REVOLUTION ANALYTICS

Modeling in R Module 6: Model Selection and Scoring

Model Selection and Scoring

```
infile <- file.path("data", "bank-full.csv")
BankDS <- read.table(infile, sep = ";", header = TRUE)
```

Balance versus Age

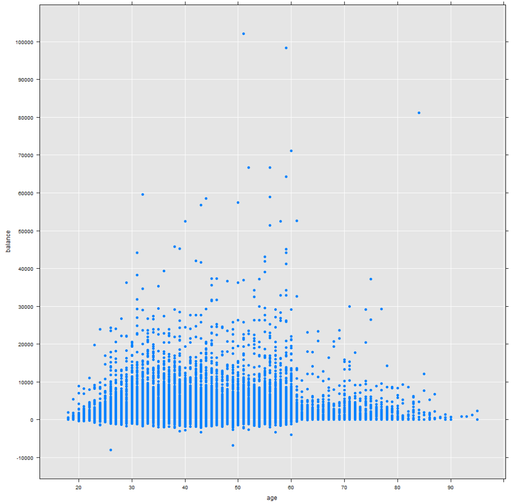


Figure:

Transformation

We can transform variables so that increments in value are proportional by taking a logarithmic transformation

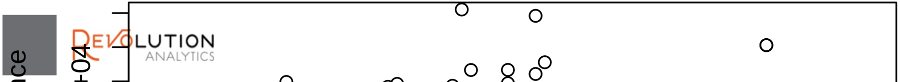
```
BankDS$logBal <- with(BankDS, log(balance))
```

```
## Warning: NaNs produced
```

```
BankSubDS <- BankDS[!is.na(BankDS$logBal) & BankDS$balance > 0, ]
```

or you can do it on the fly:

```
attach(BankSubDS)
plot(age, balance, type = "p")
```



Transfomations

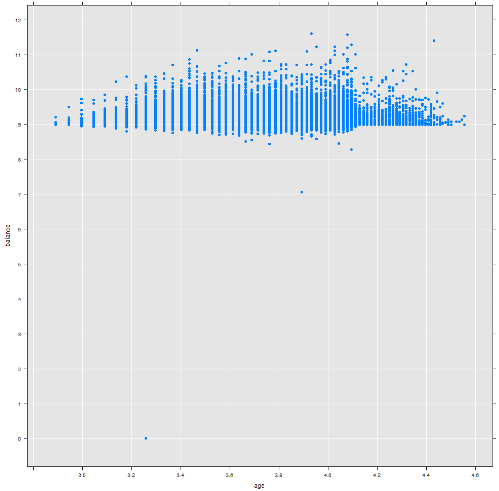


Figure:

Outliers

We can clearly see two outliers in the data after the variable transformation: the two balance values which are distinctly separated from the rest of the data in the set. Let’s remove those two values using the rxDataStep function.

We can accomplish the removal using a criterion evaluation for balance:

```
logBalNoOut <- logBal[logBal < 9]
```

We can also manipulate variables on the fly within equations by wrapping it in I():

```
linMod <- lm(logBal ~ age + I(age^2))
```

Example: Residual Standard Deviations

From the above, we can find the residuals' variance and standard deviation:

```
var(linMod$resid)
```

```
## [1] 2.849
```

```
sd(linMod$resid)
```

```
## [1] 1.688
```

## Exercise: Prediction Standard Errors

Produce a plot of the fitted values and real values of balance, using a polynomial regression of balance on phone call duration.

Recap

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

- What is a residual?
- Discuss: What ways can visual exploration enhance model dianostics

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