Data Analysis in the Health Sciences

Midterm Exam 2012 – EPIB–621

Student's Name: _____

Student's Number: _____

INSTRUCTIONS

This examination consists of 5 questions on 20 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) _____

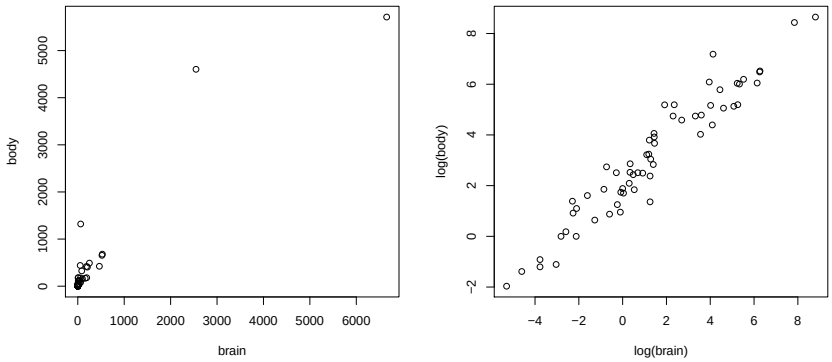
1. The data below records the average weight of the brain and body for 62 mammal species (in grams). The goal of the study is to predict the body weight of mammals as a function of their brain weight.

```
> head(data)
  Id   brain  body
1  1   3.385  44.5
2  2   0.480  15.5
3  3   1.350   8.1
4  4 465.000 423.0
5  5  36.330 119.5
6  6  27.660 115.0
> attach(data)
```

```
> brain
 [1]    3.385    0.480    1.350  465.000   36.330   27.660   14.830    1.040
 [9]    4.190    0.425    0.101    0.920    1.000    0.005    0.060    3.500
[17]    2.000    1.700 2547.000    0.023  187.100  521.000    0.785   10.000
[25]    3.300    0.200    1.410  529.000  207.000   85.000    0.750   62.000
[33] 6654.000    3.500    6.800   35.000    4.050    0.120    0.023    0.010
[41]    1.400  250.000    2.500   55.500  100.000   52.160   10.550    0.550
[49]   60.000    3.600    4.288    0.280    0.075    0.122    0.048  192.000
[57]    3.000  160.000    0.900    1.620    0.104    4.235
```

```
> body
 [1]   44.50   15.50    8.10  423.00  119.50  115.00   98.20    5.50   58.00
[10]    6.40    4.00    5.70    6.60    0.14    1.00   10.80   12.30    6.30
[19] 4603.00    0.30  419.00  655.00    3.50  115.00   25.60    5.00   17.50
[28]  680.00  406.00  325.00   12.30 1320.00 5712.00    3.90  179.00   56.00
[37]   17.00    1.00    0.40    0.25   12.50  490.00   12.10  175.00  157.00
[46]  440.00  179.50    2.40   81.00   21.00   39.20    1.90    1.20    3.00
[55]    0.33  180.00   25.00  169.00    2.60   11.40    2.50   50.40
```

The scatterplot of the data (raw scale and log-scale) is provided below:



(a) The statistician who analyzed this data chose to fit the following linear regression model:

$$\log(body) = \alpha + \beta \log(brain)$$

Explain briefly why the statistician didn't choose to fit a linear model predicting body weight (raw scale) as a function of brain weight (raw scale).

(b) The R output of the regression analysis is given below:

```
> summary(lm(log(body)~log(brain)))

Call:
lm(formula = log(body) ~ log(brain))

Residuals:
     Min       1Q   Median       3Q      Max
-1.71550 -0.49228 -0.06162  0.43597  1.94829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479     ???     22.23   <2e-16 ***
log(brain)   0.75169    0.02846   ???   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared: 0.9208,     Adjusted R-squared: 0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

As you can notice, two values are missing in the R output (??? marks). Compute the missing values above.

(c) Compute a 95% confidence interval for the slope coefficient

(d) Compute the residual value for the first mammal specie in the dataset

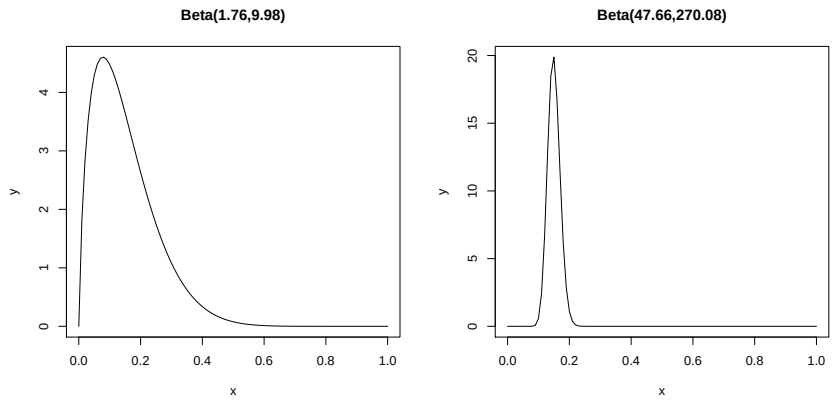
(e) Interpret the slope coefficient in the fitted model by completing the following sentences:

- *When the log-brain increases by 1 gram, the average log-body ...*

- *When the log-brain increases by 1 gram, the average body ...*

- *When the log-brain increases by 10 grams, the average body ...*

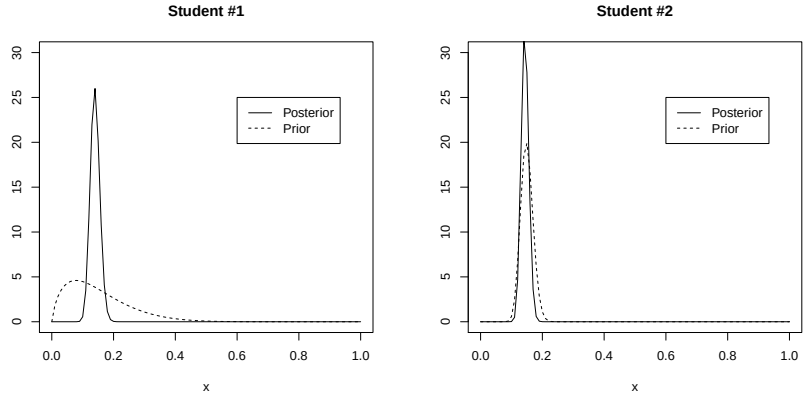
2. A small survey is carried out to estimate the prevalence of osteoarthritis in Quebec. Two EPIB621 students were asked about which prior distribution they would put on the osteoarthritis prevalence. The first student used a $Beta(1.76, 9.98)$ and the second student used a $Beta(47.66, 270.08)$. The density plots of the two distributions is given below:



(a) What is the prior expected prevalence given by the two students ?

(b) After data collection, the researchers found that there were 70 cases of osteoarthritis in 500 people surveyed. By looking at the graphs above, describe how well informed the students were about osteoarthritis prevalence, assessing at the same time their own uncertainty.

(c) The graph below gives the posterior distribution of osteoarthritis prevalence for each of the two student, along with the prior distribution.

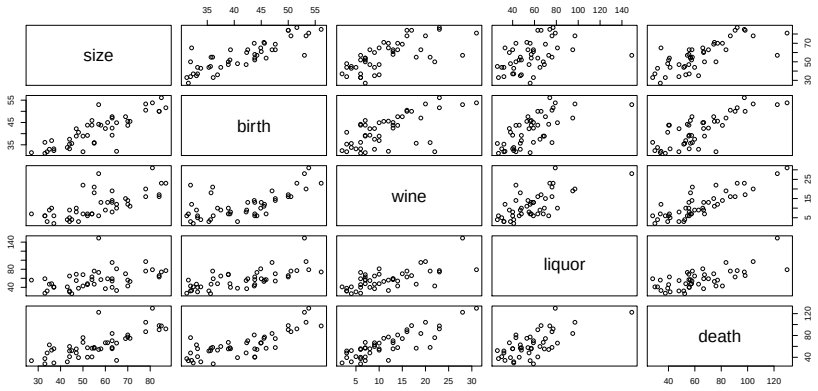


Compute the posterior expected prevalence obtained by each of the two students and explain why they differ.

3. Population and drinking data was recorded for 46 states in USA. Measured variables include:

- *size*: the size of the urban population (in percent)
- *birth*: the number of late births, i.e. births to women between 45 to 49
- *wine*: the consumption of wine per capita
- *liquor*: the consumption of hard liquor per capita
- *death*: the death rate from cirrhosis

The investigator want to investigate the risk factors associated with a death from cirrhosis. The scatterplots of the data is given below, as well as some measures of correlation between variables:




```
> cor(data)
            size     birth      wine    liquor     death
size   1.0000000 0.8432812 0.6786230 0.4402957 0.7490740
birth  0.8432812 1.0000000 0.6398407 0.6863643 0.7827244
wine   0.6786230 0.6398407 1.0000000 0.6759206 0.8446112
liquor 0.4402957 0.6863643 0.6759206 1.0000000 0.6819694
death  0.7490740 0.7827244 0.8446112 0.6819694 1.0000000
```

The confidence intervals for different regression models are given below:

```
> confint(lm(death~size))
                 2.5 %    97.5 %
(Intercept) -16.7556355 18.237103
size          0.8156752  1.415102

> confint(lm(death~birth))
                2.5 %     97.5 %
(Intercept) -71.039842 -18.096531
birth         1.975977   3.234817

> confint(lm(death~wine))
               2.5 %    97.5 %
(Intercept) 22.917596 37.751744
wine         2.310594  3.412877

> confint(lm(death~liquor))
              2.5 %     97.5 %
(Intercept) 7.485087 36.4448077
liquor      0.486901  0.9575696

> confint(lm(death~size+birth+wine+liquor))
                  2.5 %    97.5 %
(Intercept) -36.98657020 9.0603700
size         -0.39461463 0.5911864
birth        -0.02900569 2.3257598
wine          1.04810958 2.6676125
liquor       -0.22114728 0.3174877
```

(a) Comment on the changes in results between the simple and multiple regression models. Explain in particular why the late birth variable is significant in the simple model and not in the multiple model.

4. Here, we consider again the previous study on risk factors associated with a death from cirrhosis. Now a Bayesian strategy is used to analyze the data and select the best model. The R output of the *bicreg* function is shown below:

```
> output$postprob
[1] 0.7683488 0.1168226 0.1148286

> output$namesx
[1] "size"   "birth"  "wine"   "liquor"

> output$label
[1] "birthwine"       "sizebirthwine"   "birthwineliquor"

> output$r2
[1] 81.278 81.303 81.289

> output$bic
[1] -69.41438 -65.64720 -65.61277

> output$size
[1] 2 3 3


> output$postmean
  (Intercept)          size         birth          wine        liquor
-15.897208505   0.005481668   1.354532839   1.966874734   0.001938600

> output$postsd
[1] 10.23797356  0.06876643  0.30962930  0.29773110  0.03675592

> output$ols
     (Intercept)       size    birth     wine     liquor
[1,]   -16.00084 0.00000000 1.365640 1.972298 0.00000000
[2,]   -15.33613 0.04692303 1.292987 1.947158 0.00000000
[3,]   -15.77463 0.00000000 1.342830 1.950646 0.01688255

> output$se
     (Intercept)      size     birth      wine    liquor
[1,]    10.15303 0.0000000 0.2857977 0.2909153 0.0000000
[2,]    10.63615 0.1963011 0.4193937 0.3123935 0.0000000
[3,]    10.37030 0.0000000 0.3234079 0.3248586 0.1072989
```

(a) Explain briefly what each of the outputs above represent (postprob, namesx, r2, bic, size, postmean, postsd, ols and se).

(b) Write down the best model selected according to the analysis above.

5. A study wishes to examines the combinatorial effect of two medications (MedA and MedB) on blood pressure (BP). The fitted linear model is the following:

$$BP = \alpha + \beta_1 MedA + \beta_2 MedB + \beta_3 MedA \times MedB,$$

where MedA and MedB have two levels: 1 (presence) or 0 (absence).

(a) Interpret the null hypothesis $H_0 : \beta_1 = 0$

(b) Interpret the null hypothesis $H_0 : \beta_2 = 0$

(c) Interpret the null hypothesis $H_0 : \beta_1 = \beta_2$

(d) Interpret the null hypothesis $H_0 : \beta_3 = 0$