Now we want to fit the data to our chosen statistical model (linear regression) by using residual sum of squares → cost function.

- We show that minimizing a sum-of-squares error fxn could be motivated as the ML solution under an assumed Gaussian noise model!
- You can take an inference method, you will need this step because you have to compute how well you come up with beliefs.

  - Apply "maximum likelihood" method.

Select β such that it maximizes the probability of observing the data given.

$$\hat{\beta} \in \arg\max_{\beta} \Big\{ \underbrace{P(\varepsilon_1, \dots, \varepsilon_n \mid \beta)}_{\text{likelihood}} \Big\}$$

$\hat{\beta}$ = estimate value

We fit a polynomial fxn to data set by minimizing a sum-of-squares error fxN.
objective: is to find the β that maximizes the objective fxn $P(x_1, \dots, x_n \mid \theta)$

Change to log-likelihood fxn to make computation easier and take the negative to find lowest residual error

cost/error fxN $= \arg\min_{\beta} \Big[-\log P(\varepsilon_1, \dots, \varepsilon_n \mid \beta) \Big]$

* put a minus sign if you want to look for the min β.

$$= \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

// cost fxN.
// minimize the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

Change above expression into matrix notation.

$$= \arg\min_{\beta} (Y - X\beta)^T (Y - X\beta)$$

, where X is n × (d+1) matrix
Each row is an input vector x.
Y is an n × 1 of the outputs in the training set.

In order to compute the minimum condition ($\hat{\beta}$), we set the equation equal to ∅ and differentiate to minimize.

Minimizing of: $(Y - \vec{X}\beta)^T (Y - \vec{X}\beta)$

$$0 = (Y - \vec{X}\beta)^T (Y - \vec{X}\beta) \frac{d}{d\beta}$$

$$-2\vec{X}^T (Y - \vec{X}\beta) = 0$$

$$= -2\vec{X}^T Y + 2\vec{X}^T \vec{X}\beta = 0$$

$$-2X^T Y + 2X^T X \beta = 0$$
$$2X^T X \beta = 2X^T Y$$
$$X^T X \beta = X^T Y$$
$$\beta = \frac{X^T Y}{X^T X}$$

$$\frac{((d+1) \times n)(n \times 1)}{((d+1) \times n)(n \times (d+1))} = \frac{((d+1) \times 1)}{((d+1) \times (d+1))}$$

↑ matrix needs to be nonsingular!

CLOSED-FORM SOLUTION!

Find Estimate of $\hat{\beta}_{ML}$:

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

⇒ Prediction:

$$\boxed{\hat{Y} = X^T \hat{\beta}}$$

If so, then

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

X : (d+1) × n

n

d+1

β : (d+1) × 1

(d+1)

⇒

Y:

$(n \times \cancel{(d+1)})(\cancel{(d+1)} \times 1)$

n × 1