Cheat sheet

Tuesday, November 28, 2023 12:59

Write that GEE is pop level

Operations

$Var(X) = E(X^2) - [E(X)]^2 = E[(X - E(X))^2]$

$E(X \pm k) = E(X) \pm k$

$E(kX) = kE(X)$

MLR Review: $\sigma^2 = MSE = resids$

$E(X \pm Y) = E(X) \pm E(Y)$

$Var(X \pm Y) = Var(X) \pm Var(Y) \pm 2\,cov(X,Y)$

$Var(X \pm k) = Var(X)$

$Var(kX) = k^2 Var(X) = A\, Var(X)\, A'$

$E(XY) = E(X)E(Y)$ IND. ONLY

$Var(X \pm Y) = Var(X) \pm Var(Y)$ IND. ONLY

$Cov(a_1X_1 + a_2X_2, b_1Y_1 + b_2Y_2) = a_1b_1 Cov(X_1, Y_1) + a_1b_2 Cov(X_1, Y_2) + a_2b_1 Cov(X_2, Y_1) + a_2b_2 Cov(X_2, Y_2)$

$Cov(AY, BY) = A\, Var(Y)\, B'$

MLR Review

$\hat\beta_{OLS} = (X'X)^{-1}X'Y$

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

$HY = \hat Y = X\hat\beta$

$\hat\epsilon = Y - X\hat\beta$

$E(\hat Y) = X\beta$

$\hat{Var}(\hat\beta) = \sigma^2(X'X)^{-1} = Cov(\hat\beta)$

$\hat{Var}(\hat Y) = \sigma^2 H = Cov(\hat Y)$

$S^2_{pred} = X'\hat{Var}(\hat\beta)X$ ↳ variance of pred. obs.

Pred. Int: $\hat y_i \pm 1.96\hat\sigma$ which is $\hat{Var}(\hat Y) + \sigma^2$

Residual vs Fitted
- $\hat\epsilon_i$ should be independent of $\hat y_i$ (no patterns).
- Linearity: Red line should be flat.
- Variance constant (homoscedasticity).

Also LMM & GEE diagnostics

Normal Q-Q Plot
- Standardized residual $\hat\epsilon_i/(\hat\sigma\sqrt{1-h_{ii}})$ should be standard normal.
- We expect the points to follow a straight line. Check non-normality, particularly skewed tails.

Scale-Location
- Similar to residual-vs-fitted, but use $\hat\epsilon_i/(\hat\sigma\sqrt{1-h_{ii}})$. Diagnose heteroscedasticity (e.g., red line increasing).

Residual vs Leverage
- Leverage h_{ii} is how far away x_i is from other x_j.
- high leverage and outlier → problem.
- Cook's Distance: measures how much model changes when remove ith point; > 1 is problematic. (Note: when this occurs, a contour line appears in plot.)

VIF > 10 or Cooks > 1 is bad

OLS intercept only: The expected log rate per 100000 people is 2.236324 HIV cases without considering other covariates.

LMMs

Random Intercept

$y_{ij} = \beta_0 + \theta_i + x_{ij}'\beta + \epsilon_{ij}$

x_{ij} = mean subject specific diff.

$\theta_i \overset{iid}{\sim} N(0, \tau^2)$ between-group variability in intercepts aka RE variance

$\theta_i \perp \epsilon_{ij}$

$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ within-group variability aka resid. variance aka measure. error

$E[y_{ij}] = \mu + x_{ij}'\beta$

$Var[y_{ij}] = \tau^2 + \sigma^2$

$E[y_{ij} | \theta_i] = \mu + \theta_i + x'\beta$

$Var[y_{ij} | \theta_i] = \sigma^2$

In matrix: $Y = Z\theta + X\beta + \epsilon$

$\theta \sim N(0, \tau^2 I)$, $\epsilon \sim N(0, \sigma^2 I)$

Random Intercept AND Slope

$y_{ij} = \beta_0 + \theta_{0i} + (\beta_1 + \theta_{1i})x_{ij} + \epsilon_{ij}$

$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, $\theta_i \perp \epsilon_{ij}$

$\begin{bmatrix}\theta_{0i}\\ \theta_{1i}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau_1^2 & \rho\tau_1\tau_2\\ \rho\tau_1\tau_2 & \tau_2^2\end{bmatrix}\right)$

$\rho = corr(\theta_{0i}, \theta_{1i})$

τ_1^2 = between-grp variation in baseline outcome at xvar = 0

τ_2^2 = variation of random slope aka between-grp variation in the linear effect of x var

In R: Random effects:

Groups	Name	Var	Std. Dev.
ID	(Int.)	τ_1^2	
resid		σ^2	

```
> fit = lmer(arm ~ age + (1|id), data = nepal)
Linear mixed model fit by REML
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.78023  0.88330
 Residual             0.34837  0.40606

Fixed effects:
            Estimate Std. Error t value
(Intercept) 12.753799  0.109907  116.30
age          0.031859  0.002397   13.45
```

τ^2 just $\sqrt{}$ of variance aka $\sqrt{0.78023}$

We found a 0.032 cm ($CI_{95\%}$: 0.027, 0.037) increase in arm circumference per month after controlling for a child's arm circumference at birth.

We also found evidence of heterogeneity in arm circumference at birth. The estimated population-average arm circumference at birth is 12.8 cm, and the standard deviation of the random effect is 0.88 cm.

COMPARING MODELS

AIC: Rule of thumb: diff ≥ 2

LRT - ANOVA(reduced, full)

"Acquisition has a significant effect on the z-transformed correlations when controlling for gender and scanner type." P-values for LRTs of random effects tend to be too large, and similarly, the AIC can select overly simple models, which is a consequence of the null hypothesis being on the boundary of the parameter space (i.e., variance = 0).

Use MLE for these.

```
> fit2 = lmer(arm~age+lit+(1|id), data = nepal)
> summary(fit2)
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.74712  0.86436
 Residual             0.34834  0.40923

Fixed effects:
            Estimate Std. Error t value
(Intercept) 12.710685  0.109554  116.29
age          0.031789  0.002338   13.60
lit          0.890247  0.316301    2.81
```

We found literacy to be significantly associated with arm circumference as a main effect. Also note that there is a small decrease in the degree of heterogeneity (from 0.78 to 0.75). Therefore mother's literacy may help explain some of the observed between-children variation in arm circumference at birth. Also the intercept estimate 12.71 now corresponds to the population-average arm circumference at birth from mothers who are illiterate.

ICC (ρ)

$\rho = Corr(y_{ij}, y_{ij'}) = \frac{\tau^2}{\tau^2 + \sigma^2}$ for all $j \neq j'$

aka reproducibility

< 0.40 = POOR, 0.75–1.00 = EXCELLENT

$\tau^2 \to 0$, $\rho \to 0$ same intercept

$\tau^2 \to \infty$, $\rho \to 1$ separation in intercepts

$\sigma^2 \to 0$, $\rho \to 1$ no measurement error

$\sigma^2 \to \infty$, $\rho \to 0$ growing measurement error

BLUP

$\hat\theta_i = E[\theta_i | y_i] = \frac{\sigma^2\mu + \tau^2 r_i \bar y_i}{\sigma^2 + r_i\tau^2}$ useful for small samples.

REML → more conservative, SE's are often larger than MLE.

SHRINKAGE

$\hat\theta_i = \frac{(\sigma^2/\tau^2)\mu + \sum_{j=1}^{r_i} y_{ij}}{r_i + \sigma^2/\tau^2}$

τ^2 controls shrinkage & how much info borrowed across groups

$\tau^2 \to 0$, $\hat\theta_i \to \mu$ shrinks random intercepts to μ

$\tau^2 \to \infty$, $\hat\theta_i \to \bar y_i$ no shrinkage, just using raw group mean estimates

$\sigma^2 \to 0$, $\hat\theta_i \to \bar y_i$

$\sigma^2 \to \infty$, $\hat\theta_i \to \mu$

ICC FORM

Also can be written as: $\hat\theta_i = \frac{(1/\tau^2)\mu + (r_i/\sigma^2)\bar y_i}{(1/\tau^2) + (r_i/\sigma^2)} = \frac{\rho^{-1}\mu + r_i(1-\rho)^{-1}\bar y_i}{\rho^{-1} + r_i(1-\rho)^{-1}}$

Random effects are a weighted avg of $\bar y_i$ and μ w/ weights as inverse variances.

when $\rho \to 1$, $\hat\theta_i \to \bar y_i$

covariance structure ($\epsilon_{ij}^* = \theta_i + \epsilon_{ij}$)

$Cov(\epsilon_{ij}^*, \epsilon_{ij}^*) = \tau^2 + \sigma^2$

$Cov(\epsilon_{ij}^*, \epsilon_{ij'}^*) = \tau^2$

$Cov(\epsilon_{ij}^*, \epsilon_{i'j}^*) = 0$

	ϵ_{11}^*	ϵ_{12}^*	ϵ_{13}^*	ϵ_{21}^*	ϵ_{22}^*	ϵ_{23}^*
ϵ_{11}^*	$\tau^2+\sigma^2$	τ^2	τ^2	0	0	0
ϵ_{12}^*	τ^2	$\tau^2+\sigma^2$	τ^2	0	0	0
ϵ_{13}^*	τ^2	τ^2	$\tau^2+\sigma^2$	0	0	0
ϵ_{21}^*	0	0	0	$\tau^2+\sigma^2$	τ^2	τ^2
ϵ_{22}^*	0	0	0	τ^2	$\tau^2+\sigma^2$	τ^2
ϵ_{23}^*	0	0	0	τ^2	τ^2	$\tau^2+\sigma^2$

Group 1, Group 2

GLMs

In general, find g such that $g(E[y_i|x_i]) = g(\mu_i) = \beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}$. Then $\mu_i = g^{-1}(\beta_0 + \sum_{k=1}^{p}\beta_k x_{ik})$ with range $(-\infty, \infty)$. Domain = possible vals of μ_i.

BINARY outcome - LOGISTIC

$E[y_i] = \mu_i = p_i$

$Var[y_i] = V[\mu_i] = V[p_i] = p_i(1-p_i)$

$y_i \overset{IND}{\sim} Bernoulli(p_i)$ NO ERROR TERM

$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}$

$p_i \to 0$, $\log\left(\frac{p_i}{1-p_i}\right) \to -\infty$

$p_i \to 1$, $\log\left(\frac{p_i}{1-p_i}\right) \to \infty$

Logistic fxn: $p(g) = \frac{\exp(g)}{1+\exp(g)} = \frac{1}{1+\exp(-g)}$ $(-\infty, \infty) \to (0,1)$

LOGIT fxn: $g(p) = \log\left(\frac{p}{1-p}\right)$ $(0,1) \to (-\infty, \infty)$

Interpretation:

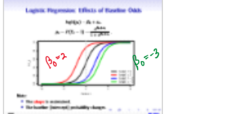
β_0 = baseline log odds

β_k = the log odds ratio associated w/ covariate k while controlling for other covariates $= \log\left[\frac{p_i^*/(1-p_i^*)}{p_i/(1-p_i)}\right]$

e^{β_1} = odds ratio

$CI = e^{(\beta_1 \pm 1.96\, SE)}$

$p_i = \mu_i$ = predicted aka estimated value $= \frac{e^{\beta_0 + \beta_1 x_1 + \dots}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots}}$



Inference: LRT. $H_0: \beta_k = 0$ ⇒ Reject H_0 if $-2\{l(\hat\beta_{reduced}) - l(\hat\beta_{FULL})\} > \chi^2_{\nu, 1-\alpha}$ with ν being diff in # params full vs. reduced

$l(\hat\beta_{FULL})$ = log-likelihood of full model

COUNT outcome - POISSON

$E[y_i] = \mu_i = \lambda_i \to P(y_i = m) = \frac{\lambda_i^m e^{-\lambda_i}}{m!}$

$Var[y_i] = V[\mu_i] = V[\lambda_i] = \lambda_i$

$y_i \overset{IND}{\sim} Poisson(\lambda_i)$ Link function is log()

$\log(\lambda_i) = \beta_0 + \beta_1 x_i$

When $x_i = 0$, $\log(\lambda_i) = \beta_0$ = baseline expected cts

e^{β_1} = relative change in count per unit change in x

% change = $100(e^{\beta} - 1)$%

Impacts are multiplicative

Diagnostics: Q-Q plot now looking at normality of standardized deviance residuals

Scale-location: want var of standardized deviance residuals constant

Deviance residual $= \epsilon_i^d = sign(y_i - \hat\lambda_i)\sqrt{d_i}$ where d_i is i^{th} term in dev. calculation

Overdispersion parameter: Scales variance to account for violations of assumption on relationship between mean & variance

Under null: $\hat\beta \sim N(0, I(\beta)^{-1}\phi)$

>1 = overdispersed, model is bad fit

<1 = underdispersed

Interaction terms:

The multiplicative effect of the interaction between ___ is $e^{\#}$ i.e. the rate increases by $e^{\#}$% relative to no interaction.

Ratios:

$e^{\text{Beta estimates}} / e^{\text{Beta estimates}}$ = # times as many ___ compared to ___

Poisson model

If y_i is the case count for state i, then y_i is independently distributed as a Poisson distribution with expected HIV case count λ_i.

The model is:

$$\log(\lambda_i) = \log(Pop100k_i) + \beta_0 + \beta_1 Poverty_i^* + \beta_2 NoDiploma_i^* + \beta_3 Uninsured_i^* + \beta_4 Vacant_i^*$$

$\log(Pop100k)$ is the log of the population in hundred thousands. This variable is the offset to account for population size differences.

β_0 is the estimated log mean number of HIV cases for state i while adjusting for the log population size (Pop100k) for the average number of households living below federal poverty level, the average population 25 years and older without a HS diploma, the average number of uninsured, and the average number of vacant housing.

β_1 is the coefficient for the predictor variable `Households living below the federal poverty level`. e^{β_1} would be interpreted as the rate ratio of HIV cases for every one unit increase in `Households living below the federal poverty level` in a model controlling all other predictor variables.

GOF test:
- H0: The model adequately fits the data.
- H1: The model does not adequately fit the data.

Test statistic: Chi squared statistic of 5597.6641

Decision: Reject the null hypothesis

Conclusion: The model does not adequately fit the data.

$\lambda_{stk} = pop_{stk} \cdot e^{\beta_0 + \beta_1 + \dots}$

e^{β_0} in offset is now per capita deaths instead of expected death ct $= \lambda_i / pop_i$

$e^{\beta_0 + \beta_1 + \dots}$ = expected ___ count [per capita if offset] for ___ in 1980...

GLMMs

Random Intercept Poisson

τ^2 = variation between groups

θ_s = group-specific dev. in baseline log expected (outcome)

CI: $e^{\beta \pm 1.96\sqrt{\tau^2}}$

$death_{stk} \sim Poisson(\lambda_{stk})$

$\theta_s \overset{iid}{\sim} N(0, \tau^2)$

$\log(E[y_{stk} | \theta_s]) = \log\lambda_{stk} = \beta_0 + \theta_s + \beta_1 I[sex_k = F] + \beta_2 I[race_k = W] + \beta_3 I[sex_k = F]I[race_k = W] + \beta_4 year_t$

- β_0: log expected count at baseline for the avg county (non-white males in 1979)
- e^{β_1} = ratio of deaths for non-white F to non-white M in model accounting for random intercepts, ...
- e^{β_3} = relative rate modification in deaths for white F
- rate ratio of ... for a 1-yr increase in a model controlling for...

Rand. Int. Logistic

What is the odds ratio and 95% confidence interval for treatment=Intervention in age group 2534 versus Control in age group 25-34?

```
control <- 2.0624
intervention <- -1.5228
q3_3 <- exp(2.0624 + -1.5228)
q3_3var <- vcov(q3_2)
q3_3var2 <- q3_3var["treatmentIntervention", "treatmentIntervention"] +
    q3_3var["treatmentIntervention:agegroup25-34", "treatmentIntervention:agegroup25-34"] +
    2 * q3_3var["treatmentIntervention", "treatmentIntervention:agegroup25-34"]
q3_3 * exp(- 1.96 * sqrt(q3_3var2))
```

WLS

$X^* = WX$ Then $Y^* = X^*\beta + \epsilon^*$ $\hat\beta_{WLS} = (X'V^{-1}X)^{-1}X'V^{-1}Y$

$WY = WX\beta + W\epsilon$ $\epsilon^* \sim N(0,1)$ $Cov(\hat\beta_{WLS}) = (X'V^{-1}X)^{-1}$

GLS

V is non-diagonal (heterosked. + corr.); $\hat\beta_{GLS} = [X'V^{-1}X]^{-1}X'V^{-1}Y$; $Cov(\hat\beta_{GLS}) = [X'V^{-1}X]^{-1}$

If no heterosc. but still corr., $V = \sigma^2 R$. $\hat\beta_{GLS} = (X^{-1}R^{-1}X)^{-1}X'R^{-1}Y$ where $\sigma^{-2}R^{-1} = V^{-1}$

Naive $Cov(\hat\beta_{GLS}) = \sigma^2(X'R^{-1}X)^{-1}$