Interaction Terms and Polynomial Regression

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai and Chris Gumb

Photo:Thea Tjolle
Scotland

Lecture Outline

Interaction Effects in Regression Models

Polynomial Regression: Extending Linear Models

Model Selection Techniques: Focus on Cross-Validation

Too many predictors and collinearity leads to **OVERFITTING!**

Game Time

If your model was a student, what would overfitting be like?

Options:

A. Studying just the night before the test

B. Memorizing every lecture, lab and OH word-for-word

C. Only studying one chapter for all subjects

D. Taking extensive notes but forgetting to actually understand the concepts

Game Time

If your model was a TF, what would overfitting be like?

Options:

A. Grading papers while wearing 3D glasses to "see the errors in a new dimension”

B. Using a "Magic 8-Ball" to decide students' grades

C. Give everyone the first letter that comes on their name. Sorry Frank

D. Subtract points for every answer that does not include the word overfitting

Too many predictors and collinearity and leads to

OVERFITTING!

Too many predictors and collinearity and leads to **OVERFITTING!**

Overfitting occurs when a model learns the training data too well, including its noise and outliers, resulting in poor performance on new, unseen data.

Lecture Outline

Interaction Effects in Regression Models

Polynomial Regression: Extending Linear Models

Model Selection Techniques: Focus on Cross-Validation

Assumptions of Linear Regression

Linearity: Relationship between variables is linear.

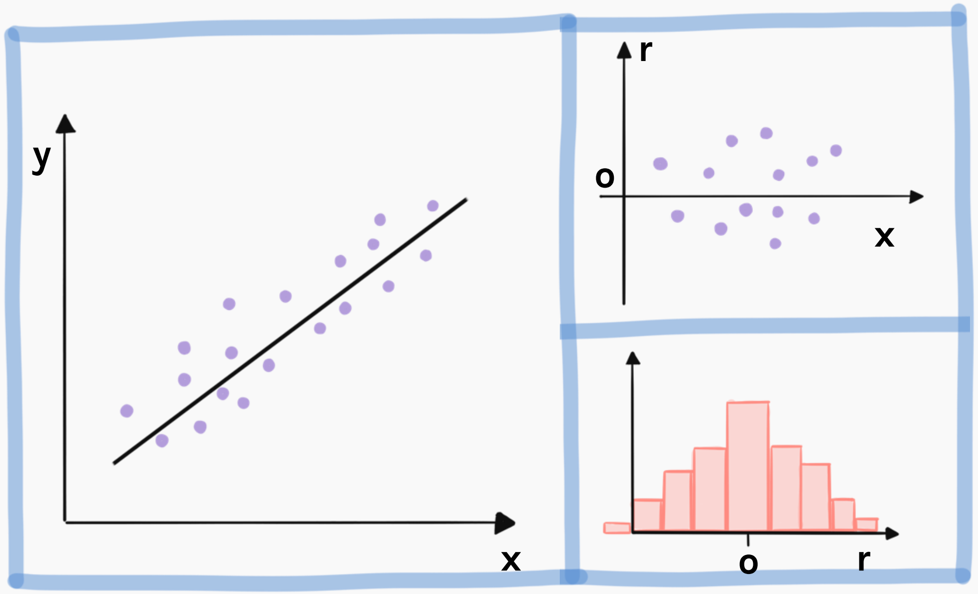
$$f(x) = \beta_0 + \beta_1 x$$

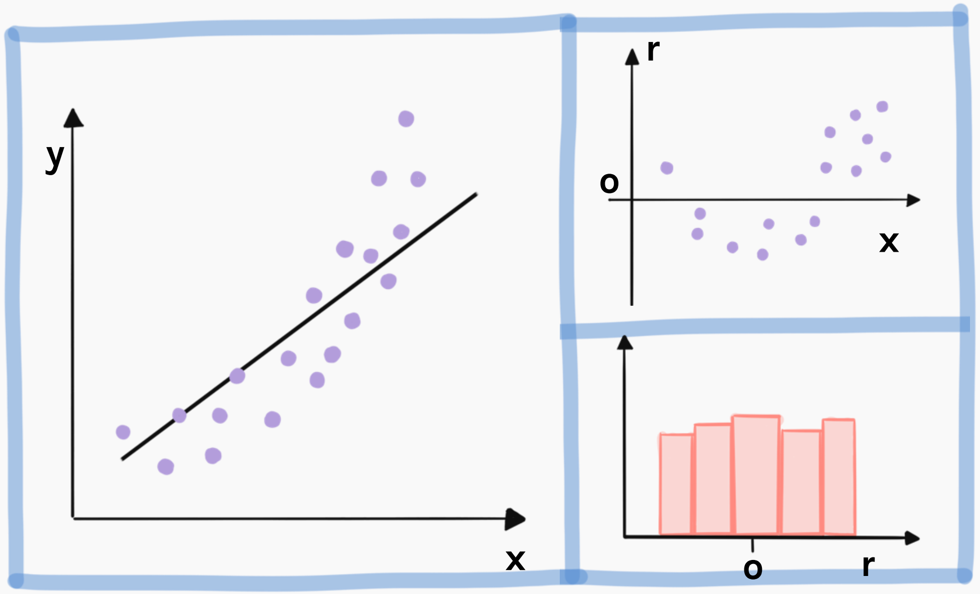
Independence: No correlation between errors and predictors.
Homoscedasticity: Constant variance of residuals.
Normality of Residuals: Residuals are normally distributed.

$$y = f(x) + \epsilon$$
$$L(\beta_0, \beta_1) = MSE$$

Other things to consider
Fixed X: Independent variables are error-free.
No Multicollinearity: Low correlation between predictors.

Residual Analysis





Linear assumption is correct. There is no obvious relationship between residuals and x. Histogram of residuals is symmetric and normally distributed.

Linear assumption is incorrect. There is an obvious relationship between residuals and x. Histogram of residuals is symmetric but **not** normally distributed.

Note: For multi-regression, we plot the residuals vs predicted y, $\hat{y}$, since there are too many x's and that could wash out the relationship.

Beyond linearity: **synergy effect** or **interaction effect**

We assumed that the average effect on *sales* of a one-unit increase in *TV*, is always β_1 regardless of the amount spent on *radio* or *newspaper*.

Synergy effect or **interaction effect** states that when an increase on the *radio budget* affects the effectiveness of the *TV* spending on *sales*.

To account for it, we simply add a term as:

Not linear term

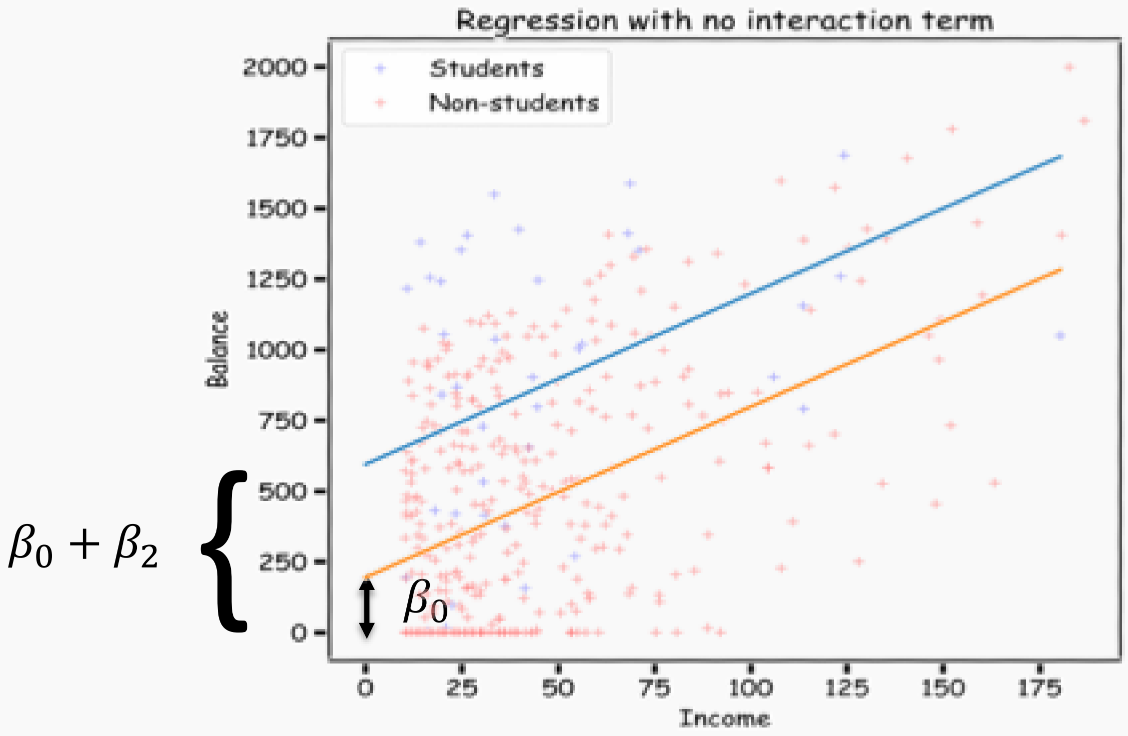
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boxed{\beta_3 X_1 X_2}$$

What does it mean? First consider the case without the interaction term

$$balance = \beta_0 + \beta_1 \times income + \beta_2 \times student$$

$$student = \begin{cases} 0 & balance = \beta_0 + \beta_1 \times income \\ 1 & balance = \beta_0 + \beta_1 \times income + \beta_2 \rightarrow balance = \overbrace{(\beta_0+\beta_2)}^{\text{intercept}} + \underbrace{\beta_1}_{\text{slope}} \times income \end{cases}$$



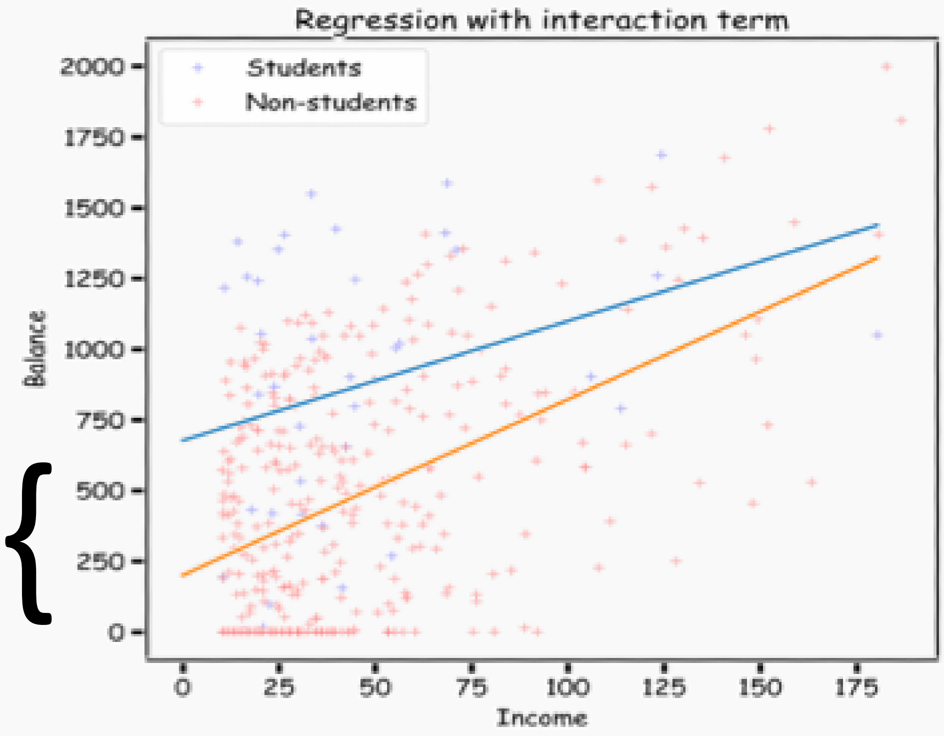
What does it mean? Next we consider the case with the interaction term

$$balance = \beta_0 + \beta_1 \times income + \beta_2 \times student + \beta_3 \times income \times student$$

$$student = \begin{cases} 0 & balance = \beta_0 + \beta_1 \times income \\ 1 & balance = \beta_0 + \beta_1 \times income + \beta_2 + \beta_3 \times income \end{cases}$$

$$\rightarrow balance = \underbrace{(\beta_0 + \beta_2)}_{\text{intercept}} + \underbrace{(\beta_1 + \beta_3)}_{\text{slope}} \times income$$

$\beta_0 + \beta_2$



⏳ Digestion Time

Too many predictors, collinearity and too many interaction terms leads to

Too many predictors, collinearity and too many interaction terms leads to

OVERFITTING!

Lecture Outline

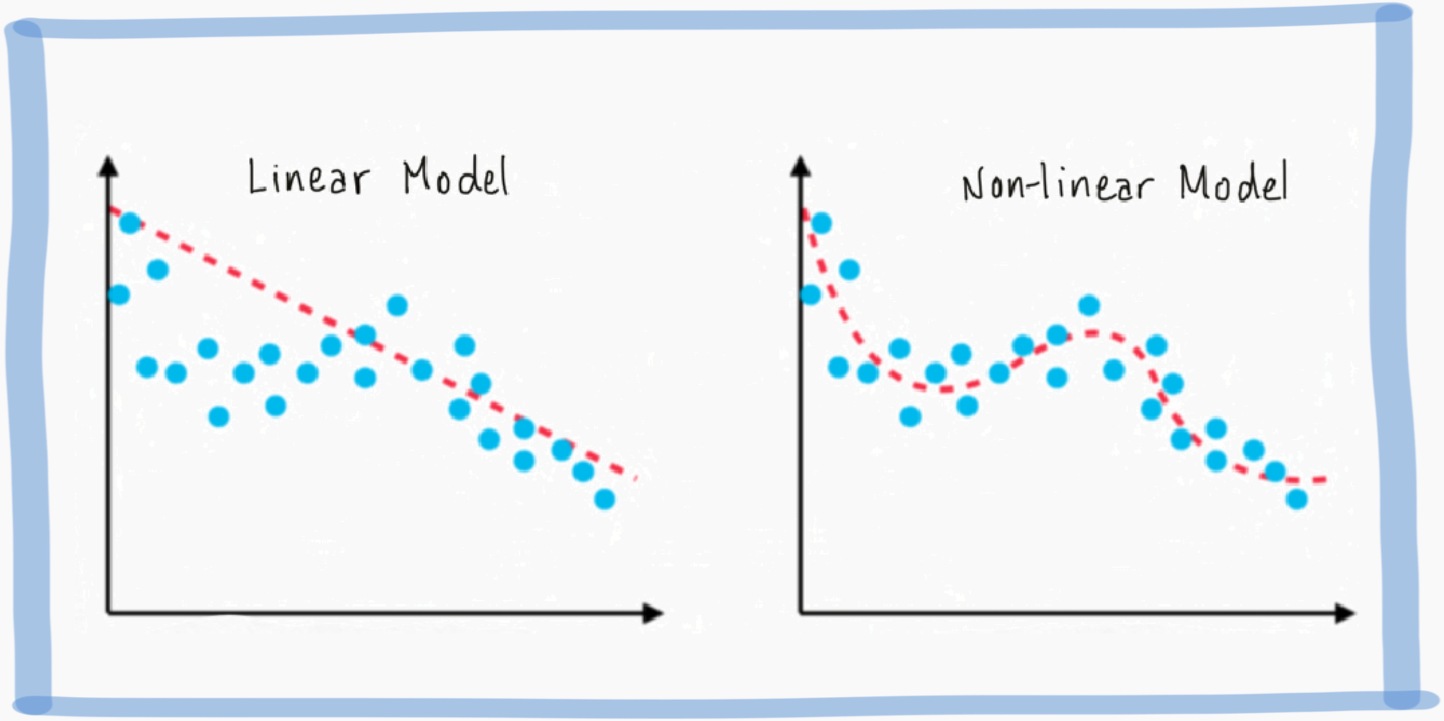
Interaction Effects in Regression Models

Polynomial Regression: Extending Linear Models

Model Selection Techniques: Focus on Cross-Validation

Fitting non-linear data

Multi-linear models can fit large datasets with many predictors. But the relationship between predictor and target isn't always linear.



We want a model:

$$y = f_{\beta}(x)$$

Where f is a non-linear function and β is a vector of the parameters of f.

Polynomial Regression

The simplest non-linear model we can consider, for a response Y and a predictor x, is a polynomial model of degree M,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

Just as in the case of multi-linear regression, polynomial regression is a special case of linear regression

HOW?

Polynomial Regression

The design matrix for a polynomial regression would be:

To the power of M

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^M \\ 1 & x_2^1 & \dots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^M \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$

Polynomial Regression

The design matrix for a polynomial regression would be:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^M \\ 1 & x_2^1 & \dots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^M \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$

This looks a lot like multi-linear regression where the predictors are powers of x!

Multi-Regression

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_y \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,J} \\ 1 & x_{2,1} & \dots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,J} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

Model Training

Give a dataset $\{(x_1, y_1), (x_2, y_2), \dots, (x$
model:

$$y = \beta_0 + \beta_1 x$$

1. We transform the data by adding

$$\tilde{x} = [1, \tilde{x}$$

where $\tilde{x}_k = x^k$

2. We find the parameter by minimizing the MSE using vector calculus yields, as in multi-linear regression

$$\boxed{\hat{\beta} = \left(\tilde{X}^T \tilde{X}\right)^{-1} \tilde{X}^T y}$$

We can also perform multi-polynomial regression in the same way.

BUT be careful:

A. Sklearn will include the interaction terms
B. The new design matrix will include the 1 column so no need to fit for intercept

g:

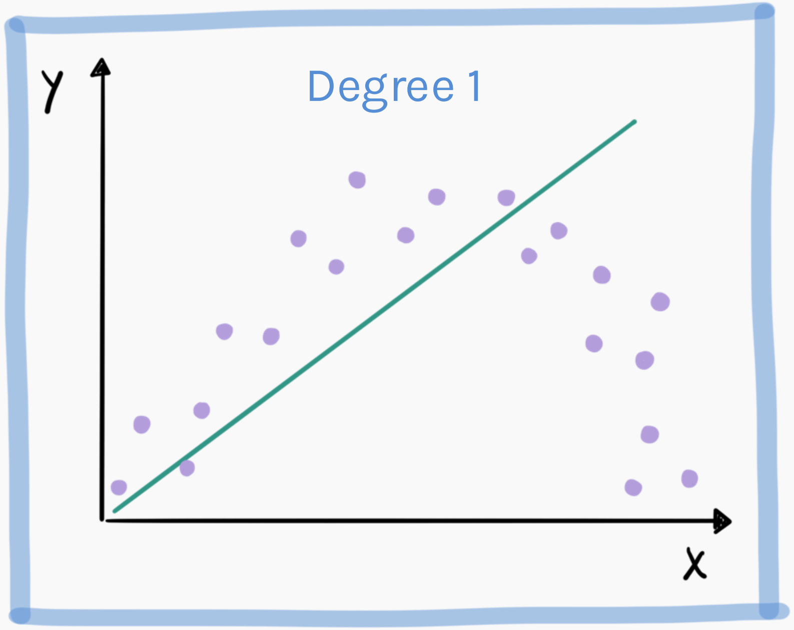
oly

?)

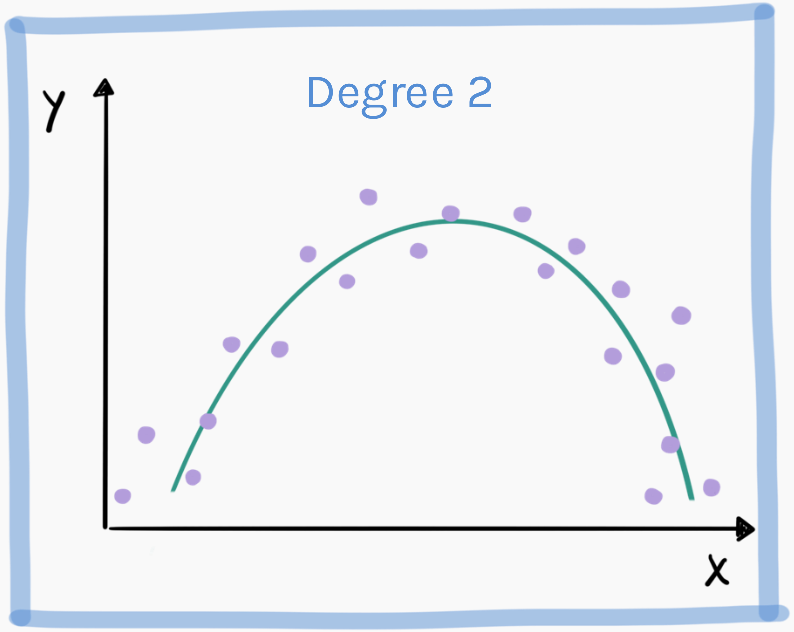
sklearn.linear_model.LinearRegression.fit()

Polynomial Regression (cont.)

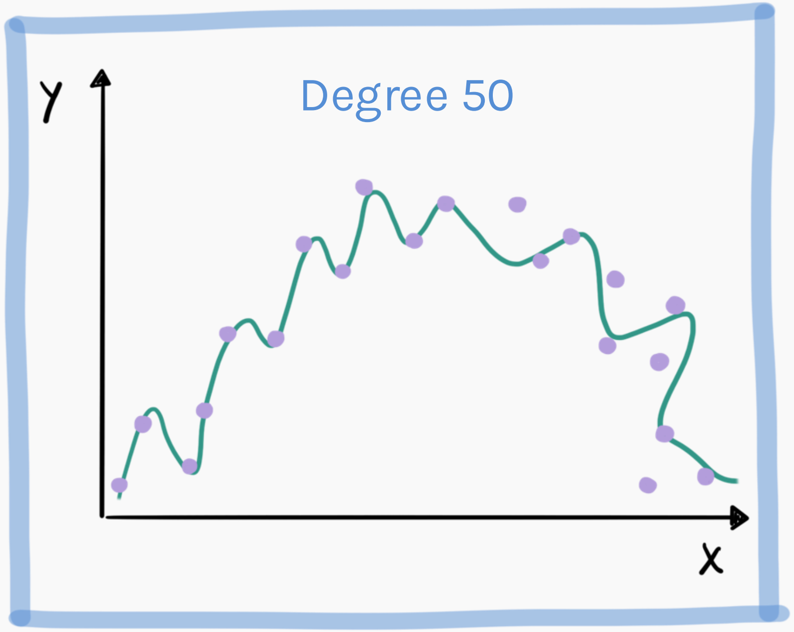
Fitting a polynomial model requires choosing a degree.



Underfitting: when the degree is too low, the model cannot fit the trend.



We want a model that fits the trend and ignores the noise.



Overfitting: when the degree is too high, the model fits all the noisy data points.

Feature Scaling

Do we need to scale out features for polynomial regression?

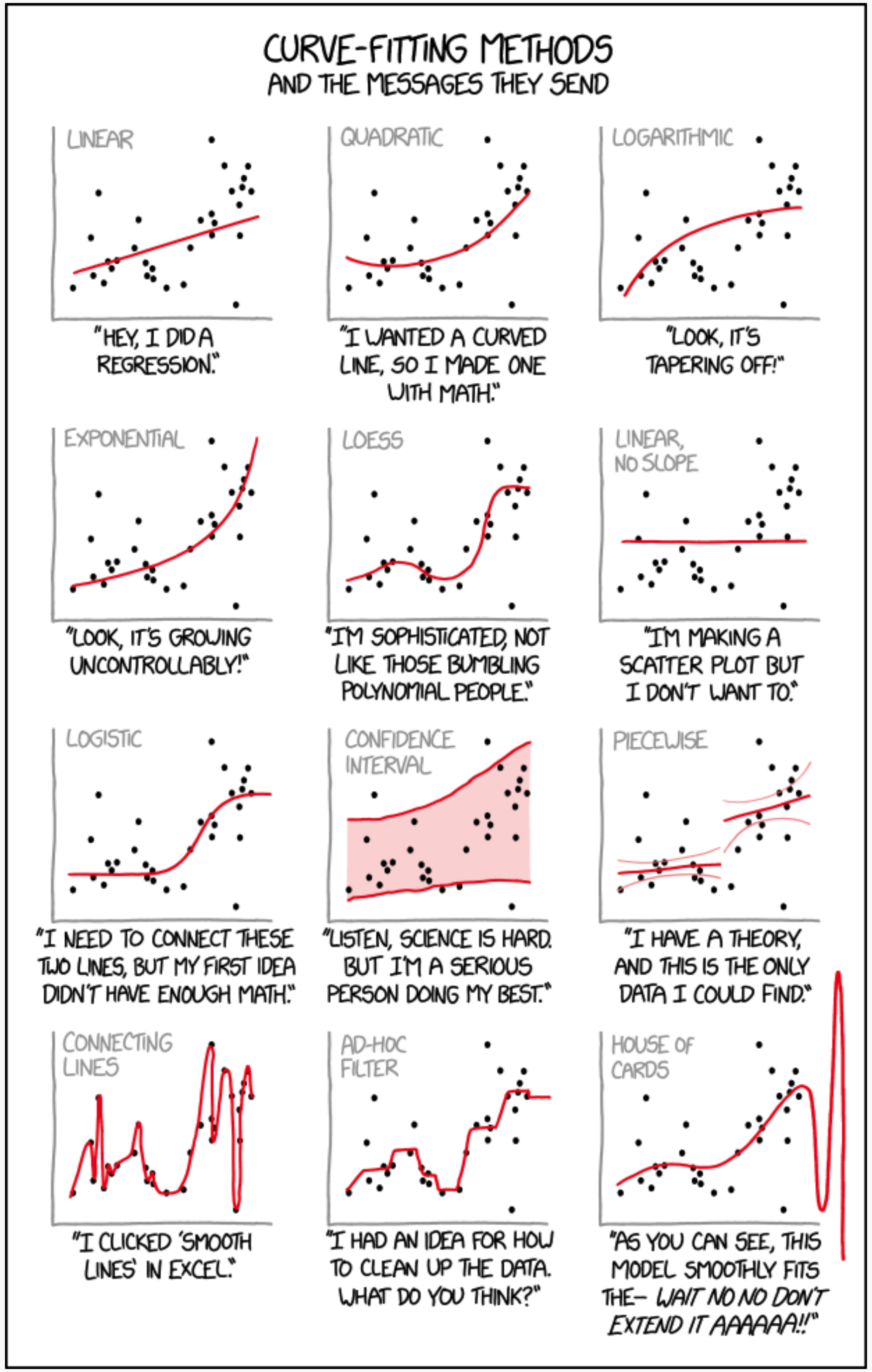
Linear regression, $Y = X\beta$, is invariant under scaling. If X is multiplied by some number λ, then β will be scaled by $\frac{1}{\lambda}$ and MSE will be identical.

However, if the range of X is small or large, then we run into troubles. Consider a polynomial degree of 20 and the maximum or minimum value of any predictor is large or small. Those numbers to the 20$^{\text{th}}$ power will be problematic.

It is always a good idea to scale X when considering polynomial regression:

$$X^{norm} = \frac{X - \bar{X}}{\sigma_X}$$

Note: sklearn's `StandardScaler()` can do this.


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– WAIT NO NO DON'T EXTEND IT AAAAAA!!"

Too many predictors, collinearity, too many interaction terms and high degree of polynomial leads to leads to

Too many predictors, collinearity, too many interaction terms and high degree of polynomial leads to leads to

OVERFITTING!

Too many predictors, collinearity, too many interaction terms and high degree of polynomial and model selection leads to

Too many predictors, collinearity, too many interaction terms and high degree of polynomial and model selection leads to

THESE ARE OVERFITTED STUDENTS!