3. Polynomial Regression

Bruce E. Shapiro

Getting Started in Machine Learning

Last revised: February 3, 2019

Goal of Polynomial Regression

Given n points

$$(x_0, y_0), (x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$$

Fit a polynomial of degree $p > n$ (usually $n \gg p$)

$$P(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_p x^p$$

that minimizes the objective function

$$\mathcal{E} = \sum_{i=0}^{n-1} (y_i - P(x_i))^2 = \sum_{i=0}^{n-1} \left(y_i - \sum_{j=0}^{p} c_j x_i^j \right)^2$$

Polynomial Regression: Derivation of Normal Equations (1)

Define the **residual error** r_i by

$$P(x_i) = y_i + r_i$$

Then substituting $(x_0, y_0), \ldots, (x_{n-1}, y_{n-1})$, where $n > p$,

$$\begin{aligned}
c_0 + c_1 x_0 + c_2 x_0^2 + \cdots + c_p x_0^p &= y_0 + r_0 \\
c_0 + c_1 x_1 + c_2 x_1^2 + \cdots + c_p x_1^p &= y_1 + r_1 \\
&\vdots \\
c_0 + c_1 x_{n-1} + c_2 x_{n-1}^2 + \cdots + c_p x_{n-1}^p &= y_{n-1} + r_{n-1}
\end{aligned}$$

Polynomial Regression: Derivation of Normal Equations (2)

Rewrite in matrix form: $\mathbf{Ac} = \mathbf{y} + \mathbf{r}$, where

$$\overbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^p \\ 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ & & & \vdots & \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^p \end{bmatrix}}^{\text{Define this as } \mathbf{A}} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_p \end{bmatrix}}_{\mathbf{c}} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}}_{\mathbf{y}} + \underbrace{\begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_{n-1} \end{bmatrix}}_{\mathbf{r}}$$

Polynomial Regression: Derivation of Normal Equations (2)

Rewrite in matrix form: $\mathbf{Ac} = \mathbf{y} + \mathbf{r}$, where

$$\overbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^p \\ 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ & & & \vdots & \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^p \end{bmatrix}}^{\text{Define this as } \mathbf{A}} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_p \end{bmatrix}}_{\mathbf{c}} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}}_{\mathbf{y}} + \underbrace{\begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_{n-1} \end{bmatrix}}_{\mathbf{r}}$$

$$\mathcal{E} = \sum_{i=0}^{n-1} (y_i - P(x_i))^2 = \sum_{i=0}^{n} (y_i - (\mathbf{Ac})_i)^2$$

Polynomial Regression: Derivation of Normal Equations (2)

Rewrite in matrix form: $\mathbf{Ac} = \mathbf{y} + \mathbf{r}$, where

$$\overbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^p \\ 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ & & & \vdots & \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^p \end{bmatrix}}^{\text{Define this as } \mathbf{A}} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_p \end{bmatrix}}_{\mathbf{c}} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}}_{\mathbf{y}} + \underbrace{\begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_{n-1} \end{bmatrix}}_{\mathbf{r}}$$

$$\mathcal{E} = \sum_{i=0}^{n-1} (y_i - P(x_i))^2 = \sum_{i=0}^{n} (y_i - (\mathbf{Ac})_i)^2 = \sum_{i=0}^{n} (\mathbf{y}\text{-}\mathbf{Ac})_i (\mathbf{y}\text{-}\mathbf{Ac})_i$$

Polynomial Regression: Derivation of Normal Equations (2)

Rewrite in matrix form: **Ac** = **y** + **r**, where

$$\overbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^p \\ 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ & & & \vdots & \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^p \end{bmatrix}}^{\text{Define this as } \mathbf{A}} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_p \end{bmatrix}}_{\mathbf{c}} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}}_{\mathbf{y}} + \underbrace{\begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_{n-1} \end{bmatrix}}_{\mathbf{r}}$$

$$\mathcal{E} = \sum_{i=0}^{n-1} (y_i - P(x_i))^2 = \sum_{i=0}^{n} (y_i - (\mathbf{Ac})_i)^2 = \sum_{i=0}^{n} (\mathbf{y}\text{-}\mathbf{Ac})_i (\mathbf{y}\text{-}\mathbf{Ac})_i$$

$$= \sum_{i=0}^{n} ((\mathbf{y}\text{-}\mathbf{Ac})^{\mathsf{T}})_i (\mathbf{y}\text{-}\mathbf{Ac})_i$$

Polynomial Regression: Derivation of Normal Equations (2)

Rewrite in matrix form: **Ac** = **y** + **r**, where

$$\overbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^p \\ 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ & & & \vdots & \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^p \end{bmatrix}}^{\text{Define this as } \mathbf{A}} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_p \end{bmatrix}}_{\mathbf{c}} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}}_{\mathbf{y}} + \underbrace{\begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_{n-1} \end{bmatrix}}_{\mathbf{r}}$$

$$\mathcal{E} = \sum_{i=0}^{n-1} (y_i - P(x_i))^2 = \sum_{i=0}^{n} (y_i - (\mathbf{Ac})_i)^2 = \sum_{i=0}^{n} (\mathbf{y}\text{-}\mathbf{Ac})_i (\mathbf{y}\text{-}\mathbf{Ac})_i$$

$$= \sum_{i=0}^{n} ((\mathbf{y}\text{-}\mathbf{Ac})^{\mathsf{T}})_i (\mathbf{y}\text{-}\mathbf{Ac})_i = (\mathbf{y}\text{-}\mathbf{Ac})^{\mathsf{T}} (\mathbf{y}\text{-}\mathbf{Ac})$$

Polynomial Regression: Derivation of Normal Equations (3)

Let

$$\mathbf{A} = \left[\begin{array}{c|c|c|c} \mathbf{a}_0 & \mathbf{a}_1 & \cdots & \mathbf{a}_p \end{array} \right]$$

where

$$\mathbf{a}_j = \begin{bmatrix} x_0^j \\ x_1^j \\ x_2^j \\ \vdots \\ x_{n-1}^j \end{bmatrix}$$

Then

$$\mathbf{Ac} = c_0\mathbf{a}_0 + c_1\mathbf{a}_1 + \cdots c_p\mathbf{a}_p$$

Hence

$$\frac{\partial \mathbf{Ac}}{\partial c_i} = \mathbf{a}_i$$

Polynomial Regression: Derivation of Normal Equations (4)

Using $\mathcal{E} = (\mathbf{y}\text{-}\mathbf{Ac})^{\mathsf{T}}(\mathbf{y}\text{-}\mathbf{Ac})$ and $\dfrac{\partial \mathbf{Ac}}{\partial c_i} = \mathbf{a}_i$,

$$0 = \frac{\partial \mathcal{E}}{\partial c_i} = (\mathbf{y}\text{-}\mathbf{Ac})^{\mathsf{T}} \frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac}) + \left[\frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac})^{\mathsf{T}}\right](\mathbf{y}\text{-}\mathbf{Ac})$$

Polynomial Regression: Derivation of Normal Equations (4)

Using $\mathcal{E} = (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}(\mathbf{y}\text{-}\mathbf{Ac})$ and $\dfrac{\partial \mathbf{Ac}}{\partial c_i} = \mathbf{a}_i$,

$$0 = \frac{\partial \mathcal{E}}{\partial c_i} = (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T} \frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac}) + \left[\frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}\right](\mathbf{y}\text{-}\mathbf{Ac})$$

$$= (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}(-\mathbf{a}_i) + \left[(-\mathbf{a}_i)^\mathsf{T}\right](\mathbf{y}\text{-}\mathbf{Ac})$$

Polynomial Regression: Derivation of Normal Equations (4)

Using $\mathcal{E} = (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}(\mathbf{y}\text{-}\mathbf{Ac})$ and $\dfrac{\partial \mathbf{Ac}}{\partial c_i} = \mathbf{a}_i$,

$$
\begin{aligned}
0 = \frac{\partial \mathcal{E}}{\partial c_i} &= (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T} \frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac}) + \left[\frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}\right](\mathbf{y}\text{-}\mathbf{Ac}) \\
&= (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}(-\mathbf{a}_i) + \left[(-\mathbf{a}_i)^\mathsf{T}\right](\mathbf{y}\text{-}\mathbf{Ac}) \\
&= -\mathbf{y}^\mathsf{T}\mathbf{a}_i + \mathbf{c}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{a}_i - \mathbf{a}_i^\mathsf{T}\mathbf{y} + \mathbf{a}_i^\mathsf{T}\mathbf{Ac}
\end{aligned}
$$

Polynomial Regression: Derivation of Normal Equations (4)

Using $\mathcal{E} = (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}(\mathbf{y}\text{-}\mathbf{Ac})$ and $\dfrac{\partial \mathbf{Ac}}{\partial c_i} = \mathbf{a}_i$,

$$
\begin{aligned}
0 = \frac{\partial \mathcal{E}}{\partial c_i} &= (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}\frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac}) + \left[\frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}\right](\mathbf{y}\text{-}\mathbf{Ac}) \\
&= (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}(-\mathbf{a}_i) + \left[(-\mathbf{a}_i)^\mathsf{T}\right](\mathbf{y}\text{-}\mathbf{Ac}) \\
&= -\mathbf{y}^\mathsf{T}\mathbf{a}_i + \mathbf{c}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{a}_i - \mathbf{a}_i^\mathsf{T}\mathbf{y} + \mathbf{a}_i^\mathsf{T}\mathbf{Ac} \\
&= -2\mathbf{a}_i^\mathsf{T}\mathbf{y} + 2\mathbf{a}_i^\mathsf{T}\mathbf{Ac}
\end{aligned}
$$

Polynomial Regression: Derivation of Normal Equations (4)

Using $\mathcal{E} = (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}(\mathbf{y}\text{-}\mathbf{Ac})$ and $\dfrac{\partial \mathbf{Ac}}{\partial c_i} = \mathbf{a}_i$,

$$
\begin{aligned}
0 = \frac{\partial \mathcal{E}}{\partial c_i} &= (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T} \frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac}) + \left[\frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}\right](\mathbf{y}\text{-}\mathbf{Ac}) \\
&= (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}(-\mathbf{a}_i) + \left[(-\mathbf{a}_i)^\mathsf{T}\right](\mathbf{y}\text{-}\mathbf{Ac}) \\
&= -\mathbf{y}^\mathsf{T}\mathbf{a}_i + \mathbf{c}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{a}_i - \mathbf{a}_i^\mathsf{T}\mathbf{y} + \mathbf{a}_i^\mathsf{T}\mathbf{Ac} \\
&= -2\mathbf{a}_i^\mathsf{T}\mathbf{y} + 2\mathbf{a}_i^\mathsf{T}\mathbf{Ac} \\
\mathbf{a}_i^\mathsf{T}\mathbf{y} &= \mathbf{a}_i^\mathsf{T}\mathbf{Ac}
\end{aligned}
$$

Polynomial Regression: Derivation of Normal Equations (4)

Using $\mathcal{E} = (\mathbf{y}\text{-}\mathbf{Ac})^{\mathsf{T}}(\mathbf{y}\text{-}\mathbf{Ac})$ and $\dfrac{\partial \mathbf{Ac}}{\partial c_i} = \mathbf{a}_i$,

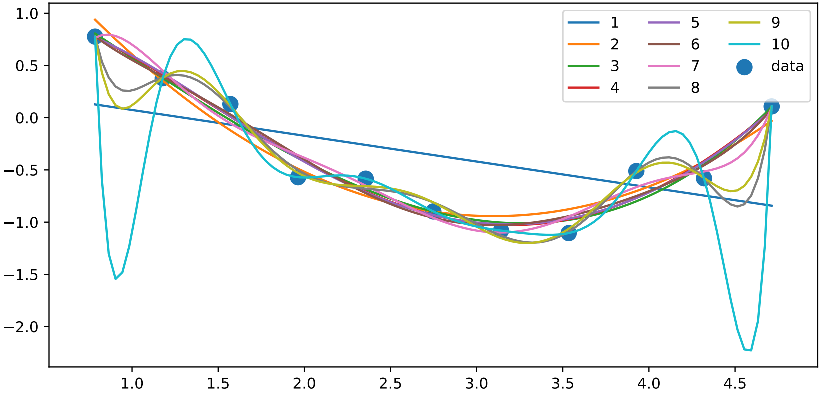
$$
\begin{aligned}
0 = \frac{\partial \mathcal{E}}{\partial c_i} &= (\mathbf{y}\text{-}\mathbf{Ac})^{\mathsf{T}} \frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac}) + \left[\frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac})^{\mathsf{T}}\right](\mathbf{y}\text{-}\mathbf{Ac}) \\
&= (\mathbf{y}\text{-}\mathbf{Ac})^{\mathsf{T}}(-\mathbf{a}_i) + \left[(-\mathbf{a}_i)^{\mathsf{T}}\right](\mathbf{y}\text{-}\mathbf{Ac}) \\
&= -\mathbf{y}^{\mathsf{T}}\mathbf{a}_i + \mathbf{c}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{a}_i - \mathbf{a}_i^{\mathsf{T}}\mathbf{y} + \mathbf{a}_i^{\mathsf{T}}\mathbf{Ac} \\
&= -2\mathbf{a}_i^{\mathsf{T}}\mathbf{y} + 2\mathbf{a}_i^{\mathsf{T}}\mathbf{Ac} \\
\mathbf{a}_i^{\mathsf{T}}\mathbf{y} &= \mathbf{a}_i^{\mathsf{T}}\mathbf{Ac} \\
\mathbf{A}^{\mathsf{T}}\mathbf{y} &= \mathbf{A}^{\mathsf{T}}\mathbf{Ac} \quad \textbf{Normal Equations} \text{ (Solve for c)}
\end{aligned}
$$

Polynomial Regression: Derivation of Normal Equations (4)

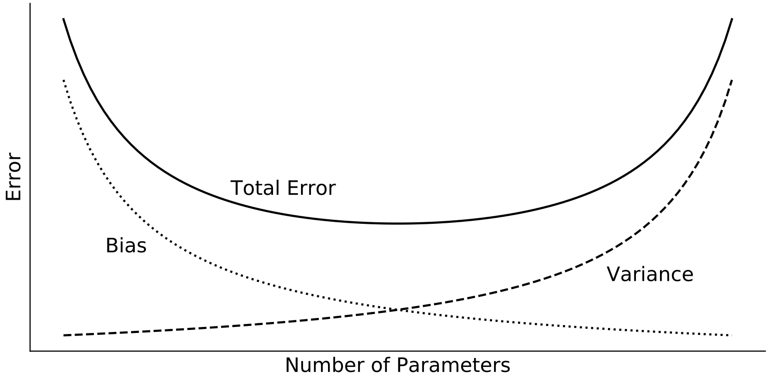
Using $\mathcal{E} = (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}(\mathbf{y}\text{-}\mathbf{Ac})$ and $\dfrac{\partial \mathbf{Ac}}{\partial c_i} = \mathbf{a}_i$,

$$
\begin{aligned}
0 = \frac{\partial \mathcal{E}}{\partial c_i} &= (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T} \frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac}) + \left[\frac{\partial}{\partial c_i}(\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}\right](\mathbf{y}\text{-}\mathbf{Ac}) \\
&= (\mathbf{y}\text{-}\mathbf{Ac})^\mathsf{T}(-\mathbf{a}_i) + \left[(-\mathbf{a}_i)^\mathsf{T}\right](\mathbf{y}\text{-}\mathbf{Ac}) \\
&= -\mathbf{y}^\mathsf{T}\mathbf{a}_i + \mathbf{c}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{a}_i - \mathbf{a}_i^\mathsf{T}\mathbf{y} + \mathbf{a}_i^\mathsf{T}\mathbf{Ac} \\
&= -2\mathbf{a}_i^\mathsf{T}\mathbf{y} + 2\mathbf{a}_i^\mathsf{T}\mathbf{Ac} \\
\mathbf{a}_i^\mathsf{T}\mathbf{y} &= \mathbf{a}_i^\mathsf{T}\mathbf{Ac} \\
\mathbf{A}^\mathsf{T}\mathbf{y} &= \mathbf{A}^\mathsf{T}\mathbf{Ac} \quad \textbf{Normal Equations} \text{ (Solve for c)} \\
\mathbf{c} &= (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{y} \quad \text{Theoretical Solution}
\end{aligned}
$$

What is the *right* value for p?



The Bias-Variance Trade-off



Bias-Variance Trade-off (Derivation)

Let $E[]$ be the expectation and var$[]$ the variance operator.
Then $E[y_i] = y_i$ and var$[y] = \sigma^2$ Further, since $E[X^2] = \text{var}[X] + E[X]^2$ for any random variable X,

$$E[(y_i - \hat{y})_i^2] = E[y_i^2 + \hat{y}_i^2 - 2y_i\hat{y}_i]$$

Bias-Variance Trade-off (Derivation)

Let $E[]$ be the expectation and var$[]$ the variance operator.
Then $E[y_i] = y_i$ and var$[y] = \sigma^2$ Further, since $E[X^2] = \text{var}[X] + E[X]^2$ for any random variable X,

$$\begin{aligned} E[(y_i - \hat{y})_i^2] &= E[y_i^2 + \hat{y}_i^2 - 2y_i\hat{y}_i] = E[y_i^2] + E[\hat{y}_i^2] - 2y_iE[\hat{y}_i] \\ &= \text{var}[y_i] + E[y_i]^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \end{aligned}$$

Bias-Variance Trade-off (Derivation)

Let $E[]$ be the expectation and var[] the variance operator.
Then $E[y_i] = y_i$ and $\text{var}[y] = \sigma^2$ Further, since $E[X^2] = \text{var}[X] + E[X]^2$ for any random variable X,

$$\begin{aligned} E[(y_i - \hat{y})_i^2] &= E[y_i^2 + \hat{y}_i^2 - 2y_i\hat{y}_i] = E[y_i^2] + E[\hat{y}_i^2] - 2y_iE[\hat{y}_i] \\ &= \text{var}[y_i] + E[y_i]^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \\ &= \sigma^2 + y_i^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \end{aligned}$$

Bias-Variance Trade-off (Derivation)

Let $E[]$ be the expectation and var$[]$ the variance operator.
Then $E[y_i] = y_i$ and var$[y] = \sigma^2$ Further, since $E[X^2] = \text{var}[X] + E[X]^2$ for any random variable X,

$$
\begin{aligned}
E[(y_i - \hat{y})_i^2] &= E[y_i^2 + \hat{y}_i^2 - 2y_i\hat{y}_i] = E[y_i^2] + E[\hat{y}_i^2] - 2y_iE[\hat{y}_i] \\
&= \text{var}[y_i] + E[y_i]^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \\
&= \sigma^2 + y_i^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \\
&= \sigma^2 + \text{var}[\hat{y}_i] + \left(y_i^2 + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] + \right)
\end{aligned}
$$

Bias-Variance Trade-off (Derivation)

Let $E[]$ be the expectation and var$[]$ the variance operator.
Then $E[y_i] = y_i$ and var$[y] = \sigma^2$ Further, since $E[X^2] = \text{var}[X] + E[X]^2$ for any random variable X,

$$
\begin{aligned}
E[(y_i - \hat{y})_i^2] &= E[y_i^2 + \hat{y}_i^2 - 2y_i\hat{y}_i] = E[y_i^2] + E[\hat{y}_i^2] - 2y_iE[\hat{y}_i] \\
&= \text{var}[y_i] + E[y_i]^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \\
&= \sigma^2 + y_i^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \\
&= \sigma^2 + \text{var}[\hat{y}_i] + \left(y_i^2 + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] + \right) \\
&= \sigma^2 + \text{var}[\hat{y}_i] + (y_i - E[\hat{y}_i])^2
\end{aligned}
$$

Bias-Variance Trade-off (Derivation)

Let $E[]$ be the expectation and var[] the variance operator.
Then $E[y_i] = y_i$ and $\text{var}[y] = \sigma^2$ Further, since $E[X^2] = \text{var}[X] + E[X]^2$ for any random variable X,

$$
\begin{aligned}
E[(y_i - \hat{y})_i^2] &= E[y_i^2 + \hat{y}_i^2 - 2y_i\hat{y}_i] = E[y_i^2] + E[\hat{y}_i^2] - 2y_iE[\hat{y}_i] \\
&= \text{var}[y_i] + E[y_i]^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \\
&= \sigma^2 + y_i^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \\
&= \sigma^2 + \text{var}[\hat{y}_i] + \left(y_i^2 + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] + \right) \\
&= \sigma^2 + \text{var}[\hat{y}_i] + (y_i - E[\hat{y}_i])^2 \\
&= \sigma^2 + \text{var}[\hat{y}_i] + E[y_i - \hat{y}_i]^2
\end{aligned}
$$

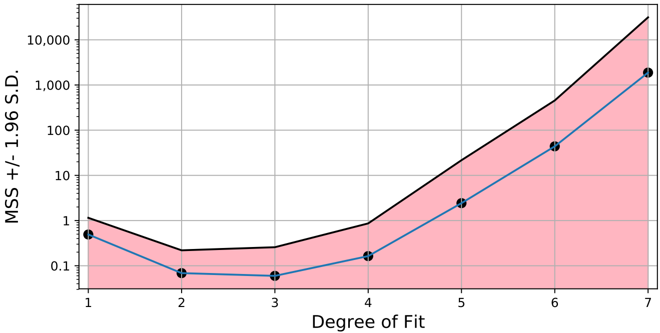
Bias-Variance Trade-off (Derivation)

Let $E[]$ be the expectation and var[] the variance operator.
Then $E[y_i] = y_i$ and $\text{var}[y] = \sigma^2$ Further, since $E[X^2] = \text{var}[X] + E[X]^2$ for any random variable X,

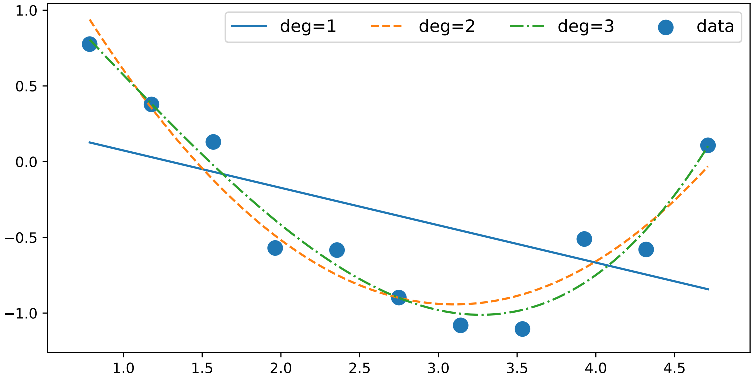
$$\begin{aligned}
E[(y_i - \hat{y})_i^2] &= E[y_i^2 + \hat{y}_i^2 - 2y_i\hat{y}_i] = E[y_i^2] + E[\hat{y}_i^2] - 2y_iE[\hat{y}_i] \\
&= \text{var}[y_i] + E[y_i]^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \\
&= \sigma^2 + y_i^2 + \text{var}[\hat{y}_i] + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] \\
&= \sigma^2 + \text{var}[\hat{y}_i] + \left(y_i^2 + E[\hat{y}_i]^2 - 2y_iE[\hat{y}_i] + \right) \\
&= \sigma^2 + \text{var}[\hat{y}_i] + (y_i - E[\hat{y}_i])^2 \\
&= \sigma^2 + \text{var}[\hat{y}_i] + E[y_i - \hat{y}_i]^2 \\
&= \underbrace{\sigma^2}_{\text{noise}} + \underbrace{\text{var}[\hat{y}_i]}_{\text{variance}} + \underbrace{E[y_i - \hat{y}_i]^2}_{\text{bias}}
\end{aligned}$$

Toy Data

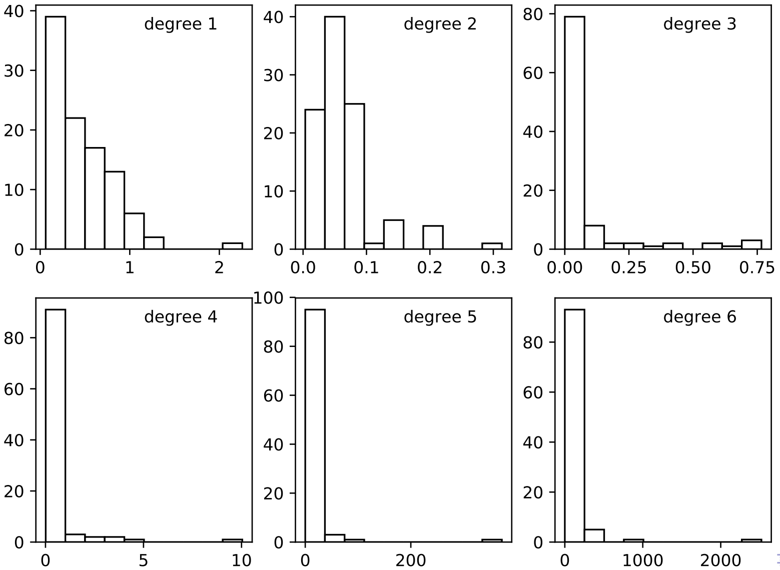
x	y
0.79	0.78
1.18	0.38
1.57	0.13
1.96	-0.57
2.36	-0.58
2.75	-0.9
3.14	-1.08
3.53	-1.11
3.93	-0.51
4.32	-0.58
4.71	0.11



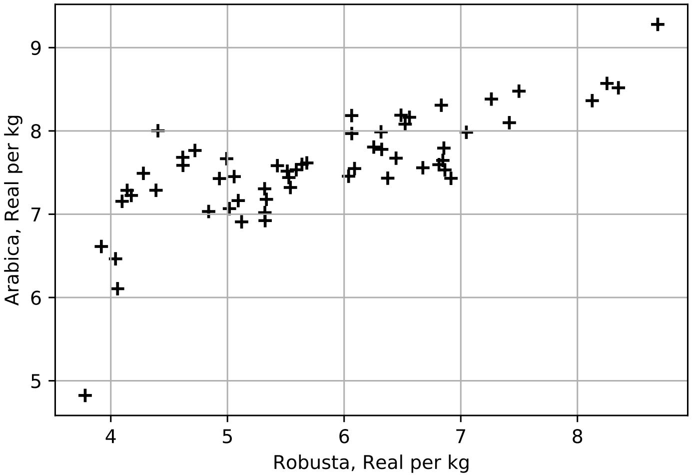
Is degree 3 better than 2?



Mean Square Error (MSE)



Price of Brazilian Coffee Beans



Perform fits on degrees 1 through maxdeg

Returns a list of maxdeg MSE values

```
import numpy as np

def fitsets(XTRAIN,YTRAIN,XTEST,YTEST,maxdeg):
    ntest=len(XTEST)
    results=[]
    for p in range(1,maxdeg+1):
        fit=np.polyfit(XTRAIN, YTRAIN,p)
        yfit=np.polyval(fit,XTEST)
        MSS=sum((yfit-YTEST)**2)/ntest
        results.append(MSS)
    return(np.array(results).T)
```

Repeat the fits through 100 times, thru degree 7

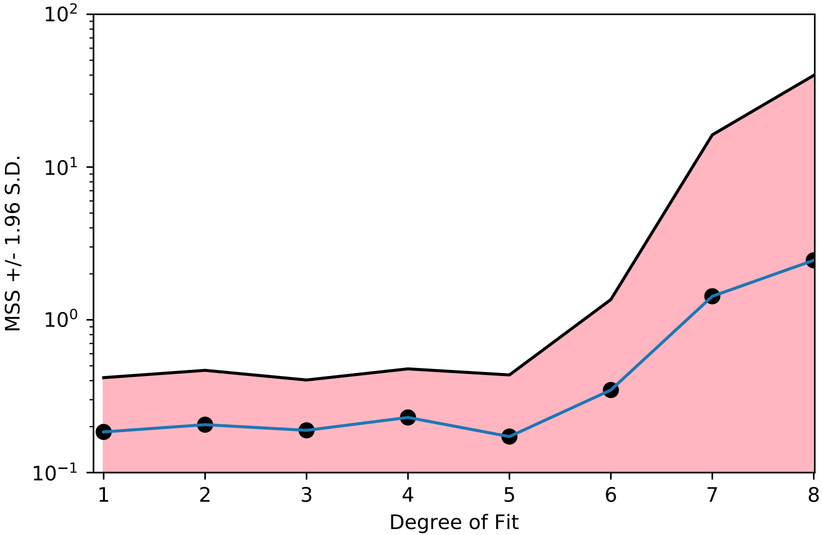
```
from sklearn.model_selection import train_test_split
y=np.array(arabica_prices)
x=np.array(robusta_prices)
nruns=100; maxdeg=7; MSS_DAT=[]
for rz in range(nruns):
    XTR,YTR,XT,YT=train_test_split(x,y,.75)
    MSS=fitsets(XTR,YTR,XT,YT,maxdeg)
    MSS_DAT.append(list(MSS))
MSS_DAT=np.array(MSS_DAT)
```

Repeat the fits through 100 times, thru degree 7

```
from sklearn.model_selection import train_test_split
y=np.array(arabica_prices)
x=np.array(robusta_prices)
nruns=100; maxdeg=7; MSS_DAT=[]
for rz in range(nruns):
    XTR,YTR,XT,YT=train_test_split(x,y,.75)
    MSS=fitsets(XTR,YTR,XT,YT,maxdeg)
    MSS_DAT.append(list(MSS))
MSS_DAT=np.array(MSS_DAT)
```

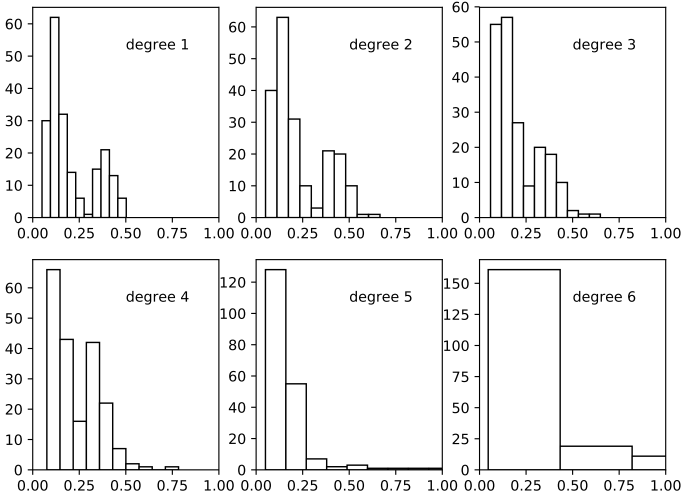
Each row of MSS_DAT has 7 values There will be 100 rows for the 100 runs. Find mean and standard deviation by column

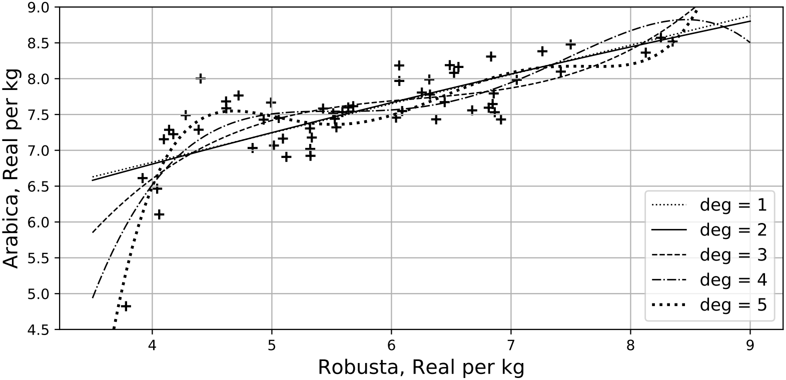
```
means=np.mean(MSS_DAT,axis=0)
stdevs=np.std(MSS_DAT,axis=0)
```



10^2
10^1
10^0
10^{-1}
MSS +/- 1.96 S.D.
1
2
3
4
5
6
7
8
Degree of Fit

Coffee Fit: MSE





9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
Arabica, Real per kg
4
5
6
7
8
9
Robusta, Real per kg
deg = 1
deg = 2
deg = 3
deg = 4
deg = 5

Citations

1. USA Foreign Agricultural Global Agricultural Information Network (GAIN) Report, BR18027, 15 Nov 2018, downloaded from https://gain.fas.usda.gov/RecentGAINPublications/CoffeeSemi-annual_SaoPauloATO_Brazil_11-5-2018.pdf