Motor Trend: mpg regression

bdanalytics

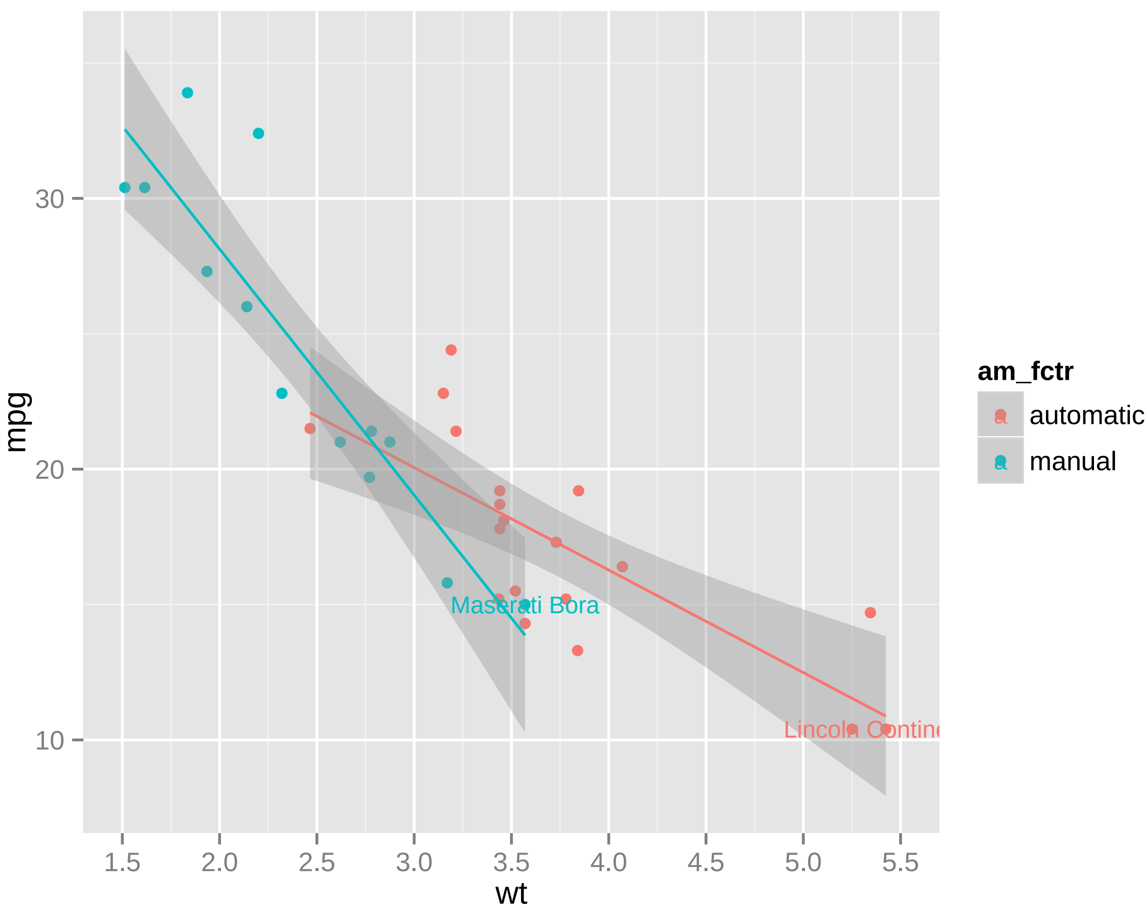
Date: (Thu) Oct 23, 2014 Data: Motor Trend Car Road Tests "mtcars {datasets}"
Source: Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411.
Time period: 1973–74 models

Synopsis:

Is an automatic or manual transmission better for MPG: Manual transmission is better for miles per gallon versus automatic transmission.
Average mpg for automatic transmission is `17` vs. `23` for manual transmission.

Quantify the MPG difference between automatic and manual transmissions: The univariate model yields `mpg ~ 17.15 + 7.25 * manual` while explaining only `33.8%` of the mpg variation.

The proposed multivariate model yields `mpg ~ 46.30 - 9.08 * wt` for cars with manual transmission and `mpg ~ 31.42 - 3.79 * wt` for cars with automatic transmission where `wt` is weight (lb/1000) of the car. This model explains `81.5%` of the mpg variation at a `99%` confidence level. The most influential data point turned out to be Maserati Bora which would increase the predicted mpg by `0.63` for cars with manual transmission without it. The most influential data point for automatic transmission was Lincoln Continental.



The model for variation from mean weight (wt) did not pass the statistical significance tests. Additional features were not statistically significant and/or explain additional mpg variation. The proposed model contains minor negative correlation of residuals with predicted values & residual heteroskedacity.

Potential next steps include:

1. Compress report to 5 pages; Knit PDF keeps crashing on my computer… extremely tedious to optimize length
2. Test other regression techniques (e.g. additive models in glm) to better quantify the relationship.

Appendix:

Import data & setup analytics: Automatic Transmission feature (am), number of cylinders (cyl), V/S (vs), number of forward gears (gear) & number of carburetors (carb) are numeric. Let's make them factors for analytics convenience.

Is an automatic or manual transmission better for MPG: Null Hypothesis (H_0): mpg is not impacted by am_fctr.
The variance by am_fctr appears to be independent.

```
print(t.test(subset(cars_df, am_fctr == "automatic")$mpg,
             subset(cars_df, am_fctr == "manual")$mpg,
             var.equal=FALSE)$conf)
```

```
## [1] -11.280194  -3.209684
```

```
## attr(,"conf.level")
## [1] 0.95
```

We reject the null hypothesis i.e. we have evidence to conclude that am_fctr impacts mpg (95% confidence). Manual transmission is better for miles per gallon versus automatic transmission.

Quantify the MPG difference between automatic and manual transmissions: Let's try the univariate model to establish a benchmark against which we can evaluate more complex models, if necessary

```
mpg_fit <- lm(mpg ~ am_fctr, data=cars_df)
print(summary(mpg_fit))
```

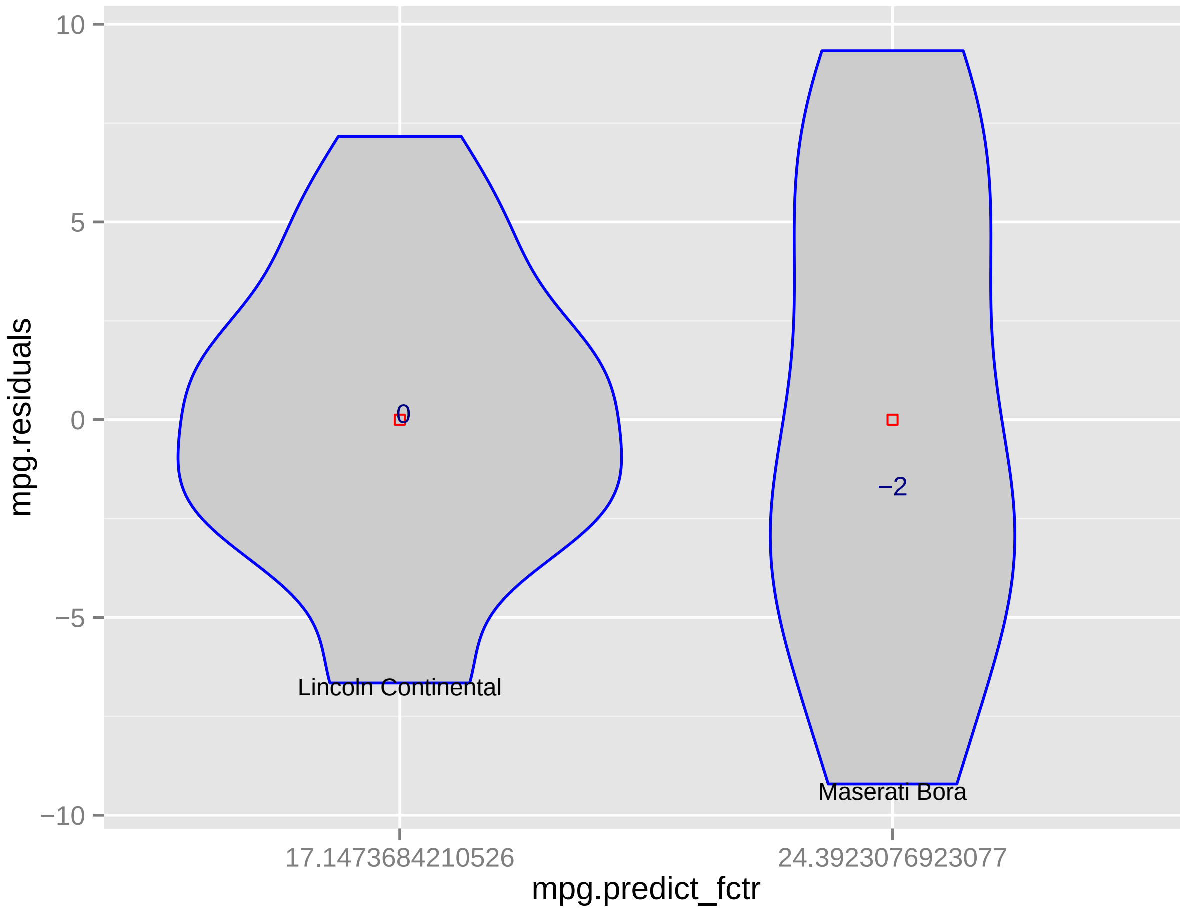
```
##
## Call:
## lm(formula = mpg ~ am_fctr, data = cars_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     17.147      1.125  15.247 1.13e-15 ***
## am_fctrmanual    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
```

This univariate model is statistically significant and explains `33.8`% of the mpg variation.

Manual Transmission provides `7.24` additional miles per gallon compared to automatic transmission. This result is highly significant - 95% confidence interval is `[ 3.64, 10.85]`.

Let's inspect the residuals for any bias

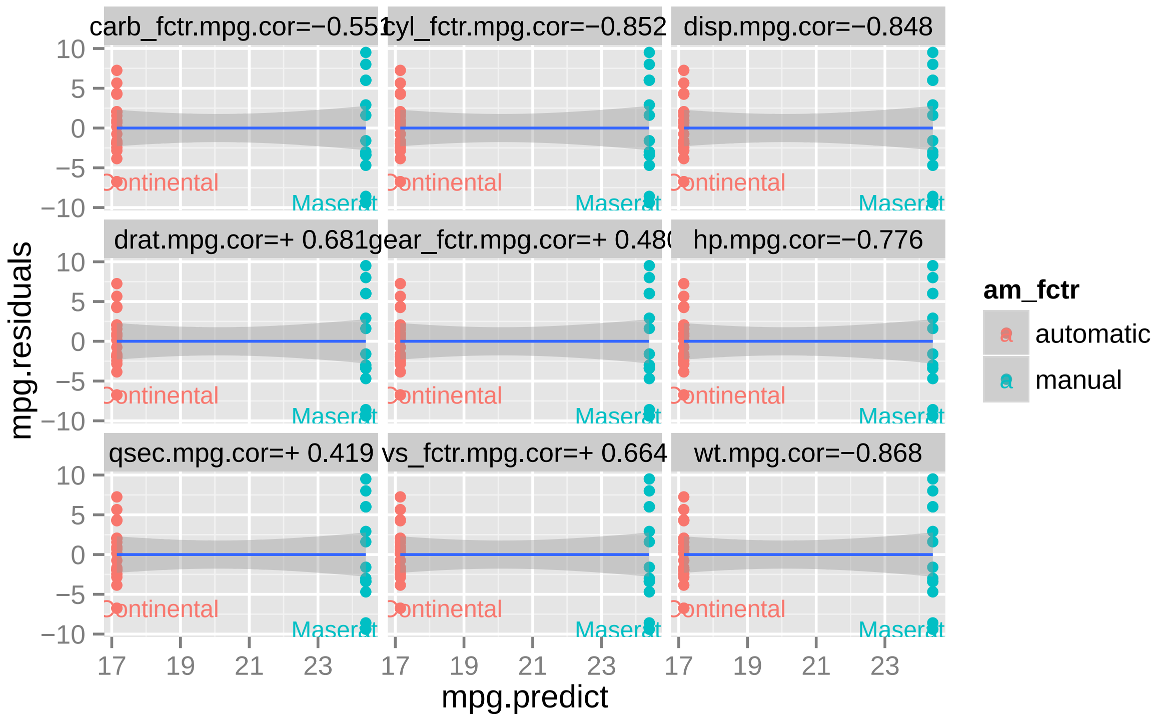
```
## Warning in myplot_violin(mpg_fit_df, "mpg.residuals", "mpg.predict"):
## xcol_name:mpg.predict is not a factor; creating mpg.predict_fctr
```



The residuals of the regression appear NOT biased with mean close to 0. However, there is some residual heteroskedacity (median residuals for manual transmission is -2). Let's examine if the residuals are impacted by features not in this model. For that, let's first inspect the correlations between mpg & each of the features and build labels that are used later in the residual plots

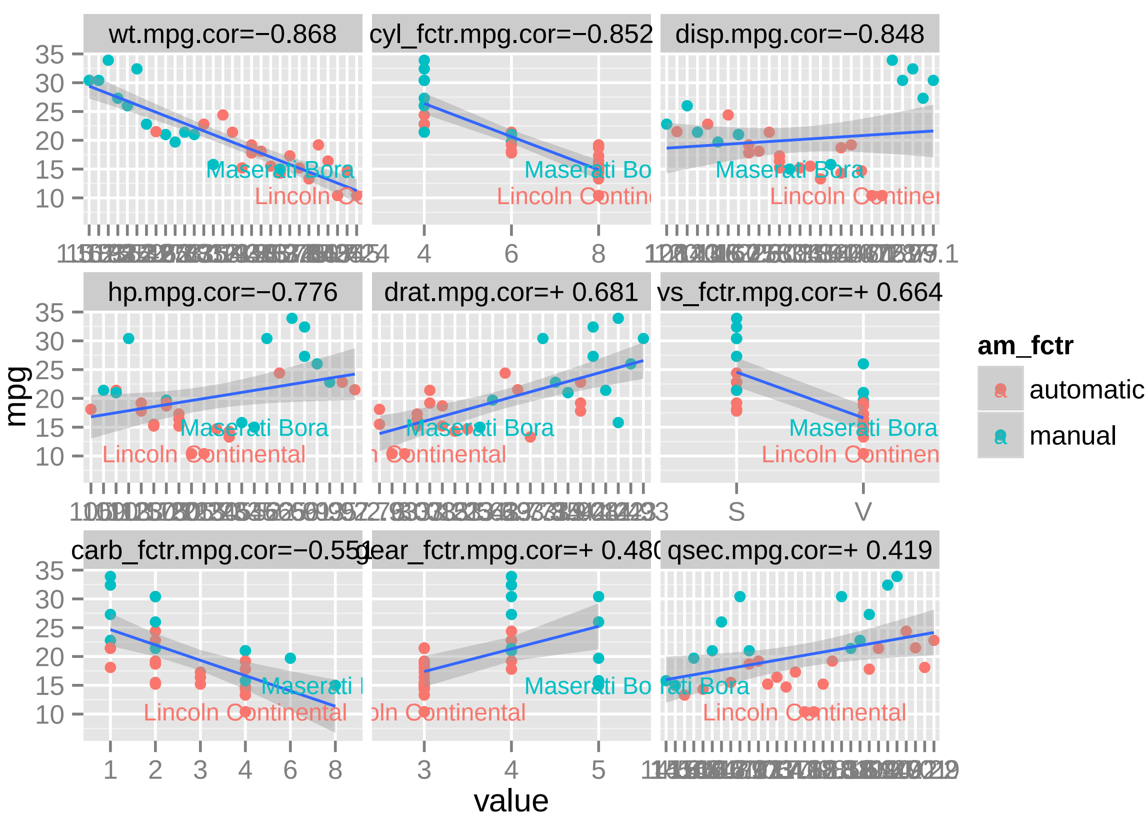
```
##         mpg.cor mpg.cor.abs feature  variable                    label
## wt   -0.8676594   0.8676594      wt        wt        wt.mpg.cor=-0.868
## cyl  -0.8521620   0.8521620     cyl  cyl_fctr   cyl_fctr.mpg.cor=-0.852
## disp -0.8475514   0.8475514    disp      disp      disp.mpg.cor=-0.848
## hp   -0.7761684   0.7761684      hp        hp        hp.mpg.cor=-0.776
## drat  0.6811719   0.6811719    drat      drat      drat.mpg.cor=+ 0.681
## vs    0.6640389   0.6640389      vs   vs_fctr   vs_fctr.mpg.cor=+ 0.664
## am    0.5998324   0.5998324      am   am_fctr   am_fctr.mpg.cor=+ 0.600
## carb -0.5509251   0.5509251    carb carb_fctr  carb_fctr.mpg.cor=-0.551
## gear  0.4802848   0.4802848    gear gear_fctr gear_fctr.mpg.cor=+ 0.480
## qsec  0.4186840   0.4186840    qsec      qsec      qsec.mpg.cor=+ 0.419
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```



None of the other variables seem to explain any more mpg variation at first glance. Let's check the actual distribution of mpg vs. am_fctr.

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

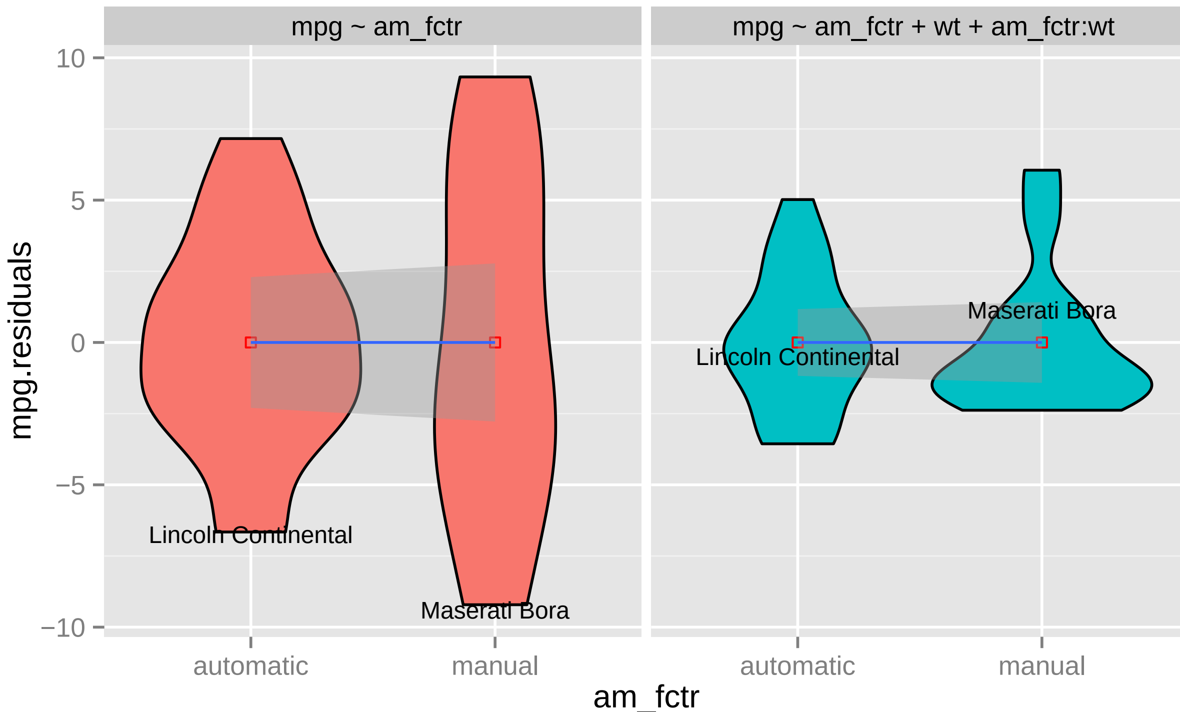


Clearly, other features can provide additional insights into the relationship between mpg & transmission. Let's add features to the simple model (mpg_fit) in order of the features correlated with mpg (highest tested first).

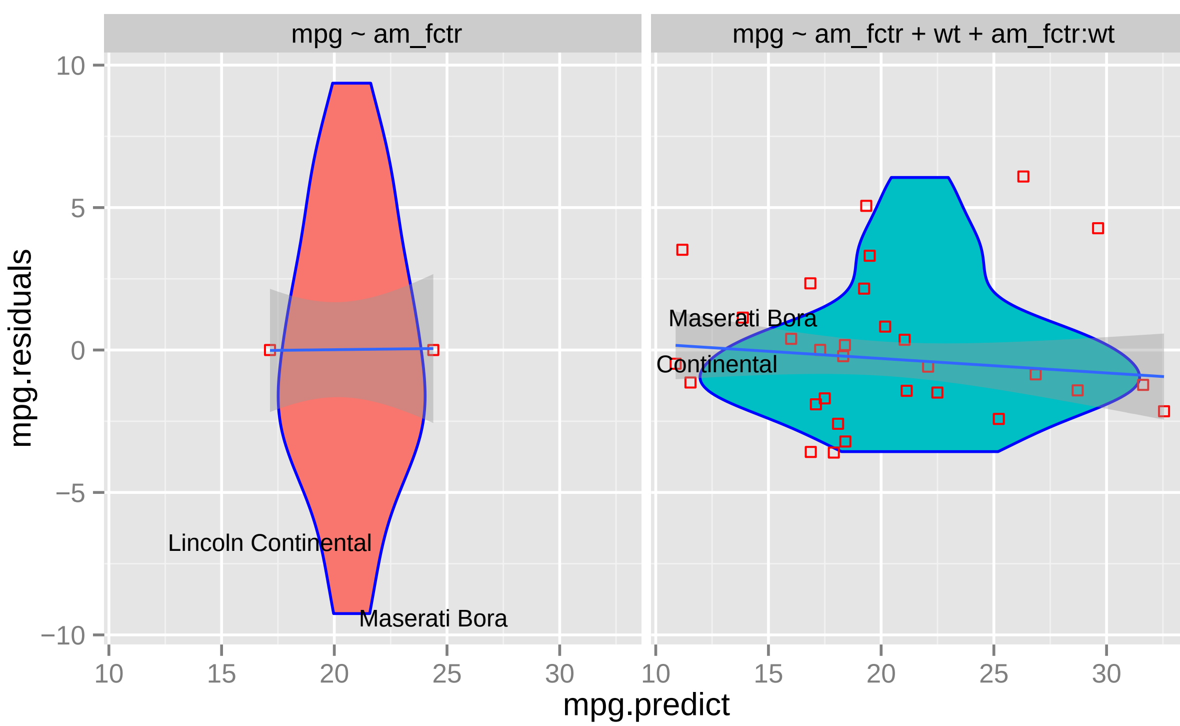
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am_fctr
## Model 2: mpg ~ am_fctr + wt
## Model 3: mpg ~ am_fctr + wt + am_fctr:wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 65.913 7.717e-09 ***
## 3     28 188.01  1     90.31 13.450  0.001017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = mpg ~ am_fctr + wt + am_fctr:wt, data = cars_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6004 -1.5446 -0.5325  0.9012  6.0909
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       31.4161     3.0201  10.402 4.00e-11 ***
## am_fctrmanual     14.8784     4.2640   3.489  0.00162 **
## wt                -3.7859     0.7856  -4.819 4.55e-05 ***
## am_fctrmanual:wt  -5.2984     1.4447  -3.667  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.591 on 28 degrees of freedom
## Multiple R-squared:  0.833,  Adjusted R-squared:  0.8151
## F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
```

The interaction model of am_fctr & wt is statistically significant (`99%` confidence) and all the model coefficients are significant (`99%` confidence). None of the other models tested piecewise for each additional feature to this model in a similar fashion crossed these thresholds. Adj-Rsq is `0.815`. Let's inspect the residuals for this model.

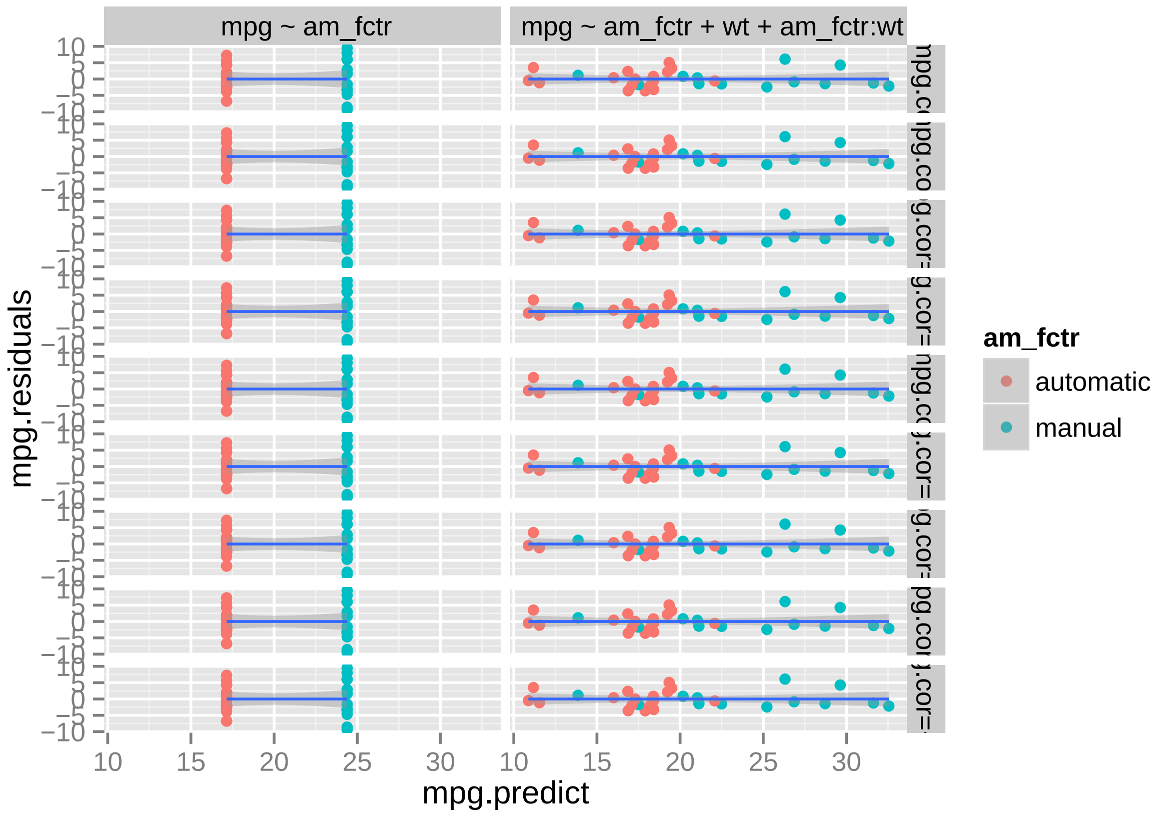


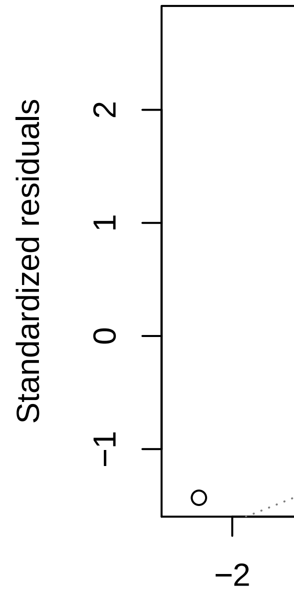
The mean of residuals is `0` for both transmission types and the heteroskedacity of the residuals is reduced significantly.



There seems to be negatively correlated residuals in the interactive wt model, although that doesn't show up when method="lm" in geom_smooth()

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```






```
##
##  Shapiro-Wilk normality test
##
## data:  mpg_wt_i_fit_plot_df$mpg.residuals
## W = 0.9354, p-value = 6.825e-10
```

```
## [1] "Max hat value:"
```

```
##                 mpg cyl disp  hp drat   wt qsec vs am gear carb am_fctr
## Maserati Bora  15   8  301 335 3.54 3.57 14.6  0  1    5    8  manual
##                cyl_fctr vs_fctr gear_fctr carb_fctr            id
## Maserati Bora         8       V         5         8 Maserati Bora
##                   id_outlier
## Maserati Bora Maserati Bora
```

```
##              (Intercept)           am_fctrmanual                      wt
## " 0.000000000000000157" "-0.966155686888044429" "-0.000000000000000028"
##         am_fctrmanual:wt
## " 0.458367018004453231"
```

The Maserati Bora is the most influential point and without it manual transmission would be `-0.97` mpg (lower) - [am_fctrmanual] and `1.6` mpg (higher) [am_fctrmanual:wt].

```
## [1] "Proposed model:  mpg ~ 31.42 + 14.88 * am_fctrmanual + -3.786 * wt + -5.298 * am_fctrmanual:wt"
```

```
##                      2.5 %    97.5 %
## (Intercept)      25.229642 37.602469
## am_fctrmanual     6.143928 23.612917
## wt               -5.395234 -2.176581
## am_fctrmanual:wt -8.257693 -2.339028
```

The proposed model is `mpg ~ 31.42 + 14.88 * am_fctrmanual + -3.786 * wt + -5.298 * am_fctrmanual:wt` where `am_fctrmanual` is 1 for manual transmission [0 for automatic], `wt` is weight (lb/1000) and `am_fctrmanual:wt` is wt for manual transmission [0 for automatic]. None of the coefficients change sign in the 95% confidence interval.

```
## R version 3.1.1 (2014-07-10)
## Platform: x86_64-apple-darwin13.1.0 (64-bit)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] splines   stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] plyr_1.8.1      reshape2_1.4    doBy_4.5-10     MASS_7.3-34
## [5] survival_2.37-7 ggplot2_1.0.0
##
## loaded via a namespace (and not attached):
##  [1] colorspace_1.2-4 digest_0.6.4     evaluate_0.5.5   formatR_1.0
##  [5] grid_3.1.1       gtable_0.1.2     htmltools_0.2.4  knitr_1.7
##  [9] labeling_0.3     lattice_0.20-29  lme4_1.1-7       Matrix_1.1-4
## [13] minqa_1.2.3      munsell_0.4.2    nlme_3.1-117     nloptr_1.0.4
## [17] proto_0.3-10     Rcpp_0.11.2      rmarkdown_0.2.54 scales_0.2.4
## [21] stringr_0.6.2    tools_3.1.1      yaml_2.1.13
```