Diagnostics & Remedial Measures for SLR (Ch 3)
Week 5 – Thursday

Applied Regression Analysis (STAT 757)

Paul J. Hurtado

18 Feb, 2016

Department of Agriculture, Nutrition and Veterinary Sciences

LECTURE

Dr. Luis Moraes

Postdoctoral Research Fellow, Dept. of Animal Science at UC Davis

Candidate for Assistant Professor Position in Beef Cattle Production

Will Present:

Mixed Model Analysis for Agriculture, Nutrition and Veterinary Sciences

February 18, 2016 at 12:00pm in FA 109

Hyperbolic Geometry of Complex Network Data

Dr. Kostia Zuev – http://www.its.caltech.edu/~zuev/
California Institute of Technology

Today from **2:30-3:30** in **AB 635**

Abstract
Recent years have witnessed an explosion of new kind of data: network data. Large datasets of social, biological, technological, and information networks are analyzed by thousands of scientists around the world, making probabilistic modeling and statistical analysis of network data a mainstream research area in statistics, computer science, social sciences, system biology, and physics. One of the fundamental questions in the study of network data is to uncover hidden evolution mechanisms that shape the structure and dynamics of large real networks. It has been empirically observed that many real networks, in spite of being very different in other respects, have heavy-tail degree distribution, high clustering, and significant community structure. Since that discovery, several mechanisms were proposed to explain some of these universal properties, but none of them captured all the three properties at once. In this talk, I will fill the gap and show how the universal properties of complex networks naturally emerge from the new mechanism, called geometric preferential attachment (GPA). I will explain how latent network geometry coupled with preferential attachment of nodes to this geometry induces power-law degree distribution, strong clustering, and community structure. Using the Internet data as an example, I will demonstrate that GPA generates networks that are similar to real networks.

Reading Data

More at:
- www.r-tutor.com/r-introduction/data-frame/data-import

- www.datacamp.com/community/tutorials/r-data-import-tutorial

and of course...
cran.r-project.org/doc/manuals/r-release/R-data.html

Example:

Download: anascombe-xl.R and anascombe.xlsx

```
## Based on http://blog.rstudio.org/2015/04/15/readxl-0-1-0/
library(readxl) # make sure to install.packages("readxl") if needed!
# Anascombe's data from the textbook. Each data set in it's own sheet:
excel_sheets("anascombe.xlsx")

## [1] "SLR"       "Quadratic" "Outlier1"  "Outlier2"

# Load sheet1
xydat=read_excel("anascombe.xlsx") # defaults to sheet=1
class(xydat)

## [1] "tbl_df"     "tbl"        "data.frame"

str(xydat)

## Classes 'tbl_df', 'tbl' and 'data.frame': 11 obs. of  2 variables:
##  $ x: num  10 8 13 9 11 14 6 4 12 7 ...
##  $ y: num  8.04 6.95 7.58 8.81 8.33 ...
```

Recall SLR Assumptions

By assuming the SLRM, you assume...

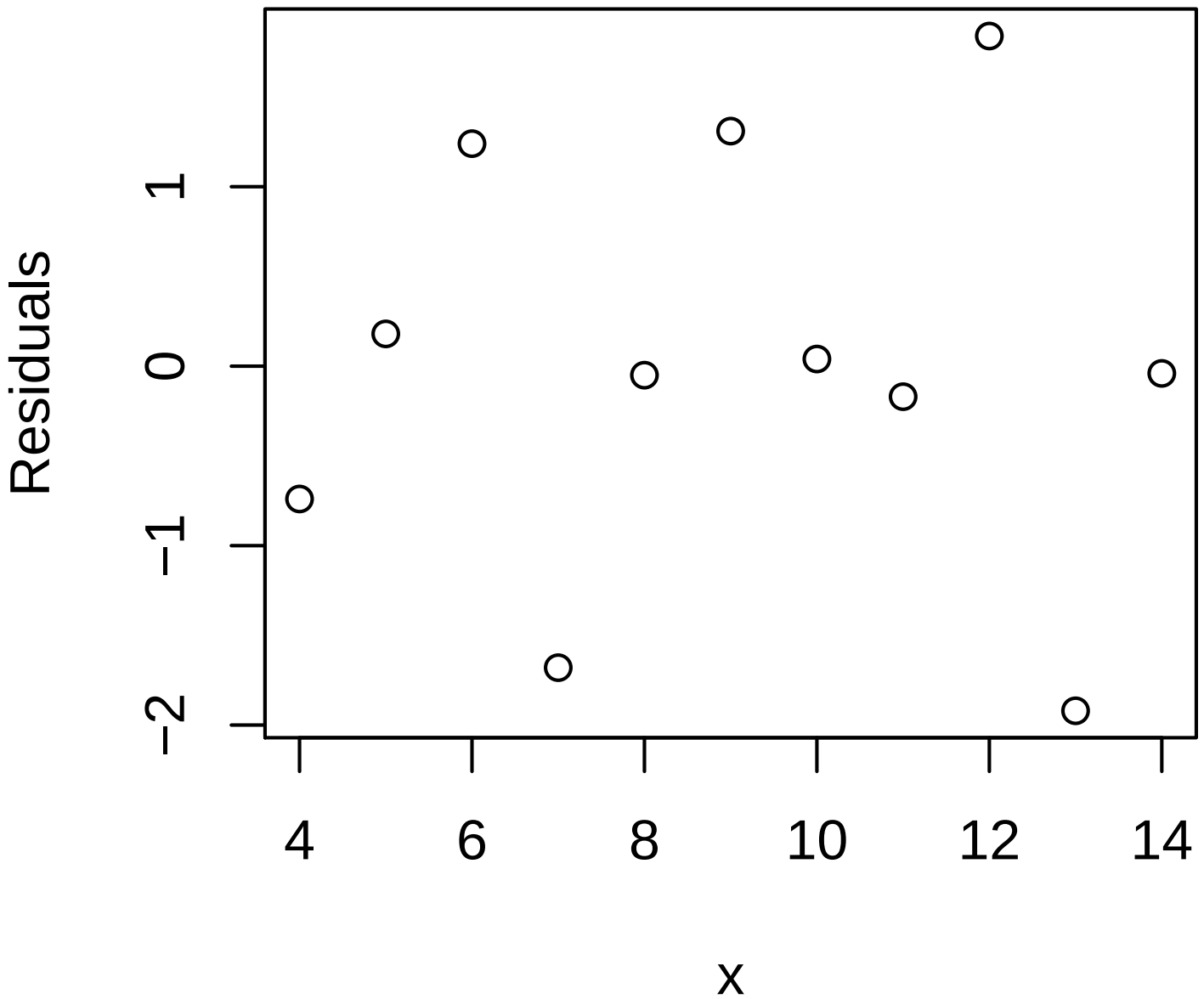
1. *All* data follow $Y|X = x_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma)$, hence $E(Y|X = x_i) = \beta_0 + \beta_1 x_i$
2. Normal errors: $e_i \sim N(0, \sigma)$
3. Independent errors e_i
4. $Var(Y|X = x_i) = Var(e_i) = \sigma^2$

Do Those Assumptions Hold?

Test using...

1. Residuals and Standardized Residuals
2. Leverage
3. Outliers
4. Correlations, etc...

Residuals $y_i - \hat{y}_i \approx e_i$



Leverage

We can quantify **leverage** with h_{ii}, where

$$\text{mean}(h_{ii}) = \frac{2}{n}$$

where a high leverage point is 2x that mean, i.e. $> 4/n$.