

# Data Analysis in the Health Sciences

Midterm Exam 2011 – EPIB-621

Student's Name: \_\_\_\_\_

Student's Number: \_\_\_\_\_

## INSTRUCTIONS

This examination consists of 5 questions on 17 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 survey is being conducted to estimate the prevalence of depression among the elderly (65 years and older) in Montreal.

(a) Provide a prior beta distribution representing your knowledge about the prevalence of depression among the elderly in Montreal. Explain your reasoning, and do not use a uniform density, as you surely do not believe that a prevalence of 99% is equally likely as a prevalence of 15%.

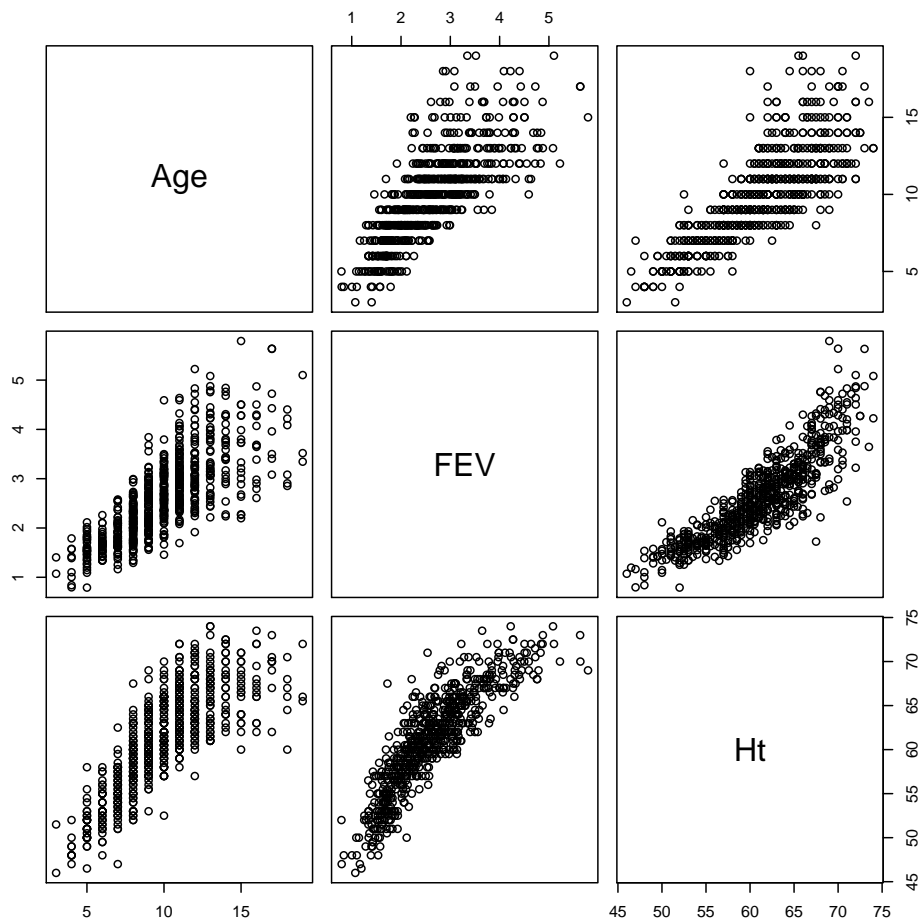
(b) Suppose that in a randomly selected sample of the Montreal elderly population, 196 out of 1000 subjects are found to be depressed. Provide your posterior density for the prevalence of depression in this population.

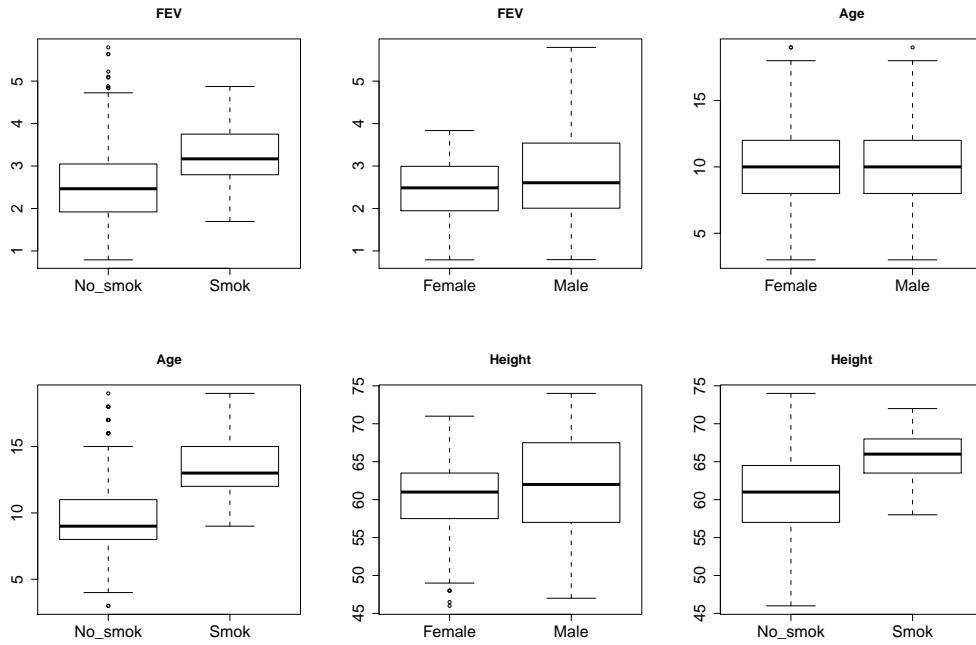
(c) Provide the mean and standard deviation of your posterior distribution given in part (b).

2. The following data set gives information on the health and smoking habits of a sample of 654 youths, aged 3 to 19, in the area of East Boston during middle to late 1970s. Five variables were measured:

- **Age:** The age of the subject in completed years
- **FEV:** The forced expiratory volume, a measure of lung capacity, in litres
- **Ht:** Height (in inches)
- **Gender:** The gender of the subject. Females are coded as 0 and males as 1.
- **Smoke:** The smoking status of the subject: 0 means a non-smoker and 1 means a smoker

The goal of the study was to investigate the relationship between respiratory function (measured by forced expiratory volume, FEV) and smoking. Some descriptive plots investigating relationships between variables are given below and on the next page:





(a) Comment on the relationships between the variables i) FEV and Age, ii) FEV and Smoking, iii) FEV and Height.

(b) Provide the insights the plots give you about potential confounding amongst the four independent variables.

(c) The statistician who analyzed these data fitted simple linear models as well as a multiple linear model including all variables. Confidence intervals for the model parameters are given in the table below. Examining the results below, the graphs given above, and keeping the main objective of the study in mind, do you think that the multiple linear regression model including all variables is appropriate? If yes, explain why. If not, state what model might you prefer to report, and why.

Model	Confidence interval			
	Age	Ht	Gender	Smoke
Age	(0.20 , 0.23)			
Ht		(0.126 , 0.13)		
Gender			(0.23 , 0.49)	
Smoke				(0.49 , 0.92)
Age+Ht+Gender+Smoke	(0.04 , 0.08)	(0.09 , 0.11)	(0.09 , 0.22)	(-0.20 , 0.02)

3. The data below come from a study investigating a new method of measuring body composition. The body fat percentage, age and gender is given for 18 adults aged between 23 and 61.

```
> fat
  Age Percent.Fat Gender
1  23      9.5      M
2  23     27.9      F
3  27      7.8      M
4  27     17.8      M
5  39     31.4      F
6  41     25.9      F
7  45     27.4      M
8  49     25.2      F
9  50     31.1      F
10 53     34.7      F
11 53     42.0      F
12 54     29.1      F
13 56     32.5      F
14 57     30.3      F
15 58     33.0      F
16 58     33.8      F
17 60     41.1      F
18 61     34.5      F
```

(a) Two simple linear models were fitted to the data. The R outputs are shown below:

```
> multiple.regression.with.ci(lm(Percent.Fat~Age))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.2209	5.0762	0.635	0.535
Age	0.5480	0.1056	5.191	8.93e-05 ***

---

Residual standard error: 5.754 on 16 degrees of freedom

Multiple R-squared: 0.6274, Adjusted R-squared: 0.6041

F-statistic: 26.94 on 1 and 16 DF, p-value: 8.93e-05

\$intercept.ci

[1] -7.540113 13.981834

\$slopes.ci

[1] 0.3241815 0.7718005

```
> multiple.regression.with.ci(lm(Percent.Fat~Gender))
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.321      1.573   20.548 6.31e-13 ***
GenderM       -16.696      3.337   -5.004 0.00013 ***
---
```

```
Residual standard error: 5.886 on 16 degrees of freedom
```

```
Multiple R-squared: 0.6101,    Adjusted R-squared: 0.5857
```

```
F-statistic: 25.04 on 1 and 16 DF,  p-value: 0.0001299
```

```
$intercept.ci
```

```
[1] 28.98681 35.65604
```

```
$slopes.ci
```

```
[1] -23.770216 -9.622642
```

(a) Interpret the values of the estimated coefficients corresponding to the Gender and Age variables.

(b) Now, a multiple linear model with an interaction term is fitted, and the R output is given below:

```
> multiple.regression.with.ci(lm(Percent.Fat~Age+Gender+Age*Gender))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	20.1116	6.2395	3.223	0.00613	**
Age	0.2401	0.1204	1.994	0.06600	.
GenderM	-29.2692	10.4098	-2.812	0.01386	*
Age:GenderM	0.5725	0.2893	1.978	0.06790	.

---

\$intercept.ci

```
[1] 6.729346 33.493927
```

\$slopes.ci

	[,1]	[,2]
[1,]	-0.01814617	0.4983065
[2,]	-51.59598452	-6.9424088
[3,]	-0.04812184	1.1930474

Estimate the fat percentage difference between a 50 year old male and a 50 year old female.

(c) What is the interpretation of the coefficient for Age alone in the multivariate model with interaction?

4. A linear regression model is run, with dependent variable  $y$  being predicted by independent regression variables  $x$ ,  $w$  and  $z$ . The bicreg program is run on these data, and the output is as follows:

```

Posterior probabilities(%):
  x      w      z
100.0  39.2 100.0

Coefficient posterior expected values:
(Intercept)          x          w          z
      -0.3546      2.8882      -0.4151      2.0920

> output$label
[1] "xz"  "xwz"

> output$postprob
[1] 0.6078369 0.3921631

> output$probne0
[1] 100.0  39.2 100.0

> output$postmean
[1] -0.3546447  2.8881931 -0.4150666  2.0919847

> output$postsd
[1] 3.1293071 0.5341566 0.6224611 0.1211106

> output$ols
      Int          x          w          z
[1,]  1.059963  2.912511  0.000000  2.092273
[2,] -2.547229  2.850501 -1.058403  2.091538

> output$se
      Int          x          w          z
[1,]  2.312528  0.5357086  0.0000000  0.1217458
[2,]  2.961876  0.5295396  0.5541594  0.1201181

```

(a) Based on the above output, what intercept and values for the coefficients of  $x$ ,  $w$  and  $z$  would you use if you wanted to make future predictions for  $y$  ?

(b) Use your equation from part (a) to predict  $y$  when  $x = 1$ ,  $w = 1$ , and  $z = -1$ .

(c) Write down a 95% confidence interval for  $w$ , based on model [2].

5. State which of the following statements are true or false, and explain why:

(a) If the residuals from a model are extremely close to normally distributed, then the model must provide a very good fit to the data.

(b) Suppose that one wants to investigate the effect on blood pressure (BP) of two medications (A and B), whose effects may vary according to age. Here, age is a variable with 4 categories: < 20, 20 – 30, 30 – 40 and > 40 years old. Treatment is a variable with 2 categories: A and B. Suppose one fits the following interaction model with R:

```
> lm(BP~Age+Treatment+Age*Treatment)
```

where Age and Treatment are both factor variables. The R output will be formatted as in the (empty) table below:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)				
Age<20				
Age20-30				
Age30-40				
TreatmentA				
Age<20:TreatmentA				
Age20-30:TreatmentA				
Age30-40:TreatmentA				

True or False: Under this model, the coefficient corresponding to “Age20-30:TreatmentA” can be interpreted as *the average BP difference between a 20-30 year old individual who receives treatment A and a 20-30 year old individual who receives treatment B.*

## 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.