Department of Statistics

Course STATS 330

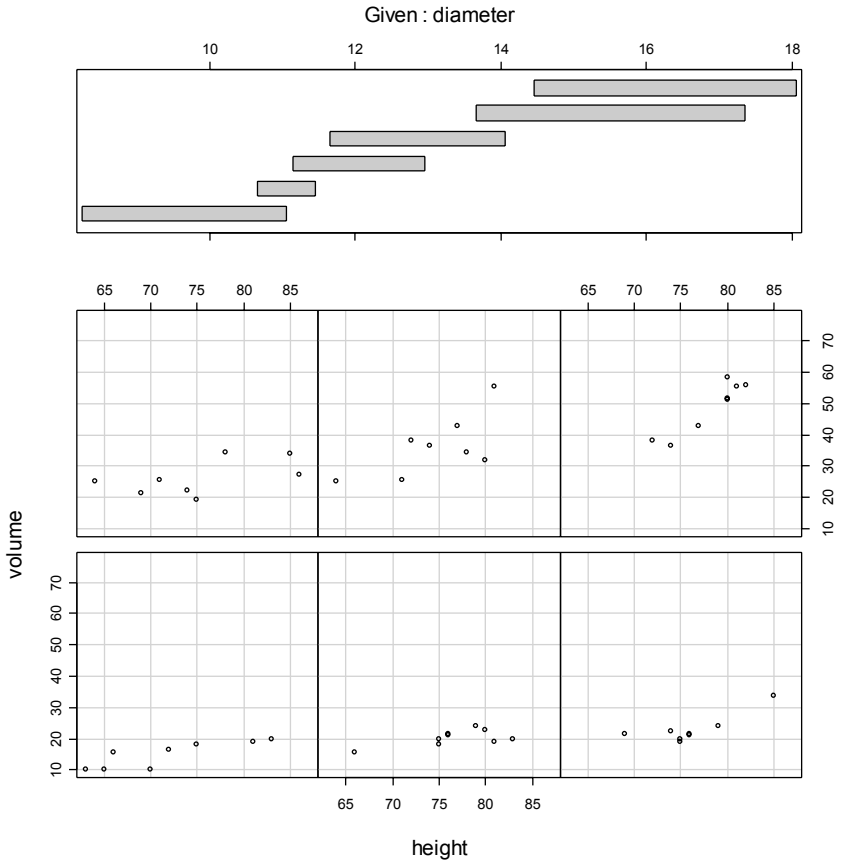
Term Test 2003. 9:00 -10:00 Friday, Sept 19, 2003

Answer all 15 questions on the answer sheet provided.

Question 1. Suppose that 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. Which of the following plots would **not** be suitable for assessing (before the model is fitted) if a linear regression model should be fitted to the data?

(1) A Box-Cox plot.1
(2) A coplot.2
(3) A 3-dimensional scatter plot.3
(4) A trellis plot.4
(5) A pairs plot.5

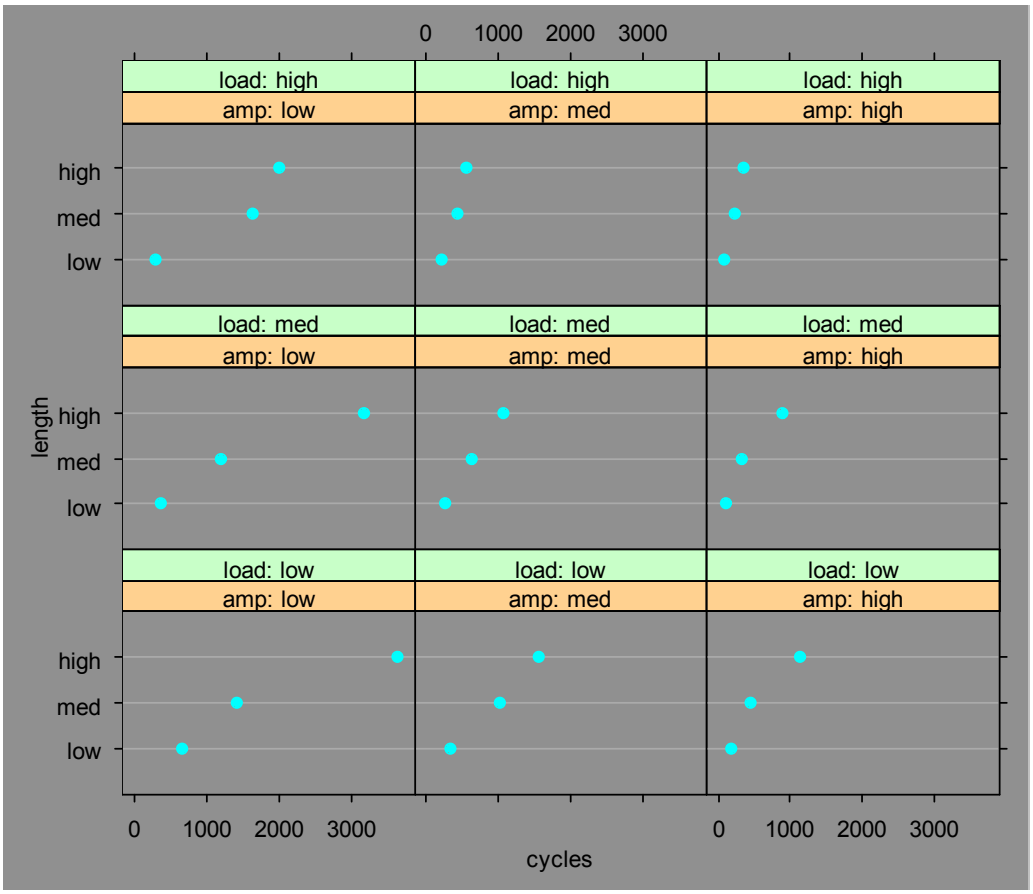
Question 2. Below is a coplot of the cherry tree data discussed in class. (Recall that this data set has three variables: volume, height and diameter.)



Which of the following is the best summary of the information in the plot?

(1) The plots indicate that the data are not planar.1
(2) The plot indicates that a linear regression model is appropriate.2
(3) The plot shows outliers that need deleting.3
(4) The bottom line of plots indicates that the data need transforming.4
(5) The plot shows that the data need transforming to equalise the variances.5

Question 3. Below 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.



Only one of the following statements is **correct**. Which one?

(1) The lower the load, the more cycles are required.1
(2) The higher (longer) the length, the fewer cycles are required.2
(3) The lower the amplitude, the fewer cycles are required.3
(4) The number of cycles required doesn't depend on load or amplitude.4
(5) There isn't enough information in the graph to decide how the factors load, amplitude and length affect the number of cycles required.5

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

(1) The mean is a linear function of the explanatory variables.1
(2) The variances are equal.2
(3) The observations are independent.3
(4) The responses are normally distributed.4
(5) There are no outliers.5

Question 5. In the following, select the **most correct** statement. The size of the regression coefficient β corresponding to a variable X in a multiple regression:

(1) Measures the change in the mean response associated with a unit increase in X, with the other variables held constant.1
(2) Measures the strength of the relationship between X and the response.2
(3) Measures the importance of X in the regression.3
(4) Measures the change in the mean response associated with a unit increase in X.4
(5) Measures the correlation between X and the response.5

Question 6. In the free fatty acid example discussed in class, the response variable was free fatty acid (`ffa`) , and the explanatory variables were `age`, `weight` and `skinfold`. Examine the R output below and then select the statement which is a valid conclusion **using this output only**.

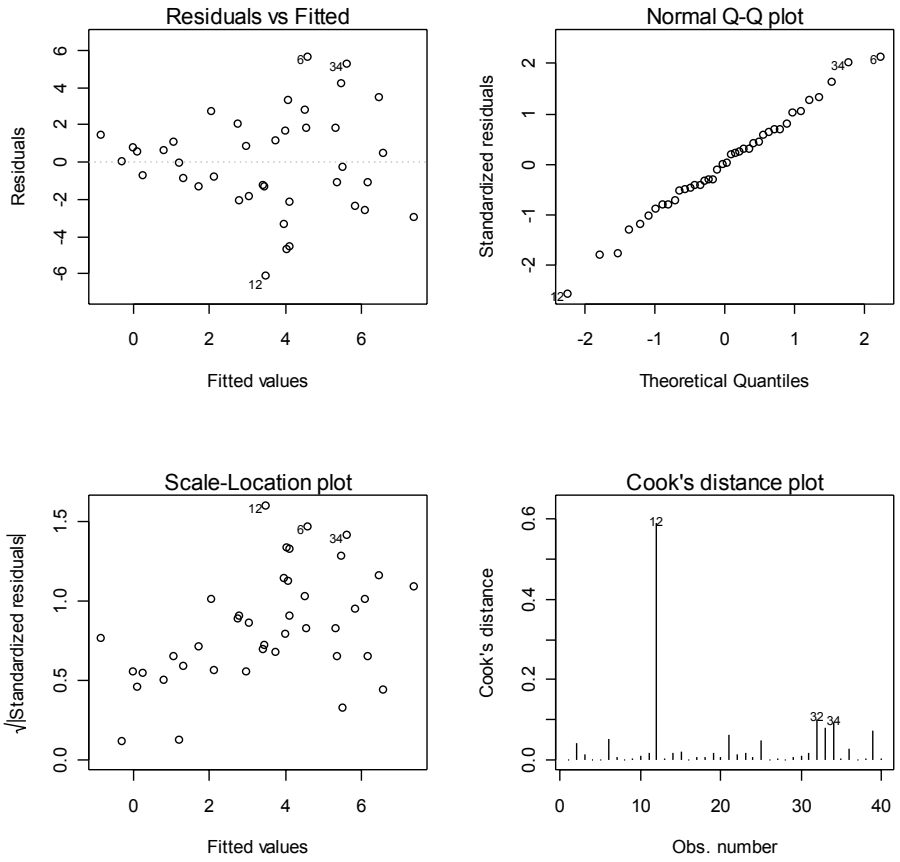
```
> model.full<- lm(ffa~age+weight+skinfold, data=fatty.df)
> summary(model.full)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.95777    1.40138   2.824  0.01222 *
age         -0.01912    0.01275  -1.499  0.15323
weight      -0.02007    0.00613  -3.274  0.00478 **
skinfold    -0.07788    0.31377  -0.248  0.80714
```

(1) As `weight` increases, and `age` and `skinfold` remain constant, the `ffa` goes down.1
(2) The variables `age` and `skinfold` can both be dropped from the model.2
(3) The variable `age` is required in the model.3
(4) The variable `weight` can be dropped from the model.4
(5) The variable `skinfold` is not related to `ffa`.5

Question 7. Which of the following statements is **not correct**?

(1) An R^2 of 0.3 means that the model doesn't fit well and that the data must be transformed.1
(2) R^2 is the ratio of the regression sum of squares to the total sum of squares.2
(3) R^2 is the square of the correlation between the fitted values and the response.3
(4) $R^2 = 0$ if and only if all the regression coefficients (except the constant term) are zero.4
(5) $R^2 = 1$ if and only if all the data lie exactly on a plane.5

Question 8. Below are several diagnostic plots of a fitted regression.



What do these plots indicate?

(1) The variances are not equal.
(2) The regression surface is not planar and the independent variables should be transformed.
(3) The errors are not normal.
(4) There are at least two outliers.
(5) There are no problems with the regression.

Question 9. In the education data discussed in class, data on educational expenditure were given for each of the 50 states in the US. The response variable was educational expenditure per capita (`educ`), and the explanatory variables were the number of 18-year-olds per 1000 population (`under18`), and percapita income (`percap`). The output below was obtained.

```
> influence.measures(educ.lm)
Influence measures of
         lm(formula = educ ~ under18 + percap, data = educ.df) :

     dfb.1. dfb.un18 dfb.prcp   dffit cov.r   cook.d    hat
10  0.06381 -0.02222 -0.16792 -0.3631 0.903 4.05e-02 0.0257
44  0.02289 -0.02948  0.00298 -0.0340 1.283 3.94e-04 0.1690
49  0.10694 -0.07892 -0.12961 -0.1765 1.080 1.05e-02 0.0496
50 -2.36876  2.23393  1.50181  2.4733 0.821 1.66e+00 0.3429
```

One of the following statements is true. Which one?

(1) Observation 50 is influential since its removal causes a big change in the coefficient for "under 18".1
(2) Observation 10 is influential.2
(3) Observation 44 is influential since its removal causes a big change in the fitted value.3
(4) Observation 49 is influential since its removal causes a big change in the coefficient for "percap".4
(5) Observation 44 is influential since its Cooks D is small.5

Hint for question 9: Points are influential if (i) Cooks D is more than $F_{3,47}(0.1)=2.204$, (ii) $|\text{DFBETAS}|>2/\sqrt{n}$, (iii) $|\text{DFFITS}|>3\sqrt{(p/(n-p))}$, (iv) $|\text{Covratio}-1|>3p/n$.

Question 10. In a regression, an explanatory variable X has a variance inflation factor of 100, and a correlation with the response of 0.8, and a p-value of 0.3. One of the following statements is true. Which one?

(1) Since the VIF is high, X is strongly related to other explanatory variables and doesn't need to be included in the regression.1
(2) Since the correlation is 0.8, the variable X must be included in the regression.2
(3) Since the VIF is high, the variable X is very important and must be included in the regression.3
(4) Since the p-value is high, the variable X is unrelated to the response.4
(5) Since the VIF is high, there must be an outlier present.5

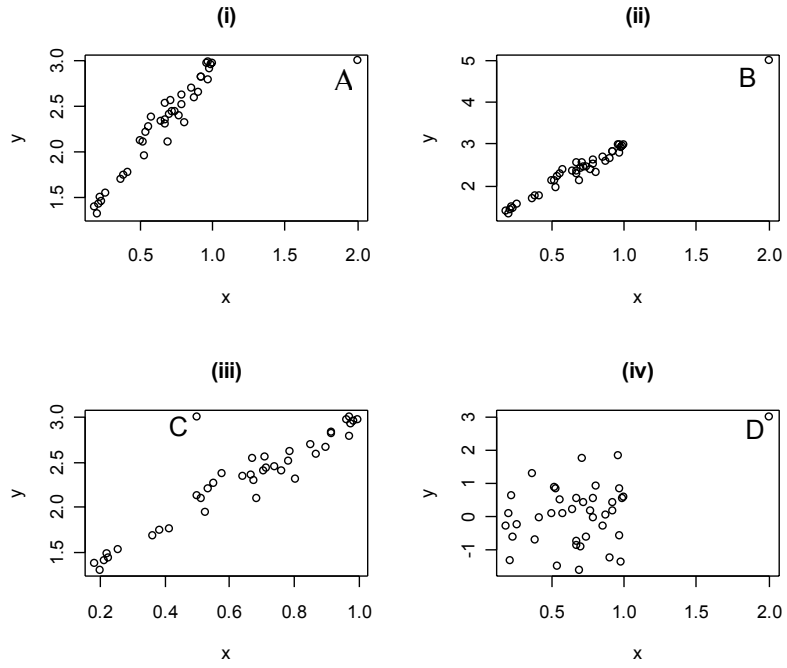
Question 11. A model was chosen for the free fatty acid data (see Question 6) using the following R code.

```
> fatty.lm <- lm(ffa ~ age + skinfold + weight, data = fatty.df)
> all.poss.regs(fatty.lm)
   rssp sigma2 adjRsq     Cp    AIC    BIC age skinfold weight
1 0.910  0.054  0.343  3.255 21.255 23.247   0        0      1
1 1.344  0.079  0.030 11.485 29.485 31.476   1        0      0
1 1.548  0.091 -0.117 15.344 33.344 35.336   0        1      0
2 0.794  0.050  0.391  3.058 21.058 24.045   1        0      1
2 0.902  0.056  0.308  5.108 23.108 26.095   0        1      1
2 1.321  0.083 -0.013 13.046 31.046 34.033   1        1      0
3 0.791  0.053  0.353  5.000 23.000 26.983   1        1      1
```

Which model is most strongly indicated by this output?

(1) **ffa ~ age + weight 1**
(2) **ffa ~ age + skinfold + weight 2**
(3) **ffa ~ age + skinfold 3**
(4) **ffa ~ skinfold + weight 4**
(5) **ffa ~ weight 5**

Question 12. Examine the plots below and select the **incorrect** statement.



Figures for Question 13

(1) In plot (iii), removing point C will decrease the R^2.1
(2) In plot (i), the point A is influential.2
(3) In plot (ii), the point B is a high-leverage point.3
(4) In plot (iv), point D is influential and has high leverage.4
(5) In plot (iii), C is not a high-leverage point.5

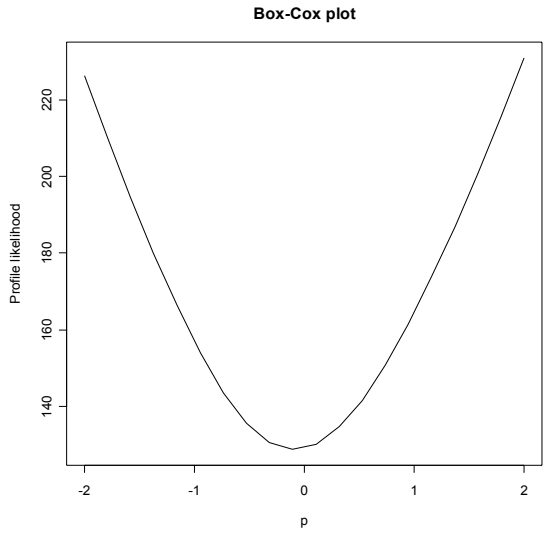
Question 13. One of the following statements is **correct**. Which one?

In a regression:

(1) If we have too many variables in a regression, the prediction error is too big.1
(2) It doesn't matter if we have too many explanatory variables.2
(3) If we have too many variables in a regression, the estimates of the coefficients are biased.3
(4) We select the model that has the biggest AIC.4
(5) We select the model with the smallest adjusted R^2.5

Question 14. In a certain regression, we suspect that transforming the response variable might improve the fit. A Box-Cox plot is shown below. What should we do?

(1) We should transform using a log.1
(2) We should transform using a reciprocal.2
(3) We should transform using a square root.3
(4) We should transform the explanatory variables, not the response.4
(5) No transformation is indicated. 5



Box – Cox plot for Question 14.

Question 15. Consider an experiment to compare three different methods A, B and C of teaching reading. Ten children were assigned to each method, and were tested before and after the course. The initial and final scores are **`start`** and **`final`** respectively. A regression model was then fitted to the data. The response variable is the final score. The explanatory variables are the method (variable **`method`**, a factor having values A, B or C) and the initial score. Study the output below and pick the **incorrect** statement.

```
> model1<-lm(final ~ method + start+ method:start)
> model2<-lm(final ~ method + start)
> anova(model2,model1)
Analysis of Variance Table

Model 1: final ~ method + start
Model 2: final ~ method + start + method:start
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     26 946.69
2     24 925.46  2     21.24 0.2754 0.7617

> summary(model2)

Call:
lm(formula = final ~ method + start)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1117  -4.3414  -0.1826   3.8537  13.1261

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8877     6.4034   1.232   0.2291
method B      4.6284     2.7058   1.711   0.0991 .
method C     -6.9582     2.7220  -2.556   0.0168 *
start         1.0391     0.1197   8.681  3.7e-09 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 6.034 on 26 degrees of freedom
Multiple R-Squared: 0.7949,     Adjusted R-squared: 0.7712
F-statistic: 33.58 on 3 and 26 DF,  p-value: 4.286e-09
```

(1) On average, students taught by method C score about 7 points more than those taught by method A.1
(2) The "parallel lines" model seems adequate: there is no need for different slopes.2
(3) There is evidence of a difference between the reading methods.3
(4) On average, students taught by method B score almost 5 points more than those taught by method A.4
(5) After the study, two new students Jack and Jill took the initial test. Jill scored 5 points more than Jack. They both then did the method A course. The model predicts that Jill will score about 5 points more than Jack on the final test.5