Predictive Analytics Lecture 3

Adam Kapelner

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

January 31 & February 1, 2017

The Coin Example from Last Class I

I want to explain the coin example from last class in the context of likelihood. Imagine you flip a coin three times and get heads, heads, tails; thus, $y_1 = 1, y_2 = 1, y_3 = 0$. There is a true probability of heads called θ. We don't know it.

What is the probability of the data? We employ the mass / density function:

$$\begin{aligned} \mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3; \theta\right) = \prod_{i=1}^{3} \mathbb{P}\left(Y_i = y_i; \theta\right) = \prod_{i=1}^{3} \theta^{y_i} \left(1-\theta\right)^{1-y_i} \\ = \quad \left(\theta^{(1)} \left(1-\theta\right)^{1-(1)}\right) \left(\theta^{(1)} \left(1-\theta\right)^{1-(1)}\right) \left(\theta^{(0)} \left(1-\theta\right)^{1-(0)}\right) \\ = \quad \theta^2 (1-\theta) \end{aligned}$$

And now we can calculate the probability of seeing the data assuming θ. Assume $\theta = 0.5$ then,

$$\mathbb{P}\left(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta = 0.5\right) = 0.5^2(1-0.5) = 0.125$$

The Coin Example from Last Class II

Now we ask the inverse question. If we saw this data $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, what is the most likely model, i.e. the most likely value of θ. We first write down the likelihood function which it's easy because it's the same as the mass / density function

$$\mathcal{L}(\theta; Y_1 = 1, Y_2 = 1, Y_3 = 0) = \mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 0; \theta) = \theta^2(1-\theta)$$

And now we pick the value of θ which maximizes the likelihood,

$$\hat{\theta} := \underset{\theta \in \Theta}{\arg\max} \; \{\mathcal{L}(\theta; x)\}$$

So we need to take the derivative

$$\frac{d}{d\theta}\left[\theta^2(1-\theta)\right] = \frac{d}{d\theta}\left[\theta^2 - \theta^3\right] = 2\theta - 3\theta^2$$

and set it equal to zero:

$$0 = 2\theta - 3\theta^2 = \theta(2 - 3\theta) \Rightarrow 0 = 2 - 3\theta \Rightarrow \hat{\theta}_{\mathrm{MLE}} = \frac{2}{3}$$

i.e. the most likely model for this data is a weighted coin with probability of heads of 2/3.

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 live in their set $\mathbb{X}$ e.g. $\in [x_m, x_M]$. What should the n values be?

Let $x_m = 0$, $x_M = 1$ and $n = 10$. The best inference for β_1 means ... $\mathbb{S}\text{E}\left[\hat{\beta}_1\right]$ 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:

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

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

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

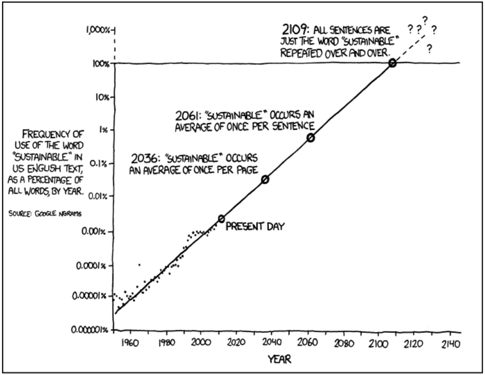
2. Note: $\mathbb{V}\mathrm{ar}\left[\hat{Y}_1, \ldots, \hat{Y}_n\right] = \sigma^2 \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$

   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

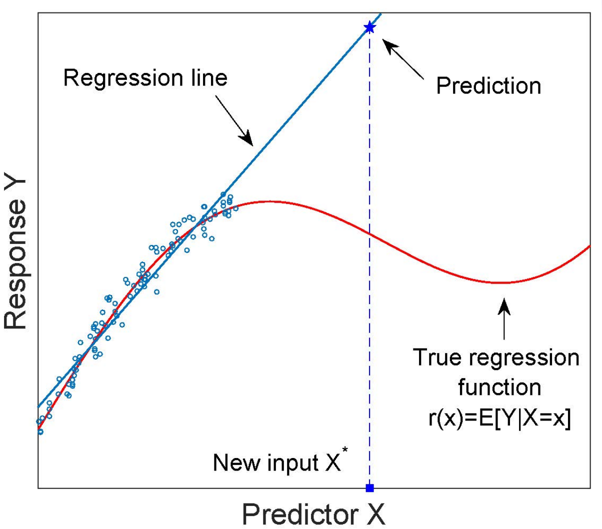
Data driven approaches are all focused on accuracy during **interpolation**.



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...

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

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):

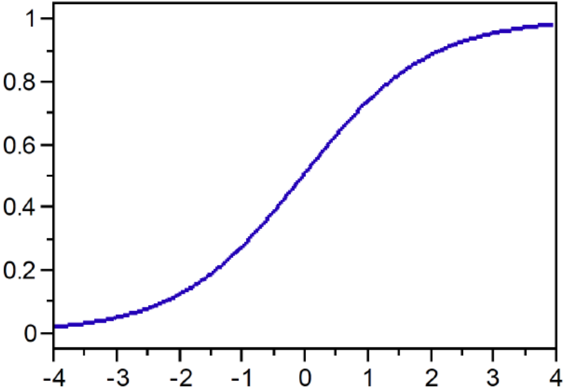
$$logit(Y=1) := \ln\left(odds(Y=1)\right) = \ln\left(\frac{p}{1-p}\right)$$

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?

$$\mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)$$

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

How? This does not have a simple, closed form solution. The computer iterates numerically usually using the log of above, since it's (1) numerically more stable and (2) the expression is easier to work with. When it "converges" on the values of the parameters that maximize the above, 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". This usually is instant on a modern computer.

Prediction with Logistic Regression

How?

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

Note the predictions are for the conditional expectation function, the probability itself. However, you may actually wish to predict the response, the 1 or the 0. What to do?

You can create a **classification rule** which allows you to make a decision about the response based on the probability. What is the most intuitive classification rule?

$$\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq 0.5 \\ 0 & \text{if} \quad \hat{p} < 0.5 \end{cases}$$

AKA the "most likely criterion". We will return to prediction and evaluation of predictive performance later but first...

Global Test in Logistic Regression

Recall in OLS regression, probability theory directly gave us t-tests and F-tests. Under the logistic regression assumptions, **we have no such analogous theory**! However, we can make use of the ... likelihood ratio test. Recall:

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

Let's now do a "whole model" / "global" / "omnibus" test:

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

So Θ would be the space of all $\beta_0, \beta_1, \ldots, \beta_p$ and Θ_R will restrict the space to only β_0 with zeroes for all other "slope" parameters.

$$LR = \frac{\max\limits_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}{\max\limits_{\beta_0} \mathcal{L}(\beta_0, \beta_1 = 0, \ldots, \beta_p = 0; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}$$

So on top the computer iterates to find $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$, plugs it in and computes the likelihood and on the bottom the computer independently iterates to find $\left\{\hat{\beta}_0\right\}$, plugs it in and computes the likelihood, then together, the LR.

Partial Tests in Logistic Regression

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped p parameters / degrees of freedom, we look at the critical $\chi^2_{p,\alpha}$ value.

Let's say we want to test something like:

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

We can again use the likelihood ratio test:

$$LR = \frac{\max\limits_{\beta_0,\beta_1,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}{\max\limits_{\beta_0,\beta_3,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1 = 0, \beta_2 = 0, \beta_3, \ldots, \beta_p = 0; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}$$

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 2 parameters / degrees of freedom, we look at the critical $\chi^2_{2,\alpha}$ value.

Individual Tests in Logistic Regression

Let's say we want to test an individual slope coefficient:

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

We can again use the likelihood ratio test:

$$LR = \frac{\max\limits_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}{\max\limits_{\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_j = 0, \beta_{j+1}, \ldots \beta_p; y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)}$$

We then look at $Q = 2\ln\left(LR\right)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 2 parameters / degrees of freedom, we look at the critical $\chi^2_{1,\alpha}$ value.

There is something special about a χ^2 r.v. with one degree of freedom. Cool fact from basic probability: $\sqrt{\chi^2_1} = Z \sim \mathcal{N}\left(0, 1\right)$. This is how JMP produces standard errors for logistic regression coefficients.

Telecom Churn Example

In marketing “churn” refers to a customer canceling their service. Studies suggest that it costs 5-10x more to acquire a new customer than to retain an old customer. Thus, predicting churn is of major interest!

Here’s a dataset from a telecom company (likely it’s churn on Verizon / AT&T / T-Mobile /Sprint’s cell-phone plan). We have 7,043 customers with 20 features. This is likely a nearly-mindless dump!! Churn is defined to be a complete cancellation of services in the next month period. Since we are predicting churn, define $y = 1$ to be churn, so the $\hat{p}$’s are estimates of probability of churning (this is just convenient).