

Regression with Categorical Predictors

To handle categorical or nominal predictors in regression analysis, dummy variables are introduced. A total of $c-1$ dummy variables are usually required to account for a categorical predictor having c levels. However, there are many different ways of coding dummy variables. R has four different ways of coding dummy variables as built-in functions: `contr.helmert`, `contr.poly`, `contr.sum`, and `contr.treatment`. By default, the reference cell coding scheme is implemented in function `contr.treatment`. To choose from these four choices, one only needs to change the `contrasts` argument in options.

We use the following artificial data to illustrate how the coding is actually done for a categorical predictor Z . Go through the following codes step by step and pay special attention to the way of generating data when a categorical predictor is involved.

```
n <- 12; x <- runif(n)
z <- rep(letters[1:4], rep(3, 4))
z.1 <- rep(c(-2.5, 3.2, 6.0, 12.66), rep(3, 4))
y <- 1 + 2*z.1 + 3*x + rnorm(n, 0, 1)
data <- data.frame(y=y, x=x, z=z)
```

The nominal variable Z has four levels: A, B, C, and D. So three dummy variables, Z_1, Z_2 , and Z_3 are needed when an intercept term is included. The data set also involves one additional continuous variable X . The model to be considered is as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_{1i} + \beta_3 Z_{2i} + \beta_4 Z_{3i} + \varepsilon_i$$

However, depending on different coding schemes, the interpretation of β_0 , β_2 , β_3 , and β_4 also differs.

- **Reference Cell Coding**

```
# By default - contr.treatment

> options()$contrasts
      unordered      ordered
"contr.treatment"  "contr.poly"
> fit <- lm(y ~ x + factor(z), x = T, data = data)
> summary(fit)
```

```
Call: lm(formula = y ~ x + factor(z), data = data, x = T)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.995 -0.283 -0.0000757  0.291  0.879
```

```
Coefficients:
```

```
      Value Std. Error t value Pr(>|t|)
```

```
(Intercept)  -4.532  0.750  -6.039  0.001
             x    4.120  0.804   5.122  0.001
             factor(z)b 12.223  0.626  19.524  0.000
             factor(z)c 17.234  0.586  29.401  0.000
             factor(z)d 29.882  0.552  54.088  0.000
```

Residual standard error: 0.657 on 7 degrees of freedom

Multiple R-Squared: 0.998

F-statistic: 777 on 4 and 7 degrees of freedom, the p-value is 2.42e-009

Correlation of Coefficients:

```
(Intercept)      x factor(z)b factor(z)c
x -0.863
factor(z)b -0.751      0.516
factor(z)c -0.675      0.404  0.600
factor(z)d -0.554      0.241  0.540      0.541
```

```
> fit$x
(Intercept)      x factor(z)b factor(z)c factor(z)d
1           1 0.7959           0           0           0
2           1 0.6539           0           0           0
3           1 0.9660           0           0           0
4           1 0.4202           1           0           0
5           1 0.3609           1           0           0
6           1 0.4296           1           0           0
7           1 0.1340           0           1           0
8           1 0.8029           0           1           0
9           1 0.5959           0           1           0
10          1 0.8449           0           0           1
11          1 0.1389           0           0           1
12          1 0.9359           0           0           1
```

• Helmert Contrasts

```
> options(contrasts=c("contr.helmert", "contr.poly"))
> fit <- lm(y ~ x + factor(z), x = T, data = data)
> fit$x
(Intercept)      x factor(z)1 factor(z)2 factor(z)3
1           1 0.7959          -1          -1          -1
2           1 0.6539          -1          -1          -1
3           1 0.9660          -1          -1          -1
4           1 0.4202           1          -1          -1
5           1 0.3609           1          -1          -1
6           1 0.4296           1          -1          -1
7           1 0.1340           0           2          -1
8           1 0.8029           0           2          -1
9           1 0.5959           0           2          -1
10          1 0.8449           0           0           3
11          1 0.1389           0           0           3
12          1 0.9359           0           0           3
```

```
> summary(fit)
```

Call: lm(formula = y ~ x + factor(z), data = data, x = T)

Residuals:

```
    Min       1Q   Median       3Q      Max
-0.995 -0.283 -0.0000757  0.291  0.879
```

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	10.303	0.511	20.164	0.000
x	4.120	0.804	5.122	0.001
factor(z)1	6.111	0.313	19.524	0.000
factor(z)2	3.707	0.157	23.643	0.000
factor(z)3	5.016	0.110	45.485	0.000

Residual standard error: 0.657 on 7 degrees of freedom
 Multiple R-Squared: 0.998
 F-statistic: 777 on 4 and 7 degrees of freedom, the p-value is 2.42e-009

Correlation of Coefficients:

	(Intercept)	x	factor(z)1	factor(z)2
x	-0.929			
factor(z)1	-0.479	0.516		
factor(z)2	-0.148	0.160	0.082	
factor(z)3	0.113	-0.122	-0.063	-0.019

In the case, obviously $Z_1 - Z_3$ are defined as follows

$Z_{1i} = -1$, if A; 1 , if B; 0 , otherwise.

$Z_{2i} = -1$, if A; 2 , if C; 0 , otherwise.

$Z_{3i} = -1$, if A; 3 , if D; 0 , otherwise.

To interpret the parameters, one is encouraged to break down the model for each category of Z .

- **Orthogonal polynomials**

```
> options(contrasts=c("contr.poly", "contr.poly"))
> fit <- lm(y ~ x + factor(z), x = T, data = data)
> fit$x
  (Intercept)      x factor(z).L factor(z).Q factor(z).C
1            1 0.7959    -0.6708         0.5    -0.2236
2            1 0.6539    -0.6708         0.5    -0.2236
3            1 0.9660    -0.6708         0.5    -0.2236
4            1 0.4202    -0.2236        -0.5     0.6708
5            1 0.3609    -0.2236        -0.5     0.6708
6            1 0.4296    -0.2236        -0.5     0.6708
7            1 0.1340     0.2236        -0.5    -0.6708
8            1 0.8029     0.2236        -0.5    -0.6708
9            1 0.5959     0.2236        -0.5    -0.6708
10           1 0.8449     0.6708         0.5     0.2236
11           1 0.1389     0.6708         0.5     0.2236
12           1 0.9359     0.6708         0.5     0.2236
```

- **Sum Contrasts**

```

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

> fit <- lm(y ~ x + factor(z), x = T, data = data)
> summary(fit)

Call: lm(formula = y ~ x + factor(z), data = data, x = T)
Residuals:
    Min       1Q   Median       3Q      Max
-0.995 -0.283 -0.0000757  0.291  0.879

Coefficients:
            Value Std. Error t value Pr(>|t|)
(Intercept)  10.303    0.511    20.164  0.000
           x    4.120    0.804     5.122  0.001
factor(z)1  -14.835    0.371   -39.959  0.000
factor(z)2   -2.612    0.361    -7.236  0.000
factor(z)3    2.399    0.334     7.173  0.000

Residual standard error: 0.657 on 7 degrees of freedom
Multiple R-Squared: 0.998
F-statistic: 777 on 4 and 7 degrees of freedom, the p-value is 2.42e-009

Correlation of Coefficients:
            (Intercept)      x factor(z)1 factor(z)2
           x -0.929
factor(z)1  0.433          -0.467
factor(z)2 -0.386          0.415 -0.462
factor(z)3 -0.176          0.190 -0.378   -0.219

> fit$x

      (Intercept)      x factor(z)1 factor(z)2 factor(z)3
1             1 0.7959             1             0             0
2             1 0.6539             1             0             0
3             1 0.9660             1             0             0
4             1 0.4202             0             1             0
5             1 0.3609             0             1             0
6             1 0.4296             0             1             0
7             1 0.1340             0             0             1
8             1 0.8029             0             0             1
9             1 0.5959             0             0             1
10            1 0.8449            -1            -1            -1
11            1 0.1389            -1            -1            -1
12            1 0.9359            -1            -1            -1

```

Notice that the estimates of β_0 , β_2 , β_3 , and β_4 are different for different coding. However, the estimate of β_1 , which explains the main effect of X , is always the same.