

Домашнее задание №3 (сдаваемое)
Дедлайн: 23:59 16 декабря

Для проверки решение ДЗ в файле формата .ipynb необходимо отправить на следующие почты: dskuznetsova_2@edu.hse.ru; darysalnikova@yandex.ru

Тема письма в формате: Фамилия ИО студента, ДЗ3

Задания выполняются на массиве `reg_hw3.dta`. Описание данных представлено ниже:

electint	Interest in EU elections (1=very low, 2=low, 3=some, 4=very high)
female	Individual gender (1=female, 0=male)
agecen50	Individual age in years (centered at 50)
lrplace	Placement on scale of left-right political attitudes (a 10-point scale with high values indicating more right wing views)
commtype	Type of community of residence (1=rural, 2=mid-sized town, 3=large town or city)

Задание 1. Подготовьте данные к последующему анализу. Оставьте в массиве только вышеприведенные переменные. При необходимости перекодируйте переменные в набор дамми. В качестве зависимой переменной выступает *electint*, для начала преобразуйте в бинарную ее таким образом, чтобы категориям 3 и 4 соответствовало значение 1 (наличие интереса к выборам), а категориям 1 и 2 – 0

Задание 2. Оцените логистическую регрессию с бинарным откликом. В качестве зависимой переменной возьмите преобразованную переменную *electint*, все остальные обозначенные в таблице переменные в исходном или преобразованном виде (см. предыдущее задание) используйте как предикторы. Проинтерпретируйте оценки коэффициентов при двух любых предикторах (выберите их самостоятельно), далее преобразуйте эти оценки в отношения шансов и также дайте интерпретацию

Задание 3. Проверьте гипотезу о согласии наблюдаемых значений (данных) и модели посредством теста Хосмера–Лемешева (Hosmer–Lemeshow test). Кратко поясните, каким образом устроена статистика теста (общий смысл), сделайте вывод на основании полученных результатов

Задание 4. Постройте ROC-кривую для оцененной модели. Проинтерпретируйте полученный график: объясните своими словами, как был построен данный график, проинтерпретируйте площадь под ROC-кривой (AUC)

Задание 5.

1. Сохраните предсказанные вероятности попадания в категорию «1» отклика в текущей регрессионной модели, задайте самостоятельно порог отсечения и представьте в качестве результата confusion matrix. По представленной таблице классификации рассчитайте меру точности (accuracy), базовой точности (baseline accuracy), ошибку первого рода, ошибку второго рода, специфичность и чувствительность (запишите в явном виде, как рассчитываются эти значения, объясните, что они содержательно показывают в контексте поставленной содержательной задачи)
2. Сравните решение (выбранный Вами порог отсечения) с классификацией наблюдений в соответствии с оптимальным порогом, выбранным на основе минимизации разницы между мерами чувствительности и специфичности

Задание 6.

1. Предложите модель, вложенную в исходную. Проверьте гипотезу об отсутствии различий между более и менее экономной моделями критерием отношения правдоподобия. Сделайте статистический и содержательный вывод
2. Предложите спецификации двух невложенных моделей. Сравните их посредством информационных критериев AIC и BIC. Сделайте статистический и содержательный вывод

Задание 7. Далее используйте electint в исходной шкале (4 категории) и оцените логистическую регрессионную модель с порядковым откликом. Сохраните предсказанные вероятности попадания в ту или иную категорию electint, выведите confusion matrix. Сделайте выводы на основе рассчитанных мер специфичности и чувствительности для разных классов

Задание 8. Проинтерпретируйте коэффициент в модели с порядковым откликом при одном из предикторов (на Ваш выбор) в терминах отношения шансов, в этой интерпретации проиллюстрируйте идею параллельности регрессий (пропорциональности шансов). Справедливо ли допущение о параллельности регрессий? Свой ответ обоснуйте