Lectures 12 and 13

An alternate path...

While our current data set concerns an observational study, some of the first applications of “regression-like” models for binomial data come from so-called bioassays

In this context, animals are exposed to varying levels of some toxic substance, and researchers want to model the probability that a given dose is lethal

In a dose-response model, the probability $\pi(x)$ that an animal dies after receiving a dose x is described by a tolerance distribution f,

$$\pi(x) = \int_{-\infty}^{x} f(s)ds$$

where $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(s)ds = 1$

Early applications

For example, if we take f(x) to be the uniform distribution over some interval, say [a,b], then we have

$$f(x) = \begin{cases} 1/(b-a) & a \leq s \leq b \\ 0 & \text{otherwise} \end{cases}$$

and so

$$\pi(x) = \int_a^s f(s)ds = \frac{x-a}{b-a} \quad \text{for} \quad a \leq x \leq b$$

This equation has the form $\pi(x) = \beta_0 + \beta_1 x$ where we need to impose constraints on the coefficients

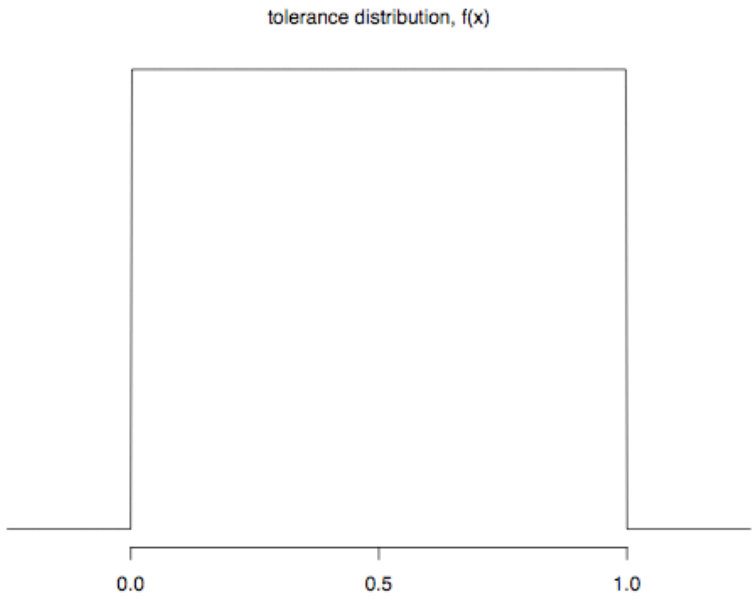
$$\beta_0 = \frac{-a}{b-a} \quad \text{and} \quad \beta_1 = \frac{1}{b-a}$$

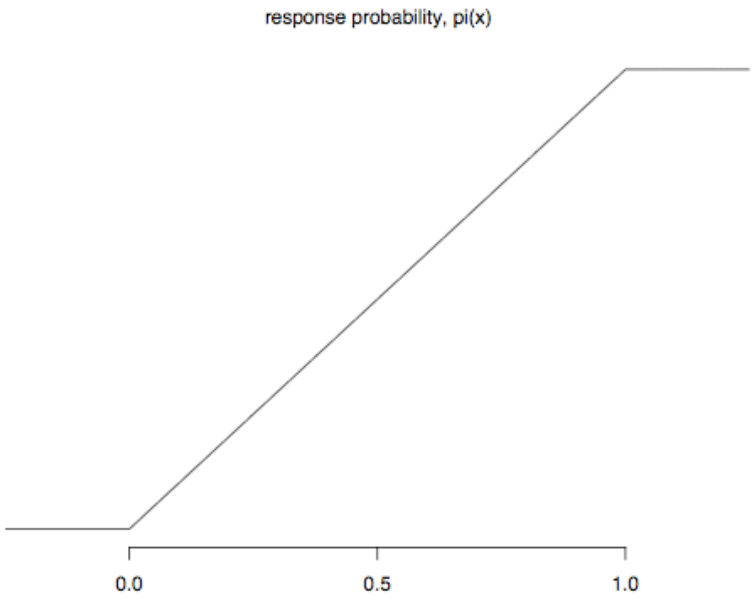
Early applications

This is a kind of linear model, describing the probability of death from a given dose -- Notice that we need to impose constraints on the coefficients to make sure that the resulting probability estimates are between 0 and 1

In terms we've described before, this model uses an identity "link" between the scale of the data (the probability $\pi(x)$) and the covariates, this time just the dose x

Given all the constraints, however, this model is rarely used...


tolerance distribution, f(x)
0.0
0.5
1.0


response probability, pi(x)
0.0
0.5
1.0

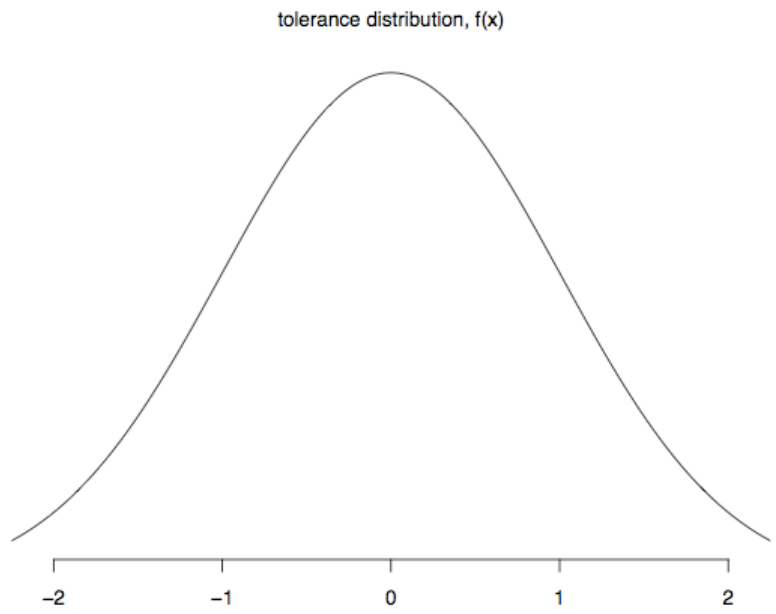
Early applications

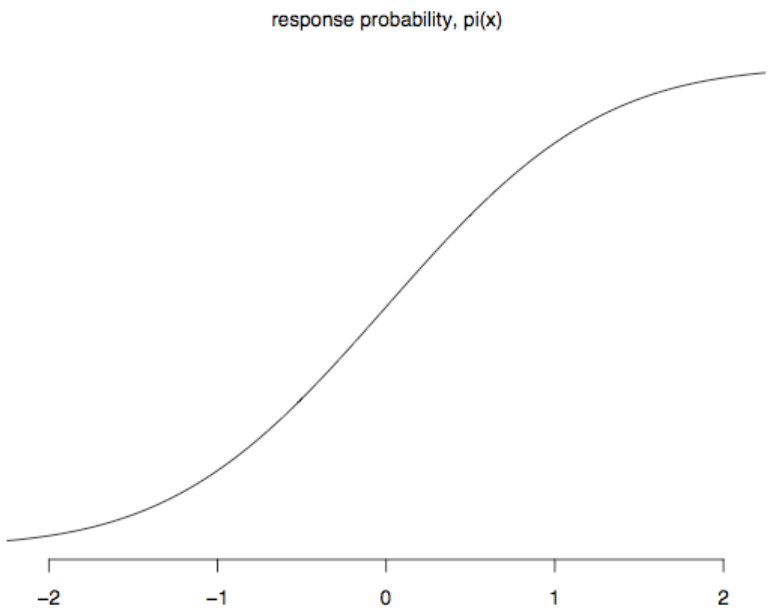
Instead, researchers started to lean on the central limit theorem; that is, if there are a large number of factors that act additively to produce a response, then a normal distribution emerges

Taking f(s) to be a normal distribution with some mean μ and standard deviation σ , we have

$$\begin{aligned}
\pi(x) &= \int_{-\infty}^{x} f(s)ds \\
&= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(s-\mu)^2/2\sigma^2} ds \\
&= \Phi\left(\frac{x-\mu}{\sigma}\right)
\end{aligned}$$

where Φ denotes the cumulative probability function for the standard normal distribution


tolerance distribution, f(x)
−2
−1
0
1
2


response probability, pi(x)
−2
−1
0
1
2

Early applications

Therefore, since

$$\pi(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

we have $\Phi^{-1}(\pi) = \beta_0 + \beta_1 x$ where $\beta_0 = -\mu/\sigma$ and $\beta_1 = 1/\sigma$

Here, the "link" between the data and the covariates is the inverse cumulative normal probability function

In this context, the link function Φ^{-1} is often called the **probit**, for probability unit

Early applications

Probit models are used in both the biological and social sciences; it's attraction is that in some applications the model is natural

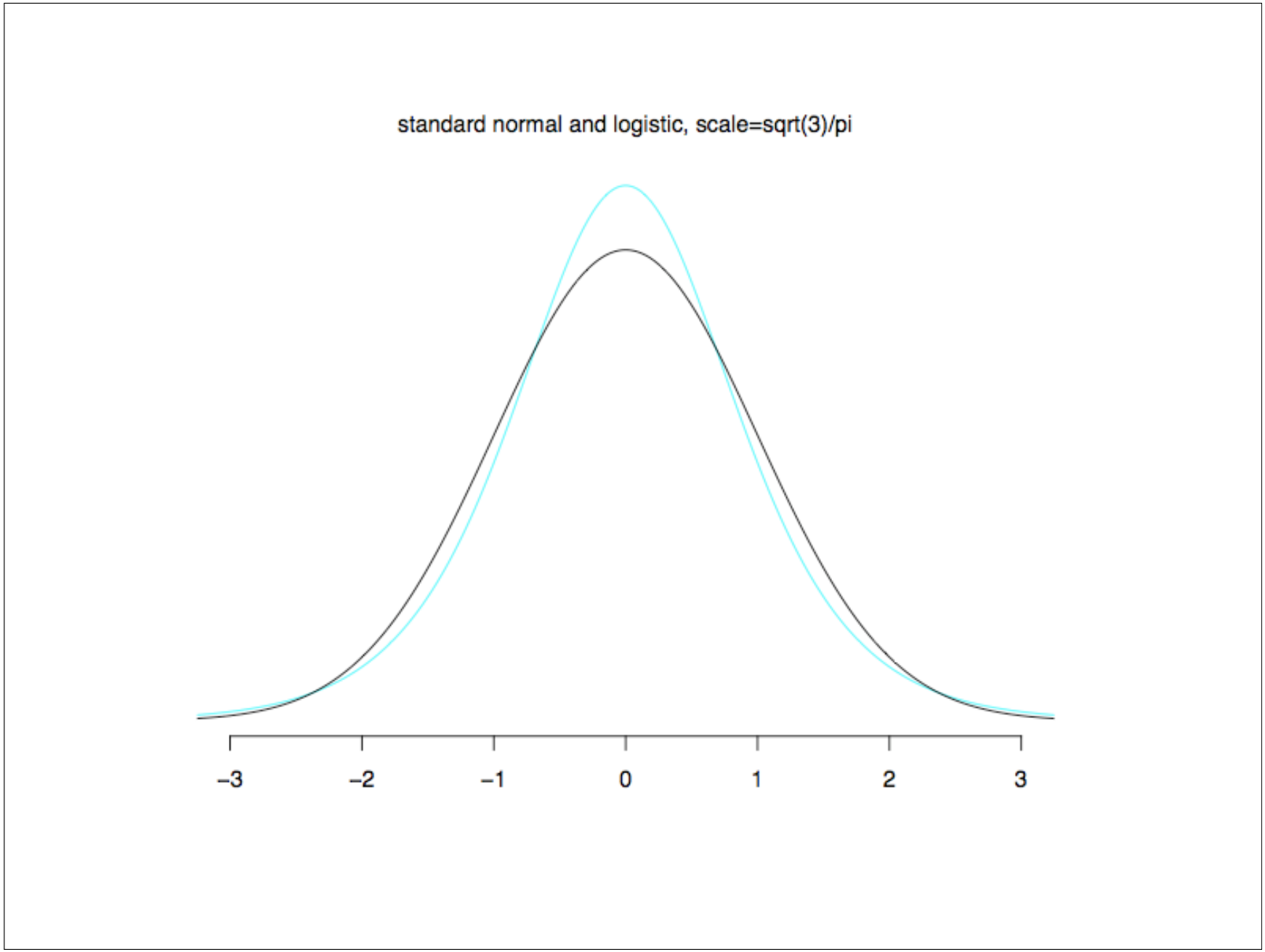
For example, $x = \mu$ is called the median lethal dose or LD50, because it is the dose that can be expected to kill half the animals; LD50 is often used as a general indicator of toxicity

And logistic regression?

We can cast logistic regression in these terms also; rather than the uniform or the normal CDF, logistic regression takes as its tolerance function, well, the logistic distribution

$$f(t) = \frac{e^{-(t-\mu)/s}}{s\left(1 + e^{-(t-\mu)/s}\right)^2}$$

The logistic is also a two-parameter location-scale family (the variance of the standard logistic with mean 0 and scale 1 is $\pi^2/3$)


standard normal and logistic, scale=sqrt(3)/pi
−3
−2
−1
0
1
2
3

Logistic regression

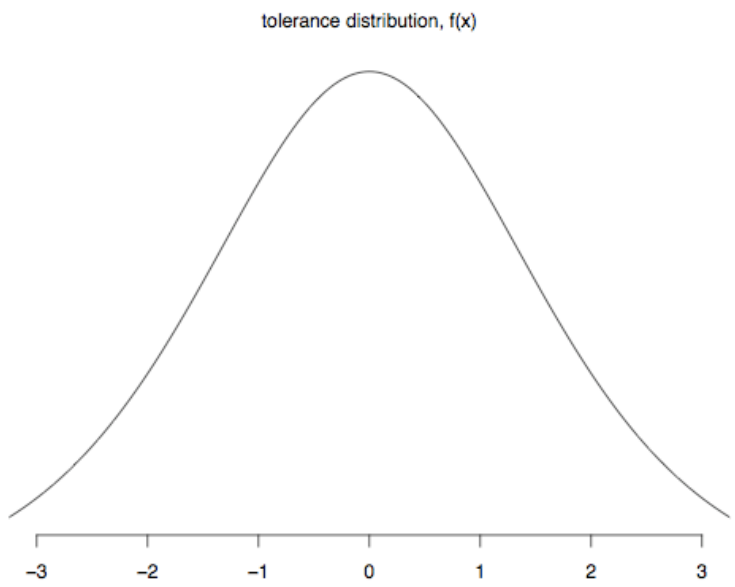
After a little calculus (and not very much, really), we find that the CDF of the logistic distribution is

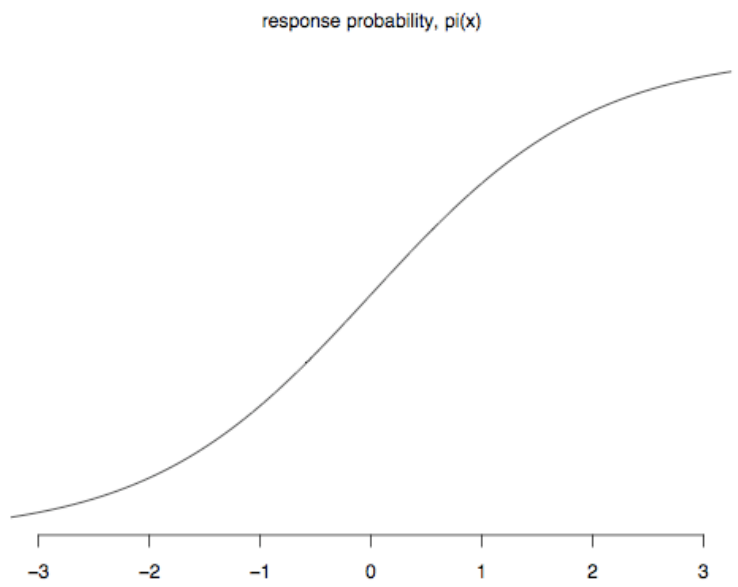
$$\begin{aligned}\int_{-\infty}^{x} f(t)dt &= \int_{-\infty}^{x} \frac{e^{-(t-\mu)/s}}{s\left(1+e^{-(t-\mu)/s}\right)^2}dt \\ &= \frac{1}{1+e^{-(x-\mu)/s}} \\ &= \frac{e^{(x-\mu)/s}}{1+e^{(x-\mu)/s}}\end{aligned}$$

and so if we define $\beta_0 = -\mu/s$ and $\beta_1 = 1/s$, then

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad \text{or} \quad \operatorname{logit}\pi(x) = \beta_0 + \beta_1 x$$

The term **logit** was coined in the 40s, with a specific reference to the probit defined earlier


tolerance distribution, f(x)
−3
−2
−1
0
1
2
3


response probability, pi(x)
−3
−2
−1
0
1
2
3

Roads to logistic regression

We approached the topic through odds ratios, making a case that this represents a natural scale for regression-style models

We took an historical view of the subject following the development of dose-response models

Finally, we will look at KL divergence and see that the logit emerges as a canonical parameter in a Bernoulli distribution

A mechanism

In addition to its pedagogical value, the dose-response framework give us another way to view our binary models

Let's introduce an unobserved (often called "latent") variable that will be used to generate the binary data that we actually observe

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i \leq 0 \end{cases}$$

$$z_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where the ϵ_i are independent and have the logistic distribution

A mechanism

If ϵ has the logistic distribution with mean 0 and scale parameter 1,
then $z = \beta_0 + \beta_1 x + \epsilon$ also has the logistic distribution, but with mean $\beta_0 + \beta_1 x$
and scale parameter 1

Therefore,

$$\text{Prob}\,(y = 1) = \text{Prob}\,(z > 0) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

using the form of the logistic distribution given earlier

A latent variable approach

We can think of the unobserved or latent variables as describing a county's propensity to vote for Obama or Clinton, say -- By collecting data, all we are able to see really is the sign of this latent variable (positive meaning we see a 1 and negative meaning a 0)

By forming a model, we might be able to see into these propensities a bit more, getting a handle on their magnitude

Logit or probit or ... ?

There's nothing worse than recognizing you have a choice where you didn't think you had one -- Or rather you had been making a choice for no good reason

The choice of link (probit, logit, etc.) is something that comes up in GLMs and there exist formal tests to help you decide the "goodness" of your link (yes, there's a test for that!) if you embed it, say, in a flexible family

For the moment, we'll treat the probit model as a historical motivator and not really discuss the choice but Pregibon (1980) is a great place to look for more information

The normal linear model

The model we worked with for the first few weeks of the course consisted of a few basic ingredients

The conditional mean. Our focus was on modeling $\mathsf{E}(\mathsf{Y}|\mathsf{X} = \mathsf{x})$ where x is a vector of p covariates and Y is a random variable representing our response

A systematic component. Here we introduced some structure to describe the dependence of the response Y on the predictors x -- We talked mainly about a simple linear model of the form

$$\beta_1 \mathsf{x}_1 + \cdots + \beta_\mathsf{p} \mathsf{x}_\mathsf{p}$$

A random component. Finally, we introduced a distribution to describe the variation we expect to see around the conditional mean -- For the normal linear model this meant adding independent normal noise with constant variance

$$\mathsf{Y} = \beta_1 \mathsf{x}_1 + \cdots + \beta_\mathsf{p} \mathsf{x}_\mathsf{p} + \epsilon \quad \text{where} \quad \epsilon \sim \mathsf{N}(0, \sigma^2)$$

Generalization

These three pieces will be important when we attempt to generalize the structure of the normal linear model -- The amazing thing is that **a large number of modeling contexts can be "handled" at one time**

To be clear, we use the term "handle" to refer to a common set of tools to define and fit a model, address basic inferential questions, perform diagnostic procedures and entertain model elaborations -- Computationally, this means that one essential approach can be applied in a large number of modeling contexts

The class of so-called Generalized Linear Models or GLMs was developed in a series of papers and a text by Nelder and McCullagh appearing in the 1970s and early 1980s

Generalized linear models

To understand the structures of these models and how they generalize the normal linear model, we'll examine each of the three components (a description of the variability in our data, a structural piece involving covariates and a link to the conditional mean)

We'll start with the random component...

The random component

To simplify things a little, let's ignore the covariates for the moment and focus on a single random variable -- For the normal linear model, our response Y has a normal distribution with mean μ and variance σ^2

Its probability function is given by

$$\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(\mathrm{y}-\mu)^2/2\sigma^2}$$

For simplicity, we'll also assume $\sigma^2 = 1$ (don't worry, it'll come back later)

The random component

Then, we can rearrange the terms its probability function to derive the following expression

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-(y-\mu)^2/2} = e^{y\mu - \mu^2/2 - y^2/2 - \log\sqrt{2\pi}}, \quad y \in \mathbb{R}$$

While this doesn't seem like much, it turns out that a fair number of other distributions we're interested in can also be written similarly...

The random component

Bernoulli trials. Here we are modeling a coin toss and our response variable Y takes on the values 0 or 1 and the probability of seeing a 1 is p -- The (discrete) probability function for Y is given by

$$f(y) = p^y(1-p)^{1-y} = e^{y\log[p/(1-p)]+\log(1-p)}, \quad y \in \{0,1\}$$

Poisson counts. Now our response variable Y has a Poisson distribution with mean λ and takes on integer values -- The (discrete) probability function for Y is given by

$$f(y) = \frac{\lambda^y e^{-\lambda}}{y!} = e^{y\log\lambda-\lambda-\log y!}, \quad y = 0,1,2,\ldots$$

Generalized linear models: An exponential family

In all three cases, the probability functions have the form

$$f(y) = e^{y\theta - b(\theta) + c(y)}$$

Those of you in Stat 200a will probably recognize this as an exponential family of probabilities -- In that context, we often refer to θ as the canonical parameter for the family

(To be completely general, McCullagh and Nelder allow for a "dispersion parameter" ϕ in their definition which would allow for the incorporation of, say, the variance term σ^2 in a normal linear model

$$f(y) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$$

For the moment, however, we'll keep to the simpler expression)

Examples

Normal. The probability function is $f(y) = e^{y\mu - \mu^2/2 - y^2/2 - \log\sqrt{2\pi}}$ so that

$$\theta = \mu, \quad b(\theta) = \theta^2/2, \quad c(y) = -(y^2/2 + \log\sqrt{2\pi})$$

Bernoulli. Y can take on just two values 0 and 1 and we had written probability function as $f(y) = e^{y\log[p/(1-p)] + \log(1-p)}$ so that

$$\theta = \log\frac{p}{1-p}, \quad b(\theta) = \log\left(1 + e^{\theta}\right), \quad c(y) = 0$$

Poisson. Finally, Y takes on non-negative integer values and has probability function $f(y) = e^{y\log\lambda - \lambda - \log y!}$ so that

$$\theta = \log\lambda, \quad b(\theta) = e^{\theta}, \quad c(y) = -\log y!$$

Examples: So what?

While all we've done is a whole lot of shuffling of algebraic expressions, it turns out that we can say quite a lot about the members of this family...

An exponential family

With this probability function, we can derive a common expression for the mean and variance of Y -- First, note that because

$$1 = \int f(y)dy = \int e^{\theta y - b(\theta) + c(y)} dy$$

(where we can swap integration for a sum if we have a discrete probability function) we have

$$b(\theta) = \log \int e^{\theta y + c(y)} dy$$

Therefore, assuming we can do the appropriate differentiation gymnastics, we find that

$$b'(\theta) = \frac{\int y e^{y\theta + c(y)} dy}{\int e^{y\theta + c(y)} dy} = \int y e^{y\theta - b(\theta) + c(y)} dy = \int y f(y) dy$$

or simply $EY = b'(\theta)$

An exponential family

Therefore, $\mathsf{E}Y = \mathsf{b}'(\theta)$ and (after taking a second derivative) $\operatorname{var} Y = \mathsf{b}''(\theta)$ -- One consequence of this is that EY is a monotone function of the canonical parameter θ

From now on, we'll refer to the mean of the family f(y) with the symbol μ -- For the Poisson case $\mu = \lambda$ and for the Bernoulli $\mu = \mathsf{p}$

All this means we can use μ as well as θ to unambiguously refer to a member of our family

Modeling and a link

Next, we come to the systematic component of our generalized linear model -- In the normal case, we were able to directly model the effect of the conditional mean on a set of covariates

$$\mu(\mathrm{x}) = \beta_1 \mathrm{x}_1 + \ldots + \beta_\mathrm{p} \mathrm{x}_\mathrm{p}$$

In the general case, this is probably not the best choice -- What can go wrong?

Modeling and a link

Consider the Bernoulli model where $\mu(x) = p(x)$ -- A direct model linking the covariates and the conditional mean would pose some difficult optimization problems

We know, for example, that no matter how we specify the dependence, p(x) can only take values in the interval [0,1] -- Specifying the value of $\beta = (\beta_1, \ldots, \beta_p)^t$ so that this remains true for any value of x can be hard (if not impossible)

Modeling and a link

A link function relates a linear predictor $\beta_1 x_1 + \cdots + \beta_p x_p$ to the conditional mean $\mu(x) = E(Y|X = x)$ -- Ideally, it provides a "sensible" mapping that removes any constraints on $\beta_1, \ldots, \beta_p$

For example, in the case of count or Poisson data, we might consider a log-link -- That is, we can set

$$\log \mu(x) = \beta_1 x_1 + \cdots + \beta_p x_p \quad \text{so that} \quad \mu(x) = e^{\beta_1 x_1 + \cdots + \beta_p x_p}$$

In the Bernoulli case, we can enforce the [0,1] constraint on p(x) with a link derived from any cumulative distribution function -- For the normal, say

$$\Phi^{-1}\big(\mu(x)\big) = \beta_1 x_1 + \cdots + \beta_p x_p \quad \text{so that} \quad \mu(x) = \Phi(\beta_1 x_1 + \cdots + \beta_p x_p)$$

The canonical link

With the GLMs we have seen so far, we find an expression for a link function comes directly from the exponential family -- Specifically, suppose we introduce our covariates on the scale of the canonical parameter

$$\theta(\mathsf{x}) = \beta_1 \mathsf{x}_1 + \cdots + \beta_\mathsf{p} \mathsf{x}_\mathsf{p}$$

Then, using the fact that $\mu = \mathsf{b}'(\theta)$ or $\theta = (\mathsf{b}')^{-1}(\mu)$, the so-called canonical link is given by $(\mathsf{b}')^{-1}$

Examples

Normal. As we expect, because $\mathsf{b}(\theta) = \theta^2/2$, $\mathsf{b}'(\theta) = \theta$ and we have the identity link $\theta = \mu$ (this is why we didn't have to think too hard about OLS)

Bernoulli. Here $\mathsf{b}(\theta) = \log(1 + \mathsf{e}^\theta)$ and $\mathsf{b}'(\theta) = \mathsf{e}^\theta/(1 + \mathsf{e}^\theta)$ so that inverting gives the "logit" link

$$\theta = \log \mathsf{p}/(1 - \mathsf{p}) = \mathsf{logit}\ \mathsf{p}$$

Poisson. Finally, for this model $\mathsf{b}(\theta) = \mathsf{e}^\theta$ and $\mathsf{b}'(\theta) = \mathsf{e}^\theta$, so we have the log-link we mentioned before $\theta = \log \lambda$

Examples

Normal. With the identity link we have

$$E(Y|X = x) = \mu(x) = \beta_1 x_1 + \cdots + \beta_p x_p$$

Bernoulli. With the logit we have

$$E(Y|X = x) = p(x) = \frac{e^{\beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_1 x_1 + \cdots + \beta_p x_p}}$$

Poisson. And under the log-link

$$E(Y|X = x) = \lambda(x) = e^{\beta_1 x_1 + \cdots + \beta_p x_p}$$

Link functions

In general, the link introduces a scale that is more “sensible” for modeling than the original scale of the conditional mean -- This construction, while introducing some indirection, will let us estimate $\beta_1, \ldots, \beta_p$ in an unconstrained way

There is nothing special about the canonical links except that (see Stat 200) there exist sufficient statistics for the parameters $\beta_1, \ldots, \beta_p$ (more later)

To sum

We've introduced an exponential family of probability functions that specify the random part of a model for our responses Y -- We can use either the canonical parameter θ or the mean μ as an index for a given model (Poisson, normal, or Bernoulli, say)

Then, we considered how best to introduce covariates into the model -- This provided us with the notion of a link function and a scale for introducing a linear predictor

After all that, we still need to consider what to do when given data from this model -- That is, how do we choose values for our "regression" parameters?

Measuring discrepancy

To fit the normal linear model, we introduced the least squares criterion and alluded to the fact that it was equivalent to maximum likelihood estimation -- For generalized linear models, we will also need **a measure of discrepancy to help us select regression parameters**

At this point we are going to part company with the usual way GLMs are developed (which appeals to likelihood theory) -- Instead we will take **a distance measure approach** that generalizes our work with OLS

Distances between probabilities

There are a number of ways to compare two densities -- Suppose that f(y) and g(y) have the same sample space, then we could look at analogs of the 1 and 2 norms for vectors

$$L_1(f,g) = \int |f(y) - g(y)| dy \quad \text{and} \quad L_2(f,g) = \sqrt{\int (f(y) - g(y))^2 dy}$$

We will instead look at a measure that has a long, long history in statistics and will admit some lovely geometric properties (the kind of thing we are now old hands with in OLS)

Kullback-Leibler divergence

Let P(y) and Q(y) denote two (for the moment, discrete) probability distributions that we'd like to compare -- Then define the KL divergence between the two as

$$K(P, Q) = \sum_{y \in \mathcal{Y}} P(y) \log \frac{P(y)}{Q(y)}$$

where $\mathcal{Y}$ is the set of elements for which either P(y) or Q(y) is non-zero and we take 0 log 0/q = 0 and p log p/0 = ∞

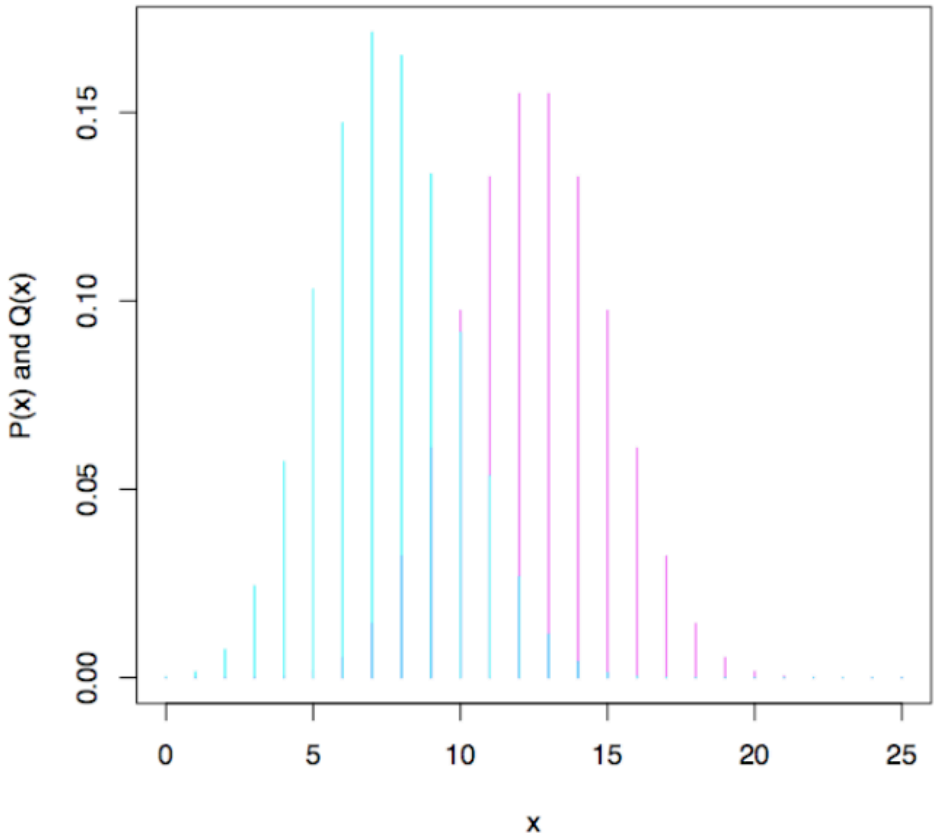
Let's apply this before thinking about it too much...

A “distance”

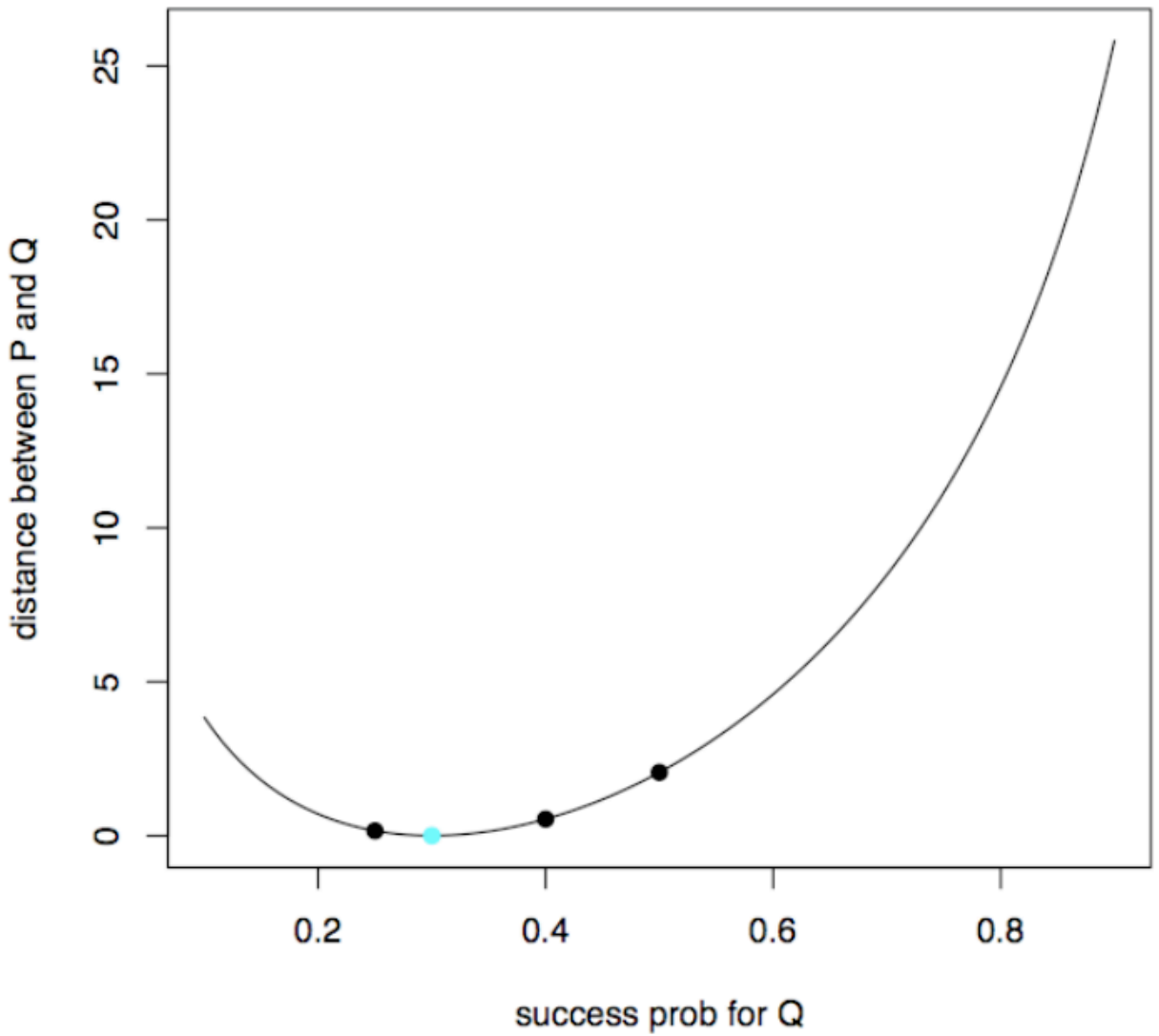
Let’s let P be the binomial distribution with n=25 and p=0.3, and take Q be a binomial with n=25 but p=0.5

On the right we compare the two distributions -- They’re clearly different (not that you needed the picture to know that)

On the next page, we expand things and look at a range of possible success probabilities for Q -- What do we notice?



```
# experiments with our new distance measure...

P = dbinom(0:25,size=25,p=0.3)
Q = dbinom(0:25,size=25,p=0.5)
sum(P*log(P/Q))

# [1] 2.057072

# intuitively, this distribution should be closer to P
Q = dbinom(0:25,size=25,p=0.4)
sum(P*log(P/Q))

# [1] 0.5400214

# intuitively, this distribution should be even closer...
Q = dbinom(0:25,size=25,p=0.25)
sum(P*log(P/Q))

# [1] 0.1600364
```


distance between P and Q
0
5
10
15
20
25
0.2
0.4
0.6
0.8
success prob for Q

ON INFORMATION AND SUFFICIENCY

By S. Kullback and R. A. Leibler

The George Washington University and Washington, D. C.

1. Introduction. This note generalizes to the abstract case Shannon's definition of information [15], [16]. Wiener's information (p. 75 of [18]) is essentially the same as Shannon's although their motivation was different (cf. footnote 1, p. 95 of [16]) and Shannon apparently has investigated the concept more completely. R. A. Fisher's definition of information (intrinsic accuracy) is well known (p. 709 of [6]). However, his concept is quite different from that of Shannon and Wiener, and hence ours, although the two are not unrelated as is shown in paragraph 2.

R. A. Fisher, in his original introduction of the *criterion of sufficiency*, required "that the statistic chosen should summarize the whole of the relevant information supplied by the sample," (p. 316 of [5]). Halmos and Savage in a recent paper, one of the main results of which is a generalization of the well known Fisher-Neyman theorem on sufficient statistics to the abstract case, conclude, "We think that confusion has from time to time been thrown on the subject by . . . , and (c) the assumption that a sufficient statistic contains all the information in only the technical sense of 'information' as measured by variance," (p. 241 of [8]). It is shown in this note that the information in a sample as defined herein, that is, in the Shannon-Wiener sense cannot be increased by any statistical operations and is invariant (not decreased) if and only if sufficient statistics are employed. For a similar property of Fisher's information see p. 717 of [6], Doob [19].

We are also concerned with the statistical problem of discrimination ([3], [17]), by considering a measure of the "distance" or "divergence" between statistical populations ([1], [2], [13]) in terms of our measure of information. For the statistician two populations differ more or less according as to how difficult it is to discriminate between them with the best test [14]. The particular measure of divergence we use has been considered by Jeffreys ([10], [11]) in another connection. He is primarily concerned with its use in providing an invariant density of *a priori* probability. A special case of this divergence is Mahalanobis' generalized distance [13].

A “distance”

If you think about the term “distance” for a moment, a couple of properties come to mind quickly

1. The distance between two objects (in this case, probability distributions) should be non-negative -- That is, $K(P, Q) \geq 0$

2. The distance between any object and itself should be zero -- This means that K(P,P) = 0

As we can see from the previous slides, these two properties seem to hold at least for the restricted family of binomial distributions we examined -- The curve on the previous slide touches zero only when P=Q

A “distance”

Afer a little more thought, it also seems reasonable to expect that measuring the distance from P to Q, or K(P,Q), should give the same value as the distance from Q to P, K(Q,P)

Unfortunately, this is not the case for the Kullback-Leibler divergence -- Which is precisely why we are referring to it as a “distance” in quotes

In many cases (like the main application we have in mind), there is a natural “direction” for computing the divergence and we don’t have to fuss about K(P,Q) or K(Q,P) -- There are, by the way, several attempts to “symmetrize” K that we won’t really touch on

A “distance”

It is often convenient to write the “distance” in terms of an expectation -- Suppose Y is a random variable with distribution P(y), then K(P,Q) is just the expectation of the random variable log P(Y)/Q(Y)

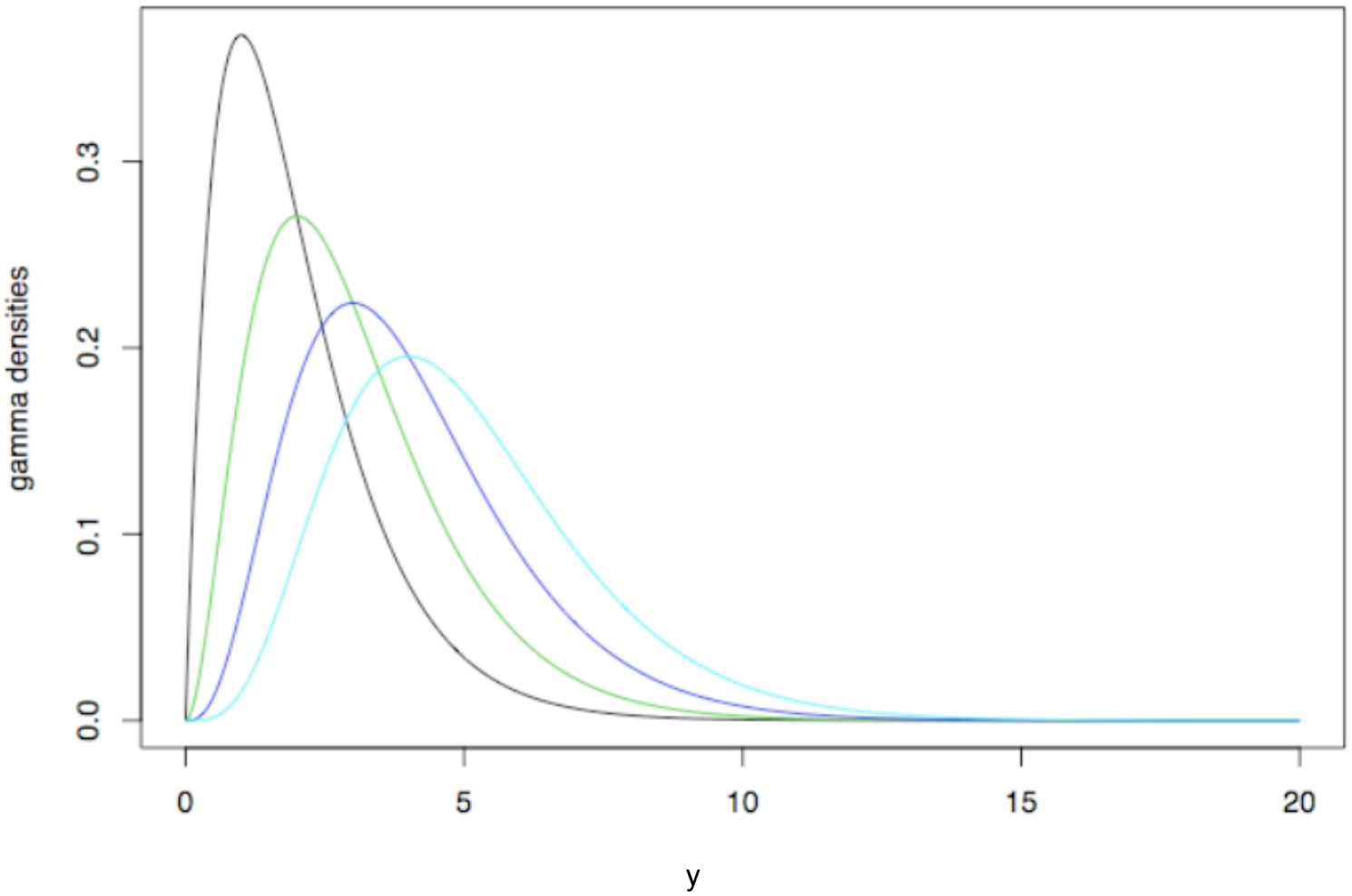
$$K(P, Q) = \sum_{y \in \mathcal{Y}} P(y) \log \frac{P(y)}{Q(y)} = E \log \frac{P(Y)}{Q(Y)}$$

This last expression helps us deal with both discrete and continuous random variables in one go (ignoring the sum versus integral shuffle)

A “distance”

With this change, we can now think about the distance between any two continuous distributions -- On the next page we have four gamma distributions (with shape parameters 2, 3, 4 and 5, and common scale 1)

Here if we let P(x) be the first gamma (plotted in black), the distances between it and the three gammas are 0.27 (green), 0.94 (blue), and 1.9 (cyan)


gamma densities
0.0
0.1
0.2
0.3
0
5
10
15
20
y

The “distance” between normals

OK, let’s flip back to our GLM notation -- Let f_{μ_1} and f_{μ_2} denote two normal distributions with means μ_1 and μ_2, respectively, each with variance 1

Let’s look at the log-ratio...

$$\begin{aligned}
\log \frac{f_{\mu_1}}{f_{\mu_2}} &= \log f_{\mu_1} - \log f_{\mu_2} \\
&= \left(y\mu_1 - \mu_1^2/2 - y^2/2 - \log\sqrt{2\pi}\right) - \left(y\mu_2 - \mu_2^2/2 - y^2/2 - \log\sqrt{2\pi}\right) \\
&= y(\mu_1 - \mu_2) - (\mu_1^2 - \mu_2^2)/2
\end{aligned}$$

The “distance” between normals

Taking an expectation of this last expression with respect to a random variable Y with probability function f_{μ_1} we have

$$\begin{aligned} K(f_{\mu_1}, f_{\mu_2}) &= E\left[Y(\mu_1 - \mu_2) - (\mu_1^2 - \mu_2^2)/2\right] \\ &= \mu_1^2 - \mu_1\mu_2 - (\mu_1^2 - \mu_2^2)/2 \\ &= (\mu_1 - \mu_2)^2/2 \end{aligned}$$

Ha! The “distance” is just the squared difference in means (well, with a factor of 2 floating around)

The “distance” between GLMs

Ok, the same basic algebra holds for Poisson and Bernoulli observations, or any GLM for that matter -- The KL divergence between two probability functions with parameters θ_1 and θ_2 is given by

$$\mathsf{K}(\mathsf{f}_{\theta_1}, \mathsf{f}_{\theta_2}) = (\,\theta_1 - \theta_2\,)\,\mu_1 - (\,\mathsf{b}(\theta_1) - \mathsf{b}(\theta_2)\,)$$

Given the equivalence between the canonical parameter θ and the conditional mean μ in our framework, from this point on, we will instead refer unambiguously to f_{μ_1} and f_{μ_2} and write

$$\mathsf{K}(\mu_1, \mu_2) = (\,\theta_1 - \theta_2\,)\,\mu_1 - (\,\mathsf{b}(\theta_1) - \mathsf{b}(\theta_2)\,)$$

(In sympathy with the normal model on the previous slide)

A “distance”

With this in hand, we can regard the random variable $\mathsf{K}(\mathsf{Y}, \mu)$ as a measure of the deviation of the random variable Y from the mean μ -- What we mean here is

$$\mathsf{K}(\mathsf{Y}, \mu) = \mathsf{E}_{\mathsf{Y}} \log \frac{\mathsf{f}_{\mathsf{Y}}(\mathsf{Y}^*)}{\mathsf{f}_{\mu}(\mathsf{Y}^*)} = \int \mathsf{f}_{\mathsf{Y}}(\mathsf{y}) \log \frac{\mathsf{f}_{\mathsf{Y}}(\mathsf{y})}{\mathsf{f}_{\mu}(\mathsf{y})} \mathsf{dy}$$

where in the middle equation Y^* has probability function f_{Y}

Again, $\mathsf{K}(\mathsf{Y}, \mu)$ is a random variable -- With each realization we have a new integral involving f_{Y} , a probability function in our family with mean Y

A “distance”

Similarly, we can think of $\mathrm{K}(\mathrm{y}, \mu)$ as an observation of the random variable $\mathrm{K}(\mathrm{Y}, \mu)$
-- In the normal case, $\mathrm{K}(\mathrm{y}, \mu) = (\mathrm{y} - \mu)^2/2$ and in general we think of
as measuring the prediction error in using μ to predict y

In this context, we think of K as our loss function, replacing OLS for the non-normal families we’ve been considering

The “distance” between GLMs

Now, suppose we use the expression for the KL divergence

$$\mathsf{K}(\mu_1, \mu_2) = (\,\theta_1 - \theta_2\,)\,\mu_1 - (\,\mathsf{b}(\theta_1) - \mathsf{b}(\theta_2)\,)$$

and apply it to $\mathsf{K}(\mathsf{y}, \mu)$ -- So, we let θ_μ be the parameter associated with μ and θ_y the parameter associated with y

Then we can write

$$(\theta_\mathsf{y} - \theta_\mu)\mathsf{y} - (\mathsf{b}(\theta_\mathsf{y}) - \mathsf{b}(\theta_\mu)) = \mathsf{y}\theta_\mathsf{y} - \mathsf{b}(\theta_\mathsf{y}) - (\mathsf{y}\,\theta_\mu \;\; - \mathsf{b}(\theta_\mu)) = \log \frac{\mathsf{f}_\mathsf{y}(\mathsf{y})}{\mathsf{f}_\mu(\mathsf{y})}$$

The “distance” between GLMs

Admittedly this is getting a little far out there, but what we’ve shown is that at least notationally

$$K(y, \mu) = \log \frac{f_y(y)}{f_\mu(y)}$$

The “distance” between GLMs

You can think of $K(y,\mu)$ as comparing two models -- The second involves the parameter μ and the other is a “saturated model” in which the data are fit exactly

This construction has very close ties with maximum likelihood estimation and, ultimately, likelihood ratio statistics

Some geometry

To put this all to work, assume we have a random variable Y with mean μ
-- Then, given a predictor η, we can use $\mathsf{K}(\mathsf{Y}, \eta)$ to measure the loss

Specifically, we can define the expected prediction error $\mathsf{E}\,\mathsf{K}(\mathsf{Y}, \eta)$,
which, in the normal case is just $\mathsf{E}(\mathsf{Y} - \eta)^2/2$, a factor of 2 away from what we've been using all quarter

Now, since

$$\mathsf{K}(\mathsf{Y}, \eta) = \log \frac{\mathsf{f}_{\mathsf{Y}}(\mathsf{Y})}{\mathsf{f}_{\eta}(\mathsf{Y})} = \log \frac{\mathsf{f}_{\mathsf{Y}}(\mathsf{Y})}{\mathsf{f}_{\mu}(\mathsf{Y})} + \log \frac{\mathsf{f}_{\mu}(\mathsf{Y})}{\mathsf{f}_{\eta}(\mathsf{Y})} = \mathsf{K}(\mathsf{Y}, \mu) + \mathsf{K}(\mu, \eta)$$

we have that

$$\mathsf{E}\,\mathsf{K}(\mathsf{Y}, \eta) = \mathsf{E}\,\mathsf{K}(\mathsf{Y}, \mu) + \mathsf{K}(\mu, \eta)$$

which in the normal case is just $\mathsf{E}\,(\mathsf{Y} - \eta)^2 = \mathsf{var}\,\mathsf{Y} + (\mu - \eta)^2$

Some geometry

This result, $\mathsf{E}\,\mathsf{K}(\mathsf{Y}, \eta) = \mathsf{E}\,\mathsf{K}(\mathsf{Y}, \mu) + \mathsf{K}(\mu, \eta)$ indicates immediately that the mean is the best predictor in terms of KL prediction error!

Introducing data

Now, suppose we have a random sample $y_1, \ldots, y_n$ from f_μ and we wish to estimate μ -- In the normal case we chose η to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \eta)^2$$

as an estimate of $E(Y - \eta)^2$

The same holds for KL prediction error -- We would like to choose η to minimize the sum of errors

$$\frac{1}{n}\sum_{i=1}^{n}K(y_i, \eta)$$

Introducing data

We can write out the error criterion and simply minimize -- The errors are

$$\frac{1}{n}\sum_i K(y_i,\eta) = \sum_i \left[(\theta_{y_i} - \theta)y_i - (b(\theta_{y_i}) - b(\theta))\right]$$

(where we have associated θ with the mean η) and taking a derivative with respect to θ and setting the result to zero, we find that the minimum occurs at

$$\bar{y} = b'(\widehat{\theta})$$

or simply $\widehat{\theta} = (b')^{-1}(\bar{y})$

Introducing data

Now, suppose our observations y_i are paired with predictors $x_{i1}, \ldots, x_{ip}$ -- Let's again form the n-by-p model matrix M, writing the ith row as $m_i = (x_{i1}, \ldots, x_{ip})^t$, and form $y = (y_1, \ldots, y_n)$

Next, we can expose or model the dependence of the mean of our response through the canonical parameter $\theta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, and set $\theta = M\beta$ -- Now, since θ is a vector, we'll abuse notation slightly and take $b(\theta)$ to be a vector $(b(\theta_1), \ldots, b(\theta_n))^t$

Following the same derivative approach on the previous slide we find that the minimizers of this expression, our estimates $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$, satisfy

$$M^t y = M^t b'(\widehat{\theta}) = M^t b'(M\widehat{\beta})$$

where we have again abused notation and let $b'(\widehat{\theta})$ be a n-vector

Introducing data

We can write the last expression in a more usable form recalling that $\widehat{\mu} = \mathrm{b}'(\widehat{\theta})$ (where we are again dealing in vectors) so that

$$\mathrm{M}^{\mathrm{t}}\mathrm{y} = \mathrm{M}^{\mathrm{t}}\mathrm{b}'(\widehat{\theta}) = \mathrm{M}^{\mathrm{t}}\widehat{\mu}$$

We'll use this shortly...

Estimation

It will turn out that there is a single, consistent way to estimate these parameters for the class of generalized linear models -- Recall that for the normal linear model, our "objective function" a parabola in $\beta_1, \ldots, \beta_p$ and we have a closed-form expression for the minimum of a parabola

In general, we aren't so lucky and we have to appeal to some other framework for minimizing the equations -- While we don't have a parabola, we do have a function that's concave

To see this, take second derivatives with respect to the parameters $\beta_1, \ldots, \beta_p$ -- We find that the Hessian matrix (yes, it's back to calculus!) is given by $M^t D M$ where

$$D = \mathrm{diag}(d_1, \ldots, d_n) \quad \text{and} \quad d_i = b''(\theta_i) = b''(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

What can you say about this matrix?

Estimation

The quantity M^tDM is a positive definite matrix (more interpretations later) and hence our “objective function” is concave -- So while not a parabola, it is bowl-shaped meaning a procedure like Newton-Raphson will be effective in finding a minimizer

We’ll see that Newton-Raphson, an iterative procedure, can be thought of as a sequence of weighted regressions, making it a direct generalization of ordinary least squares

Distances

Our divergence measure turns out to have some satisfying geometric properties -- Simon (1973) establishes a Pythagorean relationship behind minimizing the divergence measure

For the next few slides, it will be helpful to create a new symbol for our divergence over n-element vectors -- That is, given vectors $\mu = (\mu_1, \ldots, \mu_n)$ and $\eta = (\eta_1, \ldots, \eta_n)$, set

$$K_n(\mu, \eta) = \sum_{i=1}^{n} K(\mu_i, \eta_i)$$

We can also get here by computing the distance between two models for the complete data set, where we extend our exponential family by multiplication (independent responses given the inputs)

For the normal linear model, this is just $\sum_i (\mu_i - \eta_i)^2$ or squared error loss

Distances

Simon (1973) considers situations in which the divergence measures can be factored -- That is (and think about least squares now)

$$K_n(\mu_a, \mu_c) = K_n(\mu_a, \mu_b) + K_n(\mu_b, \mu_c)$$

What conditions on μ_a , μ_b and μ_c will make this work?

Distances

To answer this, we'll introduce more notation -- Define three vectors

$$\theta_a = M\beta_a, \quad \theta_b = M\beta_b, \quad \text{and} \quad \theta_c = M\beta_c$$

where the superscript indicates the first, second and third coefficient vectors associated with the conditional means

$$\mu_a = b'(\theta_a), \quad \mu_b = b'(\theta_b), \quad \text{and} \quad \mu_c = b'(\theta_c)$$

With these definitions, the conditions on the previous page become

$$\begin{aligned}(\theta_a - \theta_c)^t\mu_a - [b(\theta_a) - b(\theta_c)] \quad = \quad & (\theta_a - \theta_b)^t\mu_a - [b(\theta_a) - b(\theta_b)] \\ & +(\theta_b - \theta_c)^t\mu_b - [b(\theta_b) - b(\theta_c)]\end{aligned}$$

Distances

Simplifying this somewhat we come up with the single condition

$$(\theta_b - \theta_c)^t(\mu_a - \mu_b) = 0$$

Simon then considers deriving the result for so-called “nested hypotheses” on the regression coefficients -- We’ll speak instead about the column space of the associated model matrices

Distances

Let's consider three linear subspaces of the column space of our model matrix M -- The first H_a is the full column space, the second H_b is a subspace and the third H_c is a subspace of H_b

For simplicity, you can think of H_a as comprising the full model with p variables, H_b as being only a subset of q (q<p) variables, and H_c as being an even smaller subset of r < q variables

Distances

To connect with Simon's notation, we can take H_a to be the column space of M, the set of all vectors $M\beta_a$ for $\beta_a \in \mathbb{R}^p$ -- Simon then considers a subspace H_b defined by the set of all $(MB_{p\times q})\delta$ for $\delta \in \mathbb{R}^q$ where q < p and clearly $H_b \subset H_a$

Finally, Simon considers a third space, H_c , $(MC_{p\times r})\gamma$ for $\gamma \in \mathbb{R}^r$ where r < q and the C consists of the first r columns of B -- WIth this construction,

$H_c \subset H_b \subset H_a$

Distances

From a previous slide, we know that the minimizer over β_a of

$$K_n(y,\mu) = \frac{1}{n}\sum_{i=1}^{n} K(y_i,\mu)$$

for $\beta_a \in H_a = \mathbb{R}^p$ satisfies $M^t y = M^t \widehat{\mu}_a$ -- Similarly, the minimizers in H_b and H_c satisfy $(MB)^t y = (MB)^t \widehat{\mu}_b$ and $(MC)^t y = (MC)^t \widehat{\mu}_c$, respectively

Direct substitution of these into the condition at the top of the previous slide yields the desired result

Distances

To sum, Simon (1973) establishes the relationship

$$K_n(\widehat{\mu}_a, \widehat{\mu}_c) = K_n(\widehat{\mu}_a, \widehat{\mu}_b) + K(\widehat{\mu}_b, \widehat{\mu}_c)$$

Here we have chosen $\widehat{\mu}_a$, $\widehat{\mu}_b$ and $\widehat{\mu}_c$ using our divergence criterion to form "projections" of the data y where the regression coefficients are in each of H_a , H_b and H_c , respectively

He refers to this as the "Additivity of Forward Information" in that each of the terms in the equation involve the divergence between a subspace (first position) and one contained in it (second position)

Distances

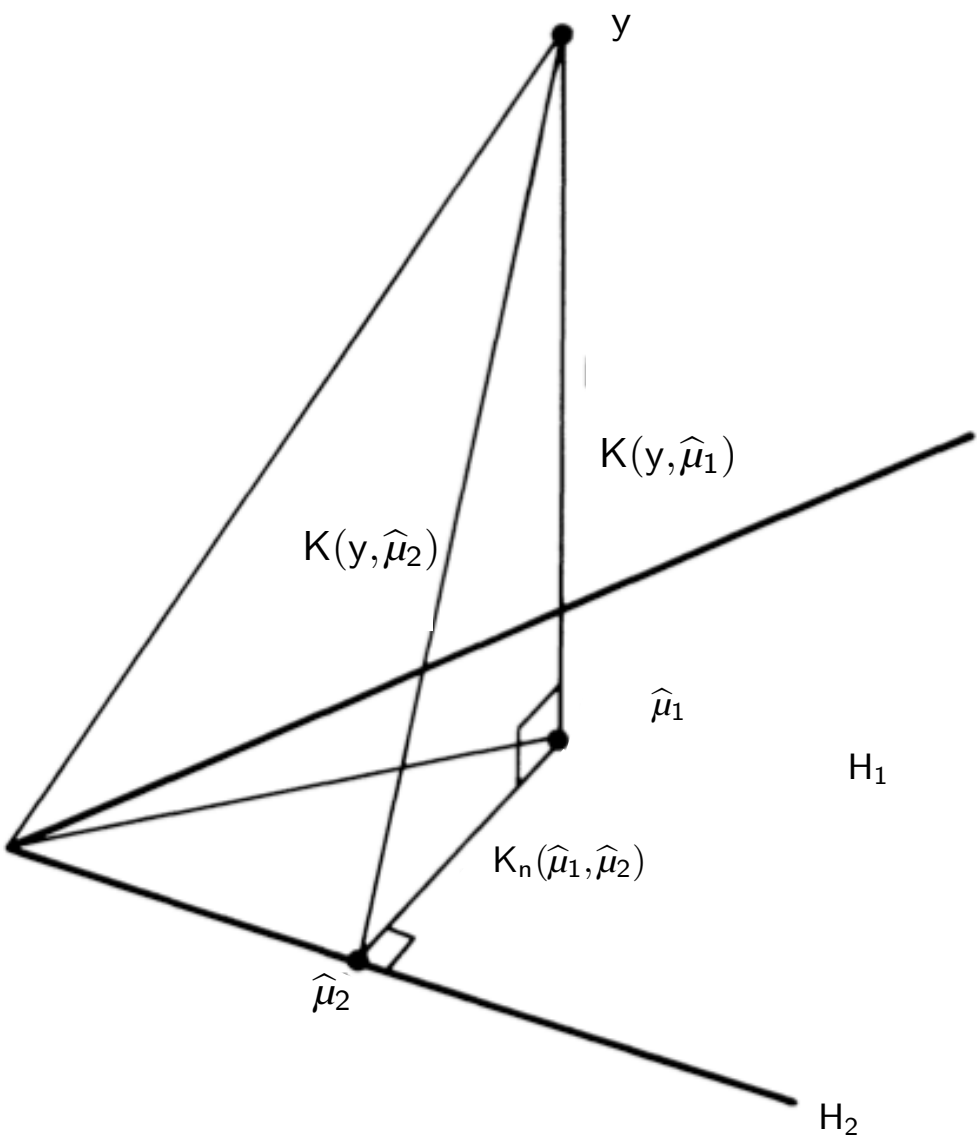
To apply this, consider a completely saturated model in which we exactly fit the data -- Then consider two regression models, one with q variables and one for which we've added more variables for a total of p (p > q)

Keeping with the notation previously (or at least subscript ordering), we'll let H_1 and H_2 denote the two underlying spaces with $H_2 \subset H_1$, and let $\widehat{\mu}_1$ and $\widehat{\mu}_2$ denote the conditional means associated with fits for these two models, the underlying regression coefficients having been chosen to minimize our divergence criterion

Then, using Simon's result we have

$$K_n(y, \widehat{\mu}_2) = K_n(y, \widehat{\mu}_1) + K_n(\widehat{\mu}_1, \widehat{\mu}_2)$$

Or in pictures...


y
$K(y,\widehat{\mu}_1)$
$K(y,\widehat{\mu}_2)$
$\widehat{\mu}_1$
H_1
$K_n(\widehat{\mu}_1,\widehat{\mu}_2)$
$\widehat{\mu}_2$
H_2

Distances

To relate this to something more familiar in terms of modeling methodology, suppose we have the normal linear model so that K refers to squared error -- Then, the relationship on the previous page can be reworded

In particular, if we let the larger model involve one or more predictors and the smaller involve just the intercept, we know that we can write

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \widehat{\mu}_i)^2 + \sum_i (\widehat{\mu}_i - \bar{y})^2$$

Or in table form

	SS	df
model	$\sum_i (\widehat{\mu}_i - \bar{y})^2$	$p - 1$
error	$\sum_i (y_i - \widehat{\mu}_i)^2$	$n - p$
total	$\sum_i (y_i - \bar{y})^2$	$n - 1$

Have we seen this before?

Distances

The Pythagorean relationship, then, the partitioning of “distances” in a sequence of projections, is at the heart of a generalization of ANOVA called ANODEV

The suffix DEV comes from the term “deviance” -- The divergence between a model with conditional means $\mu = (\mu_1, \ldots, \mu_n)^t$ and the “saturated” model that fits the data exactly (taking the conditional mean of the ith point to be y_i) is

$$D(y, \mu) = 2 K_n(y, \mu)$$

The deviance plays the role of the sum of squares in ANOVA

Distances

Using Simon's result on the previous slides, we can see that for fits using nested subspaces,

$$2\,K_n(\widehat{\mu}_1,\widehat{\mu}_2) = 2\,K_n(y,\widehat{\mu}_2) - 2\,K_n(y,\widehat{\mu}_1) = D(y,\widehat{\mu}_2) - D(y,\widehat{\mu}_1)$$

which tells us how much we have improved our fit by including the extra degrees of freedom that take us from $\widehat{\mu}_2$ to $\widehat{\mu}_1$

Deviance

A final comment on deviance -- With ANOVA and normal errors, we know that the sum of squares terms have chi-square distributions and F-tests were used from the table to assess the importance of each variable (where the order in which we entered terms was important in a non-orthogonal model)

The same can be said at least approximately for the terms in an ANODEV table -- That is, returning to the setup for Simon's results and recalling that $H_c \subset H_b \subset H_a$ with $\widehat{\mu}_c$, $\widehat{\mu}_b$ and $\widehat{\mu}_a$ being the minimum divergence estimates, we have

$$K_n(\widehat{\mu}_a, \widehat{\mu}_c) \sim \chi^2_{p-r}, \quad K_n(\widehat{\mu}_a, \widehat{\mu}_b) \sim \chi^2_{p-q}, \quad \text{and} \quad K_n(\widehat{\mu}_b, \widehat{\mu}_c) \sim \chi^2_{q-r}$$

where the distributions are only approximate

Final extension

To make a final connection between our distance measure, the sum of squares for the normal linear model and, to a lesser extent, the deviance (because it's the object usually referred to in the GLM literature), let's recall a result that led to our selection criterion Cp

Recall that by examining the expected value of RSS and comparing it to a theoretical quantity, we were able to come up with a measure of "optimism" that yielded an unbiased estimate of (in sample) prediction error

$$\mathrm{RSS}/n + 2p\widehat{\sigma}/n$$

A similar kind of reasoning can be followed for GLMs using our KL divergence measure in place of the sums of squares criterion

$$C_p = 2K_n(y,\widehat{\mu})/n + 2p\widehat{\phi}/n = D(y,\widehat{\mu})/n + 2p\widehat{\phi}/n$$

where $\widehat{\phi}$ is an estimate of the extra dispersion parameter mentioned at the beginning of these slides and is 1 for the binomial/Bernoulli family as well as the Poisson)

Example: Logistic regression

Suppose we have binary observations y_i (either 0 or 1) and predictors $x_{i1}, \ldots, x_{ip}$
-- We want to compare the distance from the data to a model for the mean

$$p(x_i) = \frac{e^{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}$$

where $p(x_i)$ is the probability that $y_i = 1$

Example: Logistic regression

In our distance measure, we are comparing this model to the so-called "saturated" model that fits the data exactly -- That is, uses the ith data point y_i as the conditional mean

So, if $y_i = 0$, then the saturated model assigns probability 0 to the value 1 and probability 1 to the value 0 -- Conversely, for $y_i = 1$, the saturate model assigns probability 1 to the value 1 and probability 0 to the value 0

If we let $s(x_i)$ denote the probability of seeing a 1 under the saturated model, then $s(x_i) = y_i$

Example: Logistic regression

Forming the KL divergence for these two distributions gives

$$\begin{aligned} & s(x_i)\log\frac{s(x_i)}{p(x_i)} + [1-s(x_i)]\log\frac{1-s(x_i)}{1-p(x_i)} \\ &= y_i\log\frac{1}{p(x_i)} + [1-y_i]\log\frac{1}{1-p(x_i)} \\ &= -y_i\log\frac{p(x_i)}{1-p(x_i)} \\ &= y_i\left(\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right) + \log\left(1+e^{\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}\right) \end{aligned}$$

or, after summing,

$$K_n(y,p) = \sum_i \left[y_i\left(\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right) + \log\left(1+e^{\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}\right)\right]$$

where p is the vector of conditional means $p = (p(x_1),\ldots,p(x_n))^t$