Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»

# Расчётно-графическая работа №4

По дисциплине «Математическая статистика»

**Михайлов Дмитрий Андреевич**
P3206
368530
**Медведев Владислав Александрович**
P3206
368508

Санкт-Петербург
2025 год

# Содержание

**Задача №1** **2**

**Задача №2** **6**

**Приложения** **8**

**Список использованных источников** **9**

# Задача №1

**Условие задачи.**

Задание представлено в 4 вариантах. Для каждого варианта требуется построить линейную модель (предполагая нормальность распределения ошибок, некоррелированность компонент, гомоскедастичность), вычислить оценки коэффициентов модели и остаточной дисперсии, построить для них доверительные интервалы, вычислить коэффициент детерминации, проверить указанные в условии гипотезы с помощью построенной линейной модели.

**Указание**: из встроенных функций разрешается пользоваться квантильными функциями и средствами для квадратичной оптимизации (иными словами, готовую обертку для построения линейной модели не использовать, максимум можете сравнить вашу реализацию с готовой)

**Вариант 1**. В файле cars93.csv представлены данные о продажах различных авто.

1. Постройте линейную модель, где в качестве независимых переменных выступают расход в городе, расход на шоссе, мощность (вместе со свободным коэффициентом), зависимой - цена.

2. Проверьте следующие подозрения:

   - Чем больше мощность, тем больше цена
   - Цена изменяется в зависимости от расхода в городе
   - Проверьте гипотезу $H_0$ о равенстве одновременно нулю коэффициентов при расходе в городе и расходе на шоссе против альтернативы $H_1 = \bar{H}_0$

**Решение.**

Построение линейной модели с зависимой переменной Price и независимыми переменными:

- MPG.city (расход в городе)
- MPG.highway (расход на шоссе)
- Horsepower (мощность)
- Свободный коэффициент (константа)

Используем метод наименьших квадратов.

$$\beta = (X^T X)^{-1} X^T y$$

После нахождения коэффициентов $\beta$ предположение делается как $\hat{y} = X\beta$.
Имеем следующее уравнение:

$$\text{Price} = \beta_0 + \beta_1 \cdot \text{MPG.city} + \beta_2 \cdot \text{MPG.highway} + \beta_3 \cdot \text{Horsepower} + \epsilon$$

где:

- $\beta_0$ — свободный коэффициент (intercept)
- $\beta_1, \beta_2, \beta_3$ — коэффициенты при соответствующих переменных

- $\epsilon$ — ошибка (остатки модели)

Метод наименьших квадратов (МНК) чувствителен к выбросам, так как он минимизирует сумму квадратов отклонений, и даже один выброс может сильно исказить коэффициенты регрессии.

**Вывод.**
После написания кода коэффициенты statsmodels и моей реализации практически полностью совпадают => модель построена верно.

**Первое подозрение.**

- Нулевая гипотеза $H_0$: Мощность не влияет на цену, т.е. коэффициент при переменной Horsepower равен 0:

$$H_0 : \beta_{\text{horsepower}} = 0$$

- Альтернативная гипотеза $H_1$: Чем выше мощность, тем выше цена, т.е. коэффициент положительный:

$$H_1 : \beta_{\text{horsepower}} > 0$$

Это односторонний t-тест на значимость одного коэффициента.
Построена линейная модель зависимости цены автомобиля (Price) от:

- расхода топлива в городе (MPG.city)
- расхода на шоссе (MPG.highway)
- мощности двигателя (Horsepower)
- и свободного коэффициента (intercept)

Коэффициенты модели вычислены вручную через формулу нормального уравнения:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Затем была рассчитана стандартная ошибка коэффициента при Horsepower и t-статистика по формуле:

$$t = \frac{\hat{\beta}_{\text{horsepower}}}{SE(\hat{\beta}_{\text{horsepower}})}$$

**Вывод.**
После выполнения всех расчётов вручную получено:

- Коэффициент при Horsepower: 0.1313
- Стандартная ошибка: 0.0161
- t-статистика: 8.1530

- Критическое значение t для уровня значимости $\alpha = 0.05$ и степеней свободы $n - p$:

$$t_{\text{кр}} = t_{1-\alpha}(n - p)$$

Если наблюдаемое значение t-статистики превышает критическое значение:

- Мы отвергаем нулевую гипотезу $H_0$
- Значит, мощность статистически значимо влияет на цену, причём влияние положительное

Решение представлено на языке Python.

**Второе подозрение.**

- Нулевая гипотеза $H_0$: Расход в городе не влияет на цену. То есть коэффициент при MPG.city равен нулю:

$$H_0 : \beta_{\text{MPG.city}} = 0$$

- Альтернативная гипотеза $H_1$: Расход в городе влияет на цену, т.е. коэффициент отличен от нуля (двусторонняя альтернатива):

$$H_1 : \beta_{\text{MPG.city}} \neq 0$$

У нас есть наша линейная модель:

$$\text{Price} = \beta_0 + \beta_1 \cdot \text{MPG.city} + \beta_2 \cdot \text{MPG.highway} + \beta_3 \cdot \text{Horsepower} + \varepsilon$$

Для проверки значимости коэффициента $\beta_1$ (при MPG.city), мы:

1. рассчитываем его оценку $\hat{\beta}_1$
2. определяем стандартную ошибку
3. вычисляем t-статистику
4. сравниваем с критическим значением t для двухстороннего теста при $\alpha = 0.05$

- Если наблюдаемое $|t|$ больше критического — гипотеза $H_0$ отвергается, расход в городе влияет на цену.
- Если $|t| \leq t_{\text{кр}}$ — гипотеза $H_0$ не отвергается, доказательств влияния нет.
- Коэффициент при MPG.city равен $\hat{\beta}_1 = -0.0386$
- t-статистика: -0.1081
- Критическое значение t при уровне значимости 5% (двусторонний тест): +-1.9870

Так как $|t|$ меньше критического значения, мы не отвергаем нулевую гипотезу.

Таким образом, нет статистически значимых оснований утверждать, что расход топлива в городе влияет на цену автомобиля.

**Третье подозрение.**

- Нулевая гипотеза $H_0$:

$$\beta_{\text{MPG.city}} = 0, \quad \beta_{\text{MPG.highway}} = 0$$

- Альтернативная гипотеза $H_1$: хотя бы один из коэффициентов не равен нулю

Это многомерная гипотеза, проверяется с помощью F-критерия сравнивая две модели:
Полная модель:

$$\text{Price} = \beta_0 + \beta_1 \cdot \text{MPG.city} + \beta_2 \cdot \text{MPG.highway} + \beta_3 \cdot \text{Horsepower} + \varepsilon$$

Упрощённая модель ($H_0$):

$$\text{Price} = \beta_0 + \beta_3 \cdot \text{Horsepower} + \varepsilon$$

Мы проверяем гипотезу о том, что коэффициенты при переменных MPG.city и MPG.highway одновременно равны нулю.

Построены две модели:

- Полная: с переменными MPG.city, MPG.highway и Horsepower
- Упрощённая: только с переменной Horsepower

Разница между остаточными суммами квадратов моделей (RSS) используется для расчёта F-статистики:

$$F = \frac{(RSS_{reduced} - RSS_{full})/q}{RSS_{full}/(n-p)}$$

Сравнивая F-статистику с критическим значением из F-распределения при уровне значимости $\alpha = 0.05$, получаем:

- Если $F > F_{crit}$, отвергаем $H_0$: влияние расхода есть
- Если $F \leq F_{crit}$, не отвергаем $H_0$: расход топлива не влияет на цену автомобиля

После получения остаточных сумм квадратов (RSS) и F-статистики мы сравниваем её с критическим значением распределения Фишера при уровне значимости $\alpha = 0.05$.

Если F-статистика превысила критическое значение, мы отвергли нулевую гипотезу, что означает:

**Вывод.**

F-статистика оказалась меньше критического значения, то нет статистических оснований утверждать, что расход топлива влияет на цену автомобиля. Решение представлено на языке Python.

# Задача №2

**Условие задачи.**

Для каждого варианта требуется проверить гипотезу о равенстве средних на каждом уровне фактора с помощью модели однофакторного дисперсионного анализа.

**Указание**: реализовать самим.

**Вариант 1.** В файле iris.csv представлены данные об ирисках. Фактор - подвид. Выходная переменная - суммарная площадь (точнее оценка площади) чашелистика и лепестка.

**Решение.**

Проверить, влияет ли категориальный фактор (подвид) на значение некоторого количественного признака (суммарная площадь чашелистника и лепестка)

Это делается путём проверки нулевой гипотезы:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

где $\mu_i$ — среднее значение признака в группе с $i$-м уровнем фактора, $k$ — количество уровней.
Альтернативная гипотеза $H_1$: средние хотя бы двух групп различаются.
Пусть:

- $X_{ij}$ — значение количественного признака для наблюдения $j$-го в $i$-й группе
- $n_i$ — число наблюдений в $i$-й группе
- $N = \sum_{i=1}^{k} n_i$ — общее число наблюдений
- $\bar{X}_i$ — среднее значение признака в $i$-й группе
- $\bar{X}$ — общее среднее значение по всем наблюдениям

**Расчётные формулы.**

- Общее среднее:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$$

- Межгрупповая сумма квадратов (SSB):

$$SSB = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2$$

  Показывает, насколько отличаются средние групп от общего среднего — то есть, влияние фактора.

- Внутригрупповая сумма квадратов (SSW):

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

Показывает естественную дисперсию внутри групп — обусловленную случайными колебаниями.

**Статистика F.**

- Число степеней свободы: $df_b = k - 1$ — межгрупповое, $df_w = N - k$ — внутригрупповое
- Средние квадраты: $MSB = \frac{SSB}{df_b}$, $MSW = \frac{SSW}{df_w}$

Общая формула F-статистики.

$$F = \frac{MSB}{MSW}$$

**Критерий принятия решения.**
Сравниваем вычисленную F-статистику с критическим значением из распределения Фишера:

$$F > F_{\text{crit}}(\alpha; df_b, df_w) \Rightarrow \text{отклоняем } H_0$$

где $\alpha$ — уровень значимости (0.05). Критическое значение находится по квантильной функции:

$$F_{\text{crit}} = F^{-1}(1 - \alpha)$$

с параметрами $df_b, df_w$.

**Принятие решения.**

- Если $F \leq F_{\text{crit}}$: фактор не оказывает значимого влияния — средние статистически одинаковы.
- Если $F > F_{\text{crit}}$: есть значимые различия между средними в разных группах — фактор влияет на результат.

Решение представлено на языке Python.

# Приложения

## Задача №1

Ссылка на исходник с кодом программы, решающей эту задачу на языке Python. [1]

## Задача №2

Ссылка на исходник с кодом программы, решающей эту задачу на языке Python. [2]

# Список использованных источников

[1] Задача №1. *URL*: Исходник с кодом, решающий задачу №1.

[2] Задача №2. *URL*: Исходник с кодом, решающий задачу №2.