Aspects of Multivariate Analysis

08/02/2023

Jing Qin

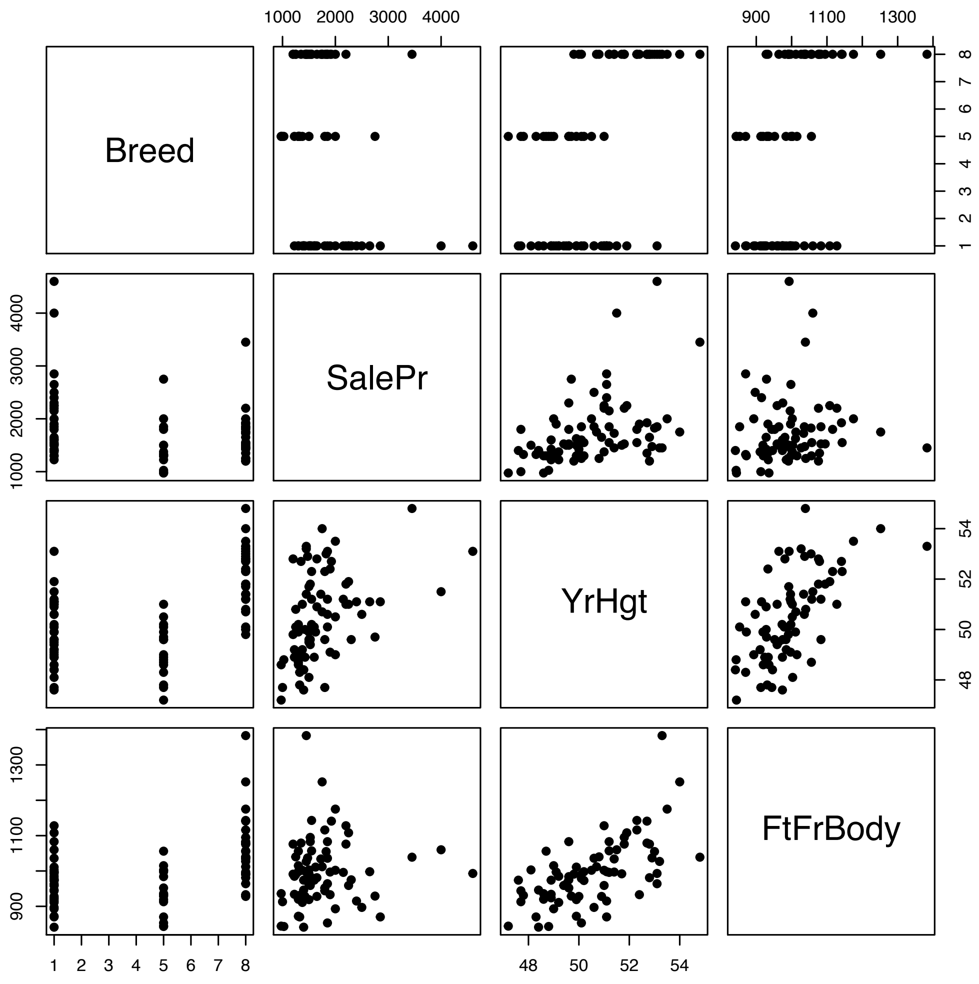
Data of the day

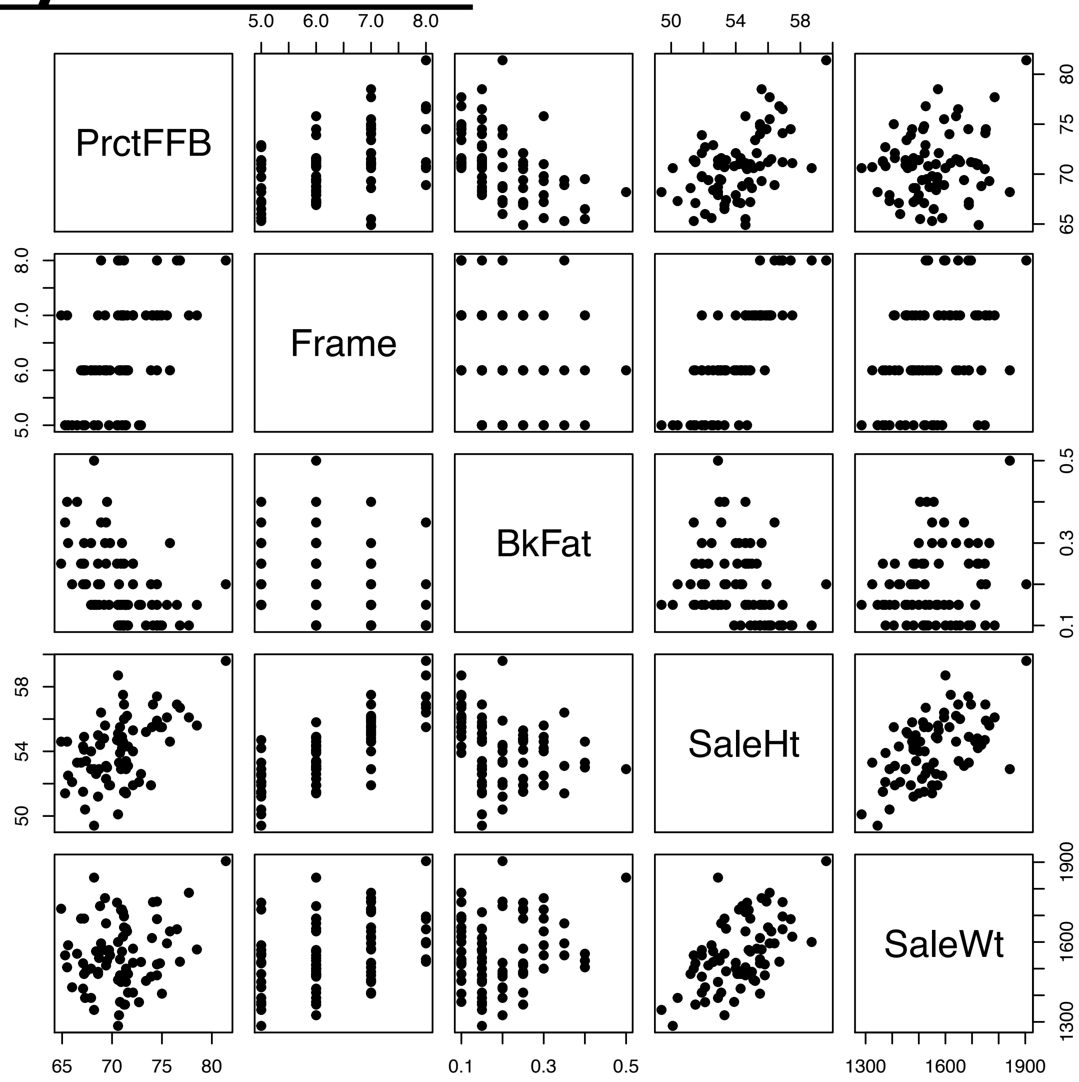
- Bull data T1-10.dat Information see Ex1.26, Table 1.10

  Rscript with Bull.R

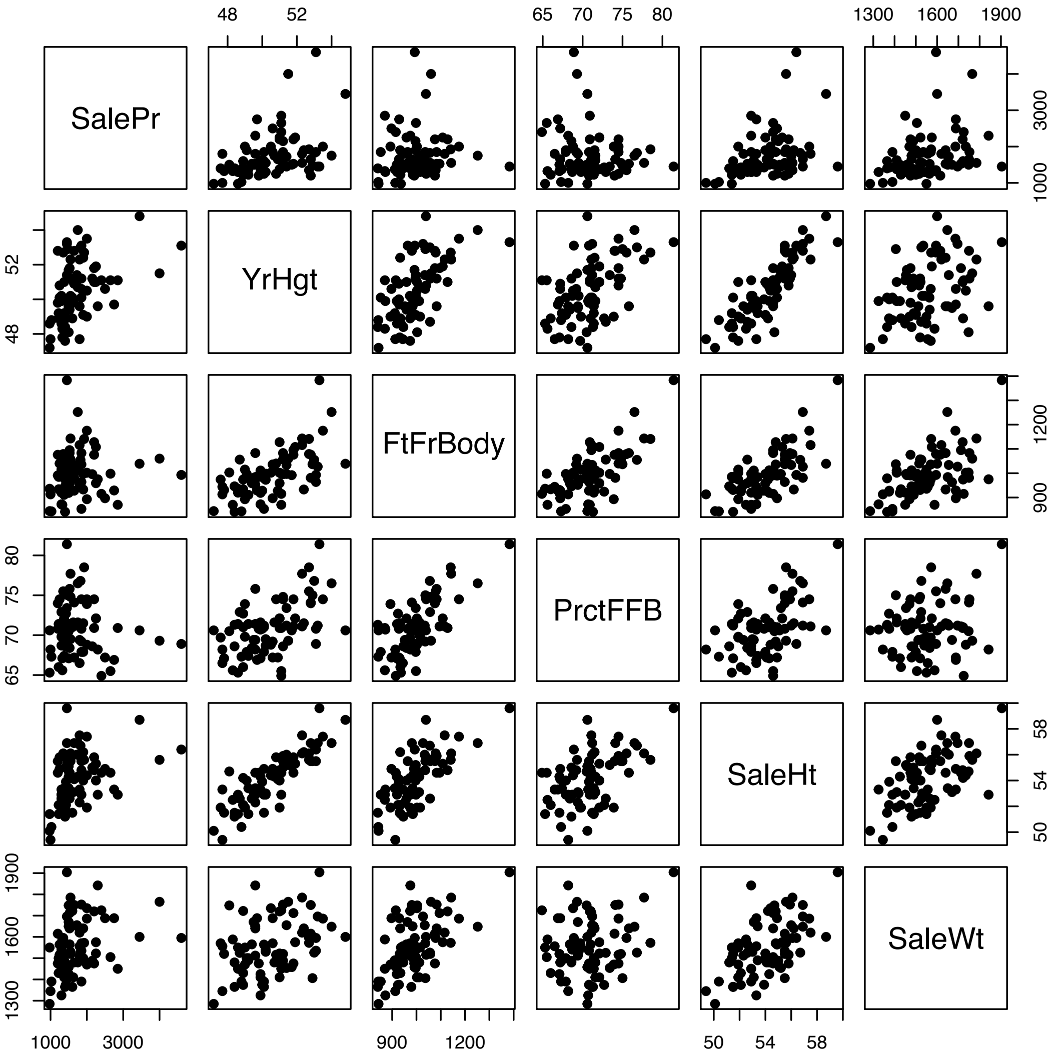
```
> head(dfBull)
  Breed SalePr YrHgt FtFrBody PrctFFB Frame BkFat SaleHt SaleWt
1     1   2200  51.0     1128    70.9     7  0.25   54.8   1720
2     1   2250  51.9     1108    72.1     7  0.25   55.3   1575
3     1   1625  49.9     1011    71.6     6  0.15   53.1   1410
4     1   4600  53.1      993    68.9     8  0.35   56.4   1595
5     1   2150  51.2      996    68.6     7  0.25   55.0   1488
6     1   1225  49.2      985    71.4     6  0.15   51.4   1500
>
```

Always have a look at your data





Numerical attributes




```
> head(dfBullnum)
  SalePr YrHgt FtFrBody PrctFFB SaleHt SaleWt
1   2200  51.0     1128    70.9   54.8   1720
2   2250  51.9     1108    72.1   55.3   1575
3   1625  49.9     1011    71.6   53.1   1410
4   4600  53.1      993    68.9   56.4   1595
5   2150  51.2      996    68.6   55.0   1488
6   1225  49.2      985    71.4   51.4   1500
```

	V1 (x_1)	V2	V3	V4	V5	V6
	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}

Data matrix: an $(n \times p)$-matrix

Data matrix one-step back to notations

$$
\mathbf{X} = \begin{array}{c}
\\
\text{Item\_1} \\ \text{Item\_2} \\ \vdots \\ \text{Item\_}j\,(\boldsymbol{x}_j^T) \\ \vdots \\ \text{Item\_}n
\end{array}
\begin{array}{c}
\begin{array}{cccccc} \text{Attribute\_1} & \text{Attribute\_2} & \cdots & \text{Attribute\_}k\,(x_k) & \cdots & \text{Attribute\_}p \end{array} \\
\left[\begin{array}{cccccc}
x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\
\vdots & \vdots & & \vdots & \cdots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np}
\end{array}\right]
\end{array}
$$

The k-th attribute is denoted by x_k

Data vector from j-th individual is denoted by $\boldsymbol{x}_j$ (bold) as a column **vector**

Sample mean vector

one-step forward to generality

```
> colMeans(dfBullnum)
    SalePr      YrHgt   FtFrBody    PrctFFB     SaleHt     SaleWt 
1742.43421   50.52237  995.94737   70.88158   54.12632 1555.28947 
```

$$
\begin{array}{cccccc}
\text{Attribute\_1} & \text{Attribute\_2} & \cdots & \text{Attribute\_}k(x_k) & \cdots & \text{Attribute\_}p
\end{array}
$$

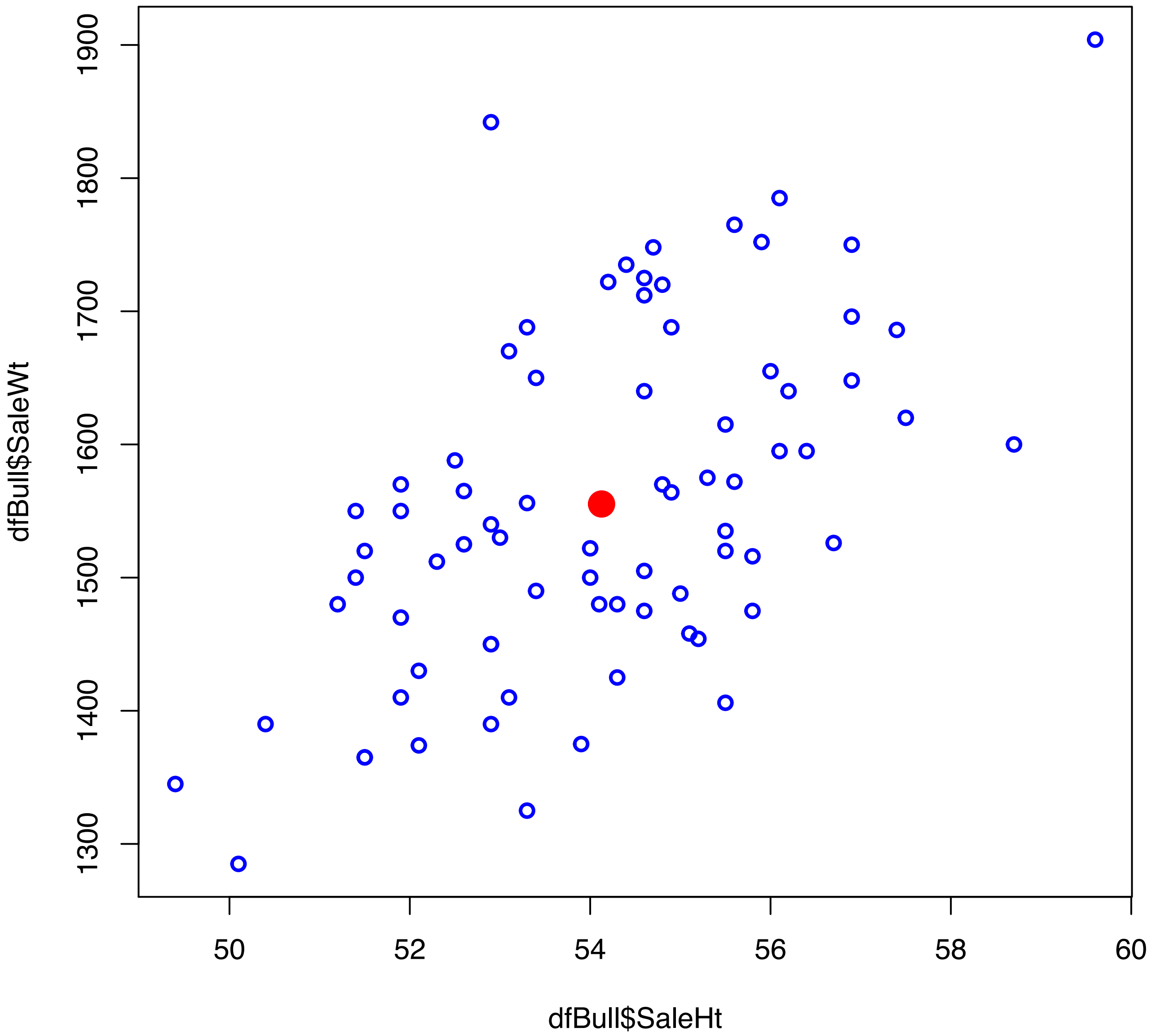
$$
\begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\
\vdots & \vdots & & \vdots & \cdots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np}
\end{bmatrix}
$$

$$\bar{x}_1 = \frac{1}{n}\sum_{j=1}^{n} x_{j1} \qquad \cdots \qquad \bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \qquad (1\text{-}1)$$

Sample mean vector center of data

$$\overline{\boldsymbol{x}} = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_k \\ \vdots \\ \overline{x}_p \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{j=1}^{n} x_{j1} \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n} x_{jk} \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n} x_{jp} \end{pmatrix}.$$

$$\overline{\boldsymbol{x}}^T = (\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_p)^T$$



Covariance and sample covariance

- Re-cap: Given two random variables (r.v.'s) Y and Z, we know population covariance of Y and Z is

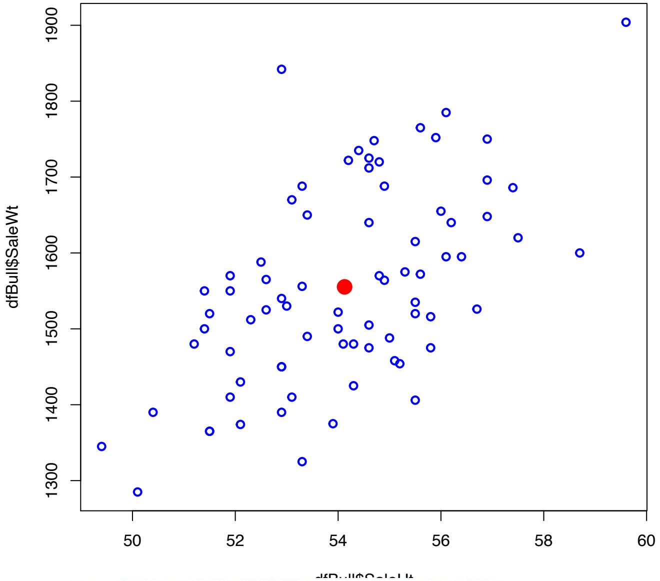
$$Cov(Y,Z) = E[(Y - E(Y)) \cdot (Z - E(Z))]$$

- Given pairs of observations accordingly, the sample covariance is

$$\begin{pmatrix} y_1 & z_1 \\ y_2 & z_2 \\ y_3 & z_3 \\ & \cdots & \\ y_n & z_n \end{pmatrix} \quad \rightarrow \quad \frac{1}{n-1}\sum_{j=1}^{n}\left[(y_j - \bar{y}) \cdot (z_j - \bar{z})\right]$$

- Note that $\frac{1}{n-1}\sum_{j=1}^{n}\left[(y_j - \bar{y}) \cdot (y_j - \bar{y})\right]$ is the *sample variance*

Sample covariance matrix (1-4) symmetric




```
> cov(dfBull$SaleHt, dfBull$SaleWt)
[1] 147.2896
```

$$s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} \left[(x_{ji} - \overline{x_i}) \cdot (x_{jk} - \overline{x_k}) \right]$$

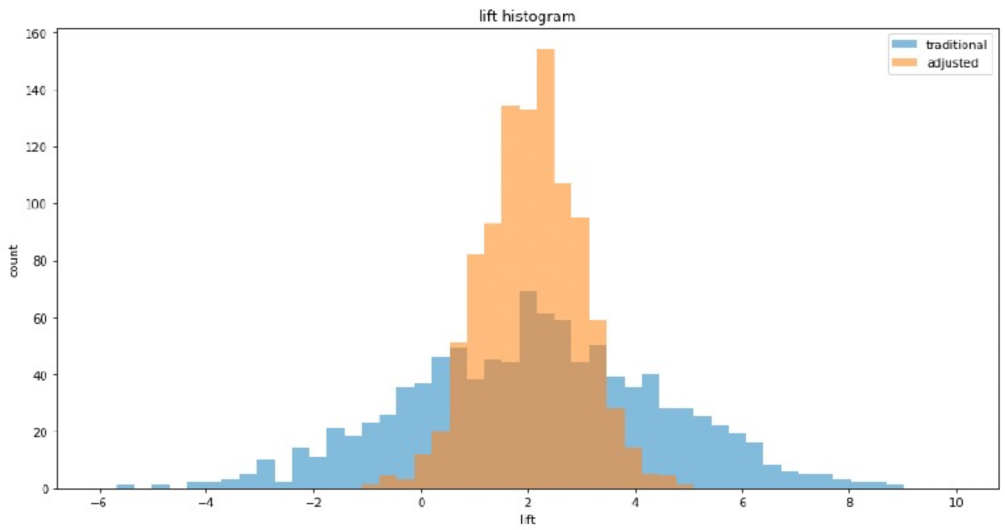
Repeat the calculation for all pairs of numerical attributes

```
> cov(dfBullnum)
              SalePr       YrHgt   FtFrBody     PrctFFB      SaleHt      SaleWt
SalePr   388133.6623  456.471491  5890.5965 -229.474561  486.968421 25645.88596
YrHgt       456.4715    2.998026   100.1305    2.960018    2.983137    82.81077
FtFrBody   5890.5965  100.130526  8594.3439  209.504351  129.940070  6680.30877
PrctFFB    -229.4746    2.960018   209.5044   10.691656    3.414225    83.92540
SaleHt      486.9684    2.983137   129.9401    3.414225    4.017965   147.28961
SaleWt    25645.8860   82.810772  6680.3088   83.925404  147.289614 16850.66175
```

Sample Variance: how data spreads

Elements on the diagonal

R cmd: `var()`



$1/n$ or $1/(n-1)$?

More commonly, one refers to $\frac{1}{n-1}(**)$ as *sample variance* and it can be calculated with R cmd `var()`.

Sample covariance matrix down to a number

- Generalized sample variance (3-12) determinant(covariance-matrix)

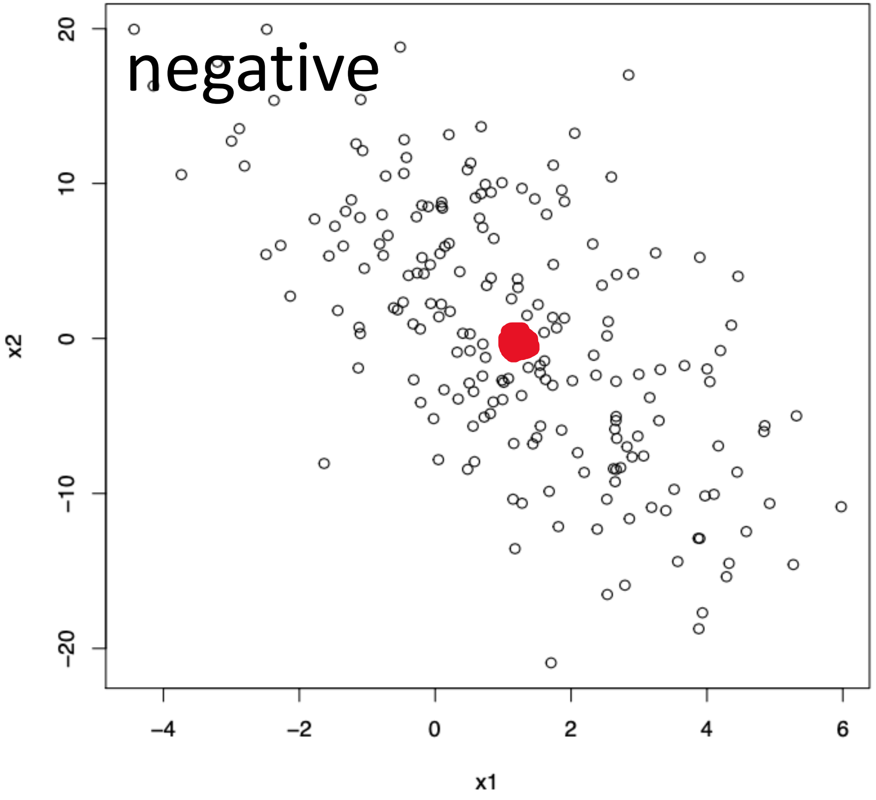
```
> det(cov(dfBullnum))
[1] 1.558357e+14
```

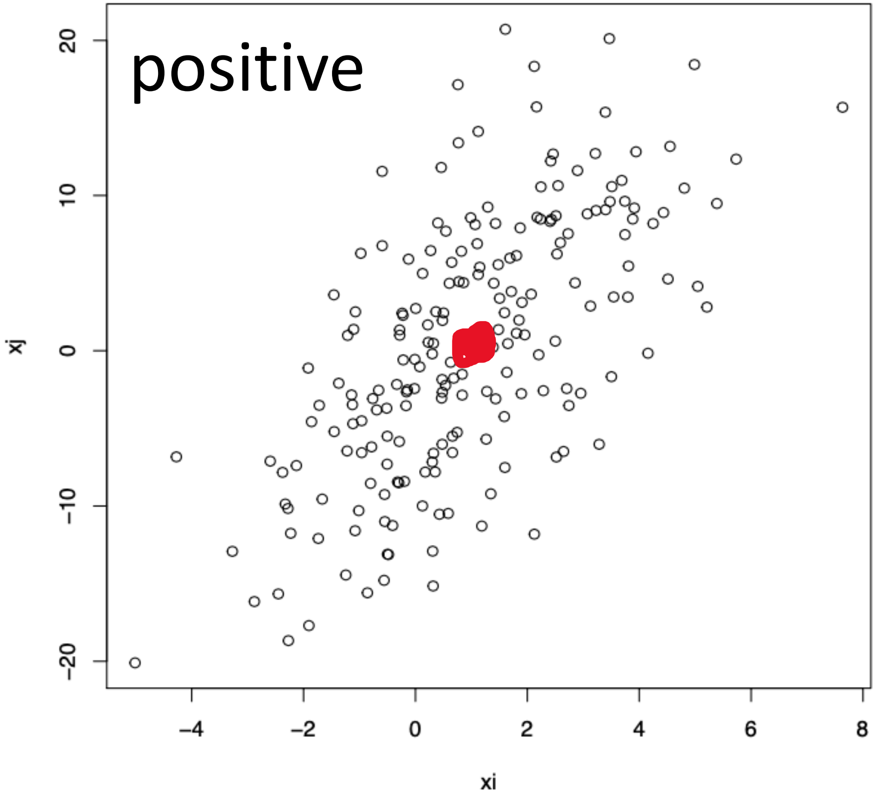
- Total sample variance (3-23) sum of the diagonal elements

```
> sum(diag(cov(dfBullnum)))
[1] 413596.4
```

More on sample covariance

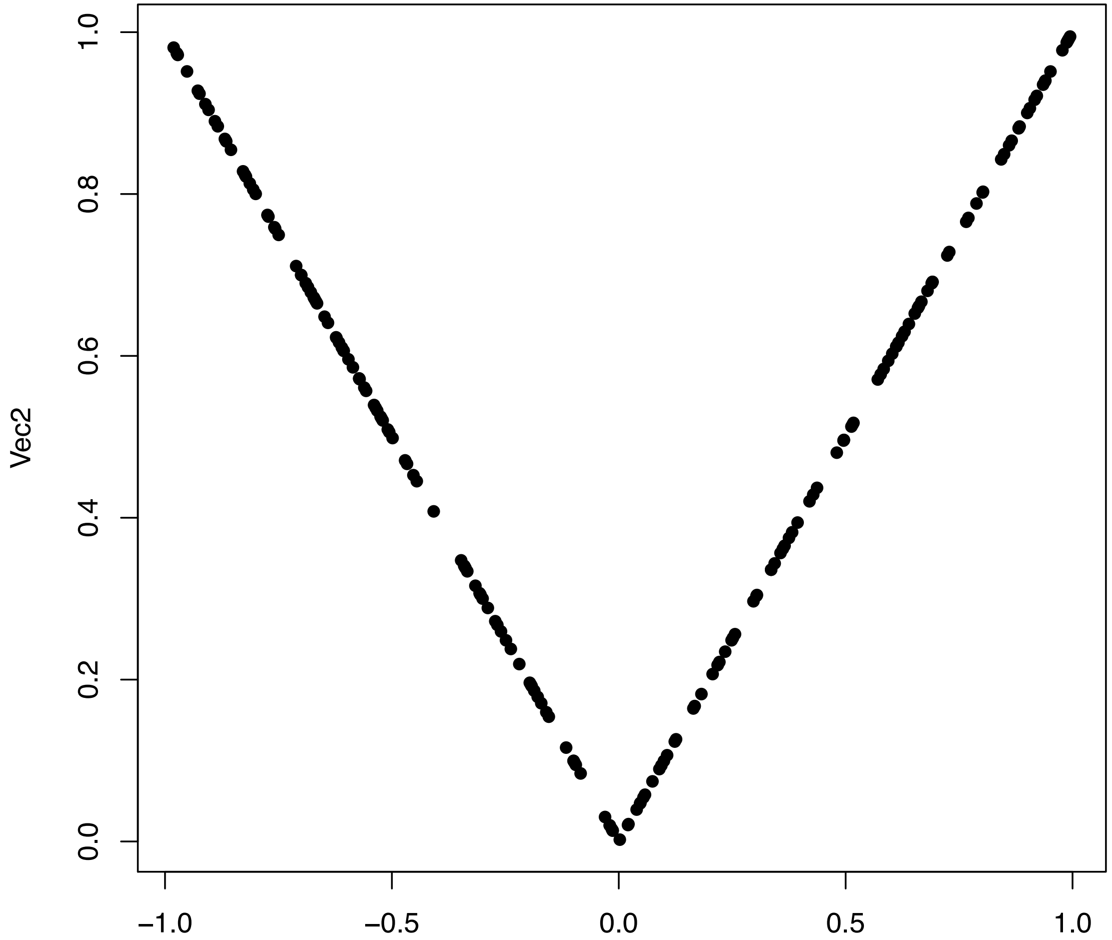
1. Linear correlation





More on sample covariance

2. Uncorrelated vs independent




```
> cov(Vec1, Vec2)
[1] 0.004503598
```

More on sample covariance

3. Depend on the scale of the data $s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} [(x_{ji} - \overline{x_i}) \cdot (x_{jk} - \overline{x_k})]$

```
> cov(dfBull$SaleHt, dfBull$SaleWt)
[1] 147.2896
> cov(dfBull$SaleHt, dfBull$SaleWt/1000)
[1] 0.1472896
```

Easy to fix with sample correlation (1-5) $r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}.$

```
> cor(dfBull$SaleHt, dfBull$SaleWt)
[1] 0.5660575
> cor(dfBull$SaleHt, dfBull$SaleWt/1000)
[1] 0.5660575
```

Sample correlation matrix free-of-scale, symmetric

1. Matrix element in between [-1, 1] $r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}.$

2. When $i = k, r_{ii} = 1$

3.
```
> cor(dfBullnum)
            SalePr     YrHgt  FtFrBody    PrctFFB    SaleHt    SaleWt
SalePr    1.0000000 0.4231607 0.1019911 -0.1126475 0.3899483 0.3171163
YrHgt     0.4231607 1.0000000 0.6237958  0.5228223 0.8595129 0.3684348
FtFrBody  0.1019911 0.6237958 1.0000000  0.6911371 0.6992519 0.5551134
PrctFFB  -0.1126475 0.5228223 0.6911371  1.0000000 0.5209146 0.1977254
SaleHt    0.3899483 0.8595129 0.6992519  0.5209146 1.0000000 0.5660575
SaleWt    0.3171163 0.3684348 0.5551134  0.1977254 0.5660575 1.0000000
```