MLR Regression

Yes

General Linear Model

$Y = X\beta + \epsilon$

PART I: some matrix operations that you might find useful

```
x <- 1:5/5;    # say, we have these x values   #
x;
```

```
[1] 0.2 0.4 0.6 0.8 1.0
```

```
var(x);         # the sample variance of those x's #
```

```
[1] 0.1
```

```r
x^2;
```

```
[1] 0.04 0.16 0.36 0.64 1.00
```

```r
mean(x);        # x_bar #
```

```
[1] 0.6
```

```r
sum((x-mean(x))^2); # sum of (x_i-x_bar)^2 i=1,2,..#
```

```
[1] 0.4
```

```r
sum((x-mean(x))^2)/(length(x)-1);  # c.f. var(x) #
```

```
[1] 0.1
```

```r
y<-x+rnorm(5)
fm<-lm(y~x)
X <- model.matrix(fm)  # this gives the design matrix  in a S-L-R model, [1, x]#
X
```

```
  (Intercept)   x
1           1 0.2
2           1 0.4
3           1 0.6
4           1 0.8
5           1 1.0
attr(,"assign")
[1] 0 1
```

```r
t(X);    # transpose of X #
```

```
              1   2   3   4 5
(Intercept) 1.0 1.0 1.0 1.0 1
x           0.2 0.4 0.6 0.8 1
attr(,"assign")
[1] 0 1
```

```r
A <- solve(t(X)%*%X);  # (X^tX)^(-1) #
A;
```

```
            (Intercept)    x
(Intercept)         1.1 -1.5
x                  -1.5  2.5
```

```r
H <- X%*%A%*%t(X);
H;    # the hat matrix #
```

```
              1             2   3             4             5
1  6.000000e-01  4.000000e-01 0.2 -6.661338e-17 -2.000000e-01
2  4.000000e-01  3.000000e-01 0.2  1.000000e-01 -1.110223e-16
3  2.000000e-01  2.000000e-01 0.2  2.000000e-01  2.000000e-01
4 -1.110223e-16  1.000000e-01 0.2  3.000000e-01  4.000000e-01
5 -2.000000e-01 -1.110223e-16 0.2  4.000000e-01  6.000000e-01
```

PART II: Multiple regression – Dwaine Studios Example

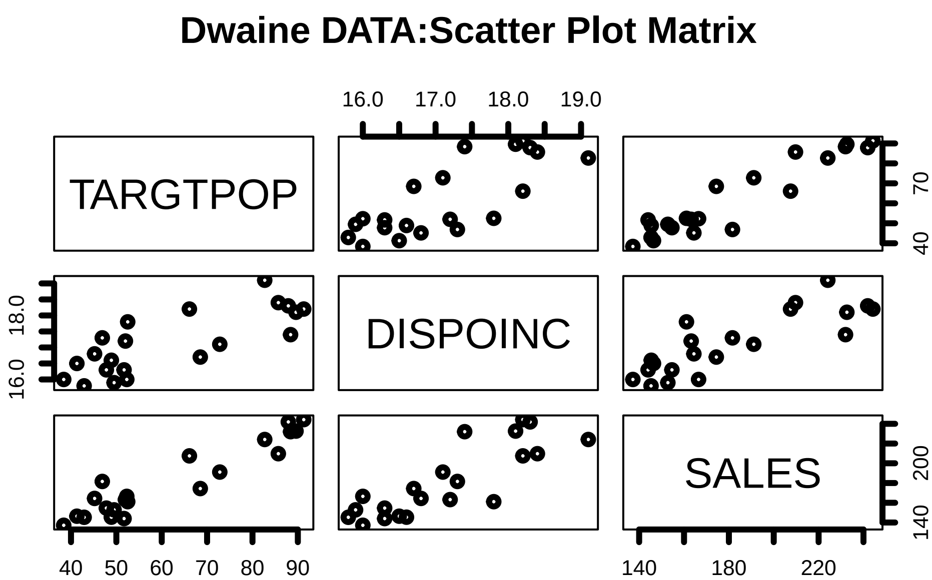
Dwaine Studios, Inc., operates portrait studios in 21 cities of medium size. These studios specialize in portraits of children. The company is considering an expansion into other cities of medium size and wishes to investigate whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the commmunity (X_1) and the per capita disposable income in the community (X_2). Data on these variables for the most recent year for the 21 cities in which Dwaine Studios is now operating are given below. Sales arae expressed in thousands of dollars and are labeled Y or SALES; the number of persons aged 16 or younger is expressed in thousands of persons and labeled X_1 or TARGTPOP for the target population; and per capita disposable personal income is expressed in thousands of dollars and labeled X_2 or DISPOINC for disposable income.

```r
ch6fi05 <- matrix(scan("CH06FI05.txt"),ncol=3, byrow=T) ;
dum <- data.frame(ch6fi05)
names(dum) <- c("TARGTPOP","DISPOINC","SALES")
# x1 <- ch6fi05[,1];
# x2 <- ch6fi05[,2];
# y <- ch6fi05[,3];
# dum <- data.frame("SALES"=y,"TARGTPOP"=x1,"DISPOINC"=x2)


#Another way to input data into R
```

```r
#dum <-read.table("CH06FI05.txt", header=FALSE,col.names = c("TARGTPOP", "DISPOINC","SALES"))

attach(dum)
# rm(x1,x2,y)
dum
```

```
   TARGTPOP DISPOINC SALES
1      68.5     16.7 174.4
2      45.2     16.8 164.4
3      91.3     18.2 244.2
4      47.8     16.3 154.6
5      46.9     17.3 181.6
6      66.1     18.2 207.5
7      49.5     15.9 152.8
8      52.0     17.2 163.2
9      48.9     16.6 145.4
10     38.4     16.0 137.2
11     87.9     18.3 241.9
12     72.8     17.1 191.1
13     88.4     17.4 232.0
14     42.9     15.8 145.3
15     52.5     17.8 161.1
16     85.7     18.4 209.7
17     41.3     16.5 146.4
18     51.7     16.3 144.0
19     89.6     18.1 232.6
20     82.7     19.1 224.1
21     52.3     16.0 166.5
```

```r
# To get Figure 6.4, p.232 #
pairs(dum, main="Dwaine DATA:Scatter Plot Matrix",lwd=3);   # (a) Scatter plot matrix
```




```r
cor(dum);                 # (b) Correlation matrix
```

```
          TARGTPOP  DISPOINC     SALES
TARGTPOP 1.0000000 0.7812993 0.9445543
DISPOINC 0.7812993 1.0000000 0.8358025
SALES    0.9445543 0.8358025 1.0000000
```

```r
# Do MLR: regress SALES on those two predictors #
fm <- lm(SALES~TARGTPOP+DISPOINC, dum)

# To get (b) Basic Data  on p237 of the text #
fi65b<-data.frame("x1"=TARGTPOP,"x2"=DISPOINC,"y"=SALES,"y_hat"=fitted(fm),"residual"=resid(
fi65b;
```

```
    x1   x2     y    y_hat    residual
1 68.5 16.7 174.4 187.1841 -12.7841146
2 45.2 16.8 164.4 154.2294  10.1705737
3 91.3 18.2 244.2 234.3963   9.8036764
4 47.8 16.3 154.6 153.3285   1.2714690
5 46.9 17.3 181.6 161.3849  20.2150722
6 66.1 18.2 207.5 197.7414   9.7585779
7 49.5 15.9 152.8 152.0551   0.7449178
```

```
8  52.0 17.2 163.2 167.8666  -4.6666316
9  48.9 16.6 145.4 157.7382 -12.3381967
10 38.4 16.0 137.2 136.8460   0.3539791
11 87.9 18.3 241.9 230.3874  11.5126289
12 72.8 17.1 191.1 197.1849  -6.0849209
13 88.4 17.4 232.0 222.6857   9.3142995
14 42.9 15.8 145.3 141.5184   3.7815611
15 52.5 17.8 161.1 174.2132 -13.1132116
16 85.7 18.4 209.7 228.1239 -18.4238900
17 41.3 16.5 146.4 145.7470   0.6530062
18 51.7 16.3 144.0 159.0013 -15.0013134
19 89.6 18.1 232.6 230.9870   1.6129777
20 82.7 19.1 224.1 230.3161  -6.2160615
21 52.3 16.0 166.5 157.0644   9.4356009
```

```
# To get (a) Multiple Regression Output  on p.237 #
summary(fm)
```

```
Call:
lm(formula = SALES ~ TARGTPOP + DISPOINC, data = dum)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.4239  -6.2161   0.7449   9.4356  20.2151 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -68.8571    60.0170  -1.147   0.2663    
TARGTPOP      1.4546     0.2118   6.868    2e-06 ***
DISPOINC      9.3655     4.0640   2.305   0.0333 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167,    Adjusted R-squared:  0.9075 
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

```
anova(fm)
```

```
Analysis of Variance Table
```

```
Response: SALES
          Df  Sum Sq Mean Sq  F value   Pr(>F)
TARGTPOP   1 23371.8 23371.8 192.8962 4.64e-11 ***
DISPOINC   1   643.5   643.5   5.3108  0.03332 *
Residuals 18  2180.9   121.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Note that the ANOVA table you get from R is a little different from that
# on p.237.  Think: How to get the ANOVA table on p.237 from R's output?

# To get (X^tX)^(-1) , p.237#
X<-model.matrix(fm)  #the design matrix

D<-solve(t(X)%*%X)
D;
```

```
            (Intercept)      TARGTPOP     DISPOINC
(Intercept) 29.72892348  0.0721834719 -1.992553186
TARGTPOP     0.07218347  0.0003701761 -0.005549917
DISPOINC    -1.99255319 -0.0055499173  0.136310637
```

```
D%*%t(X)%*%SALES     #LSE for beta's
```

```
                 [,1]
(Intercept) -68.85707
TARGTPOP      1.45456
DISPOINC      9.36550
```

```
# Load Mass package to construct confidence intervals for
# beta's
confint(fm,level=.95)
```

```
                  2.5 %    97.5 %
(Intercept) -194.9480130 57.233867
TARGTPOP       1.0096226  1.899497
DISPOINC       0.8274411 17.903560
```

```r
confint(fm,level=.9)  # c.f. p. 245 joint conf. int's for beta1 and beta2
```

```
                     5 %      95 %
(Intercept) -172.930288 35.216141
TARGTPOP       1.087317  1.821803
DISPOINC       2.318338 16.412662
```

Estimation of mean response, its confidence intervals and prediction intervals

(Dwaine ex. conti.)

Dwaine Studios as part of a possible expansion program would like to predict sales or expected (mean) sales for two new cities, with City A has $x_{h1} = 65.4$ thousands persons aged 16 or younger and per capita disposable income $x_{h2} = 17.6$ thousands dollars, and City B has $x_{h1} = 53.1$ thousands persons aged 16 or younger and per capita disposable income $x_{h2} = 17.7$ thousands dollars

```r
summary(dum)
```

```
    TARGTPOP        DISPOINC          SALES      
 Min.   :38.40   Min.   :15.80   Min.   :137.2  
 1st Qu.:47.80   1st Qu.:16.30   1st Qu.:152.8  
 Median :52.30   Median :17.10   Median :166.5  
 Mean   :62.02   Mean   :17.14   Mean   :181.9  
 3rd Qu.:82.70   3rd Qu.:18.10   3rd Qu.:209.7  
 Max.   :91.30   Max.   :19.10   Max.   :244.2  
```

```r
new<- data.frame(TARGTPOP=c(65.4,53.1),DISPOINC=c(17.6, 17.7))
new
```

```
  TARGTPOP DISPOINC
1     65.4     17.6
2     53.1     17.7
```

```r
predict(lm(SALES~TARGTPOP+DISPOINC), new, se.fit = TRUE)
```

```
$fit
       1        2 
191.1039 174.1494 
```

```
$se.fit
       1        2 
2.766798 4.598605 

$df
[1] 18

$residual.scale
[1] 11.00739
```

```r
pred.w.plim <- predict(lm(SALES~TARGTPOP+DISPOINC), new, interval="prediction",level=.95)
pred.w.clim <- predict(lm(SALES~TARGTPOP+DISPOINC), new, interval="confidence",level=.95)
pred.w.plim
```

```
       fit      lwr      upr
1 191.1039 167.2589 214.9490
2 174.1494 149.0867 199.2121
```

```r
                #c.f. p. 247 joint 90% prediction intervals for City A and B; notice that
                # level is set to be .95 in the command to get this simultaneous .90 predicti
                # intervals.
pred.w.clim     # p. 246
```

```
       fit      lwr      upr
1 191.1039 185.2911 196.9168
2 174.1494 164.4881 183.8107
```

Note: ANOVA tables in R

$\int_0^\pi sin(x)dx$

```r
# anova(fittedmodel) of R gives the EXTRA SS's, c.f. chapter 7.
# ex. anova(fm) v.s. anova(fmDT) #
fmDT <- lm(SALES~DISPOINC+TARGTPOP, dum)   # regress SALES on those two predictors
                                           # w/ different input order #
anova(fm)
```

```
Analysis of Variance Table

Response: SALES
```

```
          Df  Sum Sq Mean Sq  F value   Pr(>F)
TARGTPOP   1 23371.8 23371.8 192.8962 4.64e-11 ***
DISPOINC   1   643.5   643.5   5.3108  0.03332 *
Residuals 18  2180.9   121.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fmDT)
```

```
Analysis of Variance Table

Response: SALES
          Df  Sum Sq Mean Sq F value    Pr(>F)
DISPOINC   1 18299.8 18299.8 151.035 3.434e-10 ***
TARGTPOP   1  5715.5  5715.5  47.172 2.002e-06 ***
Residuals 18  2180.9   121.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# but, same outputs with summary(fittedmodel) #
summary(fm)
```

```
Call:
lm(formula = SALES ~ TARGTPOP + DISPOINC, data = dum)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4239  -6.2161   0.7449   9.4356  20.2151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
TARGTPOP      1.4546     0.2118   6.868    2e-06 ***
DISPOINC      9.3655     4.0640   2.305   0.0333 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167,    Adjusted R-squared:  0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

```r
summary(fmDT)
```

```

Call:
lm(formula = SALES ~ DISPOINC + TARGTPOP, data = dum)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.4239  -6.2161   0.7449   9.4356  20.2151 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -68.8571    60.0170  -1.147   0.2663    
DISPOINC      9.3655     4.0640   2.305   0.0333 *  
TARGTPOP      1.4546     0.2118   6.868    2e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167,	Adjusted R-squared:  0.9075 
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

More Graphical fun.....

```r
# Or, Load rgl package for 3D plot that can be spinned
#library(rgl)
#plot3d(TARGTPOP,DISPOINC,SALES, col="red", size=3)
```

The `echo: false` option disables the printing of code (only output is displayed).