LINEAR MODELS FOR LONGITUDINAL DATA

Sensitivity to Covariance / Correlation Model

- The linear model contains a model for the mean

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

and a model for the variance:

$$\mathrm{var}(\boldsymbol{Y}|X) = V \quad \text{or} \quad \mathrm{var}(\boldsymbol{\epsilon}) = V$$

- These lead via maximum likelihood (or WLS) for $\boldsymbol{\beta}$ to:

$$\widehat{\boldsymbol{\beta}} = (X'\widehat{V}^{-1}X)^{-1}X'\widehat{V}^{-1}\boldsymbol{y}$$

and

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (X'\widehat{V}^{-1}X)^{-1}$$

- These facts depend on being able to estimate V, which in turn depends on a correct model for V

- What happens if we get the model for V wrong? E.g., here are some possible incorrect assumptions about the v-c-c model:
  - We assume exchangeable correlation and it is really exponential plus exchangeable correlation model
  - We assume homoscedasticity across time and really the variance increases with time
  - We assume independence and the data are really correlated
- This means that we plug estimates $\widehat{V}^*$ of $V^* \neq V$ into the expressions for $\widehat{\boldsymbol{\beta}}$ and $\mathrm{var}(\widehat{\boldsymbol{\beta}})$ above, where:
  - $V =$ the **true value** of the v-c-c- matrix
  - $V^* =$ is **incorrectly-specified** v-c-c model
  - $\widehat{V}^* =$ the ReML **estimate** of the incorrectly-specified model
- Recall, letting $W = V^{*\,-1}$ be a "weight" matrix, we have the more general WLS estimator:

$$\widehat{\boldsymbol{\beta}}_W = (X'WX)^{-1}X'W\boldsymbol{y}$$

- Does $\widehat{\boldsymbol{\beta}}_W$ work?
  - It is unbiased:

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}_W|X) = \beta$$

  - It has variance

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}_W|X) = (X'WX)^{-1}\{X'WVWX\}(X'WX)^{-1}$$

    (of course, $\mathrm{var}(\widehat{\boldsymbol{\beta}}_W|X)$ is not computable because we do not know V)
  - If V^* turns out to be right (i.e., $V^* = V$ and $W = V^{-1}$) then

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}_W|X) = (X'V^{-1}X)^{-1}$$

- Loss of efficiency with incorrectly-specified V^*:

  Of all possible values of W, the one that yields the smallest variance for $\widehat{\beta}_W$ is the one with $W = V^{-1}$ (ie, BLUE)

Estimation of $\mathrm{var}(\widehat{\boldsymbol{\beta}}_W|X)$

- Recall,

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}_W|X) = (X'WX)^{-1}X'WVWX(X'WX)^{-1}$$

- Here, $W = V^{*\,-1}$, where V^* is the v-c-c model that we use to fit the model:
  - V^* is called the **working correlation** or **working v-c-c model**
  - Estimating the v-c-c parameters via ML/ReML/any other technique, using V^* **as if** it were the correct variance, we obtain:

$$\widehat{V}^* \quad \text{and} \quad \widehat{W} = \widehat{V}^{*\,-1}$$

  which in turn leads to WLS estimate

$$\widehat{\boldsymbol{\beta}}_W = (X'\widehat{W}X)^{-1}X'\widehat{W}\boldsymbol{y}$$

- How to estimate

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}_W|X) = (X'WX)^{-1}X'WVWX(X'WX)^{-1}$$

  - Estimate W: $\widehat{W} = \widehat{V}^{*\,-1}$
  - But how to estimate true V?
  - From the **block-diagonal** form of V^* and V, we can write

$$X'WVWX = \sum_i X_i'W_iV_iW_iX_i$$

  (i sums over subjects)

  - also note that

$$V_i = \mathrm{var}(\boldsymbol{\epsilon}_i) = \mathrm{E}(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i')$$

– therefore

$$\mathrm{E}\{X_i'W_i(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i')W_iX_i\} = X_i'W_i\mathrm{E}(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i')W_iX_i = X_i'W_iV_iW_iX_i$$

suggesting that we estimate $X'WVWX$ by replacing V_i with $(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i')$

– where do we get $\boldsymbol{\epsilon}_i$?

$$\boldsymbol{\epsilon}_i = \boldsymbol{y}_i - X_i\boldsymbol{\beta}$$

so we can estimate $\boldsymbol{\epsilon}_i$ by using the estimated $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - X_i\widehat{\boldsymbol{\beta}}_W$$

– finally, we can put it all together to estimate $\mathrm{var}(\widehat{\boldsymbol{\beta}}_W|X)$:

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_W|X) = (X'\widehat{W}X)^{-1}\left\{\sum_i X_i'\widehat{W}_i(\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{\epsilon}}_i')\widehat{W}_iX_i\right\}(X'\widehat{W}X)^{-1}$$

- This estimator is due to Huber (1967) and White (1980) (**Huber-White estimator**) and was also used by Liang and Zeger (1986) in their development of **generalized estimating equations**
- Also called the **sandwich** estimator
- or the **robust** or **empirical** variance estimator

  (term "empirical" is a reminder that the true V_i gets replaced by **data** $\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{\epsilon}}_i'$)
- in order for the sandwich estimator to work well, we will need a fairly large number of subjects (m). how many depends on context.

- **Example:** Protein content of cows' milk. Consider the mean model

  $$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 \texttt{mixed}_i + \beta_2 \texttt{barley}_i + \beta_3 \texttt{mixed}_i \times \texttt{week}_{ij} + \beta_4 \texttt{barley}_i \times \texttt{week}_{ij} + \alpha_j$$

  where α_j is a factor for time j, $j = 1, \ldots, 18$ weeks

- Suppose we fit this model naively using the **independence** correlation structure:

  Here is the code:

```
data temp;
set cows;
mixed = 1*(diet eq "mixed");
barley = 1*(diet eq "barley");
mixedwk = 1*(diet eq "mixed")*(week-1);
barlewk = 1*(diet eq "barley")*(week-1);
run;
```

```
proc mixed data=temp dfbw ;
        class id week;
        model prot = mixed barley mixedwk barlewk week / s;
        repeated / subject=id ;
        run;
```

(no `type=` is specified in the `repeated` statement $\longrightarrow$ an independence working correlation model with constant variance)

And here is part of the output:

Effect	Week since calving	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		3.2291	0.05261	76	61.38	<.0001
mixed		0.04391	0.03699	76	1.19	0.2388
barley		0.1172	0.03777	76	3.10	0.0027
mixedwk		0.009032	0.003830	1238	2.36	0.0185
barlewk		0.01254	0.003900	1238	3.21	0.0013
week	1	0.5529	0.06975	1238	7.93	<.0001
week	2	0.2454	0.06858	1238	3.58	0.0004

<snip>

- These standard errors are called **model-based** because they assume the specified variance V^* is correct (i.e., $V^* = V$)

- Note: In Stata, you can use following code to do the same analysis:

```
xtmixed prot mixed barley mixedwk barleywk i.week, ||
id:, noco reml
```

or

```
xtgee prot mixed barley mixedwk barleywk i.week,
c(ind)
```

- Now, we **know** that the standard errors are wrong, because the independence correlation model is wrong

- So, we refit the model using the **empirical variance estimator**:

```
proc mixed data=temp dfbw empirical ;
    ...
```

and obtain:

```
             Week
             since                   Standard
 Effect      calving   Estimate         Error     DF    t Value    Pr > |t|

 Intercept               3.2291       0.06510     76      49.60      <.0001
 mixed                  0.04391       0.07162     76       0.61      0.5416
 barley                  0.1172       0.07431     76       1.58      0.1190
 mixedwk               0.009032      0.006200   1238       1.46      0.1454
 barlewk                0.01254      0.006466   1238       1.94      0.0527
 week         1          0.5529       0.09630   1238       5.74      <.0001
 week         2          0.2454       0.08390   1238       2.92      0.0035
<snip>
```

- Parameter estimates ($\widehat{\beta}$'s) are **exactly** the same, but the standard errors are now "fixed up" to account for the possibility that the correlation model is wrong
- These standard errors are called **robust** or **empirical**

- Note: In Stata, you can use `xtgee` with option `robust` to obtain the robust variance estimator:

```
xtgee prot mixed barley mixedwk barleywk i.week,
c(ind) robust
```

- Now suppose we refit the model using the **exponential plus measurement error** correlation model, **making some attempt to get the v-c-c model correct**. We obtain

```
proc mixed data=temp dfbw ;
class id week;
model prot = mixed barley mixedwk barlewk week / s;
repeated / subject=id type=sp(pow)(week) local ;
run;
```

Effect	Week since calving	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		3.1492	0.06653	76	47.33	<.0001
mixed		0.07251	0.07452	76	0.97	0.3336
barley		0.1308	0.07600	76	1.72	0.0892
mixedwk		0.004693	0.006602	1238	0.71	0.4773
barlewk		0.009928	0.006720	1238	1.48	0.1398
week	1	0.6187	0.08884	1238	6.96	<.0001
week	2	0.3119	0.08585	1238	3.63	0.0003

<snip>

The estimates ($\widehat{\beta}$'s) are now different

– because the correlation model (hence our W matrix) is different

- Now, redo it using the empirical variance estimate **to protect ourselves in case the v-c-c model is wrong**:

```
proc mixed data=temp dfbw empirical ;
     class id week;
     model prot = mixed barley week / s;
     repeated / subject=id type=sp(pow)(week) local ;
     run;
```

and obtain

```
              Week
              since                     Standard
 Effect       calving   Estimate           Error      DF   t Value   Pr > |t|

 Intercept                3.1492         0.06345      76     49.64     <.0001
 mixed                   0.07251         0.07566      76      0.96     0.3409
 barley                   0.1308         0.07789      76      1.68     0.0971
 mixedwk                0.004693        0.006705    1238      0.70     0.4841
 barlewk                0.009928        0.007315    1238      1.36     0.1750
 week          1          0.6187          0.1000    1238      6.19     <.0001
 week          2          0.3119         0.08570    1238      3.64     0.0003
<snip>
```

Notes:

- Again, the parameter estimates are the same, but the standard errors are different
- Now the model-based and the empirical standard errors are much closer than under the independence model.
- This is a good sign: We made some attempt to get the correlation model correct, but we protect ourselves in case we get it wrong

- **Important notes:**
  - In practice, you should make some attempt to get the correlation structure approximately right. **Do not** use independence if you know it is wrong
  - You should avoid using the empirical variance estimator with less than 50 subjects. 100–200 is probably better. The cows data have only 79 subjects, so this is suspect
  - If the v-c-c model is correct, the model-based standard errors are more accurate (especially with smaller sample sizes)

    (more accurate standard errors lead to more accurate hypothesis tests and better confidence interval coveraage)

- However, if the correlation model is wrong, the model-based standard errors are biased, so the empirical standard errors are more accurate **with large sample sizes**
- Even so, for **smaller sample sizes**, if the v-c-c model is approximately correct, the model-based s.e.'s might be better than the empirical ones.
- What we have done here can be seen as examples of **GEE** estimators

Critical Concepts: True versus Estimated Standard Errors

- The **true standard error** of $\widehat{\boldsymbol{\beta}}_W$ is the standard deviation of $\widehat{\boldsymbol{\beta}}_W$ over repeated replicates of the "experiment", where "experiment" includes both the study design/data collection **and** the data analysis method (which includes specification of the working model V^*)

- A poor choice of V^* will lead to **larger true standard errors** (inefficient estimates of $\boldsymbol{\beta}$) for a given study design

- The **estimated standard error** is an estimate of the true standard error based on the data and the model

- Model-based standard errors are **estimated** standard errors assuming the v-c-c model is correct

- Robust or standard errors are **estimated** standard errors allowing for the possibility that the v-c-c model is wrong.

- A model fit with lower **estimated** standard errors does not necessarily reflect greater statistical efficiency (**true** standard errors).

- Two distinct ideas:
  - **true** standard errors reflect **(in)efficiency** due to (in)correct working model V^* for V
  - incorrect working model can lead to **biased estimates** of the true standard errors

Exploiting the Empirical Variance Estimator
Generalized Estimating Equations (GEE)

- Suppose that the **mean model** ($\boldsymbol{\beta}$) is of **scientific interest**
- And, we do not care about the **correlation model**
  - it is **not** of scientific interest
  - the correlation model is a **nuisance**
- We need correlation model to get **efficient inferences** on $\boldsymbol{\beta}$
- So we use a **working** correlation model:
  - admittedly an approximation, but yields valid inferences
  - simpler correlation structures to choose from, for example:
    - exchangeable
    - exponential or AR(1)
    - independence
  - but not:

  - exchangeable plus exponential
  - exponential plus measurement error

- We also use **simple, crude, estimators** for correlation parameters (not ML or ReML). We will discuss further when we do models for categorical data, which is where GEE is much more useful
  - Note: These are available in Stata's `xtgee`: simple estimators available for exchangeable (`corr(exch)`), for exponential (`corr(ar 1) force`), and for unstructured (`corr(uns)`) correlation models

- Using the estimated working correlation model, obtain weights $\widehat{W}_i = \widehat{V}_i^{*\,-1}$ for each subject and obtain WLS estimator $\widehat{\boldsymbol{\beta}}_W$
(The W could stand for "weights" or for "working"!)

- Then, because correlation model is "working" **and** because it may not be as well estimated:
  - use **sandwich estimator** to obtain $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_W)$ that is **robust** to misspecification or poor estimation of the correlation model
  - construct Z-tests, χ^2–tests, CI's for $\boldsymbol{\beta}$ using the robust variance estimate
- Note:

  If can get the correlation model approximately correct, can do almost as well in estimating $\boldsymbol{\beta}$ with GEE as getting the correlation model exactly correct and estimating it using ML or ReML
- **Caveat:** we must rely on much larger sample sizes so that the sandwich estimator provides a good estimate of standard errors

SUMMARY

Where Have We Been?

- The linear model for longitudinal data

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

  contains a model for the mean

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

  and a model for the variance (covariance / correlation model):

  $$\mathrm{var}(\boldsymbol{Y}|X) = \mathrm{var}(\boldsymbol{\epsilon}) = V$$

- The mean model parameters $\boldsymbol{\beta}$ have a **population-average** interpretation

  (i.e., the same interpretation as in OLS)

- One general model for the covariance is

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

where

$$\mathrm{var}(U_i) = \nu^2 \ , \quad \mathrm{var}(W_{ij}) = \delta^2 \ \text{ and } \ \mathrm{var}(Z_{ij}) = \tau^2$$

and where the W_{ij}'s are autocorrelated with

$$\mathrm{corr}(W_{ij}, W_{ik}) = \text{function}(u), \quad u = \text{lag} = |t_{ij} - t_{ik}|$$

We considered $\mathrm{corr}(W_{ij}, W_{ik}) = \alpha^u$, but others such as $\mathrm{corr}(W_{ij}, W_{ik}) = \alpha^{u^2}$ are possible

- Estimation and inferences for the covariance model parameter $\boldsymbol{\gamma}$:
  - ML: jointly estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$
    fine if mean model does not have a lot of parameters or if N large

- ReML: only have to estimate $\boldsymbol{\gamma}$

  – more robust γ-inferences

  – more flexible mean model

  good when covariance / correlation model is the focus of the study

- Model tests and evaluation: autocorrelation function, LRT, AIC, BIC; careful with one-sided hypothesis tests

- Importantly: Covariance is among the **residuals** from a **given mean model**

  different mean model $\longrightarrow$ different covariance model

• Inference in the mean model under a **given** correlation model:

- WLS estimation

- $\mathcal{F}$-tests, Wald tests and Wald CI's

- Valid even if normality of ϵ_{ij} does not hold

  – test statistics follow the same asymptotic distribution

- Inference in mean model when correlation model is (may be) wrong:
  - do ML or ReML for v-c-c- parameters $\boldsymbol{\gamma}$
  - $\widehat{\boldsymbol{\beta}}$ still unbiased
  - use HW estimator to fix-up variance
- Inference in mean model under **working** correlation model (GEE):
  - just assume that correlation model wrong (working correlation)
  - crude estimates of correlation model parameters (it is wrong, so who cares?)
  - use HW estimator to fix-up variance
  - We will talk more about GEE later for categorical data