Regression analysis

Dejan Pljevljakusic

Sunday, April 03, 2016

Executive summary

Observed average fuel consumption for *automatic* transmission is about **17.15 mpg**, while for the *manual* transmission is about **24.39 mpg**. When we fit simple regression model `mpg ~ am` we get the same results. If we try to fit some other multiple regression models both intercept and slope are changed. In our first model selection we have choose **Model 16** `mpg ~ am + hp + qsec + vs + gear + carb` as the most promising one, but when we have applied 'Step-wise model search' we get **Model S** which was simple `mpg ~ am + wt + qsec`. We finally choose to stick with this one and its regression coefficients tell us that expected fuel consumption in cars with automatic transmission is about **9.61 mpg**, and if you change your vehicle with some other that has manual transmission you can expect to drive about **2.93 miles** more for each gallon of fuel. Residuals for this model was normally distributed, what means that all error values are random.

Synopsis

Since I am a journalist of *Motor Trend* magazine my task is to explore relationship between a set of variables and fuel consumption in data set of 32 motor vehicles. The main aim of this research is to distinguish differences between automatic and manual transmission regarding miles per gallon (mpg) variable and to address the question if one is better than other. All tables and plots are presented in appendix of this document.

Exploratory data analysis

Load the data set

```
data(mtcars)
```

From `?mtcars` we can find a basic information about the `mtcars` data set. Variable that responses to type of transmission is marked as `am` (0 = automatic, 1 = manual), while variable that responses to fuel consumption is marked as `mpg` (Miles/(US) gallon). Now, let's explore what would be the mean fuel consumption values for each type of transmission (Table 1).

At this point we can observe that average fuel consumption for *automatic* transmission is about **17.15 mpg**, while for the *manual* transmission is about **24.39 mpg** (Table 1). Now, let's explore if any variables in this data set other then transmission type is correlated with fuel consumption (**mpg** variable).

From the results presented in Table 2 we can conclude that only `cyl` `disp`, `drat` `wt` and `gear` variables are in some extent correlated to the variable `am`.

Model selection

Since I was asked to check if an automatic or manual transmission type is better for fuel consumption, first I will set basic linear regression model with only `am` variable included as a predictor of `mpg` variable (outcome).

```
fit0 <- lm(mpg ~ am, data = mtcars)
```

Summary of coefficients for simple linear regression is presented in Table 3, while scatterplot of observed relationship, together with fitted linear regression line (solid red line), is shown in Figure 1. From this results we can conclude that if vehicle have *automatic* transmission fuel consumption of about **17.15 mpg** is expected, while for changing transmission type to *manual* **7.25 miles** more per gallon is expected. In other words for one gallon of fuel vehicles with *automatic* transmission can reach about 17.15 miles, while vehicles with *manual* transmission can reach about 24.39 miles. These values are the same as calculated means of fuel consumption for each transmission type separately (Table 1). The null hypothesis is that there is no linear relationship between predictor and response:

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

Since P-value of the slope (Table 3) is less than our desired type I error rate ($P < 0.05$), we can reject null hypothesis and state that transmission type had statically significant influence on fuel consumption.

Now, let's try to fit some other possible linear models that include other variables of the same data set, which could influence the impact of transmission type on fuel consumption. For this purpose I have splited this screening into two parts. First, I have examined what kind of influence on `am` variable's regression coefficient had all other potential regressors separately (Table 4). Second, I will choose regressors from the models where P-value is below our desired probability threshold for null hypothesis rejection ($P < 0.05$). These models are **4**, **7**, **8**, **9** and **10**, where additional regressors were `hp`, `vs`, `qsec`, `gear` and `carb`, respectively. Based on this information 7 new regression models are proposed in Table 5. For each of these models (**11-16**) P-value of the slope was less than 0.05, and therefore we can reject null hypothesis referring thereby that there is statistically significant relationship between predictor and outcome.

I would choose **Model 16** as the most promising since it includes all predictor variables from the previous screening, which influenced `am` variable in the way that they do not violate relationship between fuel consumption and transmission type. If we check difference between simple linear regression model (**Model 1**) and multivariable regression model (**Model 16**) we would get result indicating that it is statistically very significant (Table 6). That information give us certainty that additional regressors included in the new model significantly contribute in the influence of transmission type on the fuel consumption. Now, if we take a closer look on regression coefficients produced by our new model (**Model 16**), we can see that intercept has slightly increased (from **17.14 mpg** to **19.02 mpg**), while slope for the transmission type has decreased (from **7.25 mpg** to **3.97 mpg**). Regression line of Model 16 is presented as dashed blue line in Figure 1.

Nevertheless, if we apply 'Step-wise model search' with the `step` function for all variables in `mtcars` data set

```
fit_s <- lm(mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear + carb, data = mtcars)
step(fit, k = log(nrow(mtcars)))
```

we get result that the best model (**Model S**) that captures most of the variability in the data is simply `mpg ~ am + wt + qsec`. Regression line of Model S is presented as dash-dot green line in Figure 2. We will stick to this model and its regression coefficients are shown in Table 6. In other words after introducing new regressors into model our conclusion could be that expected fuel consumption in cars with automatic transmission is about **9.61 mpg**, and if you change your vehicle with some other that has manual transmission you can expect to drive about **2.93 miles** more for each gallon of fuel. If we observe residuals distribution for **Model S** we can conclude that it follows normal distribution curve (Figure 3), what is desirable feature and give us information that error values are random.

Appendix

Table 1: Mean values of fuel consumption for each type of transmission

Transmission type	Value	Mean (mpg)
automatic	0	17.15
manual	1	24.39

Table 2: Correlation coefficients of all variables in 'mtcars' dataset (except 'mpg') with 'am' varaible

cyl	disp	hp	drat	wt	qsec	vs	gear	carb
-0.523	-0.591	-0.243	0.713	-0.692	-0.23	0.168	0.794	0.058

Table 3: Regression coefficients of the basic model with only 'am' variable included as a predictor

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
am	7.244939	1.764422	4.106127	0.000285

Table 4: Regression coeficients for 'am' variable in linear models with addtitional regressor

Model	Estimate	Std. Error	t value	Pr(>|t|)
Model 1: mpg ~ am	7.2449	1.7644	4.1061	0.0003
Model 2: mpg ~ am + cyl	2.567	1.2914	1.9877	0.0564
Model 3: mpg ~ am + disp	1.8335	1.4361	1.2767	0.2118
Model 4: mpg ~ am + hp	5.2771	1.0795	4.8883	0
Model 5: mpg ~ am + wt	-0.0236	1.5456	-0.0153	0.9879
Model 6: mpg ~ am + drat	2.8071	2.2822	1.23	0.2286
Model 7: mpg ~ am + qsec	8.8763	1.2897	6.8827	0
Model 8: mpg ~ am + vs	6.0667	1.2748	4.7588	0
Model 9: mpg ~ am + gear	7.1416	2.9523	2.419	0.0221
Model 10: mpg ~ am + carb	7.6531	1.223	6.2579	0

Table 5: Regression coeficients for 'am' variable in linear models with addtitional multiple regressors

Model	Estimate	Std. Error	t value	Pr(>|t|)
Model 11: mpg ~ am + hp + qsec	5.9831	1.3381	4.4713	0.0001
Model 12: mpg ~ am + hp + vs	5.2985	1.0376	5.1067	0
Model 13: mpg ~ am + hp + carb	5.8802	1.1446	5.1374	0
Model 14: mpg ~ am + hp + qsec + vs	5.1496	1.4089	3.6551	0.0011
Model 15: mpg ~ am + hp + qsec + vs + gear	4.5668	1.9716	2.3162	0.0287

Model	Estimate	Std. Error	t value	Pr(>|t|)
Model 16: mpg ~ am + hp + qsec + vs + gear + carb	3.9689	1.9148	2.0728	0.0487

Table 6: Testing difference between Model 1 and Model 16 (ANOVA)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.8966	NA	NA	NA	NA
25	191.2449	5	529.6517	13.84747	0.0000016

Table 7: Regression coeficients for Model S

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
am	2.935837	1.4109045	2.080819	0.0467155
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162

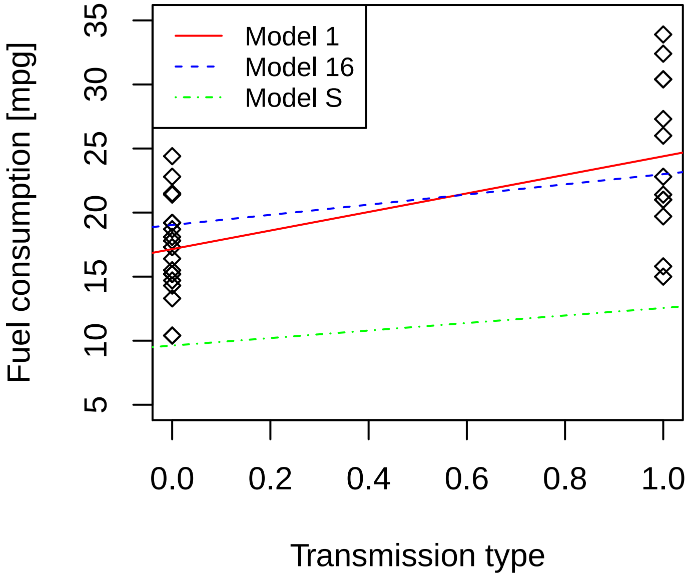

Figure 1: Relationship between fuel consumption and transmission type

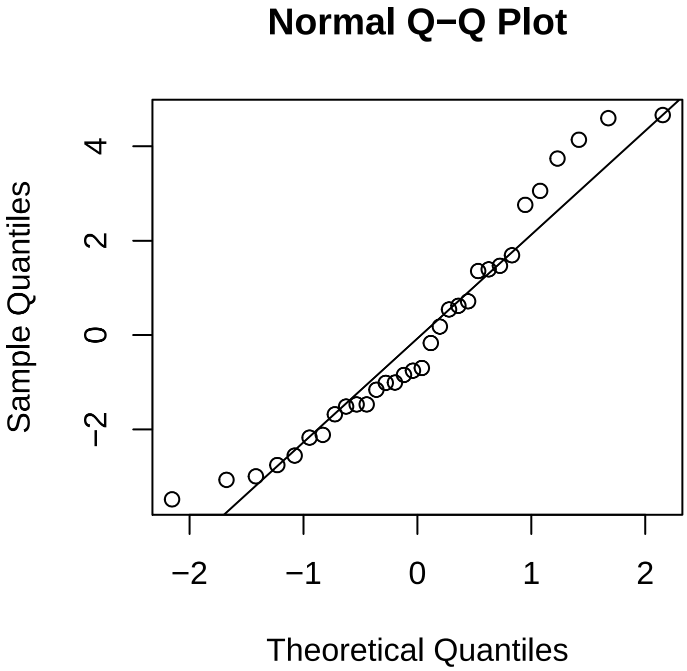

Figure 2: QQ Plot of residuals from Model S