Горяинов А.В., Замышляев А.М., Платонов Е.Н.

АНАЛИЗ ВЛИЯНИЯ ФАКТОРОВ НА УЩЕРБ ОТ ПРОИСШЕСТВИЙ НА ТРАНСПОРТЕ С ПОМОЩЬЮ РЕГРЕССИОННЫХ МОДЕЛЕЙ[1]

Статья посвящена анализу происшествий на транспорте. Для анализа влияния различных факторов на безопасность движения предлагается использовать методы регрессионного анализа. Предлагается использовать модель зависимости ущерба, нанесённого транспортным происшествием, от набора факторов, влияющих на его возникновение. Для построения модели используется нелинейная регрессионная модель с точкой разрыва. Модель строится с целью получения данных о фактическом или прогнозируемом уровне безопасности движения поездов. Эти данные необходимы для оценки достаточности мероприятий, направленных на обеспечение нормативного уровня безопасности, для минимизации ресурсов, выделяемых на решение задач безопасного движения поездов, в том числе для обоснования приоритетов при распределении ресурсов. В работе рассмотрено два примера. В первом примере приведены результаты расчетов по реальным данным о происшествиях на железнодорожном транспорте США. Второй пример построен на основе результатов численного моделировании величины ущерба и факторов.

Ключевые слова: *железнодорожный транспорт, безопасность движения, регрессионный анализ.*

Введение

Известно, что обеспечение безопасности движения является одним из приоритетных направлений функционирования железнодорожного транспорта. Федеральный Закон «О железнодорожном транспорте в Российской Федерации» определяет безопасность движения, как состояние стабильности перевозочного процесса, при котором отсутствует недопустимый риск возникновения транспортных происшествий и их последствий, влекущий за собой причинение вреда жизни и здоровью граждан, окружающей среде, имуществу физических и юридических лиц.

В работе [1] предложена методика оценки уровня опасности возникновения транспортных происшествий на основе анализа их связи с факторами, влияющими на их возникновение. Различные методы анализа безопасности на железнодорожном транспорте представлены в [2–7]. При анализе безопасности на железнодорожном транспорте кроме вероятности появления самого транспортного происшествия следует учитывать и величину ущерба, который нанесен этим происшествием. Для анализа ущерба предлагается использовать методы регрессионного анализа [8, 9].

[1] Работа выполнена при финансовой поддержке РФФИ (грант 12-07-13114-офи_м_РЖД)

1. Анализ ущерба от транспортных происшествий

О необходимости учета величины ущерба можно судить из гистограммы ущерба (см. рис. 1), нанесенного задержками поездов на железнодорожном транспорте США в 2010 году [10].

Видно, что ущерб очень сильно различается от происшествия к происшествию, поэтому учет транспортных происшествий, как равноправных наблюдений может привести к тому, что мы будем принимать меры, направленные на снижение числа происшествий, ущерб от которых невелик. Это произойдет из-за того, что число происшествий с небольшим ущербом на порядок больше, чем происшествий с большим ущербом. При этом суммарный ущерб от происшествий с небольшим ущербом, наоборот, в несколько раз меньше. Аналогичная картина наблюдается и по другим видам транспортных происшествий.

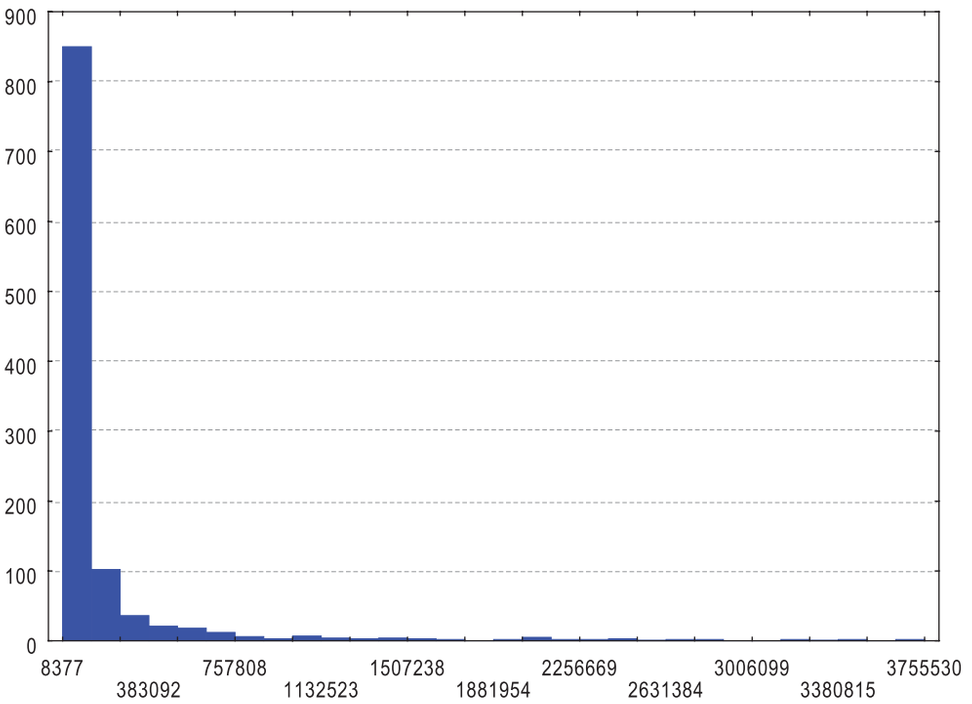

Рис. 1. Гистограмма ущерба от задержек поездов

Величину ущерба можно учесть, если построить регрессионную зависимость от набора факторов. Такие модели широко используются при анализе безопасности на транспорте [3–7].

Пусть X – скалярная переменная, описывающая величину ущерба, нанесенного транспортным происшествием некоторого вида. Это может быть ущерб от крушения поезда, бокового столкновения поездов, задержки прибытия поезда и т.д.

Мы предполагаем, что X зависит от m-мерного вектора F некоторых факторов в том смысле, что изменения вектора F вызывают вполне определенные изменения величины X.

Последнее предположение математически можно представить в виде функциональной зависимости

$$X = f(F), \quad X \in R^1, F \in R^m, \tag{1}$$

где $f(F)$ – некоторая числовая функция, зависящая от m переменных.

Одной из важнейших задач является решение проблемы определения функции $f(F)$, описывающей связь между переменной X и переменными F по результатам наблюдений за X и F.

Предположим, что у нас имеется возможность получить совместные наблюдения $\{X_k, F_k\}$, $k = 1,\ldots,n$ за X и F в n опытах.

С учетом введенной выше связи между X и F, связь между результатами отдельных наблюдений можно представить в виде

$$X_k = f(F_k) + \varepsilon_k, \quad k = 1,\ldots,n,$$

где ε_k – случайная ошибка (шум) k-го наблюдения.

Если нам удастся построить хорошее приближение к неизвестной функции $f(F)$, то мы получим возможность по значению наблюдаемого или предполагаемого вектора факторов F анализировать возможное значение ущерба от происшествия. Мы сможем понять, какие из факторов оказывают наибольшее влияние на величину ущерба и, соответственно, сосредоточиться на борьбе с этими факторами.

Естественно, без каких-либо предварительных договоренностей о возможном виде функции $f(F)$задача ее восстановления по результатам наблюдений, число которых конечно, представляется практически неразрешимой. Обычно предполагают, что функция $f(F)$ является линейной по переменным $F_1,\ldots F_m$ [8,9].

Рассмотрим линейную функцию $f(F)$ и соответственно линейную модель регрессии:

$$f(\theta, F) = \theta_0 + \theta_1 F_1 + \theta_2 F_2 + \cdots + \theta_m F_m,$$

где $\theta = [\theta_0, \theta_1, \ldots, \theta_m]^T$ – вектор неизвестных параметров.

2. Модель регрессии с точкой разрыва

В качестве альтернативы линейной функции рассмотрим кусочно-линейную модель или модель регрессии с точкой разрыва [9, 11, 12]:

$$f(\alpha, \beta, b, F) = \begin{cases} \alpha_0 + \alpha_1 F_1 + \alpha_2 F_2 + \cdots + \alpha_m F_m, & \text{если } f(\alpha, \beta, b, F) \le b, \\ \beta_0 + \beta_1 F_1 + \beta_2 F_2 + \cdots + \beta_m F_m, & \text{если } f(\alpha, \beta, b, F) > b, \end{cases} \qquad (2)$$

где $\alpha = [\alpha_0, \alpha_1, \ldots \alpha_m]^T, \beta = [\beta_0, \beta_1, \ldots \beta_m]^T$ – неизвестные параметры модели, b – неизвестная точка разрыва функции $f(\alpha, \beta, b, F)$.

Оценки неизвестных параметров будем искать как решение задач оптимизации с квадратичной функции потерь:

$$\hat{\theta} \in \arg\min_{\theta} \sum_{k=1}^{n} \left(X_k - f(\theta, F_k)\right)^2 \quad u \quad \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{b} \end{bmatrix} \in \arg\min_{\alpha,\beta,b} \sum_{k=1}^{n} \left(X_k - f(\alpha, \beta, b, F_k)\right)^2. \tag{3}$$

Оценка $\hat{\theta}$ называется оценкой метода наименьших квадратов и находится аналитически [8]. Для построения оценок в кусочно-линейной модели используются итерационные численные методы, например, метод квази – Ньютона [8, 9].

В качестве примера применения указанных моделей рассмотрим задачу восстановления зависимости ущерба от задержек поездов по данным за 2010 год для железнодорожного транспорта США [13].

Рассмотрим следующий набор факторов:

F_1 – температура по Фаренгейту, F_2 – освещенность, F_3 – погодные условия, F_4 – скорость движения поезда в момент транспортного происшествия, F_5 – тоннаж поезда.

Всего имеется 1546 наблюдений.

Линейная модель регрессии оказывается неадекватной и плохо соответствует наблюдаемым значениям. Для нее коэффициент детерминации $R^2 \in [0;1]$ оказывается равным всего лишь 0,612, а значение квадратичной функции потерь (3) в 4,05 раза больше, чем для нелинейной модели. Коэффициент детерминации показывает, в какой степени рассматриваемая модель «лучше объясняет» величину X, чем тривиальная модель (т.е. X не зависит от объясняющих переменных и $\theta = \theta_0$). Чем ближе R^2 к 1, тем большее превосходство имеет наша регрессия над тривиальной регрессией. Наоборот, если R^2 близок к нулю, то это означает, что модель линейной регрессии плохо описывает объясняемую переменную X.

Приведем результаты оценивания для кусочно-линейной регрессии. После применения метода квази-Ньютона [11,12] получены следующие результаты:

$$\hat{b} = 1006970, \quad \hat{\alpha} = \begin{bmatrix} 5080{,}6 \\ -4{,}236 \\ 1374{,}9 \\ 2963{,}0 \\ 5814{,}6 \\ 5{,}428 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} -466770 \\ -600{,}2 \\ 361254 \\ -48123 \\ 33206 \\ 26{,}95 \end{bmatrix}. \tag{4}$$

Точность оценивания можно охарактеризовать величиной объясненной дисперсии $D \in [0;1]$ и коэффициентом детерминации R^2 – чем ближе эти величины к единице, тем точнее подобранная функция f(F) объясняет наблюдения.

Для нашей модели эти характеристики равны $D = 0{,}845,\ R^2 = 0{,}919.$

Видно, что оценки параметров регрессии значительно различаются для происшествий с ущербом менее миллиона долларов и более миллиона.

К сожалению, имеющаяся статистика не содержит подробной информации о значении таких факторов, как техническое состояние поезда, дорожного полотна и сооружений, квалификация персонала и т.д. Поэтому построить адекватную регрессионную модель влияния факторов на размер ущерба от задержки поезда для железнодорожной сети США не представляется возможным.

3. Численное моделирование факторов

Для того, чтобы преодолеть недостатки имеющейся реальной статистики, проведём численное моделирование факторов, влияющих на возникновение транспортных происшествий, а также величины ущерба от этих происшествий. Моделирование можно рассматривать как основу для будущего мониторинга за факторами. После проведения мониторинга уточняются модели факторов, затем с использованием этих моделей строится модель для ущерба, которая в дальнейшем может быть использована в двух направлениях.

«Историческое» моделирование проводится, когда число реальных наблюдений недостаточно для построения адекватной модели функционирования системы и построенная модель может быть использована для увеличения объёма выборки.

Построение прогнозных значений ущерба для текущих значений факторов и прогнозных значений для теоретических значений факторов позволит проанализировать общую ситуацию с безопасностью и выработать рекомендации по наиболее проблемным факторам, влияющим на уровень безопасности движения.

Факторы, которые могут привести к транспортному происшествию, моделируются как независимые друг от друга случайные функции времени. Для каждого фактора указывается критический уровень, при превышении которого фактор начинает представлять существенную опасность. Случайные длины промежутков времени, в течение которых его значения выше (ниже) критического уровня (т.е. время ожидания пересечения фактором критического уровня), имеют для каждого из факторов экспоненциальное распределение со своим параметром. В случае, если в течение какого-то промежутка времени значения одного фактора или нескольких факторов одновременно превышают критический уровень, то считается, что это может привести к транспортному происшествию с некоторой вероятностью. Для каждого такого промежутка времени моделируется время ожидания происшествия – также случайная величина с экспоненциальным законом распределения. Параметр распределения зависит от того, какие именно факторы превысили критическое значение, иначе говоря, для каждой комбинации факторов имеется своя интенсивность появления происшествий. Происшествие считается произошедшим в том случае, когда время ожидания оказалось короче рассматриваемого промежутка (т.е. случайная величина попала в отрезок, на котором значения рассматриваемых факторов превышали критические).

Для упрощения моделирования факторов были предложены три модели:

1) Кусочно-линейная «зигзагообразная» функция. Локальные максимумы – середины промежутков времени, на которых значение фактора выше критического, локальные минимумы – середины промежутков времени, на которых значение фактора ниже критического.

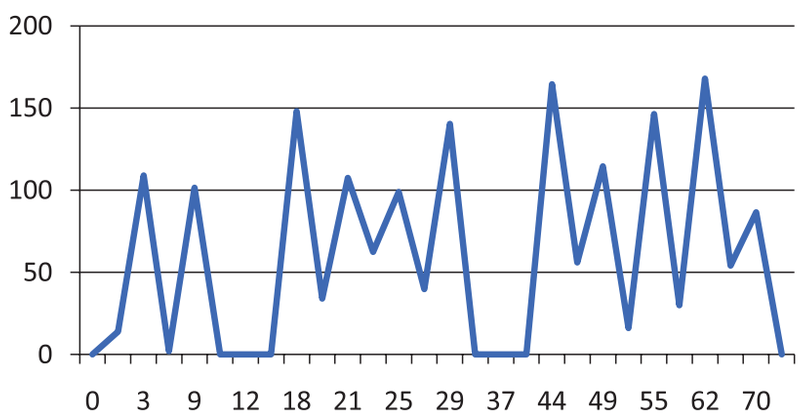

Рис. 2. Фактор первого типа

На Рис. 2 представлен график функции, моделирующей поведение фактора первого типа. Подобный вид фактора может использоваться при моделировании погодных условий и других природных факторов.

2) Кусочно-постоянная функция. На отрезках, для которых значение фактора выше критического, значение фактора – случайная величина, равномерно распределённая на отрезке от критического значения до 100. На отрезках, для которых значение фактора ниже критического, значение фактора – случайная величина, равномерно распределённая на отрезке от 0 до критического значения.

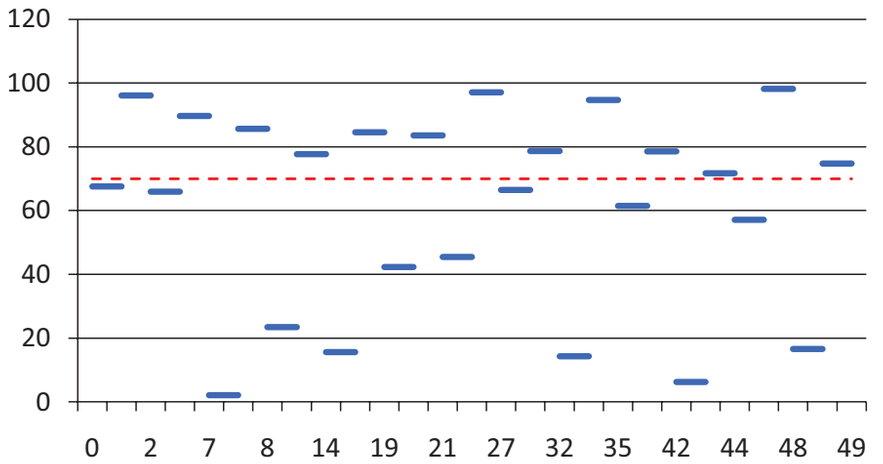

Рис. 3. Фактор второго типа

Этот вид функции может характеризовать факторы, отвечающие за техническое состояние оборудования. Значение фактора остается неизменно ниже критического уровня до тех пор, пока какое-то событие не приведет к резкому ухудшению его состояния. После этого состояние оборудования остаётся неизменным до тех пор, пока не будет устранена неисправность.

3) Факторы третьего типа моделируются схожим образом, однако зависимость в данном случае является не кусочно-постоянной, а кусочно-линейной. В начале каждого промежутка фактор, как и в предыдущем случае, принимает значение, равномерно распределённое на отрезке от критического уровня до 100 или от 0 до критического уровня, а затем линейно возрастает со случайным угловым коэффициентом (случайная величина с равномерным распределением от 0 до $\pi/6$).

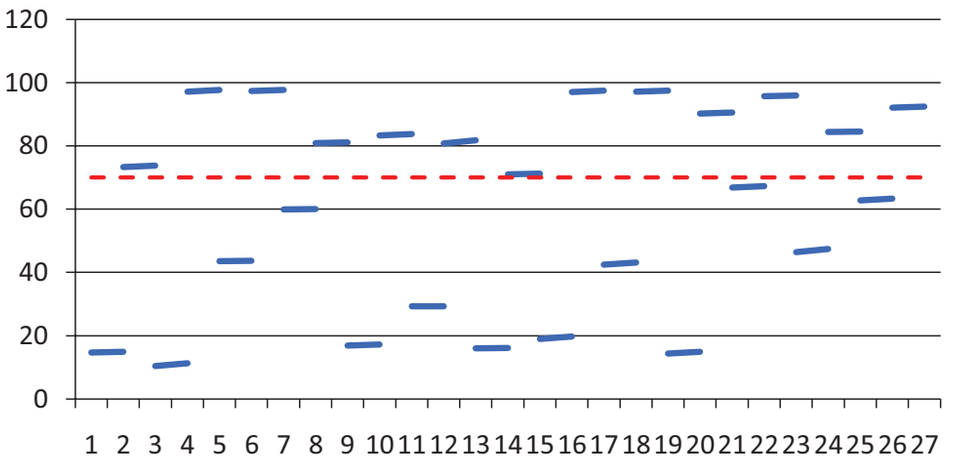

Рис. 4. Фактор третьего типа

Этот тип фактора можно использовать при моделировании состояния подвижного состава и дорожного полотна, а также человеческих факторов. Разрывы соответствуют поломкам и ремонтам, а медленное возрастание – постепенному изнашиванию оборудования.

4. Моделирование ущерба от транспортного происшествия

Из-за отсутствия большого массива реальных данных по происшествиям на российских железных дорогах, величина ущерба рассчитывалась на основе статистических данных о железнодорожных происшествиях в США. Ущерб моделируется как линейная комбинация факторов с аддитивной помехой (шумом). Шум моделируется на основе оценки закона распределения остатков при построении модели для ущерба на основе статистических данных о происшествиях в США.

Пусть X – случайная величина, плотность распределения $f(x)$ которой мы хотим оценить. Оценка плотности вероятности находится по формуле

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

где X_i – наблюдение за случайной величиной X, n – объём выборки, h – ширина окна, $K(z)$ – ядерная функция. Эту оценку часто называют оценкой Розенблатт-Парзена [16, 17].

Ядерную функцию можно выбирать в достаточно широких пределах, эта функция в основном отвечает за гладкость и дифференцируемость оценки $\hat{f}(x)$. В нашем случае в качестве ядерной функции была выбрана стандартная гауссовская плотность $K(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

В отличие от выбора ядерной функции выбор ширины окна h является ключевой проблемой при построении ядерной оценки $\hat{f}(x)$. Существуют различные способы выбора ширины окна, широко освещённые в литературе [18]. В нашем случае h выбиралась методом подстановки.

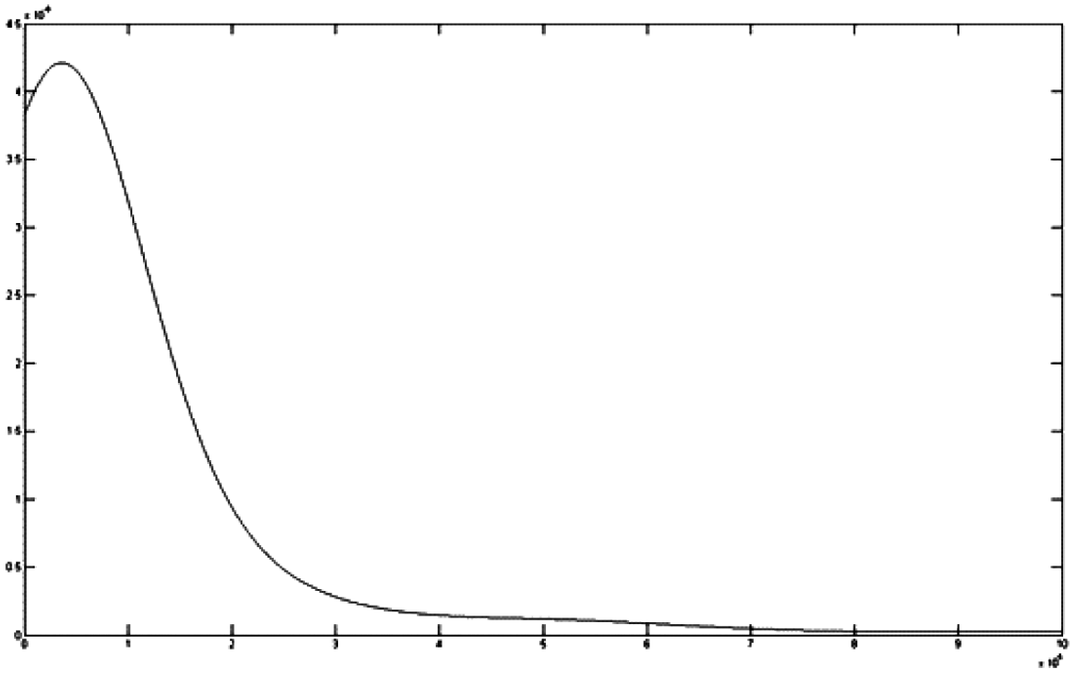

Рис. 5. Ядерная оценка плотности распределения ущерба по данным США

Итоговое распределение для шума было получено следующим образом:

- построена оценка $\hat{f}(x)$ плотности распределения ущерба по данным США;
- смоделирована реализация выборки, соответствующей случайной величине с плотностью $\hat{f}(x)$;
- построена линейная регрессия ущерба от значений факторов (реализация факторов была смоделирована в соответствии с ранее изложенным алгоритмом);
- вычислены остатки для построенной регрессии;
- на основе полученных остатков построена ядерная оценка плотности распределения шума $\hat{f}_\varepsilon(x)$;
- итоговый ущерб моделируется как линейная комбинация факторов с заданными коэффициентами плюс реализация случайной величины с плотностью $\hat{f}_\varepsilon(x)$.

5. Модельный пример

Рассмотрим линейную модель наблюдения

$$X_k = \theta_0 + \theta_1 F_1 + \theta_2 F_2 + \cdots + \theta_m F_m + \varepsilon_k \quad k = 1, \ldots, n,$$

где X_k – ущерб от k-го происшествия, F_i – фактор, влияющий на возникновение транспортного происшествия и величину ущерба, θ_i – неизвестные параметры, ε_k – случайный шум.

Смоделируем 260 наблюдений со значениями параметров линейной модели

$$\theta = \begin{bmatrix} 0 & 300 & 350 & 400 & 2500 & 500 & 100 \end{bmatrix}^T.$$

Оценка параметров линейной регрессии с помощью метода наименьших квадратов:

$$\hat{\theta} = \begin{bmatrix} 133331 & 0,027 & 0,036 & -0,091 & 0,243 & -0,011 & -0,041 \end{bmatrix}^T.$$

Коэффициент детерминации $R^2 = 0{,}28$, что говорит о плохой точности линейной модели.

Теперь по наблюдениям построим модель регрессии с точкой разрыва:

$$\hat{b} = 237768, \quad \hat{\alpha} = \begin{bmatrix} -19400 \\ 284{,}2 \\ 254{,}8 \\ 338{,}2 \\ 2121{,}1 \\ 423{,}5 \\ 136{,}2 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} 1433733 \\ -1522 \\ -1092 \\ -3921 \\ -5496 \\ -733 \\ -3311 \end{bmatrix}$$

Характеристики точности оценивания: $D = 0{,}54$, $R^2 = 0{,}74$.

Суммы квадратов остатков для обычной линейной модели в 9 раз больше, чем для модели с точкой разрыва.

Смоделируем еще 40 наблюдений с теми же значениями коэффициентов, что и ранее, и построим прогноз ущерба, используя модель с точкой разрыва.

При прогнозировании с использованием модели с точкой разрыва необходимо решить, к какой части регрессии относятся наблюдения, так как у нас нет значения ущерба и мы не можем использовать значение $\hat{b}$ для выбора вектора $\hat{\alpha}$ или $\hat{\beta}$. Для решения этой проблемы можно воспользоваться следующим алгоритмом.

1. Разделить имеющиеся наблюдения на две группы. К первой группе относятся те наборы значений факторов, для которых ущерб Y_k меньше значения $\hat{b}$. Ко второй группе относятся наборы значений факторов, для которых ущерб Y_k больше или равен, значению $\hat{b}$.

2. Вычислить для каждого фактора выборочное среднее по наборам из одной группы $\overline{F}_i^I = \frac{1}{n_1}\sum_{k=1}^{n_1} F_{i,k}$, $i = 1,\ldots,m$, где n_1 – количество наблюдения в первой группе, $F_{i,k}$ – значения фактора с номером i в k-м наблюдении. Аналогично вычисляются выборочные средние $\overline{F}_i^{II}$, $i = 1,\ldots,m$ для второй группы наблюдений.

3. При построении прогноза ущерба для наблюдений со значениями факторов $F_{1,pr},\ldots,F_{6,pr}$ выбираются оценки $\hat{\alpha}$ для наблюдений, отнесённых к первой группе и $\hat{\beta}$ для второй группы. Наблюдение относится к первой группе, если $\sum_{i=1}^{6}\left|F_{i,pr} - \overline{F}_i^I\right| \le \sum_{i=1}^{6}\left|F_{i,pr} - \overline{F}_i^{II}\right|$ и ко второй группе, если $\sum_{i=1}^{6}\left|F_{i,pr} - \overline{F}_i^I\right| > \sum_{i=1}^{6}\left|F_{i,pr} - \overline{F}_i^{II}\right|$.

Вычислим относительную погрешность прогноза по сравнению с истинными значениями ущерба, которые были также смоделированы. Обозначим прогноз через $\widetilde{X}$, тогда величина погрешности для одного наблюдения X определяется по формуле $\Delta = \frac{\left|\widetilde{X} - X\right|}{X}$.

Посчитаем среднюю относительную погрешность прогноза по 40 наблюдениям за факторами $\overline{\Delta} = \frac{1}{40}\sum_{k=1}^{40}\Delta_k$. Эта величина для модели с точкой разрыва составила 0,15. Процент ошибок при выборе группы, к которой относится наблюдения, составляет примерно 5%. Линейная модель сильно проигрывает по величине средней относительной погрешности, для неё она составила 0,52.

Стоит отметить, что точность прогноза значений для группы с большими значениями ущерба заметно уступает точности прогноза для наблюдений первой группы. Если строить прогнозы, используя только оценки $\hat{\alpha}$, то относительная погрешность составит 0.17. Дальнейшие исследования могут быть направлены на то, что после разделения наблюдения на две группы следует задуматься над тем, чтобы изменить способ построения модели для наблюдений с большими значениями ущербов.

Литература

1. **Розенберг Е.Н., Замышляев А.М., Прошин Г.Б.** Определение опасности возникновения транспортных происшествий и событий на основе контроля состояния факторов, влияющих на их возникновение // Надежность, 2009, № 3 (31). С. 37–50.

2. **Замышляев А.М., Кан Ю.С., Кибзун А.И., Шубинский И.Б.** Статистическая оценка опасности возникновения происшествий на железнодорожном транспорте // Надежность, 2012, № 2 (41). С. 104-117.

3. **Горяинов А.В., Кан Ю.С., Платонов Е.Н.** Регрессионные факторные модели для оценки рисков на железнодорожном транспорте. Труды и пленарные доклады участников конференции УКИ'12 / Научное издание. Электрон. текстовые дан. – М.:ИПУ РАН, 2012 – ISBN 978-5-91450-100-3 – С. 314-320.

4. **Xuedong Y.E., Radwan, M. Abdel-Aty.** Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model // Accident Analysis and Prevention. Vol. 37. 2005, P. 983–995.

5. **Joshua S., Garber N.** Estimating truck accident rate and involvement using linear and Poisson regression models // Transportation Planning and Technology. Vol. 15, 1990, P. 41–58.

6. **Jutaek O., Washington S.P., Doohee N.** Accident prediction model for railway-highway interfaces // Accident Analysis and Prevention. Vol. 38(2), 2006. P. 346–356.

7. **Loeb P.D., Clarke W.A.** The determinants of truck accidents // Transportation Research Part E. Vol. 43 (4), 2007. P. 442-452.

8. **Демиденко Е.З.** Линейная и нелинейная регрессии. – М.: Финансы и статистика, 1981 г.

9. **Netter J., Wasserman W., Kutner M.H.** Applied linear statistical model: Regression, analysis of variance, and experimental designs. Homewood, IL: Irwin. 1985.

10. Railroad Safety Statistics 2010, Rail Eqipment Acident Report 6180.54. U.S. Department of Transportation (http://safetydata.fra.dot.gov).

11. **Victor E. McGee, Willard T.C.** Piecewise Regression // Journal of the American Statistical Association. Vol. 65. 1970, № 331. P. 1109–1124.

12. **Martinez-Beneito M.A., García-Donato G., Salmerón D.A.** Bayesian Joinpoint regression model with an unknown number of break-points // The Annals of Applied Statistics. Vol. 5, 2011, № 3. P. 2150–2168.

13. **Rosenblatt M.** Remarks on Some Nonparametric Estimates of a Density Function, The Annals of Mathematical Statistics. Volume 27, № 3 (1956). P. 832-837.

14. **Parzen E.** On Estimation of a Probability Density Function and Mode, The Annals of Mathematical Statistics. Volume 33, № 3 (1962). P. 1065-1076.

15. **Расин Дж.** Непараметрическая эконометрика: вводный курс, Квантиль, № 4. С. 7–56.