BIO 226 Mid-Term Exam

April 9, 2013

Name:

SOLUTIONS

Department:

Instructions

1. There are <u>three</u> questions and you are asked to attempt all three questions.
2. Questions 1 is worth 40 points; Questions 2-3 are worth 30 points each.
3. Please show your work. We will give partial credit.

Question 1 (40 points). This problem considers data from a study of an anesthetic in dogs (source: Johnson and Wichern, 1982). This response was obtained at four occasions on 19 dogs. The dogs were initially given the drug pentobarbital. Each dog was sequentially administered one anesthetic treatment at each timepoint: Treatment 1 at time 1, Treatment 2 at time 2, Treatment 3 at time 3, Treatment 4 at time 4. At each occasion, the response defined as the duration of time between heartbeats, measured in milliseconds, was measured. Assume that the four measurement occasions were equally spaced in time.

Exhibit A shows selected output from longitudinal model that used a saturated structure for the regression (treating time, taking values 1-4, as a categorical variable in the regression model for the mean) and various models for the covariance matrix. The variable dog is an identifying number for the dog (taking the integer values 1 to 19). There is output for models using the following variance/covariance structures: (i) Unstructured, (ii) Toeplitz, (iii) Toeplitz with heterogeneous variances, (iv) compound symmetry with heterogeneous variances, (v) AR(1) with heterogeneous variances.

a) Explain why one might prefer REML over ML estimation of the parameters in a covariance model for these data.

[4 pts] **These data are made up of a small sample of dogs (N=19). We know that the maximum likelihood estimates of the parameters in the covariance model are biased in small samples, and as a result the REML estimates are preferrable in small sample settings.**

b) Consider the unstructured model, the Toeplitz model, and the Toeplitz model with heterogeneous variances for the covariance matrix. Among these three models, indicate which pairs of models have the property that one model is nested in the other, and provide reasoning for your claims.

[6 pts] **The Toeplitz model is nested within the unstructured model. We can see this by recognizing that we obtain the Toeplitz model when we impose the following constraints in the unstructured model:** $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$**,** $\sigma_{12} = \sigma_{23} = \sigma_{34}$**, and** $\sigma_{13} = \sigma_{24}$**.**

The Toeplitz model is nested within the Toeplitz model with heterogeneous variances. We can see this by recognizing that we obtain the Toeplitz model when we impose the following constraints in Toeplitz model with heterogeneous variances: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$**.**

The Toeplitz model with heterogeneous variances is nested within the unstructured model. We can see this by recognizing that we obtain the former from the latter by constraining the correlations $\rho_{12} = \rho_{23} = \rho_{34}$**, and** $\rho_{13} = \rho_{24}$**.**

c) Your scientific collaborator asks you to evaluate whether the Toeplitz covariance structure fits the data adequately, as compared to the Unstructured model. Define the null and alternative hypotheses corresponding to this evaluation, and describe an approach for testing this null hypothesis using the output provided. Obtain the value of the relevant test statistic for testing this null hypothesis. What do you conclude using an 0.05 level of significance?

[6 pts]

H$_0$: The Toeplitz covariance model is adequate for the observed data.

H$_1$: The Toeplitz structure is inadequate, and unstructured covariance model is necessary.

The likelihood ratio statistic testing the hypothesis is LR $= 796.2 - 785.2 = 11.0$ **on df=10-4=6. With a critical value** $\chi^2_6 = 12.59$**, this is not significant (although close), suggesting that the Toeplitz model is adequate for these data when compared against the unstructured covariance model.**

d) Now assume that the Toeplitz structure holds for the correlations in the covariance matrix. Given the information in Exhibit A, can you test whether the residual variances for the outcome differ across the four timepoints, under this Toeplitz assumption for the correlations? If so, do so and describe what you conclude from this analysis at the 0.05 level. If not, describe what additional piece of information you would need to conduct this test.

[6 pts] **Given the Toeplitz structure for the correlations hold, the condition of whether the variances differ across the four timepoints can be tested by comparing the two Toeplitz models with homogeneous and heterogeneous variances. That is, we test the null hypothesis**

H$_0$: The Toeplitz covariance model with homogeneous variances is adequate for the observed data.

H$_1$: The Toeplitz structure with homogeneous variances is inadequate, and the Toeplitz model with heterogeneous variances is necessary.

We have enough information to test this null hypothesis via a likelihood ratio test. The likelihood ratio statistic testing the hypothesis is LR $= 796.2 - 785.5 = 10.7$ **on df=7-4=3. With a critical value** $\chi^2_3 = 7.81$**, we reject the null hypothesis, and conclude that the Toeplitz model with heterogeneous variances provides a significantly better fit to the data than does the Toeplitz model with homogeneous variances.**

e) Considering your findings in (c) and (d), among the Unstructured, Toeplitz, and Toeplitz with heterogeneous variances covariance patterns, which model do you prefer for these data? Briefly justify your answer.

[4 pts] **The results in (c) and (d) suggest that the Toeplitz model with homogeneous variances is preferred to the unstructured covariance, and that the Toeplitz model with heterogeneous variances is preferred to the Toeplitz model with homogeneous variances. Just to close the loop, we test the heterogeneous model against the unstructured model. The likelihood ratio statistic testing the hypothesis is LR** $= 785.5 - 785.2 = 0.3$ **on df=10-7=3. With a critical value** $\chi^2_6 = 7.81$**, this is not even close to significant, suggesting that the Toeplitz model with heterogeneous variances is the preferred covariance model.**

f) Consider now the comparison of the model using compound symmetry with heterogeneous variances (denoted CSH) and the model using AR(1) structure with heterogeneous variances (denoted ARH(1)). Your scientific collaborator claims that these two covariance models fit the data equally well. Justify this claim.

[4 pts] **The CSH and ARH(1) models do fit the data equally well. The models contain the same number of parameters so we can directly compare the values of -2*log(L). These values are 788.2 and 788.0 for CSH and ARH(1), respectively, suggesting an almost identical quality of fit.**

g) The CSH and ARH(1) correlation structures implied by these two models assume quite different correlation structures for the covariance matrix.

(i) Describe the primary difference between these two covariance models.

[4 pts] **The difference between CSH and ARH(1) involves the structure for the correlations. CSH assumes the correlation is equal to a constant value, say ρ, for all pairs of time points. ARH(1) assumes that these correlations decay with the amount of time between measurements, specifically that the correlation between two measurements recorded at times j and k takes the value $\rho^{|j-k|}$.**

(ii) Using the output provided in Exhibit A, explain how these two models can fit the data equally well even though they imply very different assumptions about the covariance pattern.

[6 pts] **Comparing the estimated correlations in the ARH(1) model to the unconstrained estimates in the unstructured model, we see that the estimated correlation structure for ARH(1) tends to overestimate the rate of decay in correlation with increasing time separation.**

Conversely, CSH doesn't allow for decay in correlation with increasing time separation and so underestimates that compared to the unstructured structure.

Therefore, the observed decay structure is between that for ARH(1) and CSH, explaining why both fit the data equally well even though they involve very different assumptions.

Question 2 (30 points). In a recent hip replacement study, 30 patients underwent hip-replacement surgery, 13 males and 17 females. Haematocrit, the ratio of the volume of red blood cells to the volume of whole blood, was supposed to be measured for each patient at week 1, before the replacement (baseline), and then at weeks 2, 3, and 4, after the replacement; there was some missing data. The investigators' primary interest was to determine whether the mean haematocrit response following replacement was similar for men and women.

In the analysis of the hip replacement data, gender was coded as GENDER=F for female and GENDER=M for male. The measurement occasions were coded WEEK=1 for measurements at week 1, WEEK=2 for measurements at week 2, WEEK=3 for measurements at week 3, and WEEK=4 for week 4.

Exhibit B gives partial model fitting information from fitting three models to the mean responses of the hip replacement data, assuming an unstructured covariance. The displayed fit statistics are based on maximum likelihood (ML), whereas the displayed regression coefficient estimates are from the residual maximum likelihood fit (REML). The three regression models, all with gender and gender*time interactions, are:

i. the saturated model (treating WEEK as a categorical variable)

ii. the quadratic model (treating WEEK as a continuous variable centered at its mean value of 2.5)

iii. the linear trend model (treating WEEK as a continuous variable)

a) Describe which pairs of models have the property that one model is nested in the other.

[4 pts] **Model (iii) is nested within both model (ii) and model (i).**

Model (ii) is nested within model (i).

b) Using results in Exhibit B, select a parsimonous, yet adequate, model for the mean. In your answer you should present results that support your choice of model.

[10 pts]

Mean Model	**# Parameters**	**-2 LL**	**LR stat (vs. saturated)**	**df**	**p**
hline Saturated	**8**	**539.2**	**—**	**—**	**—**
Quadratic Trend	**6**	**543.4**	**4.2**	**2**	$p > .10$
Linear Trend	**4**	**578.0**	**38.8**	**4**	$p < .01$

Conclusion: Quadratic trend is defensible, but linear trend is not.

c) Based on the model you selected in part (b), what is the interpretation of the gender main effect?

[6 pts] **Due to the fact that week is centered at 2.5 in the quadratic model, the main effect of gender represents the difference in mean Haematocrit between men and women, at the zero value of the centered week variable, or in other words at week 2.5 = the mean time in the study.**

d) Exhibit B also gives the estimated regression parameters and associated standard errors from fitting the above three models by residual maximum likelihood, assuming an unstructured covariance.

Based on the model you selected in part (b) and the output in Exhibit B, provide an expression for the change in the mean response between the end of the study (time 4) and the beginning of the study (time 1), <u>for males</u>.

[10 pts] **Again remembering we need to center the week variable, this difference is**

$$
\begin{aligned}
\mu_{4F} - \mu_{1F} &= [28.7618 - 0.2975 + (-3.3502 + 0.0239) * (4 - 2.5) + (2.4990 + 1.7331) * (4 - 2.5)^2] - \\
&\quad [28.7618 - 0.2975 + (-3.3502 + 0.0239) * (1 - 2.5) + (2.4990 + 1.7331) * (1 - 2.5)^2] \\
&= (-3.3502 + 0.0239) * (1.5) - (-3.3502 + 0.0239) * (-1.5) \\
&= (3.3502 + 0.0239) * 3.0
\end{aligned}
$$

Question 3 (30 points). In a study of lifestyle and weight gain, investigators are interested in studying how the length of the commute to work affects weight gain in middle age. The investigators have access to data on a cohort of individuals measured at ages 50, 52, 54, 56, and 58. Length of commute was measured once at age 50 as a continuous variable, and assumed to be constant over the period of study. Weight was measured at each visit. Some observations are missing. The participants consisted of both women (n=250) and men (n=305).

a) Describe a model algebraically that would provide an estimate of the (linear) association between length of commute and rate of weight gain during the period of follow-up. Do not include gender in this model.

[10 pts] **This model should include a term for interaction between age and length of commute, as the coefficient of that term represents the effect of length of commute on the rate of weight gain during the eight-year period of follow-up. If we assume that weight increased linearly during the period of the study, the model should also include an intercept, a main effect for length of commute, and a main effect for age. Thus,**

$$\mathbf{Wt}_{ij} = \beta_1 + \beta_2 \mathbf{commute}_i + \beta_3 \mathbf{Age}_j + \beta_4 \mathbf{commute}_i * \mathbf{Age}_j + \epsilon_{ij},$$

where commute$_i$ **represents the measured length of commute for subject** i**, Age**$_j$ **is the age of each participant at measurement occasion** j**, and** ϵ_{ij} **are residual errors.**

b) Weight gain may depend on the length of the commute but the dependence may not be linear. One useful approach to exploring this possibility involves converting length of commute to a categorical variable with five categories and quantifying how rate of weight gain varies across the categories of length of commute.

i) Describe how you would construct such a categorical variable from the measured values of length of commute, and

[10 pts] **One popular approach is to choose cutpoints that create five groups of equal size (quintiles). One can then define a new variable, commuteclass, as a class variable or equivalently, choose a reference group and use indicator variables to identify the four other groups. We define the lowest commute length category as the reference group, and indicator variables X_{2i}, X_{3i}, X_{4i}, and X_{5i} represent dummy variables that indicate whether subject i belongs to commuter class 2, 3, 4, and 5, respectively**

ii) give the algebraic representation for the model which provides estimates of the rate of weight gain for individuals in different commute categories. For simplicity, your model should contain no terms involving gender.

[10 pts] **The model of interest is then written as:**

$$\begin{aligned}\text{Wt}_{ij} &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 \text{Age}_j \\ &\quad \beta_7 X_{2i} * \text{Age}_j + \beta_8 X_{3i} * \text{Age}_j + \beta_9 X_{4i} * \text{Age}_j + \beta_1 0 X_{5i} * \text{Age}_j + \epsilon_{ij}\end{aligned}$$

In this model, the coefficient of age represents the rate of weight gain in the reference group, while the coefficients of the interaction terms between commute categories and age represent the differences in rate of weight gain between each of the other four groups and the reference group.

EXHIBIT A

i. Unstructured Covariance Model

Estimated R Matrix for dog 1

Row	Col1	Col2	Col3	Col4
1	2819.29	3568.42	2943.50	2295.36
2	3568.42	7963.13	5303.99	4065.46
3	2943.50	5303.99	6851.32	4499.64
4	2295.36	4065.46	4499.64	4878.99

Estimated R Correlation Matrix for dog 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.7531	0.6697	0.6189
2	0.7531	1.0000	0.7181	0.6522
3	0.6697	0.7181	1.0000	0.7783
4	0.6189	0.6522	0.7783	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	dog	2819.29
UN(2,1)	dog	3568.42
UN(2,2)	dog	7963.13
UN(3,1)	dog	2943.50
UN(3,2)	dog	5303.99
UN(3,3)	dog	6851.32
UN(4,1)	dog	2295.36
UN(4,2)	dog	4065.46
UN(4,3)	dog	4499.64
UN(4,4)	dog	4878.99

Fit Statistics

-2 Res Log Likelihood	785.2
AIC (smaller is better)	805.2
AICC (smaller is better)	808.8
BIC (smaller is better)	814.6

ii. Toeplitz Correlation Structure with Homogeneous Variances

Estimated R Matrix for dog 1

Row	Col1	Col2	Col3	Col4
1	5540.55	3871.09	3423.33	3631.28
2	3871.09	5540.55	3871.09	3423.33
3	3423.33	3871.09	5540.55	3871.09
4	3631.28	3423.33	3871.09	5540.55

Estimated R Correlation Matrix for dog 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.6987	0.6179	0.6554
2	0.6987	1.0000	0.6987	0.6179
3	0.6179	0.6987	1.0000	0.6987
4	0.6554	0.6179	0.6987	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
TOEP(2)	dog	3871.09
TOEP(3)	dog	3423.33
TOEP(4)	dog	3631.28
Residual		5540.55

Fit Statistics

-2 Res Log Likelihood	796.2
AIC (smaller is better)	804.2
AICC (smaller is better)	804.8
BIC (smaller is better)	807.9

iii. Toeplitz Correlation Structure with Heterogeneous Variances

Estimated R Matrix for dog 1

Row	Col1	Col2	Col3	Col4
1	2829.09	3609.70	2944.50	2258.30
2	3609.70	8141.62	5625.42	4177.57
3	2944.50	5625.42	6870.91	4322.02
4	2258.30	4177.57	4322.02	4805.88

Estimated R Correlation Matrix for dog 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.7521	0.6679	0.6125
2	0.7521	1.0000	0.7521	0.6679
3	0.6679	0.7521	1.0000	0.7521
4	0.6125	0.6679	0.7521	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Var(1)	dog	2829.09
Var(2)	dog	8141.62
Var(3)	dog	6870.91
Var(4)	dog	4805.88
TOEPH(1)	dog	0.7521
TOEPH(2)	dog	0.6679
TOEPH(3)	dog	0.6125

Fit Statistics

-2 Res Log Likelihood	785.5
AIC (smaller is better)	799.5
AICC (smaller is better)	801.3
BIC (smaller is better)	806.1

iv. Compound Symmetric Correlation Structure with Heterogeneous Variances

Estimated R Matrix for dog 1

Row	Col1	Col2	Col3	Col4
1	2875.37	3326.23	3061.63	2638.44
2	3326.23	7883.45	5069.49	4368.76
3	3061.63	5069.49	6679.10	4021.23
4	2638.44	4368.76	4021.23	4960.27

Estimated R Correlation Matrix for dog 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.6986	0.6986	0.6986
2	0.6986	1.0000	0.6986	0.6986
3	0.6986	0.6986	1.0000	0.6986
4	0.6986	0.6986	0.6986	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Var(1)	dog	2875.37
Var(2)	dog	7883.45
Var(3)	dog	6679.10
Var(4)	dog	4960.27
CSH	dog	0.6986

Fit Statistics

-2 Res Log Likelihood	788.2
AIC (smaller is better)	798.2
AICC (smaller is better)	799.1
BIC (smaller is better)	802.9

v. First-order Autoregressive Correlation Structure with Heterogeneous Variances

Estimated R Matrix for dog 1

Row	Col1	Col2	Col3	Col4
1	2848.27	3633.82	2505.39	1565.19
2	3633.82	8199.36	5653.17	3531.70
3	2505.39	5653.17	6893.47	4306.56
4	1565.19	3531.70	4306.56	4758.36

Estimated R Correlation Matrix for dog 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.7519	0.5654	0.4252
2	0.7519	1.0000	0.7519	0.5654
3	0.5654	0.7519	1.0000	0.7519
4	0.4252	0.5654	0.7519	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Var(1)	dog	2848.27
Var(2)	dog	8199.36
Var(3)	dog	6893.47
Var(4)	dog	4758.36
ARH(1)	dog	0.7519

Fit Statistics

-2 Res Log Likelihood	788.0
AIC (smaller is better)	798.0
AICC (smaller is better)	798.9
BIC (smaller is better)	802.7

EXHIBIT B

i. Saturated Model

Fit Statistics (ML)

-2 Log Likelihood	539.2
AIC (smaller is better)	575.2
AICC (smaller is better)	583.7
BIC (smaller is better)	600.4

Solution for Fixed Effects (REML)

Effect	gender	week	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			38.6169	1.0333	28	37.37	<.0001
gender	M		3.8677	1.5466	28	2.50	0.0185
gender	F		0	.	.	.	.
week		4	-8.1757	1.2670	28	-6.45	<.0001
week		3	-7.2560	2.3333	28	-3.11	0.0043
week		2	-8.2904	1.2386	28	-6.69	<.0001
week		1	0	.	.	.	.
gender*week	M	4	-0.6881	1.9342	28	-0.36	0.7247
gender*week	M	3	-5.1710	3.5928	28	-1.44	0.1612
gender*week	M	2	-3.3173	1.8622	28	-1.78	0.0857
gender*week	M	1	0	.	.	.	.
gender*week	F	4	0	.	.	.	.
gender*week	F	3	0	.	.	.	.
gender*week	F	2	0	.	.	.	.
gender*week	F	1	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
gender	1	28	1.79	1.79	0.1807	0.1914
week	3	28	218.54	72.85	<.0001	<.0001
gender*week	3	28	6.69	2.23	0.0826	0.1069

ii. Quadratic Trend Model

Fit Statistics (ML)

-2 Log Likelihood	543.4
AIC (smaller is better)	575.4
AICC (smaller is better)	582.1
BIC (smaller is better)	597.8

Solution for Fixed Effects (REML)

Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		28.7618	1.0371	28	27.73	<.0001
gender	M	-0.2975	1.5775	28	-0.19	0.8518
gender	F	0	.	.	.	.
weekc		-3.3502	0.2415	28	-13.87	<.0001
week2c		2.4990	0.5581	28	4.48	0.0001
weekc*gender	M	0.2039	0.3817	28	0.53	0.5974
weekc*gender	F	0	.	.	.	.
week2c*gender	M	1.7331	0.8421	28	2.06	0.0490
week2c*gender	F	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
gender	1	28	0.04	0.04	0.8504	0.8518
weekc	1	28	289.71	289.71	<.0001	<.0001
week2c	1	28	63.90	63.90	<.0001	<.0001
weekc*gender	1	28	0.29	0.29	0.5932	0.5974
week2c*gender	1	28	4.24	4.24	0.0396	0.0490

iii. Linear Trend Model

Fit Statistics (ML)

-2 Log Likelihood	578.0
AIC (smaller is better)	606.0
AICC (smaller is better)	611.0
BIC (smaller is better)	625.6

Solution for Fixed Effects (REML)

Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		39.5056	0.9728	28	40.61	<.0001
gender	M	1.2510	1.5037	28	0.83	0.4125
gender	F	0	.	.	.	.
week		-2.8682	0.3049	28	-9.41	<.0001
week*gender	M	0.3785	0.4828	28	0.78	0.4396
week*gender	F	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
gender	1	28	0.69	0.69	0.4054	0.4125
week	1	28	123.17	123.17	<.0001	<.0001
week*gender	1	28	0.61	0.61	0.4330	0.4396

$\alpha = 0.05$ critical values for chi-squared distribution, for specific degrees of freedom (df)

df	Critical Value
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.06
8	15.50
9	16.92
10	18.31
11	19.68
12	21.03
13	22.36
14	23.68
15	25.00