STA 111 (Summer Session I)

Lecture Twenty-One – Multicollinearity, Variable Selection and Cross-Validation

Instructor: Olanrewaju (Michael) Akande

Department of Statistical Science, Duke University

June 23, 2016

Outline

- Questions from Last Lecture
- Nonparametric Regression
- Multicollinearity
- Variable Selection
- Cross-Validation

Recap

Recall that the multiple linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

where $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{V}[\epsilon_i] = \sigma^2$, and the ϵ_i's are independent.

The model is useful because:

- it is interpretable—the effect of each explanatory variable is captured by a single coefficient
- theory supports inference and prediction is easy
- simple interactions and transformations are easy (how?)
- dummy variables allow use of categorical information
- computation is fast.

Nonparametric Regression

We extended the multiple linear regression model to nonlinear regression, in which we fit a model of the form:

$$y_i = \beta_0 + \beta_1 g_1(x_{i2}) + \ldots + \beta_p g_p(x_{ip}) + \epsilon_i$$

where the g_j's are known transformations of the data, such as the log or $1/x$, and, as before, $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{V}[\epsilon_i] = \sigma^2$, and the ϵ_i's are independent.

This model can be further extended to nonparametric regression, in which case one does not now the functions $g_1, \ldots, g_p$ but instead must estimate these by smoothing the data.

In the real applications, the linear regression model is usually only a locally correct approximation. And it is rare that one has a strong theoretical model that prescribes specific nonlinear transformations. Thus nonparametric regression is a practical tool in many cases.

Curse of Dimensionality

Regression becomes much harder as the number of explanatory variables increases. This is called the Curse of Dimensionality (COD). The term was coined by Richard Bellman in the context of approximation theory.

The COD applies to all multivariate regressions that do not to impose strong modeling assumptions—especially the nonparametric regressions, but also those in which one tests whether a specific variable or transformed variable should be included in the model.

In terms of the sample size n and dimension p, the COD has three nearly equivalent descriptions:

- For fixed n, as p increases, the data become sparse.
- As p increases, the number of possible models explodes.
- For large p, most datasets are multicollinear.

Curse of Dimensionality (Cont'd)

To explain the model explosion aspect, suppose we restrict attention to just linear models of degree 2 or fewer. For $p = 1$ these are:

$$\mathbf{E}[Y] = \beta_0 \qquad \mathbf{E}[Y] = \beta_1 x_1 \qquad \mathbf{E}[Y] = \beta_2 x_1^2$$
$$\mathbf{E}[Y] = \beta_0 + \beta_1 x_1 \qquad \mathbf{E}[Y] = \beta_0 + \beta_2 x_1^2 \qquad \mathbf{E}[Y] = \beta_1 x_1 + \beta_2 x_1^2$$
$$\mathbf{E}[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

For $p = 2$ this set is extended to include expressions with the terms $\alpha_1 x_2$, $\alpha_2 x_2^2$, and $\gamma_{12} x_1 x_2$. For general p, combinatorics shows that the number of possible models is

$$2^{1+2p+\binom{p}{2}} - 1.$$

This increases superexponentially in p, and there is not enough sample to enable the data to discriminate among these models.

Curse of Dimensionality (Cont'd)

For the multicollinearity issue, we note that **multicollinearity** occurs when two or more of the explanatory values are highly correlated. This implies that the predictive value of the fitted model breaks down quickly as one moves away from the subspace in which the data concentrate.

We shall agree that multicollinearity occurs whenever the absolute value of the correlation between two of the explanatory variables exceeds 0.9. But this is a judgment call, and one can have multicollinearity that arises in more complex ways.

For large p with finite n, it is almost certain that two explanatory variables will have high correlation, just by chance.

Variable Selection

One wants to select a multiple regression model that only includes useful variables. Some methods are:

- Forward Selection. One starts with no variables in the model, and sequentially adds the one that best explains the current residuals (or the raw data, at the initial step). One stops when none of the remaining variables provide significant explanation.
- Backwards Elimination. Start with all the variables in the model, and sequentially removes the variable that explains the least, until a t-test shows that no further variables should be removed.
- Stepwise Regression. Alternate use of forward selection and backwards elimination.

None of these is bulletproof.

Cross-Validation

To assess model fit in complex, computer-intensive situations, the ideal strategy is to hold out a random portion of the data, fit a model to the rest, then use the fitted model to predict the response values from the values of the explanatory variables in the hold-out sample.

This allows a straightforward estimate of the error in prediction using regression. But we usually need to compare fits among *many* models. If the same hold-out sample is re-used, then the comparisons are not independent and (worse) the model selection process will tend to choose a model the overfits the hold-out sample, causing spurious optimism.

Cross-Validation (Cont'd)

Cross-validation is a procedure that balances the need to use data to select a model and the need to use data to assess prediction.

Specifically, v-fold cross-validation is as follows:

- randomly divide the sample into v portions;
- for $i = 1, \ldots, v$, hold out portion i and fit the model from the rest of the data;
- for $i = 1, \ldots, v$, use the fitted model to predict the hold-out sample;
- average the predictive mean squared error (PMSE) over the v different fits.

One repeats these steps (including the random division of the sample!) each time a new model is assessed.

The choice of v requires judgment. Often $v = 10$.