Introduction to Bayesian Data Analysis

Lecture 6: Generalized Linear Model

Julia Haaf

Summer 2025

Background

- GLMs are a big deal in frequentist stats
    - Do you use `glm` or `lme`?!?
- In Bayesian stats, we are used of thinking about our entire model more deliberately, including which model of the data to use.
- Yet, choosing between probability distributions for the data and interpreting the results requires a bit more thought than with linear models.

Overview

1. An Example
2. Model
    - Generalized Linear Model
    - Model of the Data
    - Prior
3. Estimate the Model
    - Model in `brms`
    - Interpretation of the Results
4. Model fit
    - Posterior prediction
    - Model comparison

An Example

Can mental health be predicted by daily habits and stress?

```
mentalhealth.dat <-
read.csv("data/mental_health_dataset.csv")

knitr::kable(head(mentalhealth.dat[,
c(2,6,9,10,11,12)]))%>%
  kable_styling(font_size = 16)
```

Age	Mental_Health_Condition	Stress_Level	Sleep_Hours	Work_Hours	Physical_Activity_Hours
36	No	Medium	7.1	46	5
34	Yes	Low	7.5	47	8
65	Yes	Low	8.4	58	10
34	No	Medium	9.8	30	2
22	Yes	Medium	4.9	62	5
64	Yes	High	6.3	34	0

An Example

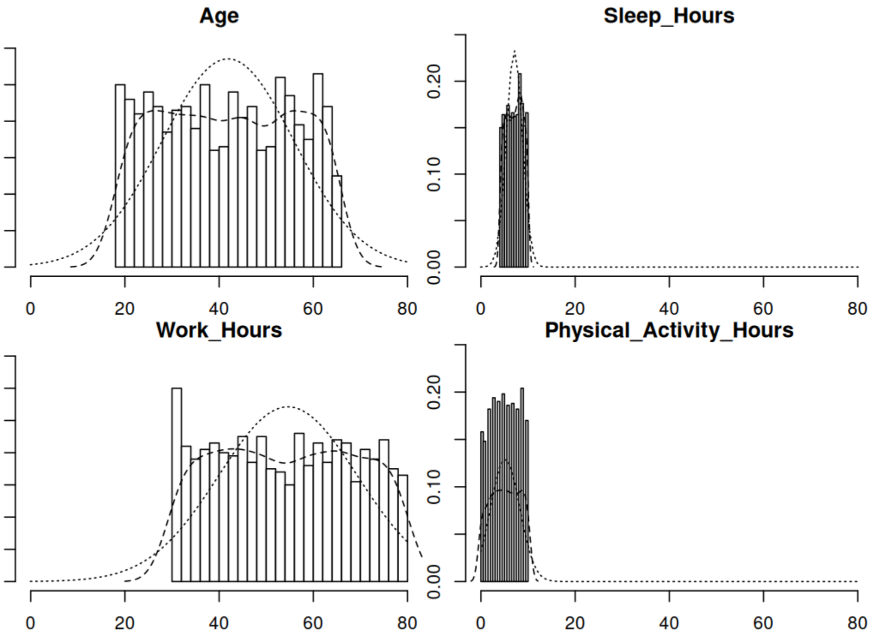
"Do you have a mental illness?"

```
table(mentalhealth.dat$Mental_Health_Condition)
```

```
##
##  No Yes
## 485 515
```

An Example

```
library("psych")
multi.hist(mentalhealth.dat[, c(2, 10:12)])
```



The Model

What does a meaningful model for this data look like?

The Model

- Linear regression:

$$Y_i \sim \text{Normal}(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2$$

- Extension:

$$Y_i \sim \text{Normal}(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,n}\beta_n$$

Problems?

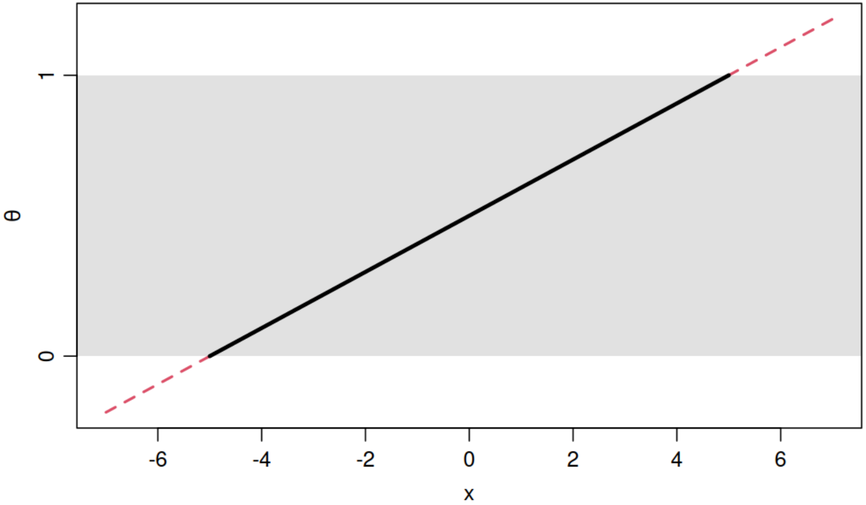
Y_i : Does the ith person have mental problems? $\rightarrow$ Yes/No (1/0)

Statistical Model

Y_i : Does the ith person have mental problems? → Yes/No (1/0)

- $Y_i \sim \text{Binomial}(n, \theta_i),$
- $0 > \theta > 1, n =?$

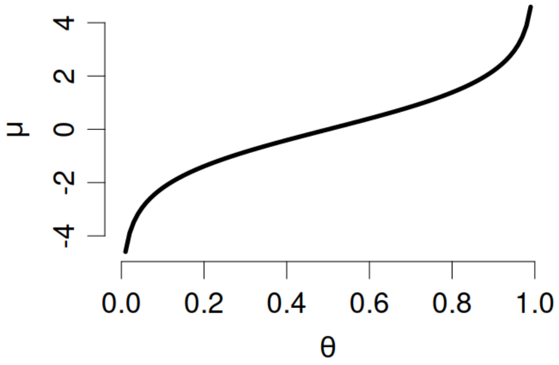
$$??? \; \theta_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,n}\beta_n \; ???$$



Generalized Linear Model

Link Function $g(\cdot)$

- Goal: Connect the linear model with the parameter to be estimated (here: probability)
- For $0, 1$ responses, the link function used is the logit transformation: $\mu_i = g(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right)$



Generalized Linear Model

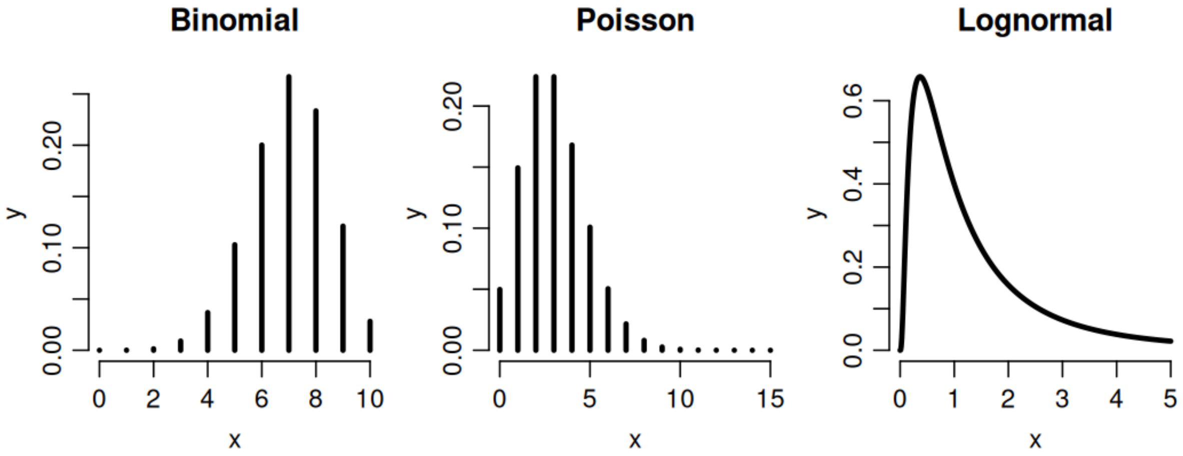
Using the link function, we go from the probability space into the "logit space"$\rightarrow$ μ can be between $-\infty$ and ∞.

Linear Model

$$\mu_i = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,n}\beta_n$$

Generalized Linear Model

The generalized linear model specifies the probability distribution of a random variable and a link function that allows flexible use of linear regression models.

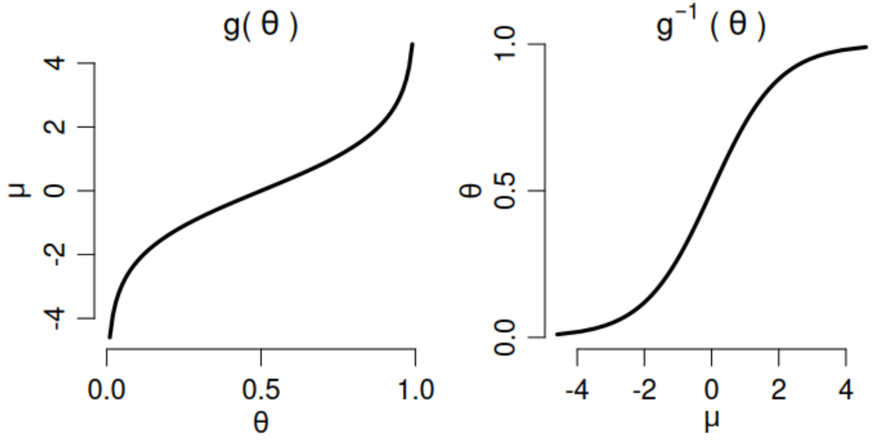


Generalized Linear Model

Inverse Function $g^{-1}(\cdot)$

- To get from μ_i back to θ_i we use the inverse of the logit function:

$$\theta_i = g^{-1}(\mu_i) = \frac{e^{\mu_i}}{1 + e^{\mu_i}}$$



Statistical Model

- $Y_i \sim \text{Binomial}(1, \theta_i)$,
- Predictors
  - $x_{i,1}$: Age
  - $x_{i,2}$: Stress level (high = 1, not high = 0)
  - $x_{i,3}$: Stress level (medium = 1, not medium = 0)
  - $x_{i,4}$: Sleep in hours
  - $x_{i,5}$: Work time per week in hours
  - $x_{i,6}$: Physical activity per week in hours
- $\mu_i = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,6}\beta_6$

Statistical Model

How Does the Model Work?

- $Y_i \sim \text{Binomial}(1, \theta_i)$,
- $\mu_i = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,6}\beta_6$
- Example:

$$\mu_i = -5 + x_{i,1}0.1 + x_{i,2}1 + x_{i,3}0.5 + x_{i,4}(-0.3) +$$
$$x_{i,5}0.1 + x_{i,6}(-0.2)$$

Statistical Model

$$\mu_i = -5 + x_{i,1}0.1 + x_{i,2}1 + x_{i,3}0.5 + x_{i,4}(-0.3) + x_{i,5}0.1 + x_{i,6}(-0.2)$$

- Age = 40
- Stress level = medium
- Sleep in hours = 7
- Work hours per week = 32
- Physical activity per week = 8

$$\mu_i = -5 + 40 \times 0.1 + 0 \times 1 + 1 \times 0.5 + 7 \times (-0.3) + 32 \times 0.1 + 8 \times (-0.2) = -1$$

Statistical Model

$$\mu_i = -5 + 40 \times 0.1 + 0 \times 1 + 1 \times 0.5 + 7 \times (-0.3) + 32 \times 0.1 + 8 \times (-0.2) = -1$$

The probability of mental problems is then $\theta_i = g^{-1}(-1) = \frac{e^{-1}}{1+e^{-1}} = 0.27$

Statistical Model

- Age = 50
- Stress level = high
- Sleep in hours = 5
- Work hours per week = 60
- Physical activity per week = 0

$$\mu_i = -5 + 50 \times 0.1 + 1 \times 1 + 0 \times 0.5 + 5 \times (-0.3) +$$
$$60 \times 0.1 + 0 \times (-0.2) = 4.5$$

The probability of mental problems is then

$$\theta_i = g^{-1}(4.5) = \frac{e^{4.5}}{1+e^{4.5}} = 0.99$$

Prior

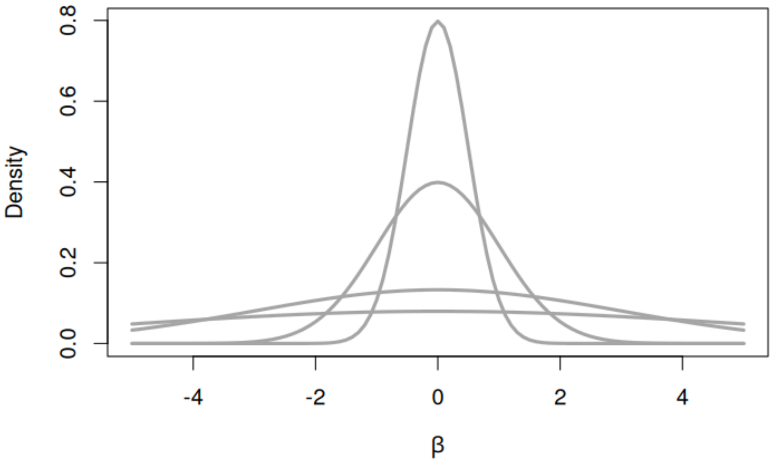
- $Y_i \sim \text{Binomial}(1, \theta_i)$,
- $\mu_i = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,6}\beta_6$

What parameters do we need in the model?

- All parameters!
  - $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$
  - Normal distribution
- What about σ?

Prior

- $Y_i \sim \text{Binomial}(1, \theta_i)$,
- $\mu_i = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,6}\beta_6$



- $x_{i,1}$: Age
- $x_{i,2}$: Stress level high
- $x_{i,3}$: Stress level med
- $x_{i,4}$: Sleep in hours
- $x_{i,5}$: Work hours
- $x_{i,6}$: Physical activity

Prior

- $Y_i \sim \text{Binomial}(1, \theta_i)$,
- $\mu_i = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,6}\beta_6$

```r
library(brms)

# default_prior(Mental_Health_Condition ~ Age +
#                  Stress_Level + Sleep_Hours + Work_Hours
#                  + Physical_Activity_Hours
#                , family = bernoulli(link = "logit")
#                , data = mentalhealth.dat)

bprior <- c(prior(normal(0, 3), class = Intercept)
            , prior(normal(0, 0.5), class = b))
```

Estimating the Model

Estimating the Model

```
model.1 <- brm(Mental_Health_Condition ~ Age +
                 Stress_Level + Sleep_Hours + Work_Hours
                 + Physical_Activity_Hours
               , family = bernoulli(link = "logit")
               , data = mentalhealth.dat
               , prior = bprior)
```

```
## Compiling Stan program...
```

```
## Trying to compile a simple C file
```

```
## Running /usr/lib/R/bin/R CMD SHLIB foo.c
## using C compiler: ‘gcc (Ubuntu 13.3.0-6ubuntu2~24.04) 13.3.0’
## gcc -I"/usr/share/R/include" -DNDEBUG   -
I"/home/juliahaaf/R/x86_64-pc-linux-gnu-library/4.4/Rcpp/include/"  -
I"/home/juliahaaf/R/x86_64-pc-linux-gnu-
library/4.4/RcppEigen/include/"  -I"/home/juliahaaf/R/x86_64-pc-linux-
gnu-library/4.4/RcppEigen/include/unsupported"  -I"/usr/lib/R/site-
library/BH/include" -I"/usr/lib/R/site-
library/StanHeaders/include/src/"  -I"/usr/lib/R/site-
library/StanHeaders/include/"  -I"/usr/lib/R/site-
library/RcppParallel/include/" -I/usr/include -DTBB_INTERFACE_NEW -
```

Estimating the Model

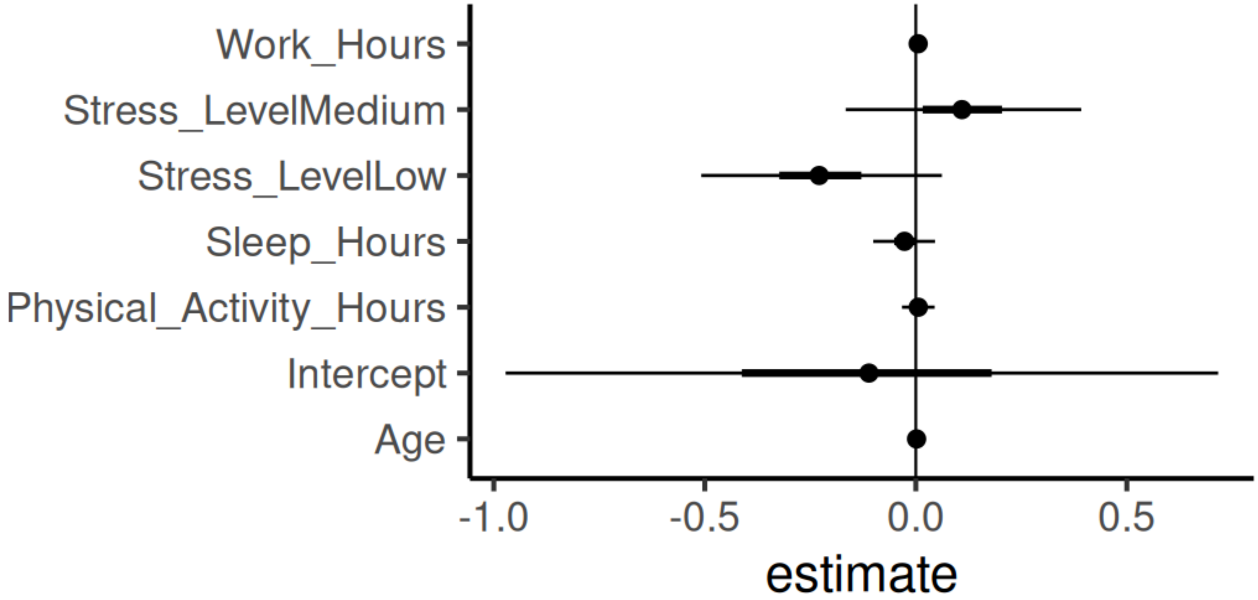
```
summary(model.1)
```

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: Mental_Health_Condition ~ Age + Stress_Level + Sleep_Hours
+ Work_Hours + Physical_Activity_Hours
##    Data: mentalhealth.dat (Number of observations: 1000)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Regression Coefficients:
##                          Estimate Est.Error l-95% CI u-95% CI Rhat
Bulk_ESS
## Intercept                   -0.12      0.44    -0.97     0.72 1.00
6260
## Age                          0.00      0.00    -0.01     0.01 1.00
5871
## Stress_LevelLow             -0.23      0.15    -0.51     0.06 1.00
4551
## Stress_LevelMedium           0.11      0.14    -0.17     0.39 1.00
4846
## Sleep_Hours                 -0.03      0.04    -0.10     0.05 1.00
6236
```

Estimating the Model

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-0.12	0.44	-0.97	0.72	1	6260.39	3361.95
Age	0.00	0.00	-0.01	0.01	1	5870.93	3008.00
Stress_LevelLow	-0.23	0.15	-0.51	0.06	1	4551.31	2957.68
Stress_LevelMedium	0.11	0.14	-0.17	0.39	1	4845.65	3399.84
Sleep_Hours	-0.03	0.04	-0.10	0.05	1	6236.45	3186.00
Work_Hours	0.01	0.00	0.00	0.01	1	6695.53	2986.99
Physical_Activity_Hours	0.01	0.02	-0.03	0.04	1	5656.69	3045.62

Estimating the Model

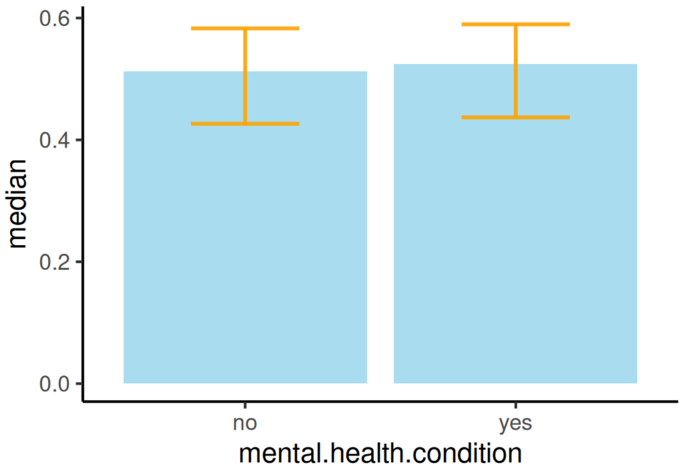


Posterior Prediction

Does the model accurately predict the response category for mental disorder?

```
post.pred <- posterior_predict(model.1) #as before
post.pred.mean <- colMeans(post.pred) #as before

# Median and prediction interval for y
post.pred.dat <- tapply(post.pred.mean
                        , mentalhealth.dat$Mental_Health_Condition
                        , quantile, probs = c(.025, .5, .975))
```



Posterior Prediction

```
# tapply gives a matrix, for ggplot we need a data frame
post.pred.dat <- as.data.frame(rbind(post.pred.dat$Yes
                                     , post.pred.dat$No))

# Naming the columns and rows
colnames(post.pred.dat) <- c("lower", "median", "upper")
post.pred.dat$mental.health.condition <- c("yes", "no")

ggplot(post.pred.dat) +
    geom_bar(aes(x = mental.health.condition
                 , y = median)
             , stat = "identity"
             , fill = "skyblue"
             , alpha = 0.7) +
    geom_errorbar(aes(x = mental.health.condition
                      , ymin=lower
                      , ymax=upper)
                  , width=0.4, colour="orange"
                  , alpha=0.9, linewidth=1.3)+
  theme_classic(base_size = 20)
```

Model Comparison

Is the model informative overall?

→ Comparison with a model without predictors

```
model.1 <- brm(Mental_Health_Condition ~ Age +
                 Stress_Level + Sleep_Hours + Work_Hours
                 + Physical_Activity_Hours
               , family = bernoulli(link = "logit")
               , data = mentalhealth.dat
               , prior = bprior
               , save_pars = save_pars(all = TRUE)
               , iter = 11000
               , warmup = 1000
               , silent = 2
               , refresh = 0)
```

```
## Running /usr/lib/R/bin/R CMD SHLIB foo.c
## using C compiler: ‘gcc (Ubuntu 13.3.0-6ubuntu2~24.04) 13.3.0’
## gcc -I"/usr/share/R/include" -DNDEBUG   -
I"/home/juliahaaf/R/x86_64-pc-linux-gnu-library/4.4/Rcpp/include/"  -
I"/home/juliahaaf/R/x86_64-pc-linux-gnu-
```

Model Comparison

```
bayes_factor(model.0, model.1)
```

```
## Warning: effective sample size cannot be calculated, has been
replaced by
## number of samples.

## Iteration: 1
## Iteration: 2
## Iteration: 3
## Iteration: 4
## Iteration: 1
## Iteration: 2
## Iteration: 3
## Iteration: 4

## Estimated Bayes factor in favor of model.0 over model.1:
1455990.93553
```

Model Comparison

- Strong evidence against effects of all predictors.
- Prevalence for a mental disorder cannot be predicted by age, life habits, and stress (for this dataset).
- Possible reasons?

:)