Linear Regression

Cody Frisby

10/21/2017

Data Summary

My data comes from the 1974 *Motor Trend* magazine that is included as one of the datasets in R (`mtcars`). There are 32 total observations with 11 total variables. Here I attempt to answer the question of whether manual or automatic is better, and what the difference is if any, for miles per gallon and what variables are significant in predicting **mpg**. I will also attempt to predict **mpg** on the test set and compare it with actual **mpg**.

I fit a linear regression model using all the variables (no interactions) and take a look at the variance inflation factors (this technique can help us determine if any of our explanatory variables are highly correlated).

cyl	16.652578
disp	57.978325
hp	14.944173
drat	5.850496
wt	58.719325
qsec	38.200271
vs	10.268670
am	5.457847
gear	9.977316
carb	10.868641

We can see that a few of the variables are going to need to be removed from the final model. Using the leaps package and `regsubsets` function from that package I narrowed down the best models based on the BICs (Bayesian information criterion). My final model is

$$mpg_i = \beta_0 + \beta_1 wt_i + \beta_2 qsec_i + \beta_3 am_i + \varepsilon_i$$

and when fit to the data our prediction function is

$$mpg = 1.9040962 - 3.5085502wt + 1.5896423qsec + 4.0427425am$$

1) Is an automatic or manual transmission better for MPGs?

The answer to this question is **manual** transmission. A **manual** transmission will get 4.0427425 more **mpg**, on average, than an **automatic** transmission, all other variables held constant.

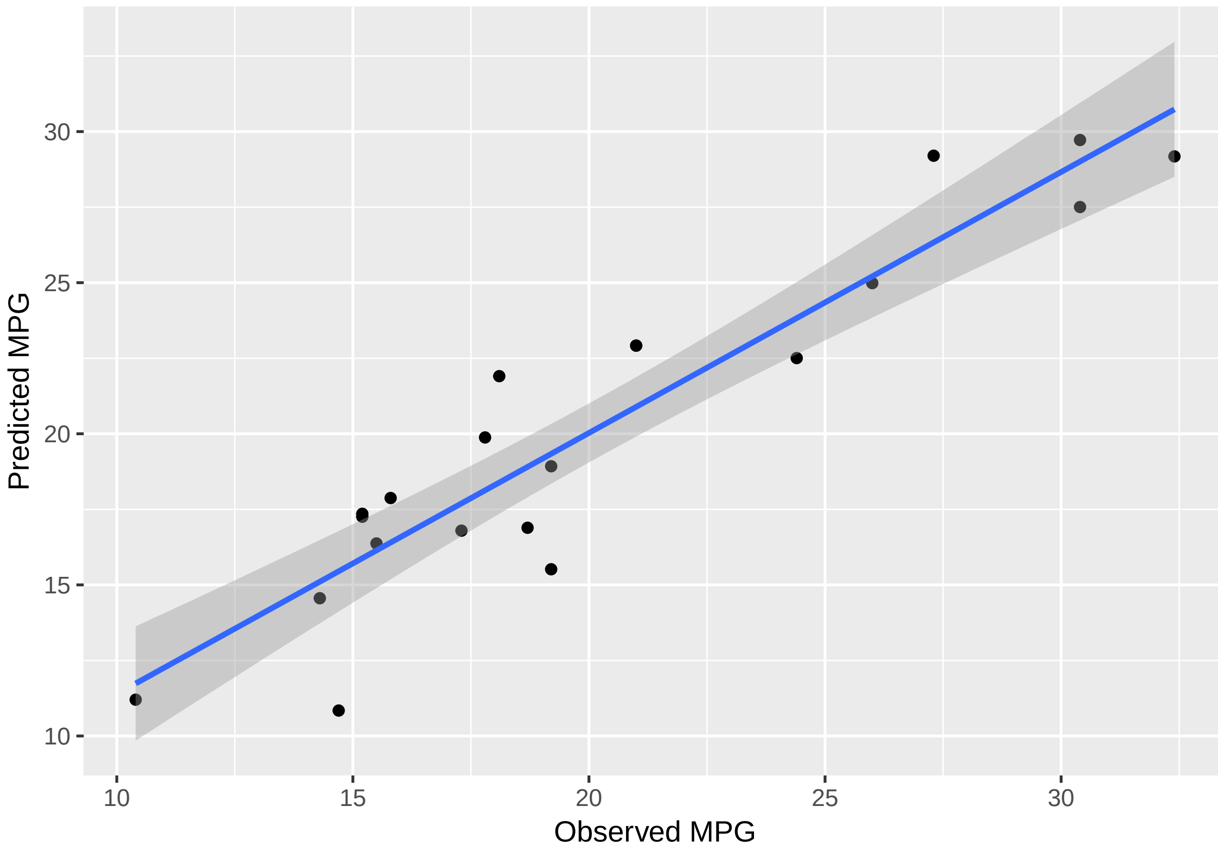
2) Difference between auto and manual?

I display the coefficients of the model I chose.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.904096	9.4625081	0.2012253	0.8429104
wt	-3.508550	0.8853891	-3.9627213	0.0010052
qsec	1.589642	0.4307739	3.6902011	0.0018161
am	4.042743	1.8414683	2.1953908	0.0423064

Here we can see that our model includes **wt** (the car's mass in 1000 lbs), **qsec** (the cars 1/4 mile time), and **am** (auto or manual transmission). There is a negative effect from **wt** meaning that as the cars' mass increases its expected **mpg**s decrease by -3.5085502. Similarly, as a cars quarter mile time increases its expected **mpg**s increase by 1.5896423. And lastly, when the car has a manual transmission, its **mpg**s are expected to be 4.0427425 higher than when it has an automatic transmission, holding the other variables constant. A 95% confidence interval is [0.157584, 7.927901], showing that at worst **manual** will beat **auto** by only 0.157584.

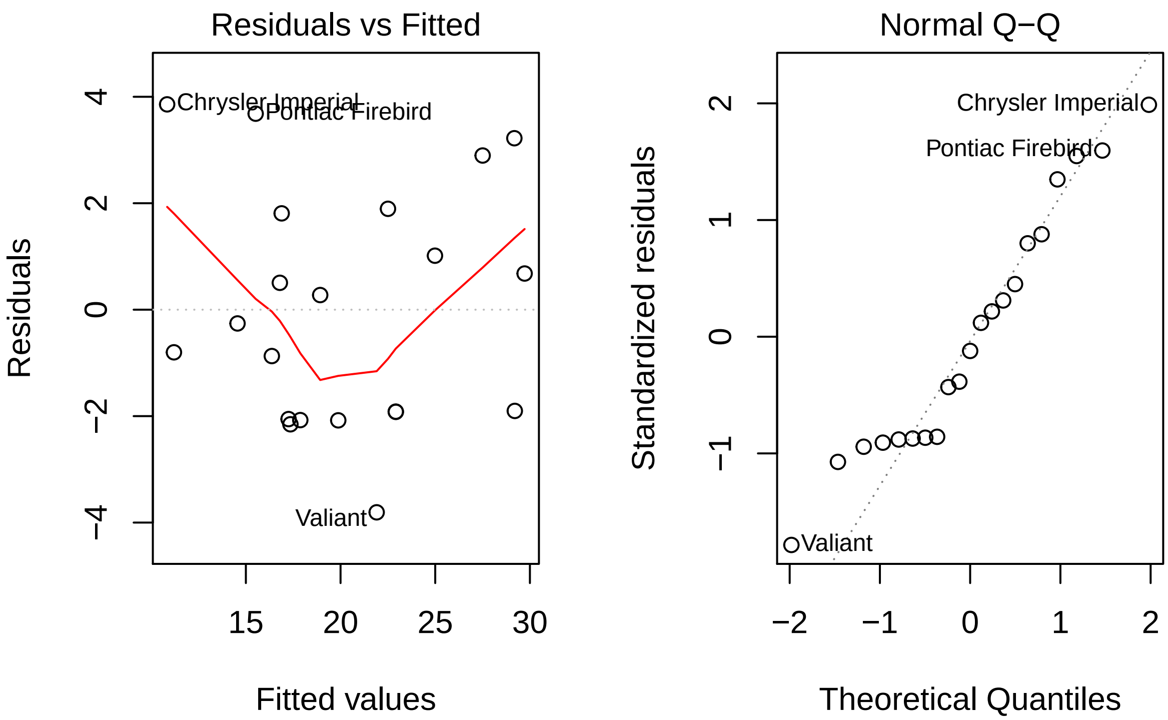
How well does the model predict on our test set?



Looking at the **observed** vs. **predicted**, our model appears to predict fairly well. The R^2 value is 0.8636.

Model Adequacy

Here, I take a look at the model residuals checking the assumptions of linear models.



Our residual plots look OK. The assumptions of equal variance, linearity, and normality are OK.

Conclusion

Although all the terms in our model are statistically significant, at the $\alpha = 0.05$ level, the confidence interval for **am** is quite wide, 0.157584, 7.927901. I chose this model over the other possibilities because it was the most parsimonious model in the top 10 (based on BIC) that contained **am** with the lowest BIC, Cp, and close to the highest adjusted R^2 value.