Regression Analysis on Data "auto-mpg"

Chenze Li, Xi Chen, Zhentao Li

2021/12/7

Outline

- Introduction and Data Preprocessing
- Linear Regression Model Construction
- PCA to solve multicollinearity
- Conclusion

Data "auto-mpg"

Variables	Definition
mpg	Miles per gallon.
cylinders	Number of cylinders in a car.
displacement	Measure of the cylinder volume swept by all of the pistons of a piston engine.
horsepower	The measure of power of a car.
weight	The weight of a car.
acceleration	Car acceleration.
model_year	Year of the car model.
origin	The origin of the car (1=USA,2=Europe,3=Japan).
carname	The specific car model.

We are interested in predicting the amount of miles per gallon cars could drive with provided data.

Data preprocessing

- Delete all rows (6 rows) with missing value.
- Delete 7 rows with 3 or 5 cylinders.
- Some key features:
  - Strong multicollinearity between displacement, horsepower, weight and acceleration.
  - Both continuous and discrete predictors.

Basic Description

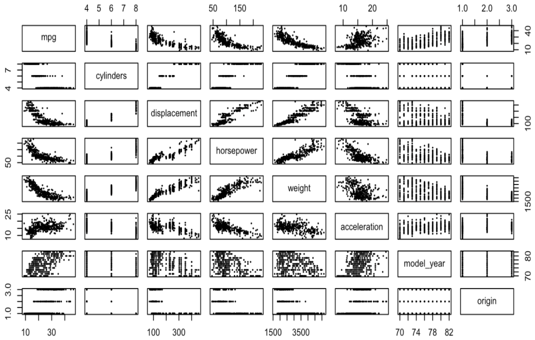


Figure: Scatter plot of data "auto-mpg"

Model building

Original model: **mpg** as a response and all other variables except **carname** as predictors. No interaction. (Linearity, Independence, Equal variance and Normal distribution)
Through boxcox, we apply log transformation to **mpg**.

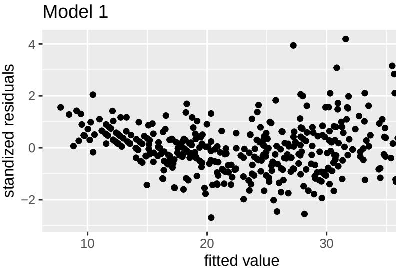


Figure: Original model without transformation

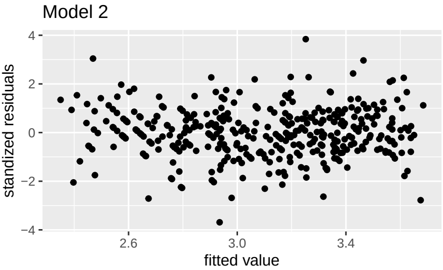


Figure: Updated model after log transformation

Figure: standardized residuals v.s. fitted values

Model building

- **Forward stepwise** method with AIC and BIC criteria to drop insignificant variables.
- Output:
  - AIC criterion : drop acceleration
  - BIC criterion : drop acceleration and displacement
- Choose the output with BIC criteria

Result

Model:

$$log(mpg) = \beta_0 + cylinders + horsepower + weight + model\ year + origin + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$

```
`model year`71 `model year`72 `model year`73 `model year`74
    0.04206910    -0.02075009    -0.02941795     0.05666581
`model year`75 `model year`76 `model year`77 `model year`78
    0.05076645     0.07519829     0.13681551     0.14027240
`model year`79 `model year`80 `model year`81 `model year`82
    0.22436391     0.32596341     0.26068069     0.29998160
```

Figure: coefficient of **model year**

Intercept	cylinders:6	cylinders:8	horsepower	weight	origin:2	origin:3
3.7733	-0.0820	-0.0360	-0.0013	-0.0002	0.0456	0.0595

Table: coefficient of other variables

Result

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
cylinders	2	31.92	15.962	1459.087	< 2e-16	***
horsepower	1	2.84	2.839	259.514	< 2e-16	***
weight	1	1.72	1.724	157.573	< 2e-16	***
model.year	12	4.23	0.353	32.242	< 2e-16	***
origin	2	0.14	0.071	6.482	0.00171	**
Residuals	366	4.00	0.011			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure: summary of Model

Inference

- Normal assumption:

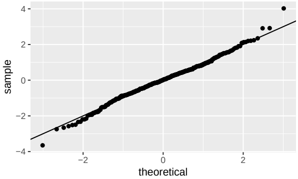


Figure: Q-Q plot

- Equal variance assumption:

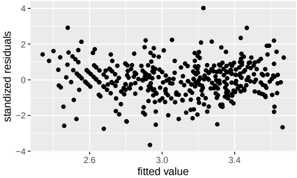


Figure: standardized residuals v.s. fitted values

Multicollinearity

	mpg	cylinders	displacement	horsepower	weight
mpg	1	-0.79	-0.82	-0.78	-0.84
cylinders		1	**0.95**	**0.85**	**0.90**
displacement			1	**0.90**	**0.94**
horsepower				1	**0.87**
weight					1

Table: Correlation Matrix of Cylinders, Displacement, Horsepower, Weight

From the correlation matrix, variables cylinders, displacement, horsepower and weight shows high positive correlation (>0.7), so risk of multicollinearity exists in this dataset. This is very intuitive. Actually, a car with more cylinders and displacement is expected to have more horsepower and bigger weight.

Multicollinearity

	GVIF
cylinders	**13.0**
displacement	**23.1**
horsepower	**10.6**
weight	**11.7**
acceleration	2.6
model year	1.3
origin	2.2

Table: GVIF of Different Variables

For further checking, we use GVIF (Generalized Variance Inflation Factor). It quantifies the severity of multicollinearity. GVIF of cylinders, displacement, horsepower and weight are bigger than 10, indicating high risk of multicollinearity. To solve the problem of multicollinearity, we can use Principle Component Analysis (PCA) to synthesize the 4 variables.

Principle Component Analysis

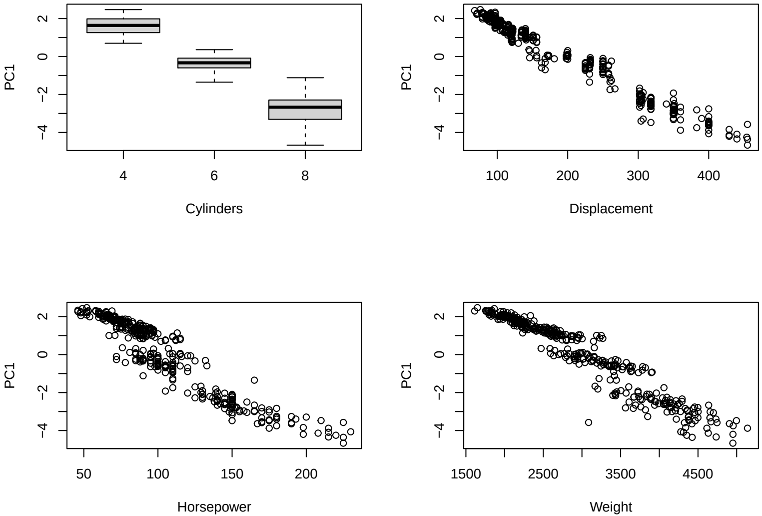
The result of PCA for cylinders, displacement, horsepower and weight is:

	PC1	PC2	PC3	PC4
Proportion of Variance	0.9264	0.0397	0.0247	0.0092

Table: PCA Results

Proportion of Variance of principle component 1 (PC1) is 0.9264, which means PC1 contains 92.6% of the information. So we treat PC1 as the composite of 4 variables with collinearity.
Then plot PC1 vs. Cylinders, Displacement, Horsepower and Weight to find what PC1 means in this dataset.

Principle Component Analysis



Principle Component Analysis

From these plots, PC1 is negatively correlated with these 4 variables.
Since a car with increased cylinders, displacement, horsepower and weight is expected as a higher performance or more fuel-consuming vehicle, PC1 is an indicator of car performance or fuel saving ability. Higher PC1 value indicates lower car performance and higher fuel saving ability.
Then we fit the new ANOVA model. After solving the problem of multicollinearity, the new model is more stable.

New model

Model:

$$log(mpg) = \beta_0 + PC1 + model\ year + origin + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$

```
model.year71 model.year72 model.year73 model.year74 model.year75 model.year76
-0.006531287 -0.066074779 -0.070227047 -0.009707565 -0.032330318  0.009800397
model.year77 model.year78 model.year79 model.year80 model.year81 model.year82
  0.08353729   0.07948941   0.18301792   0.25956265   0.19556247   0.24381521
```

Figure: coefficient of **Model Year**

Intercept	PC1	origin:2	origin:3
3.0116	0.1268	0.0439	0.0862

Table: coefficient of other variables

New model

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
             Df Sum Sq Mean Sq F value   Pr(>F)    
PC1           1  35.47   35.47 2625.09  < 2e-16 ***
model.year   12   4.12    0.34   25.43  < 2e-16 ***
origin        2   0.29    0.15   10.75 2.91e-05 ***
Residuals   369   4.99    0.01                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure: summary of Model

Conclusions

- Taking logarithm of mpg is better for model construction.
- Cylinders, Displacement, Horsepower and Weight are positively correlated. They can be considered together as cars' performance or fuel-consuming ability using PCA.
- Cars' MPG is getting smaller with cylinders, displacement, horsepower and weight grows.
- Cars produced in Japan has largest MPG in mean, while ones produced in US get smallest.
- Cars' MPG is getting larger with the model year grows.