7/9/19 Prof Zou Review

1) Logistic Regression

a) Do a t-test : $\frac{est}{s.e.}$ (wald + test, or score test)

β_0 : $\frac{-1.75}{0.06}$

b) Odds ratio 2 women 10 yrs apart: look at difference: 2 of age

odds ratio: $\exp(2 \cdot \hat{\beta}_4) = \exp(2 \cdot 0.1)$

95% CI: $2\hat{\beta}_4 \pm 1.96\sqrt{Var(2\hat{\beta}_4)}$

$= 2\hat{\beta}_4 \pm 1.96\sqrt{4 \cdot 0.1^2}$

$= 2 \cdot 0.1$

$= 2\hat{\beta}_4 \pm 1.96 \cdot 2 \cdot 0.1$

Con last year's exam, if you missed you got zero points

Then, take exp

20 yrs apart:

point estimate: $4 \cdot \hat{\beta}_4$ $\rightsquigarrow \sqrt{4^2 0.1^2}$

95% CI: $4\hat{\beta}_4 \pm 1.96 \cdot 4 \cdot 0.1$

$= [*1 \quad *2]$

95% CI of OR $\longrightarrow$ $= [\exp(*1), \exp(*2)]$

c) $\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

↳ can be applied to prospective + retrospective

if retrospective: only OR can be calculated

β_0 influenced by design (you design # of cases + controls)

if prospective: β_0 is small b/c reflects the real prevalence in the population

1c) If you calculate this out, then this is wrong it is over-inflated. B/c of study design

β_0 has a completely different meaning in case-control study and prospective.

→ Answer: You can't calculate from this study.
↳ but, if you know the prevalence (of population), then you can plug this in instead of β_0
- case control mainly used to evaluate effect of covariates

d) Hypothesis: $H_0: \beta_1 = \beta_2 \iff H_0: \underbrace{\beta_1 - \beta_2}_{=\theta} = 0$

Score test $\quad \frac{\beta_1 - \beta_2}{\text{s.e.}(\quad)} = \frac{\hat{\beta}_1 - \hat{\beta}_2}{?} \qquad \frac{\hat{\theta}}{\text{s.e.}(\hat{\theta})}$ ← asymptotically normal

missing the covariates part, so can't calculate the standard error

Can try likelihood ratio test (but are not given likelihoods of full model + reduced model)

→ Answer: No, can't do it. Need covariance (or LRT)

full model: $\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
reduced: $\text{logit}(p) = \beta_0 + \beta_1 (x_1 + x_2) + \beta_3 x_3 + \beta_4 x_4$
degrees of freedom: 1

1c) They were only looking at the percentage of cases, not looking at controls at all. So, no this doesn't tell you anything. Your study design could have been such that you had more people in that age group no matter if they are cases or controls (also this is case-control → don't know what's happening outside of this group)

2) (also can be considered logistic regression)

The information you need is in the summary data (just a repeat of the other data files)

a) $\frac{30}{75}$ ← Control estimate

S.E. : for binomial distribution:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

So 95% CI: $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

b) One summary statistic you can use: $p_1 - p_2$

so: $\hat{p}_1 - \hat{p}_2$

95% CI: $var(\hat{p}_1 - \hat{p}_2) = var(\hat{p}_1) + var(\hat{p}_2) + 0$ ← case and control are independent

You can do $\frac{p_1}{p_2}$, test if $\frac{p_1}{p_2} \neq 1$. But then, much more difficult to calculate the variance. Since you have data, you can even do the odds ratio + confidence interval. It's ok to say I ran SAS and got this.

Say if stat. sig. based on the C.I. (if spans 0 or 1)

2c) Test statistic: $\frac{\hat{p}_1 - \hat{p}_2}{s.e.(\hat{p}_1 - \hat{p}_2)}$ ← perform formal test + get p value w/ R or SAS

Or can fit logistic regression

Can use Fisher exact test or Chi-squared test

Lots of ways you can do this.

d) You have multi-ple covariates: gender + intervention

Need logistic regression

$$\text{logit}(p) = \beta_0 + \beta_1 trt + \beta_2 sex$$

↖ here, assuming that intervention is the same for male + female

instead need interaction term:

$$\text{logit}(p) = \beta_0 + \beta_1 trt + \beta_2 sex + \beta_{12} trt * sex$$

$$trt = \begin{cases} 0 & \text{control} \\ 1 & \text{intervention} \end{cases}$$

$$sex = \begin{cases} 0 & M \\ 1 & F \end{cases}$$

intervention effect: females: $\beta_1 + \beta_{12}$ — females: $\beta_0 + \beta_1 trt + \beta_2 + \beta_{12} trt$

for each gender: males: β_1

If testing if same, test $\beta_{12} = 0$, If testing works better

2d) in men: $H_0: \beta_{12} = 0$ vs $H_1: \beta_{12} < 0$

If you switch the design matrix:

$$trt = \begin{cases} 1 & \text{control} \\ 0 & \text{intervention} \end{cases}$$

$$sex = \begin{cases} 0 & m \\ 1 & f \end{cases}$$

males: $\beta_0 + \beta_1 trt$

females: $\beta_0 + \beta_1 trt + \beta_2 + \beta_{12} trt$

Intervention - control ← this is the effect of intervention

effect effect

Intervention effect for each sex:

Female:

$[\beta_0 + \beta_1(0) + \beta_2(1) + \beta_{12}(0)] - [\beta_0 + \beta_1(1) + \beta_2(1) + \beta_{12}(1)]$

Intervention − control

$= -\beta_1 - \beta_{12}$

Male:

$[\beta_0 + \beta_1(0) + \beta_2(0) + \beta_{12}(0)] - [\beta_0 + \beta_1(1) + \beta_2(0) + \beta_{12}(0)]$

$= -\beta_1$

Run logistic regression, get estimates + s.e. + run one-sided test

$H_0: \beta_{12} = 0$

vs $H_1: \beta_{12} > 0$

⚠ How to get one-sided pvalue

In SAS → can fit summary table. In R can expand.

2e) So, don't need the interaction term

$\beta_0 + \beta_1 trt + \beta_2 sex$

↑ adjusted for sex by adding to model

↑ this is the estimate

Once you read this question, d) should be more clear.

3) This will not be in our exam b/c we didn't cover ANOVA. But, should still be able to do everything except a).

a) For 1-way ANOVA that's balanced like this, only need summary statistics

switch places?

ANOVA

here, $K=3$ — samples in group — observations in each group — avg for each group

(like SS error) within SS $\Rightarrow \sum_{k=1}^{K} \sum_{i=1}^{nk} [y_{ki} - \bar{y}_k]^2$

(like SS model) between SS $\Rightarrow \sum_{k=1}^{K} \sum_{i=1}^{nk} (\bar{y}_k - \bar{y})^2$

In problem, purpose is to indicate groups

$n_1 = n_2 = n_3 = 100$

$\sum_{i=1}^{nk} [y_{ki} - \bar{y}_k]$ ← S.D. for one group

SS total = within SS + between SS

df: w/in SS: $K-1$ (2) (3 parameters in 1-way ANOVA)

(3-1)

between SS: (297)

total SS: $n-1$ (299)

3a) $H_0: \mu_1 = \mu_2 = \mu_3$

$\Longleftrightarrow$ $H_0: \mu_1 - \mu_2 = 0$

$\mu_1 - \mu_3 = 0$

F test : $\frac{\text{between}/2}{\text{w/in}/297} \sim F_{2,297}$

b) constraint on the model

$$\mu_i = \alpha_1 + \alpha_2 i$$

μ_i here is a parameter (mean of each groups)
(can use $\bar{y}_i$ to estimate μ_i)

These groups are ordered (trend test)

$\mu_1 = \alpha_1 + \alpha_2$, $\mu_2 = \alpha_1 + 2\alpha_2$, $\mu_3 = \alpha_1 + 3\alpha_3$

In part a) you have 3 parameters : μ_1, μ_2, μ_3
In part b) you have 2 parameters : α_1 and α_2

fit linear regression model. If had individual data,
the design matrix should be:

model : $y_{ij} = \alpha_1 + \alpha_2 * i + \varepsilon_{ij}$

$$X = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & 3 \end{bmatrix} \begin{matrix} \} 100 \\ \\ \} 100 \\ \\ \} 100 \end{matrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

2×300

$(X'X)^{-1} = ?$

(If smoking has no effect, test $\alpha_2 = 0$) This is a reduced model
(part a is full model)

3b) Can fit linear regression:

$$\begin{matrix}\bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3\end{matrix} \quad \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$$ ← essence matrix

get $\hat{\alpha}_1$, $\hat{\alpha}_2$ (but error is messed up b/c collapsed)

Then, go back to original formula to get error

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{100} (y_{ij} - \hat{\alpha}_1 - \hat{\alpha}_2 * i)^2$$ ← can get this, but takes a while

=

c) $\mu_i = \beta_1 + \beta_2 i + \beta_3 i^2$

Trasformation on x, linear + quadratic transformation

3 parameters: $\beta_1, \beta_2, \beta_3$

The 2 models are equivalent b/c they both have 3 parameters

If $\beta_2 = \beta_3 = 0$, then all $\mu_i = \beta_1$

testing $\mu_1 = \mu_2 = \mu_3 (= \beta_1)$, so copy the test from above.

3b) $Y_{ij} \sim N(\mu_i, \sigma^2)$

likelihood: $\prod_{i=1}^{3} \prod_{j=1}^{100} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}\right\}$

$$L(\mu_1, \mu_2, \mu_3, \sigma^2 | y) =$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{300} \prod_{i=1}^{3} \prod_{j=1}^{100} \exp\left\{-\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}\right\}$$

$$\log L \propto -\frac{300}{2}\log(\sigma^2) + \sum_{i=1}^{3}\sum_{j=1}^{100}\left(-\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}\right)$$

$$= -150\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{3}\sum_{j=1}^{100}(y_{ij}-\mu_i)^2$$

$\mu_i \rightarrow \alpha_1 + \alpha_2 * i$

expand, and see that you can get all these in terms of mean and SD

from here get MLE of σ^2

$$\hat{\sigma}^2 = \sum_{i=1}^{3}\sum_{j=1}^{100}(y_{ij} - \hat{\alpha}_1 - \hat{\alpha}_2 * i)^2 / 298$$

$(n-p)$

If you get d.f. right, you get points.