Regression Modeling


Recall that in Bayesian inference we build up an inferential model by specifying a prior and a likelihood.

$$
\pi_{S}(\theta \mid \widetilde{\mathcal{D}}) \propto \pi_{S}(\widetilde{\mathcal{D}} \mid \theta) \pi_{S}(\theta)
$$

## Recall that in Bayesian inference we build up an

 inferential model by specifying a prior and a likelihood.$$
\pi(\theta \mid \mathcal{D}) \propto \pi(\mathcal{D} \mid \theta) \pi \quad(\theta)
$$

## Likelihoods model the measurement process and are most naturally specified generatively.

$$
\pi\left(\mathcal{D} \mid \theta_{1}, \ldots, \theta_{4}\right)
$$

## Likelihoods model the measurement process and are most naturally specified generatively.

$$
\pi\left(\mathcal{D} \mid \theta_{1}, \ldots, \theta_{4}\right) \propto \quad \pi\left(\mathcal{D} \mid \theta_{1}\right)
$$

$$
\times \pi\left(\theta_{1} \mid \theta_{2}\right)
$$

$$
\times \pi\left(\theta_{2} \mid \theta_{3}\right)
$$

$$
\times \pi\left(\theta_{3} \mid \theta_{4}\right)
$$

This generative decomposition allows us to focus on modular modeling components.


$$
\pi\left(\theta_{n} \mid \theta_{n+1}\right)
$$



Many of the most common and useful modeling techniques are forms of regression.




## Foundations of Regression



## Often the data naturally separates into variates, $y$, and covariates, $x$.

$$
\mathcal{D} \rightarrow\{y, x\}
$$

## Regression models the statistical relationship between the variates and the covariates.



## Regression models the statistical relationship between the variates and the covariates.

$$
\pi(y, x \mid \theta)=\pi(y \mid x, \theta) \pi(x \mid \theta)
$$

## Regression models the statistical relationship between the variates and the covariates.

$$
\pi(y, x \mid \theta)=\pi(y \mid x, \theta) \pi(x \mid \theta)
$$

# We typically assume that the covariates are independent of the model parameters. 

$$
\pi(x \mid \theta)=\pi(x)
$$

## In which case the likelihood becomes a model of the variates conditional on the covariates.

$$
\pi(y, x \mid \theta)=\pi(y \mid x, \theta) \pi(x \mid \theta)
$$

## In which case the likelihood becomes a model of the variates conditional on the covariates.

$$
\begin{aligned}
& \pi(y, x \mid \theta)=\pi(y \mid x, \theta) \pi(x \mid \theta) \\
& \pi(y, x \mid \theta)=\pi(y \mid x, \theta) \pi(x)
\end{aligned}
$$

## In which case the likelihood becomes a model of the variates conditional on the covariates.

$$
\begin{aligned}
& \pi(y, x \mid \theta)=\pi(y \mid x, \theta) \pi(x \mid \theta) \\
& \pi(y, x \mid \theta)=\pi(y \mid x, \theta) \pi(x) \\
& \pi(y, x \mid \theta) \propto \pi(y \mid x, \theta)
\end{aligned}
$$

Covariates are often restricted to a single effective parameter through a deterministic mapping.

$$
\pi(y \mid x, \theta)=\pi\left(y \mid f\left(x, \theta_{1}\right), \theta_{2}\right)
$$

Covariates are often restricted to a single effective parameter through a deterministic mapping.

$$
\begin{aligned}
& \pi(y \mid x, \theta)=\pi\left(y \mid f\left(x, \theta_{1}\right), \theta_{2}\right) \\
& \pi(y \mid x, \theta)=\mathcal{N}(y \mid f(x, \theta), \sigma)
\end{aligned}
$$

Covariates are often restricted to a single effective parameter through a deterministic mapping.

$$
\begin{aligned}
& \pi(y \mid x, \theta)=\pi\left(y \mid f\left(x, \theta_{1}\right), \theta_{2}\right) \\
& \pi(y \mid x, \theta)=\mathcal{N}(y \mid f(x, \theta), \sigma) \\
& \pi(y \mid x, \theta)=\operatorname{Bin}(y \mid f(x, \theta), N)
\end{aligned}
$$

This immediately generalizes to multiple effective parameters.

$$
\pi(y \mid x, \theta)=\pi\left(y \mid f_{1}\left(x, \theta_{1}\right), f_{2}\left(x, \theta_{2}\right), \theta_{3}\right)
$$

This immediately generalizes to multiple effective parameters.

$$
\begin{gathered}
\pi(y \mid x, \theta)=\pi\left(y \mid f_{1}\left(x, \theta_{1}\right), f_{2}\left(x, \theta_{2}\right), \theta_{3}\right) \\
\pi(y \mid x, \theta)=\mathcal{G}\left(y \mid \alpha\left(x, \theta_{1}\right), \beta\left(x, \theta_{2}\right)\right)
\end{gathered}
$$

Linear Models


# When an effective parameter is unconstrained we can model it with a linear mapping. 

$$
f(x, \alpha, \beta)=\beta \cdot x+\alpha
$$

## Multiple covariates are commonly encapsulated in a design matrix.

$$
f(x, \alpha, \boldsymbol{\beta})=\sum_{n, i} X_{i n} \beta_{i}+\alpha
$$

## Multiple covariates are commonly encapsulated in a design matrix.



$$
f(x, \alpha, \boldsymbol{\beta})=\mathbf{X}^{T} \boldsymbol{\beta}+\alpha
$$

When the measurement model is Gaussian we recover the ubiquitous Gaussian-Linear model.


$$
\pi(y \mid \mathbf{X}, \alpha, \boldsymbol{\beta}, \sigma)=\mathcal{N}\left(y \mid \mathbf{X}^{T} \boldsymbol{\beta}+\alpha, \sigma\right)
$$

Given enough data, linear models are overconstrained and all the slopes can be fit well.


## Given enough data, linear models are overconstrained and all the slopes can be fit well.



When there are fewer data than covariates, however, linear models are subject to collinearity.


# In collinearity some of the slopes are fully determined while the others are completely undetermined. 



# Consequently (weakly) informative priors are critical for building robust linear models. 

$$
\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})
$$

Consequently (weakly) informative priors are
critical for building robust linear models.

$$
\beta_{i} \sim \mathcal{N}\left(\mu_{i}, \omega_{i}\right)
$$

# Consequently (weakly) informative priors are critical for building robust linear models. 

$$
\beta_{i} \sim \mathcal{N}(0, \omega)
$$

As with the linear model parameters, prior information for the Gaussian noise is critical.
$\pi(\sigma)=\operatorname{Half}-\mathrm{Cauchy}(0, \tau)$

As with the linear model parameters, prior information for the Gaussian noise is critical.


## General Linear Models



# Constrained effective parameters are not amenable to linear models. 

$$
\theta \in(a, b)
$$

$$
\mathbf{X}^{T} \boldsymbol{\beta}+\alpha \in(-\infty, \infty)
$$

# But we can generalize linear models with a link function. 

$$
\theta \in(a, b)
$$

$$
g\left(\mathbf{X}^{T} \boldsymbol{\beta}+\alpha\right) \in(a, b)
$$

## In the statistics literature link functions are defined by the un-constraining map.

$$
g^{-1}:(a, b) \rightarrow(-\infty, \infty)
$$

## While bounded parameters are modeled with the logit link function.

$$
\operatorname{logit}:(0,1) \rightarrow(-\infty, \infty)
$$

$$
\operatorname{logistic}\left(\mathbf{X}^{T} \boldsymbol{\beta}+\alpha\right) \in(0,1)
$$

While bounded parameters are modeled with the logit link function.


Success/failure data subject to covariates can be modeled with generalized binomial/Bernoulli models.

$$
\begin{aligned}
& \pi(y \mid \mathbf{X}, \alpha, \boldsymbol{\beta})= \\
& \quad \operatorname{Ber}\left(y \mid \operatorname{logistic}\left(\mathbf{X}^{T} \boldsymbol{\beta}+\alpha\right)\right)
\end{aligned}
$$

## Positive parameters are modeled with the $\log$ link function.

$$
\log :(0, \infty) \rightarrow(-\infty, \infty)
$$

$$
\exp \left(\mathbf{X}^{T} \boldsymbol{\beta}+\alpha\right) \in(0, \infty)
$$

## Positive parameters are modeled with the $\log$ link function.



Count data whose rate depends on covariates can be modeled with a generalized Poisson model.

$$
\begin{aligned}
& \pi(y \mid \mathbf{X}, \alpha, \boldsymbol{\beta})= \\
& \quad \operatorname{Poisson}\left(y \mid \exp \left(\mathbf{X}^{T} \boldsymbol{\beta}+\alpha\right)\right)
\end{aligned}
$$

## In some applications the Poisson likelihood is too restrictive.

$\operatorname{Poisson}(y \mid \lambda)$

$$
\begin{gathered}
\mathbb{E}[y]=\lambda \\
\operatorname{Var}[y]=\lambda
\end{gathered}
$$

# But we can incorporate overdispersion with a generalized negative binomial model. 

$\operatorname{Poisson}(y \mid \lambda)$

$$
\mathbb{E}[y]=\lambda
$$

$$
\operatorname{Var}[y]=\lambda
$$

$$
\operatorname{NegBin} 2(y \mid \mu, \phi)
$$

$$
\mathbb{E}[y]=\mu
$$

$$
\operatorname{Var}[y]=\mu+\mu^{2} / \phi
$$

But we can incorporate overdispersion with a generalized negative binomial model.


But we can incorporate overdispersion with a generalized negative binomial model.


But we can incorporate overdispersion with a generalized negative binomial model.


## Gaussian Processes



Regression is always limited by our assumption of a functional relationship between variates and covariates.


Regression is always limited by our assumption of a functional relationship between variates and covariates.


Regression is always limited by our assumption of a functional relationship between variates and covariates.


Nonparametric regression attempts to overcome this limitation by modeling entire spaces of functions.


Nonparametric regression attempts to overcome this limitation by modeling entire spaces of functions.


Gaussian processes are probability distributions over not spaces of parameters but rather spaces of functions.

$$
f \sim \mathcal{G} \mathcal{P}\left(\mu(x), k\left(x, x^{\prime}\right)\right)
$$

Gaussian processes are probability distributions over not spaces of parameters but rather spaces of functions.


Gaussian processes are probability distributions over not spaces of parameters but rather spaces of functions.


Gaussian processes are probability distributions over not spaces of parameters but rather spaces of functions.


Gaussian processes are probability distributions over not spaces of parameters but rather spaces of functions.


The covariance function defines the smoothness of functions supported by a Gaussian process.

$$
k\left(x_{1}, \ldots, x_{N}, x_{1}^{\prime}, \ldots, x_{N}^{\prime}\right)=\theta_{1} \exp \left(-\sum_{n=1}^{N} \theta_{n}\left(x_{n}-x_{n}^{\prime}\right)^{2}\right)
$$

A Gaussian likelihood is conjugate to a Gaussian process prior, yielding a Gaussian process posterior.

$$
f \sim \mathcal{G} \mathcal{P}(\mu, k)
$$

A Gaussian likelihood is conjugate to a Gaussian process prior, yielding a Gaussian process posterior.

$$
f \sim \mathcal{G} \mathcal{P}(\mu, k)
$$

$$
y \sim \mathcal{N}(f(x), \sigma)
$$

A Gaussian likelihood is conjugate to a Gaussian process prior, yielding a Gaussian process posterior.

$$
\begin{gathered}
f \sim \mathcal{G P}(\mu, k) \\
y \sim \mathcal{N}(f(x), \sigma) \\
f \sim \mathcal{G P}\left(\mu, k+\sigma^{2} \mathbb{I}\right)
\end{gathered}
$$

A Gaussian likelihood is conjugate to a Gaussian process prior, yielding a Gaussian process posterior.


A Gaussian likelihood is conjugate to a Gaussian process prior, yielding a Gaussian process posterior.


A Gaussian likelihood is conjugate to a Gaussian process prior, yielding a Gaussian process posterior.


A Gaussian likelihood is conjugate to a Gaussian process prior, yielding a Gaussian process posterior.


We can incorporate Gaussian processes into a complex likelihood by exploiting its marginalization properties.

$$
f(x) \sim \mathcal{G P}\left(\mu(x), k\left(x, x^{\prime}\right)\right)
$$

We can incorporate Gaussian processes into a complex likelihood by exploiting its marginalization properties.

$$
f(x) \sim \mathcal{G} \mathcal{P}\left(\mu(x), k\left(x, x^{\prime}\right)\right)
$$

$$
\begin{gathered}
y \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\mu_{i}=\mu\left(\mathbf{x}_{i}\right) \\
\Sigma_{i j}=k\left(\mathbf{x}_{i}, \mathbf{x}_{j}\right)
\end{gathered}
$$



