1. Let’s take MSRP ($) as the response variable and consider Wheelbase (in), Displacement (cu in), Bore (in), and Clearance (in) as potential predictors. Use scatterplots to see which variables can be appropriately used as predictors in simple linear regression.

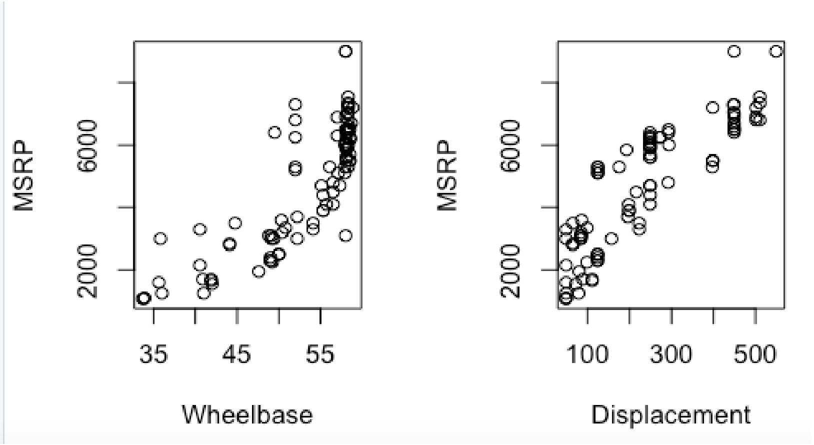
Answer

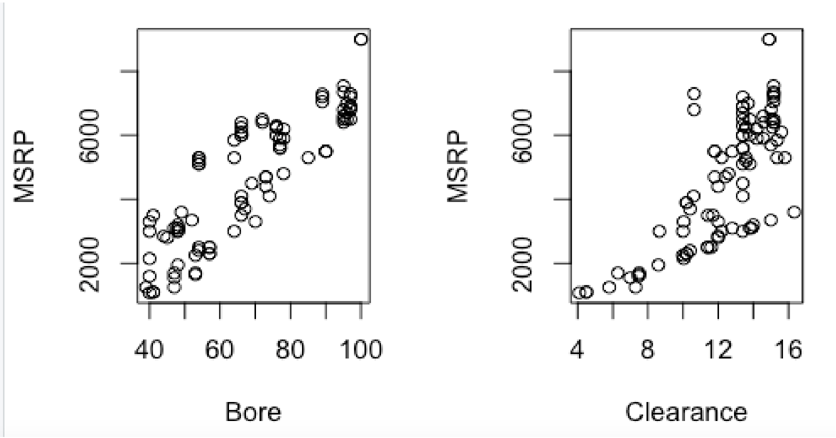
All of the scatterplots are linear except Wheelbase vs. MSRP and Clearance vs. MSRP, in which there's a curve in the Wheelbase plot and a fan shape in the Clearance plot, respectively. Therefore, these 2 variables can't be used as predictors in the simple linear regression unless transformations are performed. However, because the Displacement and Bore plots demonstrate linearity, these 2 variables can be used as predictors.

R code

```
par(mfrow = c(1,2))
plot(motor$Wheelbase, motor$MSRP, xlab = 'Wheelbase', ylab = 'MSRP')
plot(motor$Displacement, motor$MSRP, xlab = 'Displacement', ylab = 'MSRP')
plot(motor$Bore, motor$MSRP, xlab = 'Bore', ylab = 'MSRP')
plot(motor$Clearance, motor$MSRP, xlab = 'Clearance', ylab = 'MSRP')
```

R output



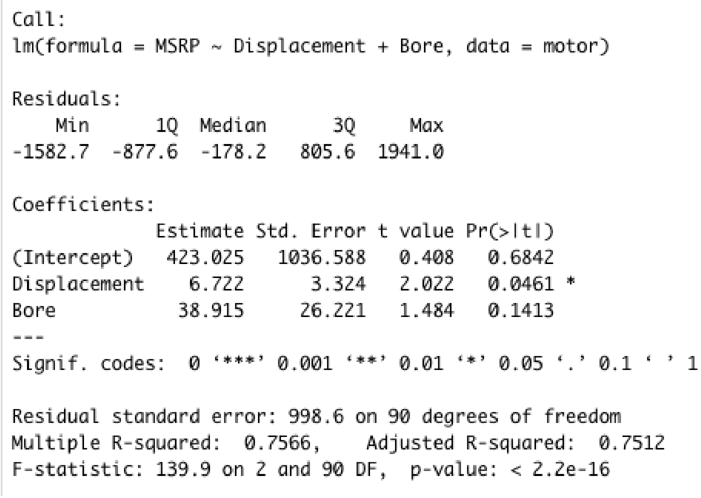


2. Build a multiple regression model for MSRP using Displacement and Bore as predictors. Write down the fitted model. Report R2 and adjusted R2 . Interpret the coefficients for Displacement and Bore.

Answer
The R^2 is 75.66%, and the adjusted R^2 is 75.12%.
Fitted model: MSRP hat = 423.025 + 6.722*Displacement + 38.915*Bore
Interpretation of Displacement: After accounting for Bore, the MSRP is expected to increase by $6.72 when Displacement increases by a unit.
Interpretation of Bore: After accounting for Displacement, the MSRP is expected to increase by $38.915 when Bore increases by a unit.

R code
```
reg1 <- lm(MSRP ~ Displacement+Bore, data=motor)
summary(reg1)
```

R output



```
Call:
lm(formula = MSRP ~ Displacement + Bore, data = motor)

Residuals:
    Min      1Q  Median      3Q     Max 
-1582.7  -877.6  -178.2   805.6  1941.0 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)   423.025   1036.588   0.408   0.6842  
Displacement    6.722      3.324   2.022   0.0461 *
Bore           38.915     26.221   1.484   0.1413  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 998.6 on 90 degrees of freedom
Multiple R-squared:  0.7566,	Adjusted R-squared:  0.7512 
F-statistic: 139.9 on 2 and 90 DF,  p-value: < 2.2e-16
```

3. Check the model you fitted in the previous question to see if it satisfies the assumptions as required in multiple regression.

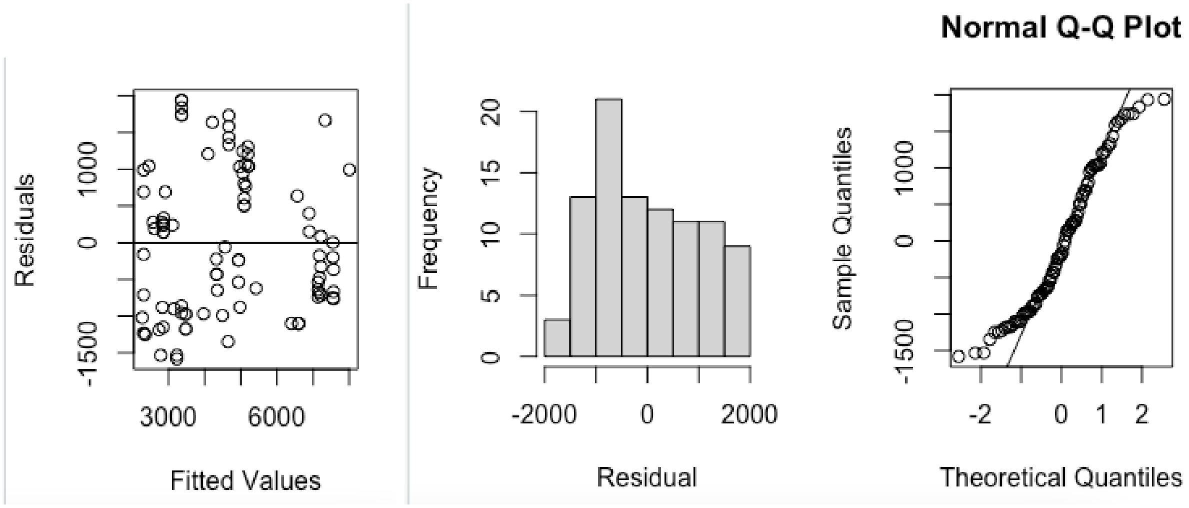
Answer
Linearity Assumption: The residual plot doesn't have any pattern or bends, so this assumption is satisfied.
Independence Assumption: The data is collected from a random sample, so this assumption is also satisfied.
Equal Variance Assumption: The plot demonstrates scatter and equal spread about the line y=0, so this assumption is satisfied.

Normal Assumption: The histogram shows a very subtle right skew and the Q-Q plot isn't very linear but because this sample size is large, the interpretation of this assumption can be interpreted more leniently.

R code

```
plot(reg1$fitted.values, reg1$residuals, xlab = 'Fitted Values', ylab = 'Residuals')
abline(0, 0)
hist(imod$residuals, main = '', xlab = 'Residual')
qqnorm(imod$residuals)
qqline(imod$residuals)
```

R output



4. Conduct a test to see if the fitted multiple regression model is statistically useful. If useful, find the predictors that make significant contributions to the MSRP in the model. Explain.

Answer

From the regression summary previously computed before, the F-test value is 140 and the p-value is less than 0.05, so there is at least one useful predictor, thereby rejecting the null hypothesis and accepting the alternative hypothesis, which is that there is at least one useful predictor. The predictor that makes a significant contribution to MSRP in the model is Displacement, because its p-value of 0.0461 is less than 0.05. On the other hand, Bore has a p-value of 0.1413, which is larger than 0.05. Therefore, Bore isn’t statistically significant with MSRP after accounting for Displacement, but it could be significant with MSRP when Displacement isn’t in the model.

R code

N/A

R output

N/A

5. Suppose we are not satisfied with the R2 given by the current model. Please propose a new multiple regression model in order to improve R2 . Compare the new model to the current one with respect to their R2 , coefficient estimates and hypothesis tests. Don’t forget to check assumptions of the new model for its validity. (Hint: We have two potential predictors Wheelbase and Clearance in the pool. Think about how to use them to improve the model.)

Answer

The new regression model I'll use is to add Clearance to the current model, and compare its summary output with the previous one.

As a result of the addition of Clearance, it helped explain an additional proportion of variation of MSRP, because the new R^2 and adjusted R^2 increased to 89.45% and 89.1%, respectively. The coefficient of Bore decreased from 39 to 9 approximately, the F-statistic increased from 140 to 251.6 which makes it more significant, and the coefficient of Displacement increased by approximately 1.

The p-value of Displacement is still statistically significant, but the p-value of Bore isn't statistically significant. Plus, its p-value increased from 0.1413 to 0.60082. Additionally, variable Clearance is statistically significant, because its p-value is < 2e-16. For the assumptions, the residual plot shows no patterns, so the Linearity and Equal Variance Assumptions are satisfied. For the Normal Condition, the histogram is normal and the Q-Q plot is linear. Therefore, this assumption is met as well. Finally, the Independence Assumption is met because the sample is randomized.

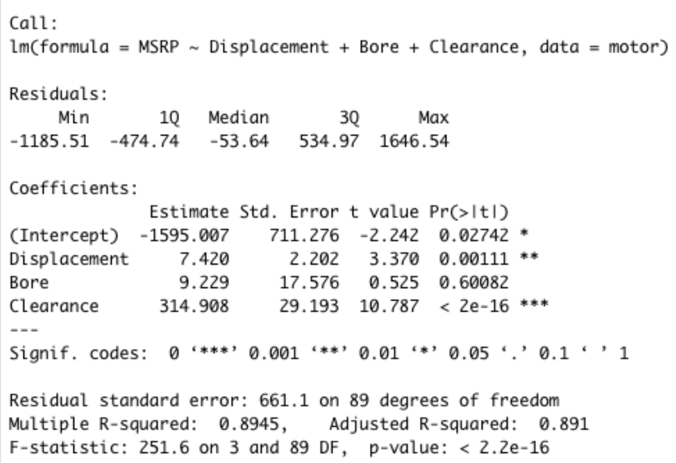
To further improve the model, I also added Wheelbase to the new regression model.

It seems that the R^2 value only increased by 0.0002 while adjusted R^2 decreased, and the F statistic decreased. The new variable Wheelhouse isn't significant, so this doesn't contribute to MSRP, after accounting for the other predictors. Moreover, the p-value of Bore increased by about 0.21 in this new model, compared to the 2nd model. Therefore, I would choose to use the second model, reg2.

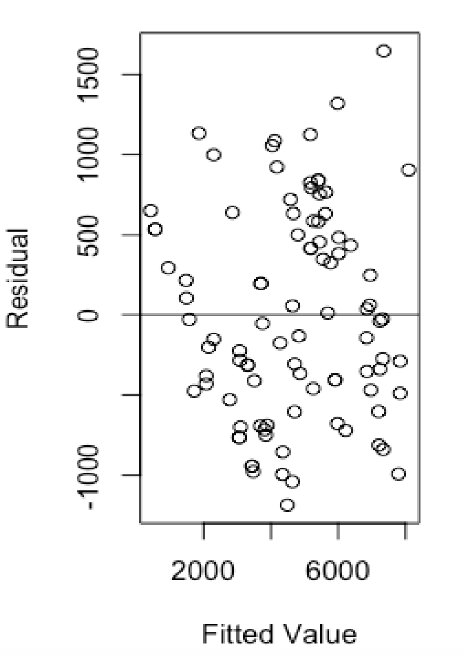
R code

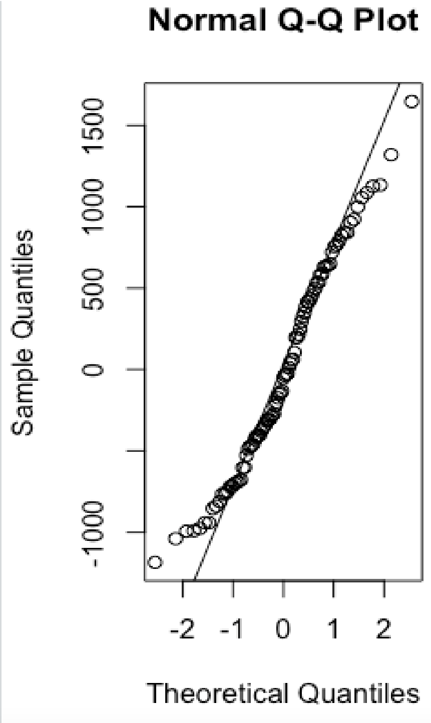
```
reg2 <- lm(MSRP ~ Displacement + Bore + Clearance, data = motor)
summary(reg2)
plot(reg2$fitted.values, reg2$residuals, xlab = 'Fitted Value', ylab = 'Residual')
abline(0, 0)
hist(reg2$residuals, main = '', xlab = 'Residual')
qqnorm(reg2$residuals)
qqline(reg2$residuals)
reg3 <- lm(MSRP ~ Displacement + Bore + Clearance + Wheelbase, data = motor)
summary(reg3)
```

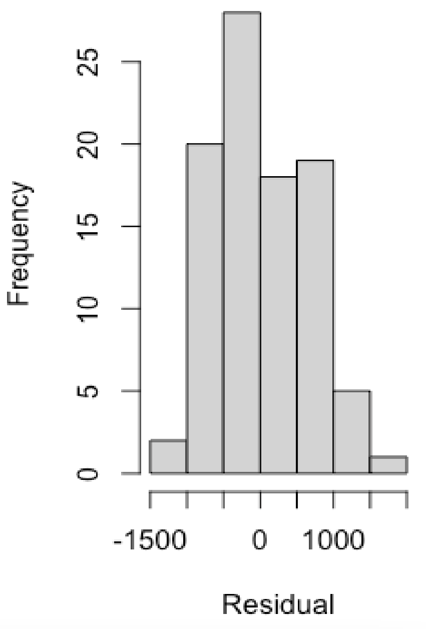
R output

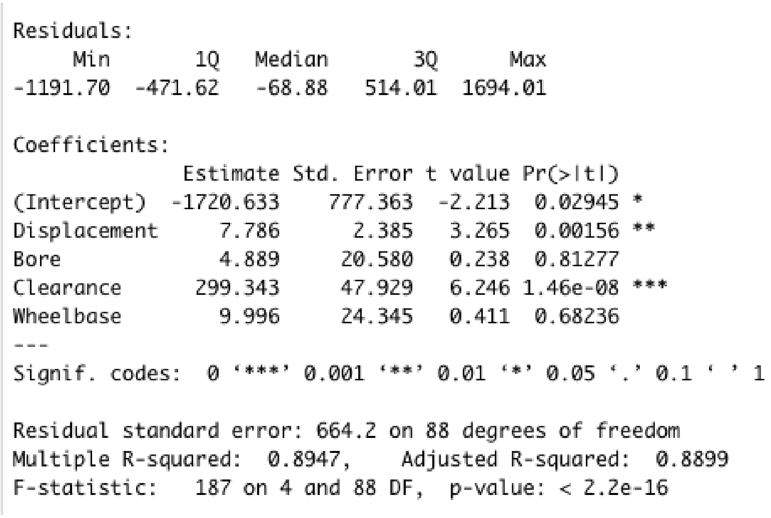


```
Call:
lm(formula = MSRP ~ Displacement + Bore + Clearance, data = motor)

Residuals:
     Min       1Q   Median       3Q      Max 
-1185.51  -474.74   -53.64   534.97  1646.54 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1595.007    711.276  -2.242  0.02742 *  
Displacement     7.420      2.202   3.370  0.00111 ** 
Bore             9.229     17.576   0.525  0.60082    
Clearance      314.908     29.193  10.787  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 661.1 on 89 degrees of freedom
Multiple R-squared:  0.8945,	Adjusted R-squared:  0.891 
F-statistic: 251.6 on 3 and 89 DF,  p-value: < 2.2e-16
```









```
Residuals:
     Min       1Q   Median       3Q      Max
-1191.70  -471.62   -68.88   514.01  1694.01

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1720.633    777.363  -2.213  0.02945 *
Displacement     7.786      2.385   3.265  0.00156 **
Bore             4.889     20.580   0.238  0.81277
Clearance      299.343     47.929   6.246 1.46e-08 ***
Wheelbase        9.996     24.345   0.411  0.68236
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 664.2 on 88 degrees of freedom
Multiple R-squared:  0.8947,    Adjusted R-squared:  0.8899
F-statistic:   187 on 4 and 88 DF,  p-value: < 2.2e-16
```

Appendix

```
par(mfrow = c(1,2))
plot(motor$Wheelbase, motor$MSRP, xlab = 'Wheelbase', ylab = 'MSRP')
plot(motor$Displacement, motor$MSRP, xlab = 'Displacement', ylab = 'MSRP')
plot(motor$Bore, motor$MSRP, xlab = 'Bore', ylab = 'MSRP')
plot(motor$Clearance, motor$MSRP, xlab = 'Clearance', ylab = 'MSRP')

reg1 <- lm(MSRP ~ Displacement+Bore, data=motor)
summary(reg1)

plot(reg1$fitted.values, reg1$residuals, xlab = 'Fitted Values', ylab = 'Residuals')
abline(0, 0)
hist(imod$residuals, main = '', xlab = 'Residual')
qqnorm(imod$residuals)
qqline(imod$residuals)

reg2 <- lm(MSRP ~ Displacement + Bore + Clearance, data = motor)
summary(reg2)
plot(reg2$fitted.values, reg2$residuals, xlab = 'Fitted Value', ylab = 'Residual')
abline(0, 0)
hist(reg2$residuals, main = '', xlab = 'Residual')
qqnorm(reg2$residuals)
qqline(reg2$residuals)
reg3 <- lm(MSRP ~ Displacement + Bore + Clearance + Wheelbase, data = motor)
summary(reg3)
```