Regression with Categorical Variables

General Principles

To study the relationship between a categorical independent variable and a continuous dependent variable, we use a *Categorical model* which applies *stratification*.

Stratification involves modeling how the k different categories of the independent variable affect the target continuous variable by performing a regression for each k category and assigning a regression coefficient for each category. To implement stratification, categorical variables are often encoded using one-hot encoding or by converting categories to indices .

Considerations

> Caution
>
> - We have the same considerations as for Regression for a Continuous Variable.
> - As we generate regression coefficients for each k category, we need to specify a prior with a shape equal to the number of categories k in the code (see comments in the code).
> - To compare differences between categories, we need to compute the distribution of the differences between categories, known as the contrast distribution. **Never compare confidence intervals or p-values directly**.

Example

Below is an example of code that demonstrates Bayesian regression with an independent categorical variable using the Bayesian Inference (BI) package. The data consist of one continuous dependent variable (*kcal_per_g*), representing the caloric value of milk per gram, and a categorical independent variable, representing species clade membership. The goal is to estimate the differences in milk calories between clades. This example is based on McElreath (2018).

Python

```python
from main import*

# Setup device--------------------------------------------------
m = bi(platform='cpu')

# Import Data & Data Manipulation ------------------------------------------------
# Import
from importlib.resources import files
data_path = files('BI.resources.data') / 'milk.csv'
m.data(data_path, sep=';')
m.index(["clade"]) # Manipulate
m.scale(['kcal_per_g']) # Scale
m.data_to_model(['kcal_per_g', "index_clade"]) # Send to model (convert to jax array)

# Define model -------------------------------------------------
def model(kcal_per_g, index_clade):
    a = m.bi.dist.normal(0, 0.5, shape=(4,), name = 'a') # shape based on the number of clade
    s = m.bi.dist.exponential( 1, name = 's')
    mu = a[index_clade]
    m.normal(mu, s, obs=kcal_per_g)


# Run mcmc -----------------------------------------------
m.fit(model) # Optimize model parameters through MCMC sampling

# Summary ------------------------------------------------
m.summary()
```

R

```r
library(BI)
m=importbi(platform='cpu')

# Load csv file
m$data(paste(system.file(package = "BI"),"/data/milk.csv", sep = ''), sep=';')
m$scale(list('kcal.per.g')) # Manipulate
m$index(list('clade')) # Scale
m$data_to_model(list('kcal_per_g', 'index_clade')) # Send to model (convert to jax array)
```

```r
# Define model ------------------------------------------------
model <- function(kcal_per_g, index_clade){
  # Parameter prior distributions
  beta = bi.dist.normal( 0, 0.5, name = 'beta', shape = c(4))  # shape based on the number o
  sigma =bi.dist.exponential(1, name = 's')
  # Likelihood
  m$normal(beta[index_clade], sigma, obs=kcal_per_g)
}

# Run mcmc ----------------------------------------------------
m$run(model) # Optimize model parameters through MCMC sampling

# Summary -----------------------------------------------------
m$summary() # Get posterior distributions
```

Mathematical Details

Frequentist formulation

We model the relationship between the categorical input feature (X) and the target variable (Y) using the following equation:

$$Y_i = \alpha + \beta_k X_i + \sigma$$

Where:

- Y_i is the dependent variable for observation i.
- α is the intercept term.
- β_k are the regression coefficients for each k category.
- X_i is the encoded categorical input variable for observation i.
- σ is the error term.

We can interpret β_i as the effect of each category on Y relative to the baseline (usually one of the categories or the intercept).

Bayesian formulation

In the Bayesian formulation, we define each parameter with priors . We can express the Bayesian regression model accounting for prior distributions as follows:

$$Y \sim Normal(\alpha + \beta_k X, \sigma)$$

$$\alpha \sim Normal(0, 1)$$

$$\beta_k \sim Normal(0, 1)$$

$$\sigma \sim Exponential(1)$$

Where:

- Y_i is the dependent variable for observation i.
- α is the prior distribution for the intercept.
- β_k are k prior distributions for k regression coefficients.
- X_i is the encoded categorical input variable for observation i.
- σ is the prior distribution for the standard deviation, ensuring it is positive.

Notes

> **Note**
>
> - We can apply multiple variables similarly to Chapter 2: Multiple Continuous Variables.
> - We can apply interaction terms similarly to Chapter 3: Interaction between Continuous Variables.

Reference(s)

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in r and Stan*. Chapman; Hall/CRC.