

# **STSCI 4780**

## **Relationships between variables: Regression**

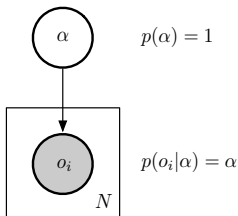
Tom Loredo, CCAPS & SDS, Cornell University

© 2020-04-14

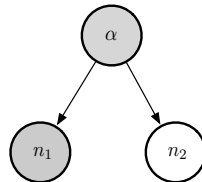
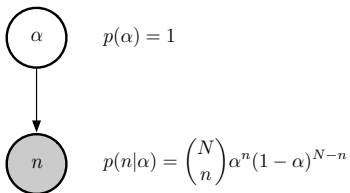
# Recap via DAGs

## Univariate data DAGs

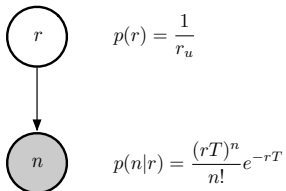
### Bernoulli outcomes



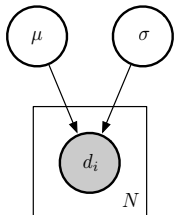
### Binomial counts — estimation, prediction



### Poisson counts



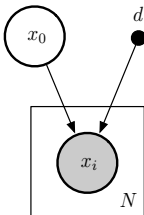
### Normal mean and standard deviation



$$p(\mu) = C$$
$$p(\sigma) \propto \frac{1}{\sigma}$$

$$p(d_i|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(d_i - \mu)^2}{2\sigma^2}}$$

### Cauchy location estimation



$$p(x_0) = C$$

$$p(x_i|x_0, d) = \frac{1}{\pi d} \frac{1}{1 + \left(\frac{x - x_0}{d}\right)^2}$$

## Common features

- *Univariate data*—binary, integer or real *scalar* samples
- *Conditionally independent* data

Most problems also univariate in parameter space

Repeated sampling problems were IID

These were all univariate parametric distribution estimation problems:

- Discrete data: Parametric PMF estimation
- Continuous data: Parametric density (PDF) estimation

## Bayes's theorem and DAGs

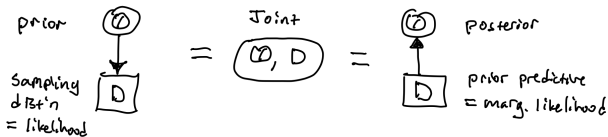
BT in terms of the joint dist'n for params + data:



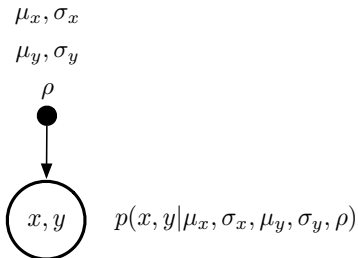
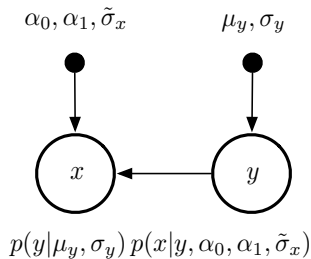
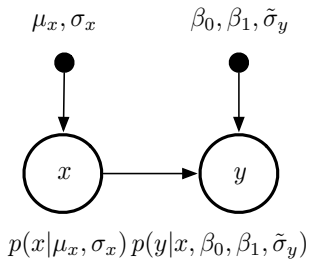
$$p(\theta, D) = p(\theta) p(D|\theta) \quad \leftarrow \text{Read off a DAG}$$

$$p(\theta|D) = \frac{p(\theta, D)}{p(D)} \propto \text{Joint}$$

BT as a "DAG equation":



## Bivariate normal sampling dist'n DAGs



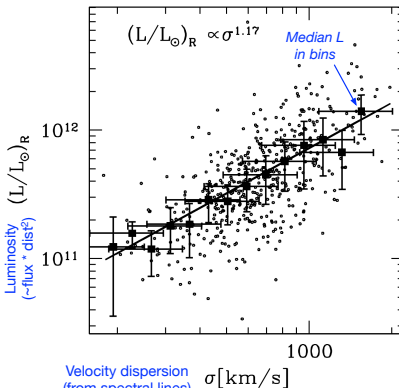
## Population studies

## Faber-Jackson relation for elliptical galaxies

**DIAGRAM BASED ON TABLE I.**  
(all female heights are multiplied by 108.)

MID-PARENTS		ADULT CHILDREN their Heights, and Deviations from 68 inches.												
Heights in inches	Deviates in inches	64	65	66	67	68	69	70	71	72	73			
		-4	-3	-2	-1	0	+1	+2	+3	+4				
72														
71	+3					Y		N						
70	+2													
69	+1													
68	0					O								
67	-1													
66	-2													

Galton (1885) "Regression Towards Mediocrity  
in Hereditary Stature"



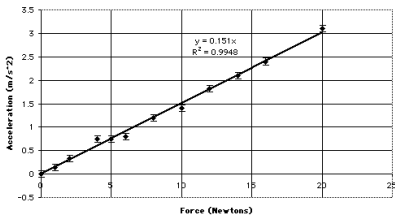
Bernardi+ (2002)

# Lec16 recap: Examples with deterministic $\times$

## Curve fitting

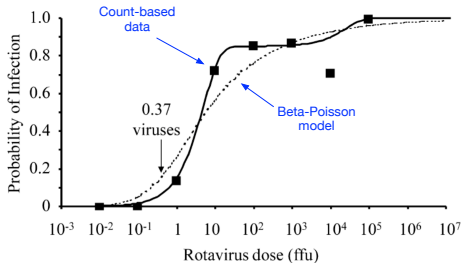
Newton's 2nd law:  $a = \frac{F}{m}$

*Apply different forces to a fixed mass*



Batesville HS AP Physics Class

Dose-response curve

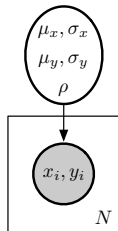
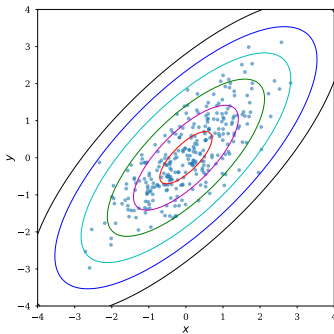


Gale (2003), "Developing risk assessments of waterborne microbial contaminations"



# Bivariate normal inference problems

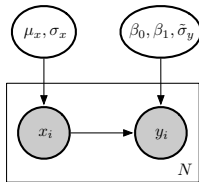
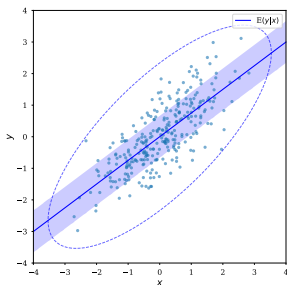
## *BVN density estimation*



$$\text{Joint: } p(\mu_x, \sigma_x, \mu_y, \sigma_y, \rho) \prod_{i=1}^N p(x_i, y_i | \mu_x, \sigma_x, \mu_y, \sigma_y, \rho)$$

$$\text{Inference: } p(\mu_x, \sigma_x, \mu_y, \sigma_y, \rho | \{x_i, y_i\}) \propto \text{Joint}$$

## BVN regression



$$\text{Joint: } p(\mu_x, \sigma_x) p(\beta_0, \beta_1, \tilde{\sigma}_y) \prod_{i=1}^N p(x_i | \mu_x, \sigma_x) p(y_i | x_i, \beta_0, \beta_1, \tilde{\sigma}_y)$$

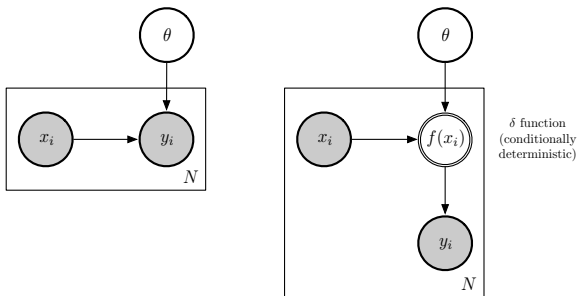
$$\begin{aligned} \text{Inference: } p(\beta_0, \beta_1, \tilde{\sigma}_y | \{x_i, y_i\}) &= \int d\mu_x \int d\sigma_x p(\mu_x, \sigma_x, \beta_0, \beta_1, \tilde{\sigma}_y | \{x_i, y_i\}) \\ &\propto p(\beta_0, \beta_1, \tilde{\sigma}_y) \prod_{i=1}^N p(y_i | x_i, \beta_0, \beta_1, \tilde{\sigma}_y) \end{aligned}$$

Note that the  $x_i$  marginal,  $p(x_i | \mu_x, \sigma_x)$ , *plays no role* in inferring the regression line (the *observed*  $x_i$  values of course play a strong role)

# Parametric regression

*Infer  $\theta$*  determining the *conditional expectation*

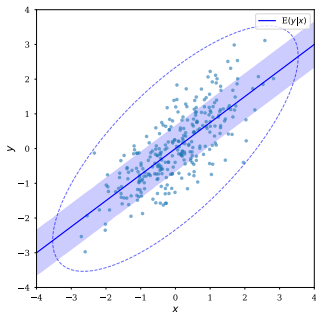
$$\mathbb{E}(y_i | x_i, \theta) = f(x_i; \theta)$$



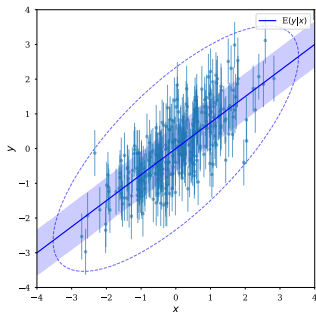
Often natural to express this via an additive error model:

$$y_i = f(x_i; \theta) + \epsilon_i; \quad \mathbb{E}(\epsilon_i) = 0$$

Population perspective



True+error perspective



Band displays the conditional prediction uncertainty for  $y$  at a given  $x$

**Error bars:** *Data* do not have errors, only *inferences* have errors:

- Error bars often signify prediction uncertainty (width of the sampling dist'n)—they belong on predictions, not observations
- View as likely range of error if one naively considers each  $y_i$  as an estimate of  $f(x_i)$  (disregarding other data!)

## Simple normal linear regression

A function-plus-error model (interpretable either as “true value plus noise” or “typical value plus dispersion”) for scalar  $x$  and  $y$ :

$$y_i = f(x_i; \theta) + \epsilon_i; \quad \epsilon_i \sim \text{Norm}(0; \sigma^2); \quad i = 1 \text{ to } N$$

$$f(x; \theta) = \sum_{\alpha=1}^M A_{\alpha} g_{\alpha}(x)$$

for *specified* set of basis functions,  $g_{\alpha}(x)$

- Parameters are  $M$  coefficients/amplitudes:  $\theta = \{A_{\alpha}\}$
- Regression function is *linear wrt  $A_{\alpha}$*  (not necessarily wrt  $x$ !)
- $M$  *basis functions*  $g_{\alpha}(x)$ 
  - Polynomials:  $\{1, x, x^2, \dots\}$  (or orthogonal polynomials)
  - Sinusoids/Fourier series:  $\{\sin(\omega x), \cos(\omega x), \dots\}$   
(with  $\omega$  fixed/known)
- PDFs for errors are *normal* (here IID), with *known*  $\sigma$

## Generalizations

- This is the *homoskedastic* case; *heteroskedastic* has variances  $\sigma_i^2$ ; more generally the errors could have a *non-diagonal covariance matrix*
- *Multiple linear regression* generalizes to multiple explanatory variables; i.e.,  $x_i \rightarrow \mathbf{x}_i$ , a vector
- *General linear models* generalize to a vector response,  $\mathbf{y}_i$
- *Generalized linear models* assume a linear model for a *nonlinear function of the conditional expectation*:

$$\ell(\mathbb{E}(y_i | \mathbf{x}_i, \theta)) = \sum_{\alpha=1}^M A_{\alpha} g_{\alpha}(x)$$

$\ell(y)$  is the *link function*, e.g.,  $\log(y)$  (Poisson regression for count data),  $\log(y/(1 - y))$  (logistic regression for binary responses)

## Likelihood function

Abbreviating  $f_i = f(x_i; \{A_\alpha\}) = f_i(\{A_\alpha\})$ ,

$$\begin{aligned} p(\{y_i\}|\{x_i\}, \{A_\alpha\}) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - f_i)^2 \right] \\ &= \frac{1}{\sigma^N (2\pi)^{N/2}} e^{-Q/2\sigma^2} \end{aligned}$$

$$\begin{aligned} Q(\{A_\alpha\}) &= \sum_{i=1}^N (y_i - f_i)^2 \\ &= \sum_{i=1}^N \left( y_i - \sum_{\alpha=1}^M A_\alpha g_{\alpha i} \right)^2 \quad \text{with } g_{\alpha i} \equiv g_\alpha(x_i) \\ &= \sum_{i=1}^N y_i^2 + \sum_{i=1}^N \left( \sum_{\alpha=1}^M A_\alpha g_{\alpha i} \right)^2 - 2 \sum_{i=1}^N y_i \sum_{\alpha=1}^M A_\alpha g_{\alpha i} \end{aligned}$$

## Vector notation

Eliminate Roman (data) indices by denoting such quantities as  $N$ -vectors:  $\vec{f} = [f_1, \dots, f_N]^T$ , etc.

Let  $\vec{u} \cdot \vec{v} \equiv \sum_i u_i v_i = \vec{v} \cdot \vec{u}$  (dot product, symmetric), and  $u^2 \equiv \vec{u} \cdot \vec{u} = \sum_i u_i^2$  (vector squared magnitude)

Model expresses  $\vec{f}$  as a sum of  $M$  basis vectors:

$$\vec{y} = \vec{f}(\{A_\alpha\}) + \vec{\epsilon}; \quad \vec{f}(\{A_\alpha\}) = \sum_{\alpha=1}^M A_\alpha \vec{g}_\alpha$$

Quadratic form is the squared magnitude of the misfit vector:

$$\begin{aligned} Q(\{A_\alpha\}) &= \left[ \vec{y} - \vec{f}(\{A_\alpha\}) \right]^2 \\ &= y^2 + f^2 - 2\vec{y} \cdot \vec{f} \\ &= y^2 + \sum_{\alpha\beta} A_\alpha A_\beta \vec{g}_\alpha \cdot \vec{g}_\beta - 2 \sum_{\alpha} A_\alpha \vec{y} \cdot \vec{g}_\alpha \end{aligned}$$



## Posterior mode

Adopt a flat prior; the posterior mode is then the maximum likelihood estimate, which satisfies (for  $\gamma = 1$  to  $M$ )

$$\left. \frac{\partial Q}{\partial A_\gamma} \right|_{A=\hat{A}} = 2 \sum_{\beta} \hat{A}_\beta \vec{g}_\beta \cdot \vec{g}_\gamma - 2 \vec{y} \cdot \vec{g}_\gamma = 0$$
$$\Rightarrow \sum_{\beta} \hat{A}_\beta \vec{g}_\beta \cdot \vec{g}_\gamma = \vec{y} \cdot \vec{g}_\gamma$$

Let  $\hat{\vec{f}} \equiv \sum_{\beta} \hat{A}_\beta \vec{g}_\beta$  (function estimate at the mode); then

$$\hat{\vec{f}} \cdot \vec{g}_\gamma = \vec{y} \cdot \vec{g}_\gamma$$

I.e., for the *maximum a posteriori* (MAP) model,

*The model's projection on each basis function  
matches the data's projection on each basis function*

## Regression geometry

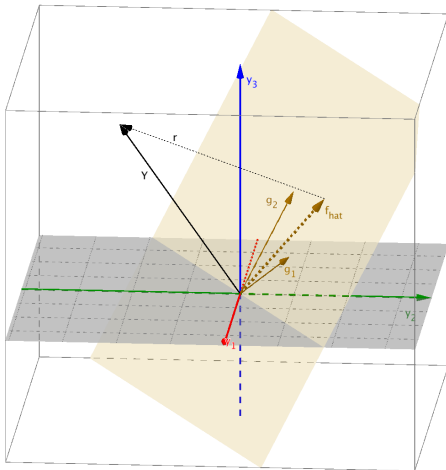
Geometry for linear regression,  
 $M = 2$  bases,  $N = 3$  samples

$$\vec{x} = [0, 1, 2]^T; \quad \vec{y} = [3, -2, 4]^T$$

$$f(x) = A_1 + A_2 x$$

$$g_1(x) = 1 \rightarrow \vec{g}_1 = [1, 1, 1]^T$$

$$g_2(x) = x \rightarrow \vec{g}_2 = [0, 1, 2]^T$$



Produced with GeoGebra Classic 5  
See "LinearModelVectors.ggb"

See: [GeoGebra.org](https://www.geogebra.org)

## Metric and mode equation

To solve for the mode, define the matrix  $\eta_{\alpha\beta}$ :

$$\eta_{\alpha\beta} \equiv \vec{g}_\alpha \cdot \vec{g}_\beta = \eta_{\beta\alpha}$$

This is a symmetric matrix; it plays the role of a metric on the  $M$ -dimensional subspace spanned by the model functions

The mode condition is now (switch  $\gamma$  to  $\alpha$  & use symmetry)

$$\sum_{\beta} \eta_{\alpha\beta} \hat{A}_{\beta} = \vec{y} \cdot \vec{g}_{\alpha}$$

The LHS describes the product of an  $M \times M$  matrix and a column vector of  $M$  components; the RHS comprises a vector of  $M$  components—this is just a matrix equation (in the  $M$ -D model space, not the  $N$ -D sample space)

$$\Rightarrow \hat{A}_{\alpha} = \sum_{\beta} [\eta^{-1}]_{\alpha\beta} \vec{y} \cdot \vec{g}_{\beta}$$

(numerically, the most stable solvers backsolve rather than invert)

### *Aside on metrics*

A metric defines dot products in terms of coordinates in an arbitrary (e.g., non-orthonormal) basis:

$$\vec{v}_1 = \sum_{\alpha} a_{\alpha} \vec{g}_{\alpha}$$

$$\vec{v}_2 = \sum_{\beta} b_{\beta} \vec{g}_{\beta}$$

$$\begin{aligned} \Rightarrow \vec{v}_1 \cdot \vec{v}_2 &= \sum_{\alpha} \sum_{\beta} a_{\alpha} b_{\beta} \vec{g}_{\alpha} \cdot \vec{g}_{\beta} \\ &= \sum_{\alpha} \sum_{\beta} a_{\alpha} b_{\beta} \eta_{\alpha\beta} \end{aligned}$$

(If the  $\vec{g}_{\alpha}$  were orthogonal, only the  $\alpha = \beta$  terms would be nonzero)

Key use is finding distance from coordinate differences:

$$\begin{aligned}d_{12}^2 &= |\vec{b} - \vec{a}|^2 \\&= \left[ \sum_{\alpha} (b_{\alpha} - a_{\alpha}) \vec{g}_{\alpha} \right] \cdot \left[ \sum_{\beta} (b_{\beta} - a_{\beta}) \vec{g}_{\beta} \right] \\&= \sum_{\alpha} \sum_{\beta} \Delta_{\alpha} \Delta_{\beta} \eta_{\alpha\beta}\end{aligned}$$

with coordinate differences  $\Delta_{\alpha} \equiv b_{\alpha} - a_{\alpha}$

In an *orthonormal* basis,  $\eta_{\alpha\beta} = \delta_{\alpha\beta}$ , the identity matrix; in this case

$$d_{12}^2 = \sum_{\alpha} \Delta_{\alpha}^2,$$

i.e., the *Pythagorean theorem*

Metrics generalize the Pythagorean theorem to non-orthonormal coordinate systems

# Connections to least squares estimation

For a flat prior and fixed  $\sigma$ , the posterior mode minimizes

$$Q(\{A_\alpha\}) = \sum_{i=1}^N [y_i - f_i(\{A_\alpha\})]^2$$

→ the flat-prior mode gives the *least squares estimates of the amplitudes*

The  $N \times M$  matrix of model vector coordinates  $[g_{\alpha i}]^T$  is the *design matrix*; it is often denoted  $\mathbf{X} = X_{i\alpha}$ , even though it consists of *response* values (the model basis in the  $y$  space—functions of  $x_i$ s)

The  $M \times M$  metric

$$\eta_{\alpha\beta} \equiv \vec{g}_\alpha \cdot \vec{g}_\beta = \sum_i g_{\alpha i} g_{\beta i} = \mathbf{X}^T \mathbf{X}$$

is sometimes called the *Gramian matrix* (or *Gram matrix*)

The mode condition

$$\sum_\beta \eta_{\alpha\beta} \hat{A}_\beta = \vec{y} \cdot \vec{g}_\alpha$$

is a set of  $M$  equations called the *normal equations* when expressed in terms of the design matrix