Regression models class project

Yevgen Yampolskiy

01/24/2015

Result summary

We observe significant difference in MPG for automatic and manual transmission. Most likely reason is that manual transmission weights less, and it is the weight impacts MPG.

Suggested model for MPG is `mpg ~ wt + cyl*hp`.

Comparison of manual and automatic transmission for MPG

Just by looking at the box plots of MPG by transmission type we see that cars with manual transmission have higher MPG. T-test confirms that the difference in mean MPG is significant. However it is more natural to expect that weight will impact MPG a lot, and indeed we can see very high correlation between MPG and weight, and not so high correlation between MPG and transimission type. If we plot MPG by weight and add coloring based on transmission type then we can obseved that cars with manual transmission weight less which explains observed difference in MPG by transmission type.

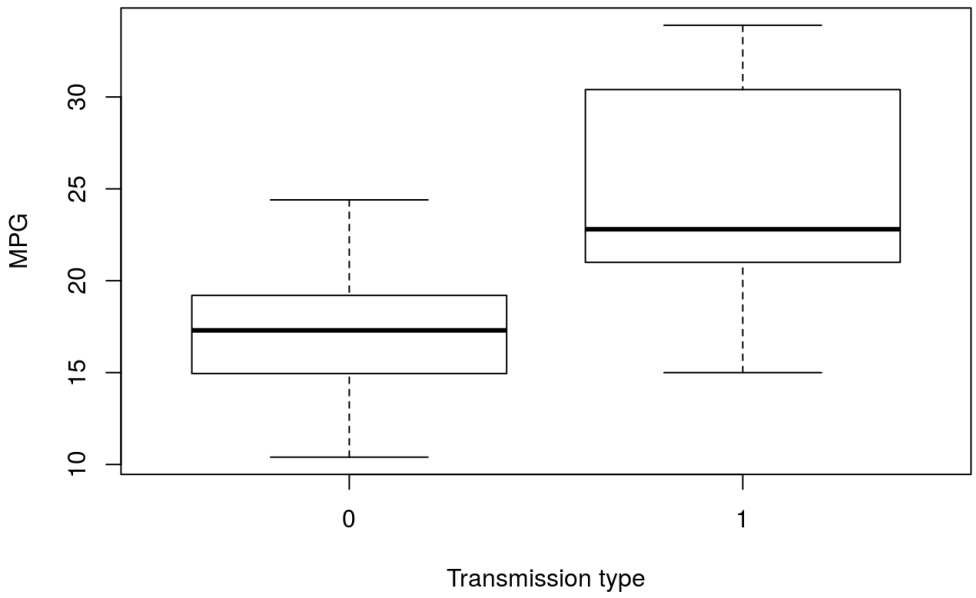
```
data(mtcars)
boxplot(mpg ~ am, data = mtcars, xlab="Transmission type", ylab="MPG")
t.test(subset(mtcars, am==0, select=mpg), subset(mtcars, select=mpg, am==1))
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(mtcars, am == 0, select = mpg) and subset(mtcars, select = mpg, am == 1)
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

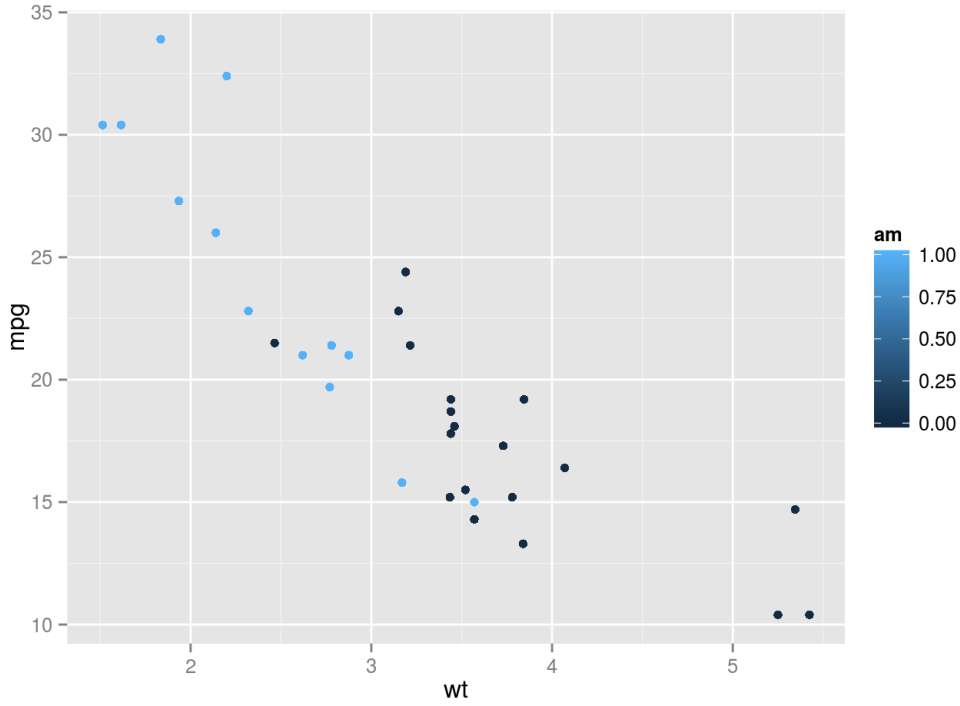
```
# print correlations of MPG with transmission type and weight
c(cor(mtcars$mpg, mtcars$am), cor(mtcars$mpg, mtcars$wt))
```

```
## [1]  0.5998324 -0.8676594
```

```
#We can see that cars with manual transmission are more lightweight
library(ggplot2)
```




```
qplot(wt, mpg, data = mtcars, col = am)
```



Choosing the model

If we feed all available variables into the model and let `R` choose the best one using `step` function then resulting model will be model `mpg ~ wt + qsec + am`. What I do not like about this model is that `qsec` is a derived variable from engine characteristic and weight, and `am` is likely related to weight. Another issue is that using 3 variables on data set with 32 observations could result into overfitting.

If we try to model MPG using only two variables (one of them must be weight) then by looking at the adjusted `R^2` values the most attractive models are `mpg ~ wt + cyl`, `mpg ~ wt + hp`, `mpg ~ wt + qsec`. I like `cyl` variable in the model as it is directly related to car physical characteristic. There is a high correlation between `cyl` and `hp`. Adding interactions between `cyl` and `hp` resulted into the model with the smallest `R^2`: `mpg ~ wt + cyl * hp`.

Residual plot does not show anything bad about this model. Note that if we compare `mpg ~ wt + qsec + am` and `mpg ~ wt + cyl * hp` models than we reduced both residual error and number of degrees of freedom.

Here is supporting `R` code:

```
data(mtcars)
```

```r
lm1 <- lm(mpg ~ ., data = mtcars)
# Let R pick a model
step(lm1)
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - cyl   1    0.0799 147.57 68.915
## - vs    1    0.1601 147.66 68.932
## - carb  1    0.4067 147.90 68.986
## - gear  1    1.3531 148.85 69.190
## - drat  1    1.6270 149.12 69.249
## - disp  1    3.9167 151.41 69.736
## - hp    1    6.8399 154.33 70.348
## - qsec  1    8.8641 156.36 70.765
## <none>              147.49 70.898
## - am    1   10.5467 158.04 71.108
## - wt    1   27.0144 174.51 74.280
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## - wt    1   27.0280 174.60 72.297
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## - wt    1   27.3799 175.22 70.410
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
##
## Step:  AIC=62.16
```

```
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## - qsec  1    20.225 180.29 63.323
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Coefficients:
## (Intercept)           wt         qsec           am
##       9.618       -3.917        1.226        2.936
```

```
# Model choosen: mpg ~ wt + qsec + am
summary(lm(formula = mpg ~ wt + qsec + am, data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
# Suggested model
model <- lm(formula = mpg ~ wt + cyl * hp, data = mtcars)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + cyl * hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3440 -1.4144 -0.6166  1.2160  4.2815
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 52.017520   4.916935  10.579 4.18e-11 ***
## wt          -3.119815   0.661322  -4.718 6.51e-05 ***
## cyl         -2.742125   0.800228  -3.427  0.00197 **
## hp          -0.163594   0.052122  -3.139  0.00408 **
## cyl:hp       0.018954   0.006645   2.852  0.00823 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.242 on 27 degrees of freedom
## Multiple R-squared:  0.8795, Adjusted R-squared:  0.8616
## F-statistic: 49.25 on 4 and 27 DF,  p-value: 5.065e-12
```

```
plot(resid(model))
abline(a=0, b=0)
```