CQF Lecture 5.7 Estimating Default Probability

Workings

Sources of PD information (risky bond)

1. In the CQF Lecture on Intensity Models, Exercises 3 considers the information provided by **the probability transition matrix**

$$\mathbf{P} = \begin{pmatrix} 1 - pdt & pdt \\ 0 & 1 \end{pmatrix}$$

The transition matrix represents a two-state Markov Chain for a simple rating model with 'default' and 'no default' states for one reference name.

The solution for a risky bond Z_I **with recovery rate assumption** is

$$\begin{aligned} Z_I &= (1-\theta)\, e^{-(r+p)(T-t)} + \theta\, e^{-r(T-t)} \\ & \quad e^{-p(T-t)} \equiv e^{-\lambda T} \qquad \text{over small time } pdt = \lambda dt \text{ and given } t = 0 \\ &\equiv (1-\theta)\, e^{-\lambda T}\, e^{-rT} + \theta\, e^{-rT} \qquad \Leftrightarrow \qquad \left[e^{-\lambda T} + (1 - e^{-\lambda T})\,\theta\right] e^{-rT} = Z_I \\ Z_I &= e^{-rT}\, \Big((1-\mathrm{RR}) e^{-\lambda T} + \mathrm{RR}\Big) \end{aligned}$$

- What is $e^{-\lambda T} \equiv e^{-\int_0^T \lambda_s ds}$? It is the probability of survival PrSurv.

  LGD $= (1-\theta)$ percentage of the notional is paid *only* in case of survival. For each unit of the notional, a risky bond pays out Recovery Rate plus the rest times PrSurv.
- If instead of default/no default we had rating transition probabilities from AAA to CCC to D states then, we would calculate *a mathematical expected value.*
- Example: a bond re-rated from BBB to A becomes 'investment grade' and should experience significant appreciation. Probability of this transition (PT) must be reflected in the fair price of a claim.

- The result on the *rhs* is useful in calculating the **credit spread** $y - r$ implied by the risky bond $Z_I = e^{-yT}$ with yield y.

$$\begin{aligned} e^{-yT} &= \left[e^{-\lambda T} + (1 - e^{-\lambda T})\,\theta\right] e^{-rT} \\ y - r &= -\ln\left[e^{-\lambda T} + (1 - e^{-\lambda T})\,\theta\right] / T. \end{aligned}$$

  If Recovery Rate $\theta = 0$ then the credit spread $y - r = \lambda$ for all tenors T.

The document prepared by CQF Faculty, Dr Richard Diamond at r.diamond@cqf.com

2. In credit derivatives markets, the payoff of **Default Put** is the price difference between the risk-free and risky bonds,

$$Z_0 - Z_I = e^{-rT} - e^{-(r+\lambda)T} = e^{-rT}\underbrace{\left(1 - e^{-\lambda T}\right)}_{}$$

There a very useful Taylor series expansion, often met in intensity models,

$$\begin{aligned} e^{-\lambda T} &\approx 1 - \lambda T + O(T^2) \\ 1 - e^{-\lambda T} &\approx 1 - (1 - \lambda T) = \lambda T \qquad \text{if } t = 0. \end{aligned}$$

making PrSurv $\approx 1 - \lambda T$ and PD $\approx \lambda T$ over relatively small timescale (expressed $\lambda\tau$).

Let's define 'a risky put' P_I as one underwritten by a risky counterparty on itself. Risky put is priced using a risky rate $r + p$, while an exchange-traded risk-neutral valued put is discounted with r. Exposure to a risky put is **the wrong-way risk**.

$$\begin{aligned} P_0 &= C_0 + Z_0 E - S \qquad \text{put-call parity} \\ &= C_I + Z_0 E - S \qquad \text{risk-free and risky calls have the same values } C_0 = C_I \\ &= \underbrace{P_I + S - Z_I E}_{} + Z_0 E - S \\ &= P_I + E(Z_0 - Z_I) \\ &= P_I + e^{-rT}(1 - e^{-\lambda T})E \end{aligned}$$

A risky put P_I written by the issuer on itself is **a hypothetical.** But results help us to see that **exchange-traded put price P_0 depends on the probability of default.**

3. The challenge to the Structural Model is about making implied volatility (of the firm's value, σ_V) to reflect the jump-to-default.

We know that the the firm's value equals to equity plus debt (accounting) $V_t = E_t + D_t$. On a per share basis, we can relate the stock price S to the firm's asset value V using recovery rate R

$$V \approx S + RD$$

Ignoring time value of the option written on the firm,

$$\sigma_V \approx \frac{\delta V}{V} \approx \frac{\delta S}{S + RD} \approx \frac{\delta S}{S}\frac{S}{S + RD} \approx \sigma_E \frac{S}{S + RD}$$

Assuming the volatility of the firm's value σ_V is a fixed quantity,

- if stock price S goes down we have to have stock volatility σ_E increasing. OTM puts become expensive.

The leverage effect is observed empirically and supports the Merton Model's explanation of volatility skew that adds a premium to the risk-free rate, $r \to (r + p)$.

Exponential family of distributions

Most familiar distributions belong to the Exponential Family: Normal, Chi-squared, Binomial, Poisson, Gamma, Beta... but, **not** Student's t. It is always possible to express any **PDF** of a distribution from the Exponential Family in *canonical form* using the key parameter θ.

$$f(y;\theta) = e^{a(y)b(\theta)+c(\theta)+d(y)}$$

An interesting fact: the variance for any of the Exponential family's distribution can be expressed using $V(\theta)$, a function of $\theta = \mathbb{E}[y]$, which is known in a general form

$$\mathbb{V}ar[y] = V(\theta)\phi$$

Because of the known general expression for $\mathbb{V}ar[y]$, we do not need to know the actual distribution of y_i to conduct **Quasi-MLE** (maximisation of quasi-likelihood). To put another way, the similarity of canonical representations for all Exponential PDFs makes it so that even if we choose a wrong distribution for y_i we are still likely to obtain robust $\boldsymbol{\beta'}$ for $y_i = \boldsymbol{X_i \beta'}$!

1. Poisson Distribution is the most easy to represent canonically. Its single intensity parameter λ is substituted by θ. The distribution for a Poisson random variable is

$$f(y;\theta) = \frac{\theta^y e^{-\theta}}{y!} = \exp\left(y\log\theta - \theta - \log y!\right)$$

$$\begin{aligned} a(y) &= y \\ b(\theta) &= \log\theta \quad \text{is the choice for the link function} \\ c(\theta) &= -\theta \\ d(y) &= -\log y \end{aligned}$$

2. Binomial Distribution also has parameter $\theta \equiv p$ buts the workings are nuanced

$$f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

application of log first and exponent second give the canonic form as

$$f(y;\theta) = \exp\left[y\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta) + \log\binom{n}{y}\right]$$

$$\begin{aligned} a(y) &= y \\ b(\theta) &= \log\left(\frac{\theta}{1-\theta}\right) \quad \text{the choice for the link function} \\ c(\theta) &= n\log(1-\theta) \\ d(y) &= \log\binom{n}{y} \end{aligned}$$

3. Normal Distribution has parameter θ representing location μ. But notice how mean μ and variance σ^2 go together in terms that are function of μ

$$
\begin{aligned}
f(y;\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\
&= \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^2 - \underbrace{2y\mu}_{} + \mu^2)\right\} \\
&= \exp\left\{\underbrace{\frac{y\mu}{\sigma^2}}_{} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}
\end{aligned}
$$

$$
\begin{aligned}
a(y) &= y \\
b(\mu) &= \frac{\mu}{\sigma^2} \quad \text{the link function is } g(\mu) = \mu \\
c(\mu) &= -\frac{\mu^2}{2\sigma^2} \\
d(y) &= -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)
\end{aligned}
$$

MLE Methodology: a reminder

1. When the assumption of Normality of residuals holds: ϵ_t is *iid* $N(0,\sigma^2)$, **the linear regression** $y_t = \hat{\boldsymbol{\beta}}\boldsymbol{x_t} + \epsilon_t$ has MLE properties. That means estimated coefficients $\hat{\boldsymbol{\beta}}$ are

   (a) consistent (i.e., close to unknown true estimates β with low tolerance) and

   (b) asymptotically efficient (i.e., their variance is known and minimised).

   Estimates $\hat{\boldsymbol{\beta}}$ in fact, maximise the following joint likelihood:

$$
\mathbf{L} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^T \exp -\frac{1}{2}\left[\frac{\epsilon_1^2}{\sigma^2} + \frac{\epsilon_2^2}{\sigma^2} + \cdots + \frac{\epsilon_T^2}{\sigma^2}\right]
$$

   Substituting $\epsilon_t = y_t - \hat{\boldsymbol{\beta}}\boldsymbol{x_t}$ and taking log gives for *an individual observation*,

$$
\log L_t = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\frac{(y_t - \hat{\boldsymbol{\beta}}\boldsymbol{x_t})^2}{\sigma^2}
$$

   The log-likelihood for a regression model is the sum of contributions from each observation $\log \mathbf{L} = \sum_{t=1}^{T} L_t$. Numerical MLE (eg, by Excel Solver) varies $\hat{\boldsymbol{\beta}}$ to maximise $\log \mathbf{L}$.

   The log-likelihood is maximised by *minimising the sum of squared errors* $\epsilon_t^2 = (y_t - \hat{\boldsymbol{\beta}}\boldsymbol{x_t})^2$.

   Exercise: consider a case with the intercept and one variable $\sum_{t=1}^{T}(y_t - (\beta_0 + \beta_1 x_t))^2 = L^*$. Apply the first-order optimisation condition, i.e., find partial derivatives and set

$$
\frac{\partial L^*}{\partial \beta_0} = 0 \quad \text{and} \quad \frac{\partial L^*}{\partial \beta_1} = 0
$$

   Obtain a system of two 'Normal equations' and confirm the well-known analytical solution result for regression coefficients $\beta_1 = \frac{\sum(x_t-\bar{x})(y_t-\bar{y})}{\sum(x_t-\bar{x})^2}$ and $\beta_0 = \bar{y} - \beta_1\bar{x}$.

2. **MLE principles.** For a known distribution, we can estimate the parameter that maximises **the probability mass**. It usually turns out to be a mean of distribution θ for the Exponential family. The key principles of an analytical MLE procedure:

1. We treat observations $\boldsymbol{Y} = [y_1, y_2, \dots, y_N]'$ as independently and identically distributed.

Their joint density (PDF) is simply the probability of observing *events together* – by multiplication rule of independent probabilities.

$$\begin{aligned} f(y_1, y_2, \dots, y_N; \theta) &= f(y_1; \theta) \times f(y_2; \theta) \times \dots \times f(y_N; \theta) \\ &= \prod_{i=1}^{N_{obs}} f(y_i; \theta) \end{aligned}$$

2. The log of the product becomes the sum of log-likelihood contributions:

$$\ell_{Y_1,\dots,Y_N}(\theta) = \log \prod_{i=1}^{N_{obs}} f(y_i; \theta) = \sum_{i=1}^{N_{obs}} \log f(y_i; \theta)$$

Let's find an analytical solution to the maximum likelihood estimation problem for a mix of independent but **identically distributed** Bernoulli variables.
The joint probability density of many default/no default events $y_i = 0, 1$ is

$$\begin{aligned} f(y_1, y_2, \dots, y_n; \theta) &= \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \\ &= \theta^{n\bar{y}} (1-\theta)^{n(1-\bar{y})} \\ &= \left[\theta^{\bar{y}} (1-\theta)^{(1-\bar{y})}\right]^n \end{aligned}$$

The log-likelihood function for the identically distributed Bernoulli variables (N draws) is

$$\begin{aligned} \ell_{Y_1,\dots,Y_N}(\theta) &= \log \left[\theta^{\bar{y}} (1-\theta)^{(1-\bar{y})}\right]^n \\ &= n \left[\bar{y} \log \theta + (1-\bar{y}) \log(1-\theta)\right] \end{aligned}$$

In general, the log-likelihood function for any of the Exponential family distributions follows a convenient linear form $\ell(\theta) = a(\bar{y})b(\theta) + c(\theta) + d(\bar{y})$.

Analytical solution to optimisation is found by equating the first derivative *wrt* θ to zero,

$$\frac{\partial}{\partial \theta} \ell_{Y_1,\dots,Y_N}(\theta) = n \left[\frac{\bar{y}}{\theta} - \frac{1-\bar{y}}{1-\theta}\right]$$

Setting the result equal to zero in order to find $\hat{\theta}$ that gives argmax $\ell_{Y_1,\dots,Y_N}(\theta)$

$$n \left[\frac{\bar{y}}{\hat{\theta}} - \frac{1-\bar{y}}{1-\hat{\theta}}\right] = 0$$

$$\bar{y}(1-\hat{\theta}) = (1-\bar{y})\hat{\theta} \quad \Rightarrow \quad \hat{\theta} = \bar{y}$$

This is consistent with what we know: average probability of default $\hat{\theta} = \bar{y}$ is a parameter that maximises likelihood. We fitted all observations to one parameter, namely 'location'. For indicator variable $y_i = 0, 1$ *the population probability of default* is PD $= \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]$.

3. The terminology of **Information Matrix** might be novel but let's consider results for multiple observations that are independent but follow same-located Normal density $f(y_i; \mu)$.

   The joint likelihood over the PDFs for independent observations becomes

$$\begin{aligned} \mathbf{L} &= f(y_1;\mu) \times f(y_2;\mu) \times \cdots \times f(y_T;\mu) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^T \exp -\frac{1}{2}\left[\frac{(y_1-\mu)^2}{\sigma^2} + \frac{(y_2-\mu)^2}{\sigma^2} + \cdots + \frac{(y_T-\mu)^2}{\sigma^2}\right] \end{aligned}$$

   The log-likelihood of the joint density simplifies to

$$\log \mathbf{L} = -\frac{T}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left[(y_1-\mu)^2 + (y_2-\mu)^2 + \cdots + (y_T-\mu)^2\right]$$

$$\frac{\partial \log \mathbf{L}}{\partial \mu} = \frac{1}{\sigma^2}\left[(y_1-\mu) + (y_2-\mu) + \cdots + (y_T-\mu)\right]$$

   (a) Equating $\frac{\partial \log \mathbf{L}}{\partial \mu} = 0$ gives the usual formula for arithmetic average – i.e., for the Normal density mean is the MLE estimate.

   (b) Let's keep differentiating *wrt* μ,

$$\frac{\partial^2 \log \mathbf{L}}{\partial \mu \partial \mu} = -\frac{T}{\sigma^2}$$

   Information matrix for any Exponential family distribution is defined as **Hessian**, in the canonical form μ is equal to distribution parameter θ

$$\mathbf{I} = -\mathbb{E}\left[\frac{\partial^2 \log \mathbf{L}}{\partial \theta \partial \theta}\right] = \frac{T}{\sigma^2}$$

   The inverse of information matrix $\mathbf{I}^{-1} = \sigma^2/T$ is an easily recognised as squared standard error $\sigma/\sqrt{T}$. If $T \to \infty$ there is no estimation error!

4. For logistic regression, the Information Matrix has been solved as

$$\mathbf{I} = -\mathbb{E}\left[\frac{\partial^2 \log \mathbf{L}}{\partial \beta_j \partial \beta_k}\right] = \mathbf{XP'X}$$

   where $\log \mathbf{L}$ is log-likelihood of Binomial density and $\mathbf{P}$ is a diagonal matrix of $p_i(1-p_i)$.

$$\mathbf{I}_{j,k} = \sum_{i=1}^{N_{obs}} p_{i,j}(1-p_{i,j}) x_{i,j} x_{i,k}$$

   This element-wise calculation of $\mathbf{I}$ with $\mathbf{P}$ as a vector reveals 'condensing' of Binomial variance.

(a) This numerical result for computation of Information Matrix has been obtained by differentiating Equation (1) as in the lecture, where

$$\frac{\partial \log \mathbf{L}}{\partial \beta_j} = \sum_{i=1}^{N_{obs}} (y_i - \hat{p}_i) x_{i,j}$$

(b) MLE estimates are found by equating a system of such equations to zero and finding such $\hat{\boldsymbol{\beta}}$ that satisfy them under the link function $p_i = \Lambda(X_i' \hat{\boldsymbol{\beta}})$. The root-finding is done numerically by a scheme such as the Newton-Raphson.

After the estimates $\hat{\boldsymbol{\beta}}$ are obtained, the asymptotic covariance matrix $\mathbf{I}^{-1}$ is calculated. Its diagonal provides squared standard errors used in significance testing.

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \mathbf{I}^{-1}\right).$$

5. The MLE framework allows for model selection by hypothesis testing. That usually involve estimating log-likelihoods for two models and comparing them.

   There are three ways to compare the log-likelihoods:

   - Wald test is alike testing difference between two means $(\theta_1 - \theta_2)$ using F-distribution.
   - Likelihood Ratio test.
   - Lagrange Multiplier test checks if $\partial \log \mathbf{L} / \partial \theta = 0$.

   These tests are described as tests for differences among *nested models*, because our restricted model (with less parameters) is nested within the original model. The null hypothesis is that the restricted model is the 'true' model, i.e., the same data is just as likely if the null hypothesis (with fewer variables in the model) is true.

6. Likelihood Ratio test can be represented within more general approach of Entropy Pooling and minimisation. For two probability distributions, generic $\tilde{f}(y)$ and reference $f(y)$ the relative information **entropy** is defined as

   $$\varepsilon[\tilde{f}(y), f(y)] = \int_{\mathbb{R}} \tilde{f}(y) \left[\log \tilde{f}(y) - \log f(y)\right] dy$$

   Imposing restrictions on $\tilde{f}(y)$ makes it to deviate from $f(y)$.

   Entropy ε is a natural measure of how distorted any (modified) distribution $\tilde{f}(y)$ is with respect to the reference $f(y)$.

   Compare with the LR test

   $$D = 2(\log f(y; \widehat{\theta}) - \log f(y; \widehat{\theta}_0))$$

   Better fit minimises Deviance implying minimal entropy.