Question 2

2010, MS-2

$n = 20$ patients. $n_1 = 10$ std
$n_2 = 10$ trt

a) Kaplan

MPH 2016

a) $E(\hat{\delta}_1 - \hat{\delta}_c) = E(\hat{\delta}_1) - E(\hat{\delta}_c)$

$= E(\bar{y}_f - \bar{y}_b) - E(\bar{x}_f - \bar{x}_b)$

$= (E\bar{y}_f - E\bar{y}_b) - (E\bar{x}_f - E\bar{x}_b)$

$= (\bar{Y}_f - \bar{Y}_b) - (\bar{X}_f - \bar{X}_b) = \delta_1 - \delta_c$

$\Rightarrow$ unbiased

b) S^2 is the same at baseline and F-U - for both samples

$Var(\hat{\delta}_1 - \hat{\delta}_c) = Var\hat{\delta}_1 + Var\hat{\delta}_c - 2cov(\delta, \delta_c)$

$cov(\bar{y}_f, \bar{y}_b) =$

$\rho = \frac{cov(x,y)}{\sqrt{Var x \cdot var y}}$

$\Rightarrow Var(\hat{\delta}_1) = Var(\bar{y}_f) + Var(\bar{y}_b) - 2cov(\bar{y}_f, \bar{y}_b)$

$= \frac{S^2}{n} + \frac{S^2}{n} - 2\frac{\rho_y\sqrt{S^2(S^2)}}{n}$

$=$

Question 4 | 2010, MS-2

$n = 25$ homes

data on PM_{10} taken $\Rightarrow$ use log

X_{Ii}, X_{oi} for i homes, the ln of PM_{10} [ ] indoors + outdoors

---

a) Estimate for the association btwn log outdoor and indoor PM_{10}

correlation $= \dfrac{cov(x,y)}{\sqrt{var x \cdot var y}}$

corr = β estimate of standardized, no int regression

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

$$= \frac{425 - 95\left(\frac{110}{25}\right) - 100\left(\frac{95}{25}\right) + \left(\frac{110}{25}\right)\left(\frac{95}{25}\right)}{\sqrt{\left(380 - 25\left(\frac{95}{25}\right)^2\right)\left(520 - 25\left(\frac{110}{25}\right)^2\right)}}$$ complex

$$= \frac{n(\Sigma xy) - \Sigma x \Sigma y}{\sqrt{(n\Sigma x^2 - (\Sigma x)^2)(n\Sigma y^2 - (\Sigma y)^2)}} = 0.2676$$

Is using correlation right?

test $\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$

$$\Sigma(X_i - \bar{X})^2$$
$$= \Sigma(X_i^2 - 2X_i\bar{X} + \bar{X}^2)$$
$$= \Sigma X_i^2 - 2n\bar{X}^2 + n\bar{X}^2$$
$$= \Sigma X_i^2 - n\bar{X}^2$$

b) $H_0: \mu_{out} = \mu_{in}$

$\bar{X}_{out} = \frac{110}{25} = 4.4$ $\quad S^2_{out} = \cancel{25*(520)} - \cancel{(110)^2} = 36$

$\bar{X}_{in} = \frac{95}{25} = 3.8$ $\quad S^2_{in} = \cancel{25 * 380 - (95)^2} = 19$

t-test for 2 sample means

$$\frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} =$$

* I'm assuming the samples are independent

~~* assuming homogeneity of variance and~~

* check what assumptions are necessary → seems like 2 cov should be required too.

c) median b/c skewed.

---

d) i. We would be missing potentially 3 low values of pollutant, and this could bias our results toward higher values

ii. use a 'lower bound' or code missing as 0 or some other low number.

Question 4/

2010, MS-2

e) 400 children . PM_{10} [ ] in homes (quartiles)
y/n recent wheezing

⇒ The one-sided p-value indicates that the proportion of children wheezing increases w/ increasing tertile ($p = 0.0398$, not <u>highly</u> significant).
(Two-sided test FTR).

---

f) dichotomized children.
estimate risk relative to other + 95% CI
based on this is the risk sign?

risk in $> 50 \mu g/m^3 = \frac{26}{112}$

$\leq 50\ mg/m^3 = \frac{29}{288}$

$RR = 2.3054$

$CI\ 95\% = (1.4233, 3.7342)$

⇒ limits do not include 1.
risk is significantly greater.

Question 3

2010, MS-2

High LDL is bad → drug to reduce

men/women

$n = 400$ $\quad n_1 = 200$ men — 100 drug, 100 none

$n_2 = 200$ wom — 100 drug, 100 none

ht, wt, age

a) $y_i = \underset{\text{cont}}{\beta_0 + \beta_1 age} + \beta_2 \underset{=1 \text{ if drug}}{trt} + \beta_3 \underset{=1 \text{ if men}}{gender} + \beta_4 \underbrace{trt \times gender}_{=1 \text{ if drug and men}}$

drug	gender	$E(y_i)$
$=0$	$=0$	$\beta_0 + \beta_1 (age) +$
$=0$	$=1$	$\beta_0 + \beta_1 (age) + \beta_3$
$=1$	$=0$	$\beta_0 + \beta_1 (age) + \beta_2$
$=1$	$=1$	$\beta_0 + \beta_1 (age) + \beta_2 + \beta_3 + \beta_4$

b) $\vec{y} = \vec{X}\vec{\beta} + \vec{e}$

i. $y = 400 \times 1$ $\quad X = 400 \times 5$ $\quad \beta = 5 \times 1$ $\quad e = 400 \times 1$

ii. $\hat{b} = (X'X)^{-1} X'y$

iii. $\hat{b} = (b, \sigma^2 (X'X)^{-1})$

c)

source	df	SS	MS	F-val	p-val
model	$p-1=4$	66308.4	16577.1	146.7	<0.0001
error	$n-p=395$	44635	113		
c. total	$n-1=399$				

$$F = \frac{MSM}{MSE}$$

---

d) Add ht, wt to model.

① ANOVA test compares (1) and (2) shows wt/ht are important →
⇒ means the larger model explains more variation in LDL levels

② insignificant Wald tests
→ both coefficients are not significantly diff from 0 in added last tests.

⇒ No, they don't necessarily contradict each other. The t-tests are assessing whether height and weight each add significant info to the model in the presence of all other variables, while the overall F-test is assessing whether the two variables jointly improve the overall model fit.

Question 3

2010, MS-2

Drug for LDL reduction.

n - 400

⇒ 200 men / 200 women

⇒ 100 drug, 100 placebo, 100 drug, 100 placebo

$$y = \beta_0 + \overset{1}{\text{age}} + \underset{=1 \text{ if drug}}{\overset{2}{\text{trt}}} + \underset{=1 \text{ if man}}{\overset{3}{\text{gender}}} + \overset{4}{\text{trt} * \text{gender}}$$

a)

drug	gender	$E(y_i)$
plac.	f	$\beta_0 + \beta_1$
plac.	m	$\beta_0 + \beta_1 + \beta_3$
drug.	f	$\beta_0 + \beta_1 + \beta_2$
drug.	m	$\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4$

b) i. $y = X\beta + \varepsilon$ $y = 400 \times 1$ $X = 400 \times 5$ $\beta = 5 \times 1$ $\varepsilon = 400 \times 1$

ii. $\hat{b} = (X'X)^{-1}X'y$

iii. $\hat{b} \sim N(b, \sigma^2 (X'X)^{-1})$

c)

Source	df	SS	MS (SS/df)	F-val (MSM/MSE)	p-val
model	4 (p-1)	66308.4	16577.1	146.7	<0.0001
error	395 (n-p)	44635	113		
C-total	399 (n-1)	110943.4			

d) Now add weight and height to the model

$y = \beta_1 + \beta_1 \text{age} + \beta_2 \text{trt} + \beta_3 g + \beta_4 \text{trt}*g + \beta_5 \text{wt} + \beta_6 \text{ht}$

test comparing 2 models ⇒ $p = 0.0098$ (wt/ht is important)

but wt and ht are insignificant in T3 testing

⇒ Does not contradict
the overall ANOVA test indicates that there is more info explained by the larger model overall.

wt and ht are not significant, after all the other vars are in the model.

wt is not <u>very</u> insignif. should be careful.

---

e) i. Is drug effect diff for m/f
⇒ yes. interaction is significant. (effect of drug differs by m/f)

ii. 'Drug has no effect in both males and females.' ⇒ INTERPRETING as: no effect of the drug overall

$H_0: \beta_2 = \beta_4 = 0$

$C = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$; $\theta_0 = (0, 0)^T$

iii. $E(y) = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{trt}) + \beta_5(\text{wt}) + \beta_6(\text{ht})$ ⇐ fem $(g=0)$

$E(y) = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{trt}) + \beta_3(g) + \beta_4(\text{trt}*g) + \beta_5(\text{wt}) + \beta_6(\text{ht})$

⇑ males $(g=1)$

f) No. conditional distr. $y|X=x$ matters.

e)

i. Drug effect diff btwn males / females?

$\Rightarrow$ yes, the interaction is significant ($p = 0.000619$)

ii. Drug has no effect in both males and females

$\Rightarrow$ drug has no effect

$\Rightarrow H_0: \beta_{drug} = \beta_{drug \times gender} = 0$

$$C = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad \theta_0 = (0\ 0)^T$$

iii. linear model for males / females

MALE $\Rightarrow E(y_i) = (\beta_0 + \beta_3) + \beta_1(age_i) + (\beta_2 + \beta_4)(trt) + \beta_5(wt) + \beta_6(ht)$

Female $E(y_i) = (\beta_0) + \beta_1(age) + \beta_2(trt) + \beta_5(wt) + \beta_6(ht)$

---

f) No. linear regression does not require any assumptions about the distribution of Y. It is important for the distribution of Y conditional on all the parameters to be normal, but that is not what is shown here.

2010, MS-7

Question 1

$$\vec{y} = \vec{X}\vec{\beta} + \vec{\varepsilon} \qquad \varepsilon_i \sim N(0, \sigma^2)$$

a)

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & -1 \\ 1 & 3 & -2 \\ 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

---

b) 2 (column rank).

Not full rank: $C1 - C2 = C3$

---

c) Re-express the model as $y = X^* \beta^* + \varepsilon$ where X^* is full rank

~~center around means?~~ ~~around a random value?~~

square x_2

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 4 \\ 1 & -1 & 4 \end{pmatrix}$$

d) $\theta_1 = \beta_1 - \beta_2$ estimable (?) To be estimable $\rightarrow$ $C = TX$

$$C = (0 \;\; 1 \;\; -1)_{a \times p} = T_{a \times n} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 4 \\ 1 & -1 & 4 \end{pmatrix}_{n \times p}$$

$T = (a_1 \; a_2 \; a_3 \; a_4)$

$$(0 \;\; 1 \;\; -1) = (a_1 + a_2 + a_3 + a_4 \, , \; a_1 + 2a_2 + 3a_3 - a_4 \, , \; a_2 + 4a_3 + 4a_4)$$

$$0 = a_1 + a_2 + a_3 + a_4$$
$$1 = a_1 + 2a_2 + 3a_3 - a_4$$
$$-1 = a_2 + 4a_3 + 4a_4$$

come up w/ some combo / matrix for a that works.

need $\Rightarrow$ x % t = c

$$0 = \sum_{i=1}^{4} a_i$$

c = t % x

e.g. $T = (1 \;\; -1 \;\; 0 \;\; 0) \Rightarrow \theta_1$ estimable

$\hat{\theta}_1 = c\hat{\beta}$ $\qquad \hat{\beta} = (X'X)^{-1}X'Y$

$\operatorname{Var}\hat{\theta}_1 = \sigma^2 c(X'X)^{-1}c'$ $\rightarrow$ MSE

how would you do this by hand?