Predictive Modeling

Department of Mathematical Sciences
Qiuying Sha, Professor

Michigan Tech

Part II Regression Models

Chapter 5. Measuring Performance in Regression Models

- Some measure of accuracy is typically used to evaluate the effectiveness of a model.

  There are different ways to measure accuracy, each with its own nuance.

- Visualizations of the model fit, particularly residual plots, are critical to understanding whether the model is fit for purpose.

5.1 Quantitative Measures of Performance

- RMSE

  - The most common method for characterizing a model's predictive capabilities is to use the root mean squared error (RMSE).

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

  - Interpretation: (on average) the residuals are from zero or as the average distance between the observed values and the model predictions.

- R^2

Another measure is the coefficient of determination, R^2.

$$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad SS_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

$$SS_{tot} = SS_{reg} + SS_{res}. \quad R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}.$$

- R^2 can be interpreted as the proportion of the information in the data that is explained by the model.
- R^2 can be also calculated as the correlation coefficient between the observed and predicted values (usually denoted by r) and $R^2 = r^2$

Michigan Tech

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

- *Notes:* R^2 is a measure of correlation, not accuracy.

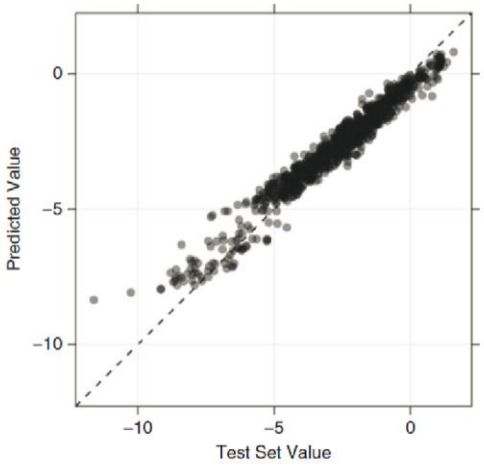

Fig. 5.1: A plot of the observed and predicted outcomes where the R^2 is moderate (51 %), but predictions are not uniformly accurate. The *diagonal grey reference line* indicates where the observed and predicted values would be equal

Figure 5.1: the R^2 is moderate (51%), but the model has a tendency to overpredict low values and underpredict high ones.

- Spearman’s rank correlation

  - To calculate this value, the ranks of the observed and predicted outcomes are obtained and the correlation coefficient between these ranks is calculated.

  - Assesses how well the relationship between two variables can be described using a monotonic function.

  - For details, see supplement from http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient

5.2 The Variance-Bias Trade-off

Suppose that we have a training set consisting of a set of points $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$, and real values y_i associated with the point x_i

We assume that there is a function, but noisy relation $y = f(\mathrm{x}) + \varepsilon$, where $E(\varepsilon) = 0$

and $Var(\varepsilon) = \sigma^2$

We want to find a function $\hat{f}(\mathrm{x})$, that approximates the true function $y = f(\mathrm{x})$ as well as possible.

Then

$$E[(y-\hat{f}(x))^2] = E[f(x) - E[\hat{f}(x)]]^2 + E\left[\left(E[\hat{f}(x)] - \hat{f}(x)\right)^2\right] + E[\varepsilon^2]$$

$$= [Bias(\hat{f}(x))]^2 + Var(\hat{f}(x)) + \sigma^2$$

where

$$Bias[\hat{f}(\mathrm{x})] = \mathrm{E}[\hat{f}(\mathrm{x})] - f(\mathrm{x})$$

and

$$\mathrm{Var}[\hat{f}(\mathrm{x})] = \mathrm{E}\left[\left(\hat{f}(\mathrm{x}) - \mathrm{E}[\hat{f}(\mathrm{x})]\right)^2\right]$$

σ^2 is usually called "irreducible noise" and cannot be eliminated by modeling.

Figure 5.2 shows extreme examples of models that are either high bias or high variance.

It is generally true that

- More complex models can have very high variance, which leads to over-fitting.
- Simple models tend to under-fit if they are not flexible enough to model the true relationship (thus high bias).
- Highly correlated predictors can lead to *collinearity* issues and this can greatly increase the model variance.
- In subsequent chapters, models will be discussed that can increase the bias in the model to greatly reduce the model variance. This is referred to as the *variance-bias trade-off*.

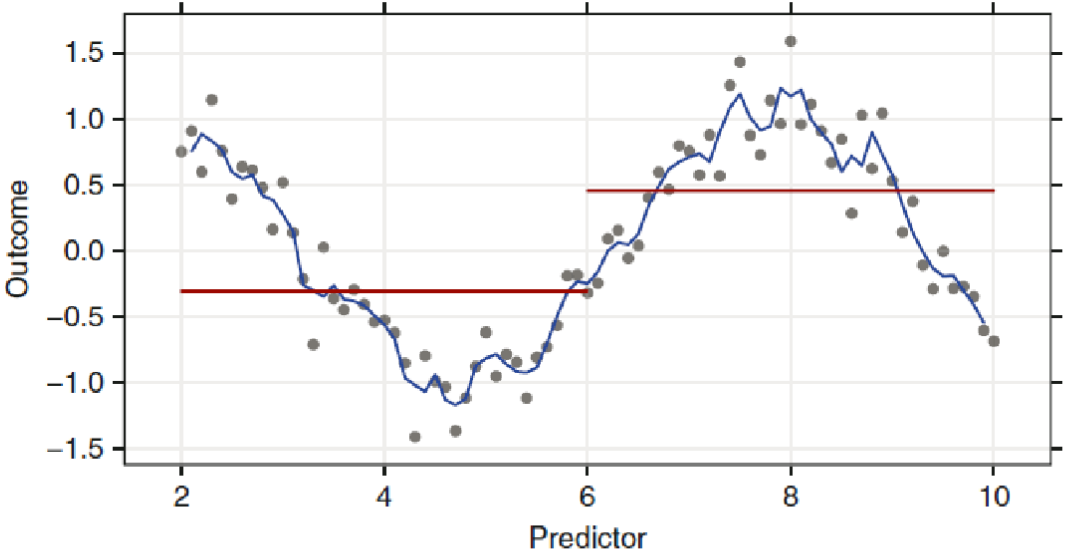

Fig. 5.2: Two model fits to a *sin* wave. The *red line* predicts the data using simple averages of the first and second half of the data. The *blue line* is a three-point moving average