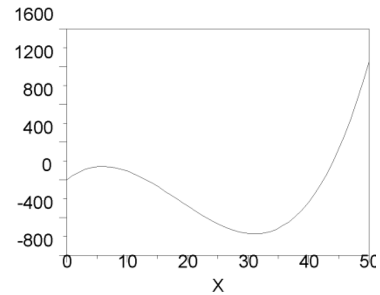
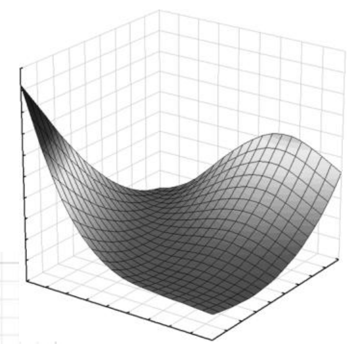
3

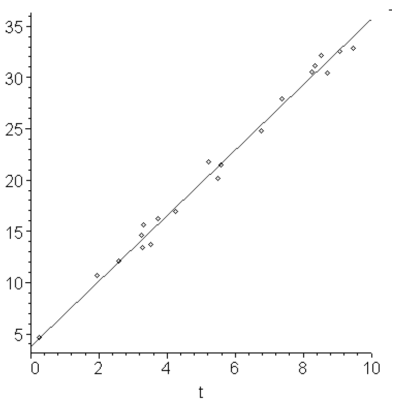
Linear Regression
Least Squares Method
Simple model evaluation

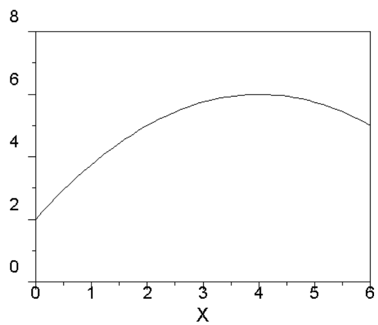
Linear regression, polynomials

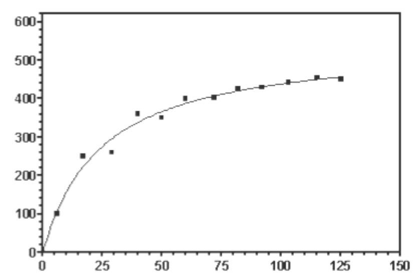
- Linear regression works not only with a line/plane:
  - 1st degree polynomial (straight line) $\hat{y} = a_0 + a_1 x$
  - 2nd degree polynomial (parabola) $\hat{y} = a_0 + a_1 x + a_2 x^2$
  - Polynomial of degree d $\hat{y} = a_0 + a_1 x + \ldots + a_2 x^d$
  - Other functions

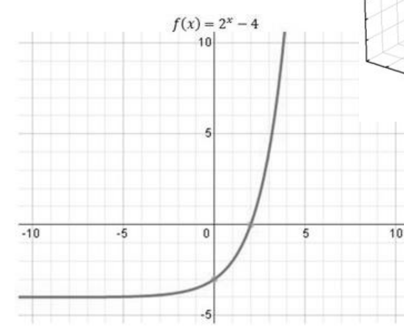


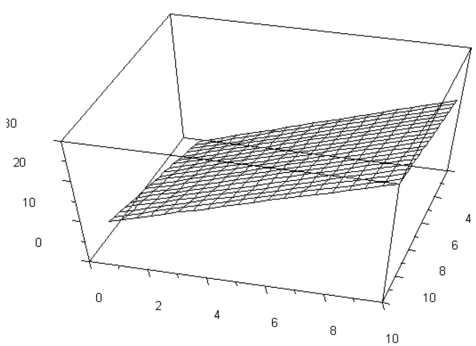


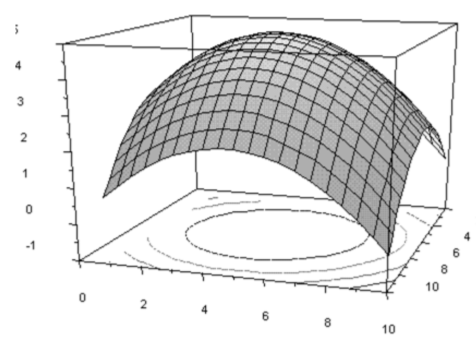


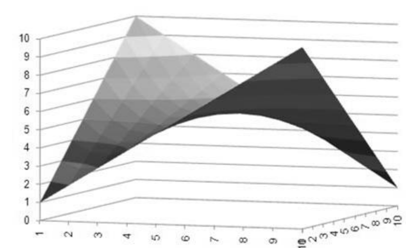


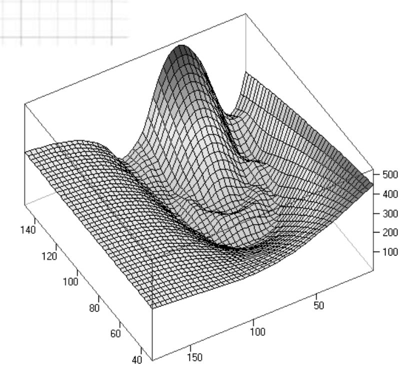










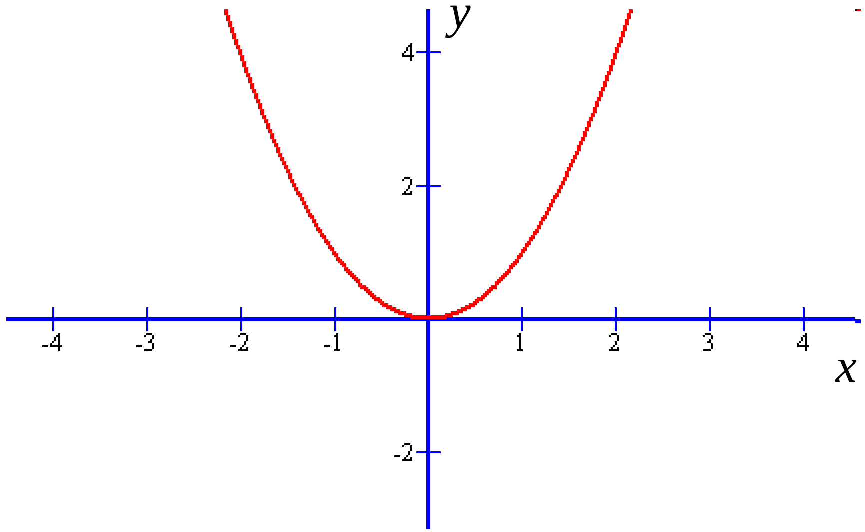


Quadratic polynomial ($d = 2$)

- Equation for parabola

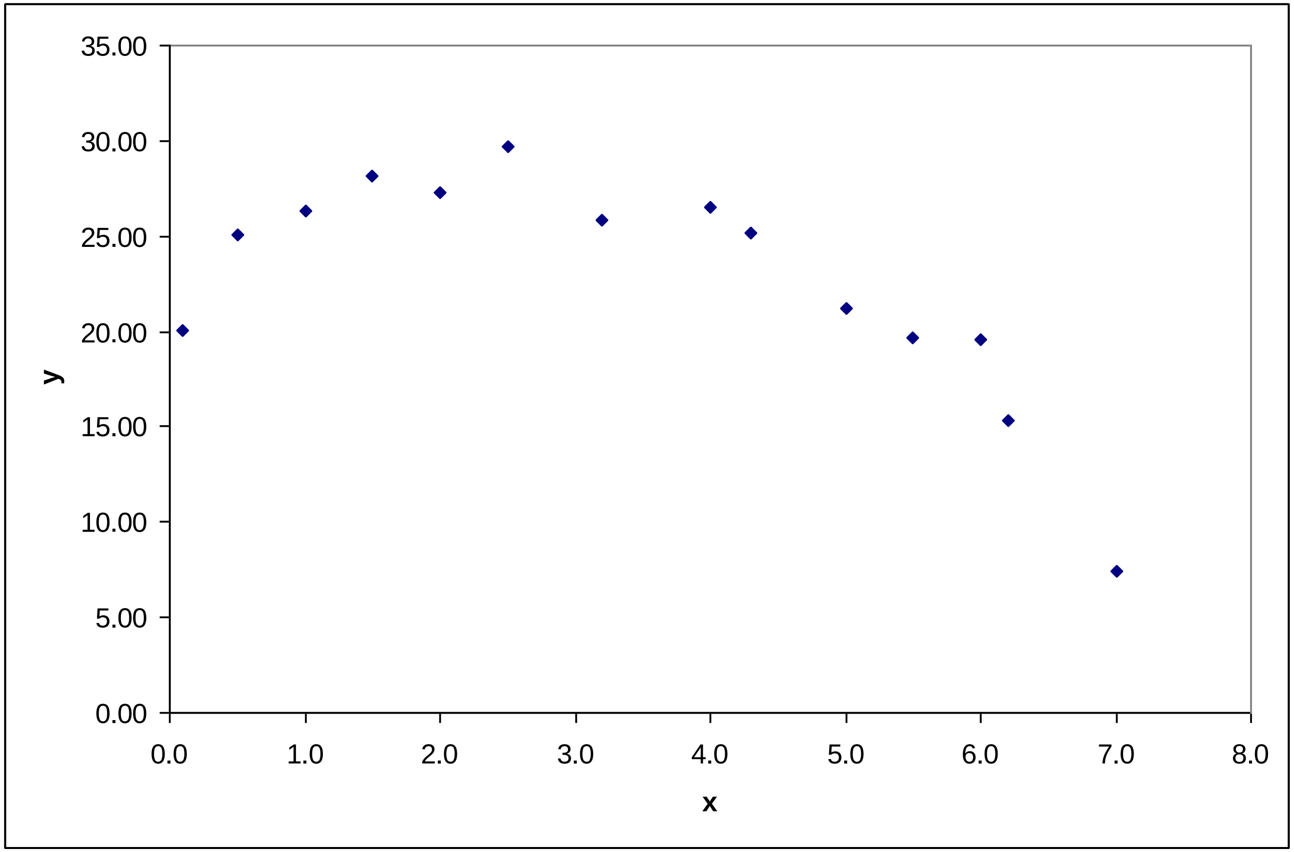
$$\hat{y} = a_0 + a_1 x + a_2 x^2$$

  - when $a_2 \to 0$, shape of parabola approaches straight line



Dataset

x *Marketing expenses*	y *Profit*
2.0	27.33
1.5	28.20
4.0	26.54
5.0	21.24
1.0	26.35
3.2	25.88
6.0	19.62
2.5	29.69
0.5	25.10
4.3	25.14
7.0	7.41
0.1	20.10
5.5	19.63
6.2	15.36



$n = 14$

$$\hat{y} = a_0 + a_1 x + a_2 x^2$$

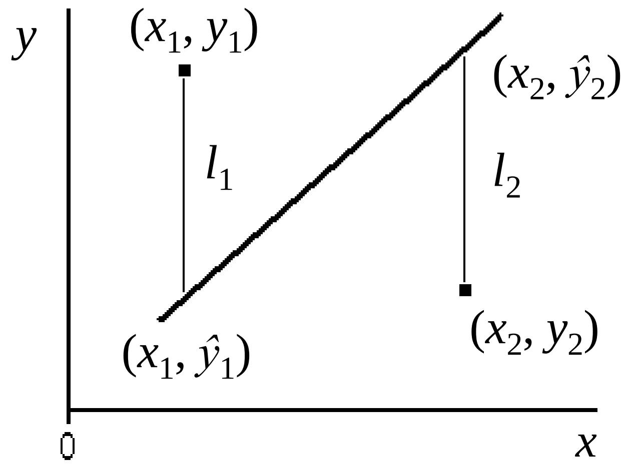
$a_0 = ?$
$a_1 = ?$
$a_2 = ?$

The same RSS criterion

- We minimize *Residual Sum of Squares (RSS)*

$$S = \sum_{i=1}^{n} (l_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min$$

$$l_i = y_i - \hat{y}_i$$



- Find parameters

$$a_0 = ?$$
$$a_1 = ?$$
$$a_2 = ?$$

$$\hat{y} = a_0 + a_1 x + a_2 x^2$$

- Solve in the same way as for the straight line in the previous lecture

Least Squares Method

Model

$$\hat{y} = a_0 + a_1 x + a_2 x^2$$

Criterion

$$S = \sum_{i=1}^{n} (l_i)^2 \to \min$$

Data

x	y
2.0	27.33
1.5	28.20
4.0	26.54
5.0	21.24
1.0	26.35
3.2	25.88
6.0	19.62
2.5	29.69
0.5	25.10
4.3	25.14
7.0	7.41
0.1	20.10
5.5	19.63
6.2	15.36

$n = 14$

- Replace l_i with $(y_i - \hat{y}_i)$

$$S = \sum_{i=1}^{n} (l_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Replace $\hat{y}_i$ with our equation's left side

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

- So we have to minimize this:

$$S = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2 \to \min$$

Partial derivatives

$$S = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2 \to \min$$

- Partial derivatives for each parameter

$$\frac{\partial S}{\partial a_0} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

$$\frac{\partial S}{\partial a_1} = -2\sum_{i=1}^{n} x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

$$\frac{\partial S}{\partial a_2} = -2\sum_{i=1}^{n} x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

- We want it to be zero as there is the minimum

$$\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

$$\sum_{i=1}^{n} x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

System of equations

- We get linear system of equations:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i \\ a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i \end{cases}$$

x	y
2.0	27.33
1.5	28.20
4.0	26.54
5.0	21.24
1.0	26.35
3.2	25.88
6.0	19.62
2.5	29.69
0.5	25.10
4.3	25.14
7.0	7.41
0.1	20.10
5.5	19.63
6.2	15.36

$n = 14$

- Insert numbers from the table

- Solve using, e.g., Gaussian elimination

Estimated parameters

Model

$$\hat{y} = a_0 + a_1 x + a_2 x^2$$

Criterion

$$S = \sum_{i=1}^{n} (l_i)^2 \rightarrow \min$$

- **Parameters:**

$$a_0 = 21.379$$

$$a_1 = 5.3613$$

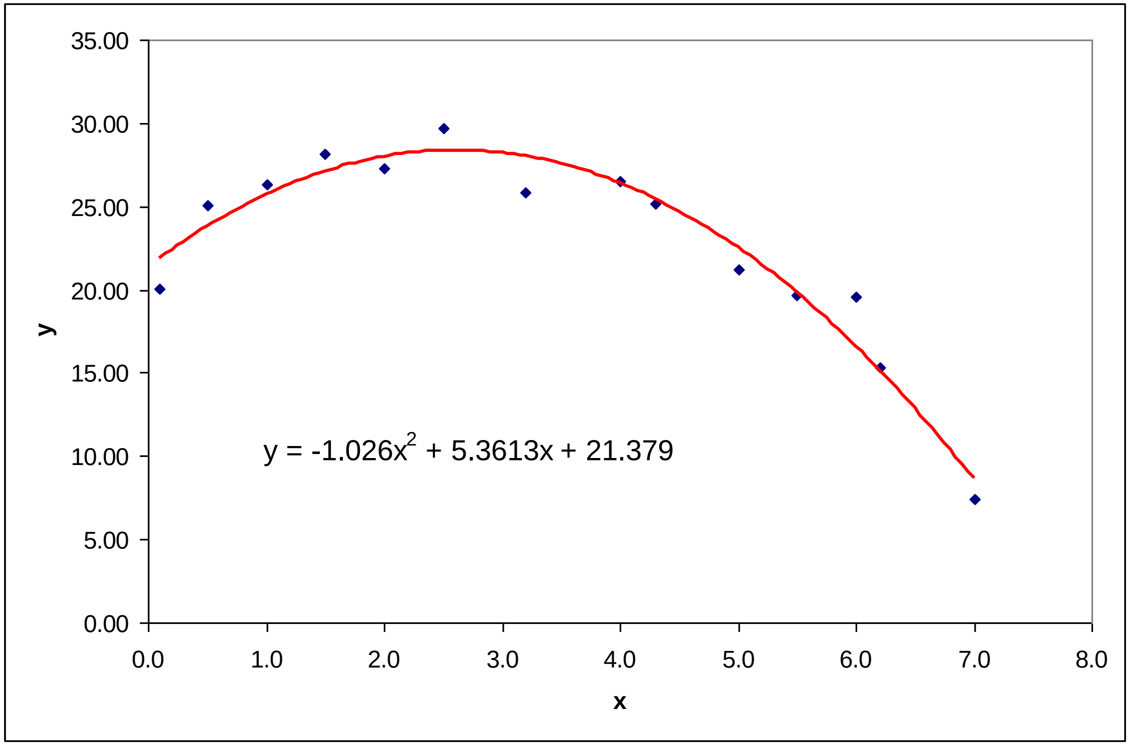
$$a_2 = -1.026$$

- **Usage:**

If $x = 3$ then

$\hat{y} = 21.379 + 5.3613*3 - 1.026*9 =$

$= 28.2$



Again: Linear model

- Any linear model can be written as a sum of features:

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m \qquad \text{or} \qquad \hat{y} = a_0 + \sum_{j=1}^{m} a_j x_j$$

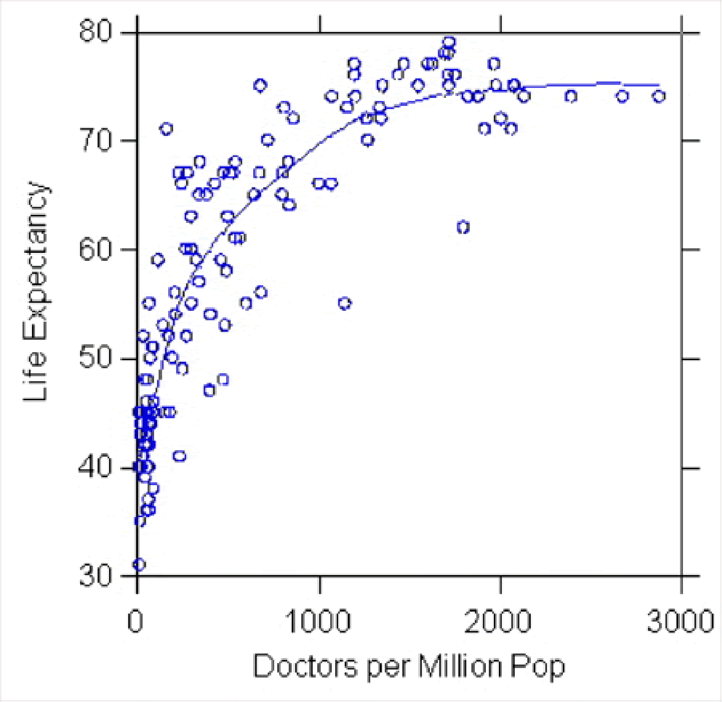
where x_j is jth feature, m is number of features, $k = m + 1$ is number of parameters. Or, by assuming $x_0 = 1$:

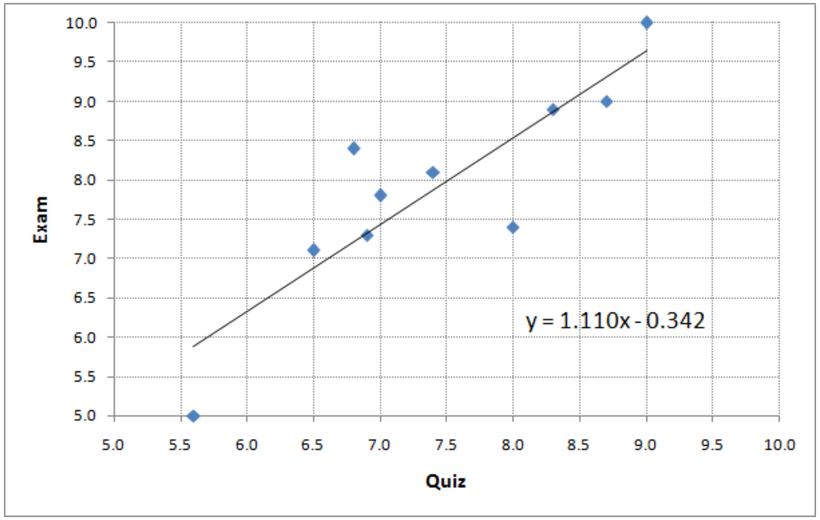
$$\hat{y} = a_0 x_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m \qquad \text{or} \qquad \hat{y} = \sum_{j=0}^{m} a_j x_j$$

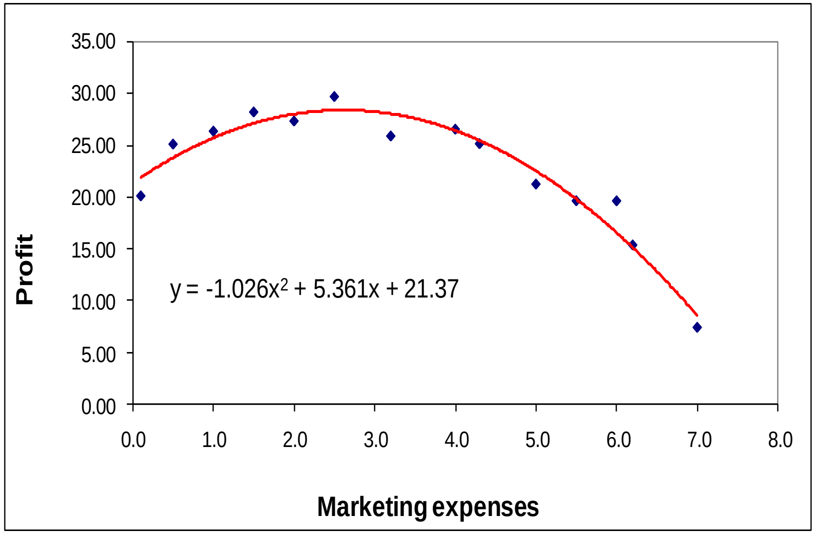
- But what if we need something more complex than just a sum of features?
  - This too can be achieved using linear regression

Feature transformation

- Why would we need to transform input variables?
  - EUR^2
  - $\log(number\_of\_doctors)$
  - ...







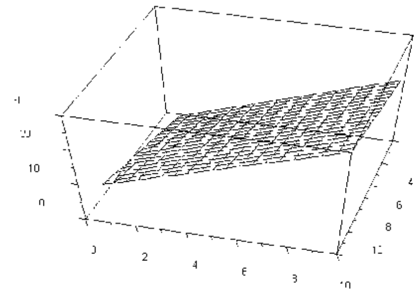
Feature transformation

- We can use our m features to make any needed feature transformations
  - *(even if we don't know which ones are needed – we will talk about that in another lecture)*

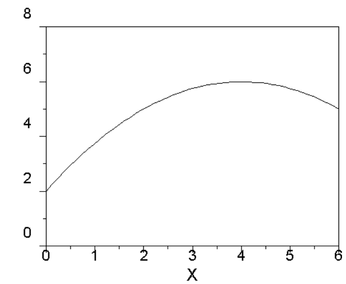
- Simply "synthetically" increase the number of features using mathematical functions that transform our original features. For instance:
  - Power (create polynomials – this is the most often used type)
  - Log
  - Exp
  - Sin
  - Cos
  - Min
  - Max
  - etc.

Examples

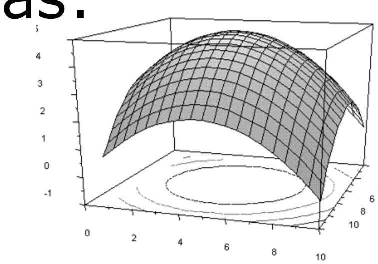
- Model $\hat{y} = a_0 + a_1x_1 + a_2x_2$ can be written as:
  $$\hat{y} = a_0\, x_0 + a_1x_1 + a_2x_2$$
  by defining: $x_0 = 1$



- Model $\hat{y} = a_0 + a_1x_1 + a_2x_1^2$ can be written as:
  $$\hat{y} = a_0\, x_0 + a_1x_1 + a_2x_2$$
  by defining: $x_0 = 1$, $x_2 = x_1^2$

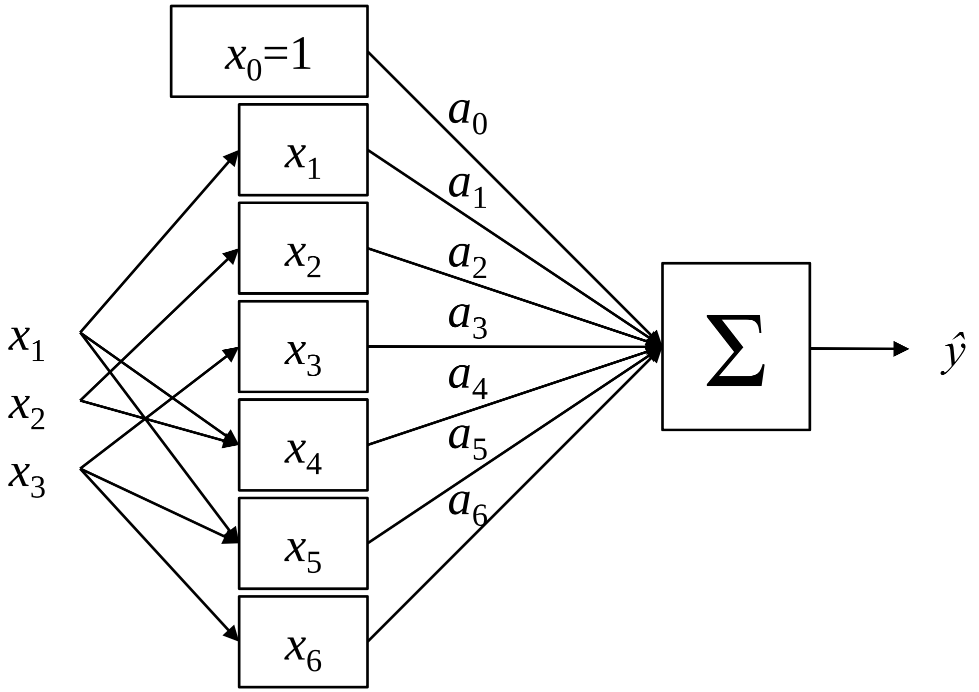


- Model $\hat{y} = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4\log(x_2) + a_5x_1x_2$ as:
  $$\hat{y} = a_0\, x_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5$$
  by defining: $x_0 = 1$, $x_3 = x_1^2$, $x_4 = \log(x_2)$, $x_5 = x_1x_2$



- If such redefinition of a model is not possible, we are talking about a non-linear model, for example:
  $$\hat{y} = a_0 + a_1\sin(a_2x_1) + x_2^{a_3}$$

Linear model



$$\hat{y} = \sum_{j=0}^{m} a_j x_j$$

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_1 x_2 + a_5 x_1 x_3 + a_6 x_3^2$$

$$\hat{y} = a_0\, x_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + a_5 x_5 + a_6 x_6$$

Least Squares Method generalized

Model: $\hat{y} = a_0 x_0 + a_1 x_1 + \ldots + a_m x_m$

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \min$$

$$S = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - a_0 x_{i0} - a_1 x_{i1} - \ldots - a_m x_{im}\right)^2 \rightarrow \min$$

$$\frac{\partial S}{\partial a_j} = (-2) \sum_{i=1}^{n} x_i (y_i - a_0 x_{i0} - a_1 x_{i1} - \ldots - a_m x_{im}) \qquad j = 0,1,\ldots,m$$

$$\begin{cases} a_0\left(\sum x_{i0}x_{i0}\right) + a_1\left(\sum x_{i1}x_{i0}\right) + \cdots + a_m\left(\sum x_{im}x_{i0}\right) = \sum x_{i0}y_i \\ a_0\left(\sum x_{i0}x_{i1}\right) + a_1\left(\sum x_{i1}x_{i1}\right) + \cdots + a_m\left(\sum x_{im}x_{i1}\right) = \sum x_{i1}y_i \\ \vdots \\ a_0\left(\sum x_{i0}x_{im}\right) + a_1\left(\sum x_{i1}x_{im}\right) + \cdots + a_m\left(\sum x_{im}x_{im}\right) = \sum x_{im}y_i \end{cases}$$

$$\begin{bmatrix} \sum x_{i0}x_{i0} & \sum x_{i1}x_{i0} & \cdots & \sum x_{im}x_{i0} \\ \sum x_{i0}x_{i1} & \sum x_{i1}x_{i1} & \cdots & \sum x_{im}x_{i1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{i0}x_{im} & \sum x_{i1}x_{im} & \vdots & \sum x_{im}x_{im} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum x_{i0}y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{im}y_i \end{bmatrix}$$

← System of equations to solve

Solving in matrix form

$$\underbrace{\begin{bmatrix} \sum x_{i0}x_{i0} & \sum x_{i1}x_{i0} & \cdots & \sum x_{im}x_{i0} \\ \sum x_{i0}x_{i1} & \sum x_{i1}x_{i1} & \cdots & \sum x_{im}x_{i1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{i0}x_{im} & \sum x_{i1}x_{im} & \vdots & \sum x_{im}x_{im} \end{bmatrix}}_{\mathbf{X}^T\mathbf{X}} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}}_{\mathbf{a}} = \underbrace{\begin{bmatrix} \sum x_{i0}y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{im}y_i \end{bmatrix}}_{\mathbf{X}^T\mathbf{y}}$$

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_{1,0} & x_{1,1} & \cdots & x_{1,m} \\ x_{2,0} & x_{2,1} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,0} & x_{n,1} & \cdots & x_{n,m} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

So we have this: $\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{X}^T\mathbf{y}$

And solution is this: $\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

Data

$\mathbf{x}_0$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{y}$
1	5.6	6.0	5.0
1	6.5	7.0	7.1
1	6.8	7.2	8.4
1	6.9	6.8	7.3
1	7.0	7.2	7.8
1	7.4	8.5	8.1
1	8.0	6.5	7.4
1	8.3	7.9	8.9
1	8.7	7.3	9.0
1	9.0	9.1	10.0

$\mathbf{x}_0=1$ is introduced to provide the a_0 parameter

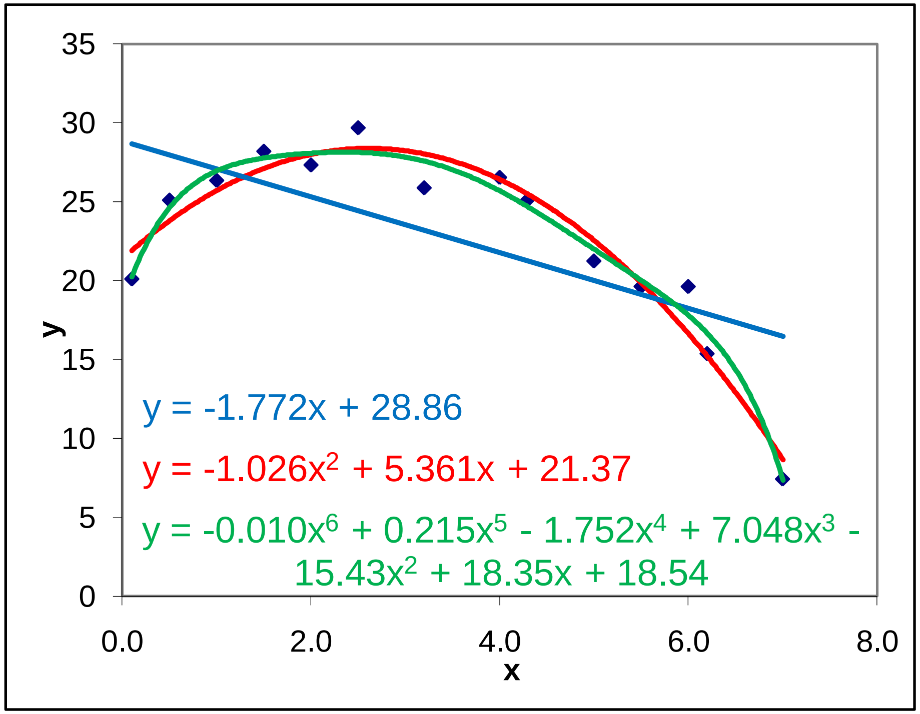
^{T} is transpose $\quad$ $^{-1}$ is inverse

Let’s begin our next topic:

Model evaluation

Regression model evaluation criteria

- "Looks good", but how exactly "good" is the model?
  - And if we have more than one model then which one should we choose?
- The simplest way – compute prediction error using the same training data



Sum of Absolute Error, SAE

$$SAE = \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

Mean Absolute Error, MAE

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

Model evaluation criteria

Sum of Squared Error, SSE

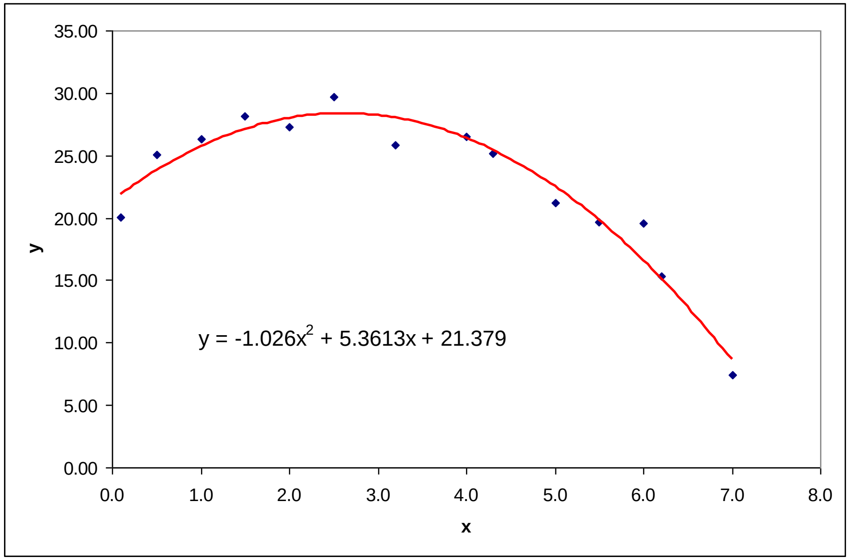
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Mean Squared Error, MSE

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Root Mean Squared Error, RMSE

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$



x	y
2.0	27.33
1.5	28.20
4.0	26.54
5.0	21.24
1.0	26.35
3.2	25.88
6.0	19.62
2.5	29.69
0.5	25.10
4.3	25.14
7.0	7.41
0.1	20.10
5.5	19.63
6.2	15.36

For our model:

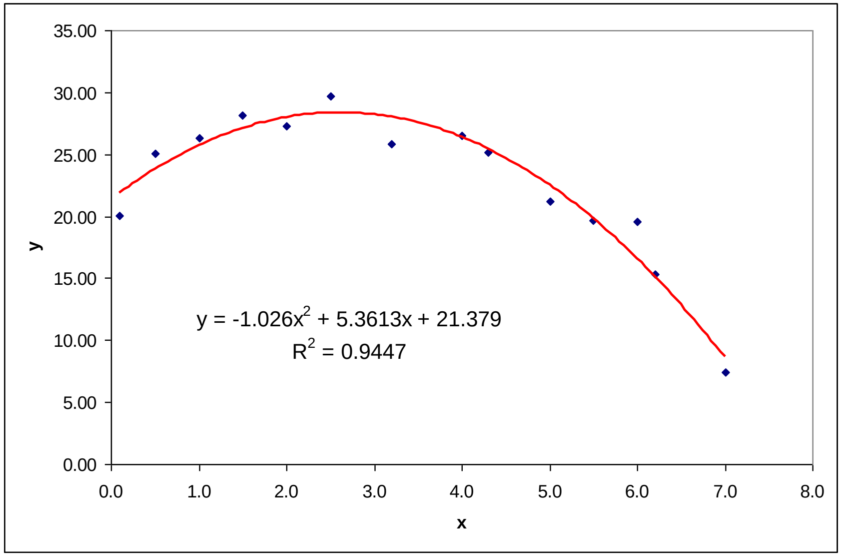
$$SSE = 25.77$$
$$MSE = 25.77 / n = 1.84$$
$$RMSE = \sqrt{1.84} = 1.36$$

All these criteria depend on the units of y *and their interpretation depends on what is* y *and on the specific problem at hand.*
They don't have universally consistent range.

Coefficient of determination, R^2

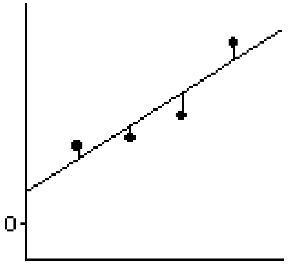
- R^2 is independent from data range – it normalizes the quadratic error rescaling it between 0 and 1

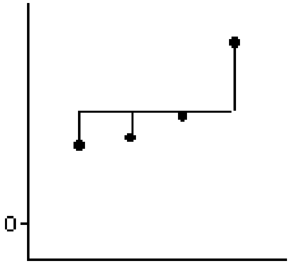
$$R^2 = 1 - \frac{SSE}{SSE_{tot}}$$



x	y
2.0	27.33
1.5	28.20
4.0	26.54
5.0	21.24
1.0	26.35
3.2	25.88
6.0	19.62
2.5	29.69
0.5	25.10
4.3	25.14
7.0	7.41
0.1	20.10
5.5	19.63
6.2	15.36

where SSE_{tot} is "total sum of squares" is variance of y around it's mean. So R^2 is 1 – ratio between variance that the model can't explain (SSE) and variance that exists in the data (SSE_{tot}).



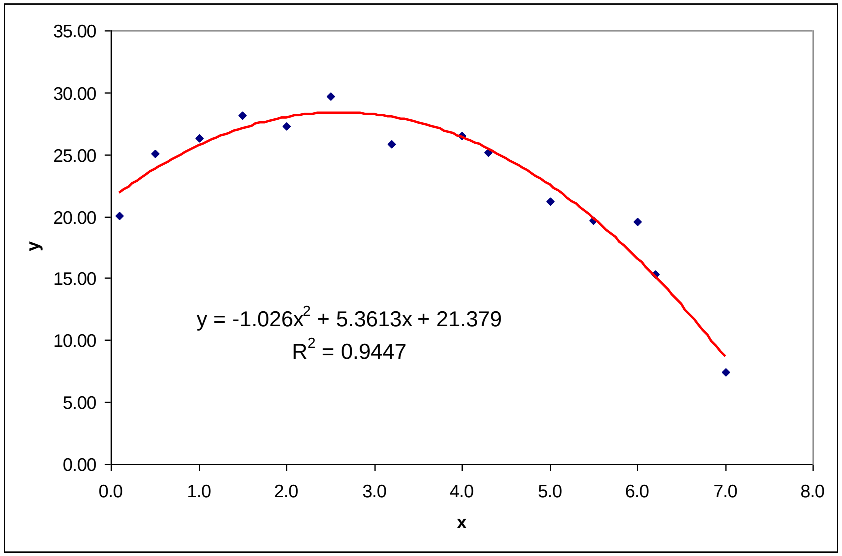


$$SSE_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

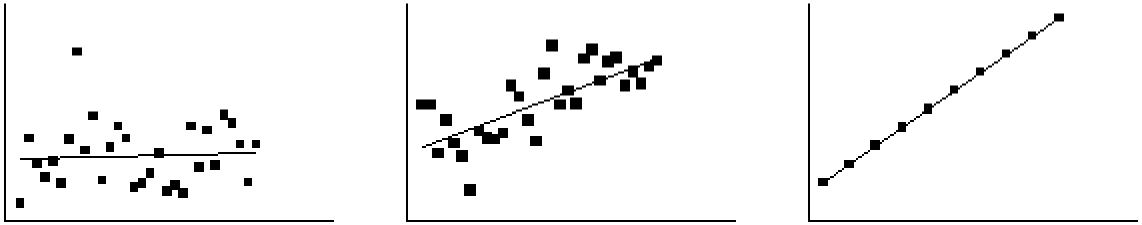
$\bar{y}$ is mean of $\mathbf{y}$

$SSE = 25.77$ $SSE_{tot} = 465.90$ $R^2 = 0.9447$

Interpreting R^2



- $R^2 = 1$ means that the model explains 100% of y variance (it perfectly fits all the data points)
- $R^2 = 0.5$ means that the model explains 50% of y variance
- $R^2 = 0$ means that the model doesn't explain anything useful about the data
- $R^2 < 0$ means that the model is worse than a simple mean of y. It is completely unsuitable for the data (such situations can occur when for example R^2 is computed on separate data other than training data)



$R^2 = 0$ $\quad$ $R^2 = 0.5$ $\quad$ $R^2 = 1.0$