# Использование метода линейной регрессии для анализа практических моделей

10 декабря 2023 г.

`https://www.overleaf.com/read/mnhfdtpjmzrv#889a60` - это ссылка на overleaf

## Введение

Целью этого литературного обзора является изучение сфер применения линейной регрессии, а также её практической пользы.

В настоящее время из-за усложнения реальных моделей, с которыми сталкиваются специалисты многих профессий, всё больше возникает потребность получить более точные функции, описывающие данные модели. Одним из фундаментальных методов является линейная регрессия, которая до сих используется повсеместно. Чтобы убедиться в данном утверждении, изучим исследования разных сфер деятельности, которые непосредственно использовали линейную регрессию для получения более точных данных.

Задачи:

- познакомиться с понятием линейная регрессия
- разобрать простые примеры её примерения
- рассмотреть практическое применение метода

Ключевые слова: линейная регрессия, зависимость, анализ данных, практическое применение

## Метод линейной регрессии

Перед тем как перейти к примерам использования линейной регрессии, уточним значение данного термина. Метод линейной регрессии - это метод

анализа данных, который предсказывает ценность неизвестных данных с помощью другого связанного и известного значения данных. Он математически моделирует неизвестную или зависимую переменную и известную или независимую переменную в виде линейного уравнения.

Например, пусть у нас есть два набора данных $X_t, Y_t$, где $t = 1, ..., n$ [12]. Наша задача подобрать такую функцию $Y = f(X) = a_1x_1 + a_2x_2 + ... + a_nx_n + a_{n+1}$, чтобы значения $y_1, y_2, ..., y_n$ были наиболее близкие. Что такое 'наиболее близкие' - это уже другая задача, связанная с тем, как правильно оценивать точность линейной регрессии. Для простоты примера возьмём самую популярную оценку - это сумма квадратов отклонений: $\sum_{i=1}^{n}(y_i - f(X)_i)^2$. Тогда, можно либо взять и представить все наборы данных, как точки в декартовой системе координат и подвигать линию - функцию $f(X)$, либо воспользоваться разными методами нахождения коэффициентом функции. И непосредственно линейной регрессией будет являться данная функция, а методом линейной регрессии - предсказание значения $Y$ с помощью функции $f(X)$.

## Описание статей

Рассмотрим более реалистичный пример из учебного пособия [12]: пусть I - это реальный доход семьи, а E - это реальные расходы. Несмотря на то, что зависимость расходов и доходов может быть разная, в учебнике приводится достаточно логичное объяснение, почему E - можно найти, через линейную функцию от дохода. (Пример из [12] $E = 4.663 + 0.686I$) Несмотря на факторы, такие как структура расходов и различный состав семьи, отклонения, особенно R - доля объяснённой дисперсии, не будут небольшие, поэтому метод линейной регрессии хорошо подходит для решения задачи - нахождения зависимости между доходами и расходами. Стоит отметить, что несмотря на популярность данного учебного пособия, оно не очень релевантно, так как не содержит много практических задач, но отлично подходит для осознания базовых методов оценки моделей, несмотря на год издания.

Перейдём к другой сфере, которая остаётся актуальна уже несколько десятков лет - это логистика. Одной из мер является VMT - vechicle miles traveled (километраж транспорта (но в некоторых странах вместо километра используют мили, что не играет особой роли. Учебное пособие [2] посвящено изучении. наилучшего метода, предсказывающего VMT относительно других факторов. Авторы [6] утверждают, что линейная регрессия оказалась

самой удобной и поэтому распространённой оценкой данной меры передвижения. В книге также упоминается более модифицированная регрессия - линейна-логарифмическая, а также гетероскедастическая и сравниваются с линейной. Пособие написано очень понятно, несмотря на большое количество терминологии. В тексте также периодически встречаются реальные ситуации, тем самым подчёркивая практичность исследования. Стоит подчеркнуть, что учебное пособие было опубликовано совсем недавно и многие данные, рассмотренных моделях, действительны.

Обратим внимание на модели, использованные в анализе данных. Авторы [10] посвятили своё издание обоснованию выбора метода линейной регрессии для изучения графиков типа box-plot. Они предоставляют точное математическое доказательство возможности получения лучших значений из box-plot по пяти параметрам (минимум, квартили и максимум), которые с правильными коэффициентами предсказывают нужные значения. Более того, они иллюстрируют пример на датасете 2023 года о Бразильском Электронном секторе. При прочтении данной книги не остаётся вопросов, связанных с верностью вычисленный, а свежие данные, взятые из достоверных источников подчёркивают актуальность исследуемой области.

Ещё одной очень распространённой сферой является Биостатистика, особенна касающаяся физиологических данных человека. В пособии [2] изучены соотношения между разными человеческими факторами и биологическими оценками и индексами. Одним из ярких и понятных примеров служит BMI - body mass index (индекс массы тела). Авторы в своей книге описали исследование: была собрана группа женщин с жизненным периодом - постменопаузы. У каждой подопечной был померен вес, возраст, период активности и изучена диета, а также был подсчитан BMI. Изучив зависимость между этими факторами и индексом массы, была выведена линейная зависимость этих факторов. Далее авторы предоставляют подробное описание похожих использований методов линейной регрессии биологами для подсчёта давления в крови и риска диабета. Правда не соответствие со временем проскальзывает сквозь текст (как контраст ко [6]), поэтому данное пособие неактуально, хотя многие фундаментальные вещи в нём хорошо обоснованы. Также стоит отметить, что авторы не вдаются в точные математические выкладки (особенно по сравнению с [10]), из-за чего у читателя возникают подозрения о достоверности информации.

Ещё одной областью применения линейной регрессии является гиперспектральная съёмка - раздел прикладной оптики, который изучает растровые изображения, каждый пиксель которых связан не с отдельным значением

интенсивности света, а с полным спектральным разложением оптической энергии в границах какого-либо частотного диапазона. В данной статье [4] изучается конкретно съёмка песка. Автор рассматривает линейные модели, которыми можно описать кадры такого вида. Несмотря на современность и релевантность статьи, текст написан сложно, особенно для читателя, не являющегося специалистом в оптике. Также статья больше теоретическая, несмотря на один очень точный эксперимент взятый и часто цитируемого источника, поэтому за прикладными примерами лучше обращаться к другим источникам, например [6].

Рассмотрим другую сферу деятельности - сейсмология. В учебном пособии [9] приводится пример линейной оптимизации FWI - sesmistic full waveform inversion (уровень сейсмических волн). Данная статья базируется на событиях 2009 годов и с помощью линейной регрессии (а точнее метода наименьших квадратов) рассчитывает соотношение наблюдаемых и симулированных сейсмических активностей. Несмотря на простые формулы, авторы делают сложные выводы, которые некоторые источники используют до сих пор. В данном пособии тоже сложный для понимания текст, полный профессионализмов. Также, важно заметить, что при этом статья достаточно актуально, в связи с участившимися землетрясениями по всему миру. Именно поэтому некоторые учебные исследования цитируют выводы из этого текста.

Следующее учебное пособие [5] посвящено WQI - water quality index (индекс качества воды). Авторы этой книги пошагово вывели линейную модели подсчёта WQI, а после на реальном примере показали её точность. В статье были предоставлены одни из самых качественных и понятных математических доказательств и построений. Несмотря на редкость данной темы, создатели пособия подобрали достоверный и понятный материал.

Перейдём к одному из самых актуальных направлений - искусственный интеллект. В статье [1] изучается авиационный буровой робот. Задачей данной статьи было показать зависимость качества авиации с аккуратностью бурового робота. Ими был использован метод линейной регрессии для подсчётов других коэффициентов, связанных с аэродинамикой. Достаточно сложный текст для прочтения и цитирования, но при является современным и подходящим для специалистов в аэрофизике. Важно отметить, что метод линейной регрессии не является центральным элементом данного пособия, но достаточно подробно разбирается в некоторых частях статьи.

Очередной пример использования линейной регрессии в логистике и ана-

лизе данных является статья [8]. Авторы описывают модель Pharmaceutical Supply Chain Network Design (Проектирование сети фармацевтической цепочки поставок) для того, чтобы ускорить доставку медикаментов и минимизировать проблемы с транспортировкой. В статье упоминается много разных аспектов и факторов, которые обрабатывают аналитики, но конкретно метод линейной регрессии используется для предсказания вероятности нехватки на складе. Текст доступен для понимания любому читателю, а также предоставляет современные базы данных в своих примерам, поэтому эта статья - отличный источник практических моделей. Но в этом пособии опускаются математические формальности.

Перейдём к ещё одной водной области практического применения метода: передача информации о батиметрии (Изучение рельефа подводной части водных бассейнов: как мирового океана, так и озёр, рек и т. д.) для навигаторов и поисковых систем. Учебное пособие [11] рассказывает о работе SDB - satellite remote sensing (спутниковое дистанционное зондирование) - достаточно популярное и до сих пор разрабатывающееся направление в наше время. Так как информации о структуре рифа хочется получить за более короткий срок и наименьшему затратами, было предложено совместить линейную регрессию и машинное обучение. В статье достаточно подробно описана задача, в которой метод имеет практическое применение. Авторы используют комплексные термины, но большая часть повествования будет понятна широкому кругу читателей. Опять же важно подчеркнуть актуальность и релевантность информации, представленной в данном пособии.

Ещё одному экономическому примеру посвящена стать [3], в которой линейная регрессия и ещё один метод сравниваются по аккуратности подсчёта предсказаний о будущих перфомансах сферы металлургии как индустрии. Авторы пособия строят две разные модели на основе прошлых данных, а потом изучают высчитанные зависимости. В статье очень чётко описывается каждый шаг. Также исследование использует реальную базу данных, взятую из достоверных источников. Авторами используются современные методы машинного обучения для показания более точной оценки обоих методов.

Приведём ещё одно исследование [7] работы некоторой модели, а точнее метеорологической. В данной статье авторы описали собственную линейную модель, которая высчитывает температуру в здании для комфорта жителей. Понятная широкому кругу читателей статья использует чёткие математические выкладки, а также приводит практический пример данной модели. Тематика этого пособия актуальная и релевантная, по этой причине,

к ней обращались другие авторы. Также данная статья централизована на методе линейной регрессии, поэтому из неё получится достать большой объём информации на тему её применения.

Подведём итоги. Метод линейной регрессии используется в огромном диапазоне областей: науке (статьи [2], [4], [9], [5], [7]), финансах и логистике ([12], [6], [3]), аналитике ([10], [8], [11], [7]) и других. Многие специалисты пользуются простотой и точностью данного метода. Приведённые в этом литературном обзоре стать показали много практических применений линейных моделей. Более точными с математической точки зрения статьи являлись [12], [10], [7]. Особо актуальные пособия - это [1], [11], [7]. Многие статьи были написаны для быстрого понимания материала, особенно [12], [2], [5], [8]. Каждый текст содержит реалистичные примеры применения метода, так что многие читатели смогут найти достаточно много информации на заданную тему.

## Список литературы


[1] Dongdong Chen и др. “Positional error compensation for aviation drilling robot based on Bayesian linear regression”. В: (2023).

[2] Vittinghoff Eric и др. “Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models (Statistics for Biology and Health)”. В: (2014), с. 69—138.

[3] Andrea Galeazzi и др. “Predicting the performance of an industrial furnace using Gaussian process and linear regression: A comparison”. В: (2023).

[4] Neal B. Gallagher и др. “Extended least squares (ELS) and generalized least squares (GLS) for clutter suppression in hyperspectral images: A theoretical description”. В: (2023).

[5] Samsul Ariffin Abdul Karim и Nur Fatonah Kamsani. “Water Quality Index Prediction Using Multiple Linear Fuzzy Regression Model”. В: (2022), с. 23—53.

[6] Asif Mahmud. “Estimation of VMT using heteroskedastic log-linear regression models”. В: (2023).

[7] Zoltán Patonai, Richárd Kicsiny и Gábor Géczi. “Multiple linear regression based model for the indoor temperature of mobile containers”. В: (2022).

[8] Shabnam Rekabi, Zeinab Sazvar и Fariba Goodarzian. “A machine learning model with linear and quadratic regression for designing pharmaceutical supply chains with soft time windows and perishable products”. В: (2023).

[9] Qianci Ren. “Seismic acoustic full waveform inversion based on the steepest descent method and simple linear regression analysis”. В: (2022).

[10] Dailys M.A. Reyes Leandro C. Souza Renata M.C.R. de Souza Adriano L.I. de Oliveira. “Parametrized linear regression for boxplot-multivalued data applied to the Brazilian Electric Sectors”. В: (2023).

[11] Pramaditya Wicaksono, Setiawan Djody Harahap и Rani Hendriana. “Satellite-derived bathymetry from WorldView-2 based on linear and machine learning regression in the optically complex shallow water of the coral reef ecosystem of Kemujan island”. В: (2023).

[12] Пересецкий А.А. Магнуса Я.Р. Катышев П.К. “Эконометрика. Начальный курс: 6 издание”. В: (2004), с. 43—58.