MS WRITTEN EXAMINATION IN BIOSTATISTICS, PART II

Friday, August 1, 2014: 9:00 AM - 3:00PM

Room: BCBS Auditorium

INSTRUCTIONS:

- This is an **OPEN BOOK** examination.
- Submit answers to **exactly 3** out of 4 questions. If you submit answers to more than 3 questions, then only questions 1-3 will be counted.
- Put answers to different questions on **separate sets of paper**. Write on **one side** of the sheet only.
- Put your code letter, **not your name**, in the upper right corner of each page, and do that just before turning in your exam. Your code is highly confidential information. Sharing it with anyone is a violation of the Honor Code.
- Return the examination with a **signed Honor Pledge form**, separate from your answers.
- You are required to answer **only what is asked** in the questions and not to tell all you know about the topics.

1. An investigator studying Type II diabetes brings you a dataset and asks for help analyzing the data. The investigator tells you he has drawn a population-based sample of elderly people (aged 65–90) in a particular community. Using blood samples and information about medication use, the investigator classified each person as diabetic or not diabetic. He also used height and weight measurements to calculate BMI (body mass index) and classified those with a BMI of at least 30 kg/m^2 as being obese. A cross-classification of diabetic and obesity status yielded:

	Diabetic	Not diabetic
Obese	250	140
Not obese	260	370

(a) What study design did the investigator (supposedly) use?

(b) Estimate the proportion of elderly people in this community who are obese and give an associated 95% confidence interval.

(c) Estimate the relative risk of diabetes for obese people relative to non-obese people and give an associated 95% confidence interval.

(d) Is diabetes associated with obesity?

In looking at the summary data you noticed that the numbers of diabetics and non-diabetics are exactly the same, which seems suspicious. You probed further and the investigator admitted that he had first identified a group of people who had diabetes and then recruited an equivalent number of people without diabetes.

(e) What is the study design that the investigator now appears to have used?

(f) With this new information about the design, estimate an appropriate measure of association between obesity and diabetes and provide an associated 95% confidence interval.

With further probing, it transpired that for each diabetes case the investigator had identified and included one person of the same age and gender who did not have diabetes.

(g) What is the study design that the investigator actually used?

(h) What additional information is needed in order to be able to calculate an appropriate measure of association for this study design?

The investigator is also interested in the effect of other factors, such as age and gender, on risk of diabetes. You managed to persuade him that this was not feasible with his current design. With your assistance, the investigator drew a population-based sample from the elderly in the same community. The data were then analyzed using PROC LOGISTIC in SAS, with diabetic status as the outcome. Below is the output from one of the analyses. The variable OBESE takes on values 0 (if not obese) or 1 (if obese), MALE is 1 (if male) or 0 (if female), OBESEMALE is 1 if a person is both male and obese and 0 otherwise, and AGE is the person's age in years.

```
             Analysis of Maximum Likelihood Estimates


                               Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq


Intercept     1     -2.2471      0.4226       28.2795        <.0001
OBESE         1      1.1137      0.0759      215.4644        <.0001
MALE          1      0.3550      0.0730       23.6318        <.0001
OBESEMALE     1     -0.2709      0.1149        5.5566        0.0184
AGE           1      0.0132     0.00552        5.7587        0.0164
```

(i) Does the association between obesity and diabetes differ by gender?

(j) For males and females separately, use the results above to estimate the odds ratio for the association between obesity and diabetes, adjusted for age. (You do not need to provide confidence intervals.)

(k) How does the odds of diabetes differ between two non-obese males if one is 5 years older than the other? (Here too you do not need to provide a confidence interval.)

Points: (a) 1, (b) 3, (c) 4, (d) 2, (e) 2, (f) 4, (g) 1, (h) 2, (i) 1, (j) 3, (k) 2.

2. We want to compare the treatment effects of three doses of one drug in terms of reduction of cholesterol level. The values of the response variable, i.e., the reduction of cholesterol level before and after taking the drug, are recorded in two batches. Let X_1 be the indicator of the second batch, i.e., $X_{1i} = 0$ or 1 if the i-th individual is from the 1st or the 2nd batch, respectively. There are three dose levels: 0, 100, and 200 mg. Let X_2 and X_3 be the indicators of doses 100 and 200, respectively. The total sample size is 90, with the following layout of doses and batches.

		Dose (mg)		
		0	100	200
Batch	1	20	20	20
	2	10	10	10

(a) First consider dose as a factor with 3 levels, and consider the following two way ANOVA model:

$$Y = \alpha_0 + \alpha_1 X_1 + \sum_{j=2}^{3} \beta_j X_j + \sum_{j=2}^{3} \gamma_j (X_1 X_j) + e,$$

where $e \sim N(0, \sigma^2)$, and $X_1 X_j$ is the product of X_1 and X_j. Write the cell mean for each combination of batch and dose (as in the table below) in terms of the parameters that appear in the above model, and interpret γ_2.

Batch	*Dose*	*Mean*
1	0	
1	100	
1	200	
2	0	
2	100	
2	200	

(b) Fitting the above model, we obtain the following estimates and ANOVA table

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    104.466      4.208  24.827  < 2e-16
batch           -6.093      7.288  -0.836  0.40551
dose100         14.919      5.951   2.507  0.01410
dose200         12.471      5.951   2.096  0.03912
batch:dose100    3.299     10.307   0.320  0.74971
batch:dose200   35.042     10.307   3.400  0.00103
```

```
           Df  Sum Sq Mean Sq F value    Pr(>F)
batch       1   894.4   _____  ______  0.115759
dose        2  9060.2   _____  ______ 1.413e-05
batch:dose  2  4992.2   _____  ______  0.001479
Residuals  84 29745.1   _____
```

Complete the ANOVA table. Are the results in the ANOVA table based on added in order test or added at last test? and why?

(c) Next we fit a model using dose as numerical (interval) covariate. Here are the estimates and ANOVA table

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.36048    3.87267  27.723  < 2e-16
batch       -10.83386    6.70766  -1.615  0.10994
dose          0.06235    0.03000   2.079  0.04063
batch:dose    0.17521    0.05196   3.372  0.00112
```

```
            Df  Sum Sq Mean Sq F value    Pr(>F)
batch        1   894.4   894.4  2.4848  0.118621
dose         1  8749.4  8749.4 24.3078 3.956e-06
batch:dose   1  4093.2  4093.2 11.3718  0.001119
Residuals   86 30955.0   359.9
```

If we write the model for the data collected in this study in matrix form as $\boldsymbol{y} = \boldsymbol{X}\eta + \boldsymbol{e}$, what are the dimensions of $\boldsymbol{y}$, $\boldsymbol{X}$, η and $\boldsymbol{e}$? In the framework of the general linear hypothesis, $\theta = C\eta = \theta_0$, specify C and θ_0 for testing the hypothesis that the regression coefficients for dose and for the dose by batch interaction are equal. Explicitly specify the null hypothesis, the test statistic, its distribution under the null hypothesis and the degrees of freedom. You do not need to calculate the test statistic.

(d) Compare the model with dose as a numerical variable to the model with dose as a factor (categorical) using an F-test. Specify the null hypothesis, calculate the test statistic and specifiy its degrees of freedom.

(e) Now instead of assuming $e \sim N(0, \sigma^2)$, we assume $e \sim N(0, 2\sigma^2)$ for batch 1 and $e \sim N(0, \sigma^2)$ for batch 2. In other words, the residual variance of batch 1 is twice that of batch 2. To account for such heterogenuous residual variance, we can estimate regression coefficients using a weighted least squares approach in which we minimize the objective function $\sum_{i=1}^{n} w_i (y_i - \boldsymbol{X}_i \eta)^2$, where $w_i = 1$ for batch 1 and $w_i = 2$ for batch 2. In other words, we want to minimize the objective function $(\boldsymbol{y} - \boldsymbol{X}\eta)^T \mathbf{W} (\boldsymbol{y} - \boldsymbol{X}\eta)$, where $\mathbf{W}$ is a diagonal matrix with diagonal elements $\{w_i\}$. Show that the estimate of η that minimizes this objective function is $\widehat{\boldsymbol{\eta}} = (\boldsymbol{X}^T \mathbf{W} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \mathbf{W} \boldsymbol{y})$.

(f) Derive the mean and the covariance of $\widehat{\boldsymbol{\eta}}$. That is, derive $E[\widehat{\boldsymbol{\eta}}]$ and $\mathrm{Cov}(\widehat{\boldsymbol{\eta}})$.

Points: (a) 5, (b) 3, (c) 4, (d) 3, (e) 5, (f) 5.

3. We have a random sample of 8 subjects from a certain population. Consider the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is blood pressure (mm Hg), $\mathbf{X}$ includes intercept, age (years) and body weight (lbs), in columns 1–3, respectively. Specifically,

$$\mathbf{y} = \begin{bmatrix} 123 \\ 129 \\ 134 \\ 132 \\ 95 \\ 113 \\ 144 \\ 106 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 20 & 138 \\ 1 & 26 & 150 \\ 1 & 21 & 142 \\ 1 & 25 & 120 \\ 1 & 27 & 142 \\ 1 & 23 & 121 \\ 1 & 29 & 138 \\ 1 & 28 & 132 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \text{and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_8 \end{bmatrix} \sim N(0, \sigma^2 \mathbf{I})$$

- The corrected total sum of squares of $\mathbf{y}$ is 1844.
- $(\mathbf{X}^T\mathbf{X})^{-1} =$

```
            intercept        age        weight
intercept    30.321481   -0.29295    -0.169229
age          -0.29295     0.0134259  -0.000303
weight       -0.1692229  -0.000303    0.0013058
```

- $\hat{\sigma}^2 = 361.33$.

(a) Is each of the following statement correct or not? If it is not correct, please explain why it is wrong and try to correct it.

i. $\boldsymbol{\beta}$ are statistics.

ii. $\boldsymbol{\epsilon}$ are parameters.

iii. $\mathbf{y}$ is a random variable following multivariate normal distribution with mean value $\mathbf{0}_{8\times 1}$ and covariance $\sigma^2\mathbf{I}_{8\times 8}$.

iv. $\hat{\sigma}^2$ is a random variable.

v. ϵ_1 is independent with ϵ_2.

(b) Fill in the following table and in particular show how you computed the standard error estimates.

		Standard		
Parameter	Estimate	Error	t value	Pr(>|t|)
intercept	118.5656			0.309
age	-0.5921			0.799
weight	0.1342			0.853

(c) Calculate an estimate of the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$.

(d) Test $H_0 : \beta_0 = \beta_1 = \beta_2$ using the general linear hypothesis approach: $H_0 : \theta = C\eta = \theta_0$. Specify C and θ_0, calculate the test statistic and specify its null distribution and the corresponding degrees of freedom. You do not need to calculate the p-value.

(e) What is the interpretation of β_0, β_1, and β_2. Does β_0 have a meaningful interpretation? Explain why. If not, how would you fix this problem?

(f) Suppose we are interested in the event of blood pressure being larger than 120 mm Hg. Let $\tilde{Y}_i = 1$ if $Y_i > 120$, and $\tilde{Y}_i = 0$ otherwise, $i = 1, \cdots, 8$. Let $p_i = Pr(Y_i > 120)$. Is p_i a parameter or a statistic? What is the distribution of $\tilde{Y}_i$? Give expressions for the mean and variance of $\tilde{Y}_i$ (you do not need compute them numerically).

(g) Estimate the odds ratio relating blood pressure being over 120 mm Hg and weight being over 132 lbs. Also give a 95% confidence interval.

Points: (a) 5, (b) 3, (c) 2, (d) 5, (e) 3, (f) 4, (g) 3.

4. An oncologist wanted to compare the risk of toxicity between two cytotoxic drugs, A and B. Each drug is given over two cycles (each cycle lasts about 3 weeks), and there is interest in comparing cycles as well.

A random sample of 100 patients were assigned at random so that 50 patients receieved drug A and the other 50 received drug B. Each patient was given two cycles of treatment, and toxicity, as a binary outcome, was recorded for each cycle.

The outcomes observed on drug A were as follows:

	No cycle 2 toxicity	Cycle 2 toxicity
No cycle 1 toxicity	48	2
Cycle 1 toxicity	7	3

Show all your computations, not just the final answer.

(a) Test the null hypothesis that, for drug A, toxicity in cycle 1 is independent of toxicity in cycle 2. Give the p-value (either the actual value or a range, e.g. $0.6 < p < 0.7$).

(b) State the assumptions required in (a), and whether you think they are satisfied or not (and justify your answer).

(c) Test the null hypothesis that, for drug A, the risk of toxicity is the same in cycles 1 and 2. Give the p-value (again, actual or a range).

(d) The data from drug B were as follows:

	No cycle 2 toxicity	Cycle 2 toxicity
No cycle 1 toxicity	34	1
Cycle 1 toxicity	6	9

Assuming that the odds ratio relating toxicity in cycle 1 to toxicity in cycle 2 is the same for the two drugs, estimate that odds ratio.

(e) The data from *only* cycle 1 were used to fit a logistic regression model (for example, using SAS PROC LOGISTIC) with toxicity as response and an indicator of drug B (drug_B: 1 if B, 0 if A) as a covariate,

$$\text{logit}P(\text{toxicity}) = \beta_1 + \beta_2\text{drug\_B}.$$

To the extent possible, fill in the following table of results from the logistic regression model.

	Parameter Estimate	S.E. Estimate
intercept	???	???
drug_B	???	???

(f) The investigator suggested using all the data (from *both* cycles) in PROC LOGISTIC to fit the model,

$$\text{logit}P(\text{toxicity}) = \gamma_1 + \gamma_2\text{drug\_B} + \gamma_3\text{cycle\_2},$$

where cycle_2 is 1 for cycle 2 and 0 for cycle 1. Comment on this approach.

(g) Use a hypothesis test to compare the distribution of toxicity in both cycles between drugs A and B. Give the p-value (actual or a range).

Points: (a)–(c) 3 each, (d)–(g) 4 each.