Quantitative Financial Research
Special Report

# Empirical Examination of Drivers of Default Risk in Prime Auto Loans

**Analysts**

**QFR**
Ahmet E. Kocagil, PhD
+1 212 908-0271
ahmet.kocagil@fitchratings.com

**Asset-Backed Securities**
Ebru Demir
+1 212 908-0682
ebru.demir@fitchratings.com



**Correction**

Changes were made to the third paragraph on page 1. Please contact the analysts listed above for further information.

**January 4, 2007**

## Summary

This study investigates the empirical drivers of default risk in the case of U.S. prime auto loans. In addition to determining the statistical drivers of default risk in a multivariate context, the report focuses on two key areas.

First, the selectivity bias resulting from a particular regressor — new/used (N/U) indicator — in the loan default risk equation is examined in detail; the presence of this indicator variable in the loan default risk regression analysis is likely to result in biased coefficient estimate for the indicator variable in the model due to correlation of unobserved disturbance terms between loan default risk and used car selection equations. Following an estimation procedure suggested by Heckman (1978, 1979), this bias can be eliminated by carrying out a two-stage estimation with a proxy created for N/U indicator as an explanatory variable in a loan's default risk model. The empirical results suggest that the N/U indicator is an endogenous variable that strongly reflects the level of borrower's income.

The second area of focus is the presence of adverse selection in subvented loans. Subvented loans are originated through manufacturer campaigns with low borrower annual percentage rates (APRs) applicable to the eligible borrowers and eligibility usually applies to the borrowers with low credit risk. The historical performance indicates lower default rates for the subvented loans compared with nonsubvented loans. However, subvented loans may still constitute more than 25% of defaulted loans due to borrowers with low bureau scores (<660). To address this problem, the subvented loan indicator variable is entered into the regression model along with the APR variable so that prediction of the riskier portion of the low APR subvented loans can be captured in the multivariate context. As a result, the model produces a positive attribute for the subvention indicator, implying that the presence of subvention is likely to increase the default probability for the set of loans that have bureau scores lower than 660, which is consistent with credit analysts' intuition.

## Data

Default risk of retail loans has been subject to scrutiny in the literature, and the research on auto loans has been more limited due to the lack of corresponding loan performance data. For this research project, a loan level data set of about 500,000 loans originated in the first-quarter 2000 was examined, with performance information received from major prime auto loan issuers. In the data set, the individual loan performances were tracked for at least five years, and several performance information measures, including delinquency and chargeoff, are provided together with a list of borrower characteristics (loan-to-value ratio [LTV], credit bureau score, and APR, among others). The loan default risk model is developed on the sample data sets received. Two different data samples not used for model development are used for validation. One validation sample was extracted from within one issuer's data sample belonging to the early vintages ("out-of-time") of 1995–1999, and the other is the "holdout" sample randomly selected and set aside during the model's development stage.

## Modeling Methodology

The default risk behavior is typically characterized as a dichotomous event, i.e. default or nondefault. Many of the credit risk models involve discrete dependent variable that can take one out of two possible values. Several methods in building classification models with binary response variables have been in practice, using either parametric or nonparametric statistical techniques. Some of the most commonly used methods include tree classification models, logistic or probit regression, neural networks, and duration models.

The econometric models are usually based on parametric formulation or logistic or probit regression approaches. Both approaches are in line with generalized linear models, in which for a response variable Y and a probability measure p,

$$g(p) = X\beta$$

Where p is the Prob(Y=$y_1$), $\beta$ is the parameter vector, X is the vector of explanatory variables, and g is the function through which p is assumed to be linearly related to the explanatory variables.

Expected default frequency can be modeled through logit transformation of event probability, p, in logistic regression:

$$g(p) = \log(p/1-p) = X\beta + u$$

*and*

$$p = \Pr ob(Y = 1) = \Lambda(X\beta) = \frac{\exp(X\beta)}{1+\exp(X\beta)}$$

Where $\Lambda(.)$ is the logit distribution function and *g (p)* is linear in explanatory variables of the vector *X*.

Probit regression is an alternative to logistic, where transformation function, *g(p)*, is expressed as either:

$$g(p) = \Phi^{-1}(p) = X\beta + u$$

*or*

$$p = \Phi(X\beta)$$

Where $\Phi^{-1}$ is the inverse of the cumulative standard normal distribution function.

Those binary models may include binary explanatory variables as a result of a self-selection (purchasing a used car versus new car) and the indicator variable may further be driven by factors that are correlated with *u*. As a result of self-selection and the endogeneity of the indicator regressor, the estimations may result in a biased coefficient, as described in the next section.

Heckman's two-stage estimation procedure has been used mainly in the context of sample selection bias to correct for potential bias resulting from using nonrandomly selected samples. It has also found applications in the context where binary models having binary independent variables that may be endogenous in the regression where endogeneity may result from self-selection. In all studies, bias due to self-selection and endogeneity of an independent variable is statistically confirmed, and corrections using two-stage estimations are shown to be effective in eliminating the bias (Arendt and Holm 2006; Zuehlke and Zeman, 1991). Some research dealt with the sample selection bias issue in consumer credit risk arena (Banasik, 2003). In the context of reject inference, the performance of approved and booked loans is examined while questioning how the rejected population of a lending institution's received applications would affect the decision mechanism if the rejected population were included.

This study begins by examining the bias due to self-selection (purchasing used car versus new car) and endogeneity of an N/U indicator for the auto loan default risk equation. A two-stage estimation was applied to correct for the bias. The next section discusses the method used and explores the procedure for auto loan default risk analysis.

## Theoretical Background: Heckman Two-Stage Estimation

Two-stage estimation methods have been applied in a wide variety of models involving qualitative endogenous variables, truncated dependent or independent variables resulting in either unobserved dependent/independent variable, or data being generated via nonrandomly selected samples. The bias resulting from using nonrandomly selected samples may arise for two reasons. First, there may be self-selection by the individuals on data units being investigated. Second, selection decisions by analysts operate in the same manner as self-selection (Heckman, 1979, p.153). In the 1979 and 1978 studies, Heckman showed the bias that results from using nonrandomly selected samples to estimate behavioral relationships and developed a two-stage estimation procedure to correct for selection bias. The method has been applied to get consistent estimators in the existence of dummy endogenous variables.

Application of the two-stage estimation procedure in this study relates to a commonly expressed risk deriving characteristic for auto loans, namely the N/U indicator (*I*). In this section, the bias arising from using the indicator variable $I$ is shown in the following paragraphs. The derivations are in line with Heckman's two-stage method with application to the study's case.

Suppose $Y_1$ is the output (default versus no default) of a used car loan resulting from a borrower who decides to purchase a used car and finance this purchase, and $Y_2$ is the output of a new car loan that the borrower finances. The two components could usually be expressed in a typical default risk equation, as follows:

$$Y = X\beta + \delta I + u$$

Where
$Y = 1$, if the loan defaulted.
$Y = 0$, if the loan has not defaulted.
$I = 1$, if a used car loan.
$I = 0$, if a new car loan.

$X\beta$ is a set of characteristics relating to loan default risk. For this model, the effect of the loan being used versus new on default risk is measured by the estimate, $\delta$ . The study hypothesizes that the dummy $I$ cannot be exogenous if the decision to buy a used car versus a new car is based on individual self-selection. In other words,

**1:** $I$ may be endogenous to the regression; the characteristics that drive the default risk of a loan may also be influencing the decision to purchase a new or used car, which results in a biased estimate for the coefficient $\delta$ .

**2:** A borrower's selection for buying a used car can also be affected by other factors, e.g. demographics, such as income or age.

Therefore, the indicator, *I*, may not be an ultimate risk attribute of a loan, as it is the outcome of borrower self-selection. If the variable *I* is endogenous, equation 1 must be estimated by two-stage techniques due to the coefficient bias resulting from the situation outlined below.

The indicator function, $I$ — individual self-selection to buy a used versus a new car — is hypothesized to be further influenced by other factors so that:

$$I = Z\gamma + \varepsilon$$

Therefore, $Z\gamma$ is the set of characteristics relating to used car selection.

The regression equation (1) and the selection equation (2) could be related for two reasons: *Y* and *I* may depend on correlated regressors (*X, Z*) and *Y* and $I$ may depend on correlated residuals $(u, \varepsilon)$ .

The study focuses on correlated residuals to show the potential bias resulting from the endogenous indicator function and, therefore, assumes that the residuals of these two equations are standard bivariate, $(u, \varepsilon \mid X, Z) \sim N(0,0,1,1,\rho)$. The theory suggests that $\rho \neq 0$ gives a biased estimate for $\delta$ , with the degree of bias depending on the level of correlation coefficient. Basically, higher correlation results in more serious bias in the coefficient estimate of the indicator *I* variable in equation (1).

The study questions the potential difference in default risk between used versus new car loans originating from a borrower's decision to purchase a new or used car. The following illustrates the rewriting of the two equations in more detail, obtaining consistent and efficient parameter estimates by maximum likelihood estimation of bivariate probit model:

$$Y_i = \beta X_i + \delta I_i + u_i \qquad (1)$$

$$I_i = \gamma Z_i + \varepsilon_i \qquad (2)$$

$$L_i(\delta, \beta, \gamma \mid Y_i, I_i, X_i, Z_i) = \Pr ob(Y_i, I_i \mid X_i, Z_i) \Pr ob(I_i \mid Z_i) \qquad (3)$$

$L$ is the likelihood function of bivariate probit regression equations. It is important to break down the contribution of the probability resulting from the first part (see Heckman, 1979 and Arendt, 2006) of the likelihood function as the second part is simply the probit estimation of the indicator equation (2).

$$\Pr ob(Y_i = 1 \mid I_i = 1, Z_i) = \Pr ob(\beta X_i + \delta I_i + u_i > 0 \mid \varepsilon_i > -\gamma Z_i)$$

$$= \int_{-\gamma Z_i}^{\infty} \Phi\left(\frac{\beta X_i + \delta I_i + \rho\varepsilon_i}{\sqrt{1-\rho^2}}\right) \frac{\phi(\gamma Z_i)}{\Phi(\gamma Z_i)} d\varepsilon \qquad (4)$$

Keeping in mind that:

$$E(Y_i \mid I_i > 0) =$$

$$\beta X_i + \delta I_i + \rho E(\varepsilon_i \mid \varepsilon_i > -\gamma Z_i) = \beta X_i + \delta I_i + \rho \frac{\phi(\gamma Z_i)}{\Phi(\gamma Z_i)}$$

The integral expression can be simplified to keep the integral proper:

$$\Pr ob(\beta X_i + \delta I_i + u_i > 0 \mid \varepsilon_i > -\gamma Z_i) \approx \Phi(\beta X_i + \delta I_i + \rho \frac{\phi(\gamma Z_i)}{\Phi(\gamma Z_i)}) \quad (5)$$

The $\frac{\phi}{\Phi}$ is the inverse Mill's ratio, in essence the bias resulting from usage of endogenous indicator in equation (1). The degree of the bias rises in absolute value with increasing $\rho$.

Unbiased estimates of interest parameters can be obtained via a two-stage mechanism, first by performing equations (1) and (2), estimating correlation coefficient $\rho$ between errors $(u, \varepsilon)$, and in the final stage, by calculating the correction factor $\frac{\phi}{\Phi}$ using the probabilities from the indicator function.

The final stage performs the regression equation (1) with the correction factor — a proxy created for the endogenous indicator variable.

## Application to Auto Loan Level Data

Out of variables included in the data sets provided, the study has determined the six most influential characteristics deriving the default probability for prime auto loans using variable elimination procedures.

The six loan default risk drivers are discussed individually in the subsections that follow. The univariate relationship of each variable to default is illustrated in the charts and tables on pages 5–7. The most influential variables for loan default risk are LTV at origination and the credit bureau score. Term at origination and APR, along with the subvented loan indicator and the proxy for the N/U indicator, are also among the strongly significant characteristics deriving the default risk for prime captive lending loan pools.

The multivariate analysis is carried through probit and logit regression estimations. Both methods give similar model discrimination power and coefficient significance. To be consistent with Heckman's two-stage procedure, probit regression results are provided. Estimation results for initial equation (1) are shown in the above right table.

### MLE of Initial Model

| Model | Direction of Relationship with Default | Weights of MLE Parameter Estimate (%) | Chi-Square Significance Test |
|---|---|---|---|
| LTV | + | 26.60 | *** |
| APR | + | 26.12 | *** |
| Bureau Score | – | 25.28 | *** |
| Term | + | 12.60 | *** |
| Subvention Ind. | + | 6.29 | *** |
| Used Ind. | – | 3.11 | *** |

***Significant at 99.99% significance level. MLE – Maximum likelihood estimation. LTV – Loan-to-value ratio. APR – Annual percentage rate. Ind. – Indicator.

All variables are strongly significant at 99.99% and listed with the order of their individual weight in the regression. LTV, APR, and credit bureau score are the most influential characteristics deriving default probability of loans. The signs of coefficients are directionally consistent, except for the coefficient of the used car loan indicator. The variable of a used car loan indicator is expected to have a positive coefficient with its relationship to default, as used car loans historically have higher default rates than new car loans. The unexpected directional relationship of the used car loan indicator is considered as a result of bias due to endogeneity of the indicator, i.e. potential relationship of indicator equation (2) and the loan default risk equation (1).

Therefore, Heckman two-stage procedure is applied where first stage estimation involves the estimation of default risk equation and the selection equation and obtaining a correlation estimate between those two regressions. The table below illustrates the probit estimation results for selection equation of the used car loan. The significance of the correlation coefficient between the errors $(u, \varepsilon)$ of two equations verifies the endogeneity of the indicator function.

The selection for used car purchase/finance is driven by borrower income and APR. Income and the probability of purchasing a used car is inversely related, while APR and the probability of purchasing

### Endogeneity of Used Ind. and Selection Equation for Used Car Purchase

| Model | Direction of Relationship with Default | Weights of MLE Parameter Estimate (%) | Chi-Square Significance Test |
|---|---|---|---|
| Log (Income) | – | 11 | *** |
| APR (%) | + | 89 | *** |
| Correlation between CoF% 1 and 2 | (0.40) | (35.66) | *** |

***Significant at 99.99% significance level. Used Ind. – Used car loan indicator. APR – Annual percentage rate. CoF% – Chargeoff frequency.

**Second Stage Loan Default Risk Equation**

| Model | Direction of Relationship with Default | Weights of MLE Parameter Estimate (%) | Chi-Square Significance Test |
|---|---|---|---|
| LTV | + | 29.58 | *** |
| Bureau Score | – | 27.98 | *** |
| Subvention Ind. | + | 13.38 | *** |
| Term Ind. | + | 12.60 | *** |
| Used Ind. Proxy | + | 10.46 | *** |
| APR | + | 5.18 | *** |

***Significant at 99.99% significance level. MLE – Maximum likelihood estimation. LTV – Loan-to-value ratio. . Ind. – Indicator. Used Ind. – Used car loan indicator. APR – Annual percentage rate.

a used car is positively related. Other variables such as LTV, credit bureau score, and term at origination are either not significant or highly correlated with either of the regressors in the selection equation. Therefore, selection equation is estimated using income and APR variables.

The second stage estimation is performed using the inverse Mill's ratio, $\frac{\phi(\gamma Z_i)}{\Phi(\gamma Z_i)}$, calculated through probabilities obtained by equation (2). The results of the second stage are presented in the table above.

LTV and bureau score continue to be the most influential variables in a loan default risk equation. The weight of APR dropped due to the proxy variable created to correct the endogeneity of the used car loan indicator. The model behaves better in assigning lower weight to APR and higher weight to the subvented loan indicator, which better captures the risk arising due to adverse selection, not due to the higher APR itself.

The model's discrimination power, measured by c-statistic and K-S statistics, shows the levels of 0.837 and 0.526, respectively.

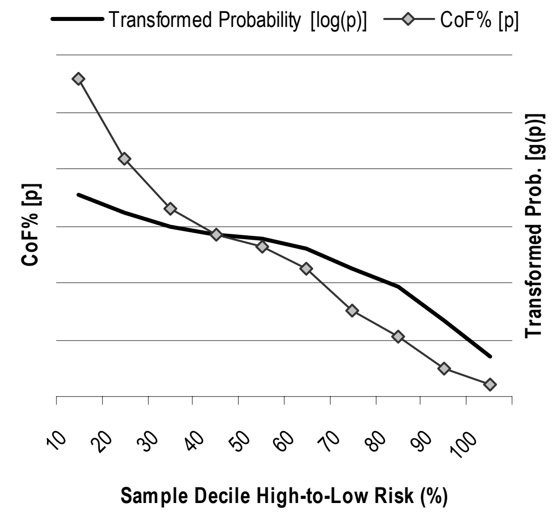
## ■ Model Variables

The model variables and their individual relationships to loan default risk are discussed in the following subsections.

### Loan-to-Value Ratio

As the most influential variable deriving auto loan default risk in this model, LTV at origination is calculated by dividing the original loan amount with the estimated dealer cost. The denominator of estimated dealer cost is a more accurate calculation for LTV, as this value is usually higher than the cash price of the car due to the borrower's specific add-ons to his/her purchase. The above right chart shows the relationship between LTV and the loan's default probability in the univariate context, in which the data sample observations are ranked by LTV and then divided into 10 equally populated subsamples (sample deciles).

**LTV: Actual [p] and Transformed [g(p)] Probabilities by Month 36**



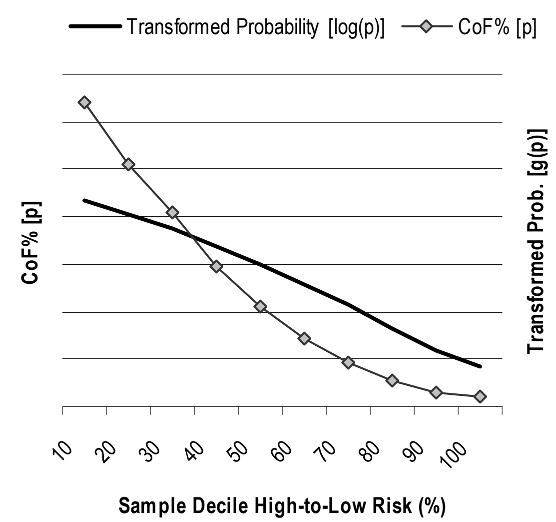

LTV – Loan-to-value ratio. CoF – **Chargeoff frequency**. Prob. – Probability.

The CoF% line illustrates the actual chargeoff frequency within the deciles of the sample, with the lowest decile representing the highest risk bucket (high LTV values). The nonlinearity in the relationship is smoothed by probit transformation of the actual probabilities shown by the transformed probability line.

### Credit Bureau Score

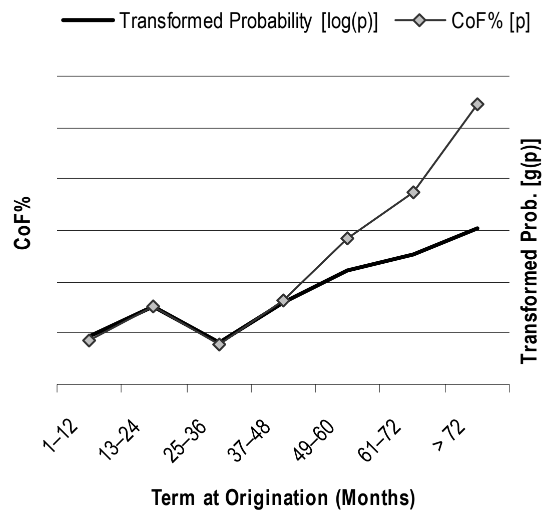
The univariate relationship of the credit bureau score to the default risk is illustrated by the chart below.

**Credit Bureau Score: Actual [p] and Transformed [g(p)] Probabilities by Month 36**




CoF – Chargeoff frequency. Prob. – Probability.

**Term at Origination: Actual [p] and Transformed [g(p)] Probabilities by Month 36**




CoF – Chargeoff frequency. Prob. – Probability.

The data sample observations are ranked by credit bureau score and then divided into sample deciles. The CoF% line illustrates the actual chargeoff frequency within the deciles of the sample, with the lowest decile representing the highest risk bucket (low credit bureau score). The nonlinearity in the relationship is captured by probit transformation of the actual probabilities shown by the transformed probability line.

### Term at Origination

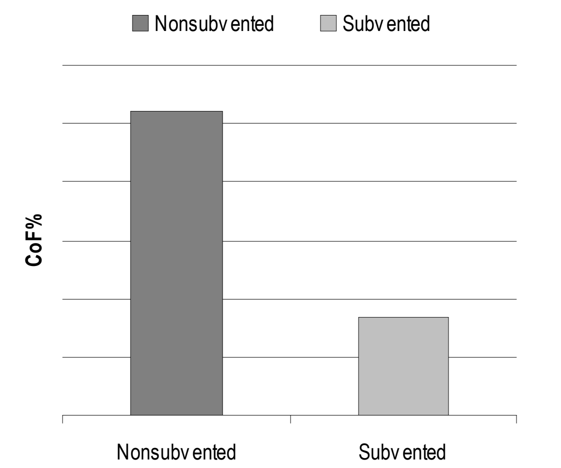
As the fourth important variable in the model, term at origination shows a slightly lower impact on default risk. Term at origination is considered to be more important on loss severity rather than default risk. The chart below shows the significance and directional consistence of the term variable in determining a loan's default risk.

The spike in the CoF% shows the loans with original terms of 13–24 months experience the highest default frequency. The univariate relationship of a variable to default is illustrated the chart above. The variable's nonlinear behavior in the univariate context may not be the same in a multivariate context.

### Subvention Indicator

Subvented loans (loans that are originated through manufacturer campaign programs with lower APRs to the eligible customers) have been important element of captive lending since early 2000. Borrowers meeting eligibility requirements are potentially low-risk customers. The above right chart shows the CoF% of subvented and nonsubvented loans in the development sample.

**Subvention Indicator: Default Frequency of Subvented vs. Nonsubvented Loans**



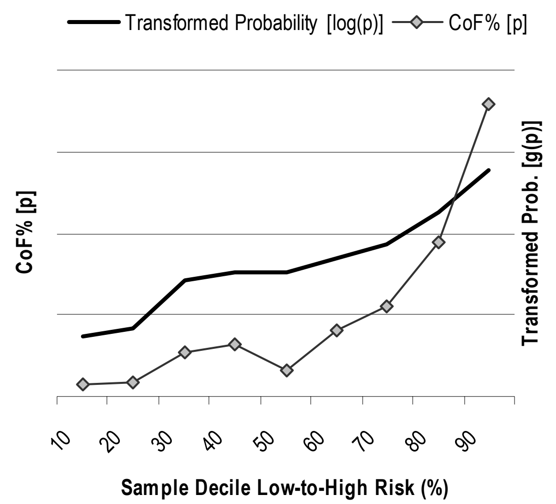

CoF – Chargeoff frequency.

Not all subvented loans are actually low-risk profile. A potential risk of subvented loans is in the case of adverse selection, where a limited number of high-risk borrowers may still be extended loans with incentives. In fact, subvented loans constitute about 27% of the total charged off loans in one issuer data set.

This portion is mainly originating from borrowers with low credit bureau scores. Subvented loan

**Subvented and Nonsubvented Loans in Charged Off Loans**

**APR: Actual [p] and Transformed [g(p)] Probabilities by Month 36**




APR – Annual percentage rate. CoF – Chargeoff frequency.

origination is a distinguishing feature for captive lending, and the portion of subvented loans in charged off loans is considerable. Adverse selection due to subvented loan originations is successively captured through a positive coefficient estimate of the subvented loan indicator in the model developed in this study.

### New/Used Indicator Proxy

The risk inherent in used car loans is captured through a proxy created for the N/U indicator. The indicator variable is hypothesized to be endogenous to the loan's default risk equation and the selection equation for used car purchase/finance is estimated separately, as discussed. The proxy, defined as the ratio of marginal to cumulative probability of selecting a used car, is positively related to loan default risk, in line with the indicator's relation to default. The variable itself is not the least but among the variables with less weight in the model.

### APR

With lenders developing advanced risk-pricing systems, APR is considered as a significant predictor of default. Although considered by some to include potential noise due to changing economic environment and market rates, business cycles, and marketing campaigns, the study views the information content of APR for determining a loan's default risk as strong and should not be ignored. The potential noise in the APR can be corrected through detailed analysis. The model — having the APR and subvention indicator characteristics as explanatory variables — controls potential risk due to adverse selection, as discussed.

## Model Predictive Ability for CoF% and Validation

The predictive ability at the development data set is tested using two out-of sample datasets. One sample is an out-of-time sample of the loans originated from Dec. 31, 1995 through Dec. 31, 1999. The second validation set is the random sample of total loans received but not used in the model development process.

Model performance is assessed through actual versus predicted cumulative CoF% at month 36 for development and out-of-sample sets.

Actual and predicted chargeoff frequencies are reasonably close. The lower chargeoff frequency observed in the hold-out-sample, compared with the development sample, is well captured by the model. Out-of-time sample validation for similar studies is considered essential but usually not performed due to the unavailability of information from different time of loan originations. The period of loan originations (1996–1999) in the out-of-time set belongs to a different economic environment and is a better validation source for the analysis compared with the holdout sample, which is a random sample of the total development sample. Marketing and lending practices, competitive pressures, volumes of loans originated, and growth rates were different in the mid-1990s compared to post-2000. More importantly, the out-of-time sample does not belong to the record low interest rate phase of the economy. The difference in actual and the model output of predicted chargeoff frequency for different vintages illustrates the robustness of the model to be used over time for different cycles.

**Loan Level Model Validation Test**
(CoF%)

| | Actual | Predicted |
|---|---|---|
| Development Sample | 4.7980 | 4.7930 |
| Holdout Sample | 2.2606 | 2.2759 |
| Out-of-Time Sample | 3.9795 | 3.6726 |

CoF% – Chargeoff frequency

## Conclusion

Self-selection and endogeneity of indicator variables may lead to seriously biased results, and Heckman's two-stage procedure accounts for bias in such models. In this report, the effect of an endogenous indicator variable resulting from self-selection of a borrower's decision to buy a used or new car in the loan's default probability regression is considered. An endogenous self-selection indicator variable may result in a seriously biased coefficient. The study shows the bias resulting from such a variable — N/U indicator — in loan default risk regression and found that some factors driving loan default risk may also affect the choice to purchase a used versus a new car in addition to other factors (e.g. income) strongly driving this choice. Therefore, the individual selection to buy a used or new car is considered in a separate modeling and a proxy for the N/U indicator created through the probabilities obtained by the selection equation (inverse Mill's ratio) is used in the loan default risk regression. The bias is eliminated, and the coefficient of the proxy variable for a used car indicator is directionally consistent (positive) with the loan's default risk.

As a result of this analysis, LTV is viewed as the most influential risk driving factor. LTV proves to be a powerful risk driver for secured consumer loans with installments (e.g. auto, mortgage loans) but even more so for auto loans. Higher LTVs at origination are attributed to higher default risk due to a lower equity set by borrower. High LTVs for mortgage loans may recover with equity valuation over time. However, with the sharp decrease in auto values over a short period, auto loans may not recover from the risk inherent in high LTVs at the origination.

The study used the APR variable in a loan's default risk equation since it is hard to avoid this variable given its strength in the model runs. As lending institutions would receive detailed information regarding borrower characteristics at the application stage to determine loan pricing, the study regards the information content within the APR as essential. Although APR may also be considered to include noise due to changing environments, the weight attributed to APR is lower when combined with the subvention indicator variable in the regression. Therefore, the study infers that this normalizes the potential overinfluence of the APR in the model. The study has further determined that the model is able to capture adverse selection due to the part of the subvented loans that could be riskier than previously thought.

Model estimations are carried via SAS statistical software, and its qualitative and limited dependent variable model procedures are used for maximum likelihood estimations (probit, logit) for single regressions. The two-stage estimation method is used to account for the bias due to self-selection and endogeneity.

The data set has limitations by including only first-quarter 2000 vintages. Borrower performance, the economic environment, and the lending practices of financing institutions have changed since then. Fitch recognizes the limitations due to the shorter vintage period of the loans in the samples provided, as well as the economic environment changes. However, Fitch believes this study has contributed to its understanding of auto loan performance behavior in the context of prime auto captive lending.

The presented modeling effort has the potential of serving as an additional tool for assessing the risk of prime auto loan pools and rating asset-backed securities backed by similar pools.

## References


- Arendt, Nielsen Jacob and Anderson Holm, 2006, "Probit Models with Binary Endogenous Regressors," Discussion Papers on Business and Economics, No.4/2006, University of Southern Denmark, Department of Business and Economics.
- Banasik J., J Crook and L Thomas, 2003, "Sample Selection Bias in Credit Scoring Models," Journal of Operational Research Society, 54, 822-832.
- Heckman James, 1978, "Dummy Endogenous Variables in a Simultaneous Equation System," Econometrica, Vol. 46, No. 6 (July 1978).
- Heckman, 1979, "Sample Selection Bias as a Specification Error," Econometrica, Vol. 47, No. 1.
- Zuehlke, W. Thomas and Allen R. Zeman, 1991, "A Comparison of Two-Stage Estimators of Censored Regression Models," The Review of Economics and Statistics, 73.