$SS_{residual} = \sum_{i=1}^{n} (y_i - \hat{B}^T x_i)^2$ $\quad B = [m \times 1]$ $\quad X_i = [i \times m]$ $\quad h^2 = R^2 = 1 - \frac{SS_{res}}{SS_{total}}$

$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

$y = B_0 + \sum_{j=1}^{m} B_j x_j + e, \quad e \sim N(0, \sigma^2) \Rightarrow$

$Var(y) = Var(B_0) + Var\sum_{j=1}^{m} B_j x_j + Var(e)$

$= Var \sum_{j=1}^{m} Var\, B_j x_j + \sigma^2$

$= \sum_{j=1}^{m} B_j^2 Var(x_j) + \sigma^2$

When standardized, $Var(y) = \sum_{j=1}^{m} B_j^2 + \sigma^2$

VARIANCE IDENTITIES

$Var(X) = E[X^2] - E[X]^2$

$Cov(X,Y) = E[XY] - E[X]E[Y]$

$Var(X+Y) = E[X+Y]^2 - E[X+Y]^2$

$= E(X^2 + 2XY + Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2$

$= E(X^2) + 2E(XY) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2$

$= E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 + 2E(XY) - 2E(X)E(Y)$

$Var(X) + Var(Y) + 2Cov(X,Y)$

$Var(X+Y) = Var(X) + Var(Y) + [Cov(X,Y)]$

$Var(cX) = E[(cX)^2] - E[cX]^2$

$= E[c^2X^2] - cE[X]^2$

$= c^2 E[X^2] - (cE[X])^2$

$= c^2 E[X^2] - c^2 E[X]^2$

$= c^2 (Var(X))$

$h^2 = R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{E(y - (B_0 + B^T x)^2)}{Var(y)}$

$= 1 - \frac{\sigma^2}{Var(y)} = 1 - \frac{\sigma^2}{\sum_{j=1}^{m} B_j^2 + \sigma^2}$

$= \frac{\sum_{j=1}^{m} B_j^2 + \sigma^2 - \sigma^2}{\sum_{j=1}^{m} B_j^2 + \sigma^2} = \frac{\sum_{j=1}^{m} B_j^2}{\sum_{j=1}^{m} (B_j^2 + \sigma^2)}$

Error squared (ϵ^2)

$\epsilon^2 = [y - (B_0 + B^T x)]^2 = \sigma^2$

$E(\epsilon^2) = E([y - (B_0 + B^T x)]^2)$

$= \sigma^2$

$\sigma^2 \leq Var(y)$

$\sigma^2 \leq \sum_{j=1}^{m} B_j^2 + \sigma^2$

$-\sigma^2 \quad -\sigma^2$

$0 \leq \sum_{j=1}^{m} B_j^2 \leftarrow$ positive semi-definite.

Ridge Regression

Matrix is invertible if $A \cdot B = I_n$

$$\hat{B} = (\tilde{X}^T \tilde{X} + \lambda I)^{-1} \tilde{X}^T y$$

$\tilde{X} = n \times m$ } $\tilde{X}^T \tilde{X} = m \times m$
$\tilde{X}^T = m \times n$ }
$y = n \times 1$ } $\tilde{X}^T y = m \times 1$
λ = scalar
$I = m \times m \Rightarrow \hat{B} = m \times 1$

Bayes Rule

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

posterior is proportional to prior times likelihood.

Prior for Bernoulli p is $P(p) = \text{Beta}(p, \alpha, \beta) \approx p^{\alpha-1}(1-p)^{\beta-1}$

$$E[p|x_1, \dots x_n] = \bar{x} \frac{n}{\alpha+\beta+n} + \frac{\alpha}{\alpha+\beta} \frac{\alpha+\beta}{\alpha+\beta+n}$$

Bernoulli Distribution

$X_1, X_2 \dots X_n \overset{iid}{\sim} \text{Ber}(p)$

$$\mathcal{L}(p) = P(x_1, \dots x_n | p) = \prod_{i=1}^{n} P(x_i|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{(1-x_i)} = p^{n\bar{x}}(1-p)^{n(1-\bar{x})}$$

$y = B^T x + \epsilon$
$y \sim N(B^T x, \sigma_0^2)$
$B_i \sim N(0, \sigma^2)$

$$P(B|D, \sigma^2, \sigma_0^2) = \frac{P(D|B, \sigma^2, \sigma_0^2)P(B)}{P(D|\sigma^2, \sigma_0^2)}$$

MAP:

$$B_{MAP} = \arg\max_B P(B|D, \sigma^2, \sigma_0^2) = \arg\max_B P(D, B, \sigma^2, \sigma_0^2)$$

If prior for B is uniform $(\frac{1}{NB} \sum_{i=1}^{N} B_i)$, MAP = MLE.

Linear Regression OLS (Probabilistic)

$y = B^T x + \epsilon$
$x = m \times n$
$B^T = 1 \times m$, $B = m \times 1$
$y = 1 \times n$

$\epsilon \sim N(0, \sigma_0^2) \Rightarrow y \sim N(B^T x, \sigma_0^2)$
We assume B to be fixed.

$$p(y|x, B, \sigma_0^2) \Rightarrow L(B, \sigma_0^2) \Rightarrow p(y|D, B, \sigma^2) \Rightarrow \prod_{i=1}^{n} p(y_i|x_i, B, \sigma_0^2)$$

$$B_{OLS} = B_{ML} = \arg\max_B L(B, \sigma_0^2) \leftarrow \underline{\text{FREQUENTIST}}$$

$$P(D, B) = P(D|B)P(B) = \prod_{i=1}^{N} P(y_i|x_i, B) \prod_{j=1}^{M} P(B_j)$$

$P(D, B)$: joint likelihood

$\prod_{i=1}^{N} P(y_i|x_i, B)$: this is the probability of D given B, $P(D|B)$

$\prod_{j=1}^{M} P(B_j)$: this is $P(B)$

$$\log P(D, B) = \sum_{i=1}^{N} \log P(y_i|x_i, B) + \sum_{j=1}^{M} \log P(B_j)$$

(with Gaussian PDF) $\Rightarrow$ $= \frac{\sum_i (B^T x_i - y_i)^2}{2\sigma_0^2} - \sum \frac{1}{2\sigma^2}(B_j^2 + \text{constant})$

Error function for ridge regression:

$$E(B) = \sum_i (B^T x_i - y_i)^2 + \lambda \|B\|_2^2 \qquad \lambda > 0, \text{ and } = \frac{\sigma_0^2}{\sigma^2}$$

regularized linear regression:

$$\arg\min_B \sum_i (B^T x_i - y_i)^2 + \lambda \|B\|_2^2 \Rightarrow B_{MAP} = (X^T X + \lambda I)^{-1} X^T y$$

When λ is 0, then this reduces to $(X^T X)^{-1} X^T y$

(x_i, y_i), $x_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}$

$X = n \times m$, $y = n \times 1$

$y = XB + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

$$(X^T X)\hat{B}_{ML} = X^T y \quad (m \times m)(m \times 1) = (m \times n)(n \times 1)$$

Gram matrix: $K = XX^T$ $\quad (n \times m)(m \times n) = n \times n$

$$\hat{B}_{ML} = \frac{X^T y}{X^T X} = (X^T y)(X^T X)^{-1}$$

K is nonparametric, and grows with the data. If $X^T X$ is invertible, then:

$$\hat{B}_{ML} = (X^T y)(X^T X)^{-1}$$
$$= (X^T y)(X^T X)(X^T X)^{-2}$$
$$= X^T \alpha \quad \text{(this is simply rearranging values).}$$

$$\alpha = \underbrace{X}_{n \times m}\underbrace{(X^T X)^{-2}}_{(n \times m)(m \times 1)}\underbrace{X^T}_{m \times n}\underbrace{y}_{n \times 1} \Rightarrow \underbrace{\alpha}_{n \times 1} = \sum_i^n \alpha_i x_i$$

$$\hat{B}_{MAP} = \underset{B}{\operatorname{argmin}} \frac{1}{2\sigma^2} (y - XB)^T (y - XB) + \lambda B^T B$$

(annotations: y: $n \times 1$, X: $n \times m$, B: $m \times 1$; $(y-XB)^T$: $1 \times n$, $(y-XB)$: $n \times 1$; $\lambda B^T B$: $(1\times1)(1\times m)(m \times 1)$)

$$\mathcal{L}(\sigma_0^2, \sigma^2) = P(y \mid X, \sigma_0^2, \sigma^2)$$
$$= \int_B P(y, B \mid X, \sigma^2, \sigma_0^2)\, dB$$
$$= \int_B P(y \mid B, X, \sigma_0^2)\, P(B \mid \sigma^2)\, dB$$

probability of the y given the other hyper parameters, integrating out one of the parameters

$$= N(0, \sigma^2 XX^T + \sigma_0^2 I_n)$$

B_{MAP} is biased estimator of B, B_{OLS} is unbiased.

estimating B given hyperparameters $\Rightarrow O(mn^2)$
but estimating hyperparameters is $O(n^3)$

$$h_m^2 = \frac{m\sigma^2}{m\sigma^2 + \sigma_0^2}$$

Workflow =

1. Take data $\{(x_i, y_i)\}$
2. model phenotype as linear:
   $y_i = B^T x_i + \epsilon_i$
   $B_j \sim N(0, \sigma^2)$ $\epsilon \sim N(0, \sigma^2)$
3. estimate hyperparameters by maximizing marginal likelihood

$$S = \{(x^{(i)}, y^{(i)})\}_{i=1}^m$$

$$p(\theta \mid S) = \frac{p(S \mid \theta)\, p(\theta)}{p(S)}$$

$$= \frac{\left(\prod_{i=1}^m p(y^{(i)} \mid x^{(i)}, \theta)\right) p(\theta)}{\int_\theta \prod_{i=1}^m p(y^{(i)} \mid x^{(i)}, \theta)\, p(\theta)\, d\theta}$$

For a logistic regression,
~~$p(\theta|S)$~~ $\Rightarrow p(y_i \mid x_i, \theta) =$

$$h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{(1-y^{(i)})}, \quad h_\theta = \frac{1}{1 + e^{-\theta^T x_i}}$$