Today's topic

- Contrast coding for categorical variables
  - Dummy coding
  - Deviation coding
  - Orthogonal coding
  - Polynomial contrasts

Example: High school and beyond survey

Two hundred observations were randomly sampled from the High School and Beyond survey, a survey conducted on high school seniors by the National Center of Education Statistics.

Response: write (standardized writing score)

Predictors: - race (four levels, Hispanic, Asian, African American, Caucasian) - readcat (category for standardized reading score)

id	write	gender	race	read	science	social science	readcat
1	70	male	white	57	52	41	(52,64]
2	121	female	white	68	59	53	(64,76]
3	86	male	white	44	33	54	(40,52]
...	...	...	...	...	...	...	...

Example: Dummy coding

Compares each level of the categorical variable to a fixed reference level.

Example: 4 treatments, each with n replicates.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, 4, \quad j = 1, \ldots, n$$

Dummy coding: $\alpha_1 = 0$.

Level of race	race.f1 (1 vs. 2)	race.f2 (1 vs. 3)	race.f3 (1 vs. 4)
1 (Hispanic)	0	0	0
2 (Asian)	1	0	0
3 (African American)	0	1	0
4 (Caucasian)	0	0	1

```
#the contrast matrix for categorical variable with four levels
contr.treatment(4)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1

#assigning the treatment contrasts to race.f
contrasts(hsb2$race.f) = contr.treatment(4)
#the regression
summary(lm(write ~ race.f, hsb2))

Residuals:
    Min     1Q Median 3Q  Max
 -23.06 -5.458 0.9724  7 18.8

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 46.4583  1.8422    25.2184  0.0000
    race.f2 11.5417  3.2861     3.5122  0.0006
    race.f3  1.7417  2.7325     0.6374  0.5246
    race.f4  7.5968  1.9889     3.8197  0.0002
```

Example: Deviation coding

Compares each level of the categorical variable to the grand mean.

Example: 4 treatments, each with n replicates.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, 4, \quad j = 1, \ldots, n$$

Deviation coding: $\sum_i \alpha_i = 0$.

Level of race	Level 1 v. Mean	Level 2 v. Mean	Level 3 v. Mean
1 (Hispanic)	1	0	0
2 (Asian)	0	1	0
3 (African American)	0	0	1
4 (Caucasian)	-1	-1	-1

```
#the contrast matrix for categorical variable with four levels
contr.sum(4)
  [,1] [,2] [,3]
1    1    0    0
2    0    1    0
3    0    0    1
4   -1   -1   -1

#assigning the deviation contrasts to race.f
contrasts(hsb2$race.f) = contr.sum(4)
#the regression
summary(lm(write ~ race.f, hsb2))

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept) 51.6784    0.9821   52.6191   0.0000
    race.f1 -5.2200    1.6314   -3.1997   0.0016
    race.f2  6.3216    2.1603    2.9263   0.0038
    race.f3 -3.4784    1.7323   -2.0079   0.0460
```

Equivalent forms of the model

- Treatment means model

$$y_{jk} = \mu_j + \epsilon_{jk},$$

where μ_j is j-th treatment mean and ϵ_{jk} represents within treatment variation (error).

Equivalent forms of the model

- Treatment means model

$$y_{jk} = \mu_j + \epsilon_{jk},$$

where μ_j is j-th treatment mean and ϵ_{jk} represents within treatment variation (error).

- Treatment difference model

$$y_{jk} = \mu + \alpha_j + \epsilon_{jk},$$

where μ is the *grand mean*, α_j represents j-th treatment effect compared to the grand mean.

For identifiability, we set $\sum_j \alpha_j = 0$.

Do treatment means and treatment difference models represent different models?

Equivalent forms of the model

- Treatment means model

$$y_{jk} = \mu_j + \epsilon_{jk},$$

where μ_j is j-th treatment mean and ϵ_{jk} represents within treatment variation (error).

- Treatment difference model

$$y_{jk} = \mu + \alpha_j + \epsilon_{jk},$$

where μ is the *grand mean*, α_j represents j-th treatment effect compared to the grand mean.

For identifiability, we set $\sum_j \alpha_j = 0$.

Do treatment means and treatment difference models represent different models?

No, they are two different parametrizations of the same model.

$$\mu_j = \mu + \alpha_j \Leftrightarrow \alpha_j = \mu_j - \mu$$

Null hypothesis for the treatment effects

What is an appropriate null hypothesis in terms of α's?

Null hypothesis for the treatment effects

What is an appropriate null hypothesis in terms of α's?

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_t = 0$$

Data decomposition approach:

Since $y_{jk} = y_{..} + (y_{j.} - y_{..}) + (y_{jk} - y_{j.})$ [Show], the model can be estimated by

$$y_{jk} = \hat{\mu} + \hat{\alpha}_j + \hat{e}_{jk},$$

where

$$\hat{\mu} = y_{..}, \quad \hat{\alpha}_j = y_{j.} - y_{..} \quad \hat{\epsilon}_{jk} = y_{jk} - y_{j.}.$$

Null hypothesis for the treatment effects

What is an appropriate null hypothesis in terms of α's?

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_t = 0$$

Data decomposition approach:

Since $y_{jk} = y_{..} + (y_{j.} - y_{..}) + (y_{jk} - y_{j.})$ [Show], the model can be estimated by

$$y_{jk} = \hat{\mu} + \hat{\alpha}_j + \hat{e}_{jk},$$

where

$$\hat{\mu} = y_{..}, \quad \hat{\alpha}_j = y_{j.} - y_{..} \quad \hat{\epsilon}_{jk} = y_{jk} - y_{j.}.$$

These estimates imply that $\hat{\mu} = n^{-1} \sum_j n_j \hat{\mu}_j$ and $\sum_{j=1}^{t} n_j \hat{\alpha}_j = 0$.

$\Longrightarrow$ The treatment effect $\alpha_j = \mu_j - \mu$ is the difference between the j-th treatment mean and the weighted mean.

Orthogonal contrasts

- A constrast in the treatment means is defined as $L = \sum_j c_j \mu_j$ where $\sum_j c_j = 0$.
- Two contrasts $L_1 = \sum_j a_j \mu_j$ and $L_2 = \sum_j b_j \mu_j$ are said to be orthogonal if $\sum_j a_j b_j = 0$.
- If the design is balanced $(n_1 = \cdots n_t = n_0)$, the estimated contrasts are uncorrelated, because

$$
\begin{aligned}
\mathrm{Cov}\left(\hat{L}_1, \hat{L}_2\right) &= \mathrm{Cov}\left(\sum_j a_j y_{j.}, \sum_j b_j y_{j.}\right) \\
&= E\left[\sum_j \sum_{j'} a_j (y_{j.} - \mu_j) b_{j'} (y_{j'.} - \mu_{j'})\right] \\
&= n_0^{-1} \sum_j a_j b_j \sigma^2 \\
&= 0
\end{aligned}
$$

Orthogonal contrasts

- If the y_{jk} are independent, then the two contrasts $\hat{L}_1 = \sum_j a_j y_{j.}$ and $\hat{L}_2 = \sum_j b_j y_{j.}$ are uncorrelated if and only if $\sum_j a_j b_j / n_j = 0$.
- We refer to contrasts satisfying $\sum_j a_j b_j / n_j = 0$ as weighted orthogonal contrasts.

Mutually orthogonal contrasts

Consider the following set of contrasts

$$
\begin{aligned}
L_1 &= l_{11}\mu_1 + l_{12}\mu_2 + \cdots + l_{1t}\mu_t \\
L_2 &= l_{21}\mu_1 + l_{22}\mu_2 + \cdots + l_{2t}\mu_t \\
\cdots & \\
L_{t-1} &= l_{(t-1)1}\mu_1 + l_{(t-1)2}\mu_2 + \cdots + l_{(t-1)t}\mu_t
\end{aligned}
$$

This set is called a set of mutually orhogonal contrasts if each contrast in the set is orthogonal to any other contrast.

$$\sum_{j=1}^{t} l_{k_1,j} l_{k_2,j} = 0, \quad \forall k_1, k_2.$$

Mutually orthogonal contrasts

- The maximum number of contrasts in a set of mutually orthogonal contrasts is $t-1$.
- A set of $t-1$ mutually orthogonal contrasts is called a complete set of orthogonal contrasts.

Mutually orthogonal contrasts

This set is called a set of mutually orthogonal contrasts if each contrast in the set is orthogonal to any other contrast.

$$\sum_{j=1}^{t} l_{k_1,j} l_{k_2,j} = 0, \quad \forall k_1, k_2.$$

- The maximum number of contrasts in a set of mutually orthogonal contrasts is $t - 1$.
- A set of $t - 1$ mutually orthogonal contrasts is called a complete set of orthogonal contrasts.

In the example, which sets are complete set of orthogonal contrasts?

- In general, for t treatments, there exists infinitely many complete sets of $t - 1$ orthogonal contrasts, but only few are useful for interpretation.

Quantitative treatments: Dose-Response modeling

Treatments: Doses of a drug; fertilizer amounts.

Reexpress treatment means as a function of dose z_j:
$\mu + \alpha_j = f(z_j; \theta)$.

Commonly used forms of f are polynomials in the dose z_j.

$$\mu + \alpha_j = \theta_0 + \theta_1 z_j + \theta_2 z_j^2 + \cdots + \theta_{t-1} z_j^{t-1}.$$

Why up to $t - 1$?

Quantitative treatments: Why are polynomials useful?

- Potential reduction in the model complexity.
- Prediction at treatment values not included in the design.
- How to decide the order?

Nested sequence of F-tests

$$
\begin{aligned}
\mathcal{M}_0 &: \theta_0 \\
\mathcal{M}_1 &: \theta_0 + \theta_1 z_j \\
\mathcal{M}_2 &: \theta_0 + \theta_1 z_j + \theta_1 z_j^2 \\
\vdots \quad &\quad \vdots \\
\mathcal{M}_{t-1} &: \theta_0 + \theta_1 z_j + \theta_1 z_j^2 + \cdots + \theta_{t-1} z_j^{t-1}
\end{aligned}
$$

SSR_k: residual sum of squares for the model that includes powers up to k, for $k = 0, \cdots, t-1$.

$SSR_{t-1} = ?$

$$
\begin{aligned}
SS_{linear} &= SS_1 = SSR_0 - SSR_1 \\
SS_{quadratic} &= SS_2 = SSR_1 - SSR_2
\end{aligned}
$$

What is a potential problem?

```r
n <- 10
set.seed(1)
x <- rep(c(1:8), each = n)
y <- 1 + 1.2*x+ 0.5*x^2+0.2*x^3 + rnorm(n, 0, 2)
show(y[1:10])
```

```
##  [1] 1.647092 3.267287 1.228743 6.090562 3.559016 1.259063 3.874858 4.376649
##  [9] 4.051563 2.289223
```

```r
show(x)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4
## [39] 4 4 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8
## [77] 8 8 8 8
```

```r
z1<-x
z2<-x^2
z3<-x^3
z4<-x^4
z5<-x^5
z6<-x^6
z7<-x^7
z <- cbind(z1, z2, z3, z4, z5, z6, z7)
```

What is a potential problem?

```r
cor(z)
```

```
##           z1        z2        z3        z4        z5        z6        z7
## z1 1.0000000 0.9761871 0.9318318 0.8865812 0.8456852 0.8099966 0.7791837
## z2 0.9761871 1.0000000 0.9876115 0.9627448 0.9348890 0.9076551 0.8823855
## z3 0.9318318 0.9876115 1.0000000 0.9929738 0.9778400 0.9597488 0.9411266
## z4 0.8865812 0.9627448 0.9929738 1.0000000 0.9956381 0.9857797 0.9734808
## z5 0.8456852 0.9348890 0.9778400 0.9956381 1.0000000 0.9971137 0.9903772
## z6 0.8099966 0.9076551 0.9597488 0.9857797 0.9971137 1.0000000 0.9980054
## z7 0.7791837 0.8823855 0.9411266 0.9734808 0.9903772 0.9980054 1.0000000
```