Diagnostics

- **Toluca Example (Page 19).** The Toluca Company manufactures refrigeration equipment as well as many replacement parts. In the past, one of the replacement parts has been produced periodically in lots of varying sizes. When a cost improvement program was undertaken, company officials wished to determine the optimum lot size for producing this part. One key input for the model to ascertain the optimum lot size was the relationship between lot size and work hours required to produce the lot. To determine this relationship, data on lot size and work hours for 25 recent production runs were obtained. This data are saved in *Toluca.csv*.

- R commands

```
rm(list=ls())           # Clears the workspace

data=read.csv("Toluca.csv",header=T)
attach(data)

##  Diagnostics for Predictor Variables ##

table(size)     # Creates a frequency table
barplot(table(data$size)) # Creates a frequency bargraph

boxplot(size) # Creates a boxplot (good for outlier detection)
size.out=c(size,200)
boxplot(size.out) # Note the presence of the outlier

plot(size,type="b") # Creates a time/sequence plot

hist(size) # Creates a histogram
bin=seq(15,125,by=10)
hist(size,breaks=bin,col="gray")

stem(size) # Creates a stem-and-leaf-plot
stem(size,scale=3)

# Dot plots
stripchart(size, method = "stack", offset = 1, at = 0, pch = 19,
           main = "Dotplot", xlab = "Lot Size")

plot(size, hours)
results=lm(hours~size)
abline(results,col="blue")

## Departures from the model ##

## 1. The regression function is not linear.
# Plot residuals againsts predictor OR residuals againsts fitted values
plot(size,results$residuals) # Note that the residual plot show no systematic pattern.
plot(results$fitted,results$residuals)  # Again, no systematic pattern.

mse=anova(results)$Mean[2] # use=2384
e.star=results$residuals/sqrt(mse)
plot(results$fitted,e.star)

## A non-linear example
z=rnorm(length(size),0,100)
x=size
y=x^2+z
```

```
plot(x,y) # Note that the relationship is curvilinear
cor(x,y) # Note that r is pretty high!
temp=lm(y~x)
plot(temp$fitted,temp$res) # Note the pattern in this residual plot

x2=x^2
temp2=lm(y~x2)
plot(temp2$fitted,temp2$res)

## 2. The error terms are not normally distributed.
# Construct a boxplot or histogram of residuals.
# Better to use Normal probability plots (QQ-plots) of residuals
qqnorm(results$residuals)
qqline(results$residuals)
shapiro.test(results$residuals)

qqnorm(temp$residuals); qqline(temp$residuals)
shapiro.test(temp$residuals)

## 3. The error terms do not have constant variance.
# Plot residuals against predictor variable
# Or plot the absolute residuals against predictor variable

z=rnorm(length(size),0,100*sqrt(size))
x=size
y=100+25*x+z
plot(x,y)   # Note the increasing variation

temp3=lm(y~x)
plot(temp3$fitted,temp3$res) # Note the increasing variation

## 4. The error terms are not independent.
# Sequence plot of the residuals
# Plot of residuals againsts time or other sequence

## 5. The model fits all but one or a few outlier observations.
# Plot residuals againsts predictor OR residuals againsts fitted values
# Box plots, stemplots, dot plots of residuals

## 6. One or several important predictor variables have been omitted from the model.

## Brown-Forsythe Test

bf=function(x,y,x.med=median(x)){
results=lm(y~x)
e=results$residuals
# x.med=median(x)

e1=e[x<=x.med]; n1=length(e1)
e2=e[x>x.med]; n2=length(e2)

d1=abs(e1-median(e1))
d2=abs(e2-median(e2))

s2=(sum((d1-mean(d1))^2)+sum((d2-mean(d2))^2))/(n1+n2-2)

t.bf=(mean(d1)-mean(d2))/(sqrt(s2)*sqrt(1/n1+1/n2))
p.val=2*pt(abs(t.bf),df=(n1+n2-2),lower.tail=F)

list(t.bf=t.bf,p.val=p.val)}
```

```
# Example: Toluca data

# Source in the function bp.R
bf(size,hours) # t_obs=1.3165 and p-value=0.201

# Example: Simulated data with increasing variance
x=sample(seq(20,120,by=10),100,replace=T)
z=rnorm(length(x),0,200*sqrt(x))
y=(62.37+3.57*x)+z
plot(x,y) # Note the increasing variation

bf(x,y)

## Breusch-Pagan Test (Also known as Cook-Weisberg test)

# load the package 'lmtest'
library(lmtest)
bptest(hours~size,studentize=F)

Example 2:
x=1:50
y=(1:50)*rnorm(50)
bf(x,y)
bptest(y~x)

temp4=lm(y~x)
plot(temp4$fit,temp4$res)
```