Section 3.5 Correlated Predictors

Load needed packages.

```
library(Stat2Data)
library(mosaic)
library(ggplot2)
library(car) #used to get VIFs
```

EXAMPLE 3.15 House Prices in NY

Create a dataframe for **HousesNY** and look at the structure of the data.

```
data("HousesNY")
str(HousesNY)
```

```
## 'data.frame':    53 obs. of  5 variables:
##  $ Price: num  57.6 120 150 143 92.5 ...
##  $ Beds : int  3 6 4 3 3 2 2 4 4 3 ...
##  $ Baths: num  2 2 2 2 1 1 2 3 2.5 2 ...
##  $ Size : num  0.96 2.79 1.7 1.2 1.33 ...
##  $ Lot  : num  1.3 0.23 0.27 0.8 0.42 0.34 0.29 0.21 1 0.3 ...
```

EXAMPLE 3.15 CHOOSE

```
cor(HousesNY$Price, HousesNY$Size)
```

```
## [1] 0.5121029
```

```
cor(HousesNY$Price, HousesNY$Beds)
```

```
## [1] 0.4191355
```

EXAMPLE 3.15 FIT simple linear regression models

```
modelSize=lm(Price~Size,data=HousesNY)
summary(modelSize)
```

```
##
## Call:
## lm(formula = Price ~ Size, data = HousesNY)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.572 -28.829  -1.346  29.430  75.338
```

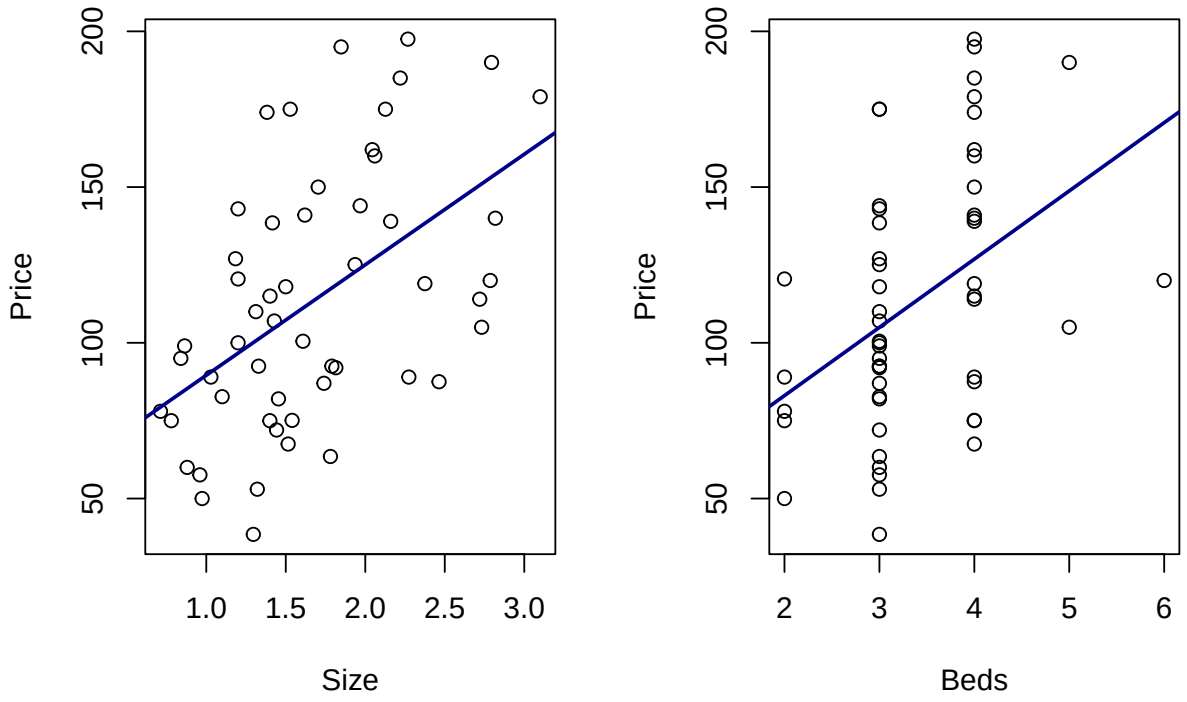
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.078     14.832   3.646 0.000625 ***
## Size          35.489      8.335   4.258 8.87e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.93 on 51 degrees of freedom
## Multiple R-squared:  0.2622, Adjusted R-squared:  0.2478
## F-statistic: 18.13 on 1 and 51 DF,  p-value: 8.865e-05
```

```
modelBeds=lm(Price~Beds,data=HousesNY)
summary(modelBeds)
```

```
##
## Call:
## lm(formula = Price ~ Beds, data = HousesNY)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.453 -32.953  -5.048  33.142  70.642
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.239     23.161   1.694  0.09632 .
## Beds          21.905      6.644   3.297  0.00179 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.98 on 51 degrees of freedom
## Multiple R-squared:  0.1757, Adjusted R-squared:  0.1595
## F-statistic: 10.87 on 1 and 51 DF,  p-value: 0.001785
```

FIGURE 3.28 Scatterplots of house selling price versus two predictors

```
par(mfrow=c(1,2))
par(mar=c(8,5,2,1))
plot(Price~Size,data=HousesNY)
abline(modelSize,lwd=2,col="darkblue")
plot(Price~Beds,data=HousesNY)
abline(modelBeds,lwd=2,col="darkblue")
```




```
layout(mat=c(1,1))
```

EXAMPLE 3.15 FIT multiple regression model based on both Size and Beds

```
modelSizeBeds=lm(Price~Size+Beds, data=HousesNY)
summary(modelSizeBeds)
```

```
##
## Call:
## lm(formula = Price ~ Size + Beds, data = HousesNY)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.493 -31.920   1.696  27.866  73.436
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   46.498     22.277   2.087    0.042 *
## Size          31.169     12.617   2.470    0.017 *
## Beds           4.367      9.515   0.459    0.648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.21 on 50 degrees of freedom
```

```
## Multiple R-squared:  0.2653, Adjusted R-squared:  0.236
## F-statistic:  9.03 on 2 and 50 DF,  p-value: 0.0004489
```

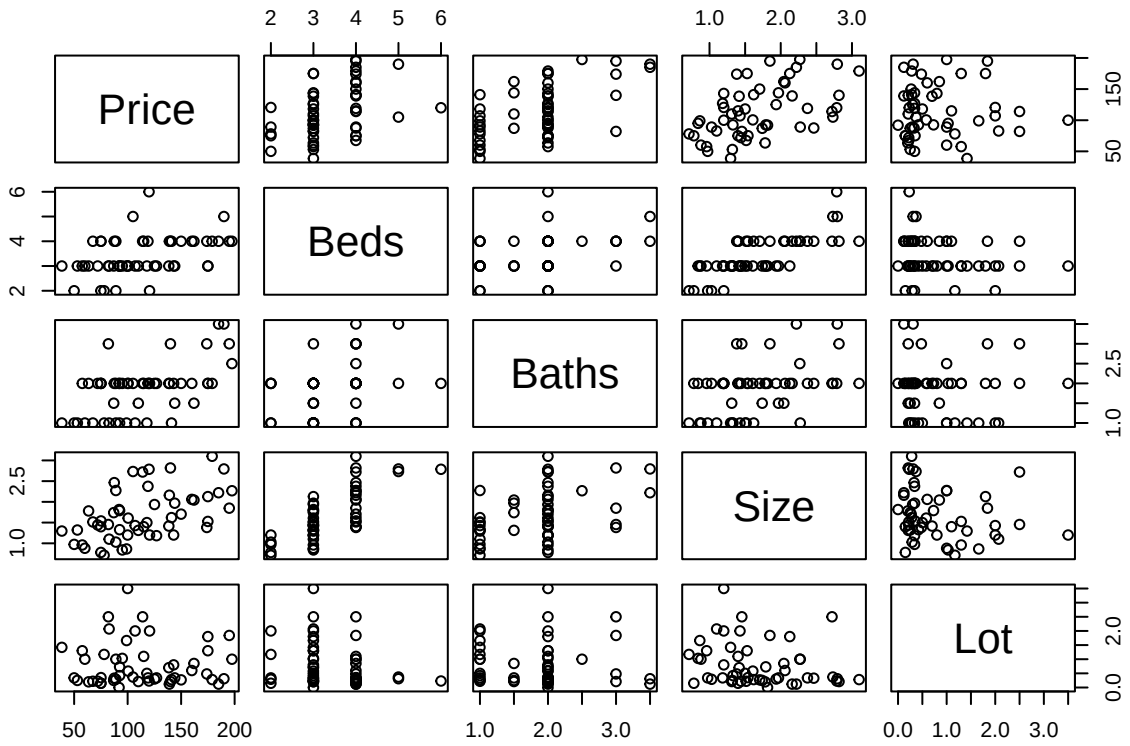
EXAMPLE 3.16 House prices in NY: correlations

```
round(cor(HousesNY),digits=3)
```

```
##         Price   Beds  Baths   Size    Lot
## Price   1.000  0.419  0.558  0.512 -0.011
## Beds    0.419  1.000  0.356  0.746 -0.211
## Baths   0.558  0.356  1.000  0.418 -0.039
## Size    0.512  0.746  0.418  1.000 -0.214
## Lot    -0.011 -0.211 -0.039 -0.214  1.000
```

FIGURE 3.29 Matrix of scatterplots for variables in **HousesNY**

```
pairs(HousesNY)
```



EXAMPLE 3.16 FIT a multiple regression model with all four predcitors

```
modelall4=lm(Price~Size+Beds+Baths+Lot, data=HousesNY)
summary(modelall4)
```

```
##
```

```
## Call:
## lm(formula = Price ~ Size + Beds + Baths + Lot, data = HousesNY)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.384 -25.910  -0.377  28.515  59.761
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.590     23.266   0.627   0.5336
## Size          22.155     11.931   1.857   0.0695 .
## Beds           2.771      8.730   0.317   0.7523
## Baths         26.238      7.844   3.345   0.0016 **
## Lot            4.621      6.184   0.747   0.4585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.03 on 48 degrees of freedom
## Multiple R-squared:  0.4134, Adjusted R-squared:  0.3645
## F-statistic: 8.457 on 4 and 48 DF,  p-value: 3.01e-05
```

```
anova(modelall4)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Size       1  23407 23407.2 21.4589 2.785e-05 ***
## Beds       1    276   276.2  0.2532  0.617125
## Baths      1  12605 12605.0 11.5558  0.001368 **
## Lot        1    609   609.1  0.5584  0.458536
## Residuals 48  52358  1090.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

EXAMPLE 3.17 Doctors and hospitals in counties: VIF

Create a dataframe for CountyHealth and look at the structure of the data.

```
data("CountyHealth")
str(CountyHealth)
```

```
## 'data.frame':    53 obs. of  4 variables:
##  $ County   : Factor w/ 53 levels "Bay, FL                                ",..: 1 2 3 4 5 6 7 8 9 10 ..
##  $ MDs      : int  351 95 260 2797 769 42 13 20 2981 83 ...
##  $ Hospitals: int  3 2 2 11 5 2 2 2 7 3 ...
##  $ Beds     : int  605 134 567 1435 976 245 33 65 1462 100 ...
```

EXAMPLE 3.17 FIT a multiple regression model for predicting sqrtMDs

```
CountyHealth$TMDs=sqrt(CountyHealth$MDs)
regmodel2=lm(TMDs~Hospitals+Beds+Hospitals*Beds, data=CountyHealth)
summary(regmodel2)
```

```
##
## Call:
## lm(formula = TMDs ~ Hospitals + Beds + Hospitals * Beds, data = CountyHealth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5188  -3.5805  -0.8423   3.4058  14.9756
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.6247854  1.7287217  -0.361    0.719
## Hospitals       3.1420035  0.6101592   5.149 4.62e-06 ***
## Beds            0.0218939  0.0025979   8.428 4.27e-11 ***
## Hospitals:Beds -0.0009755  0.0002116  -4.610 2.90e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.367 on 49 degrees of freedom
## Multiple R-squared:  0.9454, Adjusted R-squared:  0.9421
## F-statistic: 282.9 on 3 and 49 DF,  p-value: < 2.2e-16
```

Getting the variance inflation factors with the car package.

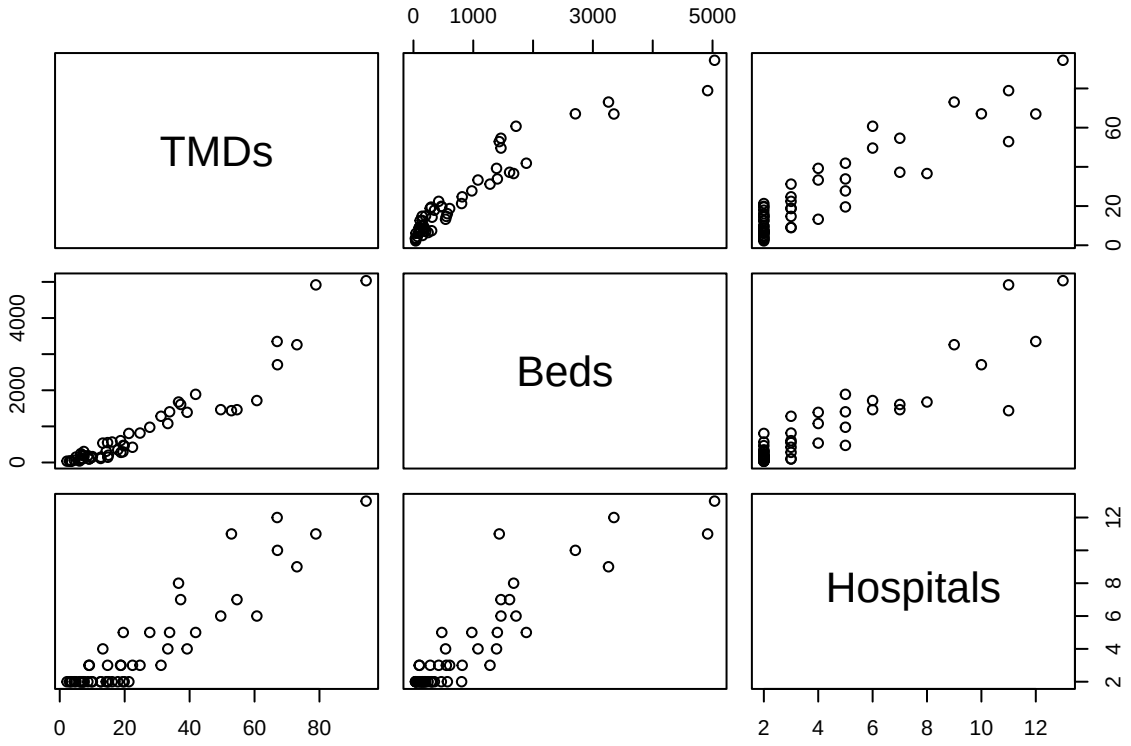
```
vif(regmodel2)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##      Hospitals           Beds Hospitals:Beds
##       6.022161      16.092174      14.418667
```

FIGURE 3.30 Scatterplot matrix of sqrt(MDs), hospitals, and hospital beds for 53 counties Note that the sqrt(MDs) is called TMDs below, since we used this code in Chapter 1.

```
pairs(CountyHealth[,5:3])
```


TMDs
Beds
Hospitals
0
1000
3000
5000
0
20
60
0
2000
4000
2
4
6
8
10
12
0
20
40
60
80
2
4
6
8
10
12