Lecture 10: Linear Regression

10.1 Statistical Learning (Supervised)

Example 10.1 Suppose that we want to analyze which factors can influence house price. We may come up with size of the house, the number of bedrooms, the number of bathrooms, the age of the house, so on and so forth.

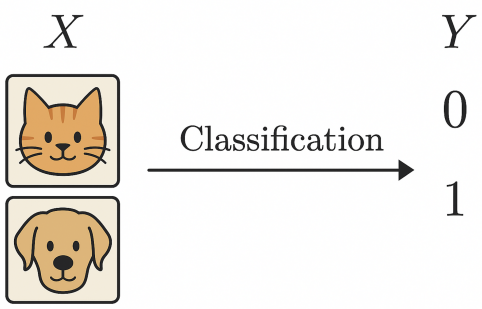
Now, based on how much we weighted the previous factors, we want to then *predict another* house price by its associated descriptive factors.

- Y: house price (in thousands of dollars) — we call it the **response** variable.
- X: a set of **predictor** variables that may influence house price.

Candidate predictors X might include:

- X_1: Size of the house (square feet)
- X_2: Number of bedrooms
- X_3: Number of bathrooms
- X_4: Age of the house (years)
- X_5: Distance to city center (km or miles)
- X_6: Lot size (square feet)
- X_7: Garage size (number of cars)
- X_8: Neighborhood quality (categorical score)
- $\vdots$

Example 10.2 Suppose that we want to distinguish between images of cats and dogs.



- Y: Binary response variable representing the class label:

$$Y = \begin{cases} 1 & \text{if the image is a dog} \\ 0 & \text{if the image is a cat} \end{cases}$$

- X: A set of input features extracted from the image.

Remark 10.1 From the examples above, it is natural to think that there exists some relationship between X and Y. In fact, X is typically chosen based on our intuitive understanding of the problem domain. More formally, we assume that the response variable Y is related to the predictor(s) X through some (possibly unknown) function f

$$Y = f(X, \epsilon),$$

where randomness, noise, or unmeasured factors in the system is explained by ϵ. The objective of statistical learning is to make the estimator $\hat{f} = \hat{f}(\mathcal{D})$, where $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ is data.

We refer to the task as a **classification problem** when the response variable Y is *categorical* or takes on a finite set of discrete values. On the other hand, if Y is a *continuous real-valued* variable, the task is known as a **regression problem**.

10.2 Simple linear regression

Definition 10.1 The simplest form of a functional relationship between X and Y is assumed to be *linear* with an *additive error*. Specifically, we model the function f as:

$$Y = f(X, \epsilon) = b_0 + b_1 X + \epsilon,$$

where b_0 is called the *intercept*, and b_1 is the *slope*. The slope b_1 has a natural interpretation: a one-unit increase in X is associated with an average change of b_1 units in Y. This is **simple linear regression**.

Example 10.3 Suppose we assume house price Y (in thousands of dollars) as a linear function of the size of the house X (in square feet)

$$Y = 50 + 0.2X$$

- $b_0 = 50$: This is the **intercept**. It represents the estimated price of a house with size zero (not meaningful in practice, but necessary for the model).
- $b_1 = 0.2$: This is the **slope**. It means that for each additional square foot of house size, the model predicts the price increases by 0.2 thousand dollars (i.e., \$200).

In more realistic settings, the data do not follow the linear equation exactly. Instead, there is stochastic variability around the linear relationship, as illustrated in the following figure: The

graph illustrates a plausible situation in which the red line captures the underlying association between X and Y well. The errors are centered around zero, meaning that the observed data points are not systematically biased upward or downward in terms of price. Suppose we do not know the true underlying system (e.g., $Y = 50 + 0.2X$), but we know that the system is linear (e.g., $Y = b_0 + b_1 X$) and are instead asked to infer the best-fitting line based solely on observed data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$. What should be our strategy for estimating this relationship?

10.2.1 Least square estimation

Definition 10.2 Given $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, to estimate the unknown relationship between X and Y, we assume a linear model of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where ε_i is the random error term for the ith observation. **Least Squares Estimation (LSE)** seeks to find the line that minimizes the total squared difference between the observed values Y_i and the values predicted by the model $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. This is done by minimizing the following objective function:

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2,$$

where RSS stands for the *Residual Sum of Squares*. Our goal is to calculate the values of $\hat{\beta}_0^{\text{LSE}}$ and $\hat{\beta}_1^{\text{LSE}}$ that minimize the RSS:

$$(\hat{\beta}_0^{\text{LSE}}, \hat{\beta}_1^{\text{LSE}}) = \arg\min_{\beta_0, \beta_1} \text{RSS}(\beta_0, \beta_1).$$

These estimates provide the best linear fit to the data in terms of minimizing squared error.

Remark 10.2 The following are the least squares estimates for β_0 and β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

How do we show this?

Proof. To find the values that minimize RSS, we take partial derivatives with respect to β_0 and β_1, set them to zero, and solve.

Step 1: Take partial derivative with respect to β_0:

$$\frac{\partial \text{RSS}}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) = 0$$

Step 2: Take partial derivative with respect to β_1:

$$\frac{\partial \text{RSS}}{\partial \beta_1} = -2 \sum_{i=1}^n X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

From Step 1:

$$\sum_{i=1}^n Y_i = n\beta_0 + \beta_1 \sum_{i=1}^n X_i$$

From Step 2:

$$\sum_{i=1}^n X_i Y_i = \beta_0 \sum_{i=1}^n X_i + \beta_1 \sum_{i=1}^n X_i^2$$

Solving for each β_0, β_1 will give the estimates. □

10.2.2 Probabilistic Background of Regression

Remark 10.3 We introduce several assumptions in order to interpret the model from a probabilistic perspective. Consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

The following assumptions are made.

1. **Linearity:** The conditional expectation of Y given X is a linear function.

$$\mathbb{E}[Y_i \mid X_i] = \beta_0 + \beta_1 X_i$$

2. **Independence:** The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent.

3. **Homoscedasticity (Constant Variance):** The error terms have the same variance:

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \text{for all } i$$

4. **Normality:** Each error term is normally distributed: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. This implies that the response variable Y_i is also normally distributed given X_i.

5. **Fixed Predictors:** The values $X_1, \ldots, X_n$ are considered fixed (non-random) in repeated samples.

Remark 10.4 Under the assumptions above, given the data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, the likelihood function for β_0, β_1, and σ^2 is given by

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right)$$

Alternatively, the log-likelihood is

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

Remark 10.5 We can easily calculate that maximum likelihood estimator for β_0 and β_1 from the following observation. For fixed σ^2,

$$\ell(\beta_0, \beta_1) = -\frac{n}{2}\log\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 = c_0 + c_1 \text{RSS}(\beta_0, \beta_1),$$

where $c_1 < 0$. Therefore, maximizing likelihood is the same as minimize RSS. Then, how do we get the MLE of σ^2? Take the partial derivative with respect to σ^2 and set to 0. Then once $\hat{\beta}_0$ and $\hat{\beta}_1$ have been obtained, plug them into your partial derivative equation, and the estimate of σ^2 is found to be the residual sum of squares:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Is it an unbiased estimator? Can we get unbiased estimator of σ^2? The following is unbiased.

$$\hat{\sigma}^2_{\text{UE}} = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Why is this unbiased? Under the normality assumption, the scaled RSS follows a chi-square dis-

tribution with $n-2$ degrees of freedom

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \sim \chi^2(n-2).$$

This result is critical for deriving the sampling distribution of the error variance estimator $\hat{\sigma}^2$.

Remark 10.6 We've shown the MLE of β_0, β_1 is equivalent to LSE of β_0, β_1. Are they unbiased?

Proof. **Unbiasedness of $\hat{\beta}_1$:** Since $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, then:

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i.$$

Using this in the formula for $\hat{\beta}_1$, we compute the expectation:

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}\right] = \beta_1,$$

since the numerator becomes a covariance between X and Y, and $\mathbb{E}[\bar{Y}] = \beta_0 + \beta_1\bar{X}$. □

Proof. **Unbiasedness of $\hat{\beta}_0$:** Using the result above

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{Y} - \hat{\beta}_1\bar{X}] = \mathbb{E}[\bar{Y}] - \bar{X}\cdot\mathbb{E}[\hat{\beta}_1] = (\beta_0 + \beta_1\bar{X}) - \bar{X}\cdot\beta_1 = \beta_0.$$

□

Theorem 10.1 *$\hat{\beta}_0^{\text{MLE}}, \hat{\beta}_1^{\text{MLE}}$ and $\hat{\sigma}^2_{\text{UE}}$ are UMVUEs of β_0, β_1, and σ^2.*