REVOLUTION ANALYTICS

Modeling in Revolution R Enterprise
Module 2: Linear Regression

July 25, 2014

Regression

Let's start this course off by talking about regression. Basically, regression establishes a prediction model where one variable to computed based on the values of the other variables.

For instance, perhaps we want to model the age of working class Americans. Hypothetically, we can suppose that a higher amount of education and experience would most likely take time to receive, and therefore people with greater experience or education should have a older age. Assuming these conditions, we would conclude that both education and experience are positively correlated to age, meaning that an increase in either of the two variables would produce a resultant increase in age.

Regression

Keep in mind, however, that the models always contain some degree of variance and bias, and consequently no model is perfect. Nevertheless, we hope that the models we construct provide us with useful information, and in this manner predictive modeling forms an ever more important component of the modern world.

Simple Regression

The most basic form of regression is Simple Linear Regression, defined as the least squares estimator of a linear regression model using only a single predictor, or explanatory, variable.

Mathematically, suppose that we have some arbitrary number of data points, say n, defined as the set $\{y_i, x_i\}$, where $i = 1, 2, \ldots, n$. Then

$$y = \alpha + \beta x$$

represents a line where α is the intercept of the y-axis and β is the slope of the line. Minimizing the sum of squared residuals, this best-fit line will attempt to explain our response variable y by using the predictor variable x as best as it can!

Residuals

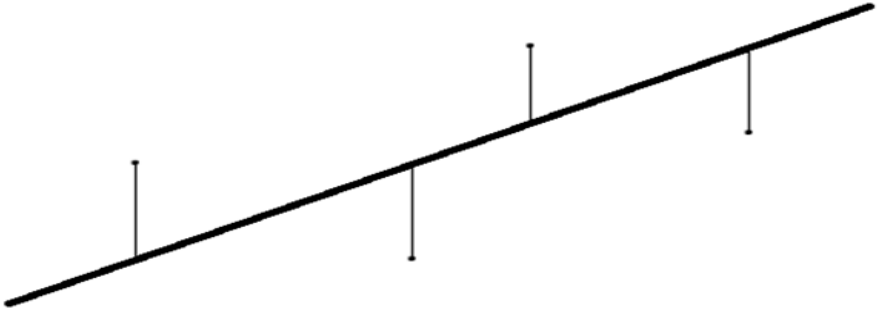
Residuals are closely related to statistical errors - they measure the difference between the estimated “model” value of a point, and its actual observed value in the sample.

- For example, suppose that we predicted the weather to be 70 degrees yesterday, but in actuality it was measured as 72.3 degrees. Then our residual would be defined as 2.3 degrees, the deviation between the estimated value and the observed value.

Least Squares

What exactly is the “Least Squares” approach?

- Least squares minimizes the sum of squared residuals of the linear regression model. The residuals in the diagram below represent the line between the estimated model (the linear model) and the observed point.



rxLinMod

The rxLinMod function fits a linear model to data. Basically, you need to define a regression formula of the dependent variable (Y) against so many independent predictor variables (X_i, for i predictor variables), and also define the data set that these values come from:

```
rxLinMod(formula, data, ...)
```

- Also, note that we can subset and transform our data set in virtually any analysis function in RevoScaleR, including rxLinMod. Use the transforms sub-functional call, and define the list modifying the variables.

Example: Simple Regression

Let's try to estimate the account balance of a customer by his or her age from the Bank data source. In this manner, we are performing simple linear regression, using only one variable to predict an outcome of another. We can do this by selecting balance against age.

```
infile <- file.path("data", "BankXDF.xdf")
BankDS <- RxXdfData(file = infile)
linMod <- rxLinMod(balance ~ age, data = BankDS)
```

```
## Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds
## Rows Read: 10000, Total Rows Processed: 20000, Total Chunk Time: 0.002 seconds
## Rows Read: 10000, Total Rows Processed: 30000, Total Chunk Time: 0.001 seconds
## Rows Read: 10000, Total Rows Processed: 40000, Total Chunk Time: 0.002 seconds
## Rows Read: 5211, Total Rows Processed: 45211, Total Chunk Time: Less than .001 seconds
## Computation time: 0.016 seconds.
```

Example: Simple Regression

```
summary(linMod)
```

```
## Call:
## rxLinMod(formula = balance ~ age, data = BankDS)
##
## Linear Regression Results for: balance ~ age
## Data: BankDS (RxXdfData Data Source)
## File name: data/BankXDF.xdf
...
```

From the above half of the function call, we can see the number of valid observations in the data set, consistent with the output of the rxGetInfo function. Further, we notice the dependent variable balance, based on the single variable age. We obtain the coefficient estimate of age along with an intercept value.

Example: Simple Regression

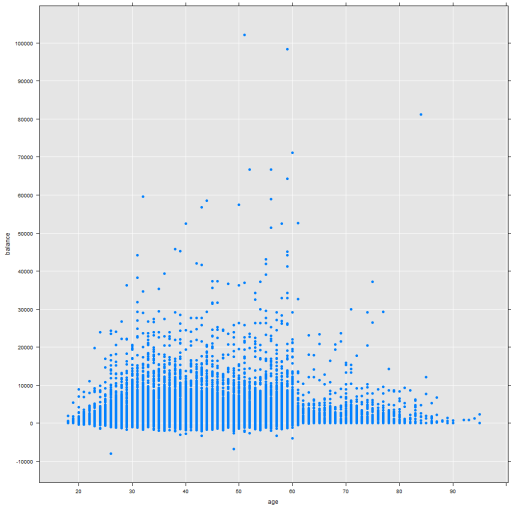
Using the coefficient estimate from above, we can mathematically construct the model, as shown below:

$$\textit{balance} = 214.52 + 28.04 \cdot \textit{age}$$

Hence, an increase of one year in age should, according to our model, produce an increase in balance equivalent to 28.04 euros!

- How accurate is this model, or more specifically, how accurate are our predictions made from using this model?

Example: Model Evaluation



Example: Model Evaluation

According to our previous plot, we can see that as age increase, for a time balance increases as well. In fact, we can confidently state that, on average, an increase in age produces an increase in balance.

- But clearly, this trend is not universal. In fact, around age 60, this relationship seems to reverse itself, and as one becomes older balance appears to decrease.

Example: Model Evaluation

From the previous analysis, we can make some inferences on our model:

- Our model represents the average trend between balance and age. That is, on average as age increases, we should expect balance to increase as well.
- A linear model does not appear to best represent the data trend. Indeed, a line is a poor representation of data that appear to follow a different type of curve.
- Our model has an R-squared value of percent, which indicates that our model represents only about 1 percent of the observed data. We will study this statistic in greater detail in a later section.

Accordingly, our model, while providing a decent average trend, is weak. We should be weary, especially at later ages, making accurate predictions with this model.

Exercise: Simple Regression

Let's try to estimate a customer's balance based on the duration of the advertising phone call. Expectedly, we can probably guess that this relationship will be weak at best; nevertheless, let's try it and perform basic model evaluation, as we did above, on the results. Openly talk about your confidence in making predictions based on your model.

Exercise: Solution

Constructing the simple linear model should be relatively straightforward. Simply define the customer balance, balance, against the duration of the advertising phone call, or duration. Optionally, we may also define the number of blocks per computation to be computed, to increase accuracy and decrease computation time, or vice versa:

```
linMod <- rxLinMod(balance ~ duration, data = BankDS)
```

Exercise: Solution

```
summary(linMod)
```

The above output gives us the dependent variable, balance, and the independent variable, duration, along with the number of valid observations. Again from the above, we obtain the coefficient estimate of duration along with an intercept value.

Exercise: Solution

Using the coefficient estimate from above, we can mathematically construct the model, as shown below:

$$\textit{balance} = 1296 + 0.2549 \cdot \textit{duration}$$

Hence, an increase of one second in an advertising phone call's duration should, according to our model, produce an increase in balance equivalent to 0.2549 euros!

- Again, how accurate is this model, or more specifically, how accurate are our predictions made from using this model?

Exercise: Solution

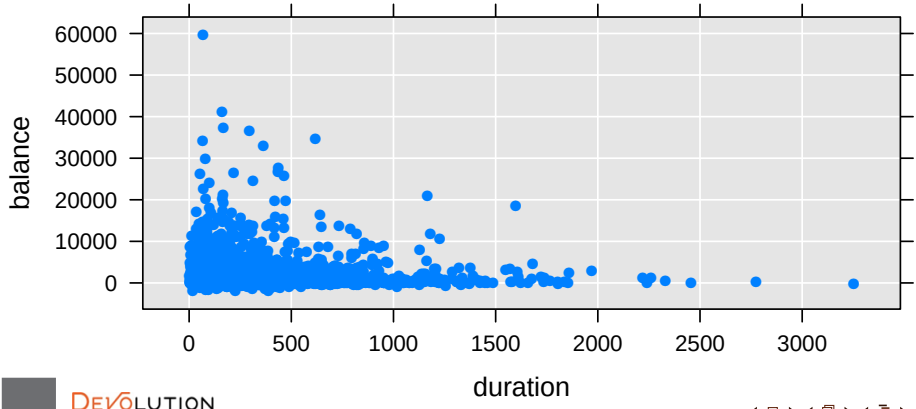
Let's plot a scatterplot as before, but due to the large number of data in our figure we will only plot a randomly selected subset of our sample space. In this case, we will define the subset as ten-percent of the original data set, using the rowSelection parameter in the rxDataStep function:

Exercise: Solution

```
df <- rxDataStep(inData = BankDS, rowSelection = as.logical(rbinom(.rxNumRows,
    1, 0.1)) == TRUE)
rxLinePlot(balance ~ duration, type = "p", data = df)
```

Exercise: Solution

```
## Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.025 seconds
## Rows Read: 10000, Total Rows Processed: 20000, Total Chunk Time: 0.020 seconds
## Rows Read: 10000, Total Rows Processed: 30000, Total Chunk Time: 0.027 seconds
## Rows Read: 10000, Total Rows Processed: 40000, Total Chunk Time: 0.023 seconds
## Rows Read: 5211, Total Rows Processed: 45211, Total Chunk Time: 0.019 seconds
```



Exercise: Solution

According to our model, we can see that as the duration of the advertising phone call increases, balance on average increases slightly as well. However, our plot once again does not follow a linear trend - instead, it appears as though the majority of phone calls are lower in the extremes and closer to the minimal values, just as are the distribution of balances.

- There are some outliers, nonetheless, where some very short phone calls corresponded to high balances, and alternatively long call durations resulted in low balances.

Exercise: Solution

From the previous analysis, we can make some inferences on our model:

- Our model represents the average trend between balance and duration. That is, on average as the call duration increases, we should expect a slight increase in balance as well.
- A linear model does not appear to best represent the data trend. Indeed, a line is a poor representation of data that appear to follow a different type of curve.
- Our model has a very small R-squared value, which indicates that our model represents an extremely small section of observable data. We will study this statistic in greater detail in a later section.

Accordingly, our model, while providing a decent average trend, is weak. We should be wary, especially at later ages, making accurate predictions with this model.

Multivariate Regression

But perhaps more than one predictor variable better explains the response! In this case we should use a similar technique, Multivariate Regression.

Suppose again that we have n data points, but this time they are classified as the set $\{Y_i, X_{i1}, X_{i2}, \ldots, X_{ip}\}$ so we have $p + 1$ terms (don't forget the explanatory variable Y!). Then multivariate regression defines the response or explanatory variable as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

where ϵ_i is defined as the error term for any i^{th} observation.

Multivariate Regression

Again, we use the rxLinMod function; however, we can add additional predictor variables to the formula portion of our function call:

```
rxLinMod(y ~ x1 + x2 + x3 + ..., data, ...)
```

Let's explore this capability in an example!

Example: Multivariate Regression

This time, let's model the customer bank balance against multiple variables: the age of the customer, the duration of time the customer has maintained the account, whether or not the customer is married, and whether or not the customer owns a house.

```
multiLinMod <- rxLinMod(balance ~ age + duration + marital + housing, data = BankDS)
```

```
## Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.002 seconds
## Rows Read: 10000, Total Rows Processed: 20000, Total Chunk Time: 0.005 seconds
## Rows Read: 10000, Total Rows Processed: 30000, Total Chunk Time: 0.003 seconds
## Rows Read: 10000, Total Rows Processed: 40000, Total Chunk Time: 0.003 seconds
## Rows Read: 5211, Total Rows Processed: 45211, Total Chunk Time: 0.003 seconds
## Computation time: 0.027 seconds.
```

Example: Multivariate Regression

```
summary(multiLinMod)
```

Again, here is the output specifying what variables are being used to model balance, as well as the number of observations.

Example: Multivariate Regression

- Some of the variables added to the above model are positively correlated to balance - that is, an increase in that variable would produce a corresponding increase in the balance variable. These variables are indicated by their positive coefficients.
- Further, variables with a negative coefficient are negatively correlated to balance. An increase in that variable would produce a corresponding decrease in the balance variable.

Regression Terms

Up to this point we have largely ignored the other variables associated with the output of our regression models. Since we now have a basic understanding of linear regression, let's address a few of these topics.

- Dropped variables
  - In our above multivariate regression model, some of the terms that were added in the original model have not been included in the final model.
  - These terms were determined to not have a significant impact on our regression model. That is, a reasonable change in one of the above variables does not produce a corresponding change in the independent variable, or balance.

Regression Terms

How do we know which variables are significant in our model and which variables are not?

- To determine significant variables, we have to talk about the p-value. Basically, for each variable added in the above model, we perform a test evaluating whether the model would be any different leaving that variable out (something called the Null Hypothesis). The p-value is the result of this test, and it tells us the probability that the observed test statistic (the variable) is at least as extreme as the model not including the test statistic (again the Null Hypothesis).

Regression Terms

So what does this imply about our model?

Well, let's consider the variable age. The p-value associated with age is equal to 2.22e-16, an extremely small number. This can be interpreted as the probability of observing the impact of this variable occurring in a model not including this variable to determine balance. In a sense, the probability of observing a model as extreme as the one including age, in the modeling holding age out, would be near impossible! We can comfortably say, then, that we are confident in the significance of including the age term in our linear model.

Coefficient of Determination

Another important term to consider in modeling is the Coefficient of Determination, otherwise referred to as R-squared. This value tells us the how well our model fits with the observed data. Let's start off with an example:

The following plot shows two linear models each graphed next to a number of observed data points.

- In the first model, residuals are relatively small, indicating that our model fits our observation points relatively well.
- In the second model, residuals grow increasingly large as we move towards the right, indicating that our model does not fit our observation points very well.

R-squared

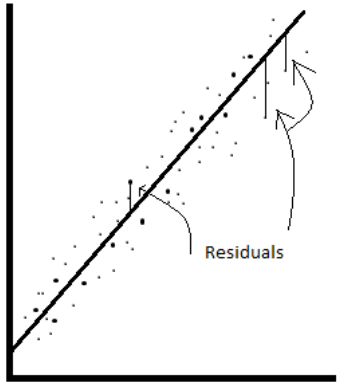


Figure 1

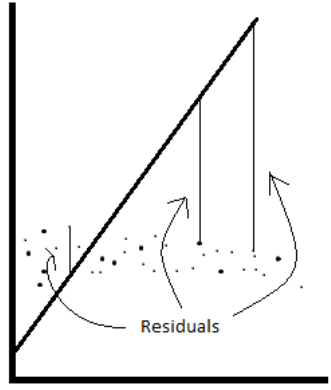


Figure 2

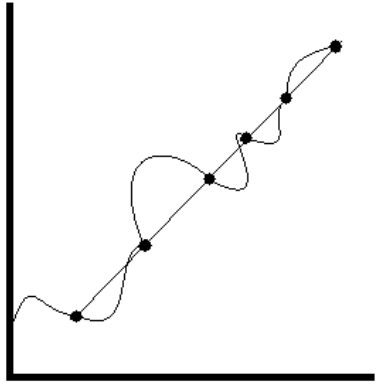
R-squared

Mathematically, computing R-squared involves the consideration of these residual values, but we can say conceptually that the first model in the previous plot fits its observed data much better than the second model fits its data. Consequently, the first model would have a much higher R-squared value in comparison to the second model.

Don't be fooled though! Just because a model has a high R-squared value, or in other words fits its observed data points very well, does not necessarily imply that predictions made from this model will be particularly accurate. For example, consider the plot on the following slide:

R-squared

Perfect R-squared, but
bad prediction



R-squared

- In the previous plot, the observed data points clearly follow a linear trend. However, the crazy function we have used to model the observed data points interpolates them perfectly. Consequently, the R-squared value for our crazy function would equal one - a perfect fit. However, making predictions based on this function would be inaccurate, since a new observation would fall on the straight line.

Complex Formulas and Higher Order Terms

Perhaps the data follow a non-linear trend. In some cases, it is better to model higher order variables in regression. However, such modeling can be misleading, since evaluating your model on another data set gathered from the same population will most likely lead to entirely different conclusions! For this reason, as in the previous example for R-squared, it is often better to limit higher ordered terms.

In notation, our model would look something like this:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots$$

where X^2 represents the second power of the predictor variable X. This trend may be continued; however, be aware of over-fitting your data.

Higher Order Terms

In R, to indicate a higher order term we must use the I() notation, or the conversion of objects. In this fashion, we can indicate higher order terms by raising the variable to the desired exponent, as shown below:

```
rxLinMod(y ~ x + I(x^2) + I(x^3) + ..., data)
```

Example: Higher Order Terms

Let's go back to our balance prediction based on age. Remember how the plot curved in a non-linear fashion? In this case, let's refit our original model to include a higher degree term for age - in fact, a second degree term.

```
highMod <- rxLinMod(balance ~ age + I(age^2), data = BankDS)
```

```
## Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds
## Rows Read: 10000, Total Rows Processed: 20000, Total Chunk Time: 0.004 seconds
## Rows Read: 10000, Total Rows Processed: 30000, Total Chunk Time: 0.003 seconds
## Rows Read: 10000, Total Rows Processed: 40000, Total Chunk Time: 0.003 seconds
## Rows Read: 5211, Total Rows Processed: 45211, Total Chunk Time: 0.002 seconds
## Computation time: 0.026 seconds.
```

Example: Higher Order Terms

```
summary(highMod)
```

```
## Call:
## rxLinMod(formula = balance ~ age + I(age^2), data = BankDS)
##
## Linear Regression Results for: balance ~ age + I(age^2)
## Data: BankDS (RxXdfData Data Source)
## File name: data/BankXDF.xdf
...
```

Example: Higher Order Terms

While we are only modeling balance with one variable, we are modeling that variable with respect to a second-degree curve in addition to its linear term. As a result, we have two separate terms: a linear age term and a second-degree age term.

Notice that both the linear and the second-degree terms for age are given as significant. Furthermore, the second degree term is even determined to be more significant than the linear term - as would be expected given the non-linear trend in data from the previous plot.

Example: Higher Order Terms

From the above coefficients, we can model the equation mathematically below:

$$\textit{balance} = 1295.71 - 23.76 \cdot \textit{age} + 0.58 \cdot \textit{age}^2$$

We can interpret the second-order term by considering an increase in age - an increase in one year, from age to (age + 1), causes balance to change by:

$$\begin{aligned} 0.58 \cdot (\textit{age} + 1)^2 - 0.58 \cdot \textit{age}^2 &= 0.58 \cdot (\textit{age}^2 + 2 \cdot \textit{age} + 1) - 0.58 \cdot \textit{age}^2 \\ &= 2 \cdot 0.58 \cdot \textit{age} + 0.58 \end{aligned}$$

Example: Higher Order Terms

Hence, just considering the second-order term, an increase of one year in age produces a corresponding increase of $1.16 \cdot \textit{age} + 0.58$ in balance - which is a non-linear term! That is, the increase in balance is determined based on the value of age. If we are talking about the difference in balance between ages 25 and 26, then, solely based on the second-order term, balance would increase by $1.16 \cdot 25 + 0.58 = 29.58$ euros.

Example: Higher Order Terms

Another thing to notice is that our age term now has a negative coefficient, which is different from our previous model. In this way, the linear and second order term work opposite of each other and produce a more accurate model of our observed points.

Example: Higher Order Terms

Finally, how much better is our new model compared to before? Well, our R-squared value went up by about half a percent, meaning that our model now accounts for about an additional half a percent of observed data. A slight improvement, but still a far cry from a very accurate model.

- Can you think of any additional improvements we can make to our model?

Recap

Let's review some of the concepts covered in this module:

- What is simple linear regression?
  - Why is it the most basic form of linear regression?
- How does multivariate regression differ from simple linear regression?
- What is p-value?
- What does R-squared refer to?
- How might higher order terms overfit data?

REVOLUTION ANALYTICS

Thank you

Revolution Analytics is the leading commercial provider of software and support for the popular open source R statistics language.

www.revolutionanalytics.com, 1.855.GET.REVO, Twitter: @RevolutionR