Practical Topic 8

Willem Vervoort

2017-04-18

```
# root dir
knitr::opts_knit$set(root.dir =
                     "z:/willem/teaching/envx2001/otherdata")
library(pander)
library(knitr)
```

Objectives

- Review linear regression by redoing simple linear regression
- Learn to use the correlation matrix as an exploratory data analysis tool;
- Learn to perform MLR and interpret the results using (i) R and (ii) Excel

DATA: Data_Topic9_2016.xls

EXERCISE 1 - PERFORMING SLR IN R looking at toxicity in peanuts

Data: *Peanuts* worksheet

The data comprise, for 34 batches, the average level of the fungal contaminant aflatoxin in a sample of 120 pounds of peanuts and the percentage of non-contaminated peanuts in the whole batch. The data were collected with the aim of being able to predict the percentage of non-contaminated peanuts ('percent') from the aflatoxin level ('toxin') in a sample.

Read in the data using read.csv() after exporting the data as a csv file.

```
Peanuts <- read.csv("Data_peanuts_Week8.csv")
pander(Peanuts[1:5,], caption="First 5 lines of the Peanuts data set")
```

Table 1: First 5 lines of the Peanuts data set

Percent	Toxin
99.97	3
99.98	4.7
99.98	8.3
99.97	9.3
99.96	9.9

a) In R make a scatter plot (using `plot(data$Toxin,data$Percent)`) of the data. Describe the relationship between the two variables. Would you say that the percentage of non-contaminated peanuts in a batch could be predicted accurately from the level of aflatoxin in a sample via a linear relationship?

b) Use simple linear regression (`lm()`) in R to fit a straight line to the data. What is the fitted model?

c) Comment on the overall fit of the regression.

d) Is toxin a significant predictor of percentage non-contaminated peanuts?

e) Interpret the slope parameter in terms of quantifying the relationship between toxin and percent.

EXERCISE 2 – PERFORMING MLR IN R AND EXCEL

Data: *Corn* worksheet, saved as *2017_CornData_Topic8.csv* In this data

- y = P content of corn
- x1 = inorganic P content of soil
- x2 = organic P content of soil
- n = 17 sites

(The original data had 18 sites, one is removed here.)

```
Corn <- read.csv("2017_CornData_Topic8.csv")
pander(Corn[1:5,], caption="First 5 lines of the Corn data set")
```

Table 2: First 5 lines of the Corn data set

CornP	InorgP	OrgP
64	0.4	53
60	0.4	23
71	3.1	19
61	0.6	34
54	4.7	24

(i) Examination of correlations and significance

Some people find it difficult to visually interpret graphical summaries of data in more than 2 dimensions; however, 3-dimensional surface plots are reasonably common in statistics although not usually in descriptive statistics.

Instead we will examine the pairwise correlations to "get a feel" for the data. and we will we will make a 3-dimensional surface plot using the package `lattice`.

Using R, we can calculate the correlation matrix quite easily. Note the use of `round()` to limit the number of significant digits.

```
round(cor(Corn),3)
```

```
##         CornP InorgP  OrgP
## CornP   1.000  0.720 0.212
## InorgP 0.720  1.000 0.399
## OrgP    0.212  0.399 1.000
```

Regrettably this does not tell us anything about the significance of the correlations. We can test the individual correlations by using the function `cor.test()`, but then we have to do each variable.

```
with(Corn,cor.test(CornP,InorgP))
```

```
##
##  Pearson's product-moment correlation
##
## data:  CornP and InorgP
## t = 4.0192, df = 15, p-value = 0.001115
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3661783 0.8920037
## sample estimates:
##       cor
## 0.7200866
```

These results tell you that the correlation between CornP and InorgP is significant and that the p-value for this is 0.0011149. To do all the correlations together use the function `rcorr()`, which is in the package `Hmisc`, but the data have to be a `matrix`. You might have to first install the package:

```
#install.packages("Hmisc")
require(Hmisc, quiet = T)
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

Then run:
`rcorr(as.matrix(Corn))`

Tasks:

1. What do the results tell you? The top part repeats the correlations which you also got with `cor()`, but the bottom part gives the p-values. Which of the correlations are significant? If you would construct a regression model would you select both variables or just one?

2. If we were to fit a single predictor model involving EITHER InorgP OR OrgP, then which model would be more successful? (Hint, the r^2 is exactly that for a single predictor regression, the square of the correlation, r).

simple 3-D plot

A 3-D plot can be made using the function `levelplot()` in lattice. Here we plot the OrgP and InorgP in the axes and the levels in the plot are CornP.

```
require(lattice, quiet = T)
levelplot(CornP ~ InorgP + OrgP, data = Corn, col.regions = topo.colors(100))
```

It is clear that the 3-D surface plot does not have colours everywhere, but this relates of course to the underlying data. In this case we don't have continuous data in both directions, so the response (the colour) is only plotted where we have input variables.

We will now use regression to estimate the joint effects of both inorganic phosphorus and organic phosphorus on the phosphorus content of corn.

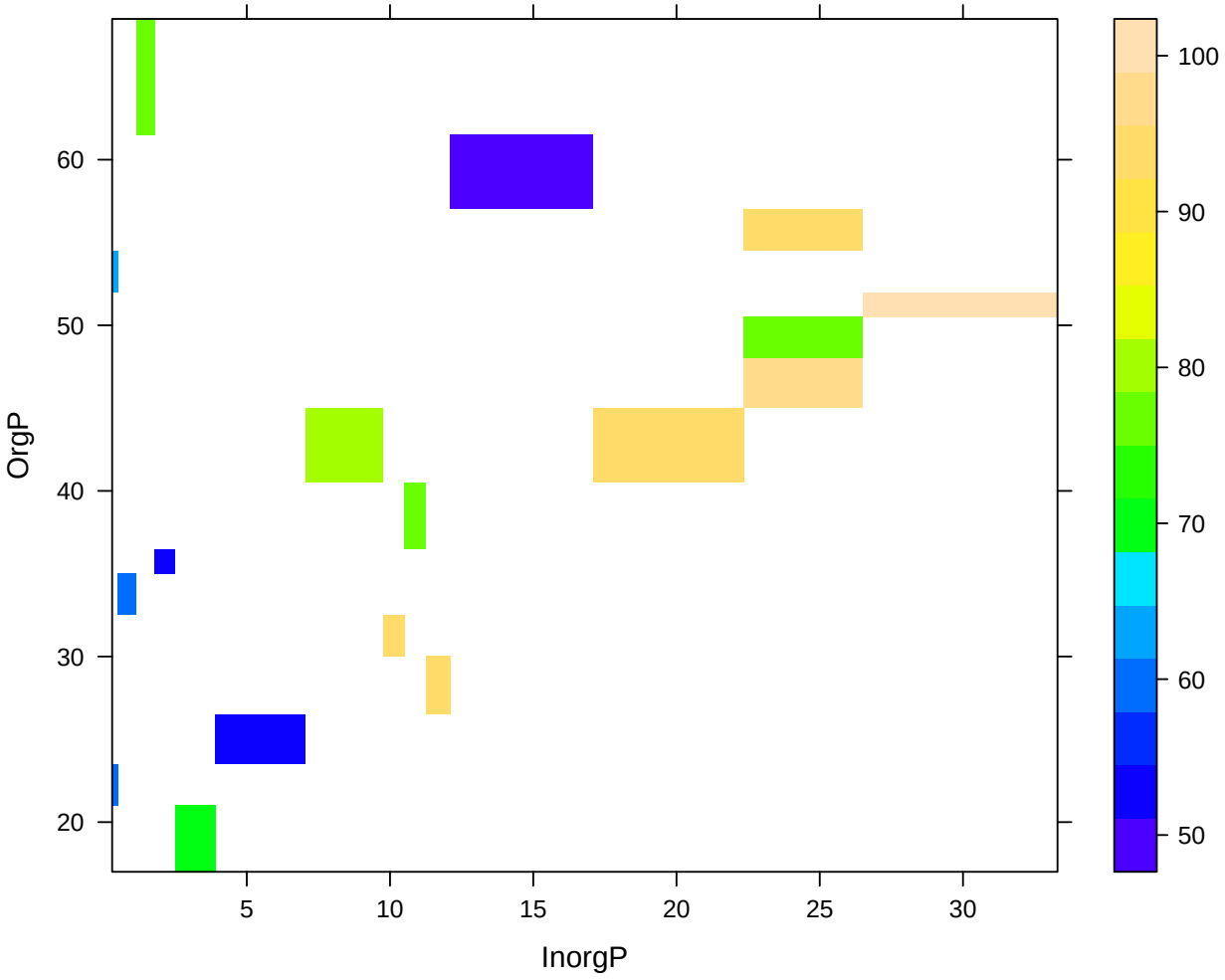

Figure 1: A 3D plot of the Corn data

(ii) MLR in R

This is fairly simple and follows the same structure as simple linear regression and uses `lm()`.

In this case:

```
MLR.Corn <- lm(CornP~InorgP + OrgP,data=Corn)
# run anova() to see the anova table
anova(MLR.Corn)
```

```
## Analysis of Variance Table
##
## Response: CornP
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## InorgP     1 2295.23 2295.23 15.2922 0.001569 **
## OrgP       1   29.95   29.95  0.1995 0.661947
## Residuals 14 2101.29  150.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# run summary to see the parameter estimates
summary(MLR.Corn)
```

```
##
## Call:
## lm(formula = CornP ~ InorgP + OrgP, data = Corn)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.282  -4.428   2.645   4.949  16.946
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.4654     9.8496   6.748 9.35e-06 ***
## InorgP        1.2902     0.3428   3.764  0.00209 **
## OrgP         -0.1110     0.2486  -0.447  0.66195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.25 on 14 degrees of freedom
## Multiple R-squared:  0.5253, Adjusted R-squared:  0.4575
## F-statistic: 7.746 on 2 and 14 DF,  p-value: 0.005433
```

Firstly though, let's check the assumptions of regression are met via residual diagnostics. We demonstrated this in class, you can simply use `plot()` on the regression model object (`MLR.Corn`) to get the plots, but we limit ourselves to plot 1,2 and 5 in the output, which is defined in `which=c(1,2,5)`. We also plot the histogram of the resiudals using `hist(resid(MLR.Corn))`. The `par(mfrow=c(2,2))` simply splits the plot into 4 components and allows you to plot everything together. `par(mfrow=c(1,1))` puts the default back

```
par(mfrow=c(2,2))
plot(MLR.Corn,which=c(1,2,5))
hist(resid(MLR.Corn))
```

```
par(mfrow=c(1,1))
```

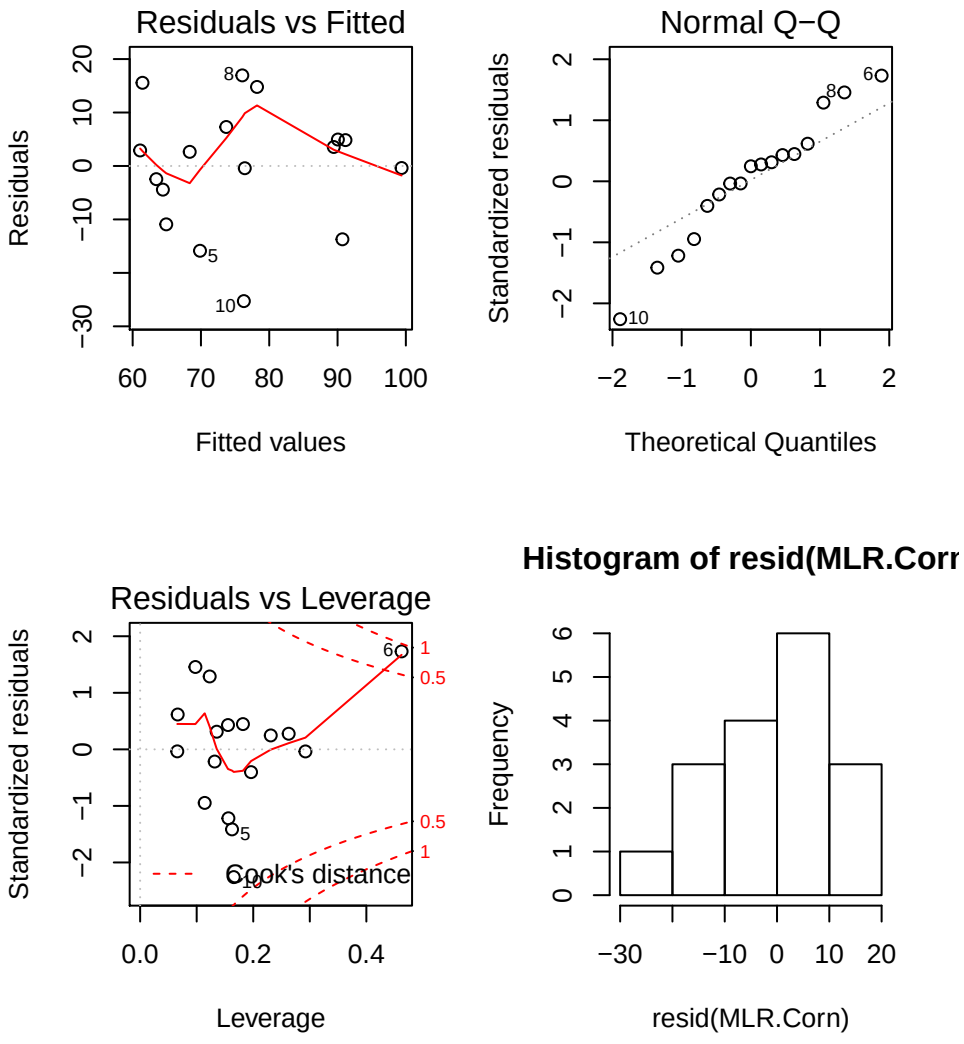

Figure 2: Diagnostic plots for Corn data

Task

Are there any apparent problems with normality of CornP residuals or equality of variance for this small data set?

(iii) MLR in Excel

(remember to activate the Data Analysis add in). Excel requires each predictor variable to be side by side. The following screen capture requests a bivariate regression (predictors in columns B and C) with residual plots and tests of normality.

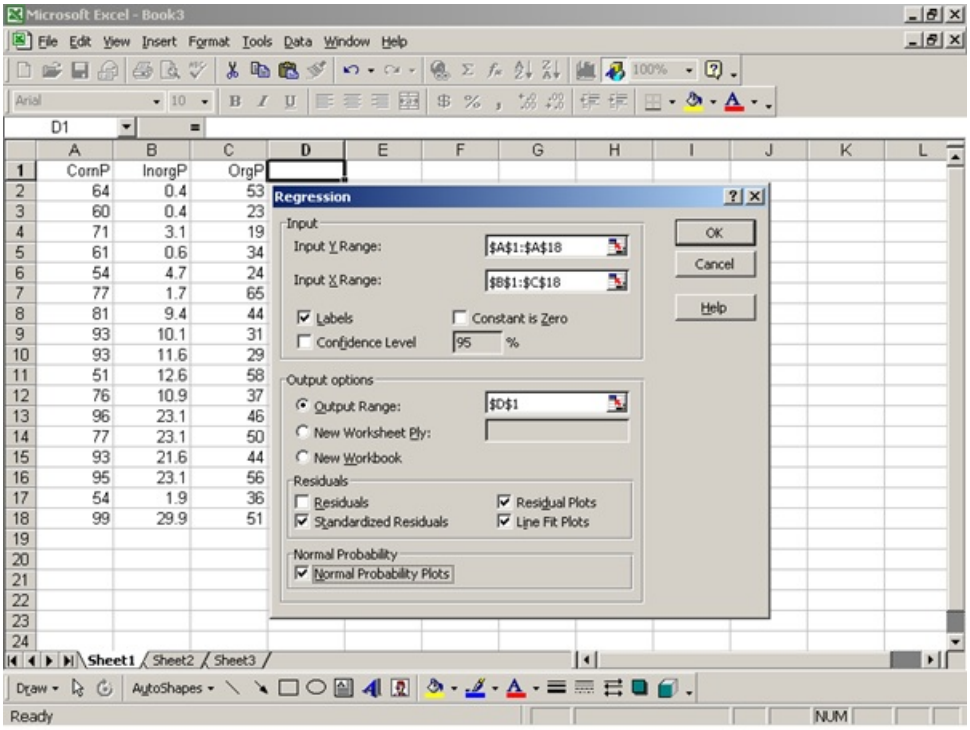

Figure 3: MLR in Excel.

The output is as follows (we are not showing the plots however):

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.724769
R Square	0.52529
Adjusted R Square	0.457474
Standard Error	12.25121
Observations	17

ANOVA

	df	SS	MS	F	Significance F
Regression	2	2325.179	1162.59	7.745836	0.00543251
Residual	14	2101.291	150.0922		
Total	16	4426.471			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	66.4654	9.849614	6.748021	9.35E-06	45.3400644	87.59075
InorgP	1.29019	0.342765	3.764071	0.002094	0.55503272	2.025348
OrgP	-0.11104	0.24859	-0.44667	0.661947	-0.6442092	0.422136

RESIDUAL OUTPUT

Observation	Predicted CornP	Residuals	Standard Residuals
1	61.10	2.90	0.25
2	64.43	-4.43	-0.39
3	68.36	2.64	0.23
4	63.46	-2.46	-0.22
5	69.86	-15.86	-1.38
6	61.44	15.56	1.36
7	73.71	7.29	0.64
8	76.05	16.95	1.48
9	78.21	14.79	1.29
10	76.28	-25.28	-2.21
11	76.42	-0.42	-0.04
12	91.16	4.84	0.42
13	90.72	-13.72	-1.20
14	89.45	3.55	0.31
15	90.05	4.95	0.43
16	64.92	-10.92	-0.95
17	99.38	-0.38	-0.03

PROBABILITY OUTPUT

Percentile	CornP
2.94	51
8.82	54
14.71	54
20.59	60
26.47	61
32.35	64
38.24	71
44.12	76
50.00	77
55.88	77
61.76	81
67.65	93
73.53	93
79.41	93
85.29	95
91.18	96
97.06	99

Figure 4: MLR output in Excel.

EXERCISE 3 – THE CORRELATION MATRIX AS A TOOL FOR EXPLORATORY DATA ANALYSIS

Data: *Loyn* worksheet, *2017_Loyn.csv*

This dataset is from Loyn (1987) which we are using in the lectures. Fragmentation of forest habitat has an impact of wildlife abundance. This study looked at the relationship between bird abundance (bird ha-1) and the characteristics of forest patches at 56 locations in SE Victoria. The predictor variables are:

• Altitude (m) [ALT]* • Year when the patch was isolated (years) [YRS.ISOL] • Grazing (coded 1-5 which is light to heavy) [GRAZE] • Patch area (ha) [AREA] • Distance to nearest patch (km) [DIST] • Distance to largest patch (km) [LDIST]

- The name is [ ] is the one used in the Excel worksheet.

In this exercise we will focus on 2 predictors (YRS.ISOL and AREA). Bring these and the response (ABUND) into R by exporting to a csv file and reading into R using read.csv().

1. Examine the histograms of each using `hist()` in R. Comment on the assumptions for regression being met.
2. Calculate the correlation matrix using rcorr() from the Hmisc package or simply cor(). Are the predictors useful?
3. Examine the scatterplot matrix using pairs().
4. The AREA predictor has a small number of observations with very large values. Apply a log10 transformation. Why are you doing this?
5. Repeat steps (1) – (3) using the transformed value of AREA.

EXERCISE 4 – MODELLING BIRD ABUNDANCE

Data: *Loyn2* worksheet, *2017_Loyn2.csv*

This worksheet has 6 predictor variables but note the 3 new names [L10AREA, L10DIST, L10LDIST]. They are the log10 transformed versions of AREA, DIST and LDIST. Each of these variables was log transformed as they had a small number of larger observations with high leverage. Import into R using `read.csv()`

Use R for the following analysis

1. Obtain the correlation between ABUND and all of the 6 predictor variables using `cor()`. Based on these, what would you expect to be the best single predictor of ABUND?
2. Use multiple regression to see whether ABUND can be predicted from L10AREA and GRAZE. Is there a significant relationship? Are the assumptions met? We are using these 2 predictors as they have the largest absolute correlations. Use `lm()` and specify the model as `ABUND~L10AREA + GRAZE`.
3. How good is the model based on the (i) r^2 (ii) adjusted r^2? Use `summary()`.
4. Which variable(s) has the most significant effect(s)? (Refer specifically to the t probabilities in the table of predictors and their estimated parameters or coefficients in the output of `summary()`). Interpret the p-values in terms of dropping predictor variables.
5. Repeat the multiple regression, but this time include YRS.ISOL as a predictor variable (it has the 3rd largest absolute correlation). This will allow you to assess the effect of YRS.ISOL with the other

variables taken into account. Compare the r^2 and adjusted r^2 values with those you calculated for the 2 predictor model.

6. Which is the better model? Why?