Data Analysis in the Health Sciences

Midterm Exam 2008 – EPIB–621

Student's Name: _____

Student's Number: _____

INSTRUCTIONS

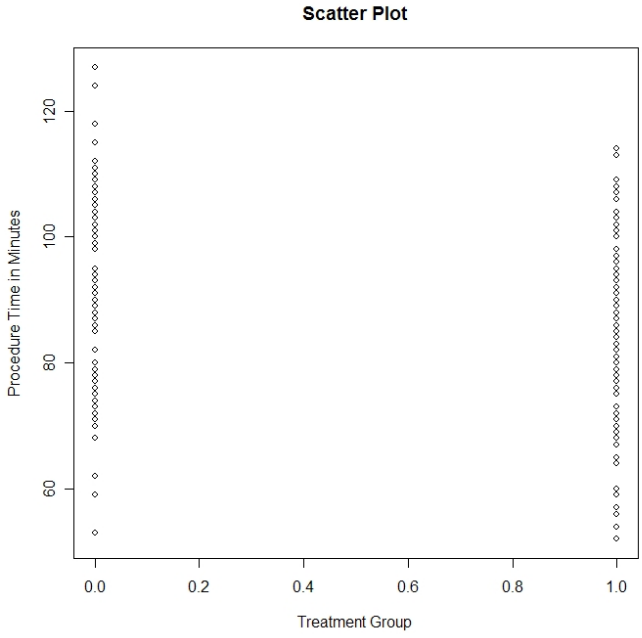
This examination consists of 5 questions on 14 pages, including this one. Please write your answers (NEATLY) in the spaces provided. Fully explain all of your answers. Normal tables are at the back of this exam. Each question is worth 10 points, for a total of 50.

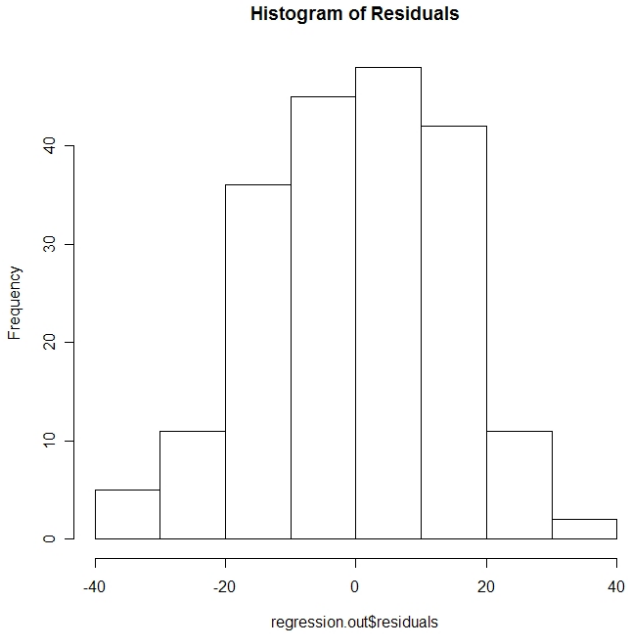
1. _____
2. _____
3. _____
4. _____
5. _____

Total (out of 50) _____

1. Medications are sometimes administered prior to surgery with the aim of reducing certain surgical complications. One crude measure of the rate of complications during surgery is procedure time.

Suppose that a clinical trial is carried out, with one treatment arm being given a new medication intended to reduce surgical complications (group 1), and the other arm given a placebo (group 0). Procedure times are tracked for all subjects, in minutes. Results from a linear model, a scatter plot of the data and a histogram of the residuals are given below.






```
> summary(regression.out)
Call: lm(formula = procedure.time ~ group)

Residuals:
   Min     1Q Median     3Q    Max
-37.52 -10.52   0.48  10.70  36.48

Coefficients:
             Estimate    Std. Error t value  Pr(>|t|)
(Intercept)   90.520      1.422     63.641   <2e-16 ***
group         -4.860      2.012     -2.416   0.0166 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.22 on 198 degrees of freedom
Multiple R-Squared: 0.02864,    Adjusted R-squared: 0.02373
F-statistic: 5.837 on 1 and 198 DF,  p-value: 0.01660
```

(a) Provide the estimated mean procedure time and approximate 95% confidence interval for the placebo group.

(b) Provide the estimated mean procedure time for the treatment group.

(c) Provide the between group difference with approximate 95% confidence interval, and from this, provide an opinion about the effectiveness of the medication as far as procedure times are concerned.

(d) From the information given, do the assumptions behind linear regression seem satisfied here?

2. A survey of Canadian adults is conducted, where data on demographic variables age (in years), sex (% female, females coded as 1) and education (years of school completed) are collected. In addition, data on incomes (in $1000) are collected, and used as the outcome in a linear regression model.

The correlation matrix of all variables collected is given below, followed by the univariate and multivariate outputs from a linear model with income as the outcome, and all demographic variables (age, sex, income) as independent variables. Average age is 45 years old, 50% are females, and average education is 8 years.

Correlation matrix:

	Age	Sex	Education	Income
Age	1	0.2	0.4	0.6
Sex	0.2	1	-0.15	0.3
Education	0.4	-0.15	1	0.5
Income	0.6	0.3	0.5	1

Linear regression results:

	Univariate Results		Multivariate Results	
	Coefficient	95% Confidence Interval	Coefficient	95% Confidence Interval
Age	0.5	(0.4, 0.6)	0.3	(0.15, 0.45)
Sex	-2.0	(-1.5, -2.5)	-1.9	(-1.4, -2.4)
Education	1.5	(1.2, 1.8)	1.2	(0.7, 2.3)

(a) Provide an interpretation for the multivariate coefficient and confidence interval for age.

(b) Predict the mean difference in income for a female aged 32, with 12 years of education, compared to a male aged 35 with 10 years of education.

(c) Using the regression results, discuss any confounding that may have occurred during these analyses.

3. One researcher studies the relationship between age and bone mineral density (BMD) in 1000 women aged 20 to 40 years old, estimating the following regression line:

$$BMD = 0.8 + 0.01 \times age$$

with 95% confidence interval for the slope of age being (0.009, 0.011).

A second researcher studies the same relationship, finding this regression equation

$$BMD = 1.6 - 0.01 \times age$$

in 1000 women aged 40 to 60 years old, with 95% confidence interval for the slope of age being (-0.011, -0.009).

Suppose you are now given data from all 2000 women participating in these two studies.

(a) Draw a graph of these two regression lines (both lines on the same plot).

(b) What would you guess the slope would be if you ran a simple linear regression line using data from all 2000 women?

(c) What type of terms might you add to a multiple linear regression model in order to analyze these data as a whole? Write down the regression equation for this model.

4. A small survey is carried out to estimate the prevalence of osteoarthritis in Quebec. Before the study is conducted, the researchers believe that the prevalence of osteoarthritis in Quebec is likely to be between 10% and 20%. After data collection, the researchers find that there are 70 cases of osteoarthritis in 500 people surveyed.

(a) State a prior distribution that reasonably represents the researchers prior opinion about the prevalence of osteoarthritis.

(b) Write down the researchers posterior distribution, after seeing their data.

(c) Provide the posterior mean and standard deviation from your answer in (b).

5. An outcome variable Y is standardized, that is, the variable

$$Y^* = \frac{Y - \overline{Y}}{s_y}$$

is created, where $\overline{Y}$ is the mean of Y, here equal to 1, and s_y is the standard deviation of Y, here equal to 5. A linear regression of Y^* is then run using independent variable X, with the following results:

```
Call:
lm(formula = y.standardized ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.39100 -0.64269  0.03024  0.59484  1.89114

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01814    0.06292   0.288    0.773
x            0.44644    0.06092   7.328  5.8e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8892 on 198 degrees of freedom
Multiple R-Squared: 0.2134,     Adjusted R-squared: 0.2094
F-statistic:  53.7 on 1 and 198 DF,  p-value: 5.796e-12

>  confint(lm(y.standardized ~ x))
                 2.5 %    97.5 %
(Intercept) -0.1059435 0.1422249
x            0.3263021 0.5665810
```

(a) Provide an interpretation of the slope of this regression equation.

(b) Provide an interpretation of the intercept of this regression equation. Is the value obtained for the intercept what you would expect? Explain why or why not.

(c) Predict the mean of Y (not Y^*) when $x = -1$.

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.