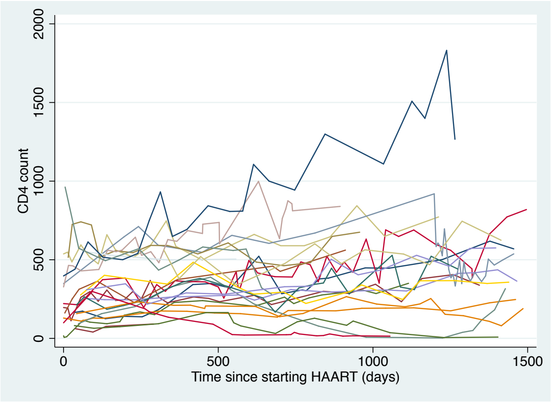
Longitudinal Data
Fall 2014

Naïve Analysis of Longitudinal Data

(Major Themes)



Instructors

Nick Jewell (jewell@berkeley.edu)

GSI

Robin Mejia (mejia@nasw.org)

Major Themes

1. Dependent data – impact on estimation and inference. Why can’t I just use the tools I already know? MT1

2. Using longitudinal history in regression.
   – Avoiding, by default, treating the data like cross-sectional data. What am I missing by using the tools I already know? MT2

3. Estimating contribution of variance from different units. Can I better understand where my variation comes from? MT3

4. Efficiency. Can I make my estimators more precise? MT4

1. Dependent data

- If one takes more than one measurement on a subject $\longrightarrow$ can no longer assume all observations are statistically independent.

- Statistical inference is easier if the data consists of independent measures.

- e.g., $var(Y_{i1}+Y_{i2})=var(Y_{i1})+var(Y_{i2})+2*cov(Y_{i1},Y_{i2})$

Treating longitudinal data like cross-sectional data: inference

- Consider a simple, random (mixed?) effects model.

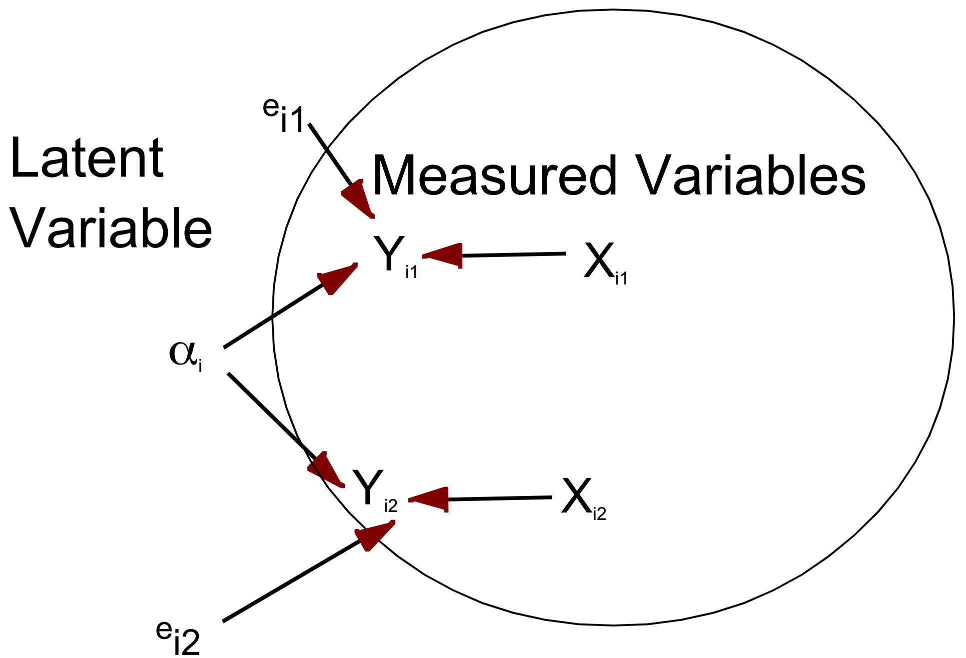
- The experiment is cd4 count measured twice on each of m randomly selected individuals.

- Model is, for the jth measurement on individual i,

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

where, $E(\alpha_i)=0$, $E(e_{ij})=0$, α_i indep. of e_{ij} and e_{i1} independent of e_{i2}.

A simple random effects model for correlation



Consequences of Ignoring Correlation

- σ_α^2 = variance between individuals (variance of α_i). *inter-individual*

- σ_e^2 = variance within an individual (variance of e_{ij}). *intra-individual*

- The correlation between measurements within an individual is:

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}$$

Correlation induced by repeated measures, cont.

- Estimate the mean as:

$$\bar{Y} = \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{2} Y_{ij}$$

- Naively estimate the variance (simple sample variance) of the average (ignoring correlation) as:

$$\hat{\text{var}}(\bar{Y}) = \frac{s^2}{N} = \frac{\left[\frac{1}{N-1}\right]\sum_{i=1}^{m}\sum_{j=1}^{2}(Y_{ij} - \bar{Y})^2}{N}$$

cont.

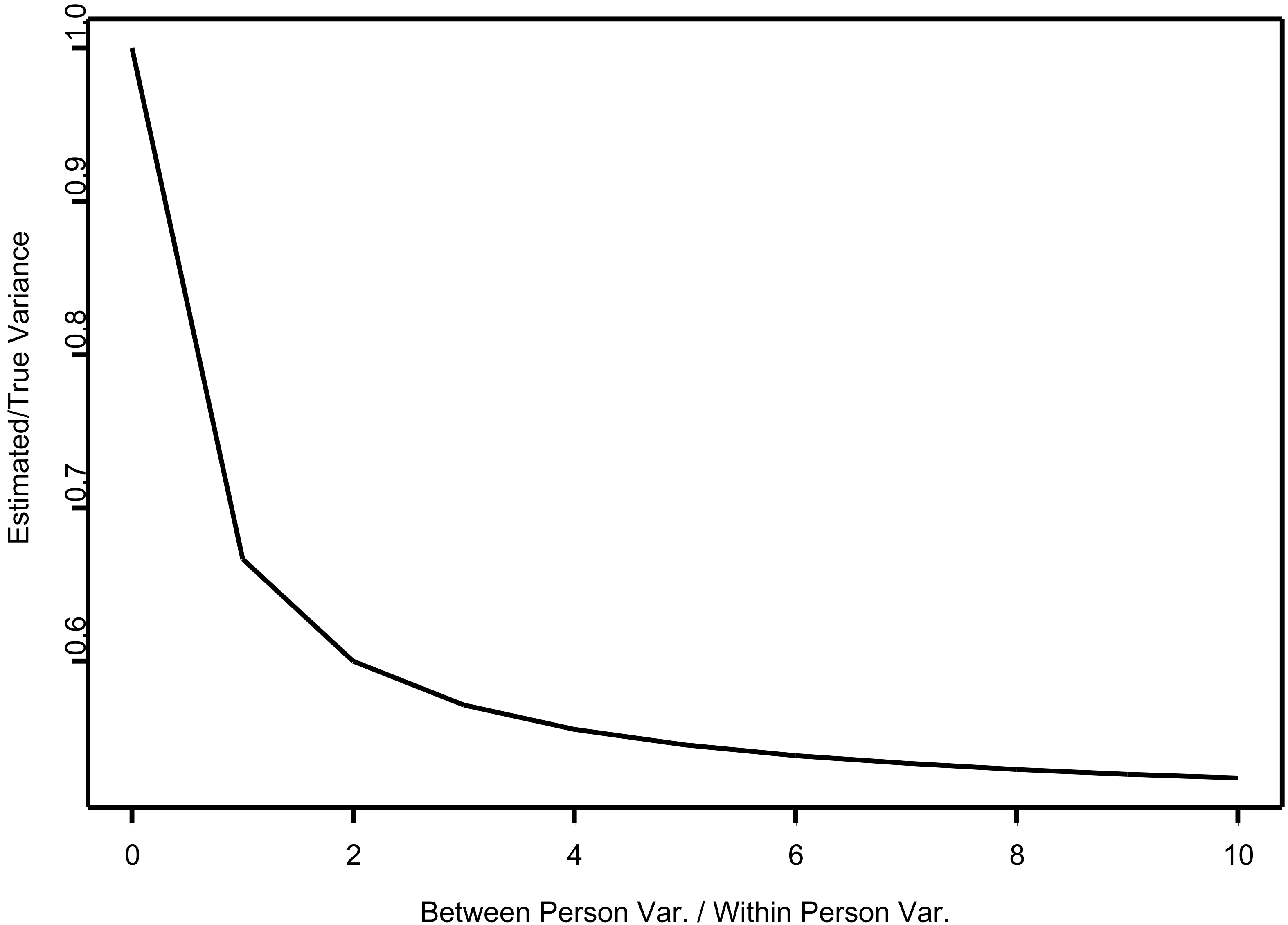
- Expected value of this variance estimate is:

$$\frac{\sigma_{\alpha}^{2} + \sigma_{e}^{2}}{2m}$$

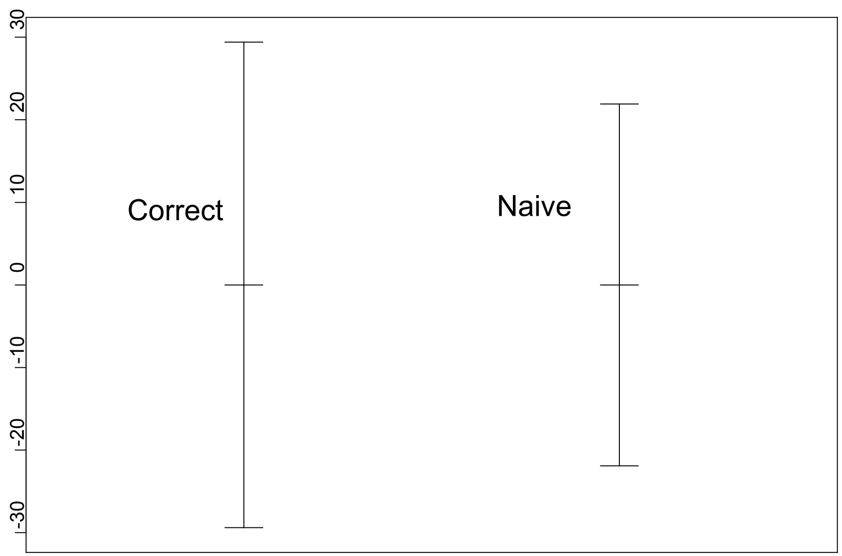
- However, because of the correlation induced by repeated measurements on the same individual, the true variance of the sample average is:

$$\frac{2\sigma_{\alpha}^{2} + \sigma_{e}^{2}}{2m}$$

Ignoring Correlation and Inference on Mean



95% CI resulting from correct and naïve estimates of variance



How does this apply to Regression? Dental Data

Potthoff & Roy (*Biometrika*, 1964):

distance (mm) from the center of the pituitary gland to the pteryomaxillary fissure

Changes in pituitary-pteryomaxillary distances during growth is important in orthodontal therapy.

The goals of the study were to describe the distance in boys & girls as simple functions of age, and then to compare the functions for boys (n = 16) & girls (n = 11).

obsno	child	age	distance	gender
1	1	8	21	0
2	1	10	20	0
3	1	12	21.5	0
4	1	14	23	0
5	2	8	21	0
6	2	10	21.5	0
7	2	12	24	0
8	2	14	25.5	0
9	3	8	20.5	0
10	3	10	24	0
11	3	12	24.5	0
12	3	14	26	0
13	4	8	23.5	0
14	4	10	24.5	0
15	4	12	25	0
16	4	14	26.5	0
17	5	8	21.5	0
18	5	10	23	0
19	5	12	22.5	0
20	5	14	23.5	0
21	6	8	20	0

Dental Data

Want to model the effect of gender and age on dental distance.

- Y_{ij} distance of the jth measurement on the ith child,
- X_{ij1} is the age of ith child, jth measurement,
- X_{ij2} is the gender of ith child, jth measurement (1 is male, 0 is female) - same for all *j* measurements.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij1} X_{ij2} + e_{ij}, \;\; E(e_{ij} \mid \mathbf{X}_{ij}) = 0$$

Ordinary Least Squares

- Mechanically:
  - Data in long format
  - Implement linear regression without regard to dependence of repeated measures.

- Can be motivated to two ways:
  - As a maximum likelihood estimate under a specific data-generating model (albeit wrong here).
  - As an estimating equation for a parametric model of a specific parameter of interest, $E(Y_{ij}|\boldsymbol{X}_{ij})$—least squares.

$$E[Y_{ij} \mid X_{ij1,} X_{ij2}] = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij1} X_{ij2}$$

MLE Justification

- Assume $e_{ij} \sim N(0,\sigma^2)$, i.i.d.

- Likelihood (*Φ* standard normal density) for each observation is:

$$L(\vec{\beta} \mid Y_{ij}, \vec{X}_{ij}) = \phi\left(\frac{Y_{ij} - \{\beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij1} X_{ij2}\}}{\sigma}\right)$$

- For all obs, log-likelihood is:

$$l(\vec{\beta} \mid \vec{Y}, \mathbf{X}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log\left\{\phi\left(\frac{Y_{ij} - \{\beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij1} X_{ij2}\}}{\sigma}\right)\right\}$$

MLE Justification, cont.

- MLE is: $\hat{\beta} = \underset{\vec{\beta}}{\operatorname{argmax}}\, l(\vec{\beta} \mid \vec{Y}, \mathbf{X})$
- Solution is the $\vec{\beta}$ that minimizes the squared residuals, or:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \{\beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij1} X_{ij2}\} \right)^2 \text{ or,}$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, \hat{\mathrm{E}}(Y - \hat{Y}(\beta))^2 = \underset{\beta}{\operatorname{argmin}}\, M\hat{S}E(\beta)$$

- This is, of course, just OLS.

Estimating Equation Motivation:
start by combining all observations into matrices

- Lump the responses of all units into one big vector $\boldsymbol{Y}=(\boldsymbol{Y}_1,\ldots\boldsymbol{Y}_m)^T$ which is an N-vector (total number of observations):

$$N = \sum_{i=1}^{m} n_i$$

- Remember can write model in matrix/vector form:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_p X_{ijp} + e_{ij} \\ &= \mathbf{X}_{ij}^T \boldsymbol{\beta} + e_{ij} \end{aligned}$$

Combining, cont.

- Model for the ith person as

$$\underset{n_i x 1}{\mathbf{Y}_i} = \underset{n_i x(p+1)}{X_i} \underset{(p+1)x1}{\boldsymbol{\beta}} + \underset{n_i x1}{\mathbf{e}_i}$$

- and for the entire data as:

$$\underset{Nx1}{\mathbf{Y}} = \underset{N x(p+1)}{X} \underset{(p+1)x1}{\boldsymbol{\beta}} + \underset{N x1}{\mathbf{e}}$$

The Variance-Covariance of $\mathbf{Y}_i$ - Most General

$$V_i = \begin{bmatrix} v_{i11} & v_{i12} & v_{i13} & v_{i14} \ldots & v_{i1n_i} \\ v_{i12} & v_{i22} & v_{i23} & v_{i24} \ldots & v_{i2n_i} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ v_{in_i1} & v_{in_i2} & v_{in_i3} & \ldots. & v_{in_in_i} \end{bmatrix}$$

The Variance-Covariance of *Y*

$$V(\mathbf{Y}) = \begin{bmatrix} V_1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & V_2 & 0 & 0 & \ldots & 0 \\ 0 & 0 & V_3 & 0 & \ldots & 0 \\ . & . & . & & & . \\ . & . & . & & & . \\ . & . & . & & & . \\ 0 & 0 & 0 & 0 & \ldots & V_m \end{bmatrix}$$

How Estimating Equation Approach Works

- MLE approach specifies the whole distribution of $Y|X$ (in this case normal)

- However, can one derive estimator just specifying the parts of the distribution one cares about, e.g., $E(Y_{ij}|\boldsymbol{X}_{ij})$?

- Maybe something about the Variance-covariance model. Weighted Least Squares?

Example 1: Compound Symmetry

- Assume that the variance of each observation, Y_{ij}, is $var(Y_{ij}) = \sigma^2$
- $cor(Y_{ij}, Y_{ij'}) = \rho$ for $j \neq j'$.
- As always, $cor(Y_{ij}, Y_{i'j'}) = 0$ for $i \neq i'$.

Example 1: V(Y) - Compound Symmetry

	Y_{11}	Y_{12}	Y_{13}	Y_{21}	Y_{22}	Y_{23}	Y_{31}	Y_{32}	Y_{33}
Y_{11}	σ^2	$\rho\sigma^2$	$\rho\sigma^2$	0	0	0	0	0	0
Y_{12}	$\rho\sigma^2$	σ^2	$\rho\sigma^2$	0	0	0	0	0	0
Y_{13}	$\rho\sigma^2$	$\rho\sigma^2$	σ^2	0	0	0	0	0	0
Y_{21}	0	0	0	σ^2 ρ	$\rho\sigma^2$	$\rho\sigma^2$	0	0	0
Y_{22}	0	0	0	$\rho\sigma^2$	σ^2	$\rho\sigma^2$	0	0	0
Y_{23}	0	0	0	$\rho\sigma^2$	$\rho\sigma^2$	σ^2	0	0	0
Y_{31}	0	0	0	0	0	0	σ^2	$\rho\sigma^2$	$\rho\sigma^2$
Y_{32}	0	0	0	0	0	0	$\rho\sigma^2$	σ^2	$\rho\sigma^2$
Y	0	0	0	0	0	0	$\rho\sigma^2$	$\rho\sigma^2$	σ^2

In this case, weighted least squares reduces to OLS (but what if ρ varied with how far apart in time the measurements in *Y* were taken?*)*

Can OLS be justified by Estimating Equation Approach?

- The OLS solution is (in matrix form):

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$$

- Does the fact that the data may be correlated ($E[e_{ij}e_{ik}] \neq 0$) make this an erroneous estimate?

- First question to ask is if this estimate is biased, i.e., $E\hat{\beta} \neq \beta$

Expected values of coefficient estimates

$$E(\hat{\beta} \mid X) = E\left\{(X^T X)^{-1} X^T \mathbf{Y}\right\} = (X^T X)^{-1} X^T E(\mathbf{Y} \mid X)$$

- Note that *E(**Y**|**X**)* is the expectation of a vector or simply the list of expectations of the components of the vector.

$$E(Y_{ij} \mid \mathbf{X}_{ij}) = E\left\{\beta_0 + \beta_1 X_{ij1} + \ldots + \beta_p X_{ijp} + e_{ij}\right\} =$$

$$= \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_p X_{ijp} + E\{e_{ij} \mid \mathbf{X}_{ij}\}$$

$$= \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_p X_{ijp} = \mathbf{X}_{ij}' \beta$$

Expectation of Vectors

- Now looking at vector of observations made on an individual.

- $E(\boldsymbol{Y}_i|\boldsymbol{X}_i) = (\boldsymbol{X}_{i1}{}^T\beta,\ldots, X_{in_i}{}^T\beta)=\boldsymbol{X}_i\beta$.

- Finally, look at the vectors stacked together.

- $E(\boldsymbol{Y}|\boldsymbol{X}) = (\boldsymbol{X}_1\beta, \boldsymbol{X}_2\beta,\ldots, \boldsymbol{X}_m\beta) = \underset{Nx(p+1)}{X}\ \underset{p+1}{\beta}$

  •

The Answer

$$E(\hat{\beta} \mid X) = (X^T X)^{-1} X^T E(\mathbf{Y} \mid X) = (X^T X)^{-1} X^T X \beta = \beta \qquad (*)$$

- That is, the estimator is unbiased, regardless of the correlation among measurements made on the same person, *i*.

- From a simple estimation view (at least with regard to bias), ignoring correlation in the linear model might be OK.

- It's when inference (standard error calculations) are made for $\hat{\beta} \rightarrow V\hat{a}r(\hat{\beta})$

* Assuming model is right, otherwise, more complicated.

Dental Data: No accounting for Correlation after OLS

```
. gen inter = age*gender

. regress distance age gender inter

      Source |       SS       df       MS              Number of obs =     108
-------------+------------------------------           F(  3,   104) =   25.39
       Model |  387.935027     3  129.311676           Prob > F      =  0.0000
    Residual |  529.757102   104  5.09381829           R-squared     =  0.4227
-------------+------------------------------           Adj R-squared =  0.4061
       Total |   917.69213   107  8.57656196           Root MSE      =  2.2569

------------------------------------------------------------------------------
    distance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4795455   .1521635     3.15   0.002     .1777996    .7812913
      gender |  -1.032102   2.218797    -0.47   0.643     -5.43206    3.367855
       inter |   .3048295   .1976661     1.54   0.126    -.0871498    .6968089
       _cons |   17.37273   1.708031    10.17   0.000     13.98564    20.75982
------------------------------------------------------------------------------
```

Are these Standard Errors , p-values and CIs ok?

Assuming nothing, what is $Var(\hat{\beta})$?

$$\mathrm{var}(\hat{\beta} \mid X) = \mathrm{var}\left\{(X^T X)^{-1} X^T \mathbf{Y}\right\} =$$

$$(X^T X)^{-1} X^T \mathrm{var}\left\{\mathbf{Y} \mid X\right\} X (X^T X)^{-1} =$$

$$(X^T X)^{-1} X^T V X (X^T X)^{-1}$$

- So, if one knows *V*, then the variance of the estimated coefficients is known.

- However, one never knows *V* so you have to estimate *V*.

- Estimating *V* will require some assumption about how the Y_{ij}' s are correlated (the form of the V_i' s).

$Var(\hat{\beta})$ cont.

$$V(\vec{Y}) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \ldots & 0 \\ 0 & \sigma^2 & 0 & 0 \ldots & 0 \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 0 & 0 & 0 & 0 \ldots & \sigma^2 \end{bmatrix}$$

- Inference returned by standard OLS assumes, the observations are assumed to be independent and *homoscedastic:* V_i is assumed to look as above.

OLS estimate of $Var(\hat{\beta})$

- To estimate σ^2, which is now the variance of the Y_{ij}'s. Note that:

$$\mathrm{var}(\mathrm{Y}_{\mathrm{ij}} \mid \mathbf{X}_{ij}) = \mathrm{var}\left\{e_{ij} \mid \mathbf{X}_{ij}\right\} = \sigma^2$$

- We don't know the e_{ij}'s, but we can estimate them from the residuals:

$$r_{ij} = Y_{ij} - \left\{\hat{\beta}_0 + \hat{\beta}_1 X_{ij1} + \ldots + \hat{\beta}_p X_{ijp}\right\}$$

OLS estimate of $Var(\hat{\beta})$

- To get a estimate of σ^2, get an average (sort of) of the squared residuals:

$$\hat{\sigma}^2 = \frac{1}{N-p} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (r_{ij})^2$$

- Then, this is plugged back in V_i to get:

$$\hat{V}_i = \hat{\sigma}^2 I_{n_i x n_i} \text{ or } \hat{V} = \hat{\sigma}^2 I_{NxN}$$

OLS estimate of $Var(\hat{\beta})$

- Finally, plug this back into (***) to get the estimated variance of the coefficients

$$
\begin{aligned}
\hat{\mathrm{var}}(\hat{\beta} \mid X) &= (X^T X)^{-1} X^T \hat{V} X (X^T X)^{-1} \\
&= (X^T X)^{-1} X^T \hat{\sigma}^2 I X (X^T X)^{-1} \\
&= \hat{\sigma}^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\
&= \hat{\sigma}^2 (X^T X)^{-1}
\end{aligned}
$$

Naïve Approach

- This is a matrix where the diagonal elements are the estimated variances of the coefficient estimates. The SE($\hat{\beta}$) are the square-roots of these estimated variances.

Dental Data: No accounting for Correlation after OLS

```
. gen inter = age*gender

. regress distance age gender inter
```

Naïve Approach

```
      Source |       SS       df       MS              Number of obs =     108
-------------+------------------------------           F(  3,   104) =   25.39
       Model |  387.935027     3  129.311676           Prob > F      =  0.0000
    Residual |  529.757102   104  5.09381829           R-squared     =  0.4227
-------------+------------------------------           Adj R-squared =  0.4061
       Total |   917.69213   107  8.57656196           Root MSE      =  2.2569

------------------------------------------------------------------------------
    distance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4795455   .1521635     3.15   0.002     .1777996    .7812913
      gender |  -1.032102   2.218797    -0.47   0.643     -5.43206    3.367855
       inter |   .3048295   .1976661     1.54   0.126    -.0871498    .6968089
       _cons |   17.37273   1.708031    10.17   0.000     13.98564    20.75982
------------------------------------------------------------------------------
```

Are these Standard Errors , p-values and CIs ok?

Biased Inference

- Ignoring the correlation, as in the standard errors returned by OLS, can give biased estimates of $\mathrm{var}(\hat{\beta})$
- This can lead to erroneous confidence intervals and tests.
- However, we can still use OLS if we can just repair the estimates of $\mathrm{var}(\hat{\beta})$.
- That means getting unbiased estimate of *V*, *so choose bigger variance-covariance model*.

Estimating V under bigger model

- First, we will take the case that measurements are time-structured and the same for each person ($n_i=n$) - the dental example.

- For this type of data, one can assume the most general structure about V (unstructured).

- The goal is to estimate the individual components of V_i (the v_{ijk}) using the residuals of OLS.

Form of the each unit's Variance-Covariance matrix, V_i.

- In order to estimate the V, we will assume that the individual covariance matrices (the V_i) are all the same: $V_1 = V_2 = \ldots = V_m$.

- That is equivalent to saying the $v_{ijk} = v_{jk}$, or the variance and covariances of equivalent observations are the same for every individual.

- That way, we will only have to estimate one V_i of in order to build V.

- Thus, we can call each $V_1, V_2, \ldots, V_m = V_0$.

The Variance-Covariance of *Y*

$$V = \begin{bmatrix} V_0 & 0 & 0 & 0 & \dots & 0 \\ 0 & V_0 & 0 & 0 & \dots & 0 \\ 0 & 0 & V_0 & 0 & \dots & 0 \\ . & . & & & \dots & \\ . & . & . & & & . \\ . & . & . & & & . \\ 0 & 0 & 0 & 0 & \dots & V_0 \end{bmatrix}$$

Estimating the components of V_0.

- If we observed the errors, then it would be easy to estimate the v_{jk}.

$$\hat{v}_{jk} = \frac{1}{m}\sum_{i=1}^{m} e_{ij}e_{ik} \ .$$

- We don't observe the errors, but we can still estimate the covariance of any two measurements on the same person using the residuals.

$$\hat{v}_{jk} = \frac{1}{m}\sum_{i=1}^{m} r_{ij}r_{ik} \ .$$

Lessons

- The LS solution is still (theoretically) unbiased even if there is residual correlation among repeated measurements on a subject (and design is balanced).

- However, the inference is not unbiased, so must account for correlation when calculating $SE(\hat{\beta})$

- One solution for regularly measured data (e.g., all subjects have same measurements at same times) is to use residuals and estimate the variance-covariance matrix of the data assuming each subject has the same matrix.

- However, can even get more flexible (robust SE's).

Robust (no Model) SE's

- Instead of assuming that very subject has the same V_i, one can estimate each subjects V-C matrix separately.

- Thus, the estimate of the variance of the first measurement on subject i is:

$$\hat{v}_{i11} = \hat{\mathrm{var}}(Y_{i1}) = r_{i1}^2 = (Y_{i1} - \hat{Y}_{i1})^2$$

- The correlation between the 1st and 2nd measurement on the ith person is estimated as: $$\hat{v}_{i12} = \hat{\mathrm{cov}}(Y_{i1}, Y_{i2}) = r_{i1} r_{i2} = (Y_{i1} - \hat{Y}_{i1})(Y_{i2} - \hat{Y}_{i2})$$

Robust SE's, cont.

- This procedure is repeated until all the variances and covariances are estimated for each subject (again, we always assume no between subject correlation).

- These are terrible estimators of the variances and covariances.

- However, do we care? No we don't!!!

- Because, they still result, under certain assumptions, in a consistent estimator of $SE(\hat{\beta})$

- The estimate, $\hat{\text{var}}(\hat{\beta})$, sums over the, $\hat{V}$ which means we gain precision by averaging over the individual estimates of the $\hat{v}_{ijk}$

Fully Nonparametric Approach

Semi-Parametric Approach (xtgee)

$$V = \begin{bmatrix} \sigma^2 & \sigma\rho_{112} & \sigma\rho_{113} & 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma\rho_{112} & \sigma^2 & \sigma\rho_{123} & 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma\rho_{113} & \sigma\rho_{123} & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & \sigma\rho_{212} & \sigma\rho_{213} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma\rho_{212} & \sigma^2 & \sigma\rho_{223} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma\rho_{213} & \sigma\rho_{223} & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & \sigma\rho_{312} & \sigma\rho_{313} \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma\rho_{312} & \sigma^2 & \sigma\rho_{323} \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma\rho_{313} & \sigma\rho_{323} & \sigma^2 \end{bmatrix}$$

Semi-robust Approach

Dental Data: No accounting for Correlation after OLS

```
. gen inter = age*gender

. regress distance age gender inter
```

Naïve Approach

```
      Source |       SS       df       MS              Number of obs =     108
-------------+------------------------------           F(  3,   104) =   25.39
       Model |  387.935027     3  129.311676           Prob > F      =  0.0000
    Residual |  529.757102   104  5.09381829           R-squared     =  0.4227
-------------+------------------------------           Adj R-squared =  0.4061
       Total |   917.69213   107  8.57656196           Root MSE      =  2.2569

------------------------------------------------------------------------------
    distance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4795455   .1521635     3.15   0.002     .1777996    .7812913
      gender |  -1.032102   2.218797    -0.47   0.643     -5.43206    3.367855
       inter |   .3048295   .1976661     1.54   0.126    -.0871498    .6968089
       _cons |   17.37273   1.708031    10.17   0.000     13.98564    20.75982
------------------------------------------------------------------------------
```

Dental Data: Correcting for Correlation Using Semi-Robust GEE

```
. xtgee distance age gender inter, i(child) cor(ind) ro
```

Semi-Robust Approach

```
GEE population-averaged model                   Number of obs      =       108
Group variable:                      child      Number of groups   =        27
Link:                             identity      Obs per group: min =         4
Family:                           Gaussian                     avg =       4.0
Correlation:                   independent                     max =         4
                                                Wald chi2(3)       =    148.77
Scale parameter:                  4.905158      Prob > chi2        =    0.0000

Pearson chi2(108):                  529.76      Deviance           =    529.76
Dispersion (Pearson):             4.905158      Dispersion         =  4.905158

                             (standard errors adjusted for clustering on child)
------------------------------------------------------------------------------
             |             Semi-robust
    distance |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4795455   .0643352     7.45   0.000     .3534507    .6056402
      gender |  -1.032102   1.404031    -0.74   0.462    -3.783952    1.719748
       inter |   .3048295   .1190935     2.56   0.010     .0714105    .5382486
       _cons |   17.37273    .739021    23.51   0.000     15.92427    18.82118
------------------------------------------------------------------------------
```

Dental Data: Correcting for Correlation using Robust Sandwich Estimator

```
regress distance age gender inter, cluster(child)
```

Fully Nonparametric Approach

```
Linear regression                                      Number of obs =     108
                                                       F(  3,     26) =   48.20
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4227
                                                       Root MSE      =  2.2569

                                   (Std. Err. adjusted for 27 clusters in child)
------------------------------------------------------------------------------
             |               Robust
    distance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4795455   .0652565     7.35   0.000     .3454087    .6136822
      gender |  -1.032102   1.424137    -0.72   0.475    -3.959459    1.895254
       inter |   .3048295    .120799     2.52   0.018     .0565236    .5531355
       _cons |   17.37273   .7496041    23.18   0.000     15.83189    18.91356
------------------------------------------------------------------------------
```

Dental Data: Correcting for Correlation using Clustered Bootstrap

```
bootstrap, reps(1000) cluster(child): regress distance age gender inter
```

Fully Nonparametric Approach

```
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
    distance |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4795455      .06735     7.12   0.000     .3475418    .6115491
      gender |  -1.032102    1.359501    -0.76   0.448    -3.696676    1.632471
       inter |   .3048295      .11794     2.58   0.010     .0736714    .5359877
       _cons |   17.37273    .7487003    23.20   0.000      15.9053    18.84015
------------------------------------------------------------------------------
```

Dental: Resulting Fit