REGRESSION AND CORRELATION

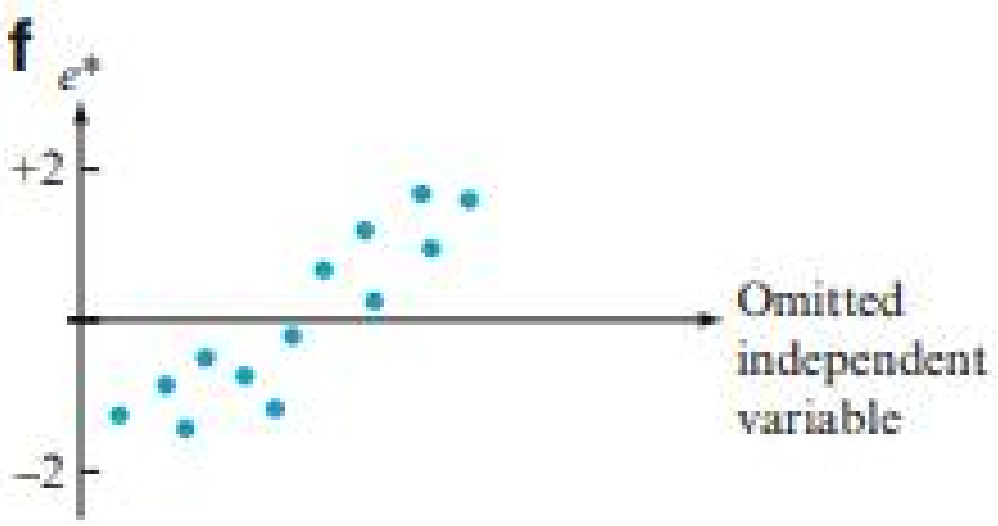
MULTIPLE REGRESSION ANALYSIS

SECTION 12.7 (DEVORE & BERK 2012)

MULTIPLE REGRESSION MODEL

Sometimes, the behavior of the dependent variable y cannot be explained by only one predictor.

We denote the number of predictors with k – if larger than 1.



Example: Let y = selling price of a house. Then we might have $k = 3$, with x_1 = size (ft^2), x_2 = age (years), and x_3 = number of rooms.

Definition: The general additive multiple regression model equation is given by

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where

- $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$ (constant variance),
- ε is normally distributed,
- the ε's associated with various observations are independent of one another $\Rightarrow$ the Y_i's are independent of one another.

ADJUSTED COEFFICIENT OF DETERMINATION

- As for the simple regression, we define

$$\text{SSE} = \sum(y_i - \hat{y})^2,\ \text{SST} = \sum(y_i - \bar{y})^2.$$

- The coefficient of (multiple) determination is

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}.$$

R^2 is interpreted as the proportion of observed variation than can by explained by the model relationship.

Note: The value of R^2 can be inflated by including predictors in the model that are relatively unimportant, or even frivolous.

ADJUSTED COEFFICIENT OF DETERMINATION

To avoid the risk of including "too many" predictors, we define the adjusted coefficient of (multiple) determination as follows:

$$R_a^2 = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{\text{SSE}/[n-(k+1)]}{\text{SST}/(n-1)}$$

$$= 1 - \frac{n-1}{n-(k+1)}\frac{\text{SSE}}{\text{SST}}$$

- In general $R_a^2 \leq R^2$ (since usually $k < n$).
- Rule of thumb: if $R_a^2 \ll R^2$, then the chosen model has too many predictors relative to the amount of data.
- R_a^2 is not a model selection criterion. However, it is often used to in-/exclude predictors in automated procedures

MODEL UTILITY TEST

The idea is similar to the model utility test for simple regression model, but we need to change null and alternative hypothesis.

- Null hypothesis H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
- Alternative hypothesis H_a: at least one $\beta_i \neq 0$ $(i = 1, \ldots, k)$
- Test statistic:

$$f = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} = \frac{\text{MSR}}{\text{MSE}}$$

  where SSR = regression sum of squares = SST − SSE

- Rejection region for a level α test: $f \geq F_{\alpha,k,n-(k+1)}$
- T-tests serve for testing separate hypothesis on single coefficients

POLYNOMIAL REGRESSION

Assume that a scatter plot shows a parabolic rather than linear space. Then it is natural to specify a quadratic regression model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Note

- we can still see the quadratic regression as a multilinear model. To see this, define $x_1 := x, x_2 := x^2$. Then, $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
- In particular, we say that quadratic (or polynomial) regression is a special case of multiple regression.

Attention

- The interpretation of the β_i's is different to the common multilinear model. This is because the value of $x_2 = x^2$ cannot be increased while $x_1 = x$ is held fixed.
- Moreover, the interpretation of regression coefficients requires extra care when some predictor variables are mathematical functions of others (keywords: multicollinearity, variance inflation factors).
- In case of polynomial regression, problems related to multicollinearity can be largely avoided by using orthogonal polynomials.

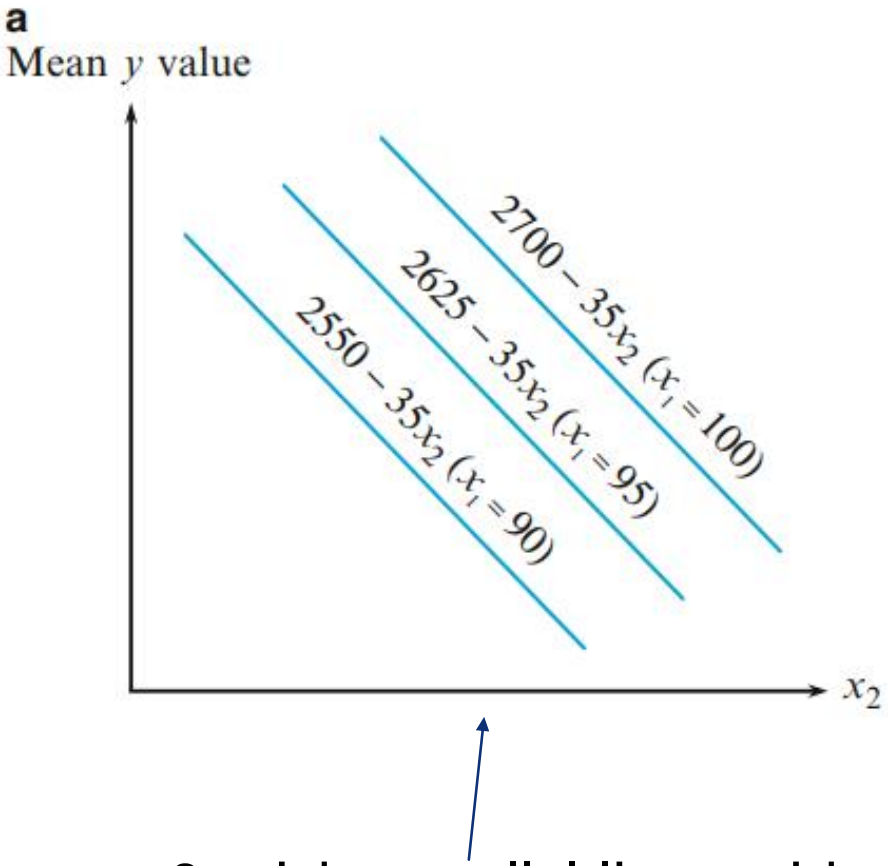
MODELS WITH INTERACTION

Example: Suppose that an industrial chemist is interested in the relationship between product yield y from a certain reaction and two independent variables x_1 = reaction temperature and x_2 = pressure at which the reaction is carried out.

The first relationship proposed is

$$Y = 1200 + 15x_1 - 35x_2 + \varepsilon$$

for temperature values between 80 and 100 in combination with pressure values ranging from 50 to 70. The population regression function gives the mean y value for any particular values of the predictor.



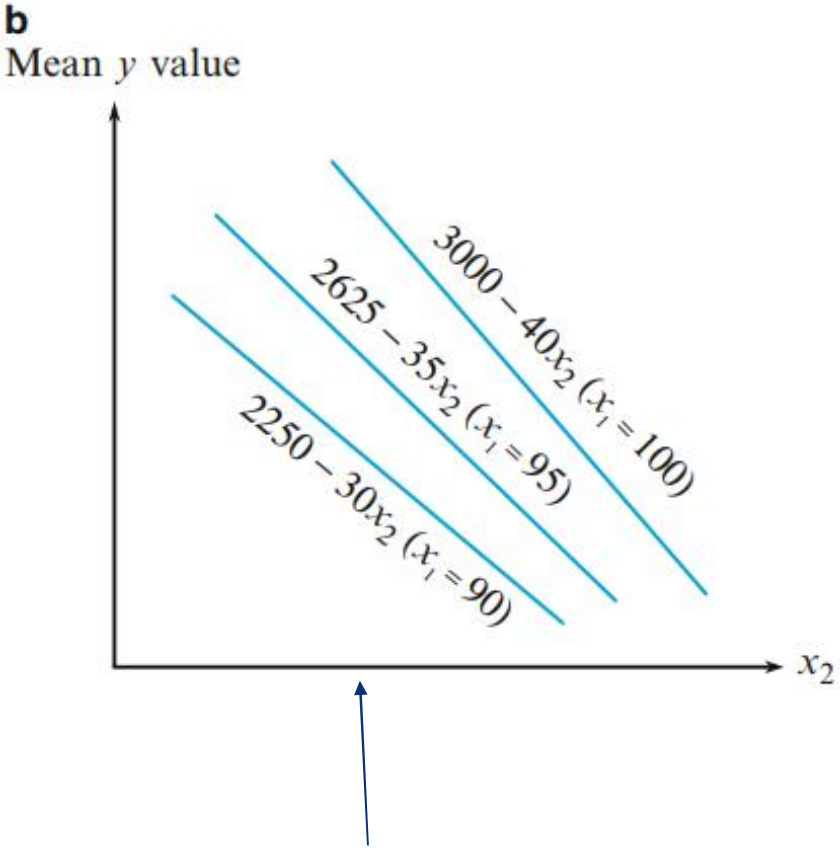
Straight parallel lines, with the same slope: −35

MODEL WITH INTERACTION

Example – continued: Now assume that the chemist has reason to doubt the appropriateness of the proposed model. He believes the following: when the pressure x_2 increases, the decline in average yield should be more rapid for a high temperature that for a low temperature.

Hence, rather than the lines being parallel, the line for a temperature of 100 should be steeper than the line for a temperature of 95. A model that has this property has a third predictor variable, $x_3 = x_1x_2$. One such model is

$$Y = -4500 + 75x_1 + 60x_2 - x_1x_2 + \varepsilon$$



Straight lines, but NOT parallel – they have different slopes

INTERACTION

Definition: If the change in the mean y value associated with a 1-unit increase in one independent variable depends on the value of a second independent variable, there is interaction between these two variables.
Denoting the two independent variables by x_1 and x_2, we can model this interaction by including as an additional predictor $x_3 := x_1 x_2$, the product of two independent variables. The model equation then becomes

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \text{ where } x_3 = x_1 x_2.$$

Definition: the full quadratic or complete second-order model is defined as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2.$$

ANALYSIS OF COVARIANCE

SECTION 12.7 (DEVORE & BERK 2012)

CATEGORICAL VARIABLES

Basic idea: using simple numerical coding, qualitative (categorical) variables can also be incorporated into a linear regression model.

Examples: type of college (private or state) or type of wood (pine, oak, or walnut) could serve as predictors.

We first focus on the case of a dichotomous variable, one with only two possible categories (e.g. male / female). We associate a dummy or indicator variable x whose possible values 0 and 1 indicate the category.

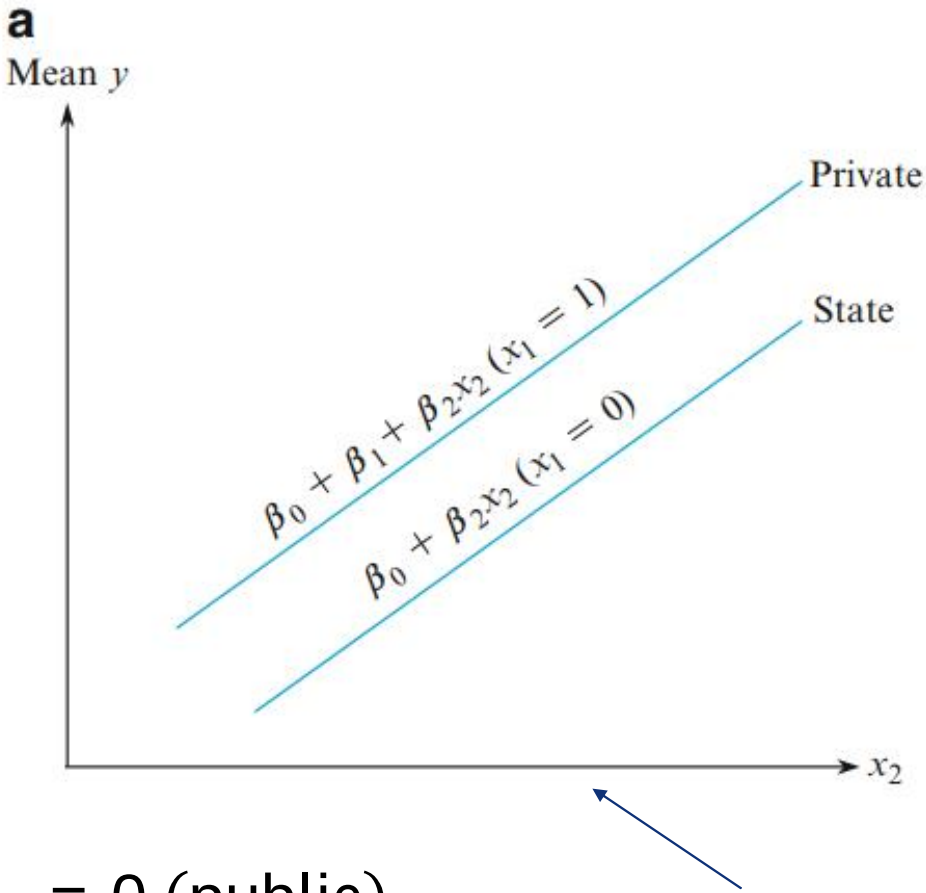
EXAMPLE 12.31

Example: Assume we have graduation rate data.
We use a model with y = graduation rate,
x_2 = average freshman SAT score, and
x_1 = a dummy variable which indicates private or public status, i.e.,

$$x_1 = \begin{cases} 1 \text{ if the university is private} \\ 0 \text{ if the university is public} \end{cases}$$

Consider the multiple regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$



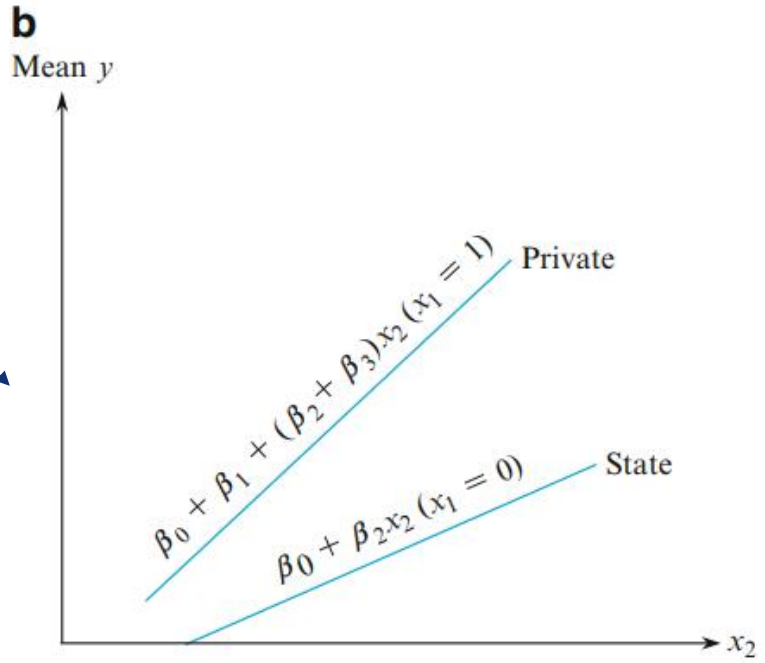
Same slope!

then:

- mean graduation rate = $\beta_0 + \beta_2 x_2$ when $x_1 = 0$ (public)
- mean graduation rate = $\beta_0 + \beta_1 + \beta_2 x_2$ when $x_1 = 1$ (private)

EXAMPLE 12.31

Different slopes!



A second possibility is a model with an interaction term:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Now the mean of graduation rates for the two types of university are given by:

- mean graduation rate $= \beta_0 + \beta_2 x_2$ when $x_1 = 0$ (public)
- mean graduation rate $= \beta_0 + \beta_1 + (\beta_2 + \beta_3) x_2$ when $x_1 = 1$ (private)

MORE THAN TWO CATEGORIES

If our predictor has three possible categories, we need to define two different dummy variables – and so forth, i.e. one dummy variable less than number of categories.

Example: Assume that we have data about the grades (y) of 200 university students. The predictors are the high school grades (x_1) and the department the students belong to, Biology, Mathematics, or Medicine. Then we need to define two predictors:

$$x_2 = \begin{cases} 0 \text{ if the student does not study Mathematics} \\ 1 \quad \text{if the student studies Mathematics} \end{cases}$$

$$x_3 = \begin{cases} 0 \text{ if the student does not study Medicine} \\ 1 \quad \text{if the student studies Medicine} \end{cases}$$

Example – continued: Then, our model becomes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = \begin{cases} \beta_0 + \beta_1 x_1 & \text{if Dep} = \text{Biology} \\ \beta_0 + \beta_1 x_1 + \beta_2 & \text{if Dep} = \text{Mathematics} \\ \beta_0 + \beta_1 x_1 + \beta_3 & \text{if Dep} = \text{Medicine} \end{cases}$$

This means:

- The biologists are captures by the model intercept β_0
- Differences between biology and mathematics are captured by β_2, and β_3 models differences between biology and medicine

ANCOVA

Definitions: Analysis that involves both quantitative and categorical predictors, as in Example 12.31, is called analysis of covariance, or ANCOVA. The quantitative variable is often called a covariate.

Note:

- Sometimes more than one covariate / categorical predictor is used.
- ANCOVA is a combination of linear regression and ANOVA.
- An ANCOVA with categorical predictors only is equivalent to an ANOVA.

IMPLEMENTATION IN R

- Use the function `lm`
- Include continuous and qualitative predictors at the same time
- Creation of dummy variables is only necessary for very particular model specifications

Example: Several model specifications are possible. Let cont be the continuous predictor and fac the qualitative predictor.

1. Varying intercept - `lm(y ~ fac + cont)`
2. Varying slope - `lm(y ~ cont + cont : fac)`
3. Varying intercept & slope - `lm(y ~ fac * cont)`

REGRESSION WITH MATRICES

SECTION 12.8 (DEVORE & BERK 2012)

When we work with multilinear regression, sometimes it is easier and more compact to write formulas using matrix notation.

Example [Multiplication of matrices]: $\begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \cdot y_1 + x_2 \cdot y_2 \\ x_3 \cdot y_1 + x_4 \cdot y_2 \end{bmatrix}$

First row

Second row

$\begin{bmatrix} 2 & 3 \\ 4 & 4 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \cdot 1 + 3 \cdot 2 \\ 4 \cdot 1 + 4 \cdot 2 \end{bmatrix} = \begin{bmatrix} 8 \\ 12 \end{bmatrix}$

THE NORMAL EQUATIONS

Suppose that we have n observations, each consisting of a y value and values of the k predictors. We then have:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{bmatrix}$$

These model equations can be written much more compactly using vectors and matrices. One usually uses the following notation:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

$k + 1$ (columns) n (rows)

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

As before, we estimate $\beta_0, \beta_1, \ldots, \beta_k$ using the principle of least squares: i.e., find $b_0, b_1, \ldots, b_n$ to minimize

$$\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})]^2 = (\boldsymbol{y} - \boldsymbol{Xb})'(\boldsymbol{y} - \boldsymbol{Xb}) = \|\boldsymbol{y} - \boldsymbol{Xb}\|^2$$

where $\boldsymbol{b} = [b_0, b_1, \ldots, b_k]'$ and $\|\boldsymbol{u}\|$ is the length of u.

After some computations, we can see that the vector of estimated coefficients is through the normal equation

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{b} = [\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{y}$$

PREDICTED VALUES

We can write the predicted values in a matrix form:

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}[\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Because $\hat{\boldsymbol{y}}$ (y-hat) is the product of $\boldsymbol{H} := \boldsymbol{X}[\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'$ and $\boldsymbol{y}$, the matrix $\boldsymbol{H}$ is called the hat matrix.

Note: a residual is defined by $y_i - \hat{y}_i$, so the vector of n residuals is given by

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}$$
$$= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

COVARIANCE MATRICES

In order to develop hypothesis tests and confidence intervals for the regression coefficients, the standard deviations of the estimated coefficients are needed.

These can be obtained from a so-called (variance-)covariance matrix. This matrix

- is a square matrix and contains
- the variances on the main diagonal and
- the covariances in the off-diagonal elements.

COVARIANCE MATRICES

Let $\boldsymbol{U} = [U_1, \dots, U_n]'$ a random vector (n-dimensional random variable) and means $\mu_1 = E(U_1), \dots, \mu_n = E(U_n)$, and let $\boldsymbol{\mu} = [\mu_1, \dots, \mu_n]'$.

The covariance matrix can then be calculated by:

$$\mathrm{Cov}(\mathbf{U}) = \begin{bmatrix} \mathrm{Cov}(U_1, U_1) & \cdots & \mathrm{Cov}(U_1, U_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(U_n, U_1) & \cdots & \mathrm{Cov}(U_n, U_n) \end{bmatrix} = \mathrm{E}\{[\boldsymbol{U} - \boldsymbol{\mu}]\,[\boldsymbol{U} - \boldsymbol{\mu}]'\}$$

When $n = 1$ this reduces to just the ordinary variance.