

Практикум. Модели с бинарным откликом

Общая постановка задачи Задание выполняется на массиве [logit_lab.sav](#). Для того, чтобы открыть файл, используйте: `pd.read_spss('logit_lab.sav', convert_categoricals = False)`

Источник данных – Pew Research Center. Тематика – восприятие степени угрозы коронавируса здоровью индивида. Описание переменных представлено ниже.

COVIDTHREAT_b_W63.5 - dependent variable

How much of a threat, if any, is the coronavirus outbreak for your personal health?

Value	Label
1	A major threat
2	A minor threat
3	Not a threat
99	No answer

F_SEX - a person's biological sex

Value	Label
1	male
2	female
99	No answer

F_AGECAТ age categories

Value	Label
1	18-29
2	30-49
3	50-64
4	65 and older
99	No answer

F_EDUCCAT2: education category

Value	Label
1	Less than high school
2	High school graduate
3	Some college, no degree
4	Associate's degree
5	College graduate/some post grad
6	Postgraduate
99	No answer

F_INCOME Family income

Value	Label
1	Less than \$10,000
2	\$10,000 to less than \$20,000
3	\$20,000 to less than \$30,000
4	\$30,000 to less than \$40,000
5	\$40,000 to less than \$50,000
6	\$50,000 to less than \$75,000
7	\$75,000 to less than \$100,000
8	\$100,000 to less than \$150,000
9	\$150,000 or more
99	No answer

Задание 1. Подготовьте данные к анализу.

1. Оставьте в массиве только вышеприведенные переменные
2. Представьте все «No answer» (код 99) как пропущенные значения
3. Преобразуйте исходные три категории зависимой переменной в две, таким образом, чтобы 1 означало восприятие угрозы коронавируса здоровью индивида как серьезной, а 0, в свою очередь, объединяло категории «a minor threat» и «no threat»

Задание 2. Оцените логит- и пробит- регрессионные модели с бинарным откликом и сравните полученные оценки. В качестве предикторов возьмите пол респондента, возрастные категории как набор дамми-переменных, уровень образования и доход семьи в качестве псевдо-интервальных переменных. Переменная – восприятие угрозы коронавируса здоровью индивида – будет выступать зависимой переменной

1. Сравните оценки пробит- и логит-модели
2. Представьте интерпретацию оценок логит-модели в терминах отношения шансов

Задание 3. Постройте ROC-кривую для оцененной модели. Проинтерпретируйте полученный график: объясните своими словами, как был построен данный график, проинтерпретируйте площадь под ROC-кривой (AUC)

- Задание 4.**
1. Сохраните предсказанные вероятности попадания в категорию восприятия угрозы коронавируса здоровью индивида как серьезной, задайте самостоятельно порог отсека и представьте в качестве результата confusion matrix. По представленной таблице классификации рассчитайте ошибку первого рода, ошибку второго рода и мощность (запишите в явном виде, как рассчитываются эти значения, объясните, что они содержательно показывают в контексте поставленной содержательной задачи)
 2. Сравните решение (выбранный Вами порог отсека) с классификацией наблюдений в соответствии с оптимальным порогом, выбранным на основе минимизации разницы между мерами чувствительности и специфичности

- Задание 5.**
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