

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Multiple Linear Regression

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
 - Use case: When no single attribute alone is good enough to predict the response variable, *i.e., the Pearson correlation between the response and each other variable is low*. However, a linear combination of attributes gives a good prediction.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,
$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$
- Question: *How do we formulate the notion of the best predictor?*

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: *How do we formulate the notion of the best predictor?*

- Answer: The one that minimizes the **Mean Squared Error** (MSE)

Find $(\theta_0, \theta_1, \dots, \theta_m)$ that minimizes $\frac{1}{n} \sum_{i=1} (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_i (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

Find $(\theta_0, \theta_1, \dots, \theta_m)$ that minimizes

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:

- $\theta = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$

- $y = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

- $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \theta = X \cdot \theta$

- So, the problem in terms of the above matrices is:

Find θ that minimizes $\frac{1}{n} \|y - X \cdot \theta\|^2$

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

$$\text{Find } (\theta_0, \theta_1, \dots, \theta_m) \text{ that minimizes } \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:
 - $\theta = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$
 - $y = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$
 - $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \theta = X \cdot \theta$
 - So, the problem in terms of the above matrices is:
Find θ that minimizes $\frac{1}{n} \|y - X \cdot \theta\|^2$
- How do we solve such an optimization problem?

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

$$\text{Find } (\theta_0, \theta_1, \dots, \theta_m) \text{ that minimizes } \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:

$$\theta = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$$

$$y = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \theta = X \cdot \theta$$

- So, the problem in terms of the above matrices is:

$$\text{Find } \theta \text{ that minimizes } \frac{1}{n} \|y - X \cdot \theta\|^2$$

- How do we solve such an optimization problem?
 1. Use calculus to find a close-form solution for θ .
 2. Use numerical hill-climbing technique such as **gradient ascent/descent**.

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

$$\text{Find } (\theta_0, \theta_1, \dots, \theta_m) \text{ that minimizes } \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$$

$$\mathbf{y} = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \boldsymbol{\theta} = \mathbf{X} \cdot \boldsymbol{\theta}$$

- So, the problem in terms of the above matrices is:

$$\text{Find } \boldsymbol{\theta} \text{ that minimizes } f(\boldsymbol{\theta}) = \frac{1}{n} \|\mathbf{y} - \mathbf{X} \cdot \boldsymbol{\theta}\|^2$$

- How do we solve such an optimization problem?

1. Use calculus to find a close-form solution for $\boldsymbol{\theta}$.
2. Use numerical hill-climbing technique such as **gradient ascent/descent**.

- Close-form solution:

- $f(\boldsymbol{\theta})$ is a nice convex function. So, it has a unique minima.
- Idea from calculus: Stack up $\frac{\partial f(\boldsymbol{\theta})}{\partial \theta_0}, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_1}, \dots, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_m}$ as a vector and equate it to the $\mathbf{0}$ vector.
- This gives: $(\mathbf{X}^T \mathbf{X}) \cdot \boldsymbol{\theta} = \mathbf{X}^T \mathbf{y}$ which gives
$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

$$\text{Find } (\theta_0, \theta_1, \dots, \theta_m) \text{ that minimizes } \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:

$$\theta = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$$

$$y = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \theta = X \cdot \theta$$

- So, the problem in terms of the above matrices is:

$$\text{Find } \theta \text{ that minimizes } f(\theta) = \frac{1}{n} \|y - X \cdot \theta\|^2$$

- How do we solve such an optimization problem?
 1. Use calculus to find a close-form solution for θ .
 2. Use numerical hill-climbing technique such as **gradient ascent/descent**.

• Gradient Descent:

- Start with some initial value of θ
- Repeat until convergence:
 - Find the direction along which the function $f(\theta)$ decreases the most. This happens to be $-\Delta f(\theta)$. So, we update $\theta = \theta - \alpha \cdot \Delta f(\theta)$, where α is the **learning rate**. So, $\theta = \theta + (2\alpha/n) \cdot X^T(y - X \cdot \theta)$

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

$$\text{Find } (\theta_0, \theta_1, \dots, \theta_m) \text{ that minimizes } \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:

$$\theta = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$$

$$y = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \theta = X \cdot \theta$$

- So, the problem in terms of the above matrices is:

$$\text{Find } \theta \text{ that minimizes } f(\theta) = \frac{1}{n} \|y - X \cdot \theta\|^2$$

- How do we solve such an optimization problem?
 - Use calculus to find a close-form solution for θ .
 - Use numerical hill-climbing technique such as **gradient ascent/descent**.

Gradient Descent:

- Start with some initial value of θ
- Repeat until convergence:
 - Find the direction along which the function $f(\theta)$ decreases the most. This happens to be $-\Delta f(\theta)$. So, we update $\theta = \theta - \alpha \cdot \Delta f(\theta)$, where α is the **learning rate**. So, $\theta = \theta + (2\alpha/n) \cdot X^T (y - X \cdot \theta)$

Variants:

- Batch GD: *Use all the data*
- Stochastic GD (SGD): *Use single data point*
- Mini-batch GD: *Use small batches*

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: *How do we formulate the notion of the best predictor?*

- Find θ that minimizes $f(\theta) = \frac{1}{n} (\mathbf{y} - X \cdot \theta)^2$

- Question: *How do we evaluate the prediction?*

- **Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
- **R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - Interpretation: How well does the model explain the variance in the response variable.

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: *How do we formulate the notion of the best predictor?*

- Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$

- Question: *How do we evaluate the prediction?*

- **Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
- **R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - Interpretation: How well does the model explain the variance in the response variable.
- **Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.
 - Interpretation: This measure penalizes the models that uses many features without improving prediction accuracy.

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: *How do we formulate the notion of the best predictor?*
 - Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$
- Question: *How do we evaluate the prediction?*
 - **Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
 - **R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - **Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.
- Question: *How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?*

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: *How do we formulate the notion of the best predictor?*

- Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$

- Question: *How do we evaluate the prediction?*

- **Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
- **R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
- **Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.

- Question: *How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?*

- Balance explainability (R^2) and complexity (# predictors).
- Guidelines:
 - If the R^2 does not improve too much on addition of new predictors, then avoid adding this new predictor
 - If two variables are correlated, then using both will not improve accuracy.
 - Forward selection: Start with highly correlated variable and add more variables as long as R^2 improves significantly.

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: *How do we formulate the notion of the best predictor?*
 - Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$
- Question: *How do we evaluate the prediction?*
 - **Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
 - **R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - **Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.
- Question: *How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?*
 - Balance explainability (R^2) and complexity (# predictors).
- Question: *What do we do if some predictors are categorical variables (i.e., non-numerical)?*

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: *How do we formulate the notion of the best predictor?*
 - Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$
- Question: *How do we evaluate the prediction?*
 - **Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
 - **R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - **Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.
- Question: *How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?*
 - Balance explainability (R^2) and complexity (# predictors).
- Question: *What do we do if some predictors are categorical variables (i.e., non-numerical)?*
 - Use **one-hot encoding**: Instead of one variable with 3 possible categories, use three 0/1 variables.
 - Sometimes it may even make sense to use discrete numerical variables as categorical variables.

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the best predictor?

- Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - X \cdot \theta)^2$

- Question: How do we evaluate the prediction?

- **Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
- **R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
- **Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.

- Question: How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?

- Balance explainability (R^2) and complexity (# predictors).

- Question: What do we do if some predictors are categorical variables (i.e., non-numerical)?

- Use **one-hot encoding**: Instead of one variable with 3 possible categories, use three 0/1 variables.

- Question: What are the other prediction methods using multiple attributes?

- **Nearest Neighbor Regression**: Take average of y values of k nearest neighbors
- **Polynomial Regression**: Fit a multivariate polynomial of degree > 1

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the best predictor?

- Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$

- Question: How do we evaluate the prediction?

- **Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
- **R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
- **Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n-p-1)}{\sum_i (y_i - \bar{y})^2 / (n-1)}$, where p is the number of predictor variables.

- Question: How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?

- Balance explainability (R^2) and complexity (# predictors).

- Question: What do we do if some predictors are categorical variables (i.e., non-numerical)?

- Use **one-hot encoding**: Instead of one variable with 3 possible categories, use three 0/1 variables.

- Question: What are the other prediction methods using multiple attributes?

- **Nearest Neighbor Regression**: Take average of y values of k nearest neighbors
- **Polynomial Regression**: Fit a multivariate polynomial of degree > 1

- Question: A polynomial with higher degree will always have better accuracy (i.e., low MSE). Does it mean that we should always use higher-degree polynomials for regression?

- **Bias-Variance Tradeoff**: Underfitting and overfitting negatively impact prediction. We need ways to find the "right fit"

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the best predictor?
 - Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$
- Question: How do we evaluate the prediction?
 - Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
 - R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.
- Question: How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?
 - Balance explainability (R^2) and complexity (# predictors).
- Question: What do we do if some predictors are categorical variables (i.e., non-numerical)?
 - Use **one-hot encoding**: Instead of one variable with 3 possible categories, use three 0/1 variables.
- Question: What are the other prediction methods using multiple attributes?
 - Nearest Neighbor Regression**: Take average of y values of k nearest neighbors
 - Polynomial Regression**: Fit a multivariate polynomial of degree > 1
- Question: A polynomial with higher degree will always have better accuracy (i.e., low MSE). Does it mean that we should always use higher-degree polynomials for regression?
 - Bias-Variance Tradeoff**: Underfitting and overfitting negatively impact prediction. We need ways to find the "right fit"

- Picking a very high-degree polynomial results in **overfitting**, where the model fits even the noise in the training data. This results in poor prediction on new data points.
- Question: How do we pick the right degree/model?

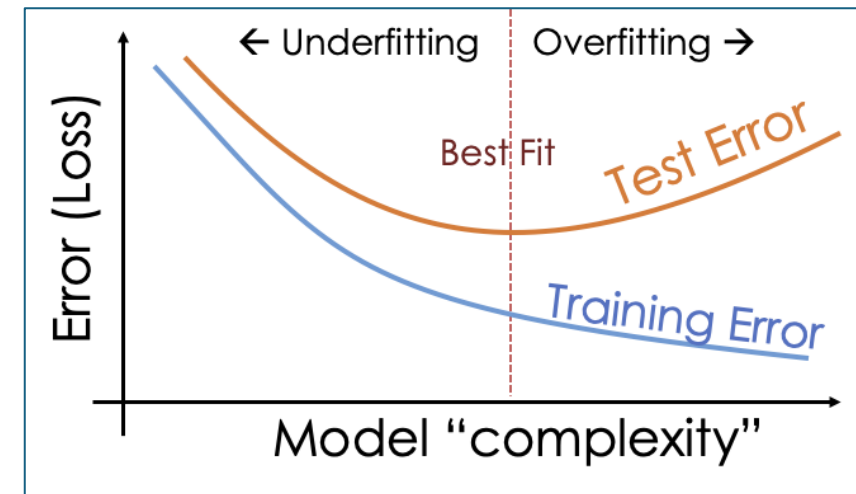
Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the best predictor?
 - Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$
- Question: How do we evaluate the prediction?
 - Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
 - R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.
- Question: How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?
 - Balance explainability (R^2) and complexity (# predictors).
- Question: What do we do if some predictors are categorical variables (i.e., non-numerical)?
 - Use **one-hot encoding**: Instead of one variable with 3 possible categories, use three 0/1 variables.
- Question: What are the other prediction methods using multiple attributes?
 - Nearest Neighbor Regression**: Take average of y values of k nearest neighbors
 - Polynomial Regression**: Fit a multivariate polynomial of degree > 1
- Question: A polynomial with higher degree will always have better accuracy (i.e., low MSE). Does it mean that we should always use higher-degree polynomials for regression?
 - Bias-Variance Tradeoff**: Underfitting and overfitting negatively impact prediction. We need ways to find the "right fit"

- Picking a very high-degree polynomial results in **overfitting**, where the model fits even the noise in the training data. This results in poor prediction on new data points.
- Question: How do we pick the right degree/model?
 - Main idea: Split the training data into a **training set** (~75-90%) and a **test set** (~10-25%). Train various models on the training set and evaluate on the test set.



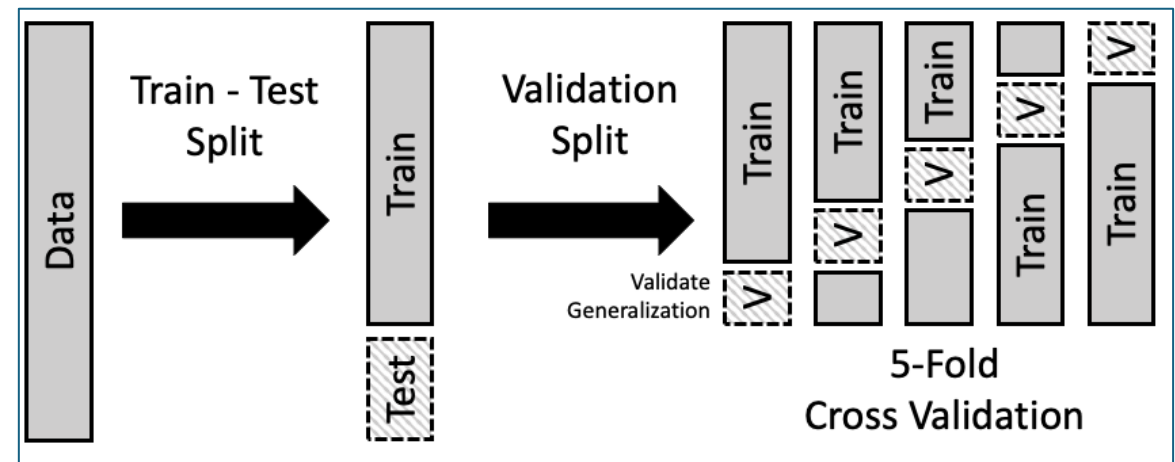
Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the best predictor?
 - Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$
- Question: How do we evaluate the prediction?
 - Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
 - R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.
- Question: How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?
 - Balance explainability (R^2) and complexity (# predictors).
- Question: What do we do if some predictors are categorical variables (i.e., non-numerical)?
 - Use **one-hot encoding**: Instead of one variable with 3 possible categories, use three 0/1 variables.
- Question: What are the other prediction methods using multiple attributes?
 - Nearest Neighbor Regression**: Take average of y values of k nearest neighbors
 - Polynomial Regression**: Fit a multivariate polynomial of degree > 1
- Question: A polynomial with higher degree will always have better accuracy (i.e., low MSE). Does it mean that we should always use higher-degree polynomials for regression?
 - Bias-Variance Tradeoff**: Underfitting and overfitting negatively impact prediction. We need ways to find the "right fit"

- Picking a very high-degree polynomial results in **overfitting**, where the model fits even the noise in the training data. This results in poor prediction on new data points.
- Question: How do we pick the right degree/model?
 - Main idea: Split the training data into a **training set** (~75-90%) and a **test set** (~10-25%). Train various models on the training set and evaluate on the test set.
 - Issue: We cannot use the test set to pick the right model
 - Solution: **Cross-validation**



Multiple Linear Regression

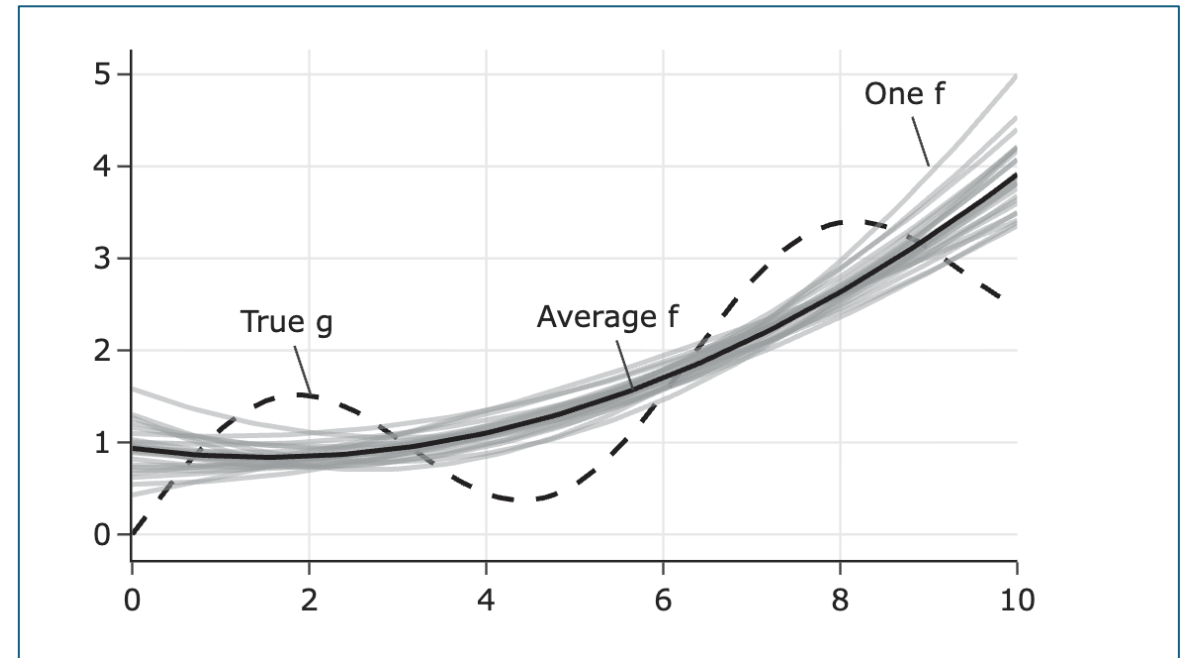
- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the best predictor?
 - Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$
- Question: How do we evaluate the prediction?
 - Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
 - R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.
- Question: How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?
 - Balance explainability (R^2) and complexity (# predictors).
- Question: What do we do if some predictors are categorical variables (i.e., non-numerical)?
 - Use **one-hot encoding**: Instead of one variable with 3 possible categories, use three 0/1 variables.
- Question: What are the other prediction methods using multiple attributes?
 - Nearest Neighbor Regression**: Take average of y values of k nearest neighbors
 - Polynomial Regression**: Fit a multivariate polynomial of degree > 1
- Question: A polynomial with higher degree will always have better accuracy (i.e., low MSE). Does it mean that we should always use higher-degree polynomials for regression?
 - Bias-Variance Tradeoff**: Underfitting and overfitting negatively impact prediction. We need ways to find the "right fit"

- Question: How do we pick the right degree/model? (**Cross-validation**)
- Question: What is the **bias-variance tradeoff**, and how does it impact model selection?

$$y = g(x) + \varepsilon$$



Expected error at $x = Bias^2 + Variance + noise$

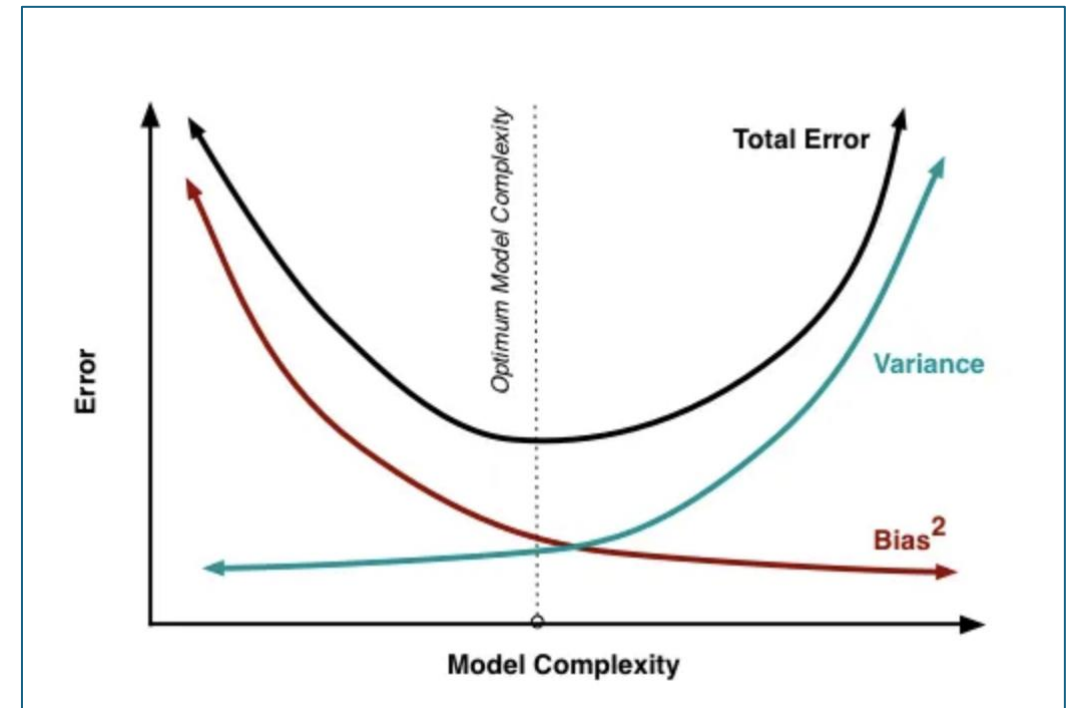
Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the best predictor?
 - Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$
- Question: How do we evaluate the prediction?
 - Root Mean Squared Error (RMSE):** $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
 - R-squared:** $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - Adjusted R^2 :** $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.
- Question: How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?
 - Balance explainability (R^2) and complexity (# predictors).
- Question: What do we do if some predictors are categorical variables (i.e., non-numerical)?
 - Use **one-hot encoding**: Instead of one variable with 3 possible categories, use three 0/1 variables.
- Question: What are the other prediction methods using multiple attributes?
 - Nearest Neighbor Regression:** Take average of y values of k nearest neighbors
 - Polynomial Regression:** Fit a multivariate polynomial of degree > 1
- Question: A polynomial with higher degree will always have better accuracy (i.e., low MSE). Does it mean that we should always use higher-degree polynomials for regression?
 - Bias-Variance Tradeoff:** Underfitting and overfitting negatively impact prediction. We need ways to find the "right fit"

- Question: How do we pick the right degree/model? (**Cross-validation**)
- Question: What is the **bias-variance tradeoff**, and how does it impact model selection?
 - Bias decreases, and variance increases as model complexity increases. Use cross-validation to find the sweet spot.



Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the best predictor?
 - Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2$
- Question: How do we evaluate the prediction?
 - **Root Mean Squared Error (RMSE)**: $RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$
 - **R-squared**: $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the target variable.
 - **Adjusted R^2** : $Adjusted - R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$, where p is the number of predictor variables.
- Question: How do we pick the predictor variables? Does it make sense to pick all available attributes as predictors?
 - Balance explainability (R^2) and complexity (# predictors).
- Question: What do we do if some predictors are categorical variables (i.e., non-numerical)?
 - Use **one-hot encoding**: Instead of one variable with 3 possible categories, use three 0/1 variables.
- Question: What are the other prediction methods using multiple attributes?
 - **Nearest Neighbor Regression**: Take average of y values of k nearest neighbors
 - **Polynomial Regression**: Fit a multivariate polynomial of degree > 1
- Question: A polynomial with higher degree will always have better accuracy (i.e., low MSE). Does it mean that we should always use higher-degree polynomials for regression?
 - **Bias-Variance Tradeoff**: Underfitting and overfitting negatively impact prediction. We need ways to find the "right fit"

- Question: How do we pick the right degree/model? (**Cross-validation**)
- Question: What is the **bias-variance tradeoff**, and how does it impact model selection?
 - Bias decreases, and variance increases as model complexity increases. Use cross-validation to find the sweet spot.
- Question: Can we model complexity with a single parameter?
 - **Ridge Regression**: Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2 + \lambda \cdot \|\theta\|_2^2$
 - **Lasso Regression**: Find θ that minimizes $f(\theta) = \frac{1}{n}(\mathbf{y} - \mathbf{X} \cdot \theta)^2 + \lambda \cdot \|\theta\|_1$
 - λ is the regularization **hyperparameter** tuned through cross-validation.
 - Other hyperparameters: k in kNN, learning rate η

End