- Module 4 -
Linear Methods for Regression

Outline

- Linear Regression Models
  - Single Feature
  - Multiple Features
  - Input Transformations
  - Polynomial Regression
  - Model Complexity

Recap

Supervised Learning: Regression (1)

- The main goal of supervised learning is to learn a model from labeled training data that allows for predictions on new unseen data. The term supervised refers to a set of samples where the desired labels are known

- Given a training set of N example (labeled) input-output pairs

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

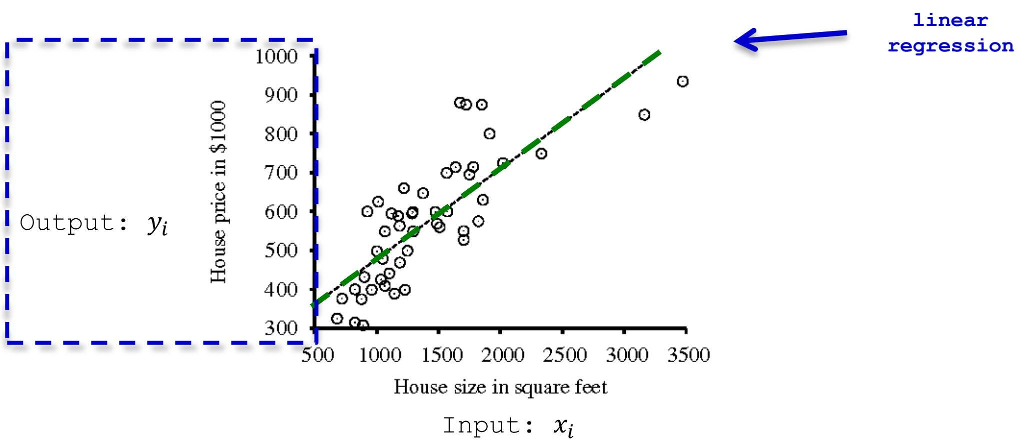
where x and y can be any value and each y_i was generated by an unknown function $y_i = f(x_i)$

➔ the goal is to discover a hypothesis h that approximates the true function f

Recap

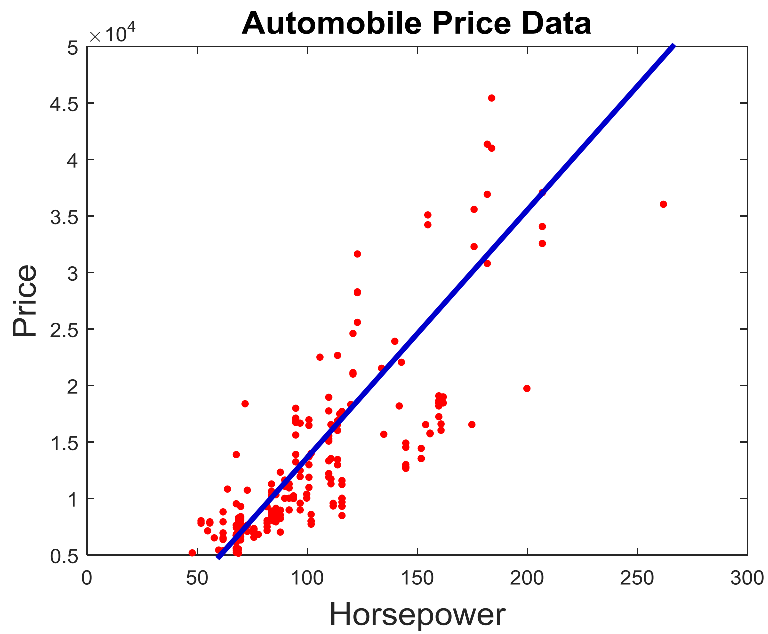
Supervised Learning: Regression (2)

- Regression is a subcategory of supervised learning where the goal is the prediction of continuous outcomes. In regression, given a number of features, p, and a continuous outcome, the objective is to find a relationship (a function $f: \mathbb{R}^p \rightarrow \mathbb{R}$) between those features to predict an outcome



Linear Regression Models (1)

- The data points shown were generated by the true function f plus observation noise



- To discover the hypothesis h, recall that the equation of a straight line is $y = mx + b$, where m is the slope and b is the y-intercept

- **Question:** How do you define a **linear regression model** to predict the price of a car given its horsepower?

LinearRegression.fit, LinearRegression.predict

Linear Regression Models (2)

- A linear regression model assumes that the **regression function** $E(Y|X)$ is linear in the inputs $X_1, \dots, X_p$

  ➔ i.e., $Y = \beta_1 X_1 + \cdots + \beta_j X_j, \beta_j \in \mathbb{R}, \forall j = 1, 2, \dots, p$

  (linear regression uses a linear equation to predict the output Y given a set of inputs X)

  - simple and often provides an adequate and interpretable description of how the inputs affect the outputs
  - for prediction purposes, they can sometimes outperform **nonlinear models**, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data
  - linear methods can be applied to **transformations** of the inputs and considerably expand their scope

- The linear model fit by **least squares** method is a prediction method that has been a mainstay of statistics and remains one of the most important tools

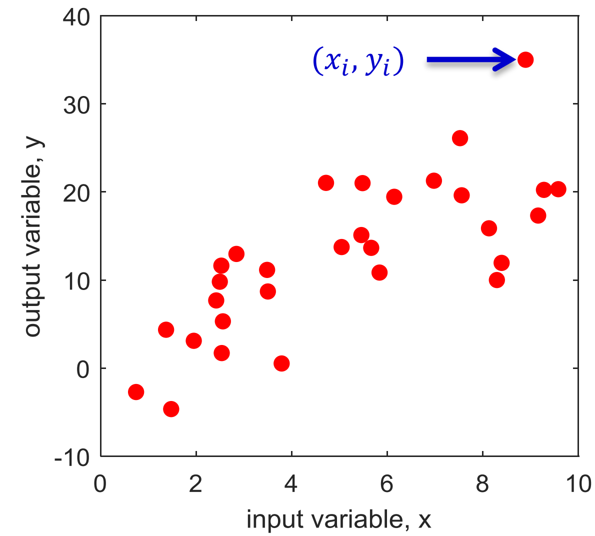
Linear Regression Models: Single Feature (3)

- Linear regression with one input (or feature)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad \forall i = 1, 2, \dots, N$$

where

N	number of samples in the data
$\hat{\beta}_0, \hat{\beta}_1$	weights (or **parameters** or **coefficients**)
x_i	input variable of the i-th sample
$\hat{y}_i$	predicted output of the i-th sample



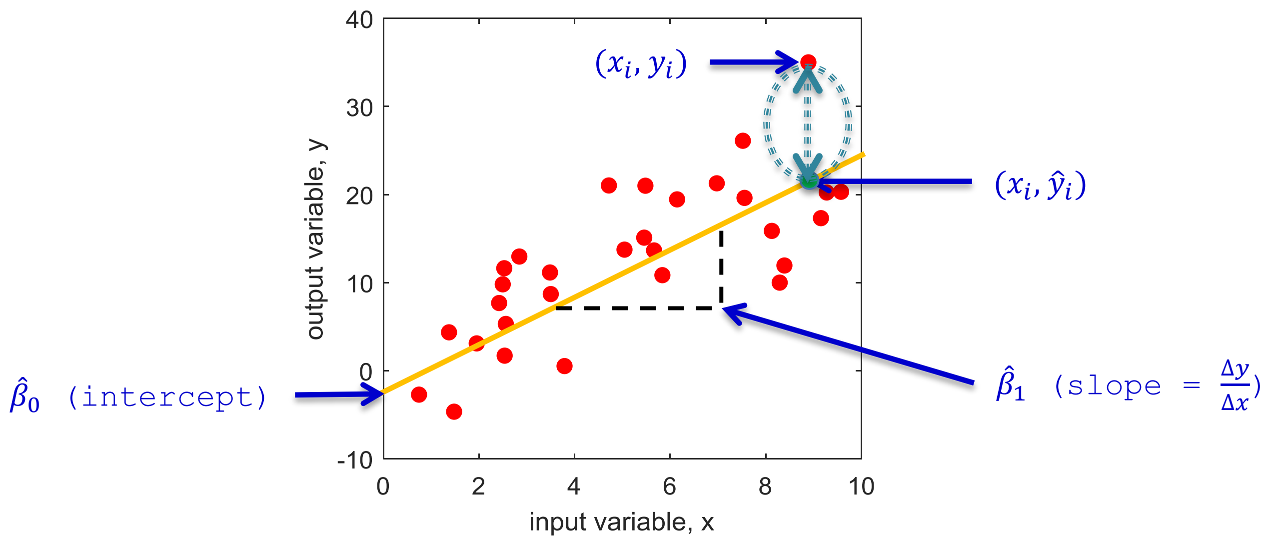
Goal: Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the distance between $\hat{y}_i$ and y_i is minimized for all i

Linear Regression Models: Single Feature (4)

- <u>Minimize</u> **residual sum of squares** (RSS), where

$$RSS\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \sum_{i=1}^{N}(y_i - \widehat{y}_i)^2$$

$$= \sum_{i=1}^{N}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2, \text{ where } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$



Linear Regression Models: Multiple Features (5)

- Linear regression with multiple inputs (or features)

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}\,, \qquad \forall i = 1, 2, \dots, N$$

where

N	number of samples in the data
p	number of features being modeled
$x_{i,j}$	j-th feature of the input variable of the i-th sample
$\hat{\beta}_j$	weight that determines how the j-th feature affects the prediction
$\hat{y}_i$	predicted output of the i-th sample

Linear Regression Models: Multiple Features (6)

- Considering all N samples in the data, we can rewrite $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}$ in matrix notation as

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$$

where

$$\hat{\boldsymbol{y}} = \underbrace{\begin{array}{c} \text{Output} \\ \begin{bmatrix} \boxed{\hat{y}_1} \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} \end{array}}_{\boldsymbol{N \times 1}}, \quad \boldsymbol{X} = \underbrace{\begin{array}{c} \begin{array}{ccccc} \hat{\boldsymbol{\beta}}_{\boldsymbol{0}} & & \text{Sample} & & \text{Feature} \end{array} \\ \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{bmatrix} \end{array}}_{\boldsymbol{N \times (1+p)}} \text{ and } \hat{\beta} = \underbrace{\begin{array}{c} \text{Weight} \\ \begin{bmatrix} \boxed{\hat{\beta}_0} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} \end{array}}_{\boldsymbol{(1+p) \times 1}}$$

Linear Regression Models: Multiple Features (7)

- Using the method of least squares to fit the linear model to the set of training data, the residual sum of squares (RSS)

$$RSS(\hat{\beta}_0, \dots, \hat{\beta}_p) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{N} \left(y_i - (\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}) \right)^2,$$

where $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}$

in matrix notation

$$RSS(\hat{\beta}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}})$$

$$= (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\hat{\beta}),$$ where $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$

Linear Regression Models: Multiple Features (8)

Goal: Find $\hat{\beta}$ such that $RSS(\hat{\beta})$ is minimized

- This is obtained by differentiating $RSS(\hat{\beta})$ with respect to $\hat{\beta}$, yielding the following **normal equations**

$$\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\hat{\beta}$$

- If $\boldsymbol{X}^T\boldsymbol{X}$ is invertible (or **nonsingular**), the unique solution is given by

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\hat{\beta}$$

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Least Squares Estimate

- Given a new input, x_{new}, the predicted value, $\hat{y}_{new}$, is given by

$$\hat{y}_{new} = \begin{bmatrix} 1 \\ x_{new} \end{bmatrix}^T \hat{\beta}$$

Linear Regression Models: Input Transformations (9)

- The inputs (or features), $X_1, \ldots, X_p$, can come from different sources
  - **quantitative** inputs
  - **transformations** of quantitative inputs, such as log, square, etc.
  - **numeric coding** of qualitative inputs

    (e.g., *fuel-type* feature is a 3-level qualitative input: gas, diesel and electric; encoded by X_{gas}, X_{diesel}, $X_{electric}$, where $X_{gas} = 1$ if the car uses gas; $X_{gas} = 0$ otherwise)
  - **interactions** between inputs, $X_3 = X_1 \cdot X_2$

    (e.g, X_3 = highway-mpg × city-mpg of a car)
  - **basis-expansions** such as $X_2 = X_1^2$, $X_3 = X_1^3$, leading to a **polynomial** representation

    (e.g., X_1 = highway-mpg of a car, X_2 = highway–mpg^2)
- Regardless of the source of X_j, the model is linear in the parameters β

Linear Regression Models: Polynomial Regression (10)

- **Polynomial regression** is a special case of the linear regression model in which the relationship between x and y is modelled as a p-th degree polynomial in x

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \cdots + \hat{\beta}_p x_i^p, \quad \forall i = 1, 2, \ldots, N$$

- Considering all N samples in the data, we can rewrite in matrix notation as

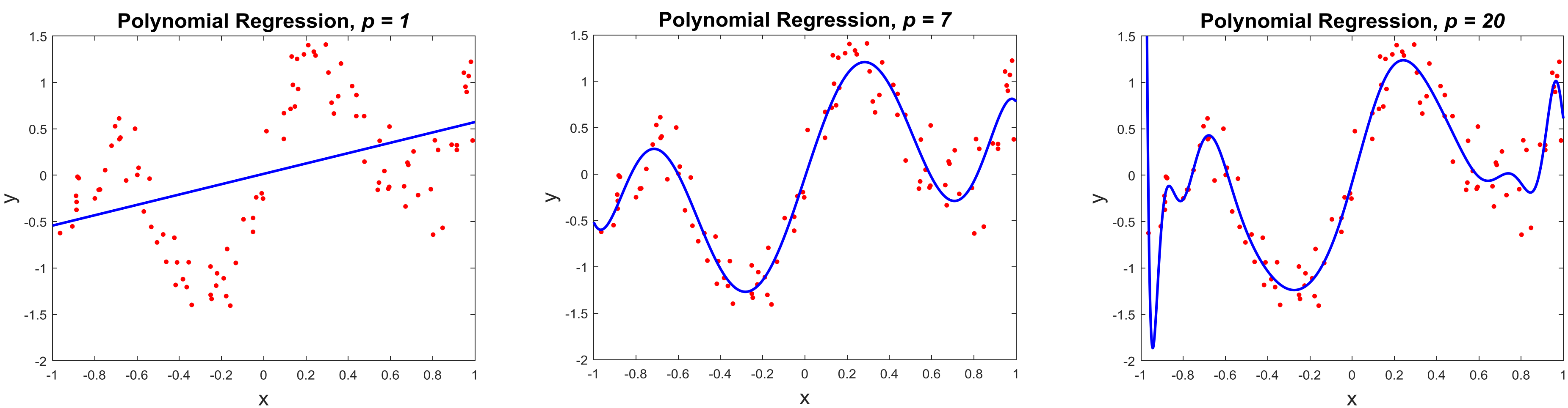
$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$$

where $\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \ldots & x_N^p \end{bmatrix}$ and $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$

- The least squares estimate for polynomial regression model remains as $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

Linear Regression Models: Model Complexity (11)

- **Model complexity** refers to the number of parameters used in the model and represents its ability to capture the patterns in the data
  - ➔ in polynomial regression models, model complexity is determined by p
- **Question:** What order of polynomial regression should be used to fit the following data?



➔ **Machine learning models** generally perform best when their complexity is appropriate for the true complexity of the task and the amount of training data provided