Post-course assessment

The work here is "light touch assessment", intended to take students up to half a week to complete. Students should talk to their supervisors to find out whether or not their department requires this work as part of any formal accreditation process (APTS itself has no resources to assess or certify students). It is anticipated that departments will decide the appropriate level of assessment locally, and may choose to drop some (or indeed all) of the parts, accordingly.

All data used in the questions on this assignment sheet can be loaded into R using the command

```
load(url("http://www.stats.gla.ac.uk/~claire/aptsassess.RData"))
```

Question 1 (P-splines). In this task we consider P-splines, which minimise the penalised least-squares criterion

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|\mathbf{D}\boldsymbol{\beta}\|^2 = \|\mathbf{y} - \mathbf{B}\boldsymbol{\beta}\|^2 + \lambda\|\mathbf{D}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{B}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{B}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\mathbf{D}^\top\mathbf{D}\boldsymbol{\beta},$$

where $\mathbf{B}$ is the matrix of B-spline basis functions, $\mathbf{D}$ is the difference matrix used in the penalty, and $\lambda \geq 0$ is the smoothing parameter.

(a) Show, by taking the derivative with respect to $\boldsymbol{\beta}$, that the minimiser of the penalised least-squares criterion is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{B}^\top\mathbf{y}.$$

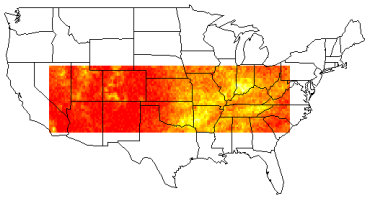
(b) Explain why we can rewrite the objective function as

$$\left\| \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{B} \\ \sqrt{\lambda}\mathbf{D} \end{pmatrix} \boldsymbol{\beta} \right\|^2.$$

(c) How can we exploit this to estimate $\boldsymbol{\beta}$ using a QR decomposition, which is numerically more stable than inverting the matrix $\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D}$?

Question 2 (Which smoother?). In the `RData` file `aptsassess` the data object `us.rain` contains data on total rainfall (mm) in March/April 2006 in a rectangular area covering most of the central US. The vectors `us.longitude` and `us.latitude` contain the corresponding longitudes and latitudes (where longitudes are in the rows and latitudes are in the columns for `us.rain`).

(a) Use the `maps` library and `image` command to produce the following display of the data:



(b) Create a dataframe with 3 columns one for each of rainfall, longitude and latitude.

(c) Use the `mgcv` library in `R` to fit an additive model with a smooth spatial surface over longitude and latitude for rainfall. Extract the fitted values and plot these using the `maps` library and `image` function.

(d) Change the smoother used here to use a spline on the sphere to estimate this surface and plot the results.

(e) Finally, replace the spatial smooth with a Gaussian process smooth, allowing the autocorrelation function to drop to zero at some point, and plot the results.

(f) Compare the AICs for the 3 models.

Question 3 (Fitting a quantile regression model). In the `RData` file `aptsassess` the data object `MaxSpeed` consists of the satellite-derived tropical cyclone lifetime-maximum wind speeds (`Wmax`) (i.e. the maximum intensities that cyclones achieve during their lifetimes) from 1977 to 2006. Researchers conjecture that tropical cyclones are getting stronger over the years. We are interested in modelling the time trend of `Wmax`.

(a) Fit a linear least squares regression regressing `Wmax` on `Year` and interpret the results.

In the following, we consider the quantile regression model $\text{Wmax}_i = \beta_0(\tau) + \beta_1(\tau)\text{Year}_i + e_i(\tau)$, where the τth conditional quantile of $e_i(\tau)$ given Year_i equals zero.

(b) Fit quantile regressions at quantile level $\tau = 0.5$, 0.75 and 0.95. Compare to the results from the least squares regression.

(c) For a quantile grid $\tau \in [0.05, 0.1, 0.15, \ldots, 0.90, 0.95]$, estimate and plot the quantile coefficients $\beta_0(\tau)$ and $\beta_1(\tau)$ against τ, respectively. Describe your findings.

(d) Suppose that due to a typo, Wmax_{20} was recorded as 200. Repeat Parts (a) and (c) with the perturbed data and compare the results.

Question 4 (Quantile regression tests). In the `RData` file `aptsassess` the data object `scores` gives 50 comprehension scores for each of three different instruction methods. In the data, `y` denotes the score and `grp` is the group indicator. Consider the quantile regression model $Q_\tau(y_i|x_i) = \beta_1(\tau) + \beta_2(\tau)x_{i1} + \beta_2(\tau)x_{i2}$ where $x_{i1} = 1$ for Method 2 and zero otherwise and $x_{i2} = 1$ for Method 3 and zero otherwise.

(a) Test the significance of $\beta_2(\tau)$ and $\beta_3(\tau)$ at the nominal level 0.05, for $\tau = 0.5$ and $\tau = 0.75$ respectively.

(b) Construct an approximate 95% confidence interval for $\beta_2(\tau) - \beta_3(\tau)$ and interpret the results. *Hint: You will need an estimate of the covariance matrix for the $\hat{\boldsymbol{\beta}}$, which can be obtained using*

`summary(fit, cov=TRUE)$cov` in `library(quantreg)`.

(c) Use pairs bootstrap to test $H_0 : \beta_2(0.75) - \beta_2(0.25) = 0$ at level 0.05. This test is equivalent to testing that the interquartile ranges are identical for the first two groups.