Chapter 5
Measuring Performance in Regression Models

For models predicting a numeric outcome, some measure of accuracy is typically used to evaluate the effectiveness of the model. However, there are different ways to measure accuracy, each with its own nuance. To understand the strengths and weaknesses of a particular model, relying solely on a single metric is problematic. Visualizations of the model fit, particularly residual plots, are critical to understanding whether the model is fit for purpose. These techniques are discussed in this chapter.

5.1 Quantitative Measures of Performance

When the outcome is a number, the most common method for characterizing a model's predictive capabilities is to use the root mean squared error (RMSE). This metric is a function of the model residuals, which are the observed values minus the model predictions. The mean squared error (MSE) is calculated by squaring the residuals, summing them and dividing by the number of samples. The RMSE is then calculated by taking the square root of the MSE so that it is in the same units as the original data. The value is usually interpreted as either how far (on average) the residuals are from zero or as the average distance between the observed values and the model predictions.

Another common metric is the coefficient of determination, commonly written as R^2. This value can be interpreted as the proportion of the information in the data that is explained by the model. Thus, an R^2 value of 0.75 implies that the model can explain three-quarters of the variation in the outcome. There are multiple formulas for calculating this quantity (Kvålseth 1985), although the simplest version finds the correlation coefficient between the observed and predicted values (usually denoted by R) and squares it.

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3_5,

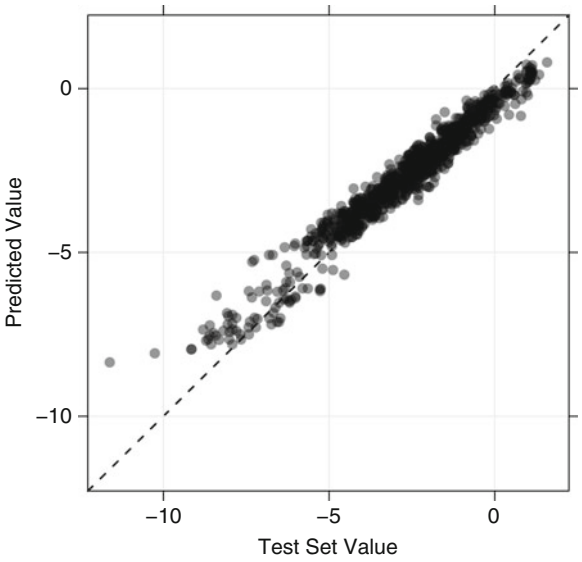

Fig. 5.1: A plot of the observed and predicted outcomes where the R^2 is moderate (51 %), but predictions are not uniformly accurate. The *diagonal grey reference line* indicates where the observed and predicted values would be equal

While this is an easily interpretable statistic, the practitioner must remember that R^2 is a measure of correlation, not accuracy. Figure 5.1 shows an example where the R^2 between the observed and predicted values is high (51 %), but the model has a tendency to overpredict low values and underpredict high ones. This phenomenon can be common to some of the tree-based regression models discussed in Chap. 8. Depending on the context, this systematic bias in the predictions may be acceptable if the model otherwise works well.

It is also important to realize that R^2 is dependent on the variation in the outcome. Using the interpretation that this statistic measures the proportion of variance explained by the model, one must remember that the denominator of that proportion is calculated using the sample variance of the outcome. For example, suppose a test set outcome has a variance of 4.2. If the RMSE of a predictive model were 1, the R^2 would be roughly 76 %. If we had another test set with exactly the same RMSE, but the test outcomes were less variable, the results would look worse. For example, if the test set variance were 3, the R^2 would be 67 %.

Practically speaking, this dependence on the outcome variance can also have a drastic effect on how the model is viewed. For example, suppose we were building a model to predict the sale price of houses using predictors such as house characteristics (e.g., square footage, number of bedrooms, number

of bathrooms), as well as lot size and location. If the range of the houses in the test set was large, say from \$60K to \$2M, the variance of the sale price would also be very large. One might view a model with a 90 % R^2 positively, but the RMSE may be in the tens of thousands of dollars—poor predictive accuracy for anyone selling a moderately priced property.

In some cases, the goal of the model is to simply rank new samples. As previously discussed, pharmaceutical scientists may screen large numbers of compounds for their activity in an effort to find "hits." The scientists will then follow up on the compounds predicted to be the most biologically active. Here, the focus is on the ranking ability of the model rather than its predictive accuracy. In this situation, determining the *rank correlation* between the observed and predicted values might be a more appropriate metric. The rank correlation takes the ranks of the observed outcome values (as opposed to their actual numbers) and evaluates how close these are to ranks of the model predictions. To calculate this value, the ranks of the observed and predicted outcomes are obtained and the correlation coefficient between these ranks is calculated. This metric is commonly known as Spearman's rank correlation.

5.2 The Variance-Bias Trade-off

The MSE can be decomposed into more specific pieces. Formally, the MSE of a *model* is

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

where y_i is the outcome and $\hat{y}_i$ is the model prediction of that sample's outcome. If we assume that the data points are statistically independent and that the residuals have a theoretical mean of zero and a constant variance of σ^2, then

$$E[\text{MSE}] = \sigma^2 + (\text{Model Bias})^2 + \text{Model Variance}, \tag{5.1}$$

where E is the expected value. The first part (σ^2) is usually called "irreducible noise" and cannot be eliminated by modeling. The second term is the squared *bias* of the model. This reflects how close the functional form of the model can get to the true relationship between the predictors and the outcome. The last term is the model variance. Figure 5.2 shows extreme examples of models that are either high bias or high variance. The data are a simulated *sin* wave. The model fit shown in red splits the data in half and predicts each half with a simple average. This model has low variance since it would not substantially change if another set of data points were generated the same

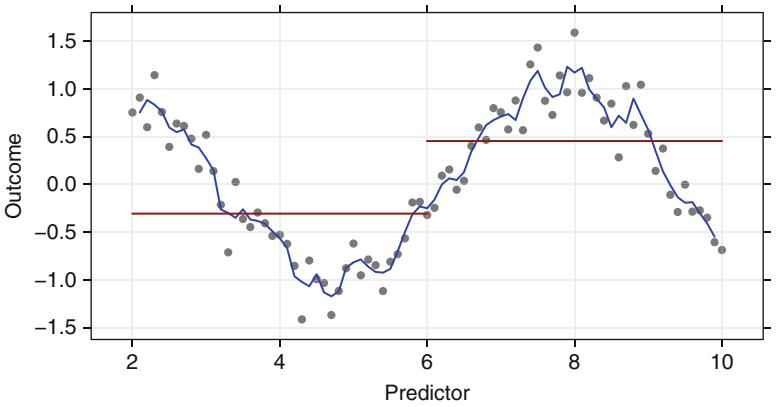

Fig. 5.2: Two model fits to a *sin* wave. The *red line* predicts the data using simple averages of the first and second half of the data. The *blue line* is a three-point moving average

way. However, it is ineffective at modeling the data since, due to its simplicity and for this reason, it has high bias. Conversely, the blue line is a three-point moving average. It is flexible enough to model the *sin* wave (i.e., low bias), but small perturbations in the data will significantly change the model fit. Because of this, it has high variance.

It is generally true that more complex models can have very high variance, which leads to over-fitting. On the other hand, simple models tend not to over-fit, but under-fit if they are not flexible enough to model the true relationship (thus high bias). Also, 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 as a way to mitigate the problem of collinearity. This is referred to as the *variance-bias trade-off*.

5.3 Computing

The following sections will reference functions from the caret package.

To compute model performance, the observed and predicted outcomes should be stored in vectors. For regression, these vectors should be numeric. Here, two example vectors are manually created to illustrate the techniques (in practice, the vector of predictions would be produced by the model function):

```
> # Use the 'c' function to combine numbers into a vector
> observed <- c(0.22,  0.83, -0.12, 0.89, -0.23, -1.30, -0.15, -1.4,
```

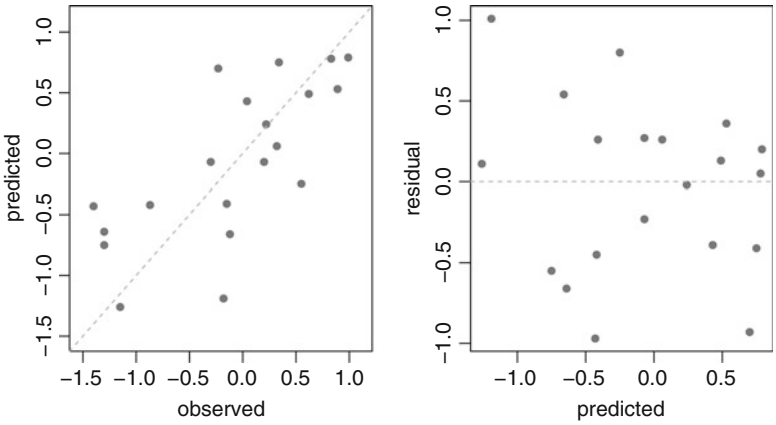

Fig. 5.3: *Left*: a plot of the observed and predicted values. *Right*: the residuals versus the predicted values

```
+               0.62,  0.99, -0.18, 0.32,  0.34, -0.30,  0.04, -0.87,
+               0.55, -1.30, -1.15, 0.20)
> predicted <- c(0.24,  0.78, -0.66,  0.53, 0.70, -0.75, -0.41, -0.43,
+                0.49,  0.79, -1.19,  0.06, 0.75, -0.07,  0.43, -0.42,
+               -0.25, -0.64, -1.26, -0.07)
> residualValues <- observed - predicted
> summary(residualValues)
     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -0.9700 -0.4200  0.0800 -0.0310  0.2625  1.0100
```

An important step in evaluating the quality of the model is to visualize the results. First, a plot of the observed values against the predicted values helps one to understand how well the model fits. Also, a plot of the residuals versus the predicted values can help uncover systematic patterns in the model predictions, such as the trend shown in Fig. 5.1. The following two commands were used to produce the images in Fig. 5.3:

```
> # Observed values versus predicted values
> # It is a good idea to plot the values on a common scale.
> axisRange <- extendrange(c(observed, predicted))
> plot(observed, predicted,
+      ylim = axisRange,
+      xlim = axisRange)
> # Add a 45 degree reference line
> abline(0, 1, col = "darkgrey", lty = 2)

> # Predicted values versus residuals
> plot(predicted, residualValues, ylab = "residual")
> abline(h = 0, col = "darkgrey", lty = 2)
```

The caret package contains functions for calculating the RMSE and the R^2 value:

```
> R2(predicted, observed)
  [1] 0.5170123
> RMSE(predicted, observed)
  [1] 0.5234883
```

There are different formulas for R^2; Kvålseth (1985) provides a survey of these. By default, the `R2` function uses the square of the correlation coefficient. Base R contains a function to compute the correlation, including Spearman's rank correlation.

```
> # Simple correlation
> cor(predicted, observed)
  [1] 0.7190357
> # Rank correlation
> cor(predicted, observed, method = "spearman")
  [1] 0.7554552
```