1. M6/1 Let $F = \hat{\beta}_1^2 \Sigma_i (x_i - \bar{x})^2 / s^2$, the F-statistic for testing $H: \beta_1 = 0$ for a straight line. using the notation of Section 6.1.5, prove that $\tilde{x}_0 - \bar{x} = \frac{F}{F + (n-2)} (\hat{x}_0 - \bar{x})$

pf: $\hat{x}_0 = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1} = \bar{x} + \frac{y_0 - \bar{y}}{\hat{\beta}_1}$, $\hat{\beta}_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$

$\tilde{x}_0$: inverse estimate $X \sim a_0 + Y a_1 + \varepsilon$

then: $\tilde{a}_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (y_i - \bar{y})^2}$, $\tilde{a}_0 = \bar{x} - \bar{y} \cdot \tilde{a}_1$

$\Rightarrow \tilde{x}_0 = \tilde{a}_0 + \tilde{a}_1 y_0$

$= \bar{x} - \bar{y} \tilde{a}_1 + y_0 \tilde{a}_1$

$= \bar{x} + \tilde{a}_1 (y_0 - \bar{y})$

$\Rightarrow \tilde{x}_0 - \bar{x} = (y_0 - \bar{y}) \cdot \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (y_i - \bar{y})^2}$

and $\hat{x}_0 - \bar{x} = (y_0 - \bar{y}) \Big/ \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$

$\Rightarrow \frac{(\tilde{x}_0 - \bar{x})}{(\hat{x}_0 - \bar{x})} = \frac{(\Sigma (x_i - \bar{x})(y_i - \bar{y}))^2}{\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2}$

that is the prove: $\frac{F}{F + (n-2)} = \frac{(\Sigma (x_i - \bar{x})(y_i - \bar{y}))^2}{\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2}$

LHS $= \frac{\hat{\beta}_1^2 \Sigma (x_i - \bar{x})^2}{\hat{\beta}_1^2 \Sigma (x_i - \bar{x})^2 + (n-2) \cdot s^2}$

$= \frac{(\Sigma (y_i - \bar{y})(x_i - \bar{x}))^2}{(\Sigma (y_i - \bar{y})(x_i - \bar{x}))^2 + (n-2) \cdot s^2 \cdot \Sigma (x_i - \bar{x})^2}$

$= \frac{(\Sigma (y_i - \bar{y})(x_i - \bar{x}))^2}{(\Sigma (y_i - \bar{y})(x_i - \bar{x}))^2 + \Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \hat{y})^2}$

$= \frac{(\Sigma (y_i - \bar{y})(x_i - \bar{x}))^2}{(\Sigma (y_i - \bar{y})(x_i - \bar{x}))^2 + \Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2 - (\Sigma (y_i - \bar{y})(x_i - \bar{x}))^2}$

$= \frac{(\Sigma (y_i - \bar{y})(x_i - \bar{x}))^2}{\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2}$

$=$ RHS

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

2. Derive the Fieller's $100(1-2\alpha)\%$ interval estimate for the ratio of means of two correlated random variables with finite variances; using the notation from the wikipedia page

Notation: $(m_L, m_U) = \frac{1}{(1-g)}\left(\frac{a}{b} - \frac{gv_{12}}{v_{22}} \pm \frac{t_{r,\alpha} s}{b}\left(v_{11} - 2\frac{a}{b}v_{12} + \frac{a^2}{b^2}v_{22} - g\left(v_{11} - \frac{v_{12}^2}{v_{22}}\right)\right)^{1/2}\right)$, for $\frac{\mu_a}{\mu_b}$

where $A \sim f(\mu_a, v_{11}\sigma^2), B \sim f(\mu_b, v_{22}\sigma^2)$, $cov(a,b) = \sigma^2 v_{12}$

$a = \bar{A}$, $b = \bar{B}$, $g = \frac{t^2_{r,\alpha} s^2 v_{22}}{b^2}$, s^2 is the unbiased estimater for σ^2 with $df = r$

pf: let $\theta = \frac{\mu_a}{\mu_b}$, $W = A - \theta B$

then: $E(W) = E(A) - E(\theta B) = \mu_a - \mu_a = 0$

$$Var(W) = var(A) + \theta^2 var(B) - 2\theta cov(A,B) = \sigma^2(v_{11} + \theta^2 v_{22} - 2\theta v_{12})$$

thus: $W = A - \theta B \sim N(0, \sigma^2(v_{11} - \theta^2 v_{22} - 2\theta v_{12}))$

and: $\frac{rs^2}{\sigma^2} \sim \chi^2_r$

$$\Rightarrow \frac{(A-\theta B)/\sigma(v_{11} + \theta^2 v_{22} - 2\theta v_{12})^{1/2}}{\sqrt{r}\, S/\sigma} \sim t_r$$

$$\frac{A - \theta B}{(r\cdot(v_{11} + \theta^2 v_{22} - 2\theta v_{12}))^{1/2}\cdot S} \sim t_r$$

therefore: the $(1-2\alpha)\cdot CI$ for $(A-\theta B)^2$ is $(t_{r,\alpha})^2 S^2(v_{11} + v_{22}\theta^2 - 2\theta v_{12})$

plug in estimation $a = \bar{A}$, $b = \bar{B}$

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

$$\theta^2 - \frac{2a}{b}\theta + \frac{a^2}{b^2} = \frac{(t_{r,\alpha}\cdot S)^2}{b^2}(v_{11} + v_{22}\theta^2 - 2\theta v_{12})$$

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

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

$$\theta = \frac{1}{2(1-g)}\left[-2\frac{gv_{12}}{v_{22}} + 2\frac{a}{b} \pm \left(4\left(\frac{gv_{12}}{v_{22}} - \frac{a}{b}\right)^2 - 4(1-g)\left(\frac{a^2}{b^2} - g\frac{v_{11}}{v_{22}}\right)\right)^{1/2}\right]$$

$$= \frac{1}{(1-g)}\left[\frac{a}{b} - \frac{gv_{12}}{v_{22}} \pm \frac{1}{2}\left(\frac{4g^2v_{12}^2}{v_{22}^2} + \frac{4a^2}{b^2} - 8\frac{gv_{12}\cdot a}{v_{22}\cdot b} - 4(1-g)\frac{a^2}{b^2} + 4(g-g^2)\frac{v_{11}}{v_{22}}\right)^{1/2}\right]$$

$$= \frac{1}{(1-g)}\left[\frac{a}{b} - \frac{gv_{12}}{v_{22}} \pm \frac{2}{2}\left((g-g^2)\frac{v_{11}}{v_{22}} - 2\frac{a}{b}\cdot\frac{gv_{12}}{v_{22}} + \frac{a^2}{b^2}g + g^2\frac{v_{12}^2}{v_{22}}\right)^{1/2}\right]$$

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

therefore: $CI = \frac{1}{(1-g)}\left[\frac{g}{b} - \frac{gv_{12}}{v_{22}} \pm \left(\frac{tr.a \cdot s}{b}\right)\left((1-g)v_{11} - 2\frac{g}{b}v_{12} + \frac{g^2}{b^2}v_{22} + g\frac{v_{12}^2}{v_{22}}\right)^{1/2}\right]$

$= \frac{1}{(1-g)}\left[\frac{g}{b} - \frac{gv_{12}}{v_{22}} \pm \left(\frac{tr.a \cdot s}{b}\right)\left(v_{11} - 2\frac{g}{b}v_{12} + \frac{g^2}{b^2}v_{22} - g\left(v_{11} - \frac{v_{12}^2}{v_{22}}\right)\right)^{1/2}\right]$

The CI is exactly the same as those from wikipedia

3. Find a 95% confidence interval for the unknown change point in a two straight line regression.

Soln: $\hat{\gamma} = -\frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\hat{\beta}_1 - \hat{\beta}_2}$

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

$E(u) = 0$

$Var(u) = \sigma^2\left\{\frac{1}{n_1} + \frac{(\bar{x}_1 - \gamma)^2}{\Sigma(x_{1i} - \bar{x}_1)^2} + \frac{1}{n_2} + \frac{(\bar{x}_2 - \gamma)^2}{\Sigma(x_{2i} - \bar{x}_2)}\right\}$

$\hat{=} \sigma^2 \cdot W$

then: 95% CI for u^2: $F_{1,n-q}^{0.95} S^2 W$

$(\hat{\alpha}_1 - \hat{\alpha}_2 + \gamma(\hat{\beta}_1 - \hat{\beta}_2))^2 = F_{1,n-q}^{.05} S^2 \cdot W$

$\Rightarrow \left[(\hat{\beta}_1 - \hat{\beta}_2)^2 - F_{1,n-q}^{.05} S^2\left(\left[\Sigma(x_{1i} - \bar{x}_1)^2\right]^{-1} + \left[\Sigma(x_{2i} - \bar{x}_2)^2\right]^{-1}\right)\right]\gamma^2$

$+ 2\left[(\hat{\alpha}_1 - \hat{\alpha}_2)(\hat{\beta}_1 - \hat{\beta}_2) + F_{1,n-q}^{.05} S^2\left(\frac{\bar{x}_1}{\Sigma(x_{1i} - \bar{x}_1)^2} + \frac{\bar{x}_2}{\Sigma(x_{2i} - \bar{x}_2)}\right)\right]\cdot\gamma$

$+ \left(1 - F_{1,n-q}^{.05} S^2\right)\left(\frac{\bar{x}_1^2}{\Sigma(x_{1i} - \bar{x}_1)^2} + \frac{\bar{x}_2^2}{\Sigma(x_{2i} - \bar{x}_2)^2} + \frac{1}{n_1} + \frac{1}{n_2}\right) + (\hat{\alpha}_1 - \hat{\alpha}_2)^2 = 0$

for simple notation: write as $c\gamma^2 + 2D\gamma + E = 0$

then 95% CI for γ is: $\left(-\frac{D}{c} \pm \frac{\sqrt{D^2 - CE}}{c}\right)$

4. Verify two results from the climate change paper.

Notation: Model: $Y(t) = a_0 + b_0 t + b(t-c)\cdot I(t>c) + \varepsilon_i$ (2)

$b = b_1 - b_0$

① Likelihood ratio test statistic: $U = \frac{(S_0 - S)/3}{S/(n-4)}$, S_0: residual sum of square from $H_0: Y(t) = a_0 + b_0 t + \varepsilon_i$

② CI for c is the c_0 s.t. $(S' - S)/(S/(n-4)) \le F_{1,n-4}(1-\alpha)$, S': residual sum of square from fitting (2) at $c = c_0$

reference: Inference in Two-Phase Regression, 1971, Hinkley (a)

Inference about the Intersection in Two-Phase Regression (b)

pf: ① write model as:

$$Y_i(t) = \begin{cases} a_0 + b_0 t + \varepsilon_i, & t \le c, \ 1 \le i \le I \\ a_1 + b_1 t + \varepsilon_i, & t > c, \ I < i \le N \end{cases} \qquad a_1 = a_0 + b_0 c - b_1 c$$

likelihood $L(y, t; a_0, b_0, a_1, b_1, c, I, \sigma)$

$$= (2\pi\sigma^2)^{-\frac{N}{2}} \cdot \exp\left\{ -\frac{1}{2\sigma^2}\sum_{j=1}^{I}(y_j - a_0 - b_0 t_j)^2 - \frac{1}{2\sigma^2}\sum_{j=I+1}^{N}(y_j - a_1 - b_1 t_j)^2 \right\}$$

$L(c) = L_I(c) \quad (t_I \le c < t_{I+1};\ I = 2, \cdots N-2)$

$L_I(\chi) = (2\pi\sigma^2)^{-\frac{N}{2}} \cdot \exp\left\{ -\frac{1}{2\sigma^2} S_I^2(\chi) \right\}$, where $S_I^2(y)$ is the residual sum of for two regression lines constrained at meet at $t = \chi$

For the residual sum of squares of a single regression:

$$S_0^2 = \sum_{j=1}^{N}(y_j - \bar{y}_N)^2 - \left[\sum_{j=1}^{N}(t_j - \bar{t})(y_j - \bar{y})\right]^2 \Big/ \sum_{j=1}^{N}(t_j - \bar{t})$$

let $Z_I^2(\chi) = S_0^2 - S_I^2(\chi)$, then the overall maximum, $Z_{\hat{I}}^2(c)$ is the likelihood ratio statistic for $H_0: \beta_0 = \beta_1$ (b)/ch 1.

test statistic: from MLR theories: $S = S_{\hat{I}}^2(c)$ follows $S/\sigma^2 \sim \chi^2(n-4)$

from strong empirical evidence, (though not theoretically proved),

$S_0 - S = Z_{\hat{I}}^2(c) \sim \chi^2(3)$ (a)/ 2.3

$\Rightarrow$ Test statistic $U = \frac{(S_0 - S)/3}{S/(n-4)} \sim F_{3,(n-4)}$ under $H_0: \beta_0 = \beta_1$

② $H_0: c = c_0$.

likelihood ratio test statistic: $\Lambda = -2\log\{L(c_0)/L(\hat{c})\} \to \chi^2(1)$ (Wilk's thm)

$$= -\frac{1}{\sigma^2}\{z^2\hat{I}_0(c_0) - z^2\hat{I}(\hat{c})\}$$

$$= -\frac{1}{\sigma^2}(S_0 - S' - S_0 + S)$$

$$= \frac{1}{\sigma^2}(S' - S) \overset{\text{asymptotically}}{\sim} \chi^2(1)$$

meanwhile: $S/\sigma^2 \sim \chi^2(n-4)$

therefore: under $H_0: c = c_0$: $\dfrac{(S'-S)}{S/(n-4)} \dot{\sim} F_{1,(n-4)}$

the asymptotic $100(1-\alpha)\%$ CI for c_0 is $\left\{c: \dfrac{(S'-S)}{S/(n-4)} \le F_{1,(n-4)}(1-\alpha)\right\}$

5.

- a. Verify that SSPE = 3.036669

```
lof <- data.frame(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))
```

$$SSPE = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2$$

```
# calculate group mean
mean_y <- lof %>%
  group_by(x) %>%
  summarise(meany = mean(y), n = n())
# calculate SSPE
(lof %>%
  left_join(mean_y, by = c("x" = "x")) %>%
  mutate(err2 = (y - meany)^2) %>%
  summarise(sum(err2)))
```

```
##   sum(err2)
## 1  3.036667
```

- b. Would a cubic polynomial fit well for the same data set? What about a quintic polynomial?

```
lof <- lof %>%
  mutate(x2 = x^2) %>%
  mutate(x3 = x^3) %>%
  mutate(x4 = x^4) %>%
  mutate(x5 = x^5)
```

```
mod1 <- lm(y ~ x, data = lof)
mod3 <- lm(y ~ x + x2 + x3, data = lof)
mod5 <- lm(y ~ x + x2 + x3 + x4 +x5, data = lof)
```

```
anova(mod1, mod3)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ x + x2 + x3
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     15 7.4150
## 2     13 4.2043  2    3.2107 4.9639 0.02502 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod1, mod5)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ x + x2 + x3 + x4 + x5
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     15 7.4150
## 2     11 4.1306  4    3.2843 2.1866 0.1376
```

```r
anova(mod3, mod5)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x + x2 + x3
## Model 2: y ~ x + x2 + x3 + x4 + x5
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     13 4.2043
## 2     11 4.1306  2  0.073615 0.098 0.9074
```

The cubic polynomial has a significant better fit than the model with only linear terms. While for the quintic polynomial model, it has neither significantly better fit than the linear model nor than the cubic polynomial model.

6. $z_1 \cdots z_k \sim N(0,1)$, $u \sim \chi^2_m(0)$, $u \perp (z_1 \cdots z_k)$, $M \triangleq \max_{1\le i\le k} \frac{|z_i|}{\sqrt{u/m}}$

$M \sim$ studentized maximum modulus distribution; $M \sim M_{k,m}$

i) Use studentized maximum modulus distribution to find a simultaneous CI for the set of all $\mu_i = \theta + \alpha_i$ for one way ANOVA, k groups, n obs/group.

under H_0: $\alpha \sim N(\mu_\alpha, \Sigma)$, $\Sigma = \frac{1}{n}\sigma^2 I_k$.

let $\bar{x}_1 \cdots \bar{x}_k$ be the group sample mean

then $[(\bar{x}_1 - \mu_1) - (\bar{x}_2 - \mu_2)], \cdots [(\bar{x}_1 - \mu_1) - (\bar{x}_k - \mu_k)] \cdots [(\bar{x}_{k-1} - \mu_{k-1}) - (\bar{x}_k - \mu_k)] \overset{iid}{\sim} N(0, \sigma^2(\frac{1}{n} + \frac{1}{n}))$

have $k^* = \frac{k(k-1)}{2}$ such pair of group mean difference.

s^2; unbiased estimater for σ^2, $(nk-k)s^2/\sigma^2 \sim \chi^2_{nk-k}(0)$

therefore: $M = \max_{1\le i<j\le k} \frac{|(\bar{x}_{i\cdot} - \mu_i) - (\bar{x}_{j\cdot} - \mu_j)| / \sqrt{\frac{2}{n}}}{s} \sim m_{(k^*, nk-k)}$

(let $m^\alpha_{k,m}$ be the upper α quatile)

$\Rightarrow$ the simultaneous CI for $\mu_i - \mu_j$ $(1\le i<j\le k)$ is: $|\bar{x}_{i\cdot} - \bar{x}_{j\cdot}| \pm s\sqrt{\frac{2}{n}} \cdot m^\alpha_{(k^*, nk-k)}$

ii) $a_1 \cdots a_k$, a set of numbers; Is it true that $\max|a_i| \le c$ iff $|\sum_{i=1}^k d_i a_i| \le c \sum_{i=1}^k |d_i|$ for all numbers $d_1, d_2, \cdots d_k$?

$\max|a_i| \le c \iff |\sum_{i=1}^k d_i a_i| \le c \sum_{i=1}^k |d_i|$

$\Rightarrow$ $|\sum_{i=1}^k d_i a_i| \le \sum_{i=1}^k |d_i a_i| \le \sum_{i=1}^k |a_i| \cdot |d_i| \le \sum_{i=1}^k c \cdot |d_i| = c \cdot \sum_{i=1}^k |d_i|$

$\Leftarrow$ suppose $|\sum_{i=1}^k d_i a_i| \le c \cdot \sum_{i=1}^k |d_i|$ is true for any $\vec{d} = (d_1 \cdots d_k)$

the it holds for any $\vec{d} = \vec{e}_i = (0, \cdots \overset{i(0)}{1}, \cdots 0)$, $\forall i_0 = 1, 2, \cdots k$

$\Rightarrow |\sum_{i=1}^k d_i a_i| = |d_{i_0} a_{i_0}| = |a_{i_0}|$

$c \cdot \sum_{i=1}^k |d_i| = c \cdot |d_{i_0}| = c$

$\Rightarrow |a_{i_0}| \le c$, $\forall i_0 = 1, 2, \cdots k \Rightarrow \max|a_i| \le c$

Yes, this conclusion is true.

iii) Utilize (ii) and studentized maximum modulus distribution, find a simultaneous set of confidence interval for $\sum_{i=1}^k d_i \mu_i$.

Soln: $\overline{x_i} - \mu_i \overset{iid}{\sim} N(0, \sigma^2/n)$, $(nk-k)S^2/\sigma^2 \sim \chi^2_{(nk-k)}$

$\Rightarrow \max_{1\le i\le k} \frac{|\overline{x_i}-\mu_i|\sqrt{n}}{S} \sim M_{k,nk-k}$

then: $(1-a)\cdot 100\%$ CI for $\max|\overline{x_i}-\mu_i|$ is $\max|\overline{x_i}-\mu_i| \le M^{\alpha}_{k,nk-k}\cdot\frac{S}{\sqrt{n}}$

where $M^{\alpha}_{k,nk-k}$ is the upper α quantile

$\max|\overline{x_i}-\mu_i| \le c \iff \left|\sum_{i=1}^{k} a_i(\overline{x_i}-\mu_i)\right| \le c\cdot\sum_{i=1}^{k}|a_i|$

$\iff \left|\sum_{i=1}^{k} a_i\overline{x_i} - \sum_{i=1}^{k} a_i\mu_i\right| \le c\cdot\sum_{i=1}^{k}|a_i|$

$\Rightarrow 100(1-a)\%$ CI for $\sum_{i=1}^{k} a_i\mu_i$ is $\sum_{i=1}^{k} a_i\overline{x_i} \pm \frac{S}{\sqrt{n}} M^{\alpha}_{k,nk-k}\sum_{i=1}^{k}|a_i|$