データサイエンス基礎講座
2015秋
【機械学習・実践編】

主催： 株式会社インプレス
企画：フューチャーブリッジパートナーズ株式会社

データサイエンス講座
第1回 統計の基礎

- データサイエンスとは？
- データサイエンスと機械学習
- Rの使い方と統計の基礎
- 統計の基礎とクロス集計
- 回帰分析

データサイエンスと機械学習

- データサイエンスと機械学習
- やりたいこと　→　機械に何かをいれると答えを出してくれる

- 機械学習（マシンラーニング）
  - 機械（マシン）が学習（ラーニング）する？
  - 機械自体は勝手に学習してくれない

- どう機械が学習するか？
  - データをもとに学習するモデルをつくる
  - 単にデータを入力すれば、勝手にモデルを作ってくれるわけではない

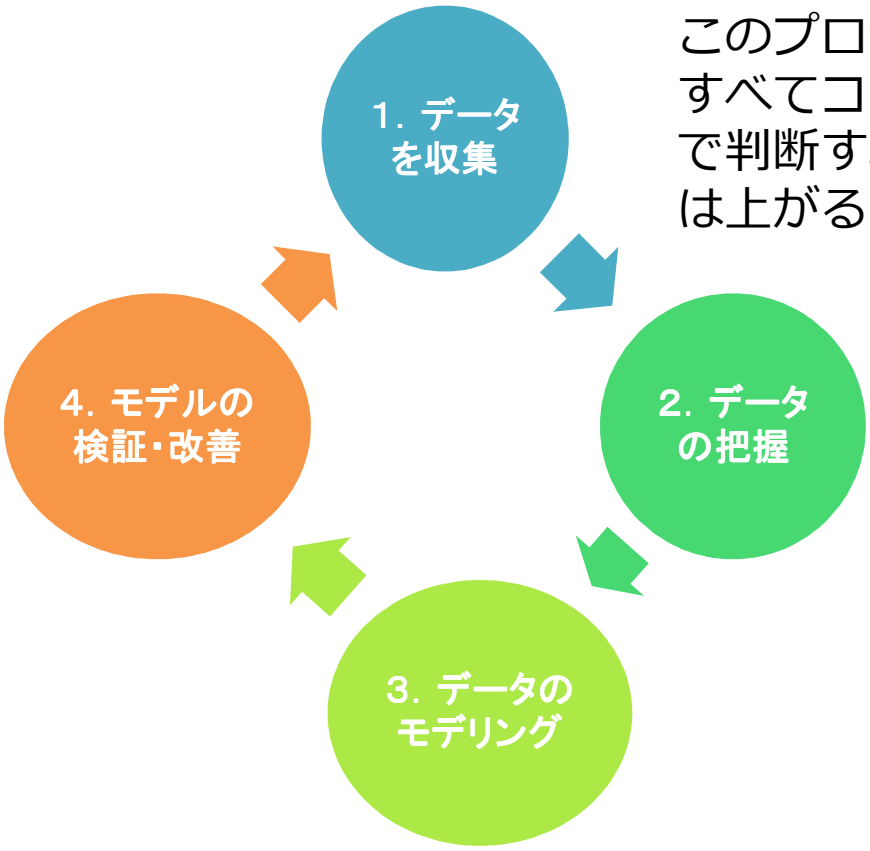
- 機械学習とは？
  - ある入力から機械がモデルに基づき自動的に識別・判定をすること

迷惑メール判定
入力＝メール
出力＝迷惑・正常メール判定

コンピュータ将棋
入力＝相手の一手
出力＝次の一手

データサイエンスと機械学習

- いま、なぜ、機械学習がアツい？
  - コンセプトはかなり昔から存在
- ネット等の普及で、「デジタル」データの収集が楽に
- くわえて、データを処理するIT基盤も進化
- 人手に頼ることなくデータサイエンスライフサイクルを実現　→ 精度の向上



このプロセスをすべてコンピュータで判断すれば、精度は上がる

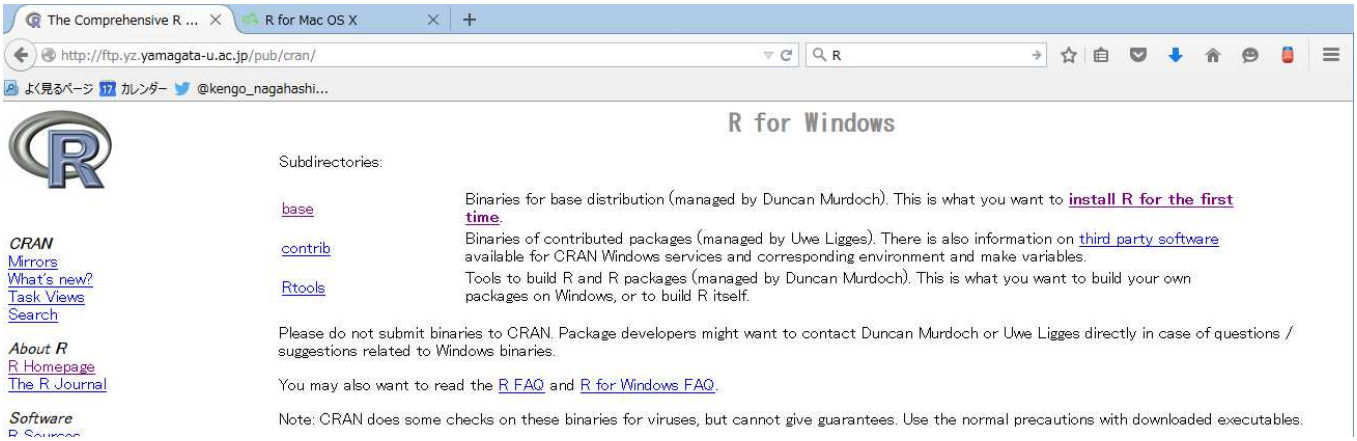
データサイエンスと機械学習

- データサイエンティストとして抑えておくべきポイント（＝今回の講座の範囲です）

1. クロス分析と確率分布
2. 回帰分析
3. 重回帰分析
4. クラスタリング分析
5. 主成分分析
6. 因子分析
7. アソシエーション分析
8. カーネル法とサポートベクターマシン
9. ロジスティクス回帰
10. アンサンブル学習

Rのインストール

- Windowsの場合
  - http://ftp.yz.yamagata-u.ac.jp/pub/cran/
  - からダウンロード
  - 最新バージョンは、R-3.2.2

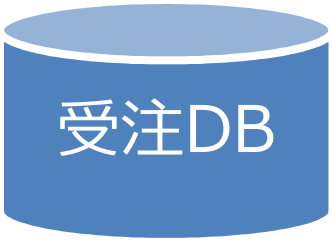


- Macの場合
  - http://ftp.yz.yamagata-u.ac.jp/pub/cran/bin/macosx/

クロス集計

- クロス集計とは？
  - データサイエンスのはじめの一歩
  - まずは、データの集計して、全体像をつかむ（クロス集計）

- 多くの企業のデータパターン



1.営業部署名
2.得意先名
3.仕入先
4.品名
5.数量
6.単価

- 多くの企業の場合、この受注DBをもとに売上予測、仕入予測をやりたいケースが多い

- まずは、データベースの全体像をつかむところから

回帰分析

◘回帰分析とは？

- 統計学の種類
  - 記述統計学　－　データを整理する（平均、分散など）　クロス集計もこの分野
  - 推測統計学　－　一部のデータ（サンプル）から全体（母集団）の状況を推測

- 回帰分析のアプローチ
  - サンプルをもとに、求めたい数値（目的変数）と入力変数（説明変数）＋係数をy=ax＋bでモデル化する
  - 例

身長	＝	体重	×	3	＋	20
目的変数		説明変数		係数		定数・切片

  - 体重のデータをもとに未知の身長をもとめる
  - 説明変数、係数、定数の決定方法　→　最小二乗法

回帰分析

◘回帰分析の流れ

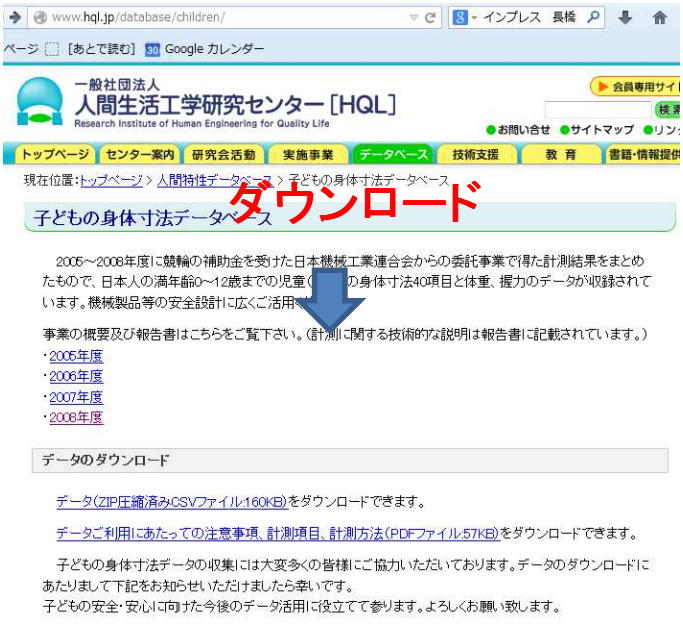
1. データの準備　＝　回帰分析の場合、目的変数に対して説明変数は一つ
2. すべてのデータが y (目的変数) ＝ ax (説明変数) + b　で説明できるとはかぎらないので、相関係数から説明変数と目的変数の相関性を確認する　→　この段階で、どの説明変数がフィットするか、仮説を立てて検証する
3. 相関系があれば、散布図を作成し、回帰直線を引くことが現実的か検討する
4. 目的変数、説明変数をもとに回帰分析を実施
5. 予測値を求めて、その残差を検討する
6. 回帰分析の結果をグラフ化して、外れ値などを検討する
7. 信頼区間と予測区間をもとめる

回帰分析

ステップ１．データの準備

- 子供の身体に関するデータをダウンロード
  - http://www.hql.jp/database/children/
  - データフォルダにあり

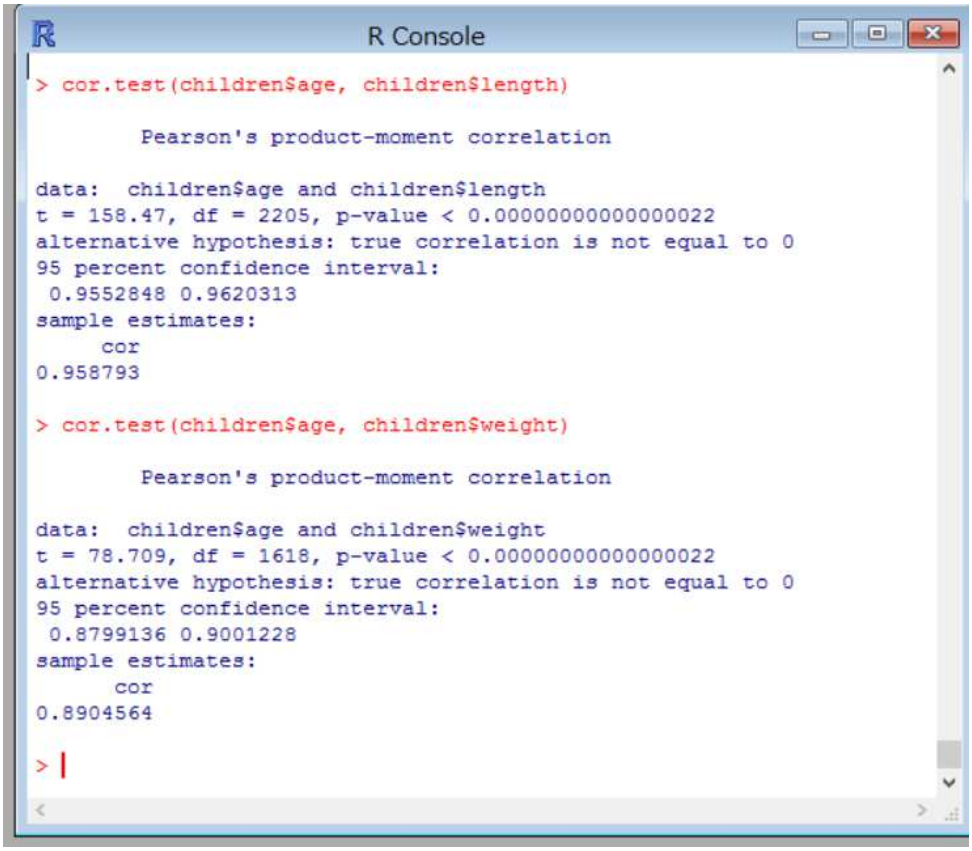
- 復習
  - children_data2005_08_130819.csvをもとに、年齢、身長、体重をそれぞれ、age,length,weightとして、別ファイルに保存
  - 保存したcsvファイルをＲに読み込みましょう
  - 読み込んだファイルについて、グラフ描画、平均、分散、相関係数を計算しましょう。
  - children = read.csv(“children.csv”,header=TRUE)



回帰分析

ステップ２．相関係数による説明変数の選定

- 目的変数　子供の年齢を説明変数から推定したい
- ２つの候補
  - 身長 children$length
  - 体重 children$weight
- cor.testでそれぞれの相関係数を求める

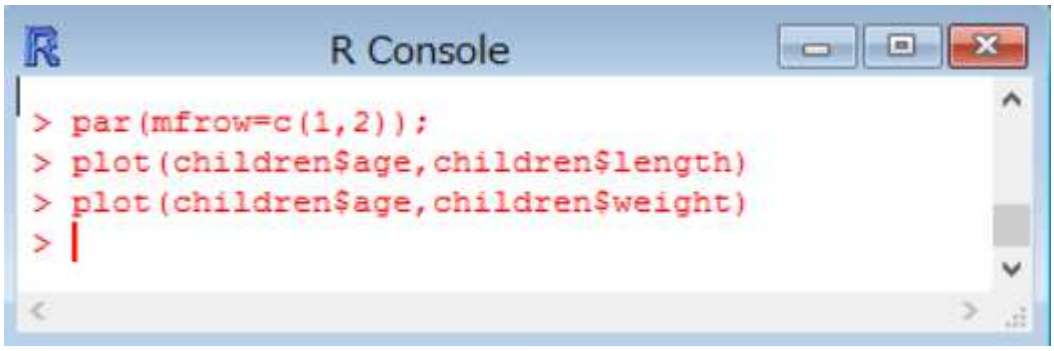


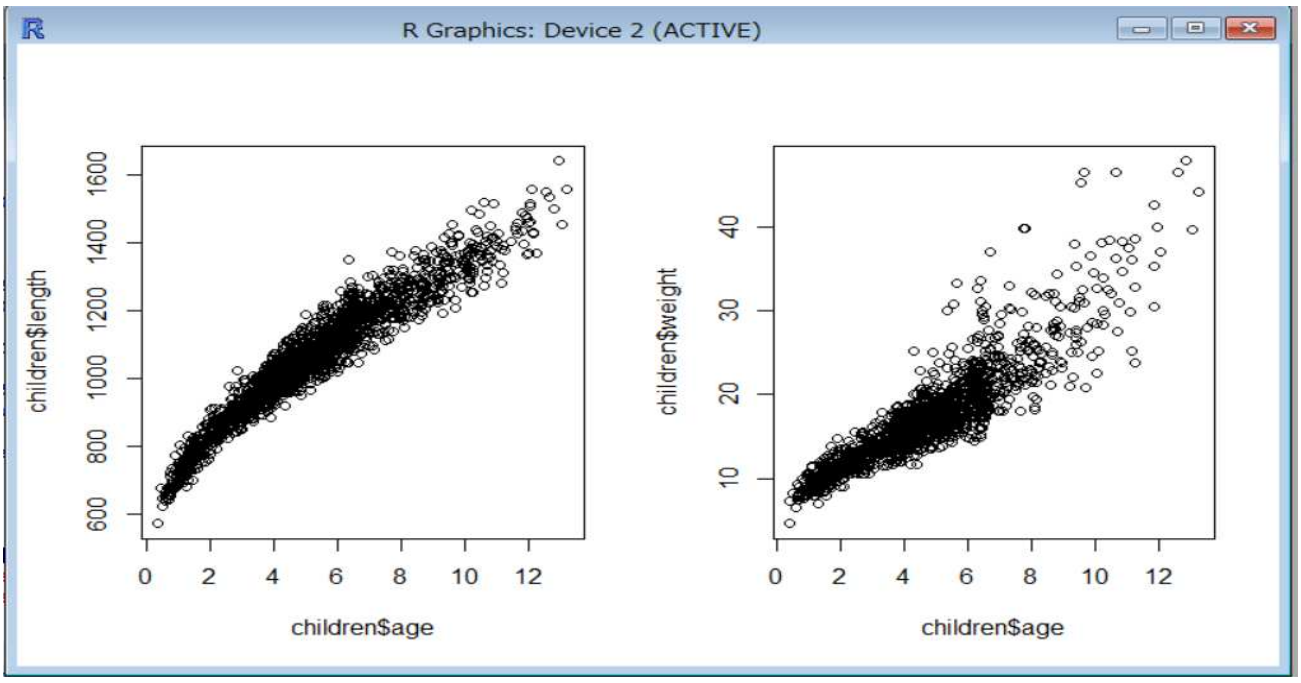
- 年齢と身長の相関係数　0.96
- 年齢と体重の相関係数　0.89
  - いずれも高い相関（R>0.8)があるものの、身長の方がよりよく説明できると仮定

回帰分析

ステップ３．散布図で確認

- par(mfrow=c(1,2));
- plot(children$age, children$length)
- plot(children$age, children$weight)





体重(weight)の場合、age 6までは相関が高いものの、それ以降の相関は低そう、一方、身長は8歳以降も相関あり

第1回のまとめ

- データサイエンス
  - 単に分析ではなく、Ｐ（仮説の設定）、Ｄ（分析）、Ｃ（検証）、Ａ（アクション）が大事
- 機械学習とデータサイエンス
  - 近年の技術進歩で、自動的にＰＤＣＡができるようになりつつあり、機械学習の重要度が増している
- Ｒの使い方
  - Ｒですべてできるわけではない。Excelが得意な分野、Pythonが得意な分野もある。ただし、パッケージはとても充実している
- クロス集計
  - データサイエンスの一歩はクロス集計から。確率分布を意識しながら、データを分析できる形にする
- 回帰分析
  - 単に回帰式ができればよいという話ではなく、モデル検証が重要

おススメ書籍

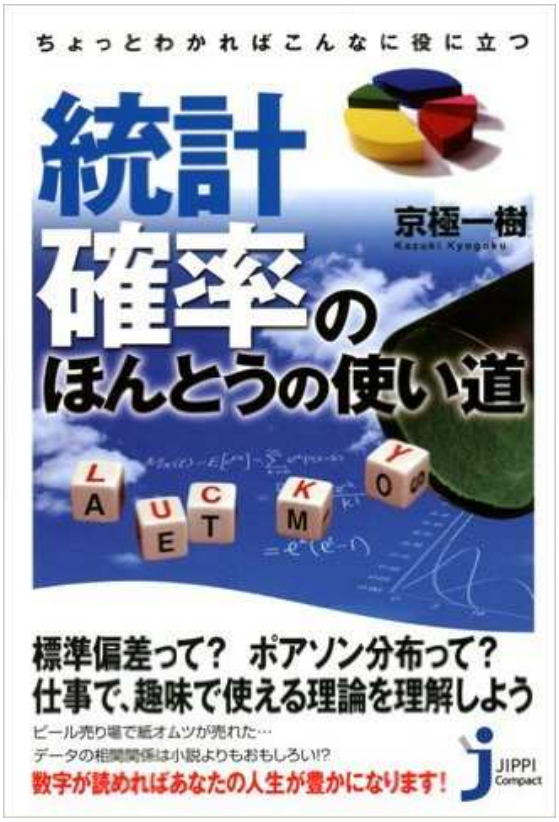
「マンガでわかる統計学」

2004年7月

高橋　信　トレンドプロ（著）

オーム社

　マンガならも統計学の初歩について広範にカバーしてあり、わかりやすい。全体像をつかむにはおすすめ。



「ちょっとわかればこんなに役に立つ 統計・確率のほんとうの使い道 (じっぴコンパクト新書)」

2012年2月

京極　一樹（著）

実業之日本社

　統計のトピックについて2ページで一つ一つ解説。ポアソン過程など比較的深いところまで言及。