stats545.1 545.1 Point estimation and variances in the linear model

Gunnar Stefansson

3. september 2022

Efnisyfirlit

1 Problem statement and estimators

1.1 Multiple linear regression problem

For y-observations, we want descriptive and predictive linear model of several variables
$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

or, rather $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i$

Formulate with matrices...

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

Note that intercept is implicit...
Statistical assumptions will be handled later!

1.1.1 Details

Consider the generic problem of fitting a model to data as a simple estimation problem. Later we will add statistical assumption in order to draw formal conclusions, but in this section we will only consider point estimation.
When collecting measurements of a dependent variable, i.e. y-observations, it is common at the same time to have measurements of several independent x-variables.

In this case one needs a descriptive and predictive linear model of several (say p) variables, i.e. a model of the form: $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$. In this notation there is no distinction between a multiplier (β_j) for a general x-measurement and the intercept. An "intercept", α, is implemented simply by setting $x_1 = 1$ and $\alpha = \beta_1$.

In practise several y-measurements will be made, say n. This can be formulated in matrix notation viz

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

where the n-vector $\mathbf{y}$ contains all the y-measurements and the $n \times p$ matrix contains all the independent variables.

1.1.2 Examples

Example 1.1. When a straight line is not an appropriate model for explaining the relationship between pairs of measurements, (x_i, y_i), it is possible to consider a quadratic response function, i.e. define the model $EY_i = \alpha + \beta x_i + \gamma x_i^2, \quad i = 1, \ldots, n$. Defining $x_{i1} = 1$, $x_{i2} = x_i$, $x_{i3} = x_i^2$, this becomes a multiple linear regression model.

This example illustrates clearly how the multiple linear regression model refers to **linearity in the unknown parameters**, not in the independent variables.

Example 1.2. Consider the following data set (from Stefansson, Skuladottir and Petursson) of indices from Icelandic waters. Here T=temperature, U=catch per unit effort of (adult) shrimp, I=index of juvenile shrimp abundance, Y=catch of shrimp, B=biomass of capelin, G=measure of growth of cod from age 4 to 5, S=biomass of spawning cod, J=biomass of juvenile (immature) cod. This forms the **ecosystem example** to be used several times in this tutorial.

```
t T U I Y B G S J
79 0.5 75.7 2313 1.1 3177 809 447 872
80 5.7 79.8 4747 3.1 2210 777 602 880
81 2.7 77.6 3217 2.1 1442 398 389 704
82 2.7 76.4 1909 1.7 1128 595 266 623
83 1.2 85.0 4368 6.1 2182 725 214 584
84 3.5 86.0 2418 12.2 3579 997 219 605
85 5.0 93.0 3930 12.2 3688 851 268 577
86 3.5 89.0 4943 17.1 3987 873 268 768
87 4.4 77.5 4309 24.6 3727 725 253 921
88 1.7 65.8 4089 20.7 2990 620 193 818
89 3.3 72.0 4994 18.1 2677 785 269 595
90 3.2 81.6 8180 19.4 2146 570 344 408
91 3.6 87.1 8406 26.1 2454 771 232 508
92 4.3 83.5 6376 27.4 3050 570 244 357
93 4.3 94.0 7192 30.1 3185 1004 224 358
94 4.7 104.6 9611 42.1 3119 675 276 292
95 0.3 87.6 9742 49.2 3700 857 380 189
```

For a data set such as this one several research questions are of interest. One such question is what factors affect the growth of cod, the predator in the system. To model cod growth as a function of the biomass of the two prey one can use the R formulation

```
G~U+B
```

and read the data with

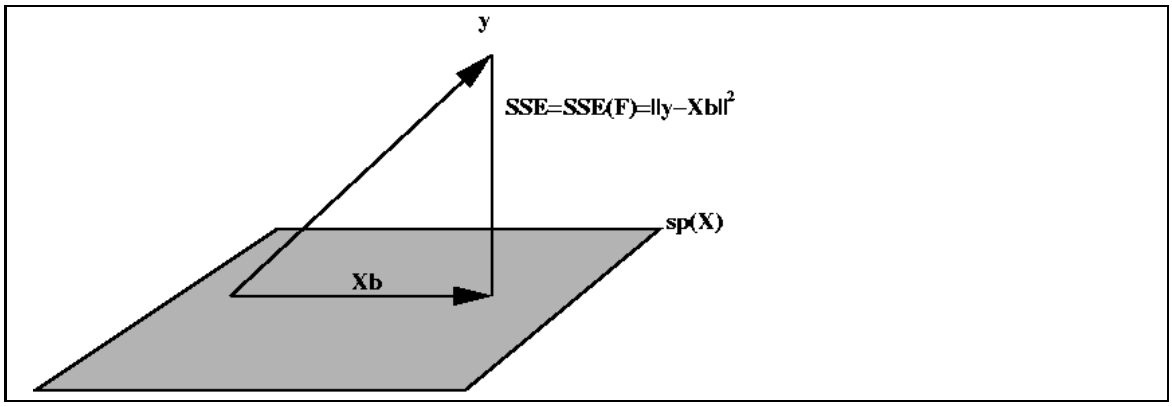
```
read.table("http://tutor-web.net/stats/stats545.1/lecture10/borecol-
   dat.txt",header=T)
```

since it is available on the web. To store the data as an R object and give it a name, a command of the form

```
m<-read.table("http://tutor-web.net/stats/stats545.1/lecture10/
   borecol-dat.txt",header=T)
```

is used.

1.2 Geometric visualization of the multiple regression problem



1.2.1 Details

The least squares problem estimates parameters, $\hat{\beta}_1, \ldots, \hat{\beta}_p$ as those values of $b_1, \ldots, b_p$ which minimise the sum of squared deviations,

$$f(b_1, \ldots, b_p) := \sum_{i=1}^{n} (y_i - (b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}))^2$$

i.e. the estimates satisfy

$$f(\hat{\beta}_1, \ldots, \hat{\beta}_p) = min_{b_1,\ldots,b_p} f(b_1, \ldots, b_p).$$

The least squares problem now becomes the same as minimizing the norm of a difference, i.e. minimize

$$||\mathbf{y} - \mathbf{Xb}||^2$$

over all vectors $\mathbf{b}$.

Notice that $\mathbf{Xb}$ is a linear combination of the column vectors of the $\mathbf{X}$-matrix. The set, V, of all such combinations forms a subspace of $\mathbb{R}^n$, commonly denoted by $span(X)$ or $sp(X)$:

$$sp(X) := \{\mathbf{Xb} \in \mathbb{R}^n : \mathbf{b} \in \mathbb{R}^p\}$$

Geometrically the problem is therefore equivalent to finding a vector $\hat{\mathbf{y}}$ in the vector space V, which is closest to $\mathbf{y}$. From a geometric viewpoint this will be seen to be the orthogonal projection of $\mathbf{y}$ onto $sp(X)$.

The solution, $\hat{\mathbf{y}}$, will be of the form of linear combinations of the columns of the $\mathbf{X}$-matrix, i.e. $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ for some vector $\hat{\beta} \in \mathbb{R}^p$. The original data vector can now be written as the sum of two vectors: $\mathbf{y} = \hat{\mathbf{y}} + (\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{X}\hat{\beta} + (\mathbf{y} - \mathbf{X}\hat{\beta})$, which will be seen to be orthogonal.

1.2.2 Examples

Example 1.3. Consider the ecosystem example from before. To set up the X matrix, three columns are needed to reflect the intercept along with the shrimp biomass effect and the capelin biomass.

To extract columns from the data from m, one can either refer to the columns by name or number. Reference by number is done with

```
> m[,c(3,6)]
      U B
1 75.7 3177
2 79.8 2210
3 77.6 1442
4 76.4 1128
...
```

but it is much simple to use column names, as in

```
cols<-m[,c("U","B")]
```

with the dplyr package this becomes even easier:

```
library(dplyr)
selcols<-select(mmm,U,I)
```

but this is not the entire **X**-matrix since the column of all ones is missing. This is easy to add, however:

```
n<-length(m$U)
one<-rep(1,n)
X<-cbind(one,selcols)
y<-m$G
```

so X and y have thus been set up. To easily manipulate the vectors in the **X**-matrix one can also extract them from the data frame:

```
U<-m$U
B<-m$B
```

In this example $n = 17$ so $\mathbf{y} \in \mathbb{R}^{17}$ and the span of the columns of the **X**-matrix is now the three-dimensional subspace of $\mathbb{R}^{17}$ spanned by the three vectors called "one", "U" and "B" in R.

1.3 Normal equations

Have

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

1.3.1 Details

Suppose $\mathbf{y} \in \mathbb{R}^n$, and V is a subspace of $\mathbb{R}^n$.

An **orthogonal projection** of $\hat{\mathbf{y}}$ onto V is a vector, $\hat{\mathbf{y}} \in V$ such that $\mathbf{y} - \hat{\mathbf{y}} \perp V$. Now consider a vector, $\hat{\mathbf{y}}$ in $V = span(\mathbf{X})$, which can then be written as $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$. Assume it is a projection, so $\mathbf{y} - \hat{\mathbf{y}} \perp V$.

Now, let $\tilde{\mathbf{y}} = \mathbf{X}\tilde{\beta}$ be any other vector in V. Then

$$||\mathbf{y} - \tilde{\mathbf{y}}||^2 = ||(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \tilde{\mathbf{y}})||^2 = ||\mathbf{y} - \hat{\mathbf{y}}||^2 + ||\hat{\mathbf{y}} - \tilde{\mathbf{y}}||^2 \geq ||\mathbf{y} - \hat{\mathbf{y}}||^2$$

and we therefore see that such an orthogonal projection is the best one can do.
It also follows that that $\hat{\mathbf{y}}$ is **unique** since the only way $\tilde{\mathbf{y}}$ can get as close is by having $||\hat{\mathbf{y}} - \tilde{\mathbf{y}}|| = 0$, which only happens when they are equal.

In conclusion, we have shown that an orthogonal projection of $\mathbf{y} \in \mathbb{R}^n$ onto V is the **unique** element in V which is closest to $\mathbf{y}$. We now need to find a way to compute the projection. Next, since the residual vector, $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta}$, is orthogonal to each vector in V it must also be orthogonal to each column vector of $\mathbf{X}$, i.e. $\mathbf{x}_i' \left(\hat{\mathbf{y}} - \mathbf{X}\hat{\beta}\right) = 0$ and therefore $\mathbf{X}' \left(\hat{\mathbf{y}} - \mathbf{X}\hat{\beta}\right) = 0$.
Thus, the following **normal equations** describe how to find the parameters of the orthogonal projection, i.e. the parameters which give the best fit:

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}.$$

In general there is no guarantee that these equations have a unique solution and this is related to the rank of the $\mathbf{X}$-matrix itself.

1.4 The solution

Solution:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

Prediction:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}.$$

Estimated residuals:

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \left(I - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y}.$$

1.4.1 Details

When the matrix $\mathbf{X}'\mathbf{X}$ is invertible, the solution is well-known:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}.$$

It should be noted, however, that in actual implementations the point estimates can be obtained using numerical techniques which do not require inverting the matrix. However, the inverse is usually needed at a later stage.

The **predicted values** are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}.$$

The estimated (or observed) **residuals** are

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \left(I - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\mathbf{y}.$$

1.4.2 Examples

Example 1.4. Consider again the ecosystem data. The $\mathbf{X}$ matrix and $\mathbf{y}$-vector are set up as before. The slope, fitted values and errors can then be computed using matrix algebra:

```
m<-read.table("http://www.hi.is/~gunnar/kennsla/alsm/alsmintro/
   borecol.dat",header=T)
selcols<-m[,c("U","B")]
n<-length(m$U)
one<-rep(1,n)
```

```
X<-cbind(one,selcols)
X
   one U B
1 1 75.7 3177
2 1 79.8 2210
3 1 77.6 1442
4 1 76.4 1128
5 1 85.0 2182
6 1 86.0 3579
7 1 93.0 3688
8 1 89.0 3987
9 1 77.5 3727
10 1 65.8 2990
11 1 72.0 2677
12 1 81.6 2146
13 1 87.1 2454
14 1 83.5 3050
15 1 94.0 3185
16 1 104.6 3119
17 1 87.6 3700
X<-as.matrix(X)
y<-m$G
b<-solve(t(X)%*%X)%*%t(X)%*%y
yhat<-X%*%b
ehat<-y-yhat
b
          [,1]
one 171.9236911
U 2.8758166
B 0.1157401
```

Example 1.5. A much better approach is to use the R functions for linear models to compute these quantities:

```
lm(G~U+B,data=m)

Call:
lm(formula = G ~ U + B, data = m)

Coefficients:
(Intercept) U B
   171.9237 2.8758 0.1157
```

Naturally, the results are the same.

1.5 Sums of squares and norms

Sum of squared errors

$$SSE = ||\hat{\mathbf{e}}||^2 = \sum_i (y_i - \hat{y}_i)^2 .$$

Denote SSE by $SSE(F)$ or $SSE(R)$ when comparing models.

1.5.1 Details

The sum of squared errors becomes

$$SSE = ||\hat{\mathbf{e}}||^2 = ||\mathbf{y} - \hat{\mathbf{y}}||^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

When comparing models, e.g. a large or "full" model and a smaller or "reduced" model, the notation is usually extended to take into account the various models in question, notably $SSE(F)$ for the full model and $SSE(R)$ for the reduced model.

1.6 Projection matrices

Projecton, "hat", matrix onto $\mathbf{V} = sp(\mathbf{X})$:

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

and onto $\mathbf{V}^{\perp} = sp(\mathbf{X})^{\perp}$:

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

1.6.1 Details

The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a projection matrix (i.e. $\mathbf{H}^2 = \mathbf{H}$), projecting $\mathbf{R^n}$ onto the subspace $\mathbf{V} := sp(\mathbf{X})$. Conversely, $\mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the projection matrix onto $\mathbf{V}^{\perp} = sp(\mathbf{X})^{\perp}$, respectively.

The matrix H is usually termed the "hat matrix", since it transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$.

Note 1.1. The diagonal elements, h_{ij}, of the hat matrix play a very important role in regression diagnostics: If a certain data point has a high value on the diagonal, then this means that it "predicts itself", i.e. is influential.

References Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. 1996. Applied linear statistical models. McGraw-Hill, Boston. 1408pp.

2 General properties of linear projections of vectors of random variables

2.1 Linear combinations of independent random variables

c a column vector
Y a vector of independent random variables
Same σ, expected values may differ, $E[\mathbf{Y}] = \mu$
Then

$$E\left[\mathbf{c}'\mathbf{Y}\right] = \mathbf{c}'\mu$$
$$V\left[\mathbf{c}'\mathbf{Y}\right] = \mathbf{c}'\mathbf{c}\sigma^2$$

2.1.1 Details

Suppose **c** a column vector and **Y** a vector of independent random variables with a common variance, σ^2, but possibly different expected values. Then the mean and variance of the linear combination, $\mathbf{c}'\mathbf{Y}$, are given by

$$E\left[\mathbf{c}'\mathbf{Y}\right] = \mathbf{c}'\mu$$
$$V\left[\mathbf{c}'\mathbf{Y}\right] = \mathbf{c}'\mathbf{c}\sigma^2$$

These results are trivial to ascertain since the components, Y_i, are independent and hence e.g.

$$\begin{aligned} V\left[\mathbf{c}'\mathbf{Y}\right] &= V\left[\sum_i c_i Y_i\right] \\ &= \sum_i c_i^2 V\left[Y_i\right] \\ &= \mathbf{c}'\mathbf{c}\sigma^2 \end{aligned}$$

A number of shortcuts have been used here and it may be useful to note them before moving on to more complex examples. Although not mentioned, the random vector **Y** has a corresponding density f, which is a function of n variables satisfying

$$f(y_1,\ldots,y_n) \geq 0 \text{ for all } y_1,\ldots y_n$$

and

$$\int \ldots \int f(y_1,\ldots,y_n)dy_1 dy_n = 1$$

(where we have assumed this is a continuous random vector).
In vector notation this becomes

$$f(\mathbf{y}) \geq 0 \text{ for all } \mathbf{y} \in \mathbb{R}^n$$

and

$$\int f(\mathbf{y})d\mathbf{y} = 1.$$

For any given i the marginal density for Y_i is obtained by integrating out the other variables

$$f_{Y_i}(y_i) = \int \ldots \int f(y_1,\ldots,y_{i-1},y_i,y_{i+1},\ldots,y_n)dy_1 \ldots dy_{i-1}dy_{i+1}\ldots dy_n.$$

If μ_i is the expected value of Y_i, then this can be computed from either the univariate or multivariate density:

$$\mu_i = E[Y_i] = \int y_i f_{Y_i}(y_i) dy_i = \int \dots \int y_i f(y_1, \dots, y_n) dy_1 \dots dy_n.$$

The **covariance** between any two random variables Y_1 and Y_2 with joint density f and means μ_1, μ_2 is

$$cov(Y_1, Y_2) := E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$$

(assuming all relevant integrals exist).
We can now use the integral definitions to show (for $n = 2$) that

$$E[a_1 Y_1 + a_2 Y_2] = E[a_1 Y_1] + E[a_2 Y_2] = a_1 E[Y_1] + a_2 E[Y_2]$$

and so forth, for arbitrary n leading to

$$E\left[\mathbf{c}'\mathbf{Y}\right] = \mathbf{c}'\mu.$$

Correspondingly we obtain for the variance

$$\begin{aligned}
V[a_1 Y_1 + a_2 Y_2] &= E\left[((a_1 Y_1 + a_2 Y_2) - E[a_1 Y_1 + a_2 Y_2])^2\right] \\
&= E\left[((a_1 Y_1 - E[a_1 Y_1]) + (a_2 Y_2 - E[a_2 Y_2]))^2\right] \\
&= E\left[(a_1 (Y_1 - \mu_1) + a_2 (Y_2 - \mu_2))^2\right] \\
&= E\left[a_1^2 (Y_1 - \mu_1)^2 + a_2^2 (Y_2 - \mu_2)^2 + 2 a_1 a_2 (Y_1 - \mu_1)(Y_2 - \mu_2)\right] \\
&= a_1^2 V[Y_1] + a_2^2 V[Y_2] + 2 a_1 a_2 cov(Y_1, Y_2).
\end{aligned}$$

This result is the basis for all the variance computations in this entire section.

2.2 Covariance between linear combinations of independent random variables

a, **b** column vectors
Y a vector of independent random variables
Same σ, expected values may differ, $E[\mathbf{Y}] = \mu$
Then

$$Cov\left[\mathbf{a}'\mathbf{Y}, \mathbf{b}'\mathbf{Y}\right] = \mathbf{a}'\mathbf{b}\sigma^2$$

2.2.1 Details

Suppose **a**, **b** are column vectors and **Y** a vector of independent random variables with a common variance, σ^2, but possibly different expected values. Then the covariance between the linear combinations, $\mathbf{a}'\mathbf{Y}$ and $\mathbf{b}'\mathbf{Y}$, is given by

$$Cov\left[\mathbf{a}'\mathbf{Y}, \mathbf{b}'\mathbf{Y}\right] = \mathbf{a}'\mathbf{b}\sigma^2$$

This follows from looking at the linear combinations as sums of components and noting that the covariance is a sum of all possible combinations, all of which are zero except where the same Y_i-combinations appear:

$$
\begin{aligned}
Cov\left[\mathbf{a}'\mathbf{Y},\mathbf{b}'\mathbf{Y}\right] &= Cov\left[\sum_i a_i Y_i, \sum_j b_j Y_j\right] \\
&= \sum_{i,j} Cov\left[a_i Y_i, b_j Y_j\right] \\
&= \sum_{i,j} a_i b_j Cov\left[Y_i, Y_j\right] \\
&= \sum_i a_i b_i Cov\left[Y_i, Y_i\right] + \sum_{i,j:i\neq j} a_i b_j Cov\left[Y_i, Y_j\right] \\
&= \sum_i a_i b_i V\left[Y_i\right] \\
&= \mathbf{a}'\mathbf{b}\sigma^2
\end{aligned}
$$

This result indicates that if the projection vectors, **a** and **b** are orthogonal, then the covariance remains zero. Note also that strictly, independence of the original variables is not required, but only zero covariance which is not the same condition in the general case.
In the case of two Gaussian random variables, it is, however, true that they have zero covariance if and only if they are independent. This can be seen by observing the bivariate Gaussian density function which neatly factors if and only if the covariance is zero.

2.3 Linear projections of independent random variables

A an $n \times n$ matrix
Y a vector of n independent random variables, mean μ, $V[Y_i] = \sigma^2$.
Then

$$E\left[\mathbf{AY}\right] = \mu$$
$$V\left[\mathbf{AY}\right] = \mathbf{AA}'\sigma^2$$

2.3.1 Details

Let **A** be a $q \times n$ matrix and **Y** an n-vector of independent random variables with common variance but possibly different expected values, then

$$E\left[\mathbf{AY}\right] = \mathbf{A}\mu$$
$$V\left[\mathbf{AY}\right] = \mathbf{AA}'\sigma^2$$

This can be derived either by considering the componentwise composition of **AY** or by writing A as a collection of row vectors and using the earlier results.

2.3.2 Examples

Example: Assuming that all expected values exist, it is easy to derive the covariance $Cov(X+Y, X-Y)$, either directly or using the above formula, assuming $V[X] = V[Y]$.

2.4 Linear combinations of dependent random variables

$\mathbf{a} \in \mathbb{R}^n$ a vector
$\mathbf{Y}$ a vector of n random variables whose variances and covariances exist as a matrix, $\Sigma = \left(\sigma_{ij}\right)$ with $\sigma_{ij} = Cov(Y_i, Y_j)$.
Then

$$V\left[\mathbf{a}'\mathbf{Y}\right] = \mathbf{a}'\Sigma\mathbf{a}$$

2.4.1 Details

Let $\mathbf{a}$ be an n-vector or $n \times 1$ matrix and $\mathbf{Y}$ an n-vector of random variables whose variances and covariances exist as a matrix, $\Sigma = \left(\sigma_{ij}\right)$, where $\sigma_{ij} = Cov(Y_i, Y_j)$.
This is a typical case when looking at linear combinations of estimates in regression, for example estimating a new point on a regression line,

$$E\left[\hat{Y_{n+1}}\right] = \hat{\alpha} + \hat{\beta}x_{n+1}$$

As before, this can be derived by studying components.

2.5 Linear transformations of dependent random variables

$\mathbf{A}$ a matrix
$\mathbf{Y}$ a vector of random variables whose variances and covariances exist as a matrix, $\Sigma = \left(\sigma_{ij}\right)$ with $\sigma_{ij} = Cov(Y_i, Y_j)$.
Then

$$V\left[\mathbf{AY}\right] = \mathbf{A}\Sigma\mathbf{A}'$$

2.5.1 Details

Let $\mathbf{A}$ be an $n \times n$ matrix and $\mathbf{Y}$ a vector of random variables whose variances and covariances exist as a matrix, $\Sigma = \left(\sigma_{ij}\right)$, where $\sigma_{ij} = Cov(Y_i, Y_j)$.
This general situation occurs in regression analysis when measurements arrive in such a fashion that they can not be assumed to be independent. Several such examples certainly exist and the theory therefore needs to be properly developed.
This is also an important result when studying distributional properties of estimators, which are typically already linear combinations of original variables and hence no longer independent.
The first step is to derive the variance of projections of such variables. As before, this can be done by studying components or by looking at vector-wise linear combinations.
We obtain

$$V\left[\mathbf{AY}\right] = \mathbf{A}\Sigma\mathbf{A}'$$

3 Expected values and variances in multiple linear regression

3.1 Expected values in multiple linear regression

Expected values in multiple linear regression
If

$$E[\mathbf{y}] = \mathbf{X}\beta$$

and

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

then

$$E[\hat{\beta}] = \beta$$

- only depends on mean structure

3.1.1 Details

The estimator in multiple linear regression $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is unbiased.

This only depends on the assumption on the mean function, not on the variance structure, nor the probability distribution around the mean. In particular, the estimator is still unbiased even if the measurements are correlated.

3.1.2 Examples

Example 3.1. Sometimes an dependent variable does not vary in a simple linear fashion as a function of two independent variables as in $EY_i = \alpha + \beta x_i + \gamma w_i$. In particular, it may become obvious that the response, as a function of x, does not have the same slope for two different values of z. In this case an **interaction model** is required: $y_i = \alpha + \beta x_i + \gamma w_i + \delta x_i w_i$. Defining $x_{i1} = 1$, $x_{i2} = x_i$, $x_{i3} = w_i$, $x_{i4} = x_i w_i$, this becomes a multiple linear regression model.

3.2 Variances in multiple linear regression

If

$$V[\mathbf{y}] = \sigma^2 \mathbf{I}$$

and

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

then

$$\begin{aligned} & V\left[\hat{\beta}\right] \\ = \; & V\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] \\ = \; & \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) V[\mathbf{y}] \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\ = \; & \ldots \\ = \; & \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Depends on true variance structure - not on p.d.f.

3.2.1 Details

If $\mathbf{X}$ is of rank p, the estimator

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

in multiple linear regression has the variance-covariance matrix:

$$V\left[\hat{\beta}\right] = V\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] = \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) V[\mathbf{y}] \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' = \ldots = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}.$$

A consequence of this is that although numerical methods exist to estimate the coefficients, the inverse is required in order to obtain the variance-covariance matrix.

Note 3.1. This depends on true variance structure - not on a Gaussian assumption.

3.2.2 Examples

Example 3.2. The **one-way analysis of variance** is the analysis of data with the model

$$\begin{aligned} y_{1j} &= \mu_1 + e_{1j} \quad j = 1, \ldots, J_1 \\ y_{2j} &= \mu_2 + e_{2j} \quad j = 1, \ldots, J_2 \\ &\vdots \\ y_{Ij} &= \mu_I + e_{Ij} \quad j = 1, \ldots, J_I, \end{aligned}$$

i.e. measurements are made on each of I means, giving a total of $n = J_1 + \ldots + J_I$ measurements.

Assuming constant variance, the least squares estimators can be derived from the matrix form of the linear model. The basic model is of the form:

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1J_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2J_2} \\ \vdots \\ \vdots \\ \vdots \\ y_{I1} \\ y_{I2} \\ \vdots \\ y_{IJ_I} \end{bmatrix} = \begin{bmatrix} 1 & 0 & & 0 \\ 1 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 1 \\ 0 & 0 & & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{bmatrix} + \mathbf{e}$$

Here it is easy to evaluate the least squares estimators, their variances and covariances from the matrix representation.

3.3 Covariances between parameter estimates

Var-cov matrices also have correlations between estimates.

Also get numerical estimates of the var-cov matrix as well as all correlations once an estimate, $\hat{\sigma}^2$, of σ^2 becomes available.

3.3.1 Details

The above derives the theoretical formulae for the variance-covariance matrix, i.e. the true var-cov matrix. Naturally, this needs to be estimated based on data since it contains an unknown parameter.

Numerical estimates of the variances and covariances are obtained once an estimate, $\hat{\sigma}^2$, of σ^2 becomes available.

Note 3.2. Note that the estimates of covariances become unbiased if estimate of σ^2 are unbiased.

3.3.2 Examples

Example 3.3. Take the case of simple linear regression, with $\mathbf{X} = [\mathbf{1}\vdots\mathbf{x}]$, $\beta = (\alpha, \beta)'$ and the model for the data is $\mathbf{y} = \mathbf{X}\beta + e$. Here it is easy to derive the theoretical variances and covariance of α and β.

Example 3.4. Revisiting the ecology example, we can evaluate the standard errors, compute t-statistics and the like with the following R commands

```
m<-read.table("http://www.hi.is/~gunnar/kennsla/alsm/alsmintro/
   borecol.dat",header=T)
selcols<-m[,c("U","B")]
n<-length(m$U)
selcols<-m[,c("U","B")]
n<-length(m$U)
one<-rep(1,n)
X<-cbind(one,selcols) # The X-matrix
y<-m$G # The y-vector
p<-length(b) # The number of regressors
SSE<-sum((y-yhat)^2)
s2<-SSE/(n-p) # The estimate of sigma^2
varb<-s2*diag(XpXinv)
seb<-sqrt(varb) # The estimated s.e. of b
data.frame(Estimate=b,se=seb,t=b/seb,p=2*(1-pt(abs(b/seb),n-p)))
      Estimate se t p
one 171.9236911 284.2704735 0.6047891 0.55499548
U 2.8758166 3.6162040 0.7952584 0.43973854
B 0.1157401 0.0404542 2.8610155 0.01257369
```

As usual, a much better approach is to use the built-in functions in R, in this case lm and summary:

```
m<-read.table("http://www.hi.is/~gunnar/kennsla/alsm/alsmintro/
   borecol.dat",header=T)
fm<-lm(G~U+B,data=m)
summary(fm)

Call:
lm(formula = G ~ U + B, data = m)

Residuals:
    Min 1Q Median 3Q Max
-195.062 -87.215 4.916 72.809 193.117

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 171.92369 284.27047 0.605 0.5550
U 2.87582 3.61620 0.795 0.4397
B 0.11574 0.04045 2.861 0.0126 *
---


Residual standard error: 125 on 14 degrees of freedom
Multiple R-squared: 0.458, Adjusted R-squared: 0.3806
F-statistic: 5.915 on 2 and 14 DF, p-value: 0.01374
```

Naturally, the answers are the same.

References Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. 1996. Applied linear statistical models. McGraw-Hill, Boston. 1408pp.