Introduction to R
Initial Analysis of Diabetes Data

Peter Kempthorne

Spring 2018

Consider the diabetes data used by Efron, Hastie, Johnstone and Tibshirani (2004):[1] observations on 442 patients, with the response of interest being a quantitative measure of disease progression one year after baseline. There are ten baseline variables—age, sex, body-mass index, average blood pressure, and six blood serum measurements.

In this note we use this data to illustrate the following basic R functions:

R Function	Purpose/description
$read.csv()$	Read in data set from a csv file
$dim()$	Display dimensions of matrix/array/data frame
$head()$ and $tail()$	Display top and bottom rows of data
$str()$	Compact display an object's structure
$apply()$	Apply function fixing array margin/dimension
$cbind()$	Bind together column vectors into matrix
$data.frame()$	Create data frame from matrix objects
$mean(), var()$	Compute sample mean and sample variance
$cor()$	Compute correlation matrix
$round(, digits = 2)$	Display values rounding to 2 decimal places
$par(mfcol = c(2, 3))$	Create 2x3 panel for plots (column ordered)
$plot()$	Generic X-Y plot
$lm()$	Fit linear model by least squares
$summary()$	Summary statistics for object (e.g., lm fit)

[1] Least Angle Regression, *Annals of Statistics*, 2004, Vol 32, No. 2 407-499

```
> # 1.  Read data into R ----
> diabetes= read.csv(file="EfronData/diabetes.csv", sep=",", header=TRUE)
> #   Display attributes of data frame
> dim(diabetes)

[1] 442  11

> head(diabetes)

  age sex  bmi map  tc   ldl hdl tch  ltg glu prog
1  59   1 32.1 101 157  93.2  38   4 2.11  87  151
2  48   0 21.6  87 183 103.2  70   3 1.69  69   75
3  72   1 30.5  93 156  93.6  41   4 2.03  85  141
4  24   0 25.3  84 198 131.4  40   5 2.12  89  206
5  50   0 23.0 101 192 125.4  52   4 1.86  80  135
6  23   0 22.6  89 139  64.8  61   2 1.82  68   97

> tail(diabetes)

    age sex  bmi   map  tc   ldl hdl  tch  ltg glu prog
437  33   0 19.5  80.0 171  85.4  75 2.00 1.72  80   48
438  60   1 28.2 112.0 185 113.8  42 4.00 2.16  93  178
439  47   1 24.9  75.0 225 166.0  42 5.00 1.93 102  104
440  60   1 24.9  99.7 162 106.6  43 3.77 1.79  95  132
441  36   0 30.0  95.0 201 125.2  42 4.79 2.23  85  220
442  36   0 19.6  71.0 250 133.2  97 3.00 2.00  92   57

> #   Use str() to display structure
> str(diabetes)

'data.frame':        442 obs. of  11 variables:
 $ age : int  59 48 72 24 50 23 36 66 60 29 ...
 $ sex : int  1 0 1 0 0 0 1 1 1 0 ...
 $ bmi : num  32.1 21.6 30.5 25.3 23 22.6 22 26.2 32.1 30 ...
 $ map : num  101 87 93 84 101 89 90 114 83 85 ...
 $ tc  : int  157 183 156 198 192 139 160 255 179 180 ...
 $ ldl : num  93.2 103.2 93.6 131.4 125.4 ...
 $ hdl : num  38 70 41 40 52 61 50 56 42 43 ...
 $ tch : num  4 3 4 5 4 2 3 4.55 4 4 ...
 $ ltg : num  2.11 1.69 2.03 2.12 1.86 1.82 1.72 1.85 1.94 2.34 ...
 $ glu : int  87 69 85 89 80 68 82 92 94 88 ...
 $ prog: int  151 75 141 206 135 97 138 63 110 310 ...

> # 2. Compute summary statistics ----
> apply(diabetes,2,summary)

          age    sex   bmi    map    tc    ldl   hdl  tch   ltg    glu  prog
Min.    19.00 0.0000 18.00  62.00  97.0  41.60 22.00 2.00 1.410  58.00  25.0
```

```
1st Qu. 38.25 0.0000 23.20  84.00 164.2  96.05 40.25 3.00 1.860  83.25  87.0
Median  50.00 0.0000 25.70  93.00 186.0 113.00 48.00 4.00 2.005  91.00 140.5
Mean    48.52 0.4683 26.38  94.65 189.1 115.40 49.79 4.07 2.016  91.26 152.1
3rd Qu. 59.00 1.0000 29.28 105.00 209.8 134.50 57.75 5.00 2.170  98.00 211.5
Max.    79.00 1.0000 42.20 133.00 301.0 242.40 99.00 9.09 2.650 124.00 346.0
```

```
> cbind(mean=apply(diabetes,2,mean),
+       sd=sqrt(apply(diabetes,2,var)))
```

```
            mean         sd
age   48.5180995 13.1090278
sex    0.4683258  0.4995612
bmi   26.3757919  4.4181216
map   94.6466063 13.8319998
tc   189.1402715 34.6080517
ldl  115.4391403 30.4130810
hdl   49.7884615 12.9342022
tch    4.0702489  1.2904499
ltg    2.0157466  0.2270465
glu   91.2601810 11.4963347
prog 152.1334842 77.0930045
```

```
> round(cor(diabetes),digits=2)
```

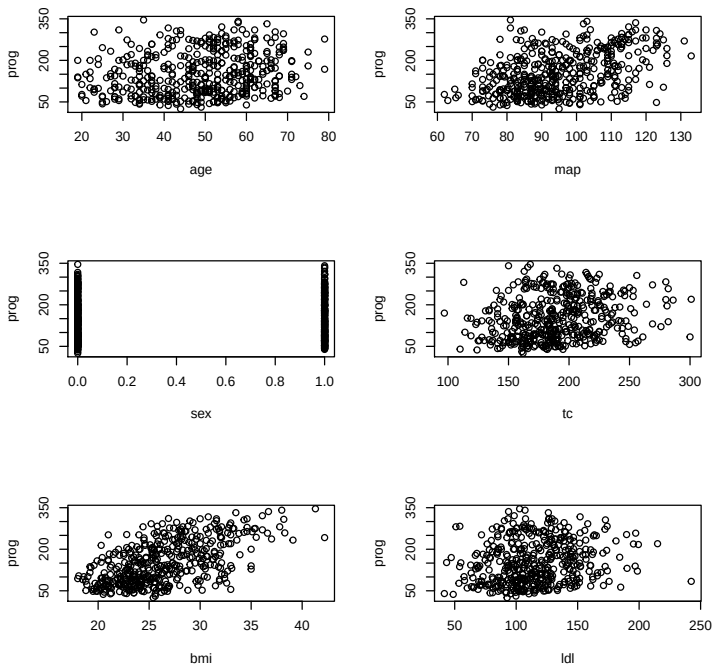
```
       age   sex   bmi   map   tc   ldl   hdl   tch   ltg   glu  prog
age   1.00  0.17  0.19  0.34 0.26  0.22 -0.08  0.20  0.27  0.30  0.19
sex   0.17  1.00  0.09  0.24 0.04  0.14 -0.38  0.33  0.15  0.21  0.04
bmi   0.19  0.09  1.00  0.40 0.25  0.26 -0.37  0.41  0.45  0.39  0.59
map   0.34  0.24  0.40  1.00 0.24  0.19 -0.18  0.26  0.39  0.39  0.44
tc    0.26  0.04  0.25  0.24 1.00  0.90  0.05  0.54  0.52  0.33  0.21
ldl   0.22  0.14  0.26  0.19 0.90  1.00 -0.20  0.66  0.32  0.29  0.17
hdl  -0.08 -0.38 -0.37 -0.18 0.05 -0.20  1.00 -0.74 -0.40 -0.27 -0.39
tch   0.20  0.33  0.41  0.26 0.54  0.66 -0.74  1.00  0.62  0.42  0.43
ltg   0.27  0.15  0.45  0.39 0.52  0.32 -0.40  0.62  1.00  0.46  0.57
glu   0.30  0.21  0.39  0.39 0.33  0.29 -0.27  0.42  0.46  1.00  0.38
prog  0.19  0.04  0.59  0.44 0.21  0.17 -0.39  0.43  0.57  0.38  1.00
```

```
> par(mfcol=c(3,2))
> names(diabetes)
```

```
 [1] "age"  "sex"  "bmi"  "map"  "tc"   "ldl"  "hdl"  "tch"  "ltg"  "glu"
[11] "prog"
```

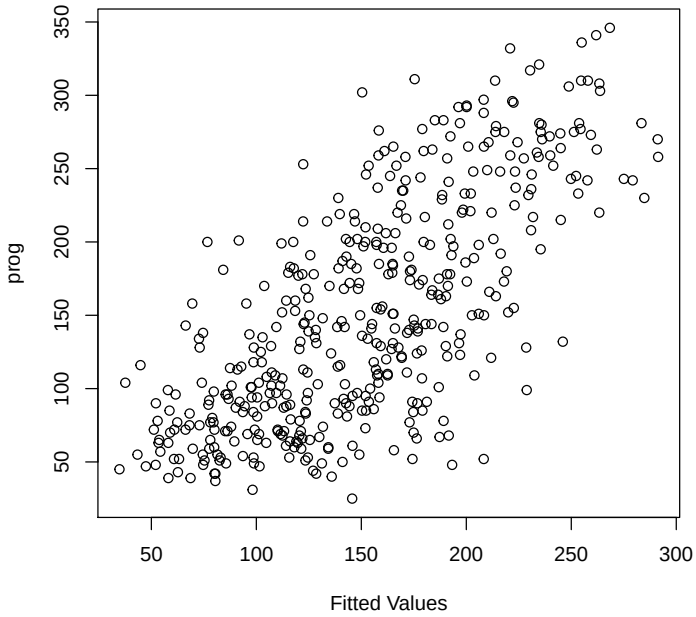
```
> #   Plot prog versus age, bmi
> plot(prog ~ age, data=diabetes)
> plot(prog ~ sex, data=diabetes)
> plot(prog ~ bmi, data=diabetes)
```

```
> plot(prog ~ map, data=diabetes)
> plot(prog ~ tc, data=diabetes)
> plot(prog ~ ldl, data=diabetes)
```




```
> # 3.  Fit linear regression model ----
> lmfit<-lm(prog ~ ., data=diabetes)
> summary(lmfit)

Call:
lm(formula = prog ~ ., data = diabetes)

Residuals:
     Min       1Q   Median       3Q      Max
-156.308  -38.402   -0.727   38.003  151.606

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -356.64395   67.01983  -5.321 1.66e-07 ***
age           -0.03529    0.21705  -0.163 0.870910
sex          -22.79233    5.83657  -3.905 0.000109 ***
bmi            5.59548    0.71746   7.799 4.75e-14 ***
map            1.11589    0.22526   4.954 1.05e-06 ***
```

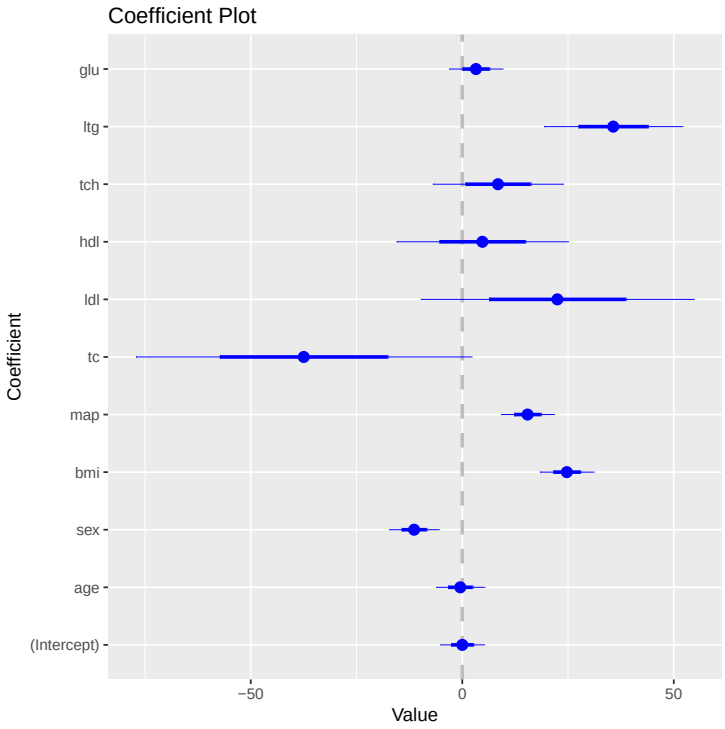
```
tc            -1.08286    0.57294  -1.890 0.059428 .
ldl            0.73914    0.53032   1.394 0.164108
hdl            0.36783    0.78274   0.470 0.638648
tch            6.54048    5.95956   1.097 0.273045
ltg          157.17606   36.04811   4.360 1.63e-05 ***
glu            0.28148    0.27332   1.030 0.303661
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.16 on 431 degrees of freedom
Multiple R-squared:  0.5176,        Adjusted R-squared:  0.5065
F-statistic: 46.25 on 10 and 431 DF,  p-value: < 2.2e-16

> # plot prog versus fitted values
> # Refit model with helpful options in lm()
> lmfit<-lm(prog ~ ., data=diabetes,x=TRUE,y=TRUE)
> par(mfcol=c(1,1))
> plot(x=lmfit$fitted.values,y=lmfit$y,
+      xlab="Fitted Values", ylab="prog")
> cor(cbind(lmfit$fitted.values, lmfit$y))

          [,1]      [,2]
[1,] 1.0000000 0.7194735
[2,] 0.7194735 1.0000000
```



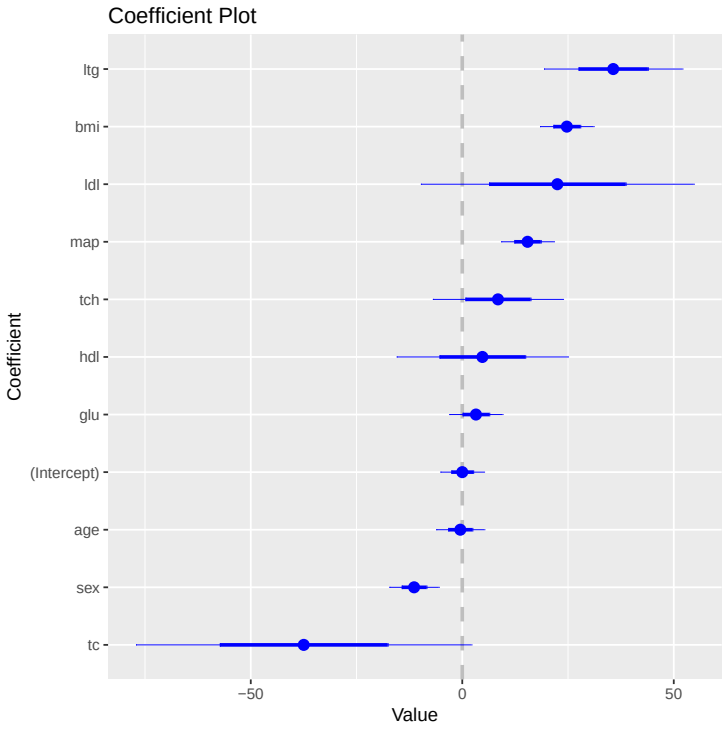

```
> # Rescale variables to have mean 0, sd 1
> diabetes0=diabetes
> names(diabetes0)

 [1] "age"  "sex"  "bmi"  "map"  "tc"   "ldl"  "hdl"  "tch"  "ltg"  "glu"
[11] "prog"

> diabetes0=apply(diabetes,2,scale)
> apply(diabetes0,2,summary)

               age        sex        bmi        map         tc        ldl
Min.    -2.252e+00 -9.375e-01 -1.896e+00 -2.360e+00 -2.662e+00 -2.428e+00
1st Qu. -7.833e-01 -9.375e-01 -7.188e-01 -7.697e-01 -7.192e-01 -6.375e-01
Median   1.130e-01 -9.375e-01 -1.530e-01 -1.190e-01 -9.074e-02 -8.020e-02
Mean     7.000e-18 -1.368e-18  1.072e-16 -4.779e-16 -2.912e-16 -1.115e-16
3rd Qu.  7.996e-01  1.064e+00  6.562e-01  7.485e-01  5.955e-01  6.267e-01
Max.     2.325e+00  1.064e+00  3.582e+00  2.773e+00  3.232e+00  4.175e+00
               hdl        tch        ltg        glu       prog
Min.    -2.148e+00 -1.604e+00 -2.668e+00 -2.893e+00 -1.649e+00
1st Qu. -7.375e-01 -8.294e-01 -6.860e-01 -6.968e-01 -8.449e-01
Median  -1.383e-01 -5.444e-02 -4.733e-02 -2.263e-02 -1.509e-01
Mean    -1.192e-16 -1.415e-16  6.008e-16  2.365e-16 -1.490e-16
```

```
3rd Qu.  6.155e-01  7.205e-01  6.794e-01  5.863e-01  7.701e-01
Max.     3.805e+00  3.890e+00  2.793e+00  2.848e+00  2.515e+00

> apply(diabetes0,2,mean)

          age           sex           bmi           map            tc
 6.999972e-18 -1.368253e-18  1.072178e-16 -4.778962e-16 -2.911564e-16
          ldl           hdl           tch           ltg           glu
-1.114637e-16 -1.191685e-16 -1.414922e-16  6.008286e-16  2.365483e-16
         prog
-1.490189e-16

> apply(diabetes0,2,var)

 age  sex  bmi  map   tc  ldl  hdl  tch  ltg  glu prog
   1    1    1    1    1    1    1    1    1    1    1

> # Coerce matrix diabetes0 to be a data frame
> # Replace the (scaled) prog variable with
> #     the mean-adjusted original
> diabetes0=data.frame(diabetes0)
> diabetes0$prog=diabetes$prog - mean(diabetes$prog)
> # Check that means are 0 and variances are apply(diabetes0,2,mean)
> apply(diabetes0,2,mean)

          age           sex           bmi           map            tc
 6.999972e-18 -1.368253e-18  1.072178e-16 -4.778962e-16 -2.911564e-16
          ldl           hdl           tch           ltg           glu
-1.114637e-16 -1.191685e-16 -1.414922e-16  6.008286e-16  2.365483e-16
         prog
-1.195733e-14

> apply(diabetes0,2,var)

     age      sex      bmi      map       tc      ldl      hdl      tch
   1.000    1.000    1.000    1.000    1.000    1.000    1.000    1.000
     ltg      glu     prog
   1.000    1.000 5943.331

> # 3.1 Refit linear model with scaled indep vars ----
>
> lmfit0=lm(prog~., data=diabetes0)
> summary(lmfit0)

Call:
lm(formula = prog ~ ., data = diabetes0)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-156.308  -38.402   -0.727   38.003  151.606

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.004e-14  2.576e+00   0.000 1.000000
age         -4.627e-01  2.845e+00  -0.163 0.870910
sex         -1.139e+01  2.916e+00  -3.905 0.000109 ***
bmi          2.472e+01  3.170e+00   7.799 4.75e-14 ***
map          1.544e+01  3.116e+00   4.954 1.05e-06 ***
tc          -3.748e+01  1.983e+01  -1.890 0.059428 .
ldl          2.248e+01  1.613e+01   1.394 0.164108
hdl          4.758e+00  1.012e+01   0.470 0.638648
tch          8.440e+00  7.691e+00   1.097 0.273045
ltg          3.569e+01  8.185e+00   4.360 1.63e-05 ***
glu          3.236e+00  3.142e+00   1.030 0.303661
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.16 on 431 degrees of freedom
Multiple R-squared:  0.5176,        Adjusted R-squared:  0.5065
F-statistic: 46.25 on 10 and 431 DF,  p-value: < 2.2e-16

>
> #   Note what statistics are same:
> #       Residual standard error
> #       Multiple R-squared
> #       t values and p-values of indep vars

> # 4.  Display coefficients (estimates and confidence intervals) ----
> # Replace FALSE by TRUE if package coefplot needs to be installed
> if (FALSE){install.packages(coefplot)}
> library(coefplot)
> coefplot(lmfit0)
```




```
> # Re-do coefplot ordered by magnitude of coefficient
> coefplot(lmfit0, sort='magnitude')
>
```


Coefficient Plot
ltg
bmi
ldl
map
tch
hdl
glu
(Intercept)
age
sex
tc
Coefficient
−50
0
50
Value