Regression Models Assignment

Introduction

We will first do exploratory data analysis to find some relationship between the variables.

Then we will fit several linear models and analyse the quality of prediction, choose the best one and find the confidence interval for the predictor *am*.

Exploratory Data Analysis

The pair plots show that *mpg* has a strong correlation (> 0.8) with *cyl*, *disp* and *wt*. The correlation between *mpg* and *am* is only 0.6 which is not very significant, and also *am* has stronger correlation with *wt* and *drat*, indicating that *am* could be confounded by *wt* and *drat*.

The box plot shows that there is a difference between the distribution of *mpg* for different *am* value.

Another plot shows *mpg* by *am* treating *cyl* as confounder. It seems *mpg* and *am* are positively correlated in each *cyl* level.

Regression Analysis

First fit a model with all predictors except *am*.

```r
lm_fit_no_am <- lm(mpg ~ . - am, data = mtcars)
```

```
summary(lm_fit_no_am)
```

```
...
Residuals:
    Min      1Q  Median      3Q     Max
-2.9886 -1.6738 -0.3834  0.9796  5.4395
...
Residual standard error: 2.68 on 22 degrees of freedom
Multiple R-squared:  0.8596,  Adjusted R-squared:  0.8022
F-statistic: 14.97 on 9 and 22 DF,  p-value: 1.855e-07
```

Then fit a linear model with all predictors.

```r
lm_fit <- lm(mpg ~ ., data = mtcars)
```

```
summary(lm_fit)
```

```
...
Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
...
am           2.52023    2.05665   1.225   0.2340
...
Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Note p-value(am) = 0.2340 >> 0.05, indicating that *am* is unlikely to be significant in predicting *mpg*.

And here is the residual plot of the two models (ignore the third one for now), and a QQ plot shows that the residuals almost follow a normal distribution.

Performing anova on lm_fit_no_am and lm_fit,

```
anova(lm_fit_no_am, lm_fit)
## ...
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     22 158.04
## 2     21 147.49  1    10.547 1.5016  0.234
```

The F-value for lm_fit compared to lm_fit_no_am is 0.234 >> 0.05, indicating that *am* does not have strong influence on *mpg*.

Next, find the VIF of the predictors of lm_fit,

```
library(car)
vif(lm_fit)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

Finally, fit another model by taking out the top two predictors with VIF > 10.

```
lm_fit_adj_vif <- lm(mpg ~ . - disp - cyl, data = mtcars)
```

```
summary(lm_fit_adj_vif)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
...
am           2.42418    1.91227   1.268   0.2176
...
Residual standard error: 2.566 on 23 degrees of freedom
Multiple R-squared:  0.8655,  Adjusted R-squared:  0.8187
F-statistic:  18.5 on 8 and 23 DF,  p-value: 2.627e-08
```

We use the model with the highest adjusted r-squared value, i.e., lm_fit_adj_vif.

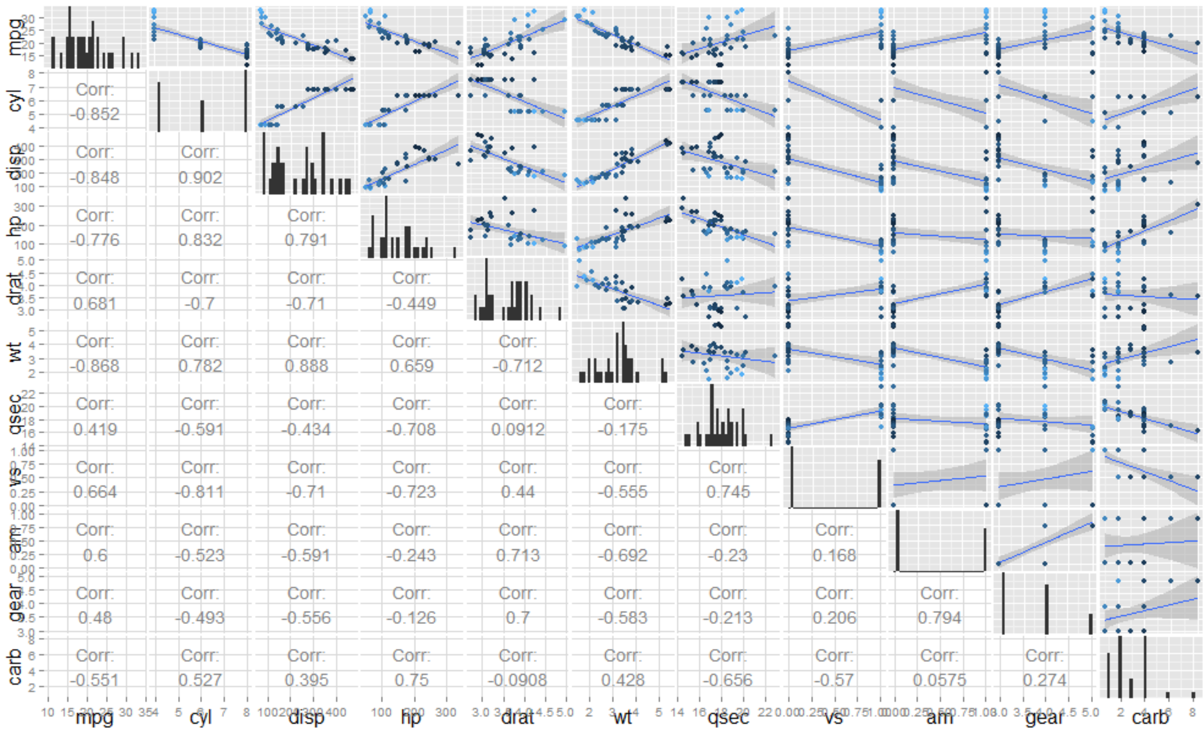
Assuming coef(am) follows a t-distribution, its α confidence interval is $[\hat{\Theta} - z\sigma, \hat{\Theta} + z\sigma]$, where $\hat{\Theta}$ = estimate mean = 2.42418, σ = standard error = 1.91227, $z = \Pr(t < \frac{\alpha}{2})$, and the degree of freedom is 23. For $\alpha = 0.05$, the interval is $[1.449181, 3.399179]$.

Overall, manual transmission seems to have a higher MPG on average, about 2.42mpg, with 95% confidence interval $[1.449181, 3.399179]$.

Appendix

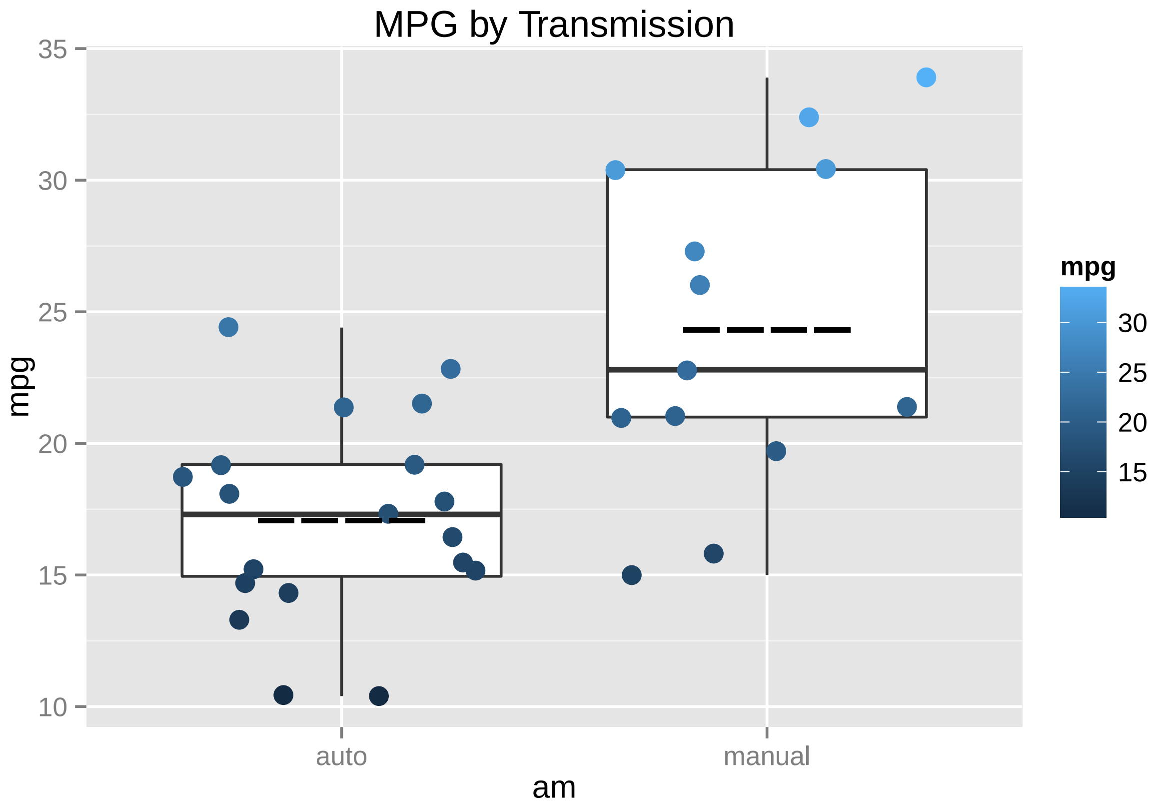
Pair plot of features

- Lighter color means higher mpg

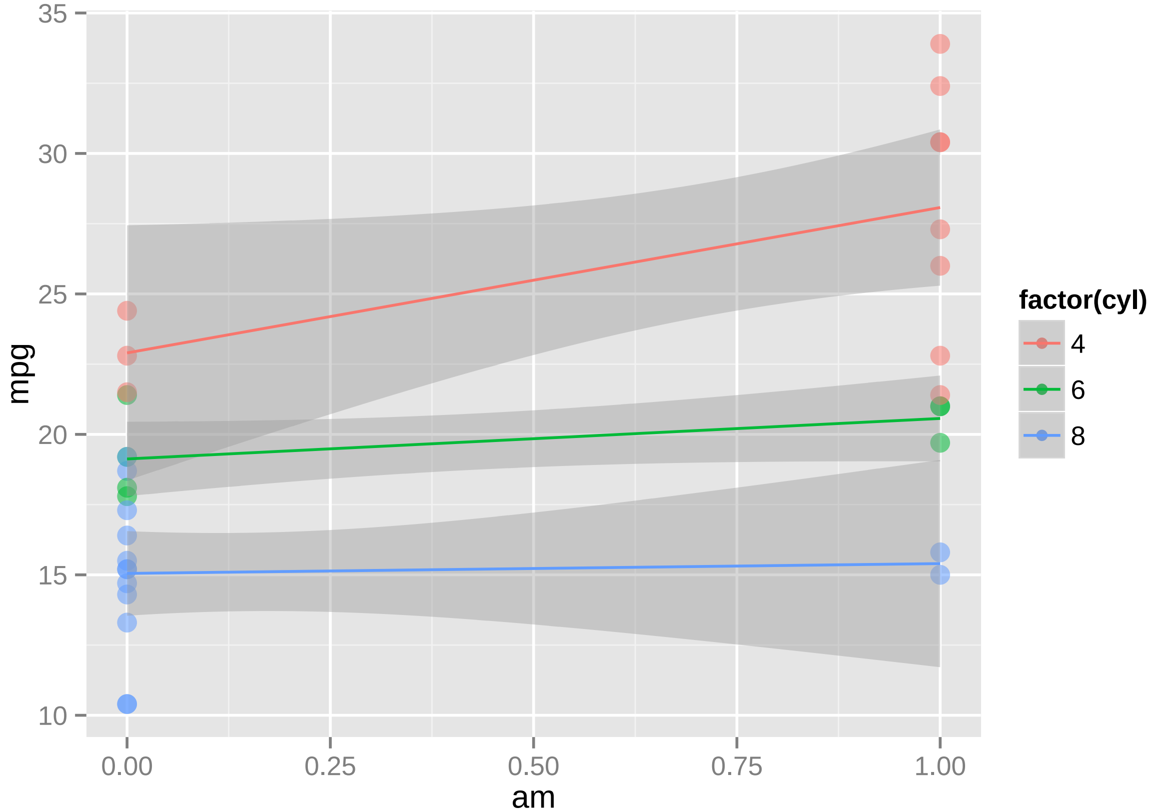


Box plot of MPG and Transmission

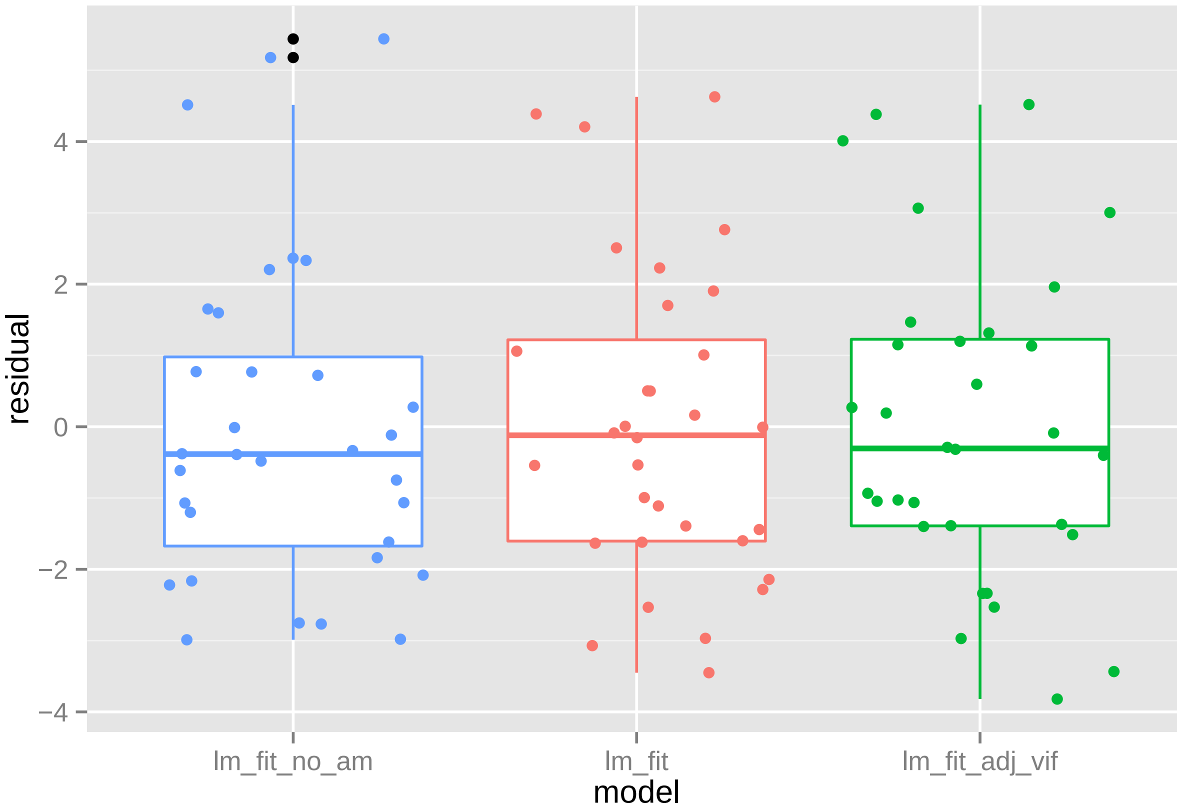
- ---- marks the mean


MPG by Transmission
35
30
25
20
15
10
mpg
auto
manual
am
mpg
30
25
20
15

MPG by Transmission across Number of Cylinders



Linear Regression Residual Plot



QQ Plot for Residuals