The Johnson-Neyman

Ann Lazar
ann.lazar@ucsf.edu

Department of Preventive and Restorative Dental Sciences
Department of Epidemiology and Biostatistics
University of California, San Francisco

February 21, 2014

Motivation
- Rheumatoid Arthritis – case-control
- High School and Beyond - survey
- CAN DO FV – Phase 3 RCT

The Johnson-Neyman (J-N)

Examples

Final Comment

Motivation: RA Case-Control Example

- Autoantibodies are present years prior to the onset of symptoms of rheumatoid arthritis (RA) and may be highly predictive of the development of RA

- Interest in determining *when* the autoantibodies are elevated prior to diagnosis of RA

Motivation: RA Case-Control Example

- Autoantibodies are present years prior to the onset of symptoms of rheumatoid arthritis (RA) and may be highly predictive of the development of RA

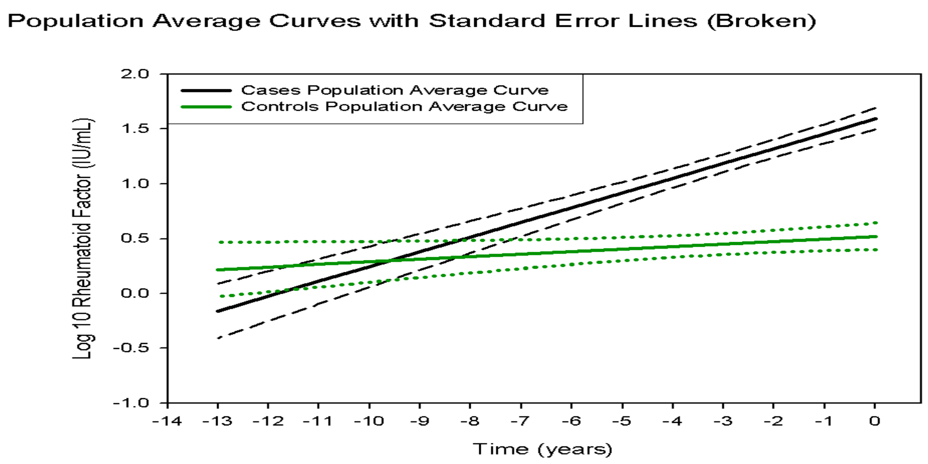
- Interest in determining *when* the autoantibodies are elevated prior to diagnosis of RA

RA Case Control Study

- SERA study
  - retrospective case-control study of RA subjects 13.67 years (4994 days) prior to diagnosis
- Study Sample
  - 166 subjects
    - 83 controls and 83 cases
      - 486 measurements
        - 1 to 4 measurements per subject

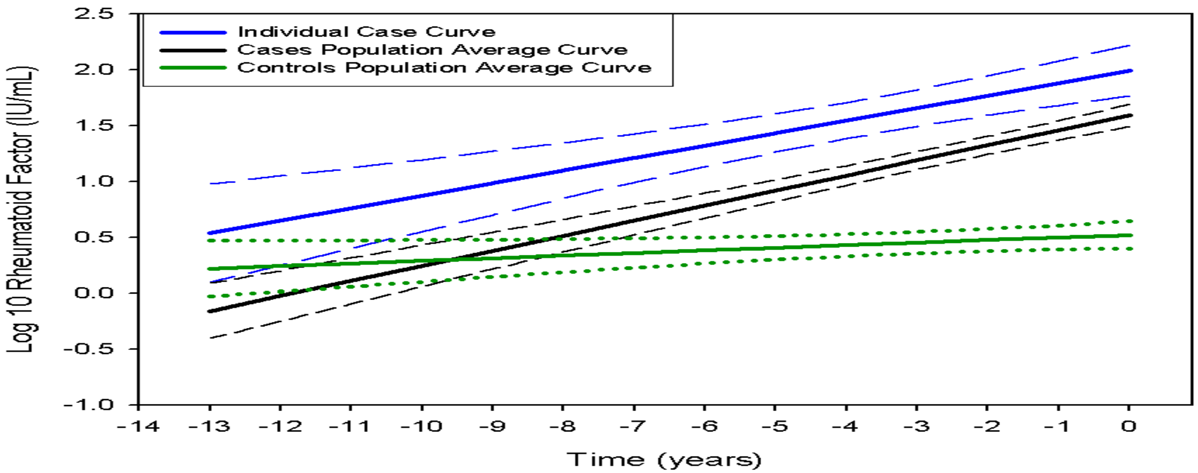
Deane KD, O'Donnell CI, Hueber W, Majka DS, Lazar AA, Derber LA, Gilliland WR, Edison JD, Norris JM, Robinson WH, Holers VM. The number of elevated cytokines and chemokines in preclinical seropositive rheumatoid arthritis predicts time to diagnosis in an age-dependent manner. Arthritis Rheum. 2010 Nov; 62(11):3161-72.

RA Case Control Study



RA Case Control Example

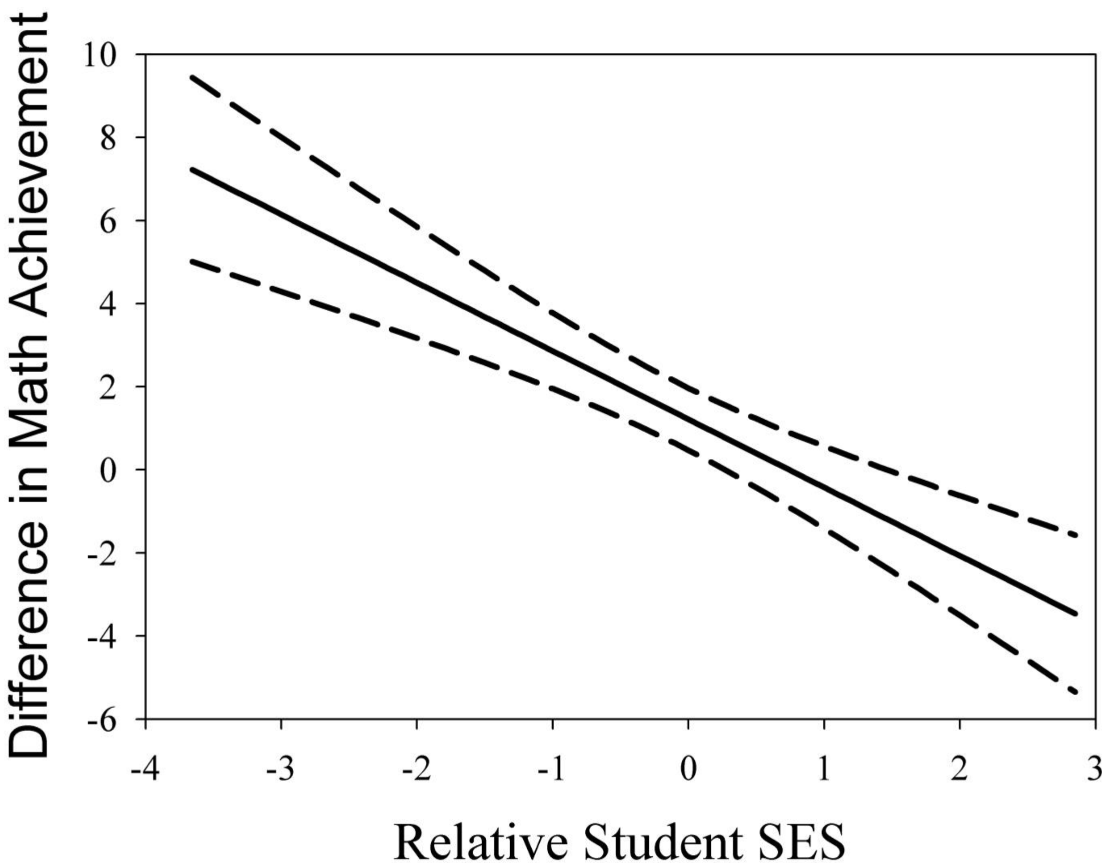
Figure 1: Population Average Curves and an
Individual RA Curve with Corresponding Standard Error Lines (Broken)



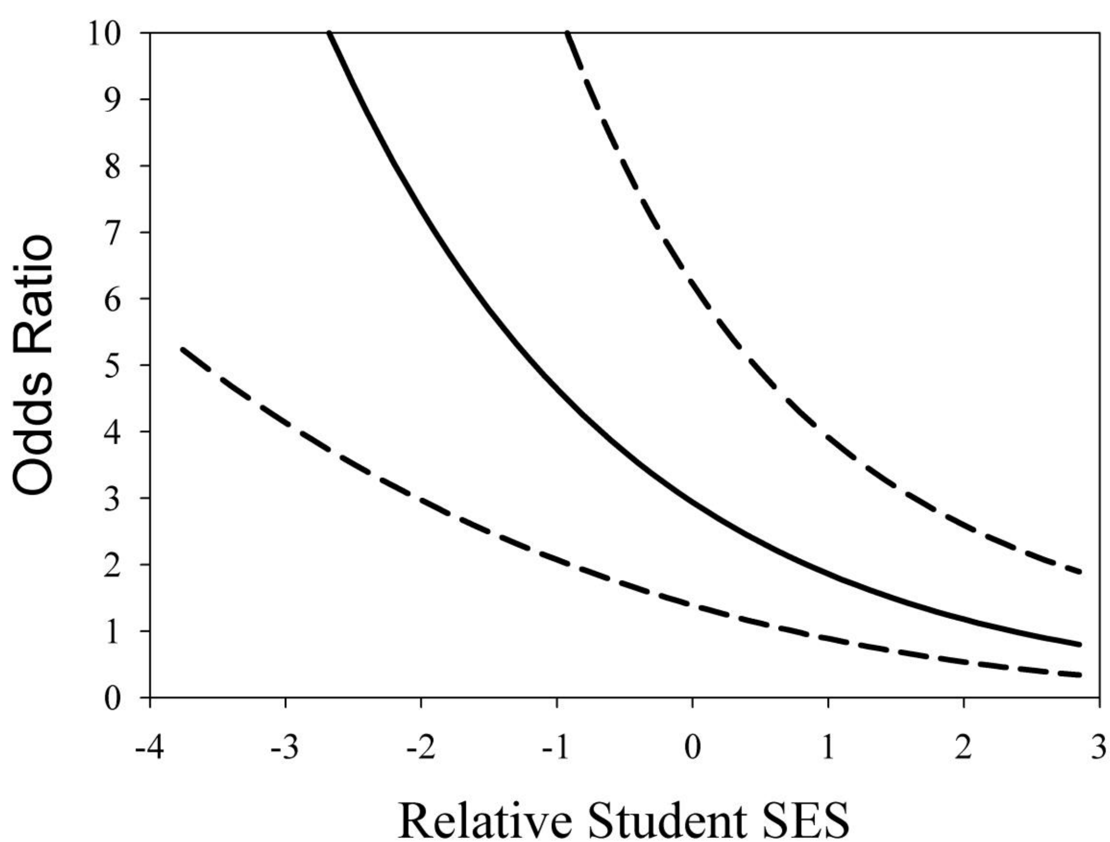
High School and Beyond Survey

- Study the relationship between mathematics achievement score (MA) and socioeconomic status score (SES) between sectors: Catholic (n=70) and public (n= 90) high schools
  - SES = a composite score of parental income, education, occupation and achievement
  - A subset of the data from the High School and Beyond Survey (HSBS) of 7,185 students nested within 160 high schools, which averaged 45 students per school
  - This study can address important questions:
    - 1) What range of SES does the mathematics achievement scores statistically differ between private and public high schools
    - 2) Which school sector does better, Catholic or Public High Schools?

High School and Beyond Survey



High School and Beyond Survey



CAN DO Fluoride Varnish (FV)

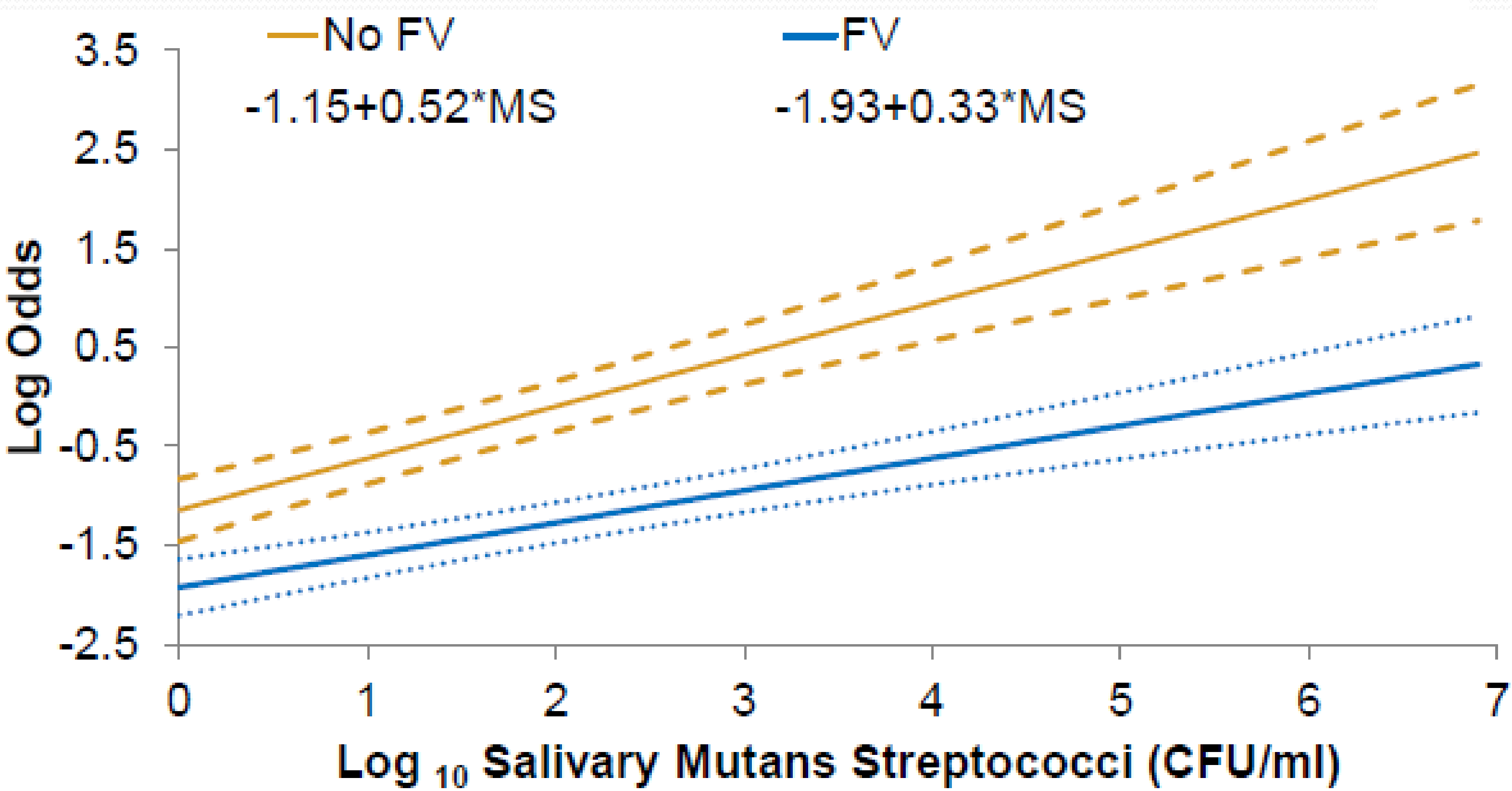
- We are interested in detecting and exploring patterns of heterogeneity of treatment effect (e.g., odds ratio).
  - Identify *people* who respond differently to treatment.
  - *Identify who* may benefit the most (or the least) from a treatment so treatment can be ***tailored*** to the individual
- We proceed by studying
  - Differences (heterogeneity) between groups for varying values of a baseline covariate based on a model

CAN DO Fluoride Varnish (FV)

- Variables
  - Primary Outcome (Dichotomous): Any Caries Incidence
  - Treatment: FV + parental counseling *vs.* counseling alone
  - Baseline Covariate (Continuous): Salivary Mutans Streptococci (MS) (CFU/ml))
- Study Sample
  - 249 caries-free children w/ baseline MS
  - 0 FV: n=89; 1F/Yr : n=78 ; 2FV/Yr : n=82
- Goal
  - At what values of MS does FV treatment result in statistically different (heterogeneity) caries incidence?
  - Who benefits (most) from treatment?

CAN DO Fluoride Varnish (FV)

Logistic Regression Model of Caries Incidence v. No Caries Incidence with Associated Standard Error Lines



Johnson-Neyman (J-N) Approach

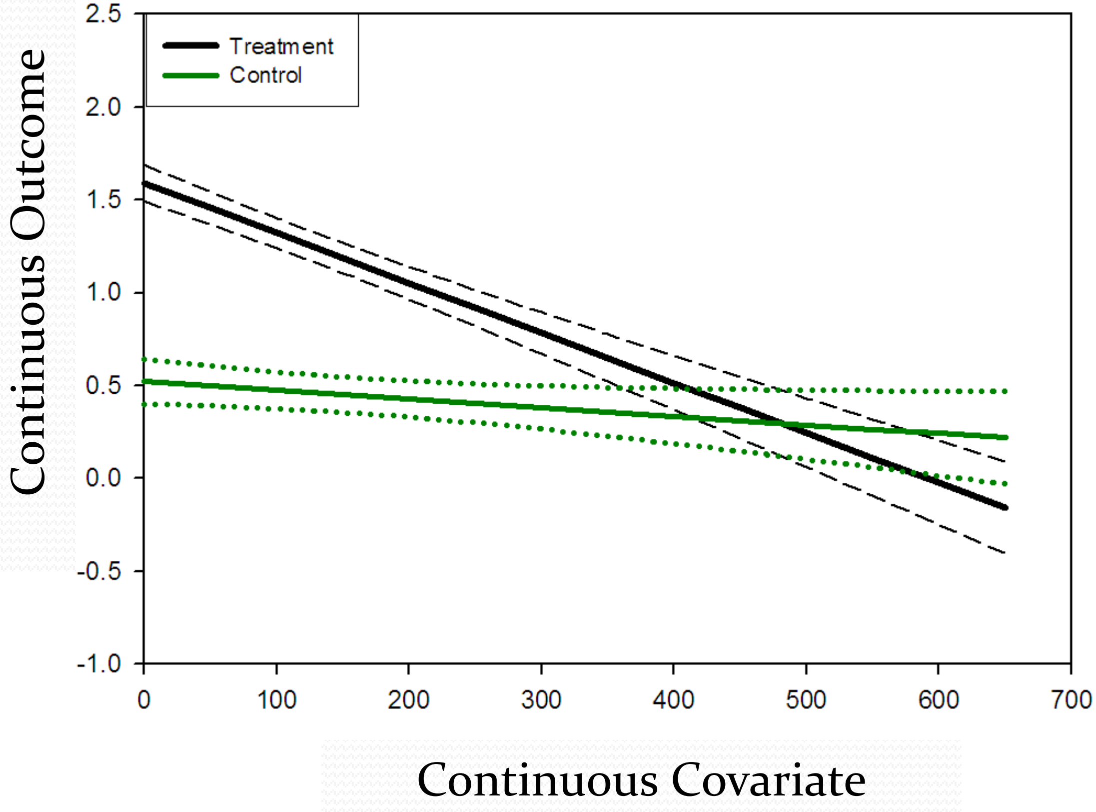
The Johnson-Neyman Approach

- **Aim:** What range of the independent variable is the difference in the outcome variable statistically significant among groups ("Significance Region")?

- Johnson-Neyman (J-N) (1936) in ANCOVA when the regression lines are not parallel
  - Solves the problem of identifying regions of significance or "Significance Region"

- Tests whether the difference in means for a particular value of the covariate between two groups is statistically significance
  - comparing computed *F* statistics to the critical value of Fisher's F distribution
  - Assumes normality and homogeneity of variance

Figure 1: Population Average Curves
with Corresponding Standard Error Lines (Broken Lines)



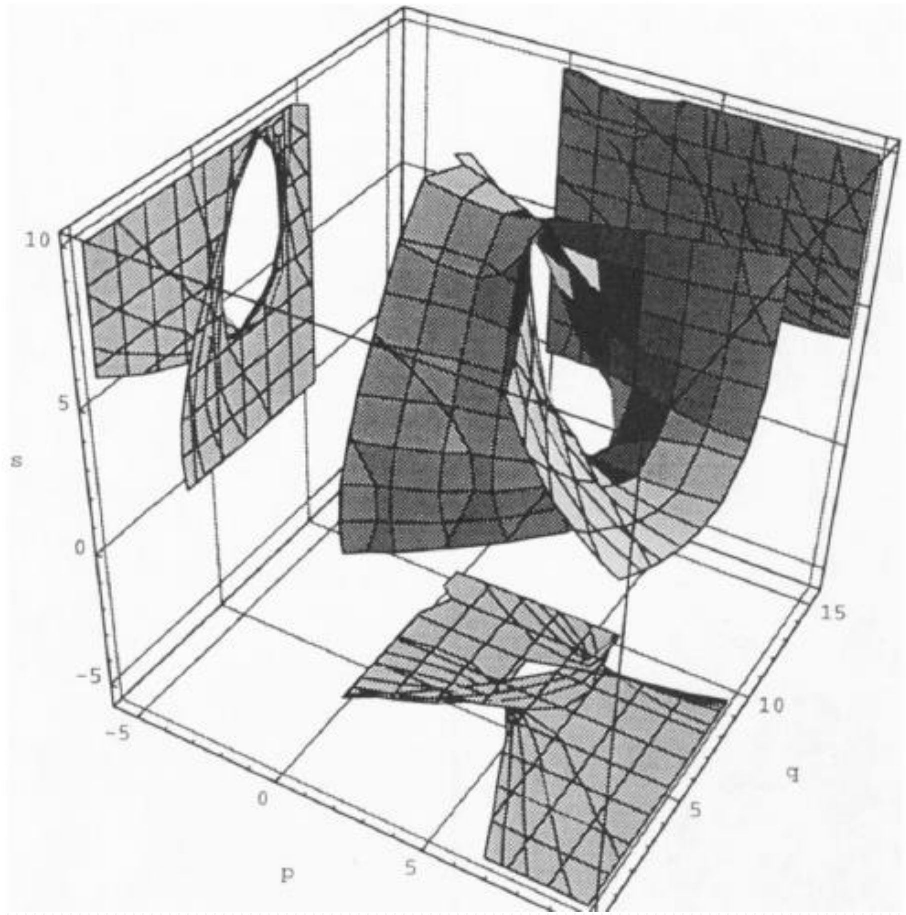
Improvements to J-N Approach

- Huitema (1980) "explicit solution", a simple formula that involves solving a quadratic equation, available for comparing several groups but only when the number of covariates is one

- For two or more covariates, the explicit solution is *thought* to be intractable

- Hunka and Leighton (1997) developed a solution for any number of possible covariates by casting the equation within a general linear model framework. But this approach cannot be solved directly, symbolic processing capabilities of computational software (e.g., Mathematica_

Improvements to J-N Approach – Hunka and Leighton (1997) 3 variables of interest



"Two Parabloids"

Improvements to J-N Approach

- Miyazaki and Maier (M-M) (2005)
  - developed a procedure for determining the significance regions for correlated Gaussian distributed data involves fitting hierarchical linear mixed models (HLM)
  - solved for significance region using the symbolic processing capabilities (e.g., Mathematica)

- Above solutions are available for comparing groups

Improvements to J-N Approach

- Significance regions (solutions) for both correlated Gaussian and non-Gaussian distributed data suitable for generalized linear (mixed) models (GL(M)M), which includes HLMs. The proposed solution can :
  - Compare subjects and groups
  - Adjust for covariates
  - Account for Multiple Testing
  - Makes use of one software package to implement the model and solution, as follows…

The GLMM Framework

- The conditional mean of Y depends on the fixed (β) and random effects (u) via the linear predictor $\eta = X\beta + Zu$, where $g\{E(Y|u)\} = \eta = X\beta + Zu$

$$H_0\!: \underbrace{\theta}_{sx1} = \underbrace{C}_{sx(p+q)} \begin{pmatrix} \underbrace{\beta}_{px1} \\ \underbrace{u}_{qx1} \end{pmatrix} = C_1\beta + C_2 u = 0$$

- $H_0\!: \theta = 0$ can be tested with a t- or F-statistic.

Let $\hat{V}_{\hat{\theta}} = C\hat{L}C'$ and $\hat{L}$ is the empirical covariance matrix of $\hat{\beta} - \beta$ and $\hat{u} - u$; s=the rank of C, the contrast matrix

$$\mathrm{t}(\hat{\theta}) = {\hat{\theta}}\Big/{\sqrt{\hat{V}_{\hat{\theta}}}} \text{ and } F(\hat{\theta}) = {\left[\underbrace{\hat{\theta}'}_{1xs} \Big(\underbrace{\hat{V}_{\hat{\theta}}}_{sxs}\Big)^{-1} \underbrace{\hat{\theta}}_{sx1}\right]}\Big/{s}$$

Solutions

Let h identify subsets of linear function of $\binom{\beta}{u}$

- $H_0\colon\ h\theta = 0$ rejected $(h\theta \neq 0)$

$$t\left(h\hat{\theta}\right) > \sqrt{sF_{1-\alpha,s,v}}$$

- Or we can use our explicit solution to identify the significance region

$$h_x\left[\theta\theta' - sF_{1-\propto,s,v}V_\theta\right]h_x > 0$$

- SAS macro available to determine the significance regions

Suppose θ_C represent the difference in intercepts; $\theta_A$denote the difference in slopes, then $\theta = \begin{pmatrix} \theta_C \\ \theta_A \end{pmatrix}$

Let $V_{\theta(C)}, V_{\theta(A)}$, and $Cov_{\theta(A)\theta(C)}$ be the empirical prediction error variance (V) and covariance (Cov) of θ_C and θ_A

$$\begin{pmatrix} 1 & X \end{pmatrix} \left[\begin{pmatrix} \theta_C^2 & \theta_C\theta_A \\ \theta_A\theta_C & \theta_A^2 \end{pmatrix} - sF_{1-\alpha,s,v} \begin{pmatrix} V_{\theta(C)} & Cov_{\theta(C)\theta(A)} \\ Cov_{\theta(A)\theta(C)} & V_{\theta(A)} \end{pmatrix} \right] \begin{pmatrix} 1 \\ X \end{pmatrix} > 0$$

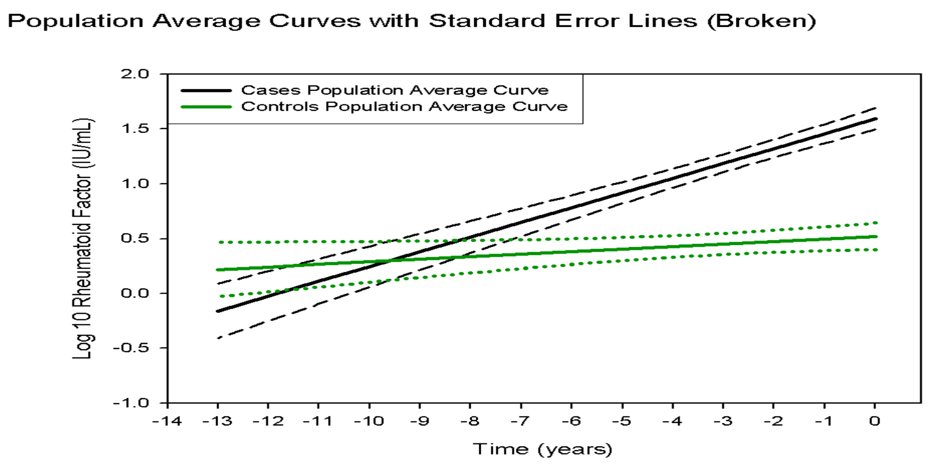
$Let\ A = \theta_A^2 - sF_{1-\alpha,s,v}V_{\theta(A)},\ \ B = 2\left(\theta_A\theta_C - sF_{1-\alpha,s,v}Cov_{\theta(A)\theta(C)}\right),$

$C = \theta_C^2 - sF_{1-\alpha,s,v}V_{\theta(C)}$, and the discriminant, $D = B^2 - 4AC$

Let $X_1 = \left[\frac{-B-\sqrt{D}}{2A}\right]$ and $X_2 = \left[\frac{-B+\sqrt{D}}{2A}\right]$

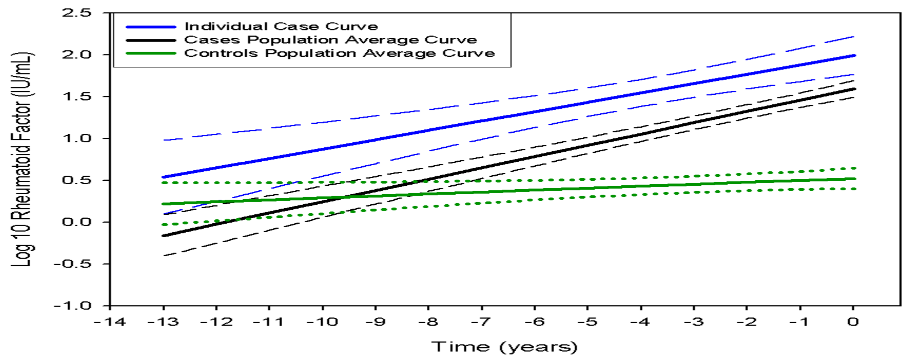
- Case I (A>0 implies D>0) : $X < X_1$ or $X > X_2$
- Case II (A<0 and D>0) : $X_2 < X < X_1$
- Case III (D < 0 and A < 0): no region
- A proof of these 3 cases has been derived

Examples-RA Case Control Study



Examples- RA Case Control Example

Figure 1: Population Average Curves and an
Individual RA Curve with Corresponding Standard Error Lines (Broken)



Step 1: Determine $\Delta(x)$

- Compare the Cases and Controls Population Average Curves

$$\Delta(x)=Y_1(x)-Y_2(x)=(\alpha_1+\beta_1 x)-(\alpha_2+\beta_2 x)=(\alpha_1-\alpha_2)+x(\beta_1-\beta_2)$$

- Compare a Case's Curve to the Controls Population Average Curve

$$\Delta(x)=Y_1(x)-Y_{11}(x)=(\alpha_2+\beta_2 x)-(\alpha_1+\beta_1 x+a_{11}+b_{11}x)=(\alpha_2-\alpha_1-a_{11})+(\beta_2-\beta_1-b_{11})x$$

- Compare a Case's Curve to their Population Average Curve

$$\Delta(x)=Y_1(x)-Y_{11}(x)=(\alpha_1+\beta_1 x)-(\alpha_1+\beta_1 x+a_{11}+b_{11}x)=-a_{11}-b_{11}x$$

Step 2: Re-write $\Delta(\,x\,)$

- From Step 1:

$$\Delta(x)=Y_1(x)-Y_2(x)=(\alpha_1+\beta_1 x)-(\alpha_2+\beta_2 x)=(\alpha_1-\alpha_2)+x(\beta_1-\beta_2)$$

- Step 2: Re-Write $\Delta(x)=\underbrace{(1 \quad x)}_{h_x}\underbrace{\begin{pmatrix}\alpha_1-\alpha_2\\ \beta_1-\beta_2\end{pmatrix}}_{\theta}=h_x\,\theta$

$$\textit{Note that } \theta=C\begin{pmatrix}\beta\\ u\end{pmatrix}=(C_1 \quad C_2)\begin{pmatrix}\beta\\ u\end{pmatrix}=C_1\beta+C_2 u=\underbrace{\begin{pmatrix}1 & 0 & -1 & 0\\ 0 & 1 & 0 & -1\end{pmatrix}}_{C_1}\underbrace{\begin{pmatrix}\alpha_1\\ \beta_1\\ \alpha_2\\ \beta_2\end{pmatrix}}_{\beta} \textit{ with } C_2=0$$

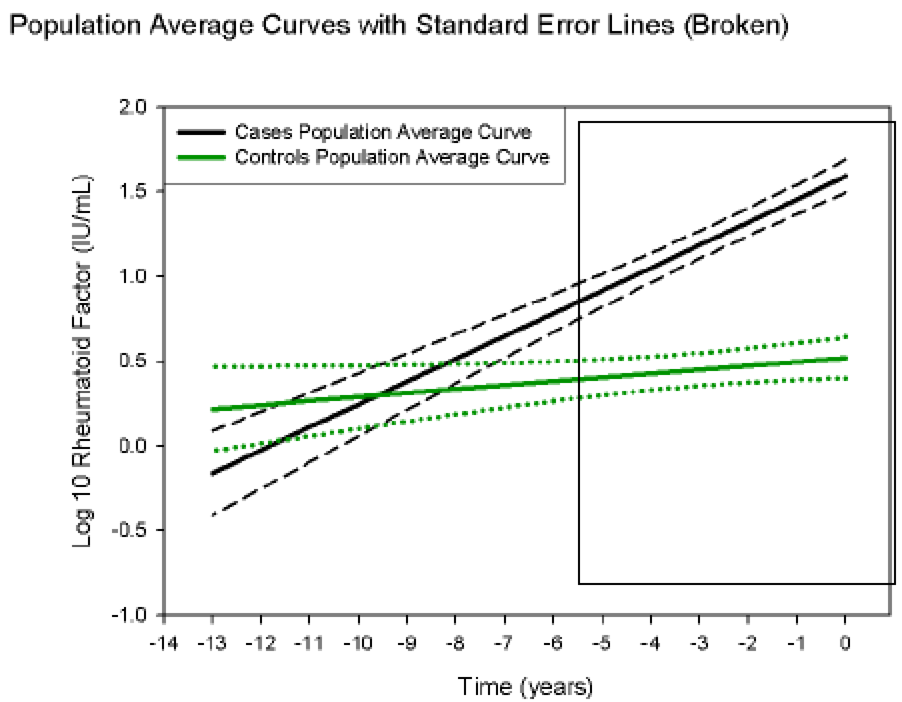
Step 3: Determine the set of x's for which the times response curves differ

- From Step 1: Determine $\Delta(x)$
  - $\Delta(x) = Y_1(x) - Y_2(x) = (\alpha_1 + \beta_1 x) - (\alpha_2 + \beta_2 x) = (\alpha_1 - \alpha_2) + x(\beta_1 - \beta_2)$
- From Step 2: Re-Write $\Delta(x)$
  - $h_x \theta = \underbrace{(1 \quad x)}_{h_x} \underbrace{\begin{pmatrix} \alpha_1 - \alpha_2 \\ \beta_1 - \beta_2 \end{pmatrix}}_{\theta} = h_x \underbrace{\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}}_{C_1} \underbrace{\begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix}}_{\beta}$
- Step 3: Determine the set of x's for which this inequality holds,
  - $t^2(h_x \hat{\theta}) = \left[\frac{h_x \hat{\theta}}{\sqrt{h_x \hat{V}_{\hat{\theta}} h'_x}} \right]^2 > sF_{1-\alpha, s, v}$

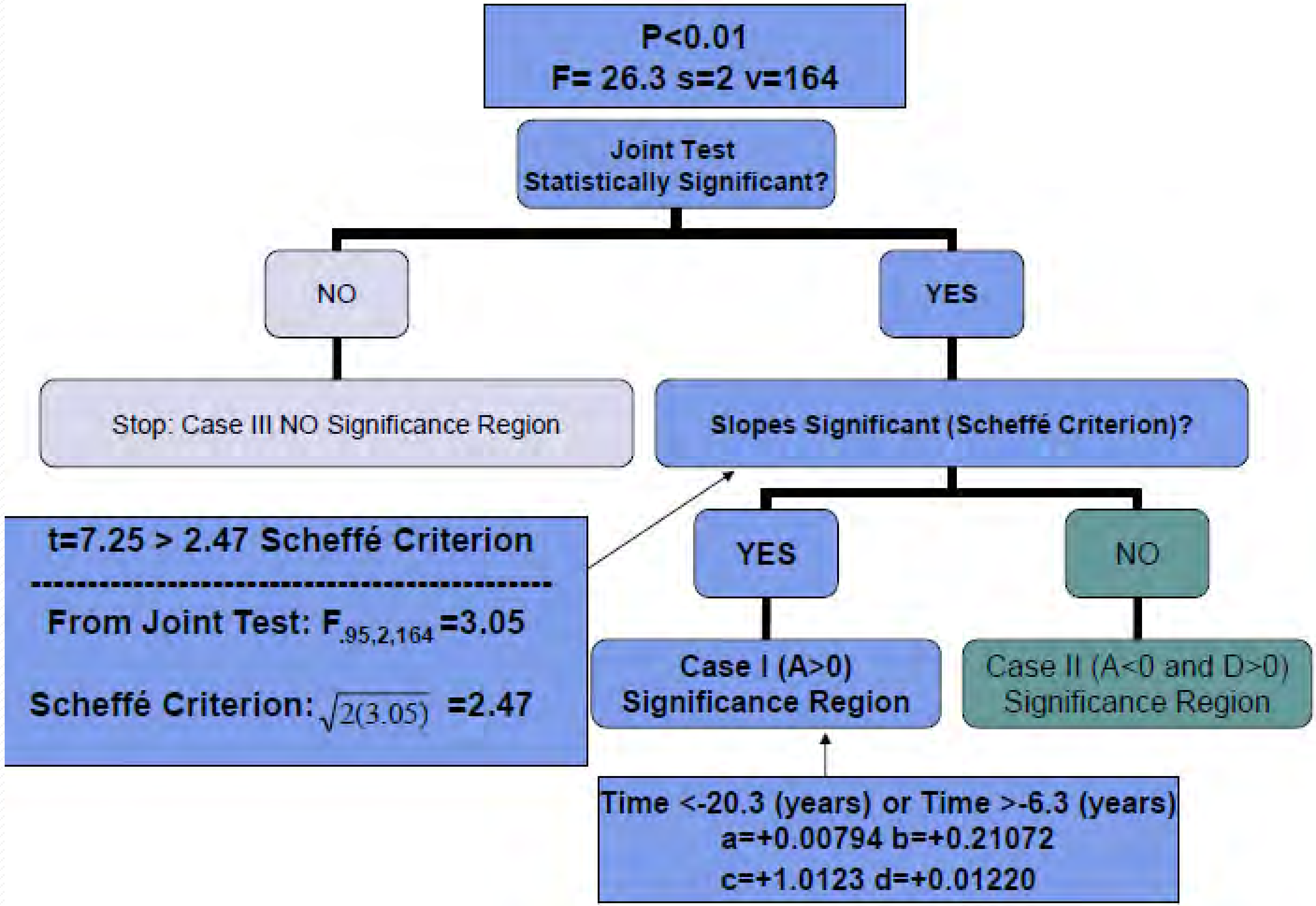
- Steps
  - √Step 1: Determine $\Delta(x)$
  - √Step 2: Re-Write $\Delta(x)$ in terms of the difference between the intercepts and the difference between the slopes in matrix notation
  - √ Step 3: Determine the set of *x*'s for which the time response curve differ
  - √ Step 4: Re-Write Step 3 in Quadratic Form such that

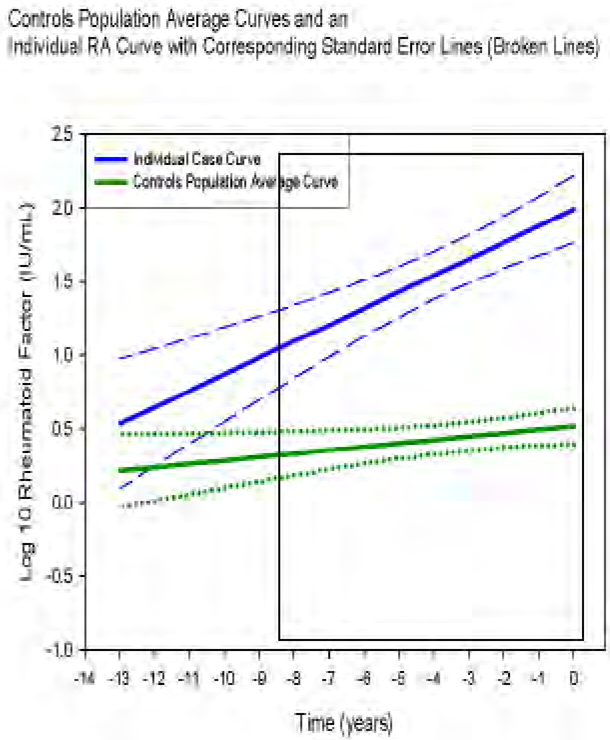
    $$h_x\left[\hat{\theta}\hat{\theta}' - sF_{1-\alpha,s,v}\hat{V}_{\hat{\theta}}\right]h_x' > 0$$

    and determine A, B, C and D from the quadratic formula, $AX^2+BX+C>0$
  - Step 5: Determine the Significance Region

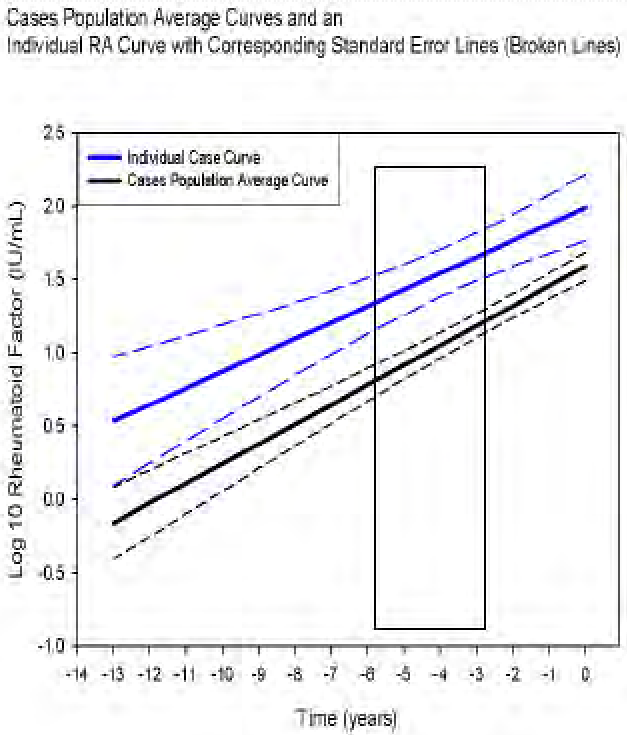


-6.3 years to diagnosis (0 years)


P<0.01
F= 26.3 s=2 v=164
Joint Test
Statistically Significant?
NO
YES
Stop: Case III NO Significance Region
Slopes Significant (Scheffé Criterion)?
t=7.25 > 2.47 Scheffé Criterion
From Joint Test: $F_{.95,2,164}$ =3.05
Scheffé Criterion: $\sqrt{2(3.05)}$ =2.47
YES
NO
Case I (A>0)
Significance Region
Case II (A<0 and D>0)
Significance Region
Time <-20.3 (years) or Time >-6.3 (years)
a=+0.00794 b=+0.21072
c=+1.0123 d=+0.01220

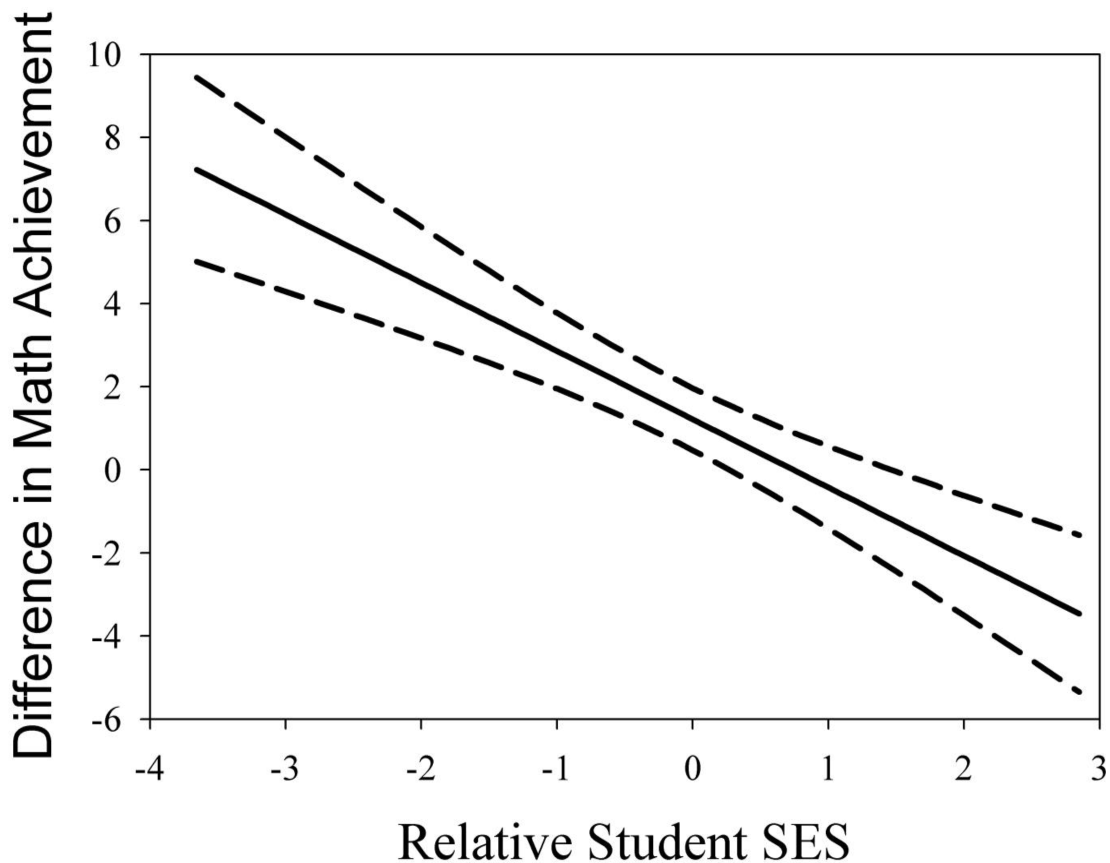


-8.3 years to 0 years (diagnosis)



-5.9 years to -2.5 years

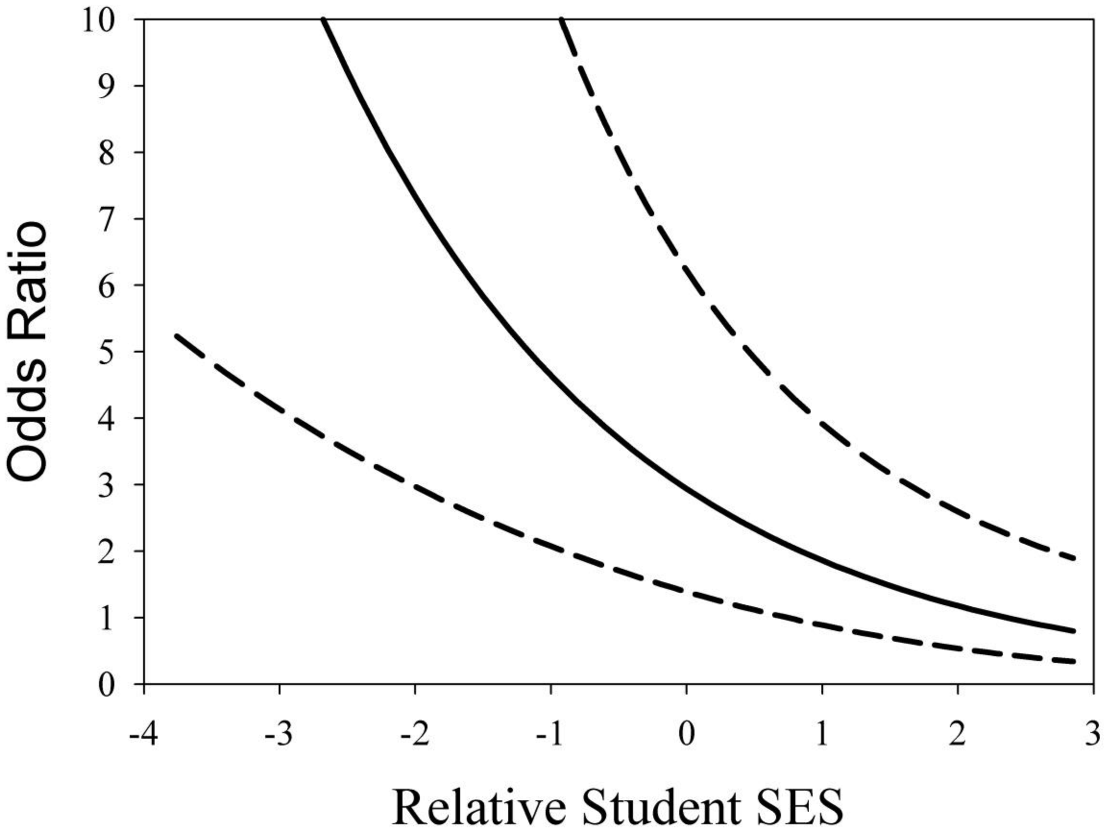
High School and Beyond Survey



1) What range of SES does mathematics achievement scores statistically differ between private and public high schools
2) Which school sector does better,

Catholic or Public High Schools?

High School and Beyond Survey



What range of SES does the mathematics achievement scores statistically differ between the Catholic high school sector and a particular public high school (for example, school 1296)

Does the Catholic high school sector have higher math scores than a particular public high school?

- Fit generalized linear mixed models (GLMM) of mathematics achievement score, y_{ij}, for the ith student in the jth cluster (school)

$$y_{ij} = \beta_{0j} + \beta_{1j}\left(SES_{ij} - \overline{SES}_j\right) + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}\,\overline{SES}_j + \gamma_{02}\ sector_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}\overline{SES}_j + \gamma_{12}\ sector_j + u_{1j}$$

- $\overline{SES}_j$ denotes the school-averaged student SES
- sector equals 1 for Catholic (C) schools or 0 for public (P) schools
- e_{ij} represents the unobservable random vector of normally distributed errors
- e_{ij} were assumed uncorrelated with the multivariate normally distributed random effects, u_{ij}, of the school mean, (u_{0j}), and SES-achievement slope (u_{1j}).

$E_C(Y_{ij}) - E_P(Y_{ij}) = \gamma_{02} + \gamma_{12}(SES_{ij} - \overline{SES}_j)$, where $\underbrace{(1 \quad x)}_{h} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{C} \underbrace{\begin{pmatrix} \gamma_{02} \\ \gamma_{12} \end{pmatrix}}_{\beta} =$

$(1 \quad x) \begin{pmatrix} \gamma_{02} \\ \gamma_{12} \end{pmatrix} = h\theta,$

- Results in (over) 3839 null hypotheses of no difference in mathematical achievement scores between Catholic and public schools for each of the 3839 distinct values of relative SES, $(SES_{ij} - \overline{SES}_j)$, or X.

*Intercepts and Slopes Outcome Model of Math Achievement Score: Continuous and Normally Distributed Outcome**

Fixed Effects	*Coefficient*	*Standard Error*	*T-Value*	*Degrees of Freedom*	*p Value*
Model for School Means (β_{0j})					
Intercept (γ_{00})	12.113	0.1988	60.93	157	<0.0001
Mean SES (γ_{01})	5.3391	0.3693	14.46	157	<0.0001
Sector (γ_{02})	1.2167	0.3064	3.97	157	<0.0001
Model for SES-Achievement (β_{1j})					
Intercept (γ_{10})	2.9388	0.1551	18.95	7022	<0.0001
Mean SES (γ_{11})	1.0389	0.2989	3.48	7022	0.0005
Sector (γ_{12})	-1.6426	0.2398	-6.85	7022	<0.0001

*Based on restricted maximum likelihood from SAS Procedure MIXED version 9.1.3

$$\hat{A} = (\hat{\gamma}_{12})^2 - 2F_{(0.95,2,7022)}\hat{V}_{\hat{\gamma}_{12}} = (-1.6426)^2 - (2 * 2.997 * 0.2398^2) = 2.353;$$

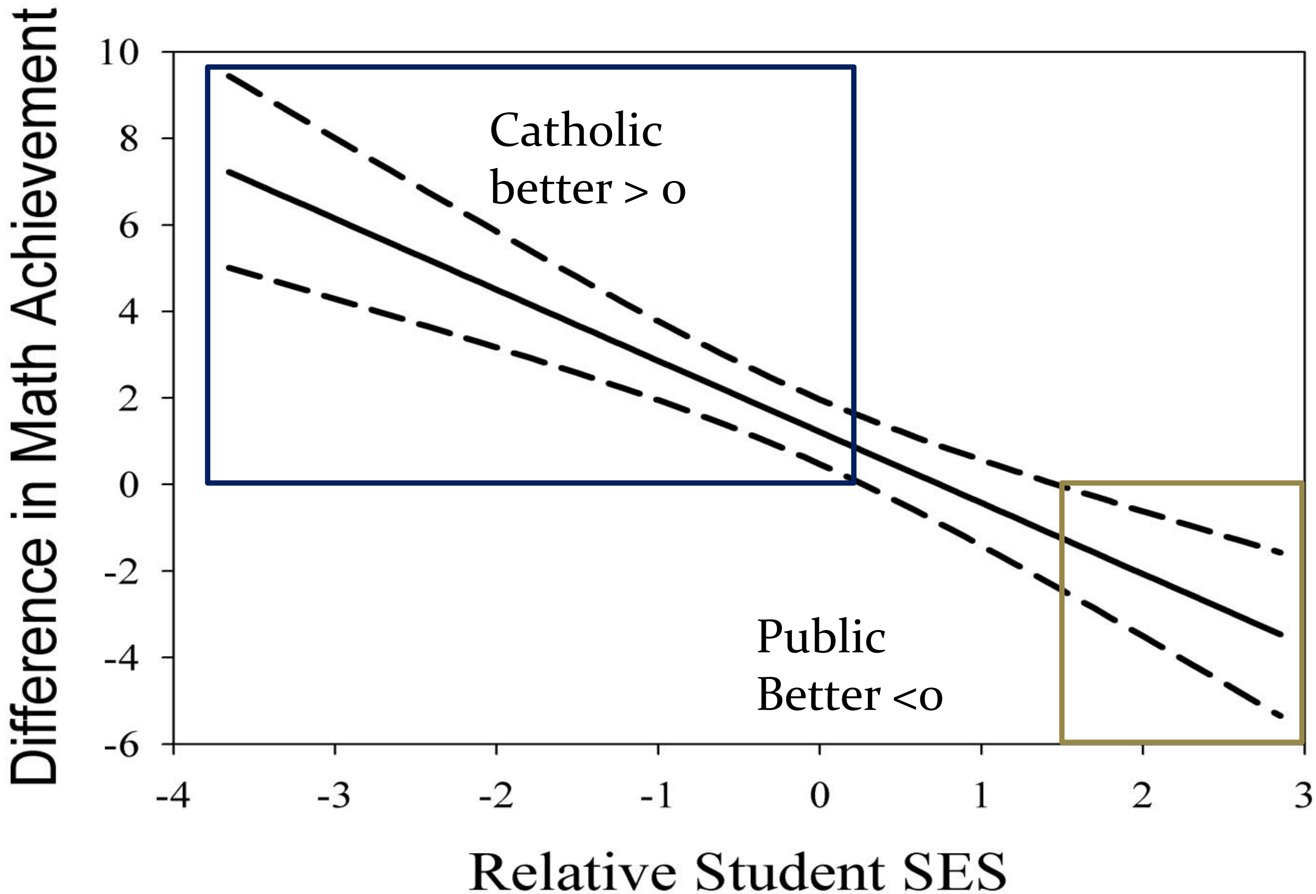
$$\hat{B} = 2 * \left\{(\hat{\gamma}_{02}\hat{\gamma}_{12}) - 2F_{(0.95,2,7022)}\hat{V}_{\hat{\gamma}_{02}\hat{\gamma}_{12}}\right\}$$

$$= 2 * \{(1.2167 * -1.6426) - (2 * 2.997 * 0.0060)\} = -4.0690;$$

$$\hat{C} = (\hat{\gamma}_{02})^2 - 2F_{(0.95,2,7022)}\hat{V}_{\hat{\gamma}_{02}} = 1.2167^2 - (2 * 2.997 * 0.3064^2) = 0.9176;$$

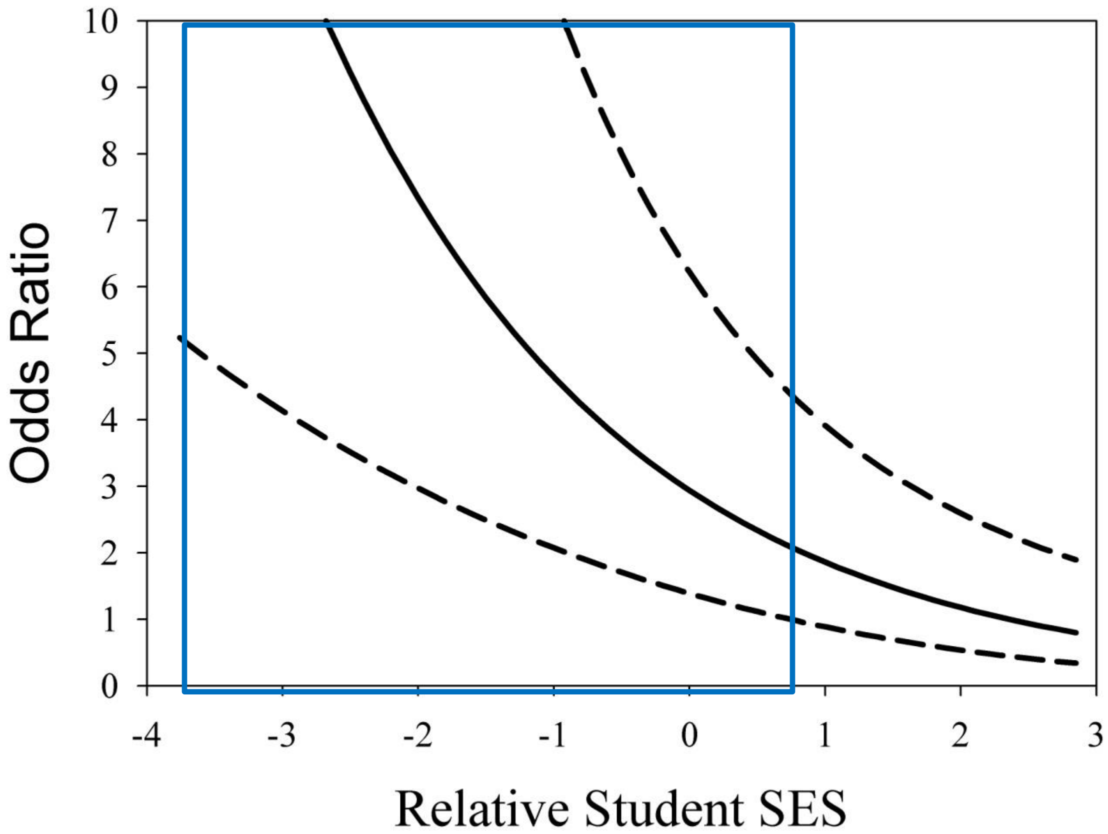
$$\hat{D} = \left(\hat{B}^2 - 4\hat{A}\hat{C}\right) = (-4.0690)^2 - 4 * 2.353 * 0.9176 = 7.920$$

After plugging in $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\widehat{D}$ for Case I; x < 0.266 or x > 1.462



Lazar, A.A. and Zerbe, GO (2011). Solutions for Determining the Significance Region Using the Johnson-Neyman Type Procedure in Generalized Linear (Mixed) Models. *Journal of Education and Behavioral Statistics* **36**(6): 699-719.

High School and Beyond Survey (U.S. High School Students)

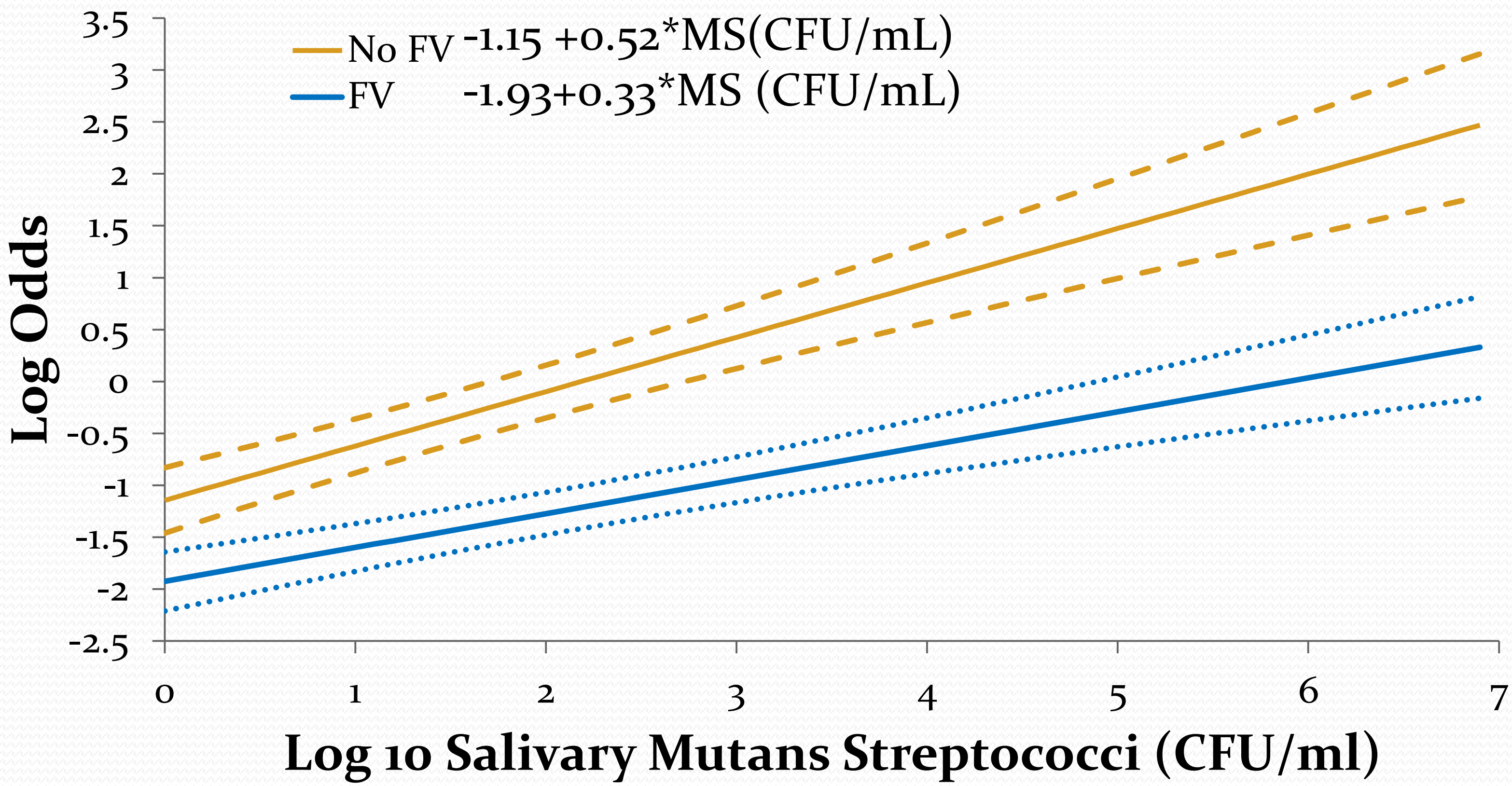


Significance Region: Relative SES : -3.657 and 0.737, students from the Catholic high school sector (n=70) have greater odds of higher mathematics achievement scores than the public school 1296

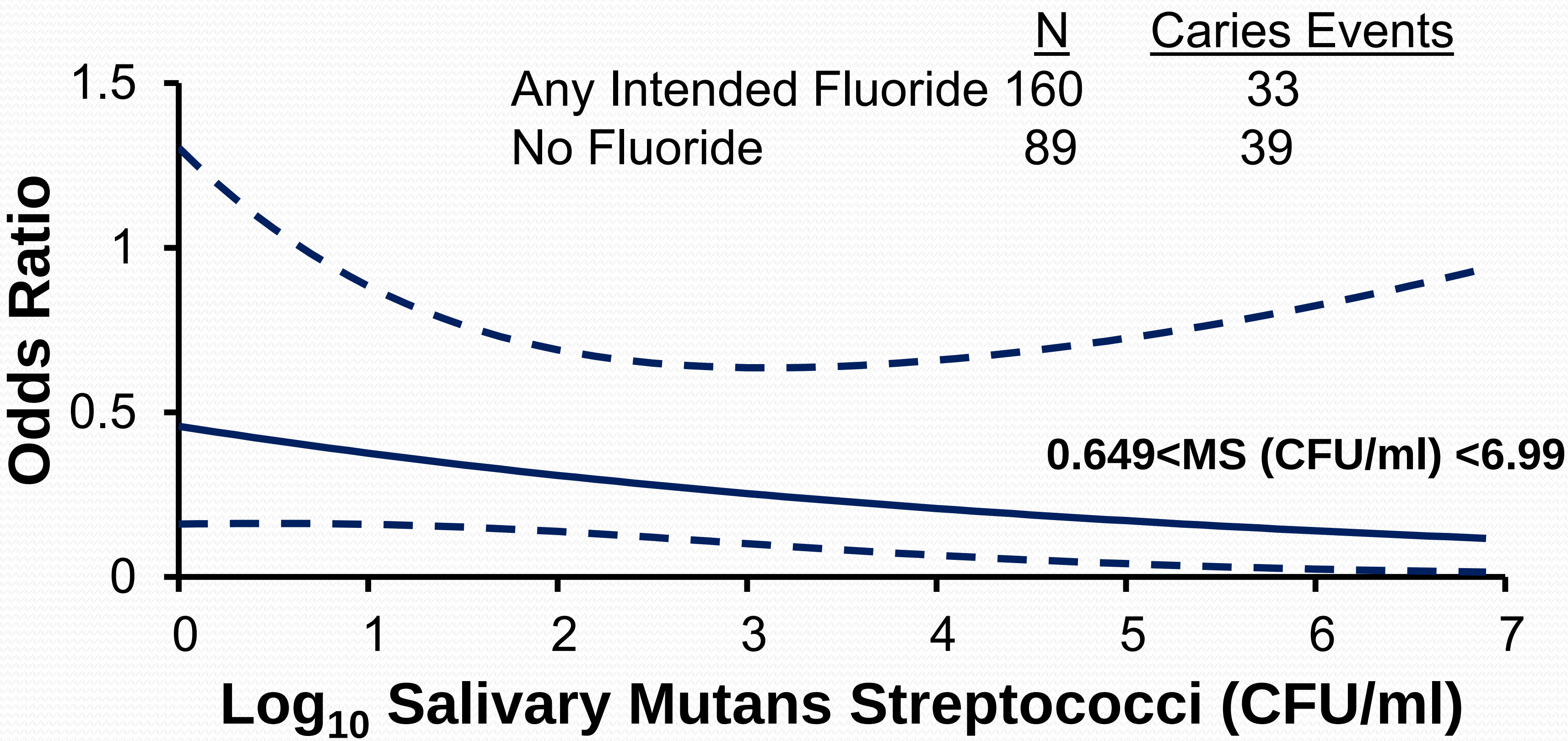
- What range of SES does the mathematics achievement scores statistically differ between the Catholic high school sector (n=70) and a particular public high school (n=1)(for example, school 1296) and

Figure: SAS macro output – OR Lazar, A.A. and Zerbe, GO (2011). Solutions for Determining the Significance Region Using the Johnson-Neyman Type Procedure in Generalized Linear (Mixed) Models. *Journal of Education and Behavioral Statistics* **36**(6): 699-719.

Efficacy of Different Fluoride Varnish Application Frequencies in Preventing Early Childhood Caries Incidence

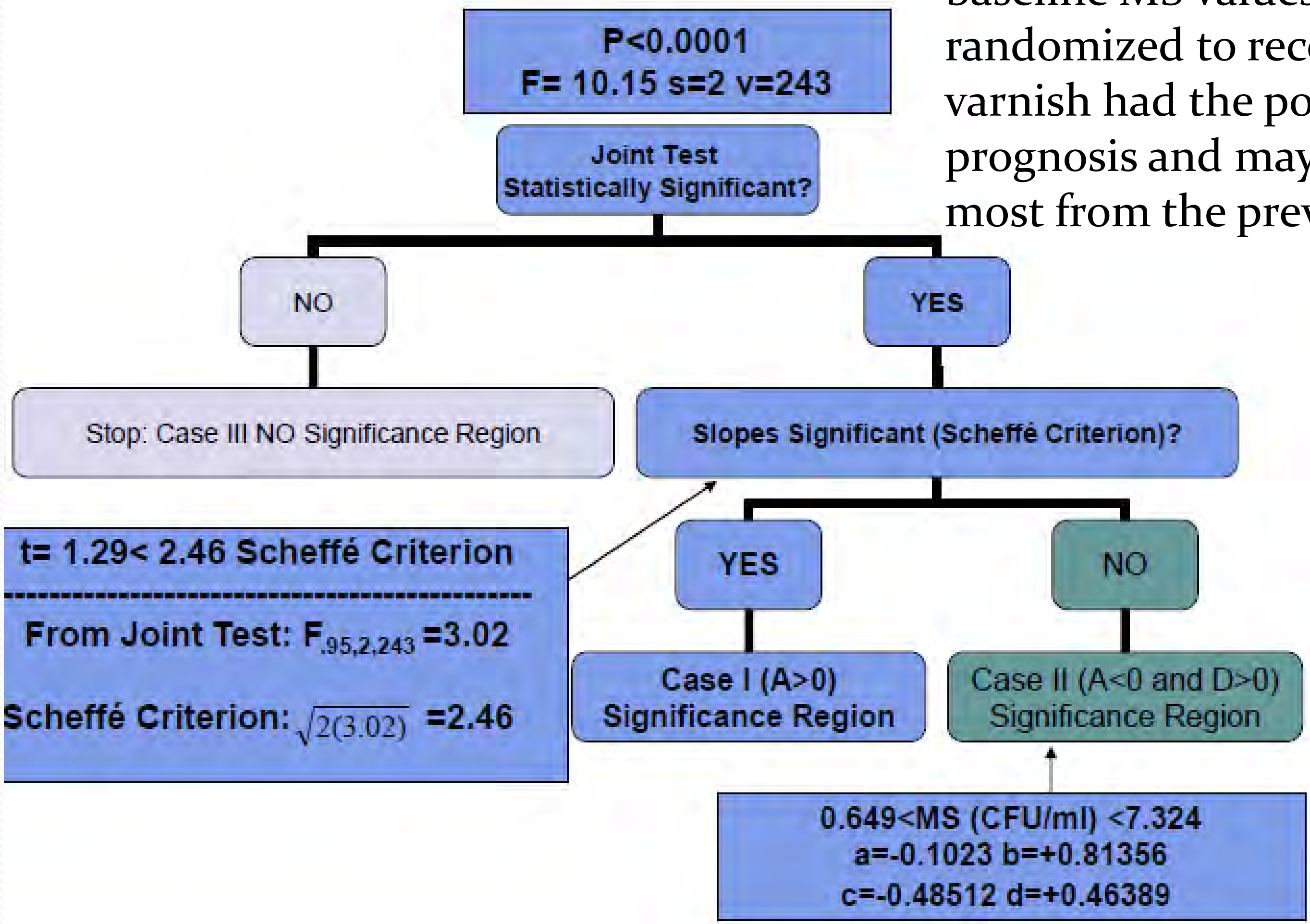


Efficacy of Different Fluoride Varnish Application Frequencies in Preventing Early Childhood Caries Incidence



Lazar, A.A., Gansky, S.A., Halstead, D.D., Slajs, A., Weintraub, J.A. Improving patient care using the Johnson-Neyman analysis of heterogeneity of treatment effects according to individuals' baseline characteristics. *Journal of Dental, Oral and Craniofacial Epidemiology*, accepted for publication.

CAN DO FV - children with higher baseline MS values who were randomized to receive fluoride varnish had the poorest dental caries prognosis and may have benefitted most from the preventive agent



Final Comment

- J-N may be an important tool in the field of personalized health care

- While J-N 'explicit solution' assumes that the covariate has linear effects, the 'grid' solution does not require linearity in the covariates

- J-N can make more detailed and substantial descriptions about the 'significance region'

J-N References

Johnson, P. O., and Neyman, J. (1936). Tests of certain linear hypothesis and their application to some educational problems. *Statistical Research Memoirs* **1**, 57-93.

Hunka, S., and Leighton, J. (1997). Defining Johnson-Neyman regions of significance in the three-covariate ANCOVA using Mathematica. *Journal of Education and Behavioral Statistics* **22**, 361-387.

Raudenbush, S. W., and Byrk, A. S. (2002). *HLMS--Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Newbury Park, CA: Sage

Miyazaki, Y., and Maier, K. (2005). Johnson-Neyman Type Technique in Hierarchical Linear Models. *Journal of Education and Behavioral Statistics* **30**, 233-259.

Lazar, A.A. and Zerbe, GO (2011). Solutions for Determining the Significance Region Using the Johnson-Neyman Type Procedure in Generalized Linear (Mixed) Models. *Journal of Education and Behavioral Statistics* **36**(6): 699-719.

Lazar, A.A., Gansky, S.A., Halstead, D.D., Slajs, A., Weintraub, J.A. (in-press, 2014) Improving patient care using the Johnson-Neyman analysis of heterogeneity of treatment effects according to individuals' baseline characteristics. *Journal of Dental, Oral and Craniofacial Epidemiology*, accepted for publication.

Weintraub JA, Ramos-Gomez F, Jue B, Shain S, Hoover CI, Featherstone JD, Gansky SA. (2006) (Original CAN DO FV report)

Deane KD, O'Donnell CI, Hueber W, Majka DS, Lazar AA, Derber LA, Gilliland WR, Edison JD, Norris JM, Robinson WH, Holers VM. The number of elevated cytokines and chemokines in preclinical seropositive rheumatoid arthritis predicts time to diagnosis in an age-dependent manner. Arthritis Rheum. 2010 Nov; 62(11):3161-72.

Acknowledgements

- Patients and participants of the RA study, High School and Beyond, CAN DO FV
- The Center to Address Disparities in Children's Oral Health (CAN DO)
- NIDCR P30 DE020752 & NCI T32 CA-09337 (AAL)
- NIH NIDCR P60DE013058, U54DE014251, U54DE019285 (CAN DO)

- Thank you for your attention.