Predictions II

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Assessing the accuracy of predictions is an important step in statistical learning.
- The objectives have shifted from inference to prediction.
  - Need methods to assess model performance.
- Some measures of prediction accuracy:
  1. Mean Absolute Error
  2. Mean Squared Error
  3. Root Mean Squared Error
  4. Mean Absolute Percentage Error

Predictions

- A **prediction** (forecast) is a statement about a future event or unknown data.
- May use previous knowledge (statistical models) to make informed predictions.
- Sometimes referred to as *statistical learning*.

Training Data

- To test the prediction accuracy, **training data** is usually selected to *train* the model(s).
- The training set is generally larger than the testing set.
- The *trained* models are used to predict the outcomes of the response found in the testing data.

Testing Data

- The **testing data** is used to assess the accuracy of the model's predictions.
- The combinations of explanatory variables are given in the testing data.
- Predictions are made based on the coefficients of the trained regression model.
- As the actual outcomes in the testing data are known, measures of prediction accuracy can be calculated.

Illustrative Example 1

- Identify the training and testing data used in Example 4 from Predictions I?

Mean Absolute Error

- The **Mean Absolute Error** (MAE) is the mean of the absolute value of the errors:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

- In R:
  - `library(DescTools)`
  - `MAE(predicted,observed)`

- *Values are expressed in units of Y.*

Example 1

- What is the MAE of the predictions we made in Example 4 from Predictions I?
- What units are this value in?
- Do these results *feel* accurate?

Mean Squared Error

- The **Mean Squared Error** (MSE) is the mean of the squared errors:

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

- In R:
  - `library(DescTools)`
  - `MSE(predicted,observed)`

- *Values are no longer expressed in units of Y.*

Example 2

- What is the MSE of the predictions we made in Example 4 from Predictions I?
- What units are this value in?
- Do these results *feel* accurate?

Root Mean Squared Error

- The **Root Mean Squared Error** (RMSE) is the square root of the mean of the squared errors:

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

- In R:
  - `library(DescTools)`
  - `RMSE(predicted,observed)`

- *Values are expressed in units of Y.*

Example 3

- What is the RMSE of the predictions we made in Example 4 from Predictions I?
- What units are this value in?
- Do these results *feel* accurate?

Relative Absolute Error

- The **Relative Absolute Error** (RAE) is the percentage :

$$RAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}_i|} \tag{4}$$

- In R:
  - `library(Metrics)`
  - `rae(observed, predicted)`

- *Values are an estimate of error percentage (0 means perfect).*

Example 4

- What is the RAE of the predictions we made in Example 4 from Predictions I?
- What units are this value in?
- Do these results *feel* accurate?

Selecting Training and Testing Data

- Sometimes there is a natural order to selection training and testing data.
    - Forecasting future events
- Otherwise, sampling may be used to select the training and testing data.
- *As results can be significantly altered by individual observations, this process should occur multiple times.*
- **Never test your model on your *training* data.**

Cross-Validation

- **Cross-validation** involves *shuffling* the data, obtaining samples, then using the different samples from the data to test and train a model on different iterations.

- At each iteration measures of prediction accuracy are recorded and saved.

- *We will cover K-fold cross validation.*

K-Fold Cross-Validation

- Algorithm:
  1. Randomly split data into K subsets.
  2. Use K-1 subsets to train the model.
  3. Use the last (left out) subset to test the model.
  4. Repeat the steps K times (until each subset has been the testing data).
  5. Generate overall measures of prediction error by taking the average of the errors.

- This can be done easily in R (*see following slides*).

- Generally, $K = 5$ or $K = 10$.

K-Fold Cross-Validation in R Example

- *The code shows the cross-validation steps using the simple linear regression model from Example 1.*
- Multiple models can be added to this loop to compare different model performances on the same folds.

K-Fold Cross-Validation in R Simplified

- There are many existing functions that can be used to perform cross-validation in R.

- We can use the caret package:
  - `library(caret)`
  - `set.seed(2020)`
  - `train.control <- trainControl(method = "cv", number = K)`
  - `model <- train(Response ~ Var.1 + ...  + Var.M, data = data, method = "lm", trControl = train.control)`
  - `print(model)` *Prints the results*

- Note: R^2 *represents the squared correlation between the observed outcome values and the predicted values by the model. Higher values imply better accuracy.*

Example 5

- Use the caret package to perform 5-fold cross validation to determine which of the two linear regression models does a better job of predicting the *Points* earned by teams in the *Football22.csv* dataset.

  1. Points = 1 + Wins + League
  2. Points = 1 + Goals_For + Goals_Against

Comments on K-Fold Cross-Validation

- There is a bias-variance trade-off associated with the choice of K in K-fold cross-validation.
  - Lower K implies higher bias and lower variance in the error estimates.
  - **K = 10 is very commonly used.**
  - *Leave-one-out cross-validation:* $K = n$

- There are other methods of evaluating model prediction accuracy:
  - Repeated K-fold cross-validation (add `repeats = m` to the `trainControl()` function).
  - May also use bootstrap resampling methods.

Underfitting

- **Underfitting** occurs when the model does not capture the underlying relationship in the data.
- If a linear model is underfit, it will generally perform poorly with regards to the Adjusted-R^2 or F-test.
- Underfit models will not generalize well to new data.
    - Generally, will not predict well.

Example 6

- Run the code to simulate data and split the data into training and testing datatsets.
- Estimate a **linear** model using the training data and Y as the response variable.
- Use your model to predict the values of Y in your testing set.
- Use `ggplot()` to plot all of your results.
- What happened?

Overfitting

- **Overfitting** occurs when the model fits the training data *too well.*
- This can occur when the model captures more than just the the overall relationship in the data.
  - Too much flexibility in the explanatory variables.
  - Too many explanatory variables.
- Overfit models will not generalize well to new data.
  - Generally, will not predict well.

Example 7

- Run the code to simulate data and split the data into training and testing datatsets.
- As we do have repeated observations, we can treat $X1$ as a factor.
- Estimate a **linear** model using the training data and Y as the response variable.
- Use your model to predict the values of Y in your testing set.
- Use `ggplot()` to plot all of your results.
- What happened?

Example 8

1. Fix your prediction model from Example 6.
2. Fix your prediction model from Example 7.

Transformed Data

- Predictions from models for transformed data should be converted back to their original units.
  - `log()` transformation $\rightarrow$ `exp(prediction)`
- **If the objective is purely prediction we do not need to ensure the assumptions hold.**
  - Violations in the assumptions *may* lead to poor prediction accuracy.
- *Reversed Box-Cox transformations do not predict the mean, it predicts the median of the distribution.*

Example 9

- Import the *Housing.csv* data into R.
- Identify three different linear regression models to predict house prices.
- Use 10-fold cross-validation to select the *best* model.

Exercise 1

- Using the *Wages.csv* dataset:
    - Estimate multiple linear regression models for the dependent variable *Salary*.
    - Use 10-fold cross-validation to test the average prediction accuracy of your models.

Exercise 2

- Take some time to train and test some prediction models for your project data.

References & Resources

1. Kaplan, Daniel T. (2017). *Statistical Modelling: A Fresh Approach. (Second Edition).* Retrieved from https://dtkaplan.github.io/SM2-bookdown/

2. Fox, J. (2015). *Applied regression analysis and generalized linear models (Third Edition).* Sage Publications.

- DescTools
- Metrics
- `trainControl()`
- `train()`