R language and data analysis: diagnostics of linear model

Qiang Shen

Jan. 9, 2018

Symbols

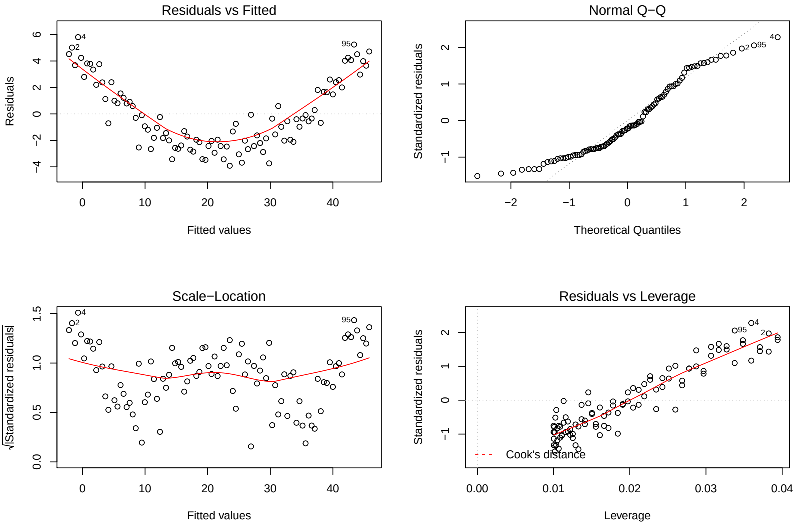
Symbol	Usage
`~`	Separates response variables on the left from the explanatory variables on the right. For example, a prediction of `y` from `x`, `z`, and `w` would be coded `y ~ x + z + w`.
`+`	Separates predictor variables.
`:`	Denotes an interaction between predictor variables. A prediction of `y` from `x`, `z`, and the interaction between `x` and `z` would be coded `y ~ x + z + x:z`.
`*`	A shortcut for denoting all possible interactions. The code `y ~ x * z * w` expands to `y ~ x + z + w + x:z + x:w + z:w + x:z:w`.
`^`	Denotes interactions up to a specified degree. The code `y ~ (x + z + w)^2` expands to `y ~ x + z + w + x:z + x:w + z:w`.
`.`	A placeholder for all other variables in the data frame except the dependent variable. For example, if a data frame contained the variables `x, y, z, and w`, then the code `y ~ .` would expand to `y ~ x + z + w`.
`-`	A minus sign removes a variable from the equation. For example, `y ~ (x + z + w)^2 - x:w` expands to `y ~ x + z + w + x:z + z:w`.
`-1`	Suppresses the intercept. For example, the formula `y ~ x -1` fits a regression of `y` on `x`, and forces the line through the origin at `x=0`.

Figure 1:

Classical linear regression

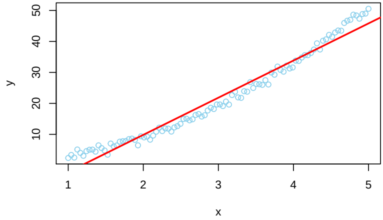
- Linearity: $Y = X\beta_0 + \epsilon$
- Full rank: $rank(X) = K$
- Exogeneity: $E(\epsilon|X) = 0$
- Spherical disturbance: $E(\epsilon\epsilon'|X) = \sigma^2 I_n$
- Normality: $\epsilon \sim N(0, \sigma^2 I_n)$

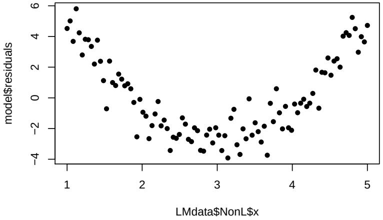
Diagnostics plot in R.

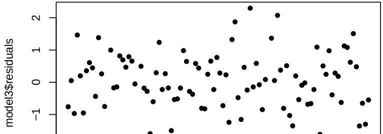


Linearity

```
##          df      AIC
## model    3 478.4558
## model2   4 269.2121
## model3   3 267.2736
```





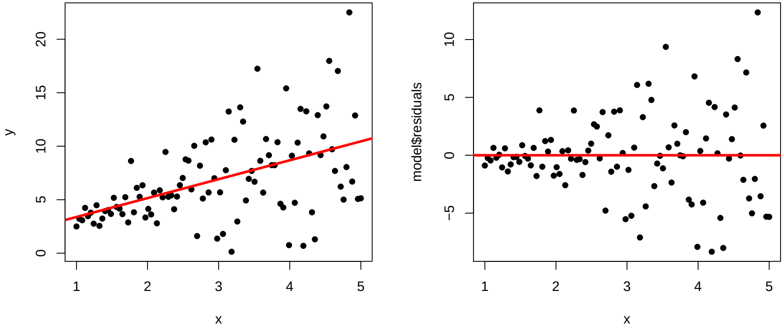


Multicollinearity

```r
data(LMdata,package='rinds')
model <-lm(y~x1+x2+x3,data=LMdata$Mult)
summary(model)$coefficients
library(car);vif(model)#variance inflation factor
model1<-step(model)
model1
summary(model1)$coe
```

Heteroskedasticity

```
par(mfrow=c(1,2))
model<-lm(y~x,data=LMdata$Hetero)
plot(y~x,data=LMdata$Hetero,pch=16)
abline(model,col='red',lwd=3)
with(LMdata$Hetero,plot(x,model$residuals,pch=16))
abline(h=0,,col='red',lwd=3)
```



Standard error

```
library(foreign)
children<- read.dta("fertil2.dta")
r1 <- lm(form <- ceb ~ age + agefbrth + usemeth,
         data=children)
summary(r1)
```

Standard error

$$var(\hat{\beta}) = \sigma_{\mu}^{2}(X'X)^{-1}$$

```r
library(foreign)
children<- read.dta("fertil2.dta")
r1 <- lm(ceb ~ age + agefbrth + usemeth,
         data=children)
X <- model.matrix(r1)
n <- dim(X)[1]
k <- dim(X)[2]
se <- sqrt(diag(solve(crossprod(X)) *
as.numeric(crossprod(resid(r1))/(n-k))))
se
```

```
## (Intercept)         age    agefbrth     usemeth
## 0.173782844 0.003448024 0.008795350 0.055429804
```

Robust standard error

$$(X'X)^{-1}X'\Sigma_\mu X(X'X)^{-1}$$

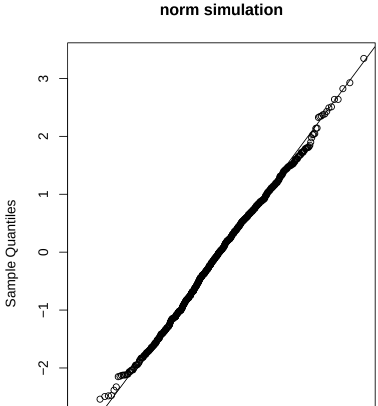
```
library(foreign)
children<- read.dta("data/fertil2.dta")
r1 <- lm(ceb ~ age + agefbrth + usemeth,
         data=children)
u <- matrix(resid(r1))
meat1 <- t(X) %*% diag(diag(crossprod(t(u)))) %*% X
dfc <- n/(n-k)
se <- sqrt(dfc*diag(solve(crossprod(X)) %*%
        meat1 %*% solve(crossprod(X))))
se
```

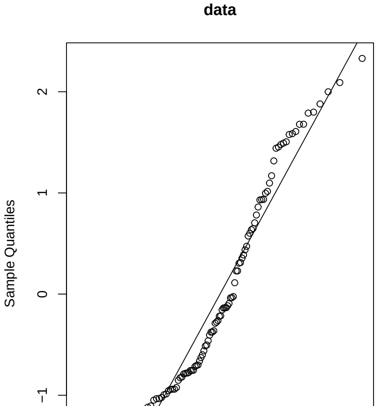
Robust standard error

```
library(foreign)
library(sandwich)
library(lmtest)
children<- read.dta("data/fertil2.dta")
model = lm( ceb ~ age + agefbrth + usemeth,data=children)
summary(model)
coeftest(model, vcov = vcovHC(model, "HC1"))#vs. Stata. ##
##https://cran.r-project.org/web/packages/sandwich/vignette
```

Normality

```
##
##  Shapiro-Wilk normality test
##
## data:  res1
## W = 0.93524, p-value = 1e-04
```





Autocorrelation

```r
data(LMdata,package='rinds')
model<-lm(y~x,data=LMdata$AC)
suppressMessages(library(lmtest))
dwtest(model)##Durbin-Watson test
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 0.65556, p-value = 2.683e-12
## alternative hypothesis: true autocorrelation is greater
```

Clustered standard error

```r
source('ols.r')
ols(ceb ~ age + agefbrth + usemeth,children)
ols(ceb ~ age + agefbrth + usemeth,children,robust=T)
ols(ceb ~ age + agefbrth + usemeth,children,
    cluster="children")
```