

Incorporating Covid in Default Rate Models



Abstract

One of the biggest issues facing econometricians and analysts of default rates today is how to incorporate data from the historical periods covering the height of the Covid-19 pandemic into their analysis. This period represents a unique challenge as the Covid-19 pandemic represented an anomalous period during which many economic indicators typically used in default rate analysis, such as unemployment and GDP, showed major movements while default rates stayed relatively flat due to government stimulus and regulations. As such, models that rely on these periods can react incongruently with the rest of the time series resulting in models that may not predict as well. However, removing this data also produces issues as many model types necessitate an unbroken time series. Furthermore, this period represents a non-trivial portion of the current business cycle.

In this paper we analyze numerous variations of two broad approaches for dealing with this issue:

- 1) Interpolation of the underlying economic data during the Covid pandemic
- 2) Introduction of binary variables to model the impact of Covid

To conduct this analysis, a time series of point-in-time default rates for a portfolio of Commercial and Industrial (C&I) loans was regressed against a wide range of economic indicators and passed through a filter to select a champion model for each of the approaches mentioned above. We then explored each of these champion models from both a quantitative and qualitative lens to discern their ability to incorporate historical periods during the Covid-19 pandemic.

Methodology

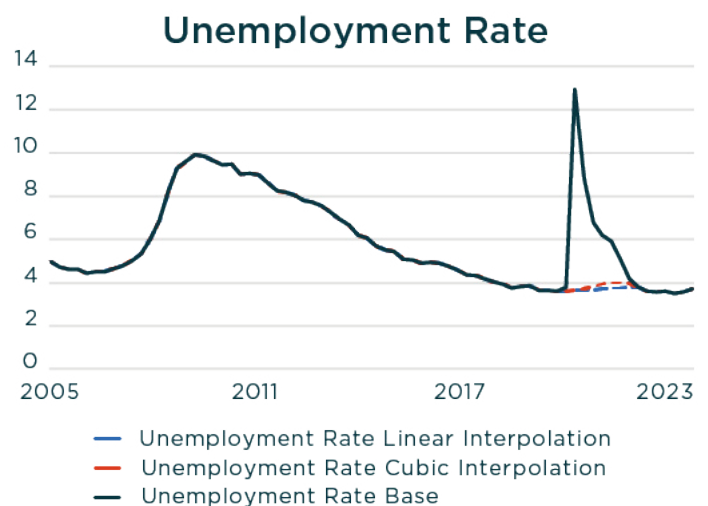
Interpolation of the underlying economic data during Covid

The first approach involves a process known as interpolation. Interpolation involves estimating and constructing new data points when existing data is either unknown or unreliable. The overall idea behind using interpolation for this analysis during the Covid-19 pandemic is that the economic variables during this period were unreliable for use in a default rate model due to government intervention, and therefore we should replace them with an approximation using interpolation that may be more suitable.

For this analysis, two forms of data interpolation methodologies were explored, linear interpolation and cubic spline interpolation. Linear interpolation simply takes two data points and draws a line between them to fill in the missing or unreliable data points between them. For example, if we had a data point with a numerical value of 5 and a subsequent data point with a numerical value of 10 and 4 data points between them that we wanted to fill in, we would calculate each of these in between data points as $5 + (10 - 5) / (4 + 1) * N$, where N is an integer representing the interim period (i.e. 1, 2, 3...).

The second method of interpolation explored, cubic spline interpolation, is more complicated but in general attempts to fit a polynomial between the two data points using the surrounding data to create a smooth curve between the points based on first and second derivatives. The University of Florida provides a more detailed explanation of the mathematics behind cubic spline interpolation here. For this analysis the “spline” function from the “stats” package in R was utilized. A graph of the unemployment rate as well as adjustments during the Covid-19 period from linear and cubic interpolation is shown in Fig. 1.0.

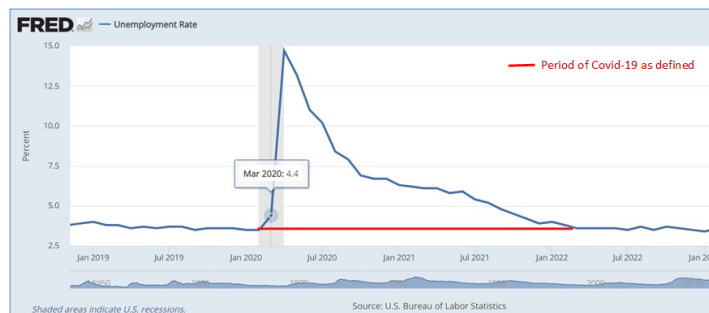
Fig. 1.0



Source: Alter Domus

In both cases, the economic data during the period from March 2020 through March 2022 was stripped and interpolated data was used in its place using the two methods described above. The selection of this period was made as it represented the period of greatest economic turmoil as illustrated in the following graph of unemployment rates from FRED (Fig. 2.0).

Fig. 2.0



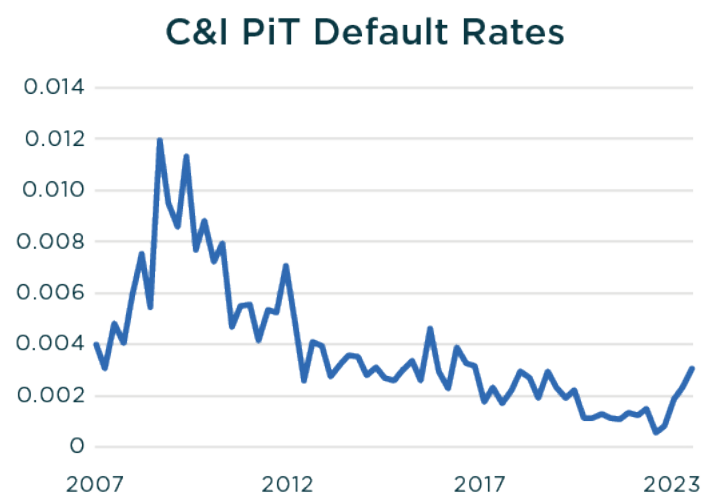
Introduction of binary variables to model the impact of Covid

The second broad approach for incorporating the historical periods during the Covid-19 pandemic involved the introduction of binary variables to model the impact of Covid. A binary variable is a variable that can take on a value of either one or zero, with a one representing the presence of the phenomenon being represented by the variable and a zero representing the lack of its presence. For example, if we were attempting to model a student's GPA based on whether they had a personal tutor, for those students that had a tutor a value of one would be applied and for all others a value of zero.

We attempted to model the 8 quarters from March 2020 until March 2022 selected as described above (Fig. 2.0) using 1, 2, 4, and 8 binary variables. With 1 binary variable a value of one was assigned during this period and a zero everywhere else. With 2 binary values the first variable was assigned ones for the first year and zeroes everywhere else, and the second binary variable was assigned ones for the second year and zeroes everywhere else. For the other variations, the same idea was utilized, subdividing the period during Covid into progressively more binary variables. The use of different numbers of binary variables is intended to capture the different phases of the economy during the Covid-19 pandemic, including 1) a sharp rise (decline) in the economic variables early, 2) a subsequent sharp return to normal levels, and 3) a slower recovery to prior levels. Including a binary variable for each quarter as in the 8 binary approach effectively removes the impact of Covid-19 on the other economic variables entirely.

To judge the ability of each of the approaches for incorporating the historical periods during the Covid-19 pandemic, a data set of quarterly observations from March 2017 through September 2023 of point-in-time probability of default rates for a portfolio of C&I loans and a wide range of economic variables was utilized. The economic variables were selected to represent a broad range of factors that could influence the propensity to default. A graph of the default rate time series and a table of the economic variables used is shown in Fig. 3.0 and Table 1.0.

Fig. 3.0



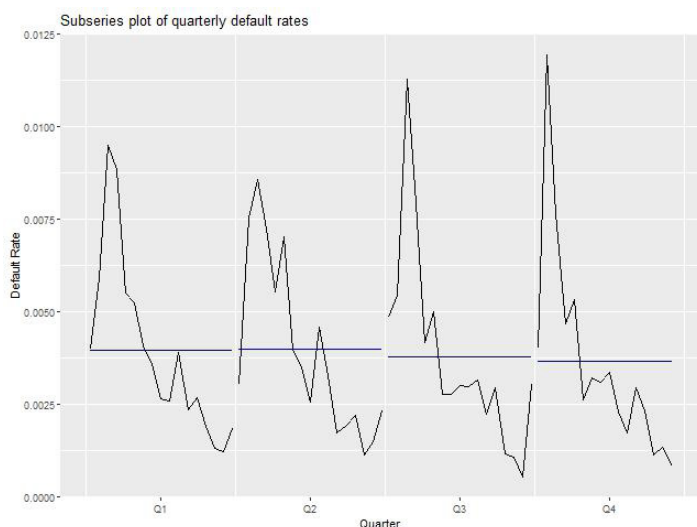
Source: Alter Domus

Table 1.0

Economic Variables
US GDP, real
US GDP for Construction, real
US Retail Sales, real
US Consumer Spending, real
US Wages, real
US Unemployment Rate
Employment in Non-Residential Construction
Term Spread Between 10-Year and 1-Year US Treasuries
The Spread Between BAA Yields and 10-Year US Treasuries
10-Year US Treasuries

The default rate time series was inspected for seasonality (none was found present, see subseries plot Fig. 4.0) and whether any transformations were required (none were deemed necessary). The economic variables were then inspected to determine whether they needed to be transformed to be on the same order of integration as the independent variable. Through that analysis it was determined that the GDP, retail sales, consumer spending, wages and employment variables would be taken as a log difference to represent a percent change, and the unemployment rate would be included both as a level and differenced variable. The spread and treasury rate variables were left unchanged. Additionally, 4 lags up to one-year of each of the economic variables were included to capture the impact of delayed effects on the default rates. A cross-correlation matrix of the economic variables is shown in Fig. 5.0.

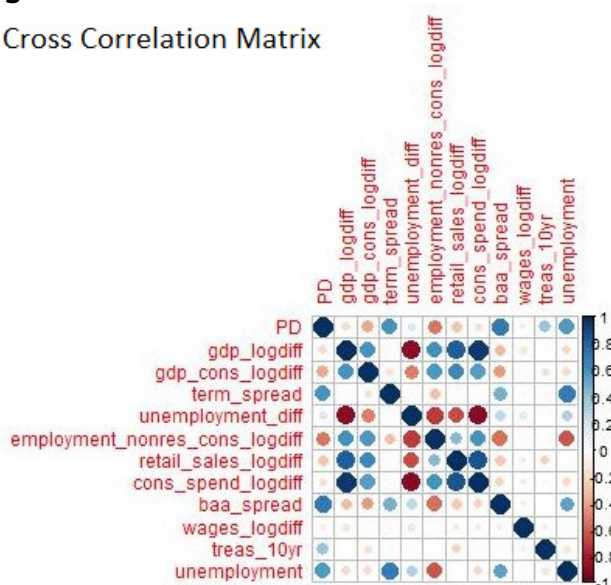
Fig. 4.0



Source: Alter Domus

Fig. 5.0

Cross Correlation Matrix



Source: Alter Domus

All combinations of 3 variables and 2 variables were regressed against the default rate time series (approximately 28,000 model combinations) and passed through a filter that removed any model that had variables with a significance level based on the p-value of more than 0.001 and had signs of coefficients that did not match economic intuition. For example, one would expect default rates and unemployment rates to move in the same direction, so if a model showed a negative coefficient for this relationship, the model would be removed from consideration to avoid spurious correlations.

The remaining models were then tested via 6-fold 10-repeat k-fold cross-validation, which involves randomly splitting the data set into 6 parts, training on 5 parts and estimating the sixth, doing this for each partition, and repeating the process 10 times. The average root mean squared error (RMSE) from this process was taken and the model with the best (lowest) RMSE was selected as the champion model as it represents the model with the best ability to predict on unseen data.

This process was conducted for all the approaches and underlying variations. In the case of interpolating the data, a data set using interpolated values for the period from March 2020 to March 2022 was utilized; one using linear interpolation and the other using cubic spline interpolation. In the case of incorporating binary variables, the process was run while including 1, 2, 4, and 8 binary variables in addition to the 2 or 3 economic variables.

The champion models for the 6 approaches (2 variations of the interpolation approach and 4 for the binary variable approach) plus a control model that was trained using the base data and no binary variables were then analyzed in greater detail to see how they compared against one another. The 7 champion models were compared based on in-sample metrics along with a post-Covid out-of-sample test involving training the models on all data through Covid and then projecting the periods after Covid to judge their ability to predict current default rates.

Note that for the in-sample metrics, the adjustments to the economic data and inclusion of binary variables could artificially inflate the metrics, so all metrics were recalculated to only include the non-Covid observations. Further qualitative analysis was also conducted based on the intuitiveness of the models and the downsides of each approach.

Results

Table 2.0 illustrates the metrics by approach, with values in dark blue representing the best to dark red representing the worst. Adjusted R-Squared represents the amount of variance in the independent variable that is described by the dependent variables expressed as a value between 1 and 0. RMSE is the root-mean squared error and shows the square root of the average squared difference between the value the model predicted versus the true value. MAE is the mean absolute error and is the absolute value of the difference between the value the model predicted versus the true value. In general, a larger value of the Adjusted R-Squared is better and lower values of the RMSE and MAE are better.

Table 2.0

Approach	Variables Used	Adjusted R-Squared In-Sample	RMSE In-Sample	MAE In-Sample	RMSE Post-Covid Out-of-Sample Test	MAE Post-Covid Out-of-Sample Test
Control	BAA yield to 10 Treasury Spread BAA yield to 10 Treasury Spread lagged 3 qtrs 10-Year Treasuries lagged 3 qtrs	84.9709%	0.0931%	0.0749%	0.0826%	0.0728%
Linear Interpolation	BAA yield to 10 Treasury Spread BAA yield to 10 Treasury Spread lagged 3 qtrs 10-Year Treasuries lagged 3 qtrs	85.3718%	0.0918%	0.0743%	0.0722%	0.0634%
Cubic Interpolation	BAA yield to 10 Treasury Spread Change in Unemployment Rate lagged 4 qtrs 10-Year Treasuries lagged 4 qtrs	83.5596%	0.0974%	0.0755%	0.0620%	0.0478%
1 Binary Variable	BAA yield to 10 Treasury Spread BAA yield to 10 Treasury Spread lagged 3 qtrs 10-Year Treasuries lagged 3 qtrs	84.6787%	0.0931%	0.0750%	0.0817%	0.0719%
2 Binary Variables	BAA yield to 10 Treasury Spread BAA yield to 10 Treasury Spread lagged 3 qtrs 10-Year Treasuries lagged 3 qtrs	84.4437%	0.0930%	0.0757%	0.0833%	0.0733%
4 Binary Variables	BAA yield to 10 Treasury Spread BAA yield to 10 Treasury Spread lagged 3 qtrs 10-Year Treasuries lagged 3 qtrs	83.8382%	0.0930%	0.0756%	0.0831%	0.0731%
8 Binary Variables	BAA yield to 10 Treasury Spread BAA yield to 10 Treasury Spread lagged 3 qtrs 10-Year Treasuries lagged 3 qtrs	82.4635%	0.0930%	0.0756%	0.0833%	0.0734%

Source: Alter Domus


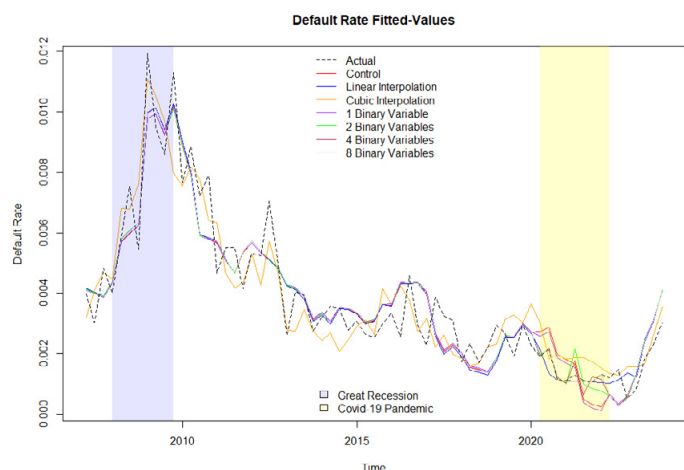
Best  Worst 

Fig. 6.0 plot shows the fitted values of each approach against the actual historical default rates (dashed lined):

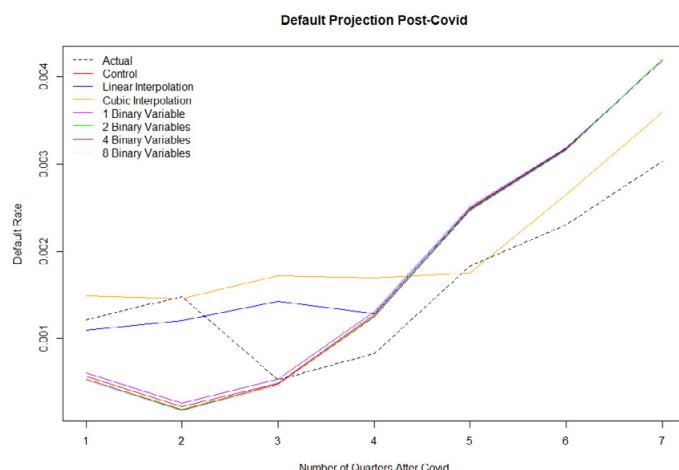
Fig. 6.0



Source: Alter Domus

The following plot shows the projected default rates after Covid compared against the true default rate (dashed line):

Fig. 7.0



Source: Alter Domus

Conclusion

In general, the models using interpolation performed better in-sample and post-covid out-of-sample (Table 2.0), with linear interpolation performing best in-sample and cubic spline interpolation performing best out-of-sample. The downside of interpolated models is they involve modifying the data, whereas binary models simply add independent variables, which may be more theoretically defensible and adds information about the relative impact of Covid on default rates. The model using a single binary variable matched the control in-sample and slightly out-performed out-of-sample (Table 2.0). Including more than one binary variable made the results worse, so it would seem allowing the economic variables to react to default rates without too much modification from an increased number of binary variables during Covid improves the ability of the models to predict during more recent times, thus fewer binary variables is preferable if this approach is selected.

Between linear and cubic interpolation, cubic spline interpolation performed better in recent history predicting out-of-sample and on visual inspection better captured default rates during the 2008 recession. Furthermore, it was the only model that picked up on a variable other than spreads or rates, as it included unemployment rates, which may be preferable for stress testing models as it lessens the impact of any single variable.

While it is ultimately on the modeler to weigh the positives and negatives of each approach, this analysis indicates interpolating the economic data during the Covid-19 pandemic period can improve the quality and ability of models to predict forward looking probability of default rates compared to leaving the data unmodified or including one or more binary variables.

Our ECRA data analyst team at Alter Domus can help determine the best approach for incorporating the period covering the Covid-19 pandemic in your data into probability of default, loss rate, and prepayment models for CECL and stress testing as well as business planning purposes. Additionally, we have pretrained models based on an extensive proprietary data set covering a wide range of product types that can be used as off-the-shelf models or to supplement your existing data.

Please check our website regularly for existing and upcoming research on topics such as:

- Why stress testing and scenario analysis are vital in assessing prepayment risk
- Prepayment rates in a rapidly changing rate environment and their impact on the CECL allowance
- The advantage of incorporating CECL models directly into the stress testing framework
- Outlook on loss rates for the volatile CRE and Consumer Credit markets

Contact us

For more information on how Alter Domus can help you to mitigate risks, please contact:

Harvey Plante

Harvey.plante@alterdomus.com

Brian Hanson

Brian.hanson@alterdomus.com