Sjekk for enkel lineær regresjon, der vi har funnet MKE tidligere

Now let's look at the pieces of the new formula:

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

$$(\mathbf{X'X})^{-1} = \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix} = \frac{1}{nSS_X} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix}$$

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

Plug these into the equation for $\mathbf{b}$:

$$\begin{aligned}
\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \frac{1}{nSS_X}\begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix}\begin{bmatrix} \sum Y_i \\ \sum X_iY_i \end{bmatrix} \\
&= \frac{1}{nSS_X}\begin{bmatrix} (\sum X_i^2)(\sum Y_i) - (\sum X_i)(\sum X_iY_i) \\ -(\sum X_i)(\sum Y_i) + n\sum X_iY_i \end{bmatrix} \\
&= \frac{1}{SS_X}\begin{bmatrix} \bar{Y}(\sum X_i^2) - \bar{X}\sum X_iY_i \\ \sum X_iY_i - n\bar{X}\bar{Y} \end{bmatrix} \\
&= \frac{1}{SS_X}\begin{bmatrix} \bar{Y}(\sum X_i^2) - \bar{Y}(n\bar{X}^2) + \bar{X}(n\bar{X}\bar{Y}) - \bar{X}\sum X_iY_i \\ SP_{XY} \end{bmatrix} \\
&= \frac{1}{SS_X}\begin{bmatrix} \bar{Y}SS_X - SP_{XY}\bar{X} \\ SP_{XY} \end{bmatrix} = \begin{bmatrix} \bar{Y} - \frac{SP_{XY}}{SS_X}\bar{X} \\ \frac{SP_{XY}}{SS_X} \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix},
\end{aligned}$$

where

$$\begin{aligned}
SS_X &= \sum X_i^2 - n\bar{X}^2 = \sum (X_i - \bar{X})^2 \\
SP_{XY} &= \sum X_iY_i - n\bar{X}\bar{Y} = \sum (X_i - \bar{X})(Y_i - \bar{Y})
\end{aligned}$$

For simple linear regression

we had

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \equiv \frac{SS_{XY}}{SS_X}$$
$$b_0 = \bar{Y} - b_1\bar{X}$$

Identisk!

Matriseformuleringen er spesielt nyttig
for multippel regresjon (p>1), men ikke bare det:

Eks. Lin. regr. gjennom origo

Modell: $y_i = \beta_1 x_i + \varepsilon_i$

Transponert design-"matrise" $\mathbf{X}^\top = [x_1\ x_2 \cdots x_n]$ som gir

$$\mathbf{X}^\top \mathbf{Y} = \sum_{i=1}^{n} x_i y_i$$

$$\mathbf{X}^\top \mathbf{X} = \sum_{i=1}^{n} x_i^2$$

og dermed bli MKE

$$\hat{\beta}_1 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Enkel lineær regresjon

med sentrerte forklaringsvariable $x_i' = x_i - \bar{x}$

$$\mathbf{X}^\top = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 - \bar{x} & x_2 - \bar{x} & \cdots & x_n - \bar{x} \end{bmatrix}$$

og dermed

$$\mathbf{X}^\top \mathbf{Y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n (x_i - \bar{x}) y_i \end{bmatrix}$$

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n (x_i - \bar{x}) \\ \sum_{i=1}^n (x_i - \bar{x}) & \sum_{i=1}^n (x_i - \bar{x})^2 \end{bmatrix} = \begin{bmatrix} n & 0 \\ 0 & s_{xx} \end{bmatrix}$$

Dermed

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{s_{xx}} \end{bmatrix}$$

og

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \begin{bmatrix} \frac{\sum_{i=1} y_i}{n} \\ \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{s_{xx}} \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{s_{xy}}{s_{xx}} \end{bmatrix}$$

Den predikerte regresjonslinja blir dermed

$$\hat{y}_i = \bar{y} + \frac{s_{xy}}{s_{xx}}(x_i - \bar{x}) = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i$$

(som ved enkel lineær regresjon uten sentrering av x_i).

Eksempel 12.32, p=2

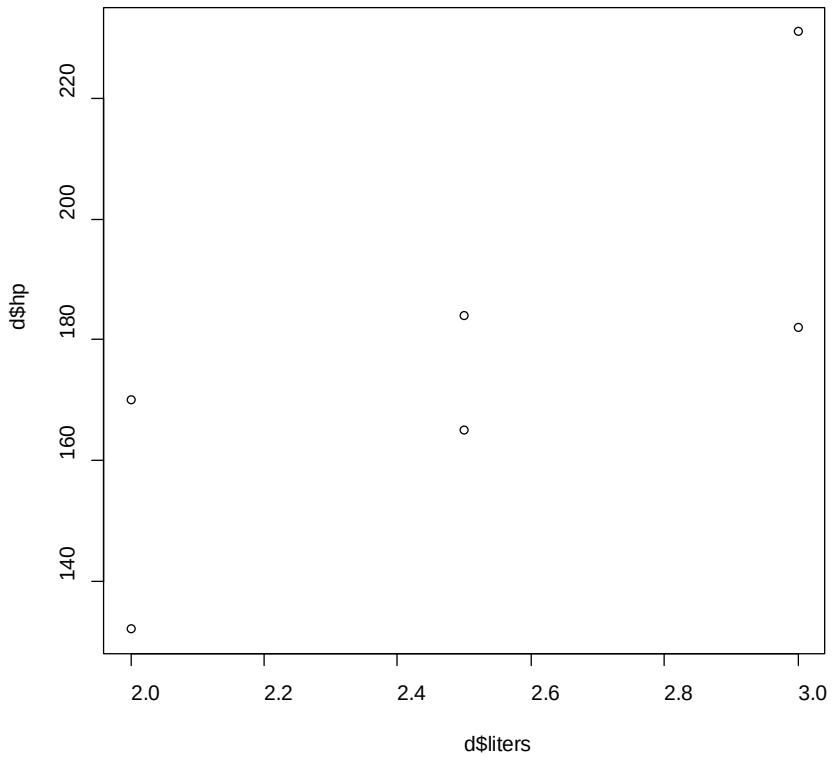
Prediker bilens hestekrefter vha motorvolum og bensintype

n=6 observasjoner (6 biler)

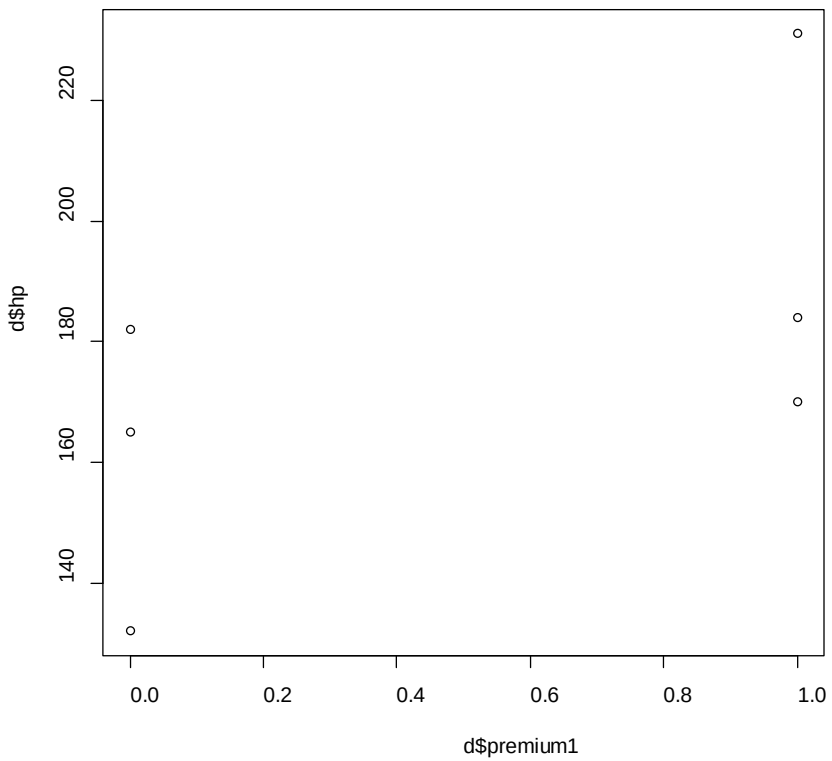
Hestekrefter (hp)	y
Motorvolum (l)	x1
Drivstoff (kategorisk)	x2

```
> d = read.table("exmp12-32.txt",header=TRUE,sep=",")
> #Liste opp navnene på de ulike variable
> names(d)
[1] "make"     "hp"       "liters"   "fuel"     "premium1"
> d
        make                 hp     liters      fuel         premium1
1     Dodge Neon         132       2.0      regular          0
2      Ford Focus SVT   170       2.0      premium        1
3      BMW                 184       2.5      premium        1
4      Subaru              165       2.5       regular         0
5      Saturn              182       3.0       regular         0
6      Jaguar              231       3.0       premium        1

> plot(d$liters,d$hp)
> plot(d$premium1,d$hp)
```



Hp mot motorvolum



Hp mot drivstoff

```
> #Beregning av beta.hat direkte

> y = matrix(d$hp,ncol=1)
> X = cbind(1,d$liters,d$premium1)

> print(cbind(X,y))
      [,1]  [,2]  [,3]  [,4]
[1,]    1   2.0    0    132
[2,]    1   2.0    1    170
[3,]    1   2.5    1    184
[4,]    1   2.5    0    165
[5,]    1   3.0    0    182
[6,]    1   3.0    1    231
```

```
> print(t(X)%*%X)
      [,1]   [,2]   [,3]
[1,]    6   15.0   3.0
[2,]   15   38.5   7.5
[3,]    3    7.5   3.0
> print(t(X)%*%y)
      [,1]
[1,] 1064.0
[2,] 2715.5
[3,]  585.0
>
> betahat = solve(t(X)%*%X,t(X)%*%y)
> print(betahat)
       [,1]
[1,] 20.91667
[2,] 55.50000
[3,] 35.33333
```

solve(a,b) løser ax=b

```
> #Bruk av lm kommando
> fit = lm(hp~liters+premium1,data=d)
> summary(fit)

Call:
lm(formula = hp ~ liters + premium1, data = d)

Residuals:
      1       2       3       4       5       6
 0.0833  2.7500 -11.0000  5.3333 -5.4167  8.2500

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   20.917      23.628    0.885   0.44123
liters        55.500       9.209    6.027   0.00916 **
premium1      35.333       7.519    4.699   0.01823 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.209 on 3 degrees of freedom
Multiple R-squared: 0.9511,     Adjusted R-squared: 0.9186
F-statistic:  29.2 on 2 and 3 DF,  p-value: 0.0108
```