DEPARTMENT OF STATISTICS

Course STATS 330/772: Advanced Statistical Modelling/Special Topic in Regression

Term Test: 8.00am - 9:00am, Thursday Sept 14, 2006

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. A data set consists of a continuous variable Y, and two explanatory variables X (which is continuous) and A, which is a factor having 3 levels. We want to see if the relationship between X and Y changes with the different levels of A. Which of the following pieces of R code would produce the most informative plot?

   (zz) `xyplot(Y~X|A)`
   (1) `bwplot(Y~A|X)`
   (1) `dotplot(Y~A|X)`
   (1) `pairs(data.frame(X,Y,A))`
   (1) `plot(lm(Y~X))`

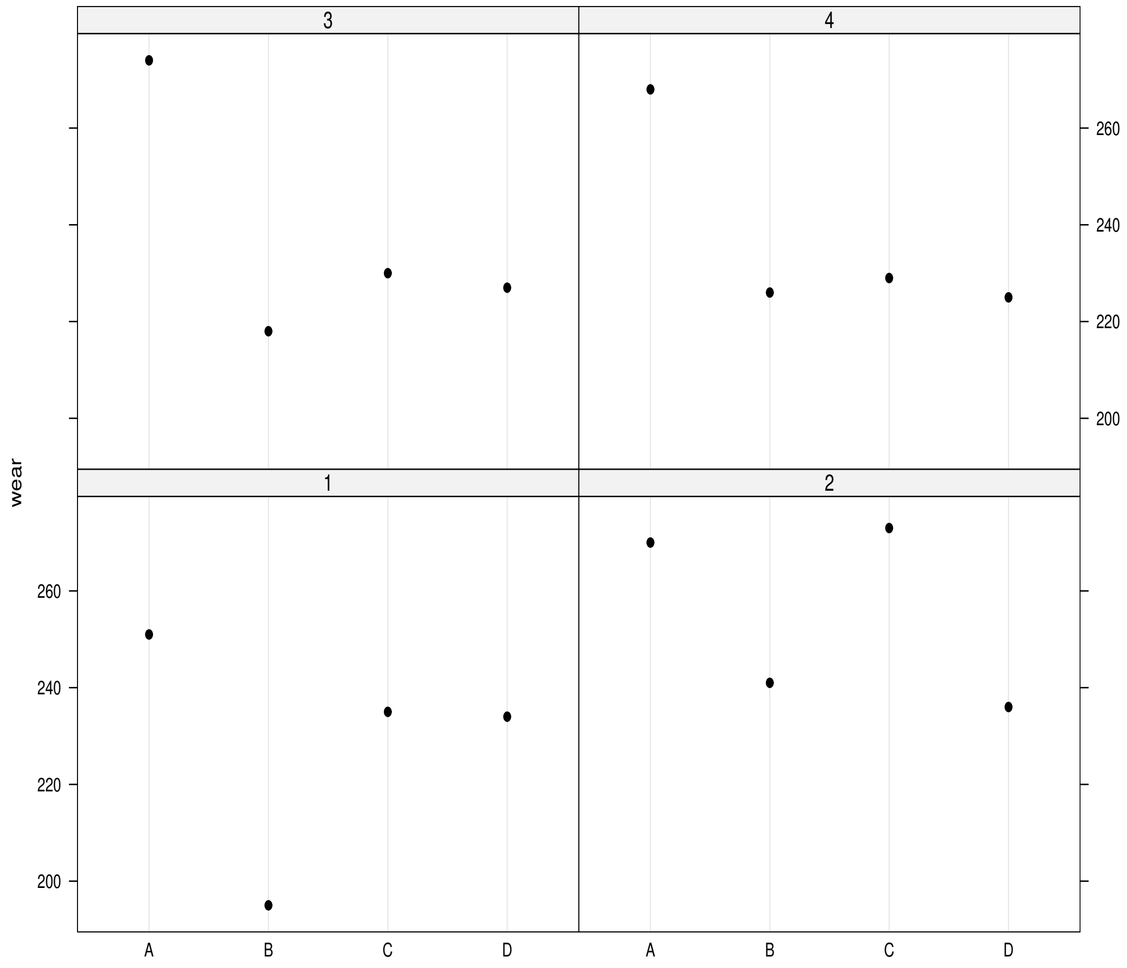

Figure 1: Trellis plot for Question 2.

2. Figure 1 shows data gathered in an experiment to compare the wear resistance of four types of material. The different types of material were fed into a wear testing machine and the amount of wear (i.e. loss of weight) recorded. Four samples could be processed

at the same time and the position of these samples in the machine may be important. The data set contains the following variables:

position: The position number 1-4,

material: The material A-D,

wear: Loss of weight in over the testing period.

A trellis plot of these data is shown in Figure 1. Which of the following is **FALSE**?

(zz) Materials C and D are similar in position 2.

(1) Overall, material A seems to have the most wear.

(1) Overall, material B seeems to have the least wear.

(1) B has more wear than D in position 2.

(1) B has less wear than D in position 3.

Questions 3-8 are concerned with air pollution data gathered on 41 US cities over a period of one year. The variables measured are

SO2: Average annual sulphur dioxide content of the air (Mcg/cubic metre);

temp: Average annual temperature, in degrees F;

logmfg: Log of the number of manufacturing firms;

logpop: Log of the population;

wind: Average annual wind speed, in mph;

precip: Average annual rainfall, in inches;

raindays: Number of days having preciptiation.

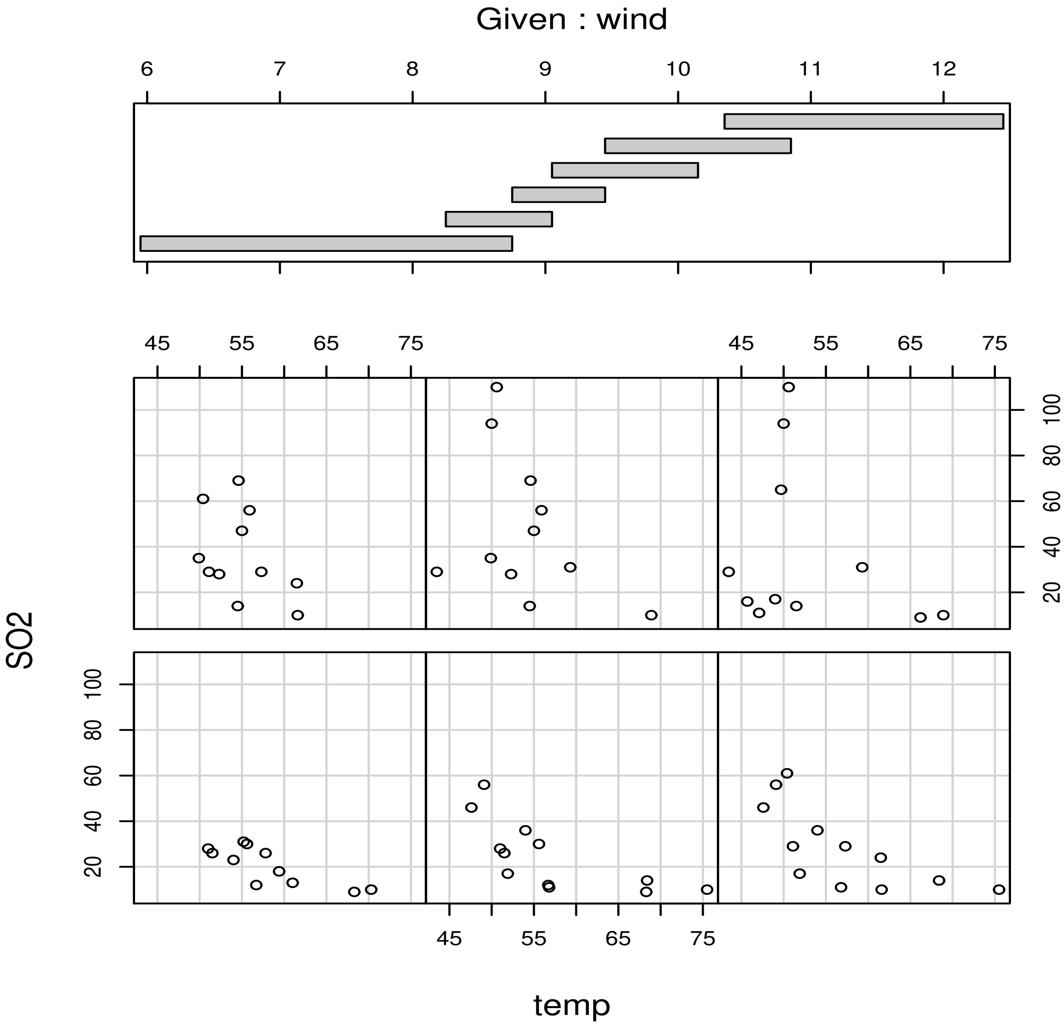

Figure 2: Coplot for Question 3.

3. A coplot of the variables `SO2`, `temp` and `wind` is shown in Figure 2. Which is the **wrong** interpretation?

   (zz) The data appear to be planar.
   (1) Some data points appear in more than one plot.
   (1) The relationship between `SO2` and `temp` gets weaker as wind increases.
   (1) As wind increases, the `SO2` at low temperatures tends to increase.
   (1) `SO2` tends to be higher at low temperatures.

4. A regression using `SO2` as the response and the all the six other variables as explanatories was fitted. Some diagnostic plots are shown in Figure 3.

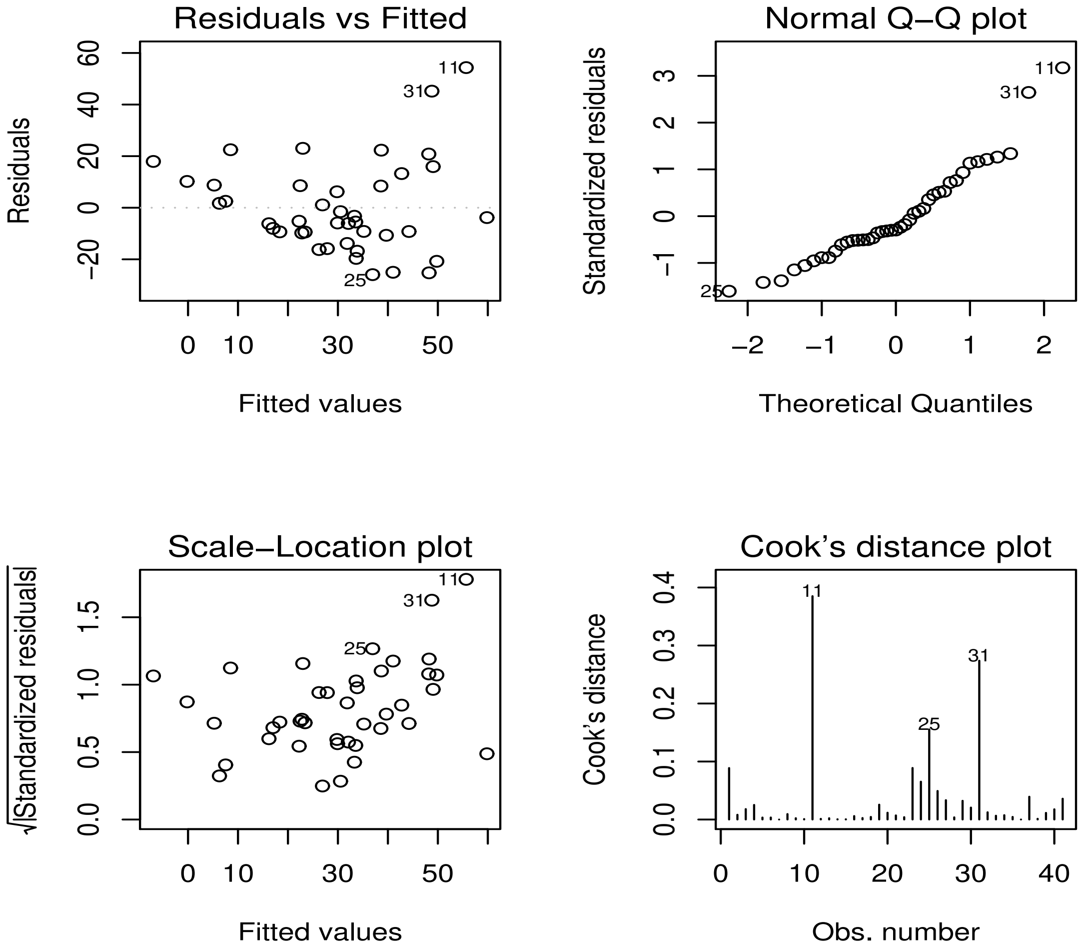

Figure 3: Diagnostic plots for Question 4.

Which of the following is the best interpretation of Figure 3?

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

(1) The bulk of the data are not normally distributed.

(1) There are no high leverage points in the data.

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

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

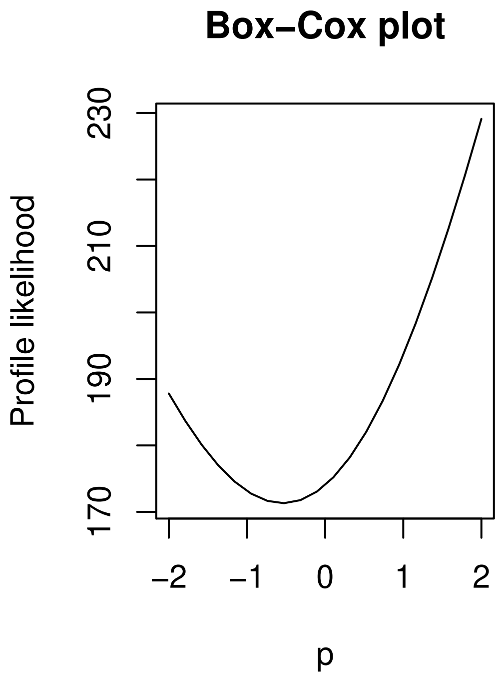

Figure 4: Diagnostic plot for Question 5.

5. A further diagnostic plot is shown in Figure 4. Which is the best interpretation of this plot?

   (zz) The response SO2 should be transformed with a power of -1/2.

   (1) No transformation of the response is indicated.

   (1) A transformation of wind and temp is indicated.

   (1) The response SO2 should be transformed with a log transformation.

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

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

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2287353  0.1594324  -1.435  0.16051
temp         0.0076280  0.0022036   3.462  0.00147 **
logmfg      -0.0283669  0.0187164  -1.516  0.13886
logpop       0.0096756  0.0226013   0.428  0.67128
wind         0.0215641  0.0063551   3.393  0.00177 **
precip      -0.0017808  0.0012843  -1.387  0.17461
raindays    -0.0001181  0.0005548  -0.213  0.83273
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.04943 on 34 degrees of freedom
Multiple R-Squared: 0.6048,     Adjusted R-squared: 0.5351
F-statistic: 8.673 on 6 and 34 DF,  p-value: 9.403e-06
```

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 0.008 units with each 1 degree increase in temperature.
(1) The mean response will increase by about 0.009 units with each extra person in the population.
(1) Both the variables `logmfg` and `logpop` should be removed from the model.
(1) The regression model assumptions are violated - the R^2 is only 60%.
(1) The variable `wind` could be removed from the model.

7. Some influence plots for the model fitted in Question 6 are shown in Figure 5. Which of the following statements is **NOT** a correct interpretation of these plots?

(zz) Observation 25 is having a big effect on the standard errors.
(1) Observation 25 is having an effect on most of the regression coefficients.
(1) Observation 31 is having an effect on some of the regression coefficients.
(1) Observation 25 has the biggest residual.
(1) Observation 37 is influential.

8. A model selection exercise was carried on on these data, producing the following output:

```
  rssp sigma2 adjRsq     Cp    AIC    BIC    CV  temp logmfg logpop wind precip raindays
1 0.139  0.004  0.322 19.889 60.889 64.317 0.015    1      0      0    0      0        0
2 0.122  0.003  0.392 14.731 55.731 60.871 0.014    1      0      0    1      0        0
3 0.099  0.003  0.492  7.430 48.430 55.284 0.012    1      0      0    1      1        0
4 0.084  0.002  0.558  3.206 44.206 52.774 0.010    1      1      0    1      1        0
5 0.083  0.002  0.548  5.045 46.045 56.327 0.011    1      1      1    1      1        0
6 0.083  0.002  0.535  7.000 48.000 59.995 0.012    1      1      1    1      1        1
```

Which model should we choose?

(zz) A model using all the explanatory variables except `logpop` and `raindays`.
(1) A model using all the explanatory variables.
(1) A model using `temp` and `wind` only.
(1) A model using `temp` only.
(1) A model using all the explanatory variables except `raindays`.

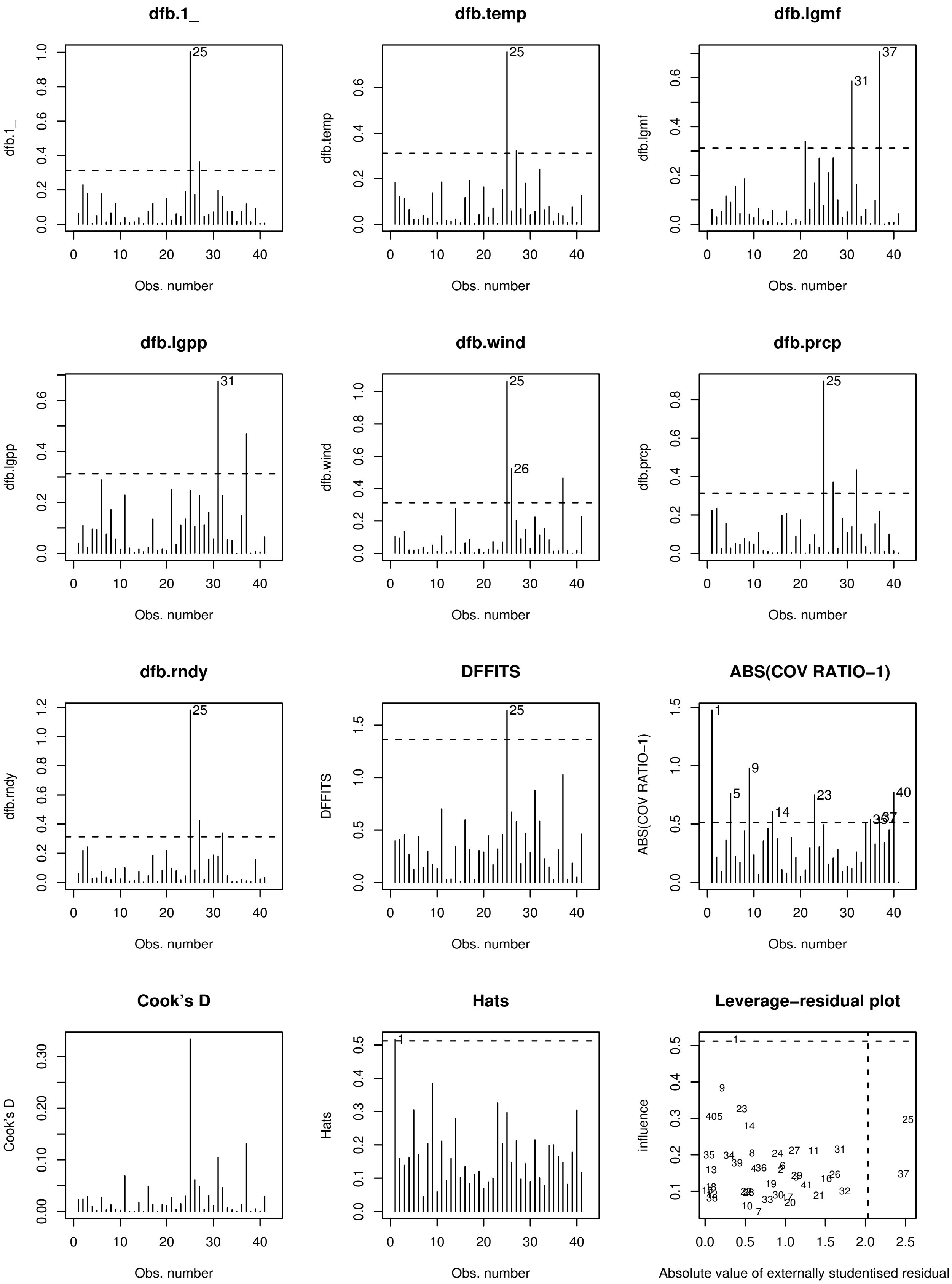

Figure 5: Influence plots for Question 7.

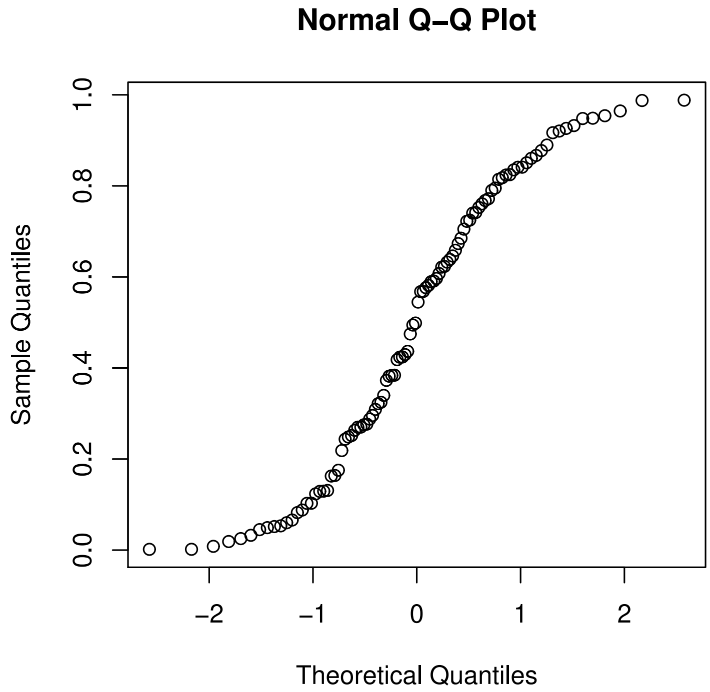

Figure 6: Normal plot for Question 9.

9. 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, short-tailed distribution.

   (1) The errors appear normal.

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

   (1) The errors seem to not be independent.

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

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

    (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 a big prediction error.

    (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.

11. In a regression having observations taken sequentially in time, it was suspected that some serial correlation in the errors might be present. Which of the following conclusions are indicated by the graphs in Figure 7?

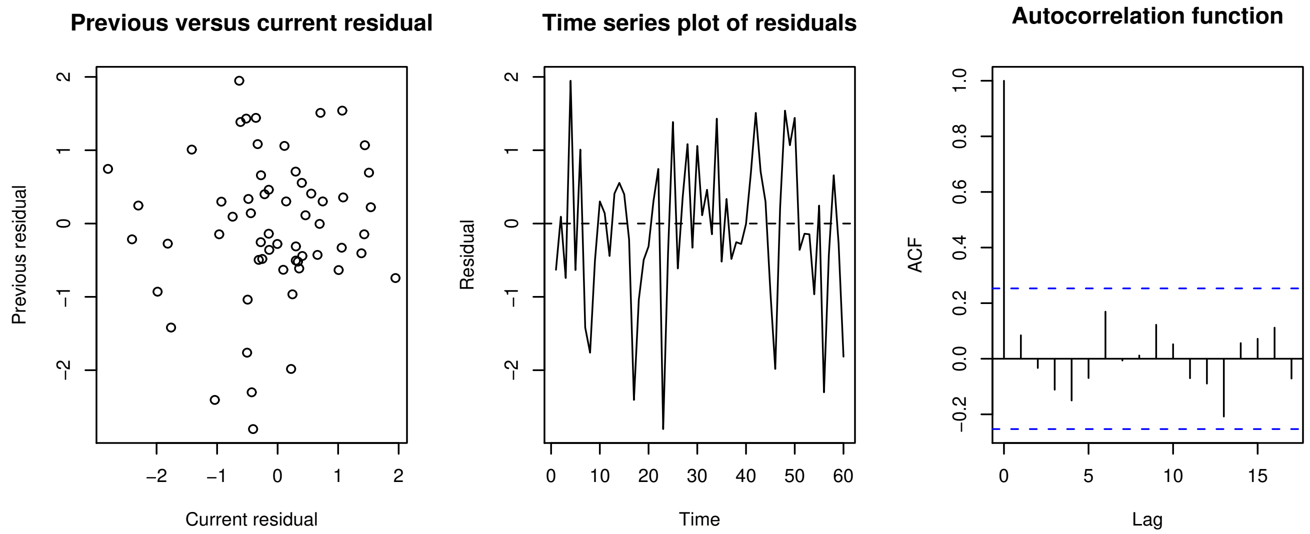

Figure 7: Diagnostic plots for Question 11.

(zz) There is no serial correlation.
(1) There is strong positive serial correlation.
(1) There is weak positive serial correlation.
(1) There is strong negative serial correlation.
(1) There is weak negative serial correlation.

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

(zz) Deleting a point that is a large outlier but has low leverage will tend to decrease the residual sum of squares.
(1) Deletion of a high influence point will always decrease the residual sum of squares.
(1) n-fold (leave one out) cross-validation is a better method than ten-fold cross-validation.
(1) The normality assumption is the most important assumption in the regression model.
(1) The Box-Cox plot gives a method for transforming explanatory variables.

The last three questions concern the following data, which was recorded on 39 students. The variables are

height: Height in metres,

weight: Weight in kg,

sex: Gender (M/F).

A plot of the data are shown in Figure 8.

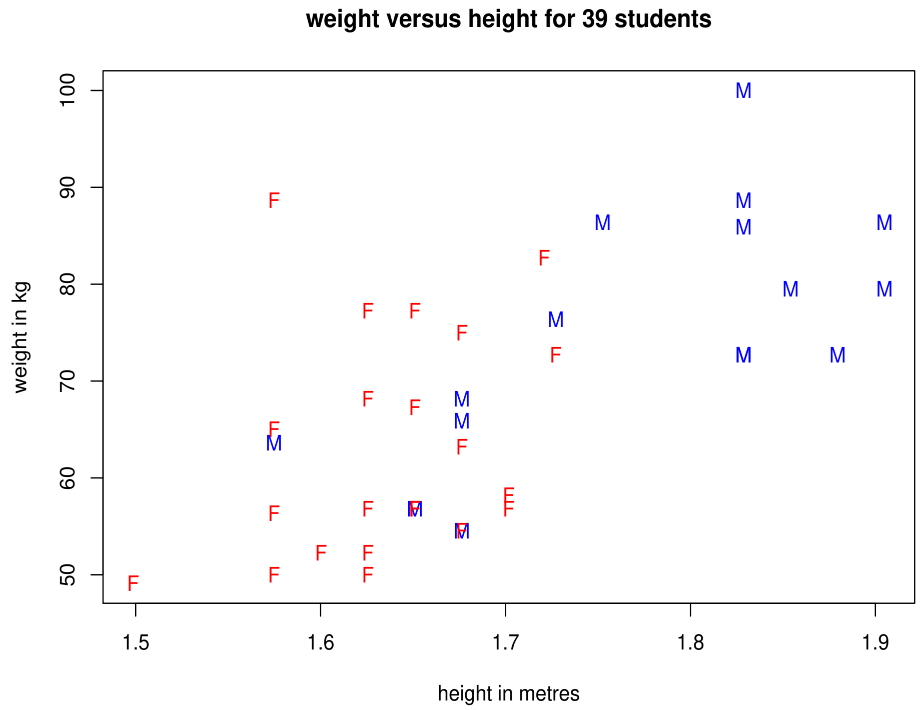

Figure 8: Weight versus height for 39 students. M=male, F=female.

13. We fitted the "parallel lines" model `weight ~ height + sex`, and got the following output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -59.399     36.260  -1.638  0.11010
height        74.657     22.105   3.377  0.00177 **
sexm           2.403      4.557   0.527  0.60117
---
Residual standard error: 10.34 on 36 degrees of freedom
Multiple R-Squared: 0.4177,     Adjusted R-squared: 0.3854
F-statistic: 12.91 on 2 and 36 DF,  p-value: 5.917e-05
```

All of the following interpretions are incorrect except one. Which one is **correct**? (Assume that we are using the "baseline" definition of effects, and the factor levels are in alphabetical order.)

(zz) The intercept of the"female" line is -59.399.
(1) On average, females are 2.4 kg heavier than males.
(1) The error variance is 10.34.
(1) Because the R^2 is so low, none of the variables contribute to the regression.
(1) There is no relationship between gender (sex) and weight.

14. We also fitted the "non-parallel lines" model, and got the output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -35.00      66.77  -0.524    0.603
height         59.76      40.76   1.466    0.152
sexm          -33.39      81.99  -0.407    0.686
height:sexm    21.31      48.75   0.437    0.665

Residual standard error: 10.45 on 35 degrees of freedom
Multiple R-Squared: 0.4209,     Adjusted R-squared: 0.3713
F-statistic:  8.48 on 3 and 35 DF,  p-value: 0.0002284
```

Which is the **correct** interpretation?

(zz) The slope of the "male" line is 81.07.
(1) The intercept of the"male" line is -33.39.
(1) For small heights, the "male" line is above the "female" line.
(1) Since the R^2 for this model is greater than that for the "parallel lines" model, we should use this model.
(1) Since the error variance for this model is greater than that for the "parallel lines" model, we should use this model.

15. In the code below, we compare two possible models. What is the hypothesis being tested by the p-value 0.7945?

```
> model1.lm = lm(weight ~ height*sex, data = students.df)
> model2.lm = lm(weight ~ height, data = students.df)
> anova(model2.lm, model1.lm)
Analysis of Variance Table

Model 1: weight ~ height
Model 2: weight ~ height * sex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     37 3876.3
2     35 3825.7  2      50.6 0.2315 0.7945
```

(zz) Gender is not required in the model, provided height is included.

(1) There is no interaction between sex and height.

(1) The male and female lines are parallel.

(1) The parallel lines model does not fit well.

(1) The male and female lines have the same intercept.