SUBJECT-SPECIFIC LINEAR MODELS FOR LONGITUDINAL DATA

What we have been doing:

- **Marginal/Population-average (PA)** linear model:

$$\boldsymbol{Y}_i = X_i'\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$

  - All variation in Y_{ij} not explained by $\boldsymbol{x}_{ij}$ is captured by ϵ_{ij}
  - $\boldsymbol{\epsilon}_i$ is mean-zero, uncorrelated with X_i
  - So that:

$$\mathrm{E}(Y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}$$

    and therefore $\boldsymbol{\beta}$ has a **population-average** interpretation equivalent to that in OLS models for independent data
  - Main role of longitudinal (versus cross-sectional) data in PA models:
    – increase **statistical efficiency** of inferences on $\boldsymbol{\beta}$
    – study the variance-covariance-correlation model

- **Conditional/Subject-specific** models have interpretations explicitly tailored to longitudinal data in order to answer questions such as:
  - What is the effect of time on Y_{ij} in terms of subject-specific trajectories (changes) over time?
  - What is the effect of covariates varying within subject, allowing subjects to act as their own conrol?
- General **subject-specific** linear model for longitudinal data:

$$Y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{d}'_{ij}\boldsymbol{U}_i + Z_{ij}$$

where

$$Z_{ij} \sim N(0, \tau^2)$$

and the Z_{ij}'s are independent of X_i, $\boldsymbol{d}_{ij}$, $\boldsymbol{U}_i$ and of each other

- Now, $\boldsymbol{d}_{ij}$ is also a set of observed covariates for the ith subject at the jth time (like $\boldsymbol{x}_{ij}$)
  - $\boldsymbol{d}_{ij}$ is usually a subset of what is in $\boldsymbol{x}_{ij}$
  - $\boldsymbol{d}_{ij}$ has coefficient $\boldsymbol{U}_i$, which is **subject specific** — each subject has his own personal $\boldsymbol{U}_i$
  - the $\boldsymbol{U}_i$'s **vary across subjects**
  - **Special case: random intercept** if $\boldsymbol{d}_{ij} = 1$, we have personal intercept $\boldsymbol{U}_i$ for each subject i

- Note importantly that if we just examine the ith subject,

$$\mathrm{E}(Y_{ij}|\boldsymbol{U}_i, \boldsymbol{d}_i, \boldsymbol{x}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{d}'_{ij}\boldsymbol{U}_i$$

  which is a **subject specific** model for that subject

- **Example:** Arm circumference in Nepalese children as a function of age and sex. Let $Y_{ij} = \texttt{arm}_{ij}$:

$$Y_{ij} = \beta_0 + \beta_1 \texttt{age}_{ij} + \beta_2 \texttt{sex}_i + \beta_3 \texttt{age}_{ij} \times \texttt{sex}_i + U_{i1} + U_{i2}\texttt{age}_{ij} + Z_{ij}$$

  – “common effect” (my term) covariates:

$$\boldsymbol{x}'_{ij} = (1, \texttt{age}_{ij}, \texttt{sex}_i, \texttt{age}_{ij} \times \texttt{sex}_i)$$

  – “subject-specific effect” (again, my term) covariates:

$$\boldsymbol{d}'_{ij} = (1, \texttt{age}_{ij})$$

  – personal intercept U_{i1} and personal slope U_{i2}

- This model could be rewritten in **hierarchical linear model form** as

$$Y_{ij} = (\beta_0 + \beta_2 \texttt{sex}_i + U_{i1}) + (\beta_1 + \beta_3 \texttt{sex}_i + U_{i2})\texttt{age}_{ij} + Z_{ij} \ ,$$

  (sex is a between-subject covariate)
  so that for subject i

  – the **subject-specific** intercept is $\beta_0 + \beta_2 \texttt{sex}_i + U_{i1} = b_{i0}$

  – the **subject-specific** slope with respect to age is $\beta_1 + \beta_3 \texttt{sex}_i + U_{i2} = b_{i1}$

- Therefore, the model for subject i is

$$Y_{ij} = b_{i0} + b_{i1}\texttt{age}_{ij} + Z_{ij}$$

- This is a **subject-specific** model:

- Each subject has **own slope** and **own intercept**
  – Both determined by subjects' sex and some deviation U_{i1} and U_{i2}

- A simpler model is

$$Y_{ij} = (\beta_0 + \beta_2 \texttt{sex}_i + U_{i1}) + \beta_1 \texttt{age}_{ij} + Z_{ij}$$

  which allows:

  – a **subject-specific** intercept

  – a **common slope** for all subjects

- Each subject has **same slope** with respect to age, and **different intercepts** determined by subjects' sex and U_{i1}

- Two ways to handle the U_i coefficients (different assumptions required):

  – Fixed effects (FE) model: Treat them as fixed parameters (like the β's) (**fixed effects** model)

    * One U_i parameter for each subject i

  – Random effects (RE) model: Treat them as random variables from a distribution, independent of X_i (**random effects** model)

    * Eg, $U_i \sim N(0, \nu^2)$, and U_i is independent of X_i and Z_{ij}

Example: "Random effect" Subject-specific Models

- **Example:** Arm circumference in Nepalese children

  Suppose we consider the **random intercept** model for arm circumference

$$
\begin{aligned}
Y_{ij} &= \beta_0 + \beta_1 \texttt{wt}_{ij} + \beta_2 \texttt{age}_{ij} + \beta_3 \texttt{sex}_i + U_i + Z_{ij} \\
&= (\beta_0 + \beta_3 \texttt{sex}_i + U_i) + \beta_1 \texttt{wt}_{ij} + \beta_2 \texttt{age}_{ij} + Z_{ij}
\end{aligned}
$$

  Particular interest: coefficient β_1 of weight

- Then each subject has his / her own intercept

$$b_{0i} = \beta_0 + \beta_3 \texttt{sex}_i + U_i,$$

  β_1 and β_2 are common slopes with respect to weight and age

- The model assumes that each subject in the population has the **same slope**, but that these slopes capture the **within-subject** change in the mean arm circumference for a unit change in weight

- In such a **random effects** model, we assume:
  – $U_i \sim N(0, \nu^2)$
  – U_i is **independent** of X_i (and the Z_{ij}'s) (a **key** assumption)

- (As we have seen in previous notes) the model can be estimated via ML or ReML in software:

```
proc mixed data=nepal method=ML;
class id sex;
model arm=wt age sex/ s;
random intercept/ subject=id;
run;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	id	0.4106
Residual		0.1275

Solution for Fixed Effects

Effect	sex	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		8.9626	0.1633	195	54.89	<.0001
wt		0.6630	0.02326	678	28.51	<.0001
age		-0.05949	0.003620	678	-16.43	<.0001
sex	1	-0.2920	0.09528	678	-3.06	0.0023
sex	2	0	.	.	.	.

- This model is estimated assuming:
  - random intercepts U_i are **normally-distributed** among subjects
  - U_i is **independent** of the covariates $\boldsymbol{x}_{ij}$ for that subject

Example: "Fixed effect" Subject-specific Models

- Suppose instead that we wanted to fit the **same model**, but:
  - did not want to assume random effects U_i independent of X_i
  - did not want to assume normally distributed random effects
  - did not really care about the effect of sex
- Then we can fit the same model (formally), assuming that the U_i's are not random, but rather **fixed quantities** for each subject

  This is called a **fixed effects** model in econometrics
- We will work more on **random effects** models later

"Fixed effect" Subject-specific Models

Why fit a model with the U_i's fixed? What do we gain?

- the U_i's model **heterogeneity** in level of arm circumference across subjects

- this heterogeneity arises from unobserved factors on each subject, e.g.:
  – genetic factors
  – environmental factors such as nutrition

- we would like to measure the effect of weight on arm circumference **adjusting** for these things
  –o.w. they may confound the arm circumference-weight relationship

- We would like to assess how arm circumference responds **within subjects** to variation in weight
  – each person act as his/her own control

- suppose we had measures of these things — we might fit the model:

  $$Y_{ij} = \beta_0 + \beta_1 \texttt{wt}_{ij} + \beta_2 \texttt{age}_{ij} + \beta_3 \texttt{sex}_i + \beta_4 \texttt{genetics}_i + \beta_5 \texttt{nutrition}_i + Z_{ij}$$

  and obtain the adjusted coefficient β_1 for weight

- since we do not get to observe all of these things, suppose we just lump them all together into U_i:

  $$U_i = \beta_4 \texttt{genetics}_i + \beta_5 \texttt{nutrition}_i + \beta_3 \texttt{sex}_i$$

  and fit the model

  $$Y_{ij} = \beta_0 + \beta_1 \texttt{wt}_{ij} + \beta_2 \texttt{age}_{ij} + U_i + Z_{ij}$$

- Note: $\texttt{sex}_i$ is absorbed in U_i

- Note: Any constant can be added to β_0 and subtract from U_i

  – We can restrict $\sum_{i=1}^{m} U_i = 0$ in estimation

  $\rightarrow$ β_0 is average of personal intercepts $U_i + \beta_0$.

- then we have automatically adjusted for genetics, nutrition, etc., without including them in the model!

  $\rightarrow$ **subject-level** effect U_i adjusts for unobserved subject-level factors

  $\rightarrow$ each subject is his/her **own control**

- however, the U_i now contains **confounding variables** on the relationship of arm circumference to weight:

  $\rightarrow U_i$ associated with both arm circumference **and** weight

  $\rightarrow U_i$ is **not independent** of covariates $\boldsymbol{x}_{ij}$! (as in random effects models)

- treating the U_i's as **fixed quantities** and trying to estimate the model skirts this violation of the independence of random effects assumption

- fixed effects models restrict our focus to **only within-subject covariation** of predictor and response

Estimation of Fixed Effects Model (Classic Approach)

- Starting with

$$Y_{ij} = \beta_0 + \beta_1 \texttt{wt}_{ij} + \beta_2 \texttt{age}_{ij} + \beta_3 \texttt{sex}_i + U_i + Z_{ij}$$

- Average the LHS for each subject i to get $\bar{Y}_i$
- Average the RHS for each subject i to get

$$\beta_0 + \beta_1 \mathrm{avg}(\texttt{wt})_i + \beta_2 \mathrm{avg}(\texttt{age})_i + \beta_3 \texttt{sex}_i + U_i + \bar{Z}_i$$

- Now subtract to obtain **within-subject deviations**

$$\begin{aligned} Y_{ij} - \bar{Y}_i &= \{\beta_0 + \beta_1 \texttt{wt}_{ij} + \beta_2 \texttt{age}_{ij} + \beta_3 \texttt{sex}_i + U_i + Z_{ij}\} \\ &\quad -\{\beta_0 + \beta_1 \mathrm{avg}(\texttt{wt})_i + \beta_2 \mathrm{avg}(\texttt{age})_i + \beta_3 \texttt{sex}_i + U_i + \bar{Z}_i\} \\ &= \beta_1(\texttt{wt}_{ij} - \mathrm{avg}(\texttt{wt})_i) + \beta_2(\texttt{age}_{ij} - \mathrm{avg}(\texttt{age})_i) + (Z_{ij} - \bar{Z}_i) \end{aligned}$$

- **Note:** Can see in this model form that the model isolates the **within-subject covariability** of arm circumference, weight and age
- Fitting this model with OLS is one way to estimate a fixed effects model
- To estimate U_i (with restriction of $\sum_{i=1}^{m} U_i = 0$):
  - Within each subject i, by averaging

    $$Y_{ij} = \beta_0 + \beta_1 \texttt{wt}_{ij} + \beta_2 \texttt{age}_{ij} + U_i + Z_{ij},$$

    we can estimate $\beta_0 + U_i$ by

    $$\widehat{\beta_0 + U_i} = \bar{y}_i - \bar{\boldsymbol{x}}_i' \widehat{\boldsymbol{\beta}},$$

    where $\bar{\boldsymbol{x}}_i' \widehat{\boldsymbol{\beta}}$ includes subject-averaged effects of `wt` and `age`.

– β_0 can be estimated by averaging $\widehat{\beta_0 + U_i}$:

$$\widehat{\beta}_0 = \frac{1}{m} \sum_{i=1}^{m} \widehat{\beta_0 + U_i}$$

– Fixed effects U_i can be "estimated" as:

$$\widehat{U}_i = (\bar{y}_i - \bar{\boldsymbol{x}}_i' \widehat{\boldsymbol{\beta}}) - \widehat{\beta}_0$$

- SAS code:

```
proc glm data=nepal;
absorb id;
class sex;
model arm = wt age sex/ solution;
contrast 'F test for wt and age'
wt 1 ,
age 1 ;
run;
```

Number of Observations Read	1000
Number of Observations Used	877

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	198	1033.244986	5.218409	42.06	<.0001
Error	678	84.120106	0.124071		
Corrected Total	876	1117.365092			

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
F test for wt and age	2	108.0030639	54.0015319	435.25	<.0001

Parameter		Estimate	Standard Error	t Value	Pr > |t|
wt		0.8010690922	0.03096592	25.87	<.0001
age		-.0754262423	0.00466085	-16.18	<.0001
sex	1	0.0000000000 B	.	.	.
sex	2	0.0000000000 B	.	.	.

- the effect of sex is not estimable
- the estimated coefficient of weight is about 25 percent higher in the fixed effects formulation than in the random effects formulation, suggesting some confounding by unobserved between subject factors
- Degrees of freedom (df) calculation for fixed effects model:
  - observations: 877
  - subject-specific intercepts (number of subjects): 197
  - within-subject predictors: 2

  final denominator df: 877 - 197 - 2 = 678

- The overall model F-test above the main output tests whether all within-subject model terms are zero, i.e.

$$H_0 : \beta_1 = \beta_2 = 0$$

versus

$$H_A : \text{ at least one of } \beta_2, \beta_3 \neq 0$$

Therefore, this is on 2 numerator df

- The denominator df 678 is used in the overall model F-test as well as for any F- or t-test generated from the model, e.g.,

```
contrast 'F test for wt and age'
wt 1 ,
age 1 ;
```

would generate F-test with 2 numerator df and 678 denominator df

- Note: F-test for fixed-effects model is same to OLS, by considering U_i as coefficients of dummy variables.

- An alternative way in SAS:

```
proc glm data=nepal;
class id;
model arm = wt age id / solution p noint;
ods output ParameterEstimates=parm PredictedValues=pred;
run;
```

```
                                          Standard
Parameter                  Estimate          Error    t Value    Pr > |t|

wt                      0.801069092     0.03096592      25.87      <.0001
age                    -0.075426242     0.00466085     -16.18      <.0001
id          120011      7.131001505     0.29489058      24.18      <.0001
id          120012      6.660281955     0.31420165      21.20      <.0001
<snip>
```

- $\widehat{\beta_0 + U_i}$ are estimated for each subject i

- Let's check $\mathrm{corr}(\hat{U}_i, \boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}})$

```
*obtain U (include beta0);
data parm;
    set parm;
    if (Parameter eq "wt")or(Parameter eq "age")then delete;
    id=input(substr(parameter,11),12.0);
    rename estimate=U;
    keep id estimate;
run;

*obtain X*beta;
data pred;
    merge nepal(keep=id) pred;
    keep id Predicted;
run;
data tmp;
    merge pred parm;
    by id;
    Xbeta=predicted-U;
run;

*calculate corr(U_i,X*beta);
```

```
proc corr data=tmp;
var U Xbeta;
run;
```

```
                Pearson Correlation Coefficients
                   Prob > |r| under H0: Rho=0
                      Number of Observations

                              U           Xbeta

              U         1.00000        -0.56602
                                         <.0001
                            985             877

              Xbeta    -0.56602         1.00000
                         <.0001
                            877             877
```

- Correlation between $\widehat{U}_i$ and the fitted linear predictor $\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}}$ is rather large (-0.56)
  – suggesting substantial confounding of the relationships between arm circumference and weight and age in a non-fixed effects model (either a marginal model or a random effects model)

- Another way in SAS using proc panel (include F test for Fixed Effects U):

```
data nepal2;
set nepal;
  if(id=360162)or(id=360431)or(id=360432)then delete;
        *all arm are missing for these subjects;
  obs + 1;
  by id;
  if first.id then obs = 1;
run;
proc panel data = nepal2;
   id id obs;
   model arm = wt age sex/ fixone;
run;
```

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	7.082318	0.2232	31.73	<.0001	Intercept
wt	1	0.801069	0.0310	25.87	<.0001	
age	1	-0.07543	0.00466	-16.18	<.0001	
sex	0	0	.	.	.	

F Test for No Fixed Effects

Num DF	Den DF	F Value	Pr > F
196	678	13.94	<.0001

- The F-test that all $U_i = 0$ tests the hypothesis

$$H_0 : U_i = 0, i = 1, \ldots, m$$

versus

$$H_A : \text{at least one } U_i \neq 0$$

  - Null: all subjects share the same intercept β_0
  - This test is on 196 df (ie, numerator DF$=m-1$) because there are 196 constraints (note we set $\sum_{i=1}^{m} U_i = 0$).
  - The test rejects with $F = 13.94$ and we conclude that there are differences among subjects in subject-specific intercepts

- Note: Stata code for FE model: `xtreg arm wt age sex , fe`

Explore Alternative Way
Partial regression for fixed effects models
(Decompose between-/within-subject effects)

- Fitting the linear FE model is equivalent to doing partial linear regression on a model with a separate intercept $(\beta_0 + U_i)$ on each subject
  – like including a **dummy variable** $U_1, \ldots, U_m$ for each of m subjects

- Partial regression:
  1. both y_{ij}'s and each covariate $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})$ is regressed on the set of subject-level dummy variables
     – ie, between-subject values: eg, $\bar{x}_{i\cdot}$ (average within subject)
  2. y_{ij} residuals and x_{ij} residuals are obtained from these regressions
     – ie, within-subject deviation: eg, $(x_{ij} - \bar{x}_{i\cdot})$ (longitudinal effect)
  3. the y_{ij} residuals are regressed on the set of x_{ij} residuals

- This will give you exactly the same β coefficients as using the within-subject deviation method above.
  – The within-subject deviations are **in fact** the residuals from the dummy variable regression:
  Within-subject deviations

$$Y_{ij} - \bar{Y}_i \quad = \quad \beta_1(\texttt{wt}_{ij} - \mathrm{avg}(\texttt{wt})_i) + \beta_2(\texttt{age}_{ij} - \mathrm{avg}(\texttt{age})_i) + (Z_{ij} - \bar{Z}_i)$$

- However, final regression will not know to account for the fact that m separate coefficients (for dummy variables) have already been estimated before β
  – need to adjust DF

- Standard errors are incorrect because it is assumed that the residuals are independent
  – i.e., denominator DF $= 877 - 2 = 875$ (wrong!).

Summary
A few general notes for FE

- When you include a fixed subject-specific intercept, you cannot estimate the effects for any subject-level covariates
- FE model sacrifices **between-subject** information to avoid the assumption of independence of U_i's from the $\boldsymbol{X}_i$'s
- All between-subject variation in the data is absorbed by the subject-specific dummies
  – Effects of independent variables are solely within-subject effects
  – There needs to be enough within-subject variation for the estimator to be meaningful
- FE models may have too many subjects
  $\rightarrow$ Too many dummy variables for model specification.
  $\rightarrow$ decreases the degrees of freedom for adequately statistical tests.

- Autocorrelation over time is not solved by FE

- FE model is equivalent to estimating a separate intercept for each subject
  – possible to include other fixed effects such as slopes, but estimation is complicated
  $\rightarrow$ may require non-standard software

- we will revisit these models later on when we do models for binary and count data

Compare FE and RE models

- FE model:
  - U_i's are allowed to be correlated with between subject components of $\boldsymbol{X}_i$'s
  - No assumption on distribution of U_i

- Random effects (RE) model
  - Independence assumption between U_i's and $\boldsymbol{X}_i$'s
    – is much safer when the within-subject covariates are fixed by design, or are the same for all subjects (e.g., week)
  - Normality distribution of U_i
    – Some numerical studies claim evidence of robustness of RE model on normality assumption violation
- Hausman test (1978)
  – Null hypothesis: U_i are not correlated with the regressors X
  – A significant test result $\rightarrow$ RE model should be rejected in favor of the FE model.
  – R code examples for FE/RE model and Hausman test (https://www.princeton.edu/ otorres/Panel101R.pdf)