Lecture #19

- So far, we have considered problems of the form "is there sufficient evidence to reject a claim H_0: ~ in favor of H_1: ~"
- A different problem in statistics (closer to ML) is "predict Y given X" for random quantities X, Y. This is the regression problem.
- Of course, we hope and expect Y and X to be related in some sense — this is captured by the conditional distribution $f_{Y|X}(y|x)$. If X is highly predictive of Y, we would expect $f_{Y|X}(y|x)$ to be very localized in y for a fixed x.
- The extreme case is when x determines y, i.e. Y is no longer random when conditioned on X. This is a helpful theoretical case (and may be realistic in idealized physical settings), but we will typically see it does not hold in most messy real-world settings.
- In theory, the most reasonable goal in most cases is to estimate

$$\mathbb{E}(Y|X=x) = \int_{\mathbb{R}} y \cdot f_{Y|X}(y|x)\,dy$$

let's start by assuming both X and Y are $\mathbb{R}$-valued.

- Again, if knowing $X=x$ uniquely determines Y, then we can imagine

the relationship between (X,Y) is governed by a function $y = h(x)$. Then the conditional density in this case is just a Dirac mass:

$$f_{Y|X}(y|x) = \delta_{h(x)}(y)$$

needs to be interpreted in the sense of measures...

$$\rightarrow \text{"}=\text{"} \begin{cases} +\infty, & y = h(x) \\ 0, & y \neq h(x), \end{cases}$$

so that in this special case $\mathbb{E}(Y \mid X = x)$

$$= \int_{\mathbb{R}} y \cdot f_{Y|X}(y|x)\, dy$$

$$= \int_{\mathbb{R}} y \delta_{h(x)}(y)\, dy$$

$$= h(x),$$

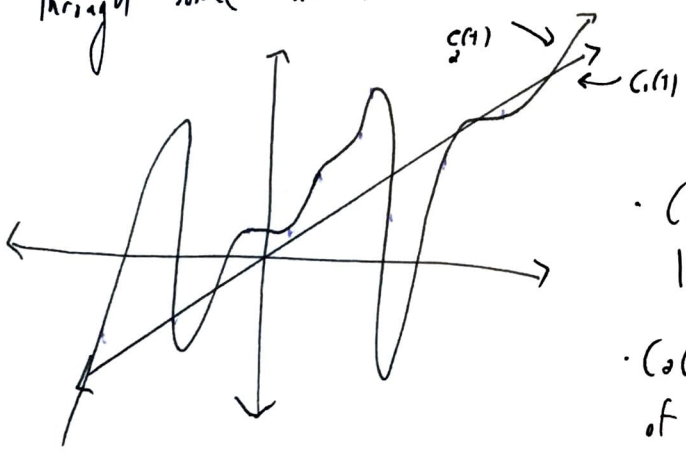
as we would hope.

- Let $r(x) := \mathbb{E}(Y \mid X = x)$ be the regression function. We typically want to estimate $r(x)$ given iid samples of the form $\{(x_i, y_i)\}_{i=1}^n$ drawn from the joint distribution of (X, Y).

- There are many, many ways to do this. It is an ongoing topic of academic and industrial research. We will focus on a few classical approaches.

Linear Regression

- Many regression methods coincide with some kind of "curve fitting", i.e. given observations $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^2$, find a curve $C(t)$ from some family that passes through some (or all) of the observations:



- $C_1(t)$ is from the family of linear functions
- $C_2(t)$ is from the family of polynomials of any degree.

- We can compute a good curve by the following very generic approach. Let $\mathcal{F}$ be a family of functions/curves (e.g. $\mathcal{F}$ is the space of polynomials of some degree). Given observations $\{(x_i, y_i)\}_{i=1}^n$, let us choose a curve-fitting function $C^*(t) \in \mathcal{F}$ as

$$C^*(t) = \underset{C(t) \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} |C(x_i) - y_i|^2$$

This chooses an element of $\mathcal{F}$ that yields the least sum-of-squares error in

predicting the data.

- This is elegant but has at least two major drawbacks:
  1.) Impossible to solve for arbitrary $\mathcal{F}$.
  2.) Fitting the observed data well does not guarantee good prediction on unseen data.

- Let's focus on the special case of $\mathcal{F}_0 = \{f_0(t) = c_0 \mid c_0 \in \mathbb{R}\}$ the space of constant functions. Very boring, but illustrative. Then we are solving

$$c_0^* = \underset{c_0 \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} |y_i - c_0|^2$$

HW: $c_0^* = \frac{1}{n} \sum_{i=1}^{n} y_i$.

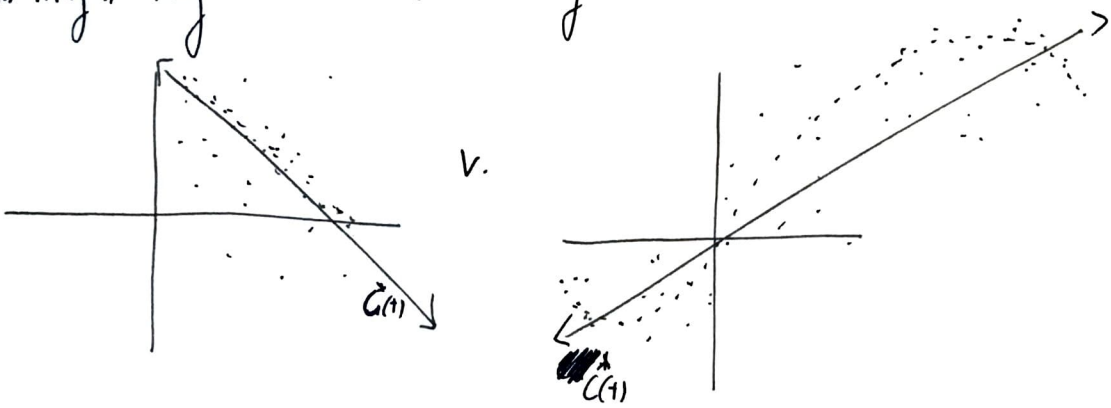
- So, the space of degree 0 polynomials is pretty unhelpful. What about the space of degree 1 polynomials? Let $\mathcal{F}_1 = \{f_1(t) = c_1 t + c_0 \mid c_0, c_1 \in \mathbb{R}\}$. Then our goal is to learn coefficients

$$(c_1^*, c_0^*) = \underset{(c_1, c_0) \in \mathbb{R}^2}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} |y_i - c_1 x_i - c_0|^2$$

HW: $c_1^* = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$, $\quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

$$c_0^* = \bar{y} - c_1^* \cdot \bar{x}$$

- This may or may not be a reasonable thing to do for the data:



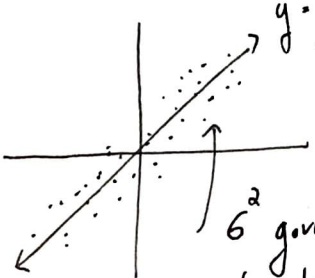
- The fit on the left is clearly more indicative of overall pattern than on the right. We can consider statistical models well-suited to the linear regression scheme as follows.

Defn: The simple linear regression model is $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where:

(1.) $\beta_0, \beta_1 \in \mathbb{R}$

(2.) $\mathbb{E}(\varepsilon_i \mid X_i) = 0$, $\mathrm{Var}(\varepsilon_i \mid X_i) = \sigma^2$ for some $\sigma^2 \in \mathbb{R}$.

- The idea is β_0, β_1 are defining a linear trend in the data, and ε_i is corrupting noise in the outputs:



σ^2 governs how much a typical observation deviates from the linear pattern

Remark: $\sigma^2 = 0$ corresponds to y deterministic given x, with $y = \beta_0 + \beta_1 x$.

- Under the SLRM, our curve-fitting coefficients match the "best estimates" on (β_0, β_1), if our goal is to make the noise estimates $\hat{\varepsilon}_i$ as little as possible.

- That is, given estimates $(\hat{\beta}_0, \hat{\beta}_1)$, we immediately get an estimate on $y_i | x_i$ via $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Then the noise estimate is $-(\hat{y}_i - y_i)$

$$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= \hat{\varepsilon}_i$$

- The least squares procedure under the SLRM is to choose $\hat{\beta}_0, \hat{\beta}_1$ to minimize $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$. The following follows (mostly) from the curve-fitting calculation.

Theorem: (1.) The estimates on (β_0, β_1) that minimize $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(2.) An unbiased estimate of σ^2 is $\frac{1}{(n-2)} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$.

Proof: HW.