Using predictions to compare groups in regression models for binary outcomes[*]

J. Scott Long[†] and Sarah A. Mustillo[‡]

March 5, 2018

Abstract

Methods for group comparisons using predicted probabilities and marginal effects on probabilities are developed for regression models for binary outcomes. Unlike approaches based on the comparison of regression coefficients across groups, the methods we propose are unaffected by the identification of the coefficients and are expressed in the natural metric of the outcome probability. While we develop our approach using the logit model with two groups, we consider how our interpretive framework can be used with a broad class of regression models and can be extended to any number of groups.

[*] We thank Long Doan, Trent Mize, Rich Williams, and an anonymous reviewer for their comments.

[†] jslong@iu.edu, Departments of Sociology & Statistics, Indiana University, Bloomington, IN 47401.

[‡] Sarah.A.Mustillo.5@nd.edu; Department of Sociology, University of Notre Dame, Notre Dame, IN 46556.

Using predictions to compare groups in regression models for binary outcomes

1 Introduction

Regression models comparing groups are used in many areas of research to answer two types of questions. First, do groups differ in the level of the outcome after adjusting for differences in observed characteristics? For example, do white and nonwhite respondents have different probabilities of reporting good health, controlling for age, income, education, and other characteristics? Second, does the effect of a regressor on the outcome differ across groups? For example, does obesity have the same effect on being diagnosed with diabetes for white and nonwhite respondents?

To answer these questions, models are fit that allow the regression coefficients to differ by group. To test if the coefficients are equal across groups, a Wald test is used (Chow, 1960). For example, suppose that we are considering the effect of x_k on y for white and nonwhite respondents, where β_k^W and β_k^N are the coefficients of interest. If $H_0\colon \beta_k^W = \beta_k^N$ is rejected, we conclude that the effects differs across groups. While this approach can be used with linear regression and some other models (Liao, 2002), Allison (1999) shows that since the regression coefficients in binary logit or probit are only identified to a scale factor, standard tests of the equality of coefficients are invalid. He develops tests that address the identification problem by adding untestable, auxiliary assumptions that the coefficients for some regressors are equal across groups. While his test addresses the identification problem, we believe that in most applications it is substantively more useful to understand whether the marginal effects of x_k on the probability of the outcome are the same in both groups than whether the regression coefficients for x_k are equal. Critically, in binary probit and logit, the equality of regression coefficients across groups does not imply that the marginal effects of a regressor on the probability are equal. In this paper, we develop methods for group comparisons using tests of the equality of probabilities conditional on the regressors and tests of the equality of marginal effects on the probability. Since probabilities are identified, these tests do not require additional identifying assumptions.

Our paper focuses on methods for comparing groups using the logit and probit models for binary outcomes for several reasons. First, these are the most commonly used models for binary outcomes. Second, much of the recent work on group comparisons has addressed the issue of how group comparisons must deal with the scalar identification of the regression coefficients in binary logit and probit. It is important to understand how our approach is related to this work. Our methods, however, can be generalized to any regression model for which it is possible to make conditional predictions and estimate marginal effects. In models where the regression coefficients are expressed in the outcome metric of interest, such as linear regression or the linear probability model, the regression coefficients are often the effects of interest. Even in these cases, our approach is necessary when there is a nonlinear relationship between a regressor and the outcome or to explore statistical interactions more generally. For example, in a linear regression model that includes age and age-squared as regressors, the marginal effect of age is computed using the coefficients for both age and age-squared and the conclusions depend on the values of regressors at which the test is made. Similarly, in any model in which predictions can be made, our methods for comparing predictions and marginal effects can be used. This includes models such as negative binomial regression, ordered logistic regression, and multinomial logit. With some adjustments, our approach can also be used with non-parametric regression. Generalizations are considered in section 5.

The substantive advantages of tests based on predicted probabilities are not without costs. Tests comparing regression coefficients are simple to apply since the only hypothesis of interest is whether the coefficients are equal across groups and the results do not depend on the values of the regressors in the model. For example, when examining racial differences in the regression coefficient for obesity on the onset of diabetes, the conclusion is either that the coefficients are the same or that they are not. With methods based on probabilities, including marginal effects, the conclusions depend on the values of the regressors where the comparison is made. For example, there might be no difference in the probability of diabetes for nonwhites and whites who have low income and a high school education, while the probabilities might differ for those with high income and a college degree. Similarly, the size of the marginal effect of a regressor on the outcome probability depends on the value of the regressor where the effect is computed as well as the values of all other variables in the model (Long, 1997). For example, the effect of obesity on diabetes for a 70-year-old, married man could be the same for both groups (i.e., the null hypothesis is not rejected), while the effect of obesity for a 50-year-old, single women could be significantly larger if the woman was white than if she was nonwhite. While conclusions about group differences in the effect of obesity on diabetes are more complex than those from testing group differences in regression coefficients, they also have the potential to provide more useful insights into the substantive process being studied.

The question of how to assess group differences in logit and other nonlinear models has attracted considerable attention across many disciplines, both directly in research dealing with statistical methods for comparing groups and indirectly in research about interactions in nonlinear models (e.g., Ai and Norton 2003, Buis 2010, Kendler and Gardner 2010, Norton et al. 2004). This work shows that conventional techniques are flawed in several ways, with research focusing on two approaches to group comparisons. One approach makes comparisons of regression coefficients that are in the metric of an underlying latent variable. Allison (1999) shows that group differences in unobserved heterogeneity invalidate traditional tests for comparing regression coefficients across groups. Allison (1999) and Williams (2009) developed new tests for comparing regression coefficients that account for differences in unobserved heterogeneity. Breen et al. (2014) present methods for group comparisons of correlations between the latent outcome and each regressor. Both the Allison and Breen approach compare effects that are not in the metric of the outcome probability. Kuha and Mills (2018) argue that these methods are only relevant when the latent outcome is of substantive interest. A second approach compares groups in terms the probability or the odds of the outcome. Long (2005, 2009) uses graphs to compare conditional probabilities and provides tests of the equality of probabilities across groups. Other researchers considered the comparison of odds ratios and marginal effects. Mood (2010) reviews how unobserved heterogeneity affects the group comparison of odds ratios and regression coefficients and argues that marginal effects of regressors on the probability are substantively more informative. Landerman et al. (2011) demonstrate how regression coefficients for interaction terms in logit models for panel data can provide misleading results about group differences in rates of change in the outcome. They recommend comparing group differences in average marginal effects of regressors on the probability of the outcome rather than the comparison of odds ratios. Mustillo et al. (2012) develop tests for group differences in growth trajectories in longitudinal mixed models for counts and argue that the interpretation of group-by-time interaction terms are misleading and suggest the comparison of average marginal effects on the rate.

In this paper, we build on exiting research to develop a general framework for the comparison of groups in regression models in terms of the probability of the outcome and marginal effects of regressors on the probability. Our predictive methods produce graphs and tables that can answer substantively motivated questions about group differences. Our methods do not involve deriving new tests, but rather the use of standard methods of specifying models, testing predictions across groups, and comparing marginal effects in ways that avoid traps of misinterpretation common in the substantive literature. The next section explains how the identification of regression coefficients affects group comparisons in the binary logit and probit models and how this issue is avoided by methods based on predictions. Section 3 presents methods for comparing conditional predictions and marginal effects. Each methods is illustrated in section 4 where we compare white and nonwhite respondents in models predicting being diagnosed with diabetes and reporting having good health. Section 5 discusses briefly how our approach can be used with any regression model in which predictions and marginal effects can be computed and explains why it is especially useful in models where the outcome has a nonlinear relationship to regressors, including linear regression when nonlinearities are included on the right hand side of the model.

2 Scalar identification in binary logit and probit

The identification of regression coefficients is critical for understanding group comparisons in logit and probit models. To explain this we begin by reviewing how coefficients are compared across groups in linear regression (Chow 1960). To simplify the presentation, we use two groups with two regressors, but the results can be easily generalized to G groups and K regressors. Let y be a continuous, observed dependent variable regressed on x_1 and x_2 for groups defined by $g=0$ and $g=1$. We begin by fitting separate regressions which allows the regression coefficients and error variances to differ by group:

$$\begin{aligned}\text{Group 0:}\quad & y = \beta_0^0 + \beta_1^0 x_1 + \beta_2^0 x_2 + \varepsilon_0 \text{ where } Var(\varepsilon_0) = \sigma_0^2\\ \text{Group 1:}\quad & y = \beta_0^1 + \beta_1^1 x_1 + \beta_2^1 x_2 + \varepsilon_1 \text{ where } Var(\varepsilon_1) = \sigma_1^2\end{aligned}$$

To assess whether the effect of x_k is the same for both groups, we test the hypothesis $H_{\beta_k}\colon \beta_k^0 = \beta_k^1$ using a Wald or likelihood ratio test. If H_{β_k} is rejected, we conclude that the effect of x_k differs by group.

If y is binary, the corresponding regression equations are

$$\begin{aligned}\text{Group 0:}\quad & \Pr_0(y=1 \mid x_1, x_2) = F(\beta_0^0 + \beta_1^0 x_1 + \beta_2^0 x_2)\\ \text{Group 1:}\quad & \Pr_1(y=1 \mid x_1, x_2) = F(\beta_0^1 + \beta_1^1 x_1 + \beta_2^1 x_2)\end{aligned}$$

where F is the normal cumulative density function for the probit model and the logistic cumulative density function for the logit model. While it seems that we could assess whether the effect of x_k is the same for both groups by testing $H_{\beta_k}\colon \beta_k^0 = \beta_k^1$, such tests are invalid since regression coefficients in the binary regression model are only identified up to a scale factor (Amemiya 1981, 1489; Maddala 1983, 23; McKelvey and Zavoina 1975). Following Allison (1999), this can be shown by deriving the model using a latent dependent variable y^* that is related to x_1 and x_2 through the equation

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{1}$$

where the error ε has mean 0 and variance σ^2. When the latent y^* is greater than 0, y is observed as 1; otherwise, y is 0. For example, if a person's propensity y^* to have diabetes exceeds 0, she is diagnosed with diabetes and $y = 1$. If her propensity is at or below 0, she is not diagnosed with diabetes and $y = 0$.[1] The probability that $y = 1$ conditional on x_1 and x_2 is the proportion of the distribution of y^* that is greater than 0:

$$\Pr(y=1 \mid x_1, x_2) = \Pr(y^* > 0 \mid x_1, x_2)$$

Substituting the right-hand-side of equation 1 for y^* and rearranging terms, the probability can be expressed in terms of the error:

$$\Pr(y=1 \mid x_1, x_2) = \Pr(\varepsilon \leq \beta_0 + \beta_1 x_1 + \beta_2 x_2 \mid x_1, x_2) \tag{2}$$

For a model with a single regressor, figure 1 shows that the probability at specific values of x is the shaded area of the error distribution above $y^* = 0$. To compute this area we must know the mean, variance, and mathematical form of the error distribution. The error is assumed to be logistic for logit and normal for probit. As with the linear regression model, the mean is assumed to be 0. The variance, however, leads to an identification problem for the βs.

— Figures 1 and 2 here —

In linear regression the residuals $y_i - \widehat{y}_i$ are used to estimate the variance of the errors. This cannot be done with logit or probit since y_i^* is unobserved. To understand the implications of this, consider what happens when we multiply equation 1 by an arbitrary, unknown constant δ:

$$\delta y^* = (\delta\beta_0) + (\delta\beta_1)\, x_1 + (\delta\beta_2)\, x_2 + \delta\varepsilon \tag{3}$$

Using the notation $\gamma_k \equiv \delta\beta_k$, $\widetilde{y}^* \equiv \delta y^*$, and $\widetilde{\varepsilon} \equiv \delta\varepsilon$, equation 3 can be written as

$$\widetilde{y}^* = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \widetilde{\varepsilon} \tag{4}$$

and equation 2 as

$$\Pr(y=1 \mid x_1, x_2) = \Pr(\widetilde{\varepsilon} \leq \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 \mid x_1, x_2) \tag{5}$$

Since all that we did was multiply both sides of the inequality by δ and change notation, the probabilities in equation 5 are *exactly* the same as those in equation 2. However, since δ is

[1] While the latent variable derivation of the logit and probit models is compelling way to illustrate difficulties in comparing regression coefficients across groups, the same problem arises when the model is derived in other ways. (Allison 1999, 190; Breen and Karlson 2013, 170).

unknown, there is no way to distinguish between the true β coefficients that generate y^* and the rescaled γ coefficients. The effects of the change in scaling are shown in figure 2 that was created by multiplying the equation for y^* in figure 1 by $\delta = 2$. The intercept, slope, and standard deviation of the error are δ times larger, while the probabilities represented by the shaded proportion of the error distribution are unaffected by the change in scale.

Since the β coefficients are only identified to a scale factor, they cannot be estimated without assuming a value for the variance of the error. For probit, the usual assumption is that $\sigma^2_{\text{Assumed}} = 1$, which implies that $\delta = \sigma_{\text{Assumed}}/\sigma = 1/\sigma$ in equation 3. For logit, $\sigma^2_{\text{Assumed}} = \pi^2/3$, which implies that $\delta = \pi/\sqrt{3}\sigma$. As illustrated in figures 1 and 2, multiplying y^* by δ rescales the β coefficients while $\Pr(y = 1 \mid x_1, x_2)$ is unaffected. We cannot estimate the βs in equation 1 since σ is unknown, but we can estimate the re-scaled γs in equation 4 since the value of the variance is assumed. The effect of the assumed value of the error variance is seen when you compare results from logit and probit. The estimated coefficients for logit are approximately $\pi/\sqrt{3}$ times larger than those from probit, while the predicted probabilities are nearly identical. The probabilities are not exactly the same and the coefficients are not exactly $\pi/\sqrt{3}$ larger in logit since the shapes of the logistic and normal distributions are slightly different (see Long 1997, 47-50).

The scalar identification of the regression coefficients led Allison (1999) to conclude: "Unless we are willing to assume that the [error] variance is constant across groups, the standard tests for cross-group differences in the [γ] coefficients tell us nothing about differences in the [β] coefficients." To understand why identification affects tests of the equality of coefficients, consider the equations for y^*:

$$\text{Group 0:} \quad y^* = \beta_0^0 + \beta_1^0 x_1 + \beta_2^0 x_2 + \varepsilon_0 \text{ where } Var\left(\varepsilon_0\right) = \sigma_0^2 \tag{6}$$

$$\text{Group 1:} \quad y^* = \beta_0^1 + \beta_1^1 x_1 + \beta_2^1 x_2 + \varepsilon_1 \text{ where } Var\left(\varepsilon_1\right) = \sigma_1^2 \tag{7}$$

Since we cannot estimate the error variances, we assume $\sigma_g^2 = 1$ for probit or $\sigma_g^2 = \pi^2/3$ for logit. This is done by multiplying equation 6 by $\delta_0 = \sigma_{\text{Assumed}}/\sigma_0$ and equation 7 by $\delta_1 = \sigma_{\text{Assumed}}/\sigma_1$:

$$\begin{aligned}
\text{Group 0:} \quad \delta_0 y^* &= \left(\delta_0\beta_0^0\right) + \left(\delta_0\beta_1^0\right) x_1 + \left(\delta_0\beta_2^0\right) x_2 + \delta_0\varepsilon_0 \text{ where } Var\left(\delta_0\varepsilon_0\right) = \sigma^2 \\
\text{Group 1:} \quad \delta_1 y^* &= \left(\delta_1\beta_0^1\right) + \left(\delta_1\beta_1^1\right) x_1 + \left(\delta_1\beta_2^1\right) x_2 + \delta_1\varepsilon_1 \text{ where } Var\left(\delta_1\varepsilon_1\right) = \sigma^2
\end{aligned}$$

Since δ_0 and δ_1 cannot be estimated, we rewrite the equations in terms of the γs which can be estimated:

$$\text{Group 0:}\quad \widetilde{y}_0^* = \gamma_0^0 + \gamma_1^0 x_1 + \gamma_2^0 x_2 + \widetilde{\varepsilon}_0 \text{ where } \mathit{Var}\,(\widetilde{\varepsilon}) = \sigma^2 \tag{8}$$

$$\text{Group 1:}\quad \widetilde{y}_1^* = \gamma_0^1 + \gamma_1^1 x_1 + \gamma_2^1 x_2 + \widetilde{\varepsilon}_1 \text{ where } \mathit{Var}\,(\widetilde{\varepsilon}) = \sigma^2 \tag{9}$$

After fitting the model, we can test H_{γ_k}: $\gamma_k^0 = \gamma_k^1$ which is equivalent to testing H_{γ_k}: $\delta^0 \beta_k^0 = \delta^1 \beta_k^1$. However, we want to test H_{β_k}: $\beta_k^0 = \beta_k^1$ which requires knowing the relative size of the error variances in the two groups. The test proposed by Allison (1999) obtains this information by *assuming* that $\beta_j^0 = \beta_j^1$ for at least one regressor. For example, if we assume that $\beta_j^0 = \beta_j^1$, then

$$\frac{\gamma_j^0}{\gamma_j^1} = \frac{\delta_0 \beta_j^0}{\delta_1 \beta_j^1} = \frac{(\sigma_{\text{Assumed}}/\sigma_0)\,\beta_j^0}{(\sigma_{\text{Assumed}}/\sigma_1)\,\beta_j^1} = \frac{\sigma_1}{\sigma_0} \tag{10}$$

is the relative magnitudes of the σ_gs. As illustrated in section 4.3.2, the results of the test for H_{β_k}: $\beta_k^0 = \beta_k^1$ depend on which β_js are assumed to be equal across groups, sometimes leading to opposite conclusions.

Tests of the equality of probabilities or marginal effects on the probability do not require additional assumptions since identical predictions are obtained using the βs from equations 6 and 7:

$$\begin{aligned}
\text{Group 0:}\quad \Pr_0(y=1 \mid x_1, x_2) &= \Pr_0(\varepsilon \le \beta_0^0 + \beta_1^0 x_1 + \beta_2^0 x_2 \mid x_1, x_2) \\
\text{Group 1:}\quad \Pr_1(y=1 \mid x_1, x_2) &= \Pr_1(\varepsilon \le \beta_0^1 + \beta_1^1 x_1 + \beta_2^1 x_2 \mid x_1, x_2)
\end{aligned}$$

or the γs from equations 8 and 9:

$$\begin{aligned}
\text{Group 0:}\quad \Pr_0(y=1 \mid x_1, x_2) &= \Pr_0(\widetilde{\varepsilon} \le \gamma_0^0 + \gamma_1^0 x_1 + \gamma_2^0 x_2 \mid x_1, x_2) \\
\text{Group 1:}\quad \Pr_1(y=1 \mid x_1, x_2) &= \Pr_1(\widetilde{\varepsilon} \le \gamma_0^1 + \gamma_1^1 x_1 + \gamma_2^1 x_2 \mid x_1, x_2)
\end{aligned}$$

While comparing groups using probabilities and marginal effects on probabilities does not required untestable, identification assumptions needed to test regression coefficients, this is far from the only advantage. Conclusions about the equality of regression coefficients in logit and probit are generally less useful than conclusions in the natural metric of probabilities, as they do not represent the substantive size of the effect. For example, knowing whether the effect of obesity on the probability of diabetes is the same for for whites and nonwhites is more useful than knowing if the regression coefficients, which in logit are in the metric of the log odds of diabetes, are the same.

3 Using probabilities to compare groups

Differences in the probability of the outcome and differences in the marginal effects of regressors on the probability emphasize different ways in which groups can differ. Probabilities show how outcomes differ under specific conditions. For example, is diabetes more prevalent for obese men who are white than those with similar characteristics who are nonwhite? This is illustrated across a range of ages by the two probability curves in figure 3 where age and age-squared are included as regressors (discussed further in section 4). The two-headed vertical arrow compares the probability of diabetes for whites and nonwhites who are 75 years old. Marginal effects examine whether a regressor has the same effect on the probability of the outcome for both groups. For example, does obesity have the same health cost for whites as it does for nonwhites? The arrows show the change in probability as age increases from 55 to 60 for each group. While group differences in probabilities and group differences in marginal effects on probabilities are related, you cannot draw conclusions about one from the other. For example, being obese could lead to a larger increase in the probability of diabetes for whites than nonwhites even though the probability of diabetes is greater for nonwhites than whites.

— Figure 3 here —

The next three subsections present methods for testing group differences in probabilities and marginal effects. The following notation is used. The vector $\mathbf{x}$ contains K regressors with the regression coefficients for group g in the vector $\boldsymbol{\gamma}^g$. We use γs rather than βs since predictions are made from the parameters that are estimated after identification assumptions have been made. We replace $\Pr_g(y{=}1\,|\,\mathbf{x})$ with the more compact notation $\pi(\mathbf{x}, g)$:

$$\text{Group 0:} \quad \pi(\mathbf{x}, g = 0) = F(\mathbf{x}'\boldsymbol{\gamma}^0) \tag{11}$$

$$\text{Group 1:} \quad \pi(\mathbf{x}, g = 1) = F(\mathbf{x}'\boldsymbol{\gamma}^1) \tag{12}$$

where F is the normal cumulative density function for the probit model and the logistic cumulative density function for the logit model. Although estimating models separately by group is conceptually appealing, we fit a single model for both groups which makes post-estimation computations simpler and is necessary for obtaining the correct standard errors when using a complex sampling design (West et al., 2008):

$$\pi\left(\mathbf{x}, g\right) = F\left(\left[g \times \mathbf{x}'\boldsymbol{\gamma}^1\right] + \left[(1{-}g) \times \mathbf{x}'\boldsymbol{\gamma}^0\right]\right) \tag{13}$$

In this model, $\pi(\mathbf{x}, g = 0) = F(0 + \mathbf{x}'\boldsymbol{\gamma}^0)$ and $\pi(\mathbf{x}, g = 1) = F(\mathbf{x}'\boldsymbol{\gamma}^1 + 0)$. While the same regressors are typically included for both groups, a regressor can be eliminated for one group by constraining $\gamma_k^g = 0$. Standard errors for predicted probabilities and marginal effects on probabilities are computed with the delta method (Agresti 2013, 72-77; Bishop et al. 1975, 486-497).

3.1 Group comparisons of probabilities

The most basic way to compare groups is to estimate probabilities at the same values of the regressors and test if the predictions are equal. Let $\mathbf{x}^*$ contain specific values of the xs. The difference between groups 0 and 1 in the probability at $\mathbf{x} = \mathbf{x}^*$ is the *group difference* in π:

$$\frac{\Delta\pi(\mathbf{x} = \mathbf{x}^*)}{\Delta\, g} = \pi(\mathbf{x} = \mathbf{x}^*, g{=}1) - \pi(\mathbf{x} = \mathbf{x}^*, g{=}0) \tag{14}$$

To test $H_0\colon \pi(\mathbf{x}{=}\mathbf{x}^*, g{=}0) = \pi(\mathbf{x}{=}\mathbf{x}^*, g{=}1)$, we can test if $\Delta\pi(\mathbf{x} = \mathbf{x}^*)/\Delta\, g$ is 0. Note that the group difference is simply the discrete change with respect to group.

Group differences in probabilities can be used in a variety of ways. At the most basic level, we can test whether whites and nonwhites differ in the probability of diabetes. Adding a layer of complexity, we can test, for example, whether forty-year-old white men have the same probability of diabetes as forty-year-old nonwhite men. Comparisons at multiple values of one or more regressors can be presented in tables. For example, racial differences in diabetes could be shown for men and women at different levels of education. For continuous regressors, plots are often more effective. For example, do nonwhites and whites differ in the probability of diabetes as they age? These methods are illustrated in sections 4.1 and 4.2.

Group differences in conditional probabilities can be measured in other ways. For example, in medicine and epidemiology, the relative risk ratio, also called the adjusted risk ratio, is often used (Bender and Kuss 2010, Greenland 1987, Norton et al. 2013). Instead of computing the *difference* in conditional probabilities across groups using equation 14, the relative risk *ratio* is the ratio of probabilities which indicate the risk of the event:

$$\text{RRR} = \frac{\pi(\mathbf{x} = \mathbf{x}^*, g{=}1)}{\pi(\mathbf{x} = \mathbf{x}^*, g{=}0)}$$

If the groups have the same risk, then the RRR equals

$$1whichcorrespondstoagroupdifferenceof$$

0. While we find group differences in conditional probabilities to be the most useful way to compare groups, our approach can be modified to use the relative risk ratio or any other

measure of group differences in the probability of the outcome.

3.2 Group comparisons of marginal effects

The marginal effect of x_k is the change in the probability of the outcome for a change in x_k, holding other variables at specific values. There are two varieties of marginal effects. A *marginal change*, sometimes called a partial change, is the change in the probability for an infinitely small change in x_k. A *discrete change* or *first difference* is the change in the probability for a discrete or finite change in x_k. In the following discussion, we focus on discrete changes since we find them to be more useful substantively, but our methods can also be used with marginal changes. The critical idea is that one variable is changing while other variables are not.

For group g, the discrete change with respect to x_k is the change in the probability as x_k changes from *start* to *end* while holding other variables at specific values:

$$\frac{\Delta\pi(\mathbf{x}{=}\mathbf{x}^*, g)}{\Delta x_k(\mathit{start} \rightarrow \mathit{end})} = \pi(x_k{=}\mathit{end}, \mathbf{x}{=}\mathbf{x}^*, g) - \pi(x_k{=}\mathit{start}, \mathbf{x}{=}\mathbf{x}^*, g) \tag{15}$$

Vector $\mathbf{x}^*$ contains values for all regressors except x_k whose value is determined by *start* and *end*. If the regressors includes polynomials or interactions, these variables change in tandem. For example, if $x_{\text{agesq}}{=}x_{\text{age}}{\times}x_{\text{age}}$, then x_{agesq} must change from 100 to 121 when x_{age} changes from 10 to 11.

To compare effects across groups, the discrete change of x_k is estimated for each group and we test if the effects are equal:

$$H_0\colon \frac{\Delta\pi(\mathbf{x}{=}\mathbf{x}^*, g{=}1)}{\Delta x_k(\mathit{start} \rightarrow \mathit{end})} = \frac{\Delta\pi(\mathbf{x}{=}\mathbf{x}^*, g{=}0)}{\Delta x_k(\mathit{start} \rightarrow \mathit{end})} \tag{16}$$

Equivalently, we can estimate the *group difference in discrete changes* with respect to x_k, which is the second difference

$$\frac{\Delta^2\pi(\mathbf{x}{=}\mathbf{x}^*)}{\Delta x_k(\mathit{start} \rightarrow \mathit{end})\,\Delta g} = \frac{\Delta\pi(\mathbf{x}{=}\mathbf{x}^*, g{=}1)}{\Delta x_k(\mathit{start} \rightarrow \mathit{end})} - \frac{\Delta\pi(\mathbf{x}{=}\mathbf{x}^*, g{=}0)}{\Delta x_k(\mathit{start} \rightarrow \mathit{end})} \tag{17}$$

The hypothesis that the effect of x_k is the same for both groups is

$$H_0\colon \frac{\Delta^2\pi(\mathbf{x}{=}\mathbf{x}^*)}{\Delta x_k(\mathit{start} \rightarrow \mathit{end})\,\Delta g} = 0 \tag{18}$$

Since the value of the discrete change of x_k depends on the values of the regressors where the change is estimated (Long and Freese, 2006, 244-246), a critical decision is how to summarize the effect. Two approaches are commonly used. First, the discrete change is estimated at representative values of the xs, referred to as a discrete change at representative values (DCR). When means are used as the representative values, the effect is called the discrete change at the mean (DCM). Second, the average discrete change (ADC) is the average of the discrete changes computed conditionally on the observed values of the xs for each observation. DCRs and ADCs highlight different ways in which groups can differ and the choice of which measure to use depends on your substantive question. This issue is discussed in section 3.5 after we formally define these measures of discrete change.

3.3 Discrete change at representative values (DCR)

A DCR is computed at values of the regressors that represent some aspect of the sample that is of substantive interest. For group g the discrete change of x_k evaluated at $\mathbf{x}=\mathbf{x}^*$ is

$$\frac{\Delta\pi(\mathbf{x}=\mathbf{x}^*,g)}{\Delta x_k(\mathit{start}\rightarrow\mathit{end})}=\pi(x_k=\mathit{end},\mathbf{x}=\mathbf{x}^*,g)-\pi(x_k=\mathit{start},\mathbf{x}=\mathbf{x}^*,g)$$

For a continuous variable we can compute the effect of changing x_k from any starting value to any ending value. For example, we could increase x_k from its mean to the mean plus one standard deviation holding other variables at their means, which is the discrete change at the mean (DCM). To compare effects across groups we estimate group differences in DCMs using equation 17:

$$\frac{\Delta^2\pi(\mathbf{x}=\overline{\mathbf{x}})}{\Delta x_k(\overline{x}_k\rightarrow\overline{x}_k+s_k)\,\Delta g}=\frac{\Delta\pi(\mathbf{x}=\overline{\mathbf{x}},g=1)}{\Delta x_k(\overline{x}_k\rightarrow\overline{x}_k+s_k)}-\frac{\Delta\pi(\mathbf{x}=\overline{\mathbf{x}},g=0)}{\Delta x_k(\overline{x}_k\rightarrow\overline{x}_k+s_k)}$$

If x_k is binary, the group difference in the effect of x_k when $\mathbf{x}=\mathbf{x}^*$ is:

$$\frac{\Delta^2\pi(\mathbf{x}=\mathbf{x}^*)}{\Delta x_k(0\rightarrow 1)\,\Delta g}=\frac{\Delta\pi(\mathbf{x}=\mathbf{x}^*,g=1)}{\Delta x_k(0\rightarrow 1)}-\frac{\Delta\pi(\mathbf{x}=\mathbf{x}^*,g=0)}{\Delta x_k(0\rightarrow 1)}$$

To test if the effects are the same in both groups, we test if the group difference in effects is 0. Since DCRs compare the effect of a variable at the *same* values of the regressors for both groups, they do *not* reflect group differences in the distribution of the regressors. This important point is discussed in detail after we consider the ADC.

3.4 Average discrete change (ADC)

The average discrete change of x_k is the average of the discrete change of x_k computed for each observation in a given group using the observed values of the covariates. Let $\pi(x_{ik},\mathbf{x}_i,g)$ be the probability at the observed values for the i^{th} observation in group g, noting in particular the value of x_k. For observation i in group g, the discrete change of x_k is

$$\frac{\Delta\pi(\mathbf{x}{=}\mathbf{x}_i, g)}{\Delta x_k(start_i \to end_i)} = \pi(x_k{=}end_i, \mathbf{x}{=}\mathbf{x}_i, g) - \pi(x_k{=}start_i, \mathbf{x}{=}\mathbf{x}_i, g)$$

The start and end values can be defined in a variety of ways. For a continuous variable we might compute the effect when x_k increases by δ from its observed value x_{ik}:

$$\frac{\Delta\pi(\mathbf{x}{=}\mathbf{x}_i, g)}{\Delta x_k(x_{ik} \to x_{ik}{+}\delta)} = \pi(x_k{=}x_{ik}{+}\delta, \mathbf{x}{=}\mathbf{x}_i, g) - \pi(x_k{=}x_{ik}, \mathbf{x}{=}\mathbf{x}_i, g)$$

While δ is often 1 or a standard deviation, other values can be used. It is also possible to change x_k between the same two values for every observation, such as increasing age from 60 to 65 or changing a binary variable from 0 to 1. For group g,

$$\frac{\Delta\pi(\mathbf{x} = \mathbf{x}_i, g)}{\Delta x_k(start \to end)} = \pi(x_k{=}end, \mathbf{x}{=}\mathbf{x}_i, g) - \pi(x_k{=}start, \mathbf{x}{=}\mathbf{x}_i, g)$$

In this equation, *start* and *end* do not need the subscript i since they have the same values for all observations. When x_k is binary, we use the simpler notation $\Delta\pi(\mathbf{x}{=}\mathbf{x}_i, g)/\Delta x_k$. For example, the effect of being female is written as $\Delta\pi(\mathbf{x}{=}\mathbf{x}_i, g)/\Delta female$. The ADC for x_k in group g is the average of the discrete changes for each observation in the group:

$$\text{ADC}^g_{x_k} = \frac{1}{N_g}\sum_{i\in g}\frac{\Delta\pi(\mathbf{x}{=}\mathbf{x}_i, g)}{\Delta x_k(start_i \to end_i)}$$

Equations 16-18 are used to test if group differences in the ADC are significant.

3.5 Should you compare ADCs or DCRs?

— Figure 4 here —

The choice of whether to make group comparisons of ADCs or DCRs depends on the substantive question being asked. To illustrate what each measure of change tells you, figure 4 plots the probability of diabetes by age for whites and nonwhites from a model with age and age-squared. The squares are observations for the younger sample of whites, while the circles are observations for the older sample of nonwhites. For whites, $\text{ADC}^W_{age}{=}.03$ which is the average change in the probability for each observations as age is increases by 5 from its observed values. For nonwhites, ADC^N_{age} is close to 0 since the positive effect of age for those younger than 72 is offset by the negative effect for those older than 72. The difference in the

ADCs across groups is due primarily to group differences in the distribution of age and mask the similar shapes of the probability curves. In contrast, the DCRs computed at specific ages reflect the similar shapes of the curves. At the mean age of 60 for whites, $\mathrm{DCR}^{W}_{age} = .03$ and $\mathrm{DCR}^{N}_{age} = .04$; at the overall mean age of 67.5, $\mathrm{DCR}^{W}_{age} = .01$ and $\mathrm{DCR}^{N}_{age} = .02$; and at the mean age of 75 for nonwhites, $\mathrm{DCR}^{W}_{age} = -.01$ and $\mathrm{DCR}^{N}_{age} = -.02$. DCRs compare the shape of the probability curves at the same values of the regressors, whiles ADCs reflect both group differences in the curves and in the distribution of regressors. In this example, the ADCs suggest that the effect of age is larger for whites, yet the curves show that the rate of change in the probability of diabetes is larger for nonwhites. Indeed, two groups can have exactly the same regression coefficients with significantly different ADCs. Neither the ADC or the DCR is always better—they simply reflect different ways in which effects differ across groups as illustrated in section 4.

4 Example: Racial differences in health

The analyses we use to illustrate these techniques are based on research related to racial and ethnic differences, hereafter referred to as racial differences, in diabetes risk and self-rated health. The literatures on diabetes and self-reported health find strong, incontrovertible racial differences in both outcomes with some evidence that these differences decline with age (Hummer et al. 2004, Markides et al. 1997). Further, it is well known that obesity and physical activity are related to the risk for diabetes, but it is less clear if these variables affect racial differences. There is some evidence that racial disparities in diabetes are higher for normal and overweight individuals and lower among the obese, but few studies have tested whether physical activity benefits one group more than another (Zhang et al., 2009). Thus, we use differences in probabilities to examine racial differences in diabetes and health by age. Then we compute differences in DCRs and ADCs to examine racial disparities in the effects of aging on diabetes, differences in the impact of physical activity on the probability of diabetes, and whether racial differences in the effect of physical activity vary over the middle-to-late life course.

Our example uses data from the Health and Retirement Study (HRS), a representative sample of older adults in the US (Health and Retirement Study, 2006).[2] Approximately 22,000 individuals and their spouses were interviewed about every other year since 1992 using a multistage, clustered probability sampling design that represents non-institutionalized

[2] The Health and Retirement Study is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. To access these data you must register with HRS. Our analyses used the 2006 Fat file and version N of the RAND HRS file. The versions of these files that are currently available for download could differ from those we used.

individuals age 50 or over in the 48 contiguous states, with an over-sampling of black and hispanic Americans. Data from the 2006 wave of the HRS were extracted from the RAND HRS data files (RAND, 2014). From the 16,955 respondents who had non- zero sampling weights, we excluded 10 respondents who were too young to be in the sample, 380 who did not identify as white, black, or hispanic, 7 with incomes greater than two million, 246 who did not report body mass, and 86 with missing data for other variables in our analyses. The resulting sample includes 16,226 observations. Models were fit using Stata 14.2 with adjustments for complex sampling (StataCorp, 2015). Two-tailed tests are used in all analyses. Post-estimation analyses were computed using the `margins` command along with the `SPost13` package (Long and Freese, 2014). Similar features are being developed in R (Leeper, 2016). The do-files to construct our dataset and to reproduce our analyses can be obtained by running the command `search groupsbrm` in Stata and following the instructions for downloading files. Then, type `help groupsbrm` for details on creating the dataset.

— Table 1 here —

Table 1 contains descriptive statistics for the variables in our models. Our group variable is the race of the respondent, comparing whites to nonwhites who include blacks and those of Hispanic ethnicity. Other racial and ethnic groups were excluded due to inadequate representation in the HRS sample. Two outcome variables are used. Self-rated health recoded responses to "Would you say your health is excellent, very good, good, fair, or poor?" to equal 1 if health was good, very good, or excellent, else 0. Diabetes is a respondent's self-report of whether diabetes was diagnosed by a physician. Independent variables include age, gender, education, income, marital status, obesity, and physical activity. Education is measured as having a high school degree or higher compared to not completing high school. Income is in thousands of dollars with an inverse hyperbolic sine transformation used to reduce the skew (Burbidge et al., 1988). Physical activity is measured as exercising more than three times a month compared to exercising less often. Following guidelines by the US Centers for Disease Control and Prevention (Pi-Sunyer et al. 1998), obesity is defined as having a body mass index of 30 or more.

— Tables 2 and 3 here —

Tables 2 and 3 contain estimates of the γ_k^g coefficients from two models. We chose these outcomes and specifications to illustrate how our methods can be used both when the relationship between a regressor and the outcome is nearly linear and when a more complex nonlinear relationship exists. The model for good health is easy to interpret since the relationship between age and health is nearly linear and the model does not include interactions among the regressors. The model for diabetes includes age-squared and an interaction between age and activity which makes the interpretation more challenging, but is a more realistic modeling scenario. Recall that one of our research questions is whether there are racial differences in the probability of diabetes at different ages (e.g., are racial difference larger at younger ages than older ages), whether there are racial differences in the effects of growing older on diabetes, and whether racial differences in the effects of growing older vary by the level of physical activity.

Our analyses begin with graphs to explore group differences in the outcomes by age. The discrete change with respect to race, referred to simply as the racial difference, is used to test if differences in the outcomes are statistically significant. Next, the discrete change with respect to age is used to test if the effects of age differ by race. Then we use graphs to examine the more complex effects of age on diabetes for those with different levels of activity. Tables are used to examine racial differences in the effects of gender and obesity on diabetes, where discrete changes for gender and obesity are compared across race using racial differences in the discrete changes for these regressors. Finally, racial differences in scalar measures of the effect of a regressor are considered using discrete changes and regression coefficients.

4.1 Graphs of probabilities and group differences in probabilities

Graphs show both group differences in predictions across values of a regressor and differences in the effects of a regressor by group. To explain this method of interpretation, we begin with a simple example showing how whites and nonwhites differ in reporting good health at different ages. We then extend this approach to a more complicated model for diabetes that includes age and age-squared along with interactions between age and a respondent's level of physical activity.

We know from table 1 that on average whites report better health than nonwhites and now want to consider whether racial disparities in health change with age. For each group probabilities are computed at ages from 50 to 90 with other variables held at their means. Figure 5 shows that while whites have a higher probability of reporting good health at all ages, differences steadily decrease from .10 at age 50 to less than .04 at 90. Racial differences at each age are used to test if these differences in probabilities are significant (see equation 14):

$$\frac{\Delta\pi(age{=}p, \mathbf{x}{=}\overline{\mathbf{x}})}{\Delta\, white}$$

Figure 6 plots these differences in the probability of good health along with the 95% confidence interval. When the confidence interval crosses 0, as it does around 85, the difference between whites and nonwhites in the probability of good health is not significant.

— Figures 5 and 6 here —

Since the probabilities curves in figure 5 are nearly linear, the effects of age can be summarized by the discrete change in the probability of good health as age increase from 50 to 90 for each group g (see equation 15):

$$\frac{\Delta\pi(\mathit{white}{=}g, \mathbf{x}{=}\overline{\mathbf{x}})}{\Delta \mathit{age}(50 \to 90)} = \pi(\mathit{age}{=}90, \mathbf{x}{=}\overline{\mathbf{x}}, \mathit{white}{=}g) - \pi(\mathit{age}{=}50, \mathbf{x}{=}\overline{\mathbf{x}}, \mathit{white}{=}g)$$

Group differences in the effect of age are computed as a second difference (see equation 17):

$$\frac{\Delta\pi(\mathbf{x}{=}\overline{\mathbf{x}})}{\Delta \mathit{age}(50 \to 90)\, \Delta \mathit{white}} = \frac{\Delta\pi(\mathbf{x}{=}\overline{\mathbf{x}}, \mathit{white}{=}1)}{\Delta \mathit{age}(50 \to 90)} - \frac{\Delta\pi(\mathbf{x}{=}\overline{\mathbf{x}}, \mathit{white}{=}0)}{\Delta \mathit{age}(50 \to 90)} \tag{19}$$

Using these measures, we find that as age increases from 50 to 90 the probability of good health decreases more rapidly for whites (.13) than than nonwhites (.07), but the difference is not significant (p=.17). While figure 6 shows tests of racial difference in the probability of good health at each age, the discrete changes examine racial difference in the effect of age on the probability of good health.

To illustrate how graphs can be used to examine more complex differences between groups, figure 7 plots the probability of diabetes by age, where the bell shaped curves reflect the inclusion of age and age-squared in the model. For both groups the probability of diabetes increases from age 50 to 75 after which the probability decreases. While whites have a smaller probability of diabetes at all ages, the difference is smallest at 50 where it is about .04, increases to a maximum of .12 at 75, and then decreases to .08 at 90. This pattern of the racial differences is plotted in figure 8 which shows how differences increase from age 50 to 75 followed by a gradual decrease. Racial differences are significant at all ages except 90 where the confidence interval includes 0.

— Figures 7 and 8 here —

The change in the size of the effect of race over age occurs because the rate of increase in diabetes is larger for nonwhites than whites from ages 50 to 75 at which point the rate of decrease with age is more rapid for nonwhites. To test this formally, we compute the discrete change for age for each group and test if they are equal (see equation 17). From 50 to 60 diabetes increases by .11 for nonwhites compared to .06 for whites, a significant difference of .05 (p=.01). From 80 to 90 the probability decreased by .05 for whites and .09 for nonwhites, a difference that is not significant (p=.27).

— Figure 9 here —

Using graphs to examine group differences over the range of a continuous regressor can be extended to show the effects of other variables. Returning to our research questions, suppose we want to determine whether the racial differences in diabetes that we found in figure 7 vary by a person's level of physical activity. Or, to put it another way, are the benefits of physical activity different for nonwhites and whites over the life course? Answering this question illustrates both the complexity and the advantages of our approach to interpretation. It can be difficult to find the most effective way to visualize how group differences in an outcome over age vary by the levels of another variable, yet doing so allows us to test more complex hypotheses than traditional methods. The first step is to graph the probability of diabetes for whites and nonwhites by level of activity. This is done in figure 9 where open circles represent nonwhites who are inactive with solid circles for those who are active. Similarly, inactive and active whites are represented by open and solid squares. While the graph contains all of the information that we need for our research question, the trends are difficult to see due to the complexity of the graph. A more effective approach is to create plots that show differences between the probability curves. There are two ways that we can proceed that emphasize different aspects of our research question. First, we can examine racial differences in diabetes over age conditional on level of activity by plotting the difference between the probability curves for whites and nonwhites for those who are active (solid circles and squares) and for those who are inactive (hollow circles and squares). Second, we can examine the effects of activity by plotting the discrete change of activity by race, which is the difference between the curves for whites (solid and open squares) and the curves for nonwhites (solid and open circles).

— Figure 10 here —

Figure 10 plots $\Delta\pi(\mathit{age}{=}p, \mathit{active}{=}q, \mathbf{x}{=}\overline{\mathbf{x}})/\Delta \mathit{white}$ (see equation 14), which is the racial difference in the probability of diabetes by level of activity over age. Since adding confidence intervals to the figure leads to overlapping lines that are confusing, a dashed line is used to indicate when a difference is *not* significant. The graph shows that while the benefits of being white occur both for those who have an active lifestyle and those who do not, the strength and timing of the benefits differ by the level of activity. For those who are not active (open diamonds), the advantages for whites increase from age 50 to 70 before decreasing thereafter. Differences are significant at all ages except 90. For those who are active (solid diamonds), the same pattern occurs, but the effects are weaker at younger ages than they are for those who are inactive. The differences increase from age 50 to 80, becoming statistically significant at age 57. At age 80 the differences begin to decrease and are no longer significant.

— Figure 11 here —

Figure 11 re-expresses the information from figure 9 to focus on the effects of activity for each group. While being active benefits members of both groups, the benefits occur differently for whites and nonwhites. For whites (open triangles) the protective effect of activity is smaller (i.e., less negative) at younger ages and increases in magnitude until age 90. For nonwhites (solid triangles), the effect gets stronger from age 50 to 60 before decreasing till age 90; after age 76 the effects are not significant. Tests of racial differences in the effect of activity are significant at the .10 level between ages 55 and 61, reach significance at the .05 level at age 58 where the difference reaches its maximum of .044, and are not significant at other ages.

Finally, another way to think of the effects of race and activity is to note that the health deficit for being nonwhite is roughly equal to the benefits of being active. This is seen in figure 9 by comparing the line for inactive whites (hollow squares) and active nonwhites (solid circles). The probabilities differ by -.01 at age 50 with a maximum of .05 at age 75, but none of the differences are significant.

4.2 Tables of probabilities and group differences in probabilities

— Table 4 here —

Tables are an effective way to show how probabilities vary over categories of a few regressors. Suppose that we are interested in whether racial differences in diabetes vary by gender and obesity, with a focus on the adverse effects of obesity. One way to approach this is to compute the probability of diabetes conditional on all combinations of race, gender, and obesity, holding other variables at their means. In table 4 these probabilities are presented in rows 1 and 2 of columns 1 through 4. The last row shows racial differences in the probabilities of diabetes (see equation 14):

$$\frac{\Delta\pi(\mathit{female}=p, \mathit{obese}=q, \mathbf{x}=\overline{\mathbf{x}})}{\Delta\, \mathit{white}}$$

To simplify notation, in the rest of this section we exclude $\mathbf{x}=\overline{\mathbf{x}}$ from $\pi()$. While we are holding other variables at the mean, these variables could be held at other values.

The probabilities of whites being diagnosed with diabetes are smaller than those for nonwhites for all combinations of obesity and gender. The largest racial differences, shown in row 3, is $-.126$ for women who are not obese and the smallest difference is a nonsignificant $-.044$ for obese men. We can test whether the racial differences for men and women are equal for a given level of obesity by estimating a second differences (see equation 17):

$$\frac{\Delta\pi(obese = q)}{\Delta female\ \Delta white} = \frac{\Delta\pi(female = 1, obese = q)}{\Delta\ white} - \frac{\Delta\pi(female = 0, obese = q)}{\Delta\ white}$$

Computing these differences, which are not included in the table, we find that the effect of race for obese men and women differ by $-.068 = (-.112 - -.044)$ which is significant at the .02 level. For those who are not obese, the gender difference is smaller and not significant ($p = .13$).

Next, we consider the effects of obesity on diabetes. The probabilities in rows 1 and 2 of the first four columns show that being obese is associated with a higher incidence of diabetes. To formalize these findings, we estimate the discrete change of obesity conditional on gender and race holding other variables at their means (see equation 15):

$$\frac{\Delta\pi(female = p, white = r)}{\Delta\ obese}$$

These effects, presented in rows 1 and 2 of columns 5 and 6, show that obesity significantly increases the probability of diabetes by about .16 for all groups except white men where the effect is .21. To test if there are racial differences in the effects of obesity, we estimate second differences (see equation 17):

$$\frac{\Delta\pi(female = p)}{\Delta obese\ \Delta white} = \frac{\Delta\pi(female = p, white = 1)}{\Delta\ obese} - \frac{\Delta\pi(female = p, white = 0)}{\Delta\ obese}$$

The results, shown in the last row of columns 5 and 6, indicate that racial differences in the effects of obesity are small and not significant for women, but larger and marginally significant for men ($p = .09$). To test whether the effect of obesity is the same for men and women, we estimate the second differences with respect to gender, which are shown in rows 1 and 2 of column 7:

$$\frac{\Delta\pi(white = r)}{\Delta\ obese\ \Delta\ female} = \frac{\Delta\pi(female = 1, white = r)}{\Delta\ obese} - \frac{\Delta\pi(female = 0, white = r)}{\Delta\ obese}$$

The effect of obesity for whites is .04 larger ($p < .001$) for men than women, but the gender difference is small and nonsignificant for nonwhite respondents.

The idea of a second difference can be extended to compare any two effects, such as whether obesity has the same effect for white men and nonwhite women:

$$H_0\colon \frac{\Delta\pi(\mathit{female}=0, \mathit{white}=1)}{\Delta\ \mathit{obese}} = \frac{\Delta\pi(\mathit{female}=0, \mathit{white}=0)}{\Delta\ \mathit{obese}}$$

or whether the effects of obesity are the same for all combinations of gender and race:

$$\begin{aligned} H_0\colon \frac{\Delta\pi(\mathit{female}=0, \mathit{white}=0)}{\Delta\ \mathit{obese}} &= \frac{\Delta\pi(\mathit{female}=0, \mathit{white}=1)}{\Delta\ \mathit{obese}} \\ &= \frac{\Delta\pi(\mathit{female}=1, \mathit{white}=0)}{\Delta\ \mathit{obese}} = \frac{\Delta\pi(\mathit{female}=1, \mathit{white}=1)}{\Delta\ \mathit{obese}} \end{aligned}$$

which is rejected at the .001 level.

In table 4 the effects of obesity were computed for each combination of race and gender holding other variables at their means. While this allows us to make comparisons where only race and gender change, is it reasonable to estimate effects for each group at the overall means when table 1 shows significant racial differences in the distribution of the regressors? An alternative approach that reflects group differences in the distribution of regressors is to compare the ADC of obesity for each group defined by race and gender. ADCs reflect group differences in the regressors, but do not show whether the effects of obesity are similar across groups for individuals with the same characteristics (see section 3.5). Another approach is to compare the effects for each group holding other variables at the means specific to each group (see Long and Freese 2014 for a discussion of local means). The most effective approach for comparing effects depends on the specific questions motivating the research.

4.3 Comparing summary measures of effect

The methods in the last two sections showed how to use predictions and marginal effects to address specific questions motivating one's researcher. In this section we consider methods for summarizing the effect of each regressor across groups. These measures are the counterpart to regression coefficients that are routinely used in linear regression to summarize the effect of each variable, assuming there are no polynomial terms. In logit and probit there are several measures that should be considered and some that should be avoided. We begin by comparing marginal effects across groups using discrete changes, which we believe is the most generally useful approach. Next we examine methods based on the comparison of the regression coefficients and illustrate their limitations.

4.3.1 Comparing marginal effects

— Table 5 here —

The discrete change at the mean (equation 3.3) and the average discrete change (equation 3.4) are standard measures of the effect of a variable.[3] Table 5 contains estimates for both measures along with racial differences in the effects (see equation 17) and p-values from testing whether the effects are equal. Consider the ADC of being female from panel 1 of the table. On average being female significantly decreases the probability of diabetes by .051 ($p<$.001) for white respondents, with a decrease to .015 ($p<$.001) for nonwhites. As shown in columns 5 and 6, these effects differ by .036, which is significant at the .10 level but not the .05 level. The results using the discrete change at the mean from panel B are nearly identical. The ADC and DCM do not always lead to the same conclusions as illustrated by the effect of a five-year increase in age. Using the ADC we conclude that the average effect of age is significantly larger for nonwhites than whites (p=.002), while using the DCM, we conclude that for an average respondent the effect of age does not differ for whites and nonwhites (p=.169). As discussed in section 3.5, conclusions based on ADCs and DCRs can be quite different depending on the model specification and group differences in the distribution of regressors. The "best" measure is the one that characterizes that aspect of groups differences that are of greatest substantive interest. In our experience, it is useful to compute both the DCM and the ADC. If your conclusions differ depending which measure you use, determine why the measures differ before using either and then decide which is most appropriate for your research question.

4.3.2 Comparing regression coefficients

Given the complexities of summarizing the effects of regressors using marginal effects on the probability, tests of the equality of the regression coefficients are appealing in their simplicity. While such tests are often used, there are several issues to consider. First, regression coefficients are in a metric that is not as substantively useful as probabilities. Comparing the effects of x_k on y^* or on the log-odds is rarely as useful as comparing how a regressor affects the probability. Second, while odds ratios are in a more natural metric than the log-odds or y^*, odds ratios can be especially misleading in group comparisons. If the odds ratio for x_k is identical for both groups, the effect of x_k on the probability can be

[3] While marginal change showing the instantaneous rate of change in the probability can also be used, we prefer the discrete change because of its simpler interpretation. If the probability curve is being evaluated in a region that is approximately linear for a unit change in the regressor, the marginal change will approximate a discrete change of one.

very different in the two groups. For example, consider probabilities of the outcome for two groups at two values of regressor x_1:

$$\begin{array}{ll} \pi(x_1=0, g=1) = .2 & \pi(x_1=0, g=0) = .010 \\ \pi(x_1=1, g=1) = .4 & \pi(x_1=1, g=0) = .026 \end{array}$$

For group 1, the discrete change of x_1 is $.2 = .4 - .2$ with an odds ratio for x_1 of $2.67 = (.4/.6)/(.2/.8)$. For group 0, the discrete change is $.016 = .026 - .010$ with an odds ratio of $2.64 = (.026/.974)/(.01/.99)$. Even though x_1 has a far bigger impact on the probability of y for group 1 than group 0 (.2 versus .016), the odds ratios are nearly identical. Third, if the models includes interactions or polynomials, there is no simple way to express the total effect of those variables in using regression coefficients or odds ratios, while marginal effects on the probability can be computed that simultaneously to take into account all of the coefficients that include the variable. For example, the marginal effect of age simultaneously accounts for changes in age and age-squared. Finally, even if these issues are not a concern in your application, you must deal with the scalar identification of regression coefficients, which we referred to earlier . To illustrate this important issue, we consider alternative tests of the hypothesis that the regression coefficients for gender and obesity are equal for whites and nonwhites.

Columns 9 and 10 of table 3 contain results from standard Wald tests of $H_0\colon \gamma_k^{\mathrm{W}} = \gamma_k^{\mathrm{N}}$. If these tests were appropriate, which they are not due to the scalar identification of the coefficients, we would conclude that the protective effects of being female are significantly larger for whites than nonwhites ($p < .01$) and that the health costs of obesity are significantly greater for whites than nonwhites ($p < .01$). This contradicts our conclusions using discrete changes in table 5. The ADC for obesity is .022 larger for whites than nonwhites, but the difference is not significant ($p = .41$); similarly the DCM is .032 larger for whites ($p = .26$). The ADC and DCM for being female are about .04 less negative for nonwhites than whites, but the differences are not significant ($p = .09$; $p = .07$).

Allison (1999) showed that standard tests comparing regression coefficients across groups confound groups differences in the coefficients with groups differences in residual variation (see section 2). He proposed a test of the equality of coefficients across groups that is unaffected by group differences in unobserved heterogeneity. This is accomplished by assuming that the regression coefficients for one or more regressors are equal across groups (see equation 10). While assuming the equality of regression coefficients across groups deals with the problem caused by the scalar identification of the coefficients, the results of the test depend on which variables are assumed to have equal effects. To illustrate this, we test the racial equality of regression coefficients for gender and obesity using three models that differ by which coefficients are assumed to be equal across groups.

— Table 6 here —

In model M_1 the coefficients for *ihsincome* are constrained to be equal; in M_2 all coefficients that include *age* or *active* are constrained (both *age* and *active* are constrained due to the *age* by *active* interaction); and in M_3 the coefficients for *obese* are constrained. Models in which the regression coefficients for being female, graduating from high school, or being married were constrained did not converge. The results are shown in table 6. Racial differences in the regression coefficients for *female* are significant ($p = .013$) when constraints are imposed on coefficients involving either *age* or *active*, but not when constraints are imposed on the coefficients for *ihsincome* ($p{=}.164$) or *obese* ($p{=}.120$). The regression coefficients for *obese* are similar in magnitude and not significantly different ($p{=}.810$) when the coefficients for *ihsincome* are constrained in M_1, but the difference is larger and marginally significant ($p{=}.064$) in M_2 when the coefficients for *age* and *active* are constrained.

Even though conclusions from the tests vary depending on which coefficients are assumed to be equal across groups, when a single pair of coefficients are constrained to be equal, such as M_1 and M_3, the predicted probabilities and marginal effects are *exactly* equal to those from the full model. That is, the models are empirically indistinguishable. Williams (2009) showed that Allison's test can be computed using the heterogeneous choice model, also know as the location scale model. A heterogeneous choice model predicting the variance of the error for each group using a single equality constraint on the coefficients for the regressors is simply a reparameterization of the model without constraints in coefficients that assumes unobserved heterogeneity is equal for both groups. When multiple constraints are imposed as in M_2, the predictions have a correlation of .999 with the full model. Williams (2009) showed that Allison's approach to allow group differences in unobserved heterogeneity can be extended by allowing regressors beyond group membership to predict the variance of the errors. When we attempted Williams' procedure using different sets of covariates, many models did not converge and the interpretation of the coefficients of those that did converge changed based upon the identifying assumptions made.

Allison's approach deals with the residual variation issue by allowing error variances to differ by group but requires assuming that the effects of at least one regressor are the same across groups. While decisions about which coefficients to constrain can affect the substantive conclusions, past research or substantive theory is unlikely to provide insights into which constraints should be used and there is no statistical test to help you decide which coefficients should be assumed to be equal. Making an *ad hoc* decision that some regression coefficients are equal can lead to incorrect conclusions. Indeed, in our example we did not have substantive reasons to constrain particular coefficients and tried multiple specifications that lead to substantively different conclusions. The risk of specifying the inappropriate identifying assumptions does not exist at the level of the probabilities. More importantly from our perspective, tests of the equality of effects across groups are most useful when those effects are measured in the metric of the probability of the outcome.

5 Summary and generalizations

In this paper we have developed methods for comparing groups using predicted probabilities and marginal effects on probabilities for the binary logit and probit models. Since these models are nonlinear, conclusions on whether groups differ in the probability of the outcome or in the effect of a regressor depend on where in the data the groups are compared and how the effects of regressors are summarized. Deciding how to make comparisons requires careful consideration based on a substantive understanding of the process being modeled and the questions being asked. While this is much harder than routine tests of the equality of regression coefficients, the more complex task of comparing predictions and marginal effects provides substantive insights that reflect the complexity of the substantive application.

While our paper focuses on comparing two groups using binary logit and probit, our methods can be used with any regression model where predictions and marginal effects can be estimated along with standard errors. In models with more than two outcome categories, our methods can be applied to the probability of each outcome. In models such as binary logit or ordinal probit, an advantage of our approach is that it deals with the scalar identification of regression coefficients without additional, untestable assumptions. But, even in models where the coefficients are identified, comparing groups using predictions and marginal effects have advantages in any model in which the outcome has a nonlinear relationship with regressors. Even in linear models, whether the outcome is continuous or binary, our methods are useful if nonlinearities, such as quadratics or interaction terms, are included in the model. In these cases, our approach provides insights that go beyond those provided by tests of regression coefficients. For example, if we used a linear probability model for diabetes, we would include age and age-squared as regressors due to the quadratic relationship between age and diabetes. For example, suppose that we have the linear probability model

$$\begin{aligned}\text{Group 0:}\quad & y = \beta_0^0 + \beta_{age}^0 x + \beta_{age^2}^0 age^2 + \beta_z^0 z + \varepsilon_0\\ \text{Group 1:}\quad & y = \beta_0^1 + \beta_{age}^1 x + \beta_{age^2}^1 age^2 + \beta_z^0 z + \varepsilon_0\end{aligned}$$

where z represent any other regressors. Since age and age^2 are regressors, the marginal effect of age depends on β_{age}^g and $\beta_{age^2}^g$ as well as the values of age, age^2 and z. In this case, testing the group equality of marginal effects is more useful than testing the equality of regression coefficients. If age^2 was not in the model, our approach using marginal effects is equivalent to testing if $\beta^0_{age} = \beta^1_{age}$.

Finally, while our examples were relatively simple and we compared only two groups, the same methods can be used with any number of groups and in models with many regressors including higher order interactions. Further, our methods can be extended beyond testing for group differences to interpreting interactions more generally, such as dichotomous by dichotomous interactions, continuous by continuous interactions, and three-way interactions.

References

Agresti, A. 2013. *Categorical Data Analysis. 3rd Edition.* Third edition ed. New York: Wiley.

Ai, C., and E. C. Norton. 2003. Interaction terms in logit and probit models. *Economics letters* 80(1): 123–129.

Allison, P. D. 1999. Comparing logit and probit coefficients across groups. *Sociological Methods & Research* 28(2): 186–208.

Amemiya, T. 1981. Qualitative response models: a survey. *Journal of Economic Literature* 19: 1483–1536.

Bender, R., and O. Kuss. 2010. Methods to calculate relative risks, risk differences, and numbers needed to treat from logistic regression. *Journal of clinical epidemiology* 63(1): 7.

Bishop, Y., S. Fienberg, and P. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Breen, R., A. Holm, and K. B. Karlson. 2014. Correlations and nonlinear probability models. *Sociological Methods & Research* 43(4): 571–605.

Breen, R., and K. B. Karlson. 2013. *Counterfactual Causal Analysis and Nonlinear Probability Models*, 167–187. Dordrecht, Netherlands: Springer.

Buis, M. L. 2010. Stata tip 87: Interpretation of interactions in non-linear models. *The stata journal* 10(2): 305–308.

Burbidge, J. B., L. Magee, and A. L. Robb. 1988. Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association* 83(401): 123–127.

Chow, G. 1960. Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28: 591–605.

Greenland, S. 1987. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 125(5): 761–768.

Health and Retirement Study. 2006. *Public use dataset.* Ann Arbor, MI.: Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740).

Hummer, R. A., M. R. Benjamins, and R. G. Rogers. 2004. *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*, chap. Racial and Ethnic Disparities in Health and Mortality Among the U.S. Elderly Population. Washington, DC: National Academies Press.

Kendler, K. S., and C. O. Gardner. 2010. Interpretation of interactions: guide for the perplexed. *The British Journal of Psychiatry* 197(3): 170–171.

Kuha, J., and C. Mills. 2018. On group comparisons with logistic regression models. *Sociological Methods & Research* 1–28.

Landerman, L. R., S. A. Mustillo, and K. C. Land. 2011. Modeling repeated measures of dichotomous data: testing whether the within-person trajectory of change varies across levels of between-person factors. *Social science research* 40(5): 1456–1464.

Leeper, T. J. 2016. *margins: An R port of Statas margins command.* R package version 0.2.0.

Liao, T. F. 2002. *Statistical Group Comparison*, vol. 29. New York: Wiley.

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*, vol. 7 of *Advanced Quantitative Techniques in the Social Sciences.* Thousand Oaks, CA: Sage.

———. 2005. Group comparisons in nonlinear models using predicted outcomes.

———. 2009. Group comparisons in logit and probit using predicted probabilities.

Long, J. S., and J. Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata. Second Edition.* College Station, Texas: Stata Press.

———. 2014. *Regression Models for Categorical Dependent Variables Using Stata. Third Edition.* College Station, Texas: Stata Press.

Maddala, G. 1983. *Limited-dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Markides, K. S., L. Rudkin, R. J. Angel, and D. V. Espino. 1997. Health status of Hispanic elderly. *Racial and ethnic differences in the health of older Americans* 285–300.

McKelvey, R. D., and W. Zavoina. 1975. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology* 4: 103–120.

Mood, C. 2010. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review* 1–16.

Mustillo, S., L. R. Landerman, and K. C. Land. 2012. Modeling longitudinal count data: Testing for group differences in growth trajectories using average marginal effects. *Sociological Methods & Research* 41(3): 467–487.

Norton, E. C., M. M. Miller, L. C. Kleinman, et al. 2013. Computing adjusted risk ratios and risk differences in Stata. *Stata J* 13(3): 492–509.

Norton, E. C., H. Wang, and C. Ai. 2004. Computing interaction effects and standard errors in logit and probit. *The Stata Journal* 4(2): 154–167.

Pi-Sunyer, F. X., D. M. Becker, C. Bouchard, R. Carleton, G. Colditz, W. Dietz, J. Foreyt, R. Garrison, S. Grundy, B. Hansen, et al. 1998. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults. *American Journal of Clinical Nutrition* 68(4): 899–917.

RAND. 2014. *RAND HRS Data, Version N.* Santa Monica, CA: Produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration.

StataCorp. 2015. *Stata 14 Survey Data Reference Manual.* College Station, TX: Stata Press.

West, B. T., P. Berglund, S. G. Heeringa, et al. 2008. A closer examination of subpopulation analysis of complex-sample survey data. *Stata J* 8(4): 520–531.

Williams, R. 2009. Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups. *Sociological Methods & Research* 37(4): 531–559.

Zhang, Q., Y. Wang, and E. S. Huang. 2009. Changes in racial/ethnic disparities in the prevalence of Type 2 diabetes by obesity level among US adults. *Ethnicity & health* 14(5): 439–457.

6 Figures

Figure 1: The link between $y^* = \beta_0 + \beta_1 x + \varepsilon$ and $\Pr(y\!=\!1 \mid x)$ with $Var(\varepsilon) = \sigma^2$.

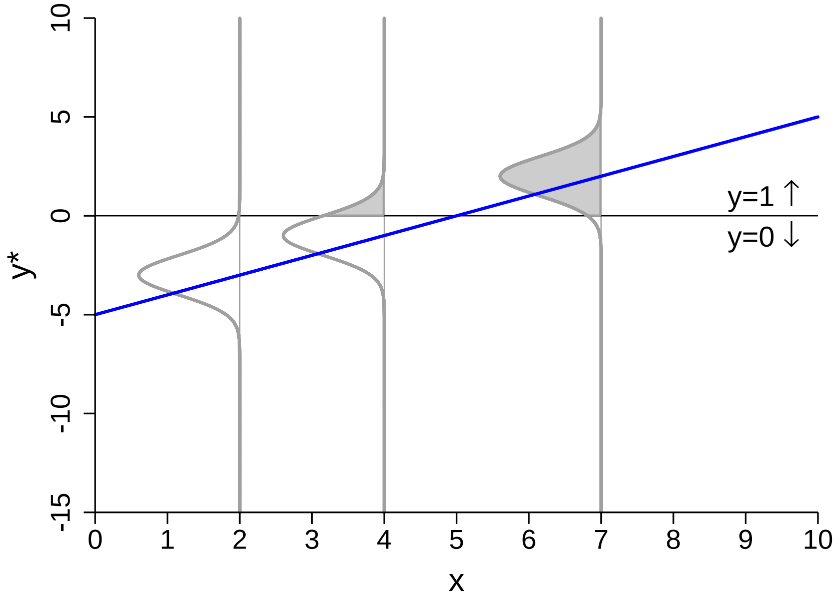


Figure 2: The link between $\delta y^* = (\delta\beta_0) + (\delta\beta_1)x + \delta\varepsilon$ and $\Pr(y\!=\!1 \mid x)$ with $Var(\delta\varepsilon) = \delta^2\sigma^2$.

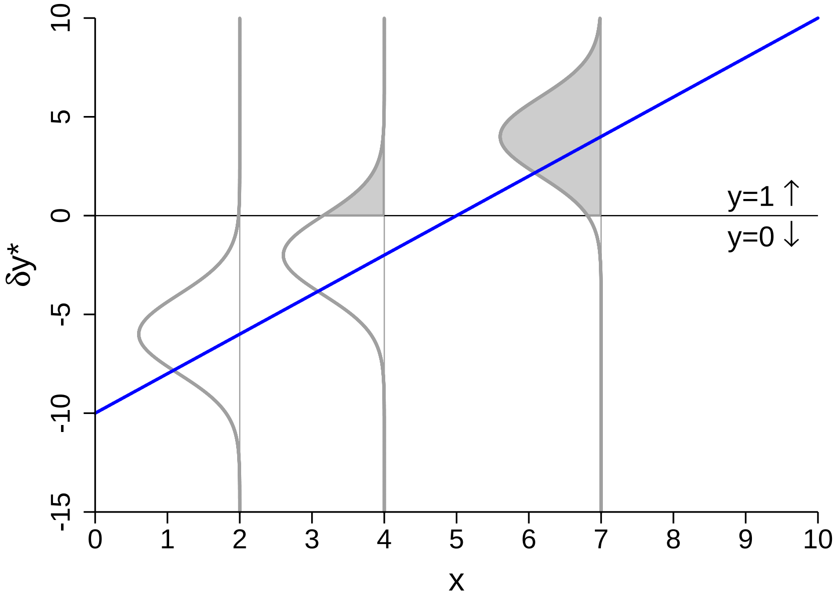


Figure 3: Group comparisons of probabilities and marginal effects.

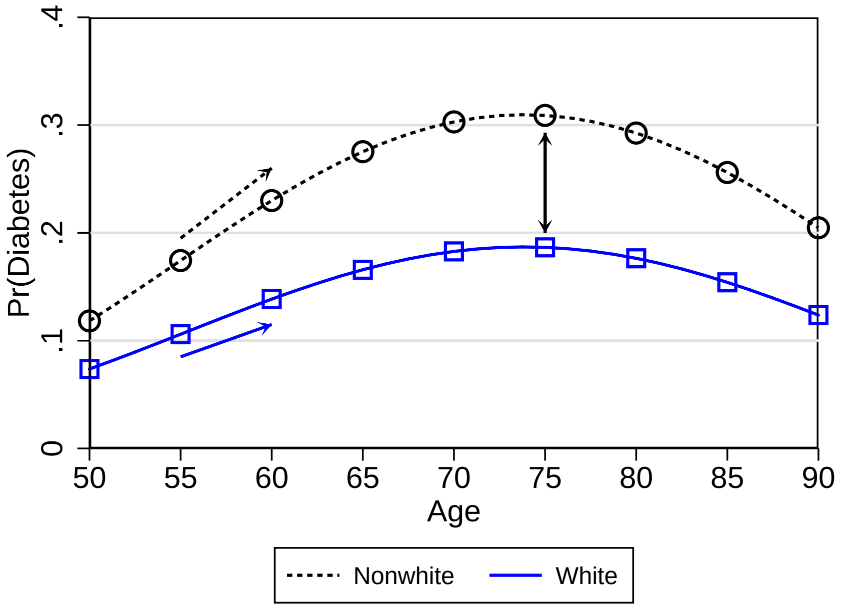


Figure 4: Group differences in the discrete change of age.

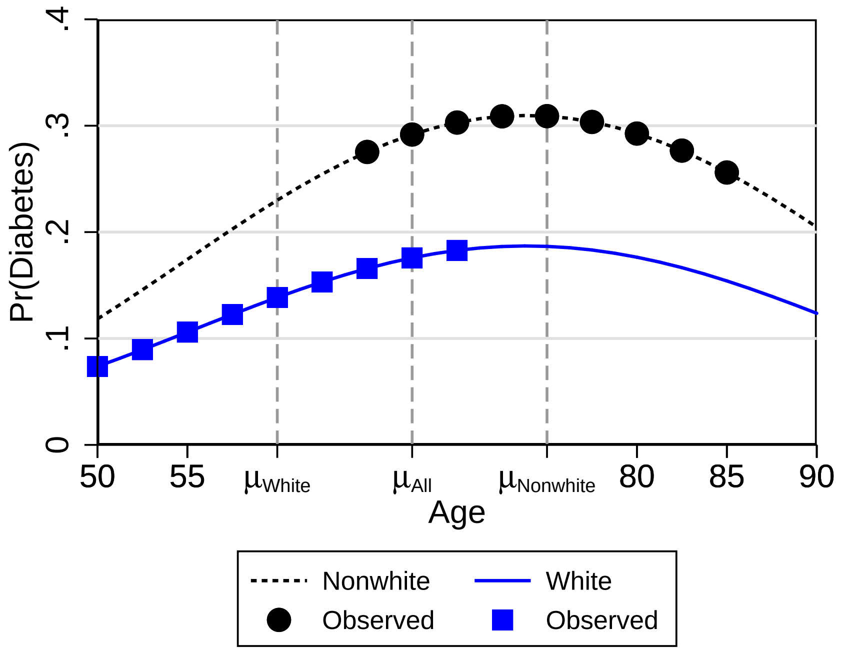


Figure 5: Probability of good health for whites and nonwhites by age.

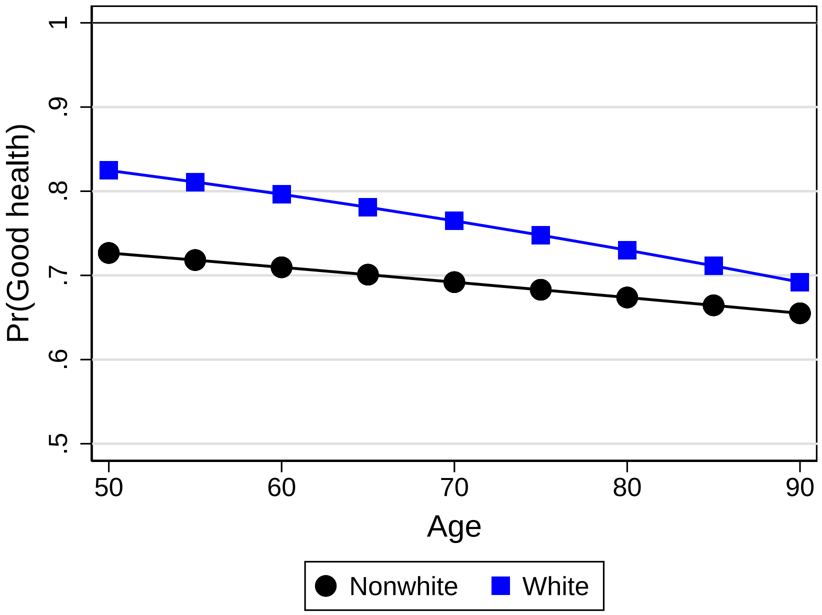


Figure 6: Racial differences in good health by age.

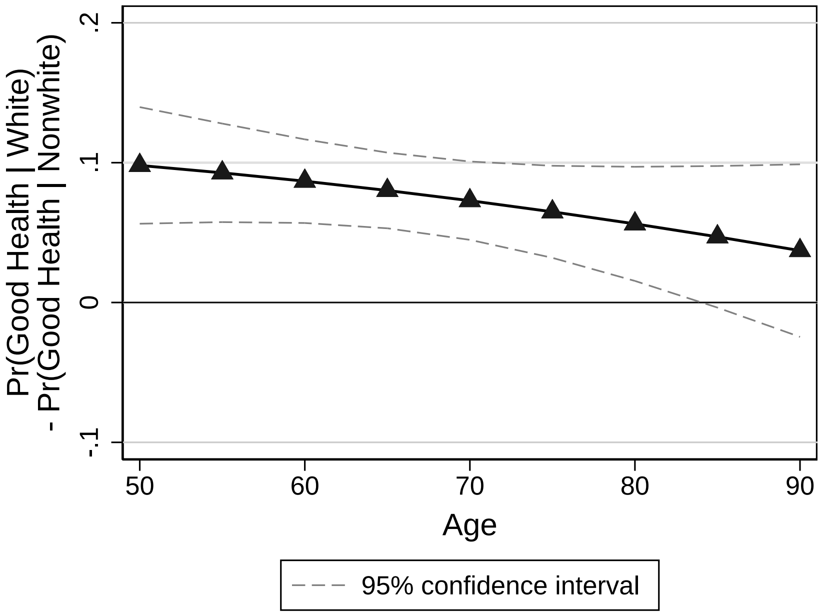


Figure 7: Probability of diabetes for whites and nonwhites by age.

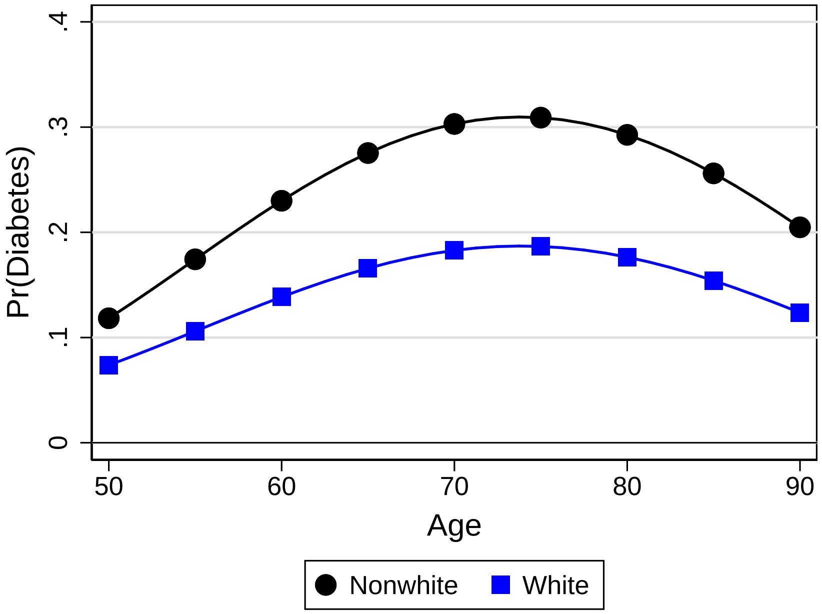


Figure 8: Racial differences in diabetes by age.

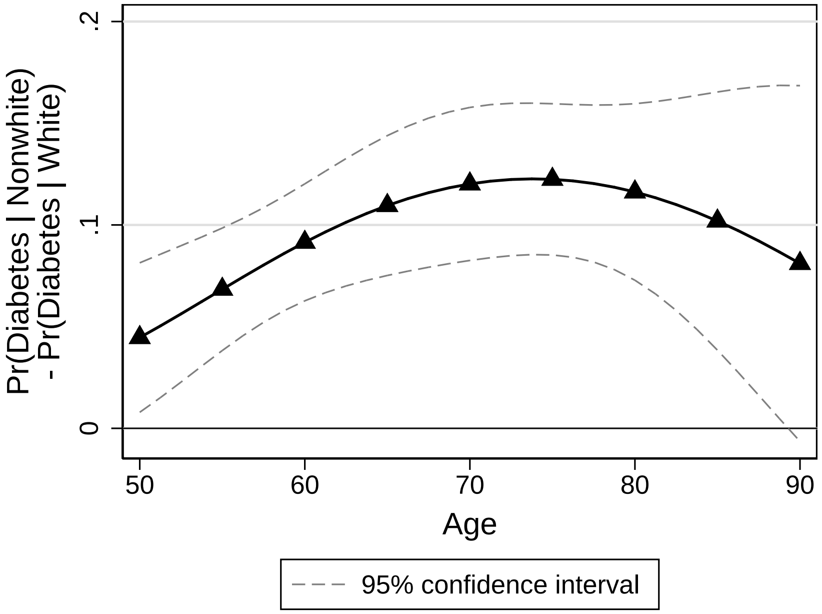


Figure 9: Probability of diabetes for blacks and whites by age and physical activity.

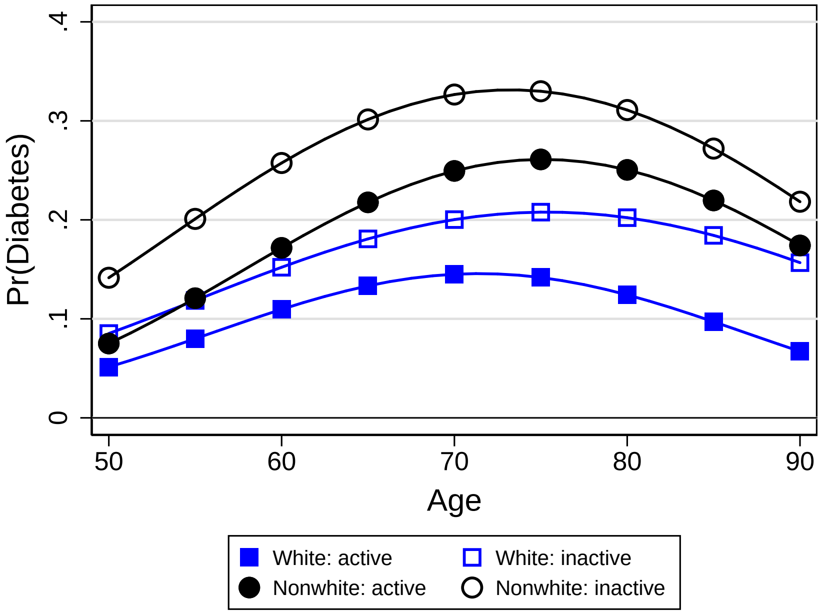


Figure 10: Racial differences in diabetes by age and physical activity. Dashed lines indicate that the difference in probabilities is not significant at the .05 level.

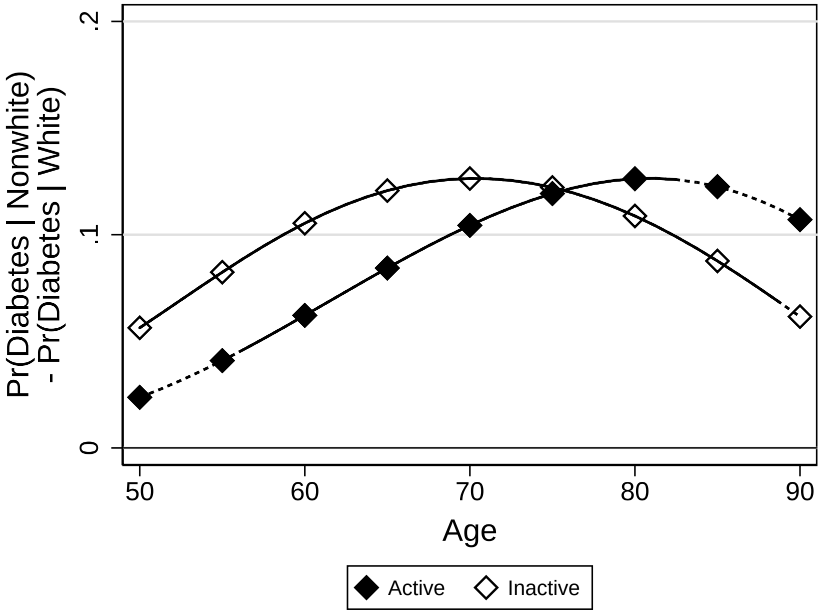
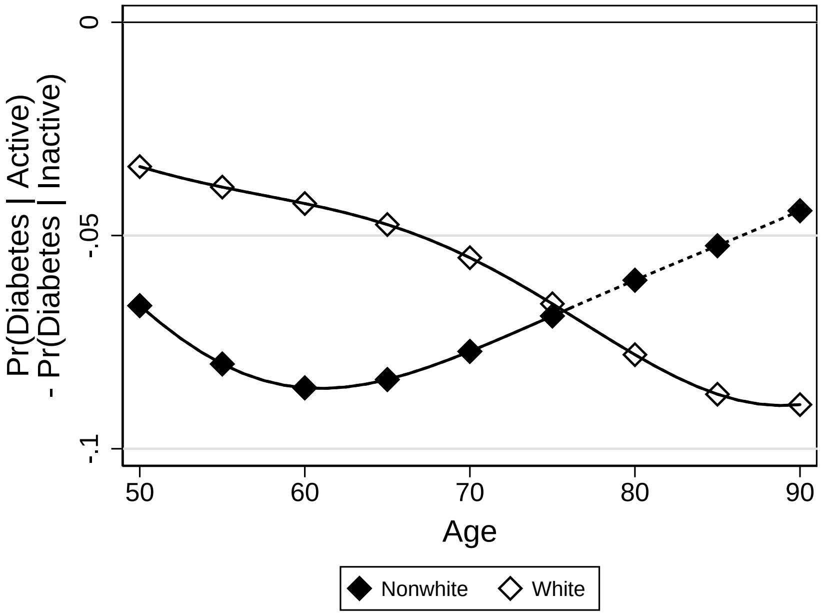


Figure 11: Effects of activity by race and age. Dashed lines indicate that the difference in probabilities is not significant at the .05 level.



7 Tables

Table 1: Descriptive statistics (N=16,226)

	White		Nonwhite		Difference†	
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Δ Mean	p
goodhlth	0.769	——	0.565	——	0.205	<.001
diabetes	0.162	——	0.281	——	-0.120	<.001
female	0.532	——	0.575	——	-0.043	<.001
highschool	0.853	——	0.563	——	0.289	<.001
married	0.692	——	0.541	——	0.150	<.001
income	74.280	99.389	41.013	58.376	33.268	<.001
ihsincome	4.523	1.001	3.809	1.164	0.714	<.001
age	66.514	10.421	64.099	9.677	2.415	<.001
active	0.303	——	0.223	——	0.080	<.001
obese	0.286	——	0.390	——	-0.104	<.001
N	12,427		3,799			

Note: † Δ Mean is the group difference in the means; p is the significance level from testing if means are equal.

Table 2: Logit model for good health (N=16,226).

	White				Nonwhite				H_0: $\beta^{\mathrm{W}} = \beta^{\mathrm{N}}$	
Variable	1: β^{W}	2: OR$^{\mathrm{W}}$	3: t	4: p	5: β^{N}	6: OR$^{\mathrm{N}}$	7: t	8: p	9: F	10: p
Constant	-0.488	——	-1.75	0.086	-1.541	——	-4.00	<.001	4.71	0.034
female	0.144	1.155	2.89	0.006	-0.118	0.888	-1.37	0.176	6.96	0.011
highschool	0.800	2.225	13.45	<.001	0.816	2.262	9.19	<.001	0.02	0.891
married	-0.056	0.945	-0.96	0.341	-0.169	0.844	-1.56	0.124	0.70	0.406
ihsincome	0.556	1.744	15.67	<.001	0.583	1.792	10.06	<.001	0.16	0.695
age	-0.018	0.982	-6.17	<.001	-0.008	0.992	-1.83	0.072	3.47	0.068
obese	-0.573	0.564	-11.16	<.001	-0.361	0.697	-3.25	0.002	3.10	0.084

Note: OR is the odds ratio. Tests of $H_0\colon \beta^{\mathrm{W}} = \beta^{\mathrm{N}}$ are shown for didactic purposes.

Table 3: Logit model for diabetes (N=16,226).

	White				Nonwhite				H_0: $\beta^{\mathrm{W}} = \beta^{\mathrm{N}}$	
Variable	1: β^{W}	2: OR$^{\mathrm{W}}$	3: t	4: p	5: β^{N}	6: OR$^{\mathrm{N}}$	7: t	8: p	9: F	10: p
Constant	-9.627	——	-6.66	<.001	-11.169	——	-5.79	<.001	0.40	0.529
female	-0.400	0.670	-6.53	<.001	-0.079	0.924	-0.80	0.427	7.27	0.009
highschool	-0.253	0.776	-3.60	0.001	-0.142	0.867	-1.42	0.160	0.92	0.342
married	0.070	1.073	0.98	0.333	0.064	1.066	0.60	0.548	0.00	0.960
ihsincome	-0.189	0.828	-5.50	<.001	-0.131	0.877	-3.93	<.001	1.61	0.210
age	0.243	1.274	6.18	<.001	0.299	1.348	5.36	<.001	0.65	0.422
agesq	-0.002	0.998	-5.97	<.001	-0.002	0.998	-5.15	<.001	0.78	0.382
active	-4.048	0.017	-1.04	0.302	-2.557	0.078	-0.42	0.678	0.04	0.849
activeXage	0.115	1.122	0.98	0.331	0.052	1.053	0.28	0.783	0.07	0.791
activeXagesq	-.0009	0.999	-1.02	0.310	-.0003	1.000	-0.21	0.835	0.11	0.737
obese	1.163	3.199	17.01	<.001	0.740	2.095	6.59	<.001	9.35	0.003

Note: OR is the odds ratio. Tests of $H_0\colon \beta^{\mathrm{W}} = \beta^{\mathrm{N}}$ are shown for didactic purposes.

Table 4: The effects of obesity on diabetes by race and gender.

	Probability of diabetes						
	Women		Men		Effect of obesity		
	1: Obese	2: Not	3: Obese	4: Not	5: Women	6: Men	7: Difference
1: White	0.278	0.107	0.365	0.152	0.170*	0.212*	-0.042*
2: Nonwhite	0.389	0.233	0.408	0.248	0.156*	0.161*	-0.005
3: Racial difference	-0.112*	-0.126*	-0.044	-0.096*	-0.014	-0.052†	-0.038*

Note: Other variables held at their means. $* = p < .01$; $\dagger = p < .10$ for two-tailed test.

Table 5: Average discrete change and discrete change at the means for logit model for diabetes (N=16,226).

Panel 1: Average discrete change

	White		Nonwhite		Difference	
Variable	1: ADC^{W}	2: p	3: ADC^{N}	4: p	5: ADC	6: p
female	-0.051	<0.001	-0.015	0.431	-0.036	0.089
highschool	-0.033	0.001	-0.027	0.160	-0.006	0.763
married	0.009	0.330	0.012	0.546	-0.003	0.875
ihsincome	-0.022	<0.001	-0.024	<0.001	0.002	0.779
age	0.011	<0.001	0.027	<0.001	-0.016	0.002
active	-0.053	<0.001	-0.084	<0.001	0.031	0.110
obese	0.169	<0.001	0.146	<0.001	0.022	0.408

Panel B: Discrete change at the mean

	White		Nonwhite		Difference	
Variable	1: DCM^{W}	2: p	3: DCM^{N}	4: p	5: DCM	6: p
female	-0.057	<0.001	-0.016	0.431	-0.041	0.070
highschool	-0.038	0.001	-0.029	0.162	-0.008	0.716
married	0.010	0.328	0.013	0.545	-0.003	0.895
ihsincome	-0.025	<0.001	-0.026	<0.001	0.001	0.916
age	0.014	<0.001	0.023	<0.001	-0.009	0.169
active	-0.049	<0.001	-0.083	0.006	0.034	0.319
obese	0.190	<0.001	0.158	<0.001	0.032	0.264

Note: The effect of age is for a five-year change.

Table 6: Testing the equality of regression coefficients in models for diabetes (N=16,226).

			Coefficients constrained to be equal					
	Full model		M_1: ihsincome		M_2: age, active†		M_3: obese	
Variable	$\Delta\beta$	p	$\Delta\beta$	p	$\Delta\beta$	p	$\Delta\beta$	p
female	-0.321	0.009	-0.199	0.164	-0.363	0.013	-0.176	0.120
obese	0.423	0.003	0.069	0.810	0.530	0.064	0.000‡	1.000‡

Note: Results for the full model are from table 3. Tests from models with coefficients constrained to be equal were estimated with a location scale model. † All coefficients involving age and/or active are constrained to be equal. ‡ Coefficient for obese are constrained to be equal.