1 Probability Theory

Thm 1.3 (LTP): $A_1,...,A_2$ partition of S and $B \subset S$, then $P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i)$.
Thm 1.4(Bayes' rule): $A_1, A_2, \ldots$ partition of S, B any set. Then for each $i = 1,2,\ldots$, $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)}$.
Def. 1.5 (Sigma Algebra): Collection of subsets of S is a *sigma algebra* $\mathcal{B}$ if it satisfies: (1) $\emptyset \in \mathcal{B}$, (2) If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$, and (3) if $A_1, A_2, \ldots \in \mathcal{B}$ then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$.
Mutual independence $\Rightarrow$ *pairwise independence*, but not $\Leftarrow$.

2 Random Variables

Thm 2.5 (Jensen's inequalities): Suppose $g(x)$ convex, then $E(g(X)) \geq g(E(X))$ if existent. Strict unless X degenerate or g linear.
Def 2.10 (MGF): $X \sim F_X$, $t \in \mathbb{R}$. Then $M_X(t) = E(e^{tX})$ given it exists in some nbh of 0.
Thm 2.7: If $M_X(t)$ exists, then $E(X^n) = \frac{\partial^n}{\partial t^n} M_X(0)$.

3 Multivariate Distributions

Def 3.1: n-dimensional rvec is $f : S \to \mathbb{R}^n$.

3.1 Bivariate Random Vectors

Define probability functions on Borel sigma algebra of $\mathbb{R}^2$.
Need to assume $E(|g(X,Y)|) < \infty$.
Joint $\Rightarrow$ Marginal: $F_X(x) = lim_{y\to\infty} F_{X,Y}(x,y)$ and $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,v)dv$.

3.2 Continuous Distributions

Conditional Expectation: $E(g(Y)|X = x) = \sum_{y \in (Y)} g(y) f_{Y|X}(y|x)$ or $= \int_{-\infty}^{\infty} g(y) f_{Y|X}(y|x)dy$.
Thm 3.1 (LIE): Y, X rvs, then $E(Y) = E_X(E_{Y|X}(Y|X))$.
Law of iterated variance: $Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$.

3.3 Independence

Def 3.4: (X,Y) rvec, X,Y *independent* if $\forall x \in \mathbb{R}, y \in \mathbb{R}$ we have $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.
Thm 3.2: X,Y independent $\Leftrightarrow$ for any two bounded $g,h : \mathbb{R} \to \mathbb{R}$ we have $E(g(X)g(Y)) = E(g(X))E(h(Y))$.
Thm 3.3: X,Y independent, $g(X)$ and $g(Y)$ independent.

4 Sampling

4.1 Distribution of the t-ratio

With $\{X_i\}_{i=1}^{\infty}$ rs of $X_i \sim N(\mu, \sigma^2)$ we have $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma/} \sim N(0,1)$. Then the *t-ratio* $\frac{\overline{X}_n - \mu}{\frac{1}{\sqrt{n}} S_n} = \frac{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{S_n^2/\sigma^2}} \sim t_{n-1}$. Invert to construct CI using t-quantiles.

5 Asymptotic Theory

5.1 Inequalities

Thm 5.1 (Markov's Inequality): X r.v., $g : \mathbb{R} \to [0,\infty)$, then $\forall \epsilon > 0$, $P(g(X) > \epsilon) \leq \frac{E(g(X))}{\epsilon}$.
Cor 5.1 (Chebyshev's Inequality): X r.v., then $\forall \epsilon > 0$, $P(|X - E(X)| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}$.

5.2 Modes of Convergence

Def 5.2: $plim_{n\to\infty} X_n = X \leftrightarrow \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$.
Def 5.4: $\{X_n\}_{n=1}^{\infty}$ converges in *distribution* to X $\leftrightarrow \lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for every continuity point of x of $F_X(\cdot)$.
Def 5.5: $\{X_n\}_{n=1}^{\infty}$ converges in *mean square* to $X \leftrightarrow \lim_{n\to\infty} E[(X_n - X)^2] = 0$.
Thm 5.2: $X_n \xrightarrow{m.s.} \Rightarrow X_n \xrightarrow{p} X$. Proof by Chebyshev's inequality. The reverse is not true, consider $X_n \in 0, \sqrt{n}$ with probabilities $1 - 1/n, 1/n$.
Thm 5.3: $X_n \xrightarrow{p} \Rightarrow X_n \xrightarrow{d} X$. Proof uses definition of $\xrightarrow{p}$ and continuity. The reverse is generally *not true*, consider $X_n = Z \sim N(0,1)$ and $X, Z \sim N(0,1)$, have $F_{X_n}(x) = F_X(x)$ but $P(|Z - X| \geq \epsilon) > 0$. Exception: $X_n \xrightarrow{d} c \in \mathbb{R} \Rightarrow X_n \xrightarrow{p} c$.

5.3 Law of Large Numbers

Thm 5.6 (LLN i.i.d): $\{X_i\}_{i=1}^{\infty}$ seq. of iid rvs from F_X with $\mu = E(X)$ exist and finite. Then $\overline{X}_n \xrightarrow{p} \mu$.
Convergence Criteria: Need a combination of three assumptions: (1) finite mean and/or variance (no LLN for Cauchy), (2) bounds on asymptotic variance (e.g. not growing too fast with i), (3) restricted dependence.

5.4 Central Limit Theorem

Thm 5.7 (Lindeberg-Levy CLT): $\{X_i\}_{i=1}^{\infty}$ seq. of iid rvs from F_X, μ and σ^2 finite. Then $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$.
Thm 5.9 (Berry-Esseen): $\{X_i\}_{i=1}^{\infty}$ seq. of iid rvs from F_X, μ and σ^2 finite and $\lambda = E(|X - E(X)|^3)$ exist and finite. Let $Z \sim N(0,1)$. Then $|P\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq x\right) - P(Z \leq x)| \leq \frac{C\lambda}{\sigma^3 \sqrt{n}}$.

5.5 Convergence of Random Vectors

Def 5.7: $\mathbf{X}_n \xrightarrow{p} \mathbf{X} \leftrightarrow \lim_{n\infty} P(\|\mathbf{X}_n - \mathbf{X}\| < \epsilon) = 1$.
Def 5.8: $\mathbf{X}_n \xrightarrow{ms} \mathbf{X} \leftrightarrow \lim_{n\infty} E(\|\mathbf{X}_n - \mathbf{X}\|^2) = 0$.
Def 5.9: $\mathbf{X}_n \xrightarrow{d} \mathbf{X} \leftrightarrow \lim_{n\infty} F_{\mathbf{X}_n}(x) = F_{\mathbf{X}}(x)$ for every continuity point x of $F_{\mathbf{X}}(\cdot)$.
Thm 5.10 (Cramér-Wold): $\{\mathbf{X}_n\}_{n=1}^{\infty}$ seq. of K-dimensional random vectors. Then, $\forall \lambda \in \mathbb{R}^{\mathbb{K}}$ we have $\lambda' \mathbf{X}_n \xrightarrow{d} \lambda' \mathbf{X} \leftrightarrow \mathbf{X}_n \xrightarrow{d} \mathbf{X}$.

5.6 CMT and Slutzky's

Thm 5.11 (CMT): Let $\{\mathbf{X}_n\}_{n=1}^{\infty}$ be a sequence of K-dim. rvecs $\mathbf{X}$ K-dim rvec, and $g : \mathbb{R}^{\mathbb{K}} \to \mathbb{R}$ with discontinuity points D such that $P(\mathbf{X} \in D) = 0$.
(a) $\mathbf{X}_n \xrightarrow{p} \mathbf{X} \Rightarrow g(\mathbf{X}_n) \xrightarrow{p} g(\mathbf{X})$.
(b) $\mathbf{X}_n \xrightarrow{d} \mathbf{X} \Rightarrow g(\mathbf{X}_n) \xrightarrow{d} g(\mathbf{X})$.
Implication: Sums and products of convergent sequences converge. Does *not* hold for *mean square* convergence.
Thm 5.12 (Slutzky's): X_n, Y_n seq of rvs with $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c \in \mathbb{R}$, then $X_n + Y_n \xrightarrow{d} X + c$ and $X_n Y_n \xrightarrow{d} cX$, and if $c \neq 0$, $X_n / Y_n \xrightarrow{d} X/c$.
Extension to rvecs: $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{p} \mathbf{C} \in \mathbb{R}^{K\times K}$, $\mathbf{C}$ invertible, then $\mathbf{Y}_n^{-1}\mathbf{X}_n \xrightarrow{d} \mathbf{C}^{-1}\mathbf{X}$.
Example CMT: $\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}\right)^2 \xrightarrow{d} N(0,1)^2 = \chi_1^2$.
Thm 5.13 (Delta-Method): X_n seq of rvs with LL-CLT applying. $g : \mathbb{R} \to \mathbb{R}$ *continuously diff.* at μ with $g'(\mu) \neq 0$. Then $\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, g'(\mu)^2 \sigma^2)$. Proof: CMT and Slutzky's applied to Taylor's/intermediate value theorem.

5.7 Interval Estimation

Suppose $\{X_i\}_{i=1}^{n}$ is a seq of iid random variables with μ, σ^2 finite. Then an asymptotically valid CI for μ is given by $CI = \left[\overline{X}_n \pm \frac{z_{1-\alpha/2}}{\sqrt{n}} S_n\right]$ where S_n is a consistent estimator of σ and $P(\mu \in CI) \to 1 - \alpha$. Proof: CLT, CMT, Slutzky.

5.8 Moment-Based Estimation

Parameter of interest: $\theta = h(E(g(X)))$ (simple case: X, θ scalars and $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ cont. diff.).
Moment-based estimator: $\hat{\theta}_n = h\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right)$. *Consistency*: LLN and CMT.
Large-sample distribution: If $Var(g(X)) < \infty$ CLT applies so $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i) - E(g(X))\right) \xrightarrow{d} N(0, Var(g(X)))$. By the *delta-method* if $h'(g(E(X))) \neq 0$ we have $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, h'(E(g(X))^2 Var(g(X)))$.

6 Maximum Likelihood Estimation

Def 6.1 (likelihood function): $L_n(\theta) = \prod_{i=1}^{n} f(\mathbf{x}_i; \theta)$.
Equivalently, we define the *log-likelihood function* as $\log(L_n(\theta))$.
Thm 6.1: Suppose $\mathbf{X}$ is a random vector with pdf or pmf $f(\mathbf{x}; \theta_0)$. Then $E(\log(f(\mathbf{x}; \theta))) \geq E(ln(f(\mathbf{X}; \theta))), \forall \theta in \Theta$.
Thm 6.2: For $\tau(\theta)$ and $\hat{\theta}_n$ MLE of θ, we have $\tau(\hat{\theta}_n)$ is MLE of $\tau(\theta)$.

6.1 Distribution of the MLE

MLE Limit Distribution: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, A^{-1}BA^{-1})$ with $A = E_\theta[-\frac{\partial^2}{\partial\theta\partial'\theta} ln(f(\mathbf{X}_i; \theta)]$ and $B = E_\theta[\frac{\partial}{\partial\theta} ln(f(\mathbf{X}_i; \theta)) \frac{\partial}{\partial\theta'} ln(f(\mathbf{X}_i; \theta))]$. B *var-cov* matrix of the score. A is *Fisher information*.
Thm 6.3: Under weak reg. cond. (diff; interch. integr./diff.) we have $A = B$.

6.2 CRLB

We could try to define best estimator in terms of MSE. However, MSE might depend on θ (e.g. $\overline{X}_n$ vs. 1, the latter dominates for $\theta = 1$).
Progress: Focus on *unbiased* estimators.
Thm 6.4: $\{X_i\}$ rs from $f(x; \theta)$, $\hat{\theta}_n$ estimator of θ. Then under some reg conds $Var_\theta[\hat{\theta}_n] = \frac{(\frac{\partial}{\partial\theta} E_\theta[\hat{\theta}_n])^2}{nE_\theta[(\frac{\partial}{\partial\theta}\log(f(X;\theta)))^2]}$.
Relative efficiency: $E_\theta[(\hat{\theta}_{1,n} - \theta)^2] \leq E_\theta[(\hat{\theta}_{2,n} - \theta)^2]$ for all $\theta \in \Theta$ and strict for some.
Asymptotic efficiency: Asymptotic distribution often implies *asymptotically unbiased*, efficiency than means attaining CRLB asymptotically.

7 Hypothesis Testing

7.1 Basics

Def 7.1: A *hypothesis* is a statement about the population distribution.
Def 7.2: H_0 (null hypothesis) and H_1 (alternative hypothesis) are the complementary hypothesis. We write $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ with Θ_k mutually exclusive and exhaustive.
Simple hypothesis: Θ_0 is singleton. *Composite* hypothesis: Θ_1 more than one value.
Def 7.3: A *hypothesis test* is a rule when to reject H_0 (in favor of H_1) given the data. (*Accepting* H_0 is weird, e.g. what about $\theta_0 + \epsilon$?.)

7.2 Size and Power

T-I error: Reject H_0 although in fact true.
T-II error: *Not* reject H_0 although in fact false.
Error rates: *Probabilities* of making these errors (errors are random because they depend on the sample). Usually **trade-off** between I and II.
Def 7.4 (Power function): $\beta(\theta) = P_\theta(rejectH_0)$.
T-I error rate: $\beta(\theta)$ for any $\theta \in \Theta_0$.
T-II error rate: $\beta(\theta)$ for any $\theta \in \Theta_1$.
Def 7.5/7.6: For $\alpha \in [0,1]$, a test is *level* α if $\sum_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$ (*size*: equality).

7.3 Test statistics and critical values

Goal: Derive statistic T and reject iff $T > c_\alpha$ controlling $sup_{\theta\in\Theta_0} P_\theta(T > c_\alpha) \Rightarrow$ need $F_T(t)$.
Ex 7.2 (Two-sided T): $X \sim N(\mu, \sigma^2)$ so $\theta = (\mu, \sigma^2)$. Test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ use $T = \frac{\sqrt{n}|\overline{X}_n - \mu_0|}{S_n} |t_{n-1}|$ and reject for $T > c_\alpha = t_{n-1,1-\alpha/2}$. By construction $sup_{\theta\in\Theta_0} P_\theta(T > c_\alpha) = \alpha$. Note this holds for all $\sigma^2 \in \Gamma$ thus a test of level and size α.
Ex (One-sided T): $T = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n}$ with $c_\alpha = t_{n-1,1-\alpha}$ or $Z = -T$ and c_α unchanged (symmetry). Intuition: want to reject for large $\mu > \mu_0$.
Deriving $\beta(\theta)$: (1) add and subtract (true) μ, (2) look at behavior as μ changes.
Def (p-value): For any realization T^*, $p^* = inf\{p \in [0,1] : T^* > c_p\}$. Intuition: smallest α for which we would still reject.
Under H_0, $p \sim Unif[0,1]$ (require $P(p^* < \alpha) = \alpha$, i.e. want $Pr_\theta(rejectH_0) < \alpha$), but holds $\forall\alpha$.
p-value with simple H_0: If F_0 is strictly increasing, $p* = 1 - F_0(T*)$ (again: $p\ _{H_0}Unif[0,1]$).

7.4 Hypothesis Testing and CIs

Test-inversion: Assume test $H_0 : \theta = \theta_0$ (note: this is some H_0) and have test s.t. $P_{\theta_0}(rejectH_0) = \alpha$ (size α). Assume can perform for any $\theta_0 \in \Theta$. Then we have $CS = \{\theta_0 \in \Theta : notrejectH_0 : \theta = \theta_0\}$ with $P_\theta(\theta \in CS) = 1 - \alpha$ (*true* θ).
We can also do the reverse: From any CS with coverage rate $1 - \alpha$ can construct size α test as $reject \Leftrightarrow \theta_0 \notin CS$.
Ex. one-sided CI: Testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ (or $H_0 : \mu \leq \mu_0$) for normal case we have $CS = \{\mu_0 \in [\overline{X}_n - \frac{t_{1-\alpha,n-1}}{\sqrt{n}} S_n, \infty)\}$.

7.5 Asymptotic Approximations

Asymptotic argument: No parametric model $f_X(x; \theta)$, but, e.g., moments: $H_0 : E(X) = \mu$. $T \xrightarrow{d} |N(0,1)|$ and we can use $\Phi^{-1}(x)$ to control α asymptotically. In particular, $P(T > z_{1-\alpha/2}) \to \alpha$ under H_0.
Hypotheses: Set of distributions $\mathcal{P}$ with $\mathcal{P}_0 \subset \mathcal{P}$ set of distributions consistent with H_0.
Def 7.7 (Asymptotic power function): $\beta^\alpha(P) = lim_{n\to\infty} \beta_n(P)$.
Def 7.8/7.9: test with $\beta^a(P)$ is *asymptotic level* α if $sup_{P\in P_0} \beta^a(P) \leq \alpha$ (*size*: equality).
Def 7.10: Test *consistent* against alternative $P \in P_1$ if $\beta^a(P) = 1$.
Example: $\mathcal{P} = \{P : E(X), E(X^2) < \infty\}$ and $\mathcal{P}_0 = \{P : E(X) = 1\} \subset \mathcal{P}$ and $\mathcal{P}_1 : \{P : E(X) \neq 1\} \subset \mathcal{P}$.
Problem: $\beta^a(P)$ might not be informative about finite sample (e.g. $H_0 : \mu = \mu_0 + \epsilon$).

8 Regression, Causality, and Identification

8.1 Basics

Conditional mean function: $m(X) = E[Y|X]$.
Error: $e = Y - E[Y|X]$, which implies $Y = m(X) + e$
By definition: $E[e|X] = 0$.

8.2 Linear Regression Model

Linear m(X): $m(X) = X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k + \beta_{k+1}$
Linear means linear *in the parameters*.
Saturated model: Includes an indicator for each level of the regressor(s). In this case, the model *is* the CEF.

8.3 Causality

Outcome Y, observed regressors X, and *unobserved* variables U. Then $Y = g(X,U)$ where g is the *structural function*.
Causal effect: Change of X_i from x_0 to x_1 holding U_i fixed, i.e. $g(x_1, U_i) - g(x_0, U_i)$.
Marginal causal effect: $\frac{\partial}{\partial x} g(x_0, U_i)$ if x is scalar.
Heterogenous causal effects: Generally, effects depend on value of U_i, implying a *distribution* of causal effects.
Average marginal effects: $E[\frac{\partial}{\partial x} g(x_0, U)]$.
Average treatment effect: $E[g(x_1, U)] - E[g(x_0, U)]$ when x is discrete.
Average structural function: $ASF(x) = E[g(x, U)]$.
In general, $ASF(x) \neq E[Y|X = x] = E[g(x,U)|X = x]$ because X, U might not be independent.
Potential outcome: $Y(x) = g(x, U)$.

8.4 Causality in the Linear Model

Suppose scalar x, so $Y = g(X,U) = X\beta_1 + \beta_2 + U$.
Then β_1 is the marginal and average treatment effect. Does *not* require mean independence.
But: $m(X)$ might have a different slope coefficient. Suppose $E[U|X] = X\gamma_1 + \gamma_2$ implying $E[Y|X] = X(\beta_1 + \gamma_1) + (\beta_2 + \gamma_2)$.

Three assumptions required:
(1) Linearity: Marginal effect does not depend on value of x.
(2) Homogeneity/additivity: Marginal effect does not depend on U; X, U enter additively.
(3) Exogeneity: Mean of U independent of X, $E[U|X] = c = 0$.
Relaxation: (1) Interaction terms, polynomials, (3) additional regressors/IV/panel.
For (2): introduce individual slope coefficients $\beta_{1,i}$. Could try to estimate distribution (difficult), never individual slopes. Different target: $E[\beta_{1,i}]$.
Assume $E[U_i|X_i] = 0$ and $E[\beta_{1,i}] = E[\beta_{1,i}]$, then $E[Y_i|X_i] = X_i E[\beta_{1,i}] + \beta_2$.

8.5 Identification

What does the joint distribution of observables tell about parameters?
Regression model: $Y = X_1\beta_1 + \ldots + X_{k-1}\beta_{k-1} + \beta_k + e = X'\beta + e$ with $E[e|X] = 0$.
Identification of β: $E[Xe] = 0 \Leftrightarrow E[X(Y - X'\beta)] = 0 \Leftrightarrow E[XY] = E[XX']\beta \Leftrightarrow \beta = E[XX']^{-1}E[XY]$, where $E[XX']$ needs *full rank*.
Underidentification: $E[XX']$ *not* full rank, $\exists \gamma \in \mathbb{R}, \gamma \neq 0$ s.t. $E[XX']\gamma = 0 \Rightarrow \gamma' E[XX']\gamma = 0 \Rightarrow E[(X'\gamma)^2] = 0$, i.e. $X'\gamma = 0$ with probability 1. Thus $\forall c \in \mathbb{R}$, $Y = X'\beta + e = X'(\beta + c\gamma) + e$.
Violations of full rank: X_k linear combination of other X_j.

Distributions

Normal: $E(X) = \mu$, $Var(X) = \sigma^2$. Sum of two independent Normals is Normal.
MVN: $\sim N(\mu, \Sigma)$. Any linear combinations are Normal. $(X, Y) \sim N(\mu, \Sigma)$, then $X \perp\!\!\!\perp Y \Leftrightarrow Cov(X, Y) = 0$.
Uniform: $X \sim Unif(a, b)$, $F_X(x) = \frac{x-a}{b-a}$, $f_X(x) = \frac{1}{b-a}$, $E(X) = \frac{1}{2(b-a)}$, $Var(X) = \frac{1}{12}(b-a)^2$. $\hat{b}_{MLE} = max\{X_1, \ldots, X_n\}$ (min for a); $\hat{b}_{MM} = 2\overline{X}_n$.
Uniform Order Statistics: $U_{(k)} \sim Beta(k, n+1-k)$ with $E(U_{(k)}) = \frac{k}{n+1}$. **Exponential**: $X \sim Expo(\theta)$ then $E(X^k) = k!\theta^k$, so $E(X) = \theta$ and $Var(X) = \theta^2$. $\hat{\theta}_{MLE} = \overline{X}_n$.
Pareto: $X \sim Pareto(\alpha)$ then $E(X^k)$ only exists if $\alpha > k$. Given that, $E(X) = \frac{\alpha}{\alpha-1}$ and $Var(X) = \frac{\alpha}{(1-\alpha)^2(\alpha-2)}$. $Y = log(X)\ Expo(\alpha)$. 20-80 rule: $\alpha = \frac{\ln 5}{\ln 4} \approx 1.16$.
t: If $Y \sim N(0,1)$ and $Z \sim \chi^2_{n-1}$ and $X \perp\!\!\!\perp Z$ then $\frac{N}{Z}\ t_{n-1}$.
Cauchy: $X, Y \sim N(0,1)$ with $X \perp\!\!\!\perp Y$, then $\frac{X}{Y}\ Cauchy(0,1)$. Expectation and variance undefined. $X \sim Cauchy(0,1)$ then $X \sim t_1$.

9 Least Squares

9.1 Interpretations

Y scalar outcome, $X \in \mathbb{R}^k$, with $E[XX']$ full rank, then $\beta = E[XX']E[XY]$. We can write $Y = X'\beta + e$ with $e = Y - X'\beta$. **(1) Slope of conditional mean**: $Y = X'\beta + e$, $E[e|X] = 0$. β purely *descriptive*.
(2) Marginal causal effect: U scalar r.v. Causal relationship with structural function $Y = X'\beta + U$. Then β is *marginal causal effect*. Requires: linearity, homogeneity/additivity. Under exogeneity ($E[U|X] = 0$) we have $E[Y|X] = X'\beta$ and β is identified under full rank.
(3) Average causal effect: U scalar r.v., but $B \in \mathbb{R}^k$ random vector. Model causal relationship $Y = X'B + U$. Marginal causal effect B is random. Assume $E[B|X] = E[B] = \beta$ and $E[U|X] = 0$. Then β is *average causal effect*. $E[Y|X] = X'\beta$ and identified under full rank.
(4) Best linear approximation: Suppose $m(X) = E[Y|X]$ may be non-linear, want best linear approximation. Solve $\min_{b \in \mathbb{R}^k} E[(m(X) - X'b)^2]$. Can show solution coincides with BLP of Y given X solving $\min_{b \in \mathbb{R}^k} E[(Y - X'b)^2]$. Solution: $b* = \beta = E[XX']^{-1}E[XY]$ (given full rank). **(5) Projection**: Let $\beta = E[XX']^{-1}E[XY]$ and define $e = Y - X'\beta$. If $m(X)$ non-linear, $E[e|X]$ may depend on X. However, $E[Xe] = 0$ (FOC). Then $X'\beta$ is the projection of Y onto the space spanned by linear combinations of X, and β is the *projection coefficient* (under full rank).

9.2 Estimation: Sample analogues

$\beta = E[XX']^{-1}E[XY]$ under full rank. Replace expectations by sample analogues:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right).$$

Consistency follows from a LLN (with i.i.d. sample) and CMT.

9.3 Estimation: Least squares estimator

$\{Y_i, X_i\}_{i=1}^n$ random sample.
$\hat{\beta} := \arg\min_{b \in \mathbb{R}^k} \sum_{i=1}^n (Y_i - X_i'b)^2$.
The unique solution under full rank is the sample analogue estimator.
Fitted values: $\hat{Y}_i = X_i'\hat{\beta}$.
Residuals: $\hat{e} = Y_i - \hat{Y}_i$.
No full rank: No unique solution. Then $\exists c \in \mathbb{R}^k$ s.t. $X_i'c = 0 \forall i$, $c \neq 0$. **Interpretation**: $Y_i = X_i'\beta + \alpha + e_i$. Then

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\frac{1}{n}(X_i - \overline{X}_n)^2}$$

i.e. sample covariance over variance.

9.4 Algebraic Properties

Projection: Decomposition of Y into $X'\beta$ and $Y - X'\beta$ with $E[X'\beta(Y - X'\beta)] = 0$. Analogues results for sample version.
Residuals orthogonal to regressors: $\sum_{i=1}^n X_i\hat{e}_i = 0$ (by FOCs).
Residuals some to zero: If X_i contains a constant, we have $\sum_{i=1}^n \hat{e}_i = 0$ (by FOC). **Residuals orthogonal to fitted values**: $\sum_{i=1}^n \hat{Y}_i\hat{e}_i = 0$ (by FOCs).
Thus can *decompose* Y into two *orthogonal* components.
R-squared: $\sum_{i=1}^n (Y_i - \overline{Y}_n)^2 = \sum_{i=1}^n (\hat{Y}_i - \overline{\hat{Y}}_i)^2 + \sum_{i=1}^n \hat{e}_i^2$, so sample variance is fitted variance plus residual variance. $R^2 :=$ fitted variance over sample variance.
Drawback: increases trivially when adding regressors $\Rightarrow$ *Adjusted* R^2: $1 - \frac{n-1}{n-k}(1 - R^2)$.

9.5 Matrix Notation

Y and e are $n \times 1$ vectors, and X an $n \times k$ matrix. Then write $Y = X\beta + e$, a system of n equations.
Also, $\sum_{i=1}^n X_i Y_i = X'Y$ and $\sum_{i=1}^n X_i X_i' = X'X$.
OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$.
Decomposition: $Y = \hat{Y} + \hat{e} = X\hat{\beta} + (Y - \hat{Y})$.
Projection: $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = PY$ where P is projection matrix.
Intuition: For any $z \in R^n$, PZ returns $n \times 1$ vector that is linear combination of columns of X (i.e. regressors). P symmetric ($P' = P$) and idempotent ($PP = P$). Also, $PX = X$, thus for any $\gamma \in \mathbb{R}^k$, $PX\gamma = X\gamma$.
Annihilator: $Y = \hat{Y} + \hat{e} = PY + MY$ where $M = I_n - P$. M symmetric and idempotent and $MX = 0$, $MP = 0$.
Thus $\hat{Y}'\hat{e} = (PY)'(MY) = YPMY = 0$.
Useful property: $tr(P) = tr(X(X'X)^{-1}X') = tr(X'X(X'X)^{-1}) = tr(I_k) = k$.

9.6 Small Sample Properties