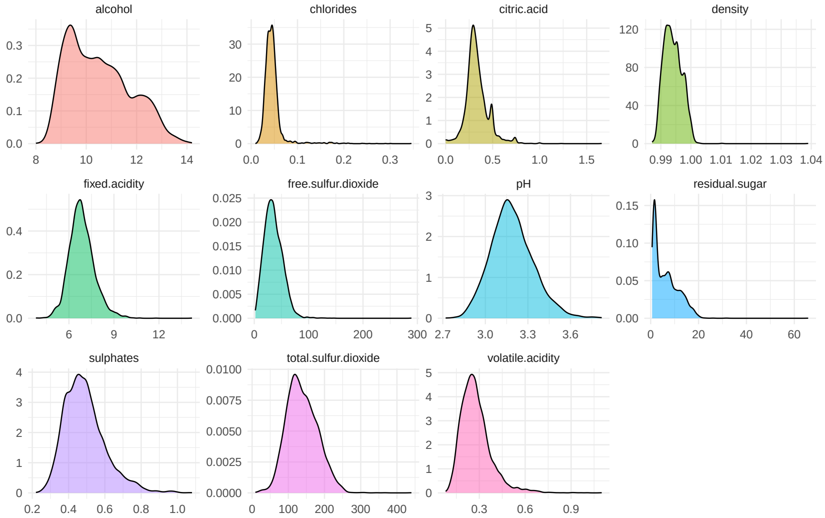
What makes wine great?

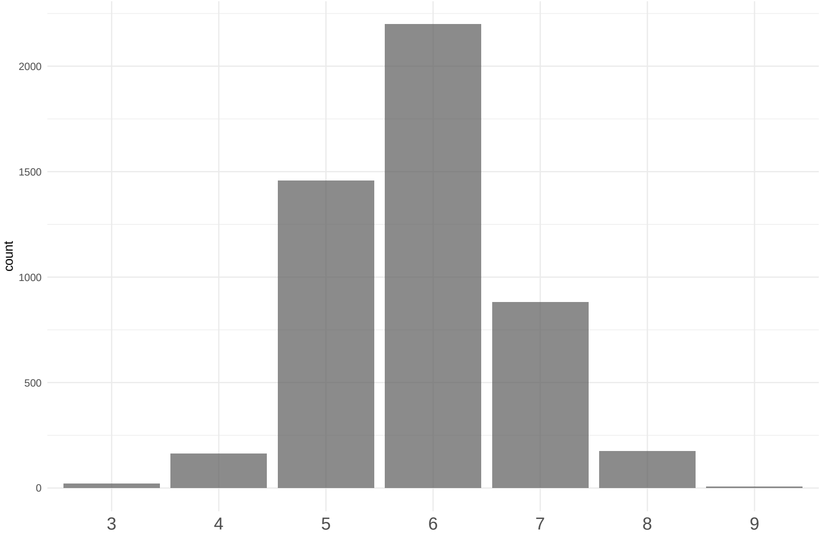
Yuga Hikida

2024-01-02

Data: Preditictive variables



Data: Response variables (quality)



How to model "quality"?

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

  $\Rightarrow$ Classification

- M_2: Continuous variable. $\textit{quality} \in [1, 10]$

  $\Rightarrow$ Regression

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

  $\Rightarrow$ 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 $c = 1, .., C$:

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

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

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

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_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:

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

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)
```