Regression Analysis

Chapter 2. The General Linear Model

Prof. dr. Thomas Neyens

Course notes: Prof. dr. Mia Hubert & Prof. dr. Stefan Van Aelst

The general linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i \tag{1}$$

for $i = 1, \ldots, n$

- β_0 is called the *intercept*.
- β_j $(j = 1, \ldots, p-1)$ are the regression *slopes*.
- X_j do not necessarily reflect the observed predictor variables, but for any function f_j of them.
- X_j do not contain any random effect or measurement error.
- We assume that the Gauss-Markov conditions are satisfied:

$$\mathrm{E}[\epsilon_i] = 0 \tag{2}$$

$$\mathrm{Var}[\epsilon_i] = \sigma^2 \tag{3}$$

$$\mathrm{E}[\epsilon_i \epsilon_j] = 0 \text{ for all } i \neq j. \tag{4}$$

The general linear model

- As the ϵ_i are random with zero mean, also Y is a random variable that satisfies:

$$E[Y|X_1, \dots, X_{p-1}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p-1} X_{p-1}$$

- Conditionally on the observed values for $X_1, \dots, X_{p-1}$, this can also be written as:

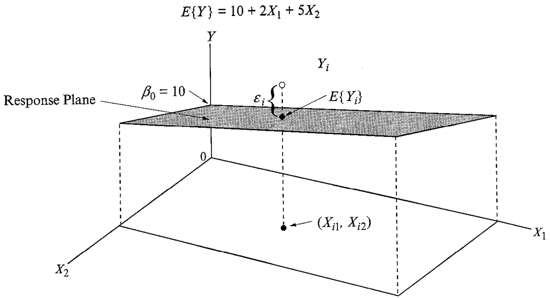
$$E[Y|\mathbf{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i,p-1} \tag{5}$$

- $\mathbf{x}_i = (1, x_{i1}, \dots, x_{i,p-1})^t$
- The first element of the $\mathbf{x}$-vector is 1, which is the x-value for the intercept.

The general linear model

- At the first-order regression model (where the X_j in (1) correspond with the observed predictor variables), we try to estimate a hyperplane in the (X, Y)-space.

FIGURE 6.1 **Response Function is a Plane—Sales Promotion Example.**



- If $p = 2$ we recover simple regression: $\mathrm{E}[Y|X] = \beta_0 + \beta_1 X$

The general linear model

- β_0 is the expected response value at $\mathbf{x}_i = (1, 0, \ldots, 0)^t$.
- β_j indicates the change in the expected value of the response Y due to a unit increase in the variable X_j *when all other predictor variables are held constant*.
- Let $\mathbf{x}_{i(j)} = (1, x_{i1}, \ldots, x_{ij}, \ldots, x_{i,p-1})^t$ and $\mathbf{x}_{i(j+1)} = (1, x_{i1}, \ldots, x_{ij} + 1, \ldots, x_{i,p-1})^t$, then from (5) it follows that

$$\mathrm{E}(Y|\mathbf{x}_{i(j)}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \ldots + \beta_{p-1} x_{i,p-1}$$
$$\mathrm{E}(Y|\mathbf{x}_{i(j+1)}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j (x_{ij} + 1) + \ldots + \beta_{p-1} x_{i,p-1}$$

- Hence $\beta_j = \mathrm{E}(Y|\mathbf{x}_{i(j+1)}) - \mathrm{E}(Y|\mathbf{x}_{i(j)})$.

Matrix notation

- We want to write (1) in matrix format.
- Before we do this, remember the following:
  - The transpose of a matrix: e.g.,

$$\underset{3x2}{\mathbf{A}} = \begin{bmatrix} 8 & 3 \\ 2 & 6 \\ 1 & 5 \end{bmatrix}, \underset{2x3}{\mathbf{A}^t} = \begin{bmatrix} 8 & 2 & 1 \\ 3 & 6 & 5 \end{bmatrix}$$

  - The sum of 2 matrices: e.g.,

$$\underset{3x2}{\mathbf{B}} = \begin{bmatrix} 2 & 2 \\ 1 & 4 \\ 3 & 1 \end{bmatrix}, \underset{3x2}{\mathbf{A}} + \underset{3x2}{\mathbf{B}} = \begin{bmatrix} 10 & 5 \\ 3 & 10 \\ 4 & 6 \end{bmatrix}$$

Matrix notation

- Remember the following:
  - The multiplication of a matrix by a scalar: e.g.,

$$\underset{3x2}{\mathbf{A}} = \begin{bmatrix} 8 & 3 \\ 2 & 6 \\ 1 & 5 \end{bmatrix}, k\underset{3x2}{\mathbf{A}} = \begin{bmatrix} 8 & 3 \\ 2 & 6 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 8k & 3k \\ 2k & 6k \\ 1k & 5k \end{bmatrix}$$

  - The multiplication of 2 matrices: e.g.,

$$\underset{2x3}{\mathbf{A}} = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 5 & 8 \end{bmatrix}, \underset{3x1}{\mathbf{B}} = \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix}, \underset{2x1}{\mathbf{AB}} = \begin{bmatrix} 26 \\ 41 \end{bmatrix}$$

Matrix notation

- Remember the following:

$$\underset{rx1}{\mathbf{0}} = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}, \underset{rx1}{\mathbf{1}} = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}, \underset{rxr}{\mathbf{I}} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ & & \dots\dots\dots & & \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix},$$

$$\underset{rxr}{\mathbf{J}} = \underset{rxr}{\mathbf{1}\mathbf{1}^t} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \\ & & \dots\dots\dots & & \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}$$

Matrix notation

- If no columns of a matrix can be expressed as a linear combination of the other, we say that the columns are *linearly independent*.
- The *rank* of a matrix is the maximum number of linearly independent columns of a matrix
  - Note that the rank can be equivalently defined as the maximum number of linearly independent rows. Hence, the rank of an *r x c* matrix cannot exceed *min(r,c)*.
- The inverse of a matrix $\mathbf{A}$ is $\mathbf{A}^{-1}$, such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

  - The inverse is only defined for square matrices.
  - The inverse of a *r x r* matrix only exists if the rank is *r* (nonsingular or full rank)

Matrix notation

- A matrix is symmetric if $\mathbf{A} = \mathbf{A}^t$
- A diagonal square matrix is a matrix whose off-diagonal elements are all zeros (e.g., $\mathbf{I}$)
- A matrix is idempotent if $\mathbf{AA} = \mathbf{A}$

Matrix notation

- Some matrices that will be useful:

$$\underset{nx1}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \underset{px1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{p-1} \end{bmatrix},$$

$$\underset{nxp}{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p-1,1} \\ 1 & x_{12} & x_{22} & \dots & x_{p-1,2} \\ & & \dots\dots\dots\dots & & \\ 1 & x_{1n} & x_{2n} & \dots & x_{p-1,n} \end{bmatrix}, \underset{px1}{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix}$$

- X is also called the *design matrix*.

Matrix notation

- Other matrices that will be useful: $\underset{pxp}{X^tX}$ and $\underset{px1}{X^tY}$
  - E.g., for the simple linear regression case:

$$\underset{2x2}{X^tX} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}, \underset{2x1}{X^t\mathbf{y}} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

- The variance-covariance matrix of a *n x 1* random vector **y** is defined as:

$$\underset{nxn}{\Sigma(\mathbf{y})} = \begin{bmatrix} \sigma^2\{y_1\} & \sigma\{y_1, y_2\} & \dots & \sigma\{y_1, y_n\} \\ \sigma\{y_2, y_1\} & \sigma^2\{y_2\} & \dots & \sigma\{y_2, y_n\} \\ & \dots\dots\dots\dots\dots\dots & & \\ \sigma\{y_n, y_1\} & \sigma\{y_n, y_2\} & \dots & \sigma^2\{y_n\} \end{bmatrix}$$

Matrix notation

- Let's write (1) in matrix format.
  - Let the vectors $\mathbf{y} = (y_1, \ldots, y_n)^t, \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^t$ and the matrix $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^t$, then (1) is equivalent to

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{6}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p-1,1} \\ 1 & x_{12} & x_{22} & \ldots & x_{p-1,2} \\ & & \ldots\ldots\ldots\ldots & & \\ 1 & x_{1n} & x_{2n} & \ldots & x_{p-1,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \ldots \\ \epsilon_n \end{bmatrix}$$

Matrix notation

- while (2), (3) and (4) correspond with

$$\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0} \tag{7}$$

$$\Sigma(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ & \dots\dots\dots & & \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 I_n. \tag{8}$$

- Here, $\Sigma(\boldsymbol{\epsilon})$ represents the variance-covariance matrix of the errors, and I_n for the $n \times n$ identity matrix.

Estimation of the regression parameters

- Any parameter estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_{p-1})^t$ yields fitted values $\hat{y}_i$ and residuals e_i:

$$\begin{aligned} e_i(\hat{\boldsymbol{\beta}}) &= y_i - \hat{y}_i \\ &= y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}. \end{aligned}$$

- The **least squares estimator** $\hat{\boldsymbol{\beta}}_{LS}$ is defined as the $\hat{\boldsymbol{\beta}}$ for which the sum of the squared residuals is minimal, or

$$\hat{\boldsymbol{\beta}}_{LS} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} e_i^2(\boldsymbol{\beta}) \tag{9}$$

- $\sum_{i=1}^{n} e_i^2(\boldsymbol{\beta})$ is called the objective function.

Estimation of the regression parameters

- Differentiating $\sum_{i=1}^{n} e_i^2(\boldsymbol{\beta})$ with respect to each β_j $(j = 0, \ldots, p-1)$ and setting the derivatives equal to zero, yields the normal equations (check this for the simple linear regression case)

$$X^t X \boldsymbol{\beta} = X^t \mathbf{y}$$

- If $\text{rank}(X) = p \leqslant n$, the solution of this linear system is given by:

$$\hat{\boldsymbol{\beta}}_{LS} = (X^t X)^{-1} X^t \mathbf{y} \tag{10}$$

- $X^t X$ is the matrix of cross-products:

$$(X^t X)_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} \tag{11}$$

$$(X^t X)_{jj} = \sum_{i=1}^{n} x_{ij}^2 \tag{12}$$

Multicollinearity

- The condition $\text{rank}(X) = p \leqslant n$ is necessary to ensure that $X^t X$ is non-singular.
- If the rank of X is exactly p (next slide left figure), the objective function is convex and hence yields a unique minimum which can be derived analytically.
- If $\text{rank}(X) < p$ (next slide right figure), there are an infinite number of LS solutions.
- More realistic: the X-variables might be strongly correlated (*multicollinearity*).
  - In such a case, the LS fit is uniquely defined, but many other parameter estimates $\hat{\beta}$ attain a residual sum of squares which is close to the minimal value of $\hat{\beta}_{LS}$ (next slide, lower figure).
  - Small changes in the data set may cause a large change in the parameter estimates.

Multicollinearity

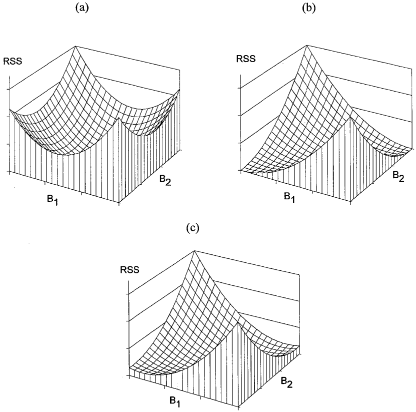


Figure 13.3. The residual sum of squares as a function of the slope coefficients B_1 and B_2. In each graph, the vertical axis is scaled so that the least-squares value of RSS is at the bottom of the axis. When, as in (*a*), the correlation between the independent variables X_1 and X_2 is small, the residual sum of squares has a well-defined minimum, much like a deep bowl. When there is a perfect linear relationship between X_1 and X_2, as in (*b*), the residual sum of squares is flat at its minimum, above a line in the B_1, B_2 plane: The least-squares values of B_1 and B_2 are not unique. When, as in (*c*), there is a strong, but less-than-perfect, linear relationship between X_1 and X_2, the residual sum of squares is nearly flat at its minimum, so values of B_1 and B_2 quite different from the least-squares values are associated with residual sums of squares near the minimum.

Multicollinearity

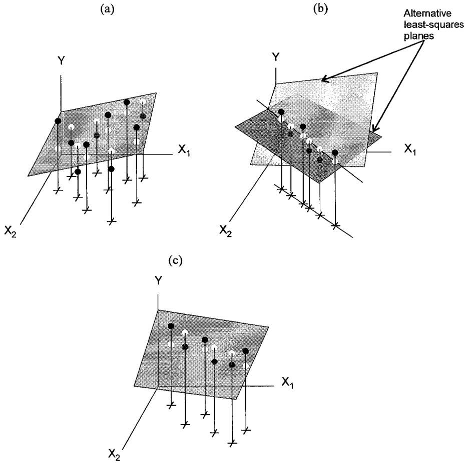


Figure 13.2. The impact of collinearity on the stability of the least-squares regression plane. In (*a*), the correlation between X_1 and X_2 is small, and the regression plane therefore has a broad base of support. In (*b*), X_1 and X_2 are perfectly correlated; the least-squares plane is not uniquely defined. In (*c*), there is a strong, but less-than-perfect, linear relationship between X_1 and X_2; the least-squares plane is uniquely defined, but it is not well supported by the data.

Properties and geometrical interpretation

- Let's denote $\hat{\beta}_{LS}$ as $\hat{\beta}$.
- Let $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^t$.
- Now,

$$\hat{\mathbf{y}} = X\hat{\beta} = X(X^tX)^{-1}X^t\mathbf{y} = H\mathbf{y} \tag{13}$$

with the **hat matrix**:

$$\boxed{H = X(X^tX)^{-1}X^t} \tag{14}$$

- Let $\mathbf{e} = (e_1, \ldots, e_n)^t$.
- Now,

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - H\mathbf{y} = (I_n - H)\mathbf{y} = M\mathbf{y} \tag{15}$$

where

$$M = I_n - H$$

- H and M are symmetric and idempotent.

Properties and geometrical interpretation

- The following relations hold:

$$
\begin{aligned}
\mathbf{e} &= M\mathbf{y} = M(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) = MX\boldsymbol{\beta} + M\boldsymbol{\epsilon} \qquad (16) \\
&= (X - X(X^tX)^{-1}X^tX)\boldsymbol{\beta} + M\boldsymbol{\epsilon} \\
&= \mathbf{0}_{n,p}\boldsymbol{\beta} + M\boldsymbol{\epsilon} = M\boldsymbol{\epsilon} \\
\Sigma(\mathbf{e}) &= \Sigma(M\boldsymbol{\epsilon}) = M\Sigma(\boldsymbol{\epsilon})M^t = \sigma^2 MI_nM^t = \sigma^2 MM^t = \sigma^2 M \qquad (17)
\end{aligned}
$$

Properties and geometrical interpretation

- The least squares residuals satisfy:

$$\sum_{i=1}^{n} e_i = 0 \tag{18}$$

$$\sum_{i=1}^{n} x_{ij} e_i = 0 \text{ for all } j = 1, \ldots, p-1 \tag{19}$$

$$\sum_{i=1}^{n} e_i \hat{y}_i = 0 \tag{20}$$

- The first two equations (18) and (19) follow from $X^t\mathbf{e} = X^t M\epsilon = \mathbf{0}_{p,n}\epsilon = \mathbf{0}_p$.
- These imply that:
  - the mean of the least squares residuals is zero.
  - the residuals are orthogonal to the design matrix X as well as to the predicted values.

Properties and geometrical interpretation

- Since $\frac{1}{n}\sum_i (y_i - \hat{y}_i) = 0$

$$
\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_i \hat{y}_i \\
&= \frac{1}{n}\sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{p-1} x_{i,p-1}) \\
&= \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \ldots + \hat{\beta}_{p-1} \bar{x}_{p-1}. \qquad (21)
\end{aligned}
$$

- LS hyperplane passes through the mean of the data points.
- Furthermore, the intercept of the LS fit will be zero if we first mean-center the data, by setting $y_i^c = y_i - \bar{y}$ and $x_{ij}^c = x_{ij} - \bar{x}_j$ for each $i = 1, \ldots, n$ and $j = 1 \ldots, p-1$.
- From (21) we see indeed that

$$\hat{\beta}_0^c = \bar{y}^c - \hat{\beta}_1^c \bar{x}_1^c - \ldots - \hat{\beta}_{p-1}^c \bar{x}_{p-1}^c = 0$$

- Since, e.g., $\frac{1}{n}\sum_i (x_{i1} - \bar{x}_1) = \bar{x}_1 - \bar{x}_1 = 0$

Properties and geometrical interpretation

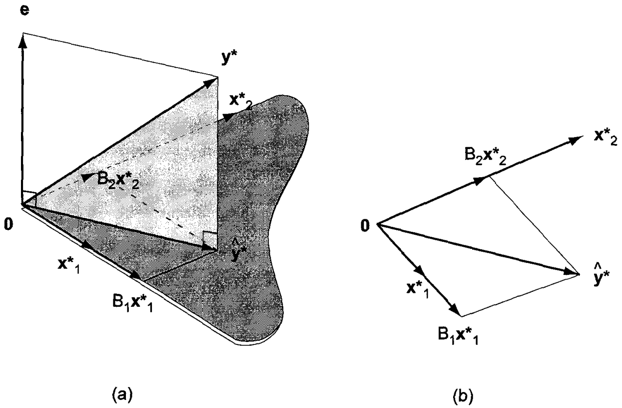


Figure 10.6. The vector geometry of least-squares fit in multiple regression, with the variables in mean-deviation form. The vectors $\mathbf{y}^*$, $\mathbf{x}_1^*$, and $\mathbf{x}_2^*$ span a three-dimensional subspace, shown in (*a*). The fitted Y vector, $\hat{\mathbf{y}}^*$, is the orthogonal projection of $\mathbf{y}^*$ onto the plane spanned by $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$. The $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane is shown in (*b*).

And what about σ?

- The variance of the errors σ^2 can be estimated by the mean squared error (MSE):

$$\boxed{\hat{\sigma}^2 = s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2.}$$

- Following (17) the variance-covariance matrix of the residuals is then estimated by

$$\hat{\Sigma}(\mathbf{e}) = s^2(I_n - H) = \mathrm{MSE}(I_n - H) \tag{22}$$

Statistical properties of the LS estimator

Under the Gauss-Markov conditions (2), (3) and (4), the following properties hold:

1. The least squares estimator $\hat{\boldsymbol{\beta}}_{LS}$ is an unbiased and consistent estimator of $\boldsymbol{\beta}$.
2. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{LS}$ is given by:
$$\Sigma(\hat{\boldsymbol{\beta}}_{LS}) = \sigma^2 (X^t X)^{-1}. \tag{23}$$
3. Gauss-Markov theorem: $\hat{\boldsymbol{\beta}}_{LS}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, i.e. any other linear and unbiased estimator of the form $A\mathbf{y}$ has a larger variance than $\hat{\boldsymbol{\beta}}_{LS}$.

Statistical properties of the LS estimator

Under the Gauss-Markov conditions (2), (3) and (4), the following properties hold:

4. The MSE s^2 is an unbiased and consistent estimator of σ^2.
5. $s^2(X^tX)^{-1}$ is an unbiased and consistent estimator of $\sigma^2(X^tX)^{-1}$.
6. If the errors $\boldsymbol{\epsilon}$ are normally distributed, $\hat{\boldsymbol{\beta}}_{LS}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$. The maximum likelihood estimator of σ^2 is given by

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{24}$$

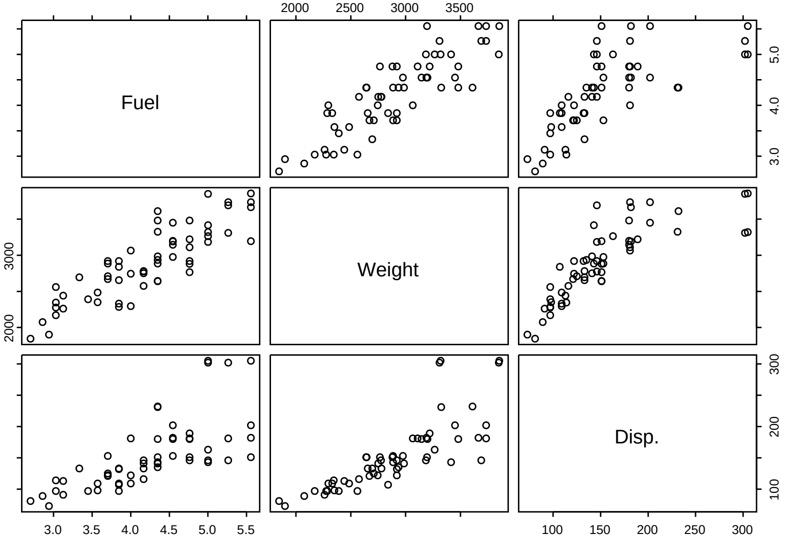
R example: Fuel

- Information of 60 cars
- 5 variables:
  - `Weight`: the weight of the car in pounds
  - `Disp.`: the engine displacement in liters
  - `Mileage`: gas mileage in miles/gallon
  - `Fuel`: fuel consumption in gallons per 100 miles (100/Mileage)
  - `Type`: a factor giving the general type of car, with levels: Small, Sporty, Compact, Medium, Large, Van
- We want to predict the fuel consumption of a car by its weight and engine displacement. The postulated model is:

$$\text{Fuel}_i = \beta_0 + \beta_1 \text{ Weight}_i + \beta_2 \text{ Disp}_i + \epsilon_i,$$

with $\epsilon_i \sim N(0, \sigma^2)$

R example: Fuel



R example: Fuel

```
Fuelfit <- lm(Fuel~Weight+Disp.)
Fuelsum <- summary(Fuelfit)
Fuelsum

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.4789731  0.3417877   1.401    0.167
Weight      0.0012414  0.0001720   7.220 1.37e-09 ***
Disp.       0.0008544  0.0015743   0.543    0.589

Residual standard error: 0.3901 on 57 degrees of freedom
Multiple R-squared: 0.7438, Adjusted R-squared: 0.7348
F-statistic: 82.75 on 2 and 57 DF,  p-value: < 2.2e-16
```

R example: Fuel

The fitted model is thus:

$$\hat{\text{Fuel}}_i = 0.48 + 0.0012 \text{ Weight}_i + 0.00085 \text{ Disp}_i$$

with $\hat{\sigma} = 0.39$.

Analysis of Variance

For an individual observation we have the identity

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

which can be illustrated for simple regression as follows.

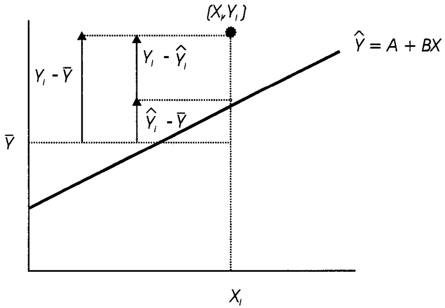


Figure 5.4. Decomposition of the total deviation $Y_i - \overline{Y}$ into components $Y_i - \hat{Y}_i$ and $\hat{Y}_i - \overline{Y}$.

Analysis of Variance

Squaring both sides of the equation and summing over all observations gives the ANOVA decomposition (check this!)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{25}$$

In other words: the total variation (SST) in the response $\mathbf{y}$ can be decomposed into an

- 'explained' component due to the regression (SSR) and an
- 'unexplained' component due to the errors (SSE).

$$\boxed{\text{SST} = \text{SSR} + \text{SSE}} \tag{26}$$

with degrees of freedom $n-1, p-1$ and $n-p$.

Analysis of Variance

The mean squares are defined as the sum of squares divided by their degrees of freedom:

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{p-1}, \mathrm{MSE} = \frac{\mathrm{SSE}}{n-p}$$

TABLE 6.1 ANOVA Table for General Linear Regression Model (6.19).

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$	$p-1$	$MSR = \frac{SSR}{p-1}$
Error	$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$
Total	$SSTO = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$	$n-1$	

Analysis of Variance

The *coefficient of multiple determination*:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{27}$$

$$= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{28}$$

- proportion of the *total variation in the response* **y** that is explained by the linear model (1)
- $0 \leq R^2 \leq 1$

Analysis of Variance

The *coefficient of multiple determination - remarks*

- In simple regression R^2 coincides with the squared correlation coefficient r^2 between $X = X_1$ and Y.
- A high value of R^2 does not necessarily imply that the fitted model is useful to make predictions.
- One can always increase R^2 by adding variables to the model. Therefore the *adjusted coefficient of determination* R_a^2 corrects for the number of variables:

$$R_a^2 = 1 - \frac{\mathrm{SSE}/(n-p)}{\mathrm{SST}/(n-1)} \tag{29}$$

Analysis of Variance

The extra sum of squares:

- The marginal *reduction in the error sum of squares* (SSE) when one or several predictor variables are added to the regression model, given that other variables are already in the model.
- E.g., when adding a new predictor X_2, to a model that already has predictor X_1, it is defined as:

$$\text{SSR}(X_2|X_1) = \text{SSE}(X_1) - \text{SSE}(X_1, X_2) \tag{30}$$

or equivalently

$$\text{SSR}(X_2|X_1) = \text{SSR}(X_1, X_2) - \text{SSR}(X_1) \tag{31}$$

Analysis of Variance

The extra sum of squares:

- Thus,

$$\mathrm{SST} = \mathrm{SSR}(X_1) + \mathrm{SSR}(X_2|X_1) + \mathrm{SSE}(X_1, X_2)$$

or as

$$\mathrm{SST} = \mathrm{SSR}(X_2) + \mathrm{SSR}(X_1|X_2) + \mathrm{SSE}(X_1, X_2)$$

- We can thus decompose the SSR of the full model into several extra sum of squares.
- The degrees of freedom associated with each sum of squares is equal to the number of variables that are added to the model.

Analysis of Variance

The extra sum of squares:

TABLE 7.3 Example of ANOVA Table with Decomposition of *SSR* for Three X Variables.

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2 \mid X_1$	$SSR(X_2 \mid X_1)$	1	$MSR(X_2 \mid X_1)$
$X_3 \mid X_1, X_2$	$SSR(X_3 \mid X_1, X_2)$	1	$MSR(X_3 \mid X_1, X_2)$
Error	$SSE(X_1, X_2, X_3)$	$n-4$	$MSE(X_1, X_2, X_3)$
Total	$SSTO$	$n-1$	

Analysis of Variance

The ANOVA analysis of the (fuel.frame) data set yields:

```
> summary(aov(Fuelfit))
            Df  Sum Sq Mean Sq  F value Pr(>F)
Weight       1 25.1388 25.1388 165.2090 <2e-16 ***
Disp.        1  0.0448  0.0448   0.2945 0.5895
Residuals   57  8.6733  0.1522
```

vs

```
> Fuelfit2<- lm(Fuel~Disp.+Weight)
> summary(aov(Fuelfit2))
            Df Sum Sq Mean Sq F value   Pr(>F)
Disp.        1 17.253  17.253  113.38 3.58e-15 ***
Weight       1  7.931   7.931   52.12 1.37e-09 ***
Residuals   57  8.673   0.152
```

Equivariance Properties

1. $\hat{\boldsymbol{\beta}}_{LS}$ is regression equivariant: for any vector $\mathbf{v}$,

$$\hat{\boldsymbol{\beta}}_{LS}(\mathbf{x}_i, y_i + \mathbf{x}_i^t \mathbf{v}) = \hat{\boldsymbol{\beta}}_{LS}(\mathbf{x}_i, y_i) + \mathbf{v}$$

2. $\hat{\boldsymbol{\beta}}_{LS}$ is scale equivariant: for any constant c,

$$\hat{\boldsymbol{\beta}}_{LS}(\mathbf{x}_i, cy_i) = c\hat{\boldsymbol{\beta}}_{LS}(\mathbf{x}_i, y_i) \text{ and } \hat{\sigma}^2_{LS}(\mathbf{x}_i, cy_i) = c^2\hat{\sigma}^2_{LS}(\mathbf{x}_i, y_i)$$

3. $\hat{\boldsymbol{\beta}}_{LS}$ is affine equivariant: for any non-singular $p \times p$ matrix,

$$\hat{\boldsymbol{\beta}}_{LS}(A\mathbf{x}_i, y_i) = (A^t)^{-1}\hat{\boldsymbol{\beta}}_{LS}(\mathbf{x}_i, y_i)$$

The standardized regression model

Calculating the inverse of X^tX will be sensitive to roundoff errors, when

1. the determinant of X^tX is close to zero (multicollinearity! We will deal with this later)
2. the elements of X^tX differ significantly in order of magnitude, which occurs when the predictor variables have substantially different magnitudes.

The standardized regression model: transform the X (and Y) variables such that the new X^tX matrix corresponds with the correlation matrix of the original X-variables (its entries are bounded by -1 and 1 and thus are less sensitive to roundoff errors).

The standardized regression model

The *correlation transformation* is defined for each observation $i = 1, \ldots, n$ and for each variable $j = 1, \ldots, p-1$ as:

$$x'_{ij} = \frac{1}{\sqrt{n-1}} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right) \tag{32}$$

$$y'_i = \frac{1}{\sqrt{n-1}} \left(\frac{y_i - \bar{y}}{s_Y} \right) \tag{33}$$

with s_j resp. s_Y the standard deviation of X_j resp. Y.

The standardized regression model

Using (11) and (12) we then obtain for the transformed variables:

$$
\begin{aligned}
((X')^t X')_{jk} &= \sum_{i=1}^{n} x'_{ij} x'_{ik} \\
&= \frac{1}{n-1} \frac{\sum (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{s_j s_k} \\
&= \frac{\mathrm{cov}(X_j, X_k)}{s_j s_k} = r_{jk} \\
((X')^t X')_{jj} &= \frac{s_j^2}{s_j s_j} = 1 \\
((X')^t y')_j &= \frac{\mathrm{cov}(X_j, Y)}{s_j s_Y} = r_{jy}
\end{aligned}
$$

with r_{jk} the simple correlation between X_j and X_k, and r_{jy} the correlation between X_j and Y.

The standardized regression model

In terms of the transformed variables, the general linear model (1)

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

now becomes

$$y_i - \bar{y} = \beta_1 (x_{i1} - \bar{x}_1) + \ldots + \beta_{p-1}(x_{i,p-1} - \bar{x}_{p-1}) + \epsilon_i$$

(drop the intercept term because mean-centering) and thus

$$\frac{y_i - \bar{y}}{s_Y} = \beta_1 \frac{s_1}{s_Y} \left(\frac{x_{i1} - \bar{x}_1}{s_1} \right) + \ldots + \beta_{p-1} \frac{s_{p-1}}{s_Y} \left(\frac{x_{i,p-1} - \bar{x}_{p-1}}{s_{p-1}} \right) + \frac{\epsilon_i}{s_Y}.$$

Divide each term by $\sqrt{n-1}$ to obtain the *standardized regression model*

$$y_i' = \beta_1' x_{i1}' + \beta_2' x_{i2}' + \ldots + \beta_{p-1}' x_{i,p-1}' + \epsilon_i' \qquad (34)$$

The standardized regression model

for $i = 1, \ldots, n$ with

$$\epsilon_i' = \frac{\epsilon_i}{\sqrt{n-1}s_Y} \tag{35}$$

$$\beta_j' = (\frac{s_j}{s_Y})\beta_j. \tag{36}$$

- β_j' are often called the *standardized regression coefficients.*
- The least squares estimates satisfy:

$$\hat{\beta}' = ((X')^t X')^{-1}(X')^t y' = R_{XX}^{-1} r_{XY}. \tag{37}$$

- R_{XX} is the correlation matrix of X
- $r_{XY} = (r_{1y}, \ldots, r_{p-1,y})^t$ contains the correlations between each predictor and the response.

Return to the estimates with respect to the original variables via:

$$\hat{\beta}_j = (\frac{s_Y}{s_j})\hat{\beta}_j' \tag{38}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \ldots - \hat{\beta}_{p-1}\bar{x}_{p-1}.$$

The standardized regression model

Example.

- Portrait sales in community (Y, expressed in \$1000)
- Number of persons aged 16 or younger in that community (X_1 in thousands of persons)
- Per capita personal income (X_2 in \$1000)

The model using the original variables yields

$$\hat{y}_i = -68.86 + 1.45x_{i1} + 9.36x_{i2}.$$

The standardized regression model yields

$$\hat{y}_i = -68.86 + 1.45x_{i1} + 9.36x_{i2}.$$

The standardized regression model

Example.

TABLE 7.5 Correlation Transformation and Fitted Standardized Regression Model—Dwaine Studios Example.

(a) Original Data

Case i	Sales Y_i	Target Population X_{i1}	Per Capita Disposable Income X_{i2}
1	174.4	68.5	16.7
2	164.4	45.2	16.8
...	...	...	...
20	224.1	82.7	19.1
21	166.5	52.3	16.0
	$\bar{Y} = 181.90$	$\bar{X}_1 = 62.019$	$\bar{X}_2 = 17.143$
	$s_Y = 36.191$	$s_1 = 18.620$	$s_2 = .97035$

(b) Transformed Data

i	Y'_i	X'_{i1}	X'_{i2}
1	−.04637	.07783	−.10205
2	−.10815	−.20198	−.07901
...	...	...	...
20	.26070	.24835	.45100
21	−.09518	−.11671	−.26336

(c) Fitted Standardized Model

$$\hat{Y}' = .7484X'_1 + .2511X'_2$$