Data Analysis in the Health Sciences

Midterm Exam 2015 – EPIB–621

Student's Name: _____

Student's Number: _____

INSTRUCTIONS

This examination consists of 6 questions on 14 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 60.

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

Total (out of 60) _____

1. Suppose you are interested in estimating the prevalence of celiac disease in Montreal. You read the literature, and find that the prevalence is previously reported to be approximately 1% in North American populations.

(a) Find the beta prior density that has a mean of 1%, and a standard deviation of 5%. The large SD is because you are unsure whether the results from past literature are accurate and apply well to the Montreal population.

(b) What is the approximate "sample size equivalent" of the beta prior density as calculated in (a)?

(c) Suppose that in a survey of 50 Montrealers, 1 of them has celiac disease. State the posterior distribution that arises from combining these data with the prior distribution from part (a).

(d) What is the mean and standard deviation of your posterior density from part (c)?

2. A researcher runs a linear regression on the following data set, where X is the independent variable and Y is the dependent variable:

X	Y
1	2
2	4
3	5
4	8
5	9
6	11

The researcher calculates the slope of the regression line to be 1.8. What is the intercept?

3. Suppose that one wants to know whether troponin is useful as a marker for cardiovascular disease (CVD). Data on troponin is collected for 50 subjects, 25 with CVD and 25 without CVD. A linear regression is run on these 50 subjects, with troponin as the dependent variable, and with an independent dummy variable coded as 1 in the presence of CVD and 0 in the absence of CVD.

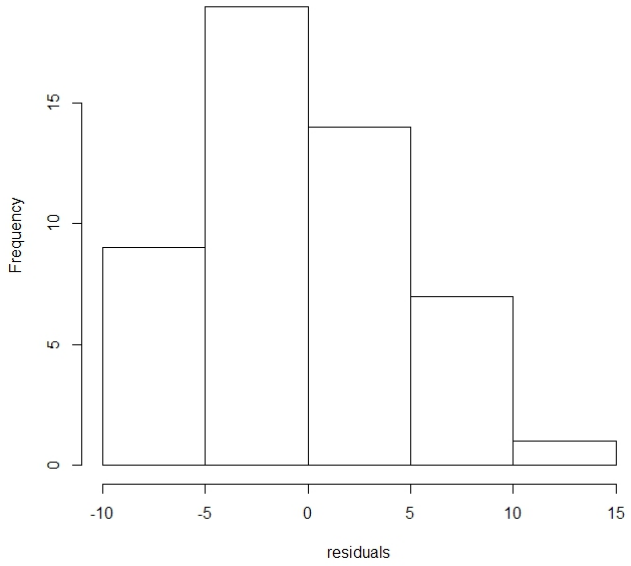
The output from the linear regression model is given below:

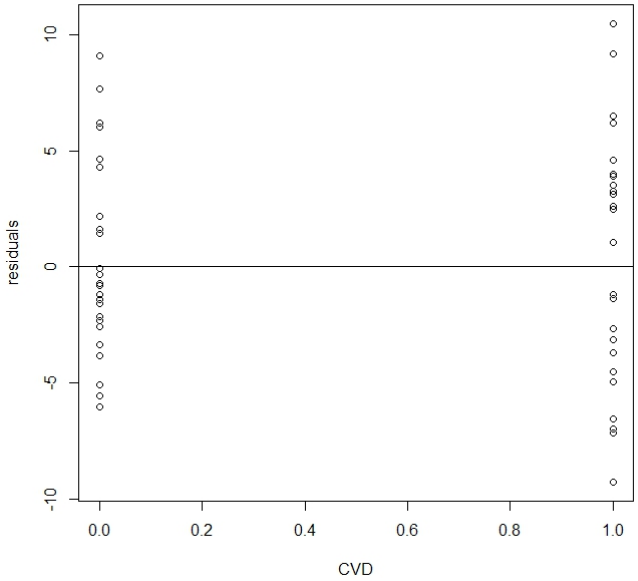
```
> out <- lm(troponin ~ CVD)
> summary(out)

Call:
lm(formula = troponin ~ CVD)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2926 -3.6250 -0.7772  3.7886 10.4881

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.161      1.006   7.119 4.82e-09 ***
CVD            3.231      1.423   2.271   0.0276 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.029 on 48 degrees of freedom
Multiple R-squared:  0.09706,   Adjusted R-squared:  0.07825
F-statistic:  5.16 on 1 and 48 DF,  p-value: 0.02764

> confint(out)
                2.5 %   97.5 %
(Intercept) 5.1388843 9.183823
CVD         0.3710791 6.091486
```

(a) What troponin value would you predict for a subject without CVD? Similarly, what troponin value would you predict for a subject with CVD?

(b) Provide an interpretation for the confidence interval for the slope coefficient for CVD. Your interpretation should be practical, and so not necessarily "technically correct."

(c) From the histogram and scatter plot of residuals, do the assumptions of linear regression seem to be satisfied for the linear model relating troponin and CVD? List three assumptions, and state whether each appears satisfied or not, and why.





4. A group of researchers are investigating the possibility of confounding between two continuous independent variables X_1 and X_2. The estimated correlation coefficient between X_1 and X_2 is 0.806. They calculate the following regression models using the bicreg program:

```
Call:
bicreg(x = x, y = y, OR = 99999999)
  2  models were selected
 Best  2  models (cumulative posterior probability =  1 ):

           p!=0    EV        SD       model 1     model 2
Intercept  100.0   0.910235  0.09306     0.91028     0.90902
X1         100.0   2.966256  0.09905     2.96398     3.02956
X2           3.5  -0.002796  0.03317       .        -0.08059

nVar                                       1           2
r2                                       0.492       0.492
post prob                                0.965       0.035
> b$mle
     (Intercept)       X1          X2
[1,]   0.9102784 2.963981  0.00000000
[2,]   0.9090223 3.029561 -0.08059439
> b$se
     (Intercept)         X1        X2
[1,]  0.09306126 0.09529741 0.0000000
[2,]  0.09312918 0.16103339 0.1594941
```

Based on all information given, do you think there is likely to be appreciable confounding between X_1 and X_2 in this regression scenario? Explain your answer.

5. Hearing loss generally increases with age, and is also affected by noise levels at work. A study is carried out that collects data on hearing loss (measured in decibels or dBs) across a wide age range, also recording whether each subject worked in a noisy or quiet environment.

The study design included the following sample sizes:

Age range	Environment	Sample size
20 - 29	Quiet	100
20 - 29	Noisy	100
30 - 39	Quiet	100
30 - 39	Noisy	100
40 - 49	Quiet	100
40 - 49	Noisy	100
50 - 59	Quiet	100
50 - 59	Noisy	100

(a) Do you think that the researchers need to concern themselves with confounding between the age and environment variables? Explain why or why not.

(b) A regression is run with hearing loss as dependent variable, and two categorical variables, age group coded as 0, 1, 2, 3 for the categories 20-29, 30-39, 40-49 and 50-59, respectively, and environment (called environ) coded as 0 for quiet and 1 for noisy. The outcome is continuous and coded in dBs, with larger negative numbers indicating greater hearing loss. The results of the regression model are given below:

```
Call:
lm(formula = dBs ~ age + environ)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.1361     0.2354 -26.063  < 2e-16
age1         -0.8054     0.2978  -2.704  0.00699
age2         -1.4945     0.2978  -5.019 6.42e-07
age3         -2.8539     0.2978  -9.583  < 2e-16
environ      -3.1477     0.2106 -14.948  < 2e-16
Residual standard error: 2.978 on 795 degrees of freedom

> confint(out)
                2.5 %     97.5 %
(Intercept) -6.598205 -5.6739377
age1        -1.389931 -0.2208146
age2        -2.079092 -0.9099753
age3        -3.438435 -2.2693186
environ     -3.561065 -2.7343752
```

Provide a clinically relevant interpretation for the results for environ variable.

(c) Suppose that the researchers considered the age variable as a continuous rather than as a categorical variable. In other words, rather than three separate estimates for age, they took the variable, still coded as 0, 1, 2, and 3, and entered that variable as a continuous variable. Provide your approximation to what the estimated coefficient for this continuous age variable would be, and explain your reasoning.

6. Continuing the scenario from problem #5, an interaction term is now added to the model. Age is returned to being a categorical variable coded with four categories as in 5 (b).

(a) Explain the motivation for including an interaction term in the model. In other words, explain what it would mean in practice if there were indeed an interaction between age and environ.

(b) The model is run returning the results below :

```
Call:
lm(formula = dBs ~ age + environ + age * environ)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.0242     0.2982 -20.205  < 2e-16
age1          -1.0995     0.4216  -2.608 0.009291
age2          -1.5482     0.4216  -3.672 0.000257
age3          -2.9535     0.4216  -7.005 5.29e-12
environ       -3.3714     0.4216  -7.996 4.55e-15
age1:environ   0.5882     0.5963   0.986 0.324237
age2:environ   0.1074     0.5963   0.180 0.857163
age3:environ   0.1992     0.5963   0.334 0.738430
Residual standard error: 2.982 on 792 degrees of freedom

> confint(out)
                   2.5 %     97.5 %
(Intercept)  -6.6094865 -5.4389697
age1         -1.9271492 -0.2717885
age2         -2.3758946 -0.7205339
age3         -3.7811544 -2.1257937
environ      -4.1990875 -2.5437268
age1:environ -0.5823250  1.7587086
age2:environ -1.0631554  1.2778782
age3:environ -0.9713226  1.3697110
```

Provide your conclusions about the interaction between age and environ.

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.