《机器学习》课程第 5 次作业

姓名：刘哲　　学号：2022103691

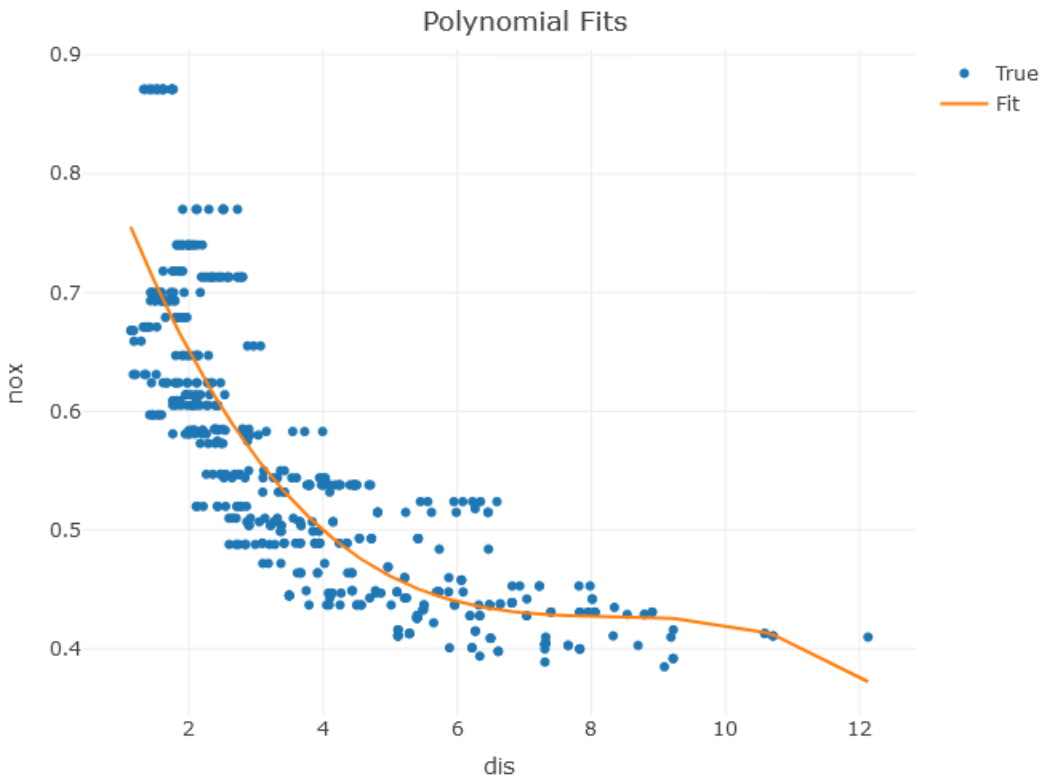
1　第9题

a

三次多项式回归的结果为

```
##
## Call:
## lm(formula = nox ~ poly(dis, degree = 3), data = boston.subset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.121130 -0.040619 -0.009738  0.023385  0.194904
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.554695   0.002759 201.021  < 2e-16 ***
## poly(dis, degree = 3)1 -2.003096   0.062071 -32.271  < 2e-16 ***
## poly(dis, degree = 3)2  0.856330   0.062071  13.796  < 2e-16 ***
## poly(dis, degree = 3)3 -0.318049   0.062071  -5.124 4.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06207 on 502 degrees of freedom
## Multiple R-squared:  0.7148, Adjusted R-squared:  0.7131
## F-statistic: 419.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

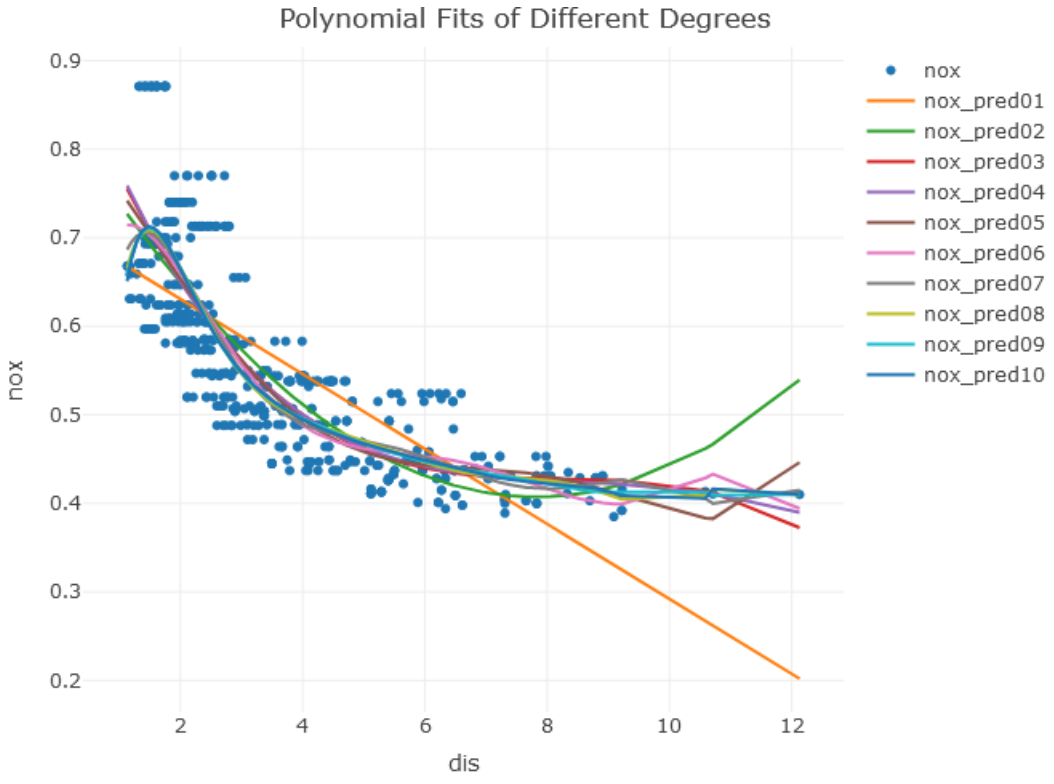
可以看出，回归系数均显著，模型显著。三次多项式回归的拟合曲线为



拟合曲线在样本数据的中间位置表现尚可，但在两端表现很差，拟合曲线出现了对样本数据明显的偏离。

b

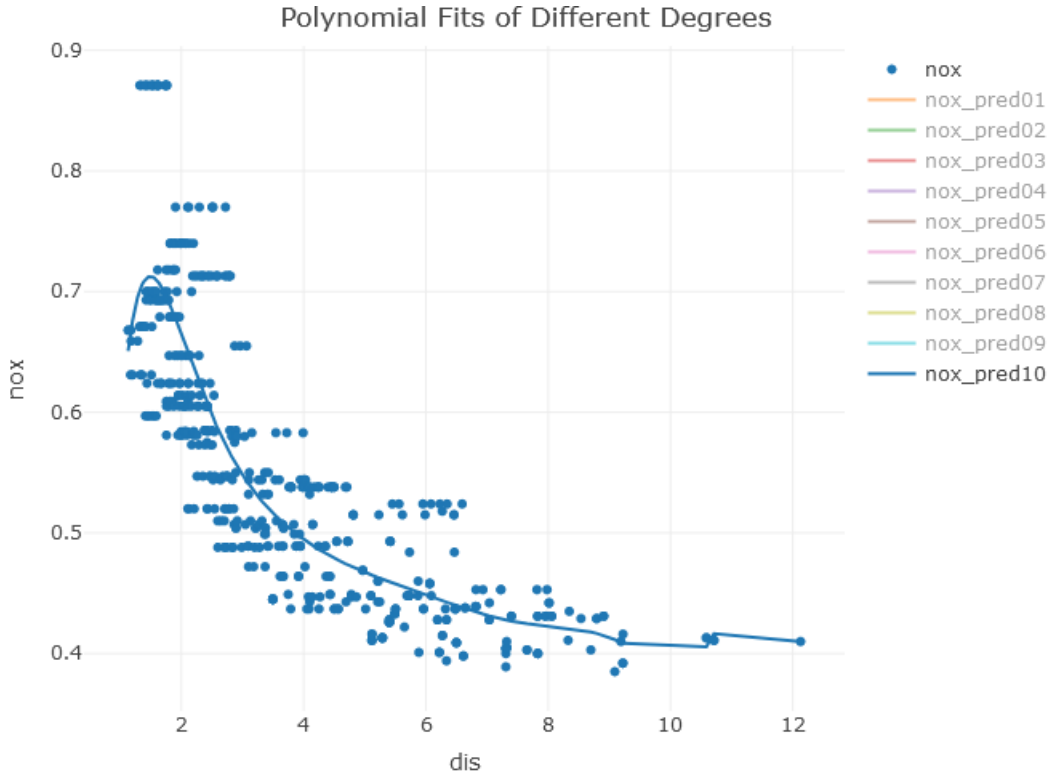
选取多项式回归的 $degree$ 为 $1 \sim 10$，分别进行多项式回归，拟合曲线分别为



计算各个回归模型的 RSS 为

Degree	1	2	3	4	5
RSS	2.7686	2.0353	1.9341	1.9330	1.9153
Degree	6	7	8	9	10
RSS	1.8783	1.8495	1.8356	1.8333	**1.8322**

根据结果可知，当 $degree = 10$ 时，多项式回归模型具有最小的 RSS，但此时可能存在比较严重的过拟合问题。拟合曲线为



c

为了降低发生过拟合的可能性，使用十折交叉验证的方法训练每个 $degree$ 下的多项式回归模型。将样本数据随机平均分为 10 组，在每个 $degree$ 下，每次选择一组数据作为验证集，其余九组作为训练集训练模型。将十折验证集的 RSS 加和作为该 $degree$ 的 RSS，结果为

Degree	1	2	3	4	5
RSS	2.7889	2.0600	**1.9594**	1.9725	2.1317
Degree	6	7	8	9	10
RSS	2.6212	5.2695	2.8932	3.8634	2.7334

根据结果可知，当 $degree = 10$ 时，多项式回归模型具有最小的 RSS，认为最优 $degree$ 为 3。拟合曲线见**a**。

由于使用交叉验证法降低了发生过拟合的可能性，该结果较为可信。

d

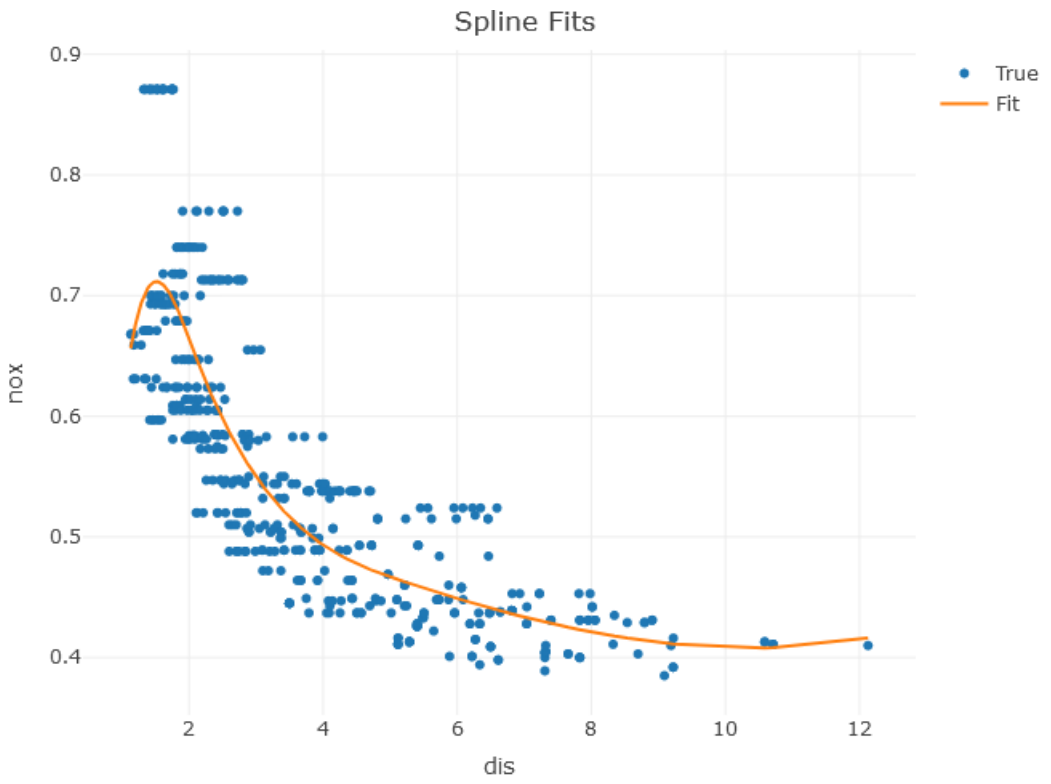
令自由度为 4，设 $knots$ 为变量 dis 的分位数

Quantile	25%	50%	75%
Value	2.1002	3.2075	5.1884

拟合回归样条的结果为

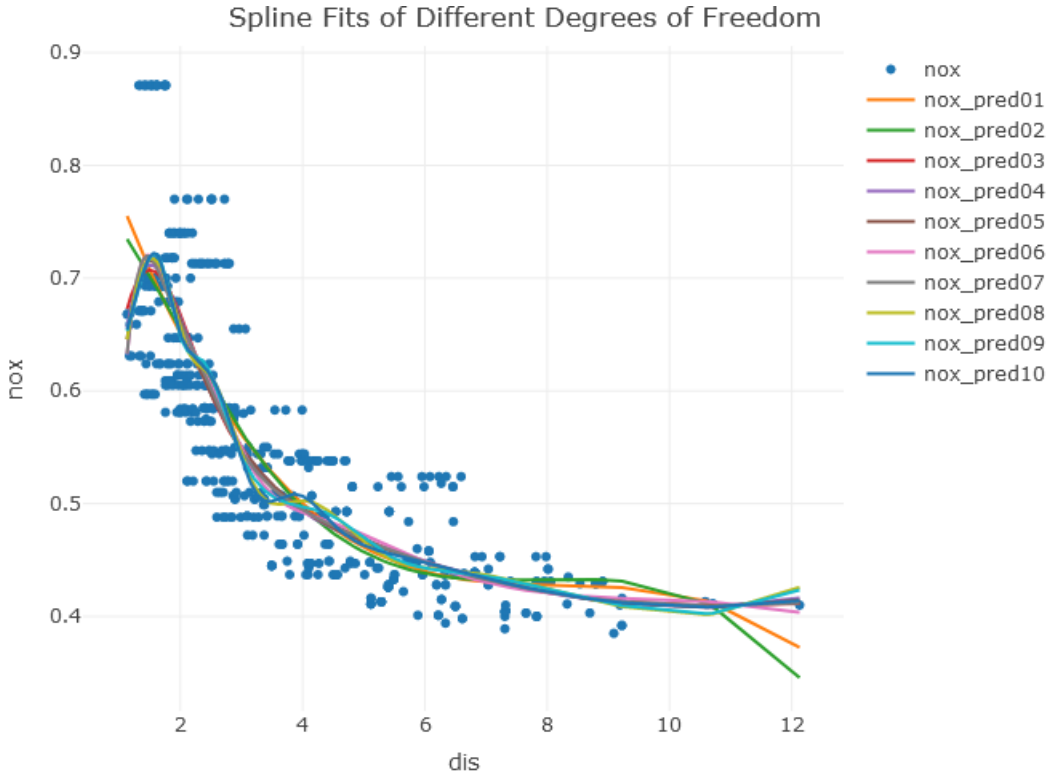
```
##
## Call:
## lm(formula = nox ~ bs(dis, df = 4, knots = boston.sp.knots, degree = 3),
##     data = boston.subset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.128538 -0.037813 -0.009987  0.022644  0.195494
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                       0.65622    0.02370
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)1  0.10222    0.03516
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)2 -0.02963    0.02338
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)3 -0.15959    0.02791
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)4 -0.22815    0.03324
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)5 -0.26272    0.04930
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)6 -0.24002    0.05434
##                                                  t value Pr(>|t|)
## (Intercept)                                       27.689  < 2e-16 ***
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)1   2.907  0.00381 **
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)2  -1.267  0.20571
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)3  -5.718 1.86e-08 ***
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)4  -6.864 1.99e-11 ***
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)5  -5.329 1.50e-07 ***
## bs(dis, df = 4, knots = boston.sp.knots, degree = 3)6  -4.417 1.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06062 on 499 degrees of freedom
## Multiple R-squared:  0.7295, Adjusted R-squared:  0.7263
## F-statistic: 224.3 on 6 and 499 DF,  p-value: < 2.2e-16
```

可以看出，不是所有回归系数都显著，但模型显著。回归样条的拟合曲线为



e

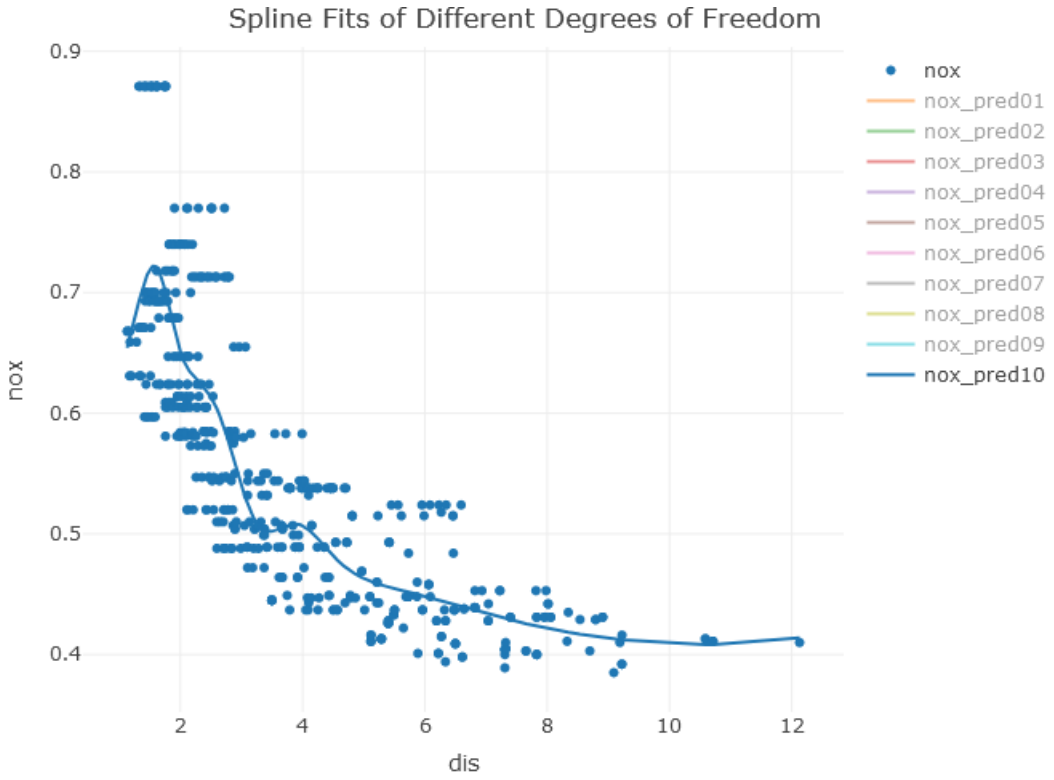
选取回归样条的自由度为 $1 \sim 10$，分别进行回归样条，拟合曲线分别为



计算各个回归模型的 RSS 为

DF	1	2	3	4	5
RSS	1.9341	1.9228	1.8402	1.8340	1.8299
DF	6	7	8	9	10
RSS	1.8170	1.8257	1.7925	1.7970	**1.7890**

根据结果可知，当 $DF = 10$ 时，回归样条模型具有最小的 RSS，但此时可能存在过拟合问题。拟合曲线为

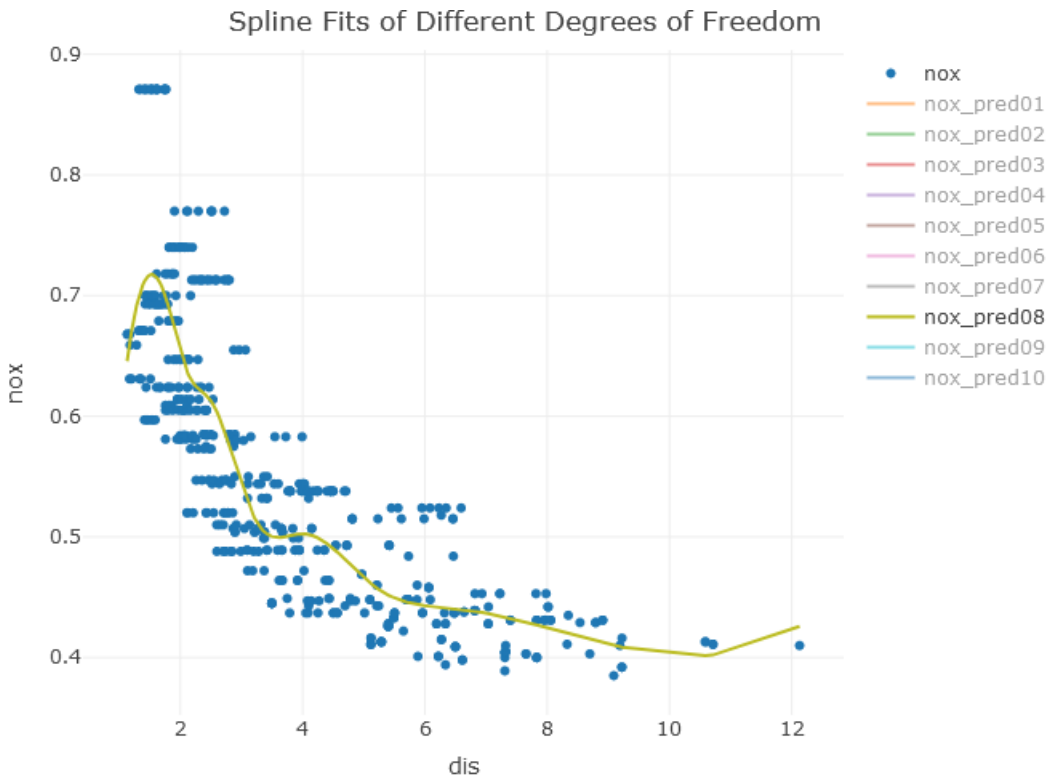


f

为了降低发生过拟合的可能性，使用十折交叉验证的方法训练每个 DF 下的回归样条模型。将样本数据随机平均分为 10 组，在每个 DF 下，每次选择一组数据作为验证集，其余九组作为训练集训练模型。将十折验证集的 RSS 加和作为该 DF 的 RSS，结果为

DF	1	2	3	4	5
RSS	1.9594	1.9747	1.8741	1.8754	1.8743
DF	6	7	8	9	10
RSS	1.8646	1.8765	**1.8633**	1.8744	1.8651

根据结果可知，当 $DF = 8$ 时，回归样条模型具有最小的 RSS，认为最优 DF 为 8。拟合曲线为



由于使用交叉验证法降低了发生过拟合的可能性，该结果较为可信。