

Biostat 250B HW3

Elvis Cui

Han Cui

Dept. of Biostat
UCLA



Reading material: (a) Seber and Lee's text, Chapter 6, sec 6.1-6.3, 6.5.

- 1 Refer to Seber and Lee's text, do Mis Ex. 6, #1.
- 2 On the webpage [https://en.wikipedia.org/wiki/Fieller's theorem](https://en.wikipedia.org/wiki/Fieller%27s_theorem), you will see a Fieller's interval estimate for the ratio of the means of two correlated random variables with finite variances. Use the same notations on the page and determine whether you agree with the $100(1-2\alpha)\%$ confidence interval for the ratio.
- 3 Find a 95% confidence interval for the unknown change point in a two straight lines regression.
- 4 Refer to the climate change paper modeled by a two phase regression and posted on the class website. Verify the statistic shown in equation (3) and the rejection rule on the right column below the line 'if'.
- 5 Refer to the handout on how to do a lack of fit test using STATA for the simple linear regression.
 A Use the data and verify that the SSPE is given by 3.036669.
 B Would a cubic polynomial fit well for the same data set? What about a quintic polynomial?
 Show your work.
- 6 Let Z_1, \dots, Z_k be i.i.d. standard normal variates, let $U \sim \chi_m^2(0)$ and U is independent of Z_1, \dots, Z_k . Define

$$M = \max_{1 \leq i \leq k} \frac{|Z_i|}{\sqrt{U/m}}.$$

We say that M has a studentized maximum modulus distribution and write $M \sim M_{k,m}$.

- i Use the studentized maximum modulus distribution to find simultaneous confidence intervals for the set of all $\mu_i = \theta + \alpha_i$ for an one way analysis of variance model with k groups and n observations in each group.
- ii Let a_1, a_2, \dots, a_k be a set of numbers. Is it true that $\max |a_i| \leq c$ if and only if $|\sum_{i=1}^k d_i a_i| \leq c \sum_{i=1}^k |d_i|$ for all numbers d_1, d_2, \dots, d_k ? Justify your answer.
- iii Use (b) and the studentized maximum modulus distribution to find a simultaneous set of confidence intervals for $\sum_{i=1}^k d_i \mu_i$.

Q1:

1. Let $F = \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 / S^2$, the F-statistic for testing $H : \beta_1 = 0$ for a straight line. Using the notation of Section 6.1.5, prove that

$$\hat{x}_0 - \bar{x} = \frac{F}{F + (n-2)} (\hat{x}_0 - \bar{x}).$$

(Hoadley [1970])

Sol. Define. $S_{XY} \equiv \sum_i (Y_i - \bar{Y})(X_i - \bar{X})$

$$S_{XX} \equiv \sum_i (X_i - \bar{X})^2.$$

$$S_{YY} \equiv \sum_i (Y_i - \bar{Y})^2.$$

$$\text{Then } \hat{x}_0 - \bar{x} = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1} = (y_0 - \bar{Y}) \frac{S_{XY}}{S_{XX}}$$

$$\hat{x}_0 - \bar{x} = (y_0 - \bar{Y}) S_{XY}/S_{YY}$$

$$(n-2) S^2 = \sum_i (Y_i - \bar{Y} - \frac{S_{XY}(X_i - \bar{X})}{S_{XX}})^2$$

$$= S_{YY} - \frac{S_{XY}^2}{S_{XX}}$$

$$F = \frac{S_{XY}^2}{S_{XX}^2} \frac{S_{XX}}{S^2} = S_{XY}^2 / (S_{XX} S^2)$$

$$\Rightarrow \frac{F}{F + (n-2)} (\hat{x}_0 - \bar{x}) = \frac{S_{XY} S_{XX}}{S_{XY}^2 + (n-2) S^2 S_{XX}} (y_0 - \bar{Y})$$

$$= \frac{S_{XY}}{S_{YY}} (y_0 - \bar{Y}) = \hat{x}_0 - \bar{x} \quad \square.$$

Q2: (Assume normality, otherwise t-dist) is ill-posed.

2. On the webpage https://en.wikipedia.org/wiki/Fieller's_theorem, you will see a Fieller's interval estimate for the ratio of the means of two correlated random variables with finite variances. Use the same notations on the page and determine whether you agree with the 100(1-2α)% confidence interval for the ratio.

Approximate confidence interval [\(edit\)](#)

Variables a and b may be measured in different units, so there is no way to directly combine the standard errors as they may also be in different units. The most complete discussion of this is given by Fieller (1954).^[1]

Fieller showed that if a and b are (possibly correlated) means of two samples with expectations μ_a and μ_b , and variances $\nu_{11}\sigma^2$ and $\nu_{22}\sigma^2$ and covariance $\nu_{12}\sigma^2$, and if $\nu_{11}, \nu_{12}, \nu_{22}$ are all known, then a $(1-\alpha)$ confidence interval (m_L, m_U) for μ_a/μ_b is given by

$$(m_L, m_U) = \frac{1}{(1-g)} \left[\frac{a}{b} - \frac{g\nu_{12}}{\nu_{22}} + \frac{t_{r,\alpha} s^2}{b} \sqrt{\nu_{11} - 2\frac{a}{b}\nu_{12} + \frac{a^2}{b^2}\nu_{22} - g \left(\nu_{11} - \frac{\nu_{12}^2}{\nu_{22}} \right)} \right]$$

where

$$g = \frac{t_{r,\alpha}^2 s^2 \nu_{22}}{b^2}.$$

Here s^2 is an unbiased estimator of σ^2 based on r degrees of freedom, and $t_{r,\alpha}$ is the α -level deviate from the Student's t-distribution based on r degrees of freedom.

Sol. $a \sim N(\mu_a, \nu_{11} \sigma^2)$,

$b \sim N(\mu_b, \nu_{22} \sigma^2)$

WLOG, $\text{Cov}(a, b) \equiv \nu_{12} \sigma^2$, $\Theta \equiv \frac{\mu_a}{\mu_b}$

Then $a - \Theta b \sim N(0, d)$, $\frac{rS^2}{\sigma^2} \sim \chi_r^2$

$$d = \nu_{11} \sigma^2 + \nu_{22} \Theta^2 \sigma^2 - 2\Theta \nu_{12} \sigma^2$$

$$\Rightarrow \frac{a - \Theta b}{\sqrt{S^2/\sigma^2}} \frac{1}{6\sqrt{\nu_{11} + \nu_{22} \Theta^2 + \nu_{12} \Theta}} \sim t_r \left(\frac{S^2/\sigma^2}{a - \frac{\mu_a}{\mu_b} b} \right)$$

$$\Rightarrow P\left(\left(\frac{a - \Theta b}{S \sqrt{\nu_{11} + \nu_{22} \Theta^2 + \nu_{12} \Theta}}\right)^2 \leq t_{r,\alpha}^2\right) = 1 - \alpha$$

\Rightarrow Solving for $(g = \frac{t_{r,\alpha}^2 S^2 \nu_{22}}{b^2})$

$$(a - \Theta b)^2 = t_{r,\alpha}^2 S^2 (\nu_{11} + \nu_{22} \Theta^2 + \nu_{12} \Theta)$$

$$= \frac{gb^2 \nu_{11}}{\nu_{22}} - \frac{2gb^2 \nu_{12} \Theta}{\nu_{22}} + gb^2 \Theta^2$$

which Simplifies to

$$(1-g)b^2 \Theta^2 - (2ab - \frac{gb^2}{\nu_{22}} \nu_{12}) \Theta + a^2 - \frac{gb^2 \nu_{11}}{\nu_{22}} = 0$$

or $(b \neq 0)$.

$$(1-g)\Theta^2 - \left(\frac{2a}{b} - \frac{2g\nu_{12}}{\nu_{22}}\right)\Theta + a^2 - \frac{g\nu_{11}}{\nu_{22}} = 0$$

$$m_U = \Theta_1 = \frac{1}{1-g} \left(\frac{a}{b} - \frac{g\nu_{12}}{\nu_{22}} + \sqrt{\left(\frac{a}{b} - \frac{g\nu_{12}}{\nu_{22}} \right)^2 - (1-g)(a^2 - \frac{g\nu_{11}}{\nu_{22}})} \right)$$

$$m_L = \Theta_2 = \frac{1}{1-g} \left(\frac{a}{b} - \frac{g\nu_{12}}{\nu_{22}} - \sqrt{\left(\frac{a}{b} - \frac{g\nu_{12}}{\nu_{22}} \right)^2 - (1-g)(a^2 - \frac{g\nu_{11}}{\nu_{22}})} \right)$$

which are the same as those given in wikipedia.

$$Q3. \text{ Let } \hat{\gamma} = -\frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\hat{\beta}_1 - \hat{\beta}_2}, \quad \gamma = -\frac{\alpha_1 - \alpha_2}{\beta_1 - \beta_2}.$$

$$U = (\hat{\alpha}_1 - \hat{\alpha}_2) + \gamma(\hat{\beta}_1 - \hat{\beta}_2)$$

Then we have

$$\mathbb{E} U = \alpha_1 - \alpha_2 + \gamma(\beta_1 - \beta_2) = 0$$

$$\begin{aligned} \text{Var}(U) &= \text{Var}(\hat{\alpha}_1 + \hat{\beta}_1 \gamma) + \text{Var}(\hat{\alpha}_2 + \hat{\beta}_2 \gamma) \\ &= S^2 \left(\frac{1}{n_1} + \frac{(\bar{x}_1 - \gamma)^2}{S_{xx_1}} + \frac{1}{n_2} + \frac{(\bar{x}_2 - \gamma)^2}{S_{xx_2}} \right) \end{aligned}$$

$$\text{where } \bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}; \quad \bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2i}$$

$$S_{xx_1} = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2; \quad S_{xx_2} = \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2$$

$$\text{Let } \omega = \frac{1}{n_1} + (\bar{x}_1 - \gamma)^2 / S_{xx_1} + \frac{1}{n_2} + (\bar{x}_2 - \gamma)^2 / S_{xx_2}$$

$$\text{Then } \text{Var}(U) \equiv S^2 \omega.$$

$$\widehat{\text{Var}(U)} = S^2 \omega$$

$$\Rightarrow \frac{U}{\sqrt{S^2 \omega}} \sim t_{n_1+n_2-4} \quad \text{or,}$$

$$\frac{U^2}{S^2 \omega} \sim F_{1, n_1+n_2-4}$$

Solving for $(\alpha = 0.05)$

$$(\hat{\alpha}_1 - \hat{\alpha}_2 + \gamma(\hat{\beta}_1 - \hat{\beta}_2))^2 = F_{1, n_1+n_2-4} \cdot S^2 \omega$$

gives γ_L & γ_U .

(ω is a function of γ , this is a quadratic equation.)

$$Q4: U = \frac{S_0 - S}{S} \cdot (n-4)$$

$$S_0 \text{ is from } T_i = a_0 + b_0 i + b(i-c) \text{IND}_{C(i)} + e_i$$

$$S \text{ is from } T_i = a_0 + b_0 i + b(i-c) \text{IND}_c(i) + e_i$$

Note S_0 is conditioned on $C = C_0$.

Note: the null model can be written as

$$T = a_0 I_1 + b_0 I_1 + b I_2 + e$$

$$\text{where } I_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad I_2 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ n-c \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

Then $\text{rank}([I_1, I_2]) = 2$ full col-rank.

$$\text{Thus, } (n-3) S^2 / 6^2 \sim \chi^2_{n-3}$$

Also, the full model has 4 parameters so asymptotically,

$$(n-4) S^2 / 6^2 \stackrel{*}{\sim} \chi^2_{n-4}.$$

To sum up,

$$\frac{S_0 - S}{S} \cdot (n-4) \xrightarrow{*} F_{1, n-4}$$

So at sig-level $(1-\alpha)$, reject H_0 if

$$\frac{S_0 - S}{S} \cdot (n-4) \leq F_{1, n-4} (1-\alpha)$$

Q 5:

```

1 x <- c(1,1,2,3.3,3.3,4,4,4,4.7,5,5.6,5.6,5.6,6,6,6.5,6.9)
2 y <- c(2.3,1.8,2.8,1.8,3.7,2.6,2.6,2.2,3.2,2,3.5,2.8,2.1,3.4,3.2,3.4,5)
3
4 data <- as.data.frame(cbind(x, y))
5
6 mod1 <- lm(y~x, data=data)
7 mod2 <- lm(y~0+factor(x), data=data)
8
9 #(a)
10 anova(mod1, mod2)
11
12 #(b)
13 mod3 <- lm(y~poly(x, 3, raw = TRUE), data = data)
14 mod4 <- lm(y~poly(x, 4, raw = TRUE), data = data)
15
16 anova(mod3, mod2)
17 anova(mod4, mod2)

```

> anova(mod1, mod2)

Analysis of Variance Table

Model 1: $y \sim x$

Model 2: $y \sim 0 + \text{factor}(x)$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	15	7.4150				
2	7	3.0367	8	4.3783	1.2616	0.3861

> anova(mod3, mod2)

Analysis of Variance Table

Model 1: $y \sim \text{poly}(x, 3, \text{raw} = \text{TRUE})$

Model 2: $y \sim 0 + \text{factor}(x)$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	13	4.2043				
2	7	3.0367	6	1.1676	0.4486	0.826

> anova(mod4, mod2)

Analysis of Variance Table

Model 1: $y \sim \text{poly}(x, 4, \text{raw} = \text{TRUE})$

Model 2: $y \sim 0 + \text{factor}(x)$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	12	4.1937				
2	7	3.0367	5	1.157	0.5334	0.7465

SSPE

rd
3 & 4th
order
regression
are no
better than

the full model.

Q6:

6 Let Z_1, \dots, Z_k be i.i.d. standard normal variates, let $U \sim \chi_m^2(0)$ and U is independent of Z_1, \dots, Z_k . Define

$$M = \max_{1 \leq i \leq k} \frac{|Z_i|}{\sqrt{U/m}}.$$

We say that M has a studentized maximum modulus distribution and write $M \sim M_{k,m}$.

- i Use the studentized maximum modulus distribution to find simultaneous confidence intervals for the set of all $\mu_i = \theta + \alpha_i$ for an one way analysis of variance model with k groups and n observations in each group.
- ii Let a_1, a_2, \dots, a_k be a set of numbers. Is it true that $\max |a_i| \leq c$ if and only if $|\sum_{i=1}^k d_i a_i| \leq c \sum_{i=1}^k |d_i|$ for all numbers d_1, d_2, \dots, d_k ? Justify your answer.
- iii Use (b) and the studentized maximum modulus distribution to find a simultaneous set of confidence intervals for $\sum_{i=1}^k d_i \mu_i$.

Sol. (i): Let $\bar{Y}_{i \cdot} = \frac{1}{n} \sum_{j=1}^n Y_{ij}$ be the sample mean of i^{th} group. Then

$$P\left(\max_{1 \leq i \leq k} \frac{|\bar{Y}_{i \cdot} - \mu_i|}{\sqrt{S^2}} \leq M_{k,m}^\alpha\right) = 1 - \alpha$$

$$\text{LHS} = P(\mu_i \in \bar{Y}_{i \cdot} \pm M_{k,m}^\alpha S) = 1 - \alpha$$

\Rightarrow A $(1-\alpha) \times 100\%$ CI for all μ_i is $(\max(\bar{Y}_{i \cdot} - M_{k,m}^\alpha S, 0), \bar{Y}_{i \cdot} + M_{k,m}^\alpha S)$.

$\forall d_1, \dots, d_k$

$$(ii) \max_{1 \leq i \leq k} |a_i| \leq c \text{ iff } \left| \sum_{i=1}^k d_i a_i \right| \leq c \sum_{i=1}^k |d_i|$$

(\Leftarrow) : Let $d_i = 1$ sequentially, we have

$$|a_i| \leq c \quad \forall i = 1, \dots, k$$

$$\Rightarrow \max_{1 \leq i \leq k} |a_i| \leq c.$$

(\Rightarrow) : This follows from Hölder's ineq.:

$$\left| \sum_{i=1}^k d_i a_i \right| \leq \left(\sum_{i=1}^k |d_i| \right) \left(\max_{1 \leq i \leq k} |a_i| \right)$$

$$\leq C \left(\sum_{i=1}^k |d_i| \right)$$

$$(i) \max_{1 \leq i \leq k} |\bar{Y}_{i \cdot} - \mu_i| \leq S \cdot M_{k,m}^\alpha$$

iff

$$\left| \sum_{i=1}^k d_i (\bar{Y}_{i \cdot} - \mu_i) \right| \leq S \cdot M_{k,m}^\alpha \left(\sum_{i=1}^k |d_i| \right)$$

\Rightarrow

$$P\left(\max_{1 \leq i \leq k} |\bar{Y}_{i \cdot} - \mu_i| \leq S \cdot M_{k,m}^\alpha\right)$$

$$= P\left(\left| \sum_{i=1}^k d_i (\bar{Y}_{i \cdot} - \mu_i) \right| \leq S \cdot M_{k,m}^\alpha \left(\sum_{i=1}^k |d_i| \right)\right)$$

$$= 1 - \alpha$$

\Rightarrow

A simultaneous set of CIs is

$$\sum_{i=1}^k d_i \mu_i \in \sum_{i=1}^k d_i \bar{Y}_{i \cdot} \pm S M_{k,m}^\alpha \left(\sum_{i=1}^k |d_i| \right) \quad \square$$