Motor trends: The effect of transmission type on fuel efficiency

John Lee

20 October 2015

criteria: Did the student interpret the coefficients correctly Did the student do some exploratory data analysis Did the student fit multiple models and detail their strategy for model selection Did the student answer the question or detail why the question was unanswerable Did the student do a residual plot and some diagnostics Did the student quantify the uncertainty in their conclusions and or perform an inference correctly Was the report brief Did the report include an executive summary Was the report done in rmd

Executive summary paragraph

main report (two pages only)

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

```
"Is an automatic or manual transmission better for MPG"
"Quantify the MPG difference between automatic and manual transmissions"
```

```
## Warning: replacing previous import by 'grid::arrow' when loading 'GGally'
```

```
## Warning: replacing previous import by 'grid::unit' when loading 'GGally'
```

The purpose of this analysis is to assess the effect of transmission type on the MPG of various cars. Without controlling for any other variables manual transmissions are typically more fuel efficent and getting more miles per gallon. Using tranmission type as a predictor for the mpg variable (see below) the mean mpg of automatic cars is 17.15 with a standard error of 1.1 and the average manual transmission is 7.2 mpg higher with a standard error of 1.8. A hypothesis test for the difference between the two categories yields a p-value < 0.001 suggesting that the transmission type strongly affects mpg; however, this description does not take into account the other confounding variables included in this dataset.

```
summary(lm(mpg~ am,data = mtcars))$coef
```

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

A model using all other variables as predictors suffers from variance inflation but indicates some of the variables that may be useful in predicting the mpg of the various cars. Indeed correlation plots using these variables demonstrate extensive correlation between many of these variables

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

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## amManual     2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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
```

```r
vif(fit_all)
```

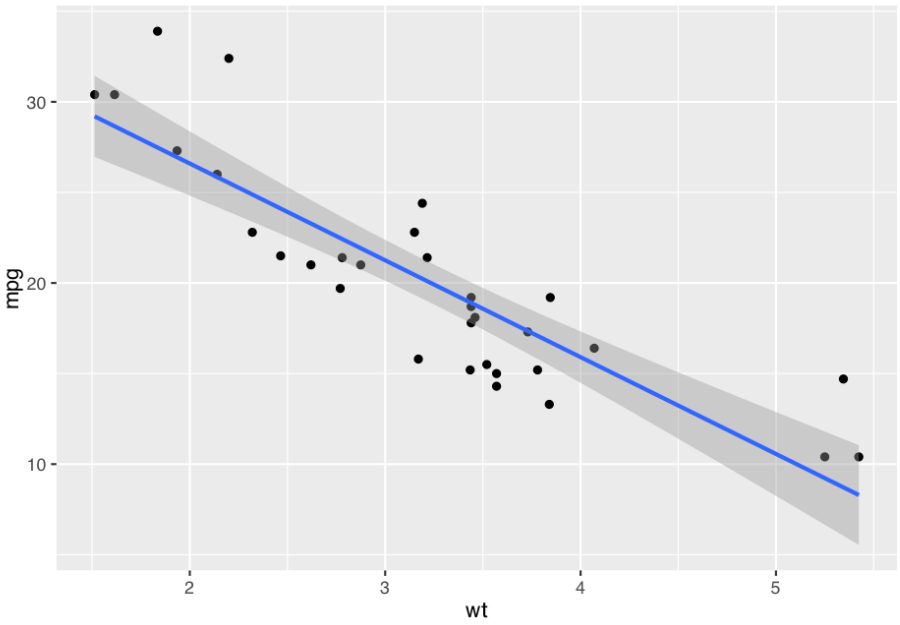
```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

Weight is inversely correlated with mpg and captures 75% of the variation in this variable. A scatterplot of these two variables with the superimposed regression line suggests that a linear relationship may not be the most accurate method of relating these two variables.

```r
fit_wt <- lm(data = mtcars, mpg ~ wt)
summary(fit_wt)[c( "coefficients","r.squared")]
```

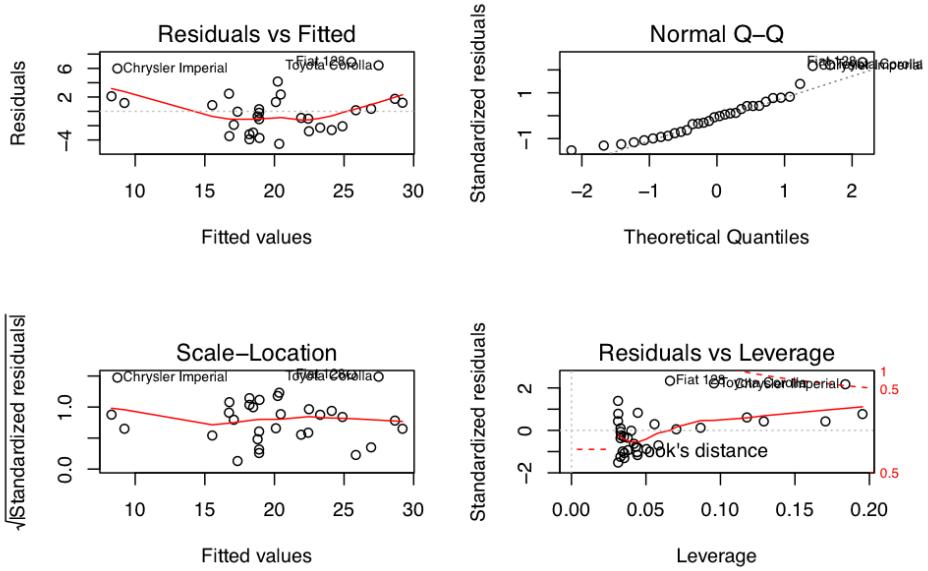
```
## $coefficients
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 37.285126   1.877627 19.857575 8.241799e-19
## wt          -5.344472   0.559101 -9.559044 1.293959e-10
##
## $r.squared
## [1] 0.7528328
```

```
ggplot(mtcars, aes(mpg,x = wt)) + geom_point() + geom_smooth(method = "lm")
```



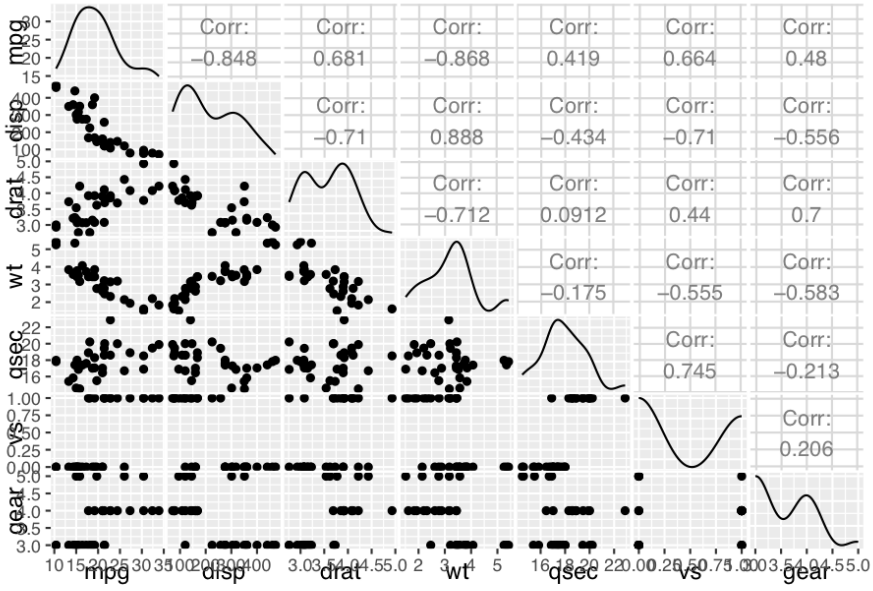
This observation is supported by

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



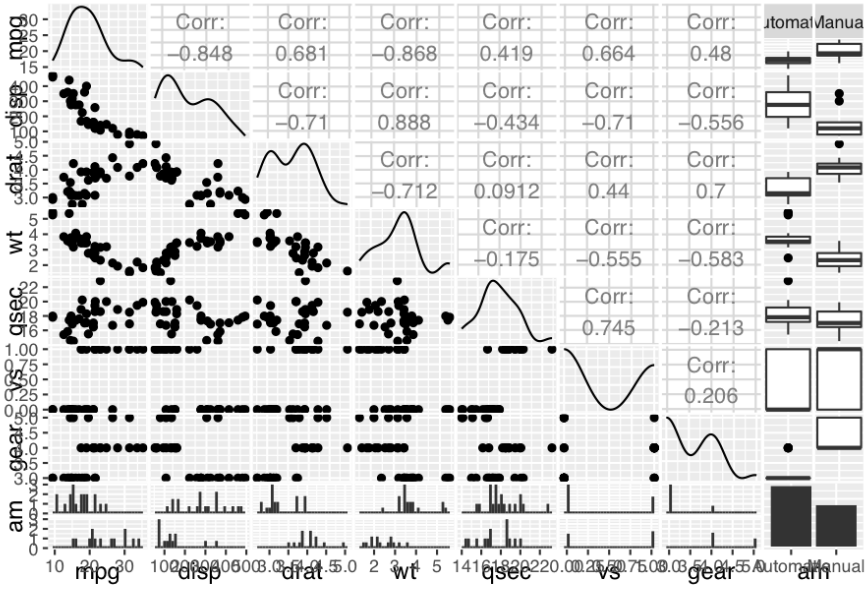
A number of details of the engine (number of carburettors, number of cylinders and displacement) are highly correlated with the power of the engine (see supplimentary figure 1). To simplify our analysis we only used the engine displacement as a measure of power fir each car. This variable is inversely correlated with mpg. A number of the other measures also showed correlations with the vehicle weight showing the strongest correlation:

```
ggpairs(select(mtcars, mpg,disp, drat,wt,qsec,vs,gear))
```




```
ggpairs(select(mtcars, mpg,disp, drat,wt,qsec,vs,gear,am))
```

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



In order to attain a parsimonious model we assessed whether transmission type still affected mpg when controlling for just weight and engine displacement:

```
fit <- lm(mpg~ wt + am, data = mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## amManual    -0.02362    1.54565  -0.015    0.988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

```r
vif(fit)
```

```
##       wt       am
## 1.921413 1.921413
```

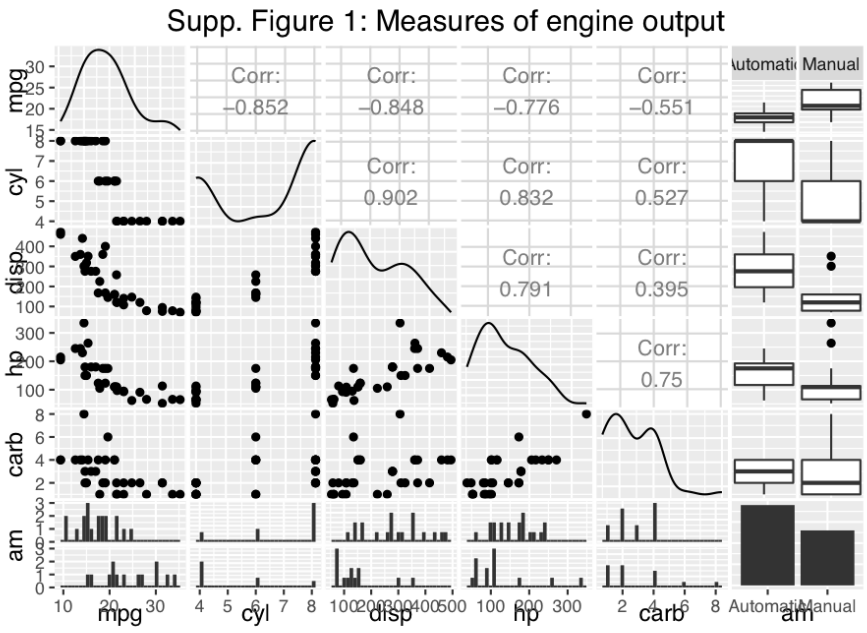
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Appendix

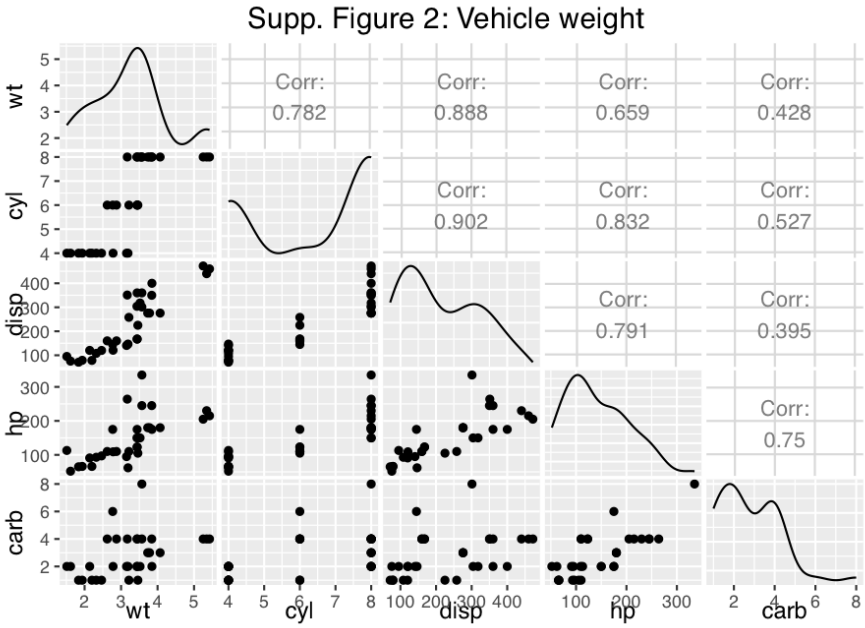
can only include figures

```r
ggpairs(select(mtcars,mpg,cyl,disp,hp,carb,am), title = "Supp. Figure 1: Measures of engine output")
```

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



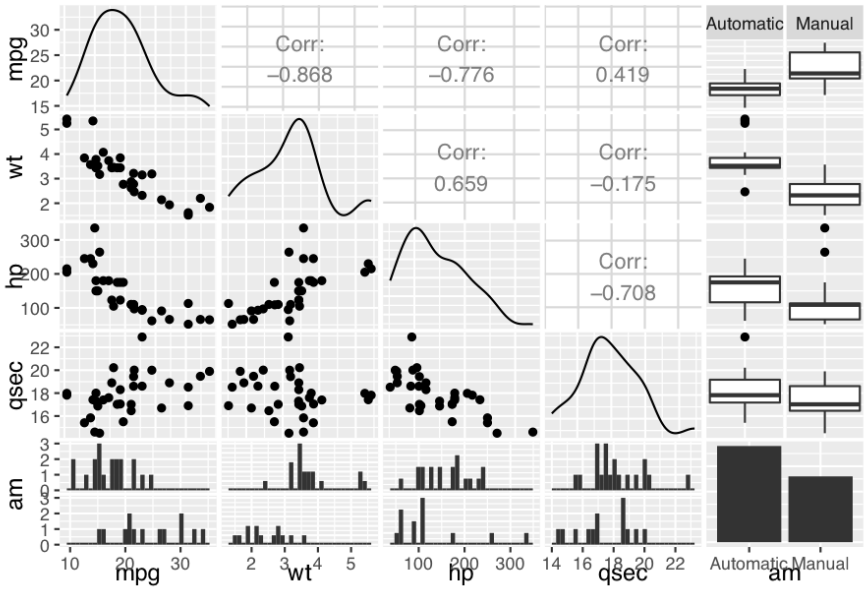

```r
ggpairs(select(mtcars,wt,cyl,disp,hp,carb), title = "Supp. Figure 2: Vehicle weight")
```




```
ggpairs(select(mtcars, mpg, wt, hp, qsec,  am), title = "Supp. Figure 3")
```

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

Supp. Figure 3




```
fit1 <- lm(data = mtcars, mpg~ disp)
fit2 <- update(fit1, mpg~ disp + carb)
fit4 <- update(fit1, mpg~ disp + carb + cyl)
fit3 <- update(fit1, mpg~ disp + carb + cyl+ hp)
anova(fit1,fit2,fit3,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ disp
## Model 2: mpg ~ disp + carb
## Model 3: mpg ~ disp + carb + cyl + hp
## Model 4: mpg ~ disp + carb + cyl
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     30 317.16
## 2     29 254.82  1    62.341 7.0515 0.01312 *
## 3     27 238.70  2    16.113 0.9113 0.41403
## 4     28 240.51 -1    -1.803 0.2040 0.65514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```