Regression Model Selection

Douglas Martins

2023-05-31

Problem Specification

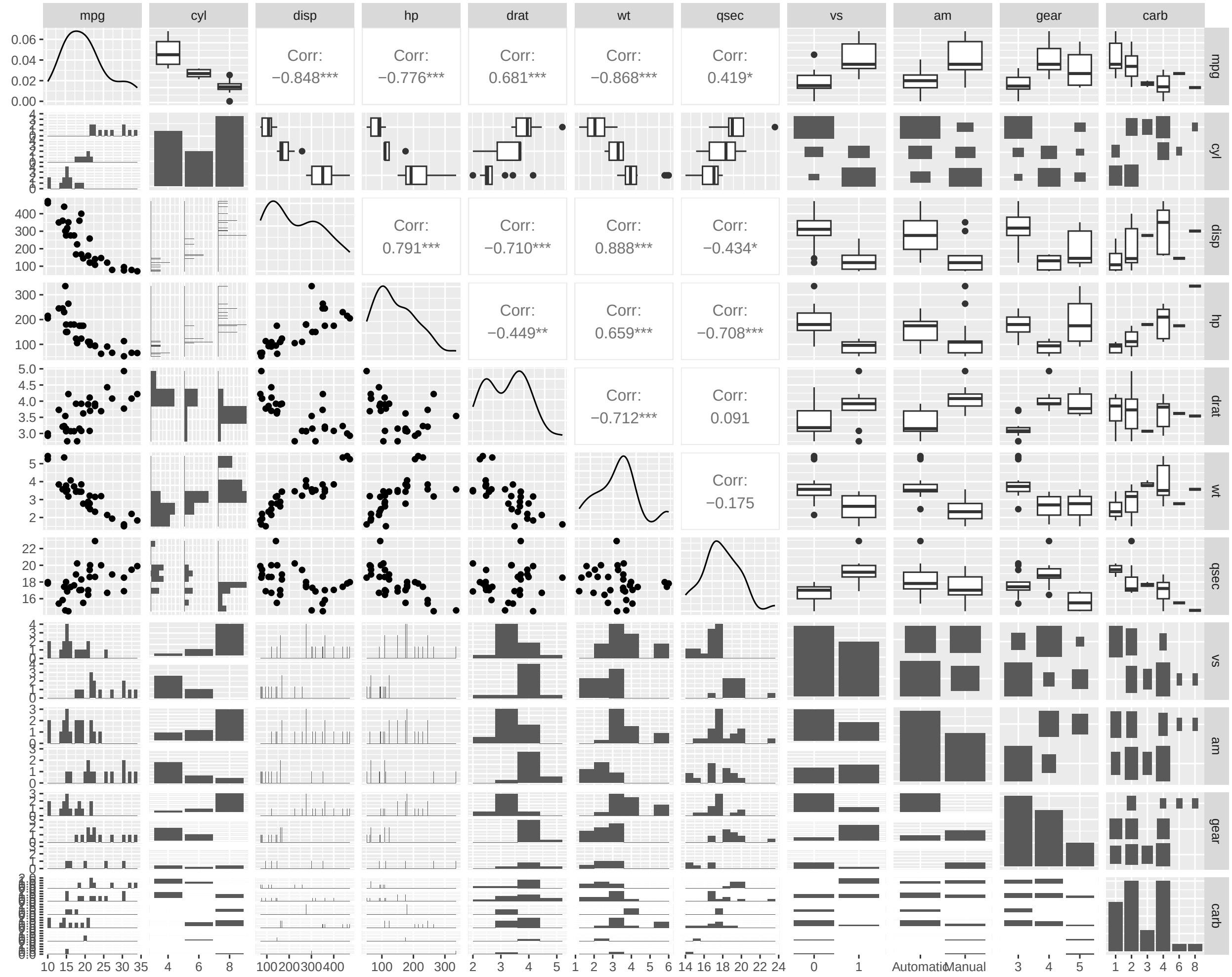
This report looks at the mtcars dataset to explore the relationship between a set of variables and fuel autonomy in MPG. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). Regression models and exploratory data analyses are used to mainly explore how automatic (am = 0) and manual (am = 1) transmissions features affect the MPG feature.

T-test shows that the performance difference between cars with automatic and manual transmission. And it is about 7 MPG more for cars with manual transmission than those with automatic transmission. Then, several linear regression models are fitted and one with highest Adjusted R-squared value is selected.

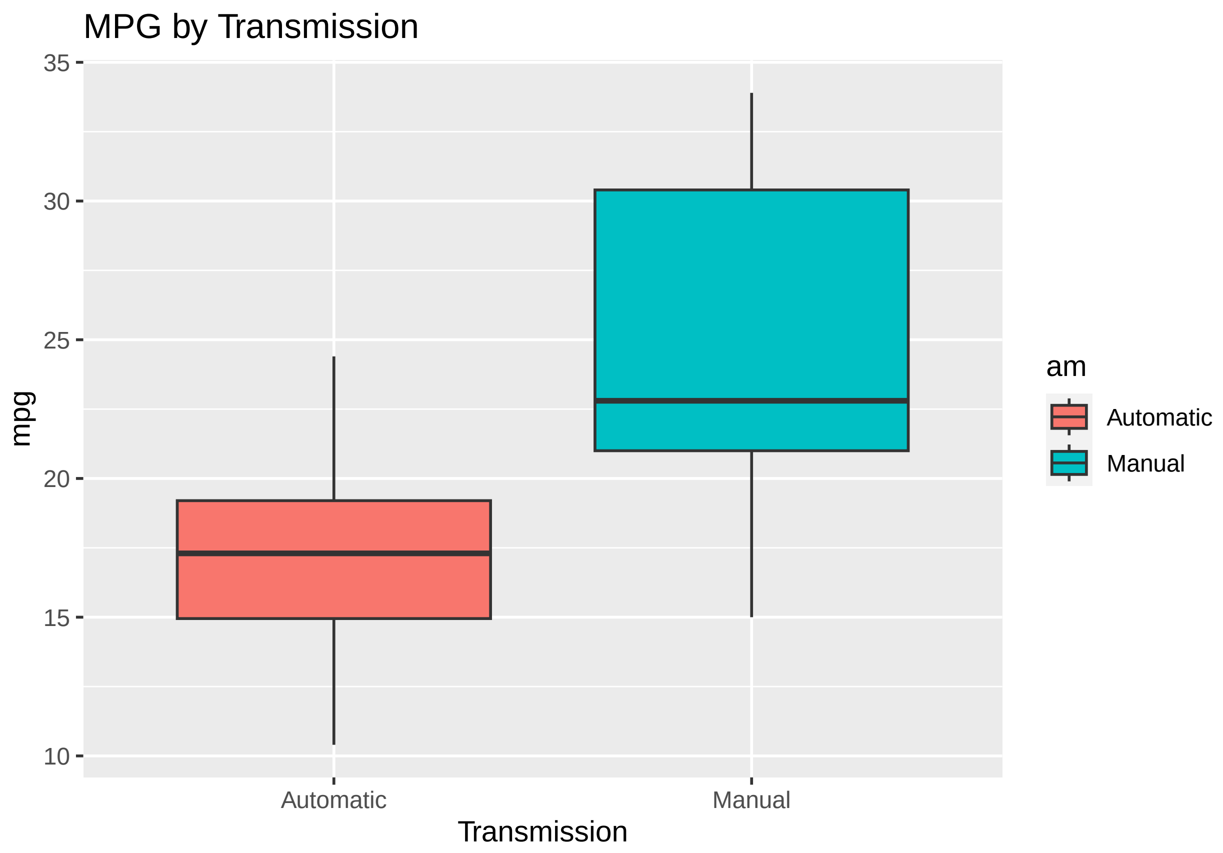
So, given that weight and quarter mile time are held constant, manual transmitted cars are 14.079 + (-4.141)*weight more MPG (miles per gallon) on average better than automatic transmitted cars. Thus, cars that are lighter in weight with a manual transmission and cars that are heavier in weight with an automatic transmission will have higher fuel economy.

Exploratory Analysis

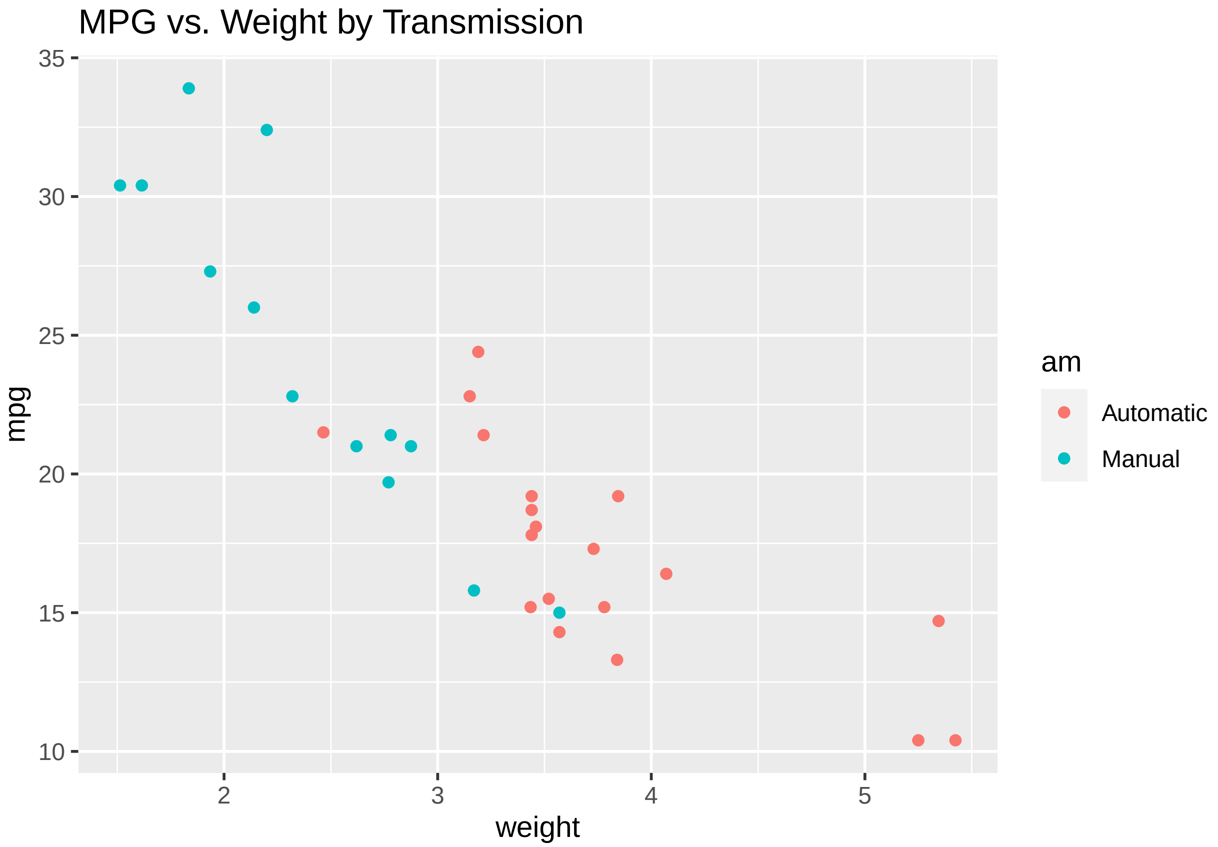
```
##       mpg        cyl         disp             hp             drat      
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695  
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597  
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930  
##        wt             qsec       vs             am      gear   carb  
##  Min.   :1.513   Min.   :14.50   0:18   Automatic:19   3:15   1: 7  
##  1st Qu.:2.581   1st Qu.:16.89   1:14   Manual   :13   4:12   2:10  
##  Median :3.325   Median :17.71                         5: 5   3: 3  
##  Mean   :3.217   Mean   :17.85                                4:10  
##  3rd Qu.:3.610   3rd Qu.:18.90                                6: 1  
##  Max.   :5.424   Max.   :22.90                                8: 1  
```



With the above plots its possible to infer that there seems to exist a relationship between fuel economy with “wt”, “disp”, “cyl”, “hp” and “am” (automatic transmission)



Looking in detail to mpg by transmission it appears to exist a relationship between type of transmission and the fuel economy of a vehicle as seen above



Considering there is a close relationship between weight and transmission we must include the variable in the model as an interaction term (wt*am)

Inference

For this step, null hypothesis is made as the MPG of the automatic and manual transmissions are within the same population (assuming the MPG has a normal distribution). Two sample T-tests are used to show it.

```
result <- t.test(mpg ~ am,df)
result$p.value
```

```
## [1] 0.001373638
```

```
result$estimate
```

```
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Since the p-value is 0.00137, the null hypothesis is rejected. So, automatic and manual transmissions are from different populations. And the mean for MPG of manual transmitted cars is about 7 more than that of automatic transmitted cars.

Model selection

```
fit_all <- lm(mpg ~ .,data=df)
summary(fit_all)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## amManual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

This model has the Residual standard error as 2.833 on 15 degrees of freedom. Adjusted R-squared value is 0.779, so the model can explain about 78% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significant level.

```
sqrt(vif(fit_all))
```

```
##           GVIF       Df GVIF^(1/(2*Df))
## cyl  11.319053 1.414214        1.834225
## disp  7.769536 1.000000        2.787389
## hp    5.312210 1.000000        2.304823
## drat  2.609533 1.000000        1.615405
## wt    4.881683 1.000000        2.209453
## qsec  3.284842 1.000000        1.812413
## vs    2.843970 1.000000        1.686407
## am    3.151269 1.000000        1.775181
## gear  7.131081 1.414214        1.634138
## carb 22.432384 2.236068        1.364858
```

When selecting all variables (except model and manufacturer, which are not relevant in this analysis) we can see high amounts for VIF for most of them, leading to a conclusion we might be inflating standard errors in adding unnecessary or correlated variables.

Then, backward selection is used to select some statistically significant variables.

```
summary(fit_step)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

The model outputted is “mpg ~ wt + qsec + am”. The Residual standard error is 2.459 on 28 degrees of freedom. The Adjusted R-squared value is 0.8336, so the model can explain about 83% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

In the exploratory step it is shown that there appears to be an interaction term between “wt” variable and “am” variable, since automatic cars are usually heavier than manual cars. Thereby, the following model including the interaction term is generated:

```
fit_wt <- update(fit_step, mpg ~ wt*am + qsec)
summary(fit_wt)
```

```
##
## Call:
## lm(formula = mpg ~ wt + am + qsec + wt:am, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## amManual      14.079      3.435   4.099 0.000341 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## wt:amManual   -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

This model has the Residual standard error as 2.084 on 27 degrees of freedom. The Adjusted R-squared value is 0.8804, so the model can explain about 88% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

For comparison with the step model, a simple model is fitted with “mpg” as the outcome variable and “am” as the predictor

```
fit_am <- update(fit_all, mpg ~ am)
summary(fit_am)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The model is selected with an analysis of the output of the anova function, which computes an analysis of variance (or deviance) tables for the supplied models.

```
anova(fit_am,fit_step,fit_wt,fit_all)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ wt + am + qsec + wt:am
## Model 4: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                   
## 2     28 169.29  2    551.61 34.3604 2.509e-06 ***
## 3     27 117.28  1     52.01  6.4795    0.0224 *  
## 4     15 120.40 12     -3.13                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova shows that adding an interaction between wt and am appears to be necessary over looking at wt and am without interaction. For steps 3 to 4 there seems to be a negative impact in adding all the others variables, as such that the selected mode is number 3, fit_wt.

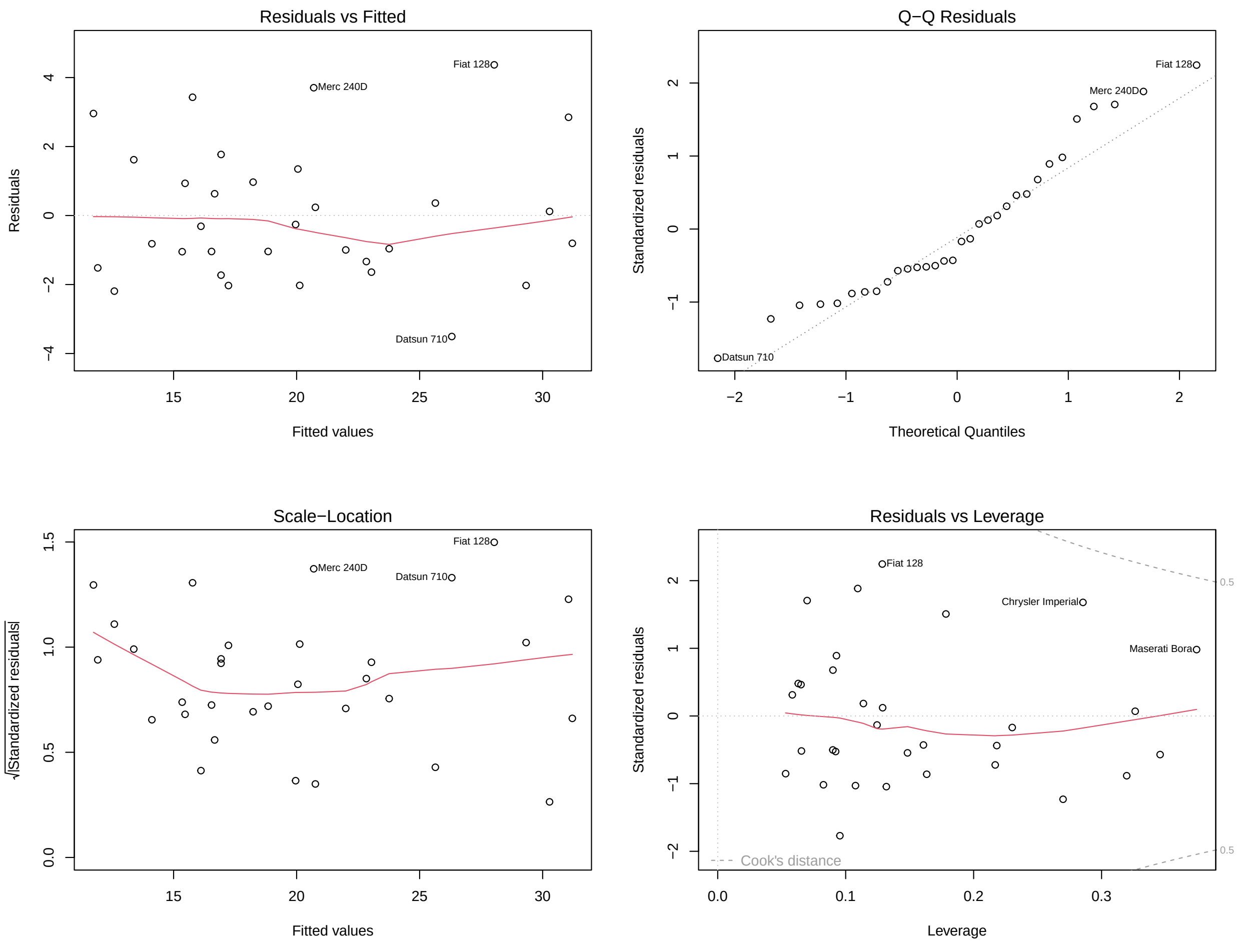
```
confint(fit_wt)
```

```
##                  2.5 %    97.5 %
## (Intercept) -2.3807791 21.826884
## wt          -4.3031019 -1.569960
## amManual     7.0308746 21.127981
## qsec         0.4998811  1.534066
## wt:amManual -6.5970316 -1.685721
```

The result shows that when “wt” (weight lb/1000) and “qsec” (quarter mile time) are held constant, cars with manual transmission add 14.079 + (-4.141)*wt more MPG (miles per gallon) on average than cars with automatic transmission. For example, a manual transmitted car that weighs 2000 lbs have 5.8 more MPG than an automatic transmitted car that has both the same weight and quarter mile time. For cars over 3400 lbs the added weight of the automatic transmission shows benefit compared to a car equipped with a manual transmission.

Residual Analysis

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



a. The Residuals vs Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.
b. The Q-Q Residuals plot indicates that the residuals are normally distributed because the points lie closely to the line.
c. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.
d. The Residuals vs Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

As for the dfbetas, the measure of how much an observation has effected the estimate of a regression coefficient, cooks distance gives a summary of the dfbetas contribution and shows there are no points with orders of magnitudes of difference.

```
cooks.distance(fit_wt)
```

```
##           Mazda RX4       Mazda RX4 Wag          Datsun 710      Hornet 4 Drive
##        0.0049793347        0.0004415911        0.0661219481        0.0090993691
##   Hornet Sportabout             Valiant          Duster 360           Merc 240D
##        0.0162616657        0.0255537430        0.0103520390        0.0872122272
##            Merc 230            Merc 280           Merc 280C          Merc 450SE
##        0.0344001241        0.0030925804        0.0037467801        0.0029871695
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood Lincoln Continental 
##        0.0012073720        0.0081445612        0.1119773985        0.0732458668 
##   Chrysler Imperial            Fiat 128         Honda Civic      Toyota Corolla 
##        0.2251059974        0.1488367000        0.0106708552        0.0986535100 
##       Toyota Corona    Dodge Challenger         AMC Javelin          Camaro Z28 
##        0.0289834124        0.0055987125        0.0186236741        0.0070393413 
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2        Lotus Europa 
##        0.0437275758        0.0330986742        0.0008691236        0.0004729663 
##      Ford Pantera L        Ferrari Dino       Maserati Bora          Volvo 142E 
##        0.0017368639        0.0005048477        0.1151986709        0.0289598448
```

Concluding, the above analyses meet the basic assumptions of linear regression and well answer the questions about the effect of manual vs automatic transmission.