BIOS663 Final Exam
May 6th 2010

Instructions: Please be as rigorous as possible in all of your answers and show all your work on the exam paper, but DON'T give irrelevant, albeit correct statements.

Remember that to report a test, you should provide H_0, the test statistic, the degrees of freedom.

You may not consult with anyone except the instructor for clarification of questions. The work you present should be your work alone. Violation of the honor code will be prosecuted (penalties may include failure of the course and expulsion from the university). Please sign the honor code pledge and submit it with your report.

Honor Code Pledge: On my honor, I have neither given nor received aid on this examination.

Name:

Signature:

Date:

1. (25pts) A new drug "B" has been developed to reduce cholesterol level. It was claimed that the new drug is more effective than the old one named "A". In a large scale study, each of these two drugs is tested on 500 patients at 5 doses, with 100 patients per dose, and thus the total sample size is 1000.

   (a) (3pts) First consider the dose variable as a factor with 5 levels, and employ an additive model:

   $$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}.$$

   Using reference cell coding, where $i = 1$, and α_1 models the effect of drug A (drug B is reference); $j = 1, 2, 3, 4$, such that β_j models the effect for dose j (dose 5 is reference); k=1,2, ..., 100, which are patient indices within one cell, and e_{ijk} indicates residual error. If we write this ANOVA model as a regression model: $y = \mathbf{X}b + e$, what is the dimension of y, $\mathbf{X}$, b and e, and for an ANOVA model, what kind of distribution e should follow?

   (b) (4pts) For the model specified in part (a), write the cell mean for each combination of drug and dose in terms of μ, α_i and β_j.

Drug	*Dose*	*Mean*
A	1	
A	2	
A	3	
A	4	
A	5	
B	1	
B	2	
B	3	
B	4	
B	5	

(c) (3pts) For the model specified in part (a), fill the following ANOVA table.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	___	_____	29250	____	<.0001
Error	___	_____	417		
Corrected Total	___	_____			

(d) (3pts) If we model the interaction between dose and drug, the model can be written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where γ_{ij} indicates interaction effects. If we write this ANOVA model as a regression model: $y = \mathbf{X}b + e$, what is the dimension of y, $\mathbf{X}$, b and e

(e) (4pts) Write the cell mean for each combination of drug and dose in terms of μ, α_i, β_j and γ_{ij}. Explain the meaning of interaction effect γ_{11} by comparing the table in question (b) and the table in this question.

Drug	*Dose*	*Mean*
A	1	
A	2	
A	3	
A	4	
A	5	
B	1	
B	2	
B	3	
B	4	
B	5	

(f) (4pts) Let μ_A and μ_B be the overall mean values of cholesterol level for drug A and B, respectively. Write down μ_A and μ_B in terms of α_i, β_j and γ_{ij}. If we want to test $H_0 : \mu_A = \mu_B$, write down H_0 in terms of α_i, β_j and γ_{ij}.

(g) (3pts) Give an example that $\mu_A = \mu_B$, but the effect of drug A and B are not the same for all the doses.

2. (15pts) Following question 1, we consider to include interval type of variables.

(a) (4pts) Now if we model dose as a interval variable, with doses equals to 1, 2, 3, 4, and 5, and fit a model of cholesterol level with additive effect of dose and drug, but no interaction, fill the following ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	___	_____	72780	_____	<.0001
Error	___	_____	416		
Corrected Total	___	_____			

(b) (1pts) Is the model in 2(a) an ANOVA model, an ANCOVA model, or a full model in each cell?

(c) (4pts) Compare the model using dose as a categorical variable and the model using dose as a interval variable by F-test. Please write down H_0, calculate F-Statistic, and give the degree of freedom of the corresponding F-distribution when H_0 is true.

Now we introduce another interval variable "age", and obtained the following output.

Dependent Variable: LDL LDL cholesterol, mg/dL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	177854.2865	35570.8573	92.38	<.0001
Error	994	382744.6685	385.0550		
Corrected Total	999	560598.9550			

R-Square	Coeff Var	Root MSE	LDL Mean
0.317258	15.32973	19.62282	128.0050

Source	DF	Type I SS	Mean Square	F Value	Pr > F
drug	1	123876.9000	123876.9000	321.71	<.0001
dose	1	21681.7710	21681.7710	56.31	<.0001
age	1	26676.8663	26676.8663	69.28	<.0001
drug*dose	1	2526.5193	2526.5193	6.56	0.0106
drug*age	1	3092.2299	3092.2299	8.03	0.0047

Source	DF	Type III SS	Mean Square	F Value	Pr > F
drug	1	188.590646	188.590646	0.49	0.4842
dose	1	4596.699264	4596.699264	11.94	0.0006
age	1	6103.211364	6103.211364	15.85	<.0001
drug*dose	1	2443.995325	2443.995325	6.35	0.0119
drug*age	1	3092.229910	3092.229910	8.03	0.0047

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	98.82355769	3.53245943	27.98	<.0001
drug	3.55006409	5.07268042	0.70	0.4842
dose	2.14467577	0.62072607	3.46	0.0006
age	0.29584942	0.07431096	3.98	<.0001
drug*dose	2.21124424	0.87770366	2.52	0.0119
drug*age	0.30090703	0.10618369	2.83	0.0047

(d) (2pts) Write down the fitted model based on the above output. Is dose treated as categorical or interval variable? Is this model an ANOVA model, an ANCOVA model, or a full model in each cell?

(e) (2pts) Write down the fitted model when drug B is used (the reference level for variable drug), using cholesterol level as response, and using drug and dose as covariates.

(f) (2pts) Write down the fitted model when drug A is used, using cholesterol level as response, and using drug and dose as covariates

3. (20 pts) In a mouse study, we are interested in tumor occurrences of 400 mice from two strains: 200 mice from B6 and 200 mice from Cast. Mice from one strain all share the same genetic background. This is a regression problem with one response, tumor occurrence, and three predictors: mouse strain (a binary variable), body weight (a continuous/interval variable), and activity index (an continuous/interval variable).

   (a) (5pts) In a simplified situation, we record 1 if a mouse has at least one tumor and 0 otherwise. Then tumor occurrence is a binary variable, and the results of a logistic regression is shown below:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	541.990	291.381
SC	545.981	315.330
-2 Log L	539.990	279.381

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	260.6084	5	<.0001
Score	200.6430	5	<.0001
Wald	80.9250	5	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.2166	1.7079	17.8538	<.0001
strain	B6	1	1.9460	1.7079	1.2983	0.2545
weight		1	-0.00269	0.0606	0.0020	0.9646
activity		1	1.7828	0.2931	36.9886	<.0001
weight*strain	B6	1	-0.0217	0.0606	0.1281	0.7205
activity*strain	B6	1	-0.8427	0.2931	8.2642	0.0040

   Please write down the fitted model in the form of $E(y_i) = f(\hat{\beta})$ based on the above SAS output, where $\hat{\beta}$ are the regression coefficient estimates. What is $Var(y_i)$?

(b) (5pts) The result in the previous logistic regression suggest weight is not important, we tried to fit the following smaller model.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	541.990	287.592
SC	545.981	303.557
-2 Log L	539.990	279.592

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	260.3980	3	<.0001
Score	200.4918	3	<.0001
Wald	80.9970	3	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
strain	1	1.6216	0.2029
activity	1	37.1344	<.0001
activity*strain	1	8.3173	0.0039

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.2628	1.1901	37.2428	<.0001
strain	B6	1	1.5155	1.1901	1.6216	0.2029
activity		1	1.7819	0.2924	37.1344	<.0001
activity*strain	B6	1	-0.8433	0.2924	8.3173	0.0039

Compared these two models in part (a) and (b) by a likelihood ratio test. Write down H_0, test-statistic, degree of freedom and the distribution of the test statistic when H_0 is true.

In a follow-up study, we took 20 mice with tumor (10 from strain B6 and 10 from Cast) and 20 mice without tumor (10 B6 + 10 Cast), and measure the expression of a gene that is important in tumor progression at three tissues of each mouse: left forebrain, left hindbrain, and right whole brain. We have altogether (20+20)*3 = 120 measurements of gene expression.

(c) (2pts) Please describe the structure of the 120*120 covariance matrix of these 120 observations. How many elements of this matrix are expected to be 0?

(d) (2pts) Here are the results of one mixed effect model, what kind of covariance structure are assumed for three expression measurements per mouse?

Estimated R Matrix for mouseID 1

Row	Col1	Col2	Col3
1	2.1015	0.6881	0.6881
2	0.6881	2.1015	0.6881
3	0.6881	0.6881	2.1015

Fit Statistics

-2 Res Log Likelihood	417.4
AIC (smaller is better)	421.4
AICC (smaller is better)	421.5
BIC (smaller is better)	424.8

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	11.11	0.0009

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
tumor	1	37	4.43	0.0421
strain	1	37	22.02	<.0001

(e) (3pts) Here are the results of the other mixed effect model, what kind of covariance structure are assumed for the three expression measurements per mouse in this model? Compare this model with previous one by a Likelihood Ratio test, write down test statistic, degree of freedom and the distribution of the test statistic when Null hypothesis is correct.

The Mixed Procedure

Estimated R Matrix for mouseID 1

Row	Col1	Col2	Col3
1	2.4998	1.3469	0.1251
2	1.3469	1.9588	0.5887
3	0.1251	0.5887	1.8423

Fit Statistics

-2 Res Log Likelihood	404.3
AIC (smaller is better)	416.3
AICC (smaller is better)	417.1
BIC (smaller is better)	426.5

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
5	24.21	0.0002

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
tumor	1	37	4.22	0.0471
strain	1	37	23.26	<.0001

(f) (3pts) Someone ignored the fact that these mouse are not independent and did a fixed effect linear regression. Compared the following results with the results from question (e), explain (i) which assumption of general linear regression is violated, (ii) why we see smaller p-values in the fixed effect linear model? (iii) Give a reasonable guess

why type I SS and type III SS in the following output are the same.

Dependent Variable: expression

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	91.9864678	45.9932339	22.26	<.0001
Error	117	241.7472351	2.0662157		
Corrected Total	119	333.7337030			

R-Square	Coeff Var	Root MSE	expression Mean
0.275628	158.7405	1.437434	0.905524

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tumor	1	15.41973044	15.41973044	7.46	0.0073
strain	1	76.56673739	76.56673739	37.06	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tumor	1	15.41973044	15.41973044	7.46	0.0073
strain	1	76.56673739	76.56673739	37.06	<.0001