mean.matrix)<-c(paste(var.name,”Q1”),paste(var.name,”Q2”),paste(var.name,”Q3”),paste(var.name,”Q4”))

colnames(mean.matrix)<-col.labels

return(mean.matrix)

}

a)

crim2<-matrix(crim,ncol=1)

colnames(crim2)<-“crim”

plot.by.quartile(crim2,cbind(zn,dis))

The Means:

zn dis

crim Q1 33.954545 5.610604

crim Q2 9.126016 4.549608

crim Q3 2.533333 2.993749

crim Q4 0.000000 1.991203



Recall that the *zn* variable is largely zeros. It looks, however, that the suburbs with zero residential land zoned for large lots have the highest per capita crime rate. These suburbs are likely urban areas where there is less room for large lots for malls, shopping centers, etc. The suburbs with the highest per capita crime rate are the closest to the five chosen employment centers.

medv2<-matrix(medv,ncol=1)

colnames(medv2)<-“medv”

plot.by.quartile(medv2,cbind(crim,zn,nox,dis,ptratio))

The Means:

crim zn nox dis ptratio

medv Q1 11.1452963 0.4065041 0.6661760 2.288350 19.72720

medv Q2 1.7302255 6.7016129 0.5494177 3.976149 19.21048

medv Q3 0.9899643 12.9173554 0.5120240 4.508082 18.06320

medv Q4 0.8120767 25.2375000 0.4931549 4.324222 16.81818







The crime rate tends to be higher in areas with lower median housing values. There are hardly any suburbs in the first quartile of housing values with any residential land zoned for large lots. Higher median housing values are associated with higher proportions of land zoned for large lots. The pollution/nitrogen oxides concentration appears to have a slight inverse relationship with median housing value. With regards to distance to employment centers, only the first quartile of median housing values stands out as being much closer than the other quartiles. Although there are many suburbs in the lower median housing value groups with low pupil-teacher ratios, it appears that an inverse relationship might exist between median housing value and pupil-teacher ratios.

The scatterplots give you an idea of what range the quartiles have (ex: fourth quartile from 25-55). The colors lead to easy identification. However, if the quartiles are all on top of each other, it can be very difficult to identify any patterns. The boxplots might be preferable in that situation. They nicely separate the distributions of the groups without worrying about overlap. *5) Build a model to predict the median house value using the other variables. Choose your variables based on both statistical significance and confounder status. That is, you may want to include a variable that you believe you should adjust for even though it is not statistically significant. Think about whether or not you should categorize any of your continuous variables. Interpret the coefficient estimates. Report 95% confidence intervals.*

Our continuous variable is *medv*, median housing value. We’ll use linear regression to predict *medv* using the other variables. Since *zn* is largely zero, we’ll use the categorized variable *zn.zero* instead. I’ll also keep the rounded version of the room variable, *rm2*.

Let’s try all variables and see what we have:

full.fit<-lm(medv~crim+zn.zero+indus+chas+nox+rm2+age+dis+rad+tax+ptratio+black+lstat)

summary(full.fit)

Call:

lm(formula = medv ~ crim + zn.zero + indus + chas + nox + rm2 +

age + dis + rad + tax + ptratio + black + lstat)

Residuals:

Min 1Q Median 3Q Max

-17.4250 -2.9798 -0.6621 1.9033 24.8107

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 39.409987 5.664280 6.958 1.37e-11 \*\*\*

crim -0.102172 0.034668 -2.947 0.00339 \*\*

zn.zero -1.062861 0.853583 -1.245 0.21378

indus 0.045757 0.068019 0.673 0.50151

chas 2.970412 0.998607 2.975 0.00311 \*\*

nox -18.482167 4.374532 -4.225 2.95e-05 \*\*\*

rm2 3.603989 0.438062 8.227 2.56e-15 \*\*\*

age 0.002651 0.014980 0.177 0.85964

dis -1.350547 0.230540 -5.858 9.60e-09 \*\*\*

rad 0.303817 0.072600 4.185 3.49e-05 \*\*\*

tax -0.010171 0.004044 -2.515 0.01227 \*

ptratio -1.026673 0.157387 -6.523 2.03e-10 \*\*\*

black 0.011619 0.002957 3.929 0.00010 \*\*\*

lstat -0.600609 0.055913 -10.742 < 2e-16 \*\*\*

---

Signif. codes: 0 `\*\*\*' 0.001 `\*\*' 0.01 `\*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 4.964 on 411 degrees of freedom

Multiple R-Squared: 0.7423, Adjusted R-squared: 0.7341

F-statistic: 91.06 on 13 and 411 DF, p-value: < 2.2e-16

Almost all variables are significant predictors. Let’s try stepwise regression.

Note: lm( ) removes all observations with one or more missing values. If a variable is removed from the model, the observations included in the model might change (because we will get back any observations that were only missing values in the removed variable) and then the AIC comparisons aren’t valid. The AIC’s from two models can only be compared if they were run on the same data.

To use step( ) fairly, we need to remove all observations with any missing data ahead of time. Recall the sub.miss variable we created on page 2.

data2<-final[-sub.miss,]

dim(data2)

[1] 425 14

Attaching our created variables:

zn.zero2<-zn.zero[-sub.miss]

rm2.2<-rm2[-sub.miss]

data2<-cbind(data2,zn.zero2,rm2.2)

full.fit.no.miss<-lm(medv~crim+zn.zero2+indus+chas+nox+rm2.2+age+dis+rad+tax+ptratio+black+lstat,data=data2)

summary(full.fit.no.miss)

fit.step<-step(full.fit.no.miss)

(Edited results)

Call:

lm(formula = medv ~ crim + chas + nox + rm2.2 + dis + rad + tax + ptratio + black + lstat, data = data2)

Coefficients:

(Intercept) crim chas nox rm2.2 dis

38.92010 -0.10167 2.93289 -17.84651 3.65645 -1.27549

rad tax ptratio black lstat

0.30027 -0.00917 -1.08669 0.01178 -0.59089

This model did not exclude any variables that I feel should be included regardless of significance. Let’s use this as our final model.

coef<-summary(fit.step)$coef

est<-coef[,1]

lower.ci<-coef[,1]-1.96\*coef[,2]

upper.ci<-coef[,1]+1.96\*coef[,2]

res<-cbind(est,lower.ci,upper.ci)

colnames(res)<-c("Est","Lower CI","Upper CI")

res

Est Lower CI Upper CI

(Intercept) 38.920096461 27.879638244 49.960554678

crim -0.101671456 -0.169400554 -0.033942357

chas 2.932889361 1.001268194 4.864510527

nox -17.846512701 -25.782629829 -9.910395573

rm2.2 3.656454136 2.822345198 4.490563074

dis -1.275487882 -1.634862769 -0.916112995

rad 0.300274459 0.165162178 0.435386741

tax -0.009169684 -0.016326899 -0.002012469

ptratio -1.086691509 -1.359990001 -0.813393017

black 0.011776693 0.006011815 0.017541571

lstat -0.590889403 -0.695508386 -0.486270420

Recall that med-hv is in units of $1000. A one-unit increase in the crime rate decreases the med-hv by $101. Bounding the Charles River increases the med-hv by $2,930. A one-unit increase in nitrogen oxide concentration decreases the med-hv by $17, 847. One extra room increase the med-hv by $3,656. A one-unit increase in the distance to employment centers decreases the med-hv by $1,275. A one-unit increase in the radial highway accessibility index increases the med-hv by $300. A $10,000 increase in the full-value property tax rate decreases the med-hv by $9. Increasing the pupil-teacher ratio by one decreases the med-hv by $1,086. A one-unit increase in the transformed percentage of blacks increases the med-hv by $11. A one percent increase in the lower status of the population decreases the med-hv by $591. Note that some of these variables are statistically significant but not that significant in real-life (changes of $9, $11, etc).

*6) Build a model to predict whether or not a suburb has a median house value of $30,000 or higher. Choose your variables based on both statistical significance and confounder status. That is, you may want to include a variable that you believe you should adjust for even though it is not statistically significant. Think about whether or not you should categorize any of your continuous variables (be wary of reference group choices). Find and interpret the odds ratios. Report 95% confidence intervals.*

We’ll use logistic regression to model the probability of having a median house value of $30,000 or higher. First, we need to create a binary response variable.

med.30<-ifelse(medv>=30,1,0)

Again, we’ll use the dataset with all observations with missing values removed.

med.30.2<-med.30[-sub.miss]

data2<-cbind(data2,med.30.2)

full.log<-glm(med.30.2~crim+zn.zero2+indus+chas+nox+rm2.2+age+dis+rad+tax+ptratio+black+lstat,data=data2,family=binomial)

summary(full.log)

Call:

glm(formula = med.30.2 ~ crim + zn.zero2 + indus + chas + nox +

rm2.2 + age + dis + rad + tax + ptratio + black + lstat,

family = binomial, data = data2)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.533996 -0.097647 -0.015610 -0.000534 2.396052

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 6.284896 7.280266 0.863 0.387984

crim 0.063450 0.043896 1.445 0.148327

zn.zero2 -0.395549 0.744290 -0.531 0.595110

indus -0.197991 0.089114 -2.222 0.026299 \*

chas 0.338821 0.926603 0.366 0.714619

nox -0.692349 6.163153 -0.112 0.910556

rm2.2 1.211241 0.471046 2.571 0.010129 \*

age 0.029555 0.017412 1.697 0.089635 .

dis -0.186024 0.182634 -1.019 0.308411

rad 0.344858 0.114390 3.015 0.002572 \*\*

tax -0.008744 0.004974 -1.758 0.078741 .

ptratio -0.644684 0.189033 -3.410 0.000649 \*\*\*

black 0.006218 0.010890 0.571 0.568026

lstat -0.724122 0.143712 -5.039 4.69e-07 \*\*\*

---

Signif. codes: 0 `\*\*\*' 0.001 `\*\*' 0.01 `\*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 383.51 on 424 degrees of freedom

Residual deviance: 111.51 on 411 degrees of freedom

AIC: 139.51

Number of Fisher Scoring iterations: 9

Far fewer variables are significant in predicting this binary outcome.

(Note: the variables would likely change if we changed our cutoff in the binary value.)

Let’s see how step handles it:

log.step<-step(full.log)

summary(log.step)

Call:

glm(formula = med.30.2 ~ indus + rm2.2 + age + rad + tax + ptratio +

lstat, family = binomial, data = data2)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.6598311 -0.1060614 -0.0202495 -0.0006116 2.4587546

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 6.661475 5.009977 1.330 0.183636

indus -0.192705 0.075251 -2.561 0.010442 \*

rm2.2 1.320509 0.455333 2.900 0.003730 \*\*

age 0.034441 0.011785 2.923 0.003472 \*\*

rad 0.384939 0.107614 3.577 0.000348 \*\*\*

tax -0.009623 0.004707 -2.044 0.040913 \*

ptratio -0.670814 0.163010 -4.115 3.87e-05 \*\*\*

lstat -0.696588 0.133453 -5.220 1.79e-07 \*\*\*

---

Signif. codes: 0 `\*\*\*' 0.001 `\*\*' 0.01 `\*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 383.51 on 424 degrees of freedom

Residual deviance: 113.82 on 417 degrees of freedom

AIC: 129.82

Number of Fisher Scoring iterations: 9

Finding Odds Ratios and 95% CIs:

or<-round(exp(log.coef[,1]),3)

lower.ci<-round(exp(log.coef[,1]-1.96\*log.coef[,2]),3)

upper.ci<-round(exp(log.coef[,1]+1.96\*log.coef[,2]),3)

res<-cbind(or,lower.ci,upper.ci)

colnames(res)<-c("OR","Lower.CI","Upper.CI")

res

OR Lower.CI Upper.CI

(Intercept) 781.703 0.043 14375394.674

indus 0.825 0.712 0.956

rm2.2 3.745 1.534 9.143

age 1.035 1.011 1.059

rad 1.470 1.190 1.815

tax 0.990 0.981 1.000

ptratio 0.511 0.371 0.704

lstat 0.498 0.384 0.647

Each unit increase in proportion of non-retail business acres per town decreases the odds of a med-hv >= $30,000 by 18%. Each extra room increases these odds by 275%. Increasing the proportion of owner-occupied units built prior to 1940 by one percent increases the odds by 3.5%; increasing the radial highway accessibility index by one unit increases the odds by 47%. Increasing the full-value property-tax rate by $10,000 decreases the odds of having a med-hv >= $30,000 by 1%. Increasing the pupil-teacher ratio by one decreases the odds by 49%; a one percent increase in the lower status of the population decreases the odds by 51%.