Response Space	Type of Modeling	g return	Example alg
$\mathcal{Y} \subseteq \mathbb{R}$	regression	$\hat{y} \in \mathcal{Y}$	OLS
$\mathcal{Y} = \{c_1, c_2, \ldots, c_K\}$	classification	$\hat{y} \in \mathcal{Y}$	KNN
if $K=2$, $\mathcal{Y} = \{\underset{c_1}{0}, \underset{c_2}{1}\}$	binary classification	$\hat{y} \in \mathcal{Y}$	SVM
$\mathcal{Y} \subseteq \mathbb{R}_{\geq 0}$	survival	$\hat{y} \in \mathcal{Y}$	Weibull regression
$\mathcal{Y} \in \{0, 1, 2, \ldots\}$	count	$\hat{y} \in \mathcal{Y}$	Poisson regression
$\mathcal{Y} \in (0,1)$	proportion	$\hat{y} \in \mathcal{Y}$	Beta regression
$\mathcal{Y} = \{c_1, c_2, \ldots, c_K\}$	probability estimation	$\vec{\hat{p}} := \begin{matrix} P(Y=c_1 \mid \vec{x}) \\ P(Y=c_2 \mid \vec{x}) \\ \vdots \\ P(Y=c_K \mid \vec{x}) \end{matrix}$	Multilogit Regression
$K=2$, $\mathcal{Y} = \{\underset{c_1}{0}, \underset{c_2}{1}\}$	probability estimation	$\hat{p} := P(Y=1 \mid \vec{x})$	Logistic Regression
$\mathcal{Y} \in \{c_1, \ldots, c_K\}$ ordinal	probability estimation		Proportional odds model

If $\mathcal{Y} = \{0, 1\}$ for all i,

$$\begin{aligned} y &= t(\vec{z}) \\ &= f(\vec{x}) + \delta \quad \text{where } \delta \in \{0, -1, +1\} \\ &= h^*(\vec{x}) + \varepsilon \quad \text{where } \varepsilon \in \{0, -1, +1\} \\ &= g(\vec{x}) + e \quad \text{where } e \in \{0, -1, +1\} \end{aligned}$$

e

		$g(\vec{x})$	
		0	1
Y	0	0	-1
	1	+1	0

false positive

false negative

How do we build a probability estimation model?

Naively, $g_0 = \bar{y}$

$\iff Y \sim \text{Bern}(t(\vec{z}))$

We now view Y as a realization from a random varible (bernoulli).
We will assue there exists a function $f_{pr}(\vec{x}) : \mathbb{R}^{p+1} \to (0,1)$
and this function is the best guess
of the probability P(Y=1|xvec) you can create with xvec.

$$\to Y \sim \text{Bern}(f_{pr}(\vec{x}) + \underbrace{t(\vec{z}) - f_{pr}(\vec{x})}_{\delta_{pr}})$$

$\Rightarrow Y \sim \text{Bern}(f_{pr}(\vec{x}))$. f_{pr} is the model we want to find.

Let's assume that all the data (all the n observations) in D are independently realized.

$$\begin{aligned} P(\mathbb{D}) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \vec{x}_1, \ldots, \vec{x}_n) \\ &\stackrel{\perp}{=} \prod_{i=1}^{n} P(Y_i = y_i \mid \vec{x}_i) \\ &= \prod_{i=1}^{n} f_{pr}(\vec{x}_i)^{y_i} (1 - f_{pr}(\vec{x}_i))^{1-y_i} \end{aligned}$$

241 Review: $V \sim \text{Bern}(\theta) = \theta^{v}(1-\theta)^{1-v}$

Now we want to "fit" f_pr using our data (learning from data paradigm). How? Is this even possible? NO. We cannot fit arbitrary functions in any dimension. We need a set of candidate functions that we can fit. Call that $\mathcal{H}_{pr}$, Each element in this set maps $\mathbb{R}^{p+1} \to (0,1)$. How about:

$$\mathcal{H}_{pr} = \{\vec{w} \cdot \vec{x} : \vec{w} \in \mathbb{R}^{p+1}\} \;?$$

We can't use this since it returns values outside (0,1), the space of legal probabilities. But... we really like wvec dot xvec because (1) easy to interpret and we have lots of intuition about it from all of our previous modeling we've done and (2) monotonic in each of the x_j's. How we do we have our cake and eat it too?

We need a function that takes wvec dot xvec and maps it into the space (0,1) i.e. $\phi : \mathbb{R} \to (0,1)$ which is called a "link function"
I think because it links the two spaces (the reals and the prob's).
We restrict the link function to be strictly increasing. Thus,

$$\mathcal{H}_{pr} = \{\phi(\vec{w} \cdot \vec{x}) : \vec{w} \in \mathbb{R}^{p+1}\}$$

These types of models are called "generalized linear models" (glm) because they retain wvec dot xvec (the linear model) but then manipulate it in some way. Which link function should we use? There are three common ones. In order of use:

1. Logistic / logit: $\phi(u) := \frac{e^u}{1+e^u} = \frac{1}{1+e^{-u}}$. Note: $1 - \phi(u) = \frac{1}{1+e^u}$

2. Probit: $\phi(u) := F_Z(u)$ i.e. the CDF of the std. normal.

3. Complementary Log-Log (cloglog)

$$\phi(u) = 1 - e^{-e^u} \Rightarrow 1 - \phi(u) = e^{-e^u} \Rightarrow \ln(1 - \phi(u)) = -e^u$$

$$\Rightarrow -\ln(1-\phi(u)) = e^u \Rightarrow u = \underbrace{\ln(-\ln}_{\text{log-log}}(\underbrace{1-\phi(u)}_{\text{complement}}))$$

Let's employ the logistic link function:

$$\mathcal{H} = \left\{\frac{1}{1+e^{-\vec{w}\cdot\vec{x}}} : \vec{w} \in \mathbb{R}^{p+1}\right\}$$

What is $\mathcal{A}$? How to get $g \in \mathcal{H}$?

Why not find the wvec that provides us the highest probability?

$$\mathcal{A} : \vec{b} := \underset{\vec{w} \in \mathbb{R}^{p+1}}{\text{argmax}} \left\{ \underbrace{\prod_{i=1}^{n} \left(\frac{1}{1+e^{-\vec{w}\cdot\vec{x}}}\right)^{y_i} \left(\frac{1}{1+e^{\vec{w}\cdot\vec{x}}}\right)^{1-y_i}}_{P(\mathbb{D})} \right\}$$

In OLS, we took the derivative and set it equal to zero to solve for bvec and we found an analytical solution. However,
there is no analytical solution here. You need to use a computer.

$$\vec{\nabla} P(\mathbb{D}) \stackrel{\text{set}}{=} \vec{0}_{p+1} \quad \text{and approximate}$$

Usually this is done with "gradient descent". Computing bvec is called "running a logistic regression". Once this is done...
we can predict using

$$\hat{p} = g_{pr}(\vec{x}) = \phi(\vec{b} \cdot \vec{x}) = \frac{1}{1+e^{-\vec{b}\cdot\vec{x}}} \quad \text{hopefully close to } f_{pr}(\vec{x})$$

$\hat{p} = \hat{P}(Y=1 \mid \vec{x})$

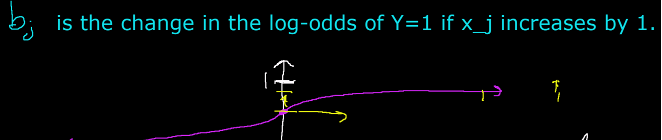
What is the interpretation of the slope coefficients (the entries in the b-vec)?

$$\hat{p} = \frac{1}{1+e^{-\vec{b}\cdot\vec{x}}} \Rightarrow \frac{1}{\hat{p}} = 1 + e^{-\vec{b}\cdot\vec{x}} \Rightarrow \frac{1}{\hat{p}} - 1 = e^{-\vec{b}\cdot\vec{x}}$$

$$\Rightarrow \frac{1-\hat{p}}{\hat{p}} = e^{-\vec{b}\cdot\vec{x}} \Rightarrow \ln\left(\frac{1-\hat{p}}{\hat{p}}\right) = -\vec{b}\cdot\vec{x} \Rightarrow \vec{b}\cdot\vec{x} = \underbrace{\ln\Big(\underbrace{\frac{\hat{p}}{1-\hat{p}}}_{\text{odds}}\Big)}_{\text{log-odds}}$$

$$\text{Odds} = \frac{\hat{p}}{1-\hat{p}}$$

$\Rightarrow b_j$ is the change in the log-odds of Y=1 if x_j increases by 1.



$$0 = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) \Rightarrow \phi(0) = \frac{1}{1+e^{-0}} = \frac{1}{1+1} = 0.5$$

log odds	prob
0	0.5
−1	$0.27 \approx \frac{1}{4}$
+1	$0.73 \approx \frac{3}{4}$
−2	$0.12 \approx \frac{1}{8}$
+2	$0.88 \approx \frac{7}{8}$
⋮	
−∞	0
+∞	1

The logistic link function is highly nonlinear.

$0 \to 1$ log odds $\iff$ 50% → 73% in prob.

$3 \to 4$ log odds $\iff$ 95% → 98% in prob.

Probability estimation models predict probabilities but we observe labels (i.e. 0 or 1). The true probabilities f_pr are unobserved! We need a metric called a "scoring rule" S that can compare a p-hat value to a y value.

A "proper scoring rule" S(p-hat, y) is one where:

$$\forall i \quad f_{pr}(\vec{x}_i) = \text{argmax}\{S(\hat{p}_i, y_i)\}$$

We will study two proper scoring rules:

1. Brier score (1950). Let $s_i := -(y_i - \hat{p}_i)^2 \leq 0$

$$\bar{s} := \frac{1}{n}\sum_{i=1}^{n} s_i \leq 0$$

2. Log scoring rule. Let $s_i := y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i) \leq 0$

$$\bar{s} = \frac{1}{n}\sum s_i \leq 0$$

These scores are used as an "R^2" of the model (but they're not between 0 and 1) in a conceptual sense. The closer to zero, the better the probability estimation model.