Regression Diagnostics

Jerry Peng

1 Statistical Terms

Overfitting Overfitting occurs when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means the model is too complex, capturing not only the underlying patterns in the data but also the noise or random fluctuations.

Underfitting Underfitting occurs when a model is too simple, both in terms of the structure or the assumptions it makes about the data. This simplicity leads to a model that does not learn adequately from the training data, missing the underlying trends.

Bias In statistical modeling, bias is the error introduced by approximating a real-world problem by a simplified model. It is the difference between the expected (or average) prediction of the model and the correct value that we are trying to predict. Calculation of bias typically involves more conceptual understanding rather than a straightforward formula because it depends on the true underlying relationships which are usually unknown.

Variance Variance measures how much the predictions for a given point vary between different realizations of the model. In other words, it captures the spread of the model's predictions. High variance can cause overfitting. Variance is calculated as the average of the squared differences from the mean:

$$Variance = \frac{\sum(x_i - \bar{X})^2}{n-1}$$

Note: High bias and low variance typically cause underfitting. Conversely, low bias and high variance are characteristics that can lead to overfitting

Coefficient In the context of a regression model, a coefficient represents the change in the dependent variable expected with a one-unit change in the predictor. Coefficients are derived from the regression model itself, based on the least squares method or other estimation approaches.

Standard Deviation Standard deviation is a measure of the amount of variation or dispersion in a set of values. It is the square root of variance and is calculated with the formula:

$$std = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}}$$

A high standard deviation indicates that the data points are widely spread out from the mean, suggesting a large amount of variability within the dataset. In practical terms, this means there is less consistency and more unpredictability in the values of the dataset.

2 Diagnose Forecasting Model

2.1 Mean Square Error (MSE)

Mean Squared Error (MSE) is a measure of the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

The MSE is calculated as follows:

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

where

- n is the number of data points
- y_i is the observed value
- $\hat{y}_i$ is the predicted value from the model

MSE is always non-negative, and values closer to zero are better. A lower MSE indicates a closer fit to the data. Since MSE is in the units squared of the variable being estimated, it is not directly interpretable in the original units of the data, which is why the root of MSE (RMSE) is often used for an interpretable error term.

2.2 Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is a frequently used measure of the differences between values predicted by a model and the values actually observed. It's particularly useful to assess the quality of a regression model, providing information on the magnitude of the error.

The formula for RMSE is:

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

where

- n is the number of data points
- y_i is the observed value
- $\hat{y}_i$ is the predicted value from the model

The squaring of the residuals, averaging them, and then taking the square root gives you the RMSE. RMSE has the same units as the observed values, which helps in interpreting the size of the error. A lower RMSE value indicates a better fit of the model to the data.

2.3 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a measure of prediction accuracy in a forecasting model. It calculates the average magnitude of errors in a set of predictions, without considering their direction. It's expressed as a percentage, which makes it easy to interpret because it describes the average error as a percentage of the actual values.

The formula for MAPE is:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

where

- n is the number of data points
- y_i is the observed value
- $\hat{y}_i$ is the predicted value from the model

The absolute value in the numerator is the absolute error, which is divided by the actual value to get the absolute percentage error. This value is then averaged over all data points to give the MAPE. A lower MAPE value indicates a better fit of the model to the data.

3 Diagnose Classification Models

3.1 Confusion Matrix

A Confusion Matrix is a table used to evaluate the performance of a classification model. It provides a detailed breakdown of true positive, true negative, false positive, and false negative predictions made by the model. The matrix is often used in machine learning and statistics to understand how well a classification model is performing and where it is making mistakes.

	Actual Positive	Actual Negative
Predicate Positive	True Positive	False Positive
Predicate Negative	False Negative	True Negative

Definitions

- **True Positive (TP):** The model correctly predicted the positive class.
- **True Negative (TN):** The model correctly predicted the negative class.
- **False Positive (FP):** The model incorrectly predicted the positive class (Type I error).
- **False Negative (FN):** The model incorrectly predicted the negative class (Type II error).

Derived Metrics Various performance metrics can be derived from the Confusion Matrix:

- **Accuracy:** (TP+TN)/(TP+TN+FP+FN)
- **Precision:** TP/(TP+FP)
- **Recall (Sensitivity):** TP/(TP+FN)
- **Specificity:** TN/(TN+FP)
- **F1 Score:** $2 \times (Precision \times Recall)/(Precision + Recall)$

The Confusion Matrix is particularly useful when the classes are imbalanced or when the costs of different types of errors vary. It provides a more detailed view of the model's performance compared to using a single metric like accuracy.

3.2 ROC Curve & AUC

ROC Curve The ROC curve is a graphical representation of the diagnostic ability of a binary classifier system as its discrimination threshold is varied. This curve plots two parameters:

- **True Positive Rate (TPR)** is a synonym for recall and is the proportion of actual positive cases that are correctly identified.

$$TPR = \frac{TP}{TP + FN}$$

- **False Positive Rate (FPR)** The proportion of actual negative cases that are incorrectly identified as positive.

$$FPR = \frac{FP}{FP + TN}$$

The ROC curve is a plot with TPR on the y-axis and FPR on the x-axis. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A model with no discriminatory ability (random guess) would produce a diagonal line from the bottom left to the top right of the plot (the line of no-discrimination).

AUC Area Under the ROC Curve provides an aggregate measure of performance across all possible classification thresholds. It ranges from 0 to 1.

- An AUC of 1 means the model is perfect at distinguishing between the two classes.
- An AUC of 0.5 suggests no discriminative ability, equivalent to random guessing.
- An AUC less than 0.5 implies worse than random predictions, but this is rare in practice.

The figure 1 shows an example of ROC curver and AUC

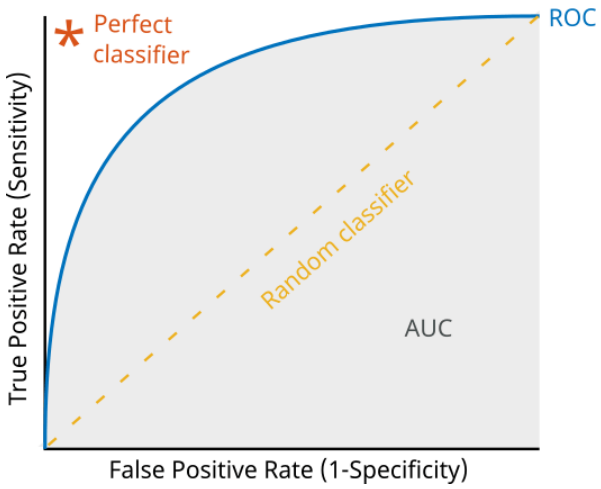

Figure 1: ROC Curve & AUC Example

3.3 Balanced Error Rate

It is a metric used to evaluate the performance of classification models, particularly in cases where the classes are imbalanced or when different types of classification errors have different costs. BER is designed to give a balanced view of the model's performance across all classes. In a binary classification problem, the Balanced Error Rate is defined as the average of the error rates for each class. Specifically, it is the average of the false positive rate and the false negative rate, or equivalently, one minus the average of the true positive rate (sensitivity) and the true negative rate (specificity).

Formula for Binary Classification

$$\text{BER} = \frac{1}{2}\left(\frac{FP}{FP+TN} + \frac{FN}{TP+FN}\right) \tag{1}$$

Formula for Multi-Class Classification For multi-class classification problems, BER is often calculated as the average of the error rates for each class, weighted by the class distribution in the test set.

$$\text{BER} = \frac{1}{N}\sum_{i=1}^{N}(\text{Error Rate of Class } i) \tag{2}$$

Importance of BER

- **Class Imbalance:** In datasets where one class significantly outnumbers the other(s), accuracy can be misleading. BER provides a more balanced measure of performance.
- **Different Error Costs:** In some applications, false positives and false negatives may have different costs. BER allows for a more balanced evaluation in such cases.
- **Fairness:** BER can be a useful metric when it's important for the model to perform equally well across different classes, such as in fairness-sensitive applications.

4 Coefficient of Determination

4.1 Fraction of Variance Unexplained(FVU)

It refers to that part of the total variability in the dataset that is not explained by the model. The total variance is the sum of squares of the differences between the observed dependent variable and its mean. This is commonly referred to as the total sum of squares (SST). The part of the variance that the model explains is called the explained sum of squares (SSE), and the part of the variance that the model does not explain is called the residual sum of squares (SSR).

The calculation formula is as follows:

$$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

$$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \tag{4}$$

$$\begin{aligned} FVU(f) &= \frac{MSE(f)}{Var(y)} \\ &= \frac{SS_{res}}{SS_{tot}} \end{aligned} \tag{5}$$

A lower FVU indicates that the model explains a larger portion of the variance, and therefore fits the data better. Conversely, a higher FVU indicates that the model does not explain a large portion of the variance.

4.2 R-squared (R^2)

R-squared, also known as the coefficient of determination, is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Mathematically, R-squared is defined as:

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

R-squared values range from 0 to 1, where 0 indicates that the model explains none of the variability of the response data around its mean, and 1 indicates that the model explains all the variability of the response data around its mean.

In terms of interpretation:

- An R^2 of 0.70 means that 70% of the variance in the dependent variable is predictable from the the independent variables.
- A higher R^2 value indicates a better fit between the model and the data.

4.3 Adjusted R-squared

Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. Unlike R-squared, which will always increase as more variables are added to the model, adjusted R-squared will only increase if the new term improves the model more than would be expected by chance.

The formula for adjusted R-squared is:

$$\text{Adjusted } R^2 = 1 - (\frac{(1 - R^2)(n - 1)}{n - k - 1})$$

where

- R^2 is the R-squared of the model

- n is the total number of observations
- k is the number of predictors(independent variables)

This adjustment accounts for the number of predictors in the model, providing a more accurate measure of the goodness-of-fit, especially when comparing models with a different number of predictors.

A higher adjusted R-squared value means that the model explains a greater proportion of the variance in the dependent variable after adjusting for the number of predictors. It indicates that the model fits the data well, considering the number of variables included.

Conversely, a lower adjusted R-squared value suggests that the model does not explain much of the variance in the dependent variable when the number of predictors is taken into account. It may indicate that additional predictors do not contribute meaningful information and could be superfluous or that the model overall has a weak fit.

4.4 Model F-statistic

The model F-statistic in the context of regression analysis is used to assess the overall significance of the model compared to a model with no independent variables (other than the constant term). It tests the null hypothesis that all the regression coefficients are equal to zero, which implies that the independent variables do not explain any of the variability in the dependent variable.

The formula for the F-statistic is:

$$F = \frac{MS_{reg}}{MS_{res}}$$

where

- MS_{reg} is the mean square regression (explained variance), calculated as the regression sum of squares (SSR) divided by the degrees of freedom of the model (number of independent variables). MS_{reg} is calculated as:

  $$MS_{reg} = \frac{SS_{reg}}{df_{reg}}$$

  where SS_{reg} is the regression sum of squares (explained sum of squares), and df_{reg} is the degrees of freedom associated with the regression, typically the number of independent variables in the model. SS_{reg} is calculated as:

  $$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

  where $\hat{y}_i$ are the predicted values from the regression equation, and $\bar{y}$ is the mean of the observed data, and n is the number of observations. df_{reg}, as the degrees of freedom for the regression, is calculated as:

  $$df_{reg} = k$$

  where k is the number of independent variables in the model.

- MS_{res} is the mean square error (residual variance), calculated as the sum of squares of residuals (SSE) divided by the error degrees of freedom (n - number of independent variables - 1). MS_{res} is calculated as:

  $$MS_{res} = \frac{SS_res}{df_res}$$

  where SS_{res} is the residual sum of squares (sum of squares of errors), and df_{res} is the degrees of freedom associated with the residual, calculated as $n-k-1$, where n s the number of observations and k is the number of predictors (not including the constant term). SS_{res} is calculated as:

  $$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

An F-statistic value greater than the critical value from the F-distribution table (for a specific alpha level, typically 0.05) indicates that the model provides a better fit to the data than the intercept-only model.

The F-statistic is then computed using these mean squares. The larger the F-statistic, the more likely it is that the observed relationship is not due to chance, implying that the independent variables significantly predict the dependent variable.

5 Regularization

Regularization is a technique to prevent overfitting in machine learning models. It controls the complexity of the model by adding a penalty term to the model's loss function, preventing the model from fitting the training data too closely. The basic idea is to add a regularization term to the loss function, which is related to the size of the model parameters (such as coefficients in linear regression or weights in neural networks). The role of the regularization term is to increase the model's loss when the complexity of the model increases (i.e., when the values of the parameters become very large).

There are various forms of regularization, with L1 regularization and L2 regularization being the most common. L1 regularization can make some parameters zero, achieving the purpose of feature selection, making the model sparse. L2 regularization, on the other hand, will make the parameters smaller but not exactly zero, preventing the model from being overly sensitive to a particular feature.

In summary, regularization improves the model's generalization capability on unseen data by sacrificing some of its performance on the training set, preventing overfitting.

L1 It is also known as Lasso (Least Absolute Shrinkage and Selection Operator). The penalty term for L1 regularization is the sum of the absolute values of the model weights. This characteristic of regularization will cause some of the model's weight parameters to shrink to zero, thereby achieving feature selection. Therefore, L1 regularization can be used to achieve sparsity in the model, meaning only a few features in the model are effective.

$$\begin{aligned} L_1 &= \frac{1}{N}\sum(y-\hat{y})^2 + \lambda\sum|b| \\ &= \frac{1}{N}\|y-\hat{y}\|_2^2 + \lambda|\theta| \end{aligned} \tag{6}$$

L2 It is also known as Ridge (Ridge Regression). The penalty term for L2 regularization is half of the sum of the squares of the model weights. Unlike L1 regularization, L2 regularization does not reduce the model weights to zero, but instead makes all weights tend uniformly towards zero. L2 regularization can prevent the model from being overly sensitive to a single feature because all features will participate in the model training to some extent.

$$\begin{aligned} L_2 &= \frac{1}{N}\sum(y-\hat{y})^2 + \lambda\sum b^2 \\ &= \frac{1}{N}\|y-\hat{y}\|_2^2 + \lambda\|\theta\|_2^2 \end{aligned} \tag{7}$$

The regularization parameter λ is a parameter that we need to set ourselves to control the strength of regularization. The larger the value of λ, the stronger the regularization, and the stricter the restriction on the model weights. When the value of λ is 0, the model becomes a regular linear regression or logistic regression."

References