Topics for today

- *Time-varying age versus baseline age in longitudinal data analyses*
- *Separating within and between-subject effects for time-varying covariates*
- *Measuring the relationship between a time-varying covariate and an outcome for longitudinal data*
- *Modeling time as a class or continuous variable*
- *Population-averaged versus subject-specific effects*

Related reading: *Interpreting parameters in longitudinal models* course notes. (Sections 4 and 6 not discussed here.)

1 Time-varying age versus baseline age in longitudinal data analyses

- Most longitudinal experiments and studies involve a relatively short amount of time, anywhere from a few weeks to a few months. Typically there is a time variable to indicate when measurements were taken.

- If one is interested in also including age of subjects as a covariate into the model, typically baseline age is used – i.e., the fixed age for subjects when they enter the study. But one question that arises is, should a time-varying age be used, or baseline age?

- Many think that baseline age is most appropriate when another time variable is already in the model. Here, we will examine both approaches mathematically, and the results may surprise you a little bit, although we will also see that you can easily estimate effects of interest from either approach.

- Consider a study in which subjects' blood pressures are observed over time (3 time points, equally spaced, no missing data). The model will include fixed effects for time, age (at start of experiment) and gender; and a random intercept for subjects. The AR(1) structure will be used to model the errors. In the model, age at start of experiment (i.e., baseline) was used. How would estimates change if you used continuous age in the model instead?

- In order to answer the question, write out the statistical models for both approaches. Note that *real_age* = *BL_age* + *time*. Thus, the model using real age (but not including gender) is

$$E(Y) = \beta_0 + \beta_1 real\_age + \beta_2 time$$
$$E(Y) = \beta_0 + \beta_1 (BL\_age\_ + time) + \beta_2 time$$
$$E(Y) = \beta_0 + \beta_1 BL\_age + (\beta_1 + \beta_2) time$$
$$E(Y) = \beta_0 + \beta_1 BL\_age + \beta_2^{new} time$$

- So we have expressed the model with real age in terms of the one that uses baseline age. Thus, the underlying models are the same. However, the interpretation of the parameters differs for the 2 approaches, as the above equations suggest.

- Considering the models above, β_1 is a between-subject age effect, regardless of whether *real_age* or *BL_age* is used; β_2 is a within-subject time (or age) effect; β_2^{new} is the combination of these. So in order to get effects of time that does not involve between-subject age effects, we use the model with *real_age*.

- As an example of when this may be of interest; the Beryllium natural history project involved evaluating the progression of illness that was not due to the aging process (specifically, we removed between-subject age effects from the time variable). Thus, real age was used in the analysis.

- In many other cases, *BL_age* is used, although many probably don't really understand the difference between the two approaches. But often there won't be a great difference unless the study is over a longer period of time. You can estimate all parameters mentioned above from either model (e.g., by including an appropriate ESTIMATE statement in PROC MIXED).

2 Separating within and between-subject effects for time-varying covariates

2.1 Methods and simpler models

- Data from many longitudinal experiments or studies are fit in regression-type models (e.g., mixed models), where each time-varying covariate is fit with one term. For example, in the Kunsberg / air pollution studies that I've been involved with, we fit health outcome models as a function of an air pollution variable (with a fixed-effect coefficient) plus other fixed and random effect terms. When personal monitors are used, the pollution variable is both subject and time-specific (i.e., subject-varying and time-varying).

- If just one term is used for the pollution variable, then we are fitting a parameter that involves pooled effects based on between-subjects differences as well as within-subject changes over time. To illustrate, consider a linear mixed model with a random intercept for subjects and fixed-effect term(s) for pollutant variable(s). Here, Y_{ij} is some health outcome measure such as FEV1 and say x_{ij} is the subject-specific (i.e., personal) pollution level for subject i on day j.

$$Y_{ij} = \beta_0 + b_{0i} + \beta x_{ij} + \varepsilon_{ij}$$

- For subject i,Y changes by an expected amount $\beta\left(x_{ij} - x_{ij'}\right)$ from day j to j'. Similarly, for day j, the expected difference between subject i and i' is $\beta\left(x_{ij} - x_{i'j}\right)$. The within and between-subject effects get pooled as there is only one slope parameter for the pollutant variable. We can obtain separate estimates for between-subject and within-subject effects. To do this, note that

$$x_{ij} = \overline{x}_i + (x_{ij} - \overline{x}_i).$$

- Thus, by fitting terms separately for $\overline{x}_i$ and $x_{ij} - \overline{x}_i$, we can determine if slope estimates differ for within and between- subject data. If there is no difference, then it is not necessary to use the separate terms. [You could also use $\overline{x}_i$ and x_{ij} as the BS and WS variables, respectively. What are the implications?]

- To further illustrate, consider a study where birth weight is the outcome for 880 mothers that had 5 children (data from CDC), and the time-varying covariate is the mother's age at each birth (denoted by x_{ij}, where i indexes subject, i=1,…,880, and j indexes the birth event j=1,…,5).

- Here are the models:

$$Y_{ij} = \beta_0 + b_{0i} + \beta x_{ij} + \varepsilon_{ij} \qquad (1)$$

$$Y_{ij} = \beta_0 + b_{0i} + \beta_B \overline{x}_i + \beta_W (x_{ij} - \overline{x}_i) + \varepsilon_{ij} \qquad (2)$$

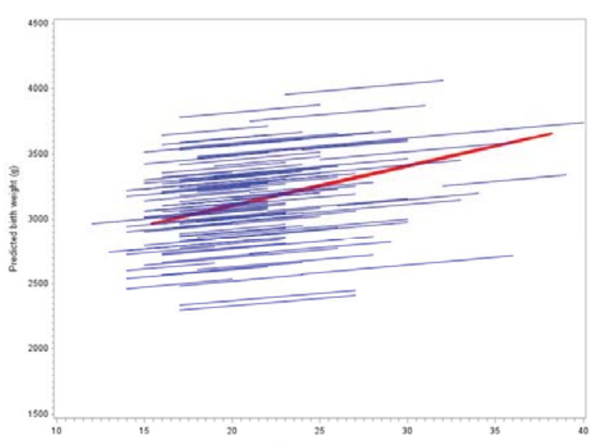
where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, $b_{i0} \sim N(0, \sigma_b^2)$ for both models.

- Fitting model (1) yields $\hat{\beta}$ =17.14 grams, which indicates that increase in birth weight increases is about 17 grams, on average, per year.

- However, when model (2) is fit, we obtain $\hat{\beta}_B$ =30.35 grams and $\hat{\beta}_W$ =11.83 grams. Thus, we would estimate that average birth weight will differ by about 30 grams between two women whose average age at birth differs by one year, and for a given woman, we would estimate that the birth weights of her children increase by an average of about 11.8 grams for each year that she ages.

- The increase estimates here may be largely due to the fact that younger women were studied – the median age at first birth was 17 years! For more detail, see Neuhaus and Kalbfleisch, *Between and within-cluster covariate effects in the analysis of clustered data*. Biometrics, 54, 638-645, 1998. Hedeker also discusses this issue on pages 72-74.

- Below are predicted values for subjects (blue) and the between-subject average fit (red) for a random sample of the 878 mothers.



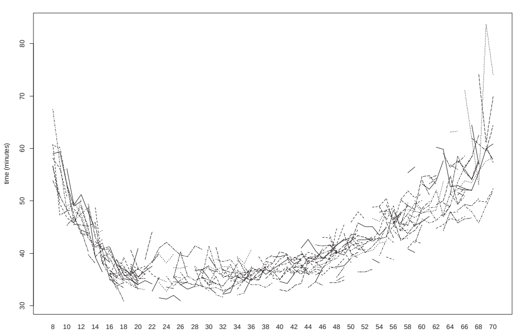
2.2 Case study: More advanced models with Bolder Boulder data

- The Bolder Boulder is a 10K race held in Boulder, Colorado on Memorial Day. The race has been run for several decades, and is one of the largest running races in the United States. Using this data, we can estimate between and within-subject changes over time. However, modeling these data requires nonlinear functions, (see *Non-normal and nonlinear* notes).

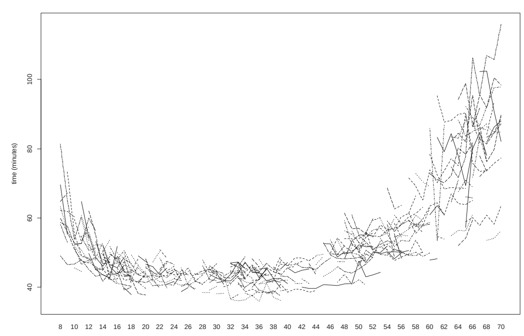
- Here we consider 12 consecutive years of data where subjects may participate in multiple years, and thus longitudinal data. In our analysis we focused on the most competitive runners in the Citizen’s (nonprofessional) race. For more specifics on subject and record inclusion, see the course notes.

- The following figure shows a spaghetti plot of a random 20% of runners, for both men and women.

(a)



(b)



- The function we use to model race time versus age is $f(x) = \alpha_0' x^{\alpha_1} e^{\alpha_2 x}$, however we can linearize the function by taking natural logs of both sides: $\ln[f(x)] = \alpha_0 + \alpha_1 \ln(x) + \alpha_2 x$, where $\alpha_0 = \ln(\alpha_0')$

- This is important when separating effects into between-subject (BS) and within-subject (WS) components, since the model will already be pretty complex.

- The function we found to work has the following form:

$$Y_{ij} = \alpha_0 + \alpha_1 \ln \bar{x}_i + \alpha_2 \bar{x}_i$$
$$+\beta_1\left(x_{ij} - \bar{x}_i\right) + \beta_2\left(x_{ij} - \bar{x}_i\right)^2 + b_{0i} + b_{1i}\left(x_{ij} - \bar{x}_i\right) + b_{2i}\left(x_{ij} - \bar{x}_i\right)^2 + \varepsilon_{ij}$$

where Y is the natural log race time, x is age, i indexes subject and j time, and $\mathbf{b}_i = (b_{0i}, b_{1i}, b_{2i})^t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, independently of $\varepsilon_{ij} \sim iid\ \mathrm{N}(0, \sigma_\varepsilon^2)$. The first line of the equation above contains the primary BS component. If we average over all races for subject i and take the expected value, we obtain

$$E(\bar{Y}_{i\cdot}) = \alpha_0 + \alpha_1 \ln \bar{x}_i + \alpha_2 \bar{x}_i + \beta_2 Ave\left(x_{ij} - \bar{x}_i\right)^2$$

For subjects who run in 9 consecutive years, the BS function becomes

$$E(\bar{Y}_{i\cdot}) = (\alpha_0 + 6.67\beta_2) + \alpha_1 \ln \bar{x}_i + \alpha_2 \bar{x}_i$$

- Since it is not given that WS functions will be constant across ages, we can determine age-specific WS functions by averaging random effects within in certain age ranges.

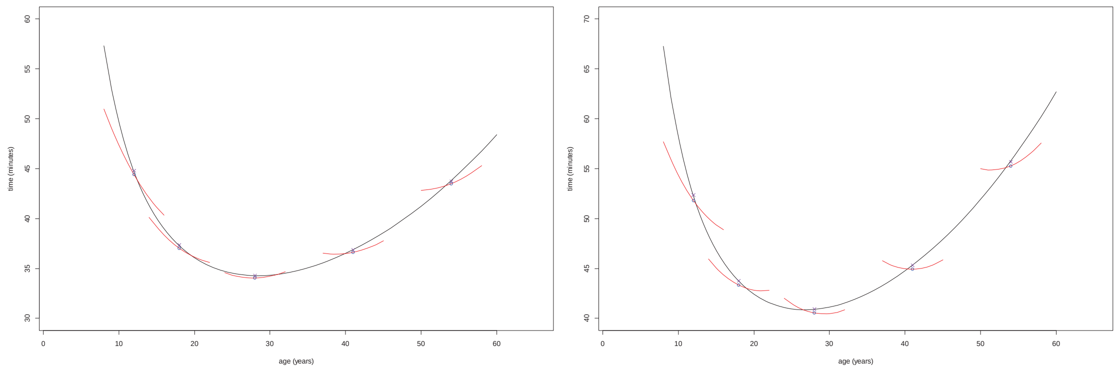
- Specifically, consider

$$E\left(Y_{ij} \middle| \mathbf{b}_i = Ave_k\{\mathbf{b}_i\}\right)$$
$$= f(\bar{x}_i) + (\beta_1 + Ave_k\{b_{1i}\})\left(x_{ij} - \bar{x}_i\right) + (\beta_2 + Ave_k\{b_{1i}\})\left(x_{ij} - \bar{x}_i\right)^2$$

where $Ave_k\{b_{1i}\}$ denotes the average of all b_{1i} within 3 years of age k. (The 3 could be changed but this was the value chosen in our analysis.)

- The resulting estimates are shown below for men and women. As before, the WS functions tend to be flatter, although in this case, they are a bit closer to the BS function.

(a) Men (b) Women



- An alternative approach to estimating WS functions is to create a parametric model that includes interaction terms between $\bar{x}_i$ and $x_{ij} - \bar{x}_i$ terms.

- Using the between and within subject functions, we can better understand how people change in performance over time.
  - Using the BS function to find ‘peak age’
  - Rates of change in BS function: ~1% per year decline at age 40, ~2% per year at age 60.
  - Within-subject changes are a bit more attenuated relative to the between-subject curve.
    - A 42-year-old male has an expected 1.0% slower time than a 41-year-old (based on the between-subject curve)
    - A 41-year-old that participates in the following year is expected to only slow by a rate of 0.3% based on the within-subject function.
  - For women, there is an even greater difference between BS and WS curves, as the graph displays.

4 Using mean cumulative versus non-cumulative responses in longitudinal models

See the course notes.

5 Modeling time as a class or continuous variable

5.1 Pro's and con's of approaches

- For some data, it is pretty clear whether time should be modeled as a class or continuous variable. For other data, it is not so clear.

- If data involves 4 or 5 time points or less, modeling time as a class variable typically yields a better fit. A wrinkle with this occurs when actual times of measurement do not meet the prescribed dates (e.g., for a '1-year' follow up, subjects might come in early or late). This has happened in several student projects. Whether the time as class variable should be abandoned really depends on the data and the degree of 'error' in times of measurement of subjects. If subjects still come in fairly close to the prescribed dates then the class variable approach may still provide a reasonable approximation.

5.2 Inference when time is modeled as a continuous variable

- We thoroughly examined tests involving *group*, *time* and *group***time* terms when time was a class variable. When time is modeled as a continuous variable, often the focus is more on estimation and features of function being estimated (e.g., slope of a linear function, minimum or maximum of a quadratic function, derivative of a curve, etc.). In order to get the minimum or maximum of a parabola, you can take the derivative of the fitted function, set it to 0 and solve for x. As an example, see the Non-normal notes and the Bolder Boulder data application.

- When there is a group variable and time is modeled as a continuous variable, you can still easily perform hypothesis tests to compare groups at particular time points (e.g., perform t-tests by including ESTIMATE statements in PROC GLM or PROC MIXED to compare genders at fixed ages since there is one degree of freedom for such tests). Some examples are shown ahead.

- Say you have two groups that you want to compare more generally. Here are some possible tests of interest:

  - Compare curves

  - Compare differences in curves over time minus intercept differences (i.e., interaction)

  - Compare highest order trend (e.g., compare quadratic trends if model is specified up through quadratic)

- To determine how to write these tests, consider the model for the Mt. Kilimanjaro data:

$$E(Y_{hij}) = \beta_0 + \beta_1 x_{ij} + \beta_2 x_i^2 + \alpha_h + \gamma_{h1} x_{ij} + \gamma_{h2} x_{ij}^2 \ ,$$

  where *h* indexes drug group, *i* indexes subject and *j* indexes the repeated measure. (We don't need an '*h*' index on *x* as long as the subject index is unique, study wide.) There is a similar model for the Bolder Boulder data in the course notes…

- Using SAS's g-inverse, the ever users become the 'reference group' since it is the highest level. The models by Diamox users (never, ever) can be expressed as

  Never users: $E(Y_{1ij}) = (\beta_0 + \alpha_1) + (\beta_1 + \gamma_{11}) x_{ij} + (\beta_2 + \gamma_{12}) x_{ij}^2$

  Ever users: $E(Y_{2ij}) = (\beta_0 + \alpha_2) + (\beta_1 + \gamma_{21}) x_{ij} + (\beta_2 + \gamma_{22}) x_{ij}^2$

- Since the higher level (ever users) is the reference group, the terms $\alpha_1,\ \gamma_{11},\ \gamma_{12}$ represent deviations of the never-user function from the ever users, for intercept, linear and quadratic terms, respectively. Thus, we can simply the function for the ever users to

  Ever users: $E(Y_{2ij}) = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2$

- In terms of the model, the overall test to compare curves can be written as

  H_0: $\alpha_1 - \alpha_2 = 0,\ \gamma_{11} - \gamma_{21} = 0,\ \gamma_{12} - \gamma_{22} = 0$.

- Similarly, the interaction test is:

$$\mathrm{H}_0\text{:}\quad \gamma_{11}-\gamma_{21}=0,\ \gamma_{12}-\gamma_{22}=0\,.$$

- If the comma is removed, the test becomes

$$\mathrm{H}_0\text{:}\quad (\gamma_{11}-\gamma_{21})+(\gamma_{12}-\gamma_{22})=0\,,$$

  which is not an interaction test. In words, the null hypothesis for this test is that the sum of the linear and quadratic deviations for the never users is 0. This test is probably not of interest, since you could have offsetting non-zero values for the linear and quadratic parameters.

- Here is the Mt. Kilimanjaro data to show how such tests could be conducted using *F*-tests (via CONTRAST statements), plus some simpler comparisons at specific time points using ESTIMATE statements.

- Note: relative to the previous code we saw, I have now added a REPEATED statement to further model the repeated measures (before I only had a RANDOM statement); recall that ‘x’ is altitude.


```
proc mixed data=alldata; class id recnum;
class diamox_ever;
model oxygen_stats= x x*x diamox_ever x*diamox_ever x*x*diamox_ever
 / outpm=outypm outp=outyp solution;
random intercept x x*x / subject=id v solution g type=un;
repeated Recnum/type=ar(1) subject=id;

estimate 'no diamox at alt=5km'
 intercept 1 x 5 x*x 25 diamox_ever 1 0 x*diamox_ever 5 0 x*x*diamox_ever 25 0;
estimate 'diamox at alt=5km'
 intercept 1 x 5 x*x 25 diamox_ever 0 1 x*diamox_ever 0 5 x*x*diamox_ever 0 25;
estimate 'diff at alt=5km'
 diamox_ever 1 -1 x*diamox_ever 5 -5 x*x*diamox_ever 25 -25;

contrast 'interaction'
 x*diamox_ever 1 -1, x*x*diamox_ever 1 -1;
contrast 'curve comparison'
 diamox_ever 1 -1, x*diamox_ever 1 -1, x*x*diamox_ever 1 -1;

ods output solutionf=sf;
ods output estimates=esty; run;
```

- For thought: in the CONTRAST statements above, what is the difference between including and excluding the ‘,’ between terms?

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
no diamox at alt=5km	80.8365	0.4687	11E3	172.47	<.0001
diamox at alt=5km	82.0397	0.1525	11E3	537.91	<.0001
diff at alt=5km	-1.2032	0.4929	11E3	-2.44	0.0147

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
interaction	2	11E3	6.53	0.0015
curve comparison	3	11E3	4.52	0.0036

- Since the higher level of factors are 0 by the way SAS computes the g-inverse, we could simplify the notation for the tests as follows:
  - Curve comparison test: H_0: $\alpha_1 = 0$, $\gamma_{11} = 0$, $\gamma_{12} = 0$.
  - Interaction test: H_0: $\gamma_{11} = 0$, $\gamma_{12} = 0$.

- You can also create a model using diamox_ever as a binary ‘continuous’ variable (i.e., do not include in the CLASS statement, but levels must be 0 and 1), and obtain the same results. In this case only 1 column is added to the X matrix and there is only 1 coefficient for diamox_ever instead of 2 (when writing contrasts and estimate statements). Can you rewrite the ESTIMATE and CONTRAST statements in this case?

- The tests for individual coefficients of the *group***time* and *group***time*^{2} terms are probably less meaningful. These tests are not like the polynomial interaction tests we considered when time was modeled as a class variable.

- When there are polynomial functions that differ by groups, tests at one specific time point, or changes between time points may be of interest.

- If applicable, so you also evaluate derivatives of functions to express rate of change. If you have a quadratic function, the derivative function expressing rate of change will be linear.

<u>Notes on DDF calculations</u>

- In our case, the default method to calculate DDF for our custom *t* and *F* tests is the ‘containment’ method. For these data, the calculated DDF is equal to the number of records used to fit the model (13369), minus the number of random effect terms per subject times the number of subjects (3*916=2748), so the DF is 13369 – 2748 = 10621. In the output, we just see 11E3 but I used ODS OUTPUT to get the exact number.

- If we use the RESIDUAL method, we essentially ignore the repeated measures and treat each record as if it came from a different subject. This would yield DDF=n–$rk(\mathbf{X})$=13369 – 6=13363.

6 *Interpreting effects for loglinear and logistic models*

See course notes and '*Non-normal I*' slides

7 *Population-averaged versus subject-specific effects*

- Beta parameters may have subject-specific or population-averaged interpretations, depending on the type of model being fit. For Normal outcomes, the interpretation is the same (which is why we have not discussed this issue yet). But for some other types of outcomes, this may not be so.

- Consider a linear mixed model with random intercept and slope for time (*x*), by subject, and fixed effects that also include an intercept and slope for time (for simplicity we won't consider other covariates).
  - The (estimated) population mean function is $E\left(Y_{ij}\right)=\beta_0+\beta_1 x_{ij}$.
  - The subject-specific function is $E\left(Y_{ij}\mid \mathbf{b}_i\right)=\beta_0+\beta_1 x_{ij}+b_{0i}+b_{1i}x_{ij}$. In particular, the 'average' subject has no deviation from the population average (b_0 and b_1 are 0), and thus the mean is $E\left(Y_{ij}\mid \mathbf{b}_i=0\right)=\beta_0+\beta_1 x_{ij}$.

- Since the population average function is equivalent to the subject specific function (for the average subject), we say that fixed effects in this case have both population-averaged and subject-specific interpretations.

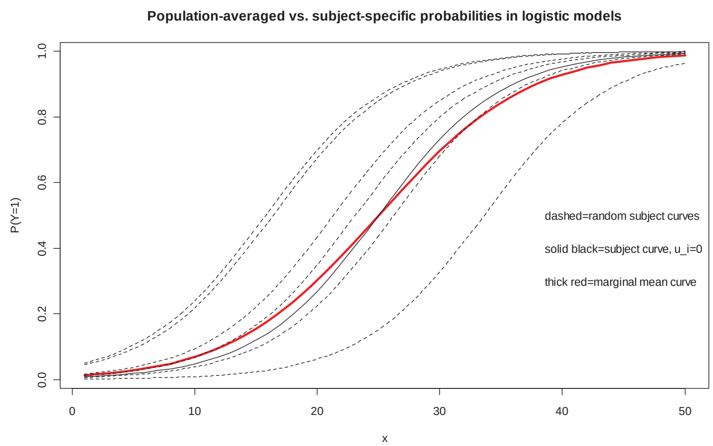
- For other distributions, these may not be equal. Consider a GzLMM where the outcome has a Poisson distribution, the fixed effects include a simple linear trend for time, and there is a random intercept for subjects:
  $g\left(\mu_{ij}\right) = \beta_0+\beta_1 x_{ij}+b_i$.

  - The conditional mean is $E\left(Y_{ij}\mid b_i=0\right)=e^{\beta_0+\beta_1 x_{ij}}$
  - The marginal mean is $E(Y_{ij})=E[E(Y_{ij}|b_i)]=E[e^{\beta_0+\beta_1 x_{ij}+b_i}]$

$$\begin{aligned} &= e^{\beta_0+0.5\sigma_b^2}e^{\beta_1 x_{ij}} \\ &= e^{\beta_0'}e^{\beta_1 x_{ij}} \end{aligned}$$

- The only difference between these means is in the intercept; it is greater in the marginal mean by the amount ½σ_b^2 compared with the conditional mean (for subjects with b_i=0). The parameter β_1 is typically of more interest and has both specific-subject and population-averaged interpretations. However, results do not necessarily generalize for more complex random effects.

- For a binary outcome, β_1 will have different interpretations, considering the model $g\left(\mu_{ij}\right) = \beta_0 + \beta_1 x_{ij} + b_i$, where g is the logit link.
- The graph below shows $E\left(Y_{ij}\right)$, red, with subject curves, $E\left(Y_{ij} \mid b_i\right)$, black. [$E\left(Y_{ij} \mid b_i = 0\right)$ is solid and the rest are dashed.] Clearly the population averaged function is not the same as the function for the average subject. Thus, both the intercept and slope parameters differ in interpretation in this case. The marginal model has parameters with population-averaged interpretations, while the conditional model has subject-specific interpretations.



- Implications for non-normal outcome models
  - GzLM using GEE can be used for Poisson or logistic regression with longitudinal data. It is particularly useful when there is serial correlation (e.g., that can be modeled using the AR(1) structure), but no random effect terms are required.
    - Since random effect terms are not in the model, the beta parameters will have population-averaged interpretations for logistic regression models.
    - For Poisson regression, parameters for predictors (other than the intercept) have both population-averaged and subject-specific interpretations if there are random-intercept differences between subjects (but not more complex random terms).
  - GzLMMs can have random effects, so the analyst has the ability to fit parameters that either have population-averaged interpretations (if appropriate random effects are not included in the model) or subject-specific interpretations (if they are included).
- For normal-theory models, the interpretations will be the same.