Deep Space Survey of Quasars

作者组号：20
2016100103004 刘梦梦 33%
2016100101003 赵允 33%
2017100101004 高原 33%

Abstract: This essay use statistic methods including multiple linear regressions to figure out the correlation among red shift, line flux, luminosity, AB1450 magnitude and rest frame equivalent width. The first-order model we have received is useful, but not perfect enough for the problem.
Key Words: Multiple linear regressions, F-test, T-test

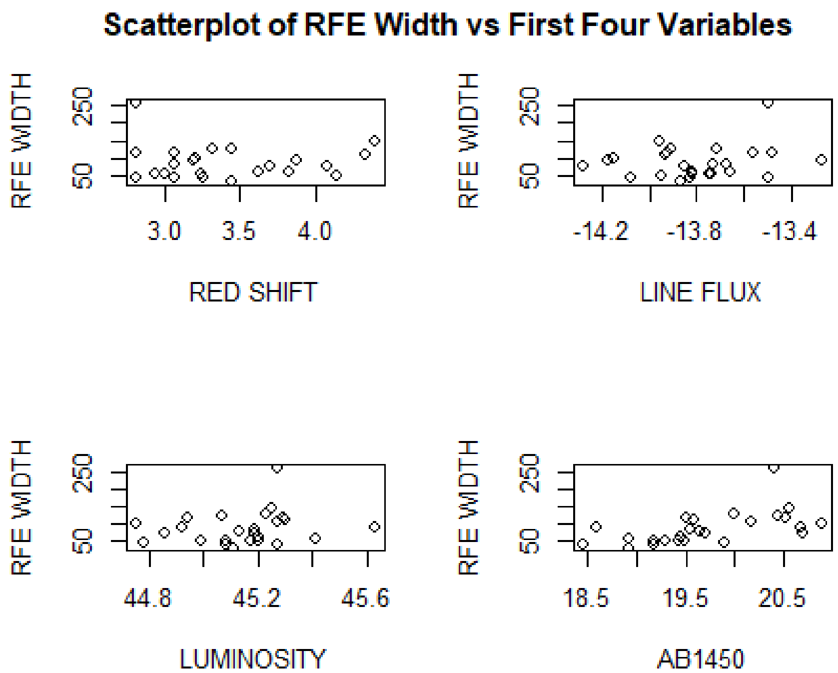
1. Introduction to the Problem

A quasar is a distant celestial object (at least 4 billion light-years away) that provides a powerful source of radio energy. The Astronomical Journal (July 1995) reported on a study of 90 quasars detected by a deep space survey. The survey enabled astronomers to measure several different quantitative characteristics of each quasar, including red shift range, line flux (erg/cm·s), line luminosity (erg/s), AB1450 magnitude, absolute magnitude, and rest frame equivalent width.

Parts of the data are listed below.

QUASAR	RED SHIFT	LINE FLUX	LUMINOSITY	AB1450	ABSMAG	RFEWIDTH
1	2.81	-13.48	45.29	19.50	-26.27	117
2	3.07	-13.73	45.13	19.65	-26.26	82
3	3.45	-13.87	45.11	18.93	-27.17	33

Firstly, we plot the raw data to observe the potential tendency.



Now we are interested in whether there is a correlation among the rest of variables with rest frame equivalent width.

2. First-order Model

2.1 Hypothesis

Firstly it is evident and straightforward for us to assume that the correlation is linear and all the variables are independent. Therefore, we have the following first-order model: considering the equivalent width of the first-order model y is the first four variables (redshift range,line flux (erg/cm·s), line luminosity (erg/s), AB1450magnitude) in the table. Then, we replace the first four variables by the independent variable $x_1, x_2 \ldots x_4$, where $\beta_1, \beta_2 \ldots \beta_4$ are the regression coefficients.

Therefore, for the first-order model y of equivalent width, we obtain the following preliminary model, where ε is a random error:

$$\mathrm{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

There are two assumptions about the random error ε. The first one is for any given set of values of $x_1, x_2 \ldots x_4$, ε has a normal probability distribution with mean equal to 0 $\left[i.e.,\ E(\varepsilon) = 0 \right]$ and variance equal to σ^2 $\left[i.e.,\ Var(\varepsilon) = \sigma^2 \right]$. Plus, the second assumption is that the random errors are independent (in a probabilistic sense).

2.2 The least squares prediction equation

With the method of least squares, we can fit the coefficients $\beta_1, \beta_2 \ldots \beta_4$ with the least square errors:

$$y = 21087.951 + 108.451 x_1 + 557.910 x_2 - 340.166 x_3 + 85.681 x_4$$

There is a must to interpret each β-*estimate* in a practical sense. Otherwise, the model can be invalid. Thus the practical interpretation is presented in the table.

Interpretations	
β_0	No meaningful interpretation when the rest of variables are 0.
β_1	As the value for the red shift range (x_1) increase by 1 unit, the value for the dependent variable equivalent width (y) increase by approximately 109 units.
β_2	As the value for the line flux (x_2) increase by 1 unit, the value for the independent variable equivalent width (y) increase by approximately 558 units.
β_3	As the value for the luminosity (x_3) increase by 1 unit, the value for the

	dependent variable y or equivalent width increase by approximately 340 units.
β_4	As the value for the magnitude (x_4) increase by 1 unit, the value for the dependent variable y or equivalent width increase by approximately 85 units.

Apart from the interpretation of coefficients, the practical sense of the coefficient of determination R^2 . Since the R^2 is 0.912 while the adjusted R^2 is 0.894, it can be concluded that the adjusted R^2 has taken into account the sample size n and the number of the model parameters. With this consideration, the adjusted R^2 is a better option, which helps explain 89.4% of the sample variation of equivalent width by the model.

2.3 Tests on the model

2.3.1 Test the overall utility

As it is shown on the above, we have got the first-order model. Next, we need to test the utility of the model statistically.

Firstly, the hypothesis of F-test is presented below:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_4 = 0$$ (All model terms are unimportant for predicting y)

$$H_a : \exists \beta_i \neq 0 \quad i = 1, ..., 4$$ (At least one model term is useful for predicting y)

Then with the help of R, we have F statistic equals to 51.72, and the $p-value$ is 2.867e-10, which is much smaller than the significance level $\alpha=0.05$. Therefore, we reject the null hypothesis and the overall model appears to be statistically useful for prediction equivalent width.

2.3.2 Test each coefficient

Next it is significant to do t-tests on each coefficient. Since we only have four independent variables, we should deal with all the coefficients.

The hypothesis is:

$$\begin{matrix} H_0\text{:} \ \ \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{matrix} \quad i = 1, ..., 4$$

The results of t-test:

	T value	p-value
Red shift	1.222	0.2359
Line flux	1.766	0.0927
Luminosity	-1.060	0.3016
AB1450	13.658	1.34e-11

Only the coefficient of AB1450 magnitude is below 0.05 significant level, which rejects the null hypothesis. However, we can’t conclude that other coefficients are important enough to explain the variation of y .

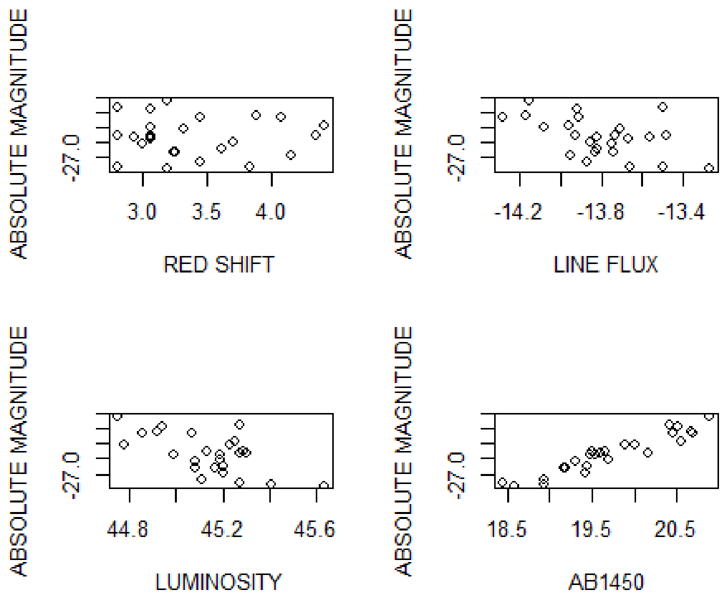
2.4 Prediction and estimation

The last part is to use the model. We consider to calculate the estimation and prediction of equivalent width when red shift is 3.85(), line flux is -13.68, luminosity is 45.42(erg/cm·s), magnitude is 18.98(erg/s).

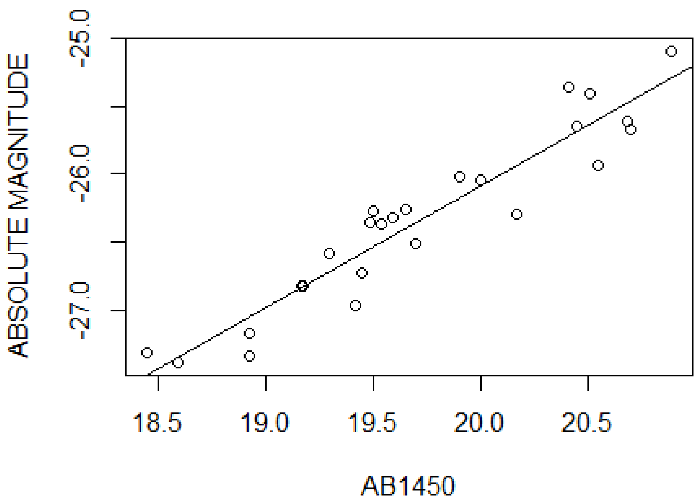
	fit	Lower	Upper
estimation	49.188	26.982	71.395
prediction	49.188	10.109	88.267

3. Other conclusions

We now are interested in the fifth variable absolute magnitude in the table. Firstly, we plot the raw data to see if there is a potential correlation.



It seems that the absolute magnitude has a linear correlation with AB1450 magnitude.



Then the correlation coefficient between AB1450 magnitude and absolute magnitude is 0.947, which indicates a strong linear correlation between the two variables.

Therefore, it is plausible that the first-order model doesn't include the fifth variable.