Some Graphical Methods for Interpreting Interactions in Logistic and OLS Regression

Peter L. Flom **Sheila M. Strauss**
National Development and Research Institutes, Inc.

In statistical models involving one dependent variable (DV) and two or more independent variables (IVs), an interaction occurs when the effect of one IV on the DV is different at different levels of another IV. The existence of an interaction makes interpretation of the model more complicated, but failing to include important interactions in the model can give misleading results. In this paper, we describe how visually examining interactions between two IVs in ordinary least squares regression and in logistic regression can aid comprehension of the interaction, and we present a tool to make such examination easier.

In modeling the relationship between a dependent variable (DV), Y, and a set of independent variables (IVs), X_1, X_2, ….X_k, if the DV is continuous and we are using ordinary least squares regression we have:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \quad (1)$$

When the DV is dichotomous, OLS regression is inappropriate; probably the most common alternative is using logistic regression (Hosmer & Lemeshow, 2000) where we have

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + ...\beta_k x_k + \varepsilon \quad (2)$$

where $\pi(x)$ is the conditional mean of Y given X. Given that Y is dichotomous, this is the same as the probability that Y = 1 (assuming that Y is coded 0, 1), or the probability of a 'success' if Y is coded 'failure/success'. The portion before the equals sign is known as the logit. Equivalently, we can model $\pi(x)$ directly as

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + ... \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + ... \beta_k x_k}} \quad (3)$$

Models (1.1) and (1.2), and, more generally, any of the class of generalized linear models (McCullagh & Nelder, 1989), assume (among other things) that the effect of each X_i (i = 1, 2, …, k) on Y is the same, regardless of the value of the other X_j (j = 1, 2, …, k; $i \neq j$); that is, that there is no interaction. We may suspect that this is not the case. Earlier research may have found interactive effects; or we may have other substantive reasons for suspecting interactions. For example, if we are examining the likelihood of being HIV positive based on a person's sex and sexual identity, we would include an interaction between the two, since the effect of being homosexual is greater for males than for females. This is usually done by adding an interaction term to the equation. While there are many possible ways to construct such a term, the most usual, and simplest, is to multiply the two IVs that we think may be involved by each other, and add that term (Harrell, 2001). For now, let us suppose (for simplicity) that our model contains only two independent variables: X_1 and X_2. Set $X_3 = X_1 X_2$ and add it to the model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ , \quad (4)$$

which allows for an interaction.

While many researchers recognize the importance of evaluating interactions, interpreting interactions is often difficult and sometimes counterintuitive. This can be so for at least two reasons. First, because the parameter for interaction is multiplied by both IVs, what appears to be a small coefficient can actually be highly meaningful, especially if one of the IVs is large. Second, when the signs of the parameters on the interaction term and the IVs are not all the same, the effect of a change in one of the IVs on the DV may not be readily apparent.

The use of graphics can facilitate the interpretation and presentation of interactions; indeed, graphics can facilitate interpretation of the results of logistic regression even in the absence of interactions (Long, 1997). While many guides to model building, variable selection, and significance testing are available (e.g. Harrell (2001), this is not the case for graphical methods of interpreting interactions. In this paper,

we therefore present some graphical methods for displaying and interpreting interactions between two independent variables, at least one of which is continuous. We provide methods for both continuous and dichotomous DVs. These graphs allow the user to examine the effect of a change in the IVs on the DV directly, without any computations. While these methods are not novel, they are under-utilized, and we are unaware of any source, which describes all of the methods described below, or their application specifically to interaction. These methods all involve plotting curves; for ordinary least squares regression, each of these curves is derived from a relatively simple equation; for logistic regression, the curves are considerably more complex. Graphs of this type have two types of uses: First, they may aid one's own analysis of data. Second, they may allow easier presentation of these findings to others.

Data

Data for this paper are drawn from the Drug Use and HIV Risk Among Youth (DUHRAY) project. DUHRAY involved a probability survey of 18-24 year old household-recruited youth in Bushwick, a low-income minority neighborhood in Brooklyn, New York with a population of approximately 100,000 in 1995. It sampled two groups of Bushwick-resident young adults: (1) a population-representative multistage household probability sample; (2) a targeted sample (Watters & Biernacki, 1989) of youth who use heroin, cocaine, crack, or inject drugs. Details of the sampling plan are available elsewhere (Flom et al., 2001).

Graphical Methods

The best method of interpreting interactions depends on the nature of the variables involved in the model. Specifically, it depends on whether the DV and IVs are dichotomous or continuous. In this section, we first present methods for a variety of types of models with two IVs, where at least one of the IVs is continuous. When both IVs are dichotomous, graphical methods are not necessary, and crosstabulations can be very useful. This is so because, in a 2 x 2 crosstabulation, it is relatively straightforward to determine main effects and interactions by hand calculation. We then present some possible extensions to cases where there are more than two IVs.

Continuous DV, One Continuous and One Dichotomous IV

In a regression model having a single DV and a single IV, a scatterplot of the IV and the DV is often useful. When we add a dichotomous IV to the model, we can make a scatterplot with two lines, one for each level of the dichotomous IV.

Example 1: In DUHRAY, we created a variable for peer objection to drug use (DROBJ), based on a factor analysis of five questions, each of which asked what proportion of your friends would object if you used a particular drug. The five drugs were marijuana, cocaine, heroin, crack, and injected drugs. We also asked about recalled childhood misbehavior, using a scale based on one devised by Windle (1993). We then modeled peer objection to drug use as a function of childhood misbehavior and sex (1 for male, 2 for female), using ordinary least squares regression, and including an interaction term. The estimated equation was

$$\mathbf{DROBJ = -0.14 - 0.0038\ Win + 0.91\ SEX - 0.036\ SEX*Win} \quad (5)$$

This yields Figure 1, from which it can be seen that, while objection to drug use decreases as reported childhood misbehavior increases, it does so faster for women than for men. Also, while women, on average, reported more objection to drug use than men did, (the mean for men was -0.10, *SD* = 0.97; for women mean = 0.15, *SD* = 1.00) the opposite was true when there was a lot of reported misbehavior. If there were no interaction, the lines would be parallel.

Dichotomous DV, One Continuous and One Dichotomous IV

Example 2: If the DV is dichotomous, we simply use the results of logistic regression rather than OLS regression to create a graphical display. For example, we modeled using hard drugs (HD) (cocaine, heroin, crack, and/or injected drugs) in the last year (yes=1, no=0) as a function of sex (male = 1, female = 2) and childhood misbehavior (win). The estimated equation was:

$$prob(HD) = \frac{e^{.354-.247*sex+.351*win+.0638*sex*win}}{1+e^{.354-.247*sex+.351*win+.0638*sex*win}} \quad (6)$$

Figure 1: Objection to drugs as a function of sex and childhood misbehavior

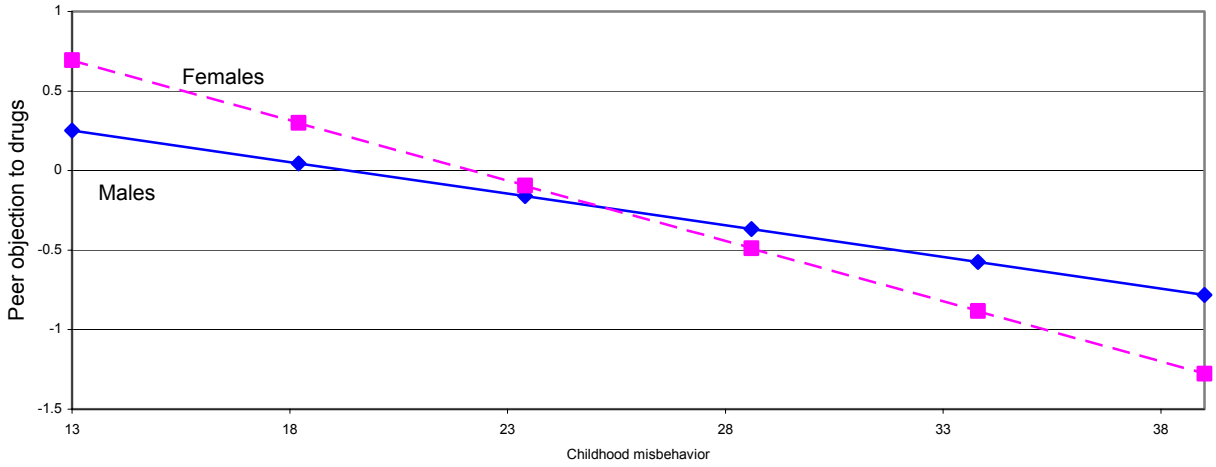


Figure 2: Probability of hard drug use as function of Windle, and sex

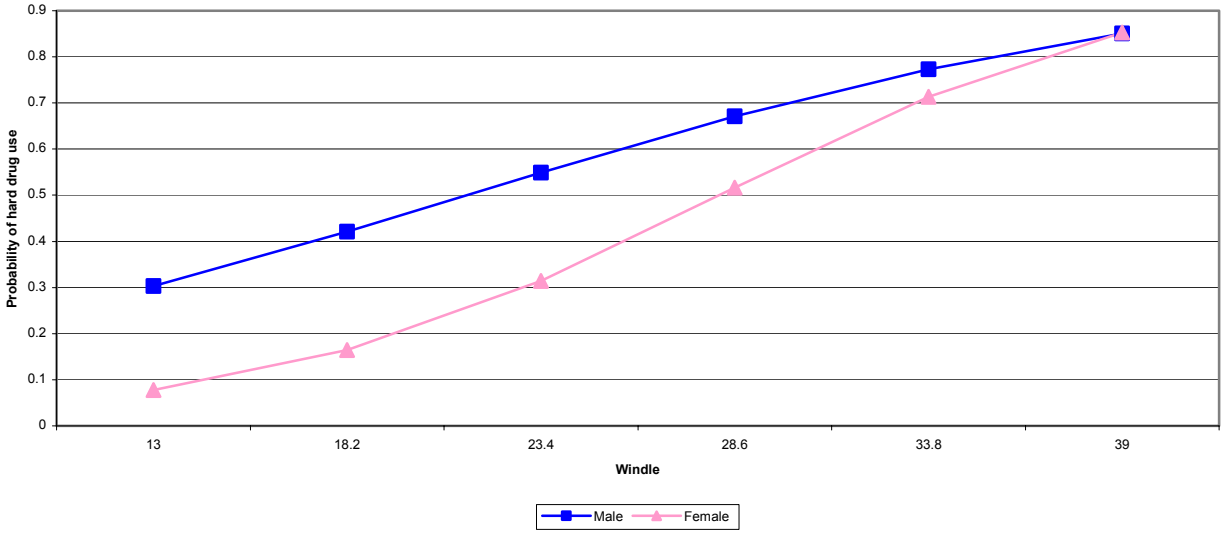


Figure 3: Objection to drugs as function of age (X) and Windle

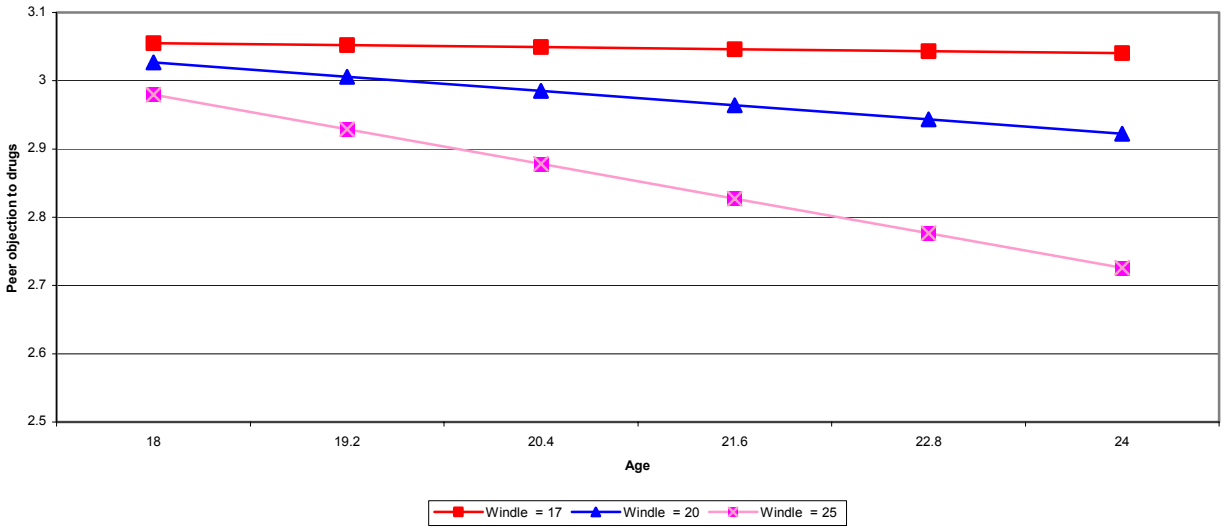
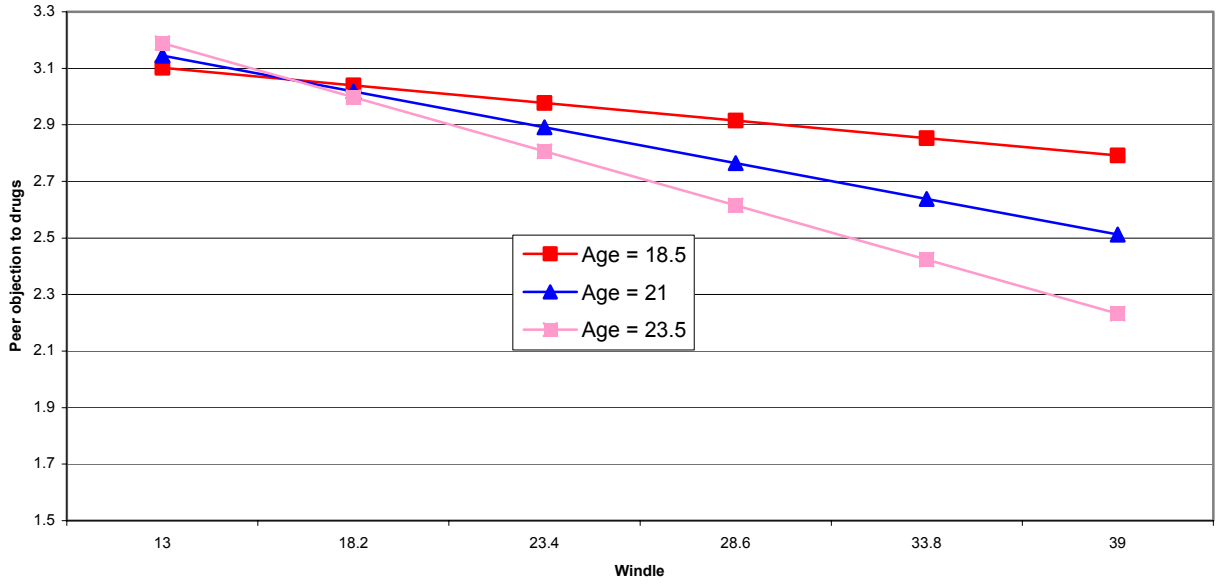


Figure 4: Objection to drugs as function of and Windle (X) and age



This yields Figure 2. If there were no interaction, the two curves would be parallel. As is, we can see signs of a moderate interaction. Although the likelihood of having used hard drugs increases with childhood misbehavior for both males and females, it increases faster for females.

Continuous DV, Two continuous IVs

When both IVs are continuous, there is no statistical reason for deciding which IV to put on the X-axis. There may be substantive reasons for choosing one, but we may need two plots to get a full sense of the interaction. One plot will have one IV on the X axis; the other plot will have the other. Also, in this case, we need to use more than two lines on the scatterplot. The exact number of lines depends on the distribution of each variable, but three is often a good compromise between comprehensibility and completeness. More lines can clutter the page, and fewer lines give an incomplete picture of the changes in the relationship. We need to pick representative values of each IV. One possible set of choices (used below) is the 25th, 50th, and the 75th percentiles.

For example, we modeled peer objection to drug use as a function of age and childhood misbehavior. The estimated equation was:

$$\textbf{DROBJ} = \textbf{-1.48 + .12*AGE + 1.24*Windle -.063*AGE*Windle} \quad (7)$$

First, we let age be the X variable, and pick the three values of Windle at the 25th, 50th, and 75th percentiles. This yields Figure 3 from which it can be seen that, while objection to drug use decreases as age increases, it does so much faster for those who reported more childhood misbehavior.

On the other hand, if we let Windle be the X variable, and choose values of age at the 25th, 50th, and 75th percentiles, we get Figure 4, from which it can be seen that the relationship between childhood misbehavior and peer objection to drug use is stronger for older subjects.

Dichotomous DV, Two Continuous IVs

If the DV is dichotomous, we again modify the above procedure by using logistic regression. We modeled the probability of having used hard drugs in the last year (HD), by age and childhood misbehavior. The estimated equation was:

$$prob(HD) = \frac{e^{5.33-.41*age-.44*win+.027*age*win}}{1+e^{5.33-.41*age-.44*win+.027*age*win}} \quad (8)$$

With age on the X axis, this yields Figure 5. This indicates that the relationship between age and drug use is stronger for subjects with higher levels of childhood misbehavior (because the slope of the line for Windle = 25 is greater than that for lower values of Windle). If there were no interaction, the lines would be parallel.

Similarly, if we put misbehavior on the X axis, and make separate lines for different ages scores, we get Figure 6. This implies that the relationship between childhood misbehavior and peer objection to drug use is stronger for older subjects.

Discussion and Conclusions

In this paper, we have presented a tool for displaying the effects of interactions involving two independent variables. While this tool is not innovative, making it more widely known and more easily implemented will, we believe, increase understanding about the nature of an interaction. In addition, it has the potential to clarify how changes in various parameters in logistic and ordinary least squares regression affect the relationship between a dependent variable and two independent variables.

While most statistical software (including SPSS, S-Plus, R, or SAS-Graph would allows production of charts similar to those produced in this paper, the graphics presented here were developed in Microsoft EXCEL. Because this software is readily available, allows the user to adjust figures and immediately see the results, and requires no programming skill, utilizing the graphics approach developed here with Microsoft EXCEL will be possible for many data analysts. It should be noted that Excel was not used to calculate the equations, but only to plot the results.

Figure 5: Probability of hard drug use as function of age (X), and Windle

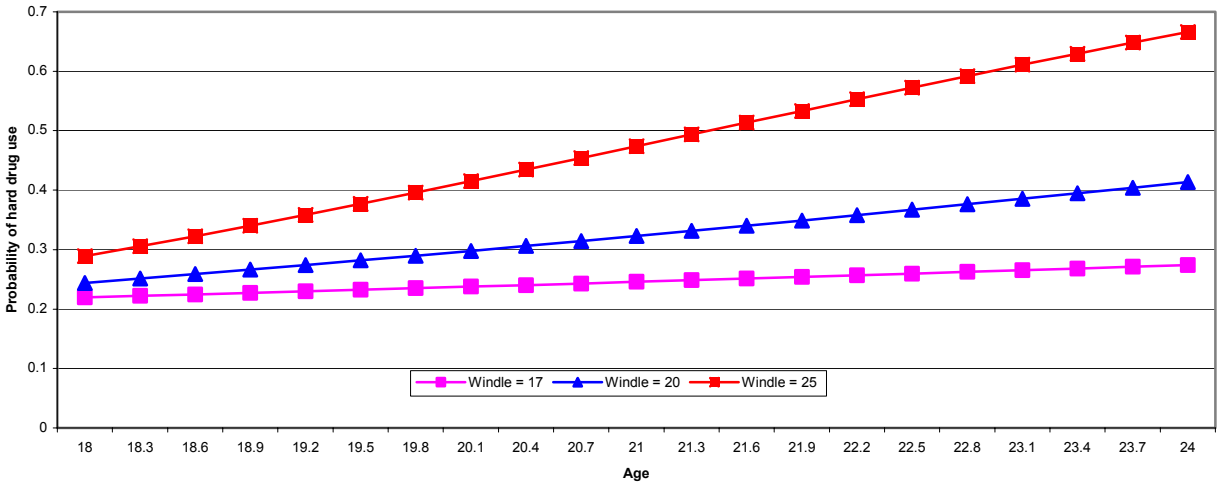
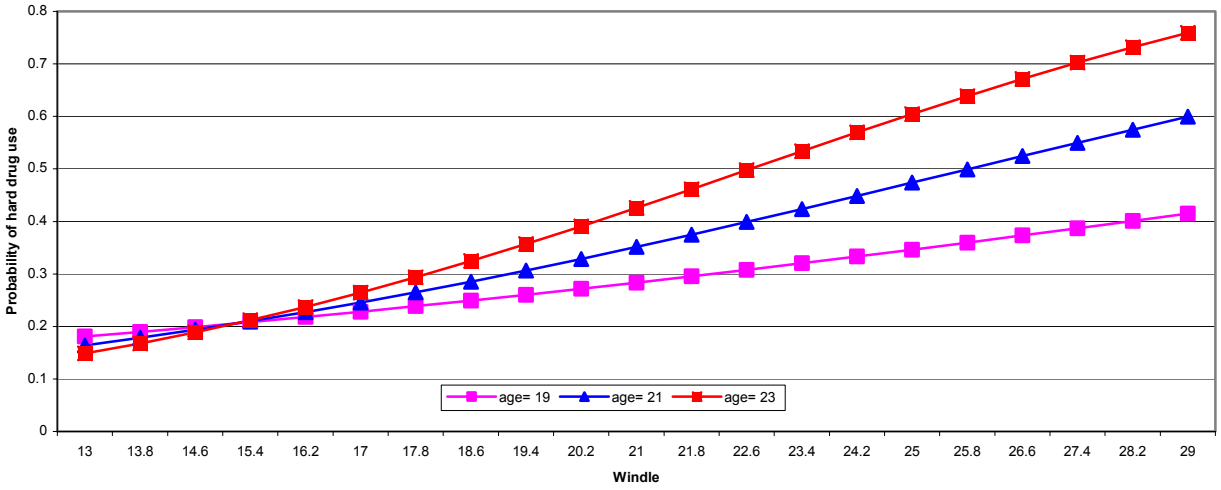


Figure 6: Probability of hard drug use as function of age and Windle(X)



References

Flom, P. L., Friedman, S. R., Kottiri, S. R., Neaigus, A., Curtis, R., Des Jarlais, D. C., Sandoval, M., & Zenilman, J. M. (2001). Stigmatized drug use, sexual partner concurrency, and other sexual risk network and behavioral characteristics of 18-24 Year old youth in a high-risk neighborhood. *Sexually Transmitted Diseases*, *28*, 598-607.

Harrell, F. E. (2001). *Regression modeling strategies*. New York: Springer-Verlag.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: John Wiley and Sons.

Long, J. S. (1997). *Regression models of categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Boca Raton, FL: Chapman & Hall.

Watters, J., & Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, *6*, 416-430.

Windle, M. (1993). A retrospective measure of childhood behavior problems and its use in predicting adolescent problem behaviors. *Journal of Studies on Alcohol*, *54*, 422-431.

Author Notes: This work was supported by NIDA grants R01 DA10411 and P30 DA 11041.

The figures in this paper were all produced in Excel. In each workbook, the worksheet contains a top section, where the user may modify the parameters, ranges of values, and so on, a bottom section, which the user should not edit, and one or two charts, which are linked to the sheet, but which may be further edited by the user. The Excel file are available at the following website:

Send correspondence to: Peter L. Flom, National Development and Research Institutes, Inc.
71 W. 23rd St., 8th Floor
New York, NY 10010
Email: flom@ndri.org