Practice Exam Part B

Example Questions

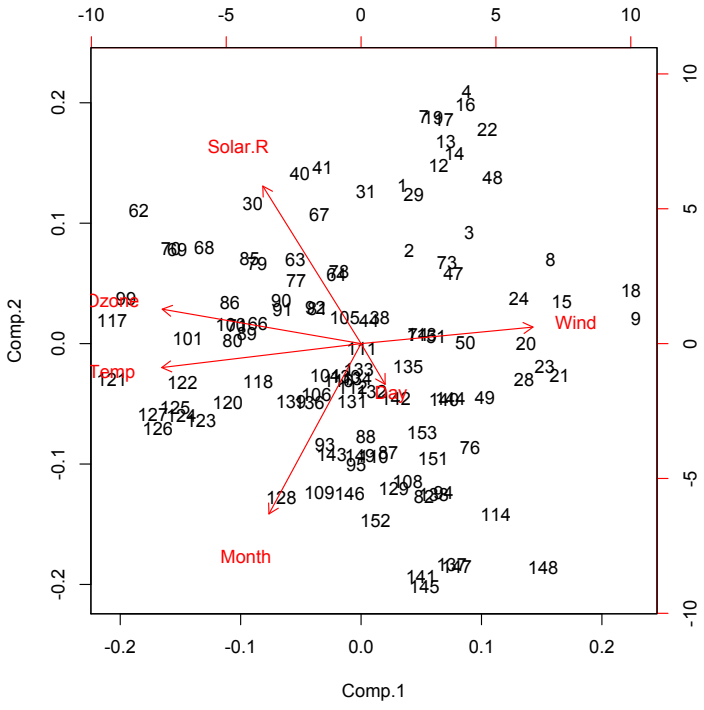
1. Suppose we have a data set called airquality. This data set has 6 variables: Ozone, Solar.R, Wind, Temp, Month, and Day.

   (a) In R how would we get a box plot of ozone?

   (b) Using R how would produce a plot to decide if ozone has a Gaussian distribution?

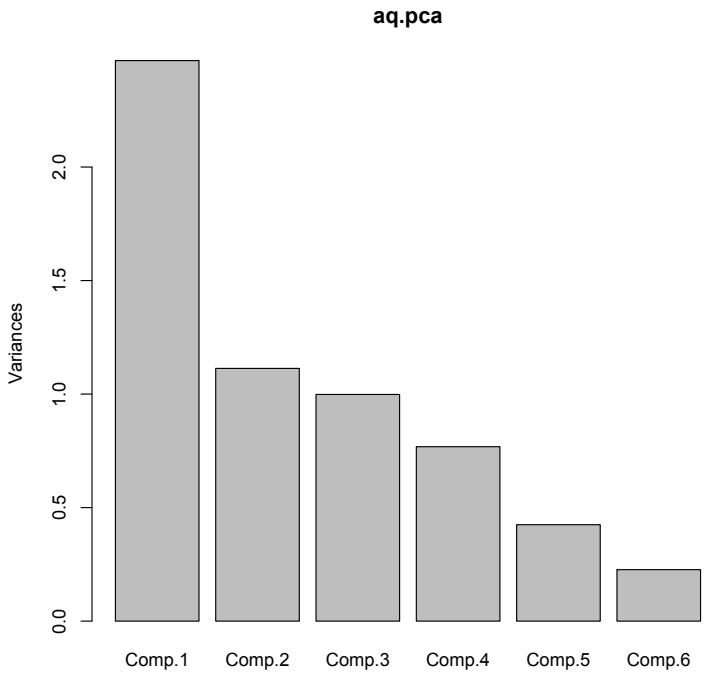
   (c) In R how would we produce a scatter plot matrix of all the variables with histograms on the main diagonal and correlations on the upper triangle?

   (d) Show how we would use R to produce the two dimensional scatter plot of the observations in the first two principal components shown below.



(e) For the scatter plot in question 1d which variables are shown to have high loadings in component 1 but low loadings in component 2?

(f) Suppose the scree plot appears as shown below. Approximately how much of the original variance is visible in the projection into the first two principal components?



(g) How would we use R to produce a main effects linear regression model with ozone as the response and the other variables as predictors?

(h) For the model in question 1g, we run a summary and get the following results:

```
F-statistic: 34.99 on 5 and 105 DF,  p-value: < 2.2e-16
```

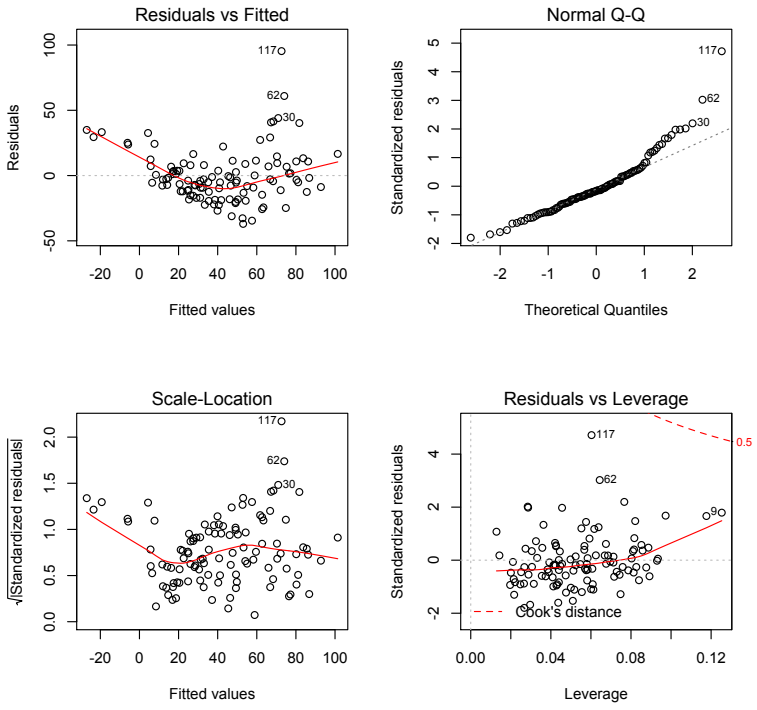
Can we reject the null hypothesis for the model utility, F-test, at a 0.05 level of significance? What does that imply?

(i) The results from the model for question 1g also show the following:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.11632   23.48249  -2.730  0.00742 **
Solar.R       0.05027    0.02342   2.147  0.03411 *
Wind         -3.31844    0.64451  -5.149 1.23e-06 ***
Temp          1.89579    0.27389   6.922 3.66e-10 ***
Month        -3.03996    1.51346  -2.009  0.04714 *
Day           0.27388    0.22967   1.192  0.23576
```

Perform t-tests for variable coefficients. Explain the results of your tests.

(j) Look at the diagnostic plots below for the model from question 1g. Do you detect any problems? If so what are they and what would you do about them?



(k) Now consider a main effects plus interaction terms model for the airquality data. Write the equation for this model.

(l) Write the R code to produce the model for question 1k.

(m) What is the hypothesis for a partial F-test to determine if we should use the interaction or main effects model?

(n) Write the R code to run the test in question 1k.

(o) We obtain the results below for the test in question 1k, where model 1 is the main effects model and model 2 is the interaction model. If we use a significance level of 0.05 what do we conclude?

```
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1    105 45683
2     95 36382 10    9300.6 2.4285 0.01271 *
```

2. The following data were analyzed by a sociologist named Duncan. He used regression analysis to predict the prestige levels of occupations for which the income and educational levels were known but for which there were no direct prestige ratings. The variables are defined as follows:
type: type of occupation - bc (blue collar), wc (white collar), or prof (professional or managerial);
income: percentage of occupational incumbents in the 1950 U.S. Census who earned more than $3500 per year;
education: percentage of occupational incumbents in 1950 who were high-school graduates;
prestige: percentage of respondents in a social survey who rated the occupation as good or better in prestige.
Output from the regression model is as follows:

```
Residuals:
    Min     1Q Median    3Q   Max
 -29.54 -6.417 0.6546 6.605 34.64

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) -6.0647  4.2719    -1.4197  0.1631
     income  0.5987  0.1197     5.0033  0.0000
```

```
  education  0.5458  0.0983     5.5554  0.0000

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-Squared: 0.8282
F-statistic: 101.2 on 2 and 42 degrees of freedom, the p-value is 1.11e-016

Correlation of Coefficients:
          (Intercept)  income
   income -0.2970
education -0.3591     -0.7245
```

(a) Write an equation for predicting prestige using this regression model.

(b) How many observations were in the data set?

(c) What is the residual sum of squares? Give a number.

(d) What percentage of the variance in the response variable were we able to explain with this model?

(e) Now write a model that includes the type variable. Show the R code to run this model.

(f) How would you perform an hypothesis test for the type variable you added to the model in question 2e. How would you run this test in R?

3. You have been asked to help develop postoperative procedures for the orthopedic ward. They interested in the occurrence and treatment of kyphosis. The data for this problem are in the data set kyphosis which has the following variables.

   Kyphosis: a factor telling whether a postoperative deformity (kyphosis) is present or absent.

   Age: the age of the child in months.

   Number: the number of vertebrae involved in the operation.

   Start: the beginning of the range of vertebrae involved in the operation.

   (a) Write the main effects logistic regression model for this problem with the estimated coefficients.

   (b) Write the R code to test to see if this model is significant.

   (c) We get the following results from our test in response to question 3b. What should we conclude?

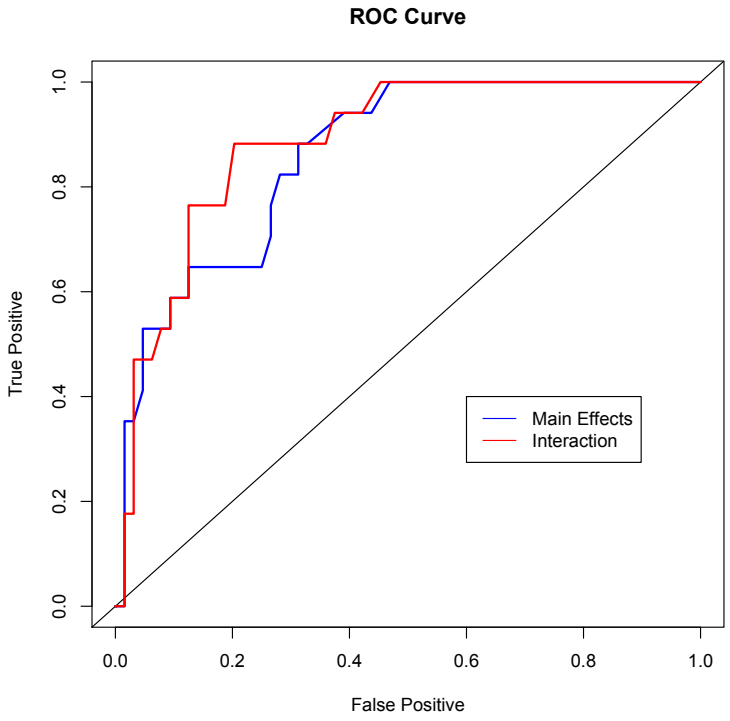
```
1        80     83.234
2        77     61.380  3   21.855 6.994e-05 ***
```

   (d) How do we test to see if interaction terms should be included? Write the test and write the R code for the test.

   (e) The summary for the model for question 3a is shown below. Explain the meaning of the coefficient on Age.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.036934   1.449575  -1.405  0.15996
Age          0.010930   0.006446   1.696  0.08996 .
Number       0.410601   0.224861   1.826  0.06785 .
Start       -0.206510   0.067699  -3.050  0.00229 **
```

   (f) The ROC curves for the training data for the main effects and interaction models are shown below. What do you conclude?



(g) Explain whether the false positive and false negative errors for this problem are of equal importance.