Data Science Notes

Bhavjot Khurana

November 5, 2025

Contents

1 Introduction

Using the freecodecamp.ord video as a reference (link: https://www.youtube.com/watch?v=XU5pw3QR
This document contains notes on data science topics covered in the video.

2 Topic 1: Linear Regression

2.1 Definition

Linear regression models the expected value of a response variable y as a linear function of a single predictor x:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{Normal}(0, \sigma^2)$$

It assumes the errors ε_i are independent, normally distributed, and have constant variance.

2.2 Key Formulas

Ordinary least squares (OLS) estimates the slope and intercept by minimizing the sum of squared residuals:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

2.3 Error Function

The residual for observation i captures the prediction error:

$$e_i = y_i - \hat{y}_i$$

Residuals tell us how far each point lies from the regression line; examining their pattern helps spot outliers or violations of model assumptions.

2.4 Mean Squared Error (MSE)

Mean Squared Error is the average of squared residuals:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Errors are squared to penalize large deviations more heavily, eliminate sign cancellation, and produce a smooth, differentiable loss function that calculus-based solvers can optimize.

2.5 How the Line is Fitted

Fitting the line involves solving the optimization problem:

$$\min_{\beta_0,\beta_1}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Setting the partial derivatives to zero yields the normal equations (matrix form: $(X^\top X)\hat{\beta} = X^\top y$). Software uses closed-form solutions for small problems or matrix decompositions (QR or SVD) for numerical stability. Gradient methods (e.g., gradient descent) offer scalable alternatives for very large datasets.

2.6 Hypothesis Testing and p-values

To test whether x helps explain y, use the t-test for the slope. The null hypothesis $H_0 : \beta_1 = 0$ indicates no linear relationship. The statistic

$$t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$$

follows a t-distribution with $n - 2$ degrees of freedom. A small p-value suggests the predictor provides statistically significant explanatory power.

2.7 Residual Standard Error (RSS and TSS)

Two sums of squares quantify variability:

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad \mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Residual Standard Error (RSE) estimates the typical size of residuals in the units of y:

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}$$

Lower RSE indicates tighter fit around the regression line.

2.8 Interpretations

- **Slope** $\hat{\beta}_1$: Expected change in y for a one-unit increase in x.
- **Intercept** $\hat{\beta}_0$: Expected value of y when $x = 0$, useful when $x = 0$ is meaningful.
- $R^2 = 1 - \frac{\textbf{RSS}}{\textbf{TSS}}$: Proportion of variance in y explained by the model.
- **Practical use**: Combine coefficient estimates with confidence or prediction intervals to communicate both central tendency and uncertainty in predictions.

3 Topic 2: Multiple Linear Regression

3.1 Definition

Multiple Linear Regression (MLR) extends the linear model to p predictors:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Coefficients measure the expected change in y for a one-unit change in the corresponding predictor while holding others constant.

3.2 F-statistic and Interpretation

The model-wide F-test compares the explained variance to unexplained variance:

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}$$

A large F-statistic (with a small p-value) indicates that, taken together, the predictors explain significantly more variation than an intercept-only model. If the F-test is not significant, it implies the collective set of predictors may not offer meaningful predictive power beyond the mean of y.

4 Topic 3: Machine Learning

4.1 Supervised Learning

Explain supervised learning here.

4.2 Unsupervised Learning

Explain unsupervised learning here.

5 Conclusion

Summarize your notes here.