Is an automatic or manual transmission better for MPG?

Executive summary

To answer the stated question exactly we should collect MPG values from all existing brands (total population). With the help of regression analysis and a sample, we can answer this question approximately, while quantifying uncertainty in our prediction. A simple regression of MPG against transmission type shows that **manual transmission better for MPG by 7.25 miles per gallon on average**. Adding more variables increases model's ability to predict MPG of a car.

Exploratory analysis

Let's explore data visually by plotting MPG separately for two groups of cars: with automatic and manual transmission (see Appendix, Figure 1: MPG vs transmission type). The cursory visual analysis does suggest that there is a difference in MPG due to transmission type. Let's validate this conjecture by regression analysis.

Regression analysis 1.

Let's perform an Ordinary Least Squares (OLS) regression of MPG ("mpg") on transmission type ("am"). It should be noted that although MPG is a continuous variable, transmission type is categorical one that can take on two states: "automatic" or "manual". To perform OLS I am treating "am" as dummy variable that can take on two values: "0" for "automatic" and "1" for manual (factorization explicit or implicit also possible). Summary of resulting regression model:

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Regression 1: Coefficient interpretation

- Intercept is mean MPG for automatic transmission (set am=0 and get average MPG for automatic transmission).
- `am` coefficient is incremental increase in MPG due to switch to manual transmission. In other words, **this is MPG difference between automatic and manual transmissions**. As before, if you set `am=1`, you will get mean MPG for cars with manual transmission. You can check these statements by comparing coefficients to calculated means:

```
ddply(mtcars, "am", function(x) mean(x$mpg))
```

```
##   am       V1
## 1  0 17.14737
## 2  1 24.39231
```

Both coefficients – `(Intercept)` and `am` – are statistically significant as evidenced by low p-values, i.e. mean MPG for automatic transmission significantly different from 0; and mean MPG for manual transmission significantly different from mean for automatic transmission (in other words, both have predictive values).

Regression 1: Quantifying model uncertainty and regression diagnostics

Predicted average MPG for automatic and manual transmission with 95% Confidence intervals are:

```
cbind(am = c(0,1), rbind(predict(ls1, newdata=data.frame(am=0), interval = "confidence"),
                         predict(ls1, newdata=data.frame(am=1), interval = "confidence")))
```

```
##   am      fit      lwr      upr
## 1  0 17.14737 14.85062 19.44411
## 1  1 24.39231 21.61568 27.16894
```

Normality of residuals distribution is not a problem here, as evidenced by Q-Q plot (see Appendix, Figure 2: Residuals Q-Q plot for Regression 1). The two real problems are:
- low explanatory ability of the model (adjusted $R^2 = 0.3385$)
- patterns in residuals distributions. See Appendix, Figure 3: Patterns in residual distribution by transmission type and additional variable, as an example of one pattern: heavier cars tend to have negative residuals (MPG overprediction).

Regression analysis 2: Adding more explanatory variables

There are two paths to adding more explanatory variables:
- **Manual**. First add to regression model all predictors that visually have linear relationships to MPG (see Appendix, Plot 3, for visualization of relationships between MPG and predictors), and then remove by hand insignificant ones, starting with least significant, unless only significant predictors are left in the model.
- **Automatic**. Let `stepAIC()` function from `MASS` package choose the "best" regression model for you (as the name of the function suggests, this is done via minimizing Akaike information criterion). As it turns out in this particular case, both approaches end up in identical model:

```
##
## 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
```

Regression 2: interpretation, diagnostics, and comparison to Regression 1

- All predictors, except intercept, are significant
- 1 sec increase in `qsec` ("1/4 mile time") will increase MPG by 1.23 while all other variables kept constant; 1000lb increase in weight will decrease MPG by 3.92; and manual transmission cars will have 2.94 Miles per gallon higher MPG on average over those with automatic transmission.
- Regression 2 has higher predictive ability as measured by adjusted R^2: 0.8336 vs 0.3385
- We are 95% confident that residuals are normal (see Appendix, Figure 5: Residuals Q-Q plot for Regression 2)
- Plot of residuals against regression components still reveals some heterescedasticity (see Appendix, Figure 6: Component + Residual Plots for Regression 2)

Appendix

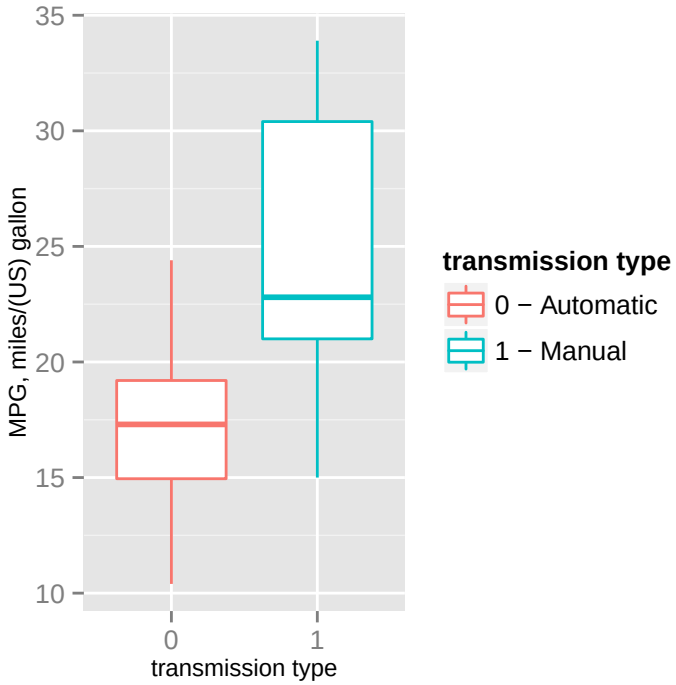

Figure 1: MPG vs transmission type

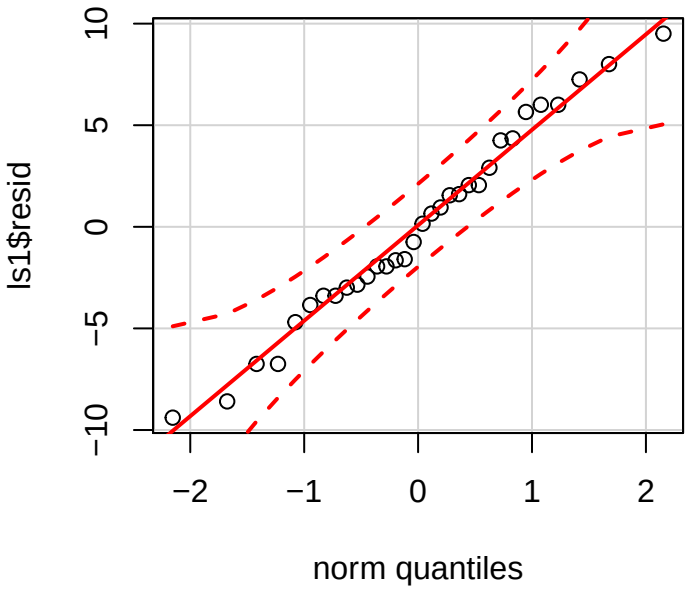

Figure 2: Residuals Q-Q plot for Regression 1

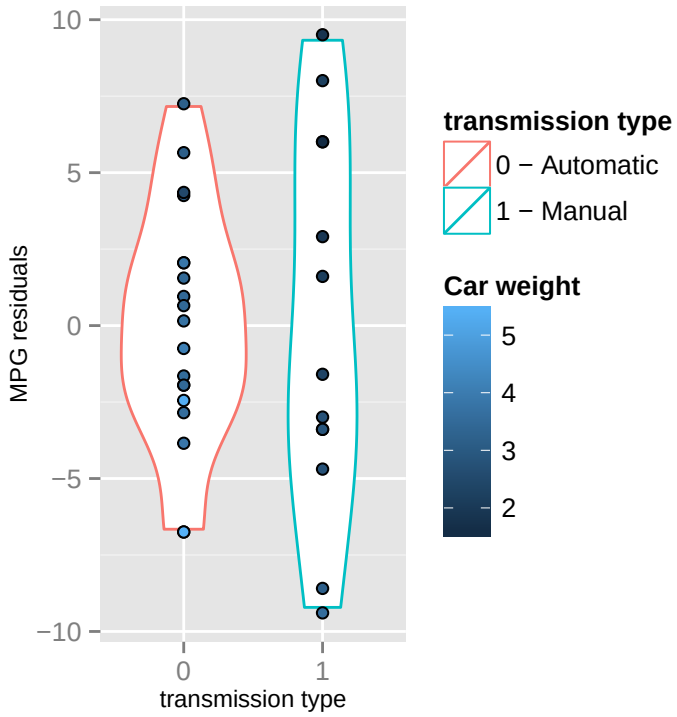

Figure 3: Patterns in residual distribution by transmission type and additional variable

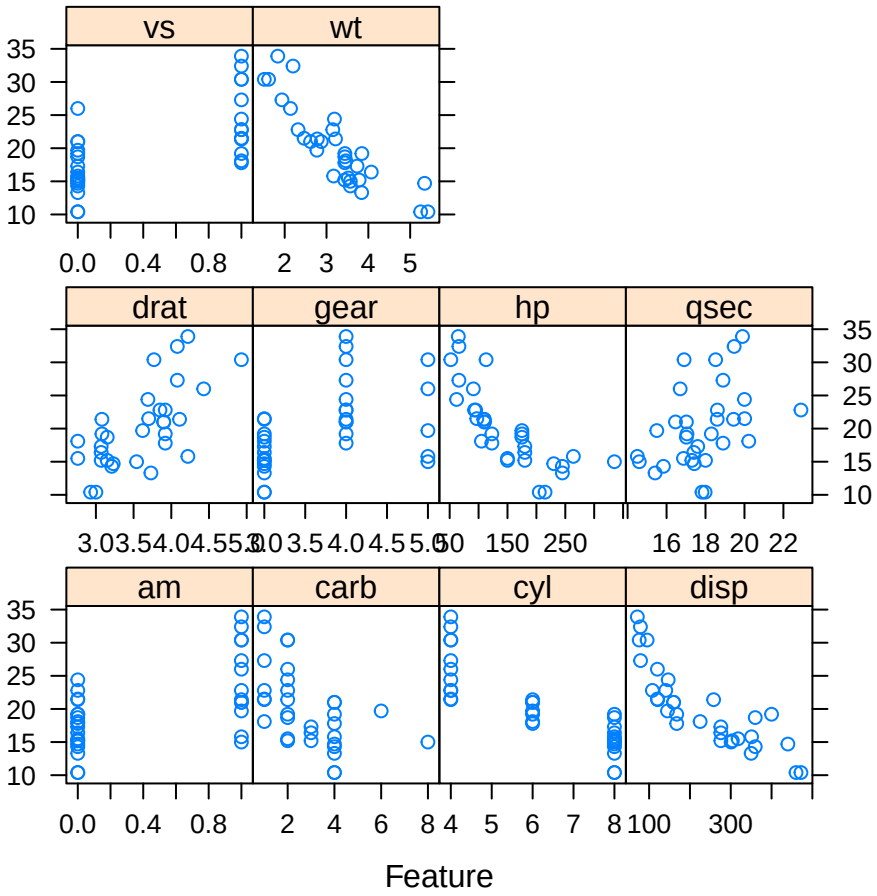

Figure 4: Feature plot: MPG vs all predictors

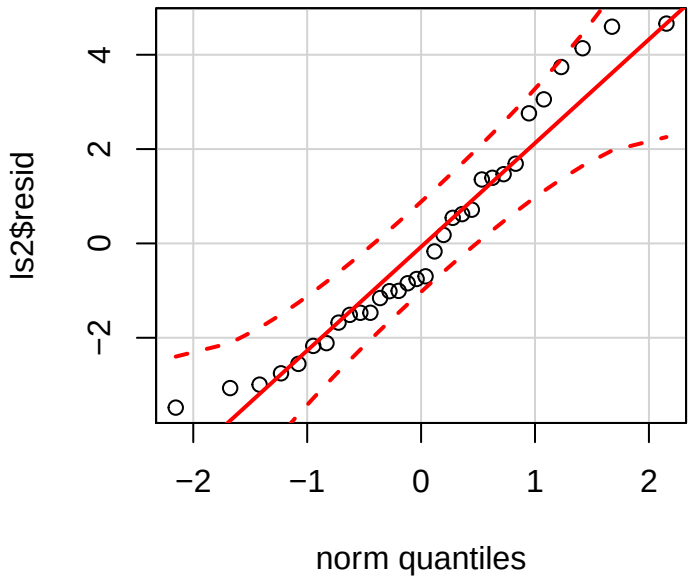

Figure 5: Residuals Q-Q plot for Regression 2

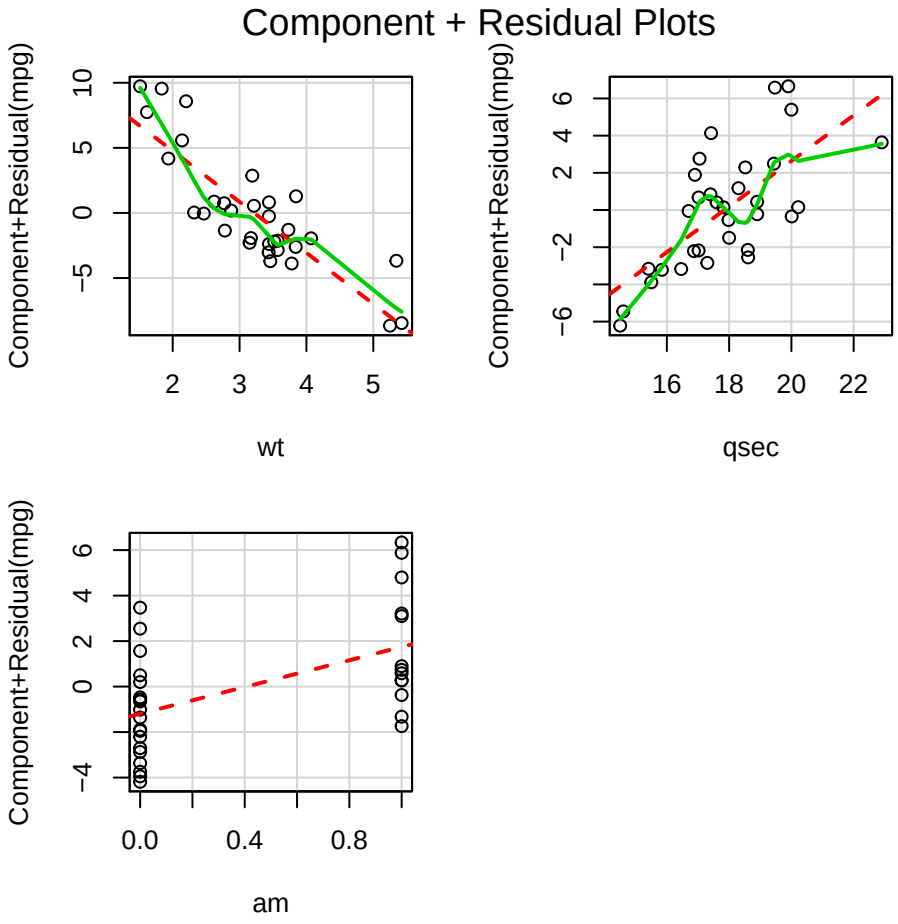

Figure 6: Component + Residual Plots for Regression 2