Data Analysis in the Health Sciences

Final Exam 2019 – EPIB–621

Student's Name: _____

Student's Number: _____

INSTRUCTIONS

This examination consists of 5 questions on 14 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 50.

1. _____
2. _____
3. _____
4. _____
5. _____

Total (out of 50) _____

1. Some researchers report the following: "In our sample of 225 subjects, the blood pressure was lowered by an average of 5 points ($p = 0.05$ using a two sided t-test)." From the information given, provide an approximate 95% confidence interval for the average amount by which blood pressure was lowered.

2. In a large cross-sectional study of depression, 10,000 subjects are asked if they are currently having symptoms of depression, along with their age and sex. The following logistic regression model is fit to the data collected:

$$logit(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where:

Y is equal to 1 if the subject was depressed and 0 otherwise,
$X1$ is the subject's sex, with 0 indicating a male and 1 a female,
$X2$ is a centered variable for age in years, centered at age = 50, and
$X3$ is the square of $X2$.

The following R output is obtained:

```
Call:
glm(formula = Y ~ X1 + X2 + X3, family = "binomial")

Deviance Residuals:
Min       1Q   Median       3Q      Max
-0.3677  -0.1873  -0.0416  -0.0034   3.6007

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.326355   0.172258 -19.310  < 2e-16 ***
X1           0.629383   0.197050   3.194  0.00140 **
X2           0.050016   0.018532   2.699  0.00696 **
X3          -0.017050   0.002243  -7.602  2.91e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1273.5  on 9999  degrees of freedom
Residual deviance: 1030.7  on 9996  degrees of freedom
AIC: 1038.7

Number of Fisher Scoring iterations: 11
```

(a) Provide an interpretation for the intercept in this model, including any applicable implications for the probability of depression for any particular age and sex combination.

(b) Provide the predicted probability that a 55 year old female will be depressed.

(c) Provide the odds ratio for sex and calculate a 95% confidence interval for this odds ratio.

3. Suppose that a prevalence survey on rheumatoid arthritis is conducted across eight different health regions, with 500 subjects studied within each region for a total sample size of 4,000. The following model is run in WinBUGS, using the data and initial values as given below:

```
model
{
 for (i in 1:8)
 {
 x[i] ~ dbin(p[i],n[i])
 logit(p[i]) <- z[i]
 z[i] ~ dnorm(mu,tau)
  }
   mu ~  dnorm(0,0.001)
   tau <- 1/(sigma*sigma)
   sigma ~ dunif(0, 20)
   y  ~ dnorm(mu, tau)
   w <- exp(y)/(1+exp(y))
}

# Data
list(n=c(500, 500, 500, 500, 500, 500, 500, 500),
     x=c(10, 12, 5, 8, 15, 18, 13, 7))
# Initial Values
list(mu=0, sigma=1)
```

In this model `n[i]` provides the sample sizes sampled within each region, and `x[i]` is the number of subjects with rheumatoid arthritis out of the `n[i]` subjects sampled within each region. The following WinBUGS output is produced upon running the above program with the above data and initial values:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
mu	-3.84	0.1842	0.003059	-4.223	-3.835	-3.497	2001	20000
p[1]	0.02088	0.004717	5.925E-5	0.01223	0.02069	0.03107	2001	20000
p[2]	0.02281	0.005024	5.477E-5	0.01422	0.02226	0.03439	2001	20000
p[3]	0.01653	0.00481	1.212E-4	0.00751	0.01652	0.02579	2001	20000
p[4]	0.01906	0.00457	8.075E-5	0.01042	0.01901	0.02822	2001	20000
p[5]	0.02569	0.00583	8.514E-5	0.01662	0.02486	0.03947	2001	20000
p[6]	0.02881	0.007052	1.37E-4	0.01849	0.02768	0.04549	2001	20000
p[7]	0.02379	0.005231	6.256E-5	0.01514	0.02312	0.03596	2001	20000
p[8]	0.01823	0.00459	9.125E-5	0.0095	0.01831	0.02716	2001	20000
sigma	0.3417	0.2296	0.006373	0.0216	0.3094	0.8796	2001	20000
w	0.02317	0.01546	1.373E-4	0.00829	0.02131	0.05048	2001	20000
y	-3.84	0.4472	0.004359	-4.784	-3.827	-2.934	2001	20000

(a) Ignoring the WinBUGS model that was run, and using the data from region [3] only, calculate the point estimate and an approximate 95% confidence interval for the prevalence of rheumatoid arthritis in region [3].

(b) Now using the results from the WinBUGS model, state the prevalence estimate and 95% credible interval that was estimated for region [3]. Comparing the two point estimates from parts (a) and (b), state which is higher and explain why there is a difference between them.

(c) State the point estimate and 95% credible interval for `w`, and explain the information given by this parameter in this model.

4. The following WinBUGS program is created and run, with results given below the program:

```
model
{
    for (i in 1:10) {
     mu[i]  <- alpha + beta*age[i]
     y[i]   ~ dnorm(mu[i],tau)
    }
    alpha    ~ dnorm(0.0, 0.001)
    beta    ~ dnorm(0.0, 0.001)
    tau      <- 1/(sigma*sigma)
    sigma  ~ dunif(0.001, 100)
 }

#Data

list(age = c(59, 52, 37, 40, 67, 43, 61, 34, 51, 58),
  y = c(143, 132, 88, NA, 177, 102, 154, NA, 131, 150))

# Results

node    mean    sd       2.5%    median  97.5%
alpha   -17.36  10.6     -38.06  -17.69  5.066
beta    2.842   0.1957    2.434   2.847   3.221
sigma   5.201   2.045     2.804   4.727   10.32
y[4]    96.25   6.397     83.84   96.16   109.7
y[8]    79.27   7.074     65.56   79.14   93.84
```

(a) Explain what the output for y[8] represents, and provide an interpretation for its mean and 95% credible interval.

(b) The same program as in 4 (a) is run again in WinBUGS, but now using the slightly different data set below. The program now terminates with an error message. Explain why.

```
#Data

list(age = c(59, 52, 37, 40, 67, 43, NA, 34, 51, 58),
  y = c(143, 132, 88, NA, 177, 102, 154, NA, 131, 150))
```

5. Researchers are deriving an equation to best predict the probability a subject will have osteoporosis based on five predictors, including age (in years), sex (1=female, 0 = male), smoking status (1 indicates a smoker, 0 indicates a non-smoker), family history of osteoporosis (1 = yes, 0 = no) and previous fracture (1 = yes, 0 = no). The researchers collect data on 2000 subjects, and run the big.glm program using a binomial family (i.e., they are running logistic regression).

They find the following results based on their sample of 2000 subjects:

```
> summary(osteo.dat)
     osteo          smoke             age            sex           fam_hist
 Min.   :0.00   Min.   :0.0000   Min.   :50.0   Min.   :0.000   Min.   :0.000
 1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:57.0   1st Qu.:0.000   1st Qu.:0.000
 Median :1.00   Median :0.0000   Median :65.0   Median :0.000   Median :0.000
 Mean   :0.52   Mean   :0.2955   Mean   :65.1   Mean   :0.484   Mean   :0.255
 3rd Qu.:1.00   3rd Qu.:1.0000   3rd Qu.:73.0   3rd Qu.:1.000   3rd Qu.:1.000
 Max.   :1.00   Max.   :1.0000   Max.   :80.0   Max.   :1.000   Max.   :1.000


    prev_frac
  Min.   :0.000
  1st Qu.:0.000
  Median :0.000
  Mean   :0.202
  3rd Qu.:0.000
  Max.   :1.000

> output <- bic.glm(osteo ~ age + sex + smoke + fam_hist + prev_frac,
glm.family = binomial, data=osteo.dat, OR = 2000)
```

```
> summary(output)
Call:
bic.glm.formula(f = osteo ~ age + sex + smoke + fam_hist + prev_frac,
data = osteo.dat, glm.family = binomial, OR = 2000)


  4  models were selected
 Best  4  models (cumulative posterior probability =  1 ):

           p!=0    EV       SD       model 1    model 2     model 3     model 4
Intercept  100    -3.898  0.3708  -3.924e+00  -3.850e+00  -3.765e+00  -3.689e+00
age        100.0   0.048  0.0054   4.860e-02   4.865e-02   4.809e-02   4.810e-02
sex        100.0   0.973  0.0960   9.760e-01   9.677e-01   9.590e-01   9.507e-01
smoke      100.0   0.669  0.1062   6.714e-01   6.689e-01   6.404e-01   6.380e-01
fam_hist    97.7   0.418  0.1263   4.280e-01   4.271e-01           .           .
prev_frac   69.8   0.252  0.1940   3.623e-01           .   3.612e-01           .

nVar                                       5           4           4           3
BIC                                -1.263e+04  -1.263e+04  -1.262e+04  -1.262e+04
post prob                           0.682       0.295       0.016       0.007
```

(a) Using the coefficients optimized for future predictions, provide the predicted probability of osteoporosis for a female subject aged 65, who smokes, and has both a family history of osteoporosis and had a previous fracture.

(b) State the average rate of osteoporosis in this sample, and compare it to the predicted probability you calculated in part (a). Explain why one of these numbers is higher than the other.

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.