STAT-S 470/670 Homework 5

Tentative Due Date: Wednesday 10/28/2015

1. UREDA, p.207, #2 (Infant mortality data).

2. Below are selected personal consumption expenditures for the U.S. (units: billions of dollars):

	1940	1945	1950	1955	1960
Food/Tobacco	22.2	44.5	59.6	73.2	86.8
Household	10.5	15.5	29.0	36.5	46.2
Medical/Health	3.53	5.76	9.71	14.0	21.1
Personal care	1.04	1.98	2.45	3.40	5.40
Educ/research	.641	.974	1.80	2.60	3.64

   (a) Fit the table via median polish. Calculate "Analog R^2".

   (b) Construct a symbols plot of the residuals. (The command is "symbols"; see example in first lecture of Olympic running time data. If doing by hand, you can make the circles (for negative residuals) and squares (for positive residuals) in 4 different sizes, for tiny (zero), small, medium, large.) Do you see any patterns?

   (c) Construct the diagnostic plot.

   (d) If your plot in (c) indicates a transformation, transform the data as suggested by the plot, and conduct a median polish on the transformed data along with "Analog R^2". Compare results with those obtained in (a).

   (e) Plot the fit (forget-it plot). Which variable, time or category, has the larger effect on the data? Discuss your findings.

3. First let us create a simulated data set in R where we create a sample of $n = 50$ observations from a statistical model $y_i = \mu(t_i) + \epsilon_i$ using uniformly spaced design points $t_i = (2t - 1)/100$ for $i = 1, \ldots 50$ and where $\mu(t) = t + 0.5\exp\{-50(t - 0.5)^2\}$. For the errors assume that $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma = 0.5$. You can generate the errors using rnorm$(50, 0, 0.5)$ in R. Create a data frame of this generated data where you label the times t in the first column and the y observations in the second column. Now pretend you don't know the true equation $\mu(t)$ but want to estimate it using a local kernel estimation

$$\hat{\mu}_\lambda(t) = \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{t - t_i}{\lambda}\right) y_i$$

for this purpose you can use ksmooth$(x, y, \text{kernel} = "\textit{normal}", \text{bandwidth})$ in R. However, first you must find the asymptotically optimal bandwidth h, which is given by

$$\lambda_{opt} = n^{-1/5} \left\{ \frac{\sigma^2 R(K)}{J_2(\mu) M_2^2} \right\}^{1/5}$$

where

$$J_2(\mu) = \int_0^1 \mu^{''}(t)^2 dt$$

$$M_2 = \int_{-1}^1 u^2 K(u) du = 1 \text{ for gaussian kernel and,}$$

$$R(K) = \int_{-1}^1 K^2(u) du = \frac{1}{2\sqrt{\pi}}$$

Compute the optimal bandwidth and then use this in the function ksmooth() to obtain a smoothed estimator $\hat{\mu}_{\lambda_{opt}}(t)$. Plot $\hat{\mu}_{\lambda_{opt}}(t)$ and $\mu(t)$ on the same graph. How does the estimator compare with the true value of $\mu(t)$?