Regression Assumptions, Diagnostics, and Fraud Detection

Regression Assumptions, Diagnostics, and Fraud Detection, Part I

There are several assumptions of linear regression.

1. The relationships are linear
2. The X variables (explanatory variables) are not correlated
3. Distribution of residuals
    a. The error terms have constant variance.
    b. The errors terms are not correlated.
    c. There are no outliers.

Regression Assumptions,
Diagnostics, and Fraud Detection, Part II

Violations	**Test**
• Nonlinearity	• RESET
• Multicollinearity	• Variance inflation factor (VIF)
• Heteroscedasticity	• Breusch-Pagan
• Serial correlation	• Durbin-Watson
• Outliers	• Bonferroni

Regression Assumptions, Diagnostics, and Fraud Detection, Part III

Violations	**Correction**
• Nonlinearity	• Box-Cox and Box-Tidwell
• Multicollinearity	• Factor analysis
• Heteroscedasticity	• Huber regression
• Serial correlation	• Prais-Winsten
• Outliers	• Drop outlier

Regression Assumptions, Diagnostics, and Fraud Detection

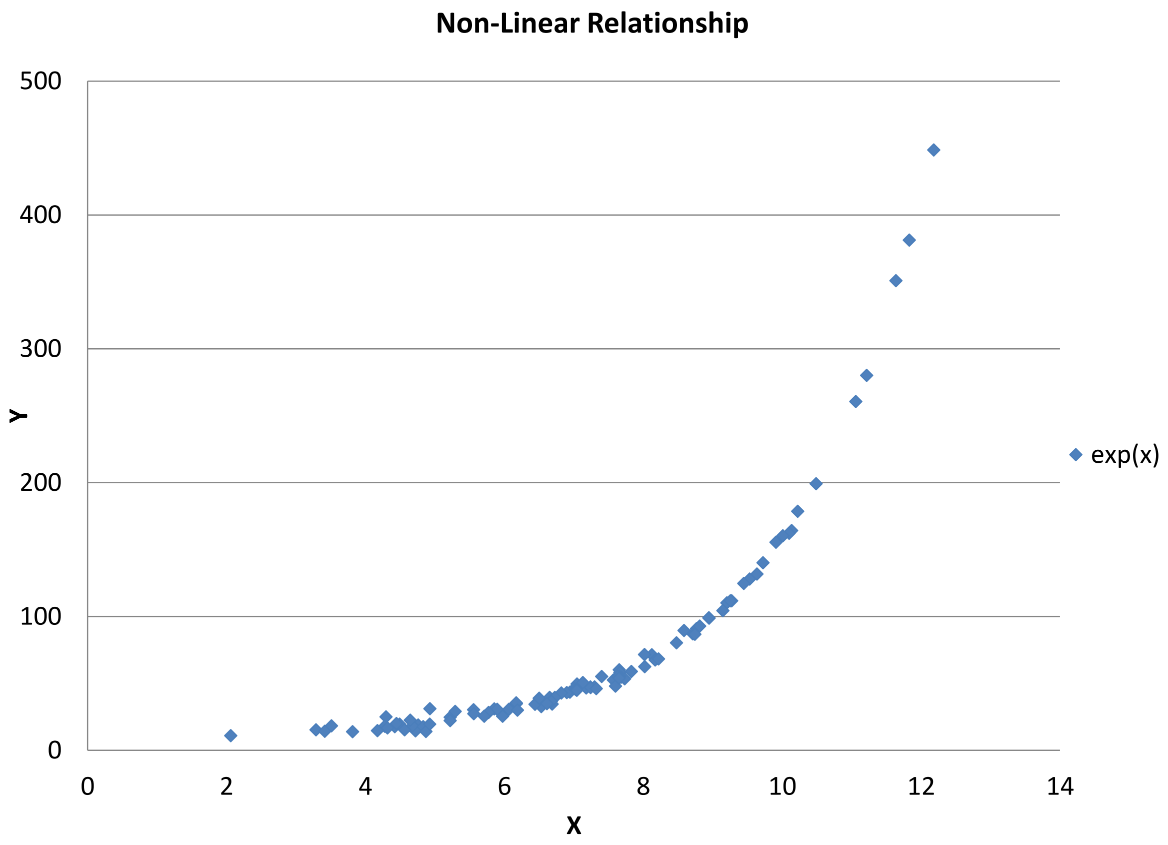
The End

R

Regression Linearity Test

R: Regression Linearity Test

- Problem: data is nonlinear
- Test: RESET
- Correction: Box-Cox and Box-Tidwell



R: Regression Linearity Test

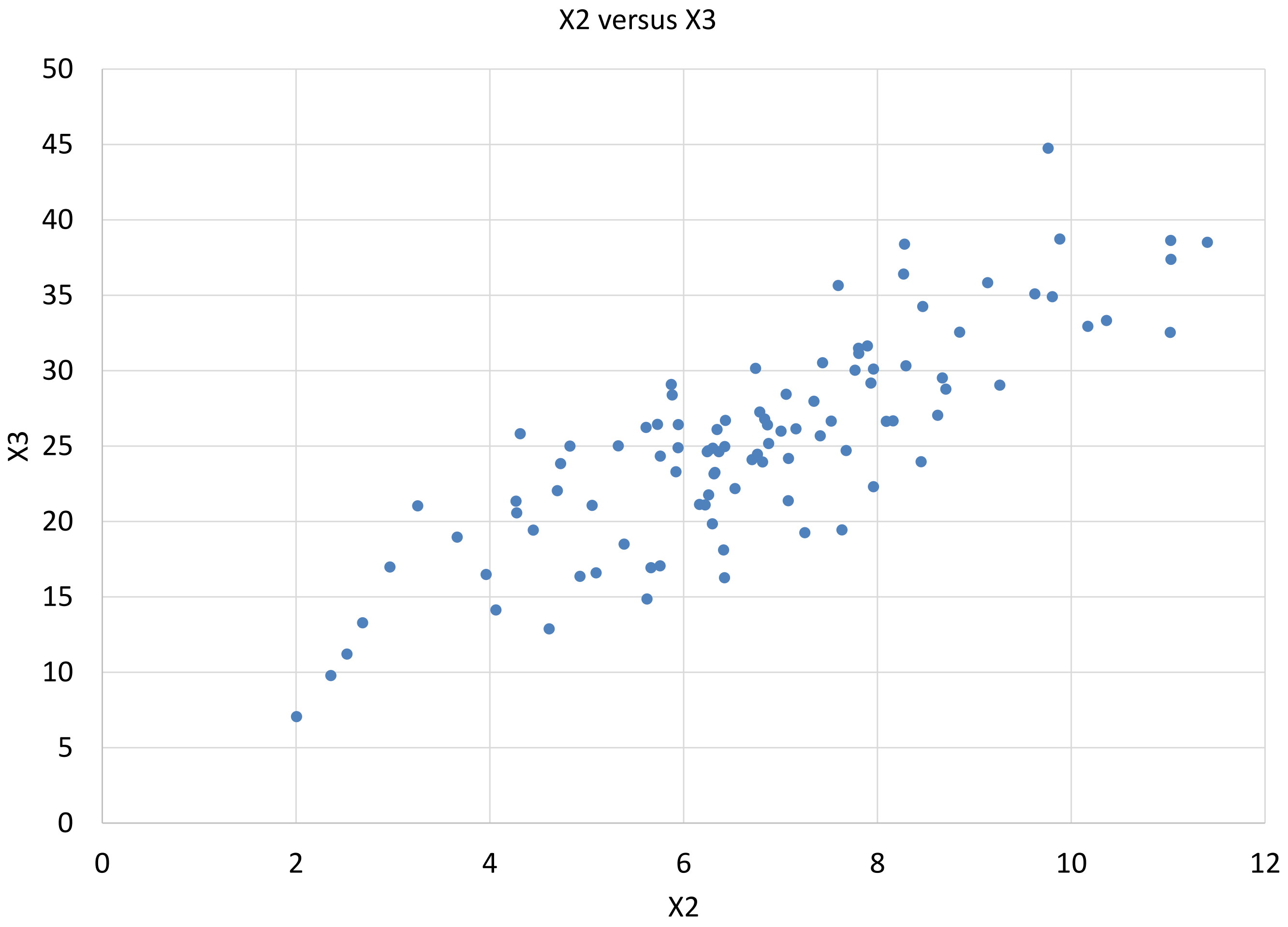
The End

R

Collinearity Test

R: Collinearity Test

- Problem: X-variables are correlated
- Test: variance inflation factor (VIF)
- Correction: factor analysis



R: Collinearity Test

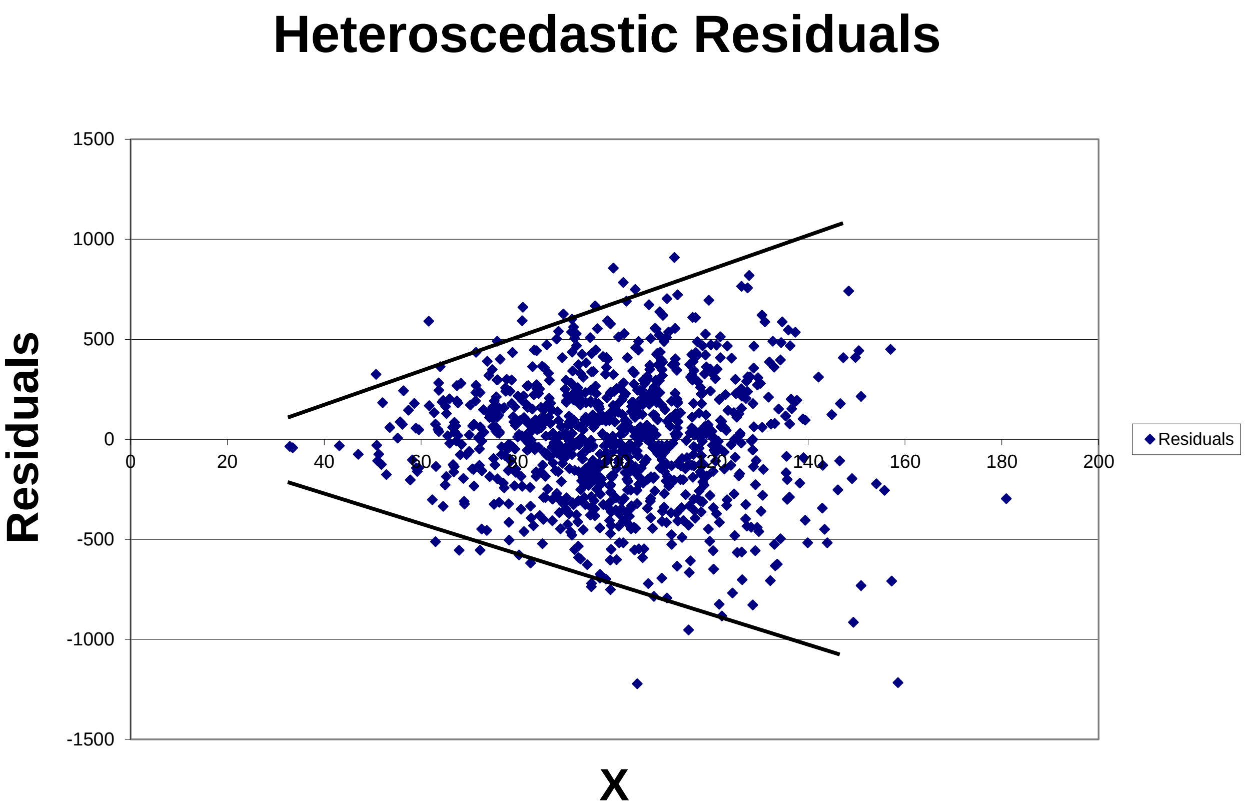
The End

R

Heteroscedasticity Test

R: Heteroscedasticity

- Problem: magnitude of residuals changes as X changes (nonconstant variance)
- Test: Breusch-Pagan
- Correction: Huber regression



R: Heteroscedasticity

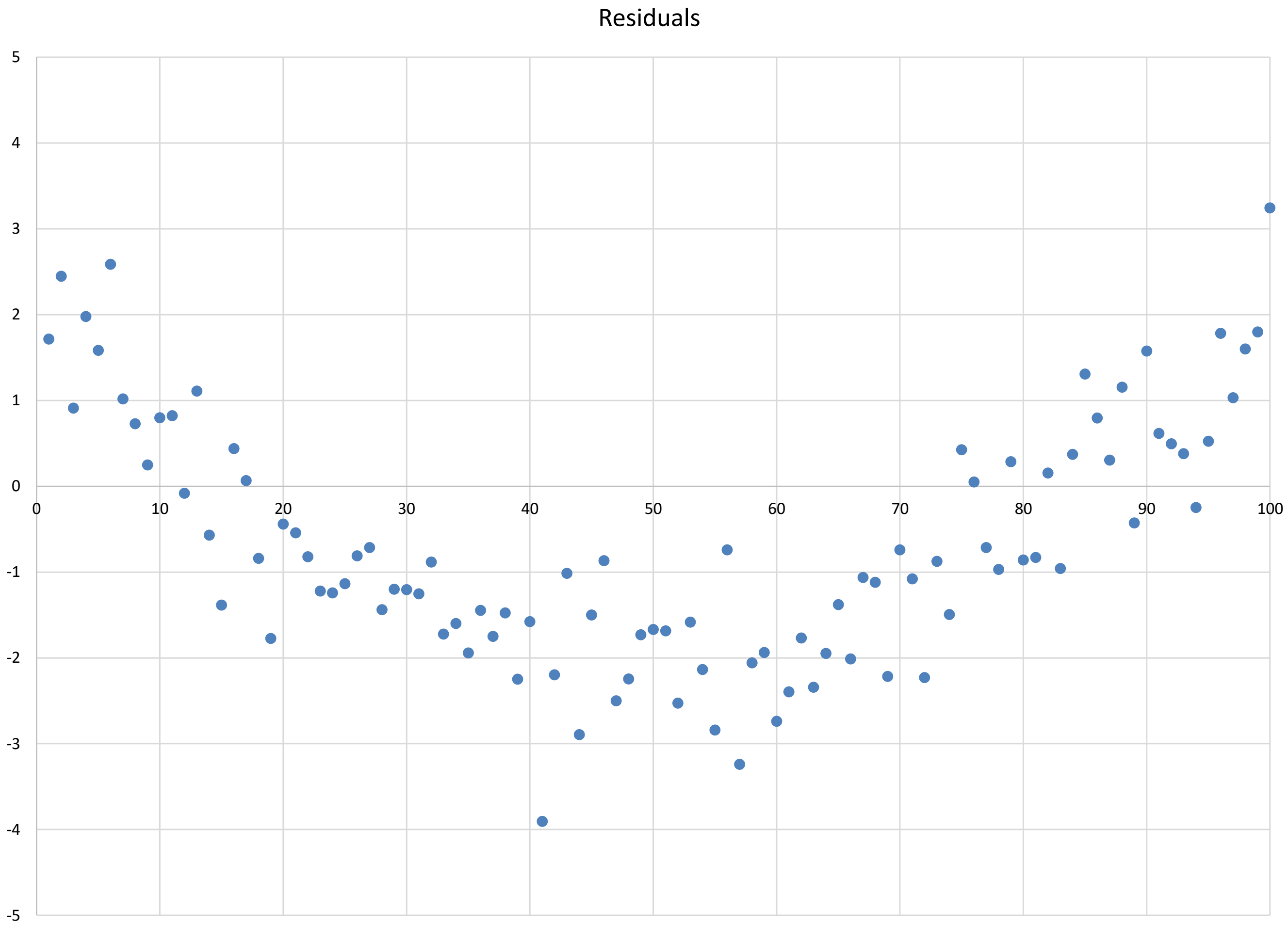
The End

R

Serial Correlation Test (Durbin-Watson)

R: Serial Correlation

- Problem: error terms are correlated
- Test: Durbin-Watson
- Correction: Prais-Winsten, Cochrane-Orcutt, ARCH, rho-differencing



R: Serial Correlation

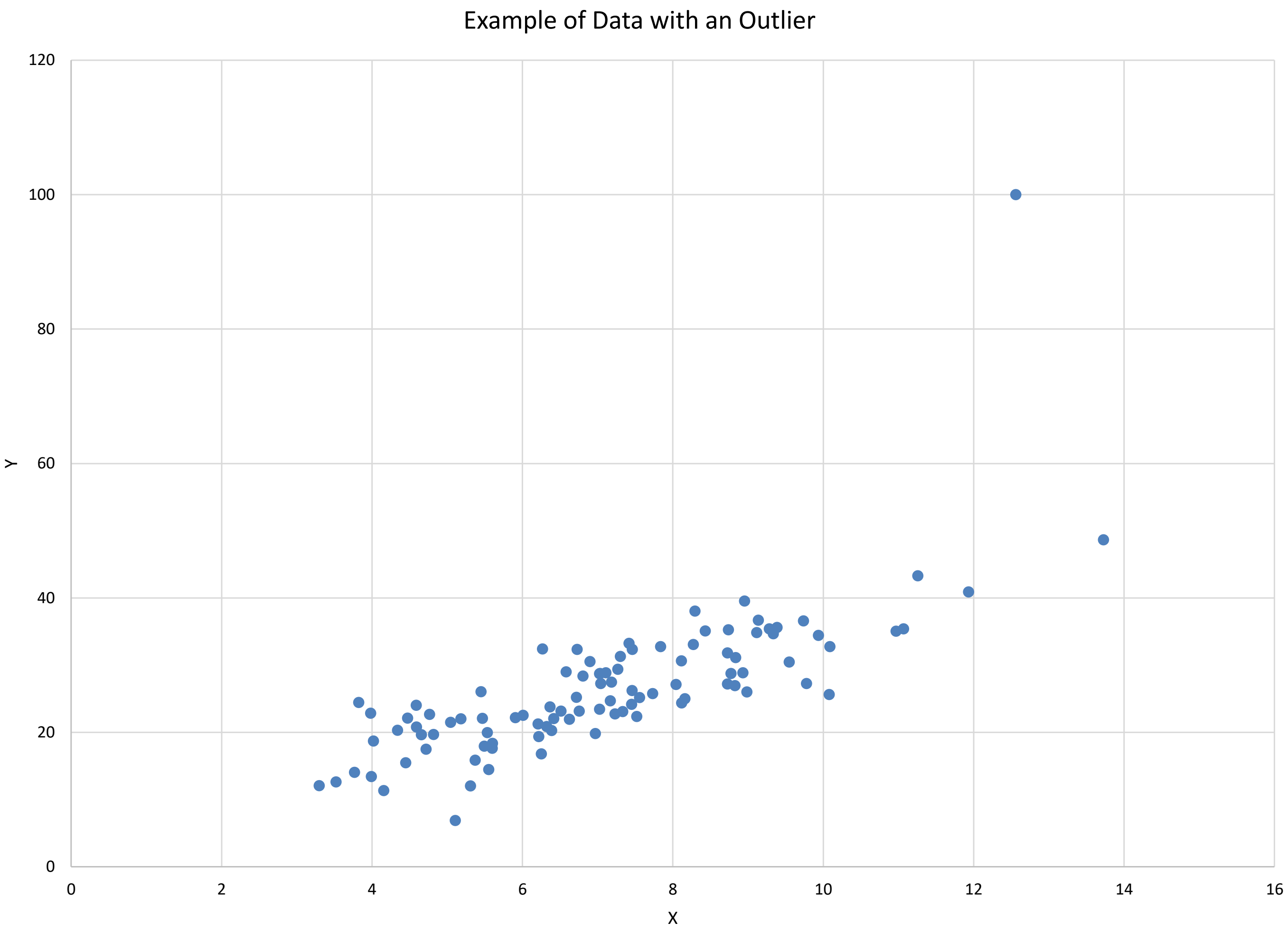
The End

R

Outlier Test

R: Outliers

- Problem: outliers are data points far from the line, artificially influencing the slope
- Test: Bonferroni
- Correction: drop outliers



R: Outliers

The End

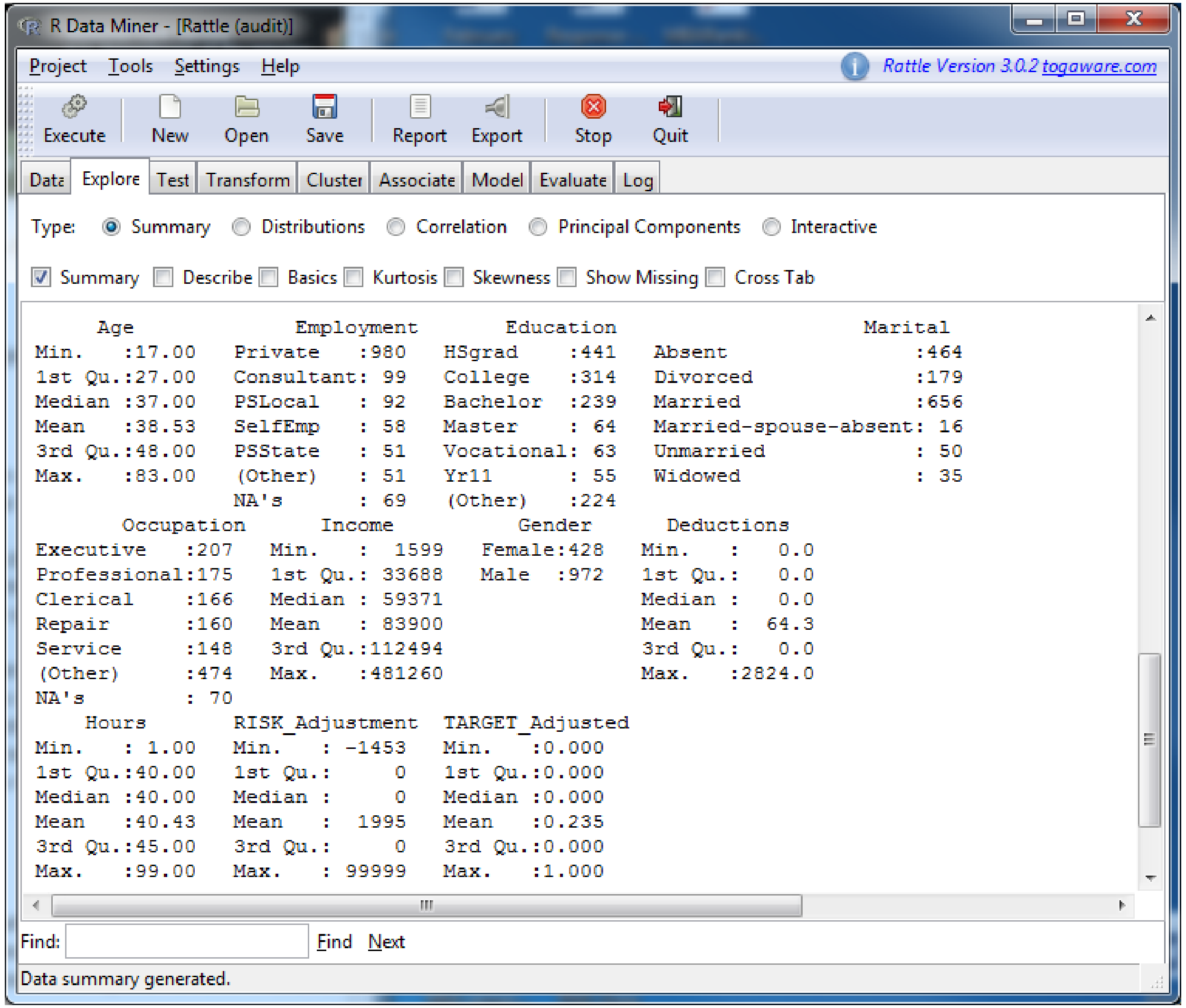
R

Data Mining Using Rattle

R: Data Mining Using Rattle

Rattle is a data mining package in R which can create summary statistics including:

- Quartiles across multiple variables simultaneously



Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

☑ Summary □ Describe □ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab

```
     Age             Employment          Education                        Marital
Min.   :17.00   Private   :980   HSgrad    :441   Absent               :464
1st Qu.:27.00   Consultant: 99   College   :314   Divorced             :179
Median :37.00   PSLocal   : 92   Bachelor  :239   Married              :656
Mean   :38.53   SelfEmp   : 58   Master    : 64   Married-spouse-absent: 16
3rd Qu.:48.00   PSState   : 51   Vocational: 63   Unmarried            : 50
Max.   :83.00   (Other)   : 51   Yr11      : 55   Widowed              : 35
                NA's      : 69   (Other)   :224
       Occupation          Income        Gender      Deductions
Executive   :207   Min.   :  1599   Female:428   Min.   :   0.0
Professional:175   1st Qu.: 33688   Male  :972   1st Qu.:   0.0
Clerical    :166   Median : 59371                Median :   0.0
Repair      :160   Mean   : 83900                Mean   :  64.3
Service     :148   3rd Qu.:112494                3rd Qu.:   0.0
(Other)     :474   Max.   :481260                Max.   :2824.0
NA's        : 70
    Hours       RISK_Adjustment  TARGET_Adjusted
Min.   : 1.00   Min.   : -1453   Min.   :0.000
1st Qu.:40.00   1st Qu.:     0   1st Qu.:0.000
Median :40.00   Median :     0   Median :0.000
Mean   :40.43   Mean   :  1995   Mean   :0.235
3rd Qu.:45.00   3rd Qu.:     0   3rd Qu.:0.000
Max.   :99.00   Max.   : 99999   Max.   :1.000
```

Find: Find Next

Data summary generated.

R: Data Mining Using Rattle

The End

R

Benford’s Law

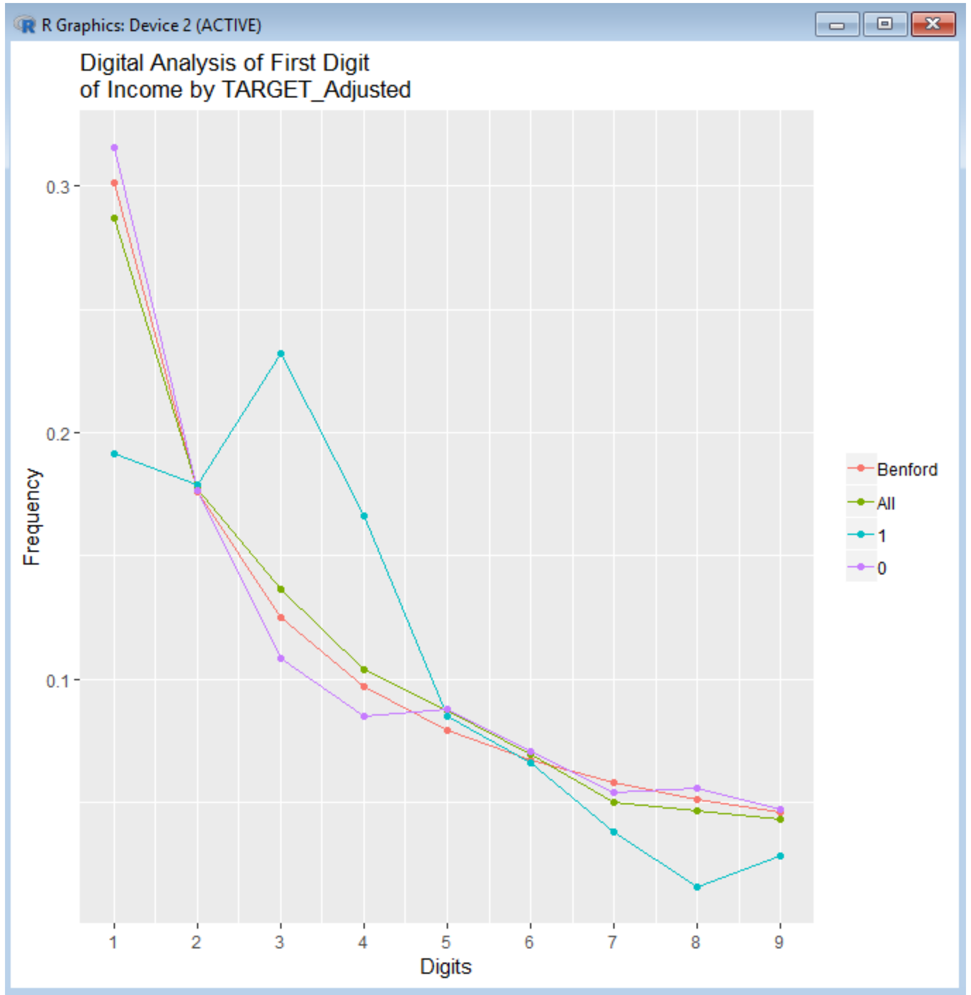
R: Benford’s Law

- Benford’s law states that financial numbers tend to start with smaller digits.
- Deviations from this distribution are often a signal of fraud.
- Benford’s law can detect fraud in expense reports, accounts receivable, accounts payable, financial statements, and income tax returns.

First digit	Frequency
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

R: Benford's Law

- Deviations from Benford's distribution can signal fraud.
- The graph on the right shows the first digit of income on tax returns
- Group 1 returns were found to be fraudulent; group 0 were not fraudulent



R: Benford’s Law

The End