DEPARTMENT OF STATISTICS

Courses STATS 330/762: Statistical Modelling/ Special Topic in Statistics

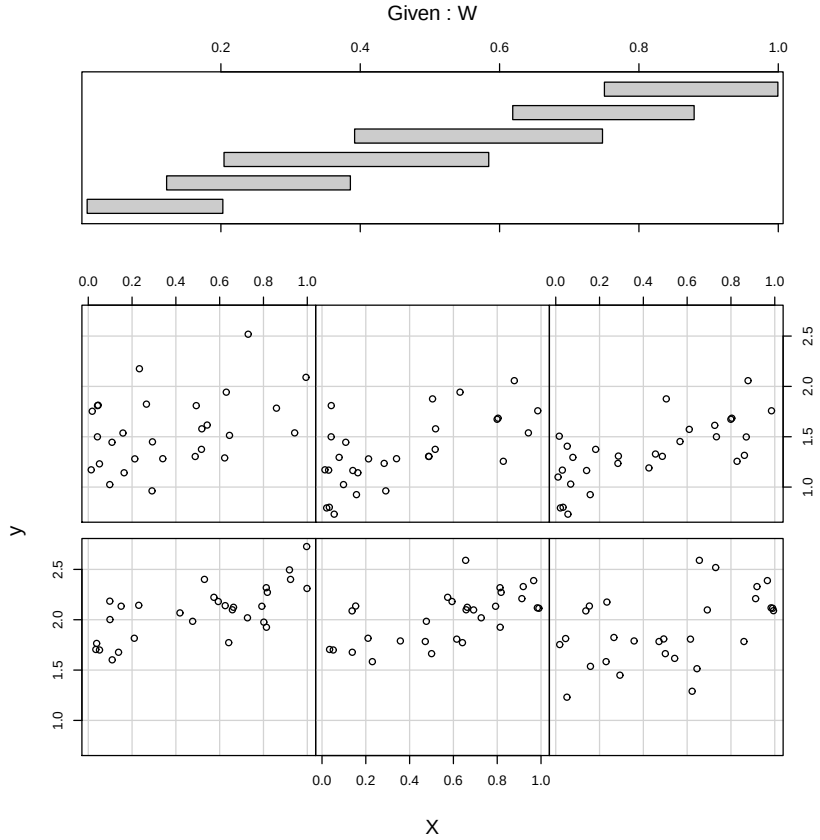
Term Test 8.00 - 9:00 Thursday Sept 13, 2012

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 we have a data set consisting of three continuous variables W, X and Y. We want to fit a regression model using Y as the response. Five types of plot are listed below. Which is the odd one out? (i.e. has a different purpose to the others)

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



2. Figure 1: Coplot for Question 2.

Figure 1 shows a coplot of the data in Question 1. Which of the following statements is **NOT** correct?

(1) The plot shows outliers that need deleting.

(2) The regression coefficient of X is positive.

(3) For fixed X, as W goes up, the response tends to go down.

(4) The plot indicates that a linear regression model is appropriate.

(5) The values of Y are less than 3.

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.

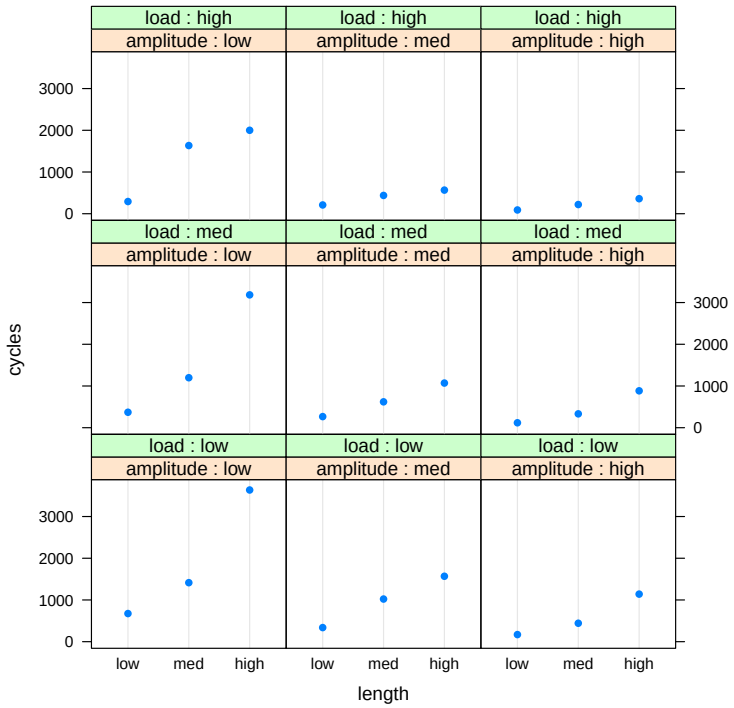


Figure 1: Trellis plot for Question 3.

Which of the following statements is **CORRECT**?

(1) The number of cycles required doesnt depend on load or amplitude

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

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

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

(5) The higher the load, the more cycles are required.

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

(1) The variances are equal.

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

(3) The responses are normally distributed.

(4) The observations are independent.

(5) The errors are uncorrelated.

5. The size of the regression coefficient β corresponding to a variable X in a multiple regression:

   (1) Measures the strength of the relationship between X and the response.

   (2) Measures the correlation between X and the response.

   (3) Measures the increase in the mean response associated with a unit increase in X, with the other variables held constant.

   (4) Measures the importance of X in the regression.

   (5) Measures the increase in the mean response associated with a unit increase in X.

6. The data for this question come from an ecological study in the Galapogos Islands.There are 30 observations corresponding to 30 islands. The variables measured are

   Species : The number of plant species found on the island, the response,

   Elevation : The highest elevation of the island (m),

   Scruz : The distance from Santa Cruz island (km),

   Adjacent : The area of the adjacent island (square km).

   Examine the R output below and then select the most **correct** statement **on the basis of this output only.**

```
Call:
lm(formula = Species ~ Elevation + Scruz + Adjacent, data = gala)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.25183   17.69201   0.806 0.427808
Elevation    0.27444    0.03137   8.749 3.17e-09 ***
Scruz       -0.21784    0.16426  -1.326 0.196320
Adjacent    -0.06744    0.01532  -4.404 0.000162 ***
---
Residual standard error: 60.02 on 26 degrees of freedom
Multiple R-squared: 0.7542,     Adjusted R-squared: 0.7258
F-statistic: 26.59 on 3 and 26 DF,  p-value: 4.393e-08
```

   (1) The estimate of the error variance is 60.02.

   (2) If the variables `Scruz` and `Adjacent` are held constant, the number of species tends to be higher in islands with higher elevation.

   (3) The variable `Scruz` should be kept in the model.

   (4) The variable `Scruz` is not related to the number of species.

   (5) The closer to `Santa Cruz`, the fewer species.

7. In a regression, which of the following statements is **NOT** correct?

   (1) An R^2 of 0.3 means that the model doesnt fit and that the data must be transformed.

   (2) $R^2 = 0$ if all the regression coefficients (except the constant term) are zero.

   (3) The R^2 cannot go down as more variables are added to the model.

   (4) $R^2 = 1$ if and only if all the data lie exactly on a plane.

   (5) The adjusted R^2 adjusts the R^2 downwards to compensate for having many variables in the model.

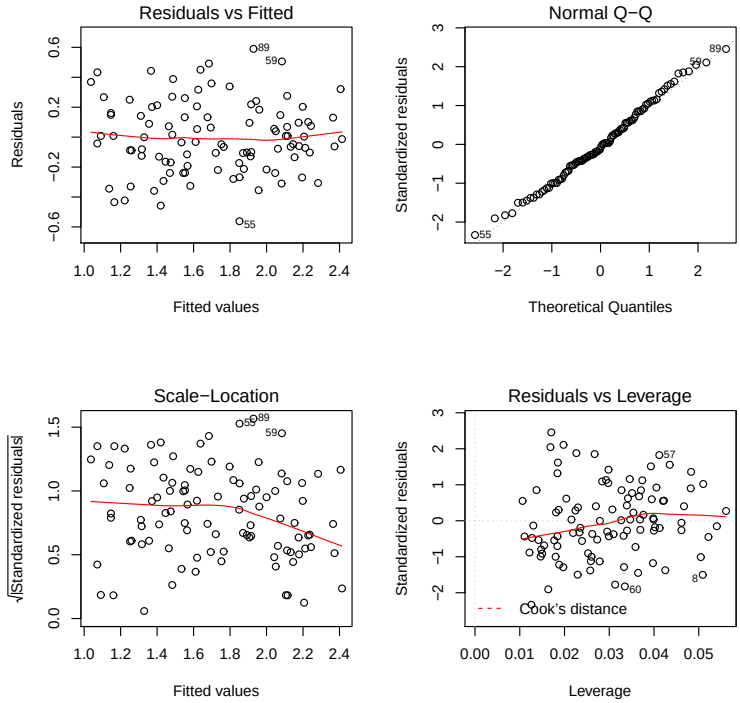


Figure 3: Diagnostic plots for Question 8.

8. Figure 3 contains several diagnostic plots for a fitted regression. What do they indicate about the regression?

   (1) There are several outliers.

   (2) The regression surface is not planar and the independent variables should be transformed.

   (3) The errors are not normal.

   (4) Nothing seems wrong with this regression.

   (5) The variances are not equal.

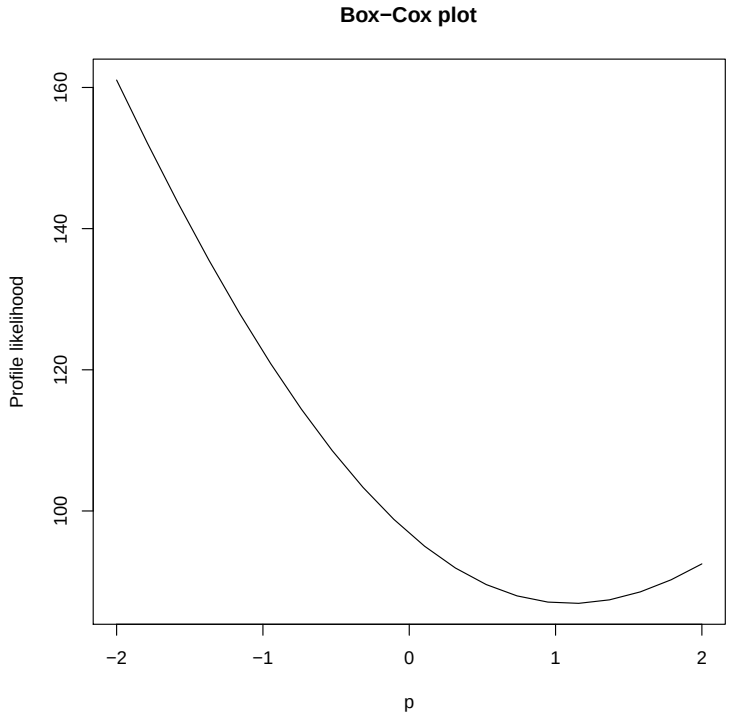


Figure 4: Box-Cox plot for Question 9.

9. In a certain regression, a colleague tells us that transforming the response variable sometimes improves the fit. A Box-Cox plot is shown in Figure 4. What should we do?

   (1) We should transform using a square root.

   (2) We should do nothing, no transformation is indicated.

   (3) We should transform using a log.

   (4) We should transform using a reciprocal.

   (5) We should transform the explanatory variables, not the response.

10. For the Galapogos data, we got the following output. Note that (i) the object `gala.lm` is the result of fitting the model using the `lm function`, and (ii) only the starred parts of the output are shown. The names at the left of the output are the names of the islands.

```
> influence.measures(gala.lm)
Influence measures of
lm(formula = Species ~ Elevation + Scruz + Adjacent, data = gala) :

               dfb.1_  dfb.Elvt dfb.Scrz dfb.Adjc   dffit  cov.r   cook.d    hat inf
Darwin      -0.072719 -0.014078  0.24659 -0.02127  0.2588  2.091 1.74e-02 0.4477   *
Fernandina  -0.096685 -0.012308  0.05945  0.86394  1.0422 17.279 2.82e-01 0.9331   *
Isabela      0.624962 -1.841755  0.14281  0.78764 -1.9425  1.138 8.34e-01 0.4619   *
SantaCruz    0.214357  1.435637 -0.69005 -0.97484  1.8863  0.114 4.98e-01 0.1422   *
Wolf         0.027515 -0.000537 -0.09347  0.01242 -0.0992  1.745 2.56e-03 0.3316   *
```

CONTINUED

Which of the following is **NOT** true?

(1) The observation for Fernandina has a high hat matrix diagonal.

(2) The observation for Isabella is influential because it is having an effect on the coefficient of Elevation.

(3) The observation for Santa Cruz is influential because it is having an effect on the coefficient of Elevation.

(4) The observation for Darwin is influential because it is having an effect on the standard errors.

(5) The observation for Fernandina is having very little effect on the standard errors.

11. In a regression, an explanatory variable X has a variance inflation factor of 100, and a correlation with the response of 0.9, and a p-value of 0.3. Which of the following is **TRUE**?

(1) Since the VIF is high, the variable X is very important and must be included in the regression.

(2) Since the correlation is 0.9, the variable X must be included in the regression.

(3) Since the VIF is high, there must be an outlier present.

(4) Since the VIF is high, X is strongly related to other explanatory variables and doesn't contribute anything extra to the prediction of the response.

(5) Since the p-value is high, the variable X is unrelated to the response.

12. A model was chosen for the Galapogos Island data (see Question 6) using the following R code. (Note that setting `best=2` in this case causes R to print out the best two 1 variable models, and the best two 2 variable models.)

```
> allpossregs(Species~Elevation + Scruz + Adjacent, data=gala, best=2)
       rssp    sigma2 adjRsq     Cp     AIC     BIC       CV Elevation Scruz Adjacent
1 173253.86  6187.638  0.529 22.092  52.092  54.894 23972.22         1     0        0
1 369919.59 13211.414 -0.005 76.682 106.682 109.484 40757.60         0     1        0
2 100003.02  3703.815  0.718  3.759  33.759  37.962 14611.59         1     0        1
2 163527.36  6056.569  0.539 21.392  51.392  55.595 23510.95         1     1        0
3  93667.16  3602.583  0.726  4.000  34.000  39.605 14911.75         1     1        1
```

If we want to use a model for prediction, which model is indicated by this output?

(1) `Species` $\sim$ `Elevation + Adjacent`

(2) `Species` $\sim$ `Scruz`

(3) `Species` $\sim$ `Elevation + Scruz`

(4) `Species` $\sim$ `Elevation`

(5) `Species` $\sim$ `Elevation + Scruz + Adjacent`

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

    (1) If we were building a model for prediction we should always use all the explanatory variables.

    (2) Suppose we are fitting a regression to estimate a particular coefficient. Using the full model is a reasonable thing to do.

    (3) Suppose we are fitting a regression to estimate a particular coefficient. If we have explanatory variables in a regression that are unrelated to the response the estimates of the coefficient of interest will be biased.

    (4) Suppose we are fitting a regression to estimate a particular coefficient. If we have explanatory variables in a regression that are unrelated to the response the standard error of the coefficient of interest will reduced.

    (5) We can eliminate all explanatory variables in a regression whose t-values are less than 2 in absolute value.

14. The data for this question refer to 125 male fruit flies. The response is the longevity of the flies. This is thought to depend on two variables (i) the length of the thorax, and (ii) the opportunity the flies have had to mate.

    The flies have been divided onto 5 groups, according to their opportunity to mate. These are designated "isolated","low","one","many","high". The group membership as been recorded (as a factor `activity` with these levels, baseline "isolated"), as has been the thorax length (variable `thorax`). The response is `longevity` (measured in days).

    The model `longevity` $\sim$ `thorax*activity` was fitted, and the output below obtained.

```
Call:
lm(formula = longevity ~ thorax * activity, data = fruitfly)

                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -50.2420    21.8012  -2.305    0.023 *
thorax              136.1268    25.9517   5.245 7.27e-07 ***
activityone           6.5172    33.8708   0.192    0.848
activitylow          -7.7501    33.9690  -0.228    0.820
activitymany         -1.1394    32.5298  -0.035    0.972
activityhigh        -11.0380    31.2866  -0.353    0.725
thorax:activityone   -4.6771    40.6518  -0.115    0.909
thorax:activitylow    0.8743    40.4253   0.022    0.983
thorax:activitymany   6.5478    39.3600   0.166    0.868
thorax:activityhigh -11.1268    38.1200  -0.292    0.771
---
Residual standard error: 10.71 on 114 degrees of freedom
Multiple R-squared: 0.6534,     Adjusted R-squared: 0.626
F-statistic: 23.88 on 9 and 114 DF,  p-value: < 2.2e-16
```

    Which of the following is **FALSE**?

(1) The model is fitting 5 non-parallel lines.

(2) There is no evidence that the slope of the "high" group is different from the "isolated" group.

(3) The line corresponding to the "isolated" group has slope 136.1268.

(4) The line corresponding to the "one" group has slope -50.2420 + 6.5172.

(5) There is no evidence that the intercept of the "one" group is different from the "isolated" group

15. Some additional output was obtained:

```
> flies.lm = lm(longevity ~ thorax*activity, data=fruitfly)
> flies1.lm = lm(longevity ~ thorax+activity, data=fruitfly)
> flies2.lm = lm(longevity ~ thorax, data=fruitfly)
>
> anova(flies1.lm, flies.lm)
Analysis of Variance Table

Model 1: longevity ~ thorax + activity
Model 2: longevity ~ thorax * activity
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1    118 13107
2    114 13083  4    24.314 0.053 0.9947

> anova(flies2.lm, flies.lm)
Analysis of Variance Table

Model 1: longevity ~ thorax
Model 2: longevity ~ thorax * activity
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    122 22742
2    114 13083  8    9658.9 10.521 5.788e-11 ***
---
> summary(flies1.lm)

Call:
lm(formula = longevity ~ thorax + activity, data = fruitfly)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -48.749     10.850  -4.493 1.65e-05 ***
thorax        134.341     12.731  10.552  < 2e-16 ***
activityone     2.637      2.984   0.884   0.3786
activitylow    -7.015      2.981  -2.353   0.0203 *
activitymany    4.139      3.027   1.367   0.1741
activityhigh  -20.004      3.016  -6.632 1.05e-09 ***
---
```

CONTINUED

```
Residual standard error: 10.54 on 118 degrees of freedom
Multiple R-squared: 0.6527,     Adjusted R-squared: 0.638
F-statistic: 44.36 on 5 and 118 DF,  p-value: < 2.2e-16
```

Which of the following statements is **FALSE**?

(1) The p-value 1.65e-05 is testing the hypothesis that the "isolated" line has zero intercept.

(2) The line

```
thorax         134.341     12.731  10.552  < 2e-16 ***
```

in the output refers to the slope of the lines in the parallel lines model.

(3) The p-value 5.788e-11 is comparing the single line model with the non-parallel line model.

(4) There seems to be no evidence that activity has an effect on longevity.

(5) The p-value 0.9947 is testing the hypothesis that the lines are parallel.