Applied Data Science
fall 2017
Session 4: Multi-variate linear regression

Instructor: Prof. Stanislav Sobolevsky
Course Assistants: Tushar Ahuja, Maxim Temnogorod

Bi-variate to multi-variate

$$y \sim x$$

$$y = w_1 x + w_0 + \varepsilon$$

$$x = (x_1, x_2, \ldots, x_n)$$

$$y = \sum_{j=1}^{n} w_j x_j + w_0 + \varepsilon$$

Intercept: $x_{n+1} = 1, w_{n+1} = w_0$

$$y = \sum_{j=1}^{n+1} w_j x_j + \varepsilon$$

$$y = w^T x + \varepsilon$$

Matrix form

$$y = x_1 + 2x_2 + 3\,x_3$$

$$x_3 = 1$$

$$w_1 = 1,\ w_2 = 2,\ w_3 = 3$$

$$y = w_1x_1 + w_2x_2 + w_3x_3 = \begin{pmatrix} w_1 & w_2 & w_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = w^T x$$

Matrix form

$$y = x_1 + 2x_2 + 3$$

$$x_1 = 1, x_2 = 1, y = 6.2$$
$$x_1 = 2, x_2 = 0, y = 4.9$$
$$x_1 = 3, x_2 = -1, y = 4.1$$
$$x_1 = 4, x_2 = -2, y = 2.9$$

$$x_3 = 1$$

$$w = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

$$Y = \begin{pmatrix} 6.2 \\ 4.9 \\ 4.1 \\ 2.9 \end{pmatrix} = Xw + E$$

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \\ 3 & -1 & 1 \\ 4 & -2 & 1 \end{pmatrix}$$

$$E = Y - Xw = \begin{pmatrix} 0.2 \\ -0.1 \\ 0.1 \\ -0.1 \end{pmatrix}$$

Least-square estimate

$$y = w^T x + \varepsilon$$

$$X = \{(x_j^i), j = 1..n, i = 1..N\}, Y = \{(y^i), i = 1..N\}$$

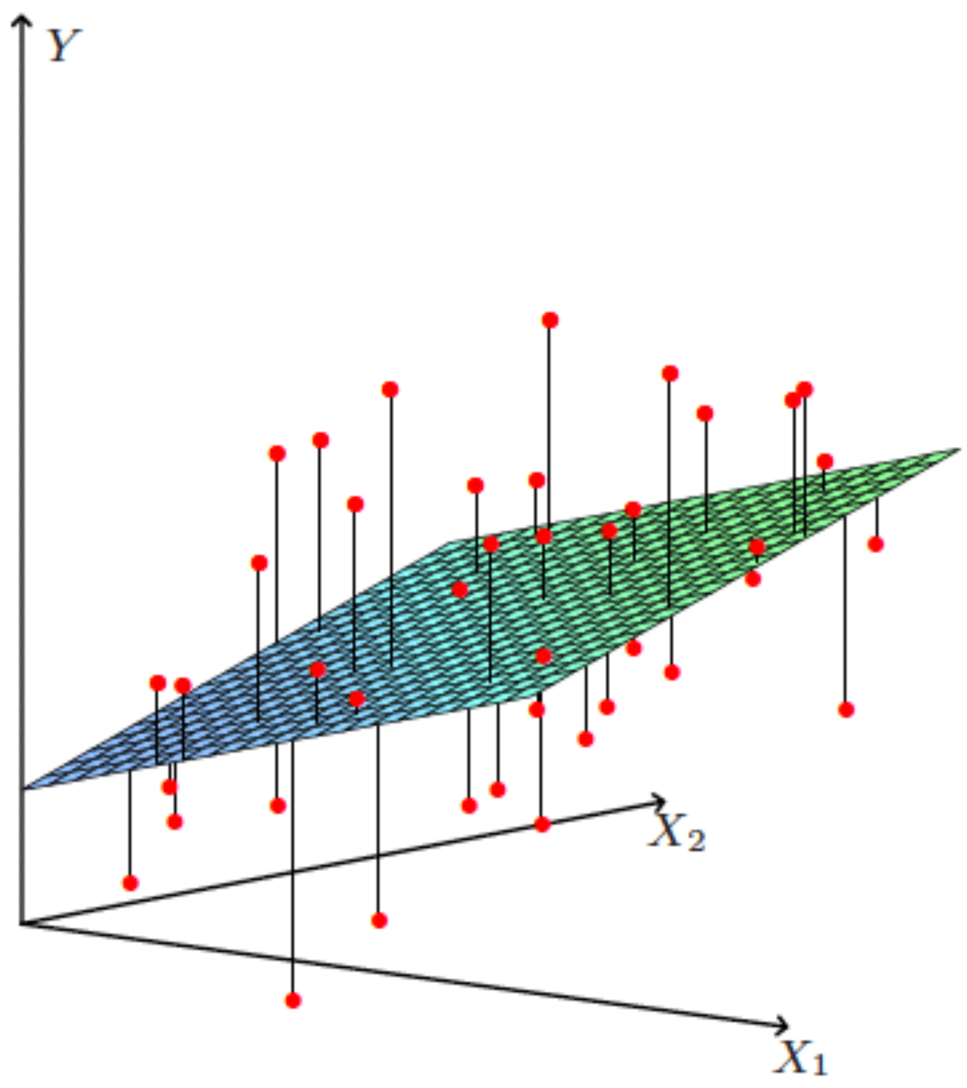
$$E = (\varepsilon_i, i = 1..N)^T = Y - Xw$$

$$RSS(w) = \sum_i \varepsilon_i^2 = \sum_i (y^i - w^T x^i)^2$$

$$RSS(w) = E^T E$$

$$RSS(w) = (Y - Xw)^T (Y - Xw)$$

$$\hat{w} = argmin_w RSS(w)$$



Hastie, *et al.*, The Elements of Statistical Learning, Data Mining, Inference and Prediction, 2nd Edition, Springer. (Free: http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf)]

Least-square estimate

$$RSS(w) = (Y - Xw)^T(Y - Xw) \qquad \hat{w} = argmin_w RSS(w)$$

$$0 = \frac{\partial RSS(\hat{w})}{dw} = -2X^T(Y - X\hat{w})$$

$$X^T Y = (X^T X)\hat{w}$$

$$\hat{w} = (X^T X)^{-1} X^T Y$$

Least-squares: geometric sense

$$y = \sum_j w_j x_j$$

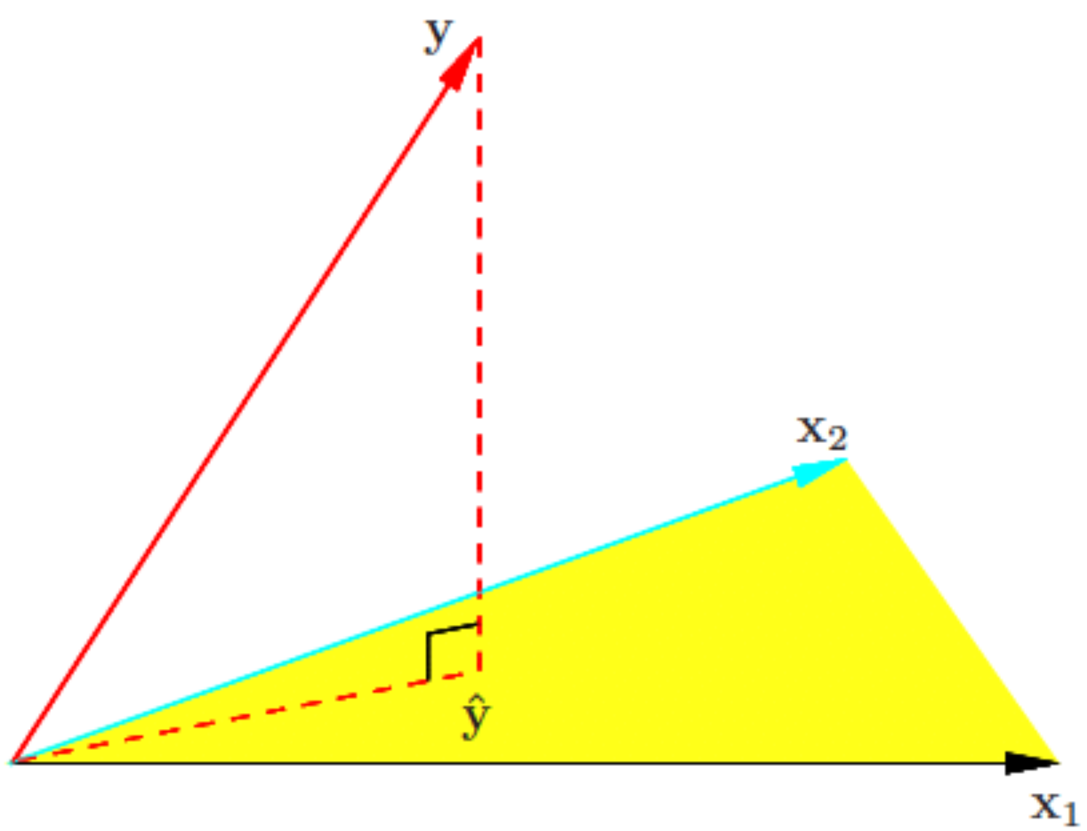
$$Y \sim \sum_j w_j X_j$$

$$\hat{Y} = X\hat{w} = X(X^TX)^{-1}X^TY$$

$$H = X(X^TX)^{-1}X^T$$

$$\hat{Y} = HY$$

$$\hat{w} = (X^TX)^{-1}X^TY$$



Hastie, *et al.*, The Elements of Statistical Learning, Data Mining, Inference and Prediction, 2nd Edition, Springer. (Free: http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf)

Linear Model - R-squared

$$R^2 = 1 - \frac{RSS}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2},$$

Linear Model - Orthogonal regressors

If you know $y \sim w_j x_j$

do you know $y \sim x = (x_1, x_2, ..., x_n)$?

$$Y = \begin{pmatrix} 5 \\ 4 \\ 3 \end{pmatrix} \quad X = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 1 & -2 \\ 0 & 2 & 0 \end{pmatrix}$$

$$y \sim 2.8x_1$$

$$y \sim 2x_2$$

$$y \sim -\frac{3}{5}x_3$$

$$y \sim 2.8x_1 - 0.6x_3$$

$$y \sim 2.5x_1 + 1.5x_2$$

Linear Model - Orthogonal regressors

If you know $y \sim w_j x_j$

you do know $y \sim x = (x_1, x_2, ..., x_n)$ if

$$< X_j, X_k >= X_j^T X_k = \sum_i x_j^i x_k^i = 0$$

Linear Model - Feature scaling

$$x_j^* = \frac{x_j - \bar{x}_j}{\sigma_j}$$

$$\bar{x}_j = E[X_j] \qquad \sigma_j = std[X_j]$$

Linear Model - Orthogonal regressors

$$< X_j, X_k > = X_j^T X_k = \sum_i x_j^i x_k^i = 0$$

$$E[X_j] = 0$$

$$0 = corr[X_j, X_k] = \frac{Cov[X_j, X_k]}{std[X_j]std[X_k]} = \frac{\frac{< X_j, X_k >}{N} - E[X_j]E[X_k]}{std[X_j]std[X_k]} = \frac{< X_j, X_k >}{std[X_j]std[X_k]}$$

$$y \sim x = (x_1, x_2, ..., x_n)$$

$$y \sim corr[X_j, Y]x_j$$

$$y \sim \sum_j corr[X_j, Y]x_j$$

Multicollinearity

$$\hat{w} = \left(X^T X\right)^{-1} X^T Y$$

$$x_1 = \sum_{j=2}^{n} k_j x_j$$

$$det(X^T X) \sim 0$$

Multicollinearity example

$$y = 2x_1$$

$$x_2 = x_1$$

$$y = 2x_2$$

$$y = x_1 + x_2$$

$$y = 0.5x_1 + 1.5x_2$$

$$y = -1000x_1 + 1001x_2$$

$$y = kx_1 + (2-k)x_2$$

Overfitting

Area	Month	Price
1010	Jan	100000
2000	Feb	200000
2990	March	300000

`Price=100.1418*Area`

Area	Month	Price
1500	Dec	150000

`Price=100.1418*1500=150212.7`

Overfitting

Area	Month	Price
1010	1	100000
2000	2	200000
2990	3	300000

`Price=100.000*Month`

Area	Month	Price
1500	12	150000

`Price=12*100.000=`
`1.200.000?`

Overfitting

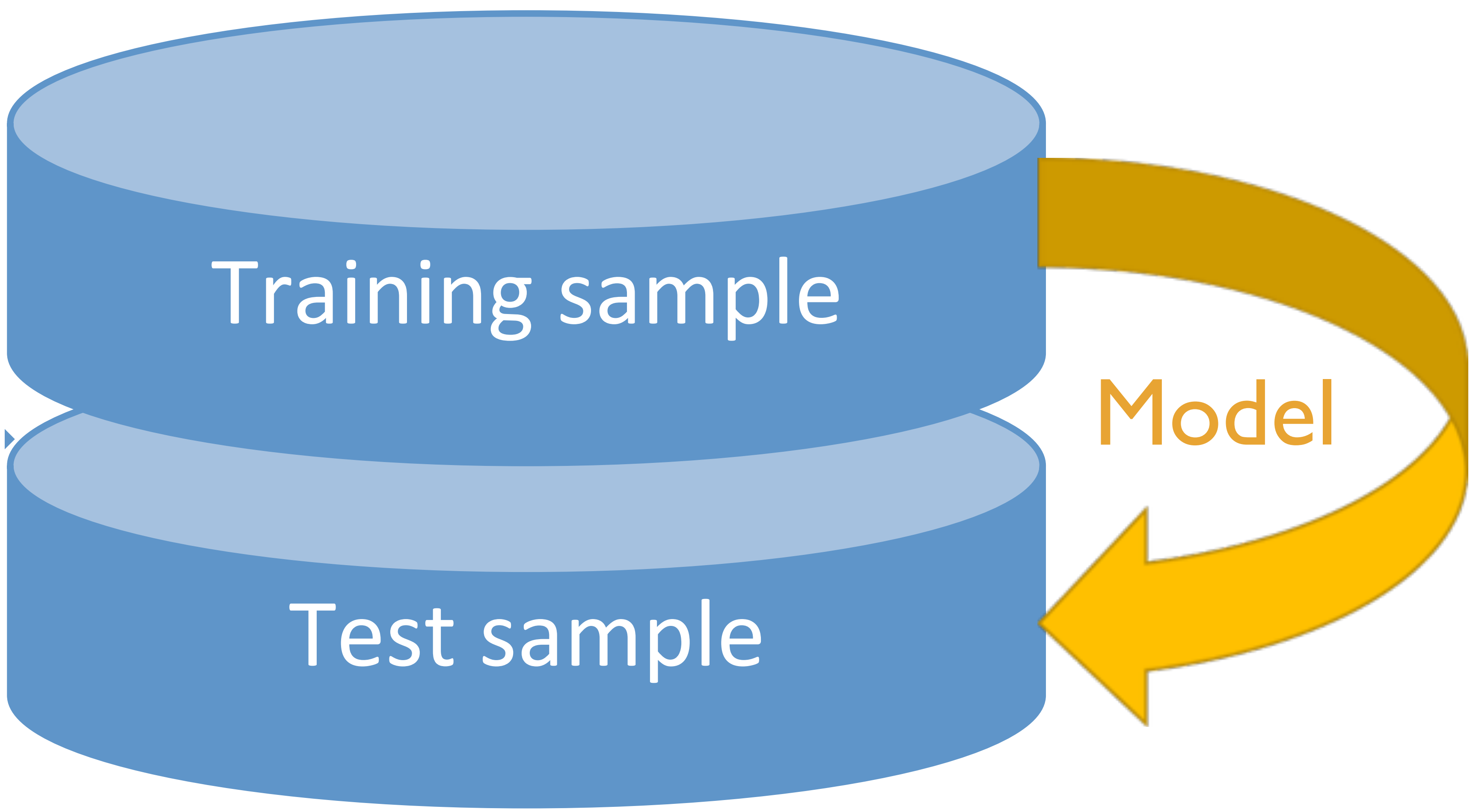
Area	Month	Price
1010	1	90000
2000	2	200000
2990	3	310000

Price=-1000*Area+1.100.000*Month

Area	Month	Price
1500	12	150000

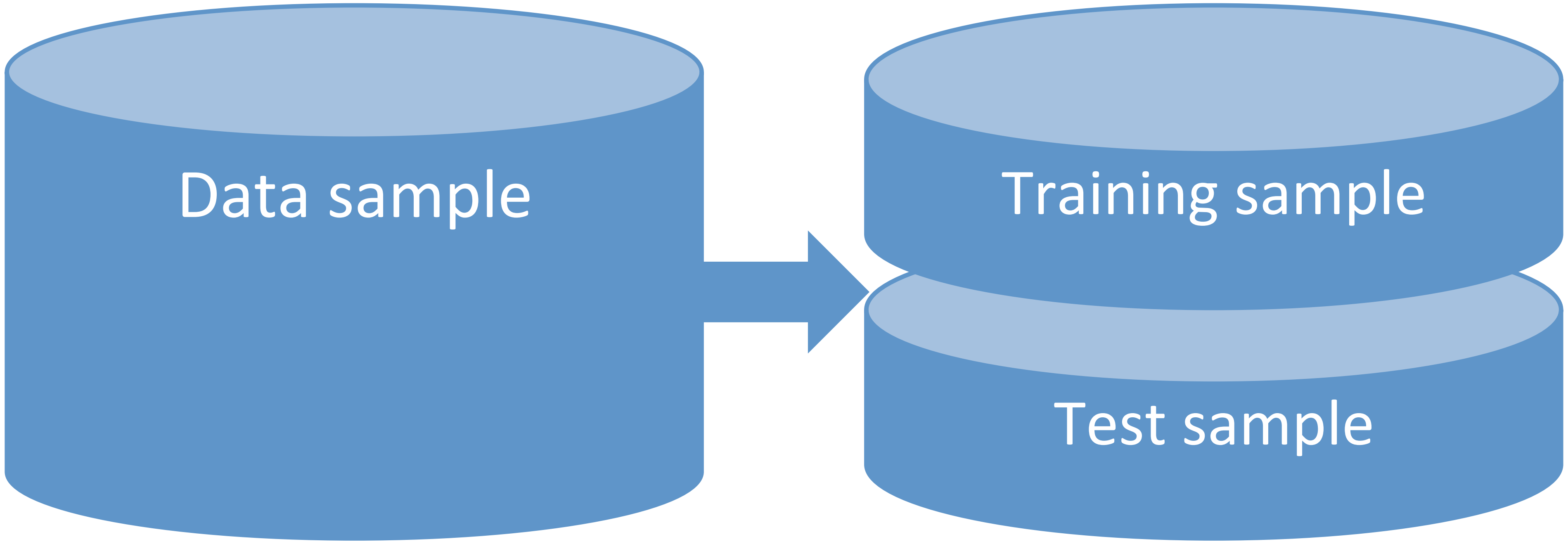
Price=-1000*1500+
1.1M*12=11.7M?

Validating regression



Fighting overfitting

No future data? Make it!



Adding more features is not always better

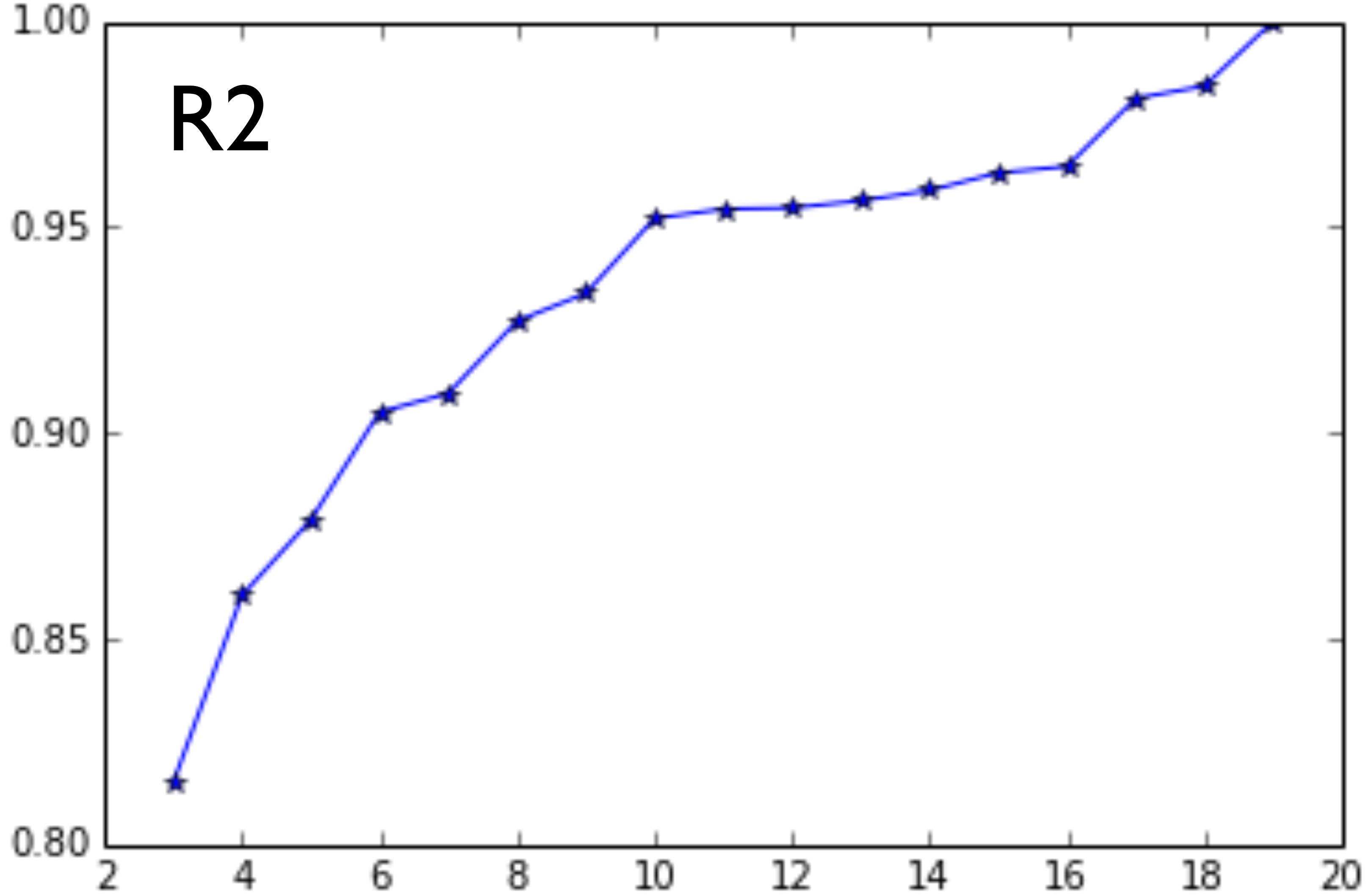
$$Y \sim X_1 + X_2 + X_3 + \mathcal{N}(0, 1.1)$$

$$Y \sim X_1 + X_2 + X_3$$

$$X_j \sim \mathcal{N}(0, 1), j = 4, 5, 6, \ldots ?$$

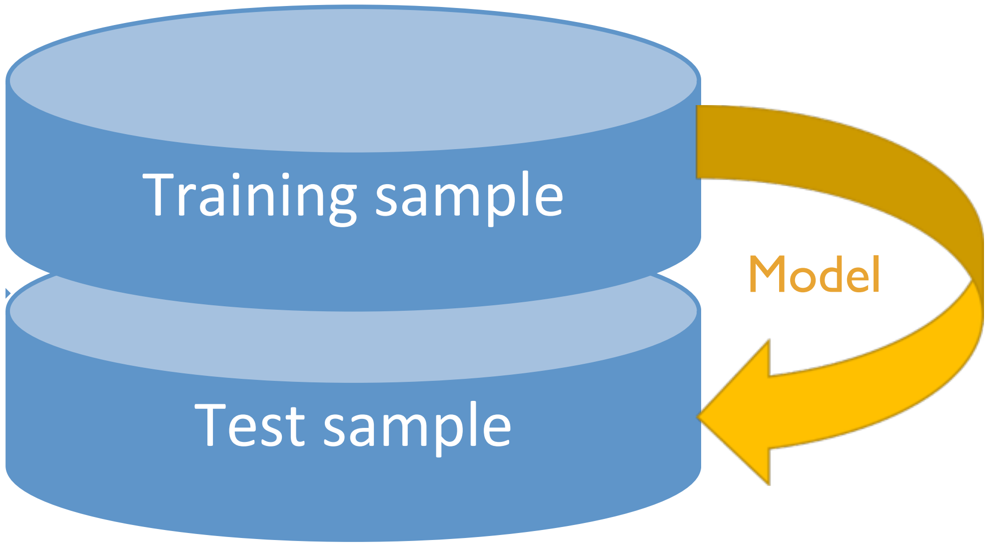
Adding more features is not always better

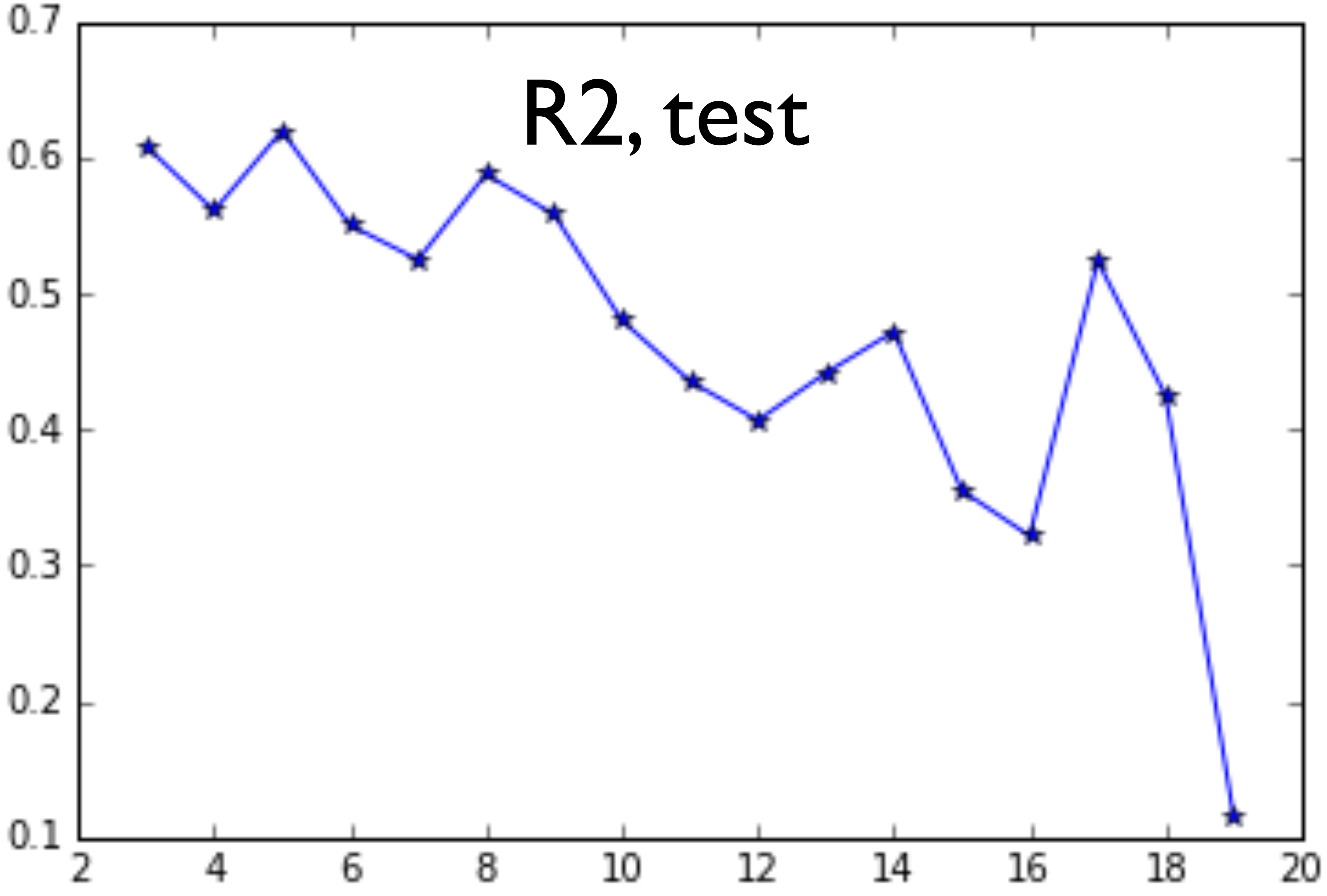
$$Y \sim \sum_{j=1}^{k} X_j, k = 3, 4, ..., 20 \qquad X_j \sim \mathcal{N}(0,1), j = 4, 5, 6, \ldots ?$$



Adding more features is not always better

$$Y \sim \sum_{j=1}^{k} X_j, k = 3, 4, ..., 20 \qquad X_j \sim \mathcal{N}(0, 1), j = 4, 5, 6, \ldots ?$$





Polynomial Models

$$y = w_2 x^2 + w_1 x + w_0$$

$$x_1 = x^2, x_2 = x, x_3 = 1$$

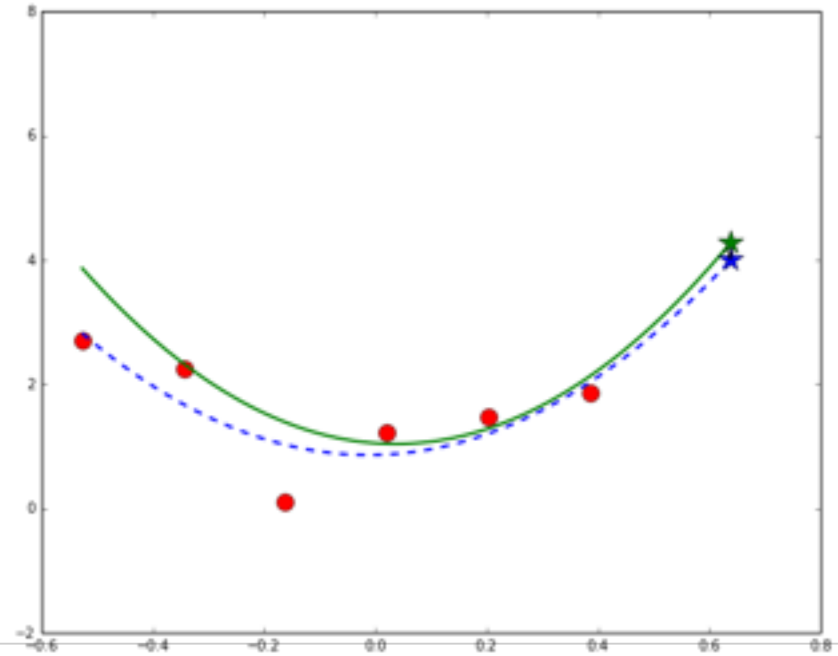
$$y \sim x_1, x_2, x_3$$

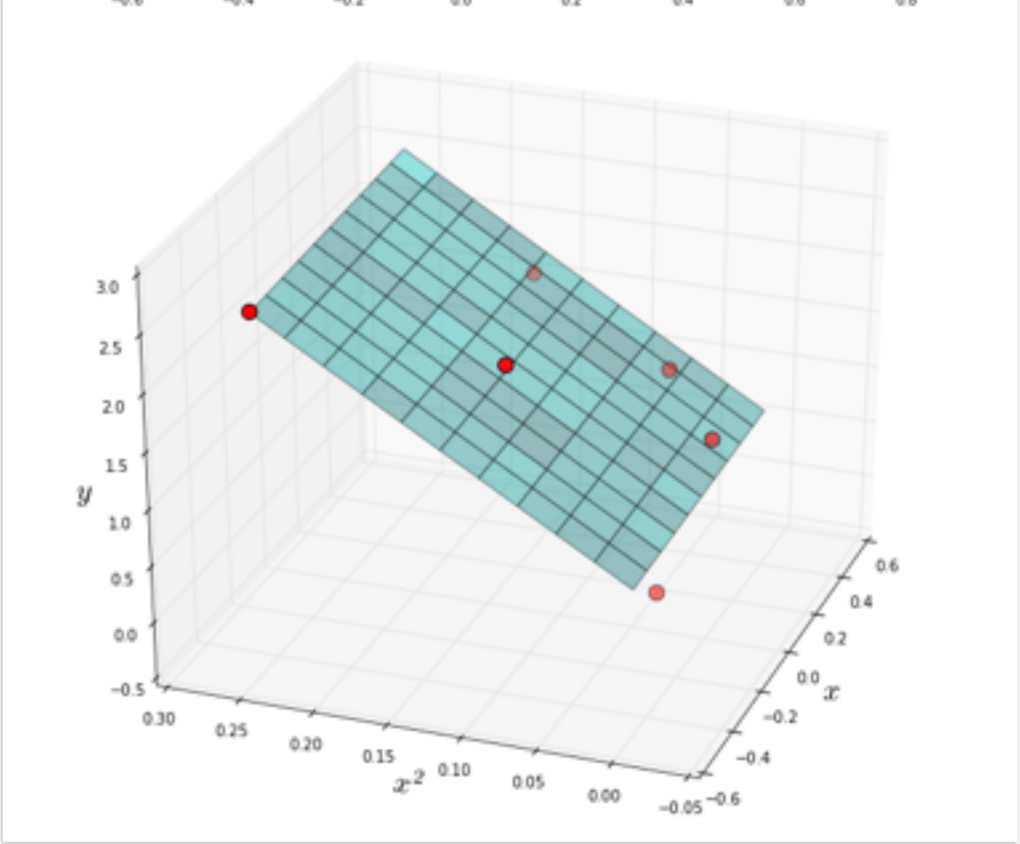
$$y = w_m x^m + w_{m-1} x^{m-1} + \ldots + w_1 x + w_0$$

$$y \sim x^m, x^{m-1}, \ldots, x, 1$$

$$y \sim w_{2,0} x_1^2 + w_{1,1} x_1 x_2 + w_{2,0} x_2^2 + w_{1,0} x_1 + w_{0,1} x_2 + w_{0,0}$$

$$y \sim 1, x_1, x_2, x_1^2, x_2^2$$





Polynomial Models - overfitting example