

# 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



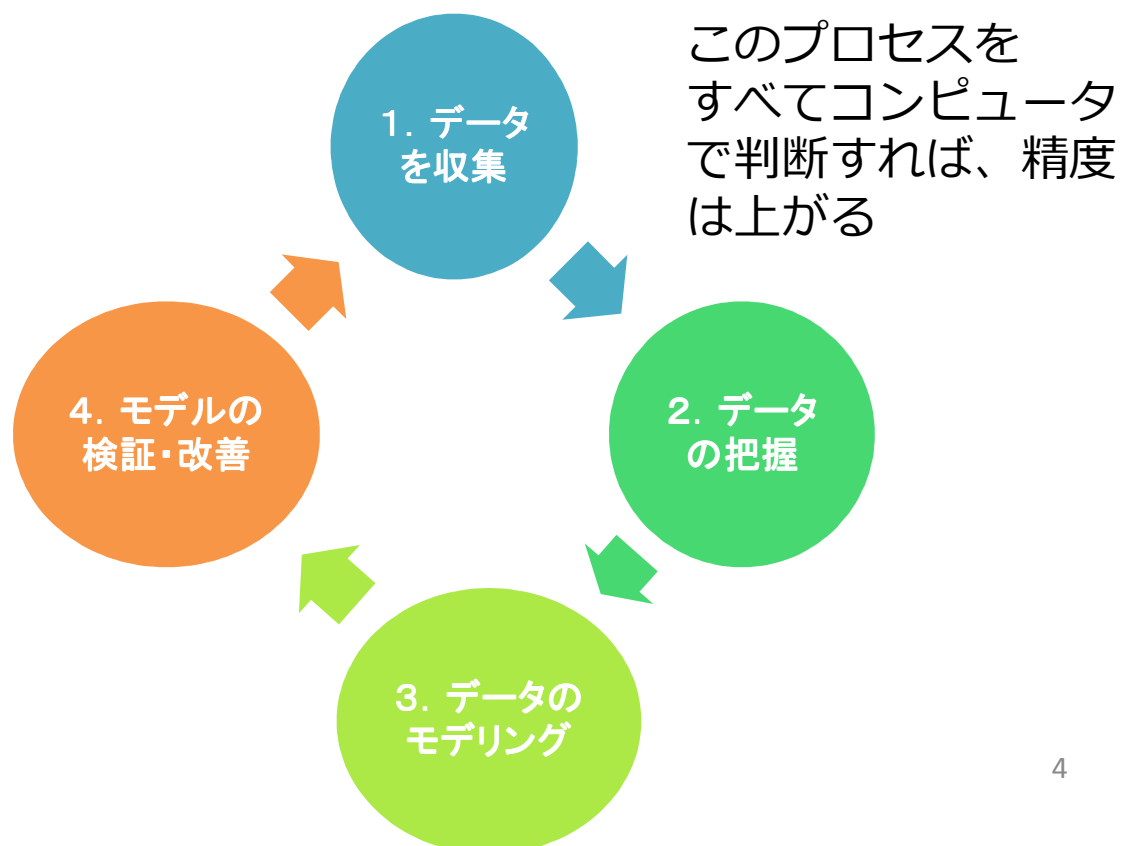
迷惑メール判定  
入力 = メール  
出力 = 迷惑・正常メール判定



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**



# データサイエンスと機械学習

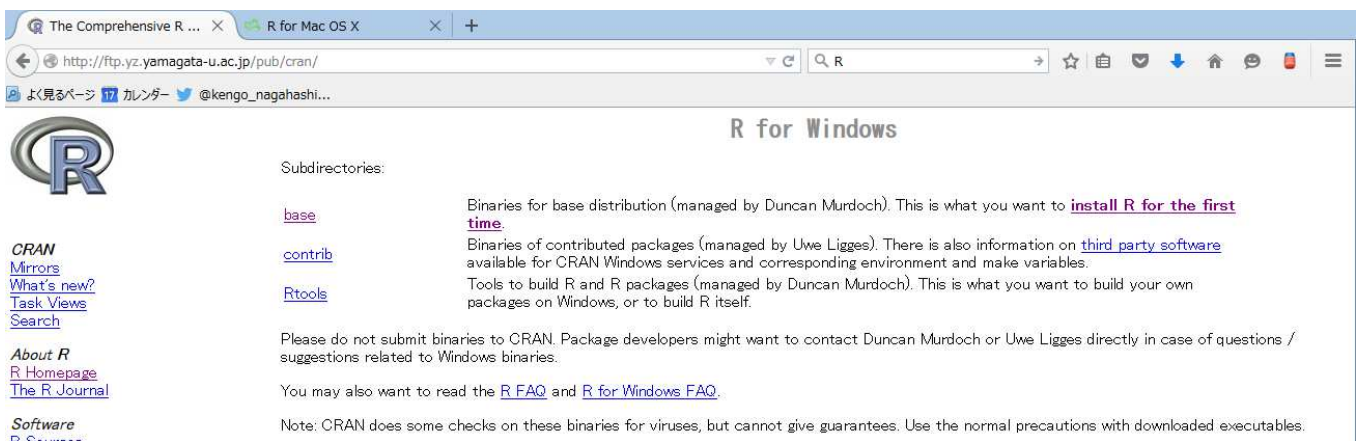
□ データサイエンティストとして抑えておくべきポイント（=今回の講座の範囲です）

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)

www.hql.jp/database/children/

一般社団法人 人間生活工学研究センター [HQL]

子どもの身体寸法データベース

**ダウンロード**

データのダウンロード

データ(ZIP圧縮済みCSVファイル160KB)をダウンロードできます。

データご利用にあたっての注意事項、計測項目、計測方法(PDFファイル57KB)をダウンロードできます。

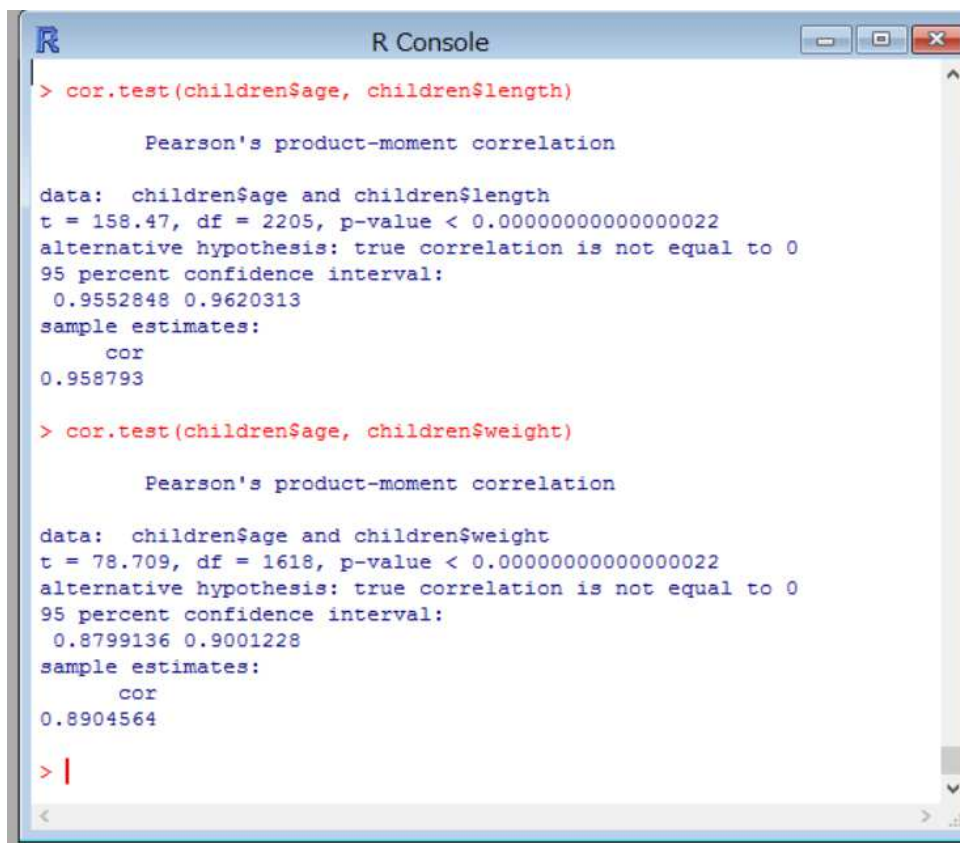
children\_data2005\_08\_130819 - Microsoft Excel

番号	計測年	計測月	性別	上半身着丈	下半身着丈	足部着衣	年齢	身長(次部)	身長(アト)	体重	肩峰高	椅骨点
1							(年)	(mm)	(mm)	(kg)	(mm)	(mm)
2	1	2005年	11月	男の子	(不明)	(不明)	(不明)	2.41	1022	20.6	784	
3	4	2005年	11月	男の子	(不明)	(不明)	(不明)	2.11	835	11.8		
4	3	2005年	11月	男の子	(不明)	(不明)	(不明)	3.56	958	13.6	741	
5	6	4	2005年	11月	女の子	(不明)	(不明)	3.2	905	12.4	698	
6	7	5	2005年	11月	男の子	(不明)	(不明)	3.2	922	16.6	798	
7	8	6	2005年	11月	男の子	(不明)	(不明)	3.2	1025	16.2	782	
8	9	7	2005年	11月	男の子	(不明)	(不明)	4.2	897	14.6	758	
9	10	8	2005年	11月	男の子	(不明)	(不明)	3.99	1033	16	790	
10	11	9	2005年	11月	男の子	(不明)	(不明)	3.74	1035	17.6	763	
11	12	10	2005年	11月	男の子	(不明)	(不明)	3.74	1057	18.5	797	
12	13	11	2005年	11月	女の子	(不明)	(不明)	3.74	894	14.3	755	
13	14	12	2005年	11月	女の子	(不明)	(不明)	3.74	917	22.9	855	
14	15	13	2005年	11月	男の子	(不明)	(不明)	3.96	1167	16.3	805	
15	16	14	2005年	11月	男の子	(不明)	(不明)	3.74	1129	15.9	793	
16	17	15	2005年	11月	男の子	(不明)	(不明)	3.92	871	12.6	668	
17	18	16	2005年	11月	男の子	(不明)	(不明)	2.37	851	13.6	648	
18	19	17	2005年	11月	男の子	(不明)	(不明)	2.17	679	12.8	676	
19	20	18	2005年	11月	男の子	(不明)	(不明)	2.56	681	12.2	657	
20	21	19	2005年	11月	男の子	(不明)	(不明)	2.38	912	13.2	707	
21	22	20	2005年	11月	男の子	(不明)	(不明)	2.38	849	10.8	643	
22	23	21	2005年	11月	女の子	(不明)	(不明)	1.82	837	11.4	628	
23	24	22	2005年	11月	女の子	(不明)	(不明)	1.91	830	11.2	627	
24	25	23	2005年	11月	男の子	(不明)	(不明)	2.27	841	13.2	640	
25	26	24	2005年	11月	男の子	(不明)	(不明)	2.54	857	12.2	649	
26	27	25	2005年	11月	男の子	(不明)	(不明)	2.32	840	10.6	650	
27	28	26	2005年	11月	男の子	(不明)	(不明)	1.41	810	11.8		
28	29	27	2005年	11月	男の子	(不明)	(不明)	1.39	795	10.4	571	
29	30	28	2005年	11月	女の子	(不明)	(不明)	1.32	756	9.3		
30	31	29	2005年	11月	女の子	(不明)	(不明)	1.5	784	9.8		568
31	32	30	2005年	11月	男の子	(不明)	(不明)	0.94	722	10.4		
32	33	31	2005年	11月	男の子	(不明)	(不明)	0.86	715	8.7		

# 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



```
R Console
> cor.test(children$age, children$length)

Pearson's product-moment correlation

data: children$age and children$length
t = 158.47, df = 2205, p-value < 0.000000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9552848 0.9620313
sample estimates:
      cor 
0.958793

> cor.test(children$age, children$weight)

Pearson's product-moment correlation

data: children$age and children$weight
t = 78.709, df = 1618, p-value < 0.000000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8799136 0.9001228
sample estimates:
      cor 
0.8904564

> |
```

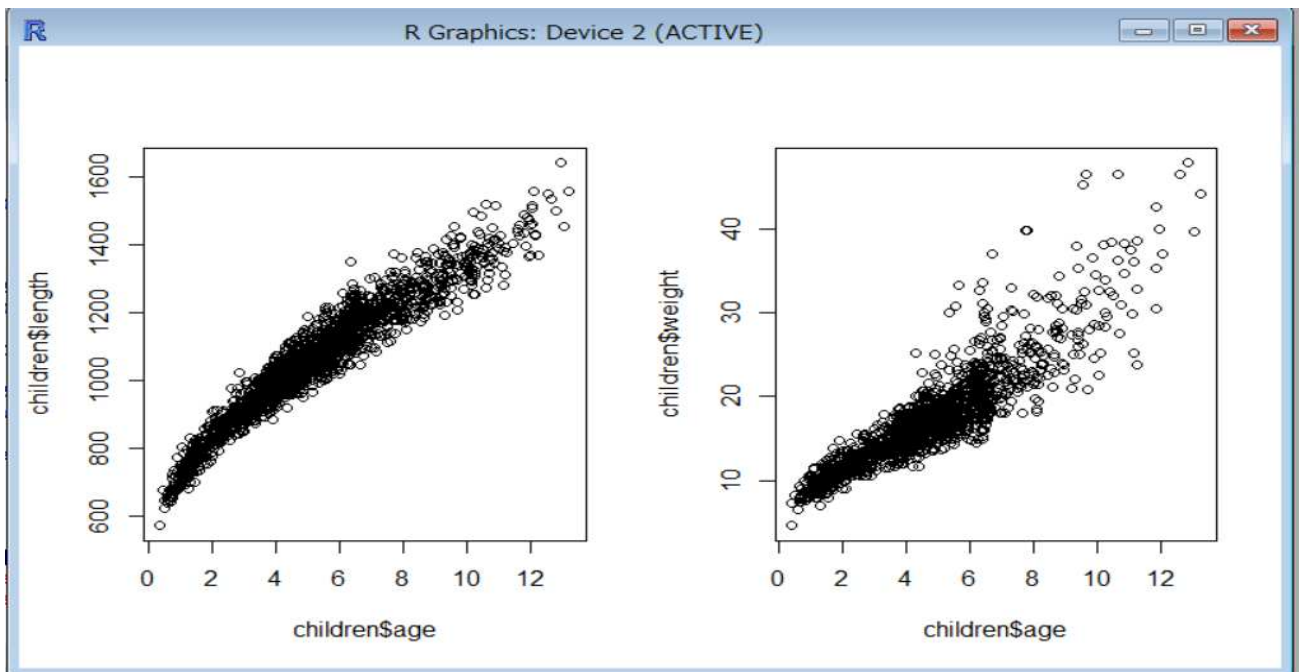
- 年齢と身長の相関係数 0.96
- 年齢と体重の相関係数 0.89
  - いずれも高い相関 ( $R > 0.8$ )があるものの、身長の方がよりよく説明できると仮定

# 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)`

```
R Console
> 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P D C A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步はクロス集計から。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

# おすすめ書籍



## 「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

マンガなが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。