# Polinomial Regression

Nel caso in cui non avessimo altre features a disposizione, o la distribuzione dei dati non fosse lineare, potremmo utilizzare una regressione polinomiale. Nel caso preso in esempio, abbiamo visto come la retta non fosse in grado di generalizzare particolarmente bene i dati. Prima di procedere con la pratica vediamo la formula della regressione polinomiale, anche se è abbastanza intuitiva e non ci dovrebbe essere nulla da spiegare:

$$\hat{y} = wx + b \underset{polinomiale}{\longrightarrow} \hat{y} = b + \sum_{i=1}^{n} w_i x^i$$

Applicando la regressione polinomiale con grado del polinomio: $n = 2$ all'esempio visto in precedenza otteniamo:

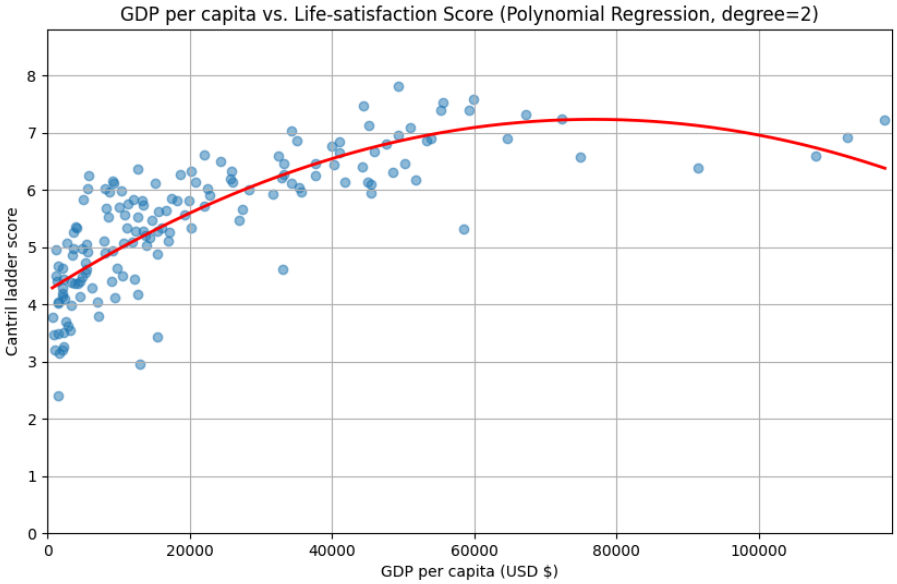


Figure 1: Plot dei dati GDP vs Life satisfaction con la regressione lineare e grado 2

Ora il modello predice un valore di 7,01 per l'Austria, più vicino al valore reale. Proviamo con gradi ancora più alti:

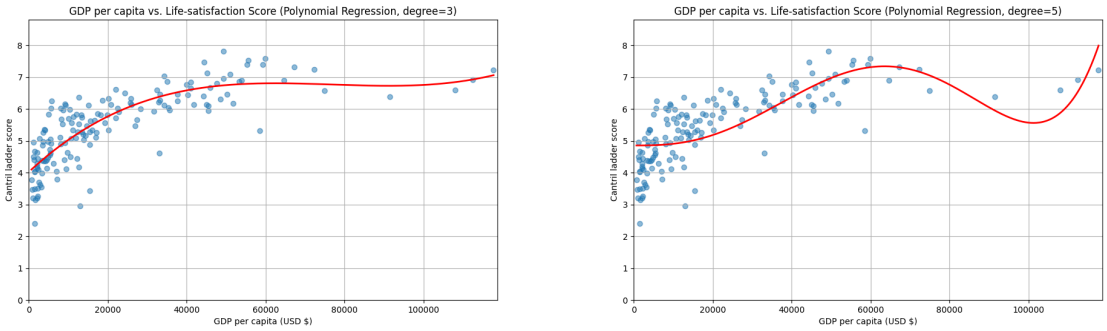


Figure 2: Plot dei dati GDP vs Life satisfaction con la regressione lineare e grado 3 e 5

Con un grado 3 il modello predice un valore di 6,79, mentre con un grado 5 il modello predice un valore di 7,21.

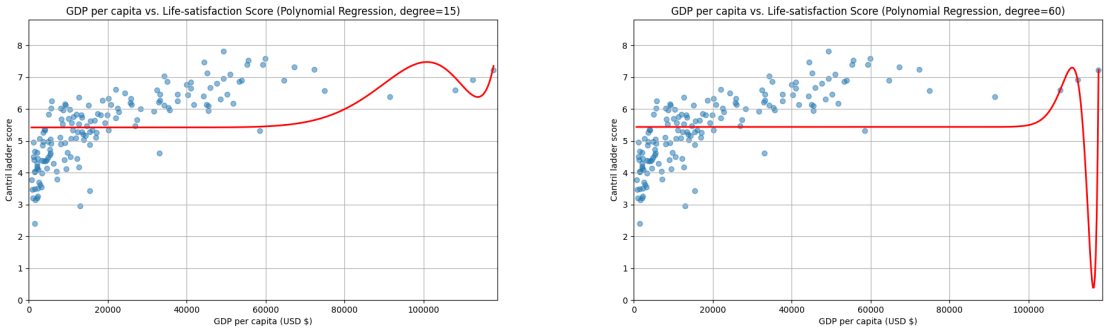


Figure 3: Plot dei dati GDP vs Life satisfaction con la regressione lineare e grado 15 e 60

Se alzialiamo ulteriormente il grado del polinomio, il modello non migliorerà; anzi dalla tabella di seguito e dalle immagini precedenti dovrebbe essere chiaro come avere un grado del polinomio più alto non implichi che il modello generalizzi meglio.

| grado | 1 | 2 | 3 | 5 | 15 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|---|---|---|---|
| predizione | 6.66 | 7.01 | 6.79 | 7.21 | 5.43 | 5.43 | 5.43 | 5.43 | 5.43 | 5.43 |
| errore | ±0.44 | ±0.09 | ±0.31 | ±0.11 | ±1.67 | ±1.67 | ±1.67 | ±1.67 | ±1.67 | ±1.67 |

Dopo questi due capitoli, dovrebbero sorgere al lettore almeno due domande:

1. Come possiamo valutare il modello in modo rigoroso?

   Per valutare il modello in maniera sistematica dovremmo innanzitutto dividere il dataset in training set e test set; in quanto il test set d'esempio è composto solo dall'Austria. Successivamente dobbiamo decidere come valutare l'errore del modello, una delle metriche più diffuse per la task di regressione è l'**errore quadratico medio** (MSE).

2. Come possiamo capire quale grado del polinomio è il migliore?

   Il grado del polinomio è un **iperparametro** del modello; e per essere selzionato al meglio viene utilizzato un **validation set**.

Nel prossimo capitolo vedremo come valutare un modello di regressione, come selezionare il grado del polinomio.