Lecture 15

polynomial regression work sheet:

Q1: Degree 4 polynomial model:

$$Y = C_0 + C_1 X_1 + C_2 X_1^2 + C_3 X_1^3 + C_4 X_1^4$$

Q2

$$\vec{Y} = A\vec{C}$$

$$Y = C_0 + C_1 X_1 + C_2 X_1^2$$

(X_1: icecream; Y: happiness)

$$\text{icecream} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

$$\text{Chocolate} = \begin{bmatrix} 99 \\ 991 \\ 9992 \\ 1992 \end{bmatrix}$$

$$\vec{Y} = A\vec{C}$$

$$\begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} \begin{bmatrix} C_0 \\ C_1 \\ C_2 \end{bmatrix}$$

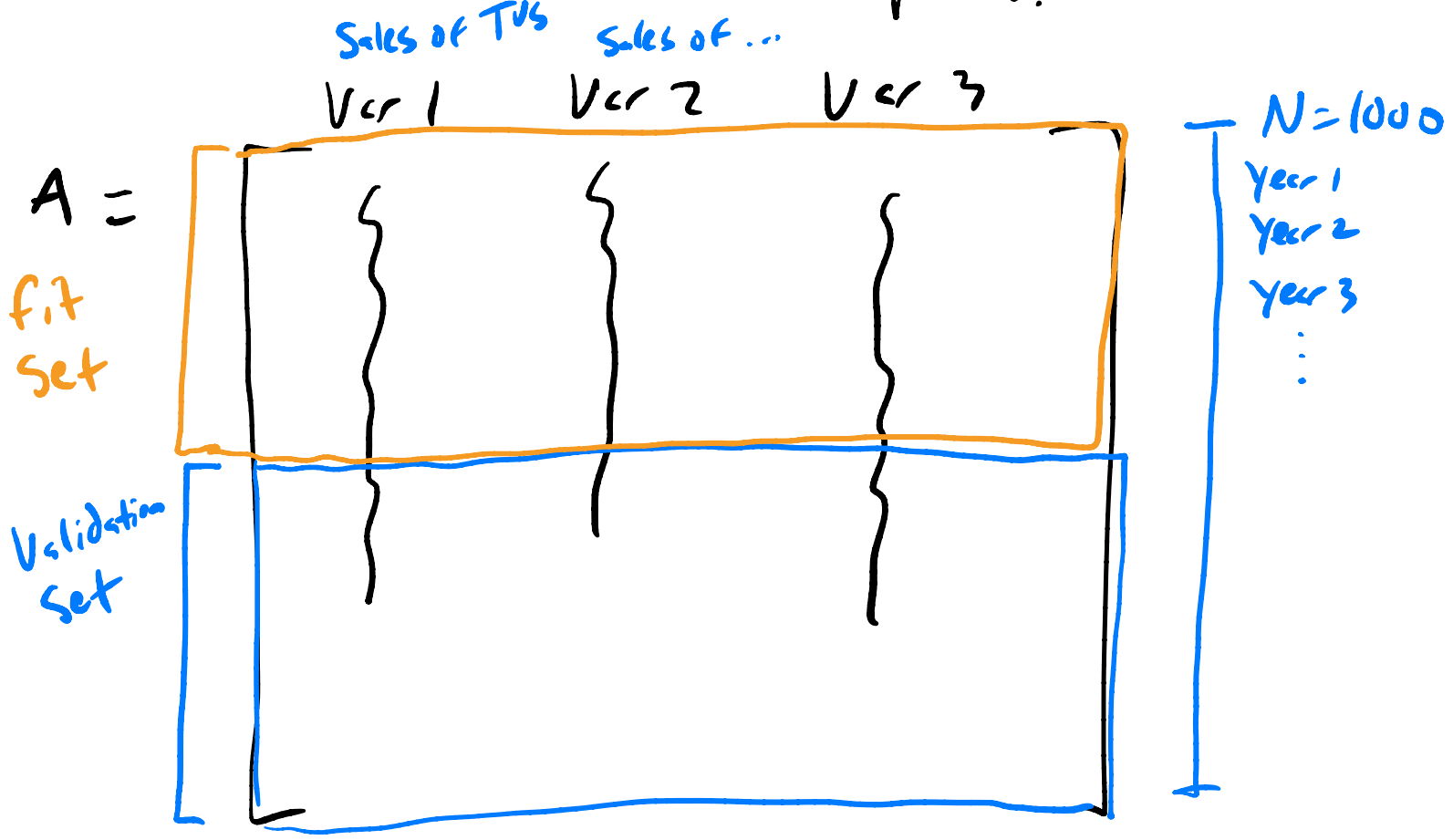
$\vec{Y}$ $\quad$ A $\quad$ $\vec{C}$

$$\text{happiness} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Fit and Validation Sets:

To help detect overfitting / prevent it, you can split your dataset into 2 parts:

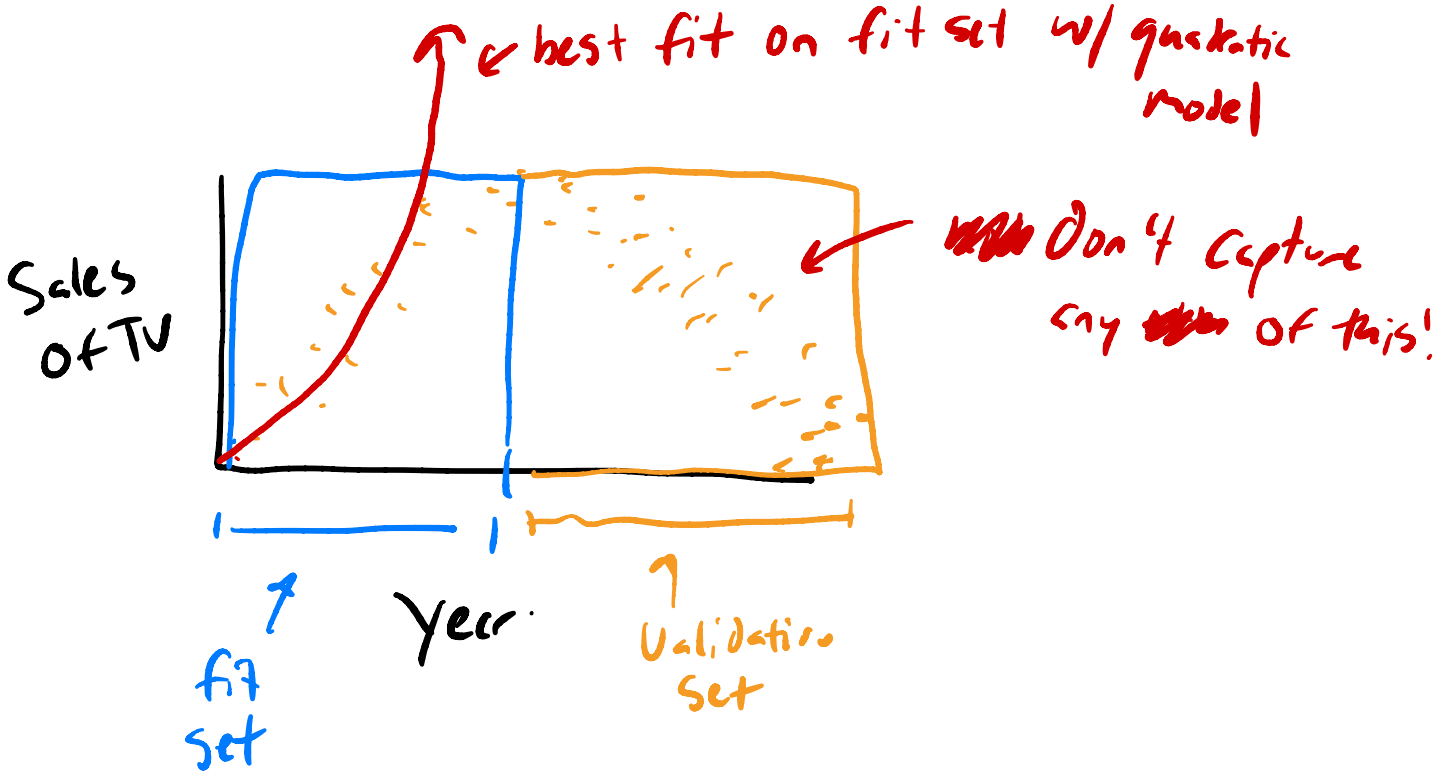
1) **Fit set**: data used to fit linear regression
2) **Validation set**: data used to check for overfitting: how well does regression fit generalize to data not used in fit process.



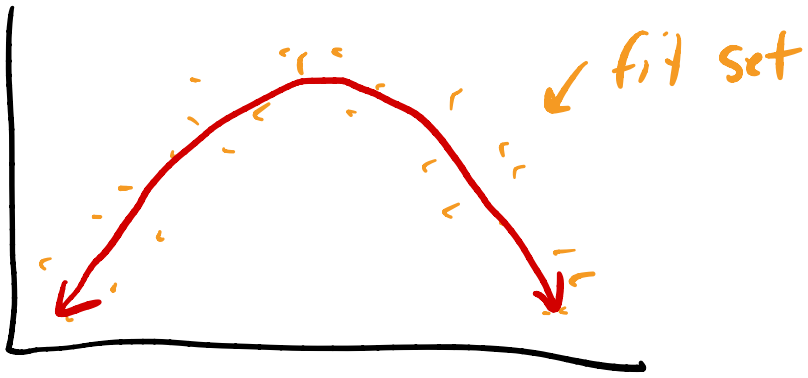
example: if N = 1000, could make 1st N = 500 fit set, remaining 500 be validation set.

★ Each sample (A) must be in either be fit set or validation, not both.

★ Usually best randomly the shuffle order of N samples (rows) before splitting A into fit/validation sets.

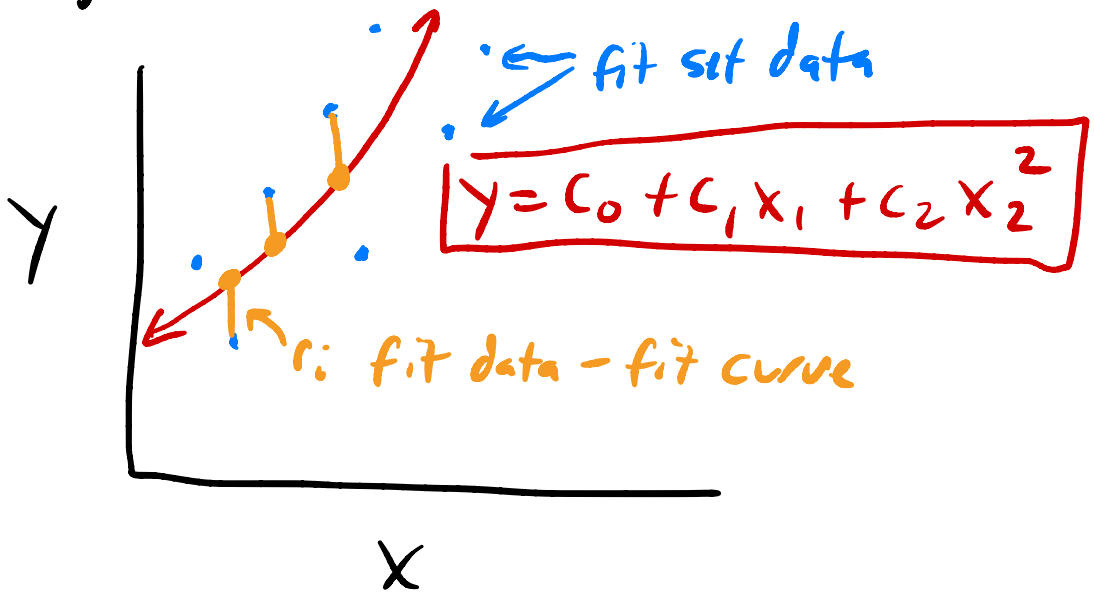


Randomly shuffle:



Fit and validation data workflow

1) Fit regression model to fit set data ⇒ get coefficients $\vec{c}$:



2) See how well I fit the fit data

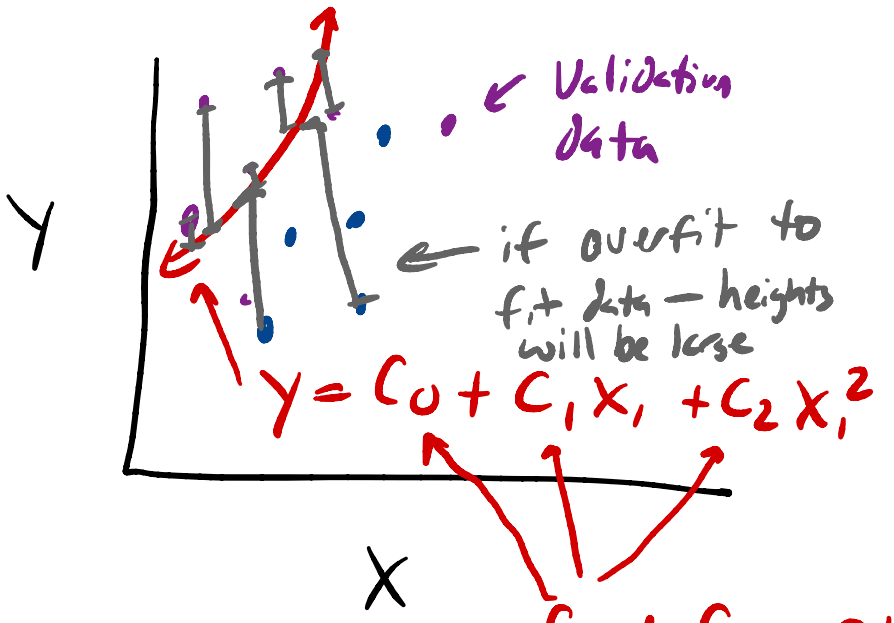
Compute residuals between

- quadratic model
- fit data

⇒ R^2, MSSE, ... get quality of fit.

3) Create a <u>new</u> plot, plot Validation data

plug in values into fitted regression model based on Validation data. ⇒ <u>Do not</u> recompute regression coefficients with Validation.



⇒ reuse C_0, C_1, C_2 determined by fit set.

4) Compute residuals between

- regression curve fit using fit data
- Validation data

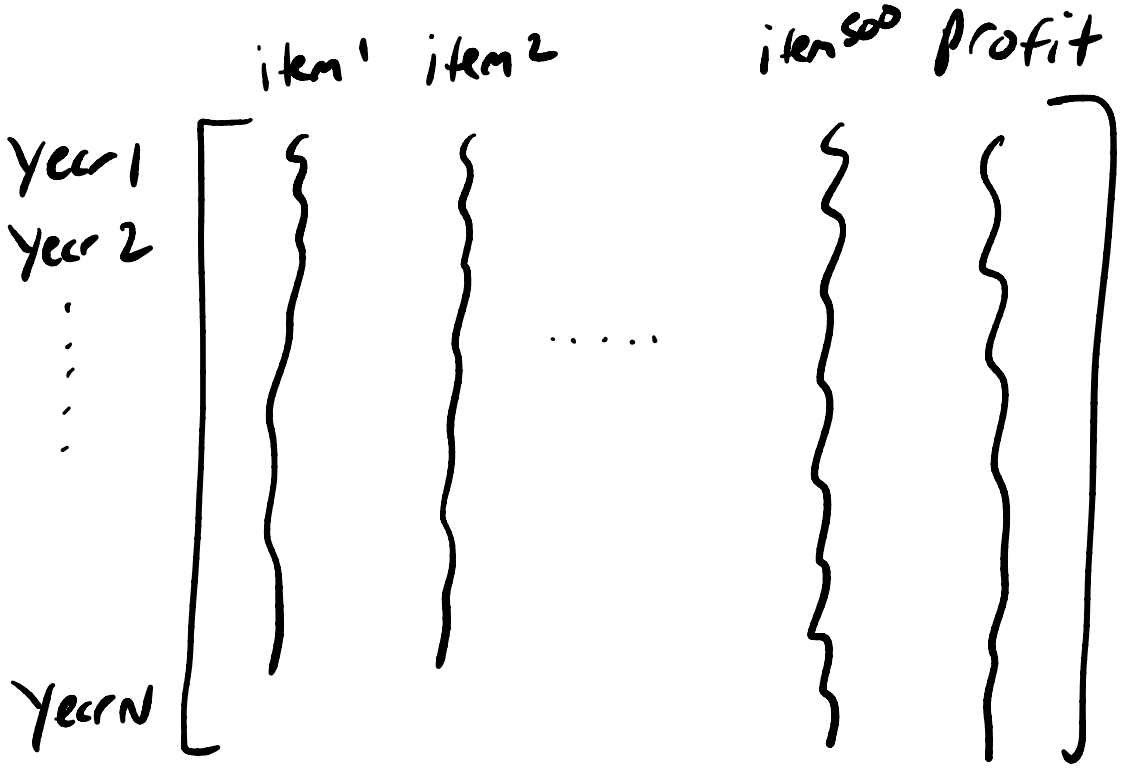
⇒ Calculate R^2, MSSE using ↑

Overfitting

fit data : $R^2 \approx 1$, MSSE ≈ 0

Validation data : $R^2 \approx 0$, MSSE big

Stepwise linear regression



Stepwise linear regression : greedy algorithm.

1) Start regression model w/ intercept only: $A = [\vec{1}]$

2) Try to add each one of the 500 Vars to your regression model:

Try: $A = [\vec{1}, \vec{x}_1]$

↑ 1 Mil Var

⇒ Calc R^2 to measure how predictive model is of $\vec{y}$ profit.

Next: $A = [\vec{1}, \vec{x}_2]$

⇒ keep track of which added Var produces the highest improvement in R^2 value.

⇒ After visiting all 500 Vars, officially add Var that improved R^2 the most.

e.g. $\vec{x}_{30}$

$A = [\vec{1}, \vec{x}_{30}]$

3) With updated model, try again to add all remaining vars (499) to regression

e.g.

$$A = [\vec{1}, \vec{x}_{30}, \vec{x}_1] \Rightarrow R^2$$

$$A = [\vec{1}, \vec{x}_{30}, \vec{x}_2] \Rightarrow R^2$$

$$\vdots$$

$$\vdots$$

4) Keep going until all variables added to large 500 var regression