Correlation, regression, and the Kelly strategy

18.600 Problem Set 8, due April 28

Welcome to your eighth 18.600 problem set! Let's think about correlations. Given a population of people who each have two attributes, like ACT and SAT scores, you can choose a member at random and interpret the attributes of the person you choose as random variables. These variables have a correlation coefficient, which you'd expect to be high for ACT and SAT scores (maybe about .87, per some site I googled).

Correlations inform beliefs. The strong observed correlations between cigarettes and early death (and many specific ailments like lung cancer) are a huge part of the of evidence that cigarettes are unhealthy. The discovery of the unhealthiness of smoking has saved millions of lives — a win for observational statistics. (Also an embarrassment, given that it took until the second half of 20th century for the case to be made persuasively.)

On the other hand, we all know the *correlation does not imply causation* cliché. The "spurious correlation" website tylervigen.com illustrates this with strong correlations between seemingly unrelated annual statistics like sociology doctorates and space launches, or pool drownings and Nicolas Cage films. The 2012 NEJM article *Chocolate consumption, Cognitive function, and Nobel Laureates* (which presents a real country-to-country correlation between chocolate consumption and Nobel prize winning) is a clever parody of the way observed correlations are used in medicine (it's only three pages; look it up). It earnestly walks the reader through causal hypotheses (brain-boosting flavanoids), reverse causal hypotheses (chocolate consumed at Nobel prize celebrations) and common demoninator hypotheses (geography, climate, socioeconomics), mostly dismissing the latter.

Alongside *correlation does not imply causation* one might add *correlation does not imply correlation*, or more precisely, reported correlation in some data set does not imply correlation in the larger population, or correlation that will persist in time. Google *study "linked to"* and scroll through a few pages of hits. Some sound fairly plausible ("walking/cycling to work linked to lower body fat") but others raise eyebrows. Clicking through, you find that sometimes the correlations are weak, the sample sizes small, the stories (at least at first glance) far fetched. A news organization's criteria for deciding which links to publicize may differ from those a careful scientist would use to decide what to take seriously and/or study further (e.g. with randomized trials). Reader beware.

This problem set will also feature discussion about the Kelly strategy, moment generating functions, regression lines, and a phenomenon called regression to the mean.

Please stop by my weekly office hours (2-249, Wednesday 3 to 5) for discussion.

A. FROM TEXTBOOK CHAPTER SEVEN:

1. Problem 51: The joint density of X and Y is given by $f(x,y) = \frac{e^{-y}}{y}$, $0 < x < y$, $0 < y < \infty$. Compute $E[X^3|Y=y]$.

Remark: To prepare for the next problem, suppose that you discover a market inefficiency in the form of a mispriced asset. Precisely, you discover an asset priced at \$10 that has a $p > 1/2$ chance to go up to \$11 over the next day or so (before reaching \$9) and a $(1-p) < 1/2$ chance to go down to \$9 (before reaching \$11). By buying r shares at \$10 and them selling when the price reaches \$9 or \$11, you have an opportunity to make a bet that will win r dollars with probability $p > 1/2$ and lose r dollars with probability $(1-p)$. Let's ignore transaction costs and bid-ask spread. (And assume that, unlike all those people who merely *think* they can recognize market inefficiencies, you *actually can.* Assume also that your wisdom was obtained legally — so no risk of an insider trading conviction!) So now you effectively have an opportunity to bet r dollars on a p coin with $p > 1/2$. The question is this: how much should you bet? In expectation you will make $pr + (1-p)(-r) = (2p-1)r$ dollars off this bet, so to maximize your expected payoff, you should bet *as much as you possibly can.* But is that really wise? If you repeatedly bet all our money on p-coins, it might not be long before you lose everything. The *Kelly strategy* (which comes from assuming utility is a logarithmic function of wealth — look it up) states that instead of betting everything, you should bet a $2p-1$ fraction of your current fortune. The next problem is a simple question about this strategy.

2. Problem 67: Consider a gambler who, at each gamble, either wins or loses her bet with respective probabilities p and $1-p$. A popular gambling system knkown as the Kelly strategy is to always bet the fraction $2p-1$ of your current fortune when $p > 1/2$. Compute the expected fortune after n gambles of a gambler who starts with x units and employs the Kelly strategy.

Remark: The next problem will help solidify your understanding of moment generating functions. These play a central role in *large deviation theory*, which in turn plays a central role in information theory, data compression, and statistical physics. In this course, we mostly use moment generating functions (and the closely related *characteristic functions*) as tools for proving the central limit theorem and the weak law of large numbers.

3. Theoretical Exercise 48: If $Y = aX + b$, where a and b are constants, express the moment generating function of Y in terms of the moment generating function of X.

B. Suppose that X and Y both have mean zero and variance one, so that $E[X^2] = E[Y^2] = 1$ and $E[X] = E[Y] = 0$.

(a) Check that the correlation coefficient between X and Y is $\rho = E[XY]$.

(b) Let r be the value of the real number a for which $E[(Y - aX)^2]$ is minimal. Show that r depends only on ρ and determine r as a function of ρ.

(c) Check that whenever Z has finite variance and finite expectation, the real number b that minimizes the quantity $E[(Z - b)^2]$ is $b = E[Z]$.

(d) Conclude that the quantity $E[(Y - aX - b)^2]$ is minimized when $a = r$ and $b = 0$.

Remark: We have shown that among all affine functions of X (i.e. all sums of the form $aX + b$ for real a and b) the one that best "approximates" Y in (in terms of minimizing expected square difference) is rX. This function is commonly called the *least squares regression line* for approximating Y (which we call a "dependent variable") as a function of X (the "independent variable"). If $r = .1$, it may seem odd that $.1X$ is considered an approximation for Y (when Y is the dependent variable) while $.1Y$ is considered an approximation for X (when X is the dependent variable). The lines $y = .1x$ and $x = .1y$ are pretty different after all. But the lines are defined in different ways, and when $|r|$ is small, the correlation is small, so that neither line is an especially *close* approximation. If $r = 1$ then $\rho = 1$ and both lines are the same (since X and Y are equal with probability one in this case).

Remark: The above analysis can be easily generalized (by simply rescaling and translating the (X, Y) plane) to the case that X and Y have non-zero mean and variances other than one. The subject known as *regression analysis* encompasses this generalization along with further generalizations involving multiple dependent variables, as well as settings where a larger collection of functions plays the role that affine functions played for us. Regressions are ubiquitous in academic disciplines that use data. Given data in a spreadsheet, you can compute and plot regression lines with the push of a button (google *spreadsheet regression*; follow instructions; or type "linear fit $\{1,3\}\{2,4\}\{4,5\}\{3,5\}$" into wolframalpha to see an example). For a more difficult setting, imagine that I give you a set of pictures, and assign a number to each picture indicating how closely it resembles a cat. If you had a nice way to approximate this function (from the set of digitally encoded pictures to the set of numbers) you could program it into your computer and enable your computer to recognize cats. Statistics and "machine learning" are hot topics, and may be part of your further coursework at MIT. (Note: even if your cat-recognizing algorithm is a sophisticated and/or incomprehensible neural net created by hundreds of tinkering MIT alumni, you will still use the math behind the "clinical trial" stories in this course when you try to *test* your algorithm's effectivness.)

C. On Smoker Planet, each person decides at age 18, according to a fair coin toss, whether or not to become a life long cigarette smoker. A person who does not become a smoker will never smoke at all and will die at a random age, the expectation of which is 75 years, with a standard deviation of 10 years. If a person becomes a smoker, that person will smoke exactly 20 cigarettes per day throughout life, and the expected age at death will be 65 years, with a standard deviation of 10 years.

(a) On this planet, let $S \in \{0, 20\}$ be cigarettes smoked daily, and let L be life duration. What is the correlation $\rho(S, L) := \mathrm{Cov}(S, L)/\sqrt{\mathrm{Var}(S)\mathrm{Var}(L)}$? Hint: start by using the $\mathrm{Var}(L) = \mathrm{Var}(E[L|S]) + E[\mathrm{Var}(L|S)]$ identity from lecture to get $\mathrm{Var}(L)$. Working out $\mathrm{Var}(S)$ shouldn't be hard. Then attack the two terms of $\mathrm{Cov}(S, L) = E[SL] - E[S]E[L]$. Note that $E[SL] = P\{S = 0\}E[SL|S = 0] + P\{S = 20\}E[SL|S = 20]$.

On Bad Celery Planet, it turns out that (through some poorly understood mechanism) celery is unhealthy. In fact, a single piecce of celery is as unhealthy as a single cigarette on Smoker

Planet. However, nobody eats 20 pieces a day for a lifetime. Everybody has a little bit, in varying amounts throughout life. Here is how that works. Each year between age 18 and age 58 a person tosses a fair coin to decide whether to be a celery eater that year. If the coin comes up heads, that person will eat, on average, one piece of celery per day for the entire year (mostly from company vegetable platters). *Given* that one consumes celery for K of the possible 40 years (celery consumption after age 58 has no effect, and everyone lives to be at least 58) one expects to live until age $75 - K/80$, with a standard deviation of 10 years. (So, indeed, eating 1 celery stick per day for the full 40 years is about 1/20 as harmful as smoking 20 cigarettes a day for a lifetime.)

(b) Write L for a person's life duration. On this planet, what is the correlation $\rho[K, L] = \mathrm{Cov}[K, L]/\sqrt{\mathrm{Var}(K)\mathrm{Var}(L)}$? Hint: start by using the $\mathrm{Var}(L) = \mathrm{Var}(E[L|K]) + E[\mathrm{Var}(L|K)]$ identity from lecture to get $\mathrm{Var}(L)$. Working out $\mathrm{Var}(K)$ shouldn't be hard. Then attack $\mathrm{Cov}(K, L) = E[KL] - E[K]E[L]$ as in (a).

Remark: The answer to (b) is really a lot smaller than the answer to (a). So small that it would be *very* hard to demonstrate this effect to the public without a really big sample size. Moreover, even if you exhibited the effect in a large sample, people might reason as follows: if those who eat more celery are statistically different from those who eat less (more health conscious, more concentrated in certain regions, more prone to also eat carrots and ranch dressing, etc.) the effects of these differences could *easily* swamp any effects of the celery itself. One can try to "control" for obvious differences in some way (e.g., with multi-variable regressions) but one cannot account for *all* of them. So even after you convincingly demonstrate correlation, it may appear (to the educated public of the planet) *very* unlikely that the health effects of celery are the actual cause. Unless celery is linked more directly to a specific ailment (e.g., death by choking on celery) or shown to be unhealthy in other ways (chemical properties, experiments on rats, etc.) the people on Bad Celery Planet may never develop a persuasive case against celery. Of course, in the real world, people worry that *lots* of products have mild carcinogenic effects: effects in the ominous "big enough to matter, small enough to be hard to demonstrate" category. Some hope that "big data" will improve our ability to detect and understand weak correlations. Otherwise, we can hope that a better basic scientific understanding of mechanisms by which harms arise might help us avoid "likely to be mildly harmful" substances whose harms are too small to observe directly.

Remark: Even when we observe large effects, it is hard to know what to make of them. Per CDC `https://www.cdc.gov/nchs/data/dvs/mortfinal2007_worktable23r.pdf` younger people in Florida died at much higher rates than younger people in Massachusetts in 2007, but people over 85 in Massachusetts died at much higher rates than people over 85 in Florida. But before one retires to Florida for the sake of longevity, one might ask: are healthier older people more prone to live in Florida (beaches and golf) and less healthy older people more prone to live in Massachusetts (family and famous hospitals)? What other population differences play a role? How persistent is the effect? If you find the answer before I'm 85, please let me know.

D. Let X be a normal random variable with mean μ and variance σ_1^2. Let Y be an independent normal random variable with mean 0 and variance σ_2^2. Write $Z = X + Y$. Let $\tilde{Y}$ be an independent random variable with the same law as Y. Write $\tilde{Z} = X + \tilde{Y}$.

(a) Compute the correlation coefficient ρ of $\tilde{Z}$ and Z.

(b) Compute $E[X|Z]$ and $E[\tilde{Z}|Z]$. Express the answer in a simple form involving ρ. Hint: consider case $\mu = 0$ first and find $f_{X,Z}(x,z)$. You know $F_X(x)$ and $f_{Z|X=x}(z)$. Alternate hint: if X_i are i.i.d. normal with variance σ^2, mean 0, and $n \geq k$ then argue by symmetry that $E[\sum_{i=1}^k X_i | \sum_{i=1}^n X_i = z] = z(k/n)$. Write $X = \sum_{i=1}^k X_i$ and $Y = \sum_{k+1}^n X_i$. Fiddle with k, n, σ^2 to handle the case that σ_1^2/σ_2^2 is rational.

Note that $E[\tilde{Z}|Z]$ is closer to $E[\tilde{Z}] = E[Z]$ than Z is. This is a case of what is called "regression to the mean." Let's tell a few stories about that. An entrant to a free throw shooting competition has a *skill level* that we denote by X, which is randomly distributed as a normal random variable with mean μ and variance 2. During the actual competition, there is an independent *luck factor* that we denote by Y, which is a normal random variable with variance 1 and mean zero. The entrant's overall score is a $Z = X + Y$. If the entrant participates in a second tournament, the new score will be $\tilde{Z} = X + \tilde{Y}$ where $\tilde{Y}$ is an independent luck factor with the same law as Y.

(c) Compute the standard deviation of Z. Given that Z is two standard deviations above its expectation, how many standard deviations above its expectation do we expect $\tilde{Z}$ to be?

Imagine that people in some large group are randomly assigned to teams of 16 people each. Each person's *skill level* is an i.i.d. Gaussian with mean 0 and standard deviation 1. The team's skill level is the sum of the individual skill levels. You can check that a team's skill level is a Gaussian random variable with mean 0 and standard deviation 4.

(d) Given that a team's total skill level is 8 (two standard deviations above the mean for teams) what do we expect the skill level of a randomly chosen team member to be?

Each drug generated by a lab has an "true effectiveness" which is a normal random variable X with variance 1 and expectation 0. In a statistical trial, there is an independent "due-to-luck effectivness" normal random variable Y with variance 1 and expectation 0, and the "observed effectiveness" is $Z = X + Y$.

(e) If we are *given* that the observed effectiveness is 2, what would we expect the observed effectiveness to be in a second independent study of the same drug?

Remark: The following is from the abstract of the Nosek reproducibility study (recall Problem Set 3) which attempted to reproduce 100 published psychology experiments: "The mean effect size (r) of the replication effects (Mr = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (Mr = 0.403, SD = 0.188), representing a substantial decline."

The fact that the effect sizes in the attempted replications were smaller those in the original studies is not surprising from a *regression to the mean* point of view. Google *Iorns reproducibility* for analogous work on cancer.

E. This problem will apply the "regression to the mean" ideas from the previous problem to a toy model for university admissions. Let's think about admissions at a (somewhat arbitrarily chosen) group of five selective universities: Harvard, Stanford, MIT, Yale and Princeton. For fall 2014, these universities all had (per usnews.com article I looked up) "yield rates" between 66 and 81 percent and class sizes between 1043 and 1654. If we refer to an admission letter to one of these five universities as a *golden ticket* then in all 9570 golden tickets were issued and 7047 were used (i.e., a total of 7047 first year students enrolled at these schools). This means there were 2523 *unused* golden tickets.

Who *had* these 2523 unused golden tickets? Somebody presumbly has a rough answer, but let's just speculate. One wild possibility is that *all* the unused tickets were held by 505 very lucky students (with 5 unused golden tickets each) who *all* crossed the Atlantic to attend Oxford and Cambridge. If this were true, then *none* of the 7047 golden ticket *users* would have an extra unused ticket. Another wild possibility is that exactly 2523 of the 7047 golden ticket *users* (about 35.8 percent) have exactly one unused ticket. In any case, the fraction of golden ticket users with an *unused ticket to spare* is somewhere between 0 and .358, which implies that the *overwhelming majority* of these 7047 entering students were accepted to the university they attend and to *none* of the other four. The stereotypical "students who apply to all five, get accepted to most" are a small minority. We would need more data to say more than that (where individuals apply, how many apply early admission, how students decide between multiple offers, how admissions criteria vary from place to place, etc.) So let's move to an imaginary (and perhaps not terribly similar) universe where the analysis is simpler. Then we'll do a little math.

In Fancy College Country there are exactly five elite universities and 40,000 elite applicants. All 40,000 applicants apply to all five universities. The *intrinsic strength* of an applicant's case is a normal random variable X with mean 0 and variance 1. When a university reads the application, the university assigns it a *score* $S = X + Y$ where Y is an independent normal random variable with mean 0 and variance 1. Think of X as the college-independent part of an application's strength and Y as the college-dependent part (perhaps reflecting the resonance of the student's background with university-specific goals, as well as the random mood of the admission team). Each student has one value X but gets an independent Y value for each university. Each university admits all applicants with scores above the 95th percentile in score distribution. Since S has variance 2, this means they admit students whose scores exceed $C = \Phi^{-1}(.95) \cdot \sqrt{2} \approx 1.6449 \cdot \sqrt{2} \approx 2.326$ where $\Phi(a) = (2\pi)^{-1/2} \int_{-\infty}^{a} e^{-x^2/2} dx$. To be admitted a student's score must exceed 2.326. Each university expects to admit 5 percent of its applicants.

(a) Compute, as a function of x, the conditional probability that the student is admitted to the first university in the list, *given* that the student's X value is x. In other words, compute the probability that $Y > C - x$.

(b) Let $A(x)$ be the conditional probability that the student is admitted to *at least one* university on the list, given that the student's X value is x. Compute $A(x)$ using Φ and C as defined above.

(c) Argue that the *overall* probability that a student is admitted to at least one university is given by $\int_{-\infty}^{\infty}(1/\sqrt{2\pi})e^{-x^2/2}A(x)dx$ and that the chance to be rejected by all universities is $\int_{-\infty}^{\infty}(1/\sqrt{2\pi})e^{-x^2/2}(1-A(x))dx$.

(d) Try to compute (c) numerically in a package like wolframalpha and report how it goes. You might (I did) have to fiddle a bit to get it to work. Here's how I did it:

1. To see how wolframalpha represents $\Phi(x)$ type in

   ```
   Integrate[(1/Sqrt[2Pi]) E^(-y^2/2), {y,-Infinity, x}]
   ```

   You get some expression involving erf, which is a close relative of Φ.
2. Click on that to get plaintext. Replace x with $(2.326 - x)$ to get

   ```
   1/2 (1+erf((2.326-x)/sqrt(2)))
   ```

3. Put parentheses about this and raise it to fifth power (to get a wolframalpha friendly expression for conditional chance to be rejected everywhere, given x), multiply by $f_X(x)$ and integrate:

   ```
   Integrate[(1/sqrt(2Pi)) E^(-x^2/2)

   (1/2 (1+erf((2.326-x)/sqrt(2))))^5, {x,-Infinity, Infinity}]
   ```

(e) Briefly justify the following conclusions. Each student has a 0.166363 chance to be accepted at least somewhere. The expected number of students admitted to at least one university is about 6655. The expected class sizes are about 1331 at each school, and each university has a typical yield rate of about .67.

Remark: If admission were completely random (each university takes a student with probability .05 independently of X) then the applicants would have a $1 - (.95)^5 \approx .2262$ chance to get accepted to at least one university. We'd expect to see $.2262 \cdot 40000 \approx 9048$ students admited to at least one university, and the yield rate for each university would be roughly .9048. If the selection process were completely determined by X (so that all universities accept exactly the *same* 2000 students) then there would be only 2000 students admitted to at least one university and the yield rate would be .2 (with class sizes of only 400). Our .67 lies between these extremes.

Remark: If some applicants applied to fewer than 5 universities (e.g., due to early admissions) yield rates might be *higher*, since fewer admits would have multiple offers. If the variance of Y were smaller, yields might be *lower* due to greater admission list overlap. If some admits chose *not* to attend one of the 5 schools, that would also decrease yields.