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 , 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.

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

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

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 $a_1, a_2, \ldots, a_k$ be a set of numbers. Is it true that $\max |a_i| \le c$ if and only if $|\sum_{i=1}^{k} d_i a_i| \le c \sum_{i=1}^{k} |d_i|$ for all numbers $d_1, d_2, \ldots, d_k$? Justify your answer.

iii Use (b) and the studentized maximum modulus distribution to find a simultaneous set of confidence intervals for $\sum_{i=1}^{k} d_i \mu_i$.

48/60

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

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

(Hoadley [1970])

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

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

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

Then $\hat{X}_0 - \bar{X} = \frac{Y_0 - \hat{\beta}_0}{\hat{\beta}_1} = (Y_0 - \bar{Y}) \frac{S_{XX}}{S_{XY}}$

$\tilde{X}_0 - \bar{X} = (Y_0 - \bar{Y}) S_{XY} / S_{YY}$

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

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

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

$\Rightarrow \frac{F}{F + (n-2)} (\hat{X}_0 - \bar{X}) = \frac{S_{XY} S_{XX}}{S_{XY}^2 + (n-2) S^2 S_{XX}} (Y_0 - \bar{Y})$

$= \frac{S_{XY}}{S_{YY}} (Y_0 - \bar{Y}) = \hat{X}_0 - \bar{X}$ ✓ □.

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 $v_{11}\sigma^2$ and $v_{22}\sigma^2$ and covariance $v_{12}\sigma^2$, and if v_{11}, v_{12}, v_{22} are all known, then a $(1-\alpha)$ confidence interval (m_L, m_U) for μ_a/μ_b is given by

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

where

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

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 $a \sim N(\mu_a, v_{11}\sigma^2)$,

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

WLOG, $Cov(a,b) \equiv v_{12}\sigma^2$, $\theta \equiv \frac{\mu_a}{\mu_b}$

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

$d = v_{11}\sigma^2 + v_{22}\theta^2\sigma^2 - 2\theta v_{12}\sigma^2$

$\Rightarrow \frac{a - \theta b}{\sqrt{s^2/\sigma^2}} \frac{1}{\sigma\sqrt{v_{11} + v_{22}\theta^2 + 2v_{12}\theta}} \sim t_r \left(s^2 \perp\!\!\!\perp a - \frac{\mu_a}{\mu_b} b \right)$

$\Rightarrow P\left(\left(\frac{a - \theta b}{s\sqrt{v_{11} - 2v_{12}\theta + v_{22}\theta^2}} \right)^2 \le t_{r,\alpha}^2 \right) = 1 - \alpha$

$\Rightarrow$ Solving for $\left(g = \frac{t_{r,\alpha}^2 s^2 v_{22}}{b^2} \right)$

$(a - \theta b)^2 \le t_{r,\alpha}^2 s^2 (v_{11} - 2v_{12}\theta + v_{22}\theta^2)$

$= \frac{g b^2 v_{11}}{v_{22}} - \frac{2g b^2 v_{12}}{v_{22}}\theta + g b^2 \theta^2$

which simplies to

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

or $(b \neq 0)$

$(1-g)\theta^2 - (\frac{2a}{b} - \frac{2g v_{12}}{v_{22}})\theta + a^2 - \frac{g v_{11}}{v_{22}} = 0$

$\Rightarrow$

$m_U = \theta_1 = \frac{1}{1-g}\left(\frac{a}{b} - \frac{g v_{12}}{v_{22}} + \sqrt{\left(\frac{a}{b} - \frac{g v_{12}}{v_{22}} \right)^2 - (1-g)(a^2 - \frac{g v_{11}}{v_{22}})} \right)$

$m_L = \theta_2 = \frac{1}{1-g}\left(\frac{a}{b} - \frac{g v_{12}}{v_{22}} - \sqrt{\left(\frac{a}{b} - \frac{g v_{12}}{v_{22}} \right)^2 - (1-g)(a^2 - \frac{g v_{11}}{v_{22}})} \right)$

which are the same as those given in wiki-pedia. ✓

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

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

Then we have

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

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

$$= \sigma^2\left(\frac{1}{n_1} + \frac{(\bar{x}_{1\cdot} - \gamma)^2}{S_{xx_1}} + \frac{1}{n_2} + \frac{(\bar{x}_{2\cdot} - \gamma)^2}{S_{xx_2}}\right)$$

where $\bar{x}_{1\cdot} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}$; $\bar{x}_{2\cdot} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$

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

Let $w = \frac{1}{n_1} + (\bar{x}_{1\cdot} - \gamma)^2/S_{xx_1} + \frac{1}{n_2} + (\bar{x}_{2\cdot} - \gamma)^2/S_{xx_2}$

Then $\text{Var } U \equiv \sigma^2 w$.

$$\widehat{\text{Var}}\, U = s^2 w$$

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

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

Solving for ($\alpha = 0.05$)

$$(\hat{\alpha}_1 - \hat{\alpha}_2 + \gamma(\hat{\beta}_1 - \hat{\beta}_2))^2 = F^{\alpha}_{1, n_1+n_2-4}\, s^2 w$$

gives γ_L & γ_U.

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

What is the final answer? -2

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

S_0 is from $T_i = a_0 + b_0 i + b(i-c_0)IND_{c_0}(i) + e_i$

S is from $T_i = a_0 + b_0 i + b(i-c)IND_c(i) + e_i$

Note S_0 is conditioned on $c = c_0$.

Note: the null model can be written as

$$T = a_0 \mathbf{1} + b_0 I_1 + b I_2 + e$$

where $I_1 = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \end{bmatrix}$ $I_2 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ n-c \end{bmatrix}$ $e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$

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

Thus, $(n-3)S_0^2/\sigma^2 \sim \chi^2_{n-3}$

Also, the full model has 4 parameters so asymptotically,

$$(n-4)S^2/\sigma^2 \overset{\cdot}{\sim} \chi^2_{n-4}.$$

To sum up,

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

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

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

This is a test of $H_0: c = c_0$

Not the correct eqn.

Supposed to derive eqn. (3)

-5

and

$$\frac{\hat{b}}{\sqrt{s^2\left(\frac{1}{c_r} + \frac{1}{c_{r+1}}\right)\cdot\frac{1}{n-4}}} \le t_{n-4}(1-\alpha)$$

Q5:

```r
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)
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)

data <- as.data.frame(cbind(x, y))

mod1 <- lm(y~x, data=data)
mod2 <- lm(y~0+factor(x), data=data)

#(a)
anova(mod1, mod2)

#(b)
mod3 <- lm(y~poly(x, 3, raw = TRUE), data = data)
mod4 <- lm(y~poly(x, 4, raw = TRUE), data = data)

anova(mod3, mod2)
anova(mod4, mod2)
```

```
> anova(mod1, mod2)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ 0 + 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 ~ poly(x, 3, raw = TRUE)
Model 2: y ~ 0 + 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 ~ poly(x, 4, raw = TRUE)
Model 2: y ~ 0 + 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

3rd & 4th order regression are no better than the full model.

suppose to be 5th order model

Look at fit test for 3rd & 5th order models?

– 2

Q6:

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

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

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 $a_1, a_2, \ldots, a_k$ be a set of numbers. Is it true that $\max |a_i| \le c$ if and only if $|\sum_{i=1}^k d_i a_i| \le c \sum_{i=1}^k |d_i|$ for all numbers $d_1, d_2, \ldots, d_k$? 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 $Var[\bar{Y}_{i\cdot}] = \frac{\sigma^2}{n}$

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

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

$\Rightarrow$ A $(1-\alpha)\times 100\%$ CI for all μ_i is

$$\left(\max(\bar{Y}_{i\cdot} - M^{\alpha}_{k,m}S, 0),\ \bar{Y}_{i\cdot} + M^{\alpha}_{k,m}S\right) \quad -3 \text{ not correct}$$

(ii) $\max_{1\le i\le k} |a_i| \le c$ iff $|\sum_{i=1}^k d_i a_i| \le c \sum_{i=1}^k |d_i|$ $\forall d_1, \ldots, d_k$

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

$$|a_i| \le c \quad \forall i = 1, \ldots, k$$

$$\Rightarrow \max_{1\le i\le k} |a_i| \le c. \quad \checkmark$$

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

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

$$\le c\left(\sum_{i=1}^k |d_i|\right) \quad \checkmark$$

(iii): $\max |\bar{Y}_{i\cdot} - \mu_i| \le S \cdot M^{\alpha}_{k,m}$

iff

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

$\Rightarrow$

$$P\left(\max |\bar{Y}_{i\cdot} - \mu_i| \le S \cdot M^{\alpha}_{k,m}\right)$$

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

$$= 1-\alpha$$

$\Rightarrow$

A simultaneous set of CIs is

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