stats545.1 Theory of linear models

Gunnar Stefansson

19. október 2016

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 $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

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

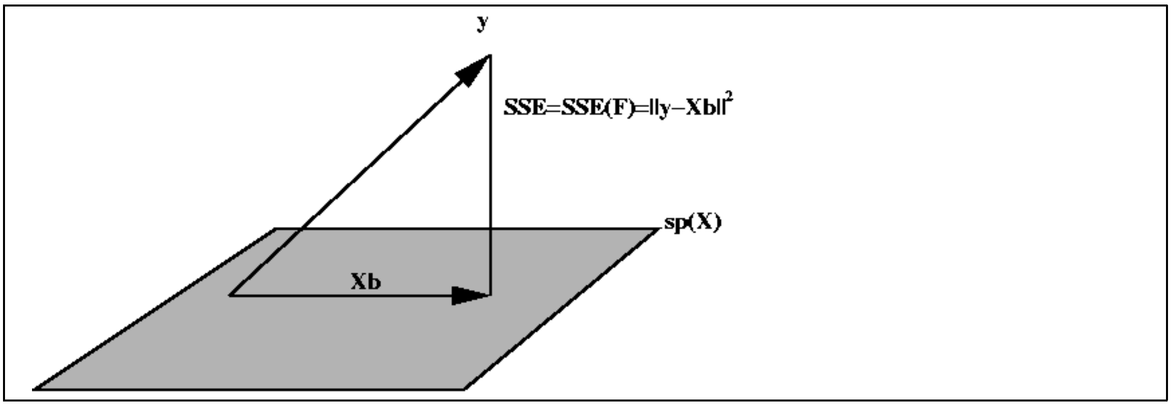
```
read.table("http://tutor-web.net/stats/stats545.1/lecture10/borecol-
   dat.txt",header=T)
```

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, $\hat{\beta}_1,\dots,\hat{\beta}_p$ as those values of $b_1,\dots,b_p$ which minimise the sum of squared deviations,

$$f(b_1,\dots,b_p) := \sum_{i=1}^{n}(y_i - (b_1x_{i1}+b_2x_{i2}+\dots+b_px_{ip}))^2$$

i.e. the estimates satisfy

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

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

$$||\mathbf{y}-\mathbf{X}\mathbf{b}||^2$$

over all vectors **b**.

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

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

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

The solution, $\hat{\mathbf{y}}$, will be of the form of linear combinations of the columns of the **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, $\mathbf{y} - \hat{\mathbf{y}} = \hat{\mathbf{y}} - \mathbf{X}\hat{\beta}$, 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 **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 **X** matrix and **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 Distributions 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.

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

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:

$$\begin{aligned}
Cov\left[\mathbf{a}'\mathbf{Y},\mathbf{b}'\mathbf{Y}\right] &= Cov\left[\sum_i a_iY_i,\sum_j b_jY_j\right] \\
&= \sum_{i,j} Cov\left[a_iY_i,b_jY_j\right] \\
&= \sum_{i,j} a_ib_jCov\left[Y_i,Y_j\right] \\
&= \sum_i a_ib_iCov\left[Y_i,Y_i\right]+\sum_{i,j:i\neq j} a_ib_jCov\left[Y_i,Y_j\right] \\
&= \sum_i a_ib_iV\left[Y_i\right] \\
&= \mathbf{a}'\mathbf{b}\sigma^2
\end{aligned}$$

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 transformations of dependent random variables

A a matrix
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

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

2.4.1 Details

Let **A** be an $n \times n$ matrix and **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

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

- 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

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

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:

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

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.

4 Orthogonal projections in multiple regression

4.1 Subspaces and degrees of freedom

Assume $rank(\mathbf{X}) = r$

We have $y - \mathbf{X}\hat{\beta} \perp \mathbf{X}\hat{\beta}$ so that $y - \mathbf{X}\hat{\beta} \in \mathbf{V}^{\perp} = \{v : v \perp sp(\mathbf{X})\}$ and $dim(\mathbf{V}^{\perp}) = n - r$.

$$\begin{aligned}\underbrace{\hat{\mathbf{e}}}_{n\times 1} &= y - \mathbf{X}\hat{\beta} = y - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')y = (\mathbf{I} - \mathbf{H})y\end{aligned}$$

and $rank(\mathbf{I} - \mathbf{H}) = dim(\mathbf{V}^{\perp}) = n - p$

4.1.1 Details

Assume $rank(\mathbf{X}) = r \leq p$ ($\mathbf{X}$ is $n \times p$).

Parameters in the model $\mathbf{y} = \mathbf{X}\beta + e_1$ are estimated with $\hat{\beta} = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}y$ if the inverse exists or in general with any $\hat{\beta}$ which is such that $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ is a projection onto the subspace $sp(\mathbf{X})$.

By definition, a projection $\hat{\mathbf{y}}$ simply corresponds to a decomposition of the original vector into two orthogonal components, i.e. writing $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$. We have $\hat{\mathbf{e}} = y - \mathbf{X}\hat{\beta} \perp \hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ so that $y - \mathbf{X}\hat{\beta} \in \mathbf{V}^{\perp} = \{v : v \perp sp(\mathbf{X})\}$ and $dim(\mathbf{V}^{\perp}) = n - r$.

$$\begin{aligned}\underbrace{\hat{\mathbf{e}}}_{n\times 1} &= y - \mathbf{X}\hat{\beta} = y - \mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}y \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'})y = (\mathbf{I} - \mathbf{H})y\end{aligned}$$

and $rank(\mathbf{I} - \mathbf{H}) = dim(\mathbf{V}^{\perp}) = n - r$

4.2 The multivariate normal and related distributions

4.2.1 Handout

Suppose $Z_1, \ldots, Z_n$ are independent Gaussian with mean zero and variance one ($Z_1, \ldots, Z_n \sim n(0,1)$ iid) so their joint density is

$$f(\xi) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\xi_i^2/2\right) = \frac{1}{(2\pi)^{n/2}} \exp(-(1/2)\xi^T\xi)$$

Let A be an invertible $n \times n$ matrix and $\mu \in \mathbb{R}^n$ and define $\mathbf{Y} = A\mathbf{Z} + \mu$.
Recall from calculus that if g is a $1-1$ function $g : \mathbb{R}^n \to \mathbb{R}^n$

$$\int f(\xi)d\xi = \int f(g(\mathbf{y}))|J|d\mathbf{y}$$

where J is the Jacobian of the transformation

$$J = \left|\frac{d\xi}{d\mathbf{y}}\right| = \left|\frac{\partial g(\mathbf{y})}{\partial \mathbf{y}}\right|$$

and the integrals are over corresponding regions.

It follows that the joint pdf of $\mathbf{Y}$ is h with $h(y) = f(g(y))|J|$.
Some linear algebra gives

$$h(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(\mathbf{y}-\mu)^T \Sigma^{-1}(\mathbf{y}-\mu))$$

where $\Sigma = AA^T$.
This leads to a natural definition of the multivariate normal distribution.
The n-dimensional random vector, $\mathbf{Y}$ is said to have a multivariate normal distribution, denoted $\mathbf{Y} \sim n(\mu, \Sigma)$ if the density of $\mathbf{Y}$ is of the form

$$h(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(\mathbf{y}-\mu)^T \Sigma^{-1}(\mathbf{y}-\mu))$$

where $\mu \in \mathbb{R}^n$ and Σ is a symmetric positive definite $n \times n$ matrix.
It is left to the reader to prove that if $\mathbf{Y} \sim n(\mu, \Sigma)$ and B is an $p \times n$ matrix of full rank p ($p < n$), then $B\mathbf{Y}$ also has a multivariate normal distribution.
If $Z \sim n(0,1)$ is standard normal, then we define the chi-squared distribution on one degree of freedom, χ^2_1 to be the distribution of $U := Z^2$ and write $U \sim \chi^2_1$. If $U_1, \ldots, U_p$ are i.i.d. χ_1, then we define χ^2_p to be the distribution of their sum: $\sum_{i=i}^{p} U_i \sim \chi^2_p$.
Finally, if $U \sim \chi^2_{\nu_1}$ and $V \sim \chi^2_{\nu_2}$ are independent, then we define the **F distribution on** ν_1 **and** ν_2 **degrees of freedom** to be the distribution of the ratio $\frac{U/\nu_1/}{V/\nu_2}$ and write

$$\frac{U/\nu_1/}{V/\nu_2} \sim F_{\nu_1,\nu_2}.$$

4.3 A basis for the span of X

Orthonormal basis, $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for $\mathbf{R}^n$:

Using Gram-Schmidt, first generate $\mathbf{u}_1, \ldots, \mathbf{u}_r$ which span $sp\{\mathbf{X}\}$, with $rank\{\mathbf{X}\} = r$ and the rest, $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$ are chosen so that the entire set, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ spans $\mathbf{R}^n$.

$$\begin{aligned} \mathbf{X}\hat{\beta} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_r\mathbf{u}_r \\ \mathbf{y} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_r\mathbf{u}_r + \hat{\zeta}_{r+1}\mathbf{u}_{r+1} + \ldots + \hat{\zeta}_n\mathbf{u}_n \end{aligned}$$

4.3.1 Details

The probability distributions can best be viewed by defining a new orthonormal basis, $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for $\mathbf{R}^n$.

This basis is defined by first generating a set of r vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ which span the space defined by $sp\{\mathbf{X}\}$, and the rest, $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$ are chosen so that the entire set, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ spans $\mathbf{R}^n$. This is obviously always possible using the method of Gram-Schmidt. This gives the following sequence of spaces and spans:

$$\begin{aligned} sp\{\mathbf{X}\} &= sp\{\mathbf{u}_1, \ldots, \mathbf{u}_r\} \\ \mathbf{R}^n &= sp\{\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \ldots \mathbf{u}_n\} \end{aligned}$$

One can then write each of $\mathbf{X}\hat{\beta}$ and $\mathbf{y}$ in terms of the new basis as follows:

$$\begin{aligned} \mathbf{X}\hat{\beta} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_r\mathbf{u}_r \\ \mathbf{y} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_r\mathbf{u}_r + \hat{\zeta}_{r+1}\mathbf{u}_{r+1} + \ldots \hat{\zeta}_n\mathbf{u}_n \end{aligned}$$

where it is well-known that $\hat{\zeta}_i = \mathbf{u}_i \cdot \mathbf{y}$.

It is important to note that the same coefficients $\hat{\zeta}_i$ are obtained for $1 \le i \le r$. This follows from considering the coefficient of $\mathbf{y}$ in the basis and noting that $\mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{e}}$ where the residual vector $\hat{\mathbf{e}}$ is orthogonal to all column vectors of $\mathbf{X}$ and therefore also to $\mathbf{u}_i$ for $1 \le i \le r$. Therefore,

$$\hat{\zeta}_i = \mathbf{u}_i \cdot \mathbf{y} = \mathbf{u}_i \cdot \mathbf{X}\hat{\beta}$$

4.4 Q-R decomposition

$$\mathbf{Q} := [\mathbf{u}_1 \mathbf{u}_2 \ldots \mathbf{u}_n]$$

is the $\mathbf{Q}$ in the Q-R decomposition of $\mathbf{X}$.

If

$$\mathbf{z} = \left(\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_n\right)$$

then

$$\mathbf{z} = \mathbf{Q}'\mathbf{y}$$

and hence

$$E[\mathbf{z}] = \mathbf{Q}'\mathbf{X}\beta$$

$$V[\mathbf{z}] = \mathbf{Q}'\sigma^2\mathbf{I}\mathbf{Q} = \sigma^2\mathbf{I}$$

4.4.1 Details

$\mathbf{Q} := [\mathbf{u}_1 \mathbf{u}_2 \ldots \mathbf{u}_n]$ is the $\mathbf{Q}$ in the Q-R decomposition of $\mathbf{X}$.

$\mathbf{Q}$ has important properties, e.g. $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ so $\mathbf{Q}^{-1} = \mathbf{Q}'$.

If

$$\mathbf{z} = \left(\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_n\right)$$

then

$$\mathbf{z} = \mathbf{Q}'\mathbf{y} \text{ and } \mathbf{y} = \mathbf{Q}\mathbf{z}$$

and hence

$$E[\mathbf{z}] = \mathbf{Q}'\mathbf{X}\beta$$

$$V[\mathbf{z}] = \mathbf{Q}'\sigma^2\mathbf{I}\mathbf{Q} = \sigma^2\mathbf{I}$$

4.5 Variances of coefficients

For each i we obtain

$$V\left[\hat{\zeta}_i\right] = \sigma^2$$

4.5.1 Details

For each i we trivially obtain

$$V\left[\hat{\zeta}_i\right] = \sigma^2$$

4.6 Normality and independence of coeffients

Note that $\hat{\zeta}_i$ are linear combinations of the various y_j since $\hat{\zeta}_i = \mathbf{u}_i \cdot \mathbf{y}$.

When the y_i are independent Gaussian random variables, $\hat{\zeta}_i$ have zero covariance and are thus also independent.

4.6.1 Details

Note that $\hat{\zeta}_i$ are linear combinations of the various y_j since $\hat{\zeta}_i = \mathbf{u}_i \cdot \mathbf{y}$. The $\hat{\zeta}_i$ have zero covariance and when the y_i are independent Gaussian random variables, the $\hat{\zeta}_i$ are also independent.

This final result uses the fact that Gaussian random variables which have zero covariance are also independent. The fact that they have zero covariance is easy to establish, but the corrollary of independence is a result from multivariate normal theory.

The normal theory is fairly simple in this case:

$$\mathbf{z} = \left(\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_n\right) = \mathbf{Q}'\mathbf{y}$$

and

$$\mathbf{y} \sim n\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$$

.
It follows that $\mathbf{z}$ is multivariate normal and from the earlier derivations of the mean and variance we have

$$\mathbf{z} \sim n\left(\mathbf{Q}'\mathbf{X}\beta, \sigma^2\mathbf{I}\right).$$

4.7 Expected values of coefficients

For $i = r+1, \ldots, n$ we obtain

$$E\left[\hat{\zeta}_i\right] = 0$$

4.7.1 Details

The expected values of the coefficients, $\hat{\zeta}_i$ depend on which space these correspond to. Define

$$\zeta_i = E\left[\hat{\zeta}_i\right]$$

and by linearity we obtain

$$\zeta_i = E\left[\mathbf{u}_i \cdot \mathbf{y}\right] = \mathbf{u}_i \cdot (\mathbf{X}\beta).$$

Now note that we have defined the basis vectors in three sets. The first is such that they span the same space as the columns of $\mathbf{Z}$. The second set complements the first to span the $\mathbf{X}$ and the last set complements the set to span all of $\mathbf{R}^n$. The basis vectors are of course all orthogonal and each basis vector is orthogonal to all vectors in spaces spanned by preceding vectors.

For $i = r+1, \ldots, n$ we obtain

$$E\left[\hat{\zeta}_i\right] = \mathbf{u}_i \cdot (\mathbf{X}\beta) = 0$$

since $\mathbf{X}\beta$ is trivially in the space spanned by the column vectors of $\mathbf{X}$ and is therefore a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and $\mathbf{u}_i$ is orthogonal to all of these.

4.8 Sums of squares and norms

$$SSE(F) = ||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 \quad = \sum_{i=p+1}^{n} \hat{\zeta}_i^2$$

4.8.1 Details

It is now quite easy to see how to form sums of squared deviations based on the new orthonormal basis, since each set of deviations corresponds to a specific portion of the space.

$$SSE(F) = ||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 \quad = \sum_{i=r+1}^{n} \hat{\zeta}_i^2$$

4.9 Degrees of freedom

$SSE(F)$has $n-r$ degrees of freedom.

4.9.1 Details

$SSE(F)$ has $n-r$ degrees of freedom.
References Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. 1996. Applied linear statistical models. McGraw-Hill, Boston. 1408pp.

5 Linear hypotheses in multiple regression

5.1 Null hypotheses and geometry

5.1.1 Details

Tests of hypotheses in linear models can be considered geometrically. The hypothesis $H_i : \beta = 0$ in simple linear regression is the question of whether the matrix

$$\mathbf{Z} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

can be used in place of $\mathbf{X}$, i.e. whether the projection of y onto $span(\mathbf{Z})$ is too much farther away from y than the projection onto $span(X)$.

5.2 Null hypotheses and matrices

5.2.1 Details

Null hypotheses are almost always concerned with how one can "reduce" or simplify the model, in this case usually whether one can reduce the number of columns in $\mathbf{X}$ or by some other means reduce the number of coefficients in the model.

5.3 Null hypothesis as matrices

Have $\underbrace{\mathbf{X}}_{n\times p}$ and $\underbrace{\mathbf{Z}}_{n\times q}$ s.t. $span(\mathbf{Z}) \subseteq span(\mathbf{X})$.
Can estimate models

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}_1$$

$$\mathbf{y} = \mathbf{Z}\gamma + \mathbf{e}_2$$

Will derive test for

$$H_0 : \mathbf{X}\beta = \mathbf{Z}\gamma$$

5.3.1 Details

Assume that $\underbrace{\mathbf{X}}_{n\times p}$ and $\underbrace{\mathbf{Z}}_{n\times q}$ are matrices such that $span(\mathbf{Z}) \subseteq span(\mathbf{X})$.
We can estimate coefficients in the model

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

and in the reduced model

$$\mathbf{y} = \mathbf{Z}\gamma + \mathbf{e}$$

We will derive tests for the general null hypothesis

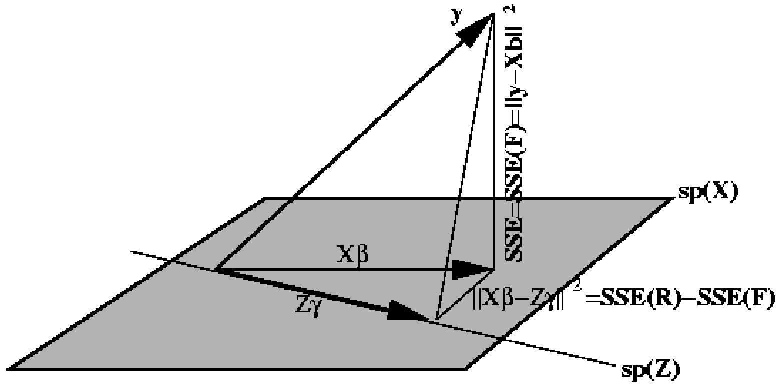
$$H_0 : \mathbf{X}\beta = \mathbf{Z}\gamma$$

which is typically some hypothesis stating that some of the coefficients in the β-vector are zero or otherwise restricted.

5.3.2 Examples

Example 5.1. In simple linear regression, $y_i = \alpha + \beta x_i + e_i$, the most common test is for $\beta = 0$.

5.4 Geometric comparisons of models



Testing linear hypotheses in linear models corresponds to projecting onto subspaces.

5.4.1 Details

Relationships between sums of squares in two linear models is best viewed geometrically.

Starting with a base model as before, $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, there is a need to investigate whether this model can be simplified in some manner. A simpler model can be denoted by $\mathbf{y} = \mathbf{Z}\gamma + \mathbf{e}$ where $\mathbf{Z}$ is a matrix, typically with fewer columns than $\mathbf{X}$, and the column vectors of $\mathbf{Z}$ span a subspace of that spanned by $\mathbf{X}$.

5.4.2 Examples

Example 5.2. A typical hypothesis test would start with a basic (full) model of the form $y_i = \alpha + \beta x_i + e_i$, wanting to test the null hypothesis $H_0 : \beta = 0$.

Define the matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{pmatrix}, \tag{1}$$

so the model in matrix notation becomes $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$.

The null hypothesis can be written as $\mathbf{y} = \mathbf{Z}\gamma + \mathbf{e}$, where

$$\mathbf{Z} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}. \tag{2}$$

5.5 Bases for the span of X

Orthonormal basis, $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for $\mathbf{R}^n$:

Using Gram-Schmidt, first generate $\mathbf{u}_1, \ldots, \mathbf{u}_r$ which span $sp\{\mathbf{Z}\}$, the next vectors, $\mathbf{u}_{q+1}, \ldots, \mathbf{u}_r$ are chosen so that $\mathbf{u}_1, \ldots, \mathbf{u}_r$ span $sp\{\mathbf{X}\}$, with $rank\{\mathbf{X}\} = r$, and the rest, $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$ are chosen so that the entire set, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ spans $\mathbf{R}^n$.

$$\begin{aligned} \mathbf{Z}\hat{\gamma} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_q\mathbf{u}_q \\ \mathbf{X}\hat{\beta} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_q\mathbf{u}_q + \hat{\zeta}_{q+1}\mathbf{u}_{q+1} + \ldots \hat{\zeta}_r\mathbf{u}_r \\ \mathbf{y} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_q\mathbf{u}_q + \hat{\zeta}_{q+1}\mathbf{u}_{q+1} + \ldots \hat{\zeta}_r\mathbf{u}_r \\ &\quad + \hat{\zeta}_{r+1}\mathbf{u}_{r+1} + \ldots \hat{\zeta}_n\mathbf{u}_n \end{aligned}$$

5.5.1 Details

The probability distributions can best be viewed by defining a new orthonormal basis, $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for $\mathbf{R}^n$. This basis is defined by first generating a set of r vectors $\mathbf{u}_1, \ldots, \mathbf{u}_q$ which span the space defined by the null hypothesis, $sp\{\mathbf{Z}\}$, where $rank\{\mathbf{Z}\} = q$, subsequently the next vectors, $\mathbf{u}_{q+1}, \ldots, \mathbf{u}_r$ are chosen so as to span the remainder of $sp\{\mathbf{X}\}$, where $rank\{\mathbf{X}\} = r$, and therefore $sp\{\mathbf{X}\} = sp\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$, and the rest, $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$ are chosen so that the entire set, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ spans $\mathbf{R}^n$. This is obviously always possible using the method of Gram-Schmidt.

This gives the following sequence of spaces and spans:

$$\begin{aligned} sp\{\mathbf{Z}\} &= sp\{\mathbf{u}_1, \ldots, \mathbf{u}_q\} \\ sp\{\mathbf{X}\} &= sp\{\mathbf{u}_1, \ldots, \mathbf{u}_q, \mathbf{u}_{q+1}, \ldots, \mathbf{u}_r\} \\ \mathbf{R}^n &= sp\{\mathbf{u}_1, \ldots, \mathbf{u}_q, \mathbf{u}_{q+1}, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \ldots \mathbf{u}_n\} \end{aligned}$$

One can then write each of $\mathbf{Z}\hat{\gamma}$, $\mathbf{X}\hat{\beta}$, $\mathbf{y}$ in terms of the new basis as follows:

$$\begin{aligned} \mathbf{Z}\hat{\gamma} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_q\mathbf{u}_q \\ \mathbf{X}\hat{\beta} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_q\mathbf{u}_q + \hat{\zeta}_{q+1}\mathbf{u}_{q+1} + \ldots \hat{\zeta}_r\mathbf{u}_r \\ \mathbf{y} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_q\mathbf{u}_q + \hat{\zeta}_{q+1}\mathbf{u}_{q+1} + \ldots \hat{\zeta}_r\mathbf{u}_r + \hat{\zeta}_{r+1}\mathbf{u}_{r+1} + \ldots \hat{\zeta}_n\mathbf{u}_n \end{aligned}$$

where it is left to the reader to see that the $\hat{\zeta}_i$-coefficients are indeed the same.

5.6 Expected values of coefficients

For $i = r+1, \ldots, n$ we obtain

$$E\left[\hat{\zeta}_i\right] = 0$$

If $H_0 : \mathbf{X}\beta = \mathbf{Z}\gamma$ is true then for $i = q+1, \ldots, r$ we obtain

$$E\left[\hat{\zeta}_i\right] = \mathbf{u}_i \cdot (\mathbf{Z}\gamma) = 0$$

5.6.1 Details

The expected values of the coefficients, $\hat{\zeta}_i$ depend on which space they correspond to.

Define

$$\zeta_i = E\left[\hat{\zeta}_i\right]$$

and by linearity we obtain

$$\zeta_i = E\left[\mathbf{u}_i \cdot \mathbf{y}\right] = \mathbf{u}_i \cdot (\mathbf{X}\beta)\,.$$

Now note that we have defined the basis vectors in three sets. The first is such that they span the same space as the columns of $\mathbf{Z}$. The second set complements the first to span the $\mathbf{X}$ and the last set complements the set to span all of $\mathbf{R}^n$. The basis vectors are of course all orthogonal and each basis vector is orthogonal to all vectors in spaces spanned by preceding vectors.

For $i = r+1, \ldots, n$ we obtain

$$E\left[\hat{\zeta}_i\right] = \mathbf{u}_i \cdot (\mathbf{X}\beta) = 0$$

since $\mathbf{X}\beta$ is trivially in the space spanned by the column vectors of $\mathbf{X}$ and is therefore a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and $\mathbf{u}_i$ is orthogonal to all of these.

If the null hypothesis that $E[\mathbf{Y}]$ **can be written as** $H_0 : \mathbf{X}\beta = \mathbf{Z}\gamma$ **is true** then for $i = q+1, \ldots, r$ we obtain

$$E\left[\hat{\zeta}_i\right] = \mathbf{u}_i \cdot (\mathbf{Z}\gamma) = 0$$

but this only holds under the null hypothesis.

5.7 Sums of squares and norms

$$\begin{aligned}
SSE(F) = ||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 &= \sum_{i=r+1}^{n} \hat{\zeta}_i^2 \\
SSE(F) - SSE(R) = ||\mathbf{Z}\hat{\gamma} - \mathbf{X}\hat{\beta}||^2 &= \sum_{i=q+1}^{r} \hat{\zeta}_i^2 \\
SSE(R) = ||\mathbf{y} - \mathbf{Z}\hat{\gamma}||^2 &= \sum_{i=q+1}^{n} \hat{\zeta}_i^2
\end{aligned}$$

5.7.1 Details

It is now quite easy to see how to form sums of squared deviations based on the new orthonormal basis, since each set of deviations corresponds to a specific portion of the space.

$$\begin{aligned} SSE(F) = ||\mathbf{y}-\mathbf{X}\hat{\beta}||^2 &= \sum_{i=r+1}^{n} \hat{\zeta}_i^2 \\ SSE(F) - SSE(R) = ||\mathbf{Z}\hat{\gamma}-\mathbf{X}\hat{\beta}||^2 &= \sum_{i=q+1}^{r} \hat{\zeta}_i^2 \\ SSE(R) = ||\mathbf{y}-\mathbf{Z}\hat{\gamma}||^2 &= \sum_{i=q+1}^{n} \hat{\zeta}_i^2 \end{aligned}$$

Since each $\hat{\zeta}_i$ is a coordinate in an orthonormal basis, this is formed as an inner product with the corresponding basis vector, i.e. $\hat{\zeta}_i = \mathbf{y} \cdot \mathbf{u}_i$.

5.8 Some probability distributions

5.8.1 Details

Suppose we have two matrices, $\mathbf{X}$ and $\mathbf{Z}$ which satisfy $rank(\mathbf{Z}) = q < p = rank(\mathbf{X})$ and $sp(\mathbf{Z}) \subseteq sp(\mathbf{X})$ (usually $\mathbf{Z}$ is $n \times q$ and $\mathbf{X}$ is $n \times p$).
Then $H_0 : E[\mathbf{Y}] = \mathbf{Z}\gamma$ is a reduction from the model $E[\mathbf{Y}] = \mathbf{X}\beta$.
Write $\mathbf{F}$ = full model and $\mathbf{R}$ = for the reduced model.
Then we have
1) $y - \mathbf{X}\hat{\beta} \perp \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\gamma}$
2) $||y-\mathbf{X}\hat{\beta}||^2$ and $||\mathbf{X}\hat{\beta}-\mathbf{Z}\hat{\gamma}||^2$ are independent
3)$\frac{||y-\mathbf{X}\hat{\beta}||^2}{\sigma^2} \sim \chi^2_{n-p}$ if the model is correct
4) $\frac{||\mathbf{X}\hat{\beta}-\mathbf{Z}\hat{\gamma}||^2}{\sigma^2} \sim \chi^2_{p-q}$ if H_0 is correct.
5) $SSE(F) = ||y-\mathbf{X}\hat{\beta}||^2$
and
$SSE(R) - SSE(F) = ||\mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\gamma}||^2$
are independent.
6) $\frac{(SSE(R)-SSE(F))/(p-q)}{SSE(F)/(n-p)} \sim F_{p-q,n-p}$
Here F_{ν_1,ν_2} is the distribution of a ratio

$$F = \frac{U/\nu_1}{V/\nu_2}$$

of independent χ^2-random variables, $U \sim \chi^2_{\nu_1}$, and $V \sim \chi^2_{\nu_2}$.

5.9 General F-tests in linear models

5.9.1 Details

In general one can compute the sum of squares from the full model, $SSE(F)$ as above and then compute the sum of squared deviations from the reduced model, $SSE(R) = ||\mathbf{y} - \mathbf{Z}\hat{\gamma}||^2$. Denote the corresponding degrees of freedom by $df(F)$ and $df(R)$, and assume that both matrices $\mathbf{Z}$ and $\mathbf{X}$ have full ranks, i.e. $rank(\mathbf{X}) = p$ and $rank(\mathbf{Z}) = r$. Then $df(F) = n-p$ and $df(R) = n-r$.

The null hypothesis can then be tested by noting that

$$F = \frac{(SSE(R)-SSE(F))/(p-r)}{SSE(F)/(n-p)} \tag{3}$$

is a realisation of a random variable from an F-distribution with $p-r$ and $n-p$ degrees of freedom under H_0.

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

6 Building a multiple regression model

6.1 Introduction

Have several independent variables
Want to select some into regression
Want to evaluate quality of resulting model
Want to improve into a final model

6.1.1 Details

Building multiple regression models includes several steps. It is, firstly, rarely pre-defined what independent variables should be included in the model, so a method for selecting these is needed. Having obtained an initial model one needs to evaluate not only the assumptions of the model but also identify possible influential observations and possibly undertake other diagnostics. Having obtained regression diagnostic, the model needs to be improved by taking these into account.

6.2 Variable selection: Measuring quality

- R^2
- AIC
- BIC
- SSE
- MSE
- *P*-values

6.2.1 Details

- R^2
- AIC
- BIC
- SSE
- MSE
- *P*-values

6.2.2 Examples

Example 6.1. Use the ecosystem data set and select a single variable in a simple linear regression to predict the growth of cod. Compare the various criteria.

6.3 Variable selection: Forward or backward

Model selection:

- All subset regression
- Forward stepwise regression
- Backwards stepwise regression

6.3.1 Details

Several methods exist to select a regression model.

All subset regression simply considers every possible combination of independent variables. Although this will indicate all possible "good" models and will certainly find the "best" model (using any given criterion), this is often not feasible.

Backwards stepwise regression starts by taking all independent variables into a single model and then dropping variables one at a time. The variable to be dropped is the one giving the least increase in SSE. This approach is often preferred, but is not feasible if the total number of variables are very large.

Forward stepwise regression selects a sequence of variables, at each stage deciding what variable to add next. The addition is based on including the variable giving the largest amount of (marginal) explained variation.

Forward stepwise regression is often augmented by allowing a variable to be dropped after a variable has been added. Thus a sequence of insertions may make an earlier variable redundant and thus dropped. Either version of forward regression is quite feasible but may lead to an incorrect or bad model since important combinations of variables may not be found.

Each approach thus has good and bad points.

6.3.2 Examples

Example 6.2. Use the ecosystem data set and conduct a forward stepwise regression to predict the growth of cod. Compare the various criteria for model selection.
R commands: add1 repeatedly - followed by anova(fm.final,fm.full)

Example 6.3. Use the ecosystem data set and conduct a backwards stepwise regression to predict the growth of cod. Compare the various criteria for model selection.
R commands: drop1 or summary - followed by anova(fm.final,fm.full)

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

Belsey, D. A., Kuh, E. and Welsh, R. E. 1980. Regression diagnostics: Identifying influential data and sources of collinearity. John. Wiley and Sons, New York. 292pp.

7 Prediction in the linear model

7.1 Prediction and prediction uncertainty

A new observation: $\mathbf{x}_h$
The prediction: $\hat{y}_h = E\left[\hat{y}_h\right] = \mathbf{x}_h\hat{\beta}$
The variance: $\sigma^2\mathbf{x}_h'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_h$
Prediction uncertainty: $V\left[y_h - \hat{y}_h\right]$

7.1.1 Details

A new observation: $\mathbf{x}_h$
The prediction: $\hat{y}_h = E\left[\hat{y}_h\right] = \mathbf{x}_h\hat{\beta}$
The variance: $\sigma^2\mathbf{x}_h'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_h$
Prediction uncertainty: $V\left[y_h - \hat{y}_h\right]$

7.1.2 Examples

Example 7.1. Age and live weight of lambs. Project: Predict the weight (with uncertainty) at a given day.

```
days weight
135 39
125 35
120 33
126 38
125 37
137 38
133 36
140 41
130 38
129 36
123 34
132 40
129 38
121 34
126 35
137 44
121 34
137 41
130 39
137 43
```

8 Estimable functions

8.1 Estimable functions: The problem

If **X** is not of full rank, then the LS problem does not have a unique solution for $\hat{\beta}$.

In general not all combinations of the form $\mathbf{c}'\hat{\beta}$ may have unique solutions.

A linear combination $\mathbf{c}'\beta$ is an **estimable function** if there is a vector of numbers, **a**, such that

$$E[\mathbf{ay}] = \mathbf{c}'\beta$$

for all β.

NB: Viewed as a function of the unknown parameter vector, β.
NB: The E-operator depends on β, could write $g(\beta) = \mathbf{c}'\beta$ and require $g(\beta) = E_\beta[\mathbf{ay}] \quad \forall\beta$ for some **a**.

8.1.1 Details

If **X** is not of full rank, then the LS problem does not have a unique solution for $\hat{\beta}$. In general not all combinations of the form $\mathbf{c}'\hat{\beta}$ may have unique solutions.

A linear combination $\mathbf{c}'\beta$ is an **estimable function** if there is a vector of numbers, **a**, such that

$$E\left[\mathbf{a}'\mathbf{y}\right] = \mathbf{c}'\beta$$

for all β.

The terminology is not accidental as the linear combination of parameters is viewed as a function of the unknown parameter vector, β. In words the requirement is simply that it is possible to obtain a linear unbiased estimator.

8.1.2 Examples

Example 8.1. The one-way layout is the simplest example giving **X**-matrices which are not of full rank when writing the model in the form

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}.$$

Example 8.2. A common issue in regression is whether the same line can be fit to two data sets or e.g. whether different slopes should be used. This can be modelled by writing

$$y_{ij} = \alpha + \beta x_{ij} + \varepsilon_{ij} \tag{4}$$

for the simple model with the same slopes and

$$y_{ij} = \alpha + \beta_i x_{ij} + \varepsilon_{ij} \tag{5}$$

for a model with different slopes in the the groups.

Alternatively one may be interested in how the slopes in the groups differ and/or in a simple evaluation of whether a single slope can be used. In this case it is reasonable to rewrite the complex model as

$$y_{ij} = \alpha + \beta x_{ij} + \beta_i x_{ij} + \varepsilon_{ij} \tag{6}$$

and the test of whether the reduced model is enough is a test of whether the β_i-values are all zero (and can be dropped).

Naturally, equation 6 is not completely determined. On the other hand, the model can easily be fit to data - most statistical packages will simply select an arbitrary LS estimate of the parameter set unless told explicitly to select a specific representation. All such solutions will lead to the same tests. The tests are really just based on comparing whether SSE(R) is too much smaller than SSE(F) and these sums are based on the LS projections onto subspaces. The projections are uniquely defined since they are based on the span, V, of the column vectors in the X-matrix. This space V does not change when columns are added, as long as these columns are linear combinations of existing ones - or when such columns are dropped.

Packages such as R will easily compare 6 and 4 with the drop1-command since 4 corresponds to deleting a term from 6. The better-determined model 5 can be compared to 4 using an anova-command in R since 4 is indeed a reduced model from 5 through a restriction of the form $\beta_1 = \ldots = \beta_I$.

8.1.3 Handout

Some further clarifications to the above definitions may be useful.

It must be emphasized that the E-operator depends on the vectors of unknown parameters, β, since the underlying model is $E[\mathbf{y}] = \mathbf{X}\beta$. One could therefore write $f(\beta) := E_\beta[\mathbf{y}]$ and $g(\beta) := c'\beta$ so the criterion of estimability would be that $f(\beta) = g(\beta) \quad \forall \beta \in \mathbf{R}^n$. This formal approach has the merit that the meaning is clear, but the notation becomes quite cumbersome.

Estimable functions are commonly denoted by the symbol ψ, e.g. $\psi = \beta_1 - \beta_2$ etc.

Note 8.1. Recall that if $\underbrace{\mathbf{X}}_{n\times p}$ with $n > p$ is of full rank if $\text{rank}(\mathbf{X}) = dim(sp(\mathbf{X})) = p$ and also $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}'\mathbf{X})$ so $\mathbf{X}$ is of full rank iff $\mathbf{X}'\mathbf{X}$ has an inverse. Hence, if $\mathbf{X}$ is of full rank we can write $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ which satisfies $E\left[\hat{\beta}\right] = \beta$.

From this we also see that

$$\begin{aligned}
\mathbf{c}'\beta &= \mathbf{c}'E\left[\hat{\beta}\right] \\
&= E\left[\mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\right] \\
&= E\left[\mathbf{a}'\mathbf{y}\right]
\end{aligned}$$

where $\mathbf{a}' = \mathbf{c}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'$. Hence any linear combination $\mathbf{c}'\beta$ is estimable is $\mathbf{X}$ is of full rank.

Conversely, if $\mathbf{X}$ is not of full rank then we can find vectors β and γ with $\beta \neq \gamma$ such that $\mathbf{X}(\beta - \gamma) = 0$ and therefore $E\left[\mathbf{y}\right] i = \mathbf{X}\beta = \mathbf{X}\gamma$ can be expressed in more than one way.

Existence of non-estimable functions are therefore an expression of the matrix not being of full rank.

8.2 Classification of estimable functions

Theorem: A parametric function $\psi = \mathbf{c}'\beta$ is estimable if and only if $\mathbf{c}' = \mathbf{a}'\mathbf{X}$ for some $\mathbf{a} \in \mathbf{R}^n$.

8.2.1 Details

Theorem 8.1. A parametric function $\psi = \mathbf{c}'\beta$ is estimable if and only if $\mathbf{c}' = \mathbf{a}'\mathbf{X}$ for some $\mathbf{a} \in \mathbf{R}^n$.

8.2.2 Examples

Example 8.3. In the linear model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, the coefficients are not all estimable.

8.2.3 Handout

Proof of theorem: By definition, $\psi = \mathbf{c}'\beta$ is estimable if and only if there is a vector $\mathbf{a} \in \mathbf{R}^n$ such that $E\left[\mathbf{a}'\mathbf{y}\right] = \mathbf{c}'\beta$ for all β. This is equivalent to requiring

$$\mathbf{a}'\mathbf{X}\beta = \mathbf{c}'\beta$$

for all β which is equivalent to

$$\mathbf{c}' = \mathbf{a}'\mathbf{X}.$$

Example 8.4. The reader should take the simple example $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad 1 \leq j \leq J_i \quad 1 \leq i \leq I$, set up the $\mathbf{X}$-matrix and consider the form of the vectors $\mathbf{a}'\mathbf{X}$ for the case $I = 2,\ J_1 = n,\ J_2 = m$.

Writing $\mathbf{a}'\mathbf{X} = (u, v, u - v)$, it is seen that the resulting estimable functions are precisely $\mu + \alpha_1$, $\mu + \alpha_2$ and $\alpha_1 - \alpha_2$.

8.3 Gauss-Markov theorem

Theorem: (Gauss-Markov theorem): Let $EY = X\beta$, $VY = \sigma^2 I$. Then every estimable function $c'\beta$ has a unique unbiased linear estimate which has minimum variance in the class of all unbiased linear estimates. This estimate can be written the form $c'\hat{\beta}$ where $\hat{\beta}$ is any LS estimator.

8.3.1 Details

A fundamental result in the theory of linear models is that estimable functions have unique unbiased linear estimates.

Lemma: If $\psi = \mathbf{c}'\beta$ is estimable and $V = sp(\mathbf{X})$ then there is a unique linear unbiased estimator of $\mathbf{c}'\beta$ of the form $\mathbf{a}'\mathbf{y}$ with $\mathbf{a} \in V$. If $\mathbf{a_0 y}$ is unbiased for $\mathbf{c}'\beta$ then $\mathbf{a}$ is the projection of $\mathbf{a_0}$ onto V.

Theorem 8.2 (Gauss-Markov theorem). Let $E[\mathbf{y}] = \mathbf{X}\beta$, $V[\mathbf{y}] = \sigma^2 I$. Then every estimable function $\mathbf{c}'\beta$ has a unique unbiased linear estimate which has minimum variance in the class of all unbiased linear estimates. This estimate can be written the form $\mathbf{c}'\hat{\beta}$ where $\hat{\beta}$ is any LS estimator.

Note 8.2. For estimable functions this is defined as the LS estimator.

8.3.2 Examples

Example 8.5. In the model $y_{ik} = \mu + \alpha_i + e_{ik}$, it is clear that parameters are not estimable but it is easy to see that $\alpha_i - \alpha_j$ are estimable.

8.3.3 Handout

Proof of lemma: Suppose $\psi = \mathbf{c}'\beta$ is estimable so we can find $\mathbf{a} \in \mathbf{R}^n$ such that $E[a'y] = \psi$.Now write $\mathbf{a} = \mathbf{a}^* + \mathbf{b}^*$ with $\mathbf{a}^* \in V$ and $\mathbf{b}^* \perp V$, i.e. we define $\mathbf{a}^*$ as the projection of $\mathbf{a}$ onto V. Then it is easy to see that $E\mathbf{b}^{*\prime}\mathbf{y} = 0$ since $\mathbf{b}^*$ is perpendicular to the columns of $\mathbf{X}$, all of which are in V. Hence $\psi = E\mathbf{a}'\mathbf{y} = E\mathbf{a}^{*\prime}\mathbf{y}$ and hence $\mathbf{a}^{*\prime}\mathbf{y}$ is unbiased for ψ and $\mathbf{a}^* \in V$.

For uniqueness of $\mathbf{a}^*$, suppose $E\alpha'\mathbf{y} = \psi$ for some $\alpha \in V$. Then $0 = E\mathbf{a}^{*\prime}\mathbf{y} - E\alpha'\mathbf{y} = (\mathbf{a}^* - \alpha)'\mathbf{X}\beta$. This holds for all $\beta \in R^p$ and hence $(\mathbf{a}^* - \alpha)'\mathbf{X} = \mathbf{0}$. Since $(\mathbf{a}^* - \alpha)$ is perpendicular to all columns of the $\mathbf{X}$-matrix, it follows that $(\mathbf{a}^* - \alpha) \in V^\perp$. But both vectors were in V to begin with, so

$$(\mathbf{a}^* - \alpha) \in V \cap V^\perp = \{0\}$$

i.e. $\mathbf{a}^* = \alpha$ so $\mathbf{a}^*$ is unique. Since $\mathbf{a}^*$ was taken as the projection of a onto V, the proof is complete.

Proof of Gauss-Markov theorem: Use the lemma to find a unique $\mathbf{a}^* \in V$ with $E\mathbf{a}^{*\prime}\mathbf{y} = \mathbf{c}'\beta$ and let $\mathbf{a}'\mathbf{y}$ be any unbiased linear estimate of ψ. Then $\mathbf{a}^* = proj_V(\mathbf{a})$ and $||\mathbf{a}||^2 = ||\mathbf{a}^*||^2 + ||\mathbf{a} - \mathbf{a}^*||^2$ so

$$V\mathbf{a}'\mathbf{y} = \mathbf{a}'\Sigma_{\mathbf{y}}\mathbf{a} = \sigma^2||\mathbf{a}||^2$$
$$= \sigma^2||\mathbf{a}^*||^2 + \sigma^2||\mathbf{a} - \mathbf{a}^*||^2 = V\mathbf{a}^*\mathbf{y} + \sigma^2||\mathbf{a} - \mathbf{a}^*||^2 \geq V\mathbf{a}^*\mathbf{y}$$

and equality holds iff $\mathbf{a} = \mathbf{a}^*$ so $\mathbf{a}^*\mathbf{y}$ is best.

Now let $\hat{\beta}$ be any least squares estimate.

Note 8.3. Note that $\mathbf{a}^* \in V$ and $\mathbf{y}-\mathbf{X}\hat{\beta} \in V^\perp$ so that $\mathbf{a}^*(\mathbf{y}-\mathbf{X}\hat{\beta}) = 0$ and therefore $\mathbf{a}^*\mathbf{y} = \mathbf{a}^*\mathbf{X}\hat{\beta}$.

Further, since $\psi = \mathbf{c}'\beta$ is estimable and $\mathbf{a}^{*\prime}\mathbf{y}$ is the unbiased linear estimate, $\mathbf{c}'\beta = E\mathbf{a}^*\mathbf{y} = \mathbf{a}^*\mathbf{X}\beta$ and this holds for any $\beta \in \mathbf{R}^p$ so $\mathbf{c}' = \mathbf{a}^{*\prime}\mathbf{X}$. Combining this with the previous paragraph, $\mathbf{a}^*\mathbf{y} = \mathbf{a}^*\mathbf{X}\hat{\beta} = \mathbf{c}'\hat{\beta}$ which concludes the proof.

8.4 Testing hypotheses in the linear model

$\underbrace{\mathbf{y}}_{n\times 1} \sim n(\underbrace{\mathbf{X}}_{n\times p}\underbrace{\beta}_{p\times 1}, \sigma^2\underbrace{\mathbf{I}}_{n\times n})$

Theorem: $\hat{\psi} \sim n\left(\psi, \Sigma_{\hat{\psi}}\right)$, $\frac{||\mathbf{y}-\mathbf{X}\hat{\beta}||^2}{\sigma^2} \sim \chi^2_{n-r}$ and these two quantities are independent.

8.4.1 Details

Let $\underbrace{\mathbf{y}}_{n\times 1} \sim n(\underbrace{\mathbf{X}}_{n\times p}\underbrace{\beta}_{p\times 1}, \sigma^2\underbrace{\mathbf{I}}_{n\times n})$
and assume $rank(\mathbf{X}) = r \leq p$.

The interest will be in obtaining some joint confidence statement on a vector, $\psi = (\psi_1,\ldots,\psi_q)$, where each $\psi_i = \mathbf{c}_i'\beta$ is an estimable function. Write $\hat{\psi} = (\hat{\psi}_1,\ldots,\hat{\psi}_q)$ for the least squares estimates with $\hat{\psi}_i = \mathbf{c}_i\hat{\beta}$ where $\hat{\beta}$ is any LS estimate and one can therefore also write $\hat{\psi}_i = \mathbf{a}_i\mathbf{y}$ for unique $a_i \in sp(\mathbf{X})$.

The above can be written more concisely as $\psi = \mathbf{C}\beta$ using obvious definitions. It follows that

$$\hat{\psi} = \mathbf{A}\mathbf{y} = \mathbf{C}\hat{\beta} \sim n(\mathbf{C}\beta, \sigma^2\mathbf{A}\mathbf{A}')$$

and the variance-covariance matrix of the estimates will be denoted

$$V[\hat{\psi}] = \Sigma_{\hat{\psi}}$$

which leads to the following theorem.

Theorem 8.3. $\hat{\psi} \sim n\left(\psi, \Sigma_{\hat{\psi}}\right)$, $\frac{||\mathbf{y}-\mathbf{X}\hat{\beta}||^2}{\sigma^2} \sim \chi^2_{n-r}$ and these two quantities are independent.

It follows that hypothesis tests can be constructed in an obvious manner for individual estimable functions.

8.4.2 Handout

Proof: Let $\{\xi_1,\ldots,\xi_n\}$ be an orthonormal basis for $\mathbb{R}^n$ such that $\{\xi_1,\ldots,\xi_r\}$ form a basis for $sp(\mathbf{X})$ and let $\hat{\zeta}_1,\ldots,\hat{\zeta}_n$ be the coordinates of $\mathbf{y}$ in this basis, so that $\hat{\zeta}_i = \xi_i \cdot \mathbf{y}$. Also define $\zeta_i = E[\hat{\zeta}_i]$. It is established that $\hat{\zeta}_i$ are independent, Gaussian with common variance σ^2.

Write $\mathbf{z} = \left(\hat{\zeta}_1, \ldots, \hat{\zeta}_n\right)'$, $\mathbf{P} = [\xi_1, \ldots, \xi_n]'$ and note that $\mathbf{P}' = [\xi_1 \ldots \xi_n]$. It is then clear that the rows of $\mathbf{P}'$ are independent so $\mathbf{P}'$ is invertible (as is $\mathbf{P}$). Clearly, $\mathbf{PP}' = \mathbf{I}$ so $\mathbf{P}'\mathbf{P} = \mathbf{I}$. Further, $\mathbf{z} = \mathbf{Py}$ and therefore $\mathbf{y} = \mathbf{P}'\mathbf{z}$.

As elsewhere, write the LS estimates of the estimable functions in the form $\hat{\psi}_i = \mathbf{a}_i'\mathbf{y}$ where $\mathbf{a}_i \in V = sp\{\xi_1, \ldots, \xi_r\}$ so that $\hat{\psi}_i = \mathbf{a}_i'\mathbf{P}'\mathbf{z}$. It follows that $\mathbf{a}_i'\mathbf{P}' = \left[\mathbf{a}_i'\xi_1 \vdots \ldots \vdots \mathbf{a}_i'\xi_n\right]$ and of these various inner products, $\mathbf{a}_i'\xi_j = 0$ if $j > r$ (since $\mathbf{a}_i \in V$) from which it is seen that

$$\mathbf{a}_i'\mathbf{P}'\mathbf{z} = \left[\mathbf{a}_i'\xi_1 \vdots \ldots \vdots \mathbf{a}_i'\xi_r \vdots 0 \ldots 0\right] \left[\hat{\zeta}_1, \ldots, \hat{\zeta}_r, \hat{\zeta}_{r+1}, \ldots, \hat{\zeta}_n\right]' = \mathbf{a}_i'\xi_1\hat{\zeta}_1 + \ldots + \mathbf{a}_i'\xi_r\zeta_r$$

i.e. the estimable functions are all formed from the first r of the $\hat{\zeta}_i$ and are all of the form

$$\hat{\psi}_i = \sum_1^r k_j\hat{\zeta}_j \tag{7}$$

for some constants $k_1, \ldots, k_r$. This important result is quite general and basically states that anything that can be estimated can be derived from $\mathbf{y}$ through the column vectors of the $\mathbf{X}$-matrix.

On the other hand it is also known that $\mathbf{X}\hat{\beta}$ is the projection of $\mathbf{y}$ onto the space spanned by $\xi_1, \ldots, \xi_r$ and therefore the residual, $\mathbf{y} - \mathbf{X}\hat{\beta}$ is in the span of $\xi_{r+1}, \ldots, \xi_n$ and in fact

$$||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 = \sum_{j=r+1}^{n} \hat{\zeta}_j^2 \tag{8}$$

All the results in the theorem follow easily from (7) and (8).

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

Scheffe, H. 1959. The analysis of variance. John Wiley and Sons, Inc, New York. 477pp.

9 Ranks, constraints and correlations in multivariate regression

9.1 Problem statement

9.1.1 Details

When $r = \underbrace{rank(\mathbf{X})}_{n\times p} < k$, the estimate $\hat{\beta}$ is not unique. Similarly, β in $E[\mathbf{Y}] = \mathbf{X}\beta$ is not unique. [1] But if the function $\psi = \mathbf{c}'\beta$ is estimable, then the number $\mathbf{c}'\beta$ is unique, i.e. the same for all β in the set $\{b : E\mathbf{Y} = \mathbf{Xb}\}$, since $\mathbf{X}\beta$ is unique and $\mathbf{c}'\beta = \mathbf{a}'\mathbf{X}\beta$ for some $\mathbf{a}$.

9.2 Constraints

To specify β uniquely we can add constraints...

9.2.1 Details

In order to specify the vector β and $\hat{\beta}$ one could simply drop some of these until the $\mathbf{X}$-matrix becomes of full rank. More generally it is possible to add constraints of the form $\underbrace{\mathbf{H}}_{t\times p}\beta = \mathbf{0}$

This can be formulated in the following manner: Suppose we have β and we want unique $\tilde{\beta}$ through $\mathbf{X}\beta = \mathbf{X}\tilde{\beta}$ and $\mathbf{H}\tilde{\beta} = \mathbf{0}$.

Theorem 9.1. $\tilde{\beta}$ is unique if $rank\left(\binom{\mathbf{X}}{\mathbf{H}}\right) = p$ and $\tilde{\beta}$ are then estimable. [2]

The reader is referred to Scheffe (1959) for the proof of the theorem.

9.2.2 Examples

Example 9.1. If $Y_{ik} \sim n(\mu + \alpha_i, \sigma 2)$, independent, with $1 \leq k \leq n_i$ and $1 \leq i \leq I$, then one can use the constraints $\sum \alpha_i = 0$.

It is a useful exercise to write the $\mathbf{X}$-matrix and $\mathbf{H}$-matrix for this problem.

9.2.3 Handout

Write $\mathbf{G} = \binom{\mathbf{X}}{\mathbf{H}}$ for the joint data and constraint matrices.

[1] This is easy to see since if $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are the columns of the $\mathbf{X}$-matrix then $E[\mathbf{Y}] = \mathbf{X}\beta$ is a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_p$ which only span a r-dimensional space and a subset of $\mathbf{x}_1, \ldots, \mathbf{x}_p$ can be used to span this space. The vector $E[\mathbf{Y}]$ can be written as a linear combination of vectors in any such subset.

Note that we obtain

$$\mathbf{G}\tilde{\beta} = \begin{pmatrix} \mathbf{X}\beta \\ \mathbf{0} \end{pmatrix}$$

and thus

$$\mathbf{G}'\mathbf{G}\tilde{\beta} = \mathbf{X}'\mathbf{X}\beta$$

where $\mathbf{G}'\mathbf{G}$ is invertible and we can write

$$\tilde{\beta} = \left(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H}\right)^{-1}\mathbf{X}'\mathbf{X}\beta \tag{9}$$

and we have

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

an unbiased estimate.

Note 9.1. The vector $\tilde{\beta}$ defined in Eq. (9) is a vector of elements, each of which is a linear function of β and each of these functions is estimable since each is of the form $\mathbf{a}'\mathbf{X}\beta$.

References ISBN: 0256117365
ISBN: 0471758345