extra homework

Haocheng Zhang

2023-03-23

```
## Load the dataset 'cars'
cars <- read.table(file='cars.txt', head=T)
```

```
## Q(1) # Remove every fifth observation for use as a test sample.
test_sample <- cars[seq(5, nrow(cars), by=5),]

# And the remaining data will be used as a training sample for futher use:
training_sample <- cars[-seq(5, nrow(cars), by=5),]
```

```
## Q(2) To perform an exploratory analysis, we first use 'summary()' function
#for an overall analysis.
summary(training_sample)
```

```
##      name                mpg             cyl             disp
##  Length:26          Min.   :10.40   Min.   :4.000   Min.   : 75.7
##  Class :character   1st Qu.:15.28   1st Qu.:4.000   1st Qu.:120.5
##  Mode  :character   Median :19.55   Median :6.000   Median :196.3
##                     Mean   :20.07   Mean   :6.077   Mean   :221.8
##                     3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:303.2
##                     Max.   :32.40   Max.   :8.000   Max.   :460.0
##        hp             drat             wt             qsec
##  Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50
##  1st Qu.: 95.5   1st Qu.:3.098   1st Qu.:2.504   1st Qu.:16.88
##  Median :111.5   Median :3.715   Median :3.203   Median :17.71
##  Mean   :145.2   Mean   :3.622   Mean   :3.168   Mean   :17.90
##  3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.570   3rd Qu.:18.90
##  Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90
##        vs               am              gear            carb
##  Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4615   Mean   :0.4231   Mean   :3.692   Mean   :2.731
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

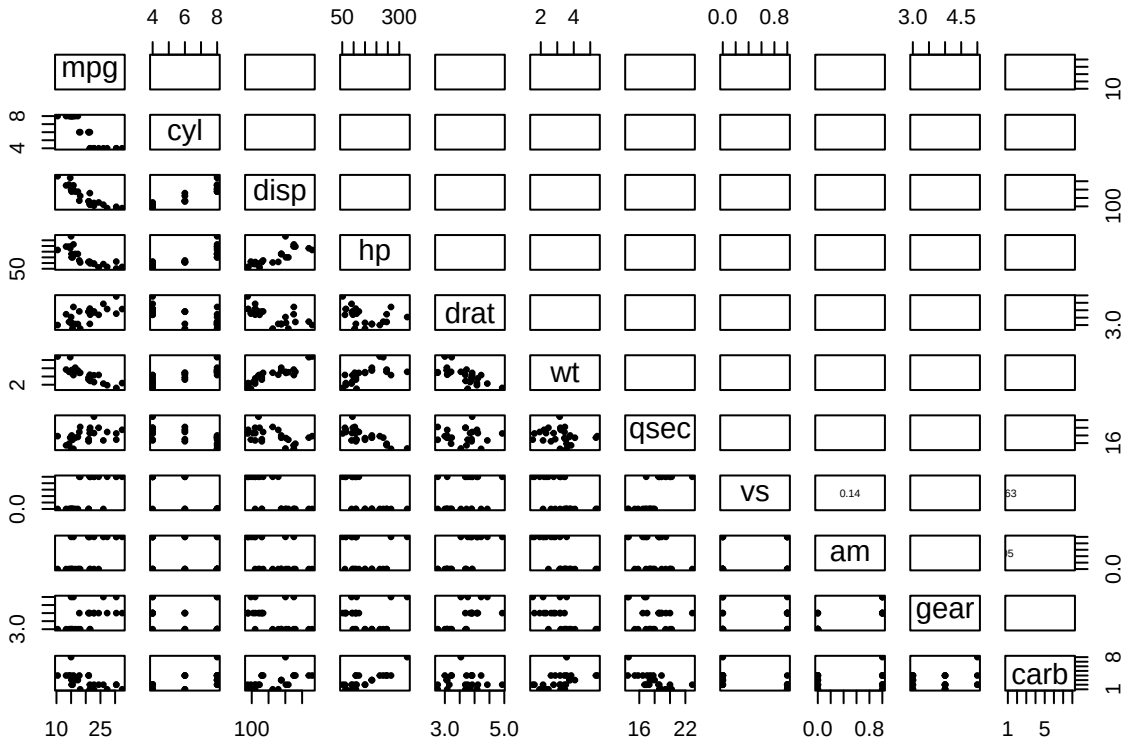
```
# 1.The average miles per gallon (mpg) for the training dataset is 20.07,
#with a minimum of 10.4 and a maximum of 32.4.
# 2.Most cars in the dataset have 6 or 8 cylinders, with a mean of 6.077.
# 3.The average displacement (disp) is 221.8, with a minimum of 75.7
#and a maximum of 460.
```

```
# 4.The average horsepower (hp) is 145.2, ranging from 52 to 335.
# 5.The average rear axle ratio (drat) is 3.622, ranging from 2.76 to 4.93.
# 6.The average weight (wt) is 3.168, with a minimum of 1.513
#and a maximum of 5.424.
# 7.The average quarter mile time (qsec) is 17.9, ranging from 14.5 to 22.9.
# 8.The vs, am, and gear variables are categorical with binary (0 or 1)
#or ordinal (3, 4, or 5) values.
# 9.The average number of carburetors (carb) is 2.731, ranging from 1 to 8.

# Then calculate the correlation matrix to understand the relationships
#between numeric variables(Exclude the 'name' column).
cor_matrix <- cor(training_sample[,-1])
print(cor_matrix)
```

```
##             mpg        cyl       disp          hp        drat         wt
## mpg   1.0000000 -0.8781485 -0.8879788 -0.77751430  0.66124336 -0.8592776
## cyl  -0.8781485  1.0000000  0.9089846  0.83050196 -0.65111857  0.7752513
## disp -0.8879788  0.9089846  1.0000000  0.82413517 -0.65054746  0.8889453
## hp   -0.7775143  0.8305020  0.8241352  1.00000000 -0.38749899  0.6489729
## drat  0.6612434 -0.6511186 -0.6505475 -0.38749899  1.00000000 -0.6853894
## wt   -0.8592776  0.7752513  0.8889453  0.64897286 -0.68538941  1.0000000
## qsec  0.4073352 -0.6007571 -0.4808054 -0.70529394  0.02358719 -0.1630227
## vs    0.6778071 -0.8129844 -0.7124245 -0.70404549  0.34999871 -0.5538998
## am    0.5861449 -0.4701960 -0.5430165 -0.20550363  0.72306718 -0.6731745
## gear  0.5113112 -0.4566258 -0.4801054 -0.09115853  0.70791536 -0.5747590
## carb -0.5623724  0.6122461  0.5332080  0.80794216 -0.07828427  0.4678306
##             qsec         vs          am        gear        carb
## mpg   0.40733515  0.6778071  0.58614488  0.51131118 -0.56237236
## cyl  -0.60075711 -0.8129844 -0.47019596 -0.45662576  0.61224609
## disp -0.48080537 -0.7124245 -0.54301645 -0.48010537  0.53320804
## hp   -0.70529394 -0.7040455 -0.20550363 -0.09115853  0.80794216
## drat  0.02358719  0.3499987  0.72306718  0.70791536 -0.07828427
## wt   -0.16302267 -0.5538998 -0.67317453 -0.57475897  0.46783065
## qsec  1.00000000  0.7423042 -0.28087102 -0.21560113 -0.66441318
## vs    0.74230418  1.0000000  0.14414999  0.18093672 -0.63230869
## am   -0.28087102  0.1441500  1.00000000  0.79668736  0.04804451
## gear -0.21560113  0.1809367  0.79668736  1.00000000  0.19999446
## carb -0.66441318 -0.6323087  0.04804451  0.19999446  1.00000000
```

```
# 1.mpg has a strong negative correlation with cyl (-0.878), disp (-0.888),
#and wt (-0.859), indicating that as the number of cylinders,
#engine displacement, and weight increase, the miles per gallon decreases.
# 2.mpg has a positive correlation with drat (0.661) and vs (0.678),
#suggesting that higher rear axle ratios and V/S values are associated
#with higher fuel efficiency.
# 3.cyl, disp, and wt are positively correlated with each other,
#indicating that cars with more cylinders, larger engine displacements,
#and heavier weights tend to have similar characteristics.
# 4.drat is positively correlated with am (0.723) and gear (0.708),
#indicating that cars with higher rear axle ratios tend to have manual
#transmissions and more forward gears.
# 5.hp has a strong positive correlation with carb (0.808), suggesting that
#cars with higher horsepower tend to have more carburetors.
```

```
# Also, we could Create scatter plots to visualize relationships between
#the response variable (mpg) and the predictor variables.
# Load the ggplot2 library for better visualizations
library(ggplot2)

# Create a scatterplot matrix
pairs(training_sample[,-1],
      lower.panel = function(x, y) {
        points(x, y, pch = 19, cex = 0.5)
      },
      upper.panel = function(x, y) {
        text(0.5, 0.5, round(cor(x, y), 2), cex = 0.5)
      },
      diag.panel = NULL)
```

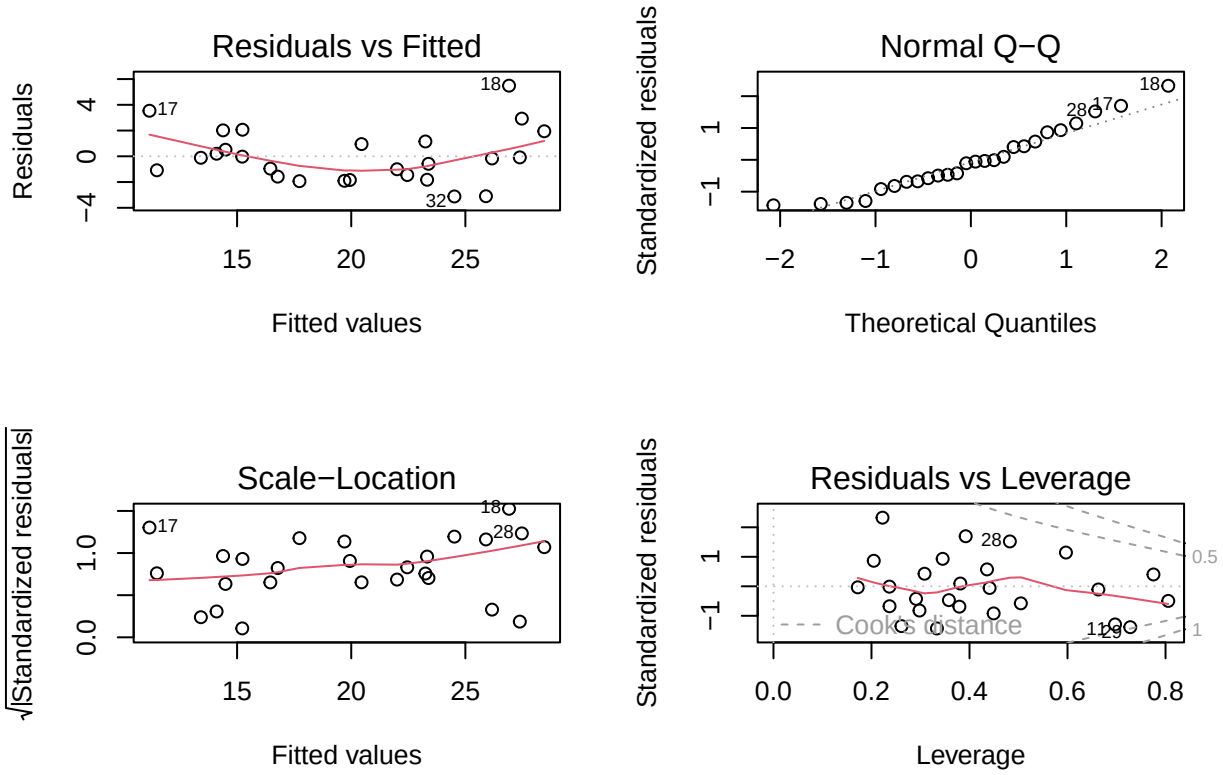



```
## Q(3)
# FIrst, fit the full model: Begin by fitting a multiple linear regression
#model using all the predictor variables (except 'name') to predict the
#response variable 'mpg'.
full_model <- lm(mpg ~ . - name, data = training_sample)
summary(full_model)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ . - name, data = training_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1191 -1.5480 -0.1406  1.1004  5.4866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.312675  20.343594   1.195    0.251
## cyl         -0.654104   1.149880  -0.569    0.578
## disp         0.005687   0.023218   0.245    0.810
## hp          -0.022785   0.024709  -0.922    0.371
## drat         0.224612   1.848097   0.122    0.905
## wt          -2.346017   2.218873  -1.057    0.307
## qsec         0.267267   0.854499   0.313    0.759
## vs           0.574131   2.381999   0.241    0.813
## am           1.805444   2.354189   0.767    0.455
## gear         0.750916   1.558551   0.482    0.637
## carb        -0.068699   0.918884  -0.075    0.941
##
## Residual standard error: 2.675 on 15 degrees of freedom
## Multiple R-squared:  0.8719, Adjusted R-squared:  0.7865
## F-statistic: 10.21 on 10 and 15 DF,  p-value: 5.107e-05
```

```r
# From the output above, the full model summary shows that the multiple
#R-squared value is 0.8719, which means that about 87.19% of the variation
#in 'mpg' can be explained by the predictor variables. However, the adjusted
#R-squared Value is only 0.7865, which which means that about 78.65% of the
#variation in 'mpg' can be explained by the predictor variablesthe. And p-values
#associated with each predictor variable are all relatively high (>= 0.05),
#indicating that none variables are statistically significant in this model.

# Second, Model diagnostics:Check the assumptions of the linear regression model
# Diagnostic plots
par(mfrow = c(2, 2))
plot(full_model)
```




```
# The red parabola indicates the trend in the residuals, and as it deviates
#from the horizontal line at y=0, suggesting that a linear model
#may not be the best fit for the data.

# And the Q-Q plot it seems that most of the points are on the diagonal line,
#which is a good sign. However, the last few points deviate from the line. This
#might indicate that there are some minor deviations from normality, but overall,
#the normality assumption seems to be mostly met.

# And in the Scale-Location plot, it seems that the points are surrounding the
#red line without any clear pattern, which is a good sign. It suggests that
#the assumption of homoscedasticity is mostly met for the current model.

# In Residuals vs Leverage plot, it seems that there is no apparent pattern,
#and most points are within the top and bottom grey dotted lines
#(0.5 Cook's distance), indicating that there are no highly influential points
#affecting the model. This is a good sign, as it suggests that the model's
#assumptions are mostly met and the model is likely reliable.

# Third, perform a backward elimination using the 'step' function:
step_model <- step(full_model, direction = "backward")
```

```
## Start:  AIC=58.86
## mpg ~ (name + cyl + disp + hp + drat + wt + qsec + vs + am +
##     gear + carb) - name
##
```

```
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.0400 107.37 56.873
## - drat  1    0.1057 107.44 56.889
## - vs    1    0.4157 107.75 56.964
## - disp  1    0.4294 107.76 56.967
## - qsec  1    0.7000 108.03 57.033
## - gear  1    1.6610 108.99 57.263
## - cyl   1    2.3154 109.65 57.419
## - am    1    4.2085 111.54 57.864
## - hp    1    6.0846 113.42 58.297
## - wt    1    7.9990 115.33 58.733
## <none>              107.33 58.864
##
## Step:  AIC=56.87
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - drat  1    0.0772 107.45 54.892
## - vs    1    0.4298 107.80 54.977
## - qsec  1    0.8191 108.19 55.071
## - disp  1    0.9742 108.35 55.108
## - gear  1    1.6962 109.07 55.281
## - cyl   1    2.7092 110.08 55.521
## - am    1    4.2981 111.67 55.894
## <none>              107.37 56.873
## - hp    1    9.9927 117.36 57.187
## - wt    1   13.9288 121.30 58.045
##
## Step:  AIC=54.89
## mpg ~ cyl + disp + hp + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.3909 107.84 52.986
## - qsec  1    0.7846 108.23 53.081
## - disp  1    0.9586 108.41 53.123
## - gear  1    1.9026 109.35 53.348
## - cyl   1    3.3538 110.80 53.691
## - am    1    4.5871 112.04 53.979
## <none>              107.45 54.892
## - hp    1    9.9204 117.37 55.188
## - wt    1   13.9717 121.42 56.070
##
## Step:  AIC=52.99
## mpg ~ cyl + disp + hp + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - disp  1    1.0959 108.94 51.249
## - gear  1    1.7750 109.61 51.411
## - qsec  1    1.7915 109.63 51.415
## - am    1    4.4761 112.32 52.044
## - cyl   1    5.1486 112.99 52.199
## <none>              107.84 52.986
## - hp    1    9.6563 117.50 53.216
## - wt    1   17.0835 124.92 54.810
```

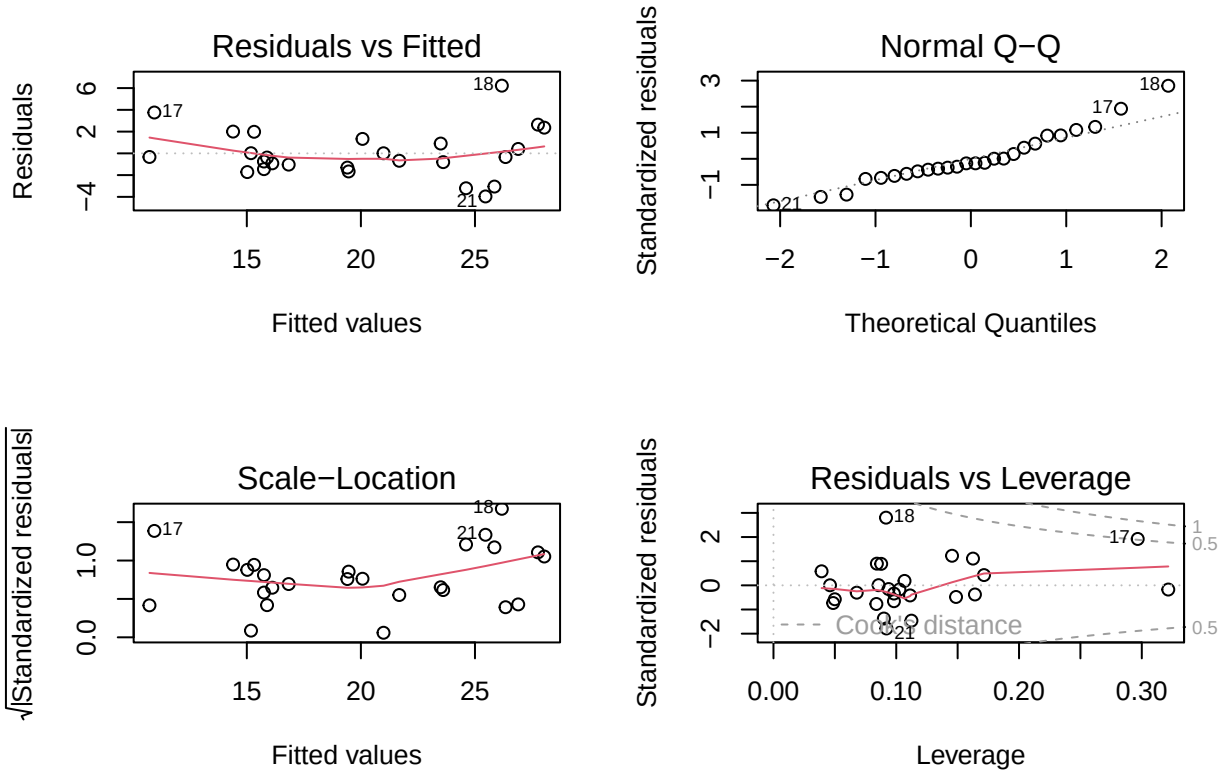
```
##
## Step:  AIC=51.25
## mpg ~ cyl + hp + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - qsec  1    1.1070 110.04 49.512
## - gear  1    1.3933 110.33 49.580
## - am    1    3.8874 112.82 50.161
## - cyl   1    4.5044 113.44 50.303
## - hp    1    8.5743 117.51 51.219
## <none>              108.94 51.249
## - wt    1   21.3735 130.31 53.907
##
## Step:  AIC=49.51
## mpg ~ cyl + hp + wt + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1    1.2356 111.28 47.803
## - am    1    2.7870 112.83 48.163
## <none>              110.04 49.512
## - hp    1   12.3247 122.37 50.272
## - cyl   1   12.8180 122.86 50.377
## - wt    1   21.2986 131.34 52.112
##
## Step:  AIC=47.8
## mpg ~ cyl + hp + wt + am
##
##        Df Sum of Sq    RSS    AIC
## - am    1     7.856 119.14 47.576
## <none>              111.28 47.803
## - hp    1    11.633 122.91 48.388
## - cyl   1    21.102 132.38 50.317
## - wt    1    23.050 134.33 50.697
##
## Step:  AIC=47.58
## mpg ~ cyl + hp + wt
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     5.694 124.83 46.790
## <none>              119.14 47.576
## - cyl   1    30.092 149.23 51.432
## - wt    1    66.314 185.45 57.082
##
## Step:  AIC=46.79
## mpg ~ cyl + wt
##
##        Df Sum of Sq    RSS    AIC
## <none>              124.83 46.790
## - wt    1    66.895 191.72 55.947
## - cyl   1    94.363 219.19 59.428
```

```
summary(step_model)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ cyl + wt, data = training_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9609 -1.2399 -0.3638  1.2217  6.2218
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.8517     1.7019  22.828  < 2e-16 ***
## cyl          -1.6795     0.4028  -4.170 0.000369 ***
## wt           -2.7070     0.7711  -3.511 0.001878 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.33 on 23 degrees of freedom
## Multiple R-squared:  0.851,  Adjusted R-squared:  0.838
## F-statistic: 65.68 on 2 and 23 DF,  p-value: 3.103e-10
```

```
#The summary of the refined step_model indicates that it's a good fit for the
#data. The adjusted R-squared value is 0.838, which suggests that the model
#explains about 83.8% of the variance in the response variable (mpg) using
#the two predictor variables (cyl and wt), which is higher than the initial
#full model's Adjusted R-squared Value that is 0.7865. The increase in the
#adjusted R-squared value suggests that the refined model is a better choice
#because it balances the trade-off between model complexity and
#model performance.Also the F-statistic is 65.68 with a p-value of
#3.103e-10, which is highly significant. This indicates that the model as a
#whole is significantly better than a model with no predictors.

# Then, plot Diagnostic plots for the step_model
par(mfrow = c(2, 2))
plot(step_model)
```




```
# By obeserving the new plots, comparing with the disgnostic plots for the
#initial model, everything is getting a little bit better, the assumptions of
#the linear regression model are met.

# In this case, we did not perform any transformation on the response or the
#predictors, because after comparing with the disgnostic plots for the
#initial model, everything is getting a little bit better,
#the assumptions of the linear regression model are met.
```

```
## Q(4)Use the predict function to generate predictions for the test dataset:
test_sample$predicted_mpg <- predict(step_model, newdata = test_sample)
print(test_sample$predicted_mpg)
```

```
## [1] 16.10353 19.46253 11.20382 27.16630 15.00718 21.27623
```

```
# Then calculate the residuals
test_sample$residuals <- test_sample$mpg - test_sample$predicted_mpg
print(test_sample$residuals)
```

```
## [1]  2.5964729 -0.2625305 -0.8038227  6.7337006  4.1928156 -1.5762332
```

```
#Residuals represent the difference between the actual and predicted values.

# Then calculate performance metrics:
```

```r
#MSE for test sample
mse_test <- mean(test_sample$residuals^2)
print(mse_test)
```

```
## [1] 12.14394
```

```r
#comparing the mse_test with the MSE for training sample
training_sample$predicted_mpg <- predict(step_model, newdata = training_sample)
training_sample$residuals <- training_sample$mpg - training_sample$predicted_mpg
mse_training <- mean(training_sample$residuals^2)
print(mse_training)
```

```
## [1] 4.801123
```

```r
# The Mean Squared Error (MSE) for the test dataset is 12.14394,
#and for the training dataset, it is 4.801123. The test MSE is higher than the
#training MSE, This difference indicates that the model is performing better
#on the training dataset than on the test dataset.

#R-squared
SST <- sum((test_sample$mpg - mean(test_sample$mpg))^2)
SSR <- sum(test_sample$residuals^2)
r_squared <- 1 - (SSR/SST)
print(r_squared)
```

```
## [1] 0.7472016
```

```r
#The R-squared value for the test dataset is 0.7472016. This means that
#approximately 74.72% of the variance in the mpg variable can be explained
#by the selected model on the test dataset.


##Overall assessment:

# 1.Variable selection: The initial full model included all predictors,
#but the stepwise variable selection process helped identify a more parsimonious
#model with only two significant predictors, cyl and wt. This simplified model
#provides a more interpretable and potentially more generalizable model,
#with less risk of overfitting.

# 2.Model diagnostics: The diagnostic plots of the selected model showed that
#the assumptions of linear regression were reasonably met, with no strong
#evidence of non-linearity, heteroscedasticity, or violation of the normality
#of residuals.

# 3.Model performance: The adjusted R-squared value for the simplified model was
#0.838 on the training dataset, which indicates that the model explains
#approximately 83.8% of the variance in mpg. The R-squared value for the test
#dataset was 0.747, which is slightly lower than the training dataset but still
#indicates decent predictive performance.
```

```
# 4.Prediction performance: The Mean Squared Error (MSE) was 4.80 for the
#training dataset and 12.14 for the test dataset. The higher MSE in the test
#dataset may suggest that the model is not perfectly generalizing to unseen data.
#However, given the small sample size, this difference might not be too
#concerning. It's important to remember that the model's performance might vary
#depending on the specific data points in the training and test datasets.

# In conclusion, the simplified model with only cyl and wt as predictors
#performed reasonably well in explaining the variance in mpg. The model
#assumptions were largely met, and the model showed decent predictive
#performance on the test dataset.
```