STAT 578 - Assignment 2 - due date is on the course outline

1. The BOD data set from A4.1 of Bates & Watts is on the course website as 'bod2'. A description of the data is attached to this assignment.

   (a) Fit the model $f(x, \boldsymbol{\theta}) = \alpha\left(1 - e^{-\beta x}\right)$ (there is a missing minus sign in B&W). Describe how the starting values are obtained. Submit the summary of the fit and a plot of the data and fitted response.

   (b) Make a perspective plot (see 'persp' in the R on-line help) of the sum of squares function over a region that contains the LSE.

   (c) On the same axes, plot: i) a linear approximation ellipsoid, ii) a likelihood region for the parameter pair. Do this with an 80% level and again with a 95% level. Comment on any noteworthy features of the data, or noteworthy differences between this analysis and the analysis of the BOD data set in class. Here you might want to refer to the profile t-plots.

   (d) Show that the $100(1-\alpha)\%$ profile region for β is of the form

   $$\left\{\beta \mid \frac{\left(\sum_{i=1}^{n}\left(1 - e^{-\beta x_i}\right) y_i\right)^2}{\sum_{i=1}^{n}\left(1-e^{-\beta x_i}\right)^2} \geq c\right\},$$

   for a constant c (which you should exhibit explicitly) depending on the data and α.

   (e) Consider testing the hypothesis that β equals some value β_0. Using each of the test statistics F_1 to F_4 discussed in class, and a 5% significance level, prepare - all on the same axes - plots of F as a function of β_0. Put a horizontal line at the critical value for the test, thus enabling you to read off confidence intervals. Plot as well the endpoints of the confidence interval obtained from the linear approximation t-statistic (so that your plot of F_1 should cross the horizontal line at these two points). Comment on any noteworthy features of these four plots. Note: these computations are probably only feasible if the conditional linearity is exploited.

2. Question 3.5, page 132 of Bates and Watts (attached). The data are on the course website as 'steady state'. A description of the data is also attached. Part (e) is poorly phrased; it should be interpreted as "apply the reparameterization of part (c) followed by the reparameterization of part (d)". In discussing the conditioning refer to the maximum condition number, i.e. the ratio of the largest to the smallest eigenvalues. Some possibly useful R commands are `control(maxiter = xxx)` as an argument of `nls(...)`, `vcov(fit)` for the covariance matrix of the parameter estimates from the nls fit, `cov2cor(mat)` to convert a covariance matrix to a correlation matrix, and `eigen(matrix, only.values = T)`.

3. (a) Recall the approximations that led to

$$F_2 \approx \frac{\boldsymbol{\varepsilon}'\left(\mathbf{H}_0 - \mathbf{H}_{1,0}\right)\boldsymbol{\varepsilon}}{p_2} \Big/ S^2 \ , \ S^2 \approx \frac{\boldsymbol{\varepsilon}'\left(\mathbf{I} - \mathbf{H}_0\right)\boldsymbol{\varepsilon}}{n-p}.$$

Show that, when $p_2 = p$, F_3 and F_4 both have the same approximate representations as F_2.

(b) Show further that, when $p_2 = p$, F_4 is *exactly* distributed as F^p_{n-p}.

(c) Consider the likelihood region on the parameter vector $\boldsymbol{\theta}$ in a nonlinear regression. Evaluate this region assuming that the model is in fact linear, and show that then it reduces to the usual confidence ellipsoid.

4. Suppose we start with a model $\mathbf{Y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$, and propose to test H: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ using a test statistic $F(\boldsymbol{\theta}_0)$. We reparameterize by introducing a $1-1$, differentiable map $\mathbf{g} : \boldsymbol{\theta} \to \boldsymbol{\phi} = \mathbf{g}(\boldsymbol{\theta})$. Define $\tilde{\boldsymbol{\eta}}(\boldsymbol{\phi}) = \boldsymbol{\eta}(\mathbf{g}^{-1}(\boldsymbol{\phi}))$ so that the model is $\mathbf{Y} = \tilde{\boldsymbol{\eta}}(\boldsymbol{\phi}) + \boldsymbol{\varepsilon}$ and we test H: $\boldsymbol{\phi} = \boldsymbol{\phi}_0$ where $\boldsymbol{\phi}_0 = \mathbf{g}(\boldsymbol{\theta}_0)$. The test is *invariant* if $F(\boldsymbol{\phi}_0) = F(\boldsymbol{\theta}_0)$. Show that the test of H: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ based on F_1 is invariant (for all $\boldsymbol{\theta}_0$) if and only if $\mathbf{g}(\cdot)$ is a *linear* transformation.

5. Let $s(x)$ be the cubic spline interpolating the function $f(x)$ at nodes $x_1 < x_2 < \cdots < x_{N-1} < x_N$. Show that, if $g(x)$ is any other twice continuously differentiable function interpolating $f(x)$ at the nodes then

$$\int_{x_1}^{x_N} \left[g''(x)\right]^2 dx \geq \int_{x_1}^{x_N} \left[s''(x)\right]^2 dx,$$

with equality iff $g(x) \equiv s(x)$ on $[x_1, x_n]$. (Hint: Consider the identity

$$\int_{x_1}^{x_N} \left[g''(x) - s''(x)\right]^2 dx = \int_{x_1}^{x_N} \left[g''(x)\right]^2 dx - 2\int_{x_1}^{x_N} \left[g''(x) - s''(x)\right] s''(x) dx - \int_{x_1}^{x_N} \left[s''(x)\right]^2 dx.$$

Show that the middle term on the rhs vanishes.)

6. Consider the local regression estimator $\hat{f}(\mathbf{x}) = (1, \mathbf{x}')\,\hat{\boldsymbol{\theta}}(\mathbf{x})$, with

$$\hat{\boldsymbol{\theta}}(\mathbf{x}) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} K_\lambda(\mathbf{x}, \mathbf{x}_i)\left(y_i - \left(1, \mathbf{x}_i'\right)\boldsymbol{\theta}\right)^2$$

for a kernel $K_\lambda(\cdot,\cdot)$. Suppose that the $\mathbf{x}_i$ are in a neighbourhood of the point $\mathbf{x}_0$ within which f is a linear function of $\mathbf{x}$. Show that then $\hat{f}(\mathbf{x}_0)$ is unbiased.

7. The fitted response in question 2 is quite complicated. Can you do as well with a nonparametric method? Prepare three plots, of Y against x_1 in each of the three temperature groups. Use `lines(...)` to superimpose the fitted response curves from 2(c). Then fit a model relating Y to (x_1, x_2) using loess (`loess(y ~x1 + x2, ... )`) and superimpose the predictions from this output. Repeat, with x_2 in place of x_1. Experiment with the choice of 'span'; let me know which one you decide upon. Here is an R function that will do the plotting. The argument `vec` will be either x_1 or x_2, `fit` will be the loess fit to which you are comparing the nls fit, and the `fit.nls` referred to in the function is the nls fit from 2(c).

```
compare = function(vec, fit) {
par(mfrow=c(3,1))
for(T in unique(temp)) {
  plot(vec[temp==T], y[temp==T], pch=19, ylab = "rate", xlab = "x",
  xlim = c(.9*min(vec), 1.1*max(vec)), ylim = c(.9*min(y), 1.1*max(y)))
  xx1 = sort(vec[temp==T])
  yy1 = predict(fit.nls)[temp==T]
  zz1 = predict(fit)[temp==T]
  yy1 = yy1[order(xx1)]
  zz1 = zz1[order(xx1)]
  lines(xx1, yy1)
  lines(xx1, zz1, lty=4)
  legend("topleft", legend = c("nls", "loess"), lty = c(1,4))
  }
}
```

8. Re-do question 7, but using gam fits instead of loess fits. Compare two fits - one using x_1, x_2 and $x_{12} = x_1 * x_2$ as the three independent variables to which splines are fitted, and another using x_1, x_2 only. Use one of the approximate tests discussed in class to determine the better of the two. Using it, prepare the same 6 plots as in #7, comparing the gam fit with the nls fit.

APPENDIX 4.

Data Sets Used in Problems

A4.1 BOD Data Set 2

Data on biochemical oxygen demand (BOD) were obtained by Marske (1967) as described in Appendix 1, Section A1.3. A second set of data is reported in Table A4.1.

A model was derived based on exponential decay with a fixed rate constant as

$$f(x, \boldsymbol{\theta}) = \theta_1(1 - e^{\theta_2 x})$$

where f is predicted biochemical oxygen demand and x is time.

Table A4.1 Biochemical oxygen demand versus time.

Time (days)	Biochemical Oxygen Demand (mg/l)	Time (days)	Biochemical Oxygen Demand (mg/l)
1	0.47	5	1.60
2	0.74	7	1.84
3	1.17	9	2.19
4	1.42	11	2.17

Q2, Asst. 2

(c) Use the starting values in a nonlinear least squares routine to find the least squares estimates for the parameters for each data set.
(d) Use incremental parameters and indicator variables to fit all of the data sets together.
(e) Simplify the model by letting some of the parameters be common to all of the data sets. Use extra sum of squares analyses to determine a simple adequate model.
(f) Write a short report about this analysis and your findings.

3.2 Use the data from Appendix 1, Section A1.14 to determine an appropriate sum of exponentials model.
(a) Plot the data on semilog paper and use the plot to determine the number of exponential terms to fit to the data.
(b) Use curve peeling to determine starting estimates for the parameters.
(c) Use the starting estimates from part (b) to fit the postulated model from part (a).

3.3 (a) Use the plot from Problem 2.6 and sketch in the curve of steepest descent from the point $\boldsymbol{\theta}^0$. Hint: The direction of steepest descent is perpendicular to the contours.
(b) Is the direction of the Gauss–Newton increment close to the initial direction of steepest descent?
(c) Calculate and plot the Levenberg increment using a conditioning factor of $k = 4$.
(d) Calculate and plot the Marquardt increment using a conditioning factor of $k = 4$.
(e) Comment on the relative directions of the Gauss–Newton, Levenberg and Marquardt increment vectors.

3.4 Use the data from Appendix 4, Section A4.3 to determine an appropriate model and to estimate the parameters.
(a) Plot the concentration versus time on semilog paper, and use the plot to determine the number of exponential terms necessary to fit the data.
(b) Use the plot and the method of curve peeling to determine starting values for the parameters.
(c) Use a nonlinear estimation routine to estimate the parameters.

3.5 Use a nonlinear estimation routine and the data and model from Appendix 4, Section A4.4 to estimate the parameters. Take note of the number of iterations required and any difficulties you encounter in each attempt.
(a) Use any approach you think is appropriate to obtain starting values for the parameters in the model.
(b) Use your starting values in a nonlinear estimation routine to estimate the parameters. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model.
(c) Reparametrize the model by centering the factor $1/x_3$, and use the equivalent starting values from part (a) to estimate the parameters. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model. What effect does this reparametrization have on the number of iterations to convergence?
(d) Reparametrize the model in part (a) using $\theta_1 = e^{\phi_1}$ and $\theta_2 = e^{\phi_2}$ and the equivalent starting values from part (a) to estimate the parameters. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model. What effect does this reparametrization have on the number of iterations to convergence?
(e) Reparametrize the model in part (b) using the same parametrization as in part (c) and the equivalent starting values from part (a) to estimate the parameters. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model. What effect does this reparametrization have on the number of iterations to convergence?

3.6 Use a nonlinear estimation routine and the data and model from Appendix 4, Section A4.5 to estimate the parameters. Take note of the number of iterations required and any difficulties you encounter in each attempt.
(a) Use any approach you think is appropriate to obtain starting values for the parameters in the model.
(b) Use your starting values in a nonlinear estimation routine to estimate the parameters. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model.
(c) Reparametrize the model in part (a) using $\theta_2 e^{-\theta_3 x} = e^{-\phi_3(x-\phi_2)}$. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model. What effect does this reparametrization have on the number of iterations to convergence?

3.7 (a) Show that the theoretical D-optimal starting design for the logistic model of Problem 3.1 consists of $\mathbf{x} = (-\infty, \theta_3 - 1.044/\theta_4, \theta_3 + 1.044/\theta_4, +\infty)^T$.
(b) Interpret the choice of the design points graphically by plotting the logistic function versus x and plotting the location of the design points on the x-axis.
(c) Plot the derivatives with respect to the parameters versus x and use these plots to help interpret the choice of the design points.

Data for Q2, Asst. 2

A4.2 Nitrendipene

Data on binding of [^{3}H] nitrendipine to sites in rat heart homogenate were obtained by Abdollah (1986). In this study, experiments were performed to investigate the competition for binding to the sites between nitrendipene (NTD), a calcium channel antagonist, and nifedipine (NIF), another calcium channel antagonist. Heart tissue was homogenated and incubated with radioactively tagged NTD at molar concentration $\approx 5\times10^{-10}$ in the presence of different concentrations of NIF, which are given in Table A4.2 as $x = \log_{10}$(NIF concentration), except for the rows with (0), for which the actual concentration was 0. The NIF has greater binding ability and so displaces the NTD. Counts on radioactive material were obtained to determine how much material was bound under different conditions. When the NIF concentration is 0, all of the radioactive NTD is bound to the sites, and so a large count is recorded: as the NIF concentration increases, it displaces NTD and so lower counts are recorded. Although the nominal NTD concentration was 5×10^{-10}, the actual concentrations were 4.76, 5.11, 4.78, and 5.02 $\times10^{-10}$ respectively, for the four tissue samples.

The proposed model is

$$f(x, \boldsymbol{\theta}) = \theta_1 + \frac{\theta_2}{1+\exp[-\theta_4(x-\theta_3)]}$$

where f is the predicted total count and x is $\log_{10}$(NIF concentration).

A4.3 Saccharin Data Set 2

Data on the concentration of saccharin in plasma were reported in Renwick (1982) and are reproduced in Table A4.3.

A4.4 Steady State Adsorption

Data on the disappearance of o-xylene as a function of oxygen concentration, inlet o-xylene concentration, and temperature, were obtained by Juusola (1971) and were further analyzed by Pritchard (1972). The data are reproduced in Table A4.4.

The postulated model is a steady state adsorption model written

$$f(\mathbf{x}, \boldsymbol{\theta}) = \frac{f_1 f_2}{f_1 + 2.2788 f_2}$$

$$f_1 = \theta_1 x_1 e^{-\theta_3/x_3}$$

$$f_2 = \theta_2 x_2 e^{-\theta_4/x_3}$$

Table A4.4 Rate of oxidation of o-xylene versus oxygen concentration (gm-mole/l), inlet o-xylene concentration (g-mole/l), and temperature (K). The reaction rate is recorded as (g-mole/g-mole eatalyst second) at standard catalyst age.

Oxygen	o-Xylene	Temp.	Rate	Oxygen	o-Xylene	Temp.	Rate
0.00502	0.000200	543	116	0.00249	0.000198	563	224
0.00499	0.000190	543	120	0.00571	0.000049	563	198
0.00504	0.000200	543	114	0.00555	0.000347	563	463
0.00505	0.000200	543	117	0.00549	0.000274	563	370
0.01000	0.000351	543	245	0.00554	0.000095	563	258
0.01010	0.000351	543	230	0.00507	0.000191	573	543
0.01030	0.000050	543	106	0.00502	0.000187	573	561
0.01040	0.000361	543	230	0.00505	0.000192	573	560
0.01010	0.000049	543	121	0.00506	0.000188	573	578
0.01010	0.000050	543	115	0.00500	0.000201	573	542
0.01010	0.000050	543	127	0.00100	0.000350	573	197
0.00570	0.000201	563	408	0.00505	0.000202	573	559
0.00552	0.000201	563	380	0.00306	0.000349	573	414
0.00551	0.000202	563	320	0.00502	0.000198	573	467
0.00551	0.000186	563	399	0.00504	0.000201	573	468
0.00554	0.000202	563	371	0.01017	0.000245	573	933
0.00553	0.000199	563	368	0.00499	0.000187	573	509
0.00108	0.000051	563	63	0.01000	0.000253	573	955
0.00707	0.000099	563	333	0.00496	0.000346	573	650
0.00554	0.000197	563	322	0.01000	0.000253	573	902
0.00605	0.000351	563	413	0.00502	0.000199	573	532
0.00552	0.000202	563	344	0.00399	0.000357	573	552
0.01016	0.000189	563	543	0.00107	0.000196	573	184
0.00552	0.000200	563	372	0.00499	0.000353	573	663
0.00603	0.000049	563	229	0.00503	0.000100	573	409
0.01000	0.000201	563	563	0.00251	0.000199	573	326
0.01010	0.000151	563	490	0.00499	0.000277	573	580
0.00805	0.000354	563	595	0.00906	0.000205	573	831
0.00552	0.000199	563	352				

where f is predicted reaction rate, x_1 is oxygen concentration, x_2 is o-xylene inlet concentration, and x_3 is temperature.