Paper SAS207-2014

Did My Coupon Campaign Accomplish Anything? An Application of Selection Models to Retailing

Gunce E. Walton and Kenneth Sanford, SAS Institute Inc.

ABSTRACT

Evaluation of the efficacy of an intervention is often complicated because the intervention is not randomly assigned. Usually, interventions in marketing, such as coupons or retention campaigns, are directed at customers because their spending is below some threshold or because the customers themselves make a purchase decision. The presence of nonrandom assignment of the stimulus can lead to over- or underestimating the value of the intervention. This can cause future campaigns to be directed at the wrong customers or cause the impacts of these effects to be over- or understated. This paper gives a brief overview of selection bias, demonstrates how selection in the data can be modeled, and shows how to apply some of the important consistent methods of estimating selection models, including Heckman's two-step procedure, in an empirical example. Sample code is provided in an appendix.

INTRODUCTION

Sample selection bias is a concern in all studies that are based on observational data—that is, data that arise outside the confines of a controlled experiment. These data are often the only information available for making causal inference in cases where a controlled experiment is not feasible. Sample selection bias presents a problem when the results of a study need to be used on a wider population than was sampled. In essence, if you collect a nonrepresentative sample, then sample selection bias could be a concern for generalizing the results. This paper lays out the problem of sample selection bias with respect to a common business application, offers several potential corrections to this problem, and highlights the impact of ignoring sample selection bias.

Coupons are a widely used method of promotion by retailers. These promotions might dramatically alter revenue, depending on the incentive effect of the coupon. To predict the effect of a promotion on revenue, you need to estimate the effects of past promotions on spending. That is, how does the use of the coupon alter customer spending? Marketing research departments at retailers are often asked to evaluate the overall impact of these promotions prior to a new campaign by using data about past campaigns. And these data contain only information about customers who previously purchased the promoted item. So where is the sample selection bias? Customers who used the coupons in the past are a nonrandom sample from the population, because the purchase decision involves two other decisions: (1) an extensive decision (to buy or not buy) and (2) an intensive decision (how much to buy). Although you are interested primarily in the intensive decision, ignoring the extensive choice might bias the results. This bias does not disappear in large samples, and this inconsistency will lead to misinformed policy decisions. The next section introduces the concepts of sample selection and incidental truncation and illustrates them by using the coupon example. Then, it introduces the model and a method of consistent estimation in the presence of sample selection, followed by presenting the data that are used in this analysis. After that, the paper compares the estimates from several different models and discusses the statistical and economic differences of the candidate estimators by using the QLIM procedure in SAS/ETS®. (See Appendix B for the example code.)

AN OVERVIEW OF SAMPLE SELECTION

Usually, when a researcher estimates the unknown parameters of an economic model by using standard econometric methods, she assumes that the data set that she is using is a random subset of the population. However, in many cases the sample might not be randomly drawn from the underlying population but instead drawn based on a selection mechanism or rule. There are a variety of selection mechanisms, and a sample that is drawn in this way is called a *selected sample*.

One example of a selection mechanism is selection on the basis of the response variable being truncated. It usually occurs when a survey of a program is designed to intentionally exclude part of the population. A classic example of this selection mechanism is the study by Hausman and Wise (1977) of the determinants of earnings. They recognized that their sample from a negative income tax experiment was truncated, because only families whose income was below 1.5 times the poverty level were allowed to participate in the program; no data were available on families whose incomes were above this threshold value. In this case, units *are* randomly drawn from the population, but data about one variable are missing for some units in the sample.

Another example of a selection mechanism is the selection that is determined by a probit model; this is known as *incidental truncation*. The incidental truncation problem is described by Gronau (1974) in his model of the wage offer and labor force participation. The model of interest is the hourly wage offer equation for people of working age. By definition, this equation is supposed to represent all people of working age, independent of whether a person is actually working or not at the time of the survey. However, the wage offer can be observed only for working people, so data about a key variable are available only for a clearly defined subset of the population.

In the incidental truncation example, people self-select into employment, so whether or not you observe the wage depends on an individual's labor supply decision. However, no matter what the selection mechanism is, whether sample selection or self-selection, you must account for the nonrandom nature of the sample that you have for estimating the equation of interest.

INCIDENTAL TRUNCATION IN COUPON AND SALES DATA

The selection problem that is due to incidental truncation is also likely to occur in coupon and sales data about a company's brand. These data include the quantities sold, the prices and coupon values that customers face, and the demographics of the customers. The interest lies in estimating $E(q_i^*|\boldsymbol{X}_i)$, where q_i^* is the quantity demanded by customer i and $\boldsymbol{X}_i$ includes the variables that explain the demand. $E(q_i^*|\boldsymbol{X}_i)$ can be viewed as an elasticity equation or a coupon sensitivity model, and q_i^* comes from the utility optimization process. You can model this economic behavior as

$$\max_{q^*} \text{ Utility}_i(q^*, \boldsymbol{X}_i) \quad \text{subject to} \quad q^* > 0$$

where $\boldsymbol{X}_i$ is an array of variables that determine the utility, such as price, coupon value, income, family size, and so on. The restriction $q_i^* > 0$ guarantees that customer i either buys a positive quantity or does not buy the product at all; that is, a purchase of negative quantity does not occur.

This process means that customer i demands the quantity q^* that maximizes her utility and she purchases this quantity if her utility is positive; otherwise, she does not purchase the product. This behavior introduces a potential sample selection problem, because the data can be observed only for customers who purchase the company's product and not for customers who have negative utility if they buy a positive quantity.

An important point to note is that the preceding selection mechanism is imposed by the customers themselves by maximizing their own utility. It is also possible to observe the same selection problem if the selection rule is imposed by the company itself. For example, if the company distributes the coupons to customers based on their demographic background that causes their spending to be below some threshold, and it collects data about only the customers who receive coupons, then these data might suffer from a potential selection problem.

MODEL AND ESTIMATION

You can model customer i's utility from a certain product as

$$U_i = \boldsymbol{X}_i\boldsymbol{\alpha} + v_i$$

where $\boldsymbol{X}_i$ is as defined earlier and $\boldsymbol{\alpha}$ is the vector of parameters that correspond to $\boldsymbol{X}_i$. Let d_i be the binary purchasing decision indicator for customer i. You can write the process that generates the data as

$$\begin{aligned} q_i^* &= \boldsymbol{X}\mathbf{1}_i\boldsymbol{\beta} + \varepsilon_i \\ d_i &= 1[U_i > 0] \end{aligned}$$

where $\boldsymbol{X1}_i$ is the vector of explanatory variables in the demand model and is usually a subset of $\boldsymbol{X}_i$, and $\boldsymbol{\beta}$ is the corresponding vector of parameters.

Note that d_i and all the explanatory variables are observed, but q_i^* is observed only when $d_i = 1$ (that is, when customer i actually purchases the product). Assume exogeneity of the explanatory variables. Let v_i and ε_i be linearly correlated with the covariance γ.

$E(q_i^* | \boldsymbol{X}_i,\ d_i = 1)$ is the equation that the company would like to estimate in order to understand its customers' sensitivity to coupon values; this is called the *equation of interest*. The equation for the purchasing decision, d_i, is called the *selection equation*.

Because q_i^* is observed only when $d_i = 1$, you estimate $E(q_i^* | \boldsymbol{X}_i,\ d_i = 1)$ along with $P(d_i = 1 | \boldsymbol{X}_i)$. Heckman (1979) and Amemiya (1985), among others, show that

$$E(q_i^* | \boldsymbol{X}_i,\ d_i = 1) = \boldsymbol{X1}_i\boldsymbol{\beta} + \gamma\lambda(\boldsymbol{X}_i\boldsymbol{\alpha})$$

where $\lambda(\cdot) \equiv \phi(\cdot)/\Phi(\cdot)$ and $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative distribution function, respectively, of standard normal distribution. The term $\lambda(\cdot)$ is the inverse Mills ratio. The preceding equation makes it clear that an ordinary least squares (OLS) regression of q_i^* on $\boldsymbol{X1}_i$ that uses the selected sample omits the term $\lambda(\boldsymbol{X}_i\boldsymbol{\alpha})$ and usually leads to inconsistent estimation of $\boldsymbol{\beta}$.

Following Heckman (1979), you can consistently estimate $\boldsymbol{\beta}$ and γ, using the selected sample, by regressing q_i^* on $\boldsymbol{X1}_i$ and $\lambda(\boldsymbol{X}_i\boldsymbol{\alpha})$. However, $\boldsymbol{\alpha}$ is unknown, so you cannot compute the additional regressor, $\lambda(\boldsymbol{X}_i\boldsymbol{\alpha})$. On the other hand, you can obtain a consistent estimator of $\boldsymbol{\alpha}$ from the probit estimation of the selection equation. This estimation procedure can be summarized in two steps:

1. Obtain the probit estimate $\hat{\boldsymbol{\alpha}}$ from the model $P(d_i = 1 | \boldsymbol{X}_i) = \Phi(\boldsymbol{X}_i\boldsymbol{\alpha})$ by using all observations. Then obtain the estimated inverse Mills ratio $\lambda(\boldsymbol{X}_i\hat{\boldsymbol{\alpha}})$ for at least the selected sample.
2. Obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}$ from the OLS regression on the selected sample q_i^* on $\boldsymbol{X1}_i$ and $\lambda(\boldsymbol{X}_i\hat{\boldsymbol{\alpha}})$ for the selected sample.

This is called *Heckman's two-step procedure*.

The standard least squares estimators of the population variance of ε and the variances of the estimated coefficients are incorrect. The correct variance estimators are given in Appendix A. The HECKIT option in PROC QLIM reports the corrected standard errors automatically. If necessary, you can obtain the uncorrected standard errors by using the HECKIT(UNCORRECTED) option. Note that a simple t test on γ is a valid test of the null hypothesis of no selection bias.

You can implement Heckman's two-step procedure in PROC QLIM when you specify the HECKIT option together with the SELECTION option in the MODEL statement.

If you are willing to make the assumption that ε_i and v_i are bivariate normal with mean 0, $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2$, $\mathrm{Cov}(\varepsilon_i, v_i) = \sigma_{\varepsilon,v_i}$, and $\mathrm{Var}(v_i) = 1$, then maximum likelihood estimation (MLE) will be more efficient than Heckman's two-step procedure under joint normality of ε_i and v_i. However, MLE is not as robust as Heckman's two-step procedure, and sometimes convergence issues occur.

The likelihood function for customer i of the selection model, which consists of the equation of interest and the selection equation, can be written as follows. (For more information, see Wooldridge 2002.)

$$\begin{aligned}\ell_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) \quad = \quad & (1 - d_i)\log[1 - \Phi(\boldsymbol{X}_i\boldsymbol{\alpha})] + d_i(\log\Phi\{[\boldsymbol{X}_i\boldsymbol{\alpha} + \sigma_{\varepsilon,v_i}\sigma_\varepsilon^{-2}(q_i^* - \boldsymbol{X1}_i\boldsymbol{\beta})] \\ & \times\ (1 - \sigma_{\varepsilon,v_i}^2\sigma_\varepsilon^{-2})^{-1/2}\} + \log\phi[(q_i^* - \boldsymbol{X1}_i\boldsymbol{\beta})/\sigma_\varepsilon] - \log(\sigma_\varepsilon))\end{aligned}$$

The log likelihood is obtained by summing $\ell_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ across all observations; d_i picks out when q_i^* is observed and therefore contains information for estimating $\boldsymbol{\beta}$.

The next section implements these two consistent estimation methods: Heckman's two-step procedure and MLE.

DATA AND RESULTS

This section applies Heckman's two-step procedure and MLE, discussed in the previous section, to partially simulated data about ketchup. By simulating some of the data, you gain full control of the true parameters that specify the selection model and hence get a better understanding of the merits of these estimators. The aim here is to demonstrate the effectiveness of these estimation methods in assessing and eliminating the self-selection bias. The model of interest is estimated using the OLS estimator also to provide an understanding of the magnitude of its bias. Then, the results of all three methods are compared.

In this section, $X1$ and X are specified. Both $X1$ and X contain the variables **NWifeInc**, **HSize**, **Age**, **InLF**, and **Educ**. Observations on these variables are generated using the actual values in the data set that comes from the University of Michigan Panel Study of Income Dynamics for 1975. The original data set is the one used in Mroz (1987).[1] Definitions of these variables are as follows:

NWifeInc: nonwife income, which is obtained by subtracting the wife's income from the total household income

HSize: household size, which is the total number of people under age 18 living in the household[2]

Age: age of the female head of household

InLF: employment indicator for the female head of household

Educ: number of years of education of the female head of household

The generated data set consists of 1,753 observations. The focus is mainly on the characteristics of the household that relate to the female head of household, because the decision of how much of the product, such as ketchup, to buy is most likely to be affected by the household characteristics related to the female head of household.

The explanatory variables **Price**, **Coup**, and **NCouPro** are also included. These are simulated and are defined as follows:

Price: price per 14-oz. bottle of ketchup

Coup: coupon value for ketchup

NCouPro: indicator of a noncoupon promotion (such as a price markdown) at time of purchase

Price changes from customer to customer, just as **Coup** does, because the price of a particular brand can vary from store to store or as a result of noncoupon promotions available in a store. **Price** is distributed as normal with mean 3 USD and standard error 1, and **Coup** is distributed as normal with mean 1 USD and standard error 2. The mean and the standard error of **Price** reflect the average price and the standard error of a 14-oz. bottle of a leading brand of ketchup sold on shopping websites such as Amazon.com. Because **Price** is normalized to the price per 14-oz. bottle, you should interpret this parameter by taking this normalization into account. The standard error of **Coup** is larger than that of **Price**, reflecting the larger variation in coupon values than the variation in prices that customers face. To make sure that negative values of **Price** or **Coup** are not observed, they are set to 0 whenever they take a negative value.

The dependent variables q^* and d are generated based on the following data generating process:

$$\begin{aligned} U_i &= 5 - 0.5Price_i + 0.6Coup_i + 0.3NWifeInc_i + 0.3HSize_i - 0.2Age_i \\ &\quad + 0.3InLF_i - 0.3Educ_i + 0.5NCouPro + v_i \end{aligned}$$

$$d_i = \begin{cases} 1 & \text{if } U_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

[1] A more complete discussion of the original data is found in Mroz (1987), Appendix 1.

[2] This variable is not in the Mroz (1987) data set. It is obtained from variables $kidslt6$ and $kidsge6$ as $kidslt6 + kidsge6 + 2$.

$$
\begin{aligned}
q_i^* &= 10 - 0.5Price_i + 0.7Coup_i + 0.5NWifeInc_i + 0.5HSize_i - 0.1Age_i \\
&\quad + 0.3InLF_i - 0.2Educ_i + \varepsilon_i \quad \text{if } d_i = 1
\end{aligned}
$$

If $d_i = 0$, q_i^* is missing. During the data generation process, if a negative value of q_i^* given $d_i = 1$ is observed, then that observation is also considered to be missing. The SAS program that generates the data appears in Appendix B.

First, you estimate the selection model by using Heckman's two-step procedure. The SAS statements for this estimation method are as follows:

```
title "Consistent Estimation Using Heckman's Two-Step Procedure";
proc qlim data=simseldat1 heckit;
     model dq = p cv nwifeinc  hsize age
                inlf educ ncprom / discrete;/* the selection equation--probit */
     model  q = p cv nwifeinc  hsize age
                inlf  educ / select(dq=1); /* the equation of interest --linear */
run;
```

The output of the first- and second-step estimations is summarized in Table 1 and Table 2, respectively.

Table 1 Output of Heckman's Two-Step Procedure: Probit

Parameter	True Value	Estimate	Unscaled Estimate	Standard Error	t Value
Intercept	5	2.382086	4.125894	0.413382	5.76
Price	–0.5	–0.240868	–0.417196	0.042977	–5.6
Coup	0.6	0.346217	0.599665	0.030581	11.32
NWifeInc	0.3	0.162452	0.281375	0.007705	21.08
HSize	0.3	0.161293	0.279368	0.029796	5.41
Age	–0.2	–0.10714	–0.185572	0.006605	–16.22
InLF	0.3	0.276589	0.479066	0.085866	3.22
Educ	–0.3	–0.157654	–0.273065	0.020603	–7.65
NCouPro	0.5	0.348068	0.602872	0.083986	4.14

Table 2 Output of Heckman's Two-Step Procedure: Linear Regression

Parameter	True Value	Estimate	Standard Error	t Value
Intercept	10	10.091337	0.467213	21.6
Price	–0.5	–0.476086	0.051288	–9.28
Coup	0.7	0.686579	0.033805	20.31
NWifeInc	0.5	0.502025	0.005416	92.7
HSize	0.5	0.505562	0.035721	14.15
Age	–0.1	–0.110637	0.00812	–13.63
InLF	0.3	0.309729	0.102318	3.03
Educ	–0.2	–0.190882	0.023086	–8.27
Mills	1.2728	1.454816	0.13097	11.11

The parameter **Mills** in Table 2 represents γ, the coefficient of $\lambda(\boldsymbol{X}_i\boldsymbol{\alpha})$, which is defined in the previous section. Note that Table 1 has a column labeled "Unscaled Estimate." In the probit estimation, the variance of the error, σ_ε^2, is normalized to 1 to ensure identification. If σ_ε^2 is in fact a positive value other than 1, then the probit parameter estimates are scaled by σ_ε. More specifically, if $\hat{\boldsymbol{\alpha}}$ is the estimator from a probit of q^* on $\boldsymbol{X}$, then $\operatorname{plim} \hat{\alpha_j} = \frac{\alpha_j}{\sigma_\varepsilon}$. Therefore, the unscaled estimates can be obtained by multiplying the probit estimates by σ_ε, the standard error of ε. In the data, $\sigma_\varepsilon = \sqrt{3}$, so the values in the "Unscaled Estimate" column are obtained by multiplying the corresponding values in the "Estimate" column by $\sqrt{3}$.

The estimate of the coefficient of **Coup** in the demand equation is of particular interest, because this is the parameter that reflects the coupon sensitivity of customers, and hence the company's decision about the

coupon promotion depends mostly on this parameter. Therefore, estimating this parameter correctly is a very important factor in making an optimal decision. In Table 2, you see that the estimates are very close to their actual values, including the coupon sensitivity parameter estimate. To understand statistically how close the coupon sensitivity parameter estimate is to the true value, a Wald test of $H_0 : \beta_C = 0.7$ is applied, where β_C is the coupon sensitivity parameter.[3] The following SAS statement, which is inserted after the MODEL statements, implements this test. Table 3 summarizes the output of the Wald test. You see that there is a failure to reject H_0 at the 5% significance level, so you can conclude that the estimate of β_C that is obtained using Heckman's two-step procedure is not statistically different from its true value.

```
test q.cv=0.7; /*Default is Wald */
```

Table 3 Test of $H_0 : \beta_C = 0.7$ Using the HECKIT Option

Test Type	Statistic	p-Value
Wald	0.20	0.6509

Overall, the parameter estimates of both equations of the selection model are very close to their corresponding true values.

Note that the t test rejects the null hypothesis that the coefficient of the inverse Mills ratio equals 0; in effect, it rejects the fact that there is no selection in the data. This is consistent with how the data were generated.

Second, the selection model is estimated by using the MLE method. The SAS statements for this estimation are as follows:

```
title "Consistent Estimation Using MLE";
proc qlim data=simseldat1;
     model dq = p cv nwifeinc  hsize age
                inlf educ ncprom / discrete;/* the selection equation--probit */
     model  q = p cv nwifeinc  hsize age
                inlf  educ / select(dq=1); /* the equation of interest --linear */
run;
```

Recall that the MLE method estimates the parameters of the selection equation and the equation of interest jointly. The output for this estimation is summarized in Table 4.

[3] By using the HECKIT(SECONDSTAGE=MLE) option, you can run a test on the parameters of the linear model.

Table 4 Results of MLE of Selection Model

	Parameter	True Value	Estimate	Standard Error	t Value
Demand Equation	Intercept	10	10.043147	0.453387	22.15
	Price	–0.5	–0.470239	0.049164	–9.56
	Coup	0.7	0.680476	0.03104	21.92
	NWifeInc	0.5	0.499585	0.004078	122.52
	HSize	0.5	0.498317	0.034453	14.46
	Age	–0.1	–0.108762	0.00703	–15.47
	InLF	0.3	0.337934	0.099331	3.4
	Educ	–0.2	–0.183938	0.022129	–8.31
	σ_ε	1.414214	1.475521	0.042463	34.75
Selection Equation	Intercept	5	2.475578	0.364598	6.79
	Price	–0.5	–0.243675	0.038625	–6.31
	Coup	0.6	0.34452	0.026882	12.82
	NWifeInc	0.3	0.16562	0.006571	25.2
	HSize	0.3	0.171717	0.026756	6.42
	Age	–0.2	–0.107622	0.005932	–18.14
	InLF	0.3	0.220305	0.07703	2.86
	Educ	–0.3	–0.168342	0.017822	–9.45
	NCouPro	0.5	0.343363	0.061864	5.55
	ρ	0.9	0.932442	0.015655	59.56

The estimates that are obtained by the MLE method are very close to their corresponding true values. Note that the unscaled parameter for the probit selection equation is reported for the reason stated earlier. The Wald test of $H_0 : \beta_C = 0.7$ is run, and the result is shown in Table 5. There is a failure to reject H_0, so you can conclude that the ML estimate of the price sensitivity parameter is not statistically different from its true value. Overall, the estimates that are obtained by the MLE method are very close to those obtained by Heckman's two-step procedure.

Table 5 Test of $H_0 : \beta_C = 0.7$ Using MLE

Test Type	Statistic	p-Value
Wald	0.40	0.5294

As discussed earlier, running a simple OLS regression on the demand equation ignores the selection problem that exists in the data set. Consequently, the OLS estimator suffers from selection bias. To show the magnitude of this bias, the selection model is estimated by OLS and the results are reported in Table 6.[4]

Table 6 OLS Estimation of the Demand Equation

Parameter	True Value	Estimate	Standard Error	t Value
Intercept	10	9.875401	0.449423	21.97
Price	–0.5	–0.383283	0.048758	–7.86
Coup	0.7	0.567184	0.029979	18.92
NWifeInc	0.5	0.467363	0.003593	130.08
HSize	0.5	0.467268	0.03443	13.57
Age	–0.1	–0.072899	0.006912	–10.55
InLF	0.3	0.229242	0.099258	2.31
Educ	–0.2	–0.155631	0.021949	–7.09
σ_ε	1.414213562	1.313653	0.034101	38.52

[4] Despite being called OLS estimates, they are actually obtained using the MLE method. However, the ML estimator is theoretically the OLS estimator under the normality assumption.

You can clearly see that the OLS estimator does not perform as well as the previous two estimators. In fact, for most of the parameters, the null hypothesis that the parameter estimate is not statistically different from its true value is rejected when you apply a Wald test at the 5% significance level. The parameters for which this null hypothesis is rejected are **Price**, **Coup**, **NWifeInc**, **Age**, and **Educ**. The test result for **Coup** is given in Table 7.

Table 7 Test of $H_0 : \beta_C = 0.7$ Using OLS

Test Type	Statistic	p-Value
Wald	19.63	< 0.0001

ECONOMIC SIGNIFICANCE OF RESULTS

The previous sections highlight the statistical differences that arise from various estimation methods. This section discusses the economic significance of employing the consistent estimators versus the inconsistent (and biased) OLS estimates. It begins with a discussion of Table 8, which compares the true **Price** and **Coup** parameter values in the demand equation with the estimates of these parameters that are obtained by the OLS, Heckman's two-step, and MLE methods (the standard errors are given in parentheses). The table displays only the coefficients and standard errors of the **Price** and **Coup** variables, because these are policy variables that are under the company's control.

The true values of the coefficients of **Price** and **Coup** are –0.5 and 0.7, respectively. As shown in previous sections, both Heckman's two-step procedure and the MLE method produce coefficients that are not statistically different from the true parameter values. The OLS estimates, which ignore the selection problem, are statistically different from the true parameter values at the 5% significance level. Because the estimates that are obtained by Heckman's two-step procedure and the MLE method are very similar, the Heckman's two-step procedure estimates are used for the remainder of this section to demonstrate the impact of using biased coefficients.

Table 8 Comparison of the Coupon Sensitivity Estimates

Parameter	True Value	OLS	Heckman's Two-Step	MLE
Price	–0.5	–0.383283	–0.470239	–0.476086
		(0.048758)	(0.049164)	(0.051288)
Coup	0.7	0.567184	0.686579	0.680476
		(0.029979)	(0.033805)	(0.03104)

From a qualitative perspective, the estimated partial effect of **Price** on quantity is biased toward 0 in OLS (–0.38 by OLS versus –0.47 by Heckman's two-step procedure). These OLS estimates suggest that customers are less price-sensitive than they actually are. Policies that are built on these biased estimates might lead the company to overestimate the impact that a change in price would have on total revenue. For this data set, a 1% increase in price leads to a 0.04% decrease in quantity under the OLS estimates but to a 0.05% decrease in quantity under the Heckman's two-step estimates. The impact of these differences is even greater in the presence of larger changes in price.

The impact of coupon use on revenue is even more interesting. The effect of **Coup** on quantity is 0.57 for OLS versus 0.69 for the Heckman's two-step estimates, implying that OLS underestimates the effect of the coupon on spending. If used in practice, these OLS estimates would undervalue the effect of the coupon. For example, the OLS estimates suggest that it would require $125 worth of additional coupons to increase sales by 1,000 units, whereas the Heckman's two-step estimates suggest that only $103 worth of coupons would be required. The interpretation of this difference is that Heckman's two-step procedure (and MLE) accurately gauges the responsiveness of the campaign. This campaign results in the need to offer fewer coupons to get the desired response.

SUMMARY

This paper provides an explanation of the sample selection bias problem as it relates to estimating the effects of both a coupon and a price change on spending. It shows that ignoring the problems of sample selection can lead to asymptotically biased results and that these results have serious implications for revenue predictions based on these models. The authors recommend that modelers exercise caution in using simple OLS methods when estimating models similar to those in this paper, because OLS methods might perform poorly in situations of sample selection.

APPENDIX A

Let N_1 be the selected subsample size and K be the number of variables in $\mathbf{X1}$. Define $\lambda_i \equiv \lambda(\boldsymbol{X}_i\boldsymbol{\alpha})$. An estimator of σ_ε^2 is given by Heckman (1979) as

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N_1}\sum_{i=1}^{N_1} e_i^2 + \hat{\gamma}^2 \frac{1}{N_1}\sum_{i=1}^{N_1} \hat{\delta}_i$$

where e_i is the residual for the ith observation obtained in step 2 and $\hat{\delta}_i = \hat{\lambda}_i^2 + \hat{\lambda}_i \mathbf{X}_i\hat{\boldsymbol{\beta}}$. Let $\mathbf{X}_*$ be an $N_1 \times (K+1)$ matrix with ith row $[\mathbf{X1}_i \;\; \lambda_i]$, and define $\mathbf{W}$ similarly with ith row $\mathbf{X}_i$. Then the estimator of the asymptotic covariance of $[\hat{\boldsymbol{\beta}}, \hat{\gamma}]$ is

$$\text{EstAsyVar}[\hat{\boldsymbol{\beta}}, \hat{\gamma}] = \hat{\sigma}_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}[\mathbf{X}_*'(\mathbf{I} - \hat{\rho}^2\hat{\boldsymbol{\Delta}})\mathbf{X}_* + \mathbf{Q}][\mathbf{X}_*'\mathbf{X}_*]^{-1}$$

where $\hat{\rho}^2 = \hat{\gamma}^2/\hat{\sigma}_\varepsilon^2$, $\hat{\boldsymbol{\Delta}} = \text{diag}(\hat{\delta}_i)$, and

$$\mathbf{Q} = \hat{\sigma}_\varepsilon^2(\mathbf{X}_*'\hat{\boldsymbol{\Delta}}\mathbf{W})\text{Est.Asy.Var}(\hat{\alpha})(\mathbf{W}'\hat{\boldsymbol{\Delta}}\mathbf{X}_*)$$

where Est.Asy.Var($\hat{\alpha}$) is the estimator of the asymptotic covariance of the probit coefficients that you obtain in step 1. When you specify the HECKIT option, the QLIM procedure uses a numerical estimated asymptotic variance.

When the HECKIT option is specified, the QLIM procedure reports the corrected standard errors for $[\hat{\boldsymbol{\beta}}, \hat{\gamma}]$ automatically. However, if you need the conventional OLS standard errors, you can specify the HECKIT(UNCORRECTED) option.

APPENDIX B

The SAS program that generates the data is as follows.

```
proc import out=work.mroz DATAFILE="C:\mroz.dta"
dbms=stata replace;
run;
data seldat;
set mroz;
keep nwifeinc hsize age inlf educ ;
hsize = kidslt6 + kidsge6 + 2;
run;
/* Generating more data using bootstrapping */
%macro generate_selection(source_dset = _last_, output_dset = generated_sample,
                          varlist = _ALL_, sample_size = 100);

data _pitemp;
   set &source_dset;
   keep &varlist;
run;

data _null_;
   dsid = open("_pitemp");
   nvar = attrn(dsid, 'NVAR');
```

```
   call symputx("number_of_variables",nvar);
   nobs = attrn(dsid, 'NOBS');
   call symputx("original_size", nobs);
run;

data gen_sample(drop = i);
  k = &original_size;
  %do it = 1 %to &number_of_variables;
  %let variable_name = %scan(&varlist,&it,%str( ));
  choice = int(ranuni(36830)*k) + 1;
  set &source_dset.(keep = &variable_name) point = choice nobs = k;
  %end;
  i + 1;
  if i > &sample_size then stop;
run;

data &output_dset;
set &source_dset gen_sample;
run;

%mend;
/* Generating 1000 more observations */
%generate_selection(source_dset = seldat, output_dset = sample,
                    varlist = nwifeinc hsize age inlf educ, sample_size = 1000);

data sim;
   keep ncprom price coup errs erri /*mu inc size age inlf educ*/;
   rho=.9; /*correlation between the errors */
   v1 = 3; /*variance of the selection model errors */
   v2 = 2; /*variance of the model of interest */
   N = 1753; /* number of observations */
   do i = 1 to N;
      /* creating correlated errors with the given variances */
      a = sqrt((1 + (sqrt(1 - rho**2)))/2);
      b = rho/(2*a);
      x1 = rannor(19283);
      x2 = rannor(19283);
      errs = sqrt(v1)*(a*x1 + b*x2);
      erri = sqrt(v2)*(b*x1 + a*x2);


      price = 3 + rannor( 19283 );  /*price of the good  */
  if ( price < 0 ) then price = .; /* price cannot be negative */
  coup = 1 + 2*rannor( 19283 ); /*coupon value  */
  if (coup < 0 ) then coup = 0; /* coupon value cannot be negative */
  /* creating the noncoupon promotion variable, which is a dummy */
  uni = ranuni(2589741); /* u ~ U[0,1] */
  if ( uni > 0.5) then ncprom = 1;
  else                 ncprom = 0;
      output;
   end;
   format price  coup 5.4;
run;

data simseldat;
merge sample sim;
run;

data simseldat1;
set simseldat;
```

```
    /* The selection rule */
    qstar = 5 + (-.5)*price + .6*coup + .3*nwifeinc + .3*hsize +
            (-.2)*age + .3*inlf + (-.3)*educ + .5*ncprom + errs;
    if (qstar > 0 ) then dq = 1;
    else            dq = 0;
    /* The model of interest */
    if (qstar > 0 ) then q = 10 + (-.5)*price + .7*coup  + .5*nwifeinc + .5*hsize
                             + (-.1)*age + .3*inlf + (-.2)*educ + erri;
    else                 q = .;
run;
```

REFERENCES

Amemiya, T. (1985), *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Gronau, R. (1974), "Wage Comparisons—a Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143.

Hausman, J. A. and Wise, D. A. (1977), "Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica*, 45, 319–339.

Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Mroz, T. A. (1987), "The Sensitivity of an Empirical Model of Married Women's Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–799.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.

ACKNOWLEDGMENTS

The author is grateful to Oleksiy Tokovenko of the Advanced Analytics Division at SAS Institute Inc. for his valuable help with the SAS DATA step.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the following author:

Gunce E. Walton
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-2366
gunce.walton@sas.com