Лабораторная работа «Множественная регрессия».
(R)

Цель работы: Для модели множественной регрессии научиться находить оценки ее параметров по методу наименьших квадратов, проводить анализ качества построенной модели.

Теоретическая часть.

Рассмотрим случай одной зависимой переменной Y и p факторов $X^{(1)}, X^{(2)}, ..., X^{(p)}$ и ограничимся рассмотрением простейшей зависимости, когда имеется n наблюдений вида

$$y_i = \sum_{j=1}^{p} x_i^{(j)} \cdot \theta_j + \varepsilon_i, \ i = \overline{1,n},$$

где θ_j, $j = \overline{1,p}$ – неизвестные параметры,

$x_i^{(j)}$, $i = \overline{1,n}$, $j = \overline{1,p}$ – i-тое значение j-того фактора. Функция регрессии (отклика) имеет вид

$$M\{Y \mid X^{(1)}, X^{(2)}, ..., X^{(p)}\} = \eta\left(X^{(1)}, X^{(2)}, ..., X^{(p)}, \theta_1, \theta_2, ..., \theta_m\right) = \sum_{j=1}^{p} x_i^{(j)} \cdot \theta_j$$

Модель может быть записана в матричном виде

$$Y = X^T \theta + \varepsilon,$$

где

$$Y = \begin{bmatrix} y_1 \\ ... \\ y_n \end{bmatrix}, \ \theta = \begin{bmatrix} \theta_1 \\ ... \\ \theta_n \end{bmatrix}, \ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ ... \\ \varepsilon_n \end{bmatrix},$$

$X = \left\| x_i^{(j)} \right\|_{p \times n}$ – матрица планирования эксперимента.

Предположим, что выполнены условия Гаусса-Маркова:

1. $M\varepsilon_i = 0$, $\forall i = \overline{1,n}$,
2. $M\varepsilon_i\varepsilon_j = \begin{cases} \sigma^2, i = j \\ 0, i \neq j \end{cases}$, $\forall i, j = \overline{1,n}$,
3. $D\varepsilon_i = \sigma^2$, $\forall i = \overline{1,n}$,

тогда

- $MY = M\left(X^T\theta + \varepsilon\right) = X^T\theta$ – вектор средних,
- $DY = M\left\{\left(Y - X^T\theta\right)\left(Y - X^T\theta\right)^T\right\} = \sigma^2 I$ – дисперсионная матрица.

МНК оценки имеют вид $\hat{\theta} = \left(XX^T\right)^{-1} XY$.

Математическое ожидание и дисперсия полученных оценок соответственно равны $M\hat{\theta} = \theta$, $D\hat{\theta} = \sigma^2\left(XX^T\right)^{-1}$.

Несмещенная оценка σ^2 определяется формулой

$$S^2 = \frac{1}{n-m}\left(Y - X^T\hat{\theta}\right)^T\left(Y - X^T\hat{\theta}\right).$$

Пусть выдвигается гипотеза H_0: $\theta_i = \theta_i^*$, $i = \overline{1,p}$, где, например, θ_i^* может быть равно истинному значению параметра или $\theta_i^* = 0$, тогда проверяется гипотеза о значимости параметра θ_i.
Тогда при

$$|t| = \left| \frac{\hat{\theta}_i - \theta_i^*}{S\sqrt{\left(XX^T\right)^{-1}_{ii}}} \right| \geq t_{1-\frac{\alpha}{2}, n-m}$$

гипотеза H_0 отклоняется.

Границы доверительного интервала для параметра θ_i, $i = \overline{1,p}$:

$$\hat{\theta}_i - t_{1-\frac{\alpha}{2}, n-m} S\sqrt{\left(XX^T\right)^{-1}_{ii}} < \theta_i < \hat{\theta}_i + t_{1-\frac{\alpha}{2}, n-m} S\sqrt{\left(XX^T\right)^{-1}_{ii}}.$$

Коэффициент детерминации

$$R^2 = 1 - \frac{S_\varepsilon^2}{S_y^2} = 1 - \frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(y_i - \bar{y}\right)^2}.$$

Проверка гипотезы об адекватности уравнения регрессии осуществляется с помощью статистики

$$F = \frac{R^2}{1-R^2}\frac{(n-p-1)}{p},$$

имеющей распределение Фишера с числом степеней свободы $f_1 = p$, $f_2 = n - p - 1$.

Если значение $F < F_{1-\alpha}\left(p, n-p-1\right)$ при заданном уровне значимости α, то принимаем нулевую гипотезу о неадекватности модели.

Задание 1. Сгенерировать полиномиальную модель регрессии, с функцией отклика, описываемой полиномом второй степени.
$y_i = a + bx_i + cx_i^2 + \varepsilon_i$, $i = 1..n$.
Оценить параметры построенной модели, проверить их значимость, оценить общую адекватность модели.
Построить 3d диаграмму рассеяния.

Задание 2. Для набора данных Rent.csv построить парную модель регрессии арендной платы от площади и множественной регрессии арендной платы на все представленные в наборе факторы, кроме района города (distirct). Определить значимые факторы. Построить модель только на значимые факторы. Построить точечный и интервальный прогноз для «своей» квартиры, задав значения факторов произвольно самостоятельно.
Построить 3d диаграмму рассеяния арендной платы от площади и этажа.