5 of 19

Regression Assumptions and Diagnostics

- Linear regression can only be performed when a set of assumptions are satisfied.
- Excel can perform regressions but has limited ability to test the assumptions.
- R has diagnostics to determine if the assumptions are met.

13 of 19

Linear Regression Assumptions

- The relationship between X and Y is linear.
- If there are multiple X variables, they are not correlated (no multicollinearity).
- The error terms (distance from the data point to the predicted line):
    - Have zero mean and constant variance (no heteroscedasticity)
    - Are independent (no serial correlation)
    - Are normally distributed (includes no outliers)

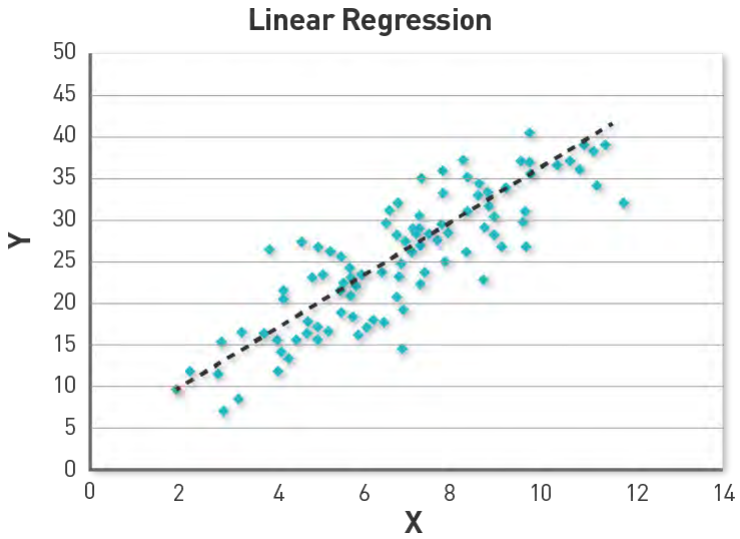
Each of these will be explained in the video for each assumption and diagnostic test.

Diagnostics and Solutions

Assumption	Diagnostic	Solution
Linearity	Ramsey	Transformation
No multicollinearity	Variance inflation factor	Combine variables or drop one
Heteroscedasticity	Breusch-Pagan	Transformation
No serial correlation	Durbin-Watson	Time series analysis
Normality/Outliers	Bonferroni outlier test	Drop outliers

View All View Keyframes

Regression Assumption: Linearity

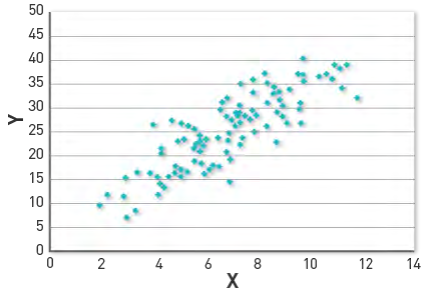


For linear regression, the data must be from a linear relationship.

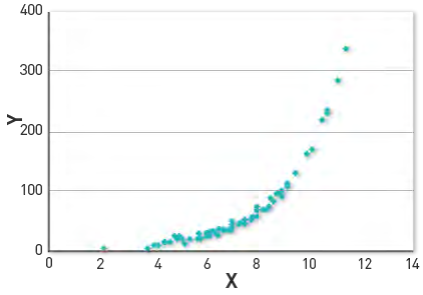
2 of 10

Linear vs. Nonlinear Relationship Illustration

Linear relationship



Nonlinear relationship



5 of 10

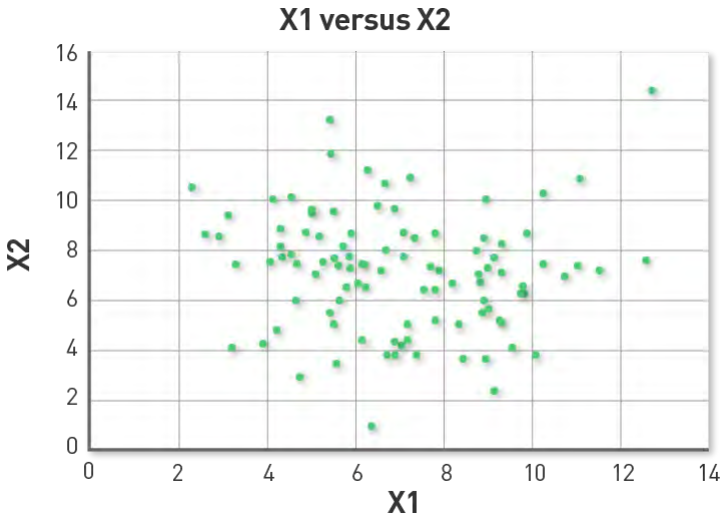
Diagnostics and Solutions

- Diagnostic test:
  - Ramsey Regression Equation Specification Error Test (RESET) (1969)
- Solution:
  - Take a transformation of the data (e.g., logarithm, square, square root, or inverse).
    - Advanced techniques (i.e., Box-Cox for Y variables, Box-Tidwell for X variables) help determine the appropriate transformation.

10 of 10

View All View Keyframes

Regression Assumption: Collinearity

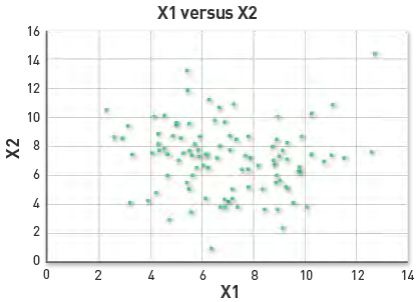


Each pair of X variables must not be correlated.

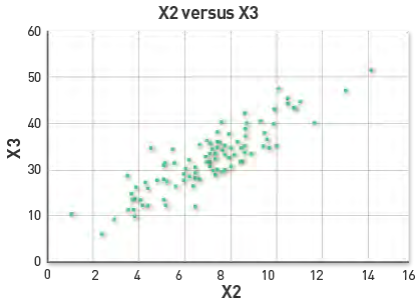
2 of 10

Not Correlated vs. Correlated: Illustration

Not correlated



Correlated



5 of 10

Diagnostics and Solution

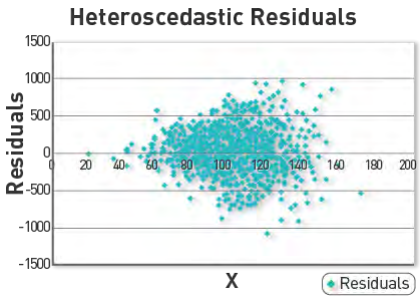
- Diagnostic test:
  - Variance inflation factor (VIF)
- Solution:
  - Drop one of the variables.
  - Use advanced techniques (e.g., factor analysis) to combine variables into one factor.

10 of 10

View All View Keyframes

Regression Assumption: Heteroscedasticity

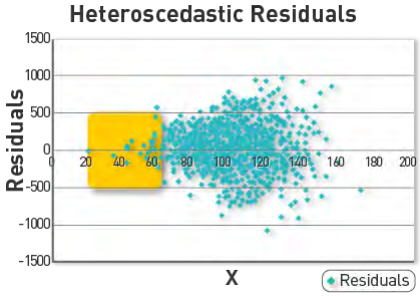
- Residuals (error terms) are the difference between the regression line and the data point.
- The error terms must have zero mean and constant variance.



3 of 12

Heteroscedasticity, Illustrated

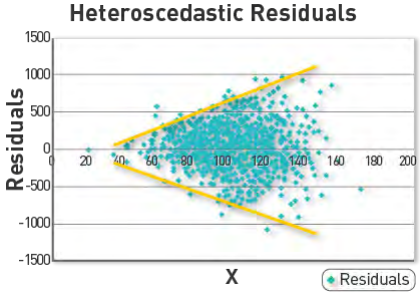
- Is there a pattern?



6 of 12

Heteroscedasticity, Illustrated (cont.)

- Is there a pattern?
- Variance increases as X increases (i.e., variance is not constant).



7 of 12

Diagnostics and Solutions

- Diagnostic test:
  - Breusch-Pagan test of heteroscedasticity

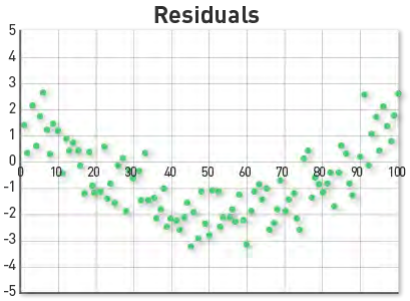
- Solution:
  - A transformation usually fixes this problem (e.g., logarithm, square, square root, or inverse).
  - If the problem still occurs after a transformation, Huber regression corrects for the error.

12 of 12

View All View Keyframes

Regression Assumption: Serial Correlation

- Error terms must not be related to each other.
- Serial correlation often occurs with time-series data.



3 of 7

Diagnostics and Solutions

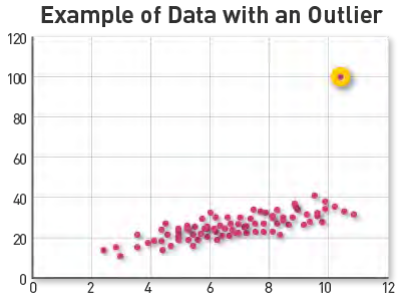
- Diagnostic test:
  - Durbin-Watson test of serial correlation
- Solution:
  - Advanced techniques include rho differencing and ARCH.

7 of 7

View All View Keyframes

Regression Assumption: Serial Correlation

- Outliers are data points that are significantly different from the other points.
- An influential outlier can twist the regression line away from its true position.



5 of 10

Diagnostics and Solutions

- Diagnostic test:
  - Bonferroni outlier test
- Solution:
  - If the data point is clearly an outlier, drop the bad data point.
  - Always report that outliers are dropped from the analysis.

10 of 10

View All View Keyframes

Data Mining

- Data mining is a data exploration process that searches for patterns or relationships within data.
- Techniques combine machine learning, statistics, and database systems.
- Some approaches include decision trees, logistic regression, perceptrons, and neural networks.
- Data mining requires new features in R, including installation of the Rattle package.

5 of 5

View All View Keyframes

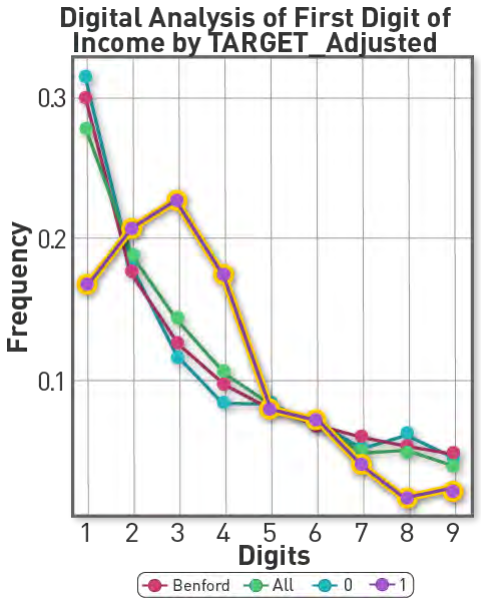
Benford's Law

- Accountants and auditors use data mining techniques to detect fraud.
  - One technique relies on Benford's law.
- Benford's law states that in financial data, small digits appear more often than larger digits at the beginning of a number.
  - E.g., 30% of all financial numbers begin with the number 1; fewer than 5% begin with the number 9.

4 of 7

Benford's Law Illustration

- If the initial digits in the data do not match the frequency specified by Benford's law, the numbers are likely fraudulent.



7 of 7

View All View Keyframes

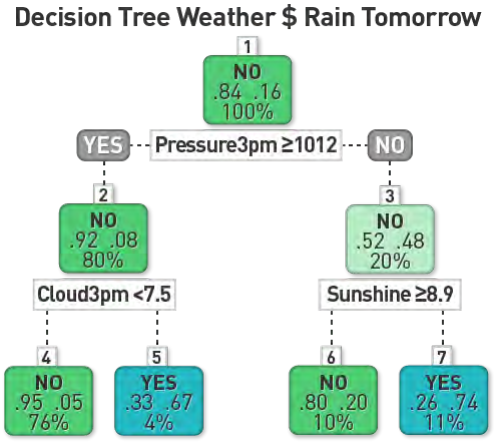
Decision Trees

- Decision trees identify which variables are most important for making decisions or predictions.
- The technique removes noise (entropy) and identifies which variables contribute the most information.
- The result is a set of rules that can assist in decision making.

4 of 8

Decision Tree Illustration: Weather Forecasting

- A simple decision tree can help predict if it will rain tomorrow.



8 of 8