Model Diagnosis

Weihao Wang

2025-02-10

Model Diagnostics for Linear Regression in R

To check the assumptions and performance of a linear regression model, we typically assess:

1. Linearity
2. Homoscedasticity (Constant Variance)
3. Normality of Residuals
4. Independence of Errors
5. Multicollinearity

```
library(MASS)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(glmnet)
```

```
## Loading required package: Matrix
##
```

```
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8
```

```r
library(leaps)

# Use variables except G1 and G2 to predicto G3
# y: G3
# x: all the other variables
data <- read.csv('math.csv', sep = ';')
data <- subset(data, select = - c(G1,G2))
str(data)
```

```
## 'data.frame':    395 obs. of  31 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

```r
set.seed(123)
training.samples <- data$G3 %>% createDataPartition(p = 0.75, list = FALSE) # caret pkg
# Uses createDataPartition() from the caret package to split the data into training (75%) and test (25%
```

```r
# The %>% operator (pipe) is from the tidyverse pkg
train.data  <- data[training.samples, ]
test.data <- data[-training.samples, ]
```

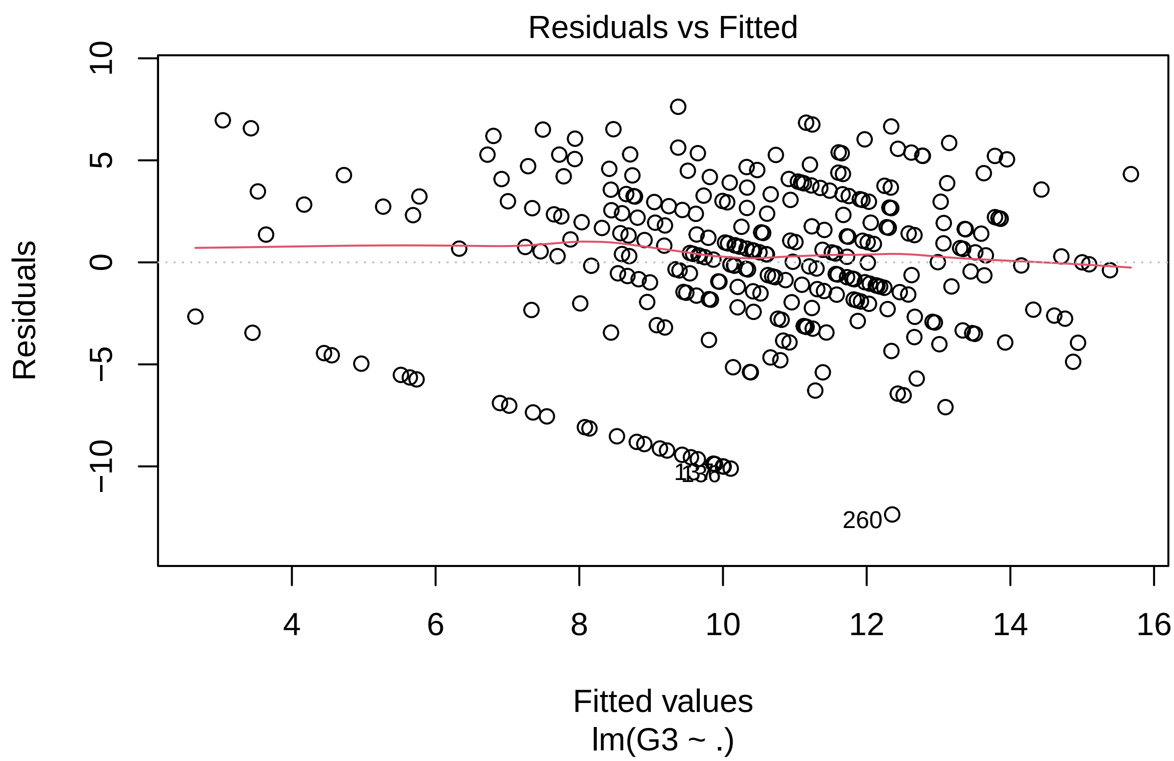
```r
fit <- lm(G3 ~ ., data = train.data)
```

A. Residual Plots for Linearity & Homoscedasticity

1. Residuals should be randomly scattered (no clear pattern).

2. The spread should be consistent across all fitted values.

```r
plot(fit, which = 1)  # Residuals vs Fitted
```

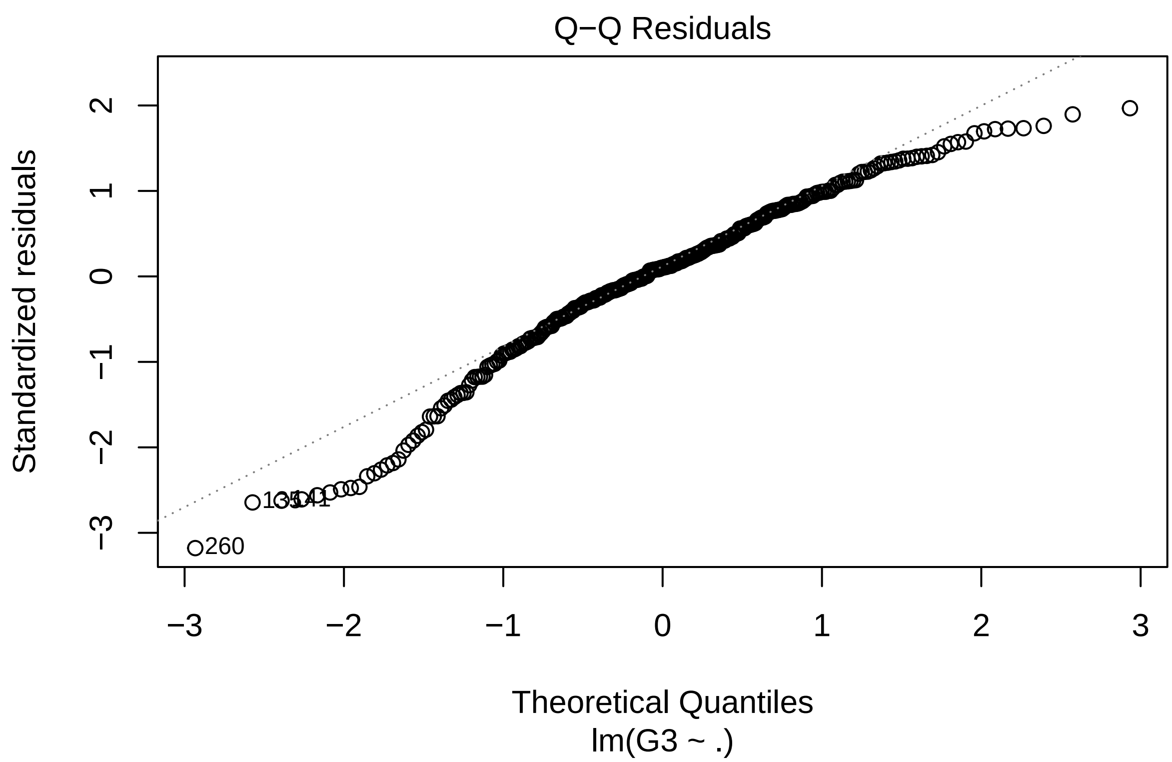



```r
# If categorical variables are treated as numeric, the model may create artificial "steps" in residuals
# If some predictors take only a few distinct values, residuals will align along specific horizontal le
# If some predictors are highly correlated, it can cause systematic patterns in residuals.
```

B. Check Normality of Residuals

Residuals should follow a normal distribution for valid hypothesis tests.

```r
plot(fit, which = 2)  # Normal Q-Q plot
```




```r
# both ends of the plot fall below the straight diagonal line, this indicates heavy-tailed (platykurtic
# Residuals may have Low Kurtosis ---> Try a Box-Cox transformation to adjust distribution
# or use rlm() robust regression from MASS
fit_robust <- rlm(G3 ~ ., data = train.data)
summary(fit_robust) # robust regression methods reduce the influence of extreme values
```

```
##
## Call: rlm(formula = G3 ~ ., data = train.data)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.619  -2.251   0.254   2.217   7.460
##
## Coefficients:
##                  Value   Std. Error t value
## (Intercept)      13.3383  4.9657     2.6861
## schoolMS         -0.0282  0.8335    -0.0339
## sexM              0.7790  0.5570     1.3986
## age              -0.3416  0.2386    -1.4314
## addressU          0.4745  0.6406     0.7407
## famsizeLE3        0.9276  0.5391     1.7205
## PstatusT          0.3388  0.8630     0.3925
## Medu              0.5470  0.3712     1.4736
## Fedu             -0.1165  0.3260    -0.3573
```

```
## Mjobhealth         1.2696  1.2135     1.0463
## Mjobother         -0.3352  0.7948    -0.4217
## Mjobservices       0.5784  0.8896     0.6501
## Mjobteacher       -1.1892  1.1754    -1.0117
## Fjobhealth         0.1457  1.5274     0.0954
## Fjobother         -0.5903  1.0572    -0.5584
## Fjobservices      -0.2410  1.0859    -0.2219
## Fjobteacher        1.1620  1.4137     0.8219
## reasonhome        -0.3444  0.6018    -0.5724
## reasonother        0.1675  0.9135     0.1833
## reasonreputation -0.0719  0.6387    -0.1125
## guardianmother     0.0281  0.5915     0.0476
## guardianother      0.3511  1.0933     0.3211
## traveltime         0.1143  0.3698     0.3091
## studytime          0.7693  0.3263     2.3579
## failures          -1.7511  0.3821    -4.5832
## schoolsupyes      -1.6345  0.7482    -2.1847
## famsupyes         -1.1835  0.5380    -2.1997
## paidyes            0.2360  0.5244     0.4501
## activitiesyes     -0.3188  0.4911    -0.6492
## nurseryyes        -0.4459  0.5948    -0.7497
## higheryes          0.7404  1.2301     0.6019
## internetyes        0.5848  0.6708     0.8718
## romanticyes       -0.9638  0.5199    -1.8537
## famrel             0.2567  0.2736     0.9380
## freetime           0.3093  0.2705     1.1436
## goout             -0.3709  0.2449    -1.5145
## Dalc              -0.2247  0.3547    -0.6336
## Walc               0.2057  0.2633     0.7814
## health            -0.2465  0.1843    -1.3380
## absences           0.0301  0.0335     0.8997
##
## Residual standard error: 3.321 on 258 degrees of freedom
```

```r
shapiro.test(residuals(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit)
## W = 0.96715, p-value = 2.655e-06
```

```r
# Residuals deviate from normality (consider transformations).
```

C. Check Homoscedasticity (Constant Variance)

Variance of residuals should be constant across fitted values.

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
bptest(fit)  # Breusch-Pagan test against heteroskedasticity
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 45.886, df = 39, p-value = 0.2083
```

```r
# p > 0.05 → No heteroscedasticity (good).
# p < 0.05 → Heteroscedasticity detected (consider transformations like log or Box-Cox).
```

D. Check for Autocorrelation (Independence of Errors)

Residuals should not be correlated over time or order.

```r
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```r
durbinWatsonTest(fit)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.04235254      1.914138   0.414
##  Alternative hypothesis: rho != 0
```

```r
# p > 0.05 → No significant autocorrelation (good).
# p < 0.05 → Autocorrelation detected (consider time-series models).
```

E. Check for Multicollinearity

High correlation between predictors can distort coefficient estimates

```r
library(car)
vif(fit)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## school     1.524719  1        1.234795
## sex        1.525539  1        1.235127
## age        1.849935  1        1.360123
## address    1.469077  1        1.212055
## famsize    1.193100  1        1.092291
## Pstatus    1.210519  1        1.100236
## Medu       3.302415  1        1.817255
## Fedu       2.374506  1        1.540943
## Mjob       4.065473  4        1.191623
## Fjob       2.615875  4        1.127722
## reason     1.658702  3        1.087998
## guardian   1.842892  2        1.165132
## traveltime 1.394992  1        1.181098
## studytime  1.419834  1        1.191568
## failures   1.517722  1        1.231959
## schoolsup  1.255873  1        1.120658
## famsup     1.357404  1        1.165077
## paid       1.349794  1        1.161806
## activities 1.189046  1        1.090434
## nursery    1.203246  1        1.096926
## higher     1.336228  1        1.155953
## internet   1.316931  1        1.147576
## romantic   1.182968  1        1.087643
## famrel     1.175115  1        1.084027
## freetime   1.396134  1        1.181581
## goout      1.484819  1        1.218532
## Dalc       2.186154  1        1.478565
## Walc       2.394839  1        1.547527
## health     1.288282  1        1.135025
## absences   1.324735  1        1.150971
```

```r
# GVIF (Generalized Variance Inflation Factor): Used for categorical variables (factors with multiple l
# GVIF^(1/(2*Df)): Adjusted VIF for easier interpretation when a factor has more than one degree of fre
# VIF < 5: No severe multicollinearity (good).
# VIF > 10: Strong multicollinearity (consider removing or combining variables).
```