Social Science Research 109 (2023) 102802

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Comparing logit & probit coefficients between nested models

Richard Williams [*], Abigail Jorgensen

University of Notre Dame, 4060 Jenkins Nanovic Halls, Department of Sociology, University of Notre Dame, Notre Dame, IN, 46556, USA

ARTICLE INFO

Keywords:
Nested models
Y-standardization
Logit & probit
Karlson/ Holm/ Breen method
Marginal effects

ABSTRACT

Social scientists are often interested in seeing how the estimated effects of variables change once other variables are controlled for. For example, a simple analysis may reveal that income differs by race – but why does it differ? To answer such a question, a researcher might estimate a model where race is the only independent variable, and then add variables such as education to subsequent models. If the original estimated effect of race declines, this may be because race affects education, which in turn affects income. What is not universally realized is that the interpretation of such nested models can be problematic when logit or probit techniques are employed with binary dependent variables. Naïve comparisons of coefficients between models can indicate differences where none exist, hide differences that do exist, and even show differences in the opposite direction of what actually exists. We discuss why problems occur and illustrate their potential consequences. Proposed solutions, such as Linear Probability Models, Y-standardization, the Karlson/Holm/Breen method, and marginal effects, are explained and evaluated.

1. Introduction

Analyses using logistic regression and probit models have appeared in Social Science Research almost since the journal began (e.g. Henretta 1979; Magidson 1981; Berk and Ray 1982; Rauma 1984; Marini 1985; McDougall and Bunce, 1986). Often, these papers, as well as papers in many other journals (e.g. Regnerus and Elder, 2003; Matthew, 2011; Kreager and Haynie 2011; Crowder et al. 2011), have focused on the analyses of nested models, which begin with simple models and gradually expand them.

The use of nested model approaches in SSR and elsewhere is not surprising. Social scientists are often interested in seeing how the effects of variables differ between nested models. For example, a simple descriptive analysis may reveal that income differs by race – but why does it differ? To answer such a question, a researcher might estimate a model where race is the only independent variable, and then add variables such as education to subsequent models. If the original estimated effect of race declines, this may be because race affects education, which in turn affects income. That is, the effect of race is *mediated*[1] by education; or, one might say that education is the *mechanism* by which race affects income.[2] Such analyses can provide a clearer picture of exactly why and how race is

* Corresponding author.
E-mail addresses: rwilliam@nd.edu (R. Williams), Abigail.Jorgensen.6@nd.edu (A. Jorgensen).

[1] We use terms like mediation much the same way as Karlson et al. (2012) and Kohler, Karlson, and Holm do (2011) do. For alternative and broader viewpoints on mediation, see, for example, VanderWeele (2015).

[2] To put it another way, researchers are often interested in distinguishing between the total, direct, and indirect effects of a variable X on an outcome Y. The estimated total effect of X is the effect of X when no other variables are in the model. The direct effect of X is the effect of X after all the mediating variables (variables affected by X which in turn affect Y) are controlled for. The indirect effect of X is the difference between the total effect and the direct effect. As we discuss later, there can also be concomitant variables that are not affected by X (e.g. gender) but which also affect the dependent variable. If these exist, they should also be controlled for when estimating total, direct, and indirect effects of X.

https://doi.org/10.1016/j.ssresearch.2022.102802
Received 1 May 2022; Received in revised form 22 September 2022; Accepted 23 September 2022
Available online 3 November 2022

related to the outcome of interest.[3] Identifying the mechanisms by which race affects income may make it possible to develop policies that change how those mechanisms work, e.g. schools might be reformed so that race had less impact on educational outcomes.

In Ordinary Least Squares (OLS) regression with continuous dependent variables, such issues are often addressed by estimating and comparing sequences of nested models. Unfortunately, these same approaches can be highly problematic when categorical and limited dependent variables like binary and ordinal are analyzed. (We primarily discuss logistic regression, but the same problems and solutions are equally true for probit and some other analyses involving limited dependent variables.) Many do not realize that there are important statistical differences in how these procedures work compared to OLS (Mood, 2010).[4] Depending on the situation, apparent differences in coefficients across models can be greatly overstated or greatly understated in probit or logistic regression models. For binary and ordinal dependent variables, naïve comparisons of coefficients between models can indicate differences where none exist, hide differences that do exist, and even show differences in the opposite direction of what actually exists.

2. The problem explained

We begin with two examples, one using hypothetical data and the other using real data, to illustrate our points. All results presented here can be replicated with the materials presented in the Appendix.

2.1. Example using hypothetical data

We constructed a data set where the dependent variable, Y, and the independent variables X1 and X2, are all 0/1 dichotomies. Each of the four possible combinations of X1 and X2 (e.g. X1 = 0, X2 = 0; X1 = 1, X2 = 0; etc.) had 125 cases. Y was constructed in such a way that the true values of the X1 and X2 coefficients in a logistic regression were equal to 4.5. We then estimated logistic regressions using X1 only, X2 only, and then X1 and X2 together. In all logistic regression tables, we present the log odds coefficients, rather the odds ratios, because comparisons across the different methods we discuss are more straightforward and easier to understand when coefficients are focused on. The results are presented in Table 1.

Usually when variables are added to a model using OLS regression, the direct effects of variables estimated in earlier stages tend to decrease in magnitude. For example, race and education may both affect income. Hence, if the first model only has race in it, and the second model adds education, the effect of race tends to be smaller in the second model, because race and education are correlated; part of the reason racial groups differ in their incomes is because they also differ in their levels of education, hence the estimated direct effects of race become smaller once education is included in the model.

However, in our example, we predict a dichotomous outcome using logistic regression rather than a continuous outcome using OLS. Here, the effects of X1 and X2 seem to increase dramatically when both variables are in the model. In the bivariate logistic regressions, the effects of X1 and X2 are each about 2.20, but in the model that has both variables, their estimated effects are more than twice as large – about 4.5.

If we saw something similar happen in an OLS regression, we would most likely suspect suppressor effects were responsible. Suppressor effects might be present if X1 and X2 were negatively correlated, but both had positive effects on the dependent variable Y. This might occur if, for example, you were to study a remedial reading program for low-income children. Income and participation in

Table 1
Logistic regressions - hypothetical data.

	Xonly	Zonly	Both
Y			
X1	**2.17*****		**4.50*****
	(0.21)		(0.60)
X2		**2.21*****	**4.54*****
		(0.21)	(0.60)
pseudo R^2	0.182	0.188	0.469
VarYstar	4.46	4.51	13.53
SDYstar	2.11	2.12	3.68
N	500	500	500

Standard errors in parentheses.
With controls added the coefficient for X1 increases by 2.34, i.e. by 108%.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

[3] Potentially, the coefficient for race might become zero after variables are added. However, it would usually be a mistake to conclude that race does NOT affect the outcome. Rather, the analysis may be demonstrating why it is that race and the outcome are related. Race affects X which in turn affects Y; if so the effect of race on the outcome is *indirect* rather than *direct*.

[4] Among other things, Mood (2010, p. 67) notes “Logistic regression estimates do not behave like linear regression estimates in one important respect: They are affected by omitted variables, even when these variables are unrelated to the independent variables in the model.” Our examples will illustrate why Mood’s claim is true.

Table 2
Linear Probability Models - Hypothetical Data

	LPM-Xonly	LPM-Zonly	LPM-Both
X1	**0.48*****		**0.48*****
	(0.04)		(0.03)
X2		0.49***	0.49***
		(0.04)	(0.03)
R^2	0.238	0.246	0.484
VarY	0.25	0.25	0.25
SDY	0.50	0.50	0.50
N	500	500	500

Standard errors in parentheses.
When controls are added the coefficient for X1 does not change.
$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

the program would be negatively correlated, because lower incomes are required to qualify for the program. But, if the program is effective, both participation in the program and higher incomes (because of the learning advantages associated with them) would likely lead to higher reading scores. Therefore, if you estimated a model in which only program participation predicted reading scores, the positive effect of the program would be suppressed because you would not be controlling for how lower incomes reduce the scores of program participants. However, if you included both program participation and income in the model, you would see that the estimated effect of the program becomes more positive (perhaps even changing signs from negative to positive) once income is added to the model. In this way, suppression is one common reason why we would see coefficients increasing in magnitude across nested models as more variables are added.

However, suppression is NOT the cause of the dramatic changes shown here. Remember that 125 cases were assigned to each of the four possible combinations of values of X1 and X2. This might happen, for example, in an experimental study where subjects are randomly assigned to treatment groups. As a result, X1 and X2 are not and cannot be correlated, i.e. since cases are randomly assigned, knowing someone's value on X1 tells you absolutely nothing about their likely value on X2. If we use OLS regression on our hypothetical data instead of logistic regression – i.e. if we estimate a Linear Probability Model – we get the results shown in Table 2.

With OLS regression, the coefficients for X1 and X2 are the same in all three models – which is what should happen in OLS regression when the independent variables are uncorrelated with each other, as is the case here.

Hence, we have a dramatic contrast. In the OLS regressions, the effects of the uncorrelated X1 and X2 are the same regardless of whether they are entered into the model individually or together. On the other hand, in the logistic regressions, the estimated effects of X1 and X2 more than double once both are in the model at the same time.

Note that some contend that the Linear Probability Model is actually better than the logistic regression model, or at least adequate, in many cases. Paul von Hippel (2015, 2017), for example, argues that the LPM coefficients are easier to interpret and that the logistic and linear models often yield results that are virtually indistinguishable. The fact that the LPM does NOT suffer from the problems with comparisons across nested models that we illustrate here is another argument in its favor. Nonetheless, given the problems that occur when binary dependent variables are analyzed via OLS (See Long and Freese, 2014; or numerous other sources for a discussion), and given that marginal effects provide an alternative way of making logistic regression results easier to interpret (Williams, 2012; see also Allison, 2017) we think the solutions we discuss later are generally superior.

In these cases, the striking changes can serve as clues to a researcher that something is wrong with the analysis. However, the clues may not always be so obvious. We now present a second example, which shows a variant of the problems we have described but which is so subtle the researcher might not even suspect that something does not seem quite right.

2.2. *Example using the ANES 2008 time series study*

In an experimental setting, such as with our hypothetical data, it would be possible to assign treatments X1 and X2 so that they would not be correlated. With survey and other observational data, however, it would be unusual to find situations where both X1 and X2 strongly affected Y and yet were uncorrelated with each other.[5] We therefore provide a second example, this time using the American National Election Studies (ANES) 2008 Time Series Study. Quoting the study's online documentation (The American National Election Studies, ANES 2015),

> This study is part of the American National Election Study (ANES), a time-series collection of national surveys fielded continuously since 1952. The American National Election Studies are designed to present data on Americans' social backgrounds, enduring political predispositions, social and political values, perceptions and evaluations of groups and candidates, opinions on questions of public policy, and participation in political life. The 2008 ANES data consists of a time series study conducted both before and after the 2008 presidential election in the United States … Like its predecessors, the 2008 ANES was

[5] An exception might be for demographic variables such as race and gender, which are usually weakly correlated but which might both have effects on an outcome of interest.

Table 3
Variables.

Variable	Description
Obama	Who did the respondent vote for in 2008? 0 = John McCain, 1 = Barack Obama
White	Race of Respondent (dichotomized). 1 = White Non-Hispanic, 0 = Not White (or else Hispanic)
Age	Age in Decades (e.g. 2.7 = 27 years old)
Income	Estimated Family Income in $10,000s (e.g. 4.6 = $46,000)
Bush	Feeling thermometer towards George W. Bush (ranges from a low of 0 to a high of 10)
Feminist	Feeling thermometer toward feminists (ranges from a low of 0 to a high of 10)

divided between questions necessary for tracking long-term trends and questions necessary to understand the particular political moment of 2008.

Table 3 shows the variables from ANES 2008 that are used in this analysis.[6] **Obama** is the dependent/outcome variable while the other variables are independen/explanatory. The Appendix shows exactly which variables were extracted from the ANES 2008 dataset and any data manipulations (e.g. recodings) that were done by us.

For this example, we are primarily interested in the effect of **white** and how it changes when the four other variables are added to the model.

In model 1 of Table 4, **Obama** is regressed on **white** alone. In model 2, **Obama** is regressed on the four control variables. In the final model, all of the independent variables are included.

On the surface at least, the results seem plausible. White people, older people, those with higher incomes, and those who felt more warmly towards George W. Bush were all less likely to vote for Barack Obama, while those who felt more warmly towards feminists were more likely to vote for Obama.

What may be somewhat surprising, however, is how the effect of being White changes – or rather does not change – once the control variables are added. In many analyses, racial differences get noticeably smaller once variables like income are controlled for, e.g. racial differences in income will become smaller once racial differences in education are accounted for. Yet, in this analysis, the effect of being White barely declines at all once other variables are added. The estimated effect of being White goes from −2.32 in the single-variable model to −2.13 in the full model, a decline of only about 8 percent. Further, the dichotomized measure of the variable **white** is indeed correlated with the other variables in the model. In this sample, White respondents tend to be somewhat older, have higher incomes, feel more warmly toward George Bush, and less warmly towards feminists.[7] A researcher might therefore reasonably wonder why the effect of **white** declines so little in these models once control variables are added.

Compare those changes to what happens when we instead estimate the models using OLS regression/Linear Probability Models in Table 5.

Once controls are added the effect of white drops from −0.44 down to −0.27, a decline of 39%. This is almost five times as large as the decline that the logistic regression coefficients suggested.

To summarize, in the first example the logistic regressions indicated that the effect of X1 doubled once another variable, X2, was

Table 4
Logistic regressions - Obama 2008 presidential vote.

	Xonly	Zonly	Both
Obama			
White	**−2.32*****		**−2.13*****
	(0.15)		(0.19)
Age		−0.18***	−0.11*
		(0.05)	(0.05)
income		−0.18***	−0.13***
		(0.02)	(0.02)
Bush		−0.63***	−0.64***
		(0.04)	(0.04)
feminist		0.32***	0.27***
		(0.04)	(0.04)
pseudo R^2	0.180	0.406	0.489
VarYstar	4.62	8.33	10.38
SDYstar	2.15	2.89	3.22
N	1372	1372	1372

Standard errors in parentheses.
With controls added the coefficient for white decreases (in magnitude) by 0.19, i.e. by 8.2%.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

[6] The original values of age, income, Bush, and feminism were divided by ten to make the output more readable.
[7] Results are available upon request, but can also easily be confirmed with the replication materials in the appendix.

Table 5
Linear probability models - Obama 2008 presidential vote.

	LPM-Xonly	LPM-Zonly	LPM-Both
White	**−0.44*****		**−0.27*****
	(0.02)		(0.02)
Age		−0.02***	−0.01*
		(0.01)	(0.01)
income		−0.02***	−0.02***
		(0.00)	(0.00)
Bush		−0.09***	−0.08***
		(0.00)	(0.00)
feminist		0.04***	0.03***
		(0.00)	(0.00)
R^2	0.213	0.428	0.496
VarY	0.23	0.23	0.23
SDY	0.48	0.48	0.48
N	1372	1372	1372

Standard errors in parentheses.
With controls added the coefficient for white decreases (in magnitude) by 0.17, i.e. by 39%.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

added to the model. Conversely, in the second example, the effect of the first variable, **white**, barely changed at all when controls were included. While these may seem like very different situations, they are actually both illustrations of the same problem: in logistic regression, naïve comparisons of coefficients across nested models are inherently problematic. In the next section, we explain why.

2.3. Statistical and intuitive explanations

We now offer both a statistical explanation and an intuitive explanation as to why these problems occur.

2.3.1. Statistical explanation

It is common to use techniques such as OLS regression to estimate models where an observed, continuous independent variable, Y, is a linear function of one or more independent X variables, i.e.

$$Y = \alpha + \sum X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

Since the residuals are uncorrelated with the Xs, it follows that

$$V(Y) = V\left(\alpha + \sum X\beta\right) + V(\varepsilon)_{= \text{Explained Variance} + \text{Residual Variance}}$$

As you add explanatory variables to a model, explained variance goes up, and the residual variance goes down. The variance of the observed variable Y stays the same.[8]

But suppose the observed Y is not continuous – instead, it is a collapsed version of an underlying unobserved variable, Y*. For example, respondents might be asked "Do you approve or disapprove of the President's health care plan? 1 = Approve, 2 = Disapprove."

Attitudes are usually not simply binary; rather, they run along a continuum, in this case ranging from strongly approve to strongly disapprove. We do not actually observe the exact values; we just observe the binary responses to the question. In effect, the underlying continuum of attitudes has been collapsed into two values. We know if an individual supports the President's health care plan, but we do not know how strong or weak that support is.

Binary & ordinal regression techniques allow us to estimate the effects of the Xs on the underlying Y*. They can also be used to see how the Xs affect the probability of being in one category of the observed Y as opposed to another.

The latent variable model in binary logistic regression can be written as

$$\text{If Y}^* >= 0, \text{y} = 1$$

$$\text{If Y}^* < 0, \text{y} = 0$$

Since Y* is a latent unobserved variable, its scaling has to be set some way; that is, some constraint needs to be placed on the model in order to identify the parameter estimates. There are actually an infinite number of ways to do this, but some approaches have clear

[8] To put it another way – As you add variables to an OLS regression, the Total Sum of Squares stays the same, but the allocation between the Model (Explained) and Residual (Unexplained) Sums of Squares shifts, i.e. adding more variables increases the Explained SS and decreases the unexplained SS by a corresponding amount. Hence, the variance of Y (which is Mean Square Total/(N – 1)) stays the same regardless of what variables are added or dropped from the model. In other words, V(Y) is a fixed quantity and does not depend on the variables in the model.

advantages over others. In logistic regression, Y* is typically identified by assuming the error terms have a standard logistic distribution. A standard logistic distribution has a mean of 0 and a variance of $\pi^2/3$, or about 3.29. Fixing the residual variance at $\pi^2/3$ has many nice statistical properties, e.g. it makes it easy to compute odds ratios and to compute the probability that a case with a given set of characteristics will have a value of 1 on the outcome variable (Williams and Quiroz, 2019).

However, when comparing coefficients across nested models, identifying the model by fixing the residual variance also has major disadvantages. Since the residuals are uncorrelated with the Xs and the error term is assumed to have a standard logistic distribution, in logistic regression it follows that

$$V(Y^*) = V(\alpha + x\beta) + V(\varepsilon_{yY^*}) = V(\alpha + x\beta) + \pi^2 / 3 = V(\alpha + x\beta) + 3.29$$

Notice an important difference between OLS and Logistic Regression.

In OLS regression with an observed variable Y, V(Y) is fixed and the explained and unexplained variances change as variables are added to the model.

But, in logistic regression with an unobserved variable Y*, $V(\varepsilon_{Y^*})$ is fixed. Therefore, if additional variables improve the fit of a model, the explained variance has to go up to reflect this because the residual variance cannot go down. As a result, the total variance of Y* has to go up too.

2.3.2. *Intuitive explanation*

An alternative, more intuitive explanation is also possible. The variance of Y* changes because estimation of the probability of success gets better and better as you add more variables, and better estimation in turn leads to more variability in the estimates. Suppose you had a single dichotomous independent variable, e.g. a variable called gender, with two options: male and female.[9] All men would have the same predicted probability of success and all women would also share the same predicted probability. For example, men might have a 40% probability of success while women have a 50% chance.

As you add more variables, however, there will start to be more variability in the individuals' predicted probabilities. So, among the men, after adding another variable the probabilities might now range between 35% and 45%, while for women the range might be 45%–55%. Hence, instead of the probabilities only ranging from 40% to 50%, adding another variable could cause the individual probabilities to range from 35% to 55%. If you are doing really well, adding more variables might even get you to the point where the predicted individual probabilities of success ranged from 0.0001 to 0.9999.

To put it another way: When relevant variables are missing, you will overestimate the probability of success for some (e.g. some men will have less than a 40% chance of success) while underestimating it for others (e.g. some women will have better than a 50% chance). As your model improves, you will get more accurate estimates of the probability of success which in turn will result in greater variability in the values of Y*.

2.3.3. *Implications of rescaling*

Further, as Y* gets rescaled, the coefficients in each model get rescaled too. As a result, comparing nested logistic regressions where Y* is measured differently is like comparing OLS regressions where the measurement of the dependent variable is different in each model, e.g. in Model 1 income is measured in dollars, in Model 2 it is measured in thousands of dollars, etc. When the scaling of the dependent variable keeps changing, comparisons of coefficients across models is challenging and potentially misleading.

This difference has important implications. Comparisons of coefficients between nested models do not work the same way in logistic regression as they do in OLS; therefore a fair and accurate comparison in an OLS model may not be fair and accurate for a similar logistic regression model. Researchers using logistic regression with nested models who are not aware of this are likely to make at least small mistakes when interpreting their results, and possibly even serious ones.

2.4. *The examples revisited*

Returning to our hypothetical example, we can now see why the logistic regressions and OLS regressions behave so differently. Note that, as Table 2 showed, in the OLS regressions the variance of Y is indeed the same in all three models (0.25). It does not matter whether you add variables to the model or delete variables; the variance of Y is fixed.

Compare that to the variance of Y* in the three logistic regression models[10] (Table 1). The variances of Y* in the three models are 4.46, 4.51, and 13.53, i.e. the variance is three times as large when both X variables are in the model. Clearly, unlike the variance of Y in an OLS regression, the variance of Y* in a logistic regression is NOT invariant across models.

Once again, why does the variance of Y* go up? Because it has to. In logistic regression, the residual variance is fixed at about 3.29,

[9] While gender has historically been treated as binary and fixed at birth – including in the ANES 2008 dataset we use here – contemporary coding schemes have started using more complex measurement. For example, in 2018 the General Social Survey (Smith and Son, 2019) added separate questions on sex assigned at birth and current gender. Respondents were able to select "Woman," "Man," "Transgender," "A gender not listed here," and "No answer." But, out of 1409 respondents, only 9 cases were identified as transgender or another alternative gender designation. However, a person at the GSS told us that the Ns may get higher, more questions are coming, and that gender scholars are excited for the new content.

[10] As noted before, in a standard logistic regression, $V(Y^*) = V(\alpha + x\beta) + \pi^2/3$. In Stata, the estimated variance of Y* can be easily computed using Long and Freese's (2014) listcoef command. Alternatively Stata users can give commands like **logit y x1 x2 x3**; **predict yhat, xb**; **sum yhat**; **scalar vary = r(Var) + _pi^2/3**.

so improvements in model fit result in increases in explained variance, which in turn result in increases in total variance. The better the fit, the more variance.

Although perhaps less obvious, rescaling is also creating problems in the Obama 2008 example presented in Table 4. In the single-variable model, the variance of Y* is 4.62. In the model with all the variables, the variance of Y* is more than twice as great, 10.38. Because adding variables rescales the variance of Y*, the coefficient for the variable **white** gets rescaled upward as well. As a result, the decline in the effect of **white** is actually greater than it appears to be based on the logistic regression coefficients.

In short, comparisons of coefficients across nested models can be misleading because the dependent variable, and with it the model coefficients, are scaled differently in each model.

3. What are possible solutions?

We have deliberately chosen to include in this paper relatively simple, easy to explain examples of the problem at hand. However the solutions we discuss can be extended to more complicated models, such when there is a series of nested models or when more than one independent variable is of interest.

3.1. Do not present nested logistic or probit regression models in the first place

One simple solution is to only present coefficients for the final model. This might be an especially good strategy when intermediate models receive only limited attention anyway. Researchers could just present chi-square contrasts or give other reasons to justify how they picked their final model and then only present those coefficients (see, for example, Hauser and Andrew 2006). If the intermediate models are not really needed anyway, then presenting them just wastes space and potentially confuses the reader because of the way coefficients get rescaled across models, e.g. the reader might be misled by how the effect of race or gender changes across models, not realizing that at least part of the change is due to how Y* gets rescaled.

However, in many situations, only presenting the final model would not be a good approach. When authors estimate sequences of models, it is often because they want to see how the effects of variables decline (or increase) after other variables are controlled for. For example, a researcher might want to know how much of the effect of race is direct and how much is indirect (e.g. race affects education and education in turn affects the dependent variable.) In such cases, researchers may want or need to estimate and present intermediate models, but they have to do so in a way that avoids the complications shown here.

3.2. Use Y-standardization

Mare (2006; also see Winship and Mare 1984) states

> When the error variance is fixed, it is also inappropriate to make within sample comparisons among the coefficients for a given covariate across equations with varying subsets of covariates. In this case, the total variance of the latent dependent variable and thus the scale of the estimated coefficients vary from model to model as a function of the different regressors that are included. *Fixing the variance of the latent dependent variable avoids this problem.* [Emphasis added].

With Y-standardization, instead of fixing the residual variance, you fix the variance of Y* at 1. To get the Y-standardized coefficients, you simply take the original logistic regression coefficients and divide them by the estimated standard deviation of Y*. For example, in the first logistic regression presented in Table 1, the coefficient of X1 was 2.17, the standard deviation of Y* was 2.11, so the Y-standardized coefficient is 2.17/2.11 = 1.03 approximately. Table 6 applies Y-standardization to our earlier hypothetical example.

The Y-standardized coefficients change much less across nested models than the original logistic regression coefficients did (although the fact that they change at all is somewhat problematic; Y-standardization seems to work better when the changes across models are not as extreme as in this case, such as in our Obama example). Conversely, Table 7 illustrates the use of Y-standardization in our Obama 2008 example.

With the Y-standardized coefficients, the effect for **white** declines by about 39 percent – almost exactly the same drop as in the earlier Linear Probability Models (Table 4), and again much larger than the logistic regression coefficients (Table 1) had suggested.

Table 6
Y standardized coefficients - hypothetical data.

	Xonly	Zonly	Both
Y			
X1	**1.02*** **		**1.22*** **
	(0.209)		(0.599)
X2		1.04***	1.23***
		(0.210)	(0.599)
N	500	500	500

bStdY coefficients; Standard errors in parentheses.
With controls added the coefficient for X1 increases by 0.2, i.e. by 19.5%.
**p* < 0.05, ***p* < 0.01, ****p* < 0.001.

Table 7
Y standardized coefficients - Obama 2008 presidential vote.

	Xonly	Zonly	Both
Obama			
white	**−1.08***** (0.149)		**−0.66***** (0.191)
Age		−0.06*** (0.0472)	−0.03* (0.0508)
income		−0.06*** (0.0215)	−0.04*** (0.0232)
Bush		−0.22*** (0.0367)	−0.20*** (0.0395)
feminist		0.11*** (0.0402)	0.08*** (0.0434)
N	1372	1372	1372

bStdY coefficients; Standard errors in parentheses.
With controls added the coefficient for white decreases (in magnitude) by 0.42, i.e. by 38.7%.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

The changes in the Y-standardized coefficients in the models using hypothetical data should give researchers pause. But, if more complicated solutions are not feasible, Y-standardization at least mitigates the problem.

3.3. The Karlson/Holm/Breen (KHB) method

The Karlson, Holm, and Breen (KHB) method is arguably the best approach for comparing coefficients across nested models (Karlson et al., 2012; Kohler et al. 2011; Breen et al., 2018, 2021).

According to KHB, their method separates changes in coefficients due to rescaling from true changes in coefficients that result from adding more variables to the model (and does a better job of doing so than Y-standardization and other alternatives). They further note that with their method the total effect of a variable can be decomposed into its direct and its indirect components.

We first present the results for the KHB method with the Obama 2008 example (Table 8), and then explain how and why it works as a solution to the problem of nested logistic regression models.[11]

The KHB approach is both ingeniously clever and surprisingly simple. Here is a step-by-step description of the basics of the KHB method.

- You regress each Z variable on X. In this case you regress age on **white**, income on **white**, etc.
- You then compute the residual from that regression for each case. In the Obama 2008 example, the Z-residual variables are age_resid, income_resid, Bush_resid, and feminist_resid.[12]
- KHB then estimate what they call the *Reduced Model.* The Reduced Model regresses Y on X *and all the Z-residual variables.* In the current example, **Obama** is regressed on **white**, age_resid, income_resid, Bush_resid, and feminist_resid.
- The *Full Model* regresses Y on X and all the Z variables. In this case **Obama** is regressed on **white**, age, income, Bush, and feminist.
- Put another way, the Reduced Model includes X and all the Z-residual variables, while the Full Model includes X and all the Z variables.

The truly remarkable feature of the KHB approach is that the Full and Reduced models are actually the same model, just parametrized differently! The two models have the exact same fit and the exact same variance of Y*. Thus, comparisons between them are NOT distorted by having Y* scaled differently. This allows the researcher to look at mediation effects in an accurate way.

- In the Reduced Model, X is, by construction, totally uncorrelated with the Z-Residual variables. Therefore none of the effect of X on Y is mediated by any other variables. That is, in the Reduced Model, the effect of X on Y is the effect of X when there are no controls applied.
- In the Full Model, X is correlated with the Z variables, and the Z variables also affect Y. Hence, the effects of X are mediated by such things as race affecting income which in turn affects the presidential vote.

In terms of results, KHB paint a very different picture than a naïve look at the logistic regression coefficients did. In the KHB Reduced Model, the coefficient for **white** is −3.51. Once controls are added, the effect of **white** drops to −2.14, a decline of 1.37, or about 39 percent. Once again, the 39 percent reduction in the effect of **white** is much greater than the 8.2 percent reduction the simple

[11] As with the other approaches we discuss, the KHB method can work with more complicated examples, e.g. there can be multiple X variables in the model whose effects are decomposed, not just a single one like is being used here. In Stata, their easy to use khb program can be downloaded and installed from the Statistical Software Components (SSC) archive.

[12] In Stata, you could give commands like **regress age white** followed by **predict age_resid, residual** for each of the Z variables.

Table 8
KHB models - Obama 2008 presidential vote.

	Reduced	Full
Obama		
white	**−3.51***** (0.22)	**−2.13***** (0.19)
age_resid	−0.11* (0.05)	
income_resid	−0.13*** (0.02)	
Bush_resid	−0.64*** (0.04)	
feminist_resid	0.27*** (0.04)	
Age		−0.11* (0.05)
Income		−0.13*** (0.02)
Bush		−0.64*** (0.04)
feminist		0.27*** (0.04)
pseudo R^2	0.489	0.489
VarYstar	10.38	10.38
SDYsta	3.22	3.22
N	1372	1372

Standard errors in parentheses.
With controls added the coefficient for white decreases (in magnitude) by 1.38, i.e. by 39.4%.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

comparison of the original logistic coefficients suggested.

In the KHB approach, the −2.14 coefficient for race represents the *direct effect* of **white** on voting. The −1.37 difference between the Reduced and Full models reflects the *indirect effect* **white** has on vote, e.g., perhaps race affects income and income affects vote, or perhaps this reflects any non-causal reasons race and the Z variables are correlated.

But what should a researcher do with variables that are not thought to be mediated or mediators? For example, race is not a cause of gender, but both race and gender may affect an outcome of interest. KHB suggest that, if X (e.g. race) is not thought to affect a variable, Z (e.g. gender), that affects Y (e.g. presidential vote), you should classify that variable as a *concomitant* variable and include it in both the Reduced and Full Models. That way the decline in the effect of X on Y across models really will reflect the indirect effect of X alongside the effect of Z.

As Table 9 shows, the KHB method also produces striking results with our hypothetical data.

The logistic regression coefficients suggested the effects of X1 and X2 more than doubled when both were in the model at the same time. KHB shows that, in fact, neither effect increases at all, which is what should happen since X1 and X2 are uncorrelated. The huge increases in the estimated effects seen before were entirely due to the rescaling of Y* that occurs when additional variables are added to a model.

3.4. Marginal effects

The previous methods examined how changes in X affected the predicted values for Y, Y*, or Y-standardized. However, other than sign and significance, the coefficients are not very intuitive. What, for example, does it mean to say that the coefficient for **white** is −2.14 in the full Obama 2008 model? It means that White individuals were less likely to vote for Obama, but how much less likely? One percent, ten percent, fifty percent, or what?

Williams (2012) and numerous others (e.g. Long and Freese 2014) have therefore suggested that marginal effects can be a great aid when interpreting the substantive meaning of effects in logit and probit models.[13] For example, as we will show shortly, rather than saying the coefficient for **white** is −2.14, we can make the much more intuitive statement that, after controlling for other variables, on average White individuals were 25 percentage points less likely than Non-White individuals to vote for Obama in 2008.

[13] Williams (2012) notes that there are different types of marginal effects, and states that he personally prefers Average Marginal Effects (AMEs) over Marginal Effects at the Means (MEMs), because AMEs are often more realistic and make better use all of the data. The khb program estimates AMEs. Nevertheless, it should be possible to estimate MEMs if the researcher prefers.

Table 9
KHB models - hypothetical data.

	Reduced	Full
Y		
X1	**4.50*****	**4.50*****
	(0.60)	(0.60)
X2		4.54***
		(0.60)
_cons	−1.77***	−4.04***
	(0.31)	(0.59)
pseudo R^2	0.469	0.469
VarYstar	13.53	13.53
SDYsta	3.68	3.68
N	500	500

Standard errors in parentheses.
The z-residual variables are included in the Reduced Model but coefficients are not shown.
When controls are added the coefficient for X1 does not change.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 10
Average marginal effects - hypothetical data.

	Xonly	Zonly	Both
X1	**0.48*****		**0.48*****
	(0.04)		(0.03)
X2		0.49***	0.49***
		(0.04)	(0.03)
N	500	500	500

Standard errors in parentheses.
When controls are added the coefficient for X1 does not change.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

The use of marginal effects and adjusted predictions has additional benefits when used with nested models. Both Mize et al. (2019) and Karlson et al. (2012) argue that changes in marginal effects across nested models have more intuitive appeal than do changes in coefficients. Mize et al. (2019) further note that rescaling is not an issue with marginal effects. In other words, marginal effects can be a way to interpret changes in nested logit and probit models that avoids the Y* rescaling problem and provides substantive insight.

Table 10 shows the marginal effects for our hypothetical example.

The marginal effects show NO changes across models.[14]

Table 11 shows the marginal effects for our Obama 2008 example.

The average marginal effect of −0.44 in the single-variable model means that White individuals were 44 percentage points less likely to vote for Obama than were Non-White individuals.[15] (see Table 6) The AME of −0.25 for **white** in the final model means that, after controls, White voters were still 25 percentage points less likely than Non-White voters to support Obama. As has been the case with all the strategies we have tried, the decline in the effect is around 40 percent, not the 8 percent figure implied by the logistic regression coefficients. This shows that part of the racial differences in support for Obama were partly due to the fact that racial groups also differ on other important variables such as income.

Mize et al. (2019) demonstrate how to do formal tests of whether marginal effects significantly differ across nested models.[16] KHB, however, argue that, instead of using the Xonly model when calculating the original marginal effect, you should use their Reduced Model (which has X and all the Z-Residual variables in it). Our experience suggests that the differences between the two approaches tend to be slight. Further, the Mize, Doan, and Long approach could possibly be tweaked to use the KHB Reduced Model.

4. How serious is this problem in practice?

There are many ways researchers use intermediate models appropriately, without comparing coefficients across nested models. Indeed, it is very common to report model fit statistics (e.g. model chi-square, BIC, chi-square contrasts) for intermediate models, and

[14] The marginal effects would change if X were continuous rather than dichotomous. Still, the changes in the marginal effects tend to be much smaller than the changes in the logistic regression coefficients.

[15] More specifically, in the sample Obama captured almost 90 percent of the Non-White vote compared to only about 45 percent of the White vote.

[16] As of this writing, the code for the Mize-Doan-Long method is available on Mize's webpage at https://www.trentonmize.com/research. The coding is fairly complicated but a simpler-to-use Stata routine may eventually be forthcoming.

Table 11
Average marginal effects - Obama 2008 presidential vote.

	Xonly	Zonly	Both
white	**−0.44***** (0.02)		**−0.25***** (0.02)
Age		−0.02*** (0.01)	−0.01* (0.01)
income		−0.02*** (0.00)	−0.01*** (0.00)
Bush		−0.08*** (0.00)	−0.07*** (0.00)
feminist		0.04*** (0.00)	0.03*** (0.00)
N	1372	1372	1372

Standard errors in parentheses.
With controls added the coefficient for white decreases (in magnitude) by 0.19, i.e. by 42.6%.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

to give coefficient estimates only for the final model. (See, for example, Hauser and Andrew, 2006). If you are not going to focus much on changes in coefficients anyway, this is a very good strategy, since presenting unnecessary and potentially misleading coefficient estimates may do more harm than good.

However, what about when researchers do present nested models? In our two examples, the naïve comparisons of nested models were highly misleading. In the hypothetical example, the effects of variables seemed to more than double once controls were added, when in reality they did not change at all. In the Obama 2008 example, the effect of being White seemed to decline only modestly, about 8 percent, when controls were added, but in reality the decline was closer to 40 percent.

The hypothetical example is, admittedly, unusual and extreme. In experimental studies it is not unusual to design the study so that the independent/treatment variables are uncorrelated. However, it may be unusual to have effects this strong. Cases that were 0 on both X1 and X2 had less than a two percent predicted probability of being a 1 on Y. Conversely when both X1 and X2 equaled 1, the predicted probability of being a 1 on Y was over 99%. Further, since the experimenter knows X1 and X2 are uncorrelated, there is little reason to run nested models with them. Also, in real world situations it would be unusual to find two X variables (besides perhaps race and gender) that both strongly affected Y while at the same time being uncorrelated with each other. This hypothetical example provides proof of concept and a warning for experimental studies, and we do not claim it to be a typical example of other situations, e.g. survey data.

The Obama 2008 example using real data is more realistic. Still, it too may be an unusual case. There are very sharp racial differences in American presidential voting, with Non-White voters (especially Black voters) being much more likely to vote for Democratic candidates. Further, race was no doubt especially important in 2008 because Obama was the first Black presidential candidate of a major political party. While studies often find major differences across racial groups, the impact of race may have been exceptionally large in the 2008 election.

So, how large has the problem been in published work? Without actually re-analyzing past papers that have used nested models, it is hard to assess how much impact these problems have actually had. We easily found dozens of recent papers that present sequences of nested models. They include potentially problematic statements such as

- "The effect of race on the dependent variable is even stronger once GPA, SES, and sex are controlled for (Model 2), indicating that when blacks and whites have equal GPAs and family SES, blacks are more likely to agree with this statement." (Matthew, 2011, p. 235)
- "Also of interest are changes in Model 1 estimates with the introduction of prior drinking. Across Models 1 and 2, the coefficient for partner drinking increases by 22 percent, suggesting that between-partner binge drinking similarity increases once prior behavior is controlled." (Kreager and Haynie, 2011, p. 751)
- "Most important, controlling for these micro-level mobility predictors accounts for a sizeable share of the association between local immigrant concentrations and inter-tract mobility among black householders, with the coefficient for this variable reduced by almost half (from 0.025 to 0.013) between Models 1 and 2." (Crowder et al., 2011, p. 36).

The numbers in such statements are probably at least a little off. What appears to be a larger effect as variables are added to models may in fact be a rescaling effect: coefficients get larger as the variance of Y* increases. Or, conversely, the decline in a coefficient's estimated value across nested models may actually be even greater than it appears to be. But, without re-analyzing the data you cannot tell whether conclusions are seriously distorted as a result.[17]

[17] Ervin (Maliq) Matthew (2011) graciously provided us with the data used for his paper. As noted, the article contains potentially problematic statements such as "The effect of race on the dependent variable is even stronger once GPA, SES, and sex are controlled for". In practice, however, we found that any potential errors were modest in Matthew's analysis. For example, his Table 7 somewhat understates how much the effect of race declines as controls are added. Some of his statements may be slightly off but not enough to undermine his major conclusions.

Overall, researchers should realize that

- Increases in the magnitudes of coefficients across nested models need not reflect suppressor effects. Instead, the increases could be due to the variance of Y* and the model coefficients being rescaled upward. Solutions such as using Y-standardization, KHB, or marginal effects can give researchers a clearer assessment of what is a suppressor effect and what is a misinterpretation of variance.
- Conversely, declines in coefficients across models can actually be understated, i.e. you will be understating how much other variables account for the estimated direct effects of the variables in the early models. Again, the solutions proposed in this paper address this problem.
- Distortions are potentially more severe when added variables strongly affect Y*, as the variance of Y* will increase more when that is the case.

5. Summary & conclusions

When you estimate a series of nested models using logit or probit, comparisons of coefficients across models may be problematic, because Y* and the model coefficients are scaled differently in each model.

In logistic regression, when additional variables are added to a model, Y* and its variance are rescaled upwards, and the coefficients are rescaled too. Real change produced by the mediating effects of the control variables can be confounded with artificial change caused by the rescaling of Y*.

Our two examples illustrated how researchers might be misled when naïvely comparing coefficients across nested models. In our hypothetical example with two uncorrelated variables, the effects of each more than doubled when they were placed in the model together. In reality, neither of their effects changed. The rescaling of Y* was entirely responsible for the differences in estimated effects.

Conversely in the Obama 2008 example, the estimated effect of being White barely declined at all, only about 8 percent, once controls were applied. Since race is correlated with the added variables (e.g. White people tend to have higher incomes than other racial groups) this seemed surprisingly small. Indeed, each of the other approaches we tried – linear probability models, Y-standardized coefficients, the KHB method, marginal effects – all showed that the decline in the effect of the variable **white** was more in the range of 40 percent. The rescaling of Y* that occurred once control variables were added caused the variable's direct effect to seem to decline much less than it actually did.

In both of our examples, a researcher could come to some very erroneous conclusions. Any naïve analysis comparing coefficient changes across nested models will probably be at least a little off, and in some situations might even be seriously off.

Without reanalyzing past papers, it is hard to assess how serious the problems have been in practive. Our guess (possibly wrong) is that distortions have not been too severe in most papers, but there probably are at least a few published papers where the problems are more serious.

What should researchers do to avoid this problem?

They may just want to not even present the results from nested models. Often people ignore everything but the final model, so why waste space on intermediate results you aren't using and which could mislead people?

If you do want to present sequences of nested models and see how coefficients change (e.g. you want to see how the direct effect a variable like **white** declines or increases as more variables are added to the model) the other approaches shown here all have strong advantages. Y-standardization and the KHB method largely control for rescaling, with KHB doing an especially good job. The use of marginal effects in recent years has grown increasingly popular, because they can make the substantive implications of logistic regression results much clearer and they do not suffer from rescaling problems with nested models.

Scholars in the past may have made some mistakes, but they nonetheless made many important contributions to our understanding of how the world works. They can be forgiven for not knowing about the potential problems with nested models or how to deal with them. Researchers today are in a much more advantaged position though. They can and should become well aware of the problems that can arise with the analysis of nested models and employ the tools that can lead to superior analyses in the future.

Acknowledgements

The ANES 2008 Time Series Study was collected and distributed by American National Election Studies. ANES is a collaboration of Stanford University and the University of Michigan. ANES 2008 was supported by the National Science Foundation under Grants SES0535334, SES-0720428 and SES-0840550. Additional funding from NSF was made available in conjunction with the Department of Homeland Security as part of grant SES-0651271. The ANES 2008 Principal Investigators were Jon A. Krosnick, Stanford University, and Arthur Lupia, University of Michigan. Vincent Hutchings at the University of Michigan was Associate Principal Investigator. Data and documentation can be downloaded at https://electionstudies.org/data-center/2008-time-series-study/. ANES graciously granted permission for us to distribute a small extract of their data so that others could easily replicate our work. ANES and its funding agencies bear no responsibility for use of the data or for interpretations or inferences based upon such uses.

Chris Quiroz developed an early version of the ANES 2008 example. Cheng Wang identified key articles that used nested models in their research. Richard Campbell, Junrong Sheng, and Steve Morgan provided helpful comments on earlier drafts, as did the two anonymous reviewers and guest editor Shawn Bauldry. The comments of the Notre Dame Spring 2022 Sociology 73994 class, Categorical Data Analysis, helped identify areas where the discussion could be and needed to be clearer. Any remaining faults are, of course, the authors' own.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssresearch.2022.102802.

References[18]

Allison, Paul, 2017. In: Defense of Logit – Part 2. Statistic Horizons Blog Page. https://statisticalhorizons.com/in-defense-of-logit-part-2/.

The American National Election Studies (ANES), 2015-11-10. ANES 2008 Time Series Study. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR25383.v3.

Berk, R.A., Ray, S.C., 1982. Selection biases in sociological data. Soc. Sci. Res. 11 (4), 352–398. https://doi.org/10.1016/0049-089x(82)90003-5.

Breen, Richard, Karlson, Kristian B., Anders, Holm, 2018. Interpreting and understanding logits, probits, and other nonlinear probability models. Annu. Rev. Sociol. 44, 39–54. https://doi.org/10.1146/annurev-soc-073117-041429.

Breen, Richard, Karlson, Kristian B., Anders, Holm, 2021. A note on a reformulation of the KHB method. Socio. Methods Res. 50 (2), 901–912. https://journals.sagepub.com/doi/full/10.1177/0049124118789717.

Crowder, K., Hall, M., Tolnay, S.E., 2011. Neighborhood immigration and native out-migration. Am. Socio. Rev. 76 (1), 25–47. https://doi.org/10.1177/0003122410396197.

Hauser, R.M., Andrew, M., 2006. Another look at the stratification of educational transitions: the logistic response model with partial proportionality constraints. Socio. Methodol. 36, 1–26. https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9531.2006.00172.x.

Henretta, J.C., 1979. Race differences in middle class lifestyle: the role of home ownership. Soc. Sci. Res. 8 (1), 63–78. https://doi.org/10.1016/0049-089x(79)90014-0.

Karlson, Kristian B., Anders, Holm, Richard, Breen, 2012. Comparing regression coefficients between same-sample nested models using logit and probit: a new method. Sociol. Methodol. August 2012 42 (1), 286–313. https://journals.sagepub.com/doi/10.1177/0081175012444861.

Kohler, U., Karlson, K.B., Holm, A., 2011. Comparing coefficients of nested nonlinear probability models. STATA J.: Prom. Commun. Statis. Stata 11 (3), 420–438. http://www.stata-journal.com/article.html?article=st0236.

Kreager, D.A., Haynie, D.L., 2011. Dangerous liaisons? Dating and drinking diffusion in adolescent peer networks. Am. Socio. Rev. 76 (5), 737–763. https://doi.org/10.1177/0003122411416934.

Long, J. S., and J. Freese. 2014. Regression Models for Categorical Dependent Variables Using Stata. 3rd ed. College Station, TX: Stata Press.

Magidson, J., 1981. Qualitative variance, entropy, and correlation ratios for nominal dependent variables. Soc. Sci. Res. 10 (2), 177–194. https://doi.org/10.1016/0049-089x(81)90003-x.

Mare, R.D., 2006. Response: statistical models of educational stratification—Hauser and Andrew's models for school transitions. Socio. Methodol. 36, 27–37. https://journals.sagepub.com/doi/abs/10.1111/j.1467-9531.2006.00173.x.

Marini, M.M., 1985. Determinants of the timing of adult role entry. Soc. Sci. Res. 14 (4), 309–350. https://doi.org/10.1016/0049-089x(85)90015-8.

Matthew, Ervin (Maliq), 2011. Effort optimism in the classroom: attitudes of black and white students on education, social structure, and causes of life opportunities. Sociol. Educ. 84 (3), 225–245. https://journals.sagepub.com/doi/10.1177/0038040711402360.

McDougall, G.S., Bunce, H., 1986. Race, moving status, and urban services in central cities. Soc. Sci. Res. 15 (1), 82–96. https://doi.org/10.1016/0049-089x(86)90005-0.

Mize, Trenton, Doan, Long, Scott Long, J., 2019. A general framework for comparing predictions and marginal effects across models. Socio. Methodol. 49 (1), 152–189. https://journals.sagepub.com/doi/full/10.1177/0081175019852763.

Mood, Carina, 2010. Logistic regression: why we cannot do what we think we can do, and what we can do about it. Eur. Socio Rev. 26 (1), 67–82. https://doi.org/10.1093/esr/jcp006.

Rauma, D., 1984. Going for the gold: prosecutorial decision making in cases of wife assault. Soc. Sci. Res. 13 (4), 321–351. https://doi.org/10.1016/0049-089x(84)90008-5.

Regnerus, Mark D., Elder, Glen H., 2003. Religion and vulnerability among low-risk adolescents. Soc. Sci. Res. 32 (4), 633–658. https://doi.org/10.1016/S0049-089X(03)00027-9. December 2003.

Smith, Tom W., Son, Jaesok, 2019. Transgender and Alternative Gender Measurement on the 2018 General Social Survey. GSS Methodology Report No. 129. Retrieved May 1, 2022. https://gss.norc.org/Documents/reports/methodological-reports/MR129%20transgender%202018.pdf.

Vanderweele, Tyler. 2015. Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press.

Von Hippel, Paul, 2015. Linear vs. Logistic Probability Models: Which is Better, and When? Statistical Horizons Blog. Retrieved April 24, 2022. https://statisticalhorizons.com/linear-vs-logistic/.

Von Hippel, Paul, 2017. When can you fit a linear probability model? More often than you think. Statistical Horizons Blog. Retrieved April 24, 2022. https://statisticalhorizons.com/when-can-you-fit/.

Williams, Richard, 2012. Using the margins command to estimate and interpret adjusted predictions and marginal effects. STATA J. 12 (2), 308–331. https://doi.org/10.1177/1536867X1201200209.

Williams, Richard A., Quiroz, Christopher, 2019. Ordinal regression models. In: Atkinson, P., Delamont, S., Cernat, A., Sakshaug, J.W., Williams, R.A. (Eds.), SAGE Research Methods Foundations. https://doi.org/10.4135/9781526421036885901.

Winship, Christopher, Mare, Robert D., 1984. Regression models with ordinal variables. Am. Socio. Rev. 49 (4), 512–525. https://doi.org/10.2307/2095465. Aug., 1984.

Richard Williams (Corresponding Author) is a Full Professor and a former Chairman of the Department of Sociology at the University of Notre Dame. He received his Ph.D. from the University of Wisconsin in 1986. His teaching and research interests include Methods and Statistics, Demography, and Urban Sociology. His work has appeared in the American Sociological Review, Social Forces, Social Problems, Demography, Sociology of Education, Journal of Urban Affairs, Cityscape, Journal of Marriage and the Family, Socius, The Stata Journal, and Sociological Methods and Research. He is a co-editor of the recently released ten volume Sage Research Methods Foundations. His work has focused on Racial, Economic and Institutional Disparities in Home Mortgage Lending; using bibliometric data to assess and evaluate academic productivity; and methodological issues in the analysis of categorical data. Contact Information: 4060 Jenkins Nanovic Halls, Department of Sociology, University of Notre Dame, Notre Dame, IN 46556. Orcid ID: https://orcid.org/0000-0001-9974-8307. Email: rwilliam@nd.edu.

Abigail Jorgensen is a PhD candidate in Sociology at the University of Notre Dame studying methodology, family, gender, culture, and politics. She is a Notebaert Fellow and the recipient of the University's 2022 Social Justice Award for her academic and community work with mothers, especially bereaved mothers. Before converting to sociology, she earned a Bachelor of Arts in both Political Science and Gender Studies at the University of Notre Dame with a minor in Theology. She earned

[18] Unless otherwise stated all links were last accessed and verified on May 1, 2022. Crossref (https://search.crossref.org/) was often used to get APA-formatted citations and the online links for papers

a master's degree in Sociology from the University of Notre Dame in May 2018 and proposed her dissertation in spring 2019. For that project, she is currently executing a four-year research design involving longitudinal interviews and surveys of 84 women in the state of Wisconsin. Contact Information: 4060 Jenkins Nanovic Halls, Department of Sociology, University of Notre Dame, Notre Dame, IN 46556. Orcid ID: https://orcid.org/0000-0003-1435-2093. Email: Abigail.Jorgensen.6@nd.edu.