Regression models Project

Kyle Maurice

December 23, 2015

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(corrplot)
library(gridExtra)
```

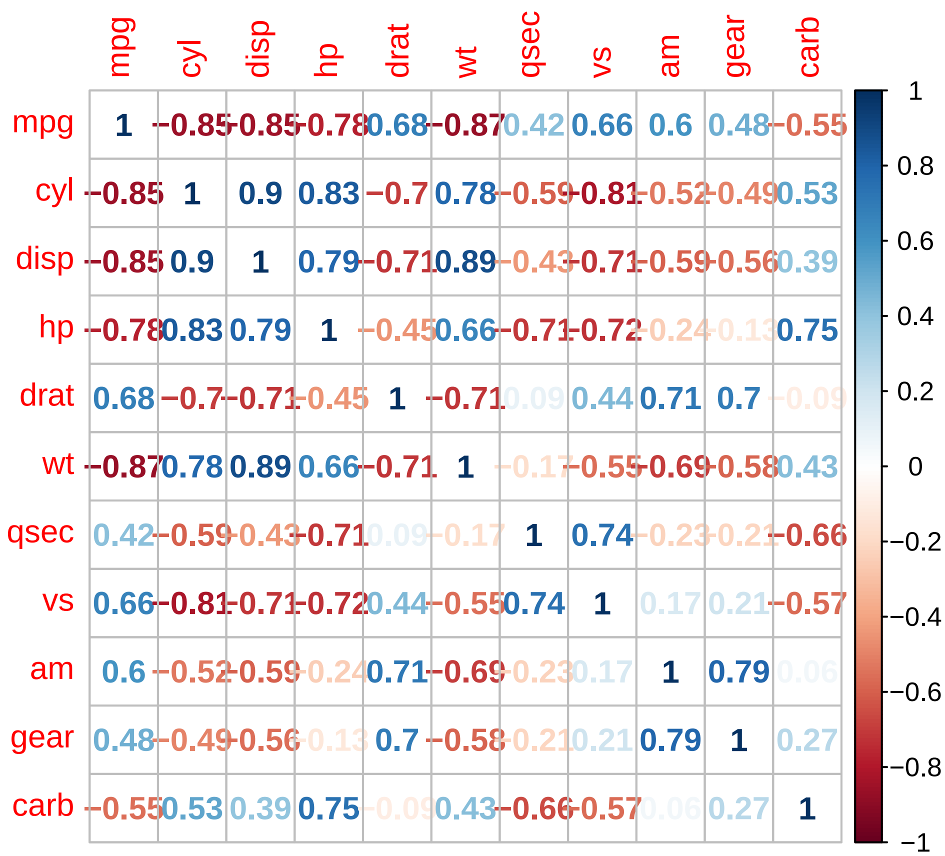
As highlighted in the table below, the types of cars represented in the data set range from american muscle and luxury cars to exotic sports cars to small economy cars. As a starting point in building the model to isolate the effect of transmission type, weight, quarter mile time, horsepower and displacement will be considered.

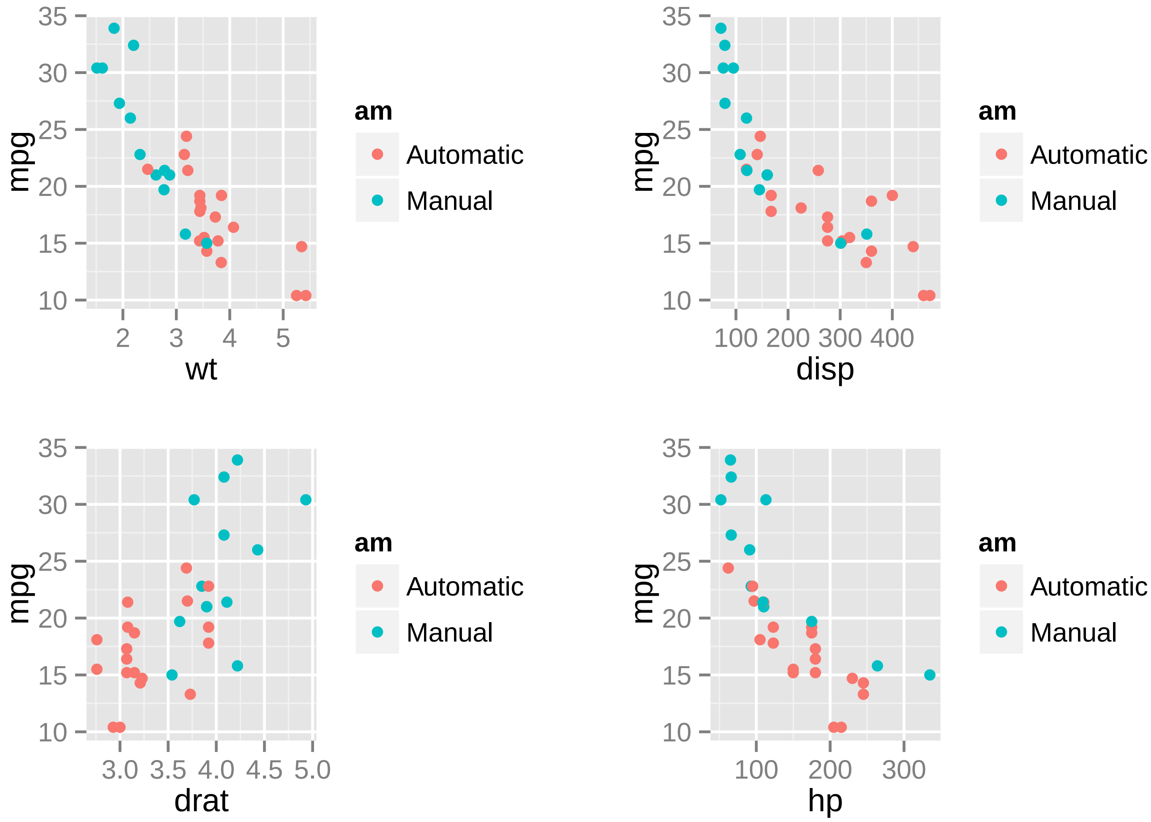
```r
mtcars
```

```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
```

```
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

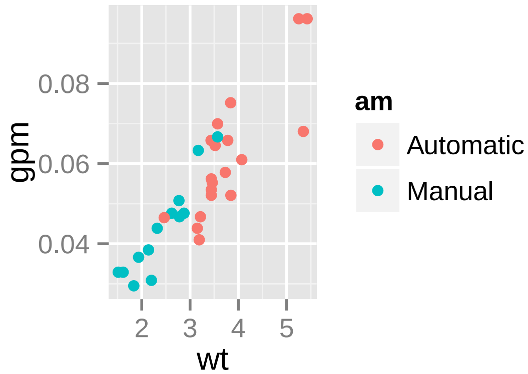
A few exploratory plots will be created to examine the relationships between the parameters of the data set.

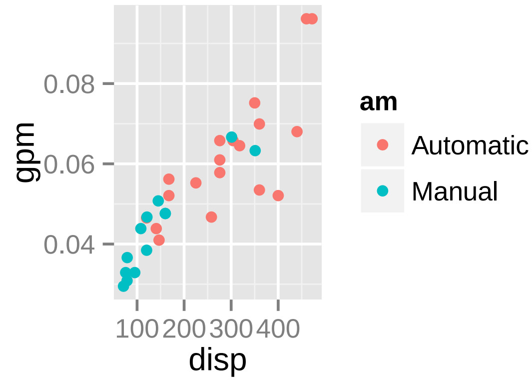


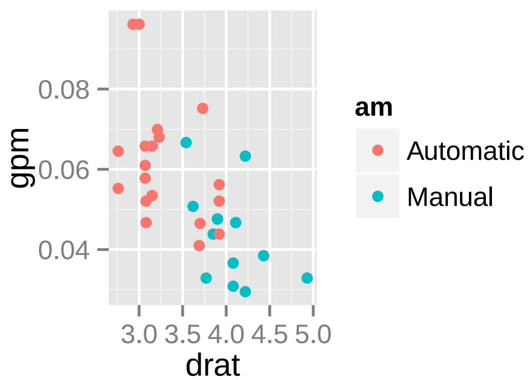


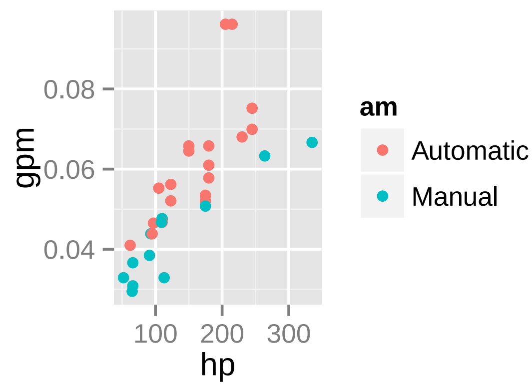
From the plots we can see a strong negative influence of displacement, cylinders, weight and horsepower. There is a weak to moderate positive influence of gears, am, qsec and vs on mpg. It is important to note that the plot of MPG with respect to weight, displacement and horsepower all look asymptotic at that approach the origin. The relationship looks like 1/x and a transformation might be helpful in removing the non linearity. As shown in the figure below, the relationships in the data are more linear when MPG is transgormed to GPM (i.e. GPM = 1/MPG).

```
mtcars$gpm <- 1/mtcars$mpg
p1<-ggplot(mtcars, aes(wt, gpm))+geom_point(aes(color=am))
p2<-ggplot(mtcars, aes(disp, gpm))+geom_point(aes(color=am))
p3<-ggplot(mtcars, aes(drat, gpm))+geom_point(aes(color=am))
p4<-ggplot(mtcars, aes(hp, gpm))+geom_point(aes(color=am))
grid.arrange(p1,p2,p3,p4, ncol=2)
```









The model should probably include all of these factors since it is not clear that they do not have an influence on the output variable that must be removed in order to examine the influence of am on mpg. The starting point will be a model with For this reason a model with all the parameters will be built and statistically insignificant factor in the models will be removed until the model is as simple as possible.

```
fit1 <- lm(I(gpm*100)~.-mpg, data = mtcars )
summary(fit1)
```

```
##
## Call:
## lm(formula = I(gpm * 100) ~ . - mpg, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70499 -0.33109  0.04737  0.38263  1.17856
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.751712   5.179898   1.110    0.279
## cyl         -0.182354   0.289195  -0.631    0.535
## disp         0.003999   0.004942   0.809    0.427
## hp           0.002876   0.006024   0.477    0.638
## drat        -0.063192   0.452565  -0.140    0.890
## wt           0.862705   0.524251   1.646    0.115
## qsec        -0.114173   0.202250  -0.565    0.578
## vs           0.090979   0.582392   0.156    0.877
```

```
## amManual     0.183750   0.569148   0.323    0.750
## gear        -0.463261   0.413238  -1.121    0.275
## carb         0.191364   0.229345   0.834    0.413
##
## Residual standard error: 0.7334 on 21 degrees of freedom
## Multiple R-squared:  0.8649, Adjusted R-squared:  0.8006
## F-statistic: 13.45 on 10 and 21 DF,  p-value: 5.15e-07
```

```r
fit2 <- update(fit1,.~. -drat)
summary(fit2)
```

```
##
## Call:
## lm(formula = I(gpm * 100) ~ cyl + disp + hp + wt + qsec + vs +
##     am + gear + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7184 -0.3323  0.0345  0.3757  1.1891
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.445921   4.588375   1.187   0.2479
## cyl         -0.171220   0.271719  -0.630   0.5351
## disp         0.003919   0.004798   0.817   0.4228
## hp           0.002950   0.005866   0.503   0.6201
## wt           0.875033   0.505116   1.732   0.0972 .
## qsec        -0.113147   0.197562  -0.573   0.5726
## vs           0.088560   0.569014   0.156   0.8777
## amManual     0.171320   0.549474   0.312   0.7581
## gear        -0.467583   0.402789  -1.161   0.2581
## carb         0.184755   0.219350   0.842   0.4087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7169 on 22 degrees of freedom
## Multiple R-squared:  0.8648, Adjusted R-squared:  0.8095
## F-statistic: 15.64 on 9 and 22 DF,  p-value: 1.253e-07
```

```r
fit3 <- update(fit2,.~. -vs)
summary(fit3)
```

```
##
## Call:
## lm(formula = I(gpm * 100) ~ cyl + disp + hp + wt + qsec + am +
##     gear + carb, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71787 -0.34588  0.03216  0.38783  1.19118
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  5.382863   4.472450   1.204   0.2410
## cyl         -0.185699   0.249822  -0.743   0.4648
## disp         0.003846   0.004673   0.823   0.4189
## hp           0.003197   0.005526   0.579   0.5685
## wt           0.867868   0.492228   1.763   0.0912 .
## qsec        -0.101939   0.180024  -0.566   0.5767
## amManual     0.153556   0.525963   0.292   0.7729
## gear        -0.464703   0.393736  -1.180   0.2500
## carb         0.181734   0.213805   0.850   0.4041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7015 on 23 degrees of freedom
## Multiple R-squared:  0.8647, Adjusted R-squared:  0.8176
## F-statistic: 18.37 on 8 and 23 DF,  p-value: 2.812e-08
```

Notice we would get rid of transmission type here if we were not intested in the effect of transission type.

```
fit4 <- update(fit3,.~. -qsec)
summary(fit4)
```

```
##
## Call:
## lm(formula = I(gpm * 100) ~ cyl + disp + hp + wt + am + gear +
##     carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6672 -0.3373  0.0484  0.3819  1.1416
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.192787   2.213856   1.442   0.1622
## cyl         -0.120989   0.218985  -0.552   0.5857
## disp         0.004781   0.004309   1.109   0.2782
## hp           0.003221   0.005447   0.591   0.5598
## wt           0.721273   0.412679   1.748   0.0933 .
## amManual     0.270939   0.476498   0.569   0.5749
## gear        -0.448591   0.387108  -1.159   0.2579
## carb         0.222806   0.198258   1.124   0.2722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6915 on 24 degrees of freedom
## Multiple R-squared:  0.8628, Adjusted R-squared:  0.8227
## F-statistic: 21.55 on 7 and 24 DF,  p-value: 6.691e-09
```

```
fit5 <- update(fit4,.~. -cyl)
summary(fit5)
```

```
##
## Call:
```

```
## lm(formula = I(gpm * 100) ~ disp + hp + wt + am + gear + carb,
##     data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61509 -0.40793  0.09338  0.30810  1.23189
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.318418   1.526436   1.519   0.1413
## disp         0.003465   0.003542   0.978   0.3372
## hp           0.002744   0.005303   0.518   0.6093
## wt           0.806330   0.377530   2.136   0.0427 *
## amManual     0.300139   0.466931   0.643   0.5262
## gear        -0.352483   0.340984  -1.034   0.3112
## carb         0.172769   0.173896   0.994   0.3300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6818 on 25 degrees of freedom
## Multiple R-squared:  0.861,  Adjusted R-squared:  0.8277
## F-statistic: 25.81 on 6 and 25 DF,  p-value: 1.439e-09
```

```
fit6 <- update(fit5,.~. -hp)
summary(fit6)
```

```
##
## Call:
## lm(formula = I(gpm * 100) ~ disp + wt + am + gear + carb, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59096 -0.34729  0.07287  0.35535  1.13407
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.451023   1.483438   1.652   0.1105
## disp         0.004894   0.002187   2.238   0.0340 *
## wt           0.726086   0.339338   2.140   0.0419 *
## amManual     0.290420   0.459937   0.631   0.5333
## gear        -0.346610   0.335962  -1.032   0.3117
## carb         0.237034   0.120008   1.975   0.0590 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6722 on 26 degrees of freedom
## Multiple R-squared:  0.8595, Adjusted R-squared:  0.8325
## F-statistic: 31.82 on 5 and 26 DF,  p-value: 2.719e-10
```

```
fit7 <- update(fit6,.~. -gear)
summary(fit7)
```

```
##
```

```
## Call:
## lm(formula = I(gpm * 100) ~ disp + wt + am + carb, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60219 -0.30124  0.00717  0.42751  1.11118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.153126   0.787038   1.465   0.1544
## disp        0.005434   0.002126   2.556   0.0165 *
## wt          0.785091   0.334883   2.344   0.0267 *
## amManual    0.056008   0.400373   0.140   0.8898
## carb        0.166173   0.098526   1.687   0.1032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.673 on 27 degrees of freedom
## Multiple R-squared:  0.8538, Adjusted R-squared:  0.8321
## F-statistic: 39.41 on 4 and 27 DF,  p-value: 6.689e-11
```

```r
fit8 <- update(fit7,.~. -carb)
summary(fit8)
```

```
##
## Call:
## lm(formula = I(gpm * 100) ~ disp + wt + am, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78089 -0.28621  0.07925  0.43032  0.97729
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.678079   0.758734   0.894   0.3791
## disp        0.005425   0.002195   2.472   0.0198 *
## wt          1.032462   0.310814   3.322   0.0025 **
## amManual    0.421606   0.347527   1.213   0.2352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6948 on 28 degrees of freedom
## Multiple R-squared:  0.8384, Adjusted R-squared:  0.8211
## F-statistic: 48.41 on 3 and 28 DF,  p-value: 3.315e-11
```

```r
fit9 <- update(fit8,.~. -carb)
summary(fit9)
```

```
##
## Call:
## lm(formula = I(gpm * 100) ~ disp + wt + am, data = mtcars)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.78089 -0.28621  0.07925  0.43032  0.97729
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.678079   0.758734   0.894   0.3791
## disp        0.005425   0.002195   2.472   0.0198 *
## wt          1.032462   0.310814   3.322   0.0025 **
## amManual    0.421606   0.347527   1.213   0.2352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6948 on 28 degrees of freedom
## Multiple R-squared:  0.8384, Adjusted R-squared:  0.8211
## F-statistic: 48.41 on 3 and 28 DF,  p-value: 3.315e-11
```

```r
anova(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8)
```

```
## Analysis of Variance Table
##
## Model 1: I(gpm * 100) ~ (mpg + cyl + disp + hp + drat + wt + qsec + vs +
##     am + gear + carb) - mpg
## Model 2: I(gpm * 100) ~ cyl + disp + hp + wt + qsec + vs + am + gear +
##     carb
## Model 3: I(gpm * 100) ~ cyl + disp + hp + wt + qsec + am + gear + carb
## Model 4: I(gpm * 100) ~ cyl + disp + hp + wt + am + gear + carb
## Model 5: I(gpm * 100) ~ disp + hp + wt + am + gear + carb
## Model 6: I(gpm * 100) ~ disp + wt + am + gear + carb
## Model 7: I(gpm * 100) ~ disp + wt + am + carb
## Model 8: I(gpm * 100) ~ disp + wt + am
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     21 11.296
## 2     22 11.306 -1  -0.01049 0.0195 0.8903
## 3     23 11.318 -1  -0.01245 0.0231 0.8805
## 4     24 11.476 -1  -0.15779 0.2934 0.5938
## 5     25 11.622 -1  -0.14596 0.2714 0.6079
## 6     26 11.747 -1  -0.12451 0.2315 0.6354
## 7     27 12.228 -1  -0.48089 0.8940 0.3551
## 8     28 13.516 -1  -1.28823 2.3950 0.1367
```