Multiple Linear Regression

Paulo S. A. Sousa

2022-03-22

Overview

- The most popular model for regression problems is the *multiple linear regression model*.
- This model is used to fit a relationship between a numerical outcome variable Y and a set of predictors $X_1, X_2, \ldots, X_p$.
- The relationship between predictors and outcome variable is assumed to be:

  $$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon,$$

  where $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are coefficients and ϵ is the noise or unexplained part.
- Data are then used to estimate the coefficients and to quantify the noise.
- In predictive modeling, the data are also used to evaluate model performance.

Overview

- Regression modeling means not only estimating the coefficients but also choosing which predictors to include and in what form.
- For example, a numerical predictor can be included as is or in logarithmic form $(\log(X))$ or in a binned form (e.g., age group).
    - Choosing the right form depends on domain knowledge, data availability, and needed predictive power.
- Multiple linear regression is applicable to numerous predictive modeling situations. Examples are:
    - Predicting customer activity on credit cards from their demographics and historical activity patterns, predicting expenditures on vacation travel based on historical frequent flyer data,
    - Predicting staffing requirements at help desks based on historical data and product and sales information,
    - Predicting sales from cross-selling of products from historical information, and
    - Predicting the impact of discounts on sales in retail outlets.

Explanatory vs. Predictive Modeling

- Our focus is typically on predicting new individual records.
- Thus, we are not interested in the coefficients themselves, nor in the "average record," but rather in the predictions that this model can generate for new records.

Estimation and Prediction

- Once we determine the predictors to include and their form, we estimate the coefficients of the regression formula from the data using a method called *ordinary least squares* (OLS).
- This method finds values $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ that minimize the sum of squared deviations between the actual outcome values (Y) and their predicted values based on that model $\left(\hat{Y}\right)$.
- To predict the value of the outcome variable for a record with predictor values $x_1, x_2, \ldots, x_p$, we use the equation

$$\hat{Y} = \hat{\hat{\beta}}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p,$$

- If we make the classical assumptions, predictions based on this equation are the best predictions possible in the sense that:
  - They will be unbiased.
  - And will have the smallest mean squared error compared to any unbiased estimates.

Reducing the Number of Predictors

There are several reasons for parsimony regarding all possible variables into a model:

- It may be expensive or not feasible to collect a full complement of predictors for future predictions.
- We may be able to measure fewer predictors more accurately (e.g., in surveys).
- The more predictors, the higher the chance of missing values in the data.
  - If we delete or impute records with missing values, multiple predictors will lead to a higher rate of record deletion or imputation.
- Parsimony is an important property of good models:
  - We obtain more insight into the influence of predictors in models with few parameters.

Reducing the Number of Predictors

- Estimates of regression coefficients are likely to be unstable, due to multicollinearity, in models with many variables.
- Regression coefficients are more stable for parsimonious models:
    - One very rough rule of thumb is to have a number of records n larger than $5\,(p+2)$, where p is the number of predictors.
- It can be shown that using predictors that are uncorrelated with the outcome variable increases the variance of predictions.
- It can be shown that dropping predictors that are actually correlated with the outcome variable can increase the average error (bias) of predictions.
- Consequently, there is a trade-off between too few and too many predictors.

Reducing the Number of Predictors

- In general, accepting some bias can reduce the variance in predictions.
- This *bias–variance trade-off* is particularly important for large numbers of predictors, because in that case, it is very likely that there are variables in the model that have small coefficients relative to the standard deviation of the noise and also exhibit at least moderate correlation with other variables.
- Dropping such variables will improve the predictions, as it reduces the prediction variance.
- This type of bias–variance trade-off is a basic aspect of most data mining procedures for regression and classification.

How to Reduce the Number of Predictors

- The first step in trying to reduce the number of predictors should always be to use domain knowledge.
- Some practical reasons for predictor elimination are:
    - The expense of collecting this information in the future;
    - Inaccuracy;
    - High correlation with another predictor;
    - Many missing values; or simply irrelevance.

Exhaustive Search

- The idea here is to evaluate all subsets of predictors.
  - Since the number of subsets for even moderate values of p is very large, after the algorithm creates the subsets and runs all the models,
    - We need some way to examine the most promising subsets and to select from them.
- The challenge is to select a model that is not too simplistic in terms of excluding important parameters (the model is *under-fit*), nor overly complex thereby modeling random noise (the model is *over-fit*).
  - Several criteria for evaluating and comparing models are based on metrics computed from the training data:
  - One popular criterion is the *adjusted* R^2, which is defined as

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right)$$

where R^2 is the proportion of explained variability in the model (in a model with a single predictor, this is the squared correlation).

Exhaustive Search

- Like R^2, higher values of R^2_{adj} indicate better fit.
- Unlike R^2, which does not account for the number of predictors used, R^2_{adj} uses a penalty on the number of predictors.
- This avoids the artificial increase in R^2 that can result from simply increasing the number of predictors but not the amount of information.
- It can be shown that using R^2_{adj} to choose a subset is equivalent to picking the subset that minimizes $\hat{\sigma}^2$.

Exhaustive Search

- A second popular set of criteria for balancing under-fitting and over-fitting are the *Akaike Information Criterion* (AIC) and Schwartz's *Bayesian Information Criterion* (BIC).
- AIC and BIC measure the goodness of fit of a model, but also include a penalty that is a function of the number of parameters in the model.
- As such, they can be used to compare various models for the same data set. AIC and BIC are estimates of prediction error based in information theory.
- Consequently, models with smaller AIC and BIC values are considered better.

Exhaustive Search

- A third criterion often used for subset selection is Mallow's C_p.
- This criterion assumes that the full model (with all predictors) is unbiased, although it may have predictors that if dropped would reduce prediction variability.
- With this assumption, we can show that if a subset model is unbiased, the average C_p value equals $p + 1$ (= number of predictors + 1), the size of the subset.
- So a reasonable approach to identifying subset models with small bias is to examine those with values of C_p that are near $p + 1$.
  - Good models are those that have values of C_p near $p + 1$ and that have small p (i.e., are of small size).

Exhaustive Search

- C_p is computed from the formula

  $$C_p = \frac{\text{SSE}}{\hat{\sigma}^2_{\text{full}}} + 2(p+1) - n,$$

  where $\hat{\sigma}^2_{\text{full}}$ is the estimated value of σ^2 in the full model that includes all predictors.

- It is important to remember that the usefulness of this approach depends heavily on the reliability of the estimate of σ^2 for the full model.
  - This requires that the training set contain a large number of records relative to the number of predictors.
  - It can be shown that for linear regression, in large samples Mallows's C_p is equivalent to AIC.

Exhaustive Search

- Finally, a useful point to note is that for a fixed size of subset, R^2, R^2_{adj}, C_p, AIC, and BIC all select the same subset.
    - In fact, there is no difference between them in the order of merit they ascribe to subsets of a fixed size.
- This is good to know if comparing models with the same number of predictors, but often we want to compare models with different numbers of predictors.

Subset Selection Algorithms

- This second method relies on a partial, iterative search through the space of all possible regression models.
- The end product is one best subset of predictors (although there do exist variations of these methods that identify several close-to-best choices for different sizes of predictor subsets).
- This approach is computationally cheaper, but it has the potential of missing "good" combinations of predictors.
- None of the methods guarantee that they yield the best subset for any criterion, such as R^2_{adj}.
- They are reasonable methods for situations with a large number of predictors, but for a moderate number of predictors, the exhaustive search is preferable.

Subset Selection Algorithms

- Three popular iterative search algorithms are:
  - *Forward selection*,
  - *Backward elimination*, and
  - *Stepwise regression*.
- In *forward selection*, we start with no predictors and then add predictors one by one.
  - Each predictor added is the one (among all predictors) that has the largest contribution to R^2 on top of the predictors that are already in it.
  - The algorithm stops when the contribution of additional predictors is not statistically significant.
- The main disadvantage of this method is that the algorithm will miss pairs or groups of predictors that perform very well together but perform poorly as single predictors.
  - This is similar to interviewing job candidates for a team project one by one, thereby missing groups of candidates who perform superiorly together (“colleagues”), but poorly on their own or with non-colleagues.

Subset Selection Algorithms

- In *backward elimination*, we start with all predictors and then at each step, eliminate the least useful predictor (according to statistical significance).
- The algorithm stops when all the remaining predictors have significant contributions.
- The weakness of this algorithm is that computing the initial model with all predictors can be time-consuming and unstable.
- *Stepwise regression* is like forward selection except that at each step, we consider dropping predictors that are not statistically significant, as in backward elimination.
- Additional ways to reduce the dimension of the data are by using principal components and regression trees.