DEPARTMENT OF STATISTICS

Course STATS 330/762: Advanced Statistical Modelling/Special Topic in Regression

Term Test: 8.00am - 9:00am, Monday August 20, 2007

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.
- A correct answer is worth one point, an incorrect answer zero points.

1. A data set consists of a continuous variable Y, and two explanatory variables X and Z (both of which are continuous). We want to see if the relationship between X and Y changes with the different values of Z. Which of the following pieces of R code would produce the most informative plot?

   (zz) `xyplot(Y~X|equal.count(Z))`

   (1) `xyplot(Y~X|Z)`

   (1) `dotplot(Y~Z|X)`

   (1) `pairs(data.frame(X,Y,A))`

   (1) `plot(lm(Y~X+Z))`

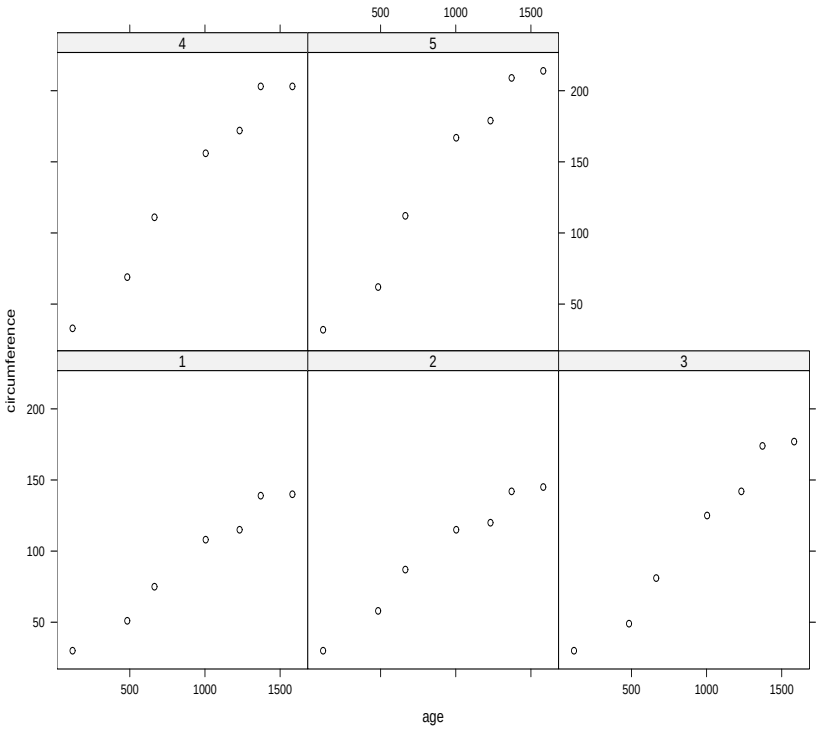

Figure 1: Trellis plot for Question 2.

2. Figure 1 shows a trellis plot of measurements made on five trees at different times. The data set contains the following variables:

   tree: A label for the tree (1-5),

   age: a numeric vector giving the age of the tree (days since 31/12/1968) at the time of the measurement,

   circumference: the trunk circumference at the time of the measurement (mm). This is "circumference at breast height", a standard measurement in forestry.

Which of the following is **FALSE**?

(zz) One of the other alternatives to this question is false.

(1) Tree 5 seems to have the biggest circumference.

(1) The rate of growth is smaller for Tree 1 than for Tree 5.

(1) The trees are labelled 1-5 in increasing order of circumference.

(1) The trees seem to be the same size at the end of 1968.

3. In the linear regression model, which is the most important assumption?

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

(1) The variances are equal.

(1) The observations are independent.

(1) The responses are normally distributed.

(1) There are no high leverage points.

4. Which of the following is **FALSE**? In a regression with two explanatory variables X and W,

(zz) The variance inflation factor for variable X can be calculated from the values of X alone.

(1) The variance inflation factor for X depends only on the correlation between X and W.

(1) The variance of the regression coefficient for X depends on the variance of X.

(1) The variance of the regression coefficient for X depends on error variance.

(1) The variance of the regression coefficient for X depends the sample size.

5. Which of the following is **FALSE**?

(zz) In a regression, adding a variable can decrease the R^2.

(1) In a regression, an R^2 of 1 means that all the points lie on a plane.

(1) In a regression, an R^2 of 0 means that all the regression coefficients except the constant term are zero.

(1) In a regression, an R^2 of 0.9 does not mean the points are planar.

(1) In a regression, R^2 is the square of the correlation between the observations and the fitted values.

6. In a regression, the variable X has a regression coefficient of -2. Which one of the following statements is **TRUE**?

   (zz) Assuming the other variables are held constant, a unit increase in X will cause the mean response to decrease by 2.

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

   (1) A unit increase in X will cause the response to increase by 2.

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

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

Questions 7-11 are concerned with air pollution data gathered daily at New York from May 1, 1973 to September 30, 1973. There are 153 days of data. The variables measured are

Ozone Ozone concentration(ppb)

Solar.R Solar Radiation (lang)

Wind Wind speed (mph)

Temp Temperature (degrees F)

7. A coplot of the variables is shown in Figure 2. Which is the **wrong** interpretation?

   (zz) Ozone goes down as solar radiation increases.

   (1) Ozone goes up as temperature goes up.

   (1) Ozone goes down as wind increases.

   (1) Ozone is highest for high temperatures, high solar radiation and low wind speeds.

   (1) Ozone ranges from roughly 0 to 150 ppb.

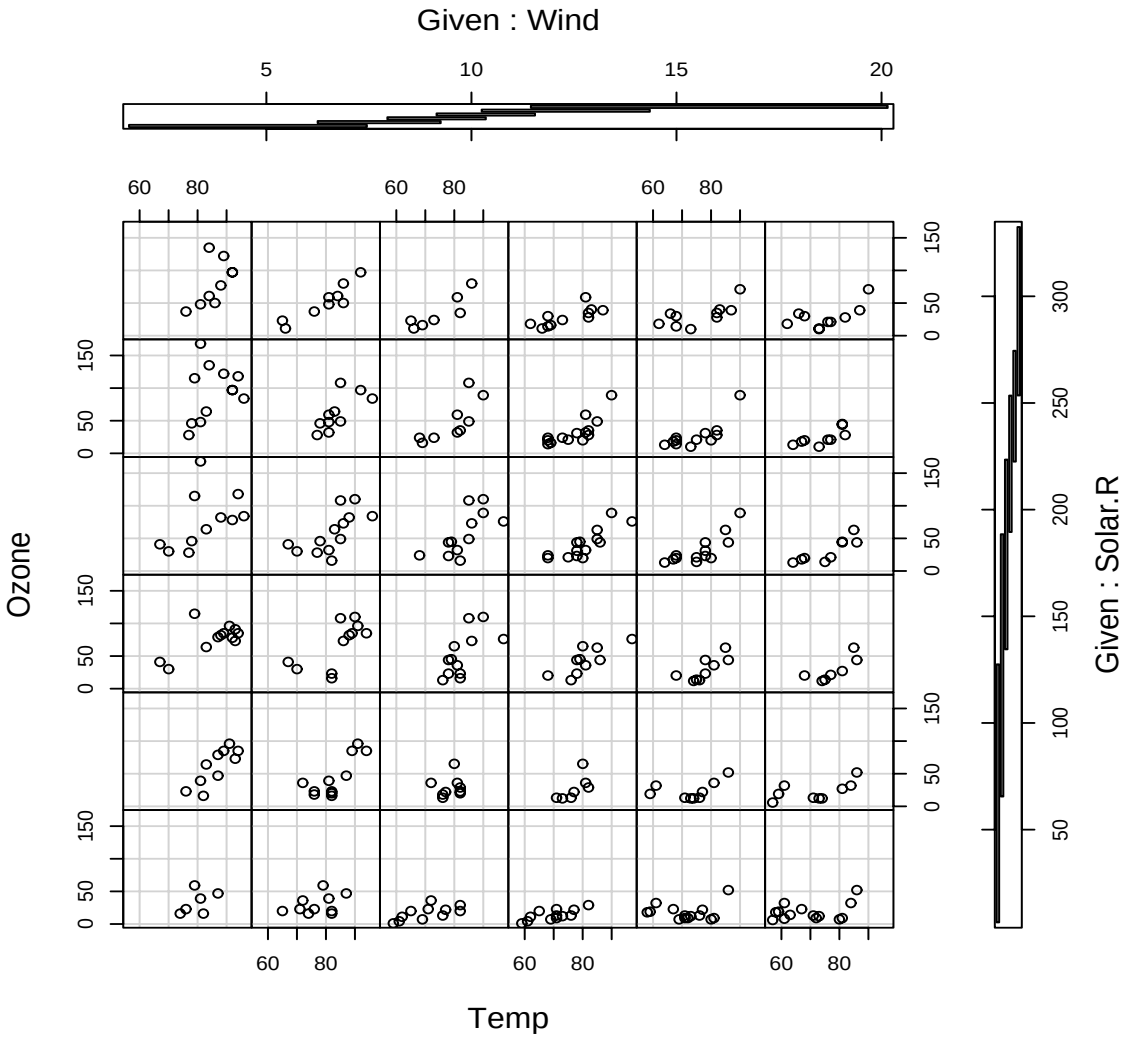

Figure 2: Coplot for Question 7.

8. A regression using `Ozone` as the response and the all the other variables as explanatories was fitted. Some diagnostic plots are shown in Figure 3.

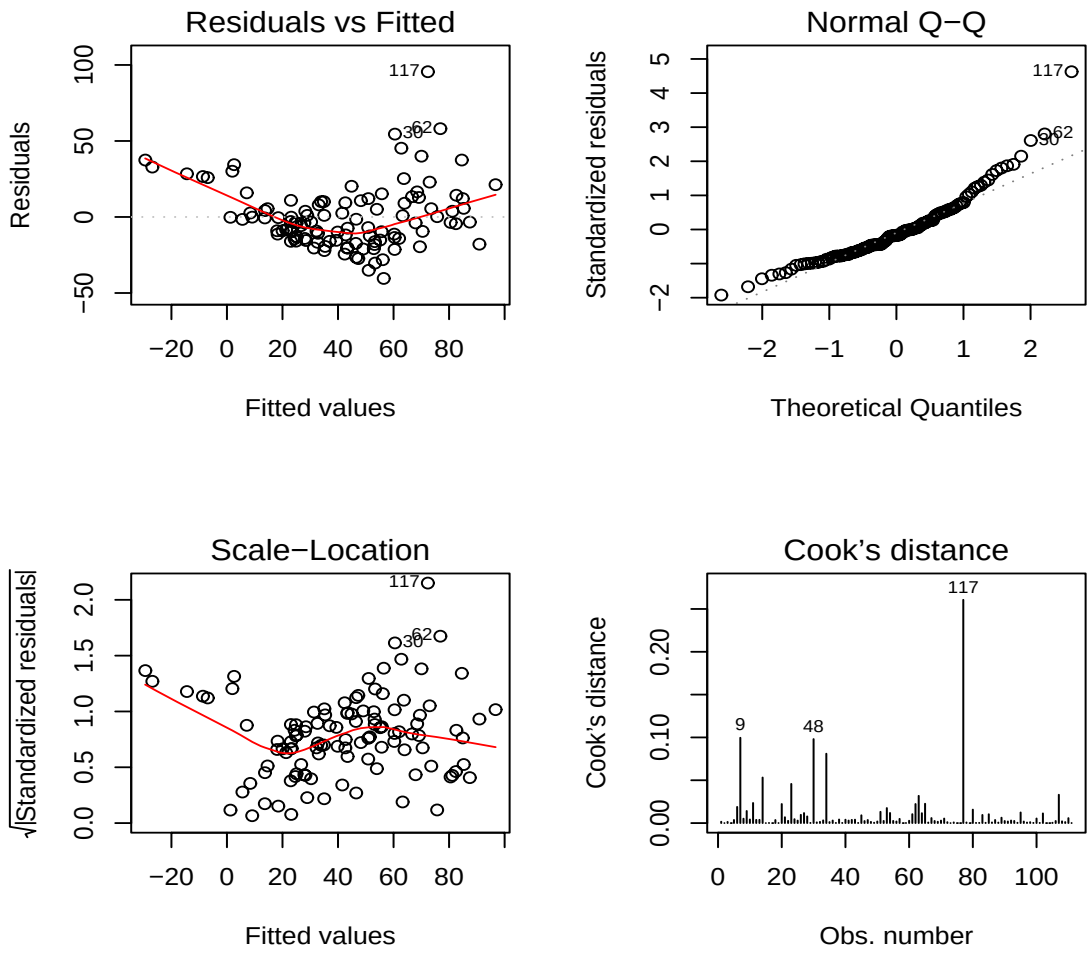

Figure 3: Diagnostic plots for Question 8.

Which of the following is **NOT** indicated by Figure 3?

(zz) There seems to be serial correlation in these data.

(1) The errors seem to be slightly right-skewed.

(1) There are outliers in the data.

(1) The variance of the errors seems to be increasing with the mean.

(1) The response is not a linear function of the expalanatory variables.

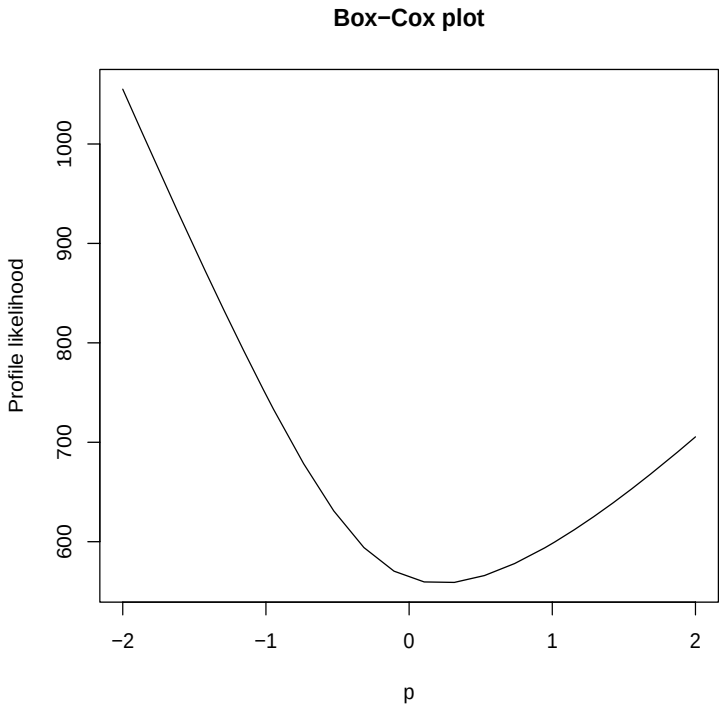

Figure 4: Box-Cox plot for Question 9.

9. A Box-Cox plot is shown in Figure 4. Which is the **BEST** interpretation of this plot?

   (zz) The response `Ozone` should be transformed with a log.

   (1) No transformation of the response is indicated.

   (1) A transformation of `Wind` and `Temp` is indicated.

   (1) The response `Ozone` should be transformed with a reciprocal transformation.

   (1) The relationship between the response and the explanatory variables is quadratic.

10. After taking some corrective action, the following summary was obtained:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2614362  0.5204961   0.502    0.617
Solar.R      0.0021904  0.0005156   4.248 4.65e-05 ***
Wind        -0.0692829  0.0145135  -4.774 5.82e-06 ***
Temp         0.0444569  0.0056785   7.829 3.95e-12 ***
---
Residual standard error: 0.4666 on 106 degrees of freedom
Multiple R-Squared: 0.6736,     Adjusted R-squared: 0.6643
F-statistic: 72.91 on 3 and 106 DF,  p-value: < 2.2e-16
```

Which of the following is a valid interpretation of this summary?

(zz) If the other variables are held constant, the mean response will increase by about .04 units with each 1 degree increase in temperature.

(1) The mean response goes down as Wind goes down.

(1) The mean response goes down as solar radiation goes up.

(1) The regression model assumptions are violated - the R^2 is only 67%.

(1) The variable `Wind` could be removed from the model.

11. Some influence plots for the original model fitted in Question 8 are shown in Figure 5. Which of the following statements is **NOT** a correct interpretation of these plots?

(zz) Observation 77 has high leverage.

(1) Observation 77 is having an effect on most of the regression coefficients.

(1) Observation 77 is having an effect on the fitted value for point 77.

(1) Observation 77 is having an effect on the standard errors.

(1) Of all the regression coefficients, the coefficient for solar radiation is least affected by outliers.

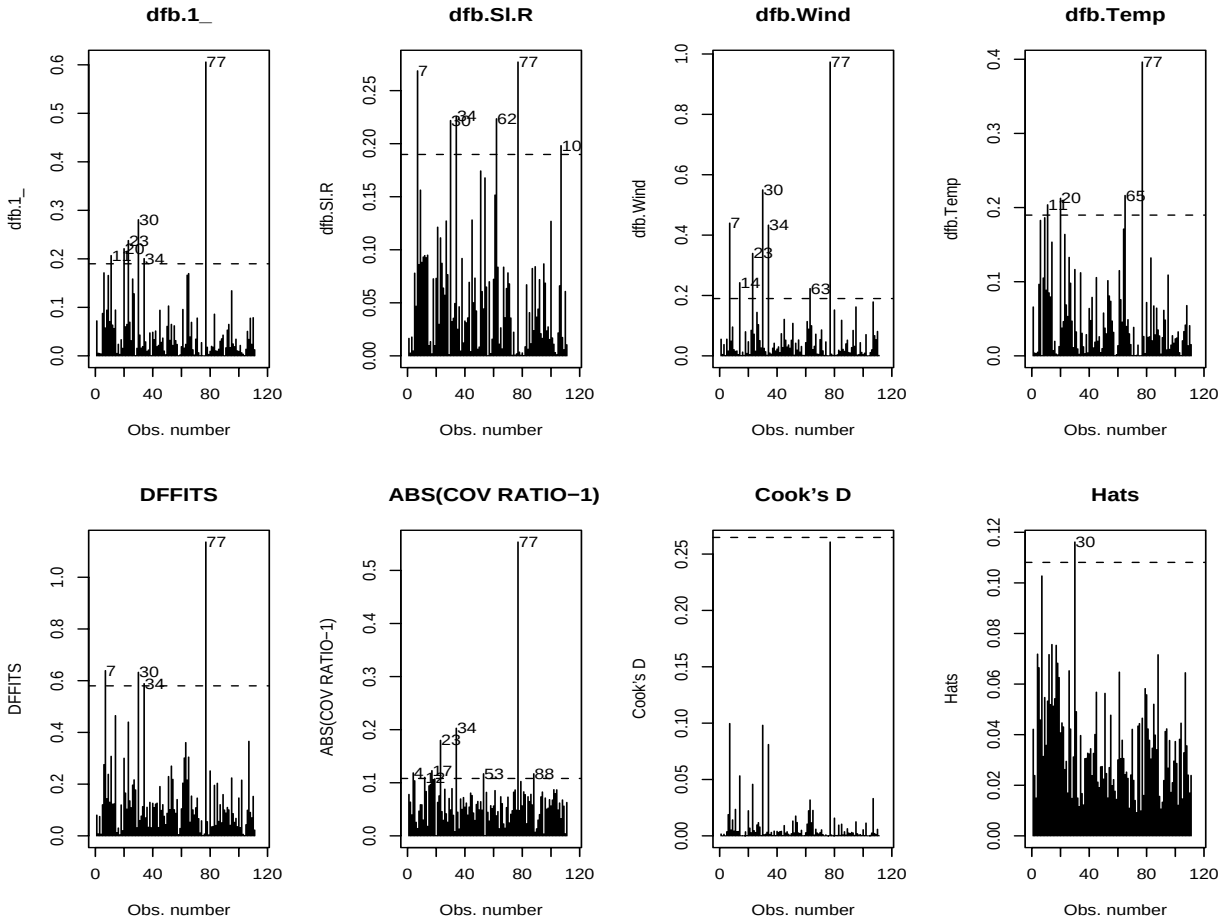

Figure 5: Influence plots for Question 11.

12. In a regression with data collected sequentially in time, we suspect serial correlation is present. We calculate a Durbin-Watson test statistic of 1.3. The values of d_l and d_u are 1.44 and 1.54 respectively. What do we conclude?

(zz) There is evidence of positive serial correlation.

(1) There is nothing to indicate the data are serially correlated.

(1) The test is inconclusive, but there is certainly no evidence of negative serial correlation.

(1) The test is inconclusive, but there is certainly no evidence of positive serial correlation.

(1) There is evidence of negative serial correlation.

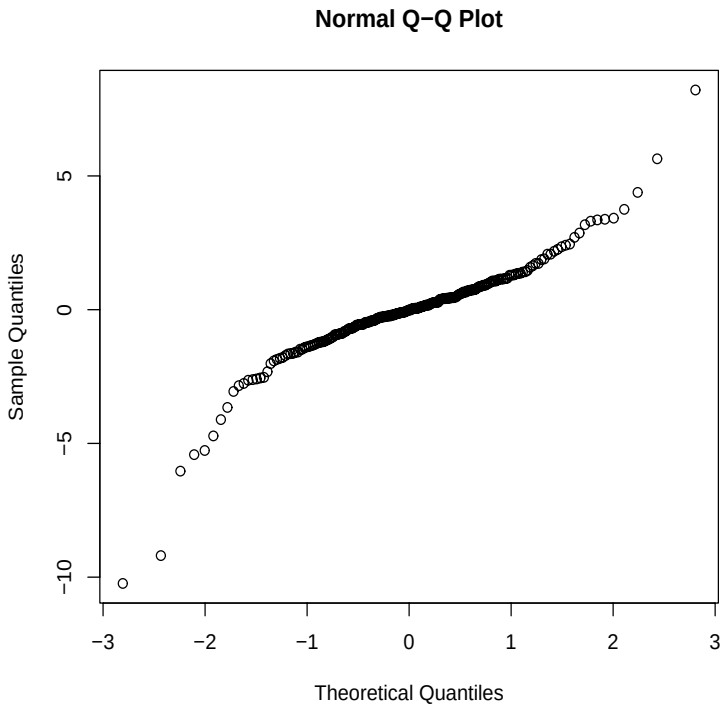

Figure 6: Normal plot for Question 13.

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

(zz) The errors appear to come from a symmetric, long-tailed 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 seem to come from a right-skew distribution.

14. Suppose we have a set of possible explanatory variables, and want to select a subset that gives a good model. Which of the following statements is **FALSE**?

    (zz) Choosing the model that has the biggest adjusted R^2 and choosing the model with the smallest residual standard error often result in different models.

    (1) An over-fitted model will usually have regression coeficients with big standard errors.

    (1) An under-fitted model will result in biased predictions.

    (1) AIC tends to select bigger models (i.e. having more variables) than BIC.

    (1) We can't use the residual sum of squares to select the best model.

15. In a regression with $k = 4$ explanatory variables and 100 observations, there is a point with a HMD of 0.2 and an externally studentised residual of 0.3. Which of the following is **TRUE**?

    (zz) From the information given, we can't tell if the point is influential or not.

    (1) The point is not a high leverage point.

    (1) The point is definitely an outlier.

    (1) "Externally Studentized" means the residual is divided by its variance.

    (1) We can't tell if the externally studentized residual is large or not as it depends on the units of measurement.