НИУ ВШЭ, «Вычислительные социальные науки»
Курс «Регрессионный анализ социально-экономических процессов», 2024 – 2025

Домашнее задание 1
Дедлайн: 23.59 15 октября 2024

Задание 1. Ниже представлены результаты анализа разложения вариации по результатам оценивания линейной регрессионной модели:

ANOVA

	sum_sq	mean_sq	df	f	PR(>F)
x	2.5771				
Residual	11.5366		490		
Total			495		

Используя информацию из данной таблицы,

1. рассчитайте коэффициент детерминации и проинтерпретируйте его
2. проверьте гипотезу о о том, что регрессия на константу (то есть, модель без объясняющих переменных) не хуже модели с предикторами, на фиксированном уровне значимости 0.05. Запишите нулевую и альтернативную гипотезы на статистическом языке, рассчитайте значение статистики, а также выберите необходимую критическую точку – квантиль – из списка ниже. Сделайте вывод.
   (a) квантиль распределения Фишера, 0.95, df1 = 490; df2 = 5: 4.373
   (b) квантиль распределения Фишера, 0.975, df1 = 490; df2 = 5: 6.029
   (c) квантиль распределения Фишера, 0.95, df1 = 6; df2 = 490: 2.117
   (d) квантиль распределения Фишера, 0.975; df1 = 6; df2 = 490: 2.434
   (e) квантиль распределения Фишера, 0.95; df1 = 5; df2 = 490: 2.232
   (f) квантиль распределения Фишера, 0.975; df1 = 5; df2 = 490: 2.592

Задание 2. В рамках исследования факторов, увеличивающих риск заболевания коронавирусной инфекцией, было показано, что у курящих чаще диагностируют коронавирус. Одна группа исследователей объяснила этот результат тем, что курение повышает риск потери обоняния, а потеря обоняния как один из симптомов коронавируса может, в свою очередь, выступить основанием для человека сделать тест на коронавирус, что увеличивает вероятность обнаружить заболевание. Другая группа исследователей придерживаются альтернативной позиции, считая, что курение приводит к большей подверженности заболеть коронавирусной инфекцией. Постройте по описанному контексту подходящий граф (на графе разделяйте заболевание коронавирусной инфекцией и показатель, диагностирована ли коронавирусная инфекция). Позволит ли включение показателя потери обоняния в модель в качестве контрольной переменной определить, действительно ли курение приводит к большей подверженности заболеть коронавирусной инфекцией? Рассмотрите два сценария, свой ответ объясните.

Задание 3. Корреляция между показателями x и y равна 0.9. Реализуйте на основе соответствующей корреляционной матрицы (размерности 2×2) метод главных компонент, выполните задания, представленные ниже:

1. Запишите применительно к данной задаче характеристическое уравнение и полученные на основе него собственные числа
2. Выполните необходимые расчеты и запишите в явном виде матрицу поворота

Задание 4. Ниже в таблице представлены значения переменных: X, Z, Y.

X	3	0	0	-4	1
Z	-3	0	1	0	2
Y	-5	0	-1	-2	2

Получите оценки коэффициентов в регрессии Y на X и Z с помощью общей векторно-матричной формулы получения оценок коэффициентов. Представьте промежуточные расчеты, выпишите полученный вектор оценок коэффициентов и запишите спецификацию модели, подставив эти оценки в уравнение.

Задание 5. Вы работаете со следующей регрессионной моделью:

$$dropout_i = \beta_0 + \beta_1 perPupil_i + \beta_2 povertyRate_i + \beta_3 classSize_i + \varepsilon_i$$

$dropout_i$ — процент отчислившихся студентов из числа зачисленных в колледж в i-том округе; $perPupil_i$ — расходы на одного студента (тыс. долл.); $povertyRate_i$ — процент студентов из семей с доходом ниже прожиточного минимума; $classSize_i$ — среднее количество студентов в группе.
После оценивания этой модели Вы для диагностики последствий мультиколлинеарности дополнительно оценили вспомогательные регрессии. Корреляция между зависимой переменной $perPupil\_i$ и предсказанным значением зависимой переменной в одной из таких вспомогательных моделей составляет 0.78. Запишите соответствующую спецификацию вспомогательной регрессии, рассчитайте для переменной $perPupil\_i$ значение Variance Inflation Factor и проинтерпретируйте полученное значение (что VIF показывает, а также насколько критично полученное значение с точки зрения последствий для инференции в регрессионной модели).