データ解析

Rによる多変量解析入門

確率変数，回帰分析 (II)

lec20031106 下平英寿 shimo@is.titech.ac.jp

概要

確率分布の基礎，そして，回帰分析に関わる検定

- 確率変数
- 部分モデルの検定
- 信頼領域

確率分布の知識を，検定や信頼領域の計算に使う

確率変数

確率変数，確率分布関数，確率密度関数
(random variable, distribution function, density function)

- 実数値確率変数X，その実現値x

  確率分布関数 $F_X(x) = \Pr\{X \le x\}$

  確率密度関数 $f_X(x) = \dfrac{dF_X}{dx}$

- 集合Aにたいして

$$\Pr\{X \in A\} = \int_A f_X(x)\,dx = \int I_A(x) f_X(x)\,dx$$

- とくに$A = (-\infty, a]$なら

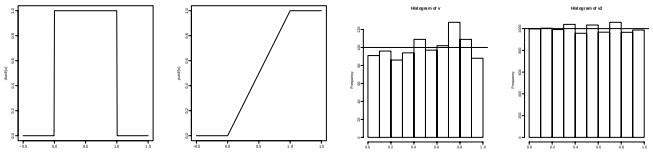
$$\Pr\{X \in A\} = \int_{-\infty}^{a} f_X(x)\,dx = F_X(a)$$

一様分布 (uinform distribution)

- 区間(a,b)の一様分布: $U(a,b)$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x \le a, \text{or } x \ge b \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \ge b \end{cases}$$



$a = 0, b = 1$

セッションファイル

edu:~shimo/class/gakubu200209/note20021106.Rt

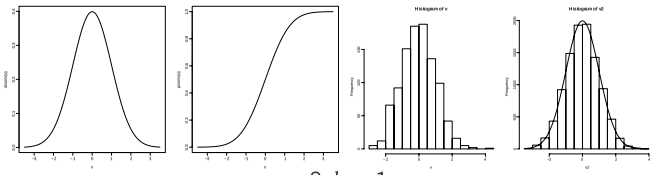
```
> ### 一様分布
> help(dunif)
> x <- seq(-0.5,1.5,length=300)
> plot(x,dunif(x),type="l")
> plot(x,punif(x),type="l")
> v <- runif(1000)
> hist(v,nclass=10)
> abline(h=100)
> v2 <- runif(10000)
> hist(v2,nclass=10)
> abline(h=1000)
> sum(v<=0.1) / length(v)
[1] 0.091
> sum(v2<=0.1) / length(v2)
[1] 0.0998
```

正規分布 (normal distribution)

- 平均a，分散bの正規分布: $N(a,b)$

$$f_X(x) = \frac{1}{\sqrt{2\pi b}} \exp\left(-\frac{(x-a)^2}{2b}\right)$$

$$F_X(x) = \int_{-\infty}^{x} f_X(s)\,ds$$



$a = 0, b = 1$


```
> ### 正規分布
> help(dnorm)
> x <- seq(-3.5,3.5,length=300)
> plot(x,dnorm(x),type="l")
> plot(x,pnorm(x),type="l")
> v <- rnorm(1000)
> hist(v,nclass=20)
> v2 <- rnorm(10000)
> hist(v2,nclass=20)
> lines(x,dnorm(x)*0.5*10000)
> sum(v<=-2) / length(v)
[1] 0.017
> sum(v2<=-2) / length(v2)
[1] 0.0237
> pnorm(-2)
[1] 0.02275013
```

中心極限定理 (central limit theorem)

- 平均μ，分散σ^2の確率変数の列（各X_iは互いに独立）

$$X_1, X_2, \ldots$$

$$E(X_i) = \mu, \quad V(X_i) = \sigma^2$$

- n個の和

$$Y_n = X_1 + \cdots + X_n$$

$$E(Y_n) = n\mu, \quad V(Y_n) = n\sigma^2$$

$$Z_n = \frac{Y_n - n\mu}{\sqrt{n\sigma^2}}$$

- Z_nは平均0，分散1．$n \to \infty$の極限でZ_nは$N(0,1)$に収束する．

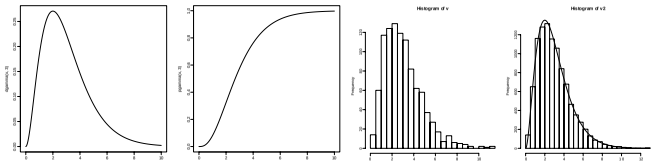
ガンマ分布 (gamma distribution)

- **shape a, scale sのガンマ分布:** $Ga(a,s)$

$$f_X(x) = \frac{1}{s\Gamma(a)}\left(\frac{x}{s}\right)^{a-1}\exp\left(-\frac{x}{s}\right), \quad x>0$$

$$\Gamma(a) = \int_0^\infty x^{a-1}\exp(-x)\,dx$$

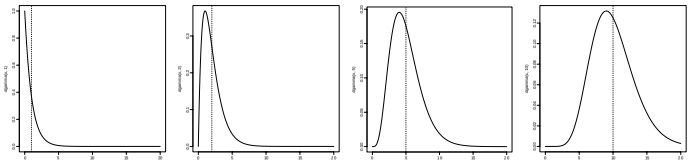
$$\Gamma(n) = (n-1)!$$



$a=3, s=1$


```
> ### ガンマ分布
> help(dgamma)
> x <- seq(0,10,length=300)
> gamma(3)
[1] 2
> plot(x,dgamma(x,3),type="l")
> plot(x,pgamma(x,3),type="l")
> v <- rgamma(1000,3)
> hist(v,nclass=20)
> v2 <- rgamma(10000,3)
> hist(v2,nclass=20)
> lines(x,dgamma(x,3)*0.5*10000)
> sum(v>7) / length(v)
[1] 0.035
> sum(v2>7) / length(v2)
[1] 0.029
> 1-pgamma(7,3)
[1] 0.02963616
```

ガンマ分布のshapeを変える($s=1$)



$a=1$ $a=2$ $a=5$ $a=10$

- **期待値（平均値）**

$$E(X) = \int_0^\infty x f_X(x)\,dx = as$$

- **分散**

$$V(X) = \int_0^\infty (x-E(X))^2 f_X(x)\,dx = as^2$$


```
> ## ガンマ関数（shapeを変えてみる）
> x <- seq(0,20,length=500)
> plot(x,dgamma(x,1),type="l"); abline(v=1,lty=3)
> plot(x,dgamma(x,2),type="l"); abline(v=2,lty=3)
> plot(x,dgamma(x,5),type="l"); abline(v=5,lty=3)
> plot(x,dgamma(x,10),type="l"); abline(v=10,lty=3)
> v <- rgamma(100000,1); c(mean(v),var(v))
[1] 1.001625 1.011608
> v <- rgamma(100000,2); c(mean(v),var(v))
[1] 2.004979 2.012081
> v <- rgamma(100000,5); c(mean(v),var(v))
[1] 5.000123 5.007556
> v <- rgamma(100000,10); c(mean(v),var(v))
[1] 10.00437 10.08082
```

カイ二乗分布 (chi-squared distribution)

- **自由度(degrees of freedom)がnのχ^2分布:** χ^2_n

$$f_X(x) = \frac{1}{2\Gamma(n/2)}\left(\frac{x}{2}\right)^{n/2-1}\exp\left(-\frac{x}{2}\right), \quad x>0$$

奇数のnに対して

$$\Gamma(n/2) = \frac{(n-2)!!\sqrt{\pi}}{2^{(n-1)/2}}, \quad (n-2)!! = (n-2)(n-4)\cdots$$

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi},\ \Gamma(\tfrac{3}{2}) = \tfrac{1}{2}\sqrt{\pi},\ \Gamma(\tfrac{5}{2}) = \tfrac{3}{4}\sqrt{\pi}$$

- **カイ二乗分布はガンマ分布の一例**
  $Ga(n/2,2)$**つまりshape:**$a=n/2$**, scale:**$s=2$**のガンマ分布**
- **「nに正の実数を許してスケールを調整したカイ二乗分布」がガンマ分布であると言うこともできる.**

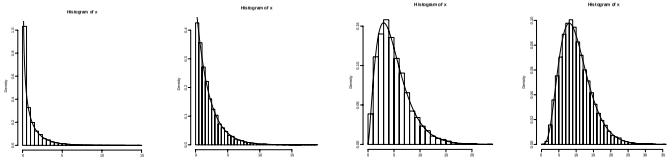
正規分布とカイ二乗分布の関係

$$Z_1, Z_2, \ldots, Z_n \sim N(0,1) \quad \textbf{i.i.d.}$$

i.i.d.=独立同分布 (independent identically distributed)

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

$$X \sim \chi^2_n$$



$n=1$ $n=2$ $n=5$ $n=10$

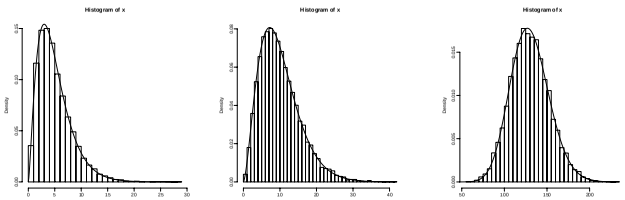

```
> ### 正規分布とカイ二乗分布の関係
> n <- 1; x <- matrix(rnorm(10000*n),n);
> y <- apply(x,2,function(v) sum(v*v)); xx <- seq(0,max(y),length=300)
> hist(y,prob=T,nclass=30); lines(xx,dchisq(xx,n))

> n <- 2; x <- matrix(rnorm(10000*n),n);
> y <- apply(x,2,function(v) sum(v*v)); xx <- seq(0,max(y),length=300)
> hist(y,prob=T,nclass=30); lines(xx,dchisq(xx,n))

> n <- 5; x <- matrix(rnorm(10000*n),n);
> y <- apply(x,2,function(v) sum(v*v)); xx <- seq(0,max(y),length=300)
> hist(y,prob=T,nclass=30); lines(xx,dchisq(xx,n))

> n <- 10; x <- matrix(rnorm(10000*n),n);
> y <- apply(x,2,function(v) sum(v*v)); xx <- seq(0,max(y),length=300)
> hist(y,prob=T,nclass=30); lines(xx,dchisq(xx,n))
```

非心カイ二乗分布

$$Z_i \sim N(a_i,1), \quad i=1,\ldots,n \quad \textbf{(独立)}$$

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

$$X \sim \chi^2_n(\delta), \quad \delta = a_1^2 + \cdots + a_n^2$$



$n=5, \delta=0$ $(a_i=0)$ $n=5, \delta=5$ $(a_i=1)$ $n=5, \delta=125$ $(a_i=5)$


```
> ### 非心カイ二乗分布
> n <- 5; a <- 0; z<- matrix(rnorm(10000*n)+a,n)
> x<- apply(z,2,function(v) sum(v*v)); zz <- seq(0,max(x),length=300)
> hist(x,prob=T,nclass=30); lines(zz,dchisq(zz,n,ncp=n*a^2))

> n <- 5; a <- 1; z<- matrix(rnorm(10000*n)+a,n)
> x<- apply(z,2,function(v) sum(v*v)); zz <- seq(0,max(x),length=300)
> hist(x,prob=T,nclass=30); lines(zz,dchisq(zz,n,ncp=n*a^2))

> n <- 5; a <- 5; z<- matrix(rnorm(10000*n)+a,n)
> x<- apply(z,2,function(v) sum(v*v)); zz <- seq(0,max(x),length=300)
> hist(x,prob=T,nclass=30); lines(zz,dchisq(zz,n,ncp=n*a^2))
```

再生性(reproductivity)

- $X_1 \sim N(a_1, b_1), X_2 \sim N(a_2, b_2)$；**(互いに独立)**
$$X_1 + X_2 \sim N(a_1 + a_2, b_1 + b_2)$$
- $X_1 \sim Ga(a_1, s), X_2 \sim Ga(a_2, s)$；**(互いに独立)**
$$X_1 + X_2 \sim Ga(a_1 + a_2, s)$$
- $X_1 \sim \chi^2_{n_1}, X_2 \sim \chi^2_{n_2}$；**(互いに独立)**
$$X_1 + X_2 \sim \chi^2_{n_1+n_2}$$

カイ二乗分布の再生性は以下のように理解できる

$(Z_1^2 + \cdots + Z_{n_1}^2) + (Z_{n_1+1}^2 + \cdots Z_{n_1+n_2}^2) = Z_1^2 + \cdots + Z_{n_1+n_2}^2$

一般的には「特性関数」を使って簡単に証明できる

F分布

$$X_1 \sim \chi^2_{n_1}, \quad X_2 \sim \chi^2_{n_2} \quad \textbf{互いに独立}$$

$$Y = \frac{X_1/n_1}{X_2/n_2}$$

このときYは自由度(n_1, n_2)のF分布に従う．

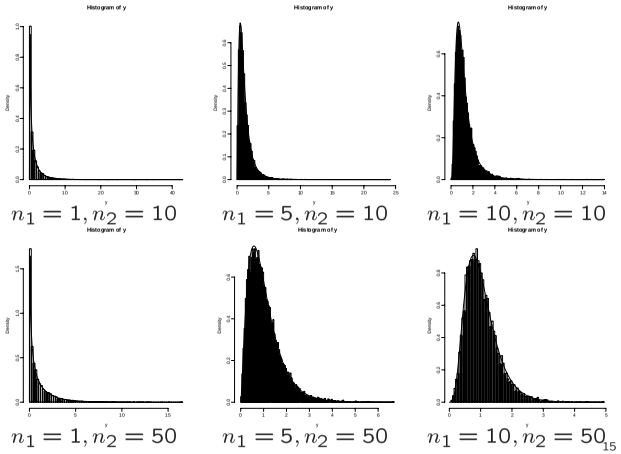
$$f_Y(y) = \frac{1}{B(n_1/2, n_2/2)} \left(\frac{n_1}{n_2}\right)^{n_1/2} \left(1 + \frac{n_1}{n_2} y\right)^{-(n_1+n_2)/2} y^{n_1/2-1}$$

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

```
> n1 <- 1; n2 <- 10; x1 <- rchisq(10000,n1); x2 <- rchisq(10000,n2)
> y <- (x1/n1)/(x2/n2) ; yy <- seq(0,max(y),length=300)
```



```
> hist(y,prob=T,breaks=100); lines(yy,df(yy,n1,n2))

> n1 <- 5; n2 <- 10; x1 <- rchisq(10000,n1); x2 <- rchisq(10000,n2)
> y <- (x1/n1)/(x2/n2) ; yy <- seq(0,max(y),length=300)
> hist(y,prob=T,breaks=100); lines(yy,df(yy,n1,n2))

> n1 <- 10; n2 <- 10; x1 <- rchisq(10000,n1); x2 <- rchisq(10000,n2)
> y <- (x1/n1)/(x2/n2) ; yy <- seq(0,max(y),length=300)
> hist(y,prob=T,breaks=100); lines(yy,df(yy,n1,n2))

> n1 <- 1; n2 <- 50; x1 <- rchisq(10000,n1); x2 <- rchisq(10000,n2)
> y <- (x1/n1)/(x2/n2) ; yy <- seq(0,max(y),length=300)
> hist(y,prob=T,breaks=100); lines(yy,df(yy,n1,n2))

> n1 <- 5; n2 <- 50; x1 <- rchisq(10000,n1); x2 <- rchisq(10000,n2)
> y <- (x1/n1)/(x2/n2) ; yy <- seq(0,max(y),length=300)
> hist(y,prob=T,breaks=100); lines(yy,df(yy,n1,n2))

> n1 <- 10; n2 <- 50; x1 <- rchisq(10000,n1); x2 <- rchisq(10000,n2)
> y <- (x1/n1)/(x2/n2) ; yy <- seq(0,max(y),length=300)
> hist(y,prob=T,breaks=100); lines(yy,df(yy,n1,n2))
```



$n_1 = 1, n_2 = 10$　$n_1 = 5, n_2 = 10$　$n_1 = 10, n_2 = 10$

$n_1 = 1, n_2 = 50$　$n_1 = 5, n_2 = 50$　$n_1 = 10, n_2 = 50$

t分布

$$Z \sim N(0, 1), \quad X \sim \chi^2_n \quad \textbf{互いに独立}$$

$$Y = \frac{Z}{\sqrt{X/n}}$$

このときYは自由度nのt分布に従う．

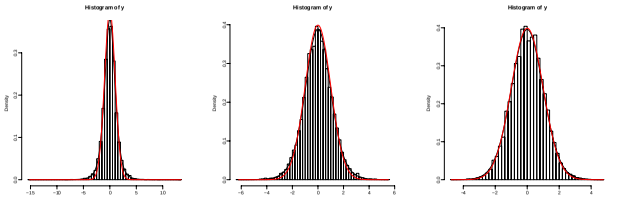
$$f_Y(y) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

なおY^2は自由度$(1, n)$のF分布に従う．

```
> ### t分布
> n <- 5; z <- rnorm(10000); x <- rchisq(10000,n); y <- z/sqrt(x/n)
> yy <- seq(min(y),max(y),length=300)
> hist(y,prob=T,nclass=50);lines(yy,dt(yy,n));lines(yy,dnorm(yy),col=2)
```



```
> n <- 10; z <- rnorm(10000); x <- rchisq(10000,n); y <- z/sqrt(x/n)
> yy <- seq(min(y),max(y),length=300)
> hist(y,prob=T,nclass=50);lines(yy,dt(yy,n));lines(yy,dnorm(yy),col=2)

> n <- 30; z <- rnorm(10000); x <- rchisq(10000,n); y <- z/sqrt(x/n)
> yy <- seq(min(y),max(y),length=300)
> hist(y,prob=T,nclass=50);lines(yy,dt(yy,n));lines(yy,dnorm(yy),col=2)
```



$n = 5$　$n = 10$　$n = 30$

部分モデルの検定

確率モデル

- **重回帰モデル**
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i; \quad i = 1, \ldots, n$$
$$\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2) \quad \textbf{i.i.d.}$$
- **共分散行列**
$$\sigma_{ij} = \mathrm{cov}(\epsilon_i, \epsilon_j)$$
$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \sigma_{ij} & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix}$$

とおくと，$\sigma_{ii} = \sigma^2$，$\sigma_{ij} = 0$，$i \neq j$である．

$$\Sigma = \sigma^2 \boldsymbol{I}_n$$

- **ベクトル表現**
$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

ただし$N_n(\mu, \Sigma)$は多変量正規分布（平均ベクトルμ，共分散行列Σ）

- **確率モデルのパラメータ（母数）**
$$\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$$

多変量正規分布

- $x_1, \ldots, x_n \sim N(0, \sigma^2)$が互いに独立なら

$$f_n(\boldsymbol{x}) = f(x_1) \cdots f(x_n)$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left[-\sum_{i=1}^{n} \frac{x_i^2}{2\sigma^2}\right] = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left[-\frac{1}{2}\boldsymbol{x}'(\sigma^2 \boldsymbol{I}_n)^{-1}\boldsymbol{x}\right]$$

- 平均ベクトルμ，共分散行列Σの多変量正規分布

$$\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$$

$$f_n(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left[-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

$n \times n$行列$\Sigma = (\sigma_{ij})$の成分σ_{ij}はx_iとx_jの共分散

- $x_1, \ldots, x_n \sim N(0, \sigma^2)$が互いに独立なら

$$\boldsymbol{x} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

回帰係数の推定量が従う分布

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}\end{aligned}$$

定理：（多変量）正規分布に従う確率変数を線形変換しても正規分布に従う

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\epsilon}) = \boldsymbol{\beta}$$

$$\begin{aligned}V(\hat{\boldsymbol{\beta}}) &= E\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right) \\ &= E\left[\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}\right\}\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}\right\}'\right] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\end{aligned}$$

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

QR分解と回帰係数の直交化

- QR分解

$$\boldsymbol{X} = [\boldsymbol{x}_0, \ldots, \boldsymbol{x}_p], \quad \boldsymbol{Q} = [\boldsymbol{q}_0, \ldots, \boldsymbol{q}_p], \quad \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_{p+1}$$

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}, \quad r_{ij} = 0, i > j$$

- 回帰係数の変数変換：$\beta \leftrightarrow \gamma$

$$\boldsymbol{\gamma} = \boldsymbol{R}\boldsymbol{\beta}, \quad \hat{\boldsymbol{\gamma}} \sim N_{p+1}(\boldsymbol{\gamma}, \sigma^2 \boldsymbol{I}_{p+1})$$

$$E(\hat{\boldsymbol{\gamma}}) = E(\boldsymbol{R}\hat{\boldsymbol{\beta}}) = \boldsymbol{R}E(\hat{\boldsymbol{\beta}}) = \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{\gamma}$$

$$\begin{aligned}V(\hat{\boldsymbol{\gamma}}) &= V(\boldsymbol{R}\hat{\boldsymbol{\beta}}) = E\left(\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{R}'\right) \\ &= \boldsymbol{R}E\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right)\boldsymbol{R}' = \boldsymbol{R}\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' \\ &= \sigma^2\boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}' = \sigma^2\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{R}'^{-1}\boldsymbol{R}' = \sigma^2\boldsymbol{I}_{p+1}\end{aligned}$$

QR分解と直交補空間

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$$

$\boldsymbol{Q} = [\boldsymbol{q}_0, \ldots, \boldsymbol{q}_p]$は$n \times (p+1)$行列．これに直交する$\boldsymbol{q}_{p+1}, \ldots, \boldsymbol{q}_{n-1}$を用意して$\boldsymbol{Q}^\perp = [\boldsymbol{q}_{p+1}, \ldots, \boldsymbol{q}_{n-1}]$とする

$\tilde{\boldsymbol{Q}} = [\boldsymbol{Q}, \boldsymbol{Q}^\perp] = [\boldsymbol{q}_0, \ldots, \boldsymbol{q}_{n-1}]$は$n \times n$の直交行列

$$\tilde{\boldsymbol{Q}}'\tilde{\boldsymbol{Q}} = \begin{bmatrix}\boldsymbol{Q}'\boldsymbol{Q} & \boldsymbol{Q}'\boldsymbol{Q}^\perp \\ \boldsymbol{Q}^{\perp\prime}\boldsymbol{Q} & \boldsymbol{Q}^{\perp\prime}\boldsymbol{Q}^\perp\end{bmatrix} = \boldsymbol{I}_n, \quad \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}' = \boldsymbol{Q}\boldsymbol{Q}' + \boldsymbol{Q}^\perp\boldsymbol{Q}^{\perp\prime} = \boldsymbol{I}_n$$

$$\hat{\boldsymbol{\gamma}} = \tilde{\boldsymbol{Q}}'\boldsymbol{y} = \begin{bmatrix}\boldsymbol{Q}'\boldsymbol{y} \\ \boldsymbol{Q}^{\perp\prime}\boldsymbol{y}\end{bmatrix} \quad \text{と}\hat{\boldsymbol{\gamma}}\text{を定義しなおす．}$$

$$\hat{\gamma}_i = \boldsymbol{q}_i'\boldsymbol{y}, \quad i = 0, \ldots, n-1$$

注意： $i = p+1, \ldots, n-1$の部分はここで新たに定義している．

QR分解と回帰係数の分布

モデル(p)

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{Q}'\boldsymbol{y} = \hat{\gamma}_0\boldsymbol{q}_0 + \cdots + \hat{\gamma}_p\boldsymbol{q}_p$$

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I}_n - \boldsymbol{Q}\boldsymbol{Q}')\boldsymbol{y} = \boldsymbol{Q}^\perp\boldsymbol{Q}^{\perp\prime}\boldsymbol{y} = \hat{\gamma}_{p+1}\boldsymbol{q}_{p+1} + \cdots + \hat{\gamma}_{n-1}\boldsymbol{q}_{n-1}$$

$$\hat{\boldsymbol{\gamma}} \sim N_n(\boldsymbol{\gamma}, \sigma^2\boldsymbol{I}_n)$$

つまり$\hat{\gamma}_0, \ldots, \hat{\gamma}_{n-1}$は平均$\gamma_0, \ldots, \gamma_{n-1}$，分散$\sigma^2$の正規分布にしたがう．

モデル(p)が正しければ $\gamma_{p+1} = \cdots = \gamma_{n-1} = 0$

モデル(k)

$$\hat{\boldsymbol{y}} = \hat{\gamma}_0\boldsymbol{q}_0 + \cdots + \hat{\gamma}_p\boldsymbol{q}_k$$

$$\boldsymbol{e} = \hat{\gamma}_{k+1}\boldsymbol{q}_{k+1} + \cdots + \hat{\gamma}_{n-1}\boldsymbol{q}_{n-1}$$

モデル(k)が正しければ $\gamma_{k+1} = \cdots = \gamma_{n-1} = 0$

しかし実際にはモデル(k)**は正しくないかもしれない．**

残差平方和 (residual sum of squares)

- モデル(k)

$$RSS^{(k)} = \|\boldsymbol{e}^{(k)}\|^2 = \hat{\gamma}_{k+1}^2 + \cdots + \hat{\gamma}_{n-1}^2$$

もしモデル(k)**が正しければ**$\gamma_{k+1} = \cdots = \gamma_{n-1} = 0$**なので**

$$RSS^{(k)}/\sigma^2 \sim \chi^2_{n-k-1}$$

一般には非心度$\sum_{i=k+1}^{n}\gamma_i^2/\sigma^2$**の非心カイ二乗分布**

- 誤差分散の不偏推定（モデル(p)が正しいと仮定して）

$$E(RSS^{(p)}) = \sigma^2(n-p-1)$$

を利用して

$$\hat{\sigma}^2 = \frac{RSS^{(p)}}{n-p-1}$$

c.f. 最尤推定は$RSS^{(p)}/n$

モデル(k)の検定

$$\frac{RSS^{(k)} - RSS^{(p)}}{\sigma^2} = \frac{\hat{\gamma}_{k+1}^2 + \cdots + \hat{\gamma}_p^2}{\sigma^2} \sim \chi^2_{p-k}(\delta)$$

$$\delta = \frac{\gamma_{k+1}^2 + \cdots + \gamma_p^2}{\sigma^2}$$

$$\frac{RSS^{(p)}}{\sigma^2} = \frac{\hat{\gamma}_{p+1}^2 + \cdots + \hat{\gamma}_{n-1}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

もしモデル(k)**が正しい，すなわち**$\delta = 0$**ならば**

$$F = \frac{(RSS^{(k)} - RSS^{(p)})/(p-k)}{RSS^{(p)}/(n-p-1)} \sim F_{p-k,n-p-1}$$

もし$\delta > 0$**ならば**F**分布から予想されるより実際の**F**は大きくなる傾向**

確率値(p**-value)** $= \Pr\{X > F\}, \quad X \sim F_{p-k,n-p-1}$

もし確率値が有意水準(5%)より小さければ仮説（モデル）を棄却

モデル(0)の検定

F**統計量は**

$$\begin{aligned}F &= \frac{(RSS^{(0)} - RSS^{(p)})/p}{RSS^{(p)}/(n-p-1)} \\ &= \frac{\hat{\gamma}_1^2 + \cdots + \hat{\gamma}_p^2}{\hat{\gamma}_{p+1}^2 + \cdots + \hat{\gamma}_{n-1}^2} \times \frac{n-p-1}{p} \\ &= \frac{R^2}{1-R^2} \times \frac{n-p-1}{p}\end{aligned}$$

ただしR^2**はモデル**(p)**の決定係数，すなわち重相関係数の二乗**

$$R^2 = \frac{\hat{\gamma}_1^2 + \cdots + \hat{\gamma}_p^2}{\hat{\gamma}_1^2 + \cdots + \hat{\gamma}_{n-1}^2}$$

帰無仮説（モデル(0)**）が正しいとき**F**統計量は自由度**$(p, n-p-1)$**の**F**分布に従う**

すでに`X`，`y`**が準備されているとして**

```
f <- mylsfit(X,y)  # 回帰分析
r2 <- cor(y,f$pred)^2  # 決定係数の計算
fs <- (r2/(1-r2)) * (nrow(X)-ncol(X)-1)/ncol(X) # F統計量
pv <- pf(fs,ncol(X),nrow(X)-ncol(X)-1,lower=F) # 確率値
```

得られた`r2`，`fs`，`pv`**を**`f$fsummary`**の内容と比較してみよ．**

```
> ax <- c("E09504","A0410302","C01301","B02101"); x <- X2000$x[,ax]
> ay <- "A05203"; y <- X2000$x[,ay]
> X2000$jitem[c(ax,ay)]
                                E09504                                A0410302
"最終学歴が大学・大学院卒の者の割合 " "未　婚　者　割　合［20～24歳・女］ "
                                C01301                                B02101
          "県民１人当たり県民所得　 "                  "年　平　均　気　温　 "
                                A05203
            "合 計 特 殊 出 生 率　 "
> f <- mylsfit(x,y)
> f$fsummary
  Mean Sum Sq R Squared  F-value Df 1 Df 2       Pr(>F)
Y  0.07061925 0.7431179 30.37478    4   42 6.678249e-12
```

モデル$(p-1)$の検定

- **モデル$(p-1)$が正しいとき，つまり$\beta_p = 0$のとき，次のF統計量は自由度$(1, n-p-1)$のF分布に従う．**

$$F = \frac{(RSS^{(p-1)} - RSS^{(p)})}{RSS^{(p)}/(n-p-1)} = \frac{\hat{\gamma}_p^2}{\hat{\sigma}^2}$$

- **$(X'X)^{-1}$の一番右下の要素をaとおくと$\hat{\beta}_p \sim N(\beta_p, \sigma^2 a)$．したがって次の$t$統計量は自由度$n-p-1$の$t$分布に従う．**

$$t = \frac{\hat{\beta}_p}{\hat{\sigma}\sqrt{a}}$$

- $r_{pp} = \pm 1/\sqrt{a}$，$\hat{\gamma}_p = r_{pp}\hat{\beta}_p$**の関係があるので，**

$$F = \frac{r_{pp}^2 \hat{\beta}_p^2}{\hat{\sigma}^2} = t^2$$

すでにX，yが準備されているとして

```
X1 <- cbind(1,X) # 定数項の列を加える
a <- diag(solve(t(X1) %*% X1))  # すべての回帰係数の「a」
1/sqrt(a) # この最後の要素とR行列の右下の要素は符号を除いて等しい
qr.R(qr(X1)) # R行列
f <- mylsfit(X,y)  # 回帰分析
s2 <- sum(f$resid^2)/(nrow(X)-ncol(X)-1) # 残差の分散の推定
ts <- f$coef/sqrt(s2*a) # すべての回帰係数のt統計量
pv <- pt(abs(ts),nrow(X)-ncol(X)-1,lower=F)*2 # 確率値
```

得られたts，pvをf$tsummaryの内容と比較してみよ．

	Estimate	Std.Err	t-value	Pr(>|t|)
Intercept	3.636118e+00	6.108140e-01	5.9529052	4.642519e-07
E09504	-1.655659e-02	6.570354e-03	-2.5198932	1.562857e-02
A0410302	-2.727500e-02	7.438516e-03	-3.6667262	6.850351e-04
C01301	1.228960e-05	4.635665e-05	0.2651098	7.922217e-01
B02101	2.012511e-02	5.098161e-03	3.9475245	2.951848e-04

信頼領域

回帰係数の線形変換の分布

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

任意の$a \times (p+1)$行列Aを使って

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

の平均と分散は

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

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

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

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

$$\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' = \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}' = \boldsymbol{R}\boldsymbol{R}^{-1}(\boldsymbol{R}')^{-1}\boldsymbol{R}' = \boldsymbol{I}_{p+1}$$

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

数値例

セッションファイル

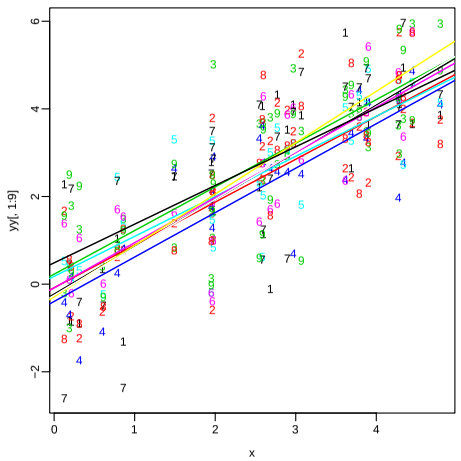
edu:~shimo/class/gakubu200209/note20021118.Rt

```
> ## 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
```

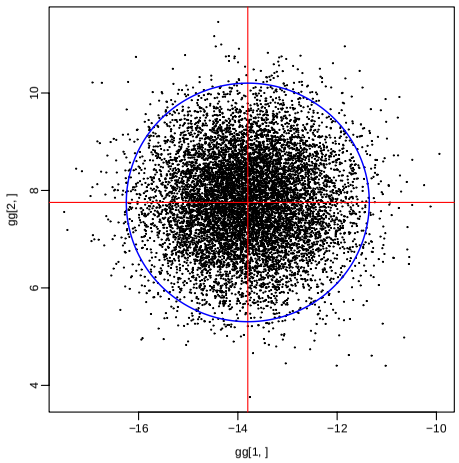


```
> b0 <- c(0,1) # 真の回帰係数
> g0 <- R1 %*% b0 # 真の直交化した回帰係数
> ## 回帰係数の推定
> gg <- t(Q1) %*% yy # 直交化した係数γの推定
> bb <- IR1 %*% gg # 回帰係数βの推定
> ## 最初のデータセット
> round(rbind(x,yy[,1]),2)
  [,1] [,2]  [,3]  [,4] [,5]  [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,1
x 1.98 0.13  0.86  0.31 4.45  0.19 2.89 2.60 2.97  4.34  3.62  4.79  3.07  4.
  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.
  [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26] [,2
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.
   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.
  [,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  4.29
  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] [,2
x  2.78  4.27  1.97  3.90  2.55  0.21  1.95  3.79  3.69  2.68  0.60   1.5  1.
   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.
  [,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)
```



$(\hat{\gamma}_0, \hat{\gamma}_1)$のプロット

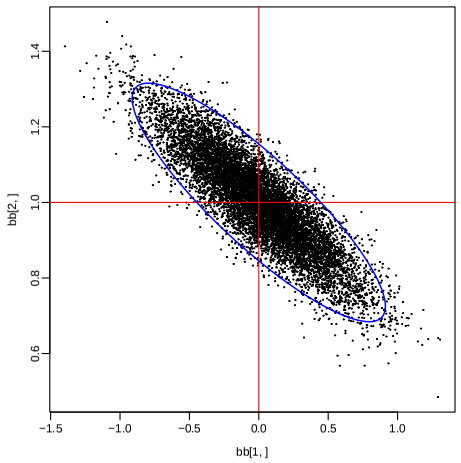
円の半径：

$r = \sigma\sqrt{\chi_2^{2\,(0.05)}}$

円内に入る確率:

$\Pr\{\|\hat{\gamma} - \gamma\| \le r\}$

$= 0.95$



$(\hat{\beta}_0, \hat{\beta}_1)$のプロット

$\hat{\boldsymbol{\beta}} = \boldsymbol{R}^{-1}\hat{\boldsymbol{\gamma}}$

楕円内に入る確率

$=0.95$

直交化した回帰係数からつくるF統計量

$$\hat{\boldsymbol{y}} = \hat{\gamma}_0\boldsymbol{q}_0 + \cdots + \hat{\gamma}_p\boldsymbol{q}_p$$
$$\boldsymbol{e} = \hat{\gamma}_{p+1}\boldsymbol{q}_{p+1} + \cdots + \hat{\gamma}_{n-1}\boldsymbol{q}_{n-1}$$

各$\hat{\gamma}_i \sim N(\gamma_i, \sigma^2)$は互いに独立.(したがって$\hat{\boldsymbol{y}}$と$\boldsymbol{e}$は互いに独立.)
モデル(p)が正しければ$\gamma_{p+1} = \cdots = \gamma_{n-1} = 0$.

$$\frac{\|\hat{\gamma} - \gamma\|^2}{\sigma^2} \sim \chi^2_{p+1}, \qquad \frac{\|\boldsymbol{e}\|^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

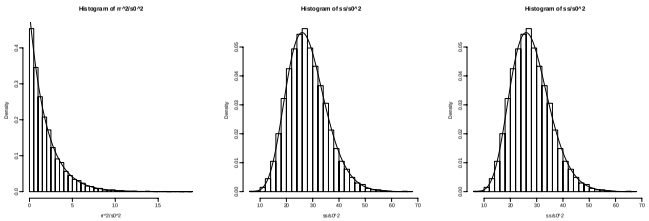
$$F = \frac{\|\hat{\gamma} - \gamma\|^2/(p+1)}{\|\boldsymbol{e}\|^2/(n-p-1)} = \frac{\|\hat{\gamma} - \gamma\|^2}{(p+1)\hat{\sigma}^2} \sim F_{p+1,n-p-1}$$

数値例

```
> ## 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)))
> ## 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)))
> ## 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
```



χ^2_2 χ^2_{28} $F_{2,28}$

直交化した回帰係数の信頼領域

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

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

$$\Pr\left\{Y > F^{(\alpha)}_{p+1,n-p-1}\right\} = \alpha, \quad 0 < \alpha < 1$$

回帰係数の信頼領域

$$C^{1-\alpha}_{\gamma}(\hat{\gamma}, \hat{\sigma}^2) = \left\{\gamma : \|\gamma - \hat{\gamma}\|^2 \le (p+1)\hat{\sigma}^2 F^{(\alpha)}_{p+1,n-p-1}\right\}$$

$$\Pr\left\{\gamma \in C^{1-\alpha}_{\gamma}(\hat{\gamma}, \hat{\sigma}^2)\right\} = 1 - \alpha$$

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

$$C^{1-\alpha}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \left\{\boldsymbol{\beta} : \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{\hat{\sigma}^2} \le (p+1)F^{(\alpha)}_{p+1,n-p-1}\right\}$$

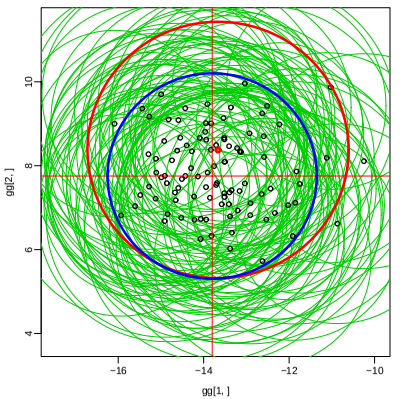
$$\Pr\left\{\boldsymbol{\beta} \in C^{1-\alpha}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)\right\} = 1 - \alpha$$

数値例

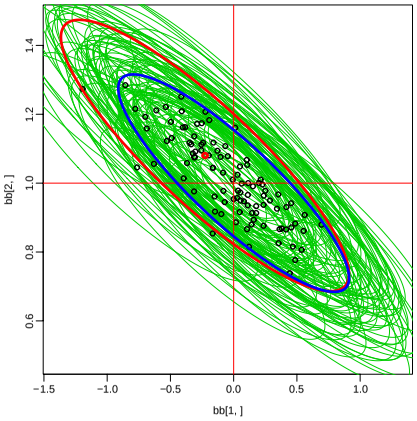
```
> ## 散布図3
> f0 <- qf(0.95,length(b0),n-length(b0)) # F分布の上側5%点
> rr0 <- sqrt(length(b0)*f0*(ss/(n-length(b0)))) # 信頼区間の半径
> sum(rr <= rr0)
[1] 9513
> sum(rr > rr0)
[1] 487
> a <- 1:100 # 最初の100点だけプロット
> sum(rr[a] > rr0[a])
[1] 7
> sum(rr[a] > r)
[1] 5
> plot(gg[1,],gg[2,],type="n") # まず座標軸だけ書く
> 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)
```



回帰係数の線形結合の同時信頼区間

$$w = \boldsymbol{a}'\boldsymbol{\beta}$$

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

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

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

シュバルツ（Schwarz）の不等式

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

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

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

回帰係数の線形結合の信頼区間

$$w = \boldsymbol{a}'\boldsymbol{\beta} = \boldsymbol{b}'\boldsymbol{\gamma}$$

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

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

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

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

回帰直線（曲面）の信頼領域

$$\boldsymbol{x}'\boldsymbol{\beta}$$

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

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

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

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

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

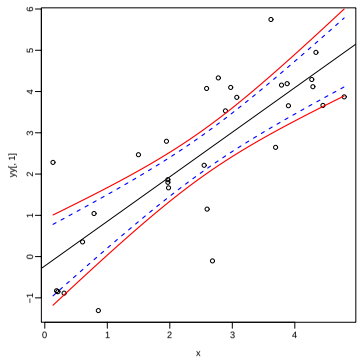
$$\Pr\left\{ \boldsymbol{x}'\boldsymbol{\beta} \in C_{\boldsymbol{x}'\boldsymbol{\beta}}^{1-\alpha}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) \right\} = 1 - \alpha, \quad \forall \boldsymbol{x}$$

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（上限）
```

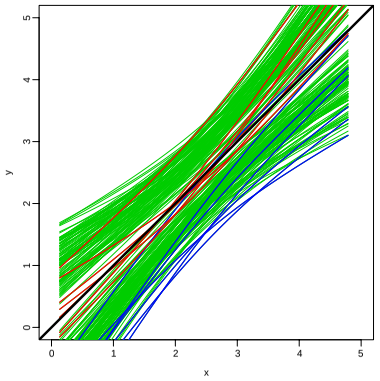


```
> lines(jx,jy-js*a1,col=2,lwd=2) #  95%同時信頼区間（下限）
> lines(jx,jy+js*a2,col=4,lty=2,lwd=2) # 95%信頼区間（上限）
> lines(jx,jy-js*a2,col=4,lty=2,lwd=2) # 95%信頼区間（下限）
```




```
> ## 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（上限）
+ lines(jx,jyy[,k]-jsss[,k]*a1,col=3,lwd=1) #  95%同時信頼区間（下限）
+ }
> 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
```



演習問題

3-1．平均ベクトルと共分散行列が次のように与えられる２次元正規分布に従う確率変数xを10000個生成し，サイズ2×10000の行列に代入せよ．また，xの散布図をplot関数を用いて描いてみよ．

$$\mu = \begin{bmatrix} 10 \\ 20 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$

ヒント： コレスキー分解$\Sigma = R'R$と$z \sim N_2(0, I_2)$から$x = R'z + \mu$と書ける．

3-2.　上記の確率変数の成分の和$y = x_1 + x_2$のヒストグラムを`hist(y,prob=T)`などを使って書き，そこにyの確率密度関数の理論曲線を重ね合わせて表示せよ．

3-3．適当な自由度のカイ二乗分布に従う確率変数を２種類以上生成し，それらの和が再びカイ二乗分布になることをヒストグラムと確率密度関数の理論曲線を使って例示せよ．

3-4.　まず次のようにデータ$(x_i, y_i), i = 1, \ldots, n$を生成し，これに３次の多項式回帰モデルを当てはめ推定した回帰係数を示せ．データの散布図に推定した曲線を重ねて表示せよ．

```
n <- 30  # データ数 n = 30
x <- runif(n,min=-3,max=3)  # x ~ U(-3,3)をn個生成
y <- 3 + 2*x + x^2 # 理論式を y = 3 + 2x + x^2  とする
y <- y + rnorm(n,mean=0,sd=1) # 誤差をN(0,1)とする.
```

3-5.　上記のデータについて決定係数R^2を計算せよ．「モデル(0)の検定」を行い，そのF統計量と確率値を示せ．それを`f$fsummary`と比較せよ．

3-6.　上記のデータについてすべての回帰係数の有意性を検定せよ．t統計量と確率値を示せ．それを`f$tsummary`と比較せよ．

3-7．信頼領域の計算で用いた

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

について$n = 30, \alpha = 0.05$とおき，$r_1(p)$と$r_2(p)$を$p = 0, 1, \ldots, 10$ 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?

3-8．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．