Quantitative Testing of Financial Risk Models

Sanghee Cho, Jerry Cline, Weiwei Shen, Michael Vallance & Jin Xia

A compendium of quantitative testing methods useful for quality assurance testing of financial risk models.

GE Global Research

Niskayuna, NY

12/15/2015

Contents

Model Risk Management

Model Risk Management

Michael Vallance, Weiwei Shen and Jerrold Cline

Table of Contents

Valuation Model Risk Management Overview

In this section we define *valuation model risk* and we look at the role of validation practices in the minimization of this risk. We define the sources of model risk categorically. We then look at how specific validation tactics are used to minimize risk, category by category. In subsequent sections, we look at the application of validation tactics to specific mathematical transformations.

We introduce a new metric, *model capability*, and we demonstrate that model capability is innately bound to the end use and the end user's tolerance for risk. We also introduce a similar metric, *model replicability*, which is used to quantitatively judge the outcome of a replication study. These metrics are used in subsequent chapters.

Manifestation of Valuation Model Risk

The purpose of financial valuation models is to estimate fair-market values for selected financial assets, for use in hedging, trading, and financial reporting.

Paraphrasing,[1] financial assets are from one of two categories. In the first category, asset prices are directly observable. These assets trade in organized markets. In the second category, prices cannot be directly observed, but need to be inferred from observable prices of related assets. This latter case is pertinent to the majority of GE's positions in debt instruments and financial derivatives; their values are related to various features of *primary* assets by use of *valuation* models. This process is called "mark-to-model," and involves mathematical models with some subjective components, exposing the process to estimation errors. Three types of estimation errors can be identified:

1. The model-derived value for the exotic asset is different than the value derived from the one "true" model, *assuming* the existence of such an indisputable *benchmark* model.
2. We observe hedging error, difference between the model-derived value of a derivative and the value of the corresponding replicating portfolio.
3. There is a difference between the mark-to-model value of the derivative and the mark-to-market price at which the same instrument is observed to trade in the marketplace, in cases where such a marketplace exists.

These differences, taken separately and together, constitute the *model risk* associated with the valuation model.

Origination of Model Risk

Derman[2] categorizes model risk as originating from seven sources:

1. Inapplicability of modeling.

[1] C. Martini and P. Henaff, Model Validation: theory, practice and perspectives, Zeliade White Paper (May, 2011).
[2] E. Derman, Model Risk, Goldman Sachs Quantitative Strategies Research Notes (April, 1996).

2. Incorrect model.
3. Correct model, incorrect mathematical representation (erroneous mathematical transformations).
4. Correct model, inappropriate use (poor calibration).
5. Badly approximated solution (inaccurate numerical methods, e.g., integration).
6. Software bugs.
7. Unstable data (non-stationary processes).[3]

The philosophy in reference 2 is echoed by the Basel Committee on Banking Supervision,[4] which states that validation includes evaluations of:

1. The model's theoretical soundness and mathematical integrity.
2. The appropriateness of model assumptions, including consistency with market practices.
3. Sensitivity analyses performed to assess the impact of variations in model parameters on fair value, including under stress conditions.
4. Accuracy in back-testing.

Limitations of Modeling (Derman category 1)

Financial model accuracy, in comparison to models in the physical sciences, is highly vulnerable to unpredictable excitations which cannot be factored (noise or nuisance variables). Practitioners must have a vision of how the underlying financial processes work, including interconnectivities. We should not attempt to model outcomes which depend largely on psychology, gamesmanship, and geopolitical events. In summary, the first defense against Derman category 1 risk (inapplicability of modeling) is a good understanding of the financial processes underlying the model.

In the associated documentation, the model designer must state all assumptions that are implicit in the model. The designer should provide literature references or data to justify such assumptions.
The designer should offer guidance as to the limits of model applicability. If the model's applicability is geographically limited, limited to a sub-class of financial instruments, or not valid during certain time periods, these boundaries must be described.

Model Accuracy by Market and Benchmark Testing (Derman category 2)

Derman risk category 2 addresses the usefulness of the model, when all is said and done. The validator must understand the theory underlying the model, and he must be aware of current industry practices and trends. If the model theory violates widely held paradigms without sound reason, then this should be treated as a material finding.

Given that the model is theoretically sound and skillfully executed, there is still risk that the model does not provide useful valuations. This risk must be addressed through out-of-sample testing; *i.e.*, validating the model's pricing accuracy versus market quotes that were not used to calibrate the model or mark-

[3] Derman does not identify poor quality data as a model risk, but could have.

[4] Supervisory guidance for assessing banks' financial instrument fair value practices, Basel Committee on Banking Supervision Consultative Document, Bank for International Settlements (BIS) (November, 2008)

to-model valuations from *benchmark*[5] models. Additionally, for derivative assets, we may compare our valuations to the values of the replicating portfolios that underlay the derivatives.

An example is the approach used by the GE Capital subsidiary GE Financial Markets.[6] Derivative valuation model output is tested monthly against benchmark model output for trades entered by GE. GE tests a subset of all trades, using a sampling process designed to select a representative sample with respect to asset type, currency, tenor, and reference indices.

Thresholds have been set on model accuracy. For illustration, these thresholds are shown in **Error! Reference source not found.**. Our internal mark-to-model valuations are compared with those of the benchmark models: the Swap Manager tool and Bloomberg Credit Default Swap Pricer tool, both available with the Bloomberg Terminal®.[7] When the absolute difference between the two valuations is larger than the threshold values shown in the table, the analyst is required to investigate the reason for the discrepancy and report the finding to the business's Valuation Committee for resolution.

Table 1: These threshold values (tolerances) are used when comparing GE's mark-to-model valuations versus those of benchmark models.[6]

Trade Type	Risk Threshold
IR Swaps	3 DV01
Cross Currency Swaps	0.25 FX01
Callable Swaps	3 Vega
Credit Default Swaps	10 Spread01
IR Caps/Floors	3 Vega

Additionally, Financial Markets uses reports sourced from TriOptima©[8], which compare GE and counterparty mark-to-model valuations for financial instruments. Again, the analyst investigates all cases where the thresholds of **Error! Reference source not found.** are violated, and reports the finding to the Valuation Committee for resolution. In the examples given, the out-of-sample testing is in the form of continuing monitoring, and is not part of the formal model validation process.

Table 2: These thresholds are used when comparing GE's mark-to-model valuations versus the mark-to-model valuations of the counterparties.6

Trade Type	Materiality Threshold	Risk Threshold
IR Swaps	$500,000	5 DV01
Cross Currency Swaps		1 FX01

[5] As used herein, *benchmark model* refers to a model that is an industry standard, an approximation of the "true" model, as described in sub-section 0.
[6] Valuations Procedure – Independent Price Verification, GE Financial Markets (June 2013).
[7] Bloomberg Terminal is a product of Bloomberg Finance, L.P. http://www.bloomberg.com/professional/products-solutions/
[8] TriOptima AB, an ICAP Group company with offices in NYC, NY. http://www.trioptima.com/services.html.

Callable Swaps		3 Vega
Credit Default Swaps	$100,000	10 Spread01
IR Caps/Floors		3 Vega

Even the best models will not predict market pricing exactly. Market pricing is influenced by factors other than those used in the models, such as current events. As an approximation, we can group these "nuisance" factors as a random variable superimposed on the primary pricing predictors. Manufacturing processes, like models, are also impacted by nuisance factors, including operator-to-operator variation, raw-material variation, and changes in ambient conditions. The produced articles and materials exhibit attributes that vary somewhat randomly from the target values. In manufacturing science, the accuracy of a process is typically expressed as *process capability*. Process capability relates to the predicted fraction of process output that is produced within the tolerances placed on the process, based on sampling statistics. Tolerances are set at levels that ensure that the process output will be useful for the downstream process or that it will satisfy consumer preferences. In financial modeling, the application of the model is called the *use case*. In the following mathematics, we assume that the use case is known, and that the accuracy requirements of the use case are known.

The worthiness of a financial model to simulate the market or a benchmark model, with adequate accuracy to satisfy the use case, can be referred to as the *model capability*. Suppose the existence of a set of *N* assets, all from the use-case asset class. These assets have a property of interest with values:

$V_i,\ i = 1,\ldots,N$

These values may be prices, or they can be another characteristic of the assets. The values are assigned by the market, or alternatively, by a benchmark process (model). This set of *N* assets is chosen to be representative of the population of assets in the use-case class.

A model of interest predicts these values. The model was constructed without reference to these *N* assets. The model predictions are:

$P_i,\ i = 1,\ldots,N$

Based on the use case, useful predictions are those that obey:

$$V_k \cdot (1 - F_L) \le P_k \le V_k \cdot (1 + F_U) \quad \text{where } k \in 1,\ldots,N \text{ [9]} \qquad \textbf{1}$$

In the above equation, V_k is assumed positive. This is non-limiting, since transformation can be employed. F_L and F_U are positive values, which are selected to fulfill business objectives. Often they are equal. Define the random variable X_i and associated statistics.

[9] We have assumed that the tolerance range, scales with value. Where, the tolerance range is absolute, independent of value, this equation and the following definitions need to be modified.

$$X_i \equiv \frac{P_i - V_i}{V_i}, i = 1,\ldots, N$$
$$\bar{X} \equiv \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \textbf{2}$$
$$S_X^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$$

We assume that X_i is normally distributed.[10] Calculate:

$$C_{ML} \equiv \frac{F_L + \bar{X}}{3S_X}$$
$$C_{MU} \equiv \frac{F_U - \bar{X}}{3S_X} \qquad \textit{2}$$
$$C_{MK} \equiv \min\left(C_{ML}, C_{MU}\right)$$

Refer to C_{MK} as *model capability*. C_{MK} is analogous to process capability C_{PK} in manufacturing science.[11,12]

Table 3: For selected values of model capability, the rate of defective model predictions is shown.

Model Capability	Defective Prediction Rate
>2	$<2 \times 10^{-9}$
>1.333	$<63 \times 10^{-6}$
>0.667	$<45.5 \times 10^{-3}$
<0	Mean error outside of tolerance limits

$C_{PK} \geq 2$ represents outstanding capability. This level of accuracy is sometimes referred to as "six nines" or "six sigma." $C_{PK} > 0.667$ is not uncommon. The model capability concept is shown graphically in Figure 1.

We cannot calculate model capability without tolerance values F_L and F_U. These tolerance values must be derived from the use case for the model. If the enterprise has no expectation or requirement for accuracy (ability to track market and/or benchmark valuations), the model capability is not, and should not, be defined. Judging by Table 1 and Table 2, GE Financial Markets does have an expectation of accuracy, and the stated thresholds can be converted to tolerances.

[10] These thresholds for C_{MK} are reasonable for normally distributed X, and will suffice for most symmetrical distributions.
[11] *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/ (July, 2015).
[12] D. Montgomery, *Introduction to Statistical Quality Contro*l, Wiley (2004).

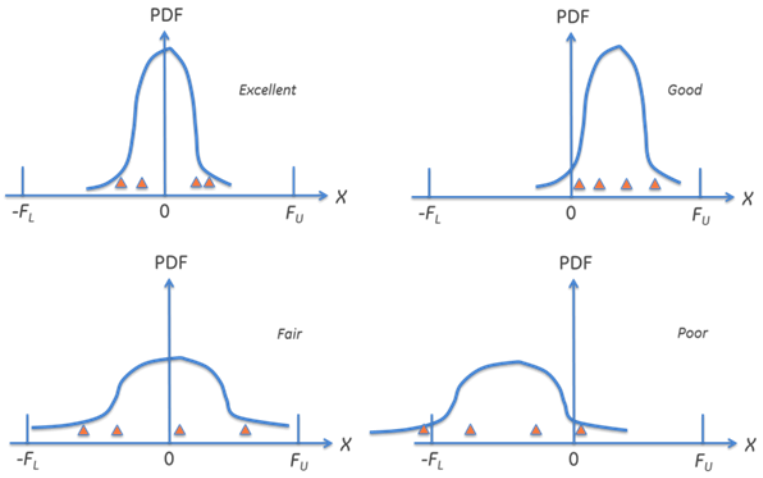

Figure 1: Four instances of process capability are shown graphically, where the triangles represent the *X* sample and the peaked curves represent the probability density functions (PDFs) associated with *X*. In the upper left graphic, the X_i values are clustered tightly around zero. In the upper right graphic, the X_i values are tightly clustered, but the mean response is offset from zero, well still within the tolerance interval. In the lower left graphic, the responses are centered, but the distribution is broad. In the lower right graphic, the mean response is offset and the distribution is broad.

We have used the expression "representative sample." In reference 6, the model values of sample trades are compared to benchmark models each month. Rather than random sampling, the sampling is done with stratified samples, where the population has been partitioned by financial instrument, by currency, by tenor, and by which reference indices were involved. A similar sampling strategy is used to compare mark-to-model pricing of trades involving GE and counterparties, wherein a third party is responsible for the analysis. In that case, sampling is again done with stratified samples, where the population is portioned by financial instrument, by counterparty, by trade size, and which reference indices were involved. A more sophisticated sampling methodology uses clustering analysis.[13]

Obviously, the larger the sample used to calculate C_{MK}, the better. The 95% confidence intervals on C_{MK} are aprroximately:[14]

$$C_{MK} \pm 0.653\sqrt{\frac{1}{n}+\frac{9C_{MK}^2}{2(n-1)}} \qquad \textbf{3}$$

A sample size of about 200 would give a 95% confidence interval of about ±10% on C_{MK}.

Replication (Derman categories 3 - 6)

Derman risk categories 3, 4, 5 and 6 (correct model, incorrect mathematical formalism; correct model, incorrect calibration; numerical methods error; software bugs) are best managed by model replication.

[13] B.S. Everitt, S. Landau, M. Leese and D. Stahl, *Cluster Analysis*, 5th Ed., Wiley (2011).

[14] A.F. Bissell, How Reliable is Your Capability Index? *Appl. Statist.* **39(3)** (1990) 331-340.

Perfect replication should result in "exact" replication of model predictions. Where the source model and the replicate model generate different values, the validator should exert due effort to identify the source of variation; the level of effort must be consistent with the available time and resources, and with the criticality of the model predictions in the business context. Once the source of variation is understood, we should consider the materiality of the difference, in consultation with the end-user of the model output. Valuation differences of up to 10% may be tolerable,[15] but this decision should not be taken by the validator alone, because the risk tolerance of the enterprise backstops this decision. If the valuation variation between the focus model and that of the replicate model is intolerable, then the observation should be treated as a material finding. Continued use of a model with a defect of any sort is a risk for the enterprise; although the defect may result in only minor distortions for the enterprise's portfolio at present time, the portfolio is dynamic, and the future impact is unbounded.

Replication can and should include sensitivity analysis. Synthetic, small perturbations of the various model inputs should result in proportional perturbations of model output for the source model and the replica. In addition, based on our knowledge of the underlying financial processes, the ratios of input and output perturbations should be reasonable in sign and magnitude. Sensitivity studies with the focus model are critical, when the values of input parameters are uncertain, dynamic, or stochastic. This is especially important when the values of input parameters cannot be measured directly in real time.

Replication is problematic in the case of "black-box," vendor-supplied models. The use of such models constitutes a risk for the enterprise, because we can never fully attest to the model's mathematical integrity and appropriateness of assumptions, as demanded by the Basel Committee. Nonetheless, validators should attempt replication. In doing so, the validator should exert due effort to reverse engineer the assumptions/choices made by the vendor; the level of effort must be consistent with the available time and resources, and with the criticality of the model predictions in the business context. As part of a recent validation for a GE Financial Markets model,[16] the validator reverse engineered a vendor-supplied model supplied by Wall Street Systems.

Replication may not be exact. This is the case when the model, in whole or part, is supplied by the vendor. In other cases, the cost of replication can be greatly reduced by using vendor-supplied software in replication, which may use modified numerical methods, relative to the subject model. We will refer to approximate replication as *casual replication*. In the case of casual replication, we suppose that the replication model can be treated as the benchmark model, and the calculations shown above in equations 1, 2, and 3 can be used to evaluate the goodness of the replication of the subject model. In this case, the figure of merit is *model replicability* R_{MK}. In this case, we still depend on the tolerances associated with the end use.

[15] K. Long, Test Acceptance Criteria (Illustration), supplied in an email to J. Cline, *et al.* (June 12, 2014).
[16] J. Lai, Swap & Debt Valuation Model (WSS), GETM-VAL-001, GMGV Model Validation Technical Report (March 2014).

$$R_{ML} \equiv \frac{F_L + \bar{X}}{3S_X}$$
$$R_{MU} \equiv \frac{F_U - \bar{X}}{3S_X} \qquad 4$$
$$R_{MK} \equiv \min\left(R_{ML}, R_{MU}\right)$$

The reader should note that model replicability is not a valid measure of model accuracy, whereas model capability is.

Data Integrity (Derman category 7)

All financial models need to be *calibrated* with market data—there is no example of an *ab initio* financial risk model. Valuation distortions related to poor data integrity constitute Derman risk category 7. There are several risks to consider:

- Data is wrong.
- Data is not representative.
- Data is non-stationary.
- Data is missing.

Non-stationary data is only relevant when back-looking, time-series analysis is extrapolated to predict future outcomes. There are many tests for stationarity. The fUnitRoots package, available for R programmers, provides several useful tests. See Pfaff's book[17] for further information.

Where sampling is required, we incur risk, because the sample may not be representative of the population. For example, when constructing an interest rate curve using a set of bond quotes or swap rates, there is a risk that the rate data used is non-representative. To identify the level of risk due to sampling distortions, we can re-sample the data, once or multiple times, and re-evaluate the assets to gage the resultant valuation variability. Much has been written about re-sampling strategies, and the "resample"[18] package is available for R users.

Wrong and missing data are very common issues. Much has been written about data cleansing, a major topic in analytics. Abken and co-workers[19] describe their approach to validation, when data are wrong and missing.

Summary

For models used to evaluate financial instruments, the potential for wrong predictions of fair market value, and/or wrong predictions of the related derivatives of value, constitutes model risk. The risk originates from one of several sources: inapplicability of modeling, wrong choice of model, poor model

[17] B. Pfaff, *Analysis of Integrated and Cointegrated Time Series with R*, Springer (November, 2005).
[18] Package "resample", http://cran.r-project.org/web/packages/resample/resample.pdf.
[19] P. Abken, I. Shi, R. Ruan and K. Zhang, Validation of GEPD_SURE Aviation v3, GE Model Validation Document (May, 2015).

mechanics and wrong calibration data. Given that the model is theoretically sound, the remaining risks can be addressed with:

1. Accuracy testing, using liquid instruments and/or benchmark models;
2. Sensitivity testing, studying the relative perturbations of inputs and outputs across the range; and
3. Replication

We have proposed two mathematical procedures, model capability and model replicability, to judge the goodness of the model under validation. These mathematical procedures depend on the availability of model performance tolerances, which should be directly traceable to the use case.

Valuation of Interest-Rate-Based Financial Instruments

Valuation of Interest-Rate-Based Financial Instruments

Michael Vallance, Jerry Cline, and Weiwei Shen

Table of Contents

Valuation of Financial Instruments

GE has positions in an array of financial instruments, which are used for lending, leasing, funding, hedging, and insurance. The interest-rate-sensitive financial assets include derivatives:[20,21]

Interest-rate swaps
Credit-default swaps
Cross-currency, interest-rate swaps
Callable interest-rate swaps
Interest-rate caps and floors
Foreign-exchange forward contracts
As well as debt instruments:

Fixed-rate bonds
Floating-rate notes
Flip-flop notes
Hybrid securities
Callable and puttable bonds
Survivor's option bonds
Preferred stock
Loans

In the case of derivative positions in hedging activities, GE adheres to FAS 133.[22] From that standard: "Fair value is the most relevant measure for financial instruments and the only relevant measure for derivative instruments." The FASB defines the fair value of a financial instrument as the amount at which the instrument could be exchanged in a current transaction between willing parties.

For uses other than hedging, GE adheres to FAS 107.[23] The standard also requires reporting of fair value. From the standard:

> "Quoted market prices, if available, are the best evidence of the fair value of financial instruments. If quoted market prices are not available, management's best estimate of fair value may be based on the quoted market price of a financial instrument with similar characteristics or on valuation techniques (for example, the present value of estimated future cash flows using a discount rate commensurate with the risks involved, option pricing models, or matrix pricing models)."

GE largely holds positions in financial instruments with no quoted prices. GE often uses financial models to value these positions. The FASB prefers to issue general, rather than specific guidance on model-

[20] GE Financial Markets Valuations Procedure – Independent Price Verification (March, 2012).
[21] J. Lai, Swap & Debt Valuation Model (WSS) GETM-VAL-001, GMGV Model Validation Technical Report (March, 2014).
[22] Accounting for Derivative Instruments and Hedging Activities, Statement of Financial Accounting Standards No. 133, Financial Accounting Standards Board (June 1998).
[23] Disclosures about Fair Value of Financial Instruments, Statement of Financial Accounting Standards No. 107, Financial Accounting Standards Board (December 1991).

based valuation, even though general guidance may result in disclosures that are less comparable from entity to entity. The board concluded that the benefits to investors and creditors of having some timely information about fair value outweigh the disadvantage of that information being less than fully comparable.

In light of the FASB guidance, we can minimally expect our models to adhere to the following:

The calculus should be transparent and reproducible.

The model output should be repeatable and stable. *I.e.*, if the input is stationary, then the output is stationary.

The theoretical underpinnings of the model should not violate widely held paradigms without valid rational.

The model's predictions should be testable, and the model should be tested periodically and across the spectrum of positions that it is meant to value. Valid testing may involve producing predictions for similar assets that are quoted on exchange markets, dealer markets, or brokered markets. Another valid testing strategy involves comparison to widely accepted benchmark models.

In the following sections, we will refer to the model under scrutiny as the *study model*. Widely accepted models (referred to as the one "true" model in Chapter 1) are called *benchmark model*s. Validator attempts to duplicate the study model are called *replicate model*s. Validator attempts to build alternative, better-performing models are called *challenger model*s.

Validation of Interest Rate Curve Calculations

Calculating zero-coupon interest rates as a function of maturity (zero-curve construction) underlies every financial instrument valuation based on discounted cash flows. The most popular approach is the bootstrap method.[24] Zero curve construction is integrated into valuation software packages such as Wallstreet Suite.[25] A principal validation practice for zero-curve construction is replication. Other tests probing theoretical soundness and stability are applicable, as described below.

Replication

Given the exact set of market instrument quotes used to construct a zero rate curve, as well as the interpolation and extrapolation rules used, curve replication should exactly match the interest rates of the production model at each and every maturity. Where the construction details are not fully disclosed, and this is not uncommon with licensed, proprietary, valuation platforms; some level of trial and error may be required to "reverse engineer" the method. By doing so, the validator can determine all assumptions and methodology involved in curve construction, and then opine on conceptual soundness, as was done in reference 21.

[24] J.C. Hull, *Fundamentals of Futures and Options Markets*, 8th Ed., Pearson (2014).

[25] Wallstreet Suite, Wallstreet Systems, http://www.wallstreetsystems.com/documents/Wallstreet-Suite-for-Corporate-Treasury.pdf.

Casual Replication

Casual replication, introduced in Chapter 1, is our name for approximate replication; casual replication offers a balance of low validation cost and low enterprise risk. Often, casual replication is the only recourse; this is the case when the study model comprises proprietary vendor software. In other cases, we may choose to use proprietary vendor software as the replication model. In our experience, vendor software is never fully transparent, so exact replication is difficult.

Selection of Tolerances for Model Replicability Calculation

In a proposal under study[26], for interest rate replication, <0.5 bps (bps = basis point = 0.01%) difference in replicated rates is "negligible," 0.5 – 5 bps difference is "moderate" disagreement, and >5 bps is a "significant" finding. While these criteria seem reasonable, a best practice is to gain agreement from the model users, because the model users best understand the use case and the associated tolerance for error (see Chapter 1).

We can examine the price impact of small rate discrepancies (Δy) on the resultant discrepancies in prices (ΔP) of fixed income instruments. Assuming that pricing of interest-rate derivatives and debt instruments is a function only of continuously compounded rate y, we can perform a Taylor series expansion around y:

$$\Delta P = \frac{dP}{dy}\Delta y + \frac{d^2P}{dy^2}\frac{\Delta y^2}{2!} + \frac{d^3P}{dy^3}\frac{\Delta y^3}{3!} + ... \qquad 5$$

Dividing both sides by P:

$$\frac{\Delta P}{P} = \frac{1}{P}\frac{dP}{dy}\Delta y + \frac{1}{P}\frac{d^2P}{dy^2}\frac{\Delta y^2}{2!} + \frac{1}{P}\frac{d^3P}{dy^3}\frac{\Delta y^3}{3!} + ...$$

or

$$\frac{\Delta P}{P} = -MD\Delta y + C\frac{\Delta y^2}{2!} + \frac{1}{P}\frac{d^3P}{dy^3}\frac{\Delta y^3}{3!} + ... \qquad 6$$

After defining:

$$MD \equiv -\frac{1}{P}\frac{dP}{dy},\ C \equiv \frac{1}{P}\frac{d^2P}{dy^2}$$

Where MD is called modified duration[27] and C is called convexity. Generally, the first term on the right side, has an outsize influence on price, and we may choose to truncate the series after this first term.

$$MD \equiv -\frac{1}{P}\frac{dP}{dy} \approx -\frac{1}{P}\frac{\Delta P}{\Delta y} \qquad 7$$

[26] K. Long, Test Acceptance Criteria (Illustration), supplied in an email to J. Cline, *et al.* (June 12, 2014).
[27] T. Coleman, A Guide to Duration, DV01, and Yield Curve Risk Transformations (January, 2011). Available at: http://ssrn.com/abstract=1733227.

Taking the approximation as exact:

$$\Delta y = -\frac{\triangle P}{P \times MD} \qquad 8$$

Equation $\Delta y = -\frac{\triangle P}{P \times MD}$ 8 is plotted in Figure 2 for five values of $\Delta P/P$: 0.5, 1, 2, 5 and 10%. For any duration, the curves define the maximum magnitude of yield replication error Δy that can be tolerated, while assuring that relative price replication error is within a given range. As demonstrated by Figure 2, errors in valuation (ΔP), propagated from zero-curve error, are more severe for longer duration, fixed-income instruments. Based on this analysis, *the tolerances on rate, as proposed in reference 26 appear conservative.*

Example to Demonstrate Model Replicability

An approach to replication acceptance criteria follows the model replicability approach introduced in Chapter 1. For example, in reference 26, ±1% relative deviation was proposed as a negligible price difference. If a representative sample of fixed-price financial instruments are priced using both zero curves, the study model curve and the replicate model curve, replicability R_{MK} can be calculated using equation 5 of Chapter 1, with $F_U = F_L = 0.01$. Better still, would be to use a tolerance directly traceable to the use case.

An illustrious example of replicability calculation is shown in Figure 3. We have used the upper and lower tolerances from the previous paragraph. The example supposes a sample of five, zero-coupon, fixed-income instruments with known modified durations *MD*. We further suppose that a replicated zero curve has been developed, and the zero-rate difference Δy between the model and the replicate has been calculated at the maturities of the five instruments. We note that X_i, as defined in equation 2 of Chapter 1, is given by:

$$X_i = \frac{\Delta P_i}{P_i} = MD_i \times \Delta y_i \qquad 9$$

The analysis shows excellent replicability between the study model and replicate model zero curves. In this example, the sample size $N = 5$ is too small for a reliable estimate of R_{MK} as discussed in Chapter 1 (see discussion associated with equation 4).

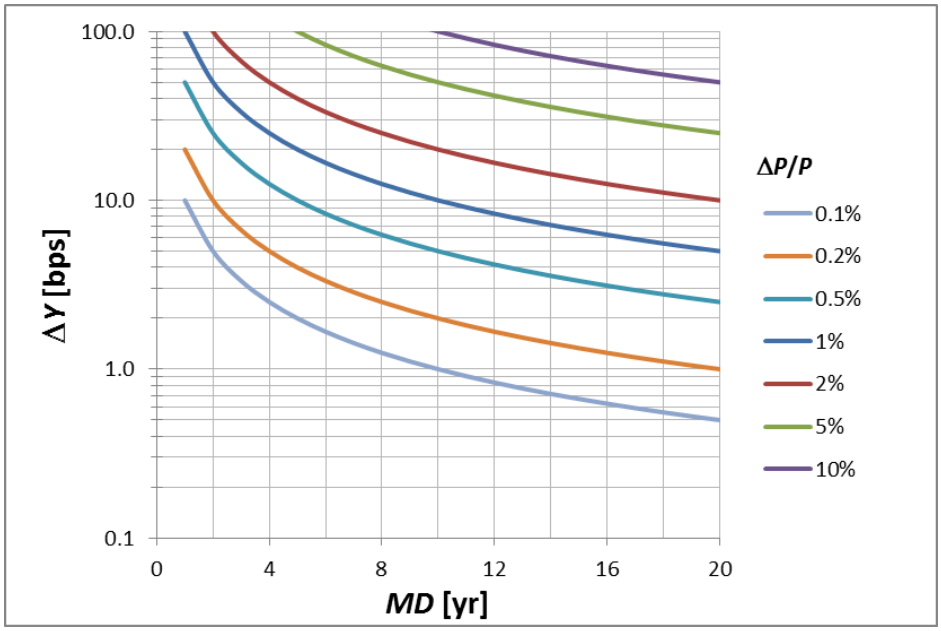

Figure 2: For a given limit on the relative absolute value of price deviation ΔP/P, the value on the logarithmic ordinate represents the maximum absolute value of yield deviation that can be tolerated.

Instrument	MD (yr)♣	ΔY (1/yr)♠	X_i				
1	3	-0.00005	0.00015				
2	5	-0.00001	0.00005				
3	7	0.00005	-0.00035				
4	9	0.00008	-0.00072				
5	11	0.00010	-0.00110				
		$\langle X \rangle$ =	-0.00039	F_L =	0.01	R_{ML} =	6.83
		S_X =	0.00047	F_U =	0.01	R_{MU} =	7.39
						R_{MK} =	6.83

♣ Assumed to be equal for the model and the replication

♠ Model yield minus replication yield

Figure 3: 5 fixed-income financial instruments, of known modified duration, are sampled from a population. The difference between the model interest rate and the replication interest rate is given in the third column. The relative price error is given in the fourth column. The resultant replicability R_{MK} is excellent.

Pricing Swaps to Zero

Zero curves are typically constructed by the bootstrap method, often using actual fixed-for-floating, interest-rate swap trades and/or quotes as calibration points. The resulting zero curve, and the forward rates derived from the zero curve, can be used to price the fixed and floating legs of these same swaps. A fixed-for-floating interest rate swap quote has the feature that the fixed and floating legs have equal

values, B_{fx} and B_{flt}, at the time of the trade or current quote. Pricing the swaps to zero is a check on the consistency of the bootstrapping process.

Initial Pricing of Fixed-for-Floating Swaps

$$B_{fx} = B_{flt}$$

$$B_{fx} = sNR\sum_{j=1}^{n_{fx}} \alpha_{j-1,j} Z_j$$

$$B_{flt} = N\sum_{i=1}^{n_{flt}} f_{i-1,i}\alpha_{i-1.i} Z_i$$

$Z_j\left(Z_i\right)$	the discount factor for time t_j
$\alpha_{j-1,j}\left(\alpha_{i-1.i}\right)$	accrual factor between dates j - 1 and j, based on the specified accrual method [yr]
$f_{i-1,i}$	(implied) forward rate between dates i - 1 and i based on specified accrual method [1/yr]
N	notional principal amount for the floating leg [$]
R	fixed coupon rate [1/yr]
s	ratio of notional principals in the fixed and floating legs, typically 1

10

If the zero curve was properly constructed, then equation 6 should be true for each of the calibration swaps.

Using Calibration Instruments to Check Zero Curve Accuracy

A general approach to validating the zero curve would be as follows. Use the primary zero curve to create a set of N_S plain, vanilla, fixed-for-floating, interest-rate swaps, where those swaps have a representative distribution of tenors. Each of these swaps should satisfy equation 6. Now price each of these swaps using the replicate zero curve, where we can expect deviation from equation 6, and define:

$$X_i = \frac{\left(B_{flt}\right)_i - \left(B_{fx}\right)_i}{\left(B_{fx}\right)_i},\ i = 1,\ldots,N_S \qquad 11$$

Use the transformed study-model error of equation 7, along with the C_{MK} methodology of Chapter 1, to calculate the replicability for this pair of curves. Tolerances should be traceable to the use case, or alternatively to a standard tolerance range, such as proposed in reference 26. In past validation studies, the tolerance half-range (= $F_U = F_L$) has often been set at 3 × DV01, the price change associated with an interest rate deviation of 3 basis points. In the present case, we would use equation 5 to translate the interest rate tolerance to price tolerance, where the modified duration is that of the ensemble set of calibration instruments.

$$F_U = F_P = \left|\frac{\Delta P_T}{P_T}\right| = \left|-\Delta y MD_T\right| = 0.0003/\text{yr} \times MD_T \qquad 12$$

In the above equation, P_T and MD_T are the value and the modified duration of the ensemble of quotes or trades.

$$P_T = \sum_{i=1}^{N_S} P_i$$

$$MD_T = \frac{1}{P_T} \sum_{i=1}^{N_S} P_i MD_i \qquad 13$$

Mathematical Correctness of Interest Rate Calculations

Hagen and West[28] provide detailed guidance to the restrictions on interest rate curves, as summarized in Table 4. We discuss validation criteria for four of these restrictions: positive forward rates; continuous, smooth forward rate curve; localization of perturbations; and forward rates stability to perturbations.

Table 4: From reference 28, five criteria are provided for well-behaved interest rate curves. The table lists a number of interpolation schemes, along with their characteristics. The reader is referred to the reference for more detail.

Yield curve type	Forwards positive?	Forward smoothness	Method local?	Forwards stable?	Bump hedges local?
Linear on discount	no	not continuous	excellent	excellent	very good
Linear on rates	no	not continuous	excellent	excellent	very good
Raw (linear on log of discount)	yes	not continuous	excellent	excellent	very good
Linear on the log of rates	no	not continuous	excellent	excellent	very good
Piecewise linear forward	no	continuous	poor	very poor	very poor
Quadratic	no	continuous	poor	very poor	very poor
Natural cubic	no	smooth	poor	good	poor
Hermite/Bessel	no	smooth	very good	good	poor
Financial	no	smooth	poor	good	poor
Quadratic natural	no	smooth	poor	good	poor
Hermite/Bessel on rt function	no	smooth	very good	good	poor
Monotone piecewise cubic	no	continuous	very good	good	good
Quartic	no	smooth	poor	very poor	very poor
Monotone convex (unameliorated)	yes	continuous	very good	good	good
Monotone convex (ameliorated)	yes	continuous	good	good	good
Minimal	no	continuous	poor	good	very poor

Forward rates must be positive.

Following Hagan and West[28] the continuously compounded rate y(t) (the zero curve) and the discount factor to present value $Z(0,t)$ are related by:

$$Z(0,t) = \exp\left[-y(t)t\right] \qquad 14$$

In the derivation below, we will ignore day-count conventions, and we will assume continuously compounded interest rates. The conclusions are relevant, regardless of the contract conventions.

The forward discount factor $Z(0,t_1,t_2)$ describes the discount factor that is applicable between future times t_1 and t_2, where $t_1 < t_2$. In the absence of arbitrage, the forward discount factor must satisfy:

$$Z(0,t_1)Z(0,t_1,t_2) = Z(0,t_2) \qquad 15$$

The forward rate $f(0,t_1,t_2)$ is defined by:

[28] P.S. Hagan and G. West, Methods for Constructing a Yield Curve, WILMOTT Magazine (May, 2008) finmod.co.za/interpreview.pdf.1,3,4,6.

$$\exp\left[-f(0,t_1,t_2)(t_2-t_1)\right]\equiv Z(0,t_1,t_2)=\frac{Z(0,t_2)}{Z(0,t_1)}<1 \qquad 16$$

Equation $\exp\left[-f(0,t_1,t_2)(t_2-t_1)\right]\equiv Z(0,t_1,t_2)=\frac{Z(0,t_2)}{Z(0,t_1)}<1$ can only be true for $f(0,t_1,t_2) > 0$; *the forward rate must be positive*. Combining the above three equations:

$$-\log\left[Z(0,t_1,t_2)\right]=f(0,t_1,t_2)(t_2-t_1)=\log\left[\frac{Z(0,t_1)}{Z(0,t_2)}\right]=y(t_2)t_2-y(t_1)t_1 \qquad 17$$

Substituting $t_1 = t$ and $t_2 = t + \varepsilon$, define the instantaneous forward rate:

$$f(t)\equiv\lim_{\varepsilon\to 0} f(0,t,t+\varepsilon)=\lim_{\varepsilon\to 0}\frac{y(t+\varepsilon)(t+\varepsilon)-y(t)t}{(t+\varepsilon)-t}=\frac{d}{dt}\left[y(t)t\right] \qquad 18$$

Equation $f(t)\equiv\lim_{\varepsilon\to 0} f(0,t,t+\varepsilon)=\lim_{\varepsilon\to 0}\frac{y(t+\varepsilon)(t+\varepsilon)-y(t)t}{(t+\varepsilon)-t}=\frac{d}{dt}\left[y(t)t\right]$ provides the transformation to calculate the instantaneous forward rate. This equation also states the *y(t)t must exhibit a positive slope with respect to t*.

Forward rate curve must be continuous

From McCulloch and Kochin[29], "A discontinuous forward curve implies either implausible expectations about future short-term interest rates, or implausible expectations about holding period returns." With reference to equation $f(t)\equiv\lim_{\varepsilon\to 0} f(0,t,t+\varepsilon)=\lim_{\varepsilon\to 0}\frac{y(t+\varepsilon)(t+\varepsilon)-y(t)t}{(t+\varepsilon)-t}=\frac{d}{dt}\left[y(t)t\right]$, this implies that *y(t)t, and therefore y(t), must be smooth*. In particular, *this precludes linear interpolation between calibration points of the zero curve.*

In fact *a smooth, forward curve is desirable, but not required*. To conform to this constraint,

$$\frac{d}{dt}\left[y(t)t\right]=y(t)+t\frac{dy}{dt} \qquad 19$$

Should also be smooth, which implies *that dy/dt should also be smooth.*

Zero-curve stability: localization

Changing one of the instruments used to bootstrap the zero curve, should only materially change the curve in the immediate vicinity of that instrument's tenor. Suppose that the zero curve $y(t)$ is bootstrapped using the yields, Y_i, $i = 1, \ldots, N_C$, on a set of N_C zero-coupon bonds or "strips," with maturities t_i, $i = 1, \ldots, N_C$, where $t_i > t_{i-1}$. Artificially change the yield on the jth bond by a small

[29] H. McCulloch and L.A. Kochin, Accurate monotonicity preserving cubic interpolation, *SIAM Journal of Scientific and Statistical Computing* **4**(4) (1983) 645-654.

perturbation ΔY, and bootstrap a synthetic zero curve $y^*(t)$, based on the modified set of bonds. A reasonable criterion for localization is:

$$\int_{t_1}^{t_{N_c}} \left| y^*(s) - y(s) \right| ds \le G \Delta Y \left(t_{j+1} - t_{j-1} \right) \qquad 20$$

where G is a factor chosen on the range [½, 1]. This is depicted in Figure 4.

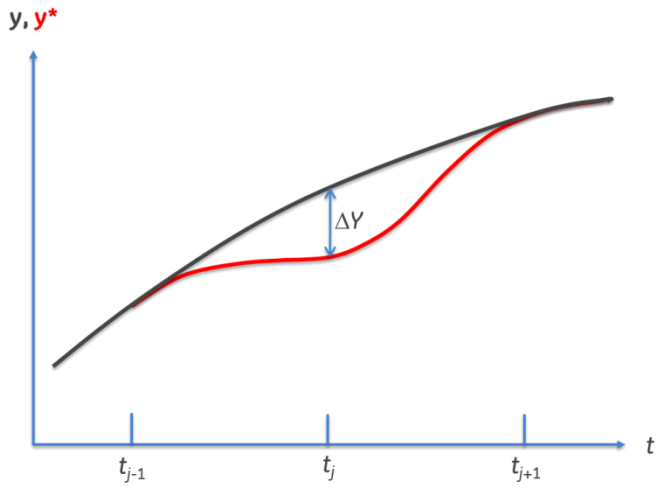

Figure 4: The calibration instrument at $t = t_j$ is synthetically perturbed to test the localization capability of the interpolation scheme.

Low values for G are associated with excellent localization of perturbations, as summarized in Table 5.

Table 5: The lower the value of G in inequality 16, the more stringent the test.

Value of G in inequality 16	Goodness of localization
0.5	Excellent
1	Satisfactory
>1	Concern

Localization and stability, as defined here, depend on the interpolation/extrapolation schemes used. Reference 28 has a detailed discussion of the characteristics of several interpolation schemes. In general, the evaluation recommended in inequality 16 (and approximations 17 below) should be repeated at each calibration point, using the focus model, not the replicate.

Forwards stability and localization

The forwards curve should be continuous, as mentioned previously; it should also be stable to perturbations of the calibration set. With reference to the above perturbation experiment, we can also calculate a synthetic instantaneous forward rate curve $f^*(t)$. We should expect the following rough guidance to be true:

$$f^*(t)-f(t)=y^*(t)-y(t)+t\frac{d}{dt}\left[y^*(t)-y(t)\right]$$

$f^*(t)-f(t)\approx 0$, for $t<t_{j-1}$ and for $t>t_{j+1}$

$$f^*(t)-f(t)\approx \Delta Y\left[\frac{t-t_{j-1}}{t_j-t_{j-1}}+\frac{t}{t_j-t_{j-1}}\right]=\Delta Y\frac{2t-t_{j-1}}{t_j-t_{j-1}}, \text{ for } t_{j-1}\le t<t_j \qquad 21$$

$$f^*(t)-f(t)\approx \Delta Y\left[\frac{t-t_j}{t_{j+1}-t_j}-\frac{t}{t_{j+1}-t_j}\right]=\Delta Y\frac{-t_j}{t_{j+1}-t_j}, \text{ for } t_j<t\le t_{j+1}$$

If the difference between the two curves is anywhere much larger in magnitude than these linearizations, there is cause for concern.

Validating Deterministic Financial Instrument Valuation Models

We define deterministic financial instruments as those where the timing of all payments and receipts throughout the lifetime of the instrument are known with certainty. Furthermore, the size of those transfers is also known or can be forecast with reference to one or more known interest rate curves, along with knowledge of spot exchange rates. This category includes hedging instruments:

Interest-rate swaps
Cross-currency, interest-rate swaps
Foreign-exchange forward contracts
As well as debt instruments:

Fixed-rate bonds
Floating-rate notes
Loans

In each case, a series of monetary exchanges b_i are expected at definite times in the future t_i. The present value B (t = 0) of the instrument is:

$$B=\sum_i b_i Z(0,t_i) \qquad 22$$

Z is the discount factor. The b_i may be positive (receipts) or negative (payments). If the payment is based on a floating rate of return, then:

$$b_i=Nf(0,\Delta t,t_i)\alpha(\Delta t,t_i) \qquad 23$$

Where N is the notional principal, f is the forward rate for the time period since the last interest settlement [t_i - Δt, t_i], and α is the time accrual associated with the same period of time. If the monetary exchange is not dollar denominated:

$$b_i=b_{fi}F_0(0,t_i) \qquad 24$$

Where b_{fi} is the monetary value in the foreign currency, and F_0 is the forward exchange rate for the foreign currency (dollar value of one unit of foreign currency), corresponding to the date of the exchange. The forward exchange rate is calculated as:

$$F_0(0,t_i) = S_0 \frac{Z_f(0,t_i)}{Z(0,t_i)} \qquad 25$$

Where S_0 is the spot rate for currency exchange, and Z_f is the applicable discount rate for the foreign currency. Given the appropriate zero-coupon interest-rate curves and forward-rate curves, the remaining complexity is involved in the day-count and other quotation conventions.

Validating Deterministic Valuation Model Accuracy

The accuracy of the model can be assessed if benchmark valuations are available for a representative set of financial instruments, either from market quotations, from counter-party valuations, or from a widely accepted benchmark model. In such cases, model capability can be calculated, as described in Chapter 1. Recall that *use-case-specific tolerances on error, which should be agreed with the model's end user, are required*.

Validating Deterministic Valuation Model Replicability

Two levels of replication are possible:

- Full replication, casual or exact, starting with the same primary data as the model developer
- Partial replication, starting with the zero and forward curves used by the model developer

The second approach may be preferred, if independent validation has been performed on the interest-rate curves.

In either case, the model valuations should be compared across a representative set of relevant, financial instruments. The comparison can be summarized by calculating the model replicability described in Chapter 1. The tolerances used for replicability may well be different from those used for model capability. In replication, we are looking for fidelity, not absolute accuracy. Tolerance standards, such as those proposed in reference 26, may be reasonable substitutes for use-case-linked tolerances, in the case of replicability studies.

Validating Stochastic Valuation Model Accuracy

We define stochastic financial instruments as those where the timing of all payments and receipts, and/or the size of those transfers have a random component. As an example, an option gives the holder the right to enter a transaction, but not the obligation. The financial assets of interest for this class of model include derivatives:[20,21]

Credit-default swaps
Callable (cancelable) interest-rate swaps
Interest-rate caps and floors
As well as debt instruments:

Callable and puttable bonds
Survivor's option bonds
Pricing the above instruments involves the use of stochastic models, since the valuations include a random component. Pricing of credit default swaps require estimates of risk-neutral default probability. Cancelable interest-rate swaps are priced like swaptions, using Black's model, the Hull-White model, and/or the Black-Karasinski model, all of which require an estimate of interest-rate volatility. The same is true of callable and puttable bonds.

The no-arbitrage models (*e.g.*, One-Factor Hull-White, Two-Factor Hull-White or Black-Karasinski) are used to predict the term structure (time dependence) of the instantaneous short rate $r(t)$. The instantaneous short rate is the interest rate at which an entity can borrow money for an infinitesimally short period of time from time t. The relationship between short rate and instantaneous forward rate is shown for the case of the One-Factor Hull-White model in equation 27. Such a model results from an assumed variational equation for dr, and the equation is called a drift-diffusion process or an Ito process, wherein the coordinates are time and a Wiener process coordinate. The process is converted to a partial differential equation (pde) using Ito's lemma, a stochastic calculus analog to the chain rule for differentiation.[30] The stochastic calculus approach is used widely in finance, particularly for options pricing.

The no-arbitrage models are formulated to exactly fit today's term structure. As an example, the drift-diffusion process for the Hull-White model is:

$$dr = \left[\theta(t) - ar\right]dt + \sigma dz \qquad 26$$

Where θ/a is the mean reversion level of the short rate, and a is the rate of mean reversion. σ^2 is the instantaneous variance of the rate.

These models include one, two or more parameters in the process formulation. One or more of these parameters can be assigned time dependence. Estimating the parameters of the term-structure model, using observable instrument pricing and/or present or past interest rate movements, is referred to as calibration. In section 0, we describe the mechanics of calibration. In most cases, we use price quotes for traded interest-rate-sensitive instruments to calibrate our models. This is referred to as risk-neutral calibration; whereas calibration based on past interest-rate movements is called real-world calibration; the choice of calibration approach depends on the intended use case.

Validating Calibration of the Term-Structure Model

The calibration procedure involves curve fitting, using the market pricing (often calculated from the quoted volatility, using Black's model) of a selected set of market-quoted instruments, most often swaptions or interest-rate caps and floors; numerous curve fitting algorithms are available.

The resulting model can now be used to calculate the prices of the N_C instruments used for calibration. Let P_i represent market pricing and P_i* represent model pricing, where $i = 1, \ldots, N_C$. Define:

[30] J.C. Hull, *Options, Futures, and Other Derivatives*, 9th Ed., Pearson (2015).

$$X_i = \frac{P_i^* - P_i}{P_i},\ i = 1,\ldots,N_C$$ 27

This transformed pricing sample is used with the model capability equations of Chapter 1 to measure calibration capability. The values of relative tolerances F_U and F_L are best related to the use case, but alternatively we may use experience-based tolerances, as proposed in reference 26.

As important as calculation of the calibration capability, as described above, is visual confirmation that the calibration is reasonable. To do this test, requires the appropriate Hull-White interest-rate tree for the class of calibrating instruments used, and conversion of market volatility quotes to instrument prices, using Black's model. This approach was demonstrated in the recent validation of GE's Credit Value Adjustment model (GETM-RR-004)[31]. The model is implemented in vended software, Adaptiv Riskbox[32]. The calibrated parameters from this model, were used to construct a Hull-White tree in Matlab, which was then used to price the calibration instruments from the model, 19 interest-rate caps with maturities from 1 to 20 years. The satisfactory comparison between the market volatility quotes and the validators' model is shown in Figure Error**! No text of specified style in document.**.

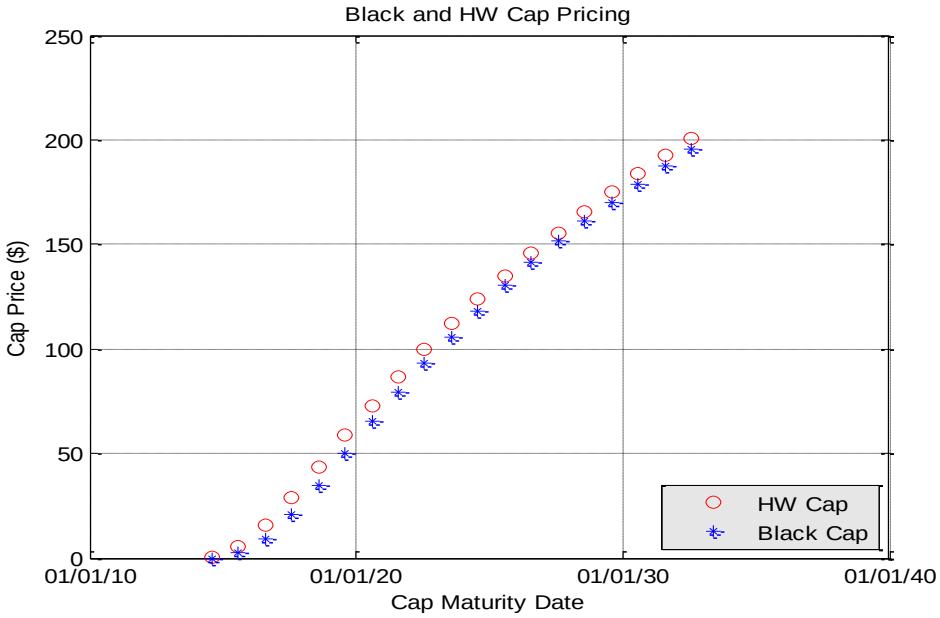

Figure **Error! No text of specified style in document.**: Plot of 19 interest-rate-cap market volatility quotations, after conversion with Black's model, shown in blue, versus price estimates from the validators' Hull-White model, built with the parameters from the original Riskbox calibration, shown in red.

[31] Z. Yang, J. Cline and J. Alvarez, Adaptiv Credit Value Adjustment GETM-RR-004 GMGV, Model Validation Technical Report (November 2013).

[32] Adaptiv Riskbox, Sungard Risk Management & Analytics, http://www.sungard.com/solutions/risk-management-analytics/enterprise-risk/adaptiv/adaptiv-riskbox.

The choice of calibration procedure is itself an art. Should the validator wish to investigate one or more alternative calibration procedures, then the validator can and should investigate the difference in pricing of the calibration instruments between the provided calibration P_i* and the challenger calibration P_i**. The relevant error parameter is:

$$X_i = \frac{P_i^* - P_i^{**}}{P_i^{**}},\ i = 1,\ldots,N_C \qquad 28$$

This error sample can be used with the model capability equations of Chapter 1 to calculate the difference between the study and challenger models, which in this case, represents the *distance* between models.[33]

A more stringent test of the applicability of the model calibration is to price an independent set of market-quoted transactions, not those used in the calibration; but, rather, a set of instruments representative of the end use. We would intuitively expect that the model capability would be lower for the pricing of independent instruments, versus its capability to price the calibration set of instruments. The set of instruments chosen should be representative, with respect to instrument type, tenor and currency, of the use case. Equation 23 is applicable.

Technical Correctness of Calibrations

As an aid to the reader, we describe the mechanics of calibration, using the Hull-White model as an example.

We state the generalized One-Factor, Hull-White, short-rate stochastic differential equation (SDE), *i.e.*, the process equation:

$$dr = \left[\theta(t) - a(t)r\right]dt + \sigma(t)dz \qquad 29$$

We include the time-dependent functions $a(t)$ for the mean reversion and $\sigma(t)$ for the volatility. $\theta(t)$ is chosen so as to exactly fit the term structure of interest rates observed currently in the market. $\theta(t)$ is a function of $a(t)$ and $\sigma(t)$. For the case of constant parameters, the One-Factor Hull-White model is given by equation $dr = \left[\theta(t) - ar\right]dt + \sigma dz$:

$$dr = \left[\theta(t) - ar(t)\right]dt + \sigma dz \qquad 30$$

The short rate is related to the instantaneous forward rate $f(0,t)$ at future time t by:[34]

$$\theta(t) = \frac{\partial f(0,t)}{\partial t} + af(0,t) + \frac{\sigma^2}{2a}\left(1 - e^{-2at}\right) \qquad 31$$

[33] C. Martini and P. Henaff, Model Validation: theory, practice and perspectives, Zeliade White Paper (May, 2011).

[34] J. Hull, Properties of Ho-Lee and Hull-White Interest Rate Models, addendum to *Options, Futures, and Other Derivatives, 9th Ed.*, Technical Note No. 31, http://www-2.rotman.utoronto.ca/~hull/technicalnotes/TechnicalNote31.pdf

To calibrate the two parameters, a and σ, given the market volatility quotes of interest-rate caps or swap options, we need to convert the volatilities to prices using Black's model, then use the prices as benchmarks for calibration.

Affine[35] term structure models, such as the Hull-White model, present appealing properties.[36] In particular, closed-form solutions for interest-rate caps/floors and efficient price approximation methods for European swap options are often available. Also, Monte Carlo simulation is relatively straightforward. In low-dimensional cases such as the One-Factor Hull-White model, path-independent products with an early exercise provision can also be evaluated efficiently using lattice methods. Finally, all kinds of interest rates can be computed readily from the short-rate factors.

In spite of the popularity of the Hull-White model, a good calibration strategy plays a central role to its performance in applications. We introduce some guidelines and practices for curve construction. Deviations from these practices may result in poor calibrations, and so should be considered carefully by the validator.

We focus on the following points:

- The choice of constant or time-dependent mean reversion and volatility.
- The choice of calibration instruments, and whether to calibrate locally or globally.
- The choice of whether to optimize on the mean reversion $a(t)$ and the volatility $\sigma(t)$ together or separately.

σ primarily influences the level of the implied volatility curves without much changing their shapes. On the other hand, even the monotonicity of the implied volatility curve can change due to the choice of the mean reversion parameter a. Calibrating the Hull-White model with constant parameters, σ and a, to a market wherein volatility exhibits a local maximum will generally yield inaccurate results.

To overcome this difficulty, adopting piecewise constant (step) functions for $a(t)$ and $\sigma(t)$ is common. The parameters are approximated as piecewise constant, and the number of segments is matched to the number of calibration instruments, with the view towards calibrating using a bootstrap-like method. This approach is less than satisfactory; the calibration parameters tend to manifest large value swings, as functions of time, resulting in instabilities, especially troublesome when pricing exotics. This approach to calibrating the Hull-White model, including the use of piecewise constant parameters, may be acceptable for many uses, but not for the purpose of pricing exotic products.

Another choice is to constrain the parameters to various time-dependent functional forms. This approach can mitigate the issues of large discontinuities. For example, one can use a logistic function to model mean reversion a, and a cubic spline to model volatility σ.

[35] An affine term structure model is a financial model that relates the discount curve to a spot rate model. It is particularly useful for inverting the yield curve—the process of determining spot rate model inputs from observable bond market data.

[36] S. Gurrieri, M. Nakabayashi and T. Wong, Calibration of Hull-White Model, Working Paper (January, 2014) http://ssrn.com/abstract=1514192.

For the second point, we should always calibrate the model to market-quoted, liquid vanilla products that are relevant to our application. The typical vanilla products are swaptions and caps/caplets. Let us give three examples where the calibration instruments are chosen with the use case in mind.

> Example 1: pricing a Bermudan swaption. The mean reversion parameter is crucial in this case. Because the Bermudan swaption provides the buyer with exercise timing flexibility during the life of the contract, the relative values associated with exercising "now" or "later" depend on the correlation of the swap rates at the different exercise time points. This correlation is controlled by the mean-reversion parameter. The calibration of the mean reversion parameter to non-callable instruments will not incorporate information related to the correlation structure, which is of necessity for callable products. Therefore, European swaptions best serve as the calibration instruments. For example, for a 10×1 Bermudan swaption the most relevant calibrating instruments are the 1×10, 2×9, 3×8, …, 10×1 co-terminal European swaptions. (An $n \times m$ swaption is an n-year European option to enter into a swap lasting for m years after option maturity.)[37]
>
> Example 2: pricing a non-callable exotic, whose payoff is LIBOR based with a coupon that is capped/floored. Calibrating $\theta(t)$ by the yield curve and the volatility of the caps is an option. As discussed in Example 1, the mean reversion parameter that captures the correlation structure of rates at different time points will considerably impact the price in a callable product. However, in this example of a non-callable exotic, calibrating to relevant caps/caplets is preferred to the use of swaptions.[38]
>
> Example 3: pricing the callable version of example 2. Since the 2007 credit crisis, discounting interest rate curves and forward rate curves are separately modeled to capture the adjustment of credit risk. As a result, a One-Factor, Hull-White model, which modeled the price of this product satisfactorily before 2007, is likely inadequate now. The current market price structure might require modeling as, at least, a two factor model.[39]

Local calibration is of the bootstrap type. This approach produces a perfect calibration to the calibrating instruments, and has the advantage of requiring only low-dimensional optimization. However, as mentioned above, this approach result in discontinuities in the parameter values, and the calibrations exhibit significant dependency on the initial values used in the optimization algorithm. In global calibration, using time-dependent, parametric expressions for the calibration parameters, the calibration parameters are simultaneously fit to all of the calibration instruments. The calibrations are not exact, and calibration requires advanced optimization approaches.

---

[37] L. Andersen and J. Andreasen. Factor dependence of Bermudan swaptions: fact or fiction*, J Financial Economics*, **62** (2001) 3-37.

[38] D. Brigo and F. Mercurio. *Interest Rate Models – Theory and Practice, With Smile, Inflation and Credit*. Springer Finance (2006).

[39] F. Mercurio, Interest Rates and The Credit Crunch: New Formulas and Market Models. Bloomberg Portfolio Research Paper No. 2010-01-FRONTIERS (February 2009) http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1332205.

For the third point, whether to calibrate both parameters simultaneously, the choice depends on the application. Optimizing only on the volatility typically produces stable, reasonably accurate results, for pricing a particular product with a known maturity and tenor. However, the resulting parameters may not be useful for pricing other products. If we want to fit the entire swaption matrix of volatilities implied by Black's model, where the swaption designated by Σ_{ij} matures at time *i* after duration *j*, optimizing mean reversion and volatility simultaneously is recommended.

As is evident from this section, the Hull-White model is not universally robust; the modeler must understand the use case when choosing his calibration method.

Use of Monte Carlo Simulation

Certain models provide analytical valuations of particular stochastic financial instruments. For example, the One-Factor, Hull-White model provides an analytical valuation of a European bond option, the right to buy or sell a particular bond at a set price at a certain point in time. Analytical expressions are not generally available across the range of interest-rate derivative instruments, and modelers will often select Monte Carlo simulation to perform such valuations.

Given a calibrated, short-rate model, we can run Monte Carlo simulation to price relevant instruments. Care needs to be taken to understand the standard errors associated with the price estimates. Using a large number of Monte Carlo simulation trials results in reduced simulation error, *i.e.,* a better estimate of instrument price. As an example, if we wish to compute *z*, the expected value of *f*(*x*), where *x* is a random variable from a known or assumed distribution:

$$z = E\left[f(x)\right] \tag{32}$$

Using n simulation trials, we can construct the following 95% confidence interval for the result:

$$z \in \left[\mu_n - 1.96\frac{\sigma_n}{\sqrt{n}},\ \mu_n + 1.96\frac{\sigma_n}{\sqrt{n}}\right] \tag{33}$$

Where the estimate of the expectation is the mean value of the trial results:

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \tag{34}$$

The x_i values are generated randomly, using the assumed or known distribution for *x*. For example, *x* might be the normally distributed Wiener diffusion coordinate in a rate process equation. The estimation error is

$$\sigma_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left[f(x_i) - \mu_n\right]^2} \tag{35}$$

This last equation shows that the numerical estimation error and the confidence intervals on the calculated expected value are proportional to the square root of 1/*n*. How many simulation trials are needed to reach a desired accuracy level? Two accuracy criteria are commonly used:

- Absolute accuracy: compute the quantity z to an accuracy ϵ;
- Relative accuracy: compute the quantity z to an accuracy $|z|\epsilon$.

Confidence intervals suggest what one must do to achieve such accuracies; *i.e.*, for absolute precision ϵ, one should choose n such that:

$$n \approx \frac{1.96^2 \sigma_n^2}{\varepsilon^2} \qquad 36$$

For the relative precision ϵ, the appropriate number of simulations is:

$$n \approx \frac{1.96^2 \sigma_n^2}{\varepsilon^2 \mu_n^2} \qquad 37$$

To determine the required number of simulations, for a particular precision, requires a reliable estimate of σ_n, which, in itself, requires a number of *test* simulations. In the following example, we wish to achieve 0.1% absolute precision (ϵ = 0.001), where $f(x)$ follows the *standard normal* distribution. By using a sufficient number of test simulations, we can achieve a good estimate of σ_n, which will actually reduce the estimated number of simulations necessary to meet the precision requirement.

Table 6: In the example, a number (first column) of test simulations of $f(x)$ are performed in order to estimate σ_n for use in equation 33. By performing 100 test simulations, versus 10, we tend to calculate a better, lower estimate of σ_n. This results in a lower number (second column) of required simulations to calculate $E[f(x)]$.

Number of Test Simulations	Number of Simulations (millions)
10	6.0
100	3.9
1000	3.9
10000	3.8
100000	3.8

Plotting μ_n versus n, as was done in Figure 5 for $f(x) = x$, is often helpful to visualize the convergence characteristics, useful for selecting an appropriate value for n. The 95% confidence intervals on the

expected value shrink as we increase the number of simulation trials.

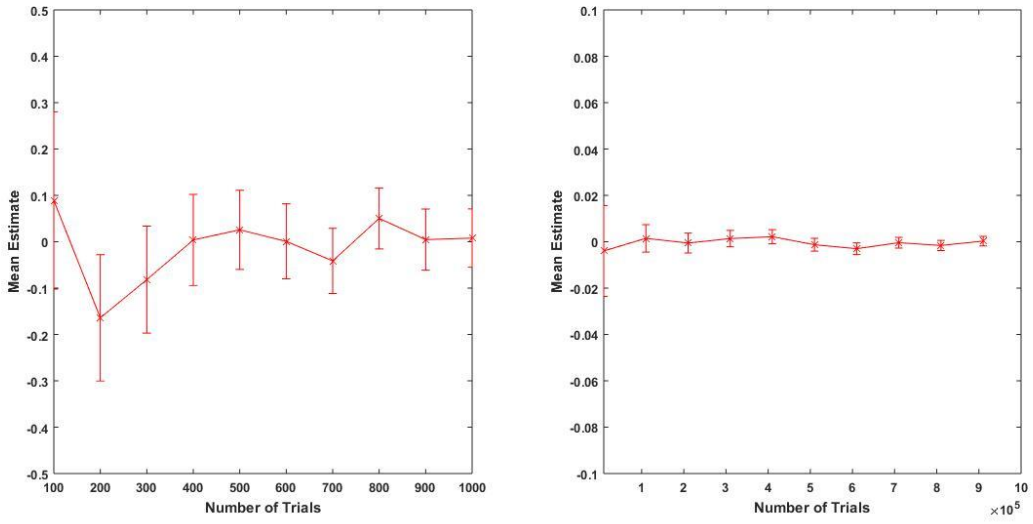

Figure 5: Monte Carlo simulations of $E[f(x)]$ where $f(x) = x$, and x is a random variable represented by the standard normal distribution. Notice that $n \in [100, 1000]$ in the left panel, and $n \in [1000, 1{,}000{,}000]$ in the right panel.

In the case of the One-Factor, Hull-White, stochastic rate process of equation 25, a simulation of the short rate r_i at time t_i is obtained by iteratively evaluating:

$$r_i = r_{i-1}\left[1 - a_i \Delta t\right] + \theta_i \Delta t + \sigma_i N(0,1)\sqrt{\Delta t} \qquad 38$$

θ_i, a_i, and σ_i are the values of these parameters at time t_i. Δt is the time step: $t_i = t_{i-1} + \Delta t$. $N(0, 1)$ is the standard normal distribution.

Check of Term-Structure Simulations

While the end use of a stochastic interest-rate model is to value an instrument, one should check intermediate results, including interest-rate trees or simulated interest-rate curves. Results should be consistent with expectations of smooth, upward sloping yield curves. As part of the validation process described in reference 31, the authors simulated the short rate, using the Hull-White model, as shown in Figure 6. Using the same calibration, they replicated the IR curve simulations from the vended model (RiskBox) using Matlab programming, resulting in visual confirmation that the two simulations were similar. Both simulations generated a kink in the simulated IR curves which required further analysis. In the case of interest-rate-tree replication, high precision is expected, since the calculations are not stochastic. In replicating simulated yield curves, model stochasticity comes into play, and many simulations are required for visual confirmation or null hypothesis testing.

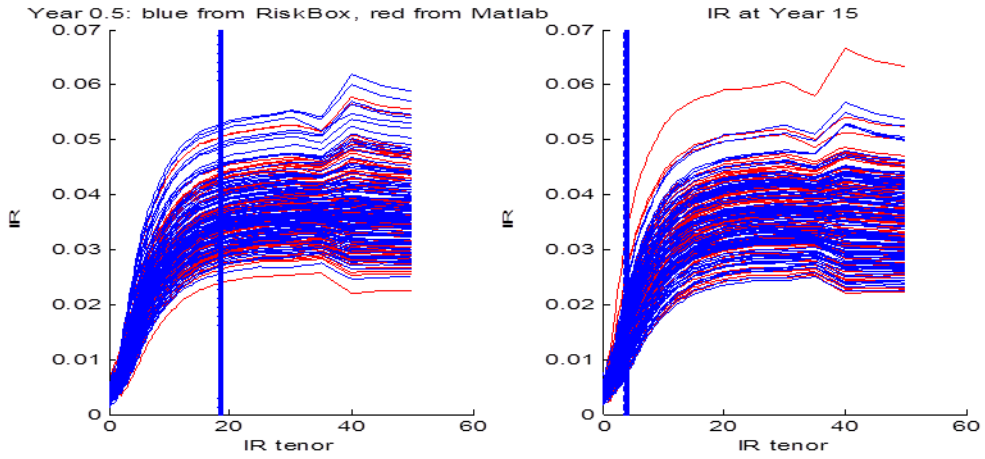

Figure 6: 100 replicated IR paths generated by Monte Carlo simulation with Matlab (red) overlaid with 100 IR paths from RiskBox.

Summary

Most of the relevant financial instruments to be valued fall into one of two categories.

In the first category are instruments that, so long as neither party to the agreement defaults, assure a stream of receipts and payments at times certain; the magnitude and direction of said transfers will typically depend on movements in the relevant zero-coupon, interest-rate curve(s). We label these deals as deterministic.

In the second category are contracts that bestow one of the parties with the right, but not the obligation, to buy or sell an instrument from the first category on a certain date, or dates, in the future, with pricing agreed at the time of the contract. Whether or not the party chooses to exercise their option depends on changes in the reference interest-rate curve(s) that occur between the date when the option is written and the date(s) when it can be exercised. We label these deals as stochastic.

For deterministic deals, the valuations depend on the calculation of discounting factors for future payments, and on the size of those future payments, in the case of floating-rate payment streams. Where those payments occur in foreign currency, future exchange rates come into play. All of these calculations require construction of zero-coupon, interest-rate curves, and proving the correctness of curve construction is a central theme in validation. Validation includes benchmark testing with liquid instruments and model replication. We maintain that the acceptance criteria in both cases should be statistical tests, model capability and model replicability, using tolerance ranges that are traceable to the use case. We noted in Chapter 1, that many deals, on the order of one hundred, should be sampled and included in these calculations, for the statistics to be valid.

We also put forward tests of theoretical soundness, including curve monotonicity, smoothness, stability and localization. These tests can be done visually, but we have provided mathematical frameworks as well. The mathematical approaches are robust and easy to automate.

For stochastic deals, the valuations are calculated from interest-rate trees (finite difference calculations) and/or from Monte Carlo simulation. Neither of these mathematical approaches will provide fair-market valuations, unless a proper calibration (determination of model parameters) is performed. Because the models are highly idealized, and the underlying assumptions are not universal, the calibration is critical. The type(s) of liquid instrument(s) used to perform the calibration, and the mathematical approach, are critical; the calibration strategy should be selected with the end use of the model in mind. Step-function parameter representations must be viewed with caution, since poor stability and poorly shaped interest-rate curves can result.

Benchmark testing, replication, and challenger models (*e.g.*, alternative calibrations) are all potent tools for validation. Again, we believe that the acceptance criteria should be the same statistical tests described for deterministic deals.

When building interest-rate trees or simulations, there are many decisions, including the choice of time step size. With Monte Carlo simulation, it is critical that we perform enough simulations to assure useful confidence intervals on the expected response.

Prepayment

Prepayment

Jin Xia and Sanghee Cho

Contents

1 Overview

This chapter describes the key quantitative components in validating a prepayment model. A prepayment model aims to study the empirical prepayment behavior of an asset, associate it with risk factors, and make forecasts of the prepayment behavior.

The commonly used metric representing prepayment behavior is the prepayment speed, or single monthly mortality (SMM) for monthly data. The prepayment speed of an asset can be defined as

$$\text{prepayment speed}(t) = \frac{\#\text{ of prepayments}(t)}{\#\text{ of deals}(t)},$$

where t generally represents a real time point or age of interest. Furthermore, if we estimate the probability of prepayment for each deal, the prepayment speed for an asset can be estimated as

$$\widehat{\text{prepayment}}\text{ speed}(t) = \frac{\sum_i \widehat{\Pr}(\text{ Deal } i \text{ is prepaid at } t)}{\#\text{ of deals at } t}.$$

Statistical models can be used to estimate the probability that the i-th deal is prepaid at t, associating with risk factors, such as macro-economic variables, amortization terms and other deal information. A most commonly used statistical model of such is the logistic regression model, which is what we focus on in this chapter. The key quantitative components in validating a prepayment model using logistic regression are discussed in the following subsections:

- Data processing
- Variable selection
- Model diagnostics
- Model performance
- Backtesting
- Small samples and rare events

Note this model output can be input to other models and analysis with the following business use cases:

- Asset-liability management
- Pricing
- Transfer pricing/match funding
- New business planning (e.g. business blueprint)
- Treasury - IRRM, Liquidity Risk, Ecap, credit, etc.

2 Logistic Regression

Logistic regression, which belongs to the family of Generalized Linear Models[40], is widely used to predict probability of prepayment. It uses a logit link function to model the probability of an event for a binary response variable. In this chapter, prepayment is the event of interest. As shown in Figure 1,

[40] Refer to (McCullagh and Nelder 1989) for detailed information about the Generalized Linear Models.

the fitted line of logistic regression has S type of curve (logit curve) that maps predictors to a value between zero and one. Logistic regression enables us to predict a probability of an event given the relevant predictors.

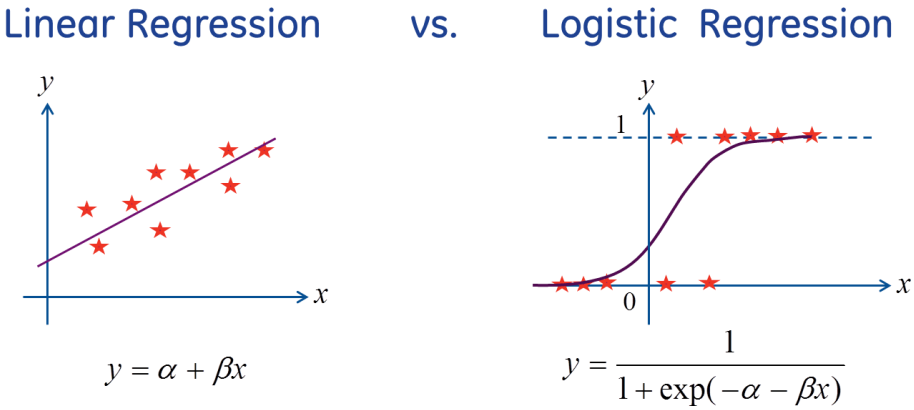

Figure 7 Difference between linear regression model vs. Logistic regression model

Let

- Y denote the binary response variable, with

$$Y = \begin{cases} 1 & \text{if prepayment happens} \\ 0 & \text{otherwise} \end{cases};$$

- $X_1, X_2, \dots, X_p$ denote the p predictors;
- π denote the probability of prepayment given $X_1, X_2, \dots, X_p$; *i.e.*, $\pi = \Pr(Y = 1 | X_1, X_2, \dots, X_p)$; and
- $\beta_0, \beta_1, \dots, \beta_p$ denote the coefficients for the predictors.

Then the logistic regression can be expressed as

$$\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

Note that the probability, after logit transformation, has linear relationships with the predictors of interest.

The model can also be presented in a matrix form. Let

- $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)'$ denote the vector of n observed response variables;
- $\mathbf{X}$ denote the design matrix, with

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix};$$

- $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_n)'$ denote the vector of prepayment probability given $\mathbf{X}$; and
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ denote the coefficients for the predictors.

Then the model can be expressed as

$$\text{logit}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta},$$

where the logit function is applied element-wise to $\boldsymbol{\pi}$.
Coefficients can be estimated by maximizing the log-likelihood (LL)

$$LL = \sum_{i=1}^{n} (1 - Y_i) \log(1 - \pi_i) + Y_i \log(\pi_i)$$

An important issue to always check is whether there are repeated observations at different levels of the predictors. Here the levels are unique values of the p-dimensional vector of predictors. The number of levels and numbers of repeated observations determine whether the asymptotic properties of certain tests are valid, as explained below. The levels are referred to as **covariate patterns** in the following context. The repeated observations are also called grouped observations.
Let

- J denote the number of covariate patterns, with $J \leq n$;
- $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_J$ denote the covariate patterns, with $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \dots, x_{jp})'$, $j = 1,2, \dots, J$;
- m_j be the number of observations for covariate pattern $\boldsymbol{x}_j$, $j = 1,2, \dots, J$, with $\sum_{j=1}^{J} m_j = n$;
- y_j be the number of events, *i.e.*, prepayments, for covariate pattern $\boldsymbol{x}_j$, $j = 1,2, \dots, J$; and
- $\hat{\pi}_j$ be the estimated probability of an event for covariate pattern $\boldsymbol{x}_j$, $j = 1,2, \dots, J$.

2.1 Data Processing

In practice, most data sets need a certain level of processing before reliable models can be fit. The processing includes formatting the data and evaluating the appropriateness of values. Here we focus on the latter. Common processing steps in evaluating data appropriateness include, but are not limited to, identifying abnormal values and treating missing data. Identifying abnormal values relies primarily on domain knowledge and is not discussed in detail here. In this section, we focus on treating missing data.

2.1.1 Missing Data

Missing data is usually a nuisance rather than a focus in model building. The model validator should evaluate the amount of missing data and whether it raises any concern for estimating and making inferences from the developed model. Two common scenarios when the amount of missing data is likely problematic are:

- The absolute amount of missing data is high.

- The absolute amount of missing data is low, and yet the data is unbalanced and missingness for the event of interest is severe.

There is no threshold on whether the amount of missing data is problematic. The model validator should make judgments using experiences on the specific type of problems. If there is uncertainty, the model validator can compare the results with and without treating missing data, and decide whether missing data is an issue and should be treated.

If the amount of missing data raises sufficient concerns, the model validator should further consider the following mechanisms of missing data (Schafer and Graham 2002) (Little and Rubin 2002):

- Missing completely at random (MCAR): The missingness does not depend on missing or observed data.
- Missing at random (MAR): The missingness depends on the observed data but not on the missing data.
- Missing not at random (MNAR): The missingness depends on the missing data.

Let D_{obs} denote the observed data, D_{mis} denote the missing data if it were observed, and D_{com} denote the complete data, i.e. $D_{\text{com}} = (D_{\text{obs}}, D_{\text{mis}})$. Let M denote the missingness. Then MCAR and MAR can be formulated as:

- MCAR: $\Pr(M|D_{\text{com}}) = \Pr(M)$
- MAR: $\Pr(M|D_{\text{com}}) = \Pr(M|D_{\text{obs}})$

The validator should evaluate the data generation process and determine the mechanism of missingness.[41]

In literature (Little and Rubin 2002) (Schafer and Graham 2002) (D. B. Rubin 1976), the following ways are suggested to deal with the above mechanisms of missingness:

- MCAR: In general, the missing data can be dropped and the model can be built on the observed data. A commonly applied method is to model the complete cases, i.e. cases without missing values. The biggest advantage of complete-case analysis is simplicity. The disadvantage is loss of precision (and bias if the missingness is not MCAR). Under MCAR, complete-case analysis is justified when the loss of precision is minimal.
- MAR: Under MAR, dropping the missing data still generates appropriate likelihood for the parameters (D. B. Rubin 1976). Any inference from the likelihood is valid. For example, the commonly used maximum likelihood estimates (MLE) are appropriate, as well as Bayesian methods. A general method for MLE in the presence of missing data is the EM algorithm (Dempster, Laird and Rubin 1977). Besides likelihood-based methods, another popular method of dealing with MAR is multiple imputation (MI). MI refers to the procedure where each missing value is replaced by a number of simulated values. As a result, a number of complete data sets are generated, which represent the imputation uncertainty. Methods for complete data are applied to each data set, and the results are combined (Schafer and Graham 2002) (D. B. Rubin 1987).

[41] Note that there is generally no way to test whether MCAR or MAR holds, given the missing data is not observed.

- MNAR: Under MNAR, an appropriate model for the missingness needs to be specified, and a large sample is needed. This is often infeasible in practice. Therefore, dealing with MNAR is usually very challenging. Here, we don't suggest any specific technique to deal with MNAR due to the complexity of the problem. Instead, we suggest for the model validator to be extremely cautious with a model developed with a MNAR data set. If the validator suspects that the departure from MAR is not severe, the validator can apply the above methods for MAR and compare the results. If the departure is apparently severe and no appropriate treatment is applied to the missing values by the model developer, the validator should challenge the feasibility of the model.

In practice, the departure from MCAR or MAR may not be severe enough to generate a significant impact on the model estimates, especially when the amount of missing data is low. In such cases, the bias from assuming MCAR or MAR may be negligible. The most commonly applied method by model developers is the complete-case analysis. As described above, complete-case analysis is only appropriate under MCAR. However, if the departure from MCAR is not severe, the bias in model estimates from assuming MCAR is probably negligible. If there is concern about assuming MCAR, the validator can apply appropriate methods as described above to treat the missing values, and compare results in order to decide whether assuming MCAR is appropriate.

2.2 Variable Selection

Purpose: A rigorous process of variable selection is needed and the model developers need to provide reasonable rationale when adding or dropping candidate variables (predictors). However, there is no exact science for this procedure and it should be always combined with statistical soundness and business intuition. For example, (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) provides step by step procedures. One doesn't have to follow the exact steps, but a logical process for variable selection must prevail. Here, plausible reasons to drop variables are listed.

2.2.1 Univariate Variable Selection

Various univariate tests can be done to examine an association between each predictor and the response variable. Hosmer, *et al.* (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) suggest to use a higher p-value cut-off (*e.g.*, 0.2 or 0.25, instead of the traditional cut-off 0.05) when we use univariate association for initial screening. If the experts believe a variable is important to the model, then that variable, even with a high p-value, should be included for further analysis.

2.2.2 Multivariate Variable Selection

When comparing candidate models, AIC and BIC are effective tools to assess the relative model performance. AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) provide summary statistics for model fit comparison. Generally, a model with lower information criteria is preferred. Hilbe (2009) provides guidance (Table 7) for selecting a preferred model. AIC and BIC can be computed as follows:.

$$AIC = -2LL + 2k$$
$$BIC = -2LL + k\log(n)$$

LL is the log-likelihood function and k is the number of parameters (usually the number of predictors + 1), representing the complexity of a model. Note that BIC penalizes heavier for the number of predictors.

Table 7 Guidance on AIC /BIC criteria

AIC (Akaike Information Criteria)		BIC (Bayesian Information Criteria)	
Difference between Models A and B (Suppose A<B)	Result if A < B	Difference between Models A and B (Suppose A<B)	Degree of Preference on A
0 < (B – A) < 2.5	No difference in models	0 < (B – A) < 2	Weak
2.5 < (B – A) < 6	Prefer A if n>256	2 < (B – A) < 8	Positive
6 < (B – A) < 9.9	Prefer A if n>64	8 < (B – A) < 10	Strong
(B – A) > 10	Prefer A	(B – A) > 10	Very Strong

2.2.3 Automatic Variable Selection (Stepwise/Backwards/Forwards)

The main idea of the automatic variable selection is to measure how each variable improves the model fit when it is added or deleted. However, decisons are based on statistical evidence alone, so careful assessment is needed. For example, consider two variables A and B with similar significance level. It is possible that B is selected over A by the algorithm, based on statistical evidence, even though the experts believe A is more important than B. Also, statistical significance for a particular variable depends on which variables are already included in the model. Automatic variable selection is not without peril.

2.2.4 Miscellaneous

There are additional considerations in variable selection:

- Multi-collinearity: Including highly correlated predictors in the model can lead to biased estimates with large standard error. Variance Inflation Factor (VIF) in linear regression indicates associations among the predictors (Menard 2002), (Allison 2000). VIF greater than 5 is cause for concern and greater than 10 suggests a serious collinearity problem (Menard 2002).
- Number of predictors in the final model: (Agresti 2013) provides a guideline based on a Monte Carlo study (Peduzzi, et al. 1996); when the number of prepayment events divided by the number of predictors is smaller than 10, then parameter estimates can be a biased, in addition to other potential issues. (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) cited (Vittinghof and McCulloch 2006); based on extensive simulations, the latter authors conclude that the "rule of 10" may be too conservative. However, (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) take issue with the latter authors' recommendation in cases where the distributions of the discrete predictors are highly skewed rather than balanced; for such cases, the rule of 10 is justified. Thus, the rule of 10 is recommended as a guideline, but acknowledge that this should not be a strict rule since

the above studies are based on simulation based study which does not cover all possible settings or data. Thus, a less stringent requirement may suffice.

- Counterintuitive coefficient sign: With limited data, it is often useful to incorporate expert knowledge.

2.3 Model Diagnostics

This section describes methods that assess the validity of fitting a logistic regression model to the data. A fitted model may be inadequate in the following ways:

- Relationships between the link logit(π) and predictors $X_1, X_2, \ldots, X_p$ are not all correctly specified. There are generally two causes of the misspecification: (1) the predictors need to be transformed in order to have linear relationships with the link; and (2) the logit link is incorrect.
- There are outliers or other strongly influential observations. They may bias the model estimates, and they are not fitted well by the model.

Three types of tools are proposed to assess the above model inadequacies. Table 8 shows which inadequacy is assessed by each tool. The tools are:

- Link specification tests
- Residual plots
- Influential observation plots

Table 8 Model Inadequacies and Diagnostic Tools

	Diagnostic Tools		
Model Inadequacies	Link specification tests	Residual plots	Influential observation plots
Misspecification of relationships between the link logit($\boldsymbol{\pi}$) and predictors $\boldsymbol{X_1, X_2, \ldots, X_p}$.	X	X	
Existence of outliers or strongly influential observations		X	X

These tools enable one to test the validity of model assumptions. They also enable one to conduct detailed analyses of the model in order to identify issues or improvement opportunities.

2.3.1 Link Specification Tests

Purpose: The logistic regression model assumes linear relationships between logit(π) and the predictors $X_1, X_2, \ldots, X_p$, as specified by the model equation

$$\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

If the linearity assumption is inappropriate, then the coefficient estimates and standard errors are biased. Therefore, it is important to assure the validity of the linearity assumption.

Two similar tests are provided to assess the linearity assumption (Hilbe 2009), namely

- Box-Tidwell test
- Pregibon Link test

1) Box-Tidwell Test

Description: The Box-Tidwell test has two steps:

- Construct the interaction[42] of each continuous predictor with its log transformation
- Fit a logistic regression with the original predictors and the above interactions

If any interaction is statistically significant, the Box-Tidwell test concludes that the linearity assumption is violated. The following criterion can be applied on each interaction.

Method	Criterion
Box-Tidwell Test	P-value < 0.05

2) Pregibon Link Test

Description: The Pregibon Link test is similar to the Box-Tidwell test, but uses the square of the hat matrix diagonal instead of the interaction term. It is carried out in two steps:

- Construct the square of the hat matrix diagonal of each continuous predictor
- Fit a logistic regression with the original predictors and the above new terms

If any new term is statistically significant, the Pregibon Link test indicates that the linearity assumption is violated.

The estimated hat matrix for logistic regression is

$$\mathbf{H}=\widehat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{1/2}.$$

Here, $\widehat{\mathbf{W}}$ is an $n \times n$ diagonal matrix

$$\widehat{\mathbf{W}} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix},$$

where $\hat{\pi}_i$ is the estimated probability of event for the i-th observation.

The following criterion can be applied to each interaction.

Method	Criterion
Pregibon Link Test	P-value < 0.05

[42] An interaction term represents the interactive effect of two or more predictors on the response. It is generally constructed as the product of these predictors. For example, let X be a continuous predictor, and the interaction of X with its log transformation is $X\log(X)$.

2.3.2 Residual Plots

This section and the next section on influential observation plot describe visual diagnostics using several normalized residuals. Visual assessment is preferred; hard thresholds are not available for these diagnostics. Visual assessment informs with respect to model adequacy and overly influential (to the model fit) observations. Where poor fit is observed, one would check for data issues or opportunity to improve the model by alternative variable selection or transformation of variables.

Purpose: The purpose of these residual plots is to assess model adequacy. An adequate model satisfies $E\{Y_i\} = \pi_i$. If such assumption is satisfied, it follows asymptotically that $E\{Y_i - \hat{\pi}_i\} = 0$ (Kutner, et al. 2005).

Description: There are three examples of residuals. Residual plots versus predictors or/and predicted probabilities with a Lowess curve can suffice (Kutner, et al. 2005). Each residual plot presents different information about the model fit; multiple residual plots are preferred. The Lowess curve of each plot is expected to be flat horizontal line with zero intercept. If significant deviations are found, influential data points or lurking variables left out of the model are typically causal. Our notation assumes the grouped case (repeated covariate patterns), but even for ungrouped data (without repeated covariate patterns), we can treat $J = n$ and $m_j = 1$ for all $j = 1, \dots, J$. However, (Agresti 2013) and (Hosmer and Lemeshow, Applied logistic regression 2000) recommend to compute residuals using grouped data if possible, especially when J is much smaller than n.

a) Pearson Residuals: The objective is to make the residuals more comparable by dividing the raw residual by an estimate of the standard deviation of the observation y_i.

$$r_j = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

b) Studentized Pearson Residuals: The objective is to fully standardize the residuals; the raw residual $(y_j - m_j\hat{\pi}_j)$ is normalized by the estimated standard error of the fitted value $\hat{\pi}_i$.

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}} = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)(1-h_j)}}$$

where, h_j is the j-th diagonal element of the estimated hat matrix $\boldsymbol{H}$ for logistic regression. $\widehat{\boldsymbol{W}}$ is the diagonal matrix wherein the j-th element is $m_j\hat{\pi}_j(1-\hat{\pi}_j)$,

$$\boldsymbol{H} = \widehat{\boldsymbol{W}}^{\frac{1}{2}}\boldsymbol{X}\left(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}^{\frac{1}{2}}$$

c) Deviance Residuals: These residuals are related to the deviance test which will be described in Section 2.4.1.

$$dev_j = sign\left(Y_j - m_j\hat{\pi}_j\right)\sqrt{2[y_j \ln\frac{y_j}{m_j\hat{\pi}_j} + \left(m_j - y_j\right)\ln\frac{m_j - y_j}{m_j\left(1-\hat{\pi}_j\right)}]}$$

2.3.3 Influential Observation Plots

Purpose: The purpose of these plots is to visualize how each observation influences the model fit. The idea is to see how goodness-of-fit test statistics, or the coefficient estimates, would change if we exclude one observation from the analysis.

Description: Identify observations that stand out in the plots, then check those observations for errors. As residual plots, each plot presents different information about the model fit; multiple plots are preferred.

a) ΔX_j^2 vs. $\hat{\pi}_j$: The ordinate is related to the Pearson Chi-Square statistic.

$$\Delta X_j^2 = \frac{r_j^2}{(1-h_j)} = r_{sj}^2$$

b) ΔD_j vs. $\hat{\pi}_j$: The ordinate is related to the deviance test.

$$\Delta D_j = dev_j^2 + \frac{r_j^2 h_j}{(1-h_j)} \approx \frac{dev_j^2}{(1-h_j)}$$

c) $\Delta\hat{\beta}_j$ vs. $\hat{\pi}_j$: The ordinate relates to the observation's influence on the estimate of coefficient β. Note that Cook's distance for logistic regression is $\Delta\hat{\beta}_j/p$.

$$\Delta\hat{\beta}_j = \frac{r_j^2 h_j}{(1-h_j)^2}$$

(Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) refer to (Martin and Pardo 2009), who derived the asymptotic distribution for Cook's distance $\Delta\widehat{\boldsymbol{\beta}}$. We can set a $\Delta\widehat{\boldsymbol{\beta}}$ threshold value to identify influential points, those that exceed the threshold. Martin and Pardo suggest using the $\chi^2_{0.5}(p+1)$ percentile as the critical value for ΔX_j^2 and $\overline{hh} \times \chi^2_{0.95}(1)$ for $\Delta\hat{\beta}_j$ where $\overline{hh}$ is the average of $h_j/(1-h_j)$. Hosmer, *et al.* (Hosmer and Lemeshow, Applied logistic regression 2000) indicate that this threshold can identify too many observations as extreme; instead they recommend focusing on observations whose values for one or more of the diagnostic statistics fall well away from the rest of the values. The following three plots are considered to be critical. Bubble plots, which plot a residual, with the size of the symbol proportional to another characteristic, can be a useful, as they introduce a third dimension to the graphical analysis.

2.4 Model Performance

This section describes methods to assess the overall goodness-of-fit of the model. A number of options are provided:

- Pearson Chi-Square, Deviance and Hosmer-Lemeshow tests
- AUC and ROC Curve / AR and CAP

The choice of method depends on the data, but at least one method should be used to assess the model performance.

The overall goodness-of-fit is robust to poor fit for a few observations. When the overall goodness-of-fit is rejected, the appropriate response is to review model diagnostics for improvement opportunities.

2.4.1 Pearson Chi-Square, Deviance and Hosmer-Lemeshow Tests

Purpose: The tests described in this section assess the goodness-of-fit of the overall model. Pearson Chi-Square test and Deviance test are classic tests for model goodness-of-fit. However, when the predictors are continuous rather than discrete, the Hosmer-Lemeshow test should be chosen.

1) Pearson Chi-Square Test

Description: The Pearson Chi-Square statistic is the sum of squares of Pearson residuals, and is defined as

$$\mathrm{X}^2 = \sum_{j=1}^{J} \left[\frac{y_j - m_j \hat{\pi}_j}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \right]^2$$

Under the null hypothesis that the fitted model is correct, X^2 follows a χ^2 distribution asymptotically with degrees-of-freedom $J - (p + 1)$. The following criterion can be applied.

Method	Criterion
Pearson Chi-Square Test	P-value < 0.05

The asymptotic χ^2 distribution is valid when each $\mathrm{m_j} \to \infty$ and J is finite, while $\mathrm{n} \to \infty$. Therefore, when one or more continuous predictors are included in the model and result in $\mathrm{J} \approx \mathrm{n}$, the Pearson Chi-Square test is inappropriate.

2) Deviance Test

Description: Similar to the Pearson Chi-Square test, the deviance test statistic is the sum of squares of deviance residuals, defined as

$$G^2 = \sum_{j=1}^{J} 2 \left[y_j \ln \frac{y_j}{m_j \hat{\pi}_j} + (m_j - y_j) \ln \frac{m_j - y_j}{m_j (1 - \hat{\pi}_j)} \right]$$

The deviance test is also the likelihood ratio test of a saturated model[43] with J parameters and the fitted model.

Under the null hypothesis that the fitted model is correct, G^2 follows a χ^2 distribution asymptotically with degrees-of-freedom $J - (p + 1)$. The following rejection criterion can be applied.

Method	Criterion

[43] A saturated model contains a unique probability estimate for each covariate pattern.

Deviance Test P-value < 0.05

As for the Pearson Chi-Square test, the asymptotic χ^2 distribution of the deviance test is valid when each $m_j \to \infty$ and J is finite, while $n \to \infty$. Therefore, when one or more continuous predictors are included in the model and result in $J \approx n$, the deviance test is inappropriate.

3) Hosmer-Lemeshow Test

Description: When the inclusion of continuous predictors causes violation of the asymptotic requirements for the Pearson Chi-Square test and deviance test, a remedy is to group the covariate patterns to form a finite number of groups with each approximately treated as one covariate pattern. The Hosmer-Lemeshow (HL) test proposes grouping based on the values of the estimated probabilities, with the following two options:

- by percentiles of the estimated probabilities
- by fixed values of the estimated probabilities

Grouping by percentiles of the estimated probabilities is preferred (Hosmer, Lemeshow, and Klar (1988)), and is commonly chosen as the default option by software packages. The choice of number of groups depends on the sample size, but 10 groups are most commonly used.

Let

- G denote the number of groups;
- $\mathrm{c_g}$ denote the number of covariate patterns in group g, $\mathrm{g} = 1,2,\dots,\mathrm{G}$;
- $\mathrm{m_{gj}}$ denote the number of observations for the j-th covariate pattern in group g, $\mathrm{j} = 1,2,\dots,\mathrm{c_g}$, $\mathrm{g} = 1,2,\dots,\mathrm{G}$;
- $\mathrm{y_{gj}}$ denote the number of events for the j-th covariate pattern in group g, $\mathrm{j} = 1,2,\dots,\mathrm{c_g}$, $\mathrm{g} = 1,2,\dots,\mathrm{G}$; and
- $\hat{\pi}_{\mathrm{gj}}$ denote the estimated probability of the event for the j-th covariate pattern in group g, $\mathrm{j} = 1,2,\dots,\mathrm{c_g}$, $\mathrm{g} = 1,2,\dots,\mathrm{G}$.

Then the HL statistic is defined as

$$\hat{C} = \sum_{g=1}^{G} \left[\frac{y'_g - m'_g \hat{\pi}'_g}{\sqrt{m'_g \hat{\pi}'_g (1 - \hat{\pi}'_g)}} \right]^2$$

where

$$y'_g = \sum_{j=1}^{c_g} y_{gj}$$

$$m'_g = \sum_{j=1}^{c_g} m_{gj}$$

$$\hat{\pi}_g = \frac{\sum_{j=1}^{c_g} m_{gj} \hat{\pi}_{gj}}{\sum_{j=1}^{c_g} m_{gj}}$$

Under the null hypothesis that the fitted model is correct, $\hat{C}$ follows a χ^2 distribution asymptotically with degrees-of-freedom $G - 2$. The following criterion can be applied.

Method	Criterion
HL Test	P-value < 0.05

2.4.2 AUC and ROC Curve / AR and CAP

Purpose: The purpose of ROC (Receiver Operator Characteristic) curve for logistic regression model is to measure the model performance from the classification (event or no event) perspective. Through varying the classification rule, we can impact the model's capability to predict prepayment. Although the primary use of a prepayment model is to predict SMM, rather than classification, the ROC curve is useful for visualizing the model's predictive power. The analysis also provides a quantitative summary statistic called AUC (Area Under the Curve).

Description: Suppose, for each observation i, we have a fitted value $\hat{\pi}_i$ from the model. We can classify whether it is prepaid or not by setting a rule as below, denoting Y = 1 for prepayment event,

$$\hat{Y}_i = \begin{cases} 1 & \hat{\pi}_i \geq \pi_0 \\ 0 & \hat{\pi}_i < \pi_0 \end{cases}$$

For a give cutoff point π_0, each observations will be classified as either 0 or 1. Then, it can be compared to the true observed event Y_i to check if the prediction rule works well. By varying the cutoff point π_0, we measure the predictive power by false positive rate and true positive rate using the whole sample. The ROC curve is constructed by plotting each associated pair of false positive rate (horizontal axis) versus true positive rate (vertical axis), resulting from sweeping the cutoff point π_0 through the entire range, as in Figure 8. If the curve passes close to the top-left corner of the plot, it means the model has a high predictive power. On the other hand, if the curve never deviates much form to the diagonal (45° line) of the plot, the model's predictive power is not better than a random guess. We can quantify this curve by calculating AUC. High AUC indicates high predictive power.

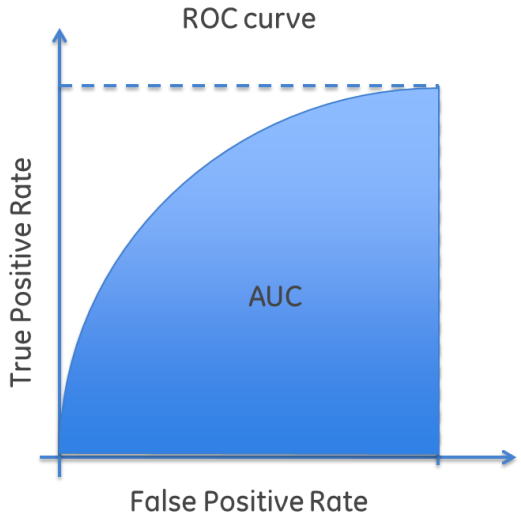

Figure 8 ROC curve example

Accuracy Ratio (AR) in Cumulative Accuracy Profie (CAP) analysis is similar to AUC in ROC. The only difference in CAP analysis is that the horizontal axis represents the total positive rate, the proportion of the sample with fitted values greater than π_0. Each ordinate measures, from the $\hat{\pi}_i \geq \pi_0$ fraction of the population, what fraction of deals were actually prepaid. AR is computed as the area under the CAP curve and above the diagonal divided by the area under the perfect prediction curve and above the diagonal. The AR value is equal to (2*AUC -1) (Engelmann, Hayden and Tasche 2003).

(Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) provides guidance on the use of summary statistics, AUC, as acceptance criteria, as in Table 9. Here we included AR in the table by calculating (2*AUC -1).

Table 9 AR and AUC criteria

AR	AUC	Criteria
> 80%	> 90%	Outstanding
60% - 80%	80% - 90%	Excellent
40% - 60%	70% - 80%	Acceptable
< 40%	< 70%	Poor

2.5 Backtesting

The purpose of out-of-sample backtesting is to assure that the model does not overfit the development data. Overfitting occurs when a statistical model describes random error or noise, instead of the underlying relationship.[44] If the model predicts the out-of-sample data well, the model is not overfit. Backtesting is essential in model assessment, especially when the model is developed to forecast the future. In this section, we discuss numerical statistics to measure the predictive performance, as well as visualization analysis of prepayment speed, defined in Section 1.

2.5.1 Numerical Statistics

This section describes assessment of fit using out-of-sample data, as described in (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013).

- Pearson Chi-Square, Deviance and Hosmer-Lemeshow tests
- Accuracy Ratio
- Mean Square Error (MSE) based Z-statistic

The methods for out-of-sample validation are similar to the assessment of model performance; the major difference is that the values of the coefficients in the model are regarded as fixed, since the coefficient estimates are independent of the out-of-sample data. This assumes that the out-of-sample data was used neither for predictor selection nor for parameter estimation.

The Pearson Chi-Square, Deviance, and Hosmer-Lemeshow tests, and the Accuracy Ratio, are the most common analyses for backtesting; they are applied in the same manner as in Section 2.4.1,

[44] See https://en.wikipedia.org/wiki/Overfitting

except that the training data is used to estimate the model coefficients, while out-of-sample data is used to measure the goodness-of-fit. Because none of these tests are definitive, analysts have considered alternative tests, including the use of MSE based Z-statistics.[45] Mean square error (MSE) is defined as:

$$S = \sum_{j=1}^{G_v} \left(y_j - m_j \pi_j\right)^2$$

where G_v is the number of covariate patterns in the out-of-sample data. In this case, we can construct an approximate z-statistic, which is normally distributed,

$$z_S = \frac{S - \sum_{j=1}^{G_v} m_j \pi_j (1 - \pi_j)}{\sigma_S}$$

where

$$\sigma_S^2 = \sum_{j=1}^{G_v} m_j \pi_j (1 - \pi_j)[1 + 2 m_j \pi_j (1 - \pi_j) - 6\pi_j (1 - \pi_j)]$$

Under the assumption that each $\mathrm{m_j} = 1$, the above equation becomes

$$\sigma_S^2 = \sum_{i=1}^{n_v} \pi_i (1 - \pi_i)(1 - 2\pi_i)^2$$

The following criterion can be applied[46]

Test	Criterion
MSE based Z-statistics	$z_S > 1.645$

2.5.2 Visualization

Another variant of back-testing is to use out-of-time sampling. Use data prior to a certain date for coefficient estimates. Predict the probability of prepayment for data after the date (out-of-time sample). Using those predictions, calculate prepayment speed for each time period (monthly, quarterly, or yearly, etc.). Compare empirical prepayment speed to predicted

$$\underset{\text{prepayment speed(t)}}{\text{empirical}} = \frac{\#\text{ of prepayments(t)}}{\#\text{ of deals(t)}} \quad \text{vs.} \quad \widehat{\text{prepayment speed(t)}} = \frac{\sum_i \widehat{\Pr}(\text{Deal i is prepaid at t})}{\#\text{ of deals at t}}$$

[45] See (Stallard 2009) and (Hosmer, et al. 1997) for comparison of other goodness of fit tests.

[46] Although (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) suggested two tailed test, it is counter-intuitive (we want small MSE as possible) and we could not find a specific reason for using two-tailed test from the literatures. We also found some other literatures that use the test with one sided test (e.g. http://www.stata.com/manuals13/rbrier.pdf and (Stallard 2009)).

t represents the time period. With t on the abscissa, and prepayment speed on the ordinate, plot two curves, one for empirical behavior and one for the forecast. A good model should have predicted prepayment speed following the actual trend reasonably.

2.6 Small Sample and Rare Events

The maximum likelihood estimate is asymptotically unbiased as sample size increases to infinity. In a small sample, the bias of a maximum likelihood estimate may not be negligible. Furthermore, the estimate is biased away from 0; therefore, it should be adjusted, or shrunk, towards 0 (McCullagh and Nelder 1989) (Firth 1993).

A related issue is when there are only a small number of the events of interest, or oppositely, when there are a small number of events of no interest. This is usually referred to as "rare events". For simplicity, we assume the event of interest (prepayment) is rarer than the event of no interest (non-prepayment) in the following context. The concept to be discussed applies to the opposite case too. The bias of the maximum likelihood estimate depends on the number of events. The smaller the number of events, the larger the bias is expected to be.

In small to medium-sized data sets or large-sized data sets with rare events, a situation may occur where the prepayment and non-prepayment observations are perfectly separated by a set of predictors. This situation is called "separation" or "monotone likelihood" (Heinze and Schemper 2002). In case of separation, at least one parameter estimate is infinite. An infinite parameter estimate can be also considered as extremely inaccurate, and is inappropriate for modeling or making inference from.

We propose some alternative estimation methods to address the issues. A significant difference between the resulting estimates and the regular maximum likelihood estimate should raise concern regarding bias or unreliable estimate.

2.6.1 Penalized Likelihood Method

Purpose: The penalized likelihood method is an approach to reducing small-sample bias in maximum likelihood estimate (Firth 1993). It is shown to be a good solution to the problem of separation (Heinze and Schemper 2002). The method is also called the "Firth method", after its author.

Description: Instead of maximizing the likelihood function, the Firth method maximizes the penalized likelihood function

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta})|I(\boldsymbol{\beta})|^{1/2}$$

where $L(\boldsymbol{\beta})$ is the original likelihood function and $|I(\boldsymbol{\beta})|^{1/2}$ is the penalty function (Firth 1993). Here $I(\boldsymbol{\beta})$ is the Fisher information matrix. Using the penalized likelihood, the estimates are calculated by solving the following score equation (Heinze and Schemper 2002):

$$U(\beta_j) + 1/2\,\mathrm{trace}\left(I(\boldsymbol{\beta})^{-1}\{\partial I(\boldsymbol{\beta})/\partial\beta_j\}\right) = 0,$$

where $U(\beta_j) = \partial \log L(\boldsymbol{\beta})/\partial\beta_j$ is the usual score equation for the maximum likelihood estimate, and $j = 0,1,\ldots p$.

For logistic regression, the above score equation is

$$\sum_{i=1}^{n}\left\{Y_i - \pi_i + h_i\left(\frac{1}{2} - \pi_i\right)\right\}X_{ij} = 0,$$

where h_i is the i-th diagonal element of the hat matrix $\mathbf{H}=\widehat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{1/2}$, with $\widehat{\mathbf{W}} = \mathrm{diag}\{\hat{\pi}_i(1-\hat{\pi}_i)\}$, and $j = 0,1,\ldots p$.

A significant difference between the parameter estimate from the Firth method and from the regular maximum likelihood estimation raises concern on the bias of the maximum likelihood estimates. There isn't any strict criterion on whether the difference between the two estimates is statistically significant. Use the comparison as an indication of potential concern; judgment is required.

2.6.2 Bias Correction Method

Purpose: According to (King and Zeng 2001), rare event data can cause bias in two ways. One is biased estimation of the coefficients and the other is bias in the estimated probability of an event (*i.e.*, prepayment). Their method provides a means to correct the bias due to maximum likelihood estimates.

Description: The bias on the estimation of coefficient can be corrected by

$$\mathrm{bias}\left(\widehat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}\boldsymbol{\xi}$$

where $\xi_i = 0.5Q_{ii}(2\hat{\pi}_i - 1)$, Q_{ii} are the diagonal elements of $\mathbf{Q} = \mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'$, and $\widehat{\boldsymbol{W}}$ is an $n \times n$ diagonal matrix wherein the i-th element is $\hat{\pi}_i(1-\hat{\pi}_i)$, where $\hat{\pi}_i$ is the estimated probability of event for the i-th observation. The bias corrected estimate is

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \mathrm{bias}(\hat{\boldsymbol{\beta}})$$

Using the bias corrected estimates above, we can estimate the probability of the event as

$$\tilde{\pi}_i = \Pr\left(Y_i = 1|\tilde{\boldsymbol{\beta}}\right) = \frac{1}{1 + e^{\mathbf{x}_i\tilde{\boldsymbol{\beta}}}}$$

which is preferable to $\hat{\pi}_i = \Pr(Y_i = 1|\hat{\boldsymbol{\beta}})$. Here, $\boldsymbol{x_i}$ is i-th row of $\boldsymbol{X}$ matrix. The reference cautions that this modification is not optimal, because it ignores the uncertainty in $\tilde{\boldsymbol{\beta}}$, which leads to underestimating the rare event. Corrected estimation is achieved by averaging over the uncertainty in $\tilde{\boldsymbol{\beta}}$. To achieve this averaging, one approach uses simulation and another approach uses an analytical approximation, as shown

$$\tilde{\pi}_i + (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1-\tilde{\pi}_i)x_i V(\tilde{\beta})x_i'$$

Here, $V(\tilde{\boldsymbol{\beta}}) = \left(\frac{n}{n+k}\right)^2 V(\hat{\boldsymbol{\beta}})$ is a variance matrix of $\tilde{\boldsymbol{\beta}}$ and $V(\hat{\boldsymbol{\beta}})$ is a variance matrix of $\hat{\boldsymbol{\beta}}$;

$$V(\widehat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{n} \pi_i(1-\pi_i)\boldsymbol{x_i}\boldsymbol{x_i}'\right]^{-1}$$

substituting $\tilde{\pi}_i$ for π_i for the calculation. You can see that if $\tilde{\pi}_i < 0.5$, the correction term is positive and we will have higher estimated probability of the event occurrence.

As with the Firth method, if there is a significant difference between the parameter estimate from the correction and from the regular maximum likelihood estimation, concern should be raised.

3 Reference

Agresti, Alan. *Categorical data analysis.* New Jersey: John Wiley & Sons, 2013.

Allison, Paul D. *Logistic regression using the SAS system: Theory and applications.* Cary, NC: SAS Institute Inc., 2000.

Dempster, A. P., N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data vis the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1977): 1-38.

Engelmann, Bernd, Evelyn Hayden, and Dirk Tasche. "Testing rating accuracy." *Risk* 16 (2003): 82-86.

Firth, David. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika*, March 1993: 27-38.

GECC Treasury Model Risk Management Leader. "Model Risk Management Procedures." 2014.

Heinze, Georg, and Michael Schemper. "A solution to the problem of separation in logistic regression." *Statistics in Medicine*, 2002: 2409-2419.

Hilbe, Joseph M. *Logistic Regression Models.* Boca Raton: Chapman & Hall/CRC Press, 2009.

Hosmer, D.W., T Hosmer, S. Le Cessie, and S. Lemeshow. "A comparison of goodness-of-fit tests for the logistic regression model." *Statistics in medicine* 16 (1997): 965-980.

Hosmer, David W., and Stanley Lemeshow. "A goodness-of-fit test for the multiple logistic regression model." *Communications in Statistics*, 1980: 1043-1069.

—. *Applied logistic regression.* Second Edition. John Wiley & Sons, 2000.

Hosmer, David W., Stanley Lemeshow, and J. Klar. "Goodness-of-fit testing for multiple logistic regression analysis when the estimated probabilities are small." *Biometrical Journal*, 1988: 911-924.

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression.* Third edition. New Jersey: John Wiley & Sons, Inc., 2013.

Jennings, D. E. "Outliers and residual distributions in logistic regression." *Journal of the American statistical association* 81 (1986): 987-990.

King, Gary, and Langche Zeng. "Logistic Regression in Rare Events Data." *Political Analysis* 9 (2001): 137-163.

Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models.* Fifth Edition. New York: McGraw-Hill/Irwin, 2005.

Lemeshow, Stanley, and David W. Hosmer. "A review of goodness-of-fit statistics for use in the development of logistic regression models." *American Journal of Epidemiology*, 1982: 92-106.

Little, Roderick J.A., and Donald B. Rubin. *Statistical Analysis with Missing Data.* Hoboken, New Jersey: John Wiley & Sons, Inc., 2002.

Martin, M., and L. Pardo. "On the asymptotic distribution of Cook's distance in logistic regression models." *Journal of Applied Statistics* 36 (2009): 1119-1146.

McCullagh, P., and J. A. Nelder. *Generalized Linear Models.* Boca Raton: Chapman and Hall/CRC, 1989.

McCullagh, Peter, and John A. Nelder. *Generalized Linear Models.* Second Edition. Boca Raton: Chapman & Hall/CRC, 1989.

Menard, Scott. *Applied logistic regression analysis.* Second Edition. Thousand oaks, CA: Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-106, 2002.

Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. "A simulation study of the number of events per variable in logistic regression analysis." *J.Clin.Epidemiol* 49, no. 12 (1996): 1373-1379.

Rubin, D. B. *Multiple imputation for nonresponse in surveys.* New York: Wiley, 1987.

Rubin, Donald B. "Inference and missing data." *Biometrika* 63 (1976): 581-592.

Schafer, Joseph L., and John W. Graham. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2002): 147-177.

Stallard, Nigel. "Simple tests for the external validation of mortality prediction scores." *Statistics in Medicine* 28 (2009): 377-388.

Vittinghof, E, and C. E. McCulloch. "Relaxing the rule of ten events per variable in logistic and Cox regression." *American Journal of Epidemiology* 165 (2006): 710-718.

Probability of Default

Probability of Default

Sanghee Cho and Jin Xia

Contents

Overview

This chapter describes the key components in validating Probability of Default (PD) models. A PD model provides a probability of default estimate of an obligor and its rating category. The models utilize an obligor's risk factors as well as macro-economic variables if needed. The models are trained on historical data of an obligor and whether it defaulted or not and its related risk factors such as financial information and qualitative evaluation.

A PD model is used in the following applications:

- Underwriting
- Estimating economic capital for finance leases and loans
- Allowance for Loan and Leases Losses

A PD model includes two components:

- Statistical model on default risk of obligors
- Calibration to a GE Obligor Rating

This chapter is organized into three sections: Model Specification, Model Diagnostics, and Calibration.

Model Specification and Estimation

This section describes model specification, estimation methods, and additional aspects that need to be considered when developing a statistical model for probability of default. The following topics are covered:

- Data processing - Missing values
- Model functional forms - Generalized linear model
- Variable selection
- Small sample and rare events

This section focuses on the statistical modeling for probability of default of each obligor that measures the risk of default associating with factors such as financial status, macro-economic variables, and other obligor information. Generalized linear models, especially the logistic regression model, are discussed since they are the most commonly used statistical models in this area. GAM (generalized additive models) and ordinal logistic regression model are also used sometimes in the business, but these models are not covered in this chapter.

Note that since categorical data is being analyzed, we use the notion of "event" in this chapter. When modeling default vs. non-default, we consider them as two different types of events, and default is the event of interest. Given the nature of obligors, we also assume that default is the rare event compared to non-default.

Data Processing

In practice, most data sets need certain levels of processing before reliable models can be fit. The processing includes formatting the data and evaluating the appropriateness of values. Here we

focus on the latter. Common processing steps in evaluating data appropriateness include, but are not limited to, identifying abnormal values and treating missing data. Identifying abnormal values relies primarily on domain knowledge and is not discussed in detail here. In this section, we focus on treating missing data.

Missing Data

Missing data is usually a nuisance rather than a focus in model building. The model validator should evaluate the amount of missing data and whether it raises any concern for estimating and making inferences from the developed model. Two common scenarios when the amount of missing data is likely problematic are:

- The absolute amount of missing data is high.
- The absolute amount of missing data is low, and yet the data is unbalanced and missingness for certain category (-ies) of data is severe.

There is no threshold on whether the amount of missing data is problematic. The model validator should make judgments using experiences on the specific type of problems. If there is uncertainty, the model validator can compare the results with and without treating missing data, and decide whether missing data is an issue and should be treated.

If the amount of missing data raises sufficient concerns, the model validator should further consider the following mechanisms of missing data (Schafer and Graham 2002) (Little and Rubin 2002):

- Missing completely at random (MCAR): The missingness does not depend on missing or observed data.
- Missing at random (MAR): The missingness depends on the observed data but not on the missing data.
- Missing not at random (MNAR): The missingness depends on the missing data.

Let D_{obs} denote the observed data, D_{mis} denote the missing data if it were observed, and D_{com} denote the complete data, i.e. $D_{\text{com}} = (D_{\text{obs}}, D_{\text{mis}})$. Let M denote the missingness. Then MCAR and MAR can be formulated as:

- MCAR: $\Pr(M|D_{\text{com}}) = \Pr(M)$
- MAR: $\Pr(M|D_{\text{com}}) = \Pr(M|D_{\text{obs}})$

The validator should evaluate the data generation process and determine the mechanism of missingness.[47]

In literature (Little and Rubin 2002) (Schafer and Graham 2002) (D. B. Rubin 1976), the following ways are suggested to deal with the above mechanisms of missingness:

- MCAR: In general, the missing data can be dropped and the model can be built on the observed data. A commonly applied method is to model the complete cases, i.e. cases without missing values. The biggest advantage of complete-case analysis is simplicity. The disadvantage is loss of precision (and bias if the missingness is not MCAR). Under MCAR, complete-case analysis is justified when the loss of precision is minimal.

[47] Note that there is generally no way to test whether MCAR or MAR holds, given the missing data is not observed.

- MAR: Under MAR, dropping the missing data still generates appropriate likelihood for the parameters (D. B. Rubin 1976). Any inference from the likelihood is valid. For example, the commonly used maximum likelihood estimates (MLE) are appropriate, as well as Bayesian methods. A general method for MLE in the presence of missing data is the EM algorithm (Dempster, Laird and Rubin 1977). Besides likelihood-based methods, another popular method of dealing with MAR is multiple imputation (MI). MI refers to the procedure where each missing value is replaced by a number of simulated values. As a result, a number of complete data sets are generated, which represent the imputation uncertainty. Methods for complete data are applied to each data set, and the results are combined (Schafer and Graham 2002) (D. B. Rubin 1987).
- MNAR: Under MNAR, an appropriate model for the missingness needs to be specified, and a large sample is needed. This is often infeasible in practice. Therefore, dealing with MNAR is usually very challenging. Here, we don't suggest any specific technique to deal with MNAR due to the complexity of the problem. Instead, we suggest for the model validator to be extremely cautious with a model developed with a MNAR data set. If the validator suspects that the departure from MAR is not severe, the validator can apply the above methods for MAR and compare the results. If the departure is apparently severe and no appropriate treatment is applied to the missing values by the model developer, the validator should challenge the feasibility of the model.

In practice, the departure from MCAR or MAR may not be severe enough to generate a significant impact on the model estimates, especially when the amount of missing data is low. In such cases, the bias from assuming MCAR or MAR may be negligible. The most commonly applied method by model developers is the complete-case analysis. As described above, complete-case analysis is only appropriate under MCAR. However, if the departure from MCAR is not severe, the bias in model estimates from assuming MCAR is probably negligible. If there is concern about assuming MCAR, the validator can apply appropriate methods as described above to treat the missing values, and compare results in order to decide whether assuming MCAR is appropriate.

Model Functional Forms

Generalized linear model[48] is widely used to predict probability of default. It uses a link function to model the probability of an event for a binary response variable. There are three commonly used link functions for the binary response: logit, probit, and complementary log-log links. This chapter describes test criteria based on the logit link function, also called the logistic regression, since it is most commonly used and studied. Most of the model evaluation tools in Section 0 and 0 are applicable for the other link functions as well. Applicability will be specified for each criterion.

As shown in Figure 1, the fitted line of logistic regression has an "S"-type of curve (logit curve) that maps predictors to a value between zero and one which enables us to predict a probability of an event given the relevant predictors. Other link functions have similar "S"-type of curves.

[48] Refer to (McCullagh and Nelder 1989) for detailed information about the generalized linear models.

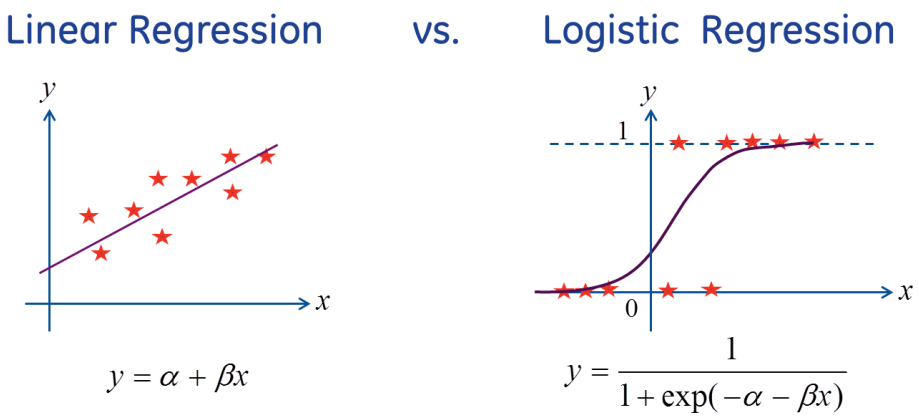

Figure 9 Difference between linear regression model vs. Logistic regression model

Let

- Y denote the binary response variable, with

$$Y = \begin{cases} 1 & \text{if defaulted} \\ 0 & \text{otherwise} \end{cases};$$

- $X_1, X_2, \ldots, X_p$ denote the p predictors;
- π denote the probability of default given $X_1, X_2, \ldots, X_p$; *i.e.*, $\pi = \Pr(Y = 1 | X_1, X_2, \ldots, X_p)$; and
- $\beta_0, \beta_1, \ldots, \beta_p$ denote the coefficients for the predictors.

Then the generalized linear model can be expressed as

$$\eta(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where $\eta(\pi)$ is called the link function. For logistic regression, $\eta(\pi)$ is defined as

$$\eta(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

Note that the probability, after logit transformation, has a linear relationship with the predictors of interest. Other alternative link functions are described in Table 10.

Table 10 Functional forms of link functions. Here, $\Phi^{-1}(.)$ is an inverse cumulative distribution function of standard normal distribution.

Link function	$\eta(\pi)$
Logit	$\log\left(\frac{\pi}{1-\pi}\right)$
Probit	$\Phi^{-1}(\pi)$
Complementary log-log	$\log(-\log(1-\pi))$

The model can also be presented in a matrix form. Let

- $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$ denote the vector of n observed response variables;

- $\mathbf{X}$ denote the design matrix, with

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix};$$

- $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_n)'$ denote the vector of default probability given $\mathbf{X}$; and
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ denote the coefficients for the predictors.

Then the model can be expressed as

$$\eta(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta},$$

where the link function is applied element-wise to $\boldsymbol{\pi}$.

Coefficients can be estimated by maximizing the log-likelihood (LL)

$$LL = \sum_{i=1}^{n} (1 - Y_i) \log(1 - \pi_i) + Y_i \log(\pi_i)$$

An important issue to always check is whether there are repeated observations at different levels of the predictors. Here the levels are unique values of the p-dimensional vector of predictors. The number of levels and numbers of repeated observations determine whether the asymptotic properties of certain tests are valid, as explained below. The levels are referred to as **covariate patterns** in the following context. The repeated observations are also called grouped observations. Let

- J denote the number of covariate patterns, with $J \leq n$;
- $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_J$ denote the covariate patterns, with $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \dots, x_{jp})'$, $j = 1,2, \dots, J$;
- m_j be the number of observations for covariate pattern $\boldsymbol{x}_j$, $j = 1,2, \dots, J$, with $\sum_{j=1}^{J} m_j = n$;
- y_j be the number of events, *i.e.*, default, for covariate pattern $\boldsymbol{x}_j$, $j = 1,2, \dots, J$; and
- $\hat{\pi}_j$ be the estimated probability of an event for covariate pattern $\boldsymbol{x}_j$, $j = 1,2, \dots, J$.

In addition, a definition of **hat matrix** is introduced since it is used multiple times throughout the chapter. The name "hat matrix" came from the linear model where it projects the data to the fitted values. For generalized linear model, a corresponding relationship holds for $\eta(\pi)$ (Agresti 2013). The equation of hat matrix for generalized linear model is defined as

$$\mathbf{H} = \widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{\frac{1}{2}}$$

where $\widehat{\mathbf{W}}$ is the diagonal matrix with elements $\left(\frac{\partial \pi_i}{\partial \eta_i}\right)^2 / Var(Y_i)$ with replacing π_i with $\hat{\pi}_i$. For example, the element of $\widehat{\mathbf{W}}$ for logistic regression is $\pi_i(1 - \pi_i)$ (or $m_j\pi_j(1 - \pi_j)$ for the grouped observations). Table 11 describes the functional form of the element of $\widehat{\mathbf{W}}$ for each link functions.

Table 11 the element of $\widehat{W}$ for different link functions

Link function	W_i in ungrouped case	W_i in grouped case
Logit link	$\pi_i(1-\pi_i)$	$m_j\pi_j(1-\pi_j)$
Probit link	$\frac{\left[\phi\left(\Phi^{-1}(\pi_i)\right)\right]^2}{\pi_i(1-\pi_i)}$	$\frac{m_j\left[\phi\left(\Phi^{-1}(\pi_j)\right)\right]^2}{\pi_j(1-\pi_j)}$
Complementary log-log link	$\frac{(1-\pi_i)[\log(1-\pi_i)]^2}{\pi_i}$	$\frac{m_j(1-\pi_j)[\log(1-\pi_j)]^2}{\pi_j}$

Variable Selection

Purpose: A rigorous process of variable selection is needed and the model developers need to provide reasonable rationale when adding or dropping candidate variables (predictors). However, there is no exact science for this procedure and it should be always combined with statistical soundness and business intuition. For example, (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) provides step by step procedures. One doesn't have to follow the exact steps, but a logical process for variable selection must prevail. Here, plausible reasons to add or to drop variables are listed.

Univariate Variable Selection

Various univariate tests can be done to examine an association between each predictor and the response variable. Hosmer, *et al.* (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) suggest to use a higher p-value cut-off (*e.g.*, 0.2 or 0.25, instead of the traditional cut-off 0.05) when we use univariate association for initial screening. If the experts believe a variable is important to the model, then that variable, even with a high p-value, should be included for further analysis.

Multivariate Variable Selection

This section describes methods that select variables in multivariate ways, meaning that it considers multiple variables simultaneously to find a subset of variables to include in the final model. Three options are discussed with a brief description below;

- Information Criteria: It is useful when comparing candidate models. Each candidate model includes some subset of variables.
- LASSO: It is an estimation method that you can include all the candidate variables in a model. This method will estimate some of the coefficients as zero so that we can exclude those variables from the final model.
- Automatic selection: It is an algorithm with multiple steps which adds or drops the candidate variables in each step.

Information Criteria

When comparing candidate models, AIC and BIC are effective tools to assess the relative model performance. AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) provide summary statistics for model fit comparison. Generally, a model with lower information criteria is preferred. Hilbe (2009) provides guidance (Table 7) for selecting a preferred candidate model, although this guidance is for logistic regression model. AIC and BIC can be computed as follows:

$$AIC = -2LL + 2k$$
$$BIC = -2LL + k\log(n)$$

LL is the log-likelihood function and k is the number of parameters (usually the number of predictors + 1), representing the complexity of a model. Note that BIC penalizes heavier for the number of predictors.

Table 12 Guidance on AIC /BIC criteria

AIC (Akaike Information Criteria)		BIC (Bayesian Information Criteria)	
Difference between Models A and B (Suppose A<B)	Result if A < B	Difference between Models A and B (Suppose A<B)	Degree of Preference on A
0 < (B – A) < 2.5	No difference in models	0 < (B – A) < 2	Weak
2.5 < (B – A) < 6	Prefer A if n>256	2 < (B – A) < 8	Positive
6 < (B – A) < 9.9	Prefer A if n>64	8 < (B – A) < 10	Strong
(B – A) > 10	Prefer A	(B – A) > 10	Very Strong

LASSO (Least Absolute Shrinkage and Selection Operator)

The LASSO (Least Absolute Shrinkage and Selection Operator) is an estimation method for the coefficients, originally applied to ordinary least squares regression. It can also be used as a variable selection method since it shrinks the coefficient estimates toward zero and some coefficient will be shrunk to zero so that we can exclude those variables from the model.

LASSO maximizes the penalized likelihood function

$$LL_{LASSO}(\beta) = LL(\beta) - \lambda \sum_j |\beta_j|$$

where $LL(\beta)$ is the original likelihood function, and $\lambda \geq 0$ is a smoothing parameter. Increasing λ results in greater shrinkage toward 0. The k-fold validation can be used for selecting λ. For each candidate value of λ, prediction error by k-fold validation is measured, and select λ that achieves the minimum. A prediction error is measured as following: divide the data into k subsamples with equal size; each time, exclude 1 subsample to estimate the coefficient and do a prediction with the excluded sample; after doing this with all k subsamples, we have prediction for the whole sample, and prediction error can be measured (Agresti 2013).

Automatic Variable Selection (Stepwise/Backwards/Forwards)[49]

The main idea of the automatic variable selection is to measure how each variable improves the model fit when it is added or deleted. However, decisions are based on statistical evidence alone, so careful assessment is needed. For example, consider two variables A and B with similar significance level. It is possible that B is selected over A by the algorithm, based on statistical evidence, even though the experts believe A is more important than B. Also, statistical significance for a particular variable depends on which variables are already included in the model. Automatic variable selection is not without peril.

[49] Forward selection adds one variable at each step, while backwards selection (elimination) starts with a full model and deletes one variable at a time. Stepwise selection is combination of both forwards and backwards selection. It allows variables to be added or deleted at each step.

Miscellaneous

There are additional considerations in variable selection:

- Multi-collinearity: Including highly correlated predictors in the model can lead to biased estimates with large standard error. Variance Inflation Factor (VIF) in linear regression indicates associations among the predictors (Menard 2002), (Allison 2000). VIF greater than 5 is cause for concern and greater than 10 suggests a serious collinearity problem (Menard 2002).

- Number of predictors in the final model: (Agresti 2013) provides a guideline for logistic regression based on a Monte Carlo study (Peduzzi, et al. 1996); when the number of defaults divided by the number of predictors is smaller than 10, then parameter estimates can be a biased, in addition to other potential issues. (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) cited (Vittinghof and McCulloch 2006); based on extensive simulations, the latter authors conclude that the "rule of 10" may be too conservative. However, (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) take issue with the latter authors' recommendation in cases where the distributions of the discrete predictors are highly skewed rather than balanced; for such cases, the rule of 10 is justified. Thus, the rule of 10 is recommended as a guideline, but acknowledge that this should not be a strict rule since the above studies are based on simulation based study which does not cover all possible settings or data. Thus, a less stringent requirement may suffice.
- Counterintuitive coefficient sign: With limited data, it is often useful to incorporate expert knowledge.

Small Sample and Rare Events

The maximum likelihood estimate is asymptotically unbiased as sample size increases to infinity. In a small sample, the bias of a maximum likelihood estimate may not be negligible. Furthermore, the estimate is biased away from 0; therefore, it should be adjusted, or shrunk, towards 0 (McCullagh and Nelder 1989) (Firth 1993).

A related issue is when there are only a small number of observations of a certain event. This is usually referred to as a "rare events" issue. The bias of the maximum likelihood estimate depends on the number of events. The smaller the number of events, the larger the bias is expected to be.

In small to medium-sized data sets or large-sized data sets with rare events, a situation may occur where the events are perfectly separated by a set of predictors. This situation is called "separation" or "monotone likelihood" (Heinze and Schemper 2002). In case of separation, at least one parameter estimate is infinite. An infinite parameter estimate can be also considered as extremely inaccurate, having infinite bias, and is inappropriate for modeling or making inference from.

The bias on the estimation of coefficient can be approximated by

$$\text{bias}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{W}}\boldsymbol{\xi}$$

where, for logistic regression, $\xi_i = 0.5Q_{ii}(2\hat{\pi}_i - 1)$, Q_{ii} are the diagonal elements of $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'$, and $\widehat{\mathbf{W}}$ is an $\mathrm{n} \times \mathrm{n}$ diagonal matrix wherein the i-th element is $\hat{\pi}_i(1 - \hat{\pi}_i)$, where $\hat{\pi}_\mathrm{i}$ is the estimated probability of event for the i-th observation. (McCullagh and Nelder 1989) provides the formulas for other link functions as well. For other link functions, use the definition of ξ_i as in Table 13 and $\widehat{\mathbf{W}}$ as in Table 11, correspondingly.

Table 13 Functional forms of ξ_i for other link functions. Here, $\eta_i = x_i\hat{\beta}$ where x_i is i-th row of X matrix, and Q_{ii} are the diagonal elements of $\mathrm{Q} = \mathrm{X}(\mathrm{X}'\widehat{\mathrm{W}}\mathrm{X})^{-1}\mathrm{X}'$, and $\widehat{\mathrm{W}}$ is an $\mathrm{n} \times \mathrm{n}$ diagonal matrix wherein the i-th element is as in Table 11.

Link function	ξ_i
Logit	$0.5Q_{ii}(2\hat{\pi}_i - 1)$
Probit	$Q_{ii}\eta_i/2$
Complementary log-log	$Q_{ii}(\exp(\eta_i) - 1)/2$

We propose some alternative estimation methods to address the issue of bias. A significant difference between the resulting estimates and the regular maximum likelihood estimate should raise concern regarding bias or unreliable estimate.

Penalized Likelihood Method

Purpose: The penalized likelihood method is an approach to reducing small-sample bias in maximum likelihood estimate (Firth 1993). It is shown to be a good solution to the problem of separation (Heinze and Schemper 2002). The method is also called the "Firth method", after its author.

Description: Instead of maximizing the likelihood function, the Firth method maximizes the penalized likelihood function

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta})|I(\boldsymbol{\beta})|^{1/2}$$

where $L(\boldsymbol{\beta})$ is the original likelihood function, and $|I(\boldsymbol{\beta})|^{1/2}$ is the penalty function (Firth 1993). Here $I(\boldsymbol{\beta})$ is the Fisher information matrix. Using the penalized likelihood, the estimates are calculated by solving the following score equation (Heinze and Schemper 2002):

$$U(\beta_j) + 1/2\,\mathrm{trace}\left(I(\boldsymbol{\beta})^{-1}\{\partial I(\boldsymbol{\beta})/\partial\beta_j\}\right) = 0,$$

where $U(\beta_j) = \partial \log L(\boldsymbol{\beta})/\partial\beta_j$ is the usual score equation for the maximum likelihood estimate, and $j = 0,1, \ldots p$.

In case of logistic regression, the above score equation is

$$\sum_{i=1}^{n}\left\{Y_i - \pi_i + h_i\left(\frac{1}{2} - \pi_i\right)\right\}X_{ij} = 0,$$

where h_i is the i-th diagonal element of the hat matrix $\mathbf{H}=\widehat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{1/2}$, with $\widehat{\mathbf{W}} = \mathrm{diag}\{\hat{\pi}_i(1 - \hat{\pi}_i)\}$, and $j = 0,1, \ldots p$.

A significant difference between the parameter estimate from the Firth method and from the regular maximum likelihood estimation raises concern on the bias of the maximum likelihood estimates. There isn't any strict criterion on whether the difference between the two estimates is statistically significant. Use the comparison as an indication of potential concern; judgment is required.

Bias Correction Method by (King and Zeng, 2001)

Purpose: According to (King and Zeng 2001), rare event data can cause bias in two ways. One is biased estimation of the coefficients, and the other is bias in the estimated probability of an event (*i.e.*, default). Their method provides a means to correct the bias due to maximum likelihood estimates.

Description: The bias corrected estimate is

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \text{bias}(\hat{\boldsymbol{\beta}})$$

Using the bias corrected estimates above, we can estimate the probability of the event as

$$\tilde{\pi}_i = \Pr(Y_i = 1|\tilde{\boldsymbol{\beta}}) = \frac{1}{1 + e^{\mathbf{x}_i\tilde{\boldsymbol{\beta}}}}$$

which is preferable to $\hat{\pi}_i = \Pr(Y_i = 1|\hat{\boldsymbol{\beta}})$. Here, $\boldsymbol{x_i}$ is i-th row of $\boldsymbol{X}$ matrix. The reference cautions that this modification is not optimal, because it ignores the uncertainty in $\tilde{\boldsymbol{\beta}}$, which leads to underestimating the rare event. Corrected estimation is achieved by averaging over the uncertainty in $\tilde{\boldsymbol{\beta}}$. To achieve this averaging, one approach uses simulation, and another approach uses an analytical approximation. The paper provides an analytical approximation only for logistic regression as shown

$$\tilde{\pi}_i + (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)x_i V(\tilde{\beta})x_i'$$

Here, $V(\tilde{\boldsymbol{\beta}}) = \left(\frac{n}{n+k}\right)^2 V(\hat{\boldsymbol{\beta}})$ is a variance matrix of $\tilde{\boldsymbol{\beta}}$, and $V(\hat{\boldsymbol{\beta}})$ is a variance matrix of $\hat{\boldsymbol{\beta}}$;

$$V(\hat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{n} \pi_i(1 - \pi_i)\boldsymbol{x_i}\boldsymbol{x_i}'\right]^{-1}$$

substituting $\tilde{\pi}_i$ for π_i for the calculation. You can see that if $\tilde{\pi}_i < 0.5$, the correction term is positive, and we will have higher estimated probability of the event occurrence.

As with the Firth method, if there is a significant difference between the parameter estimate from the correction and from the regular maximum likelihood estimation, concern should be raised.

Model Evaluation

This section describes two aspects of the model evaluation. One is model diagnostics that assess the validity of fitting a model numerically and visually. The other is to evaluate rank ordering power (discriminatory power).

Model Diagnostics

This section describes methods that assess the validity of fitting a generalized linear model to the data. A fitted model may be inadequate in the following ways:

- Relationships between the link $\eta(\pi)$ and predictors $X_1, X_2, \ldots, X_p$ are not all correctly specified. There are generally two causes of the misspecification: (1) the predictors need to be transformed in order to have linear relationships with the link; and (2) the specified link is incorrect.
- There are outliers or other strongly influential observations. They may bias the model estimates, and they are not fitted well by the model.

Three types of tools are proposed to assess the above model inadequacies. Table 8 shows which inadequacy is assessed by each tool. The tools are:

- Link specification tests
- Residual plots
- Influential observation plots

Table 14 Model Inadequacies and Diagnostic Tools

	Diagnostic Tools		
Model Inadequacies	Link specification tests	Residual plots	Influential observation plots
Misspecification of relationships between the link logit(π) and predictors $X_1, X_2, \ldots, X_p$.	X	X	
Existence of outliers or strongly influential observations		X	X

These tools enable one to test the validity of model assumptions. They also enable one to conduct detailed analyses of the model in order to identify issues or improvement opportunities.

Link Specification Tests

Purpose: The generalized linear model assumes linear relationships between $\eta(\pi)$ and the predictors $X_1, X_2, \ldots, X_p$, as specified by the model equation

$$\eta(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

If the linearity assumption is inappropriate, then the coefficient estimates and standard errors are biased. Therefore, it is important to assure the validity of the linearity assumption.

Two similar tests are provided to assess the linearity assumption (Hilbe 2009) for logistic regression, namely

- Box-Tidwell test
- Pregibon Link test

We need further analysis to check if above link specification tests applies to other link functions.

3) Box-Tidwell Test

Description: The Box-Tidwell test has two steps:

- Construct the interaction[50] of each continuous predictor with its log transformation
- Fit a logistic regression with the original predictors and the above interactions

If any interaction is statistically significant, the Box-Tidwell test concludes that the linearity assumption is violated. The following criterion can be applied on each interaction.

Method	Criterion
Box-Tidwell Test	P-value < 0.05

4) Pregibon Link Test

Description: The Pregibon Link test is similar to the Box-Tidwell test, but uses the square of the hat matrix diagonal instead of the interaction term. It is carried out in two steps:

- Construct the square of the hat matrix diagonal of each continuous predictor
- Fit a logistic regression with the original predictors and the above new terms

If any new term is statistically significant, the Pregibon Link test indicates that the linearity assumption is violated.

The estimated hat matrix for logistic regression is

$$\mathbf{H}=\widehat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{1/2}.$$

Here, $\widehat{\mathbf{W}}$ is an $n \times n$ diagonal matrix

$$\widehat{\mathbf{W}} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix},$$

where $\hat{\pi}_i$ is the estimated probability of event for the i-th observation.

The following criterion can be applied to each interaction.

Method	Criterion
Pregibon Link Test	P-value < 0.05

Residual Plots

This section and the next section on influential observation plot describe visual diagnostics using several normalized residuals. Visual assessment is preferred; hard thresholds are not available for these diagnostics. Visual assessment informs with respect to model adequacy and overly influential

[50] An interaction term represents the interactive effect of two or more predictors on the response. It is generally constructed as the product of these predictors. For example, let X be a continuous predictor, and the interaction of X with its log transformation is $X\log(X)$.

(to the model fit) observations. Where poor fit is observed, one would check for data issues or opportunity to improve the model by alternative variable selection or transformation of variables.

Purpose: The purpose of these residual plots is to assess model adequacy. An adequate model satisfies $E\{Y_i\} = \pi_i$. If such assumption is satisfied, it follows asymptotically that $E\{Y_i - \hat{\pi}_i\} = 0$ (Kutner, et al. 2005).

Description: There are three examples of residuals. Residual plots versus predictors or/and predicted probabilities with a Lowess curve can suffice (Kutner, et al. 2005). Each residual plot presents different information about the model fit; multiple residual plots are preferred. The Lowess curve of each plot is expected to be flat horizontal line with zero intercept. If significant deviations are found, influential data points or lurking variables left out of the model are typically causal. Our notation assumes the grouped case (repeated covariate patterns), but even for ungrouped data (without repeated covariate patterns), we can treat $J = n$ and $m_j = 1$ for all $j = 1, \dots, J$. However, (Agresti 2013) and (Hosmer and Lemeshow, Applied logistic regression 2000) recommend to compute residuals using grouped data if possible, especially when J is much smaller than n. All the following residuals can be applied for other link functions, while the calculation of hat matrix for studentized Pearson Residuals varies depending on the link function that is used.

d) Pearson Residuals: The objective is to make the residuals more comparable by dividing the raw residual by an estimate of the standard deviation of the observation y_i.

$$r_j = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

e) Studentized Pearson Residuals: The objective is to fully standardize the residuals; the raw residual $(y_j - m_j\hat{\pi}_j)$ is normalized by the estimated standard error of the fitted value $\hat{\pi}_i$.

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}} = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)(1-h_j)}}$$

where, h_j is the j-th diagonal element of the estimated hat matrix $\mathbf{H}$:

$$\mathbf{H} = \widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{\frac{1}{2}}$$

with $\widehat{\mathrm{W}}$ for logistic regression is the diagonal matrix wherein the j-th element is $m_j\hat{\pi}_j(1-\hat{\pi}_j)$. See Table 11 for functional form of $\widehat{\mathrm{W}}$ corresponding to each link function.

f) Deviance Residuals: These residuals are related to the deviance test which will be described in Section 0.

$$dev_j = sign(Y_j - m_j\hat{\pi}_j)\sqrt{2[y_j \ln\frac{y_j}{m_j\hat{\pi}_j} + (m_j - y_j)\ln\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}]}$$

Influential Observation Plots

Purpose: The purpose of these plots is to visualize how each observation influences the model fit. The idea is to see how goodness-of-fit test statistics, or the coefficient estimates, would change if we exclude one observation from the analysis.

Description: Identify observations that stand out in the plots, then check those observations for errors. As residual plots, each plot presents different information about the model fit; multiple plots are preferred.

d) ΔX_j^2 vs. $\hat{\pi}_j$: The ordinate is related to the Pearson Chi-Square statistic.

$$\Delta X_j^2 = \frac{r_j^2}{(1-h_j)} = r_{sj}^2$$

e) ΔD_j vs. $\hat{\pi}_j$: The ordinate is related to the deviance test.

$$\Delta D_j = dev_j^2 + \frac{r_j^2 h_j}{(1-h_j)} \approx \frac{dev_j^2}{(1-h_j)}$$

f) $\Delta\hat{\beta}_j$ vs. $\hat{\pi}_j$: The ordinate relates to the observation's influence on the estimate of coefficient β.

$$\Delta\hat{\beta}_j = \frac{r_j^2 h_j}{(1-h_j)^2}$$

(Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) recommend focusing on observations whose values for one or more of the diagnostic statistics fall well away from the rest of the values, instead of setting a threshold value to identify influential points. The following three plots are considered to be critical. Bubble plots, which plot a residual, with the size of the symbol proportional to another characteristic, can be a useful, as they introduce a third dimension to the graphical analysis.

Evaluating Rank Ordering Power

This section describes methods to assess the overall goodness-of-fit of the model, specifically rank ordering power. The accuracy of the prediction for probability of default will be discussed in Section 0. The overall goodness-of-fit is robust to poor fit for a few observations. When the overall goodness-of-fit is rejected, the appropriate response is to review model diagnostics for improvement opportunities.

AUC and ROC Curve / AR and CAP

Purpose: The purpose of ROC (Receiver Operator Characteristic) curve for generalized linear model is to measure the model performance from the classification (event or no event) perspective. Through varying the classification rule, we can impact the model's capability to predict default. The ROC curve is useful for visualizing the model's predictive power, especially rank ordering power. The analysis also provides a quantitative summary statistic called AUC (Area Under the Curve).

Description: Suppose, for each observation i, we have a fitted value $\hat{\pi}_i$ from the model. We can classify whether the obligor will be default or not by setting a rule as below, denoting Y = 1 for default,

$$\hat{Y}_i = \begin{cases} 1 & \hat{\pi}_i \geq \pi_0 \\ 0 & \hat{\pi}_i < \pi_0 \end{cases}$$

For a given cutoff point π_0, each observations will be classified as either 0 or 1. Then, it can be compared to the true observed event Y_i to check if the prediction rule works well. By varying the cutoff point π_0, we measure the predictive power by false positive rate and true positive rate using the whole sample. The ROC curve is constructed by plotting each associated pair of false positive rate (horizontal axis) versus true positive rate (vertical axis), resulting from sweeping the cutoff point π_0 through the entire range, as in Figure 8. If the curve passes close to the top-left corner of the plot, it means the model has a high discriminatory power. On the other hand, if the curve never deviates much form to the diagonal (45° line) of the plot, the model's rank ordering power is not better than a random guess. We can quantify this curve by calculating AUC. High AUC indicates high discriminatory power.

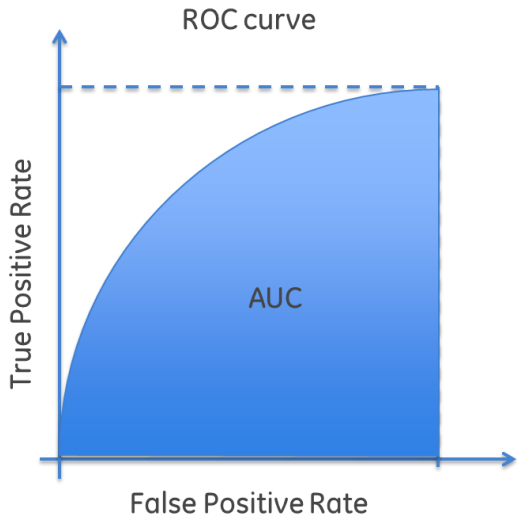

Figure 10 ROC curve example

Accuracy Ratio (AR) in Cumulative Accuracy Profile (CAP) analysis is similar to AUC in ROC. The only difference in CAP analysis is that the horizontal axis represents the total positive rate, the proportion of the sample with fitted values greater than π_0. Each ordinate measures, from the $\hat{\pi}_i \geq \pi_0$ fraction of the population, what fraction of observations was actually default. AR is computed as the area

under the CAP curve and above the diagonal divided by the area under the perfect prediction curve and above the diagonal. The AR value is equal to (2*AUC -1) (Engelmann, Hayden and Tasche 2003).

(Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013) provides guidance on the use of summary statistics, AUC, as acceptance criteria, as in Table 9. Here we included AR in the table by calculating (2*AUC -1).

Table 15 AR and AUC criteria

AR	AUC	Criteria
> 80%	> 90%	Outstanding
60% - 80%	80% - 90%	Excellent
40% - 60%	70% - 80%	Acceptable
< 40%	< 70%	Poor

Calibration

Evaluating Model Accuracy

The calibration process produces pooled PDs based on certain optimization criterion. This section focuses on evaluating the accuracy of pooled PDs, estimated vs. observed. The following methods are provided:

- Pearson Chi-Square test: Overall comparison of pooled PDs
- Binomial test: Comparison of pooled PD for each rating category
- Backtesting (Visualization): Evaluating the model accuracy using out-of-sample

Pearson Chi-Square

Purpose: The Pearson Chi-Square test conducts an overall evaluation of all the pooled PDs. The Pearson Chi-Square test is a classic method and is appropriate when the number of rating categories is not large.

Description: The Pearson Chi-Square test evaluates whether the expected frequencies of a multinomial distribution is equal to the observed frequencies. It assumes that the observations are independent, and the sample size is sufficiently large so that the distribution of the test statistic can be approximated by an asymptotic distribution. It can detect major departures from the model, but is not sensitive to small departures.

The Pearson Chi-Square statistic is the sum of squares of Pearson residuals. For binary response, let

- K denote the number of pooled PDs, i.e. rating categories;
- m_j denote the number of observations for the j-th rating category, $j = 1,2, \dots, K$;
- y_j denote the number of observed events for the j-th rating category, $j = 1,2, \dots, K$; and
- $\hat{\pi}_j$ denote the target PD for the j-th rating categories, $j = 1,2, \dots, K$.

Then the Pearson Chi-Square statistic is defined as

$$X^2 = \sum_{j=1}^{K} \left[\frac{y_j - m_j \hat{\pi}_j}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \right]^2 .$$

Under the null hypothesis that the target PDs are correct representation of the data, X^2 follows a χ^2 distribution asymptotically with degrees-of-freedom $K - 1$, denoted as χ^2_{K-1}. The p-value is computed as

$$\Pr(\chi^2_{K-1} \geq X^2).$$

The following criterion can be applied to reject the null hypothesis.

Method	Criterion
Pearson Chi-Square Test	P-value < 0.05

The asymptotic χ^2 distribution is valid when K is finite as $n \to \infty$. Therefore, when the number of rating categories is small, the Pearson Chi-Square test is appropriate.

Binomial test[51]

Binomial test is applied to each rating category at a time and examine whether the empirical default rate of the obligors is statistically not far from the target PD, which is provided in the internal rating systems. Let's denote the target PD of a certain rating category as PD

The hypothesis for the test is

H_0 (null hypothesis): the PD of a rating category is correct ($p = PD$)

H_1 (alternative hypothesis): the PD of a rating category is underestimated ($p < PD$)

Given a confidence level q (e.g. 95% or 99%) the null hypothesis is rejected if the number of defaulters k in the rating category is greater than or equal to a critical value k* which is defined as

$$k^* = \min\{k: \sum_{i=k}^{n} \binom{n}{i} PD^i (1 - PD)^{n-i} \leq 1 - q\}$$

where n is the number of obligors in the rating category. Also, there is a normal distribution approximation of k^* using central limit theorem,

$$k^* = \Phi^{-1}(q)\sqrt{nPD(1 - PD)} + nPD$$

where Φ^{-1} is the inverse cumulative distribution function of standard normal distribution. Rule of thumb to use normal approximation of binomial distribution is to check if $np > 5$ and $n(1 - p) > 5$ hold.

Method	Criterion

[51] Basel Committee on Banking Supervision (BCBS) (2005), Studies on the validation of internal rating systems, May

Binomial Test	Number of defaulters $\geq k^*$

Backtesting

The purpose of out-of-sample backtesting is to assure that the model does not overfit the development data. Overfitting occurs when a statistical model describes random error or noise, instead of the underlying relationship.[52] If the model predicts the out-of-sample data well, the model is not overfit. Backtesting is essential in model assessment, especially when the model is developed to forecast the future. In this section, we discuss numerical statistics to measure the predictive performance, as well as visualization analysis.

Numerical Statistics

This section describes assessment of fit using out-of-sample data, as described in (Hosmer, Lemeshow and Sturdivant, Applied Logistic Regression 2013).

- Accuracy Ratio
- Pearson Chi-Square test
- Binomial test

The methods for out-of-sample validation are similar to the assessment of model performance; the major difference is that the values of the coefficients in the model are regarded as fixed, since the coefficient estimates are independent of the out-of-sample data. This assumes that the out-of-sample data was used neither for predictor selection nor for parameter estimation.

The Accuracy Ratio, the Pearson Chi-Square test and the binomial test, are the most common analyses for backtesting; they are applied in the same manner as in Section 0, 0, and 0, except that the training data is used to estimate the model coefficients, while out-of-sample data is used to measure the goodness-of-fit.

Visualization

Another variant of back-testing is to use visualization. We can compare empirical default rate to predicted visually, by year or rating,

$$\underset{\text{default rate(t)}}{\text{empirical}} = \frac{\#\text{ of defaults(t)}}{\#\text{ of obligors(t)}} \quad \text{vs.} \quad \widehat{\text{default rate(t)}} = \frac{\sum_i \widehat{\Pr}(\text{ Obligor i is default at t})}{\#\text{ of obligors at t}}$$

Here, t can be each time period or each rating bucket. With t on the abscissa, and default rate on the ordinate, plot two curves, one for empirical behavior and one for the forecast. A good model should have predicted probability of default following the actual trend reasonably.

Alternative calibration methods

Alternative calibration methodologies can be useful tools to examine the calibration process. The followings are the example of those alternatives:

- QMM method: Dirk Tasche, (2012), The Art of PD Curve Calibration, working paper, Financial Services Authority.
- VDB method: M. van der Burgt, (2008), Calibrating Low-Default Portfolios Using the Cumulative Accuracy Profile, Journal of Risk Model Validation, 1(4):17-33.

[52] See https://en.wikipedia.org/wiki/Overfitting

Confusion matrix is a useful tool to compare the outcomes from each method. If most obligors concentrated around the diagonal cells, then it indicates that both calibration processes agrees the rating of most obligors. However, when the obligors are spread out off-diagonal, it indicates that the two calibration methods don't agree. In this case, further exploration is needed such as comparing empirical default rate to predicted using out-of-sample.

References

Agresti, Alan. *Categorical data analysis.* New Jersey: John Wiley & Sons, 2013.

Allison, Paul D. *Logistic regression using the SAS system: Theory and applications.* Cary, NC: SAS Institute Inc., 2000.

Dempster, A. P., N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data vis the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1977): 1-38.

Engelmann, Bernd, Evelyn Hayden, and Dirk Tasche. "Testing rating accuracy." *Risk* 16 (2003): 82-86.

Firth, David. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika*, March 1993: 27-38.

GECC Treasury Model Risk Management Leader. "Model Risk Management Procedures." 2014.

Heinze, Georg, and Michael Schemper. "A solution to the problem of separation in logistic regression." *Statistics in Medicine*, 2002: 2409-2419.

Hilbe, Joseph M. *Logistic Regression Models.* Boca Raton: Chapman & Hall/CRC Press, 2009.

Hosmer, D.W., T Hosmer, S. Le Cessie, and S. Lemeshow. "A comparison of goodness-of-fit tests for the logistic regression model." *Statistics in medicine* 16 (1997): 965-980.

Hosmer, David W., and Stanley Lemeshow. "A goodness-of-fit test for the multiple logistic regression model." *Communications in Statistics*, 1980: 1043-1069.

—. *Applied logistic regression.* Second Edition. John Wiley & Sons, 2000.

Hosmer, David W., Stanley Lemeshow, and J. Klar. "Goodness-of-fit testing for multiple logistic regression analysis when the estimated probabilities are small." *Biometrical Journal*, 1988: 911-924.

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression.* Third edition. New Jersey: John Wiley & Sons, Inc., 2013.

Jennings, D. E. "Outliers and residual distributions in logistic regression." *Journal of the American statistical association* 81 (1986): 987-990.

King, Gary, and Langche Zeng. "Logistic Regression in Rare Events Data." *Political Analysis* 9 (2001): 137-163.

Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models.* Fifth Edition. New York: McGraw-Hill/Irwin, 2005.

Lemeshow, Stanley, and David W. Hosmer. "A review of goodness-of-fit statistics for use in the development of logistic regression models." *American Journal of Epidemiology*, 1982: 92-106.

Little, Roderick J.A., and Donald B. Rubin. *Statistical Analysis with Missing Data.* Hoboken, New Jersey: John Wiley & Sons, Inc., 2002.

Martin, M., and L. Pardo. "On the asymptotic distribution of Cook's distance in logistic regression models." *Journal of Applied Statistics* 36 (2009): 1119-1146.

McCullagh, P., and J. A. Nelder. *Generalized Linear Models.* Boca Raton: Chapman and Hall/CRC, 1989.

McCullagh, Peter, and John A. Nelder. *Generalized Linear Models.* Second Edition. Boca Raton: Chapman & Hall/CRC, 1989.

Menard, Scott. *Applied logistic regression analysis.* Second Edition. Thousand oaks, CA: Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-106, 2002.

Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. "A simulation study of the number of events per variable in logistic regression analysis." *J.Clin.Epidemiol* 49, no. 12 (1996): 1373-1379.

Rubin, D. B. *Multiple imputation for nonresponse in surveys.* New York: Wiley, 1987.

Rubin, Donald B. "Inference and missing data." *Biometrika* 63 (1976): 581-592.

Schafer, Joseph L., and John W. Graham. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2002): 147-177.

Stallard, Nigel. "Simple tests for the external validation of mortality prediction scores." *Statistics in Medicine* 28 (2009): 377-388.

Vittinghof, E, and C. E. McCulloch. "Relaxing the rule of ten events per variable in logistic and Cox regression." *American Journal of Epidemiology* 165 (2006): 710-718.

Loss Given Default

Contents

Overview

Loss given default, or LGD, is usually defined as the percentage loss rate suffered by a lender on a credit exposure, if the obligor defaults. *I.e.*, even if the counterparty defaults (fails to repay the amount owed), the lender will usually succeed in recovering some percentage of the current amount owed in the process of workout or sale of the obligor's assets. The percentage is termed the recovery rate, or RR. The two percentages are related: RR = 1 – LGD. Given a portfolio of m risky deals:

$$L = \sum_{i=1}^{m} \delta_i e_i Y_i$$

$L =$ overall loss

$\delta_i =$ loss given default, $0 \le \delta_i \le 1$

$e_i =$ overall exposure

$Y_i =$ binary default indicator, $Y_i \in \{0,1\}$

Each quantity in the above equation can vary through time; in this chapter, we focus on methods that predict LGD. One method focuses on the information contained in market prices of risky instruments, and attempts to use this information for *ex-ante* estimation of future LGDs. This approach is based on Merton's structural model, in which the theory of option pricing is used.[53] This approach is not addressed in the present chapter.

In this chapter, we focus on models that estimate LGDs on the basis of historical data on realized losses. There is little open literature available for this approach, because the data is not publically available. In some models, the various assets are bucketed into classes, such as investment grade and speculative grade, and a constant LGD is attributed to each class. Such models are inherently risky due to over-simplification; for example, history indicates that probability of default (PD) and loss given default are correlated, but that the correlation varies through time, spiking during recessionary periods. Contemporary models usually include multiple firm-specific predictors combined with ambient predictors. We provide a literature review of the contemporary models.

Compared to the Probability of Default (PD), LGD was understudied partly due to the lack of data. With Basel II (Basel Committee on Banking Supervision, International Convergence of Capital Measurement and Capital Standards, 2006) established, a flourishing literature has emerged in recent years. There are major challenges in modeling LGD. First is the data limitation. Second is the bimodal or multi-modal distribution of LGD, with the modes close to the boundary values of 0 and 1. Third is that values outside of the boundaries of 0 and 1 are sometimes observed, based on the definition of LGD. Given these challenges, a variety of semi-parametric/non-parametric methods and parametric methods have been explored..

[53] See Petr Jakubik and Jakub Seidler, Estimating Expected Loss Given Default. The article can be freely downloaded from https://www.cnb.cz/en/financial_stability/fs_reports/fsr_2008-2009/FSR_2008-2009_article_4.pdf.

Some studies suggest that semi-parametric/non-parametric methods outperform traditional simple parametric methods. (Bastos, 2010) suggested that a nonparametric regression-tree model generates better out-of-sample predictions than a parametric fractional response regression model. (Loterman, Brown, Martens, Mues, & Baesens, 2012) compared ordinary least squares (OLS) regression, beta regression, robust regression, ridge regression, regression splines, neural networks, support vector machines and regression trees as well as their combinations. They concluded that the performances of these models are generally low, while observing that support vector machines and neural networks outperform traditional linear models. (Qi & Zhao, 2011) also observed that regression tree and neural network outperform fractional response regression, linear regression and other selected parametric methods. Those authors proposed an approach for modeling the bimodal distribution, including small adjustments to the 0 or 1 values, but didn't find the approach to be effective. (Li, Qi, Zhang, & Zhao, 2014) used the same data set as (Qi & Zhao, 2011), and further compared two-step, inflated beta, Tobit, censored gamma, and two-tier gamma regressions. They found that these methods all perform similarly, with moderate performances. They claimed that complex parametric models do not outperform simpler ones. These parametric methods outperform linear regression, while underperforming fractional response regression and the non-parametric methods investigated in (Qi & Zhao, 2011). (Altman & Kalotay, 2014) promoted a different semi-parametric approach. They assumed the LGD distribution to be a mixture of normal distributions, and used Markov Chain Monte Carlo to estimate the mixture components. They then used an ordered probit regression to associate the components to a number of important risk drivers, such as borrower characteristics, debt instrument characteristics and credit conditions at the time of default. Semi-parametric/non-parametric methods are less sensitive to misspecifications. In the cases of moderate or poor model performance (due to lack of important risk drivers and/or large noise in data), it is not surprising that they would outperform parametric methods.

Compared to semi-parametric/non-parametric methods, parametric methods have advantages in interpretation, prediction, and, sometimes, computation. Although claimed to have mediocre or moderate performance, some traditional simple parametric methods have been explored in the literature and applied in industry, owing to their simplicity and robustness. Some more sophisticated parametric models that better represent the LGD distribution have been found to be promising too. Linear regression has been applied to LGD modeling. The bimodal distribution of LGD challenges linear regression assumptions, particularly when implementation does not account for important risk drivers that correlate with specific LGD modes. LGD values are primarily if not all bounded by 0 and 1, which is inconsistent with the linear regression assumptions of unbounded normal distributions. A common remedy is to transform the LGD values before applying linear regression. Beta regression has also been used to model LGD, which allows the response variable to be bounded by 0 and 1 naturally. Fractional response regression (Papke & Wooldridge, 1996) has been studied frequently. The fractional response regression allows values of 0 and 1, and various conditional distributions of LGD given the explanatory variables in the regression. Censored distributions have been explored to account for the bounded nature of LGD. (Tobin, 1958) and (Goldberger, 1964) introduced the Tobit model, which uses censored normal distributions. (Sigrist & Stahel, 2011) explored censored gamma distributions. To account for significant probability mass at the boundaries, inflated models or two-part models have been studied. Inflated beta regression (Ospina & Ferrari, 2012), an improvement

from the beta regression, has been investigated. A number of two-part models have been explored. A few examples are (Ramalho, Ramalho, & Murteira, 2011), (Gürtler & Hibbeln, 2013) and (Li, Qi, Zhang, & Zhao, 2014).

Both parametric methods and semi-parametric/non-parametric methods have pros and cons. Based on a survey of GE internal LGD models, generalized linear models and linear regression are most frequently used. Generalized linear models are used to model categorized LGD, where experts create buckets of LGD and model the probability for each bucket. Linear regression, with necessary transformation of the LGD values, is used to model the continuous values of LGD. We focus on those modeling approaches that are used in the LGD models that will be in continuing use in the re-focused GE Capital enterprise. We do not advocate for any particular model, but rather, we propose quantitative testing methodologies that can be used to test the validity of the various methods that are embedded in the modeling approaches.

Multi-factor regression models, the focus of this chapter, may be simple linear regression models, where all indicators are regressed simultaneously. Staged linear regression models are also used, where the predictors for each stage are selected by business logic. Beta regression and fractional regression have also been proposed. In some modeling approaches logistic regression is used with nested linear regression. Decision tree models, combined with linear and logistic regression, are also available.

Model Specification and Evaluation

Data Processing

In practice, most data sets need certain levels of processing before reliable models can be fit. The processing includes formatting the data and evaluating the appropriateness of values. Here we focus on the latter. Common processing steps in evaluating data appropriateness include, but are not limited to, identifying abnormal values and treating missing data. Identifying abnormal values relies primarily on domain knowledge and is not discussed in detail here. In this section, we focus on treating missing data.

Missing Data

Missing data is usually a nuisance rather than a focus in model building. The model validator should evaluate the amount of missing data and determine whether it raises any concern for estimating and making inferences from the developed model. Two common scenarios when missing data raises a concern are:

- The absolute amount of missing data is high.
- The absolute amount of missing data is low, and yet the percentage of missingness for certain categories of data is high.

There is no hard and fast threshold on the amount of missing data, as this is problem dependent. The model validator should make experience-based judgment calls. To gage sensitivity, the validator can explore the common mathematical treatments (see below) of missing data and compare the outcomes.

If the amount of missing data is noteworthy, the validator should further categorize the missing data (Schafer & Graham, 2002) (Little & Rubin, 2002):

- Missing completely at random (MCAR): The missingness does not depend on missing or observed data.
- Missing at random (MAR): The missingness depends on the observed data but not on the missing data.
- Missing not at random (MNAR): The missingness depends on the missing data.

Let D_{obs} denote the observed data, D_{mis} denote the missing data if it were observed, and D_{com} denote the complete data, i.e. $D_{\text{com}} = (D_{\text{obs}}, D_{\text{mis}})$. Let M denote the missingness. Then MCAR and MAR can be formulated as:

- MCAR: $\Pr(M|D_{\text{com}}) = \Pr(M)$
- MAR: $\Pr(M|D_{\text{com}}) = \Pr(M|D_{\text{obs}})$

The validator should evaluate the data generation process and determine the mechanism of missingness.[54]

(Little & Rubin, 2002) (Schafer & Graham, 2002) (Rubin D. B., 1976), suggest the following approach to missingness:

- MCAR: Given MCAR, the observed data provide a correct sampling distribution and correct likelihood. In general, the missing data can be dropped and the model can be built on the observed data. A commonly applied method is to model the complete cases, *i.e.*, cases without missing values. The advantage of complete-case analysis is simplicity. The disadvantage is loss of precision (and bias if the missingness is not MCAR). Given MCAR, complete-case analysis is justified when the loss of precision is acceptable.
- MAR: Given MAR, dropping the missing data still generates appropriate likelihood for the parameters (Rubin D. B., 1976). Any inference from the likelihood is valid. For example, the commonly used maximum likelihood estimates (MLE) are appropriate, as well as Bayesian methods. A general method for MLE in the presence of missing data is the EM algorithm (Dempster, Laird, & Rubin, 1977). Besides likelihood-based methods, another popular method of dealing with MAR is multiple imputation (MI). MI refers to the procedure where each missing value is replaced by a number of simulated values. As a result, a number of complete data sets are generated, which represent the imputation uncertainty. Methods for complete data are applied to each data set, and the results are combined (Schafer & Graham, 2002) (Rubin D. B., 1987).
- MNAR: Under MNAR, an appropriate model for the missingness needs to be specified, and a large sample is needed. This is often infeasible in practice. Therefore, dealing with MNAR is challenging. We don't suggest any specific technique to deal with MNAR due to the complexity of the problem. Instead, we suggest for the model validator to be extremely cautious with a model developed with a MNAR data set. If the validator suspects that the departure from MAR is not severe, the validator can apply the above methods for MAR and compare the results. If the departure is severe and no rational treatment is applied to the

[54] Note that there is generally no way to test whether MCAR or MAR holds, given the missing data is not observed.

missing values by the model developer, the validator should challenge the feasibility of the model.

In practice, the departure from MCAR or MAR may not be severe enough to generate a significant impact on the model estimates, especially when the amount of missing data is low. In such cases, the bias from assuming MCAR or MAR may be negligible. The most commonly applied method by model developers is the complete-case analysis. As described above, complete-case analysis is only appropriate under MCAR. However, if the departure from MCAR is not severe, the bias in model estimates from assuming MCAR is probably negligible. If there is concern about assuming MCAR, the validator can apply appropriate methods as described above to treat the missing values, and compare results in order to decide whether assuming MCAR is appropriate.

Variable Selection

Purpose: A rigorous process of variable selection is needed and the model developers should produce rationale for adding or dropping candidate variables (predictors). However, there is no exact science for this procedure and it should be a combination of statistical soundness and business intuition. (Hosmer, Lemeshow, & Sturdivant, Applied Logistic Regression, 2013) provides step by step procedures. One toned not follow the exact steps, but a logical process for variable selection must prevail. Here, plausible reasons to add or to drop variables are listed.

Univariate Variable Selection

Univariate tests can be performed to examine the correlation between each predictor and the response variable. (Hosmer, Lemeshow, & Sturdivant, Applied Logistic Regression, 2013) suggests to use a higher p-value cut-off (*e.g.*, 0.2 or 0.25, instead of the traditional cut-off 0.05) when we use univariate association for initial screening. If the experts believe a variable is important to the model, then that variable, even with a high p-value, should be included for further analysis.

Multivariate Variable Selection

This section describes methods that select variables via multivariate analysis. Three options are discussed with a brief description below;

- Information Criteria, R^2 and Mallow's C_p are useful when comparing candidate models, where each candidate model includes some subset of variables.
- LASSO is an estimation method that simultaneously considers all of the candidate variables in a model. This method will estimate some of the coefficients as zero signalling exclusion of those variables from the final model.
- Automatic selection is an algorithm with multiple steps which adds or drops candidate variables in each step.

Information Criteria, R^2 and Mallow's C_p

When comparing candidate models, AIC and BIC are effective tools to assess the relative model performance. AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) provide summary statistics for model fit comparison. Generally, a model with lower information criteria is preferred. Hilbe (2009) provides guidance (Table 7) for selecting a preferred candidate model, although this guidance is for logistic regression model. AIC and BIC can be computed as follows:

$$AIC = -2LL + 2k$$
$$BIC = -2LL + k\log(n)$$

LL is the log-likelihood function, n is sample size, and k is the number of parameters (usually the number of predictors + 1), representing the complexity of a model. Note that BIC penalizes heavier for the number of predictors.

Table 16 Guidance on AIC /BIC criteria

AIC (Akaike Information Criteria)		BIC (Bayesian Information Criteria)	
Difference between Models A and B (Suppose A<B)	Result if A < B	Difference between Models A and B (Suppose A<B)	Degree of Preference on A
0 < (B – A) < 2.5	No difference in models	0 < (B – A) < 2	Weak
2.5 < (B – A) < 6	Prefer A if n>256	2 < (B – A) < 8	Positive
6 < (B – A) < 9.9	Prefer A if n>64	8 < (B – A) < 10	Strong
(B – A) > 10	Prefer A	(B – A) > 10	Very Strong

For linear regression models, we commonly use R^2 to evaluate overall fit of a model. Similar to information criteria, R^2 can be a useful tool to examine relative performance of the models. However, R^2 never decreases as the number of variables in the model increases. Instead of using R^2, (Kutner, Nachtsheim, Neter, & Li, 2005) suggests two other measurements similar to it. One is adjusted R^2, defined as

$$R^2_{a,p} = 1 - \frac{(n-1)}{(n-p)}\frac{SSE_p}{SST}$$

where p is the number of predictors, $\mathrm{SSE_p}$ is sum of square of error of a model, and SST is total sum of squares. The larger the value, the better the fit.

Another criterion is called Mallow's $\mathrm{C_p}$. It is an estimator of total mean squared error

$$\mathrm{C_p} = \frac{SSE_p}{MSE(P)} - (n - 2p)$$

where *MSE*(*P*) is the mean square error of a model with all possible variables (full model), and SSE_p is the sum of squares of errors for a model with a subset of the variables. When there is no bias in the regression model with the subset of variables, the expected value of $\mathrm{C_p}$ is approximately p. Thus, we would seek a model based on a subset of variables with small$\mathrm{C_p}$, approaching p [(Kutner, Nachtsheim, Neter, & Li, 2005)].

LASSO (Least Absolute Shrinkage and Selection Operator)

The LASSO (Least Absolute Shrinkage and Selection Operator) is an estimation method for the coefficients, originally applied to ordinary least squares regression. It can also be used as a variable selection method since coefficients of non-critical indicators will be shrunk to zero; we can exclude those variables from the model.

LASSO maximizes the penalized likelihood function

$$LL_{LASSO}(\beta) = LL(\beta) - \lambda \sum_j |\beta_j|$$

where $LL(\beta)$ is the original likelihood function and $\lambda \geq 0$ is a penalty parameter. Increasing λ results in greater shrinkage of $|\beta|$ toward 0. k-fold cross-validation can be used to select the optimal value of λ, and the associated model. For each candidate value of λ, mean square prediction error is measured by k-fold validation. First, divide the data into k subsamples with equal size. Then, fit a LASSO model with given λ, excluding the first subsample, and make predictions for the excluded subsample using the fitted model. Repeat this process for other subsamples. Then, we have predictions for each sample point, so that we can calculate mean square prediction error for a given λ. That allows for determination of the λ value that achieves the minimum mean square prediction error (Agresti, 2013).

Automatic Variable Selection (Stepwise/Backwards/Forwards)

Automatic variable selection measures how each variable improves the model fit when it is added or deleted. Decisions are based on statistical evidence alone, so care is needed. For example, consider two variables A and B with similar significance level. It is possible that B is selected over A by the algorithm, based on statistical evidence, even though the experts believe A is more important than B. Also, statistical significance for a particular variable depends on which variables are already included in the model. Automatic variable selection is not without peril.

Miscellaneous

There are additional considerations in variable selection:

- Multi-collinearity: Including highly correlated predictors in the model can lead to biased estimates with large standard error. Variance Inflation Factor (VIF) in linear regression indicates associations among the predictors (Menard, 2002), (Allison, 2000). VIF greater than 5 is cause for concern and greater than 10 suggests a serious collinearity problem (Menard, 2002).

- Number of predictors in the final model: (Agresti, 2013) provides a guideline for logistic regression based on a Monte Carlo study (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996); when the number of the event occurred in the data set divided by the number of predictors is smaller than 10, then parameter estimates can be a biased, in addition to other potential issues. (Hosmer, Lemeshow, & Sturdivant, Applied Logistic Regression, 2013) cited (Vittinghof & McCulloch, 2006); based on extensive simulations, the latter authors conclude that the "rule of 10" may be too conservative. However, (Hosmer, Lemeshow, & Sturdivant, Applied Logistic Regression, 2013) take issue with the latter authors' recommendation in cases where the distributions of the discrete predictors are highly skewed rather than balanced; for such cases, the rule of 10 is justified. Thus, the rule of 10 is recommended as a guideline, but not as standard practice, since the referenced studies are based on simulation, which does not cover all possible scenarios. , A less stringent requirement may suffice.
- Counterintuitive coefficient sign: With limited data, it is often useful to incorporate expert knowledge.

Generalized Linear Models

The generalized linear model[55] is widely used to predict binary events. For LGD modeling, we can discretize LGD response values into two buckets (*e.g.* LGD greater or smaller than 30%, as in the GELGD_Heller Debtor V4 model) and model the probability of a default resulting in LGD greater than

[55] Refer to (McCullagh and Nelder 1989) for detailed information about the generalized linear models.

30%. For more granular rating grade such as 6 LR rating, ordinal logistic regression can be used. See (Agresti, 2013) for more details.

Model Functional Forms

The generalized linear model uses a link function to model the probability of an event for a binary response variable. There are three commonly used link functions for the binary response: logit, probit and complementary log-log links. This chapter describes test criteria based on the logit link function, also called the logistic regression, since it is the most commonly used and studied of the three. Most of the model evaluation tools in Section 0 and Section 0 are applicable for the other link functions as well. Applicability will be specified for each criterion.

As shown in Figure 1, the fitted function in logistic regression has a characteristic S-shape (logit curve) that maps predictors to a value between zero and one which enables us to predict a probability of an event given the relevant predictors. Other link functions manifest similar S shaped curves.

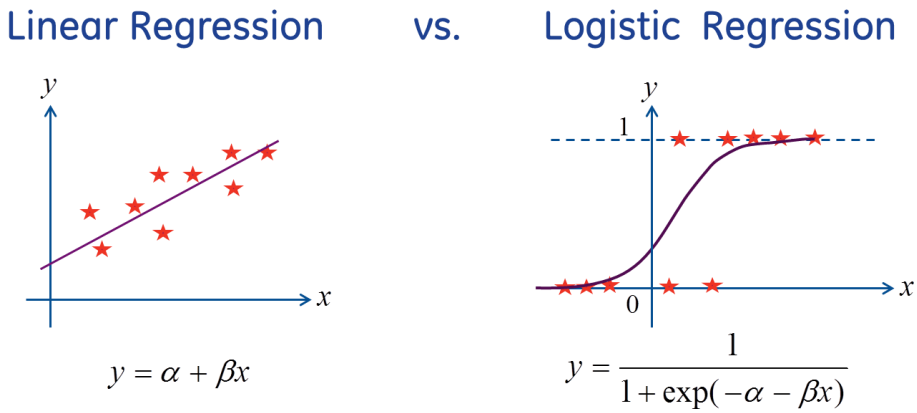

Figure 11 Difference between linear regression model vs. Logistic regression model

Here, for the purpose of discussion, we assume that the event of interest is LGD greater than 30%. Let

- Y denote the binary response variable:

$$Y = \begin{cases} 1 & \text{when LGD}\geq 30\% \\ 0 & \text{otherwise} \end{cases};$$

- $X_1, X_2, \dots, X_p$ denote the p predictors;
- π denote the probability of the event (LGD≥30%) given $X_1, X_2, \dots, X_p$; *i.e.*, $\pi = \Pr(Y = 1 | X_1, X_2, \dots, X_p)$; and
- $\beta_0, \beta_1, \dots, \beta_p$ denote the coefficients for the predictors.

Then the generalized linear model can be expressed as

$$\eta(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where $\eta(\pi)$ is called the link function. For logistic regression, $\eta(\pi)$ is defined as

$$\eta(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

Note that the probability, after logit transformation, has linear relationships with the predictors of interest. Alternative link functions are described in Table 10.

Table 17 Functional forms of link functions. Here, $\Phi^{-1}(.)$ is the inverse cumulative distribution function of the standard normal distribution.

Link function	$\eta(\pi)$
Logit	$\log\left(\frac{\pi}{1-\pi}\right)$
Probit	$\Phi^{-1}(\pi)$
Complementary log-log	$\log(-\log(1-\pi))$

The model can also be presented in matrix form. Let

- $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$ denote the vector of n observed response variables;
- $\mathbf{X}$ denote the design matrix of predictors

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix};$$

- $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)'$ denotes the vector of default probability given $\mathbf{X}$; and
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ denotes the vector of coefficients for the predictors.

Then the model can be expressed as

$$\eta(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta},$$

where the link function is applied element-wise to $\boldsymbol{\pi}$.

Coefficients can be estimated by maximizing the log-likelihood (LL)

$$LL = \sum_{i=1}^{n} (1 - Y_i)\log(1 - \pi_i) + Y_i \log(\pi_i)$$

Always check whether there are repeated observations at different levels (rows) of the predictors. Here the levels are unique values of the p-dimensional vector of predictors. The number of levels and numbers of repeated observations determine whether the asymptotic properties of certain tests are valid, as explained below. The levels are referred to as **covariate patterns** in the following context. The repeated observations are also called grouped observations.

Let

- J denote the number of covariate patterns, with $J \leq n$;
- $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_J$ denote the covariate patterns, with $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})', j = 1,2, \ldots, J$;
- m_j be the number of observations for covariate pattern $\boldsymbol{x}_j, j = 1,2, \ldots, J$, with $\sum_{j=1}^{J} m_j = n$;
- y_j be the number of events, *i.e.*, LGD≥30%, for covariate pattern $\boldsymbol{x}_j, j = 1,2, \ldots, J$; and
- $\hat{\pi}_j$ be the estimated probability of an event for covariate pattern $\boldsymbol{x}_j, j = 1,2, \ldots, J$.

In addition, a definition of **hat matrix** is introduced since it is used multiple times throughout the chapter. The name "hat matrix" came from the linear model where it projects the data to the fitted values. For generalized linear modeling, a corresponding relationship holds for $\eta(\pi)$ (Agresti, 2013). The equation that defines the hat matrix for the generalized linear model is

$$\mathbf{H} = \widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{\frac{1}{2}}$$

where $\widehat{\mathbf{W}}$ is the diagonal matrix with elements $\frac{\left(\frac{\partial \pi_i}{\partial \eta_i}\right)^2}{Var(Y_i)}$, replacing π_i with $\hat{\pi}_i$ In practice. For example, the element of $\widehat{\mathbf{W}}$ for logistic regression is $\pi_i(1-\pi_i)$ (or $m_j\pi_j(1-\pi_j)$ for the grouped observations). Table 11 describes the functional form of the element of $\widehat{\mathbf{W}}$ for each link functions.

Table 18 the element of $\widehat{W}$ for different link functions

Link function	W_i in ungrouped case	W_i in grouped case
Logit link	$\pi_i(1-\pi_i)$	$m_j\pi_j(1-\pi_j)$
Probit link	$\frac{\left[\phi\left(\Phi^{-1}(\pi_i)\right)\right]^2}{\pi_i(1-\pi_i)}$	$\frac{m_j\left[\phi\left(\Phi^{-1}(\pi_j)\right)\right]^2}{\pi_j(1-\pi_j)}$
Complementary log-log link	$\frac{(1-\pi_i)[\log(1-\pi_i)]^2}{\pi_i}$	$\frac{m_j(1-\pi_j)\left[\log(1-\pi_j)\right]^2}{\pi_j}$

Model Diagnostics

This section describes methods that assess the validity of fitting a generalized linear model to the data. A fitted model may be inadequate in the following ways:

- Relationships between the link $\eta(\pi)$ and predictors $X_1, X_2, \ldots, X_p$ are not all correctly specified. There are generally two causes of the misspecification: (1) the predictors need to be transformed in order to have linear relationships with the link; and (2) the specified link is incorrect.
- There are outliers or other strongly influential observations. They may bias the model estimates, and they are not fitted well by the model.

Three types of tools are proposed to assess the above model inadequacies. Table 8 shows which inadequacy is assessed by each tool. The tools are:

- Link specification tests
- Residual plots
- Influential observation plots

Table 19 Model Inadequacies and Diagnostic Tools

	Diagnostic Tools		
Model Inadequacies	Link specification tests	Residual plots	Influential observation plots
Misspecification of relationships between the link logit(π) and predictors $X_1, X_2, \ldots, X_p$.	X	X	
Existence of outliers or strongly influential observations		X	X

These tools enable testing of the validity of model assumptions. They also enable detailed analyses of the model in order to identify issues or improvement opportunities.

Link Specification Tests

Purpose: The generalized linear model assumes linear relationships between $\eta(\pi)$ and the predictors $X_1, X_2, \ldots, X_p$, as specified by the model equation

$$\eta(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

If the linearity assumption is inappropriate, then the coefficient estimates and standard errors are biased. Therefore, it is important to assure the validity of the linearity assumption.

Two similar tests are provided to assess the linearity assumption (Hilbe, 2009) for logistic regression

- Box-Tidwell test
- Pregibon Link test

At the time of this writing, we have not ascertained whether the above link specification tests applies to other link functions.

5) Box-Tidwell Test

Description: The Box-Tidwell test has two steps:

- Construct the interaction[56] of each continuous predictor with its log transformation
- Fit a logistic regression with the original predictors and the above interactions

If any interaction is statistically significant, based on its p-value, the Box-Tidwell test concludes that the linearity assumption is violated. The following criterion can be applied on each interaction.

Method	Criterion
Box-Tidwell Test	P-value < 0.05

6) Pregibon Link Test

Description: The Pregibon Link test is similar to the Box-Tidwell test, but uses the square of the hat matrix diagonal instead of the interaction term. It is carried out in two steps:

[56] An interaction term represents the interactive effect of two or more predictors on the response. It is generally constructed as the product of these predictors. For example, let X be a continuous predictor, and the interaction of X with its log transformation is $X\log(X)$.

- Construct the square of the hat matrix diagonal of each continuous predictor
- Fit a logistic regression with the original predictors and the above new terms.

If any new term is statistically significant, the Pregibon Link test indicates that the linearity assumption is violated.

The estimated hat matrix for logistic regression is

$$\mathbf{H}=\widehat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{1/2}.$$

Here, $\widehat{\mathbf{W}}$ is an $n \times n$ diagonal matrix

$$\widehat{\mathbf{W}} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix},$$

where $\hat{\pi}_i$ is the estimated probability of event for the i-th observation.

The following criterion can be applied to each interaction.

Method	Criterion
Pregibon Link Test	P-value < 0.05

The Pregibon Link Test is available in statistical software packages including stata (linktest) and R (pregibon).

Residual Plots

This section and the next on influential observation plots describe graphical diagnostics using several normalized residuals. Graphical assessment is preferred; hard thresholds are not available for these diagnostics. Visual assessment informs with respect to model adequacy and overly influential (to the model fit) observations. Where poor fit is observed, one would check for data issues or opportunities to improve the model by alternative variable selection or transformation of variables.

Purpose: The purpose of these residual plots is to assess model adequacy. An adequate model satisfies $E\{Y_i\} = \pi_i$. If this assumption is satisfied, it follows asymptotically that $E\{Y_i - \hat{\pi}_i\} = 0$ (Kutner, Nachtsheim, Neter, & Li, 2005).

Description: We describe three examples of residuals. Such residual plots versus predictors or/and predicted probabilities with a Lowess curve can suffice (Kutner, Nachtsheim, Neter, & Li, 2005). Each residual plot presents different information about the model fit; multiple residual plots are preferred. The Lowess curve of each plot is expected to be flat horizontal line with zero intercept. If significant deviations are found, influential data points or lurking variables left out of the model are typically causal. Our notation assumes the grouped case (repeated covariate patterns), but even for ungrouped data (without repeated covariate patterns), we can treat $J = n$ and $m_j = 1$ for all

$j = 1, \ldots, J$. However, (Agresti, 2013) and (Hosmer & Lemeshow, Applied logistic regression, 2000) recommend to compute residuals using grouped data if possible, especially when J is much smaller than n. All the following residuals can be applied for other link functions, while the calculation of hat matrix for studentized Pearson Residuals varies depending on the link function that is used.

g) Pearson Residuals: The objective is to make the residuals more comparable by dividing the raw residual by an estimate of the standard deviation of the observation y_i.

$$r_j = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

h) Studentized Pearson Residuals: The objective is to fully standardize the residuals; the raw residual $(y_j - m_j\hat{\pi}_j)$ is normalized by the estimated standard error of the fitted value $\hat{\pi}_i$.

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}} = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)(1-h_j)}}$$

where, $\mathrm{h_j}$ is the j-th diagonal element of the estimated hat matrix $\mathbf{H}$:

$$\mathbf{H} = \widehat{\mathbf{W}}^{\mathbf{1}}_{\mathbf{2}}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}'\widehat{\mathbf{W}}^{\mathbf{1}}_{\mathbf{2}}$$

where $\widehat{\mathrm{W}}$, in the case of logistic regression, is the diagonal matrix, wherein the j-th element is $m_j\hat{\pi}_j(1-\hat{\pi}_j)$. See Table 11 for the functional form of $\widehat{\mathrm{W}}$ corresponding to each link function.

i) Deviance Residuals: These residuals are related to the deviance test which will be described in Section 2.4.1.

$$dev_j = sign\left(Y_j - m_j\hat{\pi}_j\right)\sqrt{2[y_j \ln\frac{y_j}{m_j\hat{\pi}_j} + \left(m_j - y_j\right)\ln\frac{m_j - y_j}{m_j\left(1-\hat{\pi}_j\right)}]}$$

Influential Observation Plots

Purpose: The purpose of these plots is to visualize how each observation influences the model fit. The idea is to see how goodness-of-fit test statistics, or the coefficient estimates, would change if we exclude one observation from the analysis.

Description: Identify observations that stand out in the plots, then check those observations for errors. As residual plots, each plot presents different information about the model fit; multiple plots are preferred.

g) ΔX_j^2 vs. $\hat{\pi}_j$: The ordinate is related to the Pearson Chi-Square statistic.

$$\Delta X_j^2 = \frac{r_j^2}{(1-h_j)} = r_{sj}^2$$

h) ΔD_j vs. $\hat{\pi}_j$: The ordinate is related to the deviance test.

$$\Delta D_j = dev_j^2 + \frac{r_j^2 h_j}{(1-h_j)} \approx \frac{dev_j^2}{(1-h_j)}$$

i) $\Delta\hat{\beta}_j$ vs. $\hat{\pi}_j$: The ordinate relates to the observation's influence on the estimate of coefficient β.

$$\Delta\hat{\beta}_j = \frac{r_j^2 h_j}{(1-h_j)^2}$$

(Hosmer, Lemeshow, & Sturdivant, Applied Logistic Regression, 2013) recommend focusing on observations whose values for one or more of the diagnostic statistics fall well away from the rest of the values, instead of setting a threshold value to identify influential points. The following three plots are considered to be critical. Bubble plots, which plot one diagnostic statistics, with the size of the symbol proportional to another diagnostic statistics, can be a useful, as they introduce a third dimension to the graphical analysis.

Model Performance

This section describes methods to assess the overall goodness-of-fit of the model. A number of options are provided:

- Pearson Chi-Square, Deviance and Hosmer-Lemeshow tests
- AUC and ROC Curve / AR and CAP

The choice of method depends on the data, but at least one method should be used to assess the model performance.

The overall goodness-of-fit is robust to poor fit for a few observations. When the overall goodness-of-fit is rejected, the appropriate response is to review model diagnostics for improvement opportunities.

Pearson Chi-Square, Deviance and Hosmer-Lemeshow Tests

Purpose: The tests described in this section assess the goodness-of-fit of the overall model. Pearson Chi-Square test and Deviance test are classic tests for model goodness-of-fit. However, when the predictors are continuous rather than discrete, the Hosmer-Lemeshow test should be chosen.

4) Pearson Chi-Square Test

Description: The Pearson Chi-Square statistic is the sum of squares of Pearson residuals, and is defined as

$$X^2 = \sum_{j=1}^{J} \left[\frac{y_j - m_j \hat{\pi}_j}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \right]^2$$

Under the null hypothesis that the fitted model is correct, X^2 follows a χ^2 distribution asymptotically with degrees-of-freedom $J - (p + 1)$. The following criterion can be applied.

Method	Criterion
Pearson Chi-Square Test	P-value < 0.05

The asymptotic χ^2 distribution is valid when each $m_j \to \infty$ and J is finite, while $n \to \infty$. Therefore, when one or more continuous predictors are included in the model and result in $J \approx n$, the Pearson Chi-Square test is inappropriate.

5) Deviance Test

Description: Similar to the Pearson Chi-Square test, the deviance test statistic is the sum of squares of deviance residuals, defined as

$$G^2 = \sum_{j=1}^{J} 2 \left[y_j \ln \frac{y_j}{m_j \hat{\pi}_j} + (m_j - y_j) \ln \frac{m_j - y_j}{m_j (1 - \hat{\pi}_j)} \right]$$

The deviance test is also the likelihood ratio test of a saturated model[57] with J parameters and the fitted model.

Under the null hypothesis that the fitted model is correct, G^2 follows a χ^2 distribution asymptotically with degrees-of-freedom $J - (p + 1)$. The following rejection criterion can be applied.

Method	Criterion
Deviance Test	P-value < 0.05

As for the Pearson Chi-Square test, the asymptotic χ^2 distribution of the deviance test is valid when each $m_j \to \infty$ and J is finite, while $n \to \infty$. Therefore, when one or more continuous predictors are included in the model and result in $J \approx n$, the deviance test is inappropriate.

6) Hosmer-Lemeshow Test

Description: When the inclusion of continuous predictors causes violation of the asymptotic requirements for the Pearson Chi-Square and deviance tests, a remedy is to group the covariate patterns to form a finite number of groups with each approximately treated as one covariate pattern. The Hosmer-Lemeshow (HL) test proposes grouping based on the values of the estimated probabilities, with the following two options:

[57] A saturated model contains a unique probability estimate for each covariate pattern.

- by percentiles of the estimated probabilities
- by fixed values of the estimated probabilities

Grouping by percentiles of the estimated probabilities is preferred (Hosmer, Lemeshow, and Klar (1988)), and is commonly chosen as the default option by software packages. The choice of number of groups depends on the sample size, but 10 groups are most commonly used.

Let

- G denote the number of groups;
- $\mathrm{c_g}$ denote the number of covariate patterns in group g, $\mathrm{g} = 1,2,\dots,\mathrm{G}$;
- $\mathrm{m_{gj}}$ denote the number of observations for the j-th covariate pattern in group g, $\mathrm{j} = 1,2,\dots,\mathrm{c_g}$, $\mathrm{g} = 1,2,\dots,\mathrm{G}$;
- $\mathrm{y_{gj}}$ denote the number of events for the j-th covariate pattern in group g, $\mathrm{j} = 1,2,\dots,\mathrm{c_g}$, $\mathrm{g} = 1,2,\dots,\mathrm{G}$; and
- $\hat{\pi}_{\mathrm{gj}}$ denote the estimated probability of the event for the j-th covariate pattern in group g, $\mathrm{j} = 1,2,\dots,\mathrm{c_g}$, $\mathrm{g} = 1,2,\dots,\mathrm{G}$.

Then the HL statistic is defined as

$$\hat{C} = \sum_{g=1}^{G} \left[\frac{y'_g - m'_g \hat{\pi}'_g}{\sqrt{m'_g \hat{\pi}'_g (1 - \hat{\pi}'_g)}} \right]^2$$

where

$$y'_g = \sum_{j=1}^{c_g} y_{gj}$$

$$m'_g = \sum_{j=1}^{c_g} m_{gj}$$

$$\hat{\pi}_g = \frac{\sum_{j=1}^{c_g} m_{gj} \hat{\pi}_{gj}}{\sum_{j=1}^{c_g} m_{gj}}$$

Under the null hypothesis that the fitted model is correct, $\hat{C}$ follows a χ^2 distribution asymptotically with degrees-of-freedom $G - 2$. The following criterion can be applied.

Method	Criterion
HL Test	P-value < 0.05

AUC and ROC Curve / AR and CAP

Purpose: The purpose of ROC (Receiver Operator Characteristic) curve for the generalized linear model is to measure the model performance from the classification (event or no event) perspective. Through varying the classification rule, we can impact the model's capability to predict default. The ROC curve is useful for visualizing the model's predictive power, especially rank ordering power. The

analysis also provides a quantitative summary statistic called AUC (Area Under the Curve). Note that Section 0 discusses a similar analysis that takes account of actual loss, called the loss capture ratio.

Description: Suppose, for each observation i, we have a fitted value $\hat{\pi}_i$ from the model. We can classify whether the obligor will be default or not by setting a rule as below, denoting Y = 1 for default,

$$\hat{Y}_i = \begin{cases} 1 & \hat{\pi}_i \geq \pi_0 \\ 0 & \hat{\pi}_i < \pi_0 \end{cases}$$

For a given cutoff point π_0, each observations will be classified as either 0 or 1. Then, it can be compared to the true observed event Y_i to check if the prediction rule works well. By varying the cutoff point π_0, we measure the predictive power by false positive rate and true positive rate using the whole sample. The ROC curve is constructed by plotting each associated pair of false positive rate (horizontal axis) versus true positive rate (vertical axis), resulting from sweeping the cutoff point π_0 through the entire range, as in Figure 8. If the curve passes close to the top-left corner of the plot, it means the model has a high discriminatory power. On the other hand, if the curve never deviates much form to the diagonal (45° line) of the plot, the model's rank ordering power is not better than a random guess. We can quantify this curve by calculating AUC. High AUC indicates high discriminatory power.

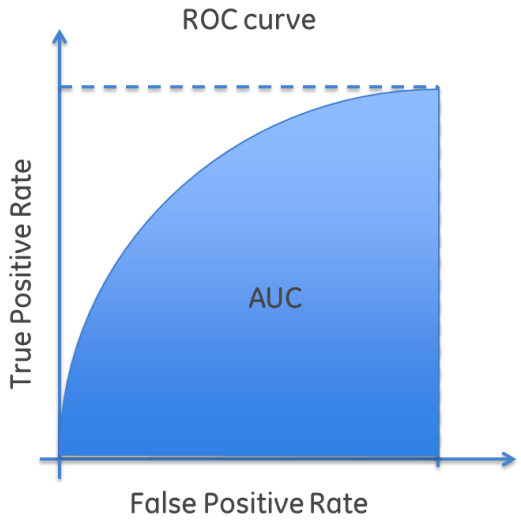

Figure 12 ROC curve example

Accuracy Ratio (AR) in Cumulative Accuracy Profile (CAP) analysis is similar to AUC in ROC. The only difference in CAP analysis is that the horizontal axis represents the total positive rate, the proportion of the sample with fitted values greater than π_0. Each ordinate measures, from the $\hat{\pi}_i \geq \pi_0$ fraction of the population, what fraction of observations was actually default. AR is computed as the area under the CAP curve and above the diagonal divided by the area under the perfect prediction curve and above the diagonal. The AR value is equal to (2*AUC -1) (Engelmann, Hayden, & Tasche, 2003).

(Hosmer, Lemeshow, & Sturdivant, Applied Logistic Regression, 2013) provides guidance on the use of the summary statistics, AUC, as acceptance criteria, as in Table 9. Here we included AR in the table by calculating (2*AUC -1).

Table 20 AR and AUC criteria

AR	AUC	Criteria
> 80%	> 90%	Outstanding
60% - 80%	80% - 90%	Excellent
40% - 60%	70% - 80%	Acceptable
< 40%	< 70%	Poor

Small Sample and Rare Events

The maximum likelihood estimate is asymptotically unbiased as sample size increases to infinity. In a small sample, the bias of a maximum likelihood estimate may not be negligible. Furthermore, the estimate is biased away from 0; therefore, it should be adjusted, or shrunk, towards 0 (McCullagh & Nelder, 1989) (Firth, 1993).

A related issue is when there are only a small number of observations of a certain event. This is usually referred to as a "rare events" issue. The bias of the maximum likelihood estimate depends on the number of events. The smaller the number of events, the larger the bias is expected to be.

In small to medium-sized data sets or large-sized data sets with rare events, a situation may occur where the events are perfectly separated by a set of predictors. This situation is called "separation" or "monotone likelihood" (Heinze & Schemper, 2002). In case of separation, at least one parameter estimate is infinite. An infinite parameter estimate can be also considered as extremely inaccurate, having infinite bias, and is inappropriate for modeling or making inference from.

The bias on the estimation of coefficients can be approximated by

$$\text{bias}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{W}}\boldsymbol{\xi}$$

where, for logistic regression, $\xi_i = 0.5Q_{ii}(2\hat{\pi}_i - 1)$, Q_{ii} are the diagonal elements of $\mathbf{Q} = \mathbf{X}\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'$, and $\hat{\mathbf{W}}$ is an $\mathrm{n} \times \mathrm{n}$ diagonal matrix wherein the i-th element is $\hat{\pi}_i(1-\hat{\pi}_i)$, where $\hat{\pi}_\mathrm{i}$ is the estimated probability of event for the i-th observation. (McCullagh & Nelder, 1989) provides the formulas for other link functions as well. For other link functions, use the definition of ξ_i in Table 13 and $\hat{\mathbf{W}}$ in Table 11.

Table 21 Functional forms of ξ_i for other link functions. Here, $\eta_i = x_i\hat{\boldsymbol{\beta}}$ where x_i is i-th row of X matrix and Q_{ii} are the diagonal elements of $\mathbf{Q} = \mathbf{X}\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'$, and $\hat{\mathbf{W}}$ is an $\mathbf{n} \times \mathbf{n}$ diagonal matrix wherein the i-th element is given in Table 11.

Link function	ξ_i
Logit	$0.5Q_{ii}(2\hat{\pi}_i - 1)$
Probit	$Q_{ii}\eta_i/2$
Complementary log-log	$Q_{ii}(\exp(\eta_i) - 1)/2$

We propose some alternative estimation methods to address the issue of bias. A significant difference between the resulting estimates and the regular maximum likelihood estimate should raise concern regarding bias or unreliable estimate.

Penalized Likelihood Method

Purpose: The penalized likelihood method is an approach to reducing small-sample bias in maximum likelihood estimates (Firth, 1993). It is a solution to the problem of separation (Heinze & Schemper, 2002). The method is also called the "Firth method," after its author.

Description: Instead of maximizing the likelihood function, the Firth method maximizes the penalized likelihood function

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta})|I(\boldsymbol{\beta})|^{1/2}$$

where $L(\boldsymbol{\beta})$ is the original likelihood function and $|I(\boldsymbol{\beta})|^{1/2}$ is the penalty function (Firth, 1993). Here $I(\boldsymbol{\beta})$ is the Fisher information matrix. Using the penalized likelihood, the estimates are calculated by solving the following score equation (Heinze & Schemper, 2002):

$$U(\beta_j) + 1/2\,\mathrm{trace}\left(I(\boldsymbol{\beta})^{-1}\{\partial I(\boldsymbol{\beta})/\partial\beta_j\}\right) = 0,$$

where $U(\beta_j) = \partial \log L(\boldsymbol{\beta})/\partial\beta_j$ is the usual score equation for the maximum likelihood estimate, and $j = 0,1,\ldots p$.

In case of logistic regression, the above score equation is

$$\sum_{i=1}^{n}\left\{Y_i - \pi_i + h_i\left(\frac{1}{2} - \pi_i\right)\right\}X_{ij} = 0,$$

where h_i is the i-th diagonal element of the hat matrix $\mathbf{H}=\widehat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{1/2}$, with $\widehat{\mathbf{W}} = \mathrm{diag}\{\hat{\pi}_i(1-\hat{\pi}_i)\}$, and $j = 0,1,\ldots p$.

A significant difference between the parameter estimate from the Firth method and from the regular maximum likelihood estimation raises concern on the bias of the maximum likelihood estimates. There isn't any strict criterion on whether the difference between the two estimates is statistically significant. Use the comparison as an indication of potential concern; judgment is required.

Bias Correction Method by (King and Zeng, 2001)

Purpose: According to (King & Zeng, 2001), rare event data can cause bias in two ways. One is biased estimation of the coefficients and the other is bias in the estimated probability of an event. Their method provides a means to correct the bias inherent in maximum likelihood estimates based on rare event data.

Description: The bias corrected estimate is

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \text{bias}(\hat{\boldsymbol{\beta}})$$

Using the bias corrected estimates above, we can estimate the probability of the event as

$$\tilde{\pi}_i = \Pr(Y_i = 1|\tilde{\boldsymbol{\beta}}) = \frac{1}{1 + e^{\mathbf{x}_i\tilde{\boldsymbol{\beta}}}}$$

which is preferable to $\hat{\pi}_i = \Pr(Y_i = 1|\hat{\boldsymbol{\beta}})$. Here, $\boldsymbol{x_i}$ is i-th row of $\boldsymbol{X}$ matrix. The reference cautions that this modification is not optimal, because it ignores the uncertainty in $\tilde{\boldsymbol{\beta}}$, which leads to underestimating the rare event. Corrected estimation is achieved by averaging over the uncertainty in $\tilde{\boldsymbol{\beta}}$. To achieve this averaging, one approach uses simulation and another approach uses an analytical approximation. The paper provides an analytical approximation only for logistic regression as shown

$$\tilde{\pi}_i + (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)x_i V(\tilde{\beta})x_i'$$

Here, $V(\tilde{\boldsymbol{\beta}}) = \left(\frac{n}{n+k}\right)^2 V(\hat{\boldsymbol{\beta}})$ is a variance matrix of $\tilde{\boldsymbol{\beta}}$ and $V(\hat{\boldsymbol{\beta}})$ is a variance matrix of $\hat{\boldsymbol{\beta}}$;

$$V(\hat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{n} \pi_i(1 - \pi_i)\boldsymbol{x_i}\boldsymbol{x_i}'\right]^{-1}$$

substituting $\tilde{\pi}_i$ for π_i for the calculation. You can see that if $\tilde{\pi}_i < 0.5$, the correction term is positive and we will have higher estimated probability of the event occurrence.

As with the Firth method, if there is a significant difference between the parameter estimate from the correction and from the unmodified maximum likelihood estimation, concern should be raised.

Linear Regression

Linear Regression can be used to model LGD after logit transformation. Instead of logit transformation, one can also consider alternative transformations, such as in **Table 10**. Linear regression is a common tool to relate response variables (observed LGD) to relevant explanatory variables. The expected (predicted) LGD would be the final output of the model transferred back to 0-100% scale.

Model Functional Forms

This section describes the mathematical formulation of such models.

Let

- Y_i denote the value of the response variable, which is observed LGD after the transformation;
- $X_{1i}, X_{2i}, \dots, X_{pi}$ denote the p predictors; and
- $\beta_0, \beta_1, \dots, \beta_p$ denote the coefficients for the predictors.

Then the linear regression model can be expressed as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i, i = 1,2, \dots, n$$

where the error terms ε_i are independent $N(0, \sigma^2)$, and are often referred to as *white noise*.

The model can also be presented in a matrix form. Let

- $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)'$ denote the vector of observed response variables;
- $\mathbf{X}$ denote the design matrix, with

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix}; \text{ and}$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ denote the coefficients for the predictors.

Then the model can be expressed as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the error terms $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ follow $N(0, \sigma^2\mathbf{I})$. Here $\mathbf{I}$ is an identity matrix.

Coefficients can be estimated by the least square method. In the least square method, the estimates of $\beta_0, \beta_1, \dots, \beta_p$ are those that minimize the following sum of squares:

$$Q = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_p X_{pi}\right)^2$$

Equivalently, the least square estimates are the solutions of the following normal equations:

$$\mathbf{X}'\mathbf{X}\boldsymbol{b} = \mathbf{X}'\boldsymbol{Y}$$

And the least square estimators are

$$\boldsymbol{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y}$$

Coefficients can also be estimated using the maximum likelihood method. The maximum likelihood estimates and least square estimates are the same for linear regression with normally distributed error terms.

Model Diagnostics

Residual Normality

Purpose: Linear regression inference assumes that the error term follows a standard normal distribution. If it is not normally distributed, the p-value for the inferences can be inaccurate.

Normal Q-Q plot (Kutner, Nachtsheim, Neter, & Li, 2005) & (Fox, 2008)

Normality of Residuals can be identified using a quantile-comparison plot, known as a Q-Q plot. The Q-Q plot visually compares the cumulative distribution of the (standardized or studentized) residuals

to a cumulative reference distribution. The points should be distributed close to the diagonal line. If non-normality exists, the plot may show a pattern. For example, concave-downward curvature at the left end indicates left-skewed distribution; concave-upward plot at the right end indicates right-sked distribution.

To construct a QQ plot for residuals, consider studentized residuals, $\tilde{e}_{(1)} < \tilde{e}_{(2)} < \cdots < \tilde{e}_{(n)}$ sorted in increasing order. Studentized residuals are defined as,

$$\tilde{e}_i = \frac{e_i}{\sqrt{MSE(1-h_i)}}$$

where e_i is residuals $Y_i - \hat{Y}_i$, MSE is mean square error from the model and h_i is defined as

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2}$$

For normal quantiles, we have, for $i = 1, \ldots, n$

$$F_i = \Phi\left(\frac{i - 0.375}{n + 0.25}\right)$$

where $\Phi(.)$ is cumulative distribution function of standard normal distribution. Then plot F_i versus $\tilde{e}_{(i)}$. Instead of standard normal distribution, one can use t-distribution with degrees of freedom (n-p-1). For residuals, one can also use standardized residuals (e_i/MSE) or raw residuals. In **Error! Reference source not found.**, R software package 'car', is used to compare studentized residuals to the t-distribution.

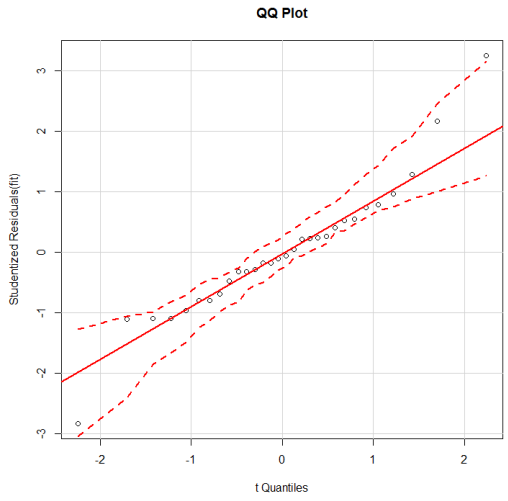

Figure 13 Example of QQ plot for residuals

Error! Reference source not found. illustrates an example of QQ plot. Data points are scattered around the diagonal line. The two dotted lines represent the "confidence envelope" suggested by Atkinson (1985). Taking into account the correlational structure of the explanatory variables, simulated sampling is employed to construct the confidence envelope. A weakness of this method is that the probability of the data points appearing outside of the envelope is greater than α (e.g. 0.05). Thus, one should not count the fraction of data points outside the interval; rather one should focus on the overall data-scatter pattern, which should be linear and matched in slope with the diagonal.

Anderson-Darling test (D'Agostino & Stephens, 1986)

Goodness of fit tests, such as the Anderson-Darling test (AD test) can be used to test the normality of the error term. The test measures distance between the empirical distribution of the data and a given cumulative distribution function that you are assuming; in this case, the normal distribution.

Description: Section 4.8.5. in (D'Agostino and Stephens) discussed specifically how to test residuals for normality. Although residuals are not independent, asymptotic distribution of the statistics below is the same as when we test normality with unknown mean and variance.

They suggested to use studentized residuals with $Z_{(i)} = \Phi(\tilde{e}_{(i)})$ where $\tilde{e}_{(i)}$ is a studentized residual.

Consider $Z_{(1)} < Z_{(2)} < \cdots < Z_{(n)}$ and $\bar{Z} = \sum Z_{(i)}/n$,

$$A^2 = -n - \frac{1}{n}\sum_i (2i-1)[log\, Z_{(i)} + log\{1 - Z_{(n+1-i)}\}]$$

The following criterion can be applied to reject the null hypothesis that the error term is normal distributed [See Table 4.7 in (D'Agostino & Stephens, 1986)].

Method	Criterion
Anderson-Darling Test	$A^2 > 0.752$

Residual Heteroskedascity

Purpose: One of the assumptions of linear regression model is that the variances of error term are the same for all observations. When the assumption is violated, although the coefficient estimates are still unbiased, standard errors for coefficient estimates are not accurate.

Residual plot

One way to visualize the presence of heteroskedasticity is to plot studentized residuals versus the predictor(s) or versus the fitted values. If the error terms all have the same variance one would expect to see that the points are scattered randomly around zero. For example, if there is a pattern such as a fan-shape, it indicates heteroskedasticity, since the variance of residuals is related to predictors or the response.

Residual plots provide other information as well. We can detect outliers, non-independence of error and nonlinearity of regression functions. When there is a pattern in a residual plot, one should check whether the model is violating any assumptions or there is a way to improve the model.

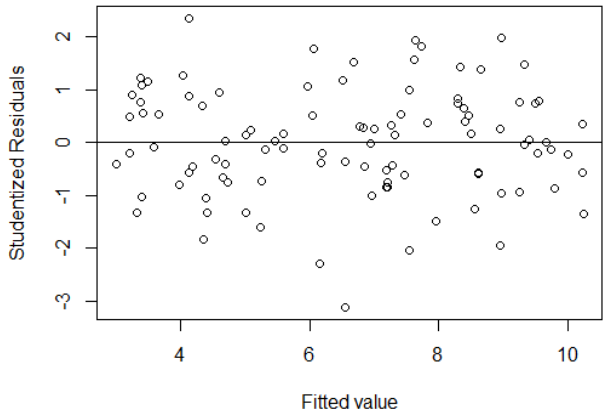

Figure 14 Example of Residual Plot

Breusch-Pagan test (Kutner, Nachtsheim, Neter, & Li, 2005) & (Fox, 2008)

Breusch and Pagan (1979) developed a score test for heteroskedasticity based on the assumption

$$\sigma_t^2 = g(\gamma_0 + \gamma_1 Z_{1t} + \cdots + \gamma_p Z_{kt})$$

where $Z_1, \ldots, Z_k$ are known variables and function g is quite general. Choice of $Z_1, \ldots, Z_k$ and function g depends on the suspected pattern that we can observed from residual plots. One simple way is to test whether the variance of error terms is related to the level of explanatory variables X. Hypothesis testing with $H_0: \gamma_1 = \cdots = \gamma_p = 0$ for a linear regression model

$$\sigma_t^2 = \gamma_0 + \gamma_1 X_{1t} + \cdots + \gamma_p X_{pt}$$

is performed fitting a regression model of e_t^2 against predictors. Note that e_t^2 is an estimate for σ_t^2. The test statistic $X_{BP}^2$is defined as

$$\mathrm{X}_{\mathrm{BP}}^2 = \frac{\mathrm{SSR}^*}{2} / \left(\frac{\mathrm{SSE}}{\mathrm{n}}\right)^2$$

where SSR^* is regression sum of squares when regressing e_{i}^2 against predictors and SSE is the error sum of squares of the main model regressing Y against predictors. If the null hypothesis is true and sample size is reasonably large, the statistic follows approximately Chi-square distribution with degrees of freedom p, the number of predictors.

The following criterion can be applied to reject the null hypothesis.

Method	Criterion
Breusch-Pagan test	$\mathrm{X}_{\mathrm{BP}}^2 > \chi^2(0.95; \mathrm{df} = \mathrm{p})$

Similarly, Cook and Weisberg (1983) suggested regressing e_i^2 against fitted values from the main model and the statistics will follow approximately Chi-square distribution with degrees of freedom 1.

Outcome Validation

This section describes how to evaluate the alignment of the estimated LGD with the empirical LGD. (Li, Bhariok, Keenan, & Santilli, 2009) discusses approaches, both graphical and quantitative, to examine the performance of the LGD model. Included are summary plots, confusion matrix and several quantitative measurements. Here, we briefly summarize the list of analyses and their purposes. For more details, see (Li, Bhariok, Keenan, & Santilli, 2009). For notation, expected LGD indicates direct model output from the LGD model, between 0 and 100%, and observed LGD indicates empirical LGD, usually between 0 and 100% (or it is censored so that the values fall between 0 and 100%).

Visualization

Graphical representation of the model captures nuanced model behavior that is typically not available from quantitative performance measurement. The paper suggests three plots:

- Scatter plot: observed LGD vs. expected LGD. We expect that the data points to cluster around the diagonal line if the model is accurate
- Histogram: check if there is any major difference between distributions of observed LGD and expected LGD
- Box and whisker plots of EAD per LGD buckets: It is useful to gauge the relative importance of the various confusion matrices.

Confusion matrix

		Predicted		
		Cat	**Dog**	**Rabbit**
Actual class	**Cat**	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

Figure 15. The confusion matrix for an imperfect method to classify cats, dogs, and rabbits (courtesy of Wikipedia). In this example, the data represents animal count, but we could have used animal mass or some other indicator.

An example confusion matrix is presented in Figure 15. In the case of LGD, there are three ways to construct confusion matrices:

- Count basis
- EAD basis (EAD basis enables us to measure the underlying exposure at risk)
- Observed loss basis

In the case of an accurate LGD model, most non-zero values to be concentrated around the diagonals if the model has high predictive power. We can describe quantitatively the tendancy for non-zero values to cluster on or near the diagonal.

- Percent matched: percentage of a perfect match (or within 1 notch diff).
- Mean absolute deviation: Use a set of weights for each cell in the confusion matrix that will be indicative of the deviation of the expected rating from the observed. Here, weights can be calculated as absolute difference between observed LGD and expected LGD for each cell i and j as below equation. A higher value means a worse fit.

$$\frac{\sum_{i,j} \text{count}_{i,j} \cdot |\text{Observed LGD}_i - \text{Expected LGD}_j|}{\sum_{i,j} \text{count}_{i,j}}$$

Other quantitative measurements

Expected Loss shortfall

Backtesting can be used to calculate the total loss difference between the observed and expected, for each transaction i in a given validation dataset (usually the time horizon is one year).

$$\frac{\sum_i \text{EAD}_i \cdot \text{Observed LGD}_i \; - \; \sum_i \text{EAD}_i \cdot \text{Expected LGD}_i}{\sum_i \text{EAD}_i \cdot \text{Observed LGD}_i}$$

This value is valuable if there is a required accuracy associated with the end use.

Loss capture ratio

The Loss capture ratio measures rank-ordering capability of the LGD model. Three curves, described below are graphed, showing the actual aggregate loss captured as a function of the modeled or observed LGD-ordered population (%).

- Model loss capture curve: the transactions are sorted along the x axis by the expected LGD (raw output of the model prior to discretizing), then aggregate observed (captured) loss, as a percentage of total loss, is plotted on the y axis.
- Ideal loss capture curve: the transactions are sorted along the x axis by the observed LGD, then aggregate observed (captured) loss, as a percentage of total loss, is plotted on the y axis.
- Random loss capture curve: 45-degree line.

Conceptually, it is very similar to Accuracy ratio. Loss capture ratio is defined as a ratio of the area between the model loss capture curve and the random loss capture curve and the area between the ideal loss capture curve and the random loss capture curve.

Reference

Agresti, A. (2013). *Categorical data analysis.* New Jersey: John Wiley & Sons.

Allison, P. D. (2000). *Logistic regression using the SAS system: Theory and applications.* Cary, NC: SAS Institute Inc.

Altman, E. I., & Kalotay, E. A. (2014). Ultimate Recovery Mixtures. *Journal of Banking & Finance*, 116-129.

Basel Committee on Banking Supervision. (2006). *International convergence of capital measurement and capital standards.* Bank for International Settlements.

Basel Committee on Banking Supervision. (2006). *International Convergence of Capital Measurement and Capital Standards.* Bank for International Settlements.

Bastos, J. A. (2010). Forecasting Bank Loans Loss-Given-Default. *Journal of Banking & Finance*, 2510-2517.

D'Agostino, R., & Stephens, M. (1986). *Goodness-of-fit techniques.* Marcel Dekker, Inc.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data vis the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*, 1-38.

Engelmann, B., Hayden, E., & Tasche, D. (2003). Testing rating accuracy. *Risk, 16*, 82-86.

Firth, D. (1993, March). Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, 27-38.

Fox, J. (2008). *Applied regression analysis and generalized linear models* (Second edition ed.). Sage.

GECC Treasury Model Risk Management Leader. (2014). *Model Risk Management Procedures.*

Goldberger, A. S. (1964). *Economic Theory.* New York: Wiley.

Gürtler, M., & Hibbeln, M. (2013). Improvements in Loss Given Default Forecasts for Bank Loans. *Journal of Banking & Finance*, 2354–2366.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 2409-2419.

Hilbe, J. M. (2009). *Logistic Regression Models.* Boca Raton: Chapman & Hall/CRC Press.

Hosmer, D. W., & Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, 1043-1069.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (Second Edition ed.). John Wiley & Sons.

Hosmer, D. W., Lemeshow, S., & Klar, J. (1988). Goodness-of-fit testing for multiple logistic regression analysis when the estimated probabilities are small. *Biometrical Journal*, 911-924.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (Third edition ed.). New Jersey: John Wiley & Sons, Inc.

Hosmer, D., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine, 16*, 965-980.

Jennings, D. E. (1986). Outliers and residual distributions in logistic regression. *Journal of the American statistical association, 81*, 987-990.

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis, 9*, 137-163.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (Fifth Edition ed.). New York: McGraw-Hill/Irwin.

Lattin, J., Carroll, J., & Green, P. (2003). *Analyzing Multivariate Data.* Duxbury.

Lemeshow, S., & Hosmer, D. W. (1982). A review of goodness-of-fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, 92-106.

Li, D., Bhariok, R., Keenan, S., & Santilli, S. (2009). Validation techniques and performance metrics for loss given default models. *The journal of risk model validation, 3*, 3-26.

Li, P., Qi, M., Zhang, X., & Zhao, X. (2014). *Further Investigation of Parametric Loss Given Default Modeling.* Office of the Comptroller of the Currency, Economics Working Paper 2014-2.

Little, R. J., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking Regression Algorithms for Loss Given Default Modeling. *International Journal of Forecasting*, 161-170.

Martin, M., & Pardo, L. (2009). On the asymptotic distribution of Cook's distance in logistic regression models. *Journal of Applied Statistics, 36*, 1119-1146.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (Second Edition ed.). Boca Raton: Chapman & Hall/CRC.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models.* Boca Raton: Chapman and Hall/CRC.

Menard, S. (2002). *Applied logistic regression analysis* (Second Edition ed.). Thousand oaks, CA: Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-106.

Ospina, R., & Ferrari, S. L. (2012). A General Class of Zero-or-One Inflated Beta Regression Models. *Computational Statistics & Data Analysis*, 1609–1623.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates. *Journal of Applied Econometrics*, 619-632.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *J.Clin.Epidemiol, 49*(12), 1373-1379.

Qi, M., & Zhao, X. (2011). Comparison of Modeling Methods for Loss Given Default. *Journal of Banking & Finance*, 2842-2855.

Ramalho, E., Ramalho, J. J., & Murteira, J. M. (2011). Alternative Estimating and Testing Empirical Stragegies for Fractional Regression Models. *Journal of Economic Surveys*, 19-68.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*, 147-177.

Sigrist, F., & Stahel, W. A. (2011). Using The Censored Gamma Distribution for Modeling Fractional Response Variables with an Application to Loss Given Default. *ASTIN Bulletin*, 673-710.

Stallard, N. (2009). Simple tests for the external validation of mortality prediction scores. *Statistics in Medicine, 28*, 377-388.

Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, 24-36.

Vittinghof, E., & McCulloch, C. E. (2006). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology, 165*, 710-718.

Stress Testing

Stress Test Models

Sanghee Cho, Jin Xia, Michael Vallance, and Jerrold Cline

Contents

Introduction

Stress Test models are required by CCAR to estimate business outcomes based on Fed Macro Scenarios. In this quantitative testing chapter we are focusing our testing on the important stress test models of the GECC verticals such as EFS and GECAS. These are mainly linear regression models and systems of regression models that are concatenated, where the output of one regression model is an input into another regression model. Given that background, this testing chapter focuses on linear regression and system-level performance of concatenated linear regressions.

Problem Formulation

In stress testing, linear regression is widely used to model the relationship among the time series of interest. This section describes the mathematical formulation of such models.

Let

- Y_t denote the value of the response variable at period t;
- $X_{1t}, X_{2t}, \dots, X_{pt}$ denote the p predictors at period t; and
- $\beta_0, \beta_1, \dots, \beta_p$ denote the coefficients for the predictors.

Then the linear regression model can be expressed as

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_p X_{pt} + \varepsilon_t, t = 1,2, \dots, T$$

where the error terms ε_t are independent $N(0, \sigma^2)$, and are often referred to as *white noise*.

The model can also be presented in a matrix form. Let

- $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_T)'$ denote the vector of observed response variables at T periods;
- $\mathbf{X}$ denote the design matrix, with

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1T} & \cdots & X_{pT} \end{bmatrix}; \text{ and}$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ denote the coefficients for the predictors.

Then the model can be expressed as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the error terms $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T)'$ follow $N(0, \sigma^2 \mathbf{I})$. Here $\mathbf{I}$ is an identity matrix.

Coefficients can be estimated by the least square method. In the least square method, the estimates of $\beta_0, \beta_1, \ldots, \beta_p$ are those that minimize the following sum of squares:

$$Q = \sum_{t=1}^{T} \left(Y_t - \beta_0 - \beta_1 X_{1t} - \cdots - \beta_p X_{pt}\right)^2$$

Equivalently, the least square estimates are the solutions of the following normal equations:

$$\mathbf{X}'\mathbf{X}\boldsymbol{b} = \mathbf{X}'\boldsymbol{Y}$$

And the least square estimators are

$$\boldsymbol{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y}$$

Coefficients can also be estimated using the maximum likelihood method. The maximum likelihood estimates and least square estimates are the same for linear regression with normally distributed error terms.

Linear Regression Model Diagnostics

Stationarity

Purpose: Economic time series are often not stationary. This results in invalid hypothesis testing results in linear regression models applied to those time series (Granger and Newbold 1973[58]).

To study from a mathematical perspective, let's consider the following widely used model of a time series Z_t

$$\phi(B)\nabla^d(Z_t - \mu) = \theta(B)\epsilon_t$$

with

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$
$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$$
$$\nabla = 1 - B$$

In the above equation, B is the backshift operator, *i.e.*, $BZ_t = Z_{t-1}$, ϵ_t is white noise, ∇ is called the differencing operator or differencing factor, $d \geq 0$ and $\phi(B)$ cannot be further factorized by $1 - B$. μ is a constant, often representing an origin, or mean if the time series Z_t is stationary. Let z be a complex variable. If all the roots of $\phi(z) = 0$ are outside the unit circle, the above model is the commonly applied ARIMA(p, d, q) model (Box and Jenkins 1970[59]).

[58] C.W.J. Granger and P. Newbold, Spurious Regression in Econometrics, *Journal of Econometrics* 2 (1974) 111-120.
[59] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control* (1970), Holden-day, San Francisco.

Whether Z_t is stationary depends on $\phi(B)$ and d. If all the roots of $\phi(z) = 0$ are outside the unit circle and $d = 0$, then Z_t is stationary. If all the roots of $\phi(z) = 0$ are outside the unit circle and $d > 0$, then Z_t is not stationary, and Z_t is said to have d unit roots. If any root of $\phi(z) = 0$ lies inside the unit circle, then Z_t is not stationary. Such roots are sometimes referred to as explosive roots.

Explosive roots are not realistic for most time series in practice, and therefore are not a focus in literature nor here. Unit roots, on the contrary, have been shown to be both practically and theoretically important. Differencing has often been found effective in dealing with non-stationarity in practice (Dickey, Bell and Miller 1986[60], Granger and Newbold 1973). A unit root is also a common theoretical implication from economic theories (Phillips and Perron 1988[61]). Therefore, this section focuses on the unit roots.

There are multiple tools in examining whether the time series has unit roots, including plots of the series and its differences, sample autocorrelation function (ACF) of the series and its differences, informal inspection of the estimated coefficients of the above model, and formal statistical tests. Although there are advantages of every tool and we encourage the validator to apply more than one tool, we explain the formal statistical tests in details here. The reason is that they are more rigorous and provide quantitative inferences.

Dickey and Fuller (1979[62], 1981[63]), Said and Dickey (1984[64], 1985[65]) conducted prominent work in testing unit roots. They developed the Dickey-Fuller test and the Augmented Dickey-Fuller test, which will be explained later. Their work focuses on the case of $d = 1$, i.e. one unit root. They assumed that other tools mentioned can detect all differencing factors except the last one, as it is the most difficult to detect (Dickey, Bell and Miller 1986). Phillips and Perron (1988)[66] modified the Dickey-Fuller test to incorporate autocorrelation and heteroscedasticity. The Phillips-Perron test will also be introduced in this section.

These tests can be used to examine if the time series of the dependent variable and independent variables are stationary individually, the violation of which may result in inappropriate testing results

[60] D. A. Dickey, W. R. Bell and R. B. Miller, Unit Roots in Time Series Models: Tests and Implications, *The American Statistician*, Vol. 40, No. 1 (Feb., 1986), pp 12-26

[61] P.C.B. Phillips and P. Perron, Testing for a Unit Root in Time Series Regression, *Biometrika* (1988), 75, 2. pp 335-46

[62] D. A. Dickey and W. A. Fuller, Distribution of the Estimators for Autoregressive Time Series With a Unit Root, *Journal of the American Statistical Association* (1979), 74, 427-431

[63] D. A. Dickey and W. A. Fuller, Likelihood Ratio Statistics for Autoregressive Time Series With a Unit Root, *Econometrica* (1981), 49, 1057-1072

[64] S. E. Said and D. A. Dickey, Testing for Unit Roots in Autoregressive Moving Average Models with Unknown Order, *Biometrika* (1984), 71, 599-607

[65] S. E. Said and D. A. Dickey, Hypothesis Testing in ARIMA(p, d, q) models, *Journal of the American Statistical Association* (1985), 80, 369-374

[66] P. C. B. Phillips and P. Perron, Testing for a Unit Root in Time Series Regression, *Biometrika* (1988), 75, 2. pp 335-46

on the coefficients. In testing for co-integration[67], these tests can be applied to determine if the residual series in the co-integration equation is stationary.

Augmented Dickey-Fuller test

The Dickey-Fuller test considers 3 autoregressive models of a time series Z_t:

1) Zero mean model: $Z_t = \rho Z_{t-1} + \epsilon_t$
2) Constant mean model: $Z_t = \mu + \rho Z_{t-1} + \epsilon_t$
3) Trend model: $Z_t = \mu + \beta t + \rho Z_{t-1} + \epsilon_t$

where ϵ_t is white noise.

The null hypothesis and alternative hypothesis are

$$H_0: \rho = 1 \text{ vs } H_a: |\rho| < 1$$

If the null hypothesis is rejected, the time series does not have a unit root and is stationary.

Two test statistics were considered

- ρ statistic: least square estimate of ρ
- τ statistic: an analogue to the t statistic in linear regression

Notations of the 2 test statistics for the 3 models are

Model	ρ Statistic	τ Statistic
Zero mean model	$\hat{\rho}$	$\hat{\tau}$
Constant mean model	$\hat{\rho}_\mu$	$\hat{\tau}_\mu$
Trend model	$\hat{\rho}_\tau$	$\hat{\tau}_\tau$

The τ statistic is

$$\hat{\tau} = (\hat{\rho} - 1)(S_{e1}^2 c_1)^{-1/2}$$
$$\hat{\tau}_\mu = (\hat{\rho}_\mu - 1)(S_{e2}^2 c_2)^{-1/2}$$
$$\hat{\tau}_\tau = (\hat{\rho}_\tau - 1)(S_{e3}^2 c_3)^{-1/2}$$

where

$$S_{ek}^2 = (T - k - 1)^{-1}[\mathbf{Z}_t'(\mathbf{I} - \mathbf{U}_k(\mathbf{U}_k'\mathbf{U}_k)^{-1}\mathbf{U}_k')\mathbf{Z}_t]$$

and c_k is the lower-right element of $(\mathbf{U}_k'\mathbf{U}_k)^{-1}, k = 1, 2, 3$, with

$$\mathbf{Z}_t = \begin{bmatrix} Z_2 \\ Z_3 \\ \vdots \\ Z_T \end{bmatrix}, \mathbf{U}_1 = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{T-1} \end{bmatrix}, \mathbf{U}_2 = \begin{bmatrix} 1 & Z_1 \\ 1 & Z_2 \\ \vdots & \vdots \\ 1 & Z_{T-1} \end{bmatrix}, \mathbf{U}_3 = \begin{bmatrix} 1 & 1 - T/2 & Z_1 \\ 1 & 2 - T/2 & Z_2 \\ \vdots & \vdots & \vdots \\ 1 & T - 1 - T/2 & Z_{T-1} \end{bmatrix}.$$

[67] R. F. Engle and C. W. J. Granger, Co-Integration and Error Correction: Representation, Estimation, and Testing, *Econometrica*, Vol. 55, No. 2 (Mar., 1987), pp 251-276

Under the null hypothesis, given that the time series is not stationary, the ρ statistic does not follow a normal distribution and the τ statistic does not follow a t-distribution as in a linear regression. Dickey and Fuller (1979) derived their rather complicated asymptotic distributions. We suggest using statistical software, such as R or SAS, to calculate the p-value of each test statistic. For the validator's information, here we provide the formulas of the asymptotic distributions:

1) Zero mean model:

$$T(\hat{\rho} - 1) \xrightarrow{L} \frac{1}{2}\Gamma^{-1}(\Psi^2 - 1)$$
$$\hat{\tau} \xrightarrow{L} \frac{1}{2}\Gamma^{-\frac{1}{2}}(\Psi^2 - 1)$$

2) Constant mean model:

$$T\left(\hat{\rho}_\mu - 1\right) \xrightarrow{L} \frac{1}{2}(\Gamma - W^2)^{-1}(\Psi^2 - 2\Psi W - 1)$$
$$\hat{\tau}_\mu \xrightarrow{L} \frac{1}{2}(\Gamma - W^2)^{-\frac{1}{2}}(\Psi^2 - 2\Psi W - 1)$$

3) Trend model:

$$T(\hat{\rho}_\tau - 1) \xrightarrow{L} \frac{1}{2}(\Gamma - W^2 - 3V^2)^{-1}[(\Psi - 2W)(\Psi - 6V) - 1]$$
$$\hat{\tau}_\tau \xrightarrow{L} \frac{1}{2}(\Gamma - W^2 - 3V^2)^{-\frac{1}{2}}[(\Psi - 2W)(\Psi - 6V) - 1]$$

where "$\xrightarrow{L}$" means "converging in distribution" and

$$\Gamma = \sum_{i=1}^{\infty} \gamma_i^2 \xi_i^2$$
$$\Psi = \sum_{i=1}^{\infty} 2^{1/2} \gamma_i \xi_i$$
$$W = \sum_{i=1}^{\infty} 2^{1/2} \gamma_i^2 \xi_i$$
$$V = \sum_{i=1}^{\infty} 2^{1/2} \left[2\gamma_i^3 - \gamma_i^2\right] \xi_i$$

with $\{\xi_i\}_{i=1}^{\infty}$ being a sequence of independent variables following the N(0, 1) distribution, and

$$\gamma_i^2 = 4[(2i - 1)\pi]^{-2}$$

$$\gamma_i = (-1)^{i+1}\sqrt{\gamma_i^2}$$

P-values can be computed using the above asymptotic distributions. Based on the p-values, we can determine whether to reject the null hypothesis. The following rejection criterion is suggested to be applied.

Method	Criterion
Dickey-Full Test	P-value < 0.05

The Augmented Dickey-Fuller test (Said and Dickey 1984, 1985) is a generalization of the above Dickey-Fuller test. It adjusts for the autocorrelation in the time series by adding lagged first differences. The new models are:

1) Zero mean model: $\nabla Z_t = \rho Z_{t-1} + \sum_{j=1}^{p} \rho_j \nabla Z_{t-j} + \epsilon_t$
2) Constant mean model: $\nabla Z_t = \mu + \rho Z_{t-1} + \sum_{j=1}^{p} \rho_j \nabla Z_{t-j} + \epsilon_t$
3) Trend model: $\nabla Z_t = \mu + \beta t + \rho Z_{t-1} + \sum_{j=1}^{p} \rho_j \nabla Z_{t-j} + \epsilon_t$

The null hypothesis and alternative hypothesis remain the same:

$$H_0: \rho = 1 \; vs \; H_a: |\rho| < 1$$

The Augmented Dickey-Fuller test also uses the ρ and τ statistics. And the distributions of the test statistics are same as those of the Dickey-Full test described above[68]. Again, we suggest using statistical software, such as R or SAS, to calculate the p-value of each test statistic in a practical use. If the null hypothesis is rejected, the time series does not include a unit root and is considered stationary. The following rejection criterion is suggested to be applied.

Method	Criterion
Augmented Dickey-Full Test	P-value < 0.05

Phillips-Perron test[69]

Phillips and Perron (1988) further modified the Dickey-Fuller test to incorporate autocorrelation and heteroscedasticity, and proposed another unit-root test. They also considered 3 models:

1) Zero mean model: $Z_t = \rho Z_{t-1} + e_t$
2) Constant mean model: $Z_t = \mu + \rho Z_{t-1} + e_t$
3) Trend model: $Z_t = \mu + \beta t + \rho Z_{t-1} + e_t$

[68] Refer to SAS/ETS(R) 9.22 User's Guide at http://support.sas.com/documentation/cdl/en/etsug/63348/HTML/default/viewer.htm#etsug_autoreg_sect026.htm.

[69] Refer to SAS/ETS(R) 9.22 User's Guide at http://support.sas.com/documentation/cdl/en/etsug/63348/HTML/default/viewer.htm#etsug_autoreg_sect026.htm.

where ϵ_t is white noise.

The null hypothesis and alternative hypothesis are

$$H_0: \rho = 1 \text{ vs } H_a: |\rho| < 1$$

If the null hypothesis is rejected, the time series does not have a unit root and is stationary.

The Phillips-Perron test has 2 test statistics, $\hat{Z}_\rho$ and $\hat{Z}_\tau$. For the zero mean case, the test statistics are

$$\hat{Z}_\rho = T(\hat{\rho} - 1) - \frac{1}{2}T^2\hat{\sigma}^2(\hat{\lambda} - \hat{\gamma}_0)/s^2$$
$$\hat{Z}_\tau = (\hat{\gamma}_0/\hat{\lambda})^{1/2}\hat{\rho} - \frac{1}{2}T\hat{\sigma}(\hat{\lambda} - \hat{\gamma}_0)/(s\hat{\lambda}^{1/2})$$

Following are the notations for the above equations. Let $\hat{\rho}$ be the OLS estimate of ρ, and let $\hat{e}_t$ denote the OLS residuals. Let $\hat{\sigma}^2$ denote the variance estimate of $\hat{\rho}$. The variance of e_t is estimated by $s^2 = \frac{1}{T-k}\sum_{i=1}^{T}\hat{e}_t^2$. The asymptotic variance of the residual sum of squares $\frac{1}{T}\sum_{i=1}^{T}\hat{e}_t^2$ is

$$\hat{\lambda} = \sum_{j=0}^{l}\kappa_j[1 - j/(l+1)\,\hat{\gamma}_j]$$

where $\kappa_0 = 1$, $\kappa_j = 2$ for $j > 0$, $\hat{\gamma}_j = \frac{1}{T}\sum_{t=j+1}^{T}\hat{e}_t\hat{e}_{t-j}$, and l is the truncation lag.

Note that the $\hat{Z}_\rho$ statistic is just the ordinary Dickey-Fuller statistic with a correction term for the autocorrelation. The correction term goes to zero asymptotically is there is no autocorrelation.

We suggest using standard statistical software, such as R or SAS, to calculate the p-value of the test statistic in a practical use. For the validator's information, we provide the asymptotic distributions of the 2 test statistics here:

1) Zero mean model:

$$\hat{Z}_\rho \Rightarrow \frac{\frac{1}{2}\{B(1)^2 - 1\}}{\int_0^1 [B(s)]^2 ds}$$
$$\hat{Z}_\tau \Rightarrow \frac{\frac{1}{2}\{B(1)^2 - 1\}}{\left\{\int_0^1 [B(s)]^2 ds\right\}^{1/2}}$$

2) Constant mean model:

$$\hat{Z}_\rho \Rightarrow \frac{\frac{1}{2}\{B(1)^2 - 1\} - B(1)\int_0^1 [B(s)]^2 ds}{\int_0^1 [B(s)]^2 ds - \left\{\int_0^1 B(s) ds\right\}^2}$$

$$\hat{Z}_\tau \Rightarrow \frac{\frac{1}{2}\{B(1)^2 - 1\} - B(1)\int_0^1 [B(s)]^2 ds}{\left\{\int_0^1 [B(s)]^2 ds\right\}^{1/2} - \left\{\int_0^1 B(s) ds\right\}^{1/2}}$$

3) Trend model:

$$[0 \quad c \quad 0]V^{-1}\begin{bmatrix} B(1) \\ (B(1)^2 - 1)/2 \\ B(1) - \int_0^1 B(s) ds \end{bmatrix}$$

with $c = I$ for $\hat{Z}_\rho$ and $c = \frac{I}{\sqrt{Q}}$ for $\hat{Z}_\tau$,

where

$$V = \begin{bmatrix} 1 & \int_0^1 B(s) ds & 1/2 \\ \int_0^1 B(s) ds & \int_0^1 [B(s)]^2 ds & \int_0^1 sB(s) ds \\ 1/2 & \int_0^1 sB(s) ds & 1/3 \end{bmatrix}$$

$$Q = [0 \quad c \quad 0]V^{-1}\begin{bmatrix} 0 \\ c \\ 0 \end{bmatrix}$$

and $B(.)$ is a standard one-dimensional Brownian motion.

P-values can be computed using the above asymptotic distributions. Based on the p-values, we can determine whether to reject the null hypothesis. The following rejection criterion is suggested to be applied.

Method	Criterion
Augmented Dickey-Full Test	P-value < 0.05

Residual Normality

Purpose: Linear regression inference assumes that the error term follows a standard normal distribution. If it is not normally distributed, the p-value for the inferences can be inaccurate.

Normal Q-Q plot[70]

[70] Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models.* Fifth Edition. New York: McGraw-Hill/Irwin, 2005

Fox, John. Applied regression analysis and generalized linear models. Second Edition.Sage, 2008.

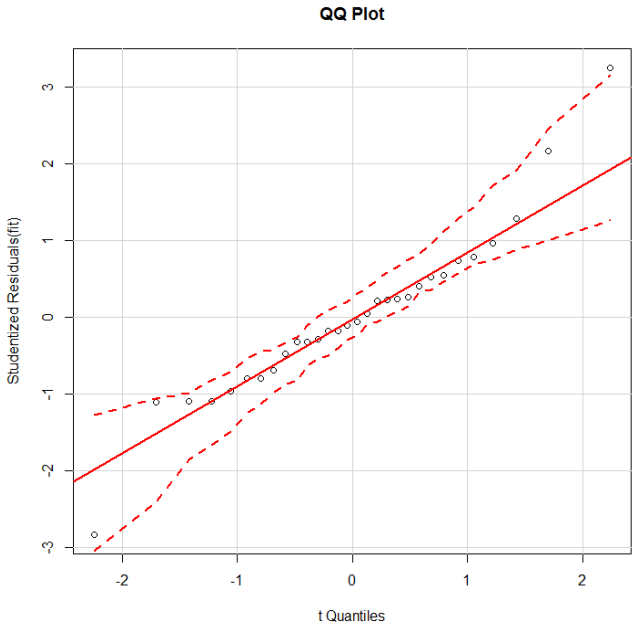

Figure 16 Example of QQ Plot

Normality of Residuals can be identified using a quantile-comparison plot, known as a Q-Q plot. The Q-Q plot visually compares the cumulative distribution of the (standardized or studentized) residuals to a cumulative reference distribution. The points should be distributed close to the diagonal line. If non-normality exists, the plot may show a pattern. For example, concave-downward curvature at the left end indicates left-skewed distribution; concave-upward plot at the right end indicates right-sked distribution.

To construct a QQ plot for residuals, consider studentized residuals, $\tilde{e}_{(1)} < \tilde{e}_{(2)} < \cdots < \tilde{e}_{(T)}$ sorted in increasing order. Studentized residuals are defined as,

$$\tilde{e}_t = \frac{e_t}{\sqrt{MSE(1-h_t)}}$$

where e_t is residuals $Y_t - \hat{Y}_t$, MSE is mean square error from the model and h_t is defined as

$$h_t = \frac{1}{T} + \frac{(X_t - \bar{X})^2}{\sum(X_t - \bar{X})^2}$$

For normal quantiles, we have, for $t = 1, \ldots, T$

$$\mathrm{F}_t = \Phi\left(\frac{t - 0.375}{n + 0.25}\right)$$

where $\Phi(.)$ is cumulative distribution function of standard normal distribution. Then plot F_t versus $\tilde{e}_{(t)}$. Instead of standard normal distribution, one can use t-distribution with degrees of freedom (T-p-

1). For residuals, one can also use standardized residuals ($e_t/\sqrt{MSE}$) or raw residuals. In Figure 1, R software package 'car', is used to compare studentized residuals to the t-distribution.

Error! Reference source not found. illustrates an example of QQ plot. Data points are scattered round the diagonal line. The two dotted lines represent the "confidence envelope" suggested by Atkinson (1985). Taking into account the correlational structure of the explanatory variables, simulated sampling is employed to construct the confidence envelope. A weakness of this method is that the probability of the data points appearing outside of the envelope is greater than α (e.g. 0.05). Thus, one should not count the fraction of data points outside the interval; rather one should focus on the overall data-scatter pattern, which should be linear and matched in slope with the diagonal.

Anderson-Darling test[71]

Goodness of fit tests, such as the Anderson-Darling test (AD test) can be used to test the normality of the error term. The test measures distance between the empirical distribution of the data and a given cumulative distribution function that you are assuming; in this case, the normal distribution.

Description: Section 4.8.5. in (D'Agostino and Stephens) discussed specifically how to test residuals for normality. Although residuals are not independent, asymptotic distribution of the statistics below is the same as when we test normality with unknown mean and variance.

They suggested to use studentized residuals with $Z_{(t)} = \Phi(\tilde{e}_{(t)})$ where $\tilde{e}_{(t)}$ is a studentized residual.

Consider $Z_{(1)} < Z_{(2)} < \cdots < Z_{(T)}$ and $\bar{Z} = \sum Z_{(t)}/T$,

$$A^2 = -T - \frac{1}{T}\sum_t (2t-1)[\log Z_{(t)} + \log\{1 - Z_{(T+1-t)}\}]$$

The following criterion can be applied to reject the null hypothesis that the error term is normal distributed [See Table 4.7 in (D'Agostino & Stephens, 1986)].

Method	Criterion
Anderson-Darling test	$A^2 > 0.752$

Residual Stationarity

Purpose: Linear regression assumes that the errors are independent. In applying a linear regression model to time series, this often implies that the residual series is stationary. Nonstationarity in the residuals results in violation of the linear regression assumption and thus inappropriate inference.

[71] R.B. D'Agostino and M.A. Stephens, Goodness-of-fit techniques, Marcel Dekker, Inc., 1986.

Description: The principles as described in the earlier section on general stationarity also apply here. There are multiple tools for examining stationarity, including plots of the time series and its differences, sample autocorrelation function (ACF) of the time series and its differences, informal inspection of the estimated coefficients of the model, and formal statistical tests. Although there are advantages of every tool and we encourage using more than one tool, the formal statistical tests should be performed at the minimum. They are more rigorous and provide quantitative inferences. We introduced two statistical tests: the Augmented Dickey-Fuller test and Phillips-Perron test. Details of these tests can be found in the earlier section.

Residual Serial Correlation

Purpose: When there is positive autocorrelation among error terms, the estimates of the variance for coefficient estimates is biased downwards, meaning that the precision of the estimates is overstated; we might wrongly conclude statistical significance.

Durbin-Watson[72]

The Durbin-Watson test is a widely used method of testing for autocorrelation. The first-order autoregressive error models assumes

$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

where u_t is white noise.

In order to test whether autocorrelation parameter ρ is zero, the null hypothesis is

$$H_0: \rho = 0$$

and the alternative hypothesis is, for testing positive autocorrelation (which business and economic application tends to show),

$$H_a: \rho > 0$$

The test statistics is:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

where e_t is residuals of the model.

Exact critical values are difficult to obtain, but Durbin-Watson obtained lower bound d_L and upper bound d_u for a decision rule[73].

- If $DW > d_u$, conclude H_0
- If $DW < d_L$, conclude H_a

[72] Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models.* Fifth Edition. New York: McGraw-Hill/Irwin, 2005

[73] The table for the upper and lower bound is in Table B.7 of (Kutner et al, 2005)

- If $d_L \leq DW \leq d_u$, the test is inconclusive.

For higher order autocorrelation testing[74], we can assume j-th order autoregressive processes

$$\epsilon_t = \rho_j \epsilon_{t-j} + u_t$$

with similar test statistics,

$$DW_j = \frac{\sum_{t=j+1}^{n} (e_t - e_{t-j})^2}{\sum_{t=1}^{n} e_t^2}$$

The tests can be performed sequentially; for j-th order autocorrelation, test $H_0: \rho_j = 0$ given $\rho_1 = \cdots = \rho_{j-1} = 0$. For positive autocorrelation ($\rho_j > 0$), refer to the p-value, which is the probability that we get test statistic less than DW_j. For negative autocorrelation ($\rho_j < 0$), refer to p-value which is the probability that we get a test statistic greater than DW_j.

The p-value can be calculated in statistical software such as SAS. For more details on the distribution of statistics, see the SAS User's Guide (http://support.sas.com/documentation/cdl/en/etsug/63348/HTML/default/viewer.htm#etsug_autoreg_sect026.htm)

The following rejection criterion is suggested to be applied.

Method	Criterion
Durbin-Watson Test	p-value < 0.05

Residual Heteroskedasticity[75]

Purpose: One of the assumptions of linear regression model is that the variances of error term are the same for all observations. When the assumption is violated, although the coefficient estimates are still unbiased, standard errors for coefficient estimates are not accurate.

Residual plot

One way to visualize the presence of heteroskedasticity is to plot studentized residuals versus the predictor(s) or versus the fitted values. If the error terms all have the same variance one would expect to see that the points are scattered randomly around zero. For example, if there is a pattern such as

[74] Refer to SAS/ETS(R) 9.22 User's Guide at http://support.sas.com/documentation/cdl/en/etsug/63348/HTML/default/viewer.htm#etsug_autoreg_sect026.htm.

[75] Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models.* Fifth Edition. New York: McGraw-Hill/Irwin, 2005
Fox, John. Applied regression analysis and generalized linear models. Second Edition.Sage, 2008.

a fan-shape, it indicates heteroskedasticity, since the variance of residuals is related to predictors or the response.

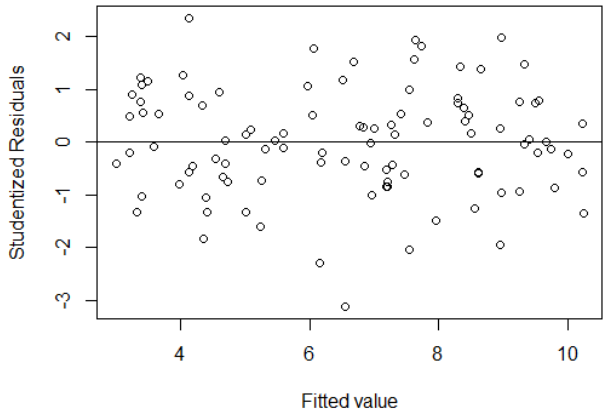

Figure 17 Studentized residuals versus the fitted values

Breusch-Pagan test

Breusch and Pagan (1979)[76] developed a score test for heteroscedasticity based on the assumption

$$\sigma_t^2 = g(\gamma_0 + \gamma_1 Z_{1t} + \cdots + \gamma_p Z_{kt})$$

where $Z_1, \ldots, Z_k$ are known variables and function g is quite general. Choice of $Z_1, \ldots, Z_k$ and function g depends on the suspected pattern that we can observed from residual plots. One simple way is to test whether the variance of error terms is related to the level of explanatory variables X. Hypothesis testing with $H_0: \gamma_1 = \cdots = \gamma_p = 0$ for a linear regression model

$$\sigma_t^2 = \gamma_0 + \gamma_1 X_{1t} + \cdots + \gamma_p X_{pt}$$

is performed fitting a regression model of e_t^2 against predictors. Note that e_t^2 is an estimate for σ_t^2. The test statistic $X_{BP}^2$is defined as

$$X_{BP}^2 = \frac{SSR^*}{2} / \left(\frac{SSE}{n}\right)^2$$

where SSR^* is the regression sum of squares when regressing e_t^2 against predictors and SSE is the error sum of squares of the main model regressing Y against predictors. If the null hypothesis is true and sample size is reasonably large, the statistic follows approximately Chi-square distribution with degrees of freedom p, the number of predictors.

The following criterion can be applied to reject the null hypothesis.

[76] Breusch, T. S., and A. R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation". Econometrica 47 (5). The Econometric Society: 1287–94.

Method	Criterion
Breusch-Pagan test	$X^2_{BP} > \chi^2(0.95; df = p)$

Similarly, Cook and Weisberg (1983)[77] suggested regressing e_t^2 against fitted values from the main model and the statistics will follow approximately Chi-square distribution with degrees of freedom 1.

System Level Uncertainties in Prediction

Chained Confidence Intervals

We have noted that a number of important stress test models (such as EFS) have cascaded or chained regressions. By this we mean there is a regression equation for prediction of one variable that is then the input to another regression equation. In this case, looking at the Rsquared and MAPE of a single regression, while valuable, may not be adequate.

The question of how good a regression needs to be is best answered by stating the requirements of accuracy in the stressed output. Say you want to estimate cash flows in a stress scenario to within +/- 20%. In the case of a single regression formula for stress testing, then looking at things like R squared and MAPE (mean absolute percentage error) may be adequate. However, if the stress test model contains cascaded regression models (where one regression result is input to another regression equation, such as EFS stress testing), then more care needs to be devoted. Here we need to think about the prediction error of the 1 stage regression. Defining the prediction error for Y* at X* as:

$$\mathrm{se}_{\mathrm{pred}}(\widehat{Y}_*) = \widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_* - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}.$$

The 1-α confidence for Y* is:

$$\widehat{Y}_* \pm z_{\alpha/2}\mathrm{se}_{\mathrm{pred}}(\widehat{Y}_*).$$

http://www.stat.cmu.edu/~roeder/stat707/lectures.pdf

Example for two concatenated regressions

$$Y_1 = \beta_1 x_1 + C_1$$

[77] Cook, R. Dennis, and Sanford Weisberg. "Diagnostics for heteroscedasticity in regression." Biometrika 70, no. 1 (1983): 1-10.

$$Y_2 = \beta_2 x Y_1 + C_2$$

First we want to examine uncertainty in Y1pred through use of the above equations, to arrive at a Ypred+ and a Ypred- (the upper and lower confidence bounds). Next we would like to substitute each value of Y1 into the second regression equation and calculate uncertainties, again using the equations above. So plugging in Y1pred+ into second regression with uncertainty we would get a Y2pred++ and a Y2pred-+. The notation here is first +/- sign is the high and low confidence bounds on Y2, and the second +/- sign is the high and low confidence bands on Y1 plugged into second regression equation. Likewise we would plug in Y1pred- to get Y2pred+- and Y2pred--. Then we would take the min and max of the four Y2pred and quantify that as our uncertainty and compare to our requirements or original goal.

While this method is conceptually simple to follow, it most likely over-estimates the uncertainty, since drawing from two 90% confidence intervals results in an overall confidence level higher than 90%. A more rigorous and correct way to quantify the uncertainty in the chained regressions is to examine the sum of the overall variances.

System Level Sum of Variances

Regression model for *Y*(*X*)

Suppose *Y* is a response and *X* is the only known factor, so *Y* is a function of *X*. We perform a linear regression, based on *n* random samples of (*X, Y*), and determine the two coefficients:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Where ε is the error. Satisfactorily we find that the error is normally distributed about zero, with no correlation between *X* and ε.

$$\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

The unbiased sample variance s^2 is:

$$s^2 = \frac{\sum_{k=1}^{n} \left(Y_k - \beta_0 - \beta_1 X_k\right)^2}{n-1}$$

The expected value of the unbiased sample variance:

$$\mathbf{E}\left(s^2\right) = \sigma^2$$

This would be the realized value for increasingly large *n*. We can do no better than to use the unbiased sample variance as an estimate for the population variance σ^2.

Regression model for $Y([x])$

We have assumed that Y is a function of a single factor X. This scenario is easily generalized to the case where Y is a function of a set of q factors x_i.

$$Y = \beta_0 + [\beta]^T [x] + \delta$$

$$[\beta] = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix} \quad [x] = \begin{bmatrix} x_1 \\ \vdots \\ x_q \end{bmatrix}$$

The unbiased sample variance becomes:

$$s^2 = \frac{\sum_{k=1}^{n} \left(Y_k - \beta_0 - [\beta]^T [x]_k \right)^2}{n-1}$$

Regression model for $W(Y)$

Suppose that W is a response and Y is the only factor. We perform a linear regression, based on m random samples of (Y, W), and determine the two coefficients.

$$W = \alpha_0 + \alpha_1 Y + \delta$$

Where δ is the error. Satisfactorily we find that the error is normally distributed about zero, with no correlation between Y and δ.

$$\delta \sim \mathrm{N}\left(0, \rho^2\right)$$

The unbiased sample variance r^2 is:

$$r^2 = \frac{\sum_{k=1}^{m} \left(W_k - \alpha_0 - \alpha_1 Y_k \right)^2}{m-1}$$

The expect value of the unbiased sample variance:

$$\mathbf{E}\left(r^2\right) = \rho^2$$

This would be the realized value for increasingly large m. We can do no better than to use the unbiased sample variance as an estimate for the population variance ρ^2.

In developing this regression model for W, we have assumed that for each measured value of W we have measured a value of Y. In fact, if Y is a modeled quantity, as described above, there is a random error ε associated with that prediction. If we include this additional source of uncertainty, then the expected value of the error, to a first order approximation, is:

$$\delta' \sim \mathbf{N}\left(0, \rho^2 + \left(\frac{\partial W}{\partial Y}\right)^2 \sigma^2\right) \approx \mathbf{N}\left(0, \rho^2 + \alpha_1^2 \sigma^2\right) \approx \mathbf{N}\left(0, r^2 + \alpha_1^2 s^2\right)$$

Regression Model for $W([y])$

Suppose again that W is the response, but there are p factors y_i, $i = 1, \ldots, p$. We perform a linear regression, based on m random samples (W, y_1, …, y_p), and determine the $1 + p$ coefficients:

$$W = \alpha_0 + [\alpha]^T [y] + \delta$$

$$[\alpha] = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} \quad [y] = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}$$

Where δ is the error. Satisfactorily we find that the error is normally distributed about zero, with no correlation between y and δ.

$$\delta \sim \mathrm{N}\left(0, \rho^2\right)$$

The unbiased sample variance r^2 is:

$$r^2 = \frac{\sum_{k=1}^{m} \left(W_k - \alpha_0 - [\alpha]^T [y]_k\right)^2}{m-1}$$

The expect value of the unbiased sample variance:

$$\mathbf{E}\left(r^2\right) = \rho^2$$

This would be the realized value for increasingly large m. We can do no better than to use the unbiased sample variance as an estimate for the population variance ρ^2.

In developing this regression model for W, we have assumed that for each measured value of W we have measured a value of $[y]$. What if, in fact, each component y_i of $[y]$ is an independently modeled quantity, from a linear regression, and for each regression there is a random, normally distributed error ε_i associated with that prediction.

$$\varepsilon_i \sim \mathrm{N}\left(0, \sigma_i^2\right), \quad i = 1, \ldots, p$$

If we include this additional source of uncertainty, then the expected value of the error, to a first order approximation, is:

$$\delta' \sim \mathbf{N}\left(0, \rho^2 + \sum_{i=1}^{p} \left(\frac{\partial W}{\partial y_i}\right)^2 \sigma_i^2\right) \approx \mathbf{N}\left(0, \rho^2 + \sum_{i=1}^{p} \alpha_i^2 \sigma_i^2\right) \approx \mathbf{N}\left(0, r^2 + \sum_{i=1}^{p} \alpha_i^2 s_i^2\right)$$

Model Capability

Several model capability indices are available. If we know the specification limits as a range, but we do not have a target value for *W*, we can use:

$$\hat{C}_M = \frac{\Delta_{SL}}{6\sqrt{r^2 + \sum_{i=1}^{p} \alpha_i^2 s_i^2}}$$

Δ_{SL} represents the tolerable range of *W,* expressed in absolute units such as $s and it is the full range (if +/- 100$, then it would be 200$). This equation optimistically assumes that the mean response of the regression equation for *W* is accurate. Model capabilities of 1.33 to 2 are reasonable; higher values are rare, and lower values are unsettling.

Outcome Analysis

Since these are stress testing models that are designed to predict an extreme scenario that does not typically occur, the typical type of back-testing (prediction versus actual) can't be performed. However, we still advise that some form of outcome analysis can be performed. Here we can take the macro variables that actually did occur and put them into the model, and then compare the model predictions with the actual business outcomes. So in that sense the model can be back tested in the observed range of macro variables.

Another problem with stress testing models is that often the training data is in a narrow range of macro variables compared to the X's going into the stress scenario. There is always risk in using a model far outside it's training range. As part of outcome analysis we can quantify the uncertainty range or prediction error using the equations in the "Chained Confidence Intervals" section of the report. These are designed to take into account that at the prediction point, X_*, the value of X is far from the mean of the training data, and will estimate higher uncertainty to account for this extrapolation. So the quantification of the prediction error can capture the model risk from using a model outside of the training data range.

It should be stated that the above assumes that the model is 'correct'. Meaning if it is a linear regression model that the relationship is indeed linear. If the assumption of linearity of the model is only accurate in the limited training range and not at the wider stress usage range, then the errors could be considerable larger than the prediction error equations would predict. Given the limited nature of the training data, it is hard to know if the model is correct, and there is always some remaining model risk. The above concept is highlighted in the graphic below.

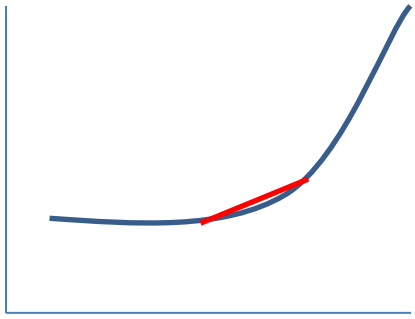

Figure 18 Example of 'wrong' model. Red section is linear fit to limited training data, blue is real relationship.

Value at Risk

Value at Risk

Michael Vallance, Jerry Cline and Weiwei Shen

Contents

1 Value at Risk

The value-at-risk (VaR) measure is summarized by the following statement:[78]

> *I am X percent certain that there will not be a loss of more than VaR dollars in the next N (business) days.*

A compact notation is VaR_X^N. VaR_X^N is routinely estimated using the probability density function of the portfolio-value change over the next *N* days.

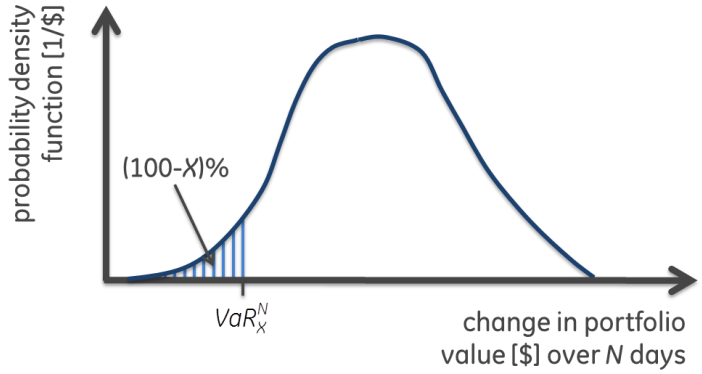

Figure 19: Graphical depiction of VaR.

The *1996 Amendment* to Basel I (implemented in 1998) requires financial institutions to hold capital for market risk as well as credit risk. The amendment distinguishes between the institution's trading book and its banking book. The banking book consists primarily of loans (and leases), and is usually revalued on a regular basis for managerial and accounting purposes. Where these assets are held to maturity, credit risk principally underlies the value at risk. On the other hand, the trading book consists of instruments that are traded (equities, bonds, swaps, forward contracts, options, etc.), and is revalued daily. In the case of the trading book, market risk principally underlies value at risk. For the trading book, the *amendment* requires the institution to maintain capital $\geq 3 \cdot VaR_{99}^{10}$.

Basel II (implemented in 2007) uses $VaR_{99.9}^{252}$ as the relevant measure for credit (and operational) risk. Further modifications to VaR calculation were introduced with Basel II.5 and Basel III.

The Basel Committee on Bank Supervision is now considering alternatives to VaR. Quoting Malz:[79]

> *VaR has dreadful limitations, both as a model and as a practice of risk managers in real-world applications. VaR has come under ferocious, and to a large extent justified, attack from many quarters.*

[78] J. Hull, *Fundamentals of Futures and Options Markets*, **8th** Ed, Pearson (2014).
[79] A. Malz, *Financial Risk Management, Models, History, and Institutions*, Wiley (2011).

Market-risk-based VaR estimates are often calculated as VaR_X^1. One-day VaR is convenient, because the amount of historical data for non-overlapped, one-day periods is typically large enough to calculate compelling value distributions. Having calculated VaR_X^1, VaR_X^N can be approximated as:

$$VaR_X^N = \sqrt{N} \cdot VaR_X^1 \tag{1}$$

This formula is exactly true when the changes in the value of the portfolio on successive days have independent, identical, normal distributions with means of zero.

In the case of market-risk-based VaR, changes in portfolio value are always linked to changes in one or more, liquid, market variables (*a.k.a.* risks), *e.g.*, exchange rates, interest rates and commodity prices.

2 Mathematical Modeling of Value at Risk

Three approaches are used to compute VaR; there are variations on each approach.

- Parametric approach
- Monte Carlo approach
- Historical approach

In the parametric approach, VaR_X^N is computed from an algebraic model of the portfolio value's variance as a function of the variances and covariances of the market-quoted variables, assuming a parametric (almost always normal), multivariate, probability density function (pdf) for changes in the market variables during the *N* days associated with VaR_X^N, The variances and covariances for the *N*-day changes in the market variables are assigned, typically based on historical data.

In the Monte Carlo approach, VaR_X^N is computed from a model of the portfolio value as a function of the market-quoted variables. As before, a parametric, multivariate, pdf is estimated for the *N*-day changes in the market-quoted variables. Hundreds or thousands of random variable draws are made from the pdf, and used to value the change in portfolio value over *N* days, creating a portfolio-value-change distribution from which VaR_X^N is estimated.

In the historical approach, VaR_X^N is again computed from a model of the portfolio value as a function of the market-quoted variables. *N*-day changes in the market-quoted variables, are sampled from a large number of historical *N*-day periods, and then used to predict changes in the present portfolio's value. By using multiple, independent, historical samples, thereby building up a distribution of portfolio value changes, VaR_X^N can be estimated.

For each of the three approaches, the validity tests are different. While numerous variants exist for each approach, we focus on the common protocols. Each approach includes assumptions about the

behavior of the market-quoted variables. Because studies have shown violations of the assumptions in most markets,[80] the choice of method depends on the modeler's judgment.

3 Quantitative Validation of VaR Models

Quantitative validation focuses on:[81]

- Sensitivity (to assumptions) testing,
- Back testing
- Stress testing

A portfolio-valuation sub-model underlies each of the above-mentioned VaR modeling approaches. Validation of the portfolio valuation sub-model will ***not*** be addressed rigorously in this chapter, since we have described techniques for the quantitative validation of portfolio valuation models previously.[82,83] We described the use of model replication, mark-to-market, and benchmark testing as key aspects of quantitative validation. We introduced the concept of model capability C_{MK}, a statistical tool that can be used to evaluate the portfolio valuation model. We also introduced several sensitivity measurements specific to interest rate curve building, and we proposed methods to evaluate the effectiveness of Monte Carlo simulation, as is often used in the valuation of financial instruments with optionality.

3.1 Replication

Replication is the validator's tool of choice for identifying model risk from Derman categories 3 -6:[82]

1. Correct model, incorrect mathematical representation (erroneous mathematical transformations).
2. Correct model, inappropriate use (poor calibration).
3. Badly approximated solution (inaccurate numerical methods, e.g., integration).
4. Software bugs.

As such, we advise the use of replication to assure that the production model complies with the model documentation, and that the numerical methods for calibration, integration and interpolation are well executed. As described in references 82 and 83, and restated above, replication should be used to validate the computational accuracy of the underlying portfolio valuation model.

Replication of the VaR model itself, the computational algorithms used to generate the pdf of future portfolio value change, may not be justified. Needless to say, if the output of the VaR model cannot be

[80] P. Dobránszky, Comparison of Historical and Parametric Value-at-Risk Methodologies, Social Science Research Network, (September 2009) https://papers.ssrn.com/sol3/Data_Integrity_Notice.cfm?abid=1508041 (downloaded on 15.09.04).
[81] G. Conn, H. Hibbert, S. Sharp and C. Trunbull, Validation of risk factor modelling in 1-year VaR capital assessments, Moody's Analytics (April 2013).
[82] M. Vallance, J. Cline and W. Shen, Chapter 1: Model Risk Management (August, 2015).
[83] M. Vallance, J. Cline and W. Shen, Chapter 2: Valuation of Interest-Rate-Based Financial Instruments (August, 2015).

rationalized, based on sensitivity testing, back testing or stress testing, then such replication may be required.

Validators will often calculate VaR in the course of sensitivity testing, using the existing production model to do so.[84] Such an exercise might be used to evaluate the impact of an alternative asset pricing-model calibration or an alternative treatment of the correlation between the prices of related assets.

3.2 Tests for the Parametric (Variance-Covariance) Approach

The advantage of the variance-covariance method is its simplicity, based upon the assumption that *N*-day changes in market-quoted variables are normally distributed.

Assumption testing includes:

- Normality testing for standardized risk factors: standard normality testing can be applied, *e.g.*, QQ-plot, Shapiro-Wilk, or Jarque-Beta. In Figure 20, the daily change in the discount rate for 52-week treasury bills is plotted for all of the business days between June 24 and September 3, 2015. Three different normality tests available in Minitab 17 have been used. At 95% confidence, we cannot reject the null hypothesis of normality, based on two of the three test statistics, although the right-hand tail is fat.

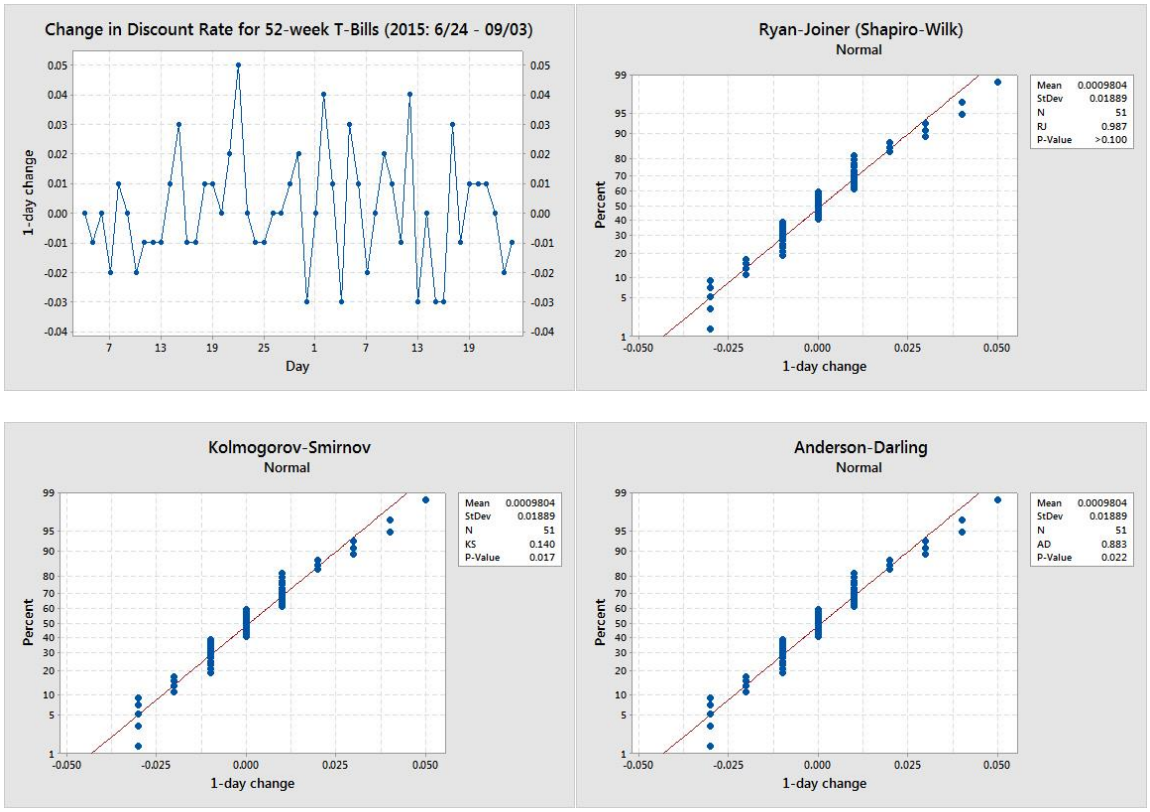

Figure 20: Normality testing of recent changes in the 52-week T-bill discount rate. Of the three tests, only the Ryan-Joiner statistic indicates that we should reject the null hypothesis at 95% confidence.

[84] N. Sinatra, G. Yates, M. Das and R. Hornick, Validation of EFS Commodity Value at Risk Model, GECC Model Validation Document (June 2015).

- Normality of changes in portfolio value: If the changes in the underlying market factors follow a multivariate normal distribution, then so too do the changes in the portfolio value, if the relationship between the market factors and the portfolio value is linear. The changes in value for assets with optionality are not linearly related to the changes in the underlying market factors; even if changes in the underlying factors are normally distributed, the same is not true for the asset values, or for portfolios containing such assets. Changes in portfolios with substantial optionality positions are better evaluated using the Monte Carlo or historical approaches.
- Autocorrelation in regression residuals: When historical data are used to generate the multivariate pdf for the market factors, regression is used to calculate the covariances between the market factors. If the residual signal from a linear regression, relating the time series of changes in two market variables, is cross-correlated with its own lagged values, then the regression model used to estimate the correlation coefficients between the changes in multiple market variables may not be valid. The Durbin-Watson test is appropriate. The Ljung-Box test, based on the autocorrelation function (ACF), and normally used for ARIMA modeling, can be adopted as an alternative. Figure 21 shows the correlation between interest rate daily changes for 20-year and 2-year Treasuries during 2015.

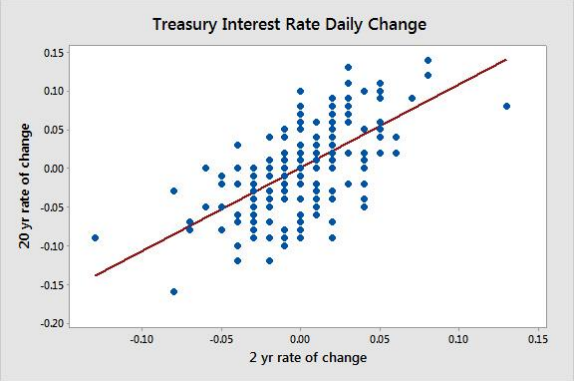

Figure 21: Plot of the daily rate of change for 20-year Treasury bond interest rates as a function of the daily rate of change of 2-year Treasury bond interest rates for the period of time from January 2 to September 4, 2015.

To test for autocorrelation, we use the Durbin-Watson test, as provided by Minitab 17. The Durbin-Watson statistic is calculated as 2.134. We look up the critical values at 95% confidence, using K = 2 (number of parameters in the regression curve) and T = 171 (number of data pairs used in the regression). The upper and lower critical values are 1.762 and 1.738.[85] Since the statistic is above the upper critical value, we are assured that autocorrelation is not present.

[85] http://web.stanford.edu/~clint/bench/dwcrit.htm (downloaded on 2015.09.08).

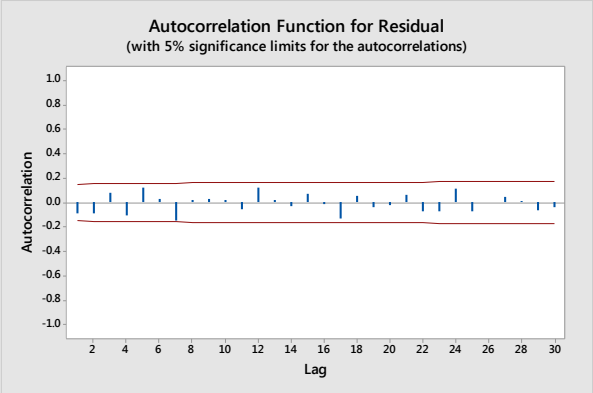

Figure 22: The ACF for the residuals of the regression model of Figure 21. The lag is in (business) days.

The residuals from the regression model fit of **Figure 21** were analyzed for autocorrelation using Minitab 17. The ACF is plotted in **Figure 22** for lags of 1 to 30 days. There is no autocorrelation at 95% confidence.

- Stationarity of statistics for changes in market variables: The augmented Dickey-Fuller test, the Phillips-Perron test, and the KPSS test can be used to test stationarity. Using 10-(business) day samples, variances of the changes in interest rate for 20-year treasury bonds were calculated for sequential periods throughout 2015 until the beginning of September. The variance time series is shown in **Figure 23**.

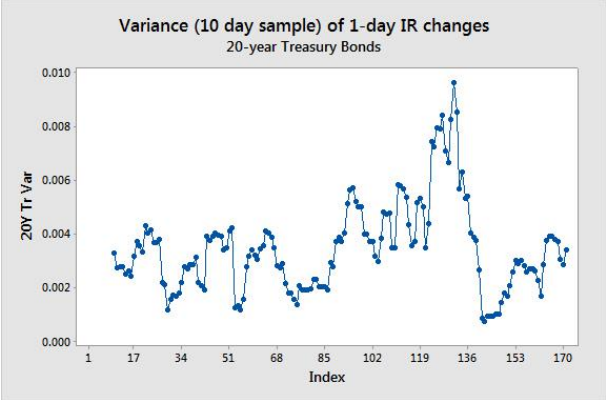

Figure 23: Time series of variances for the daily change in interest rate for 20-year Treasuries. The sample size is 10 days. The variance is calculated at the end of each business day, January 16 until September 4, 2015, using lagging daily interest rate changes.

The time series of variances were imported into R as a vector, and the augmented Dickey-Fuller statistic was calculated using adf.test from the tseries package. With p = 0.1839, the null hypothesis of non-stationary could not be rejected at 95% confidence.

- Translating $VaR_X^{N_1}$ to $VaR_X^{N_2}$: This can be done using equation 1, only if the assumptions associated with that equation are true. If those assumptions are true, additional translations are possible:

$$VaR_X^{N_1} = \sqrt{\frac{N_1}{N_2}} \cdot VaR_X^{N_2}$$
$$VaR_{X_1}^{N} = \frac{z(X_1)}{z(X_2)} \cdot VaR_{X_2}^{N} \tag{2}$$

3.3 Tests for Monte Carlo Simulation

The advantage of Monte Carlo simulation is its flexibility to incorporate nonlinearity between the market variables and the portfolio value. Given a good valuation model, we only need to ensure that the simulation error is small. However, this is challenging. If we have used the Monte Carlo simulation method with *n* draws to estimate $VaR_{X,n}^N$, by the central limit theorem:

$$\lim_{n\to\infty} VaR_{X,n}^N \in N\left(VaR_X^N, \frac{\sigma^2}{n}\right) \tag{3}$$

$N(\cdot,\cdot)$ is the normal distribution. Using the Bahadur representation:[86]

$$\sigma^2 = \frac{X(1-X)}{f\left(VaR_X^N\right)} \tag{4}$$

f is the pdf of loss/return, which will not be known with adequate precision. Given a good approximation of *f*, we could construct a confidence interval for $VaR_{X,n}^N$. Because we are interested in values of *X* near 1 (100%), $f\left(VaR_X^N\right)$ is very small, making σ^2 very large, suggesting that our confidence intervals will be wide; this is essentially the tail estimation problem. Therefore, any VaR computation based on Monte Carlo simulation without accompanying error estimation should raise a red flag.

If the model developer has not provided an estimate of error, we urge the validator to verify that the developer has used a sufficiently large number of draws *n*, and we provide a procedure, as described in reference 9. This procedure uses a jackknife estimator and sectioning. To produce a 100 (1 - δ) % confidence interval for VaR_X^N:

1. Select a sample size (number of draws) $n = m \cdot k$ with $10 \le m \le 20$.
2. Generate *n* independently and identically distributed draws of the simulation experiment.
3. Compute the sample estimator $VaR_{X,n}^N$ based on all *n* observations.
4. For each $i \in \{1, \ldots, m\}$, compute the sample $\underline{VaR}_{X,n}^N(i)$ using all observations, except those associated with the i^{th} block of *k* draws; *i.e.*, do not use draws replications indexed from $(i-1)k+1$ through ik.
5. Compute the *m* pseudo-values: for $1 \le i \le m$,

[86] P. Haas, Simulation Lecture Notes #9: Quantile Estimation, MS&E 223, Spring Quarter 2005-6, http://web.stanford.edu/class/msande223/handouts/lecturenotes09.pdf (2015.09.09).

$$VaR_X^N(i) \equiv m \cdot VaR_{X,n}^N - (m-1) \cdot \underline{VaR}_X^N(i) \tag{5}$$

6. Set

$$\begin{aligned} VaR_X^{N,J} &\equiv \frac{1}{m}\sum_{i=1}^{m} VaR_X^N(i) \\ v^J &\equiv \frac{1}{m-1}\sum_{i=1}^{m}\left[VaR_X^N(i) - VaR_X^{N,J}\right]^2 \end{aligned} \tag{6}$$

7. Compute the 100 (1 - δ) % confidence interval as

$$\left[VaR_X^{N,J} - t\sqrt{\frac{v^J}{m}},\ VaR_X^{N,J} + t\sqrt{\frac{v^J}{m}}\right] \tag{7}$$

   t is the 1 - $\delta/2$ quantile of the Student-t distribution with (m – 1) degrees of freedom.

8. If a 100 (1 - δ) % confidence interval width of D (in dollars) is desired, then the number of draws should be:

$$n = \frac{4t^2 k v^J}{D^2} \tag{8}$$

3.4 Tests for the Historical Approach

The historical approach requires no distributional assumptions about the changes in the market-quoted variables. The approach assumes that history repeats itself (stationarity, backwards and forwards). While both basic bootstrap resampling and block bootstrap resampling are relevant, the validation approaches differ. Block bootstrap resampling is used when one or more historical market factors demonstrate autocorrelation.

Tests include:

- Autocorrelation of portfolio returns: During backtesting, if VaR estimates are conditionally correct, then the fact that losses exceed VaR during a particular interval should have no predictive power regarding future excess losses, if market variable changes have been randomly sampled with replacement (basic bootstrap).[87] In section 3.1 we describe testing for autocorrelation.
- Stationarity of portfolio returns: In the case of basic bootstrap resampling, but not in the case of block bootstrap resampling, the returns are expected to be stationary. We have described tests for stationarity in section 3.1. Even when block resampling is used, to capture sequence information, we should demand an overall stationary response. This problem has been addressed by Politis and Romano.[88]
- Fidelity of resampled market factor changes and or portfolio returns: The Kolmogorov-Smirnov (K-S) test is used to compare two samples (or a sample and a given distribution)

[87] M. Pritsker, The Hidden Dangers of Historical Simulation (June 2001) http://www.federalreserve.gov/pubs/feds/2001/200127/200127pap.pdf (2015.09.09)
[88] D. Politis and J. Romano, The Stationary Bootstrap, *J. Am. Stat. Assn* **89(**428) (1994) 1303-1313.

against the null hypothesis that the two samples are drawn from the same distribution. We introduced the K-S test in Figure 20, where we used it for normality testing. As an example of K-S testing, the daily changes in 20-year Treasury interest rates in 2015 (see Figure 23) have been divided into two text files, where the individual values are separated by carriage returns. S1.txt contains the changes from January through April and S2.txt contains the changes from May through August. The K-S test, as implemented in R, was used to compare the two distributions:

```
> library("stats", lib.loc="C:/Program Files/R/R-3.1.3/library")
> setwd("C:/Users/200007942/Desktop")
> X<-scan(file = "S1.txt", what = double())
Read 82 items
> Y<-scan(file = "S2.txt", what = double())
Read 85 items
> ks.test(X,Y)

        Two-sample Kolmogorov-Smirnov test

data:  X and Y
D = 0.0697, p-value = 0.9873
alternative hypothesis: two-sided
```

Figure 24: Session record from R.

The null hypothesis is that the two data sets are representative of the same distribution. The high *p* value supports the null hypothesis.

- Optimal block length for block bootstrap sampling: When the historical-approach model uses block bootstrap sampling, the modeler must consider the optimal block length. While the research on this topic is extensive, and divisive, validation should include a review of the methodology. For example, the jackknife-after-bootstrap method is suggested to be implemented for testing.[89] The tseries package for R includes tsbootstrap, which will implement the block sampling methods of Kuensch[90] or Politis and Romano.[88]
- Sufficient number of resamples used to achieve desired precision: The required confidence interval width, in dollars, for VaR_X^N is related to the business use of the model. Given the interval width, the required number of resamples can be estimated using the jackknife estimator algorithm described for Monte Carlo simulation in section 3.3.

3.5 Backtesting

In order to evaluate the quality of the estimates, models should be backtested with appropriate methods. Backtesting is a statistical procedure where actual profits and losses from one or more time intervals are compared to corresponding VaR estimates; for example, we expect that 5% of the returns from a set of independent, *N*-day samples will be more negative than $-VaR_{95\%}^N$. The time interval used in the backtest must be "out-of-sample," *i.e.*, the data used to calibrate the VaR model should not include the data from that time interval.

[89] S.N. Lahiri, Resampling Methods for Dependent Data. Springer (2013).
[90] H. Kuensch, The Jackknife and the Bootstrap for General Stationary Observations, *The Annals of Statistics* **17** (1989) 1217-1241.

In the backtesting process, we statistically examine whether the frequency of exceptions over some specified time interval is commiserate with the selected confidence level (*X*). These types of tests are known as tests of unconditional coverage. They are straightforward tests to implement since they do not account for the sequencing or magnitude of the exceptions.

The backtest is constructed as follows:

- Split the available data into two, not necessarily equal, portions, training and testing datasets. Think of the first part as the known history and consider the second part as if it had not already happened.
- Use the first portion of the data to estimate VaR_X^N.
- Compare VaR_X^N against the actual outcomes in the second portion of the data. Actual outcomes that are worse than VaR_X^N, labeled as exceptions, are counted and stored.
- Check if the rate of exceptions is significantly different from (1 – *X*). If that case, the model may not be a useful VaR estimator at the chosen confidence level *X*. For example, if daily VaR estimates are computed at *X* = 99% confidence, during one year (252 business days) we would expect 2 - 3 VaR exceptions.

To the extent that daily returns are stochastic, the exception fraction may well differ from (1 – *X*) for any finite time interval, even in the case of a correct VaR_X^N estimate. Kupiec's proportion of failures (POF) test can be used to test the outcome of backtesting. Kupiec's test is an unconditional coverage test; each exception is assumed to occur independently of the others.[91]

Let N_O be the number of observed exceptions in a backtest of N_T periods. Define the likelihood ratio Λ:

$$\Lambda = \left(\frac{N_T - N_O}{(1-X)N_T} \right)^{N_T - N_O} \left(\frac{N_O}{XN_T} \right)^{N_O} \tag{9}$$

It can be shown that -2 log Λ is approximately, centrally chi-squared with one degree of freedom.

$$-2\log\Lambda \approx \chi^2(1,0) \tag{10}$$

At confidence level 1 – *q*, we can construct a non-rejection interval:

$$\frac{N_O}{N_T} < X_U \text{ and } \frac{N_O}{N_T} > X_L \tag{11}$$

Calculate the *q* quantile of the $\chi^2(1,0)$ distribution, $\chi_q^2(1,0)$. Then solve:

$$-2\log\Lambda = \chi_q^2(1,0) \tag{12}$$

[91] P.H. Kupiec, Techniques for Verifying the Accuracy of Risk Measurement Models, *J. Derivatives* **3** (1995) 73-84.

for X, which will produce two solutions: X_L and X_U. If $X \in [X_L, X_U]$, then we can accept our estimate of VaR_X^N at the $1 - q$ confidence level, based on the backtest.

Table 22: Non-rejection intervals $[N_T \cdot X_L, N_T \cdot X_U]$ calculated with equation (12) for various values of q and N_T.[92]

		quantile of loss q			
		0.90	0.95	0.975	0.99
N_T	125	[6, 20]	[2, 12]	[0, 8]	[0, 4]
	250	[16, 35]	[6, 20]	[2, 12]	[0, 7]
	500	[37, 64]	[16, 36]	[6, 20]	[1, 10]
	750	[59, 92]	[26, 50]	[11, 28]	[2, 14]
	1000	[81, 120]	[37, 65]	[15, 36]	[4, 17]
	1250	[104, 147]	[47, 79]	[21, 43]	[6, 20]

Compare the above POF analysis to the Basel committee's traffic light coverage test. According to different levels of exceptions, the test suggests different levels of risk capital requirement. For short, over 250 days 99% VaR, if there are 0 to 4 exceptions observed, the model falls into a green zone and is defined to be accurate as the probability of accepting an inaccurate model is quite low; 5 to 9 exceptions indicate a yellow zone and no consensus is established on whether the model is accurate or not; 10 or more times of exception indicate a red zone indicating a clear problem with the VaR model. In the yellow zone, developers are expected to provide extra evidence to support their models.

More sophisticated test protocols for backtesting adequacy have been proposed, such as conditional coverage tests,[93,94] which simultaneously test if the VaR violations are independent and if the average number of violations is correct. A full-fledged framework for backtesting may also include backtesting for the entire loss distribution and backtesting for the tails.

3.6 Stress Testing and Scenario Analysis

Stress testing and/or scenario analysis assess the vulnerability of portfolios to hypothetical events. Stress testing complements probability-based risk measures such as VaR. VaR predicts the maximum likely loss at a stated probability, but does not predict extreme losses, as might occur during an economic crisis, which will likely be larger than VaR_X^N. By contrast, stress testing provides useful estimates of extreme losses, but does not predict the likelihood of events associated with such losses.

Stress and scenario testing can play an important part in the validation of the VaR model by showing how the portfolio-value impacts, that are produced by stress and scenario tests, relate to the VaR model output. Stress and scenario testing is specified by regulators as one of the key approaches to

[92] G.A. Holton, Value-at-Risk Theory and Practice, 2nd Ed., e-book at http://value-at-risk.net.
[93] P.F. Christoffersen, Evaluating interval forecasts, *Int'l Economic Rev*. (1998) 841-862.
[94] M. Haas, New Methods in Backtesting, Financial Engineering Research center caesar, Bonn (2001), http://www.ime.usp.br/~rvicente/risco/haas.pdf (Sept, 2015).

validation.[95]

A fundamental challenge in stress and scenario testing is the identification of appropriate scenarios. The scenarios are not prescribed, nor is there a specific level of severity required. Firms are left to develop their own stress and scenario tests, taking in to account their particular business and risk profile. The stress tests and scenarios can be classified into two types:

1. **Historical stress events**, such as major stock market crashes, banking crises, and so on. Given that the stress tests are applied to today's balance sheet, an additional complexity arises in the "translation" of the historical stress event to a stress to be applied in the current economic environment. For example, if an interest rate stress is based on a historical event, when rates fell from an unusually high level, then it may be inappropriate to apply the absolute size of this shock directly to today's yield curve.

2. **Forward-looking stresses,** manifesting as dramatic, synthetic value changes of one or more market factors between the beginning and end of the VaR simulation timeframe. The development of forward-looking stresses is undoubtedly challenging, as the selection of stresses is a consensus-building exercise, and views may be sought from a wide range of stakeholders.

In both cases the portfolio-value impacts of the risk factor(s) outcomes in the specified scenario can be compared with the probability distribution produced by the firm's VaR capital model. This can be a useful way of relating the output from the model to "real-world" events, helping users of the model's output to gain a deeper understanding of the model behavior. However, as with back-testing and sensitivity testing, it is unlikely to provide absolute conclusions about model performance.

For portfolios that respond primarily to one source of risk (*i.e*., a single market variable), these synthetic stresses have been suggested by the Derivatives Policy Group and G-30 Group:[96]

- Parallel yield curve shifting by $\pm$ 100 basis points
- Yield curve twisting by $\pm$ 25 basis points
- Each of the 4 combinations of yield curve shifting and twisting
- Stock index changes of $\pm$ 10%
- Currency changes of $\pm$ 6%
- Implied volatility change by $\pm$ 20%
- Swap spread changing by $\pm$ 20 basis points

Depending on the portfolio, the validator may wish to "shock" the market variables in combinations. This cannot be undertaken willy nilly; the combinations should be plausible. Schachter provides guidance.[97]

[95] Guideline 50 of the EIOPA consultation paper CP-13/011 (March 2013).
[96] S. Allen, *Financial Risk Management: A Practitioner's Guide to Managing Market and Credit Risk, 2nd Ed*., Wiley Finance (2013).
[97] B. Schachter, How Well Can Stress Tests Complement VaR? http://gloria-mundi.com/UploadFile/2010-2/stressandevt1.pdf (September 2015).

In both test regimes, historical stress events and forward looking stress scenarios, the same portfolio valuation model underlying the VaR model is used to calculate the change in portfolio value that results from the market variables' movements. The losses, so calculated, should be compared with VaR_X^N. If the losses are larger than VaR_X^N, the value of $1 - X_S$ associated with stressed market loss should be calculated using the pdf associated with the VaR model, as illustrated in Figure 25.

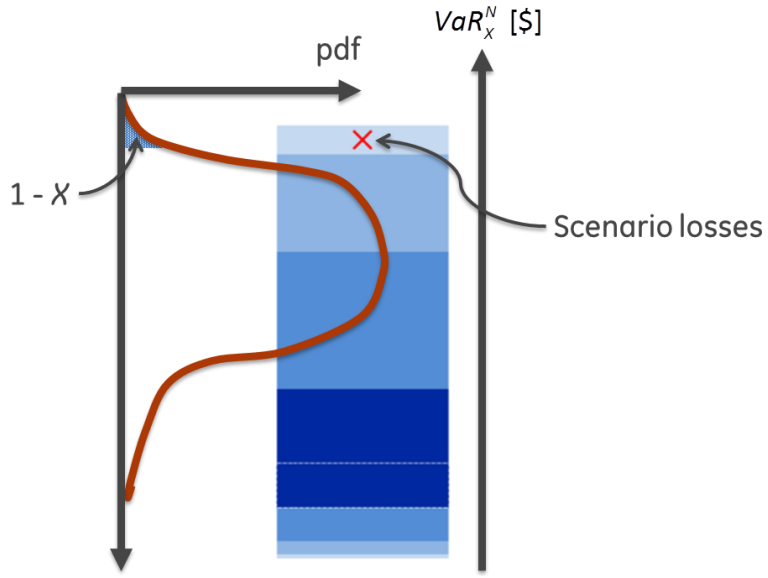

Figure 25: The portfolio losses associated with a stressed scenario are directly comparable to the value of VaR_X^N, and the underlying pdf. The losses from the stressed scenario are associated with a particular value of $1 - X$.

If the results of this analysis suggest that an event that happened in recent history, or is likely to happen in the near future, is a once-in-a-century event, based on:

$$\frac{1}{1-X_S}N$$

Then the VaR model likely is not useful for prediction of point-in-time capital requirements through periods of economic stress. This is especially true, if the VaR analysis performs thusly for a variety of historical and/or future stress scenarios.

While the precise choice of stress scenarios, whether historic or future, is inevitably subjective, the selection process can be made more systematic by using a well-produced scenario catalogue, rather than just a handful of *ad hoc* scenarios. Such a catalogue might include:[98]

- Moderate market stress scenarios of the sort that occur annually: changes in market volatility, a bond market squeeze due to fiscal surpluses, changes in the euro, a widening or falling TED spread, and other market indicator moves from recent market experience.

[98] K. Dowd, *Measuring market risk, 2nd Ed.*, Wiley Finance (2007).

- More extreme market scenarios such as repeats of major stock market crises (e.g., the 23% fall in the Dow Jones on October 19, 1987, the 48% fall in the Nikkei over 1990, etc.), or exchange-rate crises (*e.g.*, the ERM devaluations in September 1992, the fall in the peso in December 1994, the East Asian devaluations in 1997, the 40% fall in the ruble in August 1998, etc.), or a bond market crash (*e.g.*, the near doubling of US interest rates in 1994), or major country shocks (*e.g.*, the Latin American crisis in 1995, the Asian crisis in 1997, Russia in August 1998 and Brazil in 1999), or the failure or near failure of a large institution (*e.g.*, LTCM in 1998, Enron in 2001).

- Supervisory program-based scenarios, such as the Supervisory Capital Assessment Program (SCAP), the Comprehensive Capital Assessment Review (CCAR), or the Dodd—Frank Wall Street Reform and Consumer Protection Act (DFAST).

Credit Value Adjustment

Credit Value Adjustment

Michael Vallance, Jerry Cline and Weiwei Shen

Contents

1 Output of Credit Value Adjustment

Credit value adjustment (CVA) is the market value of counterparty credit risk. CVA is the difference between the risk-free portfolio value and the portfolio value accounting for counterparty defaults. Unilateral CVA calculation assumes that only the counterparty can default. Bilateral CVA calculation assumes that both the "bank" and the counterparty can default.

2 Mathematical Modeling of Credit Value Adjustment

For unilateral CVA, the mathematical expression for the CVA is:[99]

$$C = \mathbf{E}\left\{\int_0^T L(t)\left[-dH(t)\right]\right\} \tag{13}$$

In equation (13), **E** signifies expected value, T is the time span of the portfolio exposure and $L(t)$ is the loss suffered if the counterparty defaults at time t. While equation (13) describes the CVA of a single counterparty, generalization to include multiple counterparties is straightforward.

$$H(t) = \exp\left[-\int_0^t h(u)du\right] \tag{14}$$

In equation (14), $h(t)$ is the stochastic hazard rate. It represents the conditional probability of the occurrence of default in a small time interval $[t, t + dt]$, given that default has not occurred by time t. Assuming that the time to default probability is not changed by the absence of a default up until time t, then equation (13) becomes:

$$C = \int_0^T \mathbf{E}\left[L(t)\right]\left[-dH(t)\right] \approx \sum_{i=1}^{m} \frac{1}{2}\left\{\mathbb{E}\left[L(t_{i-1})\right] + \mathbb{E}\left[L(t_i)\right]\right\}\left[S(0,t_{i-1}) - S(0,t_i)\right] \tag{15}$$

In equation (15), $t_m = T$, $\mathbb{E}$ is the expected value in the risk-neutral measure, and $S(0,t)$ is the simulated survival probability to time t.

$\mathbb{E}[L(t)]$ is calculated by Monte Carlo calculation with risk-neutral, valuation models. This includes one or more stochastic, interest-rate models and net-present-value (NPV) representations of the financial instruments.

The survival probabilities S or the integral hazard rates H are calibrated from the quoted credit-default swap spreads for the obligator (if traded), the bond spread referenced to the appropriate risk-free rate (if bonds are traded), or from the appropriate rating transition matrix. When none of these are available, then survival probabilities for firms with similar credit worthiness are employed.

For example, the average hazard rate $\langle h(t)\rangle$ of a counterparty, associated with the time interval $[0, t]$, can be calculated from the counterparty's bond yield spread $s(t)$ for a bond of the same tenor.[100]

[99] *Adaptiv Analytics 2013.2 Theroy Guide*, Sungard.

$$\langle h(t)\rangle = \frac{s(t)}{1-R}$$
$$S(0,t) = \exp\left[-\langle h(t)\rangle t\right] \tag{16}$$

In equation (16) R is the recovery rate for the bond, typically 0.25 to 0.52, depending on debt seniority. Where counterparty bond spreads are available for bonds of various maturities, the hazard-rate curve can be synthesized using the same bootstrapping technique as is used for construction of the zero-coupon, yield curve.[101] In equation (4), $s(t)$ could also represent the credit default swap spread.

The model must be modified appropriately, where collateral and netting agreements are associated with the financial instruments in the portfolio. Collateral agreements reduce the value of $\mathbb{E}[L(t)]$ to an extent dependent on the terms of the agreements.

3 Quantitative Validation of Model Conceptual Soundness

The subsections of sections 0 and 4 correspond to the categories of Standards 5 and 6 in the business model risk-management template.[102] We propose quantitative testing for selected categories.

3.1 Model Theory, Design, and Construction

The CVA model, as described above, has three subcomponents:

5. Stochastic interest-rate (*e.g.*, Hull-White) model
6. Financial instrument valuation model
7. Survival probability model

[100] J.C. Hull, *Options, Futures, and other Derivatives, 9th Ed.*, Pearson (2015).
[101] G. Castellacci, On Bootstrapping Hazard Rates from CDS Spreads, http://ssrn.com/abstract=2042177 (downloaded September 2015).
[102] Model Risk Management, Model Documentation Template, V3.02, GE Capital (July 2015).

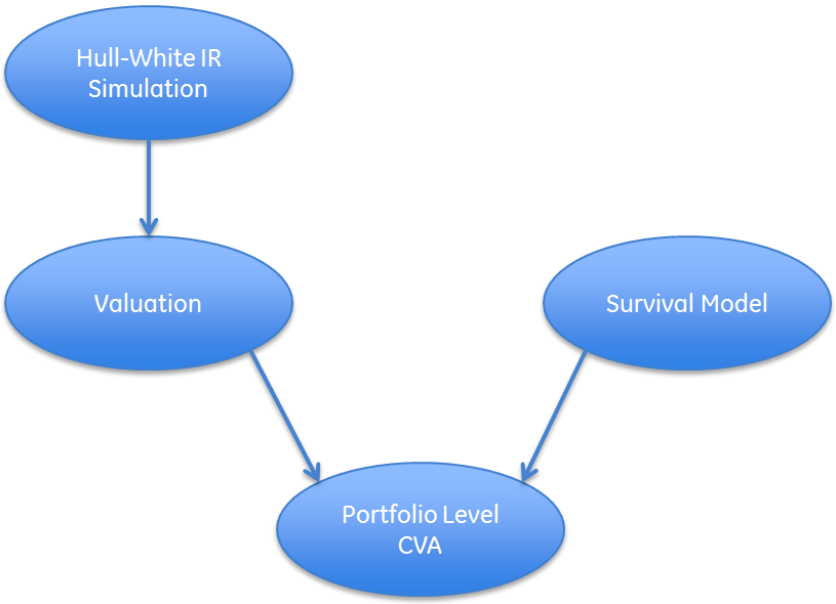

Figure 26: Flow diagram for CVA model.

In Chapter 2, section 4, we recommended tests for stochastic valuation model accuracy. In Chapter 2, sections 2 and 4, we proposed tests for interest rate curve calculations and for stochastic term structure models. In section 2 of that same chapter, we also recommended quantitative tests for portfolio valuation models. We recommend those same tests for model components 5 and 6 above.

The hazard rate curves and/or the survival probability curves developed in the production model should be reproducible. If the validator has access to the calibration data used to produce the model, we urge the validator to reproduce the curves, using the same data. To this purpose, we recommend the function cdsbootstrap, which is available in Matlab's Financial Instruments Toolbox. The appendix provides more information about this tool.

If the validator replicates one or more survival-probability curves for counterparties, using the same or different calibration data as used for the production model; we would suggest using model replicability, as defined in Chapter 1, subsection 3.3, to compare the original and replicated curves. The validator chooses a representative set of financial assets from the portfolio. The CVA for the selected assets is calculated using the production and replicated survival-probability curves, using the term-structure and valuation modeling capability of the production model. The set of differences between the two CVA estimates, asset by asset, is used, along with business-relevant tolerances, to generate an estimate of model replicability R_{MK}.

Should the validator choose to replicate the entire CVA model, the appropriate metric is still the model replicability on a characteristic set of assets. Again, this metric is only defined if the business has supplied tolerances for the model. *If the business does not feel the need to place bounds on the error of*

the model predictions, then there may be no need to quantitatively test the model. Historically, the following bounds have been used (*f* for the production model and *r* for the replicate model):[103]

$$\left|C_f - C_r\right| \le 3\langle CS01\rangle \tag{17}$$

In equation (17), $\langle CS01\rangle$ is the change in portfolio value in response to a one basis point change in the credit default swap premiums underlying the portfolio. More specifically, it is the change in portfolio value resulting from a one basis point, parallel shift in the credit curve (first derivative of portfolio value, with respect to swap premium, expressed in dollars per basis point).

We suggest that this comparison be executed without including the impact of collateral agreements. If the validator should like to test that functionality of the production model, additional testing can be done in a subsequent exercise, where the CVA can be computed including these contractual agreements. The Adaptiv model[99] evaluates CVA both ways automatically.

3.2 Model Formulation and Selection

Alternative model formulations, including alternative calibrations or segmentations, if undertaken, comprise challenger models. If the challenger model is treated as a virtual benchmark model (*i.e.,* the "one true model"), then we can compare the production model with the challenger model, using the Chapter-1 (subsection 3.2) concept of model capability, C_{MK}. This is the same calculation as used for model replicability, and the same business-related tolerances can be used.

3.3 Model's Intended Use

From the qualitative perspective, the validator can use model capability, as described in Chapter 1, subsection 3.2, to demonstrate the model's accuracy. To do so would require a benchmark model or independent valuations, such as the debt valuation adjustments (DVAs) estimated by the counterparties to the same deals.

4. Testing and Evaluation

4.1 Performance of Alternative Testing

In chapter 2, subsection 2.4, we provided guidance for testing the synthesis of interest-rate curves, following Hagan and West.[104] These guidelines apply in the present case.

In the case of validating survival-probability curves, the typical shape of these curves tends to be monotonically increasing, concave up (positive second derivative with respect to time). Monotonically increasing, concave down (negative second derivative with respect to time) expression is possible, indicating that the marginal default probability is decreasing with time.

[103] Z. Yang, J. Cline and J. Alvarez, Adaptiv Credit Value Adjustment, GMGV Validation Technical Report GETM-RR-004 (November 2013).
[104] P.S. Hagan and G. West, Methods for Constructing a Yield Curve, WILMOTT Magazine (May, 2008) finmod.co.za/interpreview.pdf.1,3,4,6.

Non-monotonic probability-survival curves indicate non-negative hazard rates for certain time intervals. While extreme market conditions may lead to such curve expression, this should serve as a red flag for the validator.

In addition, survival-probability curves must be continuous, and perturbations due to single-point changes in the calibration data should be localized, as is the case for interest-rate curves (see Chapter 2, subsection2.4).

4.2 Replication of Developer Testing

If not provided by the model developer, the validator should quantitatively examine the sensitivity of the production model to changes in credit default swap spread. In particular, selected survival-probability curves should be perturbed: at selected tenors, the underlying credit default swap spread should be artificially changed. Then the CVA is estimated. This is done for various tenors and various changes in spread, in order to develop a sensitivity map of ΔC. The results should be rational, and regions of high sensitivity should be identified—the corresponding input data may bear extra scrutiny.

Similarly, the data underlying the interest rate curve(s) should be perturbed at various tenors, and the impact on ΔC noted. Also, the parameters of the term-structure model, Hull-White or otherwise, should be perturbed, and the impact on ΔC noted.

These perturbation measurements, combined with judgment about the reliability of the calibration data, will provide a basis to judge model risk.

In Chapter 2, subsection 4.2, we opined on the estimation of error in Monte Carlo simulation. The same concerns are present in CVA estimation, since Monte Carlo simulation is used to calculate the expected value of losses, associated with counterparty default, as a function of default time. Using the Chapter-2 approach, we can develop a confidence interval for $\mathbb{E}[L(t)]$, which will be a function of time and the number of simulations. If the confidence intervals for $\mathbb{E}[L(t)]$ are broad, this translates directly into a broad confidence interval for CVA, which can be compared to the required tolerance. The remedy for broad confidence intervals is to increase the number of simulations. Observing the convergence of CVA, versus number of simulations, is an alternative approach.

Preferably, the model developer has performed backtesting with the model; if not, the validator may wish to do so. For backtesting, a representative portfolio of our firm's historical assets, preferably the whole portfolio, with timespan T, is selected, such that all of the trades have settled/defaulted in the meantime. A time zero, at least T years ago, is selected; and the CVA of the portfolio is estimated, using the present model structure with calibration data that was available at time zero. The actual losses of the portfolio can be compared to the *a priori* estimate. The desirable result is a conservative CVA estimate. Repeating this procedure with different portfolios and/or different timespans will provide a set of estimation errors that can be used to calculate model capability (see Chapter 1, subsection 3.2).

Of particular interest in backtesting are timespans that include financial shocks, since these constitute stress tests of the model.

5. Appendix: cdsbootstrap[105]

Financial Instruments Toolbox™ software supports:

Table 23: CDS Functions

Function	Purpose
cdsbootstrap	Compute default probability parameters from CDS market quotes.
cdsspread	Compute breakeven spreads for the CDS contracts.
cdsprice	Compute the price for the CDS contracts.

The market information in this example is provided in the form of running spreads of CDS contracts maturing on the CDS standard payment dates closest to 1, 2, 3, 5, and 7 years from the valuation date. The model also requires a set of zero-coupon, risk-free, interest rates with corresponding maturities, or a pre-existing zero-rate curve.[106]

```
Settle = '17-Jul-2009'; % valuation date for the CDS
MarketDates = datenum({'20-Sep-10','20-Sep-11','20-Sep-12','20-Sep-14',...
'20-Sep-16'});
MarketSpreads = [140 175 210 265 310]';
MarketData = [MarketDates MarketSpreads];
ZeroDates = datenum({'17-Jan-10','17-Jul-10','17-Jul-11','17-Jul-12',...
'17-Jul-13','17-Jul-14'});
ZeroRates = [1.35 1.43 1.9 2.47 2.936 3.311]'/100;
ZeroData = [ZeroDates ZeroRates];

[ProbData,HazData] = cdsbootstrap(ZeroData,MarketData,Settle);
```

The following code draws the bootstrapped default probability curve plotted against time, in years, from the valuation date.

```
ProbTimes = yearfrac(Settle,ProbData(:,1));
figure
plot([0; ProbTimes],[0; ProbData(:,2)])
grid on
axis([0 ProbTimes(end,1) 0 ProbData(end,2)])
xlabel('Time (years)')
ylabel('Cumulative Default Probability')
title('Bootstrapped Default Probability Curve')
```

[105] Abstracted from Mathworks Documentation: Credit Default Swap.
[106] Mathworks Documentation: cdsbootstrap.

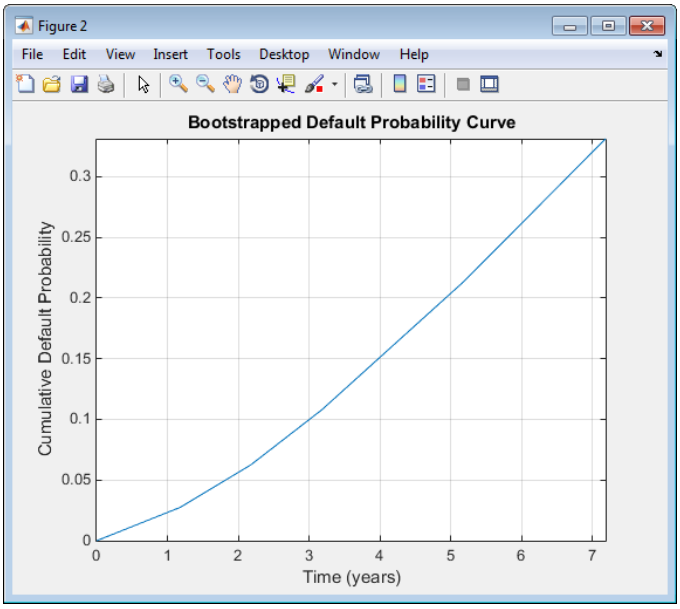

Figure 27: Cumulative default probabilities for the firm, equal to (1 – survival probability).

The associated hazard rates are returned as an optional output. The convention is that the first hazard rate applies from the settlement date to the first market date, the second hazard rate from the first to the second market date, etc., and the last hazard rate applies from the second-to-last market date onwards. The following code draws a plot of the bootstrapped hazard rates, plotted against time, in years, from the valuation date:

```
HazTimes = yearfrac(Settle,HazData(:,1));
figure
stairs([0; HazTimes(1:end-1,1); HazTimes(end,1)+1],...
[HazData(:,2);HazData(end,2)])
grid on
axis([0 HazTimes(end,1)+1 0.9*HazData(1,2) 1.1*HazData(end,2)])
xlabel('Time (years)')
ylabel('Hazard Rate')
title('Bootstrapped Hazard Rates')
```

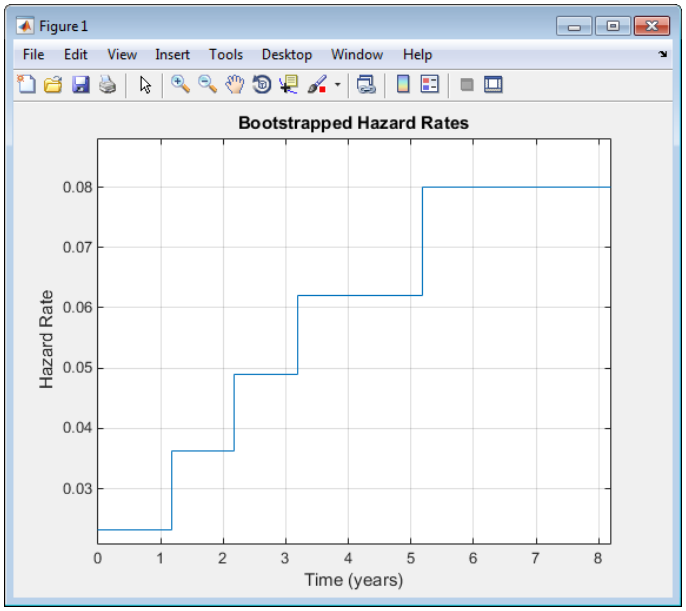

Figure 28: Hazard rate curve calculated from credit default spreads

Economic Capital

Economic Capital

Weiwei Shen and Michael Vallance

Contents

Economic capital (ECap) is the capital that shareholders should invest in the company in order to limit the probability of default to a given confidence level over a given time horizon. It is measured as the potential loss in excess of expected loss (or gain), typically over a one year time period, at a specified confidence level. Figure 1 illustrates the basic concept of economic capital given a loss distribution. The tail risk represents the amount of risk used to compute economic capital.

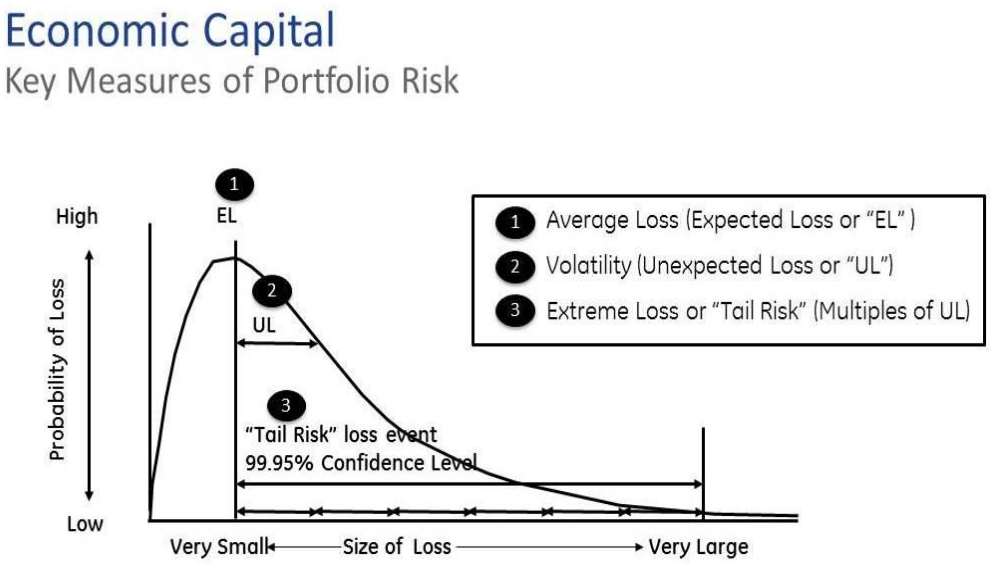

Figure 29. A schematic representation of portfolio risk.

At GECC, several different models have been adopted to compute ECap for various types of portfolios, such as Moody's RiskFrontier, GECC HQ aggregator and GEECAP Treasury IR. RiskFrontier is a vendor-supplied model; the other two are in-house models. In this chapter, we focus on common traits shared across the ECap model class. The quantitative validation tests (QVTs) that we recommend in this chapter are specific to the ECap model class; we refer to previous chapters for QVTs with broader applicability. Quantitative validation of upstream models, such as asset valuation, probability of default and loss given default are outside of the scope of the present chapter. Rather, we emphasize diversification and dependency modeling.

Negative Diversification Benefit

Diversification benefits are expected when aggregating risk from dissimilar assets, business lines or risk types. *I.e.*, through diversification, non-systematic risk should be diversified away such that the ECap calculated from the aggregate portfolio should be less than the sum of the several ECaps calculated for the individual components in the portfolio.

Denote by M the collection of random variable representing individual portfolio losses over some time interval. A risk measure is a real-valued function $\rho: M \to R$. Then $\rho(X)$ can be interpreted as the riskiness, or the amount of capital that should be held in association with a portfolio, with a loss

distribution given by X. For portfolio diversification (a portfolio of *n* components) we expect the following relationship:

$$\rho(\sum_{i=1}^{n} X_i) \leq \sum_{i=1}^{n} \rho(X_i), \quad X_i \in M,\ i = 1, \dots n \qquad 1$$

Equation 1 reflects our belief that diversification does not create extra risk.[107]

Denote by $\Lambda: M \times M \to R$ the risk capital allocation to a component of M, such that for any $X_i \in M$, and for any portfolio $Y \in M$, where Y includes X_i, we have diversification benefits at the component level:

$$\Lambda(X_i, Y) \leq \Lambda(X_i, X_i) = \rho(X_i) \qquad 2$$

In other words, the risk allocated to a sub-portfolio X_i within Y should not exceed its standalone risk. Furthermore, suppose that n risky sub-portfolios X_i, $i = 1, \dots, n$, comprise portfolio Y. The full allocation rule in risk capital allocation requires allocating the risk calculated from the total portfolio to sub-portfolios with zero remainder:

$$\sum_{i=1}^{n} \Lambda(X_i, Y) = \Lambda(Y, Y) = \rho(Y) = \rho(\sum_{i=1}^{n} X_i) \qquad 3$$

Negative diversification benefit has been observed in applications of the GECC models described in the previous section, which we would regard as a finding. Negative diversification benefit can be rooted in a variety of sources. We suggest the following tests for negative diversification benefit detection and attribution.

Risk Measure

Inappropriate choice of risk measure may result in a poorly conditioned aggregated risk assessment and poorly conditioned disaggregated risk allocation. Coherent risk measures, as a class, are intuitive and provide desirable mathematical properties. Coherent risk measures are promoted in risk management.[108] By definition, a coherent risk measure has the properties of monotonicity, sub-additivity,[109] positive homogeneity, and translational invariance. A non-coherent risk measure may generate negative diversification if it violates the sub-additivity property. Validators should examine the risk measure underlying the ECap calculation; where non-coherent measures are used, the modeler must justify the choice and demonstrate that the model is provisioned to compensate for any resulting shortcomings.

[107] Artzner, P., *et al.*, in *Risk Management: Value at Risk and Beyond*, Dempster, M.A.H., ed., Cambridge (2002). Axiom S Subadditivity, p. 152.

[108] While the risk measure is the fundamental building block of ECap calculation, there exists no consensus on the best risk measure. For example, see Artzner, P., *et al.*, in *Risk Management: Value at Risk and Beyond*, Dempster, M.A.H., ed., Cambridge (2002). p. 145.

[109] Sub-additivity is a highly desirable property for any risk measure. If regulators use non-sub-additive risk measures to set capital requirements, a financial firm might be tempted to break itself up to reduce its regulatory capital requirements, because the sum of the capital requirements of the smaller units would be less than the capital requirement of the firm as a whole.

Among typical risk measures, including standard deviation (SD), Value at Risk (VaR) and expected shortfall (ES), only ES is a coherent risk measure, in general.[110]

If the losses follow an elliptical distribution, VaR is a coherent risk measure. Where VaR is used as the risk measure, the validator should analytically confirm, of have confirmed, that the underlying loss distribution is elliptical.[111] Pure market risk may well follow an elliptical distribution, given its symmetric nature.

Given that VaR is not sub-additive, ECap calculated with VaR may result in negative diversification benefit.

$$\rho(\sum_{i=1}^{n} X_i) > \sum_{i=1}^{n} \rho(X_i) \qquad 4$$

Where VaR has been selected as the risk measure, validators should calculate ECap for a suite of representative portfolios, using the model, to assure that the model never predicts negative diversification benefit.

SD is not a tail-sensitive risk measure; SD is a poor choice for heavy tailed loss distributions. SD is non-coherent, due to the violation of monotonicity, although it does not violate sub-additivity. Use of SD as the risk measure can lead to back allocations of ECap in excess of the total exposure, in clear violation of the full allocation rule. In general, SD is a poor choice as a risk measure for ECap calculations.
For other risk measures, and especially for non-coherent risk measures, validators should require developers to provide rational and documented support of applicability.

Capital Allocation Method

A poorly selected risk allocation algorithm can lead to a negative diversification, especially for models employing a non-coherent risk measure.

The most popular algorithms are Euler's (gradient) method and Shapley's method. For a coherent risk measure, positive diversification benefit is realized with either method.

Validators should confirm that the chosen algorithm exactly satisfies full allocation; an *ad hoc* allocation rule could result in violation of full allocation.

Validators should confirm that the allocated risk is not larger than the calculated standalone risk, for each and every asset.

[110] Acerbi, C., and Tasche, D., Expected shortfall: A natural coherent alternative to value at risk, *Economic Notes* **31**(2) (2002) 379-388.
[111] The class of elliptical distributions is large; containing, as special cases, the multivariate normal, mixture normal, multivariate t, multivariate stable, Kotz and Pearson II distributions. In general, closed-form formulas for statistical testing are not available. See Zhou, G., Asset-pricing Tests under Alternative Distributions, *J. Finance* **48**(5) (1993) 1927-1942.

Negative Exposure

Negative exposure has been observed in the use of Moody's RiskFrontier. It is counterintuitive and problematic. Validators should identify such anomalous input; for large portfolios, automation can be used. During a recent validation, the total number of transactions with negative exposure was 3,700 out of 1.6 million and the EAD of those exposures was -$261 mm as shown in Figure 2.

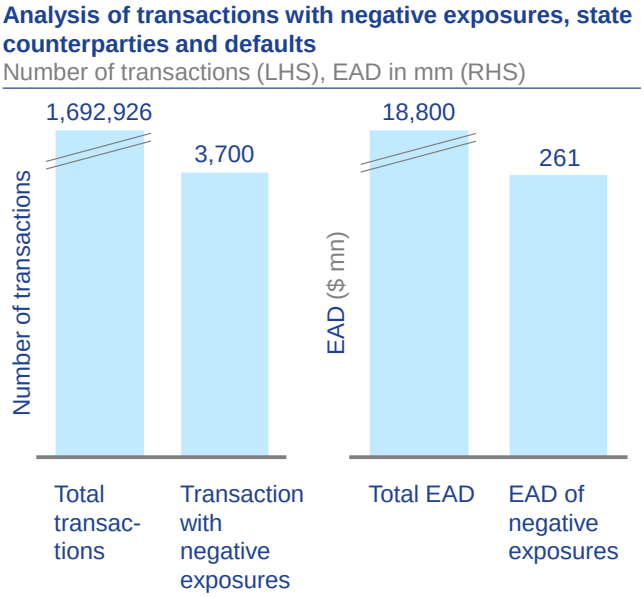

Figure 30. Anomalous negative exposures at default were common in this portfolio.

Simulation Convergence

As detailed in the Risk Analytics chapter, a large number of simulation trials are necessary to achieve precision in the calculation of VaR by simulation. If the confidence intervals on ECap are wide, the validator cannot determine whether positive or negative diversification benefit is realized, with reasonable confidence. Obtaining tight confidence bands for risk aggregation and disaggregation is a precondition for valid observations of the projected diversification benefit, positive or negative.

The following summary of a comparative study, comparing standard deviation and expected shortfall as risk factors, using Moody's RiskFrontier was performed by GE Capital. The analysts highlight several of the issues that we have described.

1. Risk factors based on distribution tail analysis are vulnerable to high noise levels. Large numbers of simulations are required.
2. Standard deviation is not sensitive to distribution tail perturbations. Such practice can lead to underestimation of risk.
3. Combining SD as the risk factor with Euler's allocation rule avoids negative diversification benefit, although other distortions are likely.
4. When using SD we risk back-allocating risk capital in excess of the total exposure.

Approach currently used by GE

	Risk contribution/ Standard deviation (RC/SD)	Tail risk contribution/Expected shortfall (TRC/ES)
Estimation noise	• Using the SD approach at 500 bp results in **a less noisy estimate** • Low simulation noise levels of the SD approach are conducive to development of the Ecap formula[1]	• Computation of ES at 1bp and 10 bp results in **noisy estimates of transaction capital** with noise diminishing as one moves away from the tail • High levels of simulation noise complicate development of the Ecap formula
Coherent risk measure	• SD is not a coherent risk measure but it **is sub-additive** • SD **is not monotonic** however the resulting simulation is less noisy	• The ES approach is a **coherent risk measure** • While the ES approach is monotonic, **the simulation of the loss distribution is complex and extremely noisy** • A scaled ES calculation with more allocation in the body to diminish noise leads to a **breakdown of the monotonicity property**
Sensitivity analysis	• SD **is more sensitive to PD** as it is based on the standard deviation of mean losses • The SD approach corresponds approximately to the 85% confidence level	• ES is **more sensitive to R^2** as it is a measure of extreme losses during a large systemic shock • Capital allocation **based solely on extreme tail events** is undesirable
Risk frontier output	• Some results are **counter-intuitive including those for SSLP and trade payables** which have high R^2's and are more sensitive to extreme events yet are allocated a higher capital in SD compared to ES	
Implement-ation in risk frontier	• RC is easier to explain as: – The standalone component is based on **intrinsic risk drivers**, PD and LGD – The interaction with the portfolio is driven by R^2 which is a measure of the relationship of a standalone component with the whole portfolio	• Using the ES approach **requires post processing** and calculation of a scaling factor between the medium tail capital calculation and extreme tail calculation

1 The Ecap formula approximates transactional Ecap output from Risk Frontier and is a regression model built on top of RF capital allocation output

Figure 31. Real-world issues with ECap analysis.Dependence Models

The overall risk of a portfolio depends not only on the individual risks associated with the component assets, but also on the dependencies among the individual risks. The dependencies among the individual risks can be roughly attributed to three sources: (a)dependencies among the market factors used to price the individual assets, (b) dependencies among the values (returns) of similar assets, (c) and dependence between values (returns) of dissimilar assets. Dependency modeling is challenging due to the non-stationary traits of markets, the lack of high quality historical data, and the limited time intervals of available data. Yet, dependence modeling is crucial; the resulting diversification benefits could amount to 40% of undiversified total economic capital.[112] In this section, we cover PD-LGD correlation in loan and leasing portfolios, inter-asset correlation, and copulas.

PD-LGD Correlation (PLC)

Empirical evidence suggests that the recovery rate is correlated with both systematic risk factors as well as with firm-dependent and asset-dependent idiosyncratic risk; *i.e.*, LGD is correlated with PD. During a recessionary period, recovery tends to be lower than during market upturn periods. Meanwhile, probability of default tends to be higher during a recessionary period than during a normal market. Moreover, a positive correlation between a firm's credit quality and recovery exists; *i.e.,* expected recovery is higher for higher credit-quality firms and lower for lower credit-quality firms. Ignoring PD-LGD correlation underestimates the risk inherent in a portfolio.

In Moody's RiskFrontier, PD-LGD correlation is built into the time-dependent, asset value A_t , which follows a lognormal distribution:

[112] The CRO Forum QIS 4 benchmarking study of 2008 suggested that diversification reduces economic capital by around 40% on average.

$$\ln\left(\frac{A_t}{A_0}\right) = \left(\mu_A - \frac{\sigma_A^2}{2}\right)t + \sigma_A\sqrt{t}B_A$$

$$B_A = R_A\phi + \sqrt{1 - R_A^2}\epsilon_A \quad 5$$

In equation 5, the systematic risk is represented by ϕ and the firm asset idiosyncratic risk by ϵ_A ; μ_A is the drift rate and σ_A is the diffusion rate of asset return.

The time-dependent recovery rate RR_t follows a lognormal distribution as well:

$$\ln\left(\frac{RR_t}{RR_0}\right) = \left(\mu_{RR} - \frac{\sigma_{RR}^2}{2}\right)t + \sigma_{RR}\sqrt{t}B_{RR}$$

$$B_{RR} = R_{RR}\phi + \sqrt{1 - R_{RR}^2}\epsilon_{RR} \quad 6$$

In equation 6, the firm's asset idiosyncratic risk is represented by ϵ_{RR}, μ_{RR} is the drift rate, and σ_{RR} is the diffusion rate associated with the recovery rate.

Thus, the PLC (short for PD-LGD correlation) can be computed as

$$\rho_{A,RR} = R_{RR}R_A + \sqrt{1 - R_A^2}\sqrt{1 - R_{RR}^2}\rho_\epsilon \quad 7$$

Backtesting of the correlation is based on simulation. To simulate the recovery rate RR_t, the analyst must use calibration to evaluate the parameters: μ_A, μ_{RR}, σ_A, σ_{RR}, R_A, R_{RR} and $\rho_{A,RR}$.

The simulated LGD distribution was calibrated to the realized LGD distribution during the 2008 downturn period, which was assumed to correspond to a one-in-20 year event;[113] the PLC parameters were calibrated by comparing the simulated LGD distribution at the 95% confidence level to the realized downturn LGD.

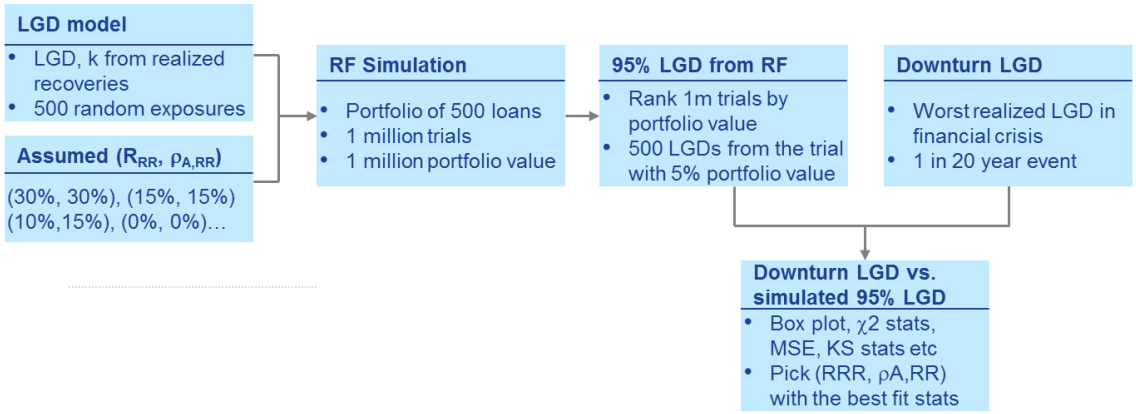

Figure 32. Flow chart for calibration of correlated PD and LGD models.

[113] E. Hayes, E. and Zhu, F., Validation of RiskFrontier Economic Capital Model (Version 2.0), GE Model Validation Document.

The flowchart describes the calibration protocol.

If backtesting is not incorporated into model development, then validation should include backtesting, in order to develop confidence in the PLC.

Sensitivity analysis, studying variation of the PLC, should also be included in validation, to assure that the relationship between ECap and PLC is stable and predictable.

The historical deal data used for PLC should be examined and compared to the present portfolio to which the PLC is applied. For example, if PLC is executed using unsecured bond recovery data, while a significant fraction of assets in the current portfolio are collateralized, then that PLC is probably not appropriate.

Correlation between Asset Returns

Moody's RiskFrontier includes a module called the Global Correlation Model, or GCorr. GCorr estimates risk correlations among obligors in a credit portfolio. Moody's documentation, third-party endorsements, and GE's own investigation find that GCorr is consistent with industry best practices.[114,115] Therefore, we use GCorr to exemplify correlation validation practices for credit portfolios.

Is GCorr's correlation engine is relevant to the asset class in the model under validation? GCorr includes 110 factors, 49 country-specific systematic factors and 61 industry-specific systematic factors. Performance of 43,000 firms over ten years is used to construct the GCorr engine. Nonetheless, validation should include a comparison of the deals in our model to those used to develop GCorr's correlation engine, to assure relevance.

In general, a multi-factor regression framework has advantages over pairwise calculations of correlation between obligor performances. In pairwise calculations, scatterplots are invaluable. Is the relationship linear? Do extreme values have undue influence on the outcome? Validation should assure that the amount of data used to calculate correlation yields the desired level of accuracy. The confidence intervals for the Pearson's product-moment correlation coefficient scale with $(n-3)^{-1/2}$, where n is the sample size.[116] The calculation of confidence intervals is typically done using bootstrap methods. Confidence interval calculation in the R programming language, use the routine *cor.test*. Where correlations calculations involve missing data, assure that the modeling methods use industry best practices (see the Prepayment chapter).

Correlation estimates should be validated by out-of-sample testing to assure that the model has not been over-fit. Metrics, such as absolute forecast errors, root-mean square errors, and correlations

[114] Neagu, R. and Santilli, S., Suh, S. and Sau, R., Evaluation of Moody's Analytics GCorr Corporate Model: GE Capital Opinion Document. 2013.
[115] Wyman, O., RiskFrontier-based Economic Capital Model, Model validation documentation (2013).
[116] Confidence Interval for Pearson's Correlation, Ch. 801, PASS Sample Size Software, http://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/PASS/Confidence_Interval_for_Pearsons_Correlation.pdf

between observed and forecasted values, should be reported. Similarly, correlations from one period should be used to estimate out-of-sample valuations. Similar outcome metrics can be used.

Risk Aggregation among Dissimilar Assets (Copulas)

Dependence structure between risk factors or between business lines must be specified and joint performance must be aggregated across all risks and lines to calculate enterprise ECap. The simplest approach is to linearly sum ECap predictions for each segment of business and/or risk factor. This conservative approach is equivalent to assuming 100% correlation; performance downturns occur in precise synchrony. For both internal decisions and external communication, the diversification benefit embedded in imperfectly correlated segments in the portfolio should be quantified. Thus, to understand the scope of the portfolio risk profile at the enterprise level, individual risks need to be aggregated by methods that account for the diversification benefits.

The dependence structure across business lines and risks is complex. Dependence structures among the tails of loss distributions must be estimated. Copulas are extensively used to capture the dependence structure in risk aggregation.[117] The approach to dependence structure estimation, as practiced in a risk aggregator developed by GE, is demonstrated below. We will focus on this framework, because it represents a common method to obtain the dependence structure in risk aggregation.

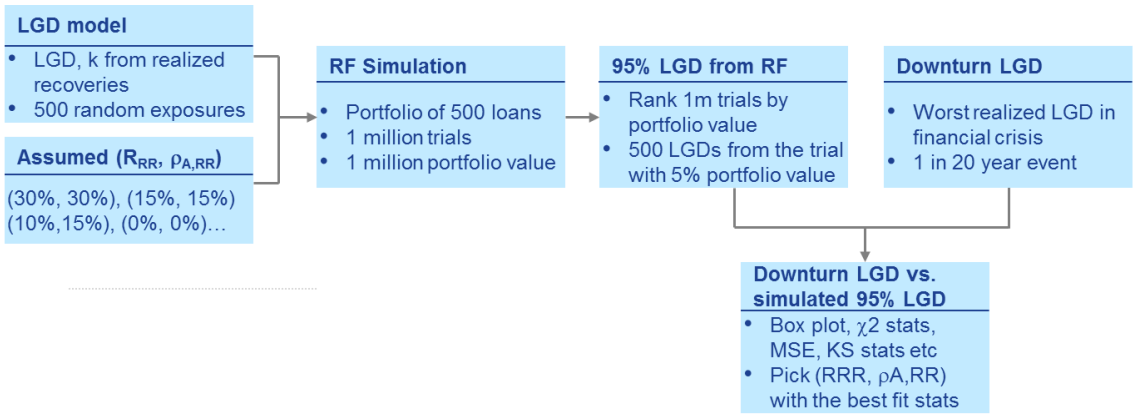

Figure 33. A flow chart for GE's multivariate distribution estimation procedure.

Underlying the copula approach is the transformation of a joint distribution into a set of marginal distributions used with a dependence function called a copula *C*. By Sklar's theorem if $F(x_1, \dots, x_n)$ is a joint distribution function with marginal risk distributions $F_1(x_1), \dots, F_n(x_n)$, then there exists a copula function $C: [0,1]^n \rightarrow [0,1]$ such that

$$F(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)) \qquad 8$$

We can produce the copula function *C* directly from the joint distribution function:

[117] Genest, C, Gendron, M. and Bourdeau-Brien, M., The advent of copulas in finance, *European J. Finance* **15**(7-8) (2009) 609-618.

$$C(u_1, \dots, u_n) = F\left(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)\right), \quad 9$$

with $u_i \in [0,1]$. Any distribution function with support on $[0,1]^n$ and a uniform marginal distribution is called a copula.

If C is a copula for $X_1, \dots, X_n$ then for every set of strictly increasing transformations, $T_1, \dots, T_n$, C is also a copula for $T(X_1), \dots, T(X_n)$. Intuitively, the copula is a relation between the quantiles of a set of random variables, rather than the original variables, and as such is invariant under monotonically increasing transformations of the raw data.

Just as marginal risk distributions $F_1(x_1), \dots, F_n(x_n)$ give an exhaustive description of $X_1, \dots, X_n$ taken separately, the joint dependence between these variables is fully and uniquely characterized by C. Therefore, an ideal method of measuring dependence might only rely on C. To measure dependence,[118] there are two well-known nonparametric measures via ranks, i.e., Spearman's rho and Kendall's tau. We focus on the latter.

Following the flow chart above, after having computed Kendall's tau, τ_n, from a sample with size n, an alternative test of independence can be implemented. H_0, the null hypothesis that the two random variables are uncorrelated, would be rejected at confidence level $\alpha = 5\%$ if

$$\sqrt{\frac{9n(n-1)}{2(2n+5)}}|\tau_n| > 1.96 \quad 10$$

There are situations for which Kendall's tau is the preferred measure of correlation:

- At least one of the variables, x or y, is measured on an ordinal scale;
- Neither x nor y is normally distributed;
- The sample size is small;
- A measure of the association between two variables is required;
- The relationship is non-linear.

Graphical tools should be used to detect and check dependence, such as the scatter plot of ranks, chi-plots and K-plots.[119]

---

[118] Let $\delta(X,Y)$ be a dependence measure. An ideal dependency measure should have several properties.

P1. Symmetry: $\delta(X,Y) = \delta(Y,X)$;

P2. Normalization: $-1 \le \delta(X,Y) \le 1$;

P3. $\delta(X,Y) = 1 \Leftrightarrow X,Y$ comonotonic; $\delta(X,Y) = -1 \Leftrightarrow X,Y$ countermonotonic;

P4. For $F : \Re \to \Re$ strictly monotonic increasing, $\delta(F(X),Y) = \delta(X,Y)$;

P5. $\delta(X,Y) = 0 \Leftrightarrow X,Y$ are independent.

[119] Genest, C. and Favr, A.-C., Everything you always wanted to know about copula modeling but were afraid to ask, *J. Hydrologic Eng'ring* **12**(4) (2007) 347-368.

Rank correlation, rather than Pearson's linear correlation, is preferred for ECap dependence calculations. The popularity of using τ_n as an estimator of the dependence parameter stems in part from the fact that closed-form expressions for the population value of Kendall's tau are available for many common parametric copula models. Gaussian or Student's t distributions do not fall into this category; as in the above flow chart, Kendall's tau can be transformed to Pearson's correlation ρ_n by

$$\rho_n = \sin\left(\frac{\pi}{2}\tau_n\right) \qquad 11$$

The flow chart implements a so-called Higham method to "distort" the correlation matrix into a positive semi-definite matrix.

The user must choose among various, copula-based, dependence structures for the data at hand. Which model provides the best fit to the observations?[120] Unless one dependence structure is undeniably best supported by the evidence, we suggest that the modeler, or the validator, try several copula forms, and note the sensitivity of the resulting ECap.

Graphical analysis, using scatter plots, can be used to judge the adequacy of a copula model ; generating a large sample from the estimated copula by Monte Carlo sampling, effectively portraying the associated copula density in two dimensions. As an example, the following figure (a) displays 100 simulated (U_i, V_i) pairs, based on an estimated copula. The six points of data, used to calibrate the copula, are represented by crosses and superimposed. Given the small size of the calibration data set, it is hard to tell from this graph whether the selected model accurately reproduces the dependence structure revealed by the six observations. To show the potential effectiveness of the procedure, the same exercise was repeated in figure (b), using an artificial Clayton copula. Here, the inappropriateness of the model is apparent; the copula was generated with $\tau = 5/6$, while, in actuality, $\tau_n = 1/15$ for the data.

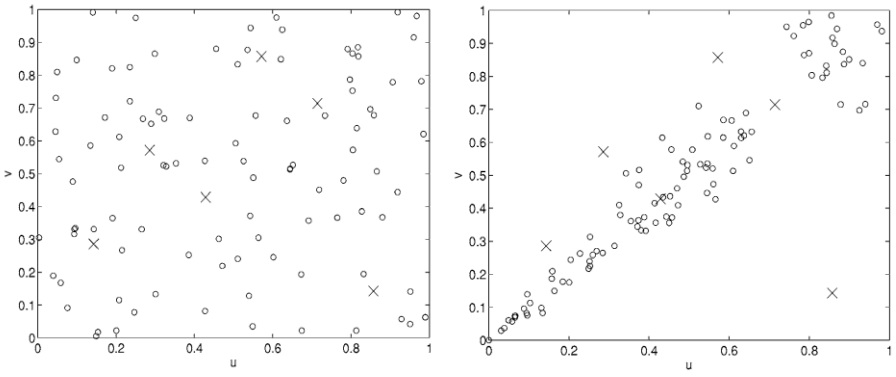

Figure 34. Correlation plots for a bivariate CDF, comparing data (Xs) to two different cupola simulations.

Formal statistical testing of the copula fit has been proposed, although the jury is still out. See the above-referenced goodness-of-fit survey paper by C. Genest, *et al*. These are rank-based versions of the

[120] Genest, C., Rémillard, B. and Beaudoin, D., Goodness-of-fit tests for copulas: A review and a power study, *Insurance: Math. & Econ.* **44**(2) (2009) 199-213.

Cramer-von Mises and Kolmogorov-Smirnov statistics. There is no consensus about which tests to choose, if any.

Index of Quantitative Test Methods