The data for this program can be found in the text file “salinity.txt” which can be read into R and plotted, as a scatterplot matrix, with the commands:

```r
data= read.table("salinity.txt",header=T)
plot(data)
```

This example comes from the “shrimp project” that was carried out in 1977-8 at the University of North Carolina. The ultimate aim of the project was to use salinity and other measurements made during the spring season to predict the summer shrimp harvest from Pamlico Sound. Accurate predictions of the shrimp harvest would, for example, allow fishermen to rig that year for other species when the shrimp harvest was predicted to be low. The spring season was divided into two-week periods.\(^1\) The statisticians working on the project wanted to use the salinity and temperature from all of the two-week periods during the spring as predictor variables. One problem was the salinity was not measured in every two-week period. Therefore, they developed model to estimate salinity in periods when it was not measured.

The variables in the data set are salinity, sallag, trend, and discharge. Salinity and sallag are the measured amounts of salinity in Pamlico Sound at the time of measurement and two weeks earlier. Trend is the number of weeks since the beginning of spring. Discharge is the discharge into the sound of fresh water from rivers. The goal of the analysis is to develop a model for estimating salinity using the other three variables. The model can be used to estimate salinity during periods when it was not measured.

The scatterplot matrix suggests that salinity increases linearly with sallag and decreases, perhaps in a nonlinear fashion, with discharge. Trend was intended as a proxy for evaporation which would increase during the spring and possibly cause salinity to increase.

1. Fit the model

\[
salinity_i = \text{sallag}_i \beta_1 + \text{trend}_i \beta_2 + s(\text{discharge}_i) + \epsilon_i, \tag{1}
\]

where \(\epsilon_1, \ldots, \epsilon_n\) are iid \(N(0, \sigma^2)\) and \(s(\cdot)\) is a linear spline with knots at the \(1/6, 2/6, \ldots, 5/6\) quantiles of discharge. Note that an intercept is included as part of the spline \(s(\cdot)\). Model (1) is called “semiparametric” since trend and sallag have linear effects but the effect of discharge is modeled nonparametrically.

Estimate the parameters in this model. Use a hierarchical prior, as discussed in class, for jumps in the first derivative of \(s(\cdot)\). Use WinBUGS for your MCMC calculations.

Include the printout of the WinBUGS results with your homework. Also plot the data and fitted values using R code such as:

\(^1\)Spring is longer in North Carolina than in Ithaca; in Ithaca there probably would be only one two-week spring period :)

plot(discharge,salinity,typ="p",col="red",pch="o",cex=1.5)
points(discharge,salinity.sim$mean$mu,col="black",pch="*",cex=2)
legend(30,14,c("data","fitted"),col=c("red","black"),
pch=c("o","*"),cex=c(1.5,1.5))
segments(discharge,salinity,discharge,salinity.sim$mean$mu,col="blue")

Of course, you'll need to change "salinity.sim$mean$mu" to correspond to the names you use. Note the use of “segments” to connect the observed and fitted values of salinity.

2. Now consider the model that is linear in all three predictors, that is, model (1) with being \( s(\cdot) \) a linear function rather than a linear spline. Estimate the posterior predictive p-value using the test measure \( \max(y_i - \mathbf{x}_i^T \beta)^2 / \text{median}(y_i - \mathbf{x}_i^T \beta)^2 \) discussed in class. Here \( y_i = \mathbf{x}_i^T \beta + \epsilon_i \) is the linear model in standard vector notation. Based on the estimated p-value, do you recommend using the linear model or the semiparametric spline model?