Contents

CHAPTER 1

Evaluating the Fit of a Linear Regression Model

The fit of a linear regression model can be evaluated using several statistical metrics. Three common ones include the residuals, the coefficient of determination (R-squared or R^2), and the root mean squared error (RMSE).

1.1 Residuals

Residuals are the differences between the observed and predicted values. For an observation i, the residual e_i is calculated as

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

where y_i is the observed value and $\hat{y}_i$ is the predicted value. By plotting these residuals against the predicted values, we can visually inspect the model's fit. Ideally, the residuals should be randomly scattered around zero, and there should be no clear pattern in the residual plot.

1.2 R-Squared (R^2)

R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that can be predicted from the independent variables. It provides a measure of how well the model's predictions fit the data. R^2 is calculated as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares. An R^2 value of 1 indicates a perfect fit, while an R^2 of 0 indicates that the model does not explain any of the variability of the response data around its mean.

1.3 Root Mean Squared Error (RMSE)

RMSE is a frequently used measure of the differences between the values predicted by a model and the values actually observed. It's the square root of the average of squared differences between prediction and actual observation. RMSE is calculated as

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

A lower RMSE indicates a better fit to the data.

It's important to note that these metrics should not be used in isolation to evaluate the model's fit. They should be used in combination along with the understanding of the underlying problem and domain knowledge.

CHAPTER 2

Linear Regression and Gradient Descent

Linear regression models can be fitted using an optimization algorithm known as gradient descent. This is especially useful when the number of features is large, making the normal equation computationally expensive.

In gradient descent, we start with an initial guess for the model parameters and iteratively update these parameters to minimize the cost function, which is usually the mean squared error (MSE) for linear regression. For a linear regression model with parameters θ, the update rule is given by

$$\theta := \theta - \alpha \nabla J(\theta)$$

where α is the learning rate and $\nabla J(\theta)$ is the gradient of the cost function evaluated at θ. For MSE, the gradient is given by

$$\nabla J(\theta) = \frac{2}{n} \mathbf{X}^T (\mathbf{X}\theta - \mathbf{y})$$

where $\mathbf{X}$ is the feature matrix, $\mathbf{y}$ is the vector of target values, and n is the number of observations.

2.1 Standardizing Inputs

Standardizing inputs can improve the performance of gradient descent. By ensuring all features have a similar scale, we can avoid a situation where the cost function has a very elongated shape, causing gradient descent to take a long time to converge.

More specifically, standardization transforms the features so they have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean and dividing by the standard deviation for each feature:

$$x_i' = \frac{x_i - \mu_i}{\sigma_i}$$

where x_i is a feature vector, and μ_i and σ_i are its mean and standard deviation, respectively.

By standardizing the inputs, each feature contributes approximately proportionately to the final distance, helping the gradient descent algorithm converge more quickly and efficiently.

Chapter 3

Generalized Linear Models

In statistics, generalized linear models (GLMs) are a flexible generalization of ordinary linear regression models for response variables that are not normally distributed. If you're already familiar with multiple linear regression, you're well on your way to understanding GLMs.

3.1 Components of a Generalized Linear Model

A GLM consists of three components:

1. A random component: This is a specification of the probability distribution of the response variable (e.g., normal, binomial, Poisson distributions, etc.). This differs from ordinary linear regression, which assumes that the response variable follows a normal distribution.
2. A systematic component: This is the linear predictor, a linear combination of the explanatory variables, just as in ordinary linear regression.
3. A link function: This is a function that connects the mean of the response variable to the linear predictor. The choice of link function depends on the nature of the response variable and the range of its possible values.

3.2 Formulation of a Generalized Linear Model

The GLM can be formulated as follows:

$$g(E(Y)) = \eta = X\beta \tag{3.1}$$

Here, Y is the response variable, X represents the matrix of explanatory variables, β is the vector of parameters to be estimated, η is the linear predictor, $E(Y)$ represents the expected value of Y, and $g(.)$ is the link function.

3.3 Fitting a Generalized Linear Model

The parameters β in a GLM are typically estimated using maximum likelihood estimation (MLE). The specifics of this process depend on the probability distribution of the response variable and the link function.

3.4 Examples of Generalized Linear Models

Examples of GLMs include:

- Logistic regression: This is a GLM with a binomial response variable and a logit link function.
- Poisson regression: This is a GLM with a Poisson response variable and a log link function.

Chapter 4

Link Functions

In generalized linear models, the link function provides the relationship between the linear predictor and the mean of the distribution function. Different choices of link function can be used to model different types of relationships. Here are a few commonly used link functions:

1. **Identity link:** The identity link function is the simplest form of link function, where the response variable is expected to be the linear combination of the predictors. This is the default link function for Gaussian family distributions.

$$g(\mu) = \mu$$

2. **Log link:** The log link function is used when modeling positive data and count data. This link function is the default for Poisson and exponential family distributions.

$$g(\mu) = \log(\mu)$$

3. **Logit link:** The logit link function is often used when modeling binary response data, and is the default link function for binomial family distributions. It gives the log-odds, or the logarithm of the odds $p/(1-p)$.

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

4. **Probit link:** The probit link function is another common choice for binary response data. It is based on the cumulative distribution function of the standard normal distribution.

$$g(\mu) = \Phi^{-1}(\mu)$$

   where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution.

5. **Inverse link:** The inverse link function is often used in modeling rates or times. It is the canonical (or default) link function for the Gamma family distributions.

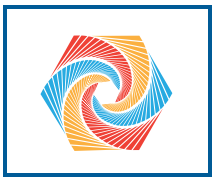
$$g(\mu) = \mu^{-1}$$

Different link functions can substantially impact the model's interpretation, so it's crucial to choose a link function that aligns with the nature of the data and the scientific question

at hand.

Appendix A

Answers to Exercises



Index