Regression Models Course Project

Paul Clark: March 5, 2017

1 Executive Summary

For its 1974 edition, US magazine *Motor Trend* has asked two questions to be addressed using data on fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). **Quesiton 1:** *"Is an automatic or manual transmission better for MPG?"* **Question 2:** *"How does MPG differ, quantitatively, between automatic and manual transmissions?"* The variables are described below:

```
 [1] "mpg  Miles/(US) gallon"          "cyl  Num cylinders"
 [3] "disp Displacement (cu.in.)"      "hp   Gross horsepower"
 [5] "drat Rear axle ratio"            "wt   Weight (1000 lbs)"
 [7] "qsec 1/4 mile time"              "vs   Engine(0=V,1=Inline)"
 [9] "am   Transmission(0=auto,1=man)" "gear Num forward gears"
[11] "carb Num carburetors"
```

We address the questions via two approaches: **Approach (a)** - *Calculation of mean differenc and two-sample T-test on `mpg` of the transmission types.* **Approach (b)** - *Adjustment of the `am` effect in (a) by regressing `mpg` on all other variables in the data.* Note: for either approach to be meaningful, this small sample of 32 cars must be representative of their populations.

We conclude from **Approach (a)** that the two groups of cars come from populations with statistically different means. If the data provided is representative, then *manual* transmission cars generally have an estimated mean `mpg` **7.2 MPG** higher than *automatic* cars.

From **Approach (b)**, we have two main conclusions: **(1)** the adjustments of the `am` effect that best predict `mpg` are `hp` and `wt`, both negative, corresponding to model $\mathbf{E}[\mathbf{mpg}] = \mathbf{M} \cdot am + \mathbf{H} \cdot hp + \mathbf{W} \cdot wt$; **(2)** the *automatic/manual* effect M, with *horsepower* and *weight* held fixed, is approximately **2 MPG**, with manual again having higher `mpg` than automatic. Adjustment due to other effects reduces the independent impact of `am` by over 5 MPG!

2 Exploratory Data Analysis

In **Figure 1**, we examine integer predictors to decide whether to treat as factors. We find value in treating `cyl` and `carb` as continuous: they show clear trends vs. other variables. From a pairs plot, **Figure 2**, we see many strong correlations, so model selection should consider variance inflation.

3 Approach (a): Two sample t-test and Inference

Figure 3 shows that in this data, `mpg` varies with `am`. Mean mileage for manual is **7.2 mpg** higher than automatic. We compute p-value and conf interval for the test of manual transmission greater.

```
t_test<-t.test(mpg~am,data=mtcars,alternative='less') # less: 2nd factor is t.test base
# note: code for processing and formatting of output suppressed
```

```
p-value = 0.07%     95% conf.int = 3.91 to Inf
```

Inference: Given the p-value, we are highly confident manual is associated with higher fuel economy in the populations from which these samples were drawn.

4 Approach (b): OLS regression

Approach **(b)** is under-specified. Having identified `mpg` and `am` as of interest, the "correct" choice of model still depends on selection of the appropriate subset of 9 other variables. This should be a function of variable/model significance, but also *Motor Trend's* interests. For significance testing, manual checking of

p-values is in-viable: at least $2^9 = 512$ models with single predictors exist. Therefore, to fully specify and make the approach manageable, we make additional assumptions.

We assume *Motor Trend* values: **(A)** Parsimony/simplicity; **(B)** Models with granular, causal variables that may clarify engineering trade-offs; **(C)** Predictiveness: good generalizability outside the training sample.

4.1 Model Search

Due to **(A)**, we consider no interactions. From **(B)**, we exclude `qsec`, a summary metric. Due to **(C)**, we rank models using the **AIC** metric, which estimates model predictiveness outside the training sample. For OLS regression, the metric is $n \cdot Log(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2) + 2k$, where $k = \#$ of parameters, an overfitting penalty. In fact, we use **AICc**, which corrects the penalty to be greater for small n. We use automated search to make evaluation of all 2^9 models feasible. Models are ranked from smallest to largest AICc. Only non-zero coefficients are shown.

```
if (!"MuMIn" %in% row.names(installed.packages())) {install.packages("MuMIn")}
library(MuMIn); mtcars$qsec <- NULL; mtcars$gear <- as.factor(mtcars$gear)
globalmodel <- lm(mpg ~ ., data = mtcars, na.action = na.fail)
bestmodels <- dredge(globalmodel, subset = ~ am) # only considers models with `am`
bestmodels[1:5,]
```

```
    (Int)  am carb  cyl    hp  vs   wt df logLik  AICc delta weight
322  34.0 2.1           -0.04     -2.9  5  -73.1 158.4   0.0   0.31
264  36.9 1.8 -0.7 -1.2           -2.5  6  -72.0 159.4   0.9   0.20
450  31.1 2.4           -0.03 1.8 -2.6  6  -72.0 159.4   1.0   0.19
326  36.1 1.5      -0.7 -0.02     -2.6  6  -72.1 159.6   1.2   0.17
262  39.4 0.2      -1.5           -3.1  5  -74.0 160.3   1.9   0.12
```

4.2 Model Inference

We investigate values of the `am` coefficient for the top models. Note the first 3 all round to 2, suggesting this is a good rough estimate of the adjusted transmission effect. Although our top model is only one AICc point lower than the next best model (model averaging is suggested via the weights, for differences less than 2), we focus attention on it, in the spirit of Assumption **(A)**.

```
            Estimate Std. Error
(Intercept)    34.00       2.64
am              2.08       1.38
hp             -0.04       0.01
wt             -2.88       0.90
```

The table contains no p-values, as after a search of 2^9 models, these would be inflated (we expect 'good' p-values often, by chance alone). However, standard errors are provided. These indicate `hp` and `wt` are relatively significant, whereas significance of the the `am` coefficient, 2.1, is low. The Estimate divided by the Std. Error, typically used as the t-statistic, is only 1.5. Two is typically the significance threshold. However, the R^2 of the top model is 84%: it explains a high degree of sample variance with only 3 covariates.

4.3 Model Diagnostics

We run base R's standard diagnostic plots, **Figure 4**. Though the smoother line in the *Residuals vs Fitted* plot shows curvature, pointing to possible quadratic terms, the trend is not pronounced except for the 3 labeled points (*Toyota Corolla, Fiat 128,* and *Chrysler Imperial*). These have notably higher MPG than the trend. The *Normal Q-Q* plot shows a somewhat right skewed distribution. But the deviations are not extreme, except for the 3 labeled points, and the lowest. The lowest, given by the code below, is *Mazda RX4*.

```
bestmodel<-lm(mpg~am+hp+wt,mtcars); row.names(mtcars)[which.min(bestmodel$residuals)]
```

The *Scale-Location* plot shows no strong heteroskedasticity. Cars getting 25-30 MPG are few, therefore the smoother line has large standard error there. In the *Residuals vs Leverage* plot, all points are inside of 0.5 in *Cook's distance*, indicating relative stability of the βs. Finaly, from package `car`, we use `vif()` to calculate the Variance Inflation Factors, to evaluate collinearity. They are all under 4, causing no reason for alarm (code suppressed to conserve space). VIFs:

```
##       am       hp       wt
## 2.271082 2.088124 3.774838
```

5 Figures

Plot provide guidance on which integer variables can be treated as continuous.

```
data(mtcars)
intplotsdf <- data.frame(mpg=mtcars$mpg, hp=mtcars$hp, cyl=mtcars$cyl, carb=mtcars$carb,
              gear=mtcars$gear, qsec=mtcars$qsec)
if (!"tidyr" %in% rownames(installed.packages())) {install.packages("tidyr")}
library(tidyr)
intplotsdf <- gather(data = intplotsdf, key = x_variable,
                     value = x_value, -mpg, -hp, -qsec)
intplotsdf <- gather(data = intplotsdf, key = y_variable,
                     value = y_value, -x_variable, -x_value)
if (!"ggplot2" %in% rownames(installed.packages())) {install.packages("ggplot2")}
library(ggplot2)
g_integer_vars <- ggplot(intplotsdf, aes(x=x_value,y=y_value)) +
                        facet_grid(y_variable ~ x_variable, scales = "free") +
                  geom_point() + geom_smooth(method = "lm") +
labs(title = "'gear' behaves as a factor, 'carb' and 'cyl' can be treated as continous")
print(g_integer_vars)
```

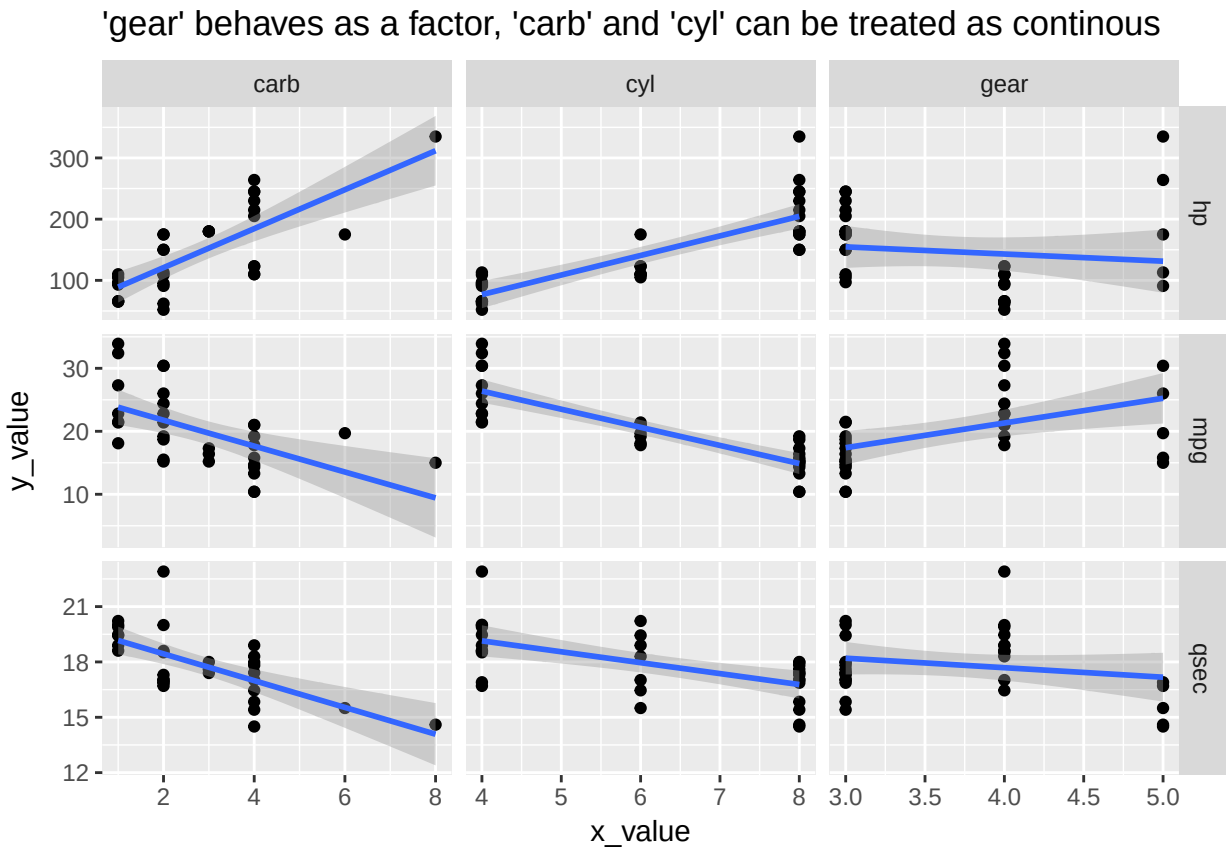

Figure 1: Plot of continuous vs. factor variables in mtcars data

The pairs plot shows many strong correlations.

```
if (!"GGally" %in% rownames(installed.packages())) {install.packages("GGally")}
library(GGally)
g_pairs <- ggpairs(mtcars, lower = list(continuous = wrap(ggally_smooth, color = "blue")),
           diag = list(continuous = "barDiag"), upper = list(continuous = wrap(ggally_cor,
```

```
            size = 3, color = "blue")), axisLabels = 'none')
print(g_pairs)
```

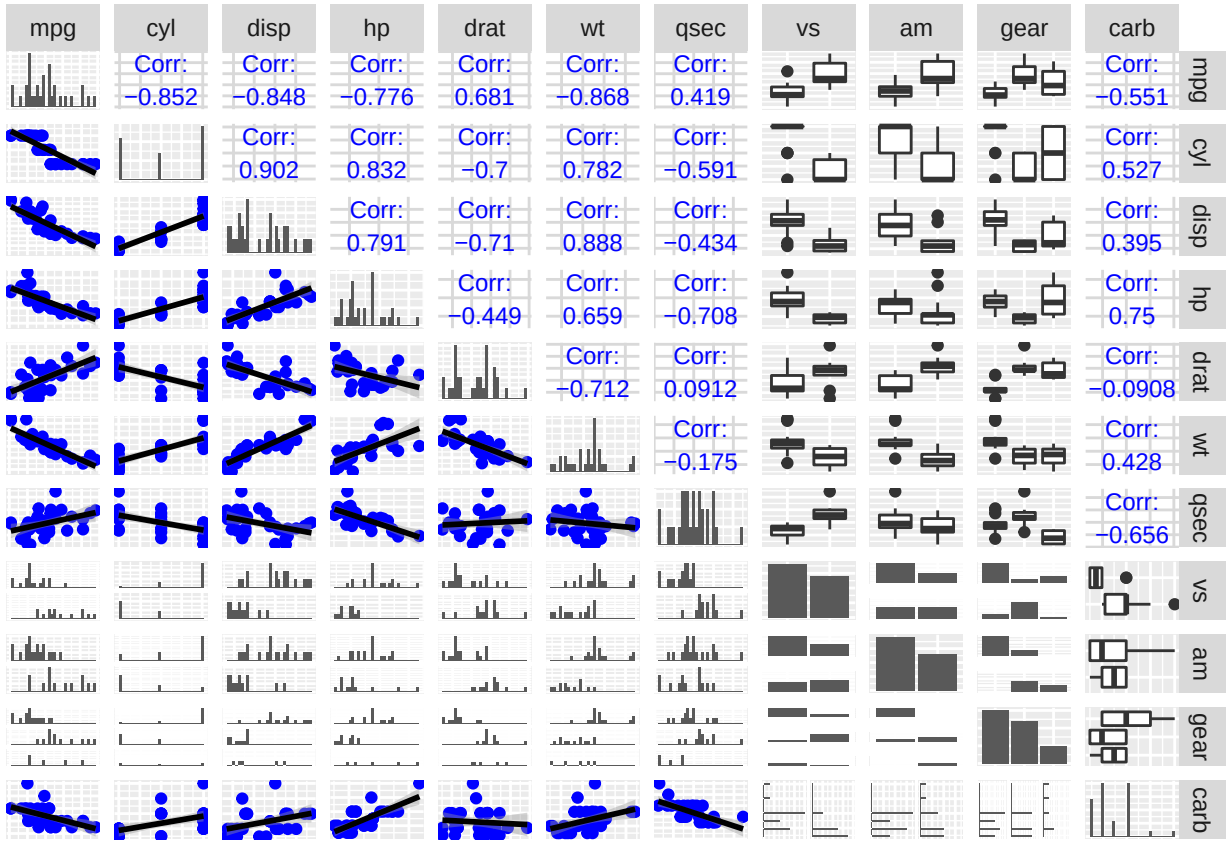

Figure 2: Pairs Plot of Motor Trend Cars Database

The violin plot depicts strong association between transmission type and `mpg`.

```
g_violin <- ggplot(mtcars, aes(x = am, y = mpg)) + geom_violin() +
        geom_dotplot(binaxis='y', stackdir='center', dotsize=1) +
        stat_summary(fun.y=mean, geom="point", shape=23, size=4, aes(fill = am)) +
        scale_fill_discrete(name="Mean MPG")+
        labs(title = "Average MPG for manual transmissions is significantly higher")
print(g_violin)
```

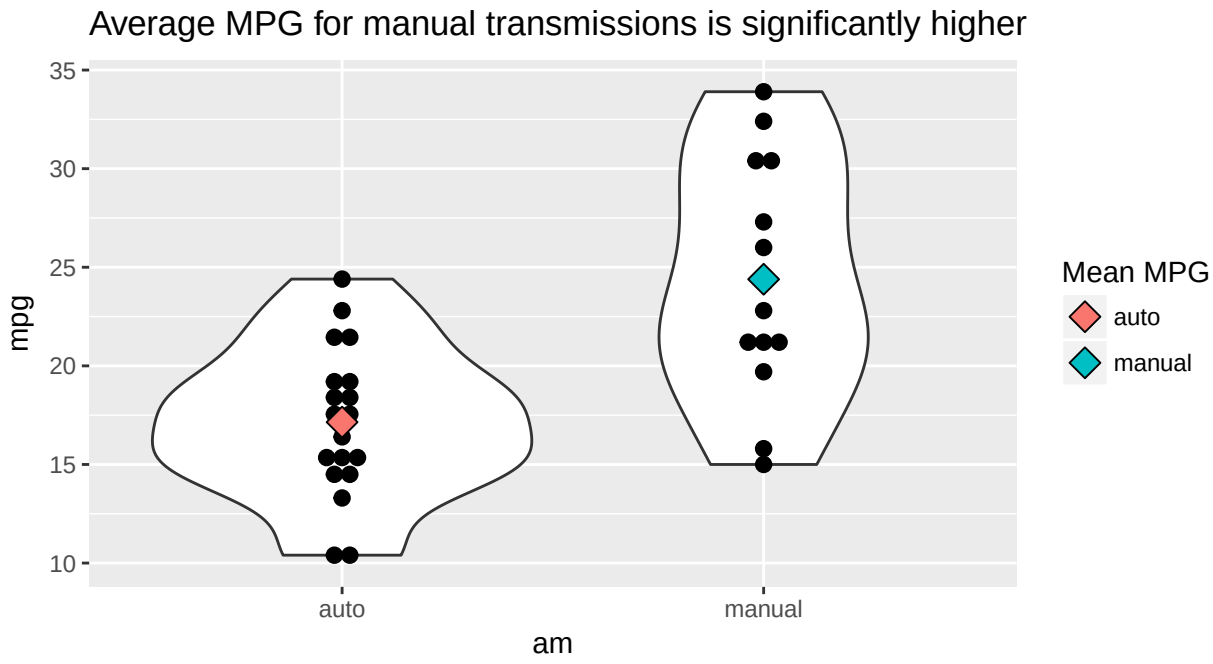

Figure 3: Violin plot of MPG vs. Transmission type

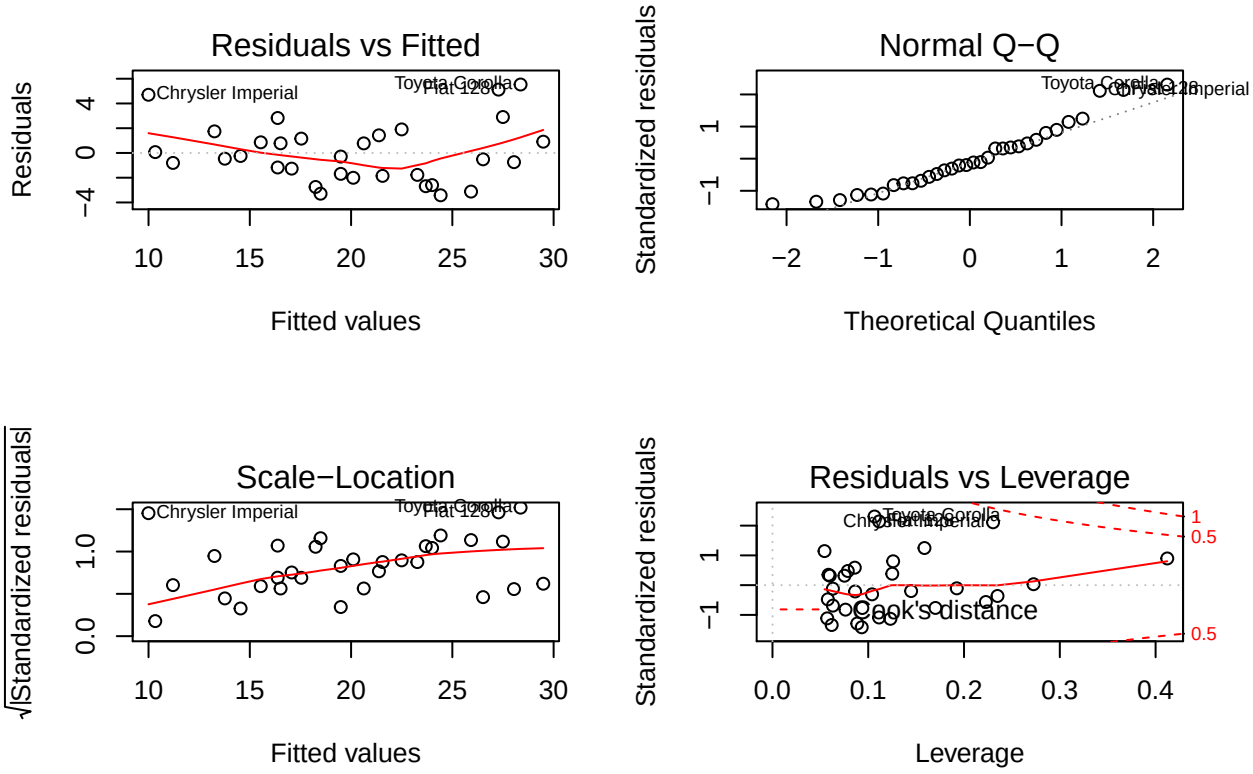

Figure 4: Base R's standard diagnostic plots guide our eye to potential problems.