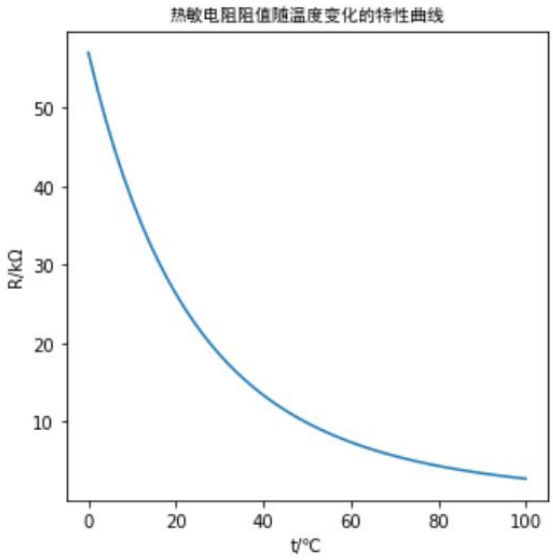
Curve fitting: perspective from machine learning

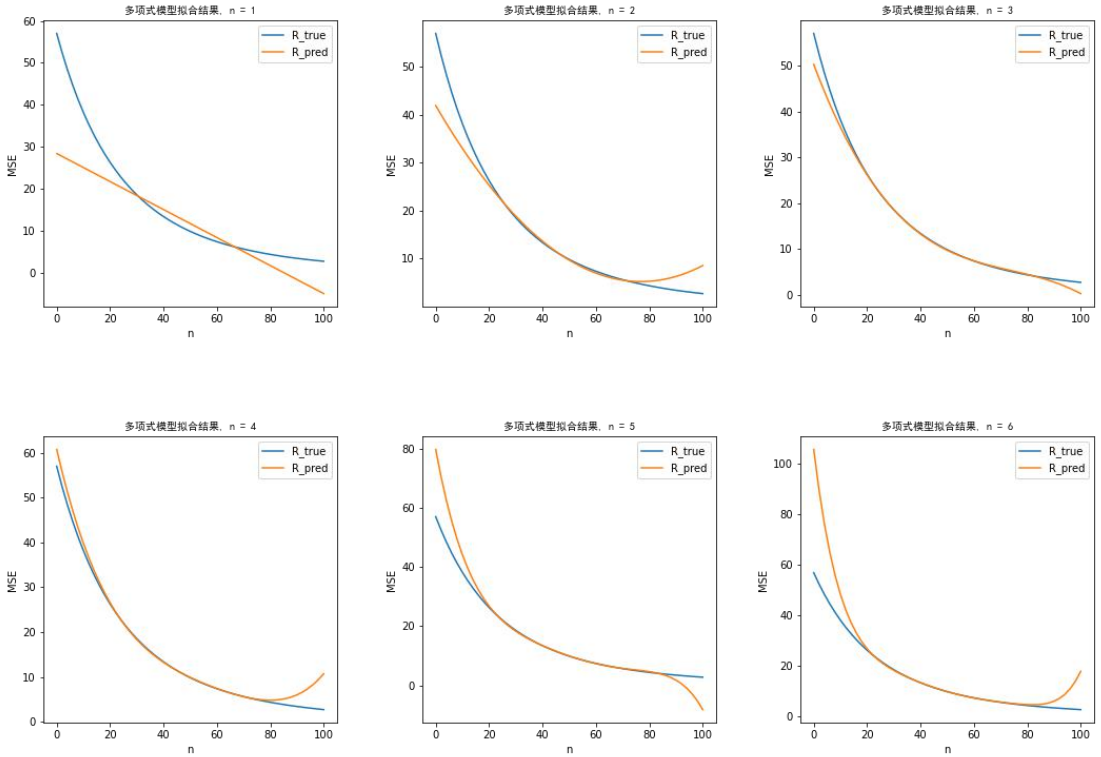
杨小诺 2018011495 自 83

问题（1）：画出该热敏电阻阻值随温度变化的特性曲线。

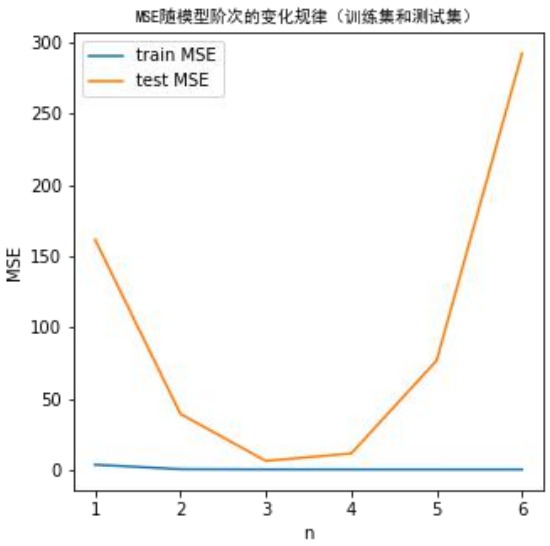


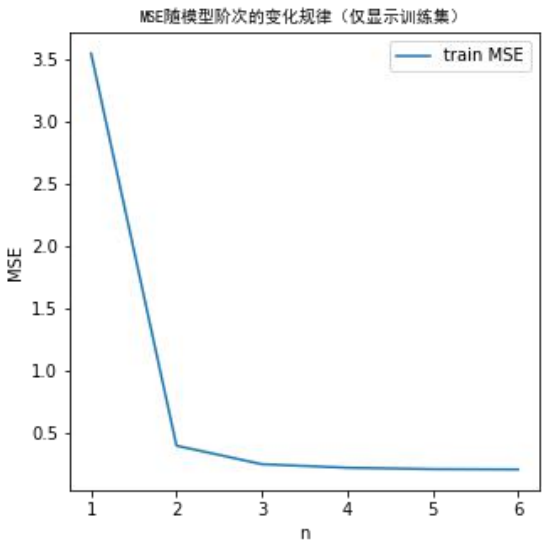
问题（2）：在（1）中获得的 20°C~80°C 范围的数据上添加噪声（零均值、标准偏差取 500Ω的高斯噪声）作为训练集；采用曲线拟合最小二乘法分别获得模型阶次 $n=1,2,3,4,5,6$ 时对应的多项式模型；分别计算训练集上和测试集上的均方误差，观察训练集和测试集上误差随模型阶次的变化规律并加以讨论。

加入噪声后进行曲线拟合实验，不同阶次时的拟合结果见下方（从左到右、从上到下依次为 n = 1，n = 2 … n = 6）。



不同阶次时的 MSE 见下方（左图为训练集和测试集上 MSE，右图为左图蓝色曲线放大结果）。





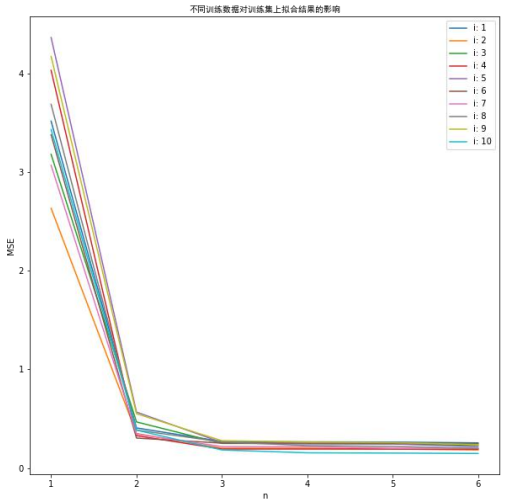
综合以上实验结果可知：

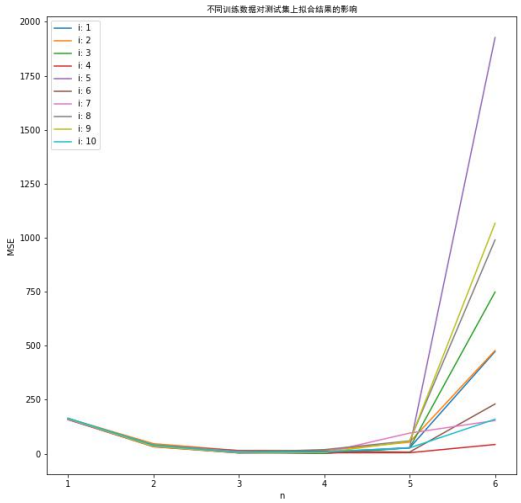
随着模型阶次 n 的增大，训练集上 MSE 逐渐减小，测试集上 MSE 先减小后增大。这是因为 n 较小时，模型过于简单，无法准确描述数据集上的特征，n 太大时，又出现了过拟合的情况，模型泛化能力不足。

从预测值曲线和实际值曲线上也能看出上面的结论，即 n 的取值要适中，不能过大或过小。

问题（3）：将（2）的内容重复十次，观察并讨论采用不同训练数据给拟合结果带来的影响。

重复实验十次，将这十次实验的结果画在同一张图上，见下方（左图为训练集上 MSE，右图为测试集上 MSE）。



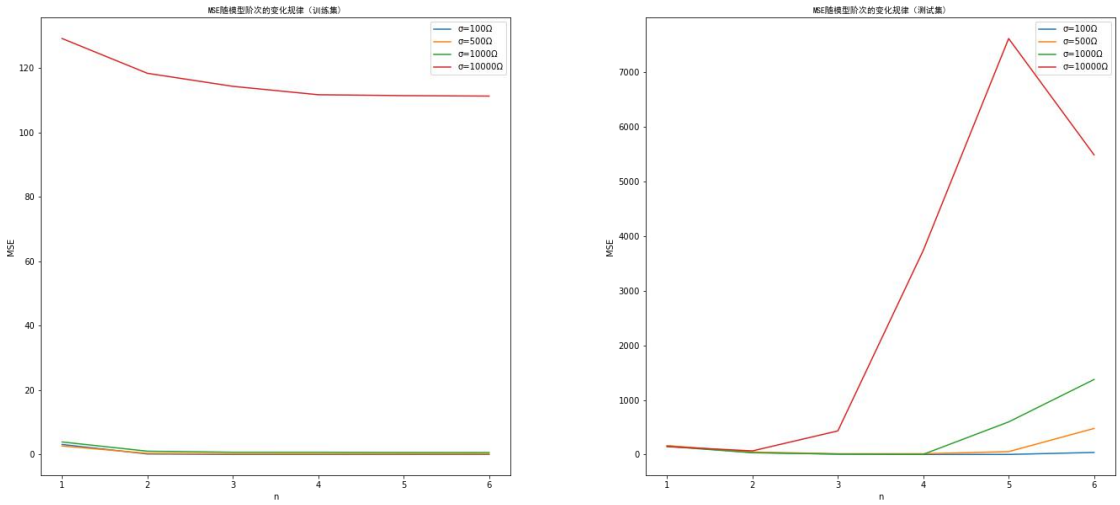


综合以上实验结果可知：

左图中不同曲线见的差距略小，而右图更大，也就是说，噪声强度对测试集拟合结果的影响大于训练集。随着模型阶次的增大，测试集上结果波动越来越大，这也说明模型越复杂，噪声带来的影响越大。

问题（4）：改变噪声强度（通过改变噪声标准差来实现），重复（2），（3），观察并讨论数据中不同噪声强度给拟合带来的影响。

改变噪声强度，重复（2），结果见下方，其中左侧为训练集上 MSE，右侧为测试集上 MSE，横坐标代表不同阶次、曲线颜色代表不同噪声强度。

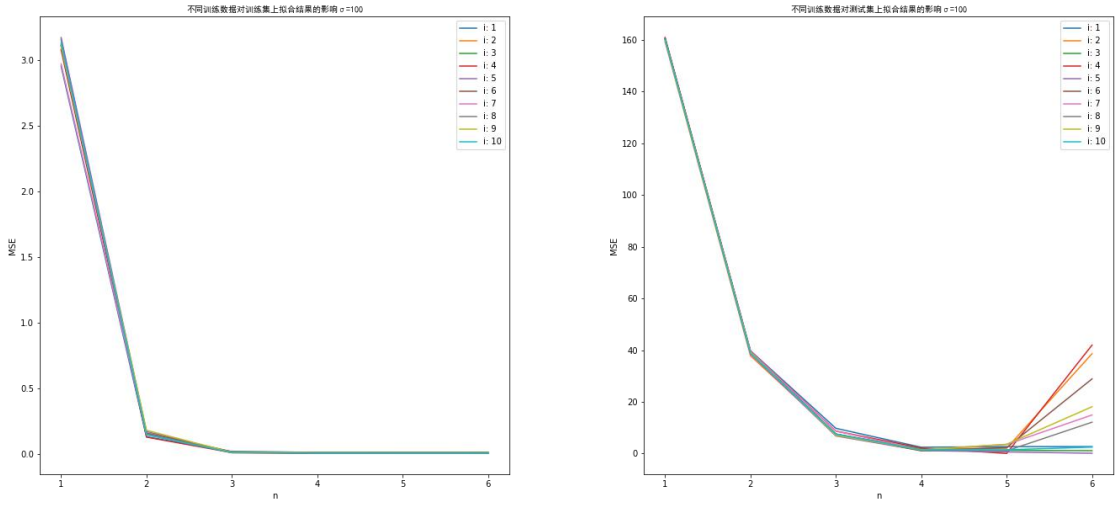


综合以上实验结果可知：

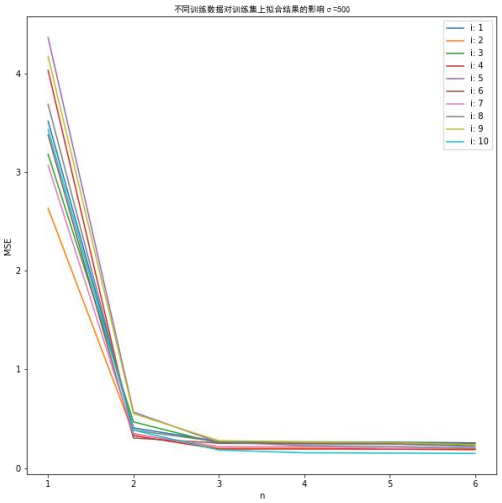
随着噪声标准差的变大，测试集和训练集的 MSE 均变大。但任一标准差下，测试集和训练集上 MSE 的变化趋势与第二问的结论都是类似的，即训练集上 MSE 随 n 增大而减小，测试集上 MSE 只有 n 大小适中使才比较低。

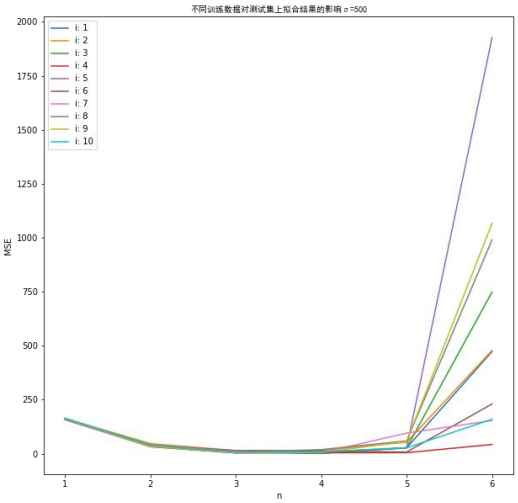
改变噪声强度，重复（3），结果见下方，左右两图为一组，其中左侧为训练集上 MSE，右侧为测试集上 MSE，横坐标代表不同阶次、曲线颜色代表不同初始数据。

σ =100 时

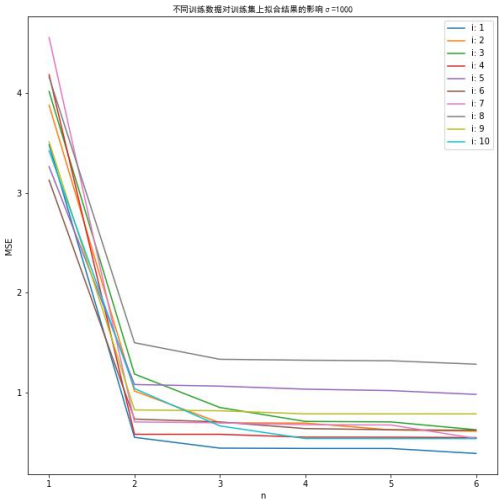


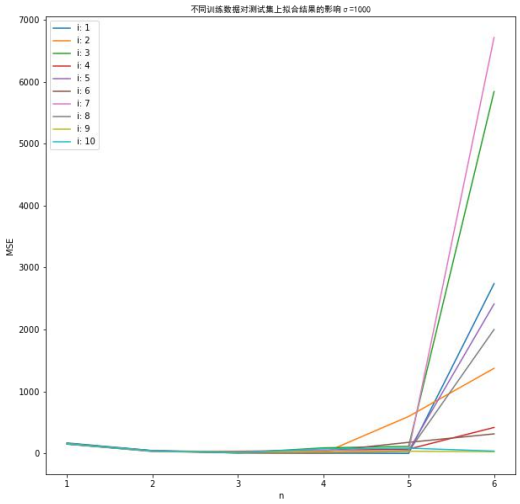
σ =500 时



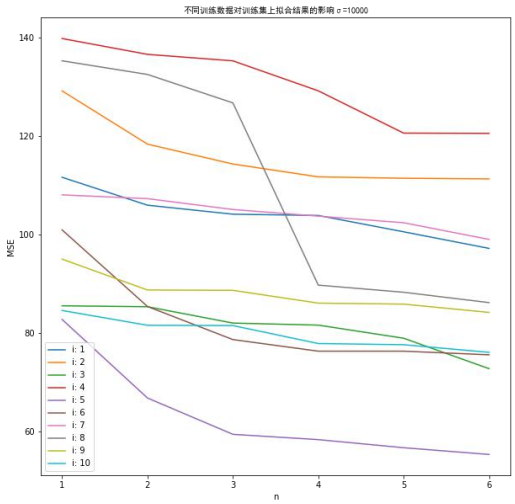


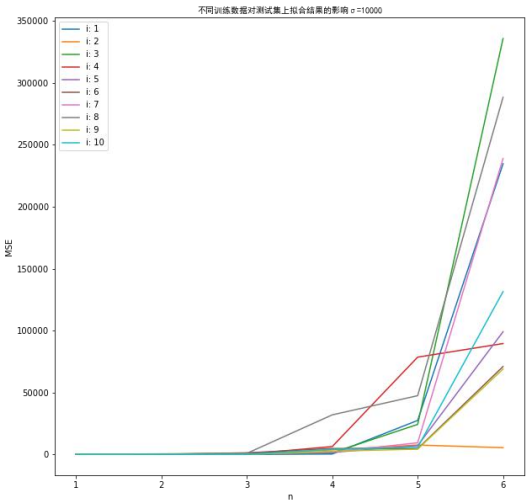
σ =1000 时





σ =10000 时



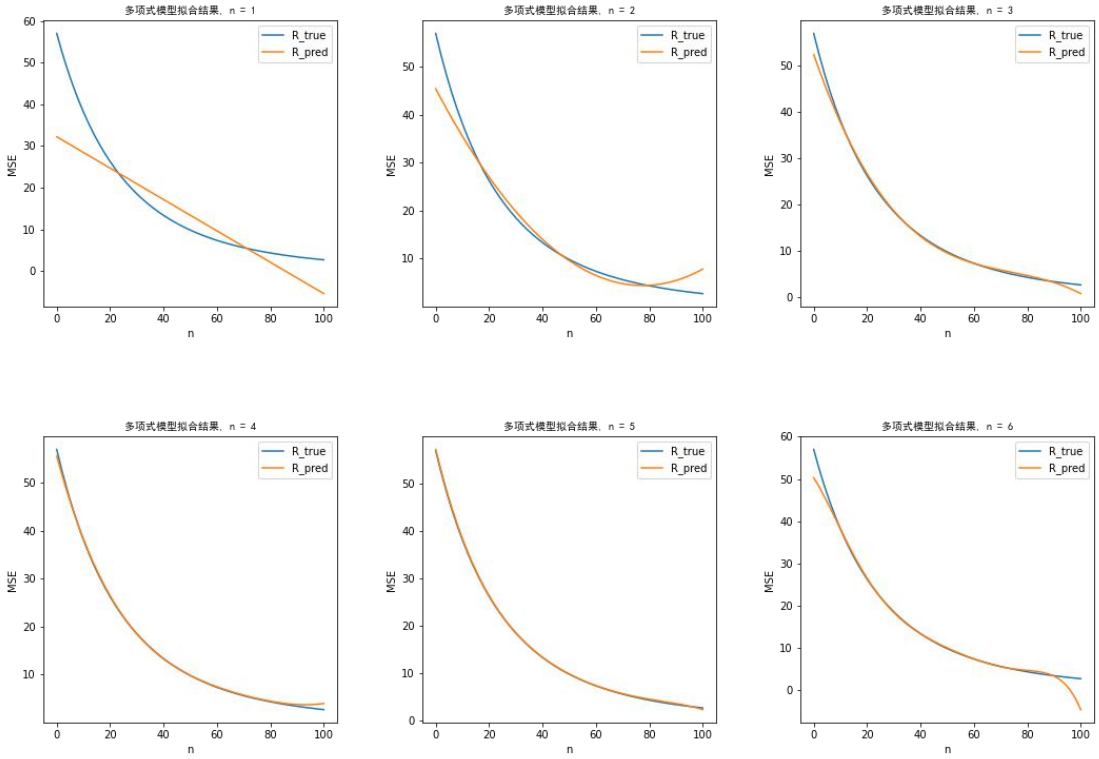


综合以上实验结果可知：

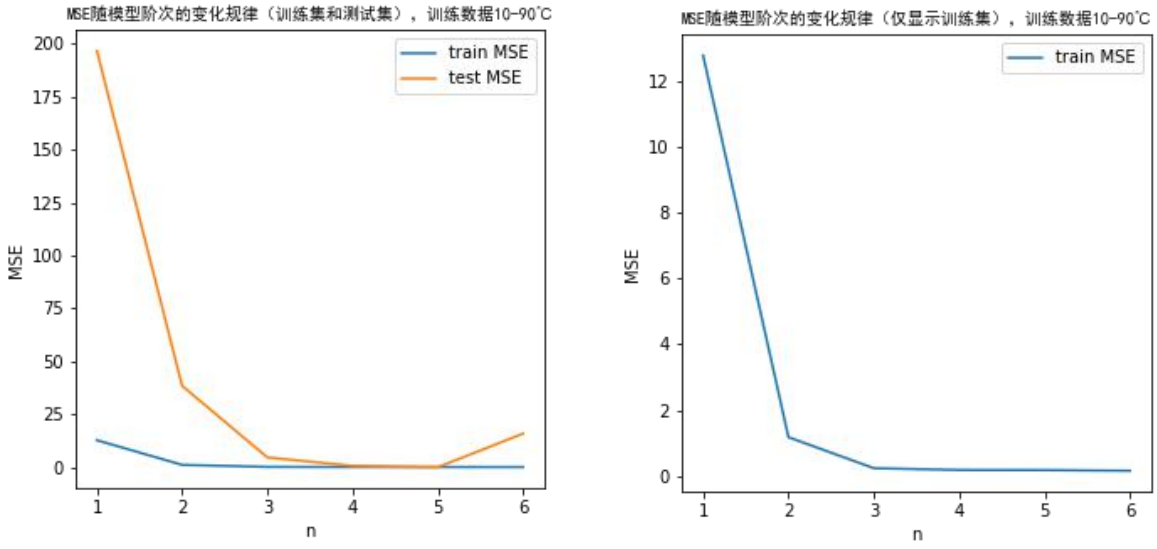
随着噪声标准差的增大，噪声对测试集和训练集的影响都在变大。同时，比较不同测试集上曲线的最低点可以知道，噪声强度不同，最优的拟合模型阶次也不同，也就是说噪声会影响模型最优参数的选择。

问题（5）：将训练集范围进行调整（扩大或缩小），重复（2），（3）内容，观察并讨论由于采用不同规模训练数据给拟合（学习）结果带来的影响。

将训练集范围调整为 10℃到 90℃（扩大），重复（2），不同阶次时的拟合结果见下方，（从左到右、从上到下依次为 n＝1，n＝2 … n＝6）。

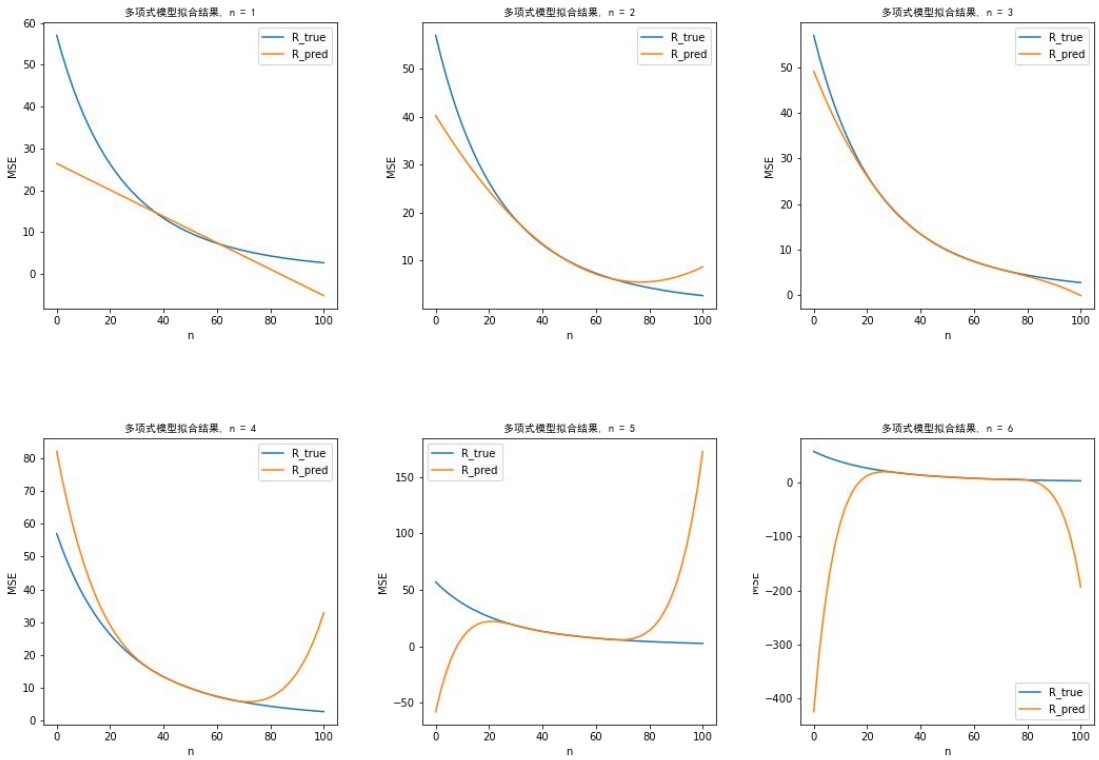


不同阶次时的 MSE 见下方，左图为训练集和测试集上 MSE，右图为左图蓝色曲线放大结果。

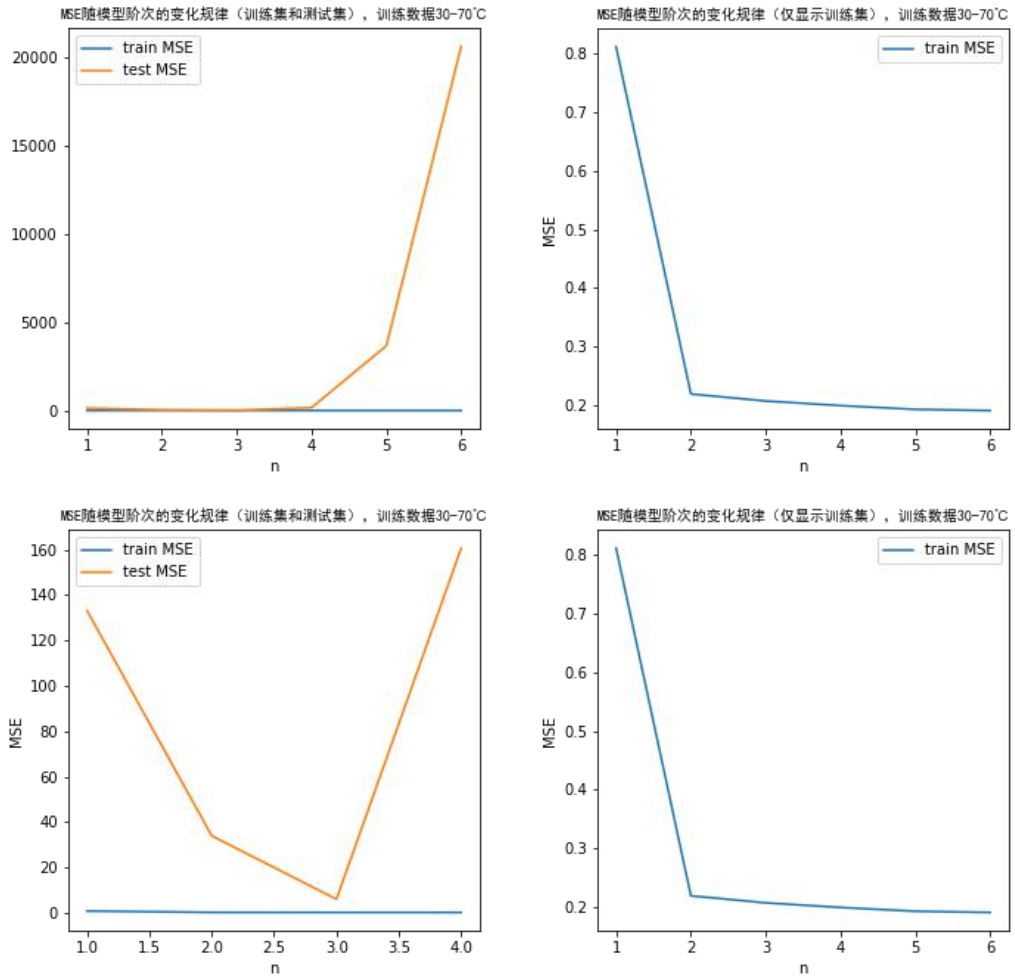


将训练集范围调整为 30℃到 70℃（缩小），重复（2），不同阶次时的拟合结果见下方，（从左到

右、从上到下依次为 n = 1，n = 2 … n = 6）。



不同阶次时的 MSE 见下方，上方两图中，左图为训练集和测试集上 MSE，右图为左图蓝色曲线放大结果，下方两图为上方两图 n=1\2\3 时的放大。

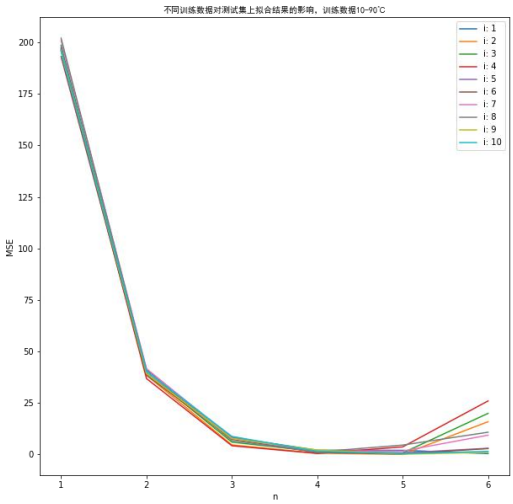


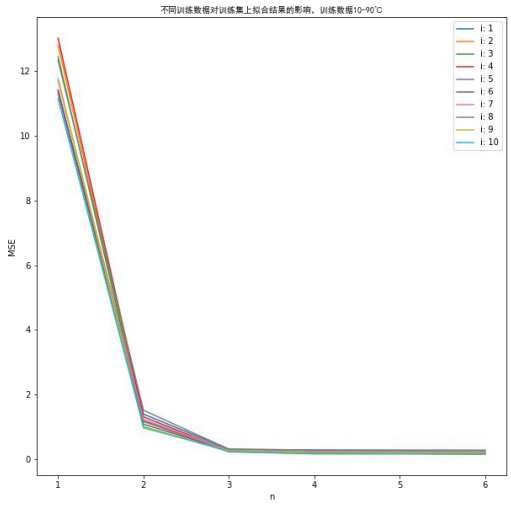
综合以上实验结果可知：

改变训练集占总数据集的比例时，MSE 曲线整体走势与第（2）问结果类似，都是与 n 相关联的、下凹型曲线；并且随着训练集占总数据集的比例增大，测试集上 MSE 的曲线呈现右移的趋势，也就是最低 MSE 对应的 n 值逐渐增大。

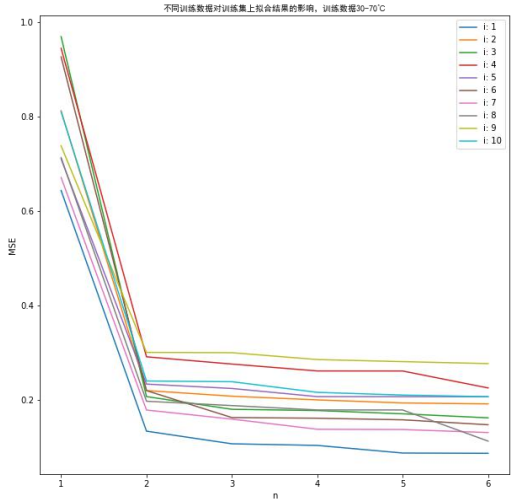
此外，随着训练集范围扩大，也就是随着用于训练的数据逐渐增多，不同阶次模型的整体效果都有提升，n 较大时这一点尤为明显。总的来说：更多的训练数据有利于提高模型拟合效果，特别地，模型较复杂时，更多的训练数据还可以防止过拟合。

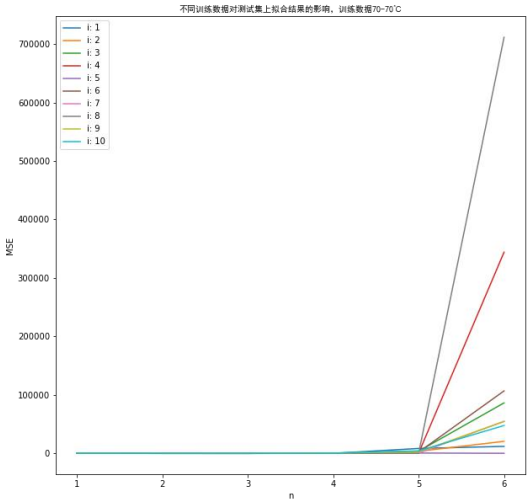
将训练集范围调整为 10℃到 90℃（扩大），重复（3），实验结果见下方（左图为训练集上 MSE，右图为测试集上 MSE，不同颜色的曲线表示不同初始数据得到的结果）。





将训练集范围调整为 30℃到 70℃（减小），重复（3），实验结果见下方（左图为训练集上 MSE，右图为测试集上 MSE，不同颜色的曲线表示不同初始数据得到的结果）。





综合以上实验结果可知：

随着训练集占总数据集的比例增大，模型的抗干扰能力增强，不同初始数据下结果的波动变小。

问题（7）：如实验前已事先了解热敏电阻测温机理并掌握其阻值与温度的关系符合（1）式所描述的模型，你将如何考虑从实验数据获得热敏电阻的阻值与温度关系模型？

由问题（1）：

$$R_T = R_{T_0} e^{\beta(\frac{1}{T}-\frac{1}{T_0})} = R_{T_0} e^{\beta(\frac{1}{t+273.15}-\frac{1}{t_0+273.15})}$$

取 ln，得：

$$\ln R_T = \beta\frac{1}{t+273.15} + \ln R_{T_0} - \beta\frac{1}{t_0+273.15}$$

令 $x = \frac{1}{t+273.15}$，$y = R_T$，则有：

$$y = \beta x + \ln R_{T_0} - \beta\frac{1}{t_0+273.15}$$

由最小二乘法，用直线拟合x，y即可。

画出 x-y 关系图如下方左图，拟合结果如下方右图。