Notes

Nathan Jung

September 19, 2024

Study linear regression assumptions and why
Study differences between supervised and unsupervised learning
What is bias-variance tradeoff
Study the metrics used by classification

1 Linear Regression

1.1 Simple Linear Regression

The basic form for linear regression is as shown:

$$Y \approx \beta_1 + \beta_0 X + \epsilon$$

The form used for prediction is as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

1.1.1 How do we estimate the coefficients?

The most common approach is *least squares*. This is done by taking the ith *residual* and finding the right values for β_0 and β_1 that will minimize the residual sum of squares:

$$\text{RSS} = e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where individual e_i is equal to:

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

1.1.2 Accuracy of the coefficient estimates

Within each linear regression formula, there is generally some random noise that will never allow the equation to fully capture the real world relationship.

We want an unbiased estimator (from the samples we take), meaning that it does not over- or under-estimate the true parameter. Therefore, theoretically if we average all the estimates we get of $\beta_0 \text{and} \beta_1$ over a large number of data sets, the average of the estimates should be spot on with the population.

To represent how far off a single estimate will be from the true estimate, we use the formula:

$$\text{Var}(\hat{\mu}) = \text{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

1.1.3 Accuracy of the model

Quality of linear regression fit is usually assessed with $residualstandarderror$ (RSE) and R^2 statistic.

RSE is an estimate of the standard deviation of σ, i.e. it is the average amount that the response will deviate from the true regression line. It uses the formula:

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RSE is considered a measure of the $lackoffit$ of the model to the data. RSE is small if $\hat{y}_i \approx y_i$ for all i, and big if there is a large residual. If you get a value of 3.26, then the actual y value deviate from true regression by 3.26 units on average. However, it is measured in units of Y so not always clear what makes a good RSE.

R^2 Statistic takes the form of a *proportion* of variance explained. Independent of Y

$$R^2 = \frac{\text{TSS - RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum (y_i - \bar{y})^2$