Applied Statistics - Spring 2012
Final Exam (resit)

August 20th, 2012

INSTRUCTIONS:

- This is a OPEN NOTES exam. You can bring all the materials distributed in class, plus the book by L. Wasserman we used.
- In addition, you can bring any MANUSCRIPT PERSONAL NOTES (other books, or copies of books are not allowed).
- You can use a calculator. Cellphone, notebooks or similar devices are not allowed.
- You have 180 minutes (3 hours) to complete the exam.
- There are 7 problems.
- The problems are not necessarily in order of difficulty. I recommend that you quickly read through all problems first, then do the problems in whatever order suits you best.
- The exam is to be done INDIVIDUALLY. Therefore discussion with your fellow colleagues is not allowed!
- A correct answer does not guarantee full credit, and a wrong answer does not guarantee loss of credit. You should clearly and concisely indicate your reasoning and **show all relevant work**. Your grade on each problem will be based on my best assessment of your level of understanding as reflected by what you have written. **JUSTIFY** your answers and be **CRITICAL** of your results.
- Please be organized in your write-up – I can't grade what I can't decipher!
- Remember to **IDENTIFY YOUR HANDOUT**.

1. Suppose we want to test the effectiveness of a new fertilizer product (Grow-Jo) against the most popular brand in the market (EverGreen). Four seeds are planted and randomly assigned one of the fertilizers, two of which to the treatment group (Grow-Jo) and the remaining two to the control group (EverGreen). After several weeks the height of each of the plants was measured: the two plants in the treatment group measured 9.5cm and 8.8cm, and the two plants in the control group measured 8.5cm and 9.1cm.

   The Wilcoxon rank-sum statistic T is defined as the sum of treatment ranks in the combined sample. So, in the present case, the observed value for the Wilcoxon rank-sum statistic equals $4 + 2 = 6$. The null hypothesis states that there is no treatment effect (in other words, there is no observable difference between Grow-Jo and EverGreen).

   (a) Compute the (exact) permutation distribution of T under the null hypothesis that there is no treatment effect (note that there are at most $\binom{4}{2}$ possible values for T).

   (b) Compute the p-value for testing the null-hypothesis against the alternative hypothesis that Grow-Jo increases plant growth. Would you recommend using Grow-Jo instead of EverGreen?

2. Let

$$X_1, \ldots, X_n \sim F \quad \text{and} \quad Y_1, \ldots, Y_m \sim G \ ,$$

   be two i.i.d. samples from two continuous distributions F and G. Furthermore assume $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$ are also independent from each other. We are interested in testing the null hypothesis $H_0 : \ F = G$ versus the alternative $H_1 : F \neq G$. This is commonly referred to as the general two-sample problem. To make things simpler you may assume both F and G are invertible and continuous.

   Let $\hat{F}_n$ and $\hat{G}_m$ denote the empirical cumulative distribution functions based respectively on the samples $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$ and define

$$K_{n,m} = \sup_{t \in \mathbb{R}} \left| \hat{F}_n(t) - \hat{G}_m(t) \right| \ .$$

   This is known as the *Kolmogorov-Smirnov two sample test statistic*.

   (a) Show that, under the null hypothesis, this statistic is distribution free. That is, if $F = G$ the distribution of $K_{n,m}$ does not depend on the shape of F (**note:** to answer this you don't need to derive an analytic form for the distribution of $K_{n,m}$).

   (b) Suppose we have a realization $k_{n,m}$ of $K_{n,m}$. Give a scheme for approximating the distribution of $K_{n,m}$ under H_0.

   (c) Now suppose either F or G has a discrete distribution. Can the bootstrap be used to approximate the distribution of $K_{n,m}$ under H_0? Explain why or why not. In case you think this can be done, explain how the resampling should be performed.

3. Suppose we have a realization of a random sample $X_1, \ldots, X_n$ from a location scale family. That is, the distribution function of each X_i is given by

$$F_{\mu,\sigma}(x) = F\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0.$$

Here F is a fixed distribution function. We use the Cramer-Von Mises test to check for goodness of fit for this location-scale family.

(a) Give a scheme or pseudo-code for approximating the p-value of the test. You don't have to given precise code, just give the essential steps. You can assume that there is a function that computes the value of the Cramer-Von Mises statistic.

(b) Is it necessary to use a bootstrap procedure for approximating the p-value, or can we do with plain Monte-Carlo simulation?

4. Consider the usual regression setting. Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be n data points and suppose that

$$Y_i = r(x_i) + \epsilon_i \ ,$$

where ϵ_i's are i.i.d. random variables such that $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i^2] = \sigma^2$. Consider the following estimator

$$\hat{r}_n(x) = \begin{cases} Y_2 & , \text{if } x = x_1 \\ Y_{n-1} & , \text{if } x = x_n \\ \frac{Y_{i-1}+Y_{i+1}}{2} & , \text{if } x = x_i, \text{ for some } i \in \{2, \ldots, n-1\} \\ \bar{Y} & \text{otherwise} \end{cases} ,$$

where $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$.

(a) Compute the weights $\ell(x) = (\ell_1(x), \ldots, \ell_n(x))$ so that $\hat{r}_n(x) = \sum_{i=1}^n \ell_i(x) Y_i$. Check also that $\sum_{i=1}^n \ell_i(x) = 1$. (Note that you need to consider 4 different cases for x).

(b) Find the smoothing matrix L (recall that $L_{ij} = \ell_j(x_i)$).

(c) Compute the effective degrees of freedom.

(d) The answer to the previous question might seem puzzling to you, as the number of effective degrees of freedom is rather small. However, as we seen when studying the quality of an estimate of the variance σ^2, the *fitted degrees of freedom* play a dominant role. Compute the fitted degrees of freedom

$$\tilde{\nu} = \text{trace}(L^T L) = \sum_{i=1}^n \|\ell(x_i)\|^2 \ .$$

Does this value seem somewhat more reasonable/intuitive?

5. Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be n data points and suppose that

$$Y_i = r(x_i) + \epsilon_i$$

where $E\epsilon_i = 0$ and $\operatorname{var} \epsilon_i = \sigma^2$. The points $x_1, \ldots, x_n$ are non-random. The error-terms are assumed to be independent. Consider $\hat{r}_{n,h}$ to be the Nadaraya-Watson estimator for r (h is the smoothing parameter). This is a linear estimator: there exists weights $\ell_1(x), \ldots, \ell_n(x)$ such that $\hat{r}_{n,h} = \sum_{i=1}^n \ell_i(x) Y_i$.

   (a) Show that $b = E[\hat{r}_{n,h}(x)] - r(x) = \sum_{i=1}^n \ell_i(x)(r(x_i) - r(x))$.

   (b) Suppose r is in fact α-Lipschitz, with $\alpha \in (0, 1)$ and the kernel function K in the definition of the Nadaraya-Watson estimator has support $[-1, 1]$. Show that $b = O(h^\alpha)$.

   (c) Show that

$$\operatorname{var}(\hat{r}_{n,h}(x)) = \sigma^2 \frac{\sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)^2}{\left(\sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)\right)^2}.$$

   (d) Suppose that $\operatorname{var}(\hat{r}_{n,h}(x)) = O(1/(nh))$ and we wish to choose the smoothing parameter h to minimise the mean squared error. What order (in terms of the sample size n) should we choose h to be?

6. Suppose $Y_1, \ldots, Y_n$ are independent and

$$Y_i \sim \operatorname{Exp}(r(x_i)).$$

Explain how we can estimate the unknown regression function $x \mapsto r(x)$ by local likelihood.

Reminder: the density of the $\operatorname{Exp}(\lambda)$ distribution is given by $f(x) = \lambda e^{-\lambda x}$, for $x \geq 0$.

7. Explain what is being calculated on each line of the following `R`-code.

```
x <- runif(n)^(1/4)
f <- function(x) sin(2*pi*x)
a <- 0.2
y <-  f(x) + rnorm(n,0,a)
out <- locfit(y ~ x, deg=2, alpha=0.4)
```