

## Distribution of Fraud

The following derives analytic expressions for  $P(F|x)$  where  $x$  is a continuous variable and will often be of the form  $x = \ln(z)$  where  $z$  is some measurable continuous variable like 'spend'.

Given that 
$$P(F | x) = \frac{1}{1 + \frac{P(x|L)}{P(x|F)} c} \text{ where } c = \frac{P(L)}{P(F)} \quad (1)$$

It is reasonable to assume that  $P(x|L)$  and  $P(x|F)$  are normally distributed. Even the sparse fraud data will be approximately normal when not over partitioned which can be avoided through clustering of categorical variables and banding of continuous variables.

We can therefore write:

$$P(F | x) = \frac{1}{1 + \left( \frac{\sigma_F e^{-\frac{1}{2} \left( \frac{x-\mu_L}{\sigma_L} \right)^2}}{\sigma_L e^{-\frac{1}{2} \left( \frac{x-\mu_F}{\sigma_F} \right)^2}} \right) c} = \frac{1}{1 + c \frac{\sigma_F}{\sigma_L} e^{\frac{1}{2} \left[ \left( \frac{x-\mu_F}{\sigma_F} \right)^2 - \left( \frac{x-\mu_L}{\sigma_L} \right)^2 \right]} \quad (2)$$

and after a bit of algebra (see appendix A) and using result A2 this can be re-written as:

$$P(F | x) = \frac{1}{1 + r e^{\frac{(x-\mu)^2}{2\sigma^2}}} \quad (3)$$

where 
$$r = c \frac{\sigma_F}{\sigma_L} e^{\frac{1}{2} \frac{(\mu_L - \mu_F)^2}{(\sigma_L^2 - \sigma_F^2)^2}} \quad \mu = \frac{\sigma_F^2 \mu_L - \sigma_L^2 \mu_F}{\sigma_F^2 - \sigma_L^2} \quad \text{and } \sigma = \sigma_L \sigma_F$$

where  $r$  is large we can then use the result B5 from appendix B and write:

$$P(F | x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{r} \quad (4)$$

## Notes

1. where  $\mu_L = \mu_F = \mu_{LF}$  and  $\sigma_L = \sigma_F = \sigma_{LF}$

then  $\mu = \mu_{LF}$  and  $r = c$  and as  $c \gg 1$  we have  $P(F | x) \approx \frac{1}{r}$  (5)

2. where  $\mu_L = \mu_F = \mu_{LF}$

then  $\mu = \mu_{LF}$  and  $r = c \frac{\sigma_F}{\sigma_L}$  and as  $O(\sigma_F) = O(\sigma_L)$  and  $c \gg 1$  then

$r$  is large and so  $P(F | x) \approx \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{r}$  (6)

3. where  $\sigma_L = \sigma_F = \sigma_{LF}$  then equation A1 reduces to:

$$T(x) = \left( \frac{x - \mu_F}{\sigma_{LF}} \right)^2 - \left( \frac{x - \mu_L}{\sigma_{LF}} \right)^2 = \frac{2(\mu_L - \mu_F)x - (\mu_L^2 - \mu_F^2)}{\sigma_{LF}^2} = \frac{2(\mu_L - \mu_F) \left( x - \frac{1}{2}(\mu_L + \mu_F) \right)}{\sigma_{LF}^2}$$

$$\text{and } P(F | x) = \frac{1}{1 + ce^{-\frac{(x-\mu)^2}{2\sigma^2}}} \quad (7)$$

$$\text{where } \mu = \frac{\mu_L + \mu_F}{2} \quad \text{and } \sigma = \frac{\sigma_{LF}^2}{4(\mu_L - \mu_F)}$$

and for large  $c$  we can use B6 and write:

$$P(F | x) \approx \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{c} \quad (8)$$

So where  $\sigma_L = \sigma_F = \sigma_{LF}$  then the distribution is exponential and

if  $\mu_L < \mu_F$  then  $P(F|x)$  grows with increasing  $x$ .

if  $\mu_L > \mu_F$  then  $P(F|x)$  decays with increasing  $x$ .

if  $\mu_L = \mu_F$  then see note 1 above.

## Appendix A

Consider the term below from equation (2)

$$T(x) = \left( \frac{x - \mu_F}{\sigma_F} \right)^2 - \left( \frac{x - \mu_L}{\sigma_L} \right)^2 = \frac{(\sigma_L^2 - \sigma_F^2)x^2 - 2(\sigma_L^2\mu_F - \sigma_F^2\mu_L)x + (\sigma_L^2\mu_F^2 - \sigma_F^2\mu_L^2)}{\sigma_L^2\sigma_F^2} \quad (\text{A1})$$

then using  $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$  we have

$$\frac{2(\sigma_L^2\mu_F - \sigma_F^2\mu_L) \pm \sqrt{4(\sigma_L^2\mu_F - \sigma_F^2\mu_L)^2 - 4(\sigma_L^2 - \sigma_F^2)(\sigma_F^2\mu_L^2 - \sigma_L^2\mu_F^2)}}{2(\sigma_L^2 - \sigma_F^2)\sigma_L^2\sigma_F^2}$$

which reduces to:  $\frac{(\sigma_L^2\mu_F - \sigma_F^2\mu_L) \pm \sigma_L\sigma_F(\mu_F - \mu_L)}{(\sigma_L^2 - \sigma_F^2)\sigma_L^2\sigma_F^2}$

Now if we set  $p = \frac{(\sigma_L^2\mu_F - \sigma_F^2\mu_L)}{(\sigma_L^2 - \sigma_F^2)}$  and  $q = \frac{\sigma_L\sigma_F(\mu_F - \mu_L)}{(\sigma_L^2 - \sigma_F^2)}$

Then, as we can always write:  $(x - (u + v))(x - (u - v)) = (x - u)^2 - v^2$

We have  $T(x) = \frac{(x - (p + q))(x - (p - q))}{\sigma_L^2\sigma_F^2} = \frac{(x - p)^2 - q^2}{\sigma_L^2\sigma_F^2}$

So substituting for  $p$  and  $q$  we have:

$$T(x) = \left( \frac{x - \frac{\sigma_L^2\mu_F - \sigma_F^2\mu_L}{\sigma_L^2 - \sigma_F^2}}{\sigma_L^2\sigma_F^2} \right)^2 - \frac{(\mu_F - \mu_L)^2}{(\sigma_L^2 - \sigma_F^2)^2} \quad (\text{A2})$$

## Appendix B

Consider a function of the form:  $p(x) = \frac{1}{1+re^{kx^2}}$  (B1)

Then if  $k > 0$  then we can show that for large  $r$  that  $p(x)$  converges to the form:

$$q(x) = \frac{e^{-kx^2}}{r} \quad (B2)$$

Let  $\frac{1}{1+re^{kx^2}} = \frac{e^{-(k+s(x))x^2}}{1+r}$  where  $s(x)$  is an arbitrary function of  $x$ . (B3)

then the LHS and RHS of B3 are coincident at  $x = 0$  and  $x = \pm\infty$  for any  $s(x)$

Rearranging B3 we have:  $1+r = e^{-s(x)x^2} e^{-kx^2} (1+re^{kx^2}) = e^{-s(x)x^2} (e^{-kx^2} + r)$  (B4)

As  $k > 0$  then we can assume that  $e^{-kx^2} \leq 1 \ll r$  and the last term in B4 can therefore be approximated by just  $r$  for large  $r$  and B4 can be written as:

$$e^{-s(x)x^2} = \frac{1+r}{r} \rightarrow 1 \text{ for large } r, \text{ which implies } s(x) \rightarrow 0 \text{ for large } r \text{ and hence } p(x) \rightarrow q(x)$$

So for large  $r$  we have:

$$p(x) = \frac{1}{1+re^{kx^2}} \rightarrow \frac{e^{-kx^2}}{1+r} \rightarrow \frac{e^{-kx^2}}{r} \quad (B5)$$

where  $x > 0$  for all  $x$  then we can also write:

$$\frac{1}{1+re^{kx}} \rightarrow \frac{e^{-kx}}{1+r} \rightarrow \frac{e^{-kx}}{r} \quad (B6)$$