Statistical Inference Summary

Transcribed by Nicholas Lyu *

April 30, 2024

Contents

1 Description and Estimation **3**

1.1 Preliminaries . . . 3

1.2 Basic statistical definitions . . . 4

2 MoM, Sample Quantile, and Asymptotics **6**

2.1 Models and Esti-(everything), Method of Moments . . . 6

2.2 Bias-variance Decomposition . . . 7

2.3 Consistency and Asymptotics . . . 7

2.4 Sample quantile Asymptotics . . . 8

3 Maximum Likelihood Estimation **10**

3.1 KL-divergence and MLE consistency . . . 10

3.2 Score function, Fisher information, and Cramér-Rao . . . 11

3.3 Asymptotics of MLE . . . 13

4 Confidence Intervals **15**

5 Regression **17**

5.1 Predictive Regression . . . 17

5.2 Descriptive regression . . . 19

6 Exponential Families and Sufficiency **21**

6.1 Exponential Families . . . 21

6.2 Sufficient Statistic . . . 22

7 Hypothesis Testing **23**

7.1 Definitions . . . 23

7.2 Likelihood-based testing . . . 24

7.3 p-value . . . 25

8 Bayesian Inference **26**

8.1 Framework and Point Estimates . . . 26

8.2 Credible Intervals, Conjugate priors . . . 27

8.3 Model choice, Marginal Likelihood, and Hypothesis Testing . . . 28

[*] This document is a summary of the main proofs and examples in Joe Blitzstein and Neil Shephard's *Stat 111: Introduction to Statistical Inference*, as of the Spring 2024 iteration of Stat 111. A few results in the course homeworks are also included.

8.4 Bayesian Prediction . . . 29
8.5 Hierarchical models and Stein's Paradox . . . 29

9 Design-based Inference 31
9.1 Sampling design . . . 31
9.2 Stratified Sampling, Horvitz-Thompson estimator . . . 32
9.3 Bootstrapping and Permutation tests . . . 33

10 Causal Inference 34
10.1 Treatment framework . . . 34
10.2 Population-based inference . . . 34
10.3 Finite sample inference . . . 35
10.4 Observational Studies and Unconfoundedness . . . 36

1 Description and Estimation

1.1 Preliminaries

Theorem 1.1 (*Markov's inequality*). for $X > 0$, $\Pr(X \geq a) \leq \mathbb{E}[X]/a$

Definition 1.1 (*Eve's law*). The law of total variance

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])$$

Intuitively, let X be a stratification of Y, then the total variance is the sum of expected variance within each class (the first term) and the variance of the classes (second term).

Proof: Apply Adam's law and regroup terms. First note that $\mathbb{E}[Y]^2 = \mathbb{E}[\mathbb{E}[Y|X]]^2$.

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \mathbb{E}[\mathbb{E}[Y^2|X]] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \\ &= \mathbb{E}[\mathrm{Var}(Y|X) + \mathbb{E}[Y|X]^2] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \\ &= \mathbb{E}[\mathrm{Var}(Y|X)] + \mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[\mathbb{E}[Y|X]]^2 = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])\end{aligned}$$

Definition 1.2 (*convergence in distribution, probability*). Given a sequence of random variables (X_j) and X, convergence in distribution is written

$$X_j \xrightarrow{d} X \iff \lim_{n\to\infty} F_{X_n}(x) = F_X(x) \text{ at all continuity points } x \text{ of } F_X$$

Convergence in probability is written

$$X_j \xrightarrow{p} X \iff \forall \epsilon > 0, \lim_{n\to\infty} \Pr(|X_n - X| \geq \epsilon) = 0$$

Almost sure convergence (strong convergence) is written

$$X_j \xrightarrow{a.s.} X \iff \Pr\left(\lim_{n\to\infty} X_n = X\right) = 1$$

Proposition 1.2. Convergence in probability implies in convergence in distribution. The converse is true when X is a constant.

Intuition: Gaussians with converging parameters converge in distribution but not probability.

Definition 1.3 (*strong law of large numbers (SLLN)*). The sample average converges almost surely to the population average: $\bar{X}_j \xrightarrow{a.s.} \mu$.

$$\Pr\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1$$

Proposition 1.3. Let μ be a constant, then $X_n \xrightarrow{a.s.} \mu$ implies that for all $\epsilon > 0$

$$\forall \epsilon > 0, \exists n_0 : |X_{\forall n > n_0} - \mu| < \epsilon$$

Proof: Unrolling the definition of the limit, $X_n \xrightarrow{a.s.} \mu$ implies

$$\Pr(\forall \epsilon > 0, \exists n_0 : |X_{\forall n > n_0} - \mu| < \epsilon) = 1$$

Taking the quantifiers out of Pr yields our claim

$$\forall \epsilon > 0, \exists n_0 : \Pr(|X_{\forall n > n_0} - \mu| < \epsilon) = 1$$

Applying to SLLN, this implies that $\forall \epsilon > 0, \exists n_0$ such that the sample mean of $> n_0$ samples is guaranteed to be within ϵ of the true mean.

Definition 1.4 (*central limit theorem*). The normalized sum of a sequence of zero-mean random variables converge to the Gaussian in distribution.

$$\lim_{n\to\infty} \frac{X_1 + \cdots X_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Theorem 1.4 (*continuous mapping theorem (CMT)*). Continuous functions preserve convergence in probability and distribution

$$X_n \xrightarrow{(p,d)} X \implies g(X_n) \xrightarrow{(p,d)} g(X)$$

Theorem 1.5 (*Slutsky's theorem*). Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then for continuous g

$$g(X_n, Y_n) \xrightarrow{d} g(X, c)$$

Note that (Y_n) must converge to a constant for this to be true. As a counterexample, for $X_n = Y_n = \mathcal{N}(0,1)$, $X_n + Y_n \xrightarrow{d} \mathcal{N}(0, \sqrt{2})$ instead of $\mathcal{N}(0,2)$.

The continuous mapping theorem guarantees consistency but does not give us any information about how the asymptotic distribution of the difference transforms under continuous maps. The delta method gives this for normal asymptotic differences.

Theorem 1.6 (*transforms of asymptotic normals (delta method)*). Given differentiable g, estimand θ and estimator $\hat{\theta}$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2) \implies \sqrt{n}\left[g(\hat{\theta}) - g(\theta)\right] \xrightarrow{d} \mathcal{N}(0, g'(\theta)^2\omega^2)$$

More conveniently, write $\hat{\theta}$ on the left hand side. For n large

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{\omega^2}{n}\right) \implies g(\hat{\theta}) \sim \mathcal{N}\left(g(\theta), g'(\theta)^2 \frac{\omega^2}{n}\right)$$

Proof: For n large, Taylor expand $g(\hat{\theta})$ about θ yielding

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$$

Substitute into the left hand side

$$\sqrt{n}\left[g(\hat{\theta}) - g(\theta)\right] = g'(\theta)\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, g'(\theta)^2\omega^2)$$

Remark 1.1. The delta method is a highly nontrivial result: the general transformation of a normal is *not* normal. We take advantage of $n \to \infty \implies |\hat{\theta} - \theta| \ll 1$, in which region g looks linear. The linear transformation of a normal is normal.

1.2 Basic statistical definitions

Definition 1.5 (*order statistics*). The order statistics of $y_1 \cdots y_n$ are the same data in increasing order

$$y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$$

Definition 1.6 (*quantile function*). Given a CDF F, its quantile function $Q : [0,1] \to \mathbb{R}$ is

$$Q(p) = F^{-1}(p) = \inf\{y \mid F(y) \ge p\}$$

Note that Q is monotonically non-decreasing, continuous, and $F(Q(p)) \ge p$.

Definition 1.7 (*sample quantile*). The p-sample quantile of $y_1 \cdots y_n$ is the order statistic

$$\hat{Q}(p) = y_{\lceil np \rceil}$$

Definition 1.8 (*empirical cdf*). The empirical CDF is $\hat{F}(y) = \frac{1}{n}\sum_{j=1}^{n} I(y_j \leq y)$. The ECDF converges to the underlying CDF: by SLLN, with probability 1

$$\lim_{n\to\infty} \hat{F}(y) = \mathbb{E}[I(Y \leq y)] = P(Y \leq y) = F(y)$$

Definition 1.9 (*treatment effect*). Let X denote assignment and Y outcome. For *the same* sample, let $Y(X = 1)$ denote the treated outcome and $Y(X = 0)$ be the outcome under control, then the random variable $\tau = Y(1) - Y(0)$ is the treatment effect. The population average's treatment effect is

$$\mathbb{E}[\tau] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

Theorem 1.7 (*causal power of randomized control trials*).

$$\mathbb{E}[\tau] = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0] \iff \text{Cov}(Y(1), X) = \text{Cov}(Y(0), X) = 0$$

Proof: $\mathbb{E}[YX] = \mathbb{E}[YX|X = 1]P(X = 1) = \mathbb{E}[Y|X = 1]P(X = 1)$, so

$$\mathbb{E}[Y|X = 1] = \frac{\mathbb{E}[YX]}{P(X = 1)} = \frac{\mathbb{E}[Y(1)]\mathbb{E}[X] + \text{Cov}(Y(1), X)}{\mathbb{E}[X]}$$

Similarly, $\mathbb{E}[Y(1 - X)] = \mathbb{E}[Y|X = 0]P(X = 0)$ and $Y(1 - X) = Y(0)(1 - X)$

$$\mathbb{E}[Y|X = 0] = \frac{\mathbb{E}[Y(1 - X)]}{P(X = 0)} = \frac{\mathbb{E}[Y(0)(1 - X)]}{1 - \mathbb{E}[X]} = \mathbb{E}[Y(0)] + \frac{\text{Cov}(Y(0), 1 - X)}{1 - \mathbb{E}[X]}$$

2 MoM, Sample Quantile, and Asymptotics

2.1 Models and Esti-(everything), Method of Moments

Definition 2.1 (*statistical model*). Given observed data $=(y_1, \cdots, y_n)$ realized from $\mathbf{Y} = (Y_1, \cdots, Y_n)$, a statistical model is set of considered candidates (models) for $\mathbf{Y}$ or $F_{\mathbf{Y}}$. A model is parametric if it can be indexed by a finite-dimensional θ.

Definition 2.2 (*estimand*). An estimand is an aspect of $F_{\mathbf{Y}}$ (e.g. p-quantile $Q_{Y_1}(p), F_{Y_1}(y)$).

Definition 2.3 (*prediction decomposition*). The joint probability density may be factorized

$$f_{\mathbf{Y}}(\mathbf{y};\theta) = f_{Y_1}(y_1;\theta)\prod_{j=2}^{n} f_{Y_j}(y_j|Y_1 = y_1, \cdots, Y_{j-1} = y_{j-1};\theta)$$

Definition 2.4 (*likelihood function*). Observing $\mathbf{y}$ the likelihood function $L(\theta;\mathbf{y})$ is

$$\theta \mapsto f_{\mathbf{Y}}(\mathbf{y};\theta)$$

Proposition 2.1. The likelihood is invariant under reparameterization $\theta \mapsto \psi$ and data transformation $\mathbf{Y} \mapsto F(\mathbf{Y})$.

$$\mathbf{Y}_{\theta=\theta_0} = \mathbf{Y}_{\psi=\psi(\theta_0)} \implies L_{\mathbf{Y}}(\theta_0, \mathbf{y}) = L_{\mathbf{Y}}(\psi(\theta_0), \mathbf{y})$$
$$L_{\mathbf{Y}}(\theta, \mathbf{y}) = L_{F(\mathbf{Y})}(\theta, F(\mathbf{y}))$$

Proof: The first claim follows from $f_{\mathbf{Y}_{\theta=\sim}}(\mathbf{y};\theta') = f_{\mathbf{Y}_{\psi=\sim}}(\mathbf{y};\psi(\theta'))$. For the second claim

$$L_{F(\mathbf{Y})}(\theta, F(\mathbf{y})) = f_{F(\mathbf{Y})}(F(\mathbf{y});\theta) = \frac{1}{F'(\mathbf{y})}F_{\mathbf{Y}}(\mathbf{y};\theta)$$

the multiplicative constant is not dependent upon θ.

Definition 2.5 (*statistic, estimator*). A statistic is a random variable $T(\mathbf{Y})$ which does not explicitly depend on unknown parameters. An estimator $\hat{\theta}$ is a statistic charged with approximating an estimand θ (a parameter of the model).

Definition 2.6 (*bias, variance*). The bias of an estimator is $\text{bias}(\hat{\theta}) = \mathbb{E}_{\mathbf{y}\sim\mathbf{Y}_\theta}[\hat{\theta}] - \theta$ (function of θ). Its variance is $\text{Var}_{\mathbf{y}\sim\mathbf{Y}_\theta}(\hat{\theta})$.

Definition 2.7 (*estimate*). An estimate is a crystallization of $\hat{\theta}$ upon observing $\mathbf{y}$.

Definition 2.8 (*method of moments*). Find α, h such that $\alpha(\theta) = \mathbb{E}[h(\mathbf{Y})]$, then

$$\hat{\theta}_{\text{MoM}} = \alpha^{-1}\left[\frac{1}{n}\sum_{j=1}^{n} h(\mathbf{y}_j)\right]$$

Distinct (α, h) satisfying our conditions may yield different MoM estimators.

Example 2.1 (*regression*). Consider i.i.d $(\mathbf{X}, \mathbf{Y})$ with the estimand $\theta = \left(\frac{\mathbb{E}[X_1Y_1]}{\mathbb{E}[X_1^2]}, \mathbb{E}[X_1^2]\right)$

$$\hat{\theta}_{\text{MoM}} = \left(\frac{(X_j)\cdot(Y_j)}{(X_j)\cdot(X_j)}, \frac{1}{n}(X_j^2)\right)$$

2.2 Bias-variance Decomposition

The example of KDE demonstrates how variance and bias may be in conflict with each other.

Definition 2.9 (*kernel density estimation*). Let (Y_j) be i.i.d with CDF $F_Y(y)$ and, the kernel density estimator (KDE) using a nonnegative normalized K and bandwidth $h > 0$ is

$$\hat{f}(y) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{Y_j - y}{h}\right)$$

Using rectangular kernel $K = I(Y \in [-1/2, 1/2])$ and let $\hat{F}$ denote the empirical CDF 1.8.

$$\hat{f}(y) = \frac{1}{nh}\sum_{j=1}^{n} I\left(Y_j \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right]\right) = \frac{\hat{F}(y + h/2) - \hat{F}(y - h/2)}{h}$$

Theorem 2.2 (*bias and variance of KDE*). Given i.i.d. (Y_j) with $f_Y(y)$ twice differentiable. For small h

$$\text{bias}(\hat{\theta}) \approx \frac{1}{24}h^2 f_Y''(y), \quad \text{Var}(\hat{\theta}) \approx \frac{f_Y(y)}{nh}$$

Definition 2.10 (*mean squared error (MSE)*). The MSE of an estimator $\hat{\theta}$ is θ-dependent

$$\text{MSE}(\hat{\theta}; \theta) = \mathbb{E}_{\mathbf{y}\sim\mathbf{Y}_\theta}[\hat{\theta}(\mathbf{y}) - \theta]$$

Theorem 2.3 (*MSE decomposition*). $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2$.

Proof: Let $V = \hat{\theta} - \theta$, then $\text{MSE}(\hat{\theta}) = \mathbb{E}[V^2] = \text{Var}(\hat{\theta} - \theta) + \mathbb{E}[\hat{\theta} - \theta]^2$.

2.3 Consistency and Asymptotics

Definition 2.11 (*consistent*). An estimator $\hat{\theta}$ is consistent with an estimand θ if it converges in probability (1.2) to θ as $n \to \infty$, written in shorthand as $\hat{\theta} \xrightarrow{p} \theta$.

$$\forall \epsilon > 0, \lim_{n\to\infty} \Pr_{\mathbf{y}\sim\mathbf{Y}_\theta}(|\hat{\theta}_{\mathbf{y},n} - \theta| \geq \epsilon) \to 0$$

Since θ is the constant, this is equivalent to $\hat{\theta} \xrightarrow{d} \theta$.

Theorem 2.4 (*a sufficient condition for consistency*). Vanishing MSE implies consistency.

Proof: Using Markov's inequality 1.1 on $(\hat{\theta} - \theta)^2$

$$\Pr(|\hat{\theta} - \theta| \geq \epsilon) = \Pr[(\hat{\theta} - \theta)^2 \geq \epsilon^2] \leq \frac{1}{\epsilon^2}\mathbb{E}[(\hat{\theta} - \theta)^2] = \frac{1}{\epsilon^2}\text{MSE}(\hat{\theta}) \to 0$$

Remark 2.1. The converse is not true: consistency does not imply $\mathbb{E}[\hat{\theta}] \to \theta$. Consider standard uniform U and $(Y_j) \sim \text{Bern}(\theta)$, let

$$\hat{\theta} = \bar{Y} + nI(U \leq 1/n)$$

Note that $\mathbb{E}(\hat{\theta}) = \theta + 1$ but $\hat{\theta} \to \theta$ since the second term occurs with vanishing probability.

Proposition 2.5. MoM estimators are consistent.

Proof: Recall the definition: here h is a multivariate function with multiple components

$$\alpha(\theta) = \mathbb{E}_\theta[h(\mathbf{Y})] \iff \hat{\theta}_{\text{MoM}} = \alpha^{-1}\left[\frac{1}{n}\sum_{j=1}^{n} h(\mathbf{y}_j)\right]$$

By the law of large numbers, $\frac{1}{n}\sum h(\mathbf{y}_j) \xrightarrow{p} \mathbb{E}_\theta[h(\mathbf{Y})]$. Applying the continuous mapping theorem yields $\hat{\theta}_{\text{MoM}} \xrightarrow{p} \theta$.

Due to the construction of MoM estimators by summation, the CLT implies that their asymptotics is normal, which allows us to apply the delta method.

Theorem 2.6 (*normal asymptotics of MoM estimators*).

$$\alpha(\theta) = \mathbb{E}_\theta[h(\mathbf{Y})] \implies \sqrt{n}(\hat{\theta}_{\text{MoM}} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}[h(\mathbf{Y})]}{(\partial_\theta \alpha)^2}\right)$$

More conveniently, isolating $\hat{\theta}_{\text{MoM}}$ yields, for large n

$$\hat{\theta}_{\text{MoM}} \sim \mathcal{N}\left(\theta, \frac{\text{Var}[h(\mathbf{Y})]}{n(\partial_\theta \alpha)^2}\right)$$

Proof: Denote by $\mathbf{Y}_j$ i.i.d copies of $\mathbf{Y}$ and assuming $\text{Var}[h(\mathbf{Y})]$ bounded, by the CLT

$$\sqrt{n}\left[\frac{1}{n}\sum_{j=1}^{n} h(\mathbf{Y}_j) - \mathbb{E}[h(\mathbf{Y})]\right] \xrightarrow{d} \mathcal{N}(0, \text{Var}[h(\mathbf{Y})])$$

Let $\beta = \mathbb{E}[h(\mathbf{Y})]$ so that $\theta = \alpha^{-1}(\beta)$, applying the delta method yields

$$\sqrt{n}\left[\alpha^{-1}\left(\frac{1}{n}\sum_{j=1}^{n} h(\mathbf{Y}_j)\right) - \alpha^{-1}(\beta)\right] = \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}[h(\mathbf{Y})]}{(\partial_\beta \alpha^{-1}(\beta))^2}\right)$$

Substituting the inverse derivative relation yields the desired equation, as claimed

$$\alpha(\theta) = \beta \implies (\partial_\theta \alpha)(\partial_\beta \theta) = 1 \implies \partial_\beta \theta = \partial_\beta \alpha^{-1}(\beta) = \frac{1}{\partial_\theta \alpha}$$

2.4 Sample quantile Asymptotics

The sample quantile provides a nontrivial example of an estimator which is not arithmetically constructed. We generalize sample quantile results for the standard normal

Proposition 2.7. The order statistics of the standard uniform is

$$Y_{(j)} = \text{Beta}(j, n-j+1), \quad \mathbb{E}[Y_{(j)}] = \frac{j}{n+1}, \quad \text{Var}[Y_{(j)}] = \frac{j(n-j+1)}{(n+1)^2(n+2)}$$

Intuitively, $P(Y_{(j)} = x) \propto x^{j-1}(1-x)^{n-j}$, giving the beta distribution.

Lemma 2.8. for i.i.d standard uniform variables (Y_j), their asymptotic distribution is

$$\lim_{n\to\infty} \sqrt{n}\left[Y_{(\lceil np \rceil)} - Q_{Y_1}(p)\right] \xrightarrow{d} \mathcal{N}(0, p(1-p))$$

Proof: Recalling $Q_{Y_1}(p) = p$, denote $Z_{j,n} = \sqrt{n}\left[Y_{(j)} - p\right]$ so that $Y_{(j)} = \dfrac{Z_{j,n}}{\sqrt{n}} + p$. Since $Y_{(j)}$ and $Z_{(j,n)}$ are related by linearly

$$\begin{aligned} f_{Z_{j,n}}(z) &\propto (f_{y_{(j)}}(z))^{j-1}(1 - f_{y_{(j)}}(z))^{n-j} \\ &= \left(p + \frac{z}{\sqrt{n}}\right)^{j-1}\left(1 - p - \frac{z}{\sqrt{n}}\right)^{n-j} \\ &= p^{j-1}(1-p)^{n-j}\left(1 + \frac{z}{p\sqrt{n}}\right)^{j-1}\left(1 - \frac{z}{(1-p)\sqrt{n}}\right)^{n-j} \end{aligned}$$

For small x we have $\log(1+x) \approx x - x^2/2$. Let $\alpha = \dfrac{z}{p\sqrt{n}}$ and $\beta = \dfrac{z}{(1-p)\sqrt{n}}$.

$$\log f_{Z_{j,n}}(z) \approx C + (j-1)\left(\alpha - \frac{\alpha^2}{2}\right) + (n-j)\left(-\beta - \frac{\beta^2}{2}\right) \to c - \frac{z^2}{2p(1-p)}$$

This is the log-density of $\mathcal{N}(0, p(1-p))$.

Lemma 2.9. Let $Q = F^{-1}$ and $f = \partial_x F$, then $\partial_p Q = \dfrac{1}{f(Q(p))}$

Proof: Let $x = Q(p)$, then $F(x) = p$ and taking ∂_p on both sides yields

$$1 = (\partial_x F)(\partial_p x) = f(x)\partial_p Q \implies \partial_p Q = \frac{1}{f(x)} = \frac{1}{f(Q(p))}$$

Theorem 2.10 (*asymptotic distribution of sample quantile*). Given i.i.d RVs (Y_j)

$$\lim_{n\to\infty} \sqrt{n}\left[Y_{(\lceil np \rceil)} - Q_Y(p)\right] \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f_Y(Q(p))^2}\right)$$

Proof: Let (U_j) be i.i.d copies of the standard uniform and $Q = Q_{Y_1}$ so that $Y = Q(U)$ (a good way to remember this is that $F_Y(Y) = U$). Also note that $Y_{(\lceil np \rceil)} = Q(U_{(\lceil np \rceil)})$ and $Q_U(p) = p$. Recalling lemma 2.8

$$\lim_{n\to\infty} \sqrt{n}\left[U_{(\lceil np \rceil)} - p\right] \xrightarrow{d} \mathcal{N}(0, p(1-p))$$

Apply the delta method with map Q and using lemma 2.9

$$\lim_{n\to\infty} \sqrt{n}\left[Q(U_{(\lceil np \rceil)}) - Q(Q_{U_1}(p))\right] \xrightarrow{d} \mathcal{N}(0, [\partial_p Q(p)]^2 p(1-p)) = \mathcal{N}\left(0, \frac{p(1-p)}{f_Y(Q(p))^2}\right)$$

On the left hand side, using $Q_{U_1}(p) = p$ yields the desired relation

$$\sqrt{n}\left[Q(U_{(\lceil np \rceil)}) - Q(p)\right] = \sqrt{n}\left[Y_{(\lceil np \rceil)} - Q_Y(p)\right]$$

Note the high variance for low values of $f_Y(x)$. It is very hard to estimate the tails.

3 Maximum Likelihood Estimation

Definition 3.1 (*Maximum likelihood estimator (MLE)*). Recalling 2.4, the MLE of θ is

$$\hat{\theta} = \text{argmax } L(\theta; \mathbf{y})$$

Proposition 3.1. The MLE is invariant under reparameterization.

Proof: Follows from the invariance of likelihood.

We show that MLE is consistent, asymptotically normal, asymptotically unbiased, and asymptotically efficient (no asymptotically unbiased estimator has lower standard error).

Definition 3.2 (*regularity conditions*). In the rest of this section we assume the following regularity conditions on our statistical model:

- Differentiable almot everywhere.
- The support of $F_{\mathbf{Y};\theta}$ is independent of θ.
- For all θ, the distribution is differentiable under the integral sign.

3.1 KL-divergence and MLE consistency

We use the notation θ^* being the estimand with data distribution $F_{\mathbf{Y};\theta^*}$.

Definition 3.3 (*Kullback-Leibler(KL)-divergence*). Given two CDF F, G

$$D_{\text{KL}}(F\|G) = \mathbb{E}_{\mathbf{Y}\sim F}\left[\log \frac{f(\mathbf{Y})}{g(\mathbf{Y})}\right] = \int \log\left(\frac{f(\mathbf{y})}{g(\mathbf{y})}\right) f(\mathbf{y})\, d\mathbf{y}$$

Note that the expectation is taken with respect to the first argument. When F is the data generating process $F_{\mathbf{Y};\theta^*}$ and $G = F_{\mathbf{Y};\theta}$ is our estimated process, this is related to likelihood

$$D_{\text{KL}}(F_{\mathbf{Y};\theta^*}\|F_{\mathbf{Y};\theta}) = \mathbb{E}_{\mathbf{Y}\sim F_{\mathbf{Y};\theta^*}}\left[\log \frac{f_{\theta^*}(\mathbf{Y})}{f_\theta(\mathbf{Y})}\right] = \mathbb{E}_{\mathbf{Y}_{\theta^*}}\left[\log L(\theta^*; \mathbf{Y}) - \log L(\theta; \mathbf{Y})\right] \tag{3.1}$$

Theorem 3.2 (*independence additivity of KL divergence*). Given independent (Y_j), then

$$D_{\text{KL}}(F_{\mathbf{Y};\theta^*}\|F_{\mathbf{Y};\theta}) = \sum_{j=1}^{n} D_{\text{KL}}(F_{Y_j;\theta^*}\|F_{Y_j;\theta})$$

Proof: Additivity of log-likelihood under independence and linearity of expectation.

Lemma 3.3. Let f, g be distributions, $\mathbb{E}_{\mathbf{Y}\sim f}[g(\mathbf{Y})/f(\mathbf{Y})] = 1$

Proof: Direct computation

$$\mathbb{E}_{\mathbf{Y}\sim f}\left[\frac{g(\mathbf{Y})}{f(\mathbf{Y})}\right] = \int \frac{g(\mathbf{Y})}{f(\mathbf{Y})} f(\mathbf{y})\, d\mathbf{y} = \int g(\mathbf{y})\, d\mathbf{y} = 1$$

Theorem 3.4 (*nonnegativity of KL-divergence*). $D_{\text{KL}}(F\|G) \geq 0$ with 0 when $F = G$.

Proof: Using the convexity of $-\log x$

$$D_{\mathrm{KL}}(F\|G) = \mathbb{E}_{\mathbf{Y}\sim F}\left[-\log\frac{g(\mathbf{Y})}{f(\mathbf{Y})}\right] \geq -\log \mathbb{E}_{\mathbf{Y}\sim F}\left[\frac{g(\mathbf{Y})}{f(\mathbf{Y})}\right] = 0$$

Equality holds if and only if $g(Y)/f(Y)$ for all Y, in which case $g = f \implies G = F$.

Corollary 3.1. Consider the expectation of likelihood θ averaged over samples

$$\zeta(\theta) = \mathbb{E}_{\mathbf{Y}\sim F_{\mathbf{Y};\theta^*}}[L(\theta;\mathbf{Y})]$$

The MLE $\hat{\theta}$ of θ maximizes $\zeta(\theta)$. The estimand maximizes the expectation of likelihood.

Proof: Using 3.1, the second term is independent while the first is minimized for $\theta = \theta^*$

$$\zeta(\theta) = -D_{\mathrm{KL}}(F_{\mathbf{Y};\theta^*}\|F_{\mathbf{Y};\theta}) - \mathbb{E}_{\mathbf{Y}_{\theta^*}}[\log L(\theta^*;\mathbf{Y})]$$

Theorem 3.5 (*consistency of MLE*). Assuming finite parameter space, i.i.d observations Y_j, and identifiability of the model (i.e. $\theta_1 \neq \theta_2 \implies F_{\theta_1} \neq F_{\theta_2}$), then $\hat{\theta} \xrightarrow{p} \theta^*$.

Proof: Using the strong law of large numbers 1.3

$$\frac{1}{n}[\log L(\theta^*;\mathbf{Y}) - \log L(\theta;\mathbf{Y})] \xrightarrow{a.s.} D_{\mathrm{KL}}(F_{Y;\theta^*}\|F_{Y;\theta})$$

Recalling the implications of almost sure converence 1.3, for every $\epsilon > 0$ there exists a sample size n_0 such that the empirical divergence is within ϵ of the true KL-divergence. By the finite parameter space assumption, choose the least admissible divergence

$$c = \min_{\theta\neq\theta^*} D_{\mathrm{KL}}(F_{Y;\theta^*}\|F_{Y;\theta}) > 0$$

Choose $0 < \epsilon < c$, then with probability one there is n_0 such that $\forall n \geq N, \theta \neq \theta^*$

$$\frac{1}{n}[\log L(\theta^*) - \log L(\theta_n)] - D_{\mathrm{KL}}(F_{\theta^*}\|F_\theta) > -\epsilon \implies \log L(\theta^*) - \log L(\theta_n) > n(c-\epsilon)$$

Then for such $n \geq N$ with the convergence threshold given by the law of large numbers,

$$L(\theta^*) \gg \log L(\theta_n \neq \theta^*)$$

By finiteness, $\hat{\theta} = \theta^*$ for all such sufficiently large n.

3.2 Score function, Fisher information, and Cramér-Rao

Definition 3.4 (*score function*). The score function of an estimate θ given data $\mathbf{y}$ is

$$s(\theta;\mathbf{y}) = \partial_\theta \log L(\theta;\mathbf{y}) = \frac{1}{L(\theta;\mathbf{y})}\partial_\theta L(\theta;\mathbf{y})$$

The MLE $\hat{\theta}$ satisfies $s(\hat{\theta};\mathbf{y}) = 0$. Intuitively, $s(\theta;\mathbf{y})$ quantifies how sensitive the likelihood of observing $\mathbf{y}$ is with respect to a given value of θ.

Theorem 3.6 (*information equality*). Under regularity conditions

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{Y}_{\theta^*}}[s(\theta^*;\mathbf{y})] = 0, \quad \mathrm{Var}[s(\theta^*;\mathbf{y})] = -\mathbb{E}\left[\partial_\theta\big|_{\theta^*} s(\theta;\mathbf{y})\right]$$

In expectation, the score is zero for $\theta = \theta^*$.

Proof: Proving the first equation first, recalling $L(\theta;\mathbf{y}) = f(\mathbf{y};\theta)$

$$\begin{aligned}\mathbb{E}[s(\theta)] &= \int \partial_\theta\left[\log L_{\mathbf{y}}(\theta)\right] f(\mathbf{y};\theta^*)\,d\mathbf{y}\\ &= \int \frac{1}{L(\theta;\mathbf{y})}\partial_\theta L(\theta;\mathbf{y}) f(\mathbf{y};\theta^*)\,d\mathbf{y}\\ &= \int \partial_\theta f(\mathbf{y};\theta)\cdot\frac{f(\mathbf{y};\theta^*)}{f(\mathbf{y};\theta)}\,d\mathbf{y}\end{aligned}$$

Specializing to $\theta = \theta^*$

$$\mathbb{E}[s(\theta^*)] = \int \partial_\theta\big|_{\theta^*} f(\mathbf{y};\theta)\,d\mathbf{y} = \partial_\theta\bigg|_{\theta^*}\int f(\mathbf{y};\theta)\,d\mathbf{y} = \partial_\theta\big|_{\theta^*} 1 = 0$$

To compute the information inequality, consider $\partial_\theta s$

$$\partial_\theta s = \partial_\theta\left(\frac{1}{L(\theta)}\partial_\theta L(\theta)\right) = \frac{1}{L(\theta)}\partial_\theta^2 L - \frac{(\partial_\theta L)^2}{L(\theta)^2} = \frac{1}{L(\theta)}\partial_\theta^2 L - (\partial_\theta \log L)^2 = \frac{1}{L(\theta)}\partial_\theta^2 L - s^2$$

The expectation value of the second term vanishes at $\theta = \theta^*$

$$\int \frac{1}{L(\theta^*)}\partial_\theta^2\big|_{\theta^*} L(\theta)\, f(\mathbf{y};\theta^*)\,d\mathbf{y} = \partial_\theta^2\bigg|_{\theta^*}\int L(\theta)\,d\mathbf{y} = 0$$

The second term yields $\mathbb{E}[s^2]$. At $\theta = \theta^*$ this is $\mathrm{Var}[s(\theta^*;\mathbf{y})]$.

Definition 3.5 (*Fisher information*). The Fisher information *in the sample* for a parameter θ with value θ^* with respect to a parametric statistical model $F_{\mathbf{Y};\theta}$ is

$$\mathcal{I}_{\mathbf{Y}}(\theta^*) = \mathrm{Var}[s(\theta^*;\mathbf{y})] = \mathbb{E}[s(\theta^*;\mathbf{y})^2] = -\mathbb{E}\left[\partial_\theta\big|_{\theta^*} s(\theta;\mathbf{y})\right] = -\mathbb{E}\left[\partial_\theta^2\big|_{\theta^*}\log L(\theta;\mathbf{y})\right]$$

Note that all the statistics here are computed w.r.t $\mathbf{y}\sim F_{\mathbf{Y};\theta^*}$. Two useful perspectives:

- $\mathbb{E}[s(\theta^*;\mathbf{y})^2]$: how sensitive, averaged over the data distribution, the observed log-likelihood is with respect to a change in parameter. The more sensitive, conversely the more "information" does model give us about an estimate.
- $-\mathbb{E}\left[\partial_\theta^2\big|_{\theta^*}\log L(\theta;\mathbf{y})\right]$: information is high when the log-likelihood is highly concave.

Theorem 3.7 (*reparameterization of Fisher information*). Let $\tau = \tau(\theta)$, then

$$\mathcal{I}_{\mathbf{Y}}(\tau) = \frac{\mathcal{I}_{\mathbf{Y}}(\theta)}{\tau'(\theta)^2}$$

When τ is very sensitive to θ, a small change in θ can yield a large change in τ, so tha variance (inversely proportional to Fisher information) for an estimator of τ is larger, see CRLB below.

Proof: The score transforms as $s(\theta;\mathbf{Y}) = \partial_\theta \log L(\theta;\mathbf{Y}) = \left[\partial_\tau \log L(\tau;\mathbf{Y})\right]\partial_\theta\tau = \tau'(\theta)s(\tau)$

$$\mathcal{I}_{\mathbf{Y}}(\tau) = \mathbb{E}[s(\tau;\mathbf{y})^2] = \frac{\mathbb{E}[s(\theta;\mathbf{y})^2]}{g'(\theta)^2} = \frac{\mathcal{I}_{\mathbf{Y}}(\theta)}{\tau'(\theta)^2}$$

Fisher information outlines fundamental limits for the MSE of estimators.

Theorem 3.8 (*Cramér-Rao lower bound (CRLB)*). Let $\hat{\theta}$ be an unbiased estimator of θ in a parametric model $F_{\mathbf{Y};\theta}$. Under regularity conditions

$$\text{MSE}(\hat{\theta}) \geq \text{Var}(\hat{\theta}) \geq \mathcal{I}_{\mathbf{Y}}(\theta)^{-1} \tag{3.2}$$

Proof: Let $S = s(\theta^*; \mathbf{Y})$. For covariance in general

$$\text{Cov}(S, \hat{\theta})^2 \leq \text{Var}(S)\text{Var}(\hat{\theta}) = \mathcal{I}(\theta^*)\text{Var}(\hat{\theta}) \implies \text{Var}(\hat{\theta}) \geq \mathcal{I}(\theta^*)^{-1}\text{Cov}(S, \hat{\theta})^2$$

Computing the left hand side, $\mathbb{E}[S] = \mathbb{E}[s(\theta^*; \mathbf{Y})]$ vanishes by 3.6.

$$\begin{aligned}\text{Cov}(S,T) = \mathbb{E}[ST] - \mathbb{E}[S]\mathbb{E}[T] = \mathbb{E}[ST] &= \int \frac{\partial_\theta f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)}\hat{\theta}(\mathbf{y})\, f(\mathbf{y};\theta)\, d\mathbf{y} \\ &= \partial_\theta \int \hat{\theta}(\mathbf{y}) f(\mathbf{y};\theta)\, d\mathbf{y} = \partial_\theta \mathbb{E}[\hat{\theta}] = \partial_\theta \theta = 1\end{aligned}$$

Theorem 3.9 (*extended CRLB*). Using the same conditions as above except $\mathbb{E}[\hat{\theta}] = g(\theta)$

$$\text{Var}(\hat{\theta}) \geq \frac{g'(\theta)^2}{\mathcal{I}(\theta)}$$

Proof: Evaluating $\text{Cov}(S, \hat{\theta})$ again, $\text{Cov}(S,T) = \partial_\theta \mathbb{E}[\hat{\theta}] = g'(\theta)$. Substitute into equation 3.2.

3.3 Asymptotics of MLE

Theorem 3.10 (*asymptotic distribution of MLE*). Let $\hat{\theta}$ be the MLE of a scalar parameter based on i.i.d observations $(Y_j) \sim F_{\mathbf{Y};\theta^*}$. Under regularity conditions,

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_Y^{-1}(\theta^*))$$

Proof: Let $\mathbf{Y}$ denote the random variable of the observed data

$$0 = s(\hat{\theta}; \mathbf{Y}) = s(\theta^*; \mathbf{Y}) + (\hat{\theta} - \theta^*)s'(\theta^*; \mathbf{Y}) + O[(\hat{\theta} - \theta^*)^2]$$

The higher-order terms vanish quickly by MLE consistency. Rearranging yields

$$\sqrt{n}(\hat{\theta} - \theta^*) \approx \frac{\sqrt{n}s(\theta^*; \mathbf{Y})}{-s'(\theta^*; \mathbf{Y})} = \frac{\sqrt{n}\left[\frac{1}{n}\sum_{j=1}^n s(\theta^*; \mathbf{Y}_j)\right]}{-\frac{1}{n}s'(\theta^*; \mathbf{Y})}$$

Applying CLT to the numerator by recalling 3.6 and definition 3.5

$$\sqrt{n}\left[\frac{1}{n}\sum_{j=1}^n s(\theta^*; \mathbf{Y}_j)\right] \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_Y(\theta^*))$$

In the numerator, by the strong law of large numbers

$$\frac{1}{n}s'(\theta^*; \mathbf{Y}) = \frac{1}{n}\sum_{j=1}^n s'(\theta^*; \mathbf{Y}_j) \xrightarrow{a.s.} \mathbb{E}[-s'(\theta^*)] = \mathcal{I}_Y(\theta^*)$$

Using Slutsky's theorem yields the asymptotic $\mathcal{I}_Y(\theta^*)^{-1}\mathcal{N}(0, \mathcal{I}_Y(\theta^*)) = \mathcal{N}(0, \mathcal{I}_Y^{-1}(\theta^*))$

Corollary 3.2. MLE is asymptotically unbiased since bias$(\hat{\theta}) \to 0$, and MLE is asymptotically efficient by asymptotically saturating the CLRB.

For multi-variate parameter, the score and fisher information are

$$s(\theta; \mathbf{y}) = \nabla_\theta \log L(\theta; \mathbf{y}), \quad I(\theta^*) = \mathrm{Var}(s(\theta^*; \mathbf{y})) = -\mathbb{E}\left[H_\theta \big|_{\theta^*} \log L(\theta; \mathbf{Y}) \right]$$

Here H_θ denotes Hessian. CLRB becomes, for $\mathrm{Var}(\hat{\theta}), \mathcal{I}(\theta^*)$ positive semidefinite operators

$$\mathrm{Var}(\hat{\theta}) - \mathcal{I}(\theta^*)^{-1} \geq 0$$

The asymptotics of MLE is

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1})$$

4 Confidence Intervals

Definition 4.1 (*interval, coverage, margin of error*). An interval estimate $C(\mathbf{y})$ of scalar θ based on the data $\mathbf{y}$ is an interval $[L(\mathbf{y}), U(\mathbf{y})]$.

- The functions L, U define an interval estimator $C(\mathbf{Y})$ as a function of the data.
- The coverage probability of the interval estimator is $\Pr(\theta \in C(\mathbf{Y}))$, with θ being the gound truth parameter of the distribution (fixed) and randomness over sampling $\mathbf{Y}$.
- The margin of error is the half-width $0.5[U(\mathbf{Y}) - L(\mathbf{Y})]$.

Definition 4.2 (*(asymptotic) pivot*). A pivot is a random variable whose distribution is known. An asymptotic (approximate) pivot is a random variable whose distribution is known in the limit $n \to \infty$.

Remark 4.1. To construct a CI for θ, the pivot needs to depend on θ, but its distribution cannot depend on unknown parameters. In contrast, a statistic cannot depend on unknown parameters, but its distribution typically depends on unknown parameters.

Example 4.1 (*normal and t*). Given $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$, construct the pivot $(\hat{\theta} - \theta)/\sigma \sim \mathcal{N}(0, 1)$. Let F denote the cdf of the standard normal, then

$$\Pr(a \le (\hat{\theta} - \theta)/\sigma \le b) = F(b) - F(a) \iff \theta \in [\hat{\theta} - b\sigma, \hat{\theta} - a\sigma]$$

For $\alpha = 0.95$ coverage, replace $b = 1.96, a = -1.96$. If σ is unknown, the quantity $(\hat{\theta} - \theta)/\hat{\sigma} \to \mathcal{N}(0, 1)$ constitutes an asymptotic pivot via CMT and Slutsky. Alternatively use the t-pivot $(\hat{\theta} - \theta)/\hat{\sigma} \sim t_{n-1}$.

$$C(\mathbf{Y}) = \bar{Y} \pm Q_{t_{n-1}}(1 - \alpha/2)\frac{\hat{\sigma}}{\sqrt{n}}$$

Example 4.2 (*binomial CI*). Given $Y \sim \text{Bin}(n, p)$. Using the MLE $\hat{p} = \bar{Y}$ with asymptotics

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

This yields a pivot with corresponding CI $C(\mathbf{Y}) = \hat{p} \pm 1.96\sqrt{p(1-p)/n}$. With p unknown, use $\hat{p}$ or conservative estimates $p = 1/2$ for an approximate CI.

Example 4.3 (*p-quantile CI*). To estimate $\theta = Q(p)$, using the MLE asymptotics

$$\sqrt{n}\frac{f_Y(\theta)}{p(1-p)}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 1)$$

yields $C(\mathbf{Y}) = \hat{\theta} \pm \dfrac{1.96}{f_Y(\theta)}\sqrt{\dfrac{p(1-p)}{n}}$. Replace $\theta \mapsto \hat{\theta}$ and $f_Y(\theta)$ with KDE estimate for approximate CI.

Example 4.4 (*variance CI*). Given $Y_1 \cdots Y_n \sim \mathcal{N}(\mu, \sigma^2)$ with both parameters unknown, recall

$$V = \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

Using this pivot yields the CI

$$Q(\alpha/2) \le V \le Q(1 - \alpha/2) \iff \sigma \in \left[\hat{\sigma}\sqrt{\frac{n-1}{Q(1-\alpha/2)}}, \hat{\sigma}\sqrt{\frac{n-1}{Q(\alpha/2)}}\right]$$

Example 4.5 (*exponential rate CI*). Given $Y_1 \cdots Y_n \sim \text{Expo}(\lambda)$ with mean $\mu = 1/\lambda$. We can:

- Use MLE asymptotics for λ.
- Use MLE asymptotics for μ to derive an equivalent CI for λ via inversion.
- $\sum Y_j = n\bar{Y} \sim \text{Gamma}(n, \lambda) \iff \lambda n\bar{Y} \sim \text{Gamma}(n, 1)$. Then $C(\mathbf{Y}) = \dfrac{1}{n\bar{Y}}[Q(0.025), Q(0.975)]$.

5 Regression

Definition 5.1 (*predictive, descriptive regression*). The central object of predictive regression is the conditional distribution $\mu(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. Descriptive regression learns a linear function $\mu(\mathbf{X})$ so that $Y - \mu_L(\mathbf{X})$ is small on average over a pair $(\mathbf{X}, Y)$; descriptive regression is concerned with the joint behavior of pairs $(\mathbf{X}, Y)$.

Definition 5.2 (*homo(hetero)-skedasticity*). Define $\sigma(\mathbf{x}) = \mathrm{Var}(Y|\mathbf{X} = \mathbf{x})$. Predictive regression is homo(hetero)-skedastic if $\sigma(\mathbf{x})$ does not vary(varies) with $\mathbf{x}$.

5.1 Predictive Regression

Definition 5.3 (*regression error*). $U(\mathbf{x}) = Y - \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$.

Theorem 5.1 (*signal-noise decomposition*). The regression error vanishes conditionally and unconditionally and is uncorrelated with each feature.

$$\mathbb{E}[U(\mathbf{X})|\mathbf{X}] = \mathbb{E}[U(\mathbf{X})] = \mathrm{Cov}(U(\mathbf{X}), X_j) = 0$$

Proof: Conditional vanishing holds by construction

$$\mathbb{E}[U(\mathbf{X})|\mathbf{X} = \mathbf{x}] = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] - \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = 0$$

Unconditional vanishing holds by Adam's law

$$\mathbb{E}[U(\mathbf{X})] = \mathbb{E}[\mathbb{E}[U(\mathbf{X})|\mathbf{X}]] = 0$$

For the covariance, since $\mathbb{E}[U(\mathbf{X})]$ vanishes

$$\mathrm{Cov}(U(\mathbf{X}), X_j) = \mathbb{E}[U(\mathbf{X})X_j] = \mathbb{E}[\mathbb{E}[X_j U(\mathbf{X})|\mathbf{X}]] = \mathbb{E}[X_j \mathbb{E}[U(\mathbf{X})|\mathbf{X}]] = 0$$

Theorem 5.2 (*variance decomposition*). The unconditional variance of error U and

$$\mathrm{Var}(U) = \mathbb{E}[\sigma^2(\mathbf{X})], \quad \mathrm{Var}(Y) = \mathbb{E}[\sigma^2(\mathbf{X})] + \mathrm{Var}[\mu(\mathbf{X})] = \mathrm{Var}(U) + \mathrm{Var}(\mu)$$

The R^2 of the predictor is the ratio of explained variance

$$R^2 = \frac{\mathrm{Var}(\mu)}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}(U)}{\mathrm{Var}(Y)}$$

Proof: $\mathrm{Var}(U) = \mathbb{E}[\mathrm{Var}(U|\mathbf{X})] + \mathrm{Var}(\cancelto{0}{\exp[U\mathbf{X}])} = \mathbb{E}[\sigma^2(\mathbf{X})]$. Similarly,

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y|\mathbf{X}]) + \mathbb{E}[\mathrm{Var}(Y|\mathbf{X})] = \mathrm{Var}[\mu(\mathbf{X})] + \mathbb{E}[\sigma^2(\mathbf{X})]$$

Definition 5.4 (*linear regression model*). $\mathbb{E}[Y|X = \mathbf{x}, \theta] = \theta \cdot (\mathbf{x}, 1)$.

Definition 5.5 (*odds, logit*). The odds of a binary random variable is the probability of success over the probability of failure: $\eta = p/(1-p)$. The logit is the log-odds with inverse

$$\lambda = \log p - \log(1-p), \quad p = \sigma(\lambda) = \frac{1}{1 - e^{-\lambda}}$$

Additionally, the variance of p w.r.t. λ is the variance

$$\partial_\lambda p = p(1-p)$$

Definition 5.6 (*logistic regression*). A binary Y invites the natural interpretation

$$\mathbb{E}[Y|\mathbf{X}] = \Pr[Y = 1|\mathbf{X}] = \mu(x)$$

The logistic regression model assumes the conditional mean

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}, \theta] = \mu(\mathbf{x}) = \sigma[\theta \cdot (\mathbf{x}, 1)]$$

The log-odds are linear in the parameters. In particular

$$\theta_j = \partial_{x_j}\lambda = \frac{1}{p(1-p)}\partial_{x_j}p$$

The parameter θ_j encodes the rate of change of the logit λ w.r.t. x_j.

Theorem 5.3 (*i.i.d. predictive regression*). Given an i.i.d. pdf $f(y|x)$ such that

$$f(\mathbf{y}|\mathbf{X} = \mathbf{x}; \theta) = \prod f(y_j|\mathbf{x}_j; \theta)$$

General results for the regression model

$$\theta = \text{argmax } s(\theta), \quad s(\theta) = \sum s_j(\theta), \quad s_j(\theta) = \partial_\theta \log f(y_j|x_j; \theta)$$

Definition 5.7 (*residuals*). The residuals are analogous to crystallizations of the error $U(\mathbf{x}) = Y - \mathbb{E}[Y|\mathbf{x}]$ upon replacing $\mathbb{E}[Y|\mathbf{x}]$ with our estimates and $\mathbf{x}, \mathbf{Y}$ with observations

$$\hat{U}_j = y_j - x_j\hat{\theta}$$

General formula relating the error to residuals: using $U_j = y_j - \mu_j = y_j - \hat{\theta}x_j$

$$\hat{U}_j = (y_j - \mu_j) - (x_j\hat{\theta} - \mu_j) = U_j - x_j(\hat{\theta} - \theta) \tag{5.1}$$

Example 5.1 (*Gaussian regression without intercept*). Fixing $Y_j|x_j \sim \mathcal{N}(\theta x_j, \sigma^2)$ then

$$\log f(y_j|x_j) \simeq -\frac{(y_j - \theta x_j)^2}{2\sigma^2}, \quad s(\theta) \simeq \frac{1}{\sigma^2}\sum x_j U_j \quad \mathcal{I}(\theta) = \frac{\sum x_j^2}{\sigma^2}, \quad \hat{\theta} = \frac{\sum x_j y_j}{\sum x_j^2}$$

View (y_j) as an approximately scaled vector for (x_j), then $\hat{\theta}$ is the scaling coefficient; as corollary $\sum x_j\hat{U}_j = 0$. Inference will condition $\mathbf{X}$ at observed values x, then

$$\mathbb{E}[\hat{\theta}|\mathbf{x}] = \frac{\sum x_j\mu_j}{\sum x_j^2}, \quad \text{Var}(\hat{\theta}|\mathbf{x}) = \frac{\sum x_j^2\sigma_j^2}{\left(\sum x_j^2\right)^2}, \quad \mu_j = \mathbb{E}[Y_j|X_j = x_j], \quad \sigma_j^2 = \text{Var}(Y_j|X_j = x_j)$$

- Least-squares as above is conditionally unbiased as long as $\mu_j = \mathbb{E}[Y_j|x_j] = \theta x_j$.
- Given full assumption $Y_j|\mathbf{x}_j \sim \mathcal{N}(\theta x_j, \sigma^2)$, Gaussian regression without intercept conditionally saturates the CRLB.
- Specializing 5.1: using $\theta x_j = \mu(x_j)$ and $\hat{\theta} - \theta = \dfrac{\sum x_j(y_j - \theta x_j)}{\sum x_j^2} \implies \hat{U}_j = U_j - x_j\dfrac{\sum x_i U_i}{\sum x_i^2}$.

Theorem 5.4 (*robust standard error*). For large n we may estimate the variance without the homoskedastic assumption by replacing $\sigma_j^2 \mapsto \hat{U}_j^2$ in the original variance formula

$$\text{Var}(\hat{\theta}|\mathbf{x}) \sim \frac{\sum x_j^2\hat{U}_j^2}{\left(\sum x_j^2\right)^2}$$

Proposition 5.5. MoM derivation. modeling $\mathbb{E}[Y|x] = \theta x$ and ssuming i.i.d (X_j, Y_j):

$$\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[Y|X]X] = \theta\mathbb{E}[X^2] \implies \hat{\theta}_{\text{MoM}} = \frac{\sum X_j Y_j}{\sum X_j^2}$$

Example 5.2 (*Gaussian regression with intercept*). Modeling $Y_j|x_j \sim \mathcal{N}(\theta_0 + \theta_1 x, \sigma^2)$

$$s(\theta_0, \theta_1) = \sum (y_j - \theta_0 - \theta_1 x_j)^2, \quad \hat{\theta}_1 = \frac{\sum (x_j - \bar{x})(y_j - \bar{y})}{\sum (x_j - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

The estimate $\hat{\theta}_1$ is conditionally Gaussian (it is linear in y_j and $y_j|x_j$ is Gaussian) with

$$\hat{\theta}_1 \sim \mathcal{N}\left(\theta_1, \frac{\sigma^2}{\sum (x_j - \bar{x})^2}\right), \quad \hat{\theta}_0 \sim \mathcal{N}\left(\theta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_j - \bar{x})^2}\right)\right), \quad \text{Cov}(\hat{\theta}_0, \hat{\theta}_1) = -\frac{\sigma^2 \bar{x}}{\sum (x_j - \bar{x})^2}$$

The standard errors of the mean $\hat{Y}_{\text{new}} = \hat{\theta}_0 + \hat{\theta}_1 x_{\text{new}}$ at a new point x_{new} and point estimate Y_{new} are

$$\text{Var}(\hat{Y}_{\text{new}}) = \sigma^2\left(\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum (x_j - \bar{x})^2}\right), \quad \text{Var}(Y_{\text{new}}) = \text{Var}(\hat{Y}_{\text{new}}) + \sigma^2$$

Proof: Using $\sum (x_j - \bar{x})(y_j - \bar{y}) = \sum (x_j - \bar{x}) y_j$ to write

$$\hat{\theta}_1 = \sum \frac{x_j - \bar{x}}{\sum (x_j - \bar{x})^2} y_j \implies \text{Var}(\hat{\theta}_1|\mathbf{x}) = \frac{\sigma^2}{\sum (x_j - \bar{x})^2}$$

Note that $\bar{Y}$ is conditionally independent of $\hat{\theta}_1$ (weighted sum of $Y_j - \bar{Y}$), results for $\hat{\theta}_0$ follow

$$\hat{\theta}_0 = \bar{Y} - \hat{\theta}_1 \bar{X} \sim \mathcal{N}\left(\theta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_j - \bar{x})^2}\right)\right)$$

For the covariance, invoke $\text{Cov}(\hat{\theta}_0, \hat{\theta}_1) = \text{Cov}(\bar{Y} - \hat{\theta}_1 \bar{x}, \hat{\theta}_1)$. Use the decorrelated decomposition

$$\hat{Y}_{\text{new}} = \hat{\theta}_0 + \hat{\theta}_1 x_{\text{new}} = \bar{Y} + \hat{\theta}_1 (x_{\text{new}} - \bar{x})$$

The result for $\text{Var}(Y_{\text{new}})$ follows from the regression variance decomposition 5.2.

5.2 Descriptive regression

Instead of modeling $\mathbb{E}[Y|\mathbf{x}]$, consider (X, Y) jointly. The central statistic is $\beta_{Y\sim X} = \dfrac{\text{Cov}(X, Y)}{\text{Var}(X)}$.

Proposition 5.6. Least squares property of regression slope

$$(\alpha, \beta_{Y\sim X}) = \text{argmin}_{a,b} L(a, b), \quad L(a, b) = \mathbb{E}[(Y - a - bX)^2]$$

Note that this does not make any parametric assumption the conditional $Y|X$.

Proof: The derivative extremization condition implies

$$\mathbb{E}[Y - a - bX] = \mathbb{E}[X(Y - a - bX)] = 0 \implies \beta_{Y\sim X} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad \alpha = \mathbb{E}[Y] - \beta_{Y\sim X}\mathbb{E}[X]$$

Definition 5.8 (*linear projection, error*). the linear projection of Y on X at $X = x$ is

$$\mu_L(x) = \mathbb{E}[Y] + \beta_{Y\sim X}(x - \mathbb{E}[X])$$

This is the best *linear* function of x for approximating Y in terms of MSE (compare to $\mathbb{E}[Y|X = x]$, which is the best general function of x that approximates Y). The linear error is the random variable

$$U_L = Y - \mu_L(X)$$

Note that $\mathbb{E}[U_L] = 0$ by construction and $\mathbb{E}[XU_L] = 0$ by the extremization condition $\mathbb{E}[X(Y-a-bX)] = 0$. Relating this to the prediction error: letting $\mu(X) = \mathbb{E}[Y|X]$

$$Y - \mu_L(X) = (Y - \mu(X)) + (\mu(X) - \mu_L(X))$$

This is the sum of the regression error $Y - \mu(X)$ (zero-meaned with certain variance) plus the linear approximation error. For linear models the second term vanish.

Theorem 5.7 (*asymptotics of descriptive slope without intercept*). Writing $U_j^L = Y_j - \theta X_j$

$$\hat{\theta} = \theta + \frac{\frac{1}{n}\sum X_j U_j^L}{\frac{1}{n}\sum X_j^2}$$

Using CLT on the numerator and LLN on the denominator

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{Var}(XU^L)}{\mathbb{E}[X^2]^2}\right) = \mathcal{N}\left(0, \frac{\mathbb{E}[X^2\sigma^2(X)]}{\mathbb{E}[X^2]^2}\right)$$

The second equality follows from recalling the predictive notation $\sigma^2(x) = \mathrm{Var}(U|X = x)$

$$\mathrm{Var}(XU) = \mathrm{Var}(\cancel{\mathbb{E}[XU|X]}) + \mathbb{E}[\mathrm{Var}(XU|X)] = \mathbb{E}[X^2\sigma^2(X)]$$

6 Exponential Families and Sufficiency

6.1 Exponential Families

Definition 6.1 (*natural exponential family*). A density is a n.e.f. if

$$f(y;\theta) = e^{\theta y - \psi(\theta)} h(y)$$

where $h(y)$ is independent of θ. Here θ is the nautral parameter.

Example 6.1 (*normal*). Recall the normal density $\mathcal{N}(\mu, \sigma^2)$

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2} = \exp\left(\frac{\mu y}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \frac{e^{-y^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

Assuming σ known, the natural parameter is $\theta = \dfrac{\mu}{\sigma^2}$ with

$$\psi(\theta) = \frac{\theta^2\sigma^2}{2} \quad h(y) = \frac{e^{-y^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

Note that $h(y)$ is the pdf of $\mathcal{N}(0, \sigma^2)$.

Example 6.2 (*binomial*). Let $Y \sim \text{Bin}(n, p)$, then

$$\begin{aligned} P(y;p) &= \binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} \exp\left[y \log p + (n-y)\log(1-p)\right] \\ &= \binom{n}{y} \exp\left(y \log \frac{p}{1-p} + n\log(1-p)\right) \end{aligned}$$

The natural parameter is $\theta = \log \dfrac{p}{1-p} = \sigma^{-1}(p)$, for σ the sigmoid function. Then

$$n\log(1-p) = n\log\left(1 - \frac{e^\theta}{1+e^\theta}\right) = n\log\frac{1}{1+e^\theta} = -n\log(1+e^\theta)$$

Rewrite the density in terms of this parameterization

$$P(y;p) = \binom{n}{y} \exp(y\theta - n\log(1+e^\theta)), \quad \psi(\theta) = n\log(1+e^\theta), \quad h(y) = \binom{n}{y} (y \in^? \mathbb{N})$$

Theorem 6.1 (*mean and variance of NEF*). Formulas for mean, variance, and skewness

$$\mathbb{E}[Y] = \psi'(\theta), \quad \text{Var}(Y) = \psi''(\theta), \quad \text{Skew}(Y) = \mathbb{E}\left[\left(\frac{Y-\mu}{\sigma}\right)^3\right] = \frac{\psi'''(\theta)}{\psi''(\theta)^{3/2}}$$

Proof: $\displaystyle\int e^{\theta y - \psi(\theta)} h(y)\, dy = 1 \iff \int e^{\theta y} h(y)\, dy = e^{\psi(\theta)}$. Different w.r.t. θ

$$\int y e^{\theta y} h(y)\, dy = \psi'(\theta) e^{\psi(\theta)} \iff \mathbb{E}[Y] = \int y e^{\theta y - \psi(\theta)} h(y)\, dy = \psi'(\theta)$$

Proof for $\text{Var}(Y)$ is obtained similarly.

Theorem 6.2 (*MLE of NEF*). Let $\mu = \psi'(\theta)$ be the mean of i.i.d (Y_j) NEFs, then

$$\hat{\mu} = \bar{Y}, \quad \hat{\theta} = (\psi')^{-1}(\bar{Y}), \quad \mathcal{I}_Y(\theta) = n\psi''(\theta) = \text{Var}(Y), \quad \mathcal{I}_Y(\mu) = \psi''(\theta)^{-1}$$

The asymptotic MLE of θ and μ are

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \psi''(\theta)^{-1})$$
$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \psi''(\theta))$$

Proof: The log-likelihood is

$$\log L(\theta; \mathbf{y}) = n(\theta\bar{y} - \psi(\theta)), \quad \partial_\theta \log L = n(\bar{y} - \psi'(\theta))$$

Set the score to 0 yields $\psi'(\hat{\theta}) = \bar{y}$. Then $\mathcal{I}_Y(\theta) = -\mathbb{E}[\partial_\theta^2 \log L] = n\psi''(\theta)$. By reparameterization, $\mathcal{I}_Y(\mu) = n/y''(\theta)$. To show that μ saturates CRLB, calculate

$$\text{Var}(\hat{\mu}) = \text{Var}(\hat{Y}) = \text{Var}(Y)/n = n/\psi''(\theta)$$

Definition 6.2 (*exponential family*). A density is an exponential family if

$$f(y; \theta) = e^{\theta T(y) - \psi(\theta)} g(y)$$

A NEF is the special case when $T(y) = y$.

6.2 Sufficient Statistic

Definition 6.3 (*sufficient statistic*). A statistic $T((Y_j) \sim F_{\mathbf{Y};\theta})$ is sufficient for θ if

$$f(\mathbf{y}; \theta) = g_\theta(T(\mathbf{y}))h(\mathbf{y}) \tag{6.1}$$

This is equivalent to $(Y_j)|T$ being independent of θ: given independence

$$f(\mathbf{y}; \theta) = \Pr(T = t; \theta) \Pr(\mathbf{y}|T = t) = g_\theta(t)h(\mathbf{y})$$

Conversely, given 6.1

$$\Pr(\mathbf{y}|t; \theta) = \frac{\Pr(\mathbf{y}, t; \theta)}{\Pr(t; \theta)} = \frac{g_\theta(t)h(\mathbf{y})}{\sum_{T(\mathbf{y}')=t} \Pr(\mathbf{y}'; \theta)} = \frac{g_\theta(t)h(\mathbf{y})}{\sum_{T(\mathbf{y}')=t} g_\theta(t)h(\mathbf{y}')} = \frac{h(\mathbf{y})}{\sum_{T(\mathbf{y}')=t} h(\mathbf{y}')}$$

The key about sufficiency is that *the data only (effectively only, i.e. coupled with θ) enter the likelihood through the sufficient statistic.*

Proposition 6.3. The sample mean $\bar{Y}$ is a sufficient statistic for θ for i.i.d. N.E.F

Proof: Apply the factorization criterion 6.1

$$f(\mathbf{y}; \theta) = e^{n(\theta\bar{y} - \psi(\theta))} h_n(\mathbf{y}) = g_\theta(\bar{y})h(\mathbf{y}), \quad g_\theta(\bar{y}) = e^{n(\theta\bar{y} - \psi(\theta))}$$

Example 6.3 (*Bernoulli*). Given $(Y_j) \sim \text{Bern}(p)$, then $\bar{p}$ is sufficient for p.

Theorem 6.4 (*Rao-Blackwell*). Let $\hat{\theta}$ be an estimator for θ and T a sufficient statistic for θ, then the Rao-Blackwellized estimator

$$\hat{\theta}_{\text{RB}} = \mathbb{E}[\hat{\theta}|T]$$

is at least as good as θ in MSE, with equality implying $\hat{\theta}$ a deterministic function of T.

Proof: Consider the bias, $\mathbb{E}[\hat{\theta}_{\text{RB}}] = \mathbb{E}[\hat{\theta}]$ by Adam's law. For the variance

$$\text{Var}(\hat{\theta}) = \mathbb{E}[\text{Var}(\hat{\theta}|T)] + \text{Var}(\mathbb{E}[\hat{\theta}|T]) = \mathbb{E}[\text{Var}(\hat{\theta}|T)] + \text{Var}(\hat{\theta}_{\text{RB}}) \geq \text{Var}(\hat{\theta}_{\text{RB}})$$

7 Hypothesis Testing

7.1 Definitions

Definition 7.1 (*statistical hypothesis*). Give $F_{\mathbf{Y};\theta}$ for $\theta \in \Theta$ unknown, a hypothesis is a logical statement $\theta \in \Theta_0 \subseteq \Theta$. It is simple if Θ_0 is a single point and composite otherwise.

Definition 7.2 (*hypothesis test, critical regions*). A (hypothesis) test specifies which observed $\mathbf{y} \sim F_{\mathbf{Y};\theta}$ leads to H_0 being rejected (otherwise retained). The retention region A is the set of possible $\mathbf{y}$ such that we decide $\mathbf{y} \in A \implies H_0$ true. Its complement is the rejection (critical) region.

Definition 7.3 (*test statistic, critical values*). $T(\mathbf{y})$ is a test statistic if the test procedure is of the form "$T(y) \notin [c_L, c_U] \implies H_0$ false". In this case (c_L, c_U) are the critical values.

Definition 7.4 (*power function*). Given $\mathbf{y} \sim F_{\mathbf{Y};\theta}$ and a test with retention region A, the power function of our test is

$$\beta(\theta) = P_{\mathbf{Y};\theta}(\mathbf{Y} \notin A)$$

This is the probability of rejecting the null hypothesis given θ.

Definition 7.5 (*sided tests*). A two-sided test is $H_0 : \theta = \theta_0$ (resp. $H_1 : \theta \neq \theta_0$). A one-sided test is of the form $H_0 : \theta \leq \theta_0$ or $H_0 : \theta \geq \theta_0$. The goal of a two-sided test is to test equality, while that of a one-sided test is to test directionality.

Definition 7.6 (*errors, (nominal) size*). Two types of errors are defined w.r.t. H_1:

$$\begin{aligned} \text{Type I error (false positive)} &: \theta \in \Theta_0 \text{ but } \mathbf{y} \notin A. \ H_1 \text{ is falsely accepted.} \\ \text{Type II error (false negative)} &: \theta \notin \Theta_0 \text{ but } \mathbf{y} \in A. \ H_1 \text{ is falsely rejected.} \end{aligned}$$

The size (or level) α of a test is the maximum possible Type I (FP) error probability.

$$\alpha = \max_{\theta \in \Theta_0} \beta(\theta)$$

A α-sized test is valid if its size is indeed α. The size of a simple test $H_0 : \theta = \theta_0$ is $\beta(\theta_0)$. Fair comparisons of tests should be conducted across fixed size.

Controlling the size of a test relies on knowing the distribution of the test statistic taken the null hypothesis. A nominal size α is the size holding under an asymptotic distribution. For a simply null, a test has nominal size α if

$$\lim_{n \to \infty} P_{\mathbf{Y};\theta_0}(\mathbf{Y} \in A^C) = \alpha$$

While constructing tests, the following theorem is often useful.

Theorem 7.1 (*t-statistic and t distribution*). The t-statistic of i.i.d. Gaussian data follows a t-distribution: Given $\mathbf{Y} = (Y_j)_{1 \cdots n}$ being i.i.d. $\mathcal{N}(\mu, \sigma^2)$

$$T(\mathbf{Y}) = \frac{\sqrt{n}(\bar{Y} - \mu)}{\hat{\sigma}} \sim t_{n-1}, \quad \hat{\sigma} = \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \bar{Y})^2$$

From now on we consider a simple H_0. Recall that a $(1-\alpha)$-CI $C(\mathbf{Y})$ satisfies

$$P_{\mathbf{Y};\theta}[\theta \in C(\mathbf{Y})] = 1 - \alpha$$

while a α-level test satisfies

$$P_{\mathbf{Y};\theta_0}[\mathbf{Y} \in A(\theta_0)] = 1 - \alpha$$

Theorem 7.2 (*test-CI duality*). Constructing α-level tests using $(1-\alpha)$-CIs, and vice versa.

Proof: Given a $(1-\alpha)$-CI $C(\mathbf{Y})$, construct the test $A(\theta_0) = \{\mathbf{Y} : \theta_0 \in C(\mathbf{Y})\}$, then

$$P[\mathbf{Y} \in A(\theta_0)] = P[\theta_0 \in C(\mathbf{Y})] = 1 - \alpha$$

Conversely, given a α-level test $A(\theta_0)$, construct $C(\mathbf{Y}) = \{\theta : \mathbf{Y} \in A(\theta)\}$, then

$$P[\theta \in C(\mathbf{Y})] = P[\mathbf{Y} \in A(\theta)] = 1 - \alpha$$

The test $A(\theta)$ needs to be α-level for all θ.

7.2 Likelihood-based testing

Consider two-sided simple test $H_0 : \theta = \theta_0$ about the population parameter. Recall MLE asymptotics

$$\hat{\theta} \sim \mathcal{N}(0, \mathcal{I}_{\mathbf{Y}}^{-1}(\theta))$$

Definition 7.7 (*Wald test*). Using the asymptotic pivot for $\hat{\theta}$ above

$$T(\mathbf{Y}) = \sqrt{\mathcal{I}_{\mathbf{Y}}(\theta_0)}(\hat{\theta} - \theta_0) \sim \mathcal{N}(0,1) \iff W(\mathbf{Y}) = \mathcal{I}_{\mathbf{Y}}(\theta_0)(\hat{\theta} - \theta_0)^2 \sim \chi^2_1$$

This yields the formula for retention region

$$A(\theta_0) = \{\mathbf{Y} : W(\mathbf{Y}) < Q_{\chi^2_1}(1-\alpha)\}$$

Definition 7.8 (*score test*). The score test uses the asymptotic distribution of the score

$$T(\mathbf{Y}) = \frac{s(\theta_0; \mathbf{Y})}{\sqrt{\mathcal{I}_{\mathbf{Y}}(\theta_0)}} \sim \mathcal{N}(0,1) \iff S(\mathbf{Y}) = \frac{s^2(\theta_0; \mathbf{Y})}{\mathcal{I}_{\mathbf{Y}}(\theta_0)} \sim \chi^2_1$$

The score asymptotics arise from the information equality 3.6 and CLT

$$\mathbb{E}[s(\theta_0)] = 0, \quad \mathrm{Var}[s(\theta_0)] = \mathcal{I}_Y(\theta_0)$$

Definition 7.9 (*likelihood ratio test*). Given L and null θ_0, the log-likelihood statistic $\Lambda(\mathbf{y})$ for the likelihood ratio test is twice the log-(likelihood ratio) (here $\hat{\theta}$ is the MLE given $\mathbf{y}$)

$$\Lambda(\mathbf{y}) = 2\left[\log L(\hat{\theta}; \mathbf{y}) - \log L(\theta_0; \mathbf{y})\right]$$

Given the null $\theta = \theta_0$, rejecting the null if $\Lambda(\mathbf{y}) > Q_{\chi^2_1}(1-\alpha)$ has nominal α-size. Note that θ_0 should not be on the boundary of the parameter space.

Proof: Under the null, $\hat{\theta} \xrightarrow{p} \theta_0$, so expanding θ_0 about θ.

$$\log L(\theta_0) \approx \log L(\hat{\theta}) + (\theta_0 - \hat{\theta})\cancelto{}{s(\hat{\theta})} + \frac{1}{2}(\theta_0 - \hat{\theta})^2 s'(\hat{\theta})$$

$$\Lambda(\mathbf{y}) \approx (\hat{\theta} - \theta_0)^2(-s'(\hat{\theta})) = [\sqrt{n}(\hat{\theta} - \theta_0)]^2 \frac{-s'(\hat{\theta})}{n}$$

The second term converges to $\mathcal{I}_{Y_1}(\theta_0)$ given i.i.d by SLLN. The first term $\xrightarrow{d} \mathcal{I}_{Y_1}(\theta_0)^{-1}\chi^2_1$.

7.3 p-value

Definition 7.10 (*p-value*). The p-value is the probability, under the null, of obtaining a test statistic at least as extreme as the observed statistic. It is a r.v. dependent upon the sampling behavior of $\mathbf{y}$.

- *Simple null:* Given $H_0 : \theta = \theta_0$, test $A(\theta_0) = \{\mathbf{y} | T(\mathbf{y}) > c\}$ and observing $\mathbf{y}$. The observed p-value is

$$p(\mathbf{y}) = P_{\mathbf{y}' \sim F_{\theta_0}}(T(\mathbf{y}') \geq T(\mathbf{y}))$$

- *General null:* let A_α be the retention region for each α such that $P_{\theta_0}(\mathbf{y} \in A_\alpha^C) = \alpha$. Observing $\mathbf{y}$, the crystallized p-value is the smallest α at which we could have rejected H_0.

Theorem 7.3 (*distribution of p-value*). Given a true simple null and continuous $T(\mathbf{Y})$, the p-value is uniformly distributed: $p(\mathbf{Y}) \sim \text{Unif}(0, 1)$.

Proof: Without loss of generality reject H_0 for T large. Observing $t_0 = T(\mathbf{y})$ and let F_T denote the cdf of T under H_0, the p-value is $p(\mathbf{y}) = 1 - F_T(T(\mathbf{y}))$. Note the second term is $\text{Unif}(0, 1)$.

8 Bayesian Inference

8.1 Framework and Point Estimates

Bayesian inference models the uncertainty about population parameter with probability.

Definition 8.1 (*prior, posterior, marginal likelihood*). We posit a *joint distribution* for

$$\mathbf{Y}, \theta$$

The conditional distribution $f(\mathbf{y}|\theta) = L(\theta; \mathbf{y})$ is the frequentist likelihood, the prior density $\pi(\theta)$ and posterior density $\pi(\theta|\mathbf{y})$. The marginal likelihood $f(\mathbf{y})$ is the data distribution.

Theorem 8.1 (*Bayes' rule*). Regarding $\mathbf{y}$ as crystallized, the posterior is proportional to the likelihood times prior, with proportionality $f(\mathbf{y})$ independent of θ.

$$\pi(\theta|\mathbf{y}) = \frac{L(\theta; \mathbf{y})\pi(\theta)}{f(\mathbf{y})}$$

Here $L(\theta; \mathbf{y}) = f(\mathbf{y}|\theta)$ is the frequentist likelihood (conditional probability), and the marginal is

$$f(\mathbf{y}) = \int d\tilde{\theta}\, L(\tilde{\theta}; \mathbf{y})\pi(\tilde{\theta}) \tag{8.1}$$

Definition 8.2 (*mean, median, and mode*). The prior (posterior) mean, median, and modes are defined in the obvious ways w.r.t. the prior (posterior) distribution. The posterior mode is the maximum a posteriori (MAP) estimator

$$\hat{\theta}_{\text{MAP}} = \text{argmax } \pi(\theta|\mathbf{y}) = \text{argmax } L(\theta; \mathbf{y})\pi(\theta)$$

Theorem 8.2 (*variational characterizations*). Posterior mean (median) minimizes MSE (absolute loss).

$$\mathbb{E}[\theta|\mathbf{y}] = \text{argmin}_{\hat{\theta}}\ \mathbb{E}[(\theta - \hat{\theta})^2|\mathbf{y}]$$
$$Q_{\theta|\mathbf{y}}(0.5) = \text{argmin}_{\hat{\theta}}\ \mathbb{E}[|\theta - \hat{\theta}||\mathbf{y}]$$

Here θ is the random variable: fixing $\mathbf{y}$ fixes a distribution $\theta|\mathbf{y}$, and the mean (median) of a distribution are simply the values which minimize the respective losses.

Proof: Given r.v. $\alpha = \theta|\mathbf{y}$, $\partial_c\mathbb{E}[(\alpha - c)^2] \propto \mathbb{E}[\alpha - c] = 0 \implies c = \mathbb{E}[\alpha]$. For the quantile result, write

$$\begin{aligned}\partial_c\mathbb{E}[|\alpha - c|] &= \partial_c \int f(\alpha')[(c - \alpha')I(c > \alpha') + (\alpha' - c)I(c < \alpha')]\, d\alpha' \\ &= \int d\alpha'\, f(\alpha')[I(c > \alpha') - I(c < \alpha')] = 1 - 2F(c) = 0 \implies c = Q(0.5)\end{aligned}$$

Theorem 8.3 (*Tweedie's formula*). Consider a Gaussian likelihood with $\mu = \theta$.

$$f(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y - \theta)^2\right]$$

The posterior mean and variance are

$$\mathbb{E}[\theta|y] = y + \sigma^2\partial_y \log f(y), \quad \text{Var}(\theta|y) = \sigma^2 + \sigma^4\partial_{y^2} \log f(y)$$

In particular, with prior $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, $Y \sim \mathcal{N}(\mu_0, \sigma^2 + \sigma_0^2)$ is of Gaussian form, with

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y|\theta]) + \mathbb{E}[\mathrm{Var}(Y|\theta)] = \sigma^2 + \mathbb{E}[\mathrm{Var}(\mu)] = \sigma^2 + \sigma_0^2$$

Substituting into Tweedie's formulae yields

$$\mathbb{E}[\theta|y] = \frac{\sigma_0^2 y + \sigma^2 \mu_0}{\sigma^2 + \sigma_0^2}, \quad \mathrm{Var}(\theta|y) = \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}$$

Proof: Given the likelihood

$$\partial_y f(y|\theta) = \frac{\theta - y}{\sigma^2} f(y|\theta)$$

Some clever manipulations yield

$$\begin{aligned} f(y)\,[\mathbb{E}[\theta|y] - y] &= f(y) \int d\theta\,(\theta - y)\pi(\theta|y) = f(y) \int d\theta\,(\theta - y)\frac{f(y|\theta)\pi(\theta)}{f(y)} \\ &= \int d\theta\,(\theta - y) f(y|\theta)\pi(\theta) = \sigma^2 \int [\partial_y f(y|\theta)]\pi(\theta)\,d\theta \\ &= \sigma^2 \partial_y \int f(y|\theta)\pi(\theta)\,d\theta = \sigma^2 \partial_y f(y) \end{aligned}$$

This yields Tweedie's formula, which can be invoked after choosing a prior to determine $f(y)$

$$\mathbb{E}[\theta|y] = y + \sigma^2 \partial_y \log f(y)$$

Similarly taking advantage of the exponential form of $f(y|\theta)$ to relate ∂_{y^2} to $f(y|\theta)$ yields variance formula.

The likelihood $\pi(\mathbf{y}|\theta)$ is easy to compute, but computing the posterior mean requires computing $P(\mathbf{y}) = \int \pi(\mathbf{y}|\theta)\,d\theta$ for $\pi(\theta|\mathbf{y})$ as well as the integral

$$\mathbb{E}[\theta|\mathbf{y}] = \int \theta\pi(\theta|\mathbf{y})\,d\theta = \frac{\int \theta\,\pi(\mathbf{y}|\theta)\pi(\theta)\,d\theta}{\int \pi(\mathbf{y}|\theta)\pi(\theta)\,d\theta}$$

8.2 Credible Intervals, Conjugate priors

Using the posterior distribution, we can directly construct a credible interval.

Definition 8.3 (*credible interval*). Given $0 < \alpha < 1$, a $(1-\alpha)$-credible interval for θ is an interval estimate $[a(\mathbf{y}), b(\mathbf{y})]$ such that

$$\Pr(\theta \in [a(\mathbf{y}), b(\mathbf{y})]) = 1 - \alpha$$

Credible intervals can be constructed directly from the posterior $\theta|\mathbf{y}$. One common choice is

$$[Q_{\theta|\mathbf{y}}(\alpha/2), Q_{\theta|\mathbf{y}}(1 - \alpha/2)]$$

Letting I be the indicator of the credible interval containing θ. The coverage probability is $1-\alpha$, averaging over sampling randomness of $\mathbf{Y}$.

$$\Pr(I = 1) = \mathbb{E}[\Pr(I|\mathbf{Y})] = \mathbb{E}[1 - \alpha] = 1 - \alpha$$

Remark 8.1. The frequentist probability for the confidence interval quantifies uncertainty over sampling $\mathbf{y}$, while the Bayesian probability quantifies uncertainty over the posterior distribution itself.

Definition 8.4 (*conjugate priors*). A family of priors is conjugate to a statistical model if computed posteriors remain in the family of priors. Several examples:

1. Beta-binomial: $Y|p \sim \text{Bin}(n,p), p \sim \text{Beta}(a,b) \implies p|y \sim \text{Beta}(a+y, b+n-y)$.
2. Gamma-Poisson: $Y|\lambda \sim \text{Pois}(\lambda), \lambda \sim \Gamma(r_0, b_0) \implies \lambda|y \sim \Gamma(r_0+y, b_0+1)$.

Theorem 8.4 (*normal-normal conjugacy*). Given statistical model $Y_j|\mu \sim \mathcal{N}(\mu, \sigma^2)$ parameterized by μ with i.i.d (Y_j) and prior $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$.

$$\mu|\mathbf{y} \sim \mathcal{N}(\mu_n, \tau_n^2), \quad \tau_n^2 = \frac{1}{n/\sigma^2 + 1/\tau_0^2}, \quad \mu_n = \tau_n^2 \left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right)$$

Proof: It suffices to consider the sufficient statistic

$$\bar{Y}|\mu \sim \mathcal{N}(\mu, \sigma^2/n)$$

since sufficiency implies $\mu|\mathbf{y} = \mu|\bar{y}$. Consider single-sample case for $Y|\mu \sim \mathcal{N}(\mu, \sigma^2), \mu \sim \mathcal{N}(\mu_0, \tau_0^2)$ with $\sigma^2, \mu_0, \tau_0^2$ known, then (ignoring constants)

$$\begin{aligned} \log \pi(\mu|y) \cong \log \pi(y|\mu) + \log \pi(\mu) &\cong -\frac{1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2\tau_0^2}(\mu-\mu_0)^2 \\ &= -\frac{1}{2}\mu^2 \left(\frac{1}{\sigma^2} + \frac{1}{\tau_0^2} \right) + \mu \left(\frac{y}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right) \end{aligned}$$

This is quadratic in μ. Matching with the Gaussian log-likelihood yields

$$\mu|y \sim \mathcal{N}(\mu_1, \tau_1^2), \quad \tau_1^2 = \frac{1}{1/\sigma^2 + 1/\tau_0^2}, \quad \mu_1 = \tau_1^2 \left(\frac{y}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right) \tag{8.2}$$

Substituting $\sigma^2 \mapsto \sigma^2/n$ into the expression above yields the theorem.

Theorem 8.5 (*normal-normal conjugacy with heteroskedasticity*). Extending to i.i.d $Y_j|\mu \sim \mathcal{N}(\mu, \sigma_j^2)$

$$\mu|(y_j) \sim \mathcal{N}(\mu_n, \tau_n^2), \quad \tau_n^2 = \frac{1}{1/\tau_0^2 + \sum 1/\sigma_j^2}, \quad \mu_n = \tau_n^2 \left(\sum \frac{y_j}{\sigma_j^2} + \frac{\mu_0}{\tau_0^2} \right)$$

8.3 Model choice, Marginal Likelihood, and Hypothesis Testing

Definition 8.5 (*Bayes factor*). Model comparisons are based on the marginal-likelihood ratio for the data

$$\beta = \frac{f(\mathbf{y}|\text{Model } A)}{f(\mathbf{y}|\text{Model } B)} \implies \frac{P(\text{Model } A|\mathbf{y})}{P(\text{Model } B|\mathbf{y})} = \frac{P(\text{Model } A)}{P(\text{Model } B)}\beta$$

Theorem 8.6 (*candidate's formula*). The Bayes factor involves calculating the marginal likelihood $f(\mathbf{y})$ (suppressing conditioning on the model). Using Baye's rule

$$f(\mathbf{y}) = \frac{f(\mathbf{y}|\theta = t) f(\theta = t)}{f(\theta = t|\mathbf{y})}$$

this holds for all values of t, so choose one which simplifies matter.

Bayesian hypothesis testing is straightforward using the posterior of the null $\Pr(\theta \in \Theta_0|\mathbf{y})$.

8.4 Bayesian Prediction

Definition 8.6 (*posterior predictive distribution*). Given that we wish to predict Y using observed data, X, the posterior predictive distribution of Y is the Bayesian conditional distribution $Y|X$.

Using parameterization θ, introduce a prior $\pi(\theta|\mathbf{X})$. Observing $(X_1, Y_1) \cdots (X_n, Y_n)$, the posterior predictive distribution is

$$f(y_{n+1}|\mathbf{x}, \mathbf{y}) = \int f(y_{n+1}|\mathbf{x}, \mathbf{y}, \theta)\pi(\theta|\mathbf{x}, \mathbf{y})\, d\theta$$

This integral consists of a posterior component $\pi(\theta|\mathbf{x}, \mathbf{y})$ and a likelihood component $f(y_{n+1}|\mathbf{x}, \mathbf{y}, \theta)$.

Example 8.1 (*normal-normal posterior predictive*). In a normal-normal model

$$(Y_j)|\mu \sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \tau_0^2)$$

The key simplification here is that $(Y_j), \mu$ is multivariate-normal, so the conditional is again a normal. Using theorem 8.4

$$\begin{aligned}
\mathbb{E}[Y_{n+1}|(Y_j)] &= \mathbb{E}[\mathbb{E}[\cancelto{\mu}{Y_{n+1}|(Y_j), \mu}]|(Y_j)] = (1 - b_n)\bar{Y}_n + b_n\mu_0, \quad b_n = \frac{\sigma^2}{\sigma^2 + n\tau_0^2} \\
\mathrm{Var}(Y_{n+1}|(Y_j)) &= \mathbb{E}[\mathrm{Var}(Y_{n+1}|(Y_j), \mu)|(Y_j)] + \mathrm{Var}(\mathbb{E}[Y_{n+1}|(Y_j), \mu]) \\
&= \mathbb{E}[\sigma^2|(Y_j)] + \mathrm{Var}(\mu|(Y_j)) = \sigma^2 + \tau_n^2
\end{aligned}$$

Then $Y_{n+1}|(Y_j) \sim \mathcal{N}(\mu_n, \tau_n^2 + \sigma^2)$.

Example 8.2 (*Poisson posterior predictive*). Observing $(Y_j)|\theta \sim \mathrm{Pois}(c_j\theta)$ with prior $\theta \sim \mathrm{Gamma}(r_0, b_0)$.

$$\theta|(Y_j) \sim \mathrm{Gamma}\left(r_0 + \sum Y_j, b_0 + \sum c_j\right)$$

via extension of 8.4. The marginal (definition 8.1 and calculated using theorem 8.6) for Y_{n+1} is

$$Y_{n+1} \sim \mathrm{NBin}\left(r_0, \frac{b_0}{b_0 + c_{n+1}}\right)$$

8.5 Hierarchical models and Stein's Paradox

Definition 8.7 (*multi-level Gaussian hierarchicel models*). With $(\sigma, \lambda_0, \lambda_0)$ known, the tuple $(\mathbf{Y}, \boldsymbol{\mu})$ form a two-level Gaussian hierarchical model if (i.i.d for each j)

$$\mu_j \sim \mathcal{N}(\gamma, \lambda_0^2), \quad Y_j \sim \mathcal{N}(\mu_j, \sigma^2)$$

In the two-level model, (Y_j) are conditionally independent given $\boldsymbol{\mu}$. Using theorem 8.4

$$\mu_j|\mathbf{y} \sim \mathcal{N}\left(\lambda_K^2\left(\frac{\gamma}{\lambda_0^2} + \frac{y_j}{\sigma^2}\right), \lambda_K^2\right), \quad \lambda_K^2 = \frac{1}{1/\lambda_0^2 + 1/\sigma^2}, \quad Y_j \sim \mathcal{N}(\gamma, \sigma^2 + \lambda_0^2)$$

The tuple $(\mathbf{Y}, \boldsymbol{\mu}, \gamma)$ form a three-level Gaussian hierarchichal level if, additionally, with (g_0, τ_0) known

$$\gamma \sim \mathcal{N}(g_0, \tau_0^2)$$

Applying theorem 8.4 on $Y_j|\gamma \sim \mathcal{N}(\gamma, \sigma^2 + \lambda_0^2)$ yields the conditional results on γ. Noting γ, μ_j are still normal and using Adam and Eve's laws to obtain the Gaussian parameters yield

$$\gamma|\mathbf{y} \sim \mathcal{N}(g_K, \tau_K^2), \quad g_K = \frac{g_0/\tau_0^2 + \frac{k\bar{y}}{\sigma^2+\lambda_0^2}}{1/\tau_0^2 + k/(\sigma^2 + \lambda_0^2)}$$

$$\mu_j|\mathbf{y} = \mathcal{N}\left(\frac{g_K/\lambda_0^2 + y_j/\sigma^2}{1/\lambda_0^2 + 1/\sigma^2}, \frac{\tau_K^2}{\lambda_0^4(1/\lambda_0^2 + 1/\sigma^2)^2} + \frac{1}{1\lambda_0^2 + 1/\sigma^2}\right), \quad \tau_k^2 = \frac{1}{1/\tau_0^2 + K/(\sigma^2 + \lambda_0^2)}$$

Definition 8.8 (*risk function, inadmissibility*). Given loss function $\text{Loss}(\theta, \hat{\theta})$ an estimator $\hat{\theta} = \hat{\theta}(\mathbf{Y})$ its associated risk function

$$R(\theta) = \mathbb{E}_{\mathbf{Y};\theta}[\text{Loss}(\theta, \hat{\theta})]$$

is the average loss incurred for $\hat{\theta}$ being the estimate of θ given $\mathbf{Y}$. An estimator $\hat{\theta}$ is inadmissible if there is another estimator whose risk function is less than or equal to that of $\hat{\theta}$ for all values of θ, with strict inequality for at least one θ.

Theorem 8.7 (*Stein's theorem*). Consider $Y_j \sim \mathcal{N}(\mu_j, \sigma^2)$ i.i.d with $K \geq 3$, μ_j unknown and σ^2 known. The MLE $\mathbf{Y} = (Y_j)$ for $\boldsymbol{\mu} = (\mu_j)_{1 \leq j \leq K}$ w.r.t. the total squared loss $\sum(\mu_j - \hat{\mu}_j)^2$ is inadmissible.

9 Design-based Inference

Randomness arises from sampling observed data from an underlying stochastic model in the model-based framework. In the design-based framework, we consider a fixed finite population $y_1 \cdots y_N$ and randomness arises from the sampling process. Estimands are correspondingly computed using finite population samples.

9.1 Sampling design

Definition 9.1 (*sampling design*). Fixing population size N and random sample size n. Label the population with IDs $1 \cdots N$ and let I_j denote the ID of the j-th unit in the random sample. A sampling design is a joint pmf of $\mathbf{I} = I_1 \cdots I_n \in \{1, \cdots, N\}^n$. The sampling design determines the statistical properties of

$$\mathbf{Y}_I = (Y_1, \cdots, Y_n) = (y_{I_1}, \cdots, Y_{I_n})$$

The descriptive estimators for μ, σ^2, F are, respectively

$$\bar{Y} = \frac{1}{n}\sum Y_j, \quad S^2 = \frac{1}{n-1}\sum (Y_j - \bar{Y})^2, \quad \hat{F}(y) = \frac{1}{n}\sum I(Y_j \le y)$$

Definition 9.2 (*equal probability sample (EPS)*). $P(I)$ describes an equal probability sample if

$$\Pr(I_j = k) = 1/N, \quad \forall j \in [n], k \in [N]$$

Theorem 9.1 (*EPS properties*). EPS provides unbiased population estimates of μ, σ^2, and $F(y)$.

$$\mathbb{E}_I[Y_j] = \mu, \quad \mathrm{Var}_I(Y_j) = \sigma^2, \quad \mathbb{E}_I[\bar{Y}] = \mu, \quad \mathbb{E}_I[\hat{F}(y)] = F(y)$$

Proof: Linearity and unrolling definitions: note that randomness arises from the index instead of y_j itself.

$$\begin{aligned}
\mathbb{E}_I[\bar{Y}] &= \frac{1}{n}\sum_{j=1}^n \mathbb{E}_I[Y_j], \quad \mathbb{E}_I[Y_j] = \mathbb{E}_I[y_{I_j}] = \sum_{k=1}^N P(I_j = k) y_k = \mu \\
\mathrm{Var}_I(Y_j) &= \mathrm{Var}_I(y_{I_j}) = \sum_k P(I_j = k)(y_k - \mu)^2 = \sigma^2 \\
\mathbb{E}_I[\hat{F}(y)] &= \frac{1}{n}\sum_{j=1}^n \mathbb{E}_I\left[I(y_{I_j} \le y)\right] = \frac{1}{n}\sum_{j=1}^n \sum_{k=1}^N P(I_j = k) I(y_k \le y) = \frac{1}{n}\sum_j F(y)
\end{aligned}$$

Definition 9.3 (*SRS with(out) replacement*). The sampling design pmfs are calculated by symmetry and counting the number of possible outcomes. They are EPS by symmetry.

$$\Pr_{\text{with}}(I) = N^{-n}, \quad \Pr_{\text{w/o}}(I) = \left(\frac{N!}{(N-n)!}\right)^{-1}$$

Theorem 9.2 (*variance and asymptotics with(out) replacement*). Results for replacement follow from i.i.d.

$$\mathrm{Var}_{\text{with}}(\bar{Y}) = \frac{\sigma^2}{n}, \quad \mathrm{Var}_{\text{with}}(\hat{F}(y)) = \frac{F(y)(1 - F(y))}{n}$$

The asymptotics follow from standard results

$$\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad \sqrt{n}(S^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, \mu_4 - \mu_2^2), \quad \sqrt{n}[\hat{F}(y) - F(y)] \xrightarrow{d} \mathcal{N}(0, F(y)(1 - F(y)))$$

Using sampling without replacement,

$$\mathrm{Cov}_{\mathrm{w/o}}(Y_j, Y_{k\neq j}) = -\frac{\sigma^2}{N-1}, \quad \mathrm{Var}_{\mathrm{w/o}}(\bar{Y}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

Proof: To compute the covariance, using $\mathbb{E}_{\mathrm{w/o}}[Y_2|Y_1] = \dfrac{1}{N-1}(N\mu - Y_1)$, we have

$$\mathbb{E}_{\mathrm{w/o}}[Y_2Y_1] = \mathbb{E}[Y_1\mathbb{E}[Y_2|Y_1]] = \frac{1}{N-1}(N\mu^2 - \mathbb{E}[Y_1^2]) = \frac{1}{N-1}(N\mu^2 - \sigma^2 - \mu^2) = \mu^2 - \frac{\sigma^2}{N-1}$$

This yields $\mathrm{Cov}(Y_1, Y_2) = \mathbb{E}[Y_1Y_2] - \mathbb{E}[Y_1]\mathbb{E}[Y_2]$. Alternatively, using

$$Y_1 + \cdots + Y_N = y_1 + \cdots y_N$$

Taking the variance of both sides, the RHS is zero (constant), while the LHS equals

$$n\sigma^2 + 2\binom{n}{2}\mathrm{Cov}(Y_1, Y_2)$$

9.2 Stratified Sampling, Horvitz-Thompson estimator

Definition 9.4 (*strata*). Partition the IDs $1\cdots N$ into L strata with index l, so $i \mapsto (j, l)$ where $j \in [N_l]$ and $l \in [L]$. The population quantities can be fine-grained per strata

$$\mu_l = \frac{1}{N_l}\sum_j y_{jl}, \quad \sigma_l^2 = \frac{1}{N_l}\sum (y_{jl} - \mu_l)^2, \quad F_l(y) = \frac{1}{N_l}\sum_j I(y_{jl} \leq y)$$

They can be collected back to population quantities (equation for σ^2 echoes Eve's law)

$$\mu = \sum_l \frac{N_l}{N}\mu_l, \quad \sigma^2 = \sum_l \frac{N_l}{N}\sigma_l^2 + \sum_l \frac{N_l}{N}(\mu_l - \mu)^2, \quad F(y) = \sum_l \frac{N_l}{N}F_l(y)$$

Definition 9.5 (*stratified sampling*). A stratified sampling design fixes $(n_l \leq N_l)$ and chooses $I_{1:n_l,l}$ and is independent across strata. It is an equal probability stratified sample if $P(I_{jl} = k) = 1/N_l$ for all j, k, l. One may view stratified sampling from the bottom up as combining populations.

Definition 9.6 (*Horvitz-Thompson estimator*). Fixing the sampling design $P(I)$, let $C_j(I)$ count the number of times ID j shows up in I and the inclusion probability $\pi_j = P(C_j \geq 1)$. Then

$$\hat{\mu}_{\mathrm{HT}} = \frac{1}{N}\sum_{j=1}^N \frac{I(C_j \geq 1)}{\pi_j}y_j = \frac{1}{N}\sum_j \frac{y_j}{\pi_j}$$

THe second sum is understood to be over observed y_j.

Theorem 9.3 (*properties of HT-estimator*). With notation as above

$$\mathbb{E}[\hat{\mu}_{\mathrm{HT}}] = \mu$$

Proof: Direct computation

$$\mathbb{E}[\hat{\mu}_{\mathrm{HT}}] = \frac{1}{N}\sum_{j=1}^N \cancelto{\pi_j}{\frac{1}{\pi_j}\mathbb{E}[I(C_j \geq 1)]}y_j = \mu$$

Onto the variance, let π_{ij} denote $P(\{i, j\} \subset I) = \mathbb{E}[I(i \in I)I(j \in I)]$ (note that $\pi_{ii} = \pi_i$)

$$\mathrm{Var}(\hat{\mu}_{\mathrm{HT}}) = \mathbb{E}[\hat{\mu}_{\mathrm{HT}}^2] - \mu^2 = \frac{1}{N^2}\sum_{i,j=1}^N \left(\frac{\mathbb{E}[I(i \in I)I(j \in I)]}{\pi_i\pi_j}y_iy_j - y_iy_j\right) = \frac{1}{N^2}\sum_{i,j=1}^N \left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)y_iy_j$$

Example 9.1 (*inclusion probabilities for SRS*). Without replacement, $\pi_j = n/N$ and $\pi_{j\neq i} = \binom{N-2}{n-2}/\binom{N}{n}$. With replacement, $\pi_j = 1 - ((N-1)/N)^n, \pi_{i,j\neq i} = \pi_i \pi_j$.

9.3 Bootstrapping and Permutation tests

Definition 9.7 (*(nonparametric) bootstrap*). Given i.i.d. $\mathbf{y} \sim F$ with n samples, create $\mathbf{Y}^*$ via SRS with replacement from $\mathbf{y}$. Both $\mathbf{y}, \mathbf{Y}^*$ contain n samples. The resampled $\mathbf{Y}^*$ is the bootstrap sample. We can similarly create bootstrap samples of statistics (estimates) $t_1^* \cdots t_B^* = t(\mathbf{y}_1^*), \cdots, t(\mathbf{y}_B^*)$. The core justification of bootstrapping is

$$\hat{F}(y) = \frac{1}{n}\sum_{j=1}^{n} I(y_j \leq y) \to F(y) \tag{9.1}$$

Generating n i.i.d draws from the empirical CDF is equivalent to drawing n SRS with replacement. Bootstrap statistical quantities are (randomness over resampling)

$$\text{Bias}_{\text{boot}}(\hat{\theta}^*) = \mathbb{E}_{\text{boot}}[\hat{\theta}^*] - \hat{\theta}, \quad \text{Var}_{\text{boot}}(\hat{\theta}^*) = \mathbb{E}_{\text{boot}}[(\hat{\theta}^* - \mathbb{E}_{\text{boot}}[\hat{\theta}^*])^2], \quad \text{MSE}_{\text{boot}}(\hat{\theta}^*) = \mathbb{E}_{\text{boot}}[(\hat{\theta}^* - \hat{\theta})^2]$$

This is analogous to original definitions with sampling randomness replaced with resampling randomness, and with ground truth estimand θ_0 replaced with the (MLE) estimate $\hat{\theta}$. The bootstrap *approximation of the standard error* is

$$\text{SE}_{\text{boot}}(\hat{\theta}^*)^2 = \mathbb{E}_{\text{boot}}[(\hat{\theta}^* - \mathbb{E}_{\text{boot}}[\hat{\theta}^*])^2]$$

This can be estimated (to arbitrary accuracy, by increasing B)

$$\widehat{\text{SE}}_{\text{boot}}(\hat{\theta}^*)^2 = \frac{1}{B-1}\sum_{b=1}^{b}\left(\hat{\theta}_b^* - \bar{\theta}^*\right)^2$$

The sources of error are

- Difference between $\text{SE}(\hat{\theta})$ and $\text{SE}_{\text{boot}}(\hat{\theta}^*)$ related to 9.1. This decreases by increasing n.
- Difference between $\text{SE}_{\text{boot}}(\hat{\theta}^*)$ and $\widehat{\text{SE}}_{\text{boot}}(\hat{\theta}^*)$. This decreases with B.

Definition 9.8 (*bootstrap confidence intervals*). Three methods

- *Normal interval*: replace analytic $\text{SE}(\hat{\theta})$ with $\widehat{\text{SE}}_{\text{boot}}(\hat{\theta}^*)$. Simply avoids calculation.
- *Percentile method*: $[\hat{\theta}^*_{\lceil 0.025B\rceil}, \hat{\theta}^*_{\lceil 0.975B\rceil}]$. Here $\mathbf{Y}$-dependence is implicit through the distribution of $\hat{\theta}^*$.
- *Bootstrap t-interval*: Bootstrap $\mathbf{y}^*_{1\leq j\leq B}$ from $\mathbf{y}$ and compute $\hat{\theta}^*_{1\leq j\leq B}$. Build pivot on

  $$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{\text{SE}_{\text{boot}}(\hat{\theta}^*)}$$

  For each j, bootstrap $\mathbf{y}^{**}_{1\leq j,k\leq B}$ to compute $\widehat{\text{SE}}_{\text{boot}}(\hat{\theta}_j^*)^2$, treating $\mathbf{y}_j^*$ as the primary data. Use

  $$[\hat{\theta} - \hat{Q}_T^*(0.975)\widehat{\text{SE}}_{\text{boot}}(\hat{\theta}^*), \hat{\theta} - \hat{Q}^*(0.025)\widehat{\text{SE}}_{\text{boot}}(\hat{\theta}_j^*)]$$

Definition 9.9 (*permutation test*). Observing $X_1 \cdots X_m \sim F_X, Y_1 \cdots Y_n \sim F_Y$ all i.i.d (inter and intra-group), we wish to test $H_0 : F_X = F_Y$ using a test statistic T such that large values of T constitute evidence against H_0 e.g. $(\bar{Y} - \bar{X})^2$. Compute $t_0 = T((X_j), (Y_k))$. Repeating B times, shuffle X, Y (both inter and intra-group) and compute t_j. Permutation does not matter under the null. Use the p-value

$$P(T \geq t_0) \approx \frac{1}{B}\sum_{j=1}^{B} I(t_j \geq t_0)$$

10 Causal Inference

10.1 Treatment framework

Definition 10.1 (*framework definitions*). Consider n units with binary treatment.

- Define the *assignment vector* $\mathbf{W} = (W_j) \in \{0,1\}^n$.
- The r.v. $Y_j(\mathbf{W})$ is the *potential outcome* for unit j. Each unit has 2^n potential outcomes.
- The *treatment effect* on unit j of the treatment change $\mathbf{W} \to \mathbf{W}'$ is $\tau_j = Y_j(\mathbf{W}') - Y_j(\mathbf{W})$.

Definition 10.2 (*SUTVA*). The Stable Unit Treatment Value Assumption (SUTVA).

- *Non-interference*: Y_j is only dependent upon W_j and not $W_{k \neq j}$.
- *Homogeneity*: treatment is effectively the same across all units.

Under SUTVA. The treatment effect $\tau_j = Y_j(1) - Y_j(0)$. The *average treatment effect* is $\bar{\tau} = n^{-1} \sum \tau_j$. Treatment effects are homogeneous (resp. heterogeneous) if $\tau_j = \bar{\tau}$ (resp. $\neq$). We also have the switching equation $Y_j = W_j Y_j(1) + (1 - W_j) Y_j(0)$.

Definition 10.3 (*assignment mechanism, randomization, RCT*). The assignment mechanism is the joint pmf of the assignments given potential outcomes.

$$P(\mathbf{W} = \mathbf{w} | \mathbf{Y}(0), \mathbf{Y}(1)) = P((w_j) | Y_1(0), \cdots, Y_n(0), Y_1(1), \cdots, Y_n(1))$$

The assignments are *randomized* if assignments are independent of potential outcomes

$$P(\mathbf{w} | \mathbf{Y}(0), \mathbf{Y}(1)) = P(\mathbf{w}) \iff \mathbf{W} \perp\!\!\!\perp \{\mathbf{Y}(0), \mathbf{Y}(1)\}$$

An experiment with randomized assignments is a randomized control trial (RCT).

Remark 10.1. The observed data of a treatment experiment is $\{(w_i, y_i(w_i))\}$. The key problem of causal inference is that $\tau_j = Y_j(1) - Y_j(0)$ requires two values, only one of which can ever be observed (i.e. $Y_j(k) | W_j = k$). Also note that $\bar{\tau}$ is a sample-specific quantity, while $\mathbb{E}[\tau_1]$ is a population quantity.

10.2 Population-based inference

The central estimand is $\mathbb{E}[\tau_1]$. We model the joint distribution (usually assuming i.i.d.)

$$\{W_j, Y_j(0), Y_j(1)\}$$

Using the randomization condition in the last step

$$\mathbb{E}[\tau_1] = \mathbb{E}[Y_1(1)] - \mathbb{E}[Y_1(0)] = \mathbb{E}[Y_1(1) | W_1 = 1] - \mathbb{E}[Y_1(0) | W_1 = 0] \equiv \theta_1 - \theta_0$$

Estimating $\theta_0 = \mathbb{E}[Y_1 | W = 0], \theta_1 = \mathbb{E}[Y_1 | W = 1]$ is straightforward when i.i.d.

Theorem 10.1 (*randomized experiment MLE*). via the setting above, using the Bernoulli trial result yields the MLE estimates which are just sample means of treatment groups.

$$\hat{\theta}_0 = \frac{\sum Y_j (1 - w_j)}{\sum 1 - w_j}, \quad \hat{\theta}_1 = \frac{\sum Y_j w_j}{\sum w_j}, \quad \widehat{\mathbb{E}[\tau_1]} = \hat{\theta}_1 - \hat{\theta}_0$$

Other MLE results follow: note that all expectations and variances are conditional on $\mathbf{W}=\mathbf{w}$ since our inference framework is $(\mathbf{Y}|\mathbf{w}), \theta_0, \theta_1$.

$$\mathbb{E}[\hat{\theta}_i] = \theta_i, \quad \mathrm{Var}(\hat{\theta}_0) = \frac{\theta_0(1-\theta_0)}{\sum 1 - w_j}, \quad \mathrm{Var}(\hat{\theta}_1) = \frac{\theta_1(1-\theta_1)}{\sum w_j}$$

The CRLB is always saturated. Results for τ_1 follow from conditional independence of $\hat{\theta}_0, \hat{\theta}_1$ given $\mathbf{W}=\mathbf{w}$.

$$\mathbb{E}[\hat{\theta}_1 - \hat{\theta}_0] = \tau_1, \quad \mathrm{Var}(\hat{\tau}_1) = \mathrm{Var}(\hat{\theta}_0) + \mathrm{Var}(\hat{\theta}_1)$$

Proof: Treating θ_0, θ_1 as population parameters, factor

$$\Pr(y_1, w_1) = \Pr(y_1|w_1)\Pr(w_1)$$

Make inference conditional on $\mathbf{w}$, then

$$\log L(\theta_0,\theta_1) = \sum_{w_j=0} \log \theta_0^{y_j}(1-\theta_0)^{1-y_j} + \sum_{w_j=1} \log \theta_1^{y_j}(1-\theta_1)^{1-y_j} = \log L_0(\theta_0) + \log L_1(\theta_1)$$

$$\log L_0(\theta_0) = \sum (1-w_j)\log \theta_0^{y_j}(1-\theta_0)^{1-y_j}, \quad \log L_1(\theta_1) = \sum w_j \log \theta_1^{y_j}(1-\theta_1)^{1-y_j}$$

Recognizing this as the standard Bernoulli with counts given by $\sum w_j$ (for $w=1$) and $\sum(1-w_j)$ (for $w=0$), the MLEs follow. For τ_1, the likelihoods separate so $\hat{\theta}_0, \hat{\theta}_1$ are conditionally independent given $\mathbf{w}$.

Theorem 10.2 (*randomized experiment Bayesian inference*). Assuming θ_0, θ_1 conditionally independent given $\mathbf{w}$ with $\theta_j|\mathbf{w} \sim \mathrm{Beta}(\alpha_j, \beta_j)$. The posteriors follow from standard bernoulli update

$$\theta_0|\mathbf{y},\mathbf{w} = \mathrm{Beta}\left(\alpha_0 + \sum_j y_j(1-w_j), \beta_0 + \sum_j (1-y_j)(1-w_j)\right) = \mathrm{Beta}(\alpha_0', \beta_0')$$

Similarly for θ_1, with $1-w_j \mapsto w_j, (\alpha,\beta)_0 \mapsto (\alpha,\beta)_1$. Then the posterior difference

$$\mathbb{E}[\theta_1 - \theta_0|\mathbf{y},\mathbf{w}] = \frac{\alpha_1'}{\alpha_1'+\beta_1'} - \frac{\alpha_0'}{\alpha_0'+\beta_0'}$$

Credible intervals can be performed in a straightforward way given the posterior.

10.3 Finite sample inference

The central estimand is the finite-sample average treatment effect

$$\bar{\tau} = \frac{1}{n}\sum_{j=1}^{n} y_j(1) - y_j(0)$$

We condition instead on the potential outcomes $\mathbf{Y}(1), \mathbf{Y}(0)$, treating them as fixed.

Remark 10.2. In population-based approach, we condition on (i.e. fix) $\mathbf{W}=\mathbf{w}$ and analyze randomness arising from $\mathbf{Y}(1), \mathbf{Y}(0)|\mathbf{W}$ to do inference on $\mathbb{E}[\tau_1]$. In the finite sample approach, we fix $\mathbf{Y}(1), \mathbf{Y}(0)$ and analyze randomness arising from $\mathbf{W}$ to do inference on $\bar{\tau}$.

Fixing the potential outcomes, the observed outcome is determined by the switching equation

$$Y_j = W_j Y_1(1) + (1-W_j)Y_j(0) \implies \Pr(\mathbf{y},\mathbf{w}|\mathbf{y}(1),\mathbf{y}(0)) = \Pr(\mathbf{w}|\mathbf{y}(1),\mathbf{y}(0))$$

Further simplifying to $\Pr(\mathbf{w})$ in a randomized setting.

Theorem 10.3 (*MoM inference of $\bar{\tau}$*). Assuming independence $\Pr(\mathbf{w}) = \prod \Pr(w_j)$. Consider

$$G_1 = \frac{W_1 Y_1}{\mathbb{E}[W_1]} - \frac{(1-W_1)Y_1}{\mathbb{E}[1-W_1]}$$

Its statistical properties are closely related to $\bar{\tau}$: onditioning on $\mathbf{y}(0), \mathbf{y}(1)$

$$\mathbb{E}[G_1] = y_1(1) - y_1(0), \quad \mathrm{Var}(G_1) = \frac{y_1(1)^2}{\mathbb{E}[W_1]} + \frac{y_1(0)^2}{1-\mathbb{E}[W_1]} - [y_1(1) - y_1(0)]^2$$

The corresponding MoM estimator is conditionally unbiased fixing $\mathbf{Y}(0), \mathbf{Y}(1)$

$$\hat{\tau}_{\text{MoM}} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{W_j Y_j}{\mathbb{E}[W_j]} - \frac{(1-W_j)Y_j}{\mathbb{E}[1-W_j]}\right), \quad \mathrm{Var}(\hat{\tau}_{\text{MoM}}) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{y_j(1)^2}{\mathbb{E}[W_j]} + \frac{y_j(0)^2}{1-\mathbb{E}[W_j]} - \tau_j^2\right) \tag{10.1}$$

A conservative estimate of the variance is

$$\widehat{\mathrm{Var}(\hat{\tau}_{\text{MoM}})} \leq \frac{1}{n^2}\sum_{j=1}^{n}\left[\frac{W_j Y_j^2}{\mathbb{E}[W_j]^2} + \frac{(1-W_j)Y_j^2}{(1-\mathbb{E}[W_j])^2}\right] = \hat{\lambda}^2 \tag{10.2}$$

Proof: Expand $Y_1 = W_1 Y_1(1) + (1-W_1)Y_1(0)$ and use binary property $W_1^2 = (1-W_1)^2 = 1, W_1(1-W_1) = 0$.

$$G_1 = \frac{W_1[W_1 Y_1(1) + (1-W_1)Y_1(0)]}{\mathbb{E}[W_1]} - \frac{(1-W_1)[W_1 Y_1(1) + (1-W_1)Y_1(0)]}{\mathbb{E}[1-W_1]} = \frac{W_1 Y_1(1)}{\mathbb{E}[W_1]} - \frac{(1-W_1)Y_1(0)}{\mathbb{E}[1-W_1]}$$

Taking the expectation yields the expectation formula. The variance formula follows from computing $\mathbb{E}[G_1^2]$ using these properties again.

Definition 10.4 (*Fisher, Neymann null*). The Fisher null is $H_0 : \forall j \in [n], \tau_j = 0$. The weaker Neymann null is $H_0 : \bar{\tau} = 0$.

Theorem 10.4 (*finite-sample testing in RCT*). The Fisher null implies $Y_j(1) = Y_j(0)$, which implies that reshuffling the assignment data (treatment or control) should have no additional effect (up to sampling randomization of $\mathbf{W}$) on $\bar{\tau}$. The randomization-based causal p-value is

$$\hat{p} = \frac{1}{B}\sum_{l=1}^{B} I(|\hat{\tau}_{\text{MoM}}(\mathbf{W}^l)| \geq |\hat{\tau}_{\text{MoM}}(\mathbf{w})|)$$

Operationally, draw B random samples of $\mathbf{W}^l$ based on the known sampling design for w, and compute $\hat{\tau}_{\text{MoM}}$ using the pairs $(W_1^l, Y_1), \cdots, (W_n^l, Y_n)$.

To test the Neyman null, use the conservative estimate of the variance 10.2.

$$\frac{\hat{\tau}_{\text{MoM}}}{\hat{\lambda}} \leq \frac{\hat{\tau}_{\text{MoM}}}{\widehat{\mathrm{Var}(\hat{\tau}_{\text{MoM}})}} \sim \mathcal{N}(0,1)$$

10.4 Observational Studies and Unconfoundedness

Definition 10.5 (*covariates*). A covariate is a pretreatment variable X whose values are determined before assignments (in observational studies, before the assignments naturally occur) such that the assignment probability $\Pr(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X})$ can depend on X.

Definition 10.6 (*unconfoundedness, popensity score, strongly ignorable*). Assignments are unconfounded (ignorable) if they are independent of potential outcomes given the conditioned covariates:

$$[\mathbf{W} \perp\!\!\!\perp \mathbf{Y}(0), \mathbf{Y}(1)]|\mathbf{X}$$

- The popensity score $\lambda(x) = \Pr(W_j = 1|X_j = x)$ is the conditional probability of receiving treatment.
- The overlap assumption is satisfied if $\forall x, \lambda(x) \in (0, 1)$.
- Assignments are strongly ignorable given $\mathbf{X}$ if ignorable and satisfies the overlap assumption.

Consider population-based analysis with (W_j, Y_j, X_j) i.i.d. with binary outcomes, then

$$\mathbb{E}[\tau_1] = \mathbb{E}[\theta_1(X)] - \mathbb{E}[\theta_0(X)], \quad \theta_j(X) = \mathbb{E}[Y_1|W_1 = j, X] = \Pr[Y_1|W_1 = j, X]$$

Theorem 10.5 (*covariate observational study MLE*). Parameterize $\theta_j(x)$ with parameters ψ_j so that

$$\begin{aligned} \log L(\psi_0, \psi_1) &= \log L_0(\psi_0) + \log L_1(\psi_1) \\ \log L_0(\psi_0) &= \sum (1 - w_j) \left[y_j \log \theta_0(x|\psi_0) + (1 - y_j) \log(1 - \theta_0(x|\psi_0))\right] \\ \log L_1(\psi_0) &= \sum w_j \left[y_j \log \theta_1(x|\psi_1) + (1 - y_j) \log(1 - \theta_1(x|\psi_1))\right] \end{aligned}$$

Compute MLEs $\hat{\theta}_0, \hat{\theta}_1$ to estimate the population average treatment effect through

$$\widetilde{\mathbb{E}[\tau_1]} = \mathbb{E}[\theta_1(X|\psi_1) - \theta_0(X|\psi_0)] = \int dx\, f_X(x) \left[\theta_1(x|\psi_1) - \theta_0(x|\psi_0)\right] \approx \frac{1}{n} \sum \theta_1(x_j|\hat{\psi_1}) - \theta_0(x_j|\hat{\psi_0})$$

Theorem 10.6 (*covariate observational study MoM*). Let $\lambda(x) = \mathbb{E}[W_1|x]$, then 10.1 becomes

$$\hat{\tau}_{\text{MoM}}(\mathbf{W}) = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{W_j Y_j}{\lambda(x_j)} - \frac{(1 - W_j) Y_j}{1 - \lambda(x_j)} \right)$$

This is unbiased, conditioning on $\mathbf{Y}(0), \mathbf{Y}(1)$ and the covariate X. The conditional variance is

$$\text{Var}[\hat{\tau}_{\text{MoM}}(\mathbf{W})] = \frac{1}{n^2} \sum_{j=1}^{n} \frac{y_j(1)^2}{\lambda(x_j)} + \frac{y_j(0)^2}{1 - \lambda(x_j)} - [y_j(1) - y_j(0)]^2$$

To estimate $\lambda(x_j)$, similarly parameterize via ψ so $\lambda(x_j|\psi)$ and estimate

$$\hat{\psi} = \text{argmax} \sum W_j \log \lambda(x_j|\psi) + (1 - W_j) \log(1 - \lambda(x_j|\psi))$$

Remark 10.3. Main idea: additionally condition on covariantes and estimate covariante-dependent quantities in the original estimates by further parameterization and MLE.