What makes wine great?

Yuga Hikida, Adya Maheshwari

2024-01-30

Task

- Prediction of quality of (white) wine (from 1, 2,.. up to 10) using physicochemical variables.
- Actually only from 3 to 9 is observed.
- Data source: Cortez,Paulo, Cerdeira,A., Almeida,F., Matos,T., and Reis,J.. (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.
- Support vector machine is used in their introductory paper.

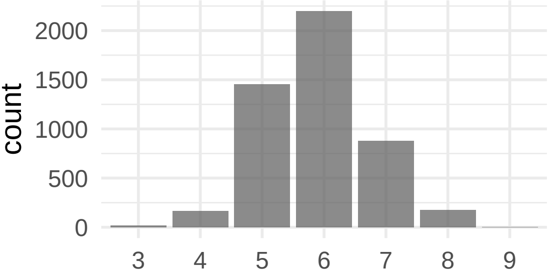

Figure 1: Histogram of quality

Data: Predictive variables

- Acidity: citric.acid, volatile.acidity
- Sweetness: residual.sugar
- Bitterness: sulphates
- Saltiness: chlorides
- Prevent oxidation and bacteria: total.sulfur.dioxide
- Literally interpretable: alcohol

Data: Preditictive variables

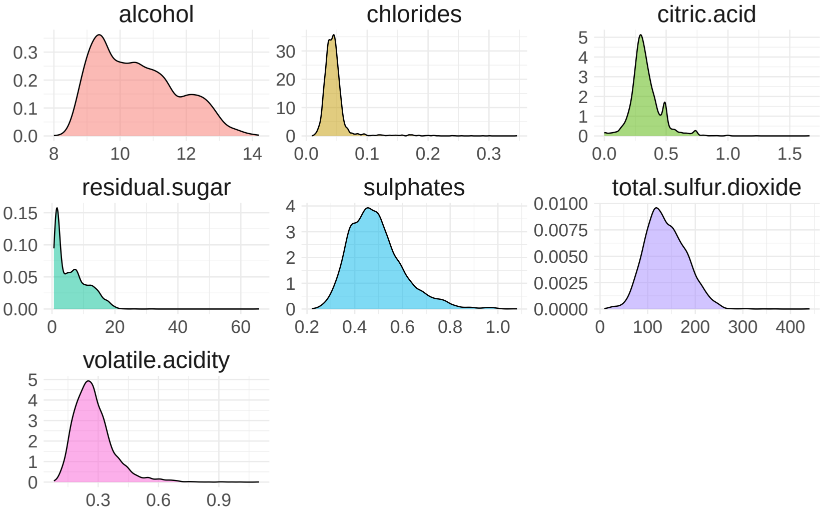

Figure 2: Density of predictive variables

How to model “quality”?

1. Categorical variable. $\textit{quality} \in \{`1`, ..., `10`\}$
- Classification
2. Continuous variable. $\textit{quality} \in [1, 10]$
- Linear Regression
3. Ordered Categorical variable. $\textit{quality} \in \{1, ..., 10\}$
- Ordinal Regression

We want to retain ordered structure of data for interpretation.

$\Rightarrow$ Linear Regression (baseline) and Ordinal Regression

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

Regression

As a baseline model.

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

```r
f <- quality ~ citric.acid + volatile.acidity +
     residual.sugar + sulphates + chlorides +
     total.sulfur.dioxide + alcohol

linear_reg <- brm(f,
                  data = d,
                  family = gaussian(),
                  prior = p_linear_reg)
```

Prior Specification

- Focus on “alcohol”: It takes from 8% to 14% (the range is 6%)
- The response takes from 3 to 9 (the range is 6)
- We don’t expect the absolute value of coefficient to be larger than 1.
- Set weakly informative prior accordingly: $\beta_{alcohol} \sim \text{Normal}(0, 0.4)$

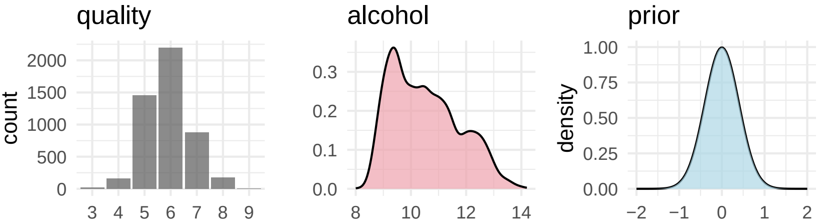

Figure 3: Distribution of (response / predictive) variables and prior distribution.

Prior Specification (cont)

We have

$$\begin{aligned}\beta_{alcohol} &\sim \text{Normal}(0, 0.4)\\ &:= \text{Normal}(0, \tau SD(y)/SD(\text{alcohol}))\end{aligned}$$

- We get scale free informativeness: $\tau \approx 0.5$
- Set prior for other variables as informative as coefficient for "alcohol". (i.e., $\beta_j \sim \text{Normal}(0, \tau SD(y)/SD(x_j))$)

Regression: Result

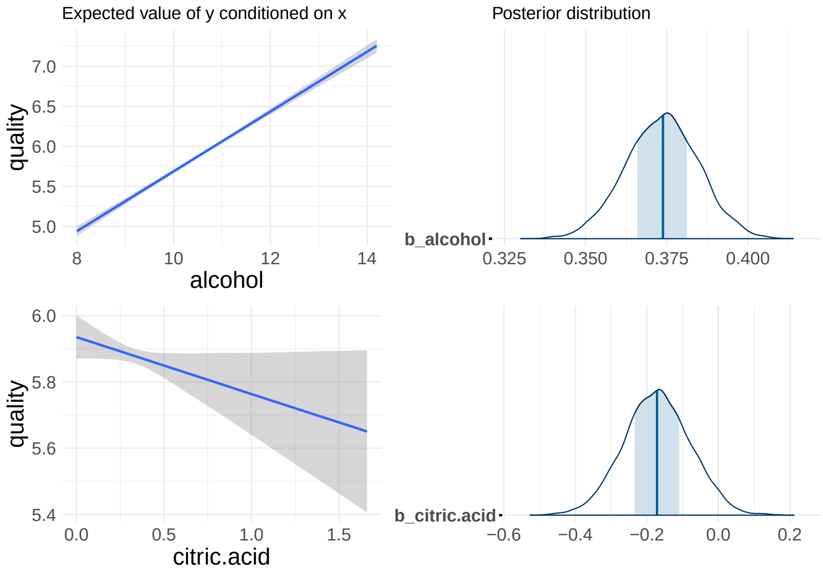

Figure 4: Result for linear regression (only for “alcohol” and “citric.acid”)

Ordinal Regression: Cumulative Model

Consider a continuous latent variable $\tilde{y}$ which determine the quality y through thresholds τ.

For $c = 2, .., C$:

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

Prior is set in the same way as regression (here assume $SD(\tilde{y}) = 1$)

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

Cumulative model: Result

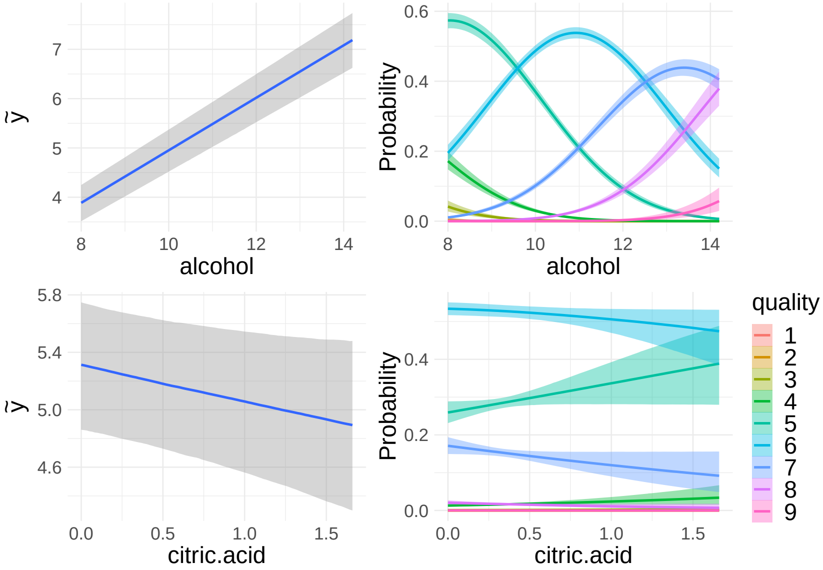

Figure 5: Result for cumulative model (only for “alcohol” and “citric.acid”)

Cumulative model: Non-equidistant

Linear regression implicitly assume equidistant among categories (quality).

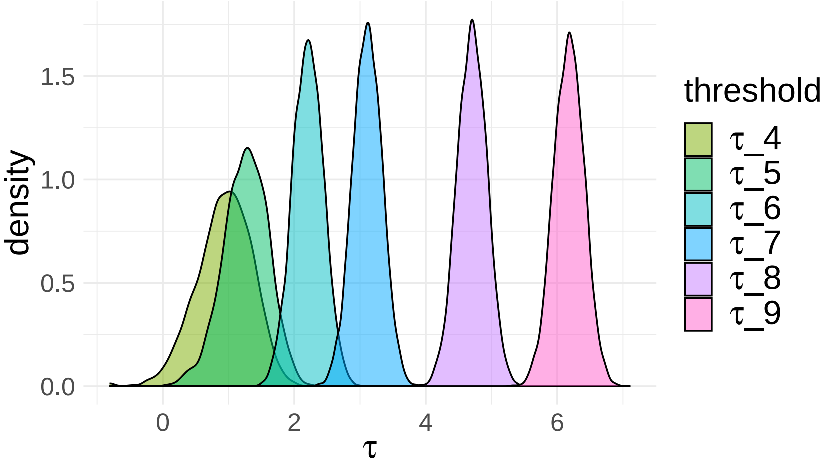

Figure 6: Posterior distribution for the thresholds τ_c

Model Comparison

Leave-one-out Cross Validation

```
loo_compare(linear_reg, cumlat)
```

```
           elpd_diff se_diff
cumlat       0.0       0.0  
linear_reg -37.8      10.0  
```

- Need to be carefully interpreted:
  - We modelled *y* differently.
- Cumulative model has lower ELPD.
  - We continue further analysis with cumulative model.

Adding non-linearity

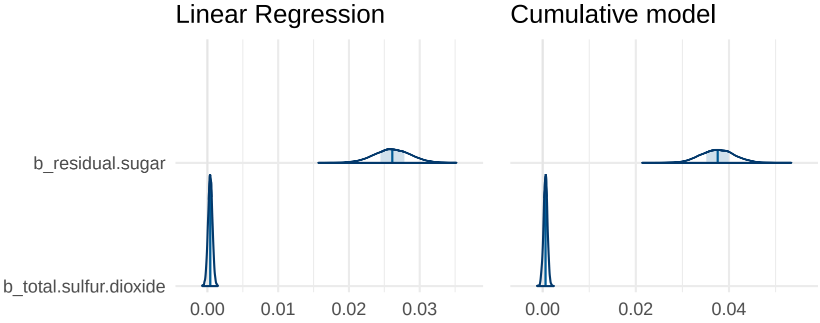

Figure 7: Posterior distributions from the two models.

- Coefficient for “residual.sugar” and “total.sulfur.dioxide” is concentrated in very small value or around zero.
- Might be due to non-linearity ⇒ use spline for the two variables.
- Does “optimal” value exist within the range of data we observed?

Spline: Result

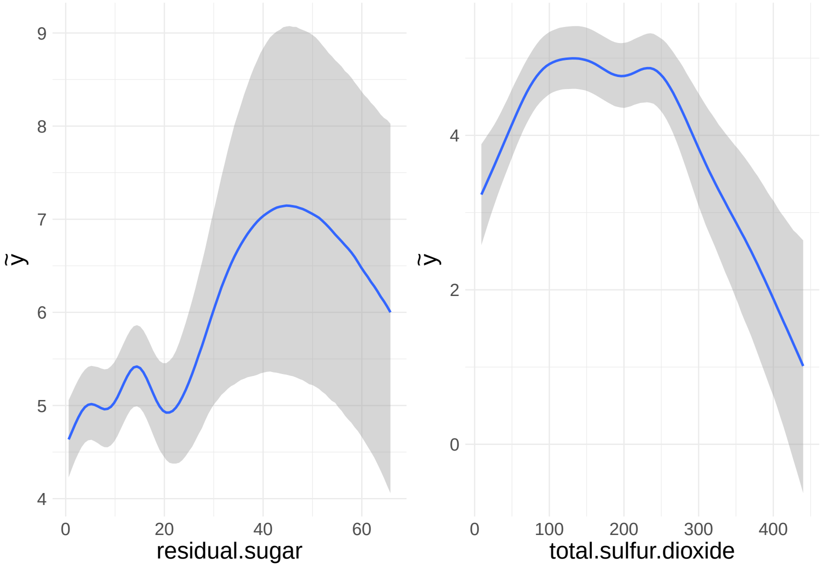

Figure 8: Result for cumulative model with spline (only variables with spline term)

Model Comparison

Leave-one-out CV

```
loo_compare(linear_reg, cumlat, cumlat_s)
```

```
           elpd_diff se_diff
cumlat_s      0.0       0.0
cumlat      -91.9      15.4
linear_reg -129.8      19.0
```

- Adding non-linearity improves ELPD.

Summary

- Positive effect of alcohol and negative effect of citric.acid on quality of wine.
- Ordinal Regression with Spline perform the best in terms of ELPD.
- Non-equidistant of quality
  - Lower quality wine tends to be more similar (quality 3, 4, and 5).
- Non-linear relationship between predictive variables and the quality.
  - The "optimal" values exist within the data rage observed (total.sulfur.dioxide).
- Further analysis: more non-linearity and synergy effects.

Yuga Hikida: yuga.hikida@tu-dortmund.de

Adya Maheshwari: adya.maheshwari@tu-dortmund.de

Appendix

Summary: Regression

```
 Family: gaussian 
  Links: mu = identity; sigma = identity 
Formula: quality ~ citric.acid + volatile.acidity + residual.sugar + sulphates + chlorides + total.sulfur
   Data: d (Number of observations: 4898) 
  Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
         total post-warmup draws = 8000

Population-Level Effects: 
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
Intercept                2.20      0.15     1.90     2.50 1.00     8186
citric.acid             -0.17      0.09    -0.35     0.01 1.00     8908
volatile.acidity        -2.12      0.11    -2.34    -1.90 1.00     8456
residual.sugar           0.03      0.00     0.02     0.03 1.00    10690
sulphates                0.44      0.10     0.25     0.64 1.00     9337
chlorides               -0.87      0.54    -1.94     0.20 1.00     7324
total.sulfur.dioxide     0.00      0.00    -0.00     0.00 1.00     8627
alcohol                  0.37      0.01     0.35     0.40 1.00     7269
                     Tail_ESS
Intercept                6717
citric.acid              5856
volatile.acidity         5703
residual.sugar           6525
sulphates                5876
chlorides                5997
total.sulfur.dioxide     6665
alcohol                  5937

Family Specific Parameters: 
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.76      0.01     0.75     0.78 1.00     9825     5584

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Summary: Cumulative model

```
 Family: cumulative
  Links: mu = probit; disc = identity
Formula: quality ~ citric.acid + volatile.acidity + residual.sugar + sulphates + chlorides + total.sulfur
   Data: d (Number of observations: 4898)
  Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
         total post-warmup draws = 8000

Population-Level Effects:
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
Intercept[1]             0.94      0.42     0.08     1.67 1.00     3026
Intercept[2]             1.26      0.34     0.54     1.89 1.00     4158
Intercept[3]             2.21      0.23     1.75     2.66 1.00     5834
Intercept[4]             3.11      0.22     2.66     3.54 1.00     5747
Intercept[5]             4.70      0.22     4.26     5.14 1.00     5618
Intercept[6]             6.18      0.23     5.72     6.62 1.00     5395
Intercept[7]             7.34      0.24     6.88     7.79 1.00     5348
Intercept[8]             8.77      0.27     8.23     9.30 1.00     5629
citric.acid             -0.25      0.13    -0.51     0.01 1.00     5557
volatile.acidity        -3.11      0.16    -3.42    -2.79 1.00     5336
residual.sugar           0.04      0.00     0.03     0.04 1.00     6909
sulphates                0.63      0.14     0.37     0.90 1.00     5755
chlorides               -1.26      0.78    -2.79     0.29 1.00     6024
total.sulfur.dioxide     0.00      0.00    -0.00     0.00 1.00     8104
alcohol                  0.53      0.02     0.50     0.57 1.00     4876
                     Tail_ESS
Intercept[1]             2646
Intercept[2]             3398
Intercept[3]             5648
Intercept[4]             5589
Intercept[5]             5566
Intercept[6]             5627
Intercept[7]             5466
Intercept[8]             5625
citric.acid              5392
```

Summary: Cumulative with Spline

```
 Family: cumulative
  Links: mu = probit; disc = identity
Formula: quality ~ s(residual.sugar) + s(total.sulfur.dioxide) + citric.acid + volatile.acidity + sulphate
   Data: d (Number of observations: 4898)
  Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
         total post-warmup draws = 8000

Smooth Terms:
                             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
sds(sresidual.sugar_1)           9.23      2.80     4.97    15.72 1.00     2470
sds(stotal.sulfur.dioxide_1)     3.47      1.18     1.78     6.37 1.00     3041
                             Tail_ESS
sds(sresidual.sugar_1)           4390
sds(stotal.sulfur.dioxide_1)     4189

Population-Level Effects:
                        Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
Intercept[1]                0.34      0.45    -0.66     1.12 1.00     4096
Intercept[2]                0.68      0.36    -0.09     1.32 1.00     5887
Intercept[3]                1.73      0.22     1.31     2.16 1.00     6835
Intercept[4]                2.69      0.20     2.30     3.09 1.00     7458
Intercept[5]                4.33      0.20     3.93     4.74 1.00     7462
Intercept[6]                5.83      0.21     5.43     6.25 1.00     7700
Intercept[7]                7.02      0.22     6.60     7.45 1.00     7773
Intercept[8]                8.46      0.26     7.97     8.98 1.00     7144
citric.acid                -0.18      0.14    -0.44     0.08 1.00     9277
volatile.acidity           -3.04      0.16    -3.35    -2.72 1.00     8595
sulphates                   0.60      0.14     0.33     0.87 1.00     8753
chlorides                  -1.58      0.78    -3.12    -0.06 1.00     9246
alcohol                     0.53      0.02     0.49     0.56 1.00     7822
sresidual.sugar_1           2.40      2.57    -2.60     7.39 1.00     5352
stotal.sulfur.dioxide_1    -0.06      2.36    -4.81     4.54 1.00     6721
                        Tail_ESS
Intercept[1]                1960
```

Prior Summary: Regression

```
           prior     class                 coef group resp dpar nlpar lb ub
          (flat)         b
  normal(0,0.36)         b              alcohol
normal(0,20.268)         b            chlorides
 normal(0,3.659)         b          citric.acid
 normal(0,0.087)         b       residual.sugar
  normal(0,3.88)         b            sulphates
  normal(0,0.01)         b total.sulfur.dioxide
 normal(0,4.393)         b     volatile.acidity
    normal(6, 5) Intercept
    normal(0, 5)     sigma                                              0
 source
default
   user
   user
   user
   user
   user
   user
   user
   user
   user
```

Prior Summary: Cumulative

```
               prior     class                 coef group resp dpar nlpar lb
              (flat)         b
     normal(0,0.406)         b              alcohol
    normal(0,22.885)         b            chlorides
    normal(0, 4.132)         b          citric.acid
     normal(0,0.099)         b       residual.sugar
      normal(0,3.88)         b            sulphates
     normal(0,0.012)         b total.sulfur.dioxide
     normal(0,4.961)         b     volatile.acidity
student_t(3, 0, 2.5) Intercept
       normal(-2, 1) Intercept                    1
    normal(-1.43, 1) Intercept                    2
    normal(-0.86, 1) Intercept                    3
    normal(-0.29, 1) Intercept                    4
     normal(0.29, 1) Intercept                    5
     normal(0.86, 1) Intercept                    6
     normal(1.43, 1) Intercept                    7
        normal(2, 1) Intercept                    8
ub  source
   default
      user
      user
      user
      user
      user
      user
      user
   default
      user
      user
      user
      user
      user
      user
```

Prior Summary: Cumulative with Spline

```
               prior     class                    coef group resp dpar nlpar
              (flat)         b
      normal(0,0.406)         b                 alcohol
    normal(0,22.885)         b               chlorides
    normal(0, 4.132)         b             citric.acid
        normal(0, 3)         b       sresidual.sugar_1
        normal(0, 3)         b stotal.sulfur.dioxide_1
      normal(0,3.88)         b               sulphates
     normal(0,4.961)         b        volatile.acidity
student_t(3, 0, 2.5) Intercept
       normal(-2, 1) Intercept                       1
    normal(-1.43, 1) Intercept                       2
    normal(-0.86, 1) Intercept                       3
    normal(-0.29, 1) Intercept                       4
     normal(0.29, 1) Intercept                       5
     normal(0.86, 1) Intercept                       6
     normal(1.43, 1) Intercept                       7
        normal(2, 1) Intercept                       8
student_t(3, 0, 2.5)       sds
student_t(3, 0, 2.5)       sds       s(residual.sugar)
student_t(3, 0, 2.5)       sds s(total.sulfur.dioxide)
lb ub       source
           default
              user
              user
              user
              user
              user
              user
              user
           default
              user
              user
              user
```

Cumulative Model (cont)

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:

$$\psi_c = \Phi(\tau_c - \eta) - \Phi(\tau_{c-1} - \eta)$$

Adding non-linearity with Spline

```r
f_s <- quality ~ citric.acid + volatile.acidity +
       sulphates + chlorides + alcohol +
       s(residual.sugar) + s(total.sulfur.dioxide)

cumlat_s <- brm(f_s,
                data = d,
                family = cumulative("probit"),
                prior = p_cumlat_s)
```

- We are particularly interested in non-linearity of these two variables.
- Other variable could be non-linear.