Introduction to Nonparametric Regression
ESRC Oxford Spring School
May 2005

Exercises on Nonparametric Regression

1. Smoothing Scatterplots

a. The data frame `Robey` in the `car` package (from Robey, Shea, Rutstein, and Morris, 1992) contains data on contraceptive use and fertility in 50 developing countries. The data set includes the following variables:

`region`	`Africa`, `Asia`, `Latin.Amer`, `Near.East`
`tfr`	Total fertility rate (children per woman)
`contraceptors`	Percentage of contraceptors among married women of childbearing age.

Construct a scatterplot of `tfr` (the response variable) against `contraceptors` (the explanatory variable). Superimpose a least-squares regression line and loess smooth on the scatterplot. How would you characterize the relationship between the two variables?

b. The data frame `SLID` in the `car` package contains data for the Province of Ontario from the 1994 wave of the Canadian Survey of Labour and Income Dynamics; there are 7425 observations in the data set, for the following variables:

`wages`	Composite hourly wage rate, in dollars, from all jobs.
`education`	Number of years of schooling.
`age`	In years.
`sex`	`Female`, `Male`
`language`	`English`, `French`, `Other`

Use scatterplots and nonparametric-regression smoothes to explore the relationship of `wages` to each of `education` and `age`. What do you conclude? If these relationships are nonlinear, can you transform them to approximate linearity? *Hint*: Begin by looking at the distributions of the quantitative variables – e.g., `hist(wages)`. *Note*: `wages` are missing for 3278 respondents, likely because most are not in the labour force.

c. In R Exercise 1, you used education data for the U.S. states and Washington D.C. to regress SAT math scores (`SATM`) on state education expenditures (`dollars`) and the percentage of graduating high-school students in the state taking the SAT exam (`percent`). Now explore the relationships among these three variables in a scatterplot matrix [e.g., `pairs(cbind(SATM, dollars, percent))`]. Make separate scatterplots for any relationships that look nonlinear, and superimpose least-squares lines and loess smoothes on the plots. Reexamine the linear least-squares multiple regression of `SATM` on `dollars` and `percent`

using component+residual plots; if you detect nonlinearity in these plots try to accommodate it by transformation or another appropriate strategy.
Suggestion: Since `percent` is (of course) a percentage, consider a logit transformation – i.e., `log(percent/(100 – percent))`.

2. Nonparametric Multiple Regression and Additive Regression[1]

a. Continuing with the state education data from Exercise 1.c, fit an additive nonparametric-regression model regressing `SATM` on `dollars` and `percent`. Then fit a general nonparametric-regression model with these predictors (i.e., permitting `dollars` and `percent` to interact). Test for the interaction and for nonlinearity in the partial relationships of `SATM` to `dollars` and `percent`. Test the partial effect of each predictor in the additive model.
Suggestion: Use the `gam` function to fit both models to facilitate their comparison, but in fitting the general model, fix the degrees of freedom for the two-dimensional term in `dollars` and `percent` to a large value; otherwise, `gam` will choose fewer *df* for the more general model (which should tell you something about the evidence for interaction!).

b. Returning to the `SLID` data from Exercise 1.b, use an additive model to regress `log(wages)` on `education`, `age`, `sex`, and `language`. Test for the linearity of the education and age effects. Test the partial effect of each predictor.

3. Generalized Additive Models

a. **Logistic regression**: In the `SLID` data set used in Exercise 2.b, `wages` are missing for 3278 of the 7425 respondents in the sample. Assuming that these are individuals who are not in the labour force, define the variable `working <- !is.na(wages)`, and then fit a semiparametric binomial generalized additive model with `working` as the response variable, smooth terms in `age` and `education`, and the factor `sex`. Test the statistical significance of each term in the model, and test for nonlinearity in the `age` and `education` effects. *Note*: It will take a few seconds for the `gam` function to do its work.

b. **Logistic regression, continued**: Now consider the possibility that sex interacts with age and education in affecting labour-force participation. It is simple to fit a linear-logit model including the interactions:

```
mod.work.glm <- glm(!is.na(wages) ~ (education + age)*sex,
   family=binomial)
Anova(mod.work.glm)
```

Fitting a GAM with an interaction between a smooth term and a factor is also possible using the `gam` function in the `mgcv` package, but it is a bit less straightforward. To facilitate tests for the interactions, I will fix the degrees of freedom for the smooth terms. You can proceed as follows:

[1] Use the `gam` function in the `mgcv` package for this and the following exercises.

```
female <- sex ==  "Female"
male <- sex == "Male"
mod.work.gam <- gam(!is.na(wages) ~
      s(education, by=female, fx=TRUE, k=4)
      + s(age, by=female, fx=TRUE, k=4)
      + s(education, by=male, fx=TRUE, k=4)
      + s(age, by=male, fx=TRUE, k=4) + sex, family=binomial)
```

Starting with this model, test for the statistical significance of the interactions.

c. **Poisson regression**: Recall from R Exercise 4, Long's Poisson regression of the number of articles published by students in the last three year of their PhD programmes (`art`) on gender (`fem`), marital status (`mar`), number of children five years old or younger (`kid5`), the prestige of the PhD department (`phd`) and the number of articles published by their mentors in a three-year period (`ment`). Redo this analysis as a semi-parametric generalized additive model, fitting smooth terms for `phd` and `ment`, and treating the rest of the predictors (including `kid5`) as factors. Test the partial effect of each predictor, and perform tests for nonlinearity for `kid5`, `phd`, and `ment`.