

Potential Problems with Using Heterogeneous Choice Models To Compare Logit and Probit Coefficients across Groups

[Note: you should read my May 2009 Sociological Methods and Research Paper, “Using Heterogeneous Choice Models To Compare Logit and Probit Coefficients Across Groups.” We’ve already covered several of the points raised in that paper. In this handout, I’ll just focus on some of the potential problems with heterogeneous choice models, and the ways in which oglm can be used to minimize (but not eliminate) concerns.]

Overview. Allison (1999) notes that comparisons of logit and probit coefficients across groups can be invalid and misleading. He proposes a procedure by which these problems can be corrected, and argues that “routine use [of this method] seems advisable” and that “it is hard to see how [the method] can be improved.” In my SMR paper I show that Allison’s method involves a special case of the heteroskedastic logit model. I argue that, as originally proposed, this method can have serious problems and should not be applied on a routine basis. However, I also note that the heteroskedastic logit model is part of a larger class of models variously known as heterogeneous choice or location-scale models. By turning to this broader and more flexible class of models, Allison’s procedure can be greatly improved.

The Problem. Allison (1999; also see Hoetker 2004) argues that we are often interested in making comparisons across groups (e.g. men versus women, blacks versus whites). **HOWEVER**, when doing logistic regression, there is a potential pitfall in cross-group comparisons that, Allison claims, has largely gone unnoticed. Unlike linear regression coefficients, coefficients in these binary regression models are confounded with residual variation (unobserved heterogeneity). Differences in the degree of residual variation across groups can produce apparent differences in coefficients that are not indicative of true differences. We have already discussed that elsewhere, so let’s look more closely at Allison’s proposed solution.

Allison’s example: Allison illustrates his point via the analysis of a data set of 301 male and 177 female biochemists (for a detailed description of the data, see Long, Allison and McGinnis 1993; the description provided here is adapted from Allison’s 1999 paper). These scientists were assistant professors at graduate universities at some point in their careers. Allison uses logistic regressions to predict the probability of promotion to associate professor. The units of analysis are person-years rather than persons, with 1,741 person-years for men and 1,056 person-years for women

TABLE 1: Results of Logit Regressions Predicting Promotion to Associate Professor for Male and Female Biochemists

Variable	Men		Women		Ratio of Coefficients	Chi-Square for Difference
	Coefficient	SE	Coefficient	SE		
Intercept	-7.6802***	.6814	-5.8420***	.8659	.76	2.78
Duration	1.9089***	.2141	1.4078***	.2573	.74	2.24
Duration squared	-0.1432***	.0186	-0.0956***	.0219	.67	2.74
Undergraduate selectivity	0.2158***	.0614	0.0551	.0717	.25	2.90
Number of articles	0.0737***	.0116	0.0340**	.0126	.46	5.37*
Job prestige	-0.4312***	.1088	-0.3708*	.1560	.86	0.10
Log likelihood	-526.54		-306.19			

* $p < .05$. ** $p < .01$. *** $p < .001$.

- The effect of # of articles on promotion is about twice as great for males (.0737) as it is for females (.0340). If accurate, this difference suggests that men get a greater payoff from their published work than do females, a conclusion that many would find troubling.
- BUT, this difference could be an artifact of differences in the residual variances. Women may have more heterogeneous career patterns, and unmeasured variables affecting the chances for promotion may be more important for women than for men. If the residual variance for women is greater, the female coefficients will be lowered.

Allison's procedure for dealing with the problem is as follows. *NOTE: Steps 1 and 2 can be estimated with Stata's built-in commands. The later steps require user-written software. See if you can replicate steps 1 and 2. Begin with the commands*

```
use "http://www.indiana.edu/~jlsoc/stata/spex_data/tenure01.dta", clear
keep if pdasample
```

- Step 1 (Shown in Table 1 above): Separate logistic regression models are estimated for each group (which Allison numbers 0 and 1). This allows the coefficients for all variables to differ across groups. The log likelihoods from the two separate models are added together. This is equivalent to running a pooled model with a dummy variable for group membership and group membership interaction terms for all variables. In Allison's Table 1, the pooled log-likelihoods are $-526.54 + -306.19 = -832.73$.

- Step 2: A model (without interaction terms) is estimated for the entire sample. A dummy variable for group membership is included. This constrains all coefficients (except the intercepts) to be the same for both groups. Allison does not show this model in his paper but he reports (p. 194) that its log-likelihood is -838.53.

- Step 3 (shown in the first half of Table 2 below): A parameter called *delta* is added to the model from Step 2. Let $G_i = 0$ for men (group 0) and 1 for women (group 1). The underlying model (Allison 1999:192) then becomes

$$y_i^* = \alpha_0 + \alpha_1 G_i + \sum_{j>1} \alpha_j x_{ij} + \sigma_i \varepsilon_i$$

The formula for the scaling factor sigma is then

$$\sigma_i = \frac{1}{1 + \delta G_i}$$

Equivalently, the formula for delta is

$$\delta = \frac{1 - \sigma_{Group1}}{\sigma_{Group1}}$$

Under Allison’s procedure, once this is done, σ_{Group0} is fixed at 1 while σ_{Group1} is free to vary. Delta, then, is an estimate of how much the disturbance standard deviation differs by group. So, for example, a delta of 1 would indicate that the disturbance variance of group 0 is 100% higher (i.e. double) the disturbance variance of group 1. A delta of -.5 indicates that the disturbance variance for group 0 is only half as large as the variance for group 1. A delta of 0 means that there are no differences in residual variation across groups. By including delta in the model, the differences in residual variation that distort cross-group comparisons are presumably controlled for. Note that this model continues to be estimated under the assumption that the underlying Alphas for both groups are equal. As shown below, the log likelihood for this model is -836.28.

- Step 4: A series of hypotheses are then tested; and based on these results, additional models may be estimated (as Allison (1999:195) notes, these tests should ideally be done “with some sort of correction for multiple comparisons”, e.g. Bonferroni adjustments).
 - Step 4A. First, you test the null hypothesis that the Alpha coefficients are the same but the residual variances differ (i.e. delta = 0). This involves a chi-square contrast between the models from steps 2 and 3. Note that this test is done *under the assumption that the underlying Alphas for both groups are equal*. As we will see, this is a critical and potentially problematic assumption. In Allison’s analysis, the likelihood ratio chi-square is 4.5 with 1 d.f., which is statistically significant (Allison explains the calculation on p. 194).
 - Step 4B. Second, if the residual variances are found to differ, a global test is done of whether any Alphas differ across groups. This involves a chi-square contrast between the models of Step 3 and Step 1. A significant chi-square value supposedly indicates that at least one coefficient differs across groups, even after controlling for differences in residual variation. Allison (p. 195) reports a LR chi-square of 7.1 with 4 d.f. (This is not significant, but Allison proceeds to the next step anyway to finish demonstrating the procedure.)

- Step 4C. Third, if it is found that at least one coefficient differs across groups, additional models can be estimated to identify the specific variables whose effects differ. This is done by adding interaction terms to the model from Step 3 and doing a chi-square contrast with the Step 3 model. (This is shown in the second half of Table 2 below.) In order to conduct these tests, however, *at least one set of coefficients must be assumed to be equal across groups*. A model with all possible group interactions and a parameter for differences in residual variation would not be identified. Again, we will see that this is a critical and potentially problematic assumption. When Allison tests whether the effect of number of articles differs by gender, he gets a likelihood ratio chi-square of 2.30 with 1 d.f. (p. 196).
- When Allison applies his procedure, the results are as follows:

TABLE 2: Logit Regressions Predicting Promotion to Associate Professor for Male and Female Biochemists, Disturbance Variances Unconstrained

Variable	<i>All Coefficients Equal</i>		<i>Articles Coefficient Unconstrained</i>	
	<i>Coefficient</i>	<i>SE</i>	<i>Coefficient</i>	<i>SE</i>
Intercept	7.4913***	.6845	-7.3655***	.6818
Female	-0.93918**	.3624	-0.37819	.4833
Duration	1.9097***	.2147	1.8384***	.2143
Duration squared	-0.13970***	.0173	-0.13429***	.01749
Undergraduate selectivity	0.18195**	.0615	0.16997***	.04959
Number of articles	0.06354***	.0117	0.07199***	.01079
Job prestige	-0.4460***	.1098	-0.42046***	.09007
$\hat{\delta}$	-0.26084*	.1116	-0.16262	.1505
Articles \times Female			-0.03064	.0173
Log likelihood	-836.28		-835.13	

* $p < .05$. ** $p < .01$. *** $p < .001$.

- The delta-hat coefficient value $-.26$ tells us that the standard deviation of the disturbance variance for men is 26 percent lower than the standard deviation for women. Ergo, women have more variable career patterns than do men, which causes their coefficients to be lowered relative to men when differences in variability are not taken into account, as in the original logistic regressions.
- The interaction term for Articles \times Female is NOT statistically significant, i.e there is no statistically significant differences if the effects of # of articles between men and women
- Allison concludes “The apparent difference in the coefficients for article counts in Table 1 does not necessarily reflect a real difference in causal effects. It can be readily explained by differences in the degree of residual variation between men and women.”

oglm replication of Allison's analysis. Glenn Hoetker has written a routine called `complot` that will estimate Allison's complete set of models. However, this can also easily be done with the user-written `oglm` (Williams 2006) which is more powerful and flexible. Here is how you can replicate Allison's 4 steps. [I will let you go over these on your own].

```
. * Step 1. Unconstrained models, all coefficients can differ by gender.
. * oglm cloning of Allison's Table 1 - results are same, except that cutpoints
. * are the negatives of the intercepts. You could also just use the
. * the logit command, but using oglm throughout makes it easier to
. * compare models
```

```
. use "http://www.indiana.edu/~jlsoc/stata/spex_data/tenure01.dta", clear
(Gender differences in receipt of tenure (Scott Long 06Jul2006))
```

```
. * Allison limited the sample to the first 10 years untenured
. keep if pdasample
(148 observations deleted)
```

```
. * Males Only
. oglm tenure year yearsq select articles prestige if male, store(step1male)
```

```
Ordered Logistic Regression                               Number of obs   =       1741
                                                         LR chi2(5)      =       302.42
                                                         Prob > chi2     =       0.0000
Log likelihood = -526.54503                               Pseudo R2      =       0.2231
```

tenure	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
year	1.908854	.2141114	8.92	0.000	1.489203	2.328504
yearsq	-.1432235	.018604	-7.70	0.000	-.1796866	-.1067604
select	.2157736	.0614129	3.51	0.000	.0954066	.3361406
articles	.0736935	.0115747	6.37	0.000	.0510076	.0963794
prestige	-.4311864	.1088151	-3.96	0.000	-.6444601	-.2179128
/cut1	7.680158	.6813939	11.27	0.000	6.344651	9.015666

```
. * Females Only
. oglm tenure year yearsq select articles prestige if female, store(step1fem)
```

```
Ordered Logistic Regression                               Number of obs   =       1056
                                                         LR chi2(5)      =       114.60
                                                         Prob > chi2     =       0.0000
Log likelihood = -306.19084                               Pseudo R2      =       0.1576
```

tenure	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
year	1.407772	.2572597	5.47	0.000	.9035524	1.911992
yearsq	-.0955919	.0219033	-4.36	0.000	-.1385215	-.0526622
select	.0551333	.0716526	0.77	0.442	-.0853033	.1955699
articles	.0339525	.012608	2.69	0.007	.0092413	.0586638
prestige	-.3707873	.1560405	-2.38	0.017	-.6766211	-.0649536
/cut1	5.841983	.8658648	6.75	0.000	4.144919	7.539047

```
. * Equivalent pooled model, using interactions.
. oglm tenure year yearsq select articles prestige f_year f_yearsq f_select
f_articles f_prestige female, store(step1)
```

```
Ordered Logistic Regression                               Number of obs   =       2797
                                                         LR chi2(11)    =       420.19
                                                         Prob > chi2    =       0.0000
Log likelihood = -832.73587                               Pseudo R2      =       0.2015
```

tenure	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
year	1.908854	.2141114	8.92	0.000	1.489203	2.328504
yearsq	-.1432235	.018604	-7.70	0.000	-.1796866	-.1067604
select	.2157736	.0614129	3.51	0.000	.0954066	.3361406
articles	.0736935	.0115747	6.37	0.000	.0510076	.0963794
prestige	-.4311864	.1088151	-3.96	0.000	-.64446	-.2179128
f_year	-.5010812	.3347032	-1.50	0.134	-1.157087	.1549251
f_yearsq	.0476316	.0287378	1.66	0.097	-.0086935	.1039567
f_select	-.1606402	.0943697	-1.70	0.089	-.3456014	.024321
f_articles	-.0397409	.0171153	-2.32	0.020	-.0732864	-.0061955
f_prestige	.0603992	.190235	0.32	0.751	-.3124545	.4332529
female	1.838174	1.101826	1.67	0.095	-.3213647	3.997712
/cut1	7.680158	.6813938	11.27	0.000	6.344651	9.015666

```
. * Step 2. Pooled model; only the intercepts differ by gender.
. * Allison refers to this model but does not present it in the paper.
```

```
. oglm tenure year yearsq select articles prestige female, store(step2)
```

```
Ordered Logistic Regression                               Number of obs   =       2797
                                                         LR chi2(6)     =       408.59
                                                         Prob > chi2    =       0.0000
Log likelihood = -838.53294                               Pseudo R2      =       0.1959
```

tenure	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
year	1.723243	.1638446	10.52	0.000	1.402113	2.044372
yearsq	-.125346	.0141187	-8.88	0.000	-.1530182	-.0976738
select	.1544061	.0460389	3.35	0.001	.0641714	.2446408
articles	.0548251	.008573	6.40	0.000	.0380223	.0716278
prestige	-.413615	.0882546	-4.69	0.000	-.5865908	-.2406393
female	-.3537514	.1320848	-2.68	0.007	-.6126329	-.0948698
/cut1	6.812655	.5290576	12.88	0.000	5.775721	7.849589

```
. * Step 3. Residual variances allowed to differ by gender.
. * Allison's model is actually a special case of a heterogeneous
. * choice model, and it is easy to compute Allison's delta using oglm.
. * Compare these results with the first half of Allison's Table 2.
```

```
. oglm tenure female year yearsq select articles prestige , het(female) store(step3)
```

```
Heteroskedastic Ordered Logistic Regression      Number of obs   =      2797
LR chi2(7)                                       =      413.09
Prob > chi2                                       =      0.0000
Pseudo R2                                        =      0.1981
Log likelihood = -836.28235
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	

tenure						
female	-.9391907	.3705243	-2.53	0.011	-1.665405	-.2129763
year	1.909544	.1996935	9.56	0.000	1.518152	2.300936
yearsq	-.1396868	.0169425	-8.24	0.000	-.1728935	-.1064801
select	.1819201	.0526572	3.45	0.001	.0787139	.2851264
articles	.0635345	.010219	6.22	0.000	.0435055	.0835635
prestige	-.4462073	.096904	-4.60	0.000	-.6361356	-.2562791

lnsigma						
female	.3022305	.146178	2.07	0.039	.0157268	.5887341

/cut1	7.490506	.6596628	11.36	0.000	6.19759	8.783421

```
. * Compute delta
. display (1 - exp(.3022305))/ exp(.3022305)
-.26083233
```

```
. * Step 4A. Test that the Alphas are = but residual variances differ.
. lrtest step2 step3, stats
```

```
Likelihood-ratio test      LR chi2(1) =      4.50
(Assumption: step2 nested in step3)  Prob > chi2 =      0.0339
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
step2	2797	-1042.828	-838.5329	7	1691.066	1732.62
step3	2797	-1042.828	-836.2824	8	1688.565	1736.055

```
. * Step 4B. Test whether any Alphas differ across groups given that
. * residual variances differ.
. lrtest step1 step3, stats
```

```
Likelihood-ratio test      LR chi2(4) =      7.09
(Assumption: step3 nested in step1)  Prob > chi2 =      0.1311
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
step3	2797	-1042.828	-836.2824	8	1688.565	1736.055
step1	2797	-1042.828	-832.7359	12	1689.472	1760.707

```
. * Step 4C. Test whether the effect of articles differs across groups.
. * First have to estimate the model with the interaction term added.
. * Compare this with the second half of Allison's Table 2.
```

```
. oglm tenure female year yearsq select articles prestige f_articles, het(female) store(step4C)
```

```
Heteroskedastic Ordered Logistic Regression      Number of obs   =      2797
                                                  LR chi2(8)      =      415.39
                                                  Prob > chi2     =      0.0000
Log likelihood = -835.13347                    Pseudo R2       =      0.1992
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	

tenure						
female	-.3780597	.4500207	-0.84	0.401	-1.260084	.5039646
year	1.838257	.2029491	9.06	0.000	1.440484	2.23603
yearsq	-.1342828	.017024	-7.89	0.000	-.1676492	-.1009165
select	.1699659	.0516643	3.29	0.001	.0687057	.2712261
articles	.0719821	.0114106	6.31	0.000	.0496178	.0943464
prestige	-.4204742	.0961206	-4.37	0.000	-.6088671	-.2320813
f_articles	-.0304836	.0187427	-1.63	0.104	-.0672185	.0062514

lnsigma						
female	.1774193	.1627087	1.09	0.276	-.141484	.4963226

/cut1	7.365286	.6547121	11.25	0.000	6.082073	8.648498

```
. * Compute delta
. display (1 - exp(.1774193))/ exp(.1774193)
-.16257142
```

```
. * Now do the formal test of the female*articles interaction term.
. lrtest step3 step4c, stats
```

```
Likelihood-ratio test      LR chi2(1) =      2.30
(Assumption: step3 nested in step4c) Prob > chi2 =      0.1296
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
step3	2797	-1042.828	-836.2824	8	1688.565	1736.055
step4c	2797	-1042.828	-835.1335	9	1688.267	1741.694

Following is a side by side comparison of the Step 3 results as presented by Allison (first half of Allison Table 2) and the replication by oglm.

Table 3: Comparison of Allison & Heteroskedastic Logit Models

	Allison	Heteroskedastic Logit
Got promoted (1 = yes, 0 = no)		
female	-0.939* (0.37)	-0.939* (0.37)
Duration	1.910*** (0.20)	1.910*** (0.20)
Duration squared	-0.140*** (0.017)	-0.140*** (0.017)
Undergraduate selectivity	0.182*** (0.053)	0.182*** (0.053)
Number of articles	0.0635*** (0.010)	0.0635*** (0.010)
Job Prestige	-0.446*** (0.097)	-0.446*** (0.097)
Intercept	-7.491*** (0.66)	-7.491*** (0.66)
delta		
female	-0.261* (0.11)	
Insigma		
female		0.302* (0.15)
N	2797	2797

Standard errors in parentheses * p<0.05, ** p<0.01, *** p<0.001

Strengths and weaknesses of Allison's Approach. Allison's procedure is logical and well thought out. One limitation is that, while Allison focuses on omitted variable bias, omitted variable bias is only one possible source of heteroskedasticity. Unfortunately, unlike OLS, uncorrected heteroskedasticity in a model with dichotomous or ordinal dependent variables results in biased parameter estimates (Yatchew & Griliches 1985; Greene 2003; Keele & Park, 2006), for reasons similar to those as Allison has already described for the case of omitted variables. Therefore, researchers who are concerned about biased parameter estimates should not confine themselves to the special case of group differences in residual variances; they should worry about any source of heteroskedasticity and its possible biasing effects.

How well, though, does Allison's approach work for the group comparisons it was designed for? Three recent simulation studies have directly or indirectly addressed this issue. These simulations suggest that, under some circumstances, Allison's procedure works fairly well – but under other circumstances it is highly, and sometimes unnecessarily, problematic.

Hoetker's simulations. Hoetker (2004) did a series of simulations where he examined the problems raised by Allison and how well Allison's method addressed them. He found (p. 17)

that “in the presence of even fairly small differences in residual variation, naive comparisons of coefficients can indicate differences where none exist, hide differences that do exist, and even show differences in the opposite direction of what actually exists.” At least in the simulations he ran, he found that Allison’s method accurately detected differences in residual variation and false differences in coefficients, and that it also accurately detected true differences in coefficients.

At first glance, Hoetker’s simulations would seem to provide powerful support for Allison’s method. A closer examination reveals, however, that almost all of his simulations assumed that

- there really were differences in residual variation across groups
- the effects of heteroskedasticity were captured by a single grouping variable.

In other words, they showed that Allison’s method worked well when the model was correctly specified. Allison’s assumptions, however, that there are differences in residual variation, and that only one grouping variable is needed to capture these differences, may be highly problematic in practice.

Keele & Park’s simulations. Keele and Park (2006) do not specifically discuss Allison’s paper, but they do look at the closely related case of the heteroskedastic probit model. Their analysis was motivated by the observation (p. 4) that

While heterogenous choice models can be used for either “curing” probit models with unequal error variances or for testing hypotheses about heterogenous choices, there is little evidence, analytical or empirical, about how well these models perform at either task.

To assess the performance of heteroskedastic probit models, Keele and Park use both Monte Carlo simulations where true parameter values are known, and a re-analysis of the Alvarez and Brehm (1995) data on abortion attitudes. They find that

- Even under ideal conditions, i.e. when the model is correctly specified, estimates from the heteroskedastic probit model are problematic. Researchers are more likely to conclude that a parameter is statistically significant when it is not. Keele & Park conclude (p. 26) that “the standard errors from heteroskedastic probit models should not be relied upon. The standard errors from these models are overly optimistic and can lead to incorrect inferences.”

- Keele and Park also found that the heteroskedastic probit model had even worse problems when the model was mis-specified. When a relevant variable was excluded from the variance equation, parameter estimates were actually more biased than when the unequal variances were ignored altogether. They concluded (p. 27) that

If researchers are only interested in the parameters from the choice model, but suspect heteroskedasticity, these models may not be the best alternative. If the error variance differs across well defined groups, specification of the variance model should be relatively easy. But *if the source of the heteroskedasticity is less clear and harder to specify, it is better to estimate a standard probit and ignore the heteroskedasticity than poorly specify a heteroskedastic model.* [emphasis added]

To review, Hoetker did simulations where the model was correctly specified and argued that under those conditions Allison’s procedure worked fairly well. However, Keele and Park

showed that, even with correct model specification, the standard errors of heteroskedastic probit models can be biased, and they further showed that serious biases can occur when relevant variables are omitted from the variance equation. A procedure, such as Allison's, that only allows for a single dichotomous variable in the variance equation would presumably make omitted variable bias more likely in many situations.

Williams' Simulations. Williams (2009) presents simulations for a third case: residual variances are the same across groups but adjustments are made for heteroskedasticity anyway, i.e. there are extraneous variables in the variance equation. In these simulations,

- we created two groups (group 0 and group 1) with equal residual variances, i.e. $\delta = 0$.
- There was a dichotomous dependent variable Y, and two independent variables, X1 and X2.
- X1 and X2 were sampled from hypothetical populations where their variances were 1 and their correlations were 0.
- For group 0, the X1 and X2 Alpha coefficients always equaled 1. For group 1, the Alpha for X2 always equaled 2, i.e. was twice as large in group 1 as it was in group 0. The constants were always fixed at 1 for both groups.
- We varied the value for Alpha 1 in group 1, starting it at .5 and increasing it gradually to 3.0, i.e. Alpha 1 was sometimes smaller in group 1 than in group 0, and sometimes larger.
- Each simulation involved 1,000 cases, with 500 members in each group.
- For each simulation, we tested (a) the hypothesis that the residual variances were equal, i.e. $\delta = 0$, and (b) the hypothesis that one or more coefficients still differed across groups even after allowing for differences in residual variation.
- We also estimated a model in which an interaction term was added, allowing the effect of X2 to differ across groups.

The results of these simulations are presented in Table 4. When viewing these results, keep in mind that the true conditions are (a) the residual variances do not differ across groups and $\delta = 0$ (b) the coefficients do differ, and (c) the interaction term is equal to 1. Ideally the results from the simulations would reflect this.

Table 4: Simulations where residual variances are equal across groups but the coefficients are not

Alphas: $\alpha_1^0 = \alpha_2^0 = 1$ $\alpha_2^1 = 2$ α_1^1 varies	Residual variances differ, Alphas assumed to be the same		% of time LR test correctly rejects hyp of equal coefficients across groups	Effect of X2 allowed to differ across groups	
	Average value of delta	% of times LR test falsely rejects hyp of equal residual variances		Average delta	Average value of interaction term
$\alpha_1^1 = 0.50$	0.591	82.4%	99.9%	-0.491	3.346
$\alpha_1^1 = 0.75$	0.608	87.5%	98.9%	-0.238	1.798
$\alpha_1^1 = 1.00$	0.649	92.3%	90.7%	0.016	1.063
$\alpha_1^1 = 1.25$	0.718	95.5%	67.7%	0.271	0.638
$\alpha_1^1 = 1.50$	0.802	98.4%	35.5%	0.522	0.359
$\alpha_1^1 = 1.75$	0.908	99.6%	11.6%	0.782	0.157
$\alpha_1^1 = 2.0$	1.023	100.0%	5.1%	1.029	0.012
$\alpha_1^1 = 2.25$	1.151	100.0%	9.7%	1.277	-0.102
$\alpha_1^1 = 2.50$	1.303	100.0%	21.5%	1.539	-0.195
$\alpha_1^1 = 2.75$	1.460	100.0%	40.3%	1.795	-0.271
$\alpha_1^1 = 3.00$	1.631	100.0%	59.8%	2.054	-0.333

Problems that the above simulations indicate:

- The hypothesis of equal residual variances is almost always rejected, even though it is true. Why does this occur?
 - Recall, as Allison pointed out, that the test is done *under the assumption that the Alpha coefficients are the same across groups*.
 - Because the assumption is not true in these simulations, and because the coefficients are constrained to be equal across groups, the only way to adjust for differences across groups is by allowing the residual variances to differ.
 - As the average value of delta indicates, as the simulated value of Alpha 1 in group 1 gets larger and larger, delta gets bigger and bigger. That is, the larger the true difference is between the coefficients in the true groups, the larger the estimate of delta is to compensate for these differences.
 - Therefore, the test for equality of residual variances is not very informative, and indeed it really doesn't test what it claims to. A significant test statistic could indicate that residual variances differ across groups, but it could just as easily indicate that coefficients differ across groups.
- We often fail to reject the hypothesis of equal coefficients, even though it is true. Why?
 - in the model that includes delta, true differences in coefficients are falsely attributed to differences in residual variation. Hence, when you contrast the model that contains delta with the model that allows all coefficients to differ across groups, the

differences in coefficients appear to be smaller than they really are, and the statistical significance of the difference is understated.

- Allison noted that his procedure would have problems when the coefficients for one group all differed by a scale factor from the other group – a situation simulated here when Alpha 1 is 2 for group 1 – but as these simulations show, the test for equal coefficients can have problems under a much broader range of conditions.
- Our estimate of the interaction term is usually biased upward or downward. This estimation is done under the assumption that at least one coefficient (in this case X1) is the same across groups. Since this is not true (except when Alpha1 = 1) some of the difference in the X2 coefficient is falsely attributed to differences in residual variance.

Summing up. Allison's procedure requires critical assumptions at two points, and when these assumptions are incorrect the results can be highly misleading.

- The test of equal residual variances requires the assumption that the coefficients are the same in both groups. When this assumption is wrong, differences in coefficients are erroneously attributed to differences in residual variance. The test of equal residual variances therefore isn't very meaningful, and it requires that we assume the very thing we eventually want to prove or disprove. The erroneous inclusion of the delta parameter further biases subsequent tests of whether coefficients differ across groups.
- Allison's procedure also requires that, if we want to test whether specific coefficients differ across groups, we must assume that at least one coefficient is the same in both populations. When this assumption is correct, you get good estimates of across-group differences in the other coefficients, but when the assumption is wrong the estimates of other coefficients are biased upward or downward. In particular, when the coefficients are all larger in one group than the other, there is a downward bias in the estimated differences across groups.

Consequences of these problems. Taken together, these findings imply that routine use of Allison's procedure can lead to serious mistakes. For example:

- In the above simulations, you aren't just falsely rejecting the hypothesis that the Alphas are equal; you are also being led to believe in an alternative hypothesis, which in this case is false, i.e. the difference between the residual variances appears to be highly significant when in reality it is not.
 - Suppose omitted variables were themselves of substantive interest to the researcher? For example, a researcher might believe that omitted variables, such as discrimination, have much more impact on women than they do on men. Or, the researcher might believe that chance & random factors play a larger role in women's lives than they do men's. Results like the above would seemingly support her position.
- It is also important to remember that, even when the null hypothesis of equal effects is correctly rejected, there is still often going to be a *downward bias* in the estimated differences between coefficients, again because part of the real differences that exist are incorrectly attributed to differences in residual variation.

- Researchers generally do not just look at significance tests; they also make substantive evaluations of what the coefficients mean.
- In the above example, the real difference of 1 in Alpha 2 across groups might be considered very important; but an estimated difference of .1 (after incorrectly adjusting for differences in residual variation) might be considered a fairly minor matter.

The above have implications for Allison's non-simulated analysis.

- In it, the coefficients for number of articles differed by .0397 in the separate logistic regressions for men and women reported in Table 1; in Table 2, after applying his procedure, the interaction was only .03064, a 23 percent decline.
- If Allison's assumption that residual variance is different for men and women is wrong, then his approach has underestimated how much more men benefit from articles than do women.
- Of course the mistake would be even worse if a researcher decides to go with the significance tests (which are very borderline) and say there are no differences whatsoever.
- This is not to say that Allison's model is wrong, but researchers should realize that if it is wrong the mistake has non-trivial consequences. A source of gender inequality that Allison himself says (p.186) "many would find troubling" would suddenly seem to become non-existent because of a procedure based on incorrect assumptions.

Improving on Allison: Heterogeneous Choice Models

As noted before, Allison's heteroskedastic logit model, with a single dichotomous variable in the variance equation, is a special case of the larger class of models that are variously known as location-scale models and heterogeneous choice models. Turning to this larger class of models offers several ways to improve on Allison's approach and hopefully overcome its most significant weaknesses.

- There is no need to limit the variance equation to a single dichotomous grouping variable. Multiple grouping variables can be used, as can continuous variables that may be sources of heteroskedasticity. Indeed, the variables in the variance equation need not even be a subset of the variables in the choice equation, as is the case with Allison's procedure. This hopefully reduces or even eliminates problems caused by specification error in the variance equation.

- The variance may itself be of substantive interest. The variance equation lets you examine the determinants of variability. Alvarez and Brehm (1995), for example, argued that individuals whose core values are in conflict will have a harder time making a decision about abortion and will hence have greater variability/error variances in their responses. In the case of Allison's example, we might be interested in whether gender or other factors affect the variability in careers.

- Allison's procedure works with a dichotomous dependent variable. Heterogeneous choice models also allow for ordinal dependent variables. There are several advantages to using ordinal variables when possible.

- As Keele and Park (2006) note, ordinal variables contain more information and models using them are much less prone to problems than are models with dichotomous dependent variables. Based on their Monte Carlo simulations, they concluded that, unlike the heteroskedastic probit model, when the model was correctly specified, “The heteroskedastic ordered probit model can be given a clean bill of health, as both the level of overconfidence and coverage rates are close to ideal.”
 - However, even for a heteroskedastic ordered probit model, they stressed the importance of the model being correctly specified; a mis-specified model, e.g. a variance equation with omitted variables might be worse than a model that made no correction at all for heteroskedasticity.
- Also, unlike with Allison’s procedure, with ordinal variables you do NOT need to make the questionable assumption that at least one coefficient is the same across groups. You do, however, need to make the assumption that the cutpoints are the same for both groups. This is a less questionable assumption, in that it implies that both groups interpret the question the same way.
 - Nonetheless, researchers should realize that the assumption may be wrong in some cases; for example, Lindeboom & Doorslaer (2004) note (p. 1084) that sometimes “sub-groups of a population use systematically different threshold levels when assessing their health, despite having the same level of ‘true’ health. These differences may be influenced by, among other things, age, sex, education, language and personal experience of illness. It means that different groups appear to ‘speak different languages’ and to use different reference points when they are responding to the same question.” Of course, any procedure can have problems if different groups interpret and answer questions differently.

VII Conclusions

Allison has provided a valuable service by alerting researchers to an important problem that has gone unnoticed by many. However, thanks, in part, to additional research that Allison’s paper helped inspire, we know that his original proposed solution can sometimes have serious problems, and, counter to his advice, should NOT be applied on a routine basis.

- Under certain conditions, Allison’s procedure can produce biased and inefficient estimates, and may be worse than doing nothing at all. Today, superior alternatives can be easily estimated using readily available commercial software.
- Luckily, Allison’s procedures can be easily modified take advantage of the broader class of heterogeneous choice models, providing a powerful, and often more appropriate, way for addressing the problems that Allison presents.

At the same time, researchers need to realize that even with these methods, mis-specified models can be problematic.

- As Keele and Park (2006) show, ordinal models can also produce misleading results when the variance equation is mis-specified. The greater flexibility of heterogeneous choice models (which, unlike Allison's procedure, allow multiple variables in the variance equation) make omitted variable bias less likely, but it is still up to researchers to think through their models carefully.
- The inclusion of extraneous variables in the variance equation could still potentially distort estimates of group differences. Again, this seems less likely with a well thought-out model involving multiple variables, but it could still happen.
- Researchers should therefore estimate models both with and without controls for heteroskedasticity, and consider whether model mis-specification could be the cause of any seemingly-major differences in conclusions.

As part of this process, researchers may wish to vary the sequence in which they estimate nested models.

- For example, Allison first added the delta parameter to his model, noted that it was significant, and then added the interaction term for gender and articles, which he concluded was insignificant. (The delta term also became insignificant once the interaction term was entered, but Allison did not note this.)
- Had he first added the interaction term for articles, he would have found that it was significant, and that the delta term was insignificant when it was next added to the model.
- In other words, the sequence of models should not automatically give preference to the hypothesis of different residual variances over the hypothesis of differing coefficients. If the sequence of models does affect the conclusions reached, researchers should at least acknowledge this in their discussion if not rethink their models altogether.

In short, comparisons of logit and probit coefficients across groups do pose challenges to researchers. However, well thought out models, modern statistical software, and the methods described here can help to make those challenges manageable.