Lecture 1. Introduction

An example multiple regression data set

Following "Hald cement data", Table B21 from page 573 of Mongomery, Peck and Vining.

	i	y	x_1	x_2	x_3	x_4
1	1	78.50	7	26	6	60
2	2	74.30	1	29	15	52
3	3	104.30	11	56	8	20
4	4	87.60	11	31	8	47
5	5	95.90	7	52	6	33
6	6	109.20	11	55	9	22
7	7	102.70	3	71	17	6
8	8	72.50	1	31	22	44
9	9	93.10	2	54	18	22
10	10	115.90	21	47	4	26
11	11	83.80	1	40	23	34
12	12	113.30	11	66	9	12
13	13	109.40	10	68	8	12

- Data consists of measurments of y, x_1, x_2, x_3, x_4 on 13 subjects.
  - y - calories per gram of cement
  - x_1 - quantity of calcium aluminate
  - x_2 - quantity of tricalcium silicate
  - x_3 - quantity of tricalcium aluminoferrite
  - x_4 - quantity of dicalcium silicate
- Goal is to predict y on the basis of the four other variables.

Multiple regression model and assumptions

- data consist of measurements on n subjects. For each subject, there is a measurement on the dependent variable y, and on each of k independent (predictor, explanatory) variables $x_1, x_2, \ldots, x_k$.
- Data: $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik}), i = 1, 2, \ldots n$
- The statistical model is

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

- which means that for subject i,

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$$

- **Assumption:**
  - the additive deviations $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are a sample from a normal population with mean 0 and variance σ^2.
  - means the ϵ_i are independent and each have the stated normal distribution
  - summarized as $\epsilon_1, \epsilon_2, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- Where $\mathbf{x} = (x_1, x_2, \ldots, x_k)$, and $\mu_{y|\mathbf{x}}$ represents the mean of y given the predictor variables $\mathbf{x}$, the model can be written as

$$y = \mu_{y|\mathbf{x}} + \epsilon$$

- read the data into R and print the data

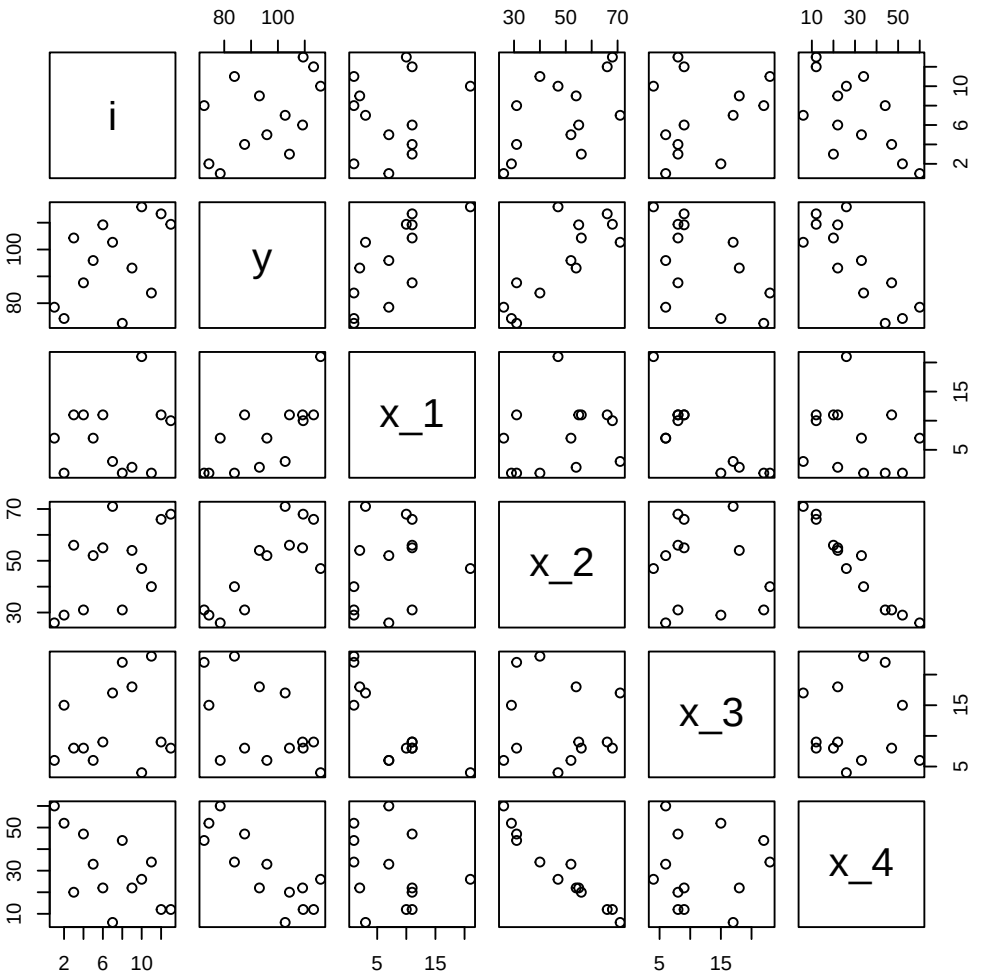
```
> cement=read.csv(
+   "http://chase.mathstat.dal.ca/~bsmith/stat3340/Data/B21.csv")
> cement
```

```
    i     y x_1 x_2 x_3 x_4
1   1  78.5   7  26   6  60
2   2  74.3   1  29  15  52
3   3 104.3  11  56   8  20
4   4  87.6  11  31   8  47
5   5  95.9   7  52   6  33
6   6 109.2  11  55   9  22
7   7 102.7   3  71  17   6
8   8  72.5   1  31  22  44
9   9  93.1   2  54  18  22
10 10 115.9  21  47   4  26
11 11  83.8   1  40  23  34
12 12 113.3  11  66   9  12
13 13 109.4  10  68   8  12
```

- For this dataset $n = 13$ and $k = 4$.

- plot the data pairwise

```
> pairs(cement)
```



- carry out a multiple linear regression of y on $x_1, \ldots, x_4$.

```
> cement.lm=lm(y~x_1+x_2+x_3+x_4,data=cement)
> summary(cement.lm)

Call:
lm(formula = y ~ x_1 + x_2 + x_3 + x_4, data = cement)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x_1           1.5511     0.7448   2.083   0.0708 .
x_2           0.5102     0.7238   0.705   0.5009
x_3           0.1019     0.7547   0.135   0.8959
x_4          -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,        Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07

> anova(cement.lm)

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
x_1        1 1450.08 1450.08 242.3679 2.888e-07 ***
```

```
x_2         1 1207.78 1207.78 201.8705 5.863e-07 ***
x_3         1    9.79    9.79   1.6370    0.2366
x_4         1    0.25    0.25   0.0413    0.8441
Residuals   8   47.86    5.98
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Chapter 2 - simple linear regression

- In **simple linear regression** there is only one predictor variable, we refer to it as x (rather than x_1).
- In this case the data are $(x_i, y_i), i = 1, \ldots, n$
- And the regression equation is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

- Some useful regression models - chapter 7 (polynomial regression) and chapter 8 (indicator variables). Note: with indicator variables, multiple regression can be used to carry out ANalysis Of VAriance (ANOVA).
- Interpretation of the output - parameter estimates, tests and confidence intervals. Chapter 2 (simple linear regression with one predictor, $k = 1$ - REVIEW) and Chapter 3 (multiple regression with k predictors)
- Were the assumptions of the regression model (ie linearity, normal errors, constant variance) appropriate, and if not, how to modify the model? Chapters 4 and 5.
- Do any observations carry undue influence in the regression? (Chapter 6)
- Multicolinearity - definition, identification, and dealing with it. (Chapter 9)
- Model building and validation. (Chapters 10 and 11)

On either your own computer, or a computer in the Learning Commons or elsewhere, run the commands used in these notes ("read.csv", "pairs", "lm" and "summary") to recreate the plot and output