Predictive Analytics Lecture 2

Adam Kapelner

Stat 422/722
at The Wharton School of the University of Pennsylvania

January 24 & 25, 2017

No Inference Possible as of Now

We haven't spoken about t or F tests. Why is that?

In order to have inference, we need to make explicit random variable model assumptions e.g.

$$Y \sim g(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma^2, \ldots)$$

must be assumed to be something like

$$Y \sim \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\, \sigma^2\right)$$

Is this a reasonable thing to do?

Back to Modeling

We said before that our model for Y was

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

assuming we can know the model, there still is $\mathcal{E}$. Where does it come from? According to determinism a la Laplace, if one knew all the causal information, there would be no error

$$y = t(z_1, z_2, \ldots)$$

i.e t is the deterministic true mathematical model.

Laplace Believes in Demons

Universal determinism and Laplace's demon

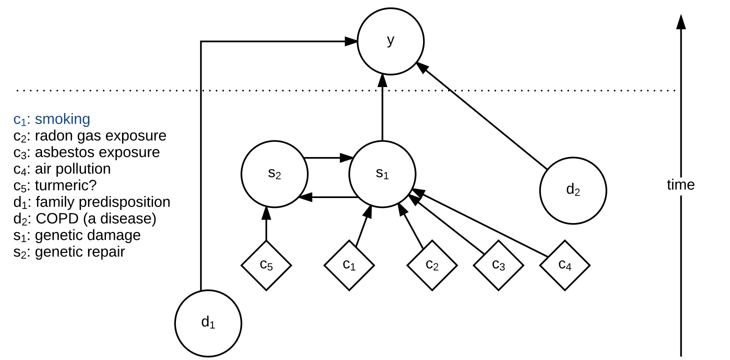
Laplace writes:

> We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it – an intelligence sufficiently vast to submit these data to analysis – it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past would be present to its eyes.
>
> (1814: 4)

The vast intelligence here described has come to be known as Laplace's demon. The idea is obviously founded on that of a human scientist (perhaps Laplace himself) using Newtonian mechanics to calculate the future paths of planets and comets. Extrapolating from this success, it was natural to suppose that a sufficiently vast intelligence could calculate the entire future course of the universe. Laplace himself relates his vast intelligence to human successes in astronomy. As he says:

> The human mind offers, in the perfection which it has been able to give to astronomy, a feeble idea of this intelligence. Its discoveries in mechanics and geometry, added to that of universal gravity, have enabled it to comprehend in the same analytical expressions the past and future states of the system of the world.
>
> (Laplace 1814: 4)

Example Lung Cancer Causal Model



Arrows represent causal directions and diamond boxes represent "manipulable" variables (more on this soon). What functions for the response would be deterministic?

$$y = t(d_1, d_2, s_1), y = t(d_1, d_2, s_2, c_1, c_2, c_3, c_4), y = t(d_1, d_2, c_1, c_2, c_3, c_4, c_5)$$

The Root Cause of Randomness

But let's say we only have information about c_1 (a contributory cause, one among many co-occurrent causes). Since we don't have all the inputs (nor the information of the states of the co-occurent causes), we cannot be sure of y. Hence we'll employ a statistical model,

$$Y \sim \text{Bernoulli}\left(f(c_1)\right)$$

where we saw before that $f(c_1 = 1) = 16\%$ and $f(c_1 = 0) = 0.4\%$ (AKA "probabilistic causation"). Thus, the response is stochastic only because we lack information. For regression,

$$y = f(x_1, \ldots, x_p) + \underbrace{t(z_1, z_2, \ldots) - f(x_1, \ldots, x_p)}_{\substack{\mathcal{E} \\ \text{(i.e. the "noise" is due to ignorance)}}}$$

Note... some believe that there is still intrinsic randomness in the universe even with all relevant information known. But we are punting on the actual philosophy...

Sidebar: Other Sources of Error

$$y = f(x_1, \ldots, x_p) + \underbrace{t(z_1, z_2, \ldots) - f(x_1, \ldots, x_p)}_{\substack{\mathcal{E} \\ \text{(i.e. the ``noise'' is due to ignorance)}}}$$

Then if we make a parametric assumption,

$$\begin{aligned} y \quad = \quad & s(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell) + \\ & \underbrace{f(x_1, \ldots, x_p) - s(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell)}_{\text{model misspecification}} + \\ & \underbrace{t(z_1, z_2, \ldots) - f(x_1, \ldots, x_p)}_{\substack{\mathcal{E} \\ \text{noise due to ignorance error}}} \end{aligned}$$

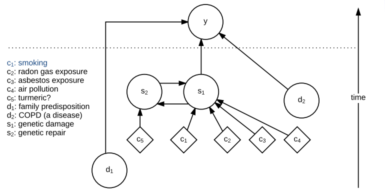
Sidebar: Other Sources of Error

Further, we then have to estimate the parameters to get a fit:

$$
\begin{aligned}
y \quad = \quad & \underbrace{\hat{s}(x_1, \ldots, x_p; \hat{\theta}_1, \ldots, \hat{\theta}_\ell)}_{\hat{y}} + \\
& \underbrace{s(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell) - \hat{s}(x_1, \ldots, x_p; \hat{\theta}_1, \ldots, \hat{\theta}_\ell)}_{\text{model / parameter estimation error}} + \\
& \underbrace{f(x_1, \ldots, x_p) - s(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell)}_{\text{model misspecification error}} + \\
& \underbrace{t(z_1, z_2, \ldots) - f(x_1, \ldots, x_p)}_{\substack{\mathcal{E} \\ \text{noise due to ignorance}}}
\end{aligned}
$$

Thus, all predictions have three sources of error. What is minimized with non-parametric machine learning?

t is Difficult to Model

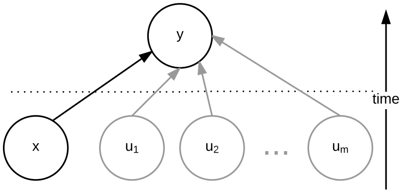


In order to get t, you'll need to know all these functions explicitly:

$$
\begin{aligned}
y &= t_y(d_1, d_2, s_1) \\
s_1 &= t_{s_1}(c_1, c_2, c_3, c_4, s_2) \\
s_2 &= f_{s_2}(c_5, s_1)
\end{aligned}
$$

which means that even if you know all the values of variables, you may not be able to properly model the response since ... you do not know the functional forms t_y, t_{s_1} and t_{s_2}.

A "Nice" Type of Ignorance



In the situation where the true model is

$$y = g(x) + h_1(u_1) + h_2(u_2) + \ldots + h_m(u_m)$$

and x is observed but $u_1, \ldots, u_m$ are the "unknowns".

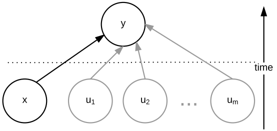
$$h_1(u_1) + h_2(u_2) + \ldots + h_m(u_m) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\sum_{k=1}^{m} \mu_k, \sum_{k=1}^{m} \sigma_k^2\right)$$

as the number of unseen variables increase (central limit theorem) and if … they're somewhat independent.

The Normal Homoskedastic Error Model

Let $\mathcal{E}_0 = \sum_{k=1}^{m} \mu_k$ and $\sigma^2_{\mathcal{E}} = \sum_{k=1}^{m} \sigma^2_k$, then

$$y = \underbrace{g(x) + \mathcal{E}_0}_{f(x)} + \mathcal{E} \quad \text{s.t.} \quad \mathcal{E} = \sum_{k=1}^{m} h_k(u_k) - \mathcal{E}_0 \sim \mathcal{N}\left(0, \sigma^2_{\mathcal{E}}\right)$$



Also, since x does not affect the other variables in any way, it cannot have an influence on their spread, hence σ^2 is not a function of x. Thus the error spread is the same everywhere across the range of x (homoskedasticity).

Parametric Worldview

We are back to the fundamental statistical problem, $Y = f(x) + \mathcal{E}$ where now we are more "okay" with the noise being normal and homoskedastic for all x.

We now invoke the parametric worldview. Within that parametric worldview, we will buy into the linear model. Thus,

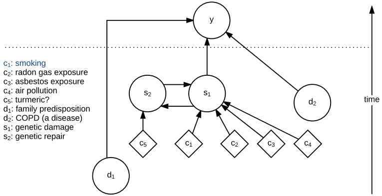
$$Y \sim \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\, \sigma^2\right)$$

But there is one more assumption...

Independence

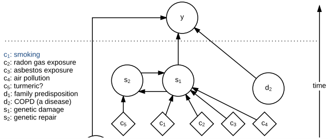
We now assume that each response is independent of every other response.

Second person:



No effect of first person's y_1 (nor any of the unobserved variables which generate the $\mathcal{E}_1$) on the second person's y (or $\mathcal{E}_2$).

First person:



If there are, we need to observe them and rotate them into our estimate of $f(x)$. Examples for this cigarette case?

The Classic OLS Assumptions

Preassuming

- linearity (the parametric assumption)

we then further assume

- independence (most important)
- homoskedasticity (less important)
- normality of $\mathcal{E}$ (least important if n is large)

in order to get inference. Changing these assumptions gives entirely new modeling techniques and inference. It is called "generalized linear model" theory.

A Different Means of Estimation

Last time, we were working on creating a fit $\hat{f}$ that means we need estimates of all the parameters:

$$\hat{f}(x_1, x_2, \ldots, x_p) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

where the unknown parameters were $\beta_0, \beta_1, \ldots, \beta_p$. Our strategy last time was to minimize SSE via a calculus to obtain $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. Why was this arbitrary?

Given the three new assumptions, we now have a completely specified joint probability distribution for our observed data,

$$\mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right)$$

where $\boldsymbol{x}_i := [x_{i1}, x_{i2}, \ldots, x_{ip}]$ i.e. the vector of all known measurements / covariates.

What's a probability? What's a likelihood?

In general, a parametric density function / mass function of a r.v. looks like the following:

$$\mathbb{P}(x; \theta) = \ldots$$

where θ are the tuning knobs on the model. We ask the question "what's the probability of this realization x (the data) assuming the density was parameterized at θ"? Now we ask the inverse question:

$$\mathcal{L}(\theta; x) = \ldots$$

that is "what's the likelihood of these parameters assuming we saw x (the data) come out the way it did"? The $\mathcal{L}()$ denotes the **likelihood function**. Of course, probability and likelihood are exactly the same numerically,

$$\mathbb{P}(x; \theta) = \mathcal{L}(\theta; x) = \ldots$$

but conceptually they couldn't be further apart!

Maximum Likelihood Estimation (MLE)

Why not just ask the very common-sense question, what θ maximizes the probability of seeing what we observe? That would be a good guess as to what θ is.

$$\hat{\theta} := \underset{\theta \in \Theta}{\arg\max} \left\{ \mathcal{L}(\theta; x) \right\} = \underset{\theta \in \Theta}{\arg\max} \left\{ \underbrace{\ln\left(\mathcal{L}(\theta; x)\right)}_{\ell(\theta; x)} \right\}$$

where Θ represents the space the parameter lives in. In our situation, Θ represents all real numbers in p dimensions. Let's do this in our example. The first step:

$$\begin{aligned} & \mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right) \\ = & \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i \mid \boldsymbol{X}_1 = \boldsymbol{x}_i\right) \end{aligned}$$

How so? Each observation is independent of every other. Recall $\mathbb{P}(ABC) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)$ if A, B and C are independent.

MLE of the Linear Model Parameters

We can continue,

$$
\begin{aligned}
&= \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i \mid \boldsymbol{X}_1 = \boldsymbol{x}_i\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y - \mathbb{E}\left[Y_i \mid \boldsymbol{X}_i\right]\right)^2\right)
\end{aligned}
$$

How? Normality and homoskedasticity of $\mathcal{E}$.

$$
= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y - \left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)\right)^2\right)
$$

How? Linearity of $\mathbb{E}\left[Y_i \mid \boldsymbol{X}_i\right]$. Now we wish to maximize the above over all possible $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$. That's the $\arg\max_{\theta \in \Theta} \left\{\mathcal{L}\left(\theta; x\right)\right\}$ step.

MLE of the Linear Model Parameters

Then, by some precalc tricks,

$$
\begin{aligned}
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\mathcal{E}_i^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\sum_{i=1}^{n} -\frac{1}{2\sigma^2}\mathcal{E}_i^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathcal{E}_i^2\right)
\end{aligned}
$$

Pick $\left\{\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_p, \hat\sigma^2\right\}$ such that the above is minimized. The solutions are called the "maximum likelihood estimates (MLE's)".

Using calculus, the solution to $\left\{\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_p\right\}$ is equivalent to minimizing SSE... What a coincidence!!

Note also: $\hat\sigma^2 = \frac{1}{n} SSE = MSE$. Why was there no $\hat\sigma^2$ until now?

The Likelihood Ratio (LR)

Imagine two models: (a) the "full" model where $\theta \in \Theta$ and (b) a reduced model where $\theta \in \Theta_R \subset \Theta$. The reduced space has q less degrees of freedom for θ to live within. Consider the ratio of the likelihoods

$$LR := \max_{\theta \in \Theta} \mathcal{L}(\theta; x) \,/\, \max_{\theta \in \Theta_R} \mathcal{L}(\theta; x)$$

representing how much more probable the full model is over the restricted model. But is this is this increase in probability statistically significant? It turns out as n gets large and under pretty forgiving conditions,

$$Q \;\; := \;\; 2 \ln (LR) \overset{\mathcal{D}}{\to} \chi^2_q$$

Testing the Simple Reduced Model

Let's test our "naive model" from Lecture 1 (always predicting $\hat{y} = \bar{y}$) versus having a model having many predictors in a linear model.

$$
\begin{aligned}
LR &= \frac{\max\limits_{\beta_0,\beta_1,\ldots,\beta_p,\sigma^2} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}{\max\limits_{\beta_0,\sigma^2} \mathcal{L}\left(\beta_0, \beta_1 = 0, \ldots, \beta_p = 0; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)} \\
&= \frac{\left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\right)^n \exp\left(-\frac{1}{2\hat{\sigma}^2} SSE\right)}{\left(\frac{1}{\sqrt{2\pi\hat{\sigma}_0^2}}\right)^n \exp\left(-\frac{1}{2\hat{\sigma}_0^2} SSE_0\right)} \\
&= \left(\frac{SSE_0}{SSE}\right)^{n/2} \underbrace{\frac{\exp\left(-\frac{n}{2SSE} SSE\right)}{\exp\left(-\frac{n}{2SSE_0} SSE_0\right)}}_{1}
\end{aligned}
$$

Testing the Simple Reduced Model

Now we build the Q statistic:

$$Q = 2\ln\left(\left(\frac{SSE_0}{SSE}\right)^{n/2}\right) = n\ln\left(\frac{SSE_0}{SSE}\right) \xrightarrow{\mathcal{D}} \chi^2_p$$

This can be used to test

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_p = 0$$
$$H_a : \text{at least one is non-zero}$$

There is another test for this you've learned about?

Omnibus F-test

$$F = \frac{\frac{SSE_0 - SSE}{p}}{\frac{SSE}{n-p}} = \frac{SSE_0 - SSE}{SSE}\frac{n-p}{p} = \left(\frac{SSE_0}{SSE} - 1\right)\frac{n-p}{p} \sim F_{p,n-p}$$

Both tests use the same test statistic, namely SSE_0/SSE (up to constants and a monotonic transformation). It is a harder proof to demonstrate they have the same power for the same n and α (but they do).

Some points

- The likelihood ratio test / F test can also test any subset of the predictors (even one).
- Thus, we now have inference for every predictor or subset of predictors i.e.
    - Hypothesis testing
    - Confidence intervals

What does inference buy you?

Previously,

$$Y \sim g(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma^2, \ldots)$$

Do not assume OLS assumptions. We picked L2 loss and minimized to get $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. What do these numbers means?

$$Y \overset{ind}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \, \sigma^2\right)$$

Assume OLS assumptions. Using MLE, we wind up minimizing L2 loss and get the same $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. What do these numbers means? Same thing, except now ... we can "test" each value and provide confidence intervals for each value. You know how "stable" each number is to the the onslaught of the noise.

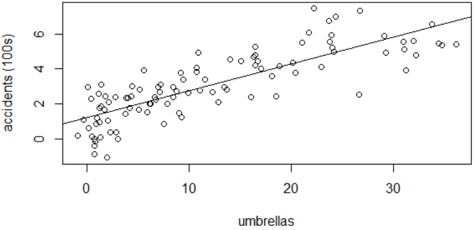
What you want to say about $\hat{\beta}_j$

[Interpret stolen bases in baseball dataset in JMP].

A change in x_j of $+1$ (a unit increase) causes / induces a β_j difference in its mean response y. Correct?

Umbrella Sales and Car Accidents

Consider a simple example. x : umbrella sales and y : car accidents. What would the relationship look like?



Does 100 more umbrellas sold *cause* 15.3 more car accidents (on average)? No... only an association (assessed by a linear correlation).

Correlation Does Not Imply Causation

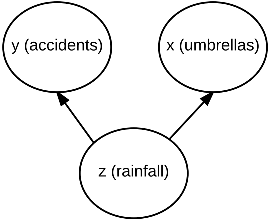
What can correlation mean?

1. There's a coincidence. How can this be?
2. They are consequence from of a common cause (the **lurking** or **counfounding** variable). How can this be?
3. There is causation
    1. x causes y (possibly with intermediates)
    2. y causes x (possibly with intermediates)
    3. x and y cause each other (cyclic)

    (recall time-boundedness property)

Controlling for the Confounder

The confounding variable is likely $z =$ rainfall.



The illustration shows that if you change x obviously y doesn't change whatsoever (causes always precede their dependent effects an assumption known as temporal boundedness)

[Show regression in R]

A Proper Interpretation of $\hat{\beta}_j$

Consider $\hat{\beta}_j$ estimates β_j. Imagine n is large and the confidence interval is really small. So basically, $\hat{\beta}_j = \beta_j \neq 0$. Interpretation?

Another object naturally observed with exactly the same features except that x_j is increased by 1 unit will have a β_j difference in its mean response y.

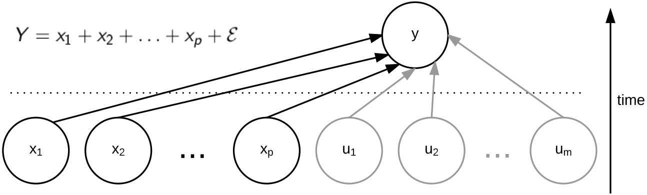
Now, the more realistic situation: $\hat{\beta}_j$ estimates β_j. Imagine n is not so large and the confidence interval is not small but we are still convinced $\beta_j \neq 0$. Interpretation?

Another object naturally observed with exactly the same features except that x_j is increased by 1 unit will have a $\hat{\beta}_j \pm \mathbb{S}\mathrm{E}\left[\hat{\beta}_j\right]$ difference in its mean response y. (Not much difference except accounting for model estimation error).

When can you say "causes"?

When can the interpretation be "causal" as follows? ~~Another object naturally observed with exactly the same features except for a change~~ **If this object in front of us has its** x_j changed by $+1$, it will ~~have~~ **cause** a $\hat{\beta}_j \pm \mathbb{S}\mathrm{E}\left[\hat{\beta}_j\right]$ difference in its mean response y.

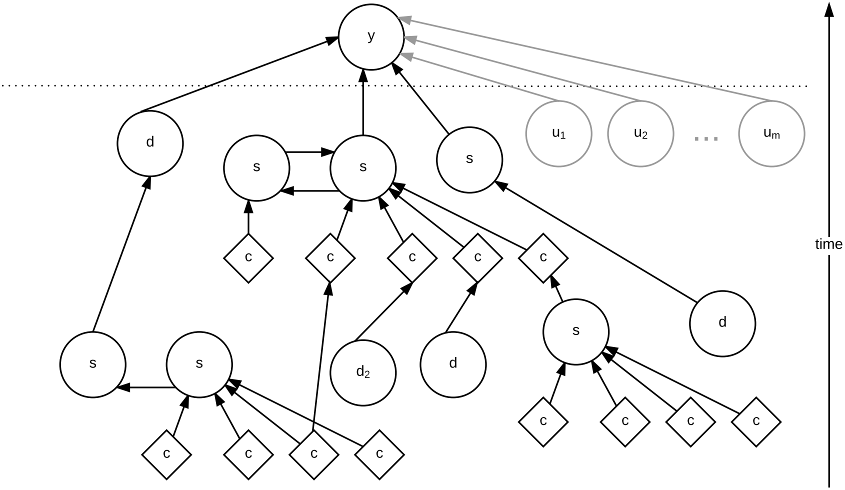
1. If we can just assume the model looks as follows:



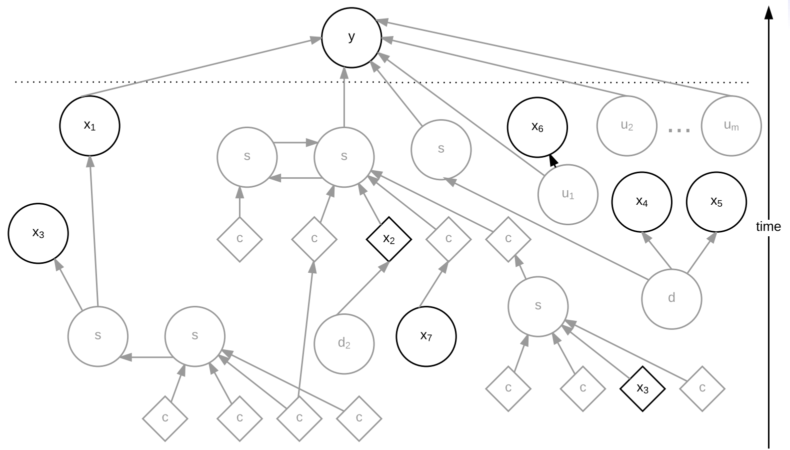
   (causal for all p features ... how can the illustration be updated for one variable?)

2. –OR– If we've run a randomized experiment manipulating x_j among the objects AND assuming an linear additive effect of x_j on y.

Consider a Realistic Model



Consider Realistic Predictors



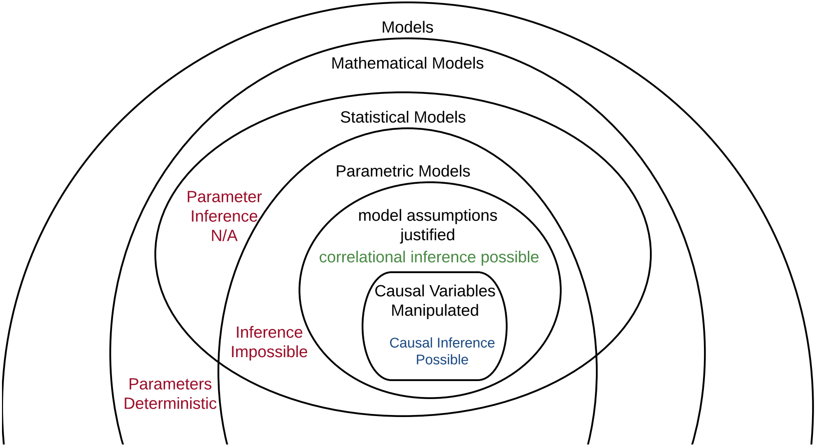
Grey variables and known to be dependent but the values are unknown and the u_k's are the "unknown unknowns".

Consider Realistic Predictors

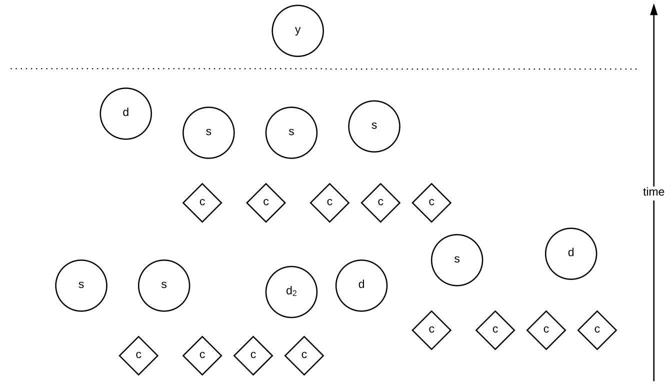
Some observations from the previous illustration:

- Maybe some of the predictors $x_1, \ldots, x_p$ are causal, but most are likely not.
- Of the ones that are not causal due to a confounder, you may have an idea of the lurking variables but it is unlikely you can measure them. Think college GPA vs SAT with confounder true IQ / ability.
- If some variables are causal, it is unlikely they have an additive causal effect; their effect is likely moderated by many other interacting variables possibly in non-linear ways.
- A linear model for y on $x_1, \ldots, x_p$ is likely far from the truth (not related to our discussion on causality).

Inference and Causality



Sidebar: Theories are Hard...

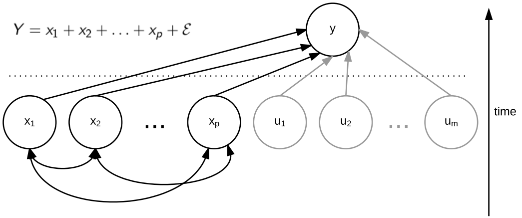


Maybe we know the predictors, but don't know the causal dependencies. How many theories are possible? 23 variables, 4 configurations between each pair, 23 possible dependences to the the response ... $= 4^{\binom{23}{2}} \times 2^{23} = 1.757 \times 10^{159}$. And that's not even counting the unknown unknowns... thus, many have said that generally speaking "science is impossible" .

More on OLS Coefficient Interpretation

The linear regression coefficient interpretation again: another object **naturally observed** with exactly the same features except that x_j is increased by 1 unit will have a $\hat{\beta}_j \pm \mathbb{SE}\left[\hat{\beta}_j\right]$ difference in its mean response y.

What do we mean by naturally observed? This other object is realized from the same joint distribution as all other observations. This means that whatever *multicollinearity* / *covariance structure* exists between the predictors, $\{\mathbb{C}\text{ov}\left[X_j, X_k\right]\}$, will give rise to the predictor values in the other object.



The Hidden “Fifth” OLS Assumption

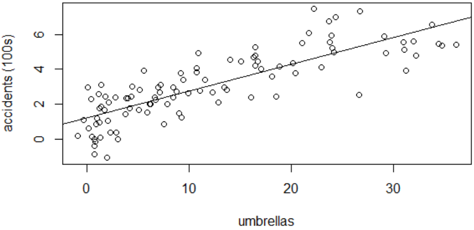
So this language “... exactly the same features except that x_j is increased by ...” is kind of absurd in the context of a strong covariance structure as ... i.e. it will be very rare to observe an observation with x_j different without any other predictor values different. Example from baseball dataset?

There is room to argue that to have these interpretations be at all realistic, we must assume there is not ... a strong multicollinearity structure between x_j and the other predictors.

But isn't getting the adjustments the whole reason we do linear regression??

But Real Correlations Still Rock

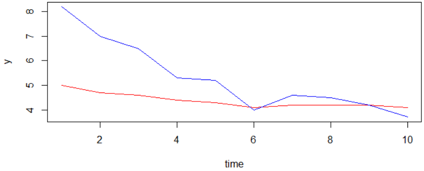
We've been beating up on correlations and their interpretations e.g. the following:



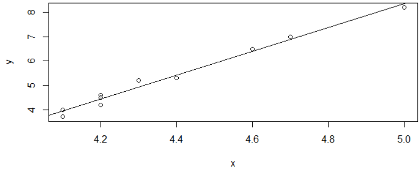
But even though higher umbrella sales do not "cause" accidents, can they still predict them? Yes, R^2 is totally agnostic to (a) if your model is true and (b) if your variables are causal or not. Predictors *truly* correlated (a causal link exists) to the response contain information about the value of the response and it doesn't matter through what channel it provides that information.

Fake / Spurious Correlations

x is margarine consumption per capita in America measured yearly for 10 years from 2000-2009, y is the divorce rate in Maine per 1000 people measured yearly for 10 years from 2000-2009



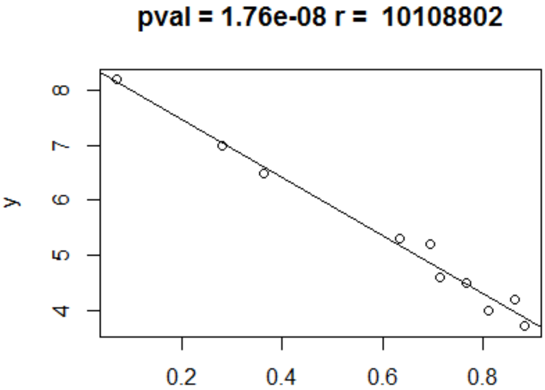
Are they linearly predictive of one another?



$R^2 \approx 99\%$, F test $p_{\text{val}} \approx 1 \times 10^{-8}$. [R demo]

Data dredging / mining / p-hacking is a dangerous enterprise

Be careful about featurization... try to at least have some inkling of an idea for a causal dependency for the response on the predictors... I "found" this using by running that demo code for a few hours...



Unintential Dredging

[JMP Baseball data] Consider all these t-tests. Is it possible some are true because I've dredged by testing all of them? Of course.

When is an individual t-test / F-test / LR test valid? When you are looking to test one single theory. Imagine you wished to test $H_a : \beta_{\text{num\_RBIs}} \neq 0$. Here's all you "see" then:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	223.11467			
batting_average	3043.1916			
on_base_pct	-3528.013			
num_runs	7.1003087			
num_hits	-2.69827			
num_doubles	1.3683081			
num_triples	-17.92163			
num_home_runs	19.483221			
num_rbi	17.415071	5.068232	3.44	0.0007*
num_walks	5.8147456			
num_str_outs	-9.585548			
num_stolen_bases	13.043869			
num_errors	-9.553269			
indic_free_agency	1372.8859			
indic_free_agent_1991	-280.7903			
indic_arb_elig	783.59223			
indic_arb_1991	352.11393			

Sidak Correction

If you "see" all of the parameter tests, what theory were you testing? Every single one... it is as if you were "looking for variables" that matter (this is called **variable selection** and we will be doing this later on in the course). But again, that $\alpha := \mathbb{P}(\text{Type I error})$ gets you... the false positives. What can we do to curb this? Do a "multiple testing correction".

Imagine K tests. The strictest correction will be when considering that they fail to reject H_0 implying that their p values are $\sim \mathrm{U}(0, 1)$. A rejection occurs at $p < \alpha$ which has probability α. This is called the false positive rate / Type I error.

We want to control the "family-wise error rate" (FWER) meaning the probability of one or more Type I errors is $\leq \alpha_{FWER}$.

$$\begin{aligned}\alpha_{FWER} &:= \mathbb{P}(\geq 1 \text{ rejection}) = 1 - \mathbb{P}(0 \text{ rejections}) = 1 - \binom{K}{0}\alpha^0 (1-\alpha)^K \\ &= 1 - (1-\alpha)^K \quad \text{(AKA the Sidak Correction)}\end{aligned}$$

Bonferroni Correction

What does the Sidak Correction assume among the K tests? Independence. Can we assume that here? No. There is multicollinearity which means $\mathbb{C}\text{ov}\left[\hat{\beta}_i, \hat{\beta}_j\right] \neq 0$. What can we do now? Call event R a rejection. Recall inclusion exclusion:

$$\begin{aligned} \mathbb{P}(R_1 \cup R_2) &= \mathbb{P}(R_1) + \mathbb{P}(R_2) - \mathbb{P}(R_1 \cap R_2) \\ \alpha_{FWER} &= \alpha + \alpha - \boxed{?} \end{aligned}$$

which can be used to demonstrate Boole's Inequality:

$$\begin{aligned} \mathbb{P}\left(\bigcup_{k=1}^{K} R_k\right) &\leq \sum_{k=1}^{K} \mathbb{P}(R_k) \\ \alpha_{FWER} &\leq \sum_{k=1}^{K} \alpha = K\alpha \end{aligned}$$

Meaning if I want the typical $\alpha_{FWER} = 5\%$, I'd better set the individual rejection at $5\%/K$. This is known as the Bonferroni Correction.

Scheffe Correction

The Bonferroni Correction is extremely conservative here. Why? Because in OLS, we know the dependence structure. We can somewhat figure out the $\mathbb{P}(R_1 \cap R_2)$ terms above. One solution from the 1950's is called Scheffe's Method:

$$\mathbb{P}\left(\frac{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\top} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{pMSE} \leq F_{\alpha,p,n-p}\right) = 1 - \alpha$$

This also account for every possible contrast you'd ever want to test e.g. $H_a : \beta_3 + \beta_7 \neq \beta_5 - \beta_2$.

I can't figure out how to do this in JMP, so if it is on the homework, we will do it in R.

Omnibus F test as a "Correction"

Recall:

$$R^2 = \frac{SSE_0 - SSE}{SSE_0} = \ldots = 1 - \left(1 + F\frac{p-1}{n-p}\right)^{-1}$$

$$\begin{aligned} F &= \frac{\frac{SSE_0 - SSE}{p}}{\frac{SSE}{n-p}} = \frac{SSE_0 - SSE}{SSE}\frac{n-p}{p} = \ldots \\ &= \underbrace{\frac{R^2}{1-R^2}}_{\text{ratio of variance explained to unexplained}} \times \underbrace{\frac{n-p}{p}}_{\text{penalty for too many features}} \end{aligned}$$

[R Demo]

Hypothesis Testing: a Review

Conceptually, let's act out the introduction of data assuming H_0, H_a and some predetermined level α.

$$H_0 : \text{UFOs do not exist}$$
$$H_a : \text{UFOs do exist}$$

and the inverse:

$$H_0 : \text{UFOs do exist}$$
$$H_a : \text{UFOs do not exist}$$

"Flipping" the null and the research hypothesis represents a completely different framing. The Type II error is now controlled for.

Hypothesis Testing: a Review

For regression, we can consider the same:

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

and the inverse:

$$H_0 : \beta_j \neq 0$$
$$H_a : \beta_j = 0$$

[R Demo]

An Equivalence Test

We are trying to prove $\beta_j = 0$ so we first assume $\beta_j \neq 0$ and wait until we have enough evidence (an "equivalence test"). Can you think of a situation you would need this type of control?

We first define δ, a margin of practical equivalence, so if $\beta \in [-\delta, \delta]$ than practically speaking we believe it to be zero. You need to set δ yourself. Then we run two tests at level α:

$$H_0 : \beta_j \geq \delta \qquad\qquad H_0 : \beta_j \leq -\delta$$
$$H_a : \beta_j < \delta \qquad\qquad H_a : \beta_j > -\delta$$

This is known as TOST (two one sided tests) which is equivalent to taking the intersection of two α-sized one sided confidence intervals, i.e. a two sided confidence interval at level 2α. Thus, we reject H_0 if:

$$CI_{\beta_j, 1-\alpha} := \left[\hat{\beta}_j \pm t_{\alpha, n-p-1} \mathbb{S}\mathrm{E}\left[\hat{\beta}_j\right]\right] \in [-\delta, \delta]$$

[R demo]

Dataframe Design

We spoke a lot about featurization i.e. selecting the columns in the dataframe (these are the predictors to measure). Once we did this, we can then go out and sample observations and then measure each for their predictor values.

But we didn't speak at all about selecting the observations themselves (assuming you have some modicum of control of selecting your data). Two things to consider:

1. **Generalizability** refers to the ability of the model to generalize, or be **externally valid** when considering new observations. This comes down to sampling observations from the same population as your new data you wish to predict (pretty obvious). Sometimes difficult in practice!
2. Optimal Design

Optimal Design for Inferring one Slope

Question: assume OLS and that we only care about inference for β_1. We can sample any x values $\in [x_m, x_M]$. What should the n values be?

Let $x_m = 0$, $x_M = 1$ and $n = 10$. The best inference for β_1 means ... $\mathbb{S}\mathrm{E}\left[\hat{\beta}_1\right]$ is minimum. Design strategies for the x's:

1. Random sampling
2. Uniform spacing: $\{0, 0.111, 0.222, \ldots, 0.999\}$
3. Something else?

[R demo]

Optimal Design: Split Between Extremes

Recall the formula from Stat 102 / 613:

$$\mathbb{S}\mathrm{E}\left[\hat{\beta}_1\right] = \sqrt{\frac{MSE}{(n-1)s_x^2}}$$

How can we make this small?

1. Maximize n (duh)
2. Minimize the numerator, MSE i.e. minimize the SSE. Can we do this? Yes by picking the closest $\hat{\beta}_1$ to β_1 (which we already do).
3. Maximize the denominator $(n-1)s_x^2$. Since n is already maximized, we can pick $x_1, \ldots, x_n$ to maximize s_x^2, the sample variance of the predictor. How? Put half of the x's at x_m and the other half at x_M thereby maximizing the distance from the x's to $\bar{x}$.

Optimal Design of Linear Models

We seek the best linear approximation of $f(x)$ which is $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. We pick the $\boldsymbol{x}$'s to give us the best linear approximation. What criteria? JMP gives two ways:

1. Note: $\mathbb{V}\text{ar}\left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right] = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$

   D-optimality: maximize $\left|\boldsymbol{X}^T \boldsymbol{X}\right|$ — this maximizes the variance-covariance among the parameter estimates.

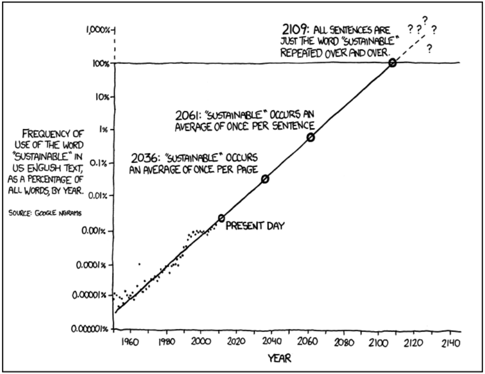
2. Note: $\mathbb{V}\text{ar}\left[\hat{Y}_1, \ldots, \hat{Y}_n\right] = \sigma^2 \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T$

   I-optimality: minimize the average prediction variance over the design space.

[R Demo] What did we learn? For linear models with no polynomials or interactions, keep the observations as close to the minimimums and maximums as possible. For linear models with polynomials and interactions (more non-parametric than parametric), keep most towards the minimums and maximums and some in the center of the input space.

Extrapolation

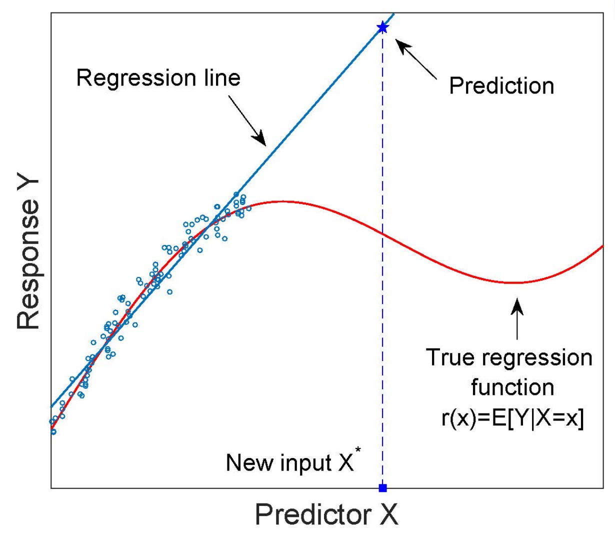
Data driven approaches are all focused on accuracy during **interpolation**.



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

Extrapolation brings trouble. It is important to ask the question for a new observation $\boldsymbol{x}^*$ if it is within the space of $\boldsymbol{x}$'s in the historical data. (Hardly anyone does this... but you should)! Be aware that extrapolation methods of different algorithms differ considerably! [R Demo]

Reconciliation of those Silly Cartoons



Modeling Categorical Responses

Previously the response y was continuous and via the OLS assumptions we obtained the statistical model,

$$Y \overset{ind}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\, \sigma^2\right)$$

If the response y is categorical, can we still use this? No... the only elements in the support of the r.v. Y are the levels only. [JMP Churn]

First, assume Y is binary i.e. zero or one. The model we use is...

$$Y \sim \text{Bernoulli}\left(f(x_1, \ldots, x_p)\right)$$

since $\mathbb{E}\left[Y \mid x_1, \ldots, x_p\right] = f(x_1, \ldots, x_p)$, then f is still the conditional expectation function like before except now it varies only within $[0, 1]$ and it is the same as $\mathbb{P}\left(Y = 1 \mid x_1, \ldots, x_p\right)$.

Linear $f(x)$?

We can model $f(x)$ as the simple linear function but this returns values smaller than 0 and larger than 1 and thus it cannot be the conditional expectation function! Why? Lines vary between $(-\infty, +\infty)$.

We need a "link function" to connect the linear function to the restricted support of the response:

$$\lambda(f_{\mathbb{R}}(x_1, \ldots, x_p)) = f(x_1, \ldots, x_p)$$

And the parametric assumption would be

$$\lambda(s_{\mathbb{R}}(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell)) = s(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell)$$

And assuming a linear form:

$$\lambda(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) = ?$$

Choice of λ?

We just need $\lambda : \mathbb{R} \to [0, 1]$. There are infinite λ's to choose from. I've only seen three used:

1. Logistic link: $\lambda(w) = \frac{e^w}{1+e^w}$ (most common)
2. Inverse normal (probit) link: $\lambda(w) = \Phi^{-1}(w)$ where Φ is the normal CDF function (somewhat common)
3. Complementary Log-log (cloglog) link: $\lambda(w) = \ln(-\ln(w))$ (rare!)

Let's investigate what the first one means. Define $p := \mathbb{P}(Y = 1)$. We can think about probability in another way:

$$odds(Y = 1) := \frac{p}{1-p}$$

So if odds = 4:1, what is p? This means that the probability of the event happening is four times more likely than the complement happening. Or... of 4+1 runs, 4 will be a yes. What is the range of odds? $[0, \infty)$.

Why Logistic Link is Interpretable

Now let's take the log odds (called the logit function):

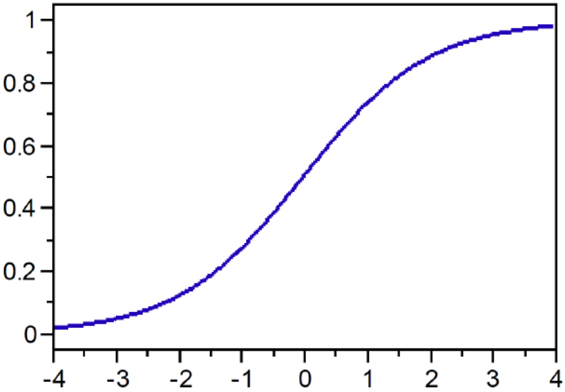
$$logit(Y = 1) := \ln\left(odds(Y = 1)\right) = \ln\left(\frac{p}{1-p}\right)$$

What is the range of the logit function? All of $\mathbb{R}$. Hence, we can now set this equal to our $s_{\mathbb{R}}$ function. In the linear modeling context,

$$\begin{aligned}
\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p &= logit(Y = 1) = \ln\left(\frac{p}{1-p}\right) \\
e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} &= \frac{p}{1-p} \\
(1-p)e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} &= p \\
e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} &= p + pe^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} \\
p &= \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}} = \lambda(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)
\end{aligned}$$

Thus, a change in the linear model becomes a linear change in log-odds. This is (I would say) the most interpretable link function situation we've got.

The Logistic Function



How to Obtain a Model Fit

A model fit would mean we estimate $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. We initially did this estimation for regression (continuous y) by defining a loss function, SSE, and finding the optimal solution via calculus. What do we do now??

Likelihood to the rescue. First the "logistic regression assumptions"

1. Linear-Logistic conditional expectation
2. Independence

$$
\begin{aligned}
& \mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right) \\
= & \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i \mid \boldsymbol{X}_1 = \boldsymbol{x}_i\right)
\end{aligned}
$$

How?

Maximum Likelihood Estimates

$$= \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

How?

$$= \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}} \right)^{y_i} \left(1 - \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}} \right)^{1-y_i}$$

How? This does not have a simple, closed form solution. The computer iterates numerically and converges on the values of the parameters that maximize the above and these are shipped to you as $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. This is called "running a logistic regression".

Prediction with Logistic Regression

How?

$$\hat{p} = \hat{p}(x_1^*, \ldots, x_p^*) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p}}$$