Multiple Linear Regression

• Extension of the simple linear regression model to two or more independent variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

$$\text{Expression} = \text{Baseline} + \text{Age} + \text{Tissue} + \text{Sex} + \text{Error}$$

• Partial Regression Coefficients: βi ≡ effect on the dependent variable when increasing the i th independent variable by 1 unit, holding all other predictors constant.

Underfitting and Overfitting

• Underfitting occurs when a model cannot accurately capture the dependencies among data, usually as a consequence of its own simplicity.

• It often yields a low R2 with known data and bad generalization capabilities when applied with new data.

• Overfitting happens when a model learns both dependencies among data and random fluctuations.

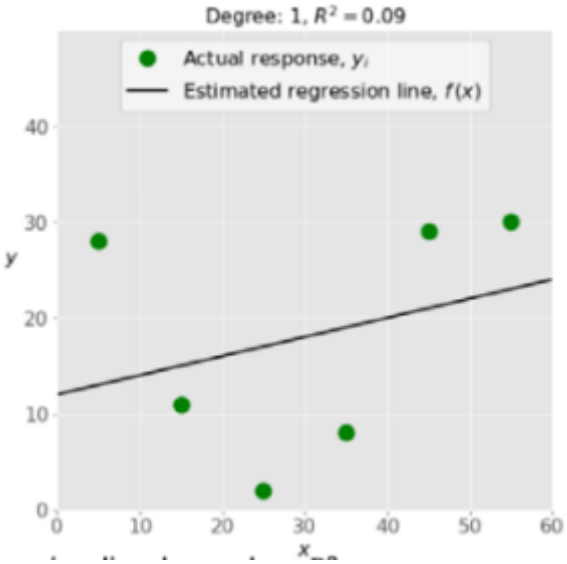
• In other words, a model learns the existing data too well.

• Complex models, which have many features or terms, are often prone to overfitting.

• When applied to known data, such models usually yield high R2.

• However, they often don’t generalize well and have significantly lower R2 when used with new data.

Linear Regression: Underfitting Example

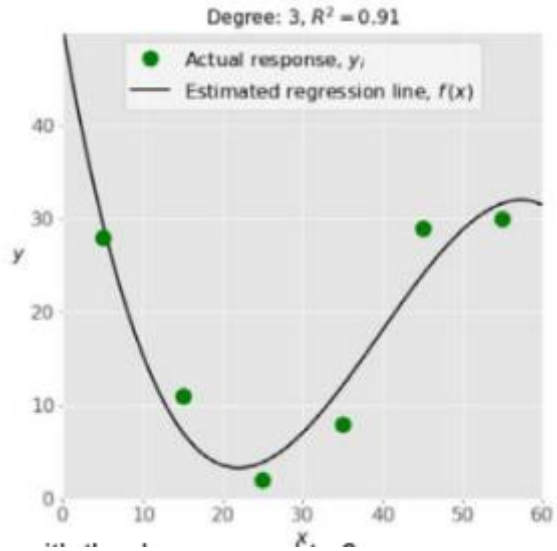


• This linear regression line has a low R2.

• The straight line can't take into account the fact that the actual response increases as x moves away from 25 towards zero.

• This is likely an example of underfitting.

Linear Regression: Overfitting Example



• Linear regression with the degree equal to 3.

• The value of R2 is higher than in the preceding cases.

• This model behaves better with known data than the previous ones.

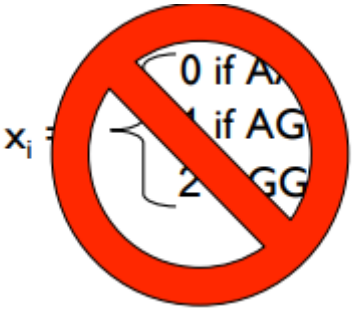
• However, it shows some signs of overfitting, especially for the input values close to 60 where the line starts decreasing, although actual data don't show that.

Categorical Independent Variables

• Qualitative variables are easily incorporated in regression framework through dummy variables

• Simple example: sex can be coded as 0/1

• What if my categorical variable contains three levels:



• Previous coding would result in colinearity

• Solution is to set up a series of dummy variable. In general, for k levels you need k-1 dummy variables

$$x_1 = \begin{cases} 1 \text{ if AA} \\ 0 \text{ otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ if AG} \\ 0 \text{ otherwise} \end{cases}$$

	x_1	x_2
AA	1	0
AG	0	1
GG	0	0