

Regression analysis on mpg in mtcars dataset

Executive summary

We fit a linear model using multiple regression to (briefly) explore the relationship between mileage (mpg) and predictors in the `mtcars` dataset from Hocking (1976). We are interested especially in transmission effects. Our model uses weight and horsepower-to-weight ratio to explain mpg. It does not permit inference based on transmission. Higher mileage is strictly observed on auto transmission cars in the data, but weight is a major confounder.

Exploratory Analysis

`mtcars` is a complete motor vehicle dataset courtesy of Hocking (1976), assembled from 32 vehicles and comprised of numeric and factor variables.

We observe highly disparate weight distributions based on transmission (Fig 2). In fact, multicollinearity occurs between many of the numeric variables (Fig. 3). There are no NA values in the data, and summary exploration of each variable does not reveal obvious or egregious errors in coding.

Ordinary least squares regression

For mpg response y_i we seek to fit a response using ordinary least squares regression of the form:

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ki} + \epsilon_i$$

Here x_{ki} are regressors (predictors) of choice, $\beta_0, \beta_1, \dots, \beta_p$ are coefficients chosen to minimize residual squared error, and ϵ_i are error terms. Standard OLS regression assumptions apply. The regressors are not linearly related, and error terms are normal and heteroscedastic.

We use R to perform OLS (in particular, the `lm` function). Full code is omitted for brevity, but present in the source `.Rmd` file (rendered to PDF via `knitr`).

Model selection

We seek a **parsimonious** model. Physically, we might expect `wt` (weight) and `hp` (horsepower) to be a promising start. The `mpg ~ wt` relationship looks quite reasonably linear (Fig. 1). Using both `wt` and `hp`, compared to only `wt`, raises adj. R^2 from 0.745 to 0.815. However, ANOVA analysis of the F ratio in comparison to `lm(mpg ~ wt * hp, mtcars)` has $p = 0.001 < 0.05$; suggesting meaningful interaction.

To more uniquely express power, [Henderson, Velleman](#) (1981) proposes `hp/wt`. This is weakly correlated to `wt` (`cor = 0.054`), and ANOVA $p = 0.509 > 0.05$, suggesting interaction may be ignored.

We invite further model comparison based on residual sum of squares criteria:

##	predictors	adj.R.squared	ANOVA.p.vs.proposed
## 1	wt only	0.745	0.0018387770138565
## 2	proposed	0.812	--
## 3	proposed + cyl	0.834	0.0695468160533971
## 4	proposed + disp	0.805	0.897341271995695
## 5	proposed + am	0.818	0.169677656748996
## 6	proposed + (all remaining)	0.763	0.849560846311976

We observe the proposed model is superior to a `wt` only model, has high adjusted R squared, and no other compared models offer significant improvement. Cylinders comes close, but we appeal to parsimony and also to the strong relationship between cylinders and weight (Fig 2).

Residuals and diagnostics

We perform a standard series of residuals diagnostics plots (Figs. 4 and 5) to evaluate the model fit and assumptions. The results are reasonable. We propose that vehicle selection, rather than model, are the source of leverage concerns.

1. **Residuals vs Fitted** is modestly in the center, although not at the edges.
2. **Normal QQ** plot shows that the residuals are fairly normal (except at edges), as assumed.
3. **Scale-Location** plot shows fair homoscedasticity, as assumed (again departing most at edges).
4. **Residuals vs Leverage** reveals a couple of points approaching Cook's distance and may have undue influence.
5. **Hat values** plot, further illustrating degree of outsize influence of some observations.

Results, inference, and interpretation

Our proposed model uses `wt` (units of 1000 lbs) and `hp/wt` as regressors.

```
coef(summary(lm(mpg ~ wt + I(hp/wt), mtcars)))
```

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	41.4842	2.0237	20.50	0.0000000000000000084
##	wt	-5.2554	0.4804	-10.94	0.00000000000827780910
##	I(hp/wt)	-0.0989	0.0289	-3.43	0.00183877701385649298

Both `wt` and `hp/wt` have a meaningful ($p < 0.05$) effect in the model on `mpg`. Controlling for `hp/wt`, a 1000-lb increase in `wt` is modeled to affect `mpg` by a quantity estimated with 95% confidence to be within the interval -4.273, -6.238. For an increase in hp-per-1000-lbs of 100, the 95% confidence interval (holding `wt` fixed) for the estimated effect on `mpg` is -3.992, -15.797. Heavier, and more overpowered cars are inferred to reduce `mpg`. (*n.b.* 95% CIs are obtained by `coeff +/- qt(.975, fit$df) * standard error`).

Automatic vs manual transmission

Note that our linear model does not include transmission; its estimated effect has poor $p > 0.05$, and so inference from this estimation is not permissible.

```
coef(summary(lm(mpg ~ wt + I(hp/wt) + factor(am), mtcars)))[4, ]
```

##	Estimate	Std. Error	t value	Pr(> t)
##	1.95	1.39	1.41	0.17

This is unsurprising; the two transmission groups have quite distinct weight makeup (Fig 2). A linear model using *only* transmission as a predictor is a poor fit ($R^2 = 0.36$). Weight is a strong predictor and has a strong relationship with transmission, and thus inference on transmission from this data is unlikely to be achieved.

Mean difference is 7.245 MPG in favor of automatic cars, but this is solely from observed data, not inference.

Appendix

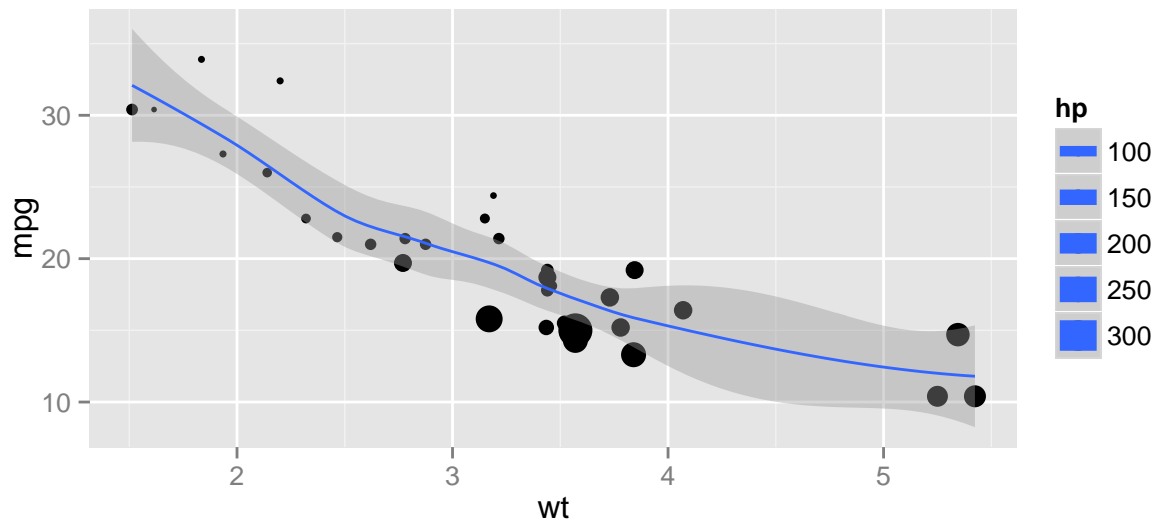
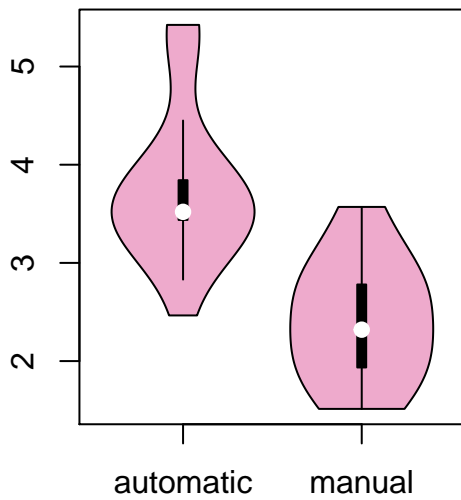


Fig. 1

mtcars: wt (1000 lbs) by am



mtcars: wt (1000 lbs) by cylinder

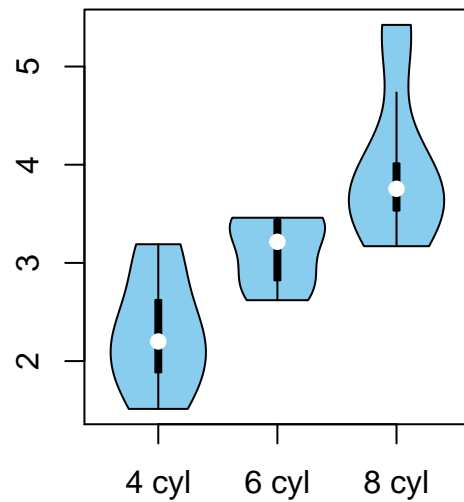


Fig. 2

mtcars correlations (corrplot)

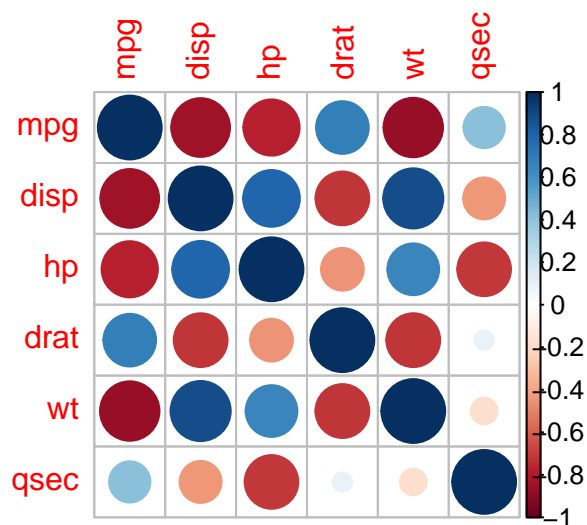


Fig. 3

Observation hat values in model fit

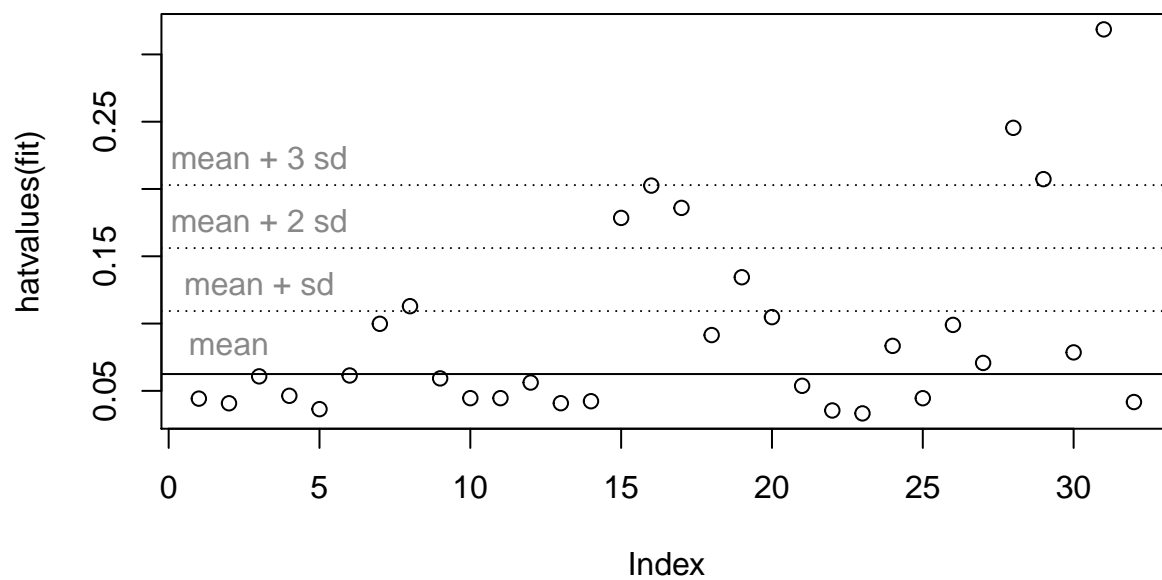


Fig. 4

Residuals Diagnostics Plots

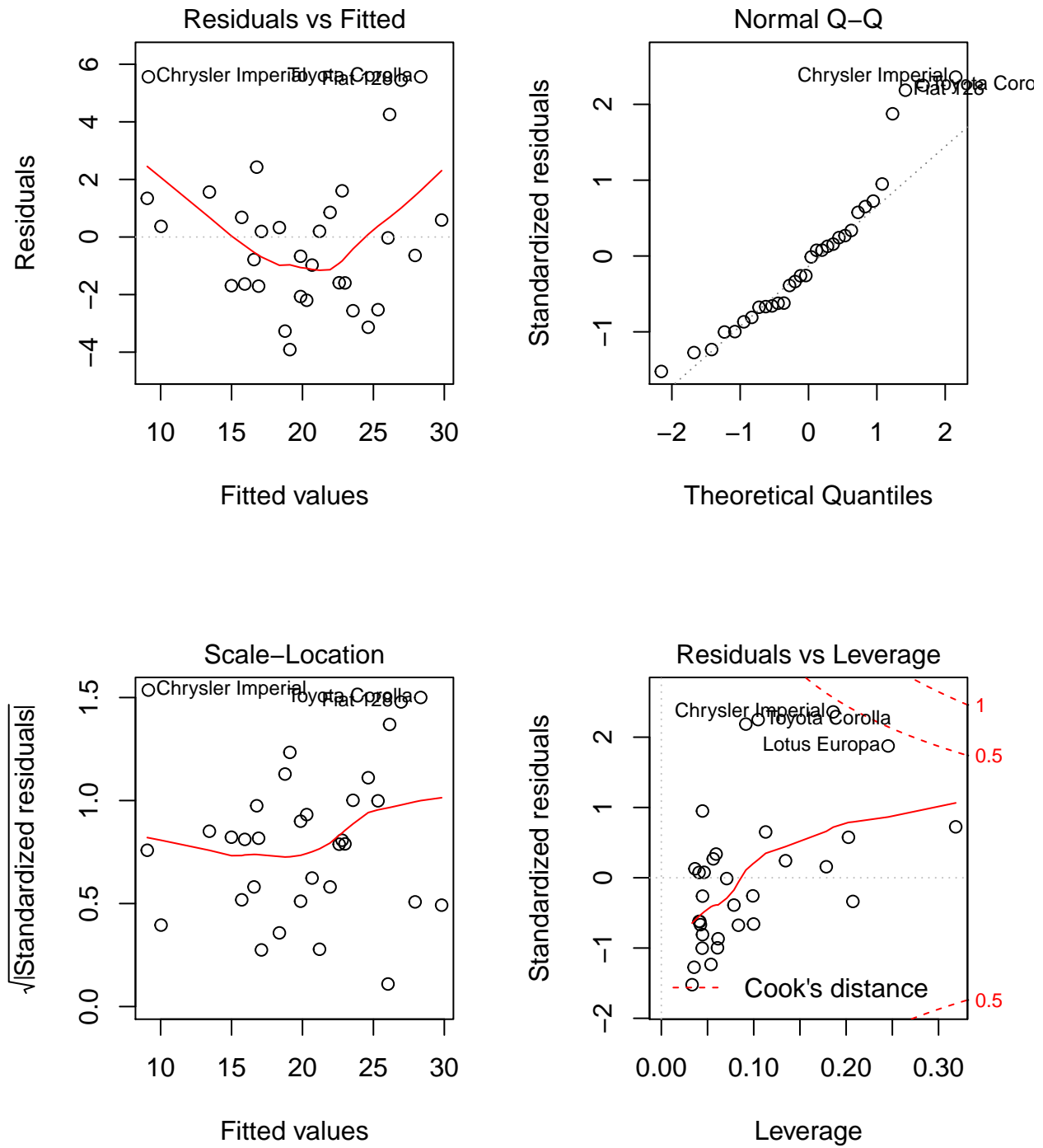


Fig. 5