Data Analysis in the Health Sciences

Midterm Exam 2007 – EPIB–621

Student's Name: _____

Student's Number: _____

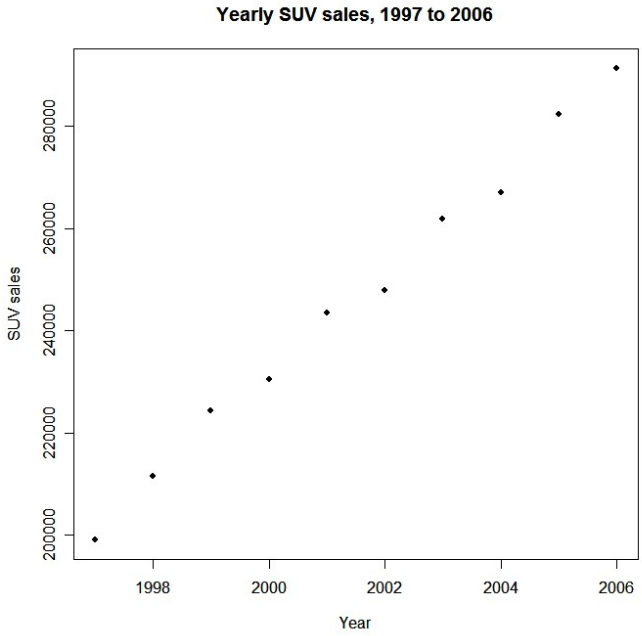
INSTRUCTIONS

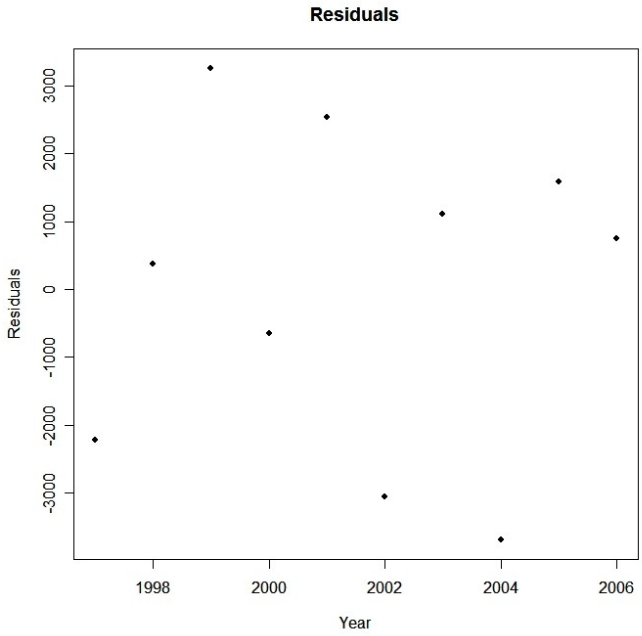
This examination consists of 5 questions on 13 pages, including this one. Please write your answers (NEATLY) in the spaces provided. Fully explain all of your answers. Each question is worth 10 points, for a total of 50.

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

Total (out of 50) _____

1. Sales of Sports Utility Vehicles (SUVs) have been increasing in Canada, despite environmental concerns. Statistics Canada tracks these sales over time. Suppose that annual SUV sales (Y) from 1997 to 2006 (X) are entered into a simple linear regression model, and the results of the regression are summarized below:






```
> summary(lm(SUV ~ year))

Call: lm(formula = SUV ~ year)

Residuals:
    Min      1Q  Median      3Q     Max
-3692.4 -1821.5   559.3  1460.4  3261.3

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.964e+07  5.493e+05  -35.76 4.10e-10 ***
year         9.936e+03  2.744e+02   36.20 3.71e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2493 on 8 degrees of freedom
Multiple R-Squared: 0.9939,     Adjusted R-squared: 0.9932
F-statistic:  1311 on 1 and 8 DF,  p-value: 3.712e-10
```

(a) State the intercept of the regression line, and provide an interpretation for the intercept of the regression line.

(b) State the slope of the regression line, and provide an interpretation for the slope of the regression line.

(c) Provide an approximate confidence interval for the slope (Suggestion: You can assume the sample size is large enough to use the normal approximation rather than the t-distribution). Provide an interpretation for this confidence interval (**NOT** a "technical" interpretation, but rather one that comments on the substantive problem).

(d) Based on the information given, is it reasonable to say that the assumptions required by the linear regression model appear satisfied?

2. A regression with outcome variable Y and three independent variables, X_1, X_2, and X_3 is run, with results from the bicreg program given below.

```
> output <- bicreg(x=x.matrix, y=y, OR=1e100)
> summary(output)

Call:
bicreg(x = x.matrix, y = y, OR = 1e+100)


  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
Int  100.0  -0.05210  0.09576  -0.05266   -0.04973    -0.07071    -0.09204    -0.03618
x1   100.0   1.15284  0.11045   1.12902    1.25367     0.80252     1.13427        .
x2   100.0   1.02283  0.17751   0.97082    1.24299        .           .           .
x3   19.1   -0.02536  0.06889      .       -0.13272     0.34908        .        0.51582

nVar                                2           3           2           1           1
r2                                0.714       0.719       0.638       0.473       0.448
BIC                            -115.89610  -113.01074   -92.34553   -59.49775   -54.75579
post prob                         0.809       0.191       0.000       0.000       0.000

> output$label
[1] "x1x2"   "x1x2x3" "x1x3"   "x1"     "x3"     "x2x3"   "x2"     "NULL"
> output$postprob
[1] 8.088646e-01 1.911291e-01 6.222072e-06 4.582792e-13 4.279822e-14 6.621575e-15
[7] 6.959399e-21 5.512611e-26

> output$ols
              Int        x1         x2         x3
[1,] -0.052664863 1.1290151  0.9708204  0.0000000
[2,] -0.049734480 1.2536694  1.2429924 -0.1327184
[3,] -0.070711609 0.8025216  0.0000000  0.3490848
[4,] -0.092037030 1.1342710  0.0000000  0.0000000
[5,] -0.036181765 0.0000000  0.0000000  0.5158193
[6,] -0.043632576 0.0000000 -0.2293053  0.5874101
[7,] -0.009635495 0.0000000  0.9796291  0.0000000
[8,] -0.049164444 0.0000000  0.0000000  0.0000000

> output$se
            Int         x1        x2         x3
[1,] 0.09580516 0.08956401 0.1075196 0.00000000
[2,] 0.09550561 0.13153535 0.2366048 0.10286511
[3,] 0.10771351 0.11246712 0.0000000 0.05258603
[4,] 0.12917480 0.12088254 0.0000000 0.00000000
```

```
[5,] 0.13219876 0.00000000 0.0000000 0.05787666
[6,] 0.13254677 0.00000000 0.2487342 0.09687826
[7,] 0.15471692 0.00000000 0.1737413 0.00000000
[8,] 0.17696924 0.00000000 0.0000000 0.00000000
```

(a) Which variables are included in the best model? Which variables are included in the second best model?

(b) What model would you use (provide intercept and slopes for all parameters that will be included in that model) if you wanted to make optimal predictions?

(c) Based on the above outputs, discuss whether any variables may be confounded. Your discussion should include mention of the clues you are using to help you investigate the presence of confounding.

3. Lupus patients are often plagued by both pain and fatigue. An experiment is designed to verify whether the amount of pain felt on one day affects the amount of fatigue on the next day. Age is thought to be a possibly confounding variable, and so is also added to the model. A multiple linear regression is run with the logarithm of fatigue (logFAT) as the outcome variable, and age and pain as the independent variables. The logarithm (to the base e) of the outcome is used in order for the relationships between dependent and independent variables to be approximately linear. The standard regression output for this model is as follows:

```
> output <- lm(logFAT ~ age + pain)
> summary(output)

Call:
lm(formula = logFAT ~ age + pain)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.86375    4.55380  -0.409  0.68420
age          0.02796    0.00884   3.162  0.00274 **
pain         0.08880    0.02774   3.200  0.00246 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: .5161 on 47 degrees of freedom
Multiple R-Squared: 0.2798,     Adjusted R-squared: 0.2492
F-statistic: 9.131 on 2 and 47 DF,  p-value: 0.0004464

> confint(output)
                 2.5 %    97.5 %
(Intercept)  -7.2973135  11.0248112
age           0.0101714   0.0457502
pain          0.0329846   0.1446328
```

(a) Provide an interpretation for the estimated slope for the pain variable, including consideration of the confidence interval for this parameter.

(b) What information does the $R^2 = 0.2798$ value provide?

(c) Predict fatigue (not log fatigue) for an individual aged 40 years old with pain score of 5.

(d) Do you think the value of the intercept is most likely equal to zero? Explain why or why not.

4. An experiment is conducted where a null hypothesis ($H_0 : \mu = 0$) is tested against a two sided alternative ($H_a : \mu \neq 0$). The sample size is 100, and the observed sample mean and sample standard deviation are given by $\overline{X}$ and s, respectively. The test gives a p-value of 0.07. Later, a new test is conducted, where the sample size is increased to 1000. By coincidence, the observed sample mean and sample standard deviation are the same as in the first experiment. Would you expect the p-value to be the same, higher, or lower than the $p = 0.07$ from the first experiment? Explain your answer.

5. It is well known that quality of life declines, on average, as people age. To quantify this relationship, researchers took 100 males (sex variable coded 1) and 100 females (sex variable coded 0), all aged between 20 and 75 years old, with all 200 persons given the same quality of life questionnaire to fill out. The questionnaire ranges in value from 0 to 100, with higher scores indicating a higher quality of life. From the results of this study, the researchers performed a linear regression analysis, finding the best fitting line to be

$$QoL = 110 - 1.1 \times age + 5 \times sex - 0.3 \times age \times sex$$

where QoL represents the score on the quality of life scale for a given age.

(a) Provide interpretations for the intercept and all three slopes of this regression line.

(b) Predict the expected quality of life for a male aged 30 years old.

(c) Draw rough lines in the plot below showing how age is related to quality of life for both males and females (one line for males, one line for females).