1

Preliminary Analyses for Parent Communication Example

In this chapter, I describe the preliminary analyses I conducted for the Chapter 12 communication example of an RET with a binary outcome. After providing a refresher of the numerical example, I consider the analysis of response distributions, evaluation of imbalance during random assignment, non-linearity, outlier/leverage analyses, and evaluation of heteroscedasticity. In general, I routinely conduct analyses on the psychometric properties of my measures but I do not cover that topic here.

NUMERICAL EXAMPLE

The example in Chapter 12 has a binary outcome with continuous mediators. The data are available on the resources tab of my website. The example focuses on parental communication with young adolescent, middle school children about reasons not to have sex at this time in their lives. About 60% of parents of middle school youth in the United States have never talked with their child about sex. Research suggests that parents often are reluctant to do so because they do not feel they have enough knowledge about sex and birth control to adequately discuss the topic. Parents also tend to feel that such discussions will be embarrassing for both them and their child. My example focuses on a program aimed at parents to encourage them to discuss issues surrounding not engaging in sex at this time in their lives by addressing three factors, (1) educating parents about the advantages of engaging in such conversations, (2) providing parents with the knowledge they feel they need to have effective conversations, and (3) teaching parents strategies to reduce embarrassment. The target mediators were measured on multi-item inventories in which each item was rated on 7 point disagree-agree scales: -3 = strongly disagree, -2 = moderately disagree, -1 = slightly disagree, 0 = neither agree nor disagree, 1 = slightly agree, 2 = moderately agree, 3 = strongly agree. Scores were averaged across items; higher scores indicated (1) higher levels of perceived advantages of engaging in the conversations, (2) higher levels of perceived knowledge, and (3) beliefs that conversations about sex and pregnancy would be embarrassing.

The outcome measure was whether the parent engaged in a meaningful conversation about sex and pregnancy with his or her child in the ensuing 9 months after program participation. This was assessed by self-reports from the adolescent child of the parent at a follow-up interview. The outcome was scored 0 = parent did not engage in a conversation, 1 = parent engaged in a conversation. Each of the mediators was measured at baseline and again at program completion. The control group received exposure to materials on an

unrelated topic. The covariates measured at baseline were the biological sex of the adolescent (0 = male, 1 = female), and the overall quality of parent-adolescent communication. The latter used a multi-item scale with each item measured on a -3 to +3 disagree-agree metric, averaged across items. Higher scores indicate higher quality communication. In a real evaluation, there would be a longer list of covariates, but I use only two to keep the example manageable. The total sample size was 1,500.

The RET model, absent covariates, is in Figure 1. As noted in the main text, the binary outcome has a disturbance term in the diagram, but often it is omitted. I discuss in the main text the reasons for and against including it.

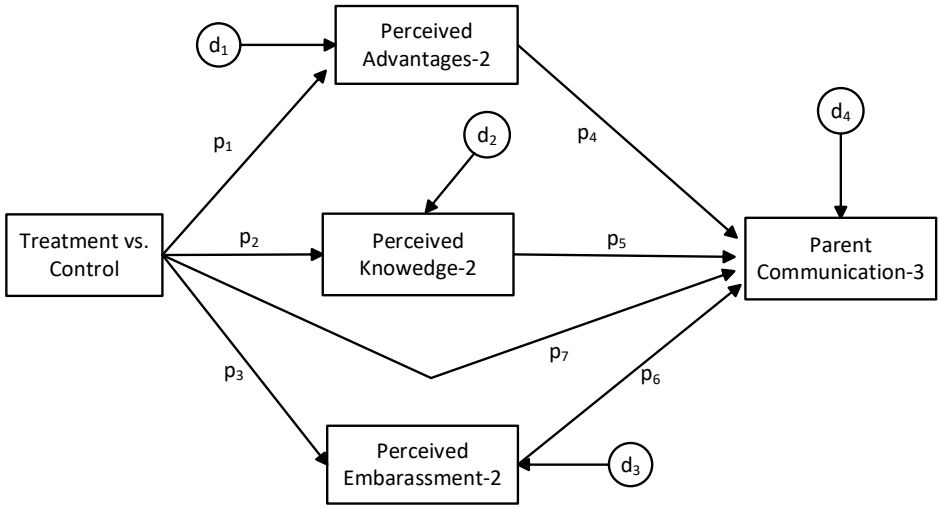


FIGURE 1. Parent communication example

The equations for the model are (note: TREAT = the treatment condition, BS is biological sex, CQ is quality of communication at baseline; PA is perceived advantages; PK is perceived knowledge, PE is perceived embarrassment, and COM is parent communication; each followed by the number 1, 2 or 3 to indicate time of assessment):

$$PA2 = a_1 + p_1\ TREAT + b_1\ BS1 + b_2\ CQ1 + b_3\ PA1 + d_1 \quad [1]$$

$$PK2 = a_2 + p_2\ TREAT + b_4\ BS1 + b_5\ CQ1 + b_6\ PK1 + d_2 \quad [2]$$

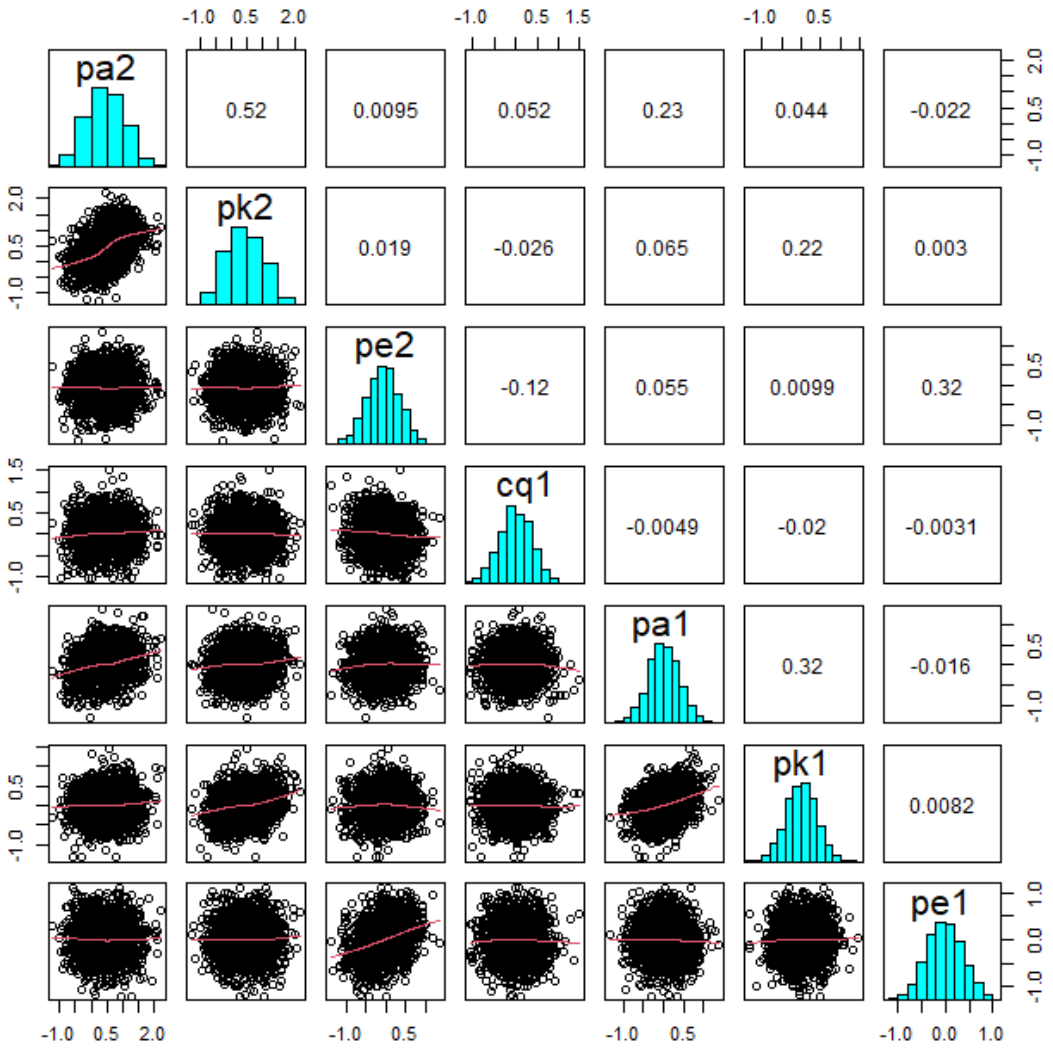
$$PE2 = a_3 + p_3\ TREAT + b_7\ BS1 + b_8\ CQ1 + b_9\ PE1 + d_3 \quad [3]$$

$$COM3 = a_4 + p_4\ PA2 + p_5\ PK2 + p_6\ PE2 + p_7\ TREAT + b_{10}\ BS1 + b_{11}\ CQ1 \quad [4]$$

Response Distributions

I first examine the frequency distribution of the binary variables to determine if there are base rate issues with them. The percentage of parents with a score of 1 on the communication outcome was 46.8% and for a score of 0 it was 53.2%. For biological sex the percent of males was 47.5% and for females it was 52.5%. Finally 49.4% of the respondents were in the intervention condition and 50.6% were in the control condition. What matters most is the absolute frequency of cases in each category, with frequencies less than 30 or so raising red flags. With an N =1500, this is not an issue.

I next examined a scatterplot matrix (using the program provided on my website) for all the continuous predictors. This matrix has histograms in the diagonal, scatterplots with smoothers in the lower triangle, and correlations in the upper triangle:



There is no distinctive non-linearity in the smooths (red lines) in the scatterplots and all of the distributions are relatively symmetric. There are no red flags here.

Imbalance

I next evaluate if there was any notable imbalance in the baseline variables during randomization as a function of the treatment condition. I calculated the mean baseline value for each measured baseline variable using SPSS and found the following (note: biological sex was scored 0 = male, 1 = female):

Baseline Variable	Intervention	Control	Eta Squared
Commun Qual	0.017	0.008	<0.01
Advantages	0.027	0.012	<0.01
Knowledge	0.020	-.003	<0.01
Embarrassment	-.002	-0.014	<0.01
Biological sex	0.520	0.530	-

What matters most is not the statistical significance of the effects (given random assignment) but rather the magnitude of the effects. Everything looks fine in this regard. The summary statistics are similar in each condition.

Viability of MLPM

I next explore if the underlying data relative to Equation 4 is compatible with a modified linear probability model (MLPM). I do so using a limited information SEM approach in conjunction with OLS regression. I first regress `COM3` onto its predictors using OLS regression in SPSS to determine if any predicted outcome probabilities are outside the values 0 to 1.00 by saving the predicted probabilities to the data file for each case, an option offered by SPSS. If there are such offending predicted scores, I might consider using the sequential least squares strategy (SLS) of Horrace and Oxaca (2003, 2006), per Chapter 5, in place of the MLPM. Uanhoro et al. (2019) found that both the MLPM and SLS strategies work well when the percent of offending scores is about 10% or less of the sample size given a true linear function between probabilities and continuous predictors. When the percent is closer to 20%, the SLS method works well, but not the MLPM. In the analyses I conducted, none of the predicted probabilities were offensive; they ranged from 0.16 to 0.81. So predicted values outside the 0 and 1 boundaries is a non-issue.

Next, I evaluated if the presumed linear function between the outcome probabilities and the predictors reasonably approximates the data. I target each quantitative predictor in

Equation 4 one at a time and conduct a series of polynomial regressions (using the full equation) to evaluate if there are statistically significant departures from linearity based on polynomials to the fifth power, sequentially. I used the program on my website called *polynomials* and regressed the binary outcome onto the relevant polynomials. There were no significant higher order terms using either traditional regression or robust regression with HC3 standard errors. If there was, this would not necessarily rule out the MLPM; it would just mean I need to modify it to accommodate the non-linearity. If the relationships are fundamentally linear, my task is simplified. The results suggested linearity.

I next examined partial component plus residual plots for each quantitative predictor in Equation 4 using the *partial residual plots* program on my website. I specified an OLS model, which serves as a stand-in for the MLPM. The program implements a type partial residual plot known as a **component plus residual plot**. These plots illustrate the relationship between a target predictor and the outcome after controlling for the other predictors in the modeled equation. Figure 1 presents an example plot for parents' perceived advantages of talking with their children about sex (PA2). The X axis in Figure 1 represents scores for the target predictor. The Y axis is the regression coefficient for the target predictor times the person's PA2 score and is called the component; it adds to this the person's residual score from the full analysis. This yields the **component plus residual value** for an individual.

The residuals in the component plus residual value contain within them the influence of all other independent factors that influence the outcome other than the linear predictors in equation. This includes any operative non-linearities from the target predictor PA2, which are ignored in the primary regression analysis because of its focus on linearity. To these residuals, we add back the (covariate adjusted) linear contribution of PA2 using the component portion of the term. This yields the component plus residual value for each individual, which is a mix of the linear and non-linear influence of PA2 on the outcome.

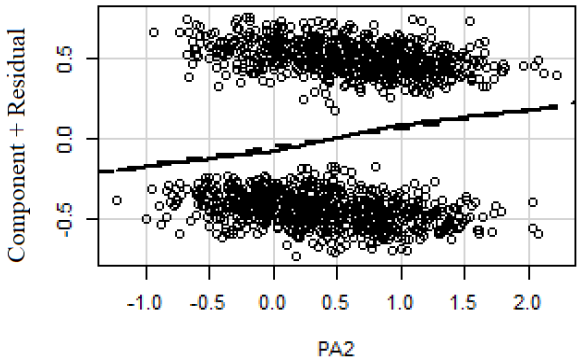


FIGURE 1. Partial residual plot for perceived advantages of communication

The plot In Figure 1 shows the best fitting line between the component plus residual values and PA2. It is the dashed line. The figure also plots a solid line smoother for the data, which captures both the linear and nonlinear influence of PA2 combined. If the smoother is functionally linear and overlays the dashed line, then this implies PA2 is linearly related to the outcome. If the lines diverge substantially, this implies non-linearity. For PA2, the lines reasonably overlap. I conducted these analyses for each quantitative predictor and found linearity to be reasonable in each case.

A few asides about residual plots for binary outcomes are worth noting. In traditional regression, the residual is the difference between peoples' observed Y scores and their predicted Y scores. For binary regression models, the observed Y scores are either 0 or 1, but the predicted scores are either predicted probabilities (per MLPM regression), predicted logits (logistic regression) or predicted probits (probit regression), all of which vary in value across the many different predictor profiles. When the predicted values are subtracted from observed scores, a common result is for two clumps of scores to appear, with the upper clump being individuals with observed scores of 1 and the lower clump those with observed scores of 0. These clumps are evident in Figure 1 for the MLPM.

As a final check, I use the running interval smoother program on my website to plot a smoother between the probability of communication and PA2 (see Figure 2). The plot does not control for the other predictors in the equation, which is a weakness. The smoother is essentially linear. The MLPM seems viable.

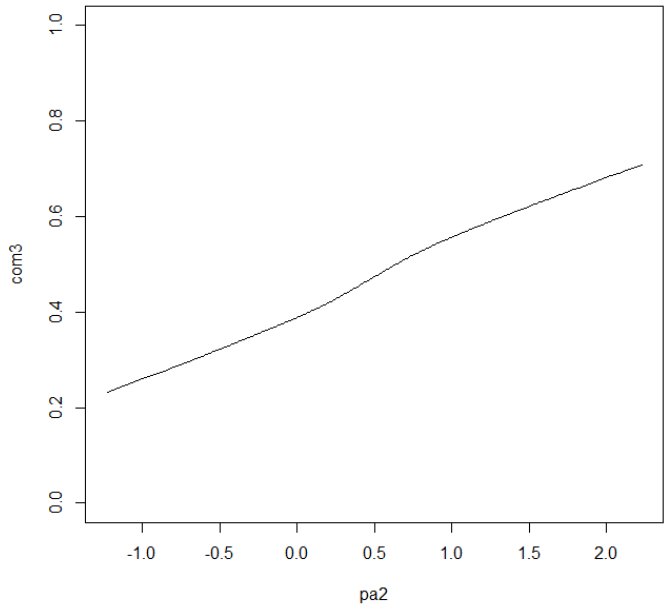


FIGURE 2. Smoother for binary outcome

Viability of Logit/Probit Approach

I also can apply the above methods to evaluate the potential applicability of logit and probit models. The question of offending probability estimates is moot because predicted probabilities outside the 0 to 1 range cannot occur with these methods. Both logit and probit models are non-linear for outcome probabilities but they assume linearity for logits and probits. I first test for linearity in probits by evaluating if there are statistically significant departures from linearity when I add polynomials to the fifth power for each continuous predictor, one predictor at a time, using probit regression as applied to Equation 4. I again used the program *polynomials* on my website. When I conducted these analyses, none of the higher order terms were statistically significant. This also was true for logistic regression.

Partial residual plots also can be used to evaluate linearity of each predictor with the logits or probits from Equation 4 using the partial residual program on my website. The plots are identical to those for the MLPM but the best fitting line is for probits or logits, hence it should be linear (the program offers an analysis for probits and logits). Figure 3 presents the probit plot for PA2. There is good correspondence between the smoother and the best fitting line, as was true for the other quantitative predictors.

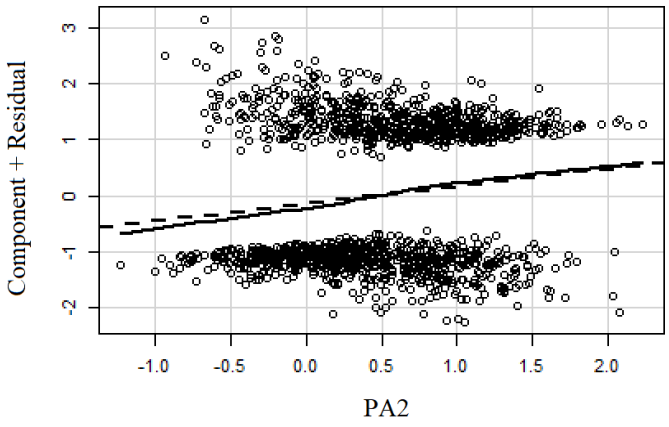


FIGURE 3. Partial residual plot for probit regression

It might seem contradictory that the partial plots can be consistent with both a MLPM and a probit model because one is inherently linear and the other is non-linear. However, this can occur if the data cover only a portion of the probit probability curve, with a linear

function operating within that portion. Such a case is shown in Figure 4 where I segment the logit and probit probability curves into three parts. As long as the range of probabilities in a study is mostly within one of these segments, the relationship between the predictor and the probability of Y is functionally linear and either a MLPM, a logit, or a probit model could be used. This was the case for the present data where the probabilities were between 0.15 and 0.80, falling primarily in Segment B. Functionally linear trends can occur in other parts of the curves besides these three segments, as illustrated by the segment in Figure 5, which slightly overlaps segment A and B. Model applicability depends not only on the abstract function (linear, logit, probit) relating Y to a predictor, but also on the particular probabilities spanned by data.

A related issue is whether to choose logit or probit regression should I decide to embrace one of these models. As noted, the underlying functions for logit and probit are technically different but they are similar enough that many statisticians view them as interchangeable in practice. Chen and Tsurumi (2010) compared five preliminary tests to help choose between logit versus probit regression and found that none of them performed well when the outcome event rate was near 0.50. When event rates were reasonably discrepant from 0.50, sample sizes well over 1,000 were necessary to discern the models and even then, test performance was suboptimal. The bottom line is that choosing between the two functions based purely on empirics can be challenging.

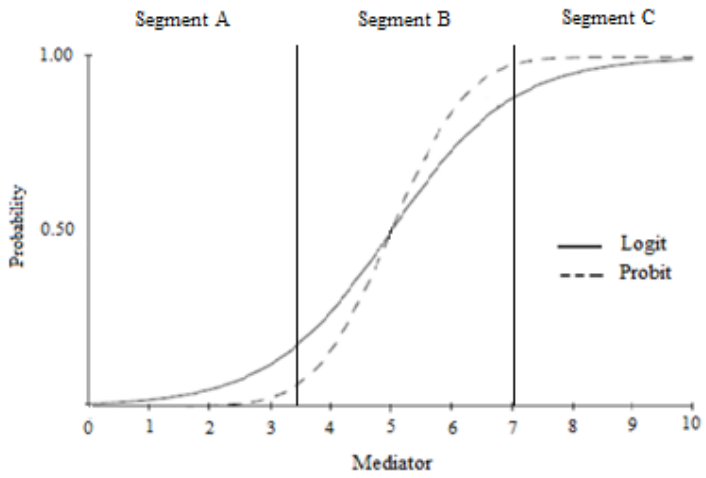


FIGURE 4. Linear segments of logit and probit probability curves

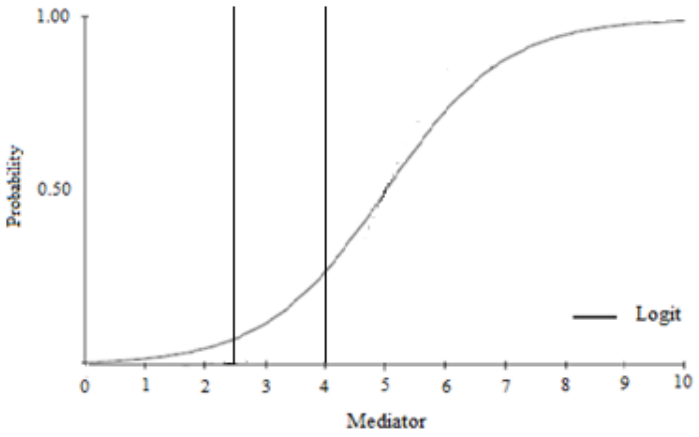


FIGURE 5. Linear segment for a logit probability curve

Because the disturbance term for the latent response probit model is normally distributed as opposed to logistically distributed, probit models often can be integrated into broader statistical theory more easily than logit models, especially SEM. For example, in the Mplus software, the traditional global fit indices are available for models that use probit regression but this is not the case for logit regression. As well, it is more straightforward to model correlated disturbances for binary endogenous variables using probit as opposed to logistic modeling with Mplus. Given this, I often use probit approaches when analyzing RETs if both logit and probit models seem viable.

Parenthetically, when I correlated individuals' predicted probabilities for Equation 4 as derived from a probit and then from a MLPM model, the squared r when predicting the probit-based probabilities from the MLPM probabilities was 0.999 with an intercept of -0.00075 ±0.001 and a slope of 1.000 ±0.002. I found similar results when I predicted the probit probabilities from the logit probabilities. Also, the partial residual program on my website applies the le Cessie–van Houwelingen–Copas–Hosmer test (Hosmer et al., 1997) for the appropriateness of a logit model for the data. Because the logit and probit curves are so similar, one can test for the fit of the logit model using this test and if the result indicates a logit model is inappropriate, in all likelihood, the probit model is as well. The logit model is deemed not viable if the z value for it is less than 0.05. In the present case, $z = 1.05$, $p < 0.30$.[1]

Based on the above, I conclude I can reasonably apply the MLPM, logit modeling, or probit modeling (and, by implication, Bayesian SEM) to the numerical example.

[1] I do not recommend the classic Hosmer-Lemeshow test for logit regression. See Allison (2013).

Outliers and Leverages

Another preliminary test I perform is to check for outliers/leverages for each equation in the model. When doing so, it is important to use approaches that address multivariate masking and that do not rely on outlier influenced referent statistics (e.g., means, standard deviations, correlations). I cannot use the robust strategy by Rousseuw and van Zomeren from Chapter 5 for Equation 4 because it requires a many-valued quantitative outcome. One strategy is to focus just on the predictor space for Equation 4 and identify unusual cases (high leverages) using a robust multivariate detection method. I used the robust outlier projection method described in Wilcox (2017) made available on the programs tab of my webpage and identified 19 high leverage cases. When I eliminated them from the analysis, the conclusions for Equation 4 in the analyses I report in the main text replicated. Disruptive leverages were not problematic. Given this, I decided to analyze the full data set.

Heteroscedasticity

As described in Chapter 5, both probit and logit regression can be expressed using a latent response framework. Doing so makes evident an important assumption of these methods of analysis (as well as ordinal regression, which relies on them). The formula for the latent response model is

$$y^* = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_k X_k + \varepsilon$$

where y* is the continuous latent variable underlying the dichotomous observed outcome measure, the X are presumed determinants of that outcome, γ_0 is the intercept (which is traiditionally fixed at zero), γ are the regression coefficients, and ε is a disturbance term that is assumed to be normally distributed in the case of probit regression and to follow a standardized logistic distribution for logistic regression. See Chapter 5 for details and explication of the model. The assumption of variance homogeneity refers to the variance of the disturbance term for the different predictor profiles, much like traditional regression, but with some notable exceptions described in Chapter 5. In OLS regression, heteroscedasticity can bias standard errors but it usually does not impact the consistency or biasedness of the coefficients per se. By contrast, heteroscedasticity can be more damaging for probit and logistic regression because it affects both standard errors and the predictor coefficients.

Different strategies have been proposed for evaluating heteroscedasticity in logit and probit models. They vary widely in their ability to detect heteroscedasticity. I like an approach known as **location-scale modeling** or **heterogenous choice modeling**. The test is not available in Mplus, but I provide a program for it on my website that can be used in

a limited information estimation context. The approach is described in depth in Keele and Park (2006), Tutz (2020), and Williams (2009,2010).

To implement location-scale modeling, we need to specify two models, a "location" model and a "scale" model. The "location" model is the model we would apply if we were conducting a traditional logistic regression analysis. In the communication example from the main text of Chapter 12, an equation that uses probit regression regresses the binary outcome parental communication onto three mediators, a treatment condition, and two covariates:

$$\text{Probit}(\text{COM3}) = a_4 + p_4\,\text{PA2} + p_5\,\text{PK2} + p_6\,\text{PE2} + p_7\,\text{T} + b_{10}\,\text{BS1} + b_{11}\,\text{CQ1}$$

and the results for the estimated coefficients of p_4 through p_7 and b_{10} and b_{11} for the probit regression in a traditional limited information estimation analysis (reported by my program) are:

```
Coefficients:
       Estimate Std. Error z value Pr(>|z|)
pa2    0.313058   0.082701   3.785 0.000153
pk2    0.298181   0.080670   3.696 0.000219
pe2   -0.375333   0.081712  -4.593 4.36e-06
treat -0.001354   0.112985  -0.012 0.990440
cq1    0.149574   0.087474   1.710 0.087279
bs1    0.242432   0.067341   3.600 0.000318
```

These results are very close to but not identical to the Mplus results reported in Chapter 12 because the Mplus analysis uses FISEM whereas we are working here with just the single equation in an LISEM sense.

The "scale" model, in contrast to the location model, specifies predictors of differences in residual variability. For example, if I expect the residual variability to be different for females as compared to males, I would use female as a predictor in the scale model (i.e., I would include `bs1`). Note that the predictors in the scale model do not have to be the same as those in the location model but often they are. Location-scale regression makes adjustments to the various estimates in the location model based on the results in the scale model based on the detected heteroscedasticity. Thus, in principle, location-scale analysis both detects and corrects for heteroscedasticity.

An important issue when using location-scale regression is the metric of the predictors of residual variability, that is the predictors in the scale portion of the model. It turns out that the coefficients in the location part of the model are conditional coefficients; they are the estimated logistic coefficients conditioned on when all predictors in the scale portion of the model equal zero. If scores of zero on the scale predictors are not meaningful,

then the coefficients for the location portion of the model are not either. For this reason, many methodologists mean center all predictors in the scale portion of the model (and, for that matter, they mean center the predictors throughout the model more generally, including in the location model) to produce meaningful zero points. This transformation does not affect the results of the scale model but, as noted, it does affect the results of the location model. Thus, having meaningful zeros does not affect the ability to detect heteroscedasticity but it does affect the ability to adjust for it. The program on my website offers an option for mean centering predictors and I used that option here, accordingly.

Because technically the outcome of the scale model are variances, statisticians recommend analyzing their logs rather than the variances per se. Here are the results for the scale portion of the model from my program:

```
log-scale coefficients:
      Estimate Std. Error z value Pr(>|z|)
pa2   -0.23210    0.20354  -1.140   0.2541
pk2    0.04271    0.23748   0.180   0.8573
pe2    0.46667    0.23410   1.993   0.0462 *
treat  0.44543    0.37668   1.183   0.2370
cq1   -0.12162    0.22291  -0.546   0.5854
bs1    0.18372    0.20597   0.892   0.3724

Exponent of scale coefficients
      pa2       pk2       pe2     treat       cq1       bs1
0.7928666 1.0436316 1.5946675 1.5611576 0.8854882 1.2016830
```

If a predictor is statistically significant, then this suggests there is heteroscedasticity in the model caused by that predictor. In the above analysis, only one predictor was statistically significant ($p < 0.462$) and it was marginally so. The exponents of the coefficients provide a sense of the magnitude of the effects. For example, for biological sex (`bs1`), a dummy variable, the residual variance for adolescent females (scored 1 on the dummy variable) was about 1.20 times larger than the residual variance for adolescent males (scored zero on the dummy variable), holding constant the other predictors in the scale model. For `pk2`, for every one unit that `pk2` increases, the residual variance is predicted to increase by a multiplicative factor of 1.59.

Here are the results from the location portion of the model that adjusts for for heteroscedasticity based on the scale model:

```
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
pa2    0.33683    0.07904   4.261 2.03e-05
pk2    0.28232    0.08172   3.455 0.000551
```

```
pe2   -0.43300    0.08064  -5.370 7.88e-08
treat -0.04275    0.10914  -0.392 0.695321
cq1    0.17931    0.08305   2.159 0.030840
bs1    0.17771    0.07137   2.490 0.012777
```

The results are fairly close to those that we observed for the unadjusted analyses reported earlier that ignored residual variance heterogeneity.

For the current data, I am probably safe ignoring the variance heterogeneity but I should make note of it in my reporting of findings. If I am uncomfortable with ignoring it, I can always report the adjusted coefficients, but there are reasons to be cautious about relying on them too heavily. In a series of simulation studies, Keele and Park (2006) found that location-scale regression is rather sensitive to specification error in either the location model or the scale model. Importantly misspecification in the location model can affect the results for the scale model and vice versa. Keele and Park (2006) conclude that "if researchers are only interested in the parameters from the choice model, but suspect heteroscedasticity, these models may not be the best alternative" and that it may be "better to estimate a standard probit and ignore the heteroscedasticity than poorly specify a heteroskedastic model." At present, we simply do not know how much differing degrees of heteroscedasticity matter. The bottom line is that there is no single best procedure for addressing the problem of heteroscedasticity in logit and probit regression, but it is something we should at least be sensitive to when we use these techniques.

REFERENCES

Chen, G. & Tsurumi, H. (2010) Probit and logit model selection. *Communications in Statistics - Theory and Methods*, 40, 159-175.

Horrace, W. C., & Oaxaca, R. L. (2003). New wine in old bottles: A sequential estimation technique for the LPM. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=383102

Horrace, W. C. & Oaxaca, R. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90, 321-327.

Keele, L. & Park, D. (2006). Difficult choices: An evaluation of heterogenous choice models. Downloaded from https://www3.nd.edu/~rwilliam/oglm/ljk-021706.pdf

Tutz, G. (2020). Modelling heterogeneity: on the problem of group comparisons with logistic regression and the potential of the heterogeneous choice model. *Advances in Data Analysis and Classification*, 14, 517–542.

Uanhoro, J., Wang, Y., & O'Connell, A. (2019) Problems with using odds ratios as effect sizes in binary logistic regression and alternative approaches. *Journal of Experimental Education*, DOI: Online at https://www.tandfonline.com/doi/full/10.1080/00220973.2019.1693328.

Wilcox, R. (2017). *Introduction to robust estimation and hypothesis testing*. Academic Press (Fourth edition).

Williams, R. (2009) Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods and Research*, 37, 531–559.

Williams, R. (2010). Fitting heterogeneous choice models with oglm. *Stata Journal*, 10, 540–567.