华东师范大学计算机科学技术系实验报告

课程名称：数据分析与数据挖掘	**年级**：大二	**实践作业成绩**：
指导教师：兰曼	**姓名**：陈越	**提交作业日期**：2017-4-23
实践编号：9	**学号**：10152130155	**实践作业编号**：3

一、 实验名称：数据分析和数据采集

台风路径预测：运用并比较多种线性回归策略提高台风路径预测的准确性。

二、 实验目的

通过根据输入数据直接编写回归函数或者调用SKlearn库的回归函数熟练掌握实现对一个目标值较好的预测方法。

通过对比多种回归函数了解并理解各种回归函数的异同，明白各种回归函数的特点。

在实现过程进一步理解并掌握各种回归算法。

通过对数据进行预测更深层地理解回归函数的特性：实现不复杂，对线性数据有比较好的预测效果。

三、 实验内容

调用编写的函数打开数据所在的文件，将数据（间隔12小时采集到的台风预报因子，台风的经纬度）分别输入到训练数据集与测试数据集中。

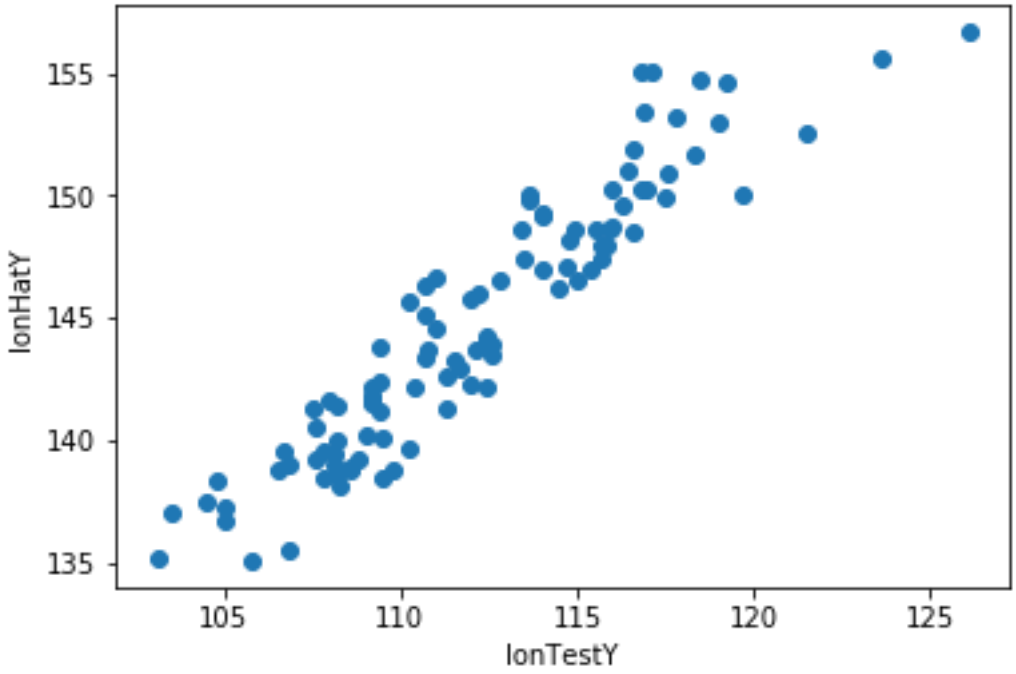
然后调用各种回归函数用训练数据集对回归参数进行训练，然后再利用回归参数对测试数据集进行预测，之后再对预测值与真实值进行误差分析从而评价该回归函数的优劣。

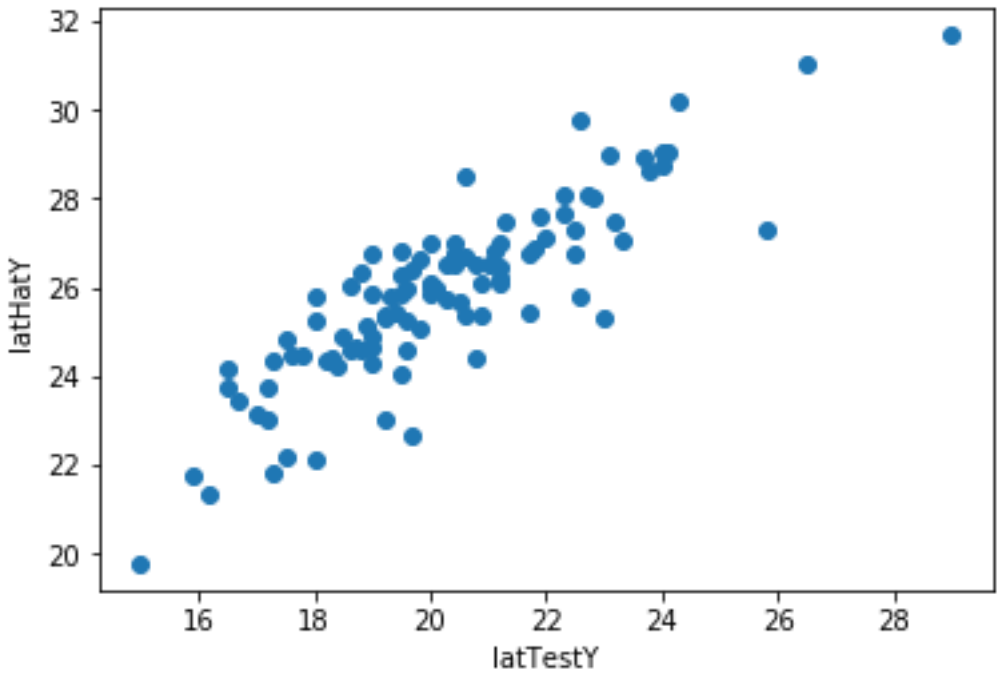
通过引用课堂内容上成熟度代码，实验主要调用了以下回归函数：标准回归训练，SKlearn里的最小二乘法线性回归算法，岭回归线性回归算法，Lasso线性回归算法，ElasticNet线性回归算法，SGDregressor回归算法。实验代码中对每一种回归算法都计算了误差平方和用于相互对比。还将真实值与预测值分别分布在x轴和y轴上，用以观察相关性。

通过多次对原始数据进行shuffle，然后调用多种不同的回归算法进行回归，记录每一次shuffle后的数据相关性与误差，如此反复多次；得到的结果再与别的回归算法进行对比，观察并分析各个算法的情况。

四、 实验结果及分析

一 准回归训练（手动编写代码）的实验结果：（80%为训练集）



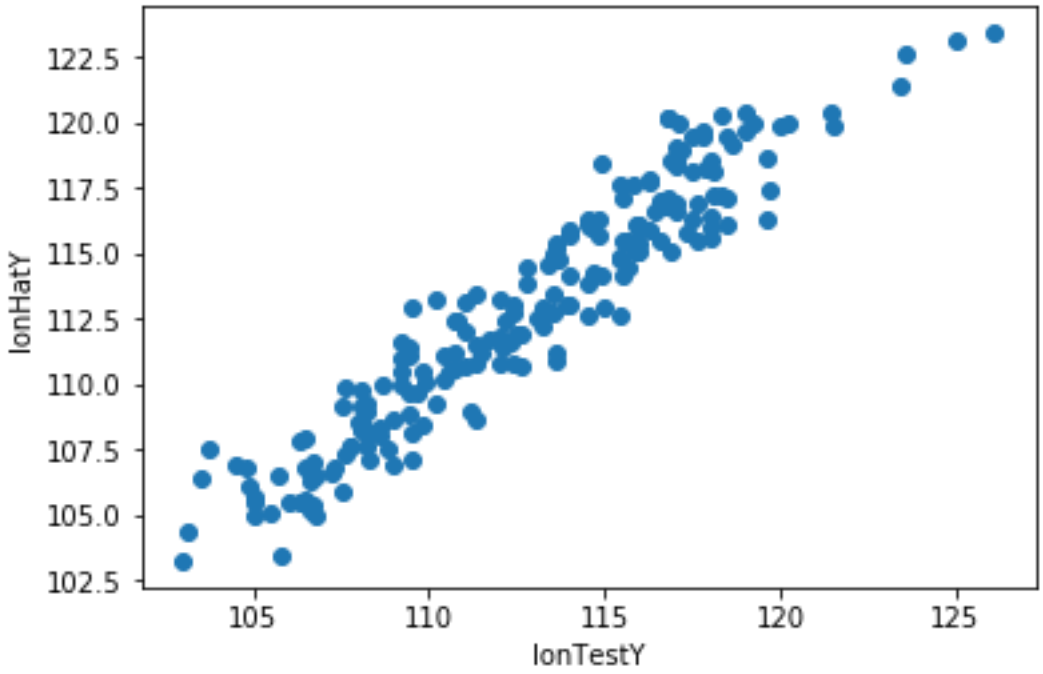


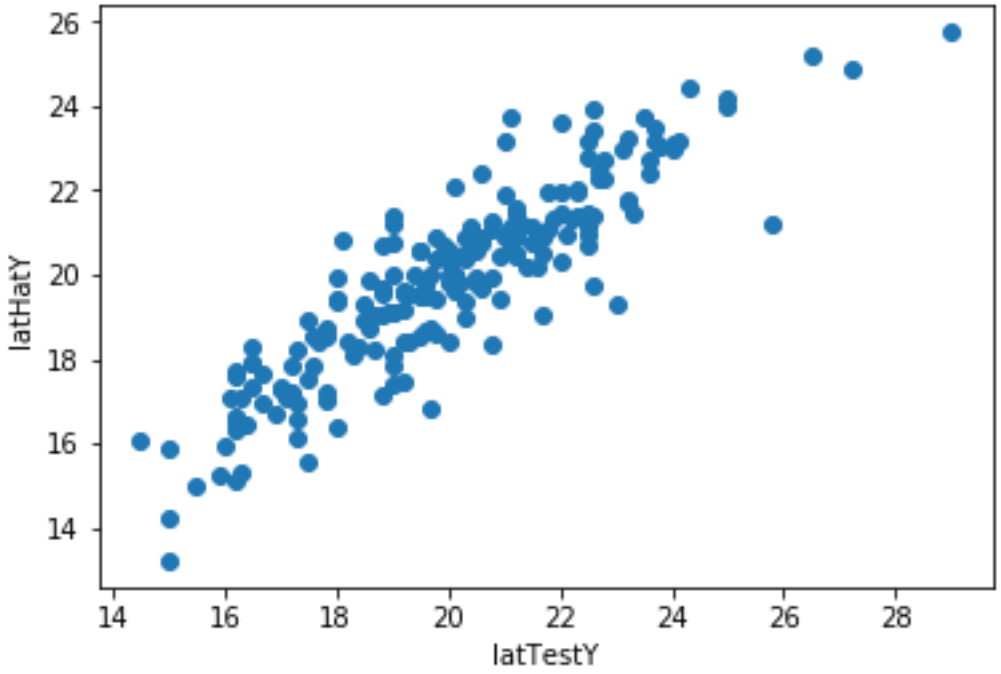
经度误差	3265.08550253
纬度误差	579.258691702
距离	364767.761797

通过反复多次试验，发现每次的实验误差都雷同，因此此处只写上一种情况。通过观察发现虽然真实值与预测值之间具有比较好的相关性，但是经度预测值与真实值的坐标数值相差很多；反映在误差上就是纬度误差十分的大，说明这个算法预测还是不够好。

二 Sklearn 的线性回归算法

实验结果（60%为训练集数据）：





经度误差	1.15470562835

纬度误差	0.862023508569
距离	158.508038823

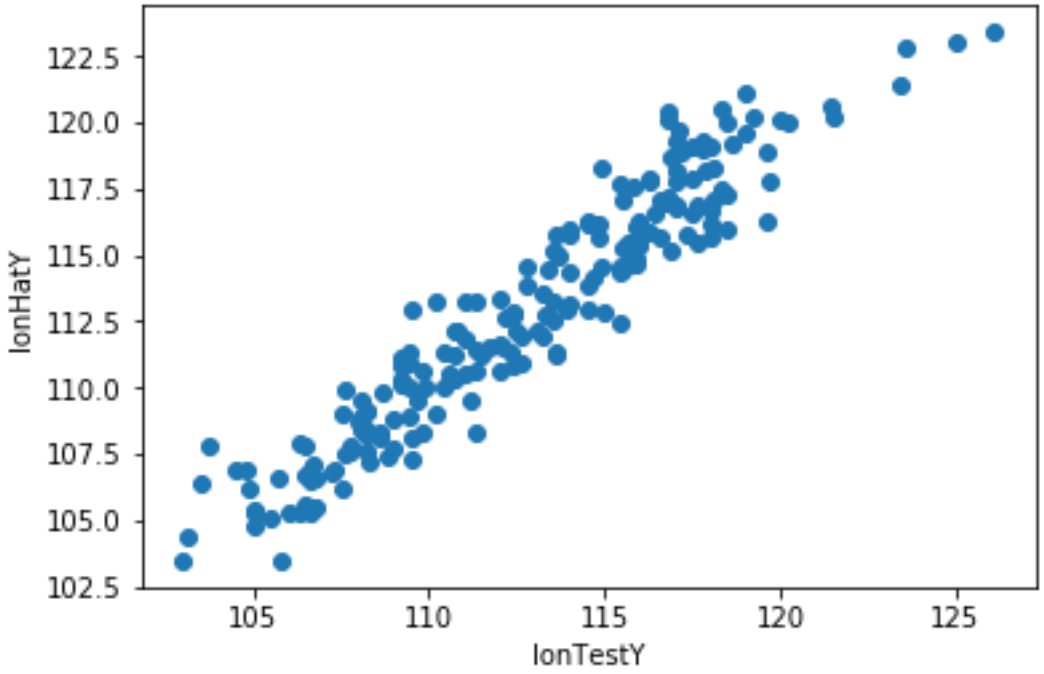
通过反复多次试验，发现每次的实验误差都雷同，因此此处只写上一种情况。

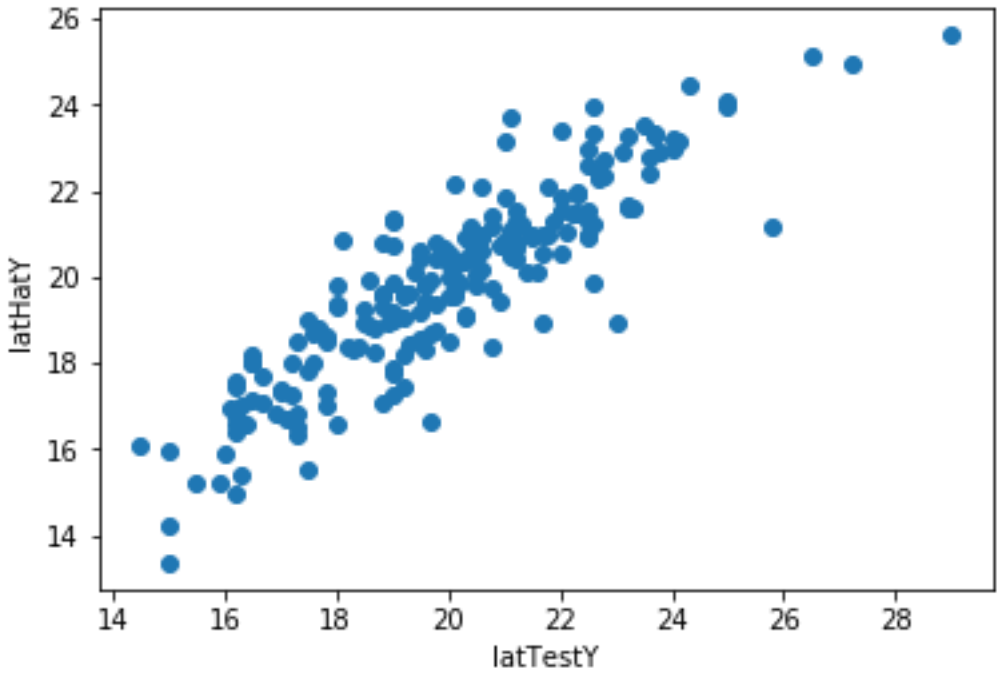
通过观察相关视图与误差大小可以发现，真实值与预测值之间具有较强的相关性并且两个误差都比较小，总的来说：拟合效果比较好。

在这里同时线性回归算法，我们可以看出 SKlearn 分析出来的结果误差比自己手动编写线性回归算法得出的结果误差要小得多。这或许在一定程度上说明了库函数的强大，毕竟库函数相比手写的函数而言，有非常多的玄学优化。

三 Sklearn 的岭回归算法

实验结果（60%为训练集数据）：





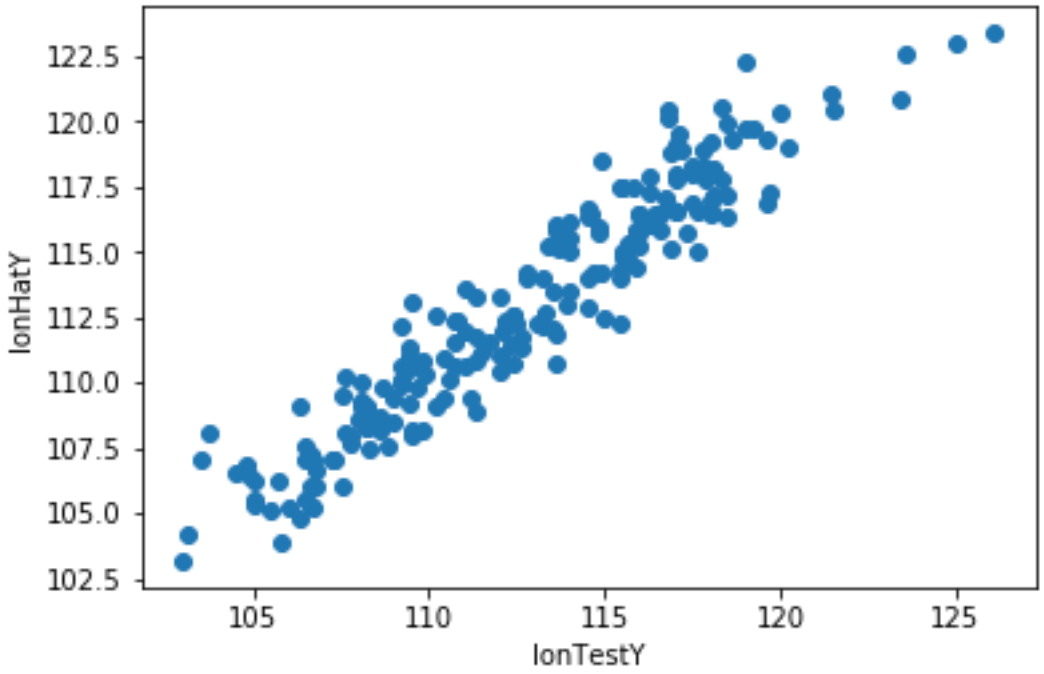
经度误差	1.14446149628
纬度误差	0.873174215789
距离	158.347486446

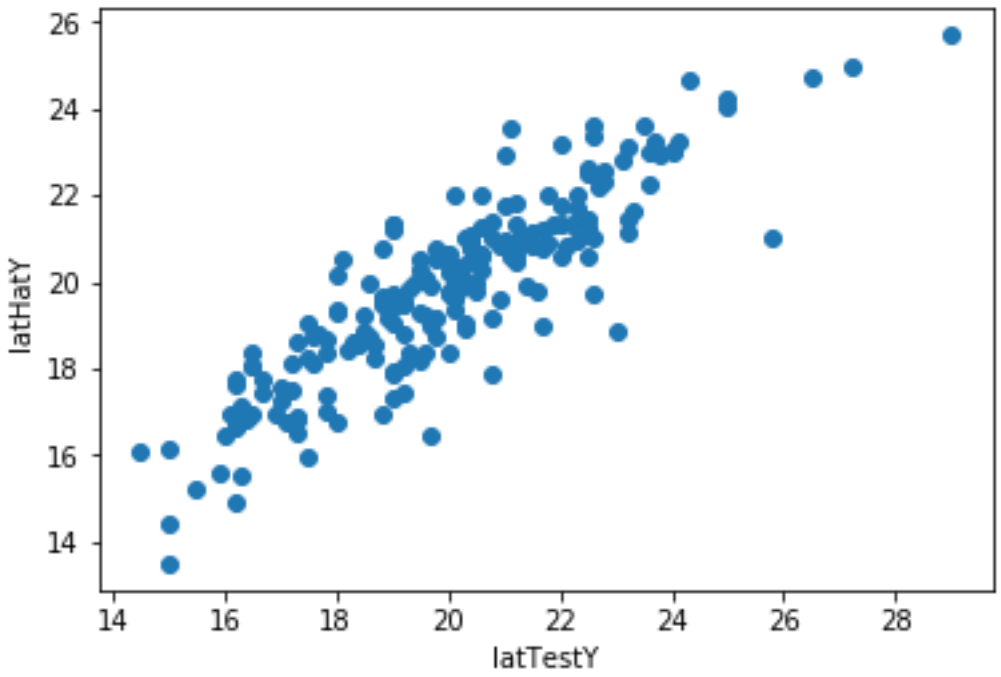
通过反复多次试验，发现每次的实验误差都雷同，因此此处只写上一种情况。

通过观察真实值与预测值相关分布以及比较实验结果误差，可以看出岭回归算法对实验的拟合效果很好。图示的相关性很好以及实验误差很小。而且算出的距离也比线性回归算法要小。总的来说，岭回归算法拟合效果很好。

四 Sklearn 的 Lasso 线性回归算法

实验结果如下（60%为训练集数据）：





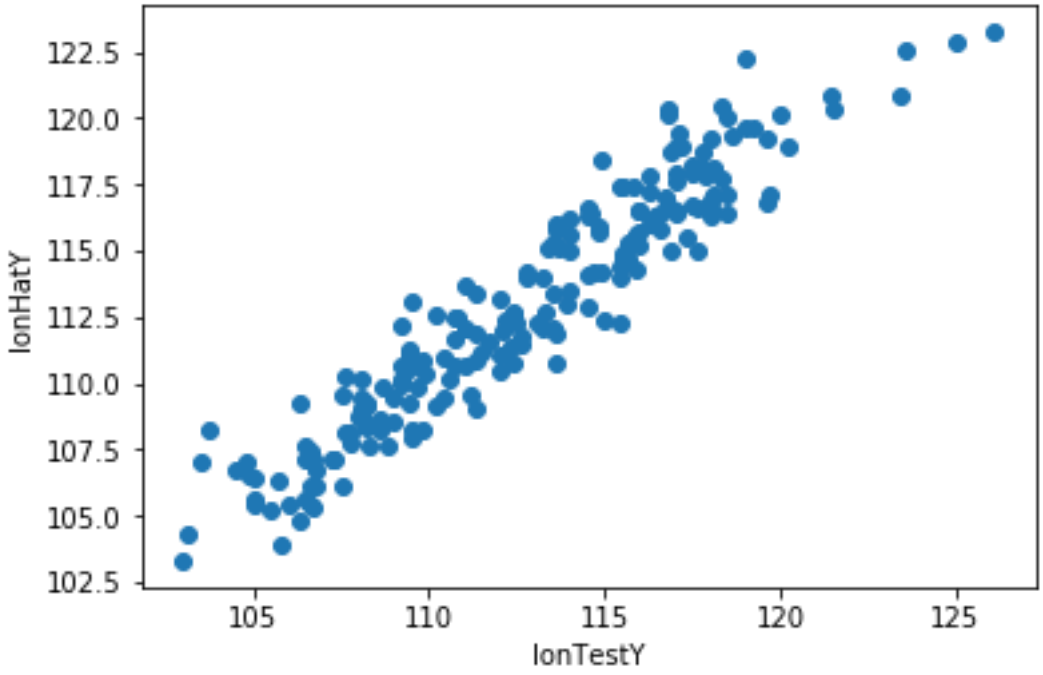
经度误差	1.17686223428
纬度误差	0.912718857663
距离	163.824696619

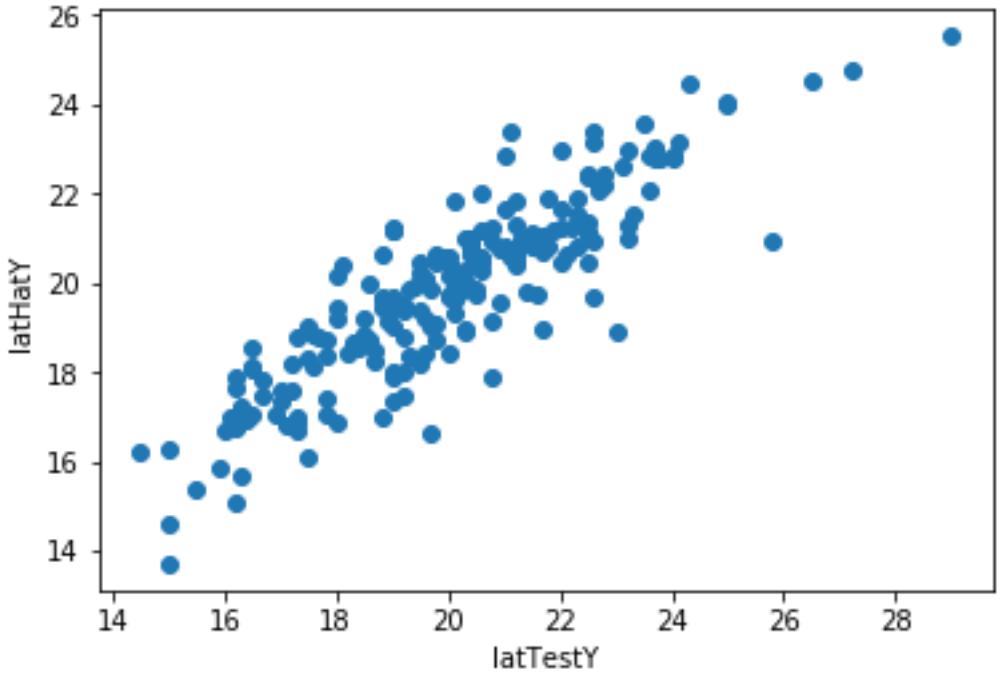
通过反复多次试验，发现每次的实验误差都雷同，因此此处只写上一种情况。

观察真实值与预测值的相关性，可以认为这两者的相关性较好。数据的误差也相对来说较小，但是和之前的算法对比，误差还是大了一点。或许在这个问题下，Lasso 算法并不是很适合这个问题。

五 Sklearn 的 ElasticNet 线性回归算法

实验结果如下（60%为训练集数据）



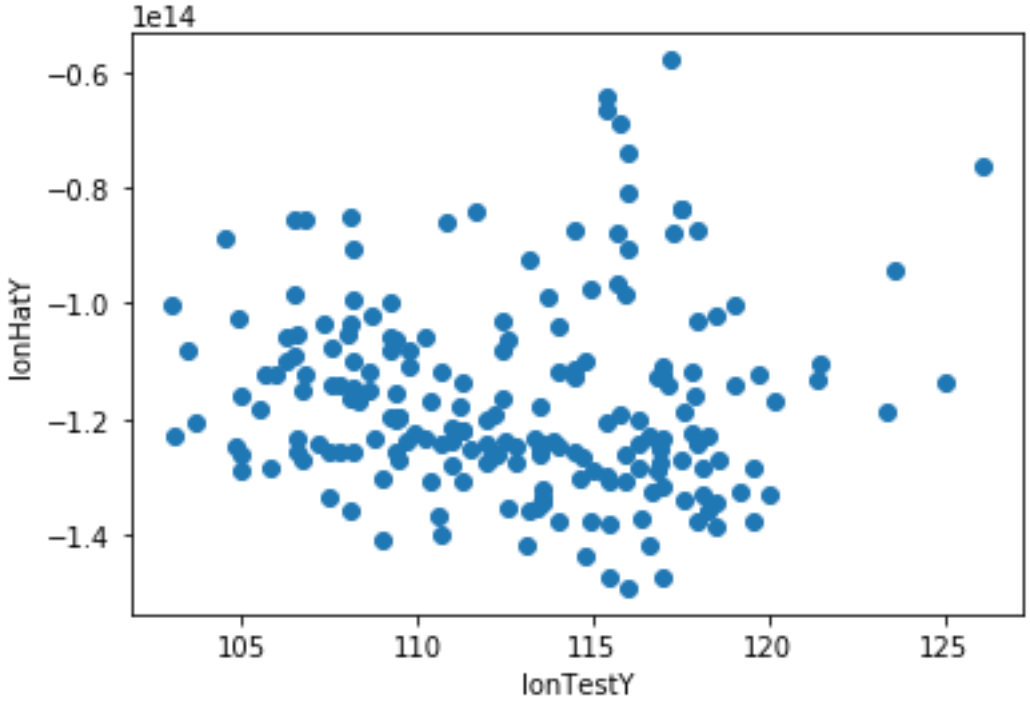


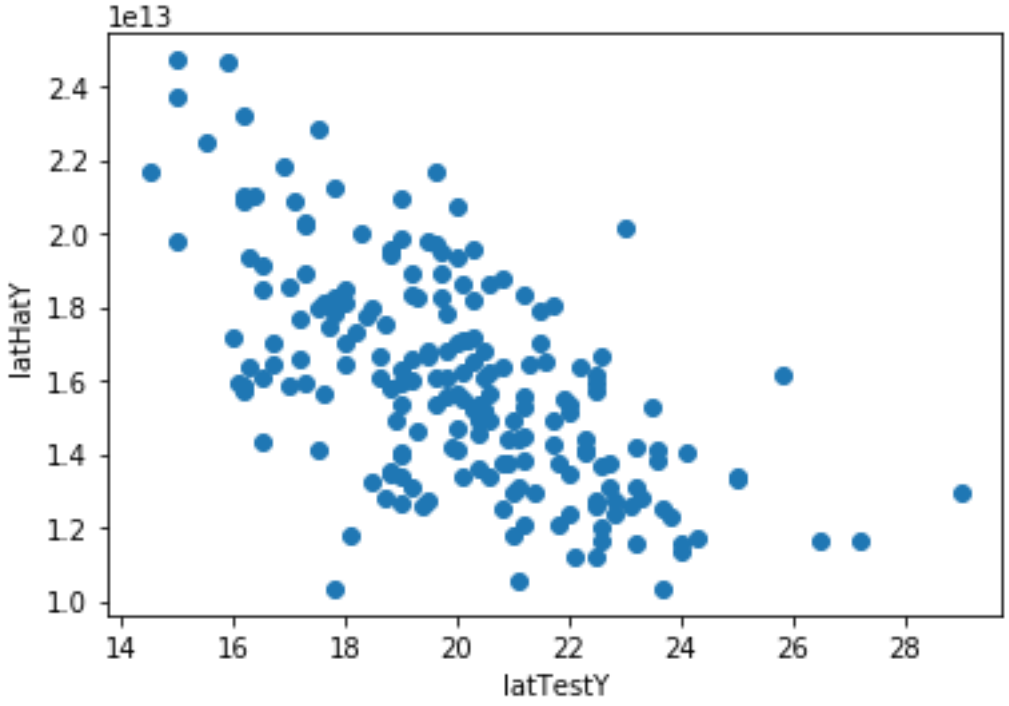
经度误差	1.15470562835
纬度误差	0.862023508569
距离	158.508038823

通过反复多次试验，发现每次的实验误差都雷同，因此此处只写上一种情况。

从真实值与预测值之间的可视化相关性来看，二者具有一定的相关性。但是经过比较实验结果误差，发现算法虽然效果较好，但是比不上岭回归算法以及线性回归算法。可能是这个问题下，ElasticNet 算法不够适用。

六 Sklearn 的 SGDRegressor 回归算法





通过画图发现相关性并不是很强，而且数据的误差也很大。这里只是略微提一提这个算法，并不做过多的分析。

以上拟合度比较好的算法分析比较：

	准回归算法	SKlearn 线性回归算法	Sklearn 岭回归算法	Sklearn Lasso 算法	SklearnElastic Net 算法
经度误差	3265.08550253	1.15470562835	1.14446149628	1.17686223428	1.17686223428
纬度误差	579.258691702	0.862023508569	0.873174215789	0.912718857663	0.912718857663
距离	364767.761797	158.508038823	158.347486446	163.824696619	163.824696619

通过表格进行纵向地比较：手写的准回归训练三个指标均是最差的，而SKlearn的岭回归算法则经度误差和距离都是最优秀的，SKlearn的线性回归算法对纬度的预测非常好。

综上所述：总而言之，SKlearn岭回归算法在本次台风预测实验问题中给出的结果最符合真实情况。比较符合这个问题的求解。

五、 问题讨论（实验过程中值得交待的事情）

本次实验代码的格式为ipynb，原因是各种结果可以中间直接打印效果直观，并且打开运行也更方便。

由于test集的特殊性，每一天有16个数据是自变量，而只有一个经度与一个纬度与之对应，规模差距太大，无法从16个数据中找出一个特别有代表性的数据与预测出来的经度纬度相对应。因此只是画出真实纬度经度与预测出来的经纬度进行对比，显示出相关性。

由于数据的特殊性，上面所有数据做过shuffle之后的情况都比不做shuffle处理的情况误差大很多。因此以上结果都是在没有做shuffle情况下得出的结果。

SGDRegressor算法得出的结果误差莫名的大，并且在实验例子中跑出来的范例结果误差也很大，所以在这一点上希望能得到助教的解答。

六、 结论

对于这个台风预测问题，SKlearn的岭回归算法给出的结果最接近真实值，因而比较符合这个问题。