BIO5312 Biostatistics
R Session 10: Regression and Correlation

Dr. Junchao Xia

Center of Biophysics and Computational Biology

Fall 2016

Correlation Matrix

➢Examples using the functions related to correlation matrix

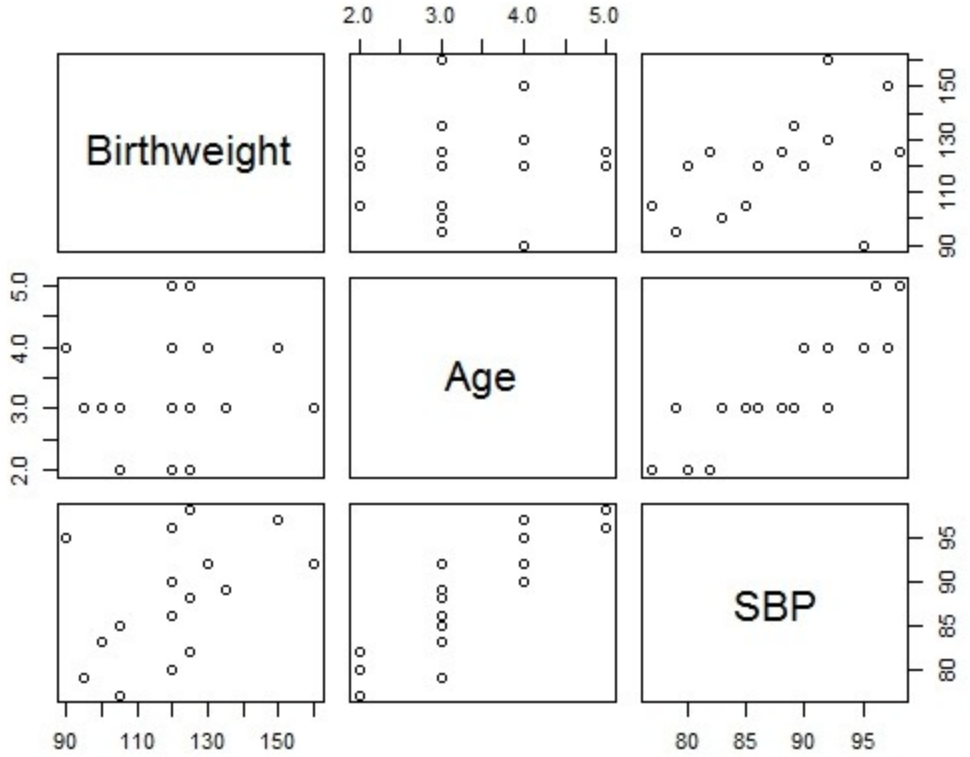
```
# set work directory
> setwd("C:/Users/Junchao/Desktop/Biostatistics_5312/2016/lab_10")
# read data from the data file
>sbp=read.table("Table11.9.DAT.txt",header=T)

#  get the correlation matrix of  for all variables in the data file
>cor(sbp)
                    ID         Birthweight         Age             SBP
ID             1.00000000  0.2595186  -0.05178087  0.1203892
Birthweight    0.25951858 1.0000000   0.10682804   0.4410894
Age           -0.05178087 0.1068280   1.00000000   0.8708424
SBP            0.12038915  0.4410894   0.87084245  1.0000000
#  get the covariance matrix
>cov(sbp)
# get the correlation matrix from the covariance matrix
>cov2cor(cov(sbp))
# extract 2,3, 4 column from sbp
>mat=sbp[c(2,3,4)]
# plotting the SBP as a function of Age
with(sbp,plot(Age,SBP))                 # or plot(sbp$Age,sbp$SBP)
# plotting the SBP as a function of Birthweight
with(sbp,plot(Birthweight,SBP))      # or plot(sbp$Birthweight,sbp$SBP)
```

Correlation Matrix (continued)

➢Examples using the functions related to correlation

get all pair plots

>pairs(mat)



Correlation Test

➢Examples using the test functions related to correlation

```
# get the help
> help("cor.test")
 ## Default S3 method:
cor.test(x, y, alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"), exact =
NULL, conf.level = 0.95, continuity = FALSE, ...)
## S3 method for class 'formula'
cor.test(formula, data, subset, na.action, ...)

#  using cor.test to perform t test
>with(sbp,cor.test(Age,SBP))
Pearson's product-moment correlation data: Age and SBP
t = 6.6287, df = 14, p-value = 1.135e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.6600822 0.9545056
sample estimates: cor 0.8708424

#  using one sided method
> with(sbp,cor.test(Age,SBP,alternative="greater", conf.level=.99))
Pearson's product-moment correlation data: Age and SBP
t = 6.6287, df = 14, p-value = 5.675e-06
alternative hypothesis: true correlation is greater than 0
99 percent confidence interval: 0.5988439 1.0000000
sample estimates: cor 0.8708424
# using spearman rank correlation
> with(sbp,cor.test(Age,SBP,alternative="greater", method="spearman",conf.level=.99))
```

Simple Linear Regression

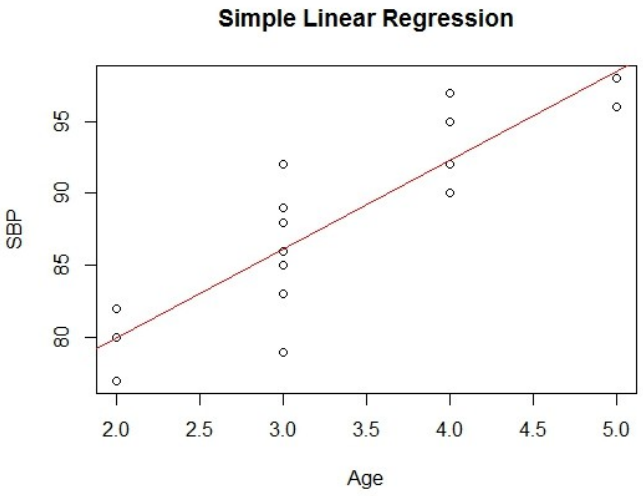
➢Examples using the functions related to simple linear regression

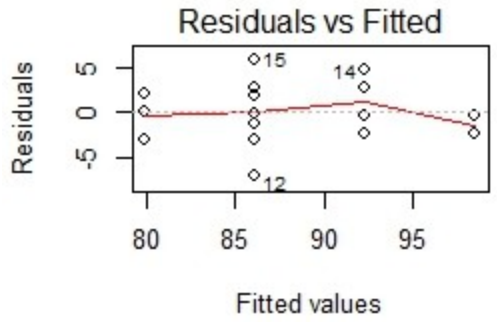
```
>help(lm)
# perform a simple linear regression between SBP and Age
>slm.out=lm(SBP~Age, data=mat)
> summary(slm.out)
Call: lm(formula = SBP ~ Age, data = mat)
Residuals: Min 1Q Median 3Q Max
-7.1395 -2.3314 -0.2163 2.1872 5.8605
Coefficients:
                  Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)       67.6791     3.1906    21.212   4.84e-12 ***
Age                6.1535     0.9283     6.629   1.13e-05 ***
---
Signif. codes:    0  ?**?0.001  ?*?0.01  ??0.05  ??0.1  ??1
Residual standard error: 3.403 on 14 degrees of freedom
Multiple R-squared: 0.7584, Adjusted R-squared: 0.7411
F-statistic: 43.94 on 1 and 14 DF, p-value: 1.135e-05
```

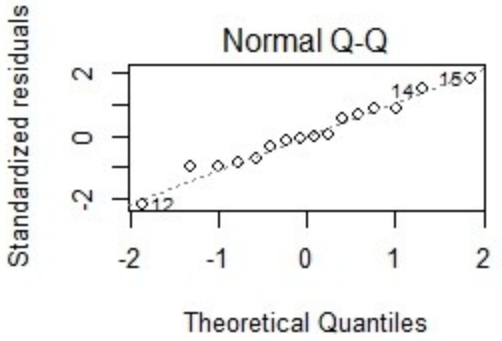
Simple Linear Regression (continued)

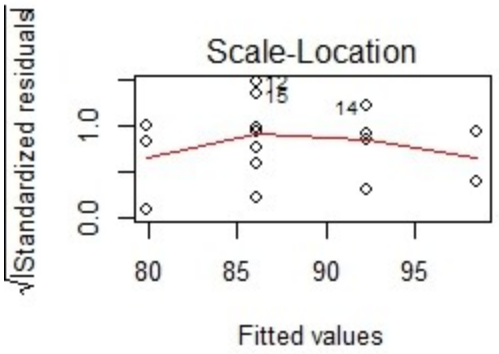
➢Examples using the plot functions to output simple linear regression results

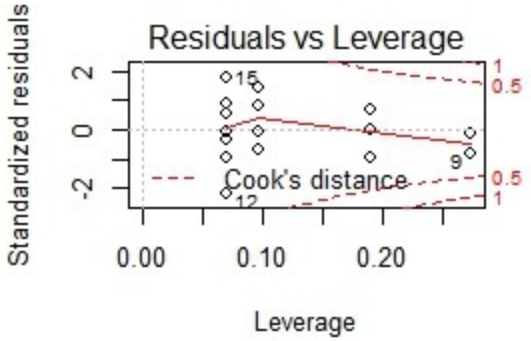
```
# plot the data and the fitting line
>plot(SBP~Age,data=mat,main="Simple Linear Regression")
>abline(slm.out, col="red")
# partition the graphics device
>par(mfrow=c(2,2))
#plot related figures from fitting
>plot(slm.out)
>par(mfrow=c(1,1))
>plot(cooks.distance(slm.out))
```











Simple Linear Regression (continued)

- Other useful functions

```
# model coefficients
>coefficients(slm.out)
# CIs for model parameters
>confint(slm.out, level=0.95)
# predicted values
>fitted(slm.out)
# residuals
>residuals(slm.out)
# anova table
>anova(slm.out)
# covariance matrix for model parameters
>vcov(slm.out)
# regression diagnostics
>influence(slm.out)
```

Multiple Linear Regression

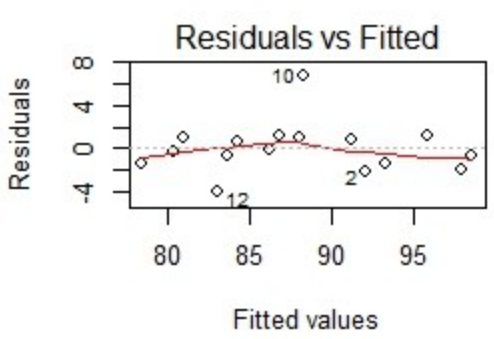
➢Examples using the functions related to simple linear regression

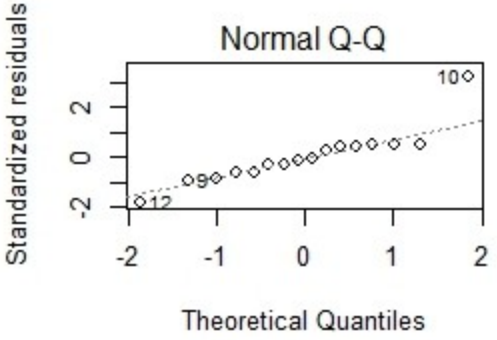
```
>help(lm)
# perform a  multiple linear regression between SBP, Age and, Birthweight
>mlm.out=lm(SBP~Age+Birthweight,data=mat)
>summary(mlm.out)
Call:
lm(formula = SBP ~ Age + Birthweight, data = mat)
Residuals:  Min  1Q  Median  3Q  Max
-4.0438 -1.3481 -0.2395 0.9688 6.6964
Coefficients:
                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)    53.45019    4.53189  11.794   2.57e-08 ***
Age             5.88772    0.68021    8.656   9.34e-07 ***
Birthweight     0.12558    0.03434    3.657     0.0029 **
---
Signif. codes:   0  ?**?0.001   ?*?0.01   ??0.05   ??0.1   ??1
Residual standard error: 2.479 on 13 degrees of freedom
Multiple R-squared: 0.8809, Adjusted R-squared: 0.8626
F-statistic: 48.08 on 2 and 13 DF, p-value: 9.844e-07
```

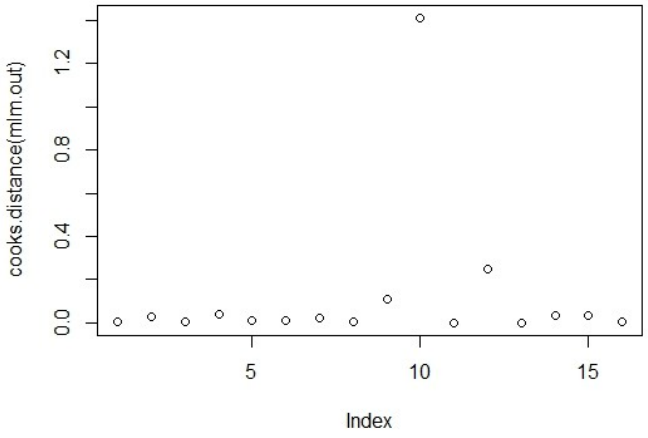
Multiple Linear Regression (continued)

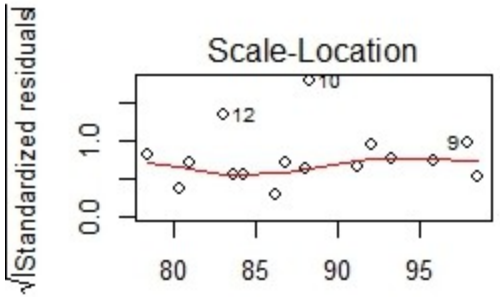
➢Examples using the plot functions to output multiple linear regression results

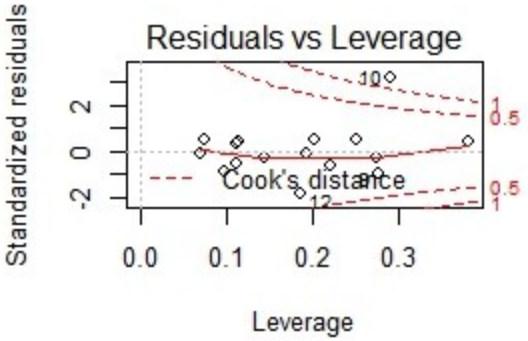
```
# plot the data and the fitting line
# partition the graphics device
>par(mfrow=c(2,2))
#plot related figures from fitting
>plot(mlm.out)
>par(mfrow=c(1,1))
>plot(cooks.distance(mlm.out))
```











ANOVA and Model Comparison

➢Examples using ANOVA to compare different models

```
> anova(slm.out)
Analysis of Variance Table
Response: SBP
           Df  Sum Sq  Mean Sq  F value      Pr(>F)
Age         1  508.82   508.82  43.939 1.135e-05 ***
Residuals  14  162.12    11.58
```

```
>anova(mlm.out)
Analysis of Variance Table
Response: SBP
            Df  Sum Sq  Mean Sq  F value      Pr(>F)
Age          1  508.82   508.82  82.784  5.322e-07 ***
Birthweight  1   82.22    82.22  13.377   0.002896 **
Residuals   13   79.90     6 .15
```

```
# perform partial F test using anova
>anova(slm.out,mlm.out)
Analysis of Variance Table
Model 1: SBP ~ Age
Model 2: SBP ~ Age + Birthweight
     Res.Df      RSS      Df    Sum of Sq          F        Pr(>F)
 1       14  162.121
 2       13   79.902   1        82.219    13.377    0.002896 **
```

ANOVA and Model Comparison

- using stepAIC to perform variable selection

```
>install.packages("MASS")
>library("MASS")
>steplm=lm(SBP~Age+Birthweight,data=mat)
>step.out =stepAIC(steplm, direction="both")
Start: AIC=31.73
SBP ~ Age + Birthweight
```

	Df	Sum of Sq	RSS	AIC
<none>			79.90	31.731
- Birthweight	1	82.22	162.12	41.052
- Age	1	460.50	540.40	60.316

The End