DOI: 10.17323/2587-814X.2021.1.7.18

Использование вневыборочных остатков Кокса–Снелл при прогнозировании наступления событий

Е.В. Румянцева
E-mail: evrumyantseva@hse.ru

К.К. Фурманов
E-mail: kfurmanov@hse.ru

Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20

Аннотация

В статье рассматривается задача оценивания прогнозной силы модели наступления события по вневыборочным данным. Данные о времени наступления событий, как правило, цензурированны справа: ожидаемое событие часто не уступает произойти за время наблюдения, из-за чего фиксируется только минимальное возможное значение прогнозируемой величины. В результате стандартные меры точности прогноза, такие как средняя абсолютная или средняя квадратическая ошибка, оказываются неприменимыми, а для измерения качества применяются коэффициенты ранговой корреляции: C-индекс Харрелла, коэффициенты Уно и Сомерса. Эти меры не отражают близости прогнозов к действительным значениям, а характеризуют только согласованность ранжировок – способность модели отличать наблюдения, в которых ожидаемое событие происходит относительно быстро, от тех наблюдений, в которых время ожидания относительно велико, из-за чего коэффициенты ранговой корреляции могут принимать высокие значения даже при сколь угодно большой систематической ошибке прогноза. Кроме того, сведение качества прогноза к корреляции или даже близости прогнозируемого и действительного значений малоудовлетворительно: время наступления редко удается оценить с определенностью, и при прогнозировании интерес представляет не только точечная оценка момента наступления, но и оценка закона распределения объясняемой величины целиком. В настоящей статье при выборе прогнозной модели предлагается дополнять сравнение коэффициентов ранговой корреляции анализом остатков Кокса–Снелл, рассчитанных для вневыборочных данных (контрольных или валидационных). Для визуального анализа предлагается применять график оценки интегрального риска остатков, а в качестве численной характеристики согласованности модели с вневыборочными данными – расстояние Колмогорова между наблюдаемым распределением остатков и экспоненциальным распределением с единичным средним, которое соответствует идеально специфицированной модели. Предлагаемый подход иллюстрируется примером выбора прогнозной модели для времени досрочного погашения договоров ипотечного кредитования.

Ключевые слова: прогнозирование; анализ наступления событий; остатки Кокса–Снелл; цензурирование.

Цитирование: Румянцева Е.В., Фурманов К.К. Использование вневыборочных остатков Кокса–Снелл при прогнозировании наступления событий // Бизнес-информатика. 2021. Т. 15. № 1. С. 7–18.
DOI: 10.17323/2587-814X.2021.1.7.18

Введение

В ряде задач статистики и анализа данных требуется оценить время ожидания определенного события. В финансовых приложениях это может быть время наступления дефолта по договору кредитования, в медицинских – время смерти или выздоровления пациента, в социальных и демографических – возраст матери на момент рождения ребенка или вступления в брак. Область прикладной статистики, занимающаяся такими задачами, называется анализом наступления событий (event-history analysis) и имеет существенные особенности, которые препятствуют применению многих устоявшихся методов регрессионного анализа. Одна из этих особенностей – цензурирование (censoring).

Продолжительность всякого исследования конечна, и за это время ожидаемое событие может не произойти. Более того, для некоторых объектов оно вообще никогда не наступит: часть заемщиков завершит выплаты по кредиту, не доходя до дефолта, часть женщин не выйдет замуж. В результате об этих объектах известно лишь то, что время ожидания для них превышает некоторое значение – длительность от начала ожидания события до конца периода наблюдения. Таким образом, наблюдения за этими объектами оказываются цензурированными справа.

Методы оценивания статистических моделей по цензурированным данным известны достаточно хорошо, классический труд в этой области – книга [1]. Измерение точности прогноза наступления событий – менее разработанная область. По-видимому, этот пробел объясняется тем, что модели наступления разрабатывались обычно не для прогноза, а с академическими целями – для определения эффективности лечения или мер социальной политики, выявления детерминант продолжительности безработицы и т.п.

За последние десять лет интерес к прогнозированию существенно возрос. С одной стороны, модели наступления событий стали чаще находить чисто прагматическое применение: примерами могут служить анализ финансовых рисков [2, 3] и продолжительность сбора средств при краудфандинге [4]. С другой стороны, распространение машинного обучения и, в частности, процедур кросс-валидации обусловило интерес к оцениванию качества моделей по их способности делать точный вневыборочный прогноз [5, 6]. При этом стандартные метрики, такие как среднеквадратическая ошибка прогноза или средняя абсолютная процентная ошибка, оказываются неприменимыми из-за цензурирования.

В настоящей статье предлагается подход к выбору прогнозной модели, опирающийся на совместное использование коэффициентов согласованности фактического и прогнозируемого времени наступления и остатков Кокса–Снелл, рассчитываемых по контрольной выборке. В качестве примера рассматривается задача построения прогнозной модели для времени досрочного погашения ипотечного кредита.

В первом разделе описываются основные понятия, которые требуют определения (либо упоминания) из-за обособленности предметной области и специфической терминологии. Во втором разделе приводится обзор характеристик точности прогноза времени наступления события. Третий раздел посвящен остаткам Кокса–Снелл, которые авторы предлагают использовать для измерения качества прогноза. Пример использования остатков для выбора наилучшей прогнозной модели приведен в четвертом разделе, за которым следует заключение.

1. Вероятностная модель наступления события

Время ожидания события моделируется неотрицательной случайной величиной , которая по усмотрению исследователя может быть как непрерывной, так и дискретной. В настоящей статье мы рассматриваем случай непрерывного времени. Распределение времени ожидания характеризуется следующими функциями, играющими ведущую роль в анализе наступления событий.

Функция дожития (survival function) $S(t)$ задает вероятность того, что время ожидания события превысит произвольный срок t:

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

Термин соответствует медицинским и актуарным приложениям, в которых анализируемое событие – это смерть пациента или застрахованного лица.

Функция риска (hazard function) $h(t)$ отражает изменение вероятности наступления события с течением времени:

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

Интегальная функция риска (integrated hazard function) $H(t)$ не имеет столь ясной интерпретации, но играет важную роль в настоящей статье:

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

Терминология зависит от области приложения, поэтому те же характеристики можно встретить и под другими названиями. В частности, функция дожития еще известна как функция надежности, а функция риска – как сила смертности или функция опасности отказа.

Внимание исследователей, как правило, сосредоточено на двух аспектах моделирования.

Первый аспект – связь вероятности наступления события со временем его ожидания. Для описания этой связи особенно удобна функция риска.

Второй аспект – связь вероятности наступления события с объясняющими переменными (ковариатами). Существует множество способов увязать распределение времени ожидания с объясняющими переменными, изложение которых можно найти в книгах [1, 7]. Здесь мы ограничимся рассмотрением четырех регрессионных моделей.

Логнормальная и обобщенная гамма-регрессия относятся к моделям ускоренного времени (accelerated failure-time models), то есть представимы в линейном виде:

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

Здесь x' – вектор-строка объясняющих переменных, β – вектор-столбец коэффициентов при них, а ε – случайная ошибка. В вектор x' включается единичный элемент, соответствующий свободному члену регрессии, так что $x'\beta = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k$. Логнормальная и гамма-регрессия отличаются только распределением случайной ошибки.

Регрессия Гомперца и модель Кокса относятся к моделям пропорциональных рисков (proportional hazards), то есть предполагают, что функция риска наступления события пропорциональна объясняющим переменным:

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

Здесь $h_0(t)$ – так называемая опорная функция риска (baseline hazard). В модели Гомперца предполагается экспоненциальное убывание или возрастание риска с течением времени: $h_0(t) = e^{\gamma t}$, где γ – параметр модели, оцениваемый наряду с коэффициентами β. Модель Кокса не накладывает ограничений на опорную функцию риска, эта функция оценивается непараметрически. Таким образом, регрессия Кокса строго постулирует характер связи риска наступления события с объясняющими переменными, как и прочие рассматриваемые модели, но в отличие от них не ограничивает зависимость риска от времени.

Методы оценивания моделей наступления событий реализованы в статистических пакетах, их описание содержится в книгах [1, 7]. Для дальнейшего изложения важно лишь то, что на основании каждой из этих моделей можно рассчитать оценки функций дожития и интегрального риска для любых значений объясняющих переменных x' и таким образом получить прогноз наступления события в заданных условиях.

2. Меры точности прогноза: Обзор

Меры средней ошибки прогноза. Под этим названием можно объединить самые распространенные характеристики точности для моделей с количественным объясняемым признаком: среднюю абсолютную ошибку (MAE), среднюю квадратическую ошибку (RMSE), среднюю абсолютную процентную ошибку (MAPE) и другие. Есть примеры их применения и в анализе наступления событий [2, 5, 8], однако эти примеры – скорее исключение. В большинстве случаев данные подвержены цензурированию, из-за которого точная величина расхождения между временем наступления события и его прогнозом неизвестна и усреднению не поддается. Проблему можно легко решить, если предположить, что распределение расхождений относится к некоторому параметрическому семейству, тогда

параметры этого распределения и математическое ожидание расхождений можно оценить методом максимального правдоподобия. Однако примеры применения такого подхода авторам статьи неизвестны (возможно, это объясняется естественным желанием исследователей избежать дополнительных предпосылок).

В работах [2, 8] рассматривается подход, при котором средняя абсолютная ошибка рассчитывается только для нецензурированных наблюдений, содержащих точные сведения о времени наступления события. Этот подход имеет существенный недостаток. Цензурирование зависит от времени наступления: чем дольше ожидание события, тем скорее наблюдение окажется цензурированным, потому что событие не наступит в течение периода наблюдения. Тогда точное значение объясняемой переменной будет известно для тех случаев, когда оно мало, а поскольку средняя абсолютная ошибка рассчитывается именно по этим наблюдениям, то предпочтение будет отдаваться моделям, занижающим время наступления события. Формально говоря, здесь цензурирование заменяется на усечение – не менее существенный вид неполноты данных [1].

Коэффициенты согласованности ранжировок. Пожалуй, самая распространенная мера качества прогноза при анализе наступления событий – коэффициент конкордации Харрелла [9], он же *C*-индекс, который можно определить следующим образом. Пусть случайные величины T_1 и T_2 отражают фактическое время наступления события в двух случайно отобранных независимых наблюдениях, а $\hat{T}_1$ и $\hat{T}_2$ представляют собой соответствующие прогнозы. Тогда коэффициент Харрелла *C* определяется выражением:

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

Одно из достоинств коэффициента – его ясная интерпретация. Пусть рассматриваются два случая с различным временем наступления события. Тогда модель с вероятностью *C* прогнозирует более быстрое наступления события именно в том случае, когда оно и в самом деле наступило быстрее. Наибольшее значение *C* равно единице и наблюдается при полном соответствии фактических и прогнозируемых ранжировок по времени (чем быстрее событие наступает согласно модели, тем меньше и действительное время наступления), наименьшее значение равно нулю и означает полное расхождение ранжировок (чем меньше предсказание модели, тем больше истинное время).

Существуют различные оценки коэффициента конкордации, учитывающие цензурирование. Помимо оригинальной статистики Харрелла [9], в последнее время стал широко применяться коэффициент Уно [10], оценивающий тот же параметр. Кроме коэффициента конкордации (1) также используется коэффициент корреляции Сомерса [11, 12].

Общее слабое место перечисленных характеристик – то, что они отражают лишь согласованность ранжировок или, иначе говоря, способность модели выявлять те случаи, в которых событие наступает относительно быстрее или медленнее. Это не есть точность прогноза в обычном понимании. Представим, что предсказанное моделью время в каждом наблюдении оказывается ровно в десять раз больше истинного. Вряд ли можно назвать десятикратно завышенный прогноз точным, однако коэффициент конкордации (и любой коэффициент корреляции рангов) будет равен единице, потому что ранжировки прогнозов и истинных значений совпадут.

Иногда именно способность модели выделять случаи относительно короткого или длительного ожидания оказывается в фокусе исследования, например, если речь идет об определении признаков скорого выздоровления или смерти [13]. Однако во многих задачах представляет интерес именно абсолютное время наступления события. Такие задачи возникают и в традиционной области применения – медицине [14], но еще более характерны для финансовых приложений, где от дефолтов и досрочных погашений зависит поток денежных средств [2, 3, 15].

Еще одна группа характеристик прогнозной силы – ***показатели классифицирующей способности***, измеряющие точность бинарного прогноза (событие наступило или не наступило к определенному сроку). Это направление стало активно развиваться в последнее десятилетие [16, 17], оно заслуживает упоминания, но подробно рассматриваться в настоящей статье не будет, так как представляет собой содержательно иной подход к задаче прогнозирования. Впрочем, коэффициент Харрелла также может рассматриваться как характеристика классифицирующей способности [18].

3. Остатки Кокса–Снелл и их применение при изучении прогнозной способности

Пусть имеется выборка $(T_1, x_1'),...,(T_n, x_n')$, где T_i – время наступления события, а x_i' – вектор объясняющих переменных в наблюдении *i*. Пусть $\hat{H}(t; x')$ – оцененная функция интегрального риска случайных величин T_i (ее можно получить из какой-либо регрессионной модели). Остатком Кокса–Снелл [19] в наблюдении *i* называется величина $r_i^{CS} = \hat{H}(T_i; x_i')$.

Если оценка $\hat{H}(t; x')$ совпадает с истинной интегральной функцией риска $H(t; x')$, то остатки Кокса–Снелл распределены по экспоненциальному закону со средним 1. В этом случае интегральная функция риска для остатков имеет вид $H^{CS}(t) = t$.

В регрессионной диагностике применяется визуальный тест, который опирается на остатки Кокса–Снелл и состоит из следующих шагов.

1. Оценивается регрессионная модель и рассчитываются остатки Кокса–Снелл для каждого наблюдения.

2. Рассчитывается оценка интегральной функции риска для остатков $\hat{H}^{CS}(t)$. При цензурировании данных T_i о времени наступления событий остатки тоже оказываются цензурированными, что требует соответствующего метода оценивания. Мы используем метод Нельсона–Аалена [20, 21].

3. Строится график зависимости оцененного интегрального риска $\hat{H}^{CS}(r_i^{CS})$ от остатков r_i^{CS}, в дальнейшем будем называть его графиком остатков Кокса–Снелл. В случае верно специфицированной модели наступления событий график выстраивается вдоль линии $H(t) = t$ (*рисунок 1а*). Пример графика для неверно подобранной модели представлен на *рисунке 1б*.

Этот тест известен и отражен в учебной литературе, соответствующие графики приводятся исследователями в статьях в подтверждение корректности предлагаемых моделей [7, 22, 23]. Мы не встречали примеров использования остатков Кокса–Снелл при изучении прогнозной способности моделей, возможные причины этого пробела указаны ниже в заключении. Далее в настоящей статье остатки Кокса–Снелл, рассчитанные по контрольной выборке, называются вневыборочными остатками в том смысле, что они характеризуют поведение прогнозной модели вне той выборки, по которой она оценивалась.

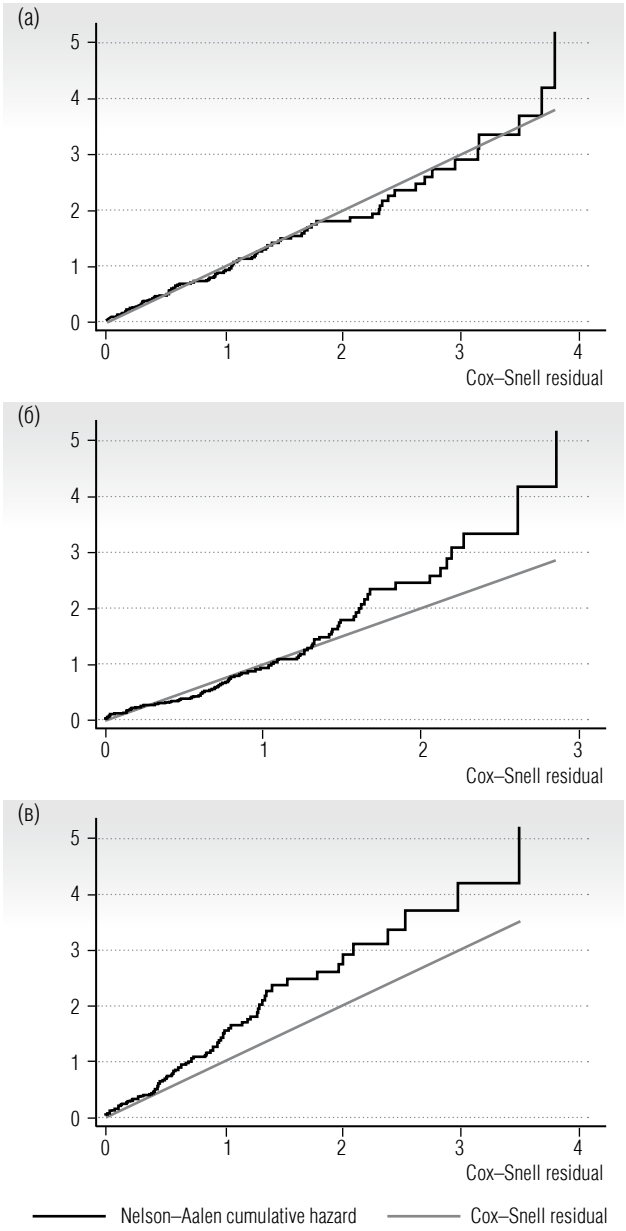


Рис. 1. Примеры графиков остатков Кокса–Снелл при (а) верной спецификации оцениваемой модели; (б) ошибочно выбранном распределении; (в) систематической ошибке прогноза

Расчет вневыборочных остатков позволяет выявить систематическую ошибку прогноза. Представим себе случай, когда модельная функция риска оказывается в *c* раз выше истинной: $h(t; x') = c\, h(t; x')$. Тогда остатки Кокса–Снелл будут завышать истинную интегральную функцию риска также в *c* раз: $r_i^{CS} = \hat{H}(T_i; x_i') = cH(T_i; x_i')$. В результате график остатков проходит выше или ниже (в зависимости от *c*), чем линия $H(t) = t$ (*рисунок 1в*). Такая ситуация практически невозможна при диагностике регрес-

сии по обучающей выборке и маловероятна даже при изучении контрольной выборки, если наблюдения для контроля отбирались случайно. Такое смещение прогноза скорее можно наблюдать при внешней валидации, когда прогнозная сила модели проверяется на новых данных.

Визуальный анализ предпочтителен при выборе модели «вручную», для задач машинного обучения удобнее иметь числовую характеристику. Ниже в качестве такой характеристики используется расстояние Колмогорова между оценкой функции дожития для остатков Кокса–Снелл $S^{CS}(t) = \exp(-\hat{H}^{CS}(t))$ и соответствующей функцией для экспоненциального распределения $S(t) = e^{-t}$:

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

Здесь $t_{\max}$ – наибольшее время ожидания события в выборке (может соответствовать и цензурированному наблюдению). Мы рассматриваем только множество $[0; t_{\max}]$, так как оценка Нельсона–Аалена не позволяет восстановить правый хвост распределения остатков Кокса–Снелл. Соответствующая функция дожития в точке $t_{\max}$ все еще отлична от нуля, и нет данных, позволяющих оценить ее для значений аргумента $t > t_{\max}$.

В следующем параграфе приводится пример применения предлагаемой метрики при построении прогнозной модели.

4. Пример: Моделирование досрочного погашения кредита

Пример опирается на данные крупной компании, занимающейся ипотечным кредитованием. Данные содержат сведения о более чем 280 тысячах договоров, заключенных в период 2001–2013 гг. Объясняемая переменная – время от заключения договора до досрочного погашения кредита. Наблюдения за объясняемой переменной цензурированы справа по следующим причинам:

✦ Прекращение периода наблюдения: пример опирается на срез данных на 1 января 2014 года, поэтому точная дата погашения действовавших на тот момент договоров не известна.

✦ Завершение срока договора: если досрочного погашения в действительности не произошло, то наблюдение за временем погашения считается цензурированным, как если бы было возможно досрочное погашение после истечения срока. Это своего рода математическая уловка, удобная при построении модели и применяемая в случаях, когда с изучаемыми объектами могут случиться взаимоисключающие события разного рода (в данном случае – досрочное погашение и истечение срока). Удобно считать, что происходят оба события, но наблюдается только первое из них.

✦ Наступление дефолта по договору. Это еще одно событие, исключающее возможность досрочного погашения.

По этим данным оцениваются модели наступления события, отличающиеся (а) предпосылками о законе распределения объясняемой величины и его связи с объясняющими переменными (логнормальная и гамма-регрессия, модели Кокса и Гомперца) и (б) набором объясняющих переменных («короткая» и «длинная» модели). «Короткая» и «длинная» модели учитывают характеристики самого кредита, основного заемщика, а также предмета ипотеки, находящегося под залогом. «Короткая» модель учитывает кредитную ставку, срок кредита, коэффициент «платеж/доход», возраст основного заемщика, тип занятости основного заемщика и количество комнат в предмете ипотеки. «Длинная» модель включает, помимо вышеперечисленных, следующие признаки: пол, семейное положение и уровень образования основного заемщика, число созаемщиков, местонахождение предмета ипотеки (положение региона в рейтинге социально-экономического положения субъектов РФ, используемом в компании, согласно которому все регионы поделены на три группы: с низким, умеренным и высоким уровнем развития), тип жилья под залогом (дом, таунхаус или квартира), отношение жилой и общей площади в предмете ипотеки, отношение совокупных денежных выплат за плановый срок жизни кредита к стоимости жилья и коэффициент «кредит/залог».

Все наблюдения случайным образом разбиты на обучающую и контрольную выборки в соотношении 60:40 соответственно.

Графики остатков Кокса–Снелл для восьми оцененных моделей для контрольной части данных приведены в Приложении. Рассчитанные по формуле (2) расстояния между наблюдаемым и теоретическим распределениями остатков, а также значения коэффициента Харрелла приведены в *таблице 1.*

Из таблицы видно, что коэффициент Харрелла практически не зависит от выбора распределения времени досрочного погашения, но меняется при добавлении в модель объясняющих переменных: «длинная» модель оказывается стабильно лучше, обе-

Таблица 1.

Характеристики точности вневыборочного прогноза моделей наступления досрочного погашения кредита

	«Короткая» модель		«Длинная» модель	
	Коэффициент Харрелла	Расстояние Колмогорова	Коэффициент Харрелла	Расстояние Колмогорова
Логнормальная	0,593	0,078	0,612	0,066
Гамма	0,593	0,015	0,610	0,009
Гомперца	0,592	0,059	0,608	0,063
Кокса	0,593	0,007	0,609	0,014

спечивая вероятность согласия прогнозов и ранжировок примерно на 0,02 больше, чем у «короткой». Конечно, исследователь может считать это расхождение несущественным, но оно устойчиво и проявляется и при прочих вариантах разбиения данных на обучающую и контрольную части. Напротив, расстояние Колмогорова между идеальным и наблюдаемым распределениями остатков Кокса–Снелл заметно меняется в зависимости от распределения: обобщенная гамма-регрессия и регрессия Кокса обеспечивают меньшее расхождение наблюдаемого и предполагаемого распределений, по сравнению с логнормальной регрессией и регрессией Гомперца, причем как в случае «короткой», так и в случае «длинной» модели.

Расхождения в прогнозах, полученных при разных предпосылках о виде распределения, представлены на *рисунке 2*. Здесь представлены оценки функций дожития для кредита с часто встречающимся набором характеристик:

✦ пол основного заемщика – мужчина;

✦ семейный статус основного заемщика – женат;

✦ образование основного заемщика – высшее образование или ученая степень;

✦ тип занятости основного заемщика – наемный работник;

✦ количество созаемщиков, включая основного заемщика – 2;

✦ возраст основного заемщика – менее 35 лет;

✦ коэффициент «платеж/доход» – от 20% до 35%;

✦ тип предмета ипотеки – квартира;

✦ местонахождение предмета ипотеки – умеренно развитый регион;

✦ количество комнат в предмете ипотеки – 2;

✦ годовая кредитная ставка – менее 11,5% (низкорисковая кредитная ставка);

✦ соотношение жилой и общей площади в предмете ипотеки – между 50% и 70%;

✦ коэффициент «кредит/залог» – от 50% до 70%;

✦ срок кредита – свыше 180 месяцев (долгосрочный кредит);

✦ отношение совокупных денежных выплат за плановый срок жизни кредита к стоимости жилья – от 1 до 1,82.

На *рисунке 2а* отражена вероятность непогашения в течение ближайших лет для только что выданного кредита, на *рисунке 2б* – для кредита, выданного пять лет назад (условная функция дожития при условии $\{T > 5\}$). По горизонтальной оси отложено число дней с момента выдачи кредита, по вертикальной – вероятность дожития. На обоих графиках оценки, полученные из модели Кокса (линия «S_cox») и гамма-регрессии («S_gamma»), практически совпадают. Линия, соответствующая регрессии Гомперца («S_gomp»), отклоняется от прочих с самого начала, предсказывая больший риск погашения. Логнормальная регрессия (линия «S_ln»), наоборот, предсказывает наименьший риск погашения, причем существенное расхождение с моделями Кокса и гамма начинается не сразу, оно заметно только для «пожилого» кредита. Графики вневыборочных остатков, приведенные в Приложении, подтверждают, что гамма-регрессия верно описывает распределение времени погашения, модель Кокса ей почти не уступает, а вот остатки регрессии Гомперца, как и остатки логнормальной модели, отклоняются от идеального экспоненциального распределения, причем в последнем случае отклонение заметно именно на правом хвосте распределения («пожилые» кредиты).

Таким образом, если при выборе прогнозной модели опираться на коэффициент Харрелла, то будет выбрана логнормальная регрессия, недооценивающая риск досрочного погашения. Принимая в расчет вневыборочные остатки Кокса–Снелл, аналитик скорее сделает выбор в пользу обобщенной

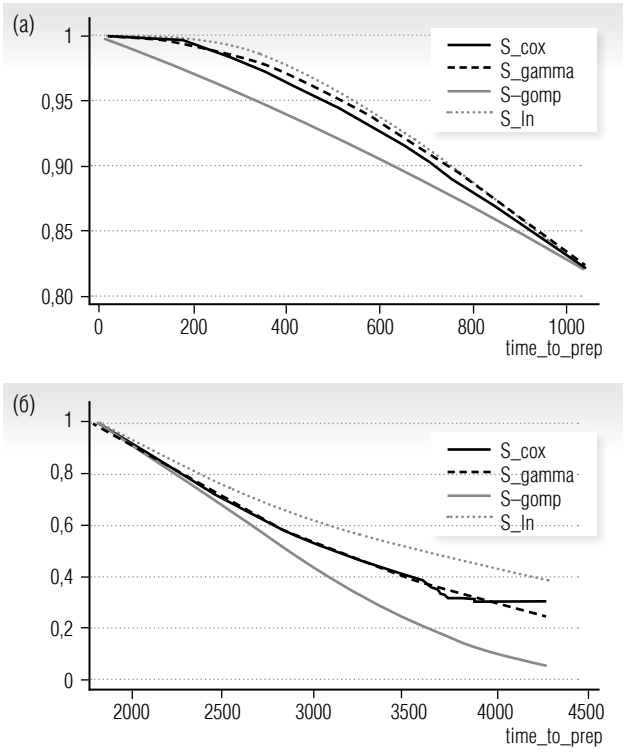


Рис. 2. Оценки функции дожития для
(а) только что выданного кредита;
(б) пятилетнего кредита

гамма-регрессии или модели Кокса, которые дают близкие прогнозы, хорошо согласующиеся с распределением времени погашения в контрольной выборке. Как видно из *рисунка 2*, оцененная вероятность погашения заметно зависит от выбора прогнозной модели. Это особенно заметно на дальнем горизонте прогнозирования (линии на *рисунке 2б* расходятся более чем на 20 процентных пунктов), но уже при расчете на один–два года вперед расхождение прогнозов превышает 5 процентных пунктов.

Заключение

Остатки Кокса–Снелл и связанный с ними визуальный тест хорошо известны и освещены не только в исследовательской, но и в учебной литературе, однако широкого применения при анализе прогнозной силы на контрольной выборке не нашли. Это важный момент, и он вызывает естественное сомнение: действительно ли от них есть польза? Приведенный пример был призван показать, что польза есть, сейчас же время сделать важную оговорку.

Сами по себе остатки Кокса–Снелл не помогают отличить точный прогноз от неточного. Они лишь показывают, насколько прогнозируемое распределение близко к тому, что наблюдается в выборке (обучающей – при обычном использовании, контрольной – если использовать так, как предлагается в настоящей статье). Модель вообще без объясняющих переменных, которая дает одинаковый прогноз во всех наблюдениях, вполне может оказаться лучше по остаткам, чем регрессия, учитывающая важные и действительно улучшающие прогноз факторы, если в первой удачнее выбрать семейство распределений (что, кстати, без объясняющих переменных сделать проще).

Коэффициенты ранговой корреляции/конкордации, наоборот, позволяют отобрать модель с удачным набором объясняющих переменных, но никак не характеризуют способность предсказать в целом распределение времени наступления события. В классических задачах регрессионного анализа этого часто и не требуется, аналитика интересует в первую очередь точечный прогноз – оценка математического ожидания или медианы. Однако время наступления событий почти никогда не удается точно охарактеризовать одним числом, и сведение прогнозирования к оценке единственного параметра непродуктивно. Здесь не меньший интерес, чем математическое ожидание, представляют квантили различных порядков, то есть, по сути, распределение целиком. Выбору распределения часто не уделяется внимание: исследователи предпочитают опираться на модель Кокса, не требующую параметризации распределения. Приведенный в настоящей статье пример показывает, что даже на выборке весьма большого объема относительно простая параметрическая регрессия может превзойти модель Кокса, склонную к переобучению. Кроме того, модель Кокса ограничивает функциональную форму связи риска наступления события с объясняющими переменными, параметрические методы позволяют ослабить это ограничение.

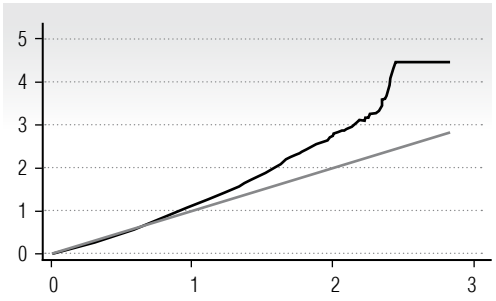
Комбинация коэффициента конкордации и остатков Кокса–Снелл позволяет оценить качество прогнозной модели и в аспекте выбора объясняющих переменных, и в аспекте выбора семейства распределений. Конечно, удобнее было бы иметь единственную характеристику, однозначно указывающую на модель с лучшей прогнозной силой, но о такой характеристике пока, по-видимому, говорить не приходится. ■

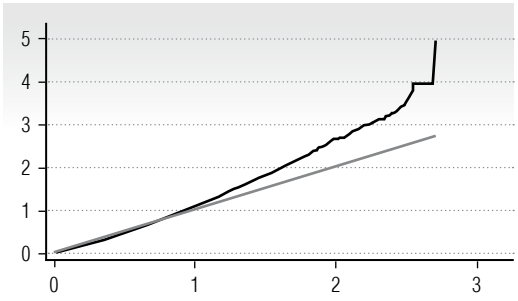
Приложение

Графики вневыборочных остатков Кокса–Снелл для моделей досрочного погашения ипотечного кредита:

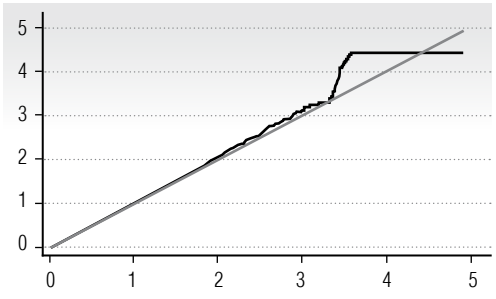
«Короткая» модель	«Длинная» модель

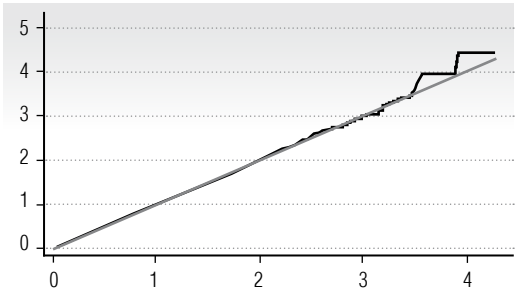
Логнормальная регрессия



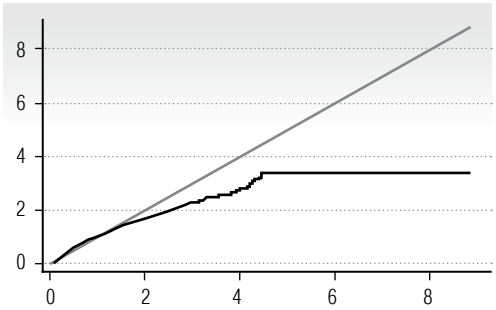


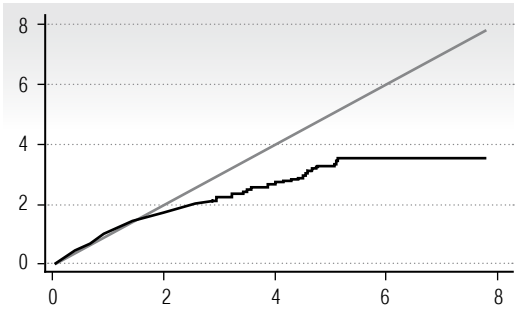
Гамма-регрессия



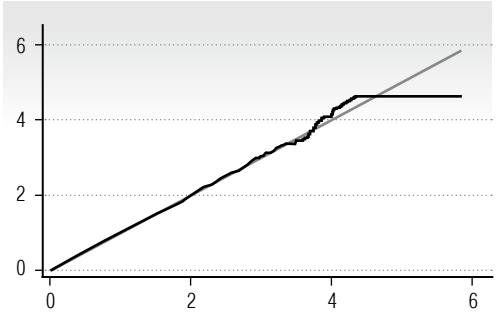


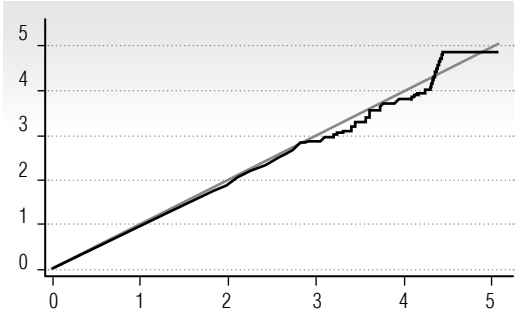
Регрессия Гомперца





Регрессия Кокса





Nelson–Aalen cumulative hazard Cox–Snell residual

Литература

1. Klein J.P., Moeschberger M.L. Survival analysis: Techniques for censored and truncated data. Second edition. Springer. 2005.
2. Zhang J., Thomas L.C. Comparison of linear regression and survival analysis using single and mixture distribution approaches in modelling LGD // International Journal of Forecasting. 2012. Vol. 28. No 1. P. 204–215. DOI: 10.1016/j.ijforecast.2010.06.002.
3. Замисный П., Козлов А. Новый подход к использованию данных БКИ при оценке вероятности дефолта и досрочного погашения // Банковское кредитование. 2018. №4. С. 4–11.
4. Rakesh V., Lee W.-C., Reddy C.K. 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. P. 257–266. DOI: 10.1145/2835776.2835793.
5. Ameri S., Fard M.J., Chinnam R.B., Reddy C.K. 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. P. 903–912. DOI: 10.1145/2983323.2983351.
6. Wang P., Li Y., Reddy C.K. Machine learning for survival analysis: A survey // ACM Computing Surveys. February 2019. Article no 110. DOI: 10.1145/3214306.
7. Cleves M.A., Gould W.W., Gutierrez R.G., Marchenko Y.U. An introduction to survival analysis using Stata. Third Edition. College Station, Texas: Stata Press, 2010.
8. Dirick L., Claeskens G., Baesens B. Time to default in credit scoring using survival analysis: a benchmark study // Journal of Operational Research Society. 2017. Vol. 68. No 6. P. 652–665. DOI: 10.1057/s41274-016-0128-9.
9. Evaluating the yield of medical tests / F.E. Harrell Jr. [et al.] // Journal of the American Medical Association. 1982. Vol. 247. No 18. P. 2543–2546. DOI: 10.1001/jama.1982.0332043004703.
10. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data / H. Uno [et al.] // Statistics in Medicine. 2011. Vol. 30. No 10. P. 1105–1117. DOI: 10.1002/sim.4154.
11. Newson R.B. Comparing the predictive powers of survival models using Harrell's C or Somers' D // The Stata Journal. 2010. Vol. 10. No 3. P. 339–358.
12. Somers R.H. A new asymmetric measure of association for ordinal variables // American Sociological Review. 1962. Vol. 27. No 6. P. 799–811.
13. Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling / J. Martinez-Romero [et al.] // BMC Genomics. 2018. No 19. Article no 857. DOI: 10.1186/s12864-018-5193-9.
14. Калинин М.Н., Хасанова Д.Р., Ибатуллин М.М. Возможные сроки начала антикоагулянтной терапии у больных с ишемическим инсультом и фибрилляцией предсердий: последующий анализ индекса геморрагической трансформации (Hemorrhagic Transformation Index) // Неврология, нейропсихиатрия, психосоматика. 2019. Т. 11. № 2. С. 12–21. DOI: 10.14412/2074-2711-2019-2-12-21.
15. Румянцева Е.В., Фурманов К.К. Моделирование времени жизни ипотечного кредита // Прикладная эконометрика. 2016. Т. 41. № 1. С. 123–143.
16. Hung H., Chiang C.T. Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data // Scandinavian Journal of Statistics. 2010. Vol. 37. No 4. P. 664–679.
17. Kamarudin A.N., Cox T., Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications // BMC Medical Research Methodology. 2017. No 17. Article no 53. DOI: 10.1186/s12874-017-0332-6.
18. Heagerty P.J., Zheng Y. Survival model predictive accuracy and ROC curves // Biometrics. 2005. Vol. 61. No 1. P. 92–105. DOI: https://doi.org/10.1111/j.0006-341X.2005.030814.x.
19. Cox D.R., Snell E.J. A general definition of residuals // Journal of the Royal Statistical Society. Series B (Methodological). 1968. Vol. 30. No 2. P. 248–275.
20. Nelson W. Theory and applications of hazard plotting for censored failure data // Technometrics. 1972. Vol. 14. No 4. P. 945–966. DOI: 10.1080/00401706.1972.10488991.
21. Aalen O. Nonparametric inference for a family of counting processes // Annals of Statistics. 1978. Vol. 6. No 4. P. 701–726.
22. Арженовский С. Социально-экономические детерминанты курения в России // Квантиль. 2006. № 1. С. 81–100.
23. Рапаков Г.Г., Горбунов В.А. Исследование методов анализа времени до события при обработке демографических данных // Вестник ВГУ. Серия: Системный анализ и информационные технологии. 2015. № 4. С. 110–120.

Об авторах

Румянцева Екатерина Владимировна

кандидат физико-математических наук;

старший преподаватель департамента прикладной экономики, факультет экономических наук, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: evrumyantseva@hse.ru

Фурманов Кирилл Константинович
кандидат экономических наук;
доцент департамента прикладной экономики, факультет экономических наук, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;
E-mail: kfurmanov@hse.ru
ORCID: 0000-0002-3433-9497

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

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

Kirill K. Furmanov
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.

Key words: forecasting; event-history analysis; Cox–Snell residuals; censoring.

Citation: Rumyantseva E.V., Furmanov K.K. (2021) Using out-of-sample Cox–Snell residuals in time-to-event forecasting. *Business Informatics,* vol. 15, no 1, pp. 7–18. DOI: 10.17323/2587-814X.2021.1.7.18

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