**Linear Regression Study Guide**

**📊 What is Linear Regression?**

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It helps predict the value of the dependent variable based on the given independent variable(s).

There are two main types:

1. **Simple Linear Regression**: One independent variable
2. **Multiple Linear Regression**: More than one independent variable

**✏️ Equation of a Linear Model**

For **Simple Linear Regression**:

y = mx + b

* **y**: Dependent variable (output)
* **x**: Independent variable (input)
* **m**: Slope (how much y changes when x changes)
* **b**: Intercept (value of y when x = 0)

For **Multiple Linear Regression**:

y = b0 + b1\*x1 + b2\*x2 + ... + bn\*xn

* **b0**: Intercept
* **b1, b2, ..., bn**: Coefficients for each independent variable

**🔄 How Does It Work?**

Linear regression finds the best-fit line by minimizing the **Sum of Squared Errors (SSE)** between actual and predicted values. This is done using:

* **Ordinary Least Squares (OLS)** for exact calculation
* **Gradient Descent** for iterative optimization

**📊 Accuracy Metrics in Linear Regression**

To evaluate how good the regression model is, we use the following metrics:

**1. Mean Absolute Error (MAE)**

Average of the absolute errors between predicted and actual values.

MAE = (1/n) \* Σ |y - ŷ|

* ✅ **Advantages**:
  + Easy to understand and interpret
  + Not heavily affected by outliers
* ❌ **Disadvantages**:
  + Doesn’t penalize large errors as strongly as squared metrics

**2. Mean Squared Error (MSE)**

Average of squared differences between predicted and actual values.

MSE = (1/n) \* Σ (y - ŷ)^2

* ✅ **Advantages**:
  + Penalizes large errors more, useful when large mistakes are more serious
* ❌ **Disadvantages**:
  + Sensitive to outliers
  + Units are squared, which may be hard to interpret directly

**3. Root Mean Squared Error (RMSE)**

Square root of MSE. Brings error units back to original scale.

RMSE = √(MSE)

* ✅ **Advantages**:
  + Same units as the original data, making interpretation easier
  + Penalizes large errors
* ❌ **Disadvantages**:
  + Still sensitive to outliers like MSE
  + Can be harder to understand than MAE for beginners

**4. R-squared (R²)**

Explains how much of the variance in the dependent variable is explained by the model.

R² = 1 - (SSE / SST)

* ✅ **Advantages**:
  + Easy to interpret as a percentage
  + Indicates overall model fit
* ❌ **Disadvantages**:
  + Can be misleading with non-linear relationships
  + Doesn’t indicate whether predictions are biased or if model assumptions are met
  + Can increase with more variables even if they don’t improve the model

**🔍 Interpretation of R²**

* **0.90 or above**: Excellent fit
* **0.75 to 0.90**: Good fit
* **0.50 to 0.75**: Moderate fit
* **Below 0.50**: Poor fit

**🔬 Assumptions of Linear Regression**

1. **Linearity**: Relationship between X and Y is linear
2. **Independence**: Observations are independent of each other
3. **Homoscedasticity**: Constant variance of errors
4. **Normality**: Errors are normally distributed
5. **No multicollinearity** (for multiple regression): Independent variables are not highly correlated

**🎓 When to Use Linear Regression?**

* Predicting sales, prices, trends, etc.
* Understanding relationships between variables
* Forecasting based on historical data

**🤖 Final Tip**

Always visualize your data and residuals, and check assumptions before trusting the results. Use metrics to evaluate your model, and choose the simplest model that works well!

2

Happy Learning!