## 練習問題 5-1

中古車価格データの全てをプロットすると，以下の図 1 のようになる。

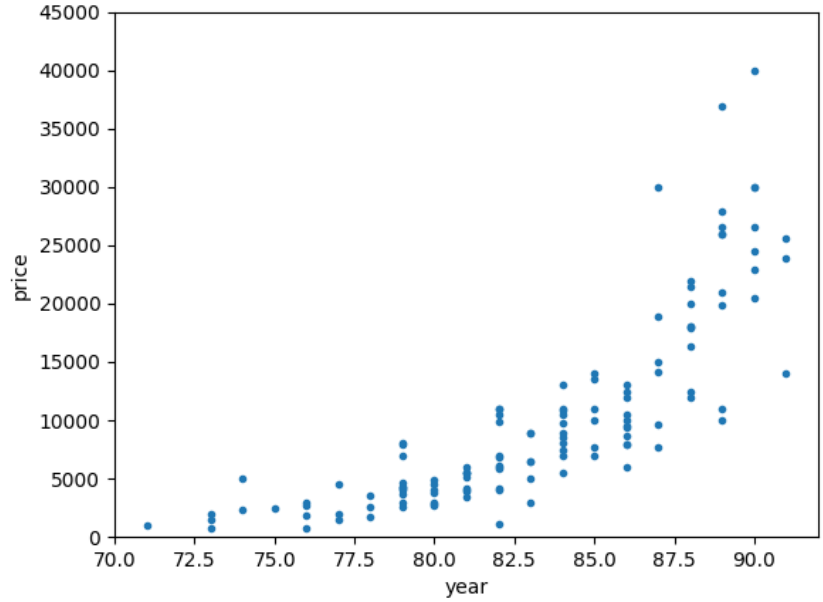


図 1: 【練習問題 5-1】 中古車価格データの完全版

これについて 1〜4 次式モデルで多項式回帰を行う。ここでは最小二乗法を用いて解を求めることとする。

モデルの優劣の決め方について述べる。まず完全版データを教師データとテストデータに分け，教師データを用いてモデルを求め，テストデータでそのモデルを評価する。教師データは乱数によって選択することとし，その乱数の seed をいくつか変えて複数回試行を行う。モデル作成時の AICc とテストデータに対する $Q$（最小二乗和）を求め，それらを単純平均することによってモデルの評価を行う。

なお，赤池情報量基準については，$n$ が小さいことから，

$$\mathrm{AICc} = n \log \frac{\hat{Q}}{n} + \frac{2kn}{n-k-1} \tag{1}$$

を用いる。

ここでは試行回数を 100 回とする。また，配布資料に倣って，教師データの数を 10 としてモデルを求めたところ，その結果は以下のようになった。なお，図は適当な seed のときのモデルをプロットしたものである。

表 1: 【練習問題 5-1】 テストデータ 10 個のときの各次元に対する AICc と $Q$ の平均値

| 次元 | AICc | $Q$ |
|---|---|---|
| 1 | 205.27 | $3.808\,51 \times 10^{9}$ |
| 2 | 215.66 | $3.804\,75 \times 10^{9}$ |
| 3 | 282.89 | $4.988\,13 \times 10^{10}$ |
| 4 | 116.06 | $3.608\,80 \times 10^{10}$ |

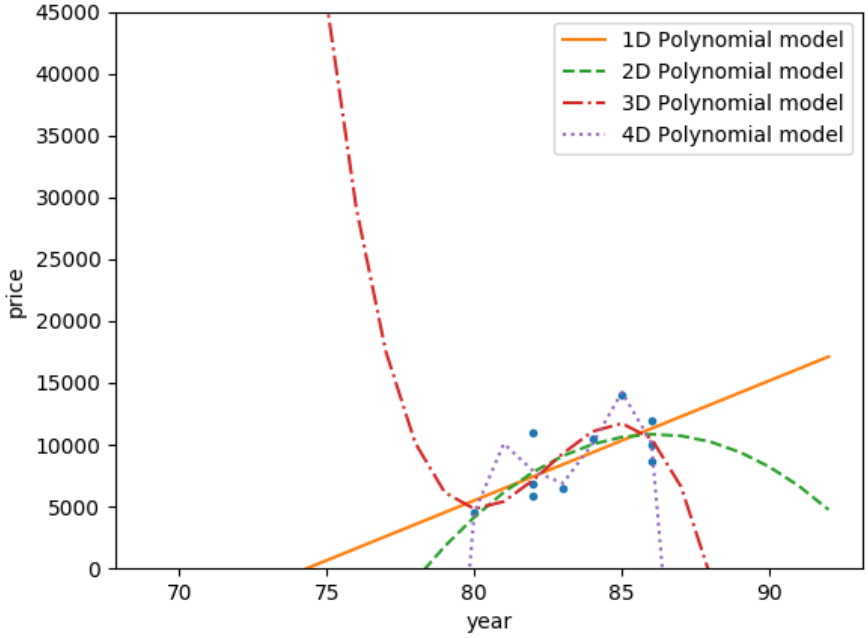


図 2: テストデータ 10 個のときの train 時の fitting の図

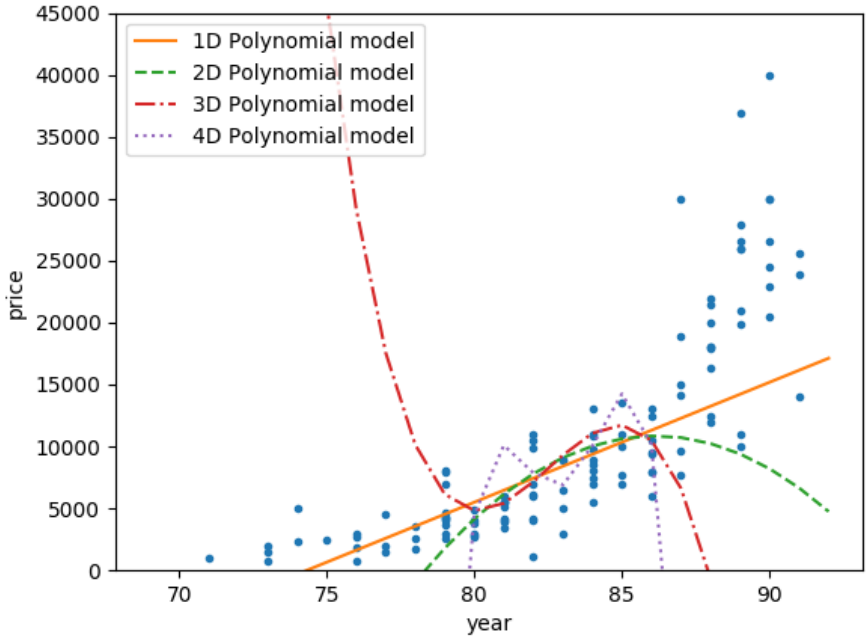


図 3: テストデータ 10 個のときの test 時の fitting の図

表 1 より，AICc の値からすれば 4 次が，$Q$ の値からすれば 1，2 次が優れているといえるが，図 2，3 を見

れば 2，4 次は妥当ではないと思われる。これは，テストデータが少なすぎるため，2 次ではうまく fit できず，また 4 次では過学習してしまって AICc が小さくなっているのだと考えられる。従って 1 次式が優れたモデルである，と言えなくもないが，元データは 1 次式というよりは曲線に見える。

以上の問題はテストデータが少なすぎたために起こっていると考えられる。そこで，全てのデータの 30% をテストに利用した場合について調べてみたところ，その結果は以下のようになった。

表 2: 【練習問題 5-1】 テストデータ 30% のときの各次元に対する AICc と $Q$ の平均値

| 次元 | AICc | $Q$ |
|---|---|---|
| 1 | 671.43 | $2.514\,12 \times 10^9$ |
| 2 | 662.66 | $1.843\,39 \times 10^9$ |
| 3 | 667.49 | $1.943\,58 \times 10^9$ |
| 4 | 676.40 | $2.323\,40 \times 10^9$ |

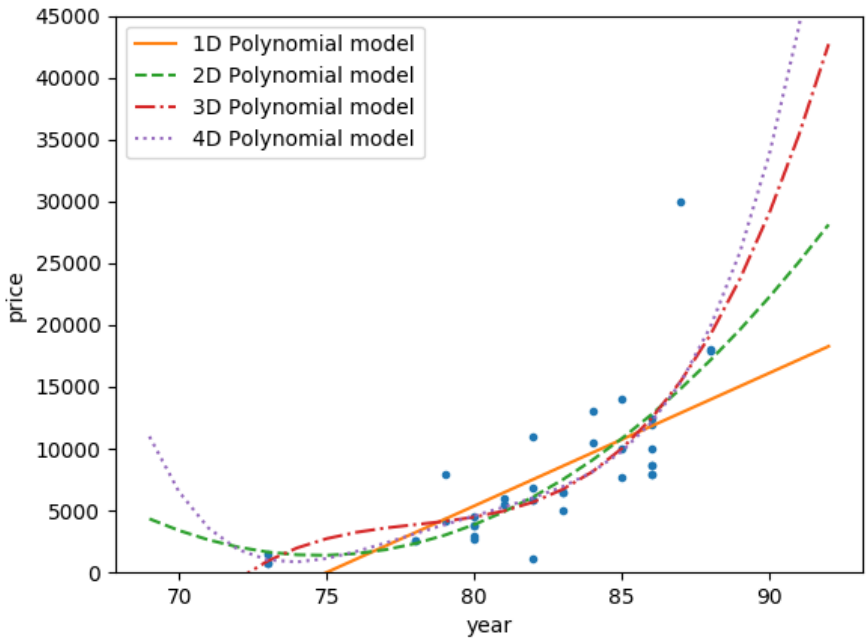


図 4: テストデータ 30% のときの train 時の fitting の図

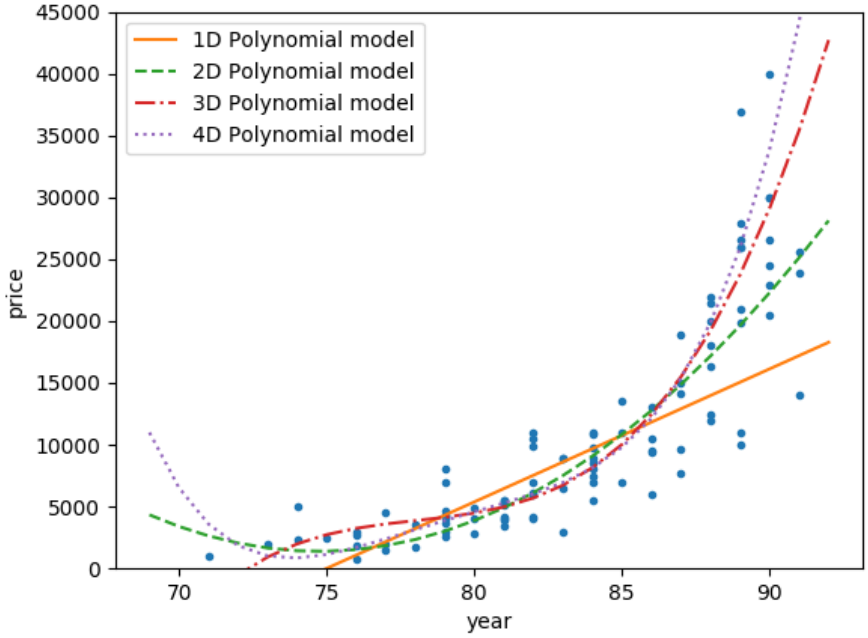


図 5: テストデータ 30% のときの test 時の fitting の図

表 2 より，テストデータを全体の 30% としたところ，AICc，$Q$ ともに 2 次が最も優れている。また，図を見ても大きな違和感はない。

以上の考察により，2 次式モデルが最も優れているといえる。