Department of Statistics

Course STATS 330

Model Answer for Assignment 3, 2004

Note: Unlike assignments 1 and 2, your answer for this assignment was expected to be in the form of a report. A sample report follows. Note that there are many possible models that could be fitted. You will get a good mark if you produce a good report, describing how you were led to choose a good model. I make no claims that my model is "best".

Developing a prediction equation for the effect of trace elements on the growth of marsh grass

Report by Alan Lee, XYZ Consultants

Executive Summary

The presence of trace elements is thought to have an effect on the growth of marsh grass. In this report we present two equations for predicting the above-ground biomass of marsh grass, given the amount of trace elements in the soil.

The first has seven variables: Eh7, pH, K, Ca, Mg, Cu and NH4.

An important consideration in the selection of a prediction equation was economy: an equation having a small number of predictor variables might be preferred if the extra variables are expensive to measure. For this reason we also suggest a four variable model with predictors pH, Mg, Cu and NH4.

Introduction

The growth of marsh grass is thought to be affected by the amount of various trace elements in the soil, together with other chemical characteristics of the soil such as Ph (Acidity), salinity (the amount of salt) . This report derives a prediction equation that may be used to predict the above-ground biomass of a plot of land, given the amount of various trace elements in the soil. The data used to develop the prediction equation were obtained from an experiment involving 45 plots of ground. For each plot, the amount of above-ground biomass was measured, together with various chemical characteristics of the soil.

Data

The data consisted of 45 records, one per plot of ground. Each record contained data on the following variables:

Bio:	the above-ground biomass of the marsh grass growing on the plot (grams per square metre);
H2S:	Free sulphide (moles);
Sal:	Salinity (%);
Eh7:	Redox potential at pH 7;
pH:	acidity of water (pH);
BUF:	Buffer acidity at pH 6.6 (meg/100 cm3);
P:	Phosphorus concentration (ppm)
K:	Potassium concentration (ppm)
Ca:	Calcium concentration (ppm)
Mg:	Magnesium concentration (ppm)
Na:	Sodium concentration (ppm)
Mn:	Manganese concentration (ppm)
Zn:	Zinc concentration (ppm)
Cu:	Copper concentration (ppm)
NH4:	Ammonium concentration (ppm)

Analysis

As a first step in the analysis, we conducted a graphical exploration of the data using a series of pairs plots. In addition, a model was fitted to the full data set in order to identify any outliers that might adversely affect the model selection process. The pairs plots were inconclusive but the full-model fit identified point 34 as an outlier. This point was set aside in the variable selection process.

We used the "all possible regressions" to choose a subset. Since the aim is ultimately prediction, we are particularly interested in the Cp criterion (which is based on prediction error), and its close relative AIC. Note that using stepwise regression can't be any better than APR under these circumstances. We are also interested in getting a cheap predictor with a small number of variables, so the BIC criterion, which tends to select simpler models, will also be of interest.

From the APR output (see the technical appendix) the best models are

On the Cp/AIC criterion: the 7 variable model with predictors Eh7, pH, K, Ca, Mg, Cu and Nh4 had the smallest AIC. It did not have the best adjusted R2, but was very close to the best (0.826 vs 0.824)

On the BIC criterion: the 4 variable model with variables pH, Mg, Cu and Nh4.

We will examine both of these models.

Fitting the 7-variable model confirms that point 34 is an outlier, so we set this point aside. The resulting model has an R^2 of 85%, and all variables significant except Ca, with a p-value of 0.08. In view of the difficulty of interpreting p-values after model selection, this variable was retained.

Diagnostic plots reveal a rather pronounced "funnel effect" in the residuals, so it we decided to transform the response. After some experimentation, a power of 0.2 was found to be satisfactory. The residual plots of the transformed model are now reasonably satisfactory, and all the variables are significant. The R2 is 86% (adjusted R^2 is 83%). The model is

Fitting the 4 variable model also indicated that point 34 is an outlier. Refitting the model without this point gave an R2 of 81%, but the residual plots again showed a funnel effect. Box-Cox plots indicated that a square-root transformation of the response might be effective.

The model using the square root of Bio as the response was fitted. The partial reasidual plots and gam plots indicated that the variables pH and Mg needs some transformation. Quadratics for both were tried and the quadratic in pH was retained.

This model (sqrt(Bio)~poly(pH,2) +Mg +Cu+NH4) has an R^2 of 82% and reasonable residual plots, although the plots are not quite as good as those of the 7-variable model (there is still a hint of a funnel effect).

Note that fitting the same model in the form

sqrt(Bio)~pH + I(pH^2) +Mg +Cu+NH4

will make identification of the quadratic coefficients easier.

The choice between these two models is a tradeoff between goodness of fit and economy. The 7 variable model has a slightly better fit (and better residual plots) but the extra expense of measuring the 3 extra variables may outweight this advantage.

Conclusions

Two models are suggested. The first has seven variables, and takes the form

$\text{sqrt(Bio)} \sim b_0 + b_1 \times \text{Eh7} + b_2 \times \text{K} + b_3 \times \text{Ca} + b_4 \times \text{pH} + b_5 \times \text{Mg} + b_6 \times \text{Cu} + b_7 \times \text{NH4}$,
where the coefficients are given by

```
Intercept    4.617e+00
Eh7          3.436e-03
K           -6.813e-04
Ca          -1.660e-04
```

```
pH           3.113e-01
Mg          -2.868e-04
Cu           2.687e-01
NH4         -4.627e-03
```

The second, with four variables, takes the form

$\text{sqrt(Bio)} \sim b_0 + b_1 \times \text{pH} + b_2 \times (\text{pH})^2 + b_3 \times \text{Mg} + b_4 \times \text{Cu} + b_5 \times \text{NH4}$,

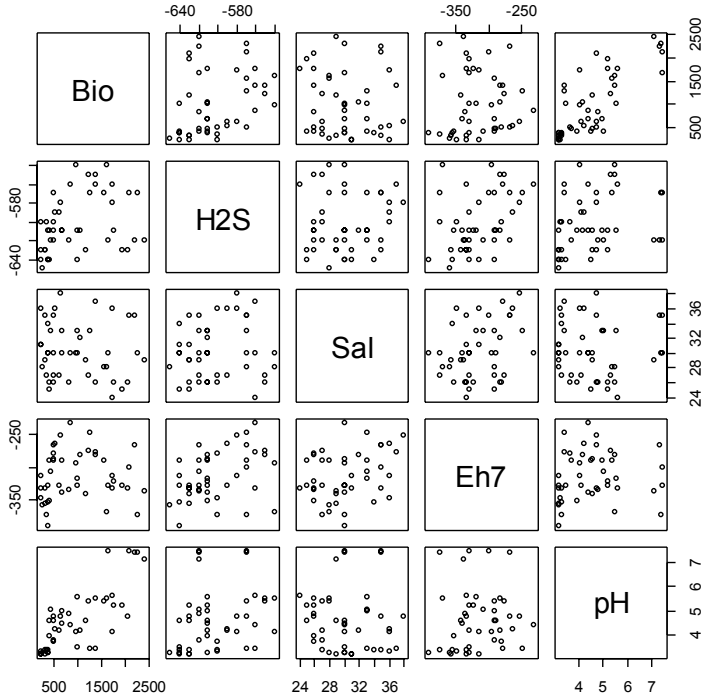
where the coefficients are given by

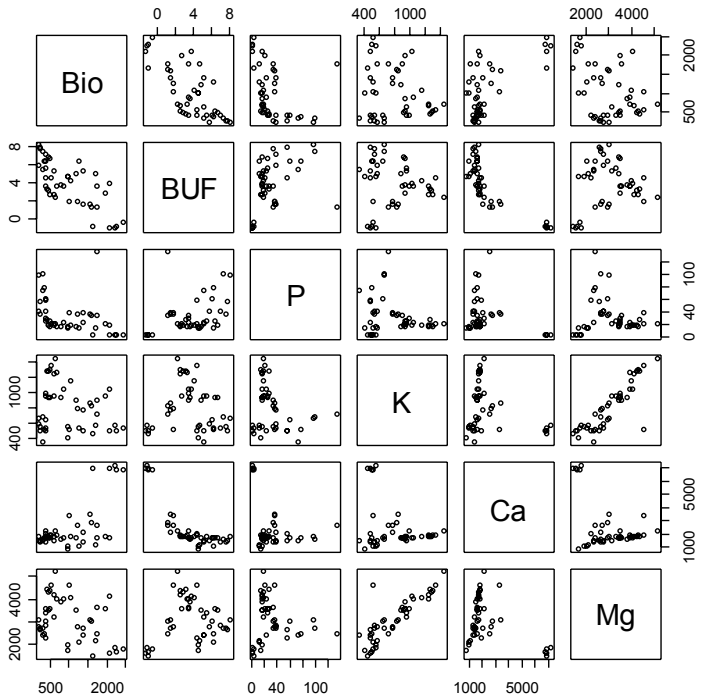
```
(Intercept)  -9.963717
 pH          20.836155
(pH^2)       -1.739093
 Mg          -0.008213
 Cu           3.545878
 NH4         -0.066484
```

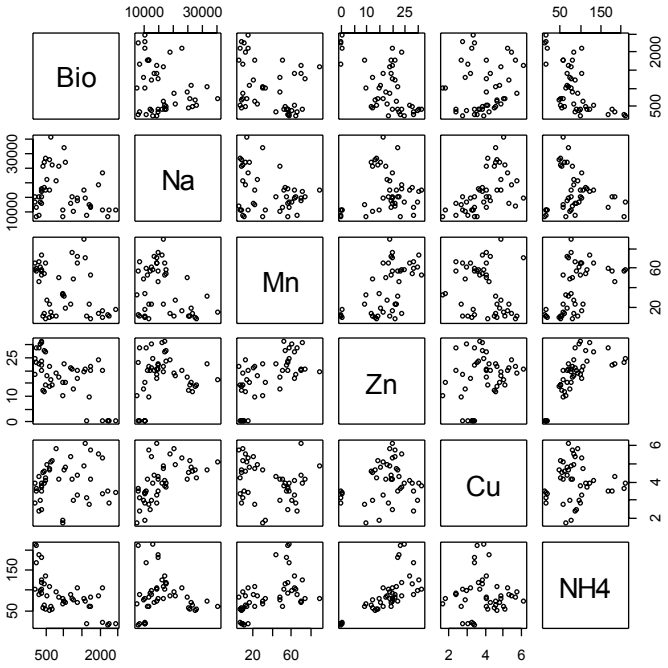
The first form is to be used unless the measurement of the additional variables is too expensive.

Technical Appendix

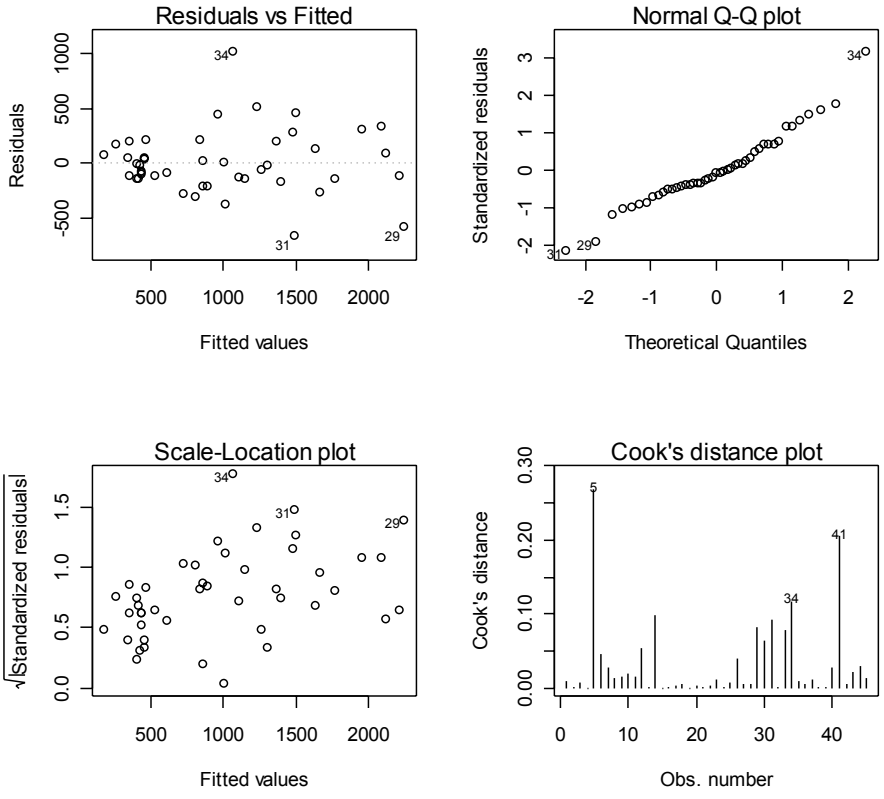
Pairs plots







Diagnostics for full model



APR output

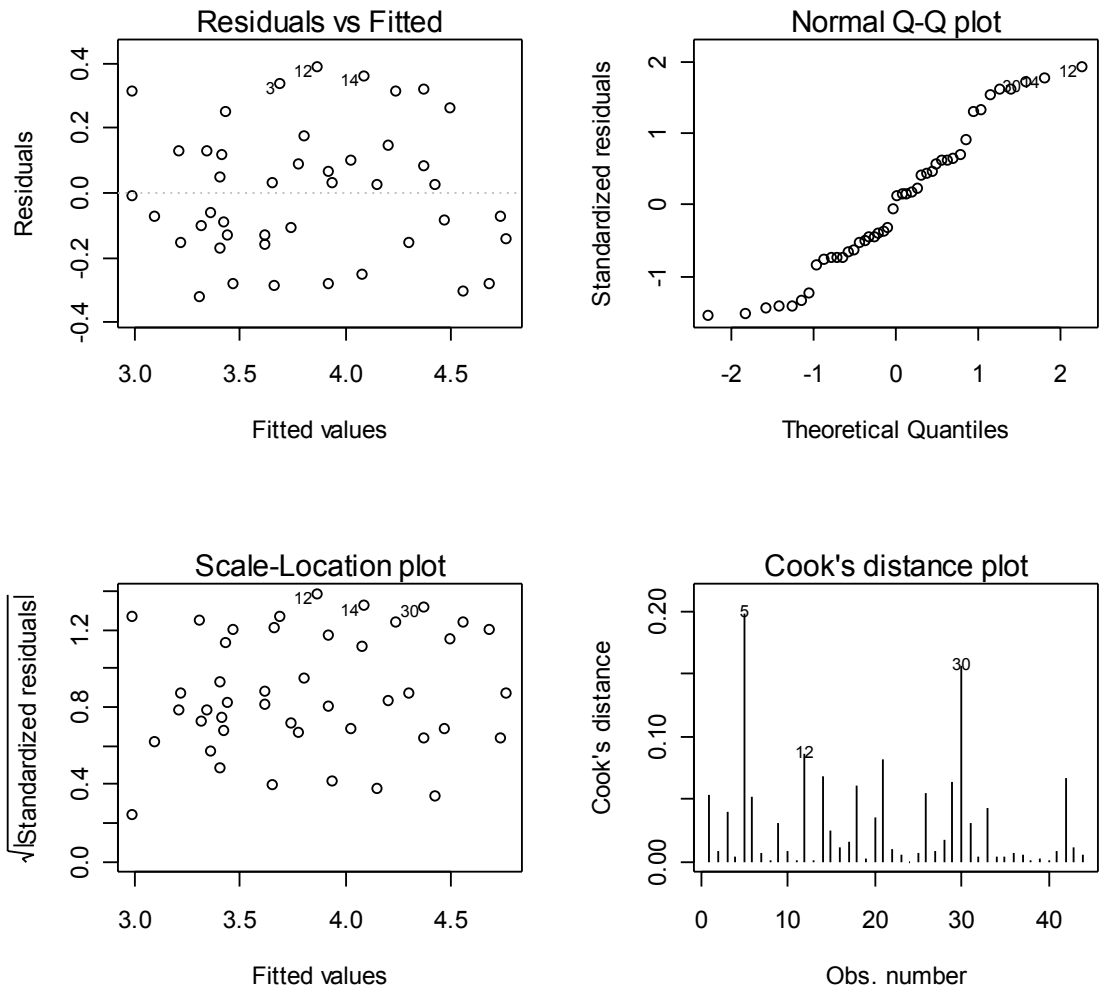
```
> all.poss.regs(stuff.34)
      rssp    sigma2 adjRsq     Cp     AIC     BIC H2S Sal Eh7 pH BUF P K Ca Mg Na Mn Zn Cu NH4
1  6619018 157595.67  0.623 37.272  81.272  84.840   0   0   0  1   0 0 0  0  0  0  0  0  0   0
1  7628782 181637.67  0.566 49.060  93.060  96.629   0   0   0  0   1 0 0  0  0  0  0  0  0   0
1  9653022 229833.86  0.450 72.692 116.692 120.260   0   0   0  0   0 0 0  1  0  0  0  0  0   0
2  5006855 122118.41  0.708 20.451  64.451  69.804   0   0   0  1   0 0 0  0  1  0  0  0  0   0
2  5122657 124942.86  0.701 21.803  65.803  71.156   0   0   0  1   0 0 0  0  0  1  0  0  0   0
2  5192786 126653.31  0.697 22.622  66.622  71.974   0   0   0  0   1 0 0  0  0  1  0  0  0   0
3  3924449  98111.23  0.765  9.815  53.815  60.952   0   0   0  1   0 0 0  1  1  0  0  0  0   0
3  4031739 100793.47  0.759 11.067  55.067  62.204   0   0   0  0   0 0 0  0  1  0  0  0  1   1
3  4127303 103182.58  0.753 12.183  56.183  63.320   0   0   0  0   1 0 0  0  1  0  0  0  1   0
4  3458104  88669.33  0.788  6.371  50.371  59.292   0   0   0  1   0 0 0  0  1  0  0  0  1   1
4  3569952  91537.24  0.781  7.676  51.676  60.597   0   0   0  1   0 0 0  1  1  0  0  0  1   0
4  3674998  94230.72  0.775  8.903  52.903  61.824   0   0   0  0   1 0 0  0  1  0  0  0  1   1
5  3166547  83330.18  0.801  4.967  48.967  59.672   0   0   1  1   0 0 0  0  1  0  0  0  1   1
5  3196821  84126.87  0.799  5.320  49.320  60.026   0   0   0  1   0 0 0  1  1  0  0  0  1   1
5  3303249  86927.60  0.792  6.563  50.563  61.268   0   0   1  1   0 0 0  1  1  0  0  0  1   0
6  2875050  77704.04  0.814  3.564  47.564  60.053   0   0   1  1   0 0 1  0  1  0  0  0  1   1
6  2978536  80500.96  0.807  4.772  48.772  61.261   0   0   1  1   0 0 0  1  1  0  0  0  1   1
6  3000866  81104.48  0.806  5.033  49.033  61.522   0   0   1  0   0 0 1  0  1  0  0  1  1   1
7  2646760  73521.11  0.824  2.899  46.899  61.172   0   0   1  1   0 0 1  1  1  0  0  0  1   1
7  2730651  75851.42  0.819  3.878  47.878  62.152   0   1   1  0   0 0 1  0  1  0  0  1  1   1
7  2802173  77838.14  0.814  4.713  48.713  62.987   0   0   1  1   0 0 1  0  1  0  0  1  1   1
8  2543837  72681.05  0.826  3.697  47.697  63.755   0   0   1  1   0 0 1  1  1  0  0  1  1   1
8  2559153  73118.67  0.825  3.876  47.876  63.934   0   0   1  1   0 1 1  1  1  0  0  0  1   1
8  2592746  74078.46  0.823  4.268  48.268  64.326   0   0   1  1   0 0 1  1  1  0  1  0  1   1
9  2506254  73713.36  0.824  5.259  49.259  67.101   0   0   1  1   0 1 1  1  1  0  0  1  1   1
9  2523082  74208.30  0.823  5.455  49.455  67.297   0   1   1  1   0 0 1  1  1  0  0  1  1   1
9  2524445  74248.38  0.822  5.471  49.471  67.313   0   0   1  1   1 0 1  1  1  0  0  1  1   1
10 2495307  75615.36  0.819  7.131  51.131  70.757   0   1   1  1   0 1 1  1  1  0  0  1  1   1
10 2498119  75700.57  0.819  7.164  51.164  70.790   0   0   1  1   1 1 1  1  1  0  0  1  1   1
10 2505242  75916.44  0.818  7.247  51.247  70.873   1   0   1  1   0 1 1  1  1  0  0  1  1   1
11 2489385  77793.27  0.814  9.062  53.062  74.472   0   1   1  1   1 1 1  1  1  0  0  1  1   1
11 2492201  77881.29  0.814  9.095  53.095  74.505   1   1   1  1   0 1 1  1  1  0  0  1  1   1
11 2494813  77962.92  0.814  9.125  53.125  74.535   0   1   1  1   0 1 1  1  1  0  1  1  1   1
12 2484751  80153.25  0.808 11.008  55.008  78.202   1   1   1  1   1 1 1  1  1  0  0  1  1   1
12 2488136  80262.44  0.808 11.047  55.047  78.242   1   1   1  1   0 1 1  1  1  0  1  1  1   1
12 2488482  80273.62  0.808 11.051  55.051  78.246   0   1   1  1   1 1 1  1  1  1  0  1  1   1
13 2484330  82810.99  0.802 13.003  57.003  81.981   1   1   1  1   1 1 1  1  1  1  0  1  1   1
13 2484424  82814.15  0.802 13.004  57.004  81.982   1   1   1  1   1 1 1  1  1  0  1  1  1   1
13 2487840  82928.00  0.802 13.044  57.044  82.022   0   1   1  1   1 1 1  1  1  1  1  1  1   1
14 2484102  85658.69  0.795 15.000  59.000  85.763   1   1   1  1   1 1 1  1  1  1  1  1  1   1
```

Summary for 7-variable model, after deleting point 34 and transforming the response.

```
lm(formula = I(Bio^0.2) ~ Eh7 + K + Ca + pH + Mg + Cu + NH4,
    data = bio.df, subset = (1:45)[-c(34)])

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.617e+00  5.348e-01   8.632 2.71e-10 ***
Eh7          3.436e-03  1.075e-03   3.196 0.002900 **
K           -6.813e-04  2.632e-04  -2.589 0.013803 *
Ca          -1.660e-04  5.669e-05  -2.929 0.005862 **
pH           3.113e-01  1.001e-01   3.111 0.003639 **
Mg          -2.868e-04  8.492e-05  -3.377 0.001772 **
Cu           2.687e-01  6.473e-02   4.151 0.000194 ***
NH4         -4.627e-03  1.311e-03  -3.529 0.001160 **
---
Residual standard error: 0.2206 on 36 degrees of freedom
Multiple R-Squared: 0.8589,     Adjusted R-squared: 0.8315
F-statistic: 31.31 on 7 and 36 DF,  p-value: 1.788e-13
```

Residual plots for the fitted 7-variable model.



Summary for fitted 4-variable model

```
lm(formula = sqrt(Bio) ~ poly(pH, 2) + Mg + Cu + NH4, data = bio.df,
    subset = (1:45)[-(34)])

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.449935   4.197551  11.066 1.90e-13 ***
poly(pH, 2)1  23.340857   9.057650   2.577  0.01398 *
poly(pH, 2)2 -20.295878   7.851003  -2.585  0.01370 *
Mg            -0.008213   0.001328  -6.183 3.19e-07 ***
Cu             3.545878   1.306650   2.714  0.00995 **
NH4           -0.066484   0.026959  -2.466  0.01829 *
---
Residual standard error: 4.642 on 38 degrees of freedom
Multiple R-Squared: 0.8184,     Adjusted R-squared: 0.7944
F-statistic: 34.24 on 5 and 38 DF,  p-value: 4.345e-13
```

Residual plots for fitted 4-variable model