Prior Pred Dist

Daniel(Kailun) Jin, John Ju

Data

```
lf_data <- read_csv("Life Expectancy Data.csv")
```

```
Rows: 2938 Columns: 22
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (2): Country, Status
dbl (20): Year, Life expectancy, Adult Mortality, infant deaths, Alcohol, pe...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
names(lf_data) <- str_replace_all(names(lf_data), pattern = " ", replacement
= "_")
```

We found data in Kaggle website, and our data is about the life expectancy for each country from 2000 to 2015. Here is our link.

data source link

Plan

Research question: Dose the percentage.expenditure has association the life expectancy?

Response:`Life_expectancy`

Key predictor:`percentage_expenditure` Expenditure on health as a percentage of Gross Domestic Product per capital(%)

Level: Country

Confounder:

`GDP`: It can decide how much we can have in percentage_expenditure, also shows the total economic developing of the country which will affect the Life_expectancy.

`Schooling`: High schooling country may have high percentage_expenditure, and high schooling also related the level of health and living habit which will affect Life_expectancy.

`Status`: Status will influence the the percentage_expenditure, and usually developed country may have high Life_expectancy.

`Year`: Life_expectancy and percentage_expenditure will change through years, we need to consider the difference of it.

Mediators

Immunization Rates(Hepatitis B, Polio, Diphtheria, Measles,HIV/AIDS):Increased health expenditure (percentage expenditure) typically enhances public health initiatives, leading to higher immunization coverage. This reduces the prevalence of infectious diseases and subsequently contributes to increased life expectancy. These indicators can help explain part of the indirect effect of health expenditure on longevity.

Infant deaths / under-five deaths: The percentage_expenditure will affect the infant deaths and under-five deaths, and them will affect the Life_expectancy.

Moderators `Status`: Different satuts of county may have different result in the same amount of percentage_expenditure.

Collider Total_expenditure: it will affect both by GDP and government.

Adult_Mortality: It will affected by Life_expectancy and other health behavior.

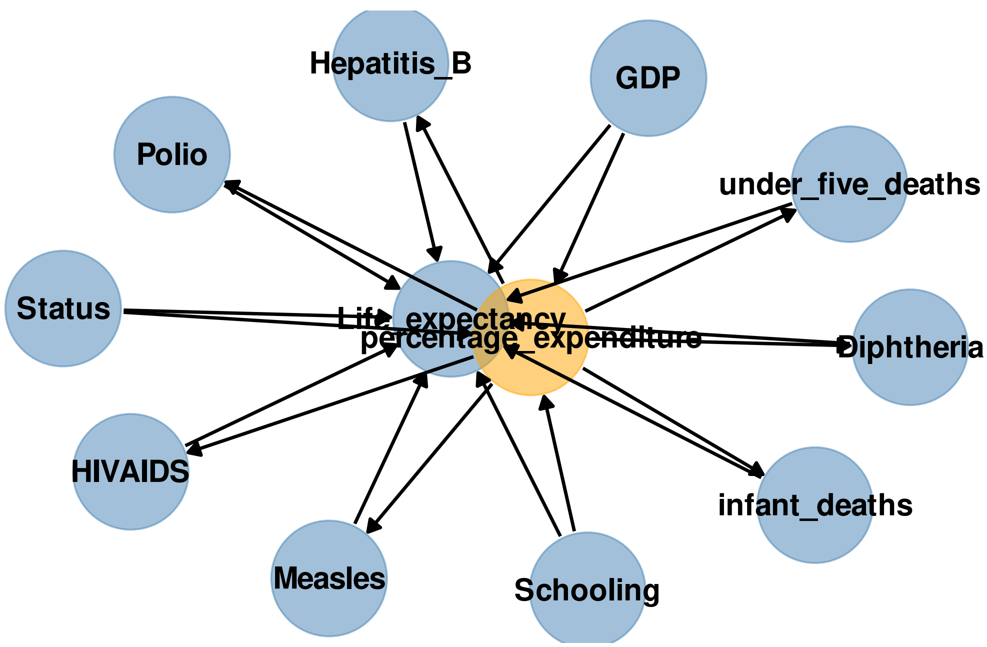
Other

`Population`: Total population in each country each year, may include.

Casual Diagrams

```
lfcd <- dagitty("dag {
  percentage_expenditure -> Life_expectancy

  GDP -> percentage_expenditure
  GDP -> Life_expectancy

  Schooling -> percentage_expenditure
  Schooling -> Life_expectancy

  Status -> percentage_expenditure
  Status -> Life_expectancy

  percentage_expenditure -> Hepatitis_B
  percentage_expenditure -> Polio
  percentage_expenditure -> Measles
  percentage_expenditure -> Diphtheria
  percentage_expenditure -> HIVAIDS

  Hepatitis_B -> Life_expectancy
  Polio -> Life_expectancy
  Measles -> Life_expectancy
```

```
  Diphtheria -> Life_expectancy
  HIVAIDS -> Life_expectancy

  percentage_expenditure -> infant_deaths
  percentage_expenditure -> under_five_deaths
  infant_deaths -> Life_expectancy
  under_five_deaths -> Life_expectancy
}")



gg_dag(lfcd,

       size = 20,

       highlight = 'percentage_expenditure')
```



We will use predictor: percentage_expenditure, GDP, Schooling, Status.

Check the Null value in the data set.(unsure what we want to do yet)

```
lf_data |>
  # for every variable,
  summarise(across(everything(),
                   # add up the number of missing values
                   ~ sum(is.na(.))))
```

```
# A tibble: 1 × 22
  Country  Year Status Life_expectancy Adult_Mortality infant_deaths Alcohol
    <int> <int>  <int>           <int>           <int>         <int>   <int>
1       0     0      0              10              10             0     194
# ℹ 15 more variables: percentage_expenditure <int>, Hepatitis_B <int>,
#   Measles <int>, BMI <int>, `under-five_deaths` <int>, Polio <int>,
#   Total_expenditure <int>, Diphtheria <int>, `HIV/AIDS` <int>, GDP <int>,
#   Population <int>, `thinness__1-19_years` <int>, `thinness_5-9_years` <int>,
#   Income_composition_of_resources <int>, Schooling <int>
```

Prior Predictive Distribution

```
zlf_data <- lf_data |>
  # scaled the predictor
  mutate(
    percentage_expenditure_scaled = as.numeric(scale(percentage_expenditure)),
    GDP_scaled = as.numeric(scale(GDP)),
    Schooling_scaled = as.numeric(scale(Schooling)),
    Status = factor(Status),
    Status_ix = as.numeric(Status))|>
  # drop any rows with NAs
   select(percentage_expenditure_scaled, GDP_scaled, Schooling_scaled, Status,
Status_ix)|>
  drop_na()
```

Model Description

Likelihood:

$$\text{Life\_expectancy}_i \sim \text{Gamma}\ (\mu_i, \sigma)$$

$$log(\mu\_i) = \beta_0 + \beta_1 \text{ percentage\_expenditure\_scaled } \_i$$

$$+ \beta_2 \text{ GDP\_scaled } \_i + \beta_3 \text{ Schooling\_scaled } \_i + \beta_4 \ [\text{Status\_ix}_i]$$

Priors:

To compress the gamma funtion, we will choose a relatively large sigma. We chose 5

$$\sigma \sim \text{Exponential}\ (5)$$

From the WHO we can know the the global Life_expectancy is about 71.4 in 2021, log(71.4) is around 4.27, let's just take 4, and standard deviation take 0.2. Sources Link

$$\beta_0 \sim \text{Normal}\ (4, 0.2)$$

For other prior because we don't know how each value affect on the Life_expectancy so we just use mean as 0, and sd as 0.2, exp(0.1) is around 1.11. So that each of them will not change it a lot.

$$\beta_1 \sim \text{Normal}\ (0, 0.1)$$

$$\beta_2 \sim \text{Normal}\ (0, 0.1)$$

$$\beta_3 \sim \text{Normal}\ (0, 0.1)$$

$$\beta_3 \sim \text{Normal}\ (0, 0.1)$$

$$\beta_4 \sim \text{Normal}\ (0, 0.1)$$

Parameter transformations:

$$\alpha = \frac{\mu^2}{\sigma^2}$$

$$\lambda = \frac{\mu}{\sigma^2}$$

```
n_sim <- 100
prior_pred_dist <- tibble(
  sim_id = c(1:n_sim)) |>
  mutate(
    b0 = rnorm(n_sim, mean = 4, sd = 0.2),         # intercept
   b1 = rnorm(n_sim, mean = 0, sd = 0.1),       # percentage_expenditure_scaled
    b2 = rnorm(n_sim, mean = 0, sd = 0.1),       # GDP_scaled
    b3 = rnorm(n_sim, mean = 0, sd = 0.1),       # Schooling_scaled
    b4 = rnorm(n_sim, mean = 0, sd = 0.1),       # Status_ix (binary: 0 or 1)
    sigma = rexp(n_sim, rate = 5)                 # dispersion
  ) |>
  rowwise()|>
  mutate(
    mu = list(exp(
      b0 +
      b1 * zlf_data$percentage_expenditure_scaled +
      b2 * zlf_data$GDP_scaled +
      b3 * zlf_data$Schooling_scaled +
      b4 * zlf_data$Status_ix
    )),
                                       percentage_expenditure_scaled          =
list(zlf_data$percentage_expenditure_scaled),
    GDP_scaled = list(zlf_data$GDP_scaled),
    Schooling_scaled = list(zlf_data$Schooling_scaled),
    Status_ix = list(zlf_data$Status_ix)
  ) |>
      unnest(cols    =    c(mu,    percentage_expenditure_scaled,    GDP_scaled,
Schooling_scaled, Status_ix)) |>
  ungroup() |>
  mutate(
    alpha = mu^2 / sigma^2,
    lambda = mu / sigma^2
  ) |>
```

```r
  rowwise() |>
  mutate(
    sim_life = rgamma(1, shape = alpha, rate = lambda)
  ) |>
  ungroup()
```

```r
gf_dens(~sim_life, group = ~sim_id,
        data = prior_pred_dist) |>
  gf_labs(title = 'Simulated Life_expectancy\n(each line is one dataset)')|>
  gf_lims(x = c(0, 130))
```

```
Warning: Removed 1143 rows containing non-finite outside the scale range
(`stat_density()`).
```