Chapter 10
Loss Given Default (LGD) and Recovery Rates

INTRODUCTION

This chapter introduces models for loss given default (LGD) and recovery estimation. It is important to note that a loss arises only in the event of default and is conditional on the default event; hence it is called loss *given* default. Exhibit 10.1 shows that loss is conditional on the default events and that the loss given default is a continuous variable with density f.

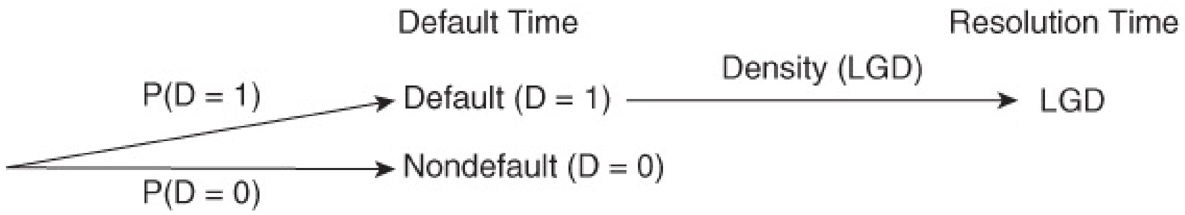


Exhibit 10.1 Conditionality of LGDs

LGDs are commonly expressed as a ratio and related to the outstanding amount or exposure at default (EAD). In other words, LGD is essentially a loss *rate* given default. The recovery rate is then $1 - LGD$.

Definition of Default

In order to unambiguously quantify the loss given default, you first need to have a well-framed definition of default. For a thorough discussion, see Van Gestel and Baesens (2009). For non-retail exposures, rating agencies such as Moody's, Standard & Poor's (S&P), and Fitch use definitions of default that, although to a large extent overlapping, are not identical. Hence, if you use different definitions of default, then of course you cannot compare the resulting default and loss rates. More specifically, there is a direct interrelation between the default definition, the default rates, and the loss or LGD values. Hence, when LGD rates are reported, it is always important to ask for the default definition adopted, to make sure you can correctly interpret and benchmark them.

Usually, a bank will distinguish among different types of defaults. An operational default is due to technical issues on the obligor side. For example, an obligor is accidentally late when making the payment. A technical default is a default due to an internal information system issue. For example, the payment was made on time, but on the wrong account. A real default is a default due to financial problems or insolvency. These are the defaults we are interested in when modeling LGD.

In case of default, various actions can take place. First, there can be a cure. This means a defaulter will pay back all outstanding debt and return to a performing or thus nondefaulter status with no accompanying loss. There could also be a restructuring or settlement, whereby the bank and the defaulter work out a recovery or repayment plan. The latter could, for example, result in an extension of the loan maturity to reduce the monthly installment amount. This usually comes with a medium loss. Finally, there could also be liquidation, repossession, or foreclosure, which implies that the bank takes full possession of the collateral asset, if available, and sells it by starting up a bankruptcy procedure. Depending upon the value of the collateral, this may come with a high loss.

When modeling LGD, it is of key importance that the default definition used is the same as for PD because PD and the LGD will be combined to calculate both expected and unexpected loss. Note that changing the default definition simultaneously impacts both the PD and the LGD. If you would, for example, relax the default definition from 90 days to 60 days in payment arrears, then the default rates and PD may increase, but the loss rates and LGD may decrease. Hence, the combined effect in terms of expected loss stays relatively constant.

Cures are those defaulters that become nondefaulters and return to performing by repaying all outstanding debt. The corresponding LGD will thus be zero, or close to zero. As already mentioned, note that this depends on the default definition. Relaxing the definition of a default, for example from 90 to 60 days, will typically increase the number of cures. In case of multiple defaults, you could opt to include only the last default event and also relate the PD and EAD to this.

Definition of LGD

The loss given default can now be defined as the ratio of the loss on an exposure due to the default of an obligor to the amount outstanding at default. As such, it is the complement of the recovery rate or, in other words, LGD equals 1 minus the recovery rate. Important to note here is that LGD focuses on economic loss, rather than accounting loss. Hence, all costs, but potentially also benefits, need to be properly taken into account when defining the LGD. Example costs are: the costs for realizing the collateral value, administrative costs incurred by sending collection letters or making telephone calls with the defaulted obligor, legal costs, and time delays in what is recovered. Also, benefits such as interest on arrears, penalties for delays, or other commissions can be considered.

LGD can be measured using various methods such as the workout method used for both corporate and retail exposures, the market approach used for corporate exposures, the implied historical LGD approach used for retail exposures, and the implied market approach used for corporate exposures. In what follows, we will discuss each of these in more detail.

The most popular method for defining LGD is the workout method, which is frequently adopted for both corporate and retail exposures. The idea here is to work out the collection process of a defaulted exposure and carefully inspect the incoming and outgoing cash flows. Both direct and indirect cash flows should be considered. Example indirect costs could be the operating costs of the workout department. These cash flows should then be discounted to the moment of default to calculate the loss. In Exhibit 10.2 you can see a simplified example.

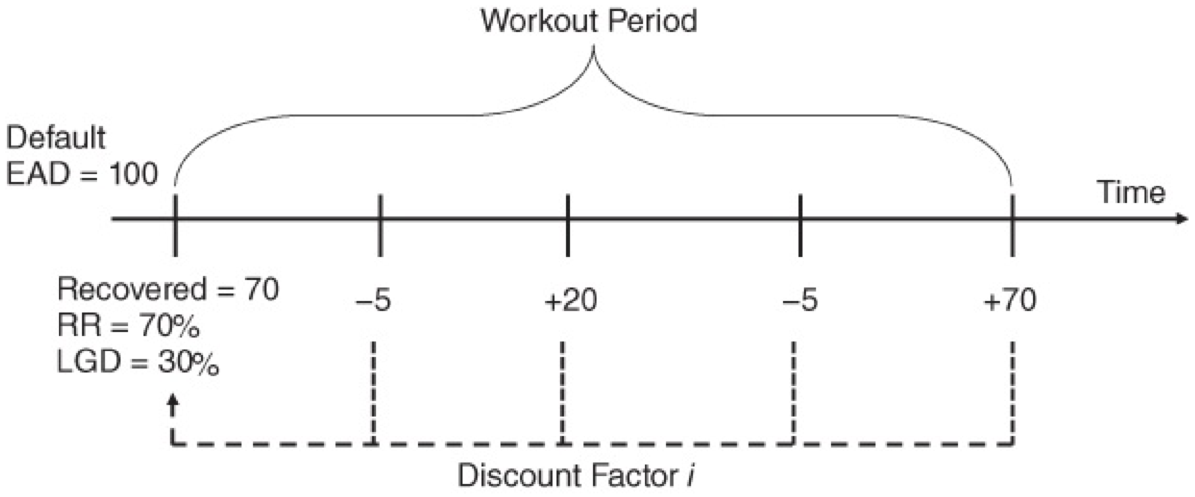


Exhibit 10.2 Workout LGDs

Let us assume an exposure goes into default with an EAD of $100. Soon after default, the collection department will contact the defaulted obligor either by telephone or by sending a collection letter. Let us assume the cost for this equals $5. This is followed by the obligor paying back $20, which is clearly not enough to cover all outstanding debt. So, the collection department contacts the obligor again at a cost of $5. Let us say that the obligor does not react, so the bank decides to materialize the collateral and receives $70 for it. We can now discount all these cash flows back to the moment of default using a discount factor, which we leave unspecified for the moment. Let us say that the discounted amount equals $70. Note that this is smaller than the sum of the four numbers, which equals $80, because of the discounting that has been applied. In other words, this means that $70 has been recovered from the $100 EAD, hereby giving a recovery rate of 70 percent and an LGD of 30 percent.

Another way to measure the LGD is by using the market approach. The idea here is to look at firms that went bankrupt and have debt securities such as bonds or loans trading in the market. Once the bankruptcy event has occurred, the bonds will become junk bonds and investors will start trading them based on what they think they will recover from the bankrupt firm. To allow for some time for the market to stabilize and absorb all information, this approach looks at the market price (e.g., one month after the bankruptcy or default event). This market price is then used as a proxy for the recovery rate, which will then also allow a calculation of the LGD as 1 minus the recovery rate. Note that this approach works only for debt securities that trade in the market, and is thus not applicable for retail exposures. This approach was followed by Moody's in its LossCalc tool (see Gupton and Stein 2005). The implied historical LGD method works by using the PD estimates and the observed losses to derive the implied LGD. In other words, it calculates the expected loss first and then makes use of the expression $EL = PD \cdot LGD$. The LGD can then be computed as the expected loss divided by the PD.

Another way to measure the LGD is the implied market LGD approach. This is a very theoretical approach and we've rarely seen it used in the industry. It analyzes the market price of risky but not defaulted bonds using asset pricing models such as structural or reduced form models. It then finds the spread above the risk-free rate, which reflects the expected loss, and backs out the LGD from there. Some of the key concerns of this approach are that the market price is only partially determined by the credit risk and also includes risk aversion premiums.

Finally, according to the Basel Capital Accord, the definition of loss used in estimating LGD is economic loss. As already stated, this means that every cash flow or cost related to the default should be properly taken into account. In the foundation internal ratings based (IRB) approach, estimates of the LGD are prescribed in the Accord. For corporates, sovereigns, and banks, the following applies: Senior claims on corporates, sovereigns, and banks not secured by recognized collateral will be assigned a 45 percent LGD. All subordinated claims on corporates, sovereigns, and banks will be assigned a 75 percent LGD; see Basel Committee on Banking Supervision (2006).

Stage I: Computing Observed LGD

When computing observed LGDs from workout cash flow observations, various issues occur, such as:

- The data set should cover at least a complete business cycle.
- The workout or resolution period needs to be defined.
- Incomplete workouts need to be handled.
- The discount rate needs to be defined.
- LGDs outside the normal range should be handled.
- Indirect costs should be included.

In what follows, we will elaborate on each of these. Given all these ingredients, the LGD is then calculated as the exposure at default (EAD) minus the present value of all cash flows, including internal costs:

$$LGD = \frac{EAD - \sum_{t=1}^{T}(CF_t/(1+r_t)^t)}{EAD}$$

where CF_t is the cash flow and r_t is the discount rate for time t. As an example, consider an EAD of $50,000 at time of default, and the stream of cash flows shown in Exhibit 10.3.

	Date of Default	Year 1	Year 2	Year 3
EAD	50,000			
Cash flows		20,000	10,000	10,000

Exhibit 10.3 Cash Flow Example

The LGD can then be computed once the workout process has finished (i.e., after year 3). If the bank uses a discount rate of, say, 5 percent per annum, the present value of recoveries at the date of default becomes $20{,}000/1.05 + 10{,}000/1.05^2 + 10{,}000/1.05^3 = 36{,}756$, and the LGD is $\frac{50{,}000-36{,}756}{50{,}000} = 26.5\%$.

Cash Flows

Recoveries

Cash recoveries refer to actual cash flows to be collected from defaulters during the workout period. Cash recoveries are easy to track since the details are typically recorded in one or more bank databases. Another form is noncash recoveries, including repossession of collateral or restructuring loans to support borrowers with their payments. Noncash recoveries are often treated case by case.

Recoveries can also be classified according to their source: product, collateral, guarantee, and residual (unsecured). Product recoveries relate to trade credit whereby the outstanding balance can be reduced if the underlying goods can be easily and quickly sold to buyers. Collateral recoveries include appraisal collateral values, carrying cost, and liquidation. Note that one collateral can be pledged for multiple facilities. In that case, collateral recovery should be reallocated specifically to each facility. Allocation approaches can be based on either pledge value or EAD of each facility. Guarantee recoveries involve a third party who is willing to pay (a part of) the outstanding balance owed in case of default. Note that guarantees can be either closely related or unrelated to the borrower. In case of a close relation (e.g., parent companies), guarantees may be treated as PD mitigants. Otherwise, if guarantees are unrelated to borrowers, they are considered LGD mitigants, meaning that they do not influence the default event, but provide support when default occurs. Unsecured recoveries include remaining parts of the assets that banks can claim after product, collateral, and guarantee recoveries.

Costs

We already mentioned that LGD represents economic loss. Hence, indirect costs should also be properly taken into account, as illustrated by the following quotes:

Committee of European Banking Supervisors (CEBS) (2005):

> Workout and collection costs should include the costs of running the institution's collection and workout department, the costs of outsourced services, and an appropriate percentage of other ongoing costs, such as corporate overhead.

Federal Register (2007):

> Cost data comprise the material direct and indirect costs associated with workouts and collections.

Federal Register (2007):

> Material indirect costs, costs of running the collection and workout department, costs of outsourced services, appropriate percentage of overhead, must be included.

The question now is how to take into account these indirect costs. Obviously, indirect costs are not tracked on a defaulter-by-defaulter basis; they need to be calculated on an aggregated level. Most banks conduct a small accounting exercise to calculate the indirect cost rate. In Exhibit 10.4 is an example.

Year	Total EAD of Files in Workout (End of Year)	Annual Recovered during Year	Internal Workout Costs per Year
2010	1,000	250	20
2011	1,500	500	28
2012	800	240	12
2013	1,250	350	27

Exhibit 10.4 Workout Costs

Suppose we have four years of data, from 2010 to 2013. The second column represents the total exposure at default of files in workout measured at the end of the year. Note that because the workout period usually lasts longer than a year, most of these numbers include double counts. In other words, in the number 1,500 measured in 2011, there are some observations that were also already included in the 1,000 measured in 2010. The next column represents the amount recovered in each year. There are no double counts here. Finally, the last column represents the aggregated internal workout costs per year. This includes the costs of the workout department, the salaries of the people working in it, the electricity, the computer hardware and software, and so on.

We can now calculate two cost rates. The first one uses the exposure at default as the denominator. The assumption here is that higher workout costs are incurred for higher exposures at default. The cost rate can now be calculated in a time-weighted or pooled way. The time-weighted cost rate is just the average for all years of the workout costs divided by the exposure at default. For our example, this becomes $1/4 * (20/1{,}000 + 28/1{,}500 + 12/800 + 27/1{,}250)$ or 1.8 percent. The pooled cost rate divides the sum of all workout costs by the sum of all exposure values. In our case, this becomes $(20 + 28 + 12 + 27)/(1{,}000 + 1{,}500 + 800 + 1{,}250)$ or 1.91 percent. A disadvantage when using this cost rate is that it has to be multiplied by the number of years the workout lasted.

Another way of calculating the cost rate is by using the amount recovered as the denominator. The assumption here is that higher workout costs are incurred for higher recoveries. Again, the cost rate can be calculated in a time-weighted or pooled way. The time-weighted cost rate is the average for all years of the workout costs divided by the recovery amounts. For our example, this becomes $1/4 * (20/250 + 28/500 + 12/240 + 27/350)$ or 6.5 percent. The pooled cost rate divides the sum of all workout costs by the sum of all recovered amounts. In our case, this becomes $(20 + 28 + 12 + 27)/(250 + 500 + 240 + 350)$ or 6.49 percent. The advantage of this approach is that it is independent of the length of the workout period because each amount was recovered during one year only. Hence, this is simpler to implement.

Discount Factor

We have already mentioned that LGD represents economic loss. Consequently, when quantifying the LGD, one should also take into account the time value of money. One dollar today is worth more than one dollar tomorrow. Hence, we should apply discounting. A key problem when applying discounting is setting the discount rate. The Basel Committee on Banking Supervision ([2005a, 2005b]) mandates that the discount rate includes the time value of money and a risk premium for undiversifiable risk:

> When recovery streams are uncertain and involve risk that cannot be diversified away, net present value calculations must reflect the time value of money and a risk premium appropriate to the undiversifiable risk. In establishing appropriate risk premiums for the estimation of LGDs consistent with economic downturn conditions, the bank should focus on the uncertainties in recovery cash flows associated with defaults that arise during the economic downturn conditions. When there is no uncertainty in recovery streams (e.g., recoveries derived from cash collateral), net present value calculations need only reflect the time value of money, and a risk free discount rate is appropriate.

A number of discount rate approaches have been proposed in the literature:

- Contract rate
- Weighted average cost of capital (WACC)
- Return on equity (ROE)
- Market return on defaulted bonds
- Equilibrium returns based on the capital asset pricing model (CAPM)

For a summary of some of these approaches, we refer to Maclachlan (2004) and Brady et al. (2006). Global Credit Data has recently undertaken a comparative study and found that the WACC and equilibrium approaches provide reasonable techniques to reflect the time value of money and systematic risk.

The contract rate is quite commonly used, although it has been criticized as it relates to the pricing at origination. Hence, it does not reflect the interest rate and price for systematic risk at the time of default and is not applicable to the distressed conditions to which recovery cash flows may be exposed.

Discount rates have recently been identified as a source of risk weight inconsistencies between financial institutions. This, and the fact that interest rates are at historic lows, means that some prudential regulators have considered imposing minimum floors. For example, the Prudential Regulation Authority of the United Kingdom stated the following (see Prudential Regulation Authority 2013):

> The PRA expects firms to ensure that no discount rate used to estimate LGD is less than 9%.

Whilst the controversy on discount rates continues, we see great merits in modeling risk-sensitive discount rates that are in line with the systematic risk that governs the realization of losses.

Workout Period

The length of the workout period can vary depending upon the type of credit, the workout policy of the financial institution, and the local regulation. Some regulators such as the Bank of International Settlements (BIS) and the Hong Kong Monetary Authority (HKMA) have provided further input on this. For example, the workout period can finish when the unrecovered value is less than 5 percent of the EAD, one year after default, at the time of repossession of the collateral, or at the time of selling off the debt to a collection agency. On average, many financial institutions have workout periods of two to three years. For a recent study and an international comparison on workout periods, see Betz, Kellner, and Rösch (2016).

Incomplete Workouts

We already briefly touched upon the issue of incomplete workouts. Incomplete workouts represent obligors that have gone into default status, and for which the workout process is still ongoing. Some regulatory authorities have provided further input about incomplete workouts. The Committee of European Banking Supervisors (CEBS), which is the predecessor of the European Banking Authority (EBA), initially mentioned in its regulation (see Committee of European Banking Supervisors (CEBS) 2005):

> Institutions should incorporate the results of incomplete workouts as data/information into their LGD estimates, unless they can demonstrate that the incomplete workouts are not relevant.

Part of this was copied by the Prudential Regulation Authority (PRA) of the United Kingdom as follows:

> In order to ensure that estimates of LGDs take into account the most up to date experience, we would expect firms to take account of data in respect of relevant incomplete workouts, i.e., defaulted exposures for which the recovery process is still in progress, with the result that the final realized losses in respect of those exposures are not yet certain.

See Prudential Regulation Authority (2013).

The most recent EU regulation does not mention anything further about incomplete workouts. Incomplete workouts can be treated in various ways. A first treatment option is to calculate the current LGD of an incomplete workout and use this as the final LGD value in the data set. This is a very conservative approach giving an upward bias to the LGDs. In other words, using this approach, the LGDs will be overestimated since additional future recoveries are likely. Note, however, that we have seen some banks systematically disregard recoveries after three or five years into their LGD calculations. Another option is to use expert or predictive models, which estimate the final LGD of an incomplete workout based on various characteristics such as date of default, percentage already collected, time of collection, and so on. The easiest method is simply to ignore incomplete workouts and include only complete workouts in the LGD modeling data set. This approach is quite commonly used in practice. Finally, you can also use survival analysis whereby the loss amount is considered as a censored variable. Note, however, that this is highly theoretical and not commonly applied in industry. For more information about this, we refer to the paper of Stoyanov (2009).

Business Cycle

The data set used for LGD modeling should cover at least a complete business cycle. The obvious question that follows is: What is a business cycle? Preferably, the data should include one or two downturn periods. This will be handy for the LGD calibration as we will discuss later. Note that you do not need to attach equal importance to each year of data. Hence, if you think data of five or seven years ago is less relevant today, you can attach a lower weight to it. Downturn periods are generally defined as periods with negative GDP growth. Hence the number of years of data required depends on the analyzed economy. For example, Japan has been in an extended economic downturn period since 1993 whereas Australia has not experienced an economic downturn since 1991.

LGDs Outside the Interval [0; 1]

Typically, any real-life LGD data set will contain negative LGDs and LGDs exceeding 100 percent. An obvious question is: Where do these extreme values come from and how should they be treated? A negative LGD is the same as a recovery rate exceeding 100 percent. There could be various reasons for this. One example is that the EAD was measured at the time of default, and the claim on the borrower increased after that because of fines or fees, and everything was recovered. In other words, the amount recovered was higher than the EAD, thereby giving a recovery rate of higher than 100 percent or a negative LGD. Another reason for a negative LGD could be a gain in collateral sales. Negative LGDs should be capped atzero. For example, the PRA expects firms to ensure that no LGD estimate is less than zero. Vice versa, LGDs exceeding 100 percent correspond to negative recovery rates. Also here, there could be various reasons for this, such as additional recovery costs were incurred and nothing was recovered. Alternatively, it could have been that additional drawings after default were considered as LGD, thereby seriously increasing the costs. Also here, it is recommended to cap such LGDs at 100 percent. The models that we will present later in this chapter provide ways of dealing with this sort of truncation and censoring.

Stage II: LGD Modeling

In a second stage, after the observed LGDs have been computed, a bank is generally interested in modeling the determinants of LGD (such as values of collaterals) and provide LGD forecasts *before* a default has happened, but also *after* a default. The remainder of this chapter details several variants of regression-type models that can be applied by a bank for modeling and forecasting LGD. The data set we use originates from a European bank and contains workout LGDs from mortgage loans that have been accordingly anonymized and preprocessed. The models can, however, also be applied to other LGD definitions, such as market LGDs.

Throughout the text we define LGD as a fraction and the recovery rate as (1 − LGD). Similar to PD models, the range of values for LGD requires specific models and considerations. First, LGD usually ranges between zero and one $[0;1]$, and values below zero or greater than one only rarely occur (or do not occur at all; see previous discussion). Second, special cases are values of exactly zero or exactly one, which indicate a zero loss or a total loss, respectively. A zero loss can occur when a default is fully cured. Third, LGDs by definition are conditional quantities and can be observed only if a default has happened. This imposes a sample selection problem when defaults and LGDs are dependent, which should be accounted for in order to avoid inconsistent parameter estimates.

Based on these considerations, this chapter starts with descriptive LGD statistics and some transformations thereof, which will be used in the empirical models. Next, marginal or stand-alone models for LGD are presented that are widely used in practice and research but do not account for the sample selection issue and LGD-PD dependence. Subsequently, the class of joint PD-LGD models that account for selection and censoring is illustrated, and differences between both are discussed. Finally, the chapter concludes with an outlook on advanced models that extend the models along several lines (e.g., random effects, downturn LGDs, and higher-order dependencies between defaults and LGDs). Exhibit 10.5 shows the different model classes.

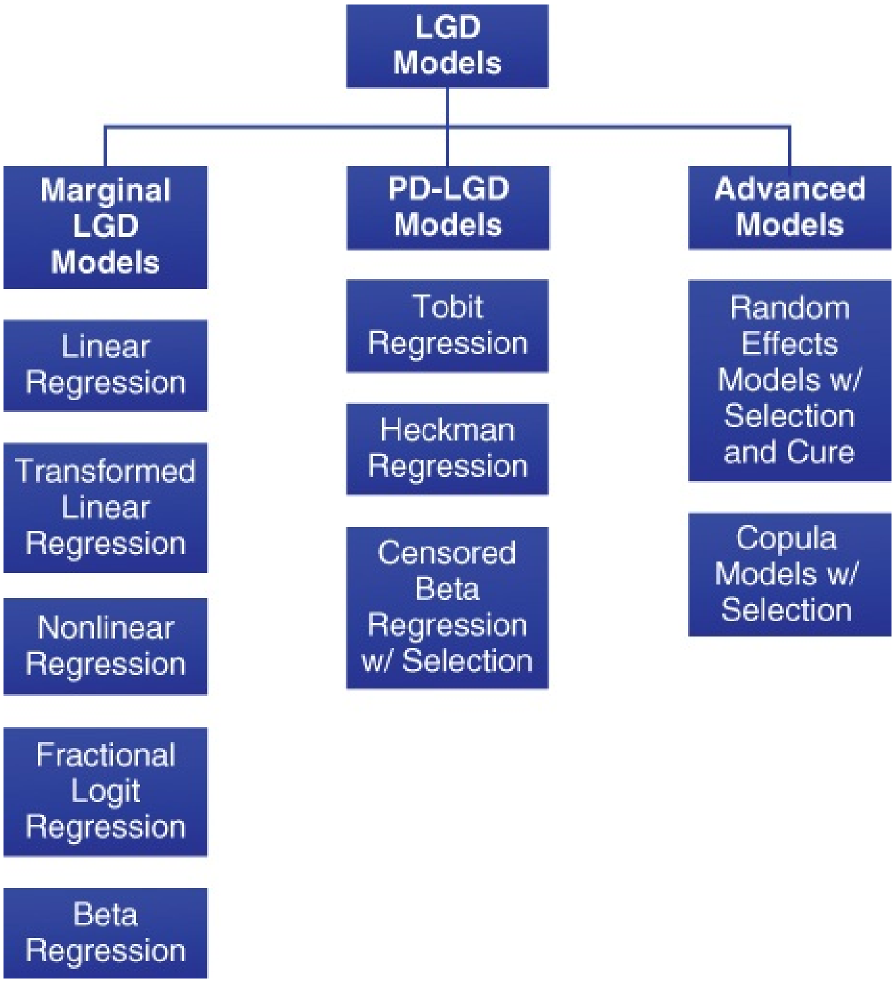

LGD Models
Marginal LGD Models
PD-LGD Models
Advanced Models
Linear Regression
Transformed Linear Regression
Nonlinear Regression
Fractional Logit Regression
Beta Regression
Tobit Regression
Heckman Regression
Censored Beta Regression w/ Selection
Random Effects Models w/ Selection and Cure
Copula Models w/ Selection

Exhibit 10.5 LGD Models in This Chapter

MARGINAL LGD MODELS

Standard models use data on LGDs for defaulted obligors only and do not consider those cases in which no default happened or treat cases with full recoveries in a special way; see Loterman et al. (2012). While this seems to be obvious at first sight and explains why these models are widely used in practice, as we will show later on this might impose inconsistency on the parameter estimates. Sometimes it is also the only way to develop models, as only LGD information is available and not information from nondefaulters. The simplest model uses LGDs as is and puts them into a linear regression. In addition, other standard models explicitly account in various ways for the fact that most LGDs are observed in the interval $[0, 1]$ only.

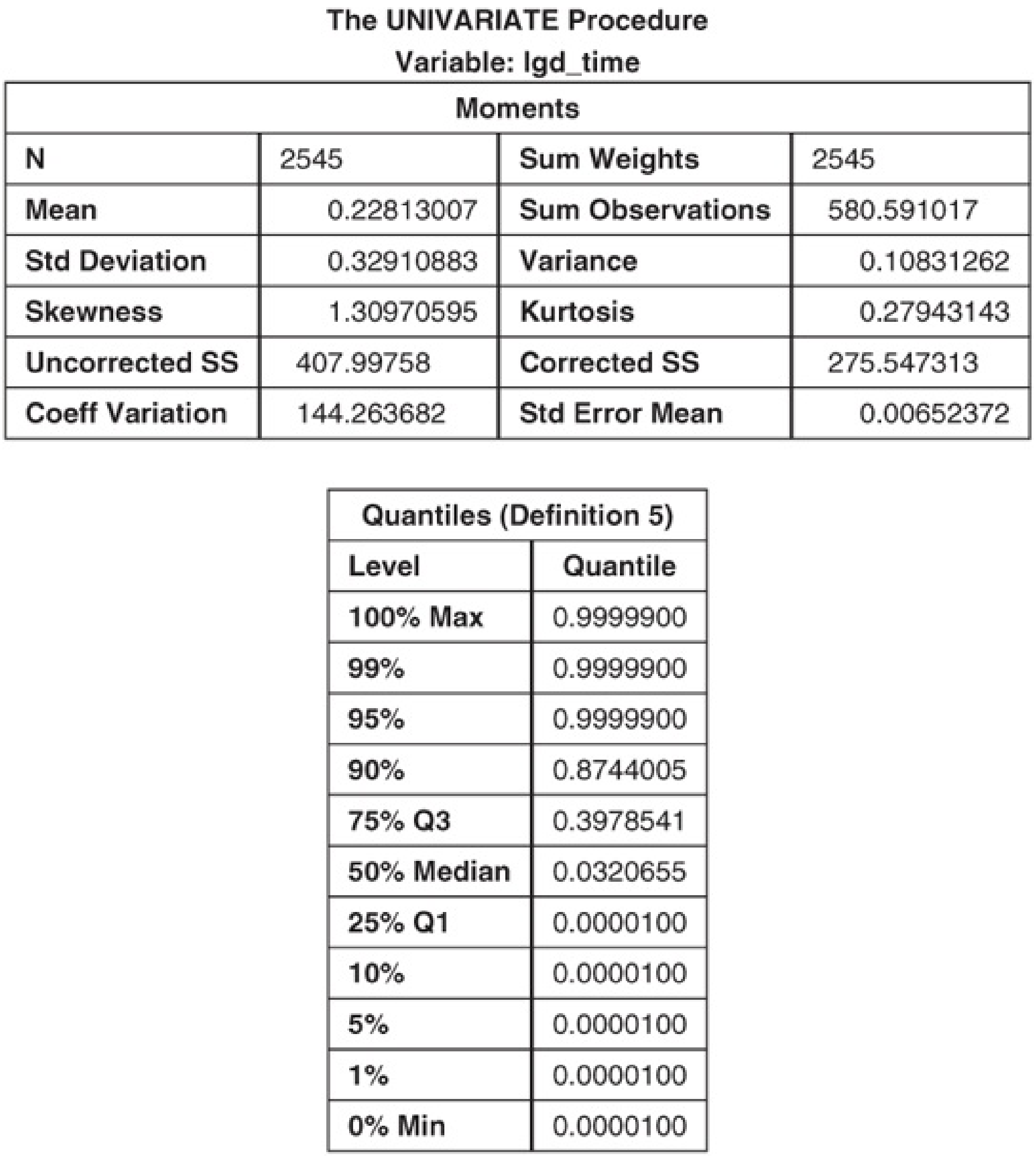
Descriptive Statistics

The data was preprocessed in such a way that only values in the interval $[0, 1]$ occur. Then, in order to show how models using transformed LGDs work, values of zero are set to 0.00001, and values of 1 are set to 0.99999. In practical applications one should carefully check all entries and should justify and report how they are treated. Given an LGD_i for each observation $i, i = 1,\ldots,n$, we compute the three transformations:

$$\begin{aligned} y_i^{logistic} &= \ln \frac{LGD_i}{1 - LGD_i} \\ y_i^{probit} &= \Phi^{-1}(LGD_i) \\ lnrr_i &= \ln(1 - LGD_i) \end{aligned}$$

which will be used as dependent variables in the upcoming regressions.

The database has 2,545 entries for LGD. Besides LGD (labeled lgdtime), it has information about loan-to-value ratios (LTVs) and a dummy variable “purpose1,” which is one if the purpose of the mortgage is to buy a house for investment purposes (rental) and zero otherwise. The following tables and figures (Exhibits 10.6 through 10.15) show the distributions for the LGDs, for their transformed values, and for LTVs. The LGD distribution is highly skewed with obvious clusters at the border values of 0.00001 and 0.99999.



The UNIVARIATE Procedure
Variable: lgd_time

Moments			
N	2545	Sum Weights	2545
Mean	0.22813007	Sum Observations	580.591017
Std Deviation	0.32910883	Variance	0.10831262
Skewness	1.30970595	Kurtosis	0.27943143
Uncorrected SS	407.99758	Corrected SS	275.547313
Coeff Variation	144.263682	Std Error Mean	0.00652372

Quantiles (Definition 5)	
Level	Quantile
100% Max	0.9999900
99%	0.9999900
95%	0.9999900
90%	0.8744005
75% Q3	0.3978541
50% Median	0.0320655
25% Q1	0.0000100
10%	0.0000100
5%	0.0000100
1%	0.0000100
0% Min	0.0000100

Exhibit 10.6 Descriptive Statistics

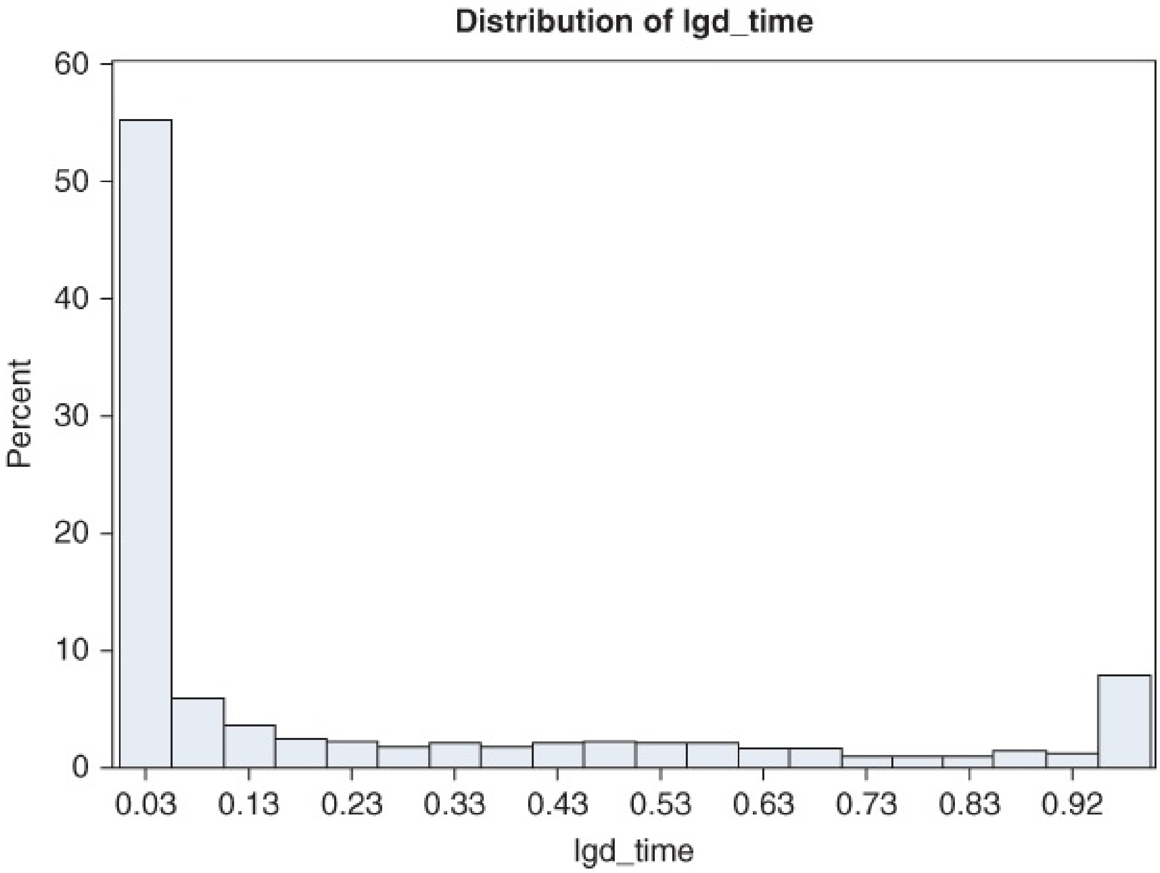
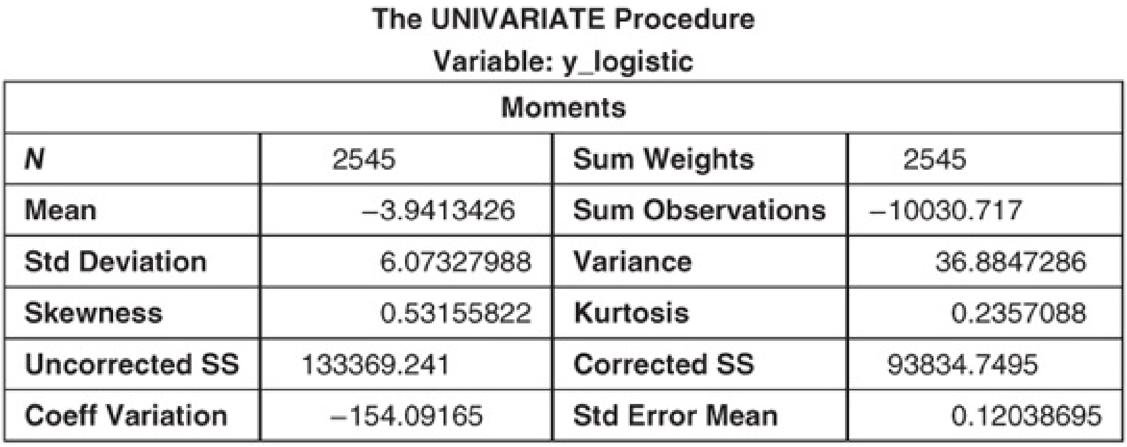


Exhibit 10.7 Descriptive Statistics



The UNIVARIATE Procedure
Variable: y_logistic

Moments			
N	2545	Sum Weights	2545
Mean	–3.9413426	Sum Observations	–10030.717
Std Deviation	6.07327988	Variance	36.8847286
Skewness	0.53155822	Kurtosis	0.2357088
Uncorrected SS	133369.241	Corrected SS	93834.7495
Coeff Variation	–154.09165	Std Error Mean	0.12038695

Exhibit 10.8 Descriptive Statistics

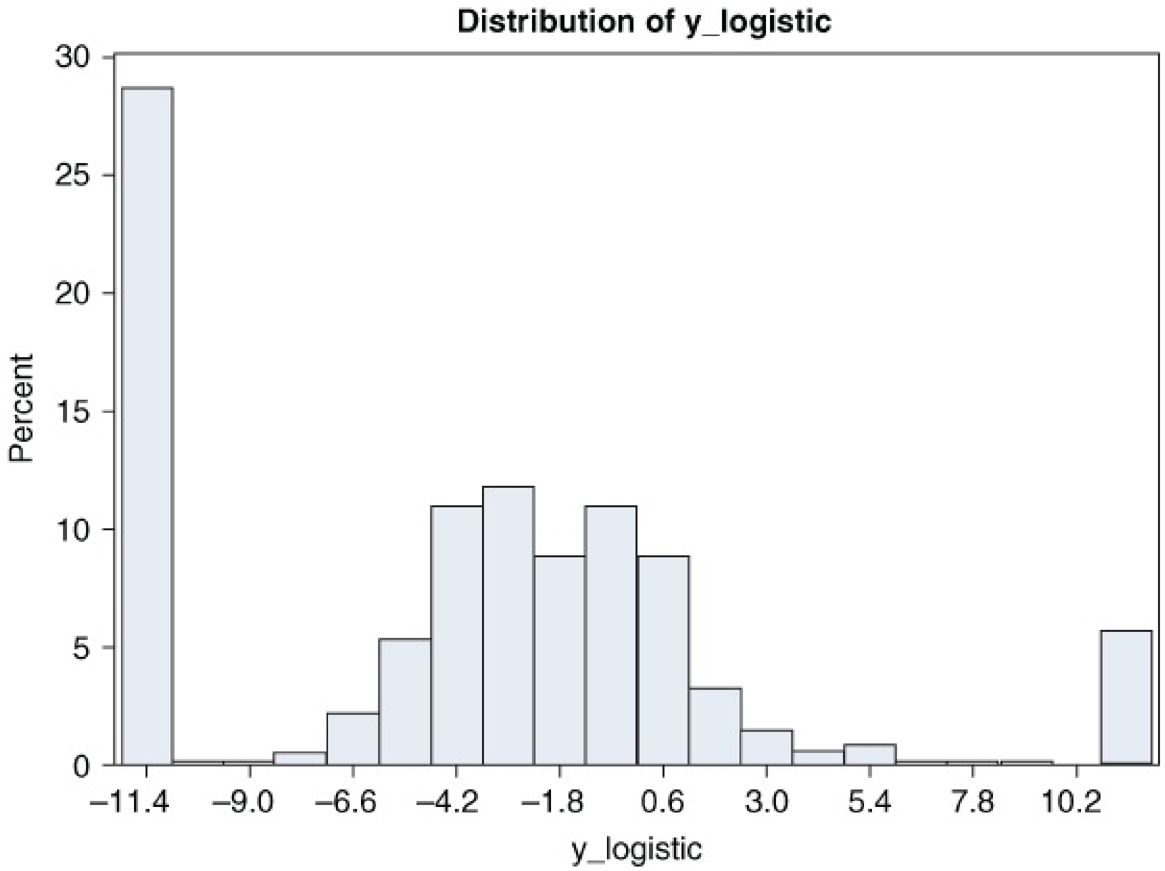
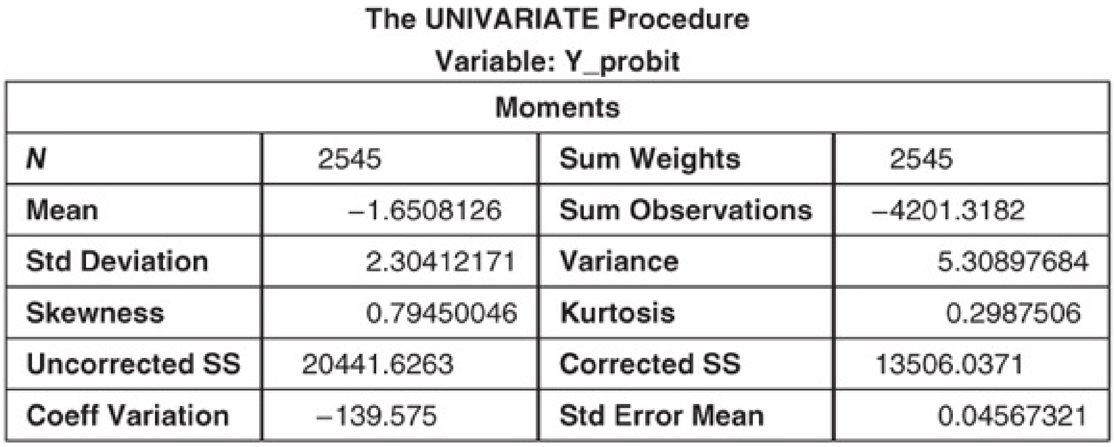


Exhibit 10.9 Descriptive Statistics



The UNIVARIATE Procedure

Variable: Y_probit

Moments			
N	2545	Sum Weights	2545
Mean	−1.6508126	Sum Observations	−4201.3182
Std Deviation	2.30412171	Variance	5.30897684
Skewness	0.79450046	Kurtosis	0.2987506
Uncorrected SS	20441.6263	Corrected SS	13506.0371
Coeff Variation	−139.575	Std Error Mean	0.04567321

Exhibit 10.10 Descriptive Statistics

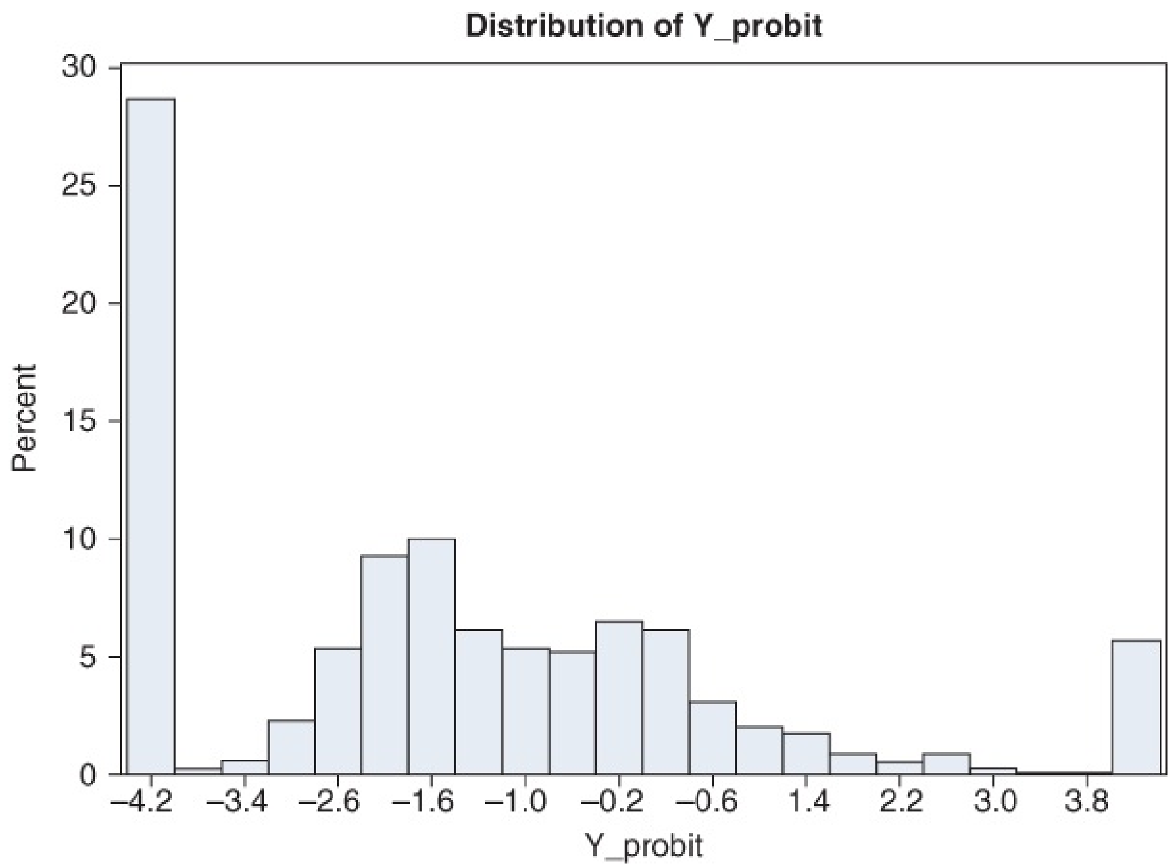
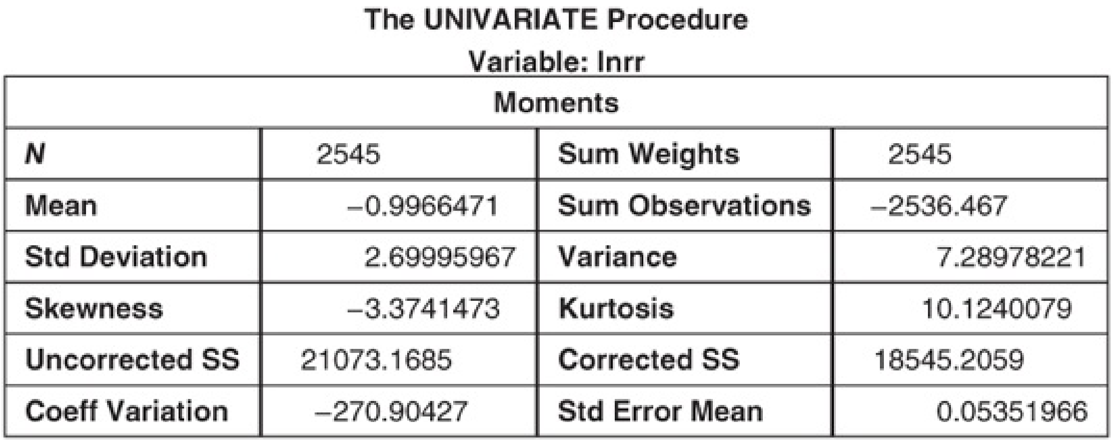


Exhibit 10.11 Descriptive Statistics



The UNIVARIATE Procedure

Variable: lnrr

Moments			
N	2545	Sum Weights	2545
Mean	−0.9966471	Sum Observations	−2536.467
Std Deviation	2.69995967	Variance	7.28978221
Skewness	−3.3741473	Kurtosis	10.1240079
Uncorrected SS	21073.1685	Corrected SS	18545.2059
Coeff Variation	−270.90427	Std Error Mean	0.05351966

Exhibit 10.12 Descriptive Statistics

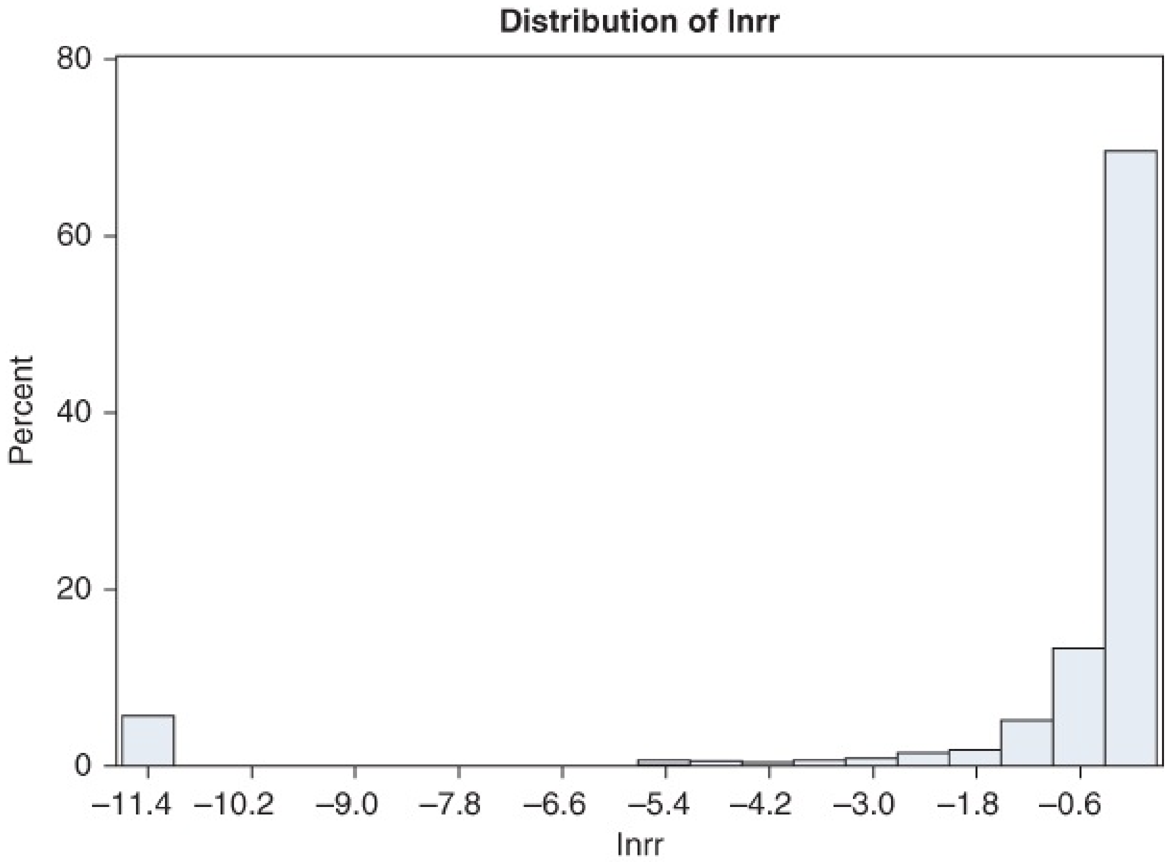
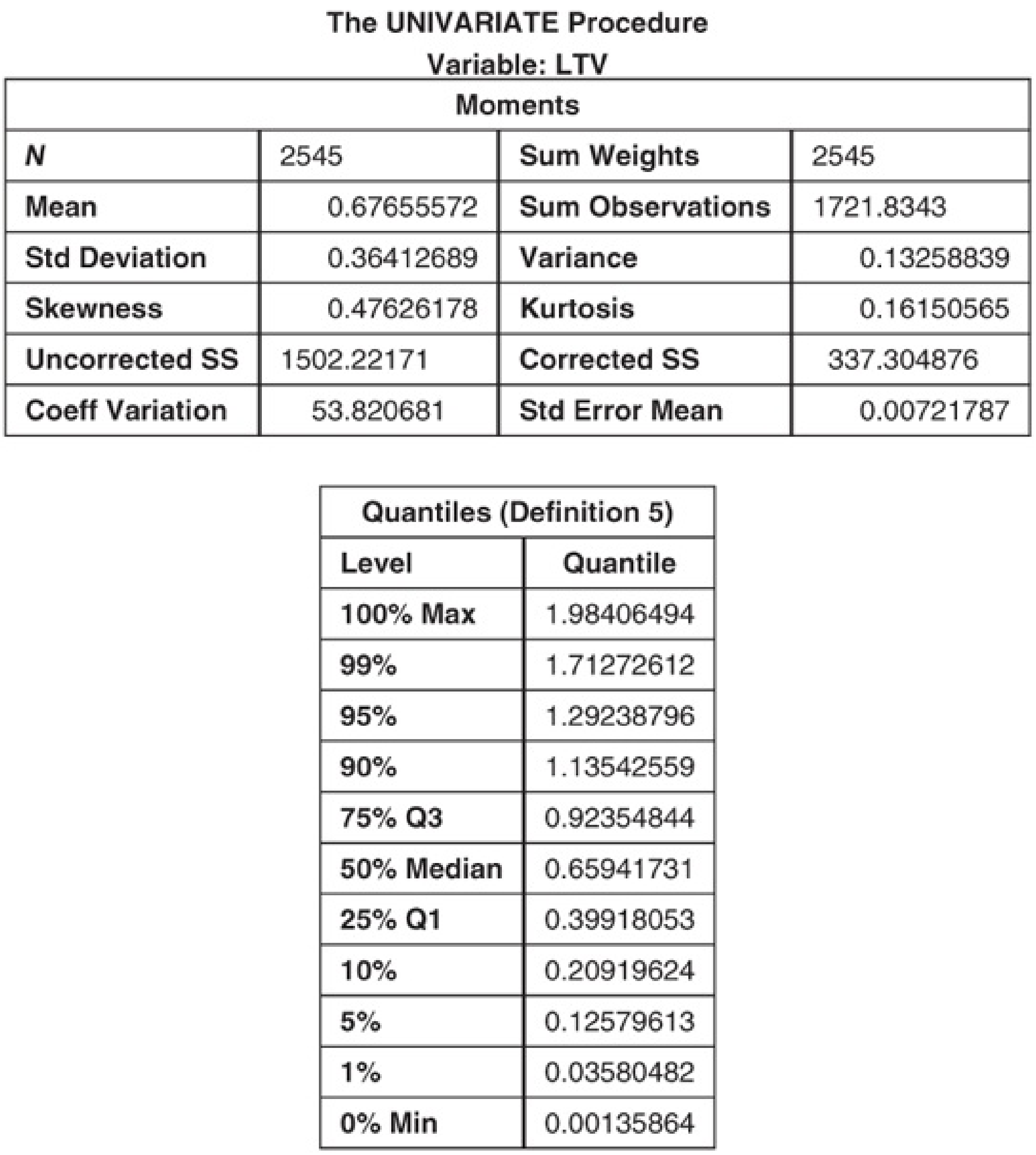


Exhibit 10.13 Descriptive Statistics



The UNIVARIATE Procedure

Variable: LTV

Moments			
N	2545	Sum Weights	2545
Mean	0.67655572	Sum Observations	1721.8343
Std Deviation	0.36412689	Variance	0.13258839
Skewness	0.47626178	Kurtosis	0.16150565
Uncorrected SS	1502.22171	Corrected SS	337.304876
Coeff Variation	53.820681	Std Error Mean	0.00721787

Quantiles (Definition 5)	
Level	Quantile
100% Max	1.98406494
99%	1.71272612
95%	1.29238796
90%	1.13542559
75% Q3	0.92354844
50% Median	0.65941731
25% Q1	0.39918053
10%	0.20919624
5%	0.12579613
1%	0.03580482
0% Min	0.00135864

Exhibit 10.14 Descriptive Statistics

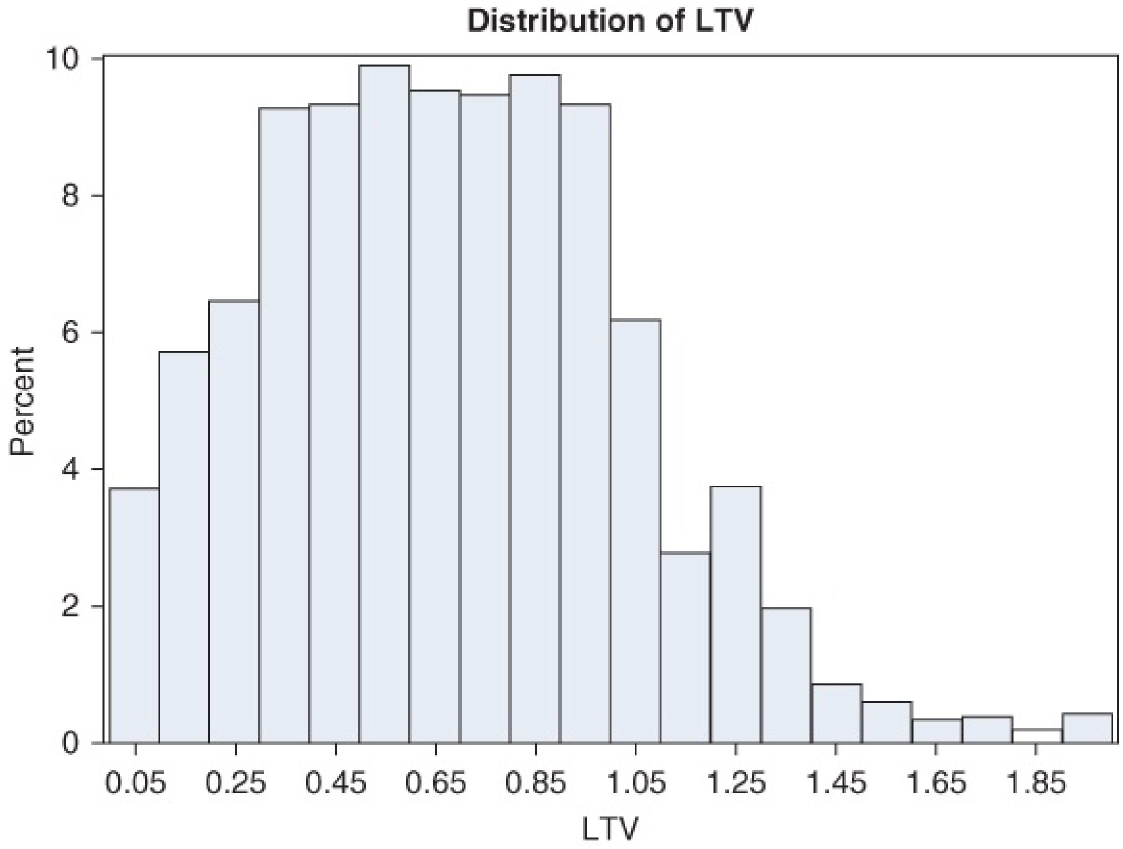


Exhibit 10.15 Descriptive Statistics

We compute descriptive statistics and histograms for the LGDs and the three r=transformations using PROC UNIVARIATE:

```
ODS GRAPHICS ON;
PROC UNIVARIATE DATA = data.lgd;
VAR  lgd_time;
HISTOGRAM;
RUN;
ODS GRAPHICS OFF;
```

Almost 30 percent of the data have an LGD at the border value. The average LGD is around 23 percent if these values are considered. The transformed variables exhibit similar clusters at the extremes:

```
ODS GRAPHICS ON;
PROC UNIVARIATE DATA = data.lgd;
VAR  y_logistic;
HISTOGRAM;
RUN;
ODS GRAPHICS OFF;

 ODS GRAPHICS ON;
 PROC UNIVARIATE DATA = data.lgd;
 VAR  y_probit;
 HISTOGRAM;
 RUN;
 ODS GRAPHICS OFF;
```

```
    ODS GRAPHICS ON;
    PROC UNIVARIATE DATA = data.lgd;
    VAR  lnrr;
    HISTOGRAM;
    RUN;
    ODS GRAPHICS OFF;
```

A key variable in our regression models is the loan-to-value ratio (LTV). We compute descriptive statistics for the LTV ratio and find that LTVs are more evenly distributed between almost zero and about 2 with a mean of 65 percent and only a smaller amount with ratios of more than 100 percent.

```
    ODS GRAPHICS ON;
    PROC UNIVARIATE DATA = data.lgd;
    VAR  LTV;
    HISTOGRAM;
    RUN;
    ODS GRAPHICS OFF;
```

Linear Regression

The first standard model is the linear ordinary least squares (OLS) regression model, which is of the form

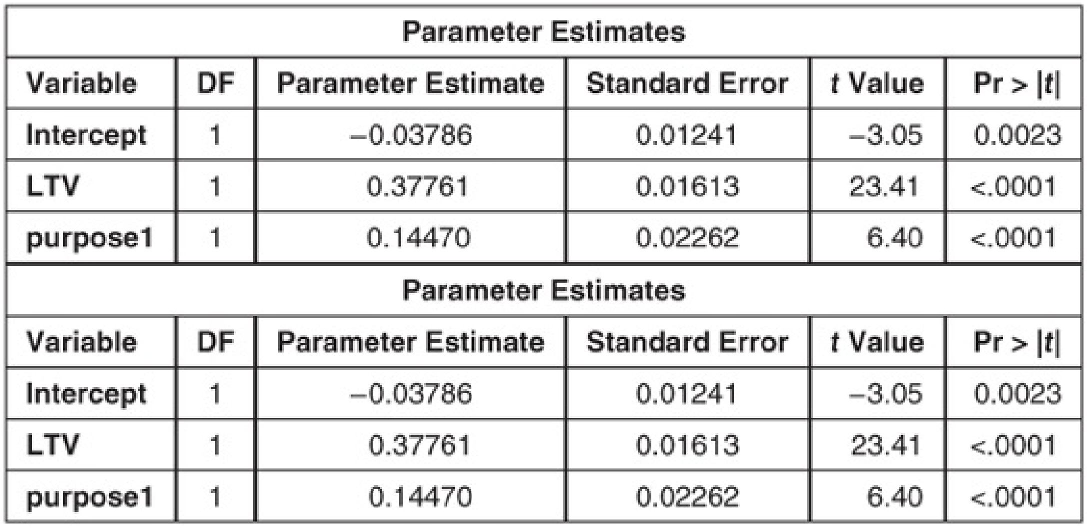
$$LGD_i = \beta' x_i + \epsilon_i \qquad i = 1, \ldots, N \tag{10.1}$$

where $\epsilon_i \sim N(0, \sigma^2)$, x_i is a vector of explanatory variables and β is a vector of unknown parameters to be estimated. We dropped the subscript "t" here, as there is no information about time in the data set. Thus, observations are treated on a cross-sectional basis. If time stamps are available, one can easily extend our models to time-varying covariates, such as macroeconomic variables.

We use PROC REG to estimate this linear regression model based on LTV and the dummy variable for renting purpose. As explained before, we set LGD values of zero to 0.00001 and values of one to 0.99999. Thus, the dependent variable has only values in $[0, 1]$. This allows transformations in the models that we show later. However, values ≤ 0 or ≥ 1 are sometimes also included in practical applications of the linear model.

```
ODS GRAPHICS ON;
PROC REG DATA=data.lgd
PLOTS(STATS= ALL)= DIAGNOSTICS;
MODEL lgd_time = LTV purpose1;
RUN;
ODS GRAPHICS OFF;
```

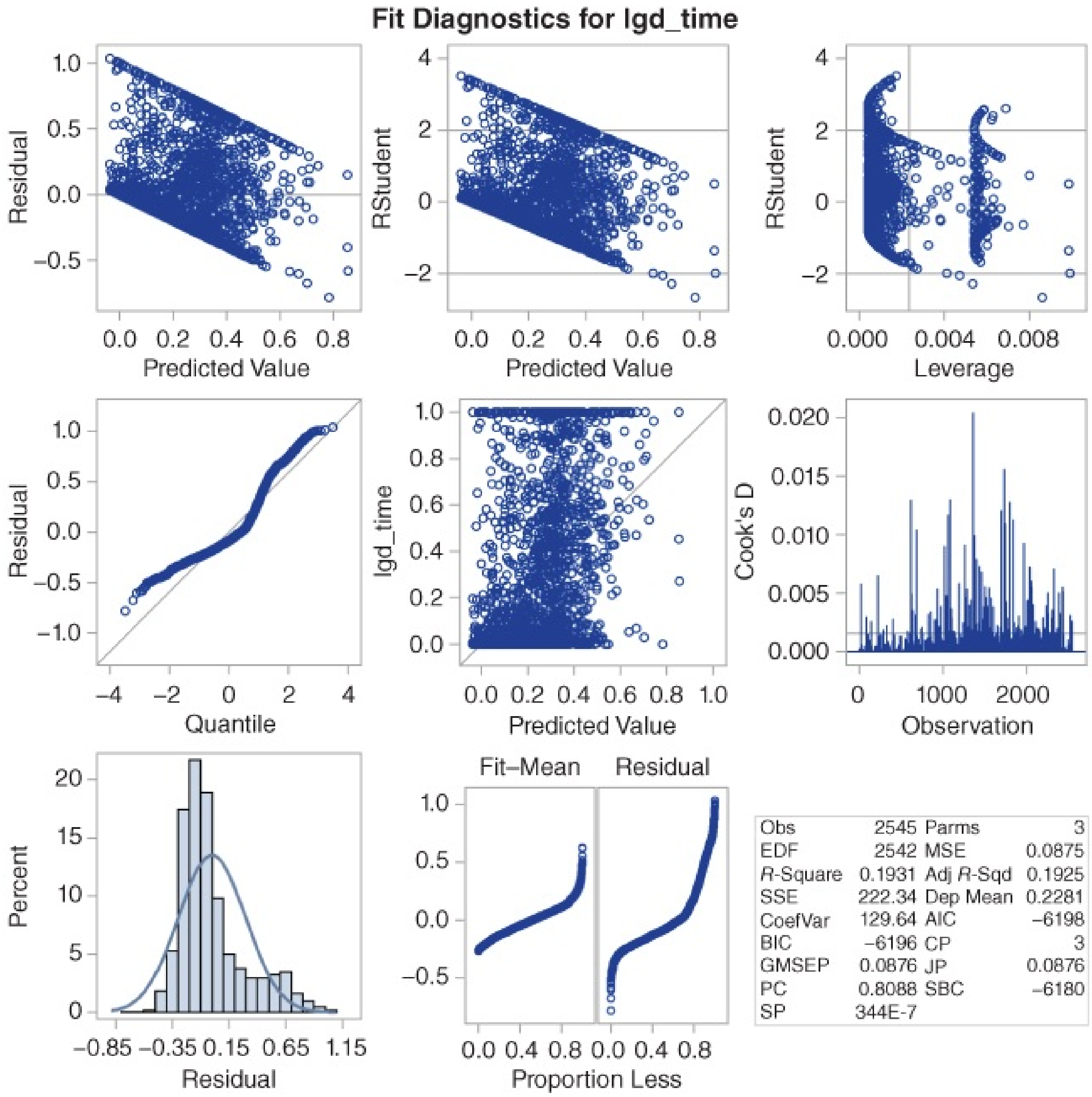
The R^2 and the fit diagnostics that are default output in PROC REG show the model fit. The explained variation (adjusted R^2) is about 19 percent for the model which uses two explanatory variables only (see Exhibits 10.16 and 10.17).



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |*t*|
Intercept	1	−0.03786	0.01241	−3.05	0.0023
LTV	1	0.37761	0.01613	23.41	<.0001
purpose1	1	0.14470	0.02262	6.40	<.0001

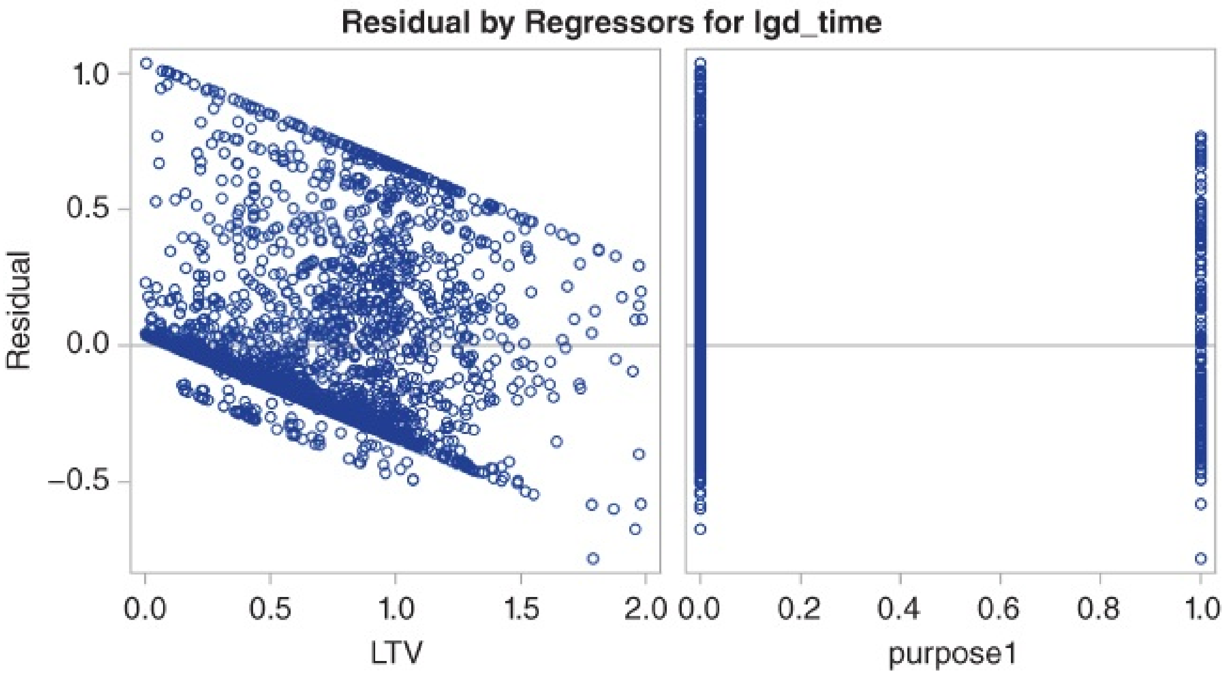
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |*t*|
Intercept	1	−0.03786	0.01241	−3.05	0.0023
LTV	1	0.37761	0.01613	23.41	<.0001
purpose1	1	0.14470	0.02262	6.40	<.0001

Exhibit 10.16 Linear Regression



Fit Diagnostics for lgd_time

Obs	2545	Parms	3
EDF	2542	MSE	0.0875
R-Square	0.1931	Adj *R*-Sqd	0.1925
SSE	222.34	Dep Mean	0.2281
CoefVar	129.64	AIC	−6198
BIC	−6196	CP	3
GMSEP	0.0876	JP	0.0876
PC	0.8088	SBC	−6180
SP	344E-7		



Residual by Regressors for lgd_time

Exhibit 10.17 Linear Regression

The SAS output shows that the coefficients for LTV and for renting are positive and statistically different from zero. In other words, higher LTVs go with higher LGDs, and lower LTVs with lower LGDs, which is in line with economic intuition. An increase of LTV by 1 percentage point increases the LGD by about 0.38 percentage points. Buying a house for renting purposes increases LGDs by about 14 percent.

The plot of realized LGDs versus predicted LGDs shows a rather weak relationship. The model seems to fit the mean well but the idiosyncratic variation around the LGD, which is not explained by the model, is still somewhat huge, which can also be seen by the quantile plot. Moreover, one sees clusters for the borderline values at 0.00001 and 0.99999 and the respective residuals. While $R^2 = 0.19$ is not too bad for two explanatory variables only, the unexplained variation of 81 percent seems to be a widespread issue in practical applications, and a major challenge is to find the proper covariates that are able to explain a higher proportion of the variation. If you have a higher number of potential explanatory covariates available, you can make use of variable selection algorithms that are implemented in SAS PROC REG, such as a forward, backward, or stepwise selection, as discussed in the credit scoring chapter.

Transformed Linear Regression

A shortcoming of the linear regression model is the assumption of normally distributed residuals, because LGDs are usually between zero and one and are thus obviously not normal by definition. This shortcoming may be addressed by using a transformation of the LGD that computes values falling into the interval $(-\infty, +\infty)$, and applying a linear regression on the transformed values. Two frequently used transformations are the logistic $\ln \frac{LGD_i}{1-LGD_i}$, and the inverse normal (probit) transformation $\Phi^{-1}(LGD_i)$. The models look as follows:

$$\ln \frac{LGD_i}{1 - LGD_i} = \beta' x_i + \epsilon_i \qquad \textbf{10.2}$$

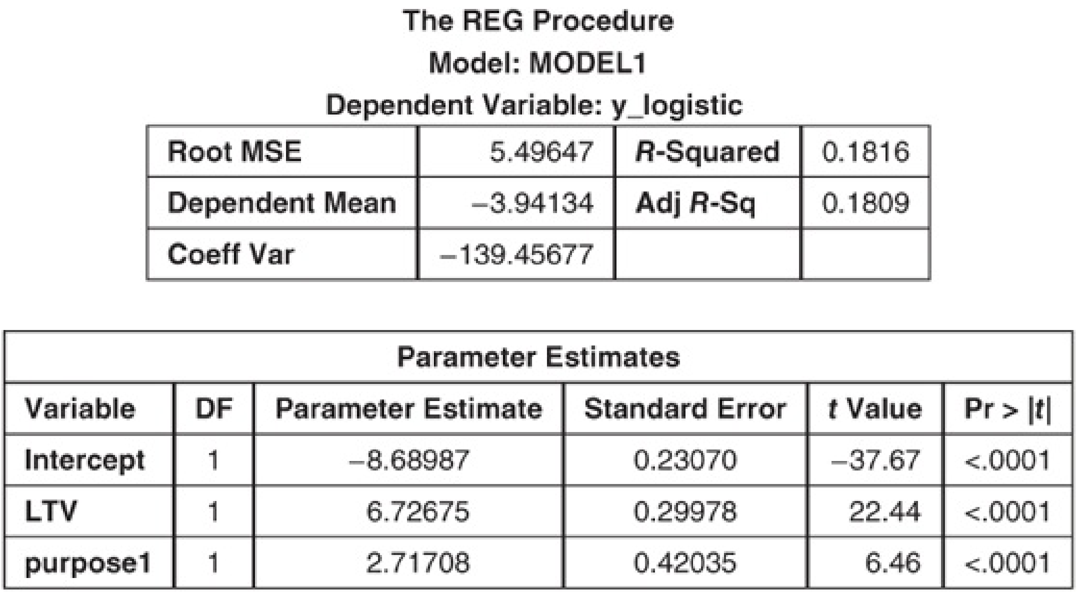
where $\epsilon_i \sim N(0, \sigma^2)$, and

$$\Phi^{-1}(LGD_i) = \beta' x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$.

The regression codes using PROC REG and outputs (Exhibits 10.18 and 10.19) for the logistic-linear regression are:

```
ODS GRAPHICS ON;
PROC REG DATA=data.lgd
PLOTS( STATS= ALL)= DIAGNOSTICS;
MODEL y_logistic = LTV purpose1;
RUN;
ODS GRAPHICS OFF;
```



The REG Procedure
Model: MODEL1
Dependent Variable: y_logistic

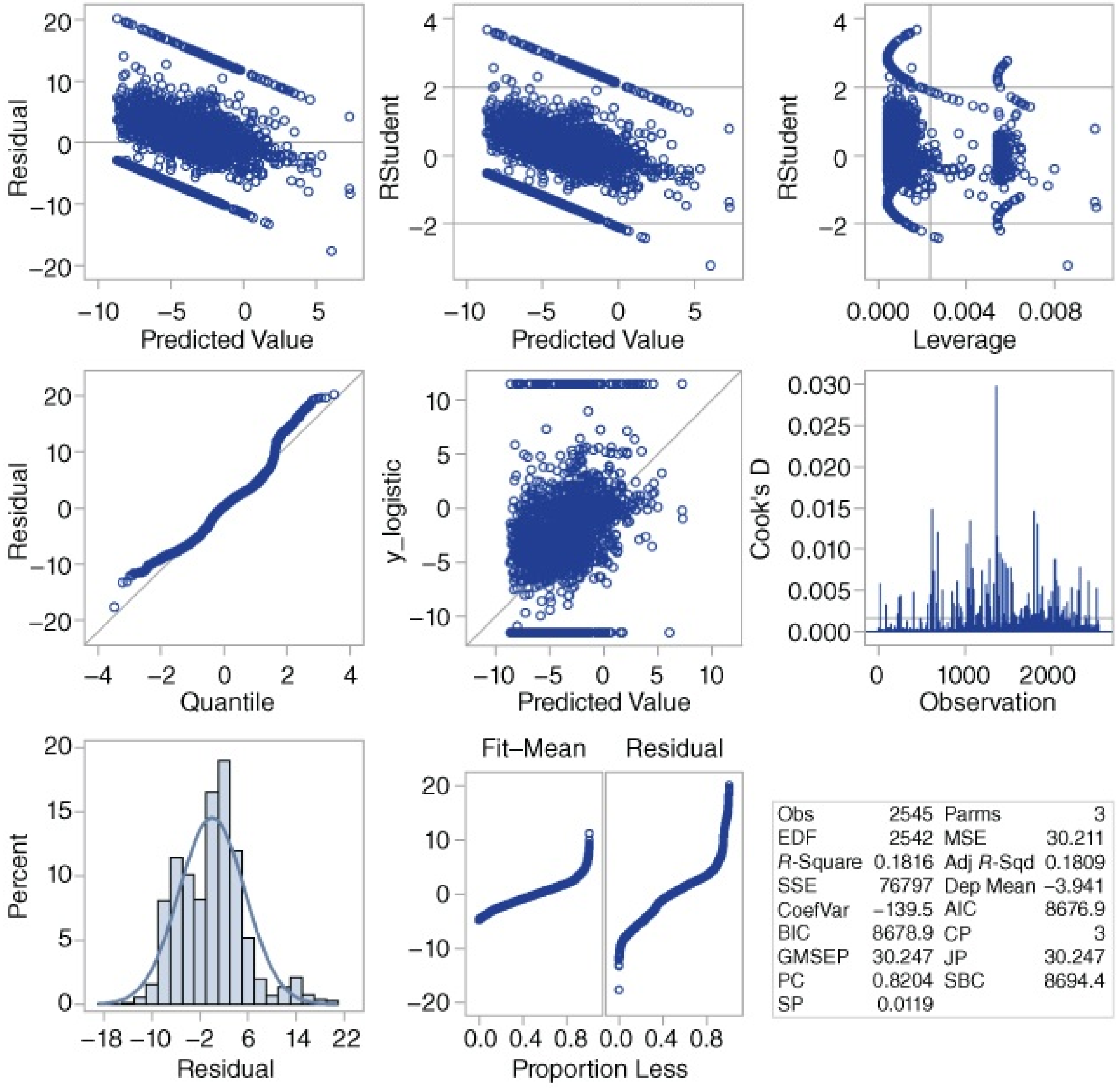
Root MSE	5.49647	*R*-Squared	0.1816
Dependent Mean	−3.94134	Adj *R*-Sq	0.1809
Coeff Var	−139.45677		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |*t*|
Intercept	1	−8.68987	0.23070	−37.67	<.0001
LTV	1	6.72675	0.29978	22.44	<.0001
purpose1	1	2.71708	0.42035	6.46	<.0001

Exhibit 10.18 Logistic-Linear Regression

Fit Diagnostics for y_logistic



Obs	2545	Parms	3
EDF	2542	MSE	30.211
R-Square	0.1816	Adj *R*-Sqd	0.1809
SSE	76797	Dep Mean	−3.941
CoefVar	−139.5	AIC	8676.9
BIC	8678.9	CP	3
GMSEP	30.247	JP	30.247
PC	0.8204	SBC	8694.4
SP	0.0119		

Residual by Regressors for y_logistic

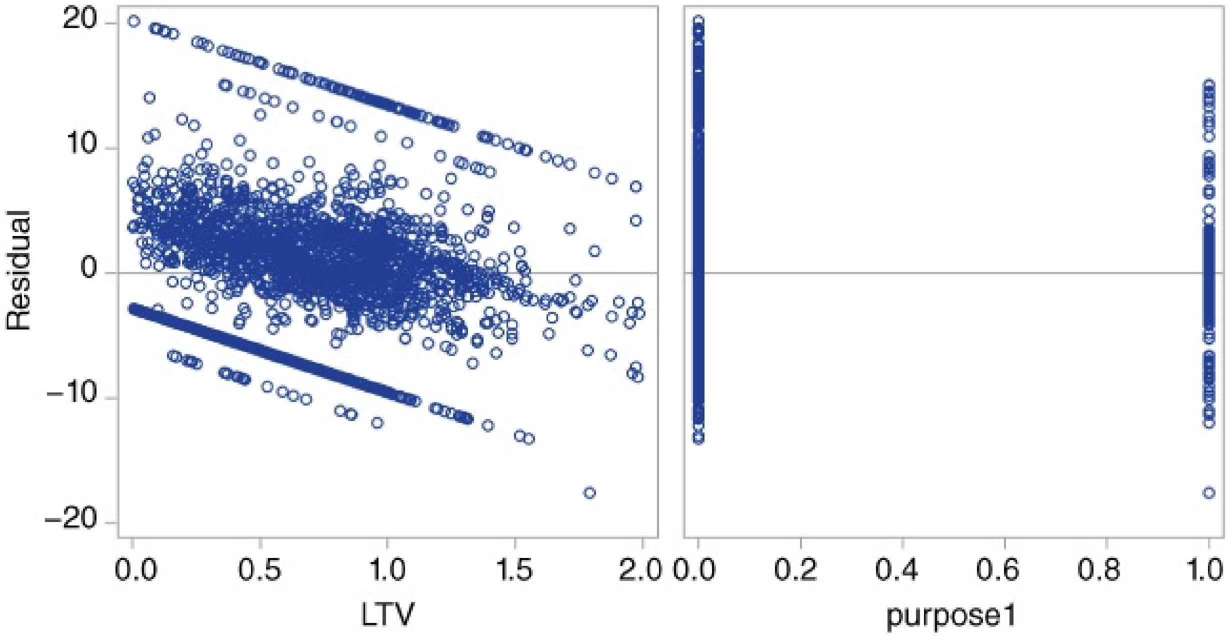
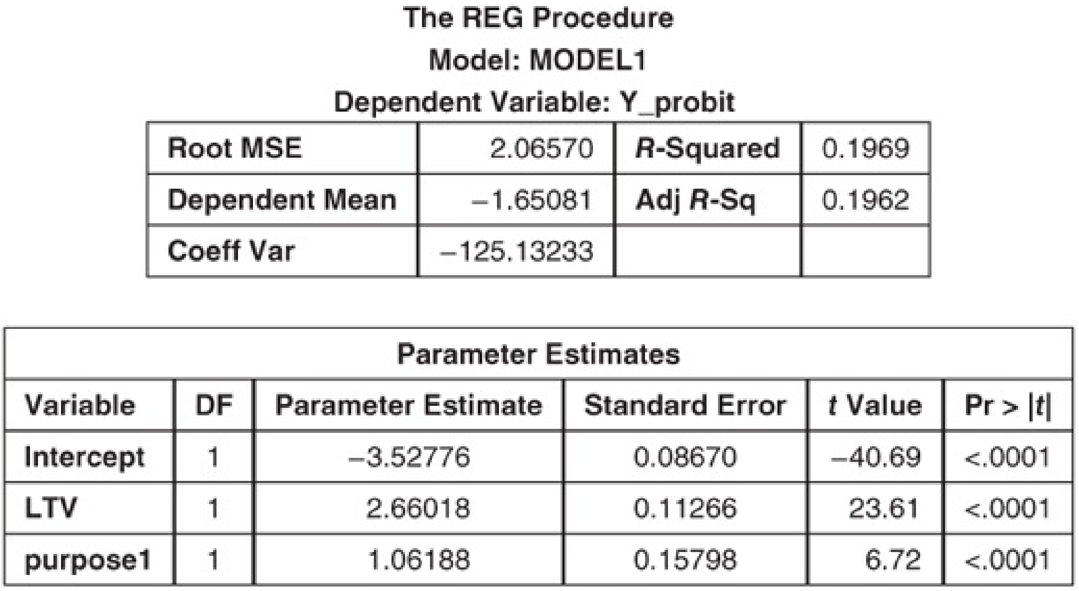


Exhibit 10.19 Logistic-Linear Regression

Both model results are basically similar to the linear regression in terms of economic and statistical significance of the parameter estimates and the model fit, although particularly the probit transformation seems to capture the distribution characteristics of the residuals somewhat better. The reason is that the linear model assumes a normal distribution whereas the other models assume transformed normal distributions, which can sometimes be more appropriate for modeling the tails of a distribution. Both values for R^2 are basically similar to those of the linear regression model.

The regression codes using PROC REG and outputs (Exhibits 10.20 and 10.21) for the probit-linear regression are:



The REG Procedure
Model: MODEL1
Dependent Variable: Y_probit

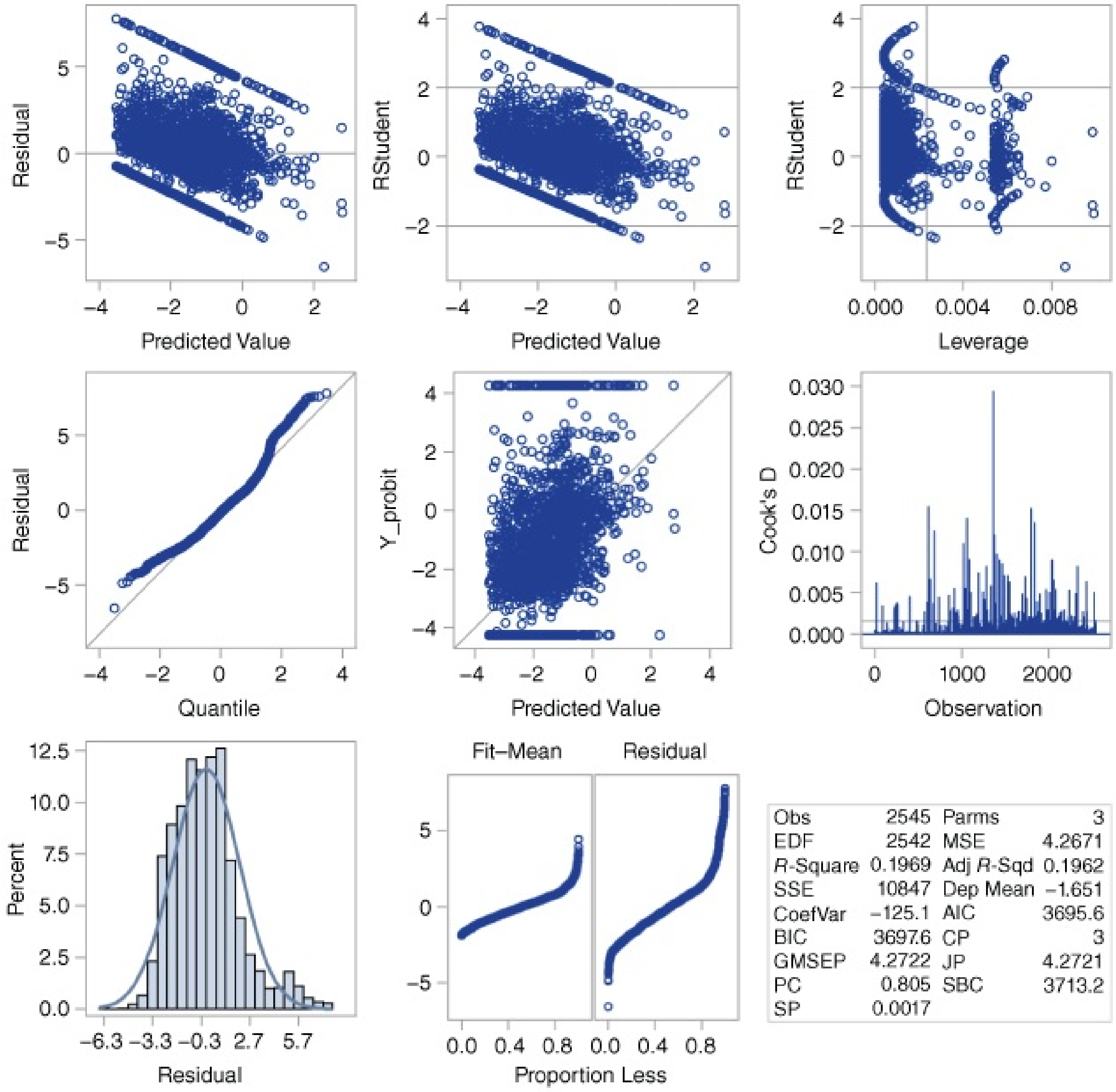
Root MSE	2.06570	*R*-Squared	0.1969
Dependent Mean	−1.65081	Adj *R*-Sq	0.1962
Coeff Var	−125.13233		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |*t*|
Intercept	1	−3.52776	0.08670	−40.69	<.0001
LTV	1	2.66018	0.11266	23.61	<.0001
purpose1	1	1.06188	0.15798	6.72	<.0001

Exhibit 10.20 Probit-Linear Regression

Fit Diagnostics for Y_probit



Obs	2545	Parms	3
EDF	2542	MSE	4.2671
R-Square	0.1969	Adj *R*-Sqd	0.1962
SSE	10847	Dep Mean	−1.651
CoefVar	−125.1	AIC	3695.6
BIC	3697.6	CP	3
GMSEP	4.2722	JP	4.2721
PC	0.805	SBC	3713.2
SP	0.0017		

Residual by Regressors for Y_probit

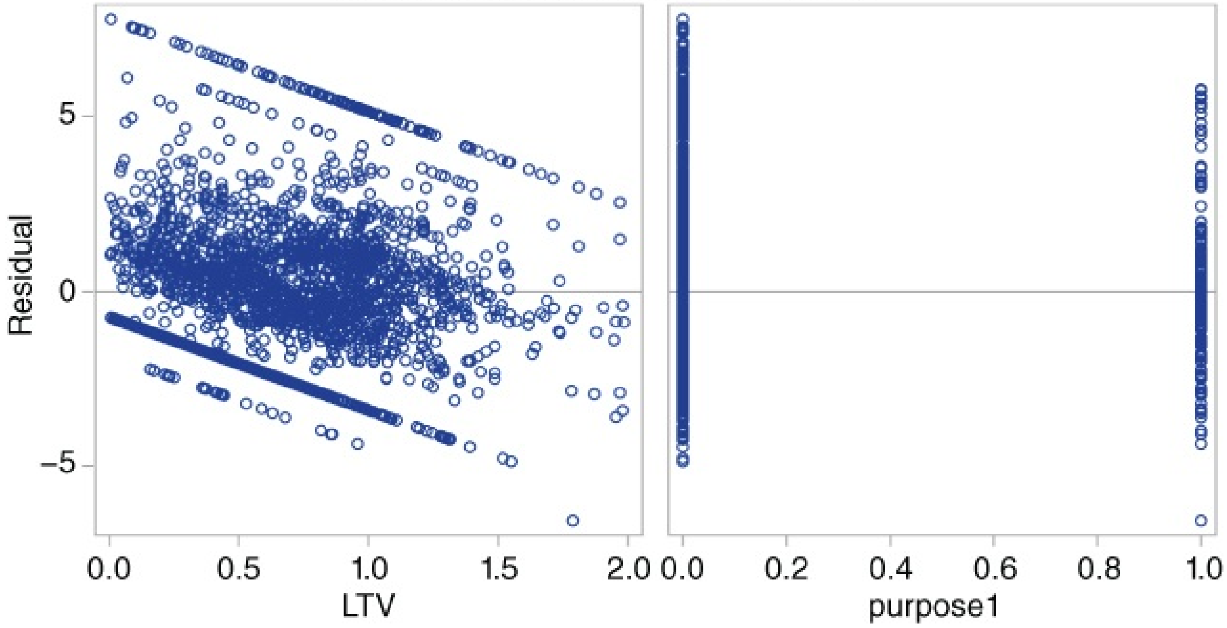


Exhibit 10.21 Probit-Linear Regression

```
ODS GRAPHICS ON;
PROC REG DATA=data.lgd
PLOTS(STATS= ALL)= DIAGNOSTICS;
MODEL y_probit = LTV purpose1;
RUN;
ODS GRAPHICS OFF;
```

Nonlinear Regression

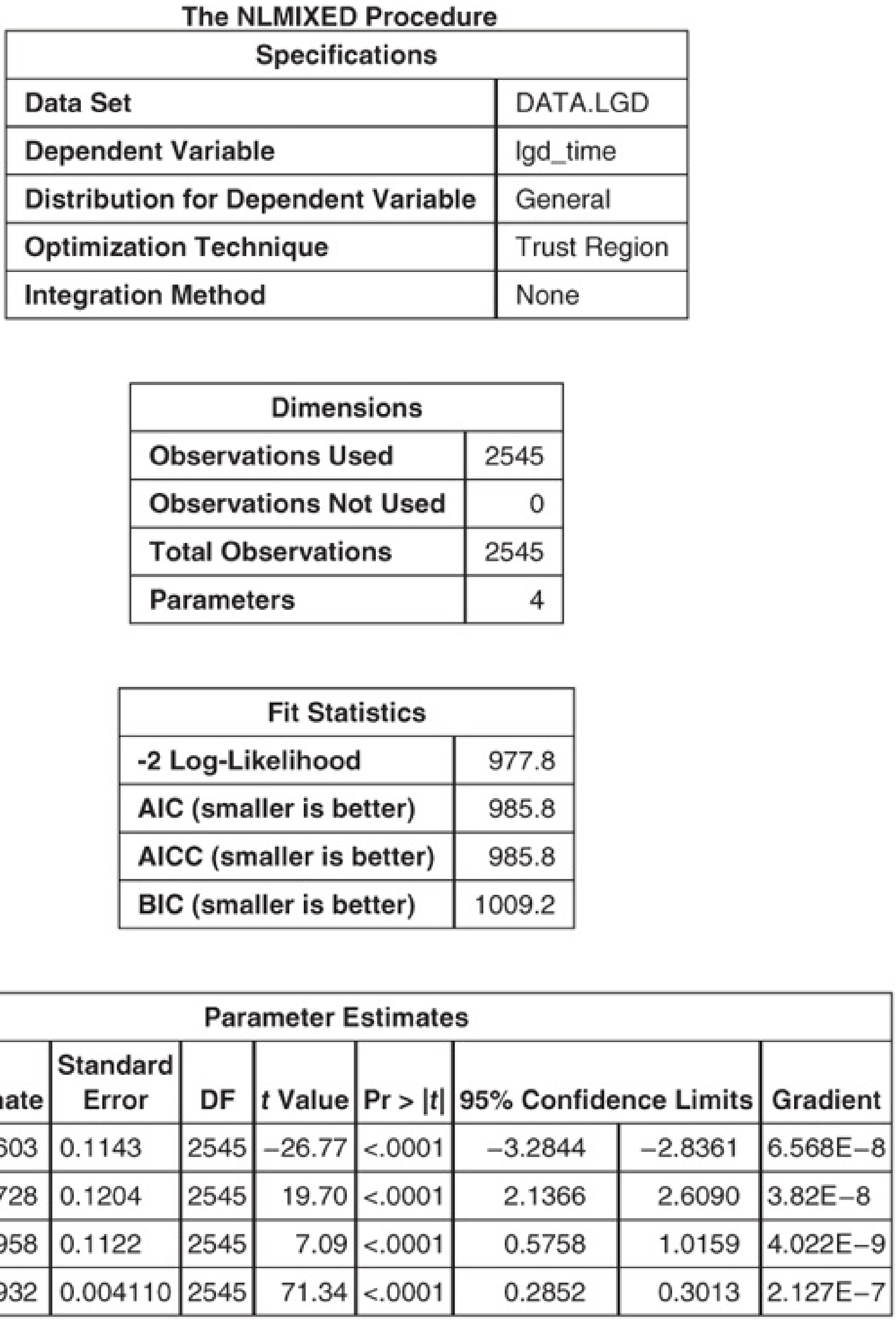
An alternative approach is to apply a nonlinear regression, which transforms the linear predictor of the explanatory variables rather than the left-hand side of the regression equation. An example model equation is:

$$LGD_i = \frac{1}{1 + \exp\ (-\boldsymbol{\beta}'\boldsymbol{x}_i)} + \epsilon_i \qquad \mathbf{10.3}$$

where a logit transformation for the predictor is used and $\epsilon_i \sim N(0, \sigma^2)$. Alternatively, another transformation, such as probit, could be used.

The model can easily be estimated by maximum likelihood using PROC NLMIXED, as shown in the following code. The parameter estimates of the model exhibit similar economical and statistical significance as in the former models, although they now enter the model in a nonlinear way. The parameter σ denotes the (remaining) volatility of the residuals after controlling for the covariates.

PROC NLMIXED has a different structure to PROC REG as it supports programming statements (between the PARMS and the MODEL statements in the following code) that allow for a more complex specification of the likelihood via the MODEL statement. The linear predictor is coded by the statement “xb = b0 + b1 * LTV + b2 * purpose1.” PROC NLMIXED estimates the parameters via the maximization of the likelihood. Furthermore, starting values for the parameters b0, b1, and b2 can be specified using the PARMS statement. (See Exhibit 10.22.) In the PROC NLMIXED line, one can include an option 'TECH=' which chooses the optimization algorithm. SAS offers several techniques here, and we choose trust region optimization as an example. Details can be found in the SAS manual; see SAS Institute Inc. (2015). The interested reader is encouraged to run the programs in this section using other techniques in order to check if and/or how they may affect the results.



The NLMIXED Procedure

Specifications	
Data Set	DATA.LGD
Dependent Variable	lgd_time
Distribution for Dependent Variable	General
Optimization Technique	Trust Region
Integration Method	None

Dimensions	
Observations Used	2545
Observations Not Used	0
Total Observations	2545
Parameters	4

Fit Statistics	
-2 Log-Likelihood	977.8
AIC (smaller is better)	985.8
AICC (smaller is better)	985.8
BIC (smaller is better)	1009.2

Parameter Estimates

Parameter	Estimate	Standard Error	DF	*t* Value	Pr > |*t*|	95% Confidence Limits		Gradient
b0	−3.0603	0.1143	2545	−26.77	<.0001	−3.2844	−2.8361	6.568E−8
b1	2.3728	0.1204	2545	19.70	<.0001	2.1366	2.6090	3.82E−8
b2	0.7958	0.1122	2545	7.09	<.0001	0.5758	1.0159	4.022E−9
Sigma	0.2932	0.004110	2545	71.34	<.0001	0.2852	0.3013	2.127E−7

Exhibit 10.22 Nonlinear Regression

```
ODS GRAPHICS ON;
PROC NLMIXED DATA = data.lgd TECH = TRUREG;
PARMS b0 = 0 b1 = 0 b2 = 0 sigma=1;
xb = b0 + b1 * LTV + b2 * purpose1 ;
mu = 1 / (1 + EXP(- xb));
lh = PDF('NORMAL', lgd_time, mu, sigma);
ll = LOG(lh);
RUN;
ODS GRAPHICS OFF;
```

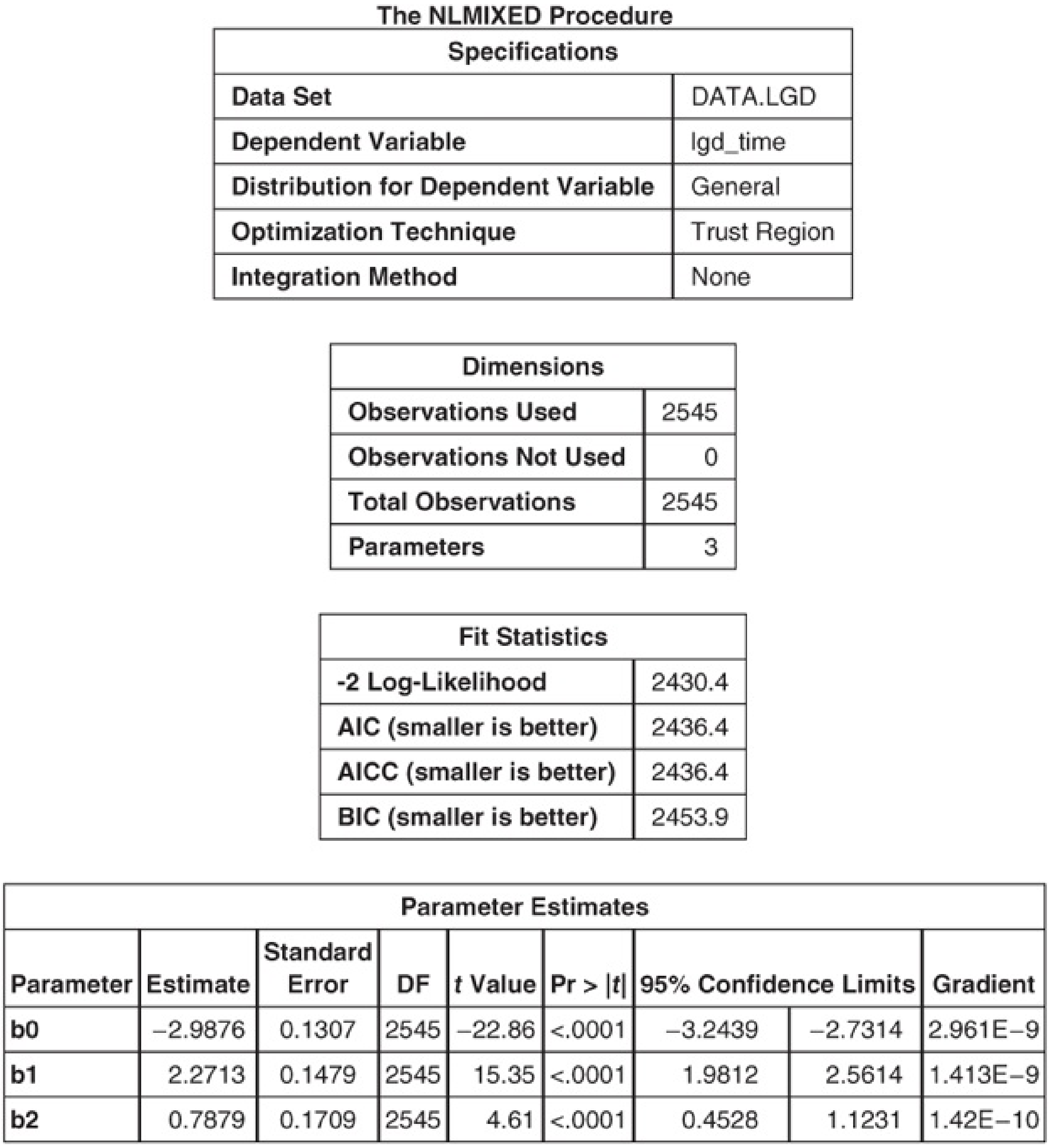
Fractional Logit Regression

The former models required specific assumptions for the residuals. A distribution-free alternative that uses quasi maximum likelihood is the fractional logit model. It requires only that the conditional mean $E(LGD_i|x_i)$ is correctly specified.

Given an assumption for the conditional mean, for example $E(LGD_i|x_i) = \frac{1}{1+\exp(-\boldsymbol{\beta}'\boldsymbol{x}_i)}$, the Bernoulli likelihood becomes:

$$L = \left(\frac{1}{1+\exp\,(-\boldsymbol{\beta}'\boldsymbol{x}_i)}\right)^{LGD_i}\left(1-\frac{1}{1+\exp\,(-\boldsymbol{\beta}'\boldsymbol{x}_i)}\right)^{1-LGD_i}$$

Alternatively, another transformation, such as probit, could be used. This is similar to binary logistic regression, as we discussed earlier. The code and the output (Exhibit 10.23) are shown next. Because the model does not assume a distribution for the residuals, only the parameters for the regression equation need to be estimated. The outcomes are similar to those of the nonlinear regression model.



The NLMIXED Procedure

Specifications	
Data Set	DATA.LGD
Dependent Variable	lgd_time
Distribution for Dependent Variable	General
Optimization Technique	Trust Region
Integration Method	None

Dimensions	
Observations Used	2545
Observations Not Used	0
Total Observations	2545
Parameters	3

Fit Statistics	
-2 Log-Likelihood	2430.4
AIC (smaller is better)	2436.4
AICC (smaller is better)	2436.4
BIC (smaller is better)	2453.9

Parameter Estimates

Parameter	Estimate	Standard Error	DF	*t* Value	Pr > |*t*|	95% Confidence Limits		Gradient
b0	−2.9876	0.1307	2545	−22.86	<.0001	−3.2439	−2.7314	2.961E−9
b1	2.2713	0.1479	2545	15.35	<.0001	1.9812	2.5614	1.413E−9
b2	0.7879	0.1709	2545	4.61	<.0001	0.4528	1.1231	1.42E−10

Exhibit 10.23 Fractional Logit Regression

```
ODS GRAPHICS ON;
PROC NLMIXED DATA = data.lgd TECH = TRUREG;
PARMS b0 = 0 b1 = 0 b2 = 0 ;
xb = b0 + b1 * LTV + b2 *purpose1 ;
mu = 1 / (1 + exp(- xb));
lh = (mu ** lgd_time) * ((1 - mu) ** (1 - lgd_time));
ll = LOG(lh);
MODEL lgd_time ~ GENERAL(ll);
RUN;
ODS GRAPHICS OFF;
```

Beta Regression

The final standard model is the beta regression model. It is related to the beta distribution that is frequently used for modeling proportions and can deal with a variety of shapes. The beta distribution for the LGD has two parameters α and β and has the form:

$$f(lgd) = \frac{1}{B(\alpha, \beta)} lgd^{\alpha-1}(1 - lgd)^{\beta-1}$$

where $B(\alpha, \beta)$ is the beta function

$$B(\alpha, \beta) = \int_0^1 lgd^{\alpha-1}(1 - lgd)^{\beta-1}\ dlgd \quad \alpha, \beta > 0$$

The beta function is related to the gamma function by

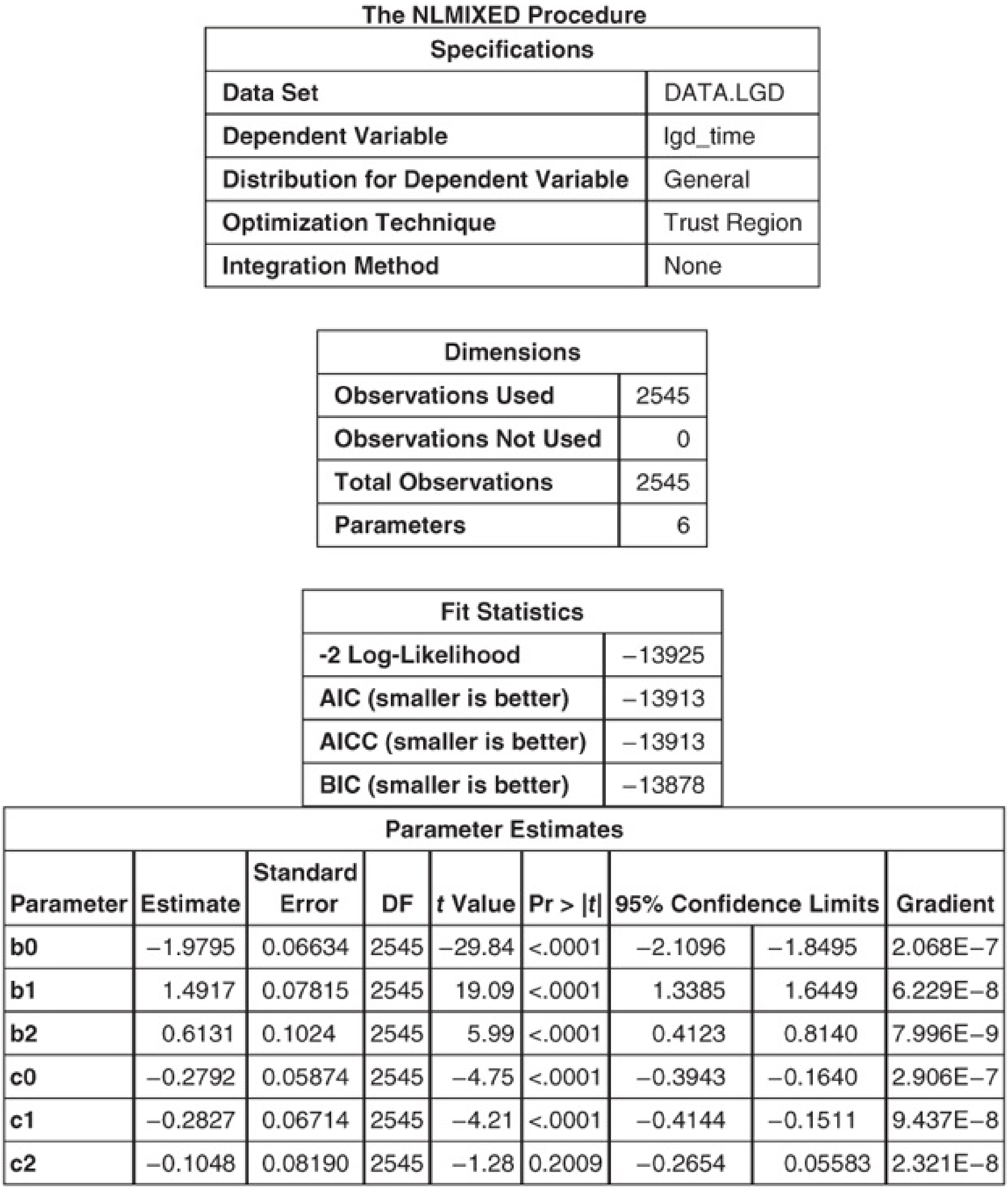
$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

We apply this relationship in the programming statements of PROC NLMIXED.

For the beta regression, both parameters are transformed into a location (mean) parameter μ and a shape parameter δ such that $\alpha = \mu * \delta$ and $\beta = (1 - \mu)\delta$, and the variance is obtained as $\sigma^2 = \frac{\mu(1-\mu)}{1+\delta}$.

A regression model is then applied to the location and the shape parameter where usually the mean is transformed to stay in the interval $(0, 1)$ as in the former models, that is $\mu = \frac{1}{1+\exp(-\beta' x_i)}$, and the location parameter is transformed by the log function to ensure that it is strictly positive; that is, $\delta = \exp\{\beta_\delta' x_i\}$ where β_δ is a vector of parameters.

Generally, the parameters can be estimated using the method of moments (MM) or maximum likelihood (ML). The likelihood is given via the previous density. The following code shows the ML estimation using PROC NLMIXED and the resulting output (Exhibit 10.24). Note that we used the same set of covariates for both parameters (location and shape), resulting in six parameters (including two constants). We leave it up to the reader to estimate other models by applying different sets of covariates for each. In our case, the economic and statistical significance of the parameters for the mean are similar to the former models with the prime distinction that we are now able to explicitly model the dispersion parameter as a function of explanatory variables. For more information about beta regression in general, we refer to Ferrari and Cribari-Neto (2004).



The NLMIXED Procedure

Specifications	
Data Set	DATA.LGD
Dependent Variable	lgd_time
Distribution for Dependent Variable	General
Optimization Technique	Trust Region
Integration Method	None

Dimensions	
Observations Used	2545
Observations Not Used	0
Total Observations	2545
Parameters	6

Fit Statistics	
-2 Log-Likelihood	−13925
AIC (smaller is better)	−13913
AICC (smaller is better)	−13913
BIC (smaller is better)	−13878

Parameter Estimates

Parameter	Estimate	Standard Error	DF	*t* Value	Pr > |*t*|	95% Confidence Limits		Gradient
b0	−1.9795	0.06634	2545	−29.84	<.0001	−2.1096	−1.8495	2.068E−7
b1	1.4917	0.07815	2545	19.09	<.0001	1.3385	1.6449	6.229E−8
b2	0.6131	0.1024	2545	5.99	<.0001	0.4123	0.8140	7.996E−9
c0	−0.2792	0.05874	2545	−4.75	<.0001	−0.3943	−0.1640	2.906E−7
c1	−0.2827	0.06714	2545	−4.21	<.0001	−0.4144	−0.1511	9.437E−8
c2	−0.1048	0.08190	2545	−1.28	0.2009	−0.2654	0.05583	2.321E−8

Exhibit 10.24 Beta Regression

```
ODS GRAPHICS ON;
PROC NLMIXED DATA=data.lgd TECH =TRUREG;
PARMS   b0 = 0 b1 = 0.001 b2 = 0.0001
c0 = 0 c1 = 0.001 c2 = 0.0001 ;
*Linear predictors;
Xb = b0 + b1 * LTV + b2 * purpose1 ;
Wc = c0 + c1 * LTV + c2 * purpose1 ;
mu = 1 / (1 + exp(-xb));
delta = EXP(Wc);
*transform to standard parameterization;
alpha = mu * delta;
```

```
beta = (1-mu) * delta;
*log-likelihood;
lh =  (GAMMA(alpha + beta) / (GAMMA(alpha) * GAMMA(beta))
    * (lgd_time ** (alpha - 1)) * ((1 - lgd_time) ** (beta - 1)));
ll = LOG(lh);
MODEL lgd_time ~ GENERAL(ll);
PREDICT mu OUT = out_mu;
PREDICT delta OUT = out_delta;
RUN;
ODS GRAPHICS OFF;
```

In order to check the predictions of the beta regression, you can use the predicted values for μ and produce a real-fit plot using PROC GPLOT of the realized LGDs versus the predicted values as done in Exhibit 10.25. Note that the same analysis may be performed for the nonlinear and the fractional logit regression. We limit our demonstration to beta regressions as they are more popular in practice.

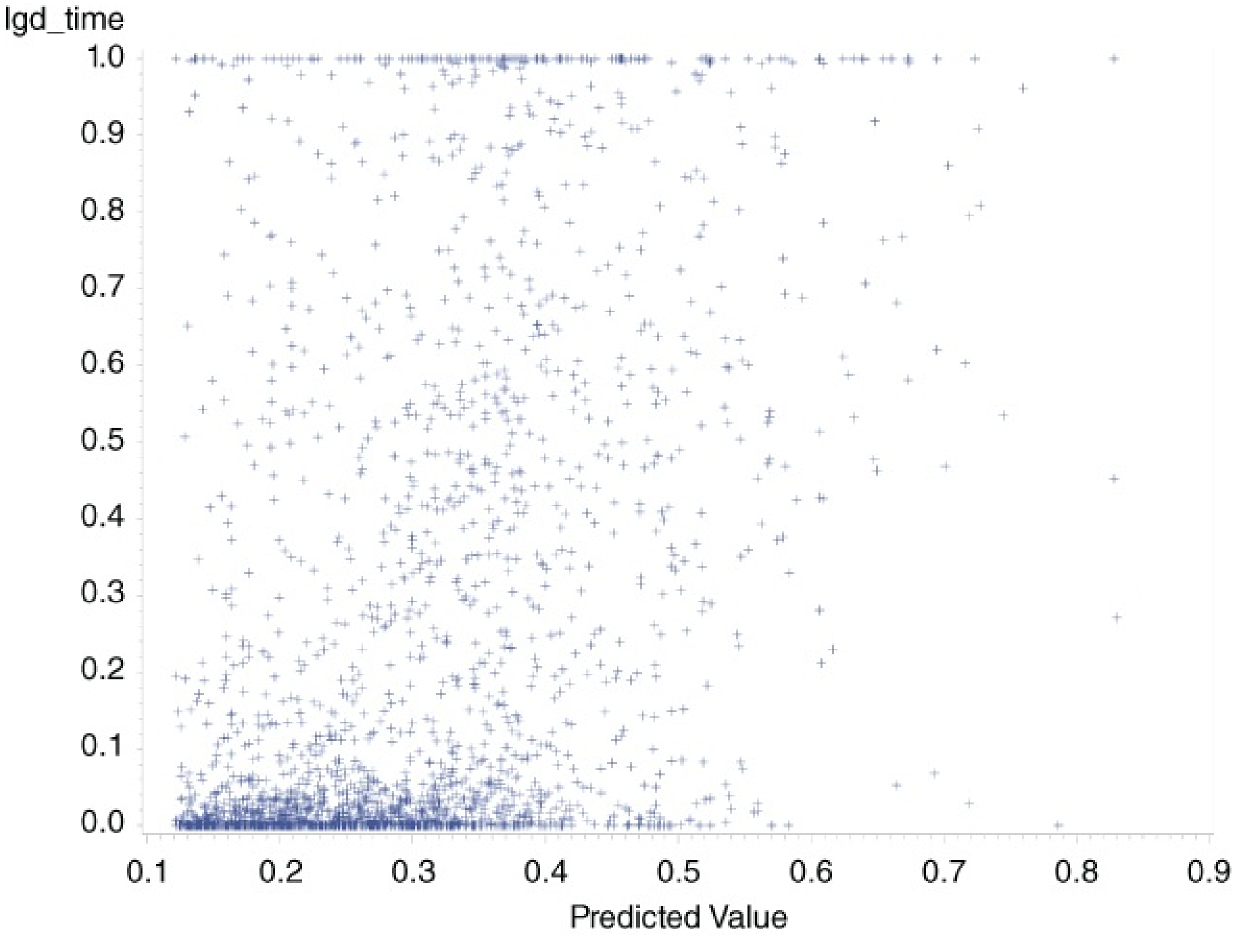
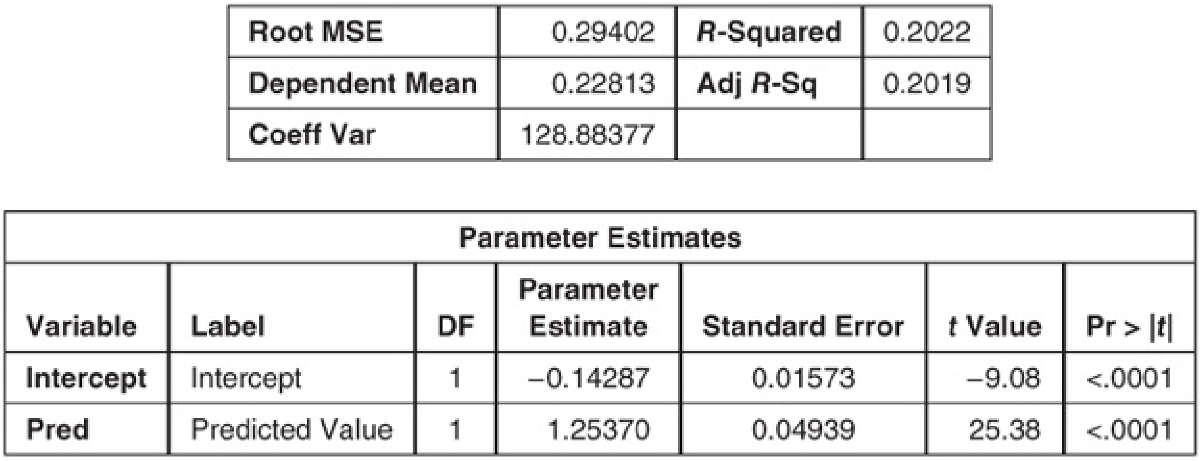


Exhibit 10.25 Real-Fit Plot of Beta Regression

Moreover, one can run a linear regression using PROC REG of realizations against predictions and compute the R^2. As shown in the output (Exhibit 10.26), the R^2 is around 20 percent, which is slightly better than the former regressions, but remember that the predictions are predictions of the mean values only and the standard deviations are modeled individually using covariates. Therefore, the realizations might be far off from the predicted means if the standard deviations are high (and the R^2 low). Modeling the individual standard deviation, however, is one advantage of the beta regression, besides modeling a more realistic distribution within $(0, 1)$ than the normal distribution. The regression also reveals that the intercept is statistically different from zero and the slope is statistically different from one (given that the value of one is more than two standard deviations away from the estimate).



Root MSE	0.29402	*R*-Squared	0.2022
Dependent Mean	0.22813	Adj *R*-Sq	0.2019
Coeff Var	128.88377		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |*t*|
Intercept	Intercept	1	−0.14287	0.01573	−9.08	<.0001
Pred	Predicted Value	1	1.25370	0.04939	25.38	<.0001

Exhibit 10.26 Real-Fit Regression of Beta Regression

```
ODS GRAPHICS ON;
PROC GPLOT DATA=out_mu;
PLOT lgd_time * PRED;
RUN;
PROC REG DATA = out_mu;
MODEL lgd_time = PRED;
RUN;
ODS GRAPHICS OFF;
```

PD-LGD MODELS

The former models use all LGDs that are observable in the sample and treat them in the same way. While widespread in academia and in practice, they exhibit the shortcoming of sample selection. Simply speaking, sample selection arises whenever an observable sample that is used for an analysis such as a regression model is the result of a selection mechanism that is not purely random and that is correlated with the variable of interest. In our data set, we have many cases where the recovery is (close to) one, which means that there is no economic loss for the bank. This could be because no default has happened or because the default has been cured. Both events create a selection mechanism for the recoveries or LGDs, and if there is some dependence between the event and the loss, the selection mechanism is not purely random. As a result from standard theory, estimators that ignore this kind of sample selection are inconsistent. In our data, all cures (i.e., observations with recoveries greater than 0.99999) and nondefaults are flagged by an event variable that is equal to zero (no default) and one otherwise (default).

Tobit Regression

A first model considered is the classic Tobit model; see Tobin (1958). It can be motivated using a classic Merton-type credit model (see Merton, 1974) as derived by Rösch and Scheule (2012). The model generally takes the form

$$\ln\ RR_i = \min\{Y_i^*, 0\} \qquad \textbf{10.4}$$

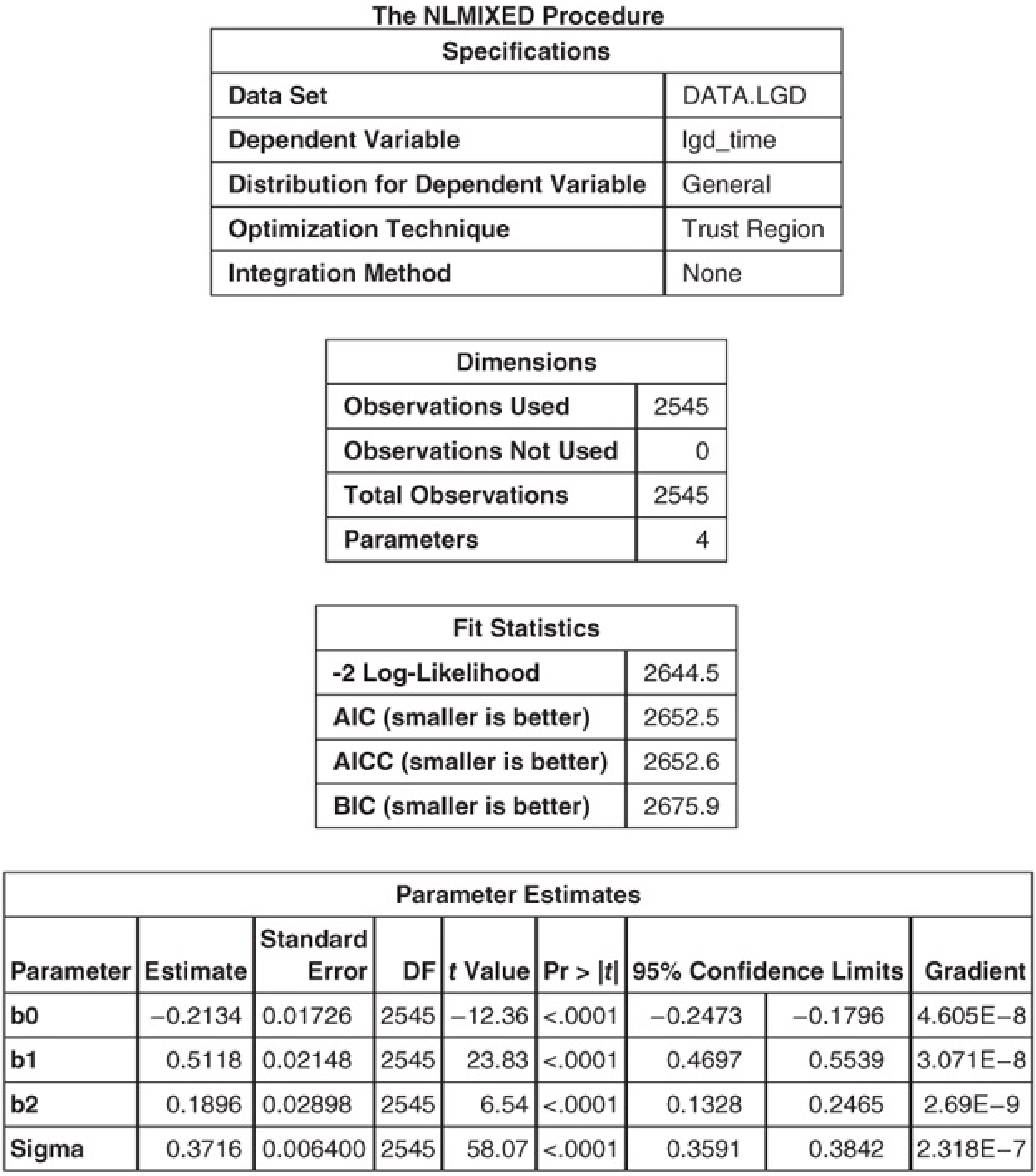
where $RR_i = (1 - LGD_i)$ is the recovery rate and Y_i^* is a latent variable generated from the classic regression model

$$Y_i^* = \beta' x_i + \epsilon_i \qquad \textbf{10.5}$$

The log transformation of the recovery rate is due to the analogy with the Merton model as shown in Rösch and Scheule (2012). Its descriptive statistics were computed and shown earlier in this chapter. In empirical applications, however, other transformations might also be used. Here, for easier comparison with the other approaches, we do not use any transformation at all; that is, we simply use LGD as the dependent variable 10. The lower bound is then censored to be 0.00001 and the equation becomes

$$LGD_i = \max\{0.00001, Y_i^*\}$$

The model acknowledges that recoveries (or LGDs, respectively) can only be observed if the underlying latent default-triggering variable crosses the threshold. The likelihood takes into account that observed values of the dependent variable are conditionally normally distributed and values on the boundary are censored values from the conditional normal distribution. In other words, in the histogram we saw earlier, the $\ln\ RR_i$ can be interpreted as the left tail of a normal distribution; see Rösch and Scheule (2012). The model can be evaluated with PROC NLMIXED and PROC QLIM, and we show that both alternatives obviously yield the same result. (See Exhibit 10.27.)



The NLMIXED Procedure

Specifications	
Data Set	DATA.LGD
Dependent Variable	lgd_time
Distribution for Dependent Variable	General
Optimization Technique	Trust Region
Integration Method	None

Dimensions	
Observations Used	2545
Observations Not Used	0
Total Observations	2545
Parameters	4

Fit Statistics	
-2 Log-Likelihood	2644.5
AIC (smaller is better)	2652.5
AICC (smaller is better)	2652.6
BIC (smaller is better)	2675.9

Parameter Estimates

Parameter	Estimate	Standard Error	DF	*t* Value	Pr > |*t*|	95% Confidence Limits		Gradient
b0	−0.2134	0.01726	2545	−12.36	<.0001	−0.2473	−0.1796	4.605E−8
b1	0.5118	0.02148	2545	23.83	<.0001	0.4697	0.5539	3.071E−8
b2	0.1896	0.02898	2545	6.54	<.0001	0.1328	0.2465	2.69E−9
Sigma	0.3716	0.006400	2545	58.07	<.0001	0.3591	0.3842	2.318E−7

Exhibit 10.27 Tobit Regression with NL Mixed

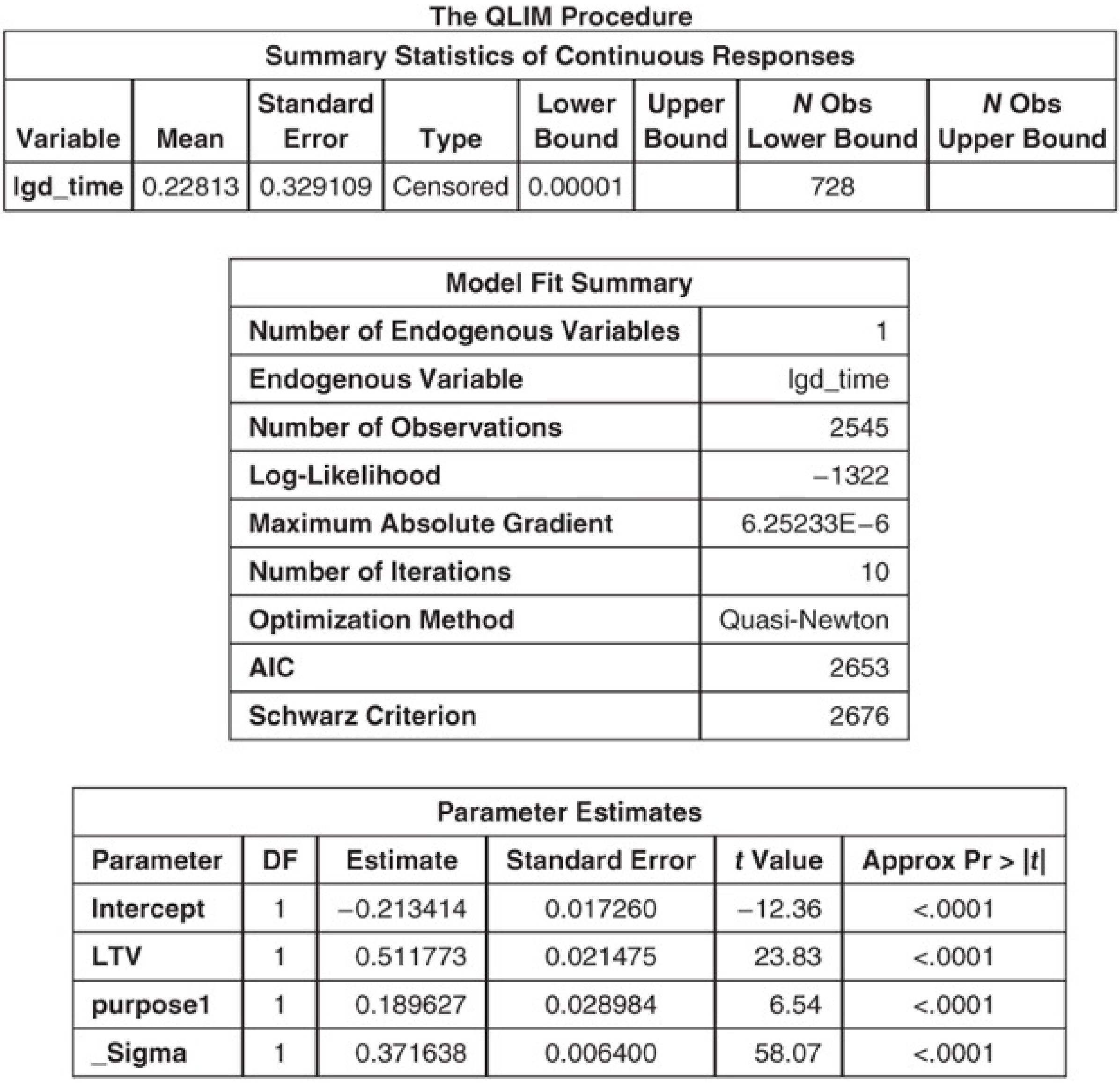
An example for a Tobit model using PROC NLMIXED is:

```
PROC NLMIXED DATA = data.lgd TECH = TRUREG;
PARMS b0 = 0 b1 = 0 b2 = 0  sigma = 1;
xb = b0 + b1 * LTV + b2 * purpose1;
IF event  = 1 THEN lh = pdf('NORMAL', lgd_time, xb, sigma);
ELSE IF event = 0 THEN lh = CDF ('NORMAL', 0, xb, sigma);
ll = LOG(lh);
MODEL lgd_time ~ GENERAL(ll);
RUN;
```

As previously, we include LTV and rental purpose as explanatory variables. Since we model the LGD as dependent variable, both coefficients have positive signs and are highly significant.

While the results are economically similar to the results from the standard nonselection models, there can, however, arise situations where the significance of the coefficients is different and variables that are not significant in standard models might become significant if the selection mechanism is taken into account.

An example for a Tobit model using PROC QLIM is (see Exhibits 10.28 and 10.29):



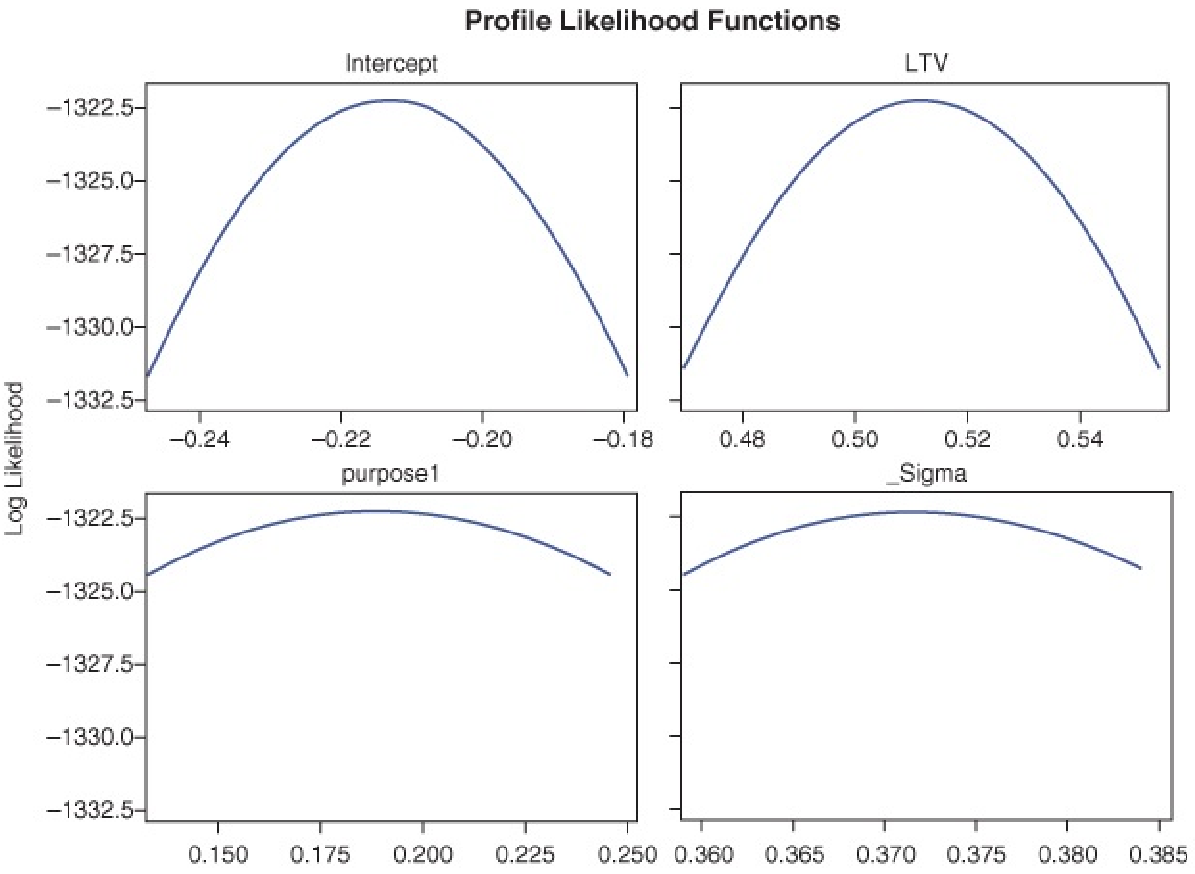
The QLIM Procedure

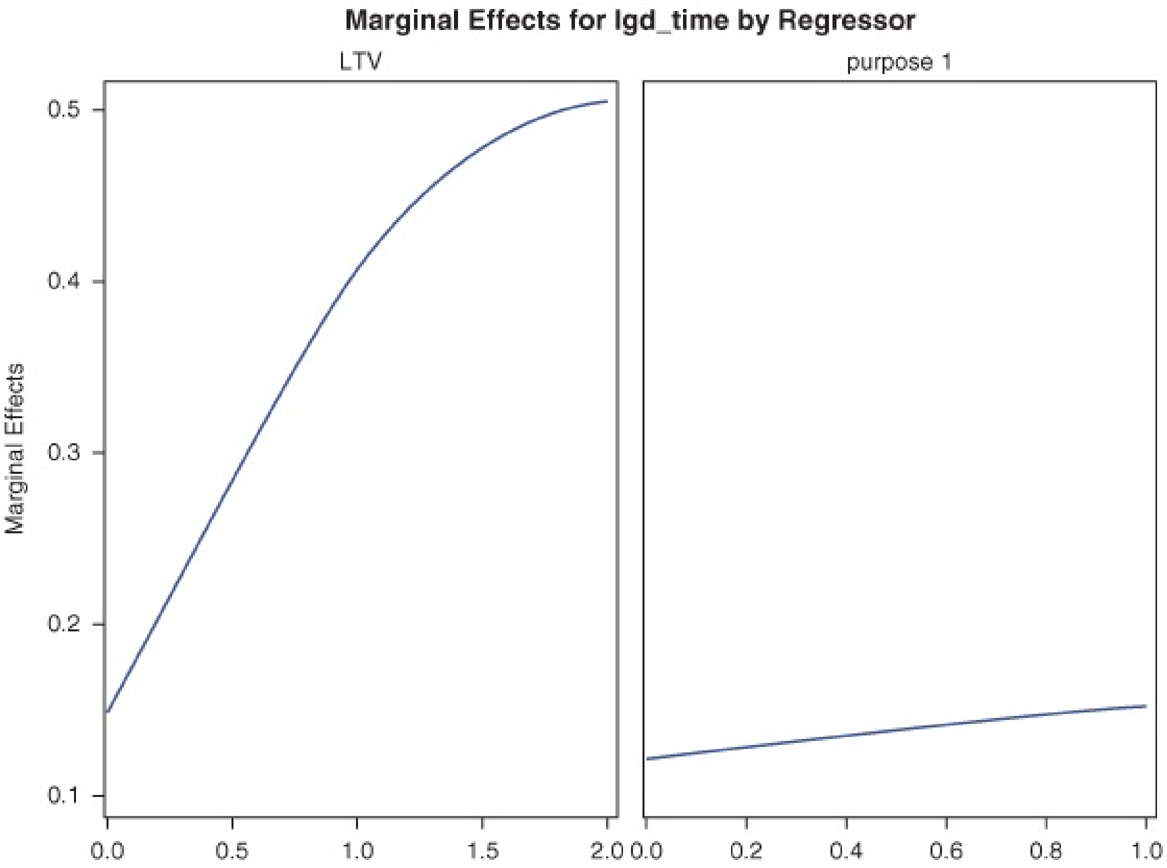
Summary Statistics of Continuous Responses							
Variable	Mean	Standard Error	Type	Lower Bound	Upper Bound	*N* Obs Lower Bound	*N* Obs Upper Bound
lgd_time	0.22813	0.329109	Censored	0.00001		728	

Model Fit Summary	
Number of Endogenous Variables	1
Endogenous Variable	lgd_time
Number of Observations	2545
Log-Likelihood	−1322
Maximum Absolute Gradient	6.25233E−6
Number of Iterations	10
Optimization Method	Quasi-Newton
AIC	2653
Schwarz Criterion	2676

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	*t* Value	Approx Pr > |*t*|
Intercept	1	−0.213414	0.017260	−12.36	<.0001
LTV	1	0.511773	0.021475	23.83	<.0001
purpose1	1	0.189627	0.028984	6.54	<.0001
_Sigma	1	0.371638	0.006400	58.07	<.0001

Exhibit 10.28 Tobit Regression with QLIM


Profile Likelihood Functions
Intercept
LTV
purpose1
_Sigma
Log Likelihood
−1322.5
−1325.0
−1327.5
−1330.0
−1332.5
−0.24
−0.22
−0.20
−0.18
0.48
0.50
0.52
0.54
0.150
0.175
0.200
0.225
0.250
0.360
0.365
0.370
0.375
0.380
0.385


Marginal Effects for lgd_time by Regressor
LTV
purpose 1
Marginal Effects
0.5
0.4
0.3
0.2
0.1
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0

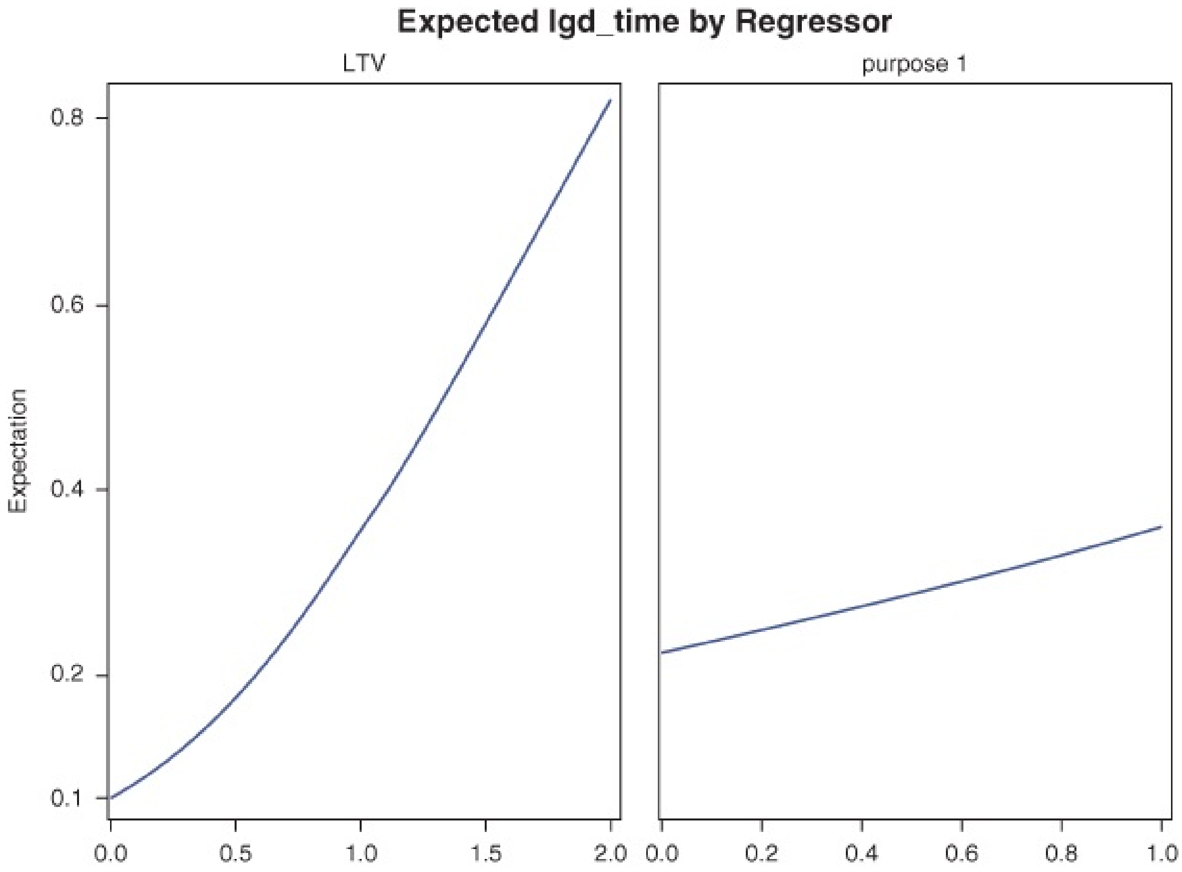


Exhibit 10.29 Tobit Regression with QLIM

```
PROC QLIM DATA=data.lgd  PLOTS=ALL ;
MODEL lgd_time =  LTV  purpose1;
ENDOGENOUS  lgd_time ~ CENSORED(LB=0.00001);
OUTPUT OUT = tobit1_out EXPECTED CONDITIONAL PROB RESIDUAL XBETA;
RUN;
```

The figures from PROC QLIM show the marginal profile likelihood functions, that is the likelihood values around the optimum, the marginal effects of the explanatory variables on the dependent variable, and the fitted expected recovery by regressor. The expression is given in Rösch and Scheule (2012) for the transformed model.

For this model, it is also possible to provide a real-fit check and an R^2, similarly as in the case of the beta regression. (See Exhibits 10.30 and 10.31.) In PROC QLIM, one can compute the expected LGDs or the conditional expected LGDs, conditional on default. Both are computed in the program by the OUTPUT statement, and the latter are plotted against the realized values and used for computation of the R^2, which is almost 20 percent. When looking at the scatter plot, again remember that the predictions are for expected values, and the still rather low R^2 and the high estimate for the volatility of about 0.37 indicate that the realizations are (randomly) far off the expectations. As the exercises in this chapter are for demonstration purposes only, it is recommended that practitioners look for other important explanatory variables.

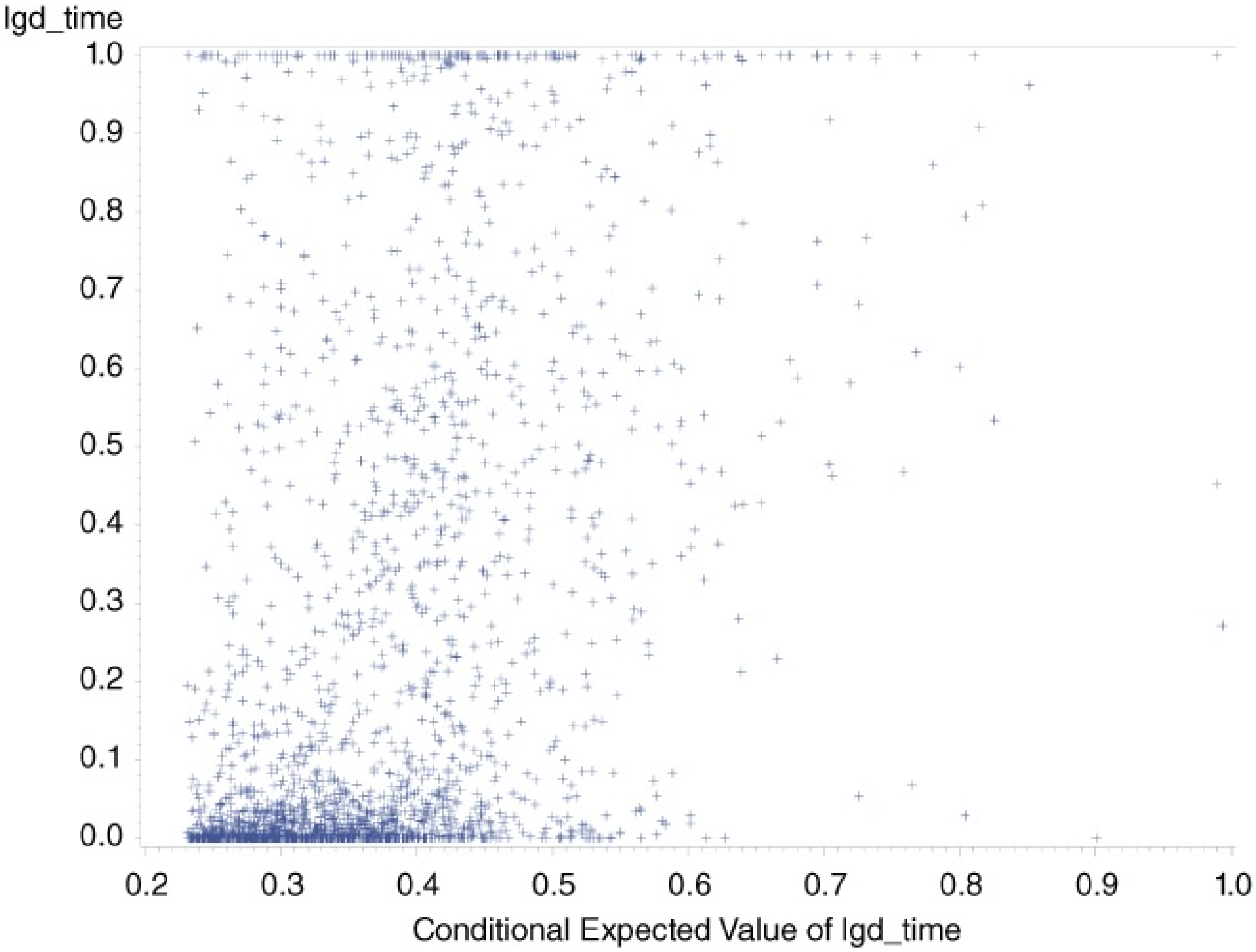
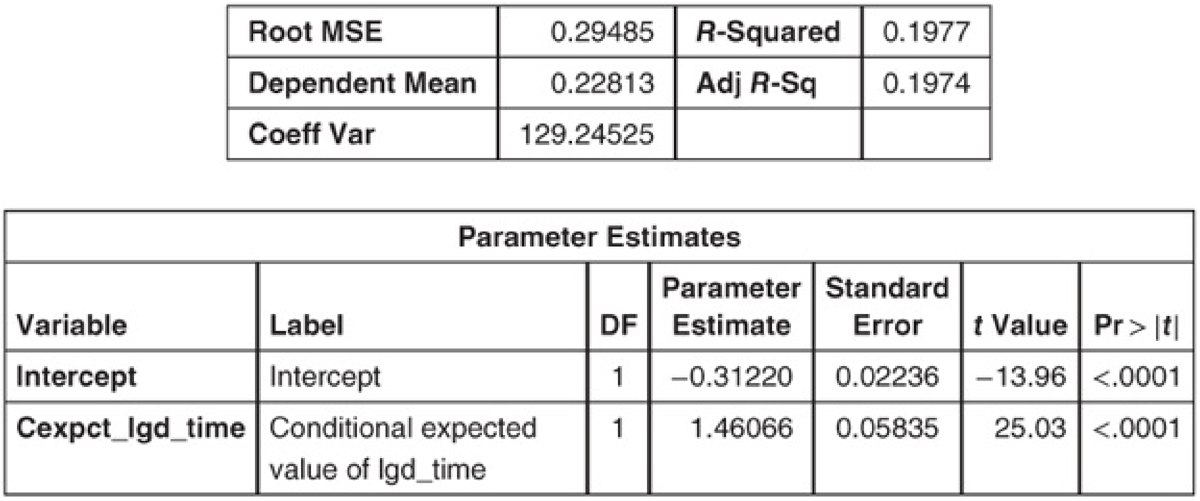


Exhibit 10.30 Real-Fit Plot of Tobit Regression



Root MSE	0.29485	***R*-Squared**	0.1977
Dependent Mean	0.22813	**Adj *R*-Sq**	0.1974
Coeff Var	129.24525		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |*t*|
Intercept	Intercept	1	−0.31220	0.02236	−13.96	<.0001
Cexpct_lgd_time	Conditional expected value of lgd_time	1	1.46066	0.05835	25.03	<.0001

Exhibit 10.31 Real-Fit Regression of Tobit Regression

```
ODS GRAPHICS ON;
PROC GPLOT DATA=Tobit1_out;
PLOT lgd_time * CExpct_lgd_time;
RUN;
PROC REG DATA = Tobit1_out;
MODEL lgd_time = CExpct_lgd_time;
RUN;
ODS GRAPHICS OFF;
```

Heckman Sample Selection Model

A generalization of the Tobit model is the Heckman model. Whereas the Tobit model uses the same mechanism for selection (i.e., a variable is observed if it crosses some threshold), the Heckman model uses two equations (see Bade, Rösch, & Scheule 2011 for an application). First, the selection process is given by the threshold model

$$D_i = \begin{cases} 1 & Z_i^* < 0 \\ 0 & Z_i^* \geq 0 \end{cases} \tag{10.6}$$

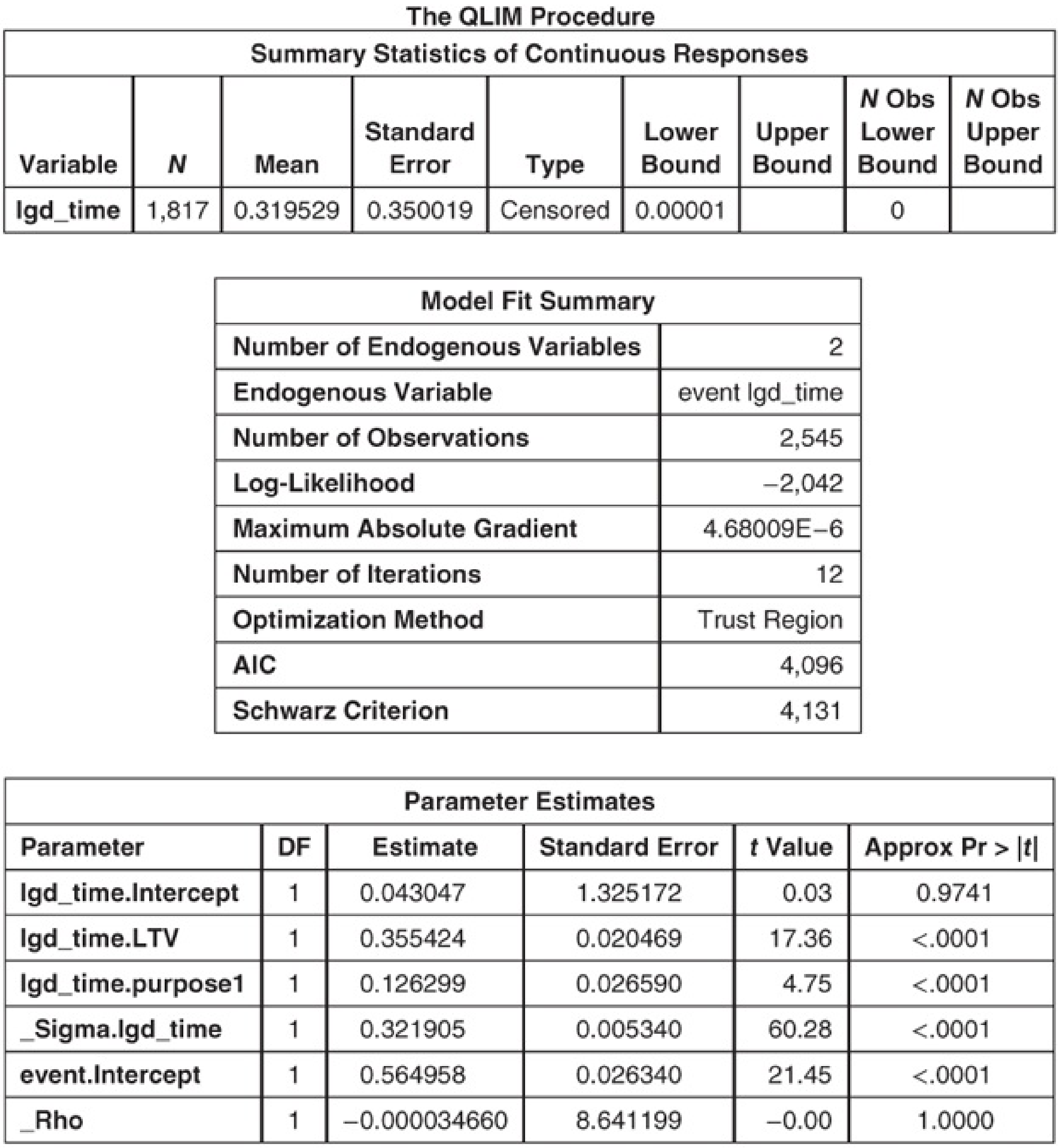
where

$$Z_i^* = \boldsymbol{\beta}' \mathbf{x}_i^z + \epsilon_i^z \tag{10.7}$$

and second, the recovery process is given as

$$\ln\ RR_i = \boldsymbol{\beta}' \mathbf{x}_i + \epsilon_i \quad if \quad D_i = 1 \tag{10.8}$$

where ϵ_i^z and ϵ_i are jointly normal with zero means, standard deviations of one and σ respectively, and correlation ρ . Selection is based on the variable D_i, and the recovery is observed when D_i has a value of 1. D_i is usually the default event; therefore recoveries or LGD are observed only in case of default. The correlation between the residuals is then the correlation between the LGD and the default-triggering process (e.g., asset value process). The code in PROC QLIM and the estimation output (Exhibit 10.32) are given next, where we again use LGD instead of $\ln\ RR$ as the dependent variable.



The QLIM Procedure

Summary Statistics of Continuous Responses								
Variable	*N*	Mean	Standard Error	Type	Lower Bound	Upper Bound	*N* Obs Lower Bound	*N* Obs Upper Bound
lgd_time	1,817	0.319529	0.350019	Censored	0.00001		0	

Model Fit Summary	
Number of Endogenous Variables	2
Endogenous Variable	event lgd_time
Number of Observations	2,545
Log-Likelihood	−2,042
Maximum Absolute Gradient	4.68009E−6
Number of Iterations	12
Optimization Method	Trust Region
AIC	4,096
Schwarz Criterion	4,131

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	*t* Value	Approx Pr > |*t*|
lgd_time.Intercept	1	0.043047	1.325172	0.03	0.9741
lgd_time.LTV	1	0.355424	0.020469	17.36	<.0001
lgd_time.purpose1	1	0.126299	0.026590	4.75	<.0001
_Sigma.lgd_time	1	0.321905	0.005340	60.28	<.0001
event.Intercept	1	0.564958	0.026340	21.45	<.0001
_Rho	1	−0.000034660	8.641199	−0.00	1.0000

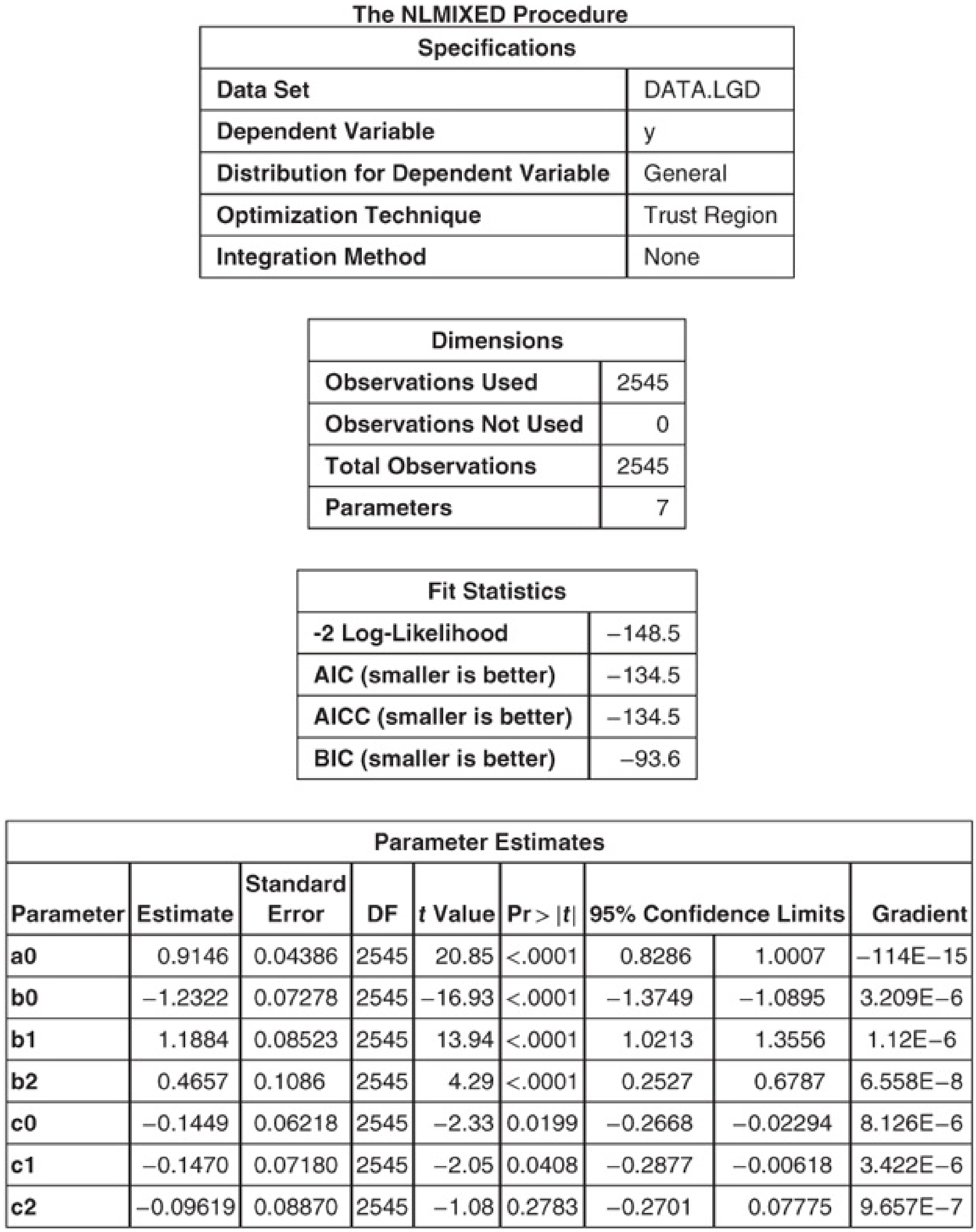
Exhibit 10.32 Heckman Regression with QLIM

The intercept for the event process gives the selection probability (it could, however, also be modeled using explanatory variables). Both explanatory variables for the LGD process are again highly significant. The output also gives an estimate for the correlation ρ between both processes that is close to zero and not significant, showing that in this case, sample selection shouldn't be a big issue. You can now also use the computed output similarly to the Tobit model and compare realizations and predicted expectations.

```
PROC QLIM DATA=data.lgd PLOTS=ALL  METHOD = TRUREG;
MODEL event = / DISCRETE;
MODEL lgd_time = LTV purpose1/ SELECT(event=1) CENSORED(lb=0.00001);
OUTPUT OUT = heckman1_out EXPECTED CONDITIONAL PROB RESIDUAL XBETA;
RUN;
```

Censored Beta Regression

The final model with sample selection is a variant of the beta regression that takes selection and censoring explicitly into account. (See Exhibit 10.33.) As in the Tobit model, censored values for the LGD are explicitly taken into account and enter the likelihood; see Liu and Zhao (2013). The model then parameterizes the mean and the dispersion of the beta distribution as well as the censoring equation (the latter with a constant only). Again, while we used the same variables for both LGD parameters, it is not a requirement to do so and the reader is invited to check other combinations for the three equations. All coefficients are statistically and economically significant.



The NLMIXED Procedure

Specifications	
Data Set	DATA.LGD
Dependent Variable	y
Distribution for Dependent Variable	General
Optimization Technique	Trust Region
Integration Method	None

Dimensions	
Observations Used	2545
Observations Not Used	0
Total Observations	2545
Parameters	7

Fit Statistics	
-2 Log-Likelihood	−148.5
AIC (smaller is better)	−134.5
AICC (smaller is better)	−134.5
BIC (smaller is better)	−93.6

Parameter Estimates

Parameter	Estimate	Standard Error	DF	*t* Value	Pr > |*t*|	95% Confidence Limits		Gradient
a0	0.9146	0.04386	2545	20.85	<.0001	0.8286	1.0007	−114E−15
b0	−1.2322	0.07278	2545	−16.93	<.0001	−1.3749	−1.0895	3.209E−6
b1	1.1884	0.08523	2545	13.94	<.0001	1.0213	1.3556	1.12E−6
b2	0.4657	0.1086	2545	4.29	<.0001	0.2527	0.6787	6.558E−8
c0	−0.1449	0.06218	2545	−2.33	0.0199	−0.2668	−0.02294	8.126E−6
c1	−0.1470	0.07180	2545	−2.05	0.0408	−0.2877	−0.00618	3.422E−6
c2	−0.09619	0.08870	2545	−1.08	0.2783	−0.2701	0.07775	9.657E−7

Exhibit 10.33 Beta Regression with Censoring and Selection

```
ODS GRAPHICS ON;
PROC NLMIXED DATA = data.lgd TECH = TRUREG;
PARMS  a0 = 0
b0 = 0 b1 = 0  b2=0
c0 = 1 c1 = 0  c2=0  ;
y = lgd_time;
xa = a0  ;
xb = b0 + b1 * LTV  + b2 * purpose1 ;
xc = c0 + c1 * LTV  + c2 * purpose1 ;
```

```
mu_xa = 1 / (1 + exp(-xa));
mu_xb = 1 / (1 + exp(-xb));
delta = EXP(xc);
alpha = mu_xb * delta;
beta = (1 - mu_xb) * delta;
IF  event=0 THEN lh = 1 - mu_xa;
ELSE lh = mu_xa * (GAMMA(alpha + beta) /
(GAMMA(alpha) * GAMMA(beta)) * (y ** (alpha - 1))
* ((1 - y) ** (beta - 1)));
ll = LOG(lh);
MODEL y ~ GENERAL(ll);
RUN;
ODS GRAPHICS OFF;
```

Which Model Should I Choose?

The final question is: Which model should I choose? While the results in this section didn't show a clear winner, some thoughts about the models presented are in order. First, it is a good idea not to limit oneself to one model only, but rather to estimate a variety of different models, to see whether the outcomes are robust with respect to the model specification. This is why we presented more models than necessary at first sight. You should take a closer look at the data and the reasons why a specific model matches the data better than another model if the outcome of one model is completely different from another model (e.g., in terms of significances of parameters or prediction power). One plausible explanation may be that the assumptions of the model (e.g., distributional assumptions) provide for a better fit. Moreover, robustness checks across models are also important for regulatory purposes.

Second, while sample selection was not a big issue in our data set, it can be an issue for other data and can lead to biased and inconsistent parameter estimates. Therefore, a selection model should always be run in parallel.

Third, when choosing among different models, you should keep in mind complexity. Sometimes, a simpler model is easier to communicate within an institution and to supervisors if it yields results that are similar to those of a more advanced, complex model.

EXTENSIONS

This chapter has focused on the most fundamental methods for modeling and forecasting LGDs and recovery rates. Depending on the data that are available for a bank and depending on econometric skills, various extensions can be applied. In our data set, we used only two explanatory variables, which yielded R^2 values of about 20 percent. If more covariates are available, you could include more information in order to get a better picture about LGDs. Examples are loan- or borrower-specific data, such as seniority, collateral, income, and so on, or macroeconomic variables, such as interest rates, unemployment rate, or GDP. For applications, see for instance Rösch and Scheule ([2004, 2005, 2009, 2012]) or Bellotti and Crook (2012).

The models can also be extended by including nonobservable random effects that can be interpreted similarly to asset correlations for default events. By modeling LGDs via observable *and* unobservable components, it is also easy to do stress-testing (by stressing the observable, e.g., macroeconomic covariates to certain levels) or to provide downturn predictions for LGDs (by stressing observable and unobservable random components). For applications in this context see Rösch and Scheule (2010).

Other selection mechanisms and other dependence structures can also be implemented. For example, a selection mechanism for default and another for cure events can be included; see Wolter and Rösch (2014). An application of various copula dependencies between PD and LGD, including the generation of a term structure of LGDs over the lifetime of a loan, is given in Krüger et al. (2015).

Finally, besides the techniques we already discussed, more complex techniques such as neural networks or support vector machines (SVMs) can also be used for LGD modeling; see Loterman et al. (2012). Although these techniques benefit from a universal approximation property and are thus very powerful, they usually suffer from a loss of interpretability. The resulting models are very complex for the human decision maker to understand. Hence, it is not recommended to use these techniques for LGD modeling, particularly for supervisory purposes.

PRACTICE QUESTIONS

1. Explain why LTV should be a good explanatory variable for LGDs.
2. Categorize the current LTV ratio and include dummy variables in a regression model for the LGD. How do you interpret the parameter estimates for the LTV ratio, and what may be the advantages and disadvantages of including a metric variable in categories (relative to a stand-alone inclusion)? Use data set lgd.
3. In the PD chapter, we discussed point-in-time (PIT) versus through-the-cycle (TTC). Could this also be an issue for LGDs? How could macroeconomic variables (such as GDP) affect LGDs?
4. Include “LTV” and “purpose1” as explanatory variables in the probit transformed regression model, and estimate a stepwise regression using the option “stepwise” in PROC REG. Interpret the results. Use data set lgd.
5. Which other variables could be included in an LGD model and should be collected by banks?
6. Discuss the issue of sample selection for LGD models.

References

Bade, B., D. Rösch, and H. Scheule. 2011. “Default and Recovery Risk Dependencies in a Simple Credit Risk Model.”*European Financial Management* 17(1), 120–144.

Basel Committee on Banking Supervision. 2005a. “Guidance on Paragraph 468 of the Framework Document.”

Basel Committee on Banking Supervision. 2005b. “The Joint Forum Credit Risk Transfer.”

Basel Committee on Banking Supervision. 2006. “International Convergence of Capital Measurement and Capital Standards: A Revised Framework, Comprehensive Version.”

Bellotti, T., and J. Crook. 2012. “Loss Given Default Models Incorporating Macroeconomic Variables for Credit Cards.” *International Journal of Forecasting* 28(1): 171–182.

Betz, J., R. Kellner, and D. Rösch. 2016. “What Drives the Time to Resolution of Defaulted Bank Loans?” *Finance Research Letters* (April).

Brady, B., P. Chang, P., Miu, B. Ozdemir, and D. Schwartz. 2006. “Discount Rate for Workout Recoveries: An Empirical Study.” Technical report, Working paper.

Committee of European Banking Supervisors (CEBS). 2005. “Guidelines on the Implementation, Validation and Assessment of Advanced Measurement (AMA) and Internal Ratings Based (IRB) Approaches.” Technical report, CP10 consultation paper.

Federal Register. 2007. “Proposed Supervisory Guidance for Internal Ratings-Based Systems for Credit Risk, Advanced Measurement Approaches for Operational Risk, and the Supervisory Review Process (Pillar 2) Related to Basel II Implementation.”

Ferrari, S., and F. Cribari-Neto. 2004. “Beta Regression for Modelling Rates and Proportions.” *Journal of Applied Statistics* 31(7), 799–815.

Gupton, G., and R. Stein. 2005. “Losscalc v2: Dynamic Prediction of LGD Modeling Methodology.” Moody's KMV Investors Services.

Krüger, S., T. Oehme, D. Rösch, and H. Scheule. 2015. “Expected Loss over Lifetime.” Working paper, University of Regensburg and University of Technology, Sydney, Australia.

Liu, W., and K. Zhao. 2013. “Statistical Models for Proportional Outcomes.” Working paper, MidWest SAS Users Group, FS-05-2013.

Loterman, G., I. Brown, D. Martens, C. Mues, and B. Baesens. 2012. “Benchmarking Regression Algorithms for Loss Given Default Modeling.” *International Journal of Forecasting* 28: 161–170.

Maclachlan, I. 2004. “Choosing the Discount Factor for Estimating Economic LGD.” In: Altman E, Resti A, Sironi A (Eds.): Recovery Risk, The Next Challenge in Credit Risk Management. Risk Books, London, 285–305.

Merton, R. C. 1974. “On the Pricing of Corporate Debt: The Risk Structure of Interest Rates.” *Journal of Finance* 29: 449–470.

Prudential Regulation Authority. 2013. “Internal Ratings Based approaches.” Working paper.

Rösch, D., and H. Scheule. 2004. “Forecasting Retail Portfolio Credit Risk.” *Journal of Risk Finance* 5(2): 16–32.

Rösch, D., and H. Scheule. 2005. “A Multi-Factor Approach for Systematic Default and Recovery Risk.” *Journal of Fixed Income* 15(2): 63–75.

Rösch, D., and H. Scheule. 2009. “Forecasting Downturn Credit Portfolio Risk.” *Financial Markets, Institutions and Instruments* 18(1), 1–26.

Rösch, D., and H. Scheule. 2010. “Downturn Credit Portfolio Risk, Regulatory Capital and Prudential Incentives.” *International Review of Finance* 10(2): 185–207.

Rösch, D., and H. Scheule. 2012. “Forecasting Probabilities of Default and Loss Rates Given Default in the Presence of Selection.” *Journal of the Operational Research Society* 65(3): 393–407.

SAS Institute Inc. 2015. *SAS/STAT 14.1 User's Guide: Technical Report*. Cary, NC: SAS Institute.

Stoyanov, S. 2009. “Application LGD Model Development.” *Credit Scoring and Credit Control XI Conference*.

Tobin, J. 1958. “Estimation of Relationships for Limited Dependent Variables.” *Econometrica* 26.

Van Gestel, T., and B. Baesens. 2009. *Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*. Oxford: Oxford University Press.

Wolter, M., and D. Rösch. 2014. “Cure Events in Default Prediction.” *European Journal of Operational Research* 238: 846–857.