INFLUENCE OF AUTOMATIC OR MANUAL TRANSMISSION ON THE CONSUMPTION IN MILES PER GALLON IN THE AUTOMOTIVE INDUSTRIE IN THE UNITED STATES

Version: V00

Date: 25-JAN-2015

GitHub Repository:
https://github.com/A6111E/datasciencecoursera/tree/master/Regression_Models

Data Source: R Data Set Package "data sets" - Data: "mtcars"

Synopsis:

In the attempt to predict gasoline mileage for 1973 - 1974 automobiles, road tests were preformed by Motor Trend US Magazine, in which gas mileage and 10 physical characteristics of various types of automobiles were recorded.

Source: Biometrics Invited Paper: The Analysis and Selection of Variables in Linear Regression by R. R. Hocking Biometrics - Vol. 32, No. 1 (Mar., 1976), pp.

DataSet Description: Extracted from Motor Trend US Magazine in 1.974 and comprises fuel consumption, 10 different aspects of automobile design and performance for 32 automobiles (1973-74 models).

- Data Frame: 32 observations - 11 variables

- Variable [mpg]: numeric - mpg (miles per US gallon)

- Variable [cyl]: numeric - Number of Cylinders

- Variable [disp]: numeric - Cylinder Displacement (cu.in)

- Variable [hp]: numeric - Gross Horsepower (hp)

- Variable [drat]: numeric - Rear Axle ratio

- Variable [wt]: numeric - Weight (lb/1000)

- Variable [qsec]: numeric - 1/4 Mile Time (sec)

- Variable [vs]: numeric - Cylinder Configuration (V - S: straight) (0 = V, 1 = S)

- Variable [am]: numeric - Transmission (0 = automatic, 1 = manual)

- Variable [gear]: numeric - Number of forward Gears

- Variable [carb]: numeric - Number of Carburetors

With this data set, this research tries to answer:

1. Is an automatic or manual transmission better for MPG (gas mileage)?
2. Quantify the MPG (gas mileage) difference between automatic and manual transmissions

Executive Summary:

- After a Exploratory Analysis and Model Finding, the best linear model, that predict the gasoline mileage for this data set is:

$$mpg = 9.62 - 3.92wt + 1.23qsec + 2.94am$$

- An Adjusted $R^2 = 0.83$, indicates that 83.36 of the variance in the outcome variable mpg can be explained by the predictors variables $wt, qsec, am$ and this linear model.
- The **Manual Transmission** is more efficient than the **Automatic Tranmission** in $2.94 miles/gal$

Exploratory Analysis:

Table 1: Data Preview

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
21.00	6.00	160.00	110.00	3.90	2.62	16.46	0.00	1.00	4.00	4.00
21.00	6.00	160.00	110.00	3.90	2.88	17.02	0.00	1.00	4.00	4.00
22.80	4.00	108.00	93.00	3.85	2.32	18.61	1.00	1.00	4.00	1.00
21.40	6.00	258.00	110.00	3.08	3.21	19.44	1.00	0.00	3.00	1.00
18.70	8.00	360.00	175.00	3.15	3.44	17.02	0.00	0.00	3.00	2.00
18.10	6.00	225.00	105.00	2.76	3.46	20.22	1.00	0.00	3.00	1.00

- The value of cylinder is a categorical variable
- Half of the cars get 19.20 miles per gallon or less (Median).
- The variability or spread for mpg is equal to $\sigma = 6.02$
- For the multivariate mpg data, there should be some relationship with the size of the engine (# of cylinders, gross horsepower or even the cylinder displacement).
- Plotting $mpg \sim cyl$, a decreasing trend on mpg as the # of cyl increases is obvious, and it should be a candidate fit a regression line.
- The summary on table "Table A1: Statistical Data Summary - Mileage per US Gallon " included on the Appendix, shows statistical data like the mean, standard deviation, variance, median, maximal and minimal values, for each type of

transmission (Automatic - Manual) in relationship with the Mileage per US Gallon (mpg).

- According with this table, the maximal and minimal values for **Manual Transmission**, are higher than corresponding values for **Automatic Transmission**, although its standard deviation is greater.
- Taking "Mileage per US Gallon (mpg)" as outcome, and using the other variables as predictors, statistical coefficients will be calculated for a linear model during a exploratory analysis.
- This can be observed on Graphic A1 included on the Appendix.

Initial Conclusions:

- **Manual Transmission**: higher throughput for the outcome "Mileage per US Gallon (mpg)" in compassion with the **Automatic Transmission**.
- **Transmision Type (am)**: according with the Statistic Coefficients Table and related graphic (please see the Appendix), the **am** variable is the **7th** place of influence on the **mpg** variable.
- According with the correlation (please see Appendix - Table A2) for the required outcome mpg and the predictors, the impact on it in descending order is: $1.wt, 2.cyl, 3.disp, 4.hp, 5.drat, 6.vs, 7.am, 8.carb, 9.gear, 10.qsec$
- The variable $qsec$ is the required time for driving $/1/4$ of mile. From physics in general, less time means more acceleration and final speed and more gas consumption. Although this variable has the smallest correlation with mpg, it can not be avoided from the model analysis.
- The model with mpg as outcome, should **NOT** have **ONLY** one predictor.
- The number of possible combinations for this linear model is 1023.
- Due to the high number of possible combinations, it's necessary to use an algorithm (Stepwise Algorithm, where the best model corresponds to the smallest AIC).

Statistical Inference

Hypothesis 1:

- Null hypothesis H_{01}: is that the residuals of the calculated linear model are normally distribute and do not have trends.
- Alternative Hypothesis H_{a1}: opposite as the null hypothesis.

Hypothesis 2:

- Null hypothesis H_{02}: is that the base model $mpg \sim am$ (trying to answer the main questions), is less representative that the best model (mpg as outcome and several variables as predictors).
- Alternative Hypothesis H_{a2}: opposite as the null hypothesis.

Modeling

- Including all variables in an initial linear model, all p-value´s are not less than 0.05, and a linear model including all variables **is not** representative.

Table 2: Summary Linear Model - All Variables

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	12.30	18.72	0.66	0.52
Cyl	-0.11	1.05	-0.11	0.92
Disp	0.01	0.02	0.75	0.46
Hp	-0.02	0.02	-0.99	0.33
drat	0.79	1.64	0.48	0.64
wt	-3.72	1.89	-1.96	0.06
qsec	0.82	0.73	1.12	0.27
vs	0.32	2.10	0.15	0.88
am	2.52	2.06	1.23	0.23
gear	0.66	1.49	0.44	0.67
carb	-0.20	0.83	-0.24	0.81

Linear Model Selection

The best fitting model for the data set has the following statistical data:

- R^2 = 0.85
- Adjusted R^2 = 0.83
- σ = 2.46
- Linear Model: $mpg = 9.62 - 3.92wt + 1.23qsec + 2.94am$
- An Adjusted R^2 = 0.83 (maximal value obtained for the best fitting linear model) value, indicates that 83.36% of the variance in the outcome variable mpg can be explained by the predictors variables: wt, $qsec$, am.
- The remaining 16.64% can be attributed to unknown, hidden or not included variables or inherent variability.

- The p_values values are all representative:

a. $pvalue(wt) = 6.95e-06$

b. $pvalue(qsec) = 2.16e-04$

c. $pvalue(am) = 4.67e-02$

- The maximal / minimal residual values are:

a. Maximal: 4.66

b. Minimal: -3.48

Statistical Inference Conclusions

- The Null hypothesis H_{01} *could not be rejected* (residuals of the calculated linear model are normally distribute and do not have trends).
- The Null hypothesis H_{02} *could be rejected* (base model $mpg \sim am$ is **NOT** representative to estimate the mpg).

Final Analysis:

- For defining which transmission, automatic or manual, is better for MPG (gas mileage) we have:

(1) **Variable [am]:** numeric - Transmission (0 = automatic, 1 = manual)

(2) **Slope [am]:** 2.94

(3) **Automatic Transmission =** $2.94 * 0 = 0$ replacing:

$$mpg = 9.62 - 3.92wt + 1.23qsec$$

(4) **Manual Transmission =** $2.94 * 1 = 2.94$ replacing:

$$mpg = 9.62 - 3.92wt + 1.23qsec + 2.94am$$

(5) **Findings:** taking wt and $qsec$ as constant values, the **Manual Transmission** adds 2.94 $miles/gal$ in comparison with the **Automatic Tranmission**

(6) In other words, the **Manual Transmission** is more efficient than the **Automatic**.

- For quantifying the MPG (gas mileage) difference between automatic and manual transmissions we have:

(1) **Variable [am]:** numeric - Transmission (0 = automatic, 1 = manual)

(2) **Slope [am]:** 2.94

(3) **Automatic Transmission =** $2.94 * 0 = 0$

(4) **Manual Transmission =** $2.94 * 1 = 2.94$

(5) **Findings:** the MPG (gas mileage) difference between automatic and manual transmissions is 2.94 $miles/gal$

Appendix

Graphic 0: Mean $mpg \sim Cylinder$

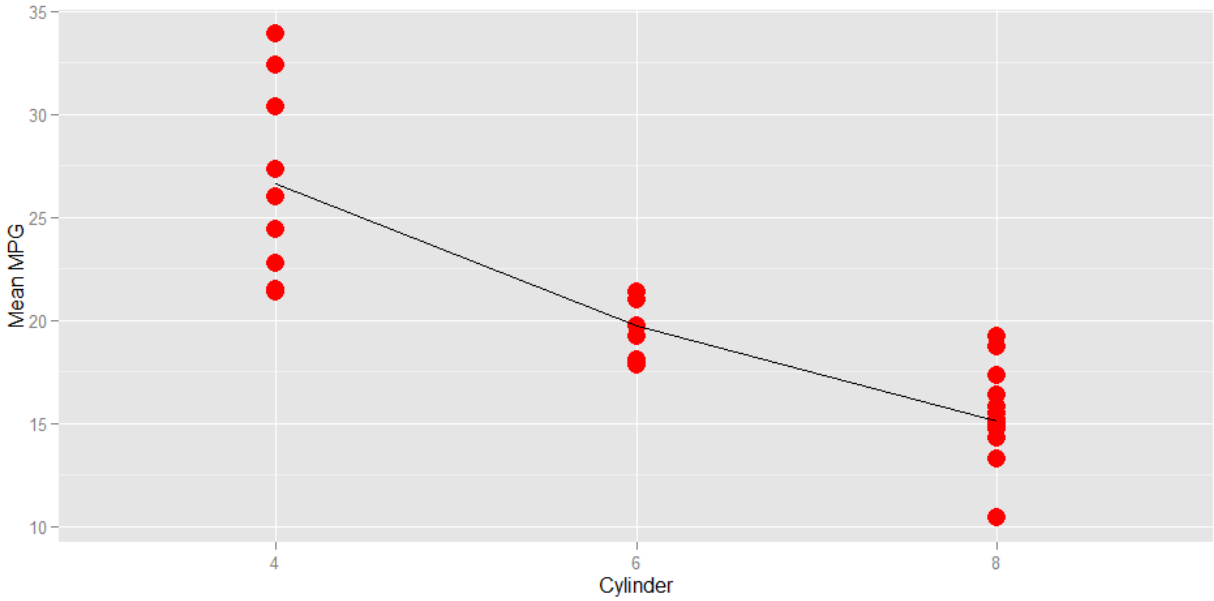


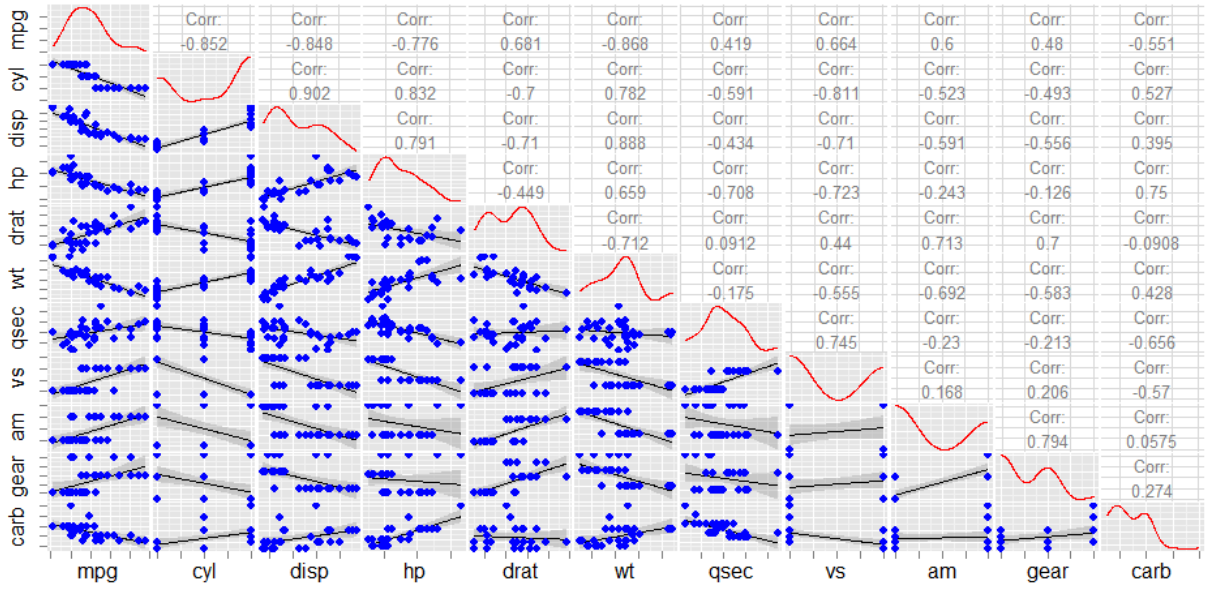
Table A1: Statistical Data Summary - Mileage per US Gallon

am	Average	Std_Deviation	Variance	Median	Max	Min
0.00	17.15	3.83	14.70	17.30	24.40	10.40
1.00	24.39	6.17	38.03	22.80	33.90	15.00

Remarks:

- Transmission Type (am): 1 (Manual Transmission) - 0 (Automatic Transmission)
- Average / Std_Deviation / Variance: statistic data for Mileage per US Gallon (mpg)
- GitHub Repository: /reports

Graphic 1: Motor Trend Car Road Tests - Linear Model



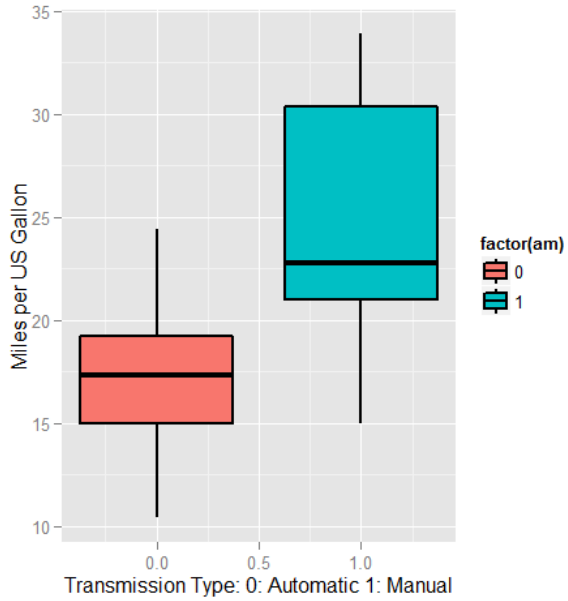


Table A2: Statistic Coefficients

Variable	Intercept_ Beta0	Slope_ Beta1	R^2	Adjusted_ R2	Correlation	Sigma
wt	37.28	-5.34	0.75	0.74	-0.87	3.05
cyl	37.88	-2.88	0.73	0.72	-0.85	3.21

disp	29.60	-0.04	0.72	0.71	-0.85	3.25
hp	30.10	-0.07	0.60	0.59	-0.78	3.86
drat	-7.53	7.68	0.46	0.45	0.68	4.49
vs	16.62	7.94	0.44	0.42	0.66	4.58
am	17.15	7.25	0.36	0.34	0.60	4.90
carb	25.87	-2.06	0.30	0.28	-0.55	5.11
gear	5.62	3.92	0.23	0.20	0.48	5.37
qsec	-5.11	1.41	0.17	0.15	0.42	5.56

Table A3: Summary Best Linear Model

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	9.6178	6.9596	1.38	0.1779
wt	-3.9165	0.7112	-5.51	0.0000
qsec	1.2259	0.2887	4.25	0.0002
am	2.9358	1.4109	2.08	0.0467

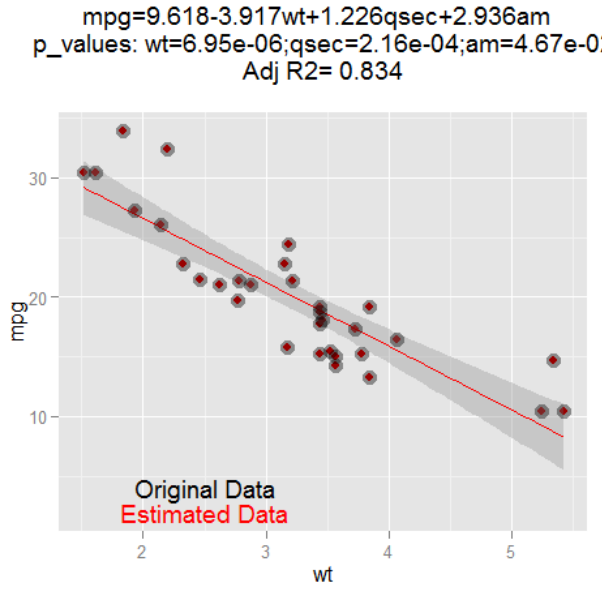
Table A4: Summary Anova - Best Linear Model

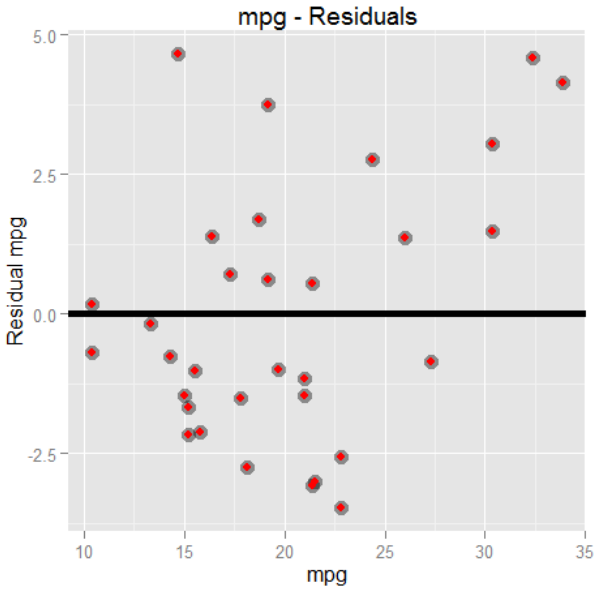
Res.Df	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr(>F)**
30	720.90				
28	169.29	2	551.61	45.62	0.0000

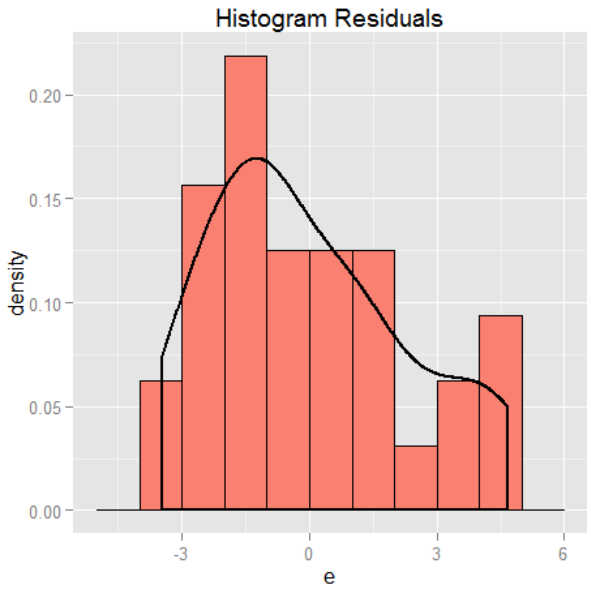
Table A5: Estimated Data vs Real Data

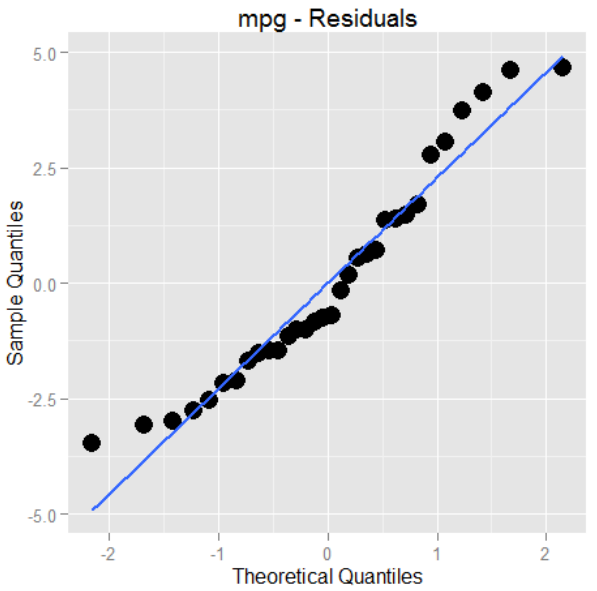
mpg	**wt**	**qsec**	**am**	**mpg_estimated**	**residuals**
21.00	2.62	16.46	1.00	22.47	-1.47
21.00	2.88	17.02	1.00	22.16	-1.16
22.80	2.32	18.61	1.00	26.28	-3.48
21.40	3.21	19.44	0.00	20.86	0.54
18.70	3.44	17.02	0.00	17.01	1.69
18.10	3.46	20.22	0.00	20.85	-2.75

Graphic A2: Best Linear Model - Residuals









Session Information

```
## R version 3.1.2 (2014-10-31)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
##
## locale:
## [1] LC_COLLATE=Spanish_Colombia.1252  LC_CTYPE=Spanish_Colombia.1252
## [3] LC_MONETARY=Spanish_Colombia.1252 LC_NUMERIC=C
## [5] LC_TIME=Spanish_Colombia.1252
```

```
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  method
s
## [8] base
##
## other attached packages:
## [1] gtools_3.4.1     car_2.0-22       GGally_0.5.0     plyr_1.8.1
## [5] gridExtra_0.9.1  knitr_1.8        xtable_1.7-4     ggplot2_1.0.0
## [9] data.table_1.9.4
##
## loaded via a namespace (and not attached):
##  [1] chron_2.3-45     colorspace_1.2-4 digest_0.6.4     evaluate_0.5.5
##  [5] formatR_1.0      gtable_0.1.2     htmltools_0.2.6  labeling_0.3
##  [9] MASS_7.3-35      munsell_0.4.2    nnet_7.3-8       proto_0.3-10
## [13] Rcpp_0.11.3      reshape_0.8.5    reshape2_1.4     rmarkdown_0.3.
3
## [17] scales_0.2.4     stringr_0.6.2    tools_3.1.2      yaml_2.1.13
```