Using out-of-sample Cox–Snell residuals in time-to-event forecasting

Ekaterina V. Rumyantseva
E-mail: evrumyantseva@hse.ru

Kirill K. Furmanov (iD)
E-mail: kfurmanov@hse.ru

National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

Abstract

The problem of assessing out-of-sample forecasting performance of event-history models is considered. Time-to-event data are usually incomplete because the event of interest can happen outside the period of observation or not happen at all. In this case, only the shortest possible time is observed and the data are right censored. Traditional accuracy measures like mean absolute or mean squared error cannot be applied directly to censored data, because forecasting errors also remain unobserved. Instead of mean error measures, researchers use rank correlation coefficients: concordance indices by Harrell and Uno and Somers' Delta. These measures characterize not the distance between the actual and predicted values but the agreement between orderings of predicted and observed times-to-event. Hence, they take almost "ideal" values even in presence of substantial forecasting bias. Another drawback of using correlation measures when selecting a forecasting model is undesirable reduction of a forecast to a point estimate of predicted value. It is rarely possible to predict the timing of an event precisely, and it is reasonable to consider the forecast not as a point estimate but as an estimate of the whole distribution of the variable of interest. The article proposes computing Cox–Snell residuals for the test or validation dataset as a complement to rank correlation coefficients in model selection. Cox–Snell residuals for the correctly specified model are known to have unit exponential distribution, and that allows comparison of the observed out-of-sample performance of a forecasting model to the ideal case. The comparison can be done by plotting the estimate of integrated hazard function of residuals or by calculating the Kolmogorov distance between the observed and the ideal distribution of residuals. The proposed approach is illustrated with an example of selecting a forecasting model for the timing of mortgage termination.

Introduction

There are various problems in statistics and data analysis that require modelling the time at which a certain event occurs. It can be the timing of a credit default in financial applications, time until death or recovery of a patient in survival analysis, the age of a woman at first marriage or age of the mother at first birth in social and demographical research. Such problems are considered in a branch of statistics called event-history analysis. This branch has important peculiarities that distinguish event-history analysis from more traditional statistics. One of these peculiarities is data censoring.

Every study lasts for a finite period of time, and the event of interest does not necessarily occur within this period. More than that, there may be objects under study that never face the event: some debtors pay off the loan, so that default never occurs for them; some women never give birth to a child. As a result, the only thing that is known about these objects is that the time-to-event exceeds a certain value which is the duration of a timespan between the start of waiting for the event and the end of the study. Such observations are called right censored.

Techniques for analyzing censored data are quite well known; the classical textbook is [1]. Measuring accuracy of time-to-event forecasts is a less developed field of study. Partly this can be explained by the fact that event history models were commonly constructed not for forecasting but for academic purposes like testing hypotheses about the efficiency of a medical treatment or social policy, revealing the individual attributes that are correlated with the duration of unemployment, etc.

The past decade has shown a growing interest in forecasting. On the one hand, nowadays event history models often find purely practical applications: financial risk assessment [2, 3], predicting the length of crowdfunding campaigns [4]. On the other hand, the expansion of machine learning and, in particular, the widespread use of cross-validation procedures has given rise to assessing the quality of statistical models by their out-of-sample predictive accuracy [5, 6]. The special thing is that commonly used accuracy measures like the mean squared error or the mean absolute percentage error are inapplicable when dealing with censored data.

The article proposes an approach to model selection that is based on combination of (1) concordance coefficients for actual and predicted timings and (2) Cox−Snell residuals. Both concordance coefficients and residuals are calculated for a test sample. The proposed approach is illustrated with an example of building a forecasting model for the timing of mortgage prepayment.

The next section describes the basic concepts that have to be defined or explained because of peculiarities of the event-history analysis. Section 2 contains a review of measures of predictive accuracy that can be applied to censored time-to-event data. Section 3 is devoted to Cox−Snell residuals and their use for model assessment. The use of Cox−Snell residuals for model selection is illustrated with a real data example in section 4, which is followed by a conclusion.

1. Probabilistic model of event occurrence

Time of event occurrence is modeled as a nonnegative random variable that can be either

discrete or continuous according to the nature of the process under study and available data. Here we consider only the continuous case. The distribution of this random variable can be characterized by the following functions that play a central role in event history analysis.

Survival function $S(t)$ reflects the probability that time-to-event exceeds the value of the argument:

$$S(t) = P(T > t).$$

The term refers to actuarial and medical applications where the event of interest is death of an insured person or a patient, so that the value of the survival function is the probability of survival until time t.

Hazard function $h(t)$ reflects changes in the probability of event occurrence over time:

$$h(t) = \lim_{\substack{\Delta \to 0 \\ \Delta > 0}} \frac{P(t < T \leq t + \Delta \mid T > t)}{\Delta}.$$

Intgrated hazard function $H(t)$ (also called cumulative hazard) does not have a clear interpretation but plays an important role in this article:

$$H(t) = \int_0^t h(s)ds.$$

Te terminology differs from one area of application to another, and the same functions are known under different names. Survival function is sometimes called reliability function, and the hazard function is also known as the mortality intensity rate or the force of mortality.

Typically, there are two aspects of event history that a researcher is interested in.

The first aspect is a relation between the probability of an event's occurrence in the near future and the time of waiting for the event. This relation is conveniently represented by the hazard function.

The second aspect is a relation between the probability of an event's occurrence and explanatory variables (covariates). There are a variety of regression models that link the distribution of time with covariates; we refer an interested reader to books [1, 7] for a detailed review. Four event-history models that are used in this article for illustration purposes are briefly described below.

Lognormal and generalized gamma regressions are special cases of an accelerated failure-time model, which means that they have a linear form representation:

$$\ln T = x'\beta + \varepsilon.$$

Here x' denotes the row vector of explanatory variables, β is the column vector of estimated coefficients, and ε stands for a random error. Apart from covariates vector x' includes a unit element that corresponds to an intercept term, so that $x'\beta = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k$. Lognormal and generalized gamma regressions differ only in assumed distribution of a random error.

Gompertz and Cox regression are proportional hazard models which means that they assume the hazard function to be proportional to covariates:

$$h(t; x', \beta) = h_0(t) \exp(x'\beta).$$

Here $h_0(t)$ denotes the so-called baseline hazard (the hazard that corresponds to zero effect of explanatory variables). The Gompertz model assumes that the hazard exponentially grows or declines with time: $h_0(t) = e^{\gamma t}$, where γ is an estimated parameter of the baseline distribution. Cox regression does not impose restrictions on the baseline hazard which is estimated via the nonparametric technique. As well as other considered models, the Cox regression restricts a functional form of relation between the time-to-event distribution and explanatory variables, but (in contrast to other models) it permits any kind of dependence between the probability of event occurrence and time.

Estimating procedures for these models are implemented in statistical packages and described in textbooks [1, 7]. The only thing

important for our purposes is that each of these regressions allows a researcher to obtain estimates of survival and integrated hazard functions for arbitrary values of explanatory variables x'. This means the possibility to predict the event's occurrence under the given conditions.

2. Predictive accuracy measures: A review

Measures based on averaging prediction errors. This group includes the most widely used metrics of forecasting accuracy for the models with quantitative response: mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE) etc. Although there are examples of their application in event-history analysis [2, 5, 8], these examples are mere exceptions. In most cases, data are subject to censoring so that the differences between actual and predicted timing of events are not precisely known and cannot be averaged. This problem can easily be solved under the assumption that the prediction errors follow a certain parametric family of distributions. In this case, one may estimate the parameters of this distribution via maximum likelihood and compute the corresponding mean. However, we have not found this approach in academic literature. A possible explanation is that researchers avoid making additional assumptions

Papers [2, 8] consider calculating the mean absolute error only for uncensored observations that contain exact timings of event occurrence. This approach has a substantial drawback because censoring depends on those timings. The longer the observer has to wait for an event, the greater is the probability that the observation period will end before the occurrence and the observation will be censored. Consequently, exact timings will be known mostly in those cases when they are small, and the mean absolute error will take into account only these observations. As a result, a model that predicts early occurrence will be preferred

to a model that gives unbiased forecasts. One can say that excluding censored observations leads to sample truncation which is no less troublesome than censoring [1].

Rank correlation coefficients and concordance indices. Harrell's concordance index [9], or C-index, is probably the most widely used predictive accuracy measure in event history analysis. Let random variables T_1 and T_2 denote times of event occurrence in two randomly chosen independent observations, and $\hat{T}_1$ and $\hat{T}_2$ denote corresponding predictions. Harrell's C-index is defined by the following expression:

$$C = P(\hat{T}_1 < \hat{T}_2 \mid T_1 < T_2). \qquad (1)$$

One advantage of this coefficient is its clear interpretability: if time-to-event differs in two cases, then C equals the probability that a model predicts a greater value in the observation with a greater actual time. The largest possible value of C is one and it is achieved when rankings of actual and predicted values are completely concordant, so that when the event of interest occurs earlier, the model always predicts earlier occurrence too. The lower bound for C-index is zero which means complete discordance of actual and predicted rankings (the earlier the event happens, the longer is predicted waiting time).

There are various estimators for the concordance coefficient in the presence of censoring. One of them is a statistic originally proposed by Harrell et al. [9], another example is Uno's estimator [10] that is gaining growing popularity. Apart from the concordance index, Somers' D correlation coefficient can be used for the same purpose [11, 12].

The mentioned metrics share a drawback. They measure only the association between rankings of actual and predicted values, which means that they assess a model's ability to distinguish the cases where an event occurs relatively early from those where the event occurs relatively late. It is not a predictive accuracy in the sense that it does not reflect the differ-

ence between real and predicted values. Suppose that the model provides forecasts that are exactly ten times greater than observed values. One would hardly call this forecast accurate, but the concordance index or any other rank correlation coefficient would achieve its highest value of one because the predicted ranking is perfectly concordant with the actual one.

Sometimes it is a model's ability to detect objects with relatively long time-to-event that draws the attention of a researcher. It can be so if, for example, the aim of the study is revealing markers of early recovery or death of patients [13]. However, in many practical cases the analyst is interested in the absolute value of an explained variable. Paper [14] provides an example of such a study in medicine, but such interest seems to be more common in financial applications where times of defaults and prepayments determine cash flow [2, 3, 15].

Another class of measures used for evaluating the predictive power of event-history models consists of ***classification metrics*** that evaluate accuracy of binary prediction (whether an event occurs in a certain period of time or not). This class has been actively developed in the last decade [16, 17] and deserves attention, but we do not review these metrics here, because they represent a substantially different approach to forecasting. However, Harrell's C-index can be considered also as a classification metric [18].

3. Cox−Snell residuals and their application to predictive power assessment

Consider a sample $(T_1, x_1'),...,(T_n, x_n')$, where T_i denotes time to event and x_i' is the vector of explanatory variables in observation i. Let $\hat{H}(t; x')$ denote an estimate for the integrated hazard function of T_i random variables (it can be obtained from some regression model). A Cox−Snell residual [19] in observation i is defined as follows: $r_i^{CS} = \hat{H}(T_i; x_i')$.

If the estimate $\hat{H}(t; x')$ coincides with the true function $H(t; x')$, then the Cox−Snell residuals follow exponential distribution with unit mean. In this case, the integrated hazard function of the residuals is $H^{CS}(t) = t$.

Below we describe the visual test that is commonly used for regression diagnostics. It is based upon the Cox−Snell residuals and is performed in three steps.

1. Estimate the regression model and compute Cox−Snell residuals for each observation.

2. Compute the estimate of the integrated hazard function of the residuals $\hat{H}^{CS}(t)$. If some observations on time-to-event are censored, the corresponding residuals are also censored, which should be taken into account. We use the Nelson−Aalen technique [20, 21] to estimate the integrated hazard from censored data.

3. Plot the estimate of the integrated hazard against residuals r_i^{CS}. Further we refer to this plot as the Cox−Snell residuals plot. In case of a correctly specified model, the integrated hazard estimate approximately lies on the line $H(t) = t$ (*Figure 1a*). An example for incorrect model specification is presented in *Figure 1b*.

This test is widely known and described in textbooks; Cox−Snell residual plots are presented by researchers to assess the goodness-of-fit of models they use [7, 22, 23]. We have not found examples of using Cox−Snell residuals for evaluating forecasting accuracy; the possible reasons are discussed in the conclusion to this article. Further Cox−Snell residuals calculated for the test sample are called out-of-sample residuals because they characterize the model's performance outside the sample used for estimation.

Out-of-sample residuals can be used to detect the systematic prediction bias. Consider a case where a fitted hazard function is c times greater than the true hazard: $h(t; x') = c\,h(t; x')$. Then the Cox−Snell residuals are also c times greater that the integrated hazard function: $r_i^{CS} = \hat{H}(T_i; x_i') = cH(T_i; x_i')$. As a result, a residual
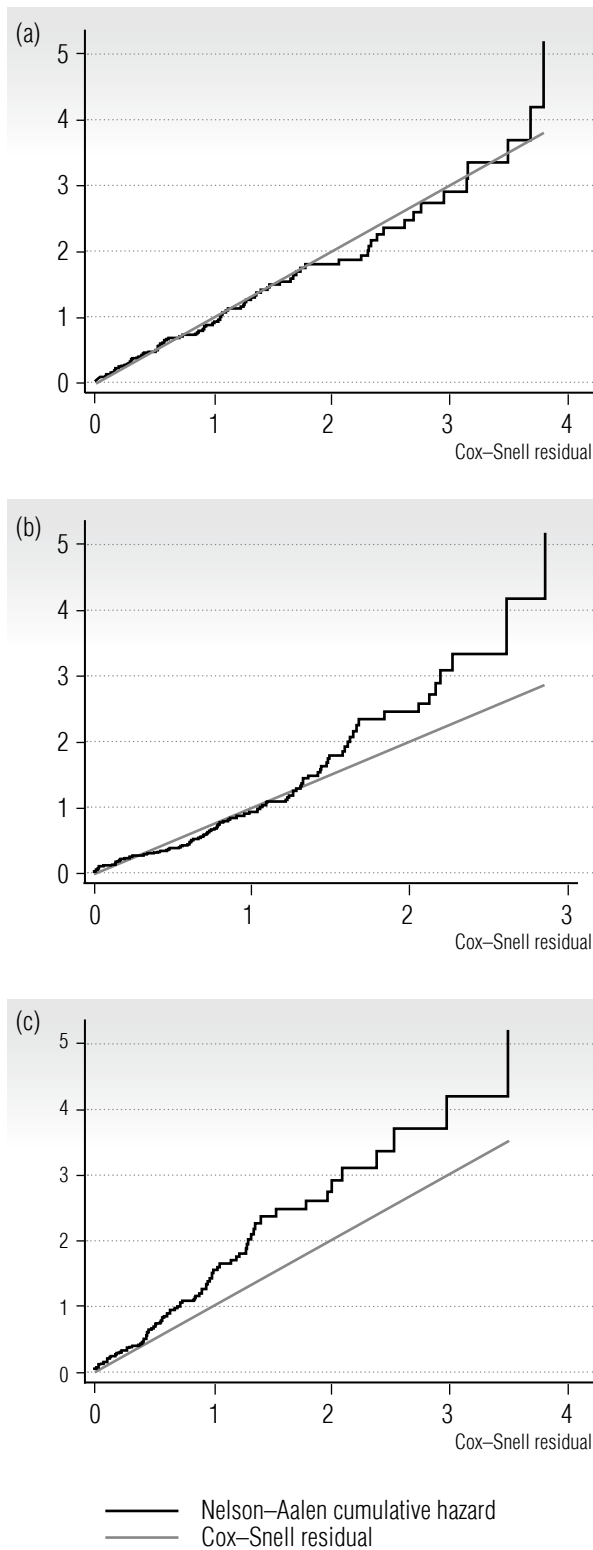
Fig. 1. Examples of Cox–Snell residuals plot for
(a) a correctly specified model;
(b) a model based on inappropriate time–to–event distribution;
(c) a model with systematic prediction bias

plot is higher or lower (depending on the value of c) than the line $H(t) = t$ (*Figure 1c*). This situation is practically impossible when assessing the regression performance in the training sample and is unlikely even when examining out-of-sample predictions if the test sample was selected at random. Such prediction bias is more likely to be found when performing external validation, evaluating the model's performance with new data.

The residual plot is a preferable tool for manual model selection, while automatic selection requires numerical measure for goodness-of-fit. Further we use the Kolmogorov distance between the survival estimate for the Cox–Snell residuals $S^{CS}(t) = \exp(-\hat{H}^{CS}(t))$ and the corresponding function for unit exponential distribution $S(t) = e^{-t}$ as such a metric:

$$KD = \sup_{t \in [0; t_{max}]} \left| \hat{S}^{CS}(t) - e^{-t} \right|. \qquad (2)$$

Here t_{max} denotes the largest time in the sample (it can be censored). We consider only the set $[0; t_{max}]$, because the Nelson−Aalen method does not allow estimating the right tail of the distribution. The corresponding survivor function at t_{max} has not reached zero yet, and there is no data to estimate it for the values of the argument greater than t_{max}.

The next section contains an example of using this metric for predictive model selection.

4. Example: Modeling mortgage prepayment

The example uses data from a large mortgage agency. The data contain more than 280,000 observations on mortgage contracts concluded from 2001 to 2013. The explained variable is the time between a conclusion of a contract and its prepayment. Observations are right-censored due to the following reasons:

◆ The end of the observation period: we use the data gathered on 1 January 2014, these data do not contain information on exact payment date for the loans that were not paid by that date.

✦ The termination of the contract: if the prepayment did not occur, then the observation is treated as right-censored as if there is a possibility of prepayment after the termination date. This is a statistical trick convenient when objects under study are exposed to mutually exclusive events (like prepayment and payment in time in our example). It is convenient to suppose that both events happen but only the first of them is observed.

✦ Mortgage default. This is another event that prevents prepayment.

This dataset is used for estimation of several event-history models that differ in (a) assumed distribution of the explained variable and its relation to the covariate vector (lognormal and generalized gamma regressions, Cox and Gompertz models), and (b) the set of explanatory variables ("short" and "long" models). Both "short" and "long" models include attributes of a loan, of a main borrower and of a subject of mortgage. The "short" model accounts for the interest rate of the loan, the credit term, payment-to-income ratio, age of the main borrower, type of employment of the main borrower and the number of rooms in the subject of mortgage. The "long" model includes, apart from all the mentioned attributes, sex, marital status and education of the main borrower, the number of co-borrowers, regional effect (measured according to the agency's rating of socio-economic development of regions of

Russian Federation that divides all the regions into three groups with low, moderate and high level of development), the type of the mortgage subject (house or apartment), the ratio of living space to total space, the ratio of total amount of planned payments to the price of the mortgage subject, and a loan-to-value ratio.

The whole dataset is randomly partitioned into training and test samples in proportion at a ratio 60:40 respectively.

Plots of out-of-sample Cox–Snell residuals for eight estimated models are given in the Appendix. The distances between observed and theoretical distributions of the residuals are calculated according to equation (2) and presented in *Table 1* among with Harrell's coefficient values.

It is seen from Table 1 that the concordance index is practically the same for different baseline distributions but depends on the choice of covariates: the "long" model stably outperforms the "short" one. The probability of concordance between actual and predicted values is greater for the "long" model by approximately 0.02. An analyst may consider it to be a minor discrepancy but it is stable. Repeating the random split into training and test samples, we found the difference between the values of the *C*-index for "long" and "short" models to be essentially the same. On the contrary, Kolmogorov's distance between ideal and observed distributions of Cox–Snell residuals substan-

Table 1.

Out-of-sample performance of mortgage prepayment models

	"Short" model		"Long" model	
	Harrell's C-index	Kolmogorov distance	Harrell's C-index	Kolmogorov distance
Lognormal	0.593	0.078	0.612	0.066
Generalized gamma	0.593	0.015	0.610	0.009
Gompertz	0.592	0.059	0.608	0.063
Cox	0.593	0.007	0.609	0.014

tially depends on the choice of the baseline distribution: generalized gamma regression and the Cox model provide better goodness-of-fit than lognormal and Gompertz regressions both in the case of "short" and "long" models.

A natural question is: do predictions obtained from these models really differ? *Figure 2* presents plots of estimated survivor functions obtained from the mentioned models for a loan with typical values of explanatory variables:

✦ gender of the main borrower: male;

✦ marital status of the main borrower: married;

✦ education of the main borrower: higher education or doctoral degree;

✦ employment type of the main borrower: private sector employee,

✦ number of co-borrowers including the main borrower: 2;

✦ age of the main borrower: less than 35 years;

✦ payment-to-income ratio: from 20% to 35%;

✦ type of the mortgage subject: apartment;

✦ location of the mortgage subject: moderately developed region;

✦ number of rooms in the mortgage subject: 2;

✦ annual interest rate: less than 11,5% (low-risk interest rate);

✦ ratio of living space to total space of the mortgage subject: from 50% to 70%;

✦ loan-to-value ratio: from 50% to 70%;

✦ credit term: more than 180 months (long term loan);

✦ ratio of total amount of planned payments to the price of the mortgage subject: from 1 to 1.82.

The probability of survival in the next several years for a newly made loan is shown in *Figure 2a*, while *Figure 2b* presents the same probability for a five-year old loan (the conditional survival function with respect to condition $\{T > 5\}$). The horizontal axis represents days after the start of the loan term, and the vertical axis represents the probability that the loan is not paid off until that time. Survival curves obtained from the Cox model ("S_cox" line) and from the generalized gamma regression ("S_gamma") practically

coincide in both figures. The curve estimated from the Gompertz regression ("S_gomp") deviates from the others and shows a greater risk of prepayment. On the contrary, the lognormal curve ("S_ln") predicts the lowest probability of prepayment, and deviates from gamma and Cox curves for an "old" loan but not in the first years of a mortgage credit. Plots of out-of-sample Cox-Snell residuals that are given in Appendix, show that the gamma model correctly predicts the distribution of prepayment times and the Cox regression performs almost as well. Residuals from the Gompertz regression have a distribution that is completely different from exponential, and the lognormal model provides residuals that lack goodness-of-fit in the right tail of the distribution.

Hence, if one chooses a model on the basis of Harrell's index, then the lognormal regression is chosen which leads to underestimation of the probability of prepayment. Taking into account out-of-sample Cox−Snell residuals, an analyst would prefer either generalized gamma regression or the Cox model, both resulting in similar forecasts that agree with the distribution of prepayment times in the test set. From *Figure 2* it follows that the estimated probability of prepayment substantially depends on the choice of forecasting model. The difference is especially large for a long-term forecast horizon: the distance between the survival curves in *Figure 2b* exceeds 20 percentage points in the right side of the plot. However, the difference between two-year-ahead forecasts already exceed 5 percentage points.

Conclusion

Cox−Snell residuals and the corresponding visual test for model specification are well-known and covered in both research papers and textbooks. However, these residuals do not seem to be a popular tool for assessing predictive accuracy. This is an important point that naturally raises a question: are they really useful? The example considered in a previous sec-
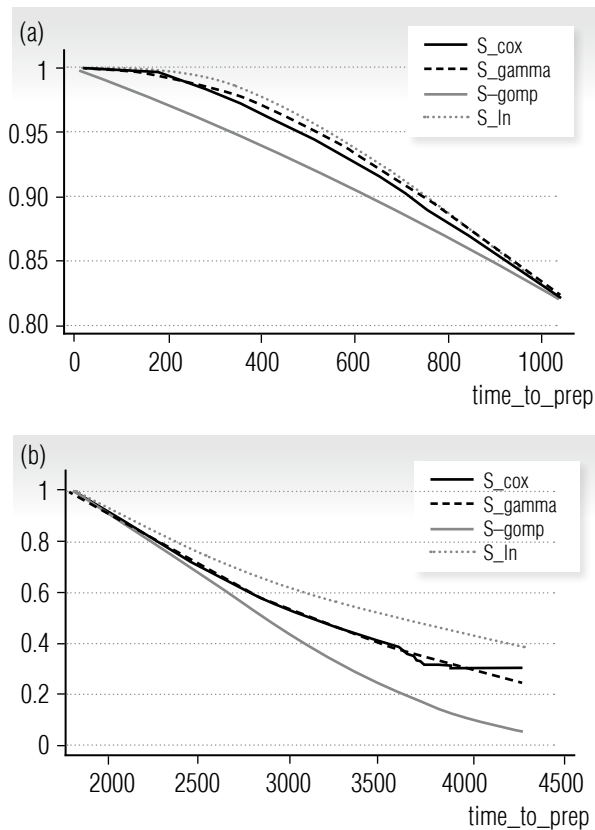
Fig. 2. Estimated survival functions for
(a) a new loan; (b) a five–year loan

On the contrary, rank correlation and concordance coefficients are a useful tool for selecting a set of explanatory variables, but do not characterize the ability of a model to predict the whole distribution of event timings. More common problems of regression analysis do not often require such prediction, and an analyst is interested mostly in a point forecast that can be obtained by estimating the mean or median of an explained variable. However, a time of event occurrence can almost never be characterized by a single number and reducing the forecast to a point estimate is counterproductive. Quantiles are of no less importance than the mean, and as the quantile function uniquely characterize the distribution of a random variable, one can say that the whole distribution is important. Often researchers do not pay much attention to distribution selection and prefer to use the Cox model that does not require parameterizing a baseline hazard. The example presented in this article demonstrates that a relatively simple parametric model can outperform Cox regression even on a very large sample. The Cox model is prone to overfit, and it restricts the functional form of relation between hazard and covariates, so that the shape of the hazard function is essentially the same for all values of explanatory variables. Parametric models can account for other forms of dependence.

Examining the concordance index and the Cox–Snell residuals plot, a researcher can assess both the choice of explanatory variables and the selected baseline distribution. This combination gives a full picture of the model's out-of-sample performance. Of course, it would be more convenient to have a single metric that would allow unambiguous ranking for a set of models according to their forecasting accuracy, but at the moment such a metric seems to be unavailable. ∎

tion was presented to demonstrate the benefits of examining out-of-sample residuals, but now it is time to make a substantial caveat.

Cox–Snell residuals alone cannot be regarded as a measure of predictive error. They only allow us to compare the actual distribution in the data with the distribution that should be observed according to the model. This comparison can be performed either on the training sample (as is usually done), or on the test sample (as we suggest for evaluating forecasting performance). Note that a model with no covariates, that yields the same prediction for all observations, may produce practically ideal exponential Cox–Snell residuals if a proper distribution for an explained variable is chosen. One need not think that such model would outperform a regression that includes informative covariates but has improper baseline distribution that would be detected by examining the residuals plot.

Appendix

Out-of-sample Cox–Snell residuals plots for mortgage prepayment models:

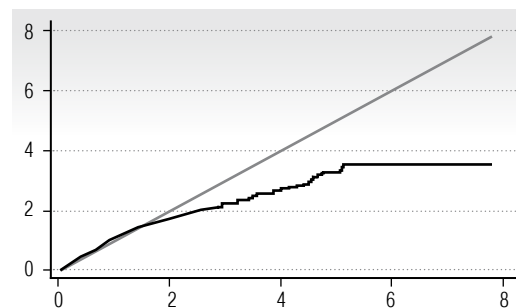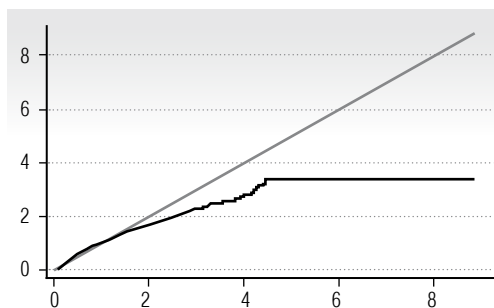| «Short» model	«Long» model

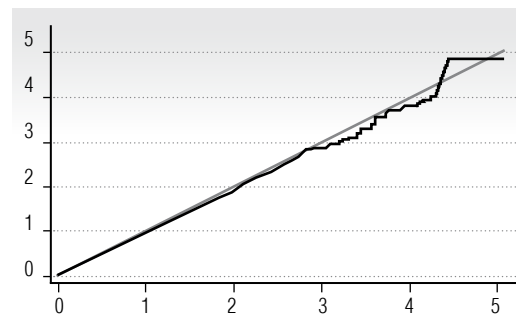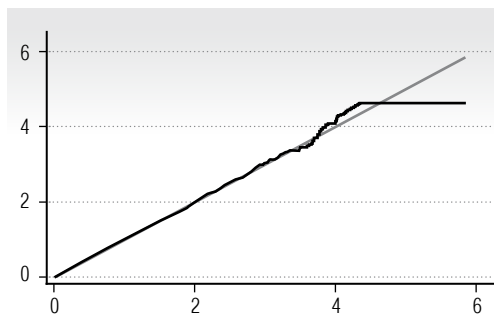**Lognormal regression**
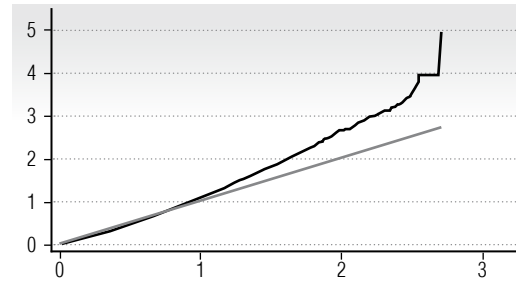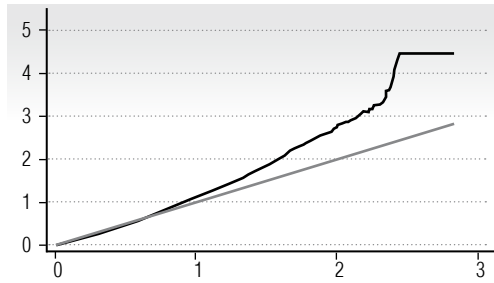
Generalized gamma regression
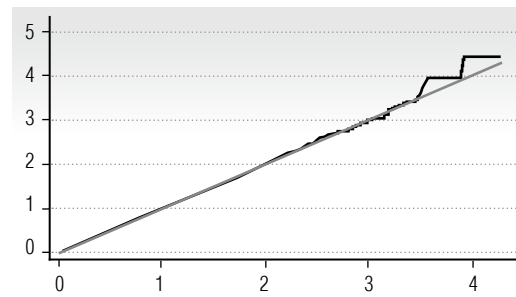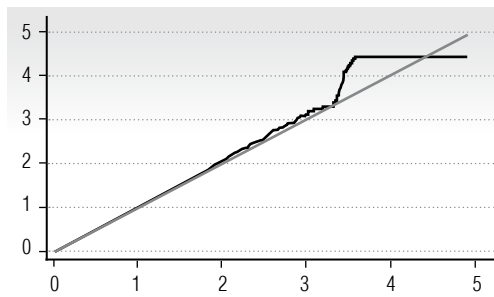
Gompertz regression

Cox model

Nelson–Aalen cumulative hazard ——— Cox–Snell residual

References

1. Klein J.P., Moeschberger M.L. (2005) *Survival analysis: Techniques for censored and truncated data*. Second edition. Springer.

2. Zhang J., Thomas L.C. (2012) Comparison of linear regression and survival analysis using single and mixture distribution approaches in modelling LGD. *International Journal of Forecasting*, vol. 28, no 1, pp. 204−215. DOI: 10.1016/j.ijforecast.2010.06.002.

3. Zamisniy P., Kozlov A. (2018) Using credit history data for estimating the probabilities of default and early termination. *Bank Crediting*, no 4. pp. 4−11 (in Russian).

4. Rakesh V., Lee W.-C., Reddy C.K. (2016) Probabilistic group recommendation model for crowdfunding domains. Proceedings of the *9th ACM International Conference on Web Search and Data Mining (WSDM 2016), San Francisco, California, USA, 22−25 February 2016*, pp. 257−266. DOI: 10.1145/2835776.2835793.

5. Ameri S., Fard M.J., Chinnam R.B., Reddy C.K. (2016) Survival analysis based framework for early prediction of student dropouts. Proceedings of the *25th ACM International Conference on Information and Knowledge Management (CIKM 2016), Indianapolis, Indiana, USA, 24−28 October 2016*, pp. 903−912. DOI: 10.1145/2983323.2983351.

6. Wang P., Li Y., Reddy C.K. (2019) Machine learning for survival analysis: A survey. *ACM Computing Surveys*, February, article no 110. DOI: 10.1145/3214306.

7. Cleves M.A., Gould W.W., Gutierrez R.G., Marchenko Y.U. (2010) *An introduction to survival analysis using Stata*. Third Edition. College Station, Texas: Stata Press.

8. Dirick L., Claeskens G., Baesens B. (2017) Time to default in credit scoring using survival analysis: a benchmark study. *Journal of Operational Research Society*, vol. 68, no 6, pp. 652−665. DOI: 10.1057/s41274-016-0128-9.

9. Harrell F.E. Jr., Califf R.M., Pryor D.B., Lee K.L., Rosati R.A. (1982) Evaluating the yield of medical tests. *Journal of the American Medical Association*, vol. 247, no 18, pp. 2543−2546. DOI: 10.1001/jama.1982.0332043004703.

10. Uno H., Cai T., Pencina M.J., D'Agostino R.B., Wei L.J. (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, vol. 30, no 10, pp. 1105−1117. DOI: 10.1002/sim.4154.

11. Newson R.B. (2010) Comparing the predictive powers of survival models using Harrell's C or Somers' D. *The Stata Journal*, vol. 10, no 3, pp. 339−358.

12. Somers R.H. (1962) A new asymmetric measure of association for ordinal variables. *American Sociological Review*, vol. 27, no 6, pp. 799−811.

13. Martinez-Romero J., Bueno-Fortes S., Martín-Merino M., de Molina A.R., De Las Rivaz J. (2018) Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling. *BMC Genomics*, no 19, article no 857. DOI: 10.1186/s12864-018-5193-9.

14. Kalinin M.N., Khasanova D.R., Ibatullin M.M. (2019) Possible timing for anticoagulation therapy initiation in ischemic stroke patients with atrial fibrillation: further analysis of the hemorrhagic transformation index. *Neurology, Neuropsychiatry, Psychosomatics*, vol. 11, no 2, pp. 12−21 (in Russian). DOI: 10.14412/2074-2711-2019-2-12-21.

15. Rumyantseva E.V., Furmanov K.K. (2016) Modeling mortgage survival. *Applied Econometrics*, vol. 41, no 1, pp. 123−143 (in Russian).

16. Hung H., Chiang C.T. (2010) Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian Journal of Statistics*, vol. 37, no 4, pp. 664−679.

17. Kamarudin A.N., Cox T., Kolamunnage-Dona R. (2017) Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, no 17, article no 53. DOI: 10.1186/s12874-017-0332-6.

18. Heagerty P.J., Zheng Y. (2005) Survival model predictive accuracy and ROC curves. *Biometrics*, vol. 61, no 1, pp. 92−105. DOI: https://doi.org/10.1111/j.0006-341X.2005.030814.x.

19. Cox D.R., Snell E.J. (1968) A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 30, no 2, pp. 248−275.

20. Nelson W. (1972) Theory and applications of hazard plotting for censored failure data. *Technometrics*, vol. 14, no 4, pp. 945−966. DOI: 10.1080/00401706.1972.10488991.

21. Aalen O. (1978) Nonparametric inference for a family of counting processes. *Annals of Statistics*, vol. 6, no 4, pp. 701−726.

22. Arzhenovsky S. (2006) Socioeconomic determinants of smoking in Russia. *Quantile*, no 1, pp. 81−100 (in Russian).

23. Rapakov G.G., Gorbunov V.A. (2015) Time-to-event analysis methods for demographics processing. *Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies*, no 4. pp. 110−120 (in Russian).

About the authors

Ekaterina V. Rumyantseva

Cand. Sci. (Phys.-Math.);

Senior Lecturer, Department of Applied Economics, National Research University Higher School of Economics,

20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: evrumyantseva@hse.ru

Kirill K. Furmanov

Cand. Sci. (Econ.);

Assistant Professor, Department of Applied Economics, National Research University Higher School of Economics,

20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: kfurmanov@hse.ru

ORCID: 0000-0002-3433-9497