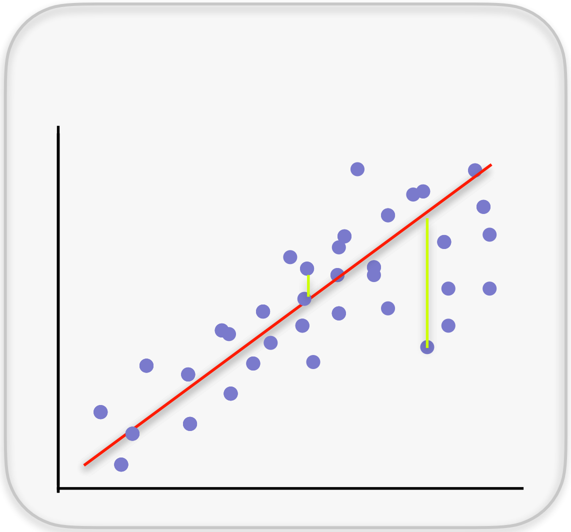
Lecture 3: Linear regression (continued)

The regression problem in matrix notation

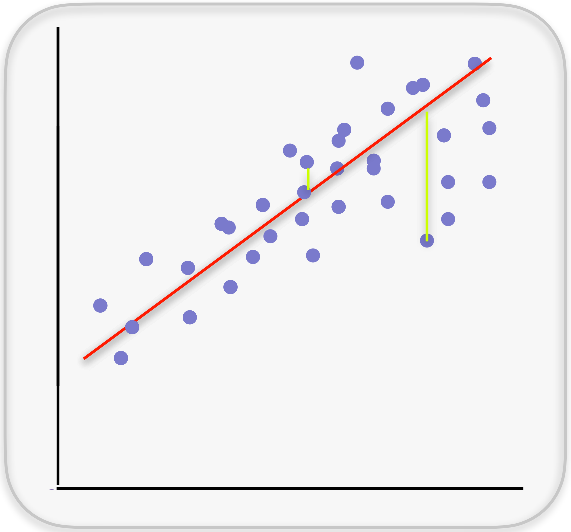
Linear model: $y_i = x_i^T w + \epsilon_i$

Least squares solution:

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

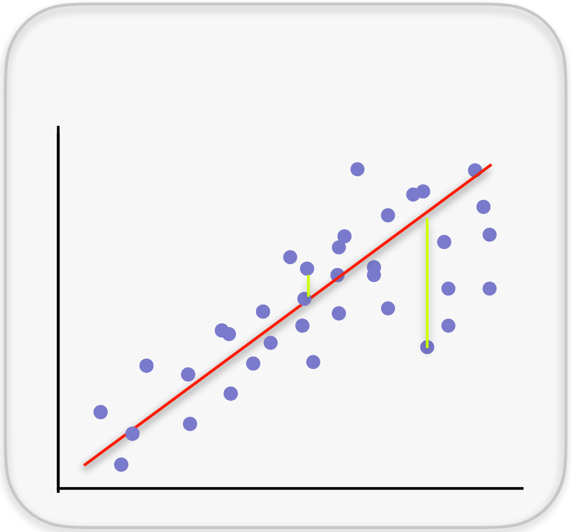


What about an offset
(a.k.a intercept)?



The regression problem in matrix notation

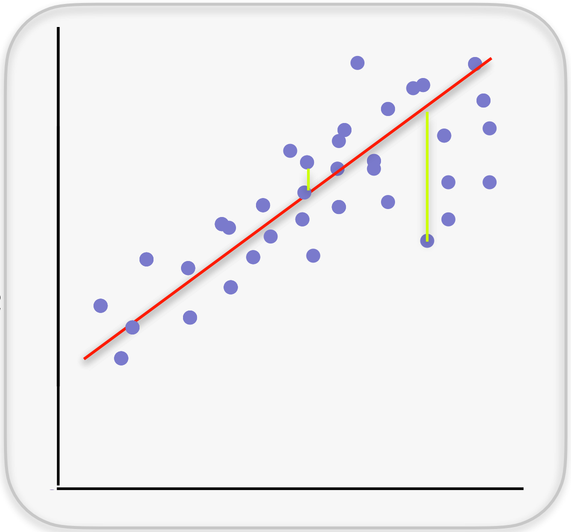
Linear model: $y_i = x_i^T w + \epsilon_i$



Least squares solution:

$$\begin{aligned}\widehat{w}_{LS} &= \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2 \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}$$

Affine model: $y_i = x_i^T w + b + \epsilon_i$



Least squares solution:

$$\begin{aligned}\widehat{w}_{LS}, \widehat{b}_{LS} &= \arg\min_{w,b} \sum_{i=1}^n \left(y_i - (x_i^T w + b)\right)^2 \\ &= \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2\end{aligned}$$

Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \|\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)\|_2^2$$

$$= \arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \underbrace{(\mathbf{y} - (\mathbf{X}w + \mathbf{1}b))^T(\mathbf{y} - (\mathbf{X}w + \mathbf{1}b))}_{\mathcal{L}(w,b)}$$

Set gradient w.r.t. w and b to zero to find the minima:

A reminder on vector calculus

$$f(w) = (Aw + b)^T(Aw + b) \implies \nabla_W f(w) = 2A^T(Aw + b)$$

Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\mathbf{X}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{X}^T\mathbf{1} = \mathbf{X}^T\mathbf{y}$$
$$\mathbf{1}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{1}^T\mathbf{1} = \mathbf{1}^T\mathbf{y}$$

If $\mathbf{X}^T\mathbf{1} = 0$, if the features have zero mean,

$$\widehat{w}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$
$$\widehat{b}_{\text{LS}} = (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\mathbf{X}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{X}^T\mathbf{1} = \mathbf{X}^T\mathbf{y}$$

$$\mathbf{1}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{1}^T\mathbf{1} = \mathbf{1}^T\mathbf{y}$$

If $\mathbf{X}^\mathbf{T}\mathbf{1} = 0$,

$$\widehat{w}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\widehat{b}_{\text{LS}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

In general, when $\mathbf{X}^T\mathbf{1} \neq 0$,

Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\mathbf{X}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{X}^T\mathbf{1} = \mathbf{X}^T\mathbf{y}$$

$$\mathbf{1}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{1}^T\mathbf{1} = \mathbf{1}^T\mathbf{y}$$

If $\mathbf{X^T1} = 0$,

$$\widehat{w}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\widehat{b}_{\text{LS}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

In general, when $\mathbf{X}^T\mathbf{1} \neq 0$,

$$\mu = \frac{1}{n}\mathbf{X}^T\mathbf{1}$$

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\mu^T$$

$$\widehat{w}_{\text{LS}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{y}$$

$$\widehat{b}_{\text{LS}} = \frac{1}{n}\sum_{i=1}^{n} y_i - \mu^T\widehat{w}_{\text{LS}}$$

Process for linear regression with intercept

Collect data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$

Decide on a **model:** $y_i = x_i^T w + b + \epsilon_i$

Choose a loss function - least squares

Pick the function which minimizes loss on data

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} \sum_{i=1}^n \left(y_i - (x_i^T w + b)\right)^2$$

Use function to make prediction on new examples x_{new}

$$\hat{y}_{\text{new}} = x_{\text{new}}^T \hat{w}_{LS} + \hat{b}_{LS}$$

Another way of dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

reparametrize the problem as $\overline{\mathbf{X}} = [\mathbf{X}, \mathbf{1}]$ and $\overline{w} = \begin{bmatrix} w \\ b \end{bmatrix}$

$$\overline{\mathbf{X}}\,\overline{w} \;=$$

Why do we use least squares (i.e. ℓ_2-loss)?

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Consider $\quad y_i = x_i^T w + \epsilon_i \quad$ where $\quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$\implies y_i \sim$

$\implies P(y_i; x_i, w, \sigma) =$

Why do we use least squares (i.e. ℓ_2-loss)?

Maximum Likelihood Estimator:

$$\begin{aligned}
\widehat{w}_{\text{MLE}} &= \arg\max_w \; \log P(\{y_i\}_{i=1}^n; \{x_i\}_{i=1}^n, w, \sigma) \\
&= \arg\max_w \; -n \log(\sigma\sqrt{2\pi}) + \sum_{i=1}^n -\frac{(y_i - x_i^T w)^2}{2\sigma^2}
\end{aligned}$$

Why do we use least squares (i.e. ℓ_2-loss)?

Maximum Likelihood Estimator:

$$
\begin{aligned}
\widehat{w}_{\text{MLE}} &= \arg\max_w \log P(\{y_i\}_{i=1}^n; \{x_i\}_{i=1}^n, w, \sigma) \\
&= \arg\max_w -n\log(\sigma\sqrt{2\pi}) + \sum_{i=1}^n -\frac{(y_i - x_i^T w)^2}{2\sigma^2} \\
&= \arg\min_w \sum_{i=1}^n (y_i - x_i^T w)^2
\end{aligned}
$$

Recall: $$\widehat{w}_{LS} = \arg\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

$$\widehat{w}_{LS} = \widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Recap of linear regression

$$\text{Data } \{(x_i, y_i)\}_{i=1}^n$$

Minimize the loss
(Empirical Risk Minimization)

Choose a loss
e.g., ℓ_2-loss: $(y_i - x_i^T w)^2$

Solve $\widehat{w}_{\text{LS}} = \arg\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$

Maximize the likelihood
(MLE)

Choose a Hypothesis class
e.g., $y_i = x_i^T w + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,\sigma^2)$

Maximize the likelihood,

$$\widehat{w}_{\text{MLE}} = \arg\max_w \left\{ -n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^n \frac{(y_i - x_i^T w)^2}{2\sigma^2} \right\}$$

Analysis of Error under additive Gaussian noise

Let's suppose $y_i = x_i^T w^* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$, then this can be written as
$\mathbf{y} = \mathbf{X}w^* + \epsilon$

$$\begin{aligned}\widehat{w}_{\text{MLE}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}w^* + \epsilon) \\ &= w^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\end{aligned}$$

Maximum Likelihood Estimator is unbiased:

Analysis of Error under additive Gaussian noise

Let's suppose $y_i = x_i^T w^* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$, then this can be written as $\mathbf{y} = \mathbf{X}w^* + \epsilon$

$$\begin{aligned}
\widehat{w}_{\mathrm{MLE}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}w^* + \epsilon) \\
&= w^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon
\end{aligned}$$

Covariance is:

Analysis of Error under additive Gaussian noise

Let's suppose $y_i = x_i^T w^* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$, then this can be written as $\mathbf{y} = \mathbf{X}w^* + \epsilon$, and the MLE is

$$\widehat{w}_{\text{MLE}} = w^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

This random estimate has the following distribution:

$$\mathbb{E}[\hat{w}_{\text{MLE}}] = w^*, \text{Cov}(\hat{w}_{\text{MLE}}) = \mathbb{E}[(\hat{w} - \mathbb{E}[\hat{w}])(\hat{w} - \mathbb{E}[\hat{w}])^T] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

$$\hat{w}_{\text{MLE}} \sim \mathcal{N}(w^*, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Interpretation: consider an example with $\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}$

The covariance of the MLE, $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, captures how each sample gives information about the unknown w^*, but each sample gives information about for different (linear combination of) coordinates and of different quality/strength

Questions?