DEPARTMENT OF STATISTICS

Course STATS 330: Advanced Statistical Modelling

Term Test: 9.00am - 10:00am, Tuesday Sept 15, 2005

INSTRUCTIONS

- Answer **ALL 15** questions on the answer sheet provided.
- All questions have a single correct answer and carry the same mark value.
- If you give more than one answer to any question you will receive zero marks for that question.
- Incorrect answers are not penalised.

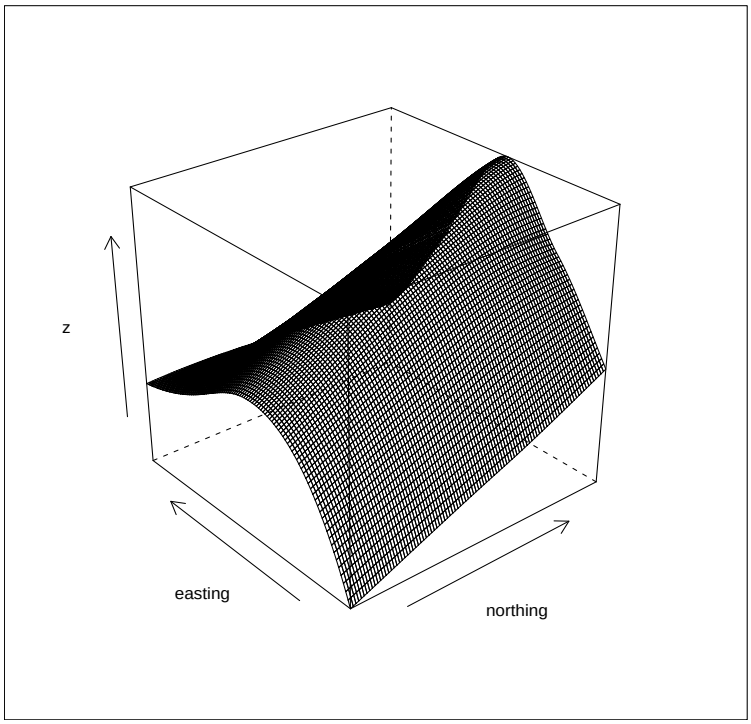

Figure 1: Wireframe plot for Question 1.

1. Figure 1 is a wireframe plot relating to a set of Australian data. The data consist of measurements of ground resistivity (a measure of the amount of salt in the soil) at a large number of locations. For each location, the "easting" and "northing" (the geographical cordinates) and the resistivity (denoted by z) are recorded. A surface was fitted and the wireframe plot represents the surface. Which of the following is **FALSE**?

   (zz) The resistivity is greatest when the value of easting is greatest.

   (1) The resistivity increases as the northing increases.

   (1) The surface could also have been represented as a coutour plot.

   (1) A regression model fitted to these data should contain polynomial terms.

   (1) The minimum resistivity occurs when the northing is least.

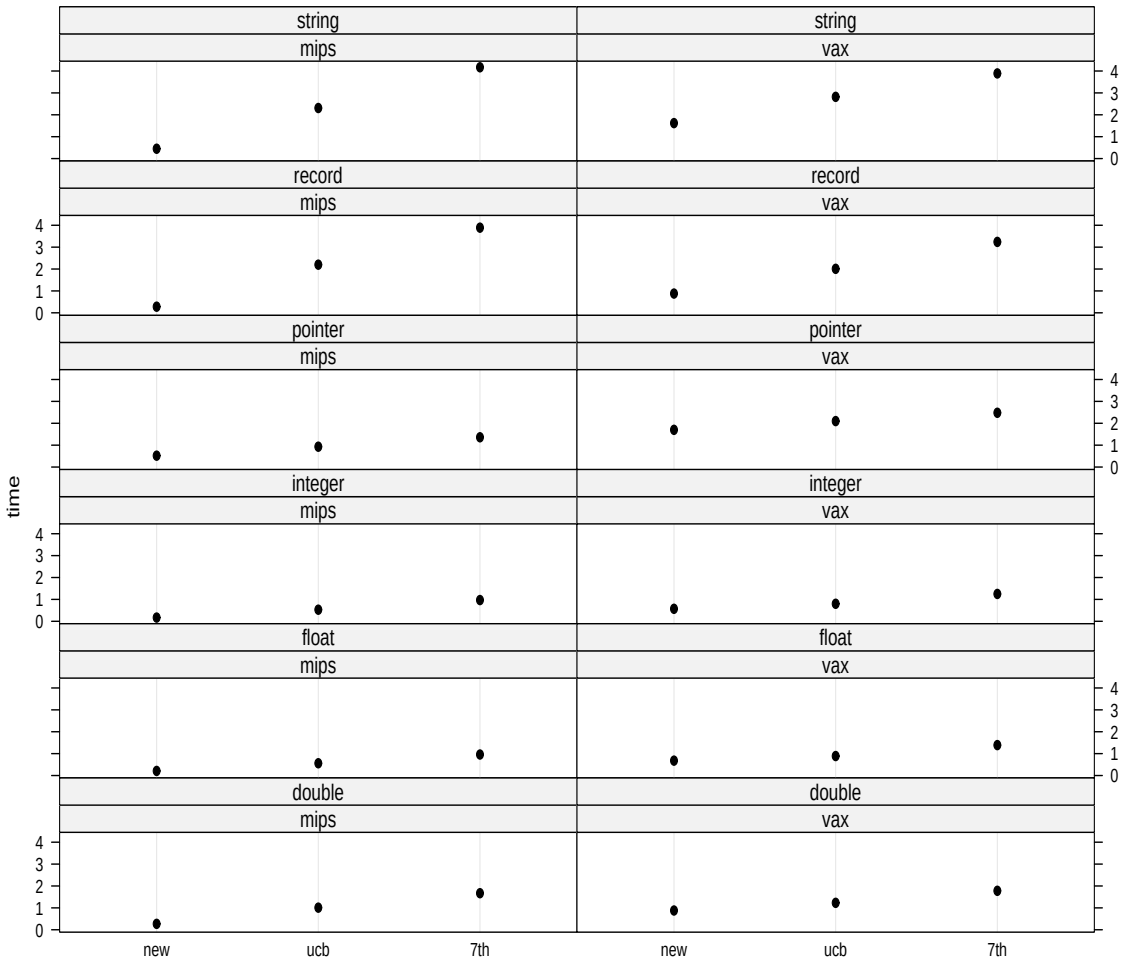

Figure 2: Trellis plot for Question 2.

2. Figure 2 shows data gathered in 1992 in an experiment to compare the speed of sorting algorithms used to sort different types of object on a computer. Three algorithms were compared, denoted here by "ucb", "7th" and "new". The algorithms were tried on two different computers, a VAX 8550 and a MIPS R3000. The objects to be sorted were of six types, namely integer, float, double, pointer, record and string. The response variable is time, the time taken to do the sort. Which of the following is **FALSE**?

   (zz) The "7th" method is the fastest.

   (1) The VAX computer seems slightly slower on the new method.

   (1) Sorting records and strings takes longer than the other types of object.

   (1) Using boxplots would not have been as effective as dotplots.

   (1) Conditioning on type of object and algorithm would have made the comparison of computers easier.

3. Which one of the following is **NOT** an assumption when fitting linear models using the R function `lm`?

   (zz) All the explanatory variables must be continuous.

   (1) The response variable must be continuous.

   (1) The mean response is a linear function of the explanatory variables.

   (1) The responses are independent.

   (1) The responses have constant variance.

4. The data for this question were collected as part of an ecological study in North Carolina. The investigators were interested in the relationship between the amount of a particular plant (Spartina) at particular sites, and the chemical properties of the soil at the site. The variables measured were

   - BIO: Biomass, a measure of the amount of Spartina at a site;
   - SAL: Salinity of the soil at the site;
   - pH: pH of the soil at the site;
   - K: Potassium concentration at the site;
   - Na: Sodium concentration at the site;
   - Zn: Zinc concentration at the site;
   - LOC: Location, one of Oak Island (OI), Smith Island (SI) or Snows Marsh (SM);
   - TYPE: Three types of Spartina vegetation (DVEG, SHRT, TALL).

   Initially, the categorical variables TYPE and LOC were ignored, and a regression with BIO as the response variable and the other (chemical) variables as explanatories was fitted, with the following results:

```
lm(formula = BIO ~ SAL + pH + K + Na + Zn, data = salinity.df)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.252e+03  1.235e+03   1.014  0.31674
SAL         -3.029e+01  2.403e+01  -1.260  0.21507
pH           3.055e+02  8.788e+01   3.477  0.00126 **
K           -2.851e-01  3.484e-01  -0.818  0.41817
Na          -8.673e-03  1.593e-02  -0.544  0.58927
Zn          -2.068e+01  1.505e+01  -1.373  0.17746

Residual standard error: 398.3 on 39 degrees of freedom
Multiple R-Squared: 0.6773,     Adjusted R-squared:  0.6359
F-statistic: 16.37 on 5 and 39 DF,  p-value: 1.082e-08
```

   Which is the **correct** interpretation?

(zz) Assuming the other variables are held constant, a unit increase in pH will cause the mean response to increase by about 305.5 biomass units.

(1) Since the regression coefficient of SAL is large, the variable should be retained in the regression.

(1) Because the regression coefficient of K is small, the variable should be deleted from the regression.

(1) Because the regression coefficient of K is small, the variable cannot be highly correlated with the response.

(1) Since the p-value associated with Na is large, the variable should be retained in the regression.

5. One investigator wanted to predict the amount of biomass at a particular site, whose pH is 4. He obtained the following R output:

```
> new.data.df<-data.frame(pH=4)
> pH.lm<- lm(BIO ~ pH , data=salinity.df)
> predict(pH.lm, new.data.df, interval="p")
          fit       lwr      upr
[1,] 754.0067 -109.9620 1617.975
> predict(pH.lm, new.data.df, interval="c")
         fit      lwr      upr
[1,] 754.0067 612.6067 895.4067
```

Which of the following is **TRUE**?

(zz) The biomass at the site is predicted to be between -109.9620 and 1617.975.

(1) The biomass at the site is predicted to be between 612.6067 and 895.4067.

(1) The prediction is invalid since the other variables are not used.

(1) The mean biomass of all sites having ph of 4.0 is estimated to be between -109.9620 and 1617.975.

(1) The standard error of the prediction is about 863.

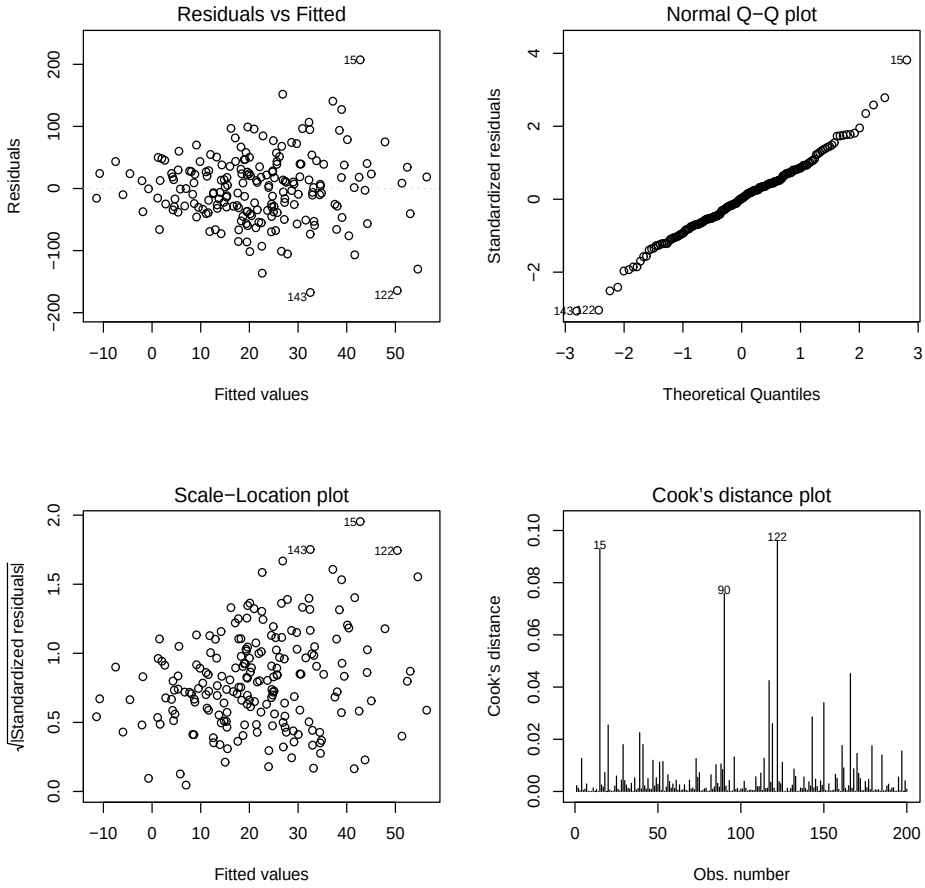

Figure 3: Diagnostic plots for Question 6.

6. Figure 3 shows some diagnostic plots obtained after fitting a regression. What, if anything, is the most important thing wrong with the regression?

   (zz) The error variances are not constant.

   (1) There are outliers in the data.

   (1) The errors are not normally distributed.

   (1) The points are not scattered about a plane.

   (1) The plots do not indicate problems with the regression.

7. For the data in Question 6, what remedial action should we take?

   (zz) We should use weighted least squares or transform the response.

   (1) We should delete the outliers.

   (1) We should not use linear regression since the data are not normal.

   (1) We should transform one or more of the explanatory variables.

   (1) We need do nothing, there are no problems with the regression.

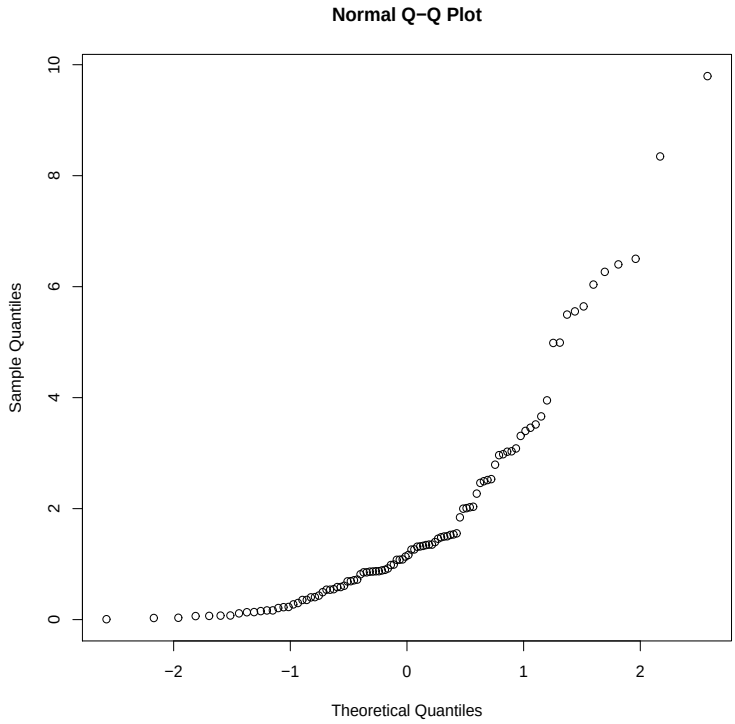

Figure 4: Normal plot for Question 8.

8. In a regression, we get the normal plot of residuals shown in Figure 6. Which of the following is **TRUE**?

   (zz) The errors seem to come from a right-skew distribution.

   (1) The errors appear normal.

   (1) The errors appear to come from a symmetric, short-tailed distribution.

   (1) The errors seem to not be independent.

   (1) The errors appear to come from a symmetric, long-tailed distribution.

9. In the biomass data described in Question 4, there were 45 observations in the data set. The (edited) influence measures display shown on the next page was obtained:

	dfb.1_	dfb.SAL	dfb.pH	dfb.K	dfb.Na	dfb.Zn	dffit	cov.r	cook.d	hat
5	6.54e-02	-0.030282	-0.107683	0.245103	-0.243854	-0.082805	-0.317801	3.318	1.73e-02	0.6507
7	-1.94e-02	0.021816	0.008607	0.025779	-0.020515	0.012555	0.031693	1.642	1.72e-04	0.2886
27	-1.10e-02	0.011965	0.012765	0.004795	-0.010075	0.005472	0.022303	1.543	8.51e-05	0.2429
28	8.97e-02	-0.094164	-0.117628	0.010580	0.037420	-0.041921	-0.200808	1.468	6.87e-03	0.2216
34	5.84e-01	-0.752085	-0.309374	-0.183422	0.532588	-0.406198	1.031879	0.325	1.45e-01	0.0975
40	-5.93e-02	0.065070	0.047404	-0.042510	0.018354	0.033227	-0.126865	1.188	2.73e-03	0.0570

One of the following statements is **TRUE**. Which one?

(zz) Observation 34 is influential.

(1) Observation 27 will have a big effect on the coefficient for SAL.

(1) Observation 27 will have a big effect on the coefficient for Zn.

(1) Observation 40 will have a big effect on the estimated standard errors.

(1) Observation 5 is not influential.

Hint for question 9: Points are influential if (i) Cook's D is more than $F_{6,39}(0.1) = 0.3593724$, (ii) $|DFBETAS| > 2/\sqrt{n} = 0.298$, (iii) $|DFFITS| > 2\sqrt{p/(n-p)} = 0.784$, (iv) $|COVRATIO - 1| > 3p/n = 0.4$.

10. Suppose we include the variable TYPE in the model for biomass. This produces the following output:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  731.82896 1072.63355   0.682  0.49932
TYPESHRT    -409.40739  133.18408  -3.074  0.00395 **
TYPETALL     319.60045  180.04276   1.775  0.08410 .
SAL          -20.74988   20.36332  -1.019  0.31483
pH           240.57471   75.50694   3.186  0.00293 **
K              0.23259    0.31912   0.729  0.47070
Na            -0.01280    0.01445  -0.885  0.38165
Zn            -8.57905   13.49883  -0.636  0.52899
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 335.1 on 37 degrees of freedom
Multiple R-Squared: 0.7833,     Adjusted R-squared: 0.7422
F-statistic:  19.1 on 7 and 37 DF,  p-value: 1.619e-10
```

Recall that the factor TYPE had 3 levels, (DVEG, SHRT, TALL). Assuming the fitted model is correct, choose the correct interpretation from the alternatives on the next page.

(zz) Sites with the SHRT type have about 409 biomass units less biomass than the DVEG type.

(1) This model is fitting three non-parallel plans through the data.

(1) This model is fitting two non-parallel plans through the data.

(1) Sites with the SHRT type have about 409 biomass units more than the TALL type

(1) The average biomass for TALL types is about 732 biomasss units.

11. Consider the biomass data described in Question 4. We want to choose a model for these data using the variables TYPE, SAL, pH, K, Na and Zn. Some R code and output shown below.

```
> salinity.lm<-lm(formula = BIO ~ TYPE + SAL + pH + K + Na + Zn, data = salinity.df)
> all.poss.regs(salinity.lm)
      rssp   sigma2 adjRsq     Cp    AIC    BIC       CV TYPESHRT TYPETALL SAL pH K Na Zn
1 7680575 178618.0  0.590 27.391 72.391 76.005 705962.0        0        0   0  1 0  0  0
2 5440725 129541.1  0.703  9.447 54.447 59.867 438706.5        0        1   0  1 0  0  0
3 4498004 109707.4  0.748  3.052 48.052 55.279 391483.9        1        1   0  1 0  0  0
4 4325492 108137.3  0.752  3.516 48.516 57.549 422999.2        1        1   1  1 0  0  0
5 4244246 108826.8  0.750  4.793 49.793 60.633 429261.3        1        1   1  1 0  0  1
6 4200581 110541.6  0.746  6.404 51.404 64.051 450891.6        1        1   1  1 1  1  0
7 4155220 112303.3  0.742  8.000 53.000 67.453 469464.5        1        1   1  1 1  1  1
```

Which statement below is **TRUE**?

(zz) The model with TYPE and pH is a good model.

(1) The full model is the best model since it has the smallest residual sum of squares.

(1) The model with TYPE and pH and SAL is a poor model since it has the biggest adjusted R^2.

(1) The model with pH alone is the best model since it has the smallest number of variables.

(1) None of these models fits well as the biggest adjusted R^2 is only 75.2%.

12. In class, we studied various criteria for model selection. Which of the following statements relating to these criteria is **FALSE**?

(zz) We can use R^2 as a criterion: the bigger the R^2, the better the model.

(1) The BIC criterion usually selects simpler models than the AIC criterion.

(1) Ten-fold cross-validation is a better method than n-fold (leave one out) cross-validation.

(1) The adjusted R^2 criterion is equivalent to using the model with the smallest estimate of error variance.

(1) The Cp criterion is based on selecting a model that predicts well.

13. Suppose in the biomass data we ignored the chemical variables and just used the variables TYPE and LOC. The table of means is given below:

```
         OI     SI     SM
DVEG  824.0  340.8 1564.8
SHRT  558.4  297.6  492.0
TALL 1280.0 2136.0 1513.6
```

Which of the following is **FALSE**? Assume that we are using the "baseline" definition of effects, and the factor levels are in alphabetical order.

(zz) All the interactions are zero.

(1) The row effect for $TYPE = SHRT$ is -265.6.

(1) The column effect for $LOC = SM$ is 740.8.

(1) The interaction for the baseline cell is zero.

(1) The overall baseline is 824.0

14. Suppose in the biomass data we now want to use all the variables. In the code below, we compare two possible models. What is the hypothesis being tested by the p-value 0.9433?

```
> model1<-lm(BIO~ TYPE*LOC + SAL  +  pH   +    K  +    Na   +   Zn,
                  data=salinity.df)
> model2<-lm(BIO~ TYPE*LOC, data=salinity.df)
> anova(model2,model1)
Analysis of Variance Table

Model 1: BIO ~ TYPE * LOC
Model 2: BIO ~ TYPE * LOC + SAL + pH + K + Na + Zn
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     36 2352787
2     31 2266304  5     86483 0.2366 0.9433
```

(zz) That the chemical variables (SAL, pH, K, Na and Zn) are unnecessary, provided the factors TYPE and LOC and their interactions are in the model.

(1) That the chemical variables (SAL, pH, K, Na and Zn) are unnecessary, provided the factors TYPE and LOC are in the model.

(1) That TYPE and LOC are not required in the model.

(1) That at least some of the variables are required in the model.

(1) That no interactions between TYPE and LOC are required in the model.

15. Suppose we want to use these data to develop a prediction formula for predicting the biomass at other unspecified locations. Which of the models listed below do you think would be most useful for this purpose? Some output from the anova function is shown to help you decide.

CONTINUED

```
> model3<-lm(BIO~ TYPE +  pH   +   SAL +  K  +    Na
                 +   Zn + LOC + TYPE:LOC, data=salinity.df)
> anova(model3)
Analysis of Variance Table

Response: BIO
          Df   Sum Sq  Mean Sq F value    Pr(>F)
TYPE       2 10875932  5437966 74.3841 1.473e-12 ***
pH         1  3797027  3797027 51.9382 4.190e-08 ***
SAL        1   172512   172512  2.3597  0.134648
K          1     3349     3349  0.0458  0.831918
Na         1   121562   121562  1.6628  0.206763
Zn         1    45361    45361  0.6205  0.436854
LOC        2   549311   274656  3.7569  0.034598 *
TYPE:LOC   4  1339605   334901  4.5810  0.005053 **
Residuals 31  2266304    73107
---
```

(zz) BIO $\sim$ TYPE + pH;

(1) BIO $\sim$ TYPE + pH + SAL + K + Na + Zn + LOC + TYPE:LOC;

(1) BIO $\sim$ pH + SAL + K + Na + Zn;

(1) BIO $\sim$ TYPE + LOC + TYPE:LOC;

(1) BIO $\sim$ pH.