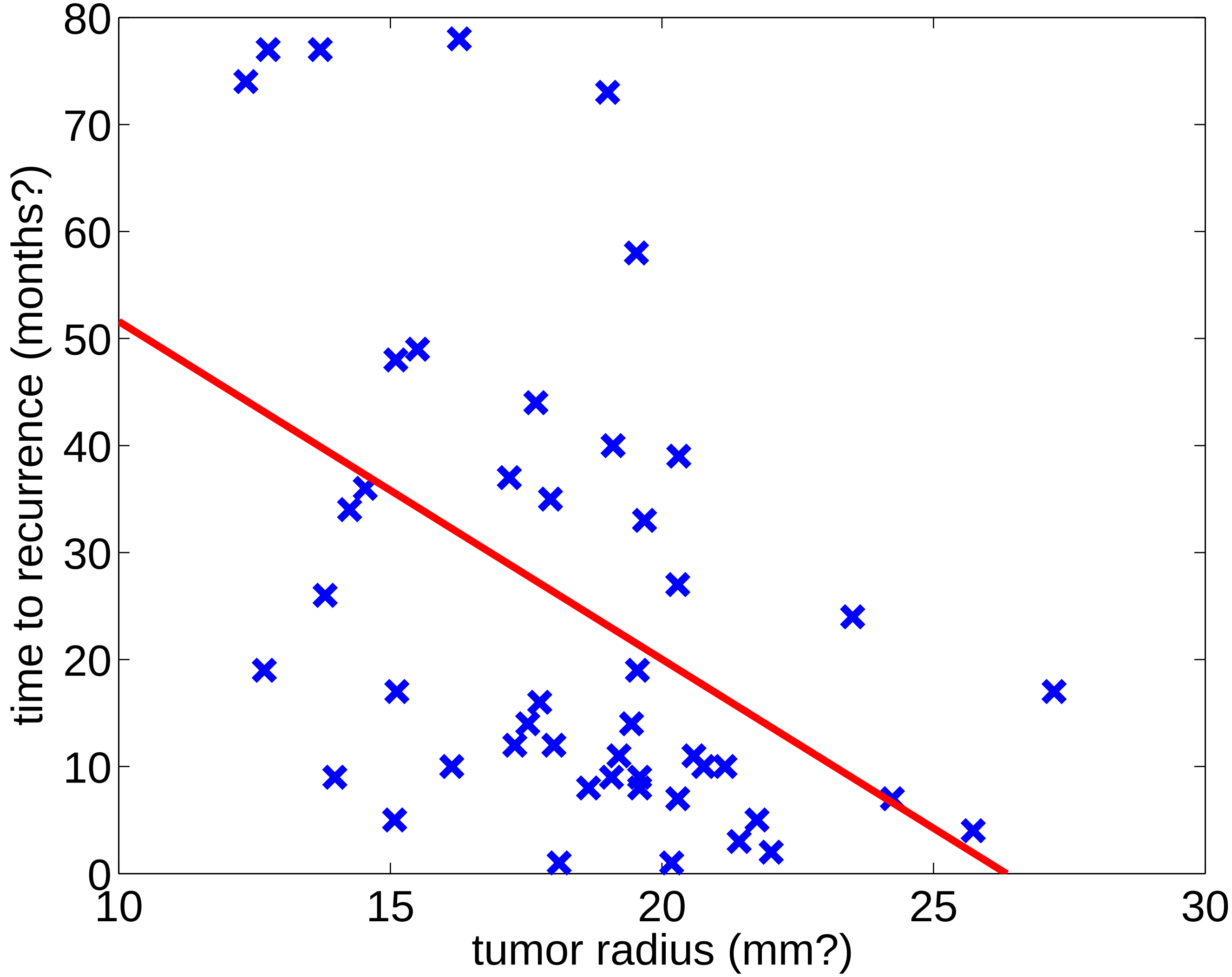
Recall – linear regression



Recall – minimizing sum-squared error

- We want a *weight vector* w such that $Xw \approx Y$, where X is the data matrix augmented with a column of 1's.
- Find w to minimize $SSQ = \sum_{i=1}^{m} (\mathbf{x}_i w - \mathbf{y}_i)^2$.
- Setting $\frac{\partial SSQ}{w_j} = 0$ for all j gives $(n+1)$ linear equations in $(n+1)$ unknowns.

The solution

- Recalling some multivariate calculus:

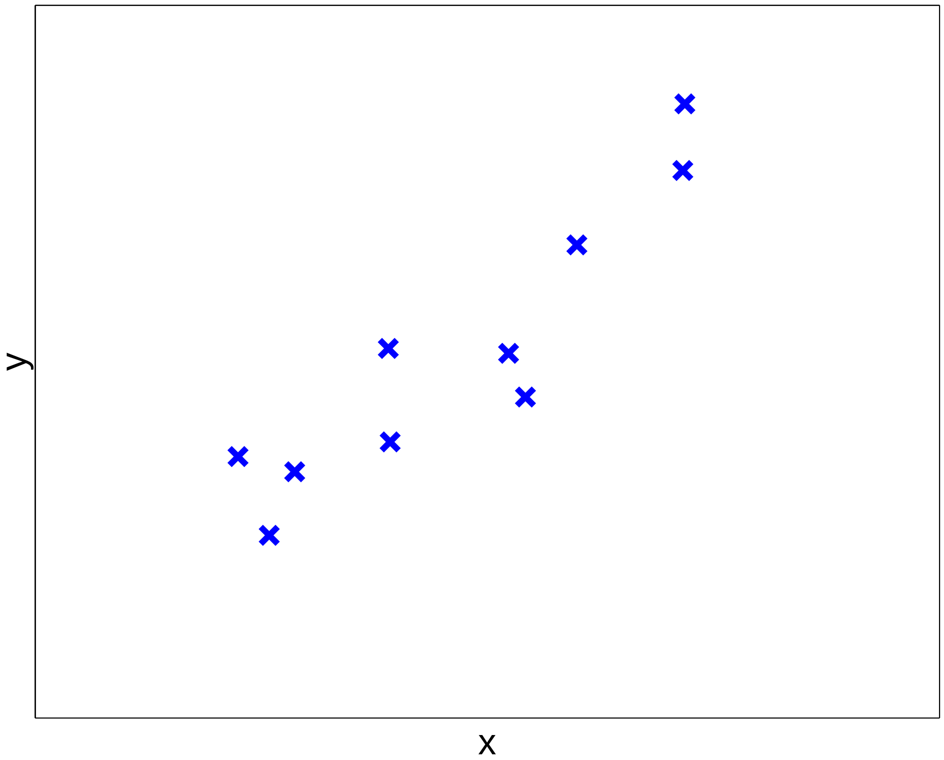
$$
\begin{aligned}
\nabla_{\mathbf{w}} SSQ &= \nabla_{\mathbf{w}} (X\mathbf{w} - Y)^T (X\mathbf{w} - Y) \\
&= \nabla_{\mathbf{w}} (\mathbf{w}^T X^T X\mathbf{w} - Y^T X\mathbf{w} - \mathbf{w}^T X^T Y - Y^T Y) \\
&= 2X^T X\mathbf{w} - 2X^T Y
\end{aligned}
$$

- Setting equal to zero:

$$
\begin{aligned}
2X^T X\mathbf{w} - 2X^T Y &= 0 \\
\Rightarrow X^T X\mathbf{w} &= X^T Y \\
\Rightarrow \mathbf{w} = (X^T X)^{-1} X^T Y &
\end{aligned}
$$

- The inverse exists if the columns of X are linearly independent.

Example of linear regression



x	y
0.86	2.49
0.09	0.83
-0.85	-0.25
0.87	3.10
-0.44	0.87
-0.43	0.02
-1.10	-0.12
0.40	1.81
-0.96	-0.83
0.17	0.43

Data matrices

$$X = \begin{bmatrix} 0.86 & 1 \\ 0.09 & 1 \\ -0.85 & 1 \\ 0.87 & 1 \\ -0.44 & 1 \\ -0.43 & 1 \\ -1.10 & 1 \\ 0.40 & 1 \\ -0.96 & 1 \\ 0.17 & 1 \end{bmatrix} \qquad Y = \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix}$$

$\underline{X^T X}$

$$X^T X =$$

$$\begin{bmatrix} 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0.86 & 1 \\ 0.09 & 1 \\ -0.85 & 1 \\ 0.87 & 1 \\ -0.44 & 1 \\ -0.43 & 1 \\ -1.10 & 1 \\ 0.40 & 1 \\ -0.96 & 1 \\ 0.17 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 4.95 & -1.39 \\ -1.39 & 10 \end{bmatrix}$$

$\underline{X^TY}$

$$X^TY =$$

$$\begin{bmatrix} 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix}$$

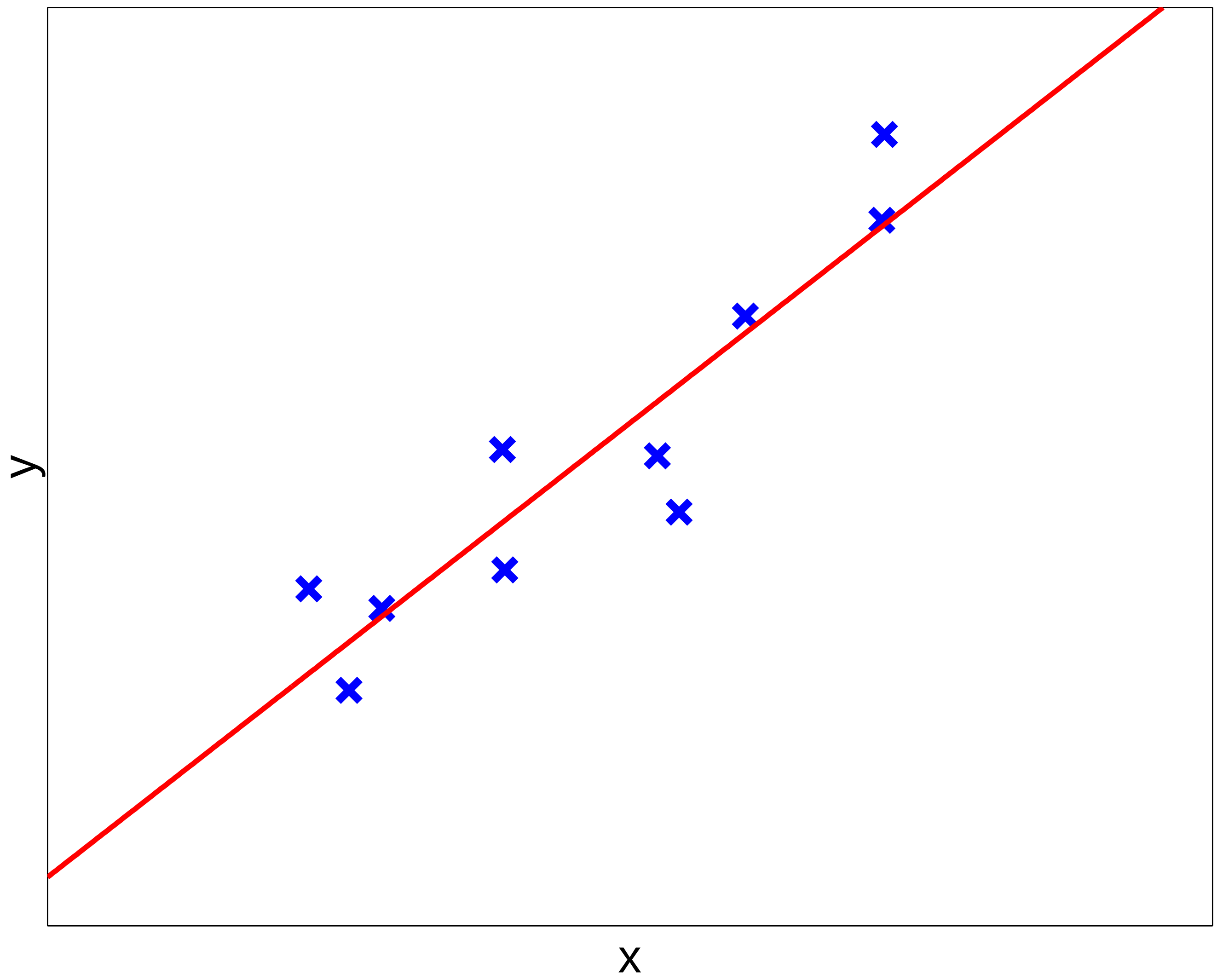
$$= \begin{bmatrix} 6.49 \\ 8.34 \end{bmatrix}$$

Solving for w

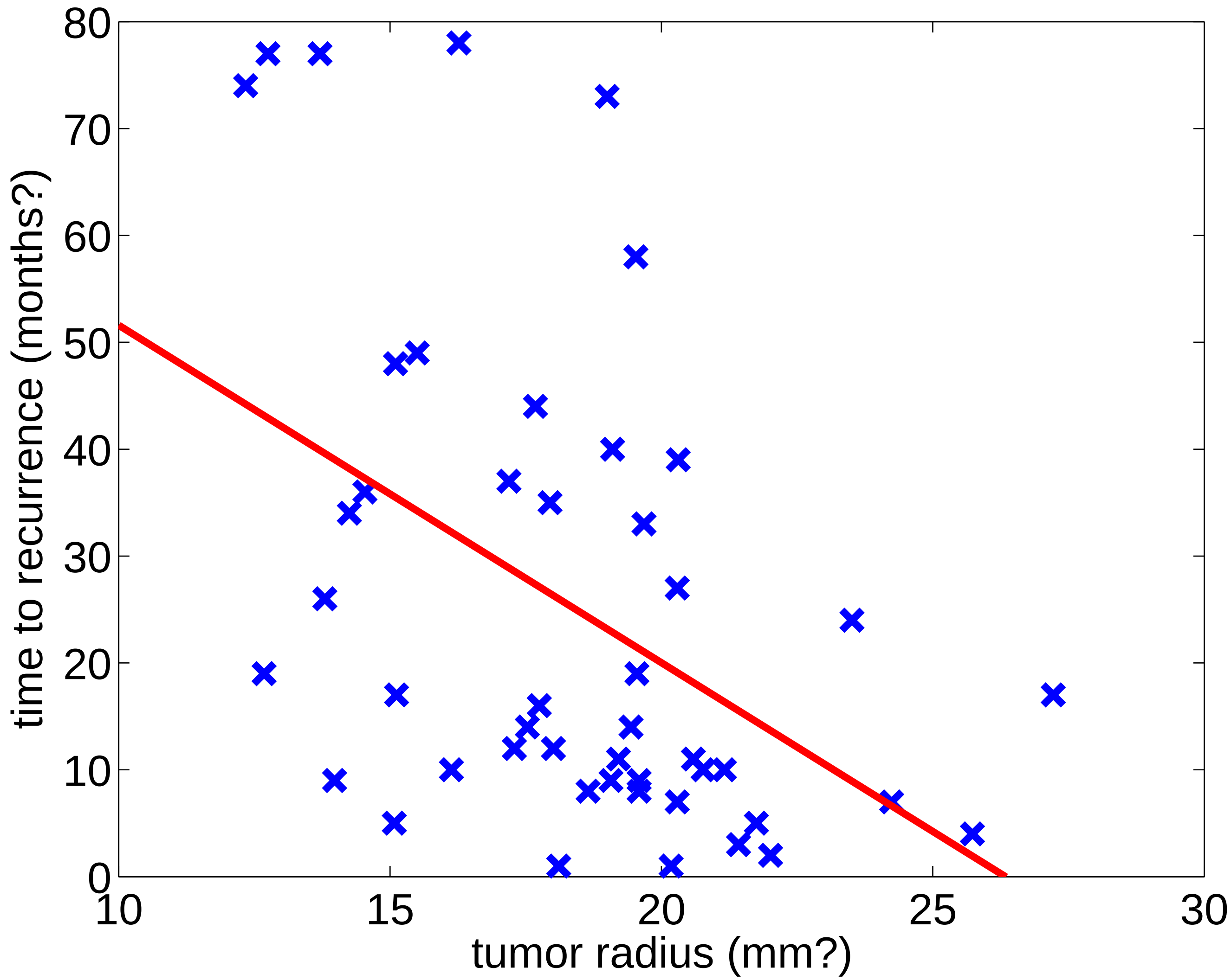
$$w = (X^T X)^{-1} X^Y = \begin{bmatrix} 4.95 & -1.39 \\ -1.39 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 6.49 \\ 8.34 \end{bmatrix} = \begin{bmatrix} 1.60 \\ 1.05 \end{bmatrix}$$

So the best fit line is $y = 1.60x + 1.05$.

Data and line $y = 1.60x + 1.05$



Predicting recurrence time based on tumor size (again)



Linear regression summary

- The optimal linear regression (minimizing sum-squared-error) can be computed in polynomial time.
- The solution is $w = (X^T X)^{-1} X^T Y$, where X is the data matrix augmented with a column of ones, and Y is the column vector of target outputs.
- What if $X^T X$ is not invertible?

Polynomial regression

Polynomial fits

- Suppose we want to fit a higher-degree polynomial to the data. (E.g., $y = w_2 x^2 + w_1 x^1 + w_0$.)
- Suppose for now that there is a single input variable per training sample.
- How do we do it?

Answer: linear regression

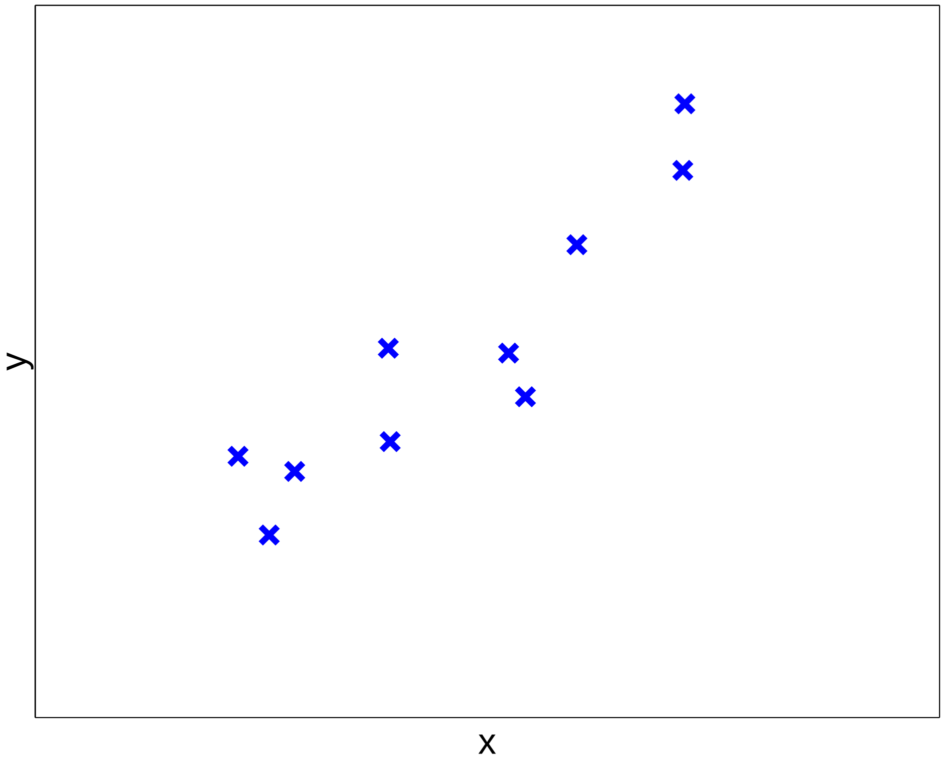
(Sometimes called polynomial regression.)

- Given data: $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$.
- Suppose we want a degree-d polynomial fit.
- Let Y be as before and let

$$X = \begin{bmatrix} x_1^d & \ldots & x_1^2 & x_1 & 1 \\ x_2^d & \ldots & x_2^2 & x_2 & 1 \\ \vdots & & \vdots & \vdots & \vdots \\ x_m^d & \ldots & x_m^2 & x_m & 1 \end{bmatrix}$$

- Solve the linear regression $Xw \approx Y$.

Example of quadratic regression



x	y
0.86	2.49
0.09	0.83
-0.85	-0.25
0.87	3.10
-0.44	0.87
-0.43	0.02
-1.10	-0.12
0.40	1.81
-0.96	-0.83
0.17	0.43

Data matrices

$$X = \begin{bmatrix} 0.75 & 0.86 & 1 \\ 0.01 & 0.09 & 1 \\ 0.73 & -0.85 & 1 \\ 0.76 & 0.87 & 1 \\ 0.19 & -0.44 & 1 \\ 0.18 & -0.43 & 1 \\ 1.22 & -1.10 & 1 \\ 0.16 & 0.40 & 1 \\ 0.93 & -0.96 & 1 \\ 0.03 & 0.17 & 1 \end{bmatrix} \qquad Y = \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix}$$

$X^T X$

$$X^T X =$$

$$\begin{bmatrix} 0.75 & 0.01 & 0.73 & 0.76 & 0.19 & 0.18 & 1.22 & 0.16 & 0.93 & 0.03 \\ 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0.75 & 0.86 & 1 \\ 0.01 & 0.09 & 1 \\ 0.73 & -0.85 & 1 \\ 0.76 & 0.87 & 1 \\ 0.19 & -0.44 & 1 \\ 0.18 & -0.43 & 1 \\ 1.22 & -1.10 & 1 \\ 0.16 & 0.40 & 1 \\ 0.93 & -0.96 & 1 \\ 0.03 & 0.17 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 4.11 & -1.64 & 4.95 \\ -1.64 & 4.95 & -1.39 \\ 4.95 & -1.39 & 10 \end{bmatrix}$$

$X^T Y$

$$X^T Y =$$

$$\begin{bmatrix} 0.75 & 0.01 & 0.73 & 0.76 & 0.19 & 0.18 & 1.22 & 0.16 & 0.93 & 0.03 \\ 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix}$$

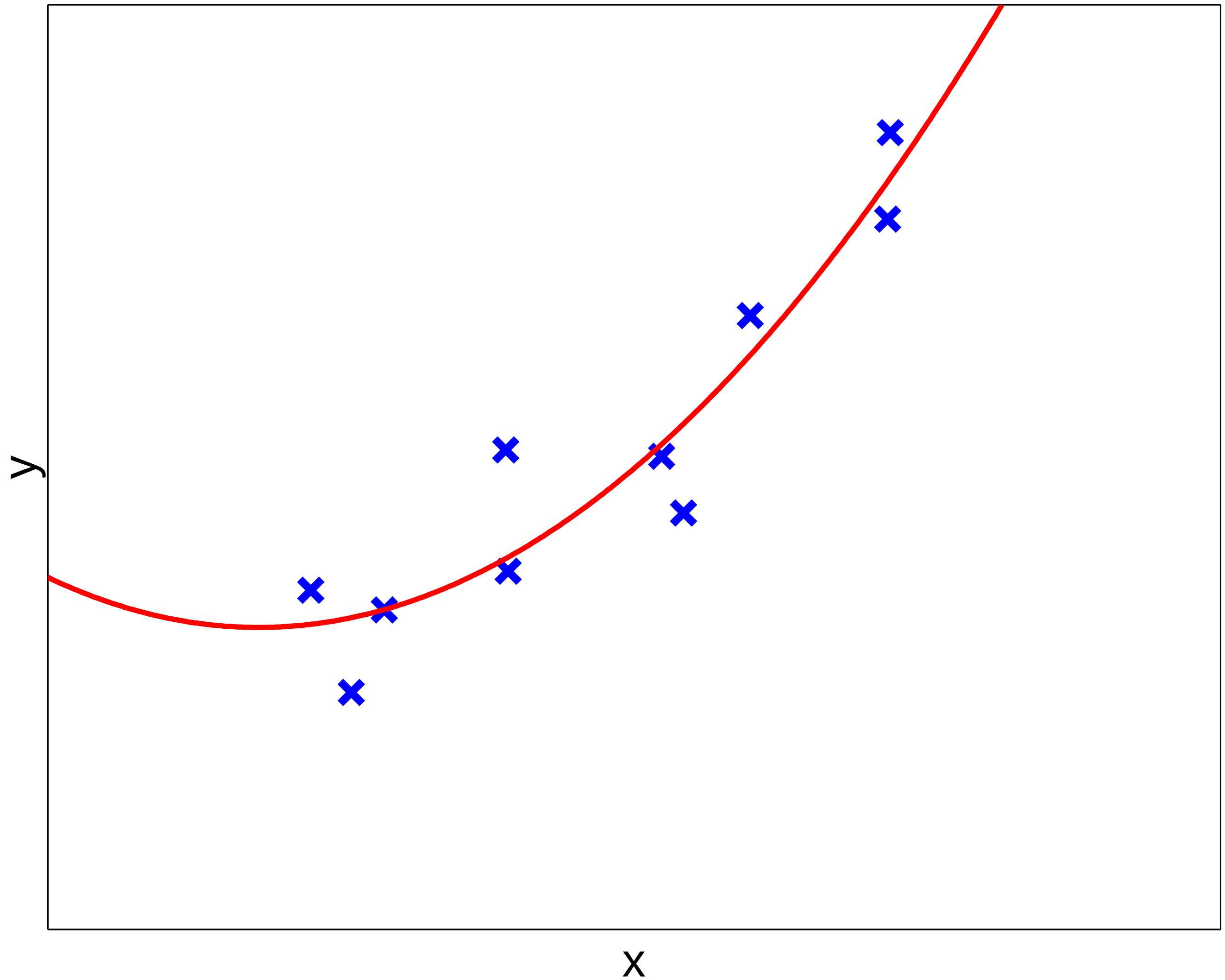
$$= \begin{bmatrix} 3.60 \\ 6.49 \\ 8.34 \end{bmatrix}$$

Solving for w

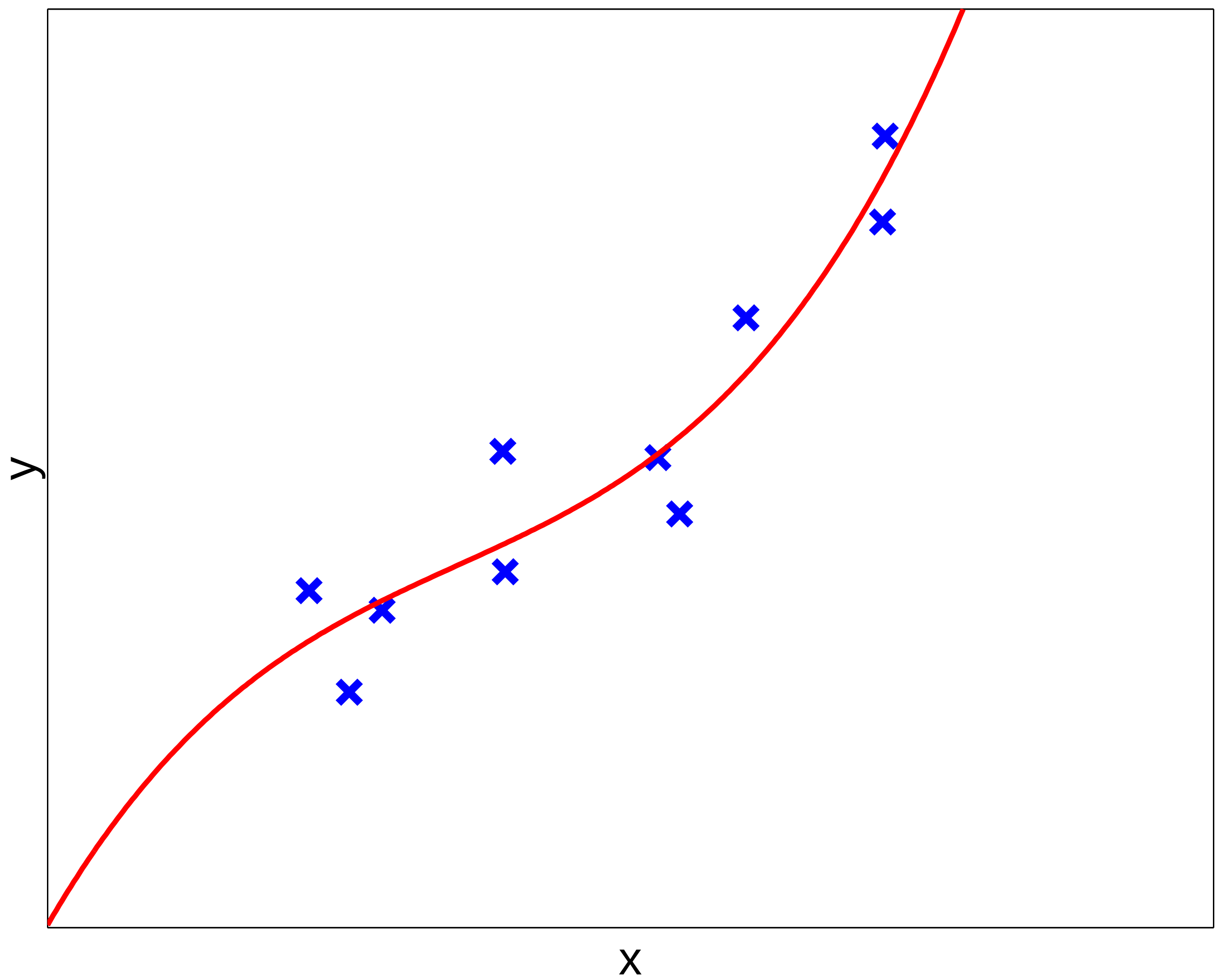
$$w = (X^T X)^{-1} X^Y = \begin{bmatrix} 4.11 & -1.64 & 4.95 \\ -1.64 & 4.95 & -1.39 \\ 4.95 & -1.39 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 3.60 \\ 6.49 \\ 8.34 \end{bmatrix} = \begin{bmatrix} 0.68 \\ 1.74 \\ 0.73 \end{bmatrix}$$

So the best order-2 polynomial is $y = 0.68x^2 + 1.74x + 0.73$.

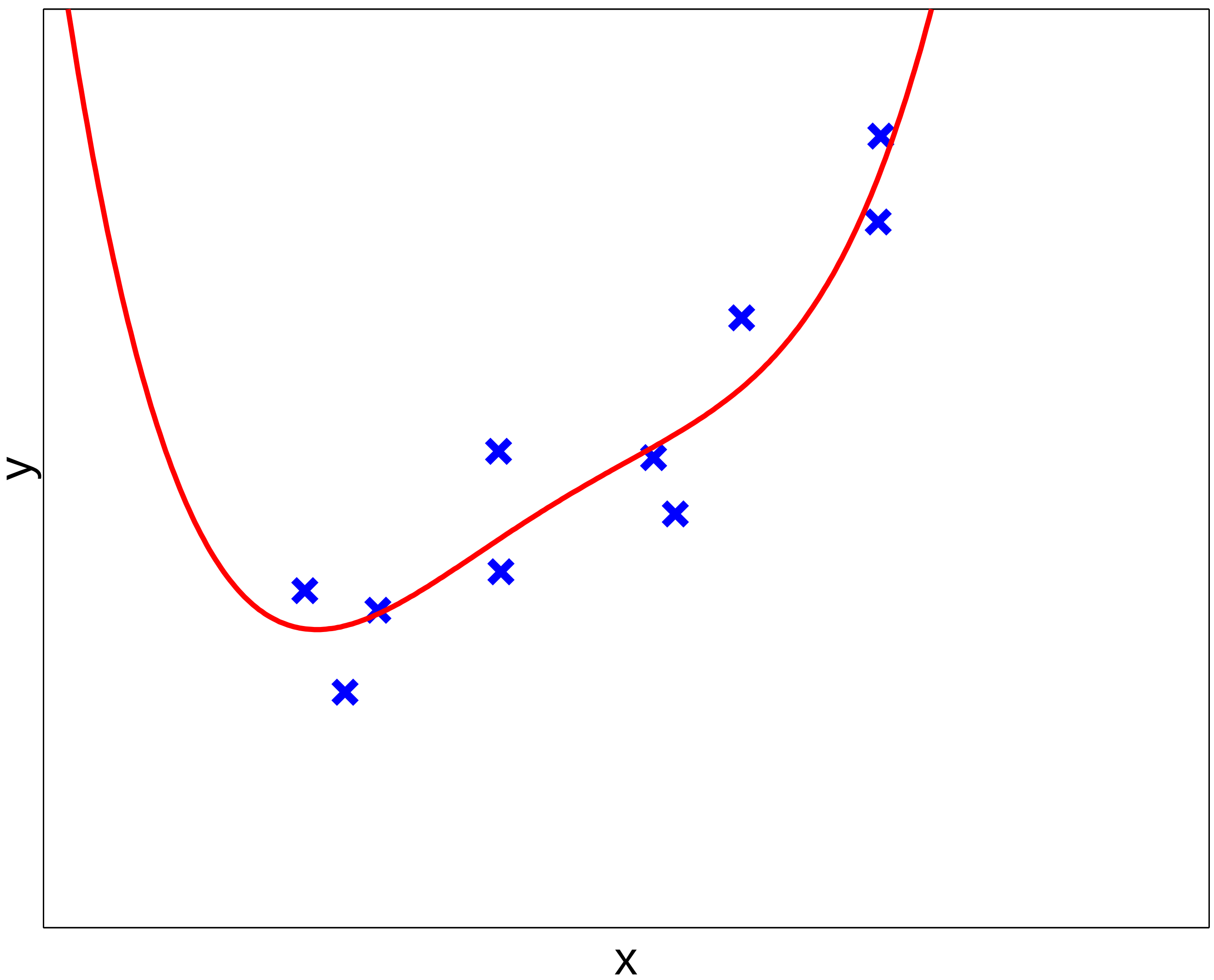
Data and curve $y = 0.68x^2 + 1.74x + 0.73$



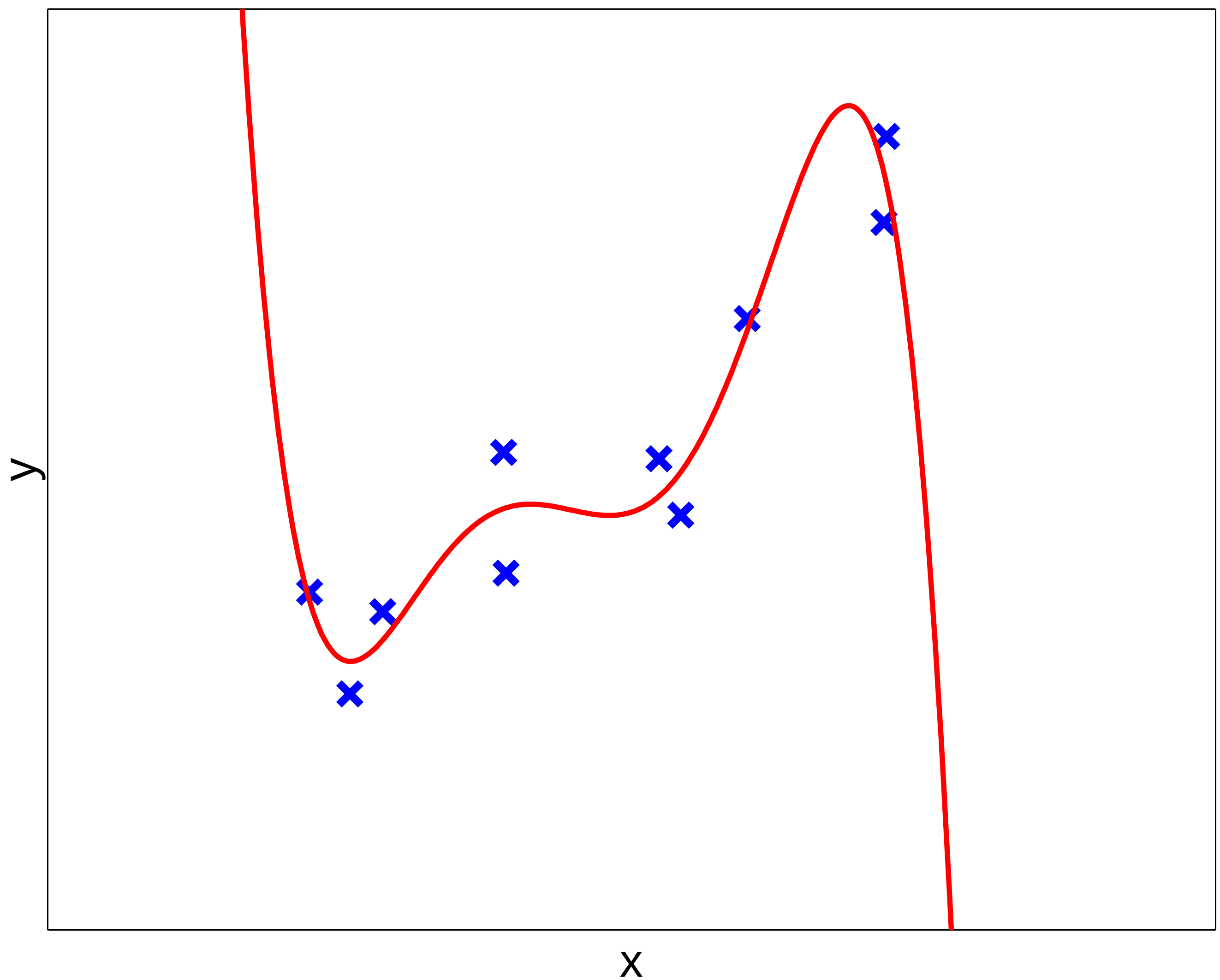
Order-3 fit



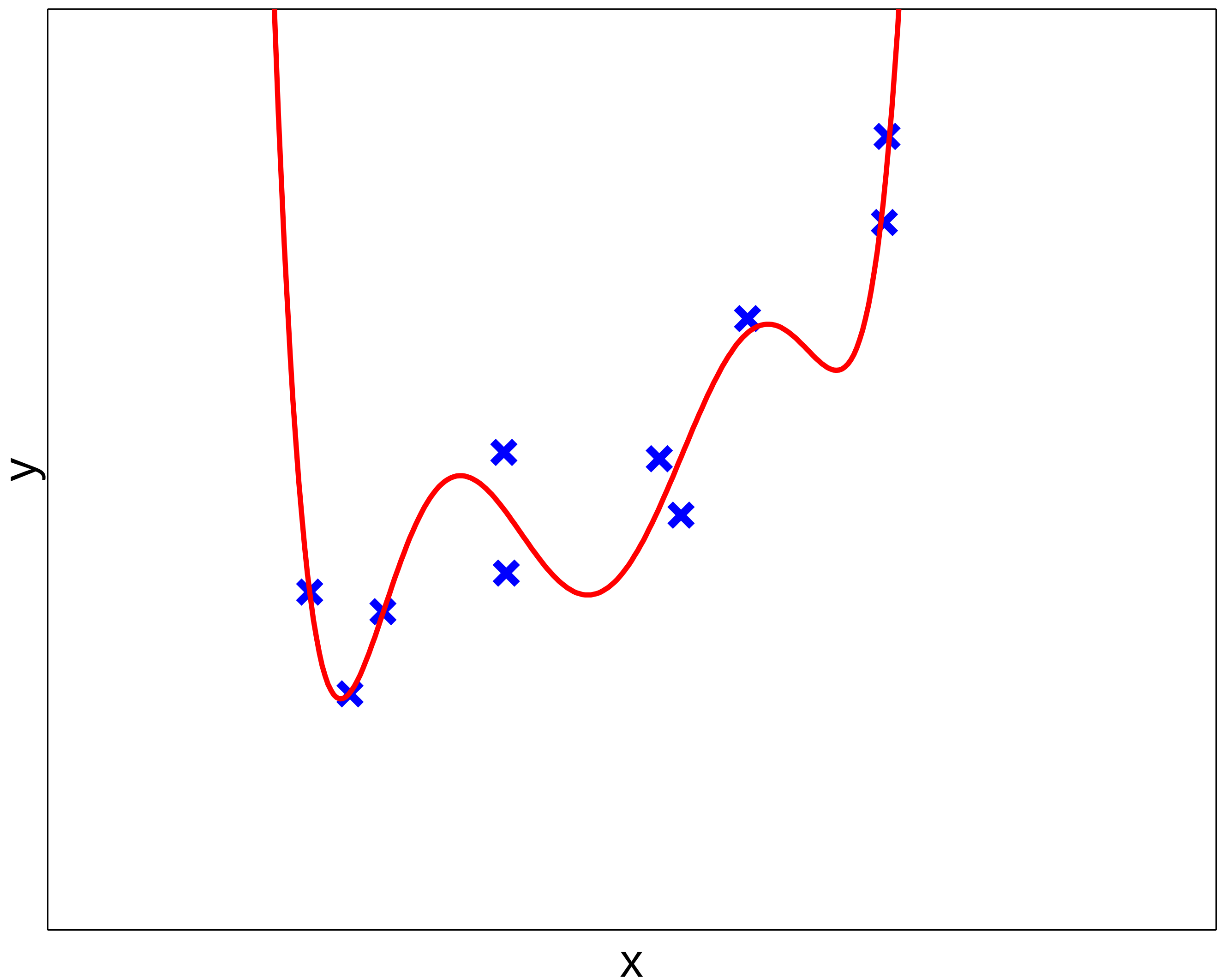
Order-4 fit



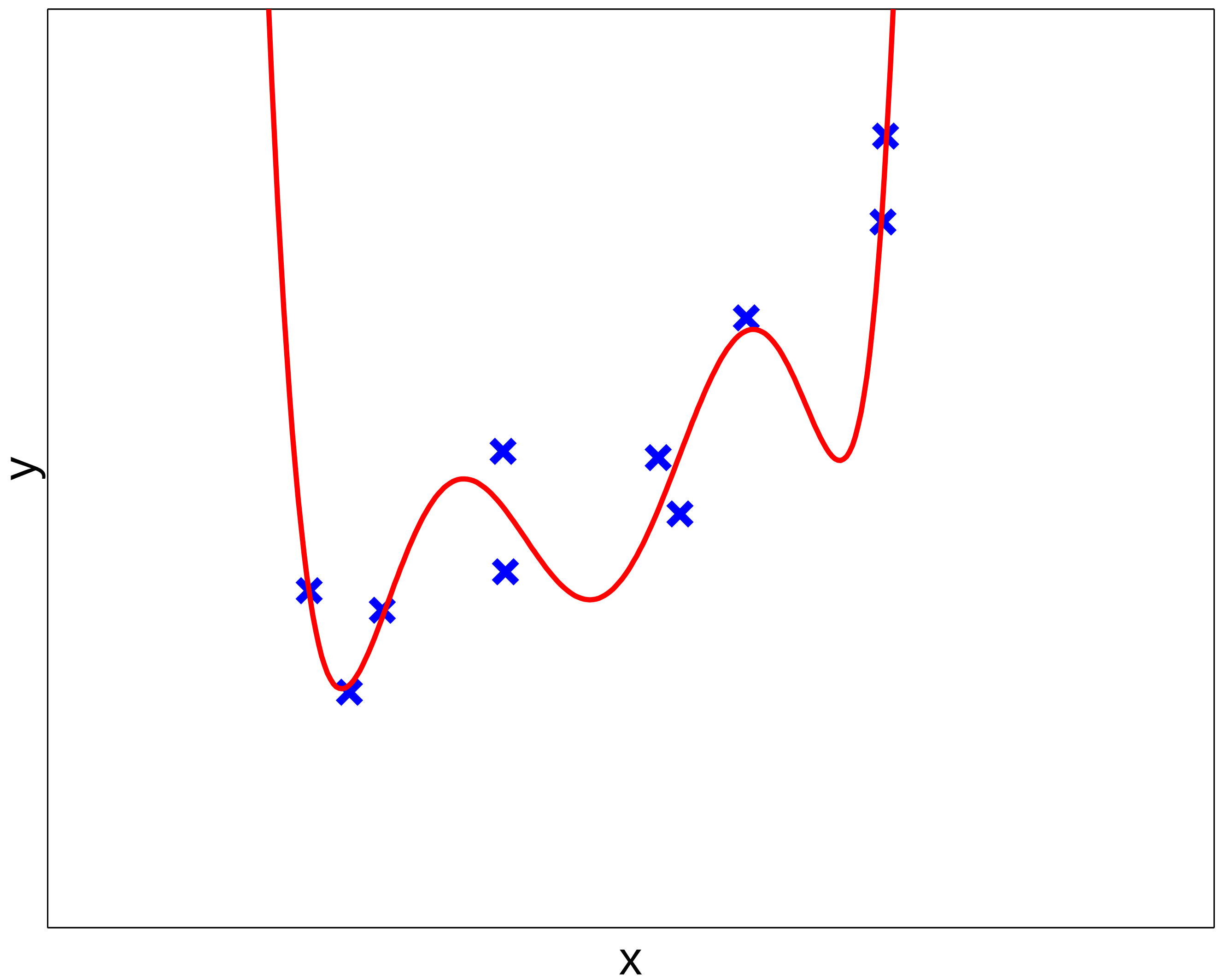
Order-5 fit



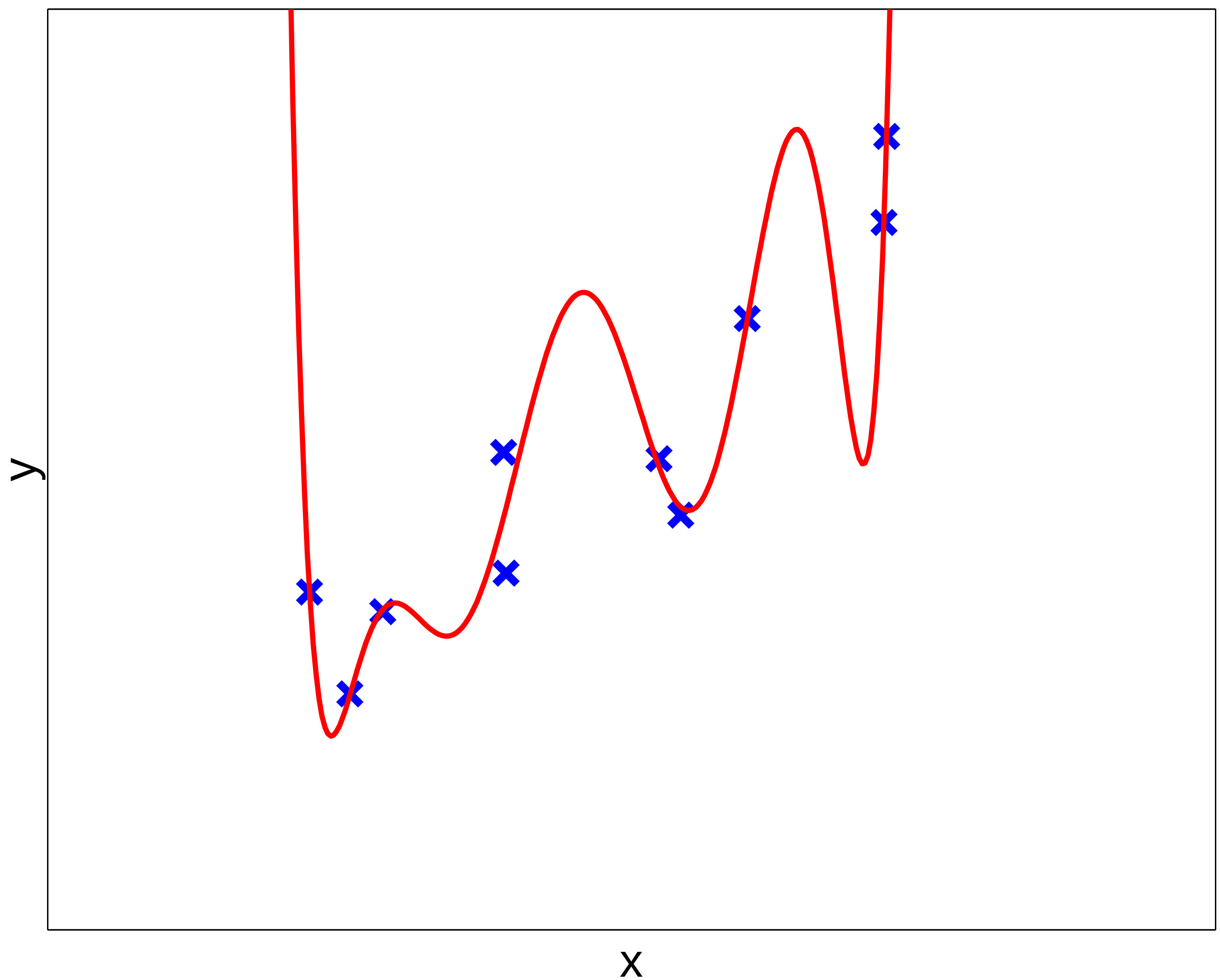
Order-6 fit



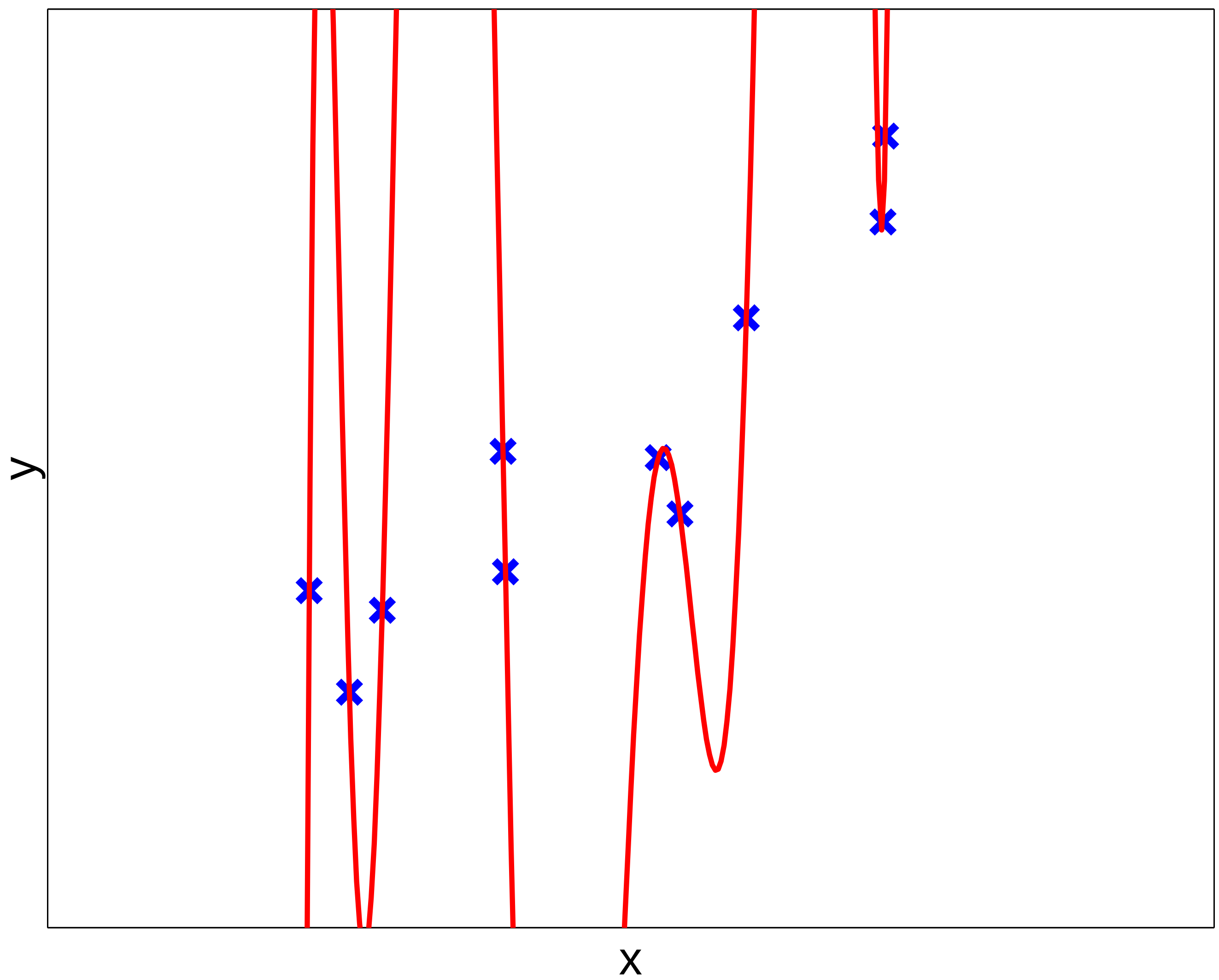
Order-7 fit



Order-8 fit



Order-9 fit



Leave-one-out cross-validation to choose order of polynomial fit

- On the same data, how can we choose the best d for an order-d polynomial fit to the data?
- One answer:
  - Use leave-one-out cross-validation to estimate the true prediction error for the best order-d fit for $d \in \{1, 2, \ldots, 9\}$.
  - Choose the d with lowest estimated true prediction error.

Estimating true error for $d = 1$

$D = \{(0.86, 2.49), (0.09, 0.83), (-0.85, -0.25), (0.87, 3.10), (-0.44, 0.87),$
$(-0.43, 0.02), (-1.10, -0.12), (0.40, 1.81), (-0.96, -0.83), (0.17, 0.43)\}$.

Iter	D_{train}	D_{valid}	$\mathcal{E}_{train}$	$\mathcal{E}_{valid}$
1	$D - \{(0.86, 2.49)\}$	$(0.86, 2.49)$	0.4928	0.0044
2	$D - \{(0.08, 0.83)\}$	$(0.09, 0.83)$	0.1995	0.1869
3	$D - \{(-0.85, -0.25)\}$	$(-0.85, -0.25)$	0.3461	0.0053
4	$D - \{(0.87, 3.10)\}$	$(0.87, 3.10)$	0.3887	0.8681
5	$D - \{(-0.44, 0.87)\}$	$(-0.44, 0.87)$	0.2128	0.3439
6	$D - \{(-0.43, 0.02)\}$	$(-0.43, 0.02)$	0.1996	0.1567
7	$D - \{(-1.10, -0.12)\}$	$(-1.10, -0.12)$	0.5707	0.7205
8	$D - \{(0.40, 1.81)\}$	$(0.40, 1.81)$	0.2661	0.0203
9	$D - \{(-0.96, -0.83)\}$	$(-0.96, -0.83)$	0.3604	0.2033
10	$D - \{(0.17, 0.43)\}$	$(0.17, 0.43)$	0.2138	1.0490
		mean:	0.2188	0.3558

Cross-validation results

d	$\mathcal{E}_{train}$	$\mathcal{E}_{valid}$
1	0.2188	0.3558
2	0.1504	0.3095
3	0.1384	0.4764
4	0.1259	1.1770
5	0.0742	1.2828
6	0.0598	1.3896
7	0.0458	38.819
8	0.0000	6097.5
9	0.0000	6097.5

- Optimal choice: $d = 2$. Overfitting beyond that.
- Why are $d = 8$ and $d = 9$ the same?

Linear and polynomial regression summary

- We can fit linear and polynomial functions in polynomial time by solving $w = (X^T X)^{-1} X^T Y$.
- We can use cross-validation to choose the best order of polynomial to fit our data.
- Issue: How many coefficients does an order-d polynomial have if there are two input variables? m input variables?
  - Often, one will use powers of individual input variables but no cross terms, or only select cross-terms (based on domain knowledge).
- The inverse $(X^T X)^{-1}$ may not exist if we have too few samples and/or try to fit too many parameters. Dimensionality reduction or "regularization" can solve this problem.