

データ解析  
Rによる多変量解析入門  
(6) 回帰分析(III)

信頼領域

lec20021114 下平英寿 shimo@is.titech.ac.jp

1

回帰係数の推定量と残差の分布

- モデル(p)

$$y = X\beta + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$$

$$\beta = [\beta_0, \dots, \beta_p]'$$

$$\hat{\beta} = (X'X)^{-1}X'y, \quad \hat{y} = X\hat{\beta}, \quad e = y - \hat{y}$$

- 以下の二つの確率変数は互いに独立

$$\frac{\hat{\beta} - \beta}{\sigma} \sim N_{p+1}(0, (X'X)^{-1}), \quad \frac{\|e\|^2}{\sigma^2} \sim \chi_{n-p-1}^2$$

2

3

回帰係数の線形変換の分布

$$\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X'X)^{-1})$$

任意の  $a \times (p+1)$  行列  $A$  を使って

$$\hat{w} = A\hat{\beta}$$

の平均と分散は

$$E(\hat{w}) = E(A\hat{\beta}) = AE(\hat{\beta}) = A\beta$$

$$V(\hat{w}) = E\{(\hat{w} - E(\hat{w}))(\hat{w} - E(\hat{w}))'\}$$

$$= E\{A(\hat{\beta} - \beta)(\hat{\beta} - \beta)'A'\} = \sigma^2 A(X'X)^{-1}A'$$

とくに  $X = QR$  と  $QR$  分解して  $A = R, \gamma = R\beta$  とおくと

$$R(X'X)^{-1}R' = R(R'R)^{-1}R' = RR^{-1}(R')^{-1}R' = I_{p+1}$$

$$\frac{\hat{\gamma} - \gamma}{\sigma} = \frac{R(\hat{\beta} - \beta)}{\sigma} \sim N_{p+1}(0, I_{p+1})$$

4

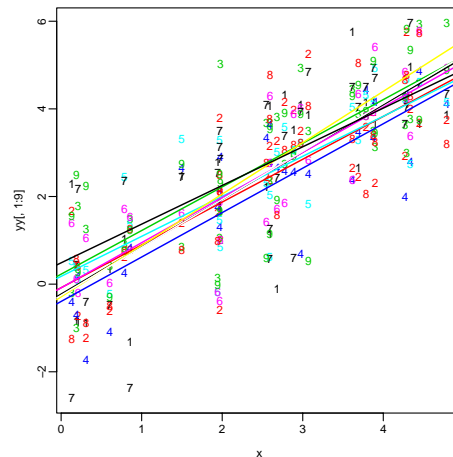
数値例

```
> ## シミュレーションデータの準備
> n <- 30 # データ数 n = 30
> B <- 10000 # シミュレーション数 B = 10000
> x <- runif(n,min=0,max=5) # x ~ U(0,5)をn個生成
> y <- x # 理論式を y = x とする
> sd0 <- 1
> ee <- matrix(rnorm(n*B,mean=0,sd=sd0),n) # 誤差をN(0,1)とする
> yy <- y + ee # y = x + e
> ## QR分解
> x1 <- cbind(1,x) # データ行列
> q1 <- qr(x1)
> Q1 <- qr.Q(q1) # Q行列
> R1 <- qr.R(q1) # R行列
> IR1 <- solve(R1) # Rの逆行列
> ## 真の係数
> b0 <- c(0,1) # 真の回帰係数
> g0 <- R1 %*% b0 # 真の直交化した回帰係数
> ## 回帰係数の推定
```

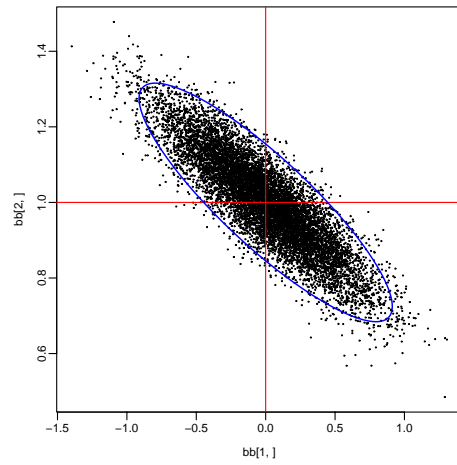
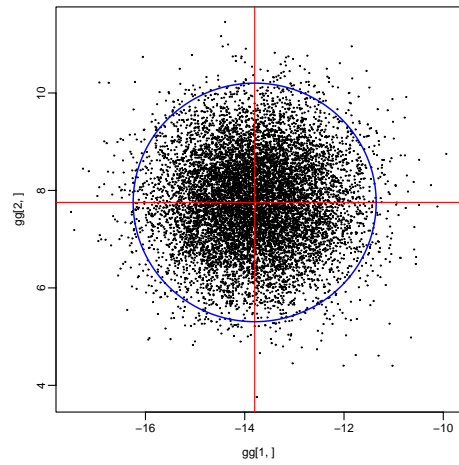
5

```
> gg <- t(Q1) %*% yy # 直交化した係数の推定
> bb <- IR1 %*% gg # 回帰係数の推定
> ## 最初のデータセット
> round(rbind(x,yy[,1]),2)
[ ,1] [ ,2] [ ,3] [ ,4] [ ,5] [ ,6] [ ,7] [ ,8] [ ,9] [ ,10] [ ,11] [ ,12] [ ,13] [ ,14]
x 1.98 0.13 0.86 0.31 4.45 0.19 2.89 2.60 2.97 4.34 3.62 4.79 3.07 4.25
1.67 2.28 -1.31 -0.89 3.66 -0.83 3.53 1.15 4.10 4.95 5.75 3.87 3.86 4.11
[ ,15] [ ,16] [ ,17] [ ,18] [ ,19] [ ,20] [ ,21] [ ,22] [ ,23] [ ,24] [ ,25] [ ,26] [ ,27] [ ,28]
x 2.78 4.27 1.97 3.90 2.55 0.21 1.95 3.79 3.69 2.68 0.60 1.50 1.11
4.33 4.29 1.80 3.65 2.21 -0.85 2.79 4.15 2.65 -0.10 0.35 2.47 1.11
[ ,28] [ ,29] [ ,30]
x 2.59 0.79 3.88
4.07 1.04 4.19
> bb[,1]
x
-0.2268940 1.0802004
> ## 2番目のデータセット
> round(rbind(x,yy[,2]),2)
[ ,1] [ ,2] [ ,3] [ ,4] [ ,5] [ ,6] [ ,7] [ ,8] [ ,9] [ ,10] [ ,11] [ ,12] [ ,13] [ ,14]
```

```
x 1.98 0.13 0.86 0.31 4.45 0.19 2.89 2.60 2.97 4.34 3.62 4.79 3.07 4.25
1.01 1.68 0.79 -1.23 3.69 0.40 3.98 2.39 3.50 4.01 2.65 3.76 5.27 5.76
[ ,15] [ ,16] [ ,17] [ ,18] [ ,19] [ ,20] [ ,21] [ ,22] [ ,23] [ ,24] [ ,25] [ ,26] [ ,27] [ ,28]
x 2.78 4.27 1.97 3.90 2.55 0.21 1.95 3.79 3.69 2.68 0.60 1.50 1.11
4.15 2.93 3.80 2.32 2.45 -0.73 1.62 3.60 2.45 3.27 -0.62 1.4 -0.11
[ ,28] [ ,29] [ ,30]
x 2.59 0.79 3.88
3.15 0.63 3.93
> bb[,2]
x
-0.08195452 0.97730787
> ## 最初の9個のデータセットの散布図
> matplot(x,yy[,1:9])
> for(k in 1:9) abline(bb[,k],col=k,lwd=2)
```



```
> ## 散布図1
> plot(gg[,1],gg[,2])
> abline(v=g0[1],col=2)
> abline(h=g0[2],col=2)
> r <- sd0*sqrt(qchisq(0.95,length(b0))) # 誤差のsd=1に注意
> r
[1] 2.447747
> i <- seq(0,2*pi,length=300)
> lines(cbind(g0[1]+r*cos(i),g0[2]+r*sin(i)),col=4,lwd=2)
> ## 確率
> rr <- apply(gg,2,function(v) sqrt(sum((v-g0)^2)))
> sum(rr<=r)
[1] 9520
> ## 散布図2
> plot(bb[,1],bb[,2])
> abline(v=b0[1],col=2)
> abline(h=b0[2],col=2)
> lines(cbind(g0[1]+r*cos(i),g0[2]+r*sin(i)) %*% t(IR1),col=4,lwd=2)
```



### 直交化した回帰係数からつくるF統計量

$$\frac{\hat{\gamma} - \gamma}{\sigma} \sim N_{p+1}(\mathbf{0}, \mathbf{I}_{p+1}), \quad \frac{\|\hat{\gamma} - \gamma\|^2}{\sigma^2} \sim \chi_{p+1}^2$$

$$\hat{\sigma}^2 = \frac{\|e\|^2}{n-p-1}, \quad \frac{\|e\|^2}{\sigma^2} \sim \chi_{n-p-1}^2$$

$$\frac{\|\hat{\gamma} - \gamma\|^2}{(p+1)\hat{\sigma}^2} \sim F_{p+1, n-p-1}$$

$$\Pr \left\{ \frac{\|\hat{\gamma} - \gamma\|^2}{(p+1)\hat{\sigma}^2} \leq F_{p+1, n-p-1}^\alpha \right\} = 1 - \alpha$$

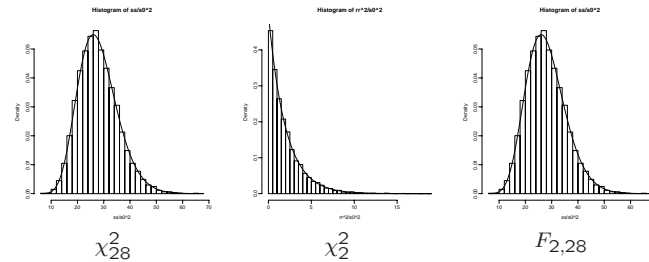
ただし  $F_{p+1, n-p-1}^\alpha$  は自由度  $(p+1, n-p-1)$  の  $F$  分布の上側  $\alpha$  点

$$\Pr \{ Y > F_{p+1, n-p-1}^\alpha \} = \alpha, \quad 0 < \alpha < 1$$

### 数値例

```
> ## 残差平方和の分布
> zz <- yy - x1 %*% bb # 残差
> ss <- apply(zz, 2, function(v) sum(v*v)) # 残差平方和
> hist(ss/s0^2, breaks=30, prob=T)
> j1 <- seq(min(ss/s0^2), max(ss/s0^2), length=300)
> lines(j1, dchisq(j1, n-length(b0)))
> ## 係数の中心からの二乗和の分布
> hist(rr^2/s0^2, breaks=30, prob=T)
> j2 <- seq(min(rr^2/s0^2), max(rr^2/s0^2), length=300)
> lines(j2, dchisq(j2, length(b0)))
> ## F統計量の分布
> ff <- (rr^2/length(b0))/(ss/(n-length(b0)))
> hist(ff, breaks=30, prob=T)
> j3 <- seq(min(ff), max(ff), length=300)
> lines(j3, df(j3, length(b0), n-length(b0)))
> f0 <- qf(0.95, length(b0), n-length(b0))
> f0
[1] 3.340386
```

```
> sum(ff>f0)
[1] 487
```



### 回帰係数の信頼領域

$$C_\gamma^{1-\alpha}(\hat{\gamma}, \hat{\sigma}^2) = \left\{ \gamma : \|\gamma - \hat{\gamma}\| \leq \hat{\sigma} \sqrt{(p+1)F_{p+1, n-p-1}^\alpha} \right\}$$

$$\Pr \{ \gamma \in C_\gamma^{1-\alpha}(\hat{\gamma}, \hat{\sigma}^2) \} = 1 - \alpha$$

$$\gamma = \mathbf{R}\beta$$

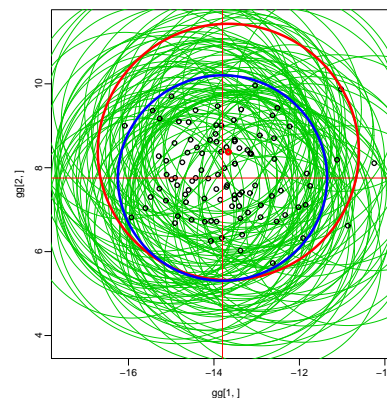
$$C_\beta^{1-\alpha}(\hat{\beta}, \hat{\sigma}^2) = \left\{ \beta : \frac{(\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})}{\hat{\sigma}^2} \leq (p+1)F_{p+1, n-p-1}^\alpha \right\}$$

$$\Pr \{ \beta \in C_\beta^{1-\alpha}(\hat{\beta}, \hat{\sigma}^2) \} = 1 - \alpha$$

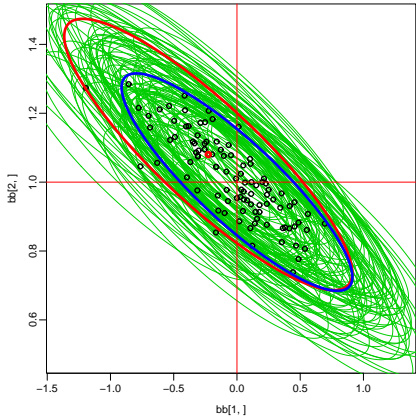
### 数値例

```
> ## 散布図3
> f0 <- qf(0.95, length(b0), n-length(b0)) # F分布の上側5%点
> rr0 <- sqrt(length(b0)*f0*(ss/(n-length(b0)))) # 信頼区間の半径
> sum(rr <= rr0)
[1] 9513
> sum(rr > rr0)
[1] 487
> a <- 1:100 # 最初の100点だけプロット
> sum(rr[a] > rr0[a])
[1] 7
> sum(rr[a] > r)
[1] 5
> plot(gg[1,], gg[2,], type="n") # まず座標軸だけ書く
> for(k in a) lines(cbind(gg[1,k]+rr0[k]*cos(i), gg[2,k]+rr0[k]*sin(i)), col=3,
> points(gg[1,a], gg[2,a], lwd=2)
> points(gg[1,1], gg[2,1], col=2, lwd=4) # 最初の1点だけ太く
> lines(cbind(gg[1,1]+rr0[1]*cos(i), gg[2,1]+rr0[1]*sin(i)), col=2, lwd=4)
> abline(v=g0[1], col=2) # 真のパラメタ値にクロスを書きそのまわりにサークル
```

```
> abline(h=g0[2], col=2)
> lines(cbind(g0[1]+r*cos(i), g0[2]+r*sin(i)), col=4, lwd=4)
```



```
> ## 散布図4
> plot(bb[1,], bb[2,], type="n") # まず座標軸だけ書く
> for(k in a) lines(cbind(gg[1,k]+rr0[k]*cos(i),
+ gg[2,k]+rr0[k]*sin(i)) %*% t(IR1), col=3, lwd=1)
> points(bb[1,a], bb[2,a], lwd=2)
> points(bb[1,1], bb[2,1], col=2, lwd=4) # 最初の1点だけ太く
> lines(cbind(gg[1,1]+rr0[1]*cos(i), gg[2,1]+rr0[1]*sin(i)) %*% t(IR1),
+ col=2, lwd=4)
> abline(v=b0[1], col=2) # 真のパラメタ値にクロスを書きそのまわりにサークル
> abline(h=b0[2], col=2)
> lines(cbind(g0[1]+r*cos(i), g0[2]+r*sin(i)) %*% t(IR1), col=4, lwd=4)
```



### 回帰係数の線形変換の同時信頼区間

$$w = a'\beta$$

$$C_w^{*1-\alpha}(\hat{\beta}, \hat{\sigma}^2) = \{w : w = a'\beta, \beta \in C_\beta^{1-\alpha}(\hat{\beta}, \hat{\sigma}^2)\}$$

$$\Pr\{w \in C_w^{*1-\alpha}(\hat{\beta}, \hat{\sigma}^2)\} \geq 1 - \alpha$$

$$w = a'R^{-1}\gamma = b'\gamma$$

シュバルツ (Schwarz) の不等式

$$|w - \hat{w}| = |b'(\hat{\gamma} - \gamma)| \leq \|b\| \cdot \|\hat{\gamma} - \gamma\|$$

で等号は  $b$  と  $\hat{\gamma} - \gamma$  が平行のときのみ .  $\|b\| = \sqrt{a'(X'X)^{-1}a}$

$$C_w^{*1-\alpha}(\hat{\beta}, \hat{\sigma}^2) = \{w : |w - \hat{w}| \leq \|b\| \hat{\sigma} \sqrt{(p+1)F_{p+1, n-p-1}^\alpha}\}$$

### 回帰係数の線形結合の信頼区間

$$w = a'\beta = b'\gamma$$

$$\hat{w} = a'\hat{\beta} = b'\hat{\gamma} \sim N(b'\gamma, \sigma^2 \|b\|^2)$$

$$\frac{\|\hat{w} - w\|^2}{\hat{\sigma}^2 \|b\|^2} \sim F_{1, n-p-1}, \quad \frac{\hat{w} - w}{\hat{\sigma} \|b\|} \sim t_{n-p-1}$$

$$C_w^{1-\alpha}(\hat{\beta}, \hat{\sigma}^2) = \{w : |w - \hat{w}| \leq \|b\| \hat{\sigma} \sqrt{F_{1, n-p-1}^\alpha}\}$$

$$\Pr\{w \in C_w^{1-\alpha}(\hat{\beta}, \hat{\sigma}^2)\} = 1 - \alpha$$

### 回帰直線 (曲面) の信頼領域

$$x'\beta$$

$$s^2 = \hat{\sigma}^2 x'(X'X)^{-1}x$$

$$C_{x'\beta}^{*1-\alpha}(\hat{\beta}, \hat{\sigma}^2) = \{x'\beta : |x'\beta - x'\hat{\beta}| \leq s \sqrt{(p+1)F_{p+1, n-p-1}^\alpha}\}$$

$$C_{x'\beta}^{1-\alpha}(\hat{\beta}, \hat{\sigma}^2) = \{x'\beta : |x'\beta - x'\hat{\beta}| \leq s \sqrt{F_{1, n-p-1}^\alpha}\}$$

$$C_{x'\beta}^{1-\alpha}(\hat{\beta}, \hat{\sigma}^2) \subset C_{x'\beta}^{*1-\alpha}(\hat{\beta}, \hat{\sigma}^2)$$

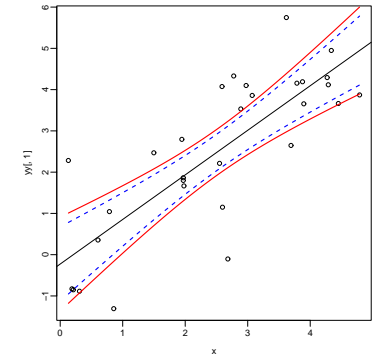
$$\Pr\{x'\beta \in C_{x'\beta}^{*1-\alpha}(\hat{\beta}, \hat{\sigma}^2), \forall x\} = 1 - \alpha$$

$$\Pr\{x'\beta \in C_{x'\beta}^{1-\alpha}(\hat{\beta}, \hat{\sigma}^2)\} = 1 - \alpha, \quad \forall x$$

### 数値例

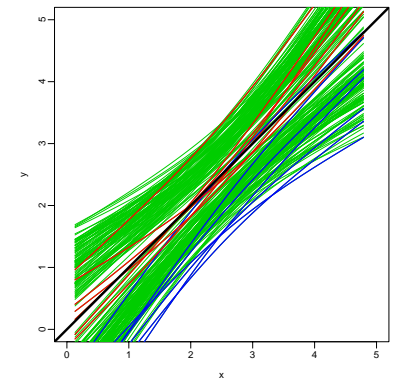
```
> ## 回帰直線の信頼領域 (最初のデータセット)
> ss[1]/(n-length(b0)) # 誤差分散の推定
[1] 1.396646
> jx <- seq(min(x),max(x),length=300) # xの範囲を等分割
> jx1 <- cbind(1, jx)
> jy <- jx1 %*% bb[,1] # 回帰直線のyの計算
> jss <- apply(jx1 %*% IR1,1,function(v)sum(v*v))
> js <- sqrt(jss*ss[1]/(n-length(b0))) # yの標準誤差
> a1 <- sqrt(length(b0)*qf(0.95,length(b0),n-length(b0))) # 同時信頼区間
> a1
[1] 2.584719
> a2 <- sqrt(qf(0.95,1,n-length(b0))) # 信頼区間
> a2
[1] 2.048407
> ## 散布図に回帰直線と信頼区間を書く
> plot(x,yy[,1]) # データ
> abline(bb[,1]) # 回帰直線
> lines(jx, jy+js*a1, col=2, lwd=2) # 95%同時信頼区間 (上側)
> lines(jx, jy-js*a1, col=2, lwd=2) # 95%同時信頼区間 (下側)
```

```
> lines(jx, jy-js*a1, col=2, lwd=2) # 95%同時信頼区間 (下側)
> lines(jx, jy+js*a2, col=4, lty=2, lwd=2) # 95%信頼区間 (上側)
> lines(jx, jy-js*a2, col=4, lty=2, lwd=2) # 95%信頼区間 (下側)
```



```
> ## 同時信頼区間と普通の信頼区間の違い
> jyy <- jx1 %*% bb # 回帰直線のyの計算をxの300点について計算
> dim(jyy)
[1] 300 10000
> jdd <- abs(jyy - jx) # 真の回帰直線との差の絶対値
> jsss <- sqrt(jss %*% t(ss)/(n-length(b0))) # yの標準誤差
> dim(jsss)
[1] 300 10000
> sum(apply(jdd <= jsss*a1,2,all))/10000 # 同時信頼区間はすべてのxで同時
[1] 0.955
> sum(apply(jdd <= jsss*a2,2,all))/10000 # 普通の信頼区間は同時ではない
[1] 0.8688
> sum(jdd <= jsss*a1)/(300*10000) # 同時信頼区間は各xを個別に見ると保守的
[1] 0.9849633
> sum(jdd <= jsss*a2)/(300*10000) # 普通の信頼区間は各xを個別に見るとOK
[1] 0.9509127
```

```
> ## 散布図5
> a <- (1:100)[apply(jdd[,1:100] > jsss[,1:100]*a1,2,any)]
> a
[1] 29 42 43 53 55 61 81
> plot(0,0,xlim=c(0,5),ylim=c(0,5),type="n",xlab="x",ylab="y")
> for(k in 1:100) { # 最初の100データセット
+ lines(jx, jyy[,k]+jsss[,k]*a1, col=3, lwd=1) # 95%同時信頼区間 (上側)
+ lines(jx, jyy[,k]-jsss[,k]*a1, col=3, lwd=1) # 95%同時信頼区間 (下側)
+ }
> for(k in a) { # 同時信頼区間が真の回帰直線を含まないもの
+ lines(jx, jyy[,k]+jsss[,k]*a1, col=2, lwd=2)
+ lines(jx, jyy[,k]-jsss[,k]*a1, col=4, lwd=2)
+ }
> abline(b0, lwd=4) # 真の回帰直線
```



第6回 課題

1. 信頼領域の計算で用いた

$$r_1(p) = \sqrt{(p+1)F_{p+1, n-p-1}^\alpha}, \quad r_2(p) = \sqrt{F_{1, n-p-1}^\alpha}$$

について  $n = 30, \alpha = 0.05$  とおき,  $r_1(p)$  と  $r_2(p)$  を  $p = 0, 1, \dots, 10$  の範囲で計算せよ.  $r_1(p)$  と  $r_2(p)$  の比較をして違いを述べよ. それは回帰直線(曲面)の信頼領域についてどのような結果をもたらすか?

2. 雪日数(B02304)を  $x$ , 最高気温(B02102)を  $y$  とする多項式回帰分析を次数  $p = 1, 2, 3$  について行え. それぞれの次数について  $x, y$  の散布図上に推定した回帰直線(曲線)とその95%同時信頼区間および95%信頼区間を重ねて描け.