Due: Not to be turned in. Answers will be posted as soon as we have them written.

1. The data in elnino.txt are records of sea surface temperature in four regions of the Pacific ocean from 1950 to current. The data labelled NINO1+2 is the average sea surface temperature for the extreme eastern equatorial Pacific. The actual temperature (in Nino1+2) is quite noisy because it includes seasonal variation. (Winter is cooler than summer, even in the tropics). The important column is labelled ANOM (just to the right of NINO1+2). This is the temperature anomaly (observed temperature - expected temperature for that month). Positive anomalies are times of El Nino conditions; negative anomalies are times of La Nina conditions. As an aside, El Nino/La Nina conditions have consequences throughout the western hemisphere. For example, La Nina conditions are associated with June droughts in Iowa.

   Throughout, assume that observations are independent given the full model. This is not true for many full models, but I don't want to deal with that complication.

   (a) One crucial question is whether there is any trend in the temperature anomaly. Fit a linear regression, E $ANOM = \beta_0 + \beta_1 Date$, where Date is Year + (Month-1)/12. Report the estimated slope and the p-value for the test of Ho: $\beta_1 = 0$.

   R Note: The variable you create should not have the name "date" because spm confuses that with the system function date().

   (b) The trend may not be linear. Fit a penalized spline curve, using spar = 5. Use this to test whether there is lack of fit to a linear trend. Report your test statistic and p-value.

   R notes:
   1) You can specify the smoothing parameter, spar, by f(date, spar=5).
   2) There is no anova() function for spm objects. If spm is an output object from spm(), the pieces you need are the residuals in spm$fit$residuals, the model df in diab.spm$aux$df.fit and the residual df in diab.spm$aux$df.res.
   3) If you get an error "Error in unique.default(x) : unique() applies only to vectors", you probably have a variable in your working directory with the same name as a variable in the El Nino data frame or the same name as an R function (e.g. date). The data frame is 2nd in the search list, so anything in the working directory supercedes it. Detach the data frame, Change the variable name (or make a second copy) in the data frame, then rerun.

   (c) Fit a penalized spline curve, using the default choice of smoothing parameter. What is the model degrees of freedom? What is the model degrees of freedom for the spline fit with spar = 5? Which fit will be "wigglier"?

   (d) Plot the anomaly over time (I suggest a line plot, not dots) and overlay the penalized spline curves for the default estimated smoothing parameter and for spar=5. Which curve is estimating long-term trends? Which curve is estimating short-term oscillations?

2. The ratio of various isotopes in fossils provides a convenient way to date fossils. The data in fossil.txt contain measurements of the ratio of two strontium isotopes (strontium ratio) for fossil brachiopods (snails) of known age (in Million years before present).

   (a) Fit a penalized regression spline using REML to estimate the unknown smoothing parameter. What is the estimated smoothing parameter? What are the model d.f. associated with this smooth?

   R Note: The estimated smoothing parameter is the "spar" value in the output from summary(spm). DF is the model df.

   (b) The default output from spm() includes the se of the predicted y for specified x. However, the investigators are interested in prediction intervals. That is they want an interval that would include 95% of the observed strontium ratios at a specified age. Consider a shell with age of 115.236 My. For that shell, the se of the fit = 8.683×10^{-6}. The estimated residual sd $\hat{\sigma} = 2.516 \times 10^{-5}$ and the residual d.f. = 91.6. Calculate the standard deviation of predicted values for age=115.236 Myr. Calculate a 95% prediction interval for the strontium ratio of shells that are 115.236 Myr's old.

The data in Iowa.csv are wheat yield data in Iowa from 1930-1962. (Yes, wheat used to be grown in Iowa. Very little, if any, grown here now.) There are also 9 potential covariates. In all parts below, we will use the data from 1930-1955 to fit models and the observations from 1956-1962 to evaluate the predictive ability of the model.

(a) Fit a OLS regression using all variables for 1930-1955 to predict yield. What is the root-mean-squared-error of prediction (rMSEP) for 1956-1962 for this model?

(b) Fit a LASSO model, initially including all possible covariates, to the yield from 1930-1955. Which variables are included in the final model?

(c) What is the rMSEP for 1956-1962 for the lasso model?

(d) Fit an additive penalized spline (i.e. f(x1) + f(x2) ...) using only the variables included in the lasso model. What is the rMSEP for this model?

(e) Look at the output from summary(spmfit). This includes information about whether a covariate is being modeled as a straight line or as a curve. For which variables is there evidence of non-linearity? Explain the reason(s) for your choice(s).

Note: No formal test of lack of fit is required.