

A Statistical Analysis on Traffic Accidents - Relations towards Yearly Fatality Rate

Renhao Cui, Kun Liu, Hongsen Shi

April 3, 2014

Applied Regression Analysis  
STAT 6450  
Spring 2014

**Abstract**

In this study, we implement the typical linear regression analysis on the traffic accident data, in order to find the relation between different variables. This includes model design, analysis, choosing the appropriate model based on several certain criteria. Although in most cases, a full model has the most power to describe the relations in terms of R square value, this study shows that some models may also be better candidates in some ways.

**Introduction**

Traffic accidents have always been serious problems across the world. It will be especially useful if the accident can be predicted, in order to control the situation and for properly distribution of forces to provide sufficient help in time. For insurance companies, it is also very helpful regarding the policy made for different circumstances and customers. However, it is barely possible to predict each individual accident as they are distinguished in almost all ways. On the other hand, statistical methods can be applied to predict the overall accident rate (in terms of fatality rate) in some meaningful ways.

In this study, we are working towards Yearly Fatality Rate and its relations with some other attributes such as population, licensed drivers, registered vehicles, and vehicle miles traveled. Also, in order to make the study more generalizable, we incorporate the data from different states (California, Ohio, Florida, and USA), with the trend of 18 years (1994 to 2011). To be more specific, in this study, we focus on the Fatality Rate per 100,000 Registered Vehicles in Ohio as a sample response variable, and research on its relations with other variables. We hope this study would reveal some factors that is related to the happening of the traffic accident, therefore to help the insurance company handle the accident and make policies in a better way.

**Assumptions**

In order to simplify the problem while keep close to the actual situation, we have the following assumptions for this study:

1. Yearly Fatality Rate is only related with population, licensed drivers, registered vehicles, and vehicle miles traveled in different ways. Some of these attributes have more impact than others.

2. Different states, as well as the whole nation, share some similarity in terms of the four variables mentioned above.

3. Data grows with time, which in this case, the year of the data collected. In other words, data has relation with the value of year.

4. When predicting a missing value from a certain year, data from the same year has the most impact.

5. When predicting the data for a future year, data from a single year has potential effect in the next year, and the next year only.

**Methods**

All the data we used is provided by The National Highway Traffic Safety Administration (NHTSA) [1]. It collects and publishes the data at national and state level records related to traffic accidents. We give the study on both the data of the country level (USA) and state level (CA, FL, and OH) from 1994 to 2011, but we only use the data from 1944 to 2010 to build the model, and use the data from 2011 as the evaluation data. Four fatality rates are under investigation in this study:

1. Yearly Fatality Rate per 100,000 Population

2. Yearly Fatality Rate per 100,000 Licensed Drivers

3. Yearly Fatality Rate per 100,000 Registered Vehicles

4. Yearly Fatality Rate per 100 Million Vehicle Miles Traveled

We apply linear regression as the statistic method, including single as well as multi-variable regression model. We consider adjusted R value, p value, along with Akaike’s information as criteria to determine the best model in different situations.

**Construction of general linear regression model**

In this study, we are particularly interested in predicting the Yearly Fatality Rate per 100,000 Registered Vehicles in Ohio using other variables as predictor. As a result, we construct a multi-variable regression model as follows:

(1)

Where:

and are parameters

are known constants

is independent 



By doing this, we are assuming that the all the errors are random and independent with time.

As the starting point, we suppose all variables have some linear relations with Yearly Fatality Rate per 100,000 Registered Vehicles in Ohio. Then in this case, a good approach would be to start with full model including all the rest variables as predictors.

**Construction of time series model**

We will start our examination of the effects of time by considering the time effect alone, using the models and methods of time-series analysis.

Three basic characteristics or features of a time series:

* secular trend, where the mean value E[X(t)] of the process seems to be changing (possibly decreasing) with time ;
* cyclic variation, where there is clear indication of a cycle or periodic pattern of variability;
* random error, where there is obviously further random varaition around any secular trend or cyclic variation -- such random changes may or may not be independent.
* serial correlation, where we expect successive Xt or X(t) to show correlation at the adjacent time points, possibly stronger the closer the time points are to each other.

In this study, we are concentrate on a discrete time series {Xt; t=0,1,2,...} and its realization {xt; t=0,1,2,...}. Consider first of all any possible trend in the process. We have:

(2)

where describes how E(Xt) varies with t (trend function) and describes random variation around the mean or trend. We would expect that E(=0 and normally assume that var(=.

In this study, we use linear regression model to estimate and eliminate the trend, which makes,

(3)

Now the trends can be turned into:

(4)

Then the autocorrelation functions (ACFs) and partial autocorrelation functions (PACFs) are examined to see which of the potential three patterns are present in the data. With autocorrelation, time series analysis can be more appropriate than multiple regression since the errors are correlated due to the patterns over time in the data.

**Model Selection**

In this study, we treat Fatality Rate per 100,000 Registered Vehicles in Ohio as the response variable, and treat all other attributes as potential predicted variables. Hereby we build several candidate models to study the relation towards the response variable. This includes one full model that tries to describe the relation using all the available variables, and some other reduced model which are picked from either human intuition about the problem or statistical analysis.

For simplicity, we defined the following notations for each attributes across states and the country:

Time in terms of year - year

For the state of California

Yearly Fatality Rate per 100,000 Population - XC1

Yearly Fatality Rate per 100,000 Licensed Drivers - XC2

Yearly Fatality Rate per 100,000 Registered Vehicles - XC3

Yearly Fatality Rate per 100 Million Vehicle Miles Traveled - XC4

For the state of Florida

Yearly Fatality Rate per 100,000 Population - XF1

Yearly Fatality Rate per 100,000 Licensed Drivers - XF2

Yearly Fatality Rate per 100,000 Registered Vehicles - XF3

Yearly Fatality Rate per 100 Million Vehicle Miles Traveled - XF4

For the state of Ohio

Yearly Fatality Rate per 100,000 Population - XO1

Yearly Fatality Rate per 100,000 Licensed Drivers - XO2

Yearly Fatality Rate per 100,000 Registered Vehicles - XO3

Yearly Fatality Rate per 100 Million Vehicle Miles Traveled - XO4

For the country of USA

Yearly Fatality Rate per 100,000 Population - XU1

Yearly Fatality Rate per 100,000 Licensed Drivers - XU2

Yearly Fatality Rate per 100,000 Registered Vehicles - XU3

Yearly Fatality Rate per 100 Million Vehicle Miles Traveled - XU4

Model selection procedures have been developed to identify a small group of regression models that are "good" according to a specified criterion. Then a detailed examination can be made of a limited number of the more promising or "candidate" models. In this paper, with the assumption that the number of observations (n) exceeds the maximum number of potential parameters (P), we consider the following criteria to assist our model selection:

* R2 or MSEp criterion

(5)

which is often referred as adjusted R2

* Mallows' Cp criterion

(6)

which is an estimate of

(7)

where and MSE(X1,..., Xp-1) is assumed to be an unbiased estimator of when the "full model" has been carefully chosen.

The Cp statistic is often used as a stopping rule for various forms of [stepwise regression](http://en.wikipedia.org/wiki/Stepwise_regression). Mallows proposed the statistic as a criterion for selecting among many alternative subset regressions. Under a model not suffering from appreciable lack of fit (bias), Cp has expectation nearly equal to p otherwise the expectation is roughly p plus a positive bias term.

* Akaike's information criterion (AICp)

(8)

which Akaike refers as "An Information Criterion" (AIC).

Since different models are built on different studies, the details of the models will be explained in the next section. In general, to make the whole process more efficient, we use three approaches to search for the candidate models: human intuition about the predictors, p-value of each predictor from a full mode, and sets of predictors picked from an analysis based on Cp criterion. Moreover, we also apply some analysis on pairwise variables to find potential multicollinearity, therefore further modify the candidate models.

**Results and Discussion**

**Analysis of relations between different variables.**

Coefficient of correlation is used to analyze the relationship between different variables. Table 1 presents the correlation between any pair of them in terms of the r value. Since the absolute values of most of the r values are relatively closed to 1, most pairs are linear associated in different levels0.

Table 1 correlation between pairs

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | XO1 | XO2 | XO3 | XO4 | XC1 | XC2 | XC3 | XC4 | XF1 | XF2 | XF3 | XF4 | XU1 | XU2 | XU3 | XU4 |
| XO1 | 1 | 0.957711 | 0.965128 | 0.985168 | 0.7889 | 0.81455 | 0.800115 | 0.785045 | 0.906918 | 0.820389 | 0.919022 | 0.923918 | 0.951063 | 0.945182 | 0.954147 | 0.929549 |
| XO2 | 0.957711 | 1 | 0.923044 | 0.949932 | 0.873592 | 0.888711 | 0.853823 | 0.8626566 | 0.924129 | 0.877096 | 0.869542 | 0.883346 | 0.949048 | 0.950227 | 0.9437300 | 0.9276211 |
| XO3 | 0.965128 | 0.923044 | 1 | 0.978091 | 0.732502 | 0.75987 | 0.842325 | 0.758224 | 0.817151 | 0.720387 | 0.940455 | 0.940684 | 0.898503 | 0.89675 | 0.9459138 | 0.9289512 |
| XO4 | 0.985168 | 0.949932 | 0.978091 | 1 | 0.79037 | 0.814614 | 0.854964 | 0.808475 | 0.872202 | 0.775574 | 0.946041 | 0.961154 | 0.933718 | 0.930775 | 0.9641232 | 0.9543396 |
| XC1 | 0.7889 | 0.873592 | 0.732502 | 0.79037 | 1 | 0.997831 | 0.908174 | 0.988743 | 0.927127 | 0.938883 | 0.792547 | 0.809925 | 0.921366 | 0.931284 | 0.8817307 | 0.9042920 |
| XC2 | 0.81455 | 0.888711 | 0.75987 | 0.814614 | 0.997831 | 1 | 0.915065 | 0.986907 | 0.939772 | 0.942659 | 0.818019 | 0.83372 | 0.938682 | 0.948418 | 0.9036386 | 0.9209828 |
| XC3 | 0.800115 | 0.853823 | 0.842325 | 0.854964 | 0.908174 | 0.915065 | 1 | 0.949676 | 0.812148 | 0.777422 | 0.891105 | 0.90956 | 0.87617 | 0.889985 | 0.9221400 | 0.9527202 |
| XC4 | 0.785045 | 0.862657 | 0.758224 | 0.808475 | 0.988743 | 0.986907 | 0.949676 | 1 | 0.891904 | 0.890339 | 0.820313 | 0.845793 | 0.908955 | 0.919971 | 0.8935282 | 0.9282756 |
| XF1 | 0.906918 | 0.924129 | 0.817151 | 0.872202 | 0.927127 | 0.939772 | 0.812148 | 0.891904 | 1 | 0.976097 | 0.854049 | 0.854766 | 0.975835 | 0.972897 | 0.9153699 | 0.9051885 |
| XF2 | 0.820389 | 0.877096 | 0.720387 | 0.775574 | 0.938883 | 0.942659 | 0.777422 | 0.890339 | 0.976097 | 1 | 0.757327 | 0.752806 | 0.925238 | 0.92703 | 0.8443406 | 0.8378480 |
| XF3 | 0.919022 | 0.869542 | 0.940455 | 0.946041 | 0.792547 | 0.818019 | 0.891105 | 0.820313 | 0.854049 | 0.757327 | 1 | 0.987607 | 0.919389 | 0.917751 | 0.9616765 | 0.959830 |
| XF4 | 0.923918 | 0.883346 | 0.940684 | 0.961154 | 0.809925 | 0.83372 | 0.90956 | 0.845793 | 0.854766 | 0.752806 | 0.987607 | 1 | 0.92311 | 0.92132 | 0.9654743 | 0.9748313 |
| XU1 | 0.951063 | 0.949048 | 0.898503 | 0.933718 | 0.921366 | 0.938682 | 0.87617 | 0.908955 | 0.975835 | 0.925238 | 0.919389 | 0.92311 | 1 | 0.998739 | 0.9769028 | 0.9662745 |
| XU2 | 0.945182 | 0.950227 | 0.89675 | 0.930775 | 0.931284 | 0.948418 | 0.889985 | 0.919971 | 0.972897 | 0.92703 | 0.917751 | 0.92132 | 0.998739 | 1 | 0.9796277 | 0.9696312 |
| XU3 | 0.954147 | 0.94373 | 0.945914 | 0.964123 | 0.881731 | 0.903639 | 0.92214 | 0.893528 | 0.91537 | 0.844341 | 0.961677 | 0.965474 | 0.976903 | 0.979628 | 1 | 0.9902204 |
| XU4 | 0.929549 | 0.927621 | 0.928951 | 0.95434 | 0.904292 | 0.920983 | 0.95272 | 0.928276 | 0.905189 | 0.837848 | 0.95983 | 0.974831 | 0.966275 | 0.969631 | 0.9902204 | 1 |

Figure 1,2,3,4 show the pair wise plot of four different variables in USA, California, Florida and Ohio. Figure 1 presents XU1, XU2, XU3 and XU4 are highly correlated since the scatter plots is close to a straight line. Figure 2 claims the linear relation between XC1 and XC2 are most significant, although all the four variables are also linearly correlated. Figure 3 shows that the pair of XF1 and XF2, XF3 and XF4 are highly correlated.

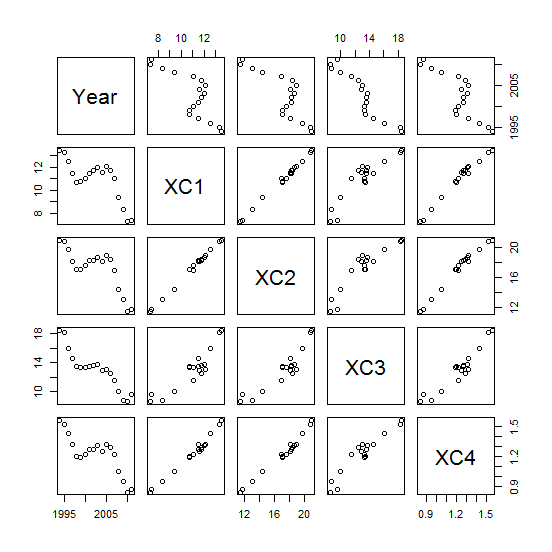
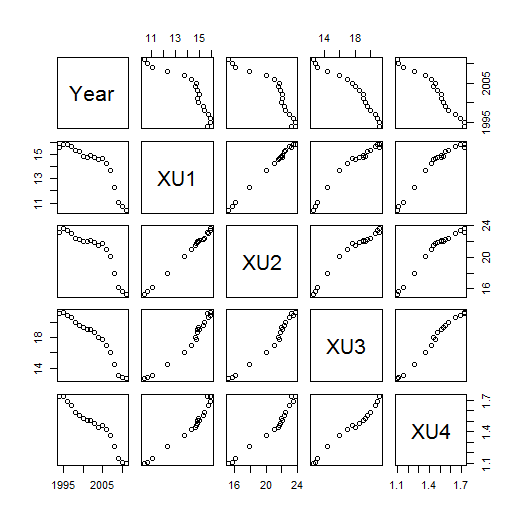


Figure 1 Pair wise plot for USA Figure 2 Pair wise plot for California

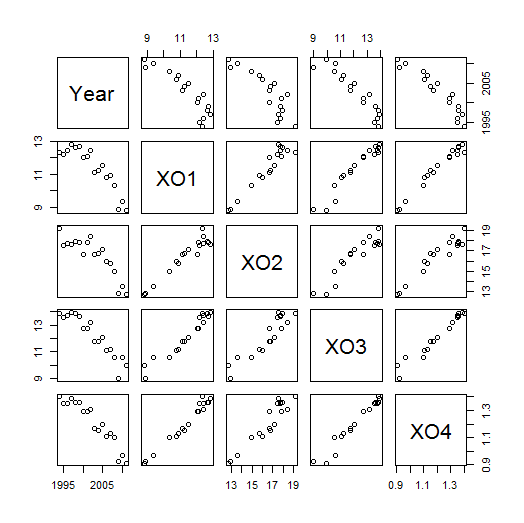
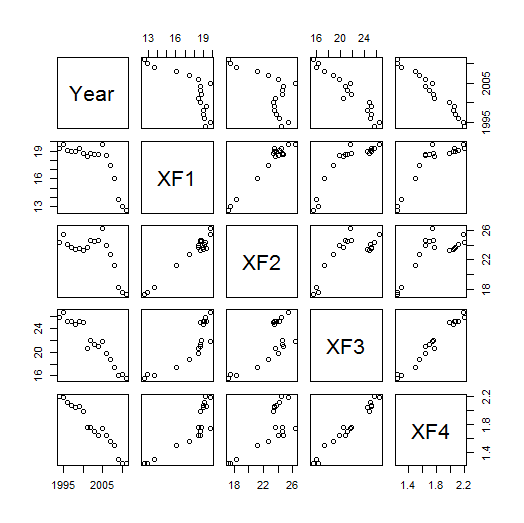
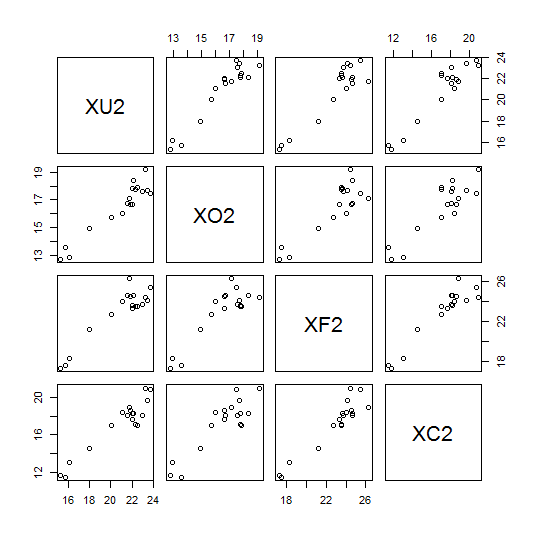


Figure 3 Pair wise plot for Florida Figure 4 Pair wise plot for Ohio

Figure 5a, 5b, 5c, 5d present the cross state relations for X1, X2, X3 and X4. We detect that except for the pair of Ohio and California, X1 variables are almost strongly positively linear associated. Moreover, except for the pair of Ohio with California, and Ohio with Florida, variables of X2, X3 and X4 are strongly positively linear associated across states.



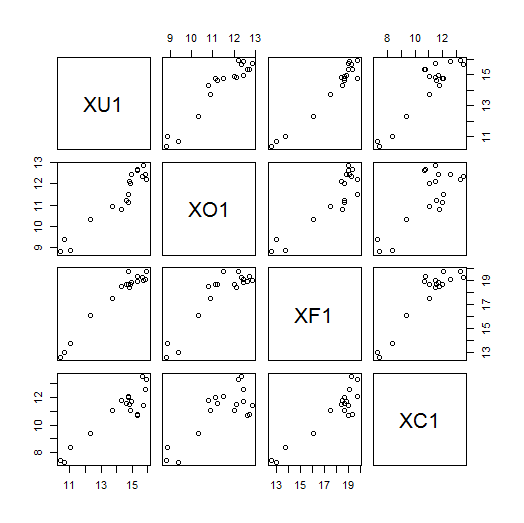
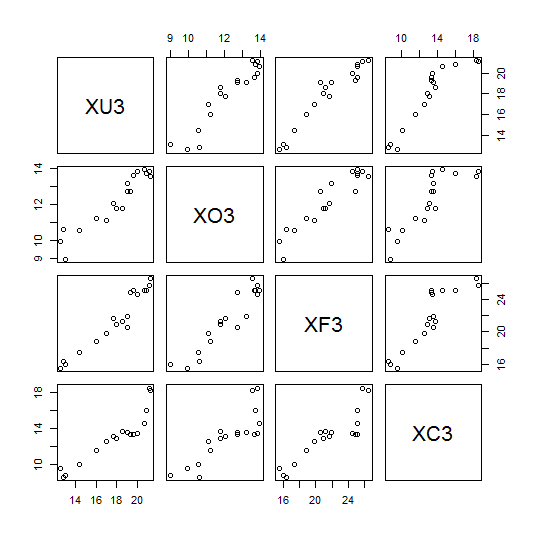
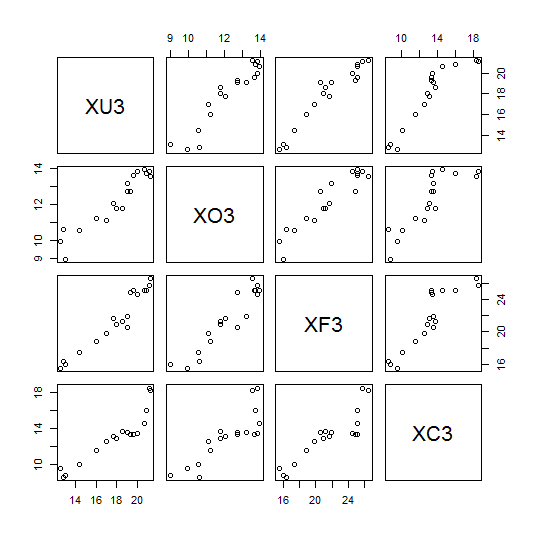


Figure 5a Cross State X1 correlation plot Figure 5b Cross State X2 correlation plot



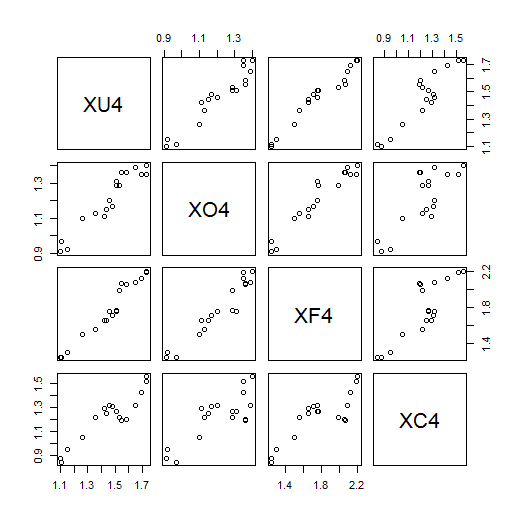


Figure 5c Cross State X3 correlation plot Figure 5d Cross State X4 correlation plot

**Time analysis in terms of Fatality Rates Per 100,000 Registered Vehicles**

Figure 6a, 6b, 6c and 6d contain the relations between Year and variables X3 in different states.

It seems that there is a trend between time and X3. As time increasing, the values of X3’s are gradually decreasing.



Figure 6a Year vs. XU3 Figure 6b Year vs. XC3



Figure 6c Year vs. XF3 Figure 6d Year vs. XO3

And it can be seen from the residual plots that the residuals seems normal and generally constant. The residuals seem to be not independent with time, which is consistent with our model.



Figure 7a Residual by Row Plot in USA

Figure 7b Residual by Row Plot in California

Figure 7c Residual by Row Plot in Florida

Figure 7d Residual by Row Plot in Ohio

**Relation study on predicting a missing data for a future year**

This study focuses on the relation in predicting the Yearly Fatality Rate per 100,000 Registered Vehicles in Ohio (XO3) for a future year, with all other variables at that year being observed.

With this setting, we start building from the full mode.

Full model: consider all other 15 attributes from the same year, that is XC1, XC2, XC3, XC4, XF1, XF2, XF3, XF4, XO1, XO2, XO4, XU1, XU2, XU3, XU4 as the predicted variables

With this full model, we then consider the case that the regression line goes through origin.

Reduced model 0: consider all other 15 attributes same as the full mode with the assumption that the regression line goes through the origin (intercept is 0)

Next we build the following reduced model 1-5 based on some human intuition and empirical experience.

Reduced model 1: assume the other three attributes within the same state and same year have the most impact, that is choosing XO1, XO2, XO4 as the predicted variables

Reduced model 2: assume the same type of attributes from other states and whole nation have the most impact, that is choosing XC3, XF3, XU3 as the predicted variables

Reduced model 3: assume the nationwide data has the most impact, that is choosing XU1, XU2, XU3, XU4 as the predicted variables

Reduced model 4: assume the nationwide data with the same type has the most impact, that is choosing XU3 as the predicted variable

Reduced model 5: assume the data grows with time in some relations, therefore year has the most impact, that is choosing year as the predicted variable

The output coefficients from the full model are shown in Table 2, ranking by variable significance based on the p-value. Thus reduced model 6-8 are chosen from the top three/seven/ten attributes from the result, with the assumption that these variables have the most impact.

Table 2 Full model estimated coefficients and p-value

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std.Error | t value | Pr(>|t|) |
| XO2 | 0.68118 | 0.05727 | 11.894 | 0.0534 |
| XU2 | 4.54729 | 0.56895 | 7.992 | 0.0792 |
| XC4 | 23.1748 | 3.26473 | 7.099 | 0.0891 |
| XC2 | -3.16311 | 0.56318 | -5.617 | 0.1122 |
| XO1 | 3.46395 | 0.67923 | 5.1 | 0.1233 |
| XU1 | -6.44549 | 1.30732 | -4.93 | 0.1274 |
| XF4 | 7.97766 | 1.69286 | 4.713 | 0.1331 |
| XO4 | -31.82412 | 6.85743 | -4.641 | 0.1351 |
| XC1 | 1.35914 | 0.49967 | 2.72 | 0.2243 |
| XU4 | -8.33452 | 3.12243 | -2.669 | 0.2282 |
| (Intercept) | -1.93833 | 0.80657 | -2.403 | 0.251 |
| XF2 | -0.21479 | 0.12825 | -1.675 | 0.3427 |
| XF1 | 0.54489 | 0.36405 | 1.497 | 0.375 |
| XF3 | -0.07736 | 0.07009 | -1.104 | 0.4686 |
| XU3 | 0.22898 | 0.25957 | 0.882 | 0.5398 |
| XC3 | 0.02859 | 0.06126 | 0.467 | 0.722 |

Reduced model 6: choosing XO2, XU2, XC4 as the predicted variables

Reduced model 7: choosing XO2, XU2, XC4, XC2, XO1, XU1, XF4 as the predicted variables

Reduced model 8: choosing XO2, XU2, XC4, XC2, XO1, XU1, XF4, XO4, XC1, XU4 as the predicted variables

More systematic approach has also been employed in this study, with Cp and adjusted R2 as selecting criterion. With the assistance of R package, we can examine a series of generated models with different index p and try to find some candidates with less predicting variables without losing too much accuracy. When using Mallows' Cp criterion, It is suggested that one should choose a subset that has Cp approaching p,[[4]](http://en.wikipedia.org/wiki/Mallows's_Cp#cite_note-4) from above, for a list of subsets ordered by increasing p. The stepwise selection based on Cp is shown in Figure 8. From this result, based on a smaller Cp value yet close to p from above, we picked out three models with p equals to 8, 9, and 11 respectively.



Figure 8 Model selection based on Mallows' Cp criterion

Reduced model 9: choosing XC1, XC4, XF3, XO2, XU1, XU2, XU3 as the predicted variables

Reduced model 10: choosing XC1, XC4, XF1, XF3, XO2, XU1, XU2, XU3 as the predicted variables

Reduced model 11: choosing XC1, XC4, XF1, XF4, XO1, XO2, XO4, XU1, XU2, XU3 as the predicted variables

By using or MSEp criterion, we achieved similar results (Figure 9) with reduced model 9-11 as our candidates.



Figure 9 Model selection based on MSEp criterion

Finally we consider the potential influence brought by multicollinearity. From the pairwise analysis (Table 1), we decide to treat the pair which has an r-value larger than 0.98 as a perfectly correlated pair, which gives us the following seven pairs:

XO1 ~ XO4, XC1 ~ XC2, XC1 ~ XC4, XC2 ~ XC4

XF3 ~ XF4, XU1 ~ XU2, XU3 ~ XU4

Then we choose the full model and reduced model 8 (which is the best model so far, and this can be shown in the following result) to remove one variable from each of the perfectly correlated pair, then generate reduced model 12 and 13.

Reduced model 12: choosing XC1, XC, XF1, XF2, XF3, XO1, XO2, XU1, XU3 as the predicted variables

Reduced model 13: choosing XO2, XC2, XO1, XU1, XF4, XC1, XU4 as the predicted variables

Table 3 summarizes the 14 reduced model derived from the full model. We now need to balance the prediction accuracy and the number of variable to pick one or more suitable models in this problem.

Table 3 Summary of selected model candidates

|  |  |  |
| --- | --- | --- |
|  | model candidates | p |
| Reduced model 0 | 0+XC1+XC2+XC3+XC4+XF1+XF2+XF3+XF4+XO1+XO2+XO4+XU1+XU2+XU3+XU4 | 15 |
| Reduced model 1 | XO1+XO2+XO4 | 4 |
| Reduced model 2 | XC3+XF3+XU3 | 4 |
| Reduced model 3 | XU1+XU2+XU3+XU4 | 5 |
| Reduced model 4 | XU3 | 2 |
| Reduced model 5 | year | 2 |
| Reduced model 6 | XO2+XU2+XC4 | 4 |
| Reduced model 7 | XO2+XU2+XC4+XC2+XO1+XU1+XF4 | 8 |
| Reduced model 8 | XO2+XU2+XC4+XC2+XO1+XU1+XF4+XO4+XC1+XU4 | 11 |
| Reduced model 9 | XC1+XC4+XF3+XO2+XU1+XU2+XU3 | 8 |
| Reduced model 10 | XC1+XC4+XF1+XF3+XO2+XU1+XU2+XU3 | 9 |
| Reduced model 11 | XC1+XC4+XF1+XF4+XO1+XO2+XO4+XU1+XU2+XU3 | 11 |
| Reduced model 12 | XC1+XC3+XF1+XF2+XF3+XO1+XO2+XU1+XU3 | 10 |
| Reduced model 13 | XO2+XC2+XO1+XU1+XF4+XC1+XU4 | 8 |

Detailed results are presented as follows:

Table 4 Summarized R-squre, AIC, residual and confidence interval for each model candidates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | adjusted R-square | AIC | residual | confidence interval |
| full model | 0.9997 | -142.09 | 0.00252 | 0.6227 |
| reduced model 0 | 1 | -111.564 | 0.00357 | 0.38804 |
| reduced model 1 | 0.9543 | -35.8592 | 0.831215 | 0.73839 |
| reduced model 2 | 0.8893 | -20.8277 | 0.932516 | 1.276696 |
| reduced model 3 | 0.8946 | -21.0203 | 0.805008 | 1.513536 |
| reduced model 4 | 0.8767 | -20.5554 | 1.044869 | 1.220235 |
| reduced model 5 | 0.8445 | -16.622 | 0.44451 | 1.149162 |
| reduced model 6 | 0.8528 | -15.9859 | 0.941241 | 1.59084 |
| reduced model 7 | 0.9782 | -46.7324 | 0.3466 | 0.80511 |
| reduced model 8 | 0.9993 | -105.344 | 0.00898 | 0.19177 |
| reduced model 9 | 0.9786 | -47.004 | 0.23574 | 0.819365 |
| reduced model 10 | 0.9866 | -47.004 | 0.09598 | 0.70955 |
| reduced model 11 | 0.9701 | -42.1949 | 0.26312 | 1.103875 |
| reduced model 12 | 0.9901 | -60.4003 | 0.17321 | 0.59596 |
| reduced model 13 | 0.9654 | -38.8332 | 0.53466 | 0.919911 |

Figure 10a Adjusted R-Square for the case of predicting same year data

Residuals and confidence intervals are generated using the data from 2010.

Figure 10b Residual for the case of predicting same year data

Figure 10c Confidence Interval for the case of predicting same year data

Figure 10d AIC for the case of predicting same year data

Adjusted R-square and AIC have been modified in the following figure, in order to show all the trend in terms of different criteria across models.

Figure 10e Combined Criteria for the case of predicting same year data

A good model should have a high adjusted R-square value, while keeping AIC, residual, and confidence interval at a small number. In this case, we believe reduced model 8 is a good candidate for this study according to the overall trend in figure 10e.

Moreover, comparing the result of full model and reduced model 12, reduced model 8 and reduced model 13, we find that considering the multicollinearity does not help in generating a better model. This might be caused by the picking of the perfectly correlated pairs and the comparing models, and should be taken more carefully in any future studies.

Finally, we use the data from 2011 to verify our approach in picking the right model. As shown in figure 11, model 8 is not necessary the best model in terms of correctly predicting the data for 2011, but is a good candidate considering the generality of the model.

Figure 11 Model evaluation for XO3 on the data of 2011

**Relation study on predicting the data in an upcoming year**

This study the relation in predicting the Yearly Fatality Rate per 100,000 Registered Vehicles in Ohio (XO3) for 2012, with all other variables being observed from 1994 to 2011. In this scenario, we are proposing to use 2011 data as predictor for 2012 value. Thus, compared to previous section, the model construction alters from using data from the same year as predictor to using data from the previous year. It is of great chance that the model would be less accurate in prediction, but we may still find use to some of the selected model which addresses more practical questions that an insurance company would concern.

Since we are using data from previous year to predict XO3, the previous XO3 observation may also be one of those predictors for its next year. But this would encounter another issue that the number of observations (n=17) is less than the maximum number of potential parameters (P). Simply adding the previous data of XO3 would results in the degree of freedom equaling to zero. So we chose to introduce XO3 by replace each of the 15 variables in the full model to find the best fitting model. Under this concern, we build the full model.

Full model: treating XC1, XC2, XC4, XF1, XF2, XF3, XF4, XO1, XO2, XO3, XO4, XU1, XU2, XU3, XU4 as the predicted variables

Similarly, by looking at all the 15 predicting variables and try to eliminate some variables which may be conceptually less relevant, we have the following reduced models.

Reduced model 0: consider all other 15 attributes from the full model, and the XO3 from the previous year (in is excluded from the full model because of the limited sample size), with the assumption that the regression line goes through the origin (intercept is 0)

Reduced model 1: assume the data from the same state has the most impact, that is choosing XO1, XO2, XO4, and XO3 from the previous year as the predicted variables

Reduced model 2: assume the same type of attributes have the most impact, that is choosing XC3, XF3, XU3, and XO3 from the previous year as the predicted variables

Reduced model 3: assume the nationwide data has the most impact, that is choosing XU1, XU2, XU3, XU4 as the predicted variables

Reduced model 4: assume the nationwide data with the same type has the most impact, that is choosing XU3 as the predicted variable

Reduced model 5: assume the data grows with time in some relations, therefore year has the most impact, that is choosing year as the predicted variable

The output coefficients of full model are shown in Table 5, ranking by variable significance based p-value. From the table we could conclude that the significance level is more close to each

other between variables compared to the previous problem. Yet we may still make a few selections based on the p-value rank. So reduced model 6 - 8 are chosen from the top three/seven/ten attributes from the result of full model, which have the most significant p values, with the assumption that these variables have the most impact.

Table 5 Full model estimated coefficients and p-value

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std.Error | t value | Pr(>|t|) |
| XC1 | -41.946 | 12.7355 | -3.294 | 0.188 |
| XC2 | 81.9376 | 27.3609 | 2.995 | 0.205 |
| XU2 | -113.3547 | 38.5952 | -2.937 | 0.209 |
| XU1 | 164.0621 | 56.4369 | 2.907 | 0.211 |
| XU4 | 198.0035 | 68.231 | 2.902 | 0.211 |
| XF4 | -194.0337 | 67.6909 | -2.866 | 0.214 |
| XO3 | 22.9278 | 8.062 | 2.844 | 0.215 |
| XC4 | -555.2677 | 196.2202 | -2.83 | 0.216 |
| XF1 | -16.3315 | 5.8195 | -2.806 | 0.218 |
| XO2 | -15.9807 | 5.6829 | -2.812 | 0.218 |
|  |  |  |  |  |
| XO1 | -82.8015 | 29.725 | -2.786 | 0.219 |
| XO4 | 768.3385 | 277.2035 | 2.772 | 0.22 |
| (Intercept) | 47.6901 | 17.2336 | 2.767 | 0.221 |
| XF2 | 6.1162 | 2.2253 | 2.748 | 0.222 |
| XF3 | 2.1389 | 0.9236 | 2.316 | 0.26 |
| XU3 | -6.0824 | 3.0698 | -1.981 | 0.298 |

Reduced model 6: choosing XC1, XC2, XU2 as the predicted variables

Reduced model 7: choosing XC1, XC2, XU2, XU4, XU1, XF4, XO3 as the predicted variables

Reduced model 8: choosing XC1, XC2, XU2, XU4, XU1, XF4, XO3, XC4, XF1, XO2 as the predicted variables

The stepwise selection based on Cp is shown in Figure 12. From this result, based on a smaller Cp value yet close to p from above, we picked out the best model as the next reduced model.



Figure 12 Model selection based on Mallows' Cp criterion

Reduced model 9: only consider XC1, XC2, XF1, XF2, XU1, XU2, XU3 as the predicted variables

By using or MSEp criterion, we achieved same results (Figure 13) with reduced model 9 as our candidates.



Figure 13 Model selection based on MSEp criterion

Then we assume the data is only affected from its value of the previous year.

Reduced model 10: choosing XO3 from the previous as the predicted variable

Applying the same method as the previous study, we generate two more models from the full model and reduced model 9, considering the problem of perfectly correlated pairs.

Reduced model 11: choosing XC1, XF1, XF2, XF3, XO1, XO2, XO3, XU1, XU3 as the predicted variables

Reduced model 12: choosing XC1, XF1, XF2, XU1, XU3 as the predicted variables

Table 6 Summary of selected model candidates

|  |  |  |
| --- | --- | --- |
|  | model candidates | p |
| Reduced model 0 | 0+XC1+XC2+XC3+XC4+XF1+XF2+XF3+XF4+XO1+XO2+XO3+XO4+XU1+XU2+XU3+XU4 | 16 |
| Reduced model 1 | XO1+XO2+XO3+XO4 | 5 |
| Reduced model 2 | XC3+XF3+XO3+XU3 | 5 |
| Reduced model 3 | XU1+XU2+XU3+XU4 | 5 |
| Reduced model 4 | XU3 | 2 |
| Reduced model 5 | year | 2 |
| Reduced model 6 | XC1+XC2+XU2 | 4 |
| Reduced model 7 | XC1+XC2+XU2+XU4+XU1+XF4+XO3 | 8 |
| Reduced model 8 | XC1+XC2+XU2+XU4+XU1+XF4+XO3+XC4+XF1+XO2 | 11 |
| Reduced model 9 | XC1+XC2+XF1+XF2+XU1+XU2+XU3 | 8 |
| Reduced model 10 | XO3 | 2 |
| Reduced model 11 | XC1+XF1+XF2+XF3+XO1+XO2+XO3+XU1+XU3 | 10 |
| Reduced model 12 | XC1+XF1+XF2+XU1+XU3 | 6 |
|  |  |  |
|  |  |  |

Detailed results are presented as follows:

Table 7 Summarized R-squre, AIC, residual and confidence interval for each model candidates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | adjusted R-square | AIC | residual | confidence interval |
| full model | 0.9792 | -67.7769 | 0.002692 | 5.568661 |
| reduced model 0 | 0.9985 | -41.4726 | 0.041544 | 12.02571 |
| reduced model 1 | 0.6873 | -1.43289 | 0.71599 | 3.378017 |
| reduced model 2 | 0.8018 | -9.82149 | 0.611079 | 2.460934 |
| reduced model 3 | 0.8609 | -15.2059 | 0.04255 | 1.795817 |
| reduced model 4 | 0.822 | -13.2213 | 0.622484 | 1.513991 |
| reduced model 5 | 0.8674 | -18.2188 | 0.113725 | 1.096448 |
| reduced model 6 | 0.7508 | -5.93346 | 0.457575 | 2.151328 |
| reduced model 7 | 0.9315 | -26.1471 | 0.144564 | 1.565604 |
| reduced model 8 | 0.9392 | -29.0705 | 0.057105 | 1.751573 |
| reduced model 9 | 0.9754 | -43.5313 | 0.031618 | 0.851622 |
| reduced model 10 | 0.7356 | -6.49172 | 0.60774 | 1.261634 |
| reduced model 11 | 0.8836 | -17.401 | 0.030123 | 2.328332 |
| reduced model 12 | 0.8941 | -19.3244 | 0.13853 | 1.664578 |

Figure 14a Adjusted R-Square for the case of predicting data for a future year

Figure 14b Residual for the case of predicting data for a future year

Figure 14c Confidence Interval for the case of predicting data for a future year

Figure 14d AIC for the case of predicting data for a future year

To better fit the data trend into one graph for analysis, we modify the value of adjusted R-square and AIC, also remove two data points from confidence interval trend since they are consider as outliers.

Figure 14e Combined Criteria for the case of predicting data for a future year

Similar to the previous study, a good model should have a high adjusted R-square value, while keeping AIC, residual, and confidence interval at a small number. In this case, we believe reduced model 9 is a good candidate for this study according to the overall trend in figure 14e.

Moreover, comparing the result of full model and reduced model 11, reduced model 9 and reduced model 12, we find the same conclusion that considering the multicollinearity does not necessarily help in generating a better model. This might be caused by the picking of the perfectly correlated pairs and the comparing models, and should be taken more carefully in any future studies.

Finally, we use the data from 2011 to verify our approach in picking the right model. As shown in figure 15, similar to the previous study, model 9 is not the best model in predicting the data for 2011. However, this cannot conclude that our approach is not correct, as using one data point for evaluation does not provide enough proof about the model in describing the relation in a broad case.

Figure 15 Model evaluation for XO3 on the data of 2011

**Predicting XO3 using specific time-series analysis**

Our first consideration is to eliminate any trend within the time series. The simple linear regression model (reduced model 5) can be used as trend model. After removing the trend, Figure 14 shows the autocorrelation function and partial autocorrelation are plotted for identification of the pattern. From the graph, no typical pattern was observed, which may indicate that this dataset may not be suitable for time-series analysis. Yet we still use auto-regression model in this study as one alternative.

The fitted auto-regressive model selected order p=0, estimated . The model prediction is shown in Figure 15.



Figure 16 autocorrelation and partial autocorrelation function



Figure 17 Prediction of auto-regression model from 2011

**Conclusion**

We implement two typical multi-variable regression analysis on two different tasks. From the candidate models we generated, a reasonable one is picked for each task considering several criteria that balance the generality and model simplicity, to describe the relation and predict the target variable. Although when using a future year to evaluate the model, it does not give the best result, we believe the whole process is still valid to provide some useful information that reveal some relations which can be used for the insurance company to better handle policy making or accident analysis.

With relatively short time periods (1994-2011), using time-series analysis may not be suitable in this case compared to multi-variable regression analysis. The autocorrelation function and partial autocorrelation plot did not show typical patterns of autocorrelation. However, with more years of observation, the time-series analysis may become feasible especially in predicting those cyclic variations.

**References**

1. Kutner, M., C. Nachtsheim, J. Neter. (2004). Applied Linear Regression Models: 4th edition. McGraw-Hill/Irwin.
2. The National Highway Traffic Safety Administration. http://www-fars.nhtsa.dot.gov
3. Vic Barnett. (2004).Environmental Statistics: Methods and Application. John Wiley & Sons, Ltd.
4. Daniel, C.; Wood, F. (1980). Fitting Equations to Data (Rev. ed.). New York: Wiley & Sons, Inc.

Attachment: R code

##### Model Selection#####

#### Problem 3 #### same year prediction ####

##### read data #####

my.data=read.csv("data3.csv",header=T)

names(my.data)=c("year","XC1","XC2","XC3","XC4","XF1","XF2","XF3","XF4","XO1","XO2","XO3","XO4","XU1","XU2","XU3","XU4")

##### full model #####

fmodel=lm(XO3~XC1+XC2+XC3+XC4+XF1+XF2+XF3+XF4+XO1+XO2+XO4+XU1+XU2+XU3+XU4,data=my.data)

##### AIC #####

library(MASS)

AICp = stepAIC(fmodel, direction="both")

AICp$anova

##### autoselection with both direction #####

step(Base, scope = list( upper=fmodel, lower=~1 ), direction = "both", trace=FALSE)

dropterm( fmodel, test = "F" )

##### Cp #####

library(leaps)

p=leaps(x=my.data[,c(2:11,13:17)], y=my.data[,12],names=names(my.data)[c(2:11,13:17)],

method="Cp")

plot(p$size,p$Cp,xlab="p", ylab="Cp")

plot(p$size[21:141],p$Cp[21:141],xlab="p", ylab="Cp")

##### adjusted R2 ######

r=leaps(x=my.data[,c(2:11,13:17)], y=my.data[,12],names=names(my.data)[c(2:11,13:17)],

method="adjr2")

plot(r$size,r$adjr2)

plot(r$size[21:141],r$adjr2[21:141])

#####selected model#####

### 1. XO3~XC1+XC4+XF3+XO2+XU1+XU2+XU3 ### P=8 ### Order=61###

s.model1=lm(XO3~XC1+XC4+XF3+XO2+XU1+XU2+XU3,data=my.data)

p$Cp[61]

r$adjr2[61]

### 2. XO3~XC1+XC4+XF1+XF3+XO2+XU1+XU2+XU3 ### P=9 ### Order=71###

s.model2=lm(XO3~XC1+XC4+XF1+XF3+XO2+XU1+XU2+XU3,data=my.data)

p$Cp[71]

r$adjr2[71]

### 3. XO3~XC1+XC4+XF1+XF4+XO1+XO2+XO4+XU1+XU2+XU3 ### P=11 ### Order=91###

s.model3=lm(XO3~XC1+XC4+XF1+XF4+XO1+XO2+XO4+XU1+XU2+XU3,data=my.data)

p$Cp[91]

r$adjr2[91]

### 4. XO3~XC1+XC2+XC4+XF1+XF2+XF4+XO1+XO2+XO4+XU1+XU2+XU3+XU4 ### P=14 ### Order=123### Cloest Cp to p from above

s.model4=lm(XO3~XC1+XC2+XC4+XF1+XF2+XF4+XO1+XO2+XO4+XU1+XU2+XU3+XU4,data=my.data)

p$Cp[123]

r$adjr2[123]

### 5. XO3~XC1+XC2+XC4+XF1+XF2+XF3+XF4+XO1+XO2+XO4+XU1+XU2+XU3+XU4 ### P=15 ### Order=131### Max adjusted R squared

s.model3=lm(XO3~XC1+XC2+XC4+XF1+XF2+XF3+XF4+XO1+XO2+XO4+XU1+XU2+XU3+XU4,data=my.data)

p$Cp[131]

r$adjr2[131]

##### Model Selection#####

#### Problem 4 #### same year prediction ####

##### read data #####

my.data=read.csv("data4.csv",header=T)

names(my.data)=c("year","XC1","XC2","XC3","XC4","XF1","XF2","XF3","XF4","XO1","XO2","XO3","XO4","XU1","XU2","XU3","XU4","Y")

##### full model #####

fmodel=lm(Y~XC1+XC2+XC4+XF1+XF2+XF3+XF4+XO1+XO2+XO3+XO4+XU1+XU2+XU3+XO3,data=my.data)

summary(fmodel)

##### AIC #####

library(MASS)

AICp = stepAIC(fmodel, direction="both")

AICp$anova

##### autoselection with both direction #####

step(Base, scope = list( upper=fmodel, lower=~1 ), direction = "both", trace=FALSE)

dropterm( fmodel, test = "F" )

##### Cp #####

library(leaps)

p=leaps(x=my.data[,c(2:3,5:17)], y=my.data[,18],names=names(my.data)[c(2:3,5:17)],

method="Cp")

plot(p$size,p$Cp,xlab="p",ylab="Cp")

min(abs(p$Cp-p$size)[abs(p$Cp-p$size)>0])

p$which[61,]

##### adjusted R2 ######

r=leaps(x=my.data[,c(2:3,5:17)], y=my.data[,18],names=names(my.data)[c(2:3,5:17)],

method="adjr2")

plot(r$size,r$adjr2xlab="p",ylab="adjusted R-square")

max(r$adjr2[1:140])

p$which[61,]

#####selected model#####

### Y~XC1+XC2+XF1+XF2+XU1+XU2+XU3 ### P=8 ### Order=61### Max adjusted R squared, Cp Close to p

s.model=lm(Y~XC1+XC2+XF1+XF2+XU1+XU2+XU3,data=my.data)

p$Cp[61]

r$adjr2[61]

####model delection#####

##### read data #####

my.data=read.csv("data4.csv",header=T)

names(my.data)=c("year","XC1","XC2","XC3","XC4","XF1","XF2","XF3","XF4","XO1","XO2","XO3","XO4","XU1","XU2","XU3","XU4","Y")

library(MASS)

##### full model #####

fmodel=lm(Y~XC1+XC2+XC4+XF1+XF2+XF3+XF4+XO1+XO2+XO3+XO4+XU1+XU2+XU3+XU4,data=my.data)

#summary(fmodel)

extractAIC (fmodel)

##### $ reduced model 0: through origin #####

rmodel0=lm(Y~0+XC1+XC2+XC3+XC4+XF1+XF2+XF3+XF4+XO1+XO2+XO3+XO4+XU1+XU2+XU3+XU4,data=my.data)

#summary(rmodel0)

extractAIC(rmodel0)

##### $ reduced model 1: within same state #####

rmodel1=lm(Y~XO1+XO2+XO3+XO4,data=my.data)

#summary(rmodel1)

extractAIC(rmodel1)

##### reduced model 2: within same type #####

rmodel2=lm(Y~XC3+XF3+XU3+XO3,data=my.data)

#summary(rmodel2)

extractAIC(rmodel2)

##### reduced model 3: US data #####

rmodel3=lm(Y~XU1+XU2+XU3+XU4,data=my.data)

#summary(rmodel3)

extractAIC(rmodel3)

##### reduced model 4: US data same type #####

rmodel4=lm(Y~XU3,data=my.data)

#summary(rmodel4)

extractAIC(rmodel4)

##### model 5:time series #####

rmodel5=lm(Y~year,data=my.data)

#summary(rmodel5)

extractAIC(rmodel5)

### p value min -> max ###

#XC1, XC2, XU2, XU4, XU1, XF4, XO3, XC4, XF1, XO2, XO3, XO4, XF2, XF3, XU3

##### reduced model 6: significant variables defined from the full model result (3) #####

rmodel6=lm(Y~XC1+XC2+XU2,data=my.data)

#summary(rmodel6)

extractAIC(rmodel6)

##### reduced model 7: significant variables defined from the full model result (7) #####

rmodel7=lm(Y~XC1+XC2+XU2+XU4+XU1+XF4+XO3,data=my.data)

#summary(rmodel7)

extractAIC(rmodel7)

##### reduced model 8: significant variables defined from the full model result (10) #####

rmodel8=lm(Y~XC1+XC2+XU2+XU4+XU1+XF4+XO3+XC4+XF1+XO2,data=my.data)

#summary(rmodel8)

extractAIC(rmodel8)

##### reduced model 9: picked from the result of Cp and adjusted R2 #####

rmodel9=lm(Y~XC1+XC2+XF1+XF2+XU1+XU2+XU3,data=my.data)

#summary(rmodel9)

extractAIC(rmodel9)

##### reduced model 10: only affected by the previous year #####

rmodel10=lm(Y~XO3, data=my.data)

#summary(rmodel10)

extractAIC(rmodel10)

##### reduced model 11: modified based on collinearity (from full model) #####

rmodel11=lm(Y~XC1+XF1+XF2+XF3+XO1+XO2+XO3+XU1+XU3,data=my.data)

#summary(rmodel11)

extractAIC (rmodel11)

##### reduced model 12: modified based on collinearity (from full model) (from reduced model 9) #####

rmodel12=lm(Y~XC1+XF1+XF2+XU1+XU3,data=my.data)

#summary(rmodel12)

extractAIC(rmodel12)

##### evaluation using 2010 data to test 2011#####

data.pred=read.csv('test4.csv',header=F)

names(data.pred)=c("year","XC1","XC2","XC3","XC4","XF1","XF2","XF3","XF4","XO1","XO2","XO3","XO4","XU1","XU2","XU3","XU4","Y")

##### full model #####

predict(fmodel, newdata=data.pred,interval='prediction')

##### reduced model : through origin #####

predict(rmodel0, newdata=data.pred,interval='prediction')

##### reduced model 1: within same state #####

predict(rmodel1, newdata=data.pred,interval='prediction')

##### reduced model 2: within same type #####

predict(rmodel2, newdata=data.pred,interval='prediction')

##### reduced model 3: US data #####

predict(rmodel3, newdata=data.pred,interval='prediction')

##### reduced model 4: US data same type #####

predict(rmodel4, newdata=data.pred,interval='prediction')

##### model 5:time series #####

predict(rmodel5, newdata=data.pred,interval='prediction')

##### reduced model 6: significant variables defined from the full model result #####

predict(rmodel6, newdata=data.pred,interval='prediction')

##### reduced model 7: significant variables defined from the full model result #####

predict(rmodel7, newdata=data.pred,interval='prediction')

##### reduced model 8: significant variables defined from the full model result #####

predict(rmodel8, newdata=data.pred,interval='prediction')

##### reduced model 9: picked from the result of Cp #####

predict(rmodel9, newdata=data.pred,interval='prediction')

##### reduced model 10

predict(rmodel10, newdata=data.pred,interval='prediction')

##### reduced model 11

predict(rmodel11, newdata=data.pred,interval='prediction')

##### reduced model 12

predict(rmodel12, newdata=data.pred,interval='prediction')

my.data=read.csv("dataAR.csv",header=T)

names(my.data)=c("XO3")

my.data=data.frame(my.data)

attach(my.data)

ts(data=my.data,start=1994, end=2010, frequency=1, deltat=1)

ts.data=ts(data=my.data,start=1994, end=2010, frequency=1, deltat=1)

ts.plot(XO3)

###fit a regressional model to remove the linear trend###

Y=ts.data

X=time(ts.data)

trend=lm(Y~X,data=ts.data)

resid.data=ts.data-predict(trend, newdata=X)

#Construct a sample autocorrelation and partial autocorrelation plots

par(mfrow=c(2,1))

acf(resid.data)

pacf(resid.data)

#############################################

## Fit AR models

#Fit a AR(p) model

ar.data <- ar(resid.data)

ar.data

#examine the residuals

par(mfrow=c(1,1))

ts.plot(ar.data$resid)

par(mfrow=c(2,1))

acf(ar.data$resid,na.action=na.omit)

pacf(ar.data$resid,na.action=na.omit)

#############################################

## Prediction

#Subset time series

data.obs=predict(trend, newdata=X)+window(resid.data)

#predict the XO3 from 2011 to 2015

N.pred=5

time.pred=read.table("pre.txt",header=F)

names(time.pred)=c("X")

resid.data.pred=(predict(ar.data, n.ahead=N.pred))$pred

data.pred= resid.data.pred+ predict(trend, newdata=time.pred)

#plot the entire observed series and the predictions

par(mfrow=c(1,1))

ts.plot(data.obs, data.pred,col=c("black", "red"), lty=c(1,2))

abline(v=2010.5, lty=2)