Comparing groups in binary regression models using predictions*

J. Scott Long† and Sarah A. Mustillo‡Ind

October 30, 2016

(groups-beta-draft1 2016-10-30.tex)

Abstract

Methods for group comparisons using predicted probabilities and marginal effects are developed for the binary regression model. Unlike tests that compare regression coefficients across groups, these methods are unaffected by the identification of the coefficients and are expressed in the natural metric of the outcome probability. While we focus on the logit model with two groups, our methods can used with most regression models with any number of groups.

*We want to thank ... for their comments.

†jslong@indiana.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.

Comparing groups in binary regression models using predictions

1 Introduction

Studying how groups differ is fundamental to research in many areas where regression models are used to answer two types of questions. First, do groups differ in the level of the outcome adjusting for differences in observed characteristics? For example, do white and nonwhite respondents with the same characteristics have different probabilities of reporting good health? 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, regressions are run for each groups where the coefficients are allowed to differ by group. In linear regression, the standard approach is to use Chow tests (1960) of the hypothesis that coefficients are equal across groups. For example, suppose that we are comparing the effect of x_k on y for white and nonwhite respondents, where $\beta^W_{x_k}$ and $\beta^N_{x_k}$ are the coefficients of interest. If H_0: $\beta^W_{x_k} = \beta^N_{x_k}$ is rejected, we conclude that the effect of x_k differs across groups. While this approach can be used with many types of models (Liao, 2002), Allison (1999) shows that since the regression coefficients in the binary regression models are only identified to a scale factor, standard tests of the equality of coefficients are invalid. He develops valid tests that use auxiliary assumptions that the effects of some regressors are equal across groups. While these tests address the identification problem, there is a more fundamental issue. In the binary regression model, regression coefficients are not expressed in the natural metric of the probability of the outcome. Substantively, it is more useful to understand whether the marginal effects of x_k on the probability of the outcome are the same for both groups than whether the regression coefficients for x_k are equal. Critically, in the binary regression model, the equality of regression coefficients across groups does not imply that the marginal effects on the probability are equal. In this paper, we develop methods for comparing groups using tests of the equality of probabilities conditional on the regressors and tests of the marginal effects of on the probability. Since probabilities are identified, additional identifying assumptions are not required. While we consider models with two groups and a binary outcome in order to simplify our presentation, the results can be generalized to more than two groups and be used with many types of regression models.

The advantages of tests based on predicted probabilities are not without costs. Tests comparing regression coefficients are simple to apply since there is only a single hypothesis of interest: are the coefficients equal across groups. The results of these tests do not depend on where you are looking in the data. For example, when examining race differences in

the regression coefficient for obesity on the onset of diabetes, the conclusion is either that the coefficients are the same or they are not. With methods based on probabilities, the conclusions depend on the specific values of the regressors where the comparison are 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, but 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 and Freese, 2014). 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 for if the woman was white than if she was nonwhite. While the conclusions about groups differences in the effect of obesity are more complex, they have the potential to be more realistic and informative.

[[**TODO:** Add brief review of related literature.]]

The next section explains how the identification of regression coefficients affects group comparisons in the binary regression model (BRM). Section 3 develops methods for comparing conditional predictions and marginal effects, including the extension of the Blinder-Oaxaca decomposition to the BRM. Each of these methods is illustrated in section 4 where we compare white and nonwhite respondents in models predicting self-rated health and being diagnosed with diabetes.

2 Identification in the binary regression model

The identification of regression coefficients is critical for understanding group comparisons in the binary regression model. 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 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 $g=1$. Separate regressions are specified for each group 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). 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 $y = 0$.

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 the logit model and normal for the probit model. 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 have done is multiply both sides of the inequality by δ and changed notation, the probabilities in equation 5 are *exactly* the same as those in equation 2. However, since δ is 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 the same in both figures.

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 $\delta {=} \pi/\sqrt{3}\sigma$. Multiplying y^* by δ rescales the β coefficients while $\Pr(y = 1 \mid x_1, x_2)$ is unaffected. Since σ is unknown, we cannot estimate the βs in equation 1. We can estimate the re-scaled γs in equation 4 since the value of the variance is assumed. The effects of the assumed variance are 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^* for two groups:

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

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

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 $\delta_1 = \sigma_{\text{Assumed}}/\sigma_1$ and equation 7 by $\delta_0 = \sigma_{\text{Assumed}}/\sigma_0$:

$$\text{Group 0:} \quad \delta_0 y^* = (\delta_0\beta_0^0) + (\delta_0\beta_1^0)\, x_1 + (\delta_0\beta_2^0)\, x_2 + \delta_0\varepsilon_0 \text{ where } Var(\delta_0\varepsilon_0) = \sigma^2$$

$$\text{Group 1:} \quad \delta_1 y^* = (\delta_1\beta_0^1) + (\delta_1\beta_1^1)\, x_1 + (\delta_1\beta_2^1)\, x_2 + \delta_1\varepsilon_1 \text{ where } Var(\delta_1\varepsilon_1) = \sigma^2$$

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

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

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

After estimation, we can test $H_{\gamma_k}\colon \gamma_k^0 = \gamma_k^1$ which is equivalent to testing $H_{\gamma_k}\colon \delta^0\beta_k^0 = \delta^1\beta_k^1$. Testing $H_{\beta_k}\colon \beta_k^0 = \beta_k^1$ requires information about 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

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

provides information on the relative magnitudes of the the σ_gs. As illustrated in section 4.4.2, the results of the test for $H_{\beta_k}\colon \beta_k^0 = \beta_k^1$ depend on which β_js are assumed to be equal across groups.

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

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

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

or the γs from equations 8 and 9:

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

The advantage of comparing groups marginal effects using probabilities is not simply avoiding the identification assumptions required to test regression coefficients. Conclusions about the equality of regression coefficients in the BRM are generally less useful than conclusions in the natural metric of probabilities. 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 effects on the propensity or 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 by the vertical arrow in figure 3 which 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 outcome for both groups. For example, does obesity have the same health cost for whites as it does for nonwhites? This is illustrated by the arrows showing the change in probability as age increases from 55 to 60 for each group. While marginal effects and probabilities are related, you cannot draw conclusions about differences in the probabilities from differences in marginal effects. 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 sections 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 $\pi(\mathbf{x}, g)$. Fitting the equations for both groups simultaneously makes post-estimation computations simpler and is necessary for obtaining the correct standard errors when using a complex sampling

design. [[groups-covbetas-2016-09-26.do]]. To do this, we specify a single equation using interactions:

$$\pi\left(\mathbf{x}, g\right) = F\left(\left[g \times \mathbf{x}'\boldsymbol{\gamma}^1\right] + \left[(1-g) \times \mathbf{x}'\boldsymbol{\gamma}^0\right]\right) \tag{10}$$

Then $\pi(\mathbf{x}, g=1) = F(\mathbf{x}'\boldsymbol{\gamma}^1 + 0)$ and $\pi(\mathbf{x}, g=0) = F(0 + \mathbf{x}'\boldsymbol{\gamma}^0)$. 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 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 $\mathbf{x}^*$ contains specific values of the xs. The difference between groups 0 and 1 in the probability at $\mathbf{x} = \mathbf{x}^*$ is the discrete change for group:

$$\frac{\Delta\pi(\mathbf{x} = \mathbf{x}^*)}{\Delta\, g} = \pi(\mathbf{x} = \mathbf{x}^*, g=1) - \pi(\mathbf{x} = \mathbf{x}^*, g=0) \tag{11}$$

To test H_0: $\pi(\mathbf{x}=\mathbf{x}^*, g=0) = \pi(\mathbf{x}=\mathbf{x}^*, g=1)$, we can test if the discrete change for group in equation 11 is 0.

Group differences in probabilities can be used in a variety of ways. Individuals with specific characteristics can be compared. For example, do 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 by presented in tables. For example, race differences in diabetes could be shown for men and women at different levels of education. For continuous regressors, plots can be used. For example, do nonwhites and whites differ in the occurrence of diabetes as they age? These methods are illustrated in sections 4.2 and 4.3.

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. While we focus on discrete changes, since we find them to be more useful substantively, 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:

$$\frac{\Delta\pi(\mathbf{x}=\mathbf{x}^*, g)}{\Delta x_k(start \rightarrow end)} = \pi(x_k = end, \mathbf{x}=\mathbf{x}^*, g) - \pi(x_k = start, \mathbf{x}=\mathbf{x}^*, g) \tag{12}$$

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 $x_{\text{agesq}} = x_{\text{age}} \times x_{\text{age}}$, when x_{age} changes from 10 to 11, then x_{agesq} changes from 100 to 121.

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

$$H_0\text{: } \frac{\Delta\pi(\mathbf{x}=\mathbf{x}^*, g=1)}{\Delta x_k(start \rightarrow end)} = \frac{\Delta\pi(\mathbf{x}=\mathbf{x}^*, g=0)}{\Delta x_k(start \rightarrow end)} \tag{13}$$

Equivalently, we can estimate a *second difference*, which is the change with respect to group in the discrete change for x_k:

$$\frac{\Delta^2\pi(\mathbf{x}=\mathbf{x}^*)}{\Delta x_k(start \rightarrow end)\,\Delta\, g} = \frac{\Delta\pi(\mathbf{x}=\mathbf{x}^*, g=1)}{\Delta x_k(start \rightarrow end)} - \frac{\Delta\pi(\mathbf{x}=\mathbf{x}^*, g=0)}{\Delta x_k(start \rightarrow end)} \tag{14}$$

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

$$H_0\text{: } \frac{\Delta^2\pi(\mathbf{x}=\mathbf{x}^*)}{\Delta x_k(start \rightarrow end)\,\Delta\, g} = 0 \tag{15}$$

Since the value of the discrete change of x_k depends on the values of the regressors where it is estimated, a critical decision is how to summarize the effect (Long and Freese, 2006, 244-246). Two approaches are commonly used. First, the discrete change is estimated at representative values of the xs. Such effects are called 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 as discussed in section 3.5.

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 for x_k evaluated at $\mathbf{x}=\mathbf{x}^*$

equals:

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

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. This is referred to as the discrete change at the mean (DCM). To compare effects across groups we estimate the second difference using equation 14:

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

When x_k is binary, the second difference 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 can test if the second difference is 0. DCRs let us compare the effect of a variable at the same values of the regressors for both groups. Group differences in DCRs do *not* reflect differences in the distribution of the regressors since the same values are used for both groups. This important point is discussed in detail after we consider the ADC.

3.4 Average discrete change (ADC)

The average discrete change for x_k is the average of the discrete changes for x_k computed for each observation. 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 for x_k is

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

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

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

where δ is often 1 or a standard deviation, but other values can be used. It is also possible to change x_k between the same two values for every observations, such as increasing age from

60 to 65 or changing a binary variable from 0 to 1. For group g:

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

When x_k is binary, the simpler notation $\Delta\pi(\mathbf{x}=\mathbf{x}_i,g)/\Delta x_k$ is used. The ADC for x_k in group g is the average of the discrete changes for each observation in that group g:

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

Equations 13-15 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 compare ADCs or DCRs depends on the question being asked. To illustrate what each type of effect tells you, figure 4 plots the probability of diabetes by age for whites and nonwhites from a model with age and age-squared. The circles are observations for the younger sample of nonwhites; the squares are observations for older sample of whites. For nonwhites, $\text{ADC}^N_{\text{age}} = .04$ which is the average change in the probability for each observations as age is increases by 5 from its observed values. For whites, $\text{ADC}^W_{\text{age}}$ is close to 0 since the positive effects of age for those younger than 72 are offset by the negative effects for those older than 72. The differences in the ADCs is due primarily to differences in the distribution of ages and masks the similar shapes of the probability curves. DCRs at specific ages reflect the similar shapes of the curves. At the mean age for nonwhites, $\text{DCR}^W_{\text{age}} = .027$ and $\text{DCR}^N_{\text{age}} = .045$; at the overall mean, $\text{DCR}^W_{\text{age}} = .011$ and $\text{DCR}^N_{\text{age}} = .018$; and at the mean for whites, $\text{DCR}^W_{\text{age}} = -.007$ and $\text{DCR}^N_{\text{age}} = -.011$. DCRs compare the shape of the probability curves at the same values of the regressors, whiles ADCs reflect both differences in the curves and the distribution of regressors. Indeed, two groups can have exactly the same regression coefficients with significantly different ADCs. Neither the ADC or the ADC is always better—they simply reflect different ways in which effects differ across groups as shown in section 4.

3.6 Blinder-Oaxaca decompositions of probabilities

A useful way to begin analyses is to determine how much of the group difference in the outcome can be attributed to differences in the distribution of the regressors. This is done

with a decomposition developed by Blinder (1973) and Oaxaca (1973) for the linear regression model that was extended to binary models by Fairlie (1999, 2005). More general methods for any nonlinear model were proposed by Yun (2004) and Bauer et al. (2007). Here we consider Fairlie's decomposition.

The decomposition begins with the *overall* difference in the probability of the outcome computed using equation 10:

$$\overline{\pi}(\mathbf{x}, g{=}1) - \overline{\pi}(\mathbf{x}, g{=}0) = \sum_{\text{if } g=1} \frac{\pi(\mathbf{x}_i, g{=}1)}{N_1} - \sum_{\text{if } g=0} \frac{\pi(\mathbf{x}_i, g{=}0)}{N_0} \tag{16}$$

where "if $g=$" indicates summing over the N_g observations in group g. In many applications, a difference in the outcome can be attributed to differences in characteristics. For example, the poorer health of nonwhites can be explained by their having lower values of regressors that positively affect health and higher values of those that negatively affect health. Accordingly, if the groups had more similar characteristics, the overall difference would be reduced. It is also possible that the overall difference is small, but would be larger or even change direction if the groups had the same characteristics.

The decomposition is constructed by estimating probabilities in two ways. Standard predictions are made using equation 10:

$$\pi(\mathbf{x}_i^g; \boldsymbol{\gamma}^g) = F(\mathbf{x}_i' \boldsymbol{\gamma}^g) \quad \text{if } g$$

where the new notation $\pi(\mathbf{x}_i^g; \boldsymbol{\gamma}^g)$ makes explicit that probabilities are computed using parameters from the model for group g with the xs observed for that group. The average probabilities in equation 16 are:

$$\overline{\pi}(\mathbf{x}^g; \boldsymbol{\gamma}^g) = \sum_{\text{if } g} \frac{\pi(\mathbf{x}_i^g; \boldsymbol{\gamma}^g)}{N_g}$$

Counterfactual predictions are made using parameters from one groups with the xs observed for the other group. For example, using the parameters for group 1 with observed data for group 0:

$$\pi(\mathbf{x}_i^0; \boldsymbol{\gamma}^1) = F(\mathbf{x}_i' \boldsymbol{\gamma}^1) \quad \text{if } g{=}0$$

The average probability if group 1 with coefficients $\boldsymbol{\gamma}^1$ had the characteristics of those in

group 0 is:

$$\overline{\pi}(\mathbf{x}^0;\boldsymbol{\gamma}^1) = \sum_{\text{if } g=0} \frac{\pi(\mathbf{x}_i^0;\boldsymbol{\gamma}^1)}{N_0}$$

Similarly, $\overline{\pi}(\mathbf{x}^1;\boldsymbol{\gamma}^0)$ is the average probability if group 0 had the characteristics of group 1. The average counterfactual predictions estimates what would happen if one group had the characteristics of the other group.

The decomposition is created by adding $\overline{\pi}(\mathbf{x}^0;\boldsymbol{\gamma}^1) - \overline{\pi}(\mathbf{x}^0;\boldsymbol{\gamma}^1) = 0$ to equation 16 and rearranging terms:

$$\overline{\pi}(\mathbf{x}^1;\boldsymbol{\gamma}^1) - \overline{\pi}(\mathbf{x}^0;\boldsymbol{\gamma}^0) = \left[\,\overline{\pi}(\mathbf{x}^1;\boldsymbol{\gamma}^1) - \overline{\pi}(\mathbf{x}^0;\boldsymbol{\gamma}^1)\,\right] + \left[\,\overline{\pi}(\mathbf{x}^0;\boldsymbol{\gamma}^1) - \overline{\pi}(\mathbf{x}^0;\boldsymbol{\gamma}^0)\,\right] \tag{17}$$

The difference in the first set of brackets compares the probabilities for group 1 computed using the observed characteristics of group 1 with the counterfactual probabilities obtained if group 1 with coefficients $\boldsymbol{\gamma}^1$ had the same characteristics as group 0. It is the amount of the overall group difference accounted for by group differences in observed characteristics. The difference in the second set of brackets is the portion of the overall difference that cannot be attributed to differences in characteristics, reflecting both differences in regression parameters and unobserved characteristics (see Jann 2008 for an excellent discussion). An equally valid decompositions uses predictions for group 0 using characteristics for group 1:

$$\overline{\pi}(\mathbf{x}^1;\boldsymbol{\gamma}^1) - \overline{\pi}(\mathbf{x}^0;\boldsymbol{\gamma}^0) = \left[\,\overline{\pi}(\mathbf{x}^1;\boldsymbol{\gamma}^0) - \overline{\pi}(\mathbf{x}^0;\boldsymbol{\gamma}^0)\,\right] + \left[\,\overline{\pi}(\mathbf{x}^1;\boldsymbol{\gamma}^1) - \overline{\pi}(\mathbf{x}^1;\boldsymbol{\gamma}^0)\,\right] \tag{18}$$

When the decomposition is used to assess discrimination, equation 17 is used when discrimination is assumed to affect group 1 and equation 18 when discrimination affects group 0. For our purposes, both forms of the decomposition are used to assess how much of the difference in outcomes can be explained by group differences in the regressors. The delta method or bootstrapping can be used to compute standard errors for each difference in the decomposition. Section 4.1 contains an example of the decomposition.

4 Example: Health and Retirement Study

Our example uses data from the Health and Retirement Study (HRS), a nationally-representative sample of older adults in the US (Health and Retirement Study, 2006).[1] Approximately

[1]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. Our analysis file can be downloaded from [[**TODO details to be added**]] after registering for use of the HRS.

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 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. [[groups-hrs-supportV5.do #6]] Analyses were conducted with Stata 14.2 using adjustments for complex sampling (StataCorp, 2015a). Two-tailed tests are used in all analyses.

— 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 and an inverse hyperbolic sine transformation is 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 [[**TODO add citation**]], 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 regression coefficients from two models. The model for good health is simpler to interpret since it does not include squared terms or interactions among regressors. The model for diabetes includes age-squared and an interaction between age and level of activity which makes the interpretation more challenging. We begin interpretation with a Blinder-Oaxaca decomposition to determine how much of the overall group differences in health outcomes can be explained by differences in the regressors. Next, graphs are used to explore group differences in the outcomes by age. The discrete change

with respect to race 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 differs by race. A more complex example of using graphs looks at the effects of age on diabetes for those with different levels of activity. Tables are then used to examine race differences in the effects of gender and obesity. In these tables, discrete changes for gender and obesity are compared across race using second differences. Finally, race differences in scalar measures of the effect of a regressor are considered using discrete changes and regression coefficients.

4.1 Decomposing group differences

The probability of good health is .205 greater for whites than nonwhites, while the probability of diabetes is .120 greater for nonwhites than whites. Table 1 shows significant race differences in the regressors, where in general nonwhites have higher levels of regressors associated with poorer health or lower levels of those associated with better health. For example, nonwhites are more likely to be obese and less likely to be active, both of which increase the probability of diabetes. The Blinder-Oaxaca decomposition (section 3.6) determines how much of the overall differences in health outcomes can be attributed to these differences in characteristics.

— Table 4 here —

Table 4 decomposes the overall differences in outcomes by computing race differences under two assumptions: (1) nonwhite respondents have the same characteristics as white respondents; and (2) whites have the same characteristics as nonwhites. If nonwhites had the characteristics of whites, the overall difference of .205 in the probability of good health would be reduced by 41 percent to .120, with a reduction of 33 percent to .137 if we assume that whites had the characteristics of nonwhites. For diabetes an even larger proportion of the .120 advantage for whites can be explained by differences in characteristics, reducing the overall difference by 67 percent if nonwhites had the characteristics of whites and by 71 if whites had the characteristics of nonwhites, with a difference of less that .04 that cannot be explained by differences in characteristics. Group differences in characteristics that affect health account for a substantial portion of the overall differences in health outcomes. Still, non-trivial and significant differences remain.

The Blinder-Oaxaca decomposition uses averages in probabilities over the sample rather than differences in probabilities at specific values of the regressors. Accordingly, the decomposition cannot answer questions such as whether the incidence of diabetes differs by race for married women who are 85 with median income. Nor, can a decomposition examine

whether the effect of a variable, say the effect of obesity on diabetes, is the same for both groups. These dimensions of groups differences are examined in the next three sections.[2]

4.2 Comparing groups using graphs

Graphs show both group differences in predictions across values of a regressor and differences in how the effects of a regressor differ across groups. To illustrate this approach, we begin with a simple example showing how whites and nonwhites differ in reporting good health at different ages. We 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 race 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. [[*groups-goodhlth-paperV8.do #3.1*]] The discrete change for race conditional on age is used to test if these differences are significant (see equation 11):

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

Figure 6 plots these race 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 computing the discrete change in the probability of good health as age increase from 50 to 90 for each group g (see equation 12):

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

Group differences in the effect of age are computed as a second difference (see equation 14):

$$\frac{\Delta\pi(\mathbf{x}=\overline{\mathbf{x}})}{\Delta\text{age}(50 \rightarrow 90)\,\Delta\text{white}} = \frac{\Delta\pi(\mathbf{x}=\overline{\mathbf{x}}, \text{white}=1)}{\Delta\text{age}(50 \rightarrow 90)} - \frac{\Delta\pi(\mathbf{x}=\overline{\mathbf{x}}, \text{white}=0)}{\Delta\text{age}(50 \rightarrow 90)}$$

[2]Blinder-Oaxaca decompositions can examine how differences in a single characteristics affect differences in the outcome (Fairlie 1999, 2005, Sinning et al. 2008).

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). [[*groups-goodhlth-paperV8.do #3.2a*]]

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 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. [[*groups-diabetes-paperV8.do #3.1*]] This informal summary of the differences between the probability curves is formalizing by plotting the discrete change for race by age in figure 8. Race differences increase from age 50 to 75 followed by a gradual decrease. Differences in diabetes are significant at all ages except 90 where the confidence interval includes 0.

— Figures 7 and 8 here —

The changing size of the effect of race at different ages occurs because the rate of increase in diabetes with age is larger for nonwhites than whites from 50 to 75 at which point the rate of decrease 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 14). [[*groups-diabetes-paperV8.do #3.2*]] 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. For example, suppose that we want to determine if the race differences in diabetes that we found in figure 7 vary by the level of physical activity. Or, to put it another way, are the benefits of activity different for nonwhites and whites over the life course? 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 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 race differences in diabetes

over age conditional on level of activity. This involves plotting the difference between the probability curves for those who are active (solid symbols) and the curves for those who are inactive (hollow symbols). Second, we can examine the effects of activity by plotting the discrete change of activity by race: this 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 race differences in the probability of diabetes by level of activity over age: $\Delta\pi(\text{age}=p, \text{active}=q, \mathbf{x}=\overline{\mathbf{x}})/\Delta\text{white}$ (see equation 11). Since adding confidence intervals to the figure leads to overlapping lines that are confusing, a dashed line is used to indicate when a race 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, 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, shown with 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 of activity occur differently. 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 75 the effects are not significant. Tests of second differences show that race differences in the effect of activity are significant at the .05 level at age 57 where the difference reaches its maximum of .043, are significant at the .10 level between ages 55 and 60, and are not significant at other ages. [[*groups-diabetes-paperV8.do #3.3c*]]

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; none of the differences are significant. [[*groups-diabetes-paperV8.do #3.4*]]

4.3 Comparing groups using tables

— Table 5 here —

Tables are an effective way to show how probabilities vary over the categories of a few regressors. Suppose that we are interested in whether race 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. These probabilities are presented in columns 1 through 4 of table 5, where the last row shows race differences in these probabilities (see equation 11):

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

In the rest of this section, we exclude $\mathbf{x}=\overline{\mathbf{x}}$ from $\pi()$ to simplify the notation.

The table shows that whites are less likely to be diagnosed with diabetes for all combinations of obesity and gender, with the largest race differences for women who are not obese and the smallest, non-significant difference for obese men. Second differences are used to test whether race differences for men and women are of equal size (see equation 14):

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

We find that the effect of race for obese men and women differ by -0.068 which is significant at the .02 level. For those who are not obese, the difference is smaller and not significant ($p=.13$). [[*groups-diabetes-paperV8.do #4.2*]]

Next, we consider the effects of obesity on diabetes. The probabilities in the first four columns show that being obese is associated with a higher incidence of diabetes. To formalize these findings, we compute the discrete change for obesity conditional on gender and race holding other variables at their means (see equation 12):

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

These effects, presented in 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 the effects of obesity are equal for whites and nonwhites, we estimate second differences (see equation 14):

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

The results, shown in the last row of columns 5 and 6, indicate that race differences in the effects of obesity are small and not significant for women, but larger and marginally significant for men ($p = .09$). We can test whether the effect of obesity is the same for men and women by computing the second difference with respect to gender:

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

which are shown in column 7. Obesity has a significantly larger effect for white men than white women, but the gender differences are small and nonsignificant nonwhite respondents. The idea of a second difference can be extended to compare any two effects, such as whether the effect of obesity is the same for white men and nonwhite women:

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

or to test the hypothesis if effects for all four groups are equal:

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

which is rejected at the .001 level. [[*groups-diabetes-paperV8.do #4.3a*]]

In table 5 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 compute effects for each group at the overall means when table 1 showed significant race differences in the distribution of the regressors? An alternative approaches that take these differences into account compute the ADC for obesity for each group defined by race and gender. These ADCs reflect group differences in the regressors, but do not show whether the effects of obesity were similar across groups for individuals with similar characteristics (see section 3.5). Another approach is to compare the effects for each group holding other variables at the means specific for each group (see Long and Freese 2014 for a discussion of local means). The most effective approach depends on the specific question motivating the research.

4.4 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 variable across groups. These measures are the counterpart to regression coefficients that a routinely used in linear regression to summarize the effect of each variable. In the binary regression model there are several measures that should be considered and some that should be avoided. We begin by considering by comparing marginal effects across groups, 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.

4.4.1 Comparing marginal effects

— Table 6 here —

The discrete change at the mean (see equation 3.3) and the average discrete change (see equation 3.4) are two standard ways to measure the effect of a variable.[3] Table 6 contains estimates for both measures along with race differences in the effects (see equation 14) and p-values from testing whether the effects are equal. Consider the ADC for 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 nonsignificant decrease of .015 for nonwhites. These effects differ by .038, 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). 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 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.

4.4.2 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,

[3]While marginal change (i.e., the instantaneous rate of change) can also be used, we prefer the discrete change because of its simpler interpretation.

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 variable affects the probability. Second, while odds ratios are in a more natural metric, odds ratios can be misleading. If the odds ratio for x_k is identical for both groups, the effect of x_k on the probability can be very different in the two groups. Third, if the models includes interactions or polynomials, there is no simple way to express the total effect of those variables in using regression coefficient or odds ratios, while marginal effects on the probability take into account the all coefficients that include the variable (e.g., the marginal effect of age 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. To illustrate this important issue, we consider alternative tests of the hypothesis the regression coefficients for gender and obesity are equal for whites and nonwhites.

Columns 9 and 10 of table 3 contain results form standard Wald tests of $H_0\colon \beta_k^{\mathrm{W}} = \beta_k^{\mathrm{N}}$. If these tests were appropriate, which they are not, 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 the conclusions using discrete changes in table 6. 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 2). While this assumption 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 regression coefficients for gender and obesity using different assumptions about which coefficients are equal across groups.

— Table 7 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; and in M_3 the coefficients for *obese* are constrained. Models which constrained the effects of being female, graduating from high school, or being married did not converge. The results are shown in table 7. Race differ-

ences 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* ($p = .163$) or *obese* ($p = .122$). 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 ($p = .063$) 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 the predicted probabilities and marginal effects are exactly equal to those from the full model. 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 had 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 the substantive results 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 to be equal can affect the substantive conclusions, it may be difficult to know whether one has the "correct" model. Making an *ad hoc* decision that some regression coefficients are equal can lead to incorrect conclusions. In our applications, we did not have strong reasons to constrain particular coefficients and tried multiple specifications that lead to substantively quite different conclusions. The risk of specifying the inappropriate identifying assumptions does not exist at the level of the probabilities.

5 Conclusions

TODO: add methods apply for interpreting interactions more generally.

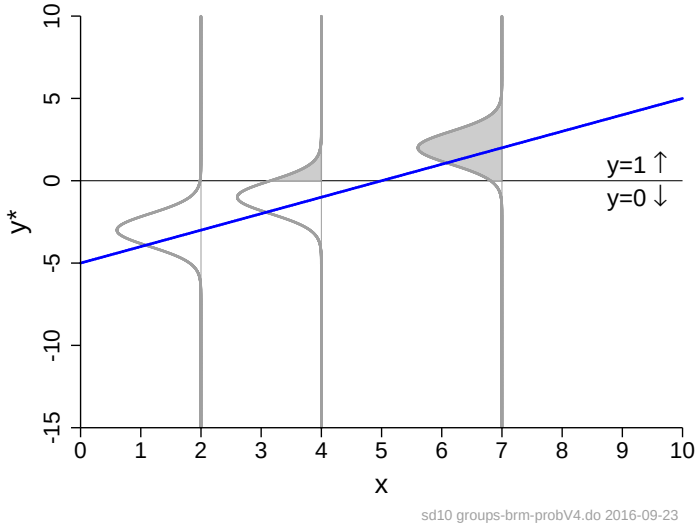
In this paper we have developed methods for comparing groups using predicted probabilities and marginal effects on probabilities for the binary regression model. Since the model is nonlinear, conclusions on whether groups differ in the probability of the outcome or in

the effect of a regressor on the where in the data the groups are compared. 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 simple tests of the equality of regression coefficients, the more complex task of comparing predictions and marginal effects provides greater substantive insights. The world is too complex the effects of regressors or levels of the outcome to always be the same or always be different across groups.

While our examples are relatively simple, the same methods can be used with any number of groups and in models with many regressors including higher order interactions. Further, the methods we use can be used more generally to interpret interactions, such as whether the effect of a variable depends on the level of another variable. Moreover, these methods can be used in any model where your software can make predictions and estimate marginal effects. We used Stata 14's (StataCorp, 2015b) `margin` command along with the `SPost13` package (Long and Freese, 2014).[4] Similar features are being developed in R (Leeper, 2016).

6 Figures

Figure 1: The link between $y^* = \beta_0 + \beta_1 x + \varepsilon$ and $\Pr(y = 1 \mid x)$ with $Var(\varepsilon) = \sigma^2$. [[*groups-brm-probV4.do*]]



[4]Sample command files are available at [[to be added]]. To obtain the HRS data you must [[to be added]].

Figure 2: The link between $\delta y^* = (\delta\beta_0)+(\delta\beta_1)x+\delta\varepsilon$ and $\Pr(y=1 \mid x)$ with $Var(\delta\varepsilon) = \delta^2\sigma^2$. [[*groups-brm-probV4.do*]]

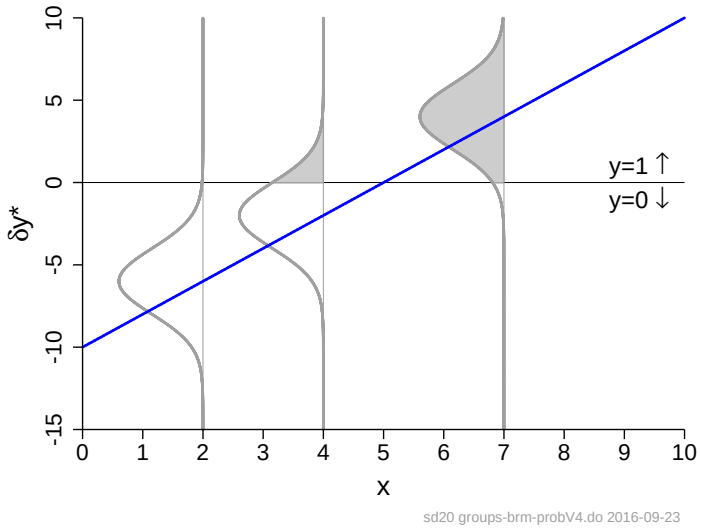


Figure 3: Group comparisons of probabilities and marginal effects. [[*groups-didactic-AMEvMEMV11.do*]]

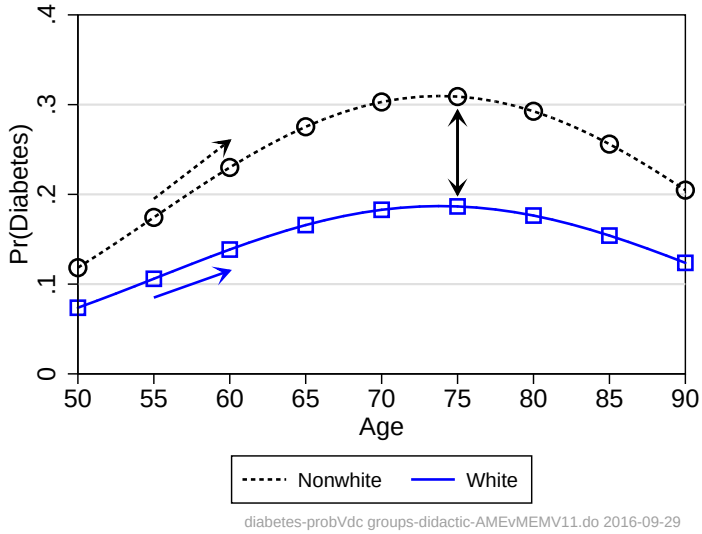


Figure 4: Group differences in the discrete change for age. [[*groups-didactic-AMEvMEMV11.do*]]

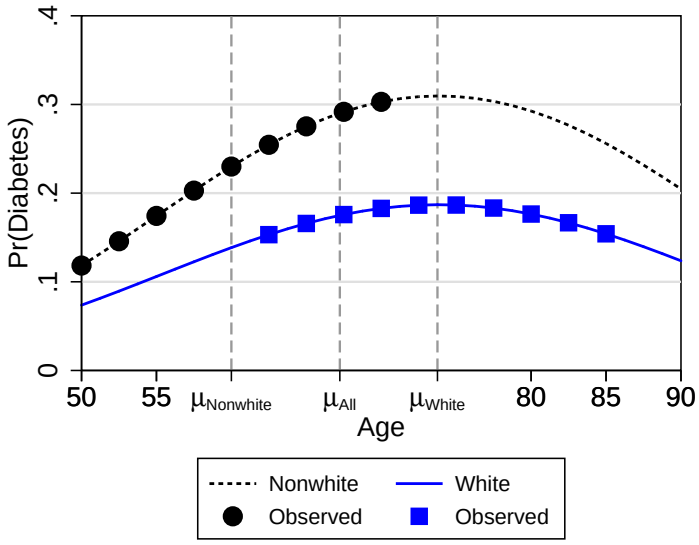


Figure 5: Probability of good health for whites and nonwhites by age. [[*groups-goodhlth-paperV8.do #3.1*]]

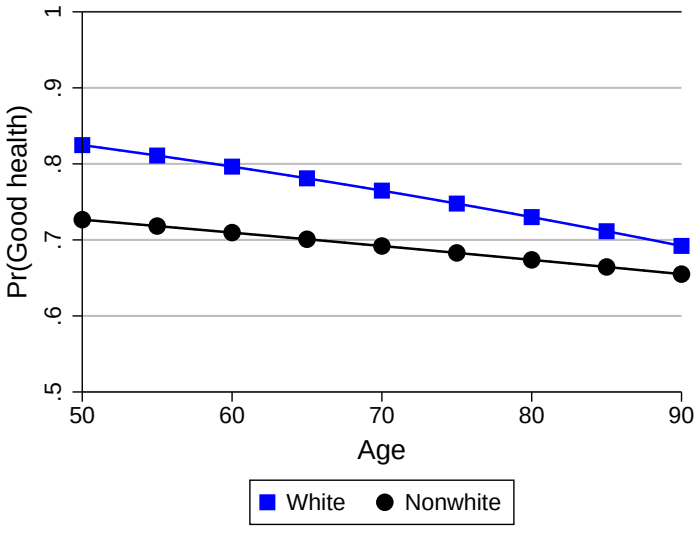


Figure 6: Race differences in good health by age. [[*groups-goodhlth-paperV8.do #3.1*]]

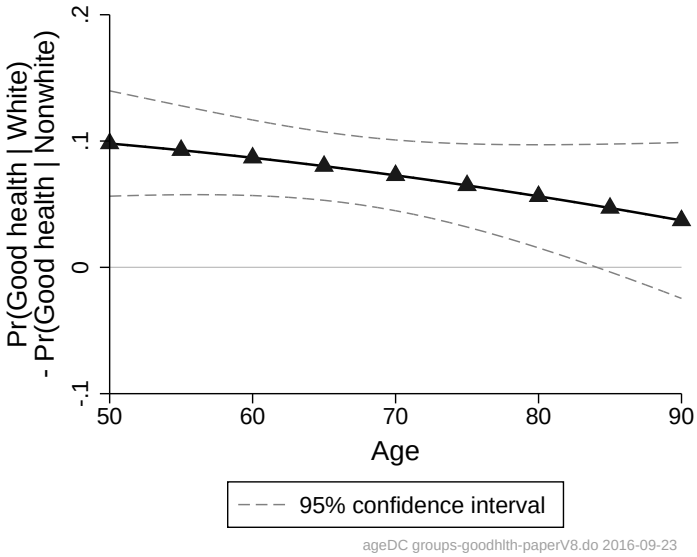


Figure 7: Probability of diabetes for whites and nonwhites by age. [[*groups-diabetes-paperV8.do #3.1*]]

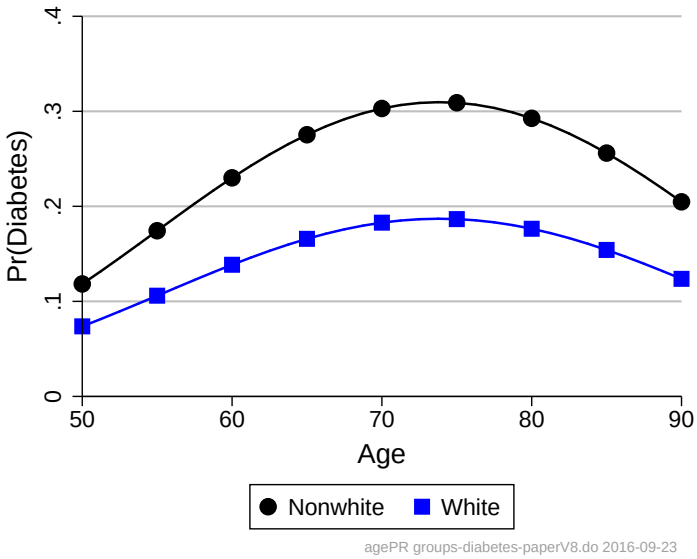


Figure 8: Race differences in diabetes for whites and nonwhites by age. [[*/groups-diabetes-paperV8.do #3.1*]]

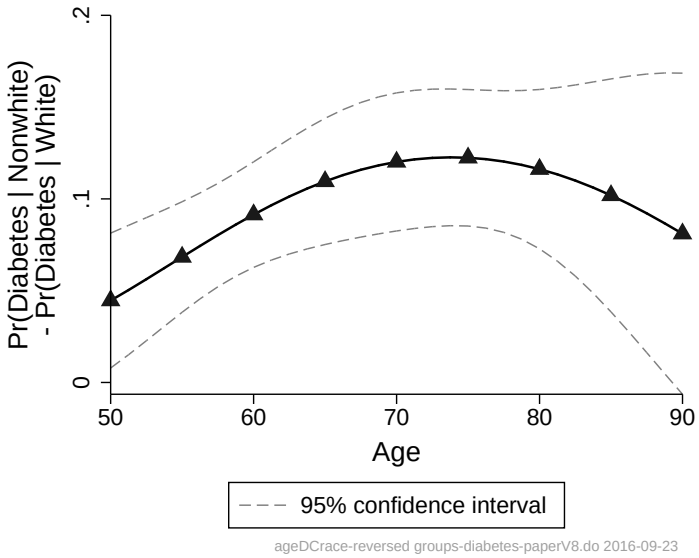


Figure 9: Probability of diabetes for blacks and whites by age and physical activity. [[*groups-diabetes-paperV8.do #3.3*]]

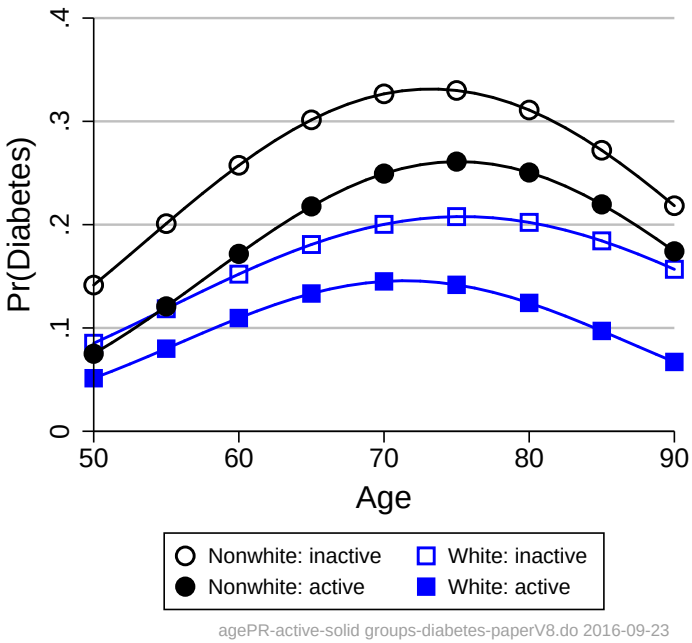


Figure 10: Race differences in diabetes by age and physical activity. Dashed lines indicate that the difference in probabilities is not significant at the .05 level. [[*groups-diabetes-paperV8.do #3.3a*]]

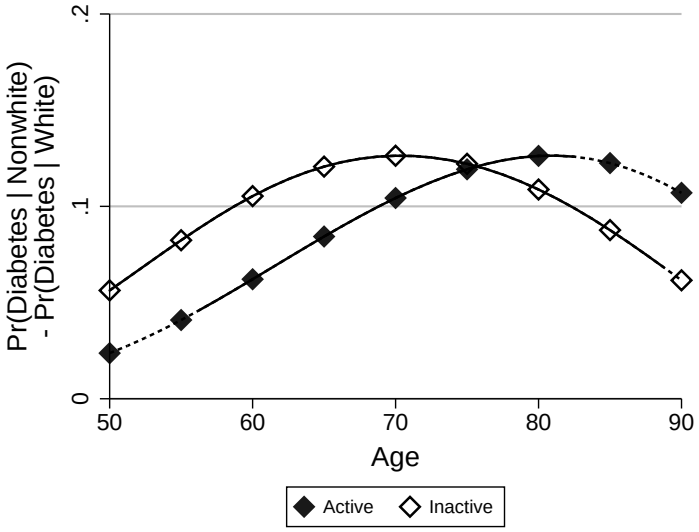
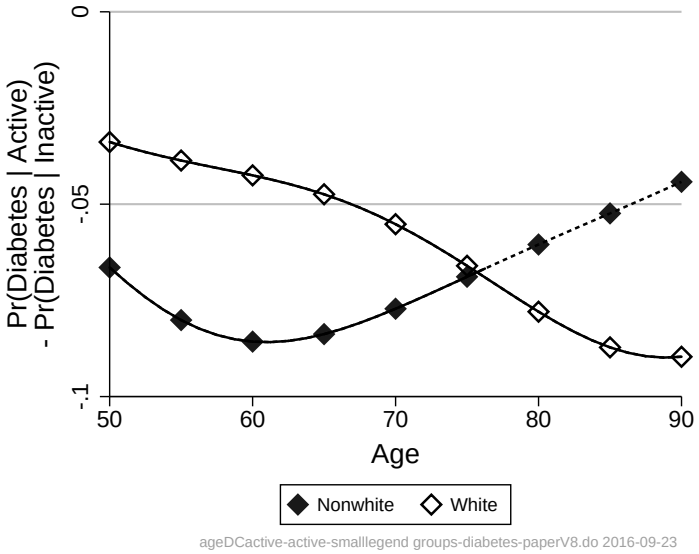


Figure 11: Effects of activity by race and age. Dashed lines indicate that the difference in probabilities is not significant at the .05 level. [[*groups-diabetes-paperV8.do #3.3b*]]



7 Tables

Table 1: Descriptive statistics (N=16,226) [[*groups-descriptive-paperV8.do #5*]]

	White		Nonwhite		Difference†	
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Amount	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: † Amount is the group difference in the means; p is the significance level from testing if means are equal.

Table 2: Logit model for reporting good health (N=16,226). [[*groups-goodhlth-paperV8.do #1.1*]]

	White				Nonwhite				H_0: $\beta^{\mathrm{W}}=\beta^{\mathrm{N}}$	
Variable	1: β^{W}	2: OR^{W}	3: t	4: p	5: β^{N}	6: OR^{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: $\beta^{\mathrm{W}}=\beta^{\mathrm{N}}$ are shown for didactic purposes.

Table 3: Logit model for diabetes (N=16,226). [[*groups-diabetes-paperV8.do #1.1*]]

	White				Nonwhite				H_0: $\beta^W = \beta^N$	
Variable	1: β^W	2: OR^W	3: t	4: p	5: β^N	6: OR^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	0.70	0.500
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 $H_0: \beta^W = \beta^N$ are shown for didactic purposes.

Table 4: Decomposition of raw differences in outcomes (N=16,226). [[*groups-diabetes-paperV8.do #2.2 groups-goodhlth-paperV8.do #2.2*]]

	Good health			Diabetes		
	W - N Difference	Percent Reduction	p	W - N Difference	Percent Reduction	p
Overall	0.205		<.001	-0.120		<.001
If N had W characteristics	0.120	41%	<.001	-0.039	67%	<.001
If W had N characteristics	0.137	33%	<.001	-0.034	71%	<.001

Note: N=nonwhite; W=white.

Table 5: The effects of obesity on diabetes by gender and race. [[*groups-diabetes-paperV8.do #4.5*]]

	Probability of diabetes						
	Women		Men		Effect of obesity		
	1: Obese	2: Not	3: Obese	4: Not	5: Women	6: Men	7: Difference
White	0.278	0.107	0.365	0.152	0.170*	0.212*	-0.042*
Nonwhite	0.389	0.233	0.408	0.248	0.156*	0.161*	-0.005
Race 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 6: Average discrete change and discrete change at the means for logit model for diabetes (N=16,226). [[*groups-diabetes-paperV8.do #5.1, 5.2*]]

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 7: Testing equality of regression coefficients in model for diabetes (N=16,226). [[*groups-diabetes-paperV8b.do #6.2 6.3*]]

			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.

References

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

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.

Bauer, T., S. Göhlmann, and M. Sinning. 2007. Gender differences in smoking behavior. *Health Economics* 16(9): 895–909.

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

Blinder, A. S. 1973. Wage discrimination: reduced form and structural estimates. *Journal of Human Resources* 436–455.

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.

Fairlie, R. W. 1999. The absence of the African-American owned business: An analysis of the dynamics of self-employment. *Journal of Labor Economics* 17(1): 80–108.

———. 2005. An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. *Journal of economic and social measurement* 30(4): 305–316.

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).

Jann, B. 2008. The Blinder-Oaxaca decomposition for linear regression models. *The Stata Journal* 8(4): 453–479.

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.

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.

Oaxaca, R. 1973. Male-female wage differentials in urban labor markets. *International Eeconomic Review* 693–709.

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.

Sinning, M., M. Hahn, T. K. Bauer, et al. 2008. The Blinder-Oaxaca decomposition for nonlinear regression models. *Stata Journal* 8(4): 480–492.

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

———. 2015b. *Stata Statistical Software: Release 14.* College Station, TX: StataCorp LP.

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

Yun, M.-S. 2004. Decomposing differences in the first moment. *Economics letters* 82(2): 275–280.