Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Теоретическая информатика и компьютерные технологии»

Домашняя работа №2

по курсу «Моделирование»

«Проверка статистических гипотез»

Студент группы ИУ9-81Б Окутин Денис Алексеевич

Преподаватель Домрачева Анна Борисовна

Москва 2025

Цель работы

Проверить гипотезу о возможности описания стохастической зависимости между переменными двупараметрической степенной функцией связи. Для оценки параметров использовать модифицированный метод моментов, для проверки гипотез - критерий Колмогорова-Смирнова..

Постановка задачи

Для исследования использован датасет, содержащий медицинские показатели пациентов, включая уровень глюкозы (Glucose) и инсулина (Insulin) в крови. Датасет включает 768 наблюдений. Glucose — уровень глюкозы в плазме крови (мг/дл) через 2 часа после приёма глюкозы. Insulin — уровень инсулина в сыворотке крови (мкЕд/мл) через 2 часа после теста. Нулевые значения в данных были скорректированы для возможности логарифмирования. Для реализации работы нужно выполнить следущие шаги:

- Загрузка и предварительная обработка данных;
- Анализ данных и удаление выбросов;
- Оценка параметров и с помощью модифицированного метода моментов;
- Проверка гипотезы о соответствии модели данным с помощью критерия Колмогорова-Смирнова;
- Визуализация результатов.

Теоретические сведения

Стохастическая зависимость характеризуется вероятностным влиянием одной переменной на распределение другой без однозначной детерминации. В отличие от функциональной связи, здесь присутствует случайная компонента. Такие зависимости анализируются методами математической статистики и эконометрики.

Для моделирования нелинейных взаимосвязей используется **двухпараметрическая функция**:

$$y = \alpha \cdot x^{\beta}$$

Линеаризация модели через логарифмирование:

$$\ln y = \ln \alpha + \beta \ln x$$

позволяет применять методы линейной регрессии. Параметр β интерпретируется как эластичность, а α — как масштабный коэффициент.

Обоснование применимости степенной зависимости:

Рассмотрим, что экспоненциальное распределение и S-образные кривые могут быть аппроксимированы функцией Вейбулла:

$$F(y) = 1 - \exp(-ay)$$

Предположим, что стохастическая зависимость между двумя случайными величинами y и x описывается функциональной зависимостью вида:

$$y = y(x) = bx^{c}$$

Тогда функция распределения примет следующий вид:

$$F(x) = F(y(x)) = 1 - \exp(-abx^{c})$$

Обозначив $d = ab$, получим:

$$F(x) = 1 - \exp(-dx^{c})$$

Таким образом, получается классическая форма функции распределения Вейбулла, где параметры d и c могут быть оценены по данным. Это демонстрирует, что степенная функция $y = bx^{c}$ естественным образом приводит к Вейбулловскому распределению, что обосновывает её применимость для описания стохастической зависимости между переменными.

Модифицированный метод моментов — это статистический метод оценки параметров вероятностных распределений. В контексте данной работы

этот метод адаптирован для оценки параметров степенной функции с использованием логарифмического преобразования. Метод основан на сопоставлении теоретических моментов распределения с их эмпирическими аналогами.Модифицированный метод моментов для оценки параметров включает:

$$\beta = \frac{\mathbb{D}[\ln y]}{\mathbb{D}[\ln x]} \tag{1}$$

$$\alpha = \exp\left(\mathbb{E}[\ln y] - \beta \cdot \mathbb{E}[\ln x]\right) \tag{2}$$

где $\mathbb{D}[\cdot]$ — выборочная дисперсия, $\mathbb{E}[\cdot]$ — выборочное среднее.

Критерий Колмогорова-Смирнова — это непараметрический статистический тест, который используется для проверки гипотезы о соответствии эмпирического распределения теоретическому. Критерий применяется для определения, насколько хорошо предсказанные значения зависимой переменной соответствуют её фактическому распределению. Статистика критерия:

$$D_n = \sup_x |F_n(x) - F(x)|$$

где $F_n(x)$ — эмперическая функция распределения, $F(x)$ — теоретическая функция распределения. Модель считается адекватной при $D_n < D_{\text{крит}}$ для заданного уровня значимости.

Практическая реализация

В фргаменте кода в листинге 1 подготавливаем данные для обучения модели.

Нулевые значения в колонках Glucose и Insulin могут быть артефактами измерения (например, неучтенные пределы чувствительности приборов). Их замена на x+1 позволяет сохранить наблюдения и избежать математических ошибок при логарифмировании.Сдвиг на +1 выполнен для корректного применения логарифмического преобразования ln(x), так как логарифм нуля не определён. Это стандартный приём при работе с данными, содержащими нулевые или отрицательные значения.

Разделение данных на обучающую (80%) и тестовую (20%) выборки проведено для проверки обобщающей способности модели. Это позволяет оценить, насколько модель адекватна на новых данных, не участвовавших в обучении.

Листинг 1: Подготовка наборов данных

```
df = pd.read_csv('diabetes.csv')
df['Glucose'] = df['Glucose']  + 1
type1_data = df['Glucose'].values

df['Insulin'] = df['Insulin']  + 1
type2_data = df['Insulin'].values

type1_train, type1_test = train_test_split(type1_data, test_size=0.2,
    random_state=42)
type2_train, type2_test = train_test_split(type2_data, test_size=0.2,
    random_state=42)

def distribution_function(sample):
    min_value = sample.min()
    max_value = sample.max()
    probs = []

    values = np.arange(min_value, max_value, 0.2)
    for value in values:
        prob = np.sum(sample < value)
        probs.append(prob / len(sample))

    return values, probs
```

В фргаменте кода в листинге 2 получаем значения коэффициентов для двупараметрической степенной функцией связи с помощью модифицированного метода моментов.

Листинг 2: Модифицированный метод моментов

```
def moment_method(sample_1, sample_2):
    log_x = np.log(sample_1)
    log_y = np.log(sample_2)

    A = np.vstack([log_x, np.ones(len(log_x))]).T
    b_lin, ln_a = np.linalg.lstsq(A, log_y, rcond=None)[0]
    a = np.exp(ln_a)
    return a, b_lin


alpha, beta = moment_method(type1_train, type2_train)
```

```
12
13 to_plot_s1 = [x - 1 for x in type1_test]
14 to_plot_s2 = [x - 1 for x in type2_test]
15
16
17 to_plot_s2.sort()
18 to_plot_s1.sort()
19 approxed = alpha * to_plot_s1 ** beta
```

В фргаменте кода в листинге 3 проверяем гипотезу о том, что между переменными возможно описать стохастическую зависимость с помощью двупараметрической степенной функции связи.

Листинг 3: Критерий Колмогорова-Смирнова

```
1  def calculate_Kolmogorov_statistic(sample_1, sample_2, alpha, beta):
2      y_sample_1 = list(map(lambda point: calculate_distribution_func(
       sample_1, point), sample_1))
3      y_sample_2 = list(map(lambda point: calculate_distribution_func(
       alpha * sample_2 ** beta, point), alpha * sample_2 ** beta))
4      max_dif = -1e10
5      for y_1 in y_sample_1:
6          for y_2 in y_sample_2:
7              if np.abs(y_1 - y_2) > max_dif:
8                  max_dif = np.abs(y_1 - y_2)
9      return np.max(np.abs(y_approx_interp - y_type1_interp))
10
11 kolmogorov_stat = calculate_Kolmogorov_statistic(type1_test, type2_test,
       alpha, beta)
12 alpha_ = 0.05
13 K_alpha = np.sqrt(-1/2 * np.log((1 - alpha_) / 2))
14
15 print(f"                                                    : {kolmogorov_stat}")
16 print(f"K_alpha: {K_alpha}")
```

Результаты работы

В ходе исследования проведен анализ стохастической зависимости уровня инсулина от концентрации глюкозы в крови с использованием двупараметрической степенной функции связи были получены следующие результаты:

На рисунке 1 и рисунке 2 представлены отдельные эмпирические функции распределения для уровня глюкозы и инсулина соответственно. Визуальный ана-

лиз подтверждает несоответствие их форм, а также они имеют разные единицы измерения, что исключает возможность совмещения на одном графике.

Рис. 1 — Функции распределения уровня глюкозы

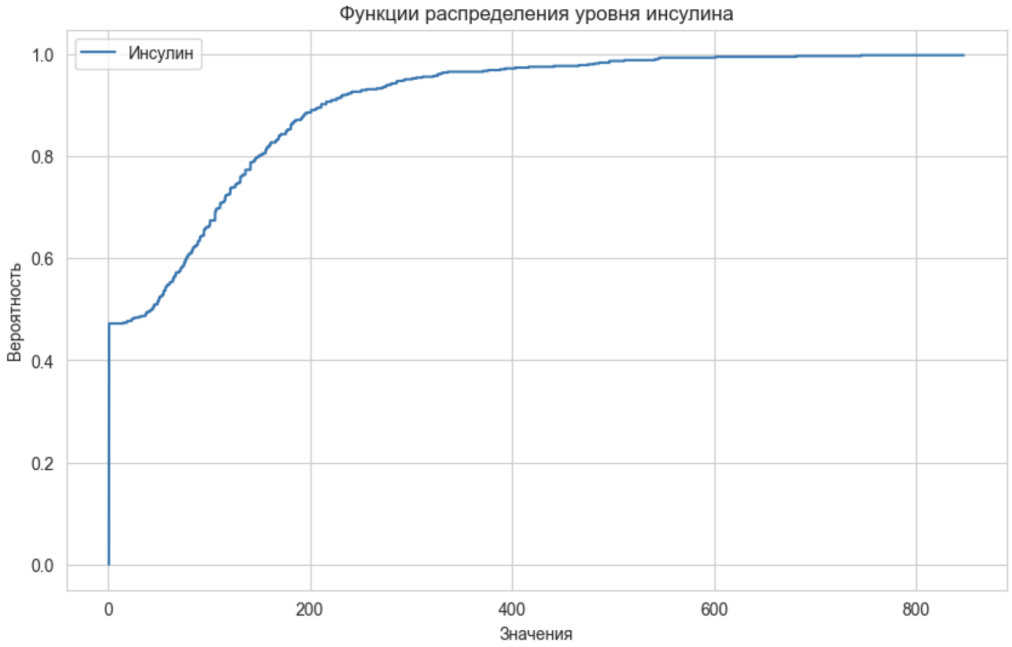

Рис. 2 — Функции распределения уровня инсулина

С помощью метода моментов были получены следующий значения для $\alpha = 0.00008$ и $\beta = 2.52$:

$$y = 0.00008 \cdot x^{2.52}$$

Рис. 3 — Функции распределения глюкозы

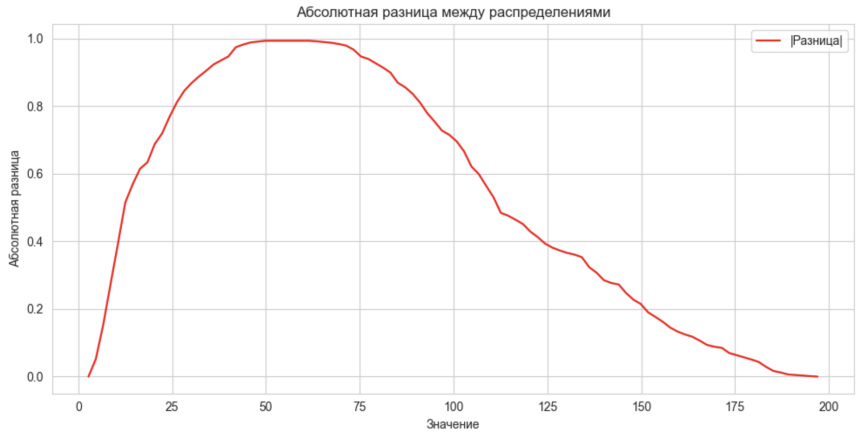

Рис. 4 — Абсолютная разница между распределениями глюкозы

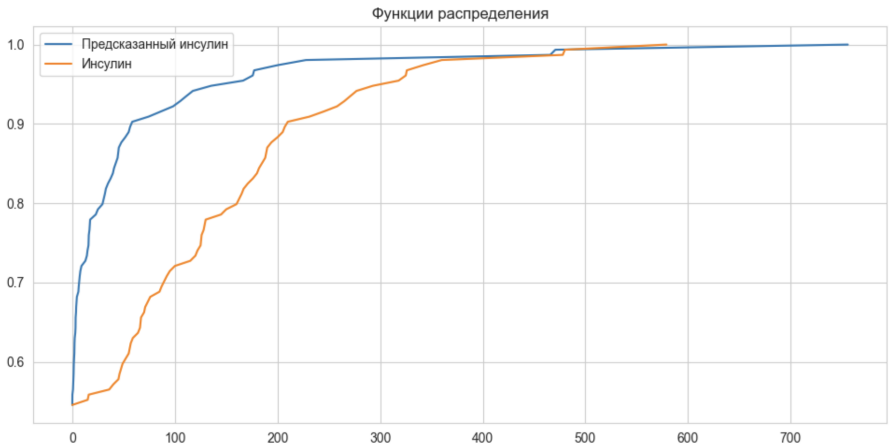

Рис. 5 — Функции распределения инсулина

Проверка соответствия модели данным на тестовой выборке с использованием критерия Колмогорова-Смирнова показала:

Рис. 6 — Абсолютная разница между распределениями инсулина

- Максимальное расхождение между эмпирическими распределениями: $D_n = 0.93$
- Критическое значение при $\alpha = 0.05$: $K_\alpha = 0.61$

Так как $D_n > K_\alpha$, гипотеза о согласии распределений *отвергается* на заданном уровне значимости.

Визуализация результатов подтвердила плохое соответствие модели данным на тестовой выборках.

Заключение

В данной работе применена двупараметрическая степенная функция для попытки описания стохастической зависимости между концентрацией инсулина и уровнем глюкозы. Как на тестовой, так и на обучающей выборках, гипотеза о возможности описания стохастической зависимости с помощью степенной функции с двумя параметрами отвергается на уровне значимости 0.05 с помощью критерия Колмогорова-Смирнова.