Special Types of Predictor

Special Types of Predictors

So far we have treated the predictors as variables being observed on a **continuous** and **ordinal** scale, and then included them in a multiple regression model in simple additive form.

We now investigate other forms of predictor terms, specifically

1. polynomial terms;
2. interactions;
3. factor predictors.

Polynomial terms

For any continuous, ordinal predictor x_1, we may consider polynomial terms

$$x_1^2, x_1^3, \ldots, x_1^k$$

for integer $k \geq 2$, or more generally x_1^α for some real value $\alpha \neq 0$.

Convention: for $k \geq 0$, if we include x_1^k, we should also include

$$x_1^2, x_1^3, \ldots, x_1^{k-1}$$

in the model.

In R, we write

```
lm(y~x1+I(x1^2)+I(x1^3))
```

where `I()` means 'identity' (i.e. "compute this as it is written").

Interactions

For any continuous, ordinal predictors x_1, an ***interaction*** term allows for the **modification** of the effect of x_1 on outcome when x_2 is included in the model.

An interaction between x_1 and x_2 is denoted

$$x_1.x_2 \qquad \text{or} \qquad x_1 : x_2$$

and can be interpreted literally as a multiplication of the two terms.

Note: this is not the same as dependence (or correlation) between x_1 and x_2.

Convention: If we include an interaction $x_1 : x_2$, we should also include the "main effects" x_1 and x_2.

Interactions (cont.)

For two continuous predictors, the model written

$$X_1 + X_2 + X_1 : X_2$$

means "main effects plus interaction"; in terms of the conditional mean, we may have

$$\mathbb{E}_{Y_i|\mathbf{X}}[Y_i|\mathbf{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1}.x_{i2}$$

that is, a $p = 4$ parameter model.

Writing this model as

$$\mathbb{E}_{Y_i|\mathbf{X}}[Y_i|\mathbf{x}_i] = \beta_0 + (\beta_1 + \beta_{12}x_{i2})x_{i1} + \beta_2 x_{i2}$$

we see that the contribution of x_{i1} in the model is incorporated in the term

$$\beta_1 + \beta_{12}x_{i2}$$

that is, the "pure" effect of x_{i1} is captured by β_1, and then this is augmented by the additional contribution due to the presence of x_{i2} in the model, captured by β_{12}.

Interactions (cont.)

In `R`, we write

```
lm(y ~ x1+x2+x1:x2)
```

for the main effect plus interaction model. We may also write this model

```
lm(y ~ x1*x2)
```

Interactions (cont.)

Higher-order interactions: we may include multiple-term interactions

$$X_1 : X_2 : X_3 \qquad X_1 : X_2 : X_3 : X_4$$

etc.

Convention: a model that contains a multi-way interaction involving k predictors should also include all the **lower order** interactions including the same predictors, and the k main effects.

For example, if you include $X_1 : X_2 : X_3$ you should also include

- X_1, X_2 and X_3;
- $X_1 : X_2$, $X_1 : X_3$ and $X_2 : X_3$.

Interactions (cont.)

In R, we write

```
lm(y ~ x1*x2*x3)
```

for the model

$$1 + X_1 + X_2 + X_3 + X_1 : X_2 + X_1 : X_3 + X_2 : X_3 + X_1 : X_2 : X_3$$

but we could also write (say)

```
lm(y ~ x1*x2+x3)
```

for the model

$$1 + X_1 + X_2 + X_3 + X_1 : X_2$$

and so on.

Interactions (cont.)

In R, we may also write

```
lm(y ~ (x1+x2+x3)^2)
```

for the model

$$1 + X_1 + X_2 + X_3 + X_1 : X_2 + X_1 : X_3 + X_2 : X_3$$

that includes all main effects and all two-way interactions.

Factor predictors

A *factor predictor* is a predictor that takes discrete values on a nominal (i.e. non-ordered) scale. We can consider these discrete values as "labels".

- **University attended:** McGill, UT, UBC, ...
- **pain relief treatment:** Tylenol, Advil, aspirin, ...
- **therapy type:** pharmacologic, behavioural, surgical, ...

The possible values that a factor can take are termed *levels*.

These are non-numeric quantities; we must convert them to numeric values in order to fit them into the linear regression modelling framework.

Factor predictors (cont.)

Suppose predictor X_1 has $M = L + 1$ levels: we

- pick a "baseline" level of the factor (say level 1), and denote its modelled mean β_0;
- for levels 2,3,…,M, write the modelled mean as

$$\beta_{l+1} = \beta_0 + \beta_l^C \qquad l = 1, 2, \ldots, M - 1.$$

The model then becomes

$$\mathbb{E}_{Y_i|\mathbf{X}}[Y_i|\mathbf{x}_i] = \beta_0 + \sum_{l=1}^{L} \beta_l^C \mathbb{1}_j(x_{i1}) = \begin{cases} \beta_1 = \beta_0 & l = 0 \\ \beta_2 = \beta_0 + \beta_1^C & l = 1 \\ \vdots & \vdots \\ \beta_M = \beta_0 + \beta_L^C & l = L \end{cases}$$

Factor predictors (cont.)

Note that in each parameterization, the model contains $M = L + 1$ parameters

$$\beta = (\beta_1, \beta_2, \ldots, \beta_M)^\top$$

or

$$\beta^{\mathsf{C}} = (\beta_0, \beta_1^{\mathsf{C}}, \beta_2^{\mathsf{C}}, \ldots, \beta_L^{\mathsf{C}})^\top.$$

We have that $\beta^{\mathsf{C}} = \mathbf{C}\beta$, where $\mathbf{C}$ is the $(M \times M)$ matrix

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ -1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & 0 & 0 & \cdots & 1 & 0 \\ -1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

In `R`, the default is to use the β^{C} '***contrast***' parameterization.

Factor predictors (cont.)

In R we may define a factor directly

```
> x<-as.factor(rep(c('A','B','C','D'),3))
> x
 [1] A B C D A B C D A B C D
Levels: A B C D
```

or using the `factor` function

```
> x<-factor(rep(1:4,3),labels=c('A','B','C','D'))
> x
 [1] A B C D A B C D A B C D
Levels: A B C D
```

or using the `gl` function

```
> x<-gl(4,1,12,labels=c('A','B','C','D'))
> x
 [1] A B C D A B C D A B C D
```

Example

```
> x<-gl(4,1,12,labels=c('Levl','Lev2','Lev3','Lev4'))
> be<-c(2,3,-1,5)
> Cmat<-diag(1,4);Cmat[2:4,1]<--1
> Cmat
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]   -1    1    0    0
[3,]   -1    0    1    0
[4,]   -1    0    0    1
> (beC<-Cmat %*% be)
     [,1]
[1,]    2
[2,]    1
[3,]   -3
[4,]    3
> mean.vec<-be[as.numeric(x)]
> set.seed(4387)
> y<-rnorm(length(x),mean.vec,1)
```

Example (cont.)

```
> data.frame(x,y)
      x          y
1  Lev1  1.8545958
2  Lev2  3.1055322
3  Lev3 -1.6076756
4  Lev4  5.5941033
5  Lev1  1.4083449
6  Lev2  2.6129410
7  Lev3 -1.3477802
8  Lev4  5.8149936
9  Lev1  0.6793912
10 Lev2  3.1493288
11 Lev3 -2.6005920
12 Lev4  5.2520114
```

Example (cont.)

```
> summary(lm(y ~ x))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3141     0.2826   4.650  0.00164 **
xLev2         1.6418     0.3996   4.108  0.00340 **
xLev3        -3.1661     0.3996  -7.922 4.68e-05 ***
xLev4         4.2396     0.3996  10.609 5.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4895 on 8 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9702
F-statistic: 120.4 on 3 and 8 DF,  p-value: 5.373e-07
```

The parameters estimated here are

$$\beta_0, \beta_1^C, \beta_2^C, \beta_3^C$$

Combining factor predictors

Consider two predictors, X_1, X_2 that have M_1 and M_2 levels respectively. The model

$$X_1 + X_2$$

says that the two factor predictors combine additively to affect the response. The full model formula is

$$\beta_0 + \underbrace{\sum_{j=1}^{M_1-1} \beta^{\mathrm{C}}_{1j} \mathbb{1}_j(x_{i1})}_{\text{main effect of } X_1} + \underbrace{\sum_{l=1}^{M_2-1} \beta^{\mathrm{C}}_{2l} \mathbb{1}_l(x_{i2})}_{\text{main effect of } X_2}$$

For each i, only **one** term in each summation is non-zero.

This model contains

$$1 + (M_1 - 1) + (M_2 - 1) = M_1 + M_2 - 1$$

parameters.

Combining factor predictors (cont.)

```
> (x1<-gl(5,1,10))
 [1] 1 2 3 4 5 1 2 3 4 5
Levels: 1 2 3 4 5
> (x2<-gl(2,5,10))
 [1] 1 1 1 1 1 2 2 2 2 2
Levels: 1 2
> be1<-c(-2,2,3,0,1)
> be2<-c(0,2)
> mean.vec<-be1[as.numeric(x1)]+be2[as.numeric(x2)]
> set.seed(4387)
> y<-rnorm(length(x1),mean.vec,1)
> data.frame(x1,x2,model.mean=mean.vec,y)
   x1 x2 model.mean          y
1   1  1         -2 -2.1454042
2   2  1          2  2.1055322
3   3  1          3  2.3923244
4   4  1          0  0.5941033
5   5  1          1  0.4083449
6   1  2          0 -0.3870590
7   2  2          4  3.6522198
8   3  2          5  5.8149936
9   4  2          2  0.6793912
10  5  2          3  3.1493288
```

Combining factor predictors (cont.)

In R:

```
> summary(lm(y~x1+x2))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.2216     0.6962  -3.191  0.03319 *
x12           4.1451     0.8988   4.612  0.00994 **
x13           5.3699     0.8988   5.974  0.00394 **
x14           1.9030     0.8988   2.117  0.10167
x15           3.0451     0.8988   3.388  0.02759 *
x22           1.9108     0.5685   3.361  0.02827 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8988 on 4 degrees of freedom
Multiple R-squared:  0.9305,    Adjusted R-squared:  0.8437
F-statistic: 10.71 on 5 and 4 DF,  p-value: 0.01967
```

Interactions and factor predictors

An interaction term between X_1 and X_2 now introduces

$$(M_1 - 1) \times (M_2 - 1)$$

new parameters that modify the effect of **each level** of each predictor on the outcome. For the model

$$X_1 + X_2 + X_1 : X_2$$

the full model formula is

$$\beta_0 + \underbrace{\sum_{j=1}^{M_1-1} \beta^{C}_{1j} \mathbb{1}_j(x_{i1})}_{\text{main effect of } X_1} + \underbrace{\sum_{l=1}^{M_2-1} \beta^{C}_{2l} \mathbb{1}_l(x_{i2})}_{\text{main effect of } X_2} + \underbrace{\sum_{j=1}^{M_1-1} \sum_{l=1}^{M_2-1} \beta^{C}_{12jl} \mathbb{1}_j(x_{i1}) \mathbb{1}_l(x_{i2})}_{\text{interaction}}$$

For each data point, only one term in each summation is non-zero.

Interactions and factor predictors (cont.)

Compare this with the model without the interaction

$$X_1 + X_2$$

the full model formula is

$$\beta_0 + \sum_{j=1}^{M_1-1} \beta^{\mathrm{C}}_{1j} \mathbb{1}_j(x_{i1}) + \sum_{l=1}^{M_2-1} \beta^{\mathrm{C}}_{2l} \mathbb{1}_l(x_{i2})$$

that is, with each parameter $\beta^{\mathrm{C}}_{12jl}$ set to zero in the previous formula.

Interactions and factor predictors (cont.)

For higher order interactions, the number of extra parameters is multiplied up; with an interaction between k predictors, we introduce

$$(M_1 - 1) \times (M_2 - 1) \times \cdots (M_k - 1)$$

new parameters.

An interaction between a **numeric** predictor X_1 and a factor predictor X_2 taking M levels introduces $(M-1)$ new parameters that describe how the expected outcome changes as a function of X_1 at each non-baseline level of the factor. For example, consider the model

$$X_1 + X_2 + X_1 : X_2$$

This model says that the expected outcome depends on both X_1 and X_2, but that the effect of X_1 is **modified** in the presence of X_2. The full model formula is

$$\beta_0 + \underbrace{\beta_1 x_{i1}}_{\text{baseline slope}} + \sum_{j=1}^{M_2-1} \beta_{2j}^{C} \mathbb{1}_j(x_{i2}) + \underbrace{\sum_{j=1}^{M_2-1} \beta_{12j}^{C} x_{i1} \mathbb{1}_j(x_{i2})}_{\text{modified slope}}$$

Contrast this with the model

$$X_1 + X_2$$

which says that X_1 has the same effect for **all** levels of the factor predictor X_2 (that is, the effect of the two predictors is **additive**. The full model formula is

$$\beta_0 + \beta_1 x_{i1} + \sum_{j=1}^{M_2 - 1} \beta_{2j}^{\mathrm{C}} \mathbb{1}_j(x_{i2})$$

That is, there is a single slope parameter β_1, but different intercepts for each of the levels of X_2.

Interactions and factor predictors (cont.)

That is, in the baseline group, the expectation is

$$\beta_0 + \beta_1 x_{i1}$$

whereas at the lth level of the factor predictor, the expectation is

$$(\beta_0 + \beta_{0l}^{\mathrm{C}}) + (\beta_1 + \beta_{1l}^{\mathrm{C}})x_{i1}$$

that is, the intercept and slope are changed from baseline.