データ解析
Rによる多変量解析入門
(5) 確率変数，回帰分析(II)

lec20021107 下平英寿 shimo@is.titech.ac.jp

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

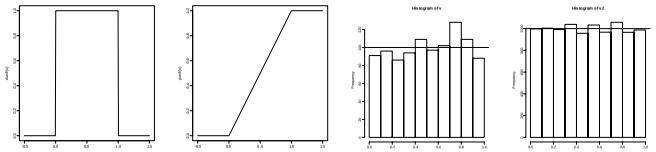
$$\Pr\{X \in A\} = \int_{-\infty}^{a} f_X(x)\,dx = F_X(a)$$

一様分布 (uinform distribution)

- 区間 (a,b) の一様分布: $U(a,b)$

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

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



$a = 0, b = 1$

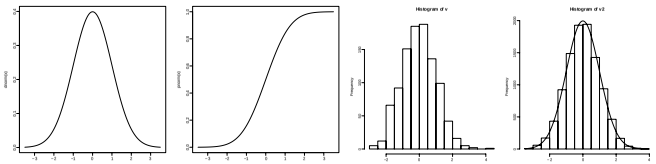

```
> ### 一様分布
> help(dunif)
> x <- seq(-0.5,1.5,length=300)
> plot(x,dunif(x),type="l")
> plot(x,punif(x),type="l")
> v <- runif(1000)
> hist(v,nclass=10)
> abline(h=100)
> v2 <- runif(10000)
> hist(v2,nclass=10)
> abline(h=1000)
> sum(v<=0.1) / length(v)
[1] 0.091
> sum(v2<=0.1) / length(v2)
[1] 0.0998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

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
```

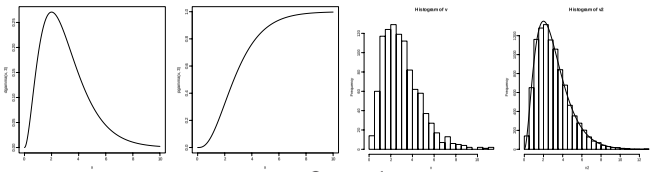
ガンマ分布 (gamma distribution)

- shape a，scale s のガンマ分布: $Ga(a,s)$

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

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

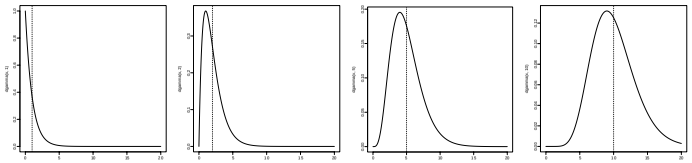
$$\Gamma(n) = (n-1)!$$



$a = 3, s = 1$


```
> ### 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

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

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

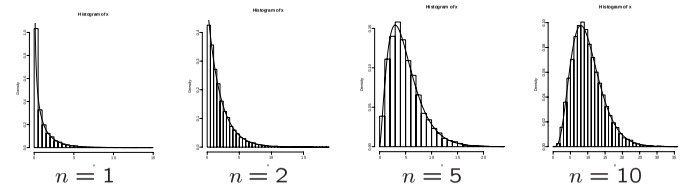
- **カイ二乗分布はガンマ分布の一例**
  $Ga(n/2, 2)$つまりshape:$a = n/2$, scale:$s = 2$のガンマ分布
- **「nに正の実数を許してスケールを調整したカイ二乗分布」がガンマ分布であると言うこともできる.**

正規分布とカイ二乗分布の関係

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

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

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



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

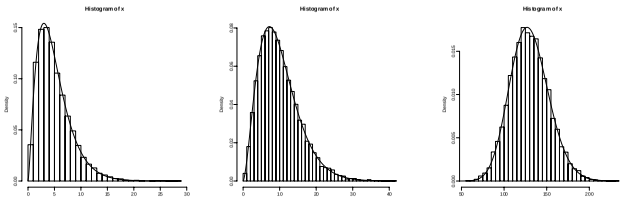

```
> ### 正規分布とカイ二乗分布の関係
> n <- 1; x <- matrix(rnorm(10000*n),n);
> y <- apply(x,2,function(v) sum(v*v)); xx <- seq(0,max(y),length=300)
> hist(y,prob=T,nclass=30); lines(xx,dchisq(xx,n))

> 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

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

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

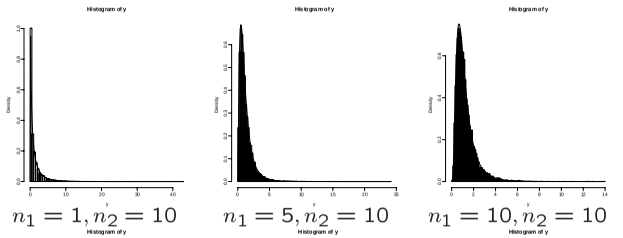
このときYは自由度(n_1, n_2)のF分布に従う.

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

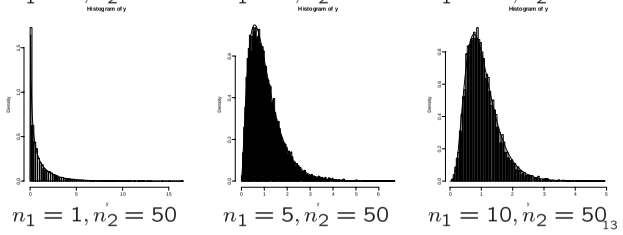
$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

```
> 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

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

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

このときYは自由度nのt分布に従う．

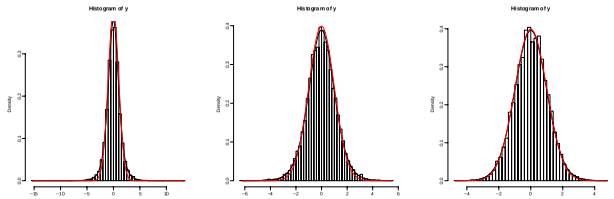
$$f_Y(y) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$$

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)
```



$n = 5$　$n = 10$　$n = 30$

回帰分析(II)

確率モデル

- **重回帰モデル**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i; \quad i = 1, \ldots, n$$

$$\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2) \quad \textbf{i.i.d.}$$

c.f. $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$

- **ベクトル表現**

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(0, \sigma^2 \boldsymbol{I}_n)$$

- **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

$$f_n(\boldsymbol{x}) = f(x_1) \cdots f(x_n) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left[-\sum_{i=1}^n \frac{x_i^2}{2\sigma^2}\right]$$

- **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

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

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

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

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

$$\tilde{\boldsymbol{Q}} = [\boldsymbol{Q}, \boldsymbol{Q}^\perp] = [\boldsymbol{q}_0, \ldots, \boldsymbol{q}_{n-1}], \quad \tilde{\boldsymbol{Q}}'\tilde{\boldsymbol{Q}} = \boldsymbol{I}_n$$

$$\boldsymbol{Q}\boldsymbol{Q}' + \boldsymbol{Q}^\perp\boldsymbol{Q}^{\perp\prime} = \boldsymbol{I}_n$$

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

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

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

QR分解と回帰係数の分布

$$\hat{\gamma}_i = \boldsymbol{q}_i'\boldsymbol{y} = \boldsymbol{q}_i'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) = \gamma_i + \boldsymbol{q}_i'\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$$

$$\gamma_i = \boldsymbol{q}_i'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{q}_i'\boldsymbol{Q}\boldsymbol{R}\boldsymbol{\beta} = \begin{cases} \sum_{j=i}^{p} r_{ij}\beta_j & i = 0, \ldots, p \\ 0 & i = p+1, \ldots, n \end{cases}$$

$$E(\hat{\gamma}_i) = \gamma_i$$

$$\mathrm{cov}(\hat{\gamma}_i, \hat{\gamma}_j) = E\left\{(\hat{\gamma}_i - \gamma_i)(\hat{\gamma}_j - \gamma_j)\right\} = \boldsymbol{q}_i' E\left\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right\}\boldsymbol{q}_j = \sigma^2\delta_{ij}$$

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

モデル(k)

$$\gamma_{k+1} = \cdots = \gamma_n = 0$$

QR分解と部分回帰

- **モデル(p)**

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

- **モデル(k)**

$$\boldsymbol{Q}^{(k)} = [\boldsymbol{q}_0, \ldots, \boldsymbol{q}_k]$$

$$\boldsymbol{\gamma}^{(k)} = \boldsymbol{Q}^{(k)\prime}\boldsymbol{y} = \begin{bmatrix} \hat{\gamma}_0 \\ \vdots \\ \hat{\gamma}_k \end{bmatrix}, \quad \hat{\gamma}_i = \boldsymbol{q}_i'\boldsymbol{y}$$

$$\hat{\boldsymbol{y}}^{(k)} = \boldsymbol{Q}^{(k)}\boldsymbol{Q}^{(k)\prime}\boldsymbol{y} = \boldsymbol{Q}^{(k)}\hat{\boldsymbol{\gamma}}^{(k)} = \hat{\gamma}_0\boldsymbol{q}_0 + \cdots + \hat{\gamma}_k\boldsymbol{q}_k$$

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

- **モデル(0)**

$$\bar{y}\boldsymbol{1}_n = \hat{\gamma}_0\boldsymbol{q}_0$$

残差平方和（residual sum of squares）

- **モデル(k)**

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

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

- **残差の推定（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

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

モデル(k)の検定

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

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

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

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

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

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

$$R^2 = \frac{\hat{\gamma}_1^2 + \cdots + \hat{\gamma}_p^2}{\hat{\gamma}_1^2 + \cdots + \hat{\gamma}_n^2}$$

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．

モデル$(p-1)$の検定

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

ただし

$$t = \frac{\hat{\beta}_p}{\hat{\sigma}/|r_{pp}|}, \quad \hat{\gamma}_p = r_{pp}\hat{\beta}_p$$

- **帰無仮説（モデル$(p-1)$）が正しいときF統計量は自由度$(1, n-p-1)$のF分布に従う．またt統計量は自由度$n-p-1$のt分布に従う．この事実を使って回帰係数β_pの有意性検定をする．**
- **仮想的に回帰係数の順序を並べ替えればすべての$\beta_0, \ldots, \beta_p$について同様の検定が行える．注意：$(X'X)^{-1}$の一番右下の要素をaとおくと，$r_{pp} = \pm 1/\sqrt{a}$の関係がある．**

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．

第５回 課題

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．

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

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．

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．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．

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.
```

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．

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．