Regression with a Categorical Independent Variable

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

> **Note**
>
> - 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 BayesForge (BF) package. The data consist of one continuous dependent variable (*kcal_per_g*), representing the caloric value of milk per gram, a categorical

independent variable (*index_clade*), representing species clade membership, and a continuous independent variable (*mass*), representing the mass of individuals in the clade. The goal is to estimate the differences in milk calories between clades. This example is based on McElreath (2018).

Python

```
from BayesForge import bf

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

# Import Data & Data Manipulation ------------------------------------------------
# Import
from importlib.resources import files
data_path = m.load.milk(only_path = True)
m.data(data_path, sep=';')
m.index(["clade"]) # Convert clade names into index
m.scale(['kcal_per_g']) # Scale

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


# Run mcmc ----------------------------------------------------
m.fit(model, progress_bar=False) # Optimize model parameters through MCMC sampling

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

```
bf v 0.0.48 package loaded

E0526 16:18:46.428920 1067541 cuda_dnn.cc:523] Loaded runtime CuDNN library: 9.1.0 but source
E0526 16:18:46.430698 1067541 cuda_dnn.cc:523] Loaded runtime CuDNN library: 9.1.0 but source

jax.local_device_count 32
```

	mean	sd	hdi_5.5%	hdi_94.5%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
a[0]	-0.34	0.33	-0.83	0.22	0.01	0.00	2866.52	2920.12	1.0
a[1]	0.60	0.29	0.15	1.08	0.01	0.00	3235.89	2685.49	1.0
a[2]	0.31	0.39	-0.29	0.94	0.01	0.01	2643.88	2599.97	1.0
a[3]	-0.17	0.44	-0.89	0.53	0.01	0.01	2852.61	2653.21	1.0
b[0]	-0.00	0.01	-0.02	0.01	0.00	0.00	3096.36	2876.63	1.0
b[1]	-0.18	0.12	-0.36	0.01	0.00	0.00	3178.46	2944.88	1.0
b[2]	0.08	0.07	-0.02	0.19	0.00	0.00	2677.26	2456.24	1.0
b[3]	-0.28	0.25	-0.68	0.11	0.00	0.00	2855.26	2808.28	1.0
s	0.77	0.12	0.59	0.96	0.00	0.00	2859.19	2708.14	1.0

R

```r
library(BayesForge)
m=importBF(platform='cpu')

# Load csv file
m$data(m$load$milk(only_path = T), sep=';')
m$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)

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

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

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

Julia

```julia
using BayesForge

# Setup device----------------------------------------------
m = importBF(platform="cpu")

# Import Data & Data Manipulation ---------------------------------------------
# Import
data_path = m.load.milk(only_path = true)
m.data(data_path, sep=';')
m.index("clade") # Convert clade names into index
m.scale(["kcal_per_g"]) # Scale

# Define model ----------------------------------------------
@BF function model(kcal_per_g, index_clade, mass)
    a = m.dist.normal(0, 0.5, shape=(4,), name = "a") # shape based on the number of clades
    b = m.dist.normal(0, 0.5, shape=(4,), name = "b")
    s = m.dist.exponential( 1, name = 's')
    mu = a[index_clade]+b[index_clade]*mass
    m.dist.normal(mu, s, obs=kcal_per_g)
end

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

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

> **Caution**
>
> For R users, when working with indices you have to ensure 1) that indices are intergers (i.e. `as.integer(index_clade)`) and, 2) that indices start at 0 (i.e. `as.integer(index_clade)-1`).

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:

$$Y \sim \text{Normal}(\alpha + \beta_K X, \sigma)$$

$$\alpha \sim \text{Normal}(0, 1)$$

$$\beta_K \sim \text{Normal}(0, 1)$$

$$\sigma \sim \text{Exponential}(1)$$

Where:

- Y_i is the dependent variable for observation i.
- α is the intercept term, which in this case has a unit-normal prior.
- β_K are slope coefficients for the K distinct independent variables categories, which also have unit-normal priors.
- X_i is the encoded categorical input variable for observation i.
- σ is a standard deviation parameter, which here has a Exponential prior that constrains it to be positive.

Notes

> **Note**
>
> - As for Multiple continuous variables, we can extend the linear regression model to include multiple categorical predictors.
> - We can apply interaction terms similarly to Interaction between continuous variables.

Reference(s)

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian course with examples in R and Stan*. Chapman; Hall/CRC.