Лабораторная работа
Регрессионный анализ данных

По данным Всемирной организации здравоохранения (ВОЗ), инсульт является второй по значимости причиной смерти в мире, на него приходится около 11% всех смертей. Этот набор данных используется для прогнозирования вероятности инсульта у пациента на основе таких входных параметров, как пол, возраст, различные заболевания и статус курения. Каждая строка в данных содержит релевантную информацию о пациенте.

Информация о переменных:

1) **id**: уникальный идентификатор;

2) **gender**: «Мужской», «Женский», «Другой»;

3) **age**: возраст пациента;

4) **hypertension**: 0, если у пациента нет гипертензии, 1, если у пациента есть гипертензия;

5) **heart_disease**: 0, если у пациента нет никаких заболеваний сердца, 1, если у пациента есть заболевание сердца;

6) **ever_married**: «Нет» или «Да»;

7) **work_type**: «Ребенок», «Государственное учреждение», «Никогда не работал», «Негосударственная (частная) компания», «Самозанятый»;

8) **Residence_type**: место проживания «Сельское» или «Городское»;

9) **avg_glucose_level**: средний уровень глюкозы в крови;

10) **bmi**: индекс массы тела;

11) **smoking_status**: «ранее курил», «никогда не курил», «курит» или «Неизвестно»* («Неизвестно» в smoking_status означает, что информация для этого пациента недоступна);

12) **stroke**: 1, если у пациента был инсульт или 0, если нет.

Задание №1. Исследовать влияние возраста человека на (age) на индекс массы тела (bmi).

1.1 Подготовить данные для анализа.

1.2. Сделать визуализацию данных.

1.3. Найти описательные статистики исходных данных. Сделать предварительные выводы о свойствах исходных данных.

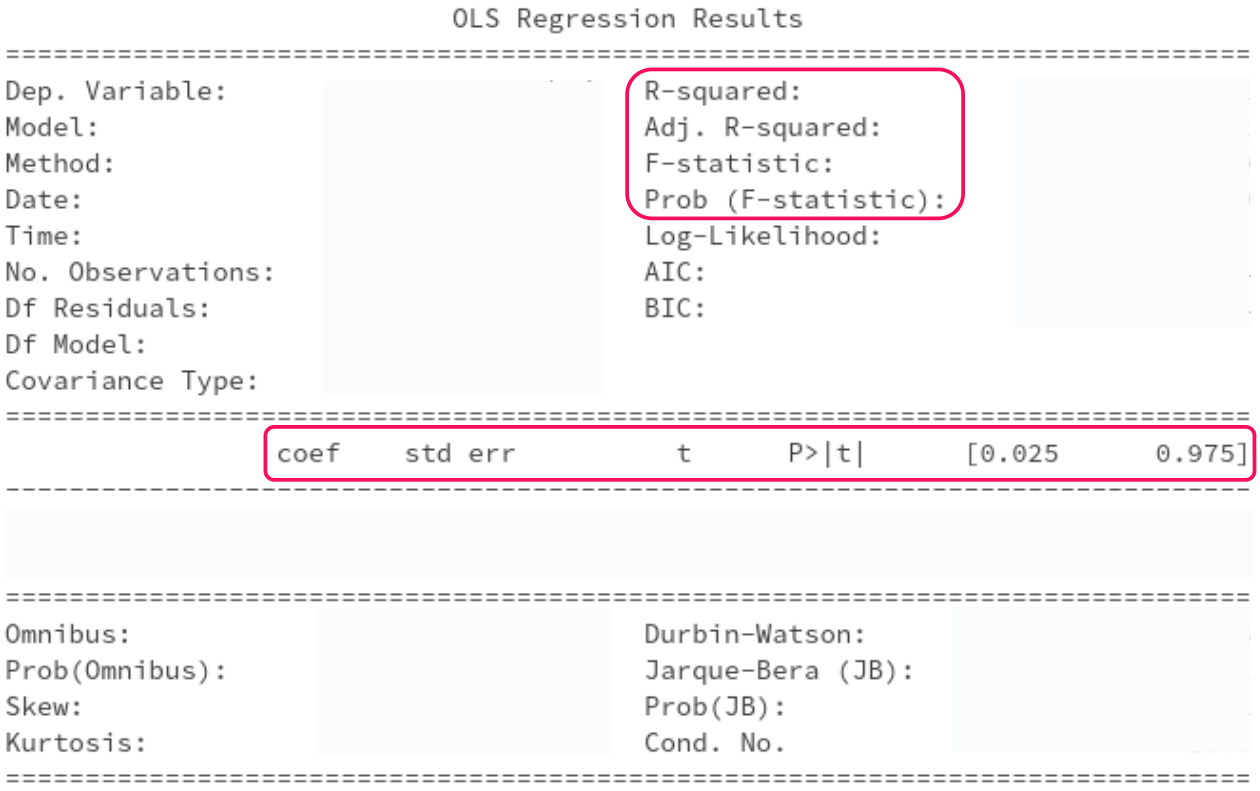
1.4. Исследовать закон распределения исходных данных и, при необходимости, преобразование исходных данных к нормальному закону распределения (не обязательно).

1.5. Выявить статистически аномальные значения (выбросы). Принять решение об их исключении.

1.6. Корреляционный анализ - исследовать корреляционную связь между исходными данными;

1.7. С помощью модуля statsmodels.formula.api (as smf) получить результаты для линейной регрессионной модели на основе обычного метода наименьших квадратов (Ordinary Least Squares):

Результат будет представлен в виде таблицы:



Для выделенных переменных понимать, что они означают, и *уметь получать* их значения самостоятельно.

1.8. На основе модулей sklearn.preprocessing (понадобится функция PolynomialFeatures) и

sklearn.linear_model (понадобится функция LinearRegression) построить полиномиальное уравнение регрессии. Вид модели выбрать самостоятельно.

1.9. Сравнить линейную и полиномиальную модель на основе квадрата отклонения предсказанных результатов от экспериментальных данных.

1.10. Построить график уравнений линейной и полиномиальной регрессии совместно с диаграммой рассеяния.

Задача №2. Выбрать из исходной таблицы любую пару из количественной и категориальной переменной. Провести исследование влияние выбранной количественной переменной на категориальную.

2.1 Подготовить данные.

2.2 Сделать визуализацию.

2.3. Провести разделение данных на обучающие и тестовые. Функция train_test_split из модуля sklearn.model_selection.

2.4 Определить параметры логистического уравнения регрессии. Для этого удобно воспользоваться модулем sklearn.linear_model (функция LogisticRegression)

2.5. На основе полученной модели получить предсказания для тестовых данных (см. п. 2.3).

2.6 Построить матрицу ошибок. Можно использовать функцию confusion_matrix из модуля sklearn.metrics.

2.7. Сделать выводы по полученным результатам.

Форма отчета -устная. Знать теорию по теме «Регрессионный анализ данных».