Final project best

yuyang chen

2023-06-12

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(latex2exp)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(readr)
library(ggplot2)
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
```

```
set.seed(302)
data1 <- read.csv("train.csv") %>%
    select(SalePrice,OverallCond,OverallQual,BsmtFinSF1,GrLivArea,GarageCars,
           FullBath,MoSold,
            BedroomAbvGr,KitchenAbvGr)
```

```
data1 <- data1 %>%
  mutate(across(where(is.character), ~parse_number(trimws(.))))
data1 <- na.omit(data1)
head(data1)
```

```
##   SalePrice OverallCond OverallQual BsmtFinSF1 GrLivArea GarageCars FullBath
## 1    208500           5           7        706      1710          2        2
## 2    181500           8           6        978      1262          2        2
## 3    223500           5           7        486      1786          2        2
## 4    140000           5           7        216      1717          3        1
## 5    250000           5           8        655      2198          3        2
## 6    143000           5           5        732      1362          2        1
##   MoSold BedroomAbvGr KitchenAbvGr
## 1      2            3            1
## 2      5            3            1
## 3      9            3            1
## 4      2            3            1
## 5     12            4            1
## 6     10            1            1
```

```
# Random select 500 sample 350 for train and 150 for test
rows_train <- sample(1:nrow(data1), 350, replace = FALSE)
train <- data1[rows_train, ]

rows_test <- sample(setdiff(1:nrow(data1), rows_train), 150, replace = FALSE)
test <- data1[rows_test, ]
```

```
constant <- 1e-10
transform <- powerTransform(train[, 1:10] + constant)
summary(transform)
```

```
## bcPower Transformations to Multinormality
##              Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## SalePrice       0.0570        0.00      -0.0641       0.1781
## OverallCond     0.6893        0.50       0.4304       0.9482
## OverallQual     1.1121        1.00       0.8553       1.3688
## BsmtFinSF1      0.0809        0.08       0.0671       0.0948
## GrLivArea       0.0115        0.00      -0.1705       0.1934
## GarageCars      0.5564        0.56       0.5034       0.6095
## FullBath        0.6920        0.69       0.5885       0.7956
## MoSold          0.8563        1.00       0.6721       1.0404
## BedroomAbvGr    0.9332        1.00       0.7613       1.1052
## KitchenAbvGr  -25.7793      -25.78     -28.4801     -23.0784
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                             LRT df       pval
## LR test, lambda = (0 0 0 0 0 0 0 0 0 0) 5170.089 10 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                             LRT df       pval
## LR test, lambda = (1 1 1 1 1 1 1 1 1 1) 6653.044 10 < 2.22e-16
```

```
train$logprice=log(train$SalePrice)
test$logprice=log(test$SalePrice)
```

```r
summary(train[,c(2:11)])
```

```
##   OverallCond     OverallQual      BsmtFinSF1       GrLivArea
##  Min.   :1.000   Min.   : 1.00   Min.   :   0.0   Min.   : 480
##  1st Qu.:5.000   1st Qu.: 5.00   1st Qu.:   0.0   1st Qu.:1132
##  Median :5.000   Median : 6.00   Median : 398.0   Median :1452
##  Mean   :5.494   Mean   : 6.12   Mean   : 459.3   Mean   :1492
##  3rd Qu.:6.000   3rd Qu.: 7.00   3rd Qu.: 732.8   3rd Qu.:1748
##  Max.   :9.000   Max.   :10.00   Max.   :1646.0   Max.   :3627
##    GarageCars       FullBath         MoSold        BedroomAbvGr
##  Min.   :0.000   Min.   :0.000   Min.   : 1.000   Min.   :0.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 5.000   1st Qu.:2.00
##  Median :2.000   Median :2.000   Median : 6.000   Median :3.00
##  Mean   :1.789   Mean   :1.603   Mean   : 6.271   Mean   :2.82
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.: 8.000   3rd Qu.:3.00
##  Max.   :3.000   Max.   :3.000   Max.   :12.000   Max.   :6.00
##   KitchenAbvGr      logprice
##  Min.   :1.000   Min.   :10.47
##  1st Qu.:1.000   1st Qu.:11.78
##  Median :1.000   Median :12.00
##  Mean   :1.057   Mean   :12.03
##  3rd Qu.:1.000   3rd Qu.:12.28
##  Max.   :3.000   Max.   :13.35
```

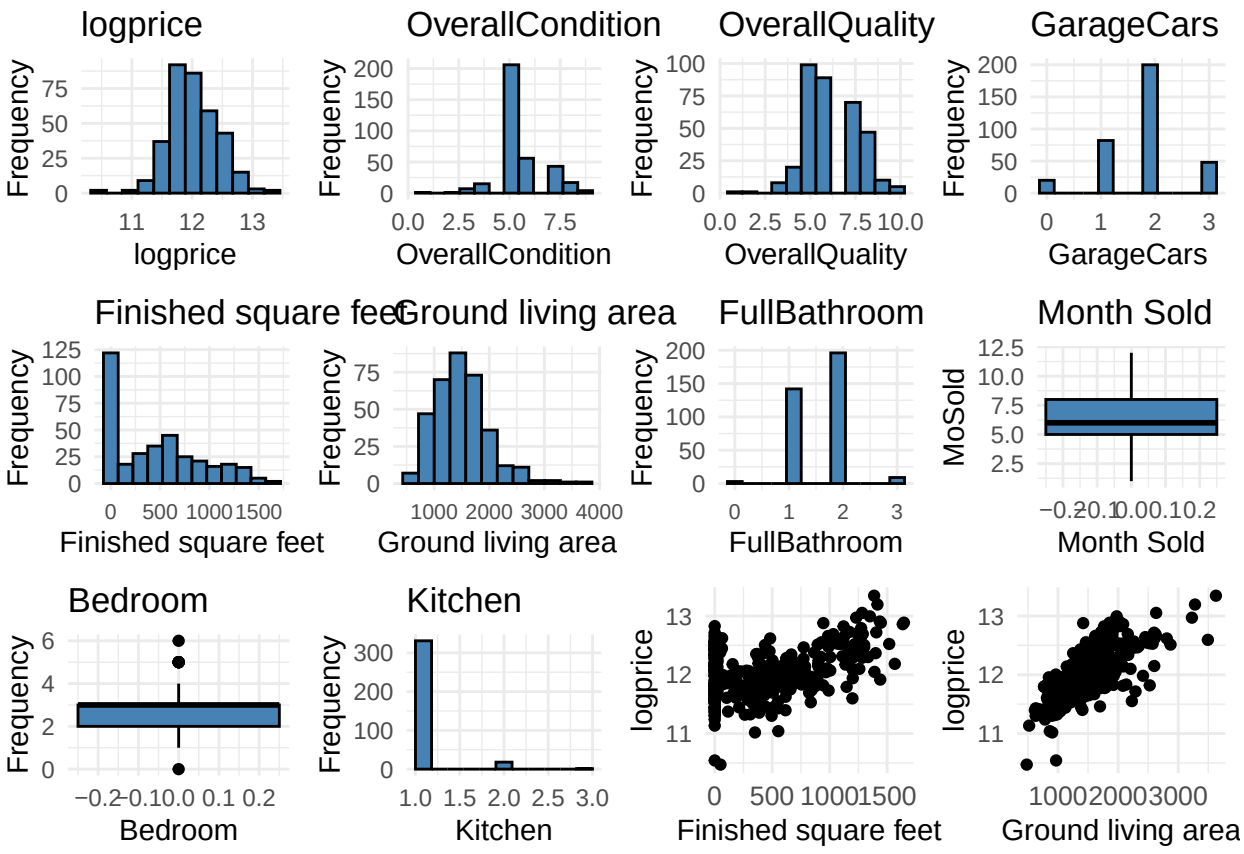
```r
attach(train)
logprice_plot <- ggplot(data = train, aes(x = logprice)) +
  geom_histogram(fill = "steelblue", color = "black", bins = 12) +
  labs(x = "logprice", y = "Frequency",
       title = "logprice") +
  theme_minimal()
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("logprice")+
  theme(plot.title = element_text(size = 13))
```

```
## List of 2
##  $ plot.title:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : num 13
##   ..$ hjust        : num 0.5
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ title     : chr "logprice"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```r
OverallCond_plot <- ggplot(data = train, aes(x = OverallCond)) +
  geom_histogram(fill = "steelblue", color = "black", bins=12) +
```

```
  labs(x = "OverallCondition", y = "Frequency",
       title = "OverallCondition") +
  theme_minimal() +
  theme(plot.title = element_text(size = 13))

OverallQual_plot <- ggplot(data = train, aes(x = OverallQual)) +
  geom_histogram(fill = "steelblue", color = "black", bins = 12) +
  labs(x = "OverallQuality", y = "Frequency",
       title = "OverallQuality") +
  theme_minimal()+
  theme(plot.title = element_text(size = 13))

BsmtFinSF1_plot <- ggplot(data = train, aes(x = BsmtFinSF1)) +
  geom_histogram(fill = "steelblue", color = "black", bins = 12) +
  labs(x = "Finished square feet", y = "Frequency",
       title = "Finished square feet") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 13),
    axis.text.x = element_text(size = 8) )

GrLivArea_plot  <- ggplot(data = train, aes(x = GrLivArea )) +
  geom_histogram(fill = "steelblue", color = "black", bins = 12) +
  labs(x = "Ground living area", y = "Frequency",
       title = "Ground living area") +
  theme_minimal()+
  theme(plot.title = element_text(size = 13),axis.text.x = element_text(size = 8) )

GarageCars_plot  <- ggplot(data = train, aes(x = GarageCars )) +
  geom_histogram(fill = "steelblue", color = "black", bins = 12) +
  labs(x = "GarageCars", y = "Frequency",
       title = "GarageCars") +
  theme_minimal()+
  theme(plot.title = element_text(size = 13))

FullBath_plot  <- ggplot(data = train, aes(x = FullBath )) +
  geom_histogram(fill = "steelblue", color = "black", bins = 12) +
  labs(x = "FullBathroom", y = "Frequency",
       title = "FullBathroom") +
  theme_minimal()+
  theme(plot.title = element_text(size = 13),axis.text.x = element_text(size = 8))


MoSold_plot <- ggplot(data = train, aes(y = MoSold)) +
  geom_boxplot(fill = "steelblue", color = "black", width = 0.5) +
  labs(x = "Month Sold", y = "MoSold",
       title = "Month Sold") +
  theme_minimal() +
  theme(plot.title = element_text(size = 13))

BedroomAbvGr_plot <- ggplot(data = train, aes(x = BedroomAbvGr)) +
  geom_boxplot(fill = "steelblue", color = "black", width = 0.5) +
  labs(x = "Frequency", y= "Bedroom",
```

```
       title = "Bedroom") +
  theme_minimal() + coord_flip()+
  theme(plot.title = element_text(size = 13))

KitchenAbvGr_plot <- ggplot(data = train, aes(x = KitchenAbvGr)) +
  geom_histogram(fill = "steelblue", color = "black", bins = 12) +
  labs(x = "Kitchen", y = "Frequency",
       title = "Kitchen") +
  theme_minimal()+
  theme(plot.title = element_text(size = 13))

scatter_plot1 <- ggplot(data = train, aes(x = BsmtFinSF1, y = logprice)) +
  geom_point() +
  labs(x = "Finished square feet", y = "logprice") +
  theme_minimal()+
  theme(axis.text.x = element_text(size = 10))

scatter_plot2 <- ggplot(data = train, aes(x = GrLivArea, y = logprice)) +
  geom_point() +
  labs(x = "Ground living area", y = "logprice") +
  theme_minimal()+
  theme(axis.text.x = element_text(size = 10),)
grid.arrange(logprice_plot, OverallCond_plot, OverallQual_plot,
             GarageCars_plot,BsmtFinSF1_plot,GrLivArea_plot
             ,FullBath_plot,MoSold_plot, BedroomAbvGr_plot,KitchenAbvGr_plot,scatter_plot1,scatter_plot
```



```
nullmod <- lm(logprice ~ 1, data = train)
fullmod <- lm(logprice ~ OverallCond + OverallQual + BsmtFinSF1 + GrLivArea +
            GarageCars + GarageCars + FullBath + MoSold + BedroomAbvGr +
              KitchenAbvGr ,data = train)
mboth = step(nullmod, scope = list(lower= formula(nullmod), upper =
                                     formula(fullmod)), direction = "both")
```

```
## Start:  AIC=-630.19
## logprice ~ 1
##
##                Df Sum of Sq    RSS      AIC
## + OverallQual   1    38.652 18.841 -1018.66
## + GrLivArea     1    29.522 27.971  -880.36
## + GarageCars    1    25.441 32.053  -832.69
## + FullBath      1    20.833 36.660  -785.68
## + BsmtFinSF1    1     9.295 48.198  -689.91
## + BedroomAbvGr  1     3.887 53.606  -652.69
## + KitchenAbvGr  1     1.756 55.737  -639.05
## + OverallCond   1     0.355 57.139  -630.36
## <none>                      57.493  -630.19
## + MoSold        1     0.227 57.266  -629.58
##
## Step:  AIC=-1018.66
## logprice ~ OverallQual
##
##                Df Sum of Sq    RSS      AIC
## + GrLivArea     1     5.752 13.089 -1144.15
## + GarageCars    1     2.850 15.991 -1074.08
## + BsmtFinSF1    1     2.525 16.316 -1067.02
## + FullBath      1     2.339 16.502 -1063.06
## + BedroomAbvGr  1     1.389 17.452 -1043.47
## <none>                      18.841 -1018.66
## + MoSold        1     0.080 18.761 -1018.16
## + KitchenAbvGr  1     0.031 18.810 -1017.24
## + OverallCond   1     0.000 18.841 -1016.67
## - OverallQual   1    38.652 57.493  -630.19
##
## Step:  AIC=-1144.15
## logprice ~ OverallQual + GrLivArea
##
##                Df Sum of Sq    RSS      AIC
## + BsmtFinSF1    1    2.5545 10.535 -1218.14
## + GarageCars    1    1.6367 11.453 -1188.90
## + KitchenAbvGr  1    0.2085 12.881 -1147.77
## + FullBath      1    0.1377 12.951 -1145.85
## + BedroomAbvGr  1    0.1286 12.960 -1145.61
## <none>                      13.089 -1144.15
## + MoSold        1    0.0656 13.024 -1143.91
## + OverallCond   1    0.0006 13.088 -1142.17
## - GrLivArea     1    5.7519 18.841 -1018.66
## - OverallQual   1   14.8822 27.971  -880.36
##
## Step:  AIC=-1218.14
## logprice ~ OverallQual + GrLivArea + BsmtFinSF1
```

```
##
##                Df Sum of Sq     RSS      AIC
## + GarageCars    1    1.3233  9.2114 -1263.12
## + FullBath      1    0.2011 10.3336 -1222.89
## + KitchenAbvGr  1    0.1028 10.4319 -1219.57
## <none>                      10.5347 -1218.14
## + MoSold        1    0.0260 10.5087 -1217.00
## + OverallCond   1    0.0179 10.5168 -1216.74
## + BedroomAbvGr  1    0.0028 10.5318 -1216.24
## - BsmtFinSF1    1    2.5545 13.0892 -1144.15
## - GrLivArea     1    5.7816 16.3163 -1067.02
## - OverallQual   1   11.8743 22.4090  -955.96
##
## Step:  AIC=-1263.12
## logprice ~ OverallQual + GrLivArea + BsmtFinSF1 + GarageCars
##
##                Df Sum of Sq     RSS     AIC
## + OverallCond   1    0.1336  9.0777 -1266.2
## + FullBath      1    0.0985  9.1128 -1264.9
## + KitchenAbvGr  1    0.0656  9.1458 -1263.6
## <none>                      9.2114 -1263.1
## + MoSold        1    0.0237  9.1877 -1262.0
## + BedroomAbvGr  1    0.0236  9.1878 -1262.0
## - GarageCars    1    1.3233 10.5347 -1218.1
## - BsmtFinSF1    1    2.2411 11.4525 -1188.9
## - GrLivArea     1    4.6571 13.8685 -1121.9
## - OverallQual   1    6.6347 15.8461 -1075.2
##
## Step:  AIC=-1266.24
## logprice ~ OverallQual + GrLivArea + BsmtFinSF1 + GarageCars +
##     OverallCond
##
##                Df Sum of Sq     RSS     AIC
## + FullBath      1    0.1061  8.9717 -1268.3
## <none>                      9.0777 -1266.2
## + KitchenAbvGr  1    0.0452  9.0325 -1266.0
## + MoSold        1    0.0153  9.0624 -1264.8
## + BedroomAbvGr  1    0.0112  9.0665 -1264.7
## - OverallCond   1    0.1336  9.2114 -1263.1
## - GarageCars    1    1.4390 10.5168 -1216.7
## - BsmtFinSF1    1    2.2928 11.3705 -1189.4
## - GrLivArea     1    4.5989 13.6767 -1124.8
## - OverallQual   1    6.5771 15.6549 -1077.5
##
## Step:  AIC=-1268.35
## logprice ~ OverallQual + GrLivArea + BsmtFinSF1 + GarageCars +
##     OverallCond + FullBath
##
##                Df Sum of Sq     RSS     AIC
## + KitchenAbvGr  1    0.0907  8.8809 -1269.9
## <none>                      8.9717 -1268.3
## + MoSold        1    0.0130  8.9587 -1266.9
## + BedroomAbvGr  1    0.0030  8.9686 -1266.5
## - FullBath      1    0.1061  9.0777 -1266.2
```

```
## - OverallCond   1    0.1412  9.1128 -1264.9
## - GarageCars    1    1.3368 10.3085 -1221.7
## - BsmtFinSF1    1    2.3438 11.3154 -1189.1
## - GrLivArea     1    2.9374 11.9090 -1171.2
## - OverallQual   1    6.0162 14.9879 -1090.7
##
## Step:  AIC=-1269.91
## logprice ~ OverallQual + GrLivArea + BsmtFinSF1 + GarageCars +
##     OverallCond + FullBath + KitchenAbvGr
##
##                Df Sum of Sq     RSS     AIC
## <none>                       8.8809 -1269.9
## + MoSold        1    0.0127  8.8682 -1268.4
## - KitchenAbvGr  1    0.0907  8.9717 -1268.3
## + BedroomAbvGr  1    0.0023  8.8787 -1268.0
## - OverallCond   1    0.1137  8.9946 -1267.5
## - FullBath      1    0.1516  9.0325 -1266.0
## - GarageCars    1    1.2464 10.1273 -1225.9
## - BsmtFinSF1    1    2.2612 11.1422 -1192.5
## - GrLivArea     1    3.0219 11.9028 -1169.4
## - OverallQual   1    5.0035 13.8844 -1115.5
```

```
formula(mboth)
```

```
## logprice ~ OverallQual + GrLivArea + BsmtFinSF1 + GarageCars +
##     OverallCond + FullBath + KitchenAbvGr
```

```
summary(mboth)
```

```
##
## Call:
## lm(formula = logprice ~ OverallQual + GrLivArea + BsmtFinSF1 +
##     GarageCars + OverallCond + FullBath + KitchenAbvGr, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66706 -0.07337  0.00828  0.09348  0.49405
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.050e+01  8.164e-02 128.624  < 2e-16 ***
## OverallQual   1.229e-01  8.850e-03  13.881  < 2e-16 ***
## GrLivArea     2.680e-04  2.485e-05  10.787  < 2e-16 ***
## BsmtFinSF1    1.880e-04  2.015e-05   9.332  < 2e-16 ***
## GarageCars    1.047e-01  1.511e-02   6.928 2.13e-11 ***
## OverallCond   1.688e-02  8.066e-03   2.092   0.0372 *
## FullBath      5.305e-02  2.196e-02   2.416   0.0162 *
## KitchenAbvGr -7.346e-02  3.930e-02  -1.869   0.0625 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1611 on 342 degrees of freedom
## Multiple R-squared:  0.8455, Adjusted R-squared:  0.8424
## F-statistic: 267.4 on 7 and 342 DF,  p-value: < 2.2e-16
```

```
model1 <- lm(logprice ~ OverallCond + OverallQual + BsmtFinSF1 + GrLivArea +
            GarageCars + GarageCars + FullBath + MoSold + BedroomAbvGr
            +KitchenAbvGr ,
            data = train)

summary(model1)
```

```
##
## Call:
## lm(formula = logprice ~ OverallCond + OverallQual + BsmtFinSF1 +
##     GrLivArea + GarageCars + GarageCars + FullBath + MoSold +
##     BedroomAbvGr + KitchenAbvGr, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66403 -0.07438  0.00982  0.08899  0.50608
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.051e+01  8.679e-02 121.087  < 2e-16 ***
## OverallCond   1.596e-02  8.197e-03   1.948   0.0523 .
## OverallQual   1.242e-01  9.083e-03  13.671  < 2e-16 ***
## BsmtFinSF1    1.883e-04  2.051e-05   9.185  < 2e-16 ***
## GrLivArea     2.622e-04  2.964e-05   8.845  < 2e-16 ***
## GarageCars    1.054e-01  1.536e-02   6.863 3.21e-11 ***
## FullBath      5.124e-02  2.231e-02   2.297   0.0223 *
## MoSold       -2.598e-03  3.539e-03  -0.734   0.4634
## BedroomAbvGr  5.319e-03  1.437e-02   0.370   0.7115
## KitchenAbvGr -7.298e-02  3.939e-02  -1.853   0.0648 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1615 on 340 degrees of freedom
## Multiple R-squared:  0.8458, Adjusted R-squared:  0.8417
## F-statistic: 207.2 on 9 and 340 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(logprice ~  OverallQual + BsmtFinSF1 + GrLivArea +
            GarageCars + FullBath,
            data = train)
summary(model2)
```

```
##
## Call:
## lm(formula = logprice ~ OverallQual + BsmtFinSF1 + GrLivArea +
##     GarageCars + FullBath, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65122 -0.07168  0.01150  0.09595  0.50082
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.051e+01  3.921e-02 268.059  < 2e-16 ***
## OverallQual 1.287e-01  8.489e-03  15.156  < 2e-16 ***
```

```
## BsmtFinSF1  1.883e-04  2.026e-05   9.294  < 2e-16 ***
## GrLivArea   2.647e-04  2.486e-05  10.644  < 2e-16 ***
## GarageCars  1.011e-01  1.489e-02   6.788 4.98e-11 ***
## FullBath    4.147e-02  2.150e-02   1.929   0.0546 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1628 on 344 degrees of freedom
## Multiple R-squared:  0.8415, Adjusted R-squared:  0.8392
## F-statistic: 365.3 on 5 and 344 DF,  p-value: < 2.2e-16
```

```
model3 <- lm(logprice ~  OverallQual + BsmtFinSF1 + GrLivArea +
            GarageCars,
            data = train)
summary(model3)
```

```
##
## Call:
## lm(formula = logprice ~ OverallQual + BsmtFinSF1 + GrLivArea +
##     GarageCars, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64906 -0.07398  0.00932  0.09536  0.51344
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.052e+01  3.918e-02 268.423  < 2e-16 ***
## OverallQual 1.318e-01  8.363e-03  15.764  < 2e-16 ***
## BsmtFinSF1  1.861e-04  2.031e-05   9.162  < 2e-16 ***
## GrLivArea   2.880e-04  2.181e-05  13.207  < 2e-16 ***
## GarageCars  1.045e-01  1.484e-02   7.040 1.04e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1634 on 345 degrees of freedom
## Multiple R-squared:  0.8398, Adjusted R-squared:  0.8379
## F-statistic: 452.1 on 4 and 345 DF,  p-value: < 2.2e-16
```

```
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: logprice
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## OverallQual   1 38.652  38.652 1447.678 < 2.2e-16 ***
## BsmtFinSF1    1  2.525   2.525   94.563 < 2.2e-16 ***
## GrLivArea     1  5.782   5.782  216.542 < 2.2e-16 ***
## GarageCars    1  1.323   1.323   49.562 1.044e-11 ***
## Residuals   345  9.211   0.027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

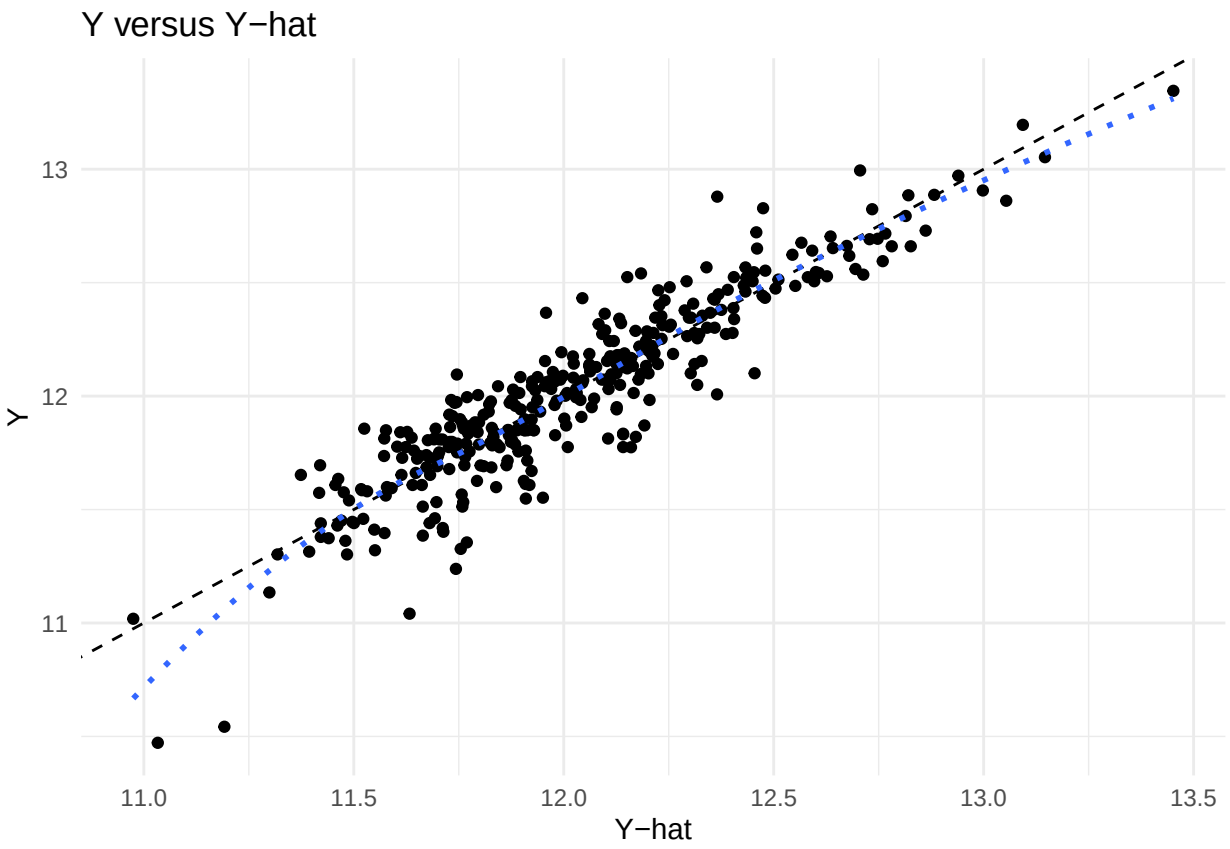
```
model3_F_test = lm(logprice ~ OverallQual + BsmtFinSF1,
            data = train)
anova(model3,model3_F_test)
```

```
## Analysis of Variance Table
##
## Model 1: logprice ~ OverallQual + BsmtFinSF1 + GrLivArea + GarageCars
## Model 2: logprice ~ OverallQual + BsmtFinSF1
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    345  9.2114
## 2    347 16.3163 -2   -7.1049 133.05 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

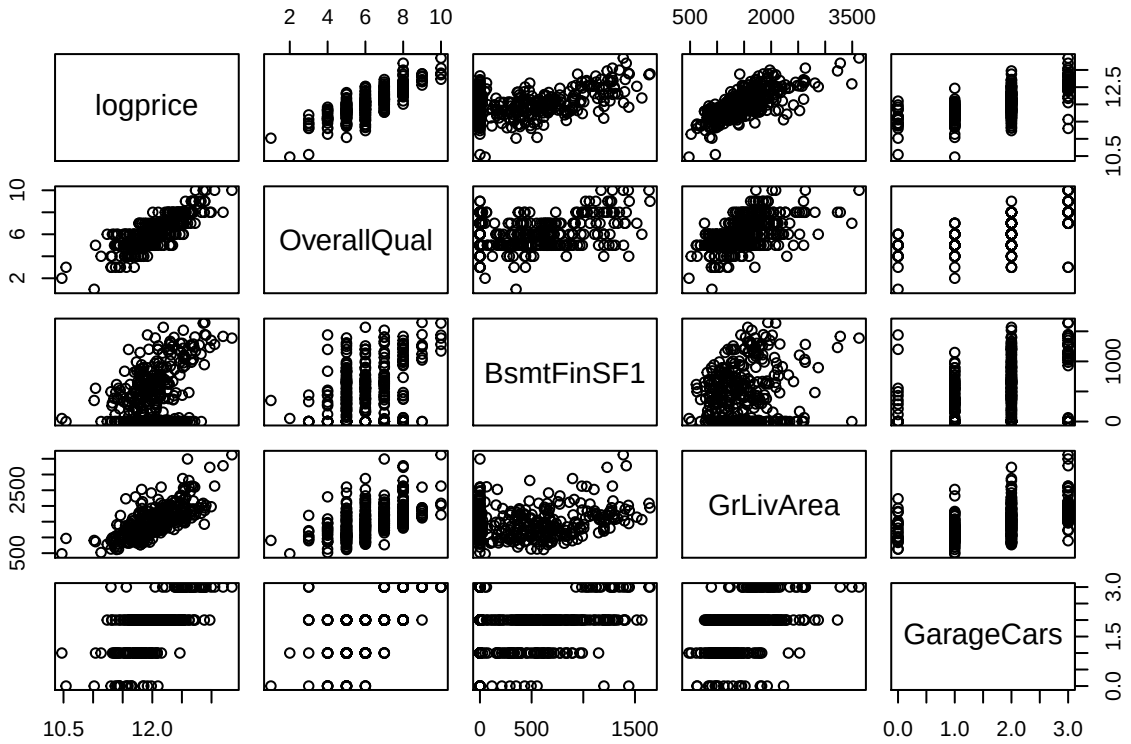
```
model5 = lm(formula = logprice ~ OverallQual + GrLivArea + BsmtFinSF1 +
    GarageCars + OverallCond + FullBath + KitchenAbvGr, data = train)
```

```
   r <- resid(model3)
# first check condition 1 and 2
#condition 1
# Create a data frame with Y and Y-hat
  comparison <- data.frame(Y = train$logprice, Y_hat = fitted(model3))
# Plot Y versus Y-hat
  ggplot(comparison, aes(x = Y_hat, y = Y)) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
    geom_smooth(method = "loess", se = FALSE, linetype = "dotted") +
    labs(x = "Y-hat", y = "Y", title = "Y versus Y-hat") +
    theme_minimal()
```

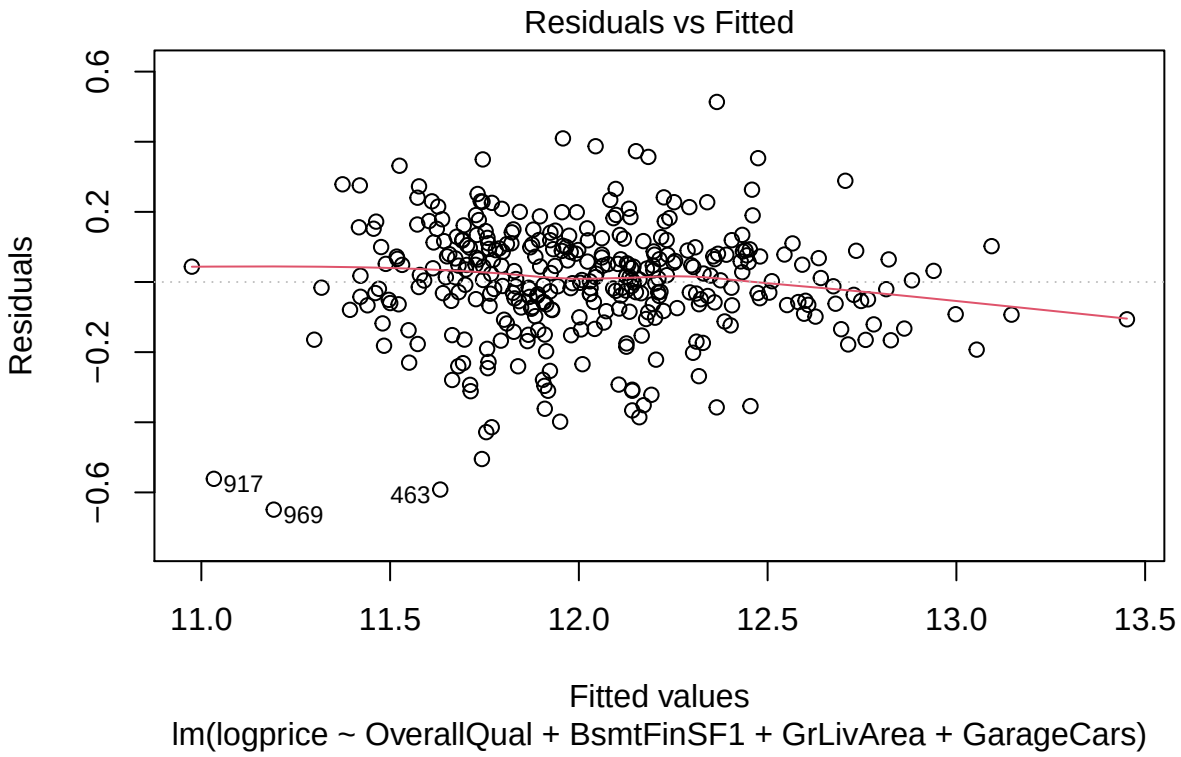
```
## `geom_smooth()` using formula = 'y ~ x'
```



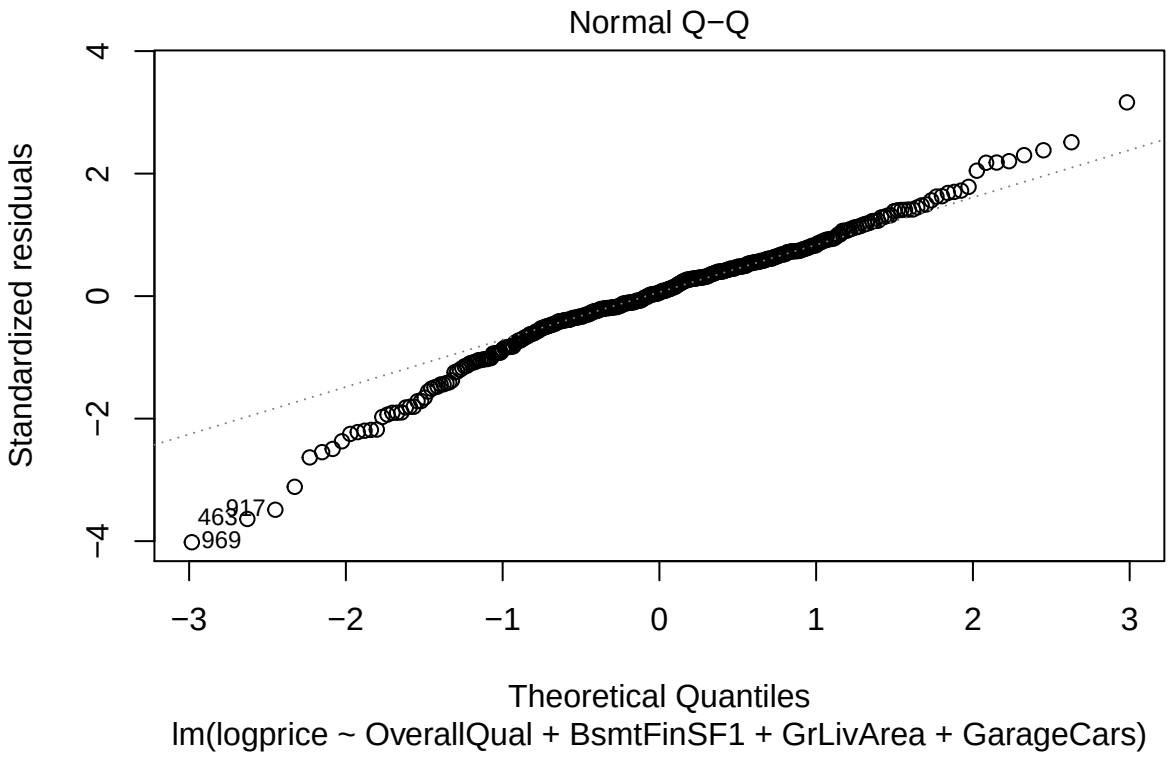
```
  r <- resid(model3)
  #condition 2
# Create a scatter plot matrix with improved appearance
data2 = data.frame(logprice, OverallQual, BsmtFinSF1,GrLivArea,GarageCars)
pairs( data2 )
```



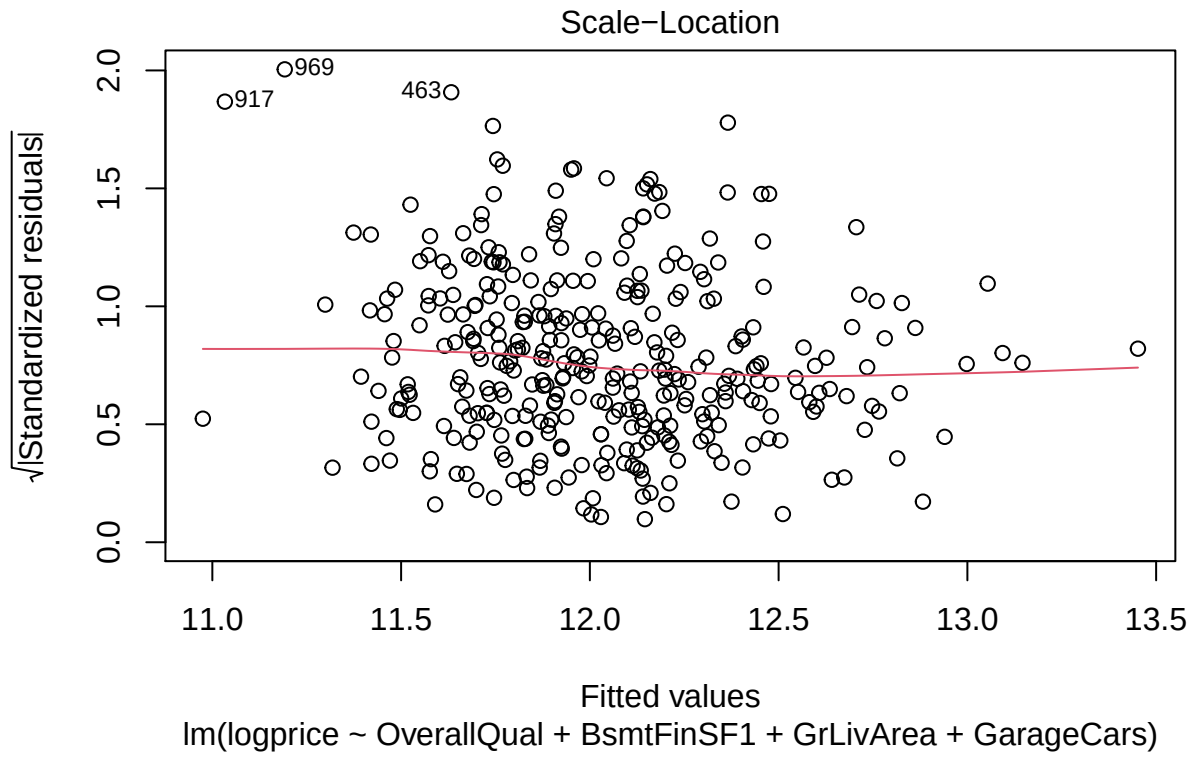

```
plot(model3, 1)  # Plot 1 with default title
```



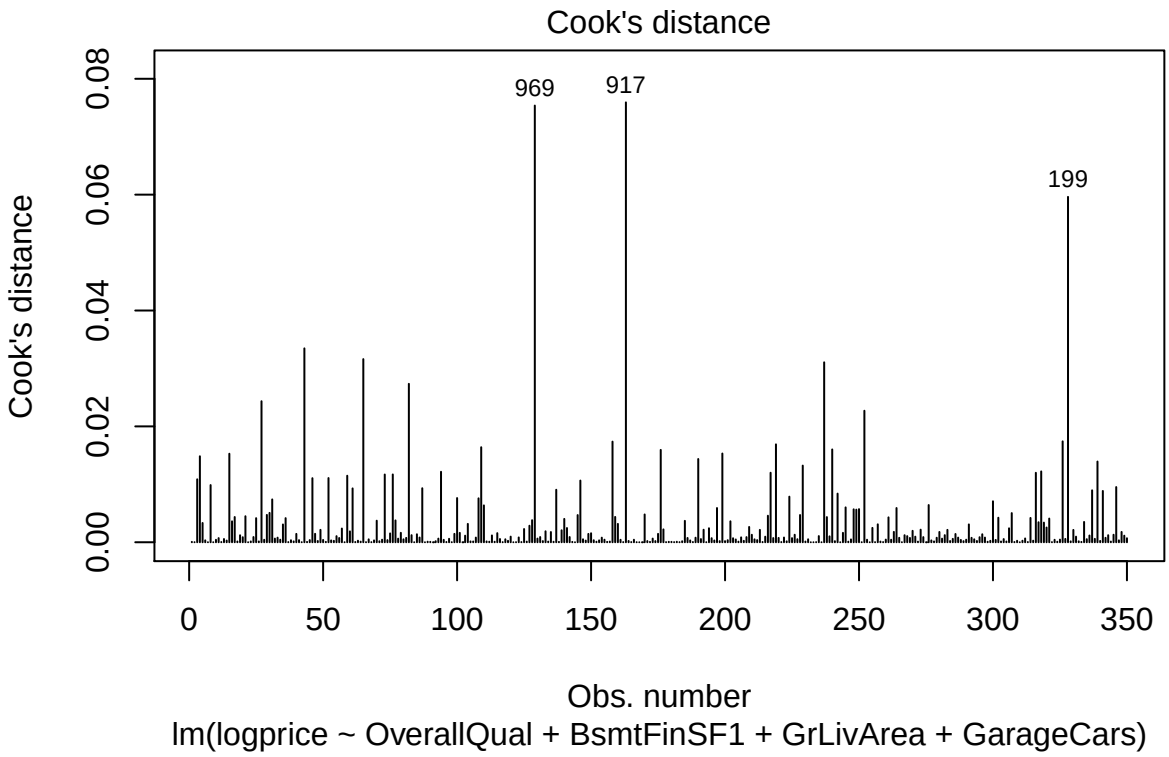

```
plot(model3, 2)
```



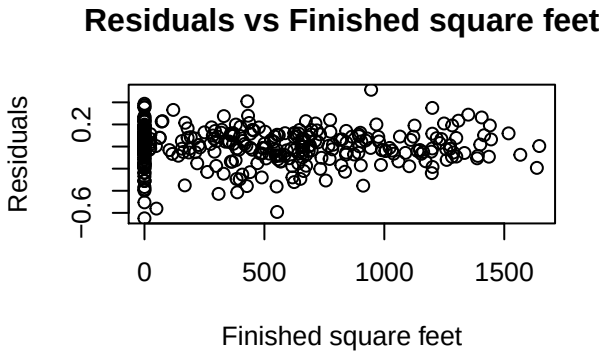

```
plot(model3, 3)
```

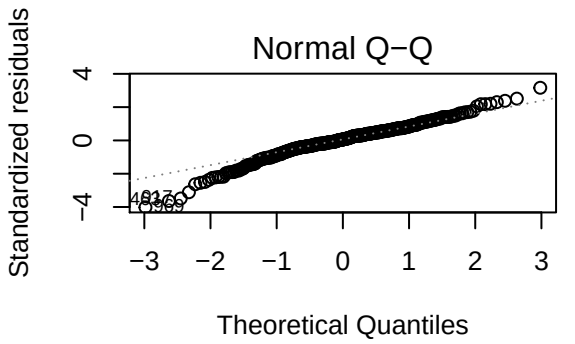


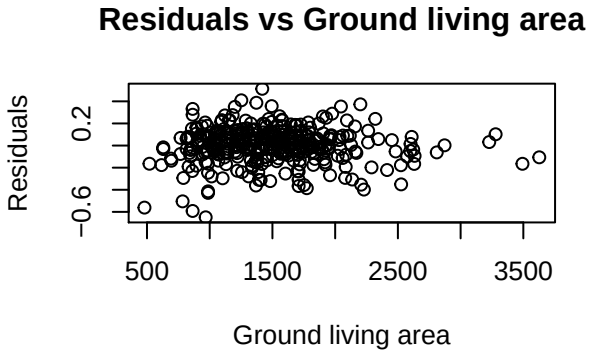

```
plot(model3, 4)
```

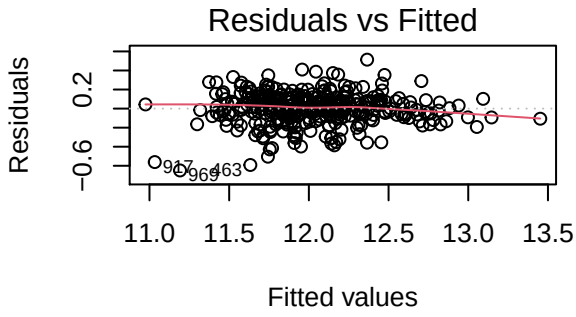



```
par(mfrow=c(2,2))
r <- model3$residuals
fit=model3$fitted.values
plot(r ~ train$BsmtFinSF1, xlab = "Finished square feet", ylab = "Residuals",
     main = "Residuals vs Finished square feet")
plot(model3, 2)
plot(r ~ train$GrLivArea, xlab = "Ground living area", ylab = "Residuals",
     main = "Residuals vs Ground living area")
plot(model3, 1)
```










```
par(mfrow=c(2,2))
```

```
r <- model3$residuals
out <- which(r > 2 | r < -2)
out
```

```
## named integer(0)
```

```
h <- hatvalues(model3)
threshold <- 2 * (length(model3$coefficients)/nrow(train))
w <- which(h > threshold)
train[w,]
```

```
##      SalePrice OverallCond OverallQual BsmtFinSF1 GrLivArea GarageCars FullBath
## 826     385000           5          10       1636      2084          3        2
## 1187     95000           5           3        440      1699          2        2
## 739     179000           5           5       1200      1200          0        3
## 516     402861           5          10       1436      2020          3        2
## 1361    189000           6           5          0      2601          2        3
## 962     272000           7           6        896      2872          2        2
## 1170    625000           5          10       1387      3627          3        3
## 1405    105000           4           3          0      1214          3        1
## 770     538000           5           8       1416      3279          3        3
## 497     430000           5           8       1231      3228          2        3
## 1091     92900           4           3          0      1040          2        2
## 1031    160000           8           5          0      1928          0        2
## 995     337500           5          10       1172      1718          3        2
## 917      35311           3           2         50       480          1        0
```

```
## 1001     82000           3           3           0       944           2         1
## 1062     81000           4           3           0       894           3         1
## 1284    139000           5           6           0      1824           0         2
## 1374    466500           5          10        1282      2633           3         2
## 1388    136000           7           6         168      2526           1         2
## 305     295000           9           7           0      3493           3         3
## 988     395192           5           9        1646      1940           3         2
## 308      89500           7           6           0      1406           0         1
## 844     141000           4           5           0      1800           0         2
## 343      87500           4           3           0      1040           2         2
## 1417    122500           6           4           0      2290           2         2
## 199     104000           6           6           0      2229           0         1
## 943     150000           3           4        1440      1440           0         2
## 376      61000           1           1         350       904           0         0
##      MoSold BedroomAbvGr KitchenAbvGr logprice
## 826       6            2            1 12.86100
## 1187      8            3            2 11.46163
## 739       3            3            1 12.09514
## 516       9            3            1 12.90635
## 1361      5            4            1 12.14950
## 962       7            4            1 12.51356
## 1170      7            4            1 13.34551
## 1405      1            3            1 11.56172
## 770       6            4            1 13.19561
## 497       5            4            1 12.97154
## 1091      6            2            2 11.43928
## 1031      7            5            2 11.98293
## 995       7            3            1 12.72932
## 917      10            1            1 10.47195
## 1001      7            2            1 11.31447
## 1062      8            2            1 11.30220
## 1284      4            4            2 11.84223
## 1374      3            2            1 13.05301
## 1388      8            5            1 11.82041
## 305       5            3            1 12.59473
## 988       4            3            1 12.88713
## 308       3            3            1 11.40199
## 844       7            6            2 11.85652
## 343       5            2            2 11.37939
## 1417      4            4            2 11.71587
## 199       7            5            1 11.55215
## 943       8            4            2 11.91839
## 376       3            1            1 11.01863
```

```
D <- cooks.distance(model3)
cutoff <- qf(0.5, length(model3$coefficients), nrow(train)-length(model3$coefficients), lower.tail=T)
influential_observations <- which(D > cutoff)
influential_observations
```

```
## named integer(0)
```

```
vif(model3)
```

```
## OverallQual  BsmtFinSF1   GrLivArea  GarageCars
##    1.872639    1.069862    1.487363    1.603288
```

```
vif(model5)
```

```
##  OverallQual    GrLivArea   BsmtFinSF1   GarageCars  OverallCond     FullBath
##     2.156650     1.985497     1.082880     1.707879     1.071109     2.001845
## KitchenAbvGr
##     1.240647
```

```
# Test model
model4 <- lm(logprice ~ OverallQual + BsmtFinSF1 + GrLivArea + GarageCars,
            data = test)

summary(model4)
```

```
##
## Call:
## lm(formula = logprice ~ OverallQual + BsmtFinSF1 + GrLivArea +
##     GarageCars, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58213 -0.06493  0.02031  0.10624  0.36712
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.048e+01  6.477e-02 161.868  < 2e-16 ***
## OverallQual 1.353e-01  1.372e-02   9.857  < 2e-16 ***
## BsmtFinSF1  1.955e-04  3.400e-05   5.751 5.06e-08 ***
## GrLivArea   2.032e-04  3.420e-05   5.942 2.00e-08 ***
## GarageCars  1.639e-01  2.173e-02   7.543 4.63e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1698 on 145 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8281
## F-statistic: 180.4 on 4 and 145 DF,  p-value: < 2.2e-16
```

```
anova(model4)
```

```
## Analysis of Variance Table
##
## Response: logprice
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## OverallQual   1 16.0200 16.0200 555.800 < 2.2e-16 ***
## BsmtFinSF1    1  1.2622  1.2622  43.792 6.590e-10 ***
## GrLivArea     1  1.8778  1.8778  65.150 2.423e-13 ***
## GarageCars    1  1.6400  1.6400  56.899 4.626e-12 ***
## Residuals   145  4.1794  0.0288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

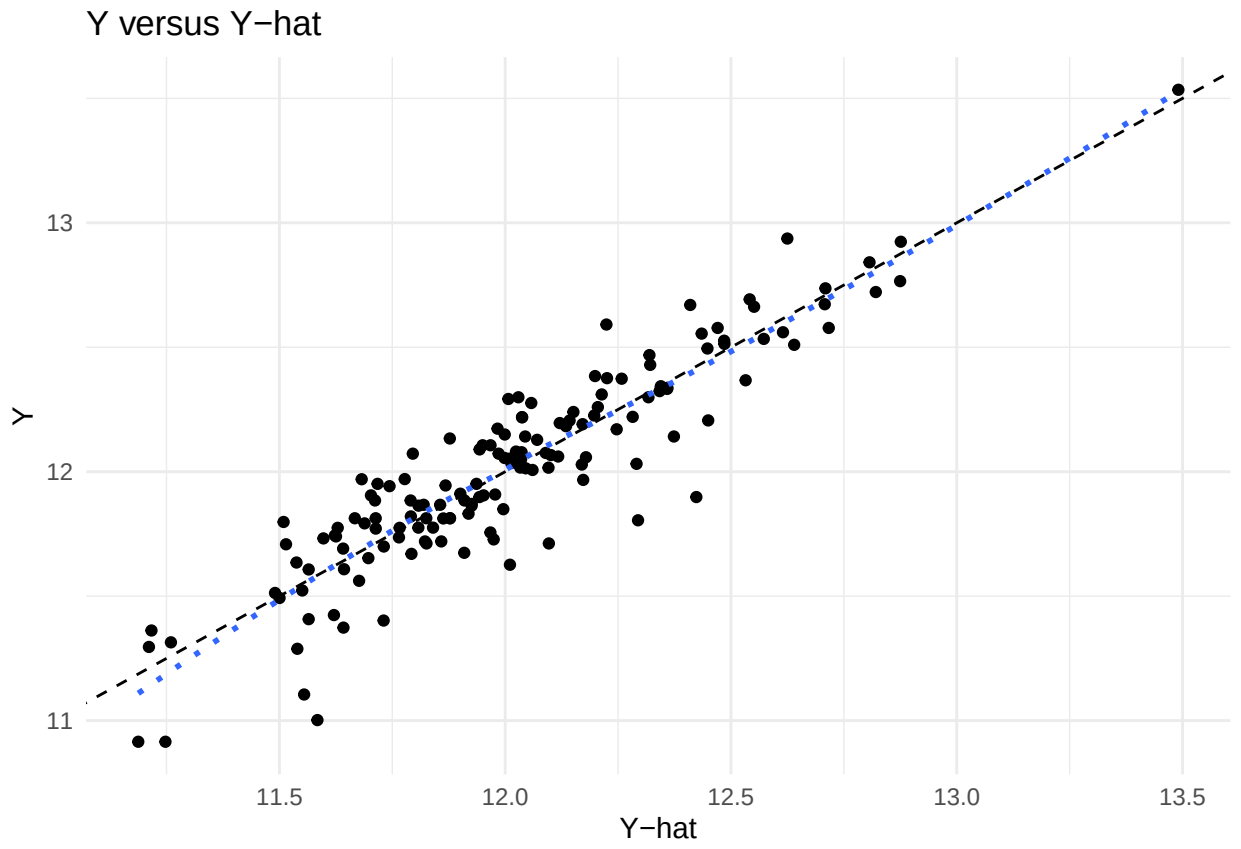
```
model4_F_test = lm(logprice ~ OverallQual + BsmtFinSF1 ,
            data = test)
anova(model4,model4_F_test)
```

```
## Analysis of Variance Table
##
```

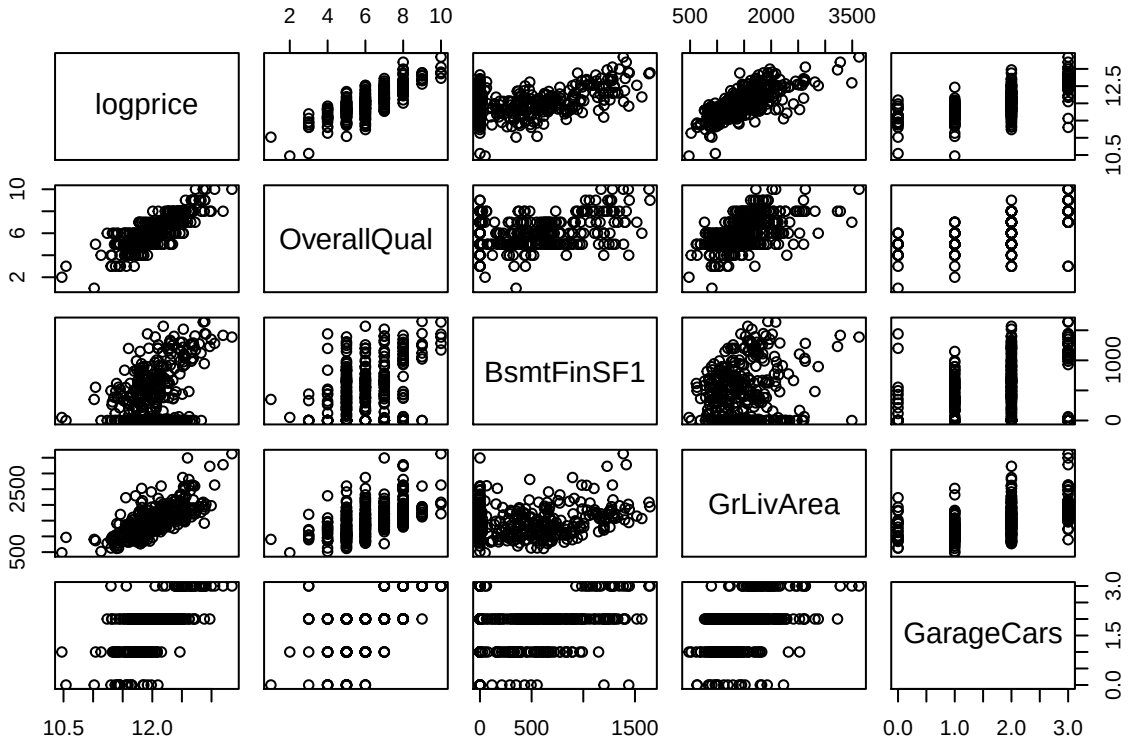
```
## Model 1: logprice ~ OverallQual + BsmtFinSF1 + GrLivArea + GarageCars
## Model 2: logprice ~ OverallQual + BsmtFinSF1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    145 4.1794
## 2    147 7.6973 -2   -3.5179 61.025 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Check condition 1 and 2 for test model
 r <- resid(model4)
comparison <- data.frame(Y = test$logprice, Y_hat = fitted(model4))
ggplot(comparison, aes(x = Y_hat, y = Y)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
  geom_smooth(method = "loess", se = FALSE, linetype = "dotted") +
  labs(x = "Y-hat", y = "Y", title = "Y versus Y-hat") +
  theme_minimal()
```

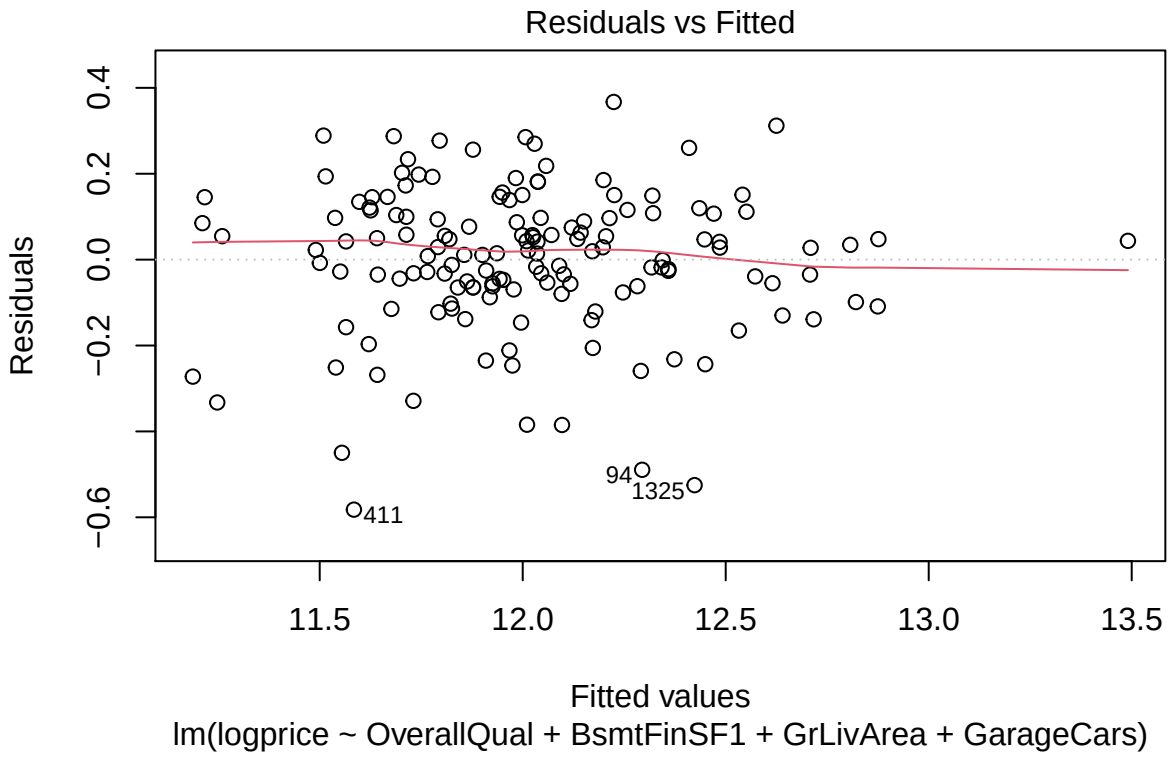
```
## `geom_smooth()` using formula = 'y ~ x'
```



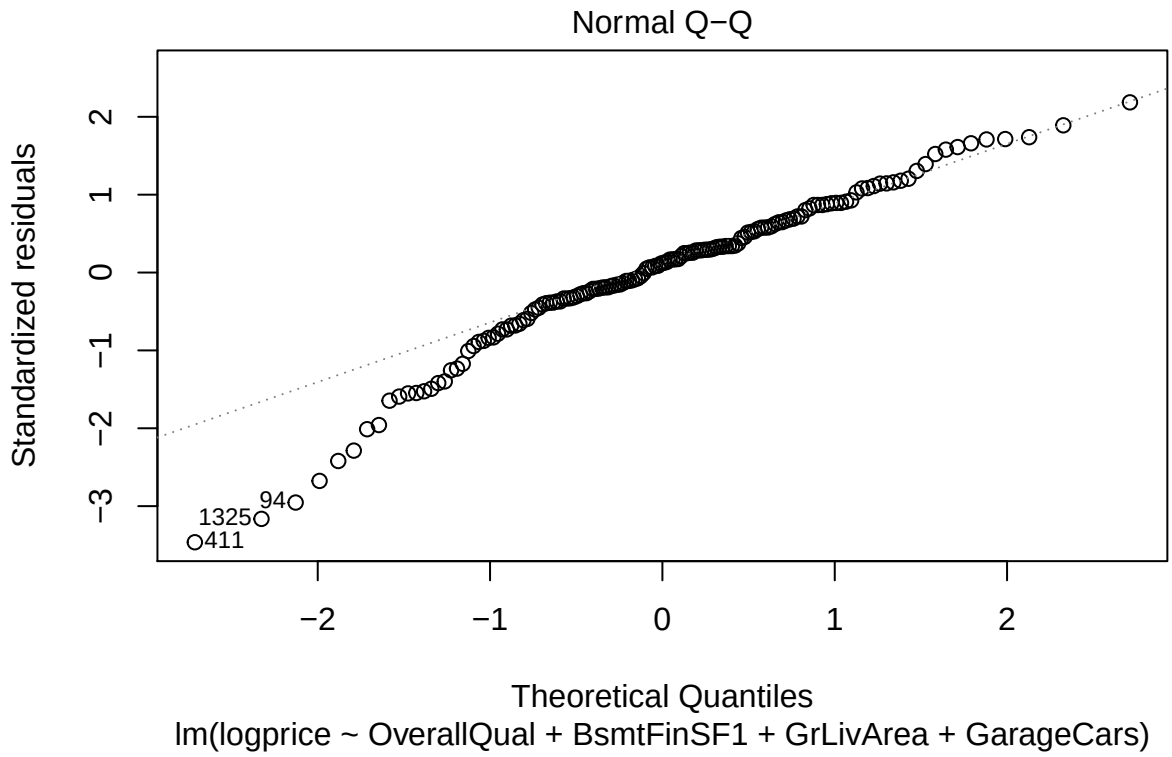

```
data2 = data.frame(logprice, OverallQual, BsmtFinSF1,GrLivArea,GarageCars)
pairs( data2 )
```



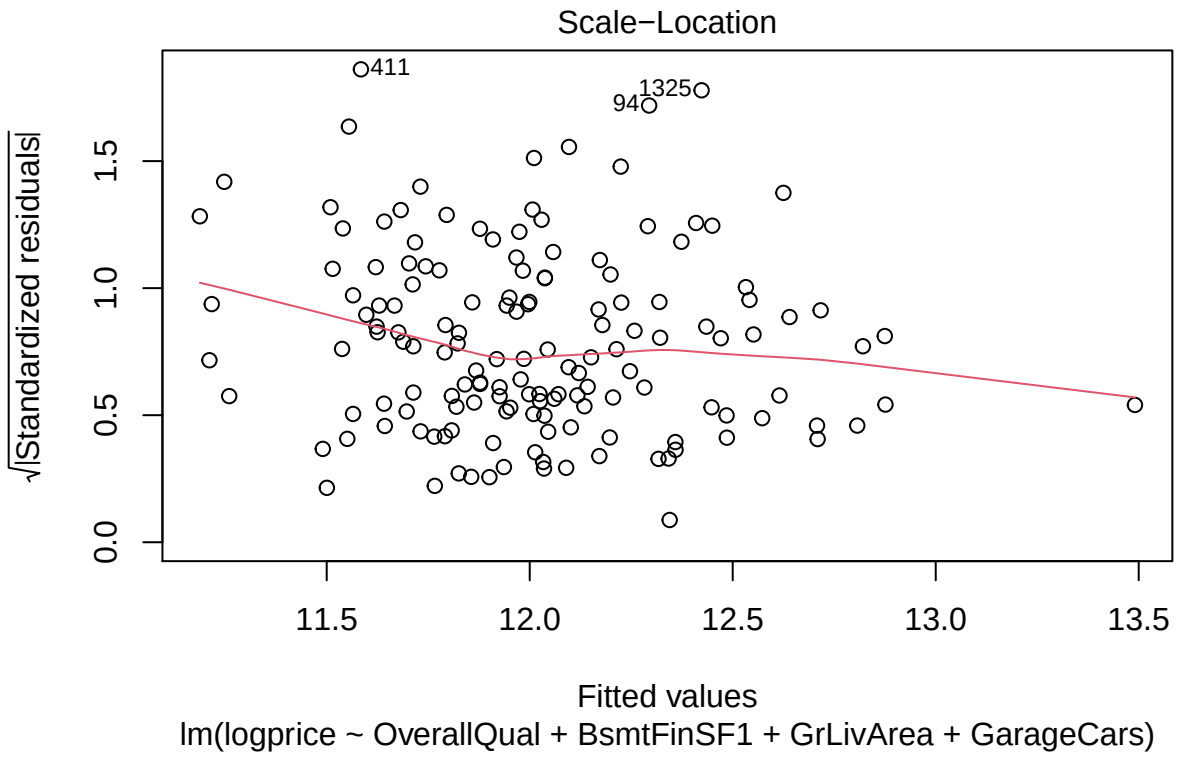

```
plot(model4,1)
```



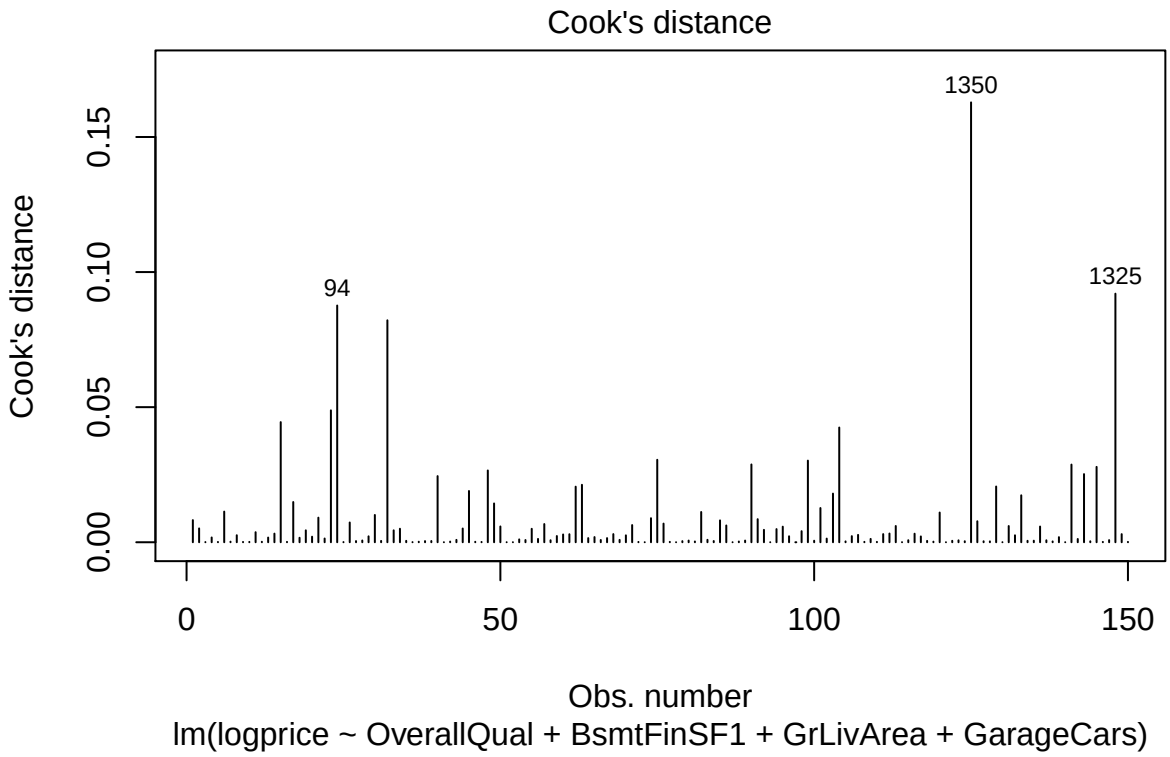

```
plot(model4,2)
```



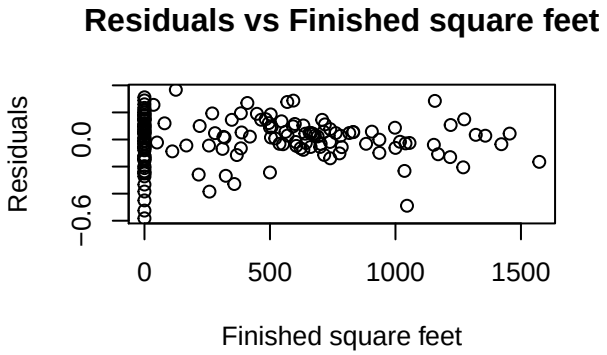

```
plot(model4,3)
```

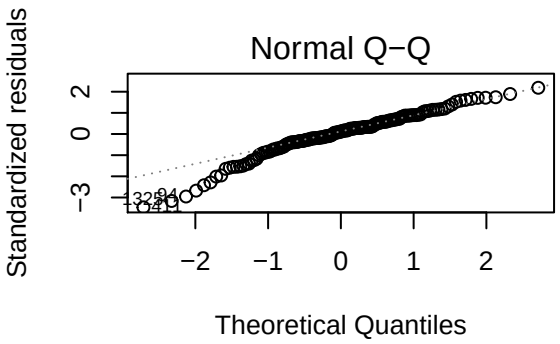


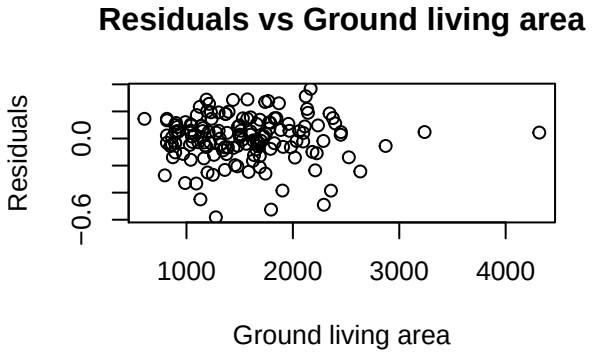

```
plot(model4,4)
```

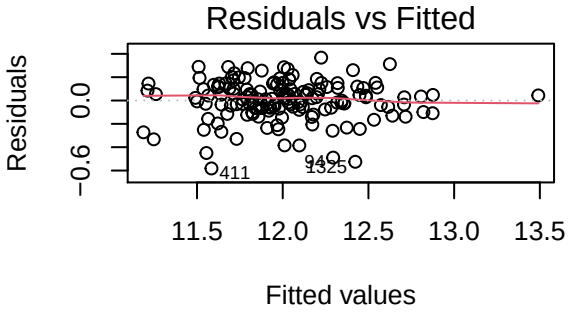



```
par(mfrow=c(2,2))
r <- model4$residuals
fit=model4$fitted.values
plot(r ~ test$BsmtFinSF1, xlab = "Finished square feet", ylab = "Residuals", main = "Residuals vs Finis
plot(model4,2)
plot(r ~ test$GrLivArea, xlab = "Ground living area", ylab = "Residuals", main = "Residuals vs Ground l
plot(model4,1)
```










```
r_test <- model4$residuals
#find outlier
out <- which(r_test > 2 | r_test < -2)
out
```

```
## named integer(0)
```

```
h <- hatvalues(model4)
threshold <- 2 * (length(model4$coefficients)/nrow(test))
w <- which(h > threshold)
test[w,]
```

```
##      SalePrice OverallCond OverallQual BsmtFinSF1 GrLivArea GarageCars FullBath
## 252     235000           5           8       1573      1625          2        2
## 1389    377500           5           9       1320      1746          3        2
## 1351    200000           5           5        500      2634          4        2
## 35      277500           5           9       1153      1561          2        2
## 1354    410000           5           8        816      3238          3        2
## 1341    123000           5           4          0       872          4        1
## 692     755000           6          10       1455      4316          3        3
## 1350    122000           5           8        259      2358          0        2
##      MoSold BedroomAbvGr KitchenAbvGr logprice
## 252      12            2            1 12.36734
## 1389     10            3            1 12.84133
## 1351      8            6            2 12.20607
## 35        8            2            1 12.53358
## 1354      3            4            1 12.92391
```

```
## 1341      6            3            1 11.71994
## 692       1            4            1 13.53447
## 1350     12            4            1 11.71178
```

```
vif(model4)
```

```
## OverallQual  BsmtFinSF1   GrLivArea  GarageCars
##    1.796140    1.057714    1.745126    1.400050
```

```
summary(test[,c(2:11)])
```

```
##   OverallCond     OverallQual       BsmtFinSF1       GrLivArea
##  Min.   :3.000   Min.   : 4.000   Min.   :   0.0   Min.   : 605
##  1st Qu.:5.000   1st Qu.: 5.000   1st Qu.:   0.0   1st Qu.:1144
##  Median :5.000   Median : 6.000   Median : 384.0   Median :1491
##  Mean   :5.527   Mean   : 6.167   Mean   : 422.4   Mean   :1537
##  3rd Qu.:6.000   3rd Qu.: 7.000   3rd Qu.: 701.5   3rd Qu.:1794
##  Max.   :9.000   Max.   :10.000   Max.   :1573.0   Max.   :4316
##    GarageCars       FullBath        MoSold        BedroomAbvGr  KitchenAbvGr
##  Min.   :0.000   Min.   :1.00   Min.   : 1.000   Min.   :1     Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.00   1st Qu.: 5.000   1st Qu.:3     1st Qu.:1.000
##  Median :2.000   Median :2.00   Median : 6.000   Median :3     Median :1.000
##  Mean   :1.827   Mean   :1.58   Mean   : 6.307   Mean   :3     Mean   :1.053
##  3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.: 8.000   3rd Qu.:3     3rd Qu.:1.000
##  Max.   :4.000   Max.   :3.00   Max.   :12.000   Max.   :6     Max.   :2.000
##     logprice
##  Min.   :10.92
##  1st Qu.:11.77
##  Median :12.01
##  Mean   :12.01
##  3rd Qu.:12.24
##  Max.   :13.53
```