Course-Project

Introduction

Here I will analyze the relationship between transmission type (automatic vs manual) on the fuel efficiency (miles per gallon) of the cars in the mtcars data set. First I will perform an exploratory data analysis, then regression modeling and statistical inference, check the reliability of the model with residual diagnostics, and finally present conclusions and assessment of uncertainty.

Exploratory Analysis

First, to get a rough idea of what kind of data and how much there is in the data set, look at the head of mtcars

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

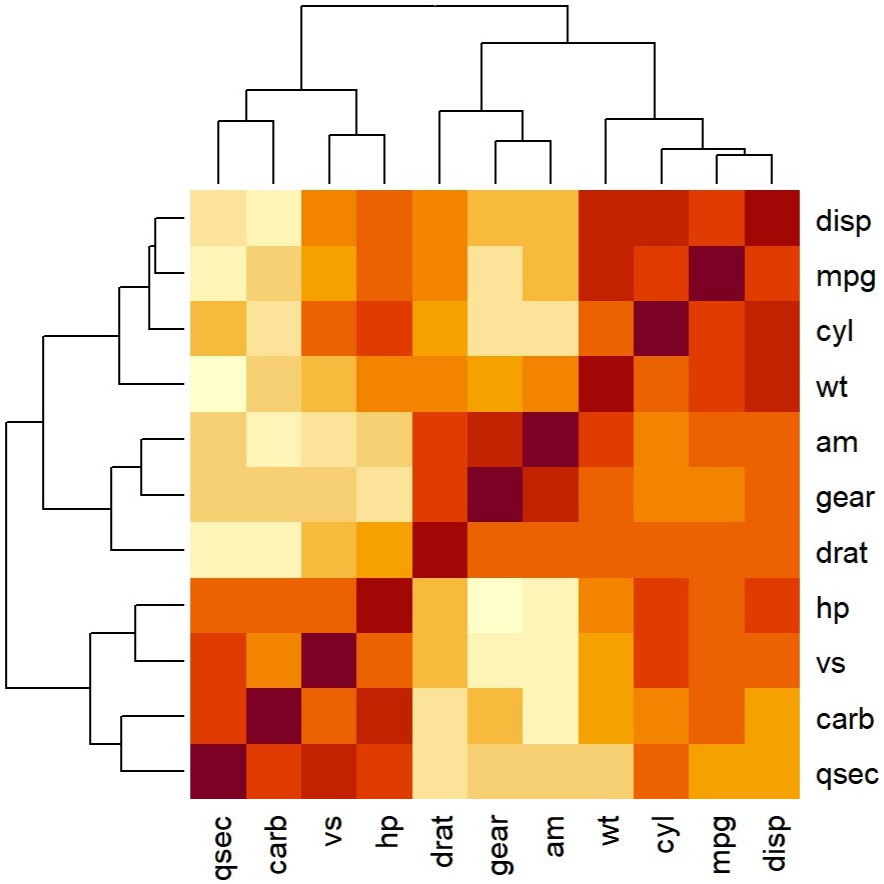
Looks like the transmission type is recorded in the field “am”, as a binary value where 1 means automatic, 0 means manual. This data is already tidy, so for a preliminary analysis we can just use a simple linear model of the correlation we hope to test.

```
mdl_a <- lm(mpg~am, mtcars)
summary(mdl_a)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

This suggests that naively, automatic transmissions seem to get 7 more miles per gallon. However, there may be bias, since other variables in this data seem likely to be correlated. To check for the possibility of bias, examine the correlation matrix, using absolute value to check strength of correlations:

```
heatmap(abs(cor(mtcars)))
```



From this, it looks like mpg and am are both significantly correlated with wt, cyl, and disp, creating the possibility of bias if these variables are omitted. am is also significantly correlated with drat and gear, but this second group is less correlated with mpg, making it unlikely to cause bias.

Regression Modeling and Statistical Inference

First, create a linear model to regress for the relationship between mpg and am, with wt, cyl, and disp included as confounding variables.

```
mdl_awcd <- lm(mpg~am+wt+cyl+disp, mtcars)
summary(mdl_awcd)$coef
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 40.898313414 3.60154037 11.3557837 8.677574e-12
## am           0.129065571 1.32151163  0.0976651 9.229196e-01
## wt          -3.583425472 1.18650433 -3.0201537 5.468412e-03
## cyl         -1.784173258 0.61819218 -2.8861142 7.581533e-03
## disp         0.007403833 0.01208067  0.6128661 5.450930e-01
```

The coefficient for disp seems quite low, suggesting it may be unnecessary if it is sufficiently dependent on the other two confounders to not introduce its own effect. To test this, use ANOVA with various combinations of confounders.

```
mdl_aw <- lm(mpg~am+wt, mtcars)
mdl_awc <- lm(mpg~am+wt+cyl, mtcars)
anova(mdl_a, mdl_aw, mdl_awc, mdl_awcd)
```

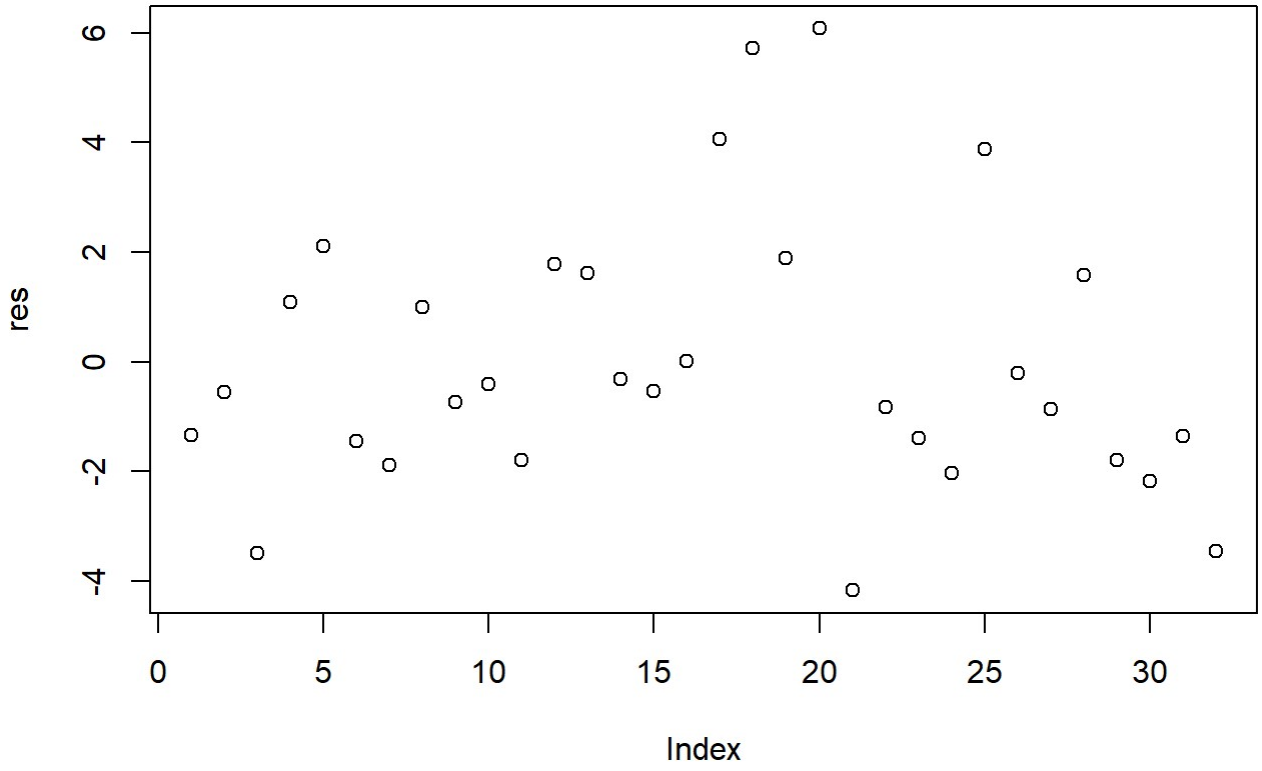
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + cyl
## Model 4: mpg ~ am + wt + cyl + disp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 63.4179 1.469e-08 ***
## 3     28 191.05  1     87.27 12.5055  0.001488 **
## 4     27 188.43  1      2.62  0.3756  0.545093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This result confirms that the am and wt confounders are necessary, but disp is not. Thus, to minimize variance inflation, use only wt and cyl as confounders for further analysis and visualization (mdl_awc).

Residual Diagnostics

Plot the residuals of the model for any signs that suggest an incorrect model.

```
res <- resid(mdl_awc)
plot(res)
```



Since there is no sign of heteroscedasticity or a visible trend, the model is likely acceptable.

Conclusions and Uncertainty

In the exploratory analysis, it initially seemed that transmission type had a major effect on mpg, since the coefficient was 7.2449393 with a p-value of 2.850207410^{-4}.

However, after checking for bias and removing the confounding effects of wt and cyl, the coefficient became 0.1764932, with a p-value of 0.8933421. This is neither strong nor significant. Thus, there is no evidence that transmission type affects gas mileage, even though it may appear so before removing bias.