# Линейные методы регрессии

Факультатив «Введение в анализ данных и машинное обучение на Python»

7 декабря 2019 г.

## 1 Определение линейной регрессии

Линейная регрессия – модель вида:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} + u_i,$$

где:

1. $u_i$ – случайная ошибка.
2. $\beta_k$ – оцениваемые параметры (коэффициенты).

Оценивается следующая спецификация модели:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + ... + \hat{\beta}_k X_{k,i}.$$

Виды линейной регрессии:

1. $Y_i = \beta_0 + \beta_1 X_{1,i} + u_i$ – парная регрессия.
2. $Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} + u_i$ – множественная регрессия.

**Важно:** линейность – по коэффициентам $\beta_j$!

### 1.1 Задание

Определите, являются ли линейной регрессией следующие модели. Если модель является линейной регрессией, определите её тип. Если модель является парной регрессией, изобразите линию истинной регрессии.

a) $Y_i = 3 + 12X_{1,i} + u_i$.

b) $Y_i = 7 + u_i$.

c) $Y_i = \beta_0 + X_{1,i} + X_{2,i}^{\beta_2} + u_i$.

d) $Y_i = X_{1,i} + \beta_2 X_{2,i} X_{3,i} + u_i$.

e) $Y_i = \beta_1 X_{1,i} + \beta_2 X_{1_i}^2 + u_i$.

f) $Y_i = \beta_0 + \beta_1 X_{1,i} + u_i$.

Истинную модель мы не знаем и никогда не узнаем! Поэтому будем думать об истинной модели в общем виде:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} + u_i$$

### 1.2 Задание

Изобразите графики оценённой линии регрессии. В случае множественной регрессии изобразите эскиз графика для каждой переменной. Поясните, чем являются коэффициенты модели: $\beta$ или $\hat{\beta}$.

a) $\hat{Y}_i = 1 + X_{1,i}$.

b) $\hat{Y}_i = 4X_{1,i}$.

c) $\hat{Y}_i = 1 + 4X_{1,i} - 2X_{2,i}$.

d) $\hat{Y}_i = 12 - 3X_{1,i} - X_{2,i} + 10X_{4,i}$.

Линейную регрессию удобно записывать в матричном виде. Истинная регрессия:

$$Y = X\beta + u,$$

где $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \dots & X_{1,k} \\ 1 & X_{2,1} & X_{2,2} & \dots & X_{2,k} \\ & & \dots & & \\ 1 & X_{n,1} & X_{n,2} & \dots & X_{n,k} \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$, $u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$.

Оценённая регрессия:

$$\hat{Y} = X\hat{\beta}.$$

### 1.3 Задание

Определите размеры всех элементов матричной записи истинной и оценённой регрессии.

### 1.4 Задание

Запишите в матричной записи следующие модели регрессии:

a) $Y_i = \beta_0 + \beta_1 X_{1,i} + u_i$.

b) $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$.

c) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i}$.

d) $\hat{Y}_i = 3 + 12X_{1,i} - 2X_{2,i} + 5X_{3,i} + 0X_{4,i}$

### 1.5 Задание

Запишите в виде линейного уравнения следующие матричные модели:

a) $Y = X\hat{\beta}, X = \begin{pmatrix} 1 & X_{1,1} \\ 1 & X_{2,1} \\ \vdots & \vdots \\ 1 & X_{n,1} \end{pmatrix}, \hat{\beta} = \begin{pmatrix} 2 \\ -4 \end{pmatrix}$.

b) $Y = X\hat{\beta}, X = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} \\ 1 & X_{2,1} & X_{2,2} \\ \vdots & \vdots & \\ 1 & X_{n,1} & X_{n,2} \end{pmatrix}, \hat{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$.

## 2 Оценка модели линейной регрессии

**Важно:** Далее будем считать, что в спецификации модели есть константный признак (вектор единиц). Выражения $e_i = Y_i - \hat{Y}_i$ называются **остатками** регрессии. Соответственно, $e = (Y - \hat{Y})$ – вектор остатков. Идея: минимизировать следующее выражение (метод наименьших квадратов, МНК), то есть функцию потерь:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \dots + \hat{\beta}_k X_{k,i})^2 \to \min_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k}.$$

В матричном виде:

$$||e||^2 = ||Y - X\hat{\beta}||^2 \to \min_{\beta}$$

### 2.1 Задание

Геометрически докажите, что верна следующая формула:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

### 2.2 Задание

Найдите оценки коэффициентов для следующей модели:

$$Y = X\beta + u.$$

$$\hat{Y} = X\hat{\beta}.$$

$$Y = \begin{pmatrix} 4 \\ 6 \\ 3 \end{pmatrix}, X = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 2 \end{pmatrix}$$

## 3 Проблемы модели линейной регрессии

1. **Мультиколлинеарность** – ситуация, когда признаки линейно зависимы, то есть

   $$X_j = a + bX_m,$$

   где $a \in \mathbb{R}$, $b \in \mathbb{R}$.

   На практике плоха и ситуация, когда признаки обладают высокой корреляцией.

   Проблемы:

   a) Хранение избыточной информации.

   b) Вредит некоторым методам машинного обучения.

   Решение: отбор признаков.

2. **Переобучение**: на практике – когда оценённые коэффициенты слишком большие.

   Решение: регуляризация:

   (a) $L_1$-регуляризация (lasso): $||Y - X\hat{\beta}||^2 + \sum_{j=1}^{k} |\beta_j| \to \min_{\beta}$.

   (b) $L_2$-регуляризация (ridge): $||Y - X\hat{\beta}||^2 + \sum_{j=1}^{k} \beta_j^2 \to \min_{\beta}$.

   $L_1$-регуляризация позволяет отобрать признаки, так как зануляет некоторые оценки коэффициентов.

3. **Проблемы со сходимостью:** на практике оценка модели происходит не по аналитической формуле, а методами оптимизации. Для лучшей сходимости рекомендуется привести признаки к одинаковому масштабу.

   Выборочное среднее:

   $$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}.$$

Выборочная дисперсия:

$$sVar = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Выборочное стандартное отклонение:

$$s\sigma = \sqrt{sVar}.$$

Масштабирование (стандартизация):

$$\tilde{X}_j = \frac{X_j - \bar{X}_j}{s\sigma_{X_j}}.$$

Также можно масштабировать к максимальному значению признака (то есть делить на максимальное значение признака):

$$\tilde{X}_{jm} = \frac{X_{jm}}{\max_m\{X_{jm}\}}$$

### 3.1 Задание

Найдите выборочное среднее, выборочную дисперсию и выборочное стандартное отклонение для следующих выборок:

a) $1, 2, 3, 4.$

b) $0, 0, 1, -2.$

c) $1, 2.$

d) $-1, 2, 2, 2, 2.$

## 4 Функционалы (метрики) качества линейной регрессии

1. Среднеквадратичная ошибка (Mean Squared Error, MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$$

2. Средняя абсолютная ошибка (Mean Absolute Error, MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i|$$

   Выше устойчивость к выбросам.

3. $RMSE = \sqrt{MSE}$

4. $R^2$ – коэффициент детерминации:

$$R^2 = \frac{||\hat{Y} - \bar{Y}||^2}{||Y - \bar{Y}||^2} = 1 - \frac{||Y - \hat{Y}||^2}{||Y - \bar{Y}||^2}.$$

   Показывает долю дисперсии, объяснённой моделью, в общей дисперсии зависимой переменной. Свойства:

   - $R^2 \in [0, 1]$ для разумных моделей.

- $R^2 = 1$ – идеальная модель.
- $R^2 = 0$ – модель предсказывает на уровне константной.
- $R^2 < 0$ – модель предсказывает хуже константной.

## 4.1 Задание

Найдите MSE, MAE, RMSE и $R^2$, если имеется следующая выдача:

a) $Y = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$, $\hat{Y} = \begin{pmatrix} 4 \\ 0 \\ 3 \end{pmatrix}$

b) $Y = \begin{pmatrix} 7 \\ -1 \\ 9 \end{pmatrix}$, $\hat{Y} = \begin{pmatrix} 5 \\ 5 \\ 5 \end{pmatrix}$

c) $Y = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}$, $\hat{Y} = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}$

d) $Y = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$, $\hat{Y} = \begin{pmatrix} 10 \\ -10 \\ -9 \end{pmatrix}$