Parametric Inference & Linear Modeling

Statistics for Data Science
CSE357 - Fall 2021

Goal of Parametric Inference

Goal: Estimate parameters, θ, for some function:

$$f(x; \theta) : \theta \in \Theta$$

$$\Theta \subset \mathbb{R}^k \qquad \theta = (\theta_1, \theta_2, ..., \theta_k)$$

Goal of Parametric Inference

Goal: Estimate parameters, θ, for some function:

$$f(x; \theta) : \theta \in \Theta$$

$$\Theta \subset \mathbb{R}^k \qquad \theta = (\theta_1, \theta_2, ..., \theta_k)$$

once θ is set, then f can be applied to any x

Goal of Parametric Inference

Goal: Estimate parameters, θ, for some function:

$$f(x; \theta) : \theta \in \Theta$$

$$\Theta \subset \mathbb{R}^k \qquad \theta = (\theta_1, \theta_2, ..., \theta_k)$$

Example models to apply parametric inference:

1) PDFs (f: x -> density) and PMFs (f: x -> probability)
2) Linear Models: f: x -> y in some form of $y = mx + b$
   (where m and b are the parameters)

once θ is set, then f can be applied to any x

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes L?

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = log\, L(\theta) = log \prod_{i=1}^{n} f(X_i; \theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes L?

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\ f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes L?

Example: $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p), then $f(x;p) = p^x(1 - p)^{1-x}$, *for* $x = 0, 1$.

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\ f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes L?

Example: $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p), then $f(x;p) = p^x(1 - p)^{1-x}$, *for x = 0, 1.*

$$L_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}, \text{where } S = \sum_i X_i$$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\ f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes L?

Example: $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p), then $f(x;p) = p^x(1 - p)^{1-x}$, *for x = 0, 1.*

$$L_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}, \text{ where } S = \sum_i X_i$$

$$l_n(p) = S\, log\, p + (n - s)\, log(1 - p)$$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\ f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes L?

Example: $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p), then $f(x;p) = p^x(1 - p)^{1-x}$, *for x = 0, 1.*

$$L_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}, \text{ where } S = \sum_i X_i$$

$$l_n(p) = S\,log\,p + (n - s)\,log(1 - p)$$

GOAL: take the derivative with respect to p and set to 0:

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\ f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes L?

Example: $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p), then $f(x;p) = p^x(1 - p)^{1-x}$, *for x = 0, 1.*

$$L_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}, \text{ where } S = \sum_i X_i$$

$$l_n(p) = S\, log\, p + (n - s)\, log(1 - p)$$

GOAL: take the derivative and set to 0 to find:

$$\boxed{\hat{p} = \frac{S}{n}}$$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Thus, $L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} =$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Thus, $L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim$ Normal(μ, σ^2), then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Thus, $L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right]$$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim$ Normal(μ, σ^2), then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Thus, $L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$l(\mu, \sigma^2) = log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right] =$ $+$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim$ Normal(μ, σ^2), then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Thus, $L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$l(\mu, \sigma^2) = log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right] = n\, log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + log \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} =$ +

=

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim$ Normal(μ, σ^2), then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Thus, $L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right] = n\, log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + log \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = n\, log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{i=1}^{n} log\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$=$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim$ Normal(μ, σ^2), then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Thus, $L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right] = n\, log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + log \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = n\, log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{i=1}^{n} log\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= n\, log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} =$$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim$ Normal(μ, σ^2), then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Thus, $L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = log\left[\left(\tfrac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right] = n\, log\left(\tfrac{1}{\sqrt{2\pi\sigma^2}}\right) + log \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = n\, log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{i=1}^{n} log\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= n\, log\left(\tfrac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} = -n\, log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Maximum Likelihood Estimation

Given data and a distribution, how does one choose the parameters?

likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} log\, f(X_i; \theta)$$

maximum likelihood estimation: What is the θ that maximizes *L*?

Example: $X \sim$ Normal(μ, σ^2), then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Thus, $L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right] = n\, log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + log \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = n\, log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{i=1}^{n} log\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= n\, log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} = -n\, log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2$$

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n\, log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n \log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

take the partial derivative with respect to μ and set equal to 0

take the derivative with respect to σ^2 and set equal to 0

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n \log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

take the partial derivative with respect to μ and set equal to 0

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0$$

take the derivative with respect to σ^2 and set equal to 0

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n \log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

take the partial derivative with respect to μ and set equal to 0

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad \sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad n\hat{\mu} = \sum_{i=1}^{n} x_i$$

take the derivative with respect to σ^2 and set equal to 0

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n\, log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

take the partial derivative with respect to μ and set equal to 0

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad \sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad n\hat{\mu} = \sum_{i=1}^{n} x_i$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

take the derivative with respect to σ^2 and set equal to 0

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n \log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

take the partial derivative with respect to μ and set equal to 0

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad \sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad n\hat{\mu} = \sum_{i=1}^{n} x_i$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

take the derivative with respect to σ^2 and set equal to 0

$$\frac{\partial l}{\partial \sigma^2} =$$

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n\, log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{n}log\, 2\pi - \frac{n}{2}log\, \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

take the partial derivative with respect to μ and set equal to 0

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad \sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad n\hat{\mu} = \sum_{i=1}^{n} x_i$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

take the derivative with respect to σ^2 and set equal to 0

$$\frac{\partial l}{\partial \sigma^2} =$$

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n\, log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 = \cancel{\frac{1}{n} log\, 2\pi} - \frac{n}{2} log\, \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

take the partial derivative with respect to μ and set equal to 0

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad \sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad n\hat{\mu} = \sum_{i=1}^{n} x_i$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

take the derivative with respect to σ^2 and set equal to 0

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n\,log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 = \cancel{\frac{1}{n}log\,2\pi} - \frac{n}{2}log\,\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

take the partial derivative with respect to μ and set equal to 0

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad \sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad n\hat{\mu} = \sum_{i=1}^{n} x_i$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

take the derivative with respect to σ^2 and set equal to 0

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n \log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 = \cancel{\frac{1}{n}\log 2\pi} - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

take the partial derivative with respect to μ and set equal to 0

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad \sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad n\hat{\mu} = \sum_{i=1}^{n} x_i$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

MLE estimate of sample mean

take the derivative with respect to σ^2 and set equal to 0

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

MLE estimate of sample variance

Maximum Likelihood Estimation

Example: $X \sim \text{Normal}(\mu, \sigma^2)$, then $f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$l(\mu, \sigma^2) = -n \log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 = \cancel{\frac{1}{n}\log 2\pi} - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

take the partial derivative with respect to μ and set equal to 0

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad \sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \qquad n\hat{\mu} = \sum_{i=1}^{n} x_i$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

MLE estimate of sample mean

take the derivative with respect to σ^2 and set equal to 0

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

biased!

MLE estimate of sample variance

Maximum Likelihood Estimation

Try yourself:

Example: $X \sim$ Exponential(λ),

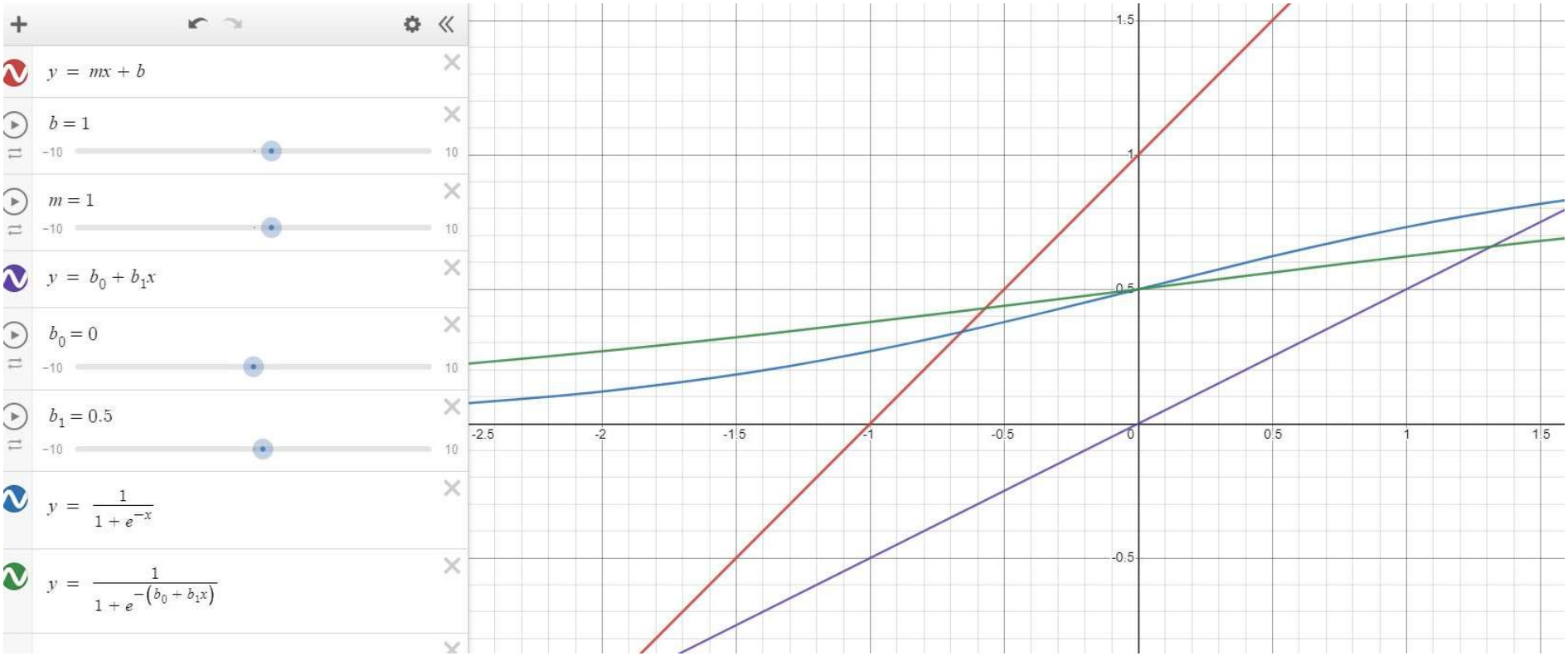
hint: should arrive at something almost familiar; then recall $\lambda = \frac{1}{\beta}$

Linear Models

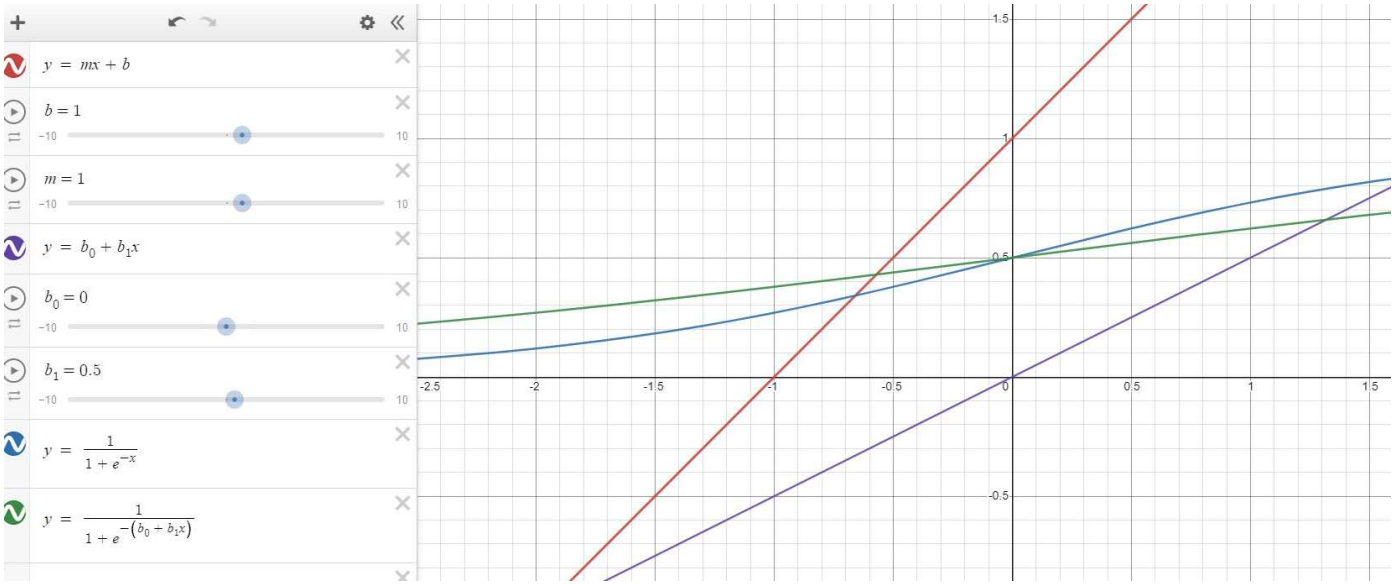
$$y = mx + b$$

Linear Models

$$y = mx + b$$

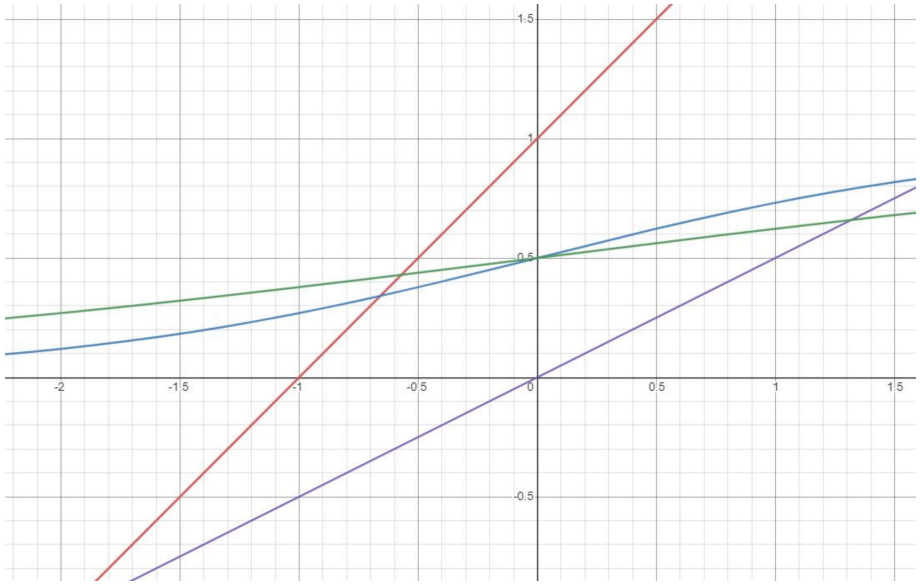


Linear Models



$$y = mx + b$$

Linear Models



- Logistic Regression
- Linear Regression
- Pearson Product-Moment Correlation
- Multiple Linear/Logistic Regression

$$y = mx + b$$

logistic regression model

MLE

interpreting coefficients beta, etc…

t-test

gradient ascent version

linear regression

Logistic Regression

$Y_i \in \{0, 1\}$; X is a **single value** and can be anything numeric.

$$P(Y_i = 1 | X_i = x) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$= \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij})}}$$

Logistic Regression

$Y_i \in \{0, 1\}$; X_i is a **single value** and can be anything numeric.

$$P(Y_i = 1 | X_i = x) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

The goal of this function is to: take in the variable *x* and return a probability that *Y* is 1.

Logistic Regression

$Y_i \in \{0, 1\}$; X can be anything numeric.

$$P(Y_i = 1 | X_i = x) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

The goal of this function is to: take in the variable *x* and return a probability that *Y* is 1.

Note that there are only three variables on the right: X_i, B_0, B_1

Logistic Regression

$Y_i \in \{0, 1\}$; X can be anything numeric.

$$P(Y_i = 1 | X_i = x) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

The goal of this function is to: take in the variable *x* and return a probability that *Y* is 1.

Note that there are only three variables on the right: X_i, B_0, B_1

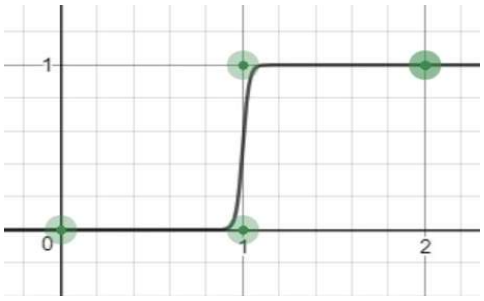
X is given. B_0 and B_1 must be learned.

Logistic Regression

$Y_i \in \{0, 1\}$; X can be anything numeric.

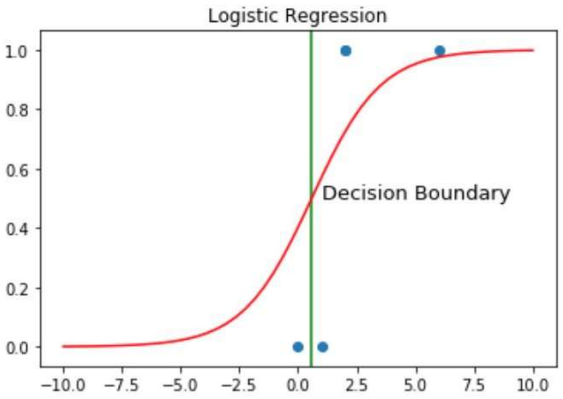
$$P(Y_i = 1 | X_i = x) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

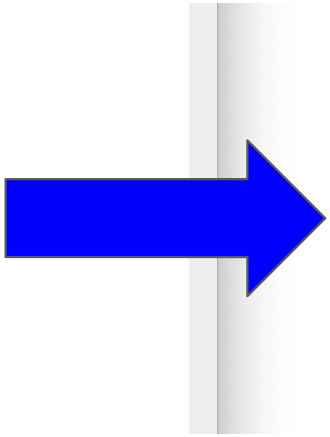
HOW? Essentially, try different B_0 and B_1 values until "best fit" to the training data (example X and Y).

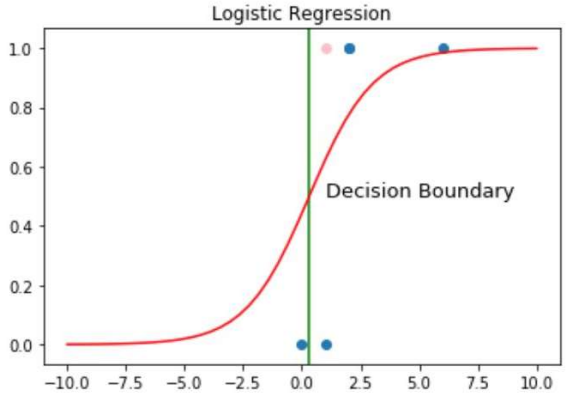


X is given. B_0 and B_1 must be learned.

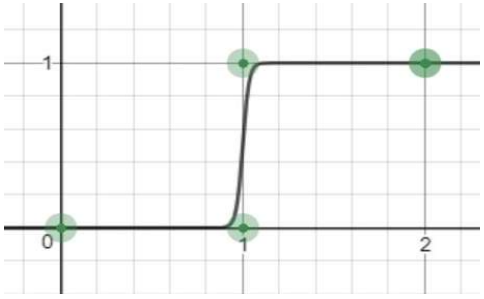
Logistic Regression







HOW? Essentially, try different B_0 and B_1 values until "best fit" to the training data (example X and Y).



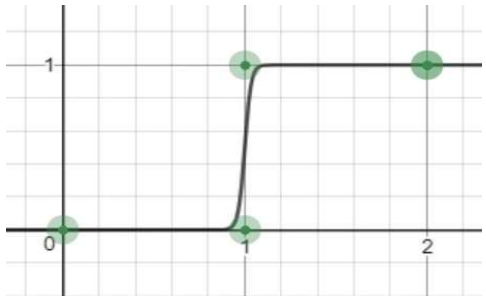
X is given. B_0 and B_1 must be <u>learned</u>.

"best fit" : whatever maximizes the Bernouli likelihood function:

$$L(\beta_0, \beta_1 | X, Y) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

$$P(Y_i = 1 | X_i = x) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

HOW? Essentially, try different B_0 and B_1 values until "best fit" to the training data (example X and Y).



X is given. B_0 and B_1 must be learned.

“best fit” : whatever maximizes the Bernouli likelihood function:

$$L(\beta_0, \beta_1 | X, Y) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

“best fit” : more efficient to maximize *log likelihood* :

“best fit” : whatever maximizes the Bernouli likelihood function:

$$L(\beta_0, \beta_1 | X, Y) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

“best fit” : more efficient to maximize *log likelihood* :

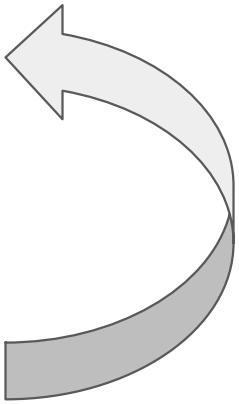
$$\ell(\beta) = \sum_{i=1}^{N} y_i log\, p(x_i) + (1-y_i) log\, (1-p(x_i))$$

"best fit" : minimize *log loss*:

$$J(\square) = -\sum_{i=1}^{N} y_i log\, p(x_i) + (1-y_i) log\, (1-p(x_i))$$

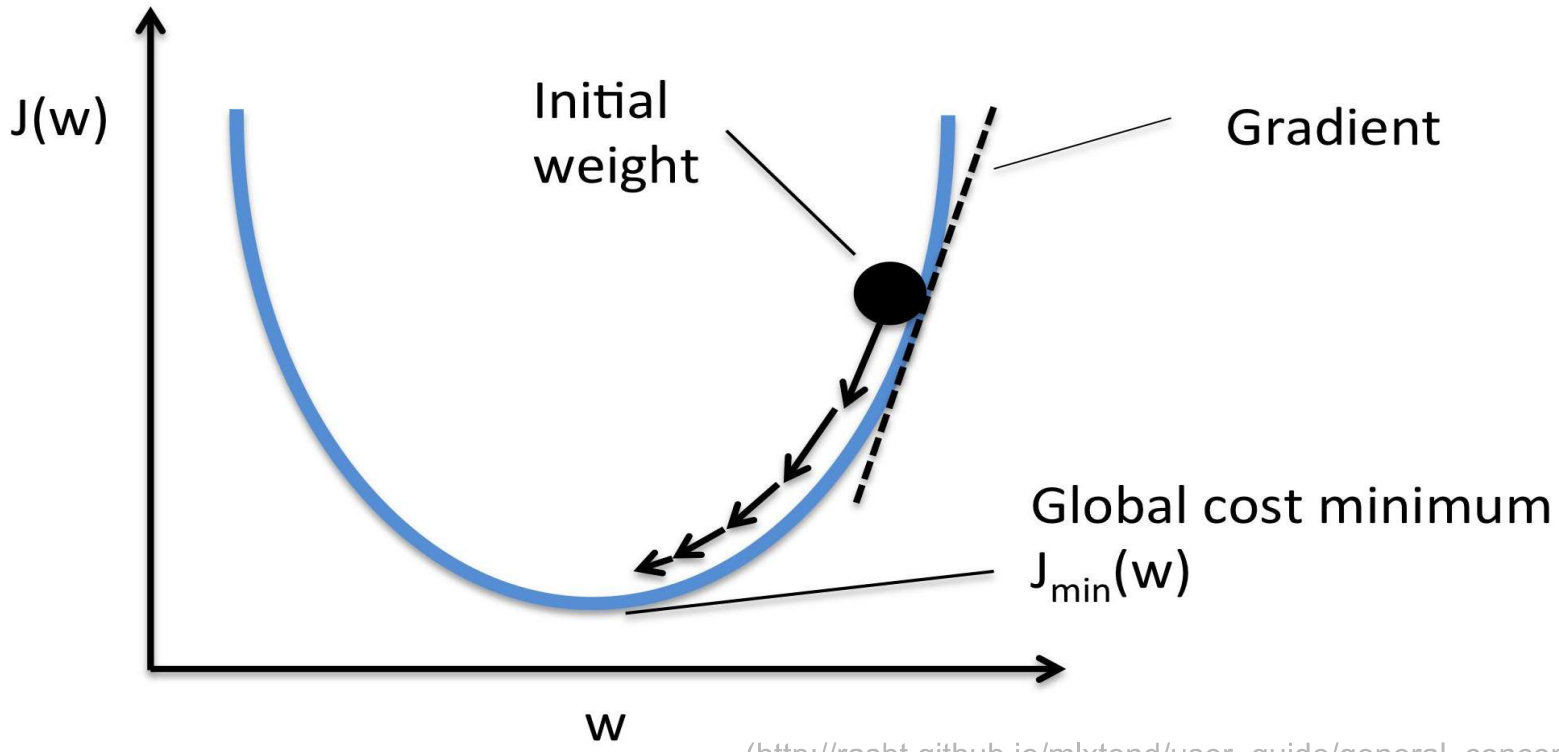
"best fit" : more efficient to maximize *log likelihood* :

$$\ell(\beta) = \sum_{i=1}^{N} y_i log\, p(x_i) + (1-y_i) log\, (1-p(x_i))$$



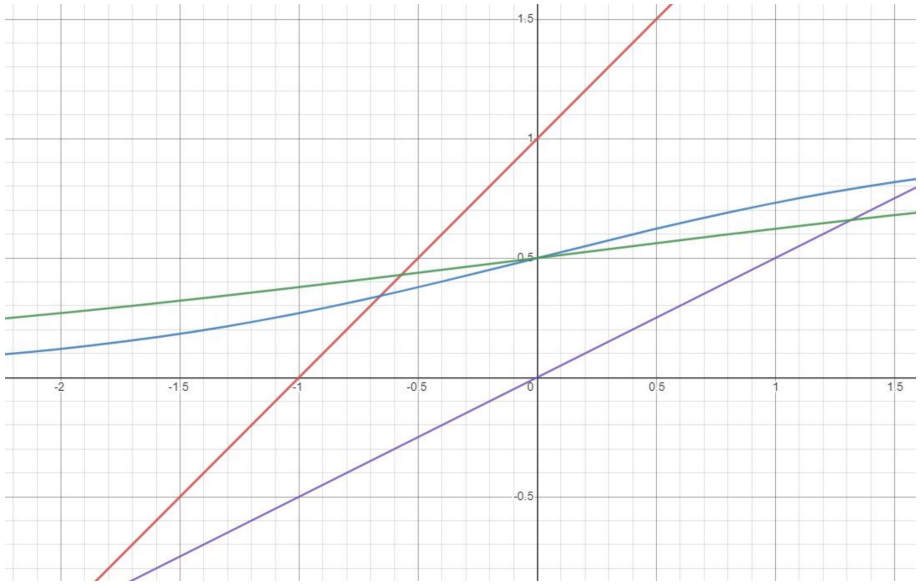
"best fit" : minimize *log loss*:

$$J(\square) = -\sum_{i=1}^{N} y_i log\, p(x_i) + (1-y_i) log\, (1-p(x_i))$$



(http://rasbt.github.io/mlxtend/user_guide/general_concepts/gradient-optimization/)

Linear Models



- Logistic Regression
- **Linear Regression**
- Pearson Product-Moment Correlation
- Multiple Linear/Logistic Regression

$$y = mx + b$$

Linear Regression

Finding a linear function based on *X* to best yield *Y.*

X = “covariate” = “feature” = “predictor” = “regressor” = “independent variable”

Y = “response variable” = “outcome” = “dependent variable”

Linear Regression

Finding a linear function based on *X* to best yield *Y.*

X = “covariate” = “feature” = “predictor” = “regressor” = “independent variable”

Y = “response variable” = “outcome” = “dependent variable”

Regression: $r(x) = \mathrm{E}(Y|X = x)$

goal: estimate the function *r*

Linear Regression

Finding a linear function based on *X* to best yield *Y.*

X = “covariate” = “feature” = “predictor” = “regressor” = “independent variable”

Y = “response variable” = “outcome” = “dependent variable”

Regression: $r(x) = \mathrm{E}(Y|X = x)$

goal: estimate the function *r*

The **expected** value of *Y*, given that the random variable *X* is equal to some specific value, x.

Linear Regression

Finding a linear function based on *X* to best yield *Y.*

X = “covariate” = “feature” = “predictor” = “regressor” = “independent variable”

Y = “response variable” = “outcome” = “dependent variable”

Regression: $r(x) = \mathrm{E}(Y|X = x)$

goal: estimate the function r

Linear Regression (univariate version): $r(x) = \beta_0 + \beta_1 x$

goal: find β_0, β_1 such that $r(x) \approx \mathrm{E}(Y|X = x)$

Linear Regression

Simple Linear Regression $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\mathbf{E}(\epsilon_i | X_i) = 0$ *and* $\mathbf{V}(\epsilon_i | X_i) = \sigma^2$

more precisely

$$r(x) = \beta_0 + \beta_1 x$$

Linear Regression

intercept slope error

Simple Linear Regression $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\mathbf{E}(\epsilon_i|X_i) = 0$ *and* $\mathbf{V}(\epsilon_i|X_i) = \sigma^2$

expected variance

Linear Regression: Estimating Params

Simple Linear Regression $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\mathbf{E}(\epsilon_i|X_i) = 0$ *and* $\mathbf{V}(\epsilon_i|X_i) = \sigma^2$

How to estimate intercept (β_0) and slope intercept (β_1)?

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the *residual sum of squares:*

$$J(\beta) = RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Linear Regression: Estimating Params

Method 1: Gradient Descent

initialize: $\hat{\beta}_0 = \hat{\beta}_1 = 0$; $\text{rss}^{(0)} = \infty$

for t in range(1, limit):

1. calculate all $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
2. if $\text{rss}^{(t-1)} - \text{rss}^{(t)} < \varepsilon$: break #converged
3. set:

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha(\sum_{i=1}^{n} \hat{Y}_i - Y_i)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha(\sum_{i=1}^{n} X_i(\hat{Y}_i - Y_i))$$

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the *residual sum of squares:*

$$J(\square) = RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Linear Regression: Estimating Params

Method 1: Gradient Descent

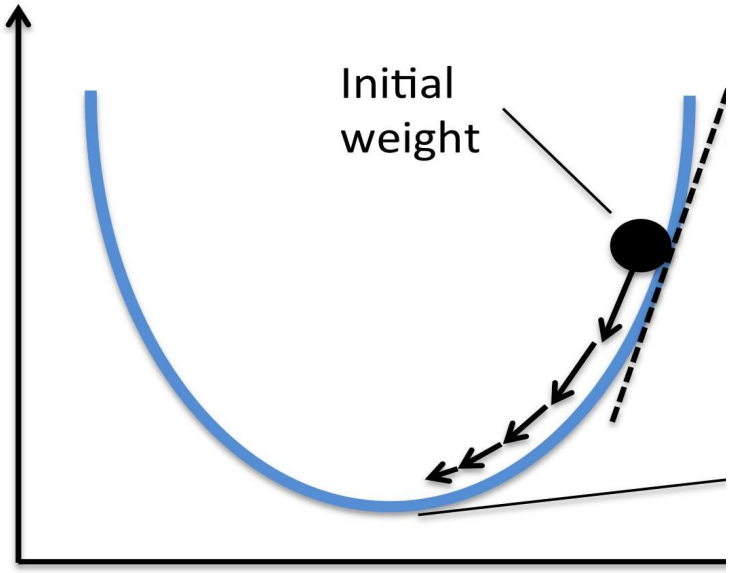
initialize: $\hat{\beta}_0 = \hat{\beta}_1 = 0$; $\text{rss}^{(0)} = \infty$

for t in range(1, limit):

1. calculate all $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
2. if $\text{rss}^{(t-1)} - \text{rss}^{(t)} < \varepsilon$: break #converged
3. set:

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha(\sum_{i=1}^{n} \hat{Y}_i - Y_i)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha(\sum_{i=1}^{n} X_i(\hat{Y}_i - Y_i))$$



(http://rasbt.github.io/mlxtend/user_guide/general_concepts/gradient-optimization/)

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the *residual sum of squares:*

$$J(\square) = RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Linear Regression: Estimating Params

Method 1: Gradient Descent

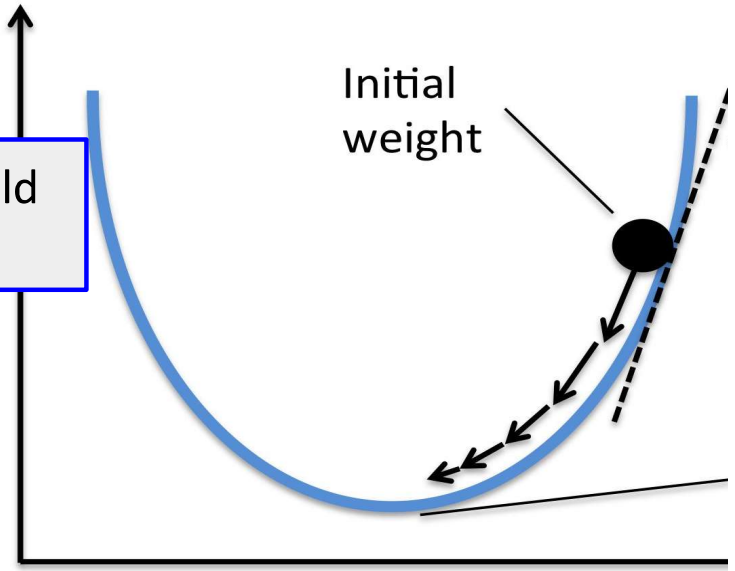
initialize: $\hat{\beta}_0 = \hat{\beta}_1 = 0$; $rss^{(0)} = \infty$

for t in range(1, limit):

1. calculate all $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
2. if $rss^{(t-1)} - rss^{(t)} < \varepsilon$: break #converged
3. set: $\hat{\beta}_0 = \hat{\beta}_0 - \alpha(\sum_{i=1}^{n} \hat{Y}_i - Y_i)$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha(\sum_{i=1}^{n} X_i(\hat{Y}_i - Y_i))$$

convergence threshold (e.g. .00001)



(http://rasbt.github.io/mlxtend/user_guide/general_concepts/gradient-optimization/)

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the *residual sum of squares:*

$$J(\square) = RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Linear Regression: Estimating Params

Method 1: Gradient Descent

initialize: $\hat{\beta}_0 = \hat{\beta}_1 = 0$; $\text{rss}^{(0)} = \infty$

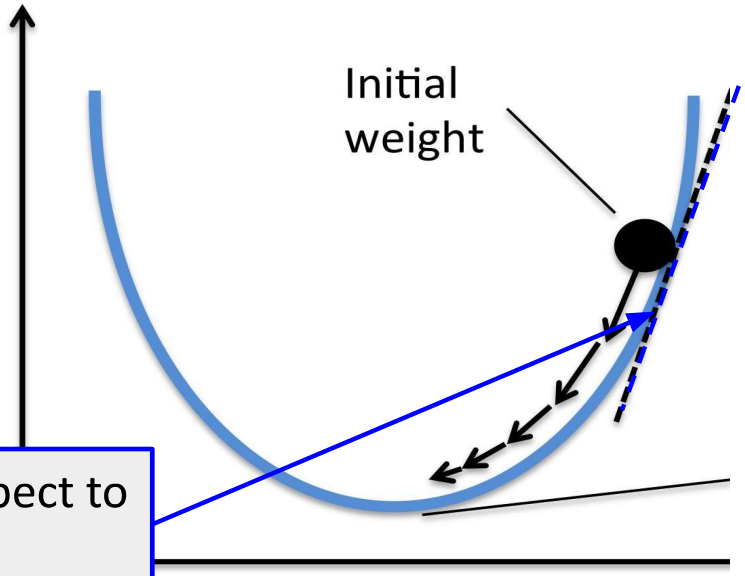
for t in range(1, limit):

1. calculate all $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
2. if $\text{rss}^{(t-1)} - \text{rss}^{(t)} < \varepsilon$: break #converged
3. set:

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha\left(\sum_{i=1}^{n} \hat{Y}_i - Y_i\right)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha\left(\sum_{i=1}^{n} X_i(\hat{Y}_i - Y_i)\right)$$

gradient with respect to given □.



(http://rasbt.github.io/mlxtend/user_guide/general_concepts/gradient-optimization/)

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the *residual sum of squares:*

$$J(\square) = RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Linear Regression: Estimating Params

Method 1: Gradient Descent

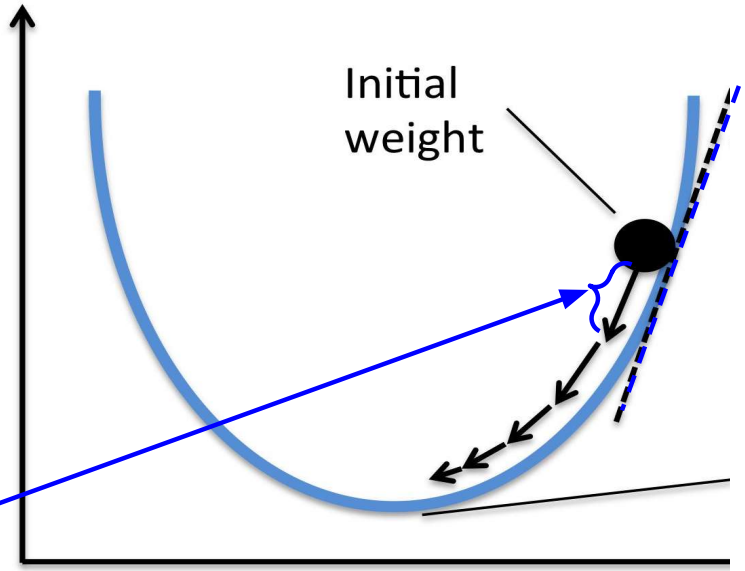
initialize: $\hat{\beta}_0 = \hat{\beta}_1 = 0$; $rss^{(0)} = \infty$

for t in range(1, limit):

1. calculate all $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
2. if $rss^{(t-1)} - rss^{(t)} < \varepsilon$: break #converged
3. set: $\hat{\beta}_0 = \hat{\beta}_0 - \alpha(\sum_{i=1}^{n} \hat{Y}$

$\hat{\beta}_1 = \hat{\beta}_1 - \alpha(\sum_{i=1}^{n} X_i(\hat{Y}$

learning rate: scales the size of the update.



(http://rasbt.github.io/mlxtend/user_guide/general_concepts/gradient-optimization/)

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the *residual sum of squares:*

$$J(\square) = RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Linear Regression: Estimating Params

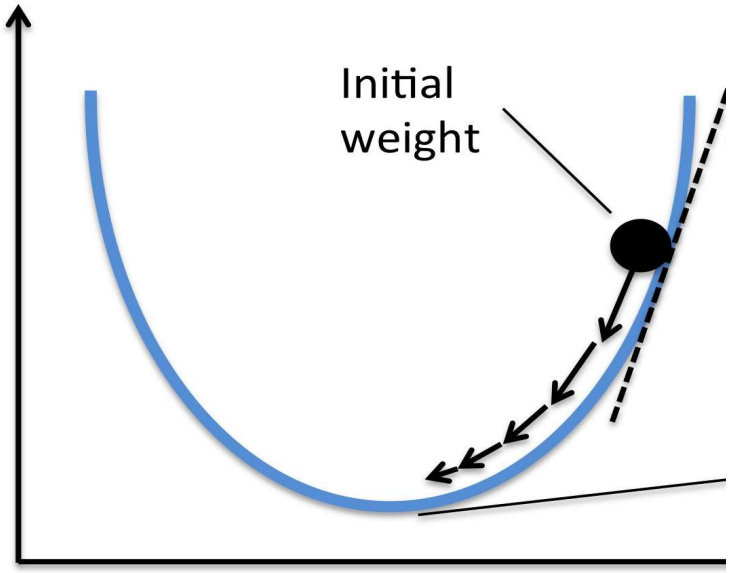
Method 1: Gradient Descent

initialize: $\hat{\beta}_0 = \hat{\beta}_1 = 0$; $\text{rss}^{(0)} = \infty$

for t in range(1, limit):

1. calculate all $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
2. if $\text{rss}^{(t-1)} - \text{rss}^{(t)} < \varepsilon$: break #converged
3. set: $\hat{\beta}_0 = \hat{\beta}_0 - \alpha(\sum_{i=1}^{n} \hat{Y}_i - Y_i)$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha(\sum_{i=1}^{n} X_i(\hat{Y}_i - Y_i))$$



(http://rasbt.github.io/mlxtend/user_guide/general_concepts/gradient-optimization/)

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the *residual sum of squares:*

$$J(\square) = RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Linear Regression: Estimating Params

Method 1: Gradient Descent

initialize: $\hat{\beta}_0 = \hat{\beta}_1 = 0$; $\text{rss}^{(0)} = \infty$

for t in range(1, limit):

1. calculate all $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
2. if $\text{rss}^{(t-1)} - \text{rss}^{(t)} < \varepsilon$: break #converged
3. set:

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha(\sum_{i=1}^{n} \hat{Y}_i - Y_i)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha(\sum_{i=1}^{n} X_i(\hat{Y}_i - Y_i))$$

Method 2: Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the *residual sum of squares:*

$$J(\square) = RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Linear Regression: Estimating Params

Like MLE:
Take the partial derivative of the loss function (RSS), and set equal to 0.

Method 2: Direct Estimates (normal equations)

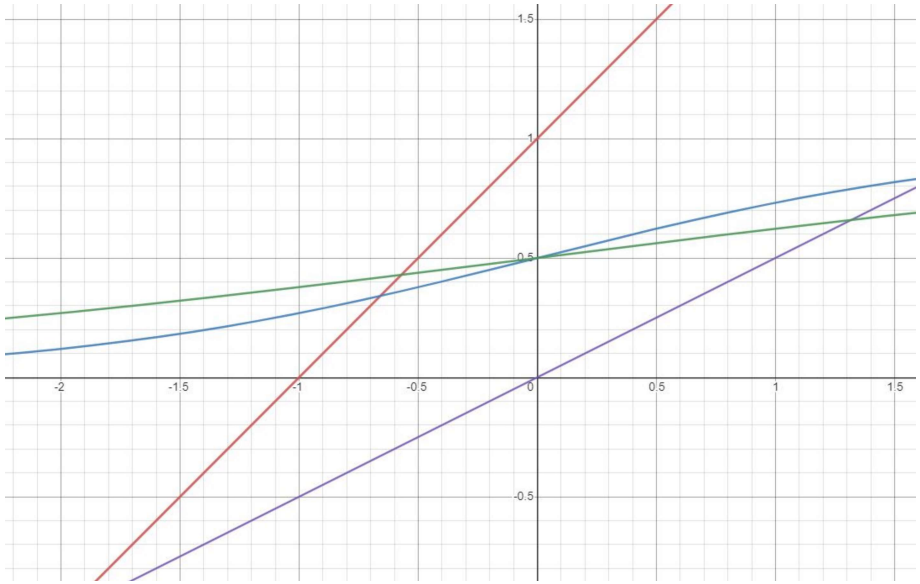
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the *residual sum of squares:*

$$J(\square) = RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Linear Models



- Logistic Regression
- Linear Regression
- **Pearson Product-Moment Correlation**
- Multiple Linear/Logistic Regression

$$y = mx + b$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} Cov(X,Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}\left((X-\bar{X})(Y-\bar{Y})\right) \end{aligned}$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} Cov(X,Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}\left((X - \bar{X})(Y - \bar{Y})\right) \end{aligned}$$

Correlation

$$r = r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y}$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} Cov(X,Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}\left((X - \bar{X})(Y - \bar{Y})\right) \end{aligned}$$

Correlation *(standardized covariance)*

$$r = r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} Cov(X,Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}\left((X-\bar{X})(Y-\bar{Y})\right) \end{aligned}$$

Correlation

$$r = r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)$$

Method 2: Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

Pearson Product-Moment Correlation

Covariance

$$Cov(X,Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$$
$$= \mathbf{E}\left((X - \bar{X})(Y - \bar{Y})\right)$$

Correlation

$$r = r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y}$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

Method 2: Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} Cov(X,Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}\left((X - \bar{X})(Y - \bar{Y})\right) \end{aligned}$$

Correlation

$$r = r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

Method 2: Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

If one standardizes *X* and *Y* (i.e. subtract the mean and divide by the standard deviation) before running linear regression, then:

??

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} Cov(X,Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}\left((X - \bar{X})(Y - \bar{Y})\right) \end{aligned}$$

Correlation

$$r = r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

Method 2: Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

If one standardizes *X* and *Y* (i.e. subtract the mean and divide by the standard deviation) before running linear regression, then:

$\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = r$ --- *i.e.* $\hat{\beta}_1$ *is the Pearson correlation!*

Multiple Linear Regression

Simple Linear Regression $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

where $\mathbf{E}(\epsilon_i|X_i) = 0$ *and* $\mathbf{V}(\epsilon_i|X_i) = \sigma^2$

expected variance

Estimated intercept and slope

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{Y}_i = \hat{r}(X_i)$$

Residual: $\hat{\epsilon}_i = Y_i - \hat{Y}_i$

Multiple Linear Regression

Suppose we have multiple X that we'd like to fit to Y at once:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

If we include and $X_{oi} = 1$ for all i, then we can say:

$$Y_i = \sum_{j=0}^{m} \beta_j X_{ij} + \epsilon_i$$

Multiple Linear Regression

Suppose we have multiple X that we'd like to fit to Y at once:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

If we include and $X_{oi} = 1$ for all i, then we can say:

$$Y_i = \sum_{j=0}^{m} \beta_j X_{ij} + \epsilon_i$$

Or in vector notation across all i:

$$Y = X\beta + \epsilon$$

where β and ϵ are vectors and X is a matrix.

Estimating β :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Multiple Linear Regression

Suppose we have multiple X that we'd like to fit to Y at once:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

If we include and $X_{oi} = 1$ for all i, then we can say: $Y_i = \sum_{j=0}^{m} \beta_j X_{ij} + \epsilon_i$

To test for significance of individual coefficient, j:

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}}}$$

Or in vector notation across all i:

$$Y = X\beta + \epsilon$$

where β and ϵ are vectors and X is a matrix.

Estimating β :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

$$s^2 = \frac{RSS}{df}$$

To test for significance of individual coefficient, j:

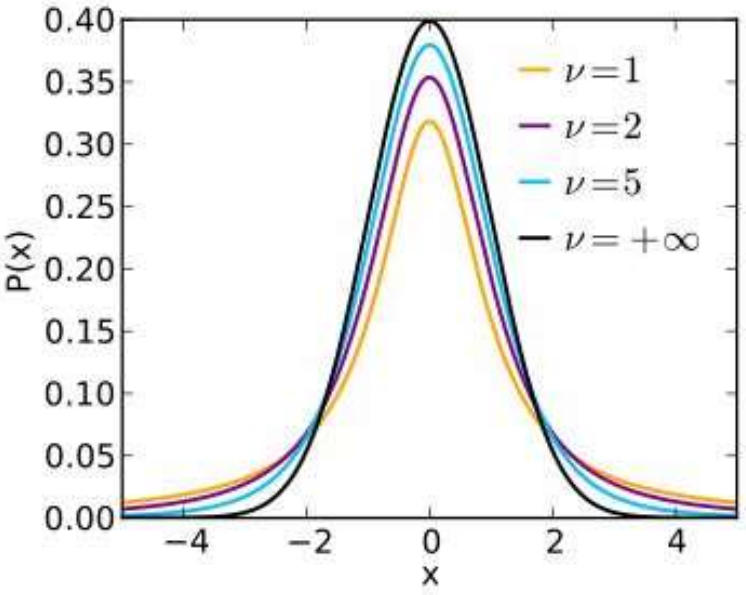
$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}}}$$

T-Test for significance of hypothesis:

1) Calculate t
2) Calculate degrees of freedom:

$$df = N - (m+1)$$

3) Check probability in a t distribution:



$$+ \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

t

T-Test for significance of hypothesis:

1) Calculate t
2) Calculate degrees of freedom:

$$df = N - (m+1)$$

3) Check probability in a t distribution: ($df = \nu$)

To test for significance of individual coefficient, j:

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}}}$$