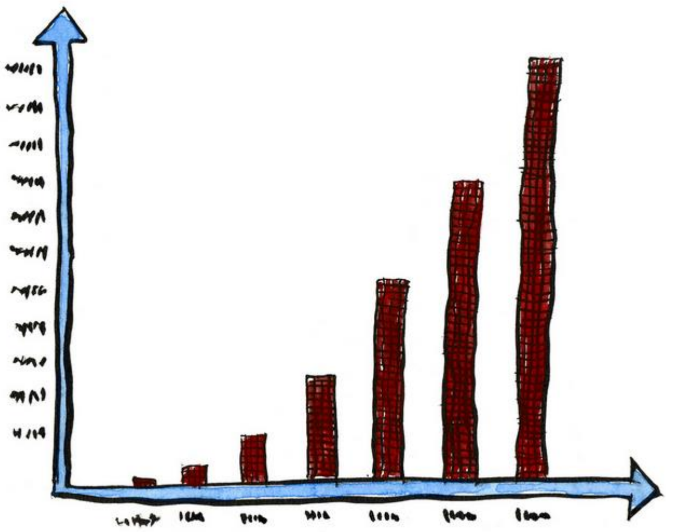
Introduction to R

R basics #2

Outline

- Univariate analysis
  - Testing for normality
- Bivariate analysis
  - Correlation
  - Regression
- Multivariate analysis
  - Correlation
  - Partial correlation
  - Regression



Dramatic increase in the amount of untrue statistics...

Univariate analysis

myDataset				
V1	V2	V3	…	Vn
0,1	4	0,8	2	Male
0,2	6	1,2	3	Female
0,8	8		6.3	Male
0,1	1	0,2	1	Male

- Looking 1 variable …
  - Histogram: single numerical variable
    - hist(V1) # histogram of V1 for all classes (Male and Female)
    - hist(V1[Vn==‘Male’]) # for Females only!
  - Density plot
    - plot(density(V1[Vn==‘Male’])) # empirical distribution
  - Boxplot: relationship between a numerical and categorical variable
    - boxplot(V1~Vn, myDataset, main = ‘…’)

Testing for normality

- Open R
- Plot the histogram of the length of the Petals for the versicolor
  - hist()
- Plot the density plot
  - lines(density())
- Does Petals.Length follow a normal distribution?
  - Using Density plots: compare visually the empirical density curve with the theoretical
    - increase the adj. parameter to smooth your density curve
    - Plot simultaneously the theoretical density curve that corresponds to the mean and sd of your data
      - Generate normally distributed data using rnorm(N data, mean, sd)
      - lines(density(), col="green")
  - Using qqplot: plot the theoretical vs the estimated quantiles
    - qqnorm(V1)
    - qqline(V1)
  - Normality tests
    - Shapiro.test(V1)
      - Null hypothesis: the distribution follows a normal distribution
      - Alternative: the distribution is different from a normal distribution
      - If p<0.1 we can accept the Alternative hypothesis therefore the distribution is significantly different from normal distribution
    - Kolmogorov-Smirnov test
      - ks.test(x, "pnorm", mean, sd)
      - Similar to Shapiro.test(V1) but mean and sd are different from the sample mean and sd
      - Test if the Sepal.Length follows a normal distribution of mean=8 and sd=1

Bivariate analysis

- Looking 2 variables …
  - Testing for normality: Kolmogorov-Smirnov
    - x= Sepal.Length for setosa, y= Sepal.Length for versicolor
    - Test if x and y follow the same distribution
    - Check if the distribution of x is stochastically smaller than that of y
      - Hint: Choose alternative = "greater“ or alternative = “less“
      - Support visually your answer
        - » plot the ecdf(x) and ecdf(y) in the same plot
  - Pairs? Did you forget already?
    - pairs()
  - Are my variables correlated?
    - cor.test(V1, V2)
      - Null hypothesis: my data are not correlated, correlation = 0
      - Alternative: Correlation is non-zero
      - Is there a significant correlation between Petal.Length and Petal.Width?

Bivariate - Regression

- Can we predict Petal.Width given the Petal.Length?
  - Make a scatterplot of the two variables
    - plot(V2~V1 , pch=20, col=as.numeric(Vn))
  - Fit a line
    - abline(V2~V1)
  - Use summary(lm(V2 ~ V1)) to
    - Write the equation of your model!
    - See the significance of your model
- Did my model fit correct the data?
  - Regression residuals should be approximately normally-distributed
    - residuals(lm(V2 ~ V1))
  - But I know how to check for normality now!
    - Choose and apply one method

Multivariate analysis

- cor.test() for pairs of variables
- Partial correlation
  - Is x and y really correlated or is there a hidden z that affects both?
    - Example:
      - z ~N[0,1]   k~N[0,2]
      - x=2z+5+0.2k
      - y=-3z+1
        - » Find the correlation between x and y
          - Use a qqplot and the cor.test
      - Is there any correlation between x and y after we perform correction??
  - Perform correction…
    - the residuals of linear regressions between the two variables should be uncorrelated
    - If they are still correlated then there is a true correlation between them
    - cor.test(residuals(lm(y ~ z)), residuals(lm(x ~ z)))

Multivariate- Regression

- Can we predict Sepal.Width given the Petal.Length?
  - summary(lm(z ~ x))
- Can we predict Sepal.Width given the Petal.Length and the Petal.Width?
  - summary(lm(z ~ x+y))
- Can we predict Sepal.Width given the Petal.Length, Petal.Width and the Sepal.Length?
  - summary(lm(z ~ y+x+p))
- Is the predictive equation significantly affected when adding predictors?
  - Adding Petal width increases R^2 by 0.2024-0.1282= 0.0244 => 2.4%
  - Adding Sepal length increases R^2 by 0.5142-0.2024= 0.3118 => 31%
- Which of the three models is the best predictor of Sepal.Width?
  - m1 = lm(z ~ x)
  - m2 = lm(z ~ y+x+p)
  - a <- anova(m1, m2)
- Plot the residuals of the two models in the same graph and check for normality