Applied Statistics - Spring 2012
Final Exam

June 4th, 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. Mark each item as T (TRUE) or F (FALSE) and explain. For example, if you think the item is FALSE, indicate how it should be altered such that it is true. An answer without justification is judges wrong.

   (a) Since a permutation test is exact, any significance level can be achieved exactly.

   (b) In the composite goodness-of-fit problem the Anderson-Darling test is always distribution free.

   (c) In the composite goodness-of-fit problem the Kolmogorov-Smirnov test is distribution free in case of location-scale families.

   (d) In constructing a confidence interval for the success parameter of a binomial distribution one should always prefer the Wald confidence interval.

   (e) The leave-one-out cross-validation statistic is a nearly unbiased estimator of the risk (=mean squared error).

   (f) Both the Wald and the Wilson confidence interval for the success parameter of a binomial distribution are based on an approximation involving the normal distribution.

   (g) For multiple linear regression the fitted degrees of freedom are always strictly larger than the effective degrees of freedom.

2. 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.

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

3. Consider a model of the form

$$Y(x) = \beta_0 + \beta_1 x + \sigma Z,$$

with $\sigma > 0$ and Z a standard normal random variable. Suppose we are interested in $\theta = P(Y(1) \leq a)$ for some $a \in \mathbb{R}$.

(a) Show that

$$\theta = \Phi\left(\frac{a - \beta_0 - \beta_2}{\sigma}\right)$$

where Φ denotes the cumulative distribution function of a standard normal random variable.

(b) Assume $Y(x_1), \ldots, Y(x_n)$ is a random sample from the postulated model. Denote the maximum likelihood estimators for the parameters by $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$ respectively. Define the following plug-in estimator for θ

$$\hat{\theta} = \Phi\left(\frac{a - \hat{\beta}_0 - \hat{\beta}_2}{\hat{\sigma}}\right)$$

Explain how you can approximate the standard error of this estimator using the parametric bootstrap. That is, give a few lines of pseudo code that upon execution would result in an approximation of the standard error of $\hat{\theta}$.

4. In class we studied a "simple" regression problem of the form

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $x_i = i/n$ and ϵ_i are independent random variables with mean zero and variance σ^2.

For every integer $m < n$ we can define a piecewise constant estimator as follows: Define the set $N_j = \{i \in \{1, \ldots, n\} : x_i \in [(j-1)/m, j/m)\}$. and set

$$\hat{f}_n(t) = \sum_{j=1}^{m} \hat{c}_j \mathbf{1}_{[(j-1)/m, j/m)}(t)$$

with $\hat{c}_j = \text{average}\,\{Y_i,\, i \in N_j\}$.

Now assume f is α-Lipschitz, for $\alpha \in (0, 1]$. That is, assume f is in the class

$$\mathcal{F}_L = \{f : [0,1] \to \mathbb{R} : |f(s) - f(t)| \leq L|t - s|^\alpha, \; \forall t, s \in [0,1]\}$$

where $L > 0$ is a constant.

Let $\bar{f}(t) = \mathbb{E}[\hat{f}_n(t)]$. Show that the squared bias, defined by

$$\int_0^1 (\bar{f}(t) - f(t))^2 dt$$

is bounded by $Cm^{-2\alpha}$ for some positive constant C.

5. Let $\hat{r}_n(x) = \sum_{i=1}^n \ell_i(x) Y_i$ be a linear smoother. Throughout this question assume that

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

   (a) Find the mean of $\hat{r}_n(x)$.

   (b) Find conditions on the weights so that $\hat{r}_n(x)$ is an unbiased estimator of $r(x) = \beta_0 + \beta_1 x$.

   (c) Suppose we decide to use a local polynomial estimator of degree p. For which values of p is $\hat{r}_n(x)$ an unbiased estimator of $r(x) = \beta_0 + \beta_1 x$?

6. 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. Consider the following estimator for r

$$\hat{r}_n(x) = \begin{cases} Y_i & \text{if } x = x_i, \quad i = 1, \ldots, n \\ \bar{Y} & \text{otherwise} \end{cases}$$

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

   (a) Find the weights $\ell(x) = (\ell_1(x), \ldots, \ell_n(x))$ so that $\hat{r}_n(x) = \sum_{i=1}^n \ell_i(x) Y_i$. (There are two cases. Case 1: $x \in \{x_1, \ldots, x_n\}$. Case 2: $x \notin \{x_1, \ldots, x_n\}$.)

   (b) Find the smoothing matrix L.

   (c) Find the effective degrees of freedom.

7. Suppose `x` and `y` are vectors containing numerical values. Consider the following R-code:

```
ord <- order(x)
y.ord <- y[ord]
varest.rice[i] <- sum((diff(y.ord))^2)/(2*(length(y)-1))
```

   What is being computed here?