Regression Models Course Project

Danny Beery

2022-09-30

This analysis will use the 'mtcars' dataset to investigate whether there is a difference in miles per gallon in automatic and manual transmission vehicles.

```
#Packages used
library(tidyr)
library(dplyr)
library(ggplot2)
library(ggpubr)
library(corrplot)
```

Summary statistics

```
#Changing variable names out of binary notation
mtcars$am[mtcars$am == 0] <- "Automatic"
mtcars$am[mtcars$am == 1] <- "Manual"

#Obtaining summary statistics for automatic and manual data
mpg_by_transmission <- mtcars %>%
  group_by(am) %>% #0 = auto, 1 = manual
  summarise(
    mean_mpg = mean(mpg, na.rm = TRUE),
    sd_mpg = sd(mpg, na.rm = TRUE),
    var_mpg = var(mpg, na.rm = TRUE)
  )
mpg_by_transmission
```

```
## # A tibble: 2 x 4
##   am        mean_mpg sd_mpg var_mpg
##   <chr>        <dbl>  <dbl>   <dbl>
## 1 Automatic     17.1   3.83    14.7
## 2 Manual        24.4   6.17    38.0
```

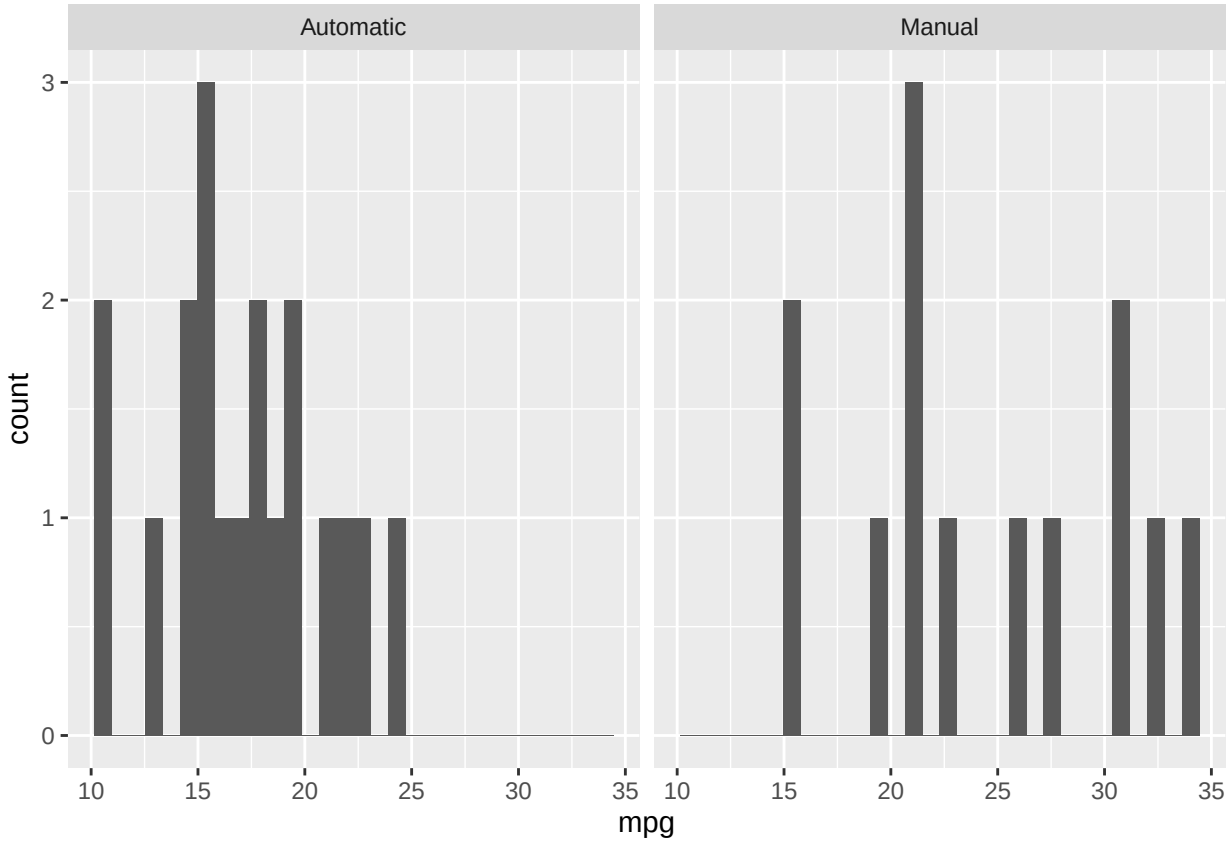
As you can see, vehicles with manual transmission appear to get better mpg than vehicles with automatic transmissions, but the variance in mpg is greater among vehicles for manual transmission.

Let's now visualize the data using a histogram and a boxplot.

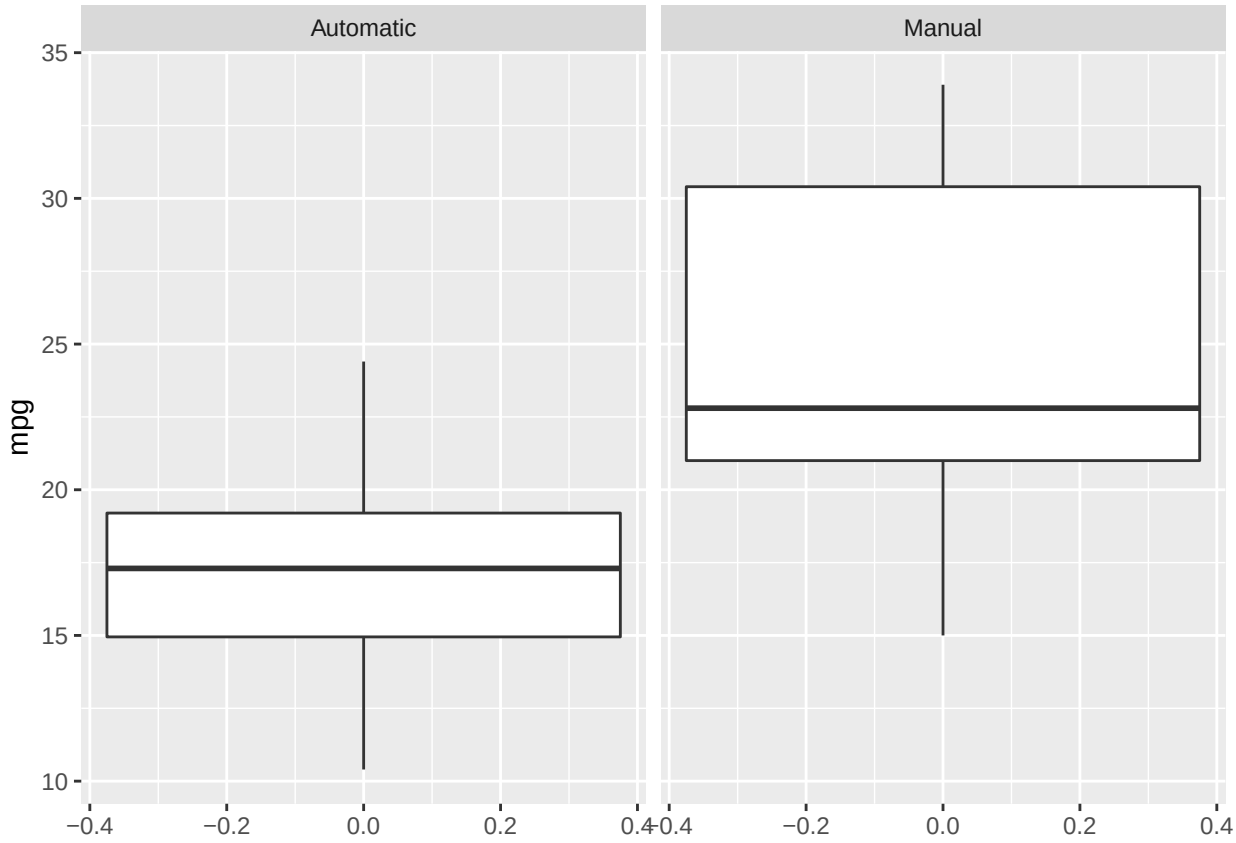
Visualing Data

```
mtcars %>%
  ggplot(aes(mpg)) +
  geom_histogram() +
  facet_wrap(~am)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
mtcars %>%
  ggplot(aes(y = mpg)) +
  geom_boxplot() +
  facet_wrap(~am)
```



Our plots confirm that vehicles with manual transmissions have a greater average miles per gallon and greater variability in this metric.

Hypothesis Testing

Now let's test the null hypothesis that there is no difference between the average mpg of manual and automatic transmission vehicle by creating a multi-variable linear regression model with mpg as the outcome.

But first, let's take a look at a simple linear regression using transmission status as a single predictor and mpg as an outcome.

Simple Regression Model

```
y <- lm(mpg ~ am, data = mtcars)
summary(y)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## 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
```

```
y$coefficients
```

```
## (Intercept)    amManual
##   17.147368    7.244939
```

As you can see from the unadjusted regression model, vehicles with manual transmissions have significantly greater miles per gallon than vehicles with automatic transmission.

Checking for Potential Confounders

Now let's see if any categorical variables may explain (confound) the association between mpg and transmission status.

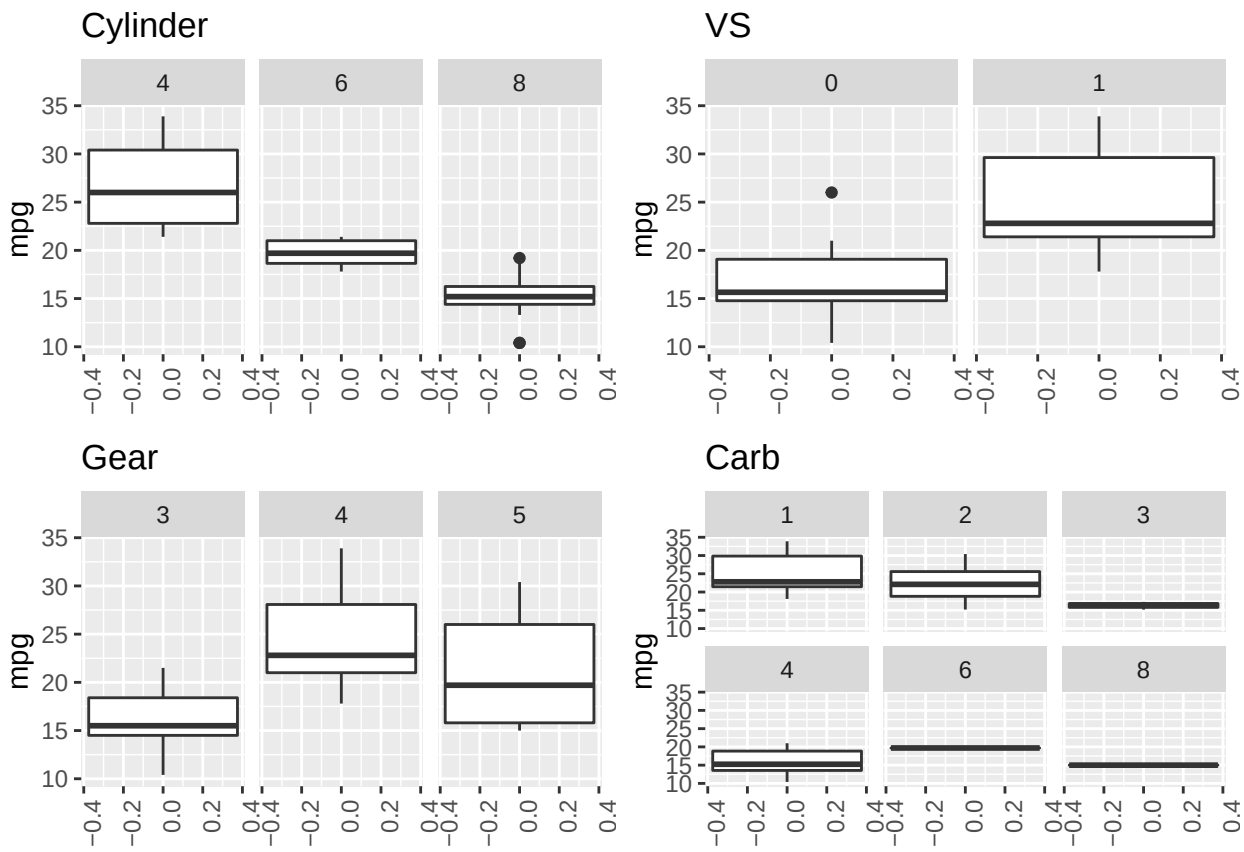
```
a <- mtcars %>%
  ggplot(aes(y = mpg)) +
  geom_boxplot() +
  facet_wrap(~cyl) +
  labs(
    title = "Cylinder"
  ) +
  theme(axis.text.x = element_text(angle=90, hjust=1))

b <- mtcars %>%
  ggplot(aes(y = mpg)) +
  geom_boxplot() +
  facet_wrap(~ vs) +
  labs(
    title = "VS"
  ) +
  theme(axis.text.x = element_text(angle=90, hjust=1))

c <- mtcars %>%
  ggplot(aes(y = mpg)) +
  geom_boxplot() +
  facet_wrap(~gear) +
  labs(
    title = "Gear"
  ) +
  theme(axis.text.x = element_text(angle=90, hjust=1))

d <- mtcars %>%
```

```
  ggplot(aes(y = mpg)) +
  geom_boxplot() +
  facet_wrap(~carb) +
  labs(
    title = "Carb"
  ) +
  theme(axis.text.x = element_text(angle=90, hjust=1))
ggarrange(a,b,c,d, ncol = 2, nrow = 2)
```



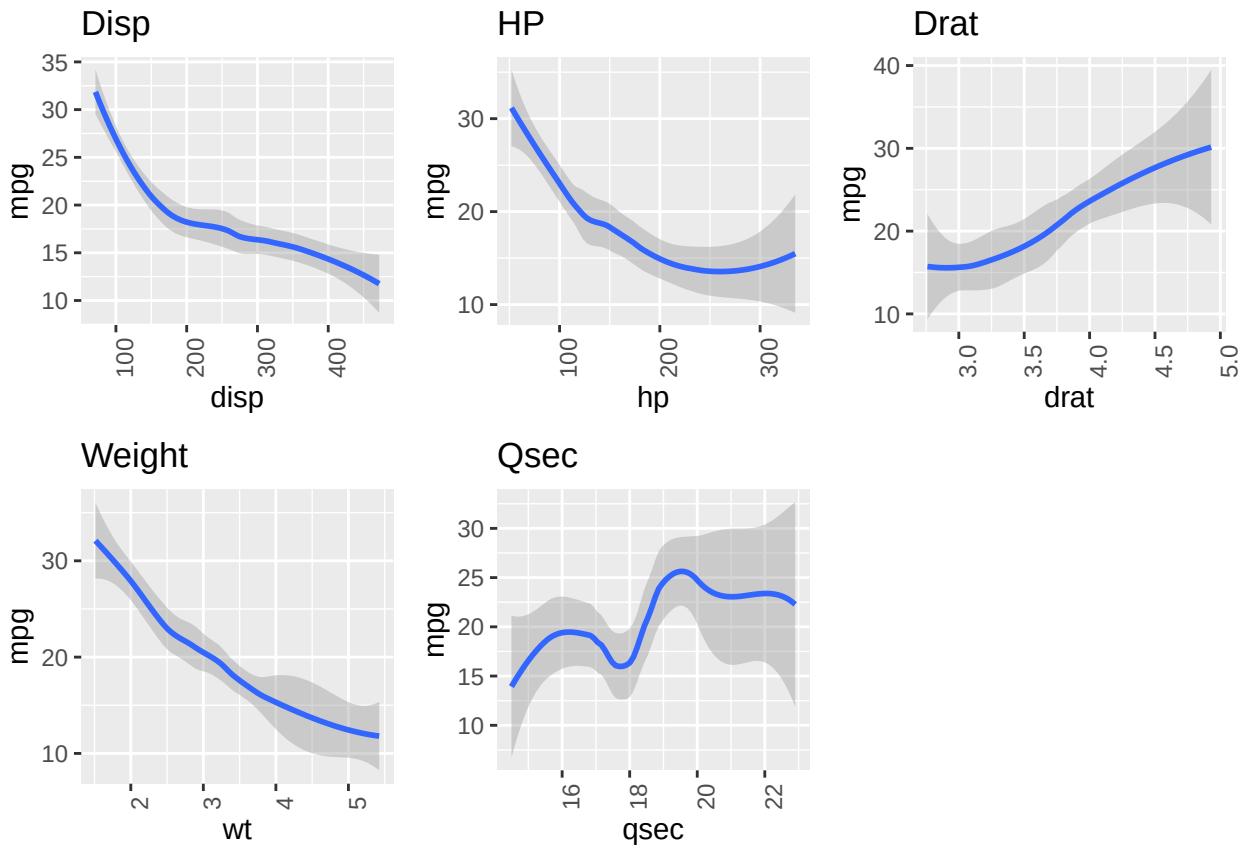
It looks like all 4 categorical variables are potentially confounding the association between transmission status and miles per gallon, thus, we should add these variables as predictors to our multivariable regression.

Now let's see if any continuous variables may explain (confound) the association between mpg and transmission status.

```
e <- mtcars %>%
  ggplot(aes(disp, mpg)) +
  geom_smooth() +
  labs(title = "Disp") +
  theme(axis.text.x = element_text(angle=90, hjust=1))
f <- mtcars %>%
  ggplot(aes(hp, mpg)) +
  geom_smooth() +
  labs(title = "HP") +
  theme(axis.text.x = element_text(angle=90, hjust=1))
g <- mtcars %>%
```

```
  ggplot(aes(drat, mpg)) +
  geom_smooth() +
  labs(title = "Drat") +
  theme(axis.text.x = element_text(angle=90, hjust=1))
h <- mtcars %>%
  ggplot(aes(wt, mpg)) +
  geom_smooth() +
  labs(title = "Weight") +
  theme(axis.text.x = element_text(angle=90, hjust=1))
i <- mtcars %>%
  ggplot(aes(qsec, mpg)) +
  geom_smooth() +
  labs(title = "Qsec") +
  theme(axis.text.x = element_text(angle=90, hjust=1))
ggarrange(e, f, g, h, i, ncol = 3, nrow = 2)
```

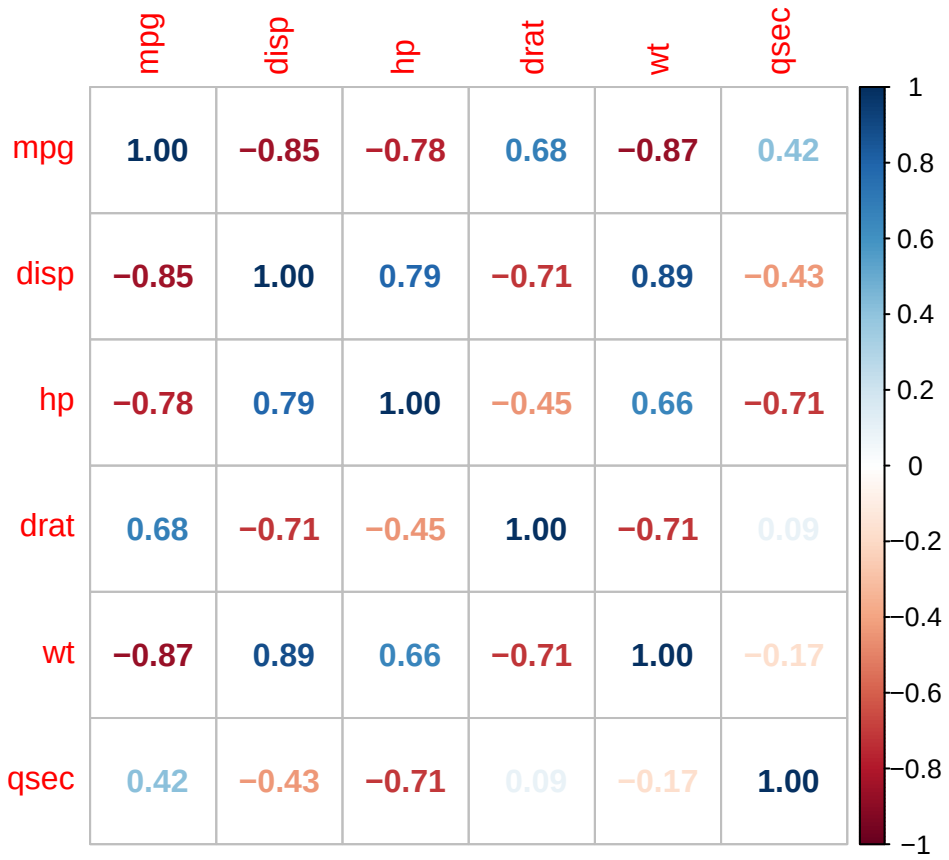
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula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




```
#Correlation Plot
mtcars_cor <- mtcars[c(1,3,4,5,6,7)]
cor(mtcars_cor)
```

```
##             mpg       disp         hp        drat         wt        qsec
## mpg   1.0000000 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403
## disp -0.8475514  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788
## hp   -0.7761684  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339
## drat  0.6811719 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476
## wt   -0.8676594  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588
## qsec  0.4186840 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000
```

```
corrplot(cor(mtcars_cor), "number")
```



From the plots and correlations above, it looks like all 4 continuous variables potentially confound the association between mpg and type of transmission. We will include these variables in our multivariate regression model.

```
y2 <- lm(data = mtcars, mpg ~ am + as.factor(cyl) + disp + hp + drat + wt + qsec + vs + gear + carb)
y2$coefficients
```

```
##     (Intercept)        amManual as.factor(cyl)6 as.factor(cyl)8            disp
##     17.81984325      2.61726546     -1.66030673      1.63743980      0.01391241
##              hp            drat              wt            qsec              vs
##     -0.04612835      0.02635025     -3.80624757      0.64695710      1.74738689
##            gear            carb
##      0.76402917      0.50935118
```

```
confint(y2)
```

```
##                         2.5 %      97.5 %
## (Intercept)     -16.19392520 51.83361170
## amManual         -1.56456843  6.79909935
## as.factor(cyl)6  -6.37937479  3.05876134
## as.factor(cyl)8  -7.36502463 10.63990423
## disp             -0.02238703  0.05021185
## hp               -0.10270005  0.01044335
## drat             -3.47074565  3.52344615
## wt               -7.65827756  0.04578242
## qsec             -0.85900473  2.15291892
## vs               -2.99332407  6.48809786
## gear             -2.27455239  3.80261072
## carb             -1.45654798  2.47525034
```

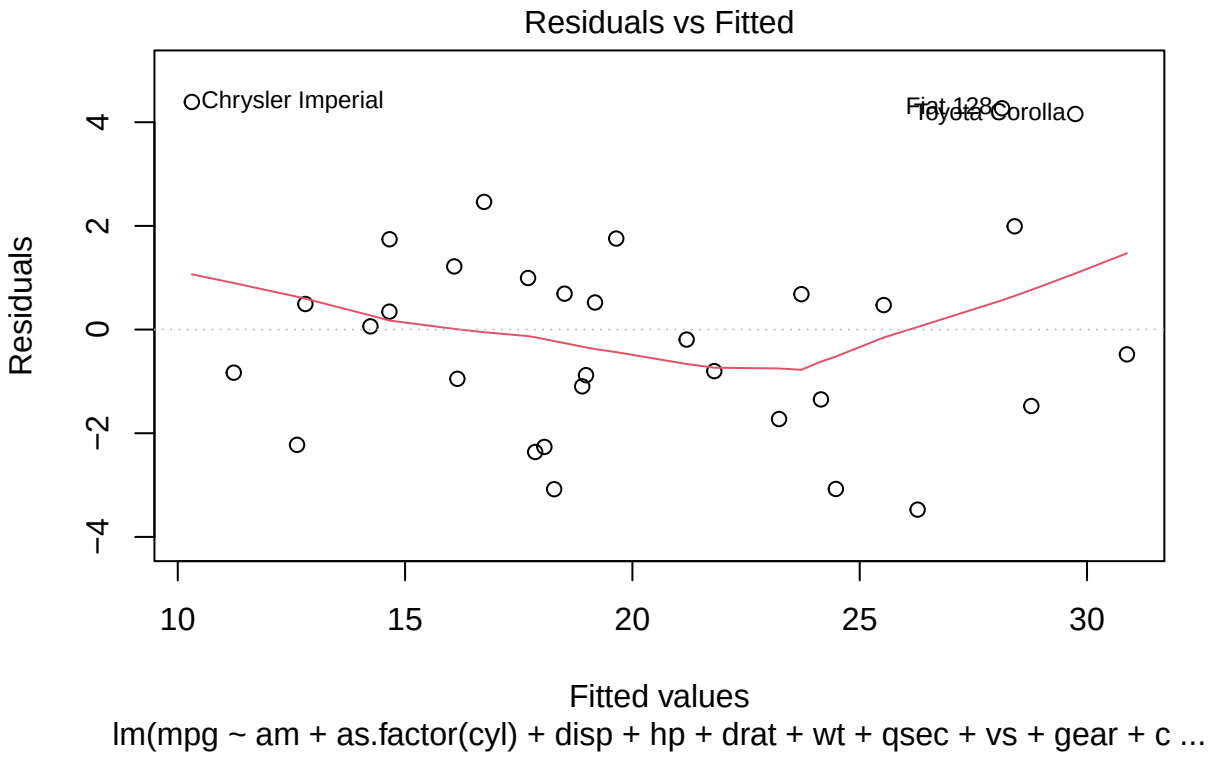
After adjusting for all the potential confounders, it appears that type of transmission has no effect on mpg (confidence interval contains 0).

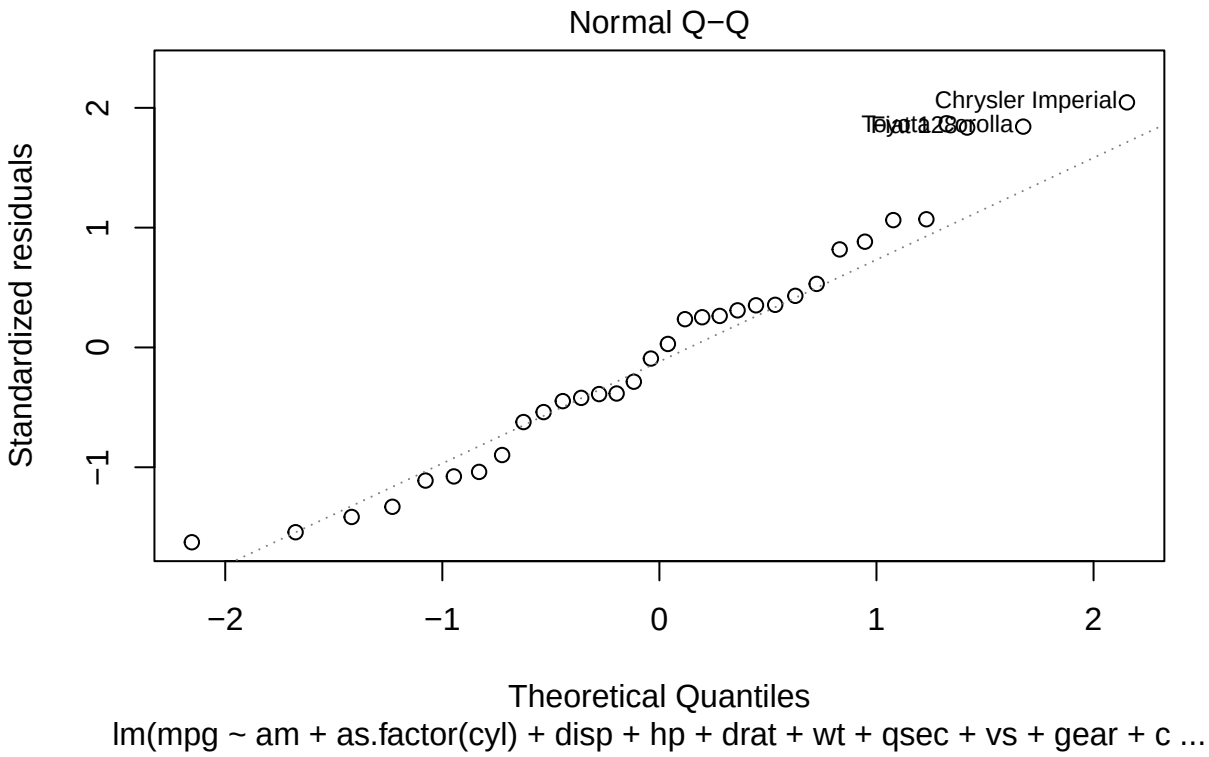
Multivariate Regression Model

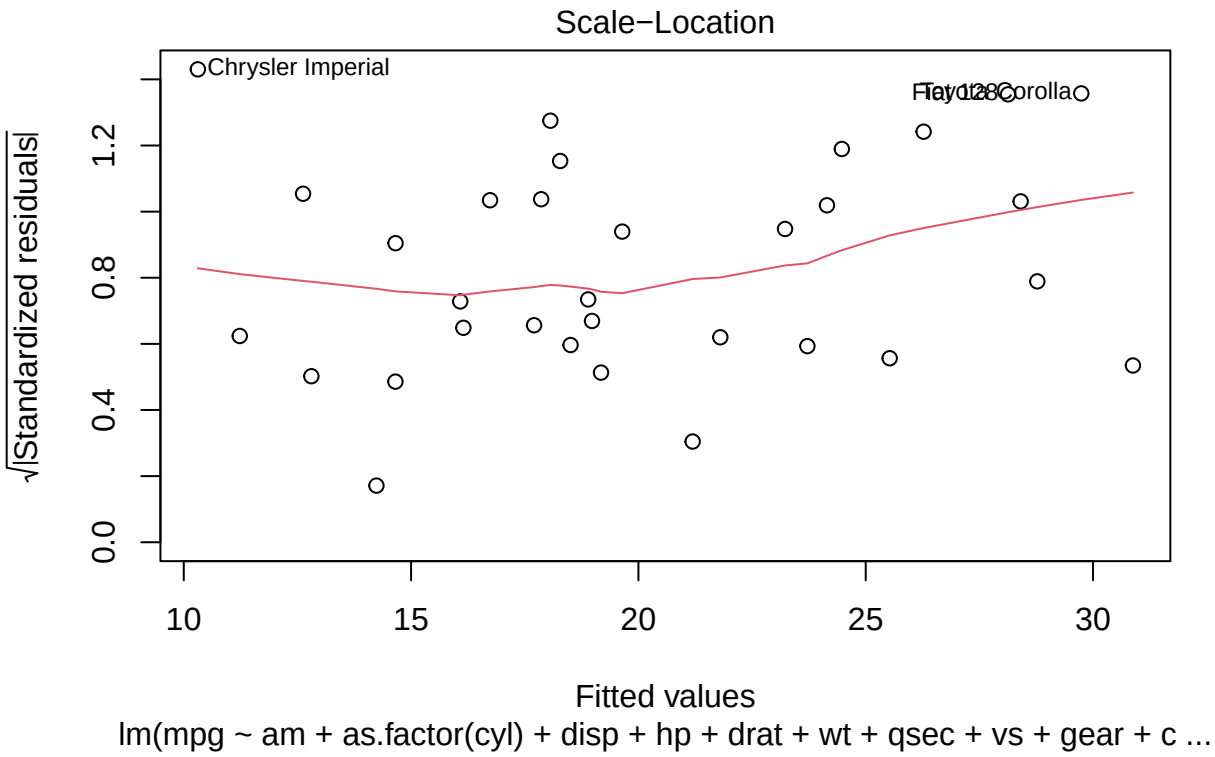
Let's check the appropriateness of our multivariate regression model.

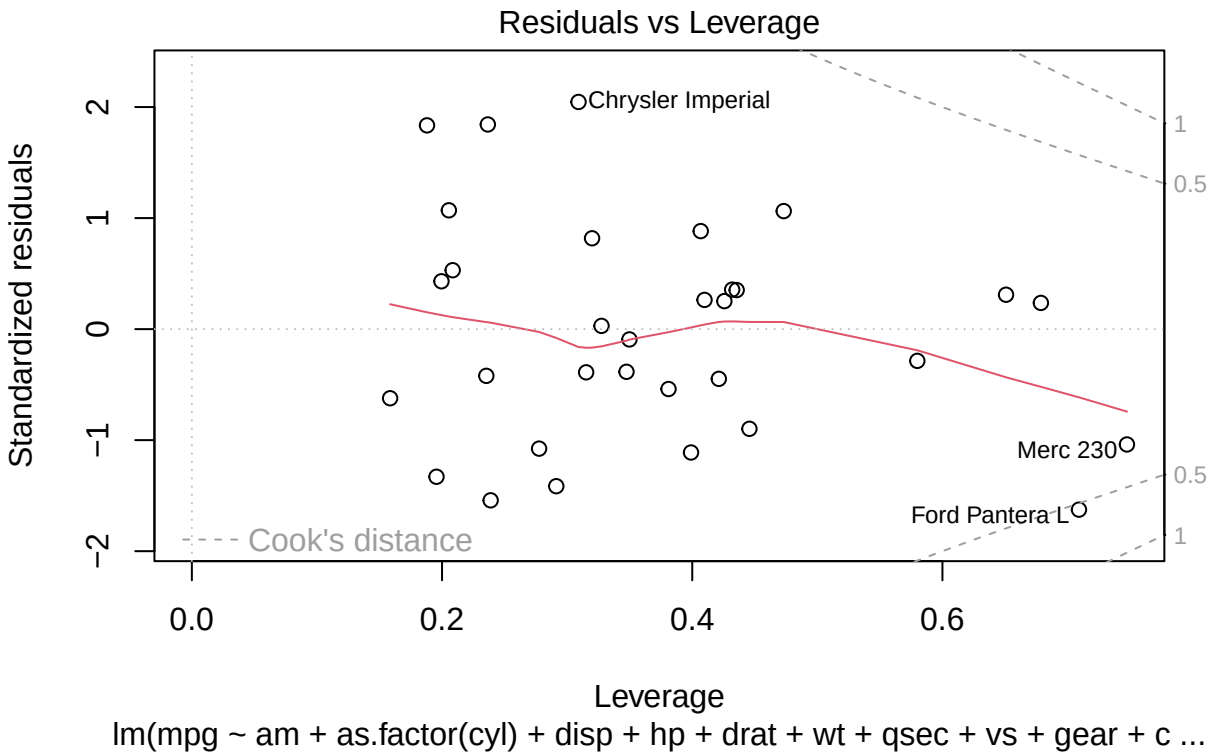
Assessing Model Fit

```
plot(y2)
```


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
4
2
0
−2
−4
10
15
20
25
30
Fitted values
lm(mpg ~ am + as.factor(cyl) + disp + hp + drat + wt + qsec + vs + gear + c ...


Normal Q−Q
Chrysler Imperial
Toyota Corolla
Standardized residuals
2
1
0
−1
−2
−1
0
1
2
Theoretical Quantiles
lm(mpg ~ am + as.factor(cyl) + disp + hp + drat + wt + qsec + vs + gear + c ...


Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
0.0
0.4
0.8
1.2
10
15
20
25
30
Fitted values
lm(mpg ~ am + as.factor(cyl) + disp + hp + drat + wt + qsec + vs + gear + c ...



As you can see from the residuals vs fitted plot and QQ plot, the linear model seems to be quite appropriate for the data.

Conclusion

After performing a multivariate regression analysis of the data, we conclude that there is no difference in miles per gallon between automatic and manual transmission vehicles. Any differences in miles per gallon between automatic and manual transmission vehicles can be explained by confounding factors.