Linear Regression (in $\mathbb{R}^2$)

- other sorts of regression:
  "find a restricted pattern, and regress data to that pattern"

-----------------------------------

Linear Least Squares

Input: $P$ subset $\mathbb{R}^2$
Output: line $y = ax + b$
  s.t. $\sum_{p \in P} (p.y - a \cdot p.x + b)^2$ minimized
  "vertical distance"

We model $p.y = f(p.x) + \varepsilon_p$ where $\varepsilon_p \sim$ Gaussian noise

- $f(p.x) = a \cdot p.x + b$
- and minimize $\sum_{p \in P} \varepsilon_p^2$

solve for $a = \text{Cov}[p.x, p.y] / \text{Var}[p.x]$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\text{Cov}[x, y] = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i - \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right) \left( \frac{1}{n} \sum_{i=1}^{n} y_i \right)$$

$$\text{Var}[x] = \text{Cov}[x, x]$$

$$a = \frac{\langle p.x, p.y \rangle}{||p.x||^2}$$

solve for $b = \bar{y} - a \cdot \bar{x}$

-------

to fit $p.y = a \cdot p.x$

Let $X = P.x$

$$a = (X^T X)^{-1} \cdot X^T \cdot y$$

and $H_X = X \cdot (X^T X)^{-1} \cdot X^T$ is the "hat" matrix

since

$$\hat{y} = X \cdot a = H_X \cdot y$$

puts the hat on $y$

but this does not allow an "intercept" value $b$.

--> first shift $P' = P - \bar{P}$

then intercept $b = 0$

can shift back later
Gauss-Markov Theorem
linear regression is optimal of all linear models
- if must have 0 expected error (unbiased)
- all errors (\(\eps_p\)) uncorrelated, have equal variances
then: above minimizes least-squared error. (minimum variance)

Are we done? (no!)

4 issues:
- robustness to outliers (from \(L_2\))
- can have less error with bias
- y-distance, not distance to line
- matrix inverse can be expensive (randomization)
- (dense)

-------------

Theil-Sen estimator

Median slope of pairs of points.
For all \(p_i = (x_i, y_i)\) and \(p_j = (x_j, y_j)\) with \(x_i < x_j\)
let \(s_{i,j} = (y_j - y_i)/(x_j - x_i)\)
Let \(a = \text{median}_{i,j} \{s_{i,j}\}\)

Let \(b = \text{median}_i \{y_i - a x_i\}\)

more robust to outliers. (up to 29.3% corruption)

+ Siegel: (for \(x_i < x_j\) for \(i<j\))
  let \(s_i = \text{median} \{ s_{j,i} \mid j<i \} \cup s_{i,j} \mid i<j \})\)
  Let \(a = \text{median}_i \{s_i\}\)
  Let \(b = \text{median}_i \{y_i - a x_i\}\)

even more robust to outliers (up to 50% corruption)

Straight-forward in \(O(n^2)\)
also \(O(n \log n)\) algorithms...

-------------

Tikhonov Regularization (ridge regression)

assume \(\bar{p.y} = 0\) and \(\bar{p.x}=0\); hence \(b = 0\)

minimize: \(\sum_{p \in P} (a \ p.x - p.y)^2 + s \ a^2\)
where \(s\) is a tunable regularization (shrinkage) parameter
trades off having some bias for having less variance (regression to mean = no correlation)

\[ \hat{a} = \frac{\langle p.x, p.y \rangle}{\langle p.x, p.x \rangle^2 + s^2} \]

where \( X = P.x \) then \( a^2 = \|sI a\|^2 \)
\[ \hat{a} = (X^T X + s^2 I)^{-1} X^T b \]

or:
\[
\begin{align*}
\text{minimize:} & \quad \sum_{p \in P} (a p.x - p.y)^2 \\
\text{s.t.} & \quad a^2 < t
\end{align*}
\]

1-1 correspondence be between each solution with s and with t
as s decreases, t increases

------------------------------------

Lasso (basis pursuit)

assume \( \bar{p.y} = 0 \) and \( \bar{p.x}=0 \); hence \( b = 0 \)

\[
\begin{align*}
\text{minimize:} & \quad \sum_{p \in P} (a p.x - p.y)^2 + s |a| \\
\text{where} & \quad s \text{is a tunable regularization (shrinkage) parameter}
\end{align*}
\]

or:
\[
\begin{align*}
\text{minimize:} & \quad \sum_{p \in P} (a p.x - p.y)^2 \\
\text{s.t.} & \quad |a| < t
\end{align*}
\]

(in higher dimensions, we'll see, when t is small, some dimensions are 0!)

Way to compute optimal solution efficiently (LAR, see later lectures)

***** Up to here, great reference: Elements of Statistical Learning
Hastie, Tshibirani, Friedman

------------------------------------

PCA -> "orthogonal distance"

Don't explain y from x, but explain relationship between x and y.
- before we assume x was correct. Now there can be error/residuals in both

In \( \mathbb{R}^2 \) first center (double center):
\[
\begin{align*}
x_i &= x_i - \bar{x} \\
y_i &= y_i - \bar{y}
\end{align*}
\]
(from now assume they are already double-centered)

Find unit vector \( v \) (i.e. \( \|v\| = 1 \)) to minimize
PCA(v) = \sum_{p \in P} \|p - \langle p,v \rangle v \|^2

note that \langle p,v \rangle is a scaler: the length of p along v.
And v is a direction from the origin.
So \langle p,v \rangle v is the "projection" of p onto v.

and p - \langle p,v \rangle v is the distance to the projection.

--------
How do we find v?
Only depends on 1 parameter (angle)
PCA(v) is convex - up to antipodes.

in higher dimensions, we (essentially) just repeat this.