REGRESSION ANALYSIS ON DATA "AUTO-MPG"

REPORT: FINAL PROJECT FOR STA 232A 2021F

Chenze Li, Xi Chen, Zhentao Li

Chenze Li: Model construction and inference in section 2

Xi Chen: Multicollinearity analysis and PCA in section 3

Zhentao Li: Data preprocessing and Ridge regression in section 3

UC Davis

December 9th 2021

1 Data preprocessing

The data "auto-mpg" contains MPG(miles per gallon) and other indexes of a car [Quinlan (1993)]. Our goal is to build a appropriate regression model for MPG as the outcome. We first delete all observations with missing value, and after checking the distribution of each predictor, we find there is of strong multicollinearity in variables weight, displacement and horsepower, which suggests a transform of linear model like PCA regression or Ridge regression could be considered.

2 Model Building

2.1 Model Construction

First, we consider the linear model without interaction between variables and treat **mpg** as the response variable(denoted as Y) and all the other variables except **carname** as predictors. But the outcome shows a problem. From the standardized residuals v.s. fitted values plot, points form a non-linear curve and width of the points are not equal along the x-axis which means unequal variance. Then we use **Box-cox** transformation to adjust our model.

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ lnY, & \text{if } \lambda = 0 \end{cases}$$

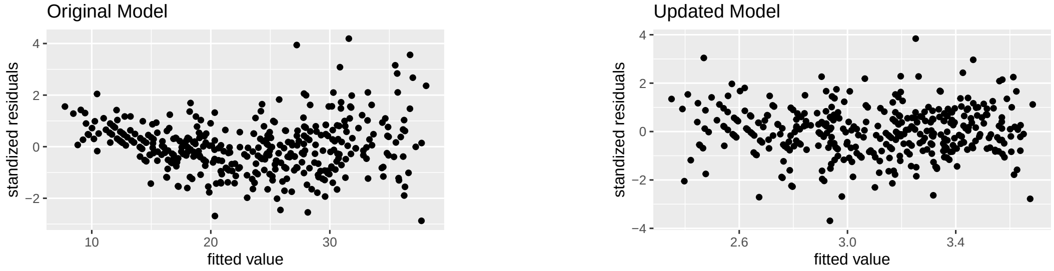

Figure 1: Original model without transformation Figure 2: Updated model after log transformation

Figure 3: standardized residuals v.s. fitted values

We make log-transformation of Y and after transformation, we find both problems are alleviated. Then we use the Akaike information criterion (a.k.a AIC) and the Bayesian information criterion (a.k.a BIC) as criteria to conduct **Forward stepwise** method. The followings are the definitions:

$$AIC = n \times log(\frac{SSE}{n}) + 2d.f.$$
$$BIC = n \times log(\frac{SSE}{n}) + log(n)d.f.$$

where n is sample size, $d.f.$ is degree of freedom and $SSE = \sum_i (Y_i - \hat{Y}_i)^2$.

From the result model, we find in AIC criterion, we drop **acceleration** and in BIC criterion, we drop **acceleration** and **displacement**. For the purpose of making model simple, we choose the model with BIC criterion.

The model is shown as below:

$$log(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \sum_{j=1}^{2} \gamma_j O_{ij} + \sum_{j=1}^{2} \alpha_j C_{ij} + \sum_{j=1}^{12} \omega_j M_{ij} + \epsilon_i$$

variable	cylinder	horsepower	weight	model year	origin
p-value	$< 2e^{-16}$	$< 2e^{-16}$	$< 2e^{-16}$	$< 2e^{-16}$	0.00171

Table 1: summary of variables

where $\epsilon_i \sim N(0, \sigma^2)$ and Y_i as **mpg**, X_{i1} as **horsepower**, X_{i2} as **weight**, O_{ij} as **origin**, C_{ij} as **cylinders** and M_{ij} as **model year**.Table 1 shows the result of F-test for each variable and since all the p-values are significantly small, all the variables are significant.

Intercept	cylinders:6	cylinders:8	horsepower	weight	origin:2
3.7733	-0.0820	-0.0360	-0.0013	-0.0002	0.0456
origin:3	model year:71	model year:72	model year:73	model year:74	model year:75
0.0595	0.0421	-0.0208	0.0294	0.0567	0.0508
model year:76	model year:77	model year:78	model year:79	model year:80	model year:81
0.0752	0.1368	0.1403	0.2244	0.3260	0.2607
model year:82					
0.3000					

Table 2: coefficient of variables

2.2 Inference

For this part, we check the normal assumption, linearity and homoscedasticity. From figure 5 and 2.2, we find the normal assumption, linearity and homoscedasticity are satisfied very well. And through figure 4, we find the model could output a good fit of the true values of **mpg**.

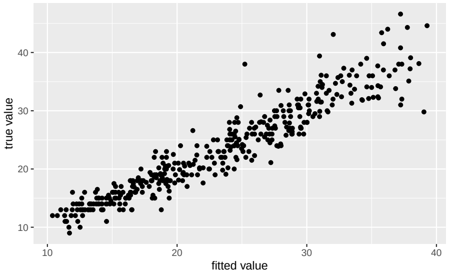

Figure 4: fitted value v.s. true value of **mpg**

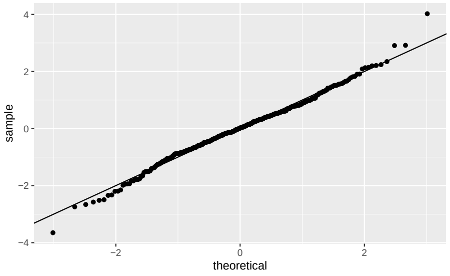

Figure 5: Q-Q plot

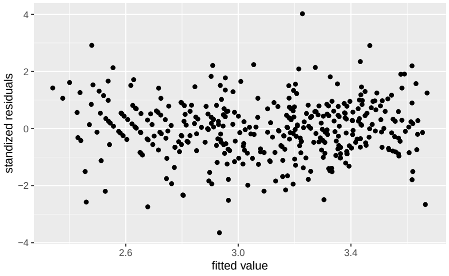

Figure 6: standardized residuals v.s. fitted values

3 Model adjustment

3.1 Multicollinearity

We check the correlation matrix of this data set, and found severe multicollinearity. Strong multicollinearity might make the design matrix X less than full rank, making $X^T X$ cannot be inverted. In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.

	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 3: 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 result is also intuitive, because a car with more cylinders and displacement is expected to have more horsepower and bigger weight.

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

Table 4: GVIF of Different Variables

For further checking, we use GVIF (Generalized Variance Inflation Factor). It quantifies the severity of multicollinearity [Fox and Monette (1992); Fox (1997)]. 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 [Venables and Ripley (2002)].

3.2 Principle Component Analysis (PCA)

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 5: 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.

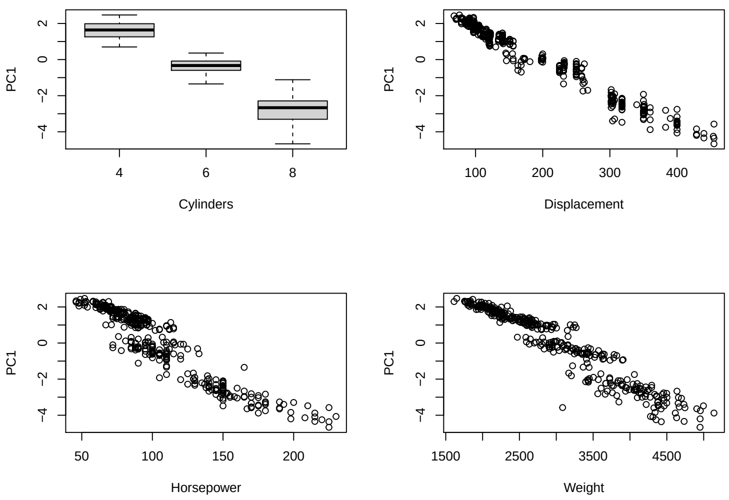

Figure 7: PC1 vs. Cylinders, Displacement, Horsepower and Weight

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.

3.3 Ridge Regression

Due to the strong multicollinearity in data, it is worth to try for a Ridge regression[Gruber (1998)] to improve our model. By applying R package **glmnet**, we provide a ideal pool of λ and use cross validation to pick up the optimal λ. As a results, we obtain a Ridge regression model, and it is necessary for us to compare it with the one of PCA method and simple linear regression.

	cylinders	displacement	horsepower	weight	acceleration	model year	origin
coef	-0.546	**-0.004**	**-0.024**	**-0.004**	-0.111	0.648	1.421

Table 6: Coefficients in Ridge regression model

From Ridge regression model, we could see that the estimation of three strongly correlated variables displacement, horsepower and weight are substantially distinguished with the ones of displacement and weight are shrinked to zero, while one of horsepower is non-zero. However, in this case $R^2 = 0.823$, which is smaller than 0.89 in simple linear regression, this fact suggests that Ridge regression does not make our model become more powerful for the response. Hence we choose PCA method rather than Ridge regression not only as a dimension-decreasing method, but also as a way to solve multicollinearity in original data.

3.4 Final Model

The new model is shown as below:

$$log(Y_i) = \beta_0 + \beta_1 X_{i1} + \sum_{j=1}^{2} \gamma_j O_{ij} + \sum_{j=1}^{12} \omega_j M_{ij} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and Y_i as **mpg**, X_{i1} as **PC1**, O_{ij} as **origin** and M_{ij} as **model year**.Table 6 shows the result of F-test for each variable for the new model and since all the p-values are significantly small, all the variables are significant.

variable	PC1	model year	origin
p-value	$< 2e^{-16}$	$< 2e^{-16}$	$2.91e^{-05}$

Table 7: summary of variables

Intercept	PC1	origin:2	origin:3	model year:71	model year:72
3.0116	0.1268	0.0439	0.0862	-0.0065	-0.0661
model year:73	model year:74	model year:75	model year:76	model year:77	model year:78
-0.0702	-0.0097	-0.0323	0.0098	0.0835	0.0795
model year:79	model year:80	model year:81	model year:82		
0.1830	0.2596	0.1956	0.2438		

Table 8: coefficient of variables

From the result, we can see that higher PC1 leads to bigger MPG values, which means MPG is getting smaller with cylinders, displacement, horsepower and weight grows. And for model year, cars after 1976 have bigger MPG, and the MPG is getting larger with the model year grows. For origin, cars made in Europe and Japan have better fuel-saving ability.

4 Conclusion

- 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.

References

J Ross Quinlan. Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning*, pages 236–243, 1993.

John Fox and Georges Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183, 1992.

John Fox. *Applied regression analysis, linear models, and related methods.* Sage Publications, Inc, 1997.

WN Venables and BD Ripley. Modern applied statistics with s. 4 edn springer-verlag. *New York*, 2002.

Marvin Gruber. *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators.* CRC Press, 1998.