Data Analysis in the Health Sciences

Final Exam 2009 – EPIB–621

Student's Name: ____

Student's Number: ____

INSTRUCTIONS

This examination consists of 8 questions on 16 pages, including this one. Tables of the normal distribution are provided on the last page. Please write your answers (NEATLY) in the spaces provided. Fully explain all of your answers. Each question is worth 10 points, for a total of 80.

1. __________
2. __________
3. __________
4. __________
5. __________
6. __________
7. __________
8. __________

Total (out of 80) __________

1. It is hypothesized that living near major highways may increase exposure to air pollutants, which in turn may lead to increased rates of rheumatoid arthritis (RA). A study of 90,000 subjects is carried out, and the following two-by-two table of data is collected:

	Distance from nearest major highway (in meters)	
RA	less than or equal to 50	50 or greater
Yes	150	700
No	9850	79300

(a) Carry out an appropriate two-sided test to determine whether there is a relationship between distance from a major highway and RA. State the null and alternative hypotheses, carry out the test, and provide a conclusion.

(b) Calculate a 95% confidence interval for the difference in RA rates for the two distance groups ($\leq 50\ m$ and $> 50\ m$). State your conclusion based on this confidence interval.

2. In a study of walking habits, the following linear regression model is fit to a set of data:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4$$

where:

$$\begin{aligned}
Y &= \text{the mean daily number of steps walked,} \\
& \text{averaged over a two week period for each subject} \\
X_1 &= \text{subjects age in years} \\
X_2 = 1 &\rightarrow \text{indicates measurements taken in fall season, 0 otherwise} \\
X_3 = 1 &\rightarrow \text{indicates measurements taken in winter season, 0 otherwise} \\
X_4 = 1 &\rightarrow \text{indicates measurements taken in spring season, 0 otherwise}
\end{aligned}$$

After analysis of data on 400 subjects (100 subjects in each season) aged 20 to 50 years old, the following point estimates are calculated:

$$\begin{aligned}
\hat{\alpha} &= 20000 \\
\hat{\beta}_1 &= -200 \\
\hat{\beta}_2 &= -100 \\
\hat{\beta}_3 &= -300 \\
\hat{\beta}_4 &= -50 \\
\hat{\beta}_5 &= -100 \\
\hat{\beta}_6 &= -200 \\
\hat{\beta}_7 &= -100
\end{aligned}$$

(a) Provide an interpretation for β_7.

(b) How many steps per day would you predict an average 30 year old person would take in the fall season?

(c) How many fewer steps per day would an average 30 year old person take in the winter compared to the summer season?

3. A group of researchers are evaluating a new treatment for migraine headaches. A clinical trial is designed to compare new Drug A to the standard treatment, Drug B. Two hundred subjects are randomized to Drug A ($n = 100$) or Drug B ($n = 100$) at the onset of their next migraine headache, and each subject reports whether their headache is gone ($Y_i = 1$) three hours later or not ($Y_i = 0$).

The researchers run the following logistic regression model on their data

$$\text{logit}(Y_i) = \alpha + \beta X_i$$

where $X_1 = 1$ if the i^{th} subject is from group A, and $X_i = 0$ if the i^{th} subject is from group B. The researchers estimate $\alpha = 0.1$ with 95% CI (0.08, 0.12), and $\beta = 0.6$, with 95% CI (0.5, 0.7).

Calculate an odds ratio and 95% CI for the effect of Drug A versus Drug B from this study.

4. The probability that asthma patients suffer an asthma attack on any given day ($Y = 1$) or not ($Y = 0$) depends in part on various measures of air quality and air temperature, as well as whether patients are classified as severe asthmatics ($S = 1$) or not ($S = 0$). Two continuous measures of air quality are given by $A1$ and $A2$, both measured on scales from 0 to 10, with higher numbers indicating poorer air quality.

Suppose that 300 randomly selected persons with asthma are followed for one day each on different days during the year. For each day, the subject reports whether they had an asthma attack that day or not, and the air quality measures and daily average temperature (T, in degrees Celsius) are recorded.

The researchers calculate a correlation matrix for their three continuous variables. The correlation matrix is:

```
          A1        A2         T
A1 1.0000000 0.9387460 0.9178854
A2 0.9387460 1.0000000 0.8646888
T  0.9178854 0.8646888 1.0000000
```

The logistic regression results using the bic.glm program are:

```
> summary(output)

Call:
bic.glm.formula(f = Y ~ A1 + A2 + T + S, data = asthma, glm.family = "binomial",
                            OR = 10^30)

14  models were selected
 Best  5  models (cumulative posterior probability =  0.9894 ):

    p!=0    EV     SD       model 1     model 2     model 3    model 4    model 5
Int 100    -3.919 1.045  -4.101e+00  -3.065e+00  -4.084e+00 -2.842e+00 -1.190e+00
A1   97.0   0.727 0.248   7.537e-01   7.124e-01   7.755e-01      .          .
A2    7.2   0.006 0.080       .           .      -2.515e-02  4.145e-01      .
T   100.0   0.345 0.071   3.501e-01   3.011e-01   3.509e-01  3.822e-01  4.389e-01
S    87.8   1.401 0.732   1.598e+00       .       1.601e+00  1.498e+00  1.488e+00

nVar                        3           2           4          3           2
BIC                     -1.584e+03  -1.580e+03  -1.578e+03 -1.576e+03  -1.575e+03
post prob                 0.806       0.111       0.047      0.016       0.009

> output$names
[1] "A1" "A2" "T"  "S"
```

```
> output$label
 [1] "A1,T,S"    "A1,T"      "A1,A2,T,S" "A2,T,S"    "T,S"
 [6] "A1,A2,T"   "A2,T"      "T"         "A1"
[10] "A1,S"      "A1,A2"     "A1,A2,S"   "A2"        "A2,S"

> output$mle
            [,1]      [,2]        [,3]      [,4]      [,5]
 [1,] -4.1006052 0.7537151  0.00000000 0.3501235 1.5982445
 [2,] -3.0650319 0.7124347  0.00000000 0.3010612 0.0000000
 [3,] -4.0841415 0.7754976 -0.02514777 0.3508818 1.6013627
 [4,] -2.8424145 0.0000000  0.41447345 0.3822355 1.4976795
 [5,] -1.1900042 0.0000000  0.00000000 0.4389060 1.4881038
 [6,] -3.0727524 0.7022309  0.01171442 0.3007594 0.0000000
 [7,] -1.9803287 0.0000000  0.40713250 0.3365197 0.0000000
 [8,] -0.3847706 0.0000000  0.00000000 0.3995131 0.0000000
 [9,] -5.4724383 1.3771444  0.00000000 0.0000000 0.0000000
[10,] -6.0610912 1.4147407  0.00000000 0.0000000 0.8586830
[11,] -5.5763508 1.2430094  0.15200718 0.0000000 0.0000000
[12,] -6.1569269 1.2815308  0.14976911 0.0000000 0.8570451
[13,] -4.6529226 0.0000000  1.12486298 0.0000000 0.0000000
[14,] -5.0797744 0.0000000  1.14230828 0.0000000 0.6947623

> output$se
           [,1]      [,2]      [,3]       [,4]      [,5]
 [1,] 0.9538962 0.2128362 0.0000000 0.06955914 0.5494020
 [2,] 0.7921597 0.1987162 0.0000000 0.06167288 0.0000000
 [3,] 0.9640919 0.2912132 0.2284706 0.06990665 0.5502603
 [4,] 0.7954493 0.0000000 0.1659538 0.06544467 0.5259452
 [5,] 0.3729248 0.0000000 0.0000000 0.05997584 0.5079964
 [6,] 0.8053955 0.2734628 0.2162258 0.06191051 0.0000000
 [7,] 0.6656227 0.0000000 0.1558285 0.05733013 0.0000000
 [8,] 0.2303562 0.0000000 0.0000000 0.05262999 0.0000000
 [9,] 0.6810708 0.1652766 0.0000000 0.00000000 0.0000000
[10,] 0.7863852 0.1715005 0.0000000 0.00000000 0.4392946
[11,] 0.7105290 0.2430059 0.2103088 0.00000000 0.0000000
[12,] 0.8084348 0.2484469 0.2103492 0.00000000 0.4390292
[13,] 0.5990318 0.0000000 0.1344980 0.00000000 0.0000000
[14,] 0.6613375 0.0000000 0.1364638 0.00000000 0.3805869

> output$postmean
[1] -3.919  0.727  0.006  0.345  1.401
```

(a) Using the results from the first (i.e., [1]) model, provide an odds ratio with 95% CI for a 1 degree change in temperature.

(b) Discuss any confounding that you think may have occurred in this analysis, based on all of the information presented.

5. (a) Continuing from the question number 4, write down the optimal model for making future predictions for the probability of asthma.

(b) Suppose that a severe asthmatic is followed on a day where $T = A1 = A2 = 0$. Using your model from (a), provide the predicted probability of asthma for that subject on that day.

(c) Again assuming $T = A1 = A2 = 0$, repeat part (b), but this time for a non-severe asthmatic subject followed for one day.

6. Suppose that whether a child will be peanut allergic ($Y = 1$) or not ($Y = 0$) at age 3 in part depends on whether their parents are allergic (dichotomous variable, $P = 0$ if neither parent is allergic, $P = 1$ if one or both parents are allergic) and the age of first introduction of peanut into the diet ($I =$ age in months, range = 4 to 36 months). In addition, there is a possible interaction between the variables P and I, so that the effect of age at introduction may depend on the parents allergic status.

A sample of 1000 children aged 3 or older is collected, and the above information collected for each. A logistic regression model is run, with the results given below (the interaction term is denoted by PI):

```
Call:
glm(formula = Y ~ P + I + PI, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5352  -0.4578  -0.2222   0.3479   2.8830

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.49334    0.75748  -7.252 4.10e-13 ***
P            1.69734    0.83509   2.033  0.04210 *
I            0.11276    0.02695   4.184 2.87e-05 ***
PI           0.08634    0.03171   2.723  0.00647 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Number of Fisher Scoring iterations: 6
```

(a) What is the odds ratio for each one month delay in introducing peanuts to the diet when neither parent is peanut allergic?

(b) What is the odds ratio for each one month delay in introducing peanuts to the diet when at least one parent is peanut allergic?

(c) What is the odds ratio for each three month delay in introducing peanuts to the diet when at least one parent is peanut allergic?

7. A clinician wishes to know the mean weights in Kg of patients visiting his clinic for the past month. He is curious to know since he has just read an article claiming that 33% of Canadians are overweight, and he is wondering if patients at his clinic are similar.

It is routine practice at this clinic to have a nurse weigh all patients prior to seeing the doctor, so, in theory, all patients should have their weight from their most recent visit recorded in their charts. However, for reasons unrelated to the patient, not all weights are recorded at all visits. For example, if the nurse is particularly busy some patients may be missed.

The clinician notices that of the 400 patients who visited in the past month, 100 have missing weights, although most of these have other information in the chart, such as weight from earlier time periods.

The clinician considers three different methods to estimate the mean weights:

(i) Ignore the 100 subjects with missing weights, and just use the data from the 300 subjects with their most recent weights recorded in their charts.

(ii) Use the most recent data available from each subject, that is, the data from the 300 subjects with recent weights recorded, and the last weight recorded for other subjects. If a few subjects do not have any weights recorded, just delete them from the analysis.

(iii) Use the the data from the 300 subjects with recent weights recorded, and create a multiple imputation model for the 100 subjects with missing data, based on whatever other information is available in their charts that could be related to weight.

State the advantages and disadvantages of each of the above three methods. Which of the above procedures do you think would be most reasonable to use in this case, and why?

8. Suppose that the Island of Montreal is divided into 5 Health Regions. It is thought that vaccination rates may differ across regions, where “vaccination” is defined as a “success” if a child is up-to-date with all recommended vaccinations by age 2. The hierarchical model below is run, with the results given below:

```
model
{
 for (i in 1:5)
{
     logit(vacc[i]) <- z[i]
     x[i] ~ dbin(vacc[i], n[i])
     z[i] ~ dnorm(mu, tau)
}

mu ~  dnorm(0,0.01)
tau <- 1/(sigma*sigma)
sigma ~  dunif(0,10)
pdiff41 <- vacc[4] - vacc[1]

}

# Inits

list(mu=0, sigma = 1)

# Data

list(x = c(8000, 15555, 14000, 26500, 12000 ),
n=c(10000, 20000, 20000, 30000, 15000))

#  Results

node    mean    sd      MC error   2.5%    median  97.5%   start   sample
mu      1.377   0.3567   0.002561  0.6846  1.378   2.076   1001    20000
pdiff41 0.08324 0.00439  3.09E-5   0.07465 0.08325 0.09188 1001    20000
sigma   0.6679  0.4622   0.006289  0.2776  0.5506  1.773   1001    20000
tau     4.207   3.415    0.02952   0.3187  3.299   12.98   1001    20000
vacc[1] 0.8     0.00399  2.91E-5   0.7922  0.8     0.8078  1001    20000
vacc[2] 0.7778  0.00291  1.991E-5  0.772   0.7778  0.7835  1001    20000
vacc[3] 0.7001  0.00323  2.278E-5  0.6937  0.7001  0.7065  1001    20000
vacc[4] 0.8832  0.00186  1.157E-5  0.8796  0.8832  0.8869  1001    20000
vacc[5] 0.8     0.00325  2.294E-5  0.7936  0.8     0.8063  1001    20000
```

(a) From the above results, what is your best estimate of the overall vaccination rate across the 5 regions?

(b) Do you think Region 1 has a different vaccination rate compared to Region 4? Explain your answer.

Normal Density Table

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table of standard normal distribution probabilities. Each number in the table provides the probability that a standard normal random variable will be less than the number indicated.