DEPARTMENT OF STATISTICS

Course STATS 330: Advanced Statistical Modelling

Term Test 9.00 - 10:00 Tuesday Sept 21, 2004

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.

1. Suppose that we have a data set consisting of a continuous response variable Y, a continuous explanatory variable X, and a categorical explanatory variable Z, each observed on 60 individuals. The variable Z has 3 levels. Which one of the following plots would be **most suitable** for portraying the relationships between the variables?

   (1) A trellis plot with 3 panels corresponding to the levels of Z, with each panel containing a scatterplot of Y versus X.

   (2) A scatterplot of X versus Z, with the value of Y shown by a colour coding.

   (3) A trellis plot with 3 panels corresponding to the levels of X, with each panel containing two boxplots.

   (4) A trellis plot consisting of 60 two-way tables.

   (5) A barchart of Y, with colour coding indicating the levels of X and Z.

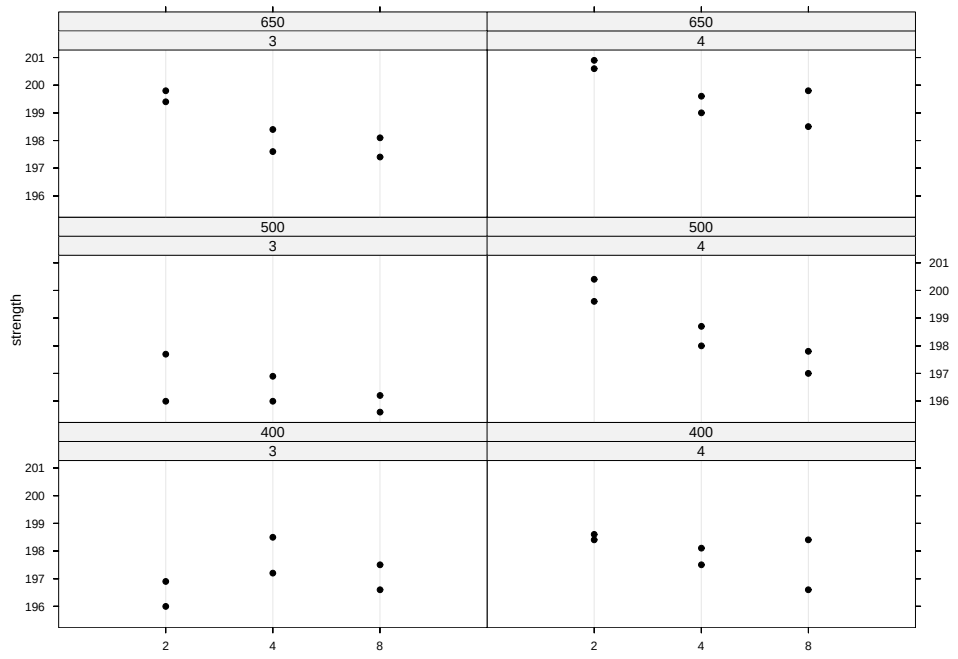

Figure 1: Trellis plot for Question 2.

2. A industrial experiment has been conducted to examine the effect of various factors on the strength of paper. The response variable is the paper strength, and the three factors considered were

   - Percent: percent of hardwood concentration in the mix, either 2%, 4% or 8%;
   - Time: Cooking time, either 3 hours or 4 hours;
   - Pressure: either 400psi, 500 psi or 650 psi.

   The trellis plot in Figure 1 displays the relationship between Strength and Percent for different Time/Pressure combinations. Which of the following statements is **FALSE**?

   (1) There are two mesurements of strength taken for each percent/time/pressure combination.

   (2) The responses for percent=2, time=3 and pressure =400 seem inconsistent with the general pattern in the rest of the data.

   (3) As the percent of hardwood goes up, the strength decreases.

   (4) Increasing the cooking time tends to increase the strength.

   (5) The cooking time seems to have no effect on strength.

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

   (1) If the residual sum of squares is one, the points all lie on a plane.

   (2) To select a model, we pick the model with the smallest residual sum of squares.

   (3) If the regression coefficients (except the constant term) are all zero, the R^2 must be zero.

   (4) The adjusted R^2 is always bigger then the R^2.

   (5) Adding an extra variable to a regression model always increases the residual sum of squares.

4. In a regression, the variable X has a regression coefficient of 3. Which one of the following statements is **TRUE**?

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

   (2) Since the p-value associated with a coefficient of 3 is small, the variable X should be retained in the regression.

   (3) Assuming the other variables are held constant, a unit change in X will cause the mean response to increase by 3.

   (4) Assuming the other variables are held constant, a unit change in X will cause the response to increase by 3.

   (5) Since the regression coefficient is large, the variable is highly correlated with the response.

5. In the petrol vapour data set discussed in class, the response variable was `hc`, the amount of hydrocarbons released into the atmosphere, and the explanatory variables were `t.temp` (the tank temperature), `p.temp` (the petrol temperature), `t.vp` (the tank vapour pressure), and `p.vp` (the petrol vapour pressure). A regression was fitted using these four explanatory variables, and the results are shown below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.16609    1.02198   0.163  0.87117
t.temp      -0.07764    0.04801  -1.617  0.10850
p.temp       0.18317    0.04063   4.508 1.53e-05 ***
t.vp        -4.45230    1.56614  -2.843  0.00526 **
p.vp        10.27271    1.60882   6.385 3.37e-09 ***

Residual standard error: 2.723 on 120 degrees of freedom
Multiple R-Squared: 0.8959,     Adjusted R-squared: 0.8925
F-statistic: 258.2 on 4 and 120 DF,  p-value: < 2.2e-16
```

   Which one of the following statements is **FALSE**? Use the above output only.

(1) The variable `t.temp` has no relationship with the amount of hydrocarbon emitted.

(2) The variable `t.temp` can be dropped from the regression.

(3) If the other variables are held constant, increasing the petrol temperature tends to increase the amount of hydrocarbon emitted.

(4) If the other variables are held constant, increasing the petrol vapour pressure tends to increase the amount of hydrocarbon emitted.

(5) If the other variables are held constant, increasing the tank vapour pressure tends to decrease the amount of hydrocarbon emitted.

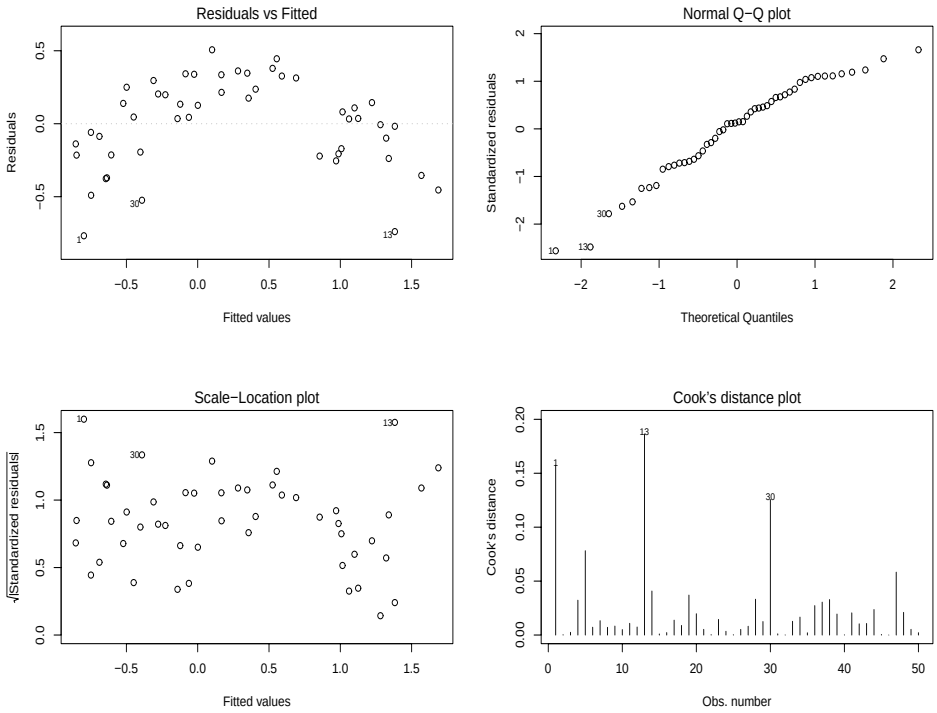

Figure 2: Diagnostic plots for Question 6.

6. Figure 2 shows some diagnostic plots obtained after fitting a regression. What, if anything, is the most important thing wrong with the regression?

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

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

(3) The errors are not normally distributed.

(4) The error variances are not constant.

(5) There are outliers in the data.

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

   (1) We should use weighted least squares.

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

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

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

   (5) We should delete the outliers.

8. After fitting a regression, which has response y and explanatory variables x_1 and x_2, we want to predict the value of y for the values $x_1 = 0.4$, $x_2 = 0.07$. We get the the following R output:

```
> fitted.model<-lm(y~x1+x2)
> new.x<-data.frame(x1=0.4, x2=0.7)
> predict(fitted.model, new.x, se=T)
$fit
[1] -0.3971443
$se.fit
[1] 0.0699711
$df
[1] 47
$residual.scale
[1] 0.3585052
> qt(0.025,47)
[1] -2.011741
```

   Which of the following is **TRUE**?

   (1) A 95% prediction interval for y is $(-0.5379, -0.2563)$.

   (2) A 95% confidence interval for the mean of y is $(-1.1319, 0.3376)$.

   (3) The R^2 is 35%.

   (4) A 95% confidence interval for the mean of y is $(-0.5379, -0.2564)$.

   (5) The estimate of σ (to 4 decimal places) is 0.1285.

9. Consider the petrol vapour data described in Question 5. We want to choose a model for these data using the R code and output shown overleaf.

```
>  vapour.lm<-lm(hc ~ p.temp + t.temp + p.vp + t.vp, data=vapour.df)
> all.poss.regs(vapour.lm)
      rssp sigma2 adjRsq      Cp     AIC     BIC p.temp t.temp p.vp t.vp
1 1513.347 12.304  0.822  83.067 208.067 213.723      0      0    1    0
1 1962.496 15.955  0.769 143.632 268.632 274.289      1      0    0    0
1 2364.791 19.226  0.721 197.879 322.879 328.536      0      0    0    1
2 1044.833  8.564  0.876  21.890 146.890 155.375      0      0    1    1
2 1089.397  8.929  0.871  27.899 152.899 161.384      1      0    1    0
2 1320.953 10.827  0.843  59.123 184.123 192.608      1      0    0    1
3  909.304  7.515  0.891   5.615 130.615 141.928      1      0    1    1
3  949.847  7.850  0.886  11.082 136.082 147.395      1      1    1    0
3 1040.640  8.600  0.875  23.325 148.325 159.638      0      1    1    1
4  889.913  7.416  0.892   5.000 130.000 144.142      1      1    1    1
```

Which model is most strongly indicated by this output?

(1) `hc ~ p.temp + t.temp + p.vp + t.vp`

(2) `hc ~ t.vp`

(3) `hc ~ p.vp`

(4) `hc ~ t.temp + p.vp + t.vp`

(5) `hc ~ p.temp + p.vp + t.vp`

10. Three different machines (A, B and C) produce monofilament fibre for a textile company. The process engineer in the company is interested in whether or not the strength of the filaments produced by the three machines are different. It is known that the strength will also depend on the diameter of the filaments. Five filaments from each machine are selected and the strength and diameter of each filiament are measured. The following output was obtained:

```
> summary(lm(strength~diameter+machine,data=fibre.df))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.360      2.961   5.862 0.000109 ***
diameter       0.954      0.114   8.365 4.26e-06 ***
machineB       1.037      1.013   1.024 0.328012
machineC      -1.584      1.107  -1.431 0.180292

Residual standard error: 1.595 on 11 degrees of freedom
Multiple R-Squared: 0.9192,     Adjusted R-squared: 0.8972
F-statistic: 41.72 on 3 and 11 DF,  p-value: 2.665e-06
```

Assuming the fitted model is correct, choose the correct interpretation from the alternatives overleaf.

(1) For a given diameter, the mean strength for machine C is 1.584 units more than for machine A.

(2) For machine B only, increasing the diameter by one unit increases the mean strength by 1.037.

(3) For all machines, increasing the diameter by one unit increases the mean strength by 0.954.

(4) There is clear evidence that the machines produce fibres of different strengths.

(5) For machine C, increasing the diameter reduces the mean strength.

11. In the vapour pressure data described in Question 5, there were 125 observations in the data set. The following (edited) influence measures display was obtained:

```
Influence measures of lm(formula = hc ~ ., data = vapour.df) :
        dfb.1  dfb.t.tm  dfb.p.tm  dfb.t.vp  dfb.p.vp    dffit cov.r   cook.d
1   -2.96e-01  0.063837  5.03e-01  0.461704 -6.12e-01 -0.76568 1.069 1.15e-01
58   1.88e-02  0.030141 -1.10e-02 -0.067463  5.46e-02 -0.09081 1.252 1.66e-03
59  -6.12e-05  0.010975 -3.77e-02 -0.023039  3.13e-02  0.04480 1.367 4.05e-04
61  -5.07e-01 -0.258243  1.03e-01  0.055914  9.77e-02  0.69341 0.894 9.25e-02
86   4.69e-02  0.232445 -7.53e-03 -0.197164  1.05e-01  0.38703 0.797 2.85e-02
```

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

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

(2) Observation 61 will have a big effect on the coefficient for tank vapour pressure.

(3) Observation 86 is influential.

(4) Observation 58 will have a big effect on the coefficient for tank vapour pressure.

(5) Observation 59 is not influential.

Hint for question 9: Points are influential if (i) Cook's D is more than $F_{5,120}(0.1) = 1.895875$, (ii) $|DFBETAS| > 2/\sqrt{n}$, (iii) $|DFFITS| > 2\sqrt{p/(n-p)}$, (iv) $|COVRATIO - 1| > 3p/n$.

12. In a regression having two explanatory variables and 50 observations taken sequentially in time, it was suspected that some serial correlation in the errors might be present. The following output was obtained, together with the graphs shown in Figure 3. Which of the following conclusions are indicated?

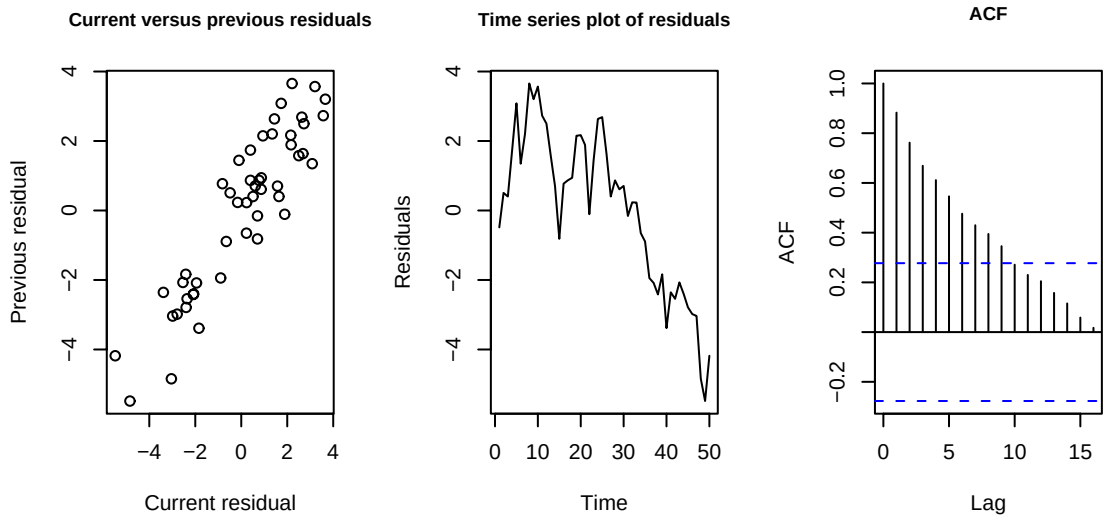

Figure 3: Diagnostic plots for Question 12.

(1) There is weak positive serial correlation.

(2) There is strong negative serial correlation.

(3) There is no serial correlation.

(4) There is strong positive serial correlation.

(5) There is weak negative serial correlation.

13. In class, we studied remedies for regression when the variances of the errors are not equal. Which of the following statements relating to these remedies is **FALSE**?

(1) If we take no action, the estimated standard errors of the regression coefficients will be incorrect.

(2) If the variances are functions of the mean, we can detect this by examining a plot of residuals versus fitted values.

(3) We could use weighted least squares with weights inversely proportional to the variances.

(4) If the variances are functions of the mean, we could try transforming the response.

(5) We should estimate the variances by smoothing the plot of residuals versus fitted values.

14. Suppose we have a regression with two explanatory variables A and B, both of which are categorical. A has levels High (baseline) and Low, while B has levels Diet 1 (baseline) and Diet 2. The population means of the four possible treatment combinations are

	$B = Diet\ 1$	$B = Diet\ 2$
$A = High$	10	15
$A = Low$	7	16

Which of the following is **FALSE**?

(1) The column effect for $B = Diet$ 2 is 5.

(2) All the interactions are zero.

(3) The overall baseline is 10.

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

(5) The row effect for $A = Low$ is -3.

15. Suppose we have a single data set coinsisting of a response y and two categorical explanatory factors A and B. Consider the null models `y ~1` (Model 1) and `y ~ A` (Model 2). In the output below, what is the hypothesis being tested by the p-value 0.03487 ?

```
> model1<-lm(y~1)
> model2<-lm(y~A)
> anova(model1,model2)
Analysis of Variance Table
Model 1: y ~ 1
Model 2: y ~ A
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     24 27.4438
2     23 22.5203  1    4.9236 5.0285 0.03487 *
```

(1) Model 2 is a satisfactory model.

(2) The group variances are unequal.

(3) Factor B is required in the model.

(4) The factor A has no effect on the response.

(5) The model assumptions are satisfied.