I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10530

Identification and Decompositions in Probit and Logit Models

Chung Choe
SeEun Jung
Ronald L. Oaxaca

JANUARY 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10530

Identification and Decompositions in Probit and Logit Models

Chung Choe
Hanyang University

SeEun Jung
Inha University

Ronald L. Oaxaca
University of Arizona, LISER, IZA and PRESAGE

JANUARY 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Identification and Decompositions in Probit and Logit Models

Probit and logit models typically require a normalization on the error variance for model identification. This paper shows that in the context of sample mean probability decompositions, error variance normalizations preclude estimation of the effects of group differences in the latent variable model parameters. An empirical example is provided for a model in which the error variances are identified. This identification allows the effects of group differences in the latent variable model parameters to be estimated.

JEL Classification: C35, J16, D81, J71
Keywords: decompositions, probit, logit, identification

Corresponding author:
Ronald Oaxaca
Department of Economics, Eller College of Management
University of Arizona
P.O. Box 210108
Tucson, AZ 85721-0108
USA
E-mail: rlo@email.arizona.edu

Introduction

The objective of decomposition methodology is to identify and estimate the separate contributions of differences in parameters and differences in characteristics when accounting for mean differences in outcome variables for two different population groups. Standard decomposition approaches appear in the context of a linear model. For nonlinear models some modification of the standard decomposition methodology is required.

Unlike when one applies the Oaxaca decomposition after OLS estimation, the estimated conditional expectations of outcome variables in nonlinear models generally are not equal to the predicted outcome values at mean characteristics. Fairlie (2005) addresses this issue and suggests a decomposition technique that allows for a detailed decomposition for the entire set of explanatory variables in the context of probit and logit models. Yun (2004) extends Fairlie's approach by proposing a method that is free from path-dependency. Later decomposition techniques are further extended to other models with discrete and limited dependent variables (Bauer and Sinning, 2008). Wolff (2012) proposes a decomposition method for non-linear models that employs simulated residuals. However, like others, the paper encounters a decomposition identification problem as the error variances in the Probit model are not identified and hence are normalized to one. The paper acknowledges the difficulty in finding a solution to the decomposition identification issue.

Our paper closely examines the decomposition identification problem in probit and logit binomial outcome models. This identification issue complicates the inferences one can draw from a decomposition into characteristics (explained) effects and parameter (unexplained) effects. The typical probit/logit model can be motivated along the lines of either a random utility model or a latent variable model. For illustrative purposes we will first examine the latent variable motivation in which a latent variable Y_i^* is defined by

$$Y_i^* = X_i\beta^* + \varepsilon_i^*,$$

where ε_i^* is distributed either $N(0, \sigma_\epsilon^2)$ or logistic$(0, \sigma_\epsilon^2)$. In the case of the logistic distribution $\sigma_\epsilon^2 = k^2\pi^2/3$, where k is a scale factor. Although we do not observe Y_i^*, we do observe the binary variable $Y_i = 1\,(Y_i^* > 0)$. Accordingly, the probit and logit models are obtained from

$$\begin{aligned}
Prob\,(Y_i = 1|X_i) &= Prob\,(X_i\beta^* + \varepsilon_i^* > 0) \\
&= Prob\,(\varepsilon_i^* > -X_i\beta^*) \\
&= Prob\,(\varepsilon_i^* < X_i\beta^*) \\
&= Prob\left(\frac{\varepsilon_i^*}{\sigma_\epsilon} < \frac{X_i\beta^*}{\sigma_\epsilon}\right) \\
&= Prob\,(\varepsilon_i < X_i\beta) \\
&= \Phi\,(X_i\beta) \text{ or } \Lambda\,(X_i\beta) \\
&= \Phi\,(I_i) \text{ or } \Lambda\,(I_i)\,,
\end{aligned}$$

where $I_i = (X_i\beta)$ is the index function, $\varepsilon_i = \dfrac{\varepsilon_i^*}{\sigma_\epsilon}$, $\beta = \dfrac{\beta^*}{\sigma_\epsilon}$, $\Phi\,(I_i)$ is the CDF for the standard normal distribution, and $\Lambda\,(I_i)$ is the CDF for the standardized logistic distribution.

Typically, the parameter σ_ϵ is not identified so it is normalized to 1 for the probit model, and the scale parameter k is normalized to 1 for the logit model ($\sigma_\epsilon = \pi/\sqrt{3}$). For most purposes these normalizations are innocuous. Unfortunately, in the context of decomposition analysis for probit and logit models these normalizations are not so innocuous. In decomposing mean differences in outcome probabilities for two populations, the natural objective would be to estimate how much group differences in the β^* parameters from the latent variable model contribute to the mean differences in outcome probabilities. Decompositions based solely on the estimated $\tilde{\beta}$ probit/logit parameters estimate the effects of group differences in $\dfrac{\beta^*}{\sigma_\epsilon}$ rather than in β^*. Consequently, group differences in the β^* parameters are confounded with group differences in the σ_ϵ parameters.

It is clear that unless the underlying theoretical latent variable or random utility model identifies the variance parameter, the decomposition ambiguity is present. On the other hand

identification could be achieved if the latent variable or random utility model contained a theoretical restriction in which one of the coefficients in the pre-normalized index function is unity. Unlike the normalization restriction on probit/logit error variances, this restriction grounded in theoretical reasoning would not be arbitrary.

Below we provide an example of a random utility model in which the probit/logit variance parameter is identified through a theoretical restriction that specifies that one of the coefficients in the index function is equal to 1.

Empirical Example

Although not the subject of a probit/logit decomposition, the mean-variance portfolio model in Jung et al. (2016) is an example in which the variance of the error in the index function is identified. The context for the portfolio model is one in which men and women in an experimental setting choose between a risky typing task characterized by exogenous spells of unemployment and a secure typing task not subject to unemployment spells. The risky job carries a risk premium for typing performance and unemployment compensation for spells of unemployment.

An individual's expected earnings from the risky job (y_{ri}) and the secure job (y_{si}) are determined according to

$$y_{ri} = \phi w_u + (1 - \phi)\gamma_r \psi_i$$

$$y_{si} = \gamma_s \psi_i,$$

where ϕ is the probability of unemployment, w_u is the amount of unemployment compensation, ψ_i is one's expected productivity, and γ_r and γ_s are the respective returns to productivity on the risky and secure jobs ($\gamma_r > \gamma_s$).

Similarly, the conditional (on ψ_i) variances of earnings from the risky job (σ^2_{ri}) and the

secure job (σ_{si}^2) can be shown to be determined according to

$$\sigma_{ri}^2 = (\phi)(1-\phi)(\gamma_r \psi_i - w_u)^2$$
$$\sigma_{si}^2 = 0.$$

The additive random utilities of the job gambles are expressed as

$$U_{si} = y_{si} + \varepsilon_{si} \text{ (secure job)}$$

$$U_{ri} = y_{ri} - \frac{\alpha}{2}\sigma_{ri}^2 + \varepsilon_{ri} \text{ (risky job)},$$

where α is the Pratt-Arrow measure of constant relative risk aversion and ε_{si} and ε_{ri} are independently and identically distributed extreme value disturbances. If we let $J_r = 1$ when the risky job is chosen (0 otherwise), the probability that one would select the risky job is given by

$$\begin{aligned}
Prob(J_{r_i} = 1) &= Prob\left(U_{ri} > U_{si}\right) \\
&= Prob\left(y_{ri} - \frac{\alpha}{2}\sigma_{ri}^2 + \varepsilon_{ri} > y_{si} + \varepsilon_{si}\right) \\
&= Prob\left(y_{ri} - y_{si} - \frac{\alpha}{2}\sigma_{ri}^2 > \varepsilon_{si} - \varepsilon_{ri}\right) \\
&= Prob\left(\frac{\varepsilon_{si} - \varepsilon_{ri}}{\sigma_\epsilon} < \frac{y_{ri} - y_{si}}{\sigma_\epsilon} - \frac{\alpha}{2\sigma_\epsilon}\sigma_{ri}^2\right) \\
&= Prob\left(\frac{\varepsilon_{si} - \varepsilon_{ri}}{\sigma_\epsilon} < I_i\right) \\
&= \Lambda\left(I_i\right),
\end{aligned}$$

where $\sigma_\epsilon = \sqrt{Var(\varepsilon_{si} - \varepsilon_{ri})}$, $I_i = \theta_1\left(y_{ri} - y_{si}\right) + \theta_2\left(\frac{-\sigma_{ri}^2}{2}\right)$, $\theta_1 = \frac{1}{\sigma_\epsilon} > 0$, and $\theta_2 = \frac{\alpha}{\sigma_\epsilon} \gtreqless 0$.

It follows that the logit standard deviation is identified from $\sigma_\epsilon = \frac{1}{\theta_1}$, and the variance from $\sigma_\epsilon^2 = \frac{1}{(\theta_1)^2}$. One can estimate α as $\tilde{\alpha} = \frac{\tilde{\theta}_2}{\tilde{\theta}_1}$. Furthermore, with this model one can directly compare the $\tilde{\alpha}_m$ and $\tilde{\alpha}_f$ risk aversion parameter estimates for males and females and

therefore identify the effect of the gender difference in the α's on the probability of choosing the risky job.

Unlike the case for probit, maximum likelihood estimation of a logit model with a constant term has the property that the sample proportion for the binary outcome variable is identical to the sample mean of the predicted probabilities. However this property does not hold in the present case because of the absence of an intercept term. Consequently, we add a remainder term to account for any deviations between sample proportions and mean probability predictions.

We can now proceed with the logit decompositions of the observed gender difference in the sample proportions of those choosing the risky job. The sample proportion and the mean predicted probability for a given group are defined by

$$\bar{P}_{rj} = \frac{\sum_{i=1}^{N_j} J_{r_i}}{N_j}$$
$$\tilde{P}_{rj} = \frac{\sum_{i=1}^{N_j} \Lambda(\tilde{I}_{ij})}{N_j},$$

where $\tilde{I}_{ij} = \tilde{\theta}_{1j}\left(y_{rij} - y_{sij}\right) + \tilde{\theta}_{2j}\left(\frac{-\sigma^2_{rij}}{2}\right)$, j=m,f. The sample proportion is equal to the mean predicted probability plus a remainder (δ_j):

$$\bar{P}_{rj} = \tilde{P}_{rj} + \delta_{rj}.$$

For a conventional decomposition one could define the counterfactual probability for women as

$$\tilde{P}^m_{rf} = \frac{\sum_{i=1}^{N_f} \Lambda(\tilde{I}^m_{if})}{N_f},$$

where $\tilde{I}^m_{if} = \tilde{\theta}_{1m}\left(y_{rif} - y_{sif}\right) + \tilde{\theta}_{2m}\left(\frac{-\sigma^2_{rif}}{2}\right)$. This counterfactual probability is an estimate

of the mean probability of choosing the risky job if women had the same θ parameters as the men but retained their own characteristics. Accordingly, the decomposition would be

$$\bar{P}_{rm} - \bar{P}_{rf} = \underbrace{\left(\tilde{P}_{rm} - \tilde{P}^{m}_{rf}\right)}_{\text{explained}} + \underbrace{\left(\tilde{P}^{m}_{rf} - \tilde{P}_{rf}\right)}_{\text{unexplained}} + \underbrace{\left(\delta_{rm} - \delta_{rf}\right)}_{\text{remainder}}.$$

The "unexplained" component of the decomposition measures the gender probability gap attributable to gender differences in <u>all</u> parameters, i.e. the α's and the σ_ϵ's. However, this decomposition does not answer the more interesting question of how much of the gender difference in the probability of choosing the risky job stems from gender differences in risk preferences.

Because the Pratt-Arrow constant relative risk aversion parameter α is identified, it is feasible to construct a decomposition that measures the contribution of gender differences in risk preferences to gender differences in the probability of choosing the risky job. Toward this end we construct a new counterfactual probability for women:

$$\tilde{P}^{\alpha_m}_{rf} = \frac{\sum_{i=1}^{N_f} \Lambda(\tilde{I}^{\alpha_m}_{if})}{N_f},$$

where $\tilde{I}^{\alpha_m}_{if} = \tilde{\theta}_{1f}\left(y_{rif} - y_{sif}\right) + \tilde{\theta}_{1f}\tilde{\alpha}_m\left(\frac{-\sigma^2_{rif}}{2}\right)$. This counterfactual probability is the estimated mean probability that women would choose the risky job if they had the same risk preference as the men but retained their own characteristics including their own logit error term variance. This scenario gives rise to the following decomposition:

$$\bar{P}_{rm} - \bar{P}_{rf} = \underbrace{\left(\tilde{P}_{rm} - \tilde{P}^{\alpha_m}_{rf}\right)}_{\text{endowments}} + \underbrace{\left(\tilde{P}^{\alpha_m}_{rf} - \tilde{P}_{rf}\right)}_{\text{risk preferences}} + \underbrace{\left(\delta_{rm} - \delta_{rf}\right)}_{\text{remainder}}.$$

The "risk preferences" component measures how much of the mean gender probability gap arises purely from gender differences in the α constant relative risk aversion parameter. The

"endowment" component of the decomposition measures how much the gender probability gap arises from a) gender differences in the difference between expected earnings in the risky job and the secure job, b) gender differences in the conditional variance of earnings from the risky job, and c) gender differences in the variance of the error term in logit model.

It is possible to further decompose the endowment component to identify the effects of gender differences in the standard deviation σ_ϵ of the logit error term. First, we introduce an additional counterfactual probability for women:

$$\tilde{P}_{rf}^{\sigma_{\epsilon m}} = \frac{\sum_{i=1}^{N_f} \Lambda(\tilde{I}_{if}^{\sigma_{\epsilon m}})}{N_f},$$

where $\tilde{I}_{if}^{\sigma_{\epsilon m}} = \tilde{\theta}_{1m}\left(y_{rif} - y_{sif}\right) + \tilde{\theta}_{1m}\tilde{\alpha}_f\left(\frac{-\sigma_{rif}^2}{2}\right)$. This counterfactual is the predicted mean probability for choosing the risky job if women had the same logit error term variance as the men but retained their own characteristics and risk preferences. We start with the previous decomposition and add and subtract the term $\left(\tilde{P}_{rf}^{\sigma_{\epsilon m}} - \tilde{P}_{rf}\right)$ to obtain

$$\bar{P}_{rm} - \bar{P}_{rf} = \underbrace{\left[\left(\tilde{P}_{rm} - \tilde{P}_{rf}^{\alpha_m}\right) - \left(\tilde{P}_{rf}^{\sigma_{\epsilon m}} - \tilde{P}_{rf}\right)\right]}_{\text{net endowment}} + \underbrace{\left(\tilde{P}_{rf}^{\sigma_{\epsilon m}} - \tilde{P}_{rf}\right)}_{\text{standard deviation}} + \underbrace{\left(\tilde{P}_{rf}^{\alpha_m} - \tilde{P}_{rf}\right)}_{\text{risk preferences}} + \underbrace{\left(\delta_{rm} - \delta_{rf}\right)}_{\text{remainder}}.$$

The "net endowment" component measures the contribution of gender differences in characteristics net of the effect of gender differences in the logit error term standard deviations, and the "standard deviation" component measures how much of the gender probability gap arises from gender differences in the logit error term standard deviation.

Table 1 reports the estimated logit job choice model. Nearly 75% of the males chose the risky job compared with about 60% of the females. Although the estimated logit parameters look quite different between males and females, only the estimated θ_1 parameter for males is statistically significant. This pattern carries over to the derived estimates of risk preference and the logit standard deviation: only the estimated standard deviation parameter for males

is statistically significant. A joint likelihood ratio test for equality of the logit parameters between males and females indicated that one could not reject the hypothesis of equal parameters, though barely (p value = 0.11). Undoubtedly, small sample sizes contributed to a lack of precision in estimating the logit models.

The implications of the estimated model for decomposition of the gender difference in the proportion who chose the risky job are presented in Table 2. The gender gap in sample probabilities is 15.2 percentage points. The first decomposition examines only the logit parameters without reference to identification of the risk preference parameter and the logit standard deviation. Only a negligible amount of the probability gap can be explained by gender differences in expected earnings and expected variance of earnings (less than 1%). On the other hand the unexplained component arising from gender differences in the estimated logit parameters is quite substantial at nearly 88%. The remainder difference is nearly 12% of the probability gap.

The second decomposition in Table 2 corresponds to the effects of gender differences in the identified risk preference parameter. The endowment effect accounts for nearly 87% of the gap while gender differences in the estimated risk preference parameter is quite small at 1.32% of the gap.

The third decomposition in Table 3 separates out the effect of gender differences in the logit standard deviation. The net endowment effect is quite modest at only about 2% of the gap. Accordingly, gender differences in the estimated logit standard deviation is substantial at nearly 85% of the probability gap. What this finding reveals is that what appeared to be a substantial unexplained gap based on the logit parameters derives from the gender difference in the error standard deviations and not gender differences in the risk aversion parameters.

One can readily generalize from the above example. When a joint test of group differences in all of the logit parameters reveals that one cannot reject the null hypothesis that there are no group differences in the parameters, then one can conclude that all of the outcome gap arises from differences in the characteristics (explained) plus differences in the remainder

term. This is the case whether or not the standard deviation parameter is identified. On the other hand if one can reject the null of equal parameters, then one cannot conduct the decomposition without identification.

The same analysis can be applied to a probit model. Indeed the mean-variance portfolio model in Jung et al. (2016) was estimated as a probit[1]. Empirically, probit and logit results are quite similar. As is usually the case with decompositions in general, decompositions for probit/logit are not unique because counterfactuals are not unique. For example we could have constructed counterfactuals in which the estimated parameters for women are applied to the characteristics of men. Nevertheless, the identification challenge and solution strategy remain the same.

References

Bauer, T. K. and M. Sinning (2008). An extension of the blinder–oaxaca decomposition to nonlinear models. *AStA Advances in Statistical Analysis 92*(2), 197–206.

Fairlie, R. W. (2005). An extension of the blinder-oaxaca decomposition technique to logit and probit models. *Journal of Economic and Social Measurement 30*(4), 305–316.

Jung, S., C. Choe, and R. L. Oaxaca (2016, August). Gender Wage Gaps and Risky vs. Secure Employment: An Experimental Analysis. IZA Discussion Papers 10132, Institute for the Study of Labor (IZA).

Wolff, F.-C. (2012). Decomposition of non-linear models using simulated residuals. *Economics Letters 116*(3), 346–348.

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

[1]The probit model was estimated by pooling the male and female samples and interacting a gender indicator variable with the conditional variance of earnings for the risky job.

Table 1: Logit Job Choice Model

	Males	Females
$\tilde{\theta_1}$	3.073***	1.200
	(0.886)	(0.896)
$\tilde{\theta_2}$	0.168	0.071
	(0.188)	(0.233)
N	103	89
Chi2	21.38	3.63
$\tilde{\alpha}$	0.055	0.060
	(0.049)	(0.160)
$\tilde{\sigma}_\epsilon$	0.325***	0.834
	(0.094)	(0.623)
Risky Job Choice	77	53
$\bar{P}_r$	0.748	0.596
$\bar{P}_{rm} - \bar{P}_{rf}$	0.152	

* p < 0.1, ** p < 0.05, *** p < 0.01

Table 2: Decompositions

Decomposition 1		
Explained	0.001	0.66%
Unexplained	0.133	87.50%
Remainder	0.018	11.84%
$\bar{P}_{rm} - \bar{P}_{rf}$	0.152	
Decomposition 2		
Endowment	0.132	86.84%
Risk Preferences	0.002	1.32%
Remainder	0.018	11.84%
$\bar{P}_{rm} - \bar{P}_{rf}$	0.152	
Decomposition 3		
Net Endowment	0.003	1.97%
Standard Deviation	0.129	84.87%
Risk Preferences	0.002	1.32%
Remainder	0.018	11.84%
$\bar{P}_{rm} - \bar{P}_{rf}$	0.152	