Maximum Value of the Correlation Between a Standard Normal Random Variable and a Binomial Random Variable

Yan Zeng

October 22, 2009

Abstract

Investigation of the range of $E[XY]$ where Y is a binomial random variable. This is a problem arising from credit risk modelling.

1 The problem

Given a standard normal random variable X and another binomial random variable Y such that $P(Y = 1) = p$, $P(Y = 0) = q := 1 - p$ $(0 < p < 1)$. *What is the maximum possible value of the correlation between X and Y?* This problem arises from credit risk modeling and is asked by Marcelo Piza at Bloomberg's quant group.

Note the standard deviation of X and Y are 1 and $\sqrt{pq}$, respectively. So the correlation $\rho(X, Y)$ is equal to

$$\rho(X,Y) = \frac{E\left\{(X - E[X])(Y - E[Y])\right\}}{\sqrt{pq}} = \frac{E[XY]}{\sqrt{pq}}.$$

Therefore, the problem is really about *what is the maximum possible value of $E[XY]$?*

2 Analysis of the problem

Denote by $f_{X|Y}(x, y)$ the conditional density of X given Y. We have

$$E[XY] = E[YE[X|Y]] = pE[X|Y = 1] = p\int_{\mathbb{R}} xf_{X|Y}(x, 1)dx.$$

We now look for the constraints $f_{X|Y}(x, 1)$ must satisfy. Define $\phi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$, we have

$$\phi(x)dx = P(X \in dx) = E[P(X \in dx|Y)] = pf_{X|Y}(x, 1)dx + qf_{X|Y}(x, 0)dx.$$

So $\phi(x)$ is the convex combination of two probability density functions. This motivates us to prove the following proposition.

Proposition 1. *A function $h(x) : \mathbb{R} \to \mathbb{R}$ can be used for $f_{X|Y}(x, 1)$ if and only if $h(x)$ satisfies the following set of conditions:*
(1) h is a probability density function, i.e. $h(x) \geq 0$ a.s. and $\int_{\mathbb{R}} h(x)dx = 1$;
(2) $h(x) \leq \frac{\phi(x)}{p}$.

Proof. The necessity is obvious. To prove the sufficiency, it suffices to show $\frac{1}{q}[\phi(x) - ph(x)]$ can be used for $f_{X|Y}(x, 0)$, i.e. $\frac{1}{q}[\phi(x) - ph(x)]$ is a probability density function. Indeed, we note condition (2) implies $\phi(x) - ph(x) \geq 0$ and

$$\int_{\mathbb{R}} \frac{1}{q}[\phi(x) - ph(x)]dx = \frac{1}{q} - \frac{p}{q} = 1.$$

Combined, we can conclude any $h(x)$ satisfying condition (1) and (2) can be used for $f_{X|Y}(x, 1)$. □

In view of Proposition 1, we can reformulate the original problem as follows. *Define* $\Lambda := \{h(x) \in \mathcal{B}(\mathbb{R}) : h(x) \geq 0, \int_{\mathbb{R}} h(x)dx = 1, h(x) \leq \frac{\phi(x)}{p}\}$, *solve the following optimization problem*:

$$h_0(x) = \arg\max_{h \in \Lambda} \int_{\mathbb{R}} xh(x)dx.$$

3 Solution

Proposition 2. *Let* $h \in \Lambda$. *Suppose there exists a pair* (x_1, x_2) *of continuity points of* $h(x)$, *such that* $x_1 < x_2$, $h(x_1) > 0$ *and* $h(x_2) < \frac{\phi(x)}{p}$. *Then we can find* $h_1(x) \in \Lambda$, *such that*

$$\int_{\mathbb{R}} xh(x)dx < \int_{\mathbb{R}} xh_1(x)dx.$$

Proof. By the continuity of $h(x)$ at x_1 and x_2, we can find $\varepsilon > 0$ and $\delta > 0$, such that $h(x) \geq \varepsilon$ on $[x_1 - \delta, x_1 + \delta]$ and $h(x) \leq \frac{\phi(x)}{p} - \varepsilon$ on $[x_2 - \delta, x_2 + \delta]$. Define

$$h_1(x) = \begin{cases} h(x) - \varepsilon & x \in [x_1 - \delta, x_1 + \delta] \\ h(x) + \varepsilon & x \in [x_2 - \delta, x_2 + \delta] \\ h(x) & \text{otherwise.} \end{cases}$$

Then $h_1 \in \Lambda$ and

$$\int_{\mathbb{R}} xh(x)dx - \int_{\mathbb{R}} xh_1(x)dx = \int_{x_1-\delta}^{x_1+\delta} \varepsilon x dx + \int_{x_2-\delta}^{x_2+\delta} (-\varepsilon)xdx = 2\delta\varepsilon(x_1 - x_2) < 0.$$

□

Corollary 1. *If* $h \in \Lambda$ *has a continuity point* x_0 *at which* $h(x_0) \in (0, \frac{\phi(x)}{p})$, *then* h *cannot be the solution to the optimization problem.*

Corollary 2. *Suppose the above optimization problem has a solution* $h_0(x)$, *then* $h_0(x)$ *must be either* 0 *or* $\frac{\phi(x)}{p}$ *at its continuity points. Furthermore, if* $h_0(x)$ *is piecewise continuous, it must have the form* $\frac{\phi(x)}{p}1_{\{x>a\}}$.

Theorem 1. *In the subclass of* Λ *consisting of piecewise continuous functions, the optimization problem has a solution*

$$h(x) = \frac{\phi(x)}{p}1_{\{x>a\}},$$

where $a = -\Phi^{-1}(p)$. *In this case the maximum value of the correlation is* $\frac{1}{\sqrt{2\pi p(1-p)}}e^{-[\Phi^{-1}(p)]^2/2}$.

Proof. It suffices to find a such that $h(x)$ thus defined is a probability density function. Indeed, we need

$$1 = \int_{\mathbb{R}} h(x)dx = \int_a^{\infty} \frac{\phi(x)}{p}dx = \frac{\Phi(-a)}{p}.$$

So $a = -\Phi^{-1}(p)$. In this case, we have

$$E[XY] = p\int_{\mathbb{R}} xh(x)dx = \int_a^{\infty} x\phi(x)dx = \frac{1}{\sqrt{2\pi}}e^{-\frac{a^2}{2}}.$$

So the maximum value of the correlation is

$$\rho_{\max}(X,Y) = \frac{1}{\sqrt{2\pi p(1-p)}}e^{-[\Phi^{-1}(p)]^2/2}.$$

□

Corollary 3. *The correlation of X and Y satisfies the following inequality*

$$-\frac{1}{\sqrt{2\pi p(1-p)}}e^{-[\Phi^{-1}(p)]^2/2} \le \rho(X,Y) \le \frac{1}{\sqrt{2\pi p(1-p)}}e^{-[\Phi^{-1}(p)]^2/2}.$$

Furthermore, the upper and lower bounds are tight.

Proof. It suffices to notice $Y_1 = 1 - Y$ itself is a binomial random variable with $P(Y_1 = 1) = 1 - p$ and $P(Y_1 = 0) = p$. Therefore

$$\min E[XY] = -\max\{-E[XY]\} = -\max E[XY_1] = \frac{-1}{\sqrt{2\pi}}e^{-[\Phi^{-1}(1-p)]^2/2} = \frac{-1}{\sqrt{2\pi}}e^{-[\Phi^{-1}(p)]^2/2}.$$

$\square$

4 Generalization

The above proof also works for a general random variable X with density function and a binomial random variable Y. Assuming X has pdf $f(x)$, cdf $F(x)$, and variance 1, we have the following result:

Theorem 2. *In the subclass of Λ consisting of piecewise continuous functions, the optimization problem has a solution*

$$h(x) = \frac{f(x)}{p}1_{\{x>a\}},$$

where $a = F^{-1}(1-p)$. In this case the maximum value of the correlation is $\frac{\int_a^\infty xf(x)dx}{\sqrt{p(1-p)}}$.

Corollary 4. *The correlation of X and Y satisfies the following inequality*

$$-\frac{\int_{F^{-1}(p)}^\infty xf(x)dx}{\sqrt{p(1-p)}} \le \rho(X,Y) \le \frac{\int_{F^{-1}(1-p)}^\infty xf(x)dx}{\sqrt{p(1-p)}}.$$

Furthermore, the upper and lower bounds are tight.

5 Numerical illustration

We use Matlab to plot the upper and lower bounds as functions of p for X being a standard normal random variable. Note Φ^{-1} approach to ∞ rather slowly, which could cause numerical instability using the original formula. Therefore, we use the change of variable ($x = \Phi^{-1}(p)$) for the plotting.

```
function plot_corr_bd

%plot_corr_bd   plots the tight bounds of the correlation between a
%               standard normal random variable and a binomial
%               random variable with paraemter p (i.e. probability of
%               being 1 is p, probability of being 0 is 1-p).
%
%               Reference
%               [1] Yan Zeng. Maximum value of the correlation between
%               a standard normal random variable and a binomial random
%               variable. October 22, 2009.
%
%               Yan Zeng, 10/20/2009.

x = -10:0.01:10; % x = nominv(p)
```

```
y = exp(-x.^2/2)./sqrt(2*pi*normcdf(x).*(1-normcdf(x)));
z = -y;

plot(x,y,'b', x,z,'r');
grid on;

end %plot_corr_bd
```

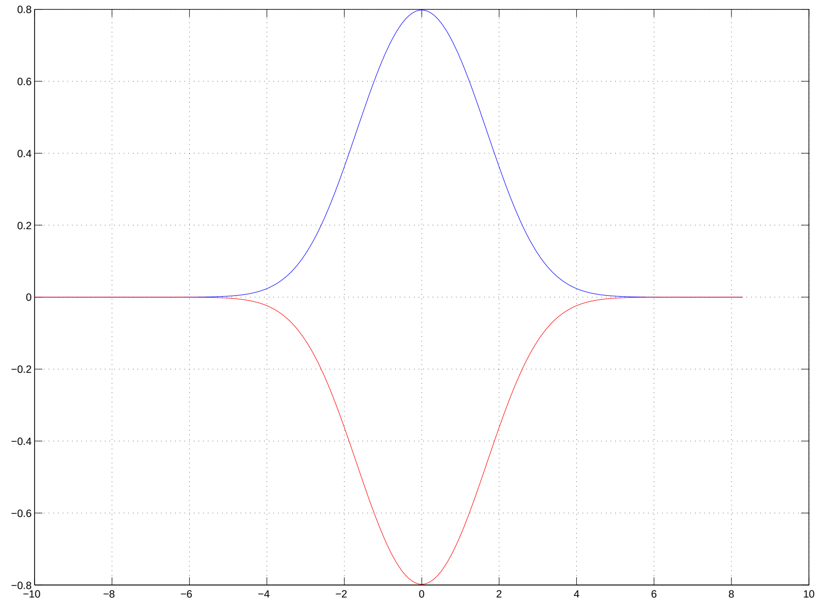

Figure 1: Bounds of the correlation between a standard normal random variable and a binomial random variable