STAT525 Car Project Report

Marco Lopez

2024-12-15

Used Car Price Data Regression Modeling

Group members: Lucas, Minh, Grace, Marco, Dang, Cuong

Introduction

Data was collected from Kelley Blue Book for several hundred 2005 used GM cars. There are some significant factors that are likely to affect the car prices such as mileage, make, model, liter, cylinder, type, etc. All cars in this dataset were less than one year old when priced and considered to be in excellent condition.

The main goal of this project is to determine how the twelve characteristics in combination contribute to the price of a 2005 used General Motors car. The end result of the project will be to create a regression model with some or all of the twelve predictor characteristics.

Variables

The variables included in the data set are as follows:

Price (suggested retail price of the used 2005 GM car in excellent condition)

Mileage (number of miles the car has been driven)

Make (manufacturer of the car such as Saturn, Pontiac, and Chevrolet)

Model (specific models for each car manufacturer such as Ion, Vibe, Cavalier)

Trim (specific type of car model such as SE Sedan 4D, Quad Coupe 2D)

Type (body type such as sedan, coupe, etc.)

Cylinder (number of cylinders in the engine)

Liter (a more specific measure of engine size)

Doors (number of doors)

Cruise (indicator variable representing whether the car has cruise control where 1 = cruise)

Sound (indicator variable representing whether the car has upgraded speakers where 1 = upgraded)

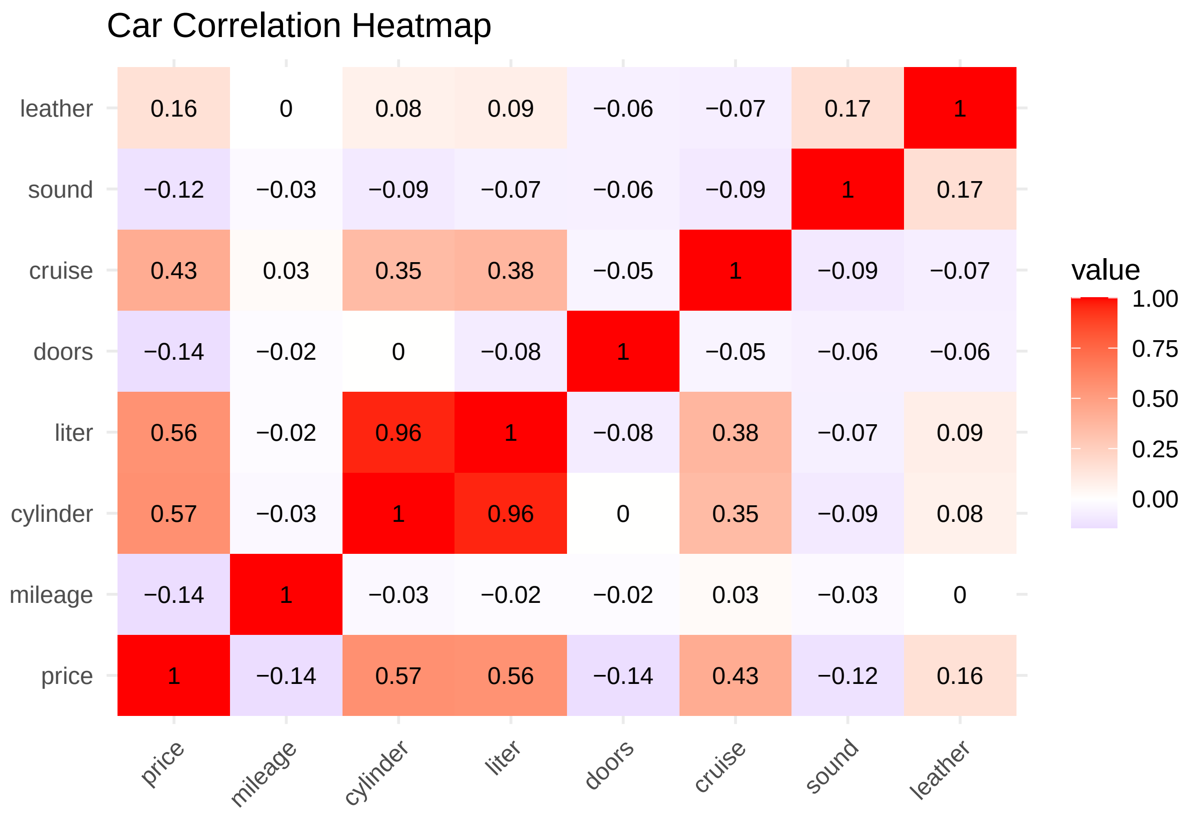
Leather (indicator variable representing whether the car has leather seats where 1 = leather)

Make, Model, Trim, and Type are the 4 categorical variables. Price, Mileage, Cylinder, Liter, Doors, Cruise, Sound, and Leather are the 8 numerical variables. The variable that we are trying to predict in the regression model is price.

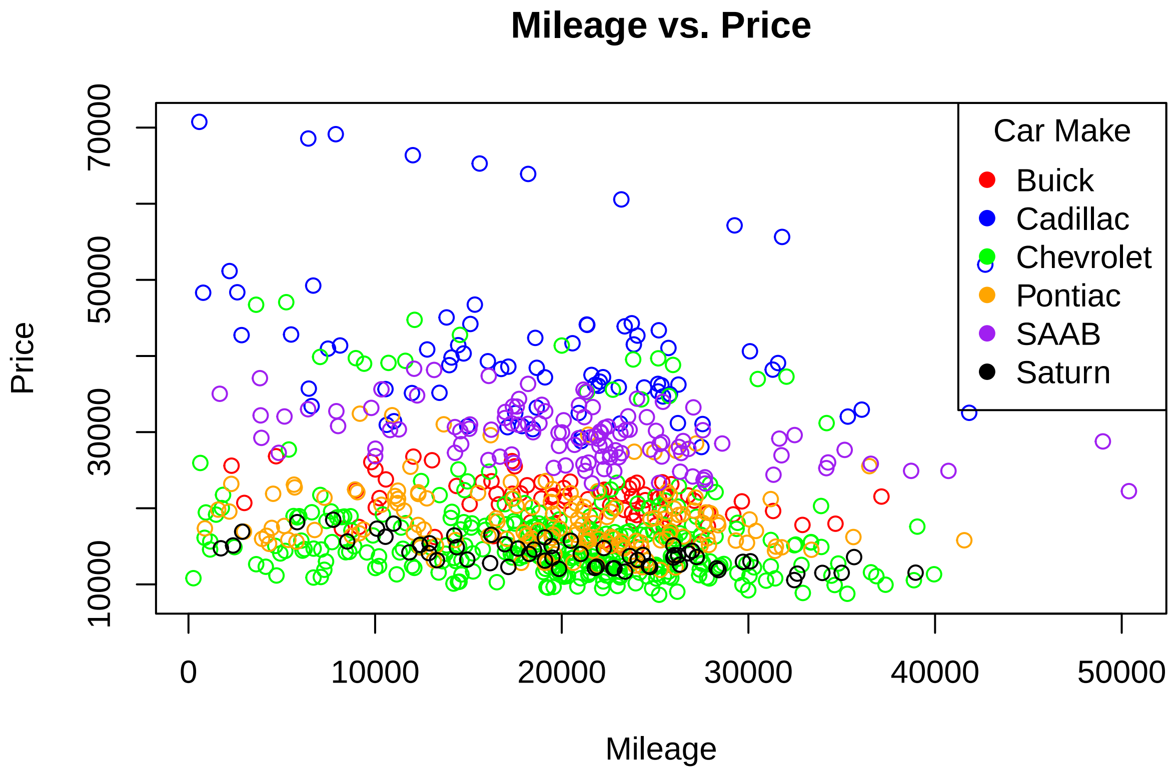
Methods/Results

Exploratory data analysis

In our exploratory data analysis, we included all 8 numerical variables, excluding the 4 categorical variables.



Based on the exploratory data analysis, we saw that mileage, doors, and sound had a slight negative correlation with price, leather had a slight positive correlation with price, and cylinder, liter, and cruise had a strong correlation with price. Additionally, we saw that cylinder and liter were highly correlated with each other (0.96).



One other thing noted in the exploratory analysis was that the Car Make (one of the categorical variables) was highly correlated with price as well. In the above plot, we notice that Cadillacs are always among the most expensive cars and Chevrolets are among the least expensive cars for example.

Model Building

```
## [1] 161  33 129   1  97 225
```

```
## [1] "Mileage"  "Make"     "Cylinder" "Liter"    "Doors"    "Sound"
```

```
## [1] 0.949475
```

In our model building, the data was split into two halves, with one half being the training set and the other half being the test set. In order to stabilize the variance and improve model assumptions, a log transformation was applied to the response variable price.

In order to determine the best combination of predictor variables, we created a linear regression model for every possible combination of predictor variables and sorted by the highest R^2 value.

After running this, we found that the best model was the one that included Mileage, Make, Cylinder, Liter, Doors, and Sound variables. This model resulted in an adjusted R^2 value of 0.9495, indicating strong predictive power.

```
## [1] 0.9089741
```

Using the test data set, we validate the results obtained from the training data set. The results indicate an R-squared of 0.909, meaning that there was a slight drop in predictive power but the R-squared of 0.909 is still quite strong.

Model Selection using the F-Test

Full model:

```
## 
## Call:
## lm(formula = logy ~ ., data = X)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.12160 -0.03077  0.00213  0.02458  0.15101 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    4.060e+00  1.719e-02 236.201  < 2e-16 ***
## Mileage       -3.510e-06  2.133e-07 -16.458  < 2e-16 ***
## MakeCadillac   2.089e-01  1.053e-02  19.835  < 2e-16 ***
## MakeChevrolet -6.589e-02  7.051e-03  -9.344  < 2e-16 ***
## MakePontiac   -3.610e-02  7.224e-03  -4.997 7.19e-07 ***
## MakeSAAB       2.856e-01  8.833e-03  32.330  < 2e-16 ***
## MakeSaturn    -5.943e-02  9.524e-03  -6.241 7.10e-10 ***
## Cylinder6     -1.283e-02  1.176e-02  -1.091    0.276    
## Cylinder8     -2.256e-02  2.358e-02  -0.957    0.339    
## Liter          1.020e-01  6.575e-03  15.514  < 2e-16 ***
## Doors4        -3.683e-02  4.532e-03  -8.126 1.71e-15 ***
## Cruise1        1.315e-03  4.991e-03   0.263    0.792    
## Sound1        -7.071e-04  3.972e-03  -0.178    0.859    
## Leather1       2.331e-03  4.318e-03   0.540    0.589
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.04934 on 790 degrees of freedom
## Multiple R-squared:  0.9245, Adjusted R-squared:  0.9232
## F-statistic:   744 on 13 and 790 DF,  p-value: < 2.2e-16
```

Trained model:

```
##
## Call:
## lm(formula = log.y.train ~ ., data = X.train.subset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.105212 -0.034881  0.002425  0.029860  0.141580
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.985e+00  2.157e-02 184.751  < 2e-16 ***
## Mileage       -3.537e-06  2.896e-07 -12.215  < 2e-16 ***
## MakeCadillac   2.284e-01  1.177e-02  19.405  < 2e-16 ***
## MakeChevrolet -3.169e-02  8.796e-03  -3.603 0.000355 ***
## Cylinder6      1.301e-02  1.760e-02   0.739 0.460376
## Cylinder8     -3.549e-02  3.264e-02  -1.087 0.277604
## Liter          1.171e-01  8.440e-03  13.871  < 2e-16 ***
## Doors4        -4.860e-02  7.262e-03  -6.692 7.62e-11 ***
## Sound1         1.251e-02  5.873e-03   2.131 0.033730 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.04819 on 393 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9495
## F-statistic:   943 on 8 and 393 DF,  p-value: < 2.2e-16
```

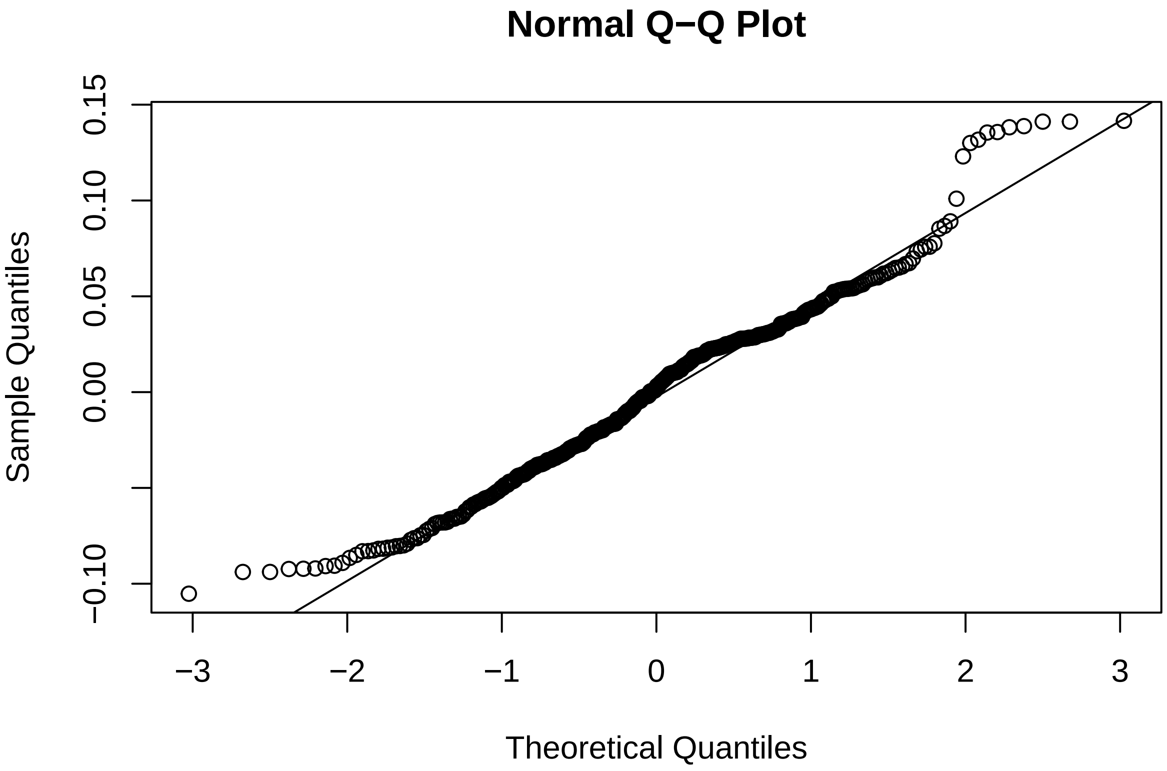
F-Test:

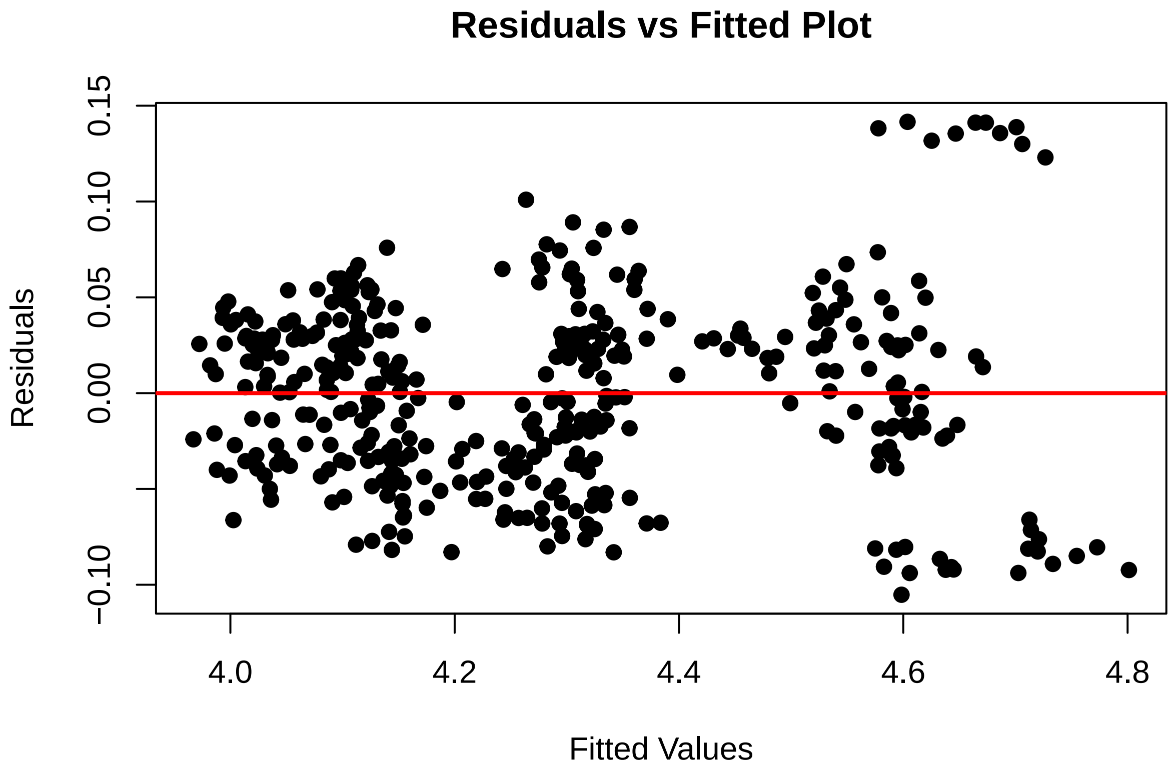
```
## F-stat: 1.045833 > F-crt: > 1.151745 => FAIL to reject the null, conclude H0
```

We run an F-test with the null hypothesis H_0 that the β's not in the small model are 0 and the alternative hypothesis H_a that at least one of the β's in the small model are non-zero.

After running the F-test, we concluded the null hypothesis H_0 and we found that the best model was the small model that included Mileage, Make (Cadillac and Chevrolet), Cylinder, Liter, Doors, and Sound variables.

Residual Diagnostics and Model Assumptions





Based on the the Normal Q-Q plot and Residuals vs. Fitted values plot, we concluded that the residuals followed nearly a straight line although we did notice slight deviations at the tails. With the result of the Residuals vs. Fitted values plot, we concluded that the residuals are primarily scattered around zero so we can assume Homoscedasticity.

Based on the two diagnostic plots (Q-Q plot and residuals vs. fitted values) we assume normality of the residuals and that the assumptions of linear regression are reasonably satisfied.

Conclusion

In conclusion, the best model was identified by choosing the one with the biggest adjusted R^2. Based on the identified best model, we conducted an F-test to confirm that the variables we dropped from the full model are insignificant and equal to zero. Finally, we looked at the Normal Q-Q plots and Residuals vs. Fitted plot to check the normality of the residuals and assumptions of linear regression.

The final model that was decided was:

$$\log(\text{Price}) = \beta_0 + \beta_1 \text{Mileage} + \beta_2 \text{Make(Cadillac)} + \beta_3 \text{Make(Chevrolet)}$$

$$+\beta_4 Cylinder(6) + \beta_5\text{Cylinder}(8) + \beta_6\text{Liter} + \beta_7\text{Doors}(4) + \beta_8\text{Sound}(1) + \epsilon$$

Appendix

R Code:

```r
library(readxl)
library(ggplot2)
library(reshape2)

car <- read_excel('C:/Users/malop/Documents/STAT525_CAR_PROJECT/car.xls')

price <- car$Price
mileage <- car$Mileage
cylinder <- car$Cylinder
liter <- car$Liter
doors <- car$Doors
cruise <- car$Cruise
sound <- car$Sound
leather <- car$Leather

model.car <- data.frame(price, mileage, cylinder, liter, doors, cruise, sound, leather)

cor_matrix <- cor(model.car)

cor_melted <- melt(cor_matrix)

ggplot(cor_melted, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(midpoint = 0, low = "blue", high = "red", mid = "white") +
  geom_text(aes(label = round(value, 2)), color = "black", size = 3) +  # Add numbers
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Car Correlation Heatmap", x = "", y = "")
```

```
make_colors <- c("Buick" = "red", "Cadillac" = "blue", "Chevrolet" = "green", "Pontiac" = "orange",
                 "SAAB" = "purple", "Saturn" = "black")
point_colors <- make_colors[car$Make]

plot(mileage, price,
     main = "Mileage vs. Price",
     xlab = "Mileage",
     ylab = "Price",
     col = point_colors)

legend("topright", legend = names(make_colors),
       col = make_colors,
       pch = 19,
       title = "Car Make")

library(readxl)

car.data <- read_excel('C:/Users/malop/Documents/STAT525_CAR_PROJECT/car.xls')

attach(car.data)

factor_col <- c("Make", "Model", "Trim", "Type", "Cylinder", "Doors", "Cruise", "Sound", "Leather") # s
car.data[factor_col] <- lapply(car.data[factor_col], factor) # convert certain col listed above to fact

X <- car.data[, setdiff(names(car.data), c("Price", "Model", "Trim", "Type"))]

y <- car.data$Price

logy <- log(y)/log(10)

log.y.train <- as.vector(logy[1:(nrow(car.data)/2)])

X.train <- X[1:(nrow(car.data)/2),]

all.subsets <- expand.grid(rep(list(c(TRUE,FALSE)), ncol(X))) #All 2^8 rows together give all subsets o
all.subsets <- as.matrix(all.subsets) # data casting--see above comment
```

```r
all.subsets <- all.subsets[-(2^(ncol(X))),] # drop the model consisting of no predictors (FALSE,FALSE)

adj.r.squared <- c()
predictors <- colnames(X.train)
for(i in 1:nrow(all.subsets)) {
  # Extract the logical vector for the current subset
  subset_logical <- all.subsets[i, ]
  # Use logical indexing to get the predictor names
  subset <- predictors[subset_logical]  # This will be a vector of column names
  #  subset <- predictors[all.subsets[i,]]
  X.train.subset <- X.train[, subset]
  adj.r.squared <- c(adj.r.squared, summary(lm(log.y.train ~ ., data = X.train.subset))$adj.r.squared)
}

ordered.adj.r.sq <- order(adj.r.squared,decreasing=TRUE)

head(ordered.adj.r.sq)

predictors[all.subsets[ordered.adj.r.sq[1],]]

subset <- predictors[all.subsets[ordered.adj.r.sq[1],]]
X.train.subset <- X.train[, subset]

print( summary(lm(log.y.train ~ ., data = X.train.subset))$adj.r.squared )

logy.test <- logy[((nrow(car.data)/2)+1):nrow(car.data)]
X.test <- X[((nrow(car.data)/2)+1):nrow(car.data),]
X.test.subset <- X.test[, subset]

lm.test.subset <- lm(logy.test ~ ., data = X.test.subset)

summary(lm.test.subset)$adj.r.squared

lm.full <- lm(logy ~ ., data = X)
summary(lm.full)
```

```r
lm.trained <- lm(log.y.train ~ ., data = X.train.subset)
summary(lm.trained)

alpha <- 0.05 # 5% significance level, 95% confidence
dfn <- (anova(lm.full)[9,1] - anova(lm.trained)[7,1] )
dfd <- anova(lm.full)[9,1]
num.temp <- (anova(lm.full)[9,2] - anova(lm.trained)[7,2] ) / dfn
denom.temp <- anova(lm.full)[9,2]  /  dfd
F.stat <- num.temp/denom.temp
F.crt <- qf((1-alpha), dfn, dfd)
if (F.stat > F.crt) {
  cat("F-stat:", F.stat, "<", "F-crt:", ">", F.crt, "=> Reject the null, conclude Ha")
} else{
  cat("F-stat:", F.stat, ">", "F-crt:", ">", F.crt, "=> FAIL to reject the null, conclude H0")
}

epsilon.hat <- resid(lm.trained)

qqnorm(epsilon.hat)
qqline(epsilon.hat)

plot(lm.trained$fitted.values, residuals(lm.trained),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Plot",
     pch = 19, col = "black")
abline(h = 0, col = "red", lwd = 2)
```