Lecture 16

Linear Regression

Linear Regression is a supervised learning algorithm used to predict a continuous output (target variable) based on one or more input features.

The core idea is to model the relationship between inputs X and output y as a linear function:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where:

- y is dependent variable (target).
- $x_1, x_2, \ldots, x_n$ are independent variables (features).
- β_0 is the intercept (bias).
- $\beta_1, \beta_2, \ldots, \beta_n$ are coefficients (weights) for each feature.
- ϵ is the error term (noise).

Goal is to learn the coefficients β that best fit the data.

The "best fit" means the coefficients minimize the difference between actual and predicted values

Assumptions of Linear Regression

1. **Linearity**
    a. The relationship between each independent variable x_j and the dependent variable y is linear.
    b. Formally, the expected value of y given the predictors X is a linear function of the X:
    $$\mathbb{E}[y|X] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$
    c. This means no nonlinear transformations or interactions unless explicitly included.
    d. If this is violated, linear regression will not capture the true relationship well and may lead to biased estimates and poor predictions.
    e. Remedy: Transform variables (e.g., logarithm, polynomial terms) or use nonlinear models

2. **Independence**
    a. The observations (rows in the dataset) are independent of each other.
    b. The residuals (errors) ϵ_i should be independent across observations.
    c. Violation occurs in time series data (autocorrelation) or clustered data.
    d. Why important? If residuals are correlated, estimated standard errors are incorrect, leading to invalid hypothesis tests and confidence intervals.
    e. Remedy: Use time series models (ARIMA), mixed models or adjust for correlation.
3. **Homoscedasticity (Constant Variance of Errors)**
    a. The variance of the residuals (errors) ϵ_i is constant across all levels of predictors.

$$Var\ \epsilon_i\ = \sigma^2$$

    b. When this is true, residuals are said to be homoscedastic.
    c. If variance changes with input variables i.e. heteroscedasticity, it violates the assumption.
    d. Implications: Coefficient estimates are still unbiased but standard errors and test statistics become unreliable.
    e. Detection: Plot residuals vs fitted values; funnel shape indicates heteroscedasticity.
    f. Remedy: Transform dependent variable (e.g., log), weighted least squares, or robust standard errors.
4. **Normality of Errors**
    a. The residuals should be approximately normally distributed:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

    b. This assumption is mainly needed for valid hypothesis testing and constructing confidence intervals.
    c. If the sample size is large, by Central Limit Theorem, the distribution of estimates tends to normal even if errors are not.
    d. Detection: Use Q-Q plots or normality tests on residuals.
    e. Remedy: Apply transformations or use bootstrap methods for inference.

5. **No Perfect Multicollinearity**
    a. The independent variables should not be perfectly correlated with each other.
    b. Perfect multicollinearity means one predictor is an exact linear combination of others:

$$x_j = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{j-1} x_{j-1} + \alpha_j x_j + \cdots$$

    c. This makes the matrix $X^T X$ non-invertible, so the coefficients cannot be uniquely estimated.
    d. Detection: Very high correlation between predictors or variance inflation factor (VIF) > 10.
    e. Remedy: Remove or combine collinear variables, use dimensionality reduction (PCA).
6. **No Omitted Variable Bias**
    a. Important predictors should be included in the model.
    b. Omitting a variable that is correlated with both the dependent variable and other predictors causes biased estimates.
    c. Always try to include relevant variables based on theory or domain knowledge.

Cost Function in Linear Regression

A dataset with n data points: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x_i is the input (can be a vector) and y_i is the corresponding output.

A hypothesis function that gives predicted values $\hat{y} = h(x)$

Goal: Evaluate how well the hypothesis h fits the data.

- Compare predicted value $h(x_i)$ to actual output y_i for each data point.
- Define a distance function $d(\cdot,\cdot)$ that measures the discrepancy between prediction and actual value.
- The cost function $L(h)$ is the sum of these discrepancies over all data:

$$L(h) = \sum_{i=1}^{n} d(h(x_i), y_i)$$

- We want to find h (or the parameters defining h) that minimizes $L(h)$.
- The typical choice of distance is squared error:

$$d(h_\beta(x_i), y_i) = (y_i - x_i^T \beta)^2$$

- So, cost function (also called Residual Sum of Squares (RSS)) is:

$$L(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

Goal: Find parameter β that minimizes this cost.

In Matrix Form

- Let X be the $n \times p$ design matrix (rows x_i^T),
- y be the $n \times 1$ vector of targets,
- β be the $p \times 1$ vector of parameters.

∴ Cost Function:

$$L(\beta) = ||y - X\beta||_2^2 = (y - X\beta)^T (y - X\beta)$$

Minimizing β

$$\frac{\partial (y - X\beta)^T (y - X\beta)}{\partial \beta}$$

Expanding terms:

$$(y - X\beta)^T (y - X\beta) = y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta$$

Setting derivative to zero:

$$y^T y - 2\beta^T X^T y + \beta^T X^T X\beta = 0$$

$$-2X^T y + 2X^T X\beta = 0$$

$$X^T X\beta = X^T y$$

$$\widehat{\beta_{LS}} = (X^T X)^{-1} X^T y$$

This is the Least Squares estimator.

Maximum Likelihood Estimation (MLE) in Linear Regression

We want to maximize the likelihood $L(h) = P(Y|h)$ — the probability of observing data Y given model h.

Recall from previous assumptions:

- Errors: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- Model: $Y = X\beta + \epsilon$

Therefore, target output Y follows a multivariate normal distribution with mean $X\beta$ and variance σ^2

$$Y \sim \mathcal{N}(X\beta, \sigma^2 I)$$

The likelihood function for observed data y is:

$$P(y|X, \beta, \sigma^2) = \frac{1}{\sqrt{(2\pi)^n \sigma^{2n}}} \exp\left(-\frac{1}{2\sigma^2} \left||y - X\beta|\right|_2^2\right)$$

The exponential term captures how well parameters β fit the data based on squared residuals.

Maximizing likelihood $L(\beta)$ is equivalent to minimizing the negative log likelihood.

Since the likelihood is monotonically related to the negative squared error, maximizing likelihood simplifies to minimizing residual sum of squares:

$$\widehat{\beta_{MLE}} = \arg\max_{\beta} L(\beta) = \arg\min_{\beta} \left||y - X\beta|\right|$$

Remarkably, the MLE coincides with the least squares estimator in linear regression under Gaussian noise assumption.

$$\widehat{\beta_{MLE}} = (X^T X)^{-1} X^T y$$

Unbiasedness of the Least Squares Estimator

$\widehat{\beta_{LS}}$ is an **unbiased estimator** of the true parameter β

i.e.

$$\mathbb{E}\left[\widehat{\beta_{LS}}\right] = \beta$$

Recall:

$$\widehat{\beta_{LS}} = (X^T X)^{-1} X^T y$$

Taking expectation on both sides:

$$\mathbb{E}\left[\widehat{\beta_{LS}}\right] = \mathbb{E}[(X^T X)^{-1} X^T y] = (X^T X)^{-1} X^T \mathbb{E}[y]$$

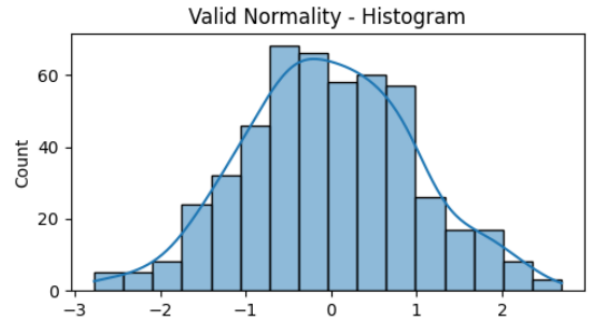
Since: $y = X\beta + \epsilon$, And noise ε has mean zero:

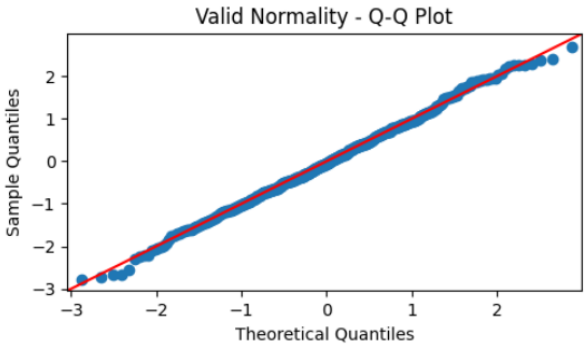
$$\mathbb{E}\left[\widehat{\beta_{LS}}\right] = (X^T X)^{-1} X^T X\beta = \beta$$

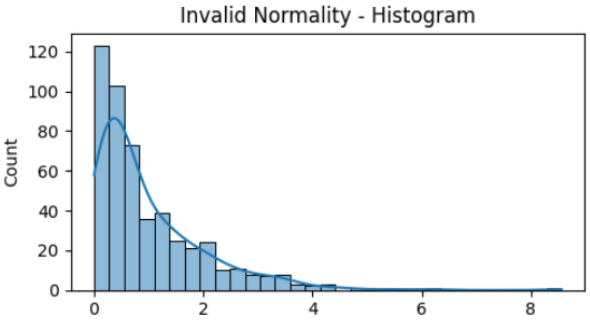
On average, over many samples from the same data generating process, the least squares estimator recovers the true parameter β.

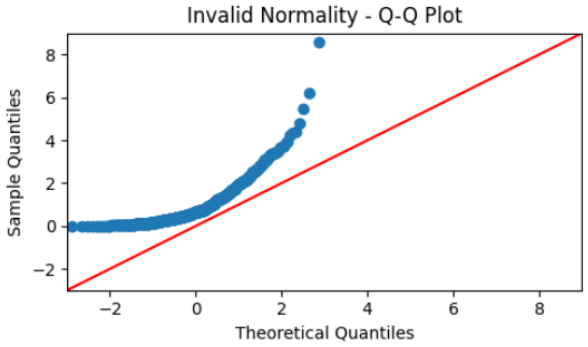
This property is important to ensure valid statistical inference.

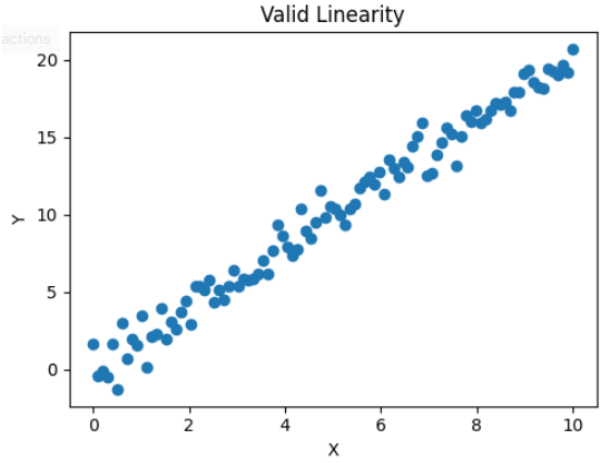
Visuals of Assumptions

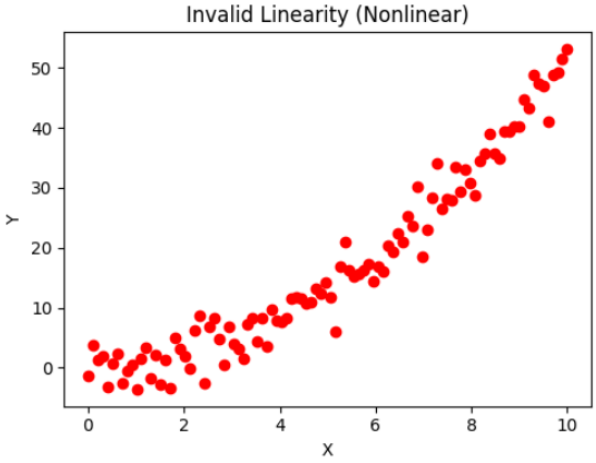


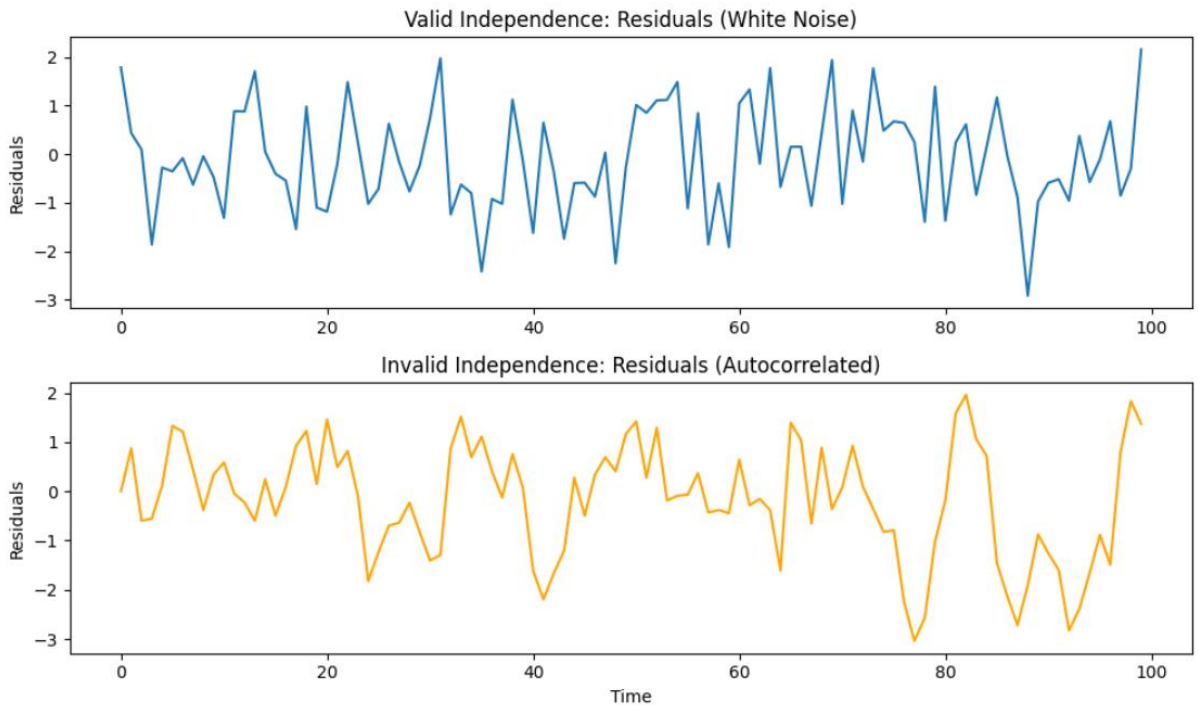












What to Look Out For

1. Normality (for parametric tests like t-test, regression residuals, etc.)

- **Valid**: Data or residuals roughly follow a bell-shaped (Gaussian) symmetric, unimodal distribution.
- **Invalid**: Strong skewness, heavy tails or multimodal distributions, strong outliers.

2. Linearity (in regression models)

- **Valid**: The relationship between independent and dependent variables looks approximately linear — scatterplots show a straight-line trend.
- **Invalid**: Curved, U-shaped, or complex nonlinear relationships.

3. Homoscedasticity (constant variance of errors)

- **Valid**: Residuals have roughly equal spread (variance) across all fitted values or levels of predictors. Residual plots show a cloud of points randomly scattered.
- **Invalid**: Residuals fan out (increasing variance) or funnel shape (heteroscedasticity), systematic changes in spread with predictor.

4. Independence

- **Valid**: No obvious pattern or correlation among observations/errors (e.g. in residual plots).
- **Invalid**: Patterns like autocorrelation in time-series data, or clustered/batched data violating independence.

5. Multicollinearity (in multiple regression)

- **Valid**: Predictor variables are not highly correlated.
- **Invalid**: Predictors strongly correlated — leads to unstable coefficient estimates.

6. Distribution Shapes in Naive Bayes Assumptions:

- Numerical features often assumed to be Gaussian-shaped per class for Gaussian Naive Bayes.
- If data per class is not bell-shaped (which may mean features are non-Gaussian, multimodal, highly skewed), this assumption is violated. Alternatives: kernel density estimations, or discretization.

7. Cluster Analysis or Mixture Models:

- Often assume data clusters are roughly spherical or elliptical.
- If clusters have irregular, elongated, or overlapping shapes, standard k-means or Gaussian mixture assumptions are violated and more flexible models might be needed.