CHAP2 线性模型（LINEAR MODEL）

——回归问题

一、基本形式

给定由d个属性描述的示例 $\boldsymbol{x}=(x_1,x_2,\ldots,x_d)$，其中 x_i 是 $\boldsymbol{x}$ 在第 i 个属性上的取值，线性模型（Linear Model）试图学得一个通过属性的线性组合来进行预测的函数，即：

$$f(\boldsymbol{x})=w_1x_1+w_2x_2+\ldots+w_dx_d+b \tag{1}$$

一般用向量的形式写成：

$$f(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}+b \tag{2}$$

线性模型形式简单、易于建模，但却蕴涵着机器学习中一些重要的基本思想， 许多功能更为强大的非线性模型(Nonlinear Model)可在线性模型的基础上通过引入层级结构或高维映射而得。此外，由于 $\boldsymbol{w}$ 直观表达了各属性在预测中的重要性，因此线性模型有很好的可解释性(Comprehensibility) / 可理解性 (Understandability) 。

本章节将介绍几种经典的线性模型，包括了**回归（Regression）**和**分类（Classification）**。

二、回归模型

2.1 线性回归模型： 一元线性回归

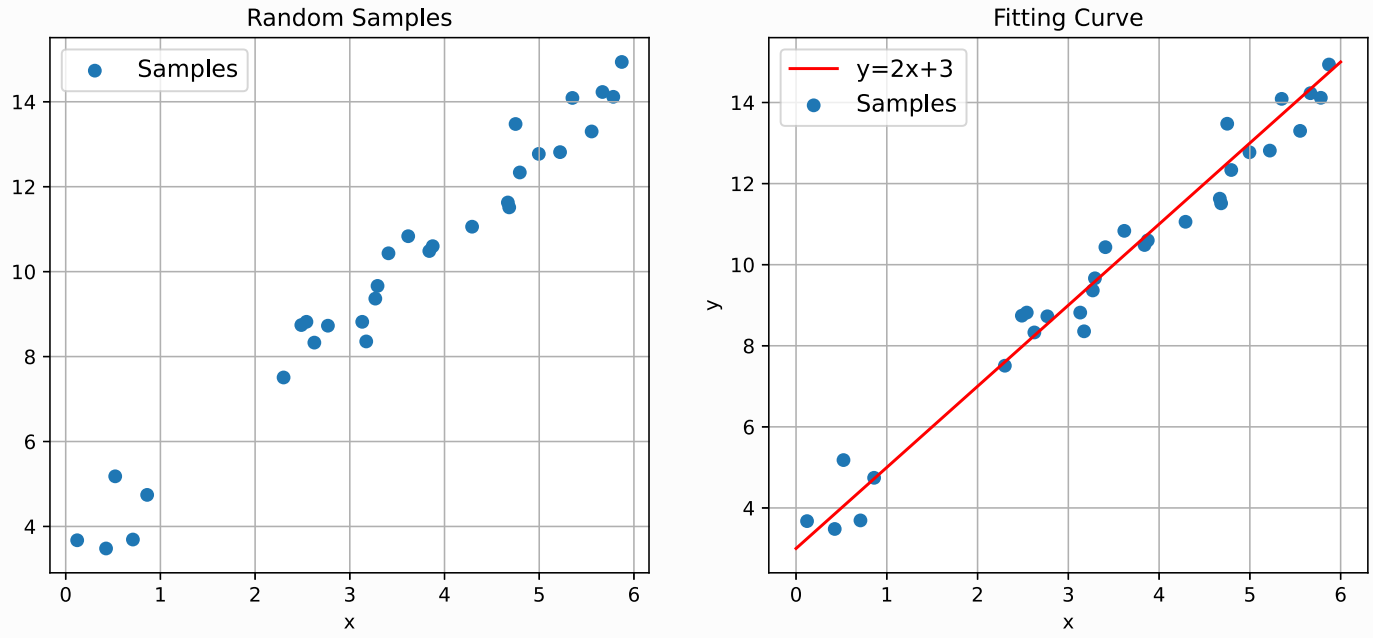
给定数据集$D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_d,y_d)\}$，$y_i\in\mathbb{R}$。“线性回归”试图学习一个线性模型以尽可能准确地预测实值输出标记。如：通过历年的出生人口数据预测2022年出生人口数量等等类似问题。

先考虑**一元线性回归**：

$$y_i=b+wx_i+\varepsilon_i，\ 其中\ \varepsilon_i\sim N(0,\sigma^2) \tag{3}$$

- 模型中，y_i 是 x_i 的线性函数（部分）加上误差项，线性部分反映了由于 x 的变化而引起的 y 的变化。
- 误差项 ε_i 是随机变量：反映了除 x 和 y 之间的线性关系之外的随机因素对 y 的影响，是不能由 x 和 y 之间的线性关系所解释的变异性。
- b 和 w 称为模型的参数。
- 任意两组残差项 ε_i 与 ε_j之间没有任何关系，即：
  $Cov(\varepsilon_i,\varepsilon_j)=0,\ i\neq j,\ \ i,j=1,2,3,\ldots$
- 任意一组残差项 ε_i 与 $\boldsymbol{x}$ 无关，即：$Cov(\varepsilon_i,\boldsymbol{x})=0,\ i=1,2,3,\ldots$

举个例子说明回归要做什么？：二维平面上，现在下左图有一些散点，我们需要找一条直线模拟它们的函数关系，使得这些散点与这条直线的距离总和最小，这条直线就是最佳回归曲线。而右图就是通过**最小二乘法**得到的回归曲线：$f(x)=2x+3$。



从数学角度来看，考虑：$y=wx+b+\varepsilon$，$\varepsilon\sim N(0,\sigma^2)$，$w$，$b$，$\sigma$ 都是不依赖于 x 的变量，如何确定 w 和 b 呢？可以利用**均方误差（MSE）**来衡量 $f(x)$ 与 y 之间的差别，我们希望的是在训练集上的模型产生的值能和真实值尽可能接近（此处暂不考虑模型复杂度、过拟合问题）。均方误差是回归任务中最常用的度量性能，因此我们尽可能让均方误差

最小化，即：

$$
\begin{aligned}
(w^*, b^*) &= \arg\min_{(w,b)} \sum_{i=1}^{d} (f(x_i) - y_i)^2 \\
&= \arg\min_{(w,b)} \sum_{i=1}^{d} (y_i - wx_i - b)^2
\end{aligned}
\tag{4}
$$

> Remark：基于均方误差最小化来进行模型求解的方法被称为“**最小二乘法**”（Least Square Method）。在高中时期，我们学习到最小二乘法就是试图找到一条直线，使得所有样本到直线上的欧氏距离之和最小。

求解 (w^*, b^*) 使得 $E_{(w,b)} = \sum_{i=1}^{d} (y_i - wx_i - b)^2$的过程称为**“线性回归”**。我们可以对 $E_{(w,b)}$ 分别对参数 w 和 b 求导，得到

$$
\begin{aligned}
\frac{\partial E_{(w,b)}}{\partial w} &= 2\left(w \sum_{i=1}^{d} x_i^2 - \sum_{i=1}^{d} (y_i - b) x_i\right), \\
\frac{\partial E_{(w,b)}}{\partial b} &= 2\left(db - \sum_{i=1}^{d} (y_i - wx_i)\right),
\end{aligned}
\tag{5}
$$

令（5）式等于 0，则可以得到最优解：

$$
\begin{aligned}
w^* &= \frac{\sum_{i=1}^{d} y_i (x_i - \bar{x})}{\sum_{i=1}^{d} x_i^2 - \frac{1}{d} \left(\sum_{i=1}^{d}\right)^2} \\
b^* &= \frac{1}{d} \sum_{i=1}^{d} (y_i - wx_i)
\end{aligned}
\tag{6}
$$

其中，$\bar{x} = \frac{1}{d} \sum_{i=1}^{d} x_i$ 为 x 的均值。

> 注：在统计学中，w^* 和 b^* 是待估计参数 w 和 b 的**无偏的、有效的、一致估计**。

为了方便计算，引入下面符号：

$$
\begin{aligned}
S_{xx} &= \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2 \\
S_{yy} &= \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}y_i^2 - \frac{1}{n}(\sum_{i=1}^{n}y_i)^2 \\
S_{xy} &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}x_iy_i - \frac{1}{n}(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)
\end{aligned} \tag{7}
$$

为了检验线性回归模型的好坏，可以引入以下变量：

TSS(Total Sum of Squares)表示实际值与期望值的总离差平方和，代表**变量的总变动程度**

$$
\mathbf{TSS} = Q_{总} = S_{yy} = \sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{8}
$$

ESS(Explained Sum of Squares)表示预测值与期望值的离差平方和，代表**预测模型拥有的变量变动程度**

$$
\mathbf{ESS} = Q_{回} = \frac{S_{xy}^2}{S_{xx}} = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{9}
$$

RSS(Residual Sum of Squares)表示实际值与预测值的离差平方和，代表**变量的未知变动程度**

$$
\mathbf{RSS} = Q_{剩} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{10}
$$

整体变量的总变动程度（TSS）可以划分为两部分：**模型模拟的变动程度（ESS）**和 **未知的变动程度（RSS）**。通常来说，预测模型拥有的变量变动程度在总变动程度中的占比越高，代表模型越准确，当RSS=0时，表示模型能完全模拟变量的总变动。

R^2 **拟合优度**表示建立的模型拥有的变动程度能模拟总变动程度的百分比，该指标越靠近 1，则模型的效果越好。

$$
R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \tag{11}
$$

> 注：有了上述统计量，参数 w^* 和 b^* 表达式可以改写为：

$$w^* = \frac{S_{xy}}{S_{xx}}, \quad b^* = \bar{y} - \bar{x} \cdot w^* \tag{12}$$

案例1：房子价值预测问题（详细见Linear_Regression.py、House_Price.ipynb、House_Price_Sklearn.ipynb）

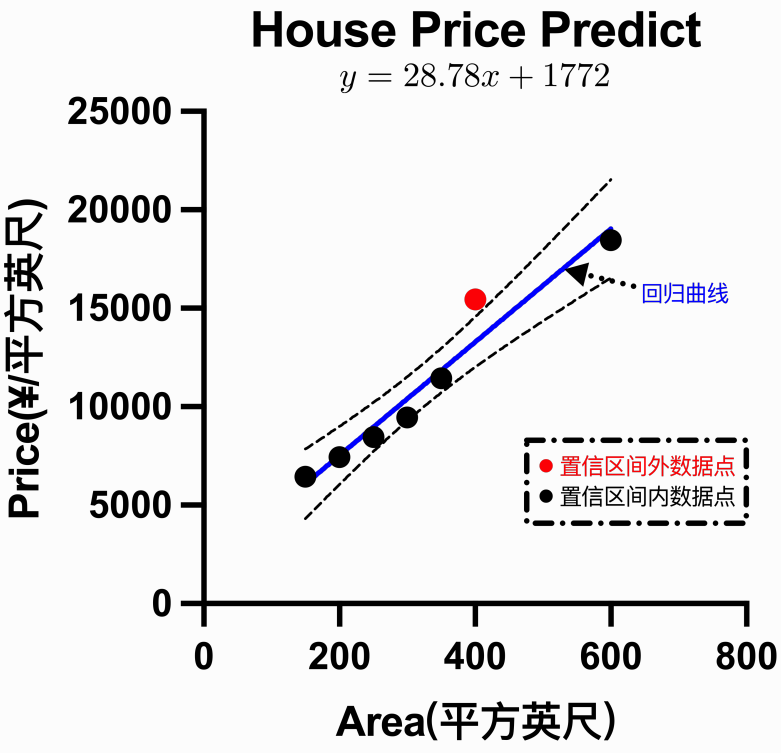
- Linear_Regression.py、House_Price.ipynb：自己编写的一元线性回归模块以及调用
- House_Price_Sklearn.ipynb：调用sklearn模块进行线性回归

均得到：$w^* = 28.78$，$b^* = 1772$，$R^2 = 0.9447$

置信区间是指由样本统计量所构造的总体参数的估计区间。在统计学中，一个概率样本的**置信区间（Confidence interval）**是对这个样本的某个总体参数的区间估计。置信区间展现的是这个参数的真实值有一定概率落在测量结果的周围的程度，其给出的是被测量参数的测量值的可信程度，即前面所要求的"一个概率"。

> 95%置信水平可以简单理解为：样本数目不变的情况下，做一百次试验，有95个置信区间包含了总体真值。

下图是房价预测作图，"虚线"之间代表95%水平的置信区间，"蓝线"代表着回归（拟合）曲线。



（95%置信水平的置信区间，使用GraphPad Prism软件验证结果并作图）

2.2 线性回归模型： 多元线性回归

更一般的情况，设数据集 $D=\{(\boldsymbol{x_1},y_1),(\boldsymbol{x_2},y_2),\ldots,(\boldsymbol{x_m},y_m)\}$， 其中 $\boldsymbol{x_i}=(x_{i1},x_{i2},\ldots,x_{id})^{\mathrm{T}}$， $y_i\in\mathbb{R}$ ，每个样本 $\boldsymbol{x_i}$ 由 d 个属性描述。此时，我们试图学得

$$f(\boldsymbol{x_i})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x_i}+b,\ 使得 f(\boldsymbol{x_i})\simeq y_i \tag{13}$$

这称为**“多元线性回归”（Multivariate Linear Regression）**。其中，回归系数 $\boldsymbol{w}=(w_1,w_2,\ldots,w_d)$，偏置项 $b\in\mathbb{R}$ 。

求解上述模型中的参数 $\boldsymbol{w}$ 和 b ，可以按照求解一元线性回归时的思路：利用最小二乘法来对 $\boldsymbol{w}$ 和 b 进行估计。首先，对于每一个样本 $(\boldsymbol{x_i},y_i)$ ，考虑该样本的**均方误差（MSE）**：

$$E_i=[\ y_i-f(\boldsymbol{x_i})\ ]^2=(y_i-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-b)^2,\ i=1,2,\ldots,m \tag{14}$$

同时， 为了计算与讨论方便，我们将 $\boldsymbol{w}$ 和 b 吸收入向量形式 $\hat{\boldsymbol{w}}=(\boldsymbol{w}^{\mathrm{T}};b)$；相应地，把数据集 D 表示为一个 $m\times(d+1)$ 大小的矩阵 $\mathbf{X}$ ，其中每一行对应着一个样本，每行的前 d 个元素对应于样本的 d 个属性值，最后一个元素为了和 $\hat{\boldsymbol{w}}$ 相乘对应，全部设置为 1 ，即：

$$\mathbf{X}=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1d}&1\\x_{21}&x_{22}&\cdots&x_{2d}&1\\\vdots&\vdots&\ddots&\vdots&\vdots\\x_{m1}&x_{m2}&\cdots&x_{md}&1\end{bmatrix}=\begin{bmatrix}\boldsymbol{x}_1^{\mathrm{T}}&1\\\boldsymbol{x}_2^{\mathrm{T}}&1\\\vdots&\vdots\\\boldsymbol{x}_m^{\mathrm{T}}&1\end{bmatrix}_{m\times(d+1)} \tag{15}$$

再把标签值（因变量值）也写成向量形式 $\boldsymbol{y}=(y_1,y_2,\ldots,y_m)$。于是，最小化所有样本的均方误差和：

$$
\begin{aligned}
\hat{\boldsymbol{w}}^* &= \arg\min_{\hat{\boldsymbol{w}}} \sum_{i=1}^{m} E_i \\
&= \arg\min_{\hat{\boldsymbol{w}}} \sum_{i=1}^{m} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i - b)^2 \\
&= \arg\min_{\hat{\boldsymbol{w}}} \sum_{i=1}^{m} [\ y_i - \hat{\boldsymbol{w}} \begin{pmatrix} x_i \\ 1 \end{pmatrix} \ ]^2 \\
&= \arg\min_{\hat{\boldsymbol{w}}} ||\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}||_2^2 \\
&= \arg\min_{\hat{\boldsymbol{w}}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})
\end{aligned} \tag{16}
$$

令 $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$，对 $\hat{\boldsymbol{w}}$ 求导得到：

$$
\begin{aligned}
\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} &= -\mathbf{X}^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (-\mathbf{X}) \\
&= -\mathbf{X}^{\mathbf{T}} \boldsymbol{y} + \mathbf{X}^{\mathbf{T}} \mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y}^{\mathrm{T}} \mathbf{X} + \hat{\boldsymbol{w}}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \\
&= -2\mathbf{X}^{\mathrm{T}} \boldsymbol{y} + 2\mathbf{X}^{\mathbf{T}} \mathbf{X}\hat{\boldsymbol{w}} \\
&= 2\mathbf{X}^{\mathrm{T}} (\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y})
\end{aligned} \tag{17}
$$

令（17）式等于 0，可以得到 $\hat{\boldsymbol{w}}$ 的最优解：

- 当 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 为满秩矩阵（Full-rank Matrix）或者正定矩阵（Positive Definite Matrix）时，最优解为

  $$
  \hat{\boldsymbol{w}}^* = (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{y} \tag{18}
  $$

  最后习得得线性回归模型为

  $$
  f(\hat{\boldsymbol{x}}_i) = \hat{\boldsymbol{x}}_i^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{y} \tag{19}
  $$

- 当 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 并非满秩矩阵时，如矩阵 $\mathbf{X}$ 的列数远远多于行数，$\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 显然不满秩。此时可以解出多个 $\hat{\boldsymbol{w}}$，并且这些 $\hat{\boldsymbol{w}}$ 都满足均方误差最小化。此时，我们可以考虑引入正则化（Regularization）项，如：

  $$
  \hat{\boldsymbol{w}}^* = \arg\min_{\hat{\boldsymbol{w}}} ||\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}||_2^2 + \frac{\lambda}{2} ||\hat{\boldsymbol{w}}||_2^2 \tag{20}
  $$

> 现实任务中的 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 往往不是满秩矩阵，例如我们在许多任务中会遇到大量的样本属性，其数目远远超过样本数，这样就会导致 $\mathbf{X}$ 的列数多于行数，再因为 $rank(\mathbf{X}) = rank(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ ，$\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 显然不会满秩。生物信息学的基因芯片数据中常有成千上万个属性，但往往只有几十个、上百个样本数。

案例2：多种因素对商品销售额的影响（详细见）

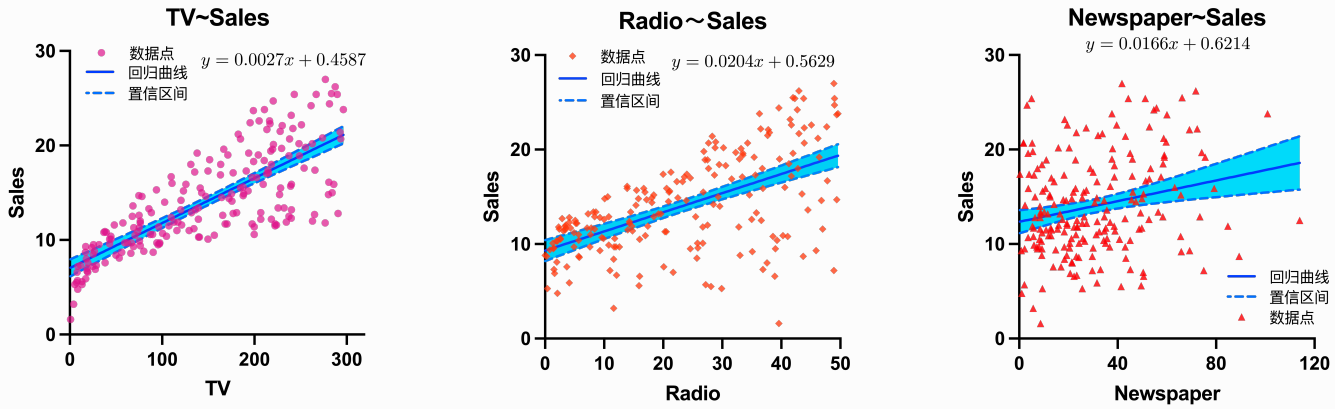
当结果值的影响因素有多个时，可以采用多元线性回归模型。例如，商品的销售额可能与电视广告投入、收音机广告投入和报纸广告投入有关系，可以有：

$$Sales = \beta_0 + \beta_1 \cdot TV + \beta_2 \cdot Radio + \beta_3 \cdot Newspaper \tag{21}$$

文件中的数据集Advertising.csv来自这儿，大家可以自行下载。

（一）分析数据

本数据集合共有200个观测值，每一个观测值对应着一个市场的情况。在这个案例中，通过不同的广告投入，预测产品销售。其中一共有3种不同广告投入，分别是：TV、Radio以及Newspaper。先单独看看它们对销售额的影响情况：

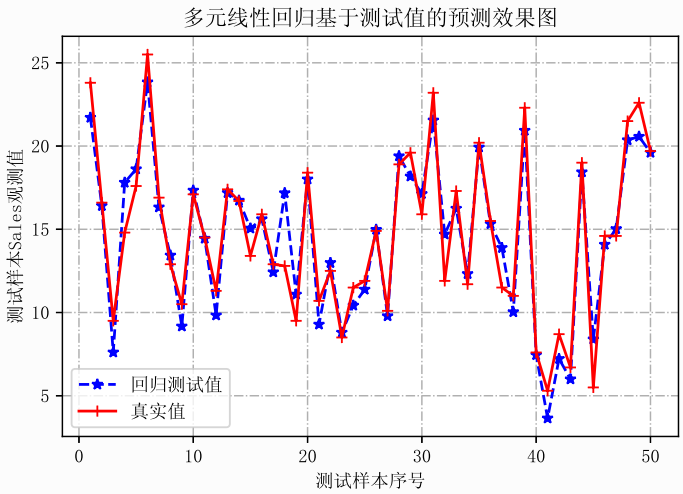


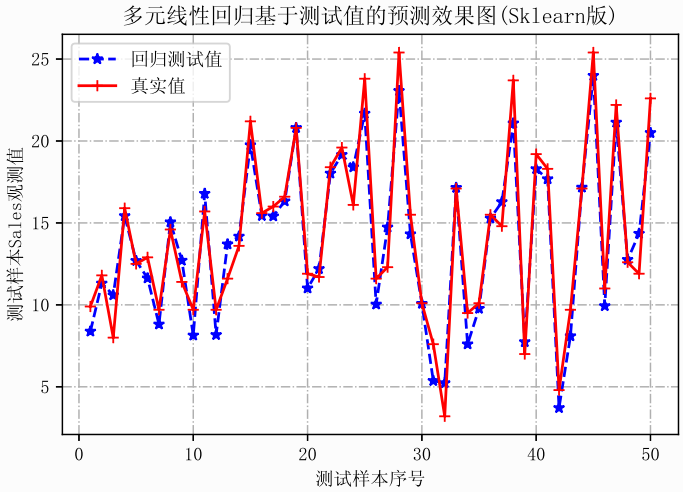
可以发现，TV这个属性和销售量的线性关系比较强，Radio次之，Newspaper稍差一些。

接下来就可以编程实现验证一下：

- Multi_Linear_Regression.py、Advertising.ipynb：自己编写的多元线性回归模块以及调用
- Advertising_sklearn.ipynb：调用sklearn模块进行多元线性回归

结果展示：基于测试集上的预测效果图，左一是自己编写的模型，右侧是使用了sklearn模块。





> 上图根据不同测试集和训练集的划分，可能会导致结果略微不同，可以设置随机数种子保证随机划分的一致性。

三、后续补充说明

岭回归与**Lasso回归**的出现是为了解决线性回归出现的过拟合以及在通过正规方程方法求解系数的过程中出现的 $\mathbf{X}$ 转置乘以 $\mathbf{X}$ 不可逆这两类问题的，这将在后续介绍**梯度下降法**之后再详细介绍。