Topics for today

- *Fitting nonlinear functions*
- *Mixture distributions*

Related reading: Sections 6-7 in Non-normal notes.

6 Using NLMIXED to fit nonlinear functions

- Up to this point we have considered linear models for the predictor part of the model (i.e., linear predictors, $\boldsymbol{\eta}=\mathbf{X}\boldsymbol{\beta}$), whether it be GLM, GzLM, LMM or GzLMM. In all of these, the 'L' stands for linear.

- Sometimes you may want to fit a predictor that is not linear. What we mean by nonlinear here is that the function is nonlinear with respect to the parameters.

- Of course, we could fit a function that does not follow a straight line but is linear with respect to the parameters (e.g., $f(x)$ = quadratic, sinusoidal, etc.) using any of the linear models methods mentioned above.

- However, once the function is not a linear combination of parameters, we cannot use the aforementioned methods to fit the function.

- As an example, considering the Bolder Boulder 10K race time data fit as a function of age. In my Master's thesis, I found that the function $f(x) = \alpha_0 x^{\alpha_1} e^{x\alpha_2}$ fit the data well. At first glance, it looks like a quadratic function might work well, although it does much worse, in terms of sum of squared errors. If you only model after age 30, the quadratic does pretty well.

- Here is how to fit the function in PROC NLMIXED and the subsequent graph. Here I fit males and females separately for simplicity. It should be noted that the data are extreme minima that should be modeled with an extreme value distribution such as the Gumbel (not Normal).

- But if we are simply interested in curve fitting, it works fine. The solution will satisfy the least squares criterion for the given function. Note that I had to specify more stringent convergence criteria to get the correct solution. Also, using initial parameter values (parms) that are relatively close to the actual solution helps.


```
proc nlmixed data=male gconv=1e-10;
 parms a0=270 a1=-1 a2=0.05 res=2;
 n=a0*(age**a1)*(exp(age*a2));
 model time~normal(n,res); run;
```

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
a0	277.23	11.1393	265	24.89	<.0001	0.05	255.30	299.17	3.676E-7
a1	-0.9052	0.01846	265	-49.05	<.0001	0.05	-0.9416	-0.8689	0.000345
a2	0.03198	0.000677	265	47.26	<.0001	0.05	0.03065	0.03332	0.003519
res	2.9792	0.2588	265	11.51	<.0001	0.05	2.4696	3.4888	-6.06E-7

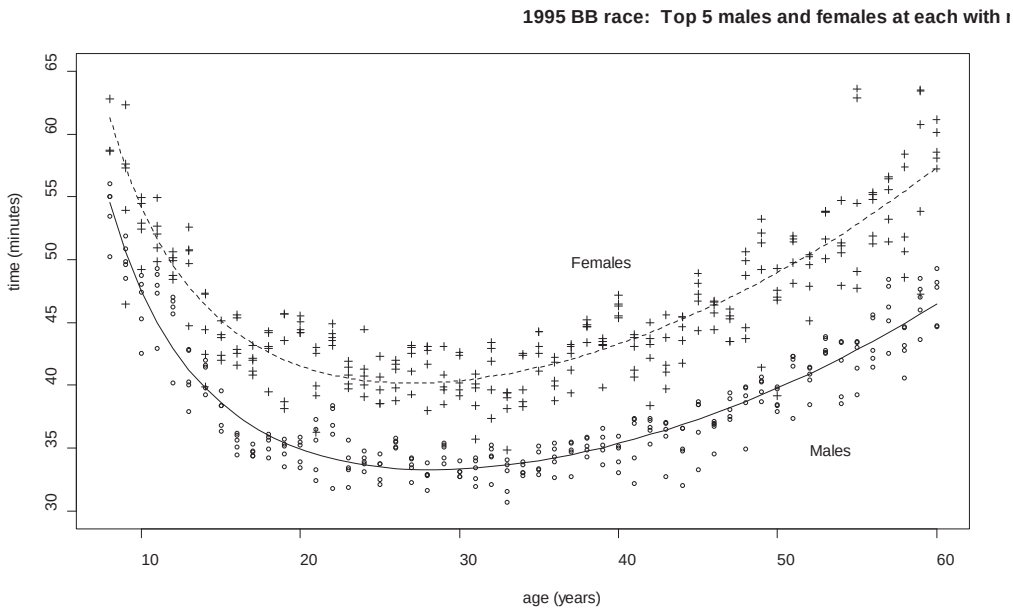
```
proc nlmixed data=female gconv=1e-10;
 parms a0=270 a1=-1 a2=0.05 res=2;
 n=a0*(age**a1)*(exp(age*a2));
 model time~normal(n,res); run;
```

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
a0	268.46	15.4156	265	17.41	<.0001	0.05	238.10	298.81	-1.02E-6
a1	-0.8290	0.02616	265	-31.69	<.0001	0.05	-0.8805	-0.7775	-0.00091
a2	0.03084	0.000948	265	32.55	<.0001	0.05	0.02897	0.03271	-0.00846
res	8.2851	0.7198	265	11.51	<.0001	0.05	6.8679	9.7023	-3.65E-6

Below is the code to graph the data in R, followed by the graph itself.

```
male<-read.table('c:/teaching/f2008 -
bios7711/data/m95t5.txt',header=F,sep="",skip=0)
female<-read.table('c:/teaching/f2008 -
bios7711/data/w95t5.txt',header=F,sep="",skip=0)
plot(male$V1,male$V2,pch=21,cex=0.75,ylim=c(30,65),
xlab="age (years)", ylab="time (minutes)",
main="1995 BB race:  Top 5 males and females at each with nonlinear
fits")
points(female$V1,female$V2,pch=3,cex=0.75)
x=c(8:60)
m=277.23*(x**-0.905)*(exp(0.03198*x)); lines(x,m,lty=1)
f=270.04*(x**-0.8317)*(exp(0.03093*x)); lines(x,f,lty=2)
text(37,50,"Females"); text(52,35,"Males")
```

- Take the derivative of $f(x)$ and set to 0 and solve for x for peak age: $-\alpha_1/\alpha_2$. With the data these turn out to be 28.3 and 26.9 for the men and women, respectively. However, since ages are truncated to the year, we should add ½ year to each to get more accurate decimal number estimates.

- Using extreme value theory models, the estimates were 27.8 and 27.4 for men and women (before adding the ½ year).

- Suppose we were to fit data across several years of Bolder Boulder races, and subjects had race times from multiple years. To model such longitudinal data, PROC NLMIXED can incorporate random effects in nonlinear models too. A simple model would include a random intercept, e.g., $E(Y|x,b_i) = b_i + \alpha_0 x^{\alpha_1} e^{x\alpha_2}$, where $b_i \sim N(0,\sigma_b^2)$. PROC NLMIXED can in fact handle more sophisticated models as well, such as those that allow the alpha parameters in the equation above to be random terms for subjects. However, in that case we would probably need a sufficient number of repeated measures (i.e., repeated races) for subjects to be able to carry out the analysis.

- As another example, consider the FDA experiment.

7 Mixture distributions

- Some distributions are more complex and cannot be modeled well using standard methods. Example (3) presented in Section 2 is one such case, for which the distribution is actually a discrete and continuous mixture. With this distribution, there are at least 3 potential questions of interest here:
  - 1) What is the probability of a non-zero observation?
  - 2) What is the mean of the non-zero values?
  - 3) What is the mean of all values?

- With the cost example, these would be:
  - What is the chance someone will incur a cost?
  - What is the mean of costs for those who do?
  - What is the overall mean, taking into account both sources (those who have some costs, and those who don't)?

- The last question would be of interest to people wanting to know population costs. For such a distribution, a simple transformation and standard analysis probably won't work. If the discrete piece is small, we can approximate the distribution with a continuous model, but if not, then the approximation may be too crude.

- Mixture distributions may be useful even if the distribution is completely discrete or continuous. For example, a zero-inflated Poisson distribution takes a standard Poisson and then adds a binomial random variable such that the probability that the mixed random variable takes on a value of 0 is increased.

- In this section we will re-examine mixture distributions, and in particular we will consider a mixture distribution where the two distributions being mixed are binomial and (approximately) normal.

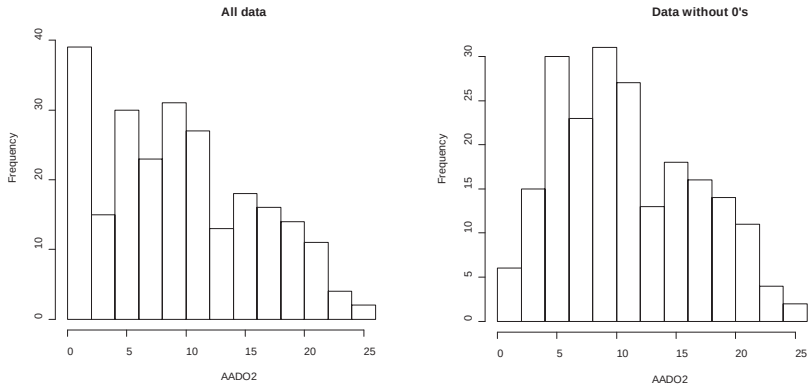
- Example: Patients with Chronic Beryllium Disease have ongoing visits at NJH to monitor their health. One measure taken during their visits is $AADO_2R$ (alveolar-arterial oxygen tension difference at rest; the lower the value, the better the health; see course notes for more detail).

- The goal is to estimate the effect of the disease on $AADO_2R$ over time, after accounting for variables known to be related to it (age, gender, height). One modeling challenge is that subjects come in on different days, have different times between visits, and don't have the same number of visits.

- Data consisted of 243 records measured on 60 subjects; the minimum and maximum numbers of measurements on subjects were 1 and 14, respectively.

- Values of $AADO_2R$ are either positive and continuous or 0. Histograms of the data are shown below. A large portion of the data (14%) are 0. For the positive values, the distribution is somewhat symmetric (lower right). Note that the histograms involve both within and between-subject data.

- The 0 values might actually be due to measurement error for this application. However, for now let's assume they are possible and accurate, in order to demonstrate the methods.



- The distribution with 'All data' above is sometimes referred to as a 'clump-at-0' distribution, where values may be continuous, if positive, or spanning a wide range of counts, but values of 0 are also possible. The $y'=ln(y+c)$ (e.g., $c=1$ or c=half of the smallest non-zero observed value) transformation does not really solve the problem since all zero values will still all be transformed to one value.

- However, if we separate out the 0 values, we are left with a distribution that is not highly skewed (graph on right) and could be modeled with a normal distribution, perhaps after logging.
- There are a few ways of simplifying the analysis.

- One is to consider the binary variable $Z=1$ if $Y>0$ and $Z=0$ if $Y=0$. We can then apply binary models (e.g., logistic regression). There are extensions of logistic regression modeling for repeated measures data, discussed ahead.

- We can also perform an analysis, conditioned on the fact that $Y>0$. In other words, just take the subset of the data where values are not 0 and use the usual normal theory model (often with log-transformed data). However, it is important to note that results refer to the conditional model ('… for values > 0') since the data being analyzed is restricted.

- Here is a synopsis of analyses that break the distribution into pieces. Note the *ntep* is time since first exposure (the variable used to indicate progression of illness); *ageep* is age at date of exercise physiology (ep) test; both of these variables are measured in years.

Naïve approach I: fit data with a linear mixed model (plus random intercept). Here is abbreviated output for the analysis. On the left is the analysis will all of the data, and on the right was the analysis with the positive values only. The analyses used 242 records (one subject had a missing height value for one visit).

The Mixed Procedure

AIC 1550.7

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	id	12.5946
Residual		27.8012

Solution for Fixed Effects

Effect		Estimate	Standard Error	DF	t Value	Pr>|t|
Intercept		-4.3141	15.5846	58	-0.28	0.7829
ntep		**0.2188**	0.07881	179	2.78	0.0061
height		0.1104	0.2204	179	0.50	0.6171
gender	F	2.0957	1.8658	179	1.12	0.2629
gender	M	0	.	.	.	.
ageep		**0.01221**	0.0757	179	0.16	0.8720

Naïve approach II: fit data with a linear mixed model, after removing the 0's (plus random intercept). There were 209 records that could be used in the analysis. In this case, the fit with the REPEATED statement to account for serial correlation did not work (did not converge). Below is the fit with the random intercept.

The Mixed Procedure

AIC 1283.1

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	id	9.3135
Residual		21.0108

Solution for Fixed Effects

Effect		Estimate	Standard Error	DF	t Value	Pr>|t|
Intercept		-20.2177	14.0436	54	-1.44	0.1557
ntep		**0.1596**	0.07823	150	2.04	0.0432
height		0.3007	0.1984	150	1.52	0.1318
gender	F	2.1020	1.6446	150	1.28	0.2032
gender	M	0	.	.	.	.
ageep		**0.1185**	0.07304	150	1.62	0.1068

Notice changes in slope estimates for *ntep* and *age* for the two analyses.

- Models that included the spatial structure to account for serial correlation were also fit, which yielded a slight improvement in the AIC using all of the data; parameter estimates did not change much, however. When restricting data to positive values, the model did not converge when using the spatial power structure. For these reasons, plus to make a more apples-to-apples comparison with models to be presented ahead, we used the model with random intercept only as the final models.

- Naïve approach III: run a logistic regression on the 0 versus non-zero values. Here, I am using Gaussian quadrature to fit a model with a random intercept (i.e., the CS structure is applied to the repeated measures). In previous analysis, this simpler structure was determined to be adequate for the data. (Recall that for non-normal outcomes, there is some limitation on types of structures that can be used. E.g., using a spatial structure with GENMOD/GEE to handle the intermittent data here would be very difficult to employ – it would require massive restructuring of the data.) We can get a comparable model fit using GENMOD/GEE using the exchangeable (i.e., CS) working covariance structure. Both model fits are given below.

Abbreviated output:

The GLIMMIX Procedure

Likelihood Approximation: Gauss-Hermite Quadrature

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	SE
Intercept	id	0.8544	0.7112

Solutions for Fixed Effects

Effect		Estimate	Standard Error	DF	t Value	Pr>|t|
Intercept		12.8978	7.0320	57	1.83	0.0719
ntep		**0.05657**	0.03227	180	1.75	0.0813
height		-0.1285	0.09541	180	-1.35	0.1796
gender	F	0.2269	0.8619	180	0.26	0.7926
gender	M	0	.	.	.	.
ageep		-0.06047	0.03240	180	-1.87	0.0636

The GENMOD Procedure

Distribution Binomial

PROC GENMOD is modeling the probability that y='1'.

Exchangeable Working Correlation: 0.0480770766

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	Z	Pr>|Z|
Intercept		12.3402	6.1176	2.02	0.0437
ntep		**0.0612**	0.0295	2.07	0.0381
height		-0.1290	0.0910	-1.42	0.1564
gender	F	0.3808	0.7569	0.50	0.6149
gender	M	0.0000	0.0000	.	.
ageep		-0.0563	0.0271	-2.08	0.0375

- With both analyses, we find that the odds of non-zero response for aa-gradient increases by approximately 0.06 for each additional year with CBD. However, each additional year in age was associated with a decrease in odds of non-zero response for $AADO_2R$ by roughly the same amount. This age effect is adjusted for progression of illness (as well as gender and height), and similarly, progression of illness is adjusted for other effects in the model. A negative effect for age that indicates an improvement in health with age is not expected and might reflect potential confounders not accounted for in the model.

- An alternative to simplifying the analyses is to use methodology that handles all the data at once. *Two-part model* and *mixture model* methods are two such methodologies.

- One such approach that uses mixture models (and is also related to two-part models) is based on work done by Janet Tooze, who developed methods to fit these two models simultaneously [Tooze JA, Grunwald GK and Jones RH. Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research*. 11(4): 341-355, 2002.] Janet was a former UCD Health Sciences student; the article was based on her dissertation work. In a nutshell, here is a description of the models and methods that she helped to develop:

- A logistic mixed regression model with a random intercept (b_{1i}) was defined for the *occurrence* part of the data (whether the value was 0 or greater than 0).

- A linear mixed model was defined for the *intensity* part of the data (the magnitude of the value, given that it was greater than 0), with random intercept b_{2i}.

- The two random intercepts were assumed to have a bivariate normal density with mean 0 and covariance $\begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right)$.

- The occurrence and intensity models were joined into one likelihood to be fit simultaneously. In her macro, fits for 2 models were given: one setting ρ to 0 (the independence model), and one allowing it to vary between -1 and 1. Goodness-of-fit statistics were included in order to determine whether allowing for correlation approved the model fit.

- For a more thorough description of the models and methods, see Tooze's article. Note that in these models, serial correlation is not taken into account. Thus, the correlation for the repeated measures is accounted for through the random intercepts. For the *AADO$_2$R* data, here is a synopsis of the model fits.

Approach IV: use Tooze's method, employing the macro that uses NLMIXED

Results from Fitting Uncorrelated Model
Response Variable: aado2r

Convergence Status:
Binomial Model - GCONV convergence criterion satisfied.
normal Model - GCONV convergence criterion satisfied.

Parm.	Name	Estimate	Std Err	Prob>|t|
a1	Intercept--bi	4.3151	1.6109	0.0096
b1	ntep--bi	**0.0559**	0.0334	0.0992
c1	ageep--bi	-0.0633	0.0344	0.0708
u1v	Var(Rndm Effect)--bi	1.1915	0.7886	0.1361
a2	Intercept--in	-17.5686	12.4856	0.1649
b2	ntep--in	**0.1579**	0.0744	0.0383
c2	ageep--in	0.1143	0.0698	0.1073
d2	height--in	0.2963	0.1879	0.1205
e2	sex--in	-2.0624	1.5497	0.1886
s2	Residual--in	21.1152	2.4651	0.0000
u2v	Var(Rndm Effect)--in	7.5401	3.2089	0.0223
u12v	Covariance	NA		

Results from Fitting Uncorrelated Model
Response Variable: aado2r

Name	Value	Sum
AIC--bi	189.53	.
AIC--normal	1287.10	1476.62
-2 Log Likelihood--bi	181.53	.
-2 Log Likelihood--normal	1273.10	**1454.62**

Results from Fitting Correlated Model
Response Variable: aado2r

Convergence Status: NOTE: GCONV convergence criterion satisfied.

Estimate	Std Err	Prob>|t|
3.8049	1.6334	0.0233
0.0583	0.0330	0.0826
-0.0555	0.0342	0.1101
1.1675	0.8126	0.1562
-17.5897	11.2282	0.1227
0.1814	0.0724	0.0150
0.0705	0.0689	0.3101
0.3179	0.1682	0.0638
-1.8265	1.3835	0.1920
21.1478	2.4364	<.0001
7.2335	3.1690	0.0261
2.8971	1.1747	0.0166

Results from Fitting Correlated Model
Response Variable: aado2r

Name	Value	Diff in -2ll	p-value
AIC	1470.11	.	.
-2 Log L	**1446.11**	**8.51**	0.0035

- The results of the model fits show a significant improvement by adding the correlation parameter (p=0.0035). This indicates that a subject with a higher (or lower) random intercept in the occurrence model tends to have a higher intercept in the intensity model.

- This can be interpreted as follows: a subject that has nonzero response will also likely have a greater magnitude of response when the value is non-zero.

- Inference for the intensity model is the same as before; that is, it is conditional on the fact that the value is greater than 0. A key parameter of interest is the slope of ntep. Here it is estimated to be 0.18; i.e. for each additional year in time since first exposure, there is an average increase in AADO2r of 0.18. This is greater than the value of 0.16 that was obtained by using the uncorrelated model (or the naïve analysis using the nonzero data and PROC MIXED).

- Note that Tooze's macro employs PROC NLMIXED and not MIXED and may have different default settings, which may explain some of the minor differences in estimates between the uncorrelated model here and the naïve model for the nonzero data.