3 Multiple Linear Regression

3.1 The Model

"Essentially, all models are wrong, but some are useful." Quote by George E.P. Box.
Models are supposed to be exact descriptions of the population, but that is assuming to much. But we have to require that they capture the essence of the relationship between the variables under investigation to be useful.

Definition 3.1.
A random variable Y fits a *Multiple Linear Regression Model*, iff there exist $\beta_0, \beta_1, \ldots, \beta_k \in \mathbb{R}$ so that for all $(x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{3.1}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Corollary:
If Y fits a *Multiple Linear Regression Model*, then for a fixed value of $\vec{x} = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$ the conditional expectation of Y given $\vec{x}$ equals

$$E(Y|\vec{x}) = E(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{3.2}$$

and the conditional variance of Y given $\vec{x}$ is

$$Var(Y|\vec{x}) = Var(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon) = Var(\varepsilon) = \sigma^2 \tag{3.3}$$

$\beta_0 \ldots \beta_k$ are called the regression coefficients, and they and σ^2 are the parameters of the model.

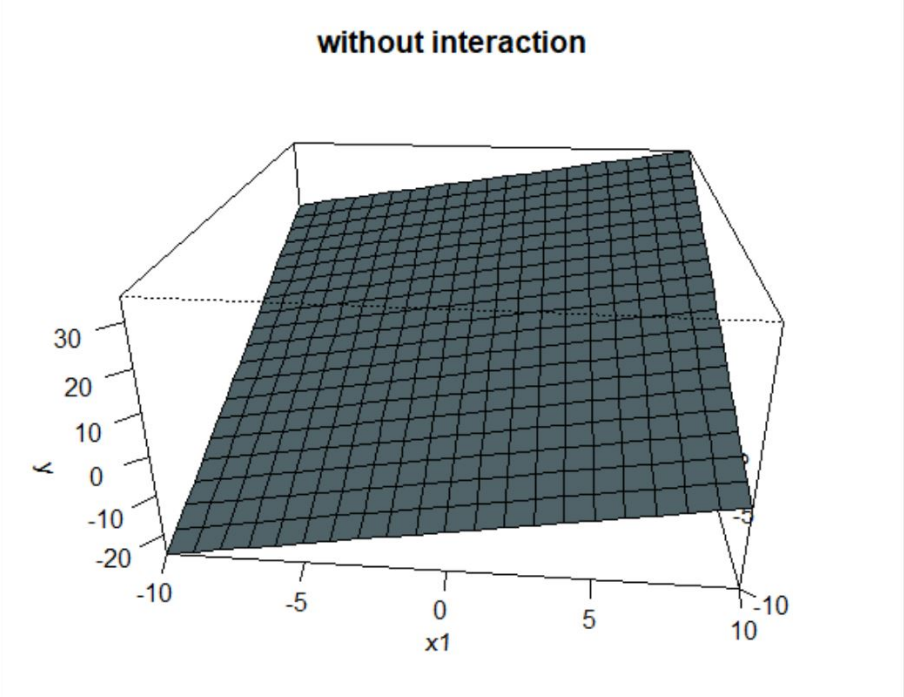
This implies that the mean of Y depends on $\vec{x}$ in a linear fashion and the variance of Y is independent from the value of $\vec{x}$, which is called homoscedasticity.

This is a *linear* model, since the model equation is linear in the parameters $\beta_0, \ldots, \beta_k$, also called the regression coefficients.

3.2 Visualize

3.2.1 Without Interaction

Let limit ourselves for the moment to two predictor variables x_1 and x_2 a third variable is Y the response variable.
In a model without error $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
Let $Y = 5 + x_1 + 2x_2$. The following diagram shows this function in a 3d-plot from R.

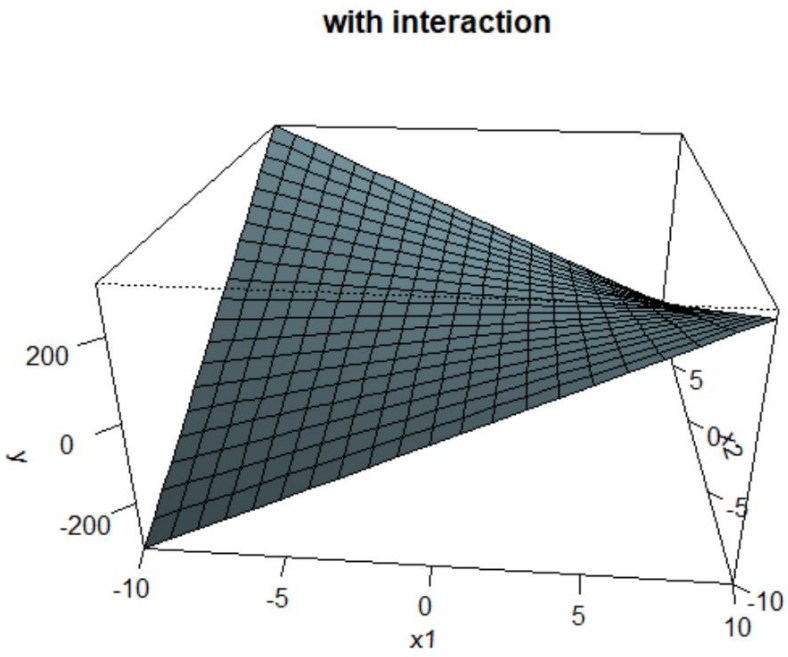


The equation does not include an interaction term, which means that values of x_2 do not have an effect on the slope in the relationship between x_1 and y, therefore the graph is a plane.

```
# R code for plotting the 3D plot
x1<-(-10:10)
x2<-(-10:10)
f <- function(x1,x2) { 5+x1 +2*x2}
y<-outer(x1,x2,f)
persp(x1, x2, y,
      theta = 5, phi = 25,
      expand = 0.5,
      col = "lightblue",
      ltheta = 120,
      shade = 0.75,
      ticktype = "detailed",
      main = "without interaction"
)
```

3.2.2 With Interaction

When interaction is present this indicates that for different values of x_2 the slope of the line relating x_1 and y can change, the graph is not a plane anymore. In the model this can be introduced by including the product term for the interacting variables, $x_1 * x_2$.
Let $Y = 5 + x_1 + 2x_2 - 3 * x_1x_2$.



Here you see that the result is no longer plane but a curved surface.

When we add an error to the model it explains that observations do not fall on the surface. The surface then represents the mean of the response variable, y.

3.3 Estimation

To learn about the population (inferential statistics) we assume a random sample has been collected, from which the goal is to infer about the population. In a first step the parameters of the model are to be estimated, using point estimators and confidence intervals.
The data can be represented in an array like

$$\begin{array}{ccccc} y_1 & x_{11} & x_{12} & \dots & x_{1k} \\ y_2 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & & \vdots & & \\ y_n & x_{n1} & x_{n2} & \dots & x_{nk} \end{array}$$

Using the data it will also be necessary to check if a multiple linear model is an appropriate description for the population, and if it is an appropriate description, what estimates for the model parameters can be found. If the MLR model is not a good fit, can we adjust the model to obtain a better fit? Another goal will be to identify outliers and check if they are causing problems, or if the model should be discarded entirely, because Y is independent from $x_1, \dots, x_k$.

Considering the above array, it seems to be reasonable to use matrices to represent the data.

3.3.1 Matrix notation

If the MLR model fits the population then the observations in the sample can be modeled by independent random variables which each fits the model equation, i.e. for $1 \leq i \leq n$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma)$$

are independent random variables.
It is assumed that the different measurements are obtained independently from each other, which is reflected by the assumption, that the Y_i are independent, which is equivalent to request that the ε_i are independent and identically distributed (iid).

Then with $\vec{Y} = (Y_1, Y_2, \ldots, Y_n)'$, $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$, $\vec{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$, and

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & & \vdots & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$$

The model for n independent observations can be given in the following form:

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon}, \quad \vec{\varepsilon} \sim \mathcal{N}(\vec{0}, \sigma^2 I_n)$$

For an explanation of the multivariate normal please see the review material "On Random Vectors".

3.4 Point Estimators

The least squares estimator is defined by the property that the total of the squared distances of the observed values to the regression function is as small as possible.

Theorem 3.1.
In the case that $X'X$ is invertible the least squares estimator for the coefficient vector $\vec{\beta}$, $\hat{\beta}$, is given by

$$\hat{\beta} = (X'X)^{-1}X'\vec{Y}$$

Proof:
Since the regression function is $f(\vec{x}) = \vec{x}\,'\vec{\beta}$, the vector of fitted values is for a given parameter vector $\vec{\beta}$ obtained by $X\vec{\beta}$. (explain) Then for $\vec{\beta} \in \mathbb{R}^{k+1}$ the total of the squared distances of $y_1, y_2, \ldots, y_n$ to the regression function is

$$(\vec{y} - X\vec{\beta})'(\vec{y} - X\vec{\beta}) = \vec{y}'\vec{y} - \vec{y}'X\vec{\beta} - (X\vec{\beta})'\vec{y} + (X\vec{\beta})'(X\vec{\beta}) = \vec{y}'\vec{y} - 2\vec{\beta}'X'\vec{y} + \vec{\beta}'X'X\vec{\beta}$$

In order to find the least squares estimator $\hat{\beta}$, which minimizes the expression above, find the partial derivatives of the expression and set them to zero, then solve for $\vec{\beta}$.
Use that for symmetric matrices A

$$\frac{\partial A\vec{x}}{\partial \vec{x}} = A, \quad \text{and} \quad \frac{\partial \vec{x}\,'A\vec{x}}{\partial \vec{x}} = 2A\vec{x}$$

The linear system giving $\hat{\beta}$:

$$-2X'\vec{y} + 2X'X\hat{\beta} = \vec{0}_k$$

which is equivalent to the **normal equation**

$$X'X\hat{\beta} = X'\vec{y}$$

In the case that $X'X$ is invertible

$$\hat{\beta} = (X'X)^{-1}X'\vec{y}$$

This means given an observed response vector $\vec{y}$, and a design matrix X of full rank, the least squares estimate for $\vec{\beta}$ is given by $\hat{\beta} = (X'X)^{-1}X'\vec{y}$ (defined in the context of a specific sample).
Embedding this into Statistical Estimation Theory means, that the least squares estimator for the MLRM is given as (defined as a random variable):

$$\hat{\beta} = (X'X)^{-1}X'\vec{Y}$$

Which concludes the proof.

$X'X$ is invertible, iff the columns of X are linearly independent.
The i^{th} column of X gives the values for the n individuals from the sample for the i^{th} predictor variable.
It follows that in the case of linearly dependent columns the least square estimator is not unique. In the case the columns are linearly dependent one usually removes columns until the design matrix has full rank, and the least squares estimator becomes unique.
This is also the reason it is disadvantageous if predictor variables have a high correlation, this results in a small determinant of $X'X$ making it hard numerically to invert the matrix and calculating the least squares estimator with high precisions.

Least squares estimator as projection

> Geometrically regression is the orthogonal projection of the vector $\vec{y} \in \mathbb{R}^n$ into the p dimensional space spanned by the columns from X.
>
> The projection $\hat{y}$ is according to Linear Algebra
>
> $$\hat{y} = X(X'X)^{-1}X'\vec{y} = X\hat{\beta}$$

The *hat matrix* is defined as

$$H = X(X'X)^{-1}X'$$

because when applied to $\vec{Y}$, it gets a hat.

The vector $\hat{y}$ gives the fitted values for observed values $\vec{y}$ from the model estimates. The model for the n observations are

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$$

where $\vec{\varepsilon}$ has en expected value of $\vec{0}$. The fitted value of $\vec{y}$, $\hat{y}$ is then

$$\hat{y} = X\hat{\beta}$$

and as random variables:

$$\hat{Y} = X\hat{\beta}$$

Lemma 1.
In the MLRM the observed residual vector for a data vector is

$$\vec{e} = \vec{y} - \hat{y} = y - X\hat{\beta} = y - X(X'X)^{-1}X'\vec{y} = (I_n - X(X'X)^{-1}X')\vec{y} = (I_n - H)\vec{y}$$

Properties of $\hat{\beta}$

Theorem 3.2.

1. $\hat{\beta}$ is an unbiased estimator for $\vec{\beta}$ ($E(\hat{\beta}) = \vec{\beta}$).
2. The covariance matrix of $\hat{\beta}$ is
$$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$$
3. $\hat{\beta}$ has a multivariate normal distribution.

Proof:

1.
$$\begin{aligned} E(\hat{\beta}) &= E\left((X'X)^{-1}X'Y\right) \\ &= (X'X)^{-1}X'E(Y) \\ &= (X'X)^{-1}X'X\vec{\beta} \\ &= I_n\vec{\beta} = \vec{\beta} \end{aligned}$$

2.
$$\begin{aligned} \mathrm{Cov}(\hat{\beta}) &= \mathrm{Cov}\left((X'X)^{-1}X'Y\right) \\ &= (X'X)^{-1}X'\mathrm{Cov}(Y)X((X'X)^{-1})' \\ &= (X'X)^{-1}X'\sigma^2 I_n X((X'X)^{-1})' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

3. We have to show that every linear combination of components of $\hat{\beta}$ is normally distributed.
   A linear combinations of the components of $\hat{\beta}$ can be written for a vector $\vec{a} \in \mathbb{R}^{k+1}$ as:
$$\vec{a}'\hat{\beta} = \vec{a}'(X'X)^{-1}X'\vec{Y} = (\vec{a}'(X'X)^{-1}X')\vec{Y}$$
   which is a linear combination of components of $\vec{Y}$. Since $\vec{Y}$ is multivariate normal, the linear combination is by definition normally distributed, which concludes the proof.

$\hat{\beta}$ has also "minimal" variance. This property has to be defined, because $\hat{\beta}$ is a random vector and therefore has a $p = k + 1$ dimensional covariance matrix.

Theorem 3.3.
(Gauss-Markov Theorem) Let $\psi = \vec{c}'\vec{\beta}$, for $\vec{c} \in \mathbb{R}^p$, be a linear combination of the parameter vector $\vec{\beta}$ in the MLR model, assume the design matrix X has full rank, then $\hat{\psi} = \vec{c}'\hat{\beta}$ is an unbiased estimator for ψ and has smallest variance in the class of linear unbiased estimators for ψ.

Proof:(Compare Faraway(2002))
Using properties from the review material "On Random Vectors" and the last theorem we get

$$E(\hat{\psi}) = E(\vec{c}'\hat{\beta}) = \vec{c}'E(\hat{\beta}) = \vec{c}'\vec{\beta} = \psi,$$

$\hat{\psi}$ Therefore $\hat{\psi}$ is an unbiased estimator for ψ.

In the next step we will now prove that $\hat{\psi}$ has the smallest variance between all linear unbiased estimators.

We start with some helpful observations: Suppose $\vec{a}'\vec{Y}$ is some linear unbiased estimate of $\vec{c}'\vec{\beta}$ so that

$$E(\vec{a}'\vec{Y}) = \vec{c}'\vec{\beta} \text{ for all } \vec{\beta} \in \mathbb{R}^p$$
$$\vec{a}'X\vec{\beta} = \vec{c}'\vec{\beta} \text{ for all } \vec{\beta} \in \mathbb{R}^p$$

from which follows that

$$\vec{a}'X = \vec{c}' \quad (1).$$

(Why?)

Now let $\vec{b} = (X'X)^{-1}\vec{c}$, then

$$\vec{c} = X'X\vec{b} \quad (2)$$

and

$$\hat{\psi} = \vec{c}'\hat{\beta} = \vec{b}'X'X\hat{\beta} = \vec{b}'X'X(X'X)^{-1}X'\vec{Y} = \vec{b}'X'\vec{Y} \quad (3)$$

Now find the variance of $\vec{a}'\vec{Y}$ and show that $Var(\vec{a}'\vec{Y})$+ something positive = $Var(\hat{\psi})$):

$$\begin{aligned} Var(\vec{a}'\vec{Y}) &= Var(\vec{a}'\vec{Y} - \vec{c}'\hat{\beta} + \vec{c}'\hat{\beta}) \\ &= Var(\vec{a}'\vec{Y} - \vec{c}'\hat{\beta}) + Var(\vec{c}'\hat{\beta}) + 2Cov(\vec{a}'\vec{Y} - \vec{c}'\hat{\beta}, \vec{c}'\hat{\beta}) \end{aligned}$$

But using (3)

$$\begin{aligned} Cov(\vec{a}'\vec{Y} - \vec{c}'\hat{\beta}, \vec{c}'\hat{\beta}) &= Cov(\vec{a}'\vec{Y} - \vec{b}'X'\vec{Y}, \vec{b}'X'\vec{Y}) \\ &= Cov((\vec{a}' - \vec{b}'X')\vec{Y}, \vec{b}'X'\vec{Y}) \\ &= (\vec{a}' - \vec{b}'X')Cov(\vec{Y})X\vec{b} \end{aligned}$$

according to "On Random Vectors" $(Cov(\vec{l}'\vec{Y}, \vec{h}'\vec{Y}) = \vec{l}'Cov(\vec{Y})\vec{h})$.
Then

$$\begin{aligned} (\vec{a}' - \vec{b}'X')Cov(\vec{Y})X\vec{b} &= (\vec{a}' - \vec{b}'X')\sigma^2 I_n X\vec{b} \\ &= \sigma^2(\vec{a}'X - \vec{b}'X'X)\vec{b} \\ &= \sigma^2(\vec{c}' - \vec{c}')\vec{b} \\ &= 0 \end{aligned}$$

using (1) and (2).
Therefore

$$Var(\vec{a}'\vec{Y}) = Var(\vec{a}'\vec{Y} - \vec{c}'\hat{\beta}) + Var(\vec{c}'\hat{\beta})$$

Since variances are always at least 0, this means that for all unbiased estimators $\vec{a}'Y$

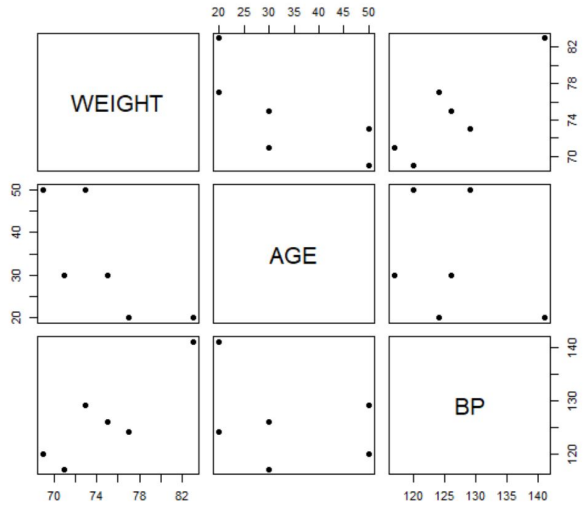
$$Var(\vec{a}'\vec{Y}) \geq Var(\vec{c}'\hat{\beta})$$

Which concludes the proof.

Example 3.1.

A small example relating age and weight to blood pressure:
The data

weight	age	blood pressure
69	50	120
83	20	141
77	20	124
75	30	126
71	30	117
73	50	129



The matrix scatterplot displays the relationship between all the variables of interest. This has limited usefulness as discussed later.
The scatterplots do not display strong linear relationships, but we will never the less propose the following model for blood pressure in the population:

$$bp = \beta_0 + \beta_1(weight) + \beta_2(age) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

First find the vector of parameter estimates for $\vec{\beta} = (\beta_0, \beta_1, \beta_2)'$, based on the sample data above:

$$X = \begin{pmatrix} 1 & 69 & 50 \\ 1 & 83 & 20 \\ 1 & 77 & 20 \\ 1 & 75 & 30 \\ 1 & 71 & 30 \\ 1 & 73 & 50 \end{pmatrix} \text{ with } X'X = \begin{pmatrix} 6 & 448 & 200 \\ 448 & 33574 & 14680 \\ 200 & 14680 & 7600 \end{pmatrix} \text{ and } (X'X)^{-1} = \begin{pmatrix} 129.777 & -1.534 & -0.452 \\ -1.534 & 0.018 & 0.005 \\ -0.452 & 0.005 & 0.002 \end{pmatrix}$$

(Last matrix found with R) With

$$\vec{y} = \begin{pmatrix} 120 \\ 141 \\ 124 \\ 126 \\ 117 \\ 129 \end{pmatrix}, \; X'\vec{y} = \begin{pmatrix} 757 \\ 56705 \\ 25040 \end{pmatrix} \text{ we get } \hat{\beta} = (X'X)^{-1}X'\vec{y} = \begin{pmatrix} -66.13 \\ 2.38 \\ 0.44 \end{pmatrix} \text{ and } \hat{y} = X\hat{\beta} = \begin{pmatrix} 119.99 \\ 140.15 \\ 125.87 \\ 125.50 \\ 115.98 \\ 129.51 \end{pmatrix}$$

Comparing the vector of fitted values, $\hat{y}$, with the measurement vector, $\vec{y}$, we can observe a close fit.

For closer examination find the residual vector:

$$\vec{e} = \vec{y} - \hat{y} = \begin{pmatrix} 0.005 \\ 0.853 \\ -1.869 \\ 0.503 \\ 1.021 \\ -0.513 \end{pmatrix}$$

The residuals giving the difference in blood pressure measured and predicted by the model, are quite small, and seem to support the model.

Interpretation of the parameter estimates:

- The mean blood pressure for a person of age 0 and weight 0 is estimated to be -66.13. This is not making sense! We would be extrapolating using this interpretation.
- The blood pressure increase in average by 2.38 mmHg for every extra year when the weight is unchanged.
- The blood pressure increase in average by 0.44 mmHg for every extra kilogram when the age is unchanged.

To be continued.

The estimator for the variance of the error in the MLR model, σ^2, is again based on the residuals.

Definition 3.2.
Let $p := k + 1$ (the number of regression parameters)

1. $\vec{e} = \vec{y} - \hat{y}$ is the vector of observed residuals.
2. The residual sum of squares is
$$SS_{Res} = \vec{e}'\vec{e}$$
3. The mean squares of error is
$$MS_{Res} = \frac{SS_{Res}}{n - p}$$

Theorem 3.4.

1. $SS_{Res} = \vec{Y}'\vec{Y} - \hat{\beta}X'\vec{Y} = \vec{Y}' \left[I_n - X(X'X)^{-1}X' \right] \vec{Y}$
2. $E(SS_{Res}) = (n - p)\sigma^2$
3. SS_{Res}/σ^2 is χ^2 distributed with $df = (n - p)$ and non centrality parameter $\lambda = 0$.

Proof:

1.

$$\begin{aligned} SS_{Res} &= \vec{e}'\vec{e} \\ &= (\vec{Y}-\hat{Y})'(\vec{Y}-\hat{Y}) \\ &= \vec{Y}'\vec{Y}-\vec{Y}'X\hat{\beta}-\hat{\beta}'X'\vec{Y}+\hat{\beta}'X'X\hat{\beta} \\ &= \vec{Y}'\vec{Y}-\vec{Y}'X\hat{\beta}-\hat{\beta}'X'\vec{Y}+\vec{Y}'X(X'X)^{-1}X'X\hat{\beta} \\ &= \vec{Y}'\vec{Y}-\vec{Y}'X\hat{\beta}-\hat{\beta}'X'\vec{Y}+\vec{Y}'X\hat{\beta} \\ &= \vec{Y}'\vec{Y}-\hat{\beta}'X'\vec{Y} \end{aligned}$$

which proves the first part. Continue from here

$$\begin{aligned} SS_{Res} &= \vec{Y}'\vec{Y}-\hat{\beta}'X'\vec{Y} \\ &= \vec{Y}'\vec{Y}-\vec{Y}X(X'X)^{-1}X'\vec{Y} \\ &= \vec{Y}'\left[I-X(X'X)^{-1}X'\right]\vec{Y} \end{aligned}$$

which concludes the second part of the equation.

2. To find the mean use $E(\vec{Y}'A\vec{Y}) = tr(A\text{Cov}(Y)) + E(\vec{Y})'AE(\vec{Y})$ from "On Random Vectors". Therefore we need the $tr(I-X(X'X)^{-1}X')$:

$$\begin{aligned} tr([I-X(X'X)^{-1}X']) &= tr(I)-tr(X(X'X)^{-1}X') \\ &= n-tr((X'X)^{-1}X'X) \\ &= n-tr(I_p) = n-p \end{aligned}$$

Using that $tr(AB) = tr(BA)$, if the matrices permit the multiplications. Then

$$\begin{aligned} E(SS_{Res}) &= E(\vec{Y}'\left[I-X(X'X)^{-1}X'\right]\vec{Y}) \\ &= tr(\left[I-X(X'X)^{-1}X'\right]\sigma^2 I) + E(\vec{Y})'\left[I-X(X'X)^{-1}X'\right]E(\vec{Y}) \\ &= \sigma^2(n-p)+\vec{\beta}'X'\left[I-X(X'X)^{-1}X'\right]X\vec{\beta} \\ &= \sigma^2(n-p)+\vec{\beta}'X'X\vec{\beta}-\vec{\beta}'X'X(X'X)^{-1}X'X\vec{\beta} \\ &= \sigma^2(n-p)+\vec{\beta}'X'X\vec{\beta}-\vec{\beta}'X'X\vec{\beta} \\ &= \sigma^2(n-p) \end{aligned}$$

3. For this proof we need the following:

   - A matrix A is called idempotent if $A = AA$.
     If A is a symmetric idempotent matrix than $I-A$ is symmetric idempotent.
     If A is idempotent, than $rank(A) = trace(A)$.
   - (From appendix C.2.4 from the text book)
     If $\vec{Y} \sim \mathcal{N}(\vec{\mu}, V)$, and $U = \vec{Y}\,'A\vec{Y}$ and either AV or VA is idempotent of rank p, than $U \sim \chi^2$, with $df = p$ and non-centrality parameter $\lambda = \vec{\mu}\,'A\vec{\mu}$.

   First show that $X(X'X)^{-1}X'$ is idempotent:

$$\begin{aligned} (X(X'X)^{-1}X')(X(X'X)^{-1}X') &= X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= X(X'X)^{-1}IX' \\ &= X(X'X)^{-1}X' \end{aligned}$$

Since $X(X'X)^{-1}X'$ is idempotent, so is $I - X(X'X)^{-1}X'$. Using that with $\vec{Z} = \frac{1}{\sigma}\vec{Y}$,

$$Cov(Z) = \frac{1}{\sigma^2}Cov(\vec{Y}) = \frac{1}{\sigma^2}\sigma^2 I = I$$

and

$$rank(I - X(X'X)^{-1}X') = trace(I - X(X'X)^{-1}X'X) = n - p$$

therefore applying the second bullet we get that

$$\begin{array}{rcll} \vec{Z}\,'(I - X(X'X)^{-1}X')\vec{Z} & \sim & \chi^2 & \text{with} \quad df = rank(I - X(X'X)^{-1}X') = n - p \\ \frac{1}{\sigma^2}\vec{Y}\,'(I - X(X'X)^{-1}X')\vec{Y} & \sim & \chi^2 & \text{with} \quad df = rank(I - X(X'X)^{-1}X') = n - p \\ \frac{1}{\sigma^2}SS_{Res} & \sim & \chi^2 & \text{with} \quad df = rank(I - X(X'X)^{-1}X') = n - p \end{array}$$

which concludes the proof.

Conclusion: $\hat{\sigma}^2 = MS_{Res}$ is an unbiased estimator for σ^2.

Continue Example. 3.1:

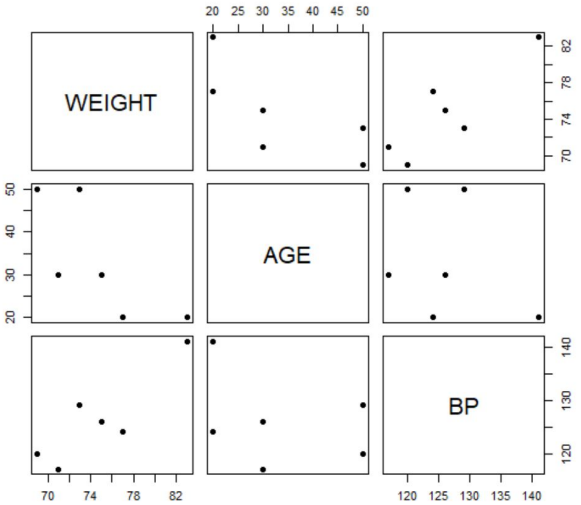
$$SS_{Res} = \vec{e}\,'\vec{e} = (0.005, 0.853, -1.869, 0.503, 1.021, -0.513)\begin{pmatrix} 0.005 \\ 0.853 \\ -1.869 \\ 0.503 \\ 1.021 \\ -0.513 \end{pmatrix} = 5.78$$

The estimated variance $\hat{\sigma}^2 = 5.78/(6-3) = 1.93$, and therefore $\hat{\sigma} = \sqrt{1.93} = 1.39$.

We estimate that the measurements vary with a standard deviation of 1.39 around the regression function.

3.5 Scatter Plots in Multiple Linear Regression

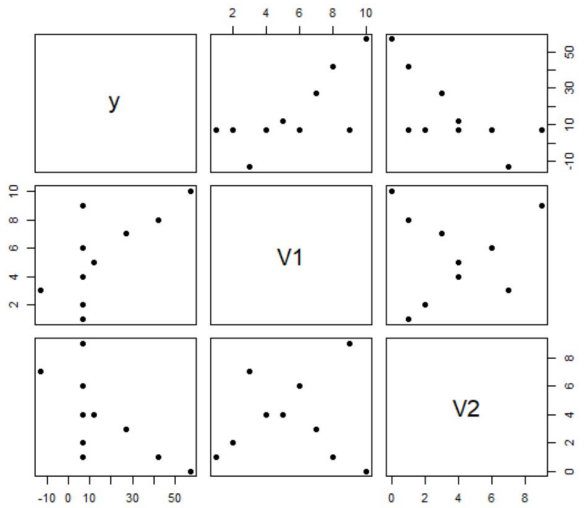
Two dimensional scatter plots as we have used in the blood pressure example are of limited usefulness. In the example we showed



The bottom left plot seems to indicate a positive association between weight and blood pressure, the bottom middle shows no apparent association between age and blood pressure, and the middle left plot shows a negative association between age and weight.
The problem arising is that when plotting the association between two variables we are ignoring the third, which can hide a more dimensional association (think of the 3D-plots)
But if we would observe a strong association between two predictors this can point to a collinearity issue, which could result in a close to singular design matrix, causing high standard errors in the estimation.
Even if there would be a perfect linear relationship between x_1, x_2 and y this would not necessarily show in the plot.

Example 3.2.

V1	V2
1	1
2	2
3	7
4	4
5	4
6	6
7	3
8	1
9	9
10	0



and $y = -7 + 5V1 - 5V2$

This is an example with only two regressors and a perfect linear association (no error, no interaction). One can imagine how this becomes even less clear in cases with more regressors, including interaction and error.