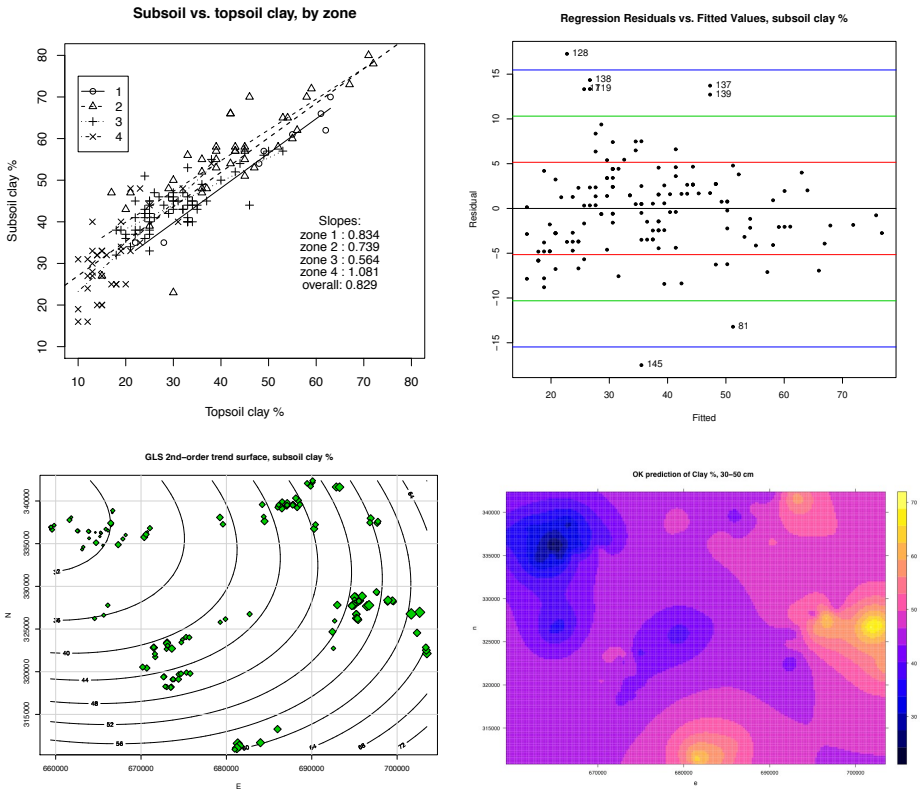
Tutorial:
An example of statistical data analysis using the R environment for statistical computing

D G Rossiter

Version 1.4; May 6, 2017



1 Introduction

This tutorial presents a data analysis sequence which may be applied to environmental datasets, using a small but typical data set of multivariate point observations. It is aimed at students in geo-information application fields who have some experience with basic statistics, but not necessarily with statistical computing. Five aspects are emphasised:

1. Placing statistical analysis in the framework of research questions;
2. Moving from simple to complex methods: first exploration, then selection of promising modelling approaches;
3. Visualising as well as computing;
4. Making correct inferences;
5. Statistical computation and visualization.

The analysis is carried out in the R environment for statistical computing and visualisation [16], which is an open-source dialect of the S statistical computing language. It is free, runs on most computing platforms, and contains contributions from top computational statisticians. If you are unfamiliar with R, see the monograph "Introduction to the R Project for Statistical Computing for use at ITC" [30], the R Project's introduction to R [28], or one of the many tutorials available via the R web page[1].

On-line help is available for all R methods using the `?method` syntax at the command prompt; for example `?lm` opens a window with help for the `lm` (fit linear models) method.

> **Note:** These notes use R rather than one of the many commercial statistics programs because R is a complete *statistical computing environment*, based on a modern computing language (accessible to the user), and with packages contributed by leading computational statisticians. R allows unlimited flexibility and sophistication. "Press the button and fill in the box" is certainly faster – but as with Windows word processors, "what you see is *all* you get". With R it may be a bit harder at first to do simple things, but you are not limited. R is completely free, can be freely-distributed, runs on all desktop computing platforms, is regularly updated, is well-documented both by the developers and users, is the subject of several good statistical computing texts, and has an active user group.

An introductory textbook with similar intent to these notes, but with a wider set of examples, is by Dalgaard [7]. A more advanced text, with many interesting applications, is by Venables and Ripley [35]. Fox [12] is an extensive explanation of regression modelling; the companion Fox and Weisberg [14] shows how to use R for this, mostly with social sciences datasets.

This tutorial follows a data analysis problem typical of earth sciences, natural and water resources, and agriculture, proceeding from visualisation and exploration through univariate point estimation, bivariate correlation and regression analysis, multivariate factor analysis, analysis of variance, and finally some geostatistics.

[1] http://www.r-project.org/

In each section, there are some *tasks*, for which a possible solution is shown as some `R code` to be typed at the console (or cut-and-pasted from the PDF version of this document, or loaded from the accompanying `.R` R code files). Then there are some *questions* to answer, based on the output of the task. Sample *answers* are found at the end of each section.

Optional sections

Some readers may want to skip more advanced sections or those that explain the mathematics behind the methods in more detail; these are marked with an asterisk '*' in the section title and in the table of contents.

Going further

These notes only scratch the surface of R's capabilities. In particular, the reader is encouraged to consult the on-line help as necessary to understand all the options of the methods used. Neither do these notes pretend to teach statistical inference; the reader should refer to a statistics reference as necessary; some good choices, depending on your background and the application, are Brownlee [3], Bulmer [4], Dalgaard [7] (general); Davis [9] (geology),Wilks [39] (meteorology); Snedecor and Cochran [31], Steel et al. [34] (agriculture); Legendre and Legendre [17] (ecology); and Webster and Oliver [38] (soil science).

See also §10, "Going further", at the end of the tutorial.

2 Example Data Set

This data set, fully described in Yemefack [40] and summarized in Yemefack et al. [41], contains 147 soil profile observations from the research area of the Tropenbos Cameroon Programme (TCP), representative of the humid forest region of southwestern Cameroon and adjacent areas of Equatorial Guinea and Gabon.

Three fixed soil layers (0–10 cm, 10–20 cm, and 30–50 cm) were sampled. The data set is from two sources. First, 45 representative soil profiles were described and sampled by genetic horizon. Soil characteristics for each of the three fixed layers were computed as weighted averages using genetic horizon thickness. Second, 102 plots from various land use/land cover types were sampled at the three fixed depths. Each of these samples was a bulked composite of five sub-samples taken with an auger in a plot diagonal basis. For both data sets, samples were located purposively and subjectively to represent soil and land use types. Laboratory analysis was by standard local methods [23].

For this exercise, we have selected three soil properties:

1. Clay content (code `Clay`), weight % of the mineral fine earth (< 2 mm);
2. Cation exchange capacity (code `CEC`), $cmol^{+}$ $(kg\ soil)^{-1}$
3. Organic carbon (code `OC`), volume % of the fine earth.

These three variables are related; in particular we know from theory and many detailed studies that the CEC of a soil depends on reactive sites, either on clay colloids or on organic complexes such as humus, where cations (such as K^{+} and Ca^{++}) can be easily adsorbed and desorbed [22, 32].

> The CEC is important for soil management, since it controls how much added artificial or natural fertiliser or liming materials will be

retained by the soil for a long-lasting effect on crop growth. Heavy doses of fertiliser on soils with low CEC will be wasted, since the extra nutrients will leach.

In addition, for each observation the following site information was recorded:

- East and North Coordinates, UTM Zone 32N, WGS84 datum, in meters (codes `e` and `n`)
- Elevation in meters above sea level (code `elev`)
- Agro-ecological zone, arbitrary code (code `zone`)
- Reference soil group, arbitrary code (code `wrb1`)
- Land cover type (code `LC`)

The soil group codes refer to Reference Groups of the World Reference Base for Soil Resources (WRB) , the international soil classification system [11]. These are presented in the text file as integer codes which correspond to three of the 31 Reference Groups identified worldwide, and which differ substantially in their properties and response to management [10]:

1. Acrisols (from the Haplic, Ferralic, and Plinthic subgroups)
2. Cambisols (from the Ferralic subgroup)
3. Ferralsols (from the Acri-ferric and Xanthic subgroups)

2.1 Loading the dataset

Note: The code in these exercises was tested with Sweave [18, 19] on R version 3.3.2 (2016-10-31), `sp` package Version: 1.2-4, `gstat` package Version: 1.1-5, and `lattice` package Version: 0.20-35 running on Mac OS X 10.6.3. So, the text and graphical output you see here was automatically generated and incorporated into LaTeX by running actual code through R and its packages. Then the LaTeX document was compiled into the PDF version you are now reading. Your output may be slightly different on different versions and on different platforms.

The dataset was originally prepared in a spreadsheet and exported as a text "comma-separated value" (CSV) file named `obs.csv`. This is a typical spreadsheet product with several inadequacies for processing in R, which we will fix up as we go along. This a tedious but necessary step for almost every dataset; so the techniques shown here should be useful in your own projects.

Task 1 : Start the R program and switch to the directory where the dataset is stored. •

Task 3 : Load the dataset into R using the `read.csv` method[2] and examine its structure. Identify each variable from the list above. Note its data type and (if applicable) numerical precision. •

```
> obs <- read.csv("obs.csv")

> str(obs)

'data.frame':        147 obs. of  15 variables:
 $ e    : int  702638 701659 703488 703421 703358 702334 681328 681508 681230 683989
 $ n    : int  326959 326772 322133 322508 322846 324551 311602 311295 311053 311685
 $ elev : int  657 628 840 707 670 780 720 657 600 720 ...
 $ zone : int  2 2 1 1 2 1 1 2 2 1 ...
 $ wrb1 : int  3 3 3 3 3 3 3 3 3 3 ...
 $ LC   : Factor w/ 8 levels "BF","CF","FF",..: 3 3 4 4 4 4 3 3 4 4 ...
 $ Clay1: int  72 71 61 55 47 49 63 59 46 62 ...
 $ Clay2: int  74 75 59 62 56 53 66 66 56 63 ...
 $ Clay5: int  78 80 66 61 53 57 70 72 70 62 ...
 $ CEC1 : num  13.6 12.6 21.7 11.6 14.9 18.2 14.9 14.6 7.9 14.9 ...
 $ CEC2 : num  10.1 8.2 10.2 8.4 9.2 11.6 7.4 7.1 5.7 6.8 ...
 $ CEC5 : num  7.1 7.4 6.6 8 8.5 6.2 5.4 7 4.5 6 ...
 $ OC1  : num  5.5 3.2 6.98 3.19 4.4 5.31 4.55 4.5 2.3 7.34 ...
 $ OC2  : num  3.1 1.7 2.4 1.5 1.2 3.2 2.15 1.42 1.36 2.54 ...
 $ OC5  : num  1.5 1 1.3 1.26 0.8 ...

> row.names(obs)

  [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"
 [11] "11"  "12"  "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"
 [21] "21"  "22"  "23"  "24"  "25"  "26"  "27"  "28"  "29"  "30"
 [31] "31"  "32"  "33"  "34"  "35"  "36"  "37"  "38"  "39"  "40"
 [41] "41"  "42"  "43"  "44"  "45"  "46"  "47"  "48"  "49"  "50"
 [51] "51"  "52"  "53"  "54"  "55"  "56"  "57"  "58"  "59"  "60"
```

[2] a wrapper for the very general `read.table` method

```
  [61] "61"  "62"  "63"  "64"  "65"  "66"  "67"  "68"  "69"  "70"
  [71] "71"  "72"  "73"  "74"  "75"  "76"  "77"  "78"  "79"  "80"
  [81] "81"  "82"  "83"  "84"  "85"  "86"  "87"  "88"  "89"  "90"
  [91] "91"  "92"  "93"  "94"  "95"  "96"  "97"  "98"  "99"  "100"
 [101] "101" "102" "103" "104" "105" "106" "107" "108" "109" "110"
 [111] "111" "112" "113" "114" "115" "116" "117" "118" "119" "120"
 [121] "121" "122" "123" "124" "125" "126" "127" "128" "129" "130"
 [131] "131" "132" "133" "134" "135" "136" "137" "138" "139" "140"
 [141] "141" "142" "143" "144" "145" "146" "147"
```

Each variable has a *name*, which the import method `read.csv` reads from the first line of the CSV file; by default the first field (here, the observation number) is used as the row name (which can be accessed with the `row.names` method) and is not listed as a variable. The suffixes `1`, `2`, and `5` on the variable name roots `Clay`, `CEC`, and `OC` refer to the lower boundary of three depths, in dm; e.g. `OC5` is the organic C content of the 30–50 cm (3–5 dm) layer.

Each variable also has a *data type*. The import method attempts to infer the data type from the format of the data. In this case it correctly found that `LC` is a *factor*, i.e. has fixed set of codes. But it identified `zone` and `wrb1` as integers, when in fact these are coded factors. That is, the 'numbers' 1, 2, ... are just codes. R should be informed of their correct data type, which is important in linear models (§5.5) and analysis of variance (§6). In the case of the soils, we can also change the uninformative integers to more meaningful abbrevations, namely the first letter of the Reference Group name:

```
> obs$zone <- as.factor(obs$zone)
> obs$wrb1 <- factor(obs$wrb1, labels=c("a", "c", "f"))
```

Q3 : *What are the names, data types and numerical precision of the clay contents at the three depths?* *Jump to A3* •

Q4 : *What are the names, data types and numerical precision of the cation exchange capacities at the three depths?* *Jump to A4* •

You can save this as an R data object, so it can be read directly by R (not imported) with the `load` method; this will preserve the corrected data types.

```
> save(obs, file="obs.RData")
```

You can recover this dataset in another R session with the command:

```
> load(file="obs.RData")
```

3 Research questions

A statistical analysis may be *descriptive*, simply reporting, visualizing and summarizing a data set, but usually it is also *inferential*; that is, statistical procedures are used as evidence to answer *research questions*. The most important of these are generally formulated by the researcher before data collection; indeed the sampling plan (number, location, strata, physical size) and data items should be motivated by the research questions. Of course, during field work or analysis other questions may suggest themselves from the data.

The data set for this case study was intended to answer at least the following research questions:

1. What are the *values* of soil properties important for agricultural production and soil ecology in the study area? In particular, the organic matter content (OM), proportion of clay vs. sand and silt (Clay), and the cation exchange capacity (CEC) in the upper 50 cm of the soil.[3]
    - OM promotes good soil structure, easy tillage, rapid infiltration and reduced runoff (hence less soil loss by surface water erosion); it also adsorbs nutrient cations and is a direct source of Nitrogen;
    - The proportion of clay has a major influence on soil structure, hardness, infiltration vs. runoff; almost all the nutrient cations not adsorbed on the OM are exchanged via the clay;
    - CEC is a direct measure of how well the soil can adsorb added cations from burned ash, natural animal and green manures, and artificial fertilizers.
2. What is the *inter-relation* (association, correlation) between these three variables? How much *total information* do they provide?
3. How well can CEC be *predicted* by OM, Clay, or both?
4. What is the depth profile of these variables? Are they constant over the first 50 cm depth; if not, how do they vary with depth?
5. Four agro-ecological zones and three major soil groups have been identified by previous mapping. Do the soil properties differ among these? If so, how much? Can the zones or soils groups be grouped or are they all different?
6. Each observation is located geographically. Is there a *trend* in any of the properties across the region? If so, how much variation does it explain, in which direction is it, and how rapidly does the property vary with distance?
7. Before or after taking any trend into account, is there any *local spatial dependence* in any of the variables?

These statistical question can then be used with knowledge of processes and causes to answer another set of research questions, more closely related to practical concerns or scientific knowledge:

[3] Note that the original data set included many more soil properties.

8. Is it necessary to do the (expensive) lab. procedure for CEC, or can it be predicted satisfactorily from the cheaper determinations for Clay and OM (or just one of these)?
9. Is it necessary to sample at depth, or can the values at depth be calculated from the values in the surface layer? If so, the cost of soil sampling could be greatly reduced.
10. Are the agro-ecological zones and/or soil maps a useful basis for predicting soil behaviour, and therefore a useful stratification for recommendations?
11. What soil-forming factor explains any regional trend?
12. What soil-forming factor explains any local spatial dependence?

Finally, the statistical questions can be used to *predict*:

13. How well can CEC be *predicted* by OM, Clay, or both?
14. What are the *expected values* of the soil properties, and the *uncertainties* of these predictions, at *unvisited locations* in the study area?

The last question can be answered by a predictive *map*.

7 Multivariate correlation and regression

In many datasets we measure several variables. We may ask, first, how are they *inter-related*? This is *multiple correlation analysis*. We may also be interested in *predicting* one variable from several others; this is *multiple regression analysis*.

7.1 Multiple Correlation Analysis

The aim here is to see how a set of variables are *inter-related*. This will be dealt with in a more sophisticated manner in *Principal Components Analysis* (§8.1) and *factor analysis* (§8.2).

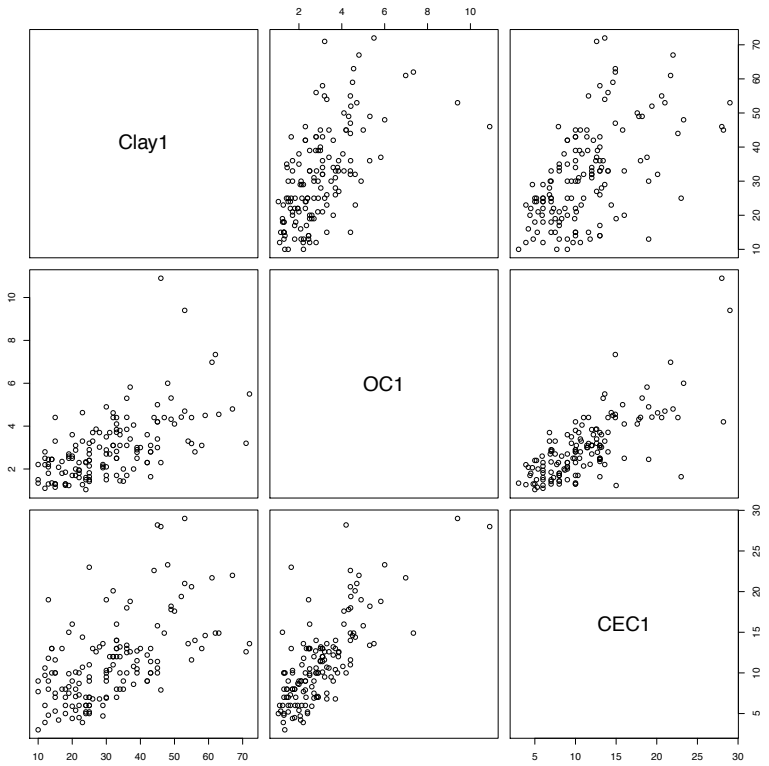
7.1.1 Pairwise simple correlations

For two variables, we used bivariate correlation analysis (§5.3). For more variables, a natural extension is to compute their *pairwise correlations* of all variables.

As explained in the next section, we expect correlations between soil cation exchange capacity (CEC), clay content, and organic carbon content.

Task 41 : Display all the bivariate relations between the three variables CEC, clay content, and organic carbon content of the 0-10cm (topsoil) layer. •

```
> pairs( ~ Clay1 + OC1 + CEC1, data=obs)
```



Q59 : *Describe the relations between the three variables.* *Jump to A59* •

The numeric strength of association is computed as for any pair of variables with a *correlation coefficient* such as Pearson's. Since these only consider two variables at a time, they are called *simple* coefficients.

Task 42 : Compute the covariances and the Pearson's correlation coefficients for all pairs of variables CEC, clay, and OC in the topsoil. •

We first must find the index number of the variables we want to plot, then we present these as a list of indices to the `cov` method:

```
> names(obs)

 [1] "e"     "n"     "elev"  "zone"  "wrb1"  "LC"    "Clay1" "Clay2"
 [9] "Clay5" "CEC1"  "CEC2"  "CEC5"  "OC1"   "OC2"   "OC5"
```

We see the target variables at positions 10, 7 and 13, so:

```
> cov(obs[c(10,7,13)])

          CEC1   Clay1     OC1
CEC1   25.9479  39.609  5.6793
Clay1 39.6092 194.213 12.5021
OC1    5.6793  12.502  2.2520

> cor(obs[c(10,7,13)])

          CEC1   Clay1     OC1
CEC1   1.00000 0.55796 0.74294
Clay1 0.55796 1.00000 0.59780
OC1   0.74294 0.59780 1.00000
```

Q60 : *Explain these in words.* *Jump to A60* •

7.1.2 Pairwise partial correlations

The simple correlations show how two variables are related, but this leaves open the question as to whether there are any underlying relations between the entire set. For example, could an observed strong simple correlation between variables X and Y be because both are in fact correlated to some underlying variable Z? One way to examine this is by *partial correlations*, which show the correlation between two variables after correcting for all others.

What do we mean by "correcting for the others"? This is just the correlation between the residuals of linear regressions between the two variables to be correlated and all the other variables. If the residuals left over after the regression are correlated, this can't be explained by the variables considered so far, so must be a true correlation between the two variables of interest.

For example, consider the relation between `Clay1` and `CEC1`as shown in the scatterplot and by the correlation coefficient ($r = 0.55$). These show a moderate

positive correlation. But, both of these are positively correlated to `OC1` ($r = 0.56$ and 0.74, respectively). Is some of the apparent correlation between clay and CEC actually due to the fact that soils with higher clay tend (in this sample) to have higher OC, and that this higher OC also contributes to CEC? This is answered by the partial correlation between clay and CEC, in both cases correcting for OC.

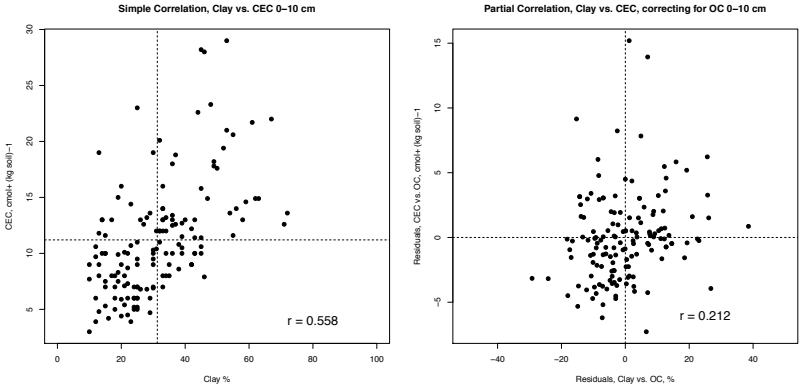
We can compute partial correlations directly from the definition, which is easy in this case with only three variables. We also recompute the simple correlations, computed above but repeated here for comparison. It's not logical (although mathematically possible) to compute the partial correlation of Clay and OC, since the "lurking" variable CEC is a result of these two, not a cause of either. So, we only consider the correlation of CEC with OC and Clay separately.

```
> cor(residuals(lm(CEC1 ~ Clay1)), residuals(lm(OC1 ~ Clay1)))

[1] 0.61538

> cor(residuals(lm(CEC1 ~ OC1)), residuals(lm(Clay1 ~ OC1)))

[1] 0.21214

> cor(CEC1, OC1)

[1] 0.74294

> cor(CEC1, Clay1)

[1] 0.55796
```

This shows that CEC is only weakly positively correlated ($r = 0.21$) to Clay after controlling for OC; compare this to the much higher simple correlation ($r = 0.56$). In other words, much of the apparent correlation between Clay and CEC can be explained by their mutual positive correlation with OC.

We can visualize the reduction in correlation by comparing the scatterplots between Clay and CEC with and without correction for OC:

```
> par(mfrow=c(1,2))
> par(adj=0.5)
> plot(CEC1 ~ Clay1, pch=20, cex=1.5, xlim=c(0,100),
+      xlab="Clay %",
+      ylab="CEC, cmol+ (kg soil)-1")
> abline(h=mean(CEC1), lty=2); abline(v=mean(Clay1), lty=2)
> title("Simple Correlation, Clay vs. CEC 0-10 cm")
> text(80, 4, cex=1.5, paste("r =",round(cor(Clay1, CEC1), 3)))
> mr.1 <- residuals(lm(CEC1 ~ OC1)); mr.2 <-residuals(lm(Clay1 ~ OC1))
> plot(mr.1 ~ mr.2, pch=20, cex=1.5, xlim=c(-50, 50),
+      xlab="Residuals, Clay vs. OC, %",
+      ylab="Residuals, CEC vs. OC, cmol+ (kg soil)-1")
> abline(h=mean(mr.1), lty=2); abline(v=mean(mr.2), lty=2)
> title("Partial Correlation, Clay vs. CEC, correcting for OC 0-10 cm")
> text(25, -6, cex=1.5, paste("r =",round(cor(mr.1, mr.2), 3)))
> par(adj=0)
> rm(mr.1, mr.2)
> par(mfrow=c(1,1))
```



The two scatterplots show that much of the apparent pattern in the simple correlation plot (left) has been removed in the partial correlation plot (right); the points form a more diffuse cloud around the centroid.

By contrast, CEC is highly positively correlated ($r = 0.62$) to OC, even after controlling for Clay (the simple correlation was a bit higher, $r = 0.74$). This suggests that OC should be the best single predictor of CEC in the topsoil; we will verify this in the next section.

The partial correlations are all smaller than the simple ones; this is because all three variables are inter-correlated. Note especially that the correlation between OC and clay remains the highest while the others are considerably diminished; this relation will be highlighted in the principal components analysis.

Simultaneous computation of partial correlations Computing partial correlations from regression residuals gets tedious for a large number of variables. Fortunately, the partial correlation can also be obtained from either the variance–covariance or simple correlation matrix of all the variables by inverting it and then standardising this inverse so that the diagonals are all 1; the off-diagonals are then the negative of the partial correlation coefficients.

Here is a small R function to do this (and give the off-diagonals the correct sign), applied to the three topsoil variables:

```
> p.cor <- function(x){
+         inv <- solve(var(x))
+         sdi <- diag(1/sqrt(diag(inv)))
+         p.cor.mat <- -(sdi %*% inv %*% sdi)
+         diag(p.cor.mat) <- 1
+         rownames(p.cor.mat) <- colnames(p.cor.mat) <- colnames(x)
+         return(p.cor.mat) }
> p.cor(obs[c(10,7,13)])

         CEC1   Clay1     OC1
CEC1  1.00000 0.21214 0.61538
Clay1 0.21214 1.00000 0.32993
OC1   0.61538 0.32993 1.00000
```

7.2 Multiple Regression Analysis

The aim here is to develop the best *predictive equation* for some predictand, given several possible predictors.

In the present example, we know that the CEC depends on reactive sites on clay colloids and humus. So it should be possible to establish a good predictive relation for CEC (the predictand) from one or both of clay and organic carbon (the predictors); we could then use this relation at sites where CEC itself has not been measured.

> Note that the type of clay mineral and, in some cases, the soil reaction are also important in modelling soil CEC; but these are similar in the sample set, so we will not consider them further.

First, we visualise the relation between these to see if the theory seems plausible in this case. This was already done in the previous section, §7.1. We saw that both predictors do indeed have some positive relation with the predictand.

To develop a predictive regression equation, we have three choices of predictors:

- Clay content
- Organic matter content
- Both Clay content and Organic matter content

The simple regressions are computed as before; the *multiple regression* with more than one predictor also uses the `lm` method, with both predictors named in the formula.

Task 43 : Compute the two simple regressions and the one multiple regression and display the summaries. Compare these with the null regression, i.e. where every value is predicted by the mean. •

```
> lmcec.null<-lm(CEC1 ~ 1); summary(lmcec.null)

Call:
lm(formula = CEC1 ~ 1)

Residuals:
   Min     1Q Median     3Q    Max
  -8.2   -3.7   -1.1    1.9   17.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.20       0.42    26.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.09 on 146 degrees of freedom

> lmcec.oc<-lm(CEC1 ~ OC1); summary(lmcec.oc)
```

```
Call:
lm(formula = CEC1 ~ OC1)

Residuals:
   Min     1Q Median     3Q    Max
 -7.28  -2.25  -0.21   1.58  15.19

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.671      0.630    5.82  3.6e-08 ***
OC1            2.522      0.189   13.37  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.42 on 145 degrees of freedom
Multiple R-squared:  0.552,        Adjusted R-squared:  0.549
F-statistic:  179 on 1 and 145 DF,  p-value: <2e-16

> lmcec.clay<-lm(CEC1 ~ Clay1); summary(lmcec.clay)

Call:
lm(formula = CEC1 ~ Clay1)

Residuals:
   Min     1Q Median     3Q    Max
-6.706 -3.351 -0.645  2.201 14.196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8262     0.8620     5.6  1.0e-07 ***
Clay1         0.2039     0.0252     8.1  2.1e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.24 on 145 degrees of freedom
Multiple R-squared:  0.311,        Adjusted R-squared:  0.307
F-statistic: 65.5 on 1 and 145 DF,  p-value: 2.11e-13

> lmcec.oc.cl<-lm(CEC1 ~ OC1 + Clay1); summary(lmcec.oc.cl)

Call:
lm(formula = CEC1 ~ OC1 + Clay1)

Residuals:
   Min     1Q Median     3Q    Max
-7.706 -2.016 -0.377  1.289 15.115

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7196     0.7179    3.79  0.00022 ***
OC1           2.1624     0.2308    9.37  < 2e-16 ***
Clay1         0.0647     0.0249    2.60  0.01015 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.36 on 144 degrees of freedom
```

```
Multiple R-squared:  0.572,        Adjusted R-squared:  0.566
F-statistic: 96.3 on 2 and 144 DF,  p-value: <2e-16
```

Q61 : *How much of the total variability of the predictand (CEC) is explained by each of the models? Give the three predictive equations, rounded to two decimals.* *Jump to A61* •

Q62 : *How much does adding clay to the predictive equation using only organic carbon change the equation? How much more explanation is gained? Does the model summary show this as a statistically-significant increase?* *Jump to A62* •

7.3 Comparing regression models

Which of these models is "best"? The aim is to explain as much of the variation in the dataset as possible with as few predictive factors as possible, i.e. a *parsimonious* model.

7.3.1 Comparing regression models with the adjusted R^2

Compare R^2

One measure which applies to the standard linear model is the "adjusted" R^2 which decreases the apparent R^2, computed from the ANOVA table, to account for the number of predictive factors:

$$R^2_{\text{adj}} \equiv 1 - \left[\frac{(n-1)}{(n-p)} \cdot (1 - R^2) \right]$$

where n is the number of observation and p is the number of coefficients.

Q63 : *What are the adjusted R^2 in the above models? Which one is highest?* *Jump to A63* •

We can see these in the model summaries (above); they can also be extracted from the model summary:

```
> summary(lmcec.null)$adj.r.squared

[1] 0

> summary(lmcec.oc)$adj.r.squared

[1] 0.54887

> summary(lmcec.clay)$adj.r.squared

[1] 0.30657

> summary(lmcec.oc.cl)$adj.r.squared

[1] 0.56618
```

7.3.2 Comparing regression models with the AIC

Compare AIC

A more general measure, which can be applied to almost any model type, is *Akaike's Information Criterion*, abbreviated AIC. The lower value is better.

```
> AIC(lmcec.null); AIC(lmcec.oc); AIC(lmcec.clay); AIC(lmcec.oc.cl)

[1] 898.81

[1] 782.79

[1] 845.98

[1] 778.02
```

Q64 : *Which model is favoured by the AIC?* *Jump to A64* •

7.3.3 Comparing regression models with ANOVA

ANOVA, F-test

A traditional way to evaluate nested models (where one is a more complex version of the other) is to compare them in an ANOVA table, normally with the more complex model listed first. We also compute the proportional reduction in the Residual Sum of Squares (RSS):

```
> (a <- anova(lmcec.oc.cl, lmcec.clay))

Analysis of Variance Table

Model 1: CEC1 ~ OC1 + Clay1
Model 2: CEC1 ~ Clay1
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1    144 1621
2    145 2609 -1      -988 87.8 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> diff(a$RSS)/a$RSS[2]

[1] 0.3787
```

The ANOVA table shows that the second model (clay only) has one more degree of freedom (i.e. one fewer predictor), but a much higher RSS (i.e. the variability not explained by the model); the reduction is about 38% compared to the simpler model. These two estimates of residual variance can be compared with an F-test. In this case the probability that they are equal is approximately zero, so it's clear the more complex model is justified (adds information).

However, when we compare the combined model with the prediction from organic matter only, we see a different result:

```
> (a <- anova(lmcec.oc.cl, lmcec.oc))

Analysis of Variance Table
```

```
Model 1: CEC1 ~ OC1 + Clay1
Model 2: CEC1 ~ OC1
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1    144 1621
2    145 1697 -1     -76.4 6.79   0.01 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> diff(a$RSS)/a$RSS[2]

[1] 0.045004
```

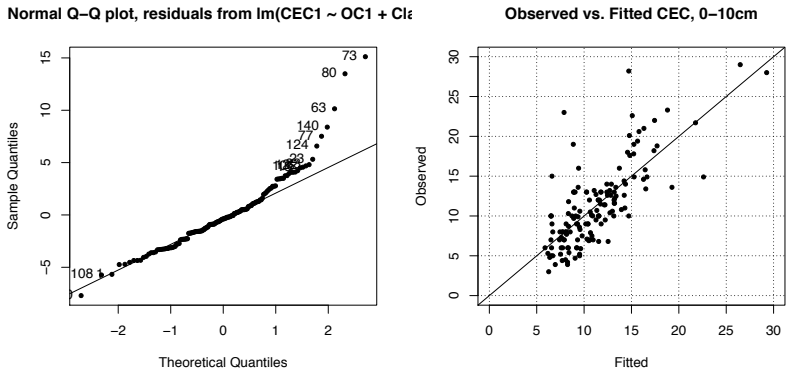
Q65 : *Which model has a lower RSS? What is the absolute and proportional difference in RSS between the combined and simple model? What is the probability that this difference is due to chance, i.e. that the extra information from the clay content does not really improve the model?* *Jump to A65* •

Regression diagnostics

Before accepting a model, we should review its diagnostics (§5.7). This provides insight into how well the model fits, and where any lack of fit comes from.

Task 44 : Display two diagnostic plots for the best model: (1) a normal quantile-quantile ("Q-Q") plot of the residuals. Identify badly-fitted observations and examine the relevant fields in the dataset, (1) predicted vs. actual topsoil CEC. •

```
> par(mfrow=c(1,2))
> tmp <- qqnorm(residuals(lmcec.oc.cl), pch=20,
+      main="Normal Q-Q plot, residuals from lm(CEC1 ~ OC1 + Clay1)")
> qqline(residuals(lmcec.oc.cl))
> diff <- (tmp$x - tmp$y)
> ### label the residuals that deviate too far from the line
> text(tmp$x, tmp$y, ifelse((abs(diff) > 3), names(diff), ""), pos=2)
> rm(tmp,diff)
> ### observed vs. fitted
> #
> plot(CEC1 ~ fitted(lmcec.oc.cl), pch=20,
+    xlim=c(0,30), ylim=c(0,30),
+    xlab="Fitted",ylab="Observed",
+    main="Observed vs. Fitted CEC, 0-10cm")
> abline(0,1); grid(col="black")
> par(mfrow=c(1,1))
```



Q66 : *Are the residuals normally distributed? Is there any apparent explanation for these poorly-modelled observations?* *Jump to A66* •

7.4 Stepwise multiple regression*

In the previous section, we examined several models individually, using our expert judgement to decide which predictors to use, and in which order. Another approach is to let R try out a large number of possible equations and select the "best" according to some criterion. One method for this is *stepwise* regression, using the `step` method.

The basic idea of `step` is to specify an initial model object, as with `lm`, and then a *scope* which specifies how variables in the full model should be added or subtracted; in the simplest case we do not specify a scope and `step` tries to eliminate all variables, one at a time, until no more can be eliminated without increasing the AIC, explained above.

We will illustrate this with the problem of predicting subsoil clay (difficult to sample) from the three topsoil parameters.

Task 45 : Set up a model to predict subsoil clay from all three topsoil variables (clay, OM, and CEC) and use `step` to see if all three are needed. •

```
> # let stepwise pick the best from a full model
> lms <- step(lm(Clay2 ~ Clay1 + CEC1 + OC1))

Start:  AIC=461.91
Clay2 ~ Clay1 + CEC1 + OC1

        Df Sum of Sq   RSS AIC
<none>                3224 462
- OC1    1        81  3305 464
- CEC1   1       179  3403 468
- Clay1  1     21078 24301 757
```

In this case we see that the full model has the best AIC (461.91) and removing any of the factors increases the AIC, i.e. the model is not as good. However,

removing either `OC1` or `CEC1` doesn't increase the AIC very much (only to 468), so although statistically valid they are not so useful.

An example with more predictors shows how variables are eliminated.

Task 46 : Set up a model to predict CEC in the 30-50 cm layer from all three variables (clay, OM, and CEC) for the two shallower layers, and use `step` to see if all six are needed. Note: this model could be applied if only the first two soil layers were sampled, and we wanted to predict the CEC value of the third layer. •

```
> lms <- step(lm(Clay5 ~ Clay1 + CEC1 + OC1 + Clay2 + CEC2 + OC2, data=obs))

Start:  AIC=420.7
Clay5 ~ Clay1 + CEC1 + OC1 + Clay2 + CEC2 + OC2

        Df Sum of Sq  RSS AIC
- CEC1   1         1 2339 419
- OC1    1         9 2347 419
- OC2    1        12 2350 419
- Clay1  1        27 2365 420
<none>               2338 421
- CEC2   1        48 2387 422
- Clay2  1      1764 4102 501

Step:  AIC=418.75
Clay5 ~ Clay1 + OC1 + Clay2 + CEC2 + OC2

        Df Sum of Sq  RSS AIC
- OC1    1        11 2350 417
- OC2    1        12 2350 417
- Clay1  1        31 2370 419
<none>               2339 419
- CEC2   1        76 2415 421
- Clay2  1      1966 4305 506

Step:  AIC=417.43
Clay5 ~ Clay1 + Clay2 + CEC2 + OC2

        Df Sum of Sq  RSS AIC
- OC2    1         5 2355 416
- Clay1  1        21 2371 417
<none>               2350 417
- CEC2   1        67 2417 420
- Clay2  1      2294 4644 516

Step:  AIC=415.77
Clay5 ~ Clay1 + Clay2 + CEC2

        Df Sum of Sq  RSS AIC
<none>               2355 416
- Clay1  1        36 2392 416
- CEC2   1        62 2417 418
- Clay2  1      2311 4666 514
```

The original AIC (with all six predictors) is 420.7; `step` examines all the variables and decides that by eliminating `CEC1` (a topsoil property) the AIC is most improved.

The AIC is now 418.75; `step` examines all the remaining variables and decides that by eliminating `OC1` the AIC is most improved; again a topsoil property is considered unimportant.

The AIC is now 417.43; `step` examines all the remaining variables and decides that by eliminating `OC2` the AIC is most improved.

The AIC is now 415.77 and all three remaining variables must be retained, otherwise the AIC increases. The final selection includes both clay measurements (0-10 and 10-20 cm) and the CEC of the second layer.

Notice from the final output that `Clay1` could still be eliminated with very little loss of information, which would leave a model with two properties from the second layer to predict the clay in the subsoil; or `CEC2` could be eliminated with a little more loss of information; this would leave the two overlying clay contents to predict subsoil clay. Either of these alternatives would be more parsimonious in terms of interpretation, although statistically just a bit weaker than the final model discovered by `step`.

7.5 Combining discrete and continuous predictors

In many datasets, including this one, we have both *discrete factors* (e.g. soil type, agro-ecological zone) and *continuous variables* (e.g. topsoil clay) which we show in one-way ANOVA and univariate regression, respectively, to be useful predictors of some continuous variable (e.g. subsoil clay). The discussion of the design matrix and linear models (§6.3) showed that both one-way ANOVA on a factor and univariate regression on a continuous predictor are just a cases of linear modelling. Thus, they can be combined in a multiple regression.

Task 47 : Model the clay content of the 20-50 cm layer from the agro-ecological zone and measured clay in the topsoil (0-10 cm layer), first separately and then as an additive model. •

```
> lm5z <- lm(Clay5 ~ zone);  summary(lm5z)

Call:
lm(formula = Clay5 ~ zone)

Residuals:
   Min     1Q Median     3Q    Max
-32.95  -5.40   0.16   3.16  24.05

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    55.00       3.21   17.14  < 2e-16 ***
zone2           0.95       3.52    0.27   0.7874
zone3         -11.16       3.41   -3.28   0.0013 **
zone4         -23.67       3.55   -6.67  5.2e-10 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.08 on 143 degrees of freedom
Multiple R-squared:  0.513,        Adjusted R-squared:  0.502
F-statistic: 50.1 on 3 and 143 DF,  p-value: <2e-16

> lm51 <- lm(Clay5 ~ Clay1); summary(lm51)

Call:
lm(formula = Clay5 ~ Clay1)

Residuals:
    Min      1Q  Median      3Q     Max
-20.626  -3.191   0.005   3.387  14.150

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.7586     1.1556    16.2   <2e-16 ***
Clay1         0.8289     0.0338    24.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.69 on 145 degrees of freedom
Multiple R-squared:  0.806,        Adjusted R-squared:  0.805
F-statistic:  602 on 1 and 145 DF,  p-value: <2e-16

> lm5z1 <- lm(Clay5 ~ zone + Clay1);  summary(lm5z1)

Call:
lm(formula = Clay5 ~ zone + Clay1)

Residuals:
   Min     1Q Median     3Q    Max
-24.09  -2.99   0.15   3.14  13.89

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.3244     2.9054    6.65  5.8e-10 ***
zone2         5.6945     2.1060    2.70   0.0077 **
zone3         2.2510     2.1831    1.03   0.3043
zone4        -0.6594     2.5365   -0.26   0.7953
Clay1         0.7356     0.0452   16.26  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.39 on 142 degrees of freedom
Multiple R-squared:  0.83,        Adjusted R-squared:  0.825
F-statistic:  173 on 4 and 142 DF,  p-value: <2e-16
```

Note the use of the + in the model specification. This specifies an *additive* model, where there is one regression line (for the *continuous* predictor) which is displaced vertically according to the mean value of the *discrete* predictor. This is sometimes called *parallel regression*. It hypothesizes that the only effect of the discrete predictor is to adjust the mean, but that the relation between the contin-

uous predictor and the predictand is then the same for all classes of the discrete predictor. Below (§7.8) we will investigate the case where we can not assume parallel slopes.

Q67 : *How much of the variation in subsoil clay is explained by the zone? by the topsoil clay? by both together? Is the combined model better than individual models? How much so?* *Jump to A67* •

Q68 : *In the parallel regression model (topsoil clay and zone as predictors), what are the differences in the means between zones? What is the slope of the linear regression, after accounting for the zones? How does this compare with the slope of the linear regression not considering zones?* *Jump to A68* •

Q69 : *Are all predictors in the combined model (topsoil clay and zone as predictors) asignificant? (Hint: look at the probability of the* t*-tests.)* *Jump to A69* •

Diagnostics We examine the residuals to see if any points were especially badly-predicted and if the residuals fit the hypothesis of normality.

Task 48 : Make a stem plot of the residuals. •

```
> stem(residuals(lm5z1))

  The decimal point is at the |

  -24 | 1
  -22 |
  -20 |
  -18 |
  -16 |
  -14 |
  -12 |
  -10 | 540
   -8 | 77104
   -6 | 10099662
   -4 | 888539854322
   -2 | 8655321009876110
   -0 | 986665432211098766655532 1
    0 | 00122334445679023444466688889
    2 | 0334488900122333345568
    4 | 0336800058
    6 | 35792244
    8 | 5
   10 | 11188
   12 | 49
```

Q70 : *Are the residuals normally-distributed? Are there any particularly bad values?* *Jump to A70* •

Clearly there are some points that are less well-modelled.

Task 49 : Display the records for these poorly-modelled points and compare their subsoil clay to the prediction. •

```
> res.lo <- which(residuals(lm5z1) < -12)
> res.hi <- which(residuals(lm5z1) > 9)
> obs[res.lo, ]

         e      n elev zone wrb1  LC Clay1 Clay2 Clay5 CEC1 CEC2 CEC5
145 695098 328237  547    2    f OCA    30    18    23    7    6    7
    OC1 OC2 OC5
145 1.5 0.8 0.8

> predict(lm5z1)[res.lo]

   145
47.086

> obs[res.hi, ]

         e      n elev zone wrb1  LC Clay1 Clay2 Clay5 CEC1 CEC2 CEC5
9   681230 311053  600    2    f  FV    46    56    70  7.9  5.7  4.5
27  679242 338073  360    3    a  FV    24    35    51  5.0  5.4 13.1
38  671039 336819  130    4    a OCA    13    23    40  4.8  3.4  3.2
42  667325 334883  243    4    a  FV    23    38    48  3.9  4.2  4.9
119 666452 337405  134    4    a  BF    21    40    48  5.4  2.6  7.5
128 699567 328185  630    2    f MCA    17    40    47  8.0  8.0  8.0
137 698928 328368  640    2    f  FV    42    61    66  9.0  9.0  8.0
139 695014 328757  560    2    f  FV    42    60    66  9.0  8.0  8.0
     OC1  OC2 OC5
9   2.30 1.36 0.9
27  1.04 0.52 0.5
38  1.30 0.34 0.2
42  1.27 0.58 0.5
119 2.00 0.60 0.4
128 1.80 0.90 0.8
137 2.30 1.30 1.0
139 2.30 1.20 1.0

> predict(lm5z1)[res.hi]

     9     27     38     42    119    128    137    139
58.856 39.229 28.228 35.583 34.112 37.524 55.913 55.913
```

Q71 : *What are the predicted and actual subsoil clay contents for the highest and lowest residuals? What is unusual about these observations?* *Jump to A71* •

7.6 Diagnosing multi-colinearity

Another approach to reducing a regression equation to its most parsiminious form is to examine the relation between the predictor variables and the predictand for *multi-collinearity*, that is, the degree to which they are themselves linearly related in the multiple regression. In the extreme, clearly if two variables are perfectly related, one can be eliminated, as it can not add information as a predictor.

This was discussed to some extent in §7.1 "Multiple correlation", but it was not clear which of the correlated variables to discard, because the predictand was not included in the analysis. For this we use the **Variance Inflation Factor** (VIF), which measures the effect of a set of explanatory variables (predictors) on the *variance* of the coefficient of another predictor, in the multiple regression equation including all predictors, i.e. how much the variance of an estimated regression coefficient is increased because of collinearity. The square root of the VIF gives the increase in the standard error of the coefficient in the full model, compared with what it would be if the target predictor were uncorrelated with the other predictors. Fox [12] has a good discussion, including a visualization.

In the standard multivariate regression:

$$Y = \sum_{0}^{k} \beta_k X_k + \varepsilon,\ X_0 = 1 \tag{5}$$

solved by ordinary least-squares, the sampling variance of an estimated regression coefficient $\hat{\beta}_j$ can be expressed as:

$$\mathrm{var}(\hat{\beta}_j) = \frac{s^2}{(n-1)s_j^2} \cdot \frac{1}{1-R_j^2} \tag{6}$$

where:

s^2 : is the estimated error variance of the residuals of the multiple regression;

s_j^2 : is the sample variance of the target variable;

R_j^2 : is the multiple coefficient of determination for the regression of the target variable X_j on the other predictors.

The **left-hand multiplicand** applies also in a single-predictor regression: it measures the imprecision of the fit compared to that of the predictor. A larger overall error variance of the regression, s^2, will, of course, always lead to a higher variance in the regression coefficient, while a larger number of observations n and a larger variance s_j^2 of the target variable will both lower the variance in the regression coefficient.

The **right-hand multiplicand**, $1/(1-R_j^2)$ applies only in multiple regression. This is the VIF: it multiplies the variance of the regression coefficient by a factor that will be larger as the multiple correlation of a target predictor with the other predictors increases. Thus the VIF increases as the target predictor does not add much information to the regression.

The VIF is computed with the `vif` function of John Fox's `car` package [13].

Task 50 : Load the `car` package and compute the VIF of the six predictors. •

```
> require(car)
> vif(lm(Clay5 ~ Clay1 + CEC1 + OC1 + Clay2 + CEC2 + OC2, data=obs))

   Clay1    CEC1     OC1   Clay2    CEC2     OC2
12.8391  4.7712  4.0944 10.3882  3.5531  3.0349
```

There is no test of significance or hard-and-fast rule for the VIF: however many authors consider $VIF \geq 5$ as a caution and $VIF \geq 10$ as a definite indication of multicolinearity. Note that this test does not tell *which* variables, of the set, each variable with a high VIF is correlated with. It could be with just one or with several taken together.

Q72 : *According to the* VIF ≥ 10 *criterion, which variables are highly correlated with the others?* *Jump to A72* •

Task 51 : Re-compute the VIF for the multiple regression without these variables, each taken out separately. •

```
> vif(lm(Clay5 ~ Clay1 + CEC1 + OC1 + CEC2 + OC2, data=obs))

 Clay1   CEC1    OC1   CEC2    OC2
2.5927 4.2208 4.0916 3.3218 3.0214

> vif(lm(Clay5 ~ Clay2 + CEC1 + OC1 + CEC2 + OC2, data=obs))

 Clay2   CEC1    OC1   CEC2    OC2
2.0978 4.5034 4.0277 3.5256 2.9037
```

Q73 : *According to the* VIF >= 10 *criterion, which variables in these reduced equations are highly correlated with the others? What do you conclude about the set of variables?* *Jump to A73* •

Since either `Clay1` or `Clay2` can be taken out of the equation, we compare the models, starting from a reduced model with each one taken out, both as full models and models reduced by backwards stepwise elimination:

First, eliminating `Clay2`:

```
> AIC(lm(Clay5 ~ Clay1 + CEC1 + OC1 + CEC2 + OC2, data=obs))

[1] 920.5

> AIC(step(lm(Clay5 ~ Clay1 + CEC1 + OC1 + CEC2 + OC2, data=obs), trace=0))

[1] 916.16
```

Second, eliminating `Clay1`:

```
> AIC(lm(Clay5 ~ Clay2 + CEC1 + OC1 + CEC2 + OC2, data=obs))
```

```
[1] 839.57

> AIC(step(lm(Clay5 ~ Clay2 + CEC1 + OC1 + CEC2 + OC2, data=obs) , trace=0))

[1] 835.2
```

Q74 : *Which of the two variables with high VIF in the full model should be eliminated?* *Jump to A74* •

Task 52 : Compute a reduced model by backwards stepwise elimination, starting from the full model with this variable eliminated. •

```
> (lms.2 <- step(lm(Clay5 ~ Clay2 + CEC1 + OC1 + CEC2 + OC2, data=obs)))

Start:  AIC=420.4
Clay5 ~ Clay2 + CEC1 + OC1 + CEC2 + OC2

        Df Sum of Sq   RSS AIC
- CEC1   1         5  2370 419
- OC1    1         5  2371 419
- OC2    1        22  2387 420
<none>                2365 420
- CEC2   1        56  2421 422
- Clay2  1     10782 13148 671

Step:  AIC=418.69
Clay5 ~ Clay2 + OC1 + CEC2 + OC2

        Df Sum of Sq   RSS AIC
- OC1    1         1  2371 417
- OC2    1        20  2390 418
<none>                2370 419
- CEC2   1        67  2437 421
- Clay2  1     11653 14023 678

Step:  AIC=416.75
Clay5 ~ Clay2 + CEC2 + OC2

        Df Sum of Sq   RSS AIC
- OC2    1        21  2392 416
<none>                2371 417
- CEC2   1        66  2437 419
- Clay2  1     11876 14247 678

Step:  AIC=416.03
Clay5 ~ Clay2 + CEC2

        Df Sum of Sq   RSS AIC
<none>                2392 416
- CEC2   1        47  2439 417
- Clay2  1     15687 18078 711

Call:
```

```
lm(formula = Clay5 ~ Clay2 + CEC2, data = obs)

Coefficients:
(Intercept)        Clay2         CEC2
     14.519        0.861       -0.199
```

Q75 : *What is the final model? What is its AIC? How do these compare with the model found by stepwise regression, not considering the VIF criterion?* *Jump to A75* •

Another approach is to compute the stepwise model starting from a full model, and then see the VIF of the variables retained in that model.

Task 53 : Compute the VIF for the full stepwise model. •

The `vif` function can be applied to a model object; in this case `lms`, computed above:

```
> vif(lms)

 Clay1  Clay2   CEC2
8.3567 8.0790 1.5327
```

Q76 : *What is the multi-colinearity in this model?* *Jump to A76* •

This again indicates that the two "clay" variables are highly redundant, and that eliminating one of them results in a more parsimonious model. Which to eliminate is evaluated by computing both reduced models and comparing their AIC.

Task 54 : Compute the AIC of this model, with each of the highly-correlated variables removed. •

We specify the new model with the very useful `update` function. This takes a model object and adjusts it according to a new formula, where existing terms are indicated by a period ('.').

```
> AIC(lms)

[1] 834.94

> AIC(update(lms, . ~ . - Clay1))

[1] 835.2

> AIC(update(lms, . ~ . - Clay2))

[1] 933.44
```

Q77 : *Which of the two "clay" variables should be eliminated? How much does this change the AIC?* *Jump to A77* •

7.11 Answers

A59 : *CEC is positively correlated with both clay and organic matter; however there more spread in the CEC-vs-clay relation. The two possible predictors (clay and organic matter) are also positively correlated.* *Return to Q59* •

A60 : *The covariances depend on the measurement scales, whereas the correlations are standardised to the range* $[-1, 1]$. *CEC is highly correlated* ($r = 0.74$) *with organic carbon and somewhat less so* ($r = 0.56$) *with clay content. The two predictors are also moderately correlated* ($r = 0.60$). *Return to Q60* •

A61 : *These are given by the adjusted* R^2: *0.3066 using only clay as a predictor* ($CEC = 4.83 + 0.20 \cdot Clay$), *0.5489 using only organic carbon as a predictor* ($CEC = 3.67 + 2.52 \cdot OC$), *and 0.5662 using both together* ($CEC = 2.72 + 2.16 \cdot OC + 0.64 \cdot Clay$). *Return to Q61* •

A62 : *The predictive equation is only a little affected: the slope associated with OC decreases from 2.52 to 2.16, while the intercept (associated with no clay or organic carbon) decreases by 0.95. Adding Clay increases* R^2 *by only* $0.5662 - 0.5489 = 0.0173$, *i.e. 1.7%. This is significant* ($p = 0.010152$) *at the* $\alpha = 0.05$ *but not the* $\alpha = 0.01$ *level.* *Return to Q62* •

A63 : *OC only:* 0.549; *Clay only:* 0.307; *Both:* 0.566. *The model with both is slightly better than the single-predictor model from OC.* *Return to Q63* •

A64 : *The AIC favours the model with both OC and clay, but this is only slightly better than the single-predictor model from OC.* *Return to Q64* •

A65 : *The combined model has the lowest RSS (necessarily); the difference is only* 76.4, *i.e. about* 12% *lower. There is a* 1% *probability that this reduction is due to chance.* *Return to Q65* •

A66 : *The residuals are not normally-distributed; both tails are too long, and there are about six serious under-predictions (observations 73, 60, 63, 140, 77, 124).*

The two observations with the most negative residuals (over-predictions), i.e. 1 and 10, are the only two with very high clay and *OC*[5]. *This suggests an* interaction *at high levels; "the whole is more than the sum of the parts".*

There seems to be no comparable explanations for the four observations with the most positive residuals (under-predictions). *Return to Q66* •

A67 : *The model explains 50% (zone); 80% (topsoil clay); 82.5% (both) of the variation*

[5] `obs[(Clay1 > 60) & (OC1 > 5.5),]`

in subsoil clay; the combined model is only a bit better than the model using only measured topsoil clay. *Return to Q67* •

A68 : *The regression lines for zones 2, 3, and 4 are adjusted by* 5.69, 2.25, *and* −0.66, *respectively, compared to zone 1. These are the mean differences. The slope is* 0.736, *which is somewhat flatter than the slope estimated without considering zones,* 0.829. *That is, some of the apparently steep slope in the univariate model is accounted for by the differences between zones. In particular zone 2, which has the higher clay values in both layers, has a higher mean, so that once this is accounted for the regression line is not "pulled" to the higher values.* *Return to Q68* •

A69 : *Topsoil clay is very highly significant (*$p \approx 0$ *that it isn't) and so is the intercept (0 clay and zone 1). Zone 2 is significantly different (*$p < 0.008$ *that it isn't) but the others are not. Note that in the one-way ANOVA by zone, zones 3 and 4 are both significantly different from zone 1 and 2, which form a group. Here we see that the inclusion of topsoil clay in the model has completely changed the relation to zone, since much of the zone effect was in fact a clay effect, i.e. zones had different average topsoil clay contents. The two predictors were* confounded. *Return to Q69* •

A70 : *The residuals are more or less normally distributed around 0, except for one very large negative residual (under-prediction) and seven large positive residuals (heavy tail)* *Return to Q70* •

A71 : *At point 145, the prediction is 23% while the actual is 47%; this is a severe under-prediction. This is an unusual observation: topsoil clay is 7% higher than both underlying layers. There are only two observations where topsoil clay exceeds subsoil clay (*`> which(Clay1 > Clay5)`*), 145 and 81, and for observation 81 the difference is only 2%.*

At point 119, the prediction is 34% while the actual is 48%; this is the largest under-prediction. Here topsoil clay is fairly low (21%) compared to the much higher subsoil values. *Return to Q71* •

A72 : *Variables* `Clay1` *and* `Clay2` *have* $VIF \geq 10$ *and are thus highly co-linear with other variables. As a set, the others are fairly independent.* *Return to Q72* •

A73 : *If either* `Clay1` *or* `Clay2` *are removed, the remaining set of five variables are fairly independent (all* $VIF < 5$*). This shows that the high VIF for* `Clay1` *and* `Clay2` *in the full model was due to the presence of the other "clay" variable. So either topsoil or subsoil clay should be included in a parsimonious model, but not both.* *Return to Q73* •

A74 : *Eliminating* `Clay1` *results in a much lower AIC. This seems logical, as subsoil clay (*`Clay2`*) is closer physically to the deep subsoil (target variable* `Clay5`*), so the processes that lead to a certain clay content would seem to be more similar.* *Return to Q74* •

A75 : *The final stepwise regression model, starting from the full set less* `Clay1`*, is*

`Clay5 ~ Clay2 + CEC2`, *with an AIC of 835.2. The model starting from the full set is* `Clay5 ~ Clay1 + Clay2 + CEC2`, *i.e. it has both clays as well as the subsoil CEC. Its AIC is 834.94. The two final models are almost the same except for the inclusion of the highly-colinear variable; their AIC is almost identical. So, the reduced model (without* `Clay1` *is preferred.* *Return to Q75* •

A76 : *Both* `Clay1` *and* `Clay2` *have* VIF > 8, *not above the threshold* VIF >= 10 *but not much below. Clearly,* `Clay1` *and* `Clay2` *are still highly-correlated.* *Return to Q76* •

A77 : *As in the previous tasks of this section, we see that* `Clay1` *can be eliminated with almost no increase in model information content as shown by the AIC.*

Return to Q77 •

A78 : *Zone 4 (blue points and line, low clay values) seems poorly-fit. A line with a lower intercept and a steeper slope would appear to fit better. So a model with interaction between classified and continuous predictor, allowing separate slopes for each class, might be better. For the other three the parallel lines seem OK.* *Return to Q78* •

A79 : *The model explains 83.4% of the variation in subsoil clay; this is slightly better than the additive model (82.5%).* *Return to Q79* •

A80 : *Additive terms for topsoil clay, the intercept (zone 1 at zero clay) and zone 3 are significant. This differs from the additive model, where zone 2 was the only zone significantly different from the intercept.* *Return to Q80* •

A81 : *The most significant interaction is* `Clay1:zone3` *but the probability that rejecting the null hypothesis of no difference in slopes is fairly high,* 0.076, *so we can't reject the null hypothesis at the conventional 95% confidence level.* *Return to Q81* •

A82 : *They certainly* appear *different, ranging from* 0.564 *in zone 3 (green points and line) to 1.081 (blue points and line), almost double. Yet the* t*-tests for the interaction terms are not significant at the 95% confidence level, so these four slopes could all be different just because of sampling error.* *Return to Q82* •

A83 : *The fundamental problems are: (1) small sample size in each zone; (2) a spread of points ("cloud" or "noise") within each zone. These two factors make it difficult to establish statistical significance.* *Return to Q83* •

A84 : *The nested model explains 83.4% of the variation in subsoil clay; this is slightly better than the additive model (82.5%) and the same as the interactions model. It is quite unlikely that the mean for zone 4 is different from zone 1.* *Return to Q84* •

A85 : *Yes, they are the same. For zone 1, the interaction model has the default slope (coefficient for* `Clay1`*) which is the same as the nested model slope for zone 1*

(coefficient for `zone1:Clay1`*). For zone 4, adding the slope difference in the interaction model (coefficient for* `Clay1:zone4`*) to the default slope (coefficient for* `Clay1`*) gives the same value as the nested model slope for zone 4 (coefficient for* `zone4:Clay1`*).* *Return to Q85* •

A*86* *: There is a big difference between the model coefficients and their significance. Without considering the covariate at all, the difference from zone 1 is (zone 4 ≫ zone 3 ≫ zone 2), the latter is not significantly different. In the nested model the differences are (zone 3 > zone 2 ≫ zone 4), the latter coefficient not significant; this is because the difference between zone 1 and 4 subsoil clay can be almost entirely explained if one knows the topsoil clay and allows separate regression lines for each zone. In the additive (parallel) model the differences are (zone 2 > zone 3 ≫ zone 4). The parallel regression line for zone 2 is significantly above that for zone 1, the others not significantly different.*
Return to Q86 •