Lecture 8

Lecture 8 - Main Points Once Again

$$y_{ij} = \underset{\sim}{x}_{ij}\underset{\sim}{\beta} + z_{ij} b_i + \epsilon_{ij}$$

$$\underset{n_i\times 1}{Y_i} = \underset{n_i\times p}{\underset{\sim}{X}_i}\ \underset{p\times 1}{\underset{\sim}{\beta}} + \underset{n_i\times q}{Z_i}\ \underset{q\times 1}{b_i} + \underset{n_i\times 1}{\epsilon_i}$$

$$\underset{\sim}{\epsilon}_i \sim G(0, \sigma^2 \underset{n_i\times n_i}{I}) \quad (\text{or } \sigma^2 R_i(\theta))$$

$$\perp$$

$$\underset{q\times 1}{b_i} \sim G(0, \underset{q\times q}{D}), \quad i = 1, \dots, m$$

$$\hat{\beta} = \left(\sum_{i=1}^{m} X_i' V_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{m} X_i' V_i^{-1} Y_i \right).$$

where $V_i = Var(Y_i) = Z_i D Z_i' + \sigma^2 I$

$$\underset{q \times 1}{\hat{b}_i} = \hat{E}(b_i \mid Y_i) = \hat{D} Z_i' (Z_i \hat{D} Z_i' + \hat{\sigma}^2 I)^{-1} \cdot (Y_i - X_i \hat{\beta})$$

EMPIRICAL BAYES (EB)

"SHRINKAGE" ESTIMATOR

Simplest case: $Y_{ij} = \beta_0 + b_i + \varepsilon_{ij}$

$b_i \sim G(0, \tau^2)$; $\varepsilon_{ij} \sim G(0, \sigma^2)$

$\mathrm{Var}\, Y_{ij} = \tau^2 + \sigma^2$; $\underset{\sim\, n_i \times n_i}{V_i} = \begin{pmatrix} \tau^2+\sigma^2 & & \tau^2 \\ & \ddots & \\ \tau^2 & & \tau^2+\sigma^2 \end{pmatrix}$

$\mathrm{Cov}(Y_{ij}, Y_{il}) = \tau^2$

$$\hat{E}(b_i \mid \underset{\sim}{Y_i}) = 0 + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma^2}\left(\bar{Y}_i - \bar{X}_i\hat{\beta}\right)$$

$$= \left(\frac{\tau^2}{\tau^2+\sigma^2}\right)\cdot \bar{Y}_i + \left(\frac{\sigma^2}{\tau^2+\sigma^2}\right)\bar{X}_i\hat{\beta} = \alpha\,\bar{Y}_i + (1-\alpha)\bar{X}_i\hat{\beta}$$

LINEAR | LOGISTIC

	LINEAR	LOGISTIC
μ_{ij}^{c}:	$E(Y_{ij} \mid b_i) = x_{ij}\beta + z_{ij}b_i$	$\text{logit}\, E(Y_{ij} \mid b_i) = x_{ij}\beta + z_{ij}b_i$
b_i:	$b_i \sim G(0, D)$	
Y_{ij}:	$Y_{ij} \sim G(\mu_{ij}^{c}, \sigma^2)$	$Y_{ij} \sim Bin(\mu_{ij}, 1)$ (no σ^2)
μ_{ij}^{M}:	$E(Y_{ij}) = x_{ij}\beta$	$E(Y_{ij}) = \int \frac{e^{x_{ij}\beta + z_{ij}b_i}}{1+e^{x_{ij}\beta + z_{ij}b_i}} \phi(b_i, D)\, db_i$

$$\neq \frac{e^{x_{ij}\beta}}{1+e^{x_{ij}\beta}}$$

Lecture 8

Outline of Lecture 8

- We will consider 2-level multi-level data where the outcome of interest is a level-1 binary variable
- PISA 2000
- Within the PISA 2000 data, we will
  - Consider logistic random intercept model to separate among and within-cluster effects
  - Review estimation and interpretation of contextual effects
  - Refresh our memory on the difference between marginal and random effects logistic model

Marginal

$$E(Y_{ij} \mid X_{ij}) = \frac{e^{X_{ij}\beta}}{1+e^{X_{ij}\beta}}$$

pop avg ef β $\neq$

Random Effects

$$E(Y_{ij} \mid \beta^*, b_i) = \frac{e^{X_{ij}\beta^* + Z_{ij} b_i}}{1+e^{X_{ij}\beta^* + Z_{ij} b_i}}$$

Cluster-specific β^* Mean

Data Example: PISA

- 2000 Program for International Student Assessment (PISA)
- Conducted by the Organization for Economic Cooperation and Development (OECD 2000)
- Annual assessment of reading, mathematics and science proficiency (rotating)
- The 2000 survey assessed proficiency in reading of 15 year olds in 43 countries.
- Focus on US sample
- Complex survey design
  - We will look at this again later in the course

Data Example: PISA

id:		school id (numeric)
pass_read:	Y_{ij}	dummy variable for being proficient in reading (1: proficient, 0: not)
female:		dummy variable for student being female
isei:		international socioeconomic index
high_school		dummy variable for highest education level by either parent being high school
college:	X_{ij}	dummy variable for highest education level by either parent being college
test_lang:		dummy variable for English being spoken at home (NOTE: the reading proficiency test is administered in English)
one_for:		dummy variable for one parent being foreign born
both_for:		dummy variable for both parents being foreign born

Exploratory Data Analysis

```
------------------------------------------------------------
id                                          school id (unique)
--------------------------------------------------------------
                  type:  numeric (long)
                 label:  id

                 range:  [1,148]                 units:  1
         unique values:  148                 missing .:  0/2069

              examples:  30     99160
                         57     99207
                         85     99249
                         116    99313
```

Exploratory Data Analysis

```
. bys id: gen copyid = _n

. bys id: egen n_student = count(id)

. summ n_student if copyid==1, detail

                          n_student
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            2              1
10%            5              1       Obs                 148
25%            9              2       Sum of Wgt.         148

50%         14.5                      Mean           13.97973
                        Largest       Std. Dev.      6.402029
75%           19             25
90%           22             25       Variance       40.98598
95%           24             25       Skewness      -.2412149
99%           25             28       Kurtosis       2.247757
```

14.5

25%tile → 9 19 75%tile 10 – IQR

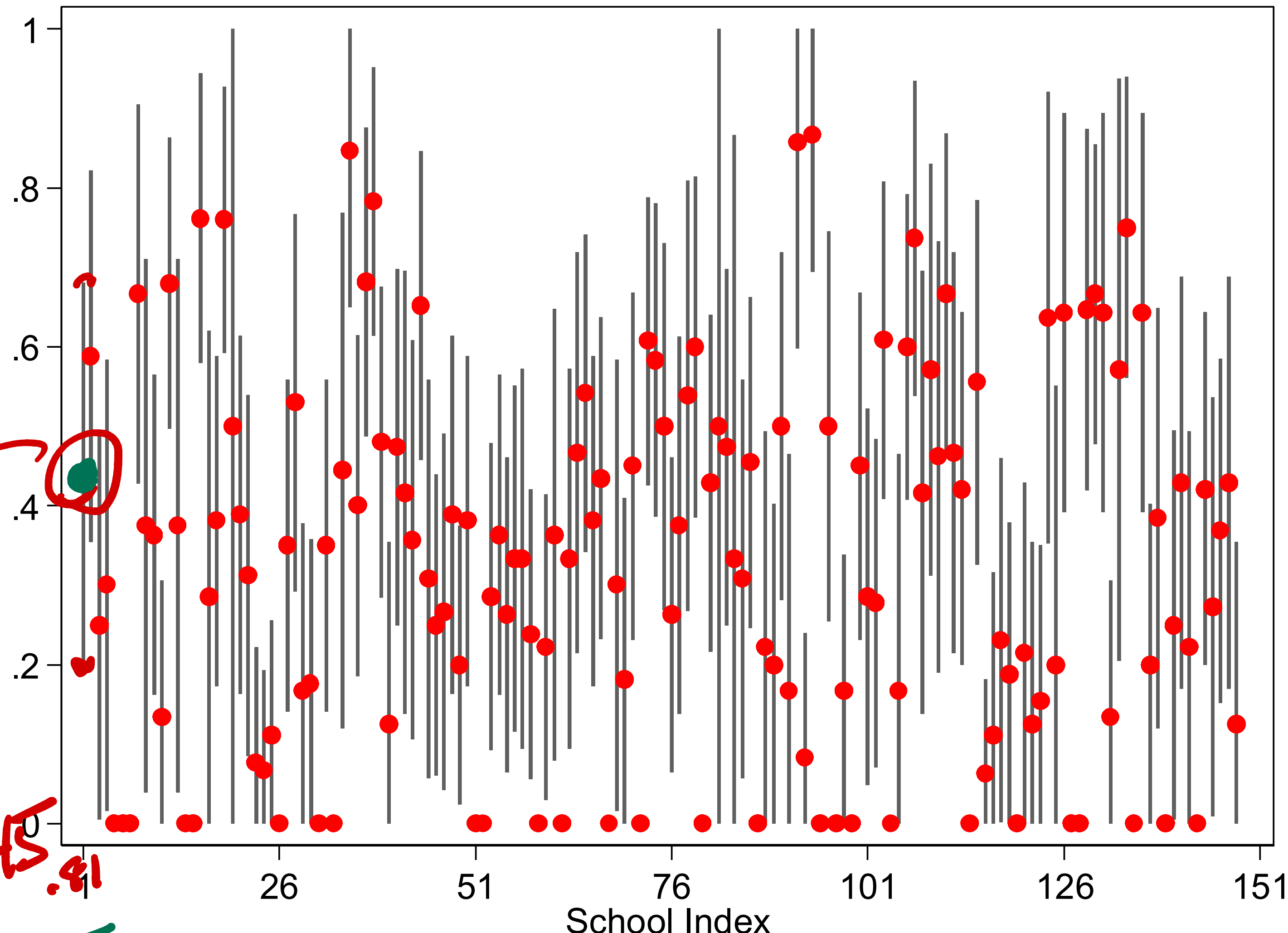
Min → 1 25 max 24 – RANGE

Tukey 5-number Summary

Exploratory Data Analysis

Characteristic	Student % or mean (SD)	School Mean [min, Q1, Q3, max]
Female	54%	54% [0%, 46%, 64%, 100%]
Highest level of education		
Less than High-school	8%	5% [0%, 0%, 16%, 100%]
High-school	34%	33% [0%, 22%, 49%, 100%]
College	58%	57% [0%, 39%, 68%, 100%]
# foreign born parents		
0	83%	90% [0%, 67%, 100%, 100%]
1	5%	0% [0%, 0%, 8%, 50%]
2	12%	5% [0%, 0%, 20%, 100%]
English at home	92%	100% [0%, 86%, 100%, 100%]
ISEI	47 (17)	45 [25, 39, 51, 68]
Proficient in Reading	39%	33% [0%, 13%, 47%, 87%]

Partitioning Among vs. Within Variation of Binary Proficiency Outcome



fraction proficient in school 1

$\hat{p}_1 \approx .45$

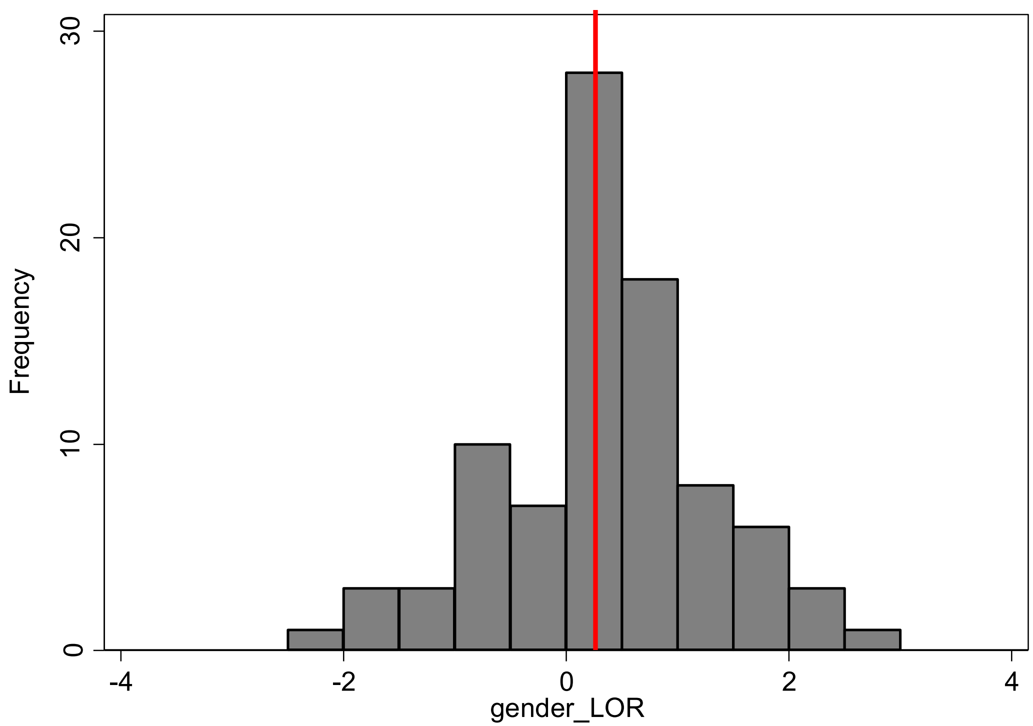
.19 .81

Within, Between and Contextual Effect

- Level-1 predictors
  - **Gender**
  - Highest level of education completed by parents
  - Number of foreign-born parents
  - **International socioeconomic index**
  - English as the primary language at home

Exploratory analysis of within-school gender effect

- Calculate the school-specific log odds ratio of reading proficiency comparing females to males

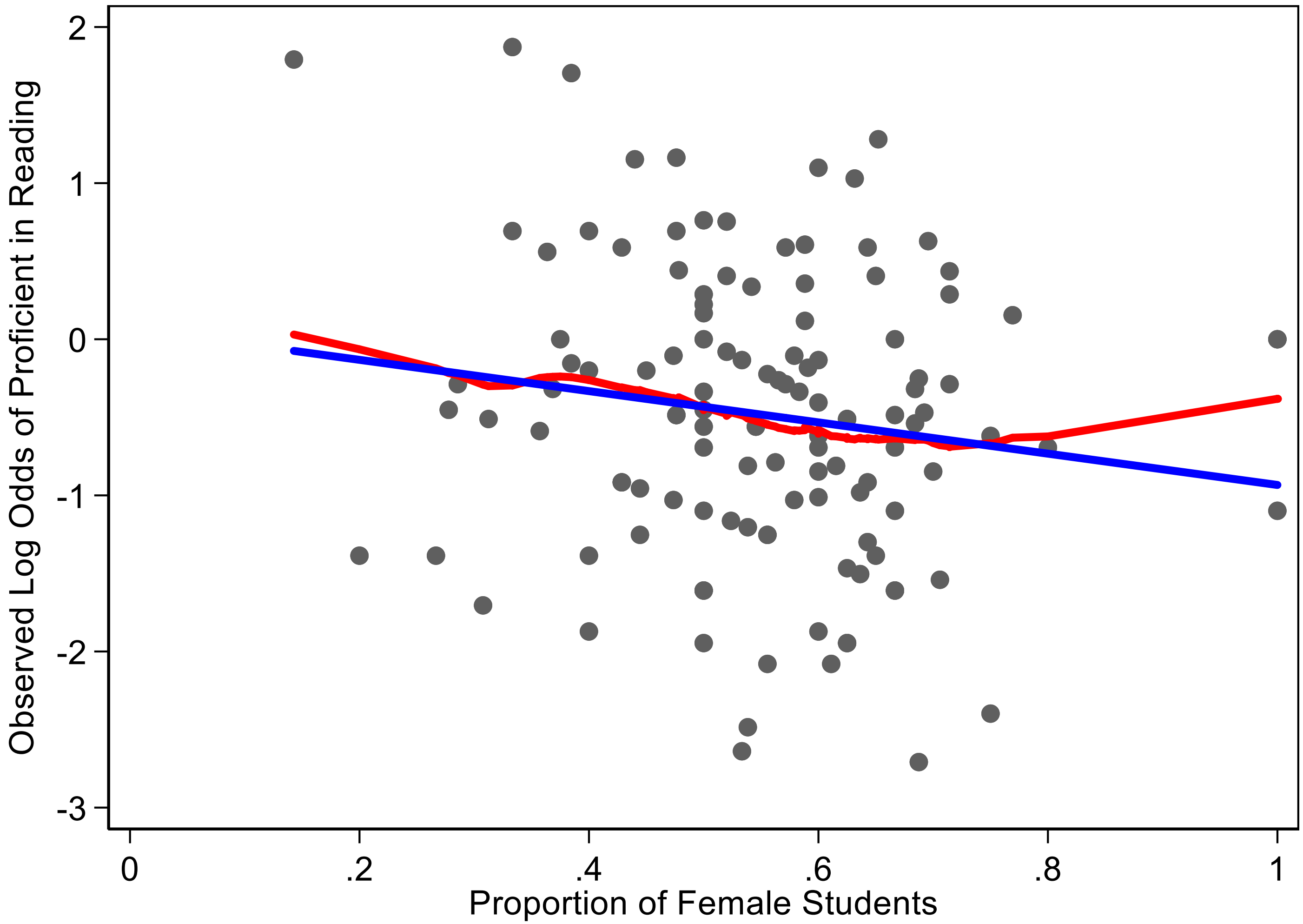


Mean = 0.32

Odds Ratio = exp(0.32) = 1.38

Within school comparison of the odds of reading proficiency comparing females to males

Exploratory analysis of between-school gender effect



Separation of Within- and Among-Cluster (School) Effects

$$\left[\log\left(\frac{Pr(Y_{ij}=1|X_{ij},\alpha_i)}{Pr(Y_{ij}=0|X_{ij},\alpha_i)} \right) - \log\left(\frac{Pr(Y_{ij}=1|X_{ij},\alpha_i)}{1-Pr(Y_{ij}=1|X_{ij},\alpha_i)} \right) \right]$$

$$Model\ 1\!:\ LogOdds(Y_{ij}=1|X_{ij}) \quad = \alpha_{1i} + \beta_1 X_{ij}$$

$$Model\ 2\!:\ LogOdds(Y_{ij}=1|X_{ij},\bar{X}_{i.}) = \alpha_{2i} + \beta_2 X_{ij} + \gamma_2 \bar{X}_{i.}$$

$$Model\ 3\!:\ LogOdds(Y_{ij}=1|X_{ij},\bar{X}_{i.}) = \alpha_{3i} + \beta_3 (X_{ij} - \bar{X}_{i.}) + \gamma_3 \bar{X}_{i.}$$

$$Model\ 4\!:\ LogOdds(Y_{ij}=1|X_{ij},\bar{X}_{i.}) = \alpha_{4i} + \beta_4 (X_{ij} - \bar{X}_{i.})$$

$$Model\ 5\!:\ LogOdds(Y_{ij}=1|X_{ij},\bar{X}_{i.}) = \alpha_{5i} + \gamma_5 \bar{X}_{i.}$$

Where

$$LogOdds(Y_{ij}=1|X_{ij}, \alpha_{ki}) = Log\left(\frac{\Pr(Y_{ij}=1|\alpha_{1i}, X_{ij})}{\Pr(Y_{ij}=0|\alpha_{1i}, X_{ij})} \right) = \alpha_{1i} + \beta_1 X_{ij}$$

$\alpha_{ki} \sim N(\alpha_k, \tau_k{}^2)$, for each model k

random

only within school effects

used both within and between school effects

Results: Gender Effect

Model	$exp(\beta_k)$	$exp(\gamma_k)$
1	1.65 (1.35, 2.01)	
2	1.68 (1.37, 2.05)	0.46 (0.13, 1.66)
3	1.68 (1.37, 2.05)	0.78 (0.22, 2.74)
4	1.68 (1.37, 2.05)	
5		0.78 (0.22, 2.69)

Model 1 interpretation: **For a given school**, the odds of being reading proficient are 1.65 times greater for female than male student.

The attempt here is to estimate the "within-schooleffect" but the result can be confounded by the "among-school effect"
- If the among-school effect is very different from the within-school effect, then the within estimate can be markedly different then the true within-school effect, confounded differences among schools.

why might within + between school effects of gender differ?

Results: Gender Effect

Model	$exp(\beta_k)$	$exp(\gamma_k)$
2	1.68 (1.37, 2.05)	0.46 (0.13, 1.66)
3	1.68 (1.37, 2.05)	0.78 (0.22, 2.74)
4	1.68 (1.37, 2.05)	

Models 2, 3 and 4 estimate the within-school gender effect!

For a given school (i.e. conditioning on the school-specific random intercept), the odds of being proficient in reading for girls are 68 percent greater than the odds for males.

The odds of being proficient in reading for girls are 1.68 times the odds for males **from the same school**.

Results: Gender Effect

Model	$exp(\beta_k)$	$exp(\gamma_k)$
2	1.68 (1.37, 2.05)	0.46 (0.13, 1.66)

$$Model\ 2:\ \ LogOdds(Y_{ij} = 1 | X_{ij}, \bar{X}_{i.}) = \alpha_{2i} + \beta_2 X_{ij} + \gamma_2 \bar{X}_{i.}$$

Comparing two schools with the same random effect value, the school with more females has a lower rate (population odds) of being proficient in reading ($\boldsymbol{exp(\gamma_k) < 1}$).

NOTE: the calculated proportion of females goes from 0 to 1. So the contextual effect of 0.46 is really comparing the scenario where we go from an all boys school to an all girls school.
- We can rescale the contextual effect, by taking $\boldsymbol{exp(\gamma_k/10)}$ = 0.93.

Comparing two schools that differ in the proportion of female students by 10% and have the same random effect values, a girl from the school with the higher proportion of female students has odds of being proficient in reading that are 7 percent smaller than a girl from the other school.

can we identify these 2 schools?
counterfactual alternative: school i with and without 10% more girls

Results: Gender Effect

Model	$exp(\beta_k)$	$exp(\gamma_k)$
2	1.68 (1.37, 2.05)	0.46 (0.13, 1.66)
3	1.68 (1.37, 2.05)	0.78 (0.22, 2.74)
5		0.78 (0.22, 2.69)

The between-cluster effect of gender is estimated to be negative!
i.e. the school log odds of being proficient in reading decreases as the proportion of females in the school increases, holding the school random effect constant (e.g. school).

Again, we can rescale $exp(\gamma_k)$ so that we are describing the between effect for a 10% increase in the proportion of female students.

Increasing the proportion of female students by 10% in a given school decreases the school-odds of reading proficiency by 2% (0.78^0.1 = 0.98)

Results: Gender Effect

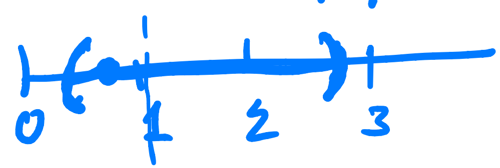
Model	$exp(\beta_k)$	$exp(\gamma_k)$
2	1.68 (1.37, 2.05)	0.46 (0.13, 1.66)
3	1.68 (1.37, 2.05)	0.78 (0.22, 2.74)
5		0.78 (0.22, 2.69)

Summarize the within and between-cluster effects of gender.

1) The within-school effect of being female is positive
2) The estimated contextual effect of gender is 0.78
   - This implies that the estimated within- and between-effects of gender ~~are the same~~. "are consistent with the pop values being the same"
   - You can test this within Model 3!

.78

.22 2.74



same within + between

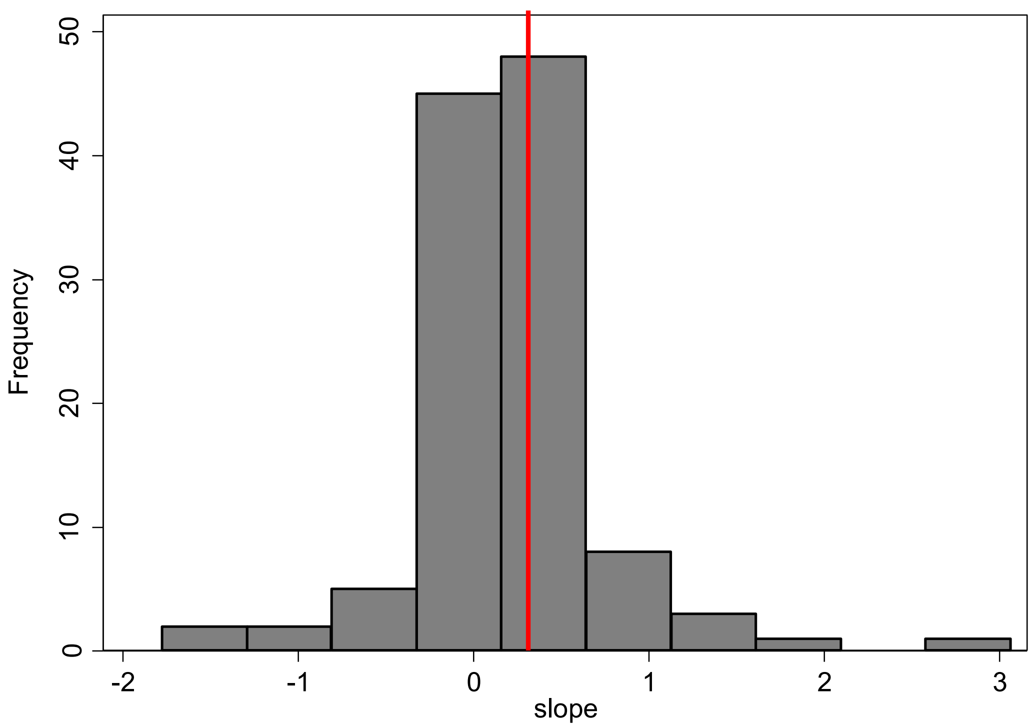
```
. test female_center = mean_female

 ( 1)  [eq1]female_center - [eq1]mean_female = 0

           chi2(  1) =    1.40
         Prob > chi2 =    0.2373
```

Within-school effects of socioeconomic index

- Fit a school specific logistic regression model for reading proficiency as a function of ISEI (scaled per 10 point increase)

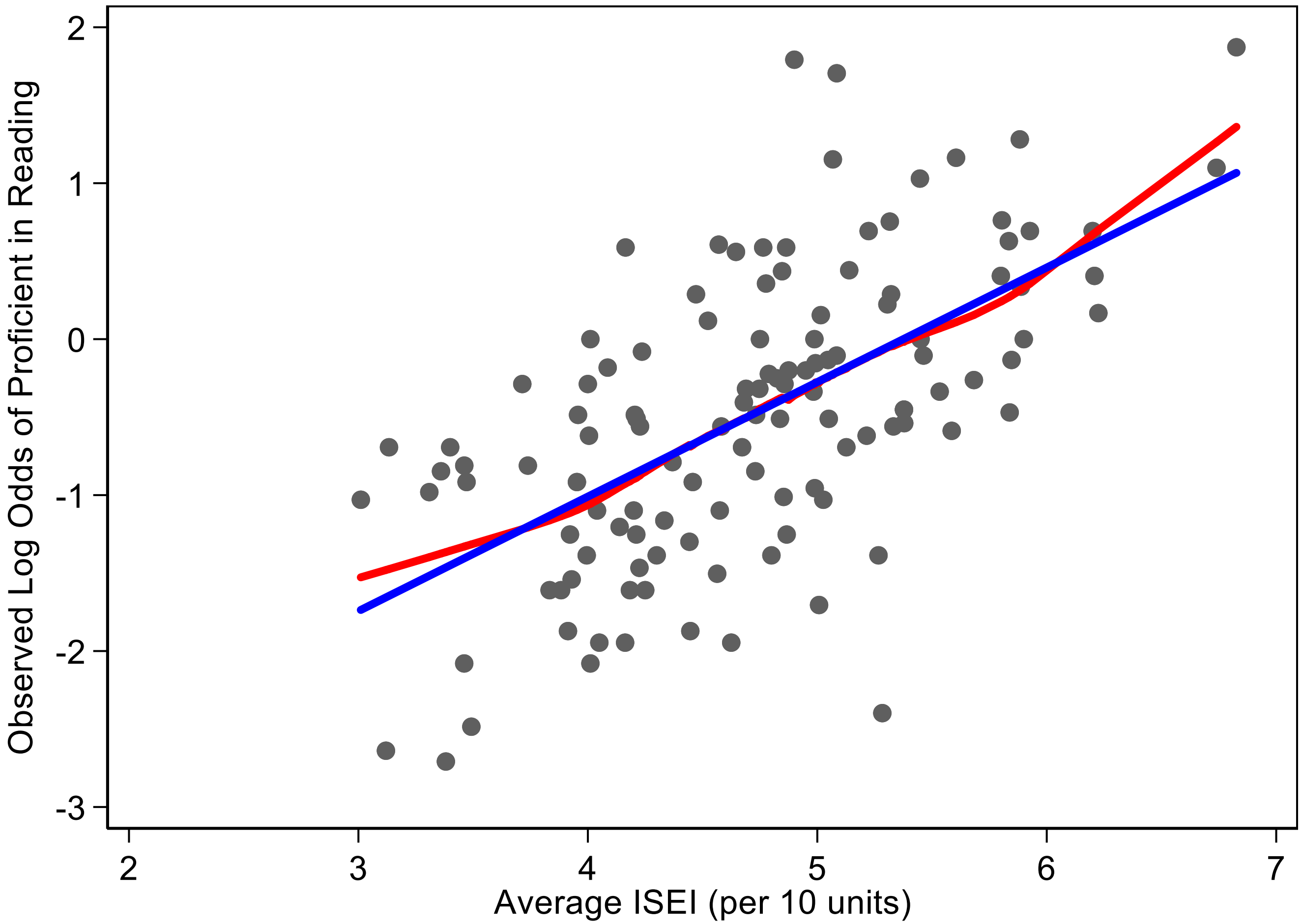


Mean = 0.20

Exp(0.20) = 1.20

Increase in odds of reading proficiency per 10 point increase in ISEI

Between-school effects of socioeconomic index



Results: ISEI Effect

Model	$exp(\beta_k)$	$exp(\gamma_k)$
1	1.29 (1.21, 1.37)	
2	1.20 (1.13, 1.28)	2.03 (1.69, 2.43)
3	1.20 (1.13, 1.28)	2.44 (2.05, 2.90)
4	1.21 (1.13, 1.28)	
5		2.41 (2.03, 2.85)

Some observations:

1) The between-school effect is roughly twice the size of the within-school effect (Model 3) and statistically significantly different ($p < 0.001$)
2) Contextual effect is positive and statistically significant
3) Model 1 total effect is larger than the within-school effect due to the large/positive between-school effect.

why might SES create larger between -vs- within school effects?

Results: ISEI Effect

Model	$exp(\beta_k)$	$exp(\gamma_k)$
2	1.20 (1.13, 1.28)	2.03 (1.69, 2.43)
3	1.20 (1.13, 1.28)	2.44 (2.05, 2.90)

Interpretations:

1. Within a school, the estimated odds of being reading proficient increase by 20 percent per 10-point increase in the ISEI.
2. For a given school, increasing the school average ISEI would results in higher odds of reading proficiency. Odds increase by a factor of 2.44 per 10-point increase in the school-average ISEI
3. Consider two students in the same s chool that have the same ISEI, the student who comes from the school with higher average ISEI has increased odds of being proficient in reading.

how?

Marginal vs. Conditional Estimates

- Fit Model 2 using logistic regression model, no random effects
- Estimation based on GEE assuming exchangeable within-school correlation structure.
- Compare the estimates from those obtained using the random intercept logistic model.
  - The RI coefficients in general will be larger in absolute value
  - The inferences (estimate/se), CI, tests) will be similar if the marginal model you specify induces the same/similar correlation structure.

Marginal vs. Conditional Estimates

Model	$exp(\beta_k)$	$exp(\gamma_k)$
RI	1.20 (1.13, 1.28) z = 5.96	2.03 (1.69, 2.43) z = 7.69
GEE	1.19 (1.12, 1.26) z = 5.95	1.94 (1.64, 2.30) z = 7.73

- In the GEE model, the within effect has ~~the same~~ similar interpretation as for the RI model.
  - Condition on a school! - RI; condition on X_{ij} and $\bar{X}_i$ values in Marginal Model
- In the GEE model, the contextual effect now does not require that the schools have similar "random effects" only that the schools have mean ISEI that differs by 10 units.

Summary of Lecture 8

- Consider logistic random intercept models to separate between and within-cluster effects
  - Similar decomposition of effects as in the linear model case
- Random effects logistic models will produce estimates that are greater in absolute value than their corresponding marginal model friends
  - Conditional interpretation
  - In two-level data, can fit a marginal model assuming exchangeable within cluster correlation