# STAT 422/722 Spring 2016 Homework #1

Professor Adam Kapelner

Due *4th floor JMHH* Thursday, February 2 5PM

(this document last updated Wednesday 25th January, 2017 at 4:03pm)

**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 or B): __

## Problem 1

These questions are about prediction and modeling theory.

(a) [easy] Give three examples of "predictions".

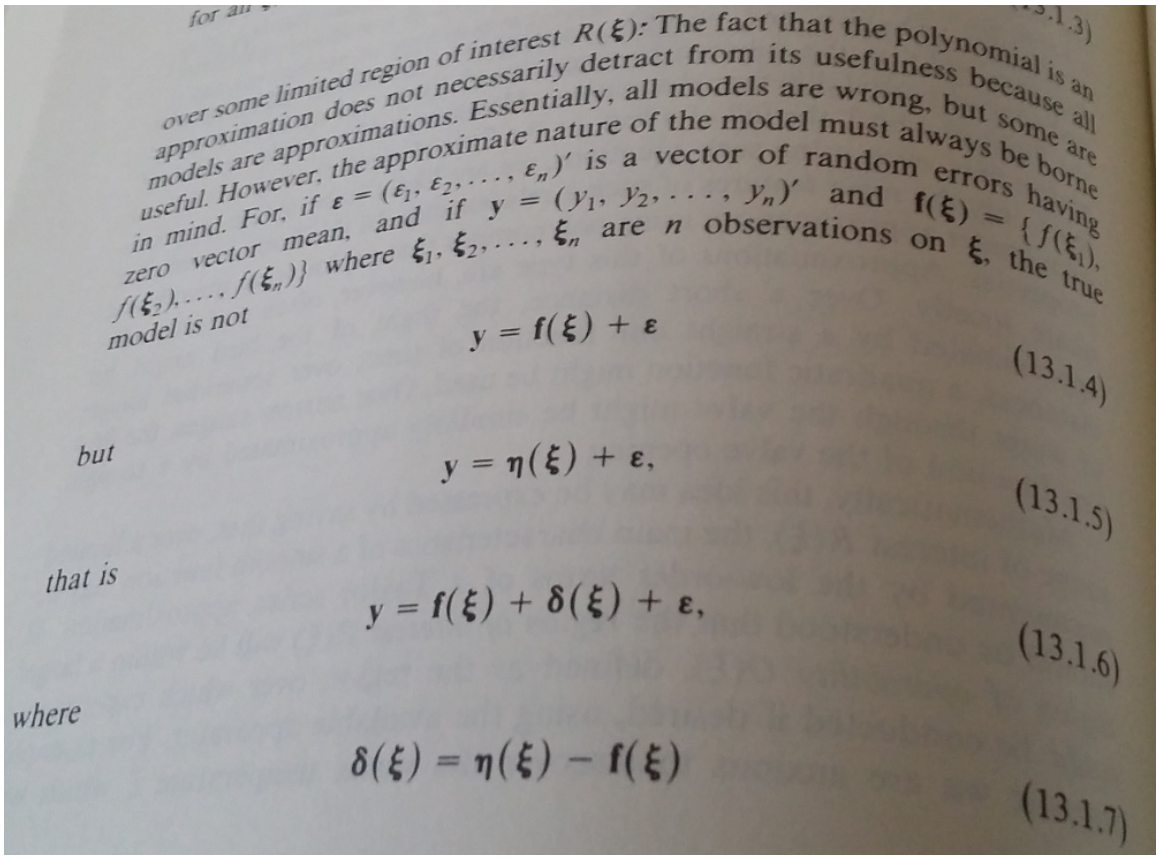
(b) [easy] Considering the etymological definitions, do we "predict" or do we "forecast"? Explain your answer.

(c) [easy] Explain what each row in a dataframe represents. Give every synonym for the rows in a dataframe. Write a sentence as to why each of the particular vocabulary words were employed here.

(d) [easy] Explain what each column in a dataframe represents. Give every synonym for the column. Write a sentence as to why each of the particular vocabulary words were employed here. In the models in this class, there will be *one* special column. Explain what its called and give definitions for all its synonyms.

(e) [easy] Explain why theories are mathematical (generally speaking).

(f) [harder] Below is an excerpt from Box and Draper (1987, page 424) and contains the famous line "all models are wrong, but some are useful". Explain what this means.



over some limited region of interest $R(\xi)$: The fact that the polynomial is an approximation does not necessarily detract from its usefulness because all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind. For, if $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ is a vector of random errors having zero vector mean, and if $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ and $\mathbf{f}(\boldsymbol{\xi}) = \{f(\xi_1), f(\xi_2), \ldots, f(\xi_n)\}$ where $\xi_1, \xi_2, \ldots, \xi_n$ are $n$ observations on $\boldsymbol{\xi}$, the true model is not

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\xi}) + \boldsymbol{\varepsilon} \tag{13.1.4}$$

but

$$\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}, \tag{13.1.5}$$

that is

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\xi}) + \boldsymbol{\delta}(\boldsymbol{\xi}) + \boldsymbol{\varepsilon}, \tag{13.1.6}$$

where

$$\boldsymbol{\delta}(\boldsymbol{\xi}) = \boldsymbol{\eta}(\boldsymbol{\xi}) - \mathbf{f}(\boldsymbol{\xi}) \tag{13.1.7}$$

(g) [harder] What did we call $\delta(\xi)$ in class (ibid, Equation 13.1.7)?

(h) [E.C.] If the true model were to be found and estimated from the a finite sample dataframe, would it be better for inference than a simple model? Yes / no explain. Would it be better for prediction? Yes / no explain.

(i) [harder] In many sciences there is a belief in the so-called "tapering effect" which means that there are large effects, then small effects, then really small effects. How can this be used to explain the success of linear regression in predicting responses with likely myriad inputs (a la slide 32, Lecture 2)?

(j) [E.C.] Why will "full reality always remain elusive in the biological sciences"? (Burnham and Anderson, 1998) What does that say about the softer sciences such as economics, psychology, sociology, etc?

(k) [harder] Explain why non-parametric models can give you lower misspecification error but possibly higher model estimation error.

(l) [harder] I got a phone call from a startup founder who described the idea for a company that predicts startup success. The founder told me their model was that they assign 5 points to a startup for having two or more C-level founders, 10 points for closing an angel round, 5 points for the first employee, etc. What kind of model is this? Why would you trust / not trust its predictions?

(m) [easy] Look at slide 32 in lecture 2 but not slide 33. Write down all observations you have about this picture.

(n) [E.C.] If you find a confounding / lurking variable $Z$, does that mean $Z$ is causal? If yes, explain; if not, provide a counterexample.

(o) [harder] Given a model with one response and 10 variables where the 10 variables are realized simultaneously but the response is realized afterwards, how many models can be posited? Ignore the fact that each functional relationship can be different. See slide 35 lecture 2.

(p) [easy] Why advantage does a "real" correlation have over a *spurious* correlation?

(q) [harder] My friend has six children, all born on Wednesday. Is this "significant"? Discuss.

## Problem 2

THIS PROBLEM IS OPTIONAL. Here we will be analyzing the theory that "skiing is dangerous".

(a) [easy] Define the response(s) and the predictors(s) in this model.

(b) [harder] Mathematize this model. Explain clearly what you are measuring and how it is measured.

(c) [easy] If you were to use a data-driven approach, what would the dataframe look like? What are the datatypes of each variable? Will the eventual model be a regression? Classification? Something else?

(d) [harder] Explain why a deterministic model for your response variable is absurd.

(e) [harder] Create a stochastic (statistical) model for the response. Pay attention to which letters are lowercase/uppercase.

(f) [harder] In this model, would the error term $\mathcal{E}$ be large or small? Explain.

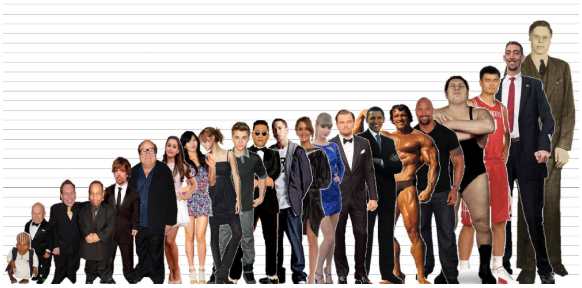
(g) [harder] If you were given leeway to collect a multidimensional representation of "skiing" (i.e. a more natural, raw representation), what would you collect?

(h) [difficult] Build a causal model (using bubbles and arrows) for the response.

(i) [difficult] Does skiing *cause* the response? If so, is it a major contributor? Explain using your diagram from (h).

## Problem 3

We will discuss confounding here. The prevailing data on wage inequality says that women are paid 90 cents on the dollar that men earn for comparable work.



(a) [easy] If you were to do a regression of $y$ : earnings on $x_1$ : employee height, what would the results look like and why? Report the p-values on each $\hat{\beta}_j$'s and the omnibus $F$ statistic.

(b) [easy] Build a causal model for $y$: earnings and $x_1$: employee height and $x_2$ : employee gender. No need to include unknown variables.

(c) [easy] If you were to do a regression of $y$ : earnings on $x_1$ : employee height and $x_2$ gender, what would the results look like? Report the p-values on each $\hat{\beta}_j$'s and the omnibus $F$ statistic. Also say if the $\hat{\beta}_j$ values changed and which direction vs. the regression in part (a).

## Problem 4

A few questions about likelihood. Imagine a simple model where you flip the same coin three times. You are modeling the response "flipping a head" with a statistical model and make a parametric assumption that the event is a Bernoulli r.v.


(a) [easy] What does $\theta$ represent in this model?

(b) [harder] Find the joint probability density / mass function for these three events.

$\mathbb{P}(Y_1, Y_2, Y_3) =$

(c) [easy] Find the likelihood function.

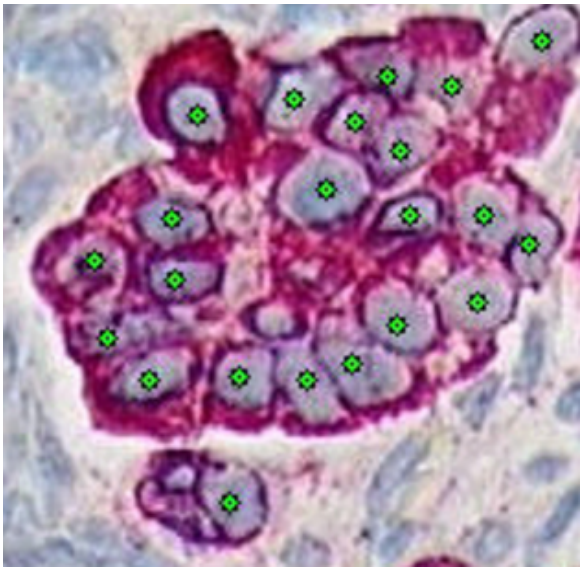
(d) [harder] Find the maximum likelihood estimator for $\theta$.

(e) [easy] If your data was heads, heads, tails (in the image above). What is the maximum likelihood estimate? This will allow you to locate the best possible model given your parametric assumptions.

(f) [E.C.] If your maximum likelihood estimate in (e) was indeed the value of $\theta$, what data would be most probable? Note: this is the inverse question of likelihood and it is meant to trick you.

## Problem 5

Here we will be considering different types of AI. Imagine you have the following problem: you are tasked with finding the centers of cancer cells in microscopic images. Images are composed of pixels which encode a color. Typical coloring schemes for computer graphics are "true color" and use about 16.8 milion different colors per pixel. Here's a typical image:



All cells are stained using immunohistochemistry using a compound which appears blue. But cancer cells in this type of immunohistochemical staining appear to have a red membrane due to a surface marker. (In the example above a green dot is placed in the center of every cancer cell to indicate to you what you are looking for; this dot is *not part of the original image*). Assume we are using a data-driven approach to solve this problem.

(a) [easy] What type of AI would work the best here? No need to discuss.

(b) [easy] What is the raw data representation?

(c) [harder] What is the unit of analysis?

(d) [harder] Why is it easy for you to find the cell centers but difficult for a computer?

(e) [difficult] Consider the situation where you employ classic machine learning. What features would you collect on the units of analysis? Enumerate and describe these features.

(f) [difficult] How would you sample to build a dataframe (collect historical data)? Explain the procedure and the goals of this step. This is known as "supervised machine learning".

(g) [harder] Once you built the dataframe, would a human be able to use that dataframe to create a predictive model? Yes / no and discuss.

(h) [easy] Considering you selected features in (e) and sampled in (f), would this entire enterprise be considered "good machine learning"? Explain.

(i) [easy] Now you have the dataframe. Given the problem context, which worldview would you select — the parametric or the nonparametric and why?

(j) [difficult] Assume you went the parametric model route and you built a linear model. Explain where it would be wrong. Be explicit by referencing your predictors in (e).

## Problem 6

These exercises will dicsuss the linear model and linear regressions.

(a) [easy] Think of three loss functions $L(e_1, \ldots, e_n)$. Do not list any that we did in class.

(b) [harder] You are building a data-driven model and choose to use the linear parametric assumption but not necessarily the other three OLS assumptions. Describe a situation where fitting this model using $L = SSE$ is *not* a good idea because it does not accuractely reflect the loss function in your situation at hand.

(c) [E.C.] Prove that the MLE of the $\beta$'s is the same solution as minimizing SSE.

(d) [E.C.] Assume that you have proven the above and plugged in those estimates to the likelihood expression. Now $\sum_{i=1}^{n} \mathcal{E}_i^2 = \sum_{i=1}^{n} e_i^2 = SSE$. Prove that $\hat{\sigma^2}_{MLE} = MSE$.

## Problem 7

We will now analyze the baseball data (`baseball.csv`). You can use any software package you wish to answer these questions.

(a) [easy] Fit a linear model with reponse variable $y$ : salary in thousands. Use all available predictors. Provide a valid interpretation on $\hat{\beta}_j$ for the feature "number of RBI's".

(b) [easy] Is this interpretation reasonable given what you know about number of RBI's and how it is related to other predictors? You may need to ask someone who knows a bit about baseball.

(c) [easy] What a causal additive model for number of RBI's make sense? Yes / no.

(d) [easy] Would you be able to make a randomized experiment to find the additive causal effect of number of RBI's? Yes / no.

(e) [easy] Some of these variables may be significant because we dredged. Why is this likely *not* the case?

(f) [E.C.] Use a likelihood ratio test to test the effect of number of RBI's and Number of Walks.