MS WRITTEN EXAMINATION IN BIOSTATISTICS, PART II

Monday, August 12, 2013: 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 the 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.
- 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. Type II diabetes is usually diagnosed by measuring glucose in blood samples. Blood is drawn in the morning after an overnight fast, so as to reduce the effect of short-term fluctuations in glucose that may occur after a meal or snack. Plasma is separated from other components of blood and glucose is measured in the plasma. A fasting plasma glucose value $\geq$126 mg/dl is regarded as evidence of type II diabetes.

   In large studies, plasma glucose is typically assayed on an automated analyzer. The calibration of the machine needs to be checked regularly to ensure that consistent results are produced. Otherwise there may be gradual or sudden changes in the measurements, for instance because of a problem with the machine or batch to batch variation in the reagents used in the assay.

   In a large epidemiology study, there is concern that there have been changes over time in the plasma glucose measurements, in particular because part of the way through the study an old machine had to be replaced by a newer model. It is also thought that, because of its age, the old machine was having increasing problems with calibration in the months leading up to its replacement. In order to investigate possible changes in calibration, the investigators have supplied data for a 12-month period consisting of the last 8 months in which the old machine was being used and the first 4 months using the newer machine.

   Let n_i denote the number of people whose samples were assayed in month i and g_{ij} the plasma glucose value for person j in month i. In the table below $\sum_j$ indicates summation over the n_i values of g_{ij} in month i.

   For any statistical tests requested below, conduct a two-sided test using $\alpha = 0.05$. Also, state explicitly any assumptions you need to make.

Month	n_i	$\sum_j i \cdot g_{ij}$	$\sum_j g_{ij}$	$\sum_j g_{ij}^2$
1	100	10659.35	10659.35	1226655.04
2	200	44888.27	22444.14	2706247.02
3	900	307721.79	102573.93	12398309.96
4	400	177226.66	44306.67	5221782.23
5	800	447022.65	89404.53	10611519.70
6	500	337366.85	56227.81	6730157.41
7	300	240296.88	34328.13	4170834.86
8	300	274306.00	34288.25	4139033.98
9	400	386044.97	42893.89	4875292.02
10	360	384625.43	38462.54	4326664.83
11	160	184535.83	16775.98	1856643.49
12	200	259033.98	21586.17	2455526.47
$\sum_{i=1}^{8}$	3500	1839488.4	394232.8	47204540
$\sum_{i=9}^{12}$	1120	1214240.2	119718.6	13514127
$\sum_{i=1}^{12}$	4620	3053728.7	513951.4	60718667

In parts (a)–(c) consider just the initial 8 months during which the old machine was being used.

(a) Write a statistical model to use to determine if there is a linear trend in glucose values over time. Explain the meaning of the parameters in your model.

(b) Estimate the parameters in the model. You may assume that $\hat{\sigma}^2 = s_{y \cdot x}^2 = 800.0$.

(c) Test whether there is a linear trend over time in the initial 8 months.

(d) Based on the results to this point, predict the mean fasting glucose value for months 9-12.

Below is part of the SAS output from a linear regression model using data from all 12 months.

```
Number of Observations Used        4620

Root MSE             27.64782    R-Square     0.0040
Dependent Mean      111.24489    Adj R-Sq     0.0038

                        Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|


Intercept     1      114.89176        0.93894     122.36      <.0001
month         1       -0.60913        0.14135      -4.31      <.0001
```

(e) Considering the whole 12-month period and the model implied by the regression output, does the mean fasting glucose value vary significantly over time? If so, describe how the mean changes over time.

(f) Now consider just the effect of the change in machine. Does the mean fasting glucose differ between the two machines?

(g) Assuming there is no change over time in months 9-12, what is the interpretation of the effect of month in the model in part (e)?

(h) The study is still on-going and some participants will have their fasting plasma glucose assayed in month 13. Predict the fasting plasma glucose value for a person seen in month 13 and give an interval that has probability 0.95 of containing that value.

(i) In order to use the glucose levels in epidemiologic analyses, the investigators are considering adjusting the measured values to eliminate the effects of changes in machine or over time. What adjustment(s) should they make?

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

2. An investigator at UNC conducted a survey of UNC students both before and after construction of a new exercise trail. Before the trail was constructed, she determined the baseline physical activity levels of a number ofstudents. After construction of the trail, she interviewed the same group of students about their physical activity levels (after construction of the trail) and collected information about the weather at the time of the second interview.

   Short descriptions of the variables of interest are provided below.

   - PAPOST: Average physical activity, measured in minutes per week, after construction of the trail.
   - PAPRE: Physical activity, measured in minutes per week, before construction of the trail (baseline).
   - RAIN: Number of days with rain in the month of the post-construction interview.
   - COLD: Number of days with high temperature of 40 degrees F or below in the month of the post-construction interview.
   - DEWPOINT: Monthly average dewpoint for the month of the post-construction interview. (Higher dewpoints indicate greater humidity.)
   - HOT: Number of days with high temperature of 80 degrees F or above in the month of the post-construction interview.

   The investigator fit the following model, with data centered as indicated, to the physical activity data:

$$\begin{aligned}\mathrm{E}[PAPOST] &= \beta_0 + \beta_1 PAPRE + \beta_2(RAIN-5) + \beta_3(COLD-1) + \beta_4(COLD-1)*(RAIN-5) \\ &\quad + \beta_5(DEWPOINT-60) + \beta_6(HOT-3) + \beta_7(HOT-3)*(DEWPOINT-60).\end{aligned}$$

   Selected SAS output is provided below.

```
                          The GLM Procedure
Dependent Variable: PAPOST
                                      Sum of
Source                     DF        Squares    Mean Square   F Value

Model                    ????     6467394.63      ????          ????
Error                     293    18530998.39      ????
Corrected Total           ????      ????

                    Source                 Pr > F

                    Model                  <.0001
                    Error
                    Corrected Total

        R-Square     Coeff Var      Root MSE    TMODTM_2 Mean
        0.258712      126.4789      251.4870         198.8372

Source                     DF      Type I SS    Mean Square   F Value

PAPRE                       1    5604896.245    5604896.245     88.62
RAIN                        1      63863.574      63863.574      1.01
COLD                        1       7939.540       7939.540      0.13
COLDRAIN                    1       7409.162       7409.162      0.12
DEWPOINT                    1     186454.304     186454.304      2.95
HOT                         1      32706.907      32706.907      0.52
HOTDEWPOINT                 1     564124.903     564124.903      8.92
```

```
                    Source                 Pr > F

                    PAPRE                  <.0001
                    RAIN                   0.3158
                    COLD                   0.7234
                    COLDRAIN               0.7324
                    DEWPOINT               0.0870
                    HOT                    0.4726
                    HOTDEWPOINT            0.0031

Dependent Variable: PAPOST

Source                      DF    Type III SS    Mean Square   F Value

PAPRE                        1    5112244.359    5112244.359     80.83
RAIN                         1      81313.633      81313.633      1.29
COLD                         1     331971.378     331971.378      5.25
COLDRAIN                     1     181222.767     181222.767      2.87
DEWPOINT                     1     523551.221     523551.221      8.28
HOT                          1     594747.218     594747.218      9.40
HOTDEWPOINT                  1     564124.903     564124.903      8.92

                    Source                 Pr > F

                    PAPRE                  <.0001
                    RAIN                   0.2578
                    COLD                   0.0227
                    COLDRAIN               0.0916
                    DEWPOINT               0.0043
                    HOT                    0.0024
                    HOTDEWPOINT            0.0031
```

```
                                        Standard
Parameter             Estimate             Error    t Value    Pr > |t|

Intercept         -452.3872405       175.3766154      -2.58      0.0104
PAPRE                0.6548320         0.0728349       8.99      <.0001
RAIN                -6.6021562         5.8226369      -1.13      0.2578
COLD               -33.7101559        14.7138403      -2.29      0.0227
COLDRAIN            -5.3415030         3.1555328      -1.69      0.0916
DEWPOINT            56.1246214        19.5069529       2.88      0.0043
HOT               -275.9392466        89.9834963      -3.07      0.0024
HOTDEWPOINT         25.7356920         8.6171532       2.99      0.0031
```

Based on this output, you may or may not be able to test all of the following hypotheses. If you are not able to test a hypothesis, explain what additional information is necessary in order to conduct the test.

(a) Fill in the missing parts with ???? in the above ANOVA table.

(b) Report a test of the hypothesis that the number of days with temperature at or above 80 degrees F in the month preceding the interview is unrelated to post-construction physical activity.

(c) Report a test of the hypothesis that the prior physical activity levels are unrelated to physical activity after construction of the trail. (If possible, report tests and interpretations for both the added-in-order and added-last tests, and compare the results of the two tests in terms of the subject matter.)

(d) Report a test of the hypothesis that the number of days with temperature at or below 40 degrees F in the month preceding the interview is unrelated to post-construction physical activity.

(e) Discuss the effects of prior physical activity levels, rain, humidity (measured by the dew point), the number of cold days, and the number of hot days on post-construction physical activity levels. Be sure to use clear language that a newspaper reporter (with no statistics expertise) writing a story about the new trail could understand.

Points: (a) 6, (b) 5, (c) 5, (d) 5, (e) 4.

3. A multiple logistic regression was fitted to data from a cardiovascular study which invetigates several potential risk factors, such as age, BMI, blood pressure and cigarette smoking for sudden death in women. Here the response is sudden death in females without prior coronary heart disease. Below is the partial output from the logistic regression:

```
variables                Coefficient            Standard Error
intercept                 -15.3
Blood Pressure(mm Hg)       0.002                 0.01
BMI                         0.06                  0.02
Smoking (cigarettes/day)    0.007                 0.02
Age (years)                 0.08                  0.02
```

(a) Assess the statistical significance of each individual risk factor and explain the practical implications of your findings.

(b) What is the odds ratio of suddent death between two women who are 10 year apart in age with the other risk factors same.

(c) Provide a 95% confidence interval for the odds ratio in part (b).

(d) Estimate the probability of sudden death for a 60 year old woman with systolic blood pressure of 110 mm Hg, BMI of 30 who smokes 20 cigarettes per day. Intepret the result.

(e) Do you have enough information to construct a 95% confidence interval for the above probability? If yes, construct one. If not, explain why.

Points: (a) 6, (b) 5, (c) 5, (d) 5, (e) 4.

4. A study was designed to investigate the effect of two treatments on weight reduction. The two treatments were dieting and exercise. There was also interest in whether the two treatments worked better, or possibly worse, together than when given separately. A random sample of 100 subjects seeking help for weight reduction were randomized to the two treatments so that 25 received dieting alone, 25 received exercise alone, 25 received both, and 25 received neither. The response was weight loss in kilograms after one month. Covariates were defined as follows: x_1 is an indicator (0=no, 1=yes) for dieting, x_2 is an indicator for exercise, and $x_3 = x_1 x_2$. Several linear regression models were fitted and the results are summarized in following table.

Model number	Estimated model	RSS	Residual d.f.
1	9.88	3654	99
2	$6.99{+}?x_1$	2821	98
3	$7.52{+}?x_2$	3097	98
4	$4.63 + 5.77x_1 + 4.72x_2$	2264	97
5	$4.69 + 5.66x_1 + 4.61x_2 + 0.23x_3$	2263	96

In what follows, show your work and justify your conclusions. Perform hypothesis tests at the 2-sided 0.05 level. Do not compute p-values unless they are explicitly requested. Specify distributions of test statistics (and d.f. when applicable).

(a) If possible, fill in the two values displayed as "?" in the table above. If not possible, state so and explain why.

(b) Test the hypothesis that diet and exercise neither enhance nor antagonize each other with respect to their effect on weight reduction. State what assumptions are required.

(c) Operating under Model 4, compute a 95% confidence interval for the effect of dieting.

For model 4, the matrix $(X'X)^{-1}$ is

$$\begin{bmatrix} 0.03 & -0.02 & -0.02 \\ -0.02 & 0.04 & 0 \\ -0.02 & 0 & 0.04 \end{bmatrix}.$$

(d) Operating under Model 4, compute a 95% confidence interval for the expected weight reduction in a subject who is dieting and following the exercise prorgram.

(e) One issue in such studies is compliance. "Compliance" means that subjects actually follow the prescribed treatment, and non-compliance means the opposite. The investigators suspected that compliance with dieting was not similar to compliance with exercise. In order to investigate this issue, the 25 subjects who were assigned to both dieting and exercise were asked to maintain a diary of how well they followed the prescribed diet and exercise. The diary data were used to classify the compliance of each subject to diet and to exercise as either "good" or "poor". The results are summarized in the following table.

	Good exercise compliance	Poor exercise compliance
Good diet compliance	14	1
Poor diet compliance	9	1

Test the hypothesis that compliance was similar for diet and exercise. Report the p-value.

(f) Write a short summary (a short paragraph) of the above results in a simple non-technical language (for example, for a local newspaper).

Points: (a) 3, (b) 3, (c) 4, (d) 6, (e) 6, (f) 3.