Social Statistics

Wrapping Up Regression

December 6, 2021

Assignment 8 Review

1. Create a new binary variable identifying respondents who think they are most likely to find a marriage or long-term partner at Middlebury with a 1 and everyone else with a 0.

```
midd_survey <- mutate(midd_survey,
                      find_partner_midd =
                           ifelse(find_partner=="Middlebury",1,0))
```

What is the mean of this new variable, and what does that value represent?

```
mean(midd_survey$find_partner_midd)
```

```
## [1] 0.05076142
```

The mean of this binary variable is the proportion of respondents who think they are most likely to find a marriage or long-term partner at Middlebury.

Assignment 8 Review

How does this vary by gender?

```r
prop.table(table(midd_survey$gender, midd_survey$find_partner_midd),1)
```

```
##
##                   0          1
##   Man   0.92746114 0.07253886
##   Other 1.00000000 0.00000000
##   Woman 0.96245734 0.03754266
```

No respondents in the "Other" gender category think they will find their partner at Middlebury. Let's take them out of the dataset so the rest of the models are easier to interpret.

```r
midd_survey <- midd_survey |>
     filter(gender!="Other") |>
     droplevels()
```

Assignment 8 Review

2. Regress the binary variable you created in #1 on gender, and interpret the coefficients.

```
model1 <-
      lm(find_partner_midd ~ gender,
      data = midd_survey)

summary(model1)
```

Assignment 8 Review

```
## 
## Call:
## lm(formula = find_partner_midd ~ gender, data = midd_survey)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.07254 -0.07254 -0.03754 -0.03754  0.96246 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.07254    0.01122   6.465 1.61e-10 ***
## genderWoman -0.03500    0.01445  -2.422   0.0156 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2205 on 970 degrees of freedom
## Multiple R-squared:  0.006009,	Adjusted R-squared:  0.004985 
## F-statistic: 5.864 on 1 and 970 DF,  p-value: 0.01563
```

Assignment 8 Review

On average, women are 3.5 percentage points less likely than men to think they will find a partner at Middlebury. This difference is significant.

Assignment 8 Review

3. Add type of housing as a control variable to the model, and interpret the coefficients.

```
model2 <-
      lm(find_partner_midd ~ gender + housing,
      data = midd_survey)

summary(model2)
```

Assignment 8 Review

```
##
## Call:
## lm(formula = find_partner_midd ~ gender + housing, data = midd_survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07733 -0.07219 -0.03719 -0.03719  0.96281
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.072194   0.011711   6.165 1.03e-09 ***
## genderWoman        -0.035008   0.014488  -2.416   0.0159 *
## housingHouse        0.005140   0.019964   0.257   0.7969
## housingOff Campus -0.005831   0.027595  -0.211   0.8327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2207 on 968 degrees of freedom
## Multiple R-squared:  0.006138,    Adjusted R-squared:  0.003058
## F-statistic: 1.993 on 3 and 968 DF,  p-value: 0.1134
```

Assignment 8 Review

Controlling for housing type, there is still a significant difference of 3.5 percentage points between men and women on average.

Or, controlling for gender, students living in dorms, in houses and off campus are equally likely to expect to find a partner at Midd, on average.

Assignment 8 Review

4. Add an interaction between gender and type of housing to the model, and interpret the coefficients.

```
model3 <-
      lm(find_partner_midd ~ gender * housing,
      data = midd_survey)

summary(model3)
```

Assignment 8 Review

```
##
## Call:
## lm(formula = find_partner_midd ~ gender * housing, data = midd_survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13462 -0.05788 -0.04719 -0.04719  0.98936
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.05788    0.01246   4.646 3.85e-06 ***
## genderWoman                   -0.01069    0.01624  -0.658  0.51059
## housingHouse                   0.07674    0.03291   2.331  0.01993 *
## housingOff Campus              0.07256    0.04747   1.528  0.12674
## genderWoman:housingHouse      -0.11329    0.04129  -2.743  0.00619 **
## genderWoman:housingOff Campus -0.11975    0.05821  -2.057  0.03995 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2197 on 966 degrees of freedom
## Multiple R-squared:  0.01697,    Adjusted R-squared:  0.01189
```

Assignment 8 Review

The differences between men and women vary across types of housing. The average differences between men and women are significantly larger for those who live in houses rather than dorms and for those who live off campus rather than in dorms.

Assignment 8 Review

5. Save the predicted values from the model in #4. Create a table with `group_by()` and `summarize()` showing how the predicted values differ by gender and housing type. Interpret any interesting results.

```r
midd_survey$pred_find_partner <- model3$fitted.values

find_partner_predictions <- midd_survey |>
     group_by(gender, housing) |>
     summarize(prob_find_partner =
      round(mean(pred_find_partner, na.rm=TRUE),3))
```

Assignment 8 Review

Gender	Housing	Proportion
Man	Dorm	0.058
Man	House	0.135
Man	Off Campus	0.130
Woman	Dorm	0.047
Woman	House	0.011
Woman	Off Campus	0.000

What are the interactions doing?

Consider the difference in the proportions between men in dorms (.058) and women in dorms (.047). That drop of .011 from men to women reflects the coefficient for women (-.011).

If there were no significant interaction terms, the drop of -.011 between men and women in dorms is what we would also expect as the drop between men and women in houses and between men and women in off campus housing. But the predicted differences in those housing types are much greater. That's because the interaction term tells us to add a certain amount to the predicted difference between men and women in those housing types.

For houses, start with .011 and add .113 (the coefficient for the womanXhouse interaction). That explains the .124 difference between men in houses (.135) and women in houses (.011).

For off campus, start with .011 and add .120 (the coefficient for the womanXoff campus interaction). That explains the .131 difference between men off campus (.130) and women off campus (.000)...with some minor rounding errors.

The results above are when we think of gender differences in the same type of housing. We could also think of housing differences within the same gender categories...

What are the interactions doing?

Consider the differences between men in dorms (.058), men in houses (.135), and men off campus (.13). Those differences simply represent the intercept (.058), the intercept plus the coefficient for houses (.058 + .077 = .135), and the intercept plus the coefficient for off campus (.058 + .073 = .131 rounded).

For women, the predictions need the interaction terms. In class, we said that among women, those in houses are 11 percentage points less likely than those in dorms to say they will find their partner at Middlebury. That's based on the interaction term alone. The actual difference across housing types will be the interaction term plus the associated coefficient for housing type.

So women in dorms have a .047 probability of saying they expect to find their partner at Middlebury. That is .036 percentage points higher than women in houses. That difference is explained by adding the coefficient for house (.077) and the interaction term for women in houses (-.113). More completely, .077 - .113 = .036.

Women in dorms have a predicted probability that is .047 higher than women off campus. That difference is explained by adding the coefficient for off campus (.073) and the interaction term for women off campus (-.120). More completely, .073 - .120 = .047.

Remember...

Interpreting the coefficients of a model with significant interaction terms is tricky. Predict from that model and then try to make sense of the predictions themselves.

Getting Output Out Of R

Final project's report should not have R output. Will discuss formatting more on Wednesday

- Figures should have titles
- Use the `huxtable` package for tables
- Let's turn to the `huxtable.Rmd` notebook...

Previewing Huxtable

	(1)	(2)	(3)
(Intercept)	0.073 ***	0.072 ***	0.058 ***
	(0.011)	(0.012)	(0.012)
Gender = Woman	-0.035 *	-0.035 *	-0.011
	(0.014)	(0.014)	(0.016)
Housing = House		0.005	0.077 *
		(0.020)	(0.033)
Housing = Off Campus		-0.006	0.073
		(0.028)	(0.047)
Woman x House			-0.113 **
			(0.041)
Woman x Off Campus			-0.120 *
			(0.058)
N. obs.	972	972	972

*** p < 0.001; ** p < 0.01; * p < 0.05.

Meier and Musick (2014)

Motivation?

Research Question?

Hypotheses?

Data?

- Dependent Variables?
- Independent Variable?
- Control Variables?

Meier and Musick (2014)

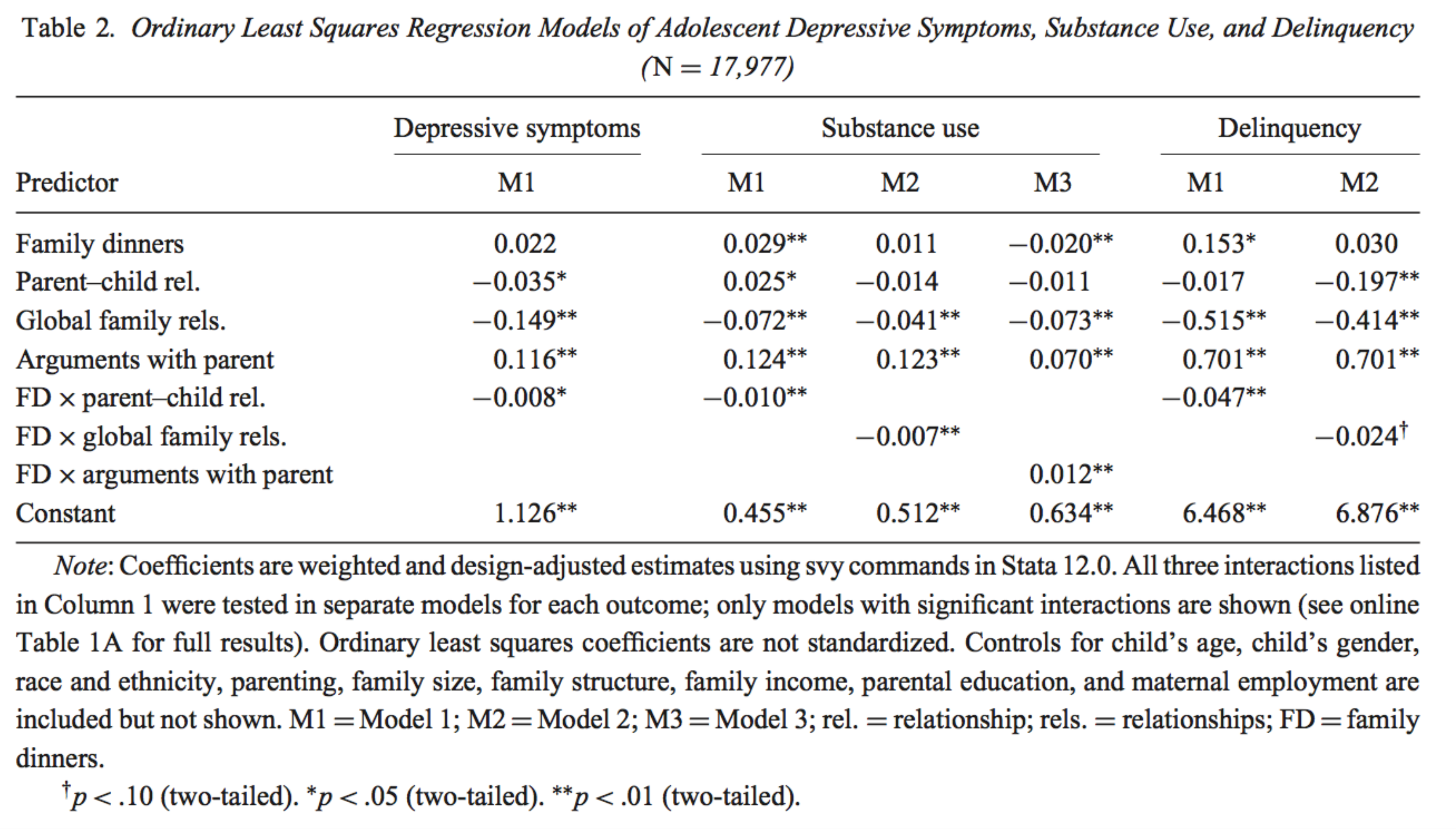

Table 2. *Ordinary Least Squares Regression Models of Adolescent Depressive Symptoms, Substance Use, and Delinquency (*N = *17,977)*

	Depressive symptoms	Substance use			Delinquency	
Predictor	M1	M1	M2	M3	M1	M2
Family dinners	0.022	0.029**	0.011	−0.020**	0.153*	0.030
Parent–child rel.	−0.035*	0.025*	−0.014	−0.011	−0.017	−0.197**
Global family rels.	−0.149**	−0.072**	−0.041**	−0.073**	−0.515**	−0.414**
Arguments with parent	0.116**	0.124**	0.123**	0.070**	0.701**	0.701**
FD × parent–child rel.	−0.008*	−0.010**			−0.047**	
FD × global family rels.			−0.007**			−0.024†
FD × arguments with parent				0.012**		
Constant	1.126**	0.455**	0.512**	0.634**	6.468**	6.876**

Note: Coefficients are weighted and design-adjusted estimates using svy commands in Stata 12.0. All three interactions listed in Column 1 were tested in separate models for each outcome; only models with significant interactions are shown (see online Table 1A for full results). Ordinary least squares coefficients are not standardized. Controls for child's age, child's gender, race and ethnicity, parenting, family size, family structure, family income, parental education, and maternal employment are included but not shown. M1 = Model 1; M2 = Model 2; M3 = Model 3; rel. = relationship; rels. = relationships; FD = family dinners.

$^{\dagger}p < .10$ (two-tailed). $^{*}p < .05$ (two-tailed). $^{**}p < .01$ (two-tailed).

Meier and Musick (2014)

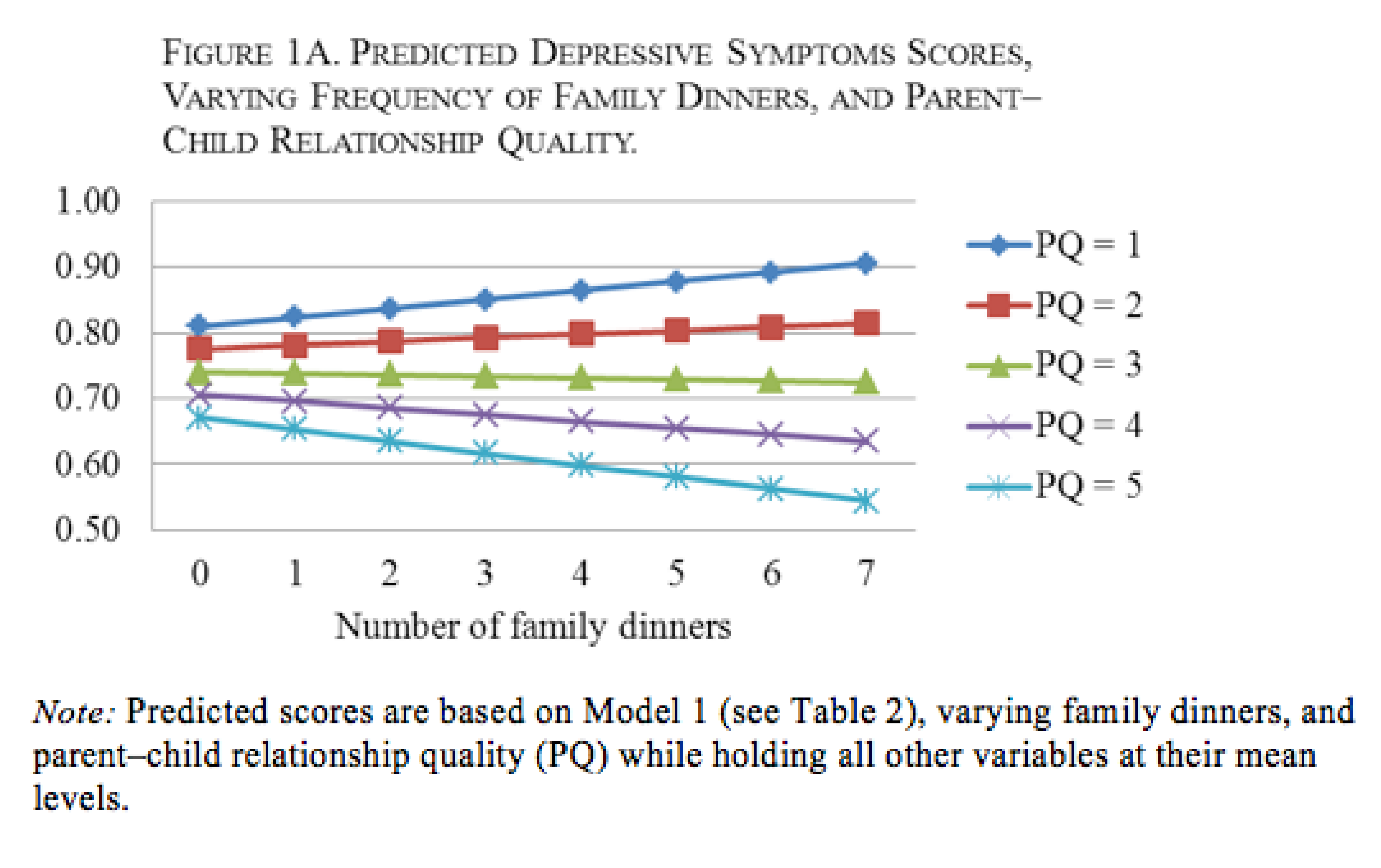

FIGURE 1A. PREDICTED DEPRESSIVE SYMPTOMS SCORES, VARYING FREQUENCY OF FAMILY DINNERS, AND PARENT–CHILD RELATIONSHIP QUALITY.

Note: Predicted scores are based on Model 1 (see Table 2), varying family dinners, and parent–child relationship quality (PQ) while holding all other variables at their mean levels.

Meier and Musick (2014)

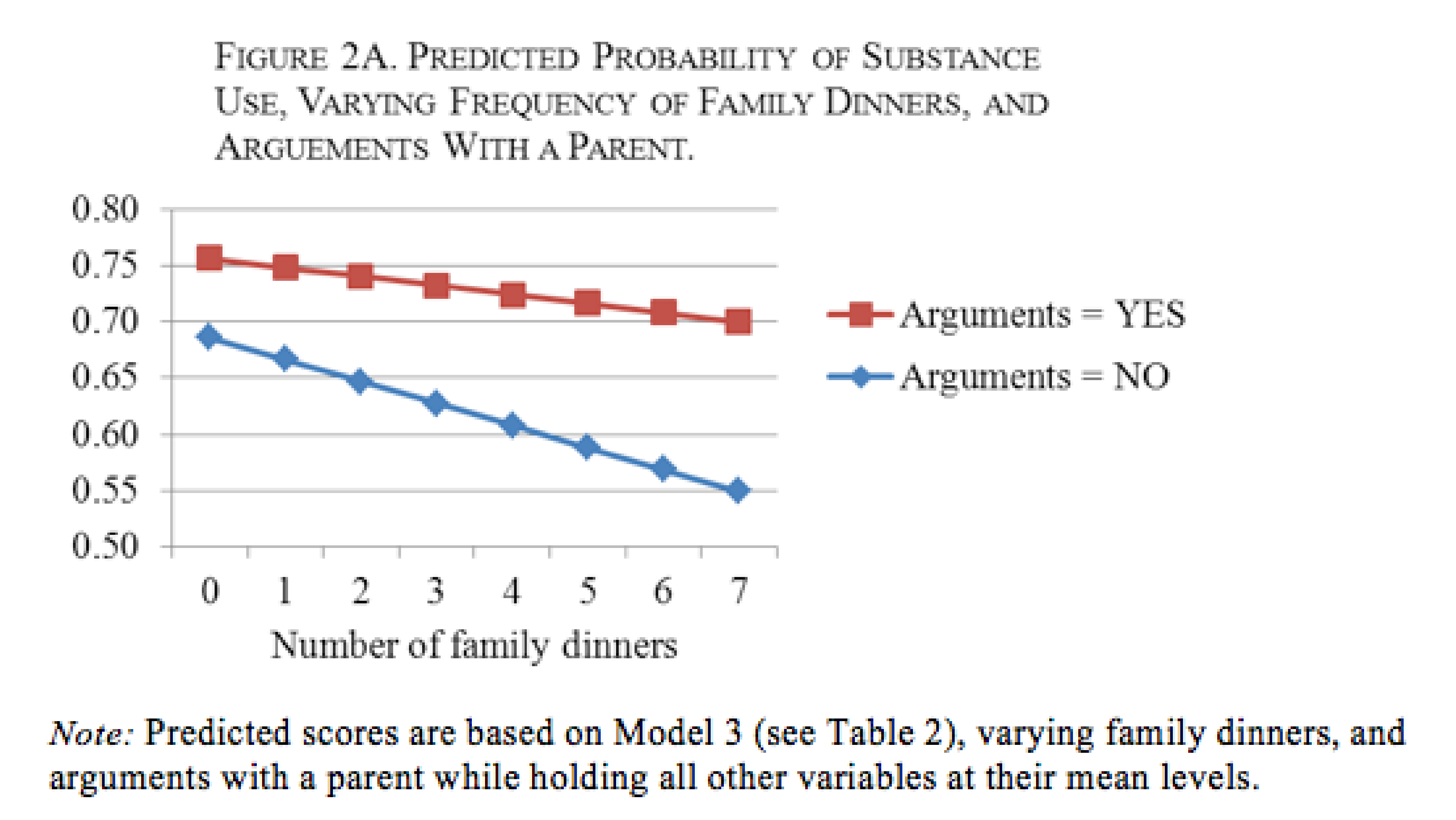

FIGURE 2A. PREDICTED PROBABILITY OF SUBSTANCE USE, VARYING FREQUENCY OF FAMILY DINNERS, AND ARGUEMENTS WITH A PARENT.

Note: Predicted scores are based on Model 3 (see Table 2), varying family dinners, and arguments with a parent while holding all other variables at their mean levels.

Meier and Musick (2014)

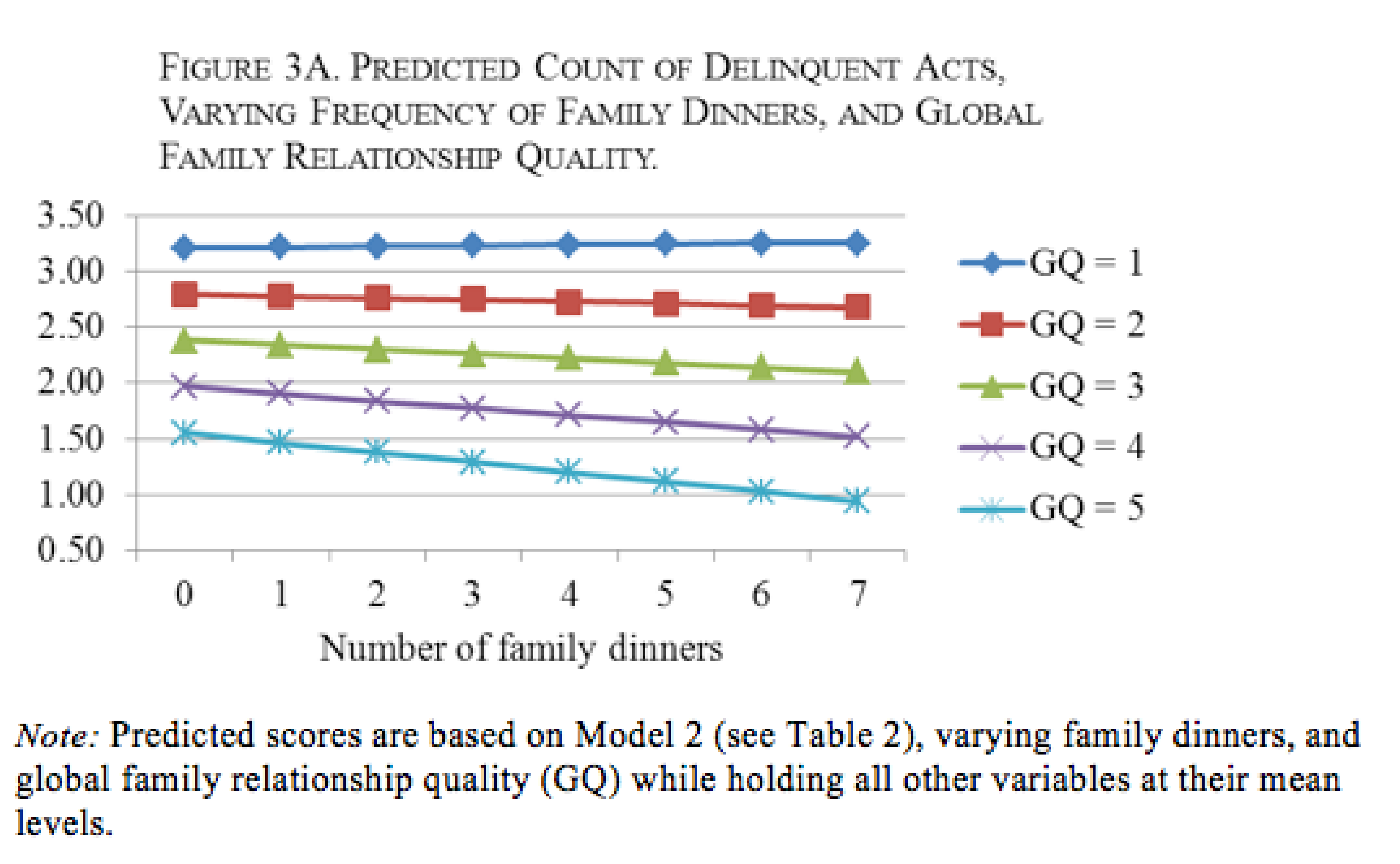

FIGURE 3A. PREDICTED COUNT OF DELINQUENT ACTS, VARYING FREQUENCY OF FAMILY DINNERS, AND GLOBAL FAMILY RELATIONSHIP QUALITY.

Note: Predicted scores are based on Model 2 (see Table 2), varying family dinners, and global family relationship quality (GQ) while holding all other variables at their mean levels.