在回归问题中，线性模型用特征的线性组合来拟合数据， $\hat{y}$ 表示模型预测值。w 表示线性组合的系数。

$$\hat{y}(w,x) = w_0 + w_1 x_1 + \cdots + w_p x_P$$

1. 普通最小二乘

用累计均方误差作为模型损失函数, w 是需要学习的权重系数向量，n为样本数目。 $w = (w_0, w_1, \cdots, w_p)^\intercal$ ， $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\intercal$ ，$X = (X_1, \cdots, X_i, \cdots, X_n)^\intercal$ 。

$$\min_w \sum_{i=1}^{n} \|Xw - y\|_2^2$$

累计均方误差是关于系数w的凸函数，损失函数的极值点即为最值点。求解最优解有两种方法，设 $J(w) = \min_w \sum_{i=1}^{n} \|Xw - y\|_2^2$ ：

1.直接计算损失函数关于 w 的导数值，在导数为0的条件下，直接解出系数 w 。

$$\frac{\partial J}{\partial w} = 2X^\intercal(Xw - y) = 0$$
$$w = (X^\intercal X)^{-1} X^\intercal y$$

2.使用梯度下降算法这一类的数值法求解。

$$w = w - \alpha \frac{\partial J}{\partial w}$$

2. 岭回归

对损失函数添加正则项，通过调整 λ 大小， λ 越大对 w 的惩罚越大， w 越趋于 0 ：

$$\min_w \sum_{i=1}^{n} \|Xw - y\|_2^2 + \lambda \|w\|_2^2$$

对正则项的两种理解:

1.从约束优化的角度理解问题，原问题可以看作是如下问题的近似。利用罚函数法求解等式约束问题，将等式约束作为惩罚项加上目标函数。 λ 越大，两个问题的解越接近，当λ趋于无穷大时，两个问题等价。正则项从这个角度可以理解为在求解过程中对 $w = 0$ 的偏好，即对简单模型的偏好程度。

$$\begin{aligned} \min_w \quad & \sum_{i=1}^{n} \|Xw - y\|_2^2 \\ subject\ to \quad & w = 0 \end{aligned}$$

2.从贝叶斯学派的观点出发将参数 w 也看成随机变量，需要知道参数 w 的先验分布 $p(w)$ 通过贝叶斯公式求出已知 x 的后验概率分布。在实际中选择 w 的最大后验点估计。

$$p(w \mid x)=\frac{p(x \mid w) \cdot p(w)}{p(x)}$$
$$w_{MAP}=arg \max_{w} \ \log p(x \mid w)+\log p(w)$$

若w 的先验分布为正态分布， $w \sim N(w; 0, \frac{1}{\lambda} I^2)$ ，线性回归模型对应分布族。

$$p(y \mid x, w)= \ N(y; w^{\top} x, I)$$

$$w_{MAP}= \ arg \max_{w} \ \frac{1}{(2\pi)^{\frac{n}{2}}|I|^{\frac{1}{2}}} \exp \left\{-\frac{1}{2}(y-w^{\top} x)^{\top} I^{-1}(y-w^{\top} x)\right\}$$

$$+\frac{1}{(2\pi)^{\frac{n}{2}}|I|^{\frac{1}{2}}} \exp \left\{-\frac{1}{2} w^{\top} \lambda I^{-1} w\right\}$$

$$w_{MAP}= \ arg \min_{w} \ \sum_{i=1}^{n}\|y-w^{\top} x\|_2^2+\lambda\|w\|_2^2$$