Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение

высшего образования

**Пермский национальный исследовательский**

**политехнический университет**

Факультет прикладной математики и механики

Кафедра «Математическое моделирование систем и процессов»

Направление 01.03.02 «Прикладная математика и информатика»

Профиль программы бакалавриата «Математическое моделирование»

**КУРСОВАЯ РАБОТА**

по дисциплине **«ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА»**

на тему

**«Регрессионные модели как инструмент анализа и прогнозирования явлений. Линейная регрессия и её парадоксы»**

Выполнил студент группы ММ-19-2б

Мельников Демид Леонидович

(Фамилия, Имя, Отчество полностью)

Проверил:

канд. Ф.-м. наук, доц. кафедры ММСП

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (Останина Т.В.)

(подпись) (ФИО)

Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2022г.

**Пермь 2022**

Содержание

[**Введение** 3](#_Toc105773804)

[**1. Подгонка прямой и Метод наименьших квадратов** 4](#_Toc105773805)

[**2. Линейная регрессионная модель** 6](#_Toc105773806)

[**2.1 Задача линейной регрессии** 8](#_Toc105773807)

[**3. Парадоксы регрессии** 10](#_Toc105773808)

[**3.1 Неоднородность данных** 10](#_Toc105773809)

[**3.2 Коррелированность предикторов** 11](#_Toc105773810)

[**3.3 Неадекватность модели** 11](#_Toc105773811)

[**3.4 Скрытый фактор** 14](#_Toc105773812)

[**Заключение** 15](#_Toc105773813)

[**Список использованных источников** 16](#_Toc105773814)

[**Приложение** 17](#_Toc105773815)

# **Введение**

Регрессией называют любую функциональную зависимость между случайными величинами, даже в тех случаях, когда измеряемые переменными являются неслучайными.

Регрессионный анализ – статистический аналитический метод, позволяющий вычислить предполагаемые отношения между зависимой переменной (одной или нескольких). Используя регрессионный анализ можно моделировать отношения между выбранными переменными, а также делать прогнозы на основе модели.

Чаще всего регрессионный анализ используют для решения таких задач как:

* Выявление конкретной независимой переменной, которая связана с зависимой.
* Выяснение отношений между зависимой и независимой переменной
* Предсказание неизвестных значений зависимой переменной

Данными в регрессионном анализе служит таблица экспериментально полученных зашумленных значений на разных наборах.

Основной целью является как можно точный прогноз на основе измеряемых (предикаторных) переменных.

# **1. Подгонка прямой и Метод наименьших квадратов**

В основе регрессионного анализа лежит подгонка пробной прямой, когда под облако (область) экспериментальных точек , полученных в соответствии с моделью:

 (1.1)

Здесь коэффициенты прямой a и b – неизвестные параметры, - подсчитанные (неслучайные) значения, - независимые и одинаково распределённые случайные ошибки, такие что .

Чаще всего для нахождения оценок коэффициентов a b и применяется метод наименьших квадратов (МНК). Суть этого метода заключается в подборе таких значений коэффициентов, чтобы значения функции было максимально близким к табличным значениями, полученным в ходе эксперимента.

Естественным условием точности подгонки прямой (1.1) служит близость к нулю всех остатков .Общую меру близости можно выбрать по-разному, но наиболее простые формулы для оценок a и b получаются, если в качестве такой меры взять:

 (1.2)

Минимум F достигается в точке , где

 (1.3)

МНК оценки обладают одним важным свойством: в случае, если вектор ошибок - нормально распределён, МНК оценка совпадает с методом максимального правдоподобия и, следовательно, является наиболее точным для больших выборок (асимптотически эффективным). А т.к. мы и задавали вектор ошибок таким образом для подгонки нашей прямой, в рассматриваемой задаче оценка, полученная методом МНК, всегда будет наиболее точной

Недостатком МНК является излишняя чувствительность МНК-оценок к выделяющимся наблюдениям, т.к. мера F зависит не от самих остатков ,а от их квадратов. Стремление уменьшить остатки в точках выбросов может привести к значительному смещению оценок параметров.

# **2. Линейная регрессионная модель**

Предположим, что (с точностью до случайных ошибок) целевая переменная Y есть линейная комбинация вида для предикаторных переменных с неизвестными коэффициентами 

Измеренная переменной Y, отвечающее заданным (не обязательно различным) значениям пердикаторных переменных, имеют вид:

 (2.1)

Можно записать модель в матричной форме:

 (2.2)

Где матрица X – матрица плана эксперимента.

Будем предполагать, что для модели (2.2) выполняются следующие допущения:

1. Столбцы матрицы X являются линейно независимыми. Т.е. в виду выполнения условия матрица X будет иметь ранг m
2. Случайные величины одинаково распределены с  и (параметр является известным) и некоррелированные: 
3. Вектор  имеет нормальное распределение

А также из выводов об оценки параметров с помощью минимизации функции [1], c учётом вышеприведённых допущений, получим:

1. Точка минимума функции является и называется МНК-оценкой.
2. Значение в точке минимума называется остаточной суммой квадратов.
3. Представление для  можно записать как . Матрицу B называют информационной.
4. Если верны Допущения 1-2, то при любом векторе оценкаимеет минимальную дисперсию в классе линейных несмещённых оценок. Если также выполняется Допущение 3, то этот класс можно расширить до множества произвольных несмещённых оценок.
5. Оценка является несмещённой, т.е. 
6. Для любого вектора несмещённой оценкой для величины служит, причём
7. Случайная величина  имеет распределение хи-квадрат с n-m степенями свободы и не зависит от оценки  и, поскольку , то статистика оценивает неизвестную дисперсию (оценка дисперсии остатков)
8. Для любого вектора случайная величинараспределена по закону Стьюдента с n-m степенями свободы.

Теперь у нас есть все данные, чтобы рассмотреть следующую задачу

## **2.1 Задача линейной регрессии**

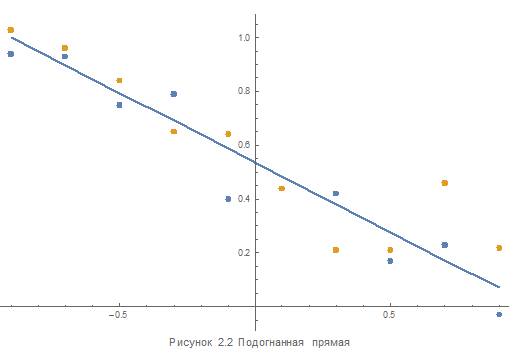
Данные в таблице (Рис.2.1) представляют собой зашумлённые значения линейной зависимости Y=a+bX. Ошибки моделировались с помощью датчика нормальных чисел. В каждом из узлов ошибки добавлялись дважды (поэтому матрица Х состоит из двух столбцов длины 20). Кроме того, одно из значений (1.1) было заменено на выброс



Задача состоит в вычислении МНК-оценок неизвестных параметров a, b и, обнаружении выброса и построении 95%-ых доверительных интервалов для параметров и значений прямой в точке предполагаемого выброса.

 Для решения этой задачи был реализован метод наименьших квадратов в среде Python (код приведён в приложении, все графики построены с помощью мат.пакета Wolfram Mathematica 12). С его помощью найдём и  и построим подогнанную прямая (Рис. 2.2):

(2.3)



Из таблицы для хи-квадрат 18 находим квантили и . Для параметра  получаем доверительный интервал:



Из таблицы для коэффициентов Стьюдента берём . Найдём границы 95%-ых доверительных интервалов для a и b:



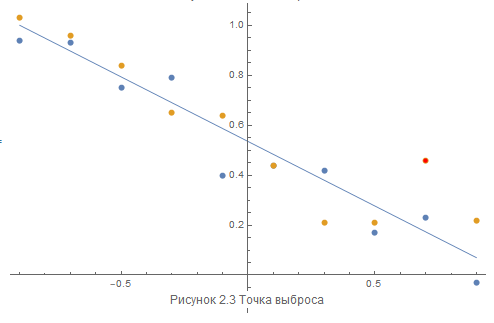
Найдём величину наибольшего остатка и оценку:



Узнаем, что этот остаток советует x (9). Получим интервал для значения прямой при таком х:



Поскольку одно из значений x (9) находится вблизи границы 0.27+0.22 = 0.49, то оно, вероятно, является выбросом (Рис. 2.3).



# **3. Парадоксы регрессии**

Есть несколько типичных ошибок (тонких мест), которые следует учитывать, применяя регрессионный анализ. Часто о низ забывают при работе реальными данными и в результате получаются неверные выводы.

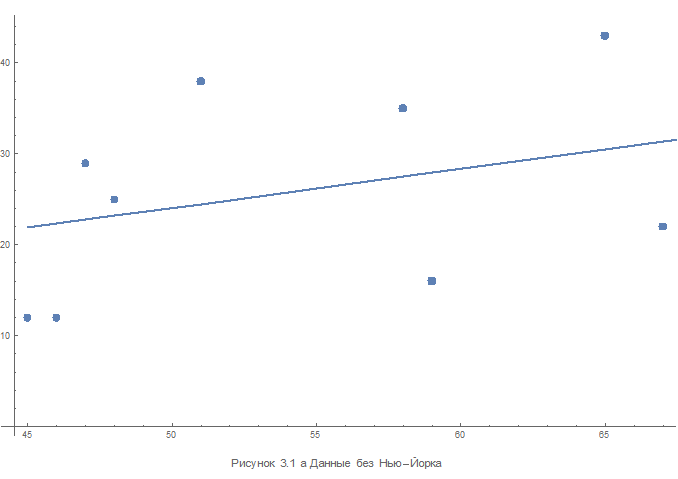
## **3.1 Неоднородность данных**

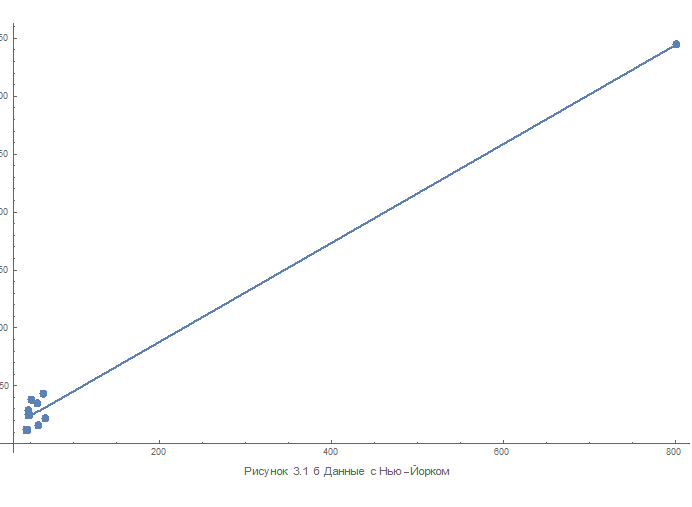
На рис. 3.1 *а* представлены данные о числе телевизионных точек Y [3] (в десят. тысяч), установленных в 1953г. в девяти городах США (Деневере, Сан-Антонио, Канзас-Сити, Сиэтле, Цинцинати, Буффало, Нью-Орлеане, Мулуоки, Хьюстоне) и о численности населения Х (в десят. тысяч) в этих городах.

Коэффициент корреляции между двумя наборами, что при выборке из 9 городов говорит нам о том, что между X и Y линейная связность достаточно мала.

Но если же к этим данным добавить соответствующие сведения о Нью-Йорке. Пересчитанный для десяти городов коэффициент корреляции  На рис. 3.2 *б* проведена регрессионная прямая.

Такой результат происходит из-за большой разницы в значениях X и Y у Нью-Йорка.





## **3.2 Коррелированность предикторов**

В случае, когда регрессионная модель включает много предикторов, некоторые из них могут оказаться приблизительно линейно связными между собой.

Сильная коррелированность предикторов опасна тем, что приводит к неустойчивости МНК-оценок. Это происходит из-за того, что столбцы матрицы Х оказываются практически линейно зависимыми, вследствие чего матрица B становится почти вырожденной, а задача поиска решений – плохо обусловленной.

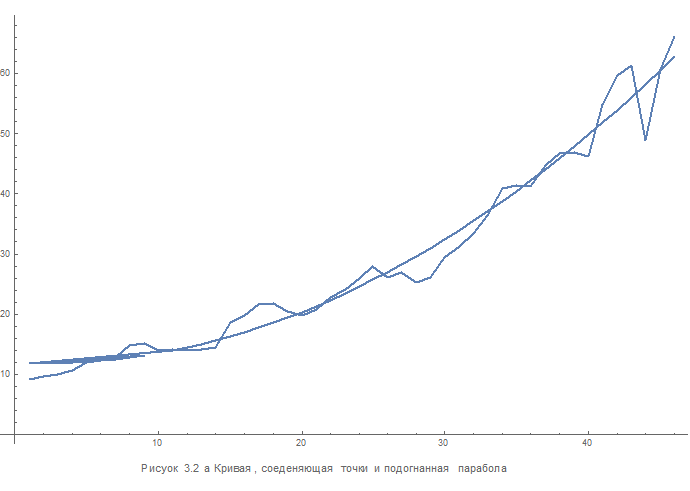
## **3.3 Неадекватность модели**

Простейшая линейная зависимость (2.1) может неадекватно описывать приведённые данные, т.е. когда между экспериментальными данными существует зависимость сложнее линейной.

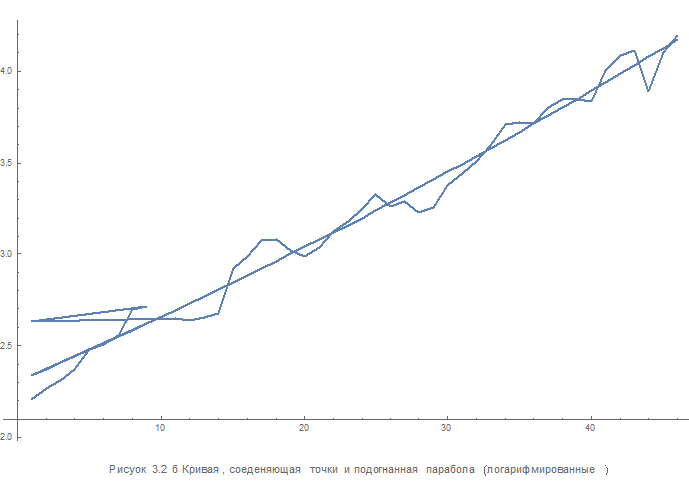
Для построения более сложной модели обычно пытаются отдельно изучить влияние каждого предиктора X на отклик Y. Для этого сглаживают двумерную область точек при помощь некоторой монотонной нелинейной функции.

Приведём пример такого парадокса [2]

В таблице на рис. 3.3 для каждого из годов T указанно количество чугуна Y (млн. тонн), которое выплавлялось за год по всему миру. Постараемся подогнать регрессионную кривую к этим данным.

Пусть Х = T-1864. На рис. 3.1 изображена кривая, соединяющая точки плоскости. На первый взгляд кажется, что эта кривая поднимается вверх значительно быстрее, чем прямая или квадратная парабола. Попытаемся сгладить кривую с помощью параболы. На этом же рисунке приведён график подогнанной параболы с помощью МНК. При этом стандартное отклонение  и остаточная сумма квадратов rss = 364.7

Поскольку заметная некоторая тенденция усиления колебаний кривой относительно параболы с ростом X. Применим преобразования Y’ = lnY. На рис. 3.2 б построена кривая, соединяющая точки. Она хорошо согласуется с подогнанной МНК-прямой с коэффициентами a = 2.203, b = 0.0413



Однако, если попытаться использовать последнюю подгонку для прогноза Y на основе формулы , то получим значение rss = 381.8, которое несколько больше, чем вычисленное ранее при помощи подгонки параболы.

Можно считать, что подгонки одинаковые по точности.

Следует отметить, что в случае ошибки при выборе типа сглаживающей кривой результаты экстраполяции могут оказаться совершенно неудовлетворительными.

## **3.4 Скрытый фактор**

Желание истолковать регрессионную модель связь как причинно-следственную может привести к парадоксам, как в нижеприведённых примерах [3].

Во время второй мировой войны англичане исследовали зависимость точности бомбометания Z от ряда факторов, в число которых входили высота бомбардировщика H, скорость ветра V, количество истребителей противника X. Как и ожидалось, Z увеличивалась при уменьшении H и V. Однако (что поначалу представлялось необъяснимым), точность бомбометания Z возрастала также и при увеличении X.

Дальнейший анализ позволил понять причину этого парадокса. Оказалось, что первоначально в модели не был включён такой фактор как Y -облачность. Он сильно влияет на Z (уменьшая точность), и на X (бессмысленно высылать истребители, если ничего не видно). Сильные отрицательные причинно-следственные связи на парах (Y, Z), (X, Y) привели к появлению положительного коэффициента при X в линейной регрессионной модели для Z.

# **Заключение**

В работе была рассмотрена постановка регрессионной модели, решена примерная задача, рассмотренные основные парадоксы регрессии.

Исследование в данной области поможет в изучении математической статистики.

# **Список использованных источников**

1. М.Б,Лагутин, Учебное пособие “Наглядная математическая статистика”, БИНОМ Лаборатория знаний, Москва, 2009г. – 474с.
2. Г.Секей, “Парадоксы в теории вероятностей и математической статистике”, М: Мир, 1990г. – 249с.
3. Ф.Миллс, “Статистические методы”, Государственное статистическое издание, Москва, 1958г. – 881с.

# **Приложение**

Ниже приведена частичная реализация метода наименьших квадратов

while k != 2:  
 if k == 0:  
 Y1 = sum(y1) / n  
 else:  
 Y1 = sum(y2) / n  
  
 for j in range(n):  
 if k == 0:  
 temp\_xy += [x0[j] \* y1[j]]  
 else:  
 temp\_xy += [x0[j] \* y2[j]]  
  
 temp\_sigma += [(x0[j] - x) \*\* 2]  
  
 XY = sum(temp\_xy) / n  
 sigma = sum(temp\_sigma) / n  
  
 b = (XY - x \* Y1) / sigma  
 a = Y1 - b \* x  
 A += [a]  
 B += [b]  
  
 k += 1  
k = 0  
a = sum(A) / 2  
b = sum(B) / 2

temp\_rss = []  
while k != 2:  
 if k == 0:  
 for i in range(n):  
 temp\_rss += [(y1[i] - a - b \* x0[i]) \*\* 2]  
 else:  
 for i in range(n):  
 temp\_rss += [(y2[i] - a - b \* x0[i]) \*\* 2]  
  
 k += 1  
  
k = 0  
rss = sum(temp\_rss)  
  
sigma = (rss / 18) \*\* 0.5  
  
print(f'rss = {rss}; sigma = {sigma}')  
  
z0 = 8.23  
z1 = 31.5  
  
ss = [(rss / z0) \*\* 0.5, (rss / z1) \*\* 0.5]  
  
ss = sorted(ss)  
  
print(f'Доверительный интевал для sigma: {ss}')