Data Analysis in the Health Sciences

Midterm Exam 2019 – EPIB–621

Student's Name: ____

Student's Number: ____

INSTRUCTIONS

This examination consists of 5 questions on 11 pages, including this one. 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. A person who had chicken pox as a child is susceptible to developing shingles later as an adult. Suppose that a study is being carried out to estimate the proportion of adults who had chicken pox as children who will develop shingles by age 70. A group of 200 subjects, all aged 70 or over and all of whom had chicken pox as children are surveyed. It is found that 60 of these subjects developed shingles as adults by age 70.

(a) Calculate a 95% confidence interval for the proportion of adults who had chicken pox as children who will develop shingles by age 70.

(b) Suppose that you do not know anything before the study is carried out about the probability of developing shingles in this population. Write down a reasonable choice of beta prior distribution to represent this lack of knowledge.

(c) Using your choice of prior from part (b) and the data given above, write down the posterior density for the probability of shingles by age 70.

(d) Calculate the mean and standard deviation for your posterior density from part (c).

2. The amount of time it takes for an ambulance to respond to an emergency call at the caller's home depends in part on how far from the city center the caller lives. The ambulance service covers an area up to 30 kilometers (km) from the city center. Some data are collected from the emergency services agency, and the following linear regression model is fit:

$$Y = \alpha + \beta_1 X + \beta_2 X^2$$

where Y is the response time in minutes, and X is the distance in km from the city center that the caller lives.

Suppose that data are collected over a one year period. Based on a sample size of 1000 emergency calls, it is estimated that $\hat{\alpha} = 5$, $\hat{\beta}_1 = 3$, and $\hat{\beta}_2 = -0.05$.

(a) Explain the meaning of α in this model, and interpret the value of 5 that is estimated for α.

(b) How long, on average, does the model predict it would take to respond to an emergency for a caller living 10 km from the city center?

(c) The researchers write an article about their study and submit it to a journal for peer review. One reviewer comments that the model must be wrong, since it predicts negative response times for callers living a distance of approximately 61.6 km away or further. Since negative response times are impossible, do you agree with this criticism of the model? Why or why not?

3. It is thought that older patients may react differently to a certain drug compared to younger patients. To estimate this effect, the following linear regression model is fit:

$$Y = \alpha + \beta_1 \times age + \beta_2 \times drug + \beta_3 \times age \times drug$$

where age is the patient's age in years, drug is a dichotomized variable to indicate whether the patient takes the drug (drug=1) or not (drug = 0), and the outcome Y is a continuous measure with higher values indicating better outcomes, and a change of 3 or more in Y being considered as clinically important.

Some data are collected and the above model is fit to the data. The R linear regression output from this model is given below:

```
Call:
lm(formula = Y ~ age + drug + age * drug)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1733 -1.8722  0.0318  2.0122  6.8349

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.77727    1.34182   2.815  0.00507 **
age          0.46982    0.02060  22.805  < 2e-16 ***
drug         4.14962    1.92410   2.157  0.03151 *
age:drug     0.06204    0.02937   2.112  0.03514 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.776 on 496 degrees of freedom
Multiple R-squared:  0.8264,    Adjusted R-squared:  0.8254
F-statistic: 787.2 on 3 and 496 DF,  p-value: < 2.2e-16
```

(a) From the information above, what must have been the sample size of the data that were input to the regression model?

(b) Calculate the 95% confidence interval for β_1, the coefficient for age in the model. Provide a "practical" (that is, not necessarily "technically correct") interpretation for the CI you just calculated.

(c) Calculate the difference in the expected outcomes Y between someone aged 50, compared to someone aged 70 who are both taking the drug.

4. Bone mineral density (BMD) is related to several factors, including age, sex and weight.

BMD is measured on 500 subjects, along with each subject's age in years, weight in pounds and sex (1 = female, 0 = male). A change of 0.02 or more in BMD is considered as clinically important. The bicreg program in R is run on these data, providing the following output:

```
> summary(output)

Call:
bicreg(x = x, y = BMD, OR = exp(1000))


8  models were selected
Best  5  models (cumulative posterior probability =  1 ):

           p!=0    EV        SD        model 1     model 2     model 3     model 4     model 5
Intercept  100.0   1.080881  0.151864   9.545e-01   1.227e+00   1.091e+00   9.755e-01   1.771e+00
age         56.3  -0.006406  0.006032  -1.188e-02       .      -6.496e-03       .       2.158e-02
sex         49.0  -0.056392  0.061568       .      -1.217e-01  -6.144e-02       .      -3.328e-01
weight     100.0   0.013926  0.002930   1.659e-02   1.081e-02   1.391e-02   1.196e-02       .

nVar                                      2           2           3           1           2
r2                                      0.663       0.663       0.664       0.639       0.638
BIC                                    -5.312e+02  -5.309e+02  -5.266e+02  -5.032e+02  -4.956e+02
post prob                               0.510       0.437       0.053       0.000       0.000

> output$ols
(Intercept)          age         sex       weight
[1,]   0.9545494 -0.011878870  0.00000000 0.01658957
[2,]   1.2272479  0.000000000 -0.12167325 0.01081445
[3,]   1.0911599 -0.006495951 -0.06143776 0.01391307
[4,]   0.9755205  0.000000000  0.00000000 0.01196341
[5,]   1.7706101  0.021576124 -0.33277494 0.00000000
[6,]   1.5580316  0.022202748  0.00000000 0.00000000
[7,]   3.1179807  0.000000000 -0.34999460 0.00000000
[8,]   2.9352835  0.000000000  0.00000000 0.00000000

> output$se
(Intercept)          age        sex       weight
[1,]  0.06461200 0.0020023123 0.00000000 0.0008717905
[2,]  0.07734393 0.0000000000 0.02060530 0.0004357748
[3,]  0.12385238 0.0046203060 0.04753181 0.0022465048
[4,]  0.06669345 0.0000000000 0.00000000 0.0004029644
[5,]  0.05959571 0.0009284023 0.01911320 0.0000000000
[6,]  0.07393848 0.0011759160 0.00000000 0.0000000000
[7,]  0.01991305 0.0000000000 0.02756148 0.0000000000
[8,]  0.01582445 0.0000000000 0.00000000 0.0000000000

> output$postprob
[1]  5.103698e-01  4.367963e-01  5.283346e-02  4.269190e-07  9.694677e-09
[6]  4.298087e-59  2.943851e-87 2.304868e-116

> output$probne0
[1]  56.3  49.0 100.0
```

(a) Using the above results, state whether you believe there was any confounding between the three independent variables age, sex and weight. Explain your answer.

(b) Using results from model 1 (that is, the first model on the list of 8 models), calculate a 95% confidence interval for the effect of age on BMD, and provide a "practical" interpretation for this effect and the CI.

5. State whether each statement is true or false, and explain why:

(a) Suppose that a clinical trial is carried out to compare two different treatments, and a p-value of 0.7 is calculated from a two-sided t-test of the null hypothesis that the two treatments are equally effective. We can safely conclude that there is no important difference between the two treatments.

(b) Linear regression models are useful only for examining linear relationships between independent and dependent variables. Therefore if the relationship is thought to be non-linear, then one cannot use linear regression techniques to investigate the relationships between independent and dependent variables.

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.