Homework 2

1. We are interested in testing whether major earthquakes are more likely during certain times of the day. The file Earthquake.csv is prepared from USGS NEIC database. It contains information about nearly all earthquakes of more than 6.0 on the Richter scale for dates from 2001 to 2012.

   (a) Draw a plot comparing the CDF of the uniform distribution to the Empirical CDF of this data.

   (b) Draw a plot of the function $G_n(t) = \sqrt{n}(\hat{F}_n(t) - t)$ where $\hat{F}_n$ is the empirical cdf

   (c) Use the ks.test function to test the null hypothesis that the data is from a uniform distribution and report the P-value. Describe your analysis and conclusion.

2. The data set GermProd.txt contains monthly measurements of industrial production in Germany. We are interested in modeling production (y) as a function of time in month (x)

   (a) Fit kernel regression to the data using bandwidth 24 and different kernels. Plot the estimates and the raw data.

   (b) Fit kernel regression to the data using different bandwidths (10,20,30) and Epanechnikov kernel. Plot the estimates and the raw data.

3. Suppose that we have $X_i, i = 1, \cdots n$ from an unknown distribution and we want to use a triangular kernel to obtain an estimate of the density.

$$K_h(x) = \begin{cases} \frac{2}{h} + \frac{4x}{h^2} & -\frac{h}{2} < x < 0 \\ \frac{2}{h} - \frac{4x}{h^2} & 0 < x < \frac{h}{2} \end{cases}$$

   Suppose the true distribution of X_i is uniform(0,1) and $h < 1$.

   (a) Find the bias of the kernel density estimator $\hat{f}(0.5)$.

   (b) Find the variance of the kernel density estimator $\hat{f}(0.5)$.

4. The following question refers to the Accept-Reject method and the notation used in class.

   (a) Why do we need f to be absolutely continuous with respect to g?

   (b) Why is it okay if g is not absolutely continuous with respect to f?

   (c) Why do we need $M > f(x)/g(x)$ for all x?