AMS 205B (Intermediate Classical Inference)

Take-Home Test 4

Absolute due date 21 Mar 2014

Here are the ground rules: this test is open-book and open-notes, and consists of two problems (essay and calculation); the essay is worth 100 points and the calculation problem is worth 50 points, for a total of 605 points.

The right answer with no reasoning to support it, or incorrect reasoning, will get half credit, so try to make a serious effort on each part of each problem (this will ensure you at least half credit). In 2012, on one take-home test in this class there were 15 true/false questions, worth a total of 150 points; one student got a score of 92 out of 150 (61%, a D−, in a graduate class where B− is the lowest passing grade) on that part of the test, for repeatedly answering just “true” or “false” with no explanation. Don’t let that happen to you.

On non-extra-credit problems, I mentally start everybody out at −0 (i.e., with a perfect score), and then you accumulate negative points for incorrect answers and/or reasoning, or parts of problems left blank. On extra-credit problems, the usual outcome is that you go forward (in the sense that your overall score goes up) or you at least stay level, but please note that it’s also possible to go backwards on such problems (e.g., if you accumulate +3 for part of an extra-credit problem but −4 for the rest of it, for saying or doing something egregiously wrong).

This test is to be entirely your own efforts; do not collaborate with anyone or get help from anyone but me. The intent is that the course lecture notes and readings should be sufficient to provide you with all the guidance you need to solve the problems posed below, but you may use other written materials (e.g., the web, journal articles, and books other than those already mentioned in the readings), provided that you cite your sources thoroughly and accurately; you will lose (substantial) credit for, e.g., lifting blocks of text directly from wikipedia and inserting them into your paper without full attribution.

If it’s clear that (for example) two people have worked together on a part of a problem that’s worth 20 points, and each answer would have earned 16 points if it had not arisen from a collaboration, then each person will receive 8 of the 16 points collectively earned (for a total score of 8 out of 20), and I reserve the right to impose additional penalties at my discretion. If you solve a problem on your own and then share your solution with anyone else (because people from your cultural background routinely do this, or out of pity, or kindness, or whatever motive you may believe you have; it doesn’t matter), you’re just as guilty of illegal collaboration as the person who took your solution from you, and both of you will receive the same penalty. This sort of thing is necessary on behalf of the many people who do not cheat, to ensure that their scores are meaningfully earned. In 2012, five people failed the class because of illegal collaboration; don’t let that happen to you.

Last ground rule: proof by Maple or some other symbolic computing package is not acceptable; when I ask you to show something, please do so by hand (you can check your results with (e.g.) Maple, but you need to do the work yourself).
1 Essay

[100 total points] In no more than three pages, summarize your principal conclusions from this course regarding the strengths and weaknesses of the frequentist and Bayesian paradigms for inference, prediction and decision-making, concentrating mainly on inference but making comments about the other two statistical activities as well. The reasoning that supports your conclusions should be specific and clearly grounded in the history of probability and statistics, as it was presented in class. In building your argument, briefly review that history, selecting highlights as appropriate from the work of the following people (obviously you don’t need to cover all of these people; in fact you won’t have enough space to do so): Pascal and Fermat, Bayes, Laplace, Gauss, Venn, Karl Pearson, Gossett, Fisher, Ramsay, de Finetti, Kolmogorov, Neyman and Egon Pearson, Richard Cox, Wald, Efron and Jaynes; comment briefly on what each of {the people on whom you choose to focus} contributed, and highlight moments when a new frequentist method was proposed and developed that was a definite improvement on previous methods (explaining in what ways the improvements were better than what came before). Make sure you cover at least the following topics: intervals based on the Central Limit Theorem, method-of-moments estimation, unbiased estimation, maximum-likelihood point estimates, observed information, confidence intervals of three kinds (approximate likelihood-based, exact small-sample, bootstrap), asymptotic relative efficiency, and hypothesis and significance testing. Conclude with a brief discussion of where frequentist inference could go from here to improve it further.

Note: As mentioned above, looking (e.g., on the web) at sources other than those from class is OK; if you want to broaden and deepen your understanding of the history of probability and statistics, above and beyond what we did this quarter, please feel free to do so as part of the process of constructing your essay. If you do bring other sources into your argument, cite them fully (and I don’t mean it’s OK to grab blocks of text from somewhere on the web and just stick them into your essay; if you borrow somebody else’s ideas or text fragments, in addition to citing these borrowed items properly You need to express them in Your own words).

2 Calculation

[505 total points] You’ll notice that, in this final test, I’m giving you less guidance on how to solve (at least some of) the problems posed; that’s intentional.

(A) [90 total points] You have a univariate data set $y = (y_1, \ldots, y_n)$; for example, consider again the U.S. family income data from Take-Home Test 3. You try a standard parametric model, such as the lognormal distribution, and you find (as you did in the take-home test) that the fit is not great. Rhetorical Question: What now?

There are two basic possibilities: change the model, or change the data. The first option represents the Bayesian impulse: try to find a model for the data set as it comes to you. The second approach was popular in the last 50 years of the 20th century, under the name exploratory data analysis (EDA), and it’s still commonly used today by frequentists (and sometimes also by Bayesians). The idea is to find a transformation of the data values such that the transformed data set does fit a standard parametric model.

For data sets (such as the U.S. family income sample) in which all of the $y_i$ values are strictly positive and the variable has a long right-hand tail, people often think of transforming
via \( w_i = \log y_i \), but other non-linear transformations are possible; for example, you could consider \( w_i = \sqrt{y_i} \) or \( w_i = \frac{1}{3} y_i \), or more generally \( w_i = y_i^\lambda \) for some \(-1 \leq \lambda \leq 1\).

1. Graphically demonstrate on the family-income data, with a reasonable number of \( \lambda \) grid points (avoiding \( \lambda = 0 \) for the time being), that as \( \lambda \) decreases from 1 to \(-1\) the transformed data set moves closer to Gaussian and then away from it; this suggests that there’s an optimal \( \lambda \) for which \( w_i = y_i^\lambda \) is closest to Gaussian. [10 points]

2. Something funny clearly happens to \( w_i = y_i^\lambda \) when \( \lambda = 0 \). Show via calculus that

\[
\lim_{\lambda \to 0} \frac{y_i^\lambda - 1}{\lambda} = \log \lambda.
\]  

This suggested to Box and Cox (1964), in a paper that’s been cited more than 8,200 times (in statistics any paper with 1,000 or more citations is regarded as a classic), that we can rescue the desirable idea of treating \( \lambda \) continuously on \([-1, 1]\) through the power-transformation family

\[
w_i = \begin{cases} 
\frac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\
\log \lambda & \text{if } \lambda = 0
\end{cases} 
\]  

(again, this only applies if \( y_i > 0 \)). Briefly explain why \( \frac{y_i^\lambda - 1}{\lambda} \) has the same transformational effect toward the Gaussian distribution as \( y_i^\lambda \) does when \( \lambda \neq 0 \), thereby verifying that (2) does indeed define a potentially useful family of transformations in which \( \lambda \) can be permitted to vary continuously from \(-1\) to \(+1\). [20 points]

3. Having found a \( \lambda \) such that \( w_i \) in (2) is close to Gaussian, it would be natural to complete the modeling with

\[
(w_i|\mu, \sigma^2) \sim \text{N}(\mu, \sigma^2) \quad (i = 1, \ldots, n); 
\]  

equations (2) and (3) together define a model in which the parameter vector is \( \theta = (\mu, \sigma, \lambda) \). Fit this model via maximum likelihood to the U.S. family income sample, obtaining point estimates and 95% confidence intervals for all three components of the \( \theta \) vector. Draw the family income histogram yet again, this time superimposing the density curve for the sampling distribution implied by (2, 3) with your MLEs for the parameter values. Does this model fit this data set better than the Gamma and Lognormal models did? Explain briefly. [60 points]

(B) [220 total points] High levels of lead (chemical symbol \( Pb \)) ingested by humans can have strong negative effects on health, especially in children. In the early part of the 20th century the two principal vectors for human exposure to lead were paint (lead was added to make the paint smoother and easier to apply) and gasoline (lead increased fuel efficiency by lubricating the cylinder walls in internal combustion engines). The U.S. government mandated a gradual reduction of lead in gasoline starting in the early 1970s, from 4.0 grams of lead per gallon to 0.5 grams in 1985 and 0.1 grams in 1986, and lead was banned completely from gasoline in 1995. Rhetorical Question: Did this cause significant reductions in lead levels in the environment, and if so, how quickly did the lead concentrations drop?

Two statisticians at Purdue University, David Moore and George McCabe, authors of *Introduction to the Practice of Statistics* (third edition, 1999), got an interesting data set,
relevant to answering this question, from an environmental scientist named Tom Siccama (Yale University School of Forestry and Environmental Studies). As Moore and McCabe describe it,

“The presence of lead in the soil of forests is an important ecological concern. One source of lead contamination [used to be] the exhaust from automobiles. In recent years this source has been greatly reduced by the elimination of lead from gasoline [in the 1970s and 1980s]. Can the effects be seen in U.S. forests? The Hubbard Brook Experimental Forest in West Thornton NH is the site of an on-going study of the forest floor. Lead measurements of [representative] samples [of soil] taken from this forest are available for a number of years. The variable of interest is lead concentration, recorded as milligrams per square meter [of topsoil, gathered down to a depth of 0.1 inches].”

The file lead.txt on the course web site contains data from this Experimental Forest in the years 1976 (59 observations), 1977 (58), 1978 (58), 1982 (68) and 1987 (70 observations); the first column is the year with 100 subtracted from it (e.g., 76), and the second column is the observed lead level (sorted from smallest to largest within year).

(1) In this part of the problem, treat year as a categorical predictor of lead, at $k = 5$ levels.

(a) Perform a graphical and numerical descriptive analysis of this data set, summarizing your findings and focusing on the following question. Fisher’s off-the-shelf ANOVA assumptions include homoscedasticity (the population standard deviations (SDs) in all of the groups being compared are supposed to be equal) and normality within each group; do these assumptions appear to be met with this data set in its raw form? Explain briefly. [20 points]

(b) Do the differences between the mean lead levels by year seem large to you in practical (environmental) terms? Explain briefly. (Hint: Relative comparisons are more useful here than absolute comparisons [why?]) [10 points]

(c) Explain why your analysis in (a) indicates that the data are crying out for a log transform, and repeat (a) on the outcome variable $w = \log y$. [20 points]

(d) Using the log lead data from (b) and writing your own code, make a table with 10 rows (one for each pairwise comparison, i.e., 1977 versus 1976, 1978 versus 1976, ..., 1987 versus 1982), and the following columns: comparison, point estimate, estimated standard error of the point estimate, 95% interval without multiple-comparisons adjustment, and 95% interval with Bonferroni adjustment. Identify which pairwise comparisons are statistically significant (i) without and (ii) with multiple-comparisons adjustment. Taking both practical and statistical significance into account, in your judgment how strong is the evidence from this data set that the U.S. gradual governmental intervention starting in the early 1970s caused significant reductions in lead levels in the environment? Explain briefly. [50 points]

(2) As interesting as the analysis in (i) has been, it potentially doesn’t make full predictive use of the year variable, because that variable is quantitative and the work in (i) treated it qualitatively.

(a) Using the formulas we developed in class and writing your own code, fit the standard simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + e_i$ (with the $e_i$ IID Gaussian
with mean 0 and variance $\sigma^2$), to predict $y = \text{lead}$ from \textbf{year} — including point estimates for the components of $\theta = (\beta_0, \beta_1, \sigma^2)$, a 95\% confidence interval for $\beta_1$, the $R^2$ value, and the root mean squared error (residual SD) — and check your answers with the \texttt{lm} command in R. Perform graphical and numerical diagnostic checks on the residuals, and briefly explain why they indicate that the model doesn’t fit well. [40 points]

(b) Repeat (a) with the outcome variable $w = \log \text{.lead}$. Is the regression slope statistically significant? Explain briefly. [30 points]

(c) In addition to wondering in (b) if the slope is large in statistical terms, it’s also important to consider whether it’s practically significant. Let $x_{\text{old}}$ be a value of the \textbf{year} variable not far from the mean of that variable, and consider the change in the predicted $y$ (i.e., \text{lead}) value associated with moving from $x_{\text{old}}$ to $x_{\text{new}} = x_{\text{old}} + 1$, based on your results in (b) in which $w$ (i.e., $\log \text{.lead}$) was regressed on $x$. Fill in the missing steps (\ldots) of the following line of reasoning, offering brief explanations in each case.

$$
\hat{w}_{\text{old}} = \left( \log y \right)_{\text{old}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{old}}
$$

$$
\hat{w}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}
$$

$$
\exp \left( \log y \right)_{\text{new}} = \exp \left( \log y \right)_{\text{old}} \cdot e^{\hat{\beta}_1}
$$

$$
\frac{\hat{y}_{\text{new}} - \hat{y}_{\text{old}}}{y_{\text{old}}} \approx \hat{\beta}_1.
$$

Use the last line of (4) to complete the following sentence: “Associated with the passing of each year from 1976 to 1987, there was a decrease in the mean lead level in the soil at the \textbf{Hubbard Brook Experimental Forest} of about ____, give or take about ____” (here, as previously in this course, “give or take” is synonymous with “standard error”). Does this match, at least approximately, the pattern observed in the sample means for each year? Does this difference (between the observed regression slope and a null slope of 0) seem large to you in practical terms? Explain briefly. [40 points]

(3) Which of the analyses in parts (B) (1) and (B) (2) do you regard as a more complete scientific summary of what the data set has to say, or do they both add useful information? Explain briefly. [10 points]

(C) [195 total points] How should regression be performed when the outcome variable is dichotomous? One leading approach is \textit{logistic regression}, which is based on the following model, when (in its simplest form) you have a single quantitative predictor variable $x = (x_1, \ldots, x_n)$ (whose values are regarded as fixed known constants) and your dichotomous outcome variable is $y = (y_1, \ldots, y_n)$:

$$
\begin{align*}
(y_i | p_i) & \overset{\text{indep}}{\sim} \text{Bernoulli}(p_i) \quad (i = 1, \ldots, n) \\
\log \left( \frac{p_i}{1 - p_i} \right) & = \beta_0 + \beta_1 x_i. 
\end{align*}
$$

(5)
(1) Express $p_i$ in terms of the linear predictor $\beta_0 + \beta_1 x_i$, and use this to show that the log-likelihood function in this model is

$$\ell(\beta_0, \beta_1 | y) = \beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \log \left(1 + e^{\beta_0 + \beta_1 x_i}\right).$$  \hspace{1cm} (6)

Does a two-dimensional sufficient statistic exist for estimating the parameter vector $\theta = (\beta_0, \beta_1)$? Explain briefly. [30 points]

(2) Can the method of moments be used in this problem to find good starting values in an iterative numerical search for the maximum-likelihood estimates (MLEs)? Explain briefly. [10 points]

(3) Show that the equations defining the MLEs of the components of $\theta$ can be expressed in the following form:

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_i, \quad \frac{1}{n} \sum_{i=1}^{n} x_i y_i = \frac{1}{n} \sum_{i=1}^{n} x_i \hat{p}_i \right\},$$  \hspace{1cm} (7)

where $\hat{p}_i$ is the maximum-likelihood estimate of $p_i$. Briefly explain why each of these equations makes good intuitive sense as guidance for how high-quality estimates of $\theta$ should behave. [20 points] Extra credit: See if you can figure out how to use these equations to define an algorithm that will successfully find the MLEs.

(4) The data set sepsis-anc-age-i2t.txt on the course web page, which arose in my collaborative work with investigators in the Division of Research at the Kaiser Permanente Northern California hospital chain, contains data on four variables measured on each of $n = 66,846$ newborn babies randomly chosen from a population $P$ to which we wish to inferentially generalize. Reading from left to right (when this data set is made into a matrix with $n$ rows and 4 columns), the first column is 1 if the baby developed SEPSIS (a full-body bacterial or viral infection: this is an extremely bad outcome for newborns [they typically either die or go into a permanent coma]) and 0 otherwise; the second and fourth columns measure two different variables (called ANC and I2T, respectively) that derive from the results of a blood test called a CBC (complete blood count), and the third column is the AGE of the baby (in hours) at which the CBC results were obtained. The idea in this part of the problem is to explore the extent to which ANC, AGE and I2T are helpful in predicting which babies will get sepsis.

(a) To get a sense of how complicated the relationship between these four variables is, download sepsis-anc-age-i2t.txt and sepsis-data-analysis.txt from the course web site into a directory, start an R session, change directory inside R to the place where the two .txt files are, and issue the command

```
source( "sepsis-data-analysis.txt" )
```

to create a useful graphical descriptive summary of the data set. Modify the code in sepsis-data-analysis.txt to create a .ps or .pdf file containing the graph that results from the source command above, and include this plot in your answer to this question (the graph uses color, but you can print it in black and white and the message will still come through). Study the code until you understand what the plot is saying, and use this understanding to describe what you observe about how SEPSIS relates to ANC, I2T and AGE. Is sepsis rare in this data set, or frequent, or in between? Does ANC appear to be predictive of sepsis? How about I2T? AGE? Explain briefly. [30 points]
Table 1: Cross-tabulation of predicted sepsis status (rows) and true sepsis status (columns) for a generic cut-point $c$.

<table>
<thead>
<tr>
<th>Predicted Sepsis</th>
<th>True Sepsis</th>
<th>1</th>
<th>0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>nTP</td>
<td>nFP</td>
<td>244</td>
<td>66,602</td>
<td>66,846</td>
</tr>
<tr>
<td>nFN</td>
<td>nTN</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Initially choose $I2T$ as your $x$ variable. Use equation (6) to find the maximum-likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and their standard errors; build an approximate 95% confidence interval for $\beta$ (writing your own code to do all of these things), and check your analysis with the `glm` command in R. Is $I2T$ a statistically significant predictor of SEPSIS? Explain briefly. [30 points]

(c) Compute the predicted sepsis probabilities $\hat{p}_i$ from your logistic regression in (b); are the equations in (7) satisfied (at least approximately) with your $\hat{p}_i$ values? If the logistic regression is extremely successful at predicting sepsis from $I2T$, then all of the $\hat{p}$ values for babies who got sepsis would be close to 1 and all of the $\hat{p}$ values for non-septic babies would be near 0; is that true here? Explain briefly. [20 points]

(d) The purpose of building predictive models in this problem is to give clinicians a rule of the form {if $\hat{p} > c$, it’s predicted that the baby will develop sepsis}, so that such babies can be identified early and treated aggressively to prevent the sepsis from happening. Of course, just as was true with ELISA in the HIV diagnosis case study many weeks ago, increasing $c$ makes the false-positive rate go down but the false-negative rate go up (where “positive” means “predicted sepsis”), and decreasing $c$ has the opposite effect. Rhetorical Question: Where should you draw the line? (Note that this is a decision-making, not inferential, issue.)

(i) Consider the utility function of the parents of one of these babies. Which is worse from their point of view — false positives or false negatives — or are they equally bad? If one is worse than the other, how much worse (a little, a medium amount, a lot)? Explain briefly. [10 points]

(ii) Now consider the utility story from the hospital’s point of view: failing to prospectively identify septic babies is bad, but needlessly treating babies that aren’t going to get sepsis is also bad for those babies (as I said above, the treatment is aggressive, and it will harm some of them). Which is worse from the hospital’s point of view — false positives or false negatives — or are they equally bad? If one is worse than the other, how much worse (a little, a medium amount, a lot)? Explain briefly. [10 points]

(iii) To illustrate the false-positive/false-negative trade-off here, consider the following three possible values of $c$ (with the $\hat{p}$ values based only on $I2T$, as in your logistic regression in (C) (4) (b) above): $c = (0.025, 0.05, 0.1)$. Choosing a value of $c$ permits you to create a dichotomous variable taking the value 1 if $\hat{p} > c$ and 0 otherwise, and this variable can be cross-tabulated against SEPSIS to enumerate the true positives (TPs), false positives (FPs), false negatives (FNs) and true negatives (TNs) arising from this decision rule. In this problem...
all such cross-tabulations will look like Table 1 above, with the column totals fixed by the observed SEPSIS values in this data set. Create the three tables corresponding to $c = (0.025, 0.05, 0.1)$ and discuss which comes closest to maximizing expected utility for (I) the parents and (II) the hospital. Do any of them represent good predictive accuracy? Explain briefly. [30 points]

(iv) How can you create better $\hat{p}$ values in this problem (where “better” means “leading to predictive rules with better false-positive and false-negative behavior”)? Explain briefly. [5 points]

Extra credit for part (C) (4): Using glm with ANC and AGE in addition to I2T (and — if you know how to do this — interaction and quadratic terms), build a better predictive model than {the one in (C) (4) (b) that was based only on I2T}, and quantify how much better your new model is than the I2T-only model in correctly identifying septic babies.