テキストデータの統計科学入門

15章　テキストの時系列分析

12NM712L 小幡智裕

概要

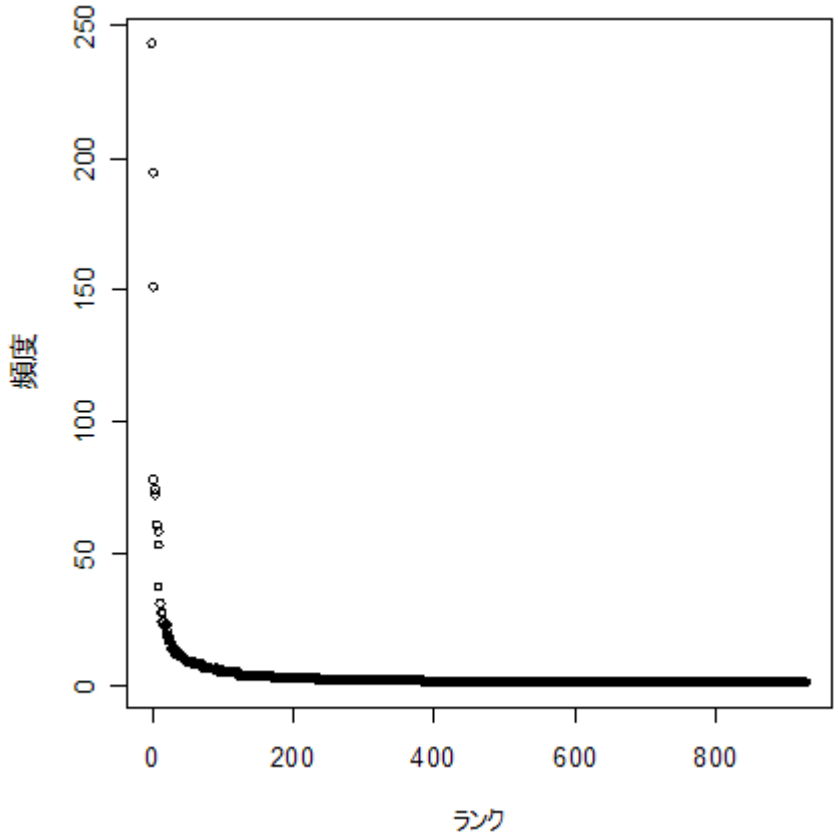
本章では、テキストにおける何らかの項目の推移の傾向を分析する基礎である回帰分析を説明し、テキストに用いられている要素の経時的変化を分析する方法について説明する

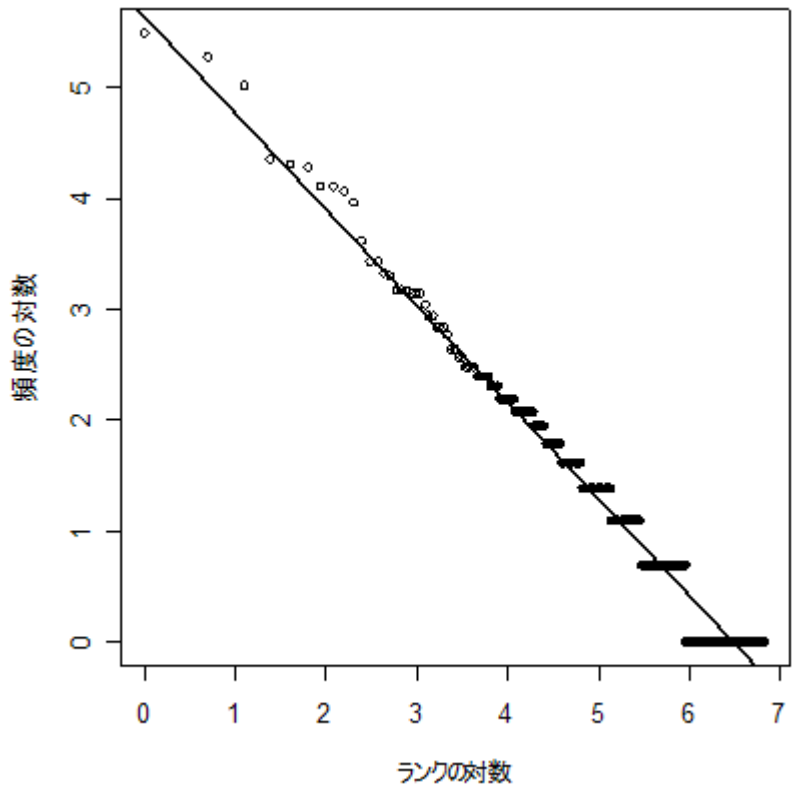
回帰分析

- 福田元総理の所信表明演説の単語のランクと頻度

単語・記号	r	f	x=log(r)	y=log(f)
の_助詞	1	143	0.0000	5.4931
を_助詞	2	194	0.6931	5.2679
に_助詞	3	151	1.0986	5.0173
ます_助動詞	4	78	1.3863	4.3567
し_動詞	5	74	1.6094	4.3041
て_助詞	6	72	1.7918	4.2767
…	…	…	…	…
野党_名詞	384	2	5.9532	0.6931
男性_名詞	385	1	5.9506	0
…	…	…	…	…
科学_名詞	928	1	6.8330	0

回帰分析(続き)


頻度
250
200
150
100
50
0
0
200
400
600
800
ランク


頻度の対数
5
4
3
2
1
0
0
1
2
3
4
5
6
7
ランクの対数

回帰分析とは

- 回帰分析
  - データの傾向をモデルにあてはめる統計分析
  - 未知の真のモデルに近似する統計モデルをデータから求めるのが目的
  - 直線を回帰直線、直線の式を回帰式とよぶ

2つの線形関係をモデル化する一般の回帰式は

$$\hat{y} = a_0 + a_1 x$$

で表される

回帰係数と最小2乗法

- 線形単回帰モデルの例

  実測データ $x_i, y_i (i = 1,2,3,\ldots,n)$ があるとき、

  真のモデルが

$$y_i = a + bx_i + \varepsilon_i$$

  であるとし、近似的な回帰式は

$$\hat{y}_i = a + bx_i$$

  であるとする

回帰係数と最小2乗法（続き）

真のモデルと回帰式との差の2乗和

$$Se = \sum (y_i - a - bx_i)^2$$

を最小とするa,bを求めると真のモデルに近似する回帰式が求まる

→このような考え方を最小2乗法という

未知数a,bを変数とした上式の連立微分方程式を解くことでa,bを求めることができる。

最小2乗法の計算

$$\frac{\partial S_e}{\partial a} = -2\sum (y - a - bx_i) = 0$$

$$\frac{\partial S_e}{\partial b} = -2\sum x_i (y - a - bx_i) = 0$$

この連立方程式を解くと

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, a = \bar{y} - b\bar{x}$$

ジップの法則と回帰式

ランクrと頻度fの対数値の散布図は線形関係であるx,yの平均はそれぞれ、5.8377,0.5732、xの分散は0.9695、xとyの共分散は-0.8388である。したがって、

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{-0.8388}{0.9695} = -0.8652$$

$$a = \bar{y} - b\bar{x} = 0.5732 - (-0.8652 \times 5.8377) = 5.6240$$

となり、この回帰係数を用いた回帰式は

$$y = 5.6240 - 0.8652x$$

である

ジップの法則と回帰式（続き）

この回帰式は対数表示であるので、回帰式

$$\log(f) = 5.6240 - 0.8652\log(r)$$

を指数変換すると

$$f = e^{5.6240 - 0.8652\log(r)} = \exp\{5.6240 - 0.8652\log(r)\}$$

になり、これがランクrと頻度fとの関係式となり、広義のジップの法則となる

散布図作成・回帰分析のRコマンド

- Rコマンド

```
FUKUDA<read.csv("http://mjin.doshisha.ac.jp/iwanami/data/FUKUDA.csv",head=T,row.names=1)
x<-log(1:nrow(FUKUDA))
y<-log(FUKUDA[,1])
plot(x,y,xlab="ランクの対数",ylab="頻度の対数")
(fukuda.lm<-lm(y~x))
abline(fukuda.lm,lw=2,col="red")
```

- 実行結果

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     5.6235      -0.8651
```

回帰結果のRコマンド

- Rコマンド

```
summary(fukuda.lm)
```

- 実行結果

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.46648 -0.10425  0.03094  0.13512  0.34418

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.623544   0.035782   157.2   <2e-16 ***
x           -0.865133   0.006044  -143.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1812 on 926 degrees of freedom
Multiple R-squared: 0.9568,     Adjusted R-squared: 0.9567
F-statistic: 2.049e+04 on 1 and 926 DF,  p-value: < 2.2e-16
```

重回帰分析

重回帰分析では観測データを次の式で表す

$$y_i = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p + \varepsilon_i$$

回帰式は次に示すような近似式である

$$\hat{y}_i = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p$$

回帰式の係数は、単回帰同様に最小2乗法で求めることができる

線形重回帰の回帰係数

- 実測値と係数を行列で示す

$$A=\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix}, Y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, E=\begin{bmatrix} \varepsilon_1 & 0 & \cdots & 0 \\ 0 & \varepsilon_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varepsilon_n \end{bmatrix}$$

- 真のモデル

$$Y = XA + E$$

- 回帰式

$$\hat{Y} = XA$$

- 最小2乗法による係数の推測値

$$A = (X^t X)^{-1} X^t Y$$

変数の選択

収集したデータの中には、回帰式にまったく役に立たない変数が含まれている場合が少なくない

変数を選択しシンプルな回帰式を構築するがよい

回帰式に用いる変数の寄与度にしたがい、変数を選択する

変数の選択方法

- 変数増加法

  寄与度が最も高い変数から順次に変数を増やしながら最善と評価される回帰式を構築する
- 変数減少法

  寄与度が最も低い変数から順次に削除しながら最善と評価される回帰式を構築する
- 変数増減法

  変数の削除と追加を繰り返しながら最善と評価される回帰式を構築する

非線形重回帰

- 非線形関数を用いる回帰分析
  多項式、指数関数、ロジスティック関数などを用いる

- 一般化線形モデル、一般化法モデル
  説明変数があまり多くない場合に用いる

- 機械学習法
  ニューラルネットワーク法、サポートベクターマシン法、集団学習といった手法

テキストの時系列分析

時間経過の順に計測したデータを時系列データと言い、
そのデータを分析するプロセスを時系列分析とよぶ.
時間とともに何がどのように変化しているかを分析する

- ニュース記事などの時系列テキストの場合、話題の推移などを分析する

- 文学作品の場合、執筆時期の推定をおこなう
  →本章では文章の執筆時期の推定に関する問題をテキストの時系列分析の例とする

執筆時期の推定

- 芥川龍之介の作品309編について形態素解析を行い、表に示す39項目について出現頻度（相対度数に変換）を集計してデータとして用いて分析する

表：39項目のリスト

の_格助詞	は_係助詞	を_各助詞	に_格助詞	て_接続助詞
と_格助詞	も_係助詞	が_格助詞	の_準体助詞	か_係助詞
ば_接続助詞	ながら_接続助詞	や_係助詞	で_接続助詞	で_格助詞
と_接続助詞	へ_格助詞	より_格助詞	だけ_副助詞	に_接続助詞
など_副助詞	まで_副助詞	から_格助詞	ばかり_副助詞	か_格助詞
が_接続助詞	ども_接続助詞	から_接続助詞	ん_準体助詞	ほど_副助詞
のみ_副助詞	しも_副助詞	とも_接続助詞	さえ_副助詞	さへ_副助詞
ん_格助詞	ど_接続助詞	その他	、_読点	

データの考察

309行×39列のデータの分散共分散行列について主成分分析を行った

早期の作品は主成分得点が第一主成分の正の方向に、晩期の作品は負の方向に配置されている傾向があった

第一主成分の正の方向には格助詞「が」と接続助詞「て」、負の方向には係助詞「は」と格助詞「の」「に」「を」が大きく寄与している

データの考察（続き）

- 係助詞「は」の使用率の図と格助詞「が」の使用率の図と相関関係のネットワークマップ

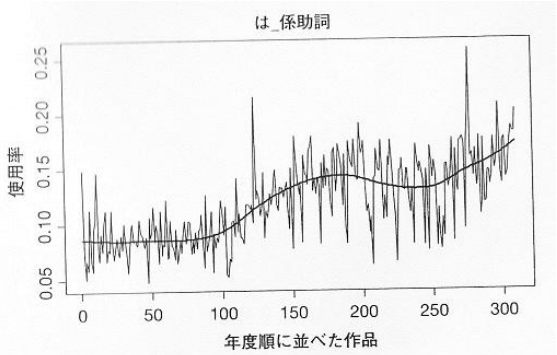


図 15.5　係助詞「は」の使用率

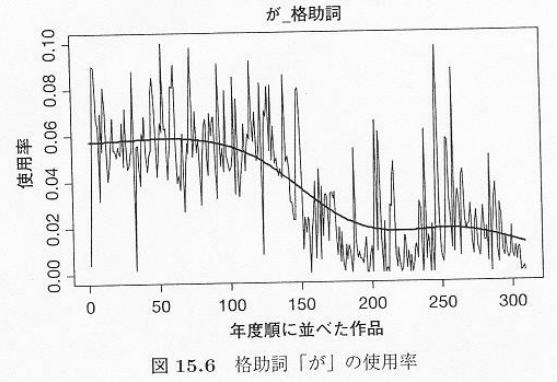


図 15.6　格助詞「が」の使用率

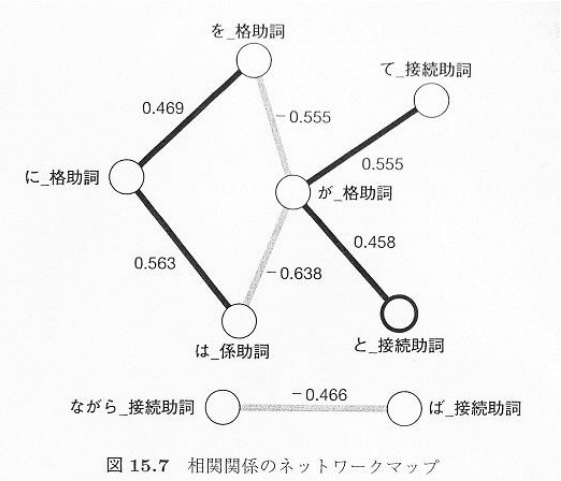


図 15.7　相関関係のネットワークマップ

線形回帰分析

データの考察より、文章に使用された要素を用いて執筆年代を推定することが可能であると見込み、回帰分析を試みた

- 39項目を独立変数とし、初出の年・月を目的変数とした
- 変数の選択にはAICにもとづいたステップワイズ法を用いた
- 回帰分析の結果

  独立変数：39→22

  線形回帰モデル：$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_{22} x_{22}$

  重相関係数：0.6284　　補正された重相関係数：0.5998

  残差の四分位数：-15.3249(Min), -1.1094(1Q), -0.1099(Median)
  1.186(3Q), 8.8921(Max)

明治時代の作品と1000文字以下の文章を取り除いた250編を用いて再び回帰分析を試みる

線形回帰分析（続き）

- Rコマンド

```
akuta<-read.csv
(http://mjin.doshisha.ac.jp/iwanami/data/akuta250.csv,head=T,row.names=1)
akuta.lm<-lm(y~.,data=akuta)
akuta.lm2<-step(akuta.lm,trace=0)
```

- 実行結果

```
Residuals:
    Min      1Q  Median      3Q     Max
 -6.3555 -0.9244 -0.0549  0.9166  7.6609


Residual standard error: 1.822 on 232 degrees of freedom
Multiple R-squared: 0.7685,     Adjusted R-squared: 0.7516
F-statistic: 45.31 on 17 and 232 DF,  p-value: < 2.2e-16
```

線形回帰分析（続き）

```
Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)   1917.649      1.772 1081.959  < 2e-16 ***
読点          -10.509      3.111   -3.379 0.000855 ***
は_係助詞      32.459      5.864    5.536 8.36e-08 ***
に_格助詞      28.731      8.057    3.566 0.000440 ***
が_格助詞     -50.212      7.561   -6.641 2.19e-10 ***
の_準体助詞    -31.446     12.596   -2.497 0.013235 *
か_係助詞      58.632     13.769    4.258 3.00e-05 ***

ば_接続助詞    32.331     17.019    1.900 0.058716 .
や_係助詞      42.364     15.465    2.739 0.006633 **
で_接続助詞    58.477     26.750    2.186 0.029813 *
へ_格助詞      72.206     33.125    2.180 0.030277 *
まで_副助詞   -119.999     61.923   -1.938 0.053854 .
ばかり_副助詞 -230.347     60.719   -3.794 0.000189 ***
が_接続助詞    122.670     41.236    2.975 0.003242 **
から_接続助詞  132.197     66.347    1.993 0.047488 *
とも_接続助詞 -283.806    108.432   -2.617 0.009444 **
さへ_副助詞   -255.127     95.339   -2.676 0.007982 **
ど_接続助詞   -175.407     85.787   -2.045 0.042015 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

テストデータを用いた回帰式の評価

250編の作品の中から10編をランダムサンプリングしたデータをテスト用とし、残りを学習データとして回帰式を作成し、その回帰式によるテストデータの予測結果と残差を求めた

- Rコマンド

```
set.seed(100)
sam<-sample(1:250,10)
test<-akuta[sam,]
train<-akuta[-sam,]
te.lm<-lm(y~.,data=train)
te.lm2<-step(te.lm,trace = 0)
te.pr<-predict(te.lm2,test[,-40])
temp<-round(cbind(test$y,te.pr,test$y-te.pr),1)
colnames(temp)<-c("初出年","予測値","残差") temp
```

テストデータを用いた回帰式の評価（続き）

- 実行結果

	初出年	予測値	残差
T08-11c江口渙氏の事	1919.9	1920.7	-0.8
T08-04bきりしとほろ上人	1919.3	1920.3	-1.0
T12-01a漱石山房の冬	1923.1	1923.4	-0.3
T05-04b虱	1916.3	1918.0	-1.6
T10-12c将軍	1922.0	1921.3	0.7
T11-01cLOSCAPR	1922.1	1922.1	0.0
T14-08d死後	1925.7	1925.9	-0.2
T09-04d沼	1920.3	1919.6	0.7
T11-09百合	1922.8	1926.3	-3.5
T06-10戯作三昧	1917.8	1919.4	-1.5

ランダムサンプリングのデータであり、偶然の可能性もあるためこのような学習とテストは繰り返しておこない、その結果を用いて評価することが必要

非線形回帰分析

本章では決定木（回帰木）とランダムフォレスト回帰法を用いる.

- 決定木のRコマンド

```
te.tr<-mvpart(y~.,data=train)
te.pr<-predict(te.tr,test)
temp<-round(cbind(test$y,te.pr,test$y-te.pr),1)
colnames(temp)<-c("初出年","予測値","残差")
temp
```

- 決定木の実行結果

	初出年	予測値	残差
T08-11c江口渙氏の事	1919.9	1919.5	0.5
T08-04bきりしとほろ上人	1919.3	1919.5	-0.1
T12-01a漱石山房の冬	1923.1	1925.5	-2.4
T05-04b虱	1916.3	1919.5	-3.1
T10-12c将軍	1922.0	1919.5	2.5
T11-01cLOSCAPR	1922.1	1919.5	2.6
T14-08d死後	1925.7	1925.5	0.2
T09-04d沼	1920.3	1919.5	0.9
T11-09百合	1922.8	1925.5	-2.7
T06-10戯作三昧	1917.8	1919.5	-1.6

非線形回帰（続き）

- ランダムフォレスト法のRコマンド

```
te.rf<-randomForest(y~.,data=train)
te.pr<-predict(te.rf,test[,-40])
temp<-round(cbind(test$y,te.pr,test$y-te.pr),1)
rownames(temp)<-rownames(test)
colnames(temp)<-c("初出年","予測値","残差")
temp
```

- ランダムフォレスト法の実行結果

	初出年	予測値	残差
T08-11c江口渙氏の事	1919.9	1922.3	-2.4
T08-04bきりしとほろ上人	1919.3	1919.9	-0.6
T12-01a漱石山房の冬	1923.1	1925.5	-2.4
T05-04b虱	1916.3	1917.8	-1.5
T10-12c将軍	1922.0	1921.7	0.3
T11-01cLOSCAPR	1922.1	1921.1	1.0
T14-08d死後	1925.7	1925.5	0.2
T09-04d沼	1920.3	1920.2	0.1
T11-09百合	1922.8	1924.8	-2.0
T06-10戯作三昧	1917.8	1919.3	-1.4

予測結果の比較分析

線形回帰(LM)、回帰木(CART)、サポートベクターマシン(SVM)、ランダムフォレスト(RF)を用いた執筆時期の推定の結果を比較

表：データの予測残差の要約

	LM	CART	SVM	RF
最小値	-8.7550	-11.6500	-8.1420	-8.6140
第1四分位数	-1.1290	-1.2250	-0.9460	-0.8889
中央値	-0.0644	0.0378	0.0411	-0.0394
第3四分位数	-0.0257	-0.0030	-0.0293	0.0047
最大値	1.1020	1.1870	1.0380	0.9169
絶対値の平均	1.5253	1.7507	1.4337	1.3509
絶対値の標準偏差	1.5176	1.8112	1.4119	1.3967

ロジスティック回帰

テキストを時系列に並べた時、考察する語句の使用率の増加傾向がS字形状をなす場合がある.このようなデータをモデリングする方法として、ロジスティック曲線に当てはめる方法がある.

- ロジスティック曲線の関数式

$$y = \frac{a}{1 + be^{cx}}$$

a,b,cが曲線のパラメータとなっている

Rにはロジスティック回帰を行う関数glmがある.ロジスティック回帰は2群非線形判別分析の手法としても用いることが可能である