Stat 415/615, Lab 3. Multiple Linear Regression (I)

Jun Lu

Stat 415/615 Regression, 2023

Contents

1 Plot the data 2

2 Multiple Linear Regression 4

3 ANOVA table 4

4 The estimated coefficients 5

5 Test of significance 5

6 Critical values for simultaneous (aka family, jointly) inference 6

7 Reminder: other useful output and functions 8

We will discuss the explanation and comments in class.

Refer to the Body Fat Example in textbook p.256. The following information is included in file CH07TA01.txt.

- 20 healthy female subjects
- Y: body fat (Underwater weighing is the "gold standard" used to obtain the measurement)
- X1: triceps skin fold thickness
- X2: thigh circumference
- X3: midarm circumference

```
# CH07TA01.txt doesn't include variable names.
bfdata<-read.table("../DataSets/CH07TA01.txt", header=F)
colnames(bfdata) <- c("triceps", "thigh", "midarm", "bodyfat")
summary(bfdata)
```

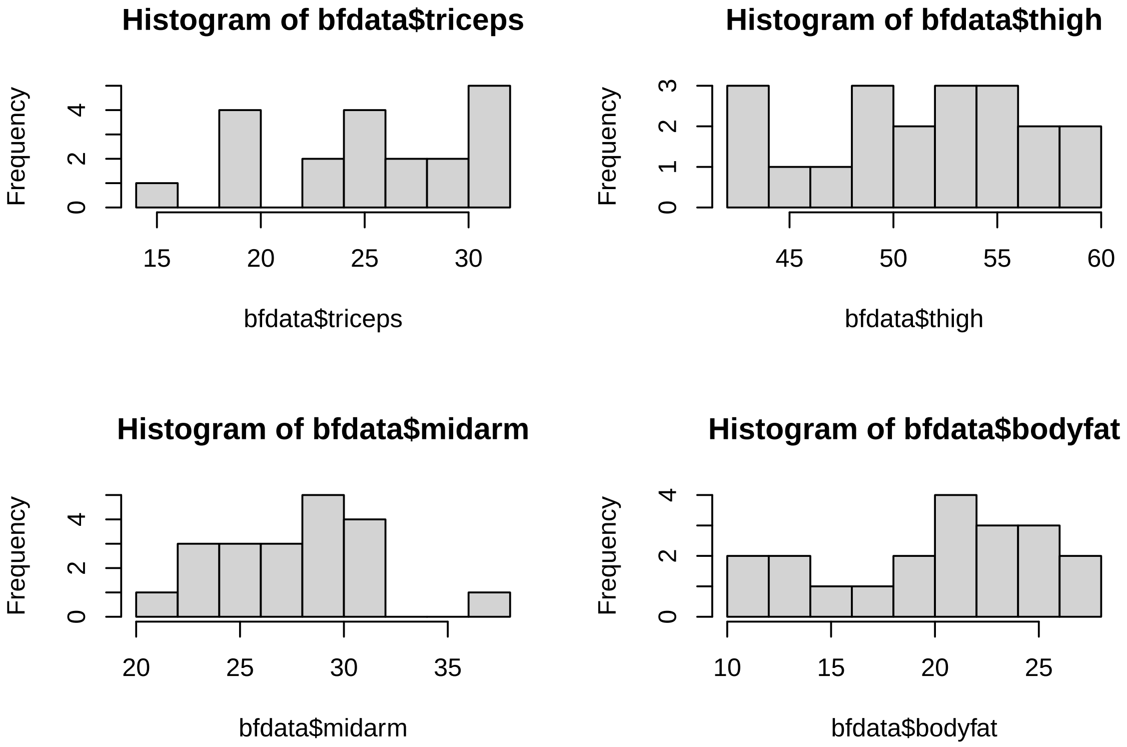
```
##     triceps          thigh           midarm         bodyfat
##  Min.   :14.60   Min.   :42.20   Min.   :21.30   Min.   :11.70
##  1st Qu.:21.50   1st Qu.:47.77   1st Qu.:24.75   1st Qu.:17.05
##  Median :25.55   Median :52.00   Median :27.90   Median :21.20
##  Mean   :25.30   Mean   :51.17   Mean   :27.62   Mean   :20.20
##  3rd Qu.:29.90   3rd Qu.:54.62   3rd Qu.:30.02   3rd Qu.:24.27
##  Max.   :31.40   Max.   :58.60   Max.   :37.00   Max.   :27.20
```

```
# BodyFat.txt has the same data, with variables names included.
bfdata<-read.table("../DataSets/BodyFat.txt", header=T)
```

1 Plot the data

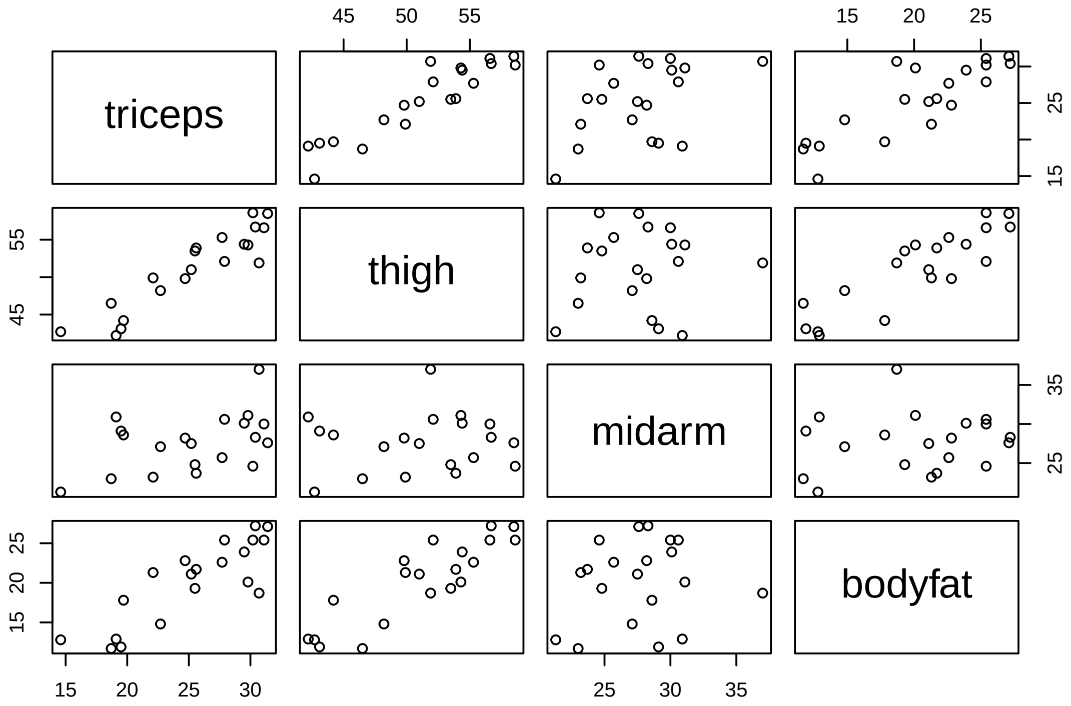
a. One variable at a time.

```
par(mfrow=c(2,2))
hist(bfdata$triceps)
hist(bfdata$thigh)
hist(bfdata$midarm)
hist(bfdata$bodyfat)
```



b. Two variable pairs.

```
pairs(bfdata)
```



```
# Correlation coefficients
cor(bfdata)
```

```
##           triceps     thigh    midarm   bodyfat
## triceps 1.0000000 0.9238425 0.4577772 0.8432654
## thigh   0.9238425 1.0000000 0.0846675 0.8780896
## midarm  0.4577772 0.0846675 1.0000000 0.1424440
## bodyfat 0.8432654 0.8780896 0.1424440 1.0000000
```

2 Multiple Linear Regression

```
bfreg1<-lm(bodyfat~triceps+thigh+midarm, data=bfdata)
summary(bfreg1)
```

```
##
## Call:
## lm(formula = bodyfat ~ triceps + thigh + midarm, data = bfdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7263 -1.6111  0.3923  1.4656  4.1277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  117.085     99.782   1.173    0.258
## triceps        4.334      3.016   1.437    0.170
## thigh         -2.857      2.582  -1.106    0.285
## midarm        -2.186      1.595  -1.370    0.190
##
## Residual standard error: 2.48 on 16 degrees of freedom
## Multiple R-squared:  0.8014, Adjusted R-squared:  0.7641
## F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

3 ANOVA table

a. Confirm how the numbers in the table are connected.

b. Confirm the computation of degree of freedom.

c. Is the model significant overall? Lay out the hypothesis, p-values, and state your conclusion in the context of the problem.

```
# Get SStotal as reference.
sstotal <- sum((bfdata$bodyfat-mean(bfdata$bodyfat))^2)
paste("SStotal = ", sstotal, "  Sample size (n) = ", nrow(bfdata))
```

```
## [1] "SStotal =  495.3895   Sample size (n) =  20"
```

```
bfregNull<-lm(bodyfat~1, data=bfdata)
anova(bfregNull, bfreg1)  # "overall significance"
```

```
## Analysis of Variance Table
##
## Model 1: bodyfat ~ 1
## Model 2: bodyfat ~ triceps + thigh + midarm
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     19 495.39
## 2     16  98.40  3    396.98 21.516 7.343e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(bfreg1)  # Sequential SS (Type I SS)
```

```
## Analysis of Variance Table
##
## Response: bodyfat
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## triceps    1 352.27  352.27 57.2768 1.131e-06 ***
## thigh      1  33.17   33.17  5.3931   0.03373 *  
## midarm     1  11.55   11.55  1.8773   0.18956    
## Residuals 16  98.40    6.15                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4 The estimated coefficients

a. Locate the estimates and their standard errors. Why are they useful?
b. Does any slope parameter appear to be significant? Is this finding counter-intuitive? Does this fining contradict the conclusion about the model from the ANOVA table?

```
summary(bfreg1)
```

```
## 
## Call:
## lm(formula = bodyfat ~ triceps + thigh + midarm, data = bfdata)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.7263 -1.6111  0.3923  1.4656  4.1277 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  117.085     99.782   1.173    0.258
## triceps        4.334      3.016   1.437    0.170
## thigh         -2.857      2.582  -1.106    0.285
## midarm        -2.186      1.595  -1.370    0.190
## 
## Residual standard error: 2.48 on 16 degrees of freedom
## Multiple R-squared:  0.8014, Adjusted R-squared:  0.7641 
## F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

5 Test of significance

Conduct test of significance to address the following questions. Clearly state the null and alternative hypothesis, the appropriate "Reduced models" when applicable, p-values and your conclusion in the context of the problem.

a. Can predictors Thigh and Midarm be both dropped from the regression model?

```
bfreg.5a<-lm(bodyfat~triceps, data=bfdata)
anova(bfreg.5a, bfreg1)
```

```
## Analysis of Variance Table
## 
## Model 1: bodyfat ~ triceps
## Model 2: bodyfat ~ triceps + thigh + midarm
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)  
## 1     18 143.120                              
## 2     16  98.405  2    44.715 3.6352 0.04995 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b. Do Thigh and Midarm have the same slope parameter?

```
bfdata$TplusM <- bfdata$thigh+bfdata$midarm
bfreg.5b<-lm(bodyfat~triceps+TplusM, data=bfdata)
anova(bfreg.5b, bfreg1)
```

```
## Analysis of Variance Table
##
## Model 1: bodyfat ~ triceps + TplusM
## Model 2: bodyfat ~ triceps + thigh + midarm
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     17 101.110
## 2     16  98.405  1    2.7046 0.4398 0.5167
```

c. Test H0: $\beta_{triceps} = 1$, $\beta_{thigh} = 2$, vs. H1: At Least One Inequality.

```
bfdata$newY <- bfdata$bodyfat-1*bfdata$triceps-2*bfdata$thigh
bfreg.5c <- lm(newY~midarm, data=bfdata)
# anova(bfreg.5c, bfreg1)  will NOT work here.
anova(bfreg.5c)
```

```
## Analysis of Variance Table
##
## Response: newY
##           Df  Sum Sq Mean Sq F value Pr(>F)
## midarm     1  114.85  114.85   0.947 0.3434
## Residuals 18 2183.00  121.28
```

```
anova(bfreg1)
```

```
## Analysis of Variance Table
##
## Response: bodyfat
##           Df Sum Sq Mean Sq F value    Pr(>F)
## triceps    1 352.27  352.27 57.2768 1.131e-06 ***
## thigh      1  33.17   33.17  5.3931   0.03373 *
## midarm     1  11.55   11.55  1.8773   0.18956
## Residuals 16  98.40    6.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
f<-(2183-98.4)/(18-16)/(6.15)
f
```

```
## [1] 169.4797
```

```
1-pf(f, 2, 16)
```

```
## [1] 1.704237e-11
```

6 Critical values for simultaneous (aka family, jointly) inference

Consider a family of 3 intervals and a family confidence level of 0.9. Assume the data set has 100 observations, and the model is $E(Y) = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ (i.e., $p = 4$).

```
family.alpha <- 0.1
g <- 3
```

```r
n <- 100
p <- 4
dfE <- n - p
```

a. Bonferroni procedure

- Applies to CI for β, CI for mean of Y, PI for individual Y, and tests of hypothesis.
- Use $1 - \alpha^* = 1 - \alpha/g$ for each CI in the family. Or $\alpha* = \alpha/g$ for each test in the family.
- Do not change the distribution used for critical value or p-value computation. (t-distribution in most cases in regression.)
- Using a **one-sided** t-calculator, $t_{(1-(\alpha^*/2), df=df_E)} = t_{(1-\alpha/(2g), df=(n-p))}$

```r
member.alpha <- family.alpha/g
B <- qt(1-member.alpha/2, df=dfE)
B
```

```
## [1] 2.159116
```

- Bonferroni procedure can be impletmented in `confint()` and `predict()` directly. We just need to specify `level = 1- member.alpha` in those functions. (Recall that, let g be the number of CIs or PIs in the family, the "member.alpha" is family.alpha/g.)

b. Working-Hotelling

- Only applies to CI for mean of Y.
- Use $W = \sqrt{pF_{(1-\alpha;\, p,\, n-p)}}$ as the critical value.

```r
W <- sqrt(p*qf(1-family.alpha, df1=p, df2=dfE))
W
```

```
## [1] 2.831505
```

- To implement Working-Hotelling procedure, one option is to use `predict()` function to get $\hat{y}$ and $se(\hat{y}_{mean})$. Then construct the intervals by $\hat{y} \pm W * se(\hat{y}_{mean})$.
- (Optional) Another method to implement Working-Hotelling procedure. First, determine the W critical value. Then, find out at which level (denote as l_2), the t_{crit} value will be the same as the W. Then, use `predict()` function and set `level=`l_2.

c. Scheffe

- Only applies to PIs for individual Y.
- Use $S = \sqrt{gF_{(1-\alpha;\, g,\, n-p)}}$ as the critical value, where g is the number of members in the family.

```r
S <- sqrt(g*qf(1-family.alpha, df1=g, df2=dfE))
S
```

```
## [1] 2.534796
```

- To implement Scheffe procedure, one option is to use `predict()` function to get $\hat{y}$ and $se(\hat{y}_{mean})$. Then compute $se(\hat{y}_{new}) = \sqrt{(se(\hat{y}_{mean}))^2 + MSE}$. The PI is constructed by $\hat{y} \pm S * se(\hat{y}_{new})$.
- (Optional) Another method to implement Scheffe procedure. First, determine the W critical value. Then, find out at which level (denote as l_2), the t_{crit} value will be the same as the W. Then, use `predict()` function and set `level=`l_2.

7 Reminder: other useful output and functions

```
bfreg1$residuals
```

```
##          1          2          3          4          5          6          7
## -2.9549896  2.5811589 -2.2866822 -3.0273199  1.1423925 -0.5437185  1.3856834
##          8          9         10         11         12         13         14
##  3.1293594  1.7051817 -1.2483822  0.8044445  2.2076913 -3.3094005  4.1276946
##         15         16         17         18         19         20
##  0.9880521  0.1725323 -0.3736041 -1.3859022 -3.7262800  0.6120883
```

```
bfreg1$fitted.values
```

```
##        1        2        3        4        5        6        7        8
## 14.85499 20.21884 20.98668 23.12732 11.75761 22.24372 25.71432 22.27064
##        9       10       11       12       13       14       15       16
## 19.59482 20.54838 24.59556 24.99231 15.00940 13.67231 11.81195 23.72747
##       17       18       19       20
## 22.97360 26.78590 18.52628 20.48791
```

```
confint(bfreg1, level=0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -94.444550 328.613940
## triceps      -2.058507  10.726691
## thigh        -8.330476   2.616780
## midarm       -5.568367   1.196247
```

```
predict(bfreg1, newdata=data.frame(midarm=c(25, 27, 30),
  triceps=c(20, 24, 26), thigh=c(45, 48, 50)), se=T,
  level= 0.9, interval="confidence")
```

```
## $fit
##        fit      lwr      upr
## 1 20.55687 13.24815 27.86560
## 2 24.95058 15.75339 34.14776
## 3 21.34688 18.72623 23.96753
##
## $se.fit
##        1        2        3
## 4.186261 5.267925 1.501044
##
## $df
## [1] 16
##
## $residual.scale
## [1] 2.479981
```

```
predict(bfreg1, newdata=data.frame(midarm=c(25, 27, 30),
  triceps=c(20, 24, 26), thigh=c(45, 48, 50)), se=F,
  level = 0.9, interval="prediction")
```

```
##        fit      lwr      upr
## 1 20.55687 12.06192 29.05183
## 2 24.95058 14.78519 35.11596
## 3 21.34688 16.28580 26.40797
```

—— **This is the end of Lab 3.** ——