

University of Oxford

More Notes
for
Linear Models

YY Teo

Michaelmas Term, 2007

*Department of Statistics, 1 South Parks Road,
Oxford OX1 3TG*

Chapter 1

Contrasts

Contrasts are used to represent categorical variables in the setting of regression, as categorical variables cannot be efficiently represented using just one coefficient (unless the categorical variable has only 2 levels, like Gender with levels of Male and Female). Compare this with numerical variable where every variable has a corresponding coefficient.

For example, if we are interested in finding how weight vary with height and gender, you may be interested in the following equations

$$\text{Weight} = \beta_0 + \beta_1 \text{Height} + \beta_2 f(\text{Gender})$$

Note that the variable Gender has only one coefficient β_2 because Gender has only 2 levels, and therefore only 1 coefficient is required (this generalises to $K - 1$ coefficients for a categorical variable with K levels). Contrasts will basically specify how the function $f(\cdot)$ will look like. The common forms of contrasts you will encounter in this course are Treatment, Helmert and Polynomial (orthogonal). We will look at each in turn.

Remember that a K -level categorical variable will have $K - 1$ coefficients and $K - 1$ contrast functions $f_i(\cdot), i = 1, \dots, K - 1$.

1.1 Treatment Contrasts

R uses treatment contrasts by default, but this is not the case in S-Plus. You can get S-PLUS to use treatment contrasts by specifying

```
> options(contrasts=c("contr.treatment", "contr.poly"))
```

For a categorical variable with 4 levels, this is equivalent to

```
> contr.treatmean(4)
  [,1] [,2] [,3]
1     0     0     0
2     1     0     0
3     0     1     0
4     0     0     1
```

The above representation basically specifies what the functions $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$ are (recall there are only 3 contrasts functions because we are considering a categorical variable with 4 levels). This means

$$\begin{aligned}f_1 &= I(\text{Level} = 2) \\f_2 &= I(\text{Level} = 3) \\f_3 &= I(\text{Level} = 4)\end{aligned}$$

where $I(M)$ represents the indicator random variable which takes 1 when the condition M is satisfied, and 0 otherwise.

For our example previously,

$$\text{Weight} = \beta_0 + \beta_1 \text{Height} + \beta_2 f(\text{Gender})$$

under treatment contrasts, $f(\cdot)$ will take value 1 when Gender = Male, and 0 when Gender = Female.

Suppose we have another example where we have individuals from 3 different populations, Africans, Chinese and Europeans, and this is coded under the variable `Nation`. Using treatment contrasts, by typing

```
> options(contrasts=c("contr.treatment", "contr.poly"))
> model.matrix(Weight ~ Height + Nation)
```

we will effectively be considering an equation of the form

$$\text{Weight} = \beta_0 + \beta_1 \text{Height} + \beta_2 \text{I}(\text{Nation} = \text{Chinese}) + \beta_3 \text{I}(\text{Nation} = \text{European})$$

1.2 Helmert Contrasts

S-PLUS uses Helmert contrasts by default, but you can specify that Helmert contrasts be used (in both S-PLUS and R) by

```
> options(contrasts=c("contr.helmert", "contr.poly"))
```

For a categorical variable with 4 levels, this is equivalent to

```
> contr.helmert(4)
  [,1] [,2] [,3]
1  -1  -1  -1
2   1  -1  -1
3   0   2  -1
4   0   0   3
```

Again we have 3 functions $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$, which is specified by the columns in the above matrix. The equation that corresponds to a variable X (only!, without other explanatory variables) of the above form is

$$y = \beta_0 + \beta_1 f_1(\cdot) + \beta_2 f_2(\cdot) + \beta_3 f_3(\cdot).$$

Thus if we observe level 1, we will obtain a fitted model of the form

$$\hat{y} = \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3,$$

and if we observe level 2, the fitted model will be

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3.$$

Level 3 corresponds to the equation

$$\hat{y} = \hat{\beta}_0 + 2\hat{\beta}_2 - \hat{\beta}_3,$$

and level 4 corresponds to the equation

$$\hat{y} = \hat{\beta}_0 + 3\hat{\beta}_3,$$

For another example, consider the previous example involving individuals from 3 continents, `Nation`, with levels given by African, Chinese, European. Using helmert contrasts, by typing

```
> options(contrasts=c("contr.helmert", "contr.poly"))
> model.matrix(Weight ~ Height + Nation)
```

we will effectively be considering an equation of the form

$$\text{Weight} = \beta_0 + \beta_1\text{Height} + \beta_2f(\text{Nation}) + \beta_3g(\text{Nation})$$

where

$$f(\text{Nation}) = \begin{cases} -1 & \text{Nation} = \text{African} \\ 1 & \text{Nation} = \text{Chinese} \\ 0 & \text{Nation} = \text{European} \end{cases}$$

and

$$g(\text{Nation}) = \begin{cases} -1 & \text{Nation} = \text{African} \\ -1 & \text{Nation} = \text{Chinese} \\ 2 & \text{Nation} = \text{European} \end{cases}$$

Therefore for an African subject, the equation will be

$$\text{Weight} = \beta_0 + \beta_1\text{Height} - \beta_2 - \beta_3 ,$$

a Chinese subject

$$\text{Weight} = \beta_0 + \beta_1\text{Height} + \beta_2 - \beta_3 ,$$

and an European subject

$$\text{Weight} = \beta_0 + \beta_1\text{Height} + 2\beta_3.$$

1.3 Polynomial Contrasts

Orthogonal polynomials are used as contrasts for ordinal (or ordered factors) terms. For example, if the explanatory variable is `Customer Service`, which has 3 levels, {Poor, Average, Good}. Then there is a natural ordering where Average is ‘greater’ (or better) than Poor, while Good is ‘greater’ (or better) than both Average and Poor. These are still levels of a categorical variables, but there is a natural ordering involved. Such categorical variables are called ordinal variables.

You can view how orthogonal polynomial contrasts (say for 4 levels) look like using the command

```
> contr.poly(4)
```

which produces

```
> contr.poly(4)
      .L   .Q   .C
1 -0.671  0.5 -0.224
2 -0.224 -0.5  0.671
3  0.224 -0.5 -0.671
4  0.671  0.5  0.224
```

The interpretation of the contrasts within the equation is identical to that for Helmert contrasts, thus the contribution by an ordered categorical variable which takes level 3 will be

$$0.224\beta_1 - 0.5\beta_2 - 0.671\beta_3.$$