Issues with Multiple Linear Regression

October 11, 2017

Looking at the data:

```
> corr.test(cystfibr, use = "pairwise",  method="pearson",  adjust="none",alpha=.05)
Call:corr.test(x = cystfibr, use = "pairwise", method = "pearson",
adjust = "none", alpha = 0.05)
Correlation matrix
age   sex height weight   bmp  fev1    rv   frc   tlc pemax
age     1.00 -0.17   0.93   0.91  0.38  0.29 -0.55 -0.64 -0.47  0.61
sex    -0.17  1.00  -0.17  -0.19 -0.14 -0.53  0.27  0.18  0.02 -0.29
height  0.93 -0.17   1.00   0.92  0.44  0.32 -0.57 -0.62 -0.46  0.60
weight  0.91 -0.19   0.92   1.00  0.67  0.45 -0.62 -0.62 -0.42  0.64
bmp     0.38 -0.14   0.44   0.67  1.00  0.55 -0.58 -0.43 -0.36  0.23
fev1    0.29 -0.53   0.32   0.45  0.55  1.00 -0.67 -0.67 -0.44  0.45
rv     -0.55  0.27  -0.57  -0.62 -0.58 -0.67  1.00  0.91  0.59 -0.32
frc    -0.64  0.18  -0.62  -0.62 -0.43 -0.67  0.91  1.00  0.70 -0.42
tlc    -0.47  0.02  -0.46  -0.42 -0.36 -0.44  0.59  0.70  1.00 -0.18
pemax   0.61 -0.29   0.60   0.64  0.23  0.45 -0.32 -0.42 -0.18  1.00
Sample Size
[1] 25
Probability values (Entries above the diagonal are adjusted for multiple tests.)
age  sex height weight  bmp fev1   rv  frc  tlc pemax
age    0.00 0.42   0.00   0.00 0.06 0.15 0.00 0.00 0.02  0.00
sex    0.42 0.00   0.42   0.36 0.51 0.01 0.19 0.38 0.91  0.16
height 0.00 0.42   0.00   0.00 0.03 0.12 0.00 0.00 0.02  0.00
weight 0.00 0.36   0.00   0.00 0.00 0.02 0.00 0.00 0.04  0.00
bmp    0.06 0.51   0.03   0.00 0.00 0.00 0.00 0.03 0.07  0.27
fev1   0.15 0.01   0.12   0.02 0.00 0.00 0.00 0.00 0.03  0.02
rv     0.00 0.19   0.00   0.00 0.00 0.00 0.00 0.00 0.00  0.12
frc    0.00 0.38   0.00   0.00 0.03 0.00 0.00 0.00 0.00  0.04
tlc    0.02 0.91   0.02   0.04 0.07 0.03 0.00 0.00 0.00  0.38
pemax  0.00 0.16   0.00   0.00 0.27 0.02 0.12 0.04 0.38  0.00
```

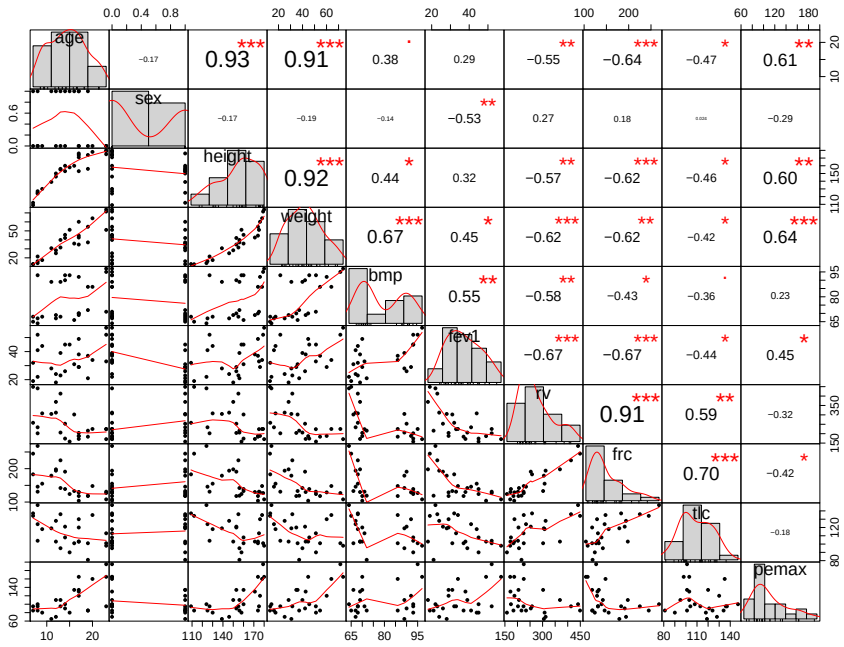

Figure 1: correlation matrix

Assume the following model is selected with minimum AIC

```
mp=lm(pemax ~ weight + bmp + fev1 + rv)
summary(mp)
```

```
Call:
lm(formula = pemax ~ weight + bmp + fev1 + rv)

Residuals:
Min      1Q Median      3Q     Max
-39.77 -11.74   4.33  15.66  35.07

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.94669   53.27673   1.200 0.244057
weight       1.74891    0.38063   4.595 0.000175 ***
bmp         -1.37724    0.56534  -2.436 0.024322 *
fev1         1.54770    0.57761   2.679 0.014410 *
rv           0.12572    0.08315   1.512 0.146178
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 22.75 on 20 degrees of freedom
Multiple R-squared:  0.6141,Adjusted R-squared:  0.5369
F-statistic: 7.957 on 4 and 20 DF,  p-value: 0.000523
```

Plotting the predicted vs responses

```
> predy = predict(mp)
> plot(predy ~ pemax,data=cystfibr,pch = 16,xlab="Actual response value",ylab="Predicted respo
> abline(0,1, col="blue", lwd=2)
```

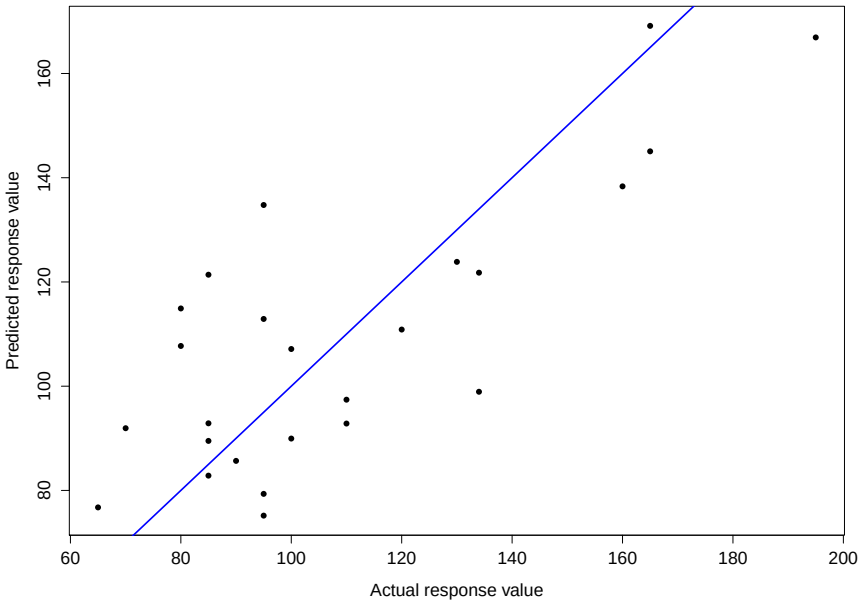

Figure 2: Predicted vs Responses

Checking outliers

```
outlierTest(mp)            # Bonferonni p-value for most extreme obs
qqPlot(mp, main="QQ Plot")       #qq plot for studentized resid
leveragePlots(mp)               # leverage plots
```

Leverages for each observation

```
> qqPlot(mp, main="QQ Plot") #qq plot for studentized resid
> leveragePlots(mp) # leverage plots
> lev = hat(model.matrix(mp))
> plot(lev)
> mtcars[lev >0.4,]
mpg cyl disp  hp drat   wt  qsec vs am gear carb
Hornet Sportabout 18.7   8  360 175 3.15 3.44 17.02  0  0    3    2
Ferrari Dino      19.7   6  145 175 3.62 2.77 15.50  0  1    5    6
```

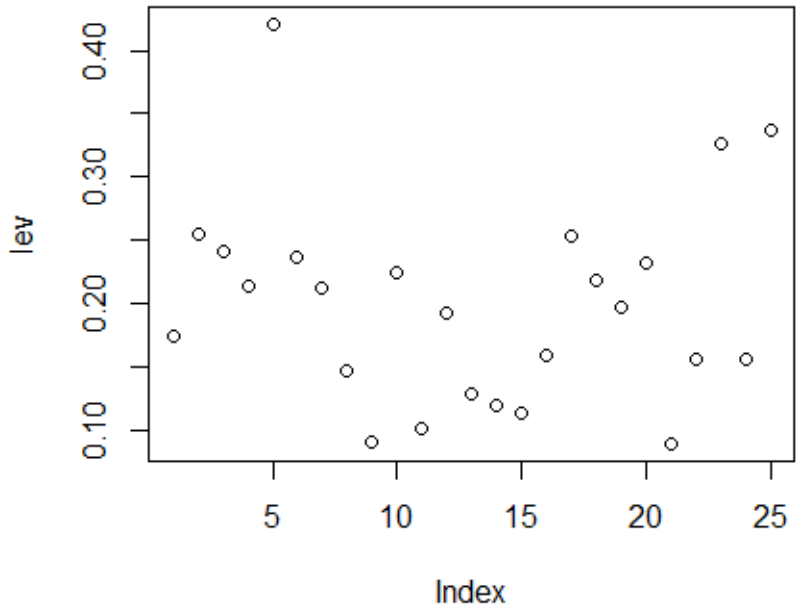

Figure 3: leverage on each obs.

Normality

```
qq.plot(mp, simulate=TRUE, line="none")
plot(density(rstudent(mp)))
```

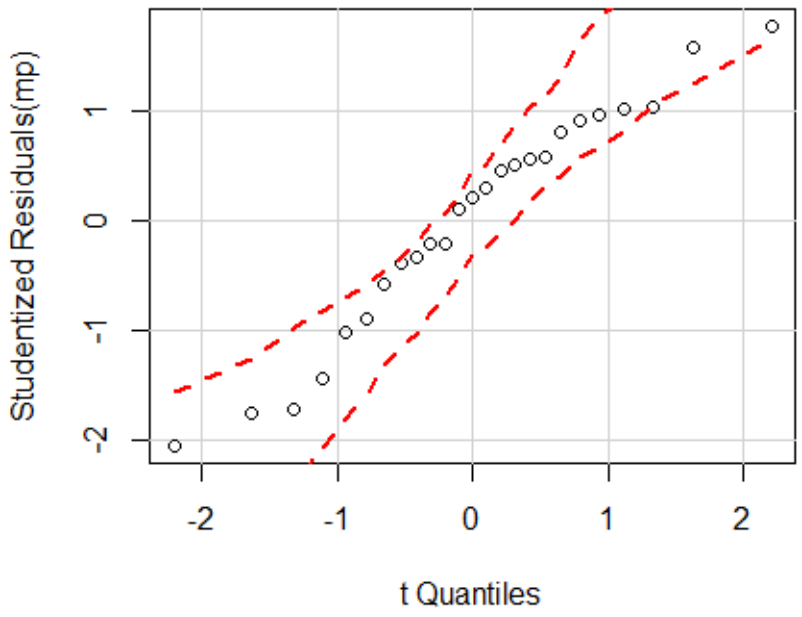

Figure 4

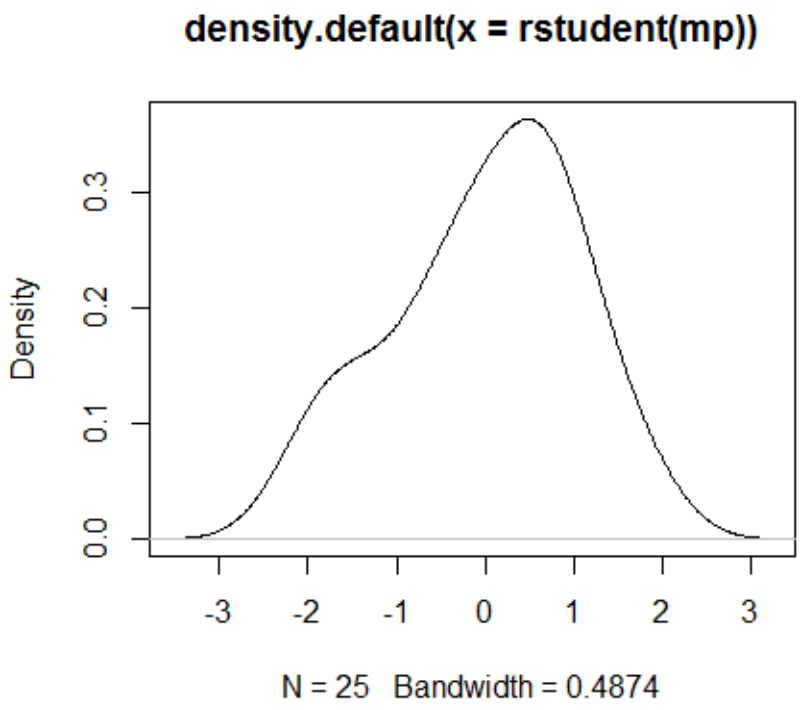

Figure 5

Constant variances

```
var.test(m1,m2) #When there are more than two groups, Bartlett's test can be used.

plot(fitted(mp), rstudent(mp), col="gray")
abline(h=0, lty=2)
lines(lowess(fitted(mp), rstudent(mp)))

> spread.level.plot(mp)

Suggested power transformation:  0.004919824
Warning message:
'spread.level.plot' is deprecated.
Use 'spreadLevelPlot' instead.
See help("Deprecated") and help("car-deprecated")

ncvTest(mp)  ###### another ways of testing constant variance
```

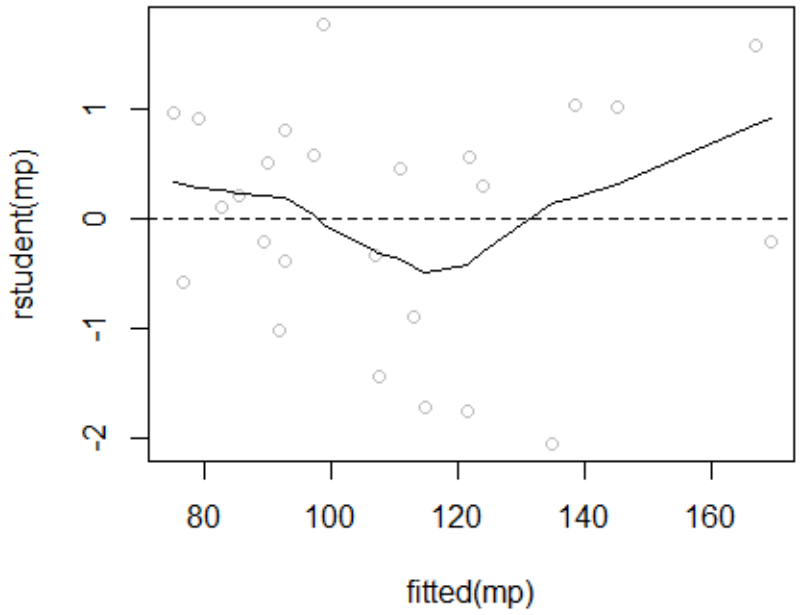

Figure 6

```
Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(cystfibr)-length(mp$coefficients)-2))
plot(mp, which=4, cook.levels=cutoff)
# Influence Plot
influencePlot(mp,id.method="identify", main="Influence Plot",
sub="Circle size is proportial to Cook's Distance" )
```

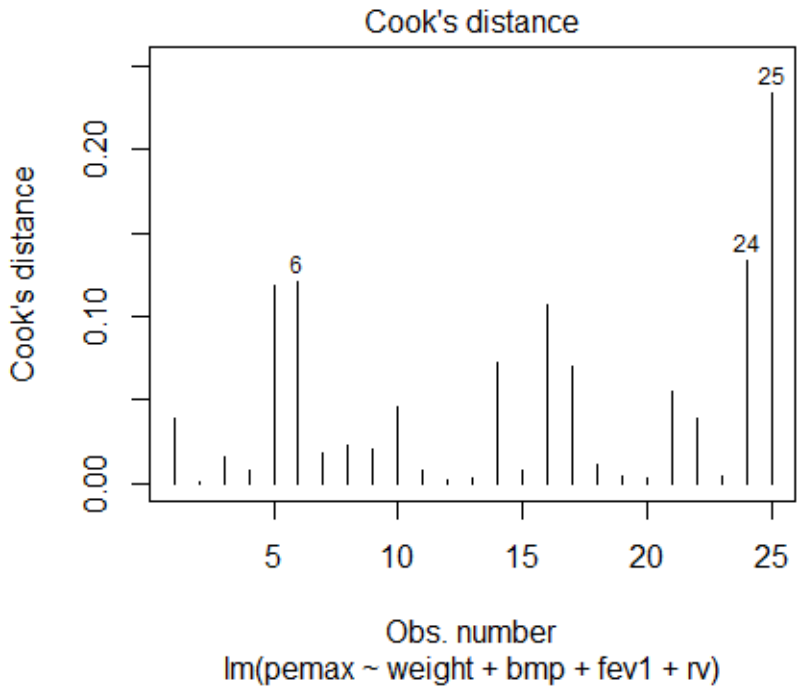
Cook's distance
Cook's distance
0.00
0.10
0.20
6
24
25
5
10
15
20
25
Obs. number
lm(pemax ~ weight + bmp + fev1 + rv)

Figure 7

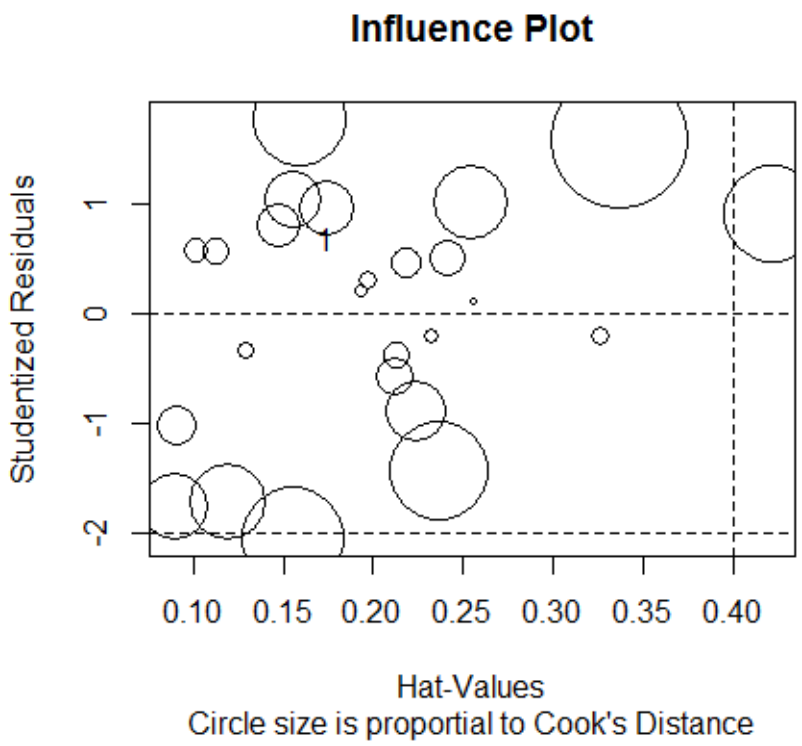

Figure 8

Issue with multicollinearity

Highly correlated explanatory variables, multicollinearity, can cause several problems when applying the multiple regression model, including:

1. It severely limits the size of the multiple correlation coe cient R because the explanatory variables are primarily attempting to explain much of the same variability in the response variable (see Dizney and Gromen [1967] for an example).

2. It makes determining the importance of a given explanatory variable (see later) di cult because the e ects of explanatory variables are confounded due to their intercorrelations.

3. It increases the variances of the regression coe cients, making use of the predicted model for prediction less stable. The parameter estimates become unreliable.

Spotting multicollinearity

amongst a set of explanatory variables might not be easy. The obvious course of action is to simply examine the correlations between these variables, but whilst this is often helpful, it is by no means foolproof more subtle forms of multicollinearity may be missed. An alternative and generally far more useful approach is to examine what are known as the variance in ation factors of the explanatory variables. The variance in ation factor VIFj for the jth variable is given by

$$VIF_j = 1/1 - R_j^2$$

where R_j^2 is the square of the multiple correlation coe cient from the regression of the j^{th} explanatory variable on the remaining explanatory variables. The *variance inflation factor* of an explanatory variable indicates the strength of the linear relationship between the variable and the remaining explanatory variables. A rough rule of thumb is that *variance inflation factors greater than 10* give some cause for concern.

```
# plot studentized residuals vs. fitted values
spreadLevelPlot(mp)
################ Evaluate Collinearity
vif(mp) # variance inflation factors
sqrt(vif(mp)) > 3 # problem?
```

how to fix the issue

- One way is to combine in some way explanatory variables that are highly correlated.
- An alternative is simply to select one of the set of correlated variables.
- Two more complex possibilities are regression on *principal components* and *ridge regression*.

Having arrived at a nal multiple regression model for a data set, it is important to go further and check the assumptions made in the modelling process.

The most useful ways of examining the residuals are graphical, and the most useful plots are

- A plot of the residuals against each explanatory variable in the model; the presence of a curvilinear relationship, for example, would suggest that a higher-order term (e.g., a quadratic) in the explanatory variable is needed in the model.
- A plot of the residuals against predicted values of the response variable; if the variance of the response appears to increase with the predicted value, a transformation of the response may be in order.
- A normal probability plot of the residuals; after all systematic variation has been removed from the data, the residuals should look like a sample from the normal distribution. A plot of the ordered residuals against the expected order statistics from a normal distribution provides a graphical check of this assumption.

Unfortunately, the simple observed- tted residuals have a distribution that is scale dependent which makes them less helpful than they might be. The problem can be overcome, however,
by using *standardised* or *studentised residuals*.
A variety of other diagnostics for regression models have been developed in the past decade or so. One that is often used is the Cooks distance statistic. This statistic can be obtained for each of the n observations and measures the change to the estimates of the regression coe cients that would result from deleting the particular observation. It can be used to identify any observations having an undue in uence of the estimation and tting process.

Rcode

```
library(ISwR)
?cystfibr              ##########################Data in IsWR package
```

```
View(cystfibr)

attach(cystfibr)    #####In order to be able to refer directly to the variables
par(mex=0.5)
pairs(cystfibr, gap=0, cex.labels=0.9)
################################################## more tools
library(psych)
corr.test(cystfibr, use = "pairwise",  method="pearson",  adjust="none",alpha=.05)
# Can adjust p-values; see ?p.adjust for options
library(PerformanceAnalytics)
chart.Correlation(cystfibr, method="pearson",histogram=TRUE,pch=16)

##################################################fitting a regression line
lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc))
anova(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc))
########################################################## comparing two models
m1<-lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
m2<-lm(pemax~age)
anova(m1,m2)           ########################################order is important!!!
anova(m2,m1)


##################################################### Model selection automated
S1=step(m1, data=cystfibr, direction="backward")
S2=step(m1, data=cystfibr, direction="forward")
S3=step(m1, data=cystfibr, direction="both")
S2$anova # display results
S1$anova
############################################## Stepwise Regression
library(MASS)
step <- stepAIC(m1, direction="both")
step$anova # display results

##########################################################model selection manually
#Model fit criteria are available to decide which model is most appropriate.
#The step function uses AIC, or optionally BIC, but there are others.
#You dont want to use multiple R-squared, because it will continue to improve
#as more terms are added into the model.
#Instead, you want to use a criterion that balances the improvement in explanatory power with r
#extraneous terms to the model.  Adjusted R-squared is a modification of R-squared that include
#Larger is better.  AIC is based on information theory and measures this balance.
#AICc is an adjustment to AIC that is more appropriate for data sets with relatively fewer obse
#BIC is similar to AIC, but penalizes more for additional terms in the model.
#Smaller is better for AIC, AICc, and BIC.
#There are differing opinions on which model fitting criteria is best to use,
#but if you have no opinion, I would recommend AICc for routine use.
#Using the step procedure to automatically find an optimal model is an option, but some people
```

```
#using an automated procedure because it might not hone in on the best model.
#Instead, you can look at the model fit criteria for competing models manually.
#There may be reasons why you wish to include or exclude some terms in the model, and it may be
#to look at different model selection criteria simultaneously.
summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc))
summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv))
summary(lm(pemax~age+sex+height+weight+bmp+fev1))
summary(lm(pemax~age+sex+height+weight+bmp))
summary(lm(pemax~age+height+weight+bmp))
summary(lm(pemax~height+weight+bmp))
summary(lm(pemax~weight+bmp))
summary(lm(pemax~weight))
##################################################Assume the following model is selected with m
mp=lm(pemax ~ weight + bmp + fev1 + rv)
summary(mp)
#######################################Simple plot of predicted values with 1-to-1 line

predy = predict(mp)
plot(predy ~ pemax,data=cystfibr,pch = 16,xlab="Actual response value",ylab="Predicted response
abline(0,1, col="blue", lwd=2)

###################################
#############################################################################checking outliears--- I
library(car)
lev = hat(model.matrix(mp))
plot(lev)
cystfibr[lev>.45,]###############################
influence.measures(mp)
influencePlot(mp)
leveragePlots(mp) # leverage plots
?leveragePlots
outlierTest(mp) # Bonferonni p-value for most extreme obs
qqPlot(mp, main="QQ Plot") #qq plot for studentized resid
######################################################################## checking normality
qq.plot(mp, simulate=TRUE, line="none")
plot(density(rstudent(mp)))
######################################################## checking constant error variance
par(mfrow=c(2,2))
plot(mp)
par(mfrow=c(1,1))
plot(fitted(mp), rstudent(mp), col="gray")
abline(h=0, lty=2)
lines(lowess(fitted(mp), rstudent(mp)))
######################################################## checking linearity

# distribution of studentized residuals
library(MASS)
sresid <- studres(mp)
```

```
hist(sresid, freq=FALSE,
main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)

############################ Evaluate homoscedasticity
# non-constant error variance test
ncvTest(mp)
# plot studentized residuals vs. fitted values
spreadLevelPlot(mp)
############### Evaluate Collinearity
vif(mp) # variance inflation factors
sqrt(vif(mp)) > 3 # problem?
################################################ Evaluate Nonlinearity

################## Test for Autocorrelated Errors
durbinWatsonTest(mp)

###################################
# Global test of model assumptions
library(gvlma)
gvmodel <- gvlma(mp)
summary(gvmodel)
```