Correlation, regression, and paradox

18.600 Problem Set 8, due November 9

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).

How high is a .87 correlation really? Imagine X, Y_1 and Y_2 are independent standard normals. Imagine that X is your "raw test-taking acumen" and $X + aY_1$ is your score on one test and $X + aY_2$ is your score on another test, where the aY_i are random noise terms. Then $\rho(X + aY_1, X + aY_2) = 1/(1 + a^2)$, which is about .86 if $a = .4$. In this case, the standard deviation of the "noise factor" is .4 times the standard deviation of the "raw ability factor," which may seem like quite a bit of noise (despite the fact that .86 seems pretty close to 1). Play the game `http://guessthecorrelation.com/` to get a sense of what different correlation levels look like.

Correlations inform beliefs. The strong observed correlations between cigarettes and early death (and 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, since it took until the second half of 20th century to make the case persuasively.)

On the other hand, we know the *correlation does not imply causation* cliché. The "spurious correlation" website `http://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) parodies 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, economics), 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 "is 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 moment generating functions, regression lines (which you will encounter often in life) and a phenomenon called regression to the mean.

On a rather different note, many of you are familiar with *Pascal's wager*. The idea is that if choosing A over B comes with a finite cost but a positive probability (however small) of an infinite payoff, then one should always choose A. Pascal's conclusion was that if living a virtuous life leads (with even a tiny probability) to an eternal reward, then it is a worthwhile sacrifice to make. A common criticism is that this kind of thinking can lead to violence (killing heretics who *might* lead souls astray, or dissidents who *might* obstruct an endless Marxist utopia) as well as virtue. A more mathematical concern is that in principle there may be many choices, each of which we expect to do an infinite amount of good (and perhaps also an infinite amount of harm) and there is no obvious mathematical way to compare the competing infinities.

The comparison difficulties associated with infinite expectations can arise even when the payoffs themselves are finite with probability one (e.g., if the utility payout is a Cauchy random variable). This problem set illustrates this point with a particularly vexing form of a famous envelope switching paradox. Interestingly, in this paradox, the conditional expectations used for decision making are all finite; but a certain *a priori* expectation is infinite, and that is the root of the paradox. I hope that you enjoy thinking about the story, and that it causes you at most a finite amount of existential angst.

A. **TEXTBOOK CHAPTER SEVEN:**

1. Problem 50: The joint density of X and Y is given by $f(x,y) = \frac{e^{-x/y}e^{-y}}{y}$, $0 < x < \infty$, $0 < y < \infty$. Compute $E[X^2|Y = y]$.

2. Theoretical Exercise 29: Let $X_1, \ldots, X_n$ be independent and identically distriuted random variables. Find

$$E[X_1|X_1 + \ldots + X_n = x].$$

Remark: If X is a random variable then the function $M_X(t) := E[e^{tX}]$ is called the moment generating function of X. Moment generating functions play a central role in *large deviation theory*, which plays a central role in information theory, data compression, and statistical physics. In this course, we 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. **LEAST SQUARES REGRESSION:** Suppose that X and Y both have mean zero and variance one, so that $E[X^2] = E[Y^2] = 1$ and $E[X] = E[Y] = 0$.

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

2. 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.

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

4. 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 is easily generalized (by 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 (copy a chart into a free spreadsheet like sheets.google.com; control click two distinct columns to highlight them; click the chart icon; then Customize, Series, Trendline... or just google *spreadsheet regression* for instructions; or type "linear fit $\{1,3\}\{2,4\}\{4,5\}\{3,5\}$" into wolframalpha for an example). For a more difficult setting, imagine you have a set of pictures, and a number indicating how closely each picture resembles a cat. If you had a nice way to approximate this function (from pictures to numbers) you could train your computer to recognize cats. Your procedure would likely be more complicated than a simple regression — it may involve *neural nets* or other *machine learning* tools. Statistics

and machine learning are hot topics, and may be part of your further coursework. (Note: even if your cat recognizing algorithm is a complex neural net refined by hundreds of tinkering MIT alumni, you will still use the math behind the "clinical trial" stories in this course when you *test* its effectivness.)

C. **REGRESSION TO MEAN:** 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 $\tilde{Z} = X + \tilde{Y}$.

1. Compute the correlation coefficient ρ of $\tilde{Z}$ and Z.

2. 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.

3. 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 9 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 3.

4. Given that a team's total skill level is 6 (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 $Z = X + Y$.

5. 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: This is from the abstract of the Nosek reproducibility study (recall Problem Set 3) which tried 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 is not surprising from a *regression to the mean* point of view. Google *Iorns reproducibility* for analogous work on cancer studies.

D. **CELERY:** On Smoker Planet, each person decides at age 18, by 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.

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

2. Write L for a person's life duration. On this planet, what is the correlation $\rho[K, L] = \text{Cov}[K, L]/\sqrt{\text{Var}(K)\text{Var}(L)}$? **Hint:** Use the $\text{Var}(L) = \text{Var}(E[L|K]) + E[\text{Var}(L|K)]$ identity from lecture to get $\text{Var}(L)$. Working out $\text{Var}(K)$ shouldn't be hard. Then attack $\text{Cov}(K, L) = E[KL] - E[K]E[L]$ as in 1. Note that $E[KL] = \sum_{k=0}^{40} P\{K = k\}E[KL|K = k]$ which is $\sum_{k=0}^{40} P\{K = k\}kE[L|K = k] = \sum_{k=0}^{40} P\{K = k\}k(75 - k/80) = E[K(75 - K/80)]$. Maybe you can argue that $E[L] = E[75 - K/80]$ and that $\text{Cov}(K, L) = \text{Cov}(K, 75 - K/80)$.

Remark: The answer to 2 is much smaller than the answer to 1. So small that it would be *very* hard to demonstrate this effect without a huge sample size. You would need several hundred thousand to be confident that you would see a statistically significant correlation. In the real world, people worry that *many* products have mild carcinogenic effects (effects in the ominous "big enough to matter, small enough to be hard to observe") category. Detectability is a big problem. Moreover, even if you observe the effect in a large sample, people will note that those who eat more celery are statistically different from those who eat less (more health conscious, more prone to eat carrots and ranch dressing, etc.) The effects of these differences could *easily* swamp any effects of the celery itself. One try to "control" for obvious differences (e.g., with multi-variable regressions) but one cannot account for *all* of them, and the question of what to *do* about observed correlation is famously hard. For example, the World Health Organization website says the following about red meat: "Eating red meat has not yet been established as a cause of cancer. However, if the reported associations were proven to be causal, the Global Burden of Disease Project has estimated that diets high in red meat could be responsible for 50,000 cancer deaths per year worldwide. These numbers contrast with about 1 million cancer deaths per year globally due to tobacco smoking, 600,000 per year due to alcohol consumption, and more than 200,000 per year due to air pollution." Shall I eat that burger or not?

E. **ADMISSIONS:** This problem will apply the "regression to the mean" ideas from the last problem to a toy model for university admissions. Think about admissions at a (somewhat arbitrarily chosen) group of five selective universities: Harvard, Stanford, Chicago, MIT and Princeton. For the most recent class, these universities all had (per collegeevaluator.com) "yield rates" between 69 and 85 percent and class sizes between 1148 and 1894. If we refer to an admission letter to one of these five universities as a *golden ticket* then in all 9600 golden tickets were issued and 7736 were used (i.e., a total of 7736 first year students enrolled at these schools). This means there were 1864 *unused* golden tickets.

Who *had* these 1864 unused golden tickets? Somebody presumably has a rough answer, but let's just speculate. One wild possibility is that *all* unused tickets were held by 373 lucky students (with 5 unused golden tickets each) who *all* chose to attend state schools instead. If this were true, then *none* of the 7736 golden ticket *users* would have an extra unused ticket. Another extreme possibility is that exactly 1864 of the 7736 golden ticket *users* (about 24 percent) have exactly one unused ticket. In any case, the fraction of golden ticket users with an

unused ticket to spare is between 0 and .24, which implies that the *overwhelming majority* of these 7736 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 tiny minority. What if I throw in Penn, Yale, Brown, Columbia and Cornell? In that case number of unused tickets is about .38 times the number of used tickets. Again the *vast majority* of the students accepted to *at least one* of these ten schools were accepted to *only one*. These even works for an appropriately chosen set of 20 schools. Why is it so hard to get into more than one high yield school? Do colleges use signals (geography, early admissions, legacy, apparent fit, etc.) to identify and accept students who "would attend if accepted" (see pset 3)? Or it is possible that universities mostly value the same thing but admissions are just subjective? Let us explore the latter possibility in 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.

1. 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 $Y > C - x$.

2. How large does X have to be for this conditional probability to exceed .05? How about .95? Find the probability that X exceeds the former threshold. And the latter. (Give numerical answers. Note the discrepancy: given their X values, *many* students have a .05 chance, but *very few* have a .95 chance. It is easier to be a contender than a sure thing.)

3. 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.

4. 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 $\int_{-\infty}^{\infty} (1/\sqrt{2\pi}) e^{-x^2/2} (1 - A(x)) dx$.

5. Try to compute 4 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:

   (a) 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.

   (b) Click on that to get plaintext. Replace x with $(2.326 - x)$ to get

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

   (c) 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}]
```

6. 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: Suppose the score were $X + aY$ for some $a > 0$. Small a means universities mainly use the same objective data (X). Large a means they mainly use subjective/idiosyncratic criteria (Y). Can admissions debates (e.g., how much to use test scores vs. school-specific essays, etc.) be framed as questions about how large a should be? Large a may have advantages (high yield rates, people with range of X values get to know each other) and disadvantages (admissions unpredictable, students have to apply many places) even aside from the question of how X and Y correlate with other things we value. Real world admissions are more complex than the toy model in this problem. Should we switch to a matching system like medical residency? `https://www.nrmp.org/matching-algorithm/`.

F. **ENVELOPES:** The following is one formulation of a famous "two envelope" paradox. Jill is a money-loving individual who, given two options, invariably chooses the one that gives her the most money in expectation. One day Harry, a trusted (and capable of delivering) individual, offers her the following deal as a gift. He will secretely toss a fair coin until the first time that it comes up tails. If there are n heads before the first tails, he will place 10^n dollars in one envelope and 10^{n+1} dollars in the second envelope. (Thus, the probability that one envelope has 10^n dollars and the other has 10^{n+1} dollars is 2^{-n-1} for $n \geq 0$.) Harry will then hand Jill the pair of envelopes (randomly ordered, indistinguishable from the outside) and invite her to choose one. After Jill chooses an envelope she will be allowed to open it. Once she does, she will be allowed to either keep the money in the first envelope or switch to the second envelope and keep whatever is in the second envelope. However, if she decides to switch, she has to pay a one dollar "switching fee."

1. If Jill finds 100 dollars in the first envelope she opens, what is the conditional probability that the other envelope contains 1000 dollars? What is the conditional probability that the other envelope contains 10 dollars?

2. If Jill finds 100 dollars in the first envelope she opens, how much money does Jill expect to win from the game if she does not switch envelopes? (Answer: 100 dollars.) How much does she expect to win (net, after the switching fee) if she *does* switch envelopes?

3. Generalize the answers above to the case that the first envelope has 10^n dollars (for $n \geq 0$) instead of 100.

4. Jill concludes from the above that, no matter what she finds in the first envelope, she will expect to earn more money if she switches envelopes and pays the one dollar switching fee. This strikes Jill as a bit odd. If she knows she will always switch envelopes, why doesn't she just take the second envelope first and avoid the envelope switching fee? How can she be maximizing her expected wealth if she spends an unnecessary "switching fee" dollar no matter what? How does one resolve this apparent paradox? (Use the hints in the Lecture 24 slides as needed. Even with hints, it may take time to make peace with this.)