# Prediction

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

April 8, 2019



## WLS and BLUE (prelude to BLUP)

Suppose that $Y$ has mean $X\beta$ and known covariance matrix $V$ (but $Y$ need not be normal). Then

$$\hat{\beta} = (X^\mathsf{T} V^{-1} X)^{-1} X^\mathsf{T} V^{-1} Y$$

is a weighted least squares estimate since it minimizes

$$(Y - X\beta)^\mathsf{T} V^{-1} (Y - X\beta).$$

It is also the best linear unbiased estimate (BLUE) - that is the unbiased estimate with smallest variance in the sense that

$$\mathbb{V}\mathrm{ar}\tilde{\beta} - \mathbb{V}\mathrm{ar}\hat{\beta}$$

is positive semi-definite for any other linear unbiased estimate $\tilde{\beta}$.



## BLUE for general parameter and $V = I$

Theorem: Suppose $\mathbb{E}Y = \mu$ is in linear subspace $M$ and $\mathbb{C}\mathrm{ov}Y = \sigma^2 I$ and $\psi = A\mu$. Then BLUE of $\psi$ is $\hat{\psi} = A\hat{\mu}$ where $\hat{\mu} = PY$ and $P$ orthogonal projection on $M$.

Key result:

$$\mathbb{C}\mathrm{ov}(\tilde{\psi} - \hat{\psi}, \hat{\psi}) = \mathbb{E}[(\tilde{\psi} - \hat{\psi})\hat{\psi}] = 0$$

for any other LUE $\psi = BY$.

Proof of theorem follows by key result:

$$\mathbb{V}\mathrm{ar}(\tilde{\psi}) = \mathbb{V}\mathrm{ar}(\tilde{\psi} - \hat{\psi}) + \mathbb{V}\mathrm{ar}\hat{\psi} \Rightarrow \mathbb{V}\mathrm{ar}(\tilde{\psi}) - \mathbb{V}\mathrm{ar}\hat{\psi} = \mathbb{V}\mathrm{ar}(\tilde{\psi} - \hat{\psi}) \geq 0.$$

Hence $\hat{\psi}$ is BLUE (here $A \geq B$ means $A - B$ positive semi definite).



Proof of key result:

Assume $\tilde{\psi}$ is LUE. I.e. $\tilde{\psi} = BY$ and $\mathbb{E}\tilde{\psi} = B\mu = A\mu$ for all $\mu \in L$. We also have $AP\mu = A\mu$ for all $\mu \in M$ (which implies that $\hat{\psi}$ is unbiased). Thus for all $w \in \mathbb{R}^p$

$$(B - AP)Pw = BPw - APw = APw - APw = 0$$

since $Pw \in M$. This implies $(B - AP)P = 0$ which gives

$$\mathbb{C}\mathrm{ov}(\tilde{\psi} - \hat{\psi}, \hat{\psi}) = \sigma^2 (B - AP) P^T A^\mathsf{T} = 0.$$

## BLUE - non-diagonal covariance matrix

Lemma: suppose $\tilde{Y} = KY$ where $K$ is an invertible matrix. If $\hat{\psi} = C\tilde{Y}$ is BLUE of $\psi$ based on data $\tilde{Y}$ then $\hat{\psi} = CKY$ is BLUE based on $Y$ as well.

Corollary: suppose $V = LL^\mathsf{T}$ is invertible and $\mu = X\beta$ where $X$ has full rank. Then BLUE of $\psi = A\mu$ is $A\hat{\mu}$ where $\hat{\mu} = X(X^T V^{-1} X)^{-1} X^T V^{-1} Y$ is WLS estimate of $\mu$.

Proof: $\tilde{Y} = L^{-1}Y$ has covariance matrix $I$ and mean $\tilde{\mu} = \tilde{X}\beta$ where $\tilde{\mu} = L^{-1}\mu$. Thus by theorem, BLUE of $A\mu = AL\tilde{\mu}$ is $AL\tilde{X}(\tilde{X}^\mathsf{T}\tilde{X})^{-1}\tilde{X}^\mathsf{T}\tilde{Y}$. Applying lemma we get BLUE based on $Y$ is $AL\tilde{X}(\tilde{X}^\mathsf{T}\tilde{X})^{-1}\tilde{X}^\mathsf{T}L^{-1}Y = A\hat{\mu}$.

Remark: $\hat{\mu}$ above is in fact orthogonal projection of $Y$ wrt. inner product $< x, y >= x^\mathsf{T} V^{-1} y$.



## †Estimable parameters and BLUE

Definition: A linear combination $a^\mathsf{T}\beta$ is estimable if it has a LUE $b^\mathsf{T} Y$.

Result: $a^\mathsf{T}\beta$ is estimable $\Leftrightarrow$ $a^\mathsf{T}\beta = c^\mathsf{T}\mu$ for some $c$.

By results on previous slides: If $a^\mathsf{T}\beta$ is estimable then BLUE is $c^\mathsf{T}\hat{\mu}$.



## Optimal prediction

$X$ and $Y$ random variables, $g$ real function. General result:

$$\mathbb{C}\mathrm{ov}(Y - \mathbb{E}[Y|X], g(X)) = \mathbb{C}\mathrm{ov}(\mathbb{E}[Y - \mathbb{E}[Y|X]|X], \mathbb{E}[g(X)|X]) + \mathbb{E}\mathbb{C}\mathrm{ov}(Y - \mathbb{E}[Y|X], g(X)|X) = 0$$

In particular, for any prediction $\tilde{Y} = f(X)$ of $Y$:

$$\mathbb{E}\big[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - f(X))\big] = 0$$

from which it follows that

$$\mathbb{E}(Y - \tilde{Y})^2 = \mathbb{E}(Y - \mathbb{E}[Y|X])^2 + \mathbb{E}(\mathbb{E}[Y|X] - \tilde{Y})^2 \geq \mathbb{E}(Y - \mathbb{E}[Y|X])^2$$

Thus $\mathbb{E}[Y|X]$ minimizes mean square prediction error.



## †Pythagoras and conditional expectation

Space of real random variables with finite variance may be viewed as a vector space with inner product and ($L_2$) norm

$$< X, Y >= \mathbb{E}(XY) \quad \|X\| = \sqrt{\mathbb{E}X^2}$$

Orthogonal decomposition (Pythagoras):

$$\|Y\|^2 = \|\mathbb{E}[Y|X]\|^2 + \|Y - \mathbb{E}[Y|X]\|^2$$

$\mathbb{E}[Y|X]$ may be viewed as projection of $Y$ on $X$ since it minimizes distance

$$\mathbb{E}(Y - \tilde{Y})^2$$

among all predictors $\tilde{Y} = f(X)$.

For zero-mean random variables, orthogonal is the same as uncorrelated.

(Grimmett & Stirzaker, Prob. and Random Processes, Chapter 7.9 good source on this perspective on prediction and conditional expectation)

Decomposition of $Y$:

$$Y = \mathbb{E}[Y|X] + (Y - \mathbb{E}[Y|X])$$

where predictor $\mathbb{E}[Y|X]$ and prediction error $Y - \mathbb{E}[Y|X]$ uncorrelated.

Thus,

$$\mathbb{V}\mathrm{ar}Y = \mathbb{V}\mathrm{ar}\mathbb{E}[Y|X] + \mathbb{V}\mathrm{ar}(Y - \mathbb{E}[Y|X]) = \mathbb{V}\mathrm{ar}\mathbb{E}[Y|X] + \mathbb{E}\mathbb{V}\mathrm{ar}[Y|X]$$

whereby

$$\mathbb{V}\mathrm{ar}(Y - \mathbb{E}[Y|X]) = \mathbb{E}\mathbb{V}\mathrm{ar}[Y|X].$$

Prediction variance is equal to the expected conditional variance of $Y$.



## BLUP

Consider random vectors $Y$ and $X$ with with mean vectors

$$\mathbb{E}Y = \mu_Y \quad \mathbb{E}X = \mu_X$$

and covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_Y & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_X \end{bmatrix}$$

Then the best *linear* unbiased predictor of $Y$ given $X$ is

$$\hat{Y} = \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(X - \mu_X)$$

in the sense that

$$\mathbb{V}\mathrm{ar}[Y - (a + BX)] - \mathbb{V}\mathrm{ar}[Y - \hat{Y}]$$

is positive semi-definite for all linear unbiased predictors $a + BX$ and '=' only if $\hat{Y} = a + BX$ (unbiased: $\mathbb{E}[Y - a - BX] = 0$).



## Prediction variance/mean square prediction error

Fact:

$$\mathbb{C}\mathrm{ov}[Y - \hat{Y}, CX] = 0 \quad \text{for all} \quad C. \tag{1}$$

Thus $\mathbb{C}\mathrm{ov}[Y - \hat{Y}, \hat{Y}] = 0$ which implies

$$\mathbb{V}\mathrm{ar}\hat{Y} = \mathbb{C}\mathrm{ov}(Y, \hat{Y}) = \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$$

It follows that mean square prediction error is

$$\mathbb{V}\mathrm{ar}[Y - \hat{Y}] = \mathbb{V}\mathrm{ar}Y + \mathbb{V}\mathrm{ar}\hat{Y} - 2\mathbb{C}\mathrm{ov}(Y, \hat{Y}) = \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$$

Proof of fact:

$$\mathbb{C}\mathrm{ov}[Y - \hat{Y}, CX] = \mathbb{C}\mathrm{ov}[Y, CX] - \mathbb{C}\mathrm{ov}[\hat{Y}, CX] = \\ \Sigma_{YX}C^{\mathsf{T}} - \Sigma_{YX}\Sigma_X^{-1}\Sigma_X C^{\mathsf{T}} = 0$$



## Proof of BLUP

By (1), $\mathbb{C}\mathrm{ov}[Y - \hat{Y}, CX] = 0$ for all $C$.

$$\mathbb{V}\mathrm{ar}[Y - (a + BX)] = \mathbb{V}\mathrm{ar}[Y - \hat{Y}] + \mathbb{V}\mathrm{ar}[\hat{Y} - (a + BX)] + \\ \mathbb{C}\mathrm{ov}[Y - \hat{Y}, \hat{Y} - (a + BX)] + \mathbb{C}\mathrm{ov}[\hat{Y} - (a + BX), Y - \hat{Y}] = \\ \mathbb{V}\mathrm{ar}[Y - \hat{Y}] + \mathbb{V}\mathrm{ar}[\hat{Y} - (a + BX)]$$

Hence $\mathbb{V}\mathrm{ar}[Y - (a + BX)] - \mathbb{V}\mathrm{ar}[Y - \hat{Y}] = \mathbb{V}\mathrm{ar}[\hat{Y} - (a + BX)]$ where right hand side is positive semi-definite.

## †BLUP as projection

$Y$ scalar for consistency with slide on $L_2$ space view. $X = (X_1, \dots, X_n)^\mathsf{T}$. Assume wlog that all variables are centered $\mathbb{E}Y = \mathbb{E}X_i = 0$ (otherwise consider prediction of $Y - \mathbb{E}Y$ based on $X_i - \mathbb{E}X_i$).

BLUP is projection of $Y$ onto *linear* subspace spanned by $X_1, \dots, X_n$ (with orthonormal basis $U_1, \dots, U_n$ where $U = \Sigma_X^{-1/2} X$):

$$\hat{Y} = \sum_{i=1}^{n} \mathbb{E}[YU_i]U_i = \Sigma_{YX}\Sigma_X^{-1}X$$

(analogue to least squares).

NB: conditional expectation $\mathbb{E}[Y|X]$ projection of $Y$ onto space of *all* variables $Z = f(X_1, \dots, X_n)$ where $f$ real function.



## Conditional distribution in multivariate normal distribution

Consider jointly normal random vectors $Y$ and $X$ with mean vector

$$\mu = (\mu_Y, \mu_X)$$

and covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_Y & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_X \end{bmatrix}$$

Then (provided $\Sigma_X$ invertible)

$$Y|X = x \sim N(\mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X), \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})$$

Proof: By BLUP

$$Y = \hat{Y} + R$$

where $\hat{Y} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$, $R = Y - \hat{Y} \sim N(0, \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})$ and $\mathbb{C}\mathrm{ov}(R, X) = 0$. By normality $R$ is independent of $X$. Given $X = x$, $\hat{Y}$ is constant and distribution of $R$ is not affected. Thus result follows.



## Optimal prediction for jointly normal random vectors

By previous result it follows that BLUP of $Y$ given $X$ coincides with $E[Y|X]$ when $(X, Y)$ jointly normal.

Hence for normally distributed $(X, Y)$, BLUP is optimal prediction.



## Conditional simulation using prediction

Suppose $Y$ and $X$ are jointly normal and we wish to simulate $Y|X = x$. By previous result

$$Y|X = x \sim \hat{y} + R$$

where $\hat{y} = \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X)$. We thus need to simulate $R$. This can be done by 'simulated prediction': simulate $(Y^*, X^*)$ and compute $\hat{Y}^*$ and $R^* = Y^* - \hat{Y}^*$.

Then our conditional simulation is

$$\hat{y} + R^*$$

Advantageous if it is easier to simulate $(Y^*, X^*)$ and compute $\hat{Y}^*$ than simulate directly from conditional distribution of $Y|X = x$

(e.g. if simulation of $(Y, X)$ easy but $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ difficult)

## Prediction in linear mixed model

Let $U \sim N(0, \Psi)$ and $Y|U = u \sim N(X\beta + Zu, \Sigma)$.

Then $\mathbb{C}\mathrm{ov}[U, Y] = \Psi Z^\mathsf{T}$ and $\mathbb{V}\mathrm{ar} Y = V = Z\Psi Z^\mathsf{T} + \Sigma$.

Thus

$$\hat{U} = \mathbb{E}[U|Y] = \Psi Z^\mathsf{T} V^{-1}(Y - X\beta)$$

NB: by Woodbury

$$\Psi Z^\mathsf{T}(Z\Psi Z^\mathsf{T} + \Sigma)^{-1} = (\Psi^{-1} + Z^\mathsf{T}\Sigma^{-1}Z)^{-1}Z^\mathsf{T}\Sigma^{-1}$$

- e.g. useful if $\Psi^{-1}$ is sparse (like AR-model).

Similarly

$$\mathbb{V}\mathrm{ar}[U - \hat{U}] = \mathbb{V}\mathrm{ar}[U|Y] = \Psi - \Psi Z^\mathsf{T} V^{-1} Z \Psi^\mathsf{T} = (\Psi^{-1} + Z^\mathsf{T}\Sigma^{-1}Z)^{-1}$$

One-way anova example at p. 186 in M & T.



## IQ example

$Y$ measurement of $IQ$, $U$ subject specific random effect:

$$Y = \mu + U + \epsilon$$

where standard deviation of $U$ and $\epsilon$ are 15 and 5 and $\mu = 100$.

Given $Y = 130$, $\mathbb{E}[\mu + U|Y = 130] = 127$.

Example of shrinkage to the mean.



## †BLUP as hierarchical likelihood estimates

Maximization of joint density ('hierarchical likelihood')

$$f(y|u;\beta)f(u;\psi)$$

with respect to $u$ gives BLUP (M & T p. 171-172 for one-way anova and p. 183 for general linear mixed model)

Joint maximization wrt. $u$ and $\beta$ gives Henderson's mixed-model equations (M & T p. 184) leading to BLUE $\hat{\beta}$ and BLUP $\hat{u}$.



## BLUP of mixed effect with unknown $\beta$

Assume $\mathbb{E}X = C\beta$ and $\mathbb{E}Y = D\beta$. Given $X$ and $\beta$, BLUP of

$$K = A\beta + BY$$

is

$$\hat{K}(\beta) = A\beta + B\hat{Y}(\beta)$$

where BLUP $\hat{Y}(\beta) = D\beta + \Sigma_{YX}\Sigma_X^{-1}(X - C\beta)$.

Typically $\beta$ is unknown. Then BLUP is

$$\hat{K} = A\hat{\beta} + B\hat{Y}(\hat{\beta})$$

where $\hat{\beta}$ is BLUE (Harville, 1991)

Proof: $\hat{K}(\beta)$ can be rewritten as

$$A\beta + B\hat{Y}(\beta) = [A + BD - B\Sigma_{YX}\Sigma_X^{-1}C]\beta + B\Sigma_{YX}\Sigma_X^{-1}X = T + B\Sigma_{YX}\Sigma_X^{-1}X$$

Note BLUE of $T = [A + BD - B\Sigma_{YX}\Sigma_X^{-1}C]\beta$ is $\hat{T} = [A + BD - B\Sigma_{YX}\Sigma_X^{-1}C]\hat{\beta}$.

Now consider a LUP $\tilde{K} = HX = [H - B\Sigma_{YX}\Sigma_X^{-1}]X + B\Sigma_{YX}\Sigma_X^{-1}X$ of K. By unbiasedness,

$$\tilde{T} = [H - B\Sigma_{YX}\Sigma_X^{-1}]X$$

is LUE of $T$. Hence $\mathbb{V}\mathrm{ar}[\tilde{T} - T] \geq \mathbb{V}\mathrm{ar}[\hat{T} - T]$. Also note by (1)

$$\mathbb{C}\mathrm{ov}[\tilde{T} - T, \hat{K}(\beta) - K] = 0 \text{ and } \mathbb{C}\mathrm{ov}[\hat{T} - T, \hat{K}(\beta) - K] = 0$$

Using this it follows that

$$\mathbb{V}\mathrm{ar}[\tilde{K} - K] \geq \mathbb{V}\mathrm{ar}[\hat{K} - K]$$

Hint: subtract and add $\hat{K}(\beta)$ both in $\mathbb{V}\mathrm{ar}[\tilde{K} - K]$ and $\mathbb{V}\mathrm{ar}[\hat{K} - K]$.



## EBLUP and EBLUE

Typically covariance matrix depends on unknown parameters.

EBLUPS are obtained by replacing unknown variance parameters by their estimates (similar for EBLUE).



## Model assessment

Make histograms, qq-plots etc. for EBLUPs of $\epsilon$ and $U$.

May be advantageous to consider standardized EBLUPS. Standardized BLUP is

$$[\mathbb{C}\mathrm{ov}\hat{U}]^{-1/2}\hat{U}$$

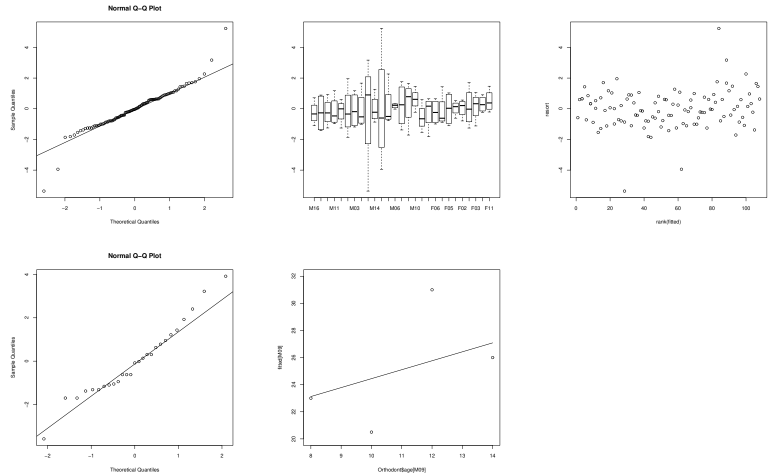


## Example: prediction of random intercepts and slopes in orthodont data

```
ort7=lmer(distance~age+factor(Sex)+(1|Subject),data=Orthodo
#check of model ort7
#residuals
res=residuals(ort7)
qqnorm(res)
qqline(res)
#outliers occur for subjects M09 and M13
#plot residuals against subjects
boxplot(resort~Orthodont$Subject)
#plot residuals against fitted values
fitted=fitted(ort7)
plot(rank(fitted),resort)
```

```
#extract predictions of random intercepts
raneffects=ranef(ort7)
#qqplot of random intercepts
qqnorm(ranint[[1]])
qqline(ranint[[1]])
#plot for subject M09
M09=Orthodont$Subject=="M09"
plot(Orthodont$age[M09],fitted[M09],type="l",ylim=c(20,32)
points(Orthodont$age[M09],Orthodont$distance[M09])
```



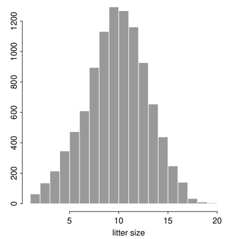




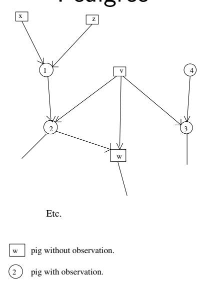

## Example: quantitative genetics (Sorensen and Waagepetersen 2003)

$X_{ij}$ size of $j$th litter of $i$th pig.

Histogram




Pedigree






$U_i$, $\tilde{U}_i$ random genetic effects influencing size and variability of $X_{ij}$:

$$X_{ij}|U_i = u_i, \tilde{U}_i = \tilde{u}_i \sim N\left(\mu_i + u_i, \exp(\tilde{\mu}_i + \tilde{u}_i)\right)$$

$$(U_1, \ldots, U_n, \tilde{U}_1, \ldots, \tilde{U}_n) \sim N(0, G \otimes A)$$

$A$: additive genetic relationship (correlation) matrix (depending on pedigree). Correlation structure derived from simple model:

$$U_{\text{offspring}} = \frac{1}{2}(U_{\text{father}} + U_{\text{mother}}) + \epsilon$$

$\Rightarrow Q = A^{-1}$ sparse ! (generalization of AR(1))

$$G = \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_{\tilde{u}} \\ \rho\sigma_u\sigma_{\tilde{u}} & \sigma_{\tilde{u}}^2 \end{bmatrix}$$

$\rho$: coefficient of genetic correlation between $U_i$ and $\tilde{U}_i$.

**NB**: high dimension $n > 6000$.

Aim: identify pigs with favorable genetic effects

## Exercises

1. Fill in the details of the proof on slide 21.
2. Verify the results on page 186 in M&T regarding BLUPs in case of a one-way anova.
3. Consider slides 15-16 and the special case of the hidden AR(1) model from the second miniproject. Explain how you can do conditional simulation of the hidden AR(1) process $U$ given the observed data $Y$ using 1) conditional simulation using prediction 2) the expressions for the conditional mean and covariance (cf. slide 14). Try to implement the solutions in R.