Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

УТВЕРЖДЕНО:
Директор
А. В. Замятин

Оценочные материалы по дисциплине

Статистический анализ данных

по направлению подготовки

01.04.02 Прикладная математика и информатика

Направленность (профиль) подготовки:
Интеллектуальный анализ больших данных

Форма обучения
Очная

Квалификация
Магистр

Год приема
2025

СОГЛАСОВАНО:
Руководитель ОП
А.В. Замятин

Председатель УМК
С.П. Сущенко

Томск – 2025

1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-1 Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;.

ОПК-3 Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями;.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-1.3 Развивает и применяет математические, естественнонаучные, социально-экономические и профессиональные знания для решения задач

ИОПК-3.1 Осуществляет сбор, обработку и анализ научно-технической информации, необходимой для решения профессиональных задач

ИОПК-3.2 Умеет работать с различными видами информации с помощью различных средств информационных и коммуникационных технологий

ИОПК-3.3 Формулирует результаты, полученные в ходе решения исследовательских задач, в виде аналитических обзоров с обоснованными выводами и рекомендациями

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:

– посещение;

– выполнение лабораторных работ на компьютере.

Посещение фиксируется на каждом занятии. Допускается 25% пропусков по уважительной причине. При большем количестве пропусков студент получается дополнительный вопрос на экзамене по пропущенным темам и/или дополнительное задание по практике.

Пример задания к лабораторной работе (ИОПК-1.3, ИОПК-3.1, ИОПК-3.2, ИОПК-3.3)

Лабораторная работа. Гетероскедастичность.

Для предложенного набора реальных данных проверить наличие гетероскедастичность и устранить ее. Выполнить следующие шаги.

- ✓ Импортировать данные из файла в R.
- ✓ Построить парную модель регрессии целевой переменной от основного количественного фактора.
- ✓ Построить диаграмму рассеяния с оцененной линией регрессии
- ✓ Вывести оцененные значения.
- ✓ Вывести значения коэффициентов уравнения регрессии.
- ✓ Построить прогноз для значений целевой переменной в соответствии с парной моделью.
- ✓ Построить модель множественной регрессии на все предложенные в наборе факторы.
- ✓ Удалить незначимые факторы и построить новую модель.
- ✓ Для всех количественных и порядковых величин построить корреляционную матрицу.

- ✓ Построить прогноз для значений целевой переменной в соответствии с множественной моделью.
- ✓ Провести анализ остатков построенных парной и множественной моделей.
- ✓ Проверить на гетероскедастичность.
- ✓ Устранить гетероскедастичность.
- ✓ Провести анализ остатков новой модели.

Лабораторная работа. Логистическая регрессия.

Задание 1. Формирование выборки

Сформировать наблюдения, связанные однофакторной логистической регрессией.

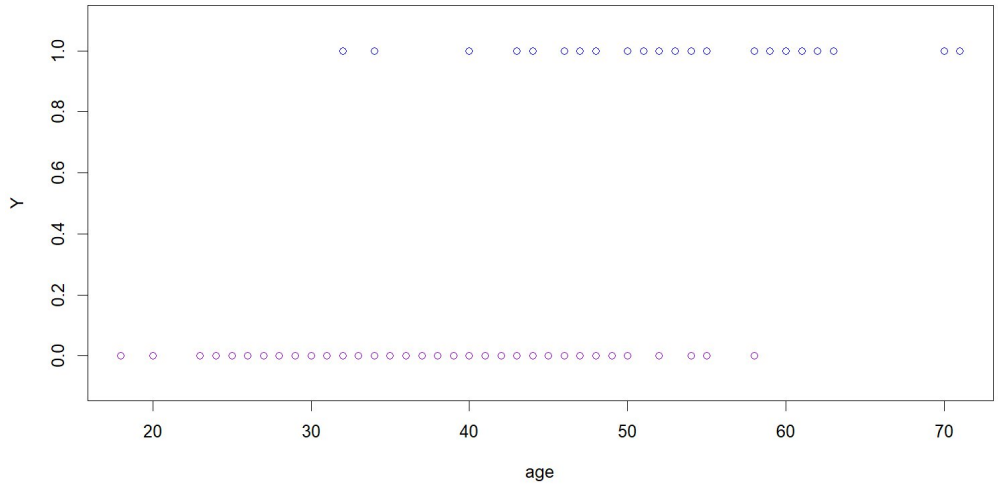
- ✓ Задать объем выборки $n = 20{:}50$.
- ✓ Значения фактора x сформировать как реализацию целочисленной равномерно распределенной случайной величины в интервале $[a,b]$.
- ✓ Задать нормально распределенный шум $\varepsilon \sim N(0,\sigma)$.
- ✓ Определить регрессионную модель

$$\Pi(x) = \frac{e^{\theta_0+\theta_1 x+\varepsilon}}{1+e^{\theta_0+\theta_1 x+\varepsilon}}.$$

- ✓ Значение бинарной зависимой переменной определить как

$$y_i = \begin{cases} 0, & \Pi(x_i) < \frac{1}{2}; \\ 1, & \Pi(x_i) \geq \frac{1}{2}. \end{cases}$$

Все параметры задать самостоятельно, ориентируясь на диаграмму рассеяния. Например, как на рисунке



Задание 2. Анализ построенной модели

- ✓ Провести лог-регрессионный анализ.
- ✓ Построить оценки параметров модели.
- ✓ Построить предсказания для произвольно заданного значения фактора.
- ✓ Оценить качество модели.
- ✓ Построить ROC-кривую.

Критерии оценивания:

Результаты лабораторной работы определяются оценками «зачтено» или «не зачтено».

Оценка «зачтено» выставляется, если работа выполнена полностью или с незначительными недочетами, код работает, студент аргументированно объясняет выбор примененных методов и полученные результаты.

Оценка «не зачтено» выставляется, если код не работает или работает с существенными ошибками, студент не может пояснить выбор методов и корректно проинтерпретировать полученные результаты. Работа отправляется на доработку.

3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

Темы для подготовки к итоговому тестированию (ИОПК-1.3, ИОПК-3.1, ИОПК-3.2, ИОПК-3.3)

Темы для подготовки к письменному экзамену.

1. Типы данных и способы их представления.
2. Параметрические критерии сравнения групп.
3. Непараметрические критерии сравнения групп.
4. Корреляционный анализ количественных данных. Коэффициент Пирсона. Z-преобразование Фишера.
5. Ранговая корреляция. Коэффициент Спирмена, Кендалла и конкордации.
6. Корреляционный анализ категоризованных данных. Коэффициент квадратичной сопряженности. Коэффициент Крамера.
7. Парная регрессии. Модель. МНК-оценки параметров.
8. Числовые характеристики оценок параметров парной регрессии.
9. Теорема Гаусса-Маркова для случая парной регрессии.
10. Проверка качества уравнения парной регрессии.
11. Нелинейные модели и линеаризация.
12. Случай смещенного шума.
13. Случай коррелированных гомоскедастичных наблюдений.
14. Случай некоррелированных гетероскедастичных наблюдений.
15. Мультиколлинеарность.
16. Фиктивные переменные.
17. Постановка задачи классификации.
18. Логистическая регрессия.
19. Метрики качества бинарного классификатора.
20. ROC-анализ.
21. Типы методов кластеризации.
22. Расстояния между объектами, расстояния между кластерами.
23. Структура временного ряда.
24. Прогнозирование во временных рядах.

Примеры вопросов итогового теста (ИОПК-1.3, ИОПК-3.1, ИОПК-3.2, ИОПК-3.3).

1. По критерию Шапиро-Уилка были получены следующие результаты.
   W = 0.9821, p-value = 0.6432
   Какой вывод можно сделать на уровне значимости 0.05?
   а) есть значимые отличия между анализируемыми совокупностями
   б) нет значимых отличий между анализируемыми совокупностями

в) выборка не противоричит нормальному закону распределения
г) выборка не подчиняется нормальному закону распределения

Ответ: в)

2. По критерию Голдфельда – Квандта были получены следующие результаты:

GQ = 6.7131, df1 = 22, df2 = 21, p-value = 2.437e-05

Какой вывод можно сделать?
а) имеется мультиколлинеарность
б) нет мультиколлинеарности
в) остатки гомоскедастичны
г) остатки гетероскедастичны

Ответ: г)

3. Собственные числа информационной матрицы равны:
7.5646e+07 3.7582e+05 2.3019e+04 3.2468e+03 4.03e+02 1.2817e+02
Какой вывод можно сделать?
а) Имеет место эффект мультиколлинеарности.
б) Нет мультиколлинеарности.
в) Недостаточно информации, чтобы сделать вывод о мультиколлинеарности.

Ответ: а)

4. В модели без свободного члена коэффициент детерминации в статистических пакетах рассчитывается по формуле:

а) $R^2 = 1 - \frac{\sum \left(y_i - \hat{y}_i\right)^2}{\sum \left(y_i - \bar{y}\right)^2}$

б) $R^2 = \frac{\sum \left(\hat{y}_i - \bar{y}\right)^2}{\sum \left(y_i - \bar{y}\right)^2}$

в) $R^2 = 1 - \frac{\frac{1}{n-m}\sum \left(y_i - \hat{y}_i\right)^2}{\frac{1}{n-1}\sum \left(y_i - \bar{y}\right)^2}$

г) $R^2 = \frac{\sum \hat{y}_i^{\,2}}{\sum y_i^{\,2}}$

Ответ: г)

5. По какой формуле можно рассчитать F1 score
а) $2\frac{Precision + Recall}{Precision \cdot Recall}$
б) $2\frac{Precision \cdot Recall}{Precision + Recall}$
в) $\frac{Precision + Recall}{2 \cdot Precision \cdot Recall}$
г) $\frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$

Ответ: б), г)

6. Мультипликативная модель временного ряда описывается уравнением

а) $y = a + bt + \varepsilon$
б) $y = at^b\varepsilon$
в) $y(t) = f(t) + \phi(t) + \psi(t) + \varepsilon(t)$
г) $y(t) = f(t) \cdot \phi(t) \cdot \psi(t) \cdot \varepsilon(t)$

Ответ: г)

Тест состоит из 20 вопросов разной сложности, за каждый правильный ответ можно получить от 1 до 3 баллов. Максимум за тест 40 баллов.
[0 ;21) неудовлетворительно
[21;28) удовлетворительно
[28;35) хорошо
[35;40] отлично

4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)

Проверка остаточных знаний проводится в форме теста.
Примеры вопросов к тесту (ИОПК-1.3, ИОПК-3.1, ИОПК-3.2, ИОПК-3.3):

1. Критерий Манна-Уитни применяется для
а) выявления различий между двумя зависимыми выборками по уровню какого-либо признака
б) выявления различий между двумя независимыми выборками по уровню какого-либо признака
в) выявления различий между несколькими независимыми выборками по уровню какого-либо признака
в) выявления различий между несколькими зависимыми выборками по уровню какого-либо признака

Ответ: б)

2. Для двух порядковых переменных при расчете коэффициента Спирмена были получены следующие результаты.

r = - 0,17558892 p = 0,0316143305

Какой вывод можно сделать при уровне значимости 0.05?

а) имеется прямая статистическая связь на уровне значимости 0,05
б) имеется обратная статистическая связь на уровне значимости 0,05
в) имеется прямая статистическая связь на уровне значимости 0,01
г) имеется обратная статистическая связь на уровне значимости 0,01
д) нет статистически значимой связи на уровне значимости 0,01
е) нет статистически значимой связи на уровне значимости 0,05
Ответы: б) , д)

3. При проведении регрессионного анализа были получены результаты

	$R = 0.9804$ $R^2 = 0.9612$ $R^2_{adj} = 0.9608$ $F(1,98) = 2428.9$ $p < 0.0000$ $S_e = 29.278$				
$n = 100$	b^*	b	S_b	$t(98)$	$p - value$
Intercept		20.33116	7.726744	2.63127	0.009882

x	0.9804	4.99870	0.101427	49.28397	0.000000

Чему равна оценка среднего ожидаемого значения зависимой переменной при значении факторной переменной равной 100. Ответ округлен до двух знаков после запятой.
а) 2038,11
б) 520,20
в) 201,63
г) 29,278

Ответ: б)

4. Коэффициент детерминации регрессионной модели может определяться как отношение:
а) остаточной суммы квадратов к общей сумме квадратов;
б) общей суммы квадратов к остаточной сумме квадратов;
в) объясненной суммы квадратов к общей сумме квадратов;
г) общей суммы квадратов к объясненной сумме квадратов;
д) остаточной суммы квадратов к объясненной сумме квадратов.

Ответ: в)

5. Метрика Accuracy в задаче классификации определяется как
а) отношение числа верно классифицированных объектов к неверно классифицированным
б) отношение числа неверно классифицированных объектов к верно классифицированным
в) отношение числа верно классифицированных к общему числу объектов
г) отношение числа неверно классифицированных к общему числу объектов

Ответ: в)

6. Долгосрочная тенденция динамики показателя - это
а) шум;
б) тренд;
в) сезонные колебания;
г) циклические колебания.

Ответ: б)

Тест для проверки остаточных знаний стоит из 10 вопросов разной сложности от 1 до 3 баллов. Максимум за тест 20 баллов. Тест оценивается на «зачтено» или «незачтено». Оценка «зачтено» ставится, если студент набрал 11 баллов и выше.

Информация о разработчиках

Кабанова Татьяна Валерьевна, канд. физ.-мат. наук, доцент, кафедра теории вероятностей и математической статистики ИПМКН ТГУ, доцент