Chapter 4 Correlation and regression analysis with R

Jixiang Wu, Associate Professor of Quantitative Genetics/Biostatistics

March 14, 2017

```
## Loading required package: MASS

## Loading required package: leaps

## Loading required package: DAAG

## Loading required package: lattice

##
## Attaching package: 'DAAG'

## The following object is masked from 'package:MASS':
##
##     hills
```

Load a data set

We will use the following R codes to load a data set, cotyldreg, from the package coursedata.

```
require(coursedata)

## Loading required package: coursedata

data(cotyldreg)
cot=cotyldreg
names(cot)

## [1] "LP" "BW" "BN" "LS" "SB" "LY"

head(cot)
```

LP	BW	BN	LS	SB	LY
41.4	5.08	634	74.3	28.3	1333
38.6	5.49	577	68.9	30.7	1221
38.0	6.21	533	66.9	35.3	1256
38.2	5.27	599	69.3	29.0	1204
40.6	5.48	748	67.4	33.0	1664
40.3	6.23	561	73.0	34.3	1407

```
summary(cot)
```

```
##        LP             BW             BN            LS             SB
##  Min.   :31.3   Min.   :3.87   Min.   :211   Min.   :55.9   Min.   :19.5
##  1st Qu.:37.2   1st Qu.:5.13   1st Qu.:475   1st Qu.:67.0   1st Qu.:27.6
##  Median :38.6   Median :5.46   Median :555   Median :72.4   Median :29.3
##  Mean   :38.4   Mean   :5.54   Mean   :551   Mean   :72.3   Mean   :29.6
##  3rd Qu.:39.9   3rd Qu.:5.89   3rd Qu.:621   3rd Qu.:76.9   3rd Qu.:31.2
##  Max.   :42.3   Max.   :7.51   Max.   :942   Max.   :89.9   Max.   :40.7
##        LY
##  Min.   : 431
##  1st Qu.:1018
##  Median :1200
##  Mean   :1167
##  3rd Qu.:1336
##  Max.   :1741
```

```
str(cot)
```

```
## 'data.frame':    256 obs. of  6 variables:
##  $ LP: num  41.4 38.6 38 38.2 40.7 ...
##  $ BW: num  5.08 5.49 6.21 5.27 5.48 ...
##  $ BN: num  634 577 533 599 748 ...
##  $ LS: num  74.3 68.9 66.9 69.3 67.4 ...
##  $ SB: num  28.3 30.7 35.3 29 33 ...
##  $ LY: num  1333 1221 1256 1204 1664 ...
```

Correlation analysis

We will use the above data for our correlation analysis

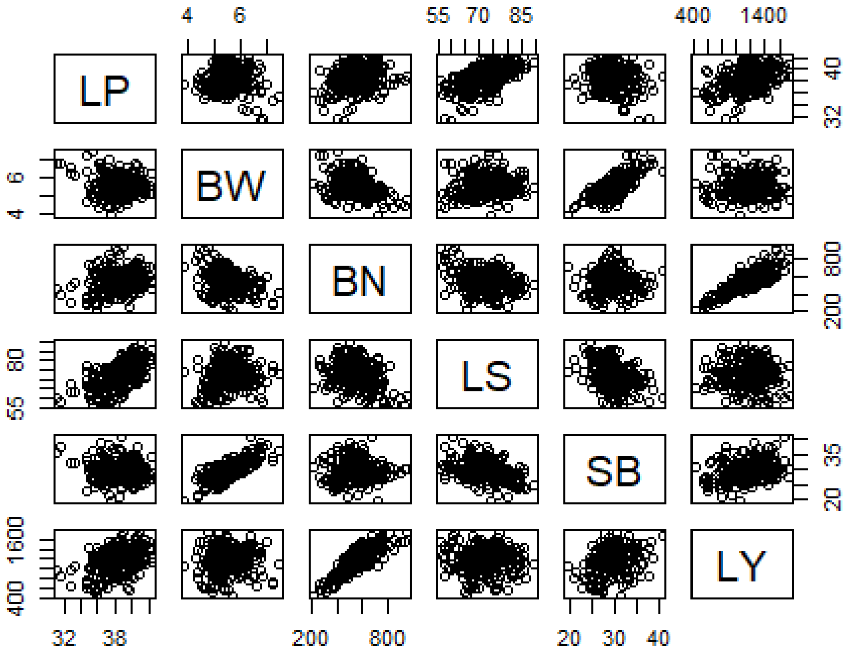
```
attach(cot)
cor(LY,BN)
```

```
## [1] 0.857
```

```
cor(cot)
```

```
##         LP      BW       BN     LS       SB      LY
## LP  1.0000 -0.0839  0.21038  0.629 -0.14569  0.4340
## BW -0.0839  1.0000 -0.37392  0.187  0.75416  0.0764
## BN  0.2104 -0.3739  1.00000 -0.286 -0.00692  0.8565
## LS  0.6294  0.1867 -0.28622  1.000 -0.38981 -0.0240
## SB -0.1457  0.7542 -0.00692 -0.390  1.00000  0.3061
## LY  0.4340  0.0764  0.85652 -0.024  0.30611  1.0000
```

```
pairs(cot)
```



Linear regression analysis

```
y=LY

reg1=lm(LY~BN)
summary(reg1)
```

```
##
## Call:
## lm(formula = LY ~ BN)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -420.2  -89.4    0.7   90.7  320.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 195.4384    37.5877     5.2  4.1e-07 ***
## BN            1.7627     0.0666    26.4  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 128 on 254 degrees of freedom
## Multiple R-squared:  0.734,  Adjusted R-squared:  0.733
## F-statistic:  700 on 1 and 254 DF,  p-value: <2e-16
```

```
reg=lm(y~LP+BW)
summary(reg)
```

```
## 
## Call:
## lm(formula = y ~ LP + BW)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##   -672   -143     18    158    586 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1266.85     316.07   -4.01  8.1e-05 ***
## LP             56.32       7.16    7.87  1.1e-13 ***
## BW             48.52      24.07    2.02    0.045 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 223 on 253 degrees of freedom
## Multiple R-squared:  0.201,  Adjusted R-squared:  0.195 
## F-statistic: 31.9 on 2 and 253 DF,  p-value: 4.56e-13
```

Linear regression with variable selection

Backward elimination

```
g=lm(LY~., data=cot)
summary(g)
```

```
## 
## Call:
## lm(formula = LY ~ ., data = cot)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -119.15  -11.68    3.44   15.96   89.88 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.34e+03   4.38e+01  -53.35   <2e-16 ***
## LP           1.51e+01   5.52e+00    2.73   0.0067 ** 
## BW           7.09e+01   3.55e+01    2.00   0.0469 *  
## BN           2.00e+00   1.95e-02  102.82   <2e-16 ***
## LS           9.79e+00   2.78e+00    3.52   0.0005 ***
## SB           2.43e+01   6.78e+00    3.59   0.0004 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 29.6 on 250 degrees of freedom
## Multiple R-squared:  0.986,  Adjusted R-squared:  0.986
## F-statistic: 3.52e+03 on 5 and 250 DF,  p-value: <2e-16
```

```
g=update(g, .~. -BW)
summary(g)
```

```
##
## Call:
## lm(formula = LY ~ LP + BN + LS + SB, data = cot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.93  -11.87    3.73   16.57   87.17
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.33e+03   4.39e+01  -53.07   <2e-16 ***
## LP           4.46e+00   1.49e+00    2.99    0.003 **
## BN           2.00e+00   1.96e-02  102.20   <2e-16 ***
## LS           1.53e+01   4.72e-01   32.35   <2e-16 ***
## SB           3.78e+01   6.68e-01   56.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.8 on 251 degrees of freedom
## Multiple R-squared:  0.986,  Adjusted R-squared:  0.986
## F-statistic: 4.35e+03 on 4 and 251 DF,  p-value: <2e-16
```

```
g=update(g, .~. -LP)
summary(g)
```

```
##
## Call:
## lm(formula = LY ~ BN + LS + SB, data = cot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -117.73  -12.03    4.76   15.58   86.18
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.27e+03   3.92e+01   -57.8   <2e-16 ***
## BN           2.03e+00   1.66e-02   122.8   <2e-16 ***
## LS           1.63e+01   3.18e-01    51.4   <2e-16 ***
## SB           3.83e+01   6.57e-01    58.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.3 on 252 degrees of freedom
## Multiple R-squared:  0.985,  Adjusted R-squared:  0.985
## F-statistic: 5.62e+03 on 3 and 252 DF,  p-value: <2e-16
```

Stepwise selection

```
g=lm(LY~., data=cot)
summary(g)
```

```
## 
## Call:
## lm(formula = LY ~ ., data = cot)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -119.15  -11.68    3.44   15.96   89.88 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.34e+03   4.38e+01  -53.35   <2e-16 ***
## LP           1.51e+01   5.52e+00    2.73   0.0067 ** 
## BW           7.09e+01   3.55e+01    2.00   0.0469 *  
## BN           2.00e+00   1.95e-02  102.82   <2e-16 ***
## LS           9.79e+00   2.78e+00    3.52   0.0005 ***
## SB           2.43e+01   6.78e+00    3.59   0.0004 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 29.6 on 250 degrees of freedom
## Multiple R-squared:  0.986,  Adjusted R-squared:  0.986 
## F-statistic: 3.52e+03 on 5 and 250 DF,  p-value: <2e-16
```

```
step(g)
```

```
## Start:  AIC=1741
## LY ~ LP + BW + BN + LS + SB
## 
##          Df Sum of Sq     RSS  AIC
## <none>                 219477 1741
## - BW      1      3501  222979 1743
## - LP      1      6554  226032 1747
## - LS      1     10908  230386 1751
## - SB      1     11294  230771 1752
## - BN      1   9281085 9500562 2704

## 
## Call:
## lm(formula = LY ~ LP + BW + BN + LS + SB, data = cot)
## 
## Coefficients:
## (Intercept)           LP           BW           BN           LS  
##    -2336.24        15.08        70.91         2.00         9.79  
##          SB  
##       24.31
```

Another stepwise selection

```
fit <- lm(y~BN+LP+BW+LS+SB,data=cot)
step <- stepAIC(fit, direction="both")
```

```
## Start:  AIC=1741
## y ~ BN + LP + BW + LS + SB
## 
##          Df Sum of Sq     RSS  AIC
## <none>                 219477 1741
## - BW      1      3501  222979 1743
## - LP      1      6554  226032 1747
## - LS      1     10908  230386 1751
## - SB      1     11294  230771 1752
## - BN      1   9281085 9500562 2704
```

```
step$anova # display results
```

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	250	219477	1741

Best subset selection

```
require(leaps)
b=regsubsets(LY~., data=cot, nbest=2)
summary(b)
```

```
## Subset selection object
## Call: regsubsets.formula(LY ~ ., data = cot, nbest = 2)
## 5 Variables  (and intercept)
##    Forced in Forced out
## LP     FALSE      FALSE
## BW     FALSE      FALSE
## BN     FALSE      FALSE
## LS     FALSE      FALSE
## SB     FALSE      FALSE
## 2 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          LP  BW  BN  LS  SB 
## 1  ( 1 ) " " " " "*" " " " "
## 1  ( 2 ) "*" " " " " " " " "
## 2  ( 1 ) " " "*" "*" " " " "
## 2  ( 2 ) " " " " "*" " " "*"
## 3  ( 1 ) "*" "*" "*" " " " "
## 3  ( 2 ) " " " " "*" "*" "*"
## 4  ( 1 ) "*" " " "*" "*" "*"
## 4  ( 2 ) " " "*" "*" "*" "*"
## 5  ( 1 ) "*" "*" "*" "*" "*"
```

Linear regression with bootstrapping

```
reg=lm(LY~.,data=cot)
summary(reg)

##
## Call:
## lm(formula = LY ~ ., data = cot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119.15  -11.68    3.44   15.96   89.88
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.34e+03   4.38e+01  -53.35   <2e-16 ***
## LP           1.51e+01   5.52e+00    2.73   0.0067 **
## BW           7.09e+01   3.55e+01    2.00   0.0469 *
## BN           2.00e+00   1.95e-02  102.82   <2e-16 ***
## LS           9.79e+00   2.78e+00    3.52   0.0005 ***
## SB           2.43e+01   6.78e+00    3.59   0.0004 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.6 on 250 degrees of freedom
## Multiple R-squared:  0.986,  Adjusted R-squared:  0.986
## F-statistic: 3.52e+03 on 5 and 250 DF,  p-value: <2e-16

bhat0=reg$coef
names(bhat0)

## [1] "(Intercept)" "LP"          "BW"          "BN"          "LS"
## [6] "SB"

N=1000
BHAT=matrix(0,N,length(bhat0))
X=cot[,-6]
head(X)
```

LP	BW	BN	LS	SB
41.4	5.08	634	74.3	28.3
38.6	5.49	577	68.9	30.7
38.0	6.21	533	66.9	35.3
38.2	5.27	599	69.3	29.0
40.6	5.48	748	67.4	33.0
40.3	6.23	561	73.0	34.3

```
n=length(cot$LY)
for(i in 1:N){
id=sample(n,replace=T)
y1=cot$LY[id]
X1=X[id,]
cot1=data.frame(y1,X1)
reg1=lm(y1~.,data=cot1)
bhat=reg1$coef
BHAT[i,]=bhat
}
colnames(BHAT)=names(bhat0)
data.frame(BHAT)[1:10,]
```

X.Intercept.	LP	BW	BN	LS	SB
-2309	0.234	-15.77	2.06	16.83	40.6
-2356	20.572	101.34	2.00	7.26	18.3
-2423	4.507	3.42	2.00	16.14	38.0
-2284	10.228	32.11	1.97	12.24	30.7
-2382	7.612	19.69	2.04	13.77	34.7
-2309	19.589	107.57	1.96	7.37	17.3
-2357	18.859	99.33	2.00	7.96	19.2
-2312	19.939	123.50	2.00	6.58	15.1
-2322	12.329	48.12	2.02	11.49	27.1
-2316	16.207	60.08	1.97	9.49	25.4

```
r=length(bhat0)
head(BHAT)

##      (Intercept)     LP     BW   BN    LS   SB
## [1,]       -2309  0.234 -15.77 2.06 16.83 40.6
## [2,]       -2356 20.572 101.34 2.00  7.26 18.3
## [3,]       -2423  4.507   3.42 2.00 16.14 38.0
## [4,]       -2284 10.228  32.11 1.97 12.24 30.7
## [5,]       -2382  7.612  19.69 2.04 13.77 34.7
## [6,]       -2309 19.589 107.57 1.96  7.37 17.3

CI=matrix(0,r,2)
for(i in 1:r){
 ci=quantile(BHAT[,i],p=c(0.025,0.975))
 CI[i,]=ci
 }
 colnames(CI)=c("LL","UL")
 rownames(CI)=names(bhat0)
data.frame(CI)
```

	LL	UL
(Intercept)	-2451.047	-2222.08
LP	0.868	34.13
BW	-24.239	211.71

BN	1.957	2.05
LS	-0.813	16.99
SB	-1.620	42.04

Linear regression with permutation

```
#cot=read.table("cotyldreg.txt",header=TRUE)
reg=lm(LY~.,data=cot)
summary(reg)
```

```
##
## Call:
## lm(formula = LY ~ ., data = cot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119.15  -11.68    3.44   15.96   89.88
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.34e+03   4.38e+01  -53.35   <2e-16 ***
## LP           1.51e+01   5.52e+00    2.73   0.0067 **
## BW           7.09e+01   3.55e+01    2.00   0.0469 *
## BN           2.00e+00   1.95e-02  102.82   <2e-16 ***
## LS           9.79e+00   2.78e+00    3.52   0.0005 ***
## SB           2.43e+01   6.78e+00    3.59   0.0004 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.6 on 250 degrees of freedom
## Multiple R-squared:  0.986,  Adjusted R-squared:  0.986
## F-statistic: 3.52e+03 on 5 and 250 DF,  p-value: <2e-16
```

```
bhat0=reg$coef
names(bhat0)
```

```
## [1] "(Intercept)" "LP"          "BW"          "BN"          "LS"
## [6] "SB"
```

```
N=1000
BHAT=matrix(0,N,length(bhat0))
X=cot[,-6]
head(X)
```

LP	BW	BN	LS	SB
41.4	5.08	634	74.3	28.3
38.6	5.49	577	68.9	30.7
38.0	6.21	533	66.9	35.3
38.2	5.27	599	69.3	29.0

40.6	5.48	748	67.4	33.0
40.3	6.23	561	73.0	34.3

```
n=length(cot$LY)
for(i in 1:N){
  id=sample(n,replace=FALSE)
  y1=cot$LY[id]
  cot1=data.frame(y1,X)
  reg1=lm(y1~.,data=cot1)
  bhat=reg1$coef
  BHAT[i,]=bhat
}
colnames(BHAT)=names(bhat0)
data.frame(BHAT)[1:10,]
```

X.Intercept.	LP	BW	BN	LS	SB
1466	15.75	158.51	0.006	-12.623	-29.55
826	65.37	408.72	0.025	-32.320	-71.50
1287	76.45	467.44	-0.219	-38.814	-92.07
1006	5.97	45.69	0.117	-3.145	-5.39
1890	-13.50	-6.73	0.009	-3.042	1.62
988	-59.51	-371.37	0.103	31.331	74.48
1283	6.70	84.21	0.120	-6.960	-13.65
1369	-37.78	-201.31	-0.084	18.234	37.00
690	-32.30	-269.49	-0.104	21.742	57.40
707	14.82	-25.84	-0.180	-0.133	4.79

```
r=length(bhat0)
head(BHAT)

##      (Intercept)     LP     BW       BN     LS     SB
## [1,]        1466  15.75  158.51  0.00552 -12.62 -29.55
## [2,]         826  65.37  408.72  0.02528 -32.32 -71.50
## [3,]        1287  76.45  467.44 -0.21891 -38.81 -92.07
## [4,]        1006   5.97   45.69  0.11747  -3.15  -5.39
## [5,]        1890 -13.50   -6.73  0.00944  -3.04   1.62
## [6,]         988 -59.51 -371.37  0.10345  31.33  74.48

CI=matrix(0,r,2)
for(i in 1:r){
 ci=quantile(BHAT[,i],p=c(0.025,0.975))
 CI[i,]=ci
 }
 colnames(CI)=c("LL","UL")
 rownames(CI)=names(bhat0)
data.frame(bhat0,CI)
```

	bhat0	LL	UL

(Intercept)	-2336.24	449.687	1908.208
LP	15.08	-88.154	88.020
BW	70.91	-574.202	582.960
BN	2.00	-0.322	0.299
LS	9.79	-44.666	44.903
SB	24.31	-112.769	110.663