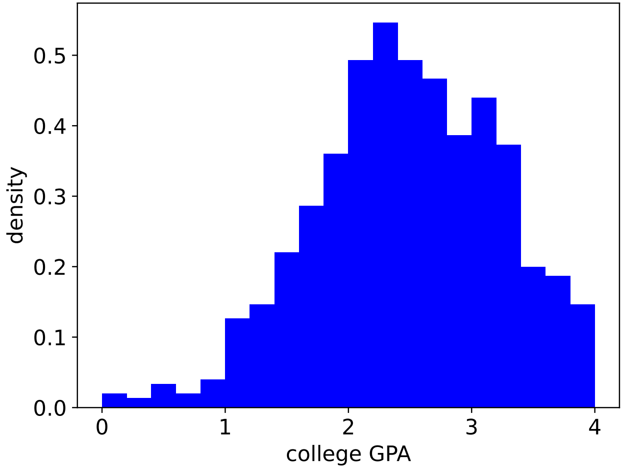
Linear regression

COMS 4771 Fall 2023

Dartmouth student dataset

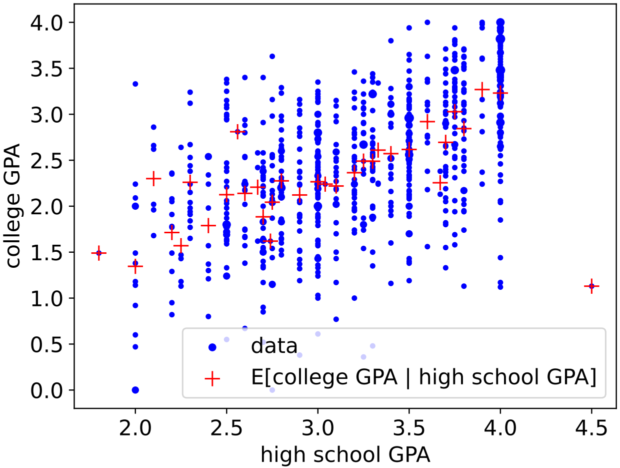
Dataset of 750 **Dartmouth students' (first-year) college GPA**[1]



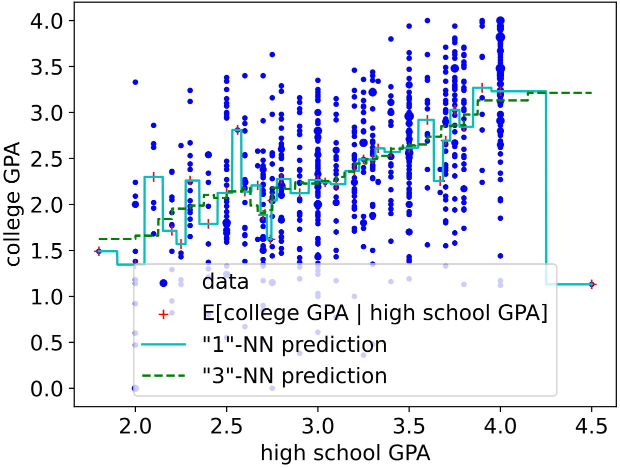
Mean 2.47
Standard deviation 0.75

[1] https://chance.dartmouth.edu/course/Syllabi/Princeton96/ETSValidation.html

Dartmouth dataset also has high school GPA of each student
Question: Is high school GPA predictive of college GPA?



Attempting to exploit “local regularity” using NN



Possible "global" modeling assumption:

- Increase in high school GPA by Δ should give an increase in (expected) college GPA by $\propto \Delta$

- In other words,
$$\mathbb{E}[\text{college GPA} \mid \text{high school GPA}]$$
is ________ function of high school GPA

Least squares linear regression

$f : \mathbb{R} \to \mathbb{R}$ is linear if it is of the form

$$f(x) = mx + b$$

for some parameters $m, b \in \mathbb{R}$

Problem: given a dataset $\mathcal{S}$ from $\mathbb{R} \times \mathbb{R}$, find (parameters of) a linear function $f(x) = mx + b$ of minimal sum of squared errors (SSE)

$$\text{sse}[m, b] = \sum_{(x,y) \in \mathcal{S}} (mx + b - y)^2$$

Method of solution is called ordinary least squares (OLS)

Minimizers of SSE must be zeros of the two partial derivative functions:

$$\frac{\partial \, \mathrm{sse}}{\partial m}[m, b] = 2 \sum_{(x,y) \in \mathcal{S}} (mx + b - y)x = 0$$

$$\frac{\partial \, \mathrm{sse}}{\partial b}[m, b] = 2 \sum_{(x,y) \in \mathcal{S}} (mx + b - y) = 0$$

Two linear equations in two unknowns

Together, the equations are called the normal equations

Equivalent form:

$$\begin{aligned} \operatorname{avg}(x^2)\, m &+ \operatorname{avg}(x)\, b &= \operatorname{avg}(xy) \\ \operatorname{avg}(x)\, m &+ b &= \operatorname{avg}(y) \end{aligned}$$

where

$$\operatorname{avg}(x) = \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} x, \qquad \operatorname{avg}(x^2) = \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} x^2,$$

$$\operatorname{avg}(xy) = \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} xy, \qquad \operatorname{avg}(y) = \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} y$$

Solution to normal equations:

$$\begin{aligned} m &= \frac{\mathrm{avg}(xy) - \mathrm{avg}(x) \cdot \mathrm{avg}(y)}{\mathrm{avg}(x^2) - \mathrm{avg}(x)^2}, \\ b &= \mathrm{avg}(y) - m \cdot \mathrm{avg}(x) \end{aligned}$$

What if $\mathrm{avg}(x^2) = \mathrm{avg}(x)^2$?

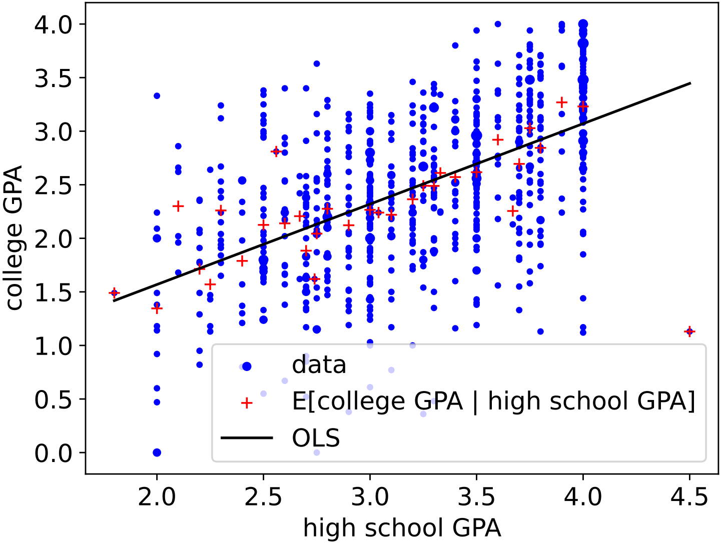
For Dartmouth dataset:

$$m = 0.751, \quad b = 0.067$$

RMSE:

$$\sqrt{\frac{1}{|\mathcal{S}|}\,\mathrm{sse}[m, b; \mathcal{S}]} = 0.629$$

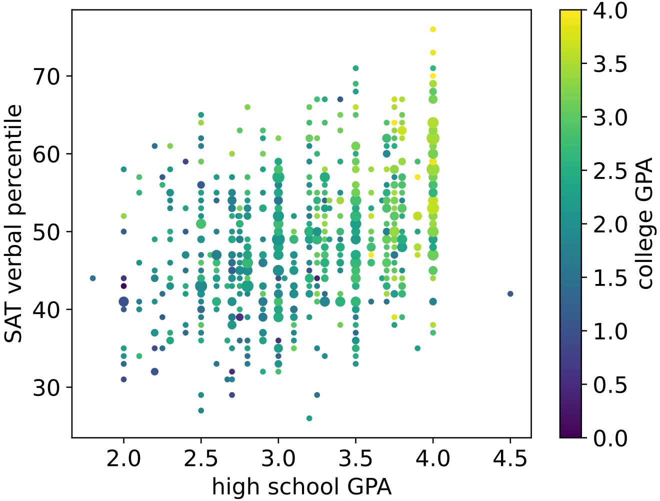
(Recall standard deviation of college GPA is 0.75)



4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
college GPA
2.0
2.5
3.0
3.5
4.0
4.5
high school GPA
data
E[college GPA | high school GPA]
OLS

Bivariate linear regression

Dartmouth dataset also includes SAT verbal percentiles



Linear function of two variables x_1 and x_2:

$$f(x_1, x_2) = m_1 x_1 + m_2 x_2 + b$$

Problem: given a dataset $\mathcal{S}$ from $\mathbb{R}^2 \times \mathbb{R}$, find (parameters of) a linear function $f(x_1, x_2) = m_1 x_1 + m_2 x_2 + b$ of minimal sum of squared errors

$$\mathrm{sse}[m, b; \mathcal{S}] = \sum_{(x_1, x_2, y) \in \mathcal{S}} (m_1 x_1 + m_2 x_2 + b - y)^2$$

Normal equations: three linear equations in three unknowns (m_1, m_2, b)

$$\begin{bmatrix} \text{avg}(x_1^2) & \text{avg}(x_1 x_2) & \text{avg}(x_1) \\ \text{avg}(x_2 x_1) & \text{avg}(x_2^2) & \text{avg}(x_2) \\ \text{avg}(x_1) & \text{avg}(x_2) & 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ b \end{bmatrix} = \begin{bmatrix} \text{avg}(x_1 y) \\ \text{avg}(x_2 y) \\ \text{avg}(y) \end{bmatrix}$$

Solve using elimination algorithm

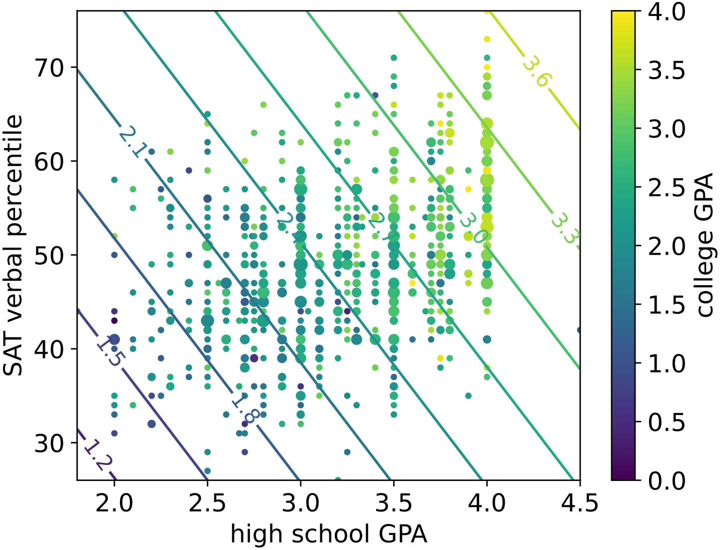
Dartmouth dataset: x_1 = high school GPA, x_2 = SAT verbal percentile

$$m_1 = 0.611, \quad m_2 = 0.024, \quad b = -0.639$$

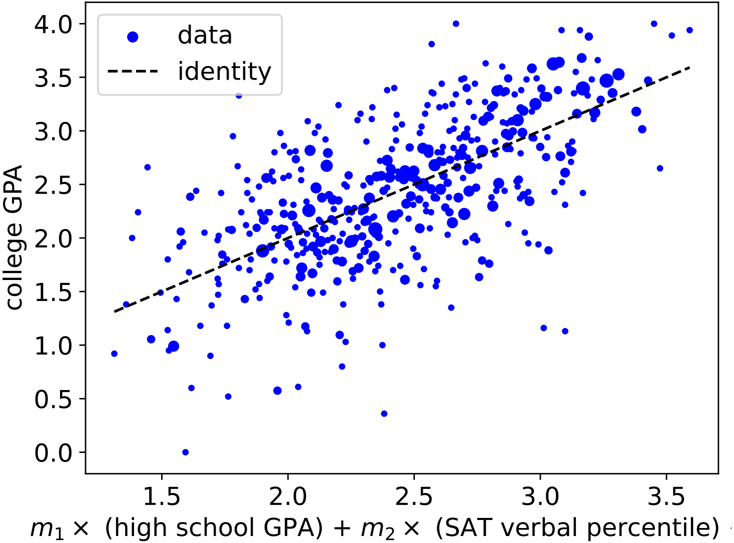
RMSE:

$$\sqrt{\frac{1}{|\mathcal{S}|}\,\mathrm{sse}[m_1, m_2, b; \mathcal{S}]} = 0.603$$

(Recall standard deviation of college GPA is 0.75)



SAT verbal percentile
70
60
50
40
30
2.0
2.5
3.0
3.5
4.0
4.5
high school GPA
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
college GPA
1.2
1.5
1.8
2.1
3.0
3.3
3.6


data
identity
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
college GPA
1.5
2.0
2.5
3.0
3.5
$m_1 \times$ (high school GPA) + $m_2 \times$ (SAT verbal percentile)

Linear algebra of ordinary least squares

(Homogeneous) linear function of d variables $x = (x_1, \ldots, x_d)$ is parameterize by d-dimensional weight vector $w = (w_1, \ldots, w_d)$:

$$f_w(x) = w^\mathsf{T} x$$

To handle inhomogeneous linear functions (i.e., affine functions), include an extra always-1 feature: $x_{d+1} = 1$

$$\begin{aligned} f_w(x) &= w^\mathsf{T} x \\ &= (w_1 x_1 + \cdots + w_d x_d) + \underline{\hspace{3cm}} \end{aligned}$$

Problem: given a dataset $\mathcal{S}$ from $\mathbb{R}^d \times \mathbb{R}$, find $w \in \mathbb{R}^d$ of minimal sum of squared errors

$$\mathrm{sse}[w; \mathcal{S}] = \sum_{(x,y) \in \mathcal{S}} (w^\intercal x - y)^2$$

Method of solution: OLS

Matrix notation: let $\mathcal{S} = ((x^{(i)}, y^{(i)}))_{i=1}^n$, and put

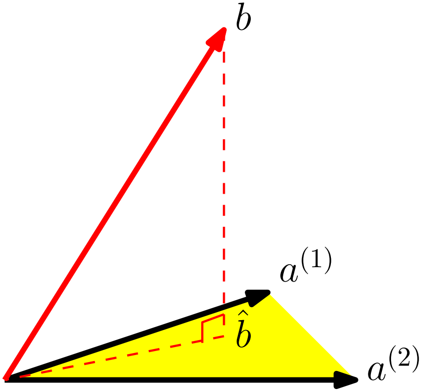
$$A = \begin{bmatrix} \longleftarrow & (x^{(1)})^\intercal & \longrightarrow \\ & \vdots & \\ \longleftarrow & (x^{(n)})^\intercal & \longrightarrow \end{bmatrix}, \quad b = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

so

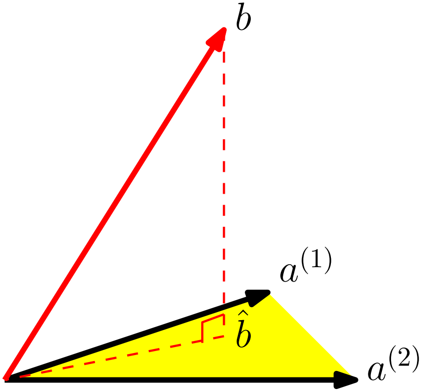
$$Aw = \begin{bmatrix} w^\intercal x^{(1)} \\ \vdots \\ w^\intercal x^{(n)} \end{bmatrix}, \quad Aw - b = \begin{bmatrix} w^\intercal x^{(1)} - y^{(1)} \\ \vdots \\ w^\intercal x^{(n)} - y^{(n)} \end{bmatrix}$$

Therefore

$$\|Aw - b\|^2 = \sum_{i=1}^n \underline{\hspace{6cm}}$$



$Aw \in \mathsf{CS}(A)$ for every $w \in \mathbb{R}^d$



How many ways to write $\hat{b}$ as a linear combination of the columns of A?

Normal equations in matrix notation

Key fact: $\mathsf{CS}(A)$ and $\mathsf{NS}(A^\intercal)$ are orthogonal complements

Summary:

- Normal equations: $(A^\intercal A)w = A^\intercal b$
- If $\text{rank}(A) = d$, then solution is unique
- Else, infinitely-many solutions
- Common choice for tie-breaking: minimum norm solution

$$\arg\min_{w \in \mathbb{R}^d} \|w\| \text{ s.t. } (A^\intercal A)w = A^\intercal b$$

```python
def learn(train_x, train_y):
  return np.linalg.pinv(train_x).dot(train_y)

def predict(params, test_x):
  return test_x.dot(params)
```

Statistical view of ordinary least squares

Normal linear regression model: Conditional distribution of Y given $X = x$ is

$$\mathrm{N}(w^{\mathsf{T}} x, \sigma^2)$$

- w and σ^2 are parameters of the model
- In this model, best possible MSE is σ^2

MLE in normal linear regression model

- Likelihood of w and σ^2:

$$L(w, \sigma^2) = \prod_{(x,y)\in\mathcal{S}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - w^{\mathsf{T}}x)^2}{2\sigma^2}\right)$$

- Log-likelihood:

$$\ln L(w, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{(x,y)\in\mathcal{S}} (y - w^{\mathsf{T}}x)^2 - \frac{|\mathcal{S}|}{2} \ln(2\pi\sigma^2)$$

- In terms of w, maximizing log-likelihood is same as minimizing SSE!

Statistical inference (example)

- Suppose you fit linear regression model to data, and find that $w \neq (0, \ldots, 0)$

  How confident are you in this finding?

Generalization

- Suppose $\mathcal{S} \overset{\text{i.i.d.}}{\sim} (X, Y)$
- OLS gives minimizer of empirical risk (for square loss, among linear functions)

$$\widehat{\text{Risk}}[w] = \frac{1}{n} \sum_{(x,y)\in\mathcal{S}} \text{loss}_{\text{sq}}(w^{\mathsf{T}} x, y)$$

  But we actually care about the (true) risk

$$\text{Risk}[w] = \mathbb{E}[\text{loss}_{\text{sq}}(w^{\mathsf{T}} X, Y)]$$

- Is empirical risk a good estimate of (true) risk?
  - Usually only if $|\mathcal{S}|$ is sufficiently large

Extreme example: $d = 1$, $|\mathcal{S}| = 2$, $\widehat{\text{Risk}}[w] = 0$