3

Dominance Analysis

This primer focuses on dominance analysis. We assume you have read the section on dominance analysis in Chapter 11, but we repeat parts of it here to set context. An issue often addressed when using multiple regression is that of identifying the relative importance of different predictors in the regression equation. Two general approaches have been used. One approach seeks to identify the subset of "important" predictors versus the subset of predictors that are "not important," but without ranking predictors beyond this two category scheme. The second approach seeks to rank order all predictors in terms of their relative importance and provides a quantitative index of the magnitude of importance of the predictors. Dominance analysis focuses on the latter. In the ensuing discussion, we assume you are familiar with the basics of multiple regression. We begin by discussing the meaning of relative importance. We then consider the role of statistical significance tests in relative importance analysis. Next, we discuss seven methods for discerning relative importance, including standardized regression coefficients, zero order correlations, squared semi-part correlations, the Platt index, stepwise regression, orthogonalization methods, and dominance analysis. Finally, we discuss the role of sampling error in relative importance analysis, the use of covariates, and the case of binary outcomes.

WHAT IS RELATIVE IMPORTANCE?

There are many ways one can define relative importance in a regression context. One approach is to define it in terms of the proportion of variability that a given predictor accounts for in a criterion - more important variables account for more variability. Another approach is to define it in terms of how much outcome variability a predictor accounts for net other predictors in the equation. A complication with this approach is that there are many ways one can define the concept of "net other predictors." Yet another approach focuses on change - if you change a predictor by a certain amount, how much does the outcome variable change? Predictors associated with more change in the outcome are more important, after adjusting for the metric difference in the predictors. A final strategy focuses on reduction of prediction error - the relative importance of predictors is dictated by the amount each predictor reduces the disparity between the predicted and observed outcome values.

Given the above, it should not be surprising that there is controversy among statisticians about how best to quantify relative importance of predictors in a regression equation. There is no one "correct" approach. Researchers often address the matter from multiple perspectives, using two or more of the approaches mentioned above. You also should keep in mind that the focus is on judging the *relative* importance of predictors not the absolute importance of a given effect. Judgments of absolute importance are inherently value laden and are far more complex than judgments of relative importance. For example, Rosenthal (1995) describes the decision to prematurely terminate a randomized trial on the effects of taking a small dose of aspirin each day on reducing heart attacks in middle age adults. The reason for the termination was because it had become so evident early in the trial that aspirin reduced heart attacks that to continue the study was deemed unethical for participants in the placebo condition. Thousands of lives every year would be saved by having physicians prescribe a small dose of aspirin each day as a preventative treatment to their middle-aged patients. Interestingly, the squared correlation indexing the effect of aspirin on heart attacks was a meager 0.0011, corresponding to a Cohen's d of less than 0.01. In the abstract, this effect size would be deemed by many researchers as trivial. However, for life and death matters and the sheer number of people affected, the effect was extremely important. Finally, relative importance indices do not take into account issues of cost, practicality or modifiability, criteria that often are important in applied contexts. For example, a personnel manager might only want to consider variables that are easy to measure when forecasting future job performance; a prevention program designer might only want to consider variables that are potentially modifiable.

STATISTICAL SIGNIFICANCE AND RELATIVE IMPORTANCE

Some researchers use statistical significance tests as an initial screen for relative importance analysis. The idea is to trim from the equation predictors that are statistically non-significant and then focus the analysis of relative importance on the retained predictors. As straightforward as this seems, the approach can be problematic. The most obvious problem is that a study may have low statistical power, leading to the finding that important predictors have non-significant coefficients. This problem is more insidious than most scientists realize. Maxwell (2000) suggests that the typical correlation between variables in psychological research is about 0.30. If five predictors in a population are each correlated 0.30 with the criterion, as well as 0.30 with each other, then the population regression coefficient for each predictor will be non-zero and equal in value to the other predictors. The sample size necessary to obtain statistical power of 0.80 for a significance test of a regression coefficient in this scenario is about 420, a sample size

that is larger than many studies use. Maxwell (2000) reported a simulation study in which a multiple regression analysis was conducted using the above scenario, but with a sample size of only 100. He found that the most frequently occurring pattern of results, occurring 45% of the time, was the case where one predictor had a statistically significant regression coefficient, but the other four did not. The next most common pattern, occurring 32% of the time, was that two of the predictors had statistically significant regression coefficients, but three did not. Thus, in a situation where each of the five predictors is of equal import in the population (that is, they all have the same correlation with the outcome and they all have the same regression coefficient), there was a high probability that only one or two of the predictors would exhibit statistical significance. Which predictors showed a statistically significant coefficient among the five predictors was essentially random. Results such as these should give theorists using smaller sample sizes pause about declaring a variable "unimportant" if it receives a statistically non-significant regression coefficient. For relative importance analysis, we generally recommend against using theoretically uninformed statistical significance tests as a screen for predictor inclusion in the analysis.

APPROACHES TO DETERMINING RELATIVE IMPORTANCE

Standardized Regression Coefficients

One common strategy used by researchers to index the relative importance of predictors is to use *standardized regression coefficients*. The larger the absolute value of the standardized coefficient for a predictor, the more important is that predictor.[1] The use of standardization presumably equates the metrics of each predictor because after standardization each has a mean of 0 and a standard deviation of 1.0. A person with a score of 1.0 means the person scored one standard deviation above the mean of the predictor. A person with a score of -0.5 means the person scored half a standard deviation below the mean. And so on. The common metric, the argument goes, makes it possible to compare the magnitude of regression coefficients.

One problem with this approach is that using standard deviations to equate metrics is dubious (Judd, McClelland & Ryan, 2009; Blanton & Jaccard, 2006). A hypothetical but somewhat "tongue-in-cheek" example clarifies this. Consider the adage "an apple a day keeps the doctor away." Suppose we analyze the number of times people see a doctor per year (mean = 2.5 and SD = 3) as a function of the number of apples and oranges they eat per week and obtain the following regression equation:

[1] Instead of the absolute value, some researchers use the square of the coefficient, which also has the effect of removing the sign of the coefficient

Number of visits = 3.0 + -1.0 Apples + -.25 Oranges.

Based on this equation, an apple a week decreases the number of visits by 1, whereas an orange a week decreases the number of visits by 0.25 (on average): It takes four oranges to accomplish what one apple accomplishes. Suppose the SD for apples is 0.50 and for oranges it is 1.0. By the logic of standard scores, half an apple is substantively equivalent to one orange because half an apple is one SD above its mean and one orange also is one SD above its mean. This so called "equating" seems arbitrary and it can distort the estimated effects of apples and oranges on doctor visits. For example, the standardized regression equation for these data is

Standardized number of visits = 0.0 + -0.17 Apples + -.08 Oranges.

and it appears, on the surface, that about two oranges has the same effect as one apple (i.e., $-0.17 \approx -0.08*2$). So, which is it? According to just the SDs, one orange is analogous to half an apple because both are 1 SD equivalents. For the unstandardized equation, four oranges equals one apple in terms of its effect on doctor visits. For the standardized equation, two (standardized) oranges equals one (two standardized) apple in terms of its effect on (standardized) doctor visits. Can we really justify using SDs to equate apples and oranges? The SDs for them seem arbitrary and the standardization masks the true effects as reflected by the unstandardized equation. Extending the argument to more substantively grounded constructs, if our predictors of adolescent delinquency are parental rule setting and maternal education, is a one standard deviation difference between people for rule setting the same as a one standard deviation difference on maternal education, metric wise? We doubt it.

Richards (1982) shows that the standardized regression coefficient for a predictor in a multiple regression analysis is not only impacted by its standard deviation but also by the standard deviations of other predictors in the equation as well (which is not the case for unstandardized coefficients). Willit, Singer and Martin (1998) note that when different samples are taken from the exact same population, their SDs can be different (due to sampling error) and these differences will affect standardized coefficients but not the unstandardized coefficients. Darlington (1968) notes that in the presence of moderately to highly correlated predictors, standardized regression coefficients tend to exaggerate the relative importance of the predictor variable most highly correlated with the dependent variable and diminish the relative importance of other variables in the model. Using standardized regression coefficients to discern relative importance is problematic.

Correlations or Squared Correlations

Another index of relative importance is the absolute zero order correlation of each predictor with the outcome or, alternatively, the square of the correlation. Darlington (1990) argues that the zero order correlation is more appropriate to use than the squared correlation, but this is controversial (see Johnson & LeBreton, 2004). The obvious weaknesses of both indices is they ignore the overlapping explained variance with the other predictors in the equation due to correlations among predictors.

Squared Semi-part Correlations

Another index of relative importance is the squared semi-part correlation between a predictor and the outcome holding constant all other predictors in the equation. Suppose we have three predictors, X1, X2 and X3, and Y is our outcome. The squared semi-part correlation for X1 is how much the squared correlation increases when Y is predicted from X1, X2 and X3 as opposed to being predicted from just X2 and X3. It represents the incremental explained variance in the outcome due to X1. The squared semi-part correlation for X2 is how much the squared correlation increases when Y is predicted from X1, X2 and X3 as opposed to being predicted from just X1 and X3. It represents the incremental explained variance in the outcome due to X2. The squared semi-part correlation for X3 is how much the squared correlation increases when Y is predicted from X1, X2 and X3 as opposed to being predicted from just X1 and X2. It represents the incremental explained variance in the outcome due to X2. This index is sometimes referred to as the *usefulness* of a predictor. The index is problematic because it ignores all the explained variance that is common to the predictors, focusing only on unique variance. Many methodologists feel that the common explained variance should be strategically factored in as well, but how best to do so is controversial (Grömping, 2015). A general mathematical property that is often sought for indices of relative importance is that the sum of the importance scores across all predictors equal the overall squared multiple correlation of the full equation (i.e., the index represents a form of variance decomposition). Squared semi-part correlations do not have this property.

Platt Index

Another index of relative importance is the Pratt index (Pratt, 1987). This index multiplies the standardized regression coefficient for a predictor by the correlation of the predictor with the outcome. Although this may seem odd, the index has interesting properties. The correlation can be thought of as an index of the “total effect” of the predictor on the outcome, including both unique effects the predictor has as well as

effects that are common to the other predictors (i.e., common explained variance). The standardized regression coefficient for a predictor reflects a “partialled effect” because it is the estimated effect of the predictor on the outcome holding constant all other predictors. The Platt index is, in essence, the total effect of the predictor weighted by its unique effect net the other predictors. Interestingly, the sum of the Platt values across all predictors will equal the overall squared multiple correlation, a desirable property for an index of relative importance.

The Platt index has been criticized on several grounds. First, all of the limitations described for standardized regression coefficients apply to the Platt index because it is based, in part, on standardized coefficients. Second, the Platt value for a predictor can be zero or near zero even when its total effect is substantial, which is somewhat counter-intuitive. If the correlation with the outcome for a predictor is large but the standardized regression coefficient is small due to collinearity, the relative importance value for the predictor will be small. More generally, if one of the components of the Platt product index is low, the index downweights considerably the contribution of the other component. Third, the correlation and regression coefficient for a predictor must be equal in sign or else a negative Platt value results, which is nonsensical. The index is intended to reflect the contribution of the predictor to the overall squared R and the lowest value it should take is zero. Sign reversals between correlations and regression coefficients typically occur in the presence of suppression dynamics (see Cohen, Cohen, West & Aiken, 2003). However, there are cases where suppressor effects can be subtle, leaving investigators scratching their heads about what to make of a negative Platt value. For example, suppose we have five predictors, each correlated 0.30 with one another. Suppose the correlations between the predictors and the outcome are X1 = 0.40, X2 = 0.50, X3 = 0.20, X4 = 0.40, X5 = 0.50. The standardized regression coefficient for X3 in this case will equal -0.10, so its Platt index will be negative.

Stepwise Regression

Another approach to gaining perspectives on the relative importance of predictors uses stepwise regression. It yields a rank order index of relative importance. With this method, a regression equation is formed consisting of one predictor and then predictors are successively added (or removed) in “steps” until a criterion (described below) is reached. The first predictor that is entered into the equation is the one with the highest zero order correlation with the outcome. The second predictor is the one that yields the largest increase in the squared multiple correlation relative to the predictor already in the equation. Once the second predictor is entered, a test of significance is performed to determine if the coefficient for the first predictor remains statistically significant. If not, it

is dropped, This sequential adding of predictors followed by re-evaluation of statistical significance of entered predictors continues until either all predictors have been entered into the equation or until the only remaining predictors do not produce a statistically significant increase in the squared multiple correlation. The variables in the final equation are designated as “important;” the variables that did not enter the final equation are deemed “unimportant.” The order in which variables entered the final equation determines the relative importance of the predictor, with earlier entry reflecting greater import. For a detailed description of the method, see Cohen et al., (2003).

This method has fallen into disrepute. One major problem is that the significance tests associated with the method are biased. They do not use the appropriate degrees of freedom (see Thompson, 1995, for details) and the statistical theory underlying such tests in intractable (Cohen et al., 2003). Relatedly, the predictor selected for entry at a given step is conditional on the variance contributions of the predictors already entered into the equation. Different entry orders will occur depending on the predictor that enters the equation first. The variable that enters the equation first can be heavily impacted by sampling error. Predictor X2 might be slightly more highly correlated with the outcome than predictor X1 in the population, but because of sampling error, predictor X1 might be slightly more highly correlated with the outcome than predictor X2 in the sample data. Given that results at subsequent steps are dependent on the predictors entered at prior steps, the results of stepwise analysis can be different purely from sampling error. This can be overcome by using large N, but doing so does not solve other problems with the method including the misleading nature of the p values and significance tests (Thompson, 1995), bias in the regression coefficients (Tibshirani, 1996), and a general failure of the approach to accurately identify variables in the true generating function for outcomes (Derksen & Keselman, 1992; Mantel, 1970). We do not recommend this approach.

Orthogonalization Methods

Yet another class of indices to assess the relative importance of predictors uses orthogonalization strategies, the most popular of which is a method known as relative weight analysis (Johnson, 2000; Johnson & LeBreton, 2004; see also Genizi, 1993). In multiple regression, when all the predictors are uncorrelated, the standardized regression coefficient for a given predictor will equal its correlation with the outcome and the squared multiple correlation will equal the sum of the square of the correlations between each predictor and the outcome. Orthogonalization methods transform the predictors (which are typically correlated) so that they are uncorrelated and satisfy these properties. More specifically, relative weight analysis uses principal components analysis as the basis for score transformation, generating a predicted score for each individual for each

orthogonal principal component (see the primer on factor analysis for details). These *component scores,* as a collective, retain their predictive power of the outcome. The component scores are then subjected to two forms of regression analysis: (1) an analysis that regresses the outcome onto each of the component scores, and (2) an analysis that regresses each predictor (separately) onto the component scores. The relative importance weight for a given predictor is a function of the squared regression coefficients in the first analysis and the squared regression coefficients in the second analysis (see Johnson, 2000, for details). The relative importance scores when summed across predictors will equal the squared multiple correlation. The scores usually are divided by the squared R and then multiplied by 100 so that the relative importance metric ranges from 0 to 100. A given score represents the percent of contribution of the predictor to the squared multiple correlation.

A weakness of the relative weight approach is that its results can be influenced by the type of orthogonalization method used. A strength is that it often yields results very similar to dominance analysis (probably the best method available – see below) but it can be used with a large number of predictors, which is not the case for dominance analysis.

A related orthogonalization method is the CAR index (Zuber & Strimmer, 2011; CAR is an abbreviation for "correlation-adjusted (marginal) correlations"). This approach achieves uncorrelated predictor representations using what is known as the Mahalanobis transform (Genizi, 1993), the technical details of which are beyond the scope of this primer. Interested readers should consult Zuber and Strimmer (2011). The CAR index for a predictor is the correlation between the outcome and the decorrelated predictors. When these correlations are squared and summed across predictors, the result will equal the overall squared multiple correlation for the full equation.

Dominance Analysis

The final approach we discuss is that of dominance analysis, which is the method described in the main text. It is probably the best of the many approaches, but the other methods might be of interest depending on how one chooses to define relative importance. Dominance analysis uses an index of importance called *general dominance* and is based on the average increase in R^2 for all subset models of equal size that include the predictor in question, X, relative to models that do not include it. The indices reflect the average unique explained variance contribution of X to the outcome across all possible subsets of independent variables. General dominance indices sum to the total R^2. Dominance analysis is the same as an approach suggested by Lindeman, Merenda and Gold (1980), so sometimes you will see it referred to as the LMG method.

As an example, with 3 predictors, (X1, X2, and X3), one can calculate the increase

in R square that X1 yields over and above X2, that X1 yields over and above X3, and that X1 yields over and above X2 and X3 together. The average of these increases is the index of general dominance for X1. This is repeated for each predictor to yield that predictor's dominance index. For the 4 predictor case, (X1, X2, X3, and X4), one calculates the increase in R square that X1 yields over and above X2, that X1 yields over and above X3, that X1 yields over and above X4, that X1 yields over and above X2 and X3 together, that X1 yields over and above X2 and X4 together, that X1 yields over and above X3 and X4 together and that X1 yields over and above X2, X3, and X4. The average of these increases is the index of general dominance for X1. Obviously, computations are intense for large numbers of predictors. As an example, for 20 predictors, the dominance method may take as long as 10 minutes to finish on a high-speed computer. Dominance analysis has the desirable property that the sum of the dominance scores across the predictors equals the full equation squared multiple correlation. It takes into account both total and unique effects and maps well onto rank orders of predictors that use fit indices other than the squared multiple correlation, such as Akaike's information criterion and the Bayesian information criterion (Azen & Budescu, 2003).

In addition to general dominance, dominance analysis has been extended to specify other forms of dominance (Azen & Budescu, 2003). For example, a predictor *completely dominates* another predictor if across all submodels of the same size the former predictor always shows a larger unique variance contribution than the latter predictor. *Conditional dominance* of one predictor over another predictor occurs if the average incremental variance for the former predictor is larger than the latter predictor across all submodels of the same subset size. Grömping (2015) has identified several limitations of these more fine grained indices and recommends just working with the general dominance indices.

THE PROBLEM OF SAMPLING ERROR

The importance indices are, of course, subject to sampling error. It is helpful to calculate a confidence interval for the index for each predictor to provide a sense of the amount of sampling error that is operative. Given the complexity of the underlying statistical theory to do so, bootstrap methods typically are used to estimate such confidence intervals. The ASA software we recommend provides this option as well as an option to conduct formal statistical tests of the differences between the indices for any two predictors (see the worked examples document).

Parenthetically, if you are willing to forego confidence intervals, then you can apply the relative importance methods to latent variable regression models that use structural equation modeling. To do so, you would use a standard SEM program to obtain an estimate of the covariance matrix among all the relevant latent variables. Then use the

ASA program option that uses the covariance matrix as input and enter the matrix for the latent variables. The importance indices will be generated based on this covariance matrix, with confidence intervals omitted.

COVARIATES

Sometimes we want to compute importance indices for a set of target predictors while holding constant one or more covariates. For example, we might want to evaluate the relative importance of different beliefs about drinking alcohol on the intent to drink among recovering addicts, holding gender constant. In traditional regression analysis, such covariates are included in the prediction equation. However, we may want the relative importance indices computed only for a subset of the predictors, namely the (covariate adjusted) beliefs about drinking. Only a few of the approaches to indexing importance can accommodate this case. Dominance analysis is one of them.

BINARY OUTCOMES

When an outcome variable is binary, it is common to apply logistic or probit regression instead of ordinary least squares multiple regression. Analytic strategies for logit models have been developed for both dominance analysis and relative weight analysis (Azen & Traxel, 2009; Tonidandel & LeBreton, 2010). The ASA software does not include these options. However, logistic regression as an approach to analyzing binary outcomes has limitations (see the Appendix to this primer). An alternative is to analyze the data using a modified linear probability model with bootstrapped significance tests and confidence intervals, which is equivalent to using the standard regression modeling for dominance analysis and relative weight analysis offered by ASA (see the video accompanying this primer). For statistical details, see the Appendix.

CONCLUDING COMMENTS

A common question asked when conducting multiple regression analyses concerns the relative importance of the predictors in influencing or predicting the outcome. This is a difficult question to answer because much depends on what one means by the term “importance.” Some researchers use statistical significance as a basis for identifying important predictors but this strategy can be undermined by the use of sample sizes that yield low statistical power. Seven commonly used methods for ordering predictors in terms of their relative importance include standardized regression coefficients, zero order correlations, squared semi-part correlations, the Platt index, stepwise regression, orthogonalization methods, and dominance analysis. Of these methods, dominance

analysis is probably the most sound but it can be challenging to apply when there are a large number of predictors. When conducting dominance analysis, you often will want to take into account sampling error as well as covariates that serve as nuisance variables that are important to control.

REFERENCES

Angrist, J. & Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Azen, R., & Traxel, N. M. (2009). Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics,* 34, 319–347.

Blanton, H. & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, 61, 27-41.

Cohen, J., Cohen, P., West, S. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69, 161-182.

Derksen, S. & Keselman, H. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45, 265–282.

Genizi, A. (1993). Decomposition of R^2 in multiple regression with correlated regressors. *Statistica Sinica*, 3, 407–420.

Grömping, U. (2015). Variable importance in regression models. *Computational Statistics*, 7, 137-152.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35, 1-19.

Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods*, 7, 238-257.

Judd, C.M., McClelland, G.H. & Ryan, C.S. (2009). *Data analysis: A model comparison approach*. New York: Routledge.

Lindeman, R.H., Merenda P.F. & Gold, R.Z. (1980). *Introduction to bivariate and multivariate analysis*. Glenview, Illinois: Scott-Foresman.

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics*, 12, 621–625.

Maxwell, S. (2000). Sample size and multiple regression analysis. *Psychological Methods*, 5, 434- 458.

Mooney, C. & Duval, R. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Thousand Oaks, CA: Sage.

Pratt, J.W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In T. Pukkila and S. Puntanen (Eds.): *Proceedings of second Tampere conference in statistics.* University of Tampere, Finland.

Richards, J. (1982). Standardized versus unstandardized regression weights. *Applied Psychological Measurement*, 6, 201-212.

Rosenthal, R. (1995). Methodology. In. A. Tesser (Ed.), *Advanced social psychology*. (pp.17-50) New York: McGraw-Hill.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55, 525-534.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B 58: 267–288.

Tonidandel, S. & LeBreton. J. (2010). Determining the relative importance of predictors in logistic regression: An extension of relative weight analysis. *Organizational Research Methods*, 13, 767-781.

Willit, J., Singer, J. & Martin, N. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology*, 10, 395-426.

Zuber, V. & Strimmer, K. (2011). High-dimensional regression and variable selection using CAR scores. *Statistical Applications in Genetics and Molecular Biology*, 10, .

APPENDIX: BINARY REGRESSION AND RELATIVE IMPORTANCE

In this appendix, we assume you are familiar with the basics of logistic regression. We begin by reviewing the difference between odds and probabilities, to help put logit regression in context. Next, we discuss the concept of conditional probabilities as a way of introducing logistic regression and the linear probability model. Finally, we discuss marginal effects in binary regression and then develop the implications of marginal effects for relative importance analysis.

Odds and Probabilities

The fundamental construct of interest in binary regression is a probability, such as the probability of contracting a disease, the probability of being fired from a job, the probability of experiencing an unintended pregnancy, or the probability of getting married. Probabilities range from 0.00 (impossible) to 1.00 (certain). In binary regression they apply to populations of individuals. If we say that the probability that a 16 year old female adolescent living in the United States will experience an unintended pregnancy while she is 16 is 0.08, then this means that for this particular population of individuals, 8% of them experience an unintended pregnancy and 92% of them do not.

Another way of expressing a probability is using odds. We convert a probability to an odds by dividing it by 1 minus the probability in question. If the probability of 50 year old men in the United States seeing a doctor in the ensuing 12 months is 0.667, then the probability of not doing this is 1 - .667 = 0.333. The ratio of these two probabilities is the odds; 0.667 / 0.333 = 2.0, or in more common vernacular, the odds are “2 to 1” or it is twice as likely that 50 year old men living in the United States will see a doctor in the next 12 months than they won’t.

An odds can be less than one. If the probability of a teenager smoking marijuana is 0.20, the odds of a teenager smoking marijuana is 0.20/0.80 = 0.25. The odds value of 0.25 means the probability of smoking marijuana is one fourth the probability of not smoking marijuana. If the probability of a Black man having a college degree is 0.25, the odds of a Black man having a college degree is 0.25./0.75 = 0.33. The odds value of 0.33 means the probability of having a college degree is one third that of not having a college degree.

Most of us are more comfortable with the concept of probability than odds, as the former are easily converted to percentages and we have more experience with them. Characterizing events in terms of probabilities is not “better” than using odds; they are just different ways of expressing the same thing.

One can convert a probability to an odds by applying the formula P(Y) / (1-P(Y)), where P(Y) refers to the probability of Y. One can convert an odds to a probability by the

formula Odds(Y) / (1 + Odds(Y)). Also, the mean of a dichotomous variable scored 0-1 will equal the proportion of scores that have a 1, which can be thought of as the probability of observing a 1.

Binary Regression and the Modeling of Conditional Probabilities

Suppose we are interested in modeling how the probability of some event occurring, Y, changes as a function of the values of one or more predictor variables, X. The most common regression models for doing so focus on conditional probabilities: For a given predictor profile, we seek to specify what the probability of Y is. This conditional probability is represented as P(Y | X), or "the probability of Y given a value of X." The symbol | is read as "given that."

Consider the simple bivariate case where the outcome is the probability an adolescent will smoke marijuana in the ensuing year and the predictor is the age of adolescents, ranging from 12 to 17. We want to characterize what the probability of Y is for adolescents who are age 12, what it is for adolescents who are age 13, what it is for adolescents who are age 14, and so on (actually, age is a continuous variable, but for pedagogical reasons, we frame it here as discrete). Each age represents a different predictor "profile" and we seek to characterize P(Y | X) at each profile, i.e. P(Y | X=12), P(Y | X= 13), P(Y | X=14), and so on.

If we have multiple predictors, then a "profile" refers to a specific combination of scores across the predictors. For example, if we predict the probability of marijuana use from age and gender, then one predictor profile is "12 year old males," another predictor profile is "12 year old females," and so on. We might seek to estimate the conditional probability P(Y | Age=12, Gender = Male), the probability P(Y | Age=12, Gender = Female), and so on.

We can express how the probability of Y, in this case smoking marijuana, varies as a function of the predictor, age, using a simple linear equation:

$$P(Y_i) = \alpha + \beta X_i$$

where $P(Y_i)$ is the probability that the specified group of adolescents smokes marijuana in the next 12 months and X_i is the age group "i" of interest (e.g., 12 years old adolescents; 13 year olds, and so on). As an example, suppose the (population) probabilities of smoking marijuana are as follows:

Age	P(Y)
12	0.025
13	0.050
14	0.075
15	0.100
16	0.125
17	0.150

The probability of smoking marijuana is 0.025 conditional on age being 12. The probability of smoking marijuana is 0.050 conditional on age being 13. And so on. The intercept for this model is -0.275 and the slope is 0.025. Note for every one unit age increases, the probability of smoking marijuana increases by 0.025 units. The intercept is meaningless in this case because it refers to an age (the probability of Y when age = 0) that is outside the range of X values.

Because the above relationship between age and the probability of smoking marijuana is linear, a reasonable model for analyzing the data is called the *linear probability model*, which explicitly assumes a linear relationship. An early strategy for implementing the linear probability model (LPM) was to use standard ordinary least squares (OLS) regression with a dichotomous outcome variable. This strategy capitalizes on the fact that the mean of a dichotomous variable scored 0-1 equals the proportion of scores that have a 1. The OLS approach is problematic, however, because for standard errors and confidence intervals to be correct, OLS requires (1) the population error scores for a given predictor profile be normally distributed (which is not the case for a dichotomous outcome), and (2) that the population error scores have equal variance across different predictor profiles (which also is not the case). One way of circumventing these parametric assumptions is to use bootstrapping to estimate confidence intervals and to conduct significance tests on the coefficients (see Mooney & Duval, 1993, for an introduction to bootstrapping). When such bootstrapping is used or when alternative forms of robust estimation are employed, the LPM is referred to as a *modified linear probability model.*

Logistic regression does not model probabilities, but rather odds. To be more precise, it models the log of odds, not odds. Let the odds of Y be represented by Odds(Y). Recall Odds(Y) is the P(Y) divided by one minus this probability. The logistic model relates the log of the odds to X as follows:

$$\text{Log}[\text{Odds}(Y_i)] = \alpha + \beta X_i$$

where log is the natural logarithm. This model posits that the log of the odds of Y is a linear function of X, whereas the linear probability model states that the probability of Y, not the log odds of Y, is a linear function of X. By modeling the log of odds as a linear function of X, the logistic model implies a non-linear relationship between (continuous) X and the probability of Y. Consider our prior example with age and the probability of smoking marijuana, where we convert the probabilities to log odds:

Age	P(Y)	Odds(Y)	ln(Odds(Y))
12	0.025	0.0256	-3.665
13	0.050	0.0526	-2.945
14	0.075	0.0811	-2.512
15	0.100	0.1111	-2.197
16	0.125	0.1429	-1.946
17	0.150	0.1765	-1.734

Note in this case, that age is *not* linearly related to the log odds of Y. The logistic model is misspecified and not appropriate for these data. If age was linearly related to the log odds of Y, then it would not be linearly related to P(Y). In this sense, the logistic model is often called a non-linear model. As will be seen, the non-linear nature of logistic regression introduces challenges if your primary focus is on probabilities.

Marginal Effects

Statisticians often make use of a concept known as a *marginal effect* in binary regression. A marginal effect is the rate of change for the probability of an outcome given a one unit increase in a predictor. Consider the following linear probability model predicting marijuana use from gender and age:

$$\text{Prob(Marijuana)} = \alpha_1 + \beta_1 \, G + \beta_2 \, \text{Age}$$

For gender, the rate of change in the probability of Y is equal to β_1 when gender is dummy coded. Note that this model assumes that the effect of gender on the probability of Y is the same at each age level and also equals the value of β_1. If this were not the case, we would need to include an interaction term between gender and age. As well, if we average these gender effects for each age group, we also will obtain the value of β_1. The marginal effect for gender is thus β_1.

Similarly, the marginal effect for age equals the value of β_2. The effect of age on the probability of Y is assumed to be the same for males and females, otherwise we would

need an interaction term. If we average the age effects for males and for females, we will obtain the value of β_2. The marginal effect for gender is thus β_1.

For logistic regression, the marginal effect is more complicated because logistic regression models log odds not probabilities. It turns out that in a logit analysis that regresses marijuana smoking onto gender and age, the effect of gender on the *probability* of smoking marijuana will differ depending on the age of the adolescent even though we do not include an interaction term between gender and age in the model. We thus cannot derive the marginal effect for gender from the logistic coefficients. Nevertheless, researchers often are interested in knowing the value of such marginal effects.

We can give an intuitive sense of how one calculates a marginal effect in logistic regression using gender from the above example. First, we calculate the logistic equation using standard statistical software. Next, we consider the first case in the data set. We treat that person as a male irrespective of what that person's gender actually is. The person's scores on all the other predictor variables in the equation are left alone but we set the gender score to "male" and then calculate the predicted probability of smoking marijuana for that person based on that person's predictor profile in the derived logistic equation. Essentially, we assume the person has the same scores s/he had on the other predictors but that s/he is a male irrespective of his or her actual gender. We then repeat this process, but this time we presume the person is female irrespective of the person's actual gender. We calculate the predicted probability of the outcome for this person using the original logistic equation under this scenario. The difference between the two calculated probabilities is the marginal effect for that particular individual. We repeat this process for every individual in the sample, calculating a marginal effect for each one. Finally, we compute the average of all these individualized marginal effects. The result is the marginal effect. In essence, we are comparing two populations, one that is all male and one that is all female, but where each population has the same distribution of values on the other predictors in the model. Because the only difference between the two populations is their gender, the logic goes, gender is the source of the differences in their likelihood of having smoked marijuana. Hence, this is the marginal effect for gender.

For continuous predictors, such as age, the same logic is used, but one calculates how the probability of the outcome changes given a one unit change in the continuous predictor based on each observed value of age while all other predictors are left equal to their observed values. The average of these individual changes is the marginal effect.

It turns out, that the marginal effect for a predictor in a logistic regression usually is very close in value to what the marginal effect for that predictor is in the linear probability model. Angrist and Pischke (2009) argue that, given this, it often is much simpler to calculate marginal effects using the modified linear probability model rather

than engaging in all the gymnastics of calculating the marginal effects in logistic regression. One basically will get the same results, they argue, but more efficiently.

Implications for Relative Importance Analysis

The implications of the above for relative importance analysis is that one usually can just use traditional linear regression to model binary outcomes given a focus on outcome probabilities (which is usually what we are interested in). The importance indices generally derived as such generally will be unbiased and, because their confidence intervals and significance tests are based on bootstrapping, the violation of population assumptions will not be problematic. One usually will not need to use specialized logistic models for relative importance analysis.