Stat 151A Final Project Report

Wine Quality

Hex Wu, Jingzhi Zhang, Martin Yang

December 19, 2025

Contents

1 Introduction

This project aims to explain differences in red and white wine quality based on physicochemical characteristics. Using a dataset from the *UC Irvine Machine Learning Repository*, we examine the relationship between chemical properties such as alcohol content, acidity, sulfur dioxide content, and sugar concentration, and taster quality ratings.

Our main research questions are: **Which physicochemical properties are most strongly related to wine quality? Are there systematic differences in these relationships between red and white wines?**

To address these questions, we use linear regression models guided by exploratory data analysis (EDA), transformations of skewed predictors, systematic model comparisons, and diagnostic checks.

2 Final Regression Model

2.1 Exploratory Data Analysis and Variable Transformation

Exploratory analysis revealed that several predictors (e.g., residual sugar, sulfur dioxide content, chlorides, sulfates, and volatile acidity) exhibited **significant right-skewed distributions and heavy tails**. Based on histogram analysis, these variables were **log-transformed** to stabilize variance and improve approximate linear relationships.

EDA also showed differences in the distribution of several predictors between red and white wines, indicating **potential interaction between wine type and selected chemical variables**.

The transformed modeling dataset (`wine_tf`) includes:

- Response: quality
- Factor: type (red, white)
- Log-transformed predictors for skewed variables
- Untransformed predictors where distributions appeared approximately symmetric

2.2 Model Specification

Based on systematic model comparison using adjusted R^2, AIC, BIC, and residual scale estimates, we selected **Model D** as our final regression model. This model balances explanatory power and interpretability while allowing key effects to vary by wine type.

The final model is

$$\begin{aligned}\text{quality} = {} & 2.244 + 0.343\,\text{alcohol} + 0.019\,\text{residual sugar} - 0.470\,\text{volatile acidity} \\ & - 0.136\,\text{chlorides} - 0.074\,\text{total sulfur dioxide} + 0.054\,\text{citric acid} \\ & + 0.726\,\text{sulphates} - 0.024\,\text{fixed acidity} - 0.003\,\text{pH} \\ & - 2.043\,\text{type}_{\text{white}} + 0.119\,(\text{residual sugar} \times \text{type}_{\text{white}}) \\ & - 0.177\,(\text{volatile acidity} \times \text{type}_{\text{white}}) \\ & - 0.044\,(\text{chlorides} \times \text{type}_{\text{white}}) + 0.231\,(\text{total sulfur dioxide} \times \text{type}_{\text{white}}) \\ & - 0.038\,(\text{citric acid} \times \text{type}_{\text{white}}) - 0.530\,(\text{sulphates} \times \text{type}_{\text{white}}) + \epsilon\end{aligned}$$

Here, residual sugar, volatile acidity, chlorides, total sulfur dioxide, citric acid, and sulphates denote the log-transformed versions of the original variables. Each interaction term is specified as a product between a predictor and wine type, which in the fitted regression model corresponds to including the predictor main effect, the wine type main effect, and their interaction.

Our final regression model relates wine quality to key physicochemical characteristics together with interaction terms between wine type and selected predictors. The model was chosen through systematic comparison of candidate models and evaluated using diagnostic checks to ensure interpretability and adequate model fit.

2.3 Model Diagnostics

We performed standard linear regression diagnostics:

- Linearity: There was no apparent systematic pattern between the residuals and fitted values, indicating that the **linear approximation is reasonable**.
- Normality: The normal Q–Q plot showed **slight deviations in the tails**, which are considered **acceptable** given the large sample size of over 6000 observations.
- Homoscedasticity: The scale-location plot indicated a **moderate degree of heteroscedasticity**. To address this, we conducted inference using heteroscedasticity-robust (HC3) standard errors.
- Influential Observations: Cook's distance and leverage diagnostics showed no observations exerting undue influence on the fitted model.

3 Explanation and Evaluation

3.1 Main Findings

Alcohol content is positively correlated with wine quality, consistent with previous empirical findings in wine science.

Significant interaction effects were observed, indicating that red and white wines respond differently to certain chemical attributes:

- The association between **residual sugar and volatile acidity** with wine quality varies depending on the wine type.
- **Sulfur dioxide and sulfates exhibit type-dependent effects**, suggesting that preservatives may have different impacts on the perceived quality of red and white wines.

These interaction terms significantly improved model fit, as evidenced by lower AIC values and higher adjusted R^2 values compared to models without interaction terms.

3.2 Limitations and Alternative Methods

Despite its advantages, this analysis has some limitations.

The model is **correlational rather than causal**; unobserved confounding factors such as production methods or regional effects may influence the results.

While logarithmic transformation can improve model performance, **nonlinear relationships may still exist** beyond quadratic or interaction effects.

Other methods, such as generalized additive models (GAMs), mixed-effects models, or explicit heteroscedasticity modeling (e.g., WLS/FGLS), could further improve inference.

4 Conclusion

In this project, we developed and evaluated a regression framework to understand the relationship between the physicochemical properties of wine and its quality, and how these relationships vary across different wine types.

Overall, the results suggest that **wine quality is most strongly associated with alcohol content and acidity-related measures, with several key factors exhibiting systematically different effects for red and white wines**, highlighting the importance of accounting for wine type when modeling and evaluating wine quality.

5 Additional Work: Alternative Models and Diagnostics

5.1 Baseline Pooled Linear Models

First, we fitted a pooled ordinary least squares (OLS) model including all transformed predictor variables and wine type as a main effect, but without interaction terms. This baseline model served as a reference for evaluating the incremental value of more complex model structures.

Although the pooled model captured several overall trends in wine quality, exploratory analysis indicated that it failed to account for systematic differences between red and white wines. Residual plots and scatter plots stratified by type suggested that the slopes of several key predictors varied by wine type, motivating further exploration of interaction terms.

5.2 Interaction Models of Increasing Complexity

Guided by exploratory data analysis, we constructed a sequence of interaction models of increasing complexity. These models progressively introduced interactions between wine type and selected physicochemical variables exhibiting strong or moderate type-dependent patterns.

We compared:

- Models containing only the strongest interactions,
- Models containing both strong and moderate interactions,
- More comprehensive interaction models including weaker but potentially meaningful interactions.

Model comparisons using adjusted R^2, AIC, and BIC showed that interaction models consistently outperformed the pooled OLS model. However, overly complex models exhibited diminishing returns relative to their increased complexity, guiding the selection of a balanced interaction structure.

5.3 Sensitivity Analysis of Correlated Predictors

Exploratory correlation analysis revealed high correlations between certain predictors, particularly between different sulfur dioxide measurements. To assess sensitivity to multicollinearity, we fitted alternative models substituting free sulfur dioxide for total sulfur dioxide while keeping the overall model structure unchanged.

The resulting models showed similar fit and coefficient patterns, indicating that the main conclusions were not driven by the specific sulfur dioxide measurement chosen. For ease of interpretation and consistency with previous findings, we retained total sulfur dioxide in the final model.

5.4 Nonlinear Extensions

To examine potential nonlinear relationships, we explored models incorporating quadratic polynomial terms for selected predictors, including alcohol content and log-transformed acidity measures.

While these nonlinear models provided slight improvements in some cases, the gains were limited and came at the cost of reduced interpretability. Residual diagnostics indicated no substantial curvature after log transformation, and polynomial terms were therefore excluded from the final model.

5.5 Summary of Additional Work

These additional analyses informed the model selection process. The baseline model established a reference point, interaction models captured meaningful type-specific effects, sensitivity analysis addressed collinearity concerns, and diagnostic checks verified model assumptions. This systematic exploration supports the final regression model presented in this report.

References

[1] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Wine Quality. *UCI Machine Learning Repository*, 2009. `https://doi.org/10.24432/C56S3T`.

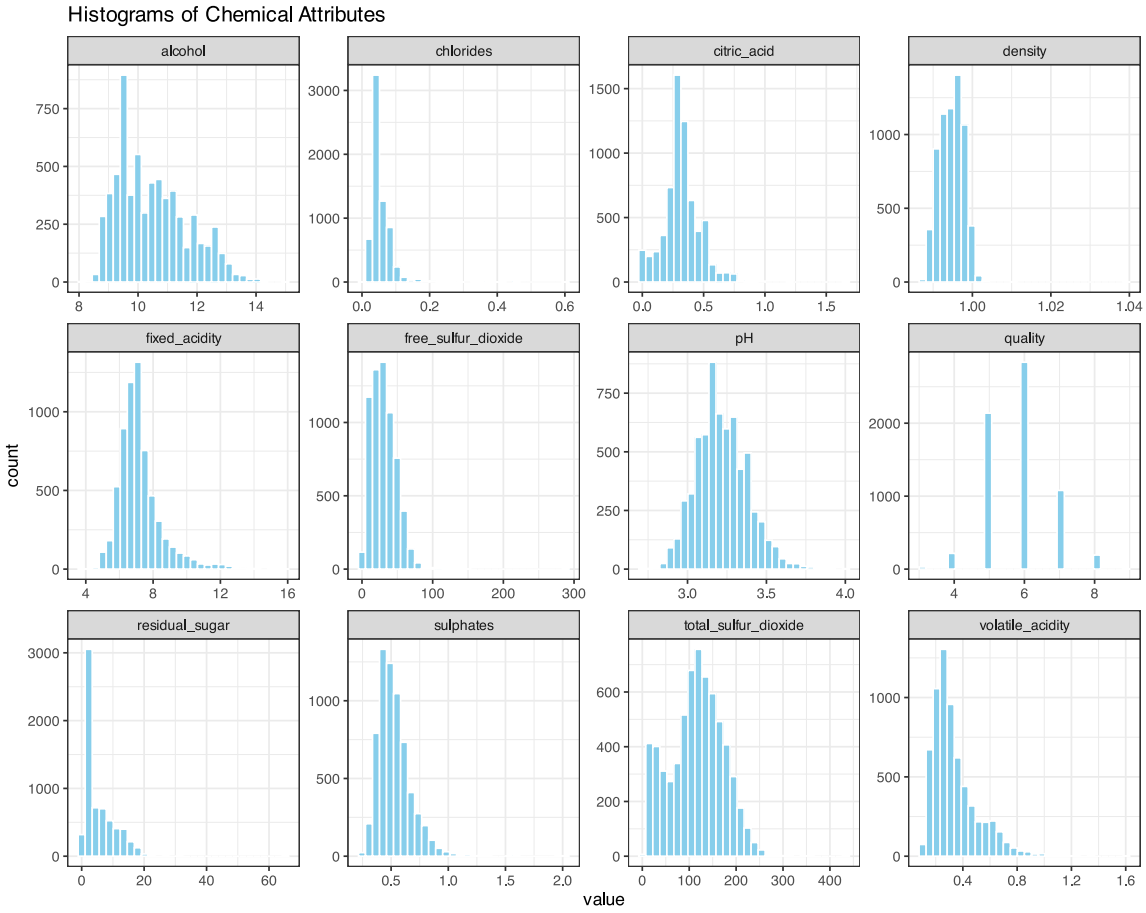
Data Processing

```
red <- read.csv("data/winequality-red.csv", sep = ";") |> mutate(type = "0")
white <- read.csv("data/winequality-white.csv", sep = ";") |> mutate(type =
"1")

wine <- bind_rows(red, white)
colnames(wine) <- str_replace_all(colnames(wine), pattern = "\\.", "_")
wine|> write_csv("data/wine.csv")
```

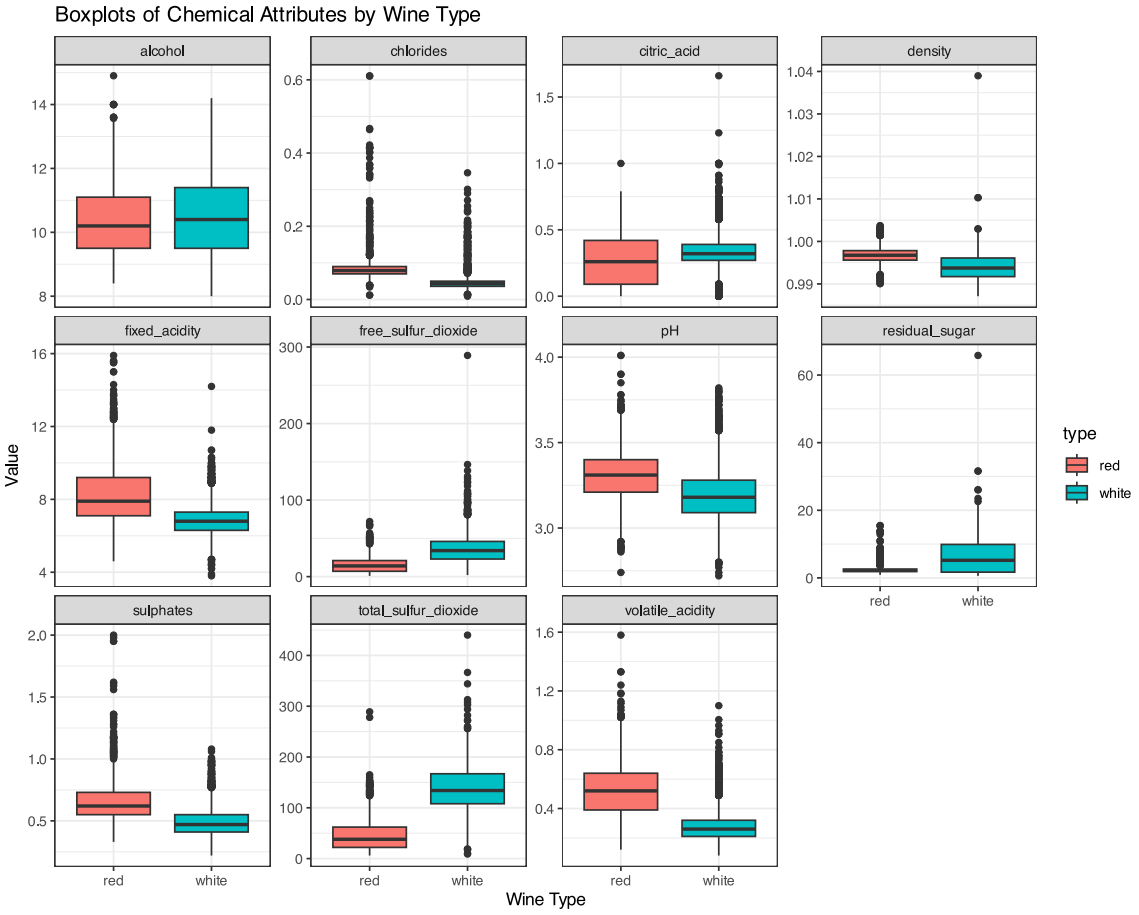
EDA

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

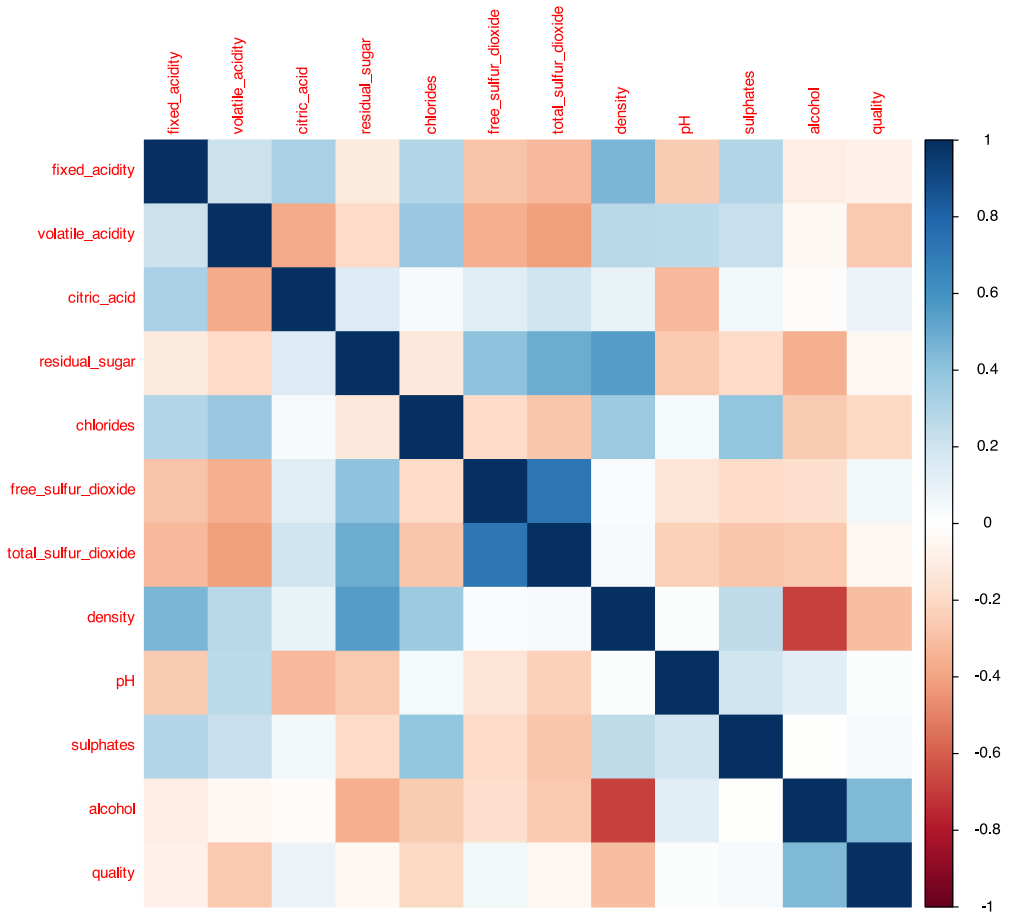
```
# only numerical variables, excluding types
wine_numeric <- wine[, sapply(wine, is.numeric)]
wine_long <- pivot_longer(wine_numeric,
                          cols = everything(),
                          names_to = "variable",
                          values_to = "value")

ggplot(wine_long, aes(x = value)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "white") +
  facet_wrap(~ variable, scales = "free") +
  theme_bw() +
  labs(title = "Histograms of Chemical Attributes")
```



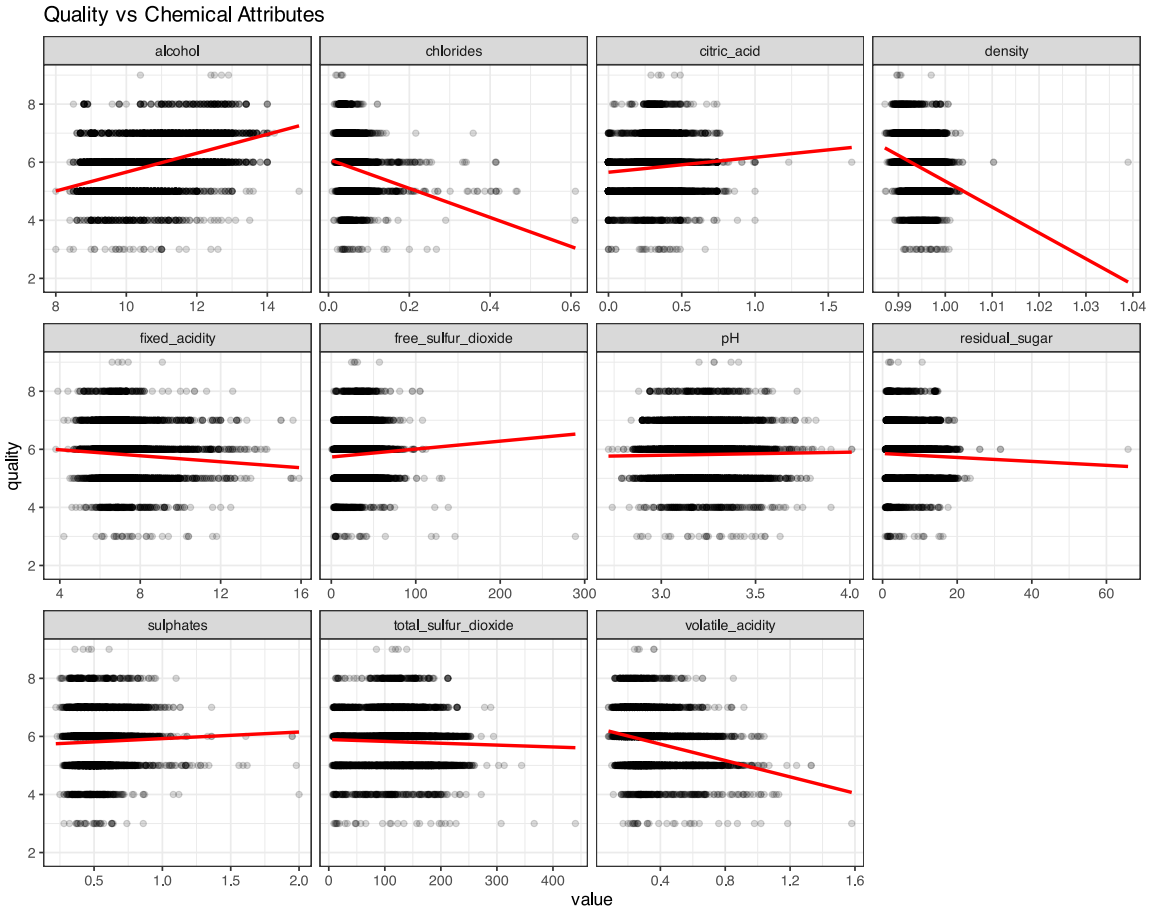

```
# Boxplot: chemical variable by type
wine_long2 <- pivot_longer(wine,
                           cols = -c(type, quality),
                           names_to = "variable",
                           values_to = "value")

ggplot(wine_long2, aes(x = type, y = value, fill = type)) +
  geom_boxplot() +
  facet_wrap(~ variable, scales = "free_y") +
  theme_bw() +
  labs(title = "Boxplots of Chemical Attributes by Wine Type",
       x = "Wine Type", y = "Value")
```




```r
# correlation heatmap
corr_matrix <- cor(wine_numeric)   # numeric variables only
corrplot(corr_matrix, method = "color", tl.cex = 0.8)
```

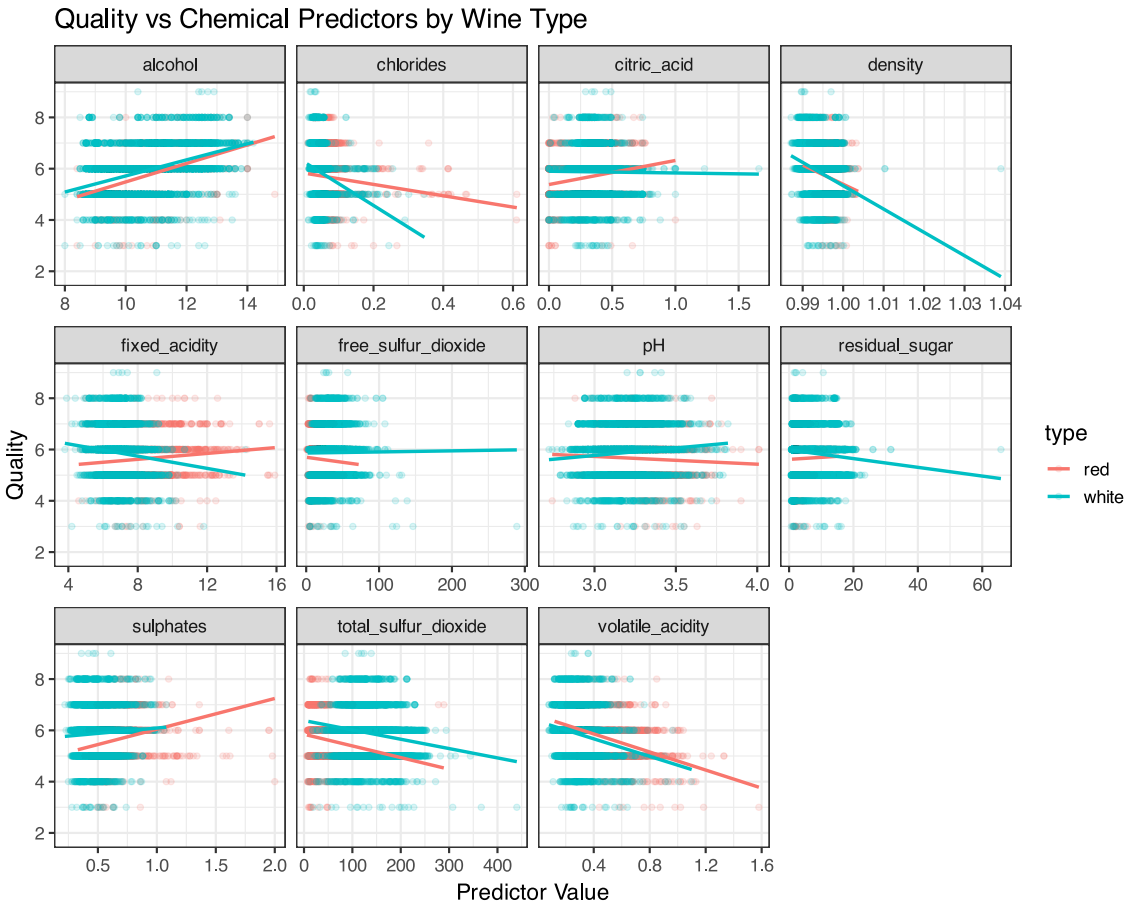



```r
# scatterplot
ggplot(wine_long2, aes(x = value, y = quality)) +
  geom_point(alpha = 0.15) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  facet_wrap(~ variable, scales = "free_x") +
  theme_bw() +
  labs(title = "Quality vs Chemical Attributes")
```

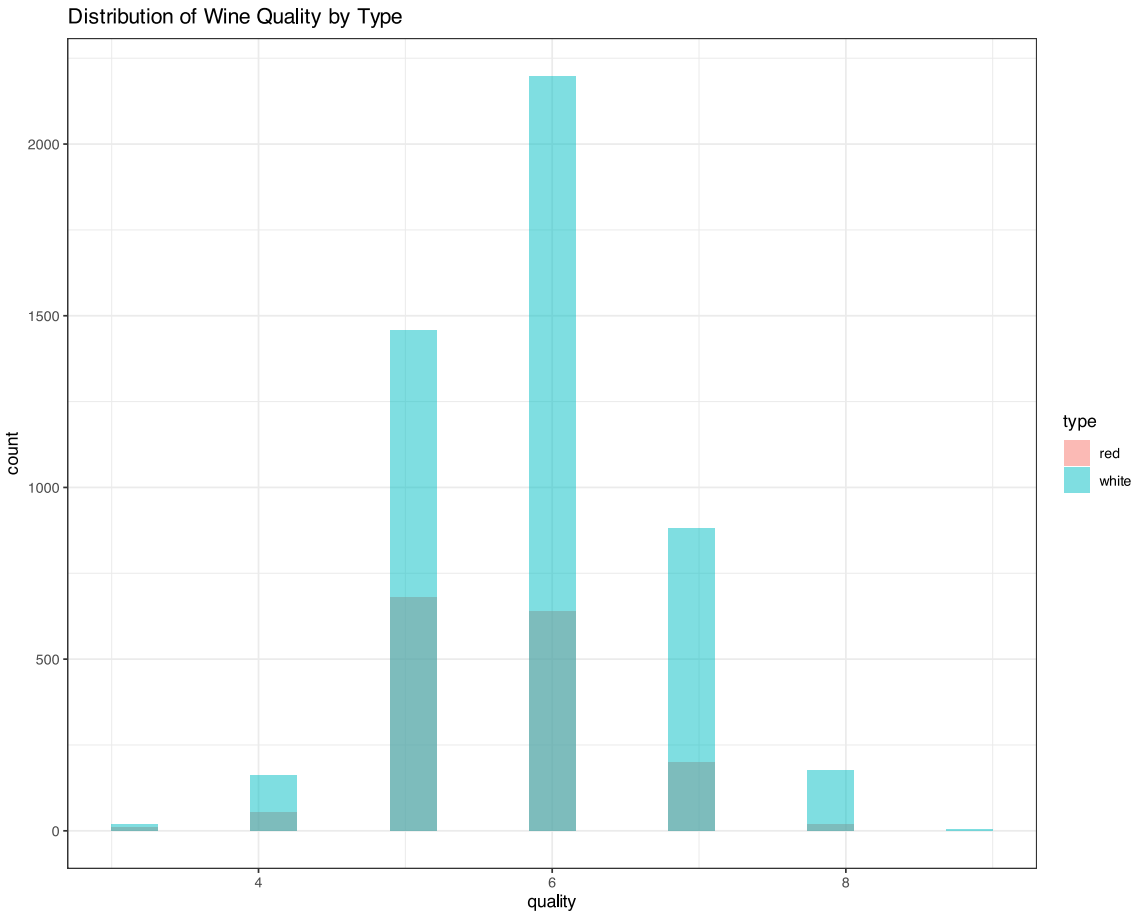



```r
# Quality VS Type
predictors <- wine %>%
  select(-quality, -type) %>%
  select(where(is.numeric)) %>%
  colnames()

wine_long <- wine %>%
  pivot_longer(
    cols = all_of(predictors),
    names_to = "variable",
    values_to = "value"
  )

ggplot(wine_long, aes(x = value, y = quality, color = type)) +
  geom_point(alpha = 0.15) +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~ variable, scales = "free_x") +
  theme_bw(base_size = 14) +
  labs(
```

```
    title = "Quality vs Chemical Predictors by Wine Type",
    x = "Predictor Value",
    y = "Quality"
  )
```




```
ggplot(wine, aes(x = quality, fill = type)) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 20) +
  theme_bw() +
  labs(title = "Distribution of Wine Quality by Type")
```



Model Candidates

```r
# Transformations
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

```r
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

```r
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

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

G

Additive nonlinear expansion model

```r
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
)
```

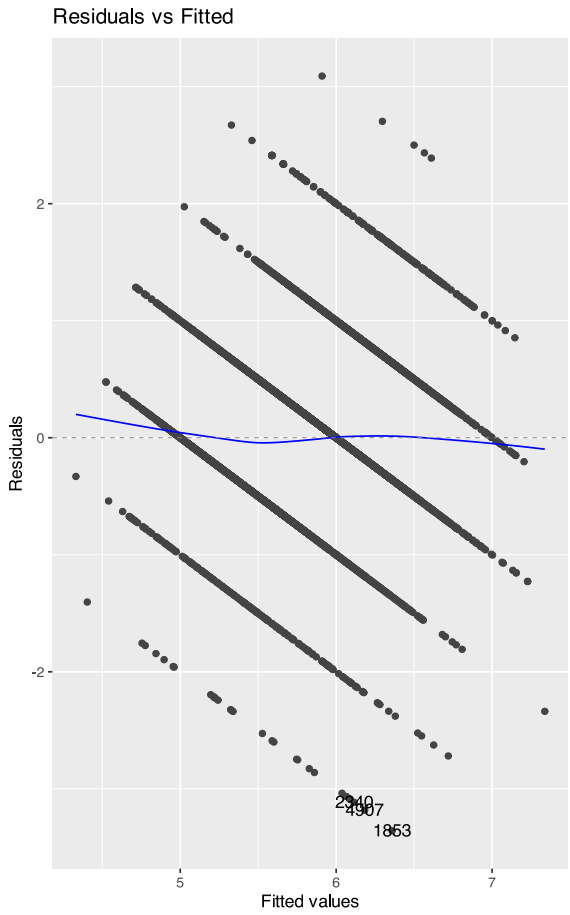
Model Comparison

```
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
```

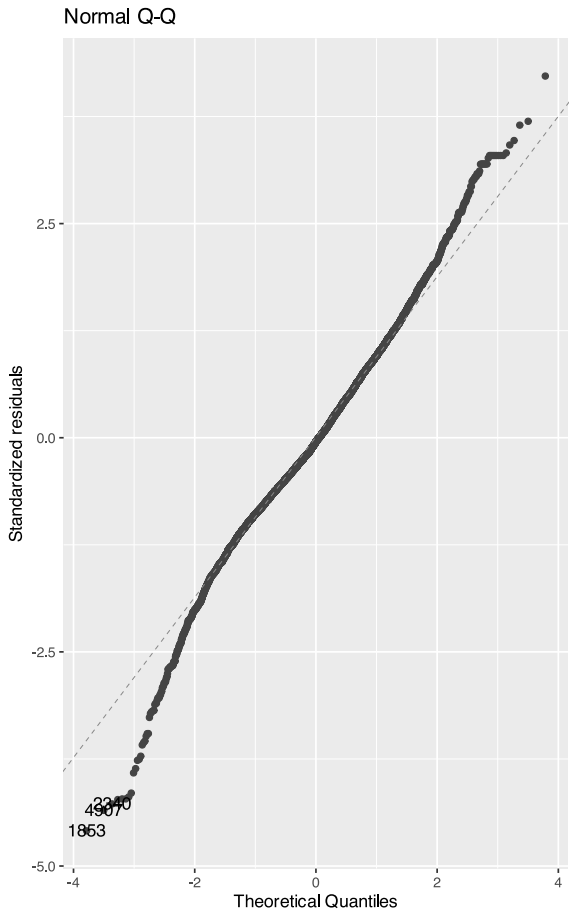
```
# A tibble: 7 × 13
  model r.squared adj.r.squared sigma statistic p.value    df logLik    AIC
  <chr>     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl>  <dbl>
1 mA        0.292         0.291 0.735      267.       0    10 -7217. 14459.
2 mB        0.286         0.285 0.738      236.       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>
```

Diagnotics

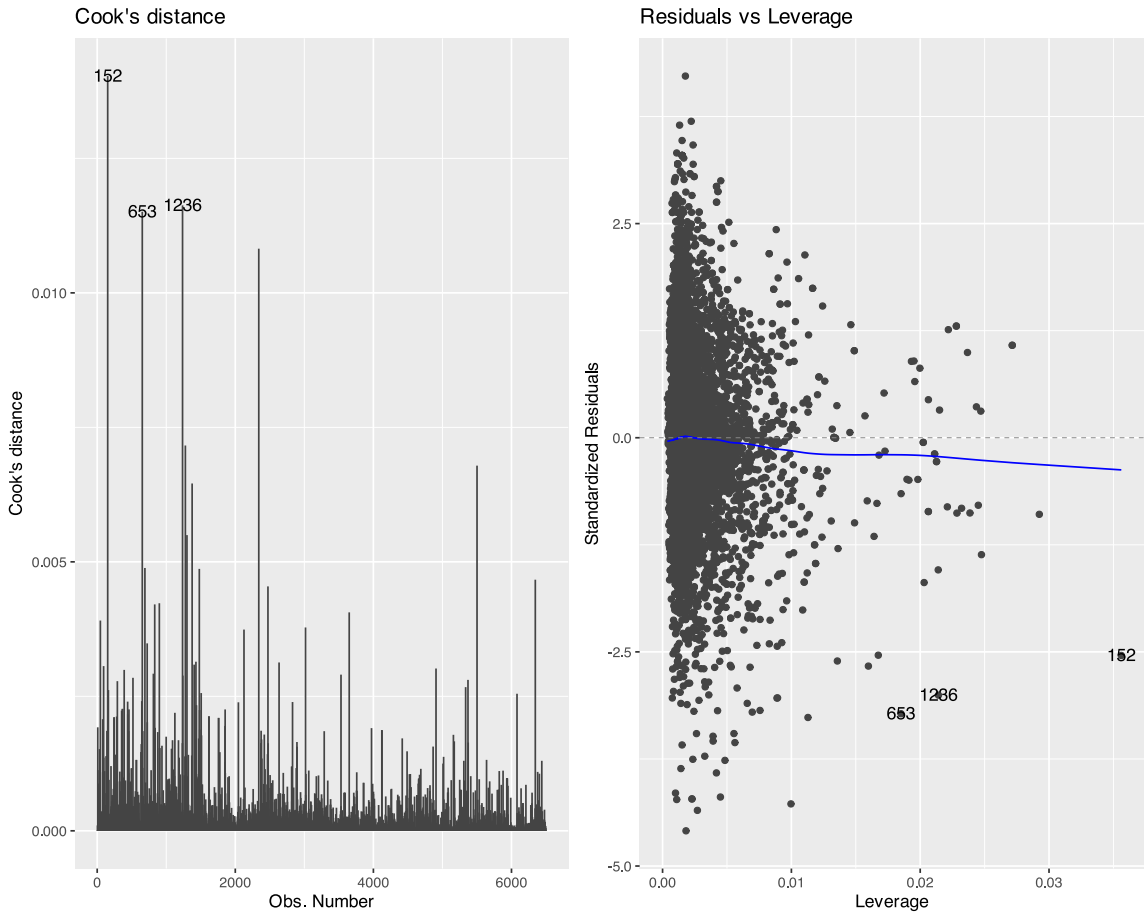
```
autoplot(mD, which = 1)
```



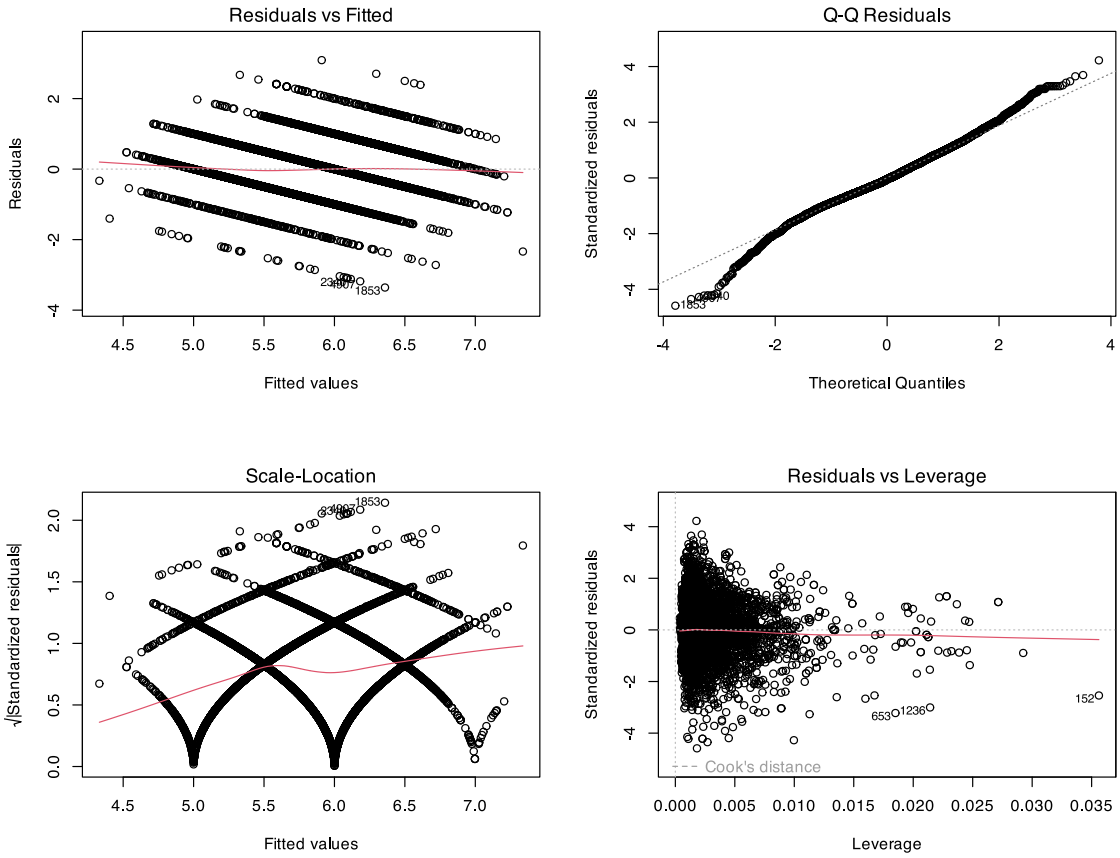

```
autoplot(mD, which = 2)
```



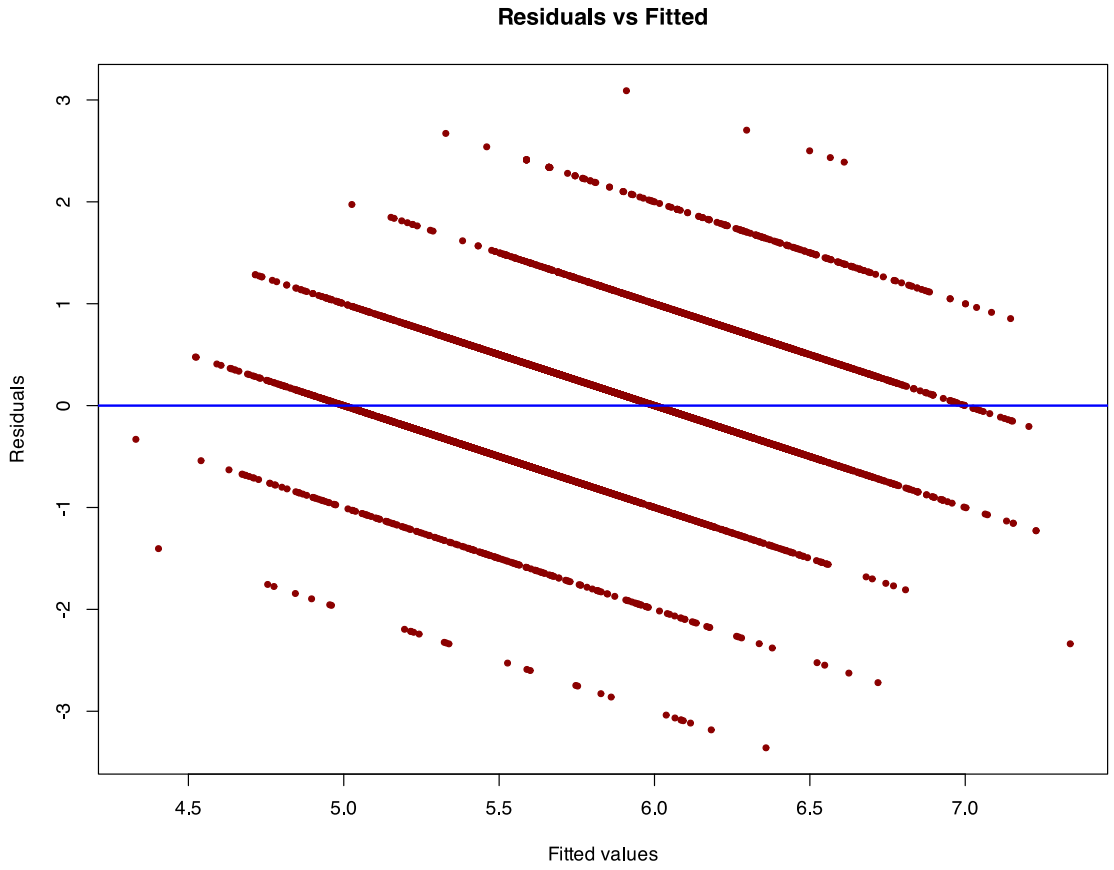

```
autoplot(mD, which = 4:5)
```



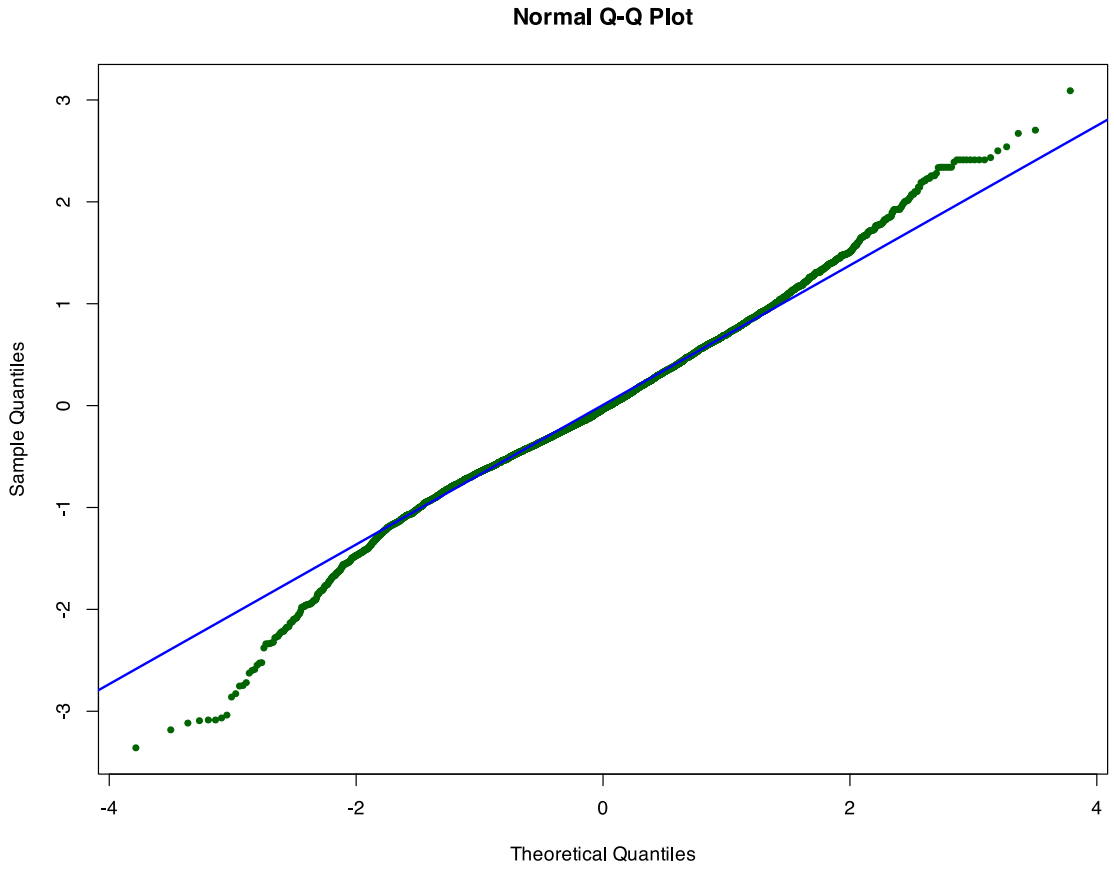

```
# Standard diagnostic panel
par(mfrow = c(2, 2))
plot(mD)
```



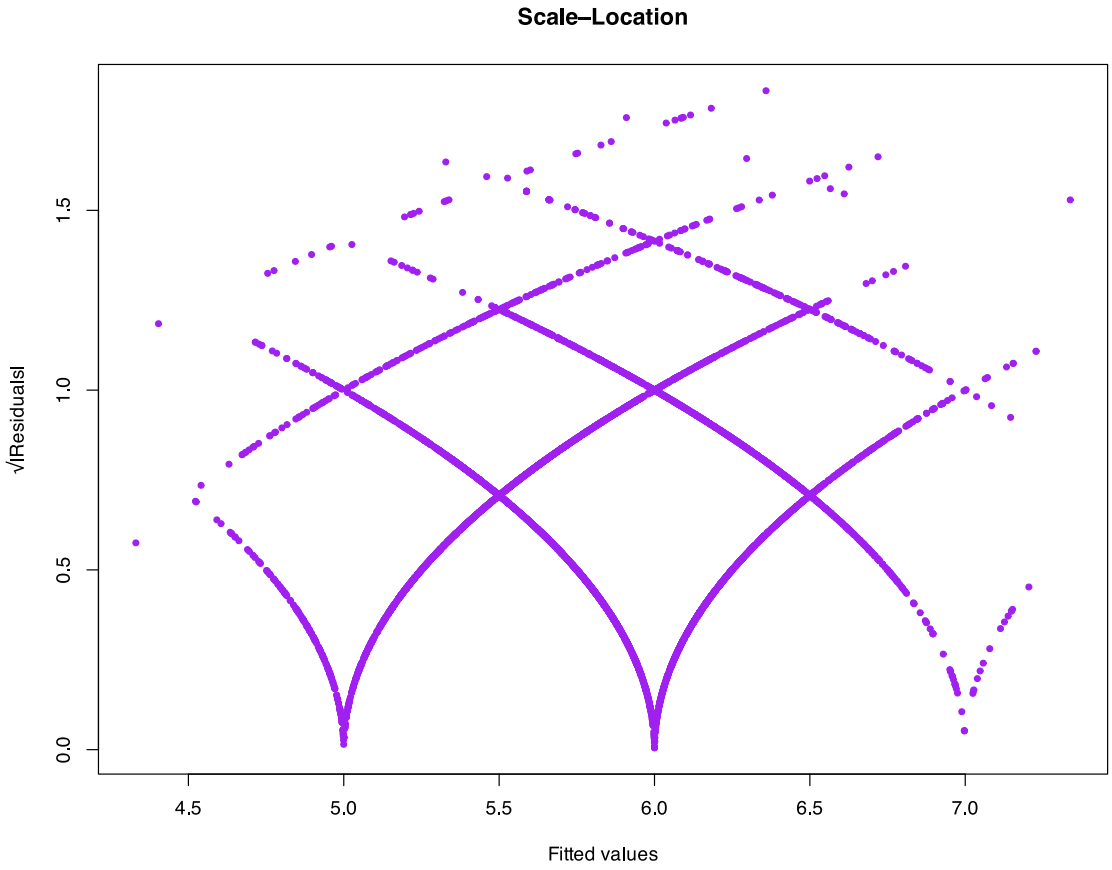

```r
par(mfrow = c(1, 1))

# Residuals vs Fitted
plot(
  fitted(mD), resid(mD),
  pch = 20, col = "darkred",
  xlab = "Fitted values",
  ylab = "Residuals",
  main = "Residuals vs Fitted"
)
abline(h = 0, col = "blue", lwd = 2)
```



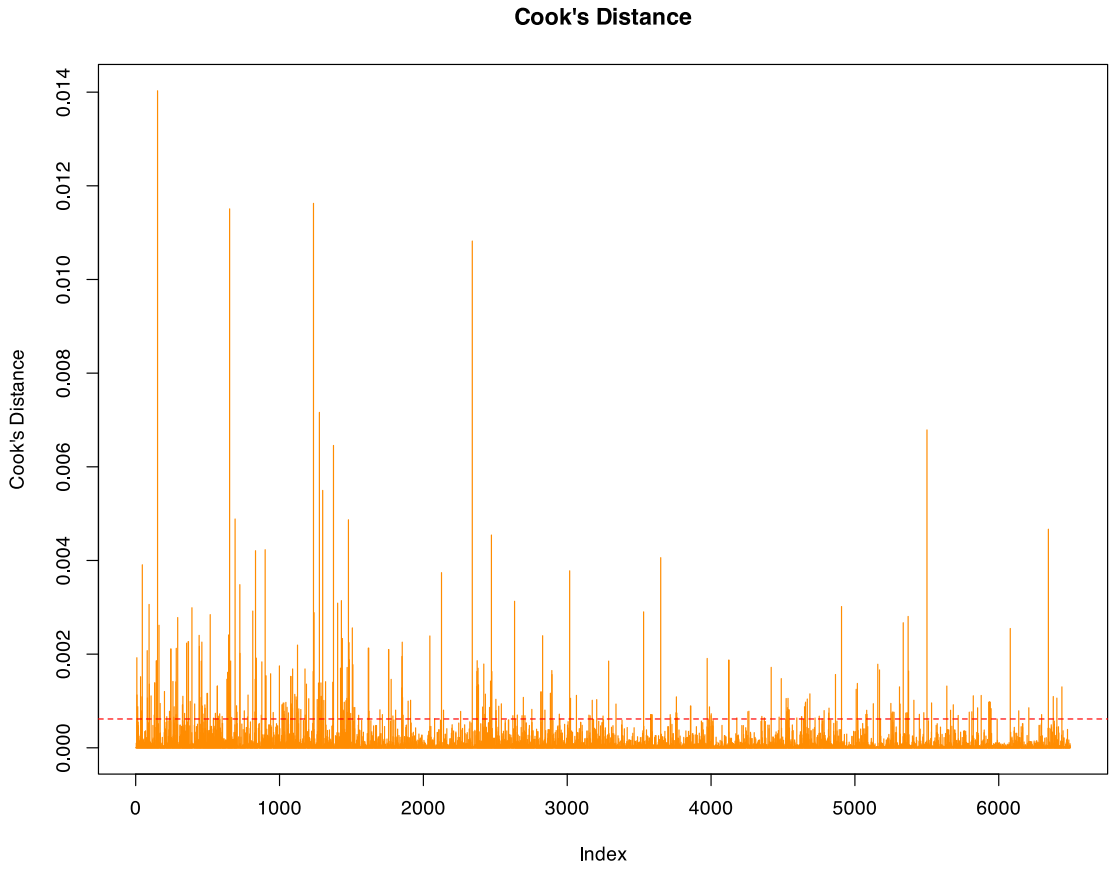

```
# QQ Plot
qqnorm(resid(mD), pch = 20, col = "darkgreen",
       main = "Normal Q-Q Plot")
qqline(resid(mD), col = "blue", lwd = 2)
```



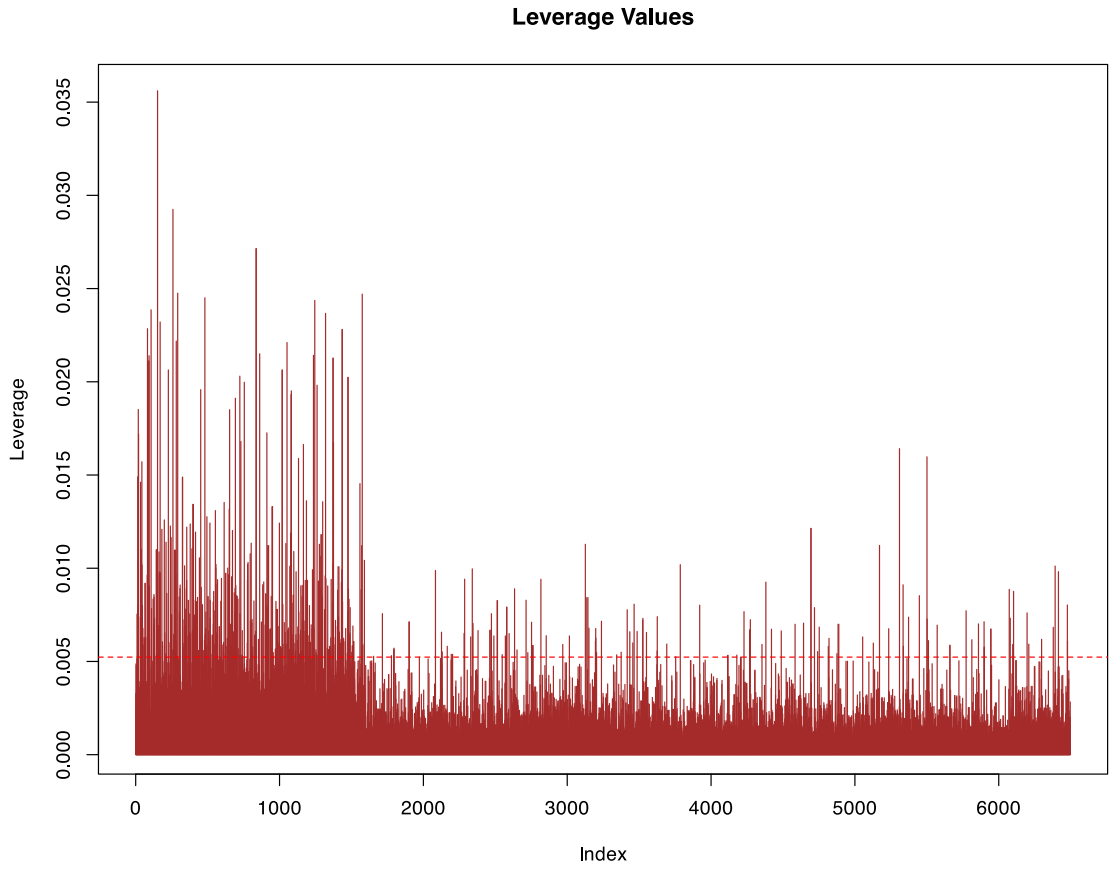

```r
# Scale–Location plot
plot(
  fitted(mD), sqrt(abs(resid(mD))),
  pch = 20, col = "purple",
  xlab = "Fitted values",
  ylab = "√|Residuals|",
  main = "Scale–Location"
)
```



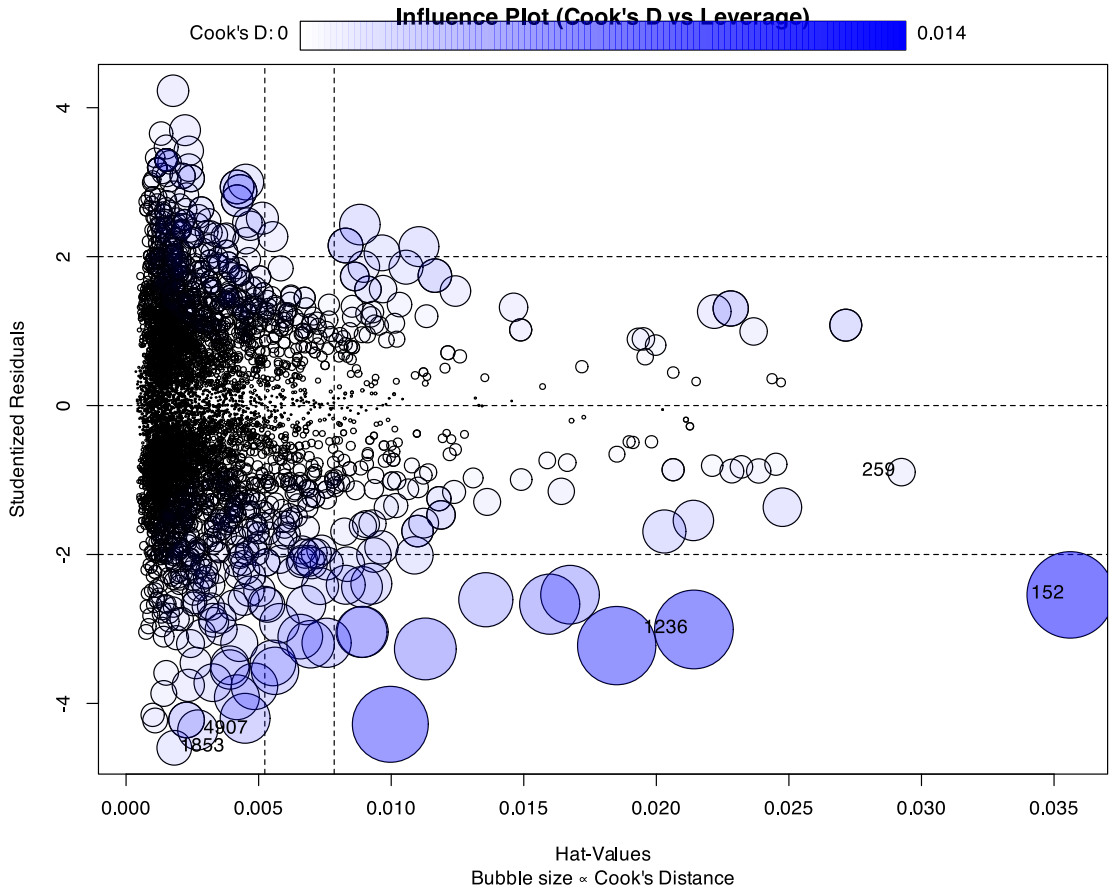

```r
# Cook's Distance
cook <- cooks.distance(mD)
n <- length(cook)
cook_cut <- 4 / n

plot(
  cook, type = "h",
  col = "darkorange",
  main = "Cook's Distance",
  ylab = "Cook's Distance"
)
abline(h = cook_cut, col = "red", lty = 2)
```




```r
# Leverage (hat values)
lev <- hatvalues(mD)
p <- length(coef(mD)) - 1
lev_cut <- 2 * (p + 1) / n

plot(
  lev, type = "h",
  col = "brown",
  main = "Leverage Values",
  ylab = "Leverage"
)
abline(h = lev_cut, col = "red", lty = 2)
```




```r
# Influence plot (optional but strong)
influencePlot(
  mD,
  main = "Influence Plot (Cook's D vs Leverage)",
  sub = "Bubble size ∝ Cook's Distance"
)
```




```
       StudRes         Hat       CookD
152  -2.5425742 0.035610662 0.014030072
259  -0.8940066 0.029246607 0.001416487
1236 -3.0069424 0.021417549 0.011626126
1853 -4.5961419 0.001818063 0.002256267
4907 -4.3563302 0.002702876 0.003017109
```