Ian Deters
Phone: 419-419-8880
E-mail: iandeters@iandeters.com
Website: www.iandeters.com

Research Statement

I am persuaded that all academic inquiry is not for its own sake, but rather to inform the actions of the inquisitor. Since mere precision is at times insufficient to inform action, I approach Data / Actuarial Science as one not only concerned with the precision of any particular model, but also its explicability. My professional goal is to create tools to swiftly build models which are simultaneously precise and explicable.

While my formal academic background is in theoretical mathematics, I have spent many years as a practitioner of Data and Actuarial Science and possess an ACAS credential from the Casualty Actuarial Society. Throughout this time I have become convinced that there is much that a rigorous mathematical approach can bring to both of these subject areas. Before proceeding, it should be mentioned that I view both Data and Actuarial Science as belonging to the subject matter of Statistics. I write this since both deal principally with the estimation of quantities related to random variables. This will be discussed in greater detail below.

Data / Actuarial Science is the intersection of Mathematics, Computer Science, and Technology. An important observation is that with any data set (e.g. $\{X_1, \dots, X_n\}$), one may perform an infinite number of computations $\left(\text{e.g. } \frac{1}{n}\sum_{k=1}^{n} X_k \text{ or } e^{\cos\left(\prod_{k=1}^{n} X_k\right)} + 54\right)$ with the data. The important question is which of these computations are useful. It should be observed that this question is almost never examined in some of the classic papers (e.g. [6], [8], [9]). These papers propose various schemes, explicate sundry computations, and examine their implications on different data sets. However, there is no discussion about why one should prefer their proposed computations to any other sequence of computation. This is precisely the question that mathematics answers. Mathematics helps to narrow the infinite field of computational possibility to sequences of meaningful and useful computations. Computer Science provides various algorithms to compute the desired quantities. Technology gives you the ability to deploy the computations.

In order to motivate my research plan, consider that almost every data scientific problem can loosely be stated as using known data to estimate some statistical object of interest. Examples of these objects include the expected value of a distribution, the cumulative distribution function of some distribution, the conditional expectation of some variable given some other set of variables, or the joint distribution function of a set of random variables. Observe that the inclusion of the word "estimate" implies there exists a notion of closeness as well as a way to select among competing estimates.

For an example, consider that in business one often seeks to estimate the average value of some quantity, (e.g. customer tenure), given the values of other variables (e.g. territory of the customer, how much business the customer has with you, etc.). In mathematical terms, one wishes to estimate the conditional expectation of tenure given other information known about the customer. If T represents customer tenure and Σ is the σ-algebra generated by all potential predictor variables, then, in typical notation, one wishes to estimate $\mathrm{E}[T|\Sigma]$. How shall one distinguish among alternatives? Assuming that $\mathrm{E}T^2 < \infty$, then it is known that $\mathrm{E}[T|\Sigma]$ minimizes $\mathrm{E}(T-f)^2$ for all square integrable, Σ - measurable functions f ([5], p. 225). Of course, one may estimate $\mathrm{E}(T-f)^2$, assuming one's observations are independent and identically distributed, for any f by computing $\frac{1}{n}\sum_{k=1}^{n}(T_k - f_k)^2$ where T_k and f_k are the kth observations of T and evaluation of f respectively.

I will now generalize the above example an enumerate my areas of inquiry.

1. As observed above, the output of any analysis, whether seeking an expected value, cumulative distribution function, joint probability density function, or conditional expectation is a function. As such, the first area of inquiry will be to specify the topology under which estimation is considered. For instance, random variables can converge uniformly, pointwise, almost surely, in probability, and in distribution. Care must be taken since some notions of convergence, such as almost sure convergence (4), do not arise from a topology. For reasons made clear subsequently, special attention will be paid to topologies which are metrizable and separable.

2. Suppose that X is the set of functions under consideration. The set X is said to be *parameterizable* if there are sequences $(d_n) \in \mathbb{N}$, $E_n \subseteq \mathbb{R}^{d_n}$, and (f_n) such that $f_n \in X^{E_n}$, f_n is continuous for $n \in \mathbb{N}$, and $\overline{\cup_{n=1}^{\infty} f(E_n)} = X$. It should be noted that if X is parameterizable, then it is separable (1). This is the reason for the earlier remark that separable topologies on X will receive special interest. If X is parameterizable, it may be assumed without loss of generality that $f_n(E_n) \subseteq f_{n+1}(E_{n+1})$ for $n \in \mathbb{N}$ (2).

3. Suppose that T is the function of interest. If X is metrizable with metric d and parameterizable, define $s_n : E_n \to \mathbb{R}$ by $s_n(x) = d(T, f_n(x))$. Since X is parameterizable, $d(T, f_n(E_n)) \to 0$. Hence, there is an $x_n \in E_n$ such that $d(T, f_n(x_n)) < d(T, f_n(E_n)) + \frac{1}{n}$. In the absence of the construction of a minimizer of s_n, the minimization of s_n may be attempted using various minimization routines. This suggests two tasks. First, if s_n is differentiable, expressions for the first and second derivatives, where they exist, of s_n may be computed. This will aid the computational efficiency of algorithms which rely upon derivatives. Second, for a given algorithm, it may be the case that a particular method, for example (3), for generating an initial value may be demonstrated to always lead to a sequence which converges to the function T. If not, at least the adequacy of different methods for generating initial values for different minimization routines may be examined empirically.

With this framework in place, the categories and descriptions of projects may now be stated.

1. Data Scientific

   (a) Determine what topologies on the set of $\sigma(X_1, \ldots, X_n)$ measurable functions, or some appropriate subset, are separable and metrizable. For instance, if $\mathrm{E}T^2 < \infty$, then $\mathrm{E}[T|X_1, \ldots, X_n] \in L^2$. Hence, the L^2 norm would yield an appropriate metric. Moreover, it is known when L^p, where $p \in (1, \infty)$, is separable in terms of the underlying measure space ([2], p. 896).

(b) Determine, for each topology, if it is parameterizable and find associated sequences (e.g. $(d_n) \in \mathbb{N}$, $E_n \subseteq \mathbb{R}^{d_n}$, and (f_n)) demonstrating it. While not a formal demonstration I think, as I have previously argued ([4]), that it is essentially sufficient to consider sets of functions dense in the space of functions continuous on the hypercube (i.e. $C([0,1]^n)$) in order to approximate $\mathrm{E}[Y|X_1, \ldots X_n]$ to any degree of accuracy. This is, for example, why neural networks are useful in Data Science. Cybenko proved in 1989 ([7]) that single layer perceptrons are dense in $C([0,1]^n)$. This task is important since each set of sequences which shows that space is parameterizable will yield a different functional form with which to approximate the conditional expectation. The more functional forms available will increase the probability that at least one of those forms will be of use to the stakeholder for interpretation. For instance, if a stakeholder finds a neural network inscrutable, perhaps they will find a stepwise function more understandable.

(c) Determine, for each set of sequences demonstrating the parameterizability of the space, the gradients and Hessians of the sequence of the aforementioned error functions (s_n).

(d) Determine, for each sequence of aforementioned error functions,

   i. a minimizer for s_n,

   ii. an initial value for a minimization routine for minimizing s_n which will ensure convergence to a global minimum, or

   iii. the empirical adequacy of different methods of generating initial values for a minimization routine for minimizing s_n.

(e) Determine the statistical properties (e.g. expected values, variance) of the quantities calculated above.

(f) Determine, in the event that the distribution of Y is known to come from a set of distributions parameterized by subsets of $\mathbb{R}^n$, what, if any, the relationship is between the quantities estimated by maximum likelihood estimation and those estimated through the minimization of s_n.

(g) Prove the density of piecewise linear functions in the space of functions continuous on the hypercube. That is, prove that functions of the form $f(x_1, \ldots, x_n) = a_0 + \sum_{j=1}^m a_j \prod_{k \in K_j} \max(0, x_k - a_{j,k})$ where $K_j \subseteq \mathbb{N} \cap [0,n]$ are dense in $C([0,1]^n)$. This is known for $n = 1$ ([3], p. 376). This will yield continuous models which have a fair degree of explicability since most people understand "connect the dots". While I suspect they are, it is not clear to me that functions of the form $f(x_1, \ldots, x_n) = a_0 + \sum_{j=1}^m a_j \prod_{k \in K_j} \max(0, x_k - a_{j,k})^p$ where $K_j \subseteq \mathbb{N} \cap [0,n]$ for general $p \in (0, \infty)$ are dense in $C([0,1]^n)$. However, if they were, this would yield $p-1$ differentiable models. The importance of this result is that it would, I believe, provide another set of functional forms for the conditional expectation.

(h) Determine the proper way to view the inverse of link or activation functions by considering them as composition operators which map the space of continuous functions to the same space. This will specify the manner in which output from various models may be transformed and what the properties of that transformation are.

(i) Prove that trees, gradient boosted trees, and random forests have the same model form as that which is specified for generalized linear models. This will be helpful in demonstrating the unity among different, seemingly disparate, modeling techniques.

(j) Create n times differentiable approximations of common functions used in models (e.g. $I_{(a,\infty)}$ and $\max(x-a,0)$) so that optimization routines which require n times differentiable functions may be applied to parameter estimation.

2. Actuarial

(a) Determine the mathematical foundation of reserving. Briefly, classical actuarial mathematics regarding reserving posits the following. Let a measure space $(\Omega, \Sigma, \mathbb{P})$ and random variables Y and Y_t on Ω such that $Y < \infty$ almost everywhere, $0 < \mathbb{P}(0 < Y)$, $t \in [0,\infty)$, $Y_0 = 0$ almost everywhere, $0 \leq Y_s \leq Y_t$ when $s \leq t$, and $\lim_{t\to\infty} Y_t = Y$ be given. For $t \in [0,\infty)$ define $\Sigma_t = \sigma(\{Y_s : s \leq t\})$. Moreover, assume that there is some function $a : (0,\infty)$ such that $\mathrm{E}[Y|\Sigma_t] = a(t)Y_t$. I seek to determine

i. how to estimate the function a,

ii. the properties of the estimators of the parameters for a, and

iii. how to expand this framework.

(b) Determine the mathematical foundation of credibility. The Actuarial equivalent of confidence interval estimation, due to the presence of random variables which act as weights (i.e. exposure) has previously been "solved" using notion of credibility. I believe that this supposed solution is not rigorous. I will attempt to prove a Central Limit Theorem which directly incorporates the exposures associated with losses or loss counts.

(c) Determine the mathematical foundation of trending. In a similar sense, I think that the trending techniques of Actuarial Science are misguided and without rigor. Briefly, I think a solution presents itself first by estimating a model $\mathrm{E}[Y|T] = f(T)$, where Y is the quantity of interest and T is time, assuming some distribution of exposure for the interval in question, and integrating $f(T)$ over that interval. I will attempt to precisely state the fundamentals of the trending problem and demonstrate that my above proposal solves the problem.

3. Miscellany

I intend to determine the details of the following.

(a) For a finite number of random variables, each of which takes a finite number of values, all possible joint distributions of these variables may be identified with the unit hypercube for an appropriate dimension. Hence, one may randomly select joint probability distributions uniformly. As an application, one may then estimate the probability of misrepresentation in one of multiple categories due to random chance, by sampling from the space of joint distributions and then sampling from that distribution. This would allow one to estimate the false positive rate of assuming that a distribution of some categorical variable differing from the distribution in the population at large would constitute discrimination. This could help to make ideas in cases such as Hazelwood School District v. United States more precise.

(b) In a similar manner, I think it is possible to generate all distribution functions using the Hilbert Cube. Hence, it is possible to truly select something "at random" by first generating a sequence of random uniform numbers from the Hilbert cube. One may then map this to a specific distribution function and create a random sample from it. As an application, any scheme which can purportedly approximate a distribution may be tested by seeing if it can detect a randomly generated distribution.

(c) I think that by using a sufficiently flexible family of absolutely continuous functions, one may estimate arbitrary distribution functions using maximum likelihood estimation.

(d) I think it is possible to give a precise meaning to the term "variable standardization". The intention of variable standardization is to place all variables on the same scale in some sense. A basic result in probability theory is that if F_x is the cumulative distribution function of an absolutely continuous random variable X, then $F_x(X)$ is uniformly distributed. Hence, if one is able to approximate the distribution function of one's predictor variables using absolutely continuous functions, then one may standardize one's variables.

(e) While not clear to me from the literature, I think that all constrained optimization problems may be shown to correspond to a sequence of unconstrained optimization problems. In particular, I believe that this applies to linear, integer, and binary programming problems.

References

[1] Billingsley, Patrick. Probability And Measure Third Edition. New York: John Wiley & Sons, 1995. Print.

[2] Bruckner, Andrew M., Bruckner, Judith B., Thomson, Brian S., Real Analysis, Second Edition, ClassicalRealAnalysis.com, 2008, xiv 656 pp.

[3] Carothers, N. L.. Real Analysis. New York: Cambridge University Press, 2000. Print.

[4] Deters, Ian. (2022, June 21 – 24). The Mathematics Of Machine Learning [Conference presentation]. Amazon re:MARS 2022, Las Vegas, NV, United States. https://www.youtube.com/watch?v=Tk1nnou9Du0

[5] Durrett, Rick. Probability And Examples Third Edition. Belmont, California: Brooks/Cole, 2005. Print.

[6] Breiman, Leo Et al. Classification And Regression Trees First Edition. New York: Chapman and Hall / CRC, 1984.

[7] Cybenko, G. (1989) Approximations by superpositions of sigmoidal functions, Mathematics of Control, Signals, and Systems, 2(4), 303–314.

[8] Friedman, Jerome (1991). Multivariate Adaptive Regression Splines, The Annals of Statistics, 19(1), 1 – 67.

[9] McCullagh, P. and Nelder J. A. Generalized Linear Models Second Edition. New York, New York: Chapman And Hall, 1989.

1 Appendix

Lemma 1 *If X is parameterizable then it is separable.*

Proof. Since $E_n \subseteq \mathbb{R}^{d_n}$, E_n is separable ([3], p. 66). For each $n \in \mathbb{N}$, let $\hat{E}_n$ be a dense, countable subset of E_n. It shall be shown that $\hat{X} = \cup_{n=1}^{\infty} f_n(\hat{E}_n)$ is dense in X. To this end, let $x \in X$ and an open set $U \subseteq X$ such that $x \in U$ be given. By definition there is some $y \in \cup_{n=1}^{\infty} f_n(E_n)$ such that $y \in U$. Hence, there is some $n \in \mathbb{N}$ and $z \in E_n$ such that $f_n(z) = y \in f_n(E_n)$. Thus, there is some sequence $(z_k) \in \hat{E}_n$ such that $\lim_{k\to\infty} z_k = z$. Therefore, $\lim_{k\to\infty} f_n(z_k) = f(z) = y$. □

Lemma 2 *If X is parameterizable then there are sequences $(d_n) \in \mathbb{N}$, $E_n \subseteq \mathbb{R}^{d_n}$, and (f_n) such that f_n is continuous and $f_n(E_n) \subseteq f_{n+1}(E_{n+1})$ for $n \in \mathbb{N}$ and $\overline{\cup_{n=1}^{\infty} f(E_n)} = X$.*

A way to think about the construction that follows is to take the sets in question, prepend an index to each set, and fill out the remaining components with zeros. This allows one to see the sets (E_n) embedded in higher dimensions.

Proof. By definition there are sequences $(d_n) \in \mathbb{N}$, $E_n \subseteq \mathbb{R}^{d_n}$, and (f_n) such that f_n is continuous for $n \in \mathbb{N}$ and $\overline{\cup_{n=1}^{\infty} f(E_n)} = X$. Let $n \in \mathbb{N}$ be given. Define $\hat{d_n} = \max(d_1, \ldots, d_n) + 1$. For $1 \leq k \leq n$ define $p_k : \mathbb{R}^{\hat{d_n}} \to \mathbb{R}^{d_k}$ as the function that satisfies the equation $\pi_j \circ p_k = \pi_{j+1}$ for $1 \leq j \leq d_k$. Define $\hat{E_n} \subseteq \mathbb{R}^{\hat{d_n}}$ by $\hat{E_n} = \cup_{k=1}^{n} \pi_1^{-1}(\{k\}) \cap p_k^{-1}(E_k) \cap_{d_k+1<j} \pi_j^{-1}(\{0\})$. Define $\hat{f_n}$ on $\pi_1^{-1}(\{k\}) \cap p_k^{-1}(E_k) \cap_{d_k+1<j} \pi_j^{-1}(\{0\})$ by $\hat{f_n}(x) = f_k(p_k(x))$. By the Pasting Lemma, $\hat{f_n}$ is continuous. Since $\hat{f_n}(\hat{E_n}) = \cup_{k=1}^{n} f_k(E_k)$, $\hat{f_n}(\hat{E_n}) \subseteq \widehat{f_{n+1}}(\widehat{E_{n+1}})$. □

Lemma 3 *If $m, y \in (1, \infty)$ and then the sequence recursively defined by $x_{n+1} = x_n - \frac{x_n^m - y}{m x_n^{m-1}}$ where $x_0 = y$ converges to $y^{\frac{1}{m}}$.*

Proof. Define the function $T : \left(\left(\frac{y(m-1)}{2m-1}\right)^{\frac{1}{m}}, \infty\right) \to \mathbb{R}$ by $T(x) = x - \frac{x^m - y}{m x^{m-1}}$ and observe that on $\left(\left(\frac{y(m-1)}{2m-1}\right)^{\frac{1}{m}}, \infty\right)$,

$$-1 = \frac{m-1}{m}\left(1 - y\left(\left(\frac{y(m-1)}{2m-1}\right)^{\frac{1}{m}}\right)^{-m}\right) < \frac{m-1}{m}(1-yx^{-m}) = T'(x) < \frac{m-1}{m} < 1.$$

By the definition of T, $T\left(\left(\left(\frac{y(m-1)}{2m-1}\right)^{\frac{1}{m}}, y^{\frac{1}{m}}\right)\right) = \left(\left(\frac{y(m-1)}{2m-1}\right)^{\frac{1}{m}}, y^{\frac{1}{m}}\right)$ and $T\left(\left(y^{\frac{1}{m}}, \infty\right)\right) = \left(y^{\frac{1}{m}}, \infty\right)$. Thus, T, restricted to any strictly smaller interval containing $y^{\frac{1}{m}}$, is a contraction mapping and, by The Contraction Mapping Principle ([3], p. 98) the sequence (x_n) will converge if $x_0 \in \left(\left(\frac{y(m-1)}{2m-1}\right)^{\frac{1}{m}}, \infty\right)$. Clearly, $x_0 = y$ meets this criteria. □

Lemma 4 *Let (Ω, Σ, μ) be a finite measure space such that for all $0 < \varepsilon$ there is a partition $P \subseteq \Sigma$ of Ω such that $0 < \mu(E) < \varepsilon$ for all $E \in P$. Let $\mathcal{F}$ be the set of real valued Lebesgue measurable functions on Ω. There is no topology on $\mathcal{F}$ such that a sequence $(f_n) \in \mathcal{F}$ converges in that topology if and only if (f_n) converges almost everywhere.*

Proof. First, it will be shown that it may be supposed that any partition in question has only finitely many elements. To see this, let $0 < \varepsilon$ be given with associated partition P. Define $P_n = \{E \in P : \frac{1}{n+1} < \mu(E) \leq \frac{1}{n}\}$ and $P_\infty = \{E \in P : 1 < \mu(E)\}$ and observe that $P = P_\infty \cup (\cup_{n=1}^{\infty} P_n)$. Since P is a partition and $\mu(\Omega) < \infty$, $|P_\infty| < \infty$ and $|P_n| < \infty$ for $n \in \mathbb{N}$. Define $a_n = \sum_{E \in P_n} \mu(E)$. Since $\sum_{n=1}^{\infty} a_n < \infty$ there is some $N \in \mathbb{N}$ such that $\sum_{n=N}^{\infty} a_n < \varepsilon$. Define $\hat{P_N} = (\cup_{n=N}^{\infty} (\cup_{E \in P_n} E))$ and $\hat{P} = P_\infty \cup \hat{P_N} \cup_{n=1}^{N-1} P_n$. Observe that $\hat{P}$ is a finite partition of Ω such that $\mu(E) < \varepsilon$ for all $E \in \hat{P}$. Hence, without loss of generality, it may be supposed that for a given ε the associated partition P is finite.

For $n \in \mathbb{N}$, let P_n be a finite partition of Ω such that $0 < \mu(E) < \frac{1}{n}$. Let $E_{n,k}$ be an enumeration of the sets of P_n and define $c_n = |P_n|$ and, when $m = k - c_n + \sum_{j=1}^{n} c_j$ for $1 \leq k \leq c_n$, $f_m \in \mathcal{F}$ by $f_m(\omega) = 1$ if $\omega \in E_{n,k}$ and $f_m(\omega) = 0$ otherwise. Observe for $n \in \mathbb{N}$ and m such that $\sum_{j=1}^{n} c_j \leq m$ that $\mu(\{\omega \in \Omega : f_m(\omega) \neq 0\}) < \frac{1}{n+1}$. Hence, (f_m) converges in measure to 0. Let $\omega \in \Omega$ be given. For $n \in \mathbb{N}$ and $\omega \in \Omega$ there is some j such that $\omega \in E_{n,j}$. Define $m_j = j - c_n + \sum_{\ell=1}^{n} c_\ell$. Thus, $f_{m_j}(\omega) = 1$. In a similar manner, another subsequence $(m_{\hat{j}})$ may be constructed such that $f_{m_{\hat{j}}}(\omega) = 0$. Hence, (f_m) does not converge pointwise anywhere on Ω.

Suppose there was a topology on $\mathcal{F}$ whose convergent sequences are those which converge almost everywhere. Since (f_m) converges in probability, for any subsequence (f_{m_k}), there is some subsequence $(f_{m_{k_\ell}})$ of (f_{m_k}) such that $(f_{m_{k_\ell}})$ converges almost everywhere ([1], Theorem 20.5). Since (f_{m_k}) was an arbitrary subsequence, (f_m) converges in the topology ([5], p. 47) and hence almost everywhere. This is a contradiction. □