Semiparametric Estimation of the Size of Oil Tanker Spills

Ayla Ogus

June 28, 1998

Abstract

This paper estimates the determinants of the size of oil tanker spills without distributional assumptions on the error terms. We employ semiparametric estimation techniques to estimate the parameters of a sample selection model and compare them to the estimates from a sample selection model with normal errors. We find that although parameter estimates are sensitive to the assumption of normality and to the semiparametric technique used. Major findings that are qualitatively supported by all methods are: groundings and collisions result in larger spills if there is a spill, but the likelihood that there will be a spill due to a grounding or collision is very low; tanker size has only a marginal effect on the probability of a spill and a dubious effect on spill size; US flag tankers and new tankers have a lower probability of causing spills, compared to foreign flag and old tankers, respectively.

1 Introduction

The sample selection model is usually estimated under the assumption that the disturbance terms are normally distributed. Since it is not always possible to work out the joint distribution of the disturbance terms when they are not jointly normally distributed, it is hard to test the sensitivity of the parameter estimates to distributional assumptions. However, bias due to misspecification of the distribution of the error terms can be substantial[1]. In this paper, we estimate a sample selection model for accidental oil spills by tankers without distributional assumptions on the error terms. We compare these parameter estimates and predictions with those generated under the normality assumption. We find that parameter estimates and predictions vary.

This paper is organized as follows. In the next section, we present some of the relevant work in the literature. In section three, we describe the model and the semiparametric estimation techniques we use, followed by a discussion of the results. A summary of major findings concludes the paper.

2 Literature Review

There is mixed evidence in the literature regarding the sensitivity of parameter estimates to distributional assumptions in the context of selection models. Ichimura (Ichimura 1987) provides Monte-Carlo evidence that bias due to misspecification of distribution can be substantial. Coslett's (Cosslett 1991) results also give evidence to this effect.

[1] For a discussion, see (Ichimura 1987) and (Cosslett 1991)

Mroz (Mroz 1987), however, tests the parameter estimates from his model of female labor supply with different distributional assumptions and finds them to be insensitive to distributional assumptions. Newey et al. (Newey, Powell, & Walker 1990) estimate the female labor supply model of Mroz (Mroz 1987) using semiparametric techniques and also find estimates that are close to Mroz's estimates.

In this paper, we estimate a sample selection model using the semiparametric estimation technique developed by Powell (Powell 1987). Powell proposes a two-step procedure analogous to Heckman's procedure (Heckman 1979) for the standard sample selection model with normal disturbances. $\sqrt{n}$-consistent and asymptotically normal, estimates are obtained if $\sqrt{n}$-consistent and asymptotically normal estimates for the sample selection effect are available. Klein and Spady (Klein & Spady 1993) and Ichimura (Ichimura 1987) propose $\sqrt{n}$-consistent and asymptotically normal estimates for the parameters of the sample selection effect. In the next section, we present the sample selection model and describe the estimation techniques we use.

3 Estimation

In this section we describe how a selection model for tanker oil spills is estimated with weak distributional assumptions for the error terms. The model is a selection model where the error distributions depend on a single index. $\sqrt{n}$-consistent and asymptotically normal estimates are obtained. This permits inference on the parameters using the central limit theorem and ensures that the estimation approach is not infinitely inefficient relative to parametric methods when the latter are correctly specified (Powell 1987).

The selection equation is:

$$z_i^* = w_i'\gamma + v_i \tag{1}$$

$$z_i = 1 \quad if z_i^* > 0; \tag{2}$$

$$z_i = 0 \quad otherwise.$$

The outcome equation is:

$$y_i = w_i'\beta + \epsilon_i, \quad observed\, only\, if\ z_i = 1; \tag{3}$$

where y_i is the natural logarithm of spill size in gallons, γ and β are vectors of coefficients, w_i is a vector of independent variables. v_i and ϵ_i are distributed independently of w_i. Since spill size is observed only when $z_i = 1$ and since $z_i = 1$ when $v_i > -w_i'\gamma$ for $z_i = 1$, the outcome equation can be rewritten as:

$$y_i = w_i'\beta + \lambda(w_i'\gamma) + u_i \tag{4}$$

where,

$$\lambda(w_i'\gamma) = E[\epsilon_i | v_i > -w_i'\gamma] \tag{5}$$

so that $E[u_i|w_i] = 0$. The additional term in (4) is referred to as the "correction" term. It corrects equation (3) by including the possible correlation between v_i and ϵ_i since $E(\epsilon_i)$ need not equal $E(\epsilon_i|v_i)$. Note that neither of the equations can contain an identifiable constant term because we do not impose any restrictions on the means of the error terms. To estimate the model, we follow a two-step procedure analogous to the two-step procedure of Heckman (Heckman 1979) for estimating the sample selection model with normal errors. Estimation of γ is based only on equation 1. We employ two different estimators for γ proposed by Ichimura (Ichimura 1987) and Klein and Spady (Klein & Spady 1993). We estimate the parameters of the outcome equation as suggested by Powell (Powell 1987). In the following sections we describe the estimators used.

3.1 The Klein and Spady Estimator

In this section we describe the estimator developed by Klein and Spady (Klein & Spady 1993). Let $P(z_i = 1|w_i'\gamma)$ denote the probability of the event $z_i = 1$ conditioned on $w_i'\gamma$. The probability function is unknown but it can be estimated nonparametrically by kernel methods. Let $\hat{P}(z_i = 1|w_i'\gamma)$ denote this estimate. The Klein and Spady estimator, referred to as the K&S estimator hereafter, maximizes the following quasi-likelihood function:

$$Q(\gamma) = \sum_i Q_i(\gamma) = \sum_i (\hat{\tau}_i/2)(z_i ln[\hat{P}(z_i = 1|w_i'\gamma)^2] + (1 - z_i) ln[(1 - \hat{P}(z_i = 1|w_i'\gamma))^2]). \quad (6)$$

The estimated probability function, $\hat{P}(z_i = 1|w_i'\gamma)$, is:

$$\hat{P}(z_i = 1|w_i'\gamma) = \frac{\hat{g}(z_i = 1, w_i'\gamma) + \hat{\delta}(z_i = 1, w_i'\gamma)}{\hat{g}(w_i'\gamma) + \hat{\delta}(w_i'\gamma)} \quad (7)$$

where $\hat{g}(z_i = 1, w_i'\gamma)$ is a kernel estimate of the joint density as in (8) and $\hat{g}(w_i'\gamma)$ is a kernel estimate of the marginal density as in (9).

$$\hat{g}(z_i = 1, w_i'\gamma) = \sum_j \frac{I(z_j = 1)}{N-1} \frac{1}{h} K\left[\frac{(w_i - w_j)'\gamma}{h}\right] \tag{8}$$

$$\hat{g}(w_i'\gamma) = \sum_j \frac{1}{N-1} \frac{1}{h} K[\frac{(w_i - w_j)\gamma}{h}] \tag{9}$$

where $I(.)$ is an indicator function, $K[.]$ is the kernel function and h is the bandwidth.

When estimated densities are too small, they pose problems for convergence. Klein and Spady use two types of trimming to avoid these problems. The estimated densities are trimmed through $\hat{\delta}$:

$$\hat{\delta}(x) = h^a[e^q/(1+e^q)] \tag{10}$$

$$\tag{11}$$

where $q = [(h^b - \hat{g}(x)/h^c]$, $0 < b < c$ and $1 > a > 2b + 2c > 0$. $\hat{\tau}_i$ trim the likelihood function by downweighting observations for which the estimated densities are too small.

$$\hat{\tau}_i \equiv \hat{\tau}_{i0}\hat{\tau}_{i1} \tag{12}$$

$$\hat{\tau}_{ik} = \{1 + exp[(h^{\epsilon/5} - \hat{g}(z_i = k, w_i'\hat{\gamma}_p))/h^{\epsilon/4}]\}^{-1}, k = 0, 1 \tag{13}$$

where $0 < \epsilon < a$ and $\hat{\gamma}_p$ is a consistent preliminary estimate which converges at rate $N^{-1/3}$.

If γ were used instead of the preliminary estimate, the gradient would depend on derivatives of τ, and this could lead to technical problems[2]. To get the preliminary estimates, we use Manski's (Manski 1975) Maximum Score Estimator (MSE) that maximizes the following function.

$$S(\gamma) = N^{-1} \sum_{n=1}^{N} [2I(z_i = 1) - 1] I(w_i' \gamma \geq 0) \tag{14}$$

Under certain conditions on the kernel function and bandwidth, this estimator is consistent and asymptotically normal. The kernel function should satisfy the following conditions:

$$\int K[u]du = 1 \tag{15}$$

$$K[u] = K[-u] \tag{16}$$

$$\int u^2 K[u]du = 0 \tag{17}$$

$$\{|\partial^r K[u]/\partial u^r|, \int |\partial^r K[u]/\partial u^r| du\} < c(r = 0, 1, 2, 3, 4) \tag{18}$$

We use the following kernel function which is of order four[3].

$$K[u] = 3/2 * k[u] - 1/2 * u^2 * k[u] \tag{19}$$

where k[u] is the standard normal density function. The bandwidth parameter h should lie

[2]For a more in depth discussion see (Klein & Spady 1993)

[3]The order of a kernel function is its first nonzero moment

between $N^{-1/6}$ and $N^{-1/8}$. Then

$$\sqrt{N}(\hat{\gamma} - \gamma_0) \sim N(0, \Delta) \tag{20}$$

$$\Delta = E\left(\left[\frac{\partial P}{\partial \gamma}\right]\left[\frac{\partial P}{\partial \gamma}\right]'\left[\frac{1}{P(1-P)}\right]\right)^{-1} \tag{21}$$

and Δ can be consistently estimated as in White (White 1982).

$$\hat{\Delta} = A(\hat{\gamma})^{-1}B(\hat{\gamma})A(\hat{\gamma})^{-1}, \tag{22}$$

$$A(\hat{\gamma}) = N^{-1}\sum \partial^2 Q(\hat{\gamma})/\partial\gamma_i\partial\gamma_j, \tag{23}$$

$$B(\hat{\gamma}) = N^{-1}\sum \partial Q(\hat{\gamma})/\partial\gamma_i \partial Q(\hat{\gamma})/\partial\gamma_j \tag{24}$$

3.2 Ichimura's Estimator

Ichimura's estimator is a minimum distance estimator and minimizes the expected conditional variance of z_i given $w'\gamma$. The intuition behind it is as follows. The variation in z has two sources, $w'\gamma$ and v. Along $w'\gamma = c$ where c is a constant, the only variation in z is because of the variation in v. It is thus natural to try and identify γ by minimizing the conditional variance. The estimator minimizes the following function:

$$J(\gamma) = N^{-1}\sum_{i=1}^{N}(z_i - \hat{E}_i(\gamma))^2 \tag{25}$$

where $\hat{E}_i(\gamma)$ is a kernel estimator for $E(z_i|w_i'\gamma)$ given by equation 26 below.

$$\hat{E}_i(\gamma) = \frac{\sum_{j \neq i} z_j K[\frac{w_i'\gamma - w_j'\gamma}{h}]}{\sum_{k \neq i} K[\frac{w_i'\gamma - w_k'\gamma}{h}]} \tag{26}$$

If the kernel function and the bandwidth are chosen appropriately, the estimator has the following asymptotic distribution:

$$\sqrt{N}(\hat{\gamma} - \gamma) \sim N(0, , ^{-1}\Phi(, ')^{-1}) \tag{27}$$

$, ^{-1}\Phi(, ')^{-1}$ can be consistently estimated by:

$$N \left[\sum_i \frac{\partial \hat{E}_i(\hat{\gamma})}{\partial \gamma} (\frac{\partial \hat{E}_i(\hat{\gamma})}{\partial \gamma})' \right]^{-1} \sum_i (z_i - \hat{E}_i(\hat{\gamma}))(z_i - \hat{E}_i(\hat{\gamma}))' \frac{\partial \hat{E}_i(\hat{\gamma})}{\partial \gamma} \left(\frac{\partial \hat{E}_i(\hat{\gamma})}{\partial \gamma} \right)' \left[\sum_i \frac{\partial \hat{E}_i(\hat{\gamma})}{\partial \gamma} (\frac{\partial \hat{E}_i(\hat{\gamma})}{\partial \gamma})' \right]^{-1} \tag{28}$$

Neither of the two estimators allow for a constant term in $w_i'\gamma$. It is further necessary to place a restriction on parameters for identification, for example set the first parameter to a scalar $|c|$.

3.3 Powell's Estimator

Given $\hat{\gamma}$, β in the outcome equation are estimated by comparing the pairs of observations for which the estimated indices are close.

$$y_i - y_j = (w_i - w_j)'\beta + (\epsilon_i - \epsilon_j) \tag{29}$$

$\hat{\beta}$ can be estimated by weighted least squares regression using $1/hK[(w_i - w_j)\hat{\gamma}/h]$ as weights. The estimates are $\sqrt{n}$-consistent if the kernel function, $K[.]$, and the bandwidth, h, are chosen appropriately. We use the same kernel function and bandwidth as in the Klein and Spady estimator. Powell (Powell 1987) derives the covariance matrix. Let

$$\xi = g(w_i'\gamma)(w_i - E(w_i|w_i'\gamma))u_i \tag{30}$$

where $g(w_i'\gamma)$ is the density of the single index,

$$\Omega = E[\lambda'(w_i'\gamma)g(w_i'\gamma)(w_i - E(w_i|w_i\gamma))^2] \tag{31}$$

and

$$\Sigma = E[g(w_i'\gamma)\ (w_i - E(w_i|w_i'\gamma))\ (w_i - E(w_i|w_i'\gamma))']. \tag{32}$$

Also suppose that $\hat{\gamma}$ can be written as

$$\hat{\gamma} = \gamma_0 + 1/N \sum_{i=1}^{N} \psi(z_i, w_i, \gamma_0) + o_p(N^{-1/2}), \tag{33}$$

such that $E(\psi(z_i, w_i, \gamma_0)) = 0$ and $E(\psi(z_i, w_i, \gamma_0)^2) < \infty$. Then

$$\sqrt{n}(\hat{\beta} - \beta) \sim N(0, V) \tag{34}$$

where

$$V = \Sigma^{-1}[E(\xi\xi') - \Omega E(\psi\xi') - E(\xi\psi')\Omega' + \Omega E(\psi\psi')\Omega'][\Sigma^{-1}]'. \tag{35}$$

Σ can be consistently estimated by:

$$1/n \sum_{i=1}^{n} \hat{g}(w_i'\hat{\gamma})(w_i - \hat{E}(w_i|w_i'\hat{\gamma}))\ (w_i - \hat{E}(w_i|w_i'\hat{\gamma})) \tag{36}$$

where $\hat{g}(w_i'\hat{\gamma})$ and $\hat{E}(w_i|w_i\hat{\gamma})$ are consistent estimators of $g(w_i'\gamma)$ and $E(w_i|w_i'\gamma)$ respectively. Ω can be consistently estimated by the following:

$$1/n \sum_{i=1}^{n} \hat{\lambda}'(w_i'\hat{\gamma})\hat{g}(w_i'\hat{\gamma})(w_i - \hat{E}(w_i|w_i'\hat{\gamma}))(w_i - \hat{E}(w_i|w_i'\hat{\gamma}))' \tag{37}$$

$E(\xi\xi')$ can be consistently estimated by:

$$1/n \sum_{i=1}^{n} \hat{g}(w_i'\hat{\gamma})^2(w_i - \hat{E}(w_i|w_i'\hat{\gamma}))(w_i - \hat{E}(w_i|w_i'\hat{\gamma}))'\hat{\sigma}^2 \tag{38}$$

where

$$\hat{\sigma}^2 = 1/n\hat{u}'\hat{u} = 1/n(y - w\hat{\beta} - \hat{\lambda}(w\gamma))'(y - w\hat{\beta} - \hat{\lambda}(w\gamma)) \tag{39}$$

$$\hat{\lambda}(w_i'\gamma) = \hat{E}(y_i|w_i'\gamma) - \hat{E}(w_i|w_i'\gamma)\hat{\beta} \tag{40}$$

$\hat{E}(y_i|w_i'\gamma)$ and $\hat{E}(w_i|w_i'\gamma)$ are kernel estimators of the form in (26). $E(\psi\xi')' = E(\xi\psi')$ can be consistently estimated by:

$$(1/n\sum_{i=1}^{n}\hat{g}(w_i'\hat{\gamma})(w_i - \hat{E}(w_i|w_i'\hat{\gamma}))\hat{u}_i)\hat{\psi}(i) \tag{41}$$

For $\hat{\gamma}$ based on Ichimura's technique

$$\psi(i) = 2,\ ^{-1}(z_i - \hat{E}_i(\gamma_0))\frac{\partial\hat{E}_i(\gamma_0)}{\partial\gamma} \tag{42}$$

$$E(\psi(i)\psi(i)') = \frac{4,\ ^{-1}\Phi(,\ ')^{-1}}{N} \tag{43}$$

For $\hat{\gamma}$ based on Klein and Spady

$$\psi(i) = \Delta\frac{\partial Q_i(\gamma_0)}{\partial\gamma} \tag{44}$$

$$E(\psi(i)\psi(i)') = \frac{\Delta}{N} \tag{45}$$

4 Discussion of Results

In this section we discuss our results. The section is organized in two parts. First parameter estimates and predictions for the selection stage are presented and compared. The normality assumption is tested and rejected. Further the results are found to be sensitive to the semiparametric technique used. Then the second-stage estimates are discussed.

4.1 First-stage estimates

Results of first-stage estimates are presented in Table 1. For identification the coefficient of the weather dummy was set to its probit estimate. The bandwidth parameter h was set to 0.3. If the disturbances for the selection equation are normally distributed, the probit estimates and the semiparametric estimates should be "close". The following test statistic can be used to test this. If the probit model is correct :

$$\sum_{i=1}^{k}\left(\frac{\hat{\gamma}_i^{prbt}-\hat{\gamma}_i^{sp}}{\sigma_{\hat{\gamma}_i^{prbt}-\hat{\gamma}_i^{sp}}}\right)\sim\chi^2(k) \tag{46}$$

where $prbt$ and sp stand for probit and semiparametric estimates respectively. The test statistic for the K&S estimates and probit estimates is 64.8121 which is above the critical values for $\chi^2(6)$. The test statistic for the Ichimura estimates is 5708.441. So the assumption of normal disturbances in the selection equation is rejected by both semiparametric estimators.

Notable differences between the parameter estimates are the signs of the significant coefficients on the flag dummy and the age variable. In all three estimation techniques, these coefficients are significant but the Ichimura estimates have different signs. The size variable is only significant for K&S estimates. The Ichimura technique also produces a significant negative coefficient on the improper maintenance dummy. These differences in parameters also yield different predictions. These are shown in Table 2. The Ichimura estimates produce

Table 1: Estimation Results for Oil Spills
Standard errors are in parenthesis

Explanatory Variables	Probit	Klein & Spady	Ichimura
Type of Accident Variables			
GRNDING	-1.8986***	-0.4916***	-5.6454***
	(0.1242)	(0.0189)	(0.2828)
COLLIS	-1.4921***	-0.4791***	-7.9669***
	(0.1364)	(0.0208)	(0.4026)
WEATHER	-0.0503	-0.0503	-0.0503
	(0.2049)	(0.0631)	(0.688)
IMPMAINT	0.0068	0.0304	-1.3397***
	(0.2057)	(0.0599)	(0.1816)
Vessel Characteristic Variables			
SIZE	-1.7e-03	-1.7e-03***	-1.12e-03
	(1.15e-03)	(0.000268)	(0.001252)
AGE	0.0103***	0.0038***	-0.0608***
	(0.0031)	(0.0008)	(0.00517)
US	-0.7598***	-0.3290***	2.9007***
	(0.0677)	(0.0437)	(0.0201)
Constant	0.3069***	NA	NA
	(0.0784)		
N	2263	2263	2263
Log Likelihood	-1178.2977	-105.6232	NA

*** significant at the 1% level.
** significant at the 5% level.
* significant at the 10% level.
Note: The probit model was also estimated without a constant and this restriction was tested and rejected by a likelihood ratio test. The value of the natural logarithm of the likelihood function with no constant term is -1186.0421.

Table 2: Probability of a Spill for Different Vessels
US flag vessels are in parenthesis

	Probit	K & S	Ichimura
Base Case[1]	0.6580	0.5224	0.6953
	(0.3622)	(0.4153)	(0.3310)
Large Vessel	0.6146	0.4879	0.6991
(100 000 tons)	(0.3197)	(0.3814)	(0.3377)
Old Vessel	0.6950	0.5313	0.2966
(25 years)	(0.4014)	(0.4278)	(0.4421)
GRNDING	0.0679	0.3700	0.0315
	(0.0122)	(0.3604)	(0.0367)
COLLIS	0.1389	0.3729	0.1624
	(0.0325)	(0.3578)	(0.0629)
WEATHER	0.6393	0.5086	0.6980
	(0.3434)	(0.3997)	(0.3350)
IMPMAINT	0.6605	0.5296	0.9489
	(0.3674)	(0.4252)	(0.3159)

[1] 15 year old vessel that weighs 32000 gross tons

a lower probability of spill conditional on an accident for old foreign flag tankers, which conflicts with the other results. The other estimates indicate that older vessels have a slightly higher probability of causing a spill, and US flag tankers have a much lower probability of causing a spill. The Ichimura estimates also produce a very high probability of a spill for accidents caused by improper maintenance for foreign flag tankers which is not replicated by the K&S and probit estimates.

Figures 1- 3 show the predicted values for the sample. It is interesting to note that the range of predictions based on K&S estimates is between 0.35 and 0.55, quite narrow in comparison to predictions from the other two methods. Ichimura's estimates produce probabilities that range from 0 to 1[4] and predictions based on probit range from 0 to 0.8.

[4] Ichimura estimates actually produce some negative probabilities

Table 3: Estimation Results for Oil Spills
Standard errors are in parenthesis

Explanatory Variables	Dependent Variable: Log Spill Size		
	Normal Errors	Kernel Estimates	
		Klein and Spady	Ichimura
Type of Accident Variables			
GRNDING	24.3496***	6.4140	NA
	(6.2962)	(26.2618)	NA
COLLIS	17.3673***	4.6331	-4.8187
	(4.8117)	(27.1600)	(14.4113)
WEATHER	0.3504	-0.4304	-0.2429
	(1.8171)	(18.8552)	(20.4737)
IMPMAINT	-0.0909	-0.0894	-1.6599
	(1.7425)	(16.7959)	(59.1457)
Vessel Characteristic Variables			
SIZE	0.02*	0.0057	0.0089
	(0.000012)	(0.1184)	(0.1362))
AGE	-0.0747*	0.0192	0.0149
	(0.0407)	(0.3085)	(2.3069)
US	4.8048**	-1.2915	-0.3569
	(2.1223)	(10.7985)	(110.6544)
Constant	10.6039***	NA	NA
	(2.7277)		
lambda	-12.5212***	NA	NA
	(4.2647)		
N	784	784	784

*** significant at the 1% level.
** significant at the 5% level.
* significant at the 10% level.

4.2 Second-stage estimates

The second-stage estimates are presented in Table 3. The estimated weights for grounding and collision incidents were very close when Ichimura estimates were used for the first-stage estimates so the grounding dummy was dropped. These results can be found in Table 4.

Table 4 compares the effect of accident types, as well as vessel characteristics such as age and size, on spill size for foreign and US flag vessels. The base case is a 15 year old tanker

Table 4: Spill Size
US flag vessels are in parenthesis

	Normal Errors	K & S	Ichimura
Base Case [1]	22.9873	24.9679	23.8301
	(7.1430)	(6.8328)	(6.1571)
Large Vessel	40.70	37.1898	43.2827
(100 000 tons)	(9.03)	(9.7575)	(11.4493)
Old Vessel	21.4466	30.0644	15.9136
(25 years)	(8.2923)	(8.3318)	(9.2154)
GRNDING	29373.9633	14420.4811	NA
	(855.6440)	(1404.7197)	NA
COLLIS	1873.5230	2443.5233	1809.7720
	(73.3556)	(291.2212)	(142.5253)
WEATHER	23.1565	16.3372	18.6044
	(6.4801)	(4.3936)	(4.8760)
IMPMAINT	21.9800	22.7211	17.7674
	(6.9259)	(6.2811)	(1.3880)

[1] 15 year old vessel that weighs 32000 gross tons

that weighs 32,000 tons. Mean spill size for the base case ranges between 23 gallons and 25 gallons for all methods for foreign flag tankers. For US flag tankers the range is even narrower, 6 to 7 gallons. Mean spill size for large tankers is twice as much, around 40 gallons for foreign flag tankers. Old vessels are predicted to have smaller spills by Ichimura and under the assumption of normality, but bigger spills if we use K&S estimates for the selection effect. For US flag tankers, all methods show an increase in spill size with age. Spills due to improper maintenance and adverse weather are smaller.

Although there are qualitative differences, the predicted mean spill sizes do not differ in a meaningful way for policy purposes in any of the above cases. For spills due to grounding and collision, they do. Mean spill for for grounding is almost 30,000 gallons for foreign tankers with the normality assumption and only half that amount based on semiparametric methods. Estimates of mean spill size for spills due to collision range between 1800 to 2500 gallons. These differences are large enough to make a difference for policy purposes. So it is important to identify which of the techniques produces more reliable estimates. The normality assumption is too restrictive and is rejected by both the semiparametric techniques. Among the semiparametric techniques we prefer to use the K&S estimator for the first stage. Ichimura has not studies how the bandwidth and the kernel function should be chosen. The choice does not matter for asymptotic properties, but could matter in small samples.

5 Conclusion

In this paper we investigate the bias due to misspecification of the distribution of the error terms in a sample selection model for oil tanker spills. We find that parameter estimates are sensitive to distributional assumptions.

Major findings that are qualitatively supported by all methods are the following. Groundings and collisions result in larger spills if there is a spill, but the likelihood that there will be a spill due to a grounding or collision is very low. Tanker size has only a marginal effect on the probability of a spill and a dubious effect on spill size. US flag tankers and new tankers have a lower probability of causing spills, compared to foreign flag and old tankers, respectively.

It would be useful to investigate how sensitive Ichimura's estimator is to different bandwidths and kernel functions for this sample. If it is found to be sensitive, this would provide further support for choosing K&S estimates over those of Ichumura.

A Linear Representation of First Stage Estimates

A.1 Ichimura

$$\hat{\gamma} = \gamma_0 - , {}^{-1} \frac{\partial J(\gamma_0)}{\partial \gamma} + o_p(n^{-1/2}) \tag{47}$$

$$\frac{\partial J(\gamma_0)}{\partial \gamma} = -1/n \sum_{i=1}^{n} 2(z_i - \hat{E}_i(\hat{\gamma}_0)) \frac{\partial \hat{E}_i(\hat{\gamma}_0)}{\partial \gamma} \tag{48}$$

A.2 Klein and Spady

$$\hat{\gamma} = \gamma_0 + \Delta 1/n \sum_{i=1}^{n} \hat{\tau}_i \hat{r}_i \hat{w}_i + o_p(n^{-1/2}) \tag{49}$$

$$\tag{50}$$

$$\hat{r}_i = \frac{z_i - \hat{P}_i}{\hat{g}(\gamma_0) + \hat{\delta}(\gamma_0)\hat{P}_i(1 - \hat{P}_i)} \tag{51}$$

$$\hat{w}_i = (\hat{g}(\gamma_0) + \hat{\delta}(\gamma_0)) \frac{\partial \hat{P}}{\partial \gamma} \tag{52}$$

$$\hat{\gamma} = \gamma_0 + \Delta 1/n \sum_{i=1}^{n} \frac{\partial Q_i(\gamma_0)}{\partial \gamma} + o_p(n^{-1/2}) \tag{53}$$

$$\tag{54}$$

References

Cosslett, S. 1991. Semiparametric estimation of a regression model with sample selectivity. In Barnett, W.; Powell, J.; and Tauches, G., eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, 175–197.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–161.

Ichimura, H. 1987. *Estimation of Single Index Models.* Ph.D. Dissertation, MIT.

Klein, R., and Spady, R. S. 1993. An efficient semiparametric estimator for binary response models. *Econometrica* 61(2):387–421.

Manski, C. F. 1975. The maximum score estimator of the stochastic utility model of choice. *Journal of Econometrics* 3:205–228.

Mroz, T. A. 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* 55:765–799.

Newey, W. K.; Powell, J.; and Walker, J. 1990. Semiparametric estimation of selection models: Some empirical results. *American Economic Association Papers and Proceedings* 324–328.

Powell, J. 1987. Semiparametric estimation of bivariate latent variable models. Working Paper 8704, SSRI, University of Wisconsin-Madison.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50:1–25.

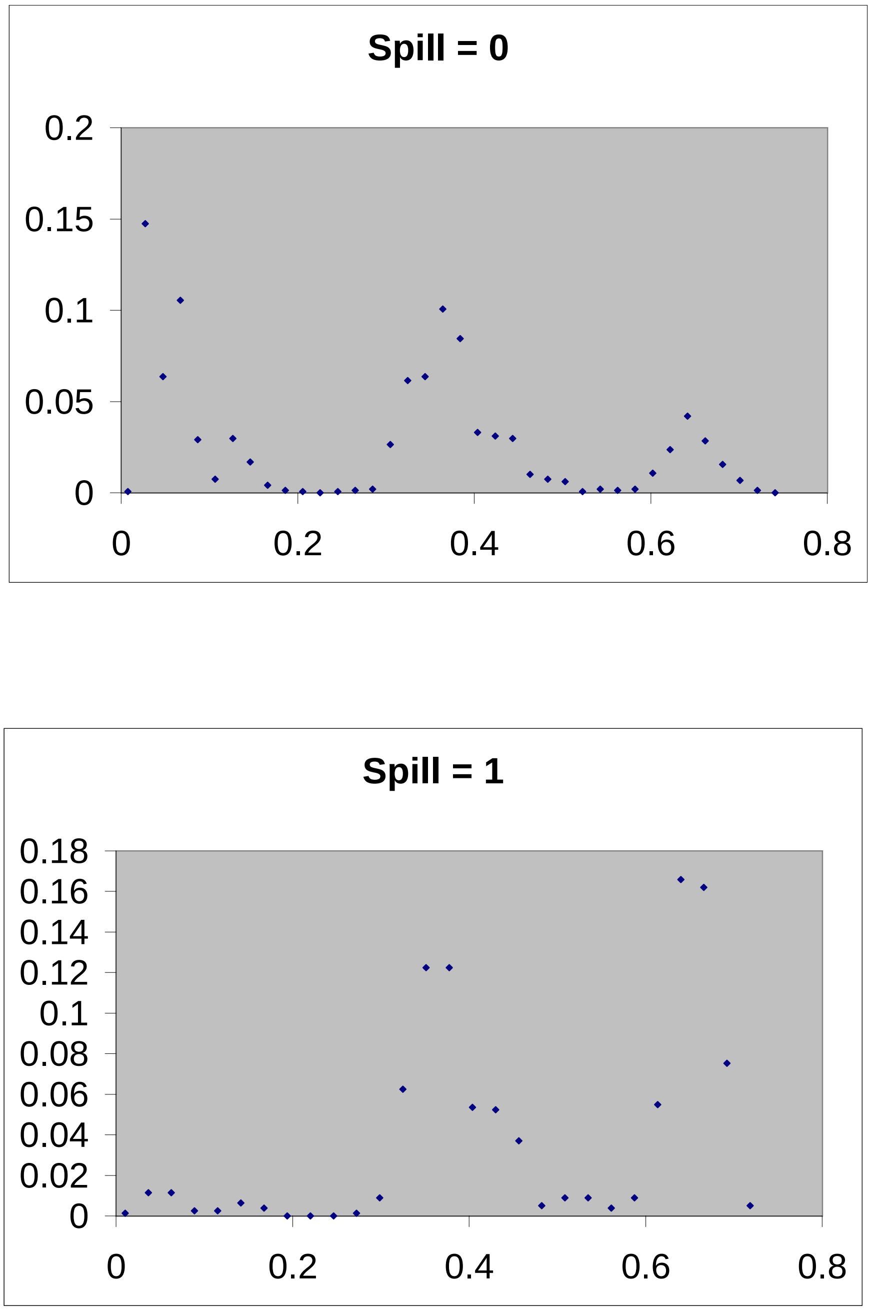
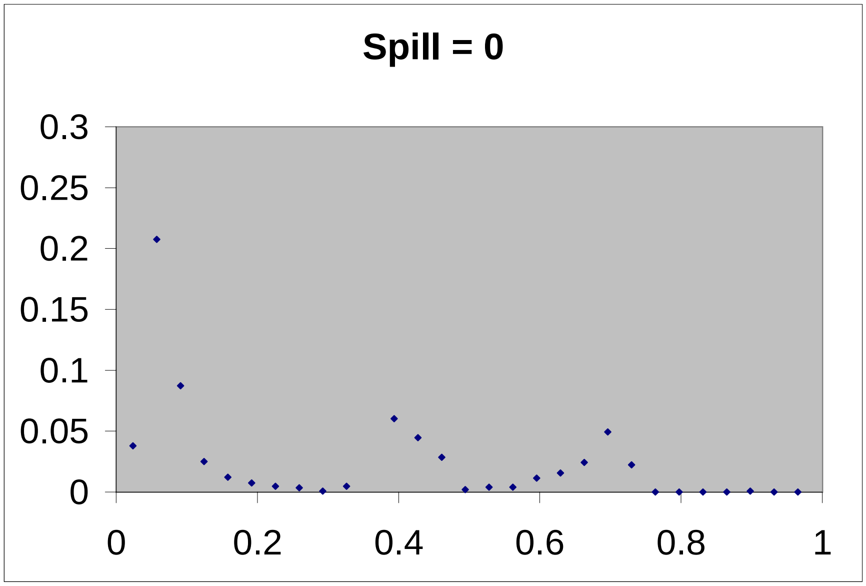


Figure 1: Probit Predicted Probabilities



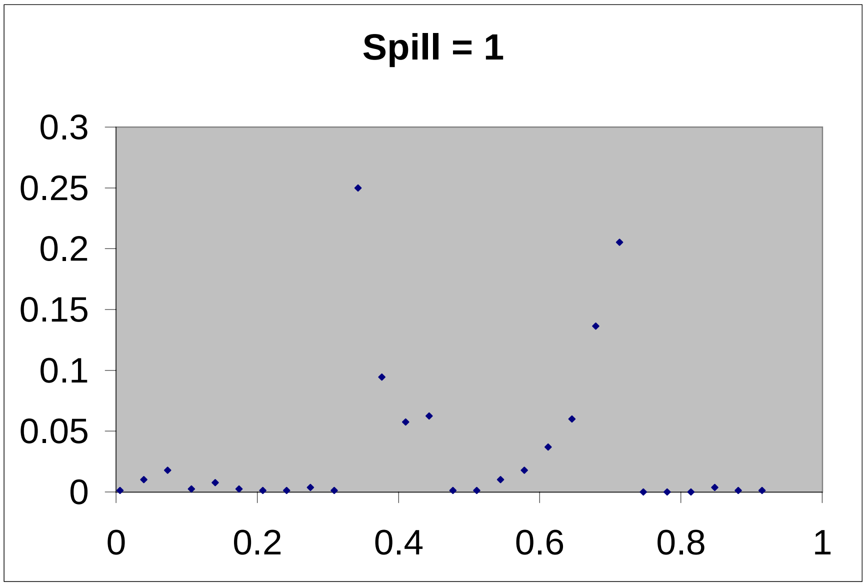
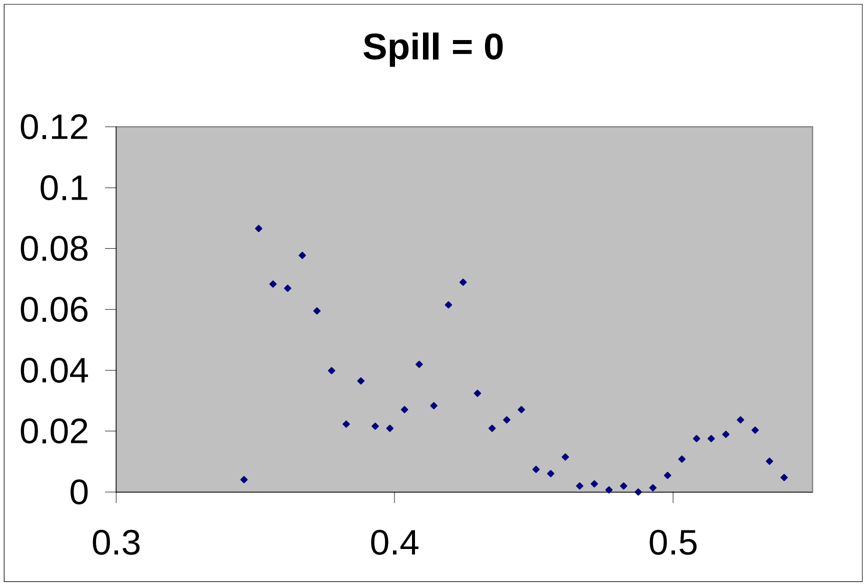
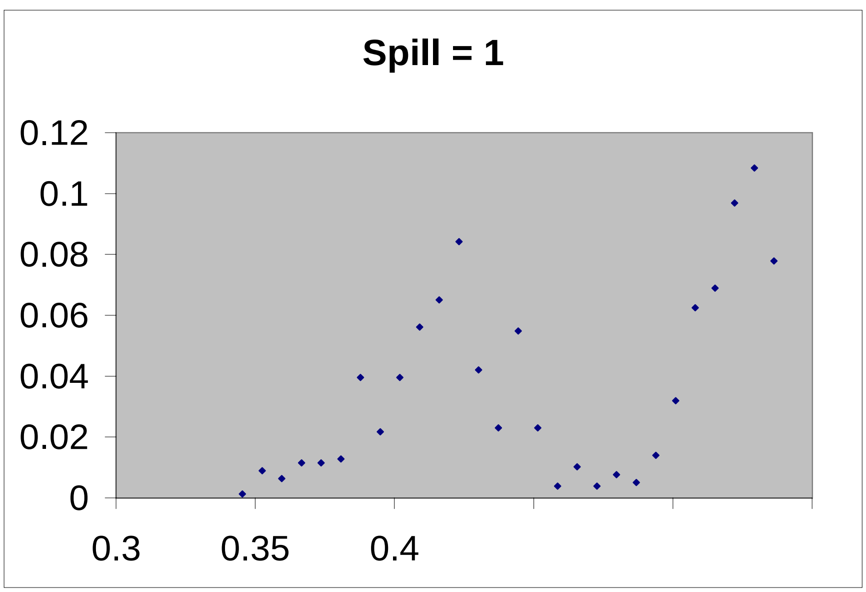


Figure 2: Semiparametric Predicted Probabilities - Ichimura


Spill = 0
0.12
0.1
0.08
0.06
0.04
0.02
0
0.3
0.4
0.5


Spill = 1
0.12
0.1
0.08
0.06
0.04
0.02
0
0.3
0.35
0.4