THE UNIVERSITY OF BRITISH COLUMBIA
DEPARTMENT OF STATISTICS

STAT 447B Methods for Statistical Learning (2014/15 Term 1)
Lab 1

17/18 September, 2014

In the previous week you were exposed to linear and non-linear regression models. In this lab, we will solidify these concepts by applying the methods to analyze a dataset.

1 Review

- The aim of a regression model is in general to find a function f that relates the covariates x to the **expected value** of the response y given the covariates, i.e. $\mathbb{E}(Y|X_1,\ldots,X_p) = f(X_1,\ldots,X_p)$. When the function f is unknown or unspecified, we are interested in estimating the function f. Such class of regression problem is referred to as **non-parametric regression**. On the other hand, **parametric regression** refers to the case where the function f is known in the sense that the unknown parameters β can completely specify the function f, $\mathbb{E}(Y|X_1,\ldots,X_p) = f(X_1,\ldots,X_p;\beta_1,\ldots,\beta_p)$.
- Even when a regression model includes non-linear functions of the covariates, it is still a **linear model** as long as the model is linear in the regression parameters. Examples of such models include:
  - **Polynomial regression**: The response is regressed on the (integral) powers of the covariates. In the case of only one covariate, this amounts to $\mathbb{E}(Y|X) = \sum_{i=0}^{K} \beta_i X^i$ for certain predetermined maximum power K.
  - **Spline regression**: This method uses piecewise polynomials and is useful to capture the local characteristics of the relationship between the response and covariates. For instance, a cubic spline model can be formulated as $\mathbb{E}(Y|X) = \sum_{i=0}^{3} \beta_i X^i + \sum_{j=1}^{K} \beta_{j+3}(X-\kappa_j)_+^3$ for K predefined knots at $\kappa_1,\ldots,\kappa_K$.
- Splines are sometimes preferred over polynomial regression because of the flexibility they provide. With polynomial regression we may need very high powers to produce such flexible fits, and this has the drawback of the fit being very unstable (wiggly) especially near the end points of the range of the covariates. Also in polynomial regression we assume that the same polynomial function works for all the range of the covariates, whereas splines allow for more localized behaviour, having coefficients that only appear (i.e. non-zero) for some values of the covariates.

2 Analysis

The dataset that we will analyze today is the US temperature data downloadable from the lab webpage. The response variable is the minimum temperature ($min.temp$) and there are two covariates in the data, namely longitude and latitude ($long$ and lat).

2.1 Visualize the data

Load the data using `read.table`. Using `plot`, draw a scatterplot of the response against each of the covariates.

Q1 What is your observation regarding the relationship between latitude and the minimum temperature? What about for the longitude and the minimum temperature?

Q2 Do you think a linear regression with respect to both lat and $long$, i.e. $min.temp_i = \beta_0 + \beta_1 lat_i + \beta_2 long_i + \epsilon_i$, is sufficient? Why?

2.2 Non-linear regression

We will now compare the fits between global polynomial regression and cubic splines with three knots. Consider the two models below:

$$min.temp_i = \beta_0 + \beta_1 lat_i + \beta_2 long_i + \beta_3 long_i^2 + \beta_4 long_i^3 + \epsilon_i \tag{1}$$

$$min.temp_i = \beta_0 + \beta_1 lat_i + \beta_2 long_i + \beta_3 long_i^2 + \beta_4 long_i^3 + \sum_{j=1}^{3} \beta_{j+4}(long_i - \kappa_j)_+^3 + \epsilon_i \tag{2}$$

Q3 Why are the models linear in lat, but much more complicated in $long$?

Q4 Fit model (1):

- Perform the regression using `lm`; search on the internet if you are unsure of how to include higher-order terms.
- (Optional during the lab) After fitting the model, obtain the 95% confidence interval for each fitted data point. You may find the function `predict` useful here. Assign this matrix to a variable.

Q5 Now fit model (2):

- Construct a vector of the knots $(\kappa_1, \kappa_2, \kappa_3)$, where κ_i should be at the $(25i)$th quantile of $long$.
- Recall that the design matrix X is of the following form:

$$X = \begin{pmatrix} 1 & lat_1 & long_1 & long_1^2 & long_1^3 & (long_1 - \kappa_1)_+^3 & (long_1 - \kappa_2)_+^3 & (long_1 - \kappa_3)_+^3 \\ 1 & lat_2 & long_2 & long_2^2 & long_2^3 & (long_2 - \kappa_1)_+^3 & (long_2 - \kappa_2)_+^3 & (long_2 - \kappa_3)_+^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & lat_n & long_n & long_n^2 & long_n^3 & (long_n - \kappa_1)_+^3 & (long_n - \kappa_2)_+^3 & (long_n - \kappa_3)_+^3 \end{pmatrix},$$

  where n is the number of observations in the data set. Create this matrix (without the first column of ones) to be used in `lm`.

- Fit this linear model using `lm` and (optionally during the lab) obtain the 95% confidence interval for each fitted data point. Again use a variable to store the results.

Q6 Calculate the sum of squared errors arising from the two models. Which one is smaller? Is this always the case and why?

Q7 (Optional during the lab) Try to plot the confidence bands against observation number and make comparisons.

Note: You are also responsible for the parts marked optional above. They are only optional during the lab due to time constraints.

2.3 Another way of visualizing the results

Since there are two covariates in this example, we may use a 3D plot to visualize the fitting results. It is also possible to inspect the fitted polynomial/spline in 2D through the following steps (taken from [1]). This method can still show the effect of a change in one unit of a covariate on the response.

1. Compute the mean of latitude ($\overline{lat}$) across all data points.
2. Plot the observed points $min.temp_i$ versus $long_i$
3. Plot $\widehat{min.temp}_i$ versus $long_i$, plugging $\overline{lat}$ into the fitted equation.

Refer to the R code on how to do this.

References

[1] David Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*, chapter 2. Cambridge University Press, 2003.