FIRST YEAR EXAM

Tuesday May 10, 2005; 9:30 - 12:30

NOTES: PLEASE READ CAREFULLY BEFORE BEGINNING EXAM!

1. Do not write solutions on the exam; please write your solutions on the paper provided.
2. Put the problem number and your assigned code on the top of **each page**.
3. Write on only **one side** of the page (solutions on the reverse side of the page will be ignored).
4. Start each problem on a **new page**.
5. It is to your advantage to show your work and explain your answers. Draw a line through work you do not want graded; you do not need to erase.
6. You should attempt to work as many parts as feasible.
7. Students in the MS program are not responsible for material from STA 205; their exams will be graded accordingly. Otherwise all problems are graded out of 10 points and carry equal weight.
8. You have 3 hours to finish.
9. This is a closed book exam. No notes are permitted.
10. The Takehome Exam is due at 10 AM May 12th and should be handed in to Krista Moyle in Room 223 Old Chemistry.

1. There is often interest in regression models in which it is assumed that one function applies in a certain range of X and that another in a different range. For example, a general segmented linear regression of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad X_i \leq t_* \tag{1}$$
$$Y_i = \alpha_0 + \alpha_1 X_i + \epsilon_i \quad X_i > t_* \tag{2}$$

with independent, identically distributed normal errors, $\epsilon_i \sim N(0, \sigma^2)$. The model is continuous if $\beta_0 + \beta_1 t_* = \alpha_0 + \alpha_1 t_*$, and discontinuous otherwise. Assume that t_* is known.

Let $I(z)$ be the indicator function with value 0 if $z \leq 0$ and 1 otherwise and define the function $(z)_+ \equiv \max(0, z)$. Consider fitting the regression model

$$Y_i = \delta_0 + \delta_1 X_i + \delta_2 I(X_i - t_*) + \delta_3 (X_i - t_*)_+ + \epsilon_i \tag{3}$$

where the errors, ϵ_i, are again independent and identically distributed $\mathrm{N}(0, \sigma^2)$.

(a) Give the relations between the parameters $\beta_0, \beta_1, \alpha_o, \alpha_1$ and $\delta_0, \delta_1, \delta_2, \delta_3$.

(b) Based on the parameterization in model (3), show how fitting model (3) can be used to test the null hypotheses

H_1**:** continuity of the segmented regression, $\beta_0 + \beta_1 t_* = \alpha_0 + \alpha_1 t_*$
H_2**:** identity of the two segments $\beta_0 = \alpha_0, \beta_1 = \alpha_1$.

(c) Air quality monitors for recording particulate levels often require adjustment and re-calibration. There are 241 measurements of the response variable Y, (log of fine particulate matter concentrations), taken a short time before and after an adjustment in the monitor. The recording time is denoted as X and the time of the adjustment of the monitor is t_*. Using the attached output, conduct the above two tests. In the output the variable $I(X_i - t_*)$ is denoted as `X.gt.t` and $(X_i - t_*)_+$ is denoted as `max.X.gt.t`. What conclusions can you draw pertaining to H_1 and H_2? What does this imply about the monitor?

```
> lm.pm = lm(Y ~ X + max.X.gt.t + X.gt.t, data=polldat)
> summary(lm.pm)

Call:
lm(formula = Y ~ X + max.X.gt.t + X.gt.t, data = polldat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.87247 -0.23428 -0.02956  0.21025  1.34733

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.1442283  0.0633881  18.051   <2e-16 ***
X            0.0003605  0.0001615   2.232   0.0265 *
max.X.gt.t  -0.0005556  0.0004455  -1.247   0.2136
X.gt.t      -0.2533756  0.3482097  -0.728   0.4675
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3543 on 237 degrees of freedom
Multiple R-Squared: 0.3604,     Adjusted R-squared: 0.3523
F-statistic: 44.51 on 3 and 237 DF,  p-value: < 2.2e-16

> anova(lm.pm)
Analysis of Variance Table

Response: Y
            Df  Sum Sq Mean Sq F value    Pr(>F)
X            1  8.6893  8.6893 69.2174 6.985e-15 ***
max.X.gt.t   1  8.0074  8.0074 63.7851 6.007e-14 ***
X.gt.t       1  0.0665  0.0665  0.5295    0.4675
Residuals  237 29.7522  0.1255
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2. Three random quantities x, y, z have joint probability density functions (pdf) of the form $p(x, y, z) = p(y|x, z)p(x)p(z)$ where $(y|x, z) \sim N(y|a + bx + cz, v)$, $x \sim N(x|m, M)$ and $z \sim N(z|h, H)$ where the parameters (a, b, c, v, m, M, h, H) are known. Derive the following distributions. *(All results used must be quoted or proven as part of your answer. Your answer may be either given as a density function or as a distribution,* $N(\mu, \sigma^2)$*.*

   (a) What is $p(y)$?

   (b) What is $p(x|y, z)$?

   (c) What is $p(y|x)$?

   (d) What is $p(x|y)$?

3. X has a geometric distribution

$$\mathsf{P}[X = x] = p\,q^x,\ x = 0, 1, 2, ...$$

with mean $\mathsf{E}X = 2$, while Y has an exponential distribution

$$\mathsf{P}[Y > y] = e^{-\lambda y},\ y > 0$$

with mean $\mathsf{E}Y = 1$, with X and Y independent.

(a) Find p and λ.

(b) What is the probability $\mathsf{P}[X < Y]$ that X is smaller than Y? Give an *exact* solution or a numeric approximation to at least four correct decimals.

(c) Find the conditional distribution of X, given that $X < Y$.

(d) Find the conditional mean $\mathsf{E}[X \mid X < Y]$ of that distribution— in simple exact form or numerically to at least four correct decimals.

4. Let $X_1 \sim \text{Bern}(\theta)$. Let δ_t be the distribution that is degenerate at t. I.e., if $Y \sim \delta_t$, then $Y = t$ with probability one. Finally, let π_0 be the uniform distribution on $[0, 1]$.

   (a) Statisticians A and B have different priors for θ. Statistician A has prior $\pi_A = (1-\epsilon_A)\pi_0 + \epsilon_A\delta_{t_A}$, while B has prior $\pi_B = (1-\epsilon_B)\pi_0 + \epsilon_B\delta_{t_B}$ for some $\epsilon_A, \epsilon_B, t_A, t_B \in [0, 1]$.

      i. Find their prior expectations $\mathsf{E}_A(\theta)$ and $\mathsf{E}_B(\theta)$.

      ii. Find the prior probability for A: $\mathsf{P}_A[X_1 = 1]$.

   (b) They both observe $X_1 = 1$. Find their posteriors $\pi_A(\theta|X_1 = 1)$ and $\pi_B(\theta|X_1 = 1)$.

   (c) Conjecture: $\mathsf{E}_A(\theta) \leq \mathsf{E}_B(\theta) \Leftrightarrow \mathsf{E}_A(\theta|X_1 = 1) \leq \mathsf{E}_B(\theta|X_1 = 1)$. Either prove the conjecture or give a counterexample.

   (d) $X_1, \ldots, X_{10} \sim$ *i.i.d.* $\text{Bern}(\theta)$. Statisticians C and D have priors $\pi_C = (1-\epsilon)\pi_0 + \epsilon\delta_0$ and $\pi_D = (1-\epsilon)\pi_0 + \epsilon\delta_1$. (They both have the same ϵ.) They observe $\sum_{i=1}^{10} X_i = 7$. Which of the following statements are true?

      i. $\mathsf{E}_C(\theta|X_1, \ldots, X_{10}) \leq \mathsf{E}_D(\theta|X_1, \ldots, X_{10})$

      ii. $\mathsf{E}_D(\theta|X_1, \ldots, X_{10}) \leq \mathsf{E}_C(\theta|X_1, \ldots, X_{10})$

      A: i only
      B: ii only
      C: both i and ii
      D: neither i nor ii

5. The random variable X can take on only three possible values, $\{1, 2, 3\}$. There are two possibilities for the probability distribution $p(x) = \mathsf{P}[X = x]$ of X, given in the following table (where they are labeled $\theta = 0$ and $\theta = 1$):

$p(x \mid \theta)$ =

	$x = 1$	$x = 2$	$x = 3$
$\theta = 0$	0.05	0.45	0.50
$\theta = 1$	0.75	0.15	0.10

(a) Find the rejection region $\mathcal{R}$ for *each* Likelihood Ratio test of $H_0 : \theta = 0$ *vs.* $H_1 : \theta = 1$, based on a single observation of X. How many distinct tests are there? For each, give the size α and the power $1 - \beta$:

(b) With a uniform prior distribution π assigning probability $1/2$ to both $\theta = 0$ and $\theta = 1$, find the posterior probability $\pi(H_0 \mid x)$ of H_0 for each of the possible values x of a single observation X.

(c) Now consider observing $N = 100$ independent observations $\{X_j\}$; let $\vec{x} = (x_1, ..., x_{100})$ be the outcome (each x_j is 1, 2, or 3). Find a minimal sufficient statistic $T(\vec{x})$ for θ. Note T might be vector valued.

(d) The most powerful test of H_0 *vs.* H_1 on the basis of $n = 100$ observations $\vec{x} = (x_1, ..., x_{100})$ will reject H_0 for large values of $c \cdot T(\vec{x})$, for some vector c. Find c (up to an arbitrary scale factor).

6. Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space with $\Omega = (0, 1]$ the unit interval, $\mathcal{F}$ the Borel sets, and P Lebesgue measure (or length). Since there are only countably many rational numbers q in Ω, it is possible to enumerate them $\{q_n\}$. Let $\{q_n\}$ be any enumeration of the rationals (so $\Omega \cap \mathbb{Q} = \{q_n\}_{n\in\mathbb{N}}$ and $q_n \neq q_m$ for $n \neq m$) and set

$$X_n(\omega) = 1_{(0,q_n]}(\omega) = \begin{cases} 1 & 0 < \omega \leq q_n \\ 0 & q_n < \omega \leq 1 \end{cases}, \qquad \omega \in \Omega.$$

Which of the following are True or False? Show why your answer is correct by sketching a proof or give a counterexample.

(a) $X_n \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ for each $n \in \mathbb{N}$, i.e., $\mathsf{E}|X_n| \leq \infty$.

(b) $\{X_n\}$ is Uniformly Integrable.

(c) There exists a subsequence $1 \leq n_1 < n_2 < n_3 < ...$ of integers $n_k \in \mathbb{N}$ along which $X_{n_k} \to 1$ almost-surely as $k \to \infty$.

(d) For some $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$, $X_n \to X$ in L_1.

First Year Exam - Takehome

Turn into Krista in Room 223 by 10 AM May 12, 2005

The following data consist of prices for ten different drugs over several years in several regions of California. Questions of interest include: Is there any evidence of quantity discounts or price premiums for quality? Are there regional differences or temporal effects? Conduct a statistical analysis of the price data and summarize your findings in a typed two page report. You may include a supplemental appendix of no more than five pages with any other key figures, output or more technical expressions to support your analysis that are not included in the main text. All figures and computer output should be clearly labeled and annotated. Any results in the appendix should be referenced in the body of the report.

The data are in the file `http://www.stat.duke.edu/programs/grad/fye/drugs.asc`; data items are all tab delimited. Definitions of variables are given in the comment field before the file header.
Header and first 4 lines:

```
Year Locat'n Drug Drg Cde Raw Qty Units Q in Grams Low $ High $ Low Pur Hi Pur Avg. $ Av Pur Pure Qty

1984 CA1 Cocaine-Powder C 1 kilo 1000 gram 50000 65000 50 75 57500 62.5 625
1984 CA1 Cocaine-Powder C 1 pound 453.6 gram 35000 35000 50 75 35000 62.5 283.5
1984 CA1 Cocaine-Powder C 1 ounce 28.35 gram 2000 2400 18 20 2200 19 5.3865
1984 CA1 Cocaine-Powder C 0.5 ounce 14.175 gram 1000 1400 18 20 1200 19 2.69325
```