Regression

Goal: predict the value of a respons variable $Y \in \mathbb{R}$ based on a feature vector / predictor $X \in \mathbb{R}^k$

- For each fixed $X = x$, there is a probability distribution $Y|X = x$
- Not realistic to assume that there is a function $f(x)$ such that $Y = f(x)|X = x$
- We look for an f that minimizes the expected error $\mathbb{E}\left[(Y - f(X))^2\right]$
  - By the tower property of conditional expectation:

$$\mathbb{E}\left[(Y - f(X))^2\right] = \mathbb{E}\left[\mathbb{E}\left[(Y - f(X))^2 | X\right]\right]$$

  - To minimize this, we can now just minimize the inner expectation for every possible $X = x$:
    - In other words, each $f(x)$ will just be set to the value a that minimizes:

$$h(a) = \mathbb{E}\left[(Y - a)^2 | X = x\right]$$

    - We can set the derivative to 0, which gives that $a = \mathbb{E}\left[Y|X = x\right]$
    - Therefore, the minimizing function is $f(x) = \mathbb{E}\left[Y|X = x\right]$
      - This is known as the **regression function** of Y onto X
- We cannot perfectly compute this regression function
  - It also does not fully capture the distribution $Y|X = x$ since it only computes the mean
  - An alternative to vanilla regression is **quantile regression**, which givesa confidence band around the mean

Linear Regression

We make the assumption that $f(x) = \mathbb{E}\left[Y|X = x\right]$ is linear in form

- $f(x) = x^T\beta$ for some ground truth $\beta = \beta^* \in \mathbb{R}^k$
- To find an estimator, we need to assume a parametric form of the distribution of $Y|X = x$
  - We'll use the Gaussian distribution, so $Y|X = x \sim \mathcal{N}(x^T\beta^*, \sigma^2)$
  - This also has the assumption that the mean function is linear and that the variance is constant across all x
- Writing out the log likelihood and maximizing it gives:
  - $\hat{\beta}^{\text{MLE}} = \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n (Y_i - X_i^T\beta)^2$
- If the Y_i s are put in a column vector and each X_i is a row in the matrix $X \in \mathbb{R}^{n \times k}$, then we can write the closed form solution as:
  - $\hat{\beta} = (X^TX)^{-1}X^TY$
  - Another interpretation of this formula is to project Y onto the column space of X

Distribution of $\hat{\beta}$

To construct confidence intervals and test hypotheses about the ground truth coefficient vector β^*, we need to know the distribution of $\hat{\beta}$

- We can write $Y = X\beta^* + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
  - Substituting this in, we get: $\hat{\beta} = \beta^* + (X^TX)^{-1}X^T\varepsilon$
  - We then have that the distribution of the noise term is:

$$(X^TX)^{-1}X^T\varepsilon \sim \mathcal{N}(0, \sigma^2(X^TX^{-1}))$$

- The final distribution of $\hat{\beta}$ is finally then:

$$\hat{\beta} \sim \mathcal{N}(\beta^*, \sigma^2(X^TX)^{-1})$$

Confidence Intervals and Hypothesis Testing

- We find the distribution of $\hat{\beta}_j$ as:

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j^*, \sigma^2 (X^T X)^{-1}_{jj})$$

- This can give rise to a confidence interval
  - $X^T X$ has size n, so this takes the rule of scaling this with $\sqrt{n}$
- However, we don't know the true variance σ
  - We can estimate it by considering the residuals $\varepsilon_i = Y - X\hat{\beta}$
  - Since Y lives in dimension n and $X\hat{\beta}$ has dimension k, this means ε has dimension $n - k$
    - Then, these residuals have $n - k$ degrees of freedom, and we have:

$$||\varepsilon||^2 \approx (n-k)\sigma^2$$

    - We can therefore approximate σ^2 as:

$$\sigma^2 \approx \frac{||\varepsilon||^2}{n-k} = \frac{1}{n-k} \sum_{i=1}^{n} \varepsilon_i^2$$

- We can use these to construct both confidence intervals and hypothesis tests