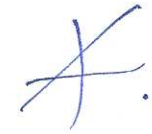
Classification / probability estimates.

want $\{0,1\}$ answer $\{0, \ldots k\}$ — want $0 \leq p \leq 1$

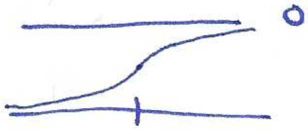
linear regression often ~~not~~ not a good fit



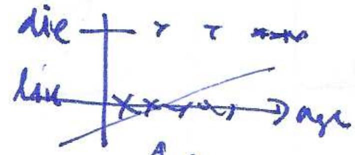
Logistic regression

Example.

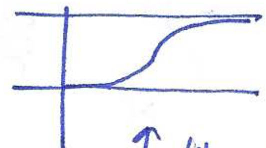
Logistic function $\frac{e^x}{1+e^x}$ $\quad 0 \leq \quad \leq 1$



0



linear model $y = mx + c$



fit a logistic function instead

$$p(x) = \frac{e^{mx+c}}{1+e^{mx+c}}.$$

if the probability of an event is p

the odds of the event are $\frac{p}{1-p}$, so $\frac{p(x)}{1-p(x)} = e^{mx+c}$.

Logistic function $\leftrightarrow$ log(odds) is linear.
logit

Estimating the coefficients $mx+c$ / $\hat{\beta}_0 + \hat{\beta}_1 x$.

instead of least squares error use <u>maximum likelihood</u>.

Likelihood function $\quad l(m,c) = \prod_{i, y_i = 1} p(x_i) \prod_{i, y_i = 0} (1 - p(x_i))$.

intuition: $l = 1$ best case, all predictions correct
$l = 0$ at least one prediction completely opposite.

<u>Fact</u>: in many cases there are fast algorithms to find m, c.

end up with $p(x) = \frac{e^{mx+c}}{1+e^{mx+c}}$ } estimates prob of event occurring

<u>Multiple logistic regression</u> (multiple dim input, one prob output).

use log odds: $\quad \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d$.

equivalent to: $p(x) = \dfrac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_d X_d}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_d X_d}}$. ②

Multinomial logistic regression (need to predict one of K things ... $\{1, \ldots K\}$).

choose a base class K.

then for $1 \le k \le K-1$ use: $$P(Y=k|X=x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kd}x_d}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{ld}x_d}}$$

for K: $$P(Y=K|X=x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{ld}x_d}}.$$

then. $$\log\left(\frac{P(Y=k|X=x)}{P(Y=K|X=x)}\right) = \beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kd}x_d.$$ ← i.e. log conditional prob are linear.

Fact doesn't matter which class you choose to be base class but coeffs β_{ij} depend on choice of base class.

Alternative softmax coding (no base class – symmetric).

$$P(Y=k|X=x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kd}x_d}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{ld}x_d}}.$$

log odds ratio is: $$\log\left(\frac{P(Y=k|X=x)}{P(Y=k'|X=x)}\right) = (\beta_{k0}-\beta_{k'0})+(\beta_{k1}-\beta_{k'1})x_1+\cdots+(\beta_{kd}-\beta_{k'd})x_d.$$

Generalised linear models · use X to predict Y ← quantitative, linear regression

← qualitative, logistic regression

space Y is a count (e.g. # of people).

Example Poisson regression

recall Poisson dist : $\mathbb{P}(X=n) = \frac{e^{-\lambda}\lambda^n}{n!}$ for $n \in \mathbb{N}_0$. $\mathbb{E}(X)=\lambda$. $\mathrm{Var}(X)=\lambda$.

$\uparrow y$

model count as a Poisson dist with parameter λ.

when $\lambda(x)$ usual choice $\lambda = e^{\beta_0+\beta_1 x_1+\dots+\beta_d x_d}$.

$\iff \log(\lambda) = \beta_0+\beta_1 x_1+\dots+\beta_d x_d$

i.e. choose $\log(\lambda)$ to be a linear function of the ~~data~~ data space.

to find β_i use maximum likelihood.

$$\ell(\beta_0,\beta_1,\dots,\beta_d) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{(y_i)!}$$

i.e. find β_i that maximise this ↑ this is the probability that you obtained that data set from the given value of λ.

Fact the β_i can be estimated efficiently

- interpretation of coefficients: β_i big means column more important: increase $x_i \mapsto x_i+1$ then $\mathbb{E}(Y)$ increases by a factor of e^{β_i}.
- Poisson model has $\lambda = \mathbb{E}(X) = \mathrm{Var}(X)$ often better fit than underlying assumption of linear regression model where var is constant.
- positive integer output, discrete, no negative values.

Generalised linear models

regression models: linear → quantities; logistic → categories; Poisson → $\mathbb{N}_0$.

common features:

- $f(X)$ predicts Y.
- model analysis usually assumes Y has
  - Gaussian dist for linear
  - Bernoulli ${}^{0}_{1}{}^{p}_{p}$ for categorical
  - Poisson for Poisson
- linear. $\mathbb{E}(Y|X) = \beta_0 + \beta_1 x_1 + \cdots \beta_d x_d$.

logistic: $\mathbb{E}(Y|X) = \mathbb{P}(Y=1|X)$.

$$= \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_d x_d}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_d x_d}} \Longleftrightarrow$$

equivalent log odds are linear

$$\log\left(\frac{\mathbb{E}(Y|X)}{1-\mathbb{E}(Y|X)}\right) = \beta_0+\beta_1 x_1+\cdots+\beta_d x_d$$

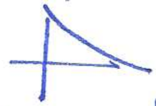
Poiss $\mathbb{E}(Y|X) = \lambda(X) = e^{\beta_0+\beta_1 x_1+\cdots+\beta_d x_d} \Longleftrightarrow \underset{\lambda(x)}{\log(\mathbb{E}(Y|X))} = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$.

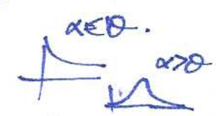
<u>notation</u> link function η transforms the expected mean so that it is linear:

linear $\eta(\mu) = \mu$ logistic $\eta(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ Poisson $\eta(\mu) = \log(\mu)$.

<u>Fact</u> these dists are all examples of exponential family dists.

other examples
- exponential dist





X_i indep exp β then $\sum X_i \sim \Gamma(n,\beta)$ $\alpha \in \mathbb{N}$ Gen Poisson.

$\Gamma(1,\beta) = \mathrm{Exp}(\beta)$.

$\Gamma(n/2, 2)$ is $\chi^2(n)$

- Gamma dist $\Gamma(\alpha,\theta)$ $f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}$ α shape, θ scale. mean $\alpha\theta$. var $\alpha\theta^2$.
- negative binomial dist

there are regression schemes for all these w/ different choice of η.

discrete analog of Γ. $k \mapsto \binom{k+r-1}{k}(1-p)^k p^r$

$NB(r,p)$ mean $\frac{r(1-p)}{p}$. var $\frac{r(1-p)}{p^2}$

Q: how many times do I need to run my Bernoulli trials before I get n wins?

$NB(r,p)$ ← prob of Bernoulli trials

# of times it occurs.