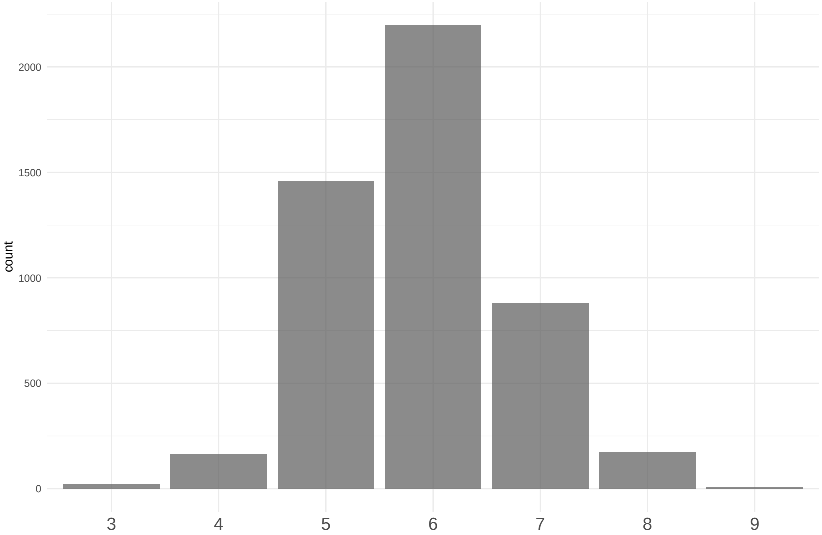
What makes wine great?

Yuga Hikida, Adya Maheshwari

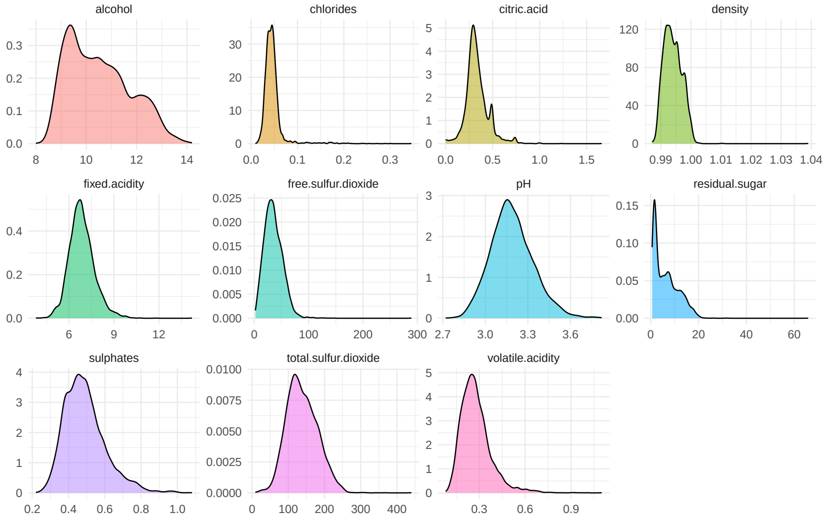
2024-01-02

Task

- Prediction of quality of (white) wine (from 1, 2,.. up to 10) using physicochemical variables.



Data: Preditictive variables



How to model "quality"?

- M_1: Categorical variable. *quality* $\in \{`1`, ..., `10`\}$

  $\Rightarrow$ Classification

- M_2: Continuous variable. *quality* $\in [1, 10]$

  $\Rightarrow$ Regression

- M_3: Ordered Categorical variable. *quality* $\in \{1, ..., 10\}$

  $\Rightarrow$ Ordinal Regression

For following slides, y for *quality* and X for (vector of) predictive variables.

M_1: Classification (1)

$$y \sim \text{categorical}(\psi_1, ..., \psi_C) \quad = \prod_{c=1}^{C} \psi_c^{I_{c(y)}}$$

where C is the number of categories ($C = 7$ for our case), $\psi_c = Pr(y = c)$ such that $\sum_{c=1}^{C} \psi_c^{I_c(y)} = 1$, and

$$I_{c(y)} = \begin{cases} 1 & y = c \\ 0 & \text{otherwise} \end{cases}$$

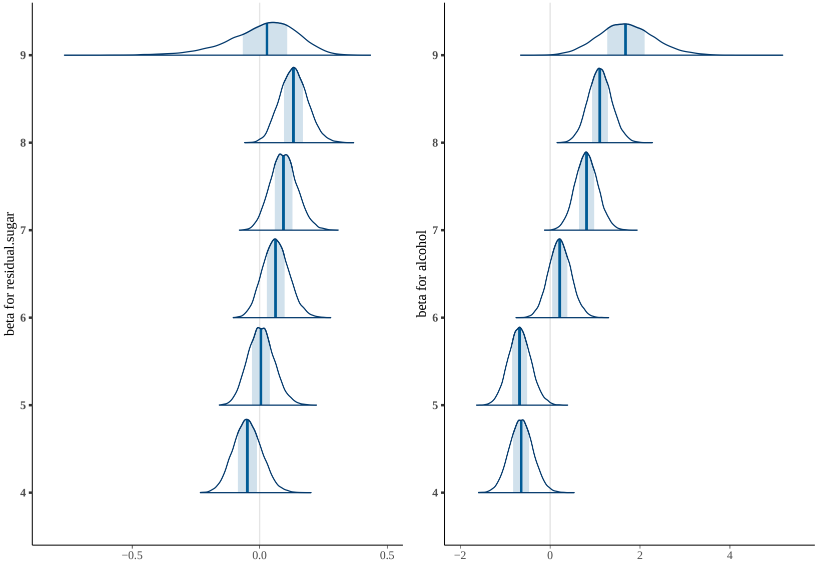
M_1: Classification (2)

For $c = 1, .., C$:

$$\begin{aligned}
\psi_c &= \mathrm{softmax}(\eta_c) \\
&= \frac{e^{\eta_c}}{\sum_{k=1}^{C} e^{\eta_k}} \\
\eta_c &= X_c \beta_c \;\; \text{where} \;\; X_c = X[y == c] \\
\beta_c &\sim \mathrm{Normal}(0, \sigma^2 I)
\end{aligned}$$

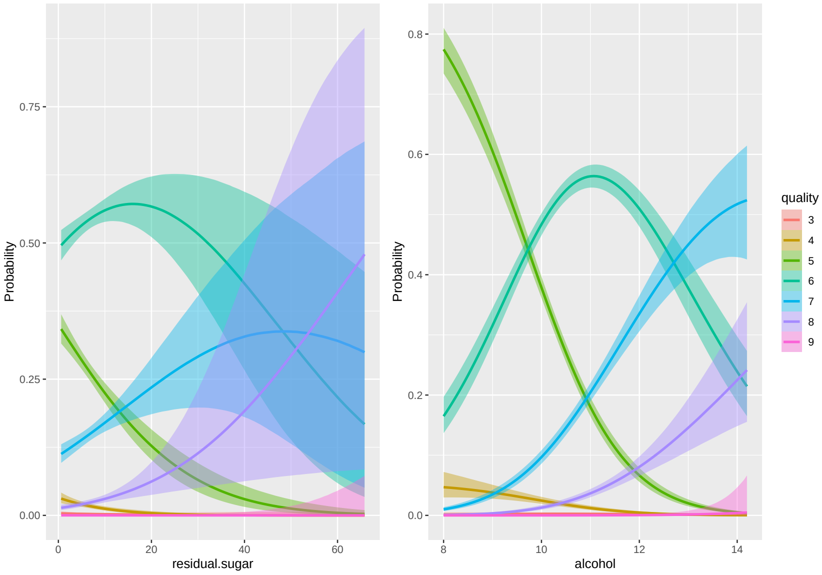
```
f <- quality ~ citric.acid + residual.sugar +
   total.sulfur.dioxide + free.sulfur.dioxide +
   chlorides + density + pH + sulphates + alcohol +
   fixed.acidity + volatile.acideity

fit1 <- brm(f,
            data = d,
            family = categorical(link = "logit"),
            prior = p1)
```

M_1: Result (1)



beta for residual.sugar
9
8
7
6
5
4
−0.5
0.0
0.5
beta for alcohol
9
8
7
6
5
4
−2
0
2
4

M_1: Result (2)



Probability
0.75
0.50
0.25
0.00
0
20
40
60
residual.sugar
Probability
0.8
0.6
0.4
0.2
0.0
8
10
12
14
alcohol
quality
3
4
5
6
7
8
9

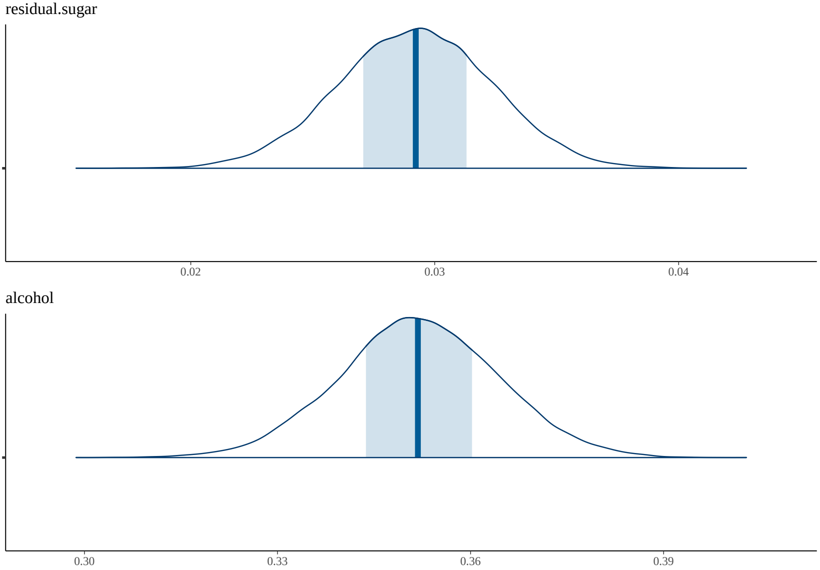
M_2: Regression

We choose to use Normal distribution but other distribution such as t-distribution can be also chosen.

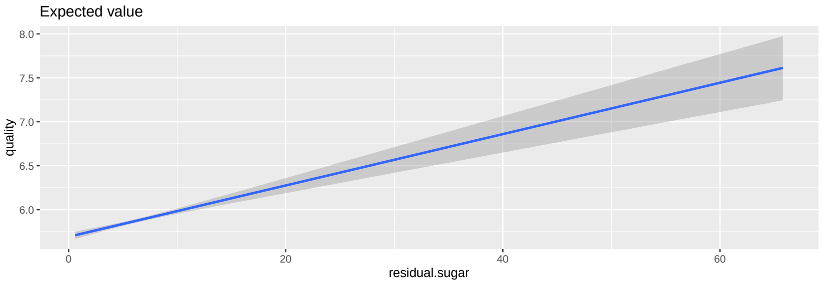
$$
\begin{aligned}
y &\sim \text{Normal}(\eta, \gamma^2) \\
\eta &= x^T \beta \\
\beta &\sim \text{Normal}(0, \sigma_\beta^2 I) \\
\gamma^2 &\sim \text{Half-normal}(0, \sigma_\gamma^2)
\end{aligned}
$$

```
fit2 <- brm(f,
            data = d,
            family = gaussian(),
            prior = p2)
```

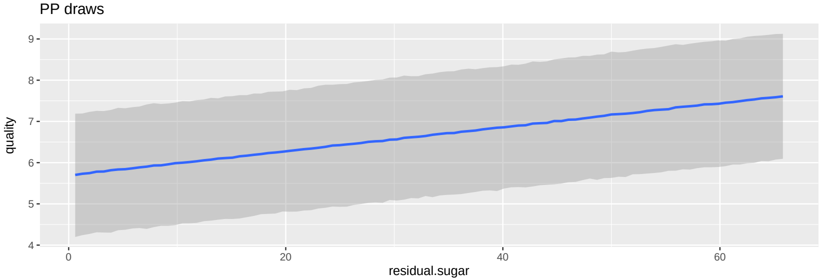
M_2: Result (1)



M_2: Result (2)



Expected value
8.0
7.5
7.0
6.5
6.0
quality
0
20
40
60
residual.sugar



PP draws
9
8
7
6
5
4
quality
0
20
40
60
residual.sugar

M_3: Ordinal Regression: Cumulative Model (1)

For $c = 1, .., C$:

$$\begin{aligned}
\psi_c &= Pr(y \leq c) - Pr(y \leq c - 1) \\
&:= Pr(\tilde{y} \leq \tau_c) - Pr(\tilde{y} \leq \tau_{c-1}) \\
\tilde{y} &= \eta + \epsilon, \ \epsilon \sim \text{Normal}(0, 1) \\
\beta &\sim \text{Normal}(0, \sigma^2 I) \\
\tau_c &\sim \text{Normal}(0, \sigma^2_{\tau_c})
\end{aligned}$$

M_3: Ordinal Regression: Cumulative Model (2)

Other expression:

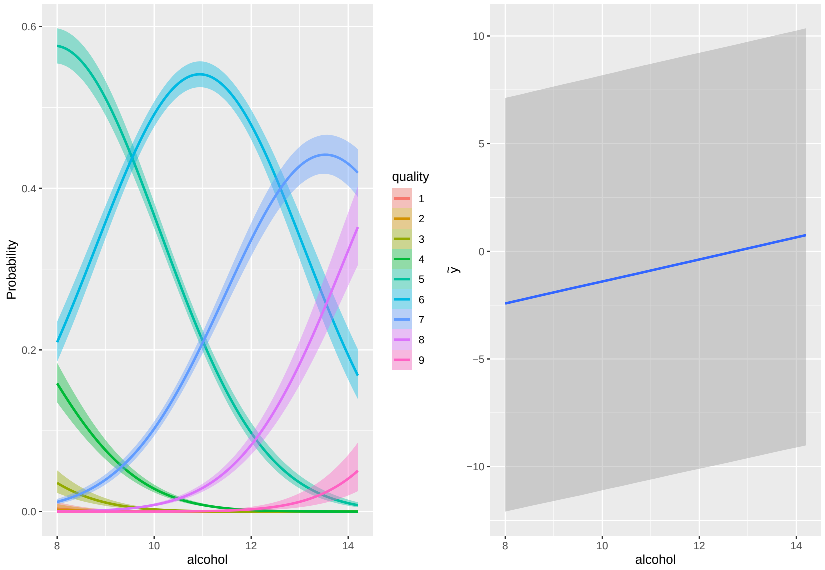
$$
\begin{aligned}
Pr(\tilde{y} \leq \tau_c) &= Pr(\eta + \epsilon \leq \tau_c) \\
&= Pr(\epsilon \leq \tau_c - \eta) \\
&= \Phi(\tau_c - \eta) \quad \Phi : \text{cdf of standard normal aka probit}
\end{aligned}
$$

Then we have:

$$
\begin{aligned}
\psi_c &= \Phi(\tau_c - \eta) - \Phi(\tau_{c-1} - \eta) \\
&\vdots
\end{aligned}
$$

```r
fit3 <- brm(f,
            data = d,
            family = cumulative("probit"),
            prior = p3)
```

M_3: Result



Probability
0.0
0.2
0.4
0.6
8
10
12
14
alcohol
quality
1
2
3
4
5
6
7
8
9
$\tilde{y}$
10
5
0
−5
−10
8
10
12
14
alcohol

Model Comparison

leave-one-out CV

```r
loo_compare(fit1, fit2, fit3)
```

```
     elpd_diff se_diff
fit1    0.0       0.0 
fit3 -144.0      26.8 
fit2 -178.4      29.4 
```

Posterior Model Probability

```r
pmp <- post_prob(fit1, fit2, fit3)
```

```
        fit1         fit2         fit3 
1.000000e+00 4.077118e-26 5.371197e-25 
```

Partial Pooling

- Adding data of red wine and do partial pooling.
-

Adding non-linearity