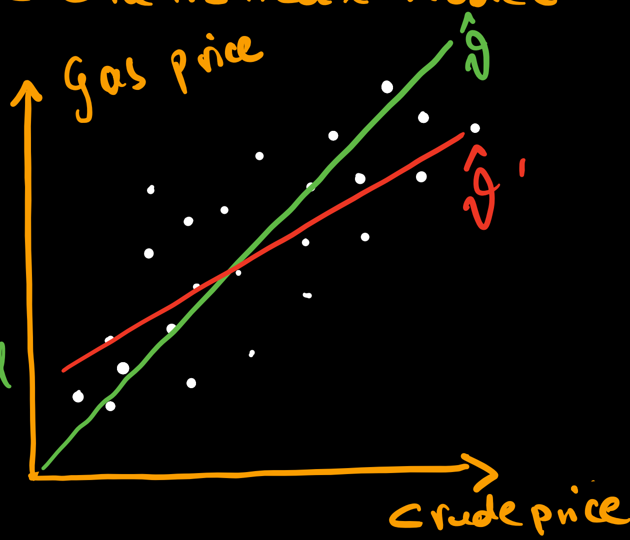
Linear Models:

Example: We collet data (x_j, y_j) on crude oil price and gas price. We consider the statistical model

$$y_j \mid X_j \sim \vartheta X_j + \varepsilon_j$$

influence of crude price on gas

noise of ter external random factors



Goal: We want to find line so as to predict data. (line to fit the data)

Idea: Estimate ϑ with the slope of the line minimizing the errors between the data.

$$\hat{\vartheta}(x,y) \text{ minimizes } \sum (y_j - \vartheta \cdot x_j)^2$$

Note: Throughout, X_j are considered fixed
if $X_j \sim P_X$ then x_j RV, results are conditional on $(X_1, \dots, X_n)$.

Example (offsets): Consider the affine model $y_j = \vartheta_0 + \vartheta_1 x_j + \varepsilon_j \quad \varepsilon_j \sim \mathcal{N}_0(0,\sigma)$

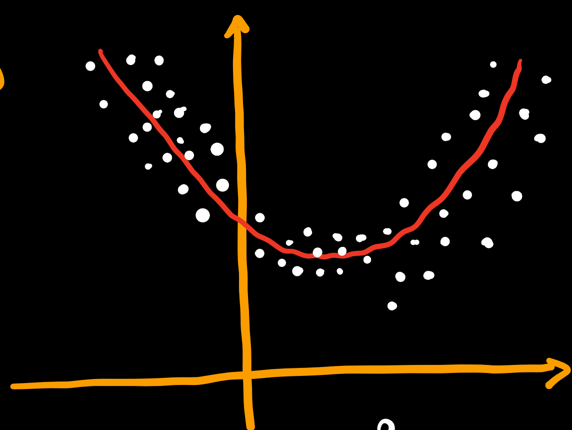
Then we can fit

$$\hat{\vartheta}(x,y) = \operatorname{argmin} \sum \left(y_j - (\vartheta_0 + \vartheta_1 x_j) \right)^2$$

$$= \operatorname{argmin} \sum \Big(\underbrace{y_j}_{\text{observed}} - \underbrace{\begin{pmatrix} \vartheta_0 \\ \vartheta_1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ x_j \end{pmatrix}}_{\text{predicted}} \Big)^2$$

Example (Quadratic): Consider now

$$y_j = \vartheta_0 + \vartheta_1 x_j + \vartheta_2 x_j^2 + \varepsilon_j$$



We can still find a model fitting our data

$$\hat{\vartheta}(x,y) = \operatorname{argmin} \sum \left(y_j - \begin{pmatrix} 1 \\ x_j \\ x_j^2 \end{pmatrix} \begin{pmatrix} \vartheta_0 \\ \vartheta_1 \\ \vartheta_2 \end{pmatrix} \right)^2$$

$\rightsquigarrow$ The model is still linear in ϑ!

Definition: A *linear model* is a statistical model where:

$$y_j = \sum_{\ell=1} \vartheta_\ell \, \Phi(x_j)_\ell + \sigma \cdot Z_j \quad j \in \{1,\dots,n\}$$

$$\mathbb{E}[Z_j] = 0$$

Unknown Parameters : $\vartheta_1, \dots, \vartheta_p$

Note: We can represent this in vector notation

$$Y = \Phi \vartheta + \sigma \cdot Z$$

where
- $\Phi_{je} = \Phi(x_j)_e$ is the design matrix
- $\vartheta = (\vartheta_1, \dots, \vartheta_p)$ and σ^2 are the parameters

We say that the linear model is underparametrized if $\operatorname{rank}(\phi) = p$

Definition: The least square estimators $\hat{\vartheta}$ of ϑ is given by:

$$\hat{\vartheta} = \underset{\vartheta \in \mathbb{R}^p}{\arg\min} \| Y - \Phi \vartheta \|_2^2$$

Lemma: In the underparametrized setting $\hat{\vartheta}$ is unique and can be written as

$$\hat{\vartheta}(Y) = (\Phi^T \Phi)^{-1} \Phi^T Y$$

Proof: $\|Y - \Phi(\vartheta)\|_2^2 = (Y - \Phi\vartheta)^T (Y - \Phi\vartheta)$

Solve $\nabla_\vartheta (Y - \Phi\vartheta)^T (Y - \Phi\vartheta) = 0$

$$\Rightarrow \quad \Phi^T \Phi \vartheta = \Phi^T y$$

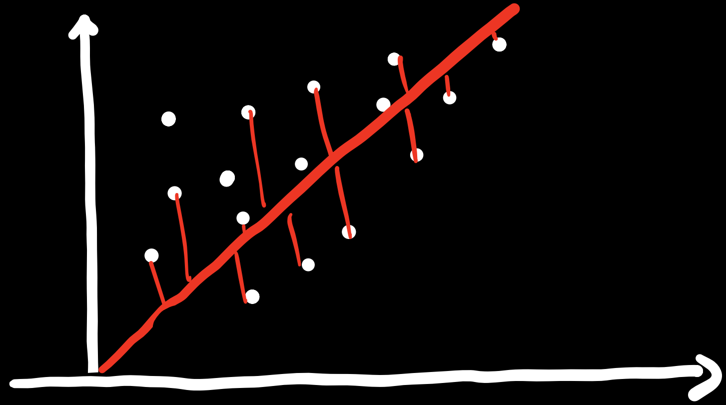
Since $\Phi^T \Phi \in \mathbb{R}^{p \times p}$ is full rank, it is invertible so

$$\hat{\vartheta}(Y) = (\Phi^T \Phi)^{-1} \Phi^T y \qquad \square$$

Note: $\hat{\vartheta}(Y)$ is linear in Y

Geometric interpretation: (least squares)

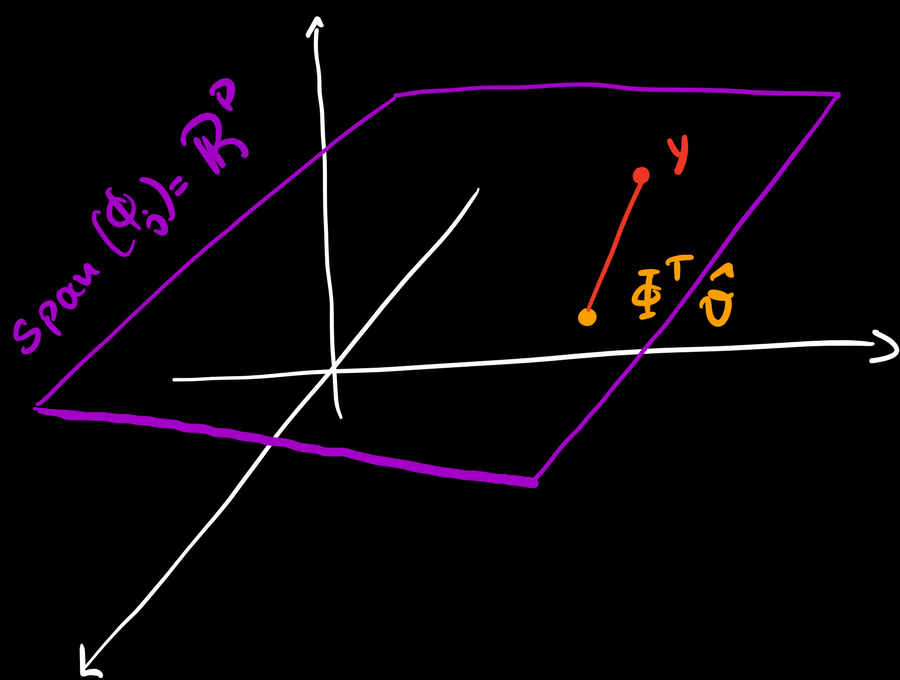
- $\operatorname{argmin} \sum (y_i - \Phi \vartheta_i)^2$: ϑ minimizes the total vertical distance between n points in $(p+1)$ dimensions. We are fitting a p-dimensional hyperplane. If we include an intercept, we fit $(p-1)$-dimensional plane offset from the origin.



- $\arg\min \| Y - \Phi \vartheta \|_2^2 = \arg\min \| Y - \sum_{j=1}^{p} \Phi_j^T \vartheta_j \|_2^2$

$\hat{\vartheta}$ contains the coefficients of the column vector of Φ. $\Phi \hat{\vartheta}$ is the orthogonal projection of Y onto the column space of Φ denoted by span(Φ_j)

we are minimizing the perpendicular distance with Y



This means that

$$\Phi\hat{\vartheta} = \Phi(\Phi^T\Phi)^{-1}\Phi^T y = \Pi_\Phi y$$

where Π_ϕ is a projection on $\mathrm{Span}_j \Phi_j$

Definition: We say that the linear model is Gaussian if $z \sim \mathcal{N}_0(0,1)$

Then $y|X \sim \mathcal{N}_0(\vartheta\phi(x), \mathbb{1}_n)$

with density:

$$f(y;\vartheta,\sigma^2) = (2\pi)^{-n/2}\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\|y-\Phi\vartheta\|^2\right)$$

$$= (2\pi)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}\|y\|^2 + \frac{1}{\sigma^2}\langle y, \phi\vartheta\rangle$$

$$- \frac{1}{2\sigma^2}(\Phi\theta)^2 - 2n\log\sigma\Big)$$

$$= \underbrace{(2\pi)^{-n/2}}_{h(x)}\exp\Big(-\frac{1}{2\sigma^2}\|y\|^2 + \langle\frac{\phi\vartheta}{\sigma^2}, \Pi_\phi y\rangle \underbrace{- \frac{1}{2\sigma^2}\|\phi\vartheta\|^2 - 2n\log\sigma}_{d(\vartheta)}$$

$\Rightarrow$ This is a 2-dim exponential family with statistics:

$T_1(Y) = \Pi_\Phi Y$ $\qquad$ $T_2(Y) = \|Y\|^2$

$c_1(\vartheta) = \frac{\Phi\vartheta}{\sigma^2}$ $\qquad$ $c_2(\vartheta) = -\frac{1}{2\sigma^2}$

$\Rightarrow$ if Φ has full rank then T_1, T_2

$T(Y) = (T_1(Y), T_2(Y))$ is sufficient and complete.

Theorem (Gauss-Markov):

The estimator $(\Phi^T\Phi)^{-1}\Phi^T Y$ is an unbiased estimator of ϑ, optimal among linear unbiased estimators and: BLUE

$$\| \Phi(\Phi^T\Phi^{-1})\Phi^T Y - Y\|^2$$

is an unbiased estimator of σ^2. In the Gaussian case, these estimators are optimal among all unbiased estimators of ϑ, $(n-k)\sigma^2$

Proof: Let VY be a linear estimator of ϑ.

$$E_{\vartheta,\sigma^2}(VY) = E_{\vartheta,\sigma^2}(V(\Phi\vartheta + \sigma Z)) = V\Phi\vartheta + \cancel{V\sigma\mathbb{E}(Z)}^{\,0}$$

So $V\Phi\vartheta = \vartheta \Rightarrow V\Phi = I_p \quad \forall\,\vartheta$

so any unbiased linear estimator satisfy

$$V\Phi = I_p = \mathbb{1}_p$$

But for the chosen $V = (\Phi^T\Phi)^{-1}\Phi$

$\Rightarrow \quad V\Phi = (\Phi^T\Phi)^{-1}\Phi^T\Phi = \mathbb{1}_p$

$\Rightarrow \quad (\Phi^T\Phi)^{-1}\Phi y$ is unbiased. Now we minimize the risk (MSE).

$$\mathbb{E}_{\theta,\sigma}(\|Vy-\theta\|^2) = \mathbb{E}(\|V\Phi\theta + \sigma z - \theta\|^2) = \sigma^2\, \mathbb{E}_{\sigma^2,\theta}(\|Vz\|^2)$$

$$= \sigma^2\, \mathbb{E}\Big(\sum_i \big(\sum_l V_{il} z_l\big)^2\Big) = \sigma^2 \sum_i \mathbb{E}\Big(\sum_{j,l} V_{ij} V_{il} z_j z_l\Big)$$

$$= \sigma^2 \sum_i V_{ij} V_{il} = \sigma^2 \sum_{il} V_{il}^2$$

Now we have the constrained optimization problem $\mathbb{E}(\|Vy-\theta\|^2)$

s.t. $\quad V\Phi = \mathbb{1}_p$

$$\mathbb{E}(\|Vy-\theta\|^2) - \lambda(V\Phi - \mathbb{1}_p) = \sigma^2\|V\|^2 - \lambda(V\Phi - \mathbb{1}_p)$$

$$\xRightarrow{\partial_V = 0} \quad 2\sigma^2 \cdot V = \lambda\Phi^T \Rightarrow V = \frac{\lambda}{2\sigma^2}\Phi^T$$

but since $\quad V\Phi = \frac{\lambda}{2\sigma}\Phi^T\Phi = \mathbb{1}_p$

$$\Rightarrow \lambda = 2\sigma^2 (\Phi^T \Phi)^{-1}$$

$$\Rightarrow Y = \frac{2\sigma^2}{2\sigma^2} (\Phi^T \Phi)^{-1} \cdot \Phi^T$$

Consinder now the estimator

$$\| \Pi_\Phi Y - Y \|^2 = \| \Pi_\Phi (\Phi \theta + \sigma Z) - (\Phi \theta + \sigma Z) \|^2$$

$$= \| (\Pi_\Phi \Phi - \Phi) \theta + \sigma (\Pi_\Phi Z - Z) \|^2$$

$$= \sigma^2 \| \Pi_\Phi \cdot Z - Z \|^2$$

if $\Pi_\Phi = \Pi_{1 \dots k} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 0 & \\ & & & & & 0 \end{pmatrix}$

$$\mathbb{E} (\| \Pi_\Phi Z - Z \|^2) = \sum_{j=k+1}^{n} \mathbb{E} (Z_k^2) = n - k$$