# GENERALIZED LINEAR MODELS

CS6140

Predrag Radivojac
KHOURY COLLEGE OF COMPUTER SCIENCES
NORTHEASTERN UNIVERSITY

Spring 2021

# Ordinary Least Squares Regression

**Given:** a set of observations $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ , $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$

**Assume:** $Y = \omega_0 + \sum_{j=1}^d \omega_j X_j + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

**Goal:** find optimal fit $f(\boldsymbol{x}) = w_0 + \sum_{j=1}^d w_j x_j$ by maximizing likelihood

**The following holds:**

1. $\mathbb{E}[Y|\mathbf{x}] = \boldsymbol{\omega}^T \mathbf{x}$
2. $Y|\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu = \boldsymbol{\omega}^T \mathbf{x}$

$\boldsymbol{\omega}, \mathbf{x} \in \mathbb{R}^{d+1}$

# Ordinary Least Squares Regression

**Given:** a set of observations $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ , $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$

**Assume:** $Y = \omega_0 + \sum_{j=1}^{d} \omega_j X_j + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

**Goal:** find optimal fit $f(\boldsymbol{x}) = w_0 + \sum_{j=1}^{d} w_j x_j$ by maximizing likelihood

**The following holds:**

1. $\mathbb{E}[Y|\mathbf{x}] = \boldsymbol{\omega}^T \mathbf{x}$
2. $Y|\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu = \boldsymbol{\omega}^T \mathbf{x}$

$\boldsymbol{\omega}, \mathbf{x} \in \mathbb{R}^{d+1}$

# EXAMPLE: TARGET VARIABLE IS A NON-NEGATIVE COUNT

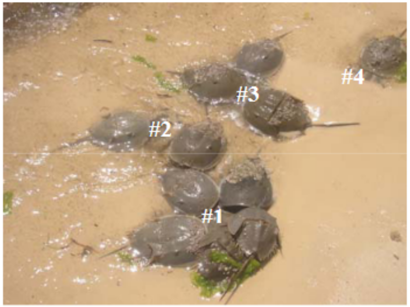


Figure 6. Four female horseshoe crabs (buried in the sand) nesting within 5 m of each other attract very different numbers of satellites. From bottom to top, female #1 has five satellite males in addition to her attached male, female #2 only has an attached male, female #3 has two satellites plus her attached male, and female #4 only has an attached male.

Schwab. MS Thesis, University of Florida, 2006.

# Example: Target Variable is a Non-negative Count

**Table 3.2. Number of Crab Satellites by Female's Color, Spine Condition, Width, and Weight**

| C | S | W | Wt | Sa | C | S | W | Wt | Sa | C | S | W | Wt | Sa | C | S | W | Wt | Sa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 3 | 28.3 | 3.05 | 8 | 3 | 3 | 22.5 | 1.55 | 0 | 1 | 1 | 26.0 | 2.30 | 9 | 3 | 3 | 24.8 | 2.10 | 0 |
| 3 | 3 | 26.0 | 2.60 | 4 | 2 | 3 | 23.8 | 2.10 | 0 | 3 | 2 | 24.7 | 1.90 | 0 | 2 | 1 | 23.7 | 1.95 | 0 |
| 3 | 3 | 25.6 | 2.15 | 0 | 3 | 3 | 24.3 | 2.15 | 0 | 2 | 3 | 25.8 | 2.65 | 0 | 2 | 3 | 28.2 | 3.05 | 11 |
| 4 | 2 | 21.0 | 1.85 | 0 | 2 | 1 | 26.0 | 2.30 | 14 | 1 | 1 | 27.1 | 2.95 | 8 | 2 | 3 | 25.2 | 2.00 | 1 |
| 2 | 3 | 29.0 | 3.00 | 1 | 4 | 3 | 24.7 | 2.20 | 0 | 2 | 3 | 27.4 | 2.70 | 5 | 2 | 2 | 23.2 | 1.95 | 4 |
| 1 | 2 | 25.0 | 2.30 | 3 | 2 | 1 | 22.5 | 1.60 | 1 | 3 | 3 | 26.7 | 2.60 | 2 | 4 | 3 | 25.8 | 2.00 | 3 |
| 4 | 3 | 26.2 | 1.30 | 0 | 2 | 3 | 28.7 | 3.15 | 3 | 2 | 1 | 26.8 | 2.70 | 5 | 4 | 3 | 27.5 | 2.60 | 0 |
| 2 | 3 | 24.9 | 2.10 | 0 | 1 | 1 | 29.3 | 3.20 | 4 | 1 | 3 | 25.8 | 2.60 | 0 | 2 | 2 | 25.7 | 2.00 | 0 |
| 2 | 1 | 25.7 | 2.00 | 8 | 2 | 1 | 26.7 | 2.70 | 5 | 4 | 3 | 23.7 | 1.85 | 0 | 2 | 3 | 26.8 | 2.65 | 0 |
| 2 | 3 | 27.5 | 3.15 | 6 | 4 | 3 | 23.4 | 1.90 | 0 | 2 | 3 | 27.9 | 2.80 | 6 | 3 | 3 | 27.5 | 3.10 | 3 |
| 1 | 1 | 26.1 | 2.80 | 5 | 1 | 1 | 27.7 | 2.50 | 6 | 2 | 1 | 30.0 | 3.30 | 5 | 3 | 1 | 28.5 | 3.25 | 9 |
| 3 | 3 | 28.9 | 2.80 | 4 | 2 | 3 | 28.2 | 2.60 | 6 | 2 | 3 | 25.0 | 2.10 | 4 | 2 | 3 | 28.5 | 3.00 | 3 |
| 2 | 1 | 30.3 | 3.60 | 3 | 4 | 3 | 24.7 | 2.10 | 5 | 2 | 3 | 27.7 | 2.90 | 5 | 1 | 1 | 27.4 | 2.70 | 6 |
| 2 | 3 | 22.9 | 1.60 | 4 | 2 | 1 | 25.7 | 2.00 | 5 | 2 | 3 | 28.3 | 3.00 | 15 | 2 | 3 | 27.2 | 2.70 | 3 |
| 3 | 3 | 26.2 | 2.30 | 3 | 2 | 1 | 27.8 | 2.75 | 0 | 4 | 3 | 25.5 | 2.25 | 0 | 3 | 3 | 27.1 | 2.55 | 0 |
| 3 | 3 | 24.5 | 2.05 | 5 | 3 | 1 | 27.0 | 2.45 | 3 | 2 | 3 | 26.0 | 2.15 | 5 | 2 | 3 | 28.0 | 2.80 | 1 |
| 2 | 3 | 30.0 | 3.05 | 8 | 2 | 3 | 29.0 | 3.20 | 10 | 2 | 3 | 26.2 | 2.40 | 0 | 2 | 1 | 26.5 | 1.30 | 0 |
| 2 | 3 | 26.2 | 2.40 | 3 | 3 | 3 | 25.6 | 2.80 | 7 | 3 | 3 | 23.0 | 1.65 | 1 | 3 | 3 | 23.0 | 1.80 | 0 |
| 2 | 3 | 25.4 | 2.25 | 6 | 3 | 3 | 24.2 | 1.90 | 0 | 2 | 2 | 22.9 | 1.60 | 0 | 3 | 2 | 26.0 | 2.20 | 3 |
| 2 | 3 | 25.4 | 2.25 | 4 | 3 | 3 | 25.7 | 1.20 | 0 | 2 | 3 | 25.1 | 2.10 | 5 | 3 | 2 | 24.5 | 2.25 | 0 |
| 4 | 3 | 27.5 | 2.90 | 0 | 3 | 3 | 23.1 | 1.65 | 0 | 3 | 1 | 25.9 | 2.55 | 4 | 2 | 3 | 25.8 | 2.30 | 0 |
| 4 | 3 | 27.0 | 2.25 | 3 | 2 | 3 | 28.5 | 3.05 | 0 | 4 | 1 | 25.5 | 2.75 | 0 | 4 | 3 | 23.5 | 1.90 | 0 |
| 2 | 2 | 24.0 | 1.70 | 0 | 2 | 1 | 29.7 | 3.85 | 5 | 2 | 1 | 26.8 | 2.55 | 0 | 4 | 3 | 26.7 | 2.45 | 0 |
| 2 | 1 | 28.7 | 3.20 | 0 | 3 | 3 | 23.1 | 1.55 | 0 | 2 | 1 | 29.0 | 2.80 | 1 | 3 | 3 | 25.5 | 2.25 | 0 |

Agresti. An introduction to categorical data analysis. Wiley. 2007.

# Poisson Regression

**Example:** the target variable is non-negative counts

1. $\log(\mathbb{E}[Y|\mathbf{x}]) = \boldsymbol{\omega}^T\mathbf{x}$
2. $Y|\mathbf{x} \sim$ Poisson$(\lambda)$, where $\lambda = e^{\boldsymbol{\omega}^T\mathbf{x}}$

**Mass function:** $$p(y|\mathbf{x},\mathbf{w}) = \frac{e^{\mathbf{w}^T\mathbf{x}y} \cdot e^{-e^{\mathbf{w}^T\mathbf{x}}}}{y!}$$

**Log-likelihood:** $$ll(\mathbf{w}) = \sum_{i=1}^{n} \mathbf{w}^T\mathbf{x}_i y_i - \sum_{i=1}^{n} e^{\mathbf{w}^T\mathbf{x}_i} - \sum_{i=1}^{n} y_i!$$

# POISSON REGRESSION: MAXIMIZE LIKELIHOOD

$$ll(\mathbf{w}) = \sum_{i=1}^{n} \mathbf{w}^T \mathbf{x}_i y_i - \sum_{i=1}^{n} e^{\mathbf{w}^T \mathbf{x}_i} - \sum_{i=1}^{n} y_i!$$

**Gradient:**

$$\begin{aligned} \frac{\partial ll(\mathbf{w})}{\partial w_j} &= \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} e^{\mathbf{w}^T \mathbf{x}_i} x_{ij} \\ &= \sum_{i=1}^{n} x_{ij} \cdot \left( y_i - e^{\mathbf{w}^T \mathbf{x}_i} \right) \\ &= \mathbf{f}_j^T \cdot (\mathbf{y} - \mathbf{p}) \end{aligned}$$

$$\nabla ll(\mathbf{w}) = \mathbf{X}^T \cdot (\mathbf{y} - \mathbf{p})$$

**Hessian:**

$$\begin{aligned} \frac{\partial^2 ll(\mathbf{w})}{\partial w_j \partial w_k} &= -\sum_{i=1}^{n} x_{ij} \cdot e^{\mathbf{w}^T \mathbf{x}_i} \cdot x_{ik} \\ &= -\mathbf{f}_j^T \cdot \mathbf{P} \cdot \mathbf{f}_k \end{aligned}$$

$$H_{ll(\mathbf{w})} = -\mathbf{X}^T \cdot \mathbf{P} \cdot \mathbf{X}$$

# Exponential Family of Distributions

$p(x|\boldsymbol{\theta}) = c(\boldsymbol{\theta})h(x)\exp\left(\sum_{i=1}^{m} q_i(\boldsymbol{\theta})t_i(x)\right)$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ is a set of parameters.

$x \in \mathbb{R}$

When $q_i(\boldsymbol{\theta}) = \theta_i$ for $\forall i$, $\theta_1, \theta_2, \ldots, \theta_m$ are called natural parameters. This gives

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= \exp\left(\sum_{i=1}^{m} \theta_i t_i(x) - a(\theta) + b(x)\right) \\ &= \exp\left(\boldsymbol{\theta}^T \mathbf{t}(x) - a(\boldsymbol{\theta}) + b(x)\right) \end{aligned}$$

where $\mathbf{t}(x) = (t_1(x), t_2(x), \ldots, t_m(x))$

# Examples of Exponential Family Members

$$p(x|\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T \mathbf{t}(x) - a(\boldsymbol{\theta}) + b(x)\right)$$

$$x \in \mathbb{R}$$

**Poisson:**

$$p(x|\lambda) = \exp\left(x \log \lambda - \lambda - \log x!\right)$$

$$\theta = \log \lambda,\ t(x) = x,\ a(\theta) = e^{\theta},\ \text{and } b(x) = -\log x!$$

**Gaussian:**

$$p(x|\mu, \sigma) = \exp\left(-\frac{x^2}{2\sigma^2} + x\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)$$

$$\boldsymbol{\theta} = (\tfrac{\mu}{\sigma^2}, -\tfrac{1}{2\sigma^2}),\ \mathbf{t}(x) = (x, x^2),\ a(\boldsymbol{\theta}) = \frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log(\tfrac{\theta_2}{\pi}),\ \text{and } b(x) = 0.$$

... and many more.

# Estimating Parameters in Exponential Family of Distributions

**Given:** $\mathcal{D} = \{x_i\}_{i=1}^n$, where $X \sim$ Exponential Family Member

**Log-likelihood:**

$$\begin{aligned} ll(\mathbf{w}) &= \log \prod_{i=1}^{n} e^{\boldsymbol{\theta}^T \mathbf{t}(x_i) - a(\boldsymbol{\theta}) + b(x_i)} \\ &= \sum_i \sum_m \theta_m t_m(x_i) - n \cdot a(\boldsymbol{\theta}) + \sum_i b(x_i) \\ &= \sum_i ll_i(\mathbf{w}) \end{aligned}$$

**Gradient:**

$$\frac{\partial ll_i(\mathbf{w})}{\partial w_j} = \sum_{k=1}^{m} \frac{\partial \theta_k}{\partial w_j} t_k(x_i) - \frac{\partial a(\boldsymbol{\theta})}{\partial w_j}$$

# Generalized Linear Model

1. $f(\mathbb{E}[Y|\mathbf{x}]) = \boldsymbol{\omega}^T \mathbf{x}$
2. $p(y|\mathbf{x}) \in$ Exponential Family

$f(\cdot) =$ link function

# Logistic Regression

1. $\text{logit}(\mathbb{E}[Y|\mathbf{x}]) = \boldsymbol{\omega}^T\mathbf{x}$
2. $Y|\mathbf{x} \sim \text{Bernoulli}(\alpha)$, where $\alpha = \frac{1}{1+e^{-\boldsymbol{\omega}^T\mathbf{x}}}$

where $\text{logit}(x) = \log \frac{x}{1-x}$

$$p(y|\mathbf{x}, \mathbf{w}) = \left(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}\right)^y \left(1 - \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}\right)^{1-y}$$