

Регрессионный анализ, введение в R.

- 1 Пусть дана регрессионная модель $Y = X\theta + \varepsilon$, где $E\varepsilon = 0$ и $Var\varepsilon = \sigma^2 V$, где V – некая неотрицательно определённая матрица. Возможно ли приведение этой модели к линейному виду? Если да, то найдите оценку параметра θ методом наименьших квадратов в исправленной модели и матрицу ковариаций этой оценки.
- 2 Пусть дана гауссовская линейная модель $Y = X\theta + \varepsilon$, где $\varepsilon \sim N(0, \sigma^2 I_n)$. Найти оценки параметров θ и σ^2 методом максимального правдоподобия и найти их смещения.
- 3 Сгенерировать нормальную выборку с параметрами 1 и 1 размера 500. Найти ее выборочные среднее и дисперсию. Найти интервал, в который должны попасть 95% наблюдений и подсчитать наблюдения, которые в него не попали.
- 4 В R откройте выданные вам данные. Выберите 3 набора регрессоров и постройте линейную регрессию по этим наборам, выведите коэффициент детерминации получившихся моделей. По каждой из моделей выведите Q-Q plot остатков модели. Можно ли считать, что остатки распределены нормально? (Графики тоже нужно прислать.)
- 5 (Acceptance-Rejection Method) Пусть X имеет плотность $f(x)$ такую, что она не имеет в явной форме выражаемой обратной функции (т.е. сразу сгенерировать случайные величины с такой плотностью мы не можем), кроме того, есть плотность $g(x)$, и генерировать случайные величины с такой плотностью мы умеем, причем $f(x)/g(x) \leq c$. Шаг алгоритма: сгенерировать величину Y с плотностью $g(x)$ и с вероятностью $f(Y)/(c(g(Y)))$ взять ее в качестве X , иначе повторить шаг алгоритма. Почему полученный метод действительно даёт величину с нужной плотностью? Сгенерировать случайную величину с плотностью $f(x) = 20x(1-x)^3$, $x \in [0, 1]$, выбрав в качестве $g(x)$ равномерную плотность на $[0, 1]$.
- 6 Прямой алгоритм отбора признаков по AIC (BIC) производится так: начинаем с 0 признаков в модели, далее подбираем модель с одним признаком, у которой наибольший AIC (BIC), потом варьируем второй признак, считая первый признак уже включенным в модель, и так далее. Откройте выданные вам данные в R и реализуйте на них (использовать `bestglm` нельзя) прямой алгоритм отбора признаков по AIC и BIC. Одинаковые ли модели получились? Как вы это объясните?