Understanding and Implementing Contrast Coding

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Assumptions in dummy coding

Contrast coding for increased statistical power

Effect coding using Simple coding

Backward difference and Helmert coding to capture linear effects of categories

Quadratic, cubic, and polynomial trends using Orthogonal Polynomial Coding

Contrast Coding

Dummy Variable Trap

If a categorical variable is used as a feature (x-variable) in linear regression

And if that categorical variable has k levels

Trap: Using k dummy variables and an intercept

Causes multi-collinearity and an unstable regression model

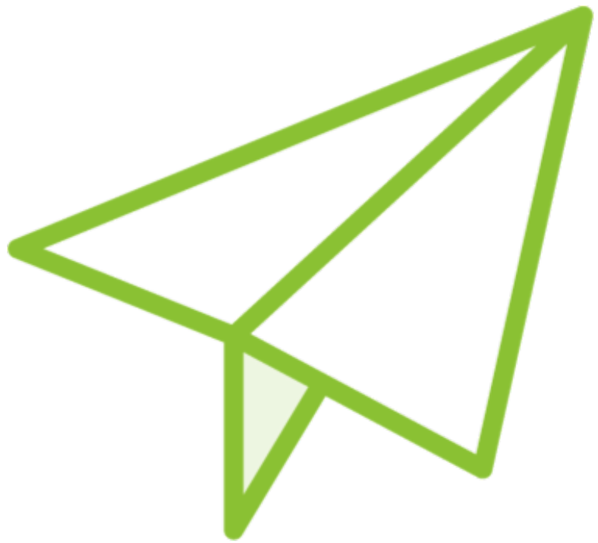
Avoiding the Dummy Variable Trap

Use either

- k dummy variables and exclude the intercept
- k-1 dummy variables and include the intercept

In either case, k levels need k variables (including the intercept)

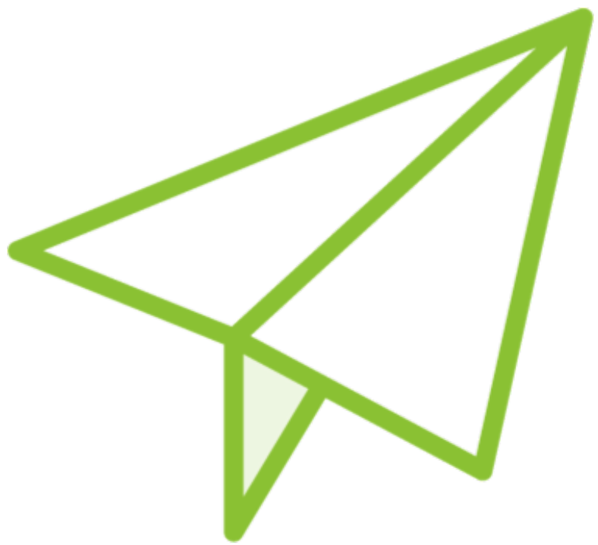
Dummy Coding



Name used for scheme with k-1 dummy variables along with intercept

Excluded level is called the reference level

Assumptions in Dummy Coding

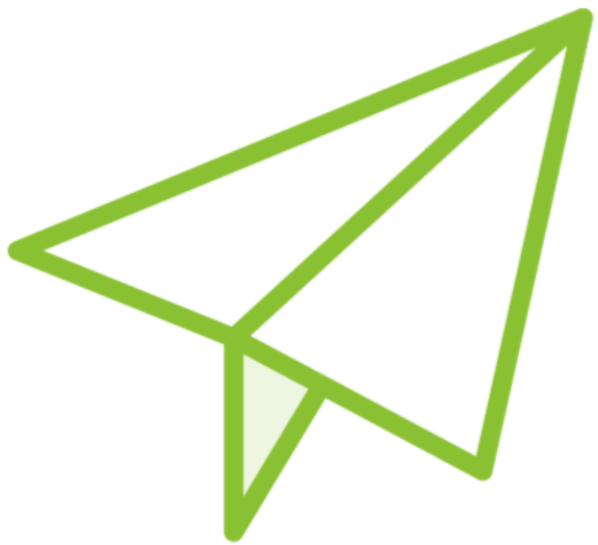


Dummy coding does not assume independence of coefficients

ANOVA makes assumptions about independence of coefficients but linear regression does not

Which is why dummy coding is most often used with linear regression

Assumptions in Dummy Coding

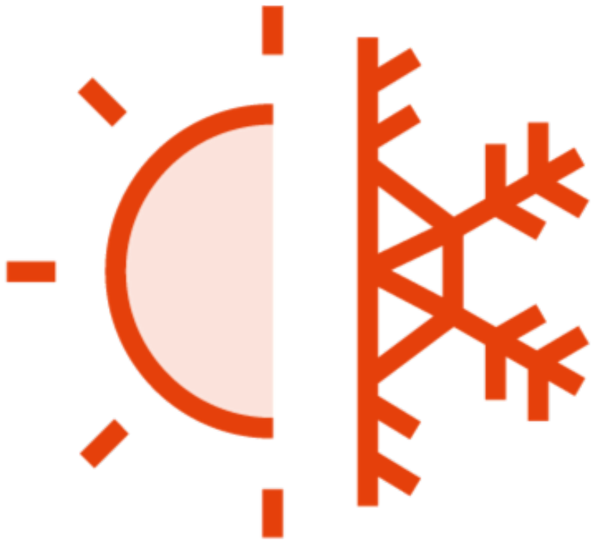


Dummy coding does not assume independence of coefficients

ANOVA makes assumptions about independence of coefficients but linear regression does not

Which is why dummy coding is most often used with linear regression

Contrast Coding



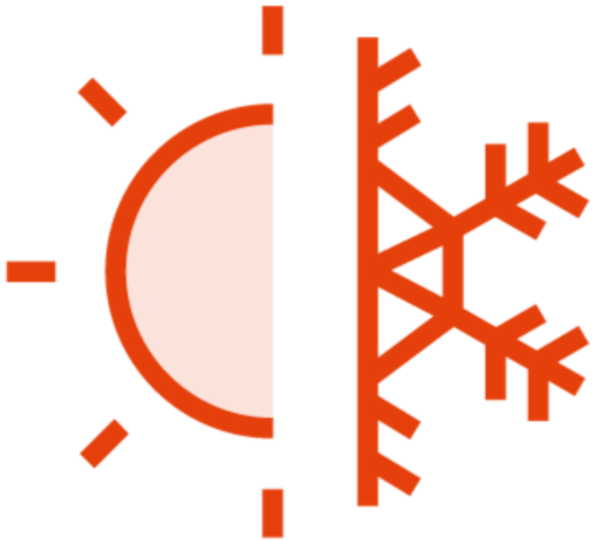
Contrast coding techniques bring out contrasts between different levels

Contrast coding for k levels is a set of k-1 functionally independent linear combinations of the reference level

Satisfies the independence of coefficients assumption of ANOVA

Contrast coding increases statistical power and provides more information about differences between categories

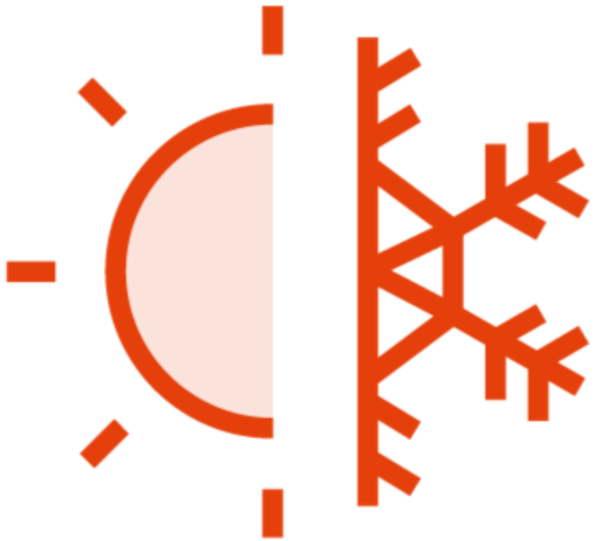
Contrast Coding



Use values other than 0 and 1 to encode categorical values

The codings for a variable should all sum to 0

Dummy vs. Contrast Coding



Dummy coding can be used for any kind of categorical variable

Contrast coding can bring out the linear effect of categories if they exist

Types of Contrast Coding

Simple	Backward Difference
Helmert	Orthogonal Polynomial

Other contrast coding techniques also exist

Types of Contrast Coding

Simple

Backward Difference

Helmert

Orthogonal Polynomial

Simple Coding with Linear Regression

Attribute	Attribute Details
Application	Compare other levels to reference
Intercept	Mean of y-values across the entire dataset
Coefficient for level(i)	Mean of y-values of level(i) - mean of y-value for reference level

Types of Contrast Coding

Simple

Backward Difference

Helmert

Orthogonal Polynomial

Backward Difference Coding with Linear Regression

Attribute	Attribute Details
Application	Compare mean of each level to mean of previous level
Intercept	Mean of y-values for all levels across the dataset
Coefficient for level(i)	Mean of y-values of level(i) - Mean of y-values for level(i-1)

Backward Difference Coding with Linear Regression

Attribute	Attribute Details
Application	Compare mean of each level to mean of previous level
Intercept	Mean of y-values for all levels across the dataset
Coefficient for level(i)	**Mean of y-values of level(i) - Mean of y-values for level(i-1)**

Useful with ordinal data, when the levels of the categorical variable are ordered in a meaningful way

Types of Contrast Coding

Simple

Backward Difference

Helmert

Orthogonal Polynomial

Helmert Coding with Linear Regression

Attribute	Attribute Details
Application	Compare mean of each level to mean of all previous levels (ordinal data only)
Intercept	Mean of means of y-values for all levels (mean of category means)
Coefficient for level(i)	Mean of y-values of level(i) - Mean of y-values for all levels up to level(i-1)

Helmert Coding with Linear Regression

Attribute	Attribute Details
Application	Compare mean of each level to mean of all previous levels (ordinal data only)
Intercept	Mean of means of y-values for all levels (mean of category means)
Coefficient for level(i)	**Mean of y-values of level(i) - Mean of y-values for all levels up to level(i-1)**

Types of Contrast Coding

Simple

Backward Difference

Helmert

Orthogonal Polynomial

Orthogonal Polynomial Coding with Linear Regression

Attribute	Attribute Details
Application	Polynomial trend in equally spaced, numeric variable
Intercept	Mean of means of y-values for all levels (mean of category means)
Coefficients	Capture linear, quadratic, and cubic effects in variable

Demo

Encoding categories using simple effect encoding

Demo

Encoding categories using backward difference encoding

Demo

Encoding categories using backward Helmert encoding

Demo

Encoding categories using backward Orthogonal Polynomial encoding

Summary

Assumptions in dummy coding

Contrast coding for increased statistical power

Effect coding using Simple coding

Backward difference and Helmert coding to capture linear effects of categories

Quadratic, cubic, and polynomial trends using Orthogonal Polynomial Coding