64/70

BIOS6643 **Homework 5** **2013**

Not to turn in:

A. Using SAS, Excel or R, create the following piecewise polynomial or spline regression functions. First, write the function and then create a graph for it.
   a. A function that is flat below x=10, and then increases by 2 units in y for each additional value once x hits 10.
   b. A function joining 2 quadratic functions, with a knot at x=10 and a clear change point at that knot.
   c. A function with 5 peaks and 6 valleys. (Note: it may be possible to create this using sinusoidal variables, but here, just use spline terms; the peaks and valleys do not need to be even.)

To turn in:

(1) Patients with Chronic Beryllium Disease have ongoing visits at NJH to monitor their health. Once measure taken during their visits is *AADO₂R* (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 *AADO₂R* 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.

a. Consider the model for *AADO₂R* as a function of time since first exposure (*ntep*, in years), age (*ageep*, in years), *height* (inches) and *gender*; the variables *ntep*, *ageep* and *height* are time-varying; a random intercept is included for subjects. Write out a statistical model.

$$Y_{ij} = \beta_0 + \beta_1 ntep_{ij} + \beta_2 ageep_{ij} + \beta_3 height_{ij} + gender_h + b_{0i} + \varepsilon_{ij}$$
$$where\ \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), b_{0i} \sim N\left(0, \sigma_{b_0}^2\right), h = 1,2; i = 1, \dots, n; j = 1, \dots, r_i.$$
$$Here\ r_i\ is\ different\ for\ each\ subjects.$$

9/10 All o.k. but just clarify whether 'gender' is a parameter or variable? (Need a parameter but we usually use greek symbols for those.)

b. Using software, fit the model. (Note: *ageep*, *height* and *gender* are 'a priori' covariates. They stay in the model regardless of their significance.)

Table 1. θ Estimates and fit statistics using linear, VC for **R**i.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
Intercept	id	12.5946
Residual		27.8012

Fit Statistics	
-2 Res Log Likelihood	1546.7
AIC (Smaller is Better)	1550.7
AICC (Smaller is Better)	1550.7
BIC (Smaller is Better)	1554.9

Solution for Fixed Effects

Effect	gender	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
ageep		0.01221	0.07570	179	0.16	0.8720
height		0.1104	0.2204	179	0.50	0.6171
gender	F	2.0957	1.8658	179	1.12	0.2629
gender	M	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
ntep	1	179	7.71	0.0061
ageep	1	179	0.03	0.8720
height	1	179	0.25	0.6171
gender	1	179	1.26	0.2629

10/10

c. Consider the same model as above, but include a spatial power covariance structure for repeated measures. Does the addition of the new $\mathbf{R}_i$ structure improve the model fit?

Table 2. θ Estimates and fit statistics using linear, Spatial Power for **R**i

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
Intercept	id	12.2551
SP(POW)	id	0.04757
Residual		28.3927

Fit Statistics	
-2 Res Log Likelihood	1542.8
AIC (Smaller is Better)	1548.8
AICC (Smaller is Better)	1548.9
BIC (Smaller is Better)	1555.1

Solution for Fixed Effects

Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		-5.4370	15.6379	58	-0.35	0.7293
ntep		0.2117	0.07918	179	2.67	0.0082
ageep		0.007747	0.07609	179	0.10	0.9190
height		0.1326	0.2214	179	0.60	0.5500
gender	F	2.1333	1.8687	179	1.14	0.2551
gender	M	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
ntep	1	179	7.15	0.0082
ageep	1	179	0.01	0.9190
height	1	179	0.36	0.5500
gender	1	179	1.30	0.2551

10/10

Based on comparing the AIC values from Table 1 and Table 2, the spatial power structure for **R** has a slightly lower AIC (1548.8 vs. 1550.7, diff=1.9) indicating a marginal improvement in fit. There are no significant changes to the fixed effect parameters and their standard errors. The variable ntep is statistically significantly associated with the aado2r (outcome) variable (p=0.0082).

d. Using relevant output and formulas (for the model in part c), estimate the correlation between $Y_{ij}|b_{0i}$ and $Y_{ij'}|b_{0i}$ for responses that are (i) 4 months apart, (ii) 1 year apart. Interpret the results for the study.

$\phi^{d_{jk}} = 0.04757^{d_{jk}}$, where d=years. $\sigma_\varepsilon^2 = 28.3927$

i. 4 months = 1/3 year. Thus $0.04757^{\frac{1}{3}} = 0.362336$. Correlation = $\frac{Cov(t1,t2)}{\sqrt{Vt1}*\sqrt{Vt2}} = \frac{\sigma_\varepsilon^2 * 0.362336}{\sqrt{\sigma_\varepsilon^2\phi^0}\sqrt{\sigma_\varepsilon^2\phi^0}} = \frac{\sigma_\varepsilon^2 * 0.780628}{\sigma_\varepsilon^2} = 0.362336.$

ii. 1 year: $\phi^1 = 0.04757. Correlation = 0.04757.$

10/10

The correlation between $Y_{ij}|b_{0i}$ and $Y_{ij'}|b_{0i}$ at 4 months apart is larger than at 1 year apart. This is similar to the decaying correlation structure of AR(1) where the repeated measures.

e. Repeat part d, but for the correlation between Y_{ij} and $Y_{ij'}$ (unconditional responses).

i. 4 months = 1/3 year. Thus $0.3693^{\frac{1}{3}} = 0.717452$. Correlation = $\frac{Cov(t1,t2)}{\sqrt{Vt1}*\sqrt{Vt2}} = \frac{\sigma_\varepsilon^2 * 0.717452}{\sqrt{\sigma_\varepsilon^2\phi^0}\sqrt{\sigma_\varepsilon^2\phi^0}} = \frac{\sigma_\varepsilon^2 * 0.780628}{\sigma_\varepsilon^2} = 0.717452.$

ii. 1 year: $\phi^1 = 0.3693. Correlation = 0.3693.$

5/10 See solutions.

The correlation between the unconditional Y_{ij} and $Y_{ij'}$ at 4 months apart is larger than at 1 year apart. But compared to the above results, the correlation is larger for both cases indicating that the correlation between the responses without the incorporation of a random intercept tends to be greater than when the random intercept components are included.

f. Include a streamlined version of the output and write a few sentences of summary of the results. Make sure to include an estimate the progression of the illness (slope of *ntep* term), with a 95% confidence interval. For *AADO2R*, remember that the lower the value, the better the health. What can be concluded about the progression of the illness over time?

Table 3. Parameter estimates and 95% CI

Solution for Fixed Effects									
Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept		-5.4370	15.6379	58	-0.35	0.7293	0.05	-36.7396	25.8656
ntep		0.2117	0.07918	179	2.67	0.0082	0.05	0.05545	0.3680
ageep		0.007747	0.07609	179	0.10	0.9190	0.05	-0.1424	0.1579
height		0.1326	0.2214	179	0.60	0.5500	0.05	-0.3042	0.5694
gender	F	2.1333	1.8687	179	1.14	0.2551	0.05	-1.5542	5.8207
gender	M	0	.	.	.	.	.	.	.

Based on the analysis of the AADO2R data, the repeated measure covariance structure (**R**) of spatial power structure was chosen (Based on lower AIC compared to VC structure) to indicate that the correlation between responses conditioned on the random intercept were 0.3624 for responses 1 year apart and 0.04757 for responses 4 months apart, indicating decaying correlation of responses between time points farther apart. The choice of the spatial power structure was appropriate because the different time since exposure (ntep) for each individuals can be accounted for along with having the time since exposure variable as a continuous variable, rather than categorical, works well to fit the correlation structure between the time points. The parameter estimate (fixed) is statistically significantly associated with AADO2R (p-value=0.0082) after adjusting for age, gender, and height with the estimate indicating 0.2117 (95% CI: 0.05545,0.3680) increase in AADO2R value (worsening health) for every 1 year increase in time since first exposure. Basically, health is worsening as time increases. 10/10

g. Note that *AADO₂R* actually is a continuous variable with a clump at 0. Such variables occur in real life and can be modeled using 2-part or mixture models. However, examine a histogram of the residuals to determine whether the current model may be acceptable. Comment on what you see. [Aside: in this case, it is likely that the 0's should have been positive since a 0 value does not appear to be theoretically possible for this variable; from what I can see they can be attributed to measurement error, which may have affected other values as well.]

Figure 1. Histogram of Residuals

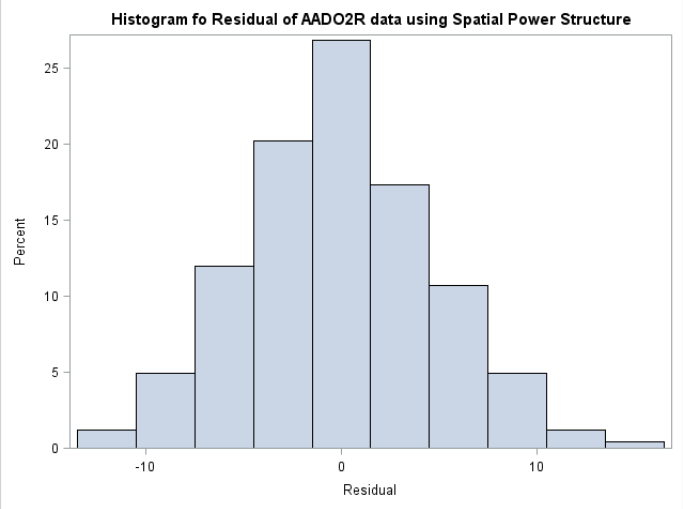
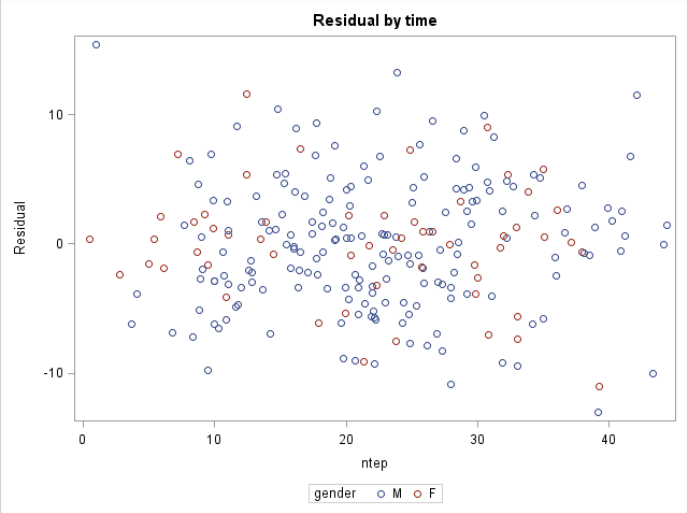


Figure 2. Residual by time



Based on the histogram (Fig 1) it seems that the errors ~ $N(0, \sigma_\varepsilon^2)$ (almost perfectly normally distributed). The residual by time plot (Fig 2) also indicate that there are no real pattern to the residual over time indicating that the model may be appropriate. So at first thought the current model may seem acceptable but if the values are not theoretically possible and the current residual plot is affected by measurement error then analysis would need to account for the potential measurement error using a new $Y_{ij}^* = Y_{ij} + w_{ij}$ where $w_{ij} \sim N(0, \sigma_w^2)$ and the regression equation would be updated accordingly to account for both σ_ε^2 and σ_w^2. Another method to resolve the issue of measurement error would be to use a mixture model where two different distributions (Truncated normal and Normal) would be used for estimation/inference.

10/10 Good.

Appendix

```
libname cbd "C:\Users\kimchon\Downloads";

*load data;
data cbd;
set cbd.be_study;
run;

data temp1;
set cbd;
nptenew = round(ntep);
run;


*fit mixed model;
proc mixed data=cbd;
class id gender;
model aado2r = ntep ageep height gender /outp=out solution;
random intercept / subject=id;
run;

*fit using spatial exponent;
proc mixed data=cbd;
class id gender;
model aado2r = ntep ageep height gender /outp=out solution;
random intercept / subject=id;
repeated / subject=id type= sp(exp)(ntep);
run;


*conditional correlation question;
```

```
*spatial power with random intercept;
proc mixed data=cbd;
class id gender;
model aado2r = ntep ageep height gender /outp=out solution cl;
random intercept / subject=id;
repeated / subject=id type= sp(pow)(ntep) rcorr r;
run;

*Figure to see what's going on;
proc sgplot data=out;
series y=pred x=ntep / group=gender;
run;

*unconditional correlation question;
proc mixed data=cbd;
class id gender;
model aado2r = ntep ageep height gender /outp=out solution;
repeated / subject=id type= sp(pow)(ntep) rcorr r;
run;

*1g: examine histogram of residual and determine if current model acceptable/appropriate;
PROC SGPLOT DATA = out;
 HISTOGRAM resid;
 TITLE "Histogram for Residual of AADO2R data using Spatial Power Structure";
RUN;

proc stdize data= out out=fig2;
  var resid;
run;

*Normal qqplot;
proc univariate data = fig2 noprint;
  var resid;
  qqplot /href = 0 vref = 0 ;
run;

*Residual by time plot;
 ods listing style=listing;

   proc sgplot data=out;
      title 'Residual by time';
      scatter x=ntep y=resid / group=gender;
   run;

      proc sgplot data=out;
      title 'AADO2R by time';
      scatter x=ntep y=aado2r / group=gender;
   run;

   proc sgplot data=out;
   title 'AADO2R by time since exposure';
   series x=ntep y=aado2r / group=id grouplc=gender name='grouping';
   keylegend 'grouping' / type=linecolor;
      run;
```

Project installment #2

Note: the approaches below are to get you start thinking. I am basically asking that some progress be made on your project, so if these are not directly applicable, just move forward and give me an update about where you are at by the middle of next week. If you would like some specific guidance, let me know.

I would just like to see a few pages including a description about what you tried, and some very streamlined output. This does not need to be finalized work; it can be 'work in progress'. You can put code in an Appendix.

- If you have not done so already, explain the type of outcome variable you have and the basic modeling approach you will use for it (e.g., linear mixed model for a continuous/normal outcome; Poisson regression for a count outcome, logistic regression for a binary outcome, etc.).

- If applicable, update or modify the graph you turned in for Installment #1.
- First attempt at modeling the data:
  - If you have a limited number of time points, try modeling time as a class variable. Look at estimates over time obtained from either the LSMEANS or ESTIMATE statements. Does it appear that some smooth function (e.g., linear, quadratic, other) could be used for time? If so, suggest it, but for now you don't need to try it.
  - If you have many time points or data for which subjects had different time points, try modeling time as continuous. Does some more complex model (polynomial, spline) seem like it might work? If so, give it a try.
  - If you do not have time as a main predictor (e.g., maybe you just have nested data), explain an approach to model it, and then try it. Explain what you found.
- Model selection:
  - Do you have clear covariates to include or will some be chosen based on model fit? (If you do any model selection, you can just use a reasonable covariance structure, one that accounts for repeated measures in some way…we will focus on selecting a 'best' covariance structure later.) Go ahead and try some models, although note that I do not want this project to be consumed by model selection of fixed effects (unless it involves the key variable of interest, such as a polynomial or spline function for time…I know some subjects will be working with such models). Remember that if you are using REML in SAS, the AIC does not penalize for fixed-effect parameters. If you just want to specify your model beforehand, that is completely fine.
  - Once you are comfortable with the fixed-effect part of the model, focus on covariance structures. What covariance structures seem like they might work? What makes most sense for your data, modeling the G matrix, R matrix or both? Before you start examining AIC's, list some potential models, and then try fitting them.