MTH 3240 Lab 11

Due Thu., May 7

1 Part A: Multiple Regression

1.1 Pesticide Biodegradation Data Set

Biodegradation of pesticides in the environment results from the activities of soil microorganisms. Degradation rates can vary due the abundance of such organisms and soil chemical properties such as organic matter content, pH, and nutrient supply.

A study was carried out to investigate the influence of these soil characteristics on the **degradation rate** of the herbicide *isoproturon*. Soil specimens collected from $\boldsymbol{n = 20}$ sites were analyzed for nitrate, potassium, phosphorus, pH, organic matter, and microbial biomass. Each soil specimen was then treated with *isoproturon* and monitored for 65 days for residues of the herbicide.

The file **degradation_rates.txt** contains data on the following variables:

Site = Collection site (**can be ignored for this analysis**)
$\mathbf{DT_{50}}$ = The time (days) required for the isoproturon concentration to decrease by 50%
Kd = Kinetic degradation, or rate constant exponential decrease in isoproturon (day^{-1}) (**can be ignored for this analysis**)
Nitrate = Nitrate concentration (mg/kg)
Potassium = Potassium concentration (mg/kg)
Phosphorus = Phosphorus concentration (mg/kg)
PH = Acidity (pH)
OrganicMatter = Organic matter (%)
Biomass = Microbial biomass (mg C/kg)

We'll carry out a **multiple regression analysis** with $\mathbf{DT_{50}}$ as the response variable to decide if any of the explanatory variables **Nitrate**, **Potassium**, **Phosphorus**, **PH**, **OrganicMatter**, or **Biomass** affect the **DT50** value.

1. After saving the file **degradation_rates.txt**, read the data into a data frame called, say, `my.data`:

```
my.file <- file.choose()        # Select the degradation_rates.txt file
my.data <- read.table(my.file, header = TRUE)
```

   Then **remove** the (unneeded) **Site** and **Kd** columns:

```
my.data$Site <- NULL
my.data$Kd <- NULL
```

2. Create a ***scatterplot matrix*** of the data using `pairs()`:

```
pairs(my.data, pch = 19)
```

3. Create a ***correlation matrix*** of the data using `cor()` (and ignoring the **Site** and **Kd** columns):

```
cor(my.data)
```

4. Carry out a ***multiple regression analysis***, with **DT50** as the response and **nitrate**, **potassium**, **phosphorus**, **pH**, **organic matter**, and **microbial biomass** as explanatory variables then look at the results using `summary()`:

```
my.reg <- lm(DT50 ~ Nitrate + Potassium + Phosphorus + PH + OrganicMatter +
             Biomass,
             data = my.data)

summary(my.reg)
```

5. Check the **normality** assumption for **residuals**[1] by making a ***normal probability plot*** of them:

```
qqnorm(my.reg$residuals)
qqline(my.reg$residuals)
```

[1] More formally, we're checking that the **errors** ϵ in the regression model are normally distributed.