Московский Государственный Университет им. М.В. Ломоносова

Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

Метод статистической верификации регрессионных моделей, основанный на перестановочных тестах.

Дипломная работа

выполнил

Дзыба Дмитрий Сергеевич

Научный руководитель

д.ф.-м.н.**Сенько Олег Валентинович**

Москва, 2014

Оглавление

1. Введение.

В различных отраслях знаний возникают задачи поиска закономерностей на основе данных полученных эмпирическим путём. Одним из самых распространённых способов описания зависимостей в биологии и медицине является использование простых линейных регрессионных моделей. Однако нередко оказывается, что использование простейших моделей не описывает полностью реально существующие зависимости. Усложнение модели при ограниченных размерах обучающих выборок приводит к эффекту переобучения, связанного с тем, что чрезмерная настройка под данные ведёт к ошибочной аппроксимации чисто случайных изменений, и, как следствие, к снижению обобщающей способности. Целью данной работы является поиск способа выбора регрессионной модели оптимальной сложности, который позволял бы делать количественные оценки необходимости использования более сложных моделей. В научной литературе этой проблеме уделено немало внимания. Предложено множество способов её решения, основанных на различных подходах. Такое разнообразие продиктовано несовершенством и ограниченностью применимости каждого из них.

Пожалуй, самыми популярными являются критерии, основанные на разбиении выборки, такие, как блоковый или поэлементный скользящий контроль и другие его разновидности. Оценка скользящего контроля сама по себе не даёт обоснованного ответа на вопрос насколько одна модель лучше другой, а требует дополнительных оценок с помощью статистических тестов.

Принцип максимальной длины описания Риссанена (MDL) связывает понятия правдоподобия и оптимальной длины описания, предлагая выбирать модели с оптимальным соотношением точности и длины описания. Оценка, получаемая в результате его применения, также не даёт объективного ответа на поставленный вопрос.

Ещё одним распространённым классом критериев, являются информационные критерии, такие как информационный критерий Акаике (AIC), байесовский информационный критерий Шварца (BIC), критерий Ханнана-Куинна (HQC) и другие. Все они предлагают выбирать модели с наименьшими вероятностями ошибки и наименьшим же числом параметров, предлагая различные схемы штрафов за увеличение этого числа, исходя из различных соображений. Однако общим является то, что значения оценок, получаемые в данных критериях, ничего не говорят о качестве модели и не могут быть использованы сами по себе, а имеют смысл лишь в сравнении при выборе модели с наименьшим значением оценки.

Теория Вапника-Червоненкиса была предложена для задачи распознавания классов. Она отталкивается от понятия ёмкости, по сути, означающего максимальное количество объектов, для которых существует алгоритм из некоторого семейства, верно классифицирующий их при произвольной разметке на классы. С её помощью можно оценить сверху ошибку модели на генеральной совокупности и выбрать модель с минимальной оценкой. При этом она также не позволяет дать точную количественную оценку необходимости выбора модели.

1.1. Метод статистической верификации регрессионных моделей, основанный на перестановочных тестах.

В данной работе использован новый критерий, основанный на статистических перестановочных тестах и принципе бритвы Оккама. В оригинале принцип бритвы Оккама выражен автором в двух высказываниях: «без необходимости не следует утверждать многое» (лат. «*Pluralitas non est ponenda sine necessitate*») и «то, что можно объяснить посредством меньшего, не следует выражать посредством большего» (лат. «*Frustra fit per plura quod potest fieri per pauciora*»). В терминах задачи поиска модели оптимальной сложности этот принцип можно записать так: «не следует использовать более сложную модель, если более простая модель исчерпывающе описывает

зависимость». Однако это что значит, исчерпывающе? Будем говорить, что модель исчерпывающе описывает зависимость в статистическом смысле, если отклонения от данной модели не могут быть объяснены с использованием какого-либо класса моделей. То есть отклонения от модели являются случайным шумом и независимы от переменных, используемых для построения модели. Является ли модель исчерпывающей в данном смысле можно проверить с помощью перестановочных тестов.

Обычно перестановочные тесты используются для установления самого факта наличия зависимости. Предположим, что рассматриваются объекты вида *(x,y)* из некоторого вероятностного пространства$(\Omega_0, \Sigma_0, \mathrm{P}_0)$. Где *x*-часть описания соответствует вектору независимых переменных, а *y*-часть – зависимой переменной. Выборки таких объектов размером n рассматриваются, как объекты нового вероятностного пространства $(\Omega_0^n, \Sigma_0^n, \mathrm{P}_0^n)$ с множеством элементарных исходов, являющимся *n*-кратным декартовым произведением множества Ω_0. Объекты в выборках считаются независимыми и имеющими одинаковое распределение.

Рассматривается нулевая гипотеза $\boldsymbol{H}_0$ о независимости значений *y* от значений векторов *x*. Тогда в качестве меры статистической достоверности наличия зависимости используется условная вероятность

$$\mathrm{P}_0^n\{T(\tilde{S}) > T(\tilde{S}_0)|\boldsymbol{H}_0, \tilde{S} \in W^p(\tilde{S}_0)\} \quad (1),$$

где $\tilde{S}_0$ – исходная выборка данных;

$W^p(\tilde{S}_0)$ – множество допустимых выборок, включающее $\tilde{S}_0$;

$T(\tilde{S})$ – функция, являющаяся мерой выборки $\tilde{S}$ от верифицируемой зависимости;

Если *x*-часть и *y*-часть описания объектов независимы, то множество новых объектов, полученных произвольной перестановкой *y*-частей описаний

заданных объектов относительно их *x*-частей, будет принадлежать множеству допустимых выборок. Поэтому в качестве $W^p(\tilde{S}_0)$ используется множество всех выборок полученных из исходной всеми возможными перестановками. В качестве $T(\tilde{S})$ чаще всего используются невязка или коэффициент корреляции между реальными данными и значениями верифицируемой зависимости (в этом случае знак неравенства в (1) меняется на противоположный). Тогда условная вероятность может быть вычислена как

$$\mathrm{P}_0^n\{T(\tilde{S}) > T(\tilde{S}_0)|\boldsymbol{H}_0, \tilde{S} \in W^p(\tilde{S}_0)\} = \frac{\mathrm{N}_1^p[T(\tilde{S}_0)]}{m!} \quad (2),$$

где $\mathrm{N}_1^p[T(\tilde{S}_0)]$ – число всех таких перестановок *y*-частей описания объектов относительно их *x*-частей, что для получаемых выборок $\tilde{S}'$ выполнено $T(\tilde{S}') > T(\tilde{S}_0)$;

На практике используется значение равное единице минус вычисленная условная вероятность, называемое *p*-значением.

Модифицируем перестановочный тест для получения статистической значимости верифицируемой модели относительно некоторой альтернативной модели. Предположим, что некоторая альтернативная модель исчерпывающе описывает зависимость в статистическом смысле, что значит, что отклонения от этой модели являются случайным шумом. Тогда рассмотрим гипотезу $\boldsymbol{H}_0$ о независимости отклонений значений *y* от альтернативной модели от значений векторов *x*. Если она верна, то отклонения можно переставлять и любая выборка полученная перестановкой отклонений является равновероятной с исходной. Тогда возможно оценить вероятность получения аппроксимации того же уровня качества, что и у верифицируемой модели, как условную вероятность вида (1), где в качестве $W^p(\tilde{S}_0)$ используется множество всех выборок полученных из исходной всеми возможными перестановками отклонений зависимой части описания

объектов от альтернативной модели. Эту вероятность можно посчитать с помощью выражения (2), которое остаётся верным для данной модификации перестановочного теста.

Если полученное в результате *p*-значение не превышает некоторый заданный доверительный уровень, то гипотеза $\boldsymbol{H}_0$ отвергается, и можно говорить о целесообразности применения верифицируемой модели, вместо альтернативной.

Таким образом, рассматривается некоторый ряд моделей в порядке увеличения их сложности. Понятие сложности задаётся исследователем и априорно не привязано к количеству параметров модели или длине её описания. Подходящей моделью оптимальной сложности считается модель минимальной сложности, которая верифицируется с помощью перестановочного теста относительно всех её более простых альтернатив, но не опровергается сама всеми более сложными моделями.

Достоинства данного метода напрямую следуют из достоинств перестановочного теста. Главным из них является то, что в результате его применения получается статистически обоснованная оценка. Кроме того, с его помощью возможна оценка необходимости применения произвольной эвристической регрессионной модели и отсутствует требование к объёму выборки наблюдений.

1.2. Задача поиска кусочно-линейной регрессионной модели зависимости концентрации паратиреоидного гормона в плазме крови от концентрации витамина D, и её обоснование.

Для иллюстрации данного метода рассматривается задача обоснования необходимости применения кусочно-линейной регрессионной модели зависимости концентрации паратиреоидного гормона в плазме крови от концентрации витамина D. Данные исходных наблюдений предоставлены «Институтом клинической эндокринологии» в структуре

«Эндокринологического научного центра министерства здравоохранения Российской Федерации». Исходная выборка состоит из 139 пар численных значений соответствующих концентраций. Концентрация витамина D измеряется в нанограммах на миллилитр кубический и обозначается в дальнейшем как *25(OH)D.* Концентрация паратиреоидного гормона измеряется в пикограммах на миллилитр кубический и обозначается как *ПТГ*.

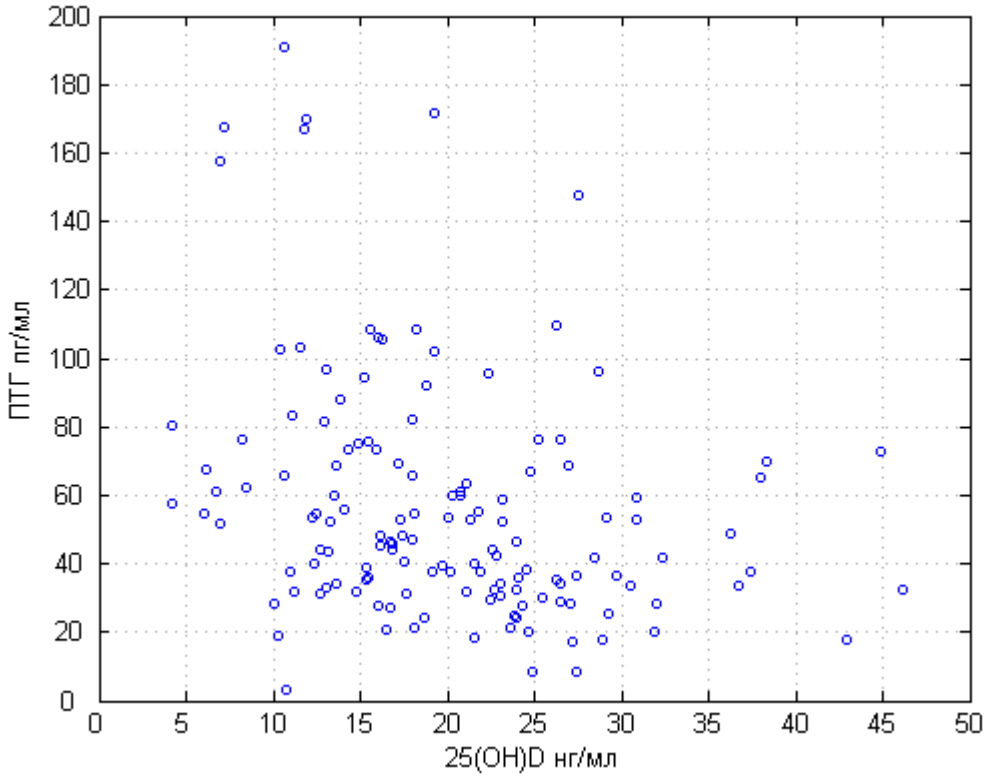

Рис.1. Данные 139 наблюдений зависимости ПТГ от 25(OH)D.

Концентрация паратиреоидного гормона в плазме крови связана с концентрацией витамина D косвенным образом, через концентрацию катионов кальция. Концентрация витамина D преимущественно зависит от внешних факторов: большая его часть поступает в организм человека с пищей и вырабатывается в коже под действием ультрафиолетовых лучей. Он улучшает абсорбцию кальция, увеличивая концентрацию катионов кальция в крови. При недостатке витамина D, как следствие, падает и концентрация катионов кальция. Чем меньше становится катионов кальция, тем больше паратиреоидного гормона вырабатывается околощитовидными железами,

который призван скомпенсировать возникший недостаток, задействуя несколько механизмов. Он увеличивает реабсорбцию кальция в почках и стимулирует выработку внутреннего витамина D. Но самым опасным является то, что он активизирует резорбцию костей, увеличивая выработку остеокластов, разрушающих костную ткань для высвобождения кальция. Выход выработки паратиреоидного гормона или получения витамина D за пределы нормы приводит к тяжёлым заболеваниям. Поэтому медиков интересует знание точной зависимости концентраций этих веществ в плазме крови.

Исходя из физиологических соображений, выдвинута гипотеза, согласно которой концентрация паратиреоидного гормона должна выходить на плато при некоторой концентрации витамина D, то есть зависимость должна хорошо описываться кусочно-линейной моделью состоящей из двух сегментов, разделённых этой точкой. Точка выхода данной зависимости на плато не известна, поэтому задача разбивается на две подзадачи: нахождение этой точки и обоснование целесообразности применения какой-либо соответствующей кусочно-линейной регрессионной модели.

2. Способы поиска точки выхода зависимости на плато.

Для поиска точки выхода на плато было разработано четыре метода. Все они исходят из предположения, что в исследуемой зависимости точка выхода на плато существует и единственна.

2.1. Подход, основанный на разности коэффициентов корреляции слева и справа относительно предполагаемой точки выхода зависимости на плато.

Данный подход основан на корреляционном анализе. Для редуцирования влияния шумовых данных, в этом подходе, в качестве зависимой переменной используется не сама концентрация *ПТГ* в плазме крови, а её натуральный логарифм *log(ПТГ)*.

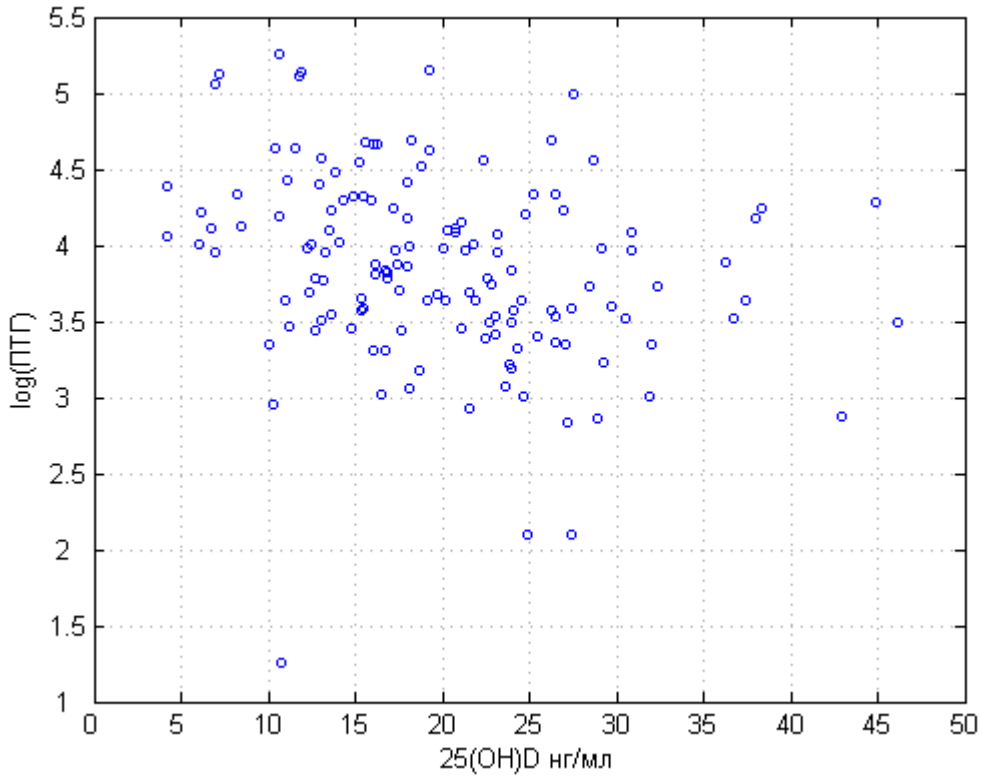

Рис.2. Данные 139 наблюдений зависимости log(ПТГ) от 25(OH)D.

В области плато концентрация *ПТГ*, а, следовательно, и значение *log(ПТГ)* не зависят от концентрации *25(OH)D*. Это значит, что на интервале значений

25(OH)D соответствующих плато коэффициент корреляции не должен статистически отличаться от 0. Вне его, напротив, должен обнаруживать статистически значимую зависимость. Поэтому коэффициент корреляции группы данных слева от некоторого значения содержания *25(OH)D* и разность коэффициентов корреляции групп данных лежащих слева и справа от этого же значения должны достигать максимума по модулю в точке, соответствующей точке $\boldsymbol{P}$ выхода зависимости на плато. Кроме этого, для достижения статистической достоверности выборочных коэффициентов корреляции необходимо, чтобы группы данных слева и справа от $\boldsymbol{P}$ были достаточно большими по размеру. Учитывая всё вышесказанное, искать точку $\boldsymbol{P}$ можно максимизируя значение:

$$\mathrm{C}(P) = |R_l| * |R_l - R_r| * \sqrt{m_l * m_r}\ (3),$$

где R_l – коэффициент корреляции *25(OH)D* и *log(ПТГ)* группы наблюдений слева от точки $\boldsymbol{P}$;

R_r – коэффициент корреляции *25(OH)D* и *log(ПТГ)* группы наблюдений справа от точки $\boldsymbol{P}$;

m_l – число наблюдений слева от точки $\boldsymbol{P}$;

m_r – число наблюдений справа от точки $\boldsymbol{P}$;

Значения этой функции вычисляются для множества точек, являющихся центрами интервалов между соседними по *25(OH)D* наблюдениями. Абсолютный максимум достигается в точке $\boldsymbol{P} = 7.7$. Однако, в этой точке m_l слишком мало ($m_l = 8$), и, учитывая большую неустойчивость значений корреляций на выборках такого размера, эту точку можно отбросить. Следующее за ним максимальное значение достигается в $\boldsymbol{P} = 24.7$.

Существенными недостатками такого подхода являются неоднородность выборок данных слева и справа от искомой точки и необходимость введения штрафа за их недостаточно большие размеры при приближении к границам

области допустимых значений ***P***. Достоинством же, то, что для вычисления каждого значения $C(P)$ используются все имеющиеся данные.

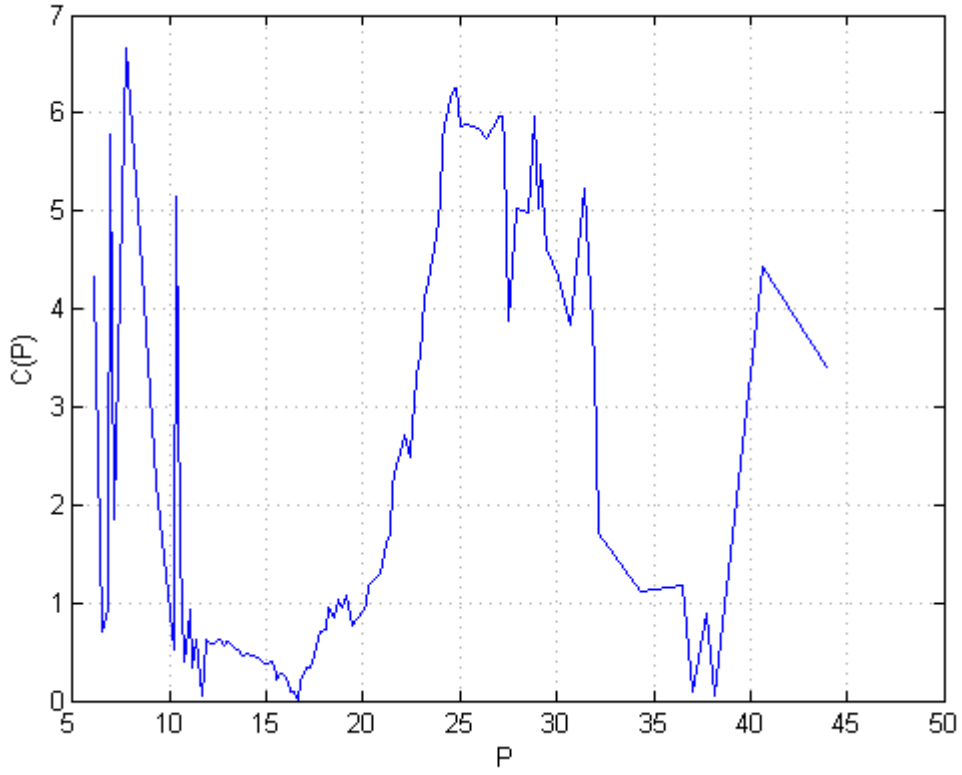

Рис.3. Зависимость величины критерия C(P) поиска точки выхода концентрации ПТГ на плато от её предполагаемого значения.

2.2. Подход, основанный на разности коэффициентов корреляции справа и слева относительно предполагаемой точки выхода зависимости на плато в окне заданной ширины.

Данный подход, также как и предыдущий основан на корреляционном анализе. Однако в отличие от него для вычисления значения оценки, является ли точка ***P*** точкой выхода зависимости на плато, используется не вся выборка, а лишь ограниченная её часть.

Назовём шириной окна $w \in \mathbb{N}$. Предположим, что точке выхода на плато будет соответствовать точка, в которой достигается максимум модуля

разности коэффициентов корреляции групп из w наблюдений слева и w наблюдений справа, ближайших к этой точке по *25(OH)D*.

Пусть $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ – множество из n наблюдений, упорядоченное по возрастанию первого значения в каждой паре. В данной задаче первое значение в каждой паре соответствует концентрации *25(OH)D*, а второе – концентрации *ПТГ*.

Пусть m – максимальное число элементов этого множества с одинаковыми значениями x. Тогда для всех w и $\boldsymbol{P}$ таких, что $\left\lfloor \frac{\boldsymbol{n}}{\boldsymbol{2}} \right\rfloor \geq w \geq m + 1$ и $x_{n-w+1} > P > x_w$ можно выделить две группы из w объектов слева и w объектов справа от точки $\boldsymbol{P}$, и определена функция

$$W(P, w) = |R_l^w - R_r^w| \ (4) \ ,$$

где R_l^w – коэффициент корреляции *25(OH)D* и *ПТГ* группы из w ближайших наблюдений слева от точки $\boldsymbol{P}$;

R_r^w – коэффициент корреляции *25(OH)D* и *ПТГ* группы из w ближайших наблюдений справа от точки $\boldsymbol{P}$;

Может оказаться, что определить w ближайших объектов с той или иной стороны однозначным образом невозможно, потому что необходимо выбрать объект или объекты из нескольких наблюдений с одинаковыми значениями x, тогда как если включить все эти объекты, то в группе станет более чем w наблюдений. В этом случае в группу включаются все объекты вызывающие неоднозначность, и функция $W(P, w)$ определяется на группах из не менее чем w ближайших наблюдений слева и не менее чем w ближайших наблюдений справа от точки $\boldsymbol{P}$.

Ширина окна w является управляющим параметром. При заданной ширине окна w значения $W(P, w)$ вычисляются для множества точек, являющихся центрами интервалов между соседними по *25(OH)D* наблюдениями,

попадающими в область определения. При разных значениях w максимум функции $W(P,w)$ может достигаться в различных значениях $\boldsymbol{P}$.

Рассмотрим среднюю ошибку нахождения точки выхода зависимости на плато для различных w на смоделированных данных. Возьмём выборку из 300 точек, состоящую из двух частей. Первая часть состоит из 100 точек (x_i, y_i) равномерно распределённых по x на интервале [0,10], с $y = -x + 10 + u$, где u имеет стандартное нормальное распределение. Вторая часть состоит из 200 точек (x_i, y_i) равномерно распределённых по x на интервале [10,30], с y из стандартного нормального распределения.

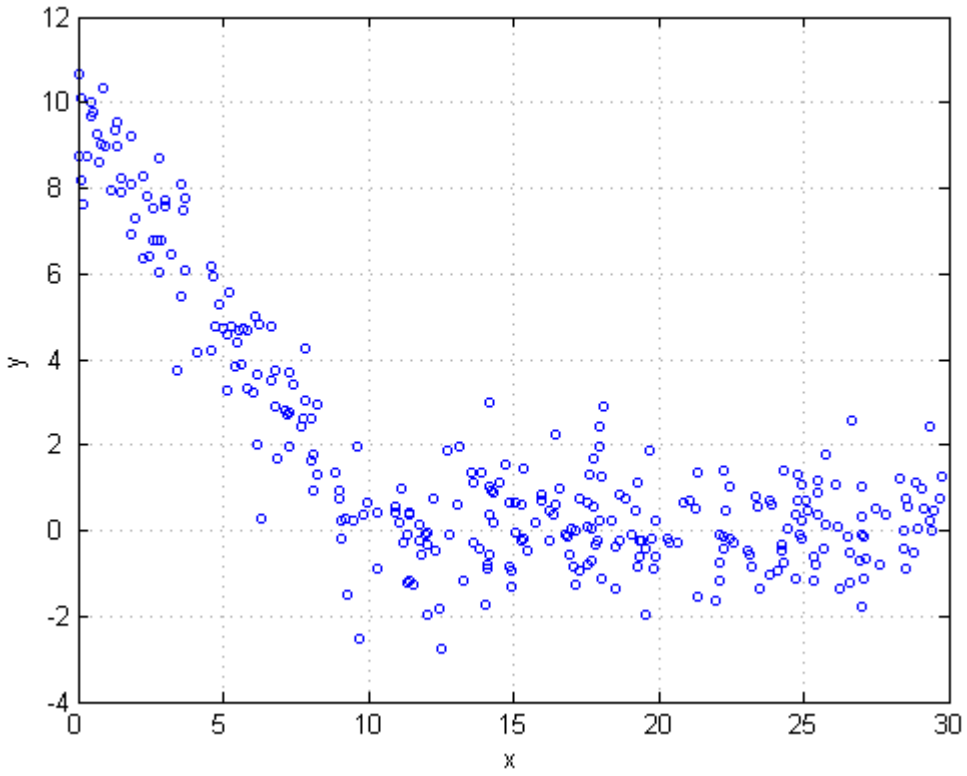

Рис.4. Пример смоделированной выборки данных

По способу моделирования эти данные выходят на плато в точке со значением $x = 10$. Для всех возможных w вычислим ошибку нахождения этой точки, как модуль разности полученного значения с известным верным значением. Повторим генерацию данных и вычисление ошибок 1000 раз, и

посчитаем суммарные, усреднённые по количеству повторений ошибки $S(w)$ для всех w.

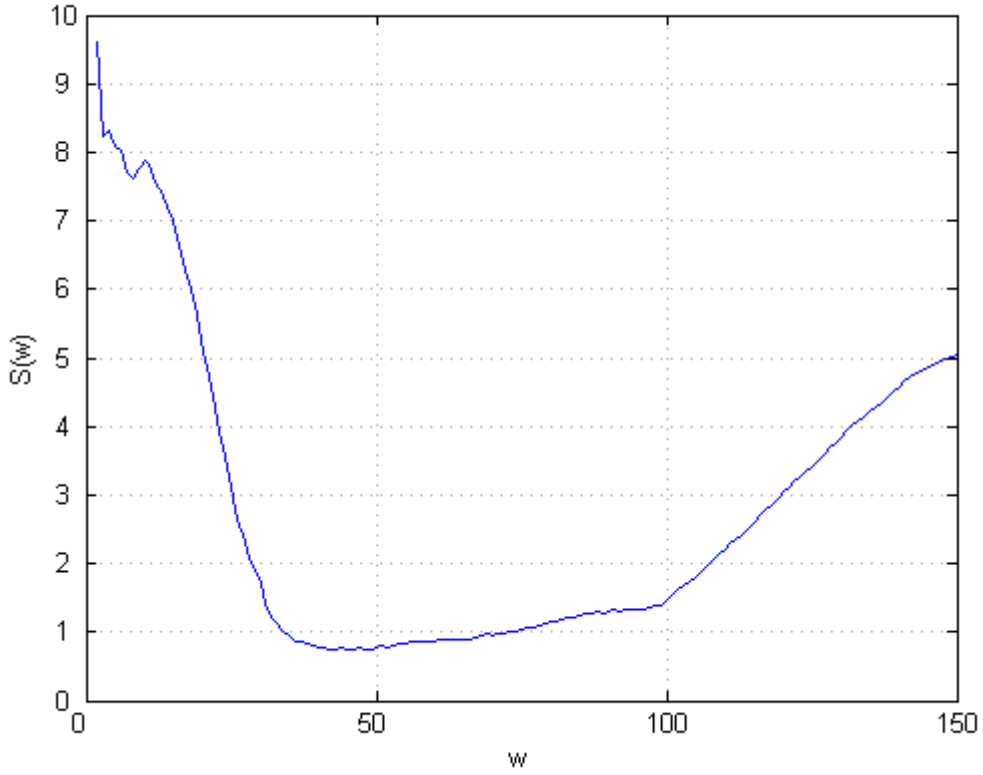

Рис.5. Зависимость значения средней ошибки $S(w)$ на тысяче повторений генерации данных, от ширины окна w.

При увеличении w область определения $W(P, w)$ по $\boldsymbol{P}$ сужается. Существует опасность, что искомая точка выхода на плато окажется за её границами. На смоделированных данных хорошо видно, что в этом случае, при ширине окна $w > 100$, ошибка начинает стремительно нарастать. Однако при чрезмерном уменьшении w уменьшается статистическая достоверность оценки, а значит будет расти и ошибка. Что видно на смоделированных при $w < 40$. Поэтому необходимо выбирать относительно средние значения ширины окна, учитывая предполагаемую область нахождения искомой точки выхода зависимости на плато. Вопрос оптимального выбора ширины окна требует дополнительного исследования.

В исходном множестве наблюдений 139 значений и не более 3 одинаковых значений концентрации *25(OH)D* подряд. Поэтому w должно лежать в интервале [4, 69].

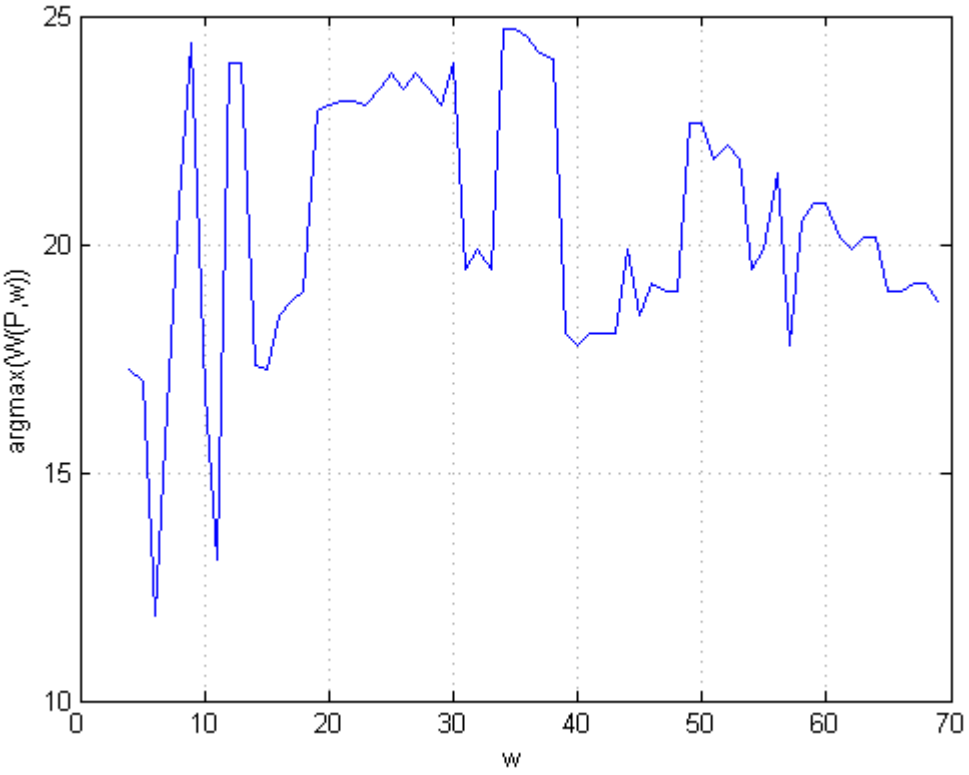

Рис.6. Зависимость найденной с помощью критерия $W(P,w)$ точки выхода концентрации ПТГ на плато от значения ширины окна w.

Для большинства значений w точка излома лежит в интервале от 17 до 25. В качестве примера возьмём $w = 30$. При этом выборе w абсолютный максимум достигается в точке $\boldsymbol{P} = 23.95$.

Недостатком такого подхода является ограничение на область поиска точки выхода на плато. Однако наличие гарантированного размера выборки по которой считаются корреляции равного ширине окна решает проблему с неоднородностью выборок справа и слева от искомой точки и позволяет отказаться от введения штрафа за их недостаточный размер, меняющего всю картину. Кроме того такой способ позволяет искать точки изменения зависимости в задачах с более чем одним изломом.

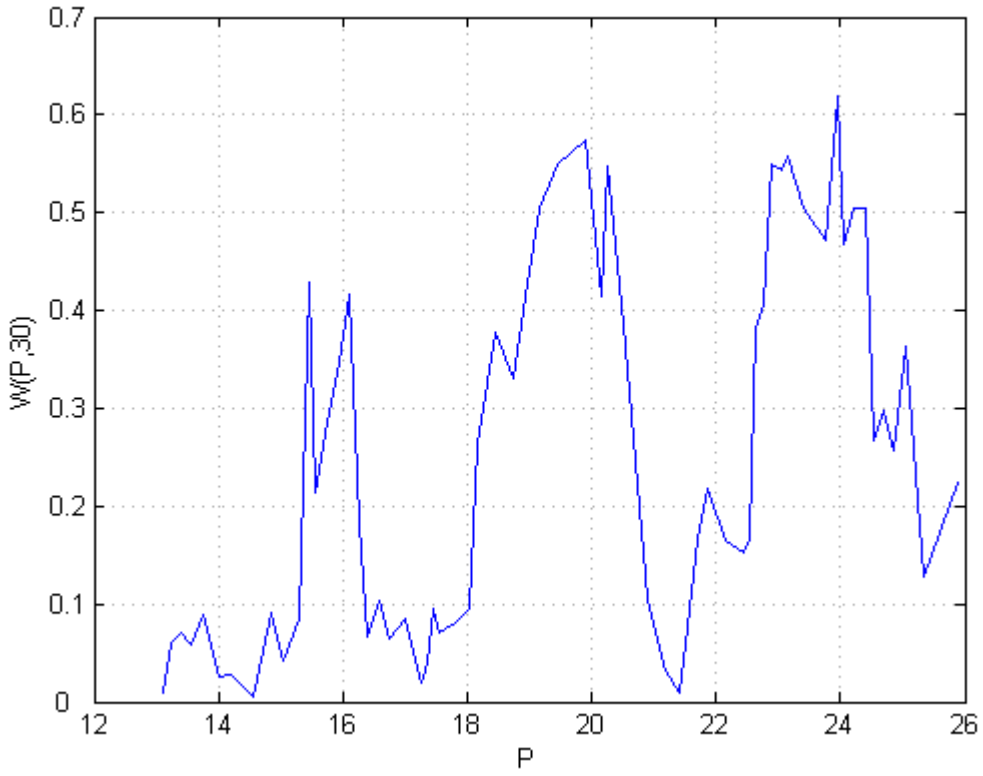

Рис.7. Зависимость величины критерия $W(P,w)$ *точки выхода концентрации ПТГ на плато от её значения, для ширины окна* $w = 30$.

2.3. Подход, основанный на минимизации суммарной невязки кусочно-линейной регрессионной модели с разрывом.

Возьмём некоторую точку ***P*** и построим с помощью метода наименьших квадратов оптимальные линейные регрессионные модели групп данных слева и справа от неё. Наименьшие значения суммарной невязки слева и справа от ***P*** будут достигаться в точках существенных изменений зависимости.

$$S(P) = S_l + S_r \ (5),$$

где S_l – суммарная невязка группы наблюдений слева от точки ***P***, то есть

$S_l = \sum_{i=1}^{n}(f_l(d_i) - g_i)^2$, $f_l(d) = v_1^l + v_2^l * d$ – оптимальная линейная регрессионная модель для данной группы;

S_r– суммарная невязка группы наблюдений справа от точки ***P***, то есть

$S_r = \sum_{i=1}^{n}(f_r(d_i) - g_i)^2$, $f_r(d) = v_1^r + v_2^r * d -$ оптимальная линейная регрессионная модель для данной группы;

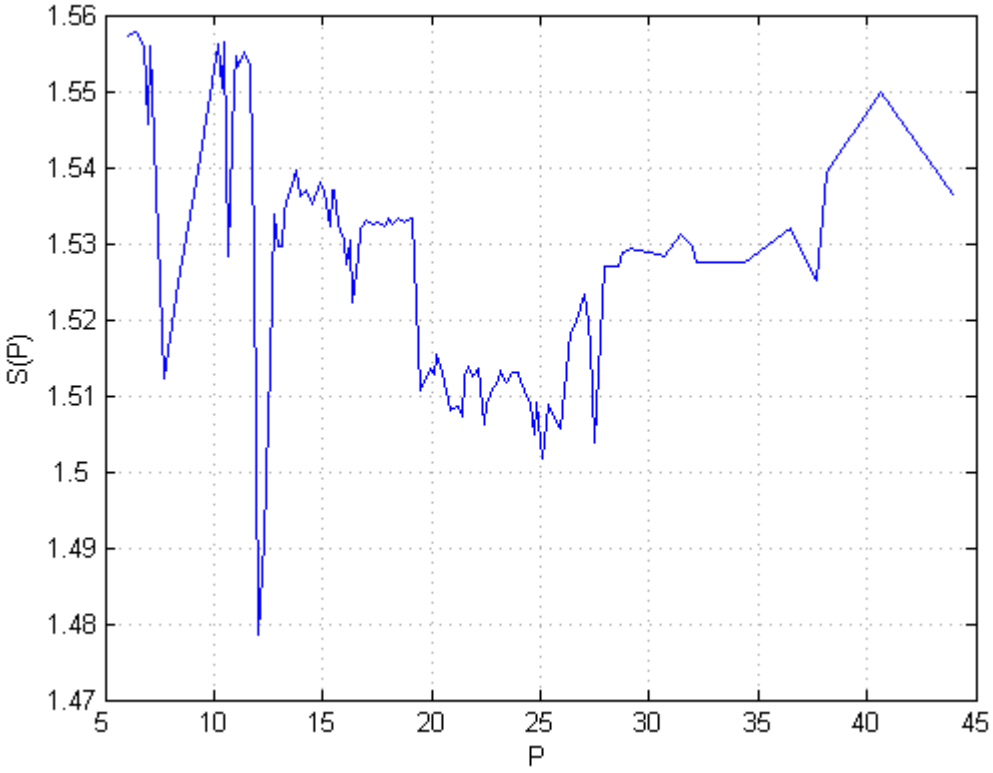

Рис.8. Зависимость величины $S(P)$ суммарной невязки оптимальных линейных регрессионных моделей групп наблюдений слева от некоторой точки и справа от неё, от значения этой точки.

Значения суммарной невязки вычисляются для множества точек, являющихся центрами интервалов между соседними по *25(OH)D* наблюдениями. Абсолютный минимум $S(P)$ достигается в $\boldsymbol{P} = 12.05$. Что существенно расходится с результатами других подходов. Однако точка пересечения полученных оптимальных линейных регрессионных моделей при данном выборе точки выхода на плато лежит далеко за границами области задаваемой множеством наблюдений по значению концентрации *25(OH)D*. Значение концентрации *25(OH)D* в точке пересечения регрессионных моделей 1.2229, тогда как минимальное значение концентрации *25(OH)D* на всём множестве наблюдений 4.2. Кроме того,

полученные регрессионные модели имеют значительный разрыв в точке предположительного выхода на плато. Он равен 32.9023, что составляет более 17.5% от размера области концентрации *ПГТ* задаваемой множеством наблюдений. С точки зрения физиологического механизма лежащего в основе исследуемой зависимости такого существенного разрыва быть не должно. Поэтому можно говорить о том, что для корректного нахождения точки выхода исследуемой зависимости на плато требования минимальности суммарной невязки недостаточно.

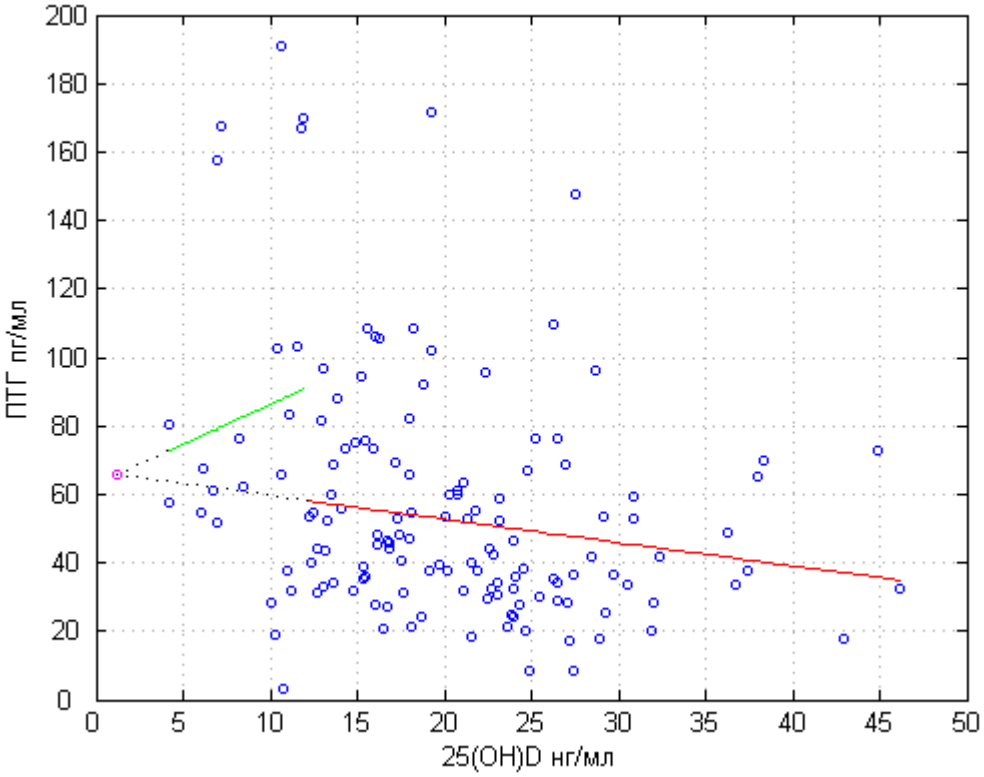

Рис.9. Зависимость ПТГ от 25(OH)D. Красная линия изображает оптимальную регрессионную зависимость справа от точки $\boldsymbol{P} = 12.05$. *Зелёная линия изображает оптимальную регрессионную зависимость слева от точки* $\boldsymbol{P} = 12.05$. *Обе линии составляют общую найденную зависимость.*

Тогда естественно наложить на модель требование непрерывности. Будем минимизировать расстояние между значениями построенных оптимальных линейных регрессионных моделей в предполагаемой точке излома.

$$D(P) = |f_l(P) - f_r(P)| \ (6),$$

где f_l – оптимальная линейная регрессионная модель для группы наблюдений слева от точки ***P***;

f_r – оптимальная линейная регрессионная модель для группы наблюдений справа от точки ***P***;

Сама по себе зависимость величины разрыва между регрессионными моделями справа и слева от некоторой точки от положения этой точки не представляет существенного значения.

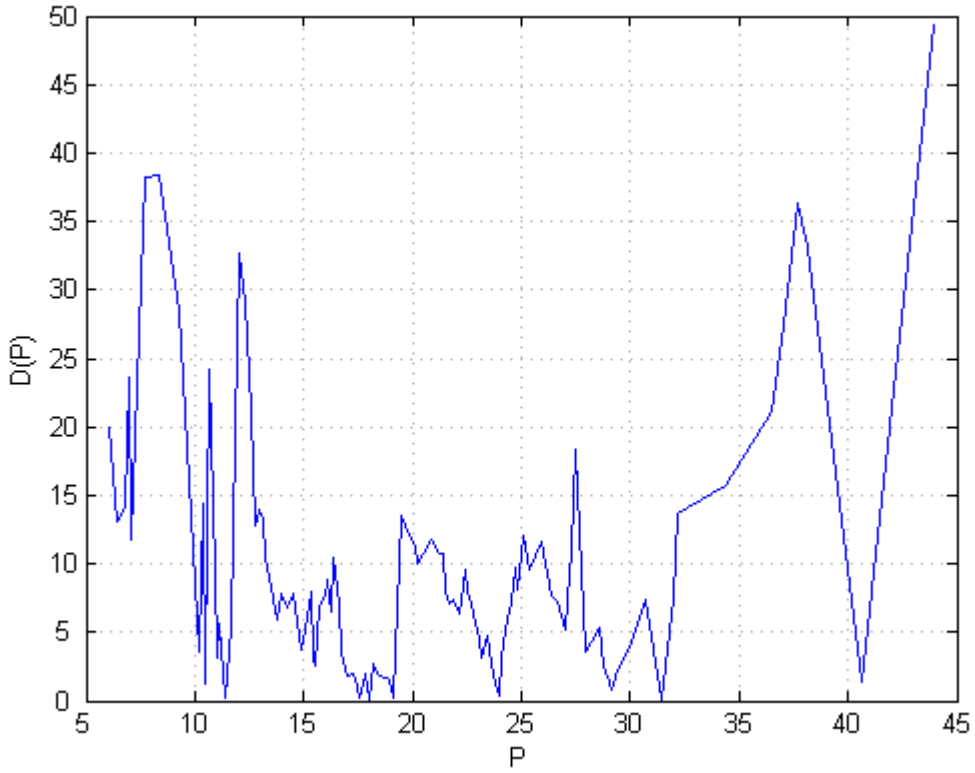

Рис.10. Зависимость величины $D(P)$ расстояния между значениями оптимальных линейных регрессионных моделей групп наблюдений слева от некоторой точки и справа от неё в этой точке, от её значения.

Чем меньше величина разрыва, тем лучше, однако, это достижение ей абсолютного минимума не является достаточным условием для наличия точки выхода на плато. Поэтому естественно построить общий взвешенный

критерий, который учтёт требования минимальности суммарной невязки и непрерывности модели.

Значения $S(P)$ и $D(P)$ имеют различную природу и чтобы иметь возможность использовать их вместе, их необходимо нормировать. Пусть

$$S_{norm}(P) = \frac{S(P) - \min(S(P))}{\max(S(P) - \min(S(P)))} \quad (7),$$

$$D_{norm}(P) = \frac{D(P) - \min(D(P))}{\max(S(P) - \min(S(P)))} \quad (8).$$

Тогда искать точку выхода на плато можно минимизируя их взвешенную сумму.

$$B(P, w_S, w_D) = S_{norm}(P) * w_S + D_{norm}(P) * w_D \quad (9),$$

где $w_S \in \mathbb{R}$, $w_S \geq 0$ – некоторый коэффициент, отвечающий за важность требования минимальности суммарной невязки;

$w_D \in \mathbb{R}$, $w_D \geq 0$ – некоторый коэффициент, отвечающий за важность требования непрерывности модели;

Для заданных коэффициентов важности требований w_S и w_D значение $B(P, w_S, w_D)$ вычисляется для середины каждого интервала между двумя соседними по содержанию *25(OH)D* наблюдениями. Будем считать оба требования равноценными, то есть возьмём w_S и w_D равными 1.

Абсолютный минимум $B(P, 1{,}1)$ достигается в точке $\boldsymbol{P} = 23.95$.

Несмотря на то, что данный подход идёт от идеи перебора возможных оптимальных регрессионных моделей с точкой излома, его вычислительная сложность по времени равна сложности первого подхода основанного на корреляционном анализе.

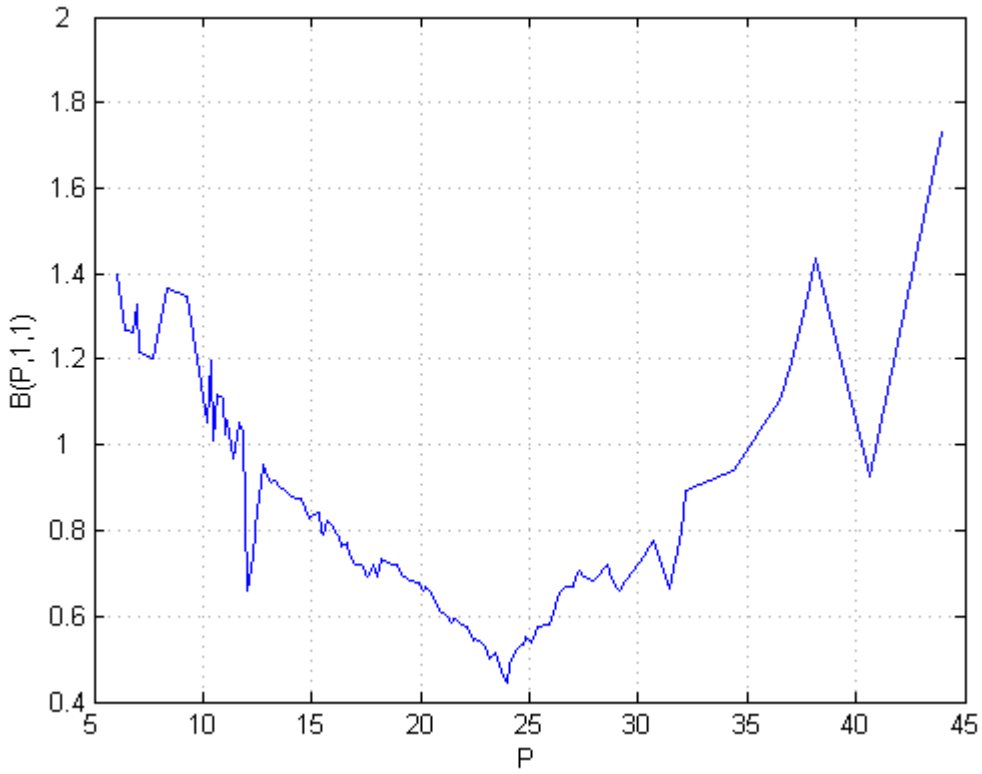

Рис.11. Зависимость величины критерия $B(P, w_S, w_D)$ *точки выхода концентрации ПТГ на плато от её значения, для коэффициентов значимости* $w_S = 1$ и $w_D = 1$.

Недостатком такого подхода является его заведомая ограниченность выбранным способом построения регрессионных моделей, тогда как подходы основанные на корреляционном анализе дают информацию, которая может быть интерпретирована шире. Такой выбор регрессионных моделей предполагает наличие разрыва, что может быть неприемлемым в некоторых задачах. Однако в заданных ограничениях данный подход даёт лучшую точность по сравнению с подходами, основанными на корреляционном анализе, что выявляется на модельных данных.

2.4. Подход, основанный на минимизации невязки кусочно-линейной непрерывной регрессионной модели.

Возьмём некоторую точку ***P*** и построим непрерывную кусочно-линейную регрессионную модель, состоящую из двух линейных сегментов разделённых

этой точкой. Наименьшие значения её невязки будут достигаться в точках $\boldsymbol{P}$ существенных изменений исходной зависимости.

Пусть $\boldsymbol{P}$ – предполагаемая точка излома. Тогда поиску оптимальной кусочно-линейной регрессионной модели при фиксированной $\boldsymbol{P}$ соответствует задача оптимизации

$$F = \sum_{x_j<\boldsymbol{P}} \left(y_j - \alpha_0^0 - \alpha_1^0 * x_j\right)^2 + \sum_{x_j>\boldsymbol{P}} \left(y_j - \alpha_0^1 - \alpha_1^1 * x_j\right)^2 \to min \ (10),$$

при условии

$$\alpha_0^0 + \alpha_1^0 * \boldsymbol{P} = \alpha_0^1 + \alpha_1^1 * \boldsymbol{P} \ (11),$$

где α_0^0 и α_1^0 – коэффициенты модели слева от $\boldsymbol{P}$;

α_0^1 и α_1^1 – коэффициенты модели справа от $\boldsymbol{P}$;

Для нахождения условного минимума F, воспользуемся методом множителей Лагранжа. Необходимым условием условного минимума F будет равенство нулю частных производных Лагранжиана

$$F + \lambda * (-\alpha_0^0 - \alpha_1^0 * \boldsymbol{P} + \alpha_0^1 + \alpha_1^1 * \boldsymbol{P}) \ (12)$$

по α_0^0, α_1^0,α_0^1 и α_1^1, при выполнении заданного ограничения. Таким образом, эта задача вырождается в систему из пяти линейных уравнений:

$$\begin{cases} m_l * \alpha_0^0 + X_l * \alpha_1^0 - \frac{1}{2} * \lambda = Y_l \\ X_l * \alpha_0^0 + s_{xx}^l * \alpha_1^0 - \frac{P}{2} * \lambda = s_{xy}^l \\ m_r * \alpha_0^1 + X_r * \alpha_1^1 + \frac{1}{2} * \lambda = Y_r \\ X_r * \alpha_0^1 + s_{xx}^r * \alpha_1^1 + \frac{P}{2} * \lambda = s_{xy}^r \\ -\alpha_0^0 - P * \alpha_1^0 + \alpha_0^1 + P * \alpha_1^1 = 0 \end{cases} (13),$$

где $m_l = \sum_{i:x_i<P} 1$; $m_r = \sum_{i:x_i>P} 1$;

$$X_l = \sum_{i:x_i<P} x_i \; ; \; X_r = \sum_{i:x_i>P} x_i \; ; \; Y_l = \sum_{i:x_i<P} y_i \; ; \; Y_r = \sum_{i:x_i>P} y_i \; ;$$

$$s_{xx}^l = \sum_{i:x_i<P} (x_i)^2 \; ; \; s_{xx}^r = \sum_{i:x_i>P} (x_i)^2 \; ; \; s_{xy}^l = \sum_{i:x_i<P} x_i * y_i \; ; \; s_{xy}^r = \sum_{i:x_i>P} x_i * y_i \; ;$$

Данная система имеет аналитическое решение.

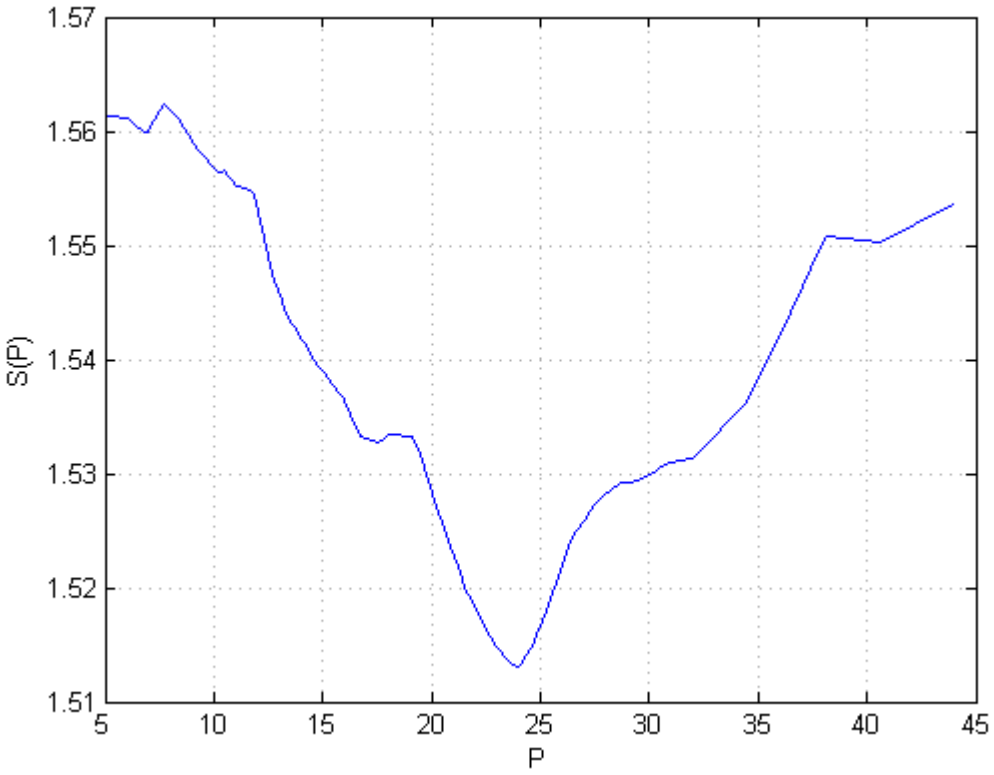

Рис. 12. Зависимость невязки S(P) оптимальной непрерывной кусочно-линейной регрессионной модели от её точки излома P.

Тогда подобная регрессионная модель строится для множества точек, являющихся центрами интервалов между соседними по *25(OH)D* наблюдениями, и для каждой из этих точек вычисляется её невязка $S(P)$.

Абсолютный минимум $S(P)$ достигается в точке $\boldsymbol{P} = 23.95$, которую мы и будем считать точкой выхода концентрации *ПТГ* на плато.

Относительным недостатком этого подхода является то, что линейные части получаемых кусочно-линейных моделей слева и справа от предполагаемых

точек выхода на плато не являются оптимальными линейными регрессиями сами по себе, в отличие от предыдущего метода. Однако неоспоримым преимуществом данного подхода является исключение самой возможности разрыва.

2.5. Сравнение различных подходов поиска точки выхода зависимости на плато на смоделированных данных.

Рассмотрим среднюю ошибку нахождения точки выхода зависимости на плато различными методами на смоделированных данных при множественных повторениях их генерации. Будем моделировать данные аналогично моделированию, описанному в пункте 2.2.

Минимальное значение суммарной усреднённой ошибки по всем w для 1000 повторений моделирования данных достигается в $w = 45$. Возьмём это значение в качестве параметра ширины окна в сравниваемом методе поиска точки выхода на плато, основанном на разности коэффициентов корреляции справа и слева относительно предполагаемой точки в окне заданной ширины.

Будем считать требования минимизации невязки и непрерывности в методе поиска точки выхода зависимости на плато, основанном на минимизации суммарной невязки кусочно-линейной регрессионной модели с разрывом, равноценными, то есть возьмём w_S и w_D равными 1.

Для всех четырёх методов поиска точки выхода зависимости на плато будем вычислять ошибку нахождения этой точки, как модуль разности полученного значения с известным верным значением. Повторим генерацию данных и вычисление ошибок 1000 раз, и посчитаем суммарные, усреднённые по количеству повторений ошибки.

По результатам моделирования наименьшую среднюю ошибку имеет метод поиска точки выхода зависимости на плато, основанный на минимизации невязки непрерывной кусочно-линейной регрессионной модели. Что является

ожидаемым, так как в своей основе моделируемые данные имеют именно непрерывную кусочно-линейную зависимость.

Метод поиска ***P***	*Средняя ошибка на 1000 повторений*
$\mathrm{C}(P)$ (в кач. зав. переменной *log(y)*)	1.5547
$W(P, 45)$	0.7531
$B(P, 1{,}1)$	0.2189
$S(P)$	0.2013

Таб.1. Средние значения ошибок нахождения точки выхода зависимости на плато с помощью различных методов поиска на смоделированных данных при их многократной генерации.

3. Регрессионные модели

С помощью описанных выше методов были найдены две точки $\boldsymbol{P} = 24.7$ и $\boldsymbol{P} = 23.95$, предположительно являющимися точками выхода концентрации *ПТГ* в плазме крови на плато.

При нахождении точки $\boldsymbol{P} = 24.7$ рассматривалась модифицированная выборка, где в качестве зависимой переменной использовался натуральный логарифм концентрации *ПТГ* в плазме крови. Поэтому естественно использовать такую же выборку при расчёте всех статистических параметров и построении регрессионных моделей для этой точки. Выборочный коэффициент корреляции группы данных слева от этой точки $R_l = -0.3027$, что указывает на существенную зависимость между *log(ПТГ)* и *25(OH)D* на данном интервале. Выборочный коэффициент корреляции группы данных справа от неё $R_r = 0.0402$, что говорит о том, что на данном интервале нет выраженной зависимости между *log(ПТГ)* и *25(OH)D.*

Для $\boldsymbol{P} = 23.95$ выборочный коэффициент корреляции группы данных слева от неё $R_l = -0.2934$, что также указывает на существенную зависимость между *ПТГ* и *25(OH)D* на этом интервале. Выборочный коэффициент корреляции группы данных справа от неё $R_r = 0.0351$, что говорит о том, что на данном интервале нет выраженной зависимости между *ПТГ* и *25(OH)D.* Таким образом, в качестве точек выхода на плато можно рассматривать обе точки.

Точка $\boldsymbol{P} = 24.7$ найдена с помощью корреляционного анализа, который не предполагает каких либо ограничений на тип регрессионной модели. В свою очередь точка $\boldsymbol{P} = 23.95$ найдена с помощью обоих методов, минимизирующих невязку для регрессионных моделей разных типов. Поэтому для каждой из этих точек можно рассматривать оба варианта: кусочно-линейную регрессионную модель с разрывом и непрерывную кусочно-линейную регрессионную модель.

3.1. Кусочно-линейная модель с разрывом в точке $P = 24.7$.

При $P \leq 24.7$ данная зависимость выражается формулой

$$log(\text{ПТГ}) = -0.0344 * 25(OH)D + 4.4837\,,$$

при $P > 24.7$

$$log(\text{ПТГ}) = 0.0045 * 25(OH)D + 3.5274$$

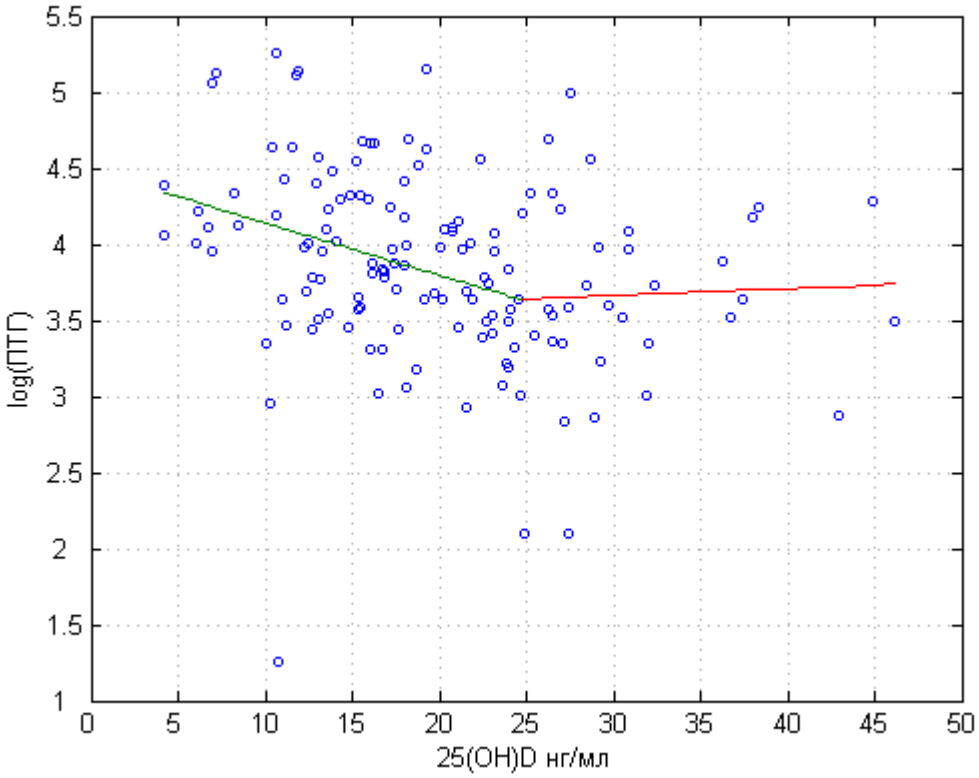

Рис.13. Зависимость ПТГ от 25(OH)D. Красная линия изображает оптимальную регрессионную зависимость справа от точки $P = 24.7$. *Зелёная линия изображает оптимальную регрессионную зависимость слева от точки* $P = 24.7$. *Обе линии составляют общую найденную зависимость.*

3.2. Непрерывная кусочно-линейная модель с изломом в точке $P = 24.7$.

При $P \leq 24.7$ данная зависимость выражается формулой

$$log(\text{ПТГ}) = -0.0342 * 25(OH)D + 4.4812\,,$$

при $\boldsymbol{P} > 24.7$

$$log(\text{ПТГ}) = 0.0049 * 25(OH)D + 3.5146$$

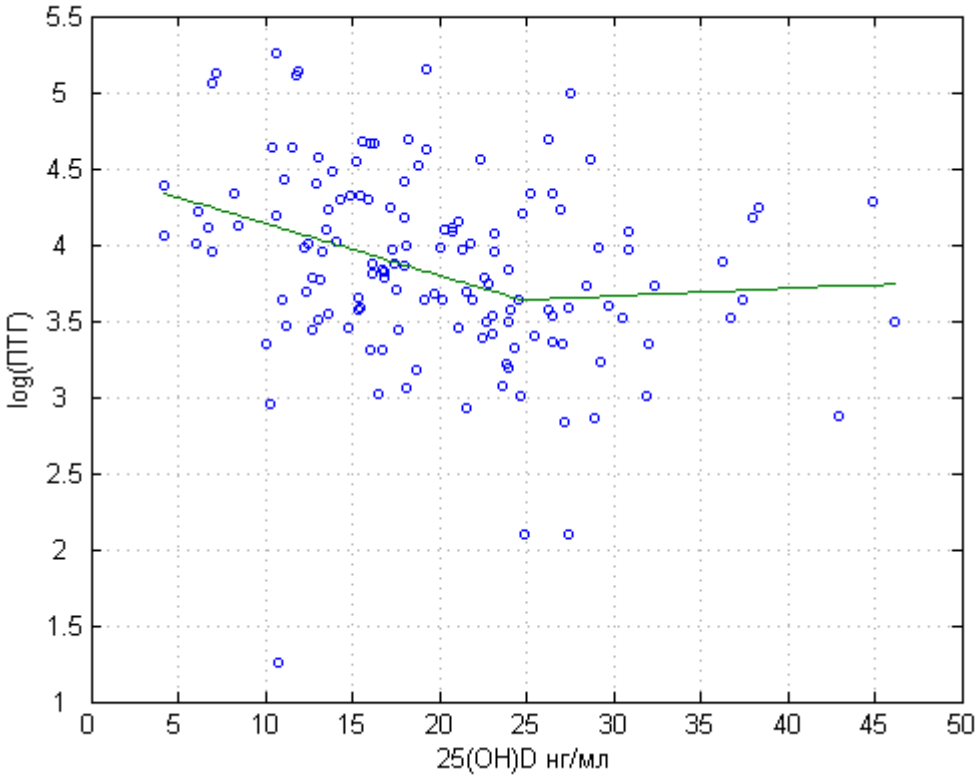

Рис.14. Зависимость log(ПТГ) от 25(OH)D. Зелёная линия изображает непрерывную кусочно-линейную регрессионную зависимость с точкой излома в $\boldsymbol{P} = 24.7$.

3.3. Кусочно-линейная модель с разрывом в точке $\boldsymbol{P} = 23.95$.

При $\boldsymbol{P} \leq 23.95$ данная зависимость выражается формулой

$$\text{ПТГ} = -2.2093 * 25(OH)D + 96.2457\,,$$

при $\boldsymbol{P} > 23.95$

$$\text{ПТГ} = 0.1687 * 25(OH)D + 39.8172$$

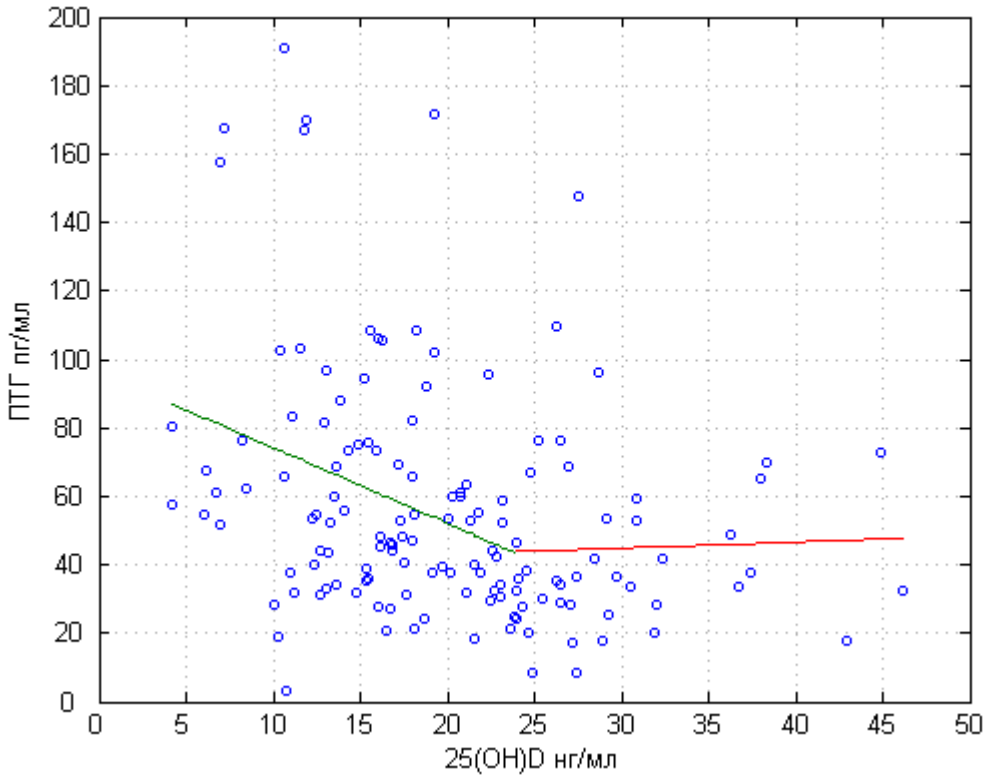

Рис.15. Зависимость ПТГ от 25(OH)D. Красная линия изображает оптимальную регрессионную зависимость справа от точки $\boldsymbol{P} = 23.95$. *Зелёная линия изображает оптимальную регрессионную зависимость слева от точки* $\boldsymbol{P} = 23.95$. *Обе линии составляют общую найденную зависимость.*

3.4. Непрерывная кусочно-линейная модель с изломом в точке $\boldsymbol{P} = 23.95$.

При $\boldsymbol{P} \leq 23.95$ данная зависимость выражается формулой

$$\text{ПТГ} = -2.1879 * 25(OH)D + 95.9713\,,$$

при $\boldsymbol{P} > 23.95$

$$\text{ПТГ} = 0.1938 * 25(OH)D + 38.9282$$

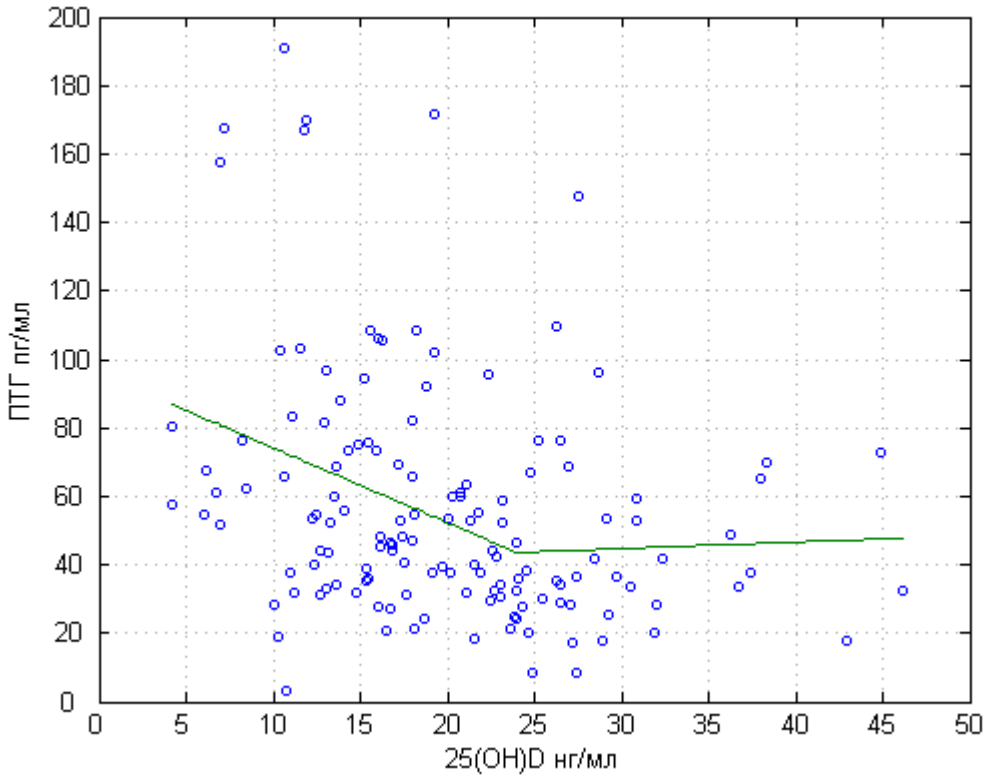

Рис. 16. Зависимость log(ПТГ) от 25(OH)D. Зелёная линия изображает непрерывную кусочно-линейную регрессионную зависимость с точкой излома в $\boldsymbol{P} = 23.95$.

Численные значения коэффициентов всех описанных зависимостей представлены до 4 знака точности после запятой.

3.5. Статистическая значимость регрессионных моделей.

Проверим статистическую значимость найденных зависимостей с помощью перестановочных тестов. Рассмотрим нулевую гипотезу $\boldsymbol{H}_0$ о независимости значений *log(ПТГ)* от *25(OH)D* в случае $\boldsymbol{P} = 24.7$ и значений *ПТГ* от *25(OH)D* в случае $\boldsymbol{P} = 23.95$. В качестве статистики $T(\tilde{S})$, характеризующей отклонение данных, естественно использовать величину невязки, так как оба рассматриваемых типа регрессионных моделей строятся на основе её минимизации. Поскольку общее число возможных перестановок *y*-частей наблюдений $N_0^p = n! = 139!$ слишком велико для полного перебора,

используется случайная генерация 10^6 перестановок и вычисление доли из них, для которой $T(\tilde{S}') > T(\tilde{S}_0)$.

P	*Тип модели*	***p***
24.7	С разрывом, (в кач. зав. переменной *log(ПТГ)*)	0.000072
24.7	Непрерывная, (в кач. зав. переменной *log(ПТГ)*)	0.000079
23.95	С разрывом	0.000041
23.95	Непрерывная	0.000037

Таб.2. Статистическая значимость кусочно-линейных регрессионных моделей различных типов для двух предполагаемых точек выхода зависимости на плато.

Во всех случаях полученная статистическая значимость $p < 0{,}0001$. То есть можно с уверенностью говорить о наличии соответствующих зависимостей. Однако возникает вопрос: насколько необходима именно кусочно-линейная модель в той или иной форме? И не является ли физиологическая зависимость, лежащая в основе данных наблюдений, целиком связанной с более простой линейной моделью? Ответ на этот вопрос может дать новый предложенный метод верификации.

4. Обоснование необходимости выбора кусочно-линейной модели.

Будем исходить из декларированного переформулированного принципа бритвы Оккама: «не следует использовать более сложную модель, если более простая модель исчерпывающе описывает зависимость».

Как показано в предыдущем разделе: данные с более чем достаточной статистической достоверностью описываются различными кусочно-линейными моделями. Более простым вариантом является лишь линейная регрессионная модель. Рассмотрим два способа построения линейной регрессионной модели: оптимальную линейную регрессионную модель и линейную регрессионную модель, ближайшую, в некотором смысле, к верифицируемой кусочно-линейной модели. Также, в качестве альтернативы со сравнимой с кусочно-линейной моделью сложностью, рассмотрим оптимальную квадратичную регрессионную модель.

4.1.Оптимальная линейная регрессионная модель.

Оптимальная линейная регрессионная модель строится с помощью метода наименьших квадратов. Для рассматриваемой задачи она выражается формулой

$$\text{ПТГ} = -1.2406 * 25(OH)D + 81.2281\,,$$

в случае же модифицированной задачи с заменённой зависимой переменной

$$log(\text{ПТГ}) = -0.0198 * 25(OH)D + 4.2553$$

4.2. Линейная регрессионная модель, ближайшая к кусочно-линейной модели.

Пусть имеется некоторая известная кусочно-линейная модель с двумя линейными сегментами разделёнными точкой $\boldsymbol{P}$. И пусть при $x \leq \boldsymbol{P}$ она выражается формулой

$$l_l(x) = \alpha_0^0 + \alpha_1^0 * x \ (14),$$

а при $x > \boldsymbol{P}$

$$l_r(x) = \alpha_0^1 + \alpha_1^1 * x \ (15)$$

Тогда ближайшей к заданной модели будем считать прямую

$$L(x) = A_0 + A_1 * x \ (16),$$

для которой достигается минимум функции

$$F(A_0, A_1) = \int\limits_{x \leq \boldsymbol{P}} \big(l_l(x) - L(x)\big)^2 \, dx + \int\limits_{x > \boldsymbol{P}} \big(l_r(x) - L(x)\big)^2 \, dx \ (17)$$

В конкретной задаче с конечной выборкой, будем считать, что область определения $l_l(x)$ задаётся интервалом $[x_{min}, \boldsymbol{P}]$, где x_{min} – минимальное значение x на всей выборке, а область определения $l_r(x)$ задаётся интервалом $[\boldsymbol{P}, x_{max}]$, где x_{max} – максимальное значение x на всей выборке. Необходимым условием минимума функции $F(A_0, A_1)$ является равенство нулю её частных производных по A_0 и A_1. Таким образом, после преобразований задача сводится к системе из двух линейных уравнений, которая имеет аналитическое решение.

Ближайшая линейная регрессионная модель строится для каждой из четырёх верифицируемых кусочно-линейных регрессионных моделей. В случае кусочно-линейной регрессионной модели с разрывом в точке $\boldsymbol{P} = 24.7$ ближайшая к ней модель выражается формулой

$$log(\text{ПТГ}) = -0.0140 * 25(OH)D + 4.1884 \, ,$$

в случае же непрерывной кусочно-линейной модели с изломом в этой же точке ближайшая модель выражается как

$$log(\text{ПТГ}) = -0.0159 * 25(OH)D + 4.2180$$

В случае кусочно-линейной регрессионной модели с разрывом в точке $\boldsymbol{P} = 23.95$ ближайшая к ней модель выражается формулой

$$\text{ПТГ} = -0.8995 * 25(OH)D + 77.5042\,,$$

в случае непрерывной кусочно-линейной модели с изломом в этой же точке

$$\text{ПТГ} = -0.9785 * 25(OH)D + 78.7488$$

4.3. Оптимальная квадратичная регрессионная модель.

Оптимальная квадратичная регрессионная модель строится с помощью метода наименьших квадратов. Для рассматриваемой задачи она выражается формулой

$$\text{ПТГ} = 0.0532 * (25(OH)D)^2 - 3.6502 * 25(OH)D + 104.4782\,,$$

в случае же модифицированной задачи с заменённой зависимой переменной

$$log(\text{ПТГ}) = 0.0008 * (25(OH)D)^2 - 0.0564 * 25(OH)D + 4.6081$$

Все численные коэффициенты моделей представлены до 4 знака точности после запятой.

4.4. Верификация кусочно-линейной регрессионной модели относительно альтернативных моделей.

Для определения статистической значимости верифицируемых моделей относительно альтернативных моделей будем использовать предложенный метод, основанный на модифицированном перестановочном тесте. Проверке подлежат четыре модели, описанные в разделе 3, относительно трёх типов моделей, описанных выше. В каждом случае рассматривается нулевая

гипотеза $\boldsymbol{H}_0$ о независимости от значения *25(OH)D* остатков значений *ПТГ* или *log(ПГТ)* при вычете значений альтернативной регрессионной модели. В качестве статистики $T(\tilde{S})$, характеризующей отклонение данных, используется величина невязки, так как все четыре верифицируемых регрессионных модели построены на основе её минимизации. Поскольку общее число возможных перестановок остатков наблюдений $N_0^p = n! = 139!$ слишком велико для полного перебора, используется случайная генерация 10^6 перестановок и вычисление доли из них, для которой $T(\tilde{S}') > T(\tilde{S}_0)$. Полученные в результате *p*-значения занесены в таблицу.

Альтернативная модель / Верифицируемая модель	Оптимальная линейная модель	Ближайшая линейная модель	Оптимальная квадратичная модель
Кусочно-линейная модель с разрывом в точке 24.7 (в кач. зав. переменной *log(ПТГ)*)	0.0253	0.0153	0.1070
Непрерывная кусочно-линейная модель с изломом в точке 24.7 (в кач. зав. переменной *log(ПТГ)*)	0.0260	0.0218	0.1057
Кусочно-линейная модель с разрывом в точке 23.95	0.0222	0.0118	0.1371
Непрерывная кусочно-линейная модель с изломом в точке 23.95	0.0220	0.0150	0.1418

Таб.3. Статистическая значимость различных кусочно-линейных регрессионных моделей относительно альтернативных моделей нескольких типов.

Все *p*-значения полученные при верификации кусочно-линейных моделей относительно линейных моделей не превышают 0.05, это значит, что гипотезы о том, что соответствующие линейные модели исчерпывающе описывают зависимость в статистическом смысле, надёжно опровергаются

всеми представленными кусочно-линейными моделями. В тоже время с помощью кусочно-линейных моделей не удаётся опровергнуть гипотезу о том, что данную зависимость исчерпывающе в статистическом смысле описывает оптимальная квадратичная модель.

Тогда аналогичным образом проверим: может ли оптимальная квадратичная модель, опровергнуть гипотезу о том, что данная зависимость исчерпывающе описывается оптимальной линейной регрессионной моделью. Задача рассматривается в двух вариантах: с исходной выборкой и модифицированной выборкой, где в качестве значений зависимой переменной вместо *ПГТ* используется *log(ПГТ)*.

Тип выборки	*p*
Исходная выборка	0.0418
Модифицированная выборка (в кач. зав. переменной *log(ПТГ)*)	0.0533

Таб.4. Статистическая значимость оптимальной квадратичной регрессионной модели относительно оптимальной линейной модели.

Для исходной выборки $p < 0.05$, то есть в этом случае оптимальная квадратичная регрессионная модель может считаться полноценной альтернативой любой из представленных кусочно-линейных моделей. Однако для модифицированной выборки $p > 0.05$, значит гипотеза о том, что оптимальная линейная модель исчерпывающе в статистическом смысле описывает зависимость не отвергается, и, следовательно, в этом случае нет необходимости применения более сложной квадратичной модели вместо линейной.

5.Заключение

Был предложен новый метод селекции регрессионных моделей с учетом их сложности, использующий принцип бритвы Оккама и перестановочные тесты. В основе метода лежит предположение о необходимости использование более сложной модели только в тех случаях, когда зависимости не могут быть исчерпывающе описаны с помощью простых моделей. Считается, что модель исчерпывающее описывает зависимость, если остатки являются случайными одинаково распределёнными независимыми величинами. Нулевая гипотеза об исчерпывающем описании проверяется на семействе выборок, полученных из исходной путём случайной перестановки остатков, с использованием тестируемой сложной модели. Разработанный метод был использован для построения оптимальной модели зависимости концентрации паратиреоидного гормона от содержания витамина D.

6. Литература

[1] Сенько О.В., «Методы эмпирического прогнозирования, основанные на устойчивых разбиениях и коллективных решениях», (стр. 130-135), (2006)

[2] Pedro Domingos, «The Role Of Occam's Razor In Knowledge Discovery», Data Mining and Knowledge Discovery, vol. 3, iss. 4, (pp. 409-425), (1999)

[3] Guilsun Kim, Ki Won Oh, Eun-Hee Jang, Mee-Kyoung Kim, Dong-Jun Lim, Hyuk Sang Kwon, Ki-Hyun Baek, Kun-Ho Yoon, Won Chul Lee, Bong Yun Cha, Kwang-Woo Lee, Ho-Young Son and Moo-Il Kang, «Relationship between Vitamin D, Parathyroid Hormone, and Bone Mineral Density in Elderly Koreans», Journal Of Korean Medical Science, vol. 27, iss. 6, (pp. 636-643), (2012)

[4] H. Akaike, «A new look at the statistical model identification», IEEE Transactions on Automatic Control, vol.19, iss. 6, (pp. 716–723), (1974)

[5] G. Schwarz, «Estimating the dimension of a model», Annals of Statistics, vol. 6, (pp. 461–464), (1978)

[6] E.J. Hannan, B.G. Quinn, «The Determination of the Order of an Autoregression», Journal of the Royal Statistical Society, Series B (Methodological), vol.41, (pp. 190–195), (1979)

[7] J. Rissanen, «Modeling by shortest data description», Automatica, vol. 14, iss. 5, (pp. 465–658), (1978)

[8] Г. Крамер, «Математические методы статистики», 2-е издание, изд. «Мир», Москва, (1975)

[9] К.Ю. Богачёв, «Практикум на ЭВМ. Методы приближения функций», изд. МГУ им. Ломоносова, Москва, (1998)