MATH 342W / 650.4 / RM742 Spring 2022 HW #1

Professor Adam Kapelner

Due 11:59PM Thursday, February 10, 2022 by email

(this document last updated 3:42am on Friday 11th February, 2022)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read the first chapter of "Learning from Data" and the introduction and Chapter 1 of Silver's book. Of course, you should be googling and reading about all the concepts introduced in class online. This is your responsibility to supplement in-class with your own readings.

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 *easy* 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 *difficult* 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 7 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. 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, print this document and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

NAME: Peter Antonaros

Problem 1

These are questions about Silver's book, the introduction and chapter 1.

(a) [easy] What is the difference between *predict* and *forecast*? Are these two terms used interchangably today?

The difference between "predict" and "forecast" is;

To predict means to give (specific) notice of an event in the future.

To forecast means to advise and plan for some possible future event(s).

Nowadays, many people use the two terms interchangeably. This is an ambiguity sacrifice we must make in our use of language; slight distortions of our actual message for ease of communication. We can still see hints of the original meanings, with the following example. Typically when we watch the news, the term used is weather forecasting, rather than weather prediction. This is because they are not stating "it will snow on February 10th at 12pm" (certainty of a set time in the future; prediction), but "there is a chance of snow next week, people should get their shovels and salt ready" (advising/planning for the future; forecasting)

(b) [easy] What is John P. Ioannidis's findings and what are its implications?

John P. Ioannidis hypothesized that most research findings are false, particularly the drug research published in academia. Bayer Laboratories confirmed this, by only being able to replicate results in about 1/3 of cases.

There are a few implications we can derive from this...

1) Research in "hot fields" often leads to competition, and the tendency to exaggerate results for the praise of being the first in this field to (insert accomplishment here) 2) Where there are large amounts of money there is a higher likelihood of corruption. If the dominating motivational factor behind the research is absolute profit, this can lead to skewed or (more cynically), intentionally false results. 3) Research needs to adhere to some universal standard in how they define terminology. If Research Group 1 classifies "safe" as having a toxicity level below X and Research Group 2 classifies "safe" as having a toxicity level below Y, then each group may find the results of the other to be irreproducible.

(c) [easy] What are the human being's most powerful defense (according to Silver)? Answer using the language from class.

Pattern Recognition and Spacial Reduction.

Humans have an amazing innate ability to recognize minute details about some situation/object/environment and determine the best response.

For example, you are walking through a forest a night, you hear a crack of a branch, and after looking, you see two glowing eyes. Your brain immediately locates the area from which the sound originated, processes the image of eyes looking at you which your brain determines is a threat, and you run away. From the given sensory metrics; the height of the glowing eyes, the loudness of the cracking branch, and the distance

from the eyes, your brain modeled multiple scenarios creating a massive event space. It then went on to simulate each event(scenario), creating a separate space of event results. The mapping of scenarios to outcomes, reduction of outcomes, and search for the optimal outcome, occurs in a fractional second. The brain and body have just ensured the highest probability of self-preservation.

(d) [easy] Information is increasing at a rapid pace, but what is not increasing?

Our understanding of how to process it.

This isn't to say we've made no progress at all but in comparison to how quickly the rate of informational growth is, one can argue our progress regarding the effective understanding of how to process it, has been and remains marginal.

(e) [difficult] Silver admits that we will always be subjectively biased when making predictions. However, he believes there is an objective truth. In class, how did we describe the objective truth? Answer using notation from class i.e. $t, f, g, h^*, \delta, \epsilon, t, z_1, \ldots, z_t, \delta, \mathbb{D}$, $\mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p,\ x_{\cdot 1}, \ldots, x_{\cdot p}, x_{1\cdot}, \ldots, x_{n\cdot}$, etc.

In class we defined the objective truth with our function f. This function is unknown and will forever remain unknown, but in an abstract sense it is the "only" correct answer. This is why we aim to minimize error and in the end reach an approximating function g.

This objective truth function f captures all features (without ignorance) required to make a completely accurate prediction.

(f) [easy] In a nutshell, what is Karl Popper's (a famous philosopher of science) definition of *science*?

Karl Popper's Theory of Falsification defines science as the ability to test a proposed theory without the guarantee that it is True or False, but that both remain valid outcomes. Karl Popper's emphasis on the fact that science should aim to disprove ensures the theory is not in the set of unfalsifiable statements, and thus in its very nature is scientific, whether right or wrong.

(g) [harder] Why did the ratings agencies say the probability of a CDO defaulting was 0.12% instead of the 28% that actually occurred? Answer using concepts from class.

Nate Silver states that an important factor as to why the ratings agencies were so off on their expected CDO default rates were because they had little to no track record to base their predictions on.

In the context of our class and modeling in general, we would say they had little historical data, which would have negatively impacted the model specifically when attempting to learn from data and validate predictions.

I would also like to add there was massive incentive to "lie" about these probabilities, which would be considered an external influence on the model. The people creating these models directly benefitted from lower default rate predictions. (I won't get too deep into this, but this is also a key factor)

(h) [easy] What is the difference between *risk* and *uncertainty* according to Silver's definitions?

Risk: There are "odds" for a particular event which are known ahead of time. This allows you to make your own informed decision about the event and how to "play" it.

Uncertainty: The odds for a particular event are not known ahead of time. This means you cannot make an informed decision ahead of time about the event, leaving you in the "dark"

Risk is good, uncertainty is bad. Risk is what spurs the economy, uncertainty grinds it to a halt. (Reworded from Silver)

(i) [difficult] How does Silver define *out of sample*? Answer using notation from class i.e. $t, f, g, h^*, \delta, \epsilon, z_1, \ldots, z_t, \delta, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}, x_{1\cdot}, \ldots, x_{n\cdot}$, etc. WARNING: Silver defines *out of sample* completely differently than the literature, than practitioners in industry and how we will define it in class in a month or so. We will explore what he is talking about in class in the future and we will term this concept differently, using the more widely accepted terminology. So please forget the phrase *out of sample* for now as we will introduce it later in class as something else. There will be other such terms in his book and I will provide this disclaimer at these appropriate times.

Silver doesn't directly define out of sample, but does so through his drunk driving example.

I believe what he is defining "out of sample" to be is, the current event you are evaluating has little to no historical record. You may have a large data set, but all the observations are under a different category.

A data set of 20,000 drives, is useful, but when trying to predict how safely you would drive when drunk, if none of those 20,000 drives were under the influence then you have no basis for your prediction.

(I do not know how to represent this with our notation, without making a mess of it, spare me)

(j) [harder] Look up *bias* and *variance* online or in a statistics textbook. Connect these concepts to Silver's terms *accuracy* and *precision*. This is another example of Silver using non-standard terminology.

Statistics Definitions (Informal, but also not plain English definitions)

Bias: Difference between expected value and actual value of a given parameter

Variance: How dispersed a set of numbers are; a measure of how far the numbers lie in relation to their mean

Silver's Definitions (My opinion is that he was making a connection with fear and greed)

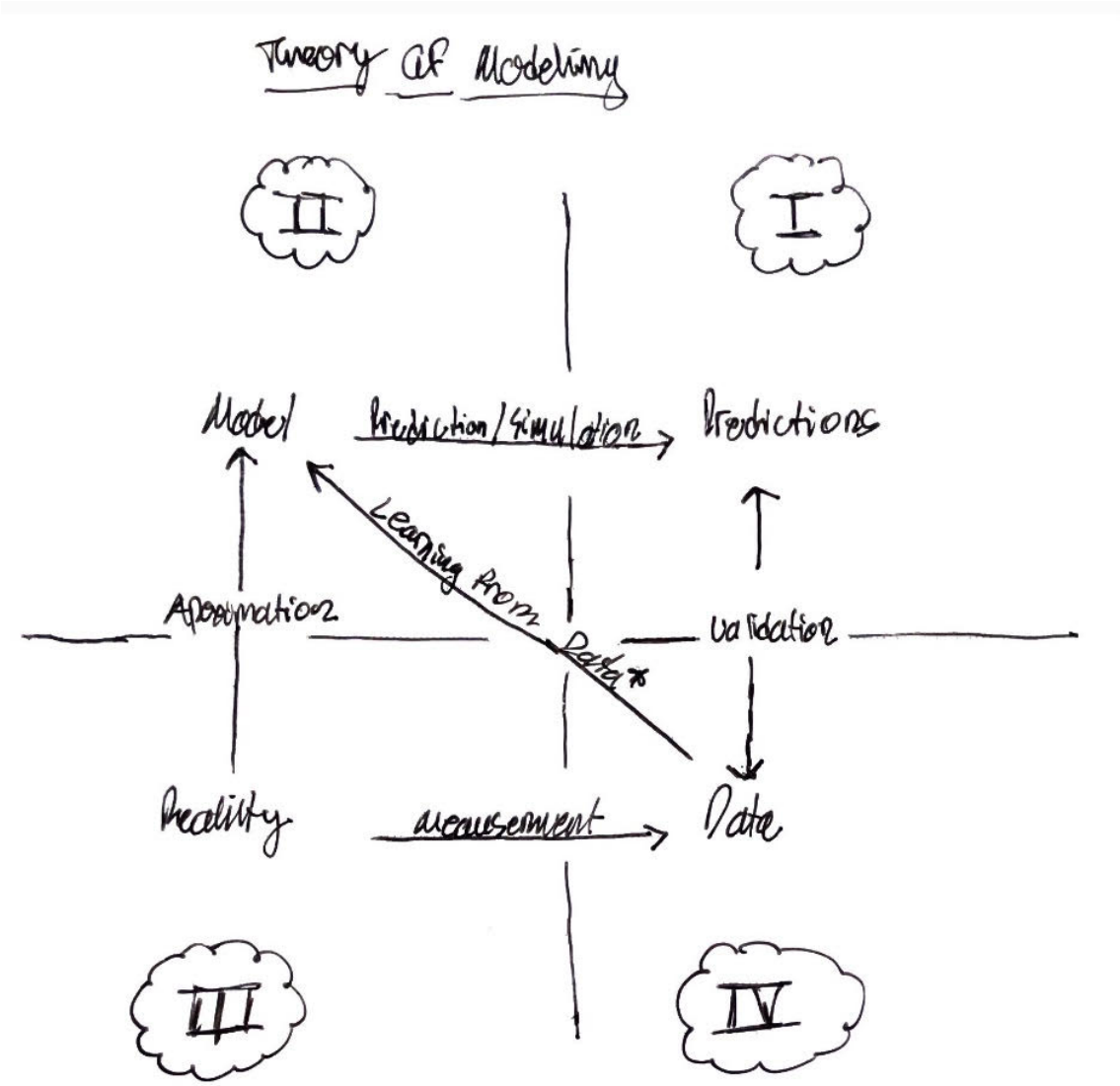
Bias("Fear"): People have an inherent fear of the difference between what they think will occur and what will actually happen. (Bias)

Variance("Greed"): Everyone has a unique level of greed leading to dispersion of "wealth" among people (Variance)

Problem 2

Below are some questions about the theory of modeling.

(a) [easy] Redraw the illustration from lecture one except do not use the Earth and a table-top globe. The quadrants are connected with arrows. Label these arrows appropriately.



(b) [easy] Pursuant to the fix in the previous question, how do we define *data* for the purposes of this class?

We define data as the natural result of measuring the phenomenon/a.

(c) [easy] Pursuant to the fix in the previous question, how do we define *predictions* for the purposes of this class?

(d) [easy] Why are “all models wrong”? We are quoting the famous statisticians George Box and Norman Draper here.

All models are wrong because they are merely approximations to the real phenomenon. They do not capture every feature with absolute accuracy, and so there will always be some inherent degree of error in the model.

(e) [harder] Why are “[some models] useful”? We are quoting the famous statisticians George Box and Norman Draper here.

Some models are useful because this error we stated above, can be minimized in such a way that the model produces results which are "very close" to the real thing. The use of "very close" here is intentionally abstract since there is a dependence on the model's application. Are we modeling weather, are we modeling Brownian Motion, they each require a unique definition of "very close", and can be considered useful in their particular application.

(f) [harder] What is the difference between a "good model" and a "bad model"?

I believe its easier to first characterize a "bad model" and from that we can say what a "good model" is. I am also defining incorrect to mean; our final error is greater than the error bound need to accurately predict output of our phenomena.

A model can be considered a bad one when, despite accurate, high quality, and diverse inputs still yields outputs (predictions) that are incorrect. A model can be considered a good one when, accurate, high quality, and diverse inputs yield correct outputs. (correct as in opposite of incorrect definition above)

Problem 3

We are now going to investigate the famous English aphorism “an apple a day keeps the doctor away” as a model. We will use this as springboard to ask more questions about the framework of modeling we introduced in this class.

(a) [easy] Is this a mathematical model? Yes / no and why.

I personally would consider this to be a Mathematical Model, albeit an abstract one. It has a clearly defined phenomena, eating apples prevents you from seeing the doctor (as much) The statement itself is easily tested; does eating 1 apple per day result in no doctors visits (or less depending on your interpretation of the statement)

We can quantify the features with measurable metrics.

In an abstract sense then yes this can be a mathematical model, but there are certainly things we can say to improve it.

(b) [easy] What is(are) the input(s) in this model?

Without assuming too much about the aphorism we can simply say the input(s) are:

Apples eaten per day

(Perhaps) Type of apple(s) eaten

I will just stick with the first one since it requires no assumptions about what the statement is directly saying. I will also use number of apples eaten per day rather than has the person eaten an apple per day that way we don't need to coerce to numeric.

(c) [easy] What is(are) the output(s) in this model?

Similar to the previous, making no assumptions about the statement, the output(s) are:

Does the doctor stay away (Yes or no)

(d) [harder] How good / bad do you think this model is and why?

I don't believe this model will be any good due to the reasons below:

We have not captured many features of the phenomena. There are too many external forces at play we haven't considered.

Our output is "sketchy". What does it really mean for the doctor to stay away? (How do we interpret and thereby quantify this)

Can we obtain a large enough dataset for this?

Obviously there are many more reasons as to why this is a poor model, but these are certainly major ones.

(e) [easy] Devise a metric for gauging the main input. Call this x_1 going forward.

x_1 = Number of apples eaten in a 24hr time period

(f) [easy] Devise a metric for gauging the main output. Call this y going forward.

y = Time between consecutive doctor visits

(g) [easy] What is $\mathcal{Y}$ mathematically?

Our output space.

(h) [easy] Briefly describe $z_1, \ldots, z_t$ in English where $y = t(z_1, \ldots, z_t)$ in this *phenomenon* (not *model*).

I am assuming 1 causal driver to simplify the phenomenon;

z_1 = has the person eaten an apple today $\in 0, 1$. We then use x_1, similar to how we used yearly salary to measure will a person payback loan at maturity (in class example)

(i) [easy] From this point on, you only observe x_1. What is the value of p?

$p = 1$, since we are only concerned with 1 regressor x_1

(j) [harder] What is $\mathcal{X}$ mathematically? If your information contained in x_1 is non-numeric, you must coerce it to be numeric at this point.

$\mathcal{X}$ is the input space/covariate space

(k) [easy] How did we term the functional relationship between y and x_1? Is it approximate or equals?

We considered the relationship between y and x_1 to be an approximate one.

(l) [easy] Briefly describe *superivised learning*.

The answer is in the term "supervised".

The human acts as a supervisor to the computer (algorithm) by training it on data that has been pre-labeled. We evaluate the input to output relationship based on this training data, and since the machine is not creating the labels itself, we consider it to be supervised learning rather than unsupervised learning.

(m) [easy] Why is *superivised learning* an *empirical solution* and not an *analytic solution*?

First we can see that the definition of empirical is to be based on/concerned with/ verified by, observation or experience.

In an abstract sense this is the exact definition of supervised learning. We have data (observations), we map inputs to outputs (concern with regard to observation), and finally validations of our mappings (observation verification). In this way we are building experiences for the machine to then repeat on new, and un-mapped set of input values.

(n) [harder] From this point on, assume we are involved in supervised learning to achieve the goal you stated in the previous question. Briefly describe what $\mathbb{D}$ would look like here.

$\mathbb{D}$ would look like... (of course with respect to our apple a day model)

$\mathbb{D} = \{\langle \vec{x}_1, y_1 \rangle, \langle \vec{x}_2, y_2 \rangle, \ldots, \langle \vec{x}_n, y_n \rangle\}$, Sample size n

$\mathbb{D} = \langle \vec{X}, \vec{y} \rangle$

$$\vec{X} = \begin{bmatrix} \leftarrow \vec{x}_1 \rightarrow \\ \leftarrow \vec{x}_2 \rightarrow \\ \leftarrow \vec{x}_n \rightarrow \end{bmatrix}, \quad \vec{y} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

(o) [harder] Briefly describe the role of $\mathcal{H}$ and $\mathcal{A}$ here.

The role of $\mathcal{H}$ is to be the space of candidate functions h that approximate f

The role of our algorithm $\mathcal{A}$ is to "select" a "good" function g which is an approximation to f and an $\in \mathcal{H}$.

(p) [easy] If $g = \mathcal{A}(\mathbb{D}, \mathcal{H})$, what should the domain and range of g be?

The domain of g is $(\mathbb{D}, \mathcal{H})$

The range of g is the outputs produced by our algorithm $\mathcal{A}$, with inputs $\mathbb{D}, \mathcal{H}$

(q) [easy] Is $g \in \mathcal{H}$? Why or why not?

Yes g is an element of $\mathcal{H}$ because we consider $\mathcal{H}$ to be the space of all possible hypotheses. We then determine g by sampling $\mathcal{H}$ for the best candidate function that approximates f.

(r) [easy] Given a never-before-seen value of x_1 which we denote x^*, what formula would we use to predict the corresponding value of the output? Denote this prediction $\hat{y}^*$.

Given a never before seen value x_1 denoted x^*, we use the formula $\hat{y}^* =$

(s) [harder] In lecture I left out the definition of f. It is the function that is the best possible fit of the phenomenon given the covariates. We will unfortunately not be able to define "best" until later in the course. But you can think of it as a device that extracts all possible information from the covariates and whatever is left over δ is due exclusively to information you do not have. Is it reasonable to assume $f \in \mathcal{H}$? Why or why not?

No it is not reasonable to assume that f is $\in \mathcal{H}$. I'm sure this is not mathematically sound, but as far as reasoning in my head goes...

f can live in what is essentially a space with dimension $\leq$ number of data points. The best function to fit the points are the points themselves. If we then say $f \in \mathcal{H}$, a fallacy arises. $\mathcal{H}$ having larger dimension than f is impossible, or better yet completely pointless. We cannot know f, and so if $f \in \mathcal{H}$ then we couldn't even possibly know our space of hypotheses.

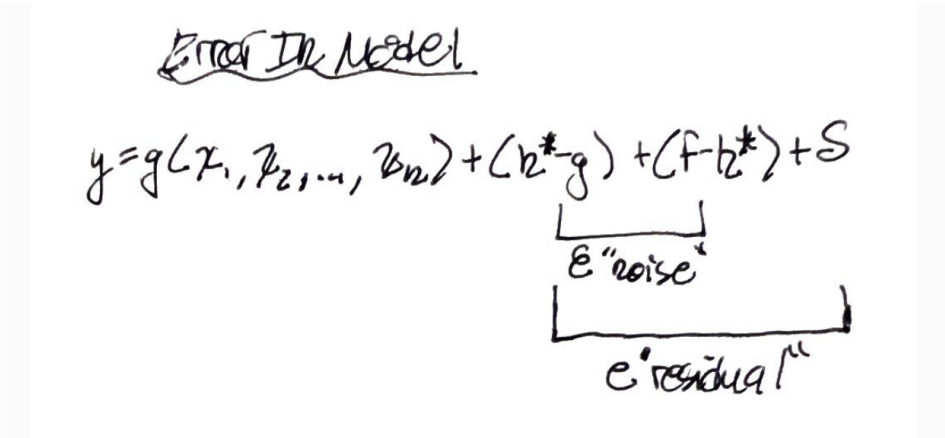
(t) [easy] In the general modeling setup, if $f \notin \mathcal{H}$, what are the three sources of error? Copy the equation from the class notes. Denote the names of each error and provide a sentence explanation of each. Denote also e and $\mathcal{E}$ using underbraces / overbraces.

Three sources of error:

Error due to Ignorance: Lack of information/observation

Error due to Misspecification: Candidate Model doesn't accurately capture f

Error due to Estimation: Candidate model doesn't agree with g



(u) [easy] In the general modeling setup, for each of the three source of error, explain what you would do to reduce the source of error as best as you can.

Three sources of error and how to reduce them...

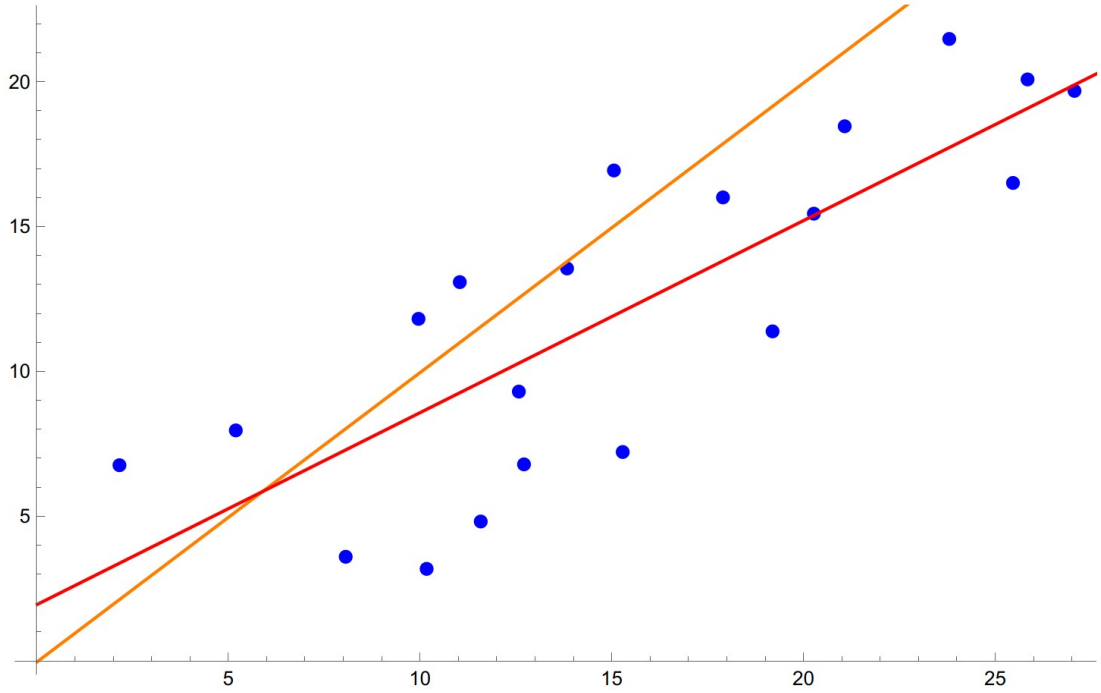
Ignorance: Collect more relevant regressors

Misspecification: Richer $\mathcal{H}$ space (larger space of possible functions)

Estimation: Run a better algorithm or collect more observations (n)

(v) [harder] In the general modeling setup, make up an f, an h^* and a g and plot them on a graph of y vs x (assume $p = 1$). Indicate the sources of error on this plot (see last question). Which source of error is missing from the picture? Why?

Plot is made using Mathematica



The blue dots denote our data points (observation)

The red line denotes our h^* (candidate model that most closely resembles f)

The orange line denotes our g (particular linear function after applying an algorithm to $\mathbb{D}$ and $\mathcal{H}$)

I would say the error missing from the picture is error due to ignorance. We don't know the quality of the observations, where they came from etc

(w) [easy] What is a null model g_0? What data does it make use of? What data does it not make use of?

The null model g_0 is a model with no features.

It makes use of an algorithm $\mathcal{A}$, y, and $\mathcal{H}$. The algorithm is typically mode(), as stated in class.

(x) [easy] What is a parameter in $\mathcal{H}$?

A parameter in $\mathcal{H}$ would be θ.

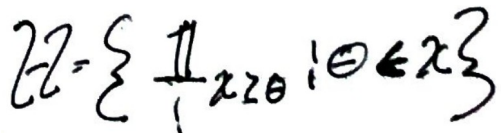
For a concrete example, I believe θ in the Threshold Model would be the value of the threshold itself.

(y) [easy] Regardless of your answer to what $\mathcal{Y}$ was above in (g), we now coerce $\mathcal{Y} = \{0, 1\}$. What would the null model g_0 be and why?

The Null Model g_0 would be the "thing" that exists in the absence of any features, only including observations. For this reason we use a mode() algorithm for g_0 so that is simply returns the most frequent occurrence of observation.

(z) [easy] Regardless of your answer to what $\mathcal{Y}$ was above in (g), we now coerce $\mathcal{Y} = \{0, 1\}$. If we use a threshold model, what would $\mathcal{H}$ be? What would the parameter(s) be?

Our $\mathcal{H}$ would be the space of all possible thresholds.



Our parameter θ is the value for which any subsequent x results in our indicator function returning 1.

(aa) [easy] Give an explicit example of g under the threshold model.

g under the threshold model is a Boolean function that returns either 0 or 1 depending on this threshold value.

An example for g with a threshold value of 0 would be...

$$g = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

Problem 4

As alluded to in class, modeling is synonymous with the entire enterprise of science. In 1964, Richard Feynman, a famous physicist and public intellectual with an inimitably captivating presentation style, gave a series of seven lectures in 1964 at Cornell University on the "character of physical law". Here is a 10min excerpt of one of these lectures about the scientific method. Feel free to watch the entire clip, but for the purposes of this class, we are only interested in the following segments: 0:00-1:00 and 3:48-6:45.

(a) [harder] According to Feynman, how does the scientific method differ from learning from data with regards to building models for reality? (0:08)

According to Feynman the scientific method starts with a guess -> compute consequences -> compare to experiment

We can see this is different from the process of learning from data. Initially we are not guessing at a new law, but determining how to model reality (feature selection). The

remaining two steps in Feynman's scientific method can be seen as quite similar (in an abstract sense) to learning from data. Computing consequences -> making predictions, and compare to experiment -> validating model to our datset.

(b) [harder] He uses the phrase "compute consequences". What word did we use in class for "compute consequences"? This word also appears in your diagram in 2a. (0:14)

Predictions/Simulations; The way I see it is that by compute consequences, we are seeing how our model performs. Since models are always wrong technically what we are doing is "computing consequences".

(c) [harder] When he says compare consequences to "experiment", what word did we use in class for "experiment"? This word also appears in your diagram in 2a. (0:29)

Validation; we compare our consequential predictions to the actual data obtained about the phenomena.

(d) [harder] When he says "compare consequences to experiment", which part of the diagram in 2a is that comparison?

Since I answered with validation in the previous question this lies as the connection between the first and fourth quadrants in our diagram.

(e) [difficult] When he says "if it disagrees with experiment, it's wrong" (0:44), would a data scientist agree/disagree? What would the data scientist further comment?

In the most literal sense yes, a Data Scientist would agree. The further comment would be something to the effect of, despite disagreeing with experiment, just how "wrong" is it? Modeling is not about being right, but to the degree of which you are wrong. Similar to Feynman stating we are never right, just sure we are wrong. I would expect a Data Scientist to conclude by saying something similar to "We are concerned with abstracting the phenomena to a finite number of features, while simultaneously minimizing the error of these abstractions"

(f) [difficult] [You can skip his UFO discussion as it belongs in a class on statistical inference on the topic of H_0 vs H_a which is *not* in the curriculum of this class.] He then goes on to say "We can disprove any definite theory. We never prove [a theory] right... We can only be sure we're wrong" (3:48 - 5:08). What does this mean about models in the context of our class?

In the context of our class, we can be sure that all of our models will be wrong. It will be impossible for us to ever find the true f (function), everything else is merely an approximation with varying accuracy.

We will also have to be careful how we interpret the results of our model. Even if the predictions are accurate we cannot say anything definite. We will have to resort to probabilistic statements. These sorts of statements account for the event that X centuries later someone, somewhere might have a counterexample to the results of your model. For this reason Feynman states we can only be sure a theory is wrong and that we can disprove any definite theory.

(g) [difficult] Further he says, "you cannot prove a *vague* theory wrong" (5:10 - 5:48). What does this mean in the context of mathematical models and metrics?

With regards to a Mathematical Model, you cannot prove a vague theory wrong would mean, if your features are difficult or impossible to quantify with REAL metrics, then you're essentially modeling nonsense. Your model will not be right nor will it be wrong, it will simply contribute to the noise surrounding the phenomena (Nate Silver in Signal and Noise said something similar). This also reminds me of your statement about depression in lecture. How do we measure depression, better yet how do we even define it? Without nailing these down with specific definitions, we cannot say much about our model. It is "shielded" by ambiguity, and according to Feynman would not be considered scientific.

(h) [difficult] He then he continues with an example from psychology. Remeber in the 1960's psychoanalysis was very popular. What is his remedy for being able to prove the vague psychology theory right (5:49 - 6:29)?

His remedy is simple; clearly define and more importantly quantify the features of the theory.

In this case, little Jimmy says his mother doesn't love him (subjective), and Feynman states he would like a measurement, specifically one you could state ahead of time of "how much love is not enough, and how much love is overindulgent". Then at least you would have a basis to compare your results/outcomes to, otherwise you are simply guessing in the dark and drawing conclusions on the basis of these faulty guesses

(i) [difficult] He then says "then you can't claim to know anything about it" (6:40). Why can't you know anything about it?

Feynman said this because there are those who claim they can't precisely define something, but at the same time maintain the stance that they know something about it. Feynman directly ties the ability to define "something" and knowing what "something" is, together. According to him, without definition you have no idea what is going on and cannot claim otherwise.