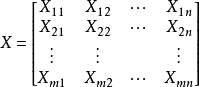
通过下图所示代码，我们希望能通过OLS即最小二乘法解决线性回归问题，这是一个极其经典的通过构造Loss Function并梯度下降从而解决回归问题的机器学习或数据挖掘算法。

此下给出最小二乘法的推导过程：

考虑超定方程组（超定指未知数小于方程个数）：https://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D199/sign=d1bb41dc80d6277fed12363111391f63/21a4462309f79052ce9892590bf3d7ca7acbd5d6.jpg

其中m代表有m个等式，n代表有 n 个未知数 https://gss3.bdstatic.com/-Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D17/sign=62db12086009c93d03f20af09e3d6cb4/b999a9014c086e06b659674d05087bf40bd1cbc7.jpg ，m>n ；

将其进行向量化后为：https://gss3.bdstatic.com/-Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D49/sign=e9d6c22fecc4b7453094b61fcefcadf2/14ce36d3d539b600d9704bb1ee50352ac65cb77e.jpg

 ， https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D88/sign=1b808bc4b5119313c343f2b86438c5ac/7e3e6709c93d70cf7304ef0bffdcd100baa12b3d.jpg ， https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D85/sign=a3688735748da9774a2f8b2eb151a361/e850352ac65c1038babb8ac4b5119313b17e89eb.jpg。

显然该方程组一般而言没有解，所以为了选取最合适的 https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D17/sign=24e9c30334fa828bd52399e4fc1f1a41/dbb44aed2e738bd42bcbf537a68b87d6267ff9ca.jpg 让该等式"尽量成立"，引入残差平方和函数Shttps://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D119/sign=f5b5ca279413b07eb9bd540935d79113/91ef76c6a7efce1b935185cca851f3deb48f65b9.jpg（在统计学中，残差平方和函数可以看成n倍的[均方误差](https://baike.baidu.com/item/%E5%9D%87%E6%96%B9%E8%AF%AF%E5%B7%AE" \t "_blank)MSE）。

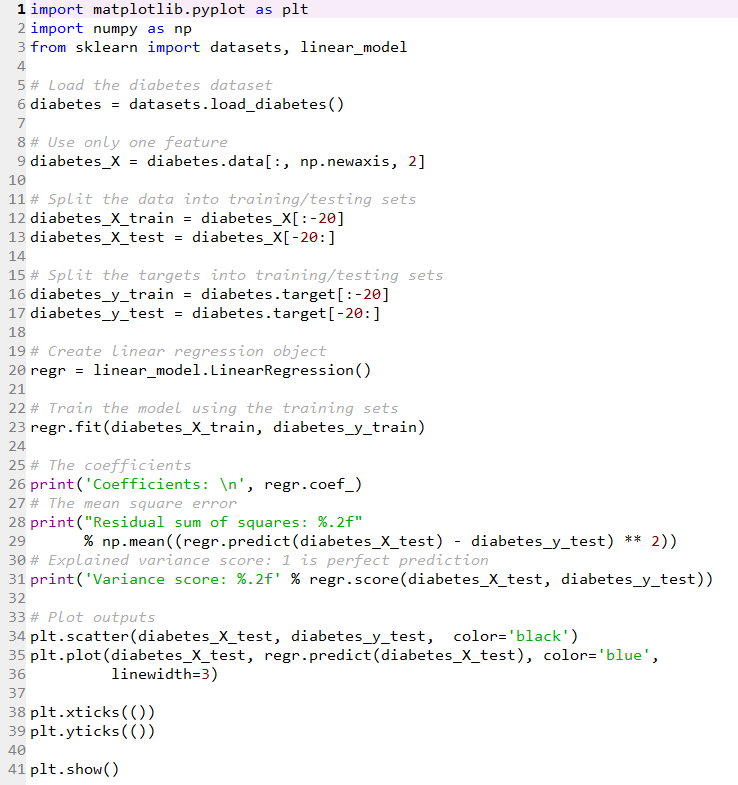
当https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D38/sign=6d7e8f35968fa0ec7bc7620527979b99/77c6a7efce1b9d16f254d767f4deb48f8c546420.jpg 时， https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D30/sign=db943f26be12c8fcb0f3f0cdfc036100/63d9f2d3572c11df00f3f225642762d0f703c292.jpg 取最小值，记作：https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D124/sign=b500c0cb5566d0167a199a2aa32ad498/728da9773912b31b3c8a8e2f8118367adbb4e1d6.jpg

通过对 https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D30/sign=db943f26be12c8fcb0f3f0cdfc036100/63d9f2d3572c11df00f3f225642762d0f703c292.jpg 进行微分 [2]  求最值，可以得到：

https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D94/sign=6b8addbd80d6277fed123e3c2938b4f7/a044ad345982b2b73c2970c236adcbef76099b06.jpg

如果矩阵 https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D34/sign=9cd5c9b98b5494ee8322091d2df53670/5bafa40f4bfbfbedc46f7e797ff0f736afc31f85.jpg 非奇异则 https://gss3.bdstatic.com/-Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D17/sign=62db12086009c93d03f20af09e3d6cb4/b999a9014c086e06b659674d05087bf40bd1cbc7.jpg 有唯一解 [3]  ：

https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D121/sign=56983e206a224f4a5399771138f69044/9213b07eca806538e8b4ec1590dda144ad34825e.jpg



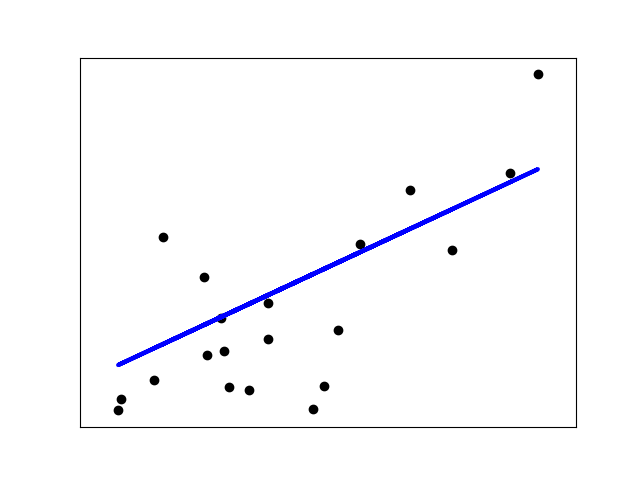
1. 我们这里引用sklearn中datasets包创造了一个简单的数据集；
2. 将数据切分为train训练数据中x与y，test测试数据中x与y；
3. 通过sklearn中linear\_model.LinearRegression()进行构造一个最小二乘方法的模型建模，并将训练数据fit入regr模型中；
4. 通过模型建立完后的api，我们可以通过regr.coef\_、方差计算、regr.score得到如下结果：

Coefficients: [938.23786125]

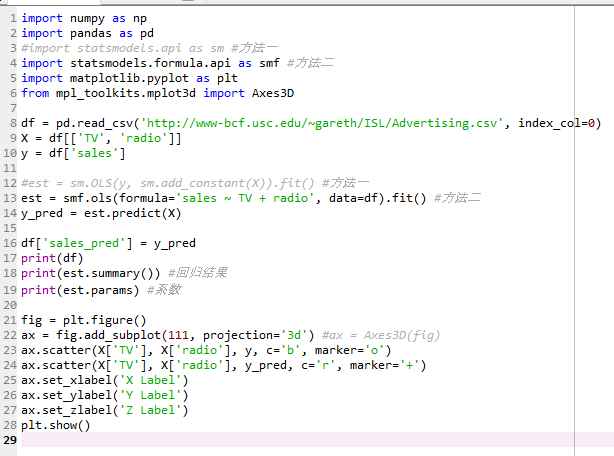
Residual sum of squares: 2548.07

Variance score: 0.47

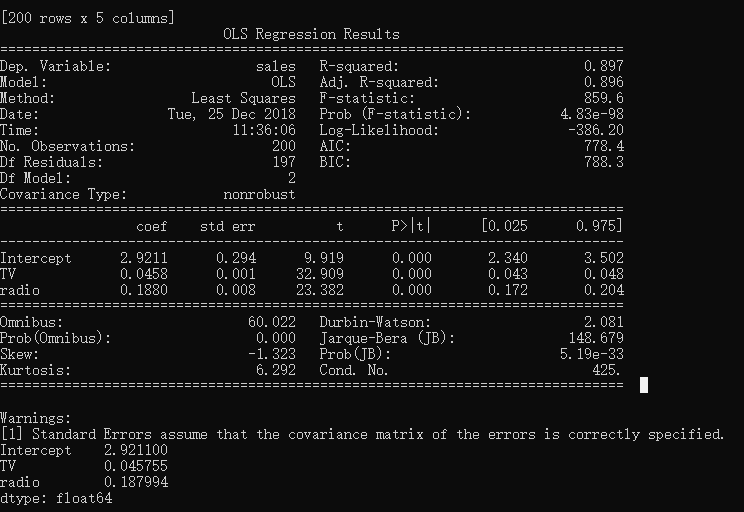
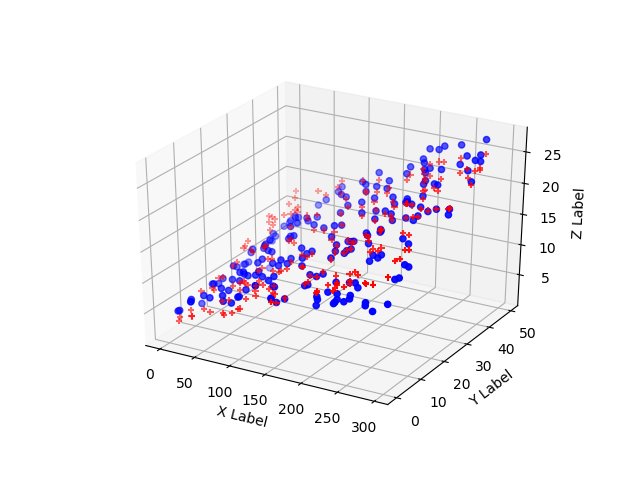
1. 最后一步我们通过matplotlib.pyplot将数据点与预测的模型展示在散点图上。

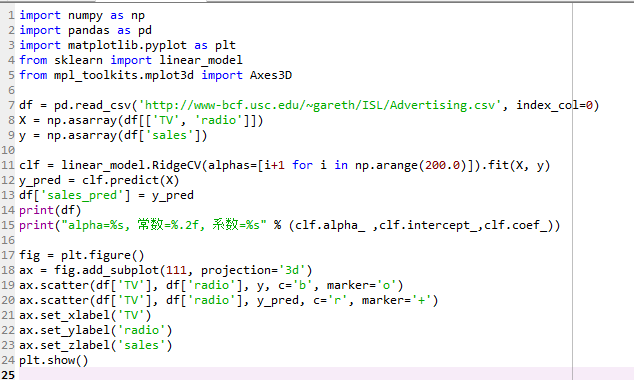


此为我们通过sklearn包中带有的数据集分析，然后我们又找到了两组经典的数据集，然后尝试着解决新环境问题。



此为我们通过csv格式文件录入数据，然后通过statsmodels（经常作为计量经济学和统计学的建模工具）进行建模，随后根据保重的summary()、params可以得到如图所示的回归结果及其系数结果，我们可以参照数据的可视化作为分析依据。



岭回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得[回归系数](https://baike.baidu.com/item/%E5%9B%9E%E5%BD%92%E7%B3%BB%E6%95%B0/10840879)更为符合实际、更可靠的回归方法，对病态数据的拟合要强于[最小二乘法](https://baike.baidu.com/item/%E6%9C%80%E5%B0%8F%E4%BA%8C%E4%B9%98%E6%B3%95/2522346)。

现在，我们已经预测了两组分别为sklearn与statsmodel进行了最小二乘法建模，那么我们是否可以进行Right Regression即岭回归进行分析，看看其的优化。于是这里改进了一下，运用了sklearn中linear\_model.RidgeCV()，进行了岭回归建模，以下是clf.alpha\_ ,clf.intercept\_,clf.coef\_从而得到的回归结果分析，同样的参照数据的可视化作为分析依据。

C:\Users\Administrator\Desktop\homework\OLS\5.png