**Regression Analysis**

Regression analysis is one of the first tools used when analyzing your dataset. It involves estimating the relationship between variables, and often it will give you an immediate insight into the organization of your data.

In this lab, we will look at tools available in R for regression analysis:

• Simple regression

• Multiple regression

• Multivariate regression

• Robust regression

# Packages

In R, there are several packages available that provide the programmer with the regression functionality. We will be using the following packages in the examples:

* chemometrics: This package has tools to analyze chemometric data (multivariate)
* MASS: This package offers modern applied statistics with S

# Simple regression

In simple regression, we try to determine whether there is a relationship between two variables. It is assumed that there is a high degree of correlation between the two variables chosen for use in regression.

For this section, we will be using the iris dataset. The iris dataset has observations of the different characteristics of iris plants. For regression, we are seeing if there is a relationship between one characteristic of iris plants and others. As mentioned, the characteristics tested will have a high degree of correlation. The iris dataset is as follows:

> data <- read.csv("http://archive.ics.uci.edu/ml/machine-learningdatabases/iris/iris.data")

Let's also clean up the data so as to be more readable:

>colnames(data) <- c("sepal\_length", "sepal\_width", "petal\_length",

"petal\_width", "species")

Now, let's look at a summary to get an overall picture:

> summary(data)

sepallength sepal\_width petal\_length

Min. :4.300 Min. :2.000 Min. :1.000

1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600

Median :5.800 Median :3.000 Median :4.400

Mean :5.848 Mean :3.051 Mean :3.774

3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100

Max. :7.900 Max. :4.400 Max. :6.900

petal\_width species

Min. :0.100 Iris-setosa :49

1st Qu.:0.300 Iris-versicolor:50

Median :1.300 Iris-virginica :50

Mean :1.205

3rd Qu.:1.800

Max. :2.500

We can look at plots of the data points to try to determine what variables appear to be related:

> plot(data)



What if we were to use petal length to predict petal width? The two plots show a nice linear relationship. The only concern would be that there appears to be two clusters.

Intuitively, there should be a strong relationship between the two. We can check this using R:

>cor(data$petal\_length,data$petal\_width)

[1] 0.9623143

And we see a high correlation between the two. If we go ahead and determine the

regression between the two, we see:

> fit <- lm(data$petal\_length ~ data$petal\_width)

> fit

Call:

lm(formula = data$petal\_length ~ data$petal\_width)

Coefficients:

(Intercept) data$petal\_width

1.093 2.224

We can display the fit information. The fit information displays four charts:

**Residuals vs Fitted**, **Normal Q-Q**, **Scale-Location**, and **Residuals vs Leverage**.

If you remember, the residuals are the difference between the observed data and the

**fitted** or projected values from the model. So, for the **Residuals vs Fitted** plot, we see

some variance.

The Qs stand for quantile, the normalized quantile of the data point versus the

actual. The **Normal Q-Q** graphic is very typical for plots The majority

of the data is in the same quantile, with some degree of variation at the foot and the

head of the plot.

The **Scale-Location** plot shows the square root of the standardized residuals as a

function of the fitted values. You can see columns of data points, as there is not

enough data to cover a wider area. There is a pretty big variety in the fitted values.

Leverage is the importance of a data point in determining the regression result.

The **Residuals vs Leverage** graphic is overlayed with Cook's distance—another measure

of importance of a data point. Overall, we see consistent importance of the data

points at all levels.

Let's display the fit information:

>par(mfrow=c(2,2)) # set the plot area to 2 plots by 2 plots

> plot(fit)



We can use the regression variables in predicting a formula (ordered in the standard

*y = mx + c* format):

petal\_length = petal\_width \* 2.224 + 1.093

We can look at the differences between the observed values and the fitted values

using the residuals() function:

> residuals(fit)

1 2 3 4

-0.138259922 -0.238259922 -0.038259922 -0.138259922

5 6 7 8

-0.283118763 -0.360689343 -0.038259922 -0.138259922

9 10 11 12

0.184169498 -0.038259922 0.061740078 0.084169498

…

There are differences for every data point with valid data (149 data points). A rough

scan doesn't reveal any outliers. However, a summary produces results:

> summary(fit)

Call:

lm(formula = data$petal\_length ~ data$petal\_width)

Residuals:

Min 1Q Median 3Q Max

-1.33171 -0.30741 -0.01956 0.25988 1.39259

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.09340 0.07384 14.81 <2e-16 \*\*\*

data$petal\_width 2.22429 0.05184 42.91 <2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4801 on 147 degrees of freedom

Multiple R-squared: 0.926, Adjusted R-squared: 0.9255

F-statistic: 1841 on 1 and 147 DF, p-value: <2.2e-16

The summary shows us several points about the regression:

* First, it shows what the model is based on (petal length and width).
* It shows the range of residuals. The residuals appear to be in a small range. It is interesting that the upper and lower bounds have the same absolute range.
* The coefficients' (in this case, we are only using one variable, so coefficient) values are presented.

We see an intercept of 1. I think that is reasonable looking back at the data.

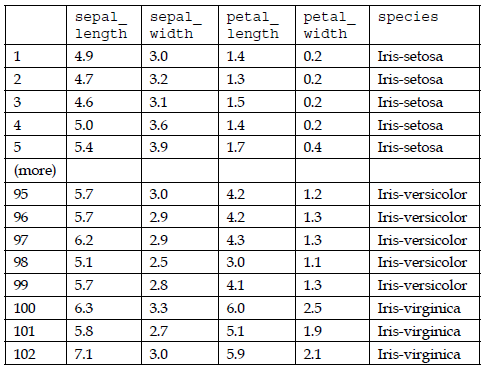
The standard error is pretty low. The probability greater than *t* value is very

low. I think there is confidence in the estimate.

* We see a petal width estimate of 2.2. Again, this looks good as seen in the raw data, with a similar low standard error and very low estimate of the difference.

The residuals vary from -1 to +1, which appears to be a broad range. Here is the

raw data:



We can see that such residual values are too extreme, at least for Iris-setosa. The

other two varieties might show a better fit or at least different regressions. We can

remove the setosa observations from the data with the following command:

> data2<- subset(data, data$species!='Iris-setosa')

We can see how the various plots look:

> plot(data2)



We can see a more clustered relationship between petal length and petal width. The setosa data was definitely responsible for the extra cluster that appeared in the earlier plots. Unfortunately, the data does appear to be more scattered.

Let's run the regression against the subset produced:

>cor(data2$petal\_length,data2$petal\_width)

[1] 0.8233476

> fit <- lm(data2$petal\_length ~ data2$petal\_width)

> summary(fit)

Call:

lm(formula = data2$petal\_length ~ data2$petal\_width)

Residuals:

Min 1Q Median 3Q Max

-0.9842 -0.3043 -0.1043 0.2407 1.2755

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.2240 0.1926 11.55 <2e-16 \*\*\*

data2$petal\_width 1.6003 0.1114 14.36 <2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4709 on 98 degrees of freedom

Multiple R-squared: 0.6779, Adjusted R-squared: 0.6746

F-statistic: 206.3 on 1 and 98 DF, p-value: <2.2e-16

The correlation dropped from 0.96 to 0.82—this is not good.

While the residual standard error has not changed, we have reduced our degrees

of freedom from 147 to 98. Reducing the standard error on the regression is a good

thing—it means we are closer to the observed data points with our modeled data.

However, the R-squared dropped significantly from 0.926 to 0.6779.

The f-statistic dropped from 1800 to 200 with the p-value unchanged with a good,

small value. Overall, I don't think we can exclude setosa from the evaluation.

I think it is important to try to test different subsets of your data to make sure they

are all truly in agreement.