Midwest Big Data Summer School: Introduction to Statistics

Kris De Brabanter

kbrabant@iastate.edu

Iowa State University
Department of Statistics
Department of Computer Science

June 20, 2016

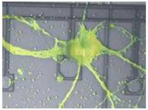
Outline

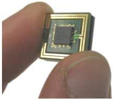
1. What is Statistics?
2. Measures of central tendency and variance
3. Data types
4. How to visualize data?
   - Boxplot
   - Histogram
5. Regression
   - Linear regression
   - Nonparametric regression

Statistics: Some examples

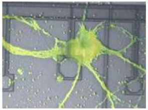
Statistics: Some examples

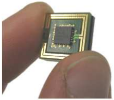
Statistics: Some examples

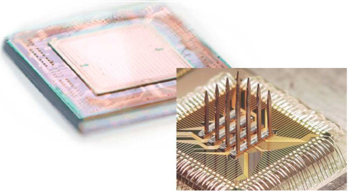




Statistics: Some examples

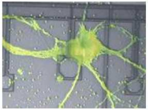





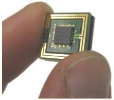


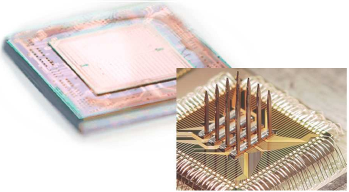
Technologies for the intelligent environment

Statistics: Some examples



Renewable Energy:
Next Generation
Solar Cells





Technologies for the
intelligent environment

What is Statistics?

- Google: "The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample."

What is Statistics?

- Google: "The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample."
- Wikipedia: "The study of the collection, analysis, interpretation, presentation, and organization of data."

What is Statistics?

- Google: “The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.”
- Wikipedia: “The study of the collection, analysis, interpretation, presentation, and organization of data.”
- Sir Arthur Lyon Bowley: ”Numerical statements of facts in any department of inquiry placed in relation to each other.“

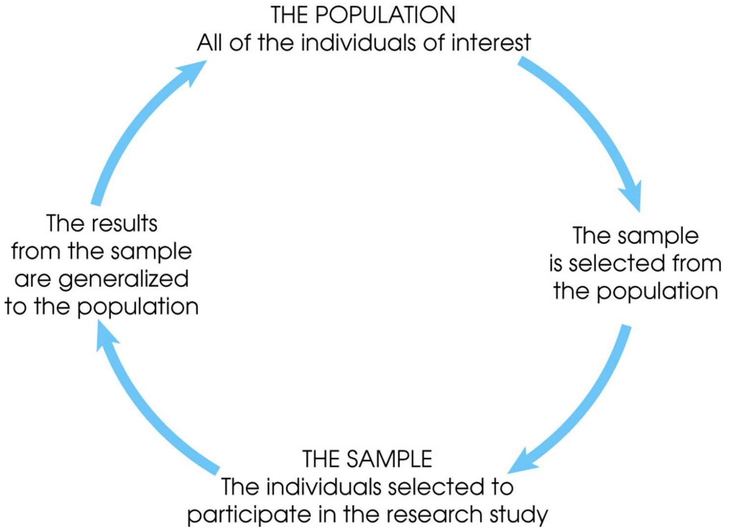
What is Statistics?

- Google: "The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample."
- Wikipedia: "The study of the collection, analysis, interpretation, presentation, and organization of data."
- Sir Arthur Lyon Bowley: "Numerical statements of facts in any department of inquiry placed in relation to each other."
- BusinessDictionary.com: "Branch of mathematics concerned with collection, classification, analysis, and interpretation of numerical facts, for drawing inferences on the basis of their quantifiable likelihood. Statistics can interpret aggregates of data too large to be intelligible by ordinary observation because such data tend to behave in regular, predictable manner..."

Descriptive vs. Inferential Statistics

- Descriptive statistics: "Analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data."
- Inferential statistics: "...makes inferences about populations using data drawn from the population. Instead of using the entire population to gather the data, the statistician will collect a sample or samples from the millions of residents and make inferences about the entire population using the sample.."

Sample vs. Population



1 What is Statistics?

2 Measures of central tendency and variance

3 Data types

4 How to visualize data?
- Boxplot
- Histogram

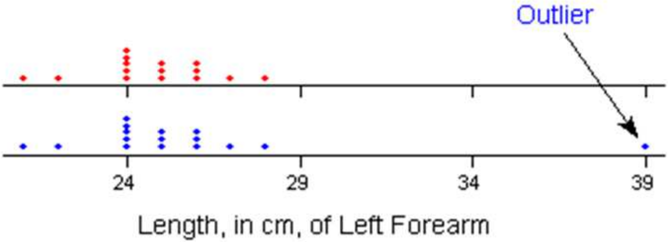
5 Regression
- Linear regression
- Nonparametric regression

Sample mean & variance: what can go wrong...

$$\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad\qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

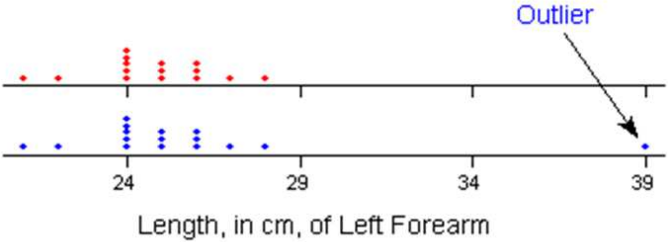
Sample mean & variance: what can go wrong...

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$



Sample mean & variance: what can go wrong...

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

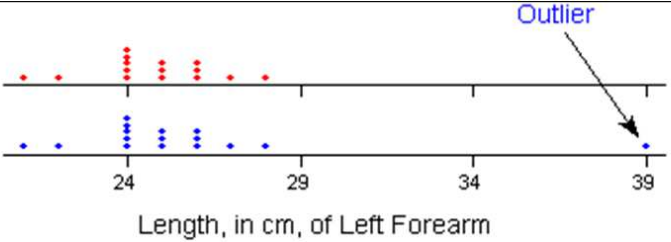


	Mean	Variance
No outlier	24.733	1.792
With outlier	25.625	3.964

Sample mean & variance: what can go wrong...

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

mean & variance NOT robust



	Mean	Variance
No outlier	24.733	1.792
With outlier	25.625	3.964

Solution: sample median & median absolute deviation

Given the order statistics: $X_{(1)} \leq \ldots \leq X_{(n)}$

$$\boxed{\mathrm{median}(X) = \begin{cases} X_{((n+1)/2)}, & \text{if } n \text{ is odd;} \\ \frac{1}{2}(X_{(n/2)} + X_{(1+n/2)}), & \text{if } n \text{ is even.} \end{cases}}$$

Solution: sample median & median absolute deviation

Given the order statistics: $X_{(1)} \leq \ldots \leq X_{(n)}$

$$\boxed{\text{median}(X) = \begin{cases} X_{((n+1)/2)}, & \text{if } n \text{ is odd;} \\ \frac{1}{2}(X_{(n/2)} + X_{(1+n/2)}), & \text{if } n \text{ is even.} \end{cases}}$$

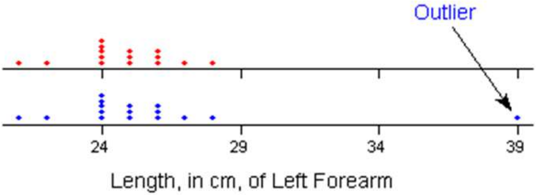
$$\boxed{\text{MAD} = \text{median}(|X_i - \text{median}(X)|)}$$

Solution: sample median & median absolute deviation

Given the order statistics: $X_{(1)} \leq \ldots \leq X_{(n)}$

$$\boxed{\mathrm{median}(X) = \begin{cases} X_{((n+1)/2)}, & \text{if } n \text{ is odd;} \\ \frac{1}{2}(X_{(n/2)} + X_{(1+n/2)}), & \text{if } n \text{ is even.} \end{cases}}$$

$$\boxed{\mathrm{MAD} = \mathrm{median}(|X_i - \mathrm{median}(X)|)}$$

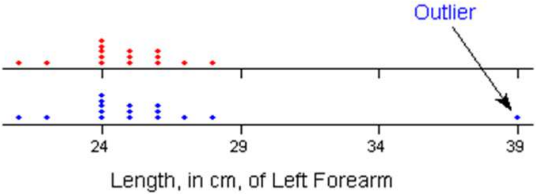


Solution: sample median & median absolute deviation

Given the order statistics: $X_{(1)} \leq \ldots \leq X_{(n)}$

$$\boxed{\text{median}(X) = \begin{cases} X_{((n+1)/2)}, & \text{if } n \text{ is odd;} \\ \frac{1}{2}(X_{(n/2)} + X_{(1+n/2)}), & \text{if } n \text{ is even.} \end{cases}}$$

$$\boxed{\text{MAD} = \text{median}(|X_i - \text{median}(X)|)}$$



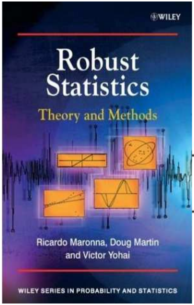
	Mean	Variance	Median	MAD
No outlier	24.733	1.792	25	1
With outlier	25.625	3.964	25	1

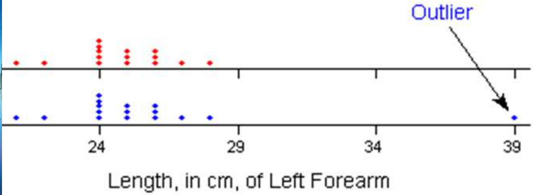
Solution: sample median & median absolute deviation

Given the order statistics: $X_{(1)} \leq \ldots \leq X_{(n)}$

$$\boxed{\operatorname{median}(X) = \begin{cases} X_{((n+1)/2)}, & \text{if } n \text{ is odd;} \\ \frac{1}{2}\left(X_{(n/2)} + X_{(1+n/2)}\right), & \text{if } n \text{ is even.} \end{cases}}$$

$$\boxed{\text{MAD} = \operatorname{median}(|X_i - \operatorname{median}(X)|)}$$





	Mean	Variance	Median	MAD
No outlier	24.733	1.792	25	1
With outlier	25.625	3.964	25	1

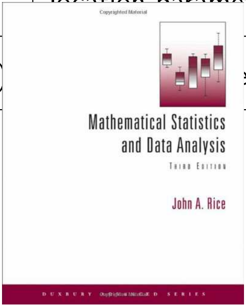
1. What is Statistics?
2. Measures of central tendency and variance
3. Data types
4. How to visualize data?
   - Boxplot
   - Histogram
5. Regression
   - Linear regression
   - Nonparametric regression

Data types

Data Type	Possible Values	Example	Permissible Statistics
binary	0,1	yes/no	mode, χ^2
categorial	1,2,...,K	blood type, color	mode, χ^2
ordinal	integer/real/order	score/rank	mode, median,...
binomial	$0, 1, \ldots, N$	# successes	mean, median,...
count	integers $(+)$	# items	Interval scales
real valued additive	real number	location parameter	mean, mode,...
real valued multiplicative	real number $(+)$	scale parameter	Interval scales

Data types

Data Type	Possible Values	Example	Permissible Statistics
binary	0,1	yes/no	mode, χ^2
categorial	1,2,. . .,K	blood type, color	mode, χ^2
ordinal	integer/real/order	score/rank	mode, median,...
binomial	$0, 1, \ldots, N$	# successes	mean, median,...
count	integers (+)	# items	Interval scales
real valued additive	real number	location parameter	mean, mode,...
real valued multiplicative	real number (+)	[illegible]er	Interval scales



How to visualize data?

1. What is Statistics?
2. Measures of central tendency and variance
3. Data types
4. How to visualize data?
   - Boxplot
   - Histogram
5. Regression
   - Linear regression
   - Nonparametric regression

Boxplot & quartiles

Definition (quartiles)

The quartiles of a ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data.

Boxplot & quartiles

Definition (quartiles)

The quartiles of a ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data.

- **First quartile** (Q_1): splits off the lowest 25% of data from the highest 75%

Boxplot & quartiles

Definition (quartiles)

The quartiles of a ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data.

- **First quartile** (Q_1): splits off the lowest 25% of data from the highest 75%
- **Second quartile** (Q_2 or median): cuts data set in half

Boxplot & quartiles

Definition (quartiles)

The quartiles of a ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data.

- **First quartile** (Q_1): splits off the lowest 25% of data from the highest 75%
- **Second quartile** (Q_2 or median): cuts data set in half
- **Third quartile** (Q_3): splits off the highest 25% of data from the lowest 75%

Boxplot & quartiles

Definition (quartiles)

The quartiles of a ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data.

- **First quartile** (Q_1): splits off the lowest 25% of data from the highest 75%
- **Second quartile** (Q_2 or median): cuts data set in half
- **Third quartile** (Q_3): splits off the highest 25% of data from the lowest 75%
- **Interquartile range**: $IQR = Q_3 - Q_1$

Boxplot & quartiles: Example

Consider the following ordered data set:

$$6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49$$

Boxplot & quartiles: Example

Consider the following ordered data set:

$$6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49$$

- $Q_1 = 15$

Boxplot & quartiles: Example

Consider the following ordered data set:

$$6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49$$

- $Q_1 = 15$
- $Q_2 = 40$

Boxplot & quartiles: Example

Consider the following ordered data set:

$$6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49$$

- $Q_1 = 15$
- $Q_2 = 40$
- $Q_3 = 43$

Boxplot & quartiles: Example

Consider the following ordered data set:

$$6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49$$

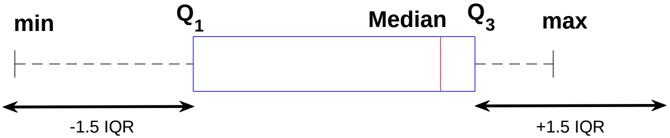
- $Q_1 = 15$
- $Q_2 = 40$
- $Q_3 = 43$
- $IQR = Q_3 - Q_1 = 43 - 15 = 28$

Boxplot & quartiles: Example

Consider the following ordered data set:

$$6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49$$

- $Q_1 = 15$
- $Q_2 = 40$
- $Q_3 = 43$
- $IQR = Q_3 - Q_1 = 43 - 15 = 28$

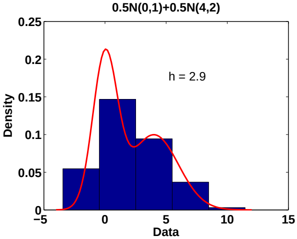


Histogram

- Graphical representation of the density of the data
- Available in each statistical software package (Matlab, R, etc.)

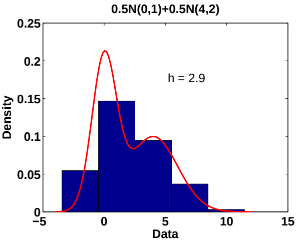
Histogram

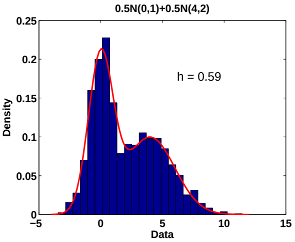
- Graphical representation of the density of the data
- Available in each statistical software package (Matlab, R, etc.)



Histogram

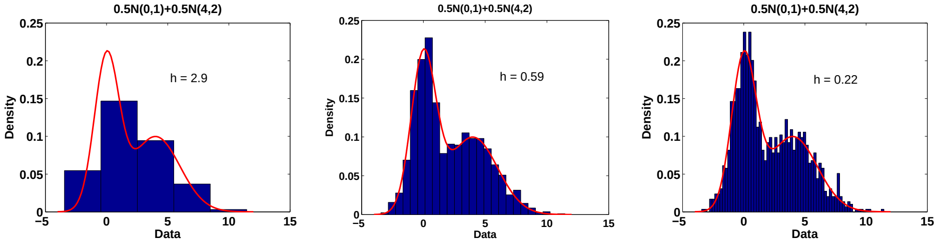
- Graphical representation of the density of the data
- Available in each statistical software package (Matlab, R, etc.)





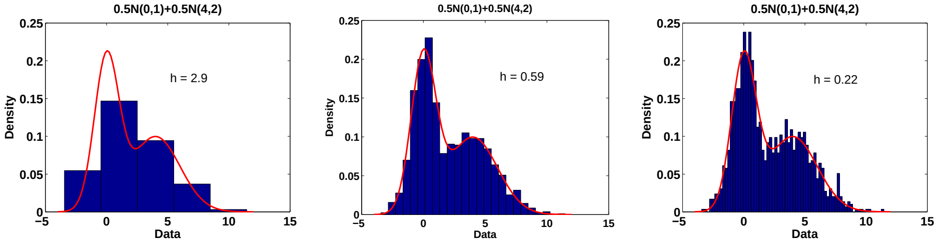
Histogram

- Graphical representation of the density of the data
- Available in each statistical software package (Matlab, R, etc.)



Histogram

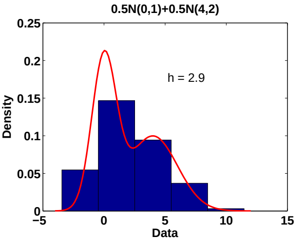
- Graphical representation of the density of the data
- Available in each statistical software package (Matlab, R, etc.)

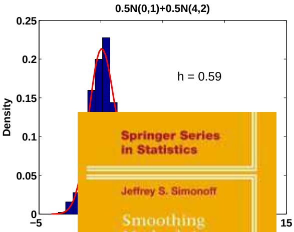


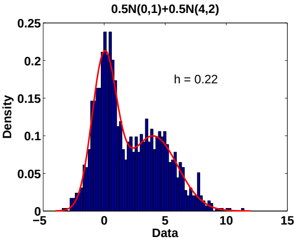
Binwidth h is crucial

Histogram

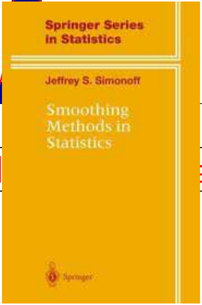
- Graphical representation of the density of the data
- Available in each statistical software package (Matlab, R, etc.)



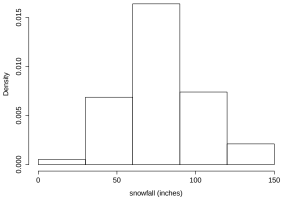




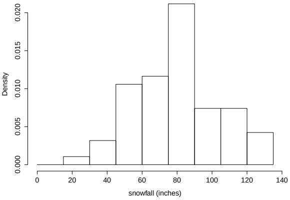
Binwid... al



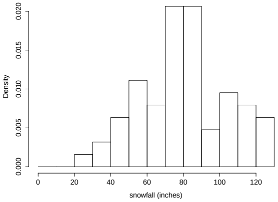
Effect of binwidth: Annual snowfall in Buffalo (NY) from 1910 to 1972



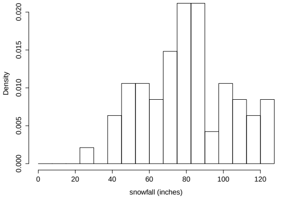
(a) $h = 30$



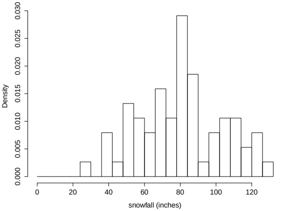
(b) $h = 15$



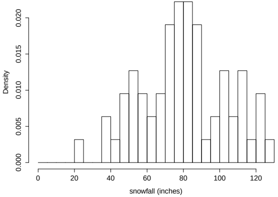
(c) $h = 10$



(d) $h = 7.5$

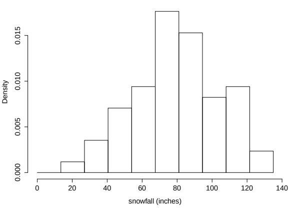


(e) $h = 6$

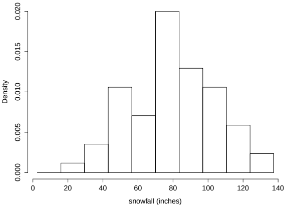


(f) $h = 5$

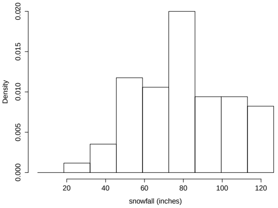
Effect of different origin: Annual snowfall in Buffalo (NY) from 1910 to 1972 with binwidth $h = 13.5$



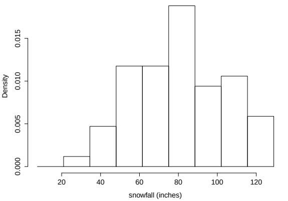
(g) $t_0 = 0$



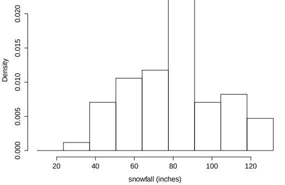
(h) $t_0 = 2.5$



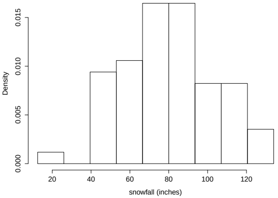
(i) $t_0 = 5$



(j) $t_0 = 7.5$



(k) $t_0 = 10$



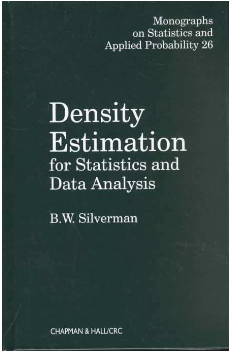
(l) $t_0 = 12.5$

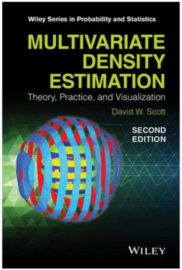
More advanced methods

In order to overcome the choice of origin, one could use average shifted histograms or kernel density estimation. Both have a parameter similar to the binwidth.

More advanced methods

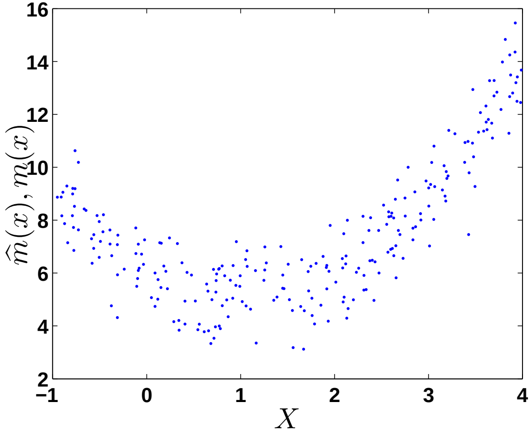
In order to overcome the choice of origin, one could use average shifted histograms or kernel density estimation. Both have a parameter similar to the binwidth.



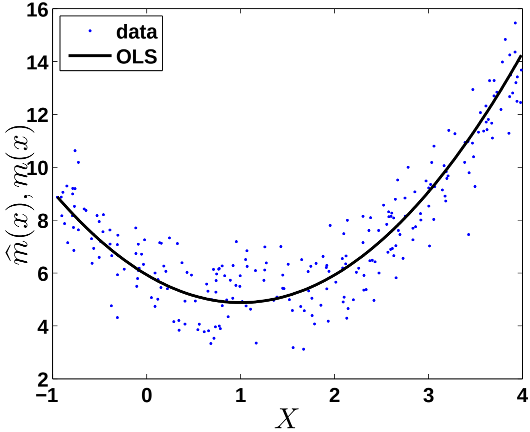


1. What is Statistics?
2. Measures of central tendency and variance
3. Data types
4. How to visualize data?
   - Boxplot
   - Histogram
5. Regression
   - Linear regression
   - Nonparametric regression

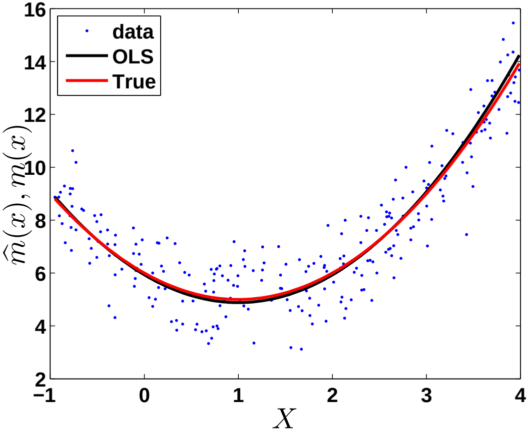
Formulation of the problem statement



Formulation of the problem statement



Formulation of the problem statement



How to find the black line?

Ordinary least squares

- Model: $Y_i = \beta_0 + \beta_1 x_i + e_i,\ 1, \ldots, n$

Ordinary least squares

- Model: $Y_i = \beta_0 + \beta_1 x_i + e_i,\ 1, \ldots, n$
- How to find parameters β_0 and β_1 given data?

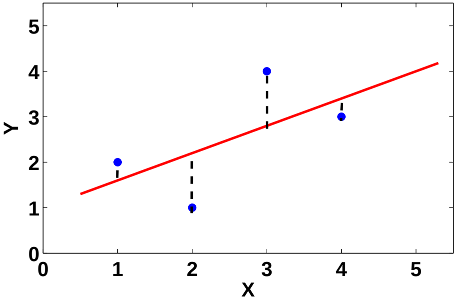
Ordinary least squares

- Model: $Y_i = \beta_0 + \beta_1 x_i + e_i, \; 1, \ldots, n$
- How to find parameters β_0 and β_1 given data?

Ordinary least squares

- Model: $Y_i = \beta_0 + \beta_1 x_i + e_i,\ 1, \ldots, n$
- How to find parameters β_0 and β_1 given data?

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1 \in \mathbb{R}^2}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$



Assumptions on the result of OLS

Definition

The **residual** $\hat{e}$ *of an observed value is the difference between the observed value and the estimated value of the quantity of interest. Mathematically*

$$\hat{e}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Assumptions on the result of OLS

Definition

The **residual** $\hat{e}$ *of an observed value is the difference between the observed value and the estimated value of the quantity of interest. Mathematically*

$$\hat{e}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

- Check visually for constant variance (homoskedasticity)

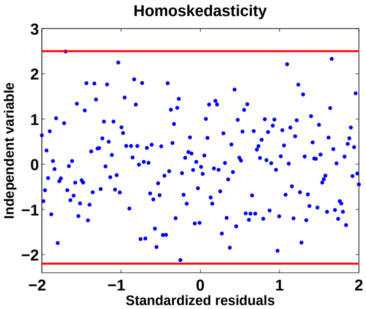
Assumptions on the result of OLS

Definition

The **residual** $\hat{e}$ *of an observed value is the difference between the observed value and the estimated value of the quantity of interest. Mathematically*

$$\hat{e}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

- Check visually for constant variance (homoskedasticity)



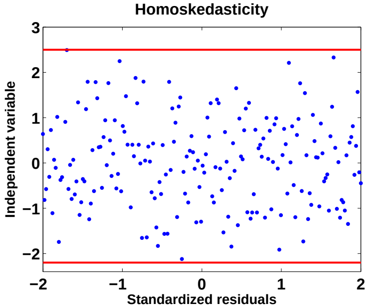
Assumptions on the result of OLS

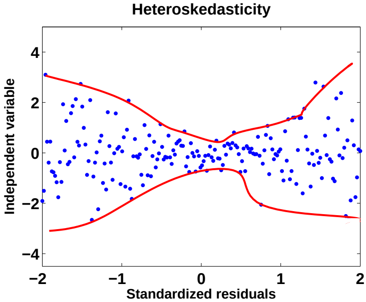
Definition

The **residual** $\hat{e}$ *of an observed value is the difference between the observed value and the estimated value of the quantity of interest. Mathematically*

$$\hat{e}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

- Check visually for constant variance (homoskedasticity)





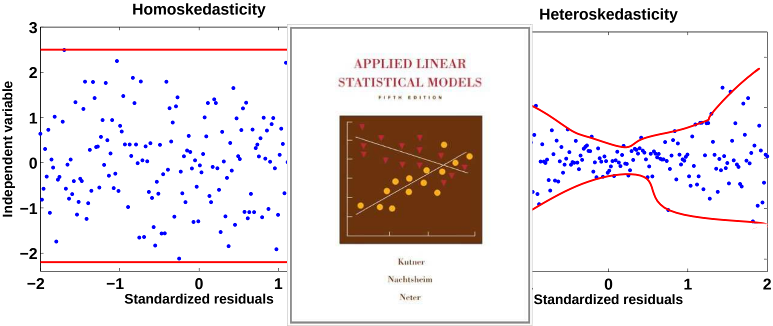
Assumptions on the result of OLS

Definition

The **residual** $\hat{e}$ *of an observed value is the difference between the observed value and the estimated value of the quantity of interest. Mathematically*

$$\hat{e}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

- Check visually for constant variance (homoskedasticity)

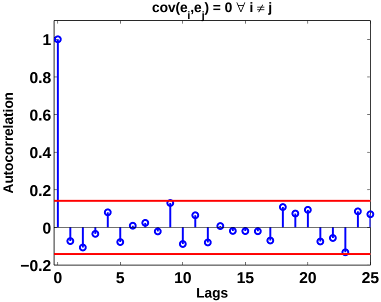


Assumptions on the result of OLS

- Plot autocorrelation function of residuals

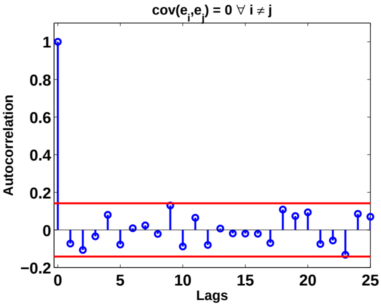
Assumptions on the result of OLS

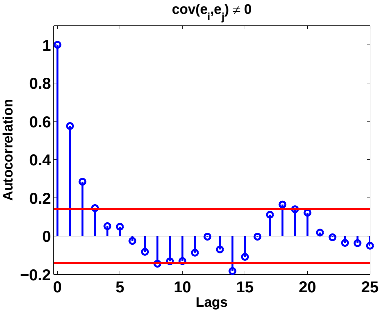
- Plot autocorrelation function of residuals



Assumptions on the result of OLS

- Plot autocorrelation function of residuals





Assumptions on the result of OLS

- Visual inspection is sometimes hard
- Hypothesis test: Breusch-Pagan, Engle's test,...
- Hypothesis test: Runs test (test of randomness)
- Cook's distance for leverage points

Assumptions on the result of OLS

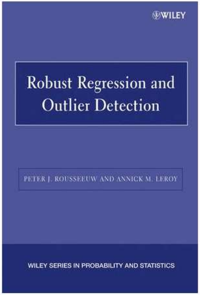
- Visual inspection is sometimes hard
- Hypothesis test: Breusch-Pagan, Engle's test,...
- Hypothesis test: Runs test (test of randomness)
- Cook's distance for leverage points

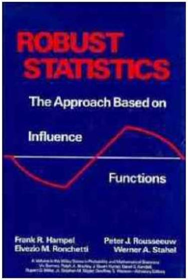
$\boxed{\text{One outlier can completely destroy the OLS estimate!!!!}}$

Assumptions on the result of OLS

- Visual inspection is sometimes hard
- Hypothesis test: Breusch-Pagan, Engle's test,...
- Hypothesis test: Runs test (test of randomness)
- Cook's distance for leverage points

One outlier can completely destroy the OLS estimate!!!!



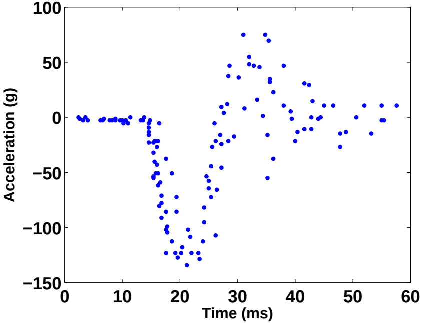


Nonparametric regression

- Why nonparametric regression?
  - Not always easy to find a suitable parametric model to explain some phenomena
  - Flexibility in data analysis

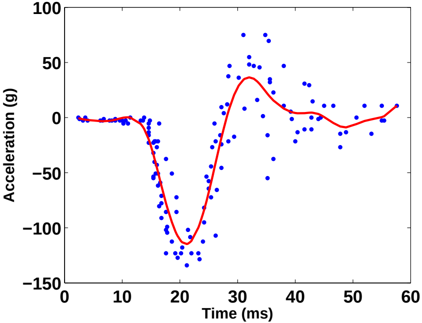
Nonparametric regression

- Why nonparametric regression?
  - Not always easy to find a suitable parametric model to explain some phenomena
  - Flexibility in data analysis



Nonparametric regression

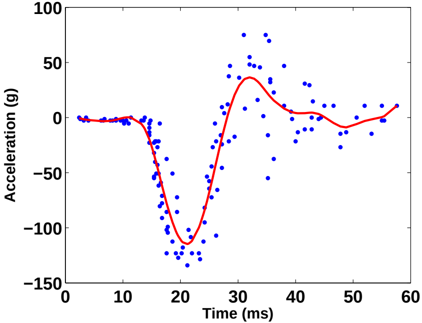
- Why nonparametric regression?
  - Not always easy to find a suitable parametric model to explain some phenomena
  - Flexibility in data analysis



  - *Let the data speak for itself*

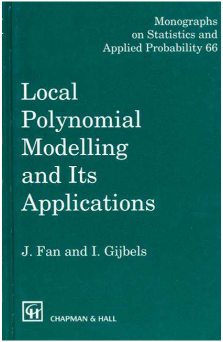
Nonparametric regression

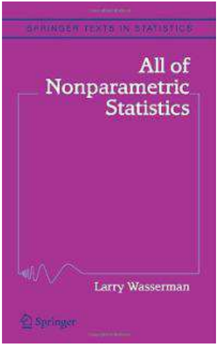
- Why nonparametric regression?
  - Not always easy to find a suitable parametric model to explain some phenomena
  - Flexibility in data analysis

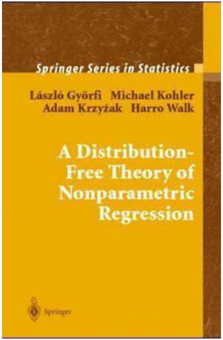


  - *Let the data speak for itself*
- Mainly developed in 1950s and 1960s
- Combination of parametric and nonparametric methods

Nonparametric regression (Cont'd)







References

- Fan J. & Gijbels I. (1996). *Local Polynomial Regression and Its Applications*, Chapman & Hall
- Györfi L, Kohler M., Krzyżak A. & Walk H. (2002). *A Distribution-Free Theory of Nonparametric Regression*, Springer
- Hampel F. R., Ronchetti E. M., Rousseeuw P. J. & Werner A. (1986), Robust Statistics: The Approach Based on Influence Functions, John Wiley & Sons
- Kutner M., Natchtsheim C., Neter J. (2004), Applied Linear Statistical Models (5th Ed.), McGraw-Hill/Irwin
- Maronna R., Martin D. & Yohai V. (2006). *Robust Statistics: Theory and Methods*, Wiley
- Rice J.A. (2007). *Mathematical Statistics and Data Analysis*, 3rd Ed., Brooks/Cole
- Rousseeuw P. J. & Leroy A. M. (2003), Robust Regression and Outlier Detection, Wiley
- Scott D. W., Multivariate Density Estimation: Theory, Practice and Visualization (2nd Ed.), Wiley, 2016
- Silverman B. W., Density Estimation for Statistics and Data Analysis, Chapman & Hall, 1986
- Simonoff J.S. (1996), *Smoothing Methods in Statistics*, Springer
- Wasserman L. (2006). *All of Nonparametric Statistics*, Springer