Regression Models - Coursera - Richard Hardy - 07/01/2020

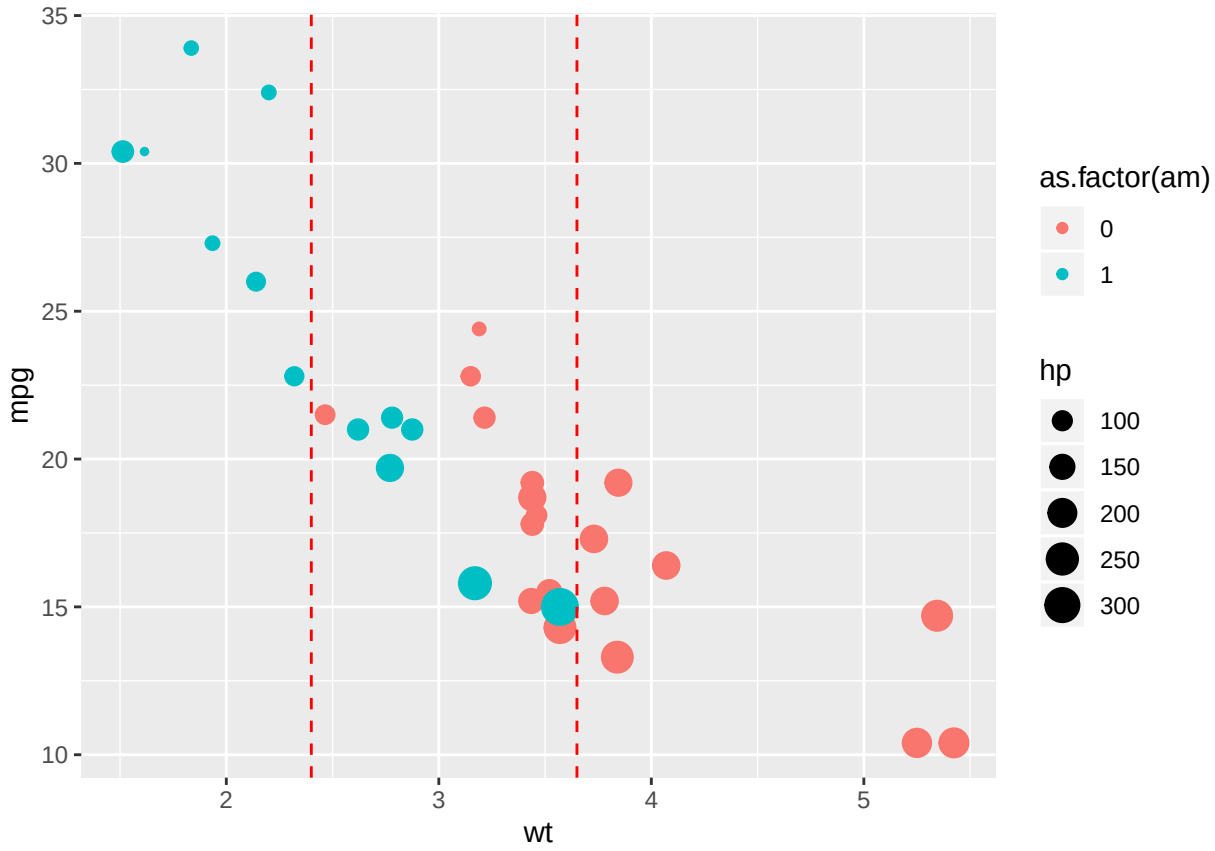
Executive Summary

utilising the `mtcars` data, exploratory analysis and linear models are used to:

- Quantify the MPG difference between automatic and manual transmissions
- determine whether automatic or manual transmission is better for MPG

Exploratory Analysis

```
plotfinal <- ggplot(mtcars, aes(y=mpg, x=wt, col=as.factor(am))) +
    geom_point(aes(size = hp)) +
    geom_vline(xintercept=2.4, col='red',linetype=2) +
    geom_vline(xintercept=3.65, col='red',linetype=2)
plotfinal
```



From the scatterplot, there is clear negative correlation between weight and mpg, and weight and transmission (manual=1 to automatic=0). And to a lesser extent, a positive correlation between weight and horsepower (see size of points increasing as weight does, cor=0.659). Cars with weight less than 2,465 lbs

are entirely manual, while the heaviest cars (>3,570lbs) are all automatic (dashed lines indicate visually, the weight boundaries within which meaningful comparisons between transmission and mpg could be made).

Given the strong association between weight and mpg, and the weaker evidence for an association between transmission and mpg **given** weight, there appears to be insufficient data to unpick an effect of transmission on mpg. However, if this sample is truly representative, any pair-wise effect of transmission on mpg might be explained entirely by weight.

Exploratory analysis shows that any pairwise effect of transmission on mpg will be confounded by the weight of the car. More light automatic transmission cars and more heavy manual transmission cars would be needed to more fully address the question.

Modelling

```
stepfit <- step(lm(data = mtcars, mpg ~ . ), trace=0)
summary(stepfit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

While stepwise selection has effectively narrowed the set of predictors amongst correlated possibilities and expectedly, highlighted weight as having the greatest effect on mpg, it does not:

- account for interaction effects.
- select variables based on practical considerations: Specifically, is acceleration pertinent as a feature? This seems more akin to an outcome caused by other factors (perhaps largely arising from a combination of weight and horsepower, whereby, light-weight, high-powered cars may accelerate more rapidly).

Given these considerations, the model is extended by re-introducing hp and adding an interaction term between weight and hp:

```
finalmod <- lm(mpg ~ am + hp + wt + qsec + hp:wt, data = mtcars)
summary(finalmod)$coef
```

```
##                Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) 35.93948381 10.265063345  3.5011458 0.0016912815
## am           0.91131006  1.400712804  0.6506045 0.5210102780
## hp          -0.09775746  0.029536329 -3.3097362 0.0027409939
## wt          -7.92300482  1.751086682 -4.5246217 0.0001178892
## qsec         0.59726932  0.392321888  1.5223961 0.1399807522
## hp:wt        0.02515920  0.008413321  2.9904001 0.0060241567
```

The effect estimate for `am` given by the final model is, with 95% confidence between -1.97 and 3.79, containing zero, with a coefficient estimate of 0.91 and $Pr(> |t|) > 0.52$. The model fails to reject the null hypothesis of no difference in mpg due to `am` at significance level $p < 0.05$ and we conclude there is no evidence for a difference in mpg due to transmission type, after accounting for the effects of weight $p < 0.001$ and gross horsepower $p < 0.01$.

Conclusions

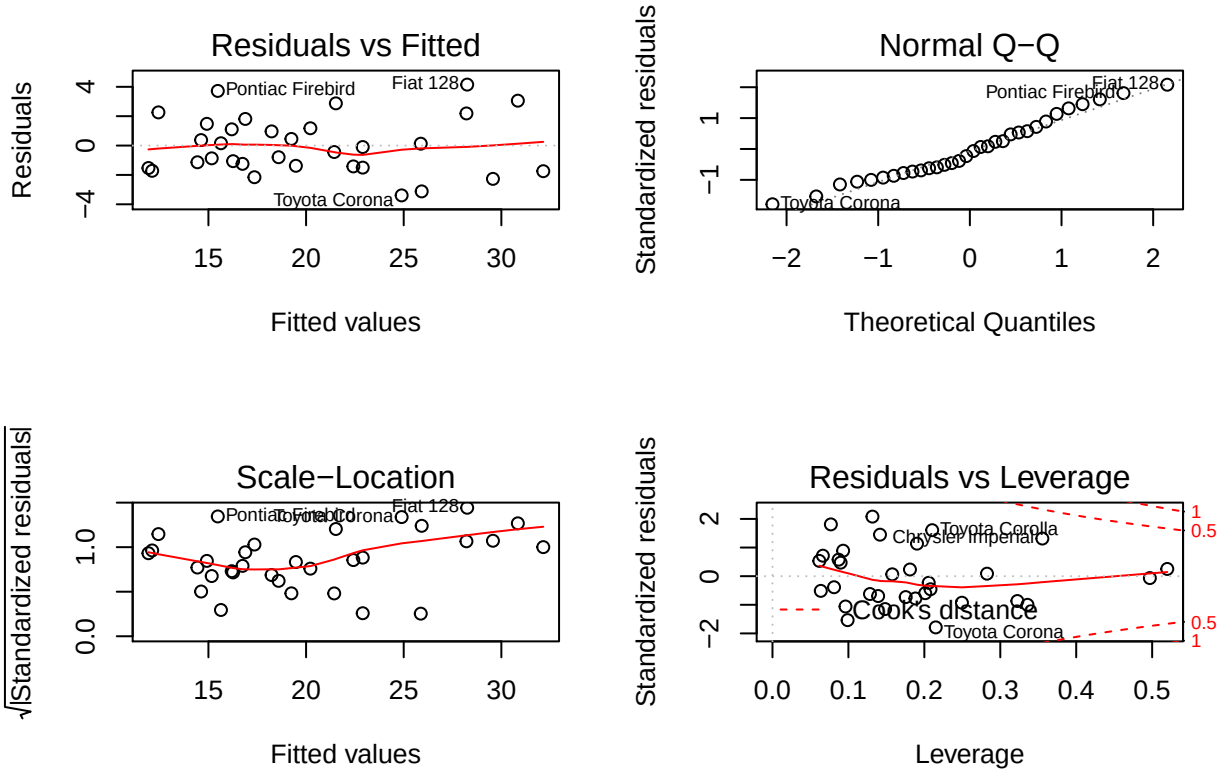
- Quantify the MPG difference between automatic and manual transmissions: According to the available data, the difference in MPG between transmission types after accounting for the confounding effects of weight and horsepower is likely 0.
- Determine whether automatic or manual transmission is better for MPG: There is no evidence in this data for a difference, any apparent difference is due predominantly to weight.

Given the strong relationship between car weight and mpg, data on manual transmission cars heavier than 3,570lbs and automatic cars lighter than 2,465lbs are required to more fully address the question.

Appendix

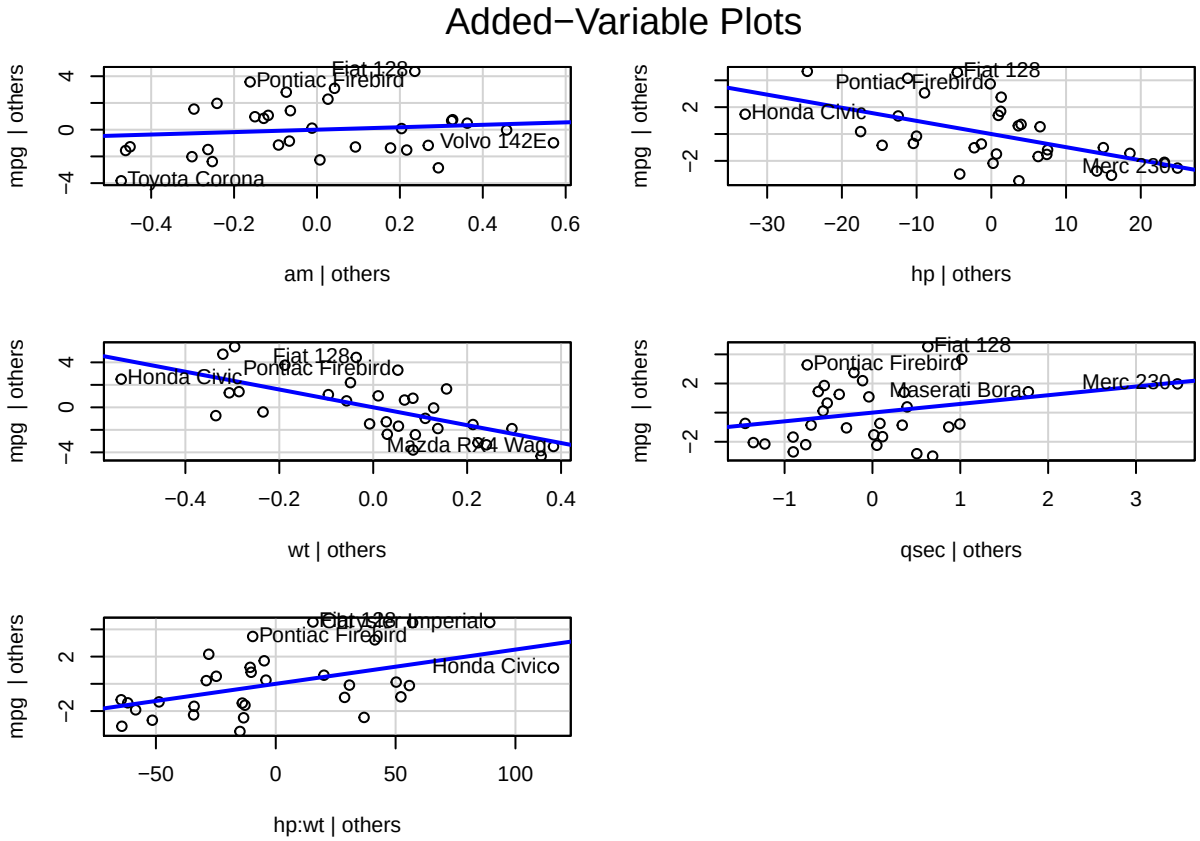
See model diagnostics:

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



The residuals vs fitted plot shows homoscedasticity indicating strong model fit. The QQ plot shows good fit across all data (with the exception of the Toyota Corona - an outlying variable by automatic transmission type. This might be addressed by excluding `am` from a predictive model for mpg, and we should, since it is not associated - but this is beyond the scope of the question) Scale-Location and Residuals vs Leverage plots show good fit, and no variables with a combination of high influence and leverage, respectively.

```
car::avPlots(finalmod)
```



Added-Variable plots indicate that all variables add information except for the `am` variable, supporting inclusion of `qsec` despite its insignificant coefficient p-value, and supporting removal of the `am` variable in further models to predict mpg.

```
vif(finalmod)
```

```
##        am        hp        wt      qsec     hp:wt
##  3.305969 27.752762 19.866182  3.325997 56.009267
```

High VIFs for wt, hp, and hp:wt are due to the inclusion of the interaction term. This is normal and expected, and does not present a problem if there is no multicollinearity with the variable of interest (am).

A model without the interaction term demonstrates this:

```
noint <- lm(mpg ~ am+hp+wt+qsec)
vif(noint)
```

```
##       am       hp       wt     qsec
## 2.541527 4.922129 3.964515 3.216021
```

Finally, the model R-squared is 0.89, explaining a large proportion of variation in mpg. Model R-squared for the initial stepwise model is 0.85.

The following plot illustrates the correlation structures between variables, pruned by the stepwise selection to form the initial model. According to the plot, beginning with a full model and using stepwise selection makes sense, since in general, each variable is correlated with mpg, as well as with each of the others. Adjusting for all variables and using stepwise removal was used to guide the exploratory analysis to focus on the most informative variables for mpg as included in the penultimate stepwise removal step: weight, hp, and qsec.

```
ggpairs(mtcars,
        lower = list(continuous="smooth"))
```