The Application of Regression Diagnosis in Outlier Detection

Mingming Chen[1, a], Meng Gao[2, b] and Jinglian Ma[1, c]

[1]School of Economics and Management, Chang'an University, Xi'an 710064, China

[2]College of Science, Chang'an University, Xi'an 710064, China

[a]changandx2013@126.com, [b]lelepingan@sina.com,[c]kunming85@126.com

Abstract. As one of the most important tasks in data mining, outlier detection may get unexpected knowledge discovery. Regression diagnosis plays an important role in detecting outliers. This paper mainly introduces the basic theory of residual analysis and impact analysis in regression diagnosis, then makes regression diagnosis analysis on a group of data which related to altitude and species amount, and uses the local weighted scatterplot smoothing method to verify the rationality of the regression model, finally gets some useful instructions of regression diagnosis on the outlier detection.

Keywords: Data mining; Regression diagnosis; outlier detection; Local weighted scatter smoothing.

1. Introduction

Regression diagnosis includes two aspects: residual analysis and influence analysis, the diagnosis of the residual analysis is about the model and the impact analysis is the diagnosis of data [1]. In linear regression, the model error is generally assumed to satisfy the Gauss-Markov hypothesis. However, in practical application, it is the first problem to investigate whether the actual data is satisfied or approximate satisfied with the Gauss-Markov hypothesis, which is the first problem to be studied. That is the rationality of the model hypothesis. Since the Gauss-Markov hypothesis is aimed at the error, and the error is unknown, it is usually used residual to discuss the estimation of error [2]. In this paper, we use the residual chart to analyze the residual error. Once the data meet the Gauss-Markov hypothesis, the next question is whether the model is suitable for all data. In practical application, it is possible that the model is fit for most data except that one or several data are not consistent with the model. At this time, we need to analyze the impact of sample data, that is, to probe the data that have a large impact on the estimates. If a sample point does not comply with a model, the rest of the sample points are followed by the model, then the sample point is strongly affected (also called an outlier) [3]. An important function of the impact analysis is to find out these strong influence data. In this paper, the use of leverage value and Cook statistics [4] to analyze.

The main idea of Local Weighted Scatter Smoothing (LOWESS) is to get a certain percentage of local data, and then to fit the polynomial regression curve, so that the law and trend of the data can be observed in the local show. But the usual regression analysis is usually based on the whole data model to describe the overall trend, while the reality is not always linear, but often non-linear. The local range of the LOWESS method is from left to right, and a continuous curve can be calculated. Obviously, the smoothness of the curve is related to the proportion of the selected data, the smaller the ratio, the more the fitting is not smooth, because of too much attention to the local property; vice versa.

2. The presentation of R software and car package

R software is a free, open source statistical analysis software, many statistical methods can be achieved in the R, and the user can understand the latest information and use of R software by the official website to get the latest version of R software and application statistics software package based on R. Among them, the car package is the official recommendation of a package and its function is to carry out regression diagnosis analysis of the actual data model [5].

The car package is not the default installation of R in the car package, you need to download and install from CARN when using car. It includes many functions, this paper mainly uses the residualPlots () and plot () to draw the residual graph, the following describes the concrete realization of the residualPlots () function.

Usage of function:

residualPlots(model, terms = ~., layout = NULL, ask, main = "", fitted = TRUE, ...)

Main parameters:

Model is linear or generalized regression model;

Terms indicates single side formula for specifying predictor variables;

layout is set layout drawing; ask represents logical variable, if it is TRUE, then ask whether to draw a residual plot, otherwise, do not ask;

Main used to set the main title of a graph;

Fitted is a logical variable, if the TRUE, painting about the residual graph of the fitted value, the default is TRUE.

3. Case Study

In this study, we chose the data of altitude and species numbers to carry out the research on the regression diagnosis, and combined with the local weighted scattered point smoothing method, and gave the application of the regression diagnosis in the detection of the abnormal value of the species number of the species, finally find the outliers. Here the dependent variable is number of species counts and independent variable is elevation altitude, specific implementation is given below.

Step 1: The relationship between the counts and altitude variables is viewed by the local weighted local smoothing method, command in R software is as follows:

```
> plot (data)
> for (i in seq(0.01, 1, length = 100))
{+     lines (lowess (data$altitude, data$counts, f = i), col = gray (i), lwd = 1.5)
+     Sys.sleep(0.15)}
```

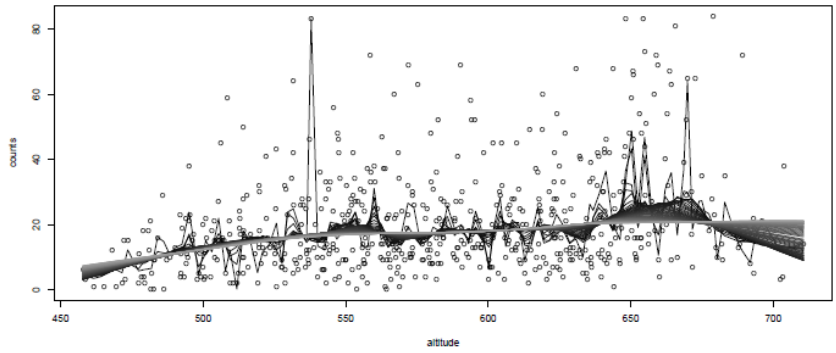

Fig. 1 Locally weighted regression scatter smoothing diagram of species number and altitude height

Table 1 Results of regression analysis

coefficient	estimated value	standard deviation	t value	p value
Intercept	-26.67224	6.45287	-4.133	4.09e-05
altitude	0.08159	0.01107	7.372	5.61e-13
Equation	$\hat{\sigma} = 15.3$	$R^2 = 0.08331$	$F = 54.35$	$P = 5.615e-13$

In Figure 1, the light color of the curve indicates that the proportion of the data is large, it is not difficult to see that the white curve is almost linear, while the black line is larger, that can be used to fit the data in a straight line. It can also be seen from Figure 1, the number of species in the three elevation deviates from the regression line is more serious. Respectively, 550 meters, 650 meters and 700 meters nearby, and the number of species in the vicinity of 650 meters above sea level at the most.

Step two: linear regression analysis, the following commands in the R software are as follows:

> lm.sol=lm(counts~altitude,data)

> summary(lm.sol)

The results are listed in Table 1, we can get the regression coefficient, intercept, standard error, F statistics and T statistics etc. The results show that the P values of the regression coefficient and regression equation are less than 0.05, that is to say the regression coefficient and regression equation are all significant. Regression equation is $\hat{y} = 0.08159x - 26.67224$. However, the correlation coefficient R^2 is small, the number of species and the altitude has a weak linear relationship, that is, the model fitting effect is not ideal. The abline() function is used to draw the fitting line in the scatter plot to see the fitting effect.

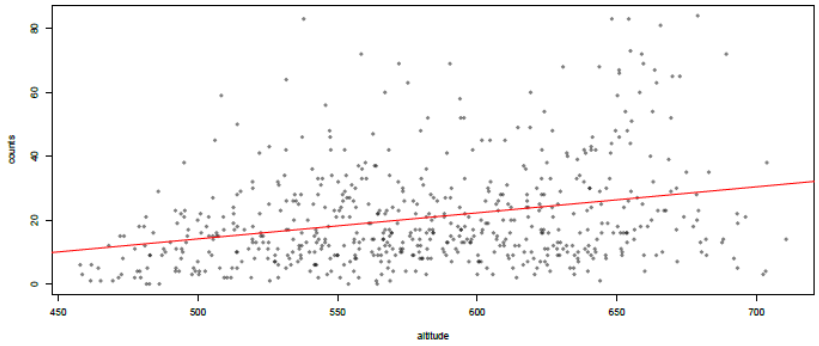

Fig. 2 Fitting straight line of the data

In Figure 2, the sample point is near the regression line, a straight line fits most of the sample points, and the number of species has a linear relationship with the altitude. But some samples are far from fitting regression line, that is, to find out the not matched data samples from the data set to analyze their impact on the model. And the correlation coefficient R^2 is small, which makes the linear model is a problem, so only according to the fitting line chart, there is not enough information to make a judgment on the fitting model. The regression analysis of the data was carried out using the residual analysis and influence analysis, and according to the relevant statistical index mark may be abnormal value of sample points, analysis and study of these points, to refit the data, improve the model, in order to get more accurate and the relationship between altitude and species number.

Step three: residual analysis, draw a residual chart. plot(lm.sol, which=1:4)

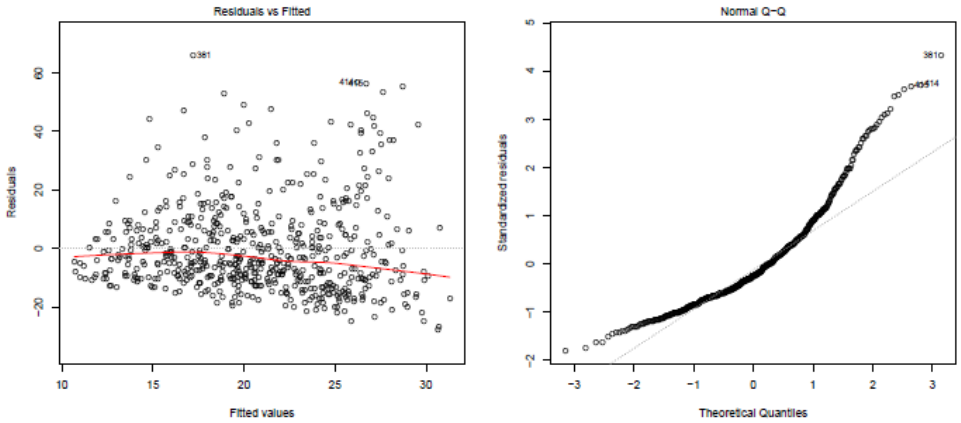

Fig.3 Common residual and residual normal QQ graph

Residual chart is an important tool for model diagnosis, from Figure 3 (left), for the average residual plots, regardless of the size of the fitted value, the residuals have the same distribution and meet the various assumptions of the model for the establishment of the vast majority of sample points. The three sample points are marked out of the absolute residuals:3 81, 414 and 415. QQ graph is used to test the normality, if the general trend of points is clearly not in a straight line in the normal QQ diagram, there is reason to doubt the assumption of normality of the error, otherwise think reasonable. Fig.3 (right) the 3 outliers are marked as the 381, 414 and 415. If the assumption of the normal

distribution of the residuals is established, normalized residuals should be approximated to obey the standard normal distribution. That is, for the standardized residuals, there should be 95% of the sample points in the interval [-2,2]. In addition, fitting values and residuals are independent of each other, the residuals are also independent of each other .In the standard residual plot, the sample points on the plane should be roughly the width of the horizontal band of 4, and not showing any trend. Fig.4 (left) marked 381, 414 and 415 of three abnormal samples; cook statistics is the diagnostic test of the analysis, the greater the value, the more likely it is the abnormal value point, the Cook statistic is the largest of 415th, 520th and 540 th sample points, and they are most likely to be outliers in the Fig.4 (right).

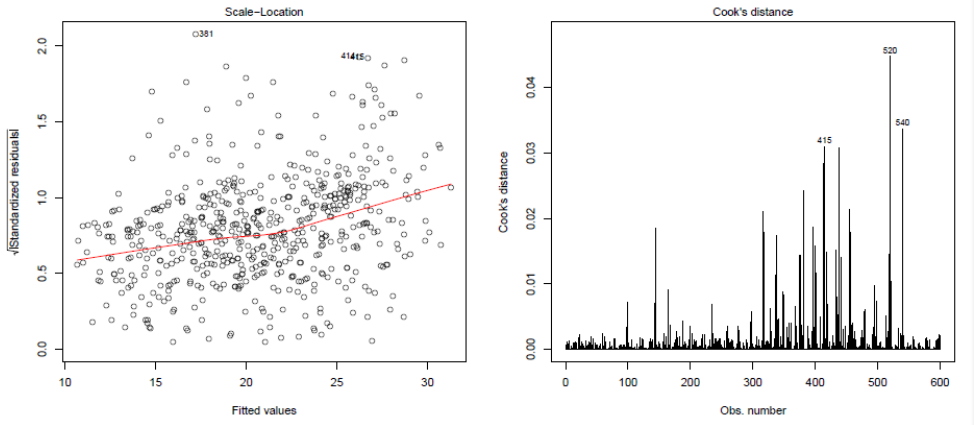

Fig.4 Standard residual and Cook statistics plot

Based on the analysis, 381, 414 and 415 these three sample points are abnormal, that is at 381, 414 and 415 sample points, and the number of species may be abnormal. For the confirmation of the abnormal points, if it is due to a recording error and other similar reasons caused, you can remove it. Otherwise, we should further analyze the abnormal points, appropriately change the model or introduce new variables, in order to establish a more suitable regression model.

4. Summary

Through the establishment of regression diagnosis several related concepts in data mining, the relationship between two dimensional variables is observed by local weighted scatter point smoothing method. Combined with statistical software R, the regression analysis of the actual sample data is carried out, and the abnormal value of the regression model is found. And the detection of outliers can be helpful to the theoretical research and practical application of the relevant issues. And the detection of outliers can be helpful to the theoretical research and practical application of the relevant issues. Therefore, for the processing and application of outliers, it is necessary to further study the work.

References

[1] John Fox. Regression Diagnostics [M]. California: SAGE publication, 2009, p. 42-63.

[2] Cook, R. D. and Sanford Weisberg. Residuals and influence in Regression [M]. New York: Chapman and Hall, 1982, p. 31-37.

[3] Zhang Jige. Abnormal Point and Influence Point in Regression Analysis [J]. Statistical Research, Vol.11 (1994) No.2, p. 43-45.

[4] Cook, R. D. Detection of influential observations in linear regression [J]. Technimetrics, Vol. 19(1977) No.1, p. 15-18.

[5] John Fox and Harvey Sanford Weisberg. An R Companion to Applied Regression [M]. California: SAGE publication, 2010, p. 285-327.