STAT 422/722 Spring 2016 Homework #2

Professor Adam Kapelner

Optionally Due *4th floor JMHH* Thursday, February 15 5PM

(this document last updated Wednesday 8th February, 2017 at 4:14pm)

Instructions and Philosophy

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The green problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the purple problems.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 10 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and the programs for compiling LaTeX is written about in the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, (1) upload `hwxx.tex` and `preamble.tex` from the correct github folder, (2) read the comments in the code as there is *one line to comment out*, (3) you should replace my name with your name and (4) your section. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, you must print this document and write in your answers. You must print after downloading and opening in Adobe reader (not from Google Chrome viewer). I do not accept homeworks not on the correctly paginated printout of this document. Write your name and section below (A or B).

You may collaborate, but hand in your own copy with your own wording. See the syllabus for more information.

NAME: _____ COURSE (422 or 722): _____

SECTION ("A" for Tuesday or "B" for Wednesday): _____

Problem 1

We will be investigating equivalence testing.

(a) [easy] In the context of linear or logistic regression, if you want to prove that a predictor has a linear effect on the response (controlling for other variables), what are the null and alternative hypotheses?

(b) [easy] In the context of linear or logistic regression, if you want to prove that a predictor does not have a linear effect on the response (controlling for other variables), what are the null and alternative hypotheses?

(c) [easy] You collect four data points

predictor	response
2.47	0.50
0.57	1.95
0.84	1.91
2.18	2.51

Test the theory in (a)

(d) [harder] Test the theory in (b). Use $\delta = 0.5$ as a margin of practical equivalence

(e) [difficult] How can you get both (c) and (d) at the same time? Discuss

Problem 2

We will be investigating dredging and multiple testing corrections. I have provided a data file for you called "xyrand.csv" located here (right click and downlod from the browser). This file is fully random data from a standard normal and thus there is no systematic connection between the column y and any of the x_j columns.

(a) [easy] Run a regression of y on the x_j's and report R^2. Why is this R^2 not *exactly* zero?

(b) [harder] Explain why RMSE is the value that it is. What number do you feel it should it be closer to?

(c) [easy] Which variables were significant and what are their significance levels? Why should any variables be significant in the first place if they're all just $\overset{iid}{\sim}$ random

realizations?

(d) [harder] Calculate the probability you see this many significant variables *or more* in a 50-predictor linear regression. Is your number of significant variables "expected"?

(e) [easy] Calculate both a Sidak and a Bonferroni corrected individual α that preserves 5% familywise error.

(f) [easy] Using the Sidak and/or the Bonferroni correction, are there any significant variables anymore? Yes/no

(g) [harder] Explain precisely what I would need to simulated in this same setup with one response and 50 predictors randomly realized to expect one significant variable if the familywise correction is employed.

(h) [harder] Report the overall F value and it's corresponding significance level. Explain how it is possible that there exist t tests which are significant for some linear predictors but the F test is not.

Problem 3

These are some conceptual questions concerning hypothesis testing, Type I errors, Type II errors and power.

(a) [easy] Given some predetermined level of α we have two ways of setting up hypothesis tests:

$$H_0 : \text{UFOs do not exist}$$
$$H_a : \text{UFOs do exist}$$

and the inverse:

$$H_0 : \text{UFOs do exist}$$
$$H_a : \text{UFOs do not exist}$$

Which set should be employed and *precisely* why?

(b) [easy] Imagine a situation where you are trying to convince your friends that UFOs exist. What do you provide to your friend to help your case? (One word).

(c) [harder] You cannot convince your friend. In the hypothesis testing framework there are two separate reasons why he is not convinced. What are they?

(d) [difficult] The Shapiro-Wilk Test of Normality is used to assess normality of a given sample of data. It is a goodness-of-fit test where

$$H_0: \text{ the data generating process is normal}$$
$$H_a: \text{ the data generating process is not normal.}$$

Usually, you want to prove normality (e.g. the case of testing the residuals from a linear regression). Why does this test reward small sample sizes?

Problem 4

These questions will be about extrapolation and generalization.

(a) [harder] Is model extrapolation and model generalizablity the same concept (lecture 3, slide 6)? Discuss.

(b) [difficult] Provide an example of a model that you use regularly that does not generalize to the observations you use to predict with it.

(c) [harder] Run lines 5–27 of the lecture 3 R demos. Which of the three models would be the worst extrapolator and why?

(d) [difficult] You are provided a new $\boldsymbol{x}^*$ with p features in which are to guess y. How would assess extrapolation? Explain.

Problem 5

These questions will be about optimal design.

(a) [easy] In the case of a simple linear regression with $n = 20$ points where x ranges from 6 to 17, what would be the optimal design?

(b) [easy] Show that for fixed n under least squares regression that the optimal design is half the points on the minimum and half the points on the maximum.

(c) [easy] Take the case of $n = 20$ and three continuous predictors ranging in $[0, 1]$. Use JMP to create an optimal design for the model with all factorial interactions and all polynomials up until degree 3. Take the optimal design right click and make data

table. Then sort the data table first by x_1 then by x_2 and by x_3. Are they all about the same? Yes/no.

(d) [difficult] The optimal design in the previous problem is what you'd probably like to do for non-parametric linear model with lots of interactions and curves. What is the takehome message in this case?

(e) [E.C.] Your goals are prediction and you have the choice bertween D-optimality and I-optimality. Which is likely better and why?

Problem 6

These questions will be about logistic regression using the Telecom Churn dataset that can be downloaded here.

(a) [easy] Give an expression for the conditional mean in a logistic regression problem with p features using the standard logistic regression assumptions.

(b) [easy] Do the multivariable logistic regression in class with target response churn (remember to delete those 6 variables which are fully collinear). Provide an interpretation

on the monthly charges coefficient.

(c) [E.C.] If we used the cloglog link function and got the same coefficient, what would be the interpretation. This one is fun!

(d) [easy] Predict the mean probability of churn for a senior citzen who has been with the company for 36 months and then predict whether or not this person will churn.

(e) [difficult] Test whether or not removing both gender and partner makes a difference in terms of linear predictive power versus the full model from (b).

(f) [difficult] In my regression output, the χ^2 value for `senior citizen` is 14.37. Calculate the standard error from this.

(g) [easy] Use the model from (b) and use JMP to compute the AUC and misclassification error.

(h) [difficult] Graph the false negative proportion versus the false positive proportion. Is this more useful than an ROC curve for the case of churn?

(i) [E.C.] Create a detection error tradeoff plot for this dataset (no need to use normal deviates for x and y axes).

(j) [harder] Imagine the cost ratio is 7.5:1 for the more costly mistake. What is the p_0 of the optimal model?

(k) [easy] Why is this p_0 less than the naive value of 50%?

(l) [harder] Create a flowchart of the optimal model in (j) similar to lecture 3, slide 35.

Problem 7

These questions will be about survival regression using the NetLixx dataset that can be downloaded here.

(a) [easy] ...