Data analysis

Ben Graham

MA930, University of Warwick

October 29, 2015

Recap: the bivariate normal distribution

- $X_1, X_2 \sim N(0,1)$ independent
- $\rho \in [-1,1]$ coefficient of correlation
- X_1 and $X_3 := \rho X_1 + \sqrt{1-\rho^2} X_2$ are bivariate normal
- $\mathrm{Cor}(X_1, X_3) = \rho$
- $\mathrm{Cor}(aX_1 + b, cX_3 + d) = \rho$
- $\mathrm{Cov}(aX_1 + b, cX_3 + d) = ac\rho$
- faithful

Multivariate normal distribution

- If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ iidrv

$$f_X(x) = (\sigma\sqrt{2\pi})^{-n} \exp\left(-\frac{(x-\mu)^T(x-\mu)}{2\sigma^2}\right)$$

  Covariance matrix is $\Sigma = \sigma^2 I_n$
- Consider $X = (X_1, \ldots, X_n)$ with mean $\mu \in \mathbb{R}^n$ and Cov(X)$= \Sigma \in R^{n\times n}$

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

- If $X \sim N(\mu, \Sigma)$ then $MX \sim N(M\mu, M\Sigma M^T)$
- If M is a rotation matrix, $M \times N(0, I_n) = N(0, I_n)$

Recap: the t-distribution

- $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ iidrv
- $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$ is independent of
- $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$,

$$\frac{(n-1)S^2}{\sigma^2} = \sigma^{-2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2_{n-1}$$

-

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Recap

- More generally, let $e_1, \ldots, e_n$ denote a collection of independent unit vectors such that
  - $\mathbb{E}\langle X, e_i\rangle \neq 0$ for $i = 1, \ldots, k$ and
  - $\mathbb{E}\langle X, e_i\rangle = 0$ for $i = k+1, \ldots, k+\ell$
  - $\mathbb{E}\langle X, e_i\rangle = 0$ for $i = k+\ell, \ldots, n$
- Independently

$$\left\|\sum_{i=1}^{k} e_i \cdot \langle X_i, e_i\rangle - \mathbb{E}X\right\|_2^2 = \left\|\sum_{i=1}^{k} [e_i \cdot \langle X_i, e_i\rangle - e_i \cdot \mathbb{E}\langle X_i, e_i\rangle]\right\|_2^2 \sim \sigma^2\chi_k^2$$

$$\left\|\sum_{i=k+1}^{k+\ell} e_i \cdot \langle X_i, e_i\rangle\right\|_2^2 \sim \sigma^2\chi_\ell^2$$

$$\left\|\sum_{i=k+\ell+1}^{n} e_i \cdot \langle X_i, e_i\rangle\right\|_2^2 \sim \sigma^2\chi_{n-k-\ell}^2$$

- Ratio of independent χ^2 $\rightarrow$Fisher's F-distribution

Linear Regression in one variable

- $Y_i = A + BX_i + \epsilon_i$, $\mathbb{E}\epsilon_i = 0$
- Least squares estimates: choose $A, B \in \mathbb{R}$ to minimize the sum of squares

$$SS = \sum_{i=1}^{n} (Y_i - A - BX_i)^2 = \|Y - \mathbb{E}Y\|_2^2 = \|\epsilon\|_2^2$$

- Solve $\frac{\partial SS}{\partial A} = 0$ and $\frac{\partial SS}{\partial B} = 0$
- $A = \bar{Y} - B\bar{X}$ and $B = Cov(X, Y)/Var(X)$

Multivariate Linear regression

- In terms of matrices:
  - Y is an $n \times 1$ column vector
  - X is an $n \times k$ matrix—the design matrix
  - β is a $k \times 1$ column vector—the parameters
  - ϵ is an $n \times 1$ column vector—mean zero
  - $Y = X\beta + \epsilon$
  - Want to minimize $SS = \sum(Y_i - (X\beta)_i)^2 = \|Y - X\beta\|_2^2$
  - Why least squares? MLE

Multivariate Linear regression

$$SS = (y - X\beta)^T(y - X\beta) = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta$$

$$= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta$$

$$\frac{\partial SS}{\partial \beta} = -2X^T y + 2(X^T X)\beta$$

Need

$$(X^T X)\beta = X^T y$$

If the inverse exists

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Hence: $\hat{\beta} \sim N(\beta, (X^T X)^{-1}\sigma^2)$

Multivariate Linear regression

In the 2d case

$$X = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$$

$$X^T X = \begin{pmatrix} n & n\bar{X} \\ n\bar{X} & \sum X_i^2 \end{pmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Let $\overline{XY} = n^{-1} \sum_{i=1}^{n} X_i Y_i$, etc

Example: Fitting a polynomial

- $Y = A + BX + CX^2 + DX^3 + \epsilon$
- Danger of overfitting!!
- Hypothesis testing per coefficient
- ANOVA model vs model
- AIC$= 2k - 2\log(\Lambda)$
- BIC$= k \log n - 2\log(\Lambda)$
- L_1 penalties on the weights (Ridge regression)
- L_2 penalties on the weights (LASSO)

t-test for coefficients

- Two models ($\Delta df = m - (m-1) = 1$)
  - $H_0 : Y = X_0\beta_0 + \epsilon,\ \beta_0 \in \mathbb{R}^{m-1}$
  - $H_1 : Y = X_1\beta_1 + \epsilon,\ \beta_1 = (\beta_0, \beta_+) \in \mathbb{R}^m$
  - i.e. Under H_0, $\beta_+ = 0$

$$S^2 = \frac{1}{n-m}\left\|Y - X_1\hat{\beta}_1\right\|_2^2, \qquad \frac{(n-m)S^2}{\sigma^2} \sim \chi^2_{n-m}$$

$$\textit{Under } H_0 : \frac{\hat{\beta}_+}{\sigma\sqrt{(X^TX)^{-1}_{++}}} \sim N(0,1)$$

$$\frac{\hat{\beta}_+}{S\sqrt{(X^TX)^{-1}_{++}}} \sim t_{n-m}$$

ANOVA

- Two models ($\Delta df = m - (m-2) = 2$)
  - $H_0 : Y = X_0\beta_0 + \epsilon,\ \beta_0 \in \mathbb{R}^{m-2}$
  - $H_1 : Y = X_1\beta_1 + \epsilon,\ \beta_1 = (\beta_0, \beta_+, \beta_*) \in \mathbb{R}^m$
  - i.e. Under H_0, $\beta_+ = \beta_* = 0$

$$S^2 = \frac{1}{n-m}\left\|Y - X_1\hat{\beta}_1\right\|_2^2, \qquad \frac{(n-m)S^2}{\sigma^2} \sim \chi^2_{n-m}$$

$$\frac{\hat{\beta}_+}{\sigma\sqrt{(X^TX)^{-1}_{++}}} \sim N(0,1) \qquad \frac{\hat{\beta}_*}{\sigma\sqrt{(X^TX)^{-1}_{**}}} \sim N(0,1)$$

Independence $\rightarrow$

$$\frac{N(0,1)^2 + N(0,1)^2}{\chi^2_{n-m}} = \frac{\chi^2_2}{\chi^2_{n-m}} = F_{2,n-m}$$

ANOVA

- a=runif(1000)
- b=runif(1000)
- c=runif(1000)
- y=a+b+c
- l0=lm(y~a)
- l1=lm(y~a+b+c)
- anova(l0,l1)

Regression towards the mean

- `library(UsingR); summary(father.son)`
- Exam results
- In R:
  - y~x
  - y~0+x
  - y=offset(x) [coefficient 1]
  - y=x+I(x^3) [I(·) "protects the inside]