ГОУ ВПО

Российско-Армянский Университет

Институт: БИОМЕДИЦИНЫ И ФАРМАЦИИ

Кафедра: Медицинской биохимии и биотехнологии

**Реферат**

**Тема: Линейная регреcсия используя Python**

Выполнил студент: Газарян А.С.

Руководитель (преподаватель): Газазян Э.А.

Ереван 2023  
  
  
Введение:

Линейная регрессия - это метод статистического анализа, который позволяет исследовать связь между двумя переменными: зависимой и независимой. Этот метод широко используется в различных областях, таких как экономика, физика, социология, биология и другие.

В данном реферате мы рассмотрим основные техники линейной регрессии, используя язык программирования Python. Мы рассмотрим корреляцию, std, t-test и slope-intercept, которые являются ключевыми аспектами линейной регрессии.

В первой части реферата мы рассмотрим понятие корреляции и ее использование в линейной регрессии. Затем мы рассмотрим понятие стандартного отклонения (std) и как оно может быть использовано для оценки точности модели линейной регрессии. В третьей части мы изучим t-test, который позволяет проверять значимость параметров линейной регрессии. И, наконец, мы рассмотрим метод slope-intercept для вычисления коэффициентов линейной регрессии.

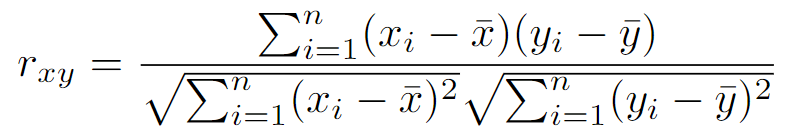
Цель данного реферата - дать читателю полное представление о линейной регрессии и ее ключевых аспектах, а также дать практические навыки использования Python для анализа данных с помощью линейной регрессии.  
  
**1. Корреляция**

Корреляция - это статистическая мера, которая измеряет степень связи между двумя переменными. Корреляция может быть положительной, если две переменные движутся в одном направлении, отрицательной, если они движутся в противоположных направлениях, или равной нулю, если между ними нет связи.

Для измерения корреляции используется коэффициент корреляции Пирсона, который определяется как отношение ковариации между двумя переменными к произведению их стандартных отклонений.

**Формула коэффициента корреляции Пирсона:**

где r\_{xy} - коэффициент корреляции Пирсона, x и y - две переменные, n - количество наблюдений, \bar{x} и \bar{y} - средние значения x и y соответственно.



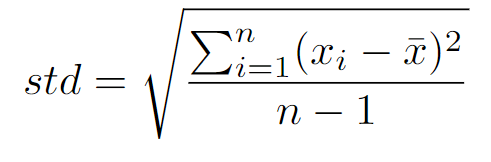
В Python можно использовать библиотеку pandas для вычисления корреляции. Для этого необходимо создать DataFrame, содержащий две переменные, и вызвать метод corr():

**Приложение 1.**  
  
В данном примере мы создаем DataFrame с двумя переменными **x** и **y**, и вычисляем коэффициент корреляции между ними. Результатом будет значение 1.0, что говорит о сильной положительной корреляции между этими переменными.

Корреляция является важным аспектом линейной регрессии, так как модель линейной регрессии использует корреляцию для определения степени связи между зависимой и независимой переменными. Если между ними существует сильная корреляция, то модель линейной регрессии будет более точной и надежной.  
  
**2. STD**

STD или стандартное отклонение - это мера разброса данных вокруг их среднего значения. STD показывает, насколько данные распределены вокруг своего среднего значения. Чем больше STD, тем больше данные распределены вокруг своего среднего значения.

Для вычисления STD используется следующая формула:



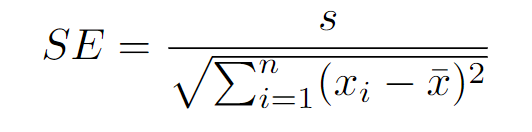
где std - стандартное отклонение, x - данные, n - количество наблюдений, \bar{x} - среднее значение данных.

В Python можно использовать библиотеку numpy для вычисления стандартного отклонения. Для этого необходимо вызвать функцию std():  
**Приложение 2.**  
  
В данном примере мы создаем список данных и вычисляем стандартное отклонение. Результатом будет значение 1.41421356, что означает, что данные распределены вокруг своего среднего значения с разбросом в 1.41.

STD является важным аспектом линейной регрессии, так как она используется для оценки точности модели линейной регрессии. Низкое значение STD означает, что данные имеют маленький разброс вокруг своего среднего значения, что указывает на более точную модель линейной регрессии. Высокое значение STD, наоборот, указывает на большой разброс данных и, следовательно, на менее точную модель линейной регрессии.  
  
  
**3. T-Test**

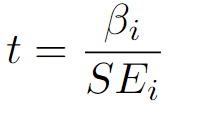
t-test - это статистический тест, который используется для проверки значимости параметров модели линейной регрессии. t-test позволяет определить, насколько значимо коэффициенты регрессии отличаются от нуля.

Для выполнения t-test необходимо вычислить стандартную ошибку (SE) для каждого коэффициента регрессии, которая определяется следующей формулой:



где s - оценка стандартного отклонения, x - данные, n - количество наблюдений, \bar{x} - среднее значение данных.

Затем можно вычислить t-статистику для каждого коэффициента, используя следующую формулу:



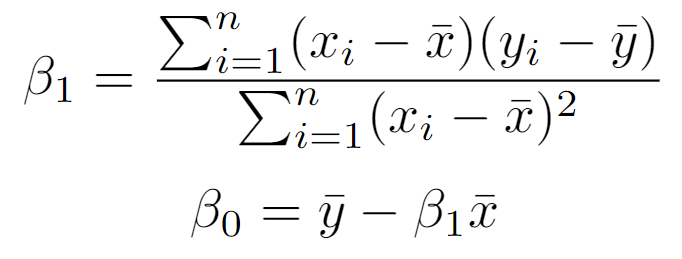
где t - t-статистика, \beta\_i - коэффициент регрессии, SE\_i - стандартная ошибка.

В Python можно использовать библиотеку statsmodels для выполнения t-test. Для этого необходимо создать модель линейной регрессии и вызвать метод summary():  
  
**Приложение 3.**  
  
В данном примере мы создаем DataFrame с двумя переменными x и y , добавляем константу к переменной x , создаем модель линейной регрессии и вызываем метод summary() для выполнения t-test. Результатом будет таблица, содержащая информацию о каждом коэффициенте регрессии, стандартной ошибке, t-статистике и p-значении.

t-test является важным аспектом линейной регрессии, так как он позволяет определить значимость параметров модели. Если p-значение меньше 0.05, то коэффициент регрессии считается значимым и его можно использовать в модели. Если же p-значение больше 0.05, то коэффициент регрессии не является значимым и его можно исключить из модели.  
  
**4. Slope-intercept**

slope-intercept - это метод вычисления коэффициентов линейной регрессии в модели y = \beta\_0 + \beta\_1x , где y - зависимая переменная, x - независимая переменная, \beta\_0 - свободный член, \beta\_1 - коэффициент наклона.

Для вычисления коэффициентов линейной регрессии по методу slope-intercept необходимо использовать следующие формулы:



где \beta\_1 - коэффициент наклона, \beta\_0 - свободный член, x и y - две переменные, n - количество наблюдений, \bar{x} и \bar{y} - средние значения x и y соответственно.

В Python можно использовать библиотеку numpy для вычисления коэффициентов линейной регрессии по методу slope-intercept. Для этого необходимо создать массивы данных, вычислить средние значения x и y , а затем применить формулы для вычисления коэффициентов:  
**Приложение 4.**

В данном примере мы создаем массивы данных x и y , вычисляем их средние значения, и применяем формулы для вычисления коэффициентов. Результатом будут значения коэффициентов \beta\_1 и \beta\_0 .

Метод slope-intercept является одним из наиболее распространенных методов вычисления коэффициентов линейной регрессии. Он прост в использовании и понимании, и может быть применен для моделей с одной независимой переменной.  
  
**Заключение:**

Линейная регрессия - это мощный инструмент статистического анализа, который позволяет исследовать связь между двумя переменными: зависимой и независимой. Мы рассмотрели четыре ключевых аспекта линейной регрессии - корреляцию, стандартное отклонение (STD), t-тест и метод slope-intercept.

Корреляция позволяет определить степень связи между двумя переменными, что является важным аспектом линейной регрессии. STD используется для оценки точности модели линейной регрессии, а t-тест позволяет проверить значимость параметров модели. Метод slope-intercept - это один из наиболее распространенных методов вычисления коэффициентов линейной регрессии.

Мы также показали, как можно использовать язык программирования Python для выполнения анализа данных с помощью линейной регрессии. Библиотеки pandas, numpy и statsmodels позволяют быстро и удобно выполнить вычисления корреляции, STD, t-теста и метода slope-intercept.

В целом, линейная регрессия является важным инструментом для анализа данных в различных областях, и понимание ее основных аспектов может помочь в принятии более обоснованных решений на основе данных.

**Пример**  
В качестве примера проанализируем курс валют USD - AMD за 20 дней на момент февраля 2023 года, и предположим котировку на 21 февраля.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 395,62 | 12 | 394,9 |
| 2 | 394,48 | 13 | 394,84 |
| 3 | 395,73 | 14 | 394,66 |
| 4 | 396,52 | 15 | 394,41 |
| 5 | 396,57 | 16 | 393,46 |
| 6 | 396,64 | 17 | 393,51 |
| 7 | 396,64 | 18 | 393,3 |
| 8 | 395,57 | 19 | 393,11 |
| 9 | 395,84 | 20 | 392,91 |
| 10 | 395,96 | 21 | x |
| 11 | 395,25 |  |  |

***Приложение 5.***Согласно программе значение x = 393.21452631578944  
На момент 21.02.23 значение котировки было = 392,41