Ann Marie Weideman,
2018 MS Exam

1 a)

HTN: $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.111}{0.031} \approx \boxed{3.58} > 1.96$

So, reject the null. Thus, there is evidence that a history of hypertension is associated with an increased probability of death.

Smoke: $t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{0.234}{0.072} = \boxed{3.25} > 1.96$

So, reject the null. Thus, there is evidence that a history of smoking is associated with an increased probability of death.

Age: $t = \frac{\hat{\beta}_3}{SE(\hat{\beta}_3)} = \frac{0.100}{0.100} = \boxed{1.00} < 1.96$

So, fail to reject the null. Thus, there is no evidence that advanced age is associated with an increased probability of death.

Remember, it increments by 5 yr.

b) $OR = \exp(d\hat{\beta}_3) = \exp(0.1 \cdot 2) \approx \boxed{1.22}$

Thus, the odds of sudden death is 1.22 higher in women who are 10 yr. older.

$95\%\ CI = \exp\left(\underbrace{d \cdot \hat{\beta}_3}_{0.2} \pm 1.96 \cdot \underbrace{d}_{2} \cdot \underbrace{SE(\hat{\beta}_3)}_{0.1}\right) = \boxed{(0.825, 1.808)}$

c) According to Dr. Zou during review.
In the situation where you have a case-control study, you cannot use the fitted logistic regression for prediction. Since you have selected on a specific outcome, the number of controls & cases are not as they are present in the population. In this case, the intercept is a result of our design & is not applicable to the population as it will tend to overestimate the probability of death.

Ann Marie Weideman,
MS Part II, 2018

1 d) According to Dr. Zou, there are two correct ways of phrasing this answer.

Method 1: We want to test $H_0: \beta_1 = \beta_2 \iff H_0: \beta_1 - \beta_2 = 0$.

We could use a score test, $\frac{\widehat{(\beta_1 - \beta_2)}}{SE(\beta_1 - \beta_2)} = \frac{(\hat{\beta}_1 - \hat{\beta}_2)}{\sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)}} = \frac{(\hat{\beta}_1 - \hat{\beta}_2)}{\sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)}}$

We don't have $Cov(\hat{\beta}_1; \hat{\beta}_2)$, we were only given $Var(\hat{\beta}_1) = SE(\hat{\beta}_1)^2$ and $Var(\hat{\beta}_2) = SE(\hat{\beta}_2)^2$.

So, we are shit out of luck. :(

Method 2: Again, want to test $H_0: \beta_1 = \beta_2 \iff H_0: \beta_1 - \beta_2 = 0$,

We could use a likelihood ratio (LR) test. This involves computing a log-likelihood ratio for the full model ($logit(\hat{p}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot HTN + \hat{\beta}_2 \cdot Smoke + \hat{\beta}_3 \cdot Age$) and a log-likelihood ratio for the reduced model ($logit(\hat{p}) = \hat{\beta}_0^* + \hat{\beta}_1^*(HTN + smoke) + \hat{\beta}_2^* \cdot Age$).

Would calculate $-2LR(reduced) - (-2LR(full)) \sim \chi^2_{df(full) - df(reduced)} \equiv \chi^2_1$

However, again we were not given the likelihood ratios for the full & reduced models, so no bueno.

1 e) Without info regarding the distribution of age in the controls, the investigator cannot make any conclusion regarding risk of death.

Ann Marie Weideman, MS Part II 2018

2 a)

First, know $Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2} Var(X) = \frac{1}{n^2} \cancel{n} \hat{p}(1-\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$

Will be able to use this eqn. for variance to derive 95% CI's.

Point Estimates: $\hat{p}_0 = P(\text{outcome}=1 \mid \text{intervention}=0) = \frac{14+5}{75} \approx \boxed{0.253}$

$\hat{p}_1 = P(\text{outcome}=1 \mid \text{intervention}=1) = \frac{20+10}{75} = \boxed{0.40}$

$95\%\ CI(p_0) = 0.253 \pm 1.96\sqrt{\frac{0.253(1-0.253)}{75}} = \boxed{(0.155, 0.351)}$

$95\%\ CI(p_1) = 0.40 \pm 1.96\sqrt{\frac{0.40(1-0.40)}{75}} = \boxed{(0.289, 0.511)}$

2 b) Can compute an OR, RR, or RD (risk difference, same as difference in proportions).
Here, will compute an RD.

$\hat{RD} = \hat{p}_1 - \hat{p}_0$ where $\hat{p}_0 = P(\text{outcome}=1 \mid \text{intervention}=0) = 0.253$ (last part)

$\hat{p}_1 = P(\text{outcome}=1 \mid \text{intervention}=1) = 0.40$ (last part)

$\Rightarrow \hat{RD} = 0.4 - 0.253 = 0.247$

$95\%\ CI(RD) = (\hat{p}_1 - \hat{p}_0) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}}$

$= 0.247 \pm 1.96\sqrt{\frac{0.4(1-0.4)}{75} + \frac{0.253(1-0.253)}{75}}$

$= \boxed{(0.099, 0.395)}$

The patients who received the intervention had almost 25 additional cases (out of 100 persons) of HbA1c below 7.5% when compared to patients who received only usual care.

Ann Marie Weideman
MS Exam II, 2018

2 c) Taking difference in proportions from 2b) have,

$H_0: p_1 - p_0 = 0$

$$Z = \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_2}}} = \frac{0.4 - 0.253}{\sqrt{\frac{0.4(1-0.4)}{75} + \frac{0.253(1-0.253)}{75}}} \approx 1.94 < 1.96$$

Using 2·pnorm(1.943, lower.tail=F) in R returns $p = 0.052$

Assuming an $\alpha = 0.05$, we fail to reject the null hypothesis & conclude that there is no evidence that the intervention is effective in lowering BS.

2 d) Have $\text{logit}(p) = \beta_0 + \beta_1 \cdot \text{intervention} + \beta_2\, \text{sex} + \beta_3 \cdot \text{intervention} \cdot \text{sex} + \varepsilon$

Have $\text{intervention} = \begin{cases} 1, & \text{intervention} \\ 0, & \text{usual care} \end{cases}$

$\text{sex} = \begin{cases} 0, & \text{female} \\ 1, & \text{male} \end{cases}$

Testing: $H_0: \beta_3 = 0$

$H_1: \beta_3 > 0$ (Know β_3 must be positive since β_3 exists if $\underset{\text{males}}{\text{sex}=1}$ & intervention $= 1$ in order for an increased logit(p) – our hypothesis.

Fitting the above model in R gives: $\hat{\beta}_3 = 0.2029$

$p = 0.7903$ (for a two-sided test)

used,

glm(Outcome ~ Intervention * sex,
data = df, family = 'binomial')

However, we need the p-value for a one-sided test. For symmetric distributions, if two-tailed p-value > 0.5, then one-tailed p-value $= 1 - \left(\frac{\text{two-tailed } p}{2}\right)$.

If two-tailed p-value < 0.5, then one-tailed p-value $= \left(\frac{\text{two-tailed } p}{2}\right)$.

Here $p_2 = 0.7903 > 0.5$, so $p_1 = 1 - \frac{0.7903}{2} \approx 0.605$

Since $p = 0.605 > \alpha = 0.05$, we fail to reject the null. There is no evidence that males respond better to the intervention.

2 e) Simply fit the model $logit(p) = \beta_0 + \beta_1 \cdot intervention + \beta_2 \cdot Sex + \varepsilon$ using glm(Outcome ~ intervention + sex, data = df, family = "binomial") in R to get a sex adjusted estimate of $\hat{\beta}_1 = 0.7165$ with associated $SE(\hat{\beta}_1) = 0.3634$.

Then, $95\% \; CI(\beta_1) = \hat{\beta}_1 \pm 1.96 \cdot SE(\hat{\beta}_1) = 0.7165 \pm \overbrace{1.96 \cdot 0.3634}^{0.712264}$

$= \boxed{(0.0042, 1.4288)}$

The confidence interval borderline contains the 0 estimate. However, at a significance level of $\alpha = 0.05$, these results would still be considered "statistically significant", and we would conclude that the intervention, when adjusted for differences in sex, has a positive effect on the outcome (increasing the probability of low blood sugar).

Ann Marie Weideman, MS Exam II
2018

3 a) $H_0: \mu_1 = \mu_2 = \mu_3 = 0$ vs. $H_1: \mu_i \neq 0$ for at least one $i = 1,2,3$

ANOVA	SS	df	F
Between	$\sum_{i=1}^{I} \sum_{j=1}^{n_i} [\bar{Y}_i - \bar{Y}]^2$	$i-1$	$\dfrac{(SS_{between}/i-1)}{(SS_{within}/n-i)}$
Within	$\sum_{i=1}^{I} \sum_{j=1}^{n_i} [Y_{ij} - \bar{Y}_i]^2$	$N-i$	
Total	$SS_{between} + SS_{within}$	$N-1$	

General formulas for I groups

ANOVA	SS	df	F
Between	70.94	2	52.24
Within	201,57	297	
Total	272.51	299	

where $\bar{Y}$ = grand mean = $\dfrac{3.78(100) + 3.23(100) + 2.59(100)}{300} = \dfrac{3.78 + 3.23 + 2.59}{3} = 3.2$

Then, $SS_{between} = \sum_{i=1}^{3} \sum_{j=1}^{100} [\bar{Y}_i - \bar{Y}]^2 = \underbrace{100}_{n_1}(\underbrace{3.78}_{\bar{Y}_1} - \underbrace{3.2}_{\bar{Y}})^2 + \underbrace{100}_{n_2}(\underbrace{3.23}_{\bar{Y}_2} - \underbrace{3.2}_{\bar{Y}})^2 + \underbrace{100}_{n_3}(\underbrace{2.59}_{\bar{Y}_3} - \underbrace{3.2}_{\bar{Y}})^2$

$= 70.94$

$SS_{within} = \sum_{i=1}^{3} \sum_{j=1}^{100} [Y_{ij} - \bar{Y}_i]^2 = (n_1 - 1) \cdot SD_1^2 + (n_2 - 1) \cdot SD_2^2 + (n_3 - 1) \cdot SD_3^2$

$= 99(0.79)^2 + 99(0.86)^2 + 99(0.82)^2$

≈ 201.57

$SS_{Total} = SS_{between} + SS_{within} = 70.94 + 201.57 = 272.51$

$df_{between} = 3 - 1 = 2$, $df_{within} = 300 - 3 = 297$, $df_{total} = 300 - 1 = 299$

$$F = \frac{SS_{between} / (i-1)}{SS_{within} / \left[\left(\sum_{i=1}^{I} n_i\right) - i\right]} = \frac{(70.94/2)}{(201.57/297)} \approx 52.24 \sim F_{2,297}$$

p-value = pf(52.24, df1 = 2, df2 = 297, lower.tail = F) = 1.61×10^{-19} in R.

Since $p = 1.61 \times 10^{-19} < \alpha = 0.05$, we reject the null hypothesis and conclude that the group means are not all identical (at least one is different).

3 b)

Ann Marie Weideman, MS Exam II, 2018

From given construct, have:

$Y = X\beta + \varepsilon$ where,

$$X = \underbrace{\begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & 3 \end{bmatrix}}_{\text{Design matrix}} \begin{array}{l} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} 100\times \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} 100\times \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} 100\times \end{array} \qquad \beta = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

$Y: 300 \times 1$

$X: 300 \times 2$

$\beta: 2 \times 1$

$\varepsilon: 300 \times 1$

Then, three steps:

① Find parameter estimates.

Fit following model in R, `lm(y~x)` where `y = c(3.78, 3.23, 2.59)` ($\uparrow$ vector)

and $\underbrace{x = 1:3}_{\text{vector containing 1, 2, and 3}}$

Will return parameter estimates of, $\boxed{\hat{\alpha}_1 = 4.39}$ and $\boxed{\hat{\alpha}_2 = -0.595}$

② Find MSE, $\hat{\sigma}^2$

According to the 663 textbook, on pg. 325,

$\hat{\sigma}^2 = SS_{within} / \underbrace{N-i}_{\substack{\text{on previous pg,} \\ \text{we wrote this} \\ \text{for df}}} = 201.57/297 \approx \boxed{0.679}$ $\leftarrow$ MSE

③ Find SEE (standard error of the estimates).

Need $SE(\hat{\alpha}_1)$ and $SE(\hat{\alpha}_2)$.

means diagonal entries of matrix $(X'X)^{-1}$

Use formula $SEE = \sqrt{\hat{\sigma}^2 \, \text{diag}(X'X)^{-1}}$. Code X in R using `X = cbind(rep(1,300), c(rep(1,100), rep(2,100), rep(3,100)))`

Then, use $(\underbrace{t(X)}_{\text{transpose of } X} \overbrace{\%*\%}^{\text{matrix mult.}} X)\,\hat{}\,\underbrace{(-1)}_{\text{inverse}}$ to get inverse SSCP, $(X'X)^{-1}$

Then, use $\text{sqrt}(0.679 \cdot (X'X)^{-1})$ and grab diagonal entries. These will be $SE(\hat{\alpha}_1)$ and $SE(\hat{\alpha}_2)$. You will find $\boxed{SE(\hat{\alpha}_1) = 0.0476}$ and $\boxed{SE(\hat{\alpha}_2) = 0.022}$.

Ann Marie W.
MS Exam II, 2018

3 c) The given hypothesis of $H_0: \beta_2 = \beta_3 = 0$ is equivalent to testing $\mu_1 = \mu_2 = \mu_3$ because if $\beta_2 = \beta_3 = 0 \Rightarrow \mu_1 = \beta_1$, $\mu_2 = \beta_1$ and $\mu_3 = \beta_1$.

This is equivalent to the F test in the one-way ANOVA table in part a).

Thus, $F \approx 52.24 \sim F_{2,297}$

p-value = pf(52.24, df1 = 2, df2 = 297, lower.tail = F) = 1.61×10^{-19} } in R

Since $p = 1.61 \times 10^{-19} < \alpha = 0.05$, we reject the null hypothesis and conclude that at least one β_i (for $i = 2, 3$) is non-zero.

3 a) $H_0: \mu_1 = \mu_2 = \mu_3 = 0$ vs. $H_1: \mu_i \neq 0$ for at least one $i = 1,2,3$

ANOVA	SS	df	F
Between	$\sum_{i=1}^{I} \sum_{j=1}^{n_i} [\bar{y}_i - \bar{y}]^2$	$i-1$	$\frac{(SS_{between}/i-1)}{(SS_{within}/n-i)}$
Within	$\sum_{i=1}^{I} \sum_{j=1}^{n_i} [y_{ij} - \bar{y}_i]^2$	$N-i$	
Total	$SS_{between} + SS_{within}$	$N-1$	

General formulas for I groups

ANOVA	SS	df	F
Between	70.94	2	52.24
Within	201.57	297	
Total	272.51	299	

where $\bar{y}$ = grand mean = $\frac{3.78(100) + 3.23(100) + 2.59(100)}{300} = \frac{3.78 + 3.23 + 2.59}{3} = 3.2$

Then, $SS_{between} = \sum_{i=1}^{3} \sum_{j=1}^{100} [\bar{y}_i - \bar{y}]^2 = \underbrace{100}_{n_1}(\underbrace{3.78}_{\bar{y}_1} - \underbrace{3.2}_{\bar{y}})^2 + \underbrace{100}_{n_2}(\underbrace{3.23}_{\bar{y}_2} - \underbrace{3.2}_{\bar{y}})^2 + \underbrace{100}_{n_3}(\underbrace{2.59}_{\bar{y}_3} - \underbrace{3.2}_{\bar{y}})^2$

$= 70.94$

$SS_{within} = \sum_{i=1}^{3} \sum_{j=1}^{100} [y_{ij} - \bar{y}_i]^2 = (n_1 - 1) \cdot SD_1^2 + (n_2 - 1) \cdot SD_2^2 + (n_3 - 1) \cdot SD_3^2$
$= 99(0.79)^2 + 99(0.86)^2 + 99(0.82)^2$

≈ 201.57

$SS_{Total} = SS_{between} + SS_{within} = 70.94 + 201.57 = 272.51$

$df_{between} = 3 - 1 = 2$, $df_{within} = 300 - 3 = 297$, $df_{total} = 300 - 1 = 299$

$F = \frac{SS_{between} / (i-1)}{SS_{within} / [(\sum_{i=1}^{I} n_i) - i]} = \frac{(70.94/2)}{(201.57/297)} \approx 52.24 \sim F_{2,297}$

p-value = pf(52.24, df1 = 2, df2 = 297, lower.tail = F) = ~~1.61 x 10^-16~~ 3.62×10^{-20} in R.

Since p = ~~1.61 x 10^-14~~ 3.62×10^{-20} $< \alpha = 0.05$, we reject the null hypothesis and conclude that the group means are not all identical (at least one is different).

Ann Marie Weideman, MS Exam II, 2018

3 b)

From given construct, have:

$Y = X\beta + \varepsilon$ where,

$Y: 300\times 1$

$X: 300\times 2$

$\beta: 2\times 1$

$\varepsilon: 300\times 1$

$$X = \underbrace{\begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & 3 \end{bmatrix}}_{\text{Design matrix}} \begin{matrix} \}\,100\times \\ \}\,100\times \\ \}\,100\times \end{matrix}, \quad \beta = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad (X'X)^{-1} = \begin{pmatrix} -0.023333 & -0.010 \\ -0.010 & 0.005 \end{pmatrix}$$

Then, three steps:

① Find parameter estimates.

$lm(\tilde{y} \sim \alpha_1 + \alpha_2 \tilde{x})$ .... (*)

Fit following model in R, where $\tilde{y} = c(3.78, 3.23, 2.59)$ ← vector

and $\tilde{x} = c(1,2,3)$

Will return parameter estimates of, $\boxed{\hat{\alpha}_1 = 4.39}$ and $\boxed{\hat{\alpha}_2 = -0.595}$, $\widetilde{RSS} = 0.00135$ ← SSE of (*)

② Find MSE, $\hat{\sigma}^2$

According to the 663 textbook, on pg. 325, $SSE = SS_{within} + 100 * \widetilde{RSS} = 201.7088$ → SSE of the actual model

$$\hat{\sigma}^2 = \cancel{SS_{within}}\ SSE / \underbrace{(N-p)}_{\text{on previous pg.}} = 201.7088/298 \approx \boxed{0.676688}$$ ← MSE

③ Find SEE (standard error of the estimates).

Need $SE(\hat{\alpha}_1)$ and $SE(\hat{\alpha}_2)$.

Use formula $SEE = \sqrt{\hat{\sigma}^2 \operatorname{diag}(X'X)^{-1}}$ ← means diagonal entries of matrix $(X'X)^{-1}$. Code X in R using X= cbind(rep(1,300), c(rep(1,100), rep(2,100), rep(3,100))

Then, use $(\underbrace{t(X)}_{\text{transpose of X}}\ \overbrace{\%*\%}^{\text{matrix mult.}}\ X)\,\wedge\underbrace{(-1)}_{\text{inverse}}$ to get inverse SSCP, $(X'X)^{-1}$

Then, use sqrt(0.677 · $(X'X)^{-1}$) and grab diagonal entries. These will be $SE(\hat{\alpha}_1)$ and $SE(\hat{\alpha}_2)$. You will find $\boxed{SE(\hat{\alpha}_1) = \cancel{0.0476}\ 0.1257}$ and $\boxed{SE(\hat{\alpha}_2) = \cancel{0.022}\ 0.05818}$.