Statistical Justifications for regression

typical model of reality:

1) Sample points from unknown prob. distr. $X_i \sim D$
2) y-values are sum of a non-random fn. and random noise
   ↳ $Y_i = g(X_i) + \varepsilon_i$

assumption: reality is described by $g(\cdot)$, deterministic

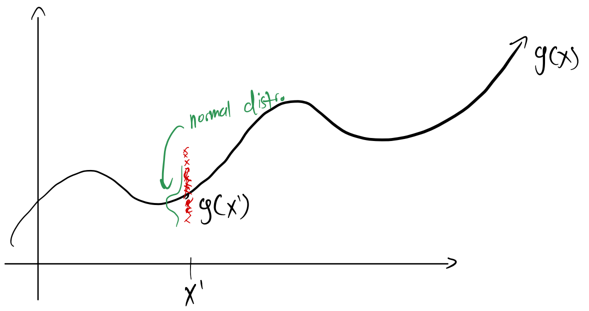
$\varepsilon_i \sim D'$, D' has zero mean

Why add ε_i?
- account for statistical error when measuring data
- for simplicity, we assume ε_i is indep. of X_i

goal of regression: find a fn h that estimates g:

$$h(\vec{x}) = E[Y \mid X = \vec{x}] = E[g(\vec{x}) + \varepsilon \mid X = \vec{x}] = g(\vec{x}) + E[\varepsilon] = g(\vec{x})$$

hypothesis   label   datapoint   0-mean



LS Regression for MLE

Suppose $\varepsilon_i \sim N(0, \sigma^2) \implies Y_i \sim N(g(x_i), \sigma^2)$

$$\ln f(Y_i) = \frac{-(Y_i - \mu)^2}{2\sigma^2} - C = \frac{-(Y_i - g(x_i))}{2\sigma^2} - C$$

log likelihood $\ell(g : X, y) = \ln(f(y_1) \cdot \ldots \cdot f(y_n))$

$$= \ln f(y_1) + \ldots + \ln f(y_n)$$

$$= \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - g(X_i))^2 - C$$

estimate on parameter g

↳ choose g that minimizes this

↳ normally-distr. noise $\implies$ use LS cost fn.

risk for hypothesis h: $R(h) = E[L] \quad \forall \vec{x} \in \mathbb{R}^d, y \in \mathbb{R}$

- With a discriminative model, we don't know distr. D of X, but want to minimize risk
- In contrast, a generative model, we estimate distr. & derive the expected loss

* to approx. the distr. w/ a discriminative model, we pretend that the sample points are the distr.

empirical distribution: discrete uniform distribution over the sample points
↳ put all sample pts in hat & pick randomly w/ equal prob.

empirical risk: expected loss under empirical distribution

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x_i), y_i) \quad , \quad \lim_{n\to\infty} \hat{R}(h) = R(h)$$

hopefully good approx of $R(h)$

empirical risk minimization: go from $\hat{R}(h)$ to $R(h)$

takeaway: this is why we minimize the sum of loss functions.

Sample Cov: Covariance matrix of empirical distribution

MLE again: what cost fn. should we use to interpolate probabilities?

- Actual prob. that X_i is in the class is y_i
- predicted prob. is $h(x_i)$

- Imagine β duplicates of pt. X_i
- $y_i \cdot \beta$ pts are in the class
- $(1-y_i)\beta$ not in class

prob we generate $y_i\beta$ copies in class

likelihood $\mathcal{L}(h; X, y) = \prod_{i=1}^{n} h(X_i)^{y_i\beta} \, (1-h(x_i))^{(1-y_i)\beta}$

$$\text{log likelihood } \ell(h) = \ln \mathcal{L}(h) = \beta \sum_{i=1}^{n} \underbrace{\left(y_i \ln h(x_i) + (1-y_i)\ln(1-h(x_i))\right)}_{\text{logistic/cross entropy loss}}$$

WOAH

maximizing this expr same as minimizing the logistic loss fn $= -\beta \sum$ logistic loss fn $L(h(x_i), y_i)$

takeaway: max likelihood $\Rightarrow$ minimize Σ logistic losses

BIAS – VARIANCE DECOMPOSITION

2 - Sources of error in hypothesis:

bias: error due to inability of hypothesis to fit to g perfectly
e.g. fitting a quadratic g w/ linear h

Variance: error due to fitting random noise in data
e.g. fit linear g w/ a linear hypothesis, yet $h \neq g$

model: - $X_i \sim D$, $\varepsilon_i \sim D'$ $Y_i = g(X_i) + \varepsilon_i$, ε_i is 0-mean
- fit h to X, Y, hope $h \approx g$ h is a RV since it depends on random X & Y

h is random $\Rightarrow$ its weights are random

Consider arbitrary pt $z \in \mathbb{R}^d$ (not necessarily a sample pt.)
& $\gamma = g(z) + \varepsilon$, $\varepsilon \sim D'$
- label at z is random due to random ε

note: $E[\gamma] = g(z)$, $Var(\gamma) = Var(\varepsilon)$ since only ε is random

risk fn when loss = squared error

$\hookrightarrow R(h) = E[L(h(z), \gamma)]$

take expectation over all possible training sets X, Y & possible values of γ

$\rightarrow R(h) = E[(h(z) - \gamma)^2]$

$= E[h(z)^2] + E[\gamma^2] - 2E[\gamma h(z)]$

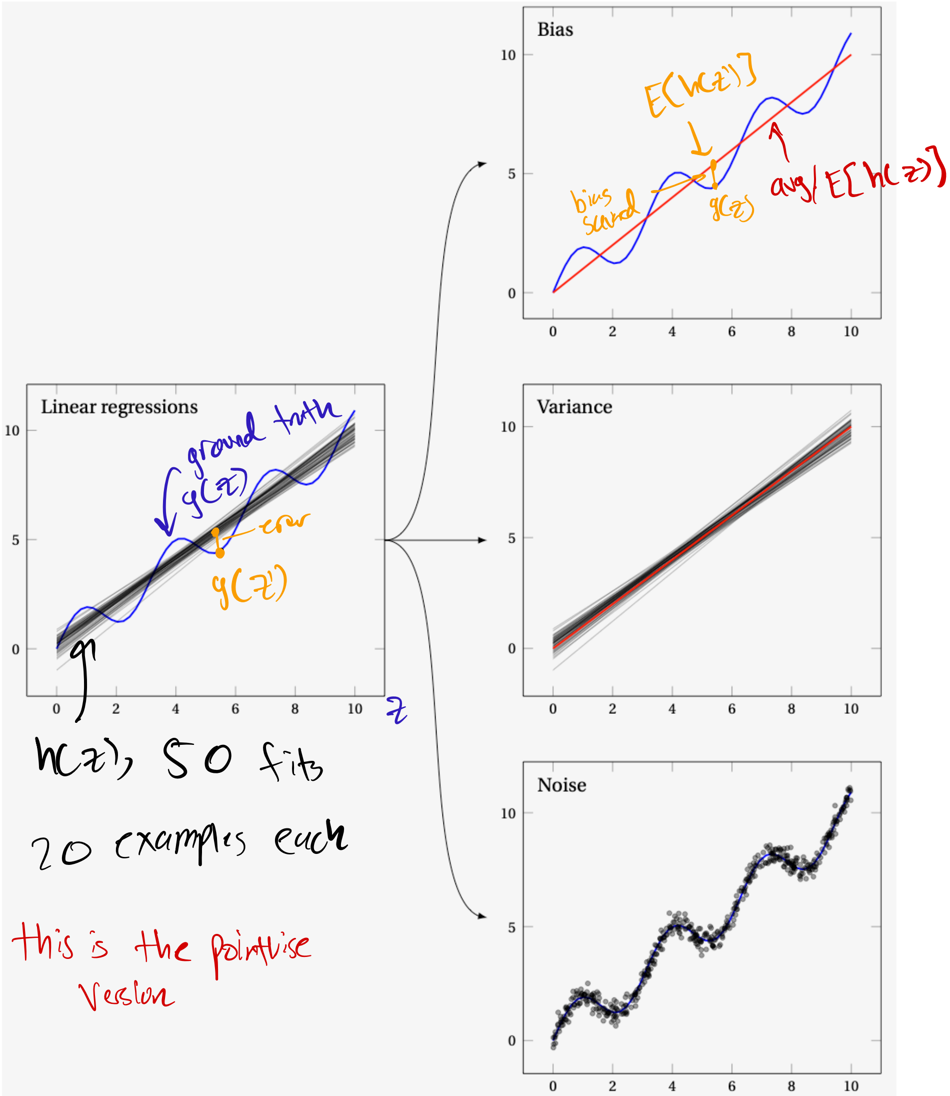
indep since h(z) only relies on X, Y
γ only random due to noise for test error

recall: $Var(x) = E[x^2] - E[x]^2$

$\rightarrow R(h) = Var(h(z)) + E[h(z)]^2 + Var(\gamma) + E[\gamma]^2 - 2E[\gamma]E[h(z)]$

uncomplete the square: $(E[h(z)] - E[\gamma])^2 + Var(h(z)) + Var(\gamma)$

$$= \underbrace{(E[h(z)] - g(z))^2}_{\text{bias}^2 \text{ of method}} + \underbrace{Var(h(z))}_{\text{var. of method}} + \underbrace{Var(\varepsilon)}_{\text{irreducible error}}$$



Mean Version: let $z \sim D$ be random variable, take mean over D of bias^2, Variance

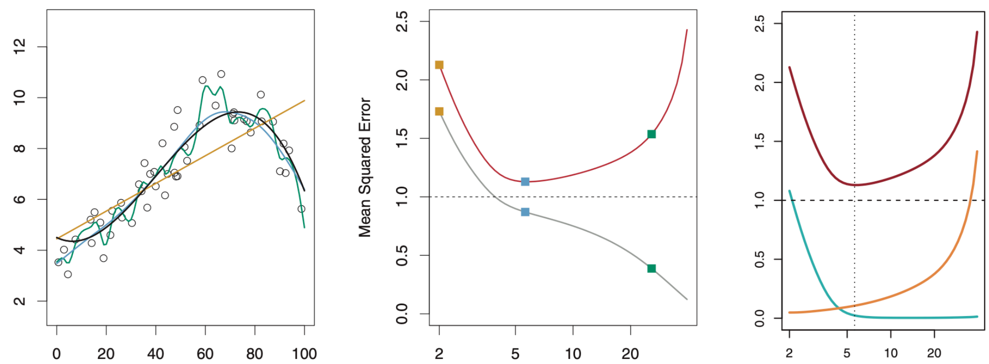
- underfitting $\to$ too much bias
- overfitting $\to$ too much Variance
- training error reflects bias, <u>not Variance</u>.
  - ↳ test error reflects both
- for many distributions, more data makes Variance go away as $n \to \infty$
- if h can fit g exactly, bias $= 0$ as $n \to \infty$
- If h can't fit g, bias is large at "most points"
- adding good feature $\to$ reduce bias
  bad feature won't increase bias in general
- adding a feature, good or bad, increases Variance

- can't reduce irreducible error

- noise in test set only affects $Var(\varepsilon)$

- noise in training set affects bias & var, not irr. error

- we can't measure bias or var on real world data precisely
  - ↳ can't know g exactly
  - ↳ but, we can test algo. by using g & making synthetic data



example: least squares linear regression

- suppose no fictitious dimension
  - ↳ decision function passes through origin

$g(\vec{z}) = \vec{v}^T\vec{z}$ (linear ground truth)

↳ if no noise, $h(\cdot)$ can be exact fit

$\vec{e}$ is noise vector, $e_i \sim N(0, \sigma^2)$

$\Longrightarrow \vec{y} = X\vec{v} + \vec{e}$, don't know $\vec{v}$ or $\vec{e}$

linear regression: $\vec{w} = X^+\vec{y} = X^+(X\vec{v} + \vec{e}) = \vec{v} + X^+\vec{e}$

want: $\vec{w} = \vec{v}$ but noise in $\vec{y}$ becomes noise in $\vec{w}$

↳ $X^+\vec{e}$ is the noise in the weights

$\rightarrow E[h(z)] - g(z) = E[\vec{w}^T\vec{z}] - \vec{v}^T\vec{z} = E[\vec{z}^T X^+ \vec{e}] = \vec{z}^T E[X^+] \cdot E[\vec{e}] = 0$

* doesn't imply $h(z) - g(z)$ always 0

- difference can be pos & neg but mean over training set should be 0
  - ↳ these deviations captured in variance

bias $=0 \implies$ perfect fit possible

but if perfect fit possible, not all models give 0 bias

↳ benefit of squared error
  - if diff. loss fn, might have nonzero bias even if we're fitting a linear h to a linear g

$$\mathrm{Var}(h(z)) = \mathrm{Var}(\vec{w}^T \vec{z}) = \mathrm{var}(\vec{z}^T \vec{v} + \vec{z}^T X^+ \vec{e}) = \mathrm{var}(\vec{z}^T X^+ \vec{e})$$

isotropic, norm. distr. $\vec{e}$

$$\mathrm{Var}(\underbrace{\vec{z}^T X^+ \vec{e}}_{\text{dot prod}}) = \sigma^2 |\vec{z}^T X^+|^2 = \sigma^2 \vec{z}^T (X^T X)^{-1} X^T X (X^T X)^{-1} \vec{z}$$

$$= \sigma^2 \vec{z}^T (X^T X)^{-1} \vec{z}$$

↳ Same as 1D gaussian along dir. of $\vec{z}^T X^+$

- If $E[X]=0$, $X^T X \to n\,\mathrm{Cov}(D)$ as $n \to \infty$

$\implies z \sim D$, $\mathrm{var}(h(z)) = \sigma^2 \frac{d}{n}$

takeaways:

- 0 bias when $h(\cdot)$ can fit ground truth $g(\cdot)$
  ↳ nice prop. of squared error loss
- Variance of residual sum of squares (RSS) decreases as $\frac{1}{n}$, increases as D, or $O(d^p)$ if we use degree p polynomials