

# Linear models with applications in R

*PUBHLTH 744: Handout 7(Estimation)*

Instructor: Andrea S. Foulkes

Division of Biostatistics and Epidemiology  
UMass School of Public Health and Health Sciences

Fall 2007

## Identifiability and estimability

Consider the general linear model  $Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}$ . We say that  $\beta$  is identifiable if knowing the mean  $E(Y)$  gives us  $\beta$ .

**Definition:** The parameterization  $\beta$  is identifiable if for any  $\beta_1$  and  $\beta_2$ ,  $f(\beta_1) = f(\beta_2)$  implies  $\beta_1 = \beta_2$ .

- ▶ For example, in the linear model setting, we have  $f(\beta) = X\beta$ .
- ▶ Note that in models for which  $r(X) = p$ , we have  $X'X$  is non-singular. Therefore,  $X\beta_1 = X\beta_2$  implies  $\beta_1 = (X'X)^{-1}X'X\beta_1 = (X'X)^{-1}X'X\beta_2 = \beta_2$  and so  $\beta$  is identifiable.

## Identifiability and estimability

**Theorem:** A function  $g(\beta)$  is identifiable if and only if  $g(\beta)$  is a function of  $f(\beta)$ .

We begin by assuming  $Cov(\epsilon) = \sigma^2 I$ . That is  $Cov(\epsilon_i, \epsilon_j) = 0$  (the errors are uncorrelated). Weighted least squares assumes the more general model in which  $Cov(\epsilon) = \Sigma$ . Our aim is to estimate  $\beta$  or some linear function of  $\beta$ ,  $\lambda'\beta$  where  $\lambda$  is a  $p \times 1$  vector of constants. For example, if  $\lambda' = (1, -1, 0, \dots, 0)$ , then  $\lambda'\beta = \beta_1 - \beta_2$ . First, we need to determine when  $\lambda'\beta$  is estimable.

## Identifiability and estimability

**Definition 7:**  $\lambda'\beta$  is estimable if there exists an  $n \times 1$  vector of constants,  $\rho$  such that

$$E(\rho'Y) = \lambda'\beta$$

Note that this is equivalent to saying that  $\lambda \in C(X')$ , since  $\lambda \in C(X')$  means

$$\lambda = X'\rho \Leftrightarrow \lambda' = \rho'X \Leftrightarrow \lambda'\beta = \rho'X\beta \Leftrightarrow \lambda'\beta = \rho'E(Y) = E(\rho'Y)$$

## Identifiability and estimability

**Definition:** An estimate  $f(Y)$  of  $\lambda'\beta$  is said to be a linear estimate for  $\lambda'\beta$  if  $f(Y) = a_0 + a'Y$ .

Note that the idea of estimability depends only on the mean function  $E(Y) = X\beta$  and not on the covariance  $Cov(Y)$ .

## Identifiability and estimability

**Example:** Consider the linear model  $Y = X\beta + \epsilon$  where

$$X = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} \text{ and } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}. \text{ Is } \beta_1 + \beta_2 \text{ estimable?}$$

Here  $\lambda' = (1, 1)$  since  $(1, 1)\beta = \beta_1 + \beta_2$  and

$C(X') = \mathcal{S} \left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\}$ . We see that  $\lambda \notin C(X')$  and so  $\beta_1 + \beta_2$  is not estimable. Note however that  $\beta_1 + 2\beta_2$  is estimable.

## Identifiability and estimability

**Definition** (system of equations): Suppose  $\Lambda$  is a  $p \times r$  matrix of constants. Then the  $r \times 1$  vector of linear functions  $\Lambda'\beta$  is estimable if and only if there exists an  $n \times r$  matrix of constants  $P$  such that  $P'X = \Lambda'$ .

Note that  $\Lambda'\beta$  is estimable if each of its components is estimable. Furthermore, if  $X$  is full rank, then  $X'X$  is invertible so we can let  $P = (X'X)^{-1}X'$  and therefore  $\Lambda' = P'X = (X'X)^{-1}(X'X) = I_{p \times p}$ . This implies that each element of  $\beta$  is estimable.

## Identifiability and estimability

**Example** Again let  $X = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix}$  and  $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ . Is

$\begin{pmatrix} \beta_1 + 2\beta_2 \\ \beta_1 + \beta_2 \\ \beta_2 \end{pmatrix}$  estimable?

In this case we have  $\Lambda'\beta = \begin{pmatrix} \beta_1 + 2\beta_2 \\ \beta_1 + \beta_2 \\ \beta_2 \end{pmatrix}$  where  $\Lambda = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}$ .

Further,  $\lambda'_1 = (1 \ 2) \in C(X')$ ; however,  $\lambda'_2 = (1 \ 1) \notin C(X')$  and  $\lambda'_3 = (0 \ 1) \notin C(X')$ . Therefore,  $\Lambda'X$  is not estimable.

## Identifiability and estimability

Alternatively, we could find an estimable function by first choosing  $P$  and letting  $\Lambda' = P'X$ . For example, if we let

$$P' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 3 & 0 \end{pmatrix}. \text{ Then we have } \Lambda' = P'X = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{pmatrix} \text{ and}$$

$\Lambda'\beta$  is estimable.

## Least squares estimation

**Definition** The least squares estimate of  $\beta$ , denoted  $\hat{\beta}$ , satisfies

$$(Y - X\hat{\beta})'(Y - X\hat{\beta}) = \min_{\beta} (Y - X\beta)'(Y - X\beta)$$

## Identifiability and estimability

Note the least squares estimator of  $\beta$  minimizes the squared Euclidean distance between  $Y$  and its mean  $\mu = X\beta$ . If  $Y \notin C(X)$  then a solution of the equation  $Y = X\beta$  does not exist in general and the least squares solution will be the closest vector to  $Y$  that lies in the  $C(X)$ . This is given by  $MY$ , the orthogonal projection of  $Y$  onto  $C(X)$ .

## Identifiability and estimability

**Theorem:**  $\hat{\beta}$  is the least squares solution to  $\beta$  if and only if  $X\hat{\beta} = MY$  where  $M = X(X'X)^{-1}X'$ .

*Proof:* Let  $\tilde{\beta}$  be an estimate of  $\beta$ . Then we have

$$\begin{aligned} & (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) \\ &= (Y - MY + MY - X\tilde{\beta})'(Y - MY + MY - X\tilde{\beta}) \\ &= (Y - MY)'(Y - MY) + (Y - MY)'(MY - X\tilde{\beta}) \\ &\quad + (MY - X\tilde{\beta})'(Y - MY) + (MY - X\tilde{\beta})'(MY - X\tilde{\beta}) \\ &= (Y - MY)'(Y - MY) + Y'(I - M)MY - Y'(I - M)X\tilde{\beta} \\ &\quad + (MY - X\tilde{\beta})'(MY - X\tilde{\beta}) \\ &= (Y - MY)'(Y - MY) + (MY - X\tilde{\beta})'(MY - X\tilde{\beta}) \end{aligned}$$

Note that both terms are positive and the first term does not depend on  $\beta$ . So this quantity is minimized as a function of  $\beta$  when  $(MY - X\tilde{\beta}) = 0$  or equivalently,  $MY = X\tilde{\beta}$ .

## Identifiability and estimability

**Corollary:**  $(X'X)^{-1}X'Y$  is a least squares estimate of  $\beta$ .

**Corollary:** The unique least squares estimate of  $\Lambda'\beta = P'X\beta$  is  $P'MY$ .

## Identifiability and estimability

**Theorem:** If  $\Lambda' = P'X$ , then  $E(P'MY) = \Lambda'\beta$ . That is, if  $\Lambda'\beta$  is estimable, then its unique least squares estimate is unbiased.

Now consider estimation of  $\sigma^2$ . We can write  $Y = MY + (I - M)Y$ , where  $MY \in C(X)$  and  $(I - M)Y \in C(X)^\perp$ . We have  $MY = MX\beta + M\epsilon = X\beta + M\epsilon$ . Similarly,  $(I - M)Y = (I - M)X\beta + (I - M)\epsilon = (I - M)\epsilon$ . Since  $(I - M)Y$  depends only on  $\epsilon$ , we construct an estimator for  $\sigma$  based on this quantity.

**Theorem:** Let  $r(X) = r$ . Then  $Y'(I - M)Y/(n - r)$  is an unbiased estimate of  $\sigma^2$ .

```
#####
# R code for fitting linear models (finding coefficient estimates and
# an estimate of the standard deviation of the model error.)
#####

# Create the design matrix X
> X <- matrix(c(1,2,1,1,2,3,4,,2,4,6,3,5),nrow=4,ncol=3)
> X
      [,1] [,2] [,3]
[1,]    1    2    4
[2,]    2    3    6
[3,]    1    4    3
[4,]    1    2    5

# Note: X'X is invertible
> solve(t(X)%*%X)
      [,1]      [,2]      [,3]
[1,] 4.107692 -4.000000e-01 -9.230769e-01
[2,] -0.400000  2.000000e-01 -2.220446e-17
[3,] -0.923077  3.279428e-16  2.692308e-01

# Define Y in the C(X)
> Y <- 6*X[,1] - X[,3]
> Y
[1] 2 6 3 1

# Fit linear model with no intercept using lm() function
# We expect our coefficient estimates to be 6 for X1 and -1 for X2 since Y is in C(X)
> mod <- lm(Y ~ -1 + X)
> mod

Call:
lm(formula = Y ~ -1 + X)
```

```

Coefficients:
      X1          X2          X3
6.000e+00  3.976e-18 -1.000e+00

# Note that residuals from this model fit are all 0
> summary(mod)

Call:
lm(formula = Y ~ -1 + X)

Residuals:
      1          2          3          4
-1.767e-15  3.534e-16  1.767e-16  8.834e-16

Coefficients:
      Estimate Std. Error  t value Pr(>|t|)
X1  6.000e+00  4.083e-15  1.470e+15  4.33e-16 ***
X2  3.976e-18  9.009e-16  4.000e-03   0.997
X3 -1.000e+00  1.045e-15 -9.567e+14  6.65e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.015e-15 on 1 degrees of freedom
Multiple R-Squared:  1,      Adjusted R-squared:  1
F-statistic: 4.107e+30 on 3 and 1 DF,  p-value: 3.627e-16

# Find coefficient estimates using least squares

> solve(t(X)%*%X)%*%t(X)%*%Y
      [,1]
[1,] 6.000000e+00
[2,] -1.221245e-15
[3,] -1.000000e+00

# Now consider Y not in the C(X)

> Y <- 6*X[,1] - X[,3] + rnorm(4,mean=0,sd=1)

```

```

> Y
[1] 3.064242 6.830868 2.202531 2.984807
> mod <- lm(Y ~ -1 + X)
> mod

Call:
lm(formula = Y ~ -1 + X)

Coefficients:
      X1      X2      X3
4.43086 -0.48516 -0.09627

> summary(mod)

Call:
lm(formula = Y ~ -1 + X)

Residuals:
      1      2      3      4
-0.011222  0.002244  0.001122  0.005611

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
X1  4.430855    0.025931  170.87  0.00373 **
X2 -0.485161    0.005722  -84.79  0.00751 **
X3 -0.096267    0.006639  -14.50  0.04383 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01279 on 1 degrees of freedom
Multiple R-Squared:  1,    Adjusted R-squared:  1
F-statistic: 1.422e+05 on 3 and 1 DF,  p-value: 0.001950

# Using the least squares solution, we get again get the same coefficient estimates
> solve(t(X)%*%X)%*%t(X)%*%Y
      [,1]
[1,] 4.43085546

```

```

[2,] -0.48516097
[3,] -0.09626745

# Estimating sigma for this model
> M <- X%*%solve(t(X)%*%X)%*%t(X)
> n <- dim(X)[1]
> r <- dim(X)[2]
> sig.sq <- t(Y)%*%(diag(1,n)-M)%*%Y/(n-r)
> sig.sq
      [,1]
[1,] 0.0001636997
> sqrt(sig.sq)
      [,1]
[1,] 0.01279452

# Finding the estimate of sigma from the model output
> attributes(summary(mod))
$names
 [1] "call"          "terms"          "residuals"      "coefficients"
 [5] "aliased"        "sigma"          "df"              "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"

$class
[1] "summary.lm"

> summary(mod)$sigma
[1] 0.01279452

# R uses the QR factorization theorem to fit models with the lm() function
> qr.coef(qr(X),Y)
[1] 4.43085546 -0.48516097 -0.09626745
> qr.solve(X,Y)
[1] 4.43085546 -0.48516097 -0.09626745
> qr.resid(qr(X),Y)
[1] -0.011221537 0.002244307 0.001122154 0.005610768

# Note that the solve() function only works if we feed it a QR object

```

```

> solve(X,Y)
Error in drop(Call("La_dgesv", a, as.matrix(b), tol, PACKAGE = "base")) :
  'A' (4 x 3) must be square
> solve(qr(X),Y)
[1] 4.43085546 -0.48516097 -0.09626745

# How does this work?  $Y=Xb \Rightarrow Q'Y=Q'QRb \Rightarrow Q'Y=Rb \Rightarrow R^{-1}Q'Y=b$ 
# (note R is always invertible since it is upper triangular.)
> Q <- qr.Q(qr(X))
> R <- qr.R(qr(X))
> Q
      [,1]      [,2]      [,3]
[1,] -0.3779645 -5.239407e-17 0.2964997
[2,] -0.7559289 -4.472136e-01 -0.4447496
[3,] -0.3779645 8.944272e-01 -0.2223748
[4,] -0.3779645 1.326632e-17 0.8153742
> R
      [,1]      [,2]      [,3]
[1,] -2.645751 -5.291503 -9.071147e+00
[2,] 0.000000 2.236068 -5.128276e-17
[3,] 0.000000 0.000000 1.927248e+00
> solve(R)
      [,1]      [,2]      [,3]
[1,] -0.3779645 -0.8944272 -1.778998e+00
[2,] 0.0000000 0.4472136 1.190005e-17
[3,] 0.0000000 0.0000000 5.188745e-01
> solve(R)%*%t(Q)%*%Y
      [,1]
[1,] 4.43085546
[2,] -0.48516097
[3,] -0.09626745

# We can use the SVD as well to find the solution since  $X=UDt(V)$  and  $[VD^{-1}t(U)][UDt(V)]=I$ 
> V <- svd(X)$v
> U <- svd(X)$u
> D <- diag(svd(X)$d)
> U %*% D %*% t(V)

```

```

      [,1] [,2] [,3]
[1,]    1    2    4
[2,]    2    3    6
[3,]    1    4    3
[4,]    1    2    5
> V %*% diag(1/diag(D)) %*% t(U) %*% Y
      [,1]
[1,] 4.43085546
[2,] -0.48516097
[3,] -0.09626745

```

# We can also use the lsfit() function to arrive at the solution

```

> lsfit(X,Y,intercept=FALSE)
$coefficients
      X1          X2          X3
4.43085546 -0.48516097 -0.09626745

$residuals
[1] -0.011221537  0.002244307  0.001122154  0.005610768

$intercept
[1] FALSE

$qr
$qt
[1] -8.28245482 -1.08485292 -0.18553128  0.01279452

$qr
      X1          X2          X3
[1,] -2.6457513 -5.291503e+00 -9.071147e+00
[2,]  0.7559289  2.236068e+00 -5.128276e-17
[3,]  0.3779645 -8.944272e-01  1.927248e+00
[4,]  0.3779645 -2.763759e-17 -7.340468e-01

$qraux
[1] 1.377964 1.447214 1.679099

```

```

$rank
[1] 3

$pivot
[1] 1 2 3

$tol
[1] 1e-07

attr(,"class")
[1] "qr"

# Now consider a design matrix that is less than full rank
> X[,2] <- 2*X[,1]
> X
      [,1] [,2] [,3]
[1,]    1    2    4
[2,]    2    4    6
[3,]    1    2    3
[4,]    1    2    5
> mod <- lm(Y ~ -1 + X)
> mod

Call:
lm(formula = Y ~ -1 + X)

Coefficients:
      X1      X2      X3
 3.46053      NA -0.09627

> summary(mod)

Call:
lm(formula = Y ~ -1 + X)

```

Residuals:

	1	2	3	4
	-0.011222	0.487405	-0.969200	0.005611

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t )
X1	3.46053	1.39524	2.480	0.131
X2	NA	NA	NA	NA
X3	-0.09627	0.39806	-0.242	0.831

Residual standard error: 0.7672 on 2 degrees of freedom

Multiple R-Squared: 0.9831, Adjusted R-squared: 0.9663

F-statistic: 58.31 on 2 and 2 DF, p-value: 0.01686

# In this case, we can find a Moore-Penrose least squares solution

```
> V1 <- svd(X)$v[,-3]
> U1 <- svd(X)$u[,-3]
> D <- diag(svd(X)$d)[1:2,1:2]
> U1 %*% D %*% t(V1)
      [,1] [,2] [,3]
[1,] 1 2 4
[2,] 2 4 6
[3,] 1 2 3
[4,] 1 2 5
> Xplus <- V1 %*% diag(1/diag(D)) %*% t(U1)
> X %*% Xplus %*% X # checking that this equals X
      [,1] [,2] [,3]
[1,] 1 2 4
[2,] 2 4 6
[3,] 1 2 3
[4,] 1 2 5
> Xplus %*% Y
      [,1]
[1,] 0.69210670
[2,] 1.38421340
[3,] -0.09626745
```

# Consider what happens if we have the first and second columns

```

# highly collinear (but not perfectly correlated.)
# Note: model fits but estimates and standard errors are not reasonable
> X[,2] <- X[,2]+rnorm(4,0,.001)
> X
      [,1]      [,2] [,3]
[1,]    1 2.000588    4
[2,]    2 4.001930    6
[3,]    1 2.000493    3
[4,]    1 1.998926    5
> mod <- lm(Y ~ -1 + X)
> mod

Call:
lm(formula = Y ~ -1 + X)

Coefficients:
      X1          X2          X3
-1654.8905   827.7181    0.6206

> summary(mod)

Call:
lm(formula = Y ~ -1 + X)

Residuals:
    1      2      3      4
-0.4503  0.4189 -0.6126  0.2252

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
X1 -1654.8905  2435.3802  -0.680  0.620
X2  827.7181  1215.5495   0.681  0.619
X3    0.6206    1.1510   0.539  0.685

Residual standard error: 0.8968 on 1 degrees of freedom
Multiple R-Squared: 0.9885,    Adjusted R-squared: 0.9539
F-statistic: 28.6 on 3 and 1 DF,  p-value: 0.1364

```

## Identifiability and estimability

**Theorem:** Suppose  $Y_{n \times 1}$  is a random vector such that  $E(Y) = \mu$  and  $Cov(Y) = \Sigma$ . Let  $A_{n \times n}$  matrix of constants. Then we have  $E(Y'AY) = \mu' A \mu + tr(A\Sigma)$ .

## Best linear unbiased estimator

Desirable properties of estimators:

- ▶ Unbiased
- ▶ Minimum variance
- ▶ Efficient
- ▶ Asymptotically normal

Now consider estimation of  $\lambda'\beta$ . We want to find the *best* linear unbiased estimate of  $\lambda'\beta$ . Note that linear means that the estimator is a linear function of  $Y$ . Best refers to minimum variance. That is, we aim to find  $a'Y$  such that  $E(a'Y)=\lambda'\beta$  and  $Var(a'Y) \leq V(b'Y)$  for any  $b \in \mathcal{R}^n$ .

## Best linear unbiased estimator

**Theorem:** (Gauss-Markov) Consider the linear model  $Y = X\beta + \epsilon$  where  $E(\epsilon) = 0$  and  $Cov(\epsilon) = \sigma^2 I$ . If  $\lambda'\beta$  is estimable, then the unique least squares estimate of  $\lambda'\beta$  is the unique best linear unbiased estimator (BLUE) of  $\lambda'\beta$ .

Proof: See Christensen, Section 2.3

## Weighted least squares

Consider the linear model  $Y = X\beta + \epsilon$  where  $E(\epsilon) = 0$  and  $Cov(\epsilon) = \sigma^2 V$  where  $V$  is a known positive definite matrix.

Examples:

- ▶ Repeated measures.  $V$  unknown.

- ▶ Split plot design.  $V = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \dots & & 1 \end{pmatrix}$  unknown.

- ▶ Family studies.  $V$  unknown.

- ▶ Observations are averages:  $\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$ ,  $i = 1, \dots, n$ . In this case, the covariance matrix is  $\sigma^2 V$  where  $V$  is a diagonal matrix and the  $i$ th diagonal element of  $V$  is  $1/m_i$ . In this case  $V$  is known and we are in the weighted least squares framework.

## Weighted least squares

Again, we aim to estimate  $\beta$  and  $\sigma$ . To do this, first note that  $V$  is positive definite and so we can write  $V = QQ'$  where  $Q$  is non-singular (Cholesky decomposition). Therefore,  $Q^{-1}VQ'^{-1} = I$  and we can re-write our model in the form:

$$Q^{-1}Y = Q^{-1}X\beta + Q^{-1}\epsilon$$

where  $E(Q^{-1}\epsilon) = Q^{-1}E(\epsilon) = 0$  and  $Cov(Q^{-1}\epsilon) = Q^{-1}\sigma^2VQ'^{-1} = \sigma^2I$ .

## Weighted least squares

The least squares estimate of  $\beta$  minimizes:

$$\begin{aligned} & (Q^{-1}Y - Q^{-1}X\beta)'(Q^{-1}Y - Q^{-1}X\beta) \\ &= (Y - X\beta)'Q'^{-1}Q^{-1}(Y - X\beta) \\ &= (Y - X\beta)'V^{-1}(Y - X\beta) \end{aligned}$$

## Weighted least squares

**Theorem:**  $\hat{\beta}$  is a weighted least squares estimate of  $\beta$  if and only if  $X(X'V^{-1}X)^{-1}X'V^{-1}Y = X\hat{\beta}$ .

Here we use the result that if  $X^* = Q^{-1}X$ , then the orthogonal projection operator onto  $X^*$  is given by

$$M^* = X^*(X^{*'}X^*)^{-1}X^{*'} = Q^{-1}X(X'V^{-1}X)^{-1}X'Q'^{-1}.$$

$M^*Y^* = X^*\beta$  then reduces to the above formula.

## Weighted least squares

**Theorem:** For any estimable function  $\lambda'\beta$  where  $\lambda' = \rho'X$ , the unique weighted least squares estimate of  $\lambda'\beta$  is  $\rho'AY$  where  $A = X(X'V^{-1}X)^{-}X'V^{-1}$ . The unique weighted least squares estimate of  $\mu = X\beta$  is  $AY$ .

Note: ICBT  $A$  is invariant to choice of generalized inverse and  $A$  is a projection operator onto  $C(X)$  along  $N(A)$ .

**Theorem:** For weighted least squares, an unbiased estimator of  $\sigma^2$  is given by  $\hat{\sigma}^2 = Y'(I - A)'V^{-1}(I - A)Y/(n - r)$

## Maximum likelihood

No distributional assumptions about  $\epsilon$  were made to obtain least squares estimates, BLUEs and weighted least squares estimates. Now we consider the setting in which  $\epsilon \sim MVN(0, \sigma^2 I)$ . This implies  $Y \sim MVN(X\beta, \sigma^2 I)$ . The joint density of  $Y$  is then given by:

$$f(Y) = (2\pi)^{-n/2} \sigma^{-n} \exp \left\{ \frac{-1}{2\sigma^2} (Y - X\beta)'(Y - X\beta) \right\}$$

## Maximum likelihood

This uses the equality  $|\sigma^2 I|^{-1/2} = \sigma^{-n}$ . The likelihood (which is proportional to  $f(y)$ ) is written:

$$\mathcal{L}(\beta, \sigma) = \sigma^{-n} \exp \left\{ \frac{-1}{2\sigma^2} (Y - X\beta)'(Y - X\beta) \right\}$$

The maximum likelihood estimates (MLEs) are given by the parameter values that maximize this equation, or equivalently, maximize the natural log of the likelihood:

$$\log \mathcal{L}(\beta, \sigma) = -n \log(\sigma) - \frac{1}{2\sigma^2} (Y - X\beta)'(Y - X\beta)$$

## Maximum likelihood

Maximizing  $\log\mathcal{L}(\beta, \sigma)$  with respect to  $\beta$  is equivalent to minimizing  $g(\beta) = (Y - X\beta)'(Y - X\beta)$ . But this is the least squares criterion. Thus, the MLE of  $\beta$  is the least squares solution.  $X\hat{\beta}_{ml} = MY$ .

To find the *ML* of  $\sigma$ , we first let  $\beta = \hat{\beta}$  and then maximize the following function with respect to  $\sigma$ .

$$\begin{aligned}\log\mathcal{L}(\hat{\beta}, \sigma) &= -n\log(\sigma) - \frac{1}{2\sigma^2}(Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= -n\log(\sigma) - \frac{1}{2\sigma^2}(Y - MY)'(Y - MY) \\ &= -n\log(\sigma) - \frac{1}{2\sigma^2}Y'(I - M)Y\end{aligned}$$

## Maximum likelihood

Taking the derivative with respect to  $\sigma$  yields:

$$\begin{aligned}\frac{\delta \log \mathcal{L}(\hat{\beta}, \sigma)}{\delta \sigma} &= -n/\sigma + 1/\sigma^3 (Y'(I - M)Y) = 0 \\ \Leftrightarrow -n\sigma^2 + Y'(I - M)Y &= 0 \\ \Leftrightarrow \hat{\sigma}_{ml}^2 &= Y'(I - M)Y/n\end{aligned}$$

Notably,  $E(\hat{\sigma}_{ml}^2) = \frac{n-r}{n}\sigma^2 \rightarrow \sigma^2$  as  $n \rightarrow \infty$  (asymptotically unbiased.)

## Minimum variance unbiased estimation

The least squares estimate of  $\rho'X\beta$  is the unique minimum variance unbiased linear estimator (Gauss-Markov theorem.) If  $\epsilon \sim MVN(0, \sigma^2 I)$  then the least squares estimator is the uniform minimum variance unbiased estimator (UMVUE) – not restricted to linear, instead covers class of all unbiased estimators