Red Wine - linear

Katie, Rita, and Chang

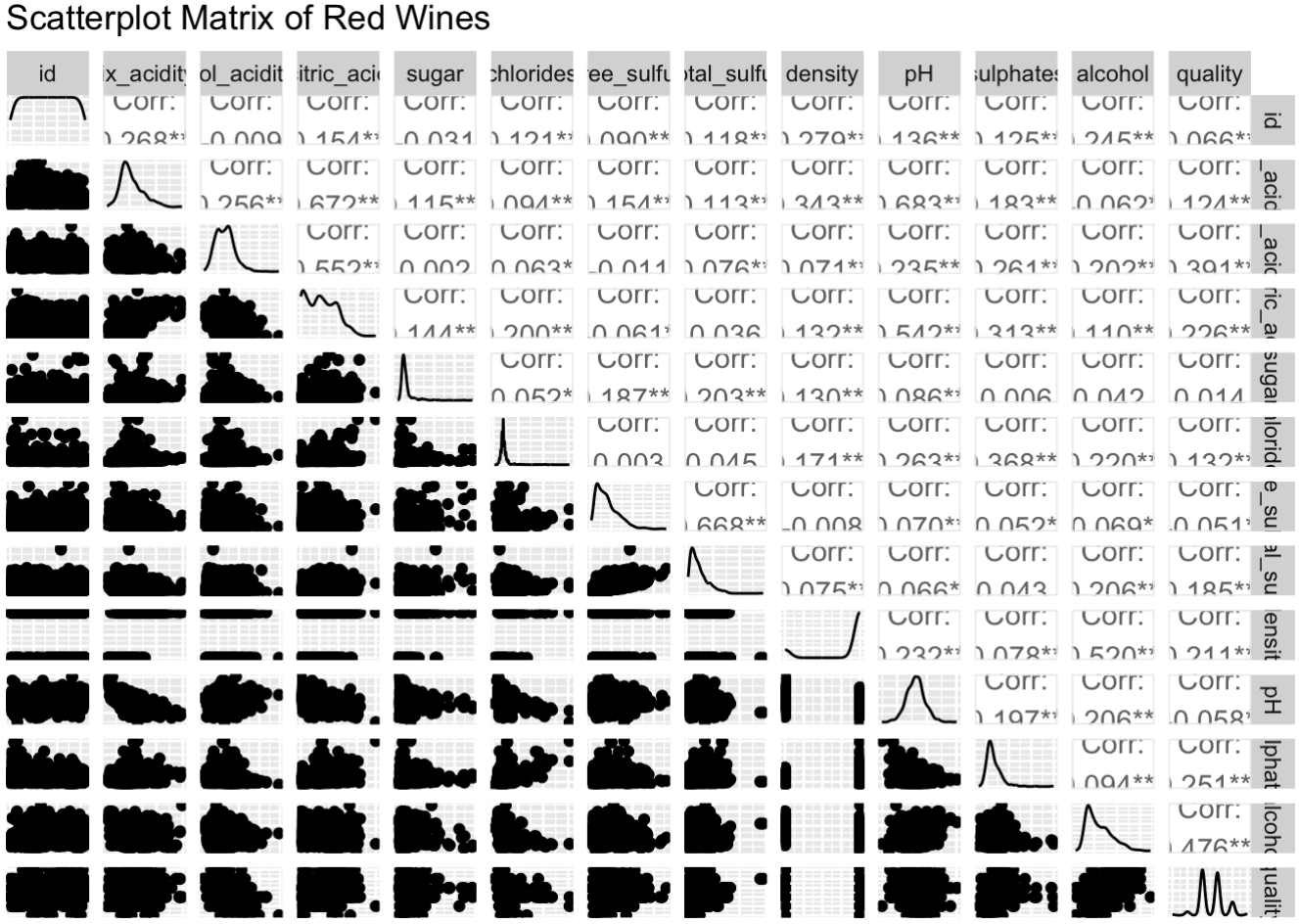
2023-10-23

Scatterplot Matrix

```
library("GGally")
ggpairs(red, axisLabels = "none",
        title = "Scatterplot Matrix of Red Wines")
```

```
# corr codes
```

Scatterplot Matrix



Stepwise Regression

```
library("MASS")
full.red <- lm(quality ~ . - id, data = red)
step.red <- stepAIC(full.red, direction = "both", trace = FALSE)
summary(step.red)
```

```
##
## Call:
## lm(formula = quality ~ vol_acidity + chlorides + free_sulfur +
##     total_sulfur + pH + sulphates + alcohol, data = red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67740 -0.36442 -0.04523  0.46104  2.03542
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4431920  0.4026109  11.036  < 2e-16 ***
## vol_acidity  -1.0066357  0.1004483 -10.021  < 2e-16 ***
## chlorides    -2.0665168  0.3962171  -5.216 2.07e-07 ***
## free_sulfur   0.0050541  0.0021246   2.379   0.0175 *
## total_sulfur -0.0034882  0.0006865  -5.081 4.19e-07 ***
## pH           -0.4854300  0.1174558  -4.133 3.77e-05 ***
## sulphates     0.8870945  0.1097169   8.085 1.22e-15 ***
## alcohol       0.2889443  0.0167839  17.216  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6475 on 1591 degrees of freedom
## Multiple R-squared:   0.36,  Adjusted R-squared:  0.3572
## F-statistic: 127.9 on 7 and 1591 DF,  p-value: < 2.2e-16
```

Stepwise Regression Model

```
library("car")
```

```
## Loading required package: carData
```

```
step <- lm(quality ~ vol_acidity + chlorides + free_sulfur + total_sulfur + pH + sulphat
es + alcohol, data = red)
summary(step)
```

```
##
## Call:
## lm(formula = quality ~ vol_acidity + chlorides + free_sulfur +
##     total_sulfur + pH + sulphates + alcohol, data = red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67740 -0.36442 -0.04523  0.46104  2.03542
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4431920  0.4026109  11.036  < 2e-16 ***
## vol_acidity  -1.0066357  0.1004483 -10.021  < 2e-16 ***
## chlorides    -2.0665168  0.3962171  -5.216 2.07e-07 ***
## free_sulfur   0.0050541  0.0021246   2.379   0.0175 *
## total_sulfur -0.0034882  0.0006865  -5.081 4.19e-07 ***
## pH           -0.4854300  0.1174558  -4.133 3.77e-05 ***
## sulphates     0.8870945  0.1097169   8.085 1.22e-15 ***
## alcohol       0.2889443  0.0167839  17.216  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6475 on 1591 degrees of freedom
## Multiple R-squared:   0.36,  Adjusted R-squared:  0.3572
## F-statistic: 127.9 on 7 and 1591 DF,  p-value: < 2.2e-16
```

```
# Check model for multicollinearity
vif(step)
```

```
##  vol_acidity    chlorides  free_sulfur total_sulfur           pH    sulphates
##     1.241548     1.328518     1.882722     1.943932     1.253520     1.318518
##      alcohol
##     1.219548
```

Forward Selection

```
library("MASS")
library("olsrr")
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:MASS':
##
##     cement
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
full.red <- lm(quality ~ . - id, data = red)
ols_step_forward_aic(full.red, details = TRUE)
```

```
## Forward Selection Method
## -----------------------
##
## Candidate Terms:
##
## 1 . fix_acidity
## 2 . vol_acidity
## 3 . citric_acid
## 4 . sugar
## 5 . chlorides
## 6 . free_sulfur
## 7 . total_sulfur
## 8 . density
## 9 . pH
## 10 . sulphates
## 11 . alcohol
##
##  Step 0: AIC = 3857.269
##  quality ~ 1
##
## ---------------------------------------------------------------------------
## Variable        DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## alcohol          1    3448.117    236.293     805.872    0.227         0.226
## vol_acidity      1    3594.696    158.927     883.238    0.152         0.152
## sulphates        1    3754.876     65.865     976.300    0.063         0.063
## citric_acid      1    3775.155     53.405     988.760    0.051         0.051
## density          1    3786.640     46.278     995.887    0.044         0.044
## total_sulfur     1    3803.523     35.707    1006.458    0.034         0.034
## chlorides        1    3831.267     18.092    1024.073    0.017         0.017
## fix_acidity      1    3834.471     16.038    1026.127    0.015         0.015
## pH               1    3853.930      3.473    1038.692    0.003         0.003
## free_sulfur      1    3855.160      2.674    1039.491    0.003         0.002
## sugar            1    3858.967      0.197    1041.969    0.000         0.000
## ---------------------------------------------------------------------------
##
##
## - alcohol
##
##
##  Step 1 : AIC = 3448.117
##  quality ~ alcohol
##
## ---------------------------------------------------------------------------
## Variable        DF      AIC       Sum Sq      RSS      R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## vol_acidity      1    3251.709    94.040    711.832    0.317         0.316
## sulphates        1    3358.288    44.977    760.896    0.270         0.269
## citric_acid      1    3385.422    31.955    773.918    0.257         0.256
## pH               1    3396.935    26.362    779.510    0.252         0.251
## fix_acidity      1    3400.496    24.624    781.248    0.250         0.249
## total_sulfur     1    3433.622     8.271    797.602    0.235         0.234
```

```
## density          1    3446.224     1.960    803.913    0.229         0.228
## chlorides        1    3448.542     0.794    805.079    0.227         0.227
## free_sulfur      1    3449.473     0.325    805.548    0.227         0.226
## sugar            1    3450.035     0.041    805.831    0.227         0.226
## ---------------------------------------------------------------------------
##
## - vol_acidity
##
##
##  Step 2 : AIC = 3251.709
##  quality ~ alcohol + vol_acidity
##
## ---------------------------------------------------------------------------
## Variable        DF      AIC       Sum Sq      RSS      R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## sulphates        1    3208.804    19.712    692.120    0.336         0.335
## total_sulfur     1    3239.286     6.392    705.441    0.323         0.322
## pH               1    3240.231     5.975    705.858    0.323         0.321
## fix_acidity      1    3240.791     5.728    706.105    0.322         0.321
## density          1    3251.435     1.012    710.821    0.318         0.317
## free_sulfur      1    3252.203     0.670    711.162    0.318         0.316
## chlorides        1    3252.589     0.498    711.334    0.317         0.316
## citric_acid      1    3253.266     0.197    711.635    0.317         0.316
## sugar            1    3253.688     0.009    711.823    0.317         0.316
## ---------------------------------------------------------------------------
##
## - sulphates
##
##
##  Step 3 : AIC = 3208.804
##  quality ~ alcohol + vol_acidity + sulphates
##
## ---------------------------------------------------------------------------
## Variable        DF      AIC       Sum Sq      RSS      R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## total_sulfur     1    3191.661     8.237    683.883    0.344         0.342
## chlorides        1    3192.287     7.969    684.151    0.344         0.342
## fix_acidity      1    3203.066     3.341    688.779    0.339         0.337
## pH               1    3203.722     3.059    689.062    0.339         0.337
## free_sulfur      1    3208.209     1.123    690.998    0.337         0.335
## citric_acid      1    3210.228     0.249    691.871    0.336         0.334
## density          1    3210.487     0.137    691.983    0.336         0.334
## sugar            1    3210.773     0.014    692.107    0.336         0.334
## ---------------------------------------------------------------------------
##
## - total_sulfur
##
##
##  Step 4 : AIC = 3191.661
##  quality ~ alcohol + vol_acidity + sulphates + total_sulfur
##
## ---------------------------------------------------------------------------
```

```
## Variable       DF      AIC       Sum Sq      RSS      R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## chlorides       1    3173.497     8.570    675.314    0.352         0.350
## pH              1    3185.874     3.322    680.561    0.347         0.345
## fix_acidity     1    3188.715     2.112    681.771    0.346         0.344
## free_sulfur     1    3190.504     1.349    682.535    0.345         0.343
## sugar           1    3193.036     0.267    683.616    0.344         0.342
## citric_acid     1    3193.493     0.072    683.812    0.344         0.342
## density         1    3193.536     0.053    683.830    0.344         0.342
## ------------------------------------------------------------------------
##
## - chlorides
##
##
##  Step 5 : AIC = 3173.497
##  quality ~ alcohol + vol_acidity + sulphates + total_sulfur + chlorides
##
## ------------------------------------------------------------------------
## Variable       DF      AIC       Sum Sq      RSS      R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## pH              1    3161.209     6.008    669.306    0.358         0.355
## fix_acidity     1    3169.736     2.429    672.885    0.354         0.352
## free_sulfur     1    3172.606     1.220    674.094    0.353         0.351
## sugar           1    3174.188     0.553    674.761    0.353         0.350
## citric_acid     1    3175.087     0.173    675.140    0.352         0.350
## density         1    3175.345     0.064    675.250    0.352         0.350
## ------------------------------------------------------------------------
##
## - pH
##
##
##  Step 6 : AIC = 3161.209
##  quality ~ alcohol + vol_acidity + sulphates + total_sulfur + chlorides + pH
##
## ------------------------------------------------------------------------
## Variable       DF      AIC       Sum Sq      RSS      R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## free_sulfur     1    3157.531     2.372    666.934    0.360         0.357
## citric_acid     1    3161.300     0.798    668.508    0.359         0.356
## sugar           1    3162.541     0.280    669.026    0.358         0.355
## fix_acidity     1    3163.181     0.012    669.294    0.358         0.355
## density         1    3163.201     0.003    669.303    0.358         0.355
## ------------------------------------------------------------------------
##
## - free_sulfur
##
##
##  Step 7 : AIC = 3157.531
##  quality ~ alcohol + vol_acidity + sulphates + total_sulfur + chlorides + pH + free_s
ulfur
##
## ------------------------------------------------------------------------
```

```
## Variable       DF      AIC       Sum Sq      RSS      R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## citric_acid     1    3158.401     0.471    666.462    0.361         0.357
## sugar           1    3159.139     0.164    666.770    0.360         0.357
## fix_acidity     1    3159.513     0.008    666.926    0.360         0.357
## density         1    3159.530     0.001    666.933    0.360         0.357
## ------------------------------------------------------------------------
##
##
## No more variables to be added.
##
## Variables Entered:
##
## - alcohol
## - vol_acidity
## - sulphates
## - total_sulfur
## - chlorides
## - pH
## - free_sulfur
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.600       RMSE                 0.647
## R-Squared               0.360       Coef. Var           11.488
## Adj. R-Squared          0.357       MSE                  0.419
## Pred R-Squared          0.352       MAE                  0.501
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ---------------------------------------------------------------------------
##                Sum of
##               Squares          DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------------
## Regression     375.231           7         53.604    127.876    0.0000
## Residual       666.934        1591          0.419
## Total         1042.165        1598
## ---------------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------
-----
##        model      Beta    Std. Error    Std. Beta       t        Sig      lower     u
pper
## ----------------------------------------------------------------------------------
```

```
-----
##  (Intercept)     4.443         0.403                11.036    0.000     3.653
5.233
##      alcohol     0.289         0.017         0.381  17.216    0.000     0.256
0.322
##  vol_acidity    -1.007         0.100        -0.224 -10.021    0.000    -1.204    -
0.810
##    sulphates     0.887         0.110         0.186   8.085    0.000     0.672
1.102
## total_sulfur    -0.003         0.001        -0.142  -5.081    0.000    -0.005    -
0.002
##    chlorides    -2.067         0.396        -0.121  -5.216    0.000    -2.844    -
1.289
##           pH    -0.485         0.117        -0.093  -4.133    0.000    -0.716    -
0.255
##  free_sulfur     0.005         0.002         0.065   2.379    0.017     0.001
0.009
## ----------------------------------------------------------------------------------
-----
```

```
##
##                            Selection Summary
## ------------------------------------------------------------------------
## Variable          AIC       Sum Sq       RSS       R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------
## alcohol         3448.117    236.293    805.872    0.22673       0.22625
## vol_acidity     3251.709    330.333    711.832    0.31697       0.31611
## sulphates       3208.804    350.045    692.120    0.33588       0.33463
## total_sulfur    3191.661    358.282    683.883    0.34379       0.34214
## chlorides       3173.497    366.852    675.314    0.35201       0.34998
## pH              3161.209    372.859    669.306    0.35777       0.35535
## free_sulfur     3157.531    375.231    666.934    0.36005       0.35723
## ------------------------------------------------------------------------
```

```
# results in same model as stepwise regression
```

Backward Elimination

```
library("MASS")
library("olsrr")
full.red <- lm(quality ~ . - id, data = red)
ols_step_backward_aic(full.red, details = TRUE)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . fix_acidity
## 2 . vol_acidity
## 3 . citric_acid
## 4 . sugar
## 5 . chlorides
## 6 . free_sulfur
## 7 . total_sulfur
## 8 . density
## 9 . pH
## 10 . sulphates
## 11 . alcohol
##
##  Step 0: AIC = 3163.532
##  quality ~ fix_acidity + vol_acidity + citric_acid + sugar + chlorides + free_sulfur
+ total_sulfur + density + pH + sulphates + alcohol
##
## ------------------------------------------------------------------------
## Variable        DF      AIC       Sum Sq      RSS      R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## density         1     3161.549     0.007    666.107    0.361        0.357
## fix_acidity     1     3161.792     0.108    666.209    0.361        0.357
## sugar           1     3162.099     0.236    666.337    0.361        0.357
## citric_acid     1     3163.081     0.645    666.746    0.360        0.356
## free_sulfur     1     3165.861     1.806    667.906    0.359        0.355
## pH              1     3171.666     4.235    670.336    0.357        0.353
## total_sulfur    1     3182.451     8.771    674.872    0.352        0.348
## chlorides       1     3183.743     9.317    675.417    0.352        0.348
## sulphates       1     3226.058    27.429    693.530    0.335        0.330
## vol_acidity     1     3242.884    34.766    700.866    0.327        0.323
## alcohol         1     3357.685    86.935    753.036    0.277        0.273
## ------------------------------------------------------------------------
##
##
## Variables Removed:
##
## - density
##
##
##   Step 1 : AIC = 3161.549
##  quality ~ fix_acidity + vol_acidity + citric_acid + sugar + chlorides + free_sulfur
+ total_sulfur + pH + sulphates + alcohol
##
## ------------------------------------------------------------------------
## Variable        DF      AIC       Sum Sq       RSS      R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## fix_acidity     1     3159.793      0.102    666.209    0.361        0.357
## sugar           1     3160.099      0.229    666.337    0.361        0.357
```

```
## citric_acid     1     3161.089      0.642    666.750    0.360         0.357
## free_sulfur     1     3163.898      1.814    667.922    0.359         0.355
## pH              1     3170.068      4.397    670.504    0.357         0.353
## total_sulfur    1     3180.491      8.781    674.889    0.352         0.349
## chlorides       1     3181.808      9.338    675.445    0.352         0.348
## sulphates       1     3224.593     27.655    693.762    0.334         0.331
## vol_acidity     1     3241.046     34.830    700.938    0.327         0.324
## alcohol         1     3423.392    119.499    785.607    0.246         0.242
## ----------------------------------------------------------------------------
##
## - fix_acidity
##
##
##   Step 2 : AIC = 3159.793
##  quality ~ vol_acidity + citric_acid + sugar + chlorides + free_sulfur + total_sulfur
+ pH + sulphates + alcohol
##
## ----------------------------------------------------------------------------
## Variable        DF       AIC       Sum Sq       RSS      R-Sq     Adj. R-Sq
## ----------------------------------------------------------------------------
## sugar           1     3158.401      0.253    666.462    0.361         0.357
## citric_acid     1     3159.139      0.561    666.770    0.360         0.357
## free_sulfur     1     3162.307      1.883    668.093    0.359         0.356
## pH              1     3174.821      7.132    673.342    0.354         0.351
## total_sulfur    1     3181.776     10.068    676.277    0.351         0.348
## chlorides       1     3182.855     10.524    676.733    0.351         0.347
## sulphates       1     3223.664     28.018    694.227    0.334         0.331
## vol_acidity     1     3242.088     36.063    702.272    0.326         0.323
## alcohol         1     3423.531    120.449    786.658    0.245         0.241
## ----------------------------------------------------------------------------
##
## - sugar
##
##
##   Step 3 : AIC = 3158.401
##  quality ~ vol_acidity + citric_acid + chlorides + free_sulfur + total_sulfur + pH +
sulphates + alcohol
##
## ----------------------------------------------------------------------------
## Variable        DF       AIC       Sum Sq       RSS      R-Sq     Adj. R-Sq
## ----------------------------------------------------------------------------
## citric_acid     1     3157.531      0.471    666.934    0.360         0.357
## free_sulfur     1     3161.300      2.045    668.508    0.359         0.356
## pH              1     3173.597      7.206    673.668    0.354         0.351
## total_sulfur    1     3179.899      9.866    676.329    0.351         0.348
## chlorides       1     3181.212     10.422    676.884    0.351         0.348
## sulphates       1     3221.735     27.795    694.258    0.334         0.331
## vol_acidity     1     3240.089     35.810    702.273    0.326         0.323
## alcohol         1     3426.106    122.450    788.912    0.243         0.240
## ----------------------------------------------------------------------------
##
## - citric_acid
```

```
##
##
##   Step 4 : AIC = 3157.531
##  quality ~ vol_acidity + chlorides + free_sulfur + total_sulfur + pH + sulphates + al
cohol
##
## ------------------------------------------------------------------------
## Variable        DF      AIC       Sum Sq       RSS      R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## free_sulfur     1     3161.209      2.372    669.306    0.358         0.355
## pH              1     3172.606      7.160    674.094    0.353         0.351
## total_sulfur    1     3181.272     10.823    677.757    0.350         0.347
## chlorides       1     3182.640     11.403    678.337    0.349         0.347
## sulphates       1     3219.918     27.403    694.337    0.334         0.331
## vol_acidity     1     3253.408     42.099    709.033    0.320         0.317
## alcohol         1     3428.681    124.239    791.173    0.241         0.238
## ------------------------------------------------------------------------
##
##
## No more variables to be removed.
##
## Variables Removed:
##
## - density
## - fix_acidity
## - sugar
## - citric_acid
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.600       RMSE                 0.647
## R-Squared               0.360       Coef. Var           11.488
## Adj. R-Squared          0.357       MSE                  0.419
## Pred R-Squared          0.352       MAE                  0.501
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## ---------------------------------------------------------------------------
##                 Sum of
##                Squares          DF    Mean Square       F          Sig.
## ---------------------------------------------------------------------------
## Regression     375.231           7         53.604    127.876    0.0000
## Residual       666.934        1591          0.419
## Total         1042.165        1598
## ---------------------------------------------------------------------------
```

```
##
##                                 Parameter Estimates
## ----------------------------------------------------------------------------
-----
##        model      Beta    Std. Error    Std. Beta       t        Sig      lower     u
pper
## ----------------------------------------------------------------------------
-----
##  (Intercept)     4.443         0.403                11.036    0.000     3.653
5.233
##  vol_acidity    -1.007         0.100       -0.224   -10.021    0.000    -1.204    -
0.810
##    chlorides    -2.067         0.396       -0.121    -5.216    0.000    -2.844    -
1.289
##  free_sulfur     0.005         0.002        0.065     2.379    0.017     0.001
0.009
## total_sulfur    -0.003         0.001       -0.142    -5.081    0.000    -0.005    -
0.002
##           pH    -0.485         0.117       -0.093    -4.133    0.000    -0.716    -
0.255
##    sulphates     0.887         0.110        0.186     8.085    0.000     0.672
1.102
##      alcohol     0.289         0.017        0.381    17.216    0.000     0.256
0.322
## ----------------------------------------------------------------------------
-----
```

```
##
##
##                       Backward Elimination Summary
## ---------------------------------------------------------------------
## Variable         AIC         RSS       Sum Sq      R-Sq      Adj. R-Sq
## ---------------------------------------------------------------------
## Full Model     3163.532    666.101    376.064    0.36085       0.35642
## density        3161.549    666.107    376.058    0.36084       0.35682
## fix_acidity    3159.793    666.209    375.956    0.36074       0.35712
## sugar          3158.401    666.462    375.703    0.36050       0.35728
## citric_acid    3157.531    666.934    375.231    0.36005       0.35723
## ---------------------------------------------------------------------
```

```
# results in same model as stepwise regression
```

Standardize the variables

```
library("tidyverse")
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ tibble  3.2.1     ✔ dplyr   1.1.3
## ✔ tidyr   1.2.0     ✔ stringr 1.4.1
## ✔ readr   2.1.2     ✔ forcats 0.5.2
## ✔ purrr   1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ dplyr::recode() masks car::recode()
## ✖ dplyr::select() masks MASS::select()
## ✖ purrr::some()   masks car::some()
```

```
library("broom")
library("mosaic")
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected
by this.
##
## Attaching package: 'mosaic'
##
## The following object is masked from 'package:Matrix':
##
##     mean
##
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
##
## The following object is masked from 'package:purrr':
##
##     cross
##
## The following objects are masked from 'package:car':
##
##     deltaMethod, logit
##
## The following object is masked from 'package:ggplot2':
##
##     stat
##
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
##
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
red_standardized <-
  red %>%
  mutate(fix_acidity = scale(fix_acidity), vol_acidity = scale(vol_acidity), citric_acid
= scale(citric_acid), sugar = scale(sugar), chlorides = scale(chlorides), free_sulfur =
scale(free_sulfur), total_sulfur = scale(total_sulfur), density = scale(density), pH = s
cale(pH), sulphates = scale(sulphates), alcohol = scale(alcohol))
```

Lasso Regression

```
library("glmnet")
```

```
## Loaded glmnet 4.1-4
```

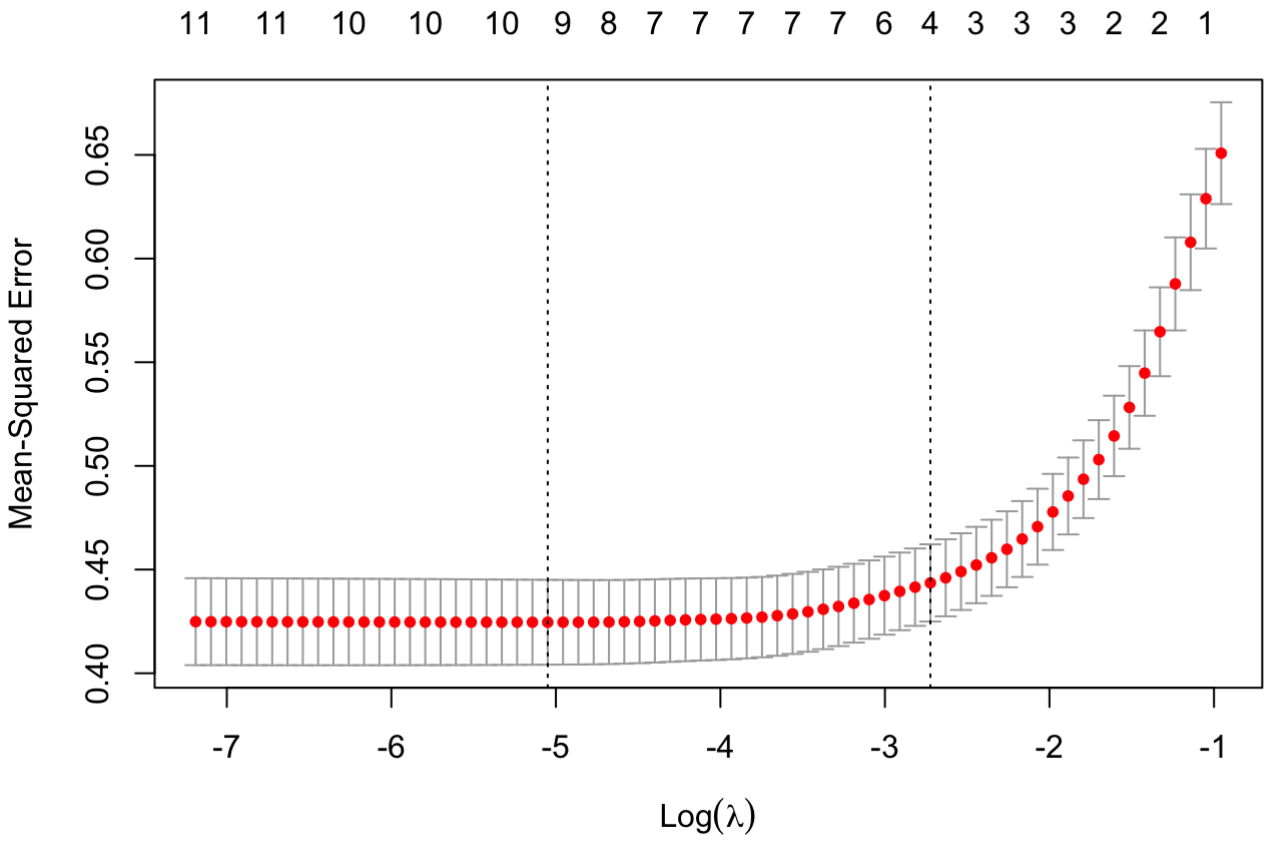
```
x = data.matrix(red_standardized [,c('fix_acidity', 'vol_acidity', 'citric_acid', 'suga
r', 'chlorides', 'free_sulfur', 'total_sulfur', 'density', 'pH', 'sulphates', 'alcoho
l')])
y = red_standardized$quality
model <- glmnet(x, y, alpha = 1)
summary(model)
```

```
##           Length Class     Mode
## a0         68    -none-    numeric
## beta      748    dgCMatrix S4
## df         68    -none-    numeric
## dim         2    -none-    numeric
## lambda     68    -none-    numeric
## dev.ratio  68    -none-    numeric
## nulldev     1    -none-    numeric
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
## offset      1    -none-    logical
## call        4    -none-    call
## nobs        1    -none-    numeric
```

```
cv_model <- cv.glmnet(x, y, alpha = 1)
best_lambda <- cv_model$lambda.min
best_lambda
```

```
## [1] 0.006412439
```

```
plot(cv_model)
```




```
best_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)
coef(best_model)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)   5.636022514
## fix_acidity   .
## vol_acidity  -0.183915770
## citric_acid  -0.003404884
## sugar         0.004308949
## chlorides    -0.086270184
## free_sulfur   0.031317528
## total_sulfur -0.094597059
## density       .
## pH           -0.064337786
## sulphates     0.142745794
## alcohol       0.304703641
```

Lasso Regression Model

```
library("car")
lassomod <- lm(quality ~ vol_acidity + sugar + chlorides + free_sulfur + total_sulfur +
pH + sulphates + alcohol, data = red_standardized)
summary(lassomod)
```

```
##
## Call:
## lm(formula = quality ~ vol_acidity + sugar + chlorides + free_sulfur +
##     total_sulfur + pH + sulphates + alcohol, data = red_standardized)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.66876 -0.36182 -0.04612  0.46409  2.04600
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.63602    0.01619 348.023  < 2e-16 ***
## vol_acidity  -0.18111    0.01806 -10.031  < 2e-16 ***
## sugar         0.01048    0.01678   0.625   0.5322
## chlorides    -0.09803    0.01870  -5.242 1.80e-07 ***
## free_sulfur   0.05173    0.02230   2.320   0.0205 *
## total_sulfur -0.11629    0.02272  -5.118 3.46e-07 ***
## pH           -0.07379    0.01823  -4.048 5.42e-05 ***
## sulphates     0.15096    0.01863   8.105 1.04e-15 ***
## alcohol       0.30659    0.01802  17.018  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6476 on 1590 degrees of freedom
## Multiple R-squared:  0.3602, Adjusted R-squared:  0.357
## F-statistic: 111.9 on 8 and 1590 DF,  p-value: < 2.2e-16
```

```
# Check model for multicollinearity
vif(lassomod)
```

```
##  vol_acidity        sugar    chlorides  free_sulfur total_sulfur           pH
##     1.242208     1.072893     1.332859     1.895322     1.967157     1.266457
##    sulphates      alcohol
##     1.321981     1.236792
```

Lasso Regression max of 5 variables

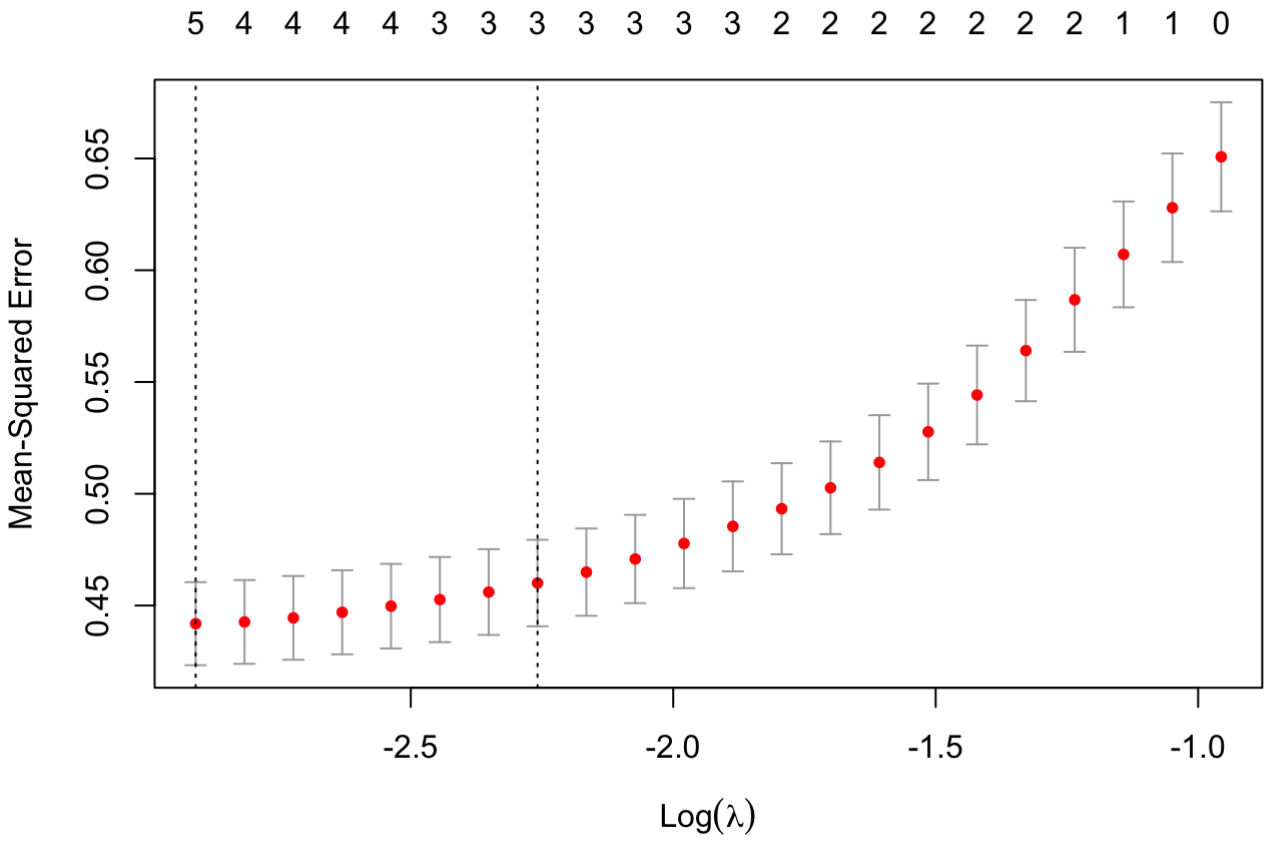
```
library("glmnet")
x = data.matrix(red_standardized [,c('fix_acidity', 'vol_acidity', 'citric_acid', 'suga
r', 'chlorides', 'free_sulfur', 'total_sulfur', 'density', 'pH', 'sulphates', 'alcoho
l')])
y = red_standardized$quality
model <- glmnet(x, y, alpha = 1)
summary(model)
```

```
##           Length Class     Mode
## a0         68    -none-    numeric
## beta      748    dgCMatrix S4
## df         68    -none-    numeric
## dim         2    -none-    numeric
## lambda     68    -none-    numeric
## dev.ratio  68    -none-    numeric
## nulldev     1    -none-    numeric
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
## offset      1    -none-    logical
## call        4    -none-    call
## nobs        1    -none-    numeric
```

```
cv_model <- cv.glmnet(x, y, alpha = 1, dfmax = 4)
best_lambda <- cv_model$lambda.min
best_lambda
```

```
## [1] 0.05448992
```

```
plot(cv_model)
```




```
best_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)
coef(best_model)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)   5.636022514
## fix_acidity   .
## vol_acidity  -0.181394342
## citric_acid   .
## sugar         .
## chlorides    -0.006935839
## free_sulfur   .
## total_sulfur -0.026885174
## density       .
## pH            .
## sulphates     0.078816028
## alcohol       0.278900058
```

Lasso Regression Model max of 5 variables

```
library("car")
lassomod5 <- lm(quality ~ vol_acidity + chlorides +  total_sulfur + sulphates + alcohol,
data = red_standardized)
summary(lassomod5)
```

```
##
## Call:
## lm(formula = quality ~ vol_acidity + chlorides + total_sulfur +
##     sulphates + alcohol, data = red_standardized)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.66489 -0.38056 -0.06617  0.44728  2.06739
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.63602    0.01628 346.140  < 2e-16 ***
## vol_acidity  -0.20406    0.01735 -11.760  < 2e-16 ***
## chlorides    -0.08273    0.01840  -4.496 7.42e-06 ***
## total_sulfur -0.07631    0.01671  -4.566 5.35e-06 ***
## sulphates     0.15597    0.01866   8.360  < 2e-16 ***
## alcohol       0.29475    0.01755  16.792  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6511 on 1593 degrees of freedom
## Multiple R-squared:  0.352,  Adjusted R-squared:   0.35
## F-statistic: 173.1 on 5 and 1593 DF,  p-value: < 2.2e-16
```

```
# Check model for multicollinearity
vif(lassomod5)
```

```
##  vol_acidity    chlorides total_sulfur    sulphates      alcohol
##     1.134978     1.276137     1.052800     1.312253     1.161413
```

Ridge Regression

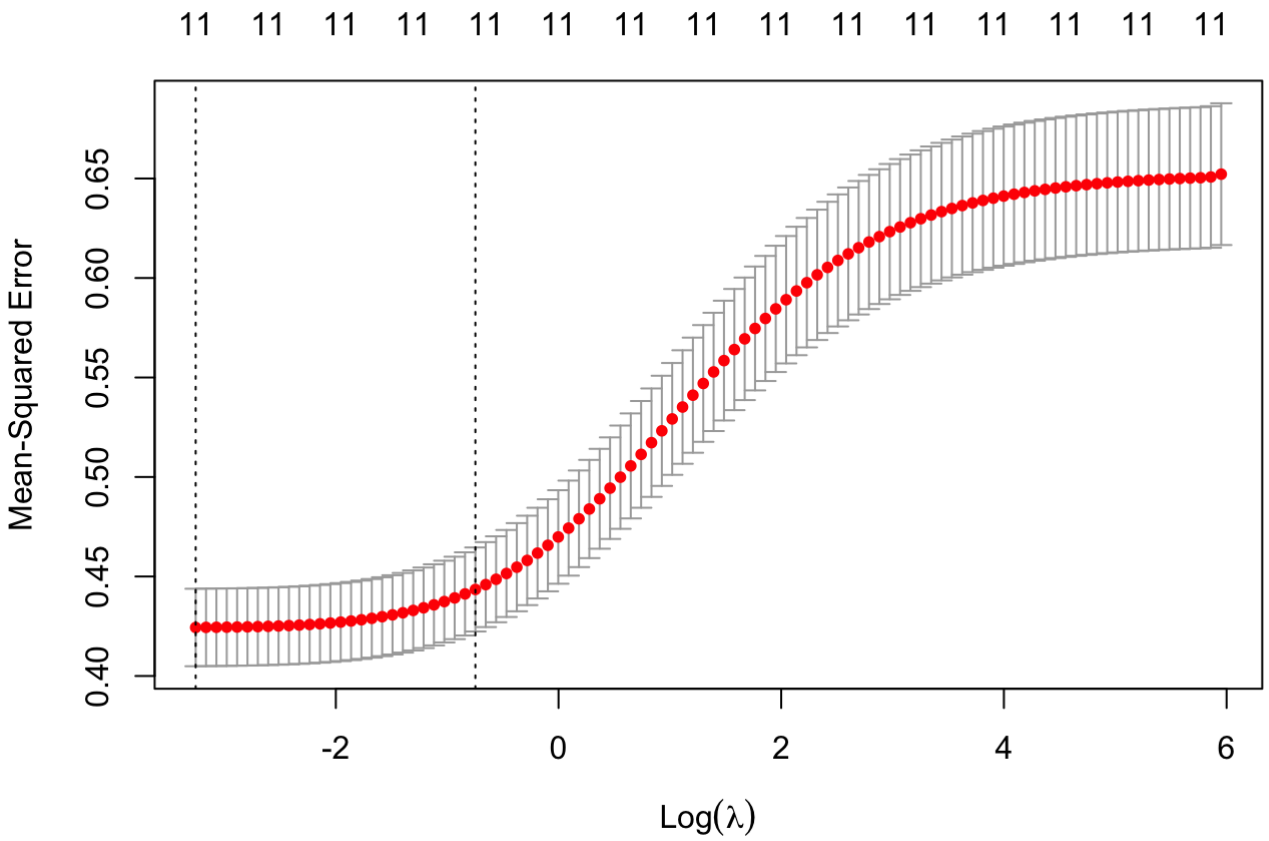
```
library("glmnet")
x = data.matrix(red_standardized [,c('fix_acidity', 'vol_acidity', 'citric_acid', 'suga
r', 'chlorides', 'free_sulfur', 'total_sulfur', 'density', 'pH', 'sulphates', 'alcoho
l')])
y = red_standardized$quality
model <- glmnet(x, y, alpha = 0)
summary(model)
```

```
##           Length Class      Mode
## a0         100   -none-     numeric
## beta      1100   dgCMatrix S4
## df         100   -none-     numeric
## dim          2   -none-     numeric
## lambda     100   -none-     numeric
## dev.ratio  100   -none-     numeric
## nulldev      1   -none-     numeric
## npasses      1   -none-     numeric
## jerr         1   -none-     numeric
## offset       1   -none-     logical
## call         4   -none-     call
## nobs         1   -none-     numeric
```

```
cv_model <- cv.glmnet(x, y, alpha = 0)
best_lambda <- cv_model$lambda.min
best_lambda
```

```
## [1] 0.03844156
```

```
plot(cv_model)
```



```
best_model <- glmnet(x, y, alpha = 0, lambda = best_lambda)
coef(best_model)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept)   5.63602251
## fix_acidity   0.01850619
## vol_acidity  -0.18602055
## citric_acid  -0.02004038
## sugar         0.01293353
## chlorides    -0.08905563
## free_sulfur   0.04021277
## total_sulfur -0.10293346
## density      -0.01374573
## pH           -0.06418478
## sulphates     0.14677572
## alcohol       0.29020924
```

Ridge Regression Model

```
library("car")
ridgemod <- lm(quality ~ vol_acidity + chlorides + total_sulfur + pH + sulphates + alcoh
ol, data = red_standardized)
summary(ridgemod)
```

```
##
## Call:
## lm(formula = quality ~ vol_acidity + chlorides + total_sulfur +
##     pH + sulphates + alcohol, data = red_standardized)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59436 -0.35904 -0.04463  0.45945  1.96250
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.63602    0.01621 347.581  < 2e-16 ***
## vol_acidity  -0.18539    0.01797 -10.315  < 2e-16 ***
## chlorides    -0.09672    0.01869  -5.174 2.58e-07 ***
## total_sulfur -0.07840    0.01665  -4.708 2.72e-06 ***
## pH           -0.06766    0.01790  -3.780 0.000163 ***
## sulphates     0.15142    0.01862   8.132 8.40e-16 ***
## alcohol       0.30936    0.01790  17.280  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6484 on 1592 degrees of freedom
## Multiple R-squared:  0.3578, Adjusted R-squared:  0.3554
## F-statistic: 147.8 on 6 and 1592 DF,  p-value: < 2.2e-16
```

```
# Check model for multicollinearity
vif(ridgemod)
```

```
##  vol_acidity    chlorides total_sulfur           pH    sulphates      alcohol
##     1.227707     1.328238     1.053963     1.217808     1.317778     1.218161
```

Select Best Model with AIC

```
library(AICcmodavg)
models <- list(step, lassomod, ridgemod, lassomod5)
mod.names <- c('Stepwise', 'Lasso', 'Ridge', 'Limited Lasso')
aictab(cand.set = models, modnames = mod.names)
```

```
##
## Model selection based on AICc:
##
##                K    AICc Delta_AICc AICcWt Cum.Wt       LL
## Stepwise       9 3157.64       0.00   0.62   0.62 -1569.77
## Lasso         10 3159.28       1.63   0.28   0.90 -1569.57
## Ridge          8 3161.30       3.65   0.10   1.00 -1572.60
## Limited Lasso  7 3173.57      15.92   0.00   1.00 -1579.75
```

```
# step is best model
```

Select Best Model with BIC

```
library("flexmix")
BIC(step)
```

```
## [1] 3205.926
```

```
BIC(lassomod)
```

```
## [1] 3212.91
```

```
BIC(ridgemod)
```

```
## [1] 3204.226
```

```
BIC(lassomod5)
```

```
## [1] 3211.137
```

```
# ridge is best model
# however, since the step model performed better with the adjusted r-squared and aic tes
ts, we will move forward with that model
```

From this point forward, we are using the stepwise regression model.

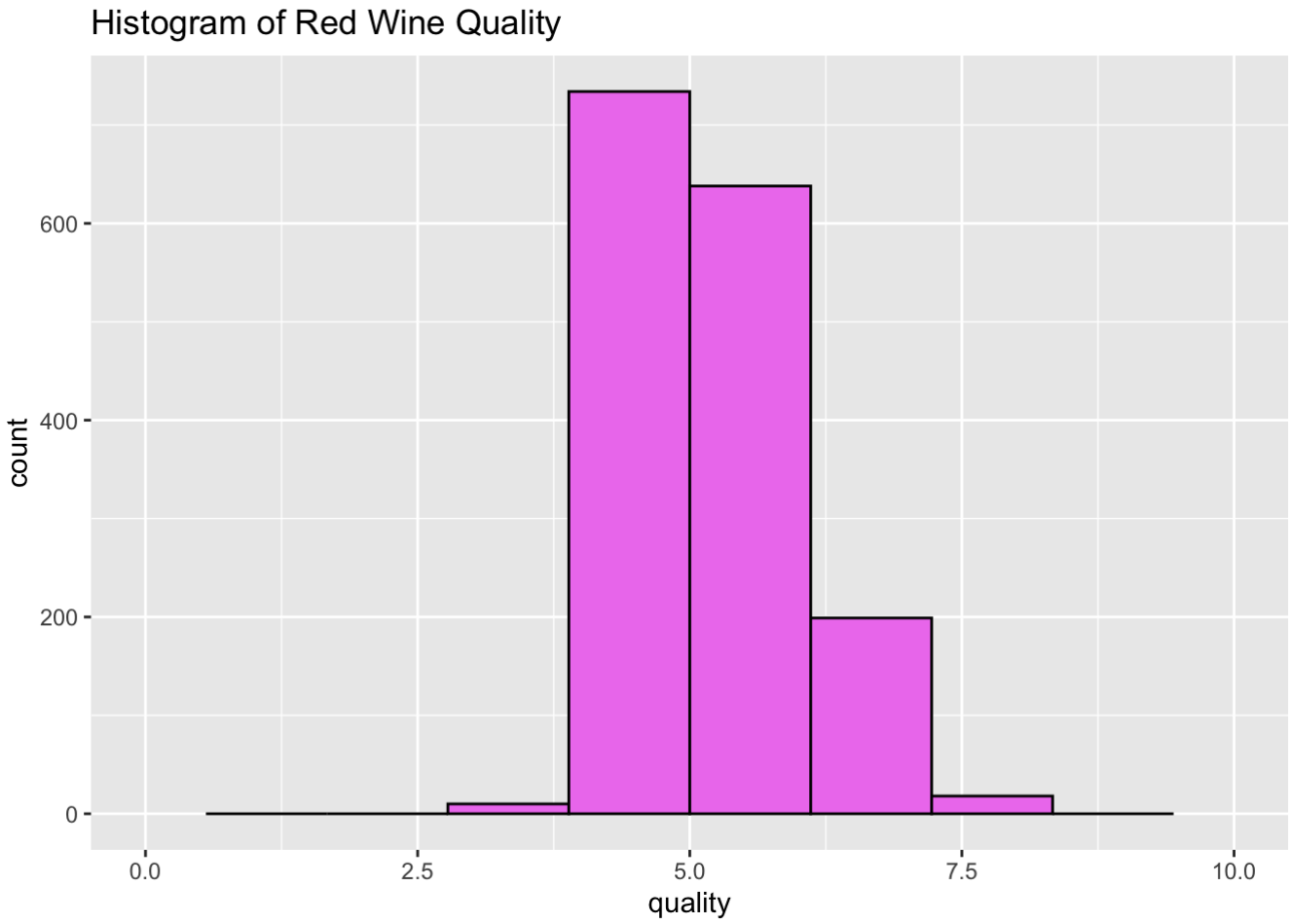
Diagnostics

```
library("broom")
diagnostics <- augment(step)
```

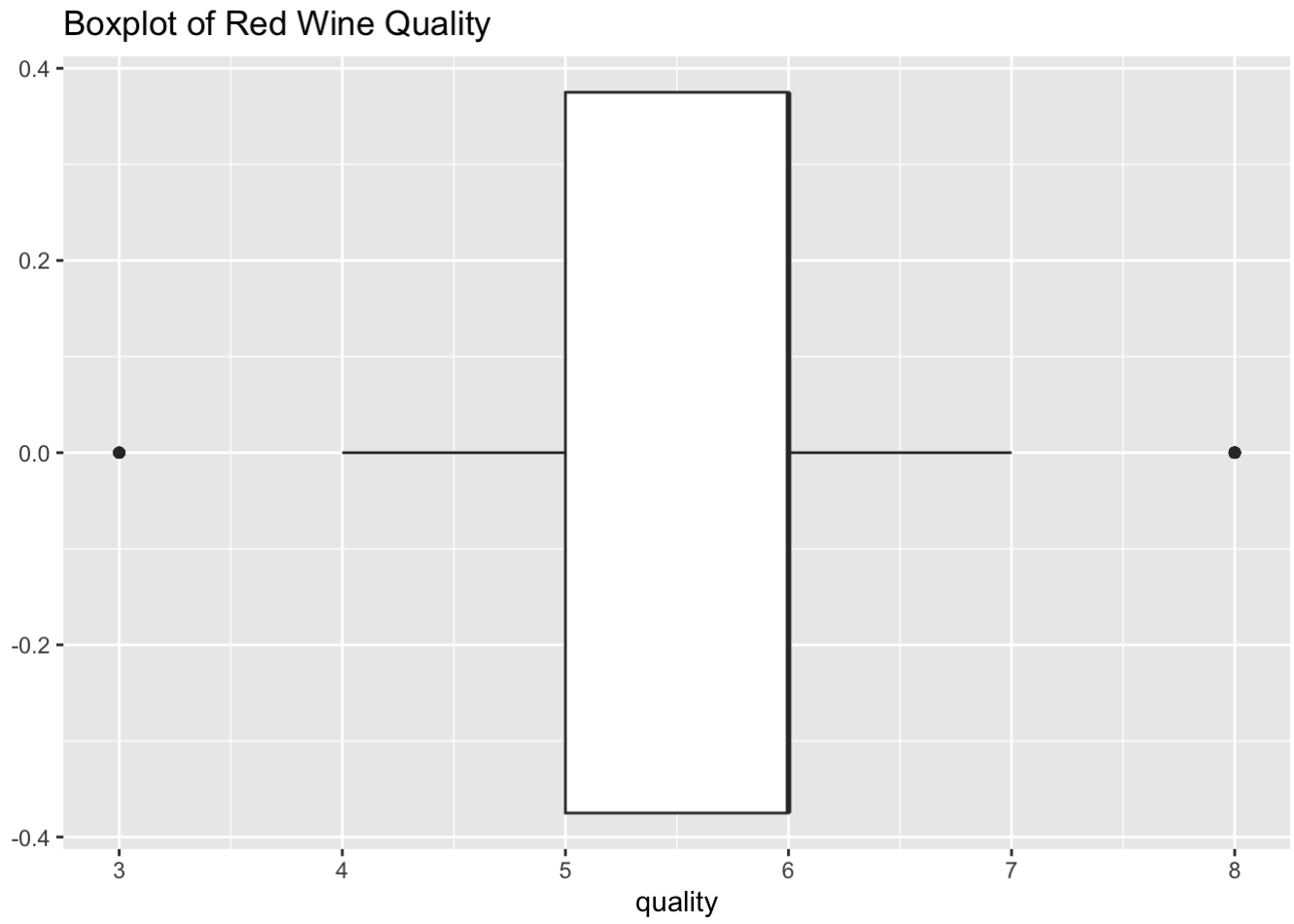
Distribution of quality

```
library("ggplot2")
ggplot(red, aes(x = quality)) + geom_histogram(bins = 10, color = "black", fill = "viole
t") + ggtitle("Histogram of Red Wine Quality") + scale_x_continuous(limits = c(0,10))
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```




```
ggplot(red, aes(x = quality)) + geom_boxplot() + ggtitle("Boxplot of Red Wine Quality")
```




```
xbar <- mean(red$quality)
xbar
```

```
## [1] 5.636023
```

```
sd <- sd(red$quality)
sd
```

```
## [1] 0.8075694
```

```
n = 1599
standard_error_mean <- sd/sqrt(n)
standard_error_mean
```

```
## [1] 0.02019555
```

```
margin <- qt(0.975,df=n-1)*sd/sqrt(n)
lowerinterval <- xbar - margin
lowerinterval
```

```
## [1] 5.59641
```

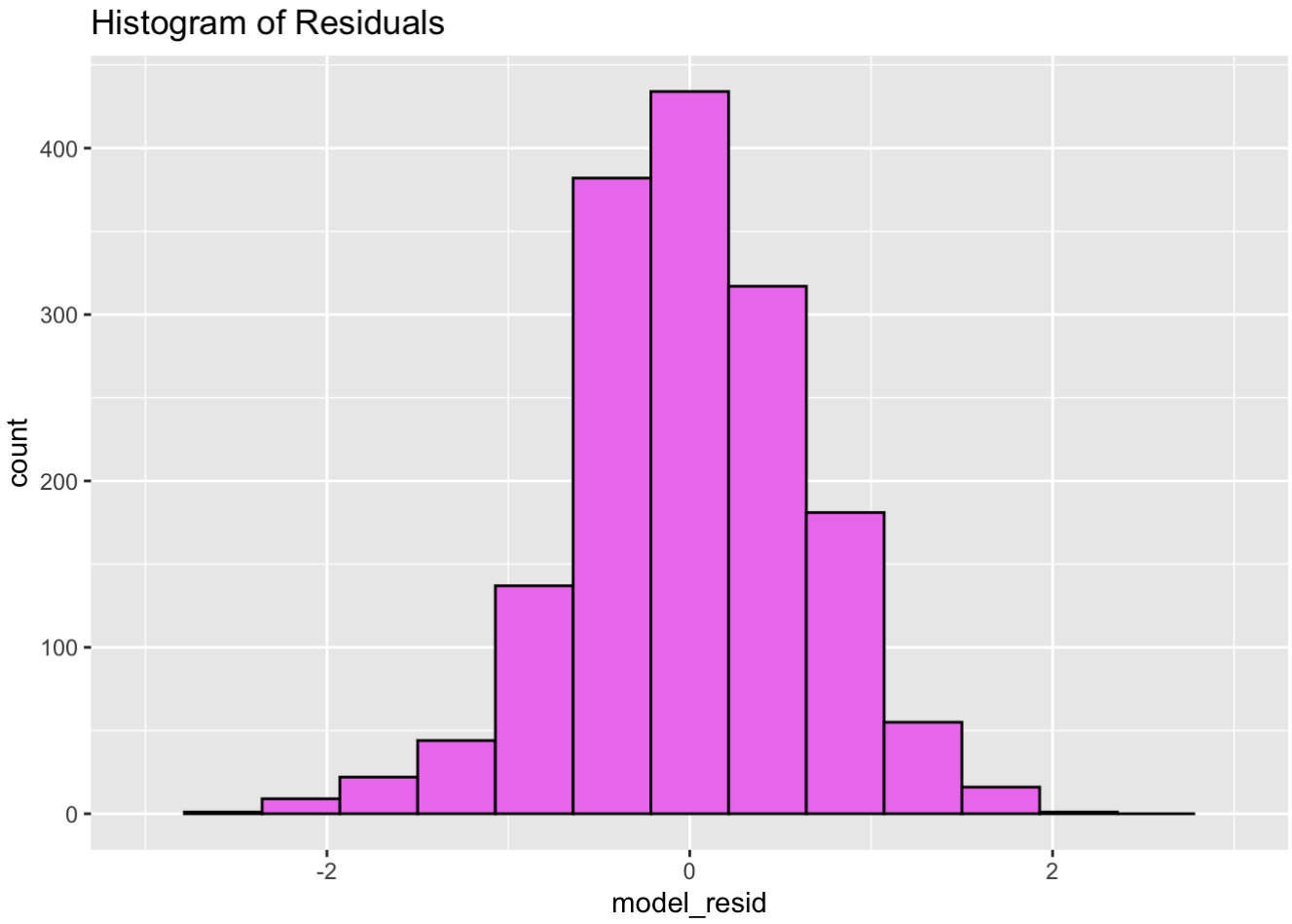
```
upperinterval <- xbar + margin
upperinterval
```

```
## [1] 5.675635
```

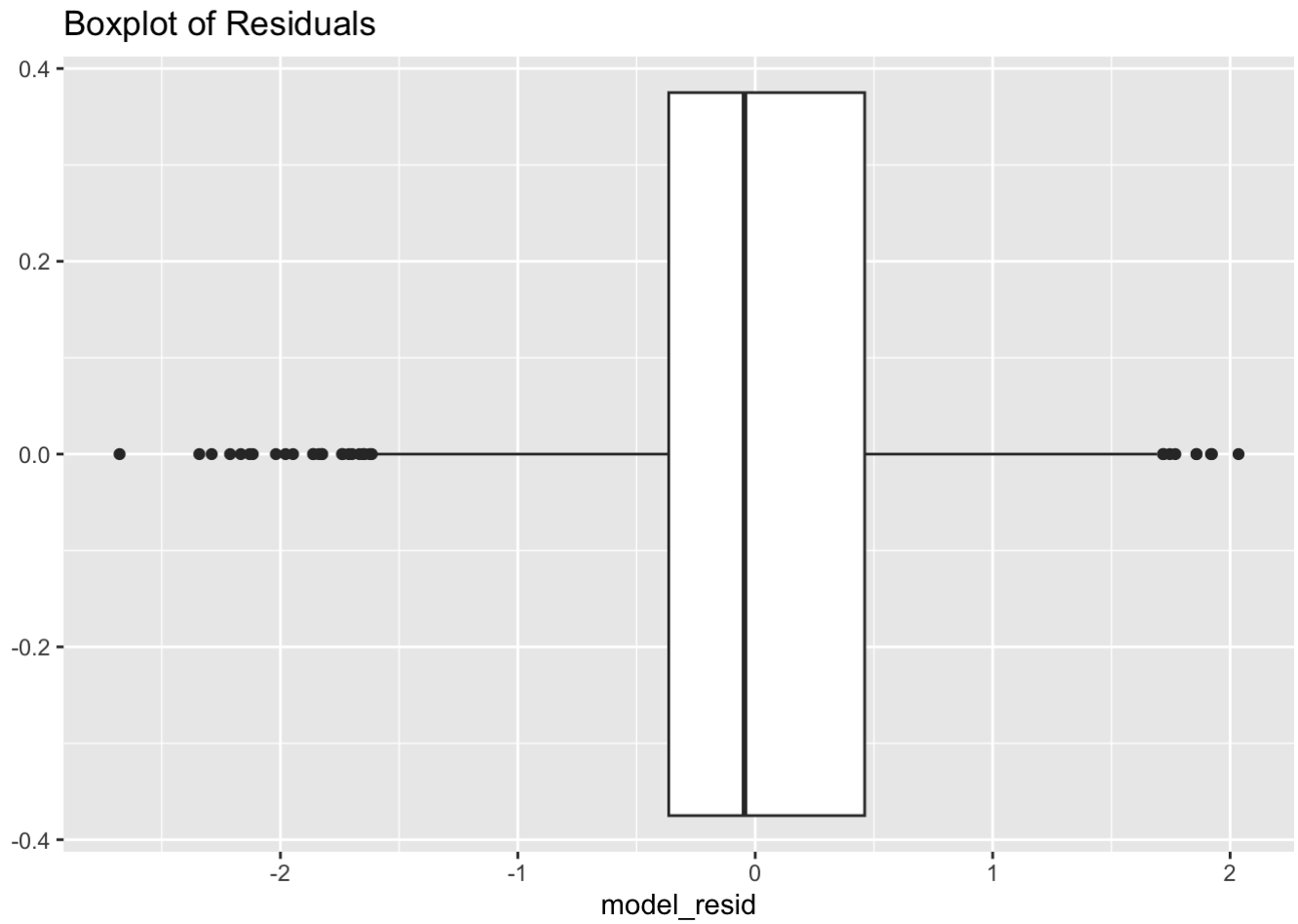
Plot Residuals

```
library("ggplot2")
model_resid = step$residuals
ggplot(red, aes(x = model_resid)) + geom_histogram(bins = 15, color = "black", fill = "v
iolet") + ggtitle("Histogram of Residuals") + scale_x_continuous(limits = c(-3,3))
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

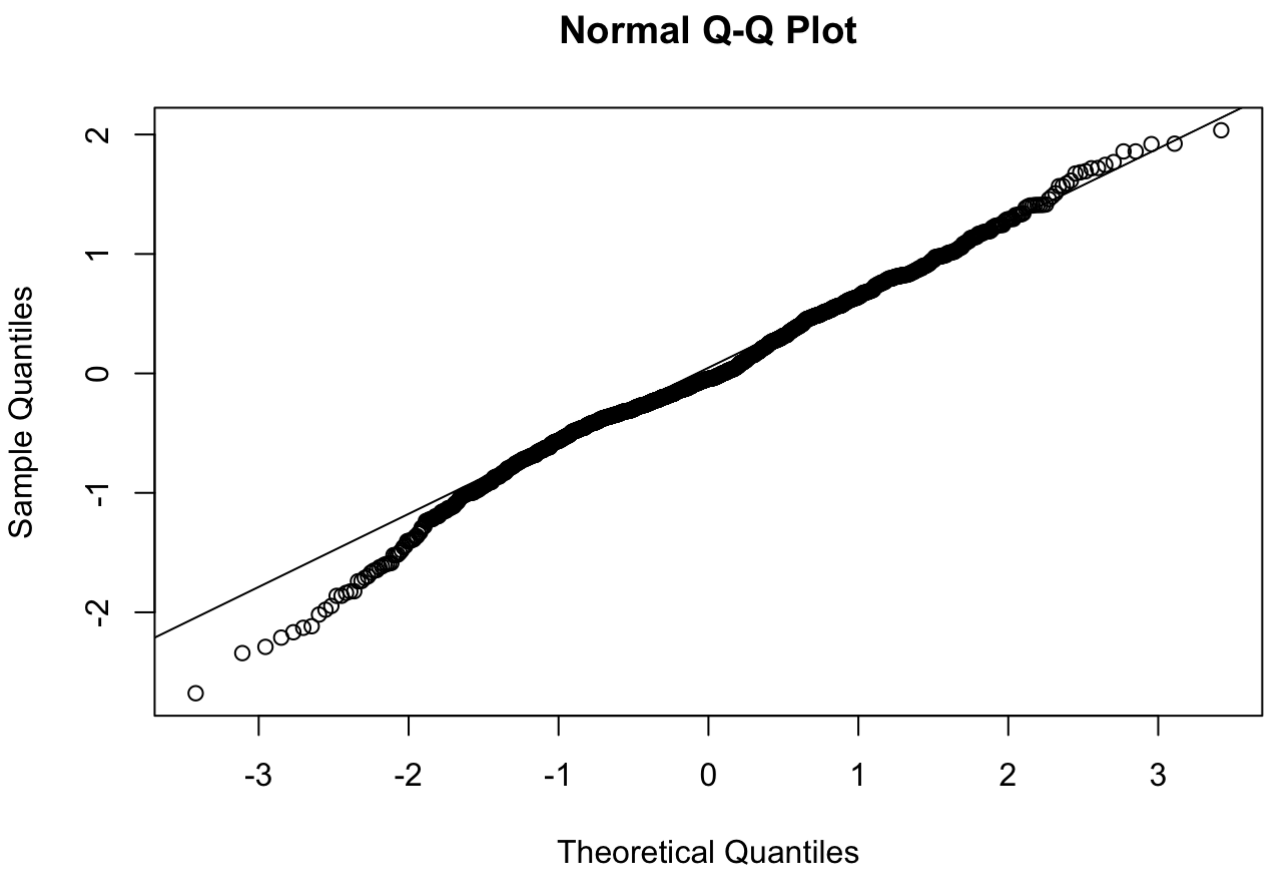



```
ggplot(red, aes(x = model_resid)) + geom_boxplot() + ggtitle("Boxplot of Residuals")
```



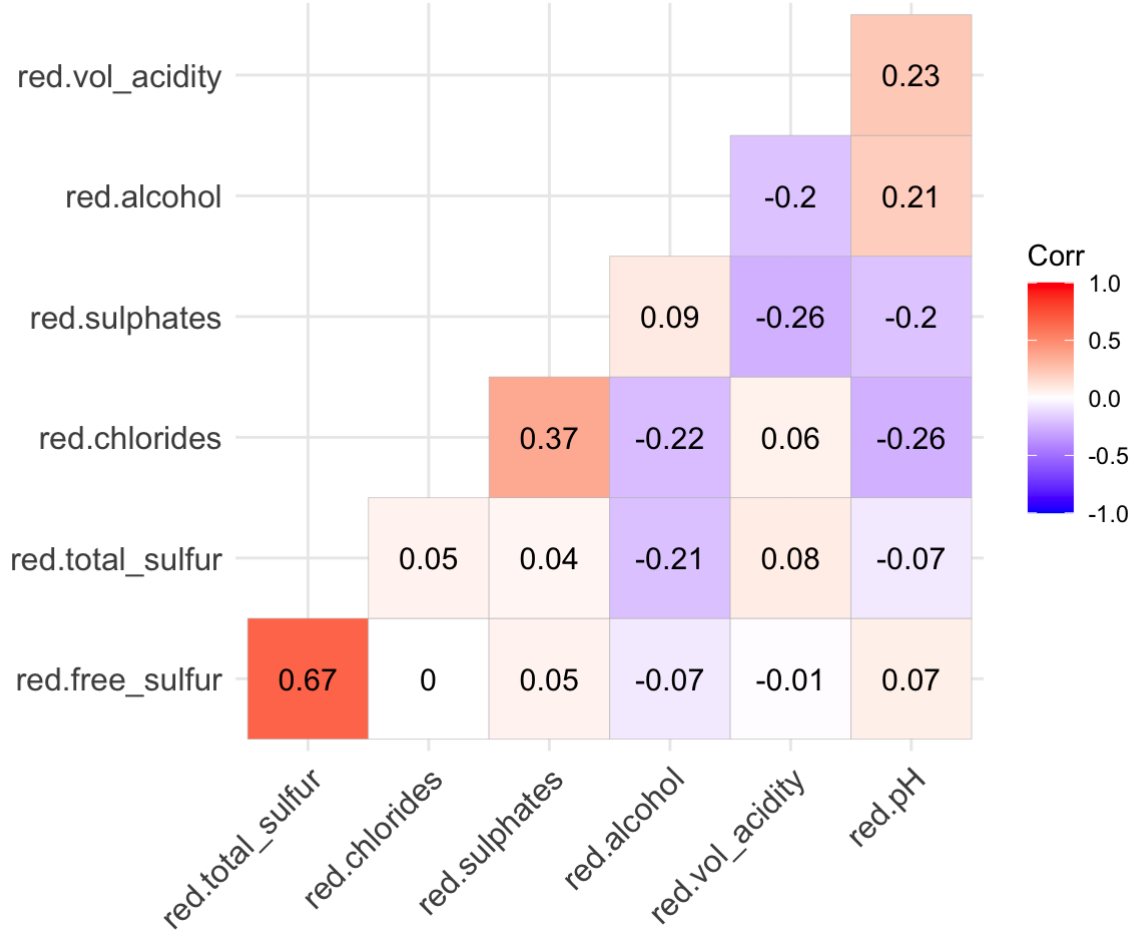
QQ Plot of Residuals - check normality

```
qqnorm(model_resid)
qqline(model_resid)
```



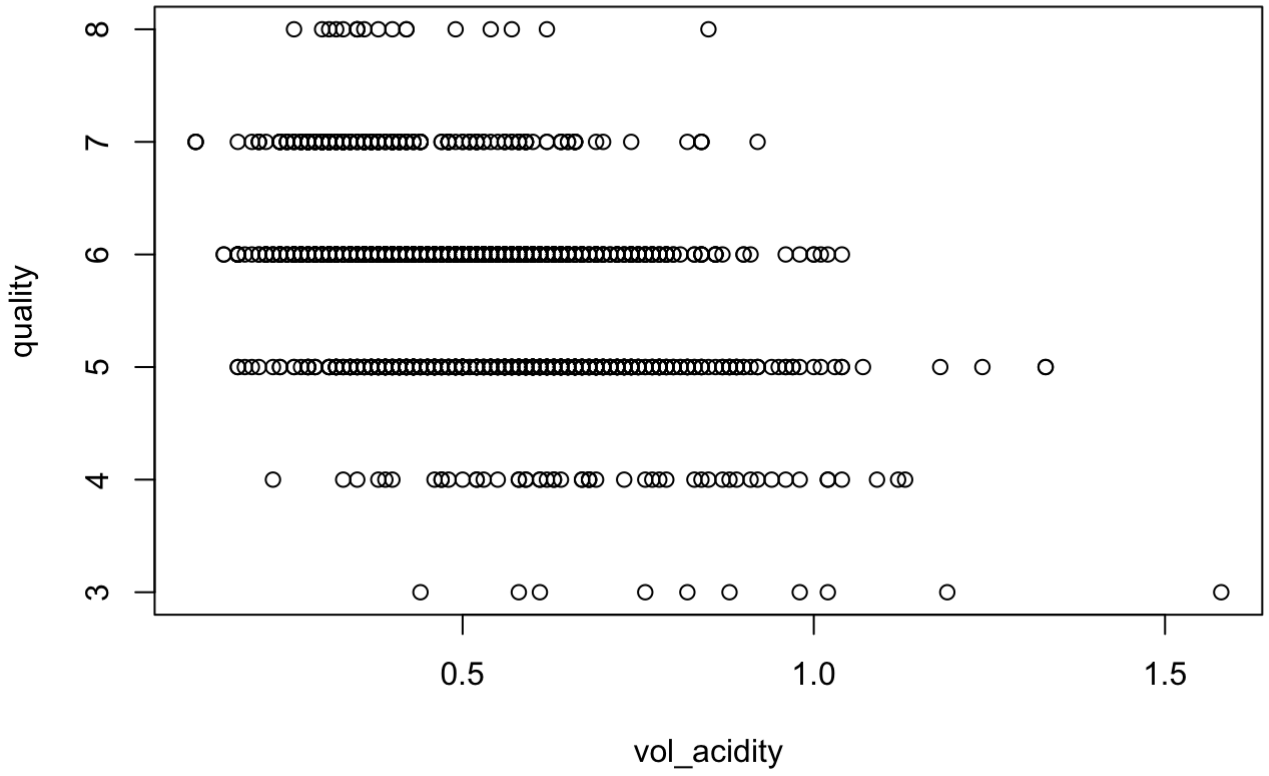
Multicollinearity Plot - check multicollinearity

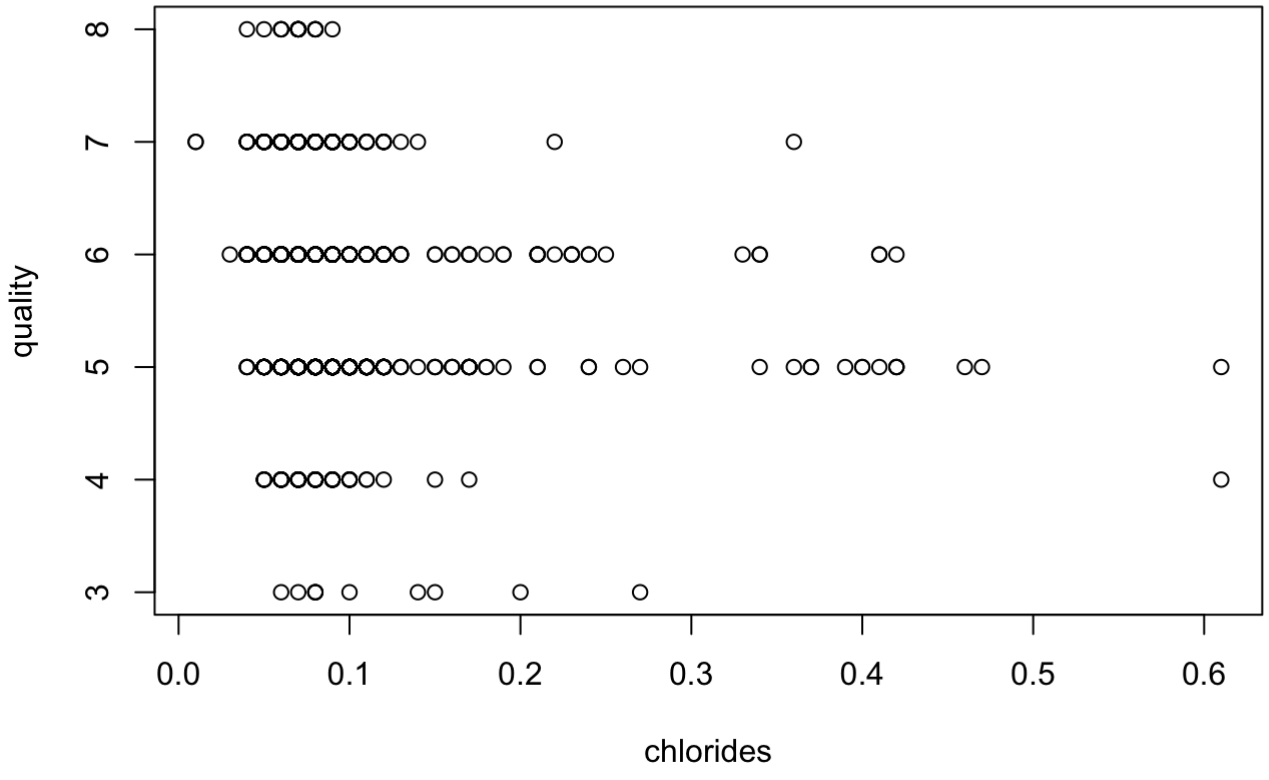
```
library("ggcorrplot")
red1 <- data.frame(red$vol_acidity, red$chlorides, red$free_sulfur, red$total_sulfur, re
d$pH, red$sulphates, red$alcohol)
corr_matrix = round(cor(red1), 2)
ggcorrplot(corr_matrix, hc.order = TRUE, type = "lower", lab = TRUE)
```

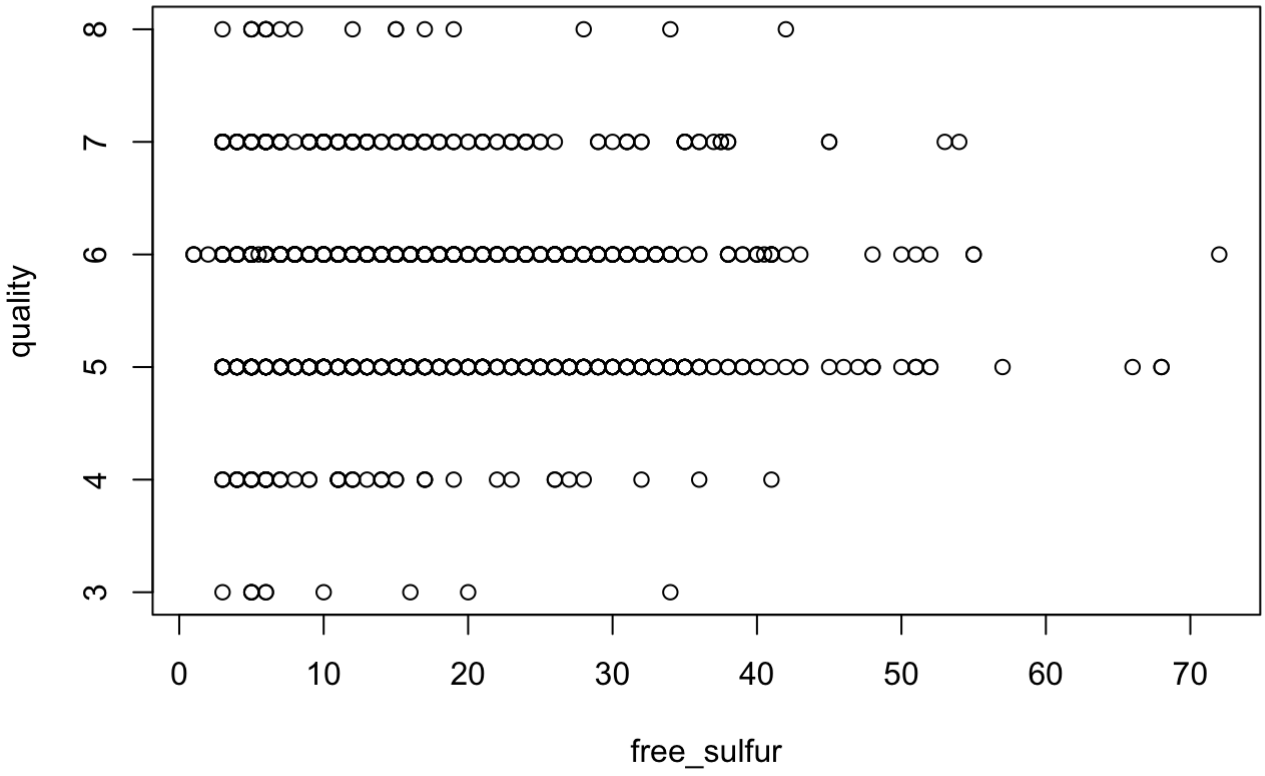


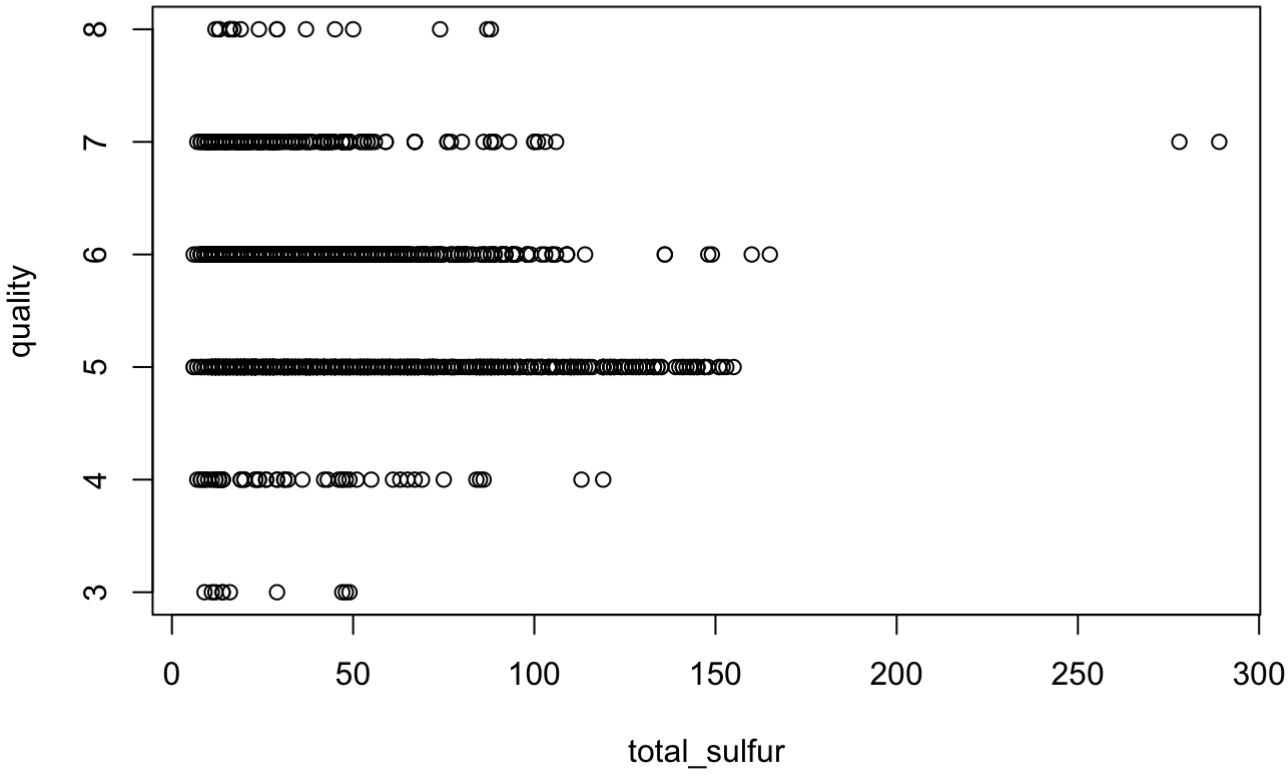
Other Plots

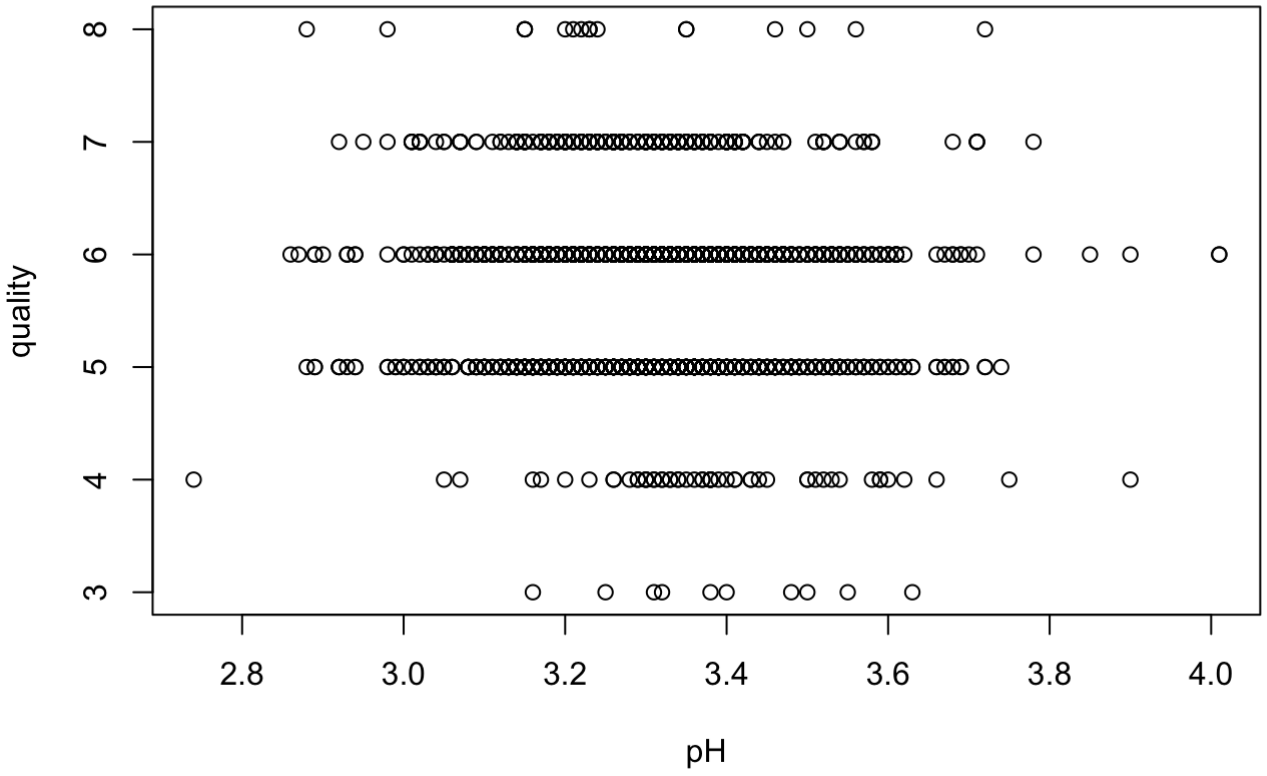
```
plot(quality ~ vol_acidity + chlorides + free_sulfur + total_sulfur + pH + sulphates + a
lcohol, data = red)
```

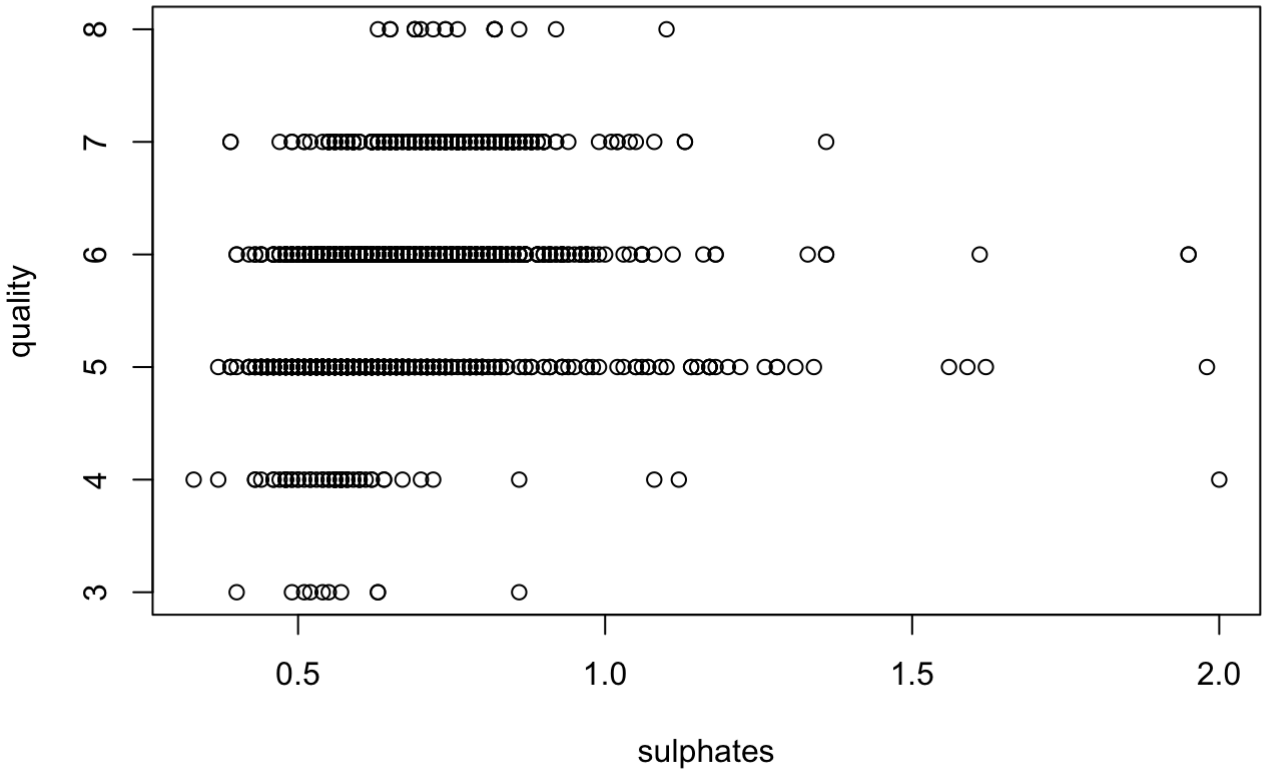

quality
3
4
5
6
7
8
0.5
1.0
1.5
vol_acidity

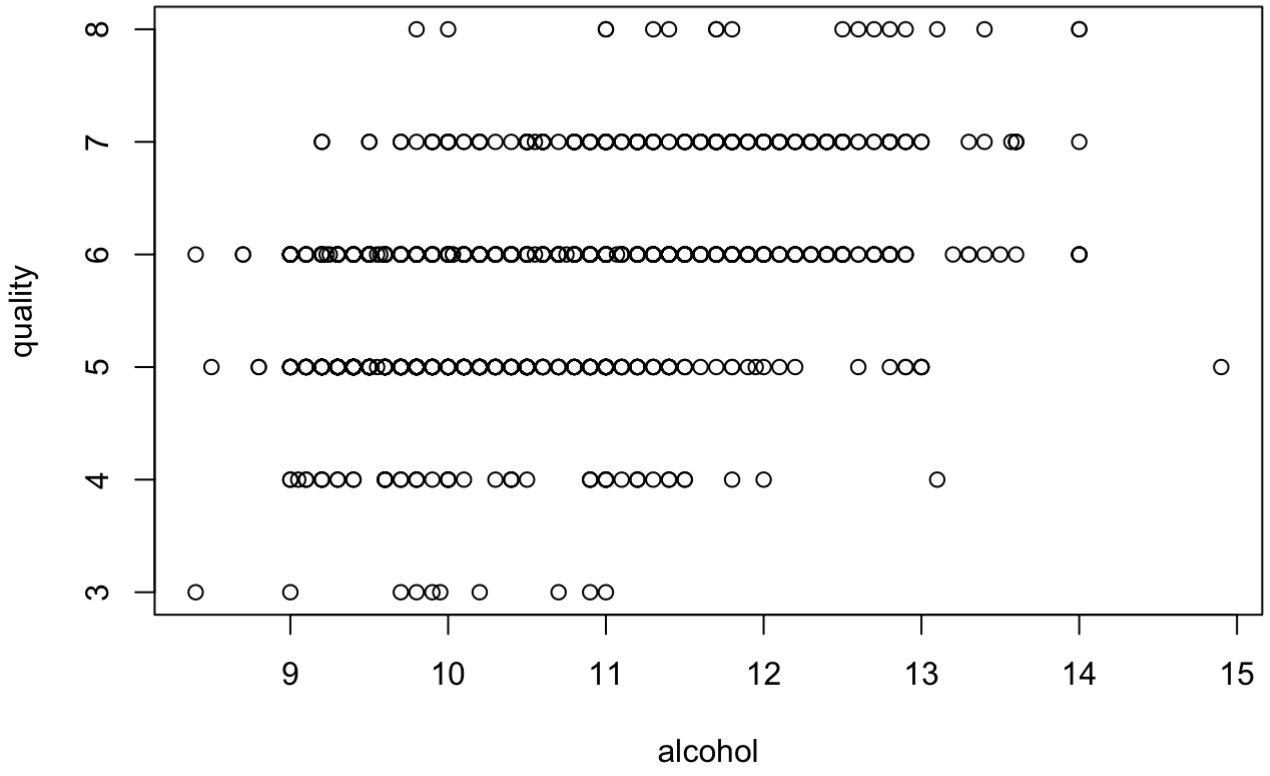

quality
3
4
5
6
7
8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides


quality
3
4
5
6
7
8
0
10
20
30
40
50
60
70
free_sulfur

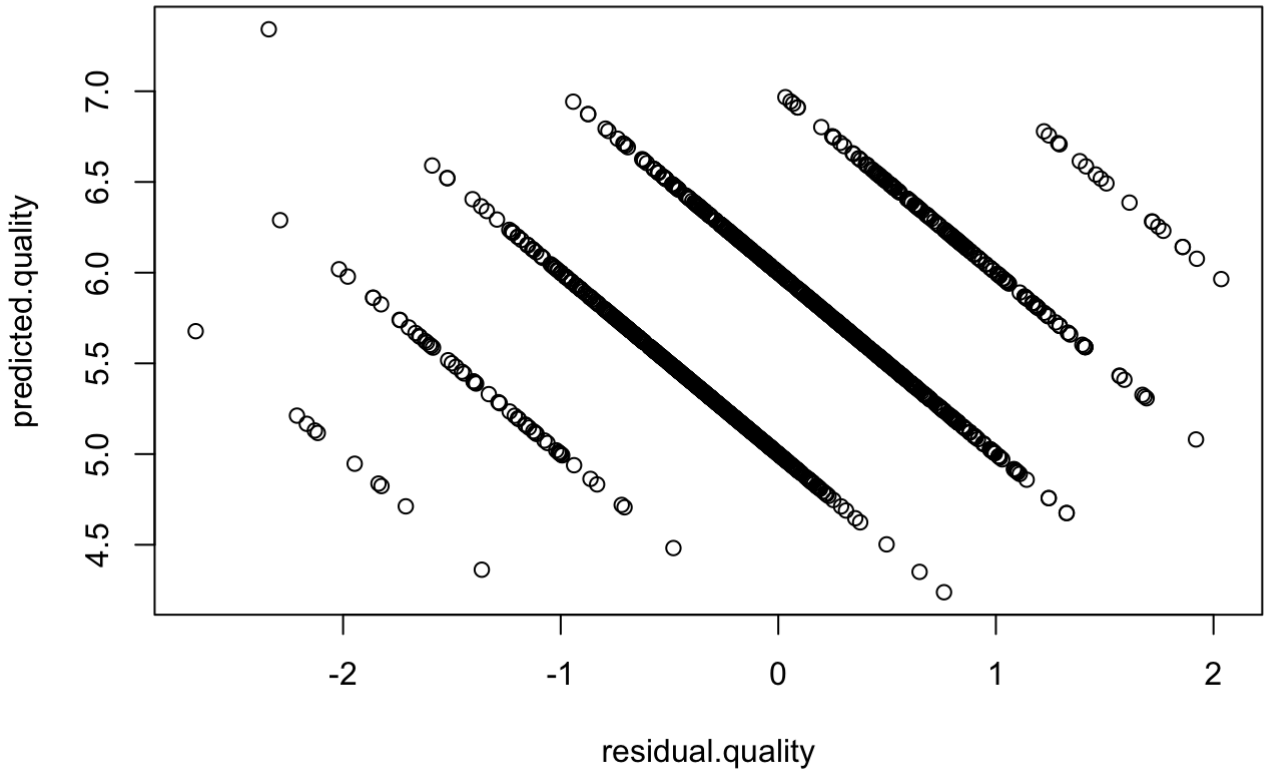

quality
3
4
5
6
7
8
0
50
100
150
200
250
300
total_sulfur


quality
3
4
5
6
7
8
2.8
3.0
3.2
3.4
3.6
3.8
4.0
pH

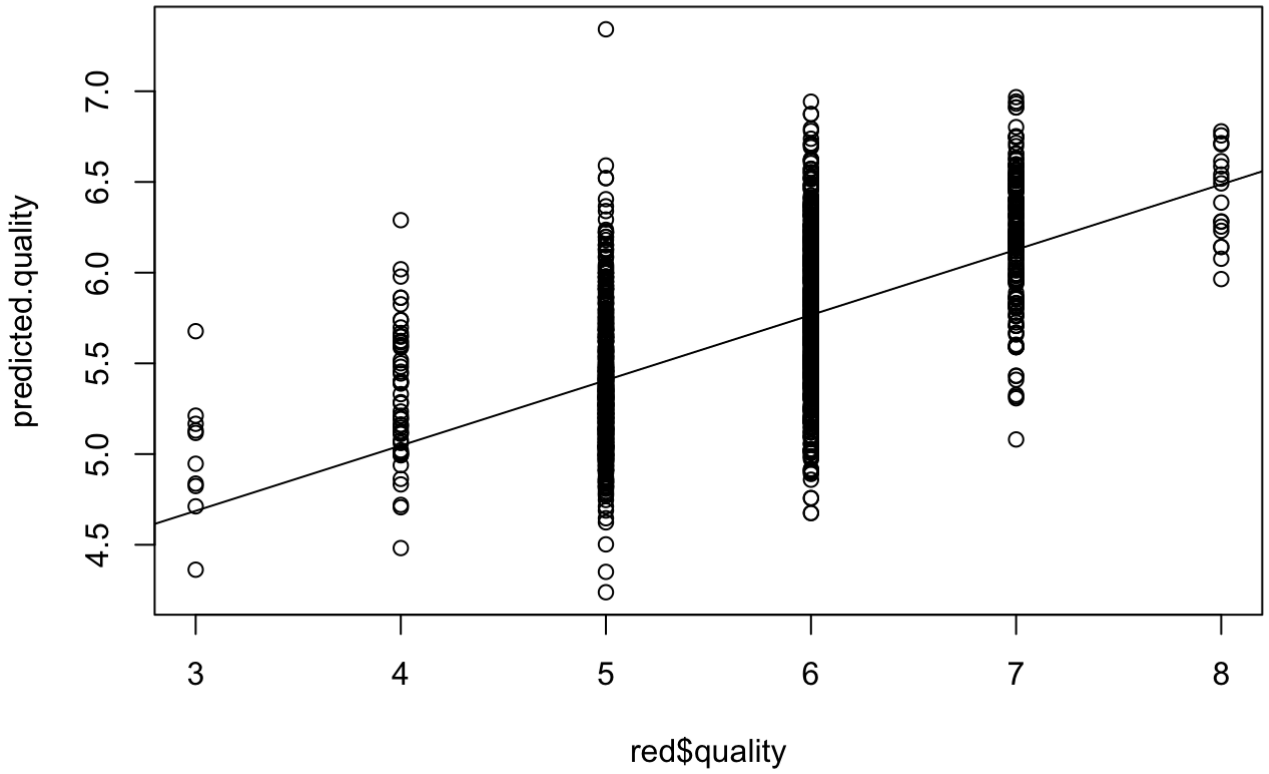

quality
3
4
5
6
7
8
0.5
1.0
1.5
2.0
sulphates



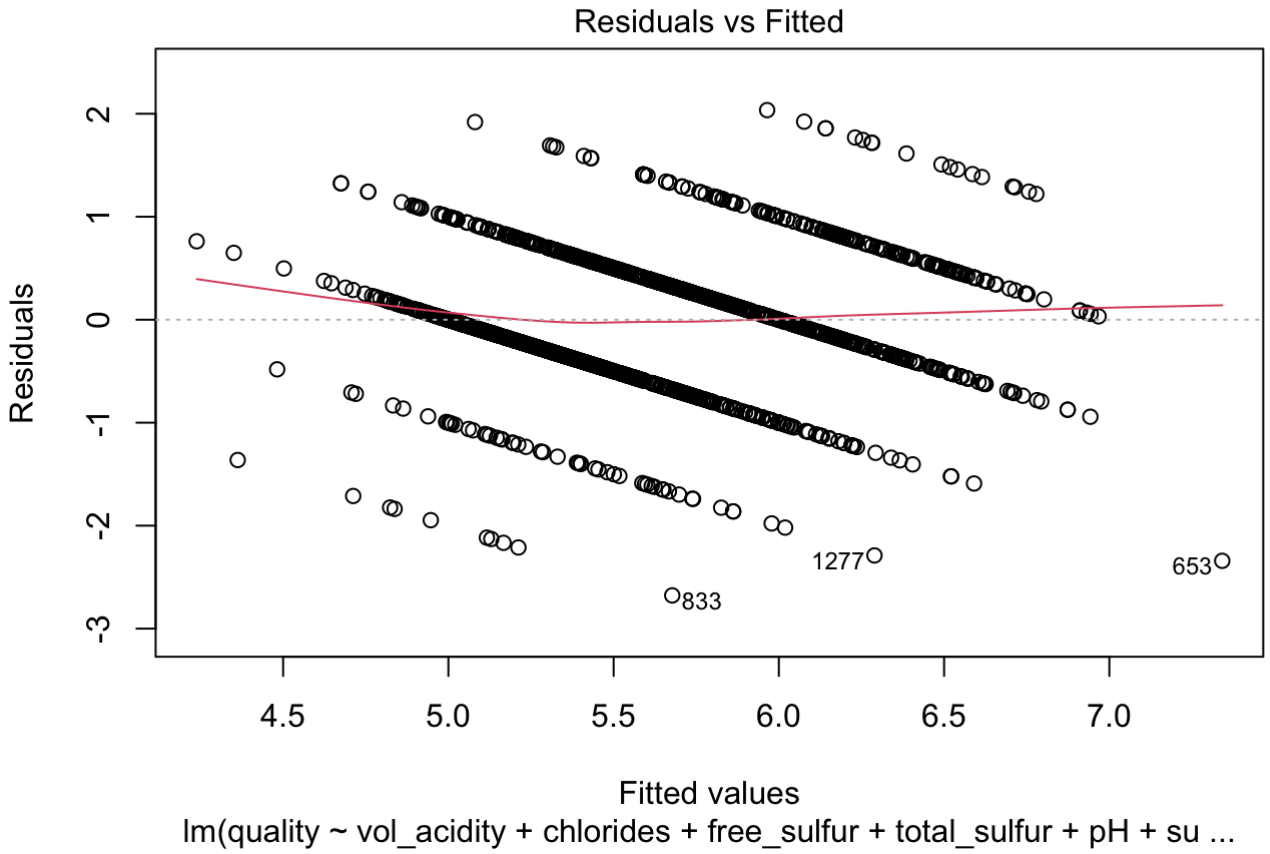

```
residual.quality = residuals(step)
predicted.quality = predict(step)
plot(residual.quality, predicted.quality)
```

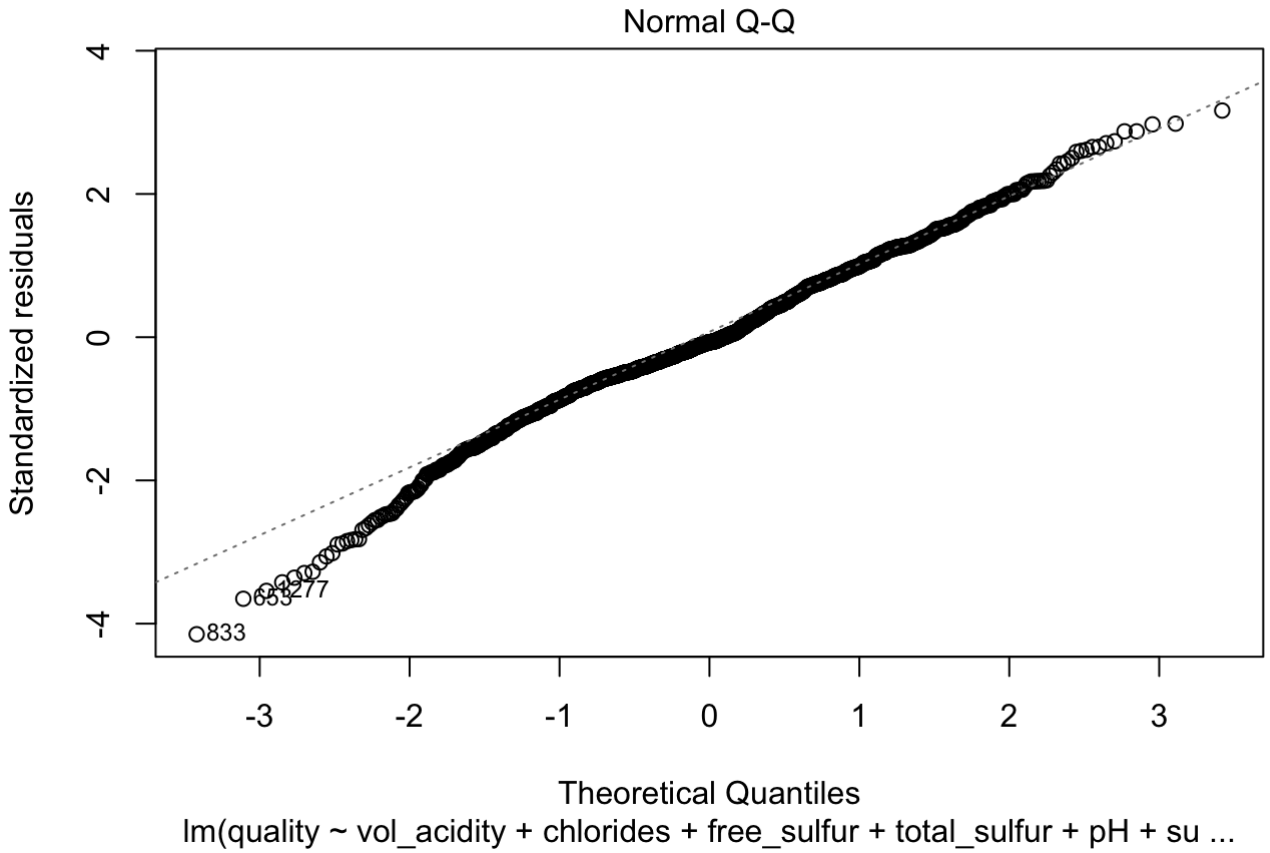


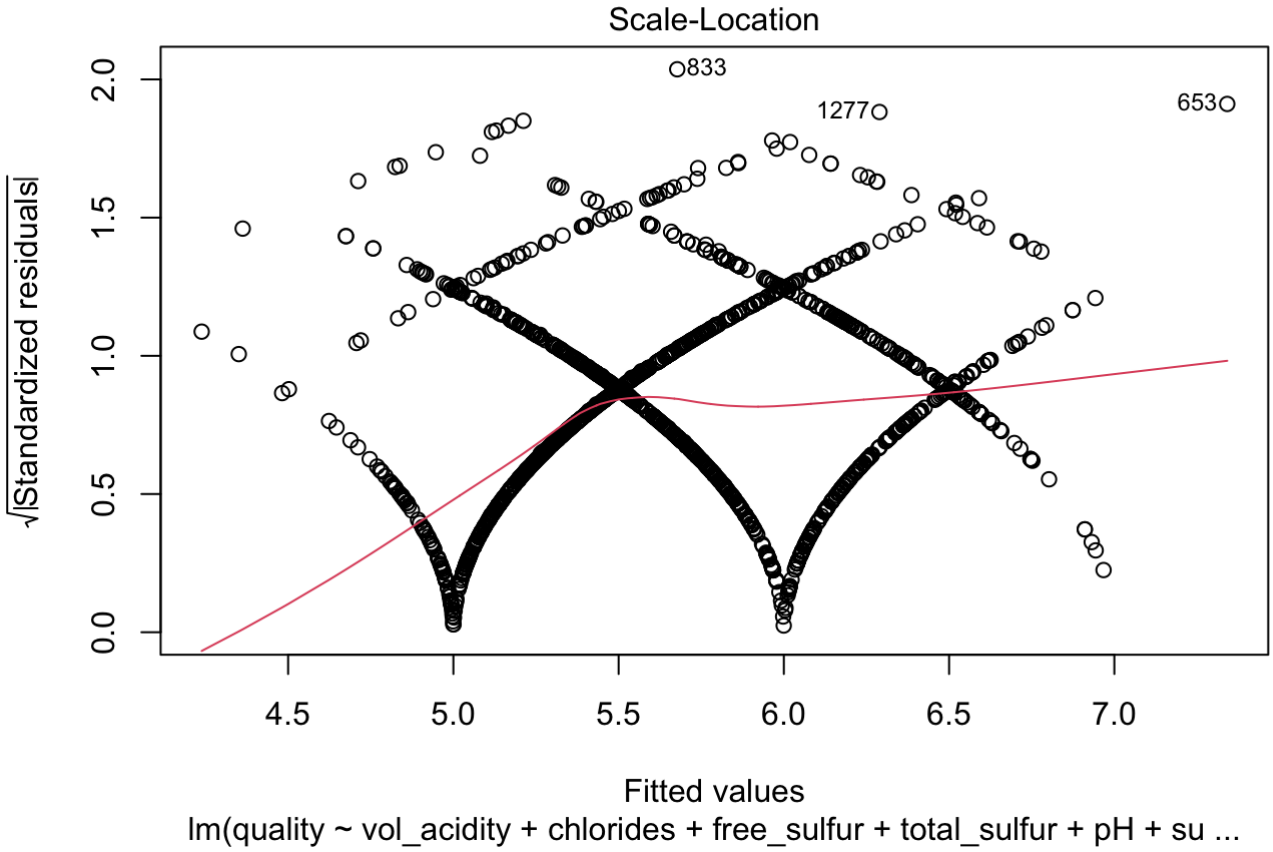

```
plot(red$quality, predicted.quality)
reg = lm(predicted.quality ~ red$quality)
abline(reg)
```

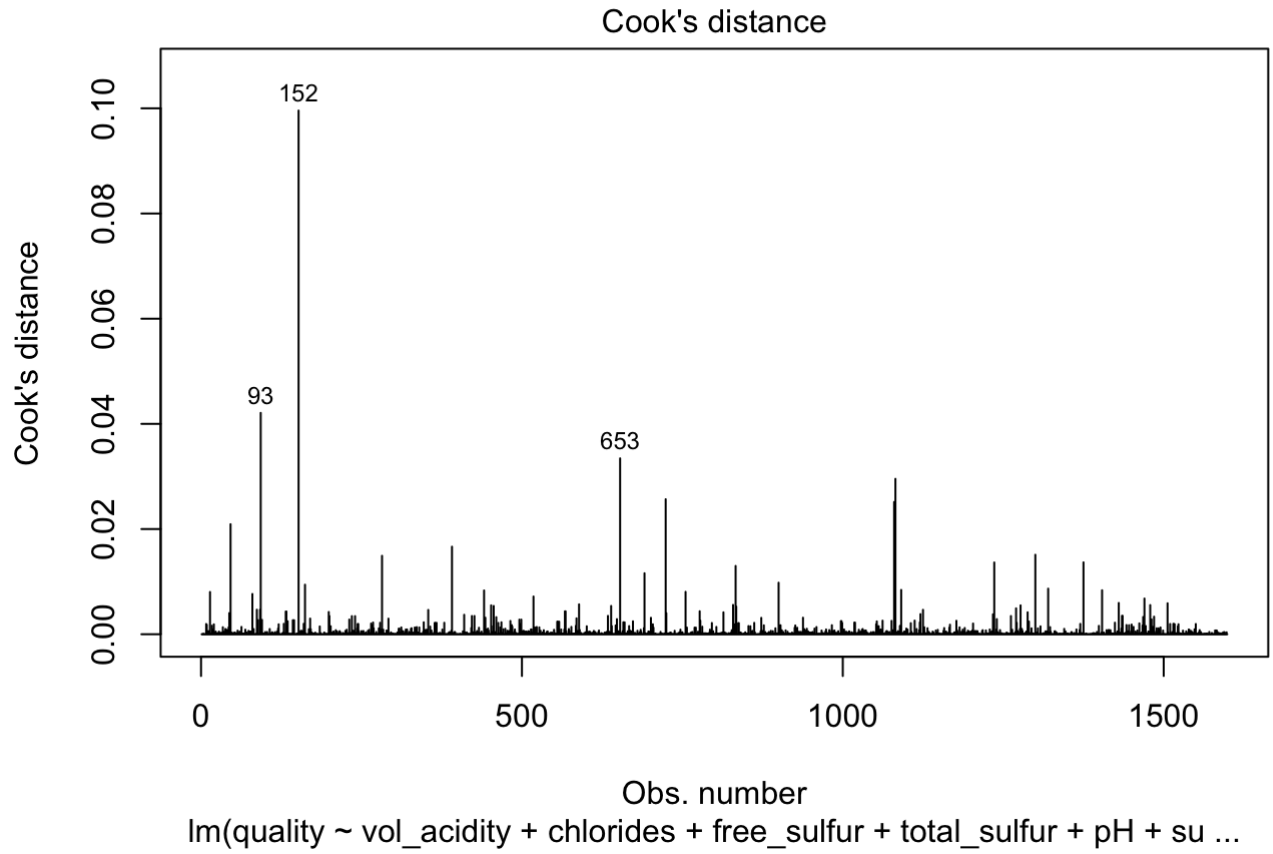


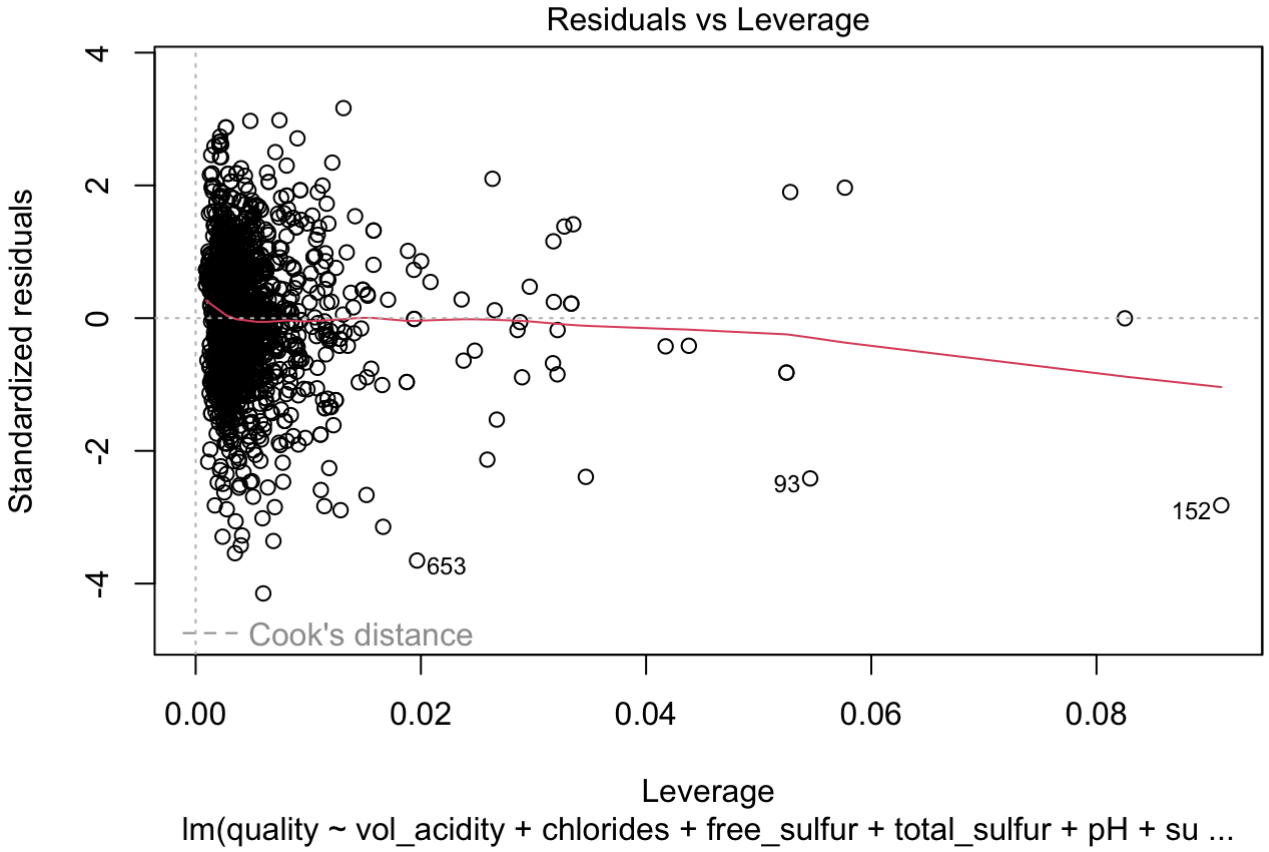

```
plot(step, which = c(1,2,3,4,5,6))
```

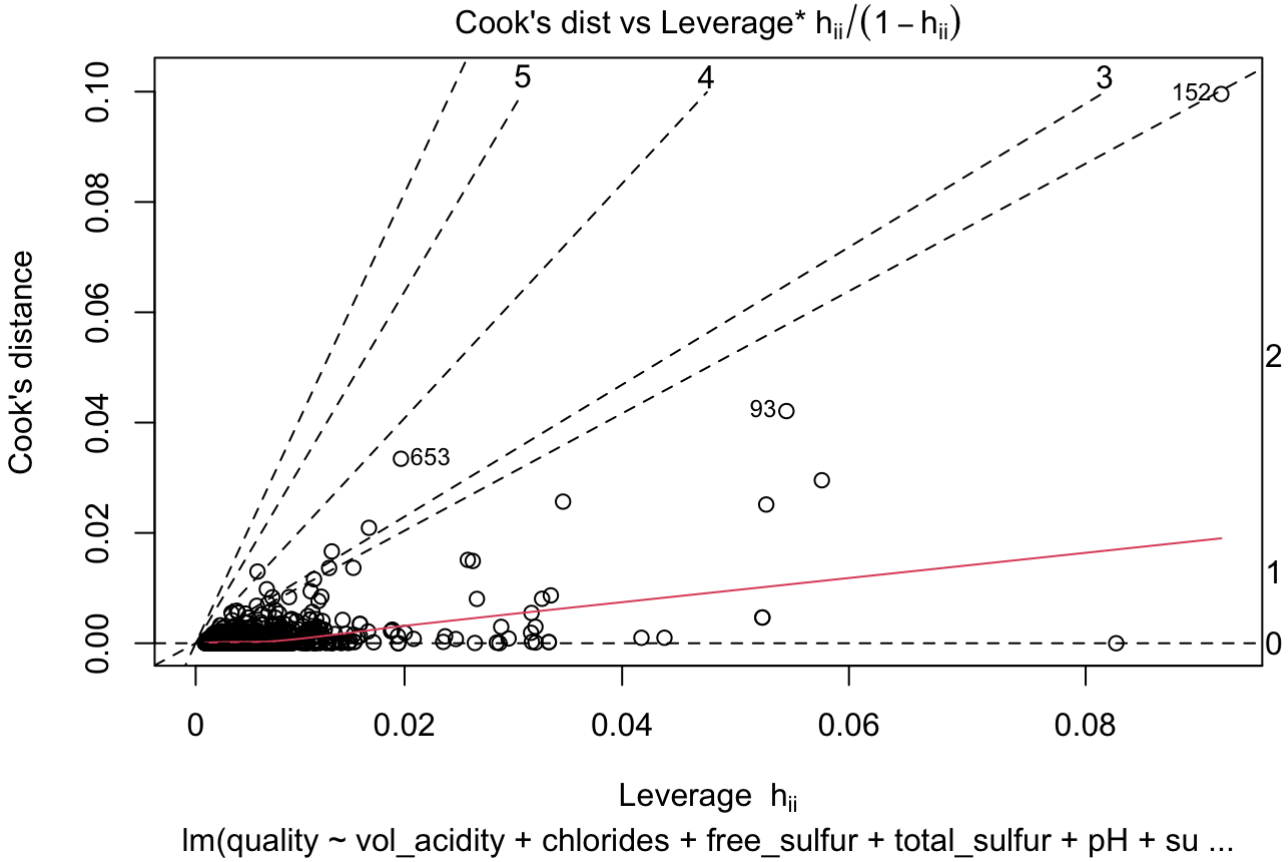

Residuals vs Fitted
Residuals
Fitted values
lm(quality ~ vol_acidity + chlorides + free_sulfur + total_sulfur + pH + su ...
1277
653
833


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(quality ~ vol_acidity + chlorides + free_sulfur + total_sulfur + pH + su ...
1277
653
833


Scale-Location
833
1277
653
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
Fitted values
lm(quality ~ vol_acidity + chlorides + free_sulfur + total_sulfur + pH + su ...

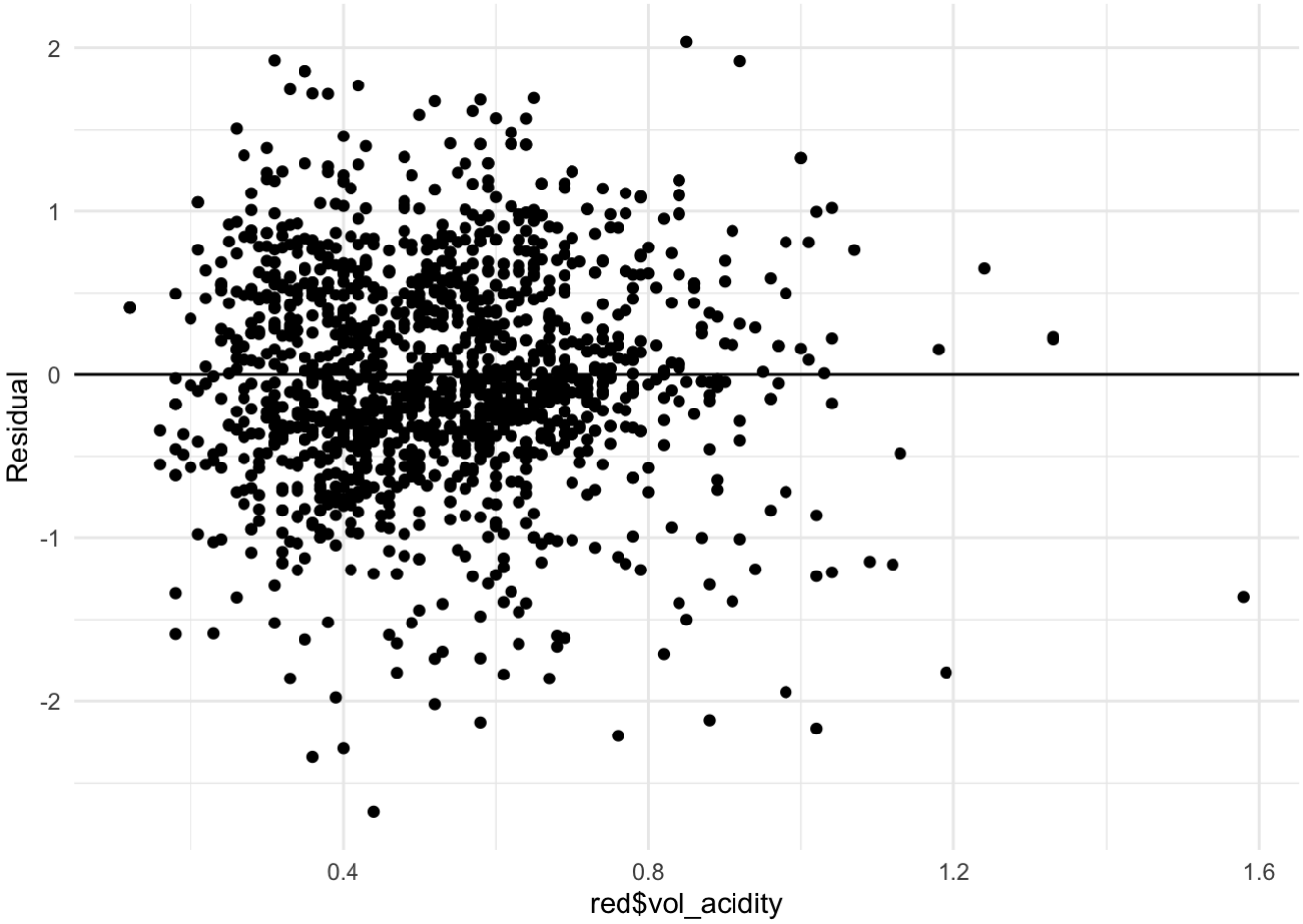

Cook's distance
152
93
653
Cook's distance
0.00
0.02
0.04
0.06
0.08
0.10
0
500
1000
1500
Obs. number
lm(quality ~ vol_acidity + chlorides + free_sulfur + total_sulfur + pH + su ...


Residuals vs Leverage
Standardized residuals
4
2
0
-2
-4
0.00
0.02
0.04
0.06
0.08
Leverage
lm(quality ~ vol_acidity + chlorides + free_sulfur + total_sulfur + pH + su ...
Cook's distance
653
93
152


Cook's dist vs Leverage* $h_{ii}/(1-h_{ii})$
Cook's distance
0.10
0.08
0.06
0.04
0.02
0.00
0
0.02
0.04
0.06
0.08
Leverage h_{ii}
lm(quality ~ vol_acidity + chlorides + free_sulfur + total_sulfur + pH + su ...
5
4
3
2
1
0
653
93
152

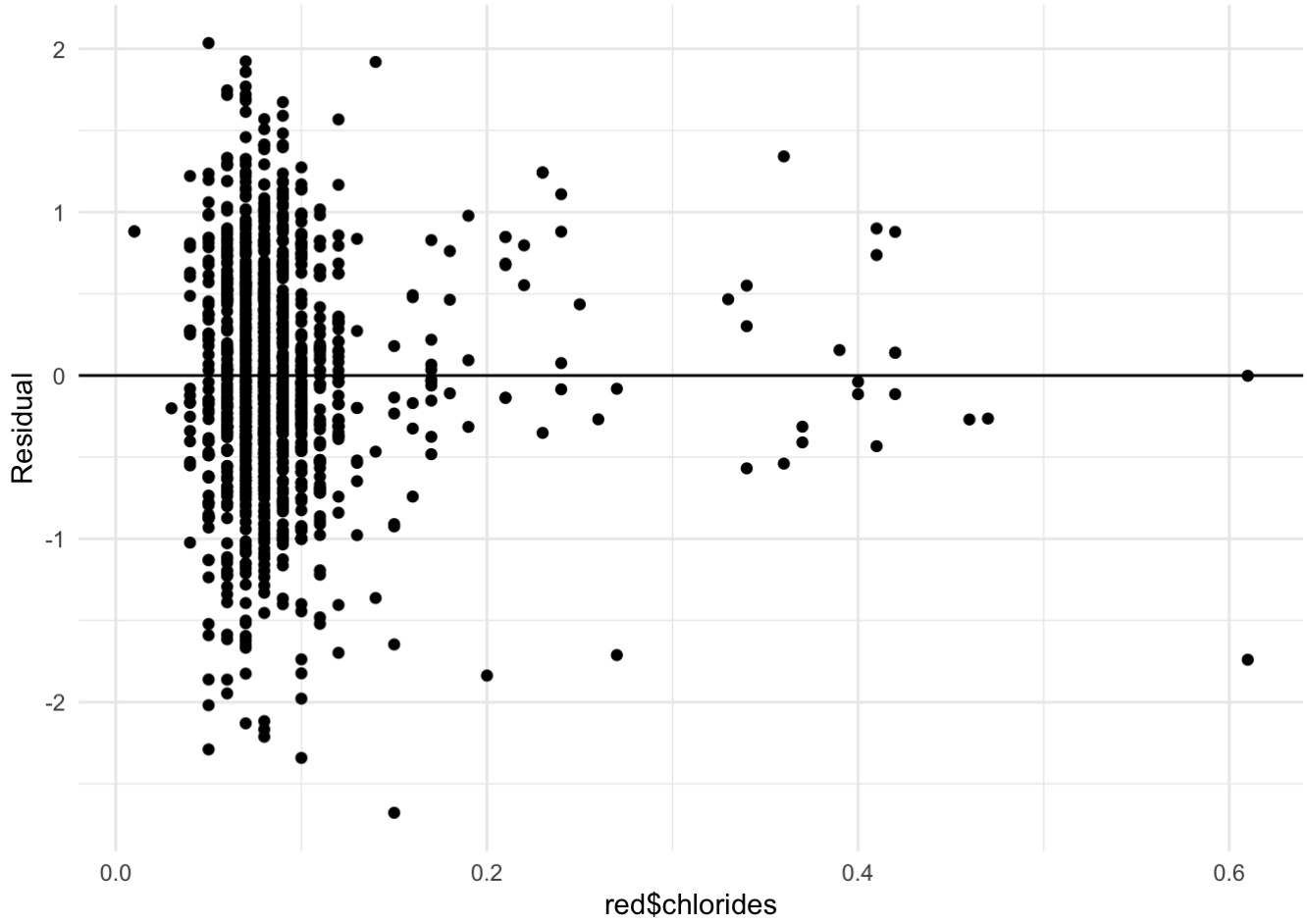
Residuals by vol_acidity

```
library("GGally")
ggplot(diagnostics) +
  geom_point(aes(x = red$vol_acidity, y = .resid)) +
  geom_hline(yintercept = 0) +
  ylab("Residual") +
  theme_minimal()
```



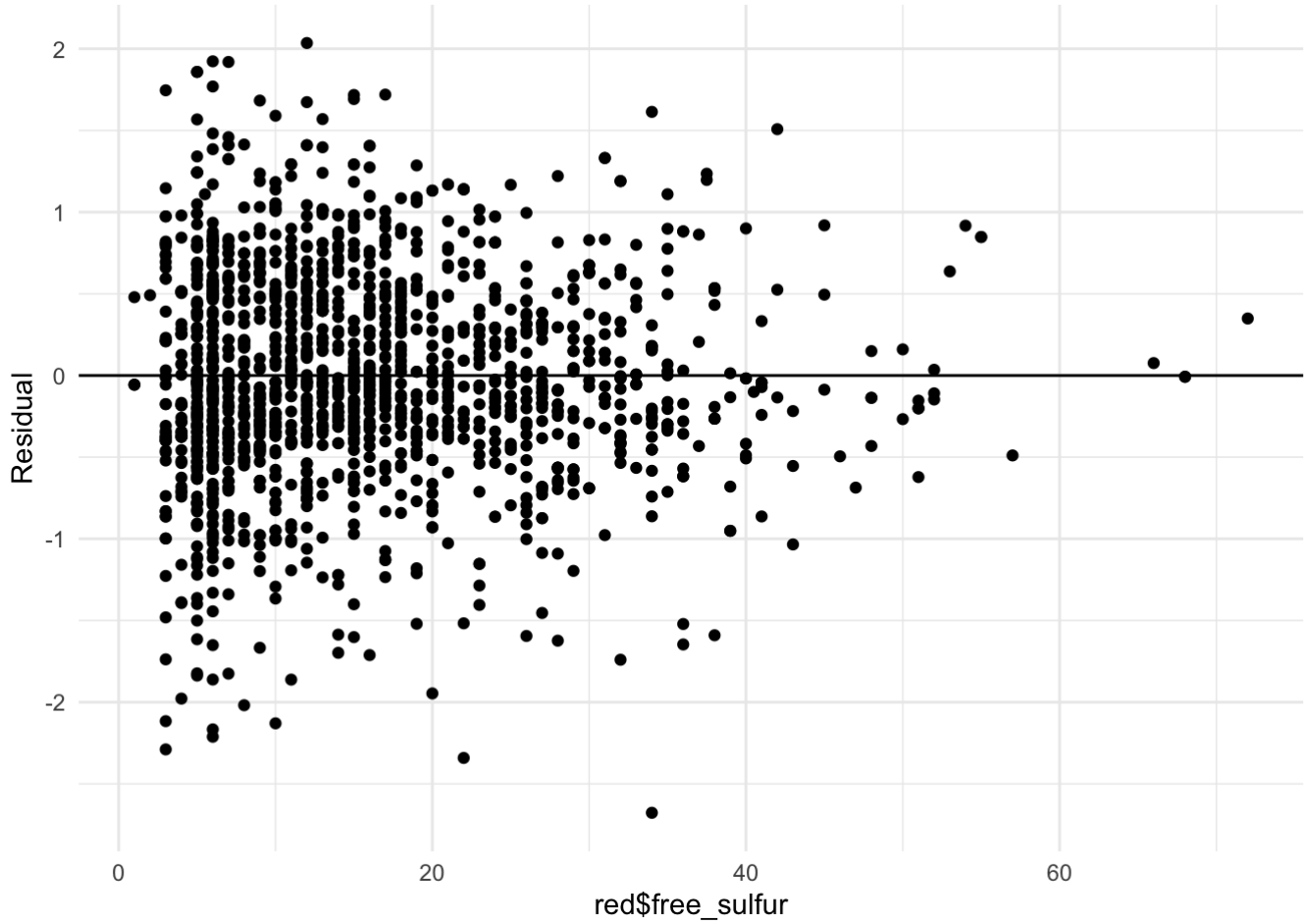
Residuals by chlorides

```
library("GGally")
ggplot(diagnostics) +
  geom_point(aes(x = red$chlorides, y = .resid)) +
  geom_hline(yintercept = 0) +
  ylab("Residual") +
  theme_minimal()
```



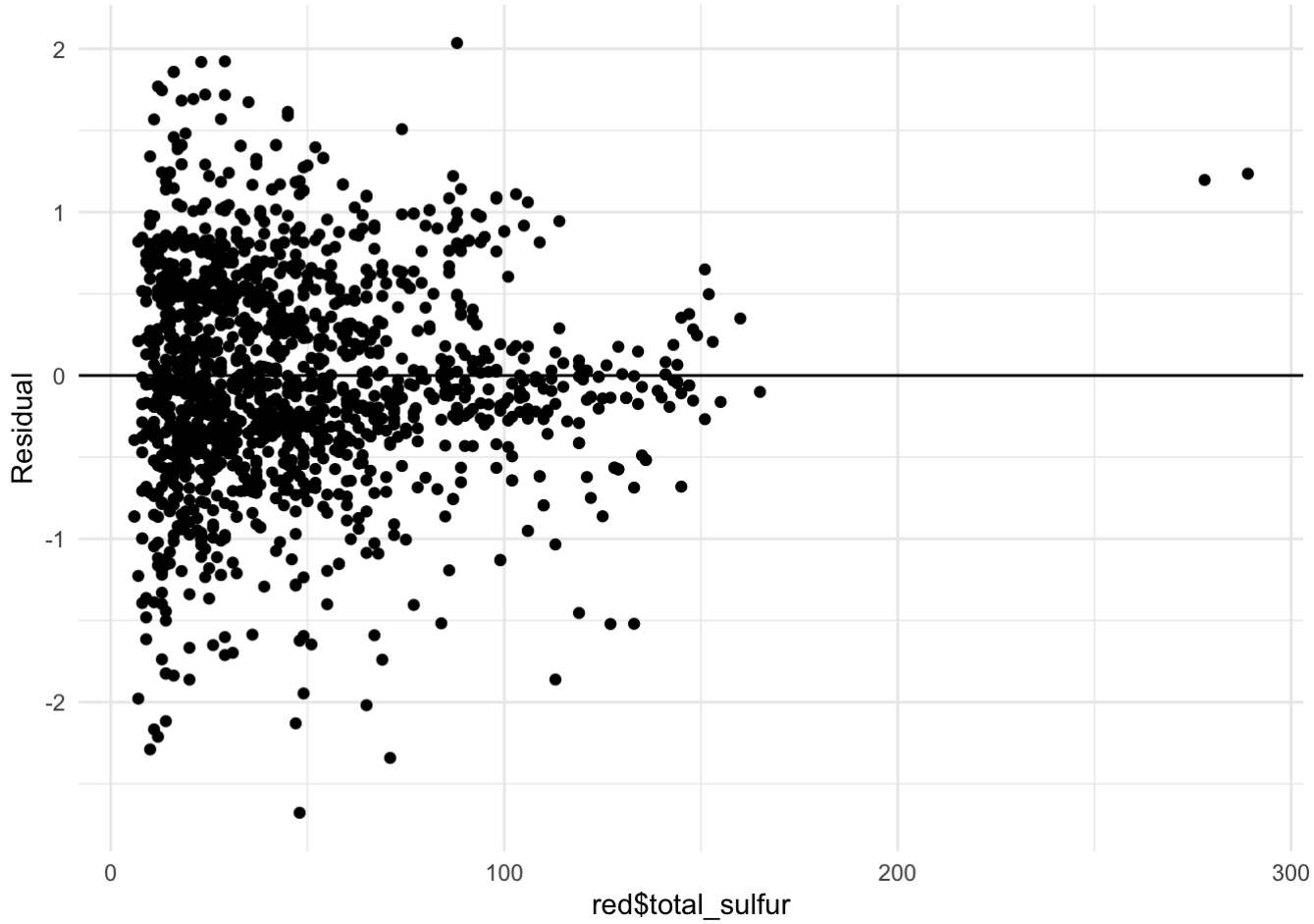
Residuals by free_sulfur

```
library("GGally")
ggplot(diagnostics) +
  geom_point(aes(x = red$free_sulfur, y = .resid)) +
  geom_hline(yintercept = 0) +
  ylab("Residual") +
  theme_minimal()
```



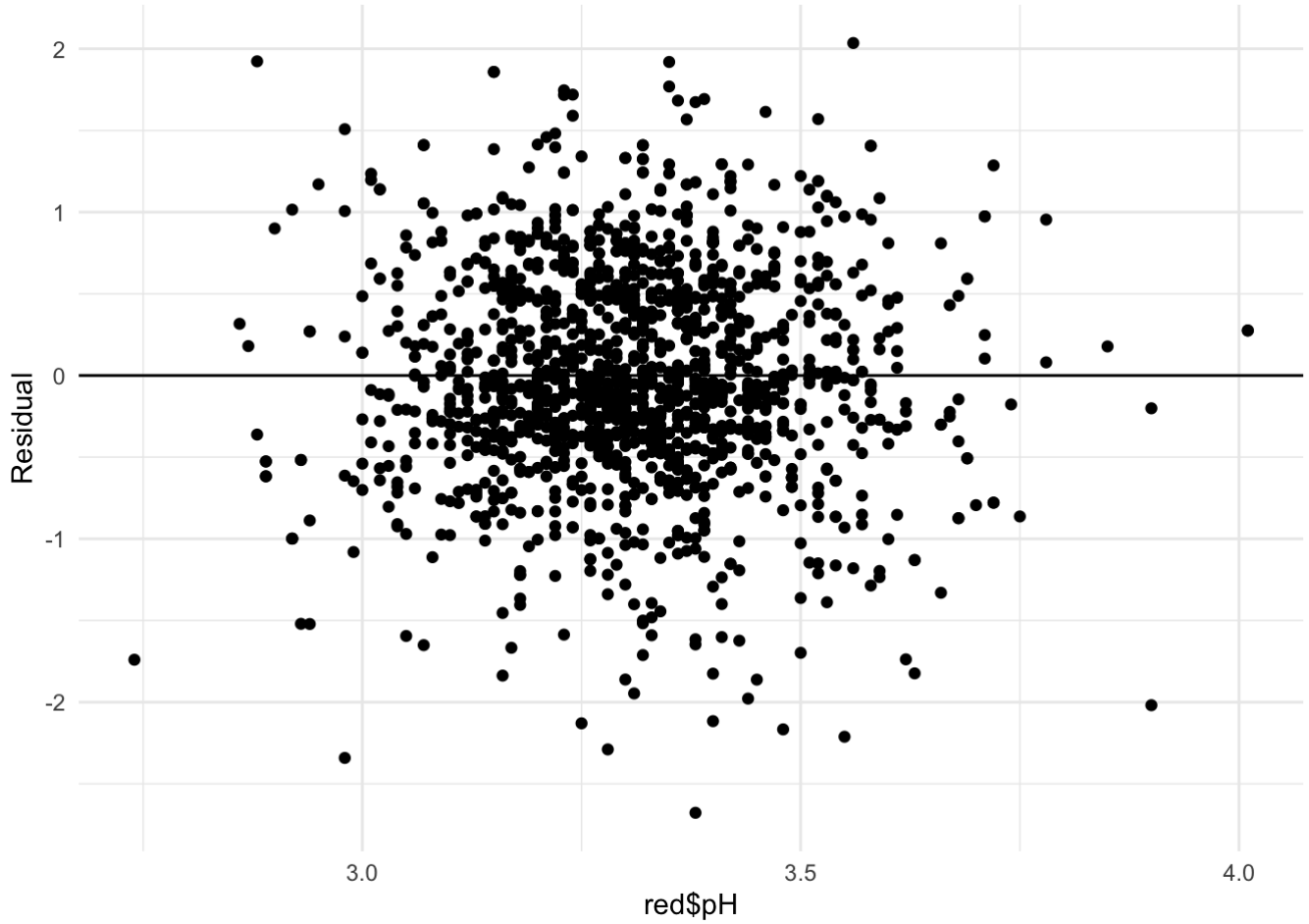
Residuals by total_sulfur

```
library("GGally")
ggplot(diagnostics) +
  geom_point(aes(x = red$total_sulfur, y = .resid)) +
  geom_hline(yintercept = 0) +
  ylab("Residual") +
  theme_minimal()
```



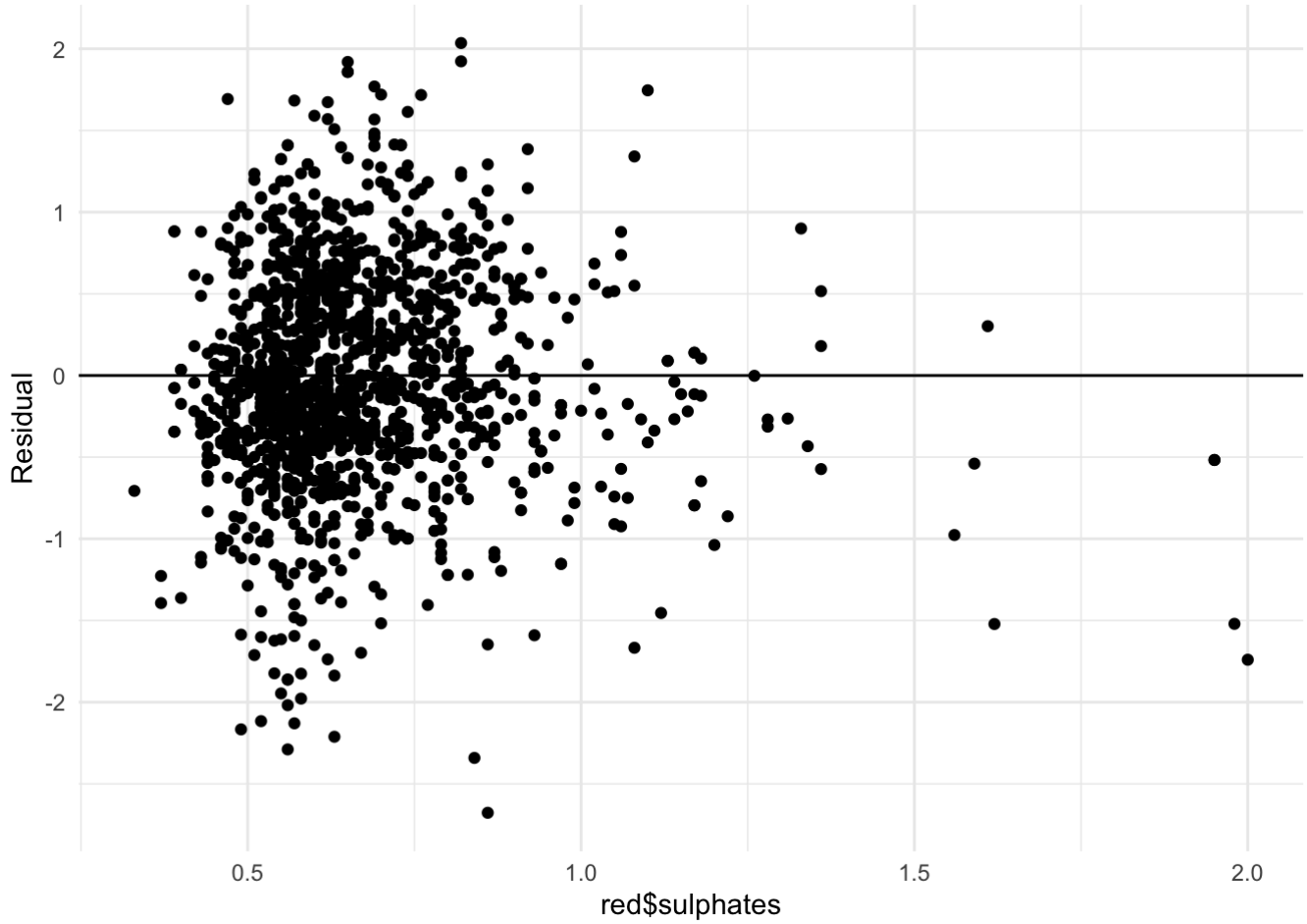
Residuals by pH

```
library("GGally")
ggplot(diagnostics) +
  geom_point(aes(x = red$pH, y = .resid)) +
  geom_hline(yintercept = 0) +
  ylab("Residual") +
  theme_minimal()
```



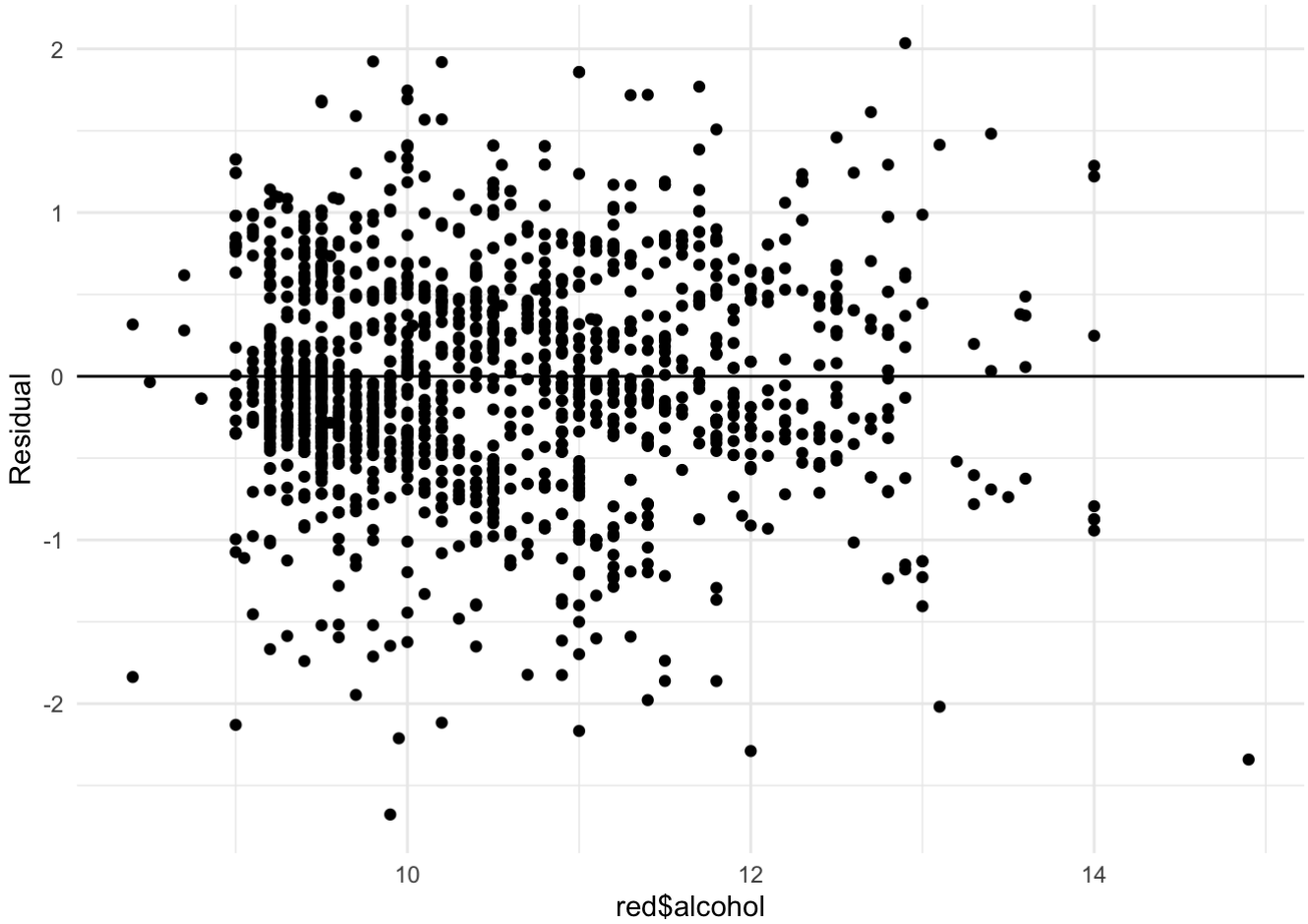
Residuals by sulphates

```
library("GGally")
ggplot(diagnostics) +
  geom_point(aes(x = red$sulphates, y = .resid)) +
  geom_hline(yintercept = 0) +
  ylab("Residual") +
  theme_minimal()
```



Residuals by alcohol

```
library("GGally")
ggplot(diagnostics) +
  geom_point(aes(x = red$alcohol, y = .resid)) +
  geom_hline(yintercept = 0) +
  ylab("Residual") +
  theme_minimal()
```


2
1
0
-1
-2
Residual
10
12
14
red$alcohol