Model Selection

Preparation & Transformation

```r
# prep
library(tidyverse)
library(ggplot2)
library(tidyr)
library(corrplot)
library(broom)
library(purrr)

wine <- read_csv("data/wine.csv")
wine <- wine |>
  mutate(
    type = factor(
      type,
      levels = c(0, 1),
      labels = c("red", "white")
    )
  )
```

```r
wine_tf <- data.frame(
  # response
  quality = wine$quality,
  type = wine$type,

  # transformed predictors
  residual_sugar_t        = log(wine$residual_sugar + 0.1),
  total_sulfur_dioxide_t  = log(wine$total_sulfur_dioxide + 1),
  free_sulfur_dioxide_t   = log(wine$free_sulfur_dioxide + 1),
  chlorides_t             = log(wine$chlorides),
  volatile_acidity_t      = log(wine$volatile_acidity),
  sulphates_t             = log(wine$sulphates),
  citric_acid_t           = log(wine$citric_acid + 0.1),

  # raw predictors kept as-is
  alcohol          = wine$alcohol,
  fixed_acidity    = wine$fixed_acidity,
  pH               = wine$pH,
  density          = wine$density
)
```

A

baseline / main effects / no interactions

```
mA <- lm(
  quality ~ alcohol + residual_sugar_t + volatile_acidity_t +
    chlorides_t + sulphates_t + citric_acid_t +
    total_sulfur_dioxide_t + fixed_acidity + pH + type,
  data = wine_tf
)
```

B

strong interactions: - volatile_acidity × type - chlorides × type - residual_sugar × type

```
mB <- lm(
  quality ~ alcohol +
    residual_sugar_t * type +
    volatile_acidity_t * type +
    chlorides_t * type +
    total_sulfur_dioxide_t +
    fixed_acidity + pH,
  data = wine_tf
)
```

C

strong + moderate interactions: - total_sulfur_dioxide × type - residual_sugar × type - volatile_acidity × type - chlorides × type

```
mC <- lm(
  quality ~ alcohol +
    residual_sugar_t * type +
    volatile_acidity_t * type +
    chlorides_t * type +
    total_sulfur_dioxide_t * type +
    fixed_acidity + pH,
  data = wine_tf
)
```

D

Strong + moderate + weak

```
mD <- lm(
  quality ~ alcohol +
    residual_sugar_t * type +
    volatile_acidity_t * type +
    chlorides_t * type +
    total_sulfur_dioxide_t * type +
    citric_acid_t * type +
    sulphates_t * type +
    fixed_acidity + pH,
  data = wine_tf
)
```

E

Replace SO_2 variable to test collinearity sensitivity

```
mE <- lm(
  quality ~ alcohol +
    residual_sugar_t * type +
    volatile_acidity_t * type +
    chlorides_t * type +
    free_sulfur_dioxide_t +     # 替代 total_SO₂
    fixed_acidity + pH,
  data = wine_tf
)
```

F

Minimal interpretable interaction model(最解释力强、科学意义最明确的交互项)： - residual sugar（白酒显著更甜→不同 slope） - volatile acidity（红酒对 VA 敏感度更高）

```
mF <- lm(
  quality ~ alcohol +
    volatile_acidity_t * type +
    residual_sugar_t * type +
    fixed_acidity + pH,
  data = wine_tf
)
```

G

Additive nonlinear expansion model

```
mG <- lm(
  quality ~ alcohol + I(alcohol^2) +
    volatile_acidity_t + I(volatile_acidity_t^2) +
    residual_sugar_t + I(residual_sugar_t^2) +
    total_sulfur_dioxide_t +
    type,
  data = wine_tf
)
```

```
models <- list(mA = mA, mB = mB, mC = mC, mD = mD, mE = mE, mF = mF, mG = mG)
model_summaries <- map_df(
  models,
  ~ glance(.x),
  .id = "model"
)

model_summaries
```

```
# A tibble: 7 × 13
  model r.squared adj.r.squared sigma statistic p.value    df logLik    AIC
  <chr>     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl>  <dbl>
1 mA        0.292         0.291 0.735      267.       0    10 -7217. 14459.
2 mB        0.286         0.285 0.738      236.       0    11 -7243. 14512.
3 mC        0.289         0.288 0.737      220.       0    12 -7230. 14489.
4 mD        0.298         0.296 0.733      172.       0    16 -7189. 14413.
5 mE        0.296         0.295 0.733      248.       0    11 -7198. 14422.
6 mF        0.284         0.283 0.739      322.       0     8 -7251. 14522.
7 mG        0.286         0.285 0.738      325.       0     8 -7243. 14505.
# ℹ 4 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>, nobs <int>
```

Model D（包含 strong + moderate interactions）是目前表现最好的模型，AIC 最低、BIC 第二低、sigma 最小，说明它在拟合质量与复杂度之间取得了最佳平衡。

以下内容仅供阅读，不要直接用❗❗

🔹English Version（for report）

Model D represents a hierarchical interaction model that incorporates both strong and moderate type-by-predictor interactions identified during the EDA stage. By allowing the effects of residual sugar, volatile acidity, chlorides, and total sulfur dioxide to vary between red and white wines, Model D captures meaningful structural differences in how chemical attributes influence sensory-perceived quality. Empirically, Model D achieves the best overall balance between model fit and complexity, with the lowest AIC, highest adjusted R^2, and smallest residual standard error among all candidate models. These results suggest that interactions between wine type and key chemical factors are essential for explaining variability in quality ratings, making Model D the strongest explanatory model for our research question.

🔹中文版本（用于报告解释）

Model D 是一个分层交互模型，纳入了在 EDA 阶段确认的强交互与中等交互项。通过允许 residual sugar、volatile acidity、chlorides 与 total sulfur dioxide 对 wine quality 的影响在红酒与白酒之间发生变化，Model D 能够捕捉"化学属性如何以不同方式影响不同类型葡萄酒的

感官评分”这一核心结构性差异。从经验结果来看，Model D 在所有候选模型中呈现出拟合度与复杂度之间的最佳平衡——具有最低的 AIC、最高的 adjusted R^2、以及最小的残差标准差。这表明 wine type 与关键化学因子之间的交互效应对于解释质量评分的变异至关重要，使 Model D 成为回答本研究问题的最具解释力的模型。