Data Analysis in the Health Sciences

Final Exam 2008 – EPIB–621

Student's Name: _____

Student's Number: _____

INSTRUCTIONS

This examination consists of 8 questions on 17 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. The following data have been collected on the development of breast cancer and age at first birth:

	Age at first birth (years)	
Cancer	less than 30	30 or greater
Yes	700	300
No	2500	500

(a) Carry out an appropriate two-sided test to determine whether there is a relationship between breast cancer and age at first birth. State the null and alternative hypotheses, carry out the test, and provide a conclusion.

(b) Calculate a 95% confidence interval for the difference in breast cancer rates within the two age at first birth groupings. State your conclusion based on this confidence interval.

2. Two surgeons are arguing about the value of a new surgical technique. Surgeon 1 is enthusiastic, and therefore has a beta(35,15) prior probability distribution on the success rate of the technique. Surgeon 2 is more pessimistic, and therefore states that his prior probability distribution is beta(15,35).

(a) Suppose that the two surgeons agree to collect some data to settle the issue. They observe 100 patients, with 60 of these having successful surgeries. What is the posterior distribution for the success rate for Surgeon 1? What is the posterior distribution for the success rate for Surgeon 2?

(b) Did Surgeon 1 or Surgeon 2 have a prior mean value which came closer to the mean success rate actually observed in the data given in part (a)?

3. A linear regression line has been fitted between the independent (x) variable age, and the dependent (y) variable Body Mass Index (BMI). The BMI value for any individual is defined to be their weight in kilograms divided by the square of their height in meters. The sample size is $n = 100$ data points. The slope of the regression line is estimated to be $b = 0.1$. The average age in this sample is $\overline{x} = 50$ years, with a standard deviation of 10 years, and the average BMI is $\overline{y} = 25$ kg/m^2, with a standard deviation of 2 kg/m^2.

(a) Provide an interpretation for the slope of the line.

(b) Is it possible to estimate the intercept of the regression line between age and BMI using the above data? If yes, provide the value of the intercept. If not, state what information is missing.

4. An investigator is studying factors associated with children having received all recommended vaccinations by age 2. The variables studies are vaccinated (vacc: yes=1, no=0), age of mother at childbirth (age is continuous), whether the child was a first born or not (first: yes=1, no=0), year of birth (continuous, from 2000 until 2005), and urban versus rural location (urban: urban=1, rural=0). After carrying out a survey and analyzing the data, the researcher finds the following logistic regression results:

```
> output<-glm(vacc ~ age + year + first + urban, family="binomial",
          data=vacc.data)
> summary(output)

Call:
glm(formula = vacc ~ age + year + first + urban, family = "binomial",
    data = vacc.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3167  -0.8364  -0.6839   1.1989   2.0625

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  12.567308 118.597649   0.106 0.915609
age           0.038211   0.016359   2.336 0.019502 *
year         -0.007235   0.059226  -0.122 0.902775
first        -0.770844   0.222486  -3.465 0.000531 ***
urban         0.793706   0.238464   3.328 0.000873 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 592.95  on 499  degrees of freedom
Residual deviance: 565.70  on 495  degrees of freedom
AIC: 575.7

Number of Fisher Scoring iterations: 4

> confint(output)
                    2.5 %       97.5 %
(Intercept) -2.199286e+02 245.72353071
age          6.326904e-03   0.07055531
year        -1.236767e-01   0.10886269
first       -1.216009e+00  -0.34206903
urban        3.237672e-01   1.26056567
```

(a) What is the odds ratio and 95% confidence interval for the odds ratio per 10 year change in mother's age at childbirth?

(b) What is your point estimate of the probability that the first child born to a mother aged 25 years old in a rural area in the year 2003 will have all recommended vaccinations by age 2?

(c) Provide an interpretation for the intercept of the model. Explain why it is a relatively high value here, with intercept ≈ 12.57.

5. Continuing the same example as in number four above, below are the results from the bic.glm program:

```
> output.bic <- bic.glm(vacc ~ age + year + first + urban,
          glm.family="binomial", data=vacc.data, OR=10000)
> summary(output.bic)

Call:
bic.glm.formula(f = vacc ~ age + year + first + urban, data = vacc.data,
        glm.family = "binomial", OR = 10000)

  16  models were selected
 Best  5  models (cumulative posterior probability =  0.9376 ):

           p!=0     model 1     model 2     model 3     model 4     model 5
Intercept  100     -8.632e-01  -1.920e+00  -6.931e-01  -1.671e+00  -1.115e+00
age        40.1        .       3.821e-02       .       3.558e-02       .
year        4.3        .           .           .           .           .
first      94.9    -7.351e-01  -7.691e-01  -7.191e-01  -7.516e-01       .
urban      89.2     7.620e-01   7.912e-01       .           .       7.423e-01

nVar                   2           3           1           2           1
BIC                -2.517e+03  -2.517e+03  -2.513e+03  -2.512e+03  -2.512e+03
post prob           0.474       0.336       0.064       0.033       0.030

> output.bic$mle
             [,1]       [,2]         [,3]       [,4]      [,5]
 [1,]  -0.8631673 0.00000000  0.000000000 -0.7350650 0.7619533
 [2,]  -1.9201813 0.03821140  0.000000000 -0.7690824 0.7911923
 [3,]  -0.6931472 0.00000000  0.000000000 -0.7191227 0.0000000
 [4,]  -1.6710898 0.03558152  0.000000000 -0.7515666 0.0000000
 [5,]  -1.1154761 0.00000000  0.000000000  0.0000000 0.7422718
 [6,]  13.5828699 0.00000000 -0.007214141 -0.7369822 0.7645038
 [7,]  12.5673077 0.03821052 -0.007234860 -0.7708443 0.7937059
 [8,]  -2.0768288 0.03434858  0.000000000  0.0000000 0.7685114
 [9,]  -0.9444616 0.00000000  0.000000000  0.0000000 0.0000000
[10,] -17.9517892 0.00000000  0.008618375 -0.7169811 0.0000000
[11,]  -1.8278153 0.03176516  0.000000000  0.0000000 0.0000000
[12,] -19.7567757 0.03559127  0.009031249 -0.7494279 0.0000000
[13,] -15.8973588 0.00000000  0.007382180  0.0000000 0.7397832
[14,] -15.7021873 0.03433886  0.006804756  0.0000000 0.7661379
[15,] -44.5701826 0.00000000  0.021786144  0.0000000 0.0000000
[16,] -46.1682103 0.03178760  0.022142735  0.0000000 0.0000000
> output.bic$se
              [,1]       [,2]       [,3]      [,4]      [,5]
 [1,]   0.13474308 0.00000000 0.00000000 0.2202667 0.2353314
 [2,]   0.47715864 0.01635915 0.00000000 0.2220036 0.2375696
```

```
 [3,]   0.12126779 0.00000000 0.00000000 0.2177887 0.0000000
 [4,]   0.46379814 0.01614120 0.00000000 0.2194401 0.0000000
 [5,]   0.11639820 0.00000000 0.00000000 0.0000000 0.2318437
 [6,] 118.24003660 0.00000000 0.05904744 0.2208455 0.2362753
 [7,] 118.59764899 0.01635874 0.05922585 0.2224860 0.2384640
 [8,]   0.47187936 0.01613769 0.00000000 0.0000000 0.2336110
 [9,]   0.09960231 0.00000000 0.00000000 0.0000000 0.0000000
[10,] 116.18529315 0.00000000 0.05801880 0.2182547 0.0000000
[11,]   0.45871171 0.01594051 0.00000000 0.0000000 0.0000000
[12,] 116.36806979 0.01614232 0.05810874 0.2198662 0.0000000
[13,] 116.17878298 0.00000000 0.05802034 0.0000000 0.2326527
[14,] 116.53786046 0.01613747 0.05820044 0.0000000 0.2344794
[15,] 114.37717486 0.00000000 0.05711811 0.0000000 0.0000000
[16,] 114.52500154 0.01594301 0.05719047 0.0000000 0.0000000
> output.bic$names
[1] "age"   "year"  "first" "urban"
> output.bic$postprob
[1]  4.742692e-01 3.360135e-01 6.402981e-02 3.340840e-02 2.985755e-02
[6]  2.136888e-02 1.513954e-02 1.317213e-02 4.581385e-03 2.895266e-03
[11] 1.519307e-03 1.512223e-03 1.346121e-03 5.931154e-04 2.203491e-04
[16] 7.323538e-05
> output.bic$label
 [1] "first,urban"          "age,first,urban"      "first"                "age,first"
 [5] "urban"                "year,first,urban"     "age,year,first,urban" "age,urban"
 [9] "NULL"                 "year,first"           "age"                  "age,year,first"
[13] "year,urban"           "age,year,urban"       "year"                 "age,year"
```

(a) Which of the above models corresponds to the model run in problem number 4 above? Compare the coefficients from this model to those given in question # 4.

(b) Considering all of the results given above, do you think there is any evidence of confounding in this analysis? Carefully explain why or why not.

(c) Using results from the second best model (model 2), what is the odds ratio and 95% confidence interval for the odds ratio per 10 year change in mother's age at childbirth? How does this compare to your answer from question # 4?

6. Suppose that the probability of being diagnosed with diabetes (1=yes, 0=no) in a population is associated with the presence of an inactive lifestyle (inactive=1, active=0) according to the following logistic regression model:

$$\text{logit}(\text{diabetes}) = -2.5 + 0.3 * \text{inactive}$$

If 50% of the population are inactive, what would be your best estimate of the overall probability of diabetes in this population?

7. Overweight and obese persons are at increased risk for gastroesophageal reflux disease. A questionnaire was given to 10,545 randomly selected women to determine the presence of gastroesophageal reflux disease. After categorizing women according to BMI, logistic-regression models were used to study the association between BMI and gastroesophageal reflux disease. The following results were found:

Table 2. Association between Body-Mass Index and Frequent Symptoms of Gastroesophageal Reflux Disease.*

Variable	No. of Women	Body-Mass Index							P for Trend
		<20.0	20.0–22.4†	22.5–24.9	25.0–27.4	27.5–29.9	30.0–34.9	≥35.0	
Mild symptoms									
No. of women with symptoms of GERD	473	13	63	99	115	64	85	34	
No. of controls	3829	314	812	917	740	439	412	195	
Univariate odds ratio (95% CI)		0.55 (0.30–1.02)	1.00	1.38 (0.99–1.92)	2.00 (1.45–2.77)	1.88 (1.30–2.72)	2.62 (1.85–3.71)	2.20 (1.41–3.43)	<0.001
Multivariate odds ratio (95% CI)		0.61 (0.33–1.14)	1.00	1.36 (0.96–1.92)	2.04 (1.45–2.88)	1.75 (1.17–2.61)	2.33 (1.59–3.43)	2.05 (1.24–3.39)	<0.001
Moderate symptoms									
No. of women with symptoms of GERD	1678	47	170	303	400	285	319	154	
No. of controls	3899	317	815	939	757	448	425	198	
Univariate odds ratio (95% CI)		0.72 (0.51–1.02)	1.00	1.54 (1.25–1.91)	2.53 (2.06–3.11)	3.05 (2.44–3.80)	3.57 (2.87–4.45)	3.68 (2.81–4.81)	<0.001
Multivariate odds ratio (95% CI)		0.70 (0.48–1.02)	1.00	1.50 (1.20–1.87)	2.36 (1.89–2.94)	2.71 (2.13–3.45)	3.18 (2.50–4.06)	3.15 (2.33–4.26)	<0.001
Severe-to-very-severe symptoms									
No. of women with symptoms of GERD	256	6	37	38	50	45	54	26	
No. of controls	3874	317	811	931	751	445	421	198	
Univariate odds ratio (95% CI)		0.43 (0.18–1.04)	1.00	0.89 (0.56–1.41)	1.46 (0.94–2.26)	2.22 (1.41–3.48)	2.76 (1.79–4.27)	2.81 (1.66–4.75)	<0.001
Multivariate odds ratio (95% CI)		0.55 (0.22–1.33)	1.00	0.94 (0.57–1.53)	1.37 (0.85–2.21)	1.92 (1.16–3.19)	2.40 (1.46–3.96)	2.36 (1.28–4.37)	<0.001

* Frequent symptoms of gastroesophageal reflux disease were defined as heartburn, acid regurgitation, or both occurring at least weekly. Severity of symptoms was defined as mild ("can be ignored if I don't think about it"), moderate ("cannot be ignored but does not affect my lifestyle"), severe ("affects my lifestyle"), and very severe ("markedly affects my lifestyle"). GERD denotes gastroesophageal reflux disease, and CI confidence interval. Multivariate odds ratios have been adjusted for age; smoking status; total activity; daily caloric intake; intake of alcohol, coffee, tea, and chocolate; use of postmenopausal hormone therapy; use of antihypertensive medication or asthma medication; and presence or absence of diabetes mellitus.

† Women with a body-mass index of 20.0 to 22.4 served as the reference population.

(a) Using results from the multivatriate model for severe to very severe symptoms, state and provide an interpretation for the odds ratio and 95% confidence interval for the effect of a BMI in the range of 25.0 to 27.4 on the outcome.

(b) Considering all of the above results, do you think that underweight subjects ($BMI < 20$) are more likely, equally likely, or less likely to experience gastroesophageal reflux compared to subjects that have normal BMI (defined as $20 \leq BMI < 25$)?

(c) The table includes both univariate and multivariate odds ratios. Comparing odds ratios from these two analyses, in general, did adjusting for the various covariates (age, smoking status, activity, etc.) tend to increase or decrease the effect of BMI on gastroesophageal reflux disease?

8. In Quebec, the type of facilities available for treatment of a heart attack varies by region, so that the probability of receiving a given treatment depends on where you live. This, in turn, affects your probability of surviving. Data are collected on survival (surv) following a heart attack in 17 regions of Quebec, and a hierarchical model is run. The program and results are given below:

```
model {
 for (j in 1:17)

{
 for (i in index1[j]:index2[j])
     {
     logit(p[i]) <- region[j]
     surv[i] ~ dbern(p[i])
     }
     region[j] ~ dnorm(mu, tau)
}

mu ~  dnorm(0,0.001)
tau <- 1/(sigma*sigma)
sigma ~  dunif(0,20)
beta  ~  dnorm(0, 0.001)
p.r1.r2 <- step(region[1] - region[2])

}

# Inits

list(region=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0), mu=0, sigma = 1)

# Data

list(index1 =c(1, 1001, 2001, 3001, 4001, 5001, 6001, 7001, 8001,
9001, 10001, 11001, 12001, 13001, 14001, 15001, 16001),
index2=c(1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000,
11000, 12000, 13000, 14000, 15000, 16000, 17000),
surv = c(1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0
1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1,
..................etc.....................................
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1,
0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1))
```

```
#  Results

node        mean     sd         MC error     2.5%    median   97.5%   start   sample
beta       -0.3317   31.67      0.3106     -62.22   -0.4716   60.24   1001    10000
mu          1.513    0.1347     0.00144     1.251    1.514    1.78    1001    10000
p.r1.r2     1.0      0.0        1.0E-12     1.0      1.0      1.0     1001    10000
region[1]   1.709    0.08679    8.566E-4    1.544    1.709    1.884   1001    10000
region[2]   0.8367   0.06843    6.609E-4    0.7048   0.8368   0.9697  1001    10000
region[3]   1.909    0.09233    8.906E-4    1.731    1.908    2.095   1001    10000
region[4]   0.841    0.06877    6.397E-4    0.7071   0.8403   0.9787  1001    10000
region[5]   2.195    0.1039     0.001023    1.995    2.193    2.404   1001    10000
region[6]   1.943    0.09513    0.001025    1.761    1.941    2.135   1001    10000
region[7]   1.671    0.08568    8.481E-4    1.505    1.67     1.841   1001    10000
region[8]   2.249    0.1049     0.001021    2.048    2.246    2.46    1001    10000
region[9]   0.99     0.07046    7.801E-4    0.8546   0.9899   1.129   1001    10000
region[10]  1.48     0.08006    7.78E-4     1.326    1.479    1.638   1001    10000
region[11]  0.8272   0.06901    6.756E-4    0.6939   0.8265   0.9612  1001    10000
region[12]  1.687    0.08646    8.97E-4     1.523    1.686    1.86    1001    10000
region[13]  1.918    0.09244    8.008E-4    1.74     1.917    2.102   1001    10000
region[14]  1.926    0.09368    9.814E-4    1.743    1.927    2.111   1001    10000
region[15]  1.205    0.07419    7.845E-4    1.063    1.204    1.352   1001    10000
region[16]  1.233    0.07457    7.162E-4    1.089    1.233    1.381   1001    10000
region[17]  1.108    0.07349    6.943E-4    0.9654   1.108    1.252   1001    10000
sigma       0.5373   0.109      0.001361    0.3721   0.5209   0.7937  1001    10000
```

(a) From the above results, what is your best estimate of the overall survival rate across all 17 regions?

(b) Do you think Region 1 has a higher survival rate compared to Region 2? 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.