Guoxin Li

Amelia Helland

STA 138 – Project 2

**INTRODUCTION**

Fisher’s *Iris* data set, or the *Iris* flower data set, is used for linear discriminant analysis, but use of this data set in [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) is uncommon, since the data set only contains two clusters with rather obvious separation. One of the clusters contains *Iris setosa*, while the other cluster contains both *Iris virginica* and *Iris versicolor* and is not separable without the species information Fisher used. This makes the data set a good example to explain the difference between supervised and unsupervised techniques in [data mining](https://en.wikipedia.org/wiki/Data_mining). Since class labels and clusters are not necessarily the same, and Fisher's linear discriminant model can only be obtained when the object species are known, we use logistic regression as a classification method for when the object species is unknown.

**MATERIAL**

*Iris* is a [multivariate](https://en.wikipedia.org/wiki/Multivariate_statistics) [data set](https://en.wikipedia.org/wiki/Data_set) introduced by [Ronald Fisher](https://en.wikipedia.org/wiki/Ronald_Fisher) in his 1936 paper, *The Use of Multiple Measurements in Taxonomic Problems*. Although it was originally presented as an example of [linear discriminant analysis](https://en.wikipedia.org/wiki/Linear_discriminant_analysis), it is also famous in data mining. It contains three different species of the Iris（Iris setosa, Iris versicolor and Iris virginica), each with fifty samples and four characteristics: sepal length, sepal width, petal length, and petal width (all numerical and measured in centimeters). Since the *Iris* flower data is built in R, we used the data directly from R.

**METHODS**

Logistic Regression is a classification method that models the probability of an observation belonging to one of two classes (classification). As such, logistic regression is typically demonstrated with a binary classification problem. Logistic Regression can also be used on problems with more than two classes (multinomial). In this case, *Iris setosa* is easy to separate from the species, so we focus on specifying *Iris virginica* and *Iris versicolor.* This is a binary classification problem, since there are two classes; each species is either *Iris virginica* (y = 1) or *Iris versicolor* (y = 0). Our objective is to determine a probability function which takes in an input Iris (petal length, petal width), returns ‘*the probability of this Iris to be virginica*’ and do the classification.

**RESULTS**

Restricting our attention to the case in which y has two values, 1 and 0, the most direct way to apply multiple logistic regression would be to use the model:

Logit(P(y=1|X1,X2,...,Xn)) = a + b1\*X1+b2\*X2+...+bn\*Xn

In our project, we try to use a simply model, so we didn’t include any interaction between the variables into our model. First, we select the best model. Because we use a simple mode, instead of using AIC or G^2 criteria to select the best model, we use another we to approach it. We separately fit the model with every single variables and plot the model. Figure 1 shows a logistic model that only contains one characteristic variable.

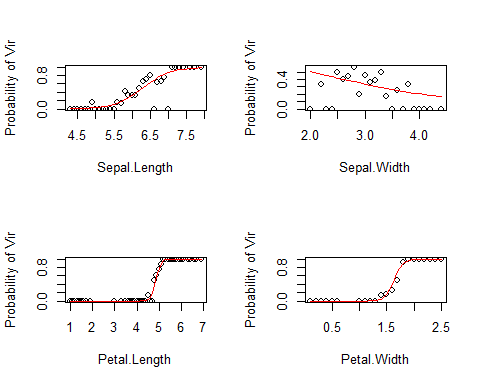


Figure 1 plots of logistic regression

From these pairwise plots, it is immediately apparent that the petals are more relevant than the sepals, since the line fit is much tighter for the petal aspects.

We also got the same result from coefficient P-values:

Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
 (Intercept) -42.638 25.707 -1.659 0.0972 .  
 Petal.Width 18.286 9.743 1.877 0.0605 .  
 Petal.Length 9.429 4.737 1.991 0.0465 \*  
 Sepal.Length -2.465 2.394 -1.030 0.3032   
 Sepal.Width -6.681 4.480 -1.491 0.1359

Table of model coefficients

The respective P-values of the Sepal.Length and Sepal.Width are 0.3032 and 0.1359, so both sepal aspects are not significant at the 0.1 level of significance. So, we only chose the Petal.Width and Petal.Length as our explanatory variables, and we set up our logistic model:

log(p(y=1| X1, X2)/(1-p(y=1| X1, X2))) = a + b1\*X1+b2\*X2

y = 1 is *Iris virginica* , X1 is petal length, X2 is petal width.

After fitting our model and obtaining its estimate model parameters, we find that all coefficients are significant at the 0.05 significance level:

Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
 (Intercept) -45.272 13.610 -3.327 0.000879 \*\*\*  
 Petal.Width 10.447 3.755 2.782 0.005405 \*\*   
Petal.Length 5.755 2.306 2.496 0.012565 \*

Table of model coefficients

As we would expect, the 95% Wald confidence intervals agree with the above conclusions:

B1: (-71.95, -18.596)

B2: (3.0872, 17.8068)

B3: (1.2352, 10.2748)

The 95% Wald confidence interval for the odds ratio for Petal.Width is given by (21.92, 54124635) fixed Petal.Length. This means when the Petal.Width increasing additional centimeter, we are 95% confident the odds of the *Iris virginica* will increase by as little as 219.2% to as much as 5412463500%. The 95% Wald confidence interval for the odds ratio for Petal.Length is given by (3.440, 28992.72) fixed Petal.Width. This means when Petal.Length increasing additional point, we are 95% confident the odds of the *Iris virginica* will increase by as little as 344% to as much as 2899272%.

From the odds results, it raises a problem that the value for odds itself is very large, and the confidence interval for the odds even larger than that. For this problem, our explanation is the data is easily to be classified, later we will show the plot of the data in figure 3. We can easily tell the different species just by looking at the data. From this, we build our logistic model:

Logit(y)= -45.272 + 10.447\*Petal.Width+5.755\*Petal.Length

The two coefficients for Petal.Width and Petal.Length are positive, so they positively relate to the *virginica* type. That is, if we have an iris that has large Petal.Width and Petal.Length values, it indicates that the species is more likely to be *virginica.* Looking at the estimate values, we can see that the effects of Petal.Width are twice as large as those of Petal.Length.

Statistics:  
 X^2 df P(> X^2)  
 Likelihood Ratio 1.273719 51 1  
 Pearson 1.283276 51 1

Table of Pearson and Likelihood Ratio test

The Chi-Square test statistic was 1.283 (df = 51) with a p-value of around to 1, the likelihood ratio test was 1.274 (df = 51) with a p-value of around to 1. Therefore at α = 05. we can not reject the null hypothesis that the model is fitting the data.

**DISCUSSION**

Linear regression can be a fast and powerful tool to model complex phenomena. However, it makes several assumptions about your data, and quickly breaks down when these assumptions (e.g. a linear relationship exists between the predictors and the dependent variable) break down. In this part, we will do some diagnostics and residual analysis to ensure that our regression does not violate basic assumptions. We will also identify outliers and assess how well our model fits the data.

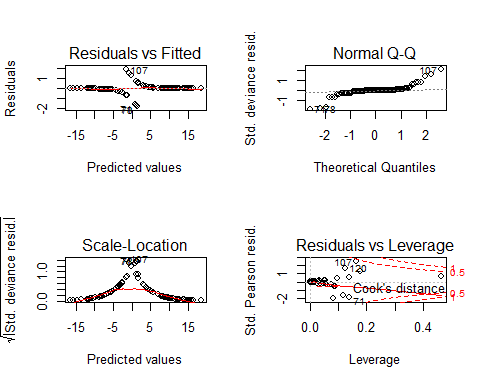


Figure 2 plots of regression diagnostics

The upper left plot shows the residuals versus the fitted values. If there was no scatter, all the points would fall exactly on the red horizontal line, but here, the points are split in two lines. Because Width and Length have a strong positive effect, small width values for an observation of *y* = 1 must have a relatively large positive residual, whereas for large width values for an observation of *y* = 0 will have a relatively large negative residual. When plotted against fitted values, a plot of the raw residuals consists merely of two lines of points which diverge to parallel.

The upper right plot is a Q-Qnorm plot of the residuals. This plot evaluates the least-squares regression assumption that the errors are normally distributed. Here, the points are very off in the two ends of the plot, but most of the points are on the gray line. We consider the errors to follow a normal distribution with a heavy tails.

The lower left plot is a scale-location plot. The regression assumes homoscedasticity, that the variance in the residuals doesn’t change as a function of x. If that assumption is correct, then the red line should be relatively flat, however the red line is not flat here—it is raised up around 0. This shows that the errors are heteroscedastic. A transformation of the data should be tried to see if this problem can be fixed.

The lower right plot shows the standardized residuals against leverage. The standardized residuals are symmetrically centered around zero and reach 2-3 standard deviations away from zero, as would be expected for a normal distribution. Leverage is a measure of how much each data point influences the regression. Here, we can see that R automatically marked the outliers, which are the 71st, 107th and 120th observations.

To continue our analysis, we calculated four different types of residuals: the Pearson residual, standardized Pearson residual, deviance residual, and the standardized deviance residual. In this section we won’t list all of the residuals, because we have 100 observations (see full data in the Appendix), but we will include the values for the 107th observation, since they are unusually large:

Pearson std.Pearson deviance std.deviance

2.245166e+00 2.455708e+00 1.8965923011 2.0744470969

Table of residuals for 107th obervation

From Pearson, std.Pearson, and std.deviance residuals, we conclude that our model doesn’t fit this particular observation very well. In fact, these values confirm the analysis from the Figure 2 plot of standardized residuals against leverage, which led us to believe that the 107th observation is an outlier.

To visualize the classification of the two iris species, we plotted our data with logistic regression in Figure 3. From this, we see that the logistic regression classified the data very well, since very few points are misclassified:

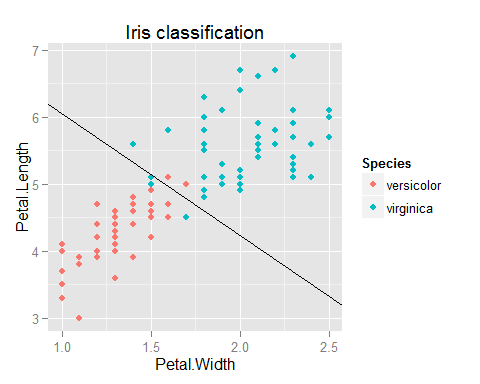


Figure 3 Plot of data classification

The black line in this plot shows where the probability of virginica crosses 0.5. The equation for the line comes from solving for the place when a + b1\*X1 + b2\*X2 = 0.

By calculating a confusion matrix to show how well our model performs, we found that only three points are misclassified in each species:

logit.predictions  
 versicolor virginica  
   
 versicolor 47 3  
 virginica 3 47

Table of confusion matrix

Since the regression plot and the confusion matrix both show that our logistic regression classifies two iris species very well, we are confident that the model is appropriate.

**CONCLUSION AND DISCUSSION**

This paper proposes logistic regression model to classify iris species because this method is good for solving binary classification problems. To keep our analysis simple, we considered a very simple logistic regression model which didn’t include interaction effects and two variables. The effect of Petal.Width and Petal.Length were found to be significant. The predicted values for π (x) increase with Petal.Width and Petal.Length. The model’s fit was analyzed by examining association of predicted and observed probabilities of success, and by using the Chi-square and likelihood goodness of fit test. Based on these results, we can not reject the hypothesis that the model is fitting the data; yet, we would be more comfortable with a model that provides more support of fit. In future work, it might be possible to improve classification accuracy by adding interaction effects and more input parameters in the logistic regression model.