Домашнее задание № 6

Регрессионный анализ.

Крайний срок сдачи: 17 декабря 2021 г., 18:10 .

Каждое задание оценивается в 2 балла.

1. В базе данных `https://stats.idre.ucla.edu/stat/data/poisson_sim.csv` представлена выборка из 200 студентов, про которых известна следующая информация:

   - количество наград, полученных за время обучения;
   - тип программы (1- профессиональная (прикладная), 2- общая, 3- академическая);
   - оценка за финальный экзамен по математике (по 100-балльной шкале).

   (i) Постройте обобщённую линейную модель, описывающую зависимость количества наград от типа программы и балла за финальный экзамен по математике. В качестве экспоненциального семейства используйте семейство распределений Пуассона. Если какие-либо факторы являются незначимыми, то постройте модель без данных факторов.

   (ii) Покажите, что тест, построенный при помощи теоремы Уилкса, позволяют отклонить гипотезу о том, что модель на самом деле является тривиальной (то есть с одинаковым значением параметра в каждой точке).

   (iii) Напишите код функции, возвращающей по заданным значениям типа программы и балла за финальный экзамен вероятности получения 0,1,2,...,6 наград.

2. Рассмотрим базу данных "LifeCycleSavings"(`https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/LifeCycleSavings.html`), содержащую информацию о среднем коэффициенте персональных сбережениях жителей 50 стран. Этот коэффициент для конкретного жителя вычисляется как отношение его совокупных личных сбережений к располагаемому доходу. Согласно гипотезе Модильяни, среднее по стране значение этого коэффициента зависит от

   - процента населения моложе 15 лет (LifeCycleSavings$pop15);
   - процента населения старше 75 лет (LifeCycleSavings$pop75);
   - располагаемого дохода на душу населения (LifeCycleSavings$dpi);
   - процентной скорости изменения располагаемого дохода на душу населения (LifeCycleSavings$ddpi).

   Представленные данные являются усреднёнными показателями за 1960–1970 гг.

   (i) Для переменных "sr"(как y-переменной) и "pop15" (как x-переменной) постройте ядерную оценку регрессии при различных вариантах выбора ядра (гауссовское ядро и ядро Епанечникова) и различных методах выбора параметра bandwidth (критерий Акаике, обобщённый метод кросс-проверки). Найдите наилучший метод в смысле наименьшей среднеквадратичной ошибки.

   (ii) Повторите вычисления для остальных трёх объясняющих переменных вместо "pop15". Выберите 2 переменные, которые по Вашему мнению наилучшим образом объясняют коэффициент персональных сбережений (в дальнейшем эти переменные будем называть V1 и V2). Объясните свой выбор.

   (iii) На основе V1 и V2 постройте многомерную регрессию методом LOESS и линейную регрессию. Разделите случайным образом все страны на 2 группы: в одну группу отнесите примерно 80 % стран, в другую - 20 %. Оцените параметры модели LOESS и линейной регрессии по большей группе и проверьте качество моделей по меньшей. Выясните, какая из построенных моделей является более точной.

3. Для базы данных из предыдущей задачи требуется найти наблюдения-выбросы. Предлагается сравнить 2 метода определения выбросов:

   (a) по диаграмме размаха отдельно по каждой из 5 переменных;

   (b) при помощи подсчёта параметров leverages: считаем выбросами все наблюдения, для которых leverages (при построении регрессии от 4-х факторов) превосходят более чем в 2 раза среднее значение этого параметра по всем наблюдениям.

   Имплементируйте оба метода и сравните качество моделей линейной регрессии, построенных на основе всех переменных, при удалении наблюдений-выбросов. Качество нужно сравнить по p-уровням тестов для коэффициентов детерминации R^2.

4. Пусть дан набор точек $(x_i, y_i), i = 1..n$. Для описания регрессионной зависимости между y_i и x_i будем использовать оценку Надарая-Ватсона

   $$\hat{r}(x) = \frac{\sum_{i=1}^{n} y_i K\big((x - x_i)/h\big)}{\sum_{i=1}^{n} K\big((x - x_i)/h\big)}$$

   с треугольным ядром

   $$K(x) = (1 - |x|) \cdot \mathbb{I}\{|x| \le 1\}$$

   и параметром $h > 0$. Для случая $n = 6$ и $x_i = i$, $\forall i = 1..6$, вычислите сглаживающую матрицу H и эффективное количество степеней свободы (след матрицы H), если

   (i) $h = 1/2$;

   (ii) $h = 3/2$.

   Комментарий. Напомним, что слаживающая матрица H - это такая матрица, что

   $$\hat{\vec{y}} = H\vec{y},$$

   где $\vec{y} = (y_1, ..., y_n)^\top, \hat{\vec{y}} = (\hat{r}(x_1), ..., \hat{r}(x_n))^\top$.

5. Рассмотрим модель линейной регрессии

$$\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nm} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \dots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix} =: X\vec{\beta} + \vec{\varepsilon}$$

с $n \geq m$. Напомним, что ключевую роль при оценивании вектора $\vec{\beta}$ играет тот факт, что матрица $Q = X^\top X$ является обратимой.

(i) Докажите, что если $x_{ij} = u_i^{j-1}, i = 1..n, j = 1..m$, где $u_1, .., u_n$ - различные значения (полиномиальная регрессия), то столбцы матрицы X линейно независимы.

(ii) Докажите, что если столбцы матрицы X линейно независимы, то матрица Q положительно определена, то есть $\vec{v}^\top Q \vec{v} > 0$ для любого ненулевого вектора $\vec{v} \in \mathbb{R}^m$.

Комментарий. Обратите внимание, что в (ii) нужно доказать строгое неравенство.

6.* Как известно, в модели линейной регрессии

$$\vec{y} = X\vec{\theta} + \vec{\varepsilon},$$

оценка вектора $\vec{\theta} \in \mathbb{R}^m$ методом наименьших квадратов равна

$$\hat{\vec{\theta}} = X\left(X^\top X\right)^{-1} X^\top \vec{y}.$$

Эта оценка приводит к предсказанным значениям

$$\hat{\vec{y}} = X\hat{\vec{\theta}}.$$

Докажите, что если один из столбцов матрицы X состоит из единиц, то коэффициент детерминации R^2 равен квадрату эмпирического коэффициента корреляции Пирсона между $\vec{y}$ и $\hat{\vec{y}}$,

$$R^2 = \left(\rho(\vec{y}, \hat{\vec{y}})\right)^2.$$