Department of Statistics

COURSE STATS 330/762

Model answers for Assignment 2, 2012.

1. *Load the data into R, and make a data frame **ozone.df** to contain the data. Check for any typographical errors (the data below may be taken to be the correct data, but the data on the web may have been corrupted). Correct as necessary. Print out the last 10 lines of the data file. [5 marks]*

To read in the data, you can use the code

```
ozone.df = read.table(
"http://www.stat.auckland.ac.nz/~lee/330/datasets.dir/ozone.txt")
```

Note that this form reads the data directly from the web page. The data set has sequence numbers so that we omit the `header=TRUE argument from the read.table function.`

A pairs plot shows no obvious outliers. It tedious to check the data frame line by line against the "correct" data on the printed question sheet. Here is a quick way, using R: First, make a text file containing the correct data by cutting and pasting from the pdf. Then, read this into R, making another data frame `correct.ozone.df`. Then type

```
> all(correct.ozone.df==ozone.df)
[1] TRUE
```

The code `correct.ozone.df==ozone.df` makes an array of logicals, with value TRUE if the corresponding elements in the two data frames match, and FALSE otherwise. The function all returns TRUE if all the elements in the array are TRUE, i.e. if all elements in the two data frames match. Thus, all the data seems OK.

To print the last 10 lines, type

```
> ozone.df[102:111,]
    ozone radiation temperature wind
102    16       201          82  8.0
103    13       238          64 12.6
104    23        14          71  9.2
105    36       139          81 10.3
106     7        49          69 10.3
107    14        20          63 16.6
108    30       193          70  6.9
109    14       191          75 14.3
110    18       131          76  8.0
111    20       223          68 11.5
```

[5 marks, 2 for reading in, 2 for checking (OK to do by hand) and 1 for printing.]

2. *Fit a regression model to these data, using **ozone** as the response. Do all variables appear to have an effect on ozone levels? What is the nature of these effects, if any? Give reasons.[5 marks]*

The following code fits the model, and prints out the regression summary:

```
> ozone.lm=lm(ozone~radiation + temperature + wind, data=ozone.df)
> summary(ozone.lm)

Call:
lm(formula = ozone ~ radiation + temperature + wind, data = ozone.df)

Residuals:
    Min      1Q  Median      3Q     Max
-40.485 -14.210  -3.556  10.124  95.600

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.23208   23.04204  -2.788  0.00628 **
radiation     0.05980    0.02318   2.580  0.01124 *
temperature   1.65121    0.25341   6.516 2.43e-09 ***
wind         -3.33760    0.65384  -5.105 1.45e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.17 on 107 degrees of freedom
Multiple R-squared: 0.6062,     Adjusted R-squared: 0.5952
F-statistic: 54.91 on 3 and 107 DF,  p-value: < 2.2e-16
```

All the variables seem significant. Interpreting the signs of the coefficients, it seems that ozone increases as radiation and temperature go up, but decreases as wind speed increases.

Note a shorthand when we are using all the variables in the data frame in the model: typing

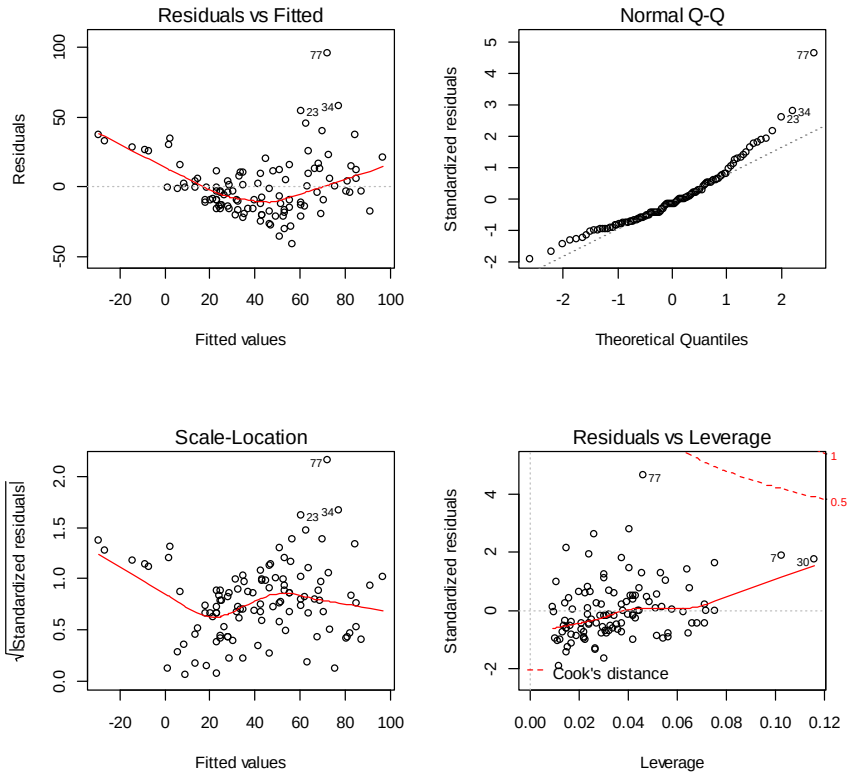
```
ozone.lm=lm(ozone~., data=ozone.df)
```

would fit the same model.

[5 marks: 2 for fitting the model, 3 for interpreting the coefficients]

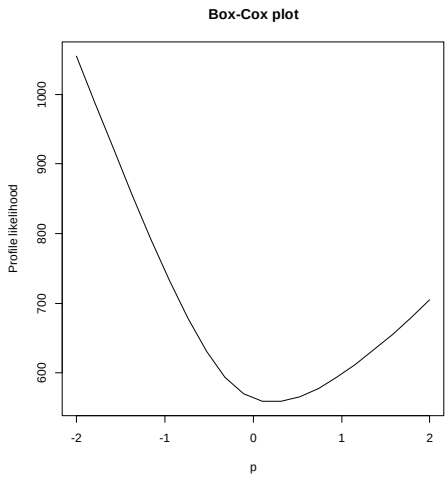
3. *Do you think that these data are suitable to be modeled by a linear regression model? If not, what action should be taken? Is the fit improved if this action is taken? [10 marks]*

Diagnostic plots of the model are shown overleaf:



From the residuals versus fitted values plot we conclude that the regression surface is curved and that there is a possible funnel effect. Some transformation seems required. Lets first try transforming the response with a power. We do a Box-Cox plot to find the correct power:

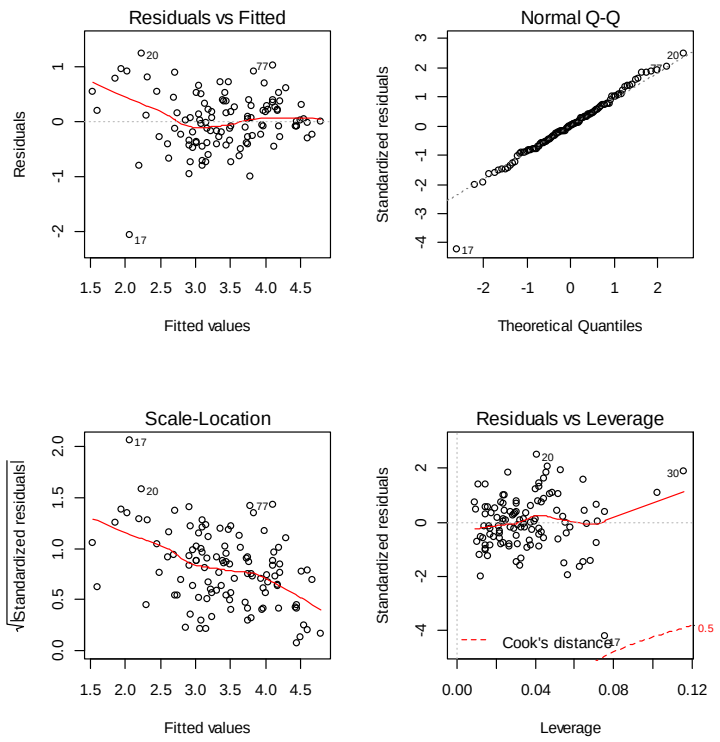
```
> par(mfrow=c(1,1))
> boxcoxplot(ozone.lm)
```



Looks like a log (power=0) is indicated. Let's transform with a log:

```
> log.ozone.lm=lm(log(ozone)~., data=ozone.df)
```

```
> plot(log.ozone.lm)
```



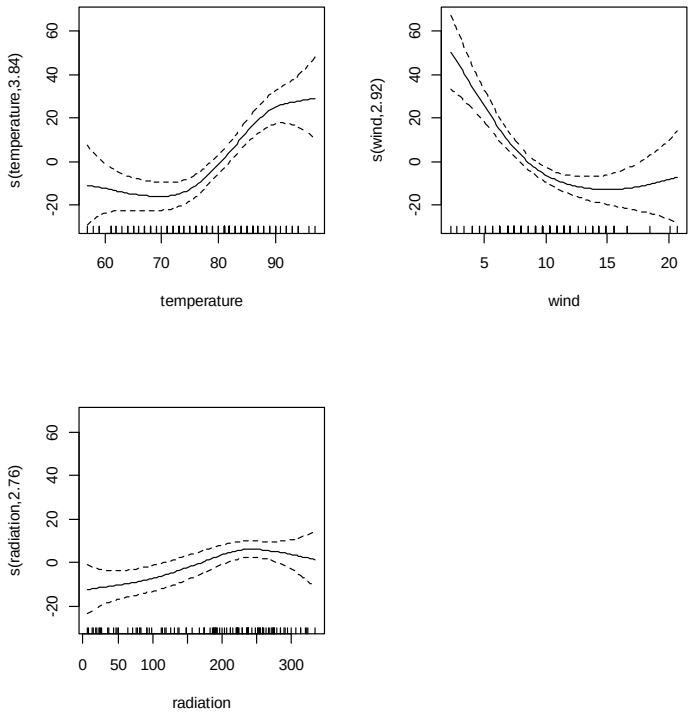
Looks better except for point 17. If we delete this point and refit we get an R^2 of 67% and reasonable residual plots, although there is still the hint of a curve.

```
Call:
lm(formula = log(ozone) ~ ., data = ozone.df, subset = -17)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02028 -0.31515 -0.00931  0.32296  1.12223

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2623617  0.5203253   0.504    0.615
radiation    0.0021899  0.0005155   4.248 4.65e-05 ***
temperature  0.0444490  0.0056771   7.830 3.94e-12 ***
wind        -0.0693092  0.0145043  -4.779 5.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4665 on 106 degrees of freedom
Multiple R-squared: 0.6737,     Adjusted R-squared: 0.6645
F-statistic: 72.95 on 3 and 106 DF,  p-value: < 2.2e-16
```

The other strategy we can follow is to transform the explanatory variables, say with polynomials. First we do a gam plot:

```
> par(mfrow=c(2,2))
> library(mgcv)
```

```
> plot(gam(ozone~s(temperature)+s(wind) + s(radiation),
  data=ozone.df))
```



Possibly temperature and radiation could be modeled by a cubic, and wind by a quadratic. Let's fit cubics to all and see what is significant:

```
  Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
  (Intercept)             42.099      1.694  24.848  < 2e-16 ***
  poly(temperature, 3)1  136.557     23.081   5.916 4.53e-08 ***
  poly(temperature, 3)2   56.869     18.789   3.027 0.003137 **
  poly(temperature, 3)3  -42.023     18.257  -2.302 0.023398 *
  poly(wind, 3)1        -127.418     21.096  -6.040 2.59e-08 ***
  poly(wind, 3)2          82.596     18.989   4.350 3.26e-05 ***
  poly(wind, 3)3         -18.087     18.040  -1.003 0.318448
  poly(radiation, 3)1     70.578     19.092   3.697 0.000355 ***
  poly(radiation, 3)2    -15.871     19.764  -0.803 0.423844
  poly(radiation, 3)3    -31.353     18.412  -1.703 0.091673 .
```

This suggests that radiation need not be transformed, and temperature and wind should be transformed with quadratics. Fitting this model

```
  ozone ~ poly(temperature,2) + poly(wind,2) + radiation
```

gives a R2 of 71.3%, with all coefficients significant. However, the residual plots show evidence of a funnel effect. We attack this by transforming the response. A Box-cox plot indicates a power of ½ i.e. a square root. The residual plots from this fit seem OK. Thus our final model is

```
  sqrt(ozone) ~ poly(temperature,2) + poly(wind,2) + radiation
```

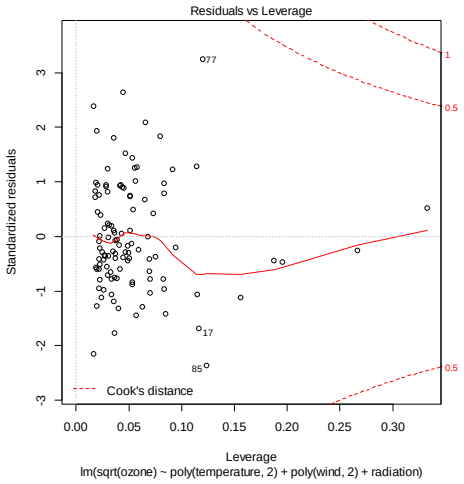
with an R2 of 73.8%:

```
> poly.ozone.lm=lm(sqrt(ozone)~poly(temperature,2) + poly(wind,2) +
+                radiation , data=ozone.df)
> summary(poly.ozone.lm)

Call:
lm(formula = sqrt(ozone) ~ poly(temperature, 2) + poly(wind,
    2) + radiation, data = ozone.df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8220 -0.8041 -0.2668  0.9519  3.8977

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.923949   0.286795  17.169  < 2e-16 ***
poly(temperature, 2)1 12.514332   1.555812   8.044 1.40e-12 ***
poly(temperature, 2)2  3.247048   1.335219   2.432 0.016714 *
poly(wind, 2)1        -8.377982   1.490745  -5.620 1.58e-07 ***
poly(wind, 2)2         5.073921   1.344071   3.775 0.000265 ***
radiation              0.005912   0.001406   4.205 5.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.279 on 105 degrees of freedom
Multiple R-squared: 0.7376,     Adjusted R-squared: 0.7251
F-statistic: 59.03 on 5 and 105 DF,  p-value: < 2.2e-16
```

[10 marks. Give the marks for an attempt to improve the model using the techniques above. The R^2 for the final model should be over 70% - deduct 3 marks if this is not the case.]

4. *Are there any data points that have high leverage or large studentised residuals? If so, are these points having an undue influence on any aspect of the fitted model? Give a full discussion with reasons. [15 marks]*

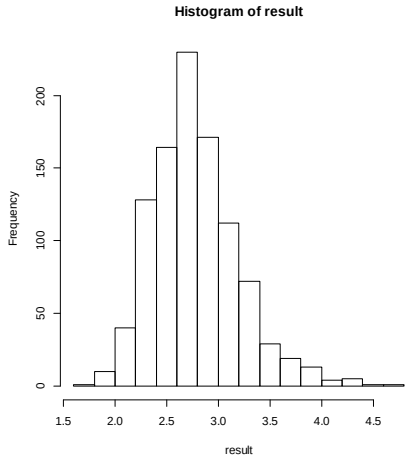
The LR plot for our final model is



The biggest studentised residual is point 77, with a value of 3.4.

```
> rstudent(poly.ozone.lm)[77]
      77
3.406769
```

Is this figure extreme? We can perform a small simulation to find out. Studentised residuals are approximately normally distributed, so the behavior of the largest one should be something like the behavior of the largest value from a normal sample of size 111. We can repeatedly generate say 1000 normal samples and calculate the maximum value (in absolute value) for each sample. We can then draw a histogram of the 1000 maximum values:



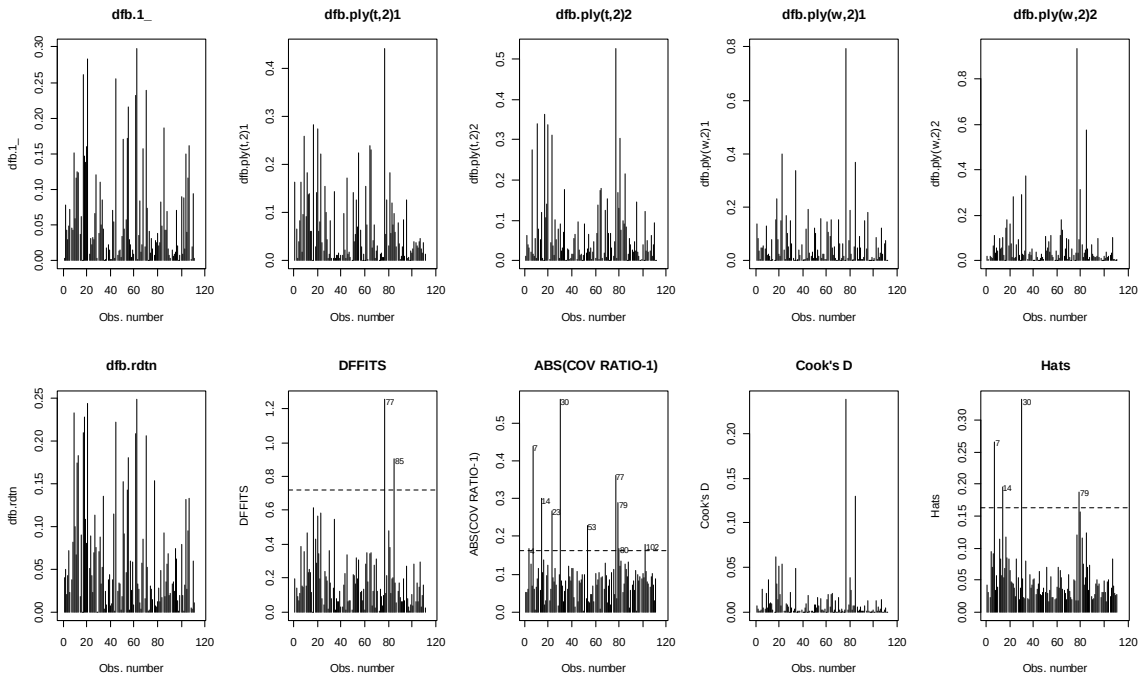

```
Thus, a value of 3.4 doesn't seem too atypical. The R code is
nsim = 1000
result = numeric(nsim)
for(i in 1:nsim)result[i] = max(abs(rnorm(111)))
hist(result)
```

Several points have HMD's over the 3p/n threshold: there are 6 parameters in our model so 3p/n = 18/111 = 1.62. The points having HMDs over 1.62 are

```
> hatvalues(poly.ozone.lm)[hatvalues(poly.ozone.lm)>0.162]
        7        14        30        79
0.2665483 0.1958113 0.3326111 0.1876183
```

However, these do not seem to be associated with large residuals.

To check if any points are having an influence on the fit, we use the function influenceplots:



Seems like some points are having an influence on the fitted values and the standard errors. In particular points 77 85, 7 and 30 seem to be the worst offenders. However, deleting these points and refitting doesn't change the interpretation of the coefficients in any material way.

[15 marks. To get the marks, I want to see an LR plot,(4 marks) and identification of the points having the largest HMD's and residuals. (3 marks each) Also a discussion of the effect points are having on the fit. (5 marks). Actual results will depend on the model fitted. I don't expect anyone to do the simulation, but if any do, give them 5 bonus marks.]

5. Do you think these data have collinearity problems? Give reasons.

This will depend on the fitted model. In terms of the original data (say using the 3 variables untransformed), we get

```
> X = ozone.df[,-1]
> cor(X)
            radiation temperature       wind
radiation   1.0000000   0.2940876 -0.1273656
temperature 0.2940876   1.0000000 -0.4971459
wind       -0.1273656  -0.4971459  1.0000000
```

So, no large correlations. This is clear from a pairs plot as well.

In terms of VIF's:

```
> diag(solve(cor(X)))
  radiation temperature        wind
   1.095241    1.431201    1.328979
```

So, all VIF's are small so no collinearity problems.

[5 marks if VIF's calculated.]