实验报告

东北林业大学

信息与计算机科学技术实验中心

|  |
| --- |
| 一、实验目的  1）理解一元线性回归和多元线性回归的数学原理，能够利用sklearn中相关库解决现实世界中的各类回归问题；  （2）掌握利用matplotlib对一元线性回归模型进行可视化的方法，并分析模型的优劣；  （3）掌握利用statsmodels进行线性回归并进行模型评价的方法；  （4）掌握利用回归模型对现实问题进行分析和解释的能力。 |
| 二、实验环境  （1）硬件：PC机；  （2）软件：Anaconda Jupyter Notebook，Spyder，Pandas |
| 三、实验内容及结果  （一）基于伊春市木材剩余物数据利用回归模型预测木材剩余物  给出伊春林区16个林业局1999年木材剩余物和年木材采伐量数据见“木材剩余物.csv”。   1. 读取伊春市木材剩余物数据集，并显示输入。   2、利用分割X和Y数据集。  3、对数据进行可视化显示。  4、利用sklearn中的线性回归模型建立回归模型，对模型进行训练，输出模型参数。  5、假设乌伊岭林业局2000年计划采伐木材20万m3，求木材剩余物的点预测值。  通过计算，置信度为0.95的2000年平均木材剩余物E(*y*2000)的置信区间是  2000 ± *t*0.05(14) *s*(E(2000)) = 7.3231 ± 2.15 × 0.6742 = [5.8736 , 8.7726]  从而得出预测结果，2000年若采伐木材20万m3，产生木材剩余物的点估计值是7.3231万m3。平均木材剩余物产出量的置信区间估计是在 [5.8736, 8.7726] 万m3之间。从而为恰当安排2000年木材剩余物的加工生产提供依据。  6、利用statsmodels实现线性回归并对模型进行评估。  代码解读：  sm.add\_constant()函数用于增加截距项，也就是增加一个常数项。  sm.OLS()函数使用OLS（普通最小二乘法）建立线性回归模型est。  est.summary()用于输出模型评估结果。  7、将模型进行可视化显示。  8、预测并查看结果  9、利用残差qq图进行回归诊断  10、进一步绘制线性回归模型诊断  实验结论：  （1）查看模型结果时，红色警告指峰度检验需要样本数大于等于20。  （2）残差qq图存在多个弯曲，不满足残差正态性，暗示模型可能需要拟合多次项。  （3）模型R-squared为0.918，可以解释91.8%的信息。  （4）采伐量每增加一个单位，剩余物增加0.4个单位。  （5）问题：估计结果中截距项没有显著性，依据实际意义可知，没有木材采伐量就没有木材剩余物，所以理论上本案例中截距项是可以取零的。但是而有些问题就不可以。例如家庭消费和收入的关系。即使家庭收入为零，消费仍然非零。一般来说，截距项的估计量没有显著性时，也不做剔除处理。  本案例剔除截距项后的估计结果如下图所示，R2从0.918上升到0.982：  2000年若采伐木材20万m3，产生木材剩余物的点估计值是7.709万m3。  (6)改进思路，模型中增加采伐量的二次项或三次项。  从中可以看出，二次项的回归模型拟合效果已经很好，三次项的优化不明显。输出二次和三次项的模型评估结果，即可发现二次项模型的R2为0.931，校正R2结果为0.921，三次项模型中分别为0.932和0.916，校正R方系数出现了下降，因此，可能会出现过拟合现象。同时，二次项和三次项模型中显著性水平P值都出现了较大的提高，可能是因为所选择的指标中还缺少影响剩余物的其他决定因素，比如采伐区域的木材径级、利用率等数据，因此，综合评估后可以采用一元线性回归模型作为最终预测模型。  （二）利用广告投入与销售量数据集进行多元线性回归  1、读取广告投入与销售量数据集，并显示数据。  2、分割X和Y数据集。  3、查看数据集的特征  4、利用散点图查看各电视、收音机和报纸广告投入量和销售额的关系。  从中可以看出，电视广告投入量与销售额的相关性最好，但是并不是线性关系，报纸广告投入量相关性最差，收音机居中。  5、建立多元线性回归模型根据各平台广告投入量预测销售额。  6、进行回归诊断，查看残差QQ图和线性模型诊断图。  从模型评估结果中可以看出，R-squared为0.897，模型可以解释89.7%的信息。截距项、TV、radio在0.01水平显著，报纸广告投入量与销售额之间相关性不明显。 |

|  |
| --- |
| 四、实验过程分析与讨论  1． 实验过程分析与讨论：  （1）使用了matplotlib库进行数据可视化。使用scatter函数绘制了X和Y的散点图。使用了sklearn库中的LinearRegression类进行线性回归。在第15行，创建了一个LinearRegression对象regr，并使用fit函数拟合了X和Y的线性回归模型，输出了回归系数和截距。  （2）使用了statsmodels库进行统计分析，使用add\_constant函数为X添加常数列，并使用OLS函数拟合了Y和X2的线性回归模型，输出了回归模型的统计摘要，使用scatter函数绘制了X和Y的散点图，并使用plot函数绘制了回归线。  （3）数据预测和误差计算使用predict函数预测了X为20时的Y值QQ图绘制使用sm.qqplot函数绘制了残差的QQ图。代码使用了sklearn库中的PolynomialFeatures类进行多项式回归。创建了一个二次项模型poly和三次项模型poly3，并使用fit\_transform函数将X转换为多项式特征。使用OLS函数拟合了一元二次和一元三次多项式回归模型，并预测了Y的值。使用scatter函数绘制了X和Y的散点图，并使用plot函数绘制了线性回归、二次项和三次项的回归线。  （4）二次和三次项的模型评估结果，校正R方系数出现了下降，因此，可能会出现过拟合现象。同时，二次项和三次项模型中显著性水平P值当出现了较大的提高，可能是因为所选择的指标中还缺少影响剩余物的其他决定因素，比如采伐区域的木材径级、利用率等数据，因此，综合评估后可以采用一元线性回归模型作为最终预测模型。 |
| 五、指导教师意见    指导教师签字：  年 月 日 |