Chapter 11: Inference on the Regression Function

Theoretical Foundations of Conformal Prediction

Yoonseo Choi

December 29, 2025

Table of Contents

- 미지의 분포 P에서 추출된 n개의 iid 데이터 $\{(X_i, Y_i)\}$가 주어졌을 때, 회귀 함수

$$\mu_P(x) = \mathbb{E}_P[Y|X = x]$$

  를 추정한다고 하자.
- **핵심 목표:** $\mu_P(x)$에 대한 분포 무관 신뢰구간을 추정하자. (distribution-free validity)
- **세부 목표 1:** 폭이 좁은(informative) 구간을 찾자.
- **세부 목표 2:** 특히, "좋은" 분포에 대해서는 구간 길이가 0으로 가는 신뢰구간을 찾자.
  - $\mathbb{E}[Leb(C(X_{n+1})] \to 0$

Definition: Distribution-free Confidence Interval

알고리즘 $\mathcal{C}$가 다음 조건을 만족하면 레벨 $1-\alpha$의 distribution free confidence interval for regression at level $1-\alpha$라고 한다.

$$\mathbb{P}(\mu_P(X_{n+1}) \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha \tag{1}$$

for **any** distribution P on $\mathcal{X} \times \mathcal{Y}$.

- Suppose $\mathcal{Y} = \mathbb{R}$ (unbounded from above and below).
- Can we perform distribution-free inference on the mean?

Theorem 11.1: The Bahadur-Savage Theorem

Let $\mathcal{Y} \subseteq \mathbb{R}$ be unbounded. Suppose $\mathcal{C}$ satisfies:

$$\mathbb{P}(\mathbb{E}_P[Y] \in \mathcal{C}) \geq 1 - \alpha$$

for any distribution P with finite mean. Then, for any P and any $y \in \mathbb{R}$:

$$\mathbb{P}(y \in \mathcal{C}) \geq 1 - \alpha$$

In particular, $\mathbb{E}[\text{Leb}(\mathcal{C})] = \infty$ (if $\alpha < 1$).

- **Y가 unbounded라면, 평균에 대한 유한한 길이의 CI를 만들 수 없다.** 따라서 11장에서는 Y가 bounded라고 가정할 것이다.
- 증명 아이디어: 분포 P를 변형하여 P와 아주 유사해서 평균은 같되, 극단적으로 튀는 값이 있는 분포 P'를 생성하고, P와 P' 간 커버리지가 유사함을 이용한다.

2. Proof of Theorem 11.1 (validity)

1. W.L.O.G, let $y \geq \mathbb{E}_P[Y]$. Fix $\epsilon > 0$.
2. Let $y' \in \mathcal{Y}$ such that $y' \geq \frac{y-(1-\epsilon)\mathbb{E}_P[Y]}{\epsilon}$.
   (This is possible because $\mathcal{Y}$ is unbounded).
3. Define distribution P' on $\mathcal{Y}$ as:
$$P' = (1-\epsilon')P + \epsilon'\delta_{y'}$$
   where $\epsilon' = \frac{y-\mathbb{E}_P[Y]}{y'-\mathbb{E}_P[Y]}$.
4. Then, $d_{TV}(P, P') \leq \epsilon' \leq \epsilon$.
   (Note: $d_{TV} \leq$ mixing weight).
5. Also, by construction, $\mathbb{E}_{P'}[Y] = y$.
6. From the validity of $\mathcal{C}$ (distribution-free guarantee):
$$\mathbb{P}_{(P')^n}(y \in \mathcal{C}) = \mathbb{P}_{(P')^n}(\mathbb{E}_{P'}[Y] \in \mathcal{C}) \geq 1-\alpha$$
7. Using the definition of Total Variation distance:
$$\begin{aligned}\mathbb{P}_{P^n}(y \in \mathcal{C}) &\geq \mathbb{P}_{(P')^n}(y \in \mathcal{C}) - d_{TV}(P^n, (P')^n) \\ &\geq \mathbb{P}_{(P')^n}(y \in \mathcal{C}) - n\epsilon' \\ &\geq 1-\alpha-n\epsilon\end{aligned}$$
8. Since ϵ is arbitrary, $\mathbb{P}_P(y \in \mathcal{C}) \geq 1-\alpha$.

2. Proof of Theorem 11.1 (interval length)

$$\begin{aligned}
\mathbb{E}_{P^n}[\text{Leb}(\mathcal{C})] &= \mathbb{E}_{P^n}\left[\int_{y\in\mathbb{R}} \mathbb{I}\{y \in \mathcal{C}\}dy\right] \\
&= \int_{y\in\mathbb{R}} \mathbb{P}_{P^n}(y \in \mathcal{C})dy \quad \text{(by Fubini's Theorem)} \\
&\geq \int_{y\in\mathbb{R}} (1-\alpha)dy = \infty \quad \text{(by Claim 1)}
\end{aligned}$$

- Q. 가정을 완화하면(ex: 모든 nonatomic P에 대해) unbounded여도 구간 길이가 유한이 될 수 있을까요?
- A. 아니요. P가 nonatomic이더라도, 증명과정의 $\delta_{y'}$ 대신 H Unif([y'-η,y'+η]를 사용한다면, $d_{TV}(P,Q) = d_{TV}(P,(1-\epsilon')P + \epsilon' H) \leq \epsilon' d_{TV}(P,H) \leq \epsilon'$이므로 이후 증명 과정을 그대로 적용하면 구간 길이가 무한임을 구할 수 있습니다.
- A. bahadur-savage thm에 대한 원 논문을 보면(The Nonexistence of Certain Statistical Procedures in Nonparametric Problems, R. R. Bahadur, Leonard J. Savage, 1956) 평균은 꼬리 확률에 민감하기 때문에 이러한 결과가 나왔다고 설명합니다. 때문에 꼬리 확률을 제어하는 가정이 있다면 unbounded여도 구간 길이가 유한이 될 수도 있다고 생각합니다. (sub gaussian 등)

Setup:

- $\mathcal{X} = \{x_1, \ldots, x_K\}$ is a finite set.
- Inference on $\mu_P(x) \rightarrow$ inference on $\mu_P(x_k)$for each k.
- X가 이산형이면 valid하면서 유한한 길이의 신뢰구간을 만들 수 있다.

Theorem 11.2

Let $\mathcal{Y} \subseteq [a, b]$. For each k, let $n_k = \sum \mathbb{I}\{X_i = x_k\}$. Define:

$$\mathcal{C}(x_k) = \begin{cases} \hat{\mu}(x_k) \pm (b-a)\sqrt{\frac{\log(2/\alpha)}{2n_k}} & (\text{if } n_k \geq 1) \\ [a, b] \quad (\text{if } n_k = 0) \end{cases}$$

Then $\mathcal{C}$ satisfies distribution-free validity (11.1), and:

$$\mathbb{E}[\text{Leb}(\mathcal{C}(X_{n+1}))] \leq 2(b-a)\sqrt{\log(2/\alpha)} \cdot \sqrt{\frac{K}{n}}$$

3. Proof of Theorem 11.2 (validity)

1. Let $p_k = \mathbb{P}_P(X = x_k)$. Condition on the training data $\mathcal{D}_n$:

$$\mathbb{P}(\mu_P(X_{n+1}) \notin \mathcal{C}(X_{n+1}) \mid \mathcal{D}_n) = \sum_{k=1}^{K} p_k \mathbb{I}\{\mu_P(x_k) \notin \mathcal{C}(x_k)\}$$

2. After marginalizing over $\mathcal{D}_n$:

$$\begin{aligned}
\mathbb{P}(\mu_P(X_{n+1}) \notin \mathcal{C}(X_{n+1})) &= \sum_{k=1}^{K} p_k \cdot \mathbb{P}(\mu_P(x_k) \notin \mathcal{C}(x_k)) \\
&= \sum_{k=1}^{K} p_k \cdot \mathbb{E}[\mathbb{P}(\mu_P(x_k) \notin \mathcal{C}(x_k) \mid n_k)] \\
&= \sum_{k=1}^{K} p_k \cdot \mathbb{E}\left[\mathbb{I}\{n_k \geq 1\} \cdot \mathbb{P}\left(|\hat{\mu}_k - \mu_P(x_k)| > (b-a)\sqrt{\frac{\log(2/\alpha)}{2n_k}} \;\middle|\; n_k\right)\right]
\end{aligned}$$

3. By Hoeffding's inequality:

$$\mathbb{P}\left(|\hat{\mu}(x_k) - \mu_P(x_k)| > (b-a)\sqrt{\frac{\log(2/\alpha)}{2n_k}} \;\middle|\; n_k\right) \leq \alpha$$

4. Thus, $\mathbb{P}(\mu_P(X_{n+1}) \notin \mathcal{C}(X_{n+1})) \leq \sum_{k=1}^{K} p_k \cdot \mathbb{E}[\mathbb{I}\{n_k \geq 1\} \cdot \alpha] \leq \alpha \sum_{k=1}^{K} p_k = \alpha$

3. Proof of Theorem 11.2 (interval length)

1. By definition, $\text{Leb}(\mathcal{C}(x_k)) \leq 2(b-a)\sqrt{\frac{\log(2/\alpha)}{n_k+1}}$.

$$\begin{aligned}
\mathbb{E}[\text{Leb}(\mathcal{C}(X_{n+1}))] &= \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{I}\{X_{n+1} = x_k\} \cdot \text{Leb}(\mathcal{C}(x_k))\right] \\
&= \sum_{k=1}^{K} p_k \cdot \mathbb{E}[\text{Leb}(\mathcal{C}(x_k))] \\
&\leq \sum_{k=1}^{K} p_k \cdot \mathbb{E}\left[2(b-a)\sqrt{\frac{\log(2/\alpha)}{n_k+1}}\right] \\
&= 2(b-a)\sqrt{\log(2/\alpha)} \cdot \sum_{k=1}^{K} p_k \cdot \mathbb{E}\left[\frac{1}{\sqrt{n_k+1}}\right]
\end{aligned}$$

2. Since $n_k \sim \text{Bin}(n, p_k)$, using fact of Binoniaml distribution and Jensen's inequality:

$$\mathbb{E}\left[\frac{1}{\sqrt{n_k+1}}\right] \leq \sqrt{\mathbb{E}\left[\frac{1}{n_k+1}\right]} \leq \sqrt{\frac{1}{p_k(n+1)}} \leq \frac{1}{\sqrt{p_k n}}$$

3. Therefore,

$$\mathbb{E}[\text{Leb}(\mathcal{C}(X_{n+1}))] \leq 2(b-a)\sqrt{\log(2/\alpha)} \sum_{k=1}^{K} p_k \frac{1}{\sqrt{p_k n}} \leq 2(b-a)\sqrt{\log(2/\alpha)}\sqrt{\frac{K}{n}}$$

Setting:

- X has a **nonatomic** distribution (e.g., continuous).
- Fundamental limits exist even if $\mathcal{Y}$ is bounded.

Surprising Connection:

- distribution-free confidence interval for $\mu_P(X_{n+1}) \supseteq$ prediction interval for Y_{n+1}.

Theorem 11.3 (Regression implies Prediction)

Suppose $\mathcal{C}$ satisfies distribution-free coverage of the regression function:

$$\mathbb{P}(\mu_P(X_{n+1}) \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$$

Then, for any P where P_X is nonatomic:

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$$

- thm 11.4: $\mu_P(X_{n+1}) \to Med_P(X_{n+1})$

- **Sampling**
  1. Let $Z^{(i)} = (X^{(i)}, Y^{(i)})$. Draw $Z^{(1)}, \ldots, Z^{(M)} \overset{i.i.d}{\sim} P$.
     Define the empirical distribution $\hat{P}_M = \frac{1}{M} \sum_{i=1}^{M} \delta_{Z^{(i)}}$.
  2. Then, for non-atomic P_X:
     $$\mu_{\hat{P}_M}(X^{(i)}) = \text{Med}_{\hat{P}_M}(X^{(i)}) = Y^{(i)}$$
     So, almost surely:
     $$\mu_{\hat{P}_M}(X) = \text{Med}_{\hat{P}_M}(X) = Y$$
- **Resampling**
  1. Let $Z_i = (X_i, Y_i) \overset{i.i.d}{\sim} \hat{P}_M$. From the validity condition of $\mathcal{C}$ (distribution-free):
     $$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid \hat{P}_M) \geq 1 - \alpha$$
  2. After marginalizing over $\hat{P}_M$:
     $$\mathbb{P}_Q(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$$
     where Q is the joint distribution of $Z_1, \ldots, Z_{n+1}$ induced by the sample-resample process.
  3. Then, by Lemma 4.15:
     $$\mathbb{P}_{P^{n+1}}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha - d_{TV}(P^{n+1}, Q) \geq 1 - \alpha - \frac{n(n+1)}{2M}$$

Recall: Lemma 4.15

Let P be a distribution on $\mathcal{Z}$, and let $m, M \geq 1$. Let P^m denote the corresponding product distribution on $\mathcal{Z}^m$. Let Q denote the distribution on $\mathcal{Z}^m$ obtained by the following process:

1. Sample $Z^{(1)}, \ldots, Z^{(M)} \overset{i.i.d}{\sim} P$, and define $\hat{P}_M = \frac{1}{M} \sum_{i=1}^{M} \delta_{Z^{(i)}}$.
2. Sample $Z_1, \ldots, Z_m \overset{i.i.d}{\sim} \hat{P}_M$.

Then:

$$d_{TV}(P^m, Q) \leq \frac{m(m-1)}{2M}$$

Theorem 11.5

Let $\mathcal{Y} = [a, b]$. Suppose $\mathcal{C}$ satisfies distribution-free coverage of μ_P. Then, for any P with nonatomic P_X and $\mathrm{Var}(Y|X) \geq \sigma_*^2$:

$$\mathbb{E}[\mathrm{Leb}(\mathcal{C}(X_{n+1}))] \geq \frac{\sigma_*^2}{b-a} \cdot 2(1-\alpha)$$

Result: The lower bound does not depend on n. Vanishing width is **impossible**.
즉, X가 연속형일 때는 valid하면서 유한한 길이의 CI를 만들 수 없다.

1. First, we claim that for any $t \in [0,1]$,

$$\frac{1}{2}\mathbb{P}(\mu_P(X_{n+1}) - t\sigma_*^2 \in \mathcal{C}(X_{n+1})) + \frac{1}{2}\mathbb{P}(\mu_P(X_{n+1}) + t\sigma_*^2 \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha \tag{11.4}$$

2. Assuming (11.4) holds, we calculate the expected Lebesgue measure:

$$\begin{aligned}
\mathbb{E}[\text{Leb}(\mathcal{C}(X_{n+1}))] &= \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{I}\{y \in \mathcal{C}(X_{n+1})\}dy\right] \\
&\geq \mathbb{E}\left[\int_{\mu_P - \sigma_*^2}^{\mu_P + \sigma_*^2} \mathbb{I}\{y \in \mathcal{C}(X_{n+1})\}dy\right] \\
&= \sigma_*^2 \mathbb{E}\left[\int_{-1}^{1} \mathbb{I}\{\mu_P + t\sigma_*^2 \in \mathcal{C}(X_{n+1})\}dt\right] \\
&= \sigma_*^2 \int_{-1}^{1} \mathbb{P}(\mu_P + t\sigma_*^2 \in \mathcal{C}(X_{n+1}))dt \quad \text{(by Fubini-Tonelli)} \\
&= \sigma_*^2 \int_{0}^{1} \left[\mathbb{P}(\mu_P - t\sigma_*^2 \in \mathcal{C}(X_{n+1})) + \mathbb{P}(\mu_P + t\sigma_*^2 \in \mathcal{C}(X_{n+1}))\right] dt \\
&\geq \sigma_*^2 \cdot 2(1-\alpha)
\end{aligned}$$

4.2 Proof of Theorem 11.5 (11. 4)

1. Fix $t \in [0,1]$. Define a joint distribution $\tilde{P}$ on $\mathcal{X} \times \{\pm 1\} \times [0,1]$ as follows:
   1. Sample $(X, B) \sim P_X \times \mathrm{Unif}(\{\pm 1\})$.
   2. Sample $Y|(X,B) \sim \tilde{P}_{Y|(X,B)}$ where
$$\tilde{P}_{Y|(X,B)} = \left(\frac{1}{2} - \frac{B \cdot t\sigma_*^2}{\mu_P^1(X) - \mu_P^0(X)}\right) P_{Y|X}^0 + \left(\frac{1}{2} + \frac{B \cdot t\sigma_*^2}{\mu_P^1(X) - \mu_P^0(X)}\right) P_{Y|X}^1$$
2. Then, we have $\mathbb{E}_{\tilde{P}}[Y|X,B] = \mu_P(X) + t\sigma_*^2 B$, and the marginal distribution $\tilde{P}(X,Y) = P(X,Y)$.
3. Next, fix $M \geq 1$. Let $(X^{(1)}, B^{(1)}), \ldots, (X^{(M)}, B^{(M)}) \overset{i.i.d.}{\sim} P_X \times \mathrm{Unif}(\{\pm 1\})$.
   Define the empirical distribution:
$$\hat{P}_M = \frac{1}{M} \sum_{i=1}^{M} \delta_{(X^{(i)}, B^{(i)})}$$
   Then define $\tilde{P}_M$ on $(X,Y) \in \mathcal{X} \times [0,1]$ as:
   - $(X,B) \sim \hat{P}_M$ (with replacement)
   - $Y|(X,B) \sim \tilde{P}_{Y|X,B}$
4. We calculate the conditional mean under $\tilde{P}_M$:
$$\mu_{\tilde{P}_M}(X^{(i)}) = \frac{\sum_{j=1}^{M} (\mu_P(X^{(j)}) + t\sigma_*^2 B^{(j)}) \cdot \mathbb{I}\{X^{(j)} = X^{(i)}\}}{\sum_{j=1}^{M} \mathbb{I}\{X^{(j)} = X^{(i)}\}}$$
$$= \mu_P(X^{(i)}) + t\sigma_*^2 B^{(i)} \quad (\because X \text{ is non-atomic, so distinct a.s.})$$

6. Since $\mathcal{C}$ satisfies distribution-free validity:

$$\mathbb{P}(\mu_{\hat{P}_M}(X_{n+1}) \in \mathcal{C}(X_{n+1}) \mid \hat{P}_M) \geq 1 - \alpha$$

7. After marginalizing over $\hat{P}_M$:

$$\mathbb{P}(\mu_P(X_{n+1}) + t\sigma_*^2 B_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$$

8. Now consider $\tilde{P}$:
   - $(X, B) \sim \hat{P}_M$ **without replacement**
   - $Y|(X, B) \sim \tilde{P}_{Y|X,B}$
   - Then, by Lemma 4.15 (Sample-Resample):

$$\mathbb{P}_{\tilde{P}^{n+1}}(\mu_P(X_{n+1}) + t\sigma_*^2 B_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha - \frac{n(n+1)}{2M}$$

   Taking $M \to \infty$, we get:

$$\mathbb{P}_{\tilde{P}^{n+1}}(\mu_P(X_{n+1}) + t\sigma_*^2 B_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$$

9. Since this event does not depend on $B_1, \ldots, B_n$ or on Y_{n+1}, we can write:

$$\mathbb{P}_{P^n \times P_X \times \mathrm{Unif}(\{\pm 1\})}(\mu_P(X_{n+1}) + t\sigma_*^2 B_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$$

10. Since $B_{n+1} \sim \mathrm{Unif}(\{\pm 1\})$, this proves Eq (11.4).

- thm 11.5에서 만든 $\tilde{P}$의 유효성 증명에 쓸 lemma

lemma 11.6

Let $Y \in [0,1]$ be a random variable with distribution P, and let Med_P and σ_P^2 denote its median and its variance. Consider the unique decomposition of P into a mixture

$$P = \frac{1}{2}P_0 + \frac{1}{2}P_1$$

such that P_0 is supported on $[0, \mathrm{Med}_P]$ and P_1 is supported on $[\mathrm{Med}_P, 1]$. Then

$$\mathbb{E}_{P_1}[Y] - \mathbb{E}_{P_0}[Y] \geq 2\sigma_P^2.$$

4.2 Proof of lemma 11.6

1. Construct P_0, P_1 using the generalized inverse CDF F^{-1}.
   - P_0 is the distribution of $F^{-1}(U)$ for $U \sim \text{Unif}[0, 0.5]$.
   - P_1 is the distribution of $F^{-1}(U)$ for $U \sim \text{Unif}[0.5, 1]$.

   Define $\mu_0 = \mathbb{E}_{P_0}[Y]$, $\mu_1 = \mathbb{E}_{P_1}[Y]$, and $\mu_P = \mathbb{E}_P[Y] = \frac{\mu_0+\mu_1}{2}$.
2. Then, with simple calculation:
$$\sigma_P^2 = \frac{1}{2} Var_{P_0}(Y) + \frac{1}{2} Var_{P_1}(Y) + \frac{1}{4}(\mu_1 - \mu_0)^2$$
3. Since P_0 is supported on $[0, \text{Med}_P]$:
$$Var_{P_0}(Y) \le \mu_0(\text{Med}_P - \mu_0) = \mu_0(\text{Med}_P - \mu_P) + \frac{1}{2}\mu_0(\mu_1 - \mu_0)$$
   Similarly, since P_1 is supported on $[\text{Med}_P, 1]$:
$$Var_{P_1}(Y) \le (1 - \mu_1)(\mu_1 - \text{Med}_P) = (1 - \mu_1)(\mu_P - \text{Med}_P) + \frac{1}{2}(1 - \mu_1)(\mu_1 - \mu_0)$$
4. Therefore,
$$\sigma_P^2 \le \frac{1}{2}|\text{Med}_P - \mu_P| + \frac{1}{4}(\mu_1 - \mu_0)$$
5. Since $\mu_0 \le \text{Med}_P \le \mu_1$, we have $|\text{Med}_P - \mu_P| \le \max(|\mu_1 - \mu_P|, |\mu_0 - \mu_P|) = \frac{1}{2}(\mu_1 - \mu_0)$. Thus, $\sigma_P^2 \le \frac{1}{2}(\mu_1 - \mu_0)$.

5. Relaxing the Target for the continuous case

Since vanishing width is impossible for $\mu_P(X)$ in the nonatomic case, we consider relaxations.

1. **Binning:** Discrete approximation of X.
2. **Blurring:** Smoothing the target function.

이 방법을 사용하면 X가 연속형일 때도 valid하면서 유한한 길이의 CI를 만들 수 있다.

Binning

- Partition $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$.
- Target: $\mu_P(\mathcal{X}_k) = \mathbb{E}[Y|X \in \mathcal{X}_k]$.

Blurring

- Target:

$$\tilde{\mu}_P(x) = \frac{\mathbb{E}_P[\mu_P(X)H(x,X)]}{\mathbb{E}_P[H(x,X)]}$$

calibration

- Target: How close the $\mathbb{E}_P[Y|\hat{\mu}(X)$ is to $\hat{\mu}(X)$

theorem 11.7

Let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} \subseteq [a, b]$, and let $\alpha \in [0, 1]$. Let $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$ be a fixed partition. Let $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P$.
For each $k \in [K]$, define $n_k = \sum_{i=1}^{n} \mathbb{I}\{X_i \in \mathcal{X}_k\}$, and let

$$\hat{\mu}(\mathcal{X}_k) = \frac{1}{n_k} \sum_{i=1}^{n} Y_i \cdot \mathbb{I}\{X_i \in \mathcal{X}_k\}$$

for each k with $n_k \geq 1$. Define

$$\mathcal{C}(\mathcal{X}_k) = \begin{cases} \hat{\mu}(\mathcal{X}_k) \pm (b-a)\sqrt{\frac{\log(2/\alpha)}{2n_k}}, & \text{if } n_k \geq 1 \\ [a, b], & \text{if } n_k = 0. \end{cases}$$

Then $\mathcal{C}$ satisfies the binned coverage validity:

$$\mathbb{P}(\mu_P(\mathcal{X}_{k(X_{n+1})}) \in \mathcal{C}(\mathcal{X}_{k(X_{n+1})})) \geq 1 - \alpha.$$

- Proof: Let $(k(X), Y) \sim \tilde{P}$ when $(X, Y) \sim P$. Then, by construction:

  $$\mu_{\tilde{P}}(\mathcal{X}_k) = \mu_{\tilde{P}}(k)$$

  So the goal of (11.6) is equivalent to the goal of the discrete case (Theorem 11.2).

Theorem 11.8

Let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} \subseteq [a, b]$, and let $\alpha \in [0, 1]$. Let $H : \mathcal{X} \times \mathcal{X} \to [0, B]$ be a function satisfying $\mathbb{E}_P[H(x, X)] > 0$ for all $x \in \mathcal{X}$.

Let $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P$, and let $U_1, \ldots, U_n \overset{i.i.d.}{\sim} \text{Unif}[0, 1]$ be drawn independently of the data. For each $x \in \mathcal{X}$, define

$$n(x) = \sum_{i=1}^{n} \mathbb{I}\{U_i \leq \frac{H(x, X_i)}{B}\}.$$

Let

$$\hat{\mu}(x) = \frac{1}{n(x)} \sum_{i=1}^{n} Y_i \cdot \mathbb{I}\{U_i \leq \frac{H(x, X_i)}{B}\}$$

if $n(x) \geq 1$. Define

$$\mathcal{C}(x) = \begin{cases} \hat{\mu}(x) \pm (b - a)\sqrt{\frac{\log(2/\alpha)}{2n(x)}}, & \text{if } n(x) \geq 1 \\ [a, b], & \text{if } n(x) = 0. \end{cases}$$

Then $\mathcal{C}$ satisfies $\mathbb{P}(\tilde{\mu}_P(x) \in \mathcal{C}(x)) \geq 1 - \alpha$ for every $x \in \mathcal{X}$.

1. Define $Q = Q_X \times P_{Y|X}$ where $\frac{dQ_X}{dP_X}(x', y) \propto H(x, x')$ (Radon-Nikodym derivative). Then,

$$\begin{aligned}\mu_{Q_Y} &= \mathbb{E}_Q[Y] = \mathbb{E}_P\left[Y \cdot \frac{dQ}{dP}(X, Y)\right] \\ &= \mathbb{E}_P\left[Y \cdot \frac{H(x, X)}{\mathbb{E}_P[H(x, X)]}\right] \\ &= \mathbb{E}_P\left[\mu_P(X) \cdot \frac{H(x, X)}{\mathbb{E}_P[H(x, X)]}\right] \\ &= \tilde{\mu}_P(x)\end{aligned}$$

2. **Next,** let $I(x) = \{i \in [n] : U_i \leq \frac{H(x, X_i)}{B}\}$ and note $n(x) = |I(x)|$.
   Then, conditional on $n(x)$, $\{(X_i, Y_i)\}_{i \in I(x)}$ are i.i.d. draws from Q.
   i.e., $\{Y_i\}_{i \in I(x)}$ are i.i.d. draws from Q_Y.
3. Therefore, by Hoeffding's inequality,

$$\mathbb{P}(\mu_P(x) \in C(x)) \geq 1 - \alpha$$

- Q. Intuitively, if $x \to \mu_p(x)$ is reasonably smooth, then we should expect $\hat{\mu}_p(x) \simeq \mu_p(x)$, as long as the kernel H is reasonably strongly localized—for instance, a Gaussian kernel with a small bandwidth h ¿ 0. 에 대해 설명해주세요.
- A. $\mu_p(x)$가 smooth하다는 것은, $x \simeq x'$라면 $\mu_p(x) \simeq \mu_p(x')$라는 것이고, kernel H가 strongly localized 하다는건, 가까운 이웃에 대해서만 H값이 크다는 것입니다. 따라서 x와 아주 유사한 x'들의 $\mu(x')$를 평균 내어서 $\hat{\mu}_p(x)$를 구하게 되므로, 이는 $\mu_p(x')$와 비슷하게 될 것입니다.

5.2 Relaxation by blurring the target: Q3

- Q. 가우시안 커널을 결정하는 방법에는 fine tuning밖에 없나요?
- A. thm 11.8의 출처 논문 [Tight distribution-free confidence intervals for local quantile regression(Jayoon Jang and Emmanuel Candès, 2023)]에 따르면, 원하는 해상도의 bandwidth를 고르되 유효표본크기가 10∼20 이상이 되도록 고르는 것을 권장합니다.
  - **Kernel Weight (L_i):** X_i와 타겟 x_0가 얼마나 가까운지를 나타냄.

$$L_i = K\left(\frac{x_0 - X_i}{h}\right)$$

  - **Effective Sample Size (n_{eff}):** 얼마나 많은 이웃 정보를 이용해서 추론하는가.

$$n_{eff} = \frac{(\sum_{i=1}^{n} L_i)^2}{\sum_{i=1}^{n} L_i^2}$$

2. Trade-off

- **h가 작으면:** High Resolution (국소적인 특징) ↔ 작은 n_{eff} (unstable/wide intervals).
- **h가 크면:** Low Resolution (전체적인 특징) ↔ 큰 n_{eff} (stable/narrow intervals).

3. Practical Guideline

- Choose h based on the **desired resolution** (how local you want to be).
- **Constraint:** Ensure that the effective sample size is sufficiently large (e.g., $n_{eff} \geq 10 \sim 20$) to guarantee validity.

5.2 Relaxation by blurring the target: Q4

- Q. 각 relaxation 방법의 장단점, 선택 기준이 궁금합니다.
- A. Binning은 간단하고 Blurring은 Binning에 비해서는 kernel도 선택해야하는 등 결정해야하는 hyperparameter가 많습니다. 만약 이를 결정할만한 충분한 정보가 있다면 Blurring을 통해 섬세한 구간 추정을 하는 것이 좋을 것 같습니다. 하지만 그렇지 않고, 특히 나이와 같이 구간 나누기가 수월한 변수라면 간편하게 Binning을 사용하는 것이 좋을 듯 합니다.

Recap:

- Chapter 4: Test-conditional predictive coverage is hard.
- Chapter 11: Regression inference is hard.

The Link:

- Define $\tilde{Y} = \mathbb{I}\{Y \notin \mathcal{C}_{init}(X)\}$.
- $\mathbb{P}(Y \in \mathcal{C}_{init}(X)|X) \geq 1 - \alpha \iff \mathbb{E}[\tilde{Y}|X] \leq \alpha$.
- Regression on $\tilde{Y}$ (estimating error probability) allows controlling conditional coverage.

6.1 From regression to test-conditional prediction

Strategy:

- Given a regression oracle $\mathcal{C}_{regr}$ for the bounded variable $\tilde{Y} \in \{0, 1\}$.
- Define prediction set $\mathcal{C}_{pred}(x)$:

$$\mathcal{C}_{pred}(x) = \begin{cases} \mathcal{C}_{init}(x), & \text{if } \mathcal{C}_{regr}(x) \subseteq [0, \alpha - \epsilon/\delta] \\ [a, b], & \text{otherwise} \end{cases}$$

- C_{regr} is CI of "error of C_{init}". If error is so big, then use conservative interval.

Proposition 11.9

$\mathcal{C}_{pred}$ satisfies relaxed test-conditional coverage:

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{pred}(X_{n+1}) | X_{n+1} \in \mathcal{X}_0) \geq 1 - \alpha$$

for any set $\mathcal{X}_0$ with $P_X(\mathcal{X}_0) > \delta$.

- 회귀구간으로 (완화된) test conditional 예측 구간을 생성할 수 있다.
- 근데 Ch4에서 유한한 길이의 (완화된) test conditional 예측 구간 생성이 어려움을 이미 확인함.(thm 4.14) 따라서 유한한 길이의 회귀구간을 만드는 것이 쉽지 않음을 간접적으로 다시 확인 가능하다.

1. By definition, the failure of the prediction set is decomposed as:

$$Y_{n+1} \notin \mathcal{C}_{pred}(X_{n+1}) \iff Y_{n+1} \notin \mathcal{C}_{init}(X_{n+1}) \quad \text{and} \quad \mathcal{C}_{regr}(X_{n+1}) \subseteq [0, \alpha - \epsilon/\delta]$$

2. Therefore, writing $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$,

$$\mathbb{P}(Y_{n+1} \notin \mathcal{C}_{pred}(X_{n+1}) \mid \mathcal{D}_n, X_{n+1} \in X_0) \leq (\alpha - \epsilon/\delta) + \delta^{-1} \cdot \mathbb{P}(\mu_{\tilde{P}}(X_{n+1}) \notin \mathcal{C}_{regr}(X_{n+1}) \mid \mathcal{D}_n)$$

3. Marginalizing over $\mathcal{D}_n$:

$$\mathbb{E}[\mathbb{P}(\mu_{\tilde{P}}(X_{n+1}) \notin \mathcal{C}_{regr}(X_{n+1}) \mid \mathcal{D}_n)] = \mathbb{P}(\mu_{\tilde{P}}(X_{n+1}) \notin \mathcal{C}_{regr}(X_{n+1})) \leq \epsilon$$

4. Therefore,

$$\mathbb{P}(Y_{n+1} \notin \mathcal{C}_{pred}(X_{n+1}) \mid X_{n+1} \in X_0) \leq (\alpha - \epsilon/\delta) + \delta^{-1} \cdot \epsilon = \alpha$$

Alternative View:

- New Target: Estimate $\hat{\mu}(x)$ and quantify uncertainty
- Validity Condition (11.9):

$$\mathbb{P}(\|\hat{\mu} - \mu_P\|_{L_1(P)} \leq \hat{\epsilon}) \geq 1 - \delta$$

  where $\hat{\epsilon} > 0$ is a function of the training data.
- **Same statistical difficulty** with achieving distribution free CI for regression
  - Choose $\hat{\mu}, \hat{\epsilon} \Leftrightarrow$ Choose $\mathcal{C}$
  - interval length **vanish** if discrete, not vanish if continuous
- 점추정 관점에서의 validity 조건과 앞서 나왔던 validity 조건이 동치이며(증명X), 점추정 관점에서도 X가 이산형일 때는 구간이 vanish, 연속형일 때는 그렇지 않음을 확인할 수 있다.

Theorem 11.10

Let P be a distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{x_1, \ldots, x_K\}$ and $\mathcal{Y} \subseteq [a, b]$, and let $\delta \in [0, 1]$. Let $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P$. Define $n_k = \sum_{i=1}^{n} \mathbb{I}\{X_i = x_k\}$. Let

$$\hat{\mu}(x_k) = \begin{cases} \frac{1}{n_k} \sum_{i=1}^{n} Y_i \cdot \mathbb{I}\{X_i = x_k\}, & \text{if } n_k \geq 1 \\ \frac{a+b}{2}, & \text{if } n_k = 0 \end{cases}$$

and define

$$\hat{\epsilon} = \frac{b-a}{\sqrt{2\delta}} \cdot \sqrt{\frac{K}{n}}.$$

Then $\hat{\mu}, \hat{\epsilon}$ satisfy the distribution-free validity condition:

$$\mathbb{P}(\|\hat{\mu} - \mu_P\|_{L_1(P)} \leq \hat{\epsilon}) \geq 1 - \delta.$$

7. Proof of Theorem 11.10 (Discrete Case)

1. Let $p_k = \mathbb{P}_P(X = X_k)$. Consider the $L_1(P)$ norm difference:

$$\|\hat{\mu} - \mu_P\|_{L_1(P)} = \mathbb{E}_P[|\hat{\mu} - \mu_P|] \leq \sqrt{\sum_{k=1}^{K} p_k(\hat{\mu}(x_k) - \mu_P(x_k))^2}$$

2. By the fact of variance (for bounded variable in $[a, b]$):

$$\mathbb{E}[(\hat{\mu}(x_k) - \mu_P(x_k))^2 \mid n_k] \leq \frac{(b-a)^2}{4n_k} \leq \frac{(b-a)^2}{2(n_k+1)}$$

3. So, taking expectation over n_k:

$$\mathbb{E}[(\hat{\mu}(x_k) - \mu_P(x_k))^2] \leq \mathbb{E}_{n_k}\left[\frac{(b-a)^2}{2(n_k+1)}\right]$$

4. Thus,

$$\begin{aligned}
\mathbb{E}[\|\hat{\mu} - \mu_P\|_{L_1(P)}^2] &\leq \mathbb{E}\left[\sum_{k=1}^{K} p_k \cdot \frac{(b-a)^2}{2(n_k+1)}\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} p_k \cdot \frac{(b-a)^2}{2 \cdot np_k}\right] \quad (\because n_k \sim \text{Binomial}) \\
&= \frac{(b-a)^2}{2} \cdot \frac{K}{n}
\end{aligned}$$

6. Therefore, by Markov's inequality:

$$\mathbb{P}\left(\|\hat{\mu}-\mu_P\|_{L_1(P)} > \frac{b-a}{\sqrt{2\delta}}\sqrt{\frac{K}{n}}\right) = \mathbb{P}\left(\|\hat{\mu}-\mu_P\|^2_{L_1(P)} > \frac{(b-a)^2}{2\delta}\frac{K}{n}\right) \leq \delta$$

Theorem 11.11

Suppose $\hat{\mu}$ and $\hat{\epsilon}$ satisfy the validity condition $\mathbb{P}(\|\hat{\mu} - \mu_P\|_{L_1(P)} \leq \hat{\epsilon}) \geq 1 - \delta$.
Then, for any distribution P on $\mathcal{X} \times [a, b]$ for which the marginal P_X is nonatomic, and for any sample size $n > 1$,

$$\mathbb{P}\left(\hat{\epsilon} \geq \frac{\mathbb{E}_P[\mathrm{Var}_P(Y|X)]}{b-a}\right) \geq 1 - \delta.$$

7. Proof of Theorem 11.11 (Continuous Case)

Note: Sample-resample 기법 사용. Theorem 11.3 증명과 유사.

1. W.L.O.G., let $[a, b] = [0, 1]$.
   Let $(X^{(1)}, Y^{(1)}), \ldots, (X^{(M)}, Y^{(M)})$ be the data, and let $\hat{P}_M$ be the empirical distribution.
2. Conditional on $\hat{P}_M$, let $((X_i, Y_i))_{i \in [n]}$ be i.i.d. draws from $\hat{P}_M$, and let $\hat{\mu}$ be trained on this data.
   Then, by validity condition:

$$1 - \delta \leq \mathbb{P}(\|\hat{\mu} - \mu_{\hat{P}_M}\|_{L_1(\hat{P}_M)} \leq \hat{\epsilon} \mid \hat{P}_M) \leq \mathbb{P}(\|\hat{\mu} - \mu_{\hat{P}_M}\|_{L_1(\hat{P}_M)} \leq c_M \mid \hat{P}_M) + \mathbb{P}(\hat{\epsilon} > c_M \mid \hat{P}_M)$$

   where

$$c_M = \mathbb{E}_P[\mathrm{Var}_P(Y|X)] - \frac{n}{4M} - \frac{1}{2\sqrt{M}}$$

   And by definition of $\hat{P}_M$:

$$\|\hat{\mu} - \mu_{\hat{P}_M}\|_{L_1(\hat{P}_M)} = \frac{1}{M} \sum_{i=1}^{M} |\hat{\mu}(X^{(i)}) - Y^{(i)}|$$

3. Therefore,

$$\mathbb{P}_{\hat{P}_M}(\hat{\epsilon} > c_M \mid \hat{P}_M) \geq 1 - \delta - \mathbb{P}_{\hat{P}_M}\left(\frac{1}{M} \sum_{i=1}^{M} |\hat{\mu}(X^{(i)}) - Y^{(i)}| \leq c_M \middle| \hat{P}_M\right)$$

7. Proof of Theorem 11.11 (Continuous Case)

5. Next, by Lemma 4.15:

$$\begin{aligned}\mathbb{P}_P(\hat{\epsilon} > c_M) &\geq \mathbb{E}[\mathbb{P}_{\hat{P}_M}(\hat{\epsilon} > c_M \mid \hat{P}_M)] - \frac{n(n-1)}{2M} \\ &\geq 1 - \delta - \frac{n(n-1)}{2M} - \mathbb{E}\left[\mathbb{P}_{\hat{P}_M}\left(\frac{1}{M}\sum_{i=1}^{M}|\hat{\mu}(X^{(i)}) - Y^{(i)}| \leq c_M \middle| \hat{P}_M\right)\right]\end{aligned}$$

6. And by Chebyshev's inequality:

$$\mathbb{E}\left[\mathbb{P}_{\hat{P}_M}\left(\cdots \leq c_M \mid \hat{P}_M\right)\right] = \mathbb{P}\left(\frac{1}{M}\sum_{i=1}^{M}|\hat{\mu}(X^{(i)}) - Y^{(i)}| \leq c_M\right) \leq \frac{1}{2\sqrt{M}}$$

7. Combining them:

$$\mathbb{P}_P(\hat{\epsilon} > c_M) \geq 1 - \delta - \frac{n(n-1)}{2M} - \frac{1}{2\sqrt{M}}$$

8. Since $\lim_{M\to\infty} c_M = \mathbb{E}_P[\mathrm{Var}_P(Y|X)]$, taking $M \to \infty$ completes the proof.

Q5.

- Q. 회귀 구간이 예측 구간보다 크다는 것을 직관적으로 설명해주세요.
- A. 일반적으로 회귀가 예측보다 쉬운 이유는, 회귀가 smoothness를 전제로 하고 있기 때문입니다. 하지만 dist free 하에서는 그렇지 않으므로 회귀가 더 쉽다고 볼 수 없습니다.
- A. 예시) 모든 값이 0이고 하나의 값만 극단적으로 큰 값을 갖는 분포가 있을 때, 예측 구간을 만든다면 [0]으로 만들어도 커버리지를 만족할 수 있겠지만 회귀 구간을 만든다면 극단적인 값도 커버하기 위해 구간을 넓게 잡을 것...
- A. 예시) 동전 던지기-예측 구간은 0, 1로 두면 되지만, 회귀 구간은 이보다 긴 구간이 될 것...