Data Analysis in the Health Sciences

Midterm Exam 2010 – EPIB–621

Student's Name: _____

Student's Number: _____

INSTRUCTIONS

This examination consists of 5 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 50.

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

Total (out of 50) _____

1. Suppose you are interested in the proportion, π, of downtown workers in Montreal whose commute to work from home lasts 45 minutes or longer each day.

(a) Before collecting any data, provide a beta prior probability density that best summarizes your knowledge about the parameter π. Explain your reasoning.

(b) Suppose that a survey of 200 downtown workers finds that 60 of them have commutes that are 45 minutes or longer, the rest all having commutes shorter than 45 minutes. State the posterior distribution that arises from combining these data with your answer from part (a).

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

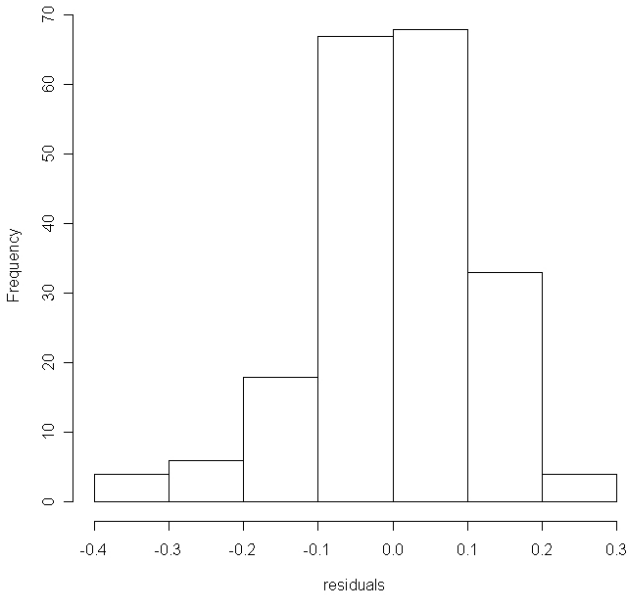
2. Suppose that the dosage of a certain drug, measured in milligrams (mg) is thought to be linearly related to the logarithm of total cholesterol levels, measured in milligrams per deciliter (mg/dL). Normal levels are considered to be 200 mg/dL or below, and elevated levels are 240 mg/dL and above. Values between these limits are moderately high. The average age of subjects in this study is 50 years old.

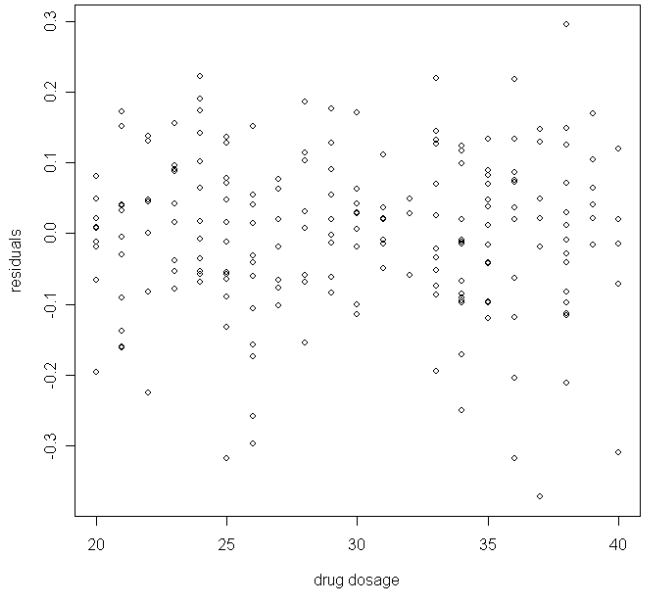
(a) The output from the linear regression model is given below:

```
> out <- lm(log.chol ~ dose)
> summary(out)
Call: lm(formula = log.chol ~ dose)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37279 -0.06578  0.01013  0.07212  0.29633

Coefficients:
             Estimate   Std. Error  t value Pr(>|t|)
(Intercept)  5.957392   0.040205   148.18    <2e-16 ***
dose        -0.014243   0.001324   -10.76    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1117 on 198 degrees of freedom
Multiple R-Squared: 0.3688,     Adjusted R-squared: 0.3656
F-statistic: 115.7  on 1 and 198 DF,  p-value: < 2.2e-16

> confint(out)
                  2.5 %      97.5 %
(Intercept)  5.87810710  6.03667731
dose        -0.01685436 -0.01163165
```

For a subject on a drug dosage of 30 mg, what would you predict the total cholesterol (not log(cholesterol)) to be?

(b) From the histogram and scatter plot of residuals, do the assumptions of linear regression seem to be satisfied for the linear model relating log(cholesterol) and drug dosage?





3. This problem continues to consider the situation outlined in question number 2. It is thought that age may mediate the effect of drug dosage, so that the following model is run:

$$log(cholesterol) = \alpha + \beta_1 \times dose + \beta_2 \times age.c + \beta_3 \times age.c \times dose$$

```
> out2 <- lm(log.chol ~ dose + age.c + age.c.dose)
> summary(out2)

Call:
lm(formula = log.chol ~ dose + age.c + age.c.dose)

Residuals:
      Min        1Q    Median        3Q       Max
-0.162979 -0.045340 -0.003401  0.044864  0.178427

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.0097474  0.0257564  233.33   <2e-16 ***
dose        -0.0163587  0.0008486  -19.28   <2e-16 ***
age.c        0.1036349  0.0046565   22.26   <2e-16 ***
age.c.dose  -0.0021123  0.0001564  -13.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07147 on 196 degrees of freedom
Multiple R-Squared: 0.9353,     Adjusted R-squared: 0.9343
F-statistic: 943.9 on 3 and 196 DF,  p-value: < 2.2e-16
```

The variable age.c is a centered version of the variable age, that is, it is constructed by taking the age of each subject and subtracting the mean value of 50 years.

(a) From the information given above, provide a 95% confidence interval for the intercept. What does the intercept represent clinically in this model?

(b) Would you have any concerns about reporting your confidence interval from (a)? Explain why or why not.

(c) What is the effect of a one mg change in dose on log(cholesterol) for a subject aged 45?

4. A group of researchers are investigating the possibility of confounding between a dichotomous variable X_1 and a continuous variable X_2, and their effects on an outcome variable Y. They calculate the following descriptive statistics:

Variable	Mean	Standard deviation	Min	Max
X_1	0.6	0.48	0	1
X_2	5	2	0	10
Y	63	10	40	105

Further, they calculate that within the subset of subjects with $X_1 = 0$, that $\overline{X}_2 = 3.5$, and within the subset of subjects with $X_1 = 1$, that $\overline{X}_2 = 6$.

The following linear regression model is estimated:

$$Y = 68.2 + 8X_1 - 2X_2$$

(a) 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.

(b) Based on all information given, provide your best estimate of the value for β_1 if the following simple linear regression model were run:

$$Y = \alpha + \beta_1 X_1$$

5. In a rehabilitation centre, researchers would like to predict how long it takes patients to walk 50 meters, based on various covariates. The outcome is measured in seconds, and possible predictor variables include:

Variable	name	coding
age	age	continuous in years
sex	sex	0 = male, 1 = female
cardiovascular disease	cardio	0 = no, 1 = yes
previous hip fracture	hip	0 = no, 1 = yes
global fitness measure (fit)	fit	scale from 0 to 10 with 0 = poorest and 10 = best fitness

The data are analyzed using the bicreg program, and the following results are observed:

```
> out <- bicreg(x, seconds, OR=100000)
> out

bicreg(x = x, y = seconds, OR = 1e+05)

> out$namesx
[1] age sex cardio hip fit

> out$probne0

[1] 100.0  99.8  8.3 14.2 30.4 100.0

> out$postmean

[1] 39.7967 1.1215 0.2033 0.9983 2.5014 2.6135

> out$postsd

[1] 15.9732 0.2495 1.2418 3.1969 4.4747 0.4935



> out$ols
            Int       age      sex   cardio       hip      fit
 [1,]  38.65101 1.1473398 0.000000 0.000000  0.000000 2.615448
 [2,]  40.46383 1.1056072 0.000000 0.000000  8.104262 2.574995
 [3,]  38.64872 1.1328754 0.000000 6.545365  0.000000 2.692838
 [4,]  44.31200 1.0410069 2.367607 0.000000  0.000000 2.601171
 [5,]  40.62198 1.0846548 0.000000 7.805405  8.823481 2.663693
 [6,]  45.43103 1.0124141 2.088086 0.000000  7.990454 2.562972
```

```
 [7,]  44.59128 1.0210182 2.485384 6.651629  0.000000 2.679107
 [8,]  45.86289 0.9861466 2.202482 7.883457  8.710630 2.651898
 [9,] 107.32274 0.0000000 9.362663 0.000000  0.000000 2.558746
[10,] 106.64651 0.0000000 8.841218 0.000000  8.834053 2.517802
[11,] 106.23744 0.0000000 9.343377 7.757345  0.000000 2.650587
[12,] 105.31481 0.0000000 8.771181 9.078562  9.638200 2.621559
[13,] 111.71319 0.0000000 0.000000 0.000000  0.000000 2.615065
[14,] 110.67213 0.0000000 0.000000 0.000000  9.988368 2.565225
[15,] 109.27332 0.0000000 0.000000 9.312992 10.803900 2.671275
[16,] 110.60753 0.0000000 0.000000 7.837480  0.000000 2.707738
[17,]  52.17278 1.1471358 0.000000 0.000000  0.000000 0.000000
[18,]  53.97447 1.1002542 0.000000 0.000000  9.104887 0.000000
> out$se
            Int       age      sex   cardio      hip       fit
 [1,] 15.438370 0.2380312 0.000000 0.000000 0.000000 0.4924561
 [2,] 15.371133 0.2375700 0.000000 0.000000 4.321787 0.4898159
 [3,] 15.420191 0.2380511 0.000000 5.408129 0.000000 0.4960151
 [4,] 17.566978 0.2852557 3.489187 0.000000 0.000000 0.4935805
 [5,] 15.329017 0.2373563 0.000000 5.401026 4.338474 0.4923022
 [6,] 17.470913 0.2839483 3.471323 0.000000 4.332972 0.4910223
 [7,] 17.545741 0.2853516 3.485996 5.416977 0.000000 0.4970117
 [8,] 17.423348 0.2837076 3.462262 5.410681 4.348732 0.4934040
 [9,]  3.336792 0.0000000 3.005311 0.000000 0.000000 0.5086367
[10,]  3.329719 0.0000000 2.994740 0.000000 4.453905 0.5053126
[11,]  3.418808 0.0000000 2.998194 5.568610 0.000000 0.5116918
[12,]  3.414108 0.0000000 2.982335 5.551062 4.462171 0.5071512
[13,]  3.090359 0.0000000 0.000000 0.000000 0.000000 0.5193698
[14,]  3.096629 0.0000000 0.000000 0.000000 4.522760 0.5148529
[15,]  3.198244 0.0000000 0.000000 5.657769 4.530436 0.5166661
[16,]  3.186167 0.0000000 0.000000 5.690342 0.000000 0.5225472
[17,] 16.239493 0.2538588 0.000000 0.000000 0.000000 0.0000000
[18,] 16.146843 0.2531198 0.000000 0.000000 4.600237 0.0000000

> out$label
 [1] "agefit"           "agehipfit"       "agecardiofit"     "agesexfit"
 [5] "agecardiohipfit"  "agesexhipfit"    "agesexcardiofit"  "agesexcardiohipfit"
 [9] "sexfit"           "sexhipfit"       "sexcardiofit"     "sexcardiohipfit"
[13] "fit"              "hipfit"          "cardiohipfit"     "cardiofit"
[17] "age"              "agehip"

> out$postprob
 [1] 5.546673e-01 2.319784e-01 8.255698e-02 4.960276e-02 4.763086e-02 1.976267e-02
 [8] 4.145945e-03 9.788572e-04 5.047180e-04 1.853217e-04 1.394268e-04 1.129378e-04
[15] 2.570613e-05 2.081029e-05 1.210496e-05 6.128430e-06
```

(a) Write down the model and model coefficients best suited for making future predictions, based on all output given.

(b) Using your model from (a), what would your mean prediction be for the number of seconds it would take a 50 year old female with no cardiovascular disease or previous hip fracture and a global fitness level of 5 to walk 50 meters?

(c) If you had to use a model with just two variables to make future predictions, which model would you use? Write down the intercept and both beta coefficients for this model. [This may sometimes be of interest to create a simple model for clinical use, for example.]

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.