Name: Student Number:

STAT 350: Summer Semester 2008

Midterm 1

9 June 2008 Instructor: Richard Lockhart

Instructions: This is an open book test. You may use notes, text, other books and a calculator. Your presentations of statistical analysis will be marked for clarity of explanation. I expect you to explain what assumptions you are making and to comment if those assumptions seem unreasonable. In general you need not finish doing arithmetic; I will be satisfied if your answers contain things like

$$27 \pm 1.96\sqrt{247.5/11},$$

but I have to be absolutely convinced you know what arithmetic to do! I want the answers written on the paper. The exam is out of 25.

1. The following story is taken from a data library called DASL: "In 1929, Edwin Hubble investigated the relationship between distance of a galaxy from the earth and the velocity with which it appears to be receding. Galaxies appear to be moving away from us no matter which direction we look." His paper records the distance (in megaparsecs) measured to 24 nebulae which lie outside our galaxy. For each of these nebulae Hubble also measured the velocity of the nebulae away from the earth (negative numbers mean the nebula is moving towards the earth).

   Here are the data:

Distance	0.032	0.034	0.214	0.263	0.275	0.275	0.45	0.5
Velocity	170	290	-130	-70	-185	-220	200	290
Distance	0.5	0.63	0.8	0.9	0.9	0.9	0.9	1
Velocity	270	200	300	-30	650	150	500	920
Distance	1.1	1.1	1.4	1.7	2	2	2	2
Velocity	450	500	500	960	500	850	800	1090

   In appendix 1 I fit the model:

$$\text{Velocity} = \beta_0 + \beta_1 \text{Distance} + \text{Error}.$$

   I give SAS code and output, R code and output and corresponding JMP output. Use that output to answer the following questions:

(a) Could the true value of the intercept term be 0? (Hubble's model called for the intercept to be 0.) [3 marks]

(b) Give a 95% confidence interval for the coefficient of Distance in our model. This coefficient is called the Hubble constant. [3 marks]

(c) The matrix $(X^TX)^{-1}$ for this model and this data set is given in Appendix 2. Use this to give a 90% confidence interval for the mean recession velocity of all nebulae whose distance from earth is 1 megaparsec. [6 marks]

(d) When I fit a cubic polynomial model

$$V = \beta_0 + \beta_1 D + \beta_2 D^2 + \beta_3 D^3 + \epsilon$$

to this data I find the error sum of squares is 1062087 whereas the error sum of squares in the straight line model above is 1193442. What is the value of the F statistic for testing the hypothesis that $\beta_2 = \beta_3 = 0$ and what are the relevant degrees of freedom? [3 marks]

2. Consider the sand / fibre / plaster hardness example from class. In that example there were 18 batches of plaster – 2 batches made from each combination of 3 sand contents S and 3 fibre contents, F. I regressed Hardness on S, S^2, F, F^2 and SF.

   (a) If I now regressed on $S, S^2, F, F^2, SF, S^3, S^2F, SF^2$ and F^3 would the error sum of squares:

      i. go up,
      ii. go down,
      iii. stay the same
      iv. or is it not possible to tell without further information?

      [1 mark for the answer and 2 for the explanation]

   (b) Suppose instead I made another 18 batches of plaster again making 2 batches of each of the combinations of S and F. If I now used all 36 data points to fit the same model as in the start of this problem would the error sum of squares:

      (a) be higher than when using the original 18 data points,
      (b) be lower than when using the original 18 data points,
      (c) be the same
      (d) or is it not possible to tell without further information?

      [1 mark for the answer and 2 for the explanation]

(c) Consider the situation described in part (b). When we fit the model to the original 18 data points we will get a certain standard error for the estimated coefficient of F. This standard error will have the form $C\sigma$ for some constant C. When we fit the same model to all 36 data points will the the new standard error be higher or lower than $C\sigma$? [1 mark for the answer, 1 for the explanation and a bonus mark for saying what the new standard error will be]

3. In a study of a large number of sets of identical twins the correlation between IQ of the first born twin and IQ of the second born twin was found to be 0.85. Mean IQs were found to be similar in the two groups, as were standard deviations. When researchers picked out those pairs where the first born twin had an IQ over 130 (about 2% of the sets of twins) they found that the second born twin had a somewhat lower IQ than the first born twin on average. The opposite happened when they took sets of twins where the first born had an IQ below 70. Is this to be expected? Explain. [2 marks for the explanation]

Appendix 1

R Code and output

```
> fit <-  lm(Velocity~Distance,data=d)

Call:
lm(formula = Velocity ~ Distance, data = d)

Coefficients:
(Intercept)     Distance
     -40.78       454.16

> summary(fit)

Call:
lm(formula = Velocity ~ Distance, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-397.96 -158.10  -13.16  148.09  506.63

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -40.78      83.44  -0.489     0.63
Distance      454.16      75.24   6.036 4.48e-06 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 232.9 on 22 degrees of freedom
Multiple R-Squared: 0.6235,Adjusted R-squared: 0.6064
F-statistic: 36.44 on 1 and 22 DF,  p-value: 4.477e-06
> anova(fit.lin)
Analysis of Variance Table

Response: Velocity
          Df  Sum Sq Mean Sq F value    Pr(>F)
Distance   1 1976648 1976648  36.438 4.477e-06 ***
Residuals 22 1193442   54247
```

SAS Code and output

```
data Hubble;
  infile 'Hubble.dat' firstobs=2;
  input Distance Velocity;
run;
proc glm;
  model Velocity = Distance;
run;
```

```
                                 The GLM Procedure

             Number of Observations Read          24
             Number of Observations Used          24

Dependent Variable: Velocity

                Sum of
Source   DF     Squares   Mean Square  F Value  Pr > F

Model     1  1976648.259  1976648.259   36.44   <.0001
Error    22  1193442.366    54247.380

Corr Tot 23  3170090.625

R-Square     Coeff Var      Root MSE   Velocity Mean
0.623531      62.42162      232.9107      373.1250


Source   DF   Type I SS   Mean Square  F Value  Pr > F
Distance  1  1976648.259  1976648.259   36.44   <.0001

                              Standard
Parameter     Estimate          Error    t Value  Pr > |t|

Intercept    -40.7836491   83.43886994  -0.49    0.6298
Distance     454.1584409   75.23710535   6.04    <.0001
```