LINEAR MODELS FOR LONGITUDINAL DATA

Modeling the Variance-Covariance-Correlation

- Models for longitudinal data involve a **mean model** and a **variance-covariance-correlation** model

- Reasons to build / test the covariance model:

  - make correct and efficient inferences on the mean model, or

  - because the covariance model is of direct interest

  - Here, we focus primarily on the first motivation

- We have already discussed two broad classes of covariance model

- **Unstructured** correlation model:

  - different correlation for each pair of times

  - requires balanced data

  - yields a large number of variance-covariance parameters

- **Covariance pattern** models, e.g.:

  - exchangeable or compound-symmetry

1

- exponential

- others (later notes and Fitzmaurice et al. Section 7.4)

- Recall that

$$Y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \epsilon_{ij}$$

and that the variance-covariance-correlation model applies to the ϵ_{ij}'s (i.e., represents the variance-covariance among the Y_{ij}'s **after** removing effects of whatever is in $\boldsymbol{x}_{ij}$

- Specifically, the v-c-c model will be a model for V_i in

$$\boldsymbol{\epsilon}_i \sim N(0, V_i)$$

- We will parameterize this model by $\boldsymbol{\gamma}$, writing $V_i = V_i(\boldsymbol{\gamma})$

- **Examples:**

  - $\boldsymbol{\gamma} = (\sigma^2, \rho)'$ in the exchangeable correlation model

  - $\boldsymbol{\gamma} = (\nu^2, \tau^2)'$ in the exchangeable correlation model (alternative parameterization)

- $\boldsymbol{\gamma} = (\delta^2, \alpha)'$ in the exponential correlation model

- $\boldsymbol{\gamma} = (\sigma_{11}, \sigma_{12}, \ldots)'$ in the unstructured correlation model (all variances and covariances, each representing its own parameter)

- The **general linear model** for longitudinal data can be written

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{\epsilon} \sim MVN(0, V)$$

where V is the **variance-covariance** matrix of $\boldsymbol{\epsilon}$ and we have concatenated all $\boldsymbol{Y}_i$'s, X_i's and $\boldsymbol{\epsilon}_i$'s from all subjects

- With **correlated data**, V is of the **block diagonal** form:

$$\mathbf{V} = \begin{pmatrix} \mathbf{V_1} & 0 & \ldots \\ 0 & \mathbf{V_2} & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

3

Estimation of the Variance-Covariance-Correlation Model
ML and ReML

- Inference on the mean model parameters β is an extension of general linear model theory (WLS, to be presented later)

- Inference on the **variance-covariance-correlation** model is also such an extension

- Basic approaches to estimation of γ:

  - maximum likelihood for joint estimation of β and γ

  - restricted maximum likelihood for estimation of γ

  - Software:
    - SAS: proc mixed
    - Stata: xtmixed
    - R: lme4 package, etc.

- Conceptually, the easiest thing to do is **maximum likelihood**:

  – Recall, the **general linear model** for longitudinal data is

  $$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  where

  $$\boldsymbol{\epsilon} \sim MVN(0, V(\boldsymbol{\gamma}))$$

  the log-likelihood is

  $$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\frac{1}{2}N\log(2\pi) - \frac{1}{2}\log|V| - \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})$$

  – This contains (mean model) parameters $\boldsymbol{\beta}$ and **covariance model** parameters $\boldsymbol{\gamma}$

  – Maximize $L(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with respect to both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} \to$ **joint estimation** of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$

  – This is what SAS `proc mixed method=ml` does (default is ReML)

- Maximum likelihood (ML) allows LRTs and Wald-type confidence intervals

- However ML is not the best approach to estimation of γ; ReML is a better approach

ReML for Covariance Parameters

- Recall: When we estimate $\sigma^2 = \text{var}(Y)$ for i.i.d. Y

  – Maximum likelihood estimator is divided by N and is biased:

  $$\hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N}$$

  – Note: if we had known $\mu = \text{E}(Y)$ instead of replacing μ by $\bar{y}$, then MLE would have been unbiased

  – We need to use $\hat{\sigma}^2_{unbiased} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$ to obtain unbiased estimator

- Recall: When we do OLS regression for independent observations,

  $$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

  where

  $$\boldsymbol{\epsilon} \sim MVN(0, \sigma^2 I_N),$$

7

– MLE for σ^2 is divided by N and is biased:

$$\hat{\sigma}^2_{MLE} = \frac{(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})}{N},$$

we can show that

$$\mathrm{E}(\hat{\sigma}^2_{MLE}) = \frac{N - p}{N}\sigma^2$$

where $p = \mathsf{rank}(X) = $ number of coefficients in $\boldsymbol{\beta}$ (if X is full rank)

– Bias arises because MLE has not taken into account that $\boldsymbol{\beta}$ is also estimated, and replace $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ in estimation of σ^2

– Need to divide by $N - p$ to obtain unbiased estimator

• ReML – **Restricted (or Residual/Reduced) Maximum Likelihood** estimation extends this idea to correct bias by ML

– Does not include $\boldsymbol{\beta}$ in likelihood

- Here is how ReML works:

  - Recall that $\boldsymbol{Y}$ is the entire vector of responses, and that

  $$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  $$\mathrm{E}(\boldsymbol{Y}) = X\boldsymbol{\beta}$$

  $$\boldsymbol{\epsilon} \sim MVN(0, V),$$

  which leads to the log-likelihood

  $$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\frac{1}{2}\log|V| - \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})$$

  - Suppose that we had a matrix K such that $K'X = 0$
    Hence

  $$K'\boldsymbol{Y} = 0 + K'\boldsymbol{\epsilon} \sim MVN(0, K'VK).$$

  - Lets use the linear transformed response $K'\boldsymbol{Y}$ and do maximum

9

likelihood; our new (ReML) log-likelihood is

$$L^*(\boldsymbol{\gamma}) = -\frac{1}{2}\log|K'VK| - \frac{1}{2}\boldsymbol{y}'K(K'VK)^{-1}K'\boldsymbol{y}$$

which does not contain $\boldsymbol{\beta}$

– ReML makes inferences about $\boldsymbol{\gamma}$ with $L^*(\boldsymbol{\gamma})$ and does not involve estimating $\boldsymbol{\beta}$ at all

• One example for such K':

$$K' = A' = I - X(X'X)^{-1}X'$$

Then

$$
\begin{aligned}
A'X &= \{I - X(X'X)^{-1}X'\}X \\
&= X - X(X'X)^{-1}X'X \\
&= 0
\end{aligned}
$$

- **Problem**: $A' = I - X(X'X)^{-1}X'$ is not full rank:
  - A' is $N \times N$, $\text{rank}(A) = N - p$
  - This will lead to singular multivariate Gaussian distribution

$$A'\boldsymbol{Y} \sim MVN(0, A'VA)$$

$$\Rightarrow |A'VA| = 0$$

  - Solution:
    * Use $K' = B'$, where $BB' = A'$, $B'B = I_{N-p}$
      (decomposition of idempotent and symmetric A)
    * Then B is full rank with $\text{rank}(B) = N - p$, and

$$B'X = B'BB'X = B'A'X = 0$$

- **Note**: It actually does not matter what you use for K, as long as

  - $\mathrm{rank}(K) = N - p$ and

  - $\mathrm{E}(K'\boldsymbol{Y}) = 0$

- More details for ReML derivation in Diggle, Heagerty, Liang and Zeger's book Sec 4.5

- Special case:

  When we have independent data with $V = \sigma^2 I_N$, here we show that ReML estimator with $K' = B'$ will be the unbiased estimator for σ^2:

  $$
  \begin{aligned}
  L^*(\sigma^2) &= -\frac{1}{2}\log|\sigma^2 B'B| - \frac{1}{2}\boldsymbol{y}'B(\sigma^2 B'B)^{-1}B'\boldsymbol{y} \\
  &= -\frac{1}{2}\log\sigma^{2(N-p)}|B'B| - \frac{1}{2\sigma^2}\boldsymbol{y}'B(B'B)^{-1}B'\boldsymbol{y} \\
  &= -\frac{1}{2}\log\sigma^{2(N-p)}|I| - \frac{1}{2\sigma^2}\boldsymbol{y}'A'\boldsymbol{y}
  \end{aligned}
  $$

  use the property of $A'\boldsymbol{y} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ and A is symmetric and idempotent, we can solve $\frac{\partial\{L^*(\sigma^2)\}}{\partial\sigma^2} = 0$ to obtain

  $$
  \hat{\sigma}^2_{ReML} = \frac{(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})}{N - p}
  $$

13

- Solving for ML/ReML:

  - Closed analytical solutions are only available in very special cases

  - A wide variety of iterative techniques for solving ML/REML equations have been proposed based on various modifications of two basic approaches (not covered here)
    * Newton-Raphson algorithm
    * EM algorithm

- **Why ReML is a great idea:**
  - Some evidence that ReML is more robust to outliers
  - it is **still maximum likelihood**, just not with the same likelihood
    * all of the **statistical theory** backing up maximum likelihood applies to ReML as well
    * you can obtain LRTs, Wald tests, and Wald CI's for the covariance parameters, just as with ML
  - Now we do not know
  $$\mathrm{E}(Y) = X\boldsymbol{\beta}$$
  without also estimating β, we **do know** that
  $$\mathrm{E}(K'\boldsymbol{Y}) = 0$$
  regardless of what the true value of $\boldsymbol{\beta}$ is (a benefit of ReML)!
  - ReML is more robust to small samples or complicated mean models than is ML

15

- Note: **ReML still depends on the specified mean model**, because K depends on covariates X

  – REML can only be compared across models with same mean model (eg, LRT)

- In particular, here is the practical implication:

  – suppose you had a particular **mean model** in mind, e.g.:

  $$\mathrm{E}(\mathtt{arm}_{ij}) = \beta_0 + \beta_1 \mathtt{wt}_{ij} + \beta_2 \mathtt{age}_{ij} + \beta_3 \mathtt{sex}_i$$

  – you want to make sure to estimate the variance / covariance / correlation model correctly

  – you remember that the correlation model depends on the mean model

  – i.e., the target of estimation and inference is a covariance model for **residual variance** in $\boldsymbol{Y}_i$ after removing effects of age, weight, sex

  – so you would like to be thorough about removing effects of weight, age and sex

– you therefore choose a more **flexible** mean model for **purposes of estimating the covariance / correlation**

e.g., you use:

$$\mathrm{E}(\texttt{arm}_{ij}) = \beta_0 + \beta_1\texttt{wt}_{ij} + \beta_2\texttt{age}_{ij} + \beta_3\texttt{sex}_i + \beta_4\texttt{age}_{ij} * \texttt{wt}_{ij} + \beta_5\texttt{age}_{ij}^2$$

to obtain covariance / correlation parameter estimates $\widehat{\gamma}_{\mathrm{ReML}}$

– you then **fix** the covariance / correlation model and parameter estimates at $\widehat{\gamma}_{\mathrm{ReML}}$ and do more careful modelling of the mean

• Such a protocol is analogous to estimating σ^2 with the "biggest" model in OLS, and then keeping the same $\widehat{\sigma}^2$ for tests, model reduction, and model evaluation

• SAS `proc mixed` have many options for variance-covariance-correlation models using ReML

Fitting Variance-Covariance-Correlation Models
SAS Examples

- **Example:** Nepalese children: arm circumference data

- **Important points:**

  - Inferences on the **covariance model** are for a **given mean model**

  - Different mean models may produce different covariance model estimates, tests and interpretations

  - In general, in the absence of sample size constraints, it is best to use the most / more flexible mean models

- For now, lets use the following mean model:

$$\mathrm{E}(\mathtt{arm}_{ij}) = \beta_0 + \beta_1 \mathtt{wt}_{ij} + \beta_2 \mathtt{age}_{ij} + \beta_3 \mathtt{sex}_i + \beta_4 \mathtt{wt}_{ij} \mathtt{age}_{ij} + \beta_5 \mathtt{age}_{ij}^2$$

even though it may be richer than necessary

Exchangeable Correlation Model

- Here is how we would set this up in SAS `proc mixed`

```
data temp;
    set nepal;
    wtage = wt*(age-36);
    age2=(age-36)*(age-36);
    age4=age/4;
run;

proc mixed data=temp dfbw  ;
    title2 "1st run" ;
    class id;
    model arm = wt age sex wtage age2 / s ;
    repeated / subject=id type=cs r;
    run;
```

- Here:

  - the `class` statement identifies categorical variables

  - the `model` statement is where you state the **mean model**

- the `repeated` statement tells SAS that observations with the same value of `id` (which must be given in the class statement) are repeated measures on the same subject

- the `type=cs` option tells SAS to fit a **compound symmetry** correlation model
  * recall: compound symmetry $=$ exchangeable
  * Equivalent to replace `repeated / subject=id type=cs r;`
    by `random intercept/ subject=id;`

- the `/ s` option says to print the estimated β coefficients

- the `/ r` option says to print out the fitted covariance matrix for the first subject (i.e., V_1) (R_i matrix in SAS is our V_i)

- `dfbw` selects an option on how to compute the degrees-of-freedom for β inferences

• Some output:

<div align="center">

The Mixed Procedure

</div>

Model Information

Data Set	WORK.TEMP
Dependent Variable	arm
Covariance Structure	Compound Symmetry
Subject Effect	id
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Class Level Information

Class	Levels	Values
id	197	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

<snip>

197 198 199 200

Dimensions

Covariance Parameters	2

```
           Columns in X                        6
           Columns in Z                        0
           Subjects                          197
           Max Obs Per Subject                 5
       Number of Observations Read          1000
       Number of Observations Used           877
       Number of Observations Not Used       123
```

Estimated R Matrix for id 1

Row	Col1	Col2	Col3	Col4
1	0.4541	0.3308	0.3308	0.3308
2	0.3308	0.4541	0.3308	0.3308
3	0.3308	0.3308	0.4541	0.3308
4	0.3308	0.3308	0.3308	0.4541

```
                 Covariance Parameter Estimates

            Cov Parm        Subject       Estimate

            CS                 id            0.3308
            Residual                         0.1233


                        Fit Statistics

       -2 Res Log Likelihood              1199.0


                  Solution for Fixed Effects


                             Standard
       Effect       Estimate      Error      DF     t Value     Pr > |t|

       Intercept      5.1785      0.4368     195      11.86      <.0001
       wt             0.7023      0.02324    676      30.23      <.0001
       age            0.01703     0.009782   676       1.74      0.0821
       sex            0.2847      0.08615    195       3.30      0.0011
       wtage         -0.00781     0.000929   676      -8.41      <.0001
       age2           0.001207    0.000145   676       8.32      <.0001
```

```
                     Type 3 Tests of Fixed Effects

                       Num        Den
           Effect       DF         DF      F Value      Pr > F

           wt            1        676       913.61      <.0001
           age           1        676         3.03      0.0821
           sex           1        195        10.92      0.0011
           wtage         1        676        70.67      <.0001
           age2          1        676        69.30      <.0001
```

- Interpretation of $\hat{\beta}_1 = 0.7023$ is the same as in an ordinary linear regression model:

  – difference in mean response for each unit difference in weight (weight slope) among persons with age of 36 (because age is centralized at 36)

- Recall, in exchangeable correlation model:

$$\epsilon_{ij} = U_i + Z_{ij}$$

- The variance parameter estimates are

$$\text{between-subject variance} = \widehat{\text{var}}(U_i) = \hat{\nu}^2 = 0.3308 \quad \text{and}$$

$$\text{within-subject variance} = \widehat{\text{var}}(Z_{ij}) = \hat{\tau}^2 = 0.1233$$

- The estimated **total variance** (for ϵ_{ij}) is therefore

$$\text{total variance} = \hat{\sigma}^2 = \hat{\nu}^2 + \hat{\tau}^2 = .3308 + .1233 = 0.4541$$

and the **within-subject** correlation coefficient is

$$\hat{\rho} = \frac{.3308}{.3308 + .1233} = 0.728$$

- Printout of R (which is our V_1 here) is helpful for interpretation

- Note: In Stata, we can use `xtmixed` to obtain the same results

  `xtmixed arm wt age sex wtage age2 || id:, nocons`
  `residual(exc) var reml nolog`

  - `|| id:  nocons` specifies the random effect portion. If we omit `nocons` option, Stata will add a random intercept.

  - `residual(exc)` specifies residual correlation structure to be exchangeable

  - `var` option request Stata to report variances rather than standard deviations

- Another way to fit this model is to request a **random intercept**. The two models should be equivalent.

- SAS code:

```
proc mixed data=temp dfbw  ;
    title2 "2nd run" ;
    class id;
    model arm = wt age sex wtage age2 / s ;
    random intercept / subject=id g;
    run;
```

   – the `random intercept` statement requests that there be a random intercept (U_i) at each subject defined by `id`.

   – the `/ g` option requests printout of $\widehat{\text{var}}(U_i)$ (what SAS calls the G matrix)

- Relevant output (all very similar to what was obtained with the `repeated ...type=cs` statement)

27

```
               Estimated G Matrix

     Row      Effect        group(id)         Col1

      1      Intercept         1             0.3308

            Covariance Parameter Estimates

        Cov Parm       Subject     Estimate

        Intercept       id          0.3308
        Residual                    0.1233

                  Fit Statistics

     -2 Res Log Likelihood            1199.0
```

- Note: In Stata, we can run the same model using
  `xtmixed arm wt age sex wtage age2 || id:,`
  `residual(ind) var reml nolog`

28

Exponential Correlation Model in SAS

- `proc mixed` is extremely flexible

- To fit the exponential correlation model, recall that (in this example),

$$\rho_{jk} = \alpha^{|t_j - t_k|}$$

  where $t_j, j = 1, \ldots, 5$ indicates **age** in units of four-months.

- The data for the first subject are:

	id	obsno	arm	age	sex
1.	1	1	14.3	41	1
2.	1	2	13.5	45	1
3.	1	3	14.5	49	1
4.	1	4	14.1	53	1
5.	1	5	.	57	1

- We need a variable that captures age in "time units" of four months (age/4)

• Then, SAS needs to know what that **time variable** is in order to fit the exponential correlation model

• Recalling that we created `age4` earlier in a data step, we specify:
```
proc mixed data=temp dfbw   ;
     title2 "3rd run" ;
     class id;
     model arm = wt age sex wtage age2 / s ;
     repeated / subject=id type=sp(pow)(age4) r;
     run;
```

• the option `type=sp(pow)(age4)` says to use the power correlation model with `age4` as the time variable
  – This option does not require time to be in equal units

• An advantage to using SAS `proc mixed` (unlike some options in Stata) is that the time correlation structure allows for unequally-spaced times

- Output

<div align="center">

The Mixed Procedure

Model Information

</div>

```
Data Set                      WORK.TEMP
Dependent Variable            arm
Covariance Structure          Spatial Power
```
<snip>

<div align="center">

Estimated R Matrix for id 1

</div>

Row	Col1	Col2	Col3	Col4
1	0.4606	0.3497	0.2655	0.2016
2	0.3497	0.4606	0.3497	0.2655
3	0.2655	0.3497	0.4606	0.3497
4	0.2016	0.2655	0.3497	0.4606

<div align="center">

Covariance Parameter Estimates

</div>

Cov Parm	Subject	Estimate
SP(POW)	id	0.7592
Residual		0.4606

Fit Statistics

-2 Res Log Likelihood	1279.0

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	4.2463	0.5062	195	8.39	<.0001
wt	0.7136	0.02387	676	29.90	<.0001
age	0.03918	0.01124	676	3.49	0.0005
sex	0.2673	0.07971	195	3.35	0.0010
wtage	-0.01027	0.001102	676	-9.32	<.0001
age2	0.001626	0.000183	676	8.87	<.0001

- Recall, exponential correlation: $\epsilon_{ij} = W_{ij}$, and

$$W_{i1} \sim N(0, \delta^2)$$

$$W_{ij} = \alpha W_{i,j-1} + E_{ij}, \quad j = 2, 3, \ldots$$

where $E_{ij} \sim N\{(0, \delta^2(1 - \alpha^2)\}$

- The total variance estimate (of ϵ_{ij}) is

$$\widehat{\mathrm{var}}(W_{ij}) = \widehat{\sigma}^2 = \widehat{\delta}^2 = 0.4606$$

- The within-subject correlation estimate for two observations separated by 4 months is

$$\widehat{\alpha} = 0.7592$$

- Can check that in R matrix $0.3497 = 0.4606 \times 0.7592$

- Note: In Stata, we can run the same model using
  `xtmixed arm wt age sex wtage age2 || id:, nocons`
  `res(ar 1, t(age)) var reml nolog`

  - However, `t(age)` must be integer here, and thus we cannot use age4 as time variable.

  - The reported $\hat{\alpha} = .9334534$ in Stata is for two observations separated by 1 month, and $\hat{\alpha}^4 = .7592$ is for 4-month interval.

- There are many other correlation functions (as a function of lag u) available in `proc mixed`, for example

  – Gaussian:
  $$\rho(u) = \alpha^{u^2} \quad (0 \le \alpha < 1)$$

  – Linear:
  $$\rho(u) = \begin{cases} 1 - \alpha|u|, & \alpha|u| \le 1 \\ 0, & \text{o.w.} \end{cases}$$

  making this a very flexible modeling environment

35

Unstructured Variance-Covariance Model in SAS

- Suppose that we did not want to make any assumptions about the variance or covariance or correlation of the repeated observations on an individual

- Then, we might fit an **unstructured** variance-covariance model

- To do this, we need a **finite number** of unique time points and relatively **good balance** among them, because:

  - variance of each time point will be estimated separately

  - correlation between each pair of two time points will be estimated separately

36

- We can illustrate this with the Nepal data, using the order of the observations as the time variable (Why wouldn't age work?)

- Note: With missing data in the middle of a series, this creates a problem because, if observation 3 is missing, then observation 4 becomes 3 and observation 5 becomes 4. So, missing data should be minimal

- Here is the SAS code, which **assumes the data are already sorted by time**:

```
proc mixed data=temp dfbw   ;
    title2 "4th run" ;
    class id;
    model arm = wt age sex wtage age2 / s ;
    repeated / subject=id type=un r rcorr;
    run;
```

  - The option `type=un` says to use an unstructured v-c-c model
  - the `r` and `rcorr` options request the first subject's estimated V_1 matrix in variance-covariance and in correlation form

37

- And some output

```
                    The Mixed Procedure

                    Model Information

      Data Set                       WORK.TEMP
      Dependent Variable             arm
      Covariance Structure           Unstructured
<snip>
                      Dimensions

      Covariance Parameters                 15
      Columns in X                           6
 <snip>
              Estimated R Matrix for id 1

        Row        Col1        Col2        Col3        Col4

         1       0.4920      0.3387      0.3104      0.3347
         2       0.3387      0.4622      0.3182      0.3538
         3       0.3104      0.3182      0.3958      0.3223
         4       0.3347      0.3538      0.3223      0.4115
```

Estimated R Correlation Matrix for id 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.7104	0.7033	0.7440
2	0.7104	1.0000	0.7439	0.8113
3	0.7033	0.7439	1.0000	0.7985
4	0.7440	0.8113	0.7985	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	id	0.4920
UN(2,1)	id	0.3387
UN(2,2)	id	0.4622
UN(3,1)	id	0.3104
UN(3,2)	id	0.3182

```
UN(3,3)        id            0.3958

UN(4,1)        id            0.3347
UN(4,2)        id            0.3538
UN(4,3)        id            0.3223
UN(4,4)        id            0.4115

UN(5,1)        id            0.3631
UN(5,2)        id            0.3837
UN(5,3)        id            0.3324
UN(5,4)        id            0.3892
UN(5,5)        id            0.5837
```

Fit Statistics

```
-2 Res Log Likelihood              1152.5
```

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5.0128	0.4494	195	11.15	<.0001
wt	0.7234	0.02319	195	31.20	<.0001
age	0.01535	0.009957	195	1.54	0.1249
sex	0.2802	0.08557	195	3.27	0.0013
wtage	−0.00795	0.000954	195	−8.33	<.0001
age2	0.001216	0.000150	195	8.11	<.0001

• Note: In Stata, we can use following code to fit unstructured model

```
. xtmixed arm wt age sex wtage age2 || id:, nocons res(uns, t(obs))
    var reml
```

Expanded Covariance Pattern Models

Exponential plus Measurement Error
Variance-Covariance Model

- We have examined some simple **covariance pattern** models.
  Hybrid covariance pattern models can be obtained by combining
  or expanding upon these

- The **exponential plus measurement error** variance-covariance
  model is given by:

$$\epsilon_{ij} = W_{ij} + Z_{ij}$$

  where:

  - W_{ij}'s follow an **exponential correlation** model with
    autocorrelation parameter α

$$W_{ij} = \alpha W_{i,j-1} + E_{ij} \ ,$$

$$\text{var}(W_{ij}) = \delta^2$$

– Z_{ij}s are all independent of each other and of the W_{ij}'s and

$$\text{var}(Z_{ij}) = \tau^2$$

- In this model:

  – W_{ij}'s represent smooth, unobserved time-varying processes contributing to the response variable (Y_{ij}).
  – Smooth in the sense that W_{ij} and W_{ik} very close together in time are nearly perfectly correlated.

  – Z_{ij} is random noise (**measurement error**) specific to each observation (time point)

- This gives rise to the total variance (for ϵ_{ij}):

$$\text{var}(\epsilon_{ij}) = \sigma^2 = \delta^2 + \tau^2$$

and correlation model

$$
\begin{aligned}
\mathrm{corr}(\epsilon_{ij}, \epsilon_{ik}) &= \frac{\mathrm{cov}(W_{ij} + Z_{ij}, W_{ik} + Z_{ik})}{\mathrm{var}(\epsilon_{ij})} \\
&= \frac{\mathrm{cov}(W_{ij}, W_{ik})}{\mathrm{var}(\epsilon_{ij})} \\
&= \frac{\mathrm{corr}(W_{ij}, W_{ik})\mathrm{var}(W_{ij})}{\mathrm{var}(\epsilon_{ij})} \\
&= \alpha^{|t_{ij} - t_{ik}|} \frac{\delta^2}{\delta^2 + \tau^2}
\end{aligned}
$$

- **At home:** check correlation, write the variance-covariance matrix
  Note how the measurement error puts extra τ^2 on the diagonal of V_i

- **Note:** Even as time lag tends to zero 0, $\mathrm{corr}(\epsilon_{ij}, \epsilon_{ik})$ will not tend to 1

- The extra noise Z_{ij} due to measurement error is sometimes called a "nugget" or a "local" error because it is "local" in time

- To fit this model in SAS, we would specify `proc mixed` just as for the exponential correlation model, with the following change:

```
repeated / subject=id type=sp(pow)(age4) local r;
```

  - where `local` requests the measurement error (nugget) term

Exchangeable plus Exponential Variance-Covariance Model

- This model is given by

$$\epsilon_{ij} = U_i + W_{ij}$$

  where:

  - U_i is constant for all observations on a subject, $\mathrm{var}(U_i) = \nu^2$

  - W_{ij}s follow an exponential correlation model with autocorrelation parameter α and $\mathrm{var}(W_{ij}) = \delta^2$, i.e.,

$$W_{ij} = \alpha W_{i,j-1} + E_{ij}$$

  - U_i independent of W_{ij}'s

- U_i represents time-invariant factors specific to subject i that affect the outcome Y_{ij}

- **At home:** Compute the correlation $\mathrm{corr}(\epsilon_{ij}, \epsilon_{ik})$ for this model

  Note how the exchangeable model puts extra ν^2 in every cell of V_i

- How is this model different from the "exponential plus measurement error" correlation model?

- This model is fitted in SAS by casting U_i as a **random intercept**

- Returning to the Nepal example, the `proc mixed` code is

```
proc mixed data=temp dfbw covtest ;
    title2 "5th run" ;
    class id;
    model arm = wt age sex wtage age2 / s ;
    random intercept / subject=id g;
    repeated / subject=id type=sp(pow)(age4) r;
    run;
```

  – `random` statement specified variables to be included as predictors

47

with coefficients that vary from subject to subject.

– The most common such variable is the **intercept**

– May also include random slope (coming later)

– `repeated` statement specifies the v-c-c model for the residual process not accounted for by the random terms

• Result

```
                        Model Information


        Data Set                    WORK.TEMP
        Dependent Variable          arm
        Covariance Structures       Variance Components,
                                    Spatial Power
        Subject Effects             id, id
        Estimation Method           REML
        Residual Variance Method    Profile
        Fixed Effects SE Method     Model-Based
        Degrees of Freedom Method   Between-Within
  <snip>
                          Dimensions
```

48

```
              Covariance Parameters              3
              Columns in X                       6
              Columns in Z Per Subject           1
              Subjects                         197
              Max Obs Per Subject                5
        Number of Observations Read           1000
         Number of Observations Used            877
         Number of Observations Not Used        123
```

Estimated R Matrix for id 1

Row	Col1	Col2	Col3	Col4
1	0.1316	0.01743	0.002309	0.000306
2	0.01743	0.1316	0.01743	0.002309
3	0.002309	0.01743	0.1316	0.01743
4	0.000306	0.002309	0.01743	0.1316

```
                      Estimated G Matrix

          Row     Effect         group(id)         Col1

           1     Intercept          1            0.3243

                Covariance Parameter Estimates


                                   Standard        Z
Cov Parm        Subject    Estimate    Error    Value      Pr Z

Intercept       id          0.3243   0.03979     8.15    <.0001
SP(POW)         id          0.1324   0.06196     2.14    0.0326
Residual                    0.1316   0.009083   14.49    <.0001

                       Fit Statistics

        -2 Res Log Likelihood          1193.9
```

```
                  Solution for Fixed Effects


                           Standard
    Effect        Estimate     Error      DF    t Value    Pr > |t|

    Intercept       5.0550    0.4486      195     11.27     <.0001
    wt              0.7069    0.02347     676     30.12     <.0001
    age             0.01905   0.01006     676      1.89     0.0587
    sex             0.2825    0.08634     195      3.27     0.0013
    wtage          -0.00809   0.000958    676     -8.44     <.0001
    age2            0.001258  0.000151    676      8.36     <.0001
```

• The estimated random effects variance is

$$\widehat{\text{var}}(U_i) = \hat{\nu}^2 = 0.3243$$

• The estimated autoregressive variance is

$$\widehat{\text{var}}(W_{ij}) = \hat{\delta}^2 = 0.1316$$

- The within-subject correlation for two observations separated by 4 months is

$$\widehat{\text{corr}}(W_{ij}, W_{i,j-1}) = \widehat{\alpha} = 0.1324$$

- The total variance estimate is therefore

$$\widehat{\sigma}^2 = \widehat{\text{var}}(\epsilon_{ij}) = \widehat{\text{var}}(U_i) + \widehat{\text{var}}(W_{ij}) = 0.3243 + 0.1316 = 0.4559$$

- This will also allow us to **test for positive autocorrelation** in the exponential part of the model, having removed the correlation due to subject-to-subject variability (U_i)
  – Next we will see how such tests are constructed

- Note: In Stata, we can use following code to fit this hybrid model

```
xtmixed arm wt age sex wtage age2 || id:, res(ar 1, t(age))
    var reml nolog
```

52

- **Final note on such hybrid v-c-c models:** Empirical features of correlation in longitudinal data are reflected in these models:

  - correlation decreases with time lag separating observations: captured by the exponential correlation part of the model

  - even observations close together in time do not have correlation one: captured by the measurement error term Z_{ij}

  - even observations far apart in time are positively correlated: captured by including a random intercept term U_i for each subject

ReML Likelihood-based Tests and CIs for Covariance Parameters

- **Example:** Nepalese children: Suppose we wanted to compare the "exponential plus exchangeable" model to the "exchangeable only" covariance model

  I.e., in the "exponential plus exchangeable" model, we wish to test

  $$H_0 : \alpha = 0 \quad \text{vs.} \quad H_A : \alpha \neq 0$$

- We generally make such comparisons under one of the more flexible mean models

  Here, we continue with the following mean model:

  $$\mathrm{E}(\mathtt{arm}_{ij}) = \beta_0 + \beta_1 \mathtt{wt}_{ij} + \beta_2 \mathtt{age}_{ij} + \beta_3 \mathtt{sex}_i + \beta_4 \mathtt{wt}_{ij} \mathtt{age}_{ij} + \beta_5 \mathtt{age}_{ij}^2$$

- In the foregoing example, we saw the SAS `proc` `mixed` code for "exchangeable plus exponential" model.

- For simple exchangeable only model, we would remove the line:
  ```
  repeated / subject=id type=sp(pow)(age4) r;
  ```
  and rerun the model

- Best statistical test: Likelihood ratio test (LRT) for nested models (actually, ReLRT for residual likelihood ratio test)

  - Recall, in LRT: $-2\{ln(L_{reduce}) - ln(L)\} \sim \chi^2_{DF}$

    - where L is likelihood of with all the parameters unrestricted
    - L_{reduce} is likelihood of reduced model
    - DF $=$ number of parameters lost in reduced model

- In this example,

  - Exponential plus exchangeable:

    ```
                        Fit Statistics


              -2 Res Log Likelihood          1193.9
    ```
    $-2ln(L)$ under random intercept, plus exponential correlation model is 1193.9

  - Exchangeable model (from earlier):
    ```
                      -2 Res Log Likelihood          1199.0
    ```
    $-2ln(L_{reduce})$ under random intercept only model is 1199.0

– Difference on 1 DF (1 parameter less in exchangeable model): $\chi^2 = 5.1$, which gives

$$P\text{-value} = \Pr(\chi_1^2 \geq 5.1) \approx 0.02$$

– Conclusion: The exponential autocorrelation α among the residuals ϵ_{ij} on the same subject is significantly greater than zero, even after accounting for a subject-level random intercept

• **Careful!** When using any software package, make sure that the log-likelihood is multiplied by 2! Otherwise, you must multiply the difference by 2 to get the χ^2 value

• **Another caution:** Do not use LRTs to compare mean models when using ReML to estimate the model.

– If must use LRT (and not $\mathcal{F}$-tests) to compare mean models: Estimate model with ML

57

• Alternative test: Wald Test

  – When the model is estimated, the score equations are solved:

$$\frac{\partial L^*(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial(\boldsymbol{\beta}, \boldsymbol{\gamma})} = 0 \quad \text{yields} \quad (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$$

  – A by-product is the **observed information**:

$$\mathcal{I}_O = -\frac{\partial^2 L^*(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial(\boldsymbol{\beta}, \boldsymbol{\gamma})\partial(\boldsymbol{\beta}, \boldsymbol{\gamma})}\Big|_{(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})}$$

  – The inverse of $\mathcal{I}_o$ is an estimate of $\mathrm{var}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$, which can be used to get standard errors for components of the variance parameter $\boldsymbol{\gamma}$

  – Tests based on these standard errors are called Wald tests

  – These are obtained with the covtest option in proc mixed

- **Example:** In the "exponential plus exchangeable" model, we may want to test whether $\alpha = 0$:

  - We fit the model and obtain $\widehat{\alpha}$

  - Then
    $$\widehat{\mathrm{se}}(\widehat{\alpha}) = \sqrt{(\mathcal{I}_o^{-1})_{\alpha,\alpha}}$$

  - Then a Z-test can be constructed:
    $$Z = \frac{\widehat{\alpha}}{\widehat{\mathrm{se}}(\widehat{\alpha})}$$

    and squared to obtain χ^2

• This produces:

<pre>
                   Covariance Parameter Estimates

                                     Standard   Z
   Cov Parm   Subject   Estimate   Error   Value   Pr Z

   Intercept   id   0.3243   0.03979   8.15   <.0001
   SP(POW)   id   0.1324   0.06196   2.14   0.0326
   Residual   0.1316 0.009083   14.49   <.0001
</pre>

• Note that $\chi^2 = 2.14^2 = 4.58$

  – not that much different to 5.1 in LRT

• This can also be used for **approximate** confidence intervals for covariance parameters:

A 95% CI for α is

$$\widehat{\alpha} \pm 1.96 \times \widehat{\text{se}}(\widehat{\alpha})$$

60

which is

$$.1324 \pm 1.96 * .06196 = [0.011, 0.254]$$

• **Note**:

– These tests are **approximate**. They require adequate sample sizes

– SAS PROC MIXED uses observed information $\mathcal{I}_O$

– Other packages may use **expected information**:

$$\mathcal{I}_E = -\mathrm{E}\left\{\frac{\partial^2 L^*(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial(\boldsymbol{\beta}, \boldsymbol{\gamma})\partial(\boldsymbol{\beta}, \boldsymbol{\gamma})}\right\}\big|_{(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})},$$

– Usually LRTs are more powerful than Wald tests, but more work to produce

Problem: Some Covariance Parameter Tests are Naturally One-sided

• **Example:** Exchangeable correlation model

$$\text{corr}(\epsilon_{ij}, \epsilon_{ik}) = \rho$$

which can be motivated by the random effects model

$$\epsilon_{ij} = U_i + Z_{ij}$$

Obviously $\text{var}(U_i) = \nu^2 \geq 0$ and so

$$\rho = \frac{\text{var}(U_i)}{\text{var}(U_i) + \text{var}(Z_{ij})} \geq 0$$

• Note: More generally, you could allow $\rho < 0$ (but not due to RE model)

- Under the **random effects model**, a significance test for the exchangeable correlation ρ would be for a **one-sided** alternative:

$$H_0 : \rho = 0 \quad \text{vs.} \quad H_A : \rho > 0$$

or

$$H_0 : \nu^2 = 0 \quad \text{vs.} \quad H_A : \nu^2 > 0$$

- **Example:** Exponential plus exchangeable:

$$\epsilon_{ij} = U_i + W_{ij}$$

where

$$W_{ij} = \alpha W_{i,j-1} + Z_{ij}$$

(if time lags are all in 1-unit intervals)

Here, if time is **continuous**, $\mathrm{corr}\{W_i(t), W_i(t+u)\} = \alpha^u$ we cannot have $\alpha < 0$, so we would again want to test

$$H_0 : \alpha = 0 \quad \text{vs.} \quad H_A : \alpha > 0$$

64

- **Problem:** LRT chi-square tests are set up for two-sided alternatives:

$$H_0 : \alpha = 0 \quad \text{vs.} \quad H_A : \alpha \neq 0$$

  - The null hypothesis value of the parameter is on the **boundary** of the parameter space (see Self and Liang, 1987)
  - The theoretical distribution of the LRT chi-square statistic X^2 comes from the square of a Z
  - Negative Z would correspond to $\widehat{\alpha} < 0$, positive Z to $\widehat{\alpha} > 0$

- As we cannot have $\widehat{\alpha} < 0$, under the null hypothesis, X^2 will only be greater than some critical value **half as often** as we expect

- **Solution**: you can (and should!) **divide your** P-value**s by 2** for one-sided covariance parameter hypothesis tests

- In the LRT for the exponential correlation α, we would then have

$$P\text{-value} = \frac{1}{2}\Pr\left(\chi^2 \geq 5.1\right) \approx \frac{1}{2}0.02 = 0.01$$

- Note: This does not change computation of Wald-based confidence intervals (bad CI!).

Selecting and Evaluating Covariance Models

• How does one choose among a set of candidate covariance models? How does one evaluate whether they fit the data well?

• If using maximum likelihood or restricted maximum likelihood estimation:

  – Perform LRTs for comparing **nested** models
    Example: Nepal data comparing the exponential plus exchangeable to the exchangeable only models

  – Akaike's Information Criterion (AIC), Bayesian information criterion (BIC) (or other information criteria).
    – Will not discuss here. With either AIC or BIC, if two models, non-nested, have the same number of parameters, can just compare the likelihood values

- Examine empirical ACF (auto-correlation function) to help select candidate models
  - must assume that $\mathrm{var}(\epsilon_{ij})$ is approximately constant and that $\mathrm{corr}(\epsilon_{ij}, \epsilon_{ij})$ depends only on the lag $|t_{ij} - t_{ik}|$, at least approximately
- A more powerful technique is to use empirical and fitted model variograms.
  - not covered here, interested students can read Diggle, Heagerty, Liang and Zeger's book

Examining and Interpreting ACF

- Considering the acf:

  - If acf declines strongly with lag, then: "exponential" component

  - If acf does not appear to $\to 0$ as lag $\to \infty$, then: "exchangeable" component

  - If acf does not appear to $\to 1$ as lag $\to 0$, then: "measurement error" component

- **Example:** Nepal data: Generate ACF in an **exploratory mode**, using OLS to generate residuals:

```
# read data
data=read.csv("nepal.csv")
data$obs=rep(1:5,200)

## Compute residuals, removing effects by OLS
data$wtage=data$wt*(data$age-36)
data$age2=(data$age-36)^2
fit<-lm(arm~wt+age+sex+wtage+age2,data)
```

```
data$armrs=rep(NA,dim(data)[1])
data$armrs[!(is.na(data$arm) | is.na(data$wt) |
    is.na(data$age) | is.na(data$sex))]=residuals(fit)

## Autocorrelation function and correlogram

# extract all pairs within each subject
data$id=as.factor(data$id)
id.list=levels(data$id)
data.pairs=list()
for(i in 1:length(id.list))
{
    subject.i=data[data$id==id.list[i],]
    data.pairs.i=gtools::combinations(dim(subject.i)[1], 2,repeats=FALSE)
        #index of all possible pairs

    data.pairs.add=data.frame(id=rep(i,dim(data.pairs.i)[1]),
        obs1=subject.i$obs[data.pairs.i[,1]],obs2=subject.i$obs[data.pairs.i[,2]],
        armrs1=subject.i$armrs[data.pairs.i[,1]],
        armrs2=subject.i$armrs[data.pairs.i[,2]])
        # all pairs within subject i
```

```
    data.pairs=rbind(data.pairs,data.pairs.add)
}
data.pairs$lag=abs(data.pairs$obs2-data.pairs$obs1)

# remove missing observations
# otherwise N in TL calculation will include missing observations
data.pairs=data.pairs[!(is.na(data.pairs$armrs1) | is.na(data.pairs$armrs2)),]

# Autocorrelation function and tolerance limit
lag.list=1:4
ACF=TL=rep(NA,length(lag.list))
for(lag in 1:length(lag.list))
{
    subset=data.pairs[which(data.pairs$lag==lag.list[lag]),4:5]
    ACF[lag]=cor(subset,use="pairwise.complete.obs")[1,2]
    TL[lag]=1.96/sqrt(dim(subset)[1])
}

# correlogram
plot(lag.list,ACF,"l",xlab="Time lag",xlim=c(1,4),ylim=c(0,1))
lines(lag.list,TL,lty=2)
legend("topright",c("Autocorrelation function","95% tolerance limit"),lty=1:2)
```

Autocorrelation function of armres
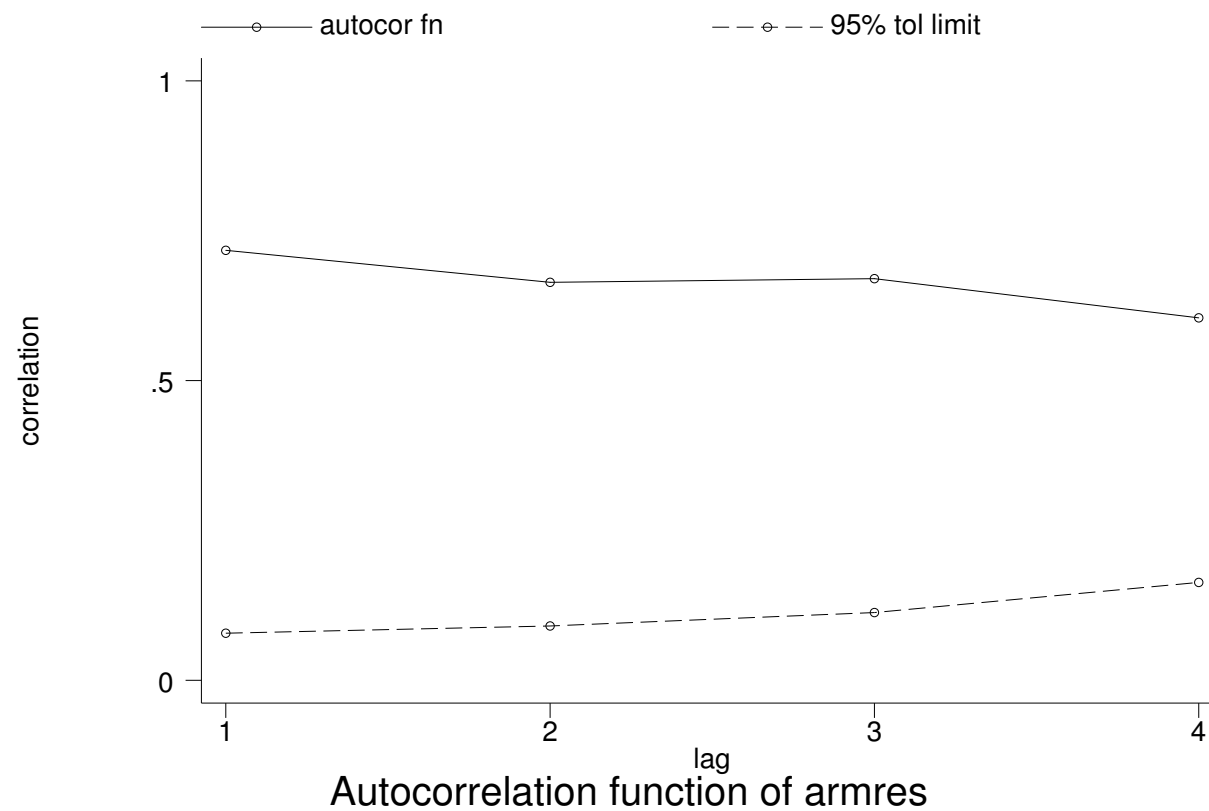
From this:

– clearly, there is an "exchangeable" component in the model

– "exponential" component might not be that important (even
  though it was significant in the test)