Regression with Categorical Variables

- Back to SLR, where we have a response variable Y and one predictor X,

$$Y \sim X.$$

- X: a categorical variable, e.g., gender, education level, etc.
- How to run regression?
- First, let's revisit the `corrosion data`.

Corrosion:

Data consist of thirteen specimens of 90/10 Cu-Ni alloys with varying iron content in percent. The specimens were submerged in sea water for 60 days and the weight loss due to corrosion was recorded in units of milligrams per square decimeter per day.

Fe Iron content in percent

Loss Weight loss in mg per square decimeter per day

Fe	loss
0.01	127.6
0.01	130.1
0.01	128.0
0.48	124.0
0.48	122.0
0.71	110.8
0.71	113.1
0.95	103.9
1.19	101.5
1.44	92.3
1.44	91.4
1.96	83.7
1.96	86.2

Fe	loss	fitted
0.01	127.6	128.5667
0.01	130.1	128.5667
0.01	128.0	128.5667
0.48	124.0	123.0000
0.48	122.0	123.0000
0.71	110.8	111.9500
0.71	113.1	111.9500
0.95	103.9	103.9000
1.19	101.5	101.5000
1.44	92.3	91.8500
1.44	91.4	91.8500
1.96	83.7	84.9500
1.96	86.2	84.9500

```
>(127.6+130.1+
  128)/3
[1] 128.5667

>(110.8 + 113.1)/2
[1] 111.95

>(92.3+91.4)/2
[1] 91.85
> (83.7+86.2)/2
[1] 84.95
```

```
> ga=lm(loss ~ factor(Fe), data=corrosion);
> summary(ga)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        128.567      0.809 158.914    0.000
factor(Fe)0.48      -5.567      1.279  -4.352    0.005
factor(Fe)0.71     -16.617      1.279 -12.990    0.000
factor(Fe)0.95     -24.667      1.618 -15.245    0.000
factor(Fe)1.19     -27.067      1.618 -16.728    0.000
factor(Fe)1.44     -36.717      1.279 -28.703    0.000
factor(Fe)1.96     -43.617      1.279 -34.097    0.000
```

How to interpret those coefficients?

One-Way ANOVA Model

group 1	y_{11},	y_{12}	y_{13}	$\cdots$	y_{1n_1}
group 2	y_{21},	y_{22}	$\cdots$	y_{2n_2}	
	$\cdots\cdots$				
group g	y_{g1},	y_{g2},	$\cdots$	y_{gn_g}	

g is $\#$ of groups,

n_i denotes $\#$ of obs in the i-th group,

and the total sample size $n = \sum_{i=1}^{g} n_i$.

- The LS fit for y_{ij} is the corresponding group mean

$$\hat{y}_{ij} = \bar{y}_{i\cdot}$$

- Residuals

$$r_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

- RSS

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2,$$

i.e., the within-group variation.

- The one-way ANOVA model is described as

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad e_{ij} \text{ iid } \sim \mathsf{N}(0, \sigma^2).$$

- The unknown parameters are

$$(\mu, \quad \alpha_1, \quad \ldots, \quad \alpha_g).$$

- For simplicity, consider a simple case $g = 2, n_1 = 3$ and $n_2 = 2$, and write the one-way ANOVA model in matrix form

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \text{err.}$$

- However, this model is over-parameterized. We need to put some constraint on μ or α_i's.

How to the code categorical variables?

- $\mu = 0$: What's the design matrix $\mathbf{X}$? How to interpret the parameters? (the default case)

- $\alpha_1 = 0$: What's the design matrix $\mathbf{X}$? How to interpret the parameters? (`contr.treatment`)

- $\sum \alpha_i = 0$: What's the design matrix $\mathbf{X}$? How to interpret the parameters? (`contr.sum`)

- Suffices to remember the default case; the interpretations are not important.

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \text{err.}$$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \text{err}$$

```
> tmp= lm(loss ~ factor(Fe)-1, data=corrosion);
> summary(tmp)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
factor(Fe)0.01 128.5667     0.8090  158.91 4.19e-12 ***
factor(Fe)0.48 123.0000     0.9909  124.13 1.84e-11 ***
factor(Fe)0.71 111.9500     0.9909  112.98 3.24e-11 ***
factor(Fe)0.95 103.9000     1.4013   74.15 4.05e-10 ***
factor(Fe)1.19 101.5000     1.4013   72.43 4.66e-10 ***
factor(Fe)1.44  91.8500     0.9909   92.70 1.06e-10 ***
factor(Fe)1.96  84.9500     0.9909   85.73 1.70e-10 ***
```

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \text{err}.$$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_2 \end{pmatrix} + \text{err} = \begin{pmatrix} \mu \\ \mu \\ \mu \\ \mu + \alpha_2 \\ \mu + \alpha_2 \end{pmatrix} + \text{err}$$

```
> ga=lm(loss ~ factor(Fe), data=corrosion);
> summary(ga)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       128.567      0.809 158.914    0.000
factor(Fe)0.48     -5.567      1.279  -4.352    0.005
factor(Fe)0.71    -16.617      1.279 -12.990    0.000
factor(Fe)0.95    -24.667      1.618 -15.245    0.000
factor(Fe)1.19    -27.067      1.618 -16.728    0.000
factor(Fe)1.44    -36.717      1.279 -28.703    0.000
factor(Fe)1.96    -43.617      1.279 -34.097    0.000
```

```
> 123-128.5667
[1] -5.5667
> 111.95-128.5667
[1] -16.6167
```

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \text{err}.$$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \end{pmatrix} + \text{err} = \begin{pmatrix} \mu + \alpha_1 \\ \mu + \alpha_1 \\ \mu + \alpha_1 \\ \mu - \alpha_1 \\ \mu - \alpha_1 \end{pmatrix} + \text{err}$$

```
> newFe = factor(Fe)
> contrasts(newFe) = contr.sum(7)
> tmp= lm(loss ~ newFe);summary(tmp)

Estimate Std. Error t value Pr(>|t|)
(Intercept) 106.5310     0.4167 255.644 2.42e-13 ***
newFe1       22.0357     0.8007  27.519 1.52e-07 ***
newFe2       16.4690     0.9354  17.607 2.15e-06 ***
newFe3        5.4190     0.9354   5.793  0.00116 **
newFe4       -2.6310     1.2555  -2.096  0.08097 .
newFe5       -5.0310     1.2555  -4.007  0.00706 **
newFe6      -14.6810     0.9354 -15.695 4.24e-06 ***
```

```
> tmp=round(ga$fitted, dig=5); tmp=unique(tmp)
[1] 128.5667 123.0000 111.9500 103.9000
101.5000  91.8500  84.9500
> mean(tmp)
[1] 106.531
> round(tmp-mean(tmp), dig=4)
[1]  22.0357  16.4690   5.4190  -2.6310  -5.0310
-14.6810 -21.5810
```

The F-test

- Are levels of the factor really different? State the hypothesis in terms of models

$$
\begin{aligned}
H_a &: \quad y_{ij} = \mu + \alpha_i + e_{ij} \\
H_0 &: \quad y_{ij} = \mu + e_{ij}
\end{aligned}
$$

- They are two nested models, then we can use F-test.

$$\frac{(\text{RSS}_0 - \text{RSS}_a)/(g-1)}{\text{RSS}_a/(n-g)} \sim F_{g-1,n-g},$$

under H_0. The test statistic can also written as

$$\frac{\sum_{i=1}^{g} n_i (y_{i\cdot} - y_{\cdot\cdot})^2/(g-1)}{\sum_{i,j}(y_{ij} - y_{i\cdot})^2/(n-g)} = \frac{\text{Between-group Variation}/(g-1)}{\text{Within-group Variation}/(n-g)}.$$

Coagulation:

Dataset comes from a study of blood coagulation times. 24 animals were randomly assigned to four different diets and the samples were taken in a random order.

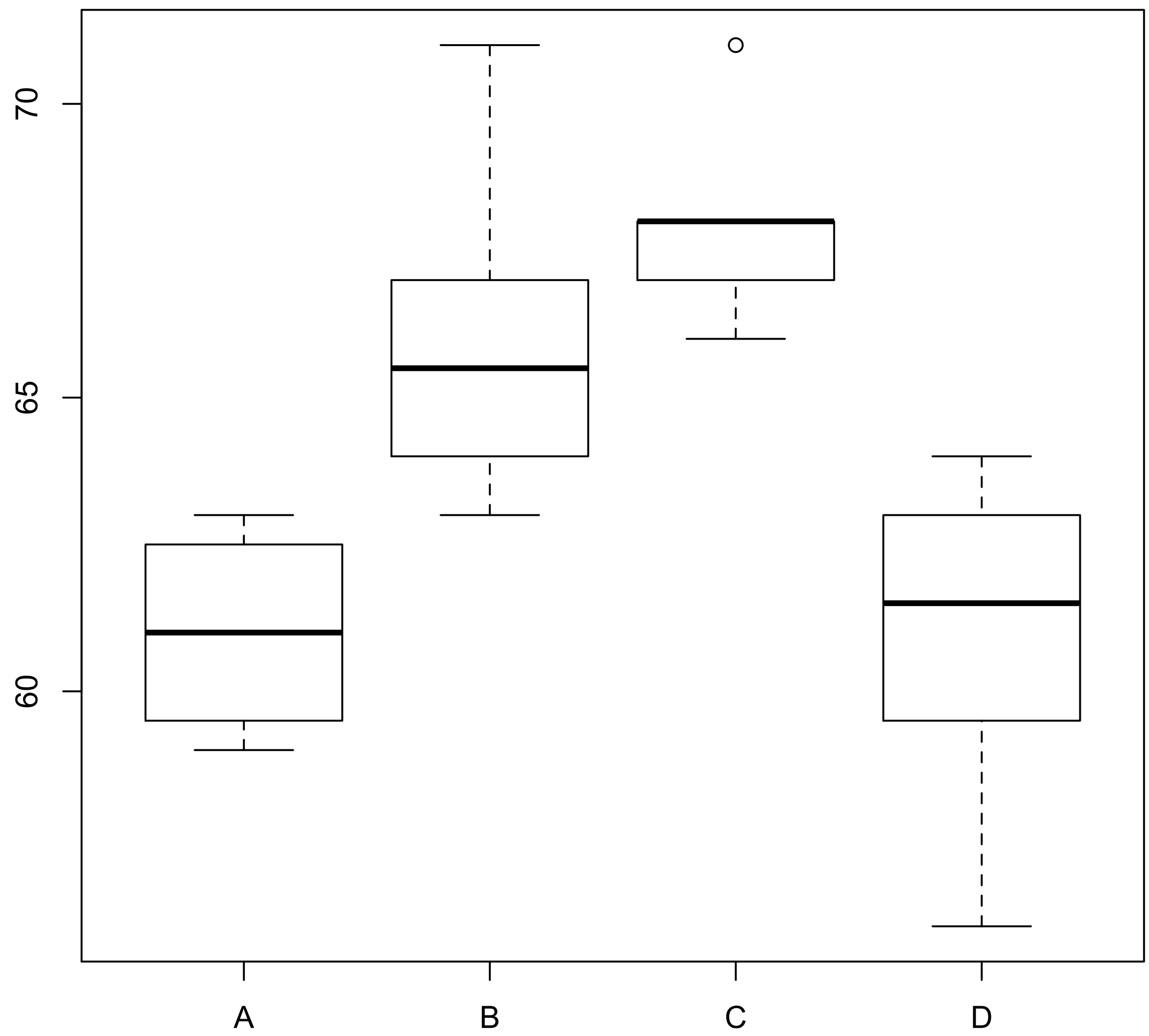
`coag` coagulation time in seconds

`diet` diet type - A,B,C or D

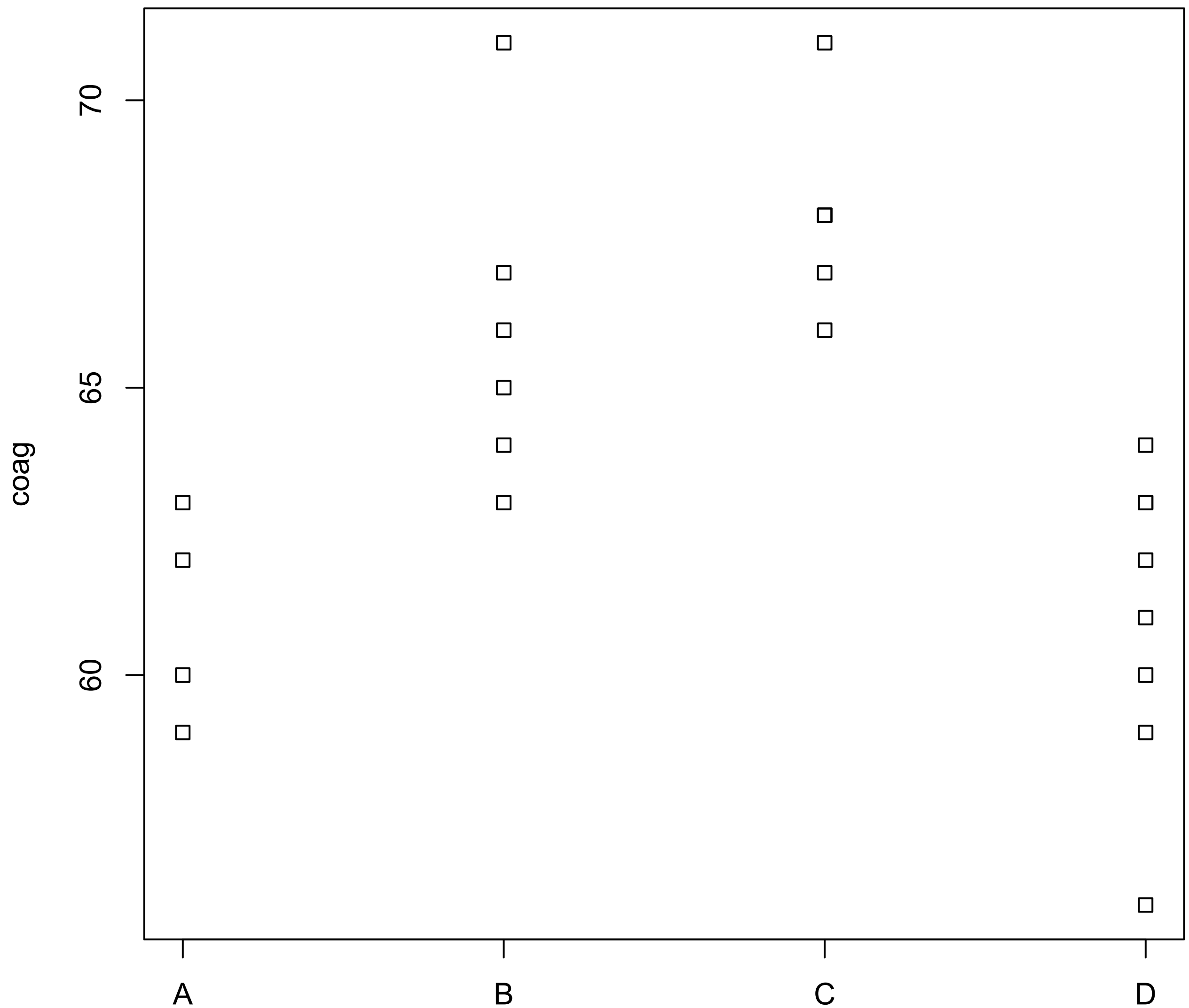
```
   coag diet
1    62    A
2    60    A
3    63    A
4    59    A
5    63    B
6    67    B
7    71    B
8    64    B
9    65    B


21   63    D
22   64    D
23   63    D
24   59    D
```

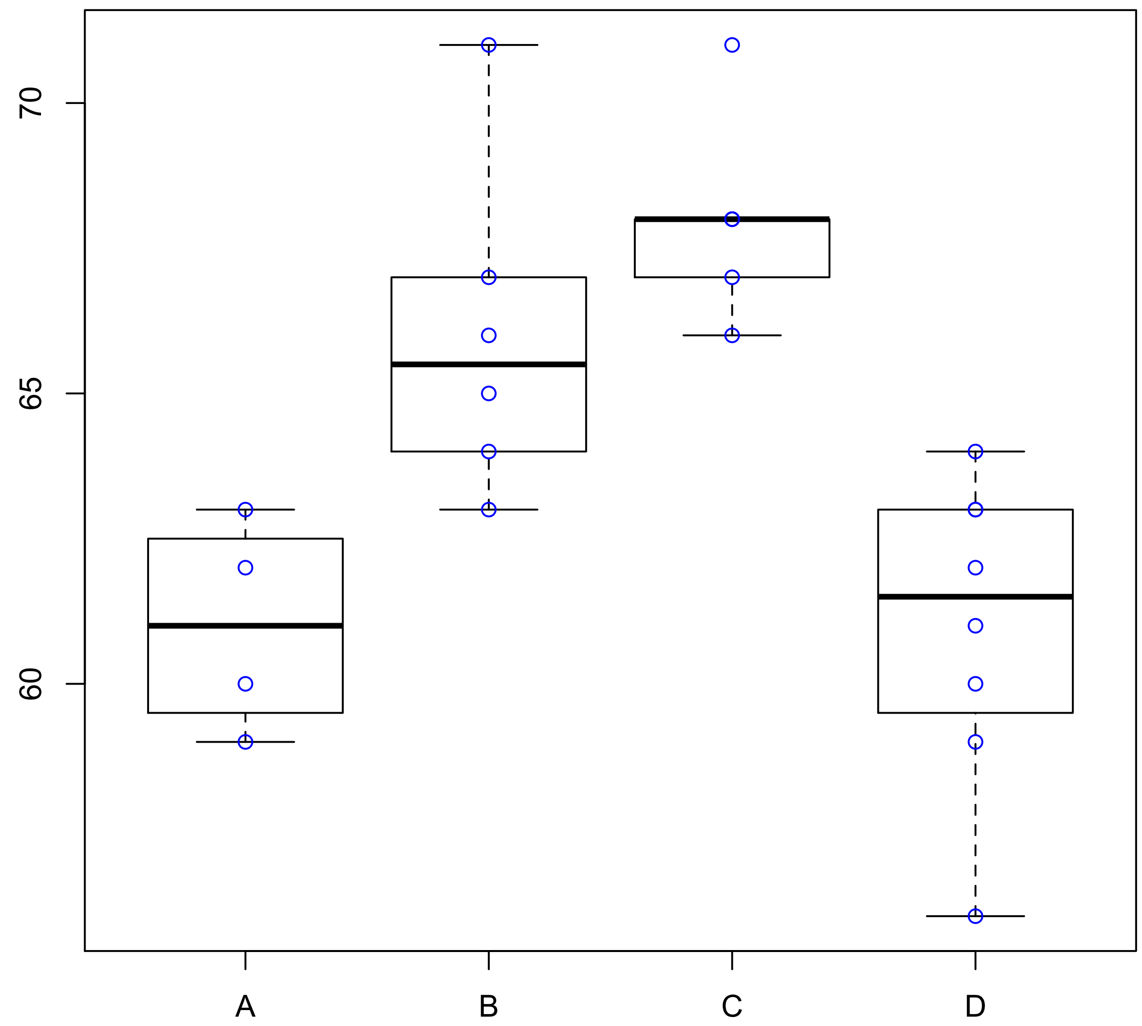
```
> attributes(diet)
$levels
[1] "A" "B" "C" "D"
```




```
> boxplot(coag ~ diet)
```




```
> stripchart(coag ~ diet, vertical=TRUE)
```




```
> boxplot(coag ~ diet, outline=FALSE)
> stripchart(coag ~ diet, vertical=TRUE,
 add=TRUE, col="blue", pch=1)
```

```
> g=lm(coag~diet)
> anova(g)
Analysis of Variance Table

Response: coag
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       3    228    76.0  13.571 4.658e-05
Residuals 20    112     5.6
```