DEPARTMENT OF STATISTICS

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

Term Test: 8.00am - 9:00am, Tuesday September 14, 2010

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. . Suppose we have a data set consisting of three continuous variables X, Y and Z. We want to fit a regression model using Y as the response and X and Z as the explanatory variables. Which of the following plots would be most suitable for assessing if a transformation of the response should be made?

   (1) A plot of residuals versus time.

   (2) An autocorrelation plot.

   (3) A normal plot.

   (4) A Leverage-Residual plot.

   (5) A Box-Cox plot.

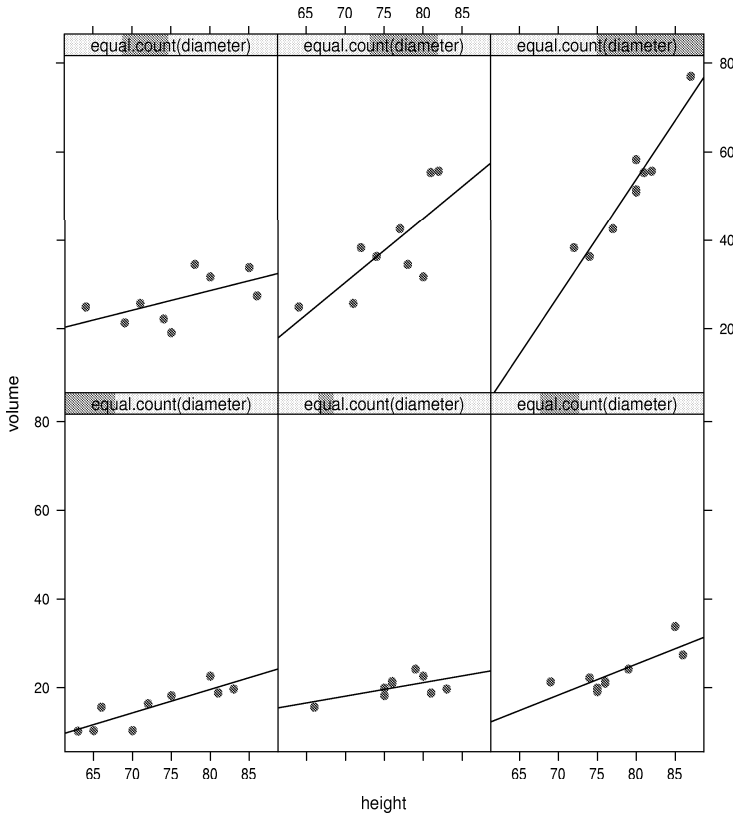

Figure 1: Trellis plot for Question 2.

2. Figure 1 shows a trellis plot of the cherry tree data discussed in class. (Recall that this data set has three variables: volume, height and diameter.) Least squares lines have been fitted to each panel.

   Which of the following is **TRUE**?

   (1) The fact that the fitted least squares lines fit well in the bottom panels suggests that the data are planar.

   (2) The gray bars above each panel show the distribution of height.

CONTINUED

(3) As diameter goes up, volume goes down.

(4) The fact that the fitted least squares lines are not parallel suggests that the regression surface is curved.

(5) As height goes up, diameter goes down.

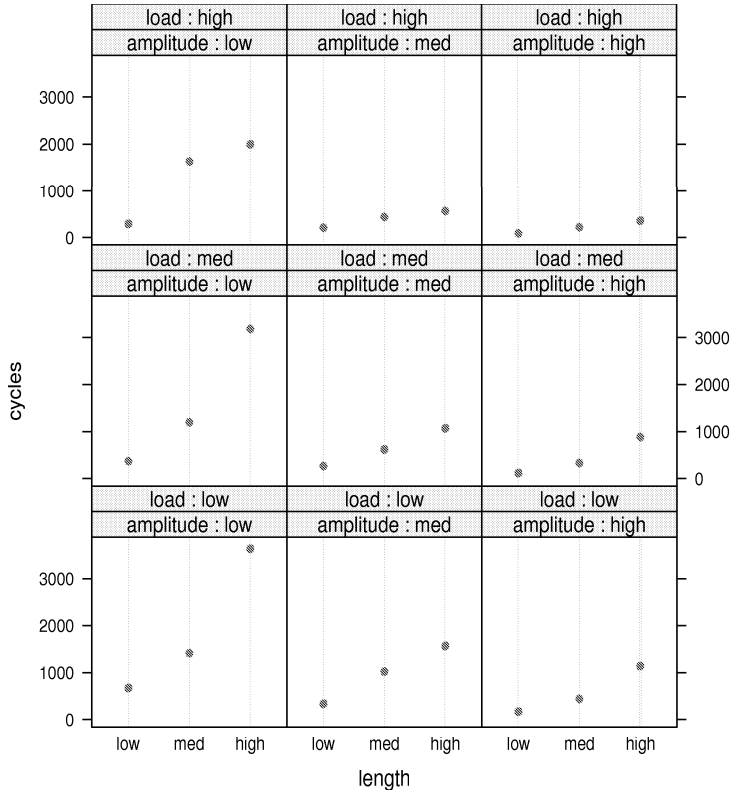

Figure 2: Trellis plot for Question 3.

3. Figure 2 is a plot of the yarn data discussed in class. The response is the number of cycles to failure, which is thought to possibly depend on the variables length, amplitude and load, each of which is a categorical variable having 3 levels, High, Medium and Low.

   Which of the following is **TRUE**?

   (1) There isnt enough information in the graph to decide how the factors load, amplitude and length affect the number of cycles required.

   (2) The higher (longer) the length, the more cycles are required.

   (3) The lower the amplitude, the fewer cycles are required.

   (4) The lower the load, the more cycles are required.

   (5) The number of cycles required doesn't depend on load or amplitude.

4. In the linear regression model, which is the least important assumption?

   (1) The variances are equal.

   (2) The regression surface is planar.

   (3) There are no outliers.

   (4) The observations are independent.

   (5) The responses are normally distributed.

5. In a regression, an explanatory variable X has a coefficient of -2, and a p-value of 0.0001. Which of the following is the best interpretation?

   (1) On average, the response will increase by 2 units when X increases by one unit, provided all other variable are held constant.

   (2) On average, the response will decrease by 2 units when X increases by one unit.

   (3) On average, the response will not change when X increases by one unit, provided all other variable are held constant.

   (4) On average, the response will decrease by 2 units when X increases by one unit, provided all other variable are held constant.

   (5) On average, the response will increase by 2 units when X increases by one unit.

6. Which of the following statements is correct?

   (1) In a regression the R^2 will be zero if all the estimated regression coefficients (apart from the constant term) are zero.

   (2) R^2 is the ratio of the regression sum of squares to the residual sum of squares.

   (3) In a regression the R^2 will be equal to 1 if all the true regression coefficients (apart from the constant term) are zero.

   (4) In exceptional cases, it is possible to get a negative R^2 value.

   (5) R^2 will be small if serial correlation is present in the data.

Questions 7-10 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 regression using `Ozone` as the response and the all the other variables as explanatories was fitted. The summary output is

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -59.79981   23.23638  -2.574  0.01145 *
radiation     0.06493    0.02339   2.775  0.00652 **
temperature   1.59354    0.25827   6.170 1.27e-08 ***
wind         -3.41456    0.65491  -5.214 9.21e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 21.22 on 106 degrees of freedom
Multiple R-squared: 0.6065,     Adjusted R-squared: 0.5953
F-statistic: 54.45 on 3 and 106 DF,  p-value: < 2.2e-16
```

Which of the following is **NOT** indicated by this output?

(1) At least one of the explanatory variables is related to the response.

(2) High temperatures tend to increase ozone, assuming constant values of the other variables.

(3) The residual sum of squares is about 47700.

(4) High winds tend to lower the ozone in the air, assuming constant values of the other variables.

(5) The estimate of the error variance is 21.22.

8. Some diagnostic plots from the regression are shown in Figure 3.

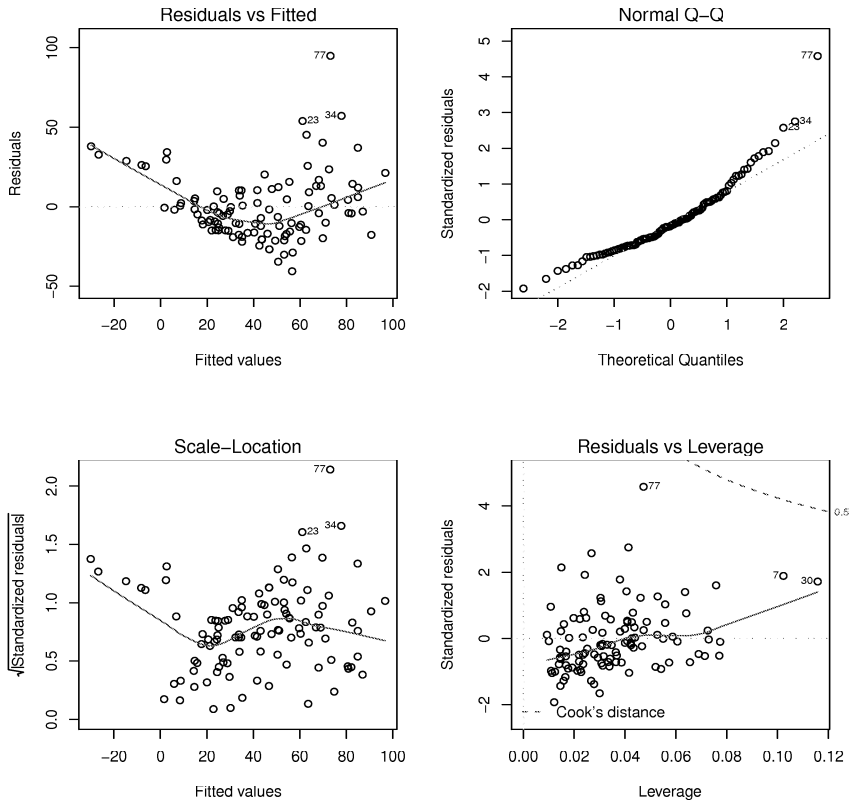

Figure 3: Diagnostic plots for Question 8.

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

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

(2) Several points have very high leverage.

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

(4) There are outliers in the data.

(5) The data are not planar.

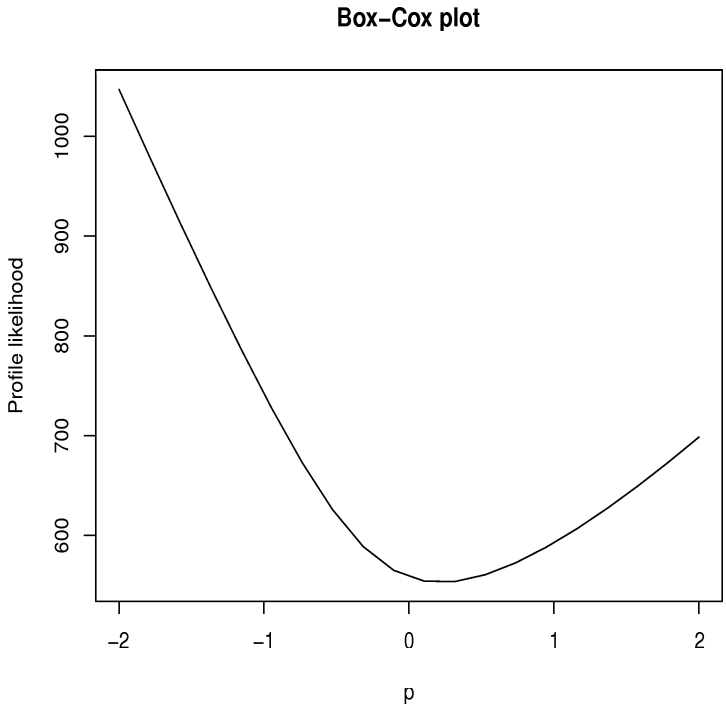

Figure 4: Box-Cox plot for Question 9.

9. A Box-Cox plot of the ozone data is shown in Figure 4. On the basis of Figures 3 and 4, which of the following is the best strategy?

   (1) No transformations need be made, but you need to delete some outliers.

   (2) Because the plot is quadratic, you should fit a polynomial in the response.

   (3) You should transform the response with a log transformation.

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

   (5) Log transformations of `radiation`, `wind` and `temperature` are indicated.

10. After taking some corrective action, a new model was fitted. Some influence plots for the new model are shown in Figure 5. Which of the following statements is **NOT** a correct interpretation of these plots?

    (1) Observation 17 is having an effect on the fitted value for point 17.

    (2) Observation 17 is having an effect on some of the standard errors.

    (3) Observation 17 is not having an appreciable effect on the the coefficient for solar radiation.

    (4) Observation 17 is not a high-leverage point.

    (5) Observation 17 is having an effect on most of the regression coefficients.

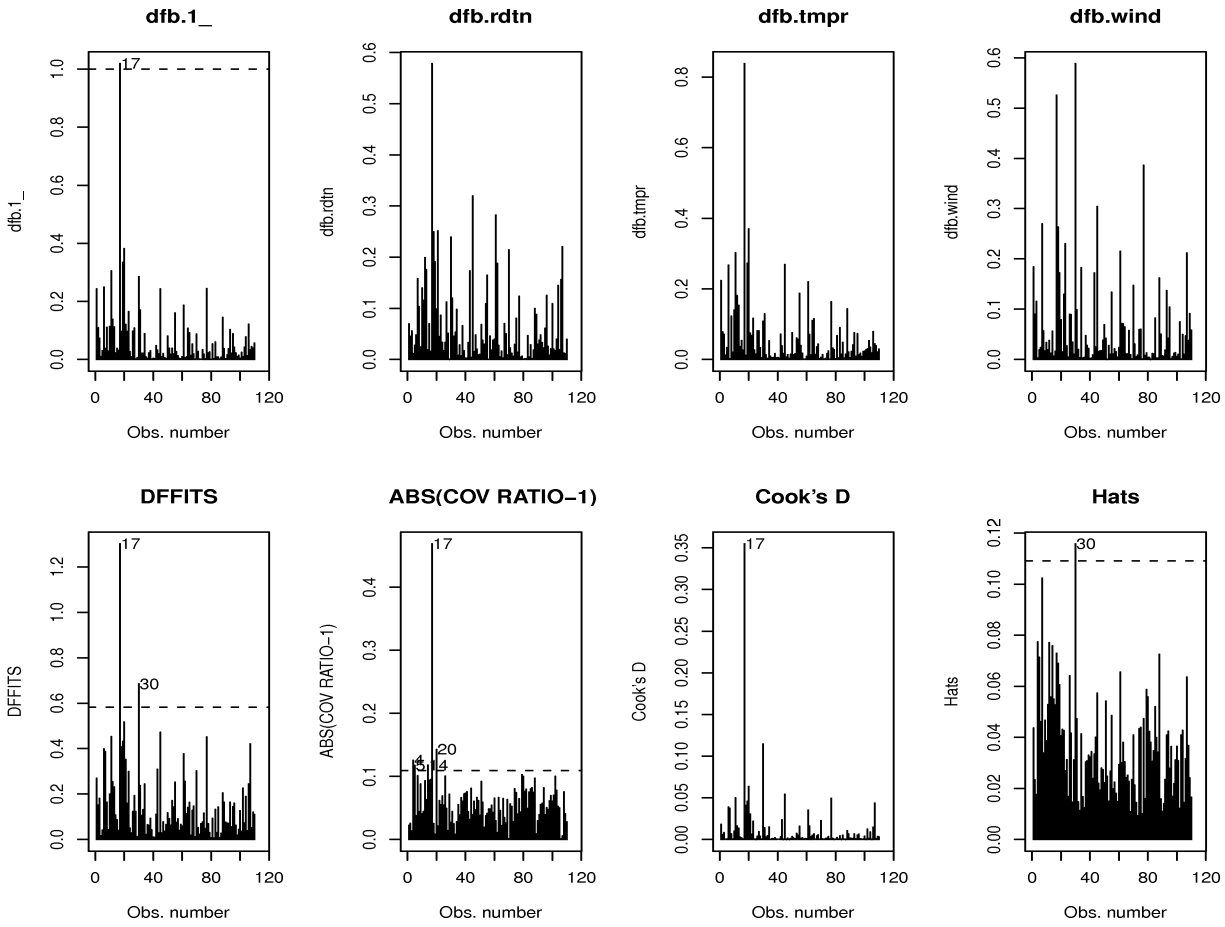

Figure 5: Influence plots for Question 10.

11. The education data studied in class consisted of several variables measured for each US state, excluding California. We fitted a regression to the response variable `educ`, (the percapita expenditure on education by a state) using three other variables (`urban`, `percap` and `under18`). Which of the following statements is **NOT** correct, based on the all possible regressions output below?

```
      rssp   sigma2 adjRsq    Cp    AIC    BIC       CV urban percap under18
1 70509.72 1500.207  0.372 9.769 58.769 62.553 6285.712     0      1       0
2 60223.81 1309.213  0.452 3.779 52.779 58.455 5602.218     0      1       1
3 57933.14 1287.403  0.461 4.000 53.000 60.567 5526.254     1      1       1
```

(1) The model `educ~percap+under18` is indicated by the AIC criterion.

(2) The model `educ~urban+percap+under18` is a reasonable model.

(3) The `sigma2` and `adjRsq` criteria indicate the same model.

(4) The model `educ~urban+percap+under18` is indicated by the BIC criterion.

(5) The model `educ~percap` is not a good model.

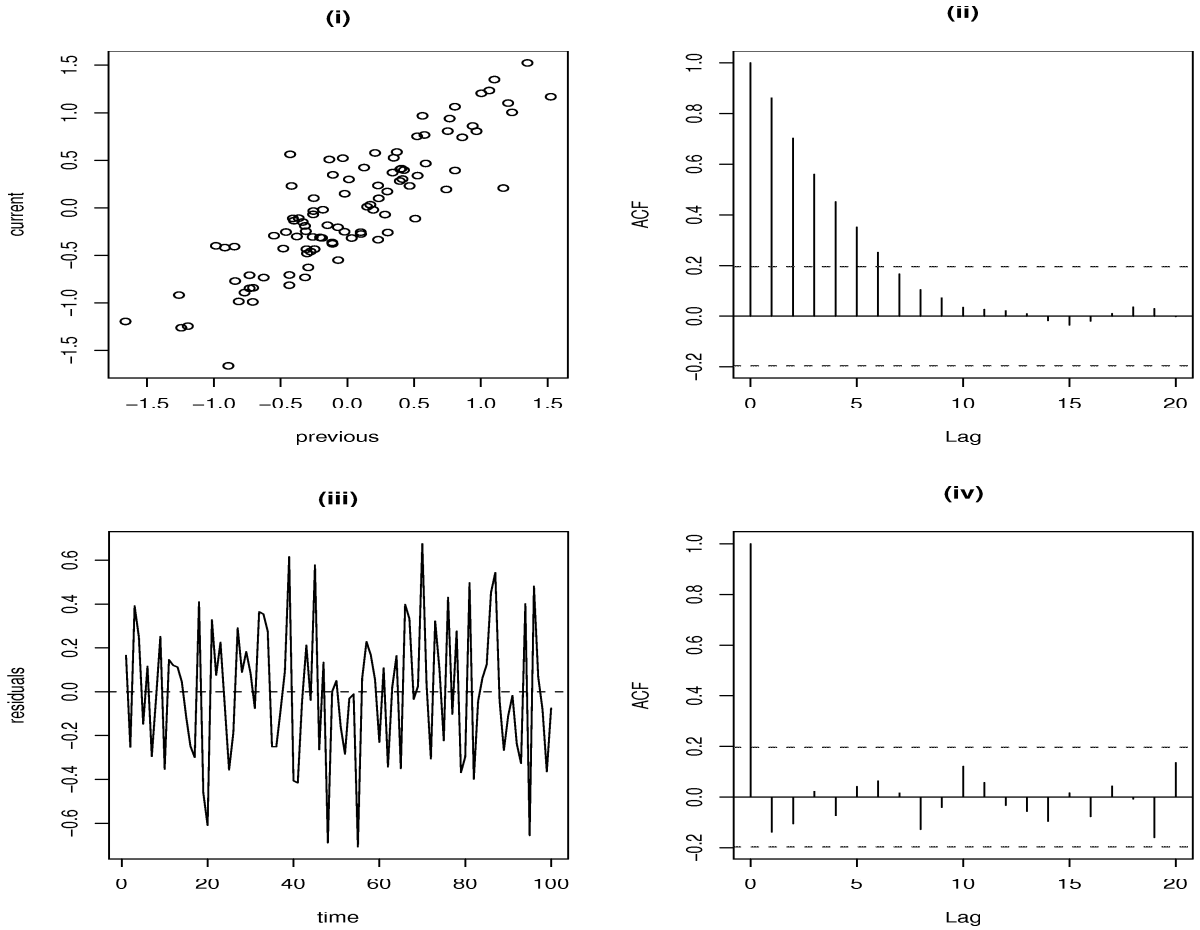

Figure 6: Diagnostic plots for Question 12.

12. Suppose we have four regressions where the data are in time order. In the plots in Figure 6, plot (i) is a plot of residuals versus the residual from the previous observation, plots (ii) and (iv) are acf plots of residuals, and plot (iii) is a time series plot of residuals (i.e. a plot of residuals versus time order). All the plots refer to separate regressions.

    Which of the following is **FALSE**?

    (1) In plot (iii), the errors seem to be uncorrelated.

    (2) Plot(iii) suggests that the residuals have zero mean.

    (3) In plot (iv), there is no evidence that the errors are correlated.

    (4) In plot(ii), the errors are strongly positively autocorrelated.

    (5) In plot(i), there is evidence of negative autocorrelation in the errors.

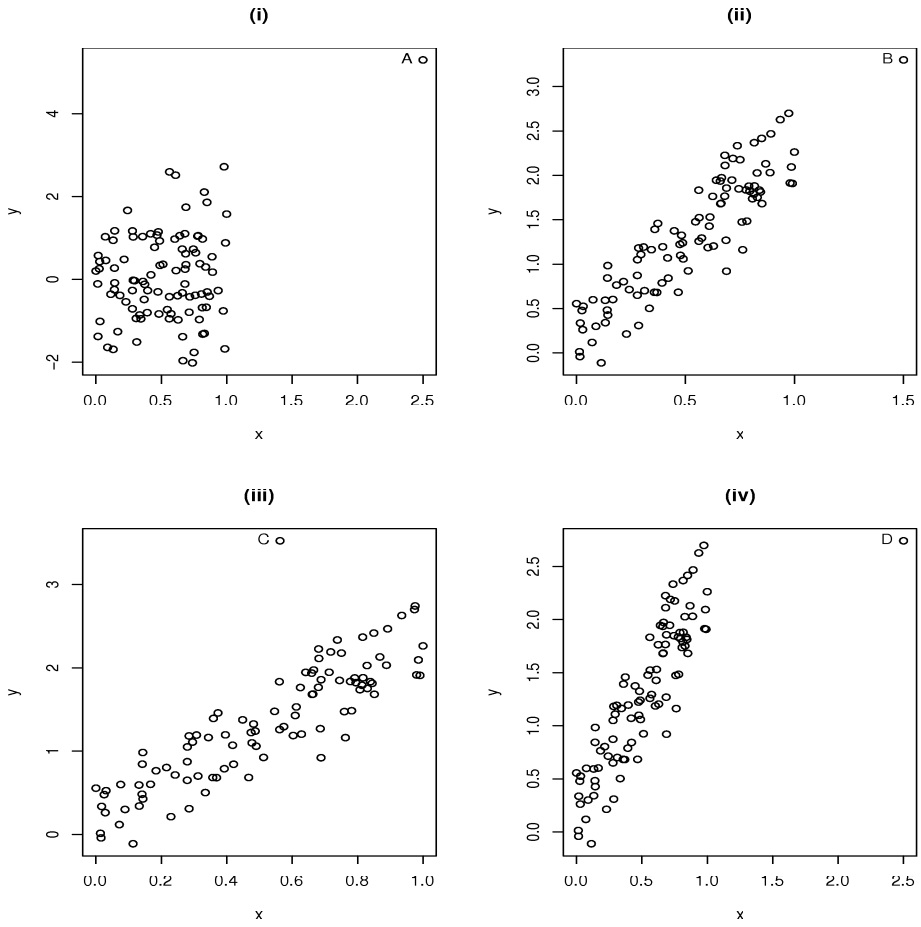

Figure 7: Scatter plots for plots for Question 13.

13. In Figure 7, we show four scatter plots of a variable `y` versus `x` for four different data sets. In each case a regression of `y` on `x` was fitted. Which of the following is **FALSE**?

(1) In plot (i), removing point A will increase the R^2.

(2) In plot (ii), the point B is a high leverage point.

(3) In plot (iii) removing point C will increase the R^2.

(4) In plot (ii) point B is not likely to be influential.

(5) In plot (iv), point D is likely to be very influential.

Questions 14 and 15 are based on the following: Consider an experiment to compare the lifetimes of two different makes of cutting tool, A and B. Ten tools of each type were tested at different cutting speeds. The lifetime of each tool and the speed at which it was rotated were recorded. A regression model was fitted to the data. The response variable is the lifetime (in hours),and the explanatory variables are the type (the variable `type`, a factor having values A and B) and the speed (in rpm).

14. We first fitted a "parallel lines" model to the data and obtained the output below:

```
> model1 = lm(lifetime~speed+type, data=tool.df)
> summary(model1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
speed       -0.02661    0.00452  -5.887 1.79e-05 ***
typeB       15.00425    1.35967  11.035 3.59e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-squared: 0.9003,     Adjusted R-squared: 0.8886
F-statistic: 76.75 on 2 and 17 DF,  p-value: 3.086e-09
```

Which of the following is **FALSE**?

(1) The output suggests that the true regression lines for A and B coincide.

(2) The fitted line for type A tools has slope -0.02661.

(3) The fitted line for type B tools has slope -0.02661.

(4) The fitted line for type B tools is 15.00425 hours above that for type A tools.

(5) The fitted line for type A tools has intercept 36.98560.

15. We then expanded the model so that the lines were no longer constrained to be parallel. We obtained the following output:

```
> model2<-lm(lifetime~speed*type, data=tool.df)
> summary(model2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.633472   7.073 2.63e-06 ***
speed       -0.020970   0.006074  -3.452  0.00328 **
typeB       23.970593   6.768973   3.541  0.00272 **
speed:typeB -0.011944   0.008842  -1.351  0.19553
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.968 on 16 degrees of freedom
Multiple R-squared: 0.9105,     Adjusted R-squared: 0.8937
F-statistic: 54.25 on 3 and 16 DF,  p-value: 1.319e-08
```

Which of the following is **FALSE**?

(1) The "parallel lines model seems adequate: there is no need for different slopes.

(2) Under model 2, a type A bit running at 500 rpm has expected lifetime 22.28976.

(3) Under model 1, a type B bit running at 500 rpm has expected lifetime 38.68485.

(4) Under model 1, a type A bit running at 500 rpm has expected lifetime 23.68060.

(5) Under model 2, a type B bit running at 500 rpm has expected lifetime 50.77335.
