Višeparametarska linearna regresija te povezanost zakona o prekidu trudnoće i stope kriminaliteta u SAD

Ćosić, Dario

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Dario Ćosić

VIŠEPARAMETARSKA LINEARNA REGRESIJA TE POVEZANOST ZAKONA O PREKIDU TRUDNOĆE I STOPE KRIMINALITETA U SAD

Diplomski rad

Voditelj rada:
prof.dr.sc.Siniša Slijepčević

Zagreb, veljača 2025.

Ovaj diplomski rad obranjen je dana ____________________ pred ispitnim povjerenstvom u sastavu:

1. ____________________________, predsjednik

2. ____________________________, član

3. ____________________________, član

Povjerenstvo je rad ocijenilo ocjenom ________________.

Potpisi članova povjerenstva:

1. ____________________________

2. ____________________________

3. ____________________________

Sadržaj

Uvod

Tema prekida trudnoće u Sjedinjenim Američkim Državama (SAD) postala je izuzetno relevantna u posljednjih nekoliko godina, osobito nakon odluke Vrhovnog suda iz 2022. godine koja je poništila dugogodišnje pravo na pobačaj. Ova odluka omogućila je pojedinim državama da samostalno reguliraju pristup pobačaju, što je izazvalo široku javnu raspravu o posljedicama takvih zakona na društvo, uključujući i potencijalne utjecaje na stopu kriminaliteta. Istraživanja o utjecaju zakona o legalizaciji pobačaja na stopu kriminaliteta u SAD-u započela su još 1980-ih godina, a najviše utjecaja imao je rad "The Impact of Legalized Abortion on Crime", [13], objavljen 2001. godine od autora Stevena Levitta i John J. Donohuea. Prema tom istraživanju, postoji značajna povezanost između legalizacije pobačaja i smanjenja stope kriminala, posebice stope nasilnog kriminala, što sugerira da bi promjene u zakonodavstvu mogle imati dalekosežne posljedice. Ponovio sam istraživanje koristeći iste podatke i kod dostupan u replikacijskom paketu kojeg je objavio jedan od autora [4], što je rezultiralo istim nalazima kao u originalnoj studiji, potvrđujući pouzdanost i reproduktivnost rezultata. Višeparametarska linearna regresija predstavlja moćan alat za analizu ovakvih složenih odnosa. Ova metoda omogućava istraživačima da istovremeno analiziraju više varijabli poticaja i utvrde njihov odnos i jačinu djelovanja na varijablu ili varijable odziva. Prvi dio rada uvest će osnovne pojmove iz teorije vjerojatnosti i matematičke statistike koji su nužni za daljnji razvoj teorije koja će obuhvatiti višeparametarsku linearnu regresiju. U drugom dijelu rada fokusirat ćemo se na analizu podataka koji se odnose na zakone o prekidu trudnoće i stope kriminaliteta u različitim američkim državama. Cilj je utvrditi postoji li statistički značajna povezanost između liberalizacije zakona o pobačaju i smanjenja stope kriminala, koristeći podatke iz razdoblja nakon legalizacije pobačaja. Osim toga, istražit ćemo i druge varijable koje bi mogle utjecati na stopu kriminaliteta, kao što su ekonomski faktori, demografske karakteristike i obrazovni sustav. U kontekstu trenutnih promjena u zakonodavstvu o pobačaju u SAD-u, važno je razumjeti kako ove promjene mogu utjecati na društvo kao cjelinu. Ovaj rad će pružiti empirijsku analizu koja može poslužiti za buduće rasprave o pravima žena, etici pobačaja i njegovim socijalnim posljedicama.

Poglavlje 1

Teorija vjerojatnosti i matematička statistika

Kako bismo u potpunosti razumjeli višeparametarsku linearnu regresiju, potrebno je poznavanje osnovnih pojmova teorije vjerojatnosti i matematičke statistike. Ovi koncepti pružaju temelj za analizu podataka i interpretaciju rezultata. Na primjer, razumijevanje distribucije podataka, varijance i kovarijance ključno je za pravilno modeliranje i procjenu parametara. U ovom poglavlju ćemo se detaljnije zabaviti s ključnim pojmovima kao što su hipoteze testiranja, intervali pouzdanosti i procjena parametara. Razumijevanje ovih koncepata omogućava nam da procijenimo kvalitetu modela i njegovu sposobnost predikcije. Također ćemo istražiti kako različite pretpostavke o podacima utječu na regresijske rezultate.

1.1 Osnovni pojmovi teorije vjerojatnosti

Cilj regresijske analize je odrediti snagu i karakter odnosa između jedne ili više zavisnih varijabli (varijabli odziva, označenih s Y) te jedne ili više nezavisnih varijabli (varijabli poticaja, označenih s X), odnosno omogućiti predviđanje ili utvrđivanje uzročnih odnosa između nezavisnih i zavisnih varijabli. Promatrano obilježje najčešće modeliramo slučajnom varijablom, a pripadajuću vjerojatnosnu distribuciju nazivamo distribucijom te slučajne varijable. Ako istodobno promatramo više statističkih obilježja onda govorimo o slučajnom vektoru, čije su komponente slučajne varijable. Neka je sada $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor, gdje je Ω neprazan skup elementarnih događaja, $\mathcal{F}$ je σ-algebra na Ω i $\mathbb{P}$ je vjerojatnost na izmjerivom skupu $(\Omega, \mathcal{F})$. Nadalje, neka je $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ izmjeriv prostor sa σ-algebrom Borelovih skupova u $\mathbb{R}^k$ za $k \geq 1, k \in \mathbb{N}$.

Definicija 1.1.1. *Funkcija $X : \Omega \to \mathbb{R}^k$ je k-dimenzionalna slučajna varijabla ako je X izmjerivo preslikavanje u paru σ-algebri $(\mathcal{F}, \mathcal{B}(\mathbb{R}^k))$, tj. ako vrijedi*

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F} \text{ za sve } B \in \mathcal{B}(\mathbb{R}^k)$$

Kao što smo napomenuli, ako je $k = 1$, X zovemo slučajna varijabla dok za $k > 1$, X zovemo slučajni vektor. Definirajmo sada i vjerojatnosnu mjeru generiranu sa X te funkciju distribucije slučajne veličine X

Definicija 1.1.2. *Neka je X k-dimenzionalna slučajna veličina na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Vjerojatnosna mjera generirana s X je $\mathbb{P}_X : \mathcal{B}(\mathbb{R}^k) \to [0, 1]$ definirana s*

$$\mathbb{P}_X(B) := \mathbb{P}(X \in B), \quad B \in \mathcal{B}(\mathbb{R}^k)$$

Funkcija distribucije slučajne veličine X je $F_X : \mathbb{R}^k \to [0, 1]$ definirana relacijom

$$F_X(B) := \mathbb{P}_X(\langle -\infty, x]), \quad x \in \mathbb{R}^k$$

Slučajne varijable se mogu svrstati u dvije osnovne kategorije, diskretne i neprekidne. Diskretne slučajne varijable mogu poprimiti samo određene, izdvojene vrijednosti. Primjeri uključuju rezultate bacanja novčića ili igraće kocke, broj djece u obitelji ili broj automobila koji prođe kroz raskrižje u određenom periodu. Ove varijable se često prikazuju pomoću funkcije vjerojatnosti koja dodjeljuje vjerojatnost svakoj mogućoj vrijednosti koju varijabla poprima. Neprekidne slučajne varijable, s druge strane, mogu poprimiti bilo koju vrijednost unutar određenog intervala. Primjeri uključuju rezultate mjerenja visine, temperature ili vremena potrebnog za obavljanje nekog zadatka. Ove varijable se opisuju pomoću funkcije gustoće vjerojatnosti, koja omogućava izračunavanje vjerojatnosti da će varijabla poprimiti vrijednost unutar određenog raspona. Više primjera diskretnih i neprekidnih slučajnih varijabli zainteresirani čitatelj može pronaći u [20].

Definicija 1.1.3. *Slučajna veličina X dimenzije k je diskretna ako postoji skup $D \subseteq \mathbb{R}^k$ koji je prebrojiv i takav da je $\mathbb{P}_X(D) = 1$. Funkcija gustoće diskretne $k-$dimenzionalne slučajne veličine X je $f_X : \mathbb{R}^k \to \mathbb{R}$ definirana formulom*

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P}_X(\{x\})$$

Definicija 1.1.4. *Slučajna veličina X dimenzije k je neprekidna ako postoji nenegativna Borelova funkcija f_X definirana na $\mathbb{R}^k$ takva da za sve $x \in \mathbb{R}^k$ vrijedi*

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f_X(y)\, d\lambda(y)$$

pri čemu je λ Lebesgueova mjera definirana na izmjerivom prostoru $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. Funkciju f_X zovemo funkcija gustoće od X.

Uvedimo sada i pojmove matematičkog očekivanja i varijance. Matematičko očekivanje predstavlja prosječnu vrijednost koju bismo mogli očekivati kada ponavljamo eksperiment koji generira slučajnu varijablu X mnogo puta. Drugim riječima, to je "težinski" prosjek svih mogućih vrijednosti koje varijabla može poprimiti, pri čemu su težine vjerojatnosti tih vrijednosti. Matematičko očekivanje je generalizacija pojma srednje vrijednosti. Varijanca mjeri koliko se vrijednosti slučajne varijable raspršuju oko njenog matematičkog očekivanja. Drugim riječima, one pokazuju koliko su podaci "rašireni" ili "koncentrirani" oko prosječne vrijednosti. Visoka varijanca sugerira da su podaci široko raspršeni, dok niska varijanca ukazuje na to da su podaci bliže srednjoj vrijednosti.

Definicija 1.1.5. *Za slučajnu varijablu definiranu na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ kažemo da ima matematičko očekivanje ako $\int_\Omega |X(\omega)|\, d\mathbb{P}(\omega) < \infty$. U tom slučaju matematičko očekivanje je*

$$\mathrm{E}(X) := \int_\Omega X\, d\mathbb{P}$$

Uz varijancu, definiramo još i standardnu devijaciju, te kovarijancu i koeficijent korelacije.

Definicija 1.1.6. *Neka su X i Y slučajne varijable za koje vrijedi* $\mathrm{E}[X^2] < +\infty$ *i* $\mathrm{E}[Y^2] < +\infty$. *Varijanca od X je*

$$\mathrm{Var}(X) := \mathrm{E}[(X - \mathrm{E}[X])^2],$$

standardna devijacija od X je

$$std(X) := \sqrt{\mathrm{Var}(X)},$$

kovarijanca od X i Y je

$$\mathrm{Cov}(X, Y) := \mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)],$$

ako su $std(X) > 0$ i $std(Y) > 0$ koeficijent korelacije od X i Y je

$$Corr(X, Y) := \frac{\mathrm{Cov}(X, Y)}{std(X)std(Y)}.$$

1.2 Osnovni pojmovi matematičke statistike

U prirodi i društvu, mnogi fenomeni su višedimenzionalni i međusobno povezani. Proširenje vjerojatnosnog prostora na statističke strukture je nužan korak u analizi složenih fenomena stvarnog svijeta. U dinamičkom okruženju, gdje nesigurnost i kompleksnost postaju norma, ovaj pristup omogućuje bolje razumijevanje i predikciju, čime obogaćuje naše znanje i sposobnost djelovanja.

Definicija 1.2.1. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor i $\mathcal{P}$ množina vjerojatnosnih mjera definiranih na $(\Omega, \mathcal{F})$. Tada je uređena trojka $(\Omega, \mathcal{F}, \mathcal{P})$ statistička struktura.*

Množinu $\mathcal{P}$ uglavnom parametriziramo konačnodimenzionalnim parametrom θ.

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$$

gdje je $\Theta \subseteq \mathbb{R}^m$, $m \geq 1$, skup svih mogućih vrijednosti parametra θ koji zovemo parametarski prostor. Ako je $m = 1$, model zovemo jednoparametarskim dok za $m > 1$ govorimo o višeparametarskom modelu.

Definicija 1.2.2. *Slučajni uzorak duljine n je niz nezavisnih jednako distribuiranih slučajnih veličina $X_1, X_2, .., X_n$*

Slučajni uzorak je temeljni koncept u statistici koji se koristi za prikupljanje informacija o populaciji. To je način na koji možemo doći do reprezentativnog uzorka iz veće grupe kako bismo dobili uvid u njene karakteristike bez potrebe da proučavamo cijelu populaciju. Zamislite da želite saznati prosječnu visinu svih učenika u nekoj školi. Umjesto da mjerite visinu svakog učenika, što može biti dugotrajno i nepraktično, možete nasumično odabrati određeni broj učenika (slučajni uzorak) i mjeriti samo njih. Ako je uzorak pravilno odabran, rezultat će vam dati dobar uvid u prosječnu visinu svih učenika. Nezavisnost slučajnih varijabli znači da ishod jedne varijable ne utječe na ishod druge. Ovo je važno jer osigurava da svaki dio uzorka doprinosi informacijama bez pristranosti ili utjecaja drugih dijelova. Kao u primjeru, ako mjerimo visine učenika, visina jednog učenika ne bi trebala utjecati na visinu drugoga. Nezavisnost također omogućava primjenu statističkih metoda koje pretpostavljaju ovu karakteristiku. Jednaka distribucija znači da su sve varijable iz uzorka izvučene iz iste distribucije.

Definicija 1.2.3. *Statistika na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ je svaka slučajna veličina koja je izmjeriva funkcija slučajnog uzorka na toj statističkoj strukturi.*

Definicija 1.2.4. *Statistika $T = t(X_1, X_2, .., X_n)$ dimenzije $k \geq 1$ je dovoljna za θ ako uvjetna distribucija slučajnog uzorka $(X_1, X_2, .., X_n)$ uz uvjet $T = y$ ne ovisi o parametru θ za svako $y \in \mathbb{R}^k$ za koje postoji ta uvjetna distribucija.*

Kada imamo dovoljnu statistiku, možemo reći da su svi podaci potrebni za procjenu parametara sažeti u toj statistici, bez gubitka informacija. Dovoljna statistika stoga omogućava sažimanje velikog skupa podataka u manji skup koji i dalje sadrži sve potrebne informacije za procjenu određenog parametra. Sama definicija dovoljne statistike uglavnom nije korisna za pronalaženje dovoljne statistike ili za provjeru je li određena statistika dovoljna. Sljedeći teorem daje karakterizaciju dovoljne statistike.

Teorem 1.2.5 (Neyman-Fisher)**.** *Neka je $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak iz modela $\mathcal{P}$, pri čemu svaki X_i pripada prostoru $\mathbb{R}^d$ (tj. podaci su d-dimenzionalni). Neka je $T = t(\mathbf{X})$ statistika dimenzije $k \geq 1$. Tada je T dovoljna statistika za parametar θ ako i samo ako postoje nenegativne funkcije*

$$g_\theta : \mathbb{R}^k \to \mathbb{R} \quad i \quad h : \mathbb{R}^{n \cdot d} \to \mathbb{R},$$

takve da se zajednička gustoća slučajnog uzorka $f_X(x; \theta)$ može zapisati kao

$$f_X(x; \theta) = g_\theta(t(x))\, h(\mathbf{X}), \quad \text{za sve } x \in \mathbb{R}^{n \cdot d}.$$

Dokaz. Dokaz sprovodimo samo za slučaj diskretnog modela. Za neprekidni slučaj pogledajte u [15].

$\Rightarrow$: Neka je $T = t(\mathbf{X})$ dovoljna statistika za θ gdje je $t : \mathbb{R}^{dn} \to \mathbb{R}^k$ Borelova funkcija. Neka je $\theta \in \Theta$ po volji te neka su $x \in \mathbb{R}^{dn}$ i $y \in \mathbb{R}^k$ takvi da je $y = t(x)$ i $f_T(y; \theta) = \mathbb{P}_\theta(T = y) > 0$. Tada je

$$f_{\mathbf{X}}(x; \theta) = \mathbb{P}_\theta(\mathbf{X} = x) = \mathbb{P}_\theta(\mathbf{X} = x, T = y = t(x)) = \mathbb{P}_\theta(\mathbf{X} = x \mid T = t(x)) \cdot \mathbb{P}_\theta(T = t(x))$$

Budući da je T dovoljna za θ, $\mathbb{P}_\theta(\mathbf{X} = x \mid T = t(x)) =: h(x)$ ne ovisi o θ. Ako stavimo u desnu stranu gornje jednakosti da je $g_\theta(y) := \mathbb{P}_\theta(T = y)$, tada imamo da je

$$f_{\mathbf{X}}(x; \theta) = \mathbb{P}_\theta(\mathbf{X} = x \mid T = t(x)) \cdot \mathbb{P}_\theta(T = t(x)) = h(x) \cdot g_\theta(t(x)),$$

dakle, tvrdnja vrijedi.

$\Leftarrow$: Pretpostavimo da tvrdnja vrijedi za slučajni uzorak $\mathbf{X}$. Neka je $x \in \mathbb{R}^{dn}$ proizvoljno. Tada je za $\theta \in \Theta$:

$$\mathbb{P}_\theta(\mathbf{X} = x) = f_{\mathbf{X}}(x; \theta) = g_\theta(t(x))h(x).$$

Neka je $y \in \mathbb{R}^k$ bilo koja vrijednost takva da je $f_T(y;\theta) = \mathbb{P}_\theta(T = y) > 0$. Tada je:

$$\mathbb{P}_\theta(T = y) = \sum_{t(\xi)=y} \mathbb{P}_\theta(\mathbf{X} = \xi) \overset{(3.3)}{=} \sum_{t(\xi)=y} g_\theta(t(\xi))h(\xi) = g_\theta(y) \sum_{t(\xi)=y} h(\xi).$$

U gornjem izrazu se sumira po svim $\xi \in \mathbb{R}^{dn}$ takvima da je $t(\xi) = y$. Ako je x takav da je $y = t(x)$, tada je:

$$\mathbb{P}_\theta(\mathbf{X} = x \mid T = y) = \frac{\mathbb{P}_\theta(\mathbf{X} = x, T = y = t(x))}{\mathbb{P}_\theta(T = y)} = \frac{\mathbb{P}_\theta(\mathbf{X} = x)}{\mathbb{P}_\theta(T = y)} =$$

$$= \frac{g_\theta(t(x))h(x)}{g_\theta(y) \sum_{t(\xi)=y} h(\xi)} = \frac{h(x)}{\sum_{t(\xi)=y} h(\xi)}.$$

Ako je x takav da je $y \neq t(x)$ tada je:

$$\mathbb{P}_\theta(\mathbf{X} = x \mid T = y) = \frac{\mathbb{P}_\theta(\mathbf{X} = x, T = y)}{\mathbb{P}_\theta(T = y)} = \frac{\mathbb{P}_\theta(\emptyset)}{\mathbb{P}_\theta(T = y)} = 0.$$

Dakle,

$$f_{X|T}(x|y) = \mathbb{P}_\theta(\mathbf{X} = x \mid T = y) = \begin{cases} \frac{h(x)}{\sum_{t(\xi)=y} h(\xi)}, & t(x) = y \\ 0, & t(x) \neq y \end{cases}$$

ne ovisi o θ pa je T dovoljna statistika za θ. □

Funkcija gustoće za n-dimenzionalni slučajni uzorak X sa slučajnim veličinama dimenzije $d, d \geq 1$, iz modela $\mathcal{P}$ je dana sa:

$$f_X(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta), \quad x = (x_1, x_2, .., x_n) \in \mathbb{R}^{dn}$$

Uočimo da bijektivna i izmjeriva transformacija čuva svojstvo dovoljnosti statistike. Primijetimo uz to da je cijeli uzorak također dovoljna statistika. Cilj je od svih dovoljnih statistika izabrati onu koja najbolje reducira količinu podataka potrebnih za procjenu parametra populacije.

Definicija 1.2.6. *Dovoljna statistika T je minimalna dovoljna statistika za θ ako za svaku drugu dovoljnu statistiku S za θ postoji izmjeriva funkcija g takva da je $T = g(S)$*

Potpuna statistika je statistika koja sadrži sve informacije o parametru populacije i ne može se dodatno sažeti bez gubitka informacija. U ovom kontekstu, ako je T potpuna statistika za parametar θ, tada nijedna funkcija od T ne može biti nezavisna od θ osim ako je ta funkcija konstantna.

Definicija 1.2.7. *Statistika T za θ je potpuna statistika ako za svaku Borelovu funkciju g za koju vrijedi $(\forall\theta \in \Theta)\mathrm{E}_\theta[g(T)] = 0$, slijedi da je $(\forall\theta \in \Theta)\mathbb{P}_\theta[g(T) = 0] = 1$*

Svaka minimalna dovoljna statistika je također dovoljna, ali nije svaka dovoljna statistika minimalna. Potpune statistike često su povezane s minimalnim dovoljnim statistikama jer pružaju sve informacije nužne za procjenu, ali ne moraju nužno biti dovoljne. No, ako je neka statistika dovoljna i potpuna, pokazuje se da je onda i minimalna dovoljna. Za detalje vidjeti **Teorem 3.22** u [11]

Teorem 1.2.8. *Neka je T dovoljna i potpuna statistika za θ. Tada vrijedi da je T minimalna dovoljna statistika za θ*

1.3 Procjena parametara

Ovo poglavlje razmatra različite metode za rješavanje problema procjene parametara, s ciljem pronalaženja optimalnog točkovnog procjenitelja za populacijske distribucije. Ove metode uključuju pristupe kao što su metoda maksimalne vjerodostojnosti, metoda najmanjih kvadrata i bayesovska procjena, a svaki od njih nudi jedinstvene prednosti ovisno o prirodi podataka i pretpostavkama o distribuciji. Osim toga, važno je razumjeti kako različite statističke tehnike utječu na preciznost i pouzdanost procjena. Na primjer, izbor između točkovnih i intervalnih procjenitelja može značajno utjecati na interpretaciju rezultata i donošenje odluka. U konačnici, učinkovita procjena parametara ne samo da poboljšava naše razumijevanje populacije, već također omogućava bolje planiranje i strategije u različitim područjima, od ekonomije do zdravstvenih znanosti.
Neka je $X = (X_1, X_2, .., X_n)$ slučajni uzorak iz parametriziranog modela $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$, gdje su $X_i,\ i = 1, 2, .., n$, dimenzije $d \geq 1$, a funkcije gustoće su parametrizirane parametrom θ dimenzije $m \geq 1$. Neka je $\tau : \Theta \to \mathbb{R}^k$ funkcija čiju vrijednost $\tau(\theta)$ želimo procijeniti iz informacije sadržane u uzorku. Procijenitelj nepoznatog parametra $\tau(\theta)$ je slučajna veličina dimenzije k definirana kao funkcija slučajnog uzorka, odnosno kao bilo koja statistika $T = t(X)$ dimenzije k. Procijenitelj nije jedinstven, stoga uvijek želimo odabrati onaj koji je u nekom smislu optimalan. Neka je sada $x = (x_1, x_2, .., x_n)$ jedna realizacija slučajnog uzorka X. Da bismo procijenili koji procjenjitelj $T = t(X)$ je optimalan, definiramo funkciju gubitka $(t(x), \tau(\theta)) \to L(t(x), \tau(\theta))$. Funkcija gubitka je slučajna veličina, stoga možemo definirati funkciju rizika $\tau(\theta) \to R(\tau(\theta))$ kao $R(\tau(\theta)) = \mathrm{E}_\theta[L(T, \tau(\theta))]$. Vrijednost

funkcije rizika označava očekivani gubitak kada nepoznati parametar $\tau(\theta)$ procjenjujemo vrijednošću procjenjitelja T, to jest mjeri koliko je dobar procjenjitelj T za $\tau(\theta)$ kad je stvarna vrijednost parametra upravo θ.

Najčešća funkcija gubitka pri procjeni $\tau(\theta)$ vrijednošću procjenjitelja $T = t(X)$ za danu realizaciju uzorka x je kvadratna greška, $L(t(x), \tau(\theta)) = |t(x) - \tau(\theta)|^2$. S obzirom na to da ne znamo koja je stvarna vrijednost parametra modela, želimo pronaći onaj procjenitelj koji ima najmanju funkciju rizika za sve $\theta \in \Theta$.

Definicija 1.3.1. *Ako procjenitelj $T = t(X)$ za $\tau(\theta)$ zadovoljava uvjet*

$$(\forall \theta \in \Theta) \quad \mathrm{E}_\theta[T] = \tau(\theta)$$

tada kažemo da je T nepristran procjenitelj.

Uspoređujući procjenitelje pomoću funkcije rizika, nemamo garanciju da najbolji procjenitelj u tom smislu postoji. Međutim, može se pokazati da u klasi nepristranih procjenitelja postoji procjenitelj koji ima najmanju srednjekvadratnu grešku, ali nepristrani procjenitelj za općenitu funkciju $\tau(\theta)$ ne mora postojati. Za detalje upućujem zainteresiranog čitatelja na [17]. Ako postoji barem jedan nepristrani procjenitelj od $\tau(\theta)$, kažemo da je funkcija procjenjiva. Nadalje, kažemo da je statistika T definirana na statističkom modelu $\mathcal{P}$ konačne varijance ako je varijanca od T konačna za svaki $\theta \in \Theta$.

Definicija 1.3.2. *Neka je $\tau(\theta)$ procjenjiva funkcija. Statistika T je nepristrani procjenitelj uniformno minimalne varijance ili UMVUE procjenitelj od $\tau(\theta)$ ako vrijedi da je $T \in \mathcal{W}_\tau$ i*

$$(\forall S \in \mathcal{W}_\tau)(\forall \theta \in \Theta) \quad \mathrm{Var}_\theta T \leq \mathrm{Var}_\theta S$$

gdje je $\mathcal{W}_\tau$ množina svih nepristranih procjenitelja za $\tau(\theta)$ konačne varijance

Ponovimo, nepristranost procjenitelja znači da je očekivana vrijednost UMVUE-a jednaka pravoj vrijednosti parametra koji se procjenjuje, dok minimalna varijanca znači da među svim nepristranim procjeniteljima, UMVUE ima najmanju varijancu što ga čini optimalnim izborom za procjenu. Također, kako veličina uzorka raste, UMVUE se približava stvarnoj vrijednosti parametra, čime se povećava njegova pouzdanost. Postoji nekoliko metoda za pronalaženje UMVUE-a. Spomenimo Lehman-Schefféov teorem, koji kaže da ako je procjenitelj funkcija dovoljnih statistika te ako je nepristran, tada je taj procjenitelj UMVUE. Dovoljna statistika sadrži

sve informacije o parametru koji se procjenjuje. Također, moguće je koristiti Rao-Blackwellov teorem. Ova metoda uključuje uzimanje nepristranog procjenitelja i njegovu transformaciju u bolji procjenitelj korištenjem dovoljne statistike, tako da novi procjenitelj ima manju ili jednaku varijancu od originalnog. U praksi, često se koriste specifične formule za izračunavanje UMVUE-a za određene distribucije. Na primjer, za normalnu distribuciju s nepoznatim srednjim vrijednostima i varijancom, srednja vrijednost uzorka predstavlja UMVUE za srednju vrijednost populacije. Postoje i druge metode traženja procjenitelja minimalne varijance koji su korisni kad ne postoje potpune i dovoljne statistike ili ih je teško pronaći. Detaljnija pojašnjenja nalaze se u [24] i [12]. Navedimo još i procjenu metodom maksimalne vjerodostojnosti.
Neka je $X = (X_1, X_2, .., X_n)$ slučajan uzorak iz statističkog modela $\mathcal{P}$.
Ako je $x = (x_1, x_2, .., x_n) \in \mathbb{R}^{dn}$ jedna realizacija slučajnog uzorka X, tada je funkcija vjerodostojnosti $L : \Theta \to \mathbb{R}$ definirana s

$$L(\theta|x) := f_X(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta), \quad \theta \in \Theta$$

Sada možemo definirati

Definicija 1.3.3. *Statistika $\hat{\theta}$ je procjenitelj maksimalne vjerodostojnosti ako vrijedi*

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta|x)$$

Kada je veličina uzorka dovoljno velika, procjenitelji maksimalne vjerodostojnosti su nepristrani, što znači da njihova očekivana vrijednost konvergira prema pravoj vrijednosti parametra. ML procjenitelji imaju minimalnu varijancu među svim nepristranim procjeniteljima kada je uzorak dovoljno velik. To se zove Cramér-Raova donja ograda. Mogu se primijeniti na širok spektar statističkih modela i distribucija. Bez obzira na to je li distribucija normalna, binomna, Poissonova ili neka druga, ML metoda može se prilagoditi za procjenu parametara. ML procjenitelji su konzistentni, što znači da kako veličina uzorka raste, procjenitelj će konvergirati prema pravoj vrijednosti parametra.U mnogim praktičnim situacijama, ML procjenitelji daju vrlo dobre rezultate čak i s malim uzorcima, posebno kada su podaci blizu normalne distribucije. Međutim, iako su ML procjenitelji asimptotski nepristrani, mogu biti pristrani kada je veličina uzorka mala. To može dovesti do netočnih procjena u situacijama s ograničenim podacima. Maksimizacija funkcije vjerodostojnosti može biti složena ili čak nemoguća za određene modele ili distribucije, posebno kada uključuju više parametara ili neobične oblike funkcije vjerodostojnosti. Ako su pretpostavke o distribuciji podataka netočne (npr., ako podaci nisu neovisni ili identično distribuirani), ML procjenitelji mogu dati loše rezultate. Iako ML procjenitelji imaju minimalnu varijancu asimptotski, u praksi može biti teško odrediti varijancu procjenitelja bez dodatnih informacija ili metoda (npr., bootstrap metoda)

1.4 Testovi statističkih hipoteza

Testiranje statističkih hipoteza je postupak koji omogućava istraživačima da donose zaključke o populaciji na temelju uzorka podataka. Ova metoda igra ključnu ulogu u analizi podataka i donošenju odluka u raznim disciplinama, uključujući društvene znanosti, biologiju, ekonomiju... U testiranju hipoteza istraživači postavljaju **statističku hipotezu**, koja predstavlja tvrdnju o nekom obilježju populacije. Ova hipoteza može biti **nulta hipoteza** H_0, koja predstavlja osnovnu tvrdnju koju želimo testirati, na primjer "nema razlike između srednjih vrijednosti dviju grupa". **Alternativna hipoteza** predstavlja alternativnu tvrdnju koju uzimamo da vrijedi ako odbacimo nultu hipotezu, na primjer "postoji razlika između srednjih vrijednosti dviju grupa". Jedna od ključnih značajki statističkog testa je da, kada ne odbacimo nultu hipotezu, ne možemo reći da je ona prihvaćena. Preciznije, to znači da test nije uspio otkriti značajna odstupanja od nje. Ovo može dovesti do potencijalnih pogrešaka:
a) Tip I pogreška: Odbacivanje nulte hipoteze kada je ona zapravo istinita,
b) Tip II pogreška: Ne odbacivanje nulte hipoteze kada ona nije istinita

Neka je ponovno $X = (X_1, X_2, .., X_n)$ slučajni uzorak iz parametarskog modela $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$, gdje su slučajne veličine dimenzije $d \geq 1$, a parametar θ dimenzije $m \geq 1$. Neka su Θ_0 i Θ_1 neprazni disjunktni skupovi koji čine particiju Θ. Pretpostavimo da želimo testirati sljedeće hipoteze:

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1 \tag{1.1}$$

Primijetimo da su postavljene hipoteze zapravo tvrdnje o distribuciji statističkog obilježja koje je predmet izučavanja. Kako bismo donijeli odluku o tome odbacujemo li nultu hipotezu ili ne, potrebno je konstruirati statistički test. Statistički test možemo promatrati kao funkciju koja svakoj realizaciji slučajnog uzorka pridružuje vrijednost 0 ili 1, gdje vrijednost 0 znači da ne odbacujemo nultu hipotezu, dok vrijednost testa 1 znači da odbacujemo nultu hipotezu u korist alternativne hipoteze.

Definicija 1.4.1. *Statistički test hipoteze H_0 u odnosu na H_1 je funkcija $\tau : \mathbb{R}^{dn} \to \{0, 1\}$.*

Definiramo kritično područje testa $C_\tau \subseteq \mathbb{R}^{dn}$ kao skup svih realizacija uzorka za koje se nulta hipoteza odbacuje u korist alternativne $C_\tau = \tau^{-1}(1) = \{x \in \mathbb{R}^{dn} : \tau(x) = 1\}$.
Nadalje, **jakost testa** tumači se kao vjerojatnost odbacivanja nulte hipoteze u korist alternativne kada je stvarni parametar jednak θ.

Definicija 1.4.2. *Jakost testa τ je funkcija $\gamma : \Theta \to [0,1]$ definirana sa*

$$\gamma_\tau(\theta) := \mathrm{E}_\theta[\tau(X)] = \mathbb{P}_\theta(X \in C_\tau), \theta \in \Theta$$

Za test hipoteza 1.1, ako uzmemo $\theta \in \Theta_0$, funkcija jakosti testa je zapravo vjerojatnost pogreške prve vrste. Uobičajeno je unaprijed zadati razinu značajnosti testa, označavamo je sa $\alpha \in \langle 0,1 \rangle$, pomoću koje ograničavamo vjerojatnosti pogreške prve vrste.

Definicija 1.4.3. *Značajnost testa je $\alpha_\tau := \sup_{\theta\in\Theta} \gamma_\tau(\theta)$*

Kažemo da test τ ima zadanu razinu značajnosti α ako mu je značajnost α_τ manja ili jednaka od α. Među svim testovima koji imaju razinu značajnosti α, želimo pronaći test s najmanjom vjerojatnošću pogreške druge vrste. Takav test ćemo zvati uniformno najjačim za zadanu razinu značajnosti.

Definicija 1.4.4. *Za statistički test τ nulte hipoteze $H_0 : \theta \in \Theta_0$ nasuprot alternativi $H_1 : \theta \in \Theta_1$ s razinom značajnosti α_τ kažemo da je uniformno najjači na razini značajnosti α ako je $\alpha_\tau \leq \alpha$ i za svaki drugi test τ' istih hipoteza takav da je $\alpha_{\tau'} \leq \alpha$ vrijedi:*

$$\gamma_{\tau'}(\theta) \leq \gamma_\tau(\theta) \quad \forall\theta \in \Theta_1$$

Ključni pristup nalaženja "dobrog" testa, tzv. *Neyman-Pearsonova teorija*, polazi od fiksne razine značajnosti α i konstruira test za koji vrijedi da je pogreška druge vrste β najmanja moguća za sve parametre specificirane alternativnom hiptezom H_1. Ključni rezultat u toj teoriji je Neyman-Pearsonova lema koja daje najbolji test (najmanje β uz fiksno α) u slučaju kada su obje hipoteze, nulhipoteza i alternativna, jednostavne hipoteze. Za zadanu razinu značajnosti, kritično se područje, a time i testna statistika, za najbolji test, odredi kao skup onih vrijednosti uzoraka za koje vrijedi da je omjer L_O/L_1 vjerodostojnosti L_0 uz H_0 i L_1 uz H_1, izraženih kao funkcije uzoraka, ograničen odozgo nekom konstantom. Ako je barem jedna od hipoteza H_0 i/ili H_1 složena, tada samo u specijalnim slučajevima, na primjer kod jednostranih hipoteza, postoji test koji je najbolji za sve parametre. U slučajevima kada najbolji test u smislu Neyman-Pearsonove teorije ne postoji, koristi se drugi pristup za nalaženje dobrih testova: metoda omjera vjerodostojnosti. Testovi dobiveni metodom omjera vjerodostojnosti su, na neki način, poopćenja testova dobivenih Neyman-Pearsonovim pristupom.

Umjesto na temelju kritičnog područja, odluke o odbacivanju ili ne odbacivanju nulte hipoteze možemo donijeti pomoću p-vrijednosti. P-vrijednost je vjerojatnost da se opazi dobiveni uzorak ili ekstremniji u slučaju da je nulta hipoteza istinita. Koristeći p-vrijednost donosimo odluku: ako je p-vrijednost manja ili jednaka od prethodno postavljene razine značajnosti α, odbacujemo nultu hipotezu u korist alternativne na razini značajnosti α, inače nultu hipotezu ne odbacujemo.

Poglavlje 2

Višeparametarska linearna regresija

Višeparametarska linearna regresija je statistička metoda koja se koristi za modeliranje odnosa između jedne zavisne varijable i više nezavisnih varijabli. Ova metoda omogućava istraživačima da identificiraju i kvantificiraju utjecaj različitih faktora na ishod, čime se olakšava donošenje informiranih odluka. Cilj višeparametarske linearne regresije je pronaći linearan model koji najbolje odgovara dostupnim podacima, što uključuje određivanje nepoznatih koeficijenata koji predstavljaju težinu svake od nezavisnih varijabli u predikciji zavisne varijable. Kvaliteta modela je od suštinske važnosti, jer dobar model omogućava precizno predviđanje vrijednosti zavisne varijable na temelju poznatih vrijednosti nezavisnih varijabli. Iako se često koristi kao empirijski model, o čemu govori [16], gdje stvarna funkcijska veza ostaje nepoznata, važno je napomenuti da specifikacija modela može biti izazovna. Ne postoji univerzalno pravilo za odabir prikladne specifikacije, što zahtijeva pažljivo razmatranje i analizu podataka kako bi se osiguralo da model adekvatno reflektira složenost odnosa među varijablama. O toj problematici vrlo detaljno piše [23].

2.1 Model višeparametarske linearne regresije

Neka su $x_1, x_2, .., x_k$ ulazne varijable. Nazivamo ih još i varijablama poticaja. Najčešće su one zadane, a Y se opaža. Neka je ϵ slučajna varijabla definirana na (Ω, $\mathcal{F}$, $\mathbb{P}$) takva da je $\mathrm{E}[\epsilon] = 0$ te $\mathrm{Var}[\epsilon] = \sigma^2 > 0$, te neka su $\beta_0, \beta_1, .., \beta_k$ konstantni realni brojevi. Linearni višeparametarski regresijski model s k prediktora definiramo sljedećom relacijom:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k + \epsilon \tag{2.1}$$

Parametar ϵ interpretiramo kao slučajnu grešku ili šum, a parametri $\beta_0, \beta_1, .., \beta_k$ predstavljaju, najčešće nepoznate, parametre modela. Slučajna greška uključuje druge faktore koji također utječu na varijablu odziva, a nisu uključeni u model.

Nadalje, neka je $n \in \mathbb{N}$ broj opažanja te neka je $x_{i1}, x_{i2}, .., x_{ik}$, $i = 1, .., n$ i-ta vrijednost ulaznih varijabli za koje je y_i pripadna vrijednost izlazne slučajne varijable Y_i. Uz to, neka su $\epsilon_i, i = 1, .., n$ nezavisne slučajne varijable s očekivanjem $\mathrm{E}[\epsilon_i] = 0$ te $\mathrm{Var}[\epsilon_i] = \sigma^2 > 0$. Tada je opći linearni regresijski model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + .. + \beta_k x_{ik} + \epsilon_i \quad i = 1, 2, .., n \tag{2.2}$$

Model 2.2 možemo zapisati i u matričnom obliku:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Vidimo da su $\mathbf{Y}$ i ϵ n-dimenzionalni slučajni vektori stupci koje zovemo vektor izlaznih podataka, odnosno vektor slučajne greške. $\mathbf{X}$ je $n \times (k + 1)$ matrica ulaznih podataka, a β je vektor stupac regresijskih koeficijenata. Pretpostavljamo da vrijedi $n > k$. Slijedi da uvedeni k-dimenzionalni model možemo zapisati u matričnom obliku:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{2.3}$$

2.2 Procjena parametara

Vrijednosti svih koeficijenata regresije β_i i vrijednost varijance σ^2 slučajne greške najčešće nam nisu poznate i njih ćemo procjenjivati iz slučajnog uzorka. U opisanom modelu višeparametarske linearne regresije, slučajni uzorak je
niz $(x_{i1}, x_{i2}, .., x_{ik}, Y_i)$, $\quad i = 1, .., n$, gdje su x_{ij} vrijednost j-te ulazne varijable u i-tom opažanju, a $Y_1, Y_2, .., Y_n$ međusobno nezavisne slučajne varijable za koje vrijedi:

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + .. + \beta_k x_{ik}, \quad \mathrm{Var}[Y_i] = \sigma^2, \quad i = 1, .., n$$

Metoda najmanjih kvadrata

Ova metoda minimizira sumu kvadriranih slučajnih grešaka, odnosno sumu kvadratnih odstupanja između stvarnih vrijednosti zavisne varijable i vrijednosti predviđenih modelom. To je najčešće korištena metoda za pronalaženje najbolje procjene $\hat{\beta}_o, \hat{\beta}_1, .., \hat{\beta}_k$ nepoznatih koeficijenata u linearnoj regresijskoj analizi. Dakle, funkciju definiranu sa:

$$S(\beta_0, \beta_1, .., \beta_k) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{2.4}$$

minimiziramo obzirom na $\beta_0, \beta_1, .., \beta_k$.

U modelu višeparametarske regresije prikladnije je funkciju definiranu sa 2.4 izraziti u matričnom zapisu:

$$S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\tau (\mathbf{Y} - \mathbf{X}\beta) \tag{2.5}$$

Neka je sada $\hat{\beta}$ procjena za nepoznati vektorski parametar β. Dakle, $\hat{\beta}$ je određena tako da vrijedi:

$$\min_{\beta} S(\beta) = S(\hat{\beta})$$

Propozicija 2.2.1. *Vrijedi:*

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

Dokaz. Procjene regresijskih koeficijenata dobivene metodom najmanjih kvadrata moraju biti stacionarne točke funkcije najmanjih kvadrata $S(\beta)$ što se iz 2.5 može pojednostavniti na jednakost:

$$\mathbf{X}^\tau \mathbf{X} \hat{\beta} = \mathbf{X}^\tau \mathbf{Y} \tag{2.6}$$

Jer je $\mathbf{X}$ neslučajna matrica dimenzije $n \times (k + 1)$ kojoj su stupci linearno nezavisni, te je matrica punog ranga, slijedi da je kvadratna matrica $\mathbf{X}^\tau \mathbf{X}$ dimenzije $k + 1$ punog ranga pa je regularna, odnosno postoji inverz. Sada iz 2.6 množenjem s lijeva inverzom od $\mathbf{X}^\tau \mathbf{X}$ dobivamo da vrijedi:

$$\hat{\beta} = (\mathbf{X}^\tau \mathbf{X})^{-1} \mathbf{X}^\tau \mathbf{Y} \tag{2.7}$$

$\square$

Nadalje, druga derivacija funkcije S je:

$$\frac{\partial S}{\partial^2 \beta} = 2\mathbf{X}^\tau \mathbf{X}$$

što je pozitivno definitna matrica neovisno o $\beta \in \mathbb{R}^{k+1}$. Prema tome, S je konveksna funkcija na skupu $\mathbb{R}^{k+1}$ pa je jedinstvena stacionarna točka ujedno i točka minimuma. Stoga je procjenjitelj metodom najmanjih kvadrata za nepoznate koeficijente linearne regresije upravo jednak 2.7.

Svojstva procjenitelja dobivenog metodom najmanjih kvadrata

Pretpostavili smo da su slučajne greške ϵ_i, $i = 1, .., n$ nezavisne slučajne varijable s očekivanjem $\mathrm{E}[\epsilon_i] = 0$ i varijancom $\mathrm{Var}[\epsilon_i] = \sigma^2 > 0$. Dakle, za slučajni vektor ϵ vrijedi:

$$\mathrm{E}[\epsilon] = 0 \tag{2.8}$$

$$\mathrm{Cov}(\epsilon) = \sigma^2 I \tag{2.9}$$

Propozicija 2.2.2. *Vektor* $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ *je linearan i nepristran procjenitelj.*

Dokaz. Za slučajnu varijablu $\hat{\beta}_i$ onda vrijedi:

$$\hat{\beta}_i = \sum_{j=1}^{n} c_{i,j} Y_j, \quad i = 1, .., p \tag{2.10}$$

pri čemu je $c_{i,j}$ element u i-tom retku i j-tom stupcu matrice $(\mathbf{X}^\tau \mathbf{X})^{-1} \mathbf{X}^\tau$. Dakle, slučajni vektor $\hat{\beta}$ je linearan procjenitelj. Nadalje:

$$\begin{aligned}
\mathrm{E}[\hat{\beta}] &= \mathrm{E}\left[(\mathbf{X}^\tau \mathbf{X})^{-1} \mathbf{X}^\tau \mathbf{Y}\right] \\
&= \mathrm{E}\left[(\mathbf{X}^\tau \mathbf{X})^{-1} \mathbf{X}^\tau (\mathbf{X}\beta + \epsilon)\right] \\
&= \mathrm{E}\left[(\mathbf{X}^\tau \mathbf{X})^{-1} \mathbf{X}^\tau \mathbf{X}\beta + (\mathbf{X}^\tau \mathbf{X})^{-1} \mathbf{X}^\tau \epsilon\right] = \beta
\end{aligned}$$

gdje zadnja jednakost slijedi zbog toga što je $\mathrm{E}[\epsilon] = 0$ i $(\mathbf{X}^\tau \mathbf{X})^{-1} \mathbf{X}^\tau \mathbf{X} = \mathbf{I}$.

Dakle, zaključujemo da je slučajni vektor $\hat{\beta}$ nepristrani procjenitelj za β. □

Nakon što smo pokazali da je procjenitelj metodom najmanjih kvadrata linearan i nepristran, zanima nas je li procjenitelj 2.10 najbolji, u smislu najmanje varijance, među svim linearnim nepristranim procjeniteljima za nepoznati parametar β_i. Sljedeći teorem nam kaže uz koje uvjete je procjenitelj dobiven metodom najmanjih kvadrata najbolji linearni nepristrani procjenitelj.

Teorem 2.2.3. *(Gauss-Markov)*
Neka je $\hat{\beta}$ procjenitelj dobiven metodom najmanjih kvadrata za parametre linearnog regresijskog modela i neka je $L : \mathbb{R}^{k+1} \to \mathbb{R}$ linearni funkcional parametara $L(\beta) = l^{\tau}\beta$. Pretpostavimo da za slučajne greške $\epsilon_i, i = 1, 2, .., n$ vrijede Gauss-Markovljevi uvjeti:

(i) $\mathrm{E}[\epsilon_i] = 0, \quad \forall i = 1, .., n$

(ii) $\mathrm{Var}[\epsilon_i] = \sigma^2, \quad \forall i = 1, .., n$

(iii) $\mathrm{Cov}[\epsilon_i, \epsilon_j] = 0 \quad \forall i \neq j$

Tada je statistika $\boldsymbol{T} = l^{\tau}\hat{\beta}$ najbolji linearni nepristrani procjenitelj za $\boldsymbol{L}(\beta)$

Dokaz. Statistika T je najbolji linearni nepristrani procjenitelj za $L(\beta)$ ako je linearna, nepristrana i u klasi svih nepristranih linearnih procjenitelja za $L(\beta)$ ima najmanju varijancu. Pokažimo prvo da je T linearan procjenitelj. Vrijedi:

$$T = l^\top\hat{\beta} = l^\top(X^\top X)^{-1}X^\top Y = c_0^\top Y,$$

gdje je $c_0 = X(X^\top X)^{-1}l \in \mathbb{R}^n$. Dakle, T je linearan procjenitelj. Zatim pokažimo da je T nepristran procjenitelj:

$$\mathbb{E}_\theta(T) = \mathbb{E}_\theta(l^\top\hat{\beta}) = l^\top\mathbb{E}_\theta(\hat{\beta}) = l^\top\beta,$$

što povlači da je T nepristran procjenitelj. Preostalo je pokazati da T ima najmanju varijancu u klasi svih nepristranih linearnih procjenitelja za $L(\beta)$. Neka je

$$U = c^\top Y$$

neki drugi nepristrani linearni procjenitelj za $L(\beta)$. Zbog nepristranosti U, vrijedi

$$l^\top\beta = \mathbb{E}_\theta(U) = \mathbb{E}_\theta(c^\top Y) = c^\top\mathbb{E}_\theta(Y) = c^\top X\beta.$$

Slijedi da je U nepristran ako i samo ako je $l^\top\beta = c^\top X\beta$, odnosno $l = X^\top c$. Računamo:

$$\begin{aligned}
\mathrm{Var}(U) - \mathrm{Var}(T) &= \mathrm{Var}(c^\top Y) - \mathrm{Var}(l^\top\hat{\beta}) = c^\top\mathrm{Cov}(Y)c - l^\top\mathrm{Cov}(\hat{\beta})l \\
&= \sigma^2\left(c^\top c - l^\top(X^\top X)^{-1}l\right) = \sigma^2\left(c^\top c - c^\top X(X^\top X)^{-1}X^\top c\right) \\
&= \sigma^2 c^\top\left(I - X(X^\top X)^{-1}X^\top\right)c = \sigma^2 c^\top Mc \geq 0,
\end{aligned}$$

pri čemu je $H = X(X^\top X)^{-1}X^\top$ ortogonalni projektor na $\mathcal{M}$, a $M = I - H$ ortogonalni projektor na ortogonalni komplement od $\mathcal{M}$. Budući da je M pozitivno semidefinitan operator, vrijedi $c^\top Mc \geq 0$. Dakle, vrijedi

$$\mathrm{Var}(T) \leq \mathrm{Var}(U).$$

□

Procjena varijance greške

Procjenitelj za nepoznati parametar σ^2 možemo izvesti iz sume kvadrata razlike između izmjerene i procijenjene vrijednosti izlazne varijable $\mathbf{Y}$. Neka su $x_1, x_2, .., x_k$ neslučajne ulazne varijable, te neka je $\hat{Y}$ procijenjena vrijednost izlazne varijable pa se prilagođeni regresijski model može zapisati kao

$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + .. + \hat{\beta}_k x_k \tag{2.11}$$

Prema tome, za vektor procijenjenih vrijednosti izlazne varijable $\hat{Y}$ vrijedi:

$$\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y} = H\mathbf{Y} \tag{2.12}$$

gdje je matrica H dimenzije $n \times n$ definirana sa $H = \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau$ ortogonalni projektor na potprostor od $\mathbb{R}^n$ razapet stupcima od X. Nadalje, s e_i označiti ćemo razliku između stvarne vrijednosti Y i procijenjene vrijednosti izlazne varijable $\hat{Y}$. Vektor $e = (e_1, e_2, .., e_n)$ zovemo vektor reziduala i pišemo $\mathbf{e} = \mathbf{Y} - \hat{Y}$. Pomoću 2.12 vektor e možemo izraziti kao:

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - H\mathbf{Y} = (I - H)\mathbf{Y} = M\mathbf{Y}$$

gdje je matrica $M = I - H$ ortogonalni projektor na ortogonalni komplement potprostora od $\mathbb{R}^n$ razapetim stupcima od X.
Definirajmo statistiku:

$$\hat{\sigma}^2 = \frac{Y^\tau M Y}{n - k - 1} \tag{2.13}$$

Propozicija 2.2.4. *Statistika $\hat{\sigma}^2$ definirana s 2.13 je nepristran procjenitelj za nepoznati parametar zajedničke varijance σ^2*

Dokaz. Vrijedi

$$\begin{aligned}
Y^T M Y &= (M(X\beta + \varepsilon), X\beta + \varepsilon) \\
&= (M\varepsilon, X\beta + \varepsilon) \\
&= (M\varepsilon, X\beta) + (M\varepsilon, \varepsilon) \\
&= (\varepsilon, MX\beta) + (M\varepsilon, \varepsilon) \\
&= (M\varepsilon, \varepsilon) = \varepsilon^T M\varepsilon.
\end{aligned}$$

Druga i predzadnja jednakost slijede iz činjenice da je $MX = 0$, a u drugom redu smo iskoristili da je $M = M^T$. Neka je m_{ij} (i, j)-ti član matrice M. Slijedi

$$Y^T M Y = \varepsilon^T M\varepsilon = \sum_{i=1}^{n} m_{ii}\varepsilon_i^2 + 2\sum_{1\le i<j\le n} m_{ij}\varepsilon_i\varepsilon_j.$$

Sada možemo računati matematičko očekivanje

$$E(Y^T MY) = \sum_{i=1}^{n} m_{ii} E(\varepsilon_i^2) + 2 \sum_{1 \leq i < j \leq n} m_{ij} E(\varepsilon_i \varepsilon_j).$$

$$= \sigma^2 \sum_{i=1}^{n} m_{ii} = \sigma^2 \mathrm{tr}(M) = \sigma^2 r(M) = \sigma^2 (n - k - 1).$$

U drugoj jednakosti iskoristili smo pretpostavke o slučajnim greškama. Prethodna jednakost vrijedi jer su rang i trag jednaki za ortogonalni projektor. Iz linearnosti matematičkog očekivanja slijedi tvrdnja propozicije:

$$\mathrm{E}\left(\frac{Y^\top MY}{n-k-1}\right) = \frac{1}{n-k-1} \mathrm{E}(Y^\top MY) = \sigma^2.$$

$\square$

Metoda maksimalne vjerodostojnosti

Ako u višedimenzionalnom linearnom regresijskom modelu uvedemo dodatnu pretpostavku o normalnoj distribuciji slučajnih grešaka, može se pokazati da metoda maksimalne vjerojatnosti daje iste procjenitelje za parametre modela kao i metoda najmanjih kvadrata. Procjene za nepoznati parametar σ^2 razlikuju se samo u faktoru.

Dakle, za višedimenzionalni linearni regresijski model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

pretpostavljamo da vrijedi $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$, tj. da su greške ε_i normalno distribuirane nezavisne slučajne varijable s očekivanjem $E[\varepsilon_i] = 0$ i konstantnom varijancom $\mathrm{Var}[\varepsilon_i] = \sigma^2$.

Funkcija gustoće normalne slučajne varijable s očekivanjem μ i varijancom σ^2 je dana izrazom:

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

Funkcija vjerojatnosti dana je sljedećim izrazom:

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(y_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2\right).$$

Korištenjem matrica, funkcija vjerojatnosti može se zapisati kao:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right).$$

Uobičajeno, radimo s logaritamskom funkcijom vjerojatnosti, koja je:

$$\ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.14}$$

Vidimo da za fiksnu vrijednost σ funkcija log-vjerojatnosti 2.14 postiže maksimum kada član $(y - X\hat{\beta})^\top(y - X\hat{\beta})$ postiže minimum. Dakle, procjene nepoznatih parametara $\beta_0, \beta_1, \ldots, \beta_k$ dobivene metodom maksimalne vjerojatnosti i metodom najmanjih kvadrata su ekvivalentne. Parcijalnim deriviranjem 2.14 po parametru σ^2 dobiva se procjena za nepoznati parametar σ^2 dobivena metodom maksimalne vjerojatnosti:

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^\top(y - X\hat{\beta})}{n}.$$

Dobivena procjena za varijancu slučajne greške σ^2 razlikuje se od procjene dobivene metodom najmanjih kvadrata po faktoru, iz čega slijedi da ML-procjenitelj nije nepristran.

2.3 Testiranje hipoteza

U kontekstu višeparametarske linearne regresije, nakon što procijenimo parametre nekom od dostupnih metoda, važno je testirati kvalitetu procjenitelja, odnosno dobivene procjene. Time zapravo provjeravamo i kvalitetu modela kojim pokušavamo opisati pojavu koju izučavamo. Navest ćemo nekoliko procedura testiranja kojima možemo odrediti kvalitetu modela te značajnost svakog od pojedinog regresora. Te procedure, i još mnogo toga, zainteresirani čitatelj može pronaći u [10]
Uz dosadašnje pretpostavke o očekivanju i varijanci slučajnih grešaka, potrebno je uvesti dodatnu pretpostavku o normalnosti slučajnih grešaka. Dakle, pretpostavljamo da slučajni vektor grešaka ϵ ima normalnu distribuciju kojoj je
n-dimenzionalni nul vektor $\mathbf{0}$ vektor očekivanja te kojoj je $\sigma^2 I$ kovarijacijska matrica. Ova dodatna pretpostavka omogućava primjenu raznih statističkih testova i metoda procjene, no njena opravdanost ovisi o specifičnim okolnostima i prirodi podataka. Postoji nekoliko načina za provjeru normalnosti grešaka. Grafički prikazi poput histograma ili Q-Q plotova mogu pomoći u vizualizaciji odstupanja od normalne distribucije. Također, statistički testovi poput Shapiro-Wilk testa ili

Kolmogorov-Smirnov testa mogu se koristiti za formalnu provjeru normalnosti. U praksi, čak i ako greške blago odstupaju od normalne distribucije, mnogi statistički testovi i dalje ostaju robusni. Naime, rezultati mogu biti pouzdani sve dok odstupanja od normalnosti nisu ekstremna. U slučajevima kada su varijable snažno asimetrične ili imaju značajne outliere, može biti potrebno koristiti alternativne metode koje ne zahtijevaju pretpostavku normalnosti, kao što su neparametarski testovi.
Iz gornje pretpostavke slijedi i da su slučajne varijable Y normalno distribuirane s očekivanjem $\beta_o + \sum_{j=1}^{k} \beta_j x_j$ i varijancom σ^2. To jest, za slučajni vektor $\mathbf{Y}$ vrijedi:

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 I) \tag{2.15}$$

Nadalje, pokazuje se da je slučajni vektor procjenitelja $\hat{\beta}$ također normalno distribuiran, s vektorom očekivanja β i kovarijacijskom matricom $\sigma^2 \left(\mathbf{X}^\tau \mathbf{X}\right)^{-1}$. Dakle, vrijedi:

$$\hat{\beta} \sim N_{k+1}(\beta, \sigma^2 \left(\mathbf{X}^\tau \mathbf{X}\right)^{-1}) \tag{2.16}$$

Uz dodatnu pretpostavku o normalnosti slučajnih grešaka, možemo također pokazati da za procjenitelja varijance $\hat{\sigma}^2$ definiranog s 2.13 vrijedi:

$$(n - k - 1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n - k - 1) \tag{2.17}$$

Naime, definirajmo Z kao standardni normalni slučajni vektor:

$$\mathbf{Z} := \frac{1}{\sigma}\epsilon \sim N_n(\mathbf{0}, I)$$

Uz oznake kao ranije tada vrijedi:

$$(n - k - 1)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\mathbf{Y}^\tau \mathbf{M} \mathbf{Y}}{\sigma^2} = \frac{\epsilon^\tau \mathbf{M} \epsilon}{\sigma^2} = \mathbf{Z}^\tau \mathbf{M} \mathbf{Z} \sim \chi^2(n - k - 1)$$

Koeficijent determinacije R^2

Koeficijent determinacije predstavlja ključni statistički pokazatelj koji se koristi za procjenu kvalitete regresijskog modela. Ovaj koeficijent izražava udio varijance zavisne varijable koji se može objasniti varijablama neovisnog modela. Vrijednosti koeficijenta determinacije R^2 se kreću između 0 i 1, pri čemu vrijednost bliža 1 sugerira da model dobro objašnjava varijancu podataka, dok vrijednost 0 ukazuje na slabiju prediktivnu moć modela. Za procjenu kvalitete regresijskog modela koristi se i korigirani koeficijent determinacije, u oznaci $\bar{R}^2$, koji predstavlja prilagođenu verziju R^2 koja uzima u obzir broj nezavisnih varijabli u modelu. Ovaj koeficijent se koristi kako bi se izbjeglo prekomjerno prilagođavanje modela (overfitting), što može nastati kada u modelu imamo višak nezavisnih varijabli koje ne doprinose značajno objašnjenju zavisne varijable. Uvedimo oznake:

$$\begin{aligned} SSE &:= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^\tau \mathbf{e} = \mathbf{Y}^\tau \mathbf{M}\mathbf{Y}, \\ SSR &:= \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2, \\ SST &:= \sum_{i=1}^{n}(Y_i - \overline{Y})^2. \end{aligned}$$

gdje je $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Veličina SSE se zove suma kvadrata grešaka (Sum of Squares Error), veličinu SSR zovemo regresijski zbroj kvadrata (Sum of Squares Regression) i ona mjeri odstupanja koja su objašnjiva regresijskim modelom, dok veličinu SST zovemo ukupni zbroj kvadrata (Total Sum of Squares). Za te veličine vrijedi sljedeća relacija koju ćemo iskazati u obliku propozicije:

Propozicija 2.3.1. *(Osnovna jednakost analize varijance za regresijski model) Vrijedi:*

$$SST = SSE + SSR$$

Dokaz. Neka je $\mathcal{N}$ potprostor od $\mathbb{R}^n$ razapet vektorom stupcem jedinica $\mathbf{1} \in \mathbb{R}^n$, te neka je $\mathcal{M}$ također potprostor od $\mathbb{R}^n$ razapet stupcima matrice $\mathbf{X}$. Tada je $\mathcal{N}$ potprostor od $\mathcal{M}$, što označavamo $\mathcal{N} \leq \mathcal{M}$. Nadalje, neka je N ortogonalni projektor na potprostor $\mathcal{N}$. Slijedi da je $N = N^\tau$, $N^2 = N$, $NM = MN = 0$, $NH = HN = N$. Veličinu SST tada možemo izraziti kao $SST = |\mathbf{Y} - \overline{Y}\mathbf{1}|^2$. Sada možemo primijeniti Pitagorin poučak na vektor $\mathbf{Y} - \overline{Y}\mathbf{1}$, definiciju projektora i navedena svojstva projektora. Vrijedi:

$$\begin{aligned} SST &= |\mathbf{Y} - \overline{Y}\mathbf{1}|^2 \\ &= |H(\mathbf{Y} - \overline{Y}\mathbf{1})|^2 + |M(\mathbf{Y} - \overline{Y}\mathbf{1})|^2 \\ &= |H\mathbf{Y} - HN\mathbf{Y}|^2 + |M\mathbf{Y} - MN\mathbf{Y}|^2 \\ &= |H\mathbf{Y} - N\mathbf{Y}|^2 + |M\mathbf{Y} - \mathbf{0}|^2 \\ &= |H\mathbf{Y} - \overline{Y}\mathbf{1}|^2 + |M\mathbf{Y}|^2 \\ &= |\hat{Y} - \overline{Y}\mathbf{1}|^2 + \mathbf{Y}^\tau M\mathbf{Y} \\ &= SSR + SSE \end{aligned}$$

□

Konačno, možemo definirati i koeficijent determinacije:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{2.18}$$

Obzirom da SST mjeri ukupnu varijabilnost izlaznih podataka, a zbroj kvadrata grešaka SSE mjeri preostalu varijabilnost izlaznih podataka nakon što se uzme u obzir utjecaj regresora, koeficijent R^2 možemo tumačiti kao udio rasipanja izlaznih podataka koji se može objasniti funkcijskom vezom ulaznih i izlaznih podataka. Zbog $0 \leq SSE \leq SST$, vrijedi i $0 \leq R^2 \leq 1$. Kao što smo već napomenuli, vrijednost koeficijenta R^2 blizu 1 ukazuje na to da je velik dio varijabilnosti u Y objašnjen regresijskim modelom. Ako je vrijednost koeficijenta determinacije R^2 bliže 0, to znači da veliki dio rasipanja izlaznih podataka otpada na rezidualno rasipanje koje nije objašnjeno regresijskim modelom. Iako razlog može biti slaba koreliranost između ulaznih i izlaznih podataka, tom slučaju imamo poticaj da razmislimo o tome trebamo li promijeniti regresijski model. Međutim, velika vrijednost R^2 ne znači nužno da je regresijski model dobar procjenitelj, posebno kada imamo mali broj podataka u odnosu na dimenziju regresijskog modela. To je jedan od razloga zbog kojeg definiramo korigirani koeficijent determinacije:

$$\overline{R^2} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

Uvođenjem nove regresorske varijable koeficijent determinacije R^2 se neće promijeniti bez obzira na doprinos nove varijable modelu. S druge strane, korigirani koeficijent determinacije $\overline{R^2}$ će se povećati kada uvedemo dodatnu regresorsku varijablu, ali samo ako dodatna varijabla smanjuje procjenu varijance slučajne greške.

Testiranje značajnosti linearnog regresijskog modela

Još jedan način na koji možemo ispitati kvalitetu modela je test značajnosti. Za detaljnije objašnjenje vidi [21]. Želimo odrediti postoji li linearna veza između izlazne varijable Y i barem jedne od ulaznih regresorskih varijabli $x_1, x_2, .., x_k$. Zbog toga ćemo postaviti sljedeće hipoteze:

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ &H_1 : \exists j \in \{1, 2, \ldots, k\} \text{ takav da } \beta_j \neq 0 \end{aligned} \tag{2.19}$$

Nulta hipoteza je da su svi regresijski koeficijenti jednaki 0, tj. da ne postoji linearna veza izlazne varijable Y i bilo koje ulazne varijable. Ulazne varijable uobičajeno je zvati faktori. Alternativna hipoteza je da je barem jedan regresijski koeficijent različit od 0. Dakle, ukoliko odbacimo nultu hipotezu možemo zaključiti da barem jedan od regresora značajno doprinosi modelu. Nadalje, neka je sada:

$$\begin{aligned} MSR &=: \frac{SSR}{k} \\ MSE &=: \frac{SSE}{n-k-1} \end{aligned} \tag{2.20}$$

Iskoristit ćemo ove definicije za definiranje testne statistike F:

$$F := \frac{MSR}{MSE}$$

Odredimo i distribuciju slučajne varijable F:

Teorem 2.3.2. *Ako vrijedi*

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

i $\epsilon \sim N_n(\boldsymbol{0}, \sigma^2 I_n)$, tada su statistike SSR i SSE nezavisne te vrijedi

$$\frac{SSR}{\sigma^2} \sim \chi^2(k) \tag{2.21}$$

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-k-1) \tag{2.22}$$

Nadalje,

$$F = \frac{MSR}{MSE} \sim F(k, n-k-1) \tag{2.23}$$

Dokaz teorema se može pronaći u [9].

Rezultati ovog testa mogu nam ukazati ima li smisla nastaviti provođenje daljnjih testiranja modela. Ukoliko ne postoji linearna povezanost izlazne varijable Y i barem neke od prediktora, trebalo bi razmisliti o izmjeni postavljenog modela. Cjelokupna procedura testa može se prikazati sljedećom ANOVA tablicom:

Izvor varijabilnosti	**Broj stupnjeva slobode**	**Zbroj kvadrata odstupanja**	**Srednje kvadratno odstupanje**	**F-statistika**
Model	k	SSR	MSR	F
Slučajna pogreška	$n-k-1$	SSE	MSE	
Ukupna varijabilnost	$n-1$	SST		

Tablica 2.1: ANOVA tablica

Testiranje značajnosti pojedinih kovarijata

Ako smo testom značajnosti linearnog modela utvrdili da postoji linearna veza između izlazne varijable Y i barem jedne od ulaznih regresorskih varijabli, preostaje nam još ispitati koji točno od regresora su značajni u našem modelu. Dakle, želimo testirati hipotezu da je pojedini regresijski koeficijent jednak 0, što znači da pripadajuća ulazna varijabla nema utjecaj na vrijednost izlazne varijable. Hipoteze za provjeru značajnosti pojedinog koeficijenta regresije su:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \tag{2.24}$$

Nulta hipoteza je tvrdnja da izlazna varijabla Y ne ovisi o regresoru x_j za fiksni $j \in 1, 2, .., k$. Prema tome, ako nulta hipoteza ne bude odbačena, možemo zaključiti da je regresor x_j višak te ga izbacujemo iz modela. Ako odbacimo nultu hipotezu, na zadanoj razini značajnosti, zaključujemo da je regresor x_j značajan u modelu te ga ne možemo izbaciti.
Iz jednakosti 2.20 slijedi da $\hat{\beta}_j$ ima normalnu distribuciju $N(\beta_j, \sigma_j^2)$ gdje je $\sigma_j^2 = \sigma^2 a_{jj}$, a a_{jj} j-ti dijagonalni element matrice $(X^\tau X)^{-1}$. Odatle slijedi da slučajna varijabla $U = \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}}$ ima standardnu normalnu distribuciju N(0,1). Iz 2.21 vidimo da slučajna varijabla $V = (n-k-1)\frac{\hat{\sigma^2}}{\sigma^2}$ ima $\chi^2(n-k-1)$ distribuciju. Slučajne varijable U i V su nezavisne pa statistika T definirana s:

$$T = \frac{U}{\sqrt{V/(n-k-1)}} = \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}} \tag{2.25}$$

ima Studentovu distribuciju s (n-k-1) stupnjeva slobode.
Pokazali smo da u slučaju kada nultu hipotezu nismo odbacili, testna statistika definirana s:

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{a_{jj}}}$$

ima Studentovu distribuciju s (n-k-1) stupnjeva slobode. Dakle, za zadanu razinu značajnosti α nultu hipotezu odbacujemo ako vrijedi $|t| > t_{\frac{\alpha}{2},n-k-1}$, to jest apsolutna vrijednost testne statistike je veća od $1 - \frac{\alpha}{2}$ kvantila Studentove distribucije s n-k-1 stupnjeva slobode. Naravno, regresijski koeficijent β_j će ovisiti i o svim drugim regresorskim varijablama $x_i, i \neq j$ koje su uključene u model. Prema tome, možemo reći da je ovo testiranje doprinosa varijable x_j u odnosu na ostale varijable uključene u model.

Testiranje značajnosti podskupa parametara

Slično kao u prethodnom odjeljku, možemo istovremeno ispitati doprinos proizvoljnog podskupa ulaznih varijabli modela. Skup od k regresorskih varijabli ćemo podijeliti na dva podskupa. Jedan će sadržavati varijable čiju značajnost doprinosa regresijskom modelu provjeravamo, a drugi sve ostale varijable. Neka je taj drugi skup kardinalnosti m. Tada će kardinalnost skupa varijabli čiji doprinos provjeravamo biti $k - m$. Potpuni model s k regresorskih varijabli zapravo možemo shvatiti kao proširenje modela koji sadrži samo m regresorskih varijabli te se tada zapravo pitamo opisuje li potpuni model pojavu koju izučavamo bolje od suženog modela. U tom slučaju hipoteze su:

$$\begin{aligned} &H_0 : \beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0 \\ &H_1 : \exists j \in \{m+1, m+2, \ldots, k\} \text{ takav da } \beta_j \neq 0 \end{aligned} \tag{2.26}$$

Dakle, nulta hipoteza sadrži tvrdnju da je model s m regresorskih varijabli dovoljan, a alternativna hipoteza tvrdnju da je potreban potpuni model. Za testnu statistiku možemo uzeti:

$$F := \frac{(SSR_m - SSR_k)/(k-m)}{SSR_k/(n-k-1)}$$

gdje je SSR_m regresijski zbroj kvadrata za suženi model s m regresorskih varijabli $x_1, x_2, .., x_m$, a SSR_k je regresijski zbroj kvadrata za potpuni model s k regresora.

Razlika $SSR_m - SSR_k$ je pozitivna jer regresijski zbroj kvadrata opada s povećanjem broja regresijskih varijabli. Testna statistika, to jest empirijski F-omjer ima $F(k-m, n-k-1)$ distribuciju. Dakle, za danu razinu značajnosti α, nultu hipotezu ćemo odbaciti ako vrijedi $F > F_{\alpha,k-m,n-k-1}$ što ukazuje da je barem jedna od varijabli $x_{m+1}, x_{m+2}, .., x_k$ značajna u modelu. Ovaj test uobičajeno se naziva parcijalni F-test jer mjeri doprinos podskupa regresora kada su ostali regresori već uključeni u model.
U slučaju $m = k - 1$, istražujemo doprinos modelu samo jedne varijable x_j kada su preostale $k - 1$ varijable $x_1, x_2, .., x_{j-1}, x_{j+1}, .., x_k$ već u modelu. Može se pokazati da je parcijalan F-test za jednu varijablu istovjetan t-statistici definiranoj u prethodnom odjeljku. Ipak, parcijalni F-test je općenitiji jer može ispitati značajnost proizvoljnog podskupa regresora. Parcijalni F-test pronalazi važnu primjenu upravo u modeliranju višeparametarske linearne regresije, odnosno u biranju najboljeg skupa regresora za model.

2.4 Pouzdani intervali

Pouzdani intervali su vrlo važan alat, posebno kada govorimo o procjeni regresijskih koeficijenata i očekivanja ciljne varijable. Ovi intervali nam pomažu da razumijemo koliko su naše procjene pouzdane. Na primjer, kada izračunamo regresijski koeficijent, pouzdani interval nam pokazuje raspon unutar kojeg možemo očekivati stvarnu vrijednost tog koeficijenta s određenim stupnjem sigurnosti. Širina pouzdanog intervala može se smatrati mjerom kvalitete našeg regresijskog modela. Uži interval ukazuje na veću preciznost i pouzdanost modela, dok širi interval može signalizirati probleme s modelom ili visoku varijabilnost u podacima. U praksi, korištenje pouzdanih intervala pomaže istraživačima da donesu bolje odluke i pravilno interpretiraju rezultate. Time se povećava vjerodostojnost analize i smanjuje rizik od pogrešnih zaključaka. Na taj način, pouzdani intervali postaju ključni alat za razumijevanje i komunikaciju rezultata statističkih analiza.
Vrijede pretpostavke kao u prethodnom potpoglavlju: slučajni vektor ϵ prati normalnu distribuciju s vektorom očekivanja 0_n te kovarijacijskom matricom $\sigma^2 I_n$.

Pouzdani intervali za nepoznate koeficijente

Nakon što smo odredili točkovne procjenitelje za regresijske koeficijente, želimo im pridružiti i pouzdane intervale na određenoj razini značajnosti α. Za konstrukciju pouzdanih intervala koristimo već definiranu statistiku 2.25 koja ima Studentovu distribuciju s $n - k - 1$ stupnjeva slobode.
Sada definiramo $(1 - \alpha) \cdot 100\%$ pouzdani interval za nepoznati parametar $\beta_j, \quad j =$

$0, 1, .., k$ je:

$$\left[\hat{\beta}_j - t_{\alpha/2,n-p}\hat{\sigma}_j, \hat{\beta}_j + t_{\alpha/2,n-p}\hat{\sigma}_j\right]$$

gdje je $\hat{\sigma}_j = \sqrt{\hat{\sigma^2} a_{jj}}$ procjena standardne devijacije regresijskog koeficijenta.

Pouzdani interval za očekivanje izlazne varijable

Osim točkovne procjene, možemo konstruirati intervalnu procjenu za očekivanje izlazne varijable Y u određenoj točki. Definiramo vektor vrijednosti prediktorskih varijabli x_0:

$$x_o = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

Želimo procijeniti vrijednost $\mathrm{E}[Y|x_0]$. Procjena vrijednosti izlazne varijable u točki x_0 jednaka je:

$$\hat{Y}_0 = x_0^\tau \hat{\beta}$$

$\hat{Y}_0$ je nepristrani procjenitelj za $\mathrm{E}[Y|x_0]$ jer vrijedi $\mathrm{E}[\hat{Y}_0] = x_0^\tau \beta = \mathrm{E}[Y|x_0]$. Uočavamo da je procjena očekivane vrijednosti od Y jednaka procjeni iznosa mjerenja Y za dani $\mathbf{x} = x_0$. Uz pretpostavku o normalnoj distribuciji slučajne greške, slijedi da je slučajna varijabla $\hat{Y}_0$ distribuirana s očekivanjem $\mathrm{E}[Y|x_0]$ i varijancom $\sigma^2 x_0^\tau (X^\tau X)^{-1} x_0$. Može se pokazati da sljedeća statistika slijedi Studentovu distribuciju s $n - k - 1$ stupnjeva slobode:

$$\frac{\hat{Y}_0 - \mathrm{E}[Y|x_0]}{\sqrt{\sigma^2 x_0^\tau (X^\tau X)^{-1} x_0}} \tag{2.27}$$

prema [11], može se pokazati da je $(1 - \alpha) \cdot 100\%$ pouzdani interval za $\mathrm{E}[Y|x_0]$, tj. za očekivanu vrijednost izlazne varijable u točki x_0 jednak

$$\left[\hat{Y}_0 - t_{\alpha/2,n-k-1}\sqrt{\sigma^2 x_0^\tau (X^\tau X)^{-1} x_0}, \hat{Y}_0 + t_{\alpha/2,n-k-1}\sqrt{\sigma^2 x_0^\tau (X^\tau X)^{-1} x_0}\right]$$

Pouzdani intervali za predviđanje novih opažanja

Linearna regresija nam omogućava da za proizvoljno odabrane vrijednosti ulaznih varijabli predvidimo vrijednost izlazne varijable. S obzirom na to da na temelju slučajnog uzorka ne možemo znati koja je prava vrijednost izlazne varijable za dane regresore, pokazuje se korisnim konstruirati pouzdane intervale. Za razliku od prethodnog poglavlja u kojem smo konstruirali intervalnu procjenu za srednju vrijednost izlazne varijable, sada ćemo konstruirati pouzdane intervale za procjenu vrijednosti izlazne varijable. Neka je sada $x_{01}, x_{02}, .., x_{0k}$ po volji odabrana vrijednost prediktorske varijable te $\hat{Y}_0$ točkovni procjenitelj za vrijednost izlazne varijable Y_0 u točki x_0 definiranoj sa:

$$x_o = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

Točkovni procjenitelj $\hat{Y}_0$ jednak je točkovnom procjenitelju za očekivanu vrijednost od Y uz dano $x = x_o$:

$$\hat{Y}_0 = x_0^\tau \hat{\beta}$$

Neka je $Y_0 = \mathbf{x_0^\tau} \hat{\beta} + \epsilon_0 \sim N(\mathbf{x_0^\tau} \hat{\beta}, \sigma^2)$ vrijednost izlazne varijable koju želimo predvidjeti. Pokazuje se da statistika:

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2(1 + x_0^\tau (X^\tau X)^{-1} x_0)}} \tag{2.28}$$

ima Studentovu distribuciju s $n - k - 1$ stupnjeva slobode iz čega proizlazi, a detaljno je raspisano u [11], da je $(1 - \alpha) \cdot 100\%$ pouzdani interval za vrijednost izlazne varijable u točki x_0 jednak:

$$\left[\hat{Y}_0 - t_{\alpha/2,n-k-1} \sqrt{\hat{\sigma}^2(1 + x_0^\tau (X^\tau X)^{-1} x_0)}, \hat{Y}_0 + t_{\alpha/2,n-k-1} \sqrt{\hat{\sigma}^2(1 + x_0^\tau (X^\tau X)^{-1} x_0)}\right]$$

2.5 Nelinearna regresija

Višeparametarska linearna regresija predstavlja temeljni koncept za shvaćanje složenijih modela, koji ne moraju biti nužno linearni. Višeparametarska linearna regresija u svom nazivu nosi pridjev linearna zato što je linearna u parametrima, to jest parametri u modelu imaju potenciju jednaku jedan. Do sada smo izučavali model koji je linearan i u parametrima i u ulaznim varijablama. U profesionalnom radu, često ćemo naići na pojave koje nećemo moći modelirati isključivo linearnom vezom između ulaznih i izlaznih varijabli. S obzirom na to da nelinearni modeli nisu tema ovoga rada, bit će spomenuti vrlo sažeto. Općenito, ako je model linearan u parametrima ali nije linearan u ulaznim varijablama, možda možemo pronaći prikladnu transformaciju koja će model svesti na model linearan u parametrima i ulaznim varijablama, te potom koristiti sve poznate rezultate iz prethodnih poglavlja. Na primjer:

$$Y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \cdot \ldots \cdot x_k^{\beta_k} \epsilon \tag{2.29}$$

Ovaj model naziva se opći multiplikativni regresijski model s k prediktora. Model je linearan u parametrima i nelinearan u ulaznim varijablama. Trivijalno je za vidjeti da je prikladna transformacija logaritmiranje:

$$\ln Y = \ln \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \ldots + \beta_k \ln x_k + \ln \epsilon \tag{2.30}$$

čime smo dobili model linearan u parametrima i u ulaznim varijablama, s time da umjesto originalnih vrijednosti koristimo logaritmirane vrijednosti zavisne i nezavisnih varijabli. U ovome modelu dobivene regresijske koeficijente $\hat{\beta}_j$ tumačimo kao srednju vrijednost postotka promjene zavisne varijable Y ako se ulazna varijabla x_j poveća za 1%, a sve ostale varijable ostanu nepromijenjene.
Posljednji primjer koji navodimo je linearni regresijski polinom u kojem se pretpostavlja da je veza između zavisne varijable Y i jedne nezavisne varijable x polinom stupnja n:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \epsilon \tag{2.31}$$

Ovaj model koji je linearan u parametrima, ali nije linearan u ulaznim varijablama, možemo uz jednostavnu supstituciju:

$$x_1^* = x, x_2^* = x^2, \ldots, x_n^* = x^n$$

shvatiti i kao model linearne regresije s n ulaznih varijabli te ponovno primjenjivati sve poznate rezultate.
Modele koji nisu linearni u parametrima ili ne postoji prikladna transformacija kojom se svode na linearni model, moramo analizirati drugim metodama.

Poglavlje 3

Pobačaj i kriminalitet u SAD

Opisano istraživanje i rezultati u ovom poglavlju temelje se na članku [13]. U svojoj studiji iz 2001. godine, Donohue i Levitt tvrdili su da je legalizacija pobačaja početkom 1970-ih značajno doprinijela smanjenju kriminala koje je zabilježeno 1990-ih. Predvidjeli su da će se, kako vrijeme prolazi, utjecaj legaliziranog pobačaja udvostručiti, rezultirajući stalnim godišnjim smanjenjem kriminala od oko 1% tijekom dva desetljeća.

Kasnija analiza koja obuhvaća podatke od 1998. do 2014. godine potvrdila je ovu predikciju, procjenjujući da je ukupni kriminal pao za 17.5% zbog legalizacije pobačaja, održavajući predviđeno godišnje smanjenje. Konkretno, od 1991. do 2014. godine, stope nasilnog i imovinskog kriminala smanjile su se za čak 50%. Legalizacija pobačaja uzrokovala je smanjenje nasilnog kriminala za 47% i imovinskog kriminala za 33%, što objašnjava značajan dio ukupnog smanjenja kriminala tijekom tog razdoblja.

3.1 Uvod

Donohue i Levitt predložili su vezu između legalizacije pobačaja i budućeg kriminala. Hipoteza je jednostavna: neželjena djeca imaju veći rizik za nepovoljne životne ishode, uključujući kriminal. Legalizacija pobačaja smanjila je broj neželjenih rođenja, što je rezultiralo manjim kriminalom među tim kohortama. Njihova analiza pokazala je da je legalizacija pobačaja možda najvažniji faktor u smanjenju kriminala 1990-ih, možda objašnjavajući polovicu pada kriminala u SAD-u između 1991. i 1997. godine. Njihova tvrdnja izazvala je brojne akademske rasprave i kritike. Iako su dokazi bili sugestivni, nisu bili opće prihvaćeni. Identifikacija procjena nije proizašla iz randomiziranog eksperimenta, već iz različitih izvora, od kojih svaki ima svoje manjkavosti. Također, vremenski okvir epidemije cracka, koja se poklopila s vršnim godinama kriminala prve kohorte izložene legaliziranom pobačaju, dodatno je otežao razlučivanje uzročnog utjecaja legalizacije pobačaja. Članak [13] analizirao je podatke o kriminalu prikupljene gotovo 20 godina nakon prvotne analize kako bi podržao ili opovrgnuo hipotezu o vezi između pobačaja i kriminala. Metodologija je jednostavna: reproduciramo glavne tablice iz Donohue i Levitt (2001), proširujući skup podataka na razdoblje od 1998. do 2014. godine. Rezultati pružaju snažnu podršku hipotezi o vezi između pobačaja i kriminala.

3.2 Pozadina problema i podatci

U 19. stoljeću, američke su države počele zabranjivati pobačaj, no te su zabrane počele popuštati krajem 1960-ih. Godine 1973., odluka Vrhovnog suda u slučaju Roe protiv Wade legalizirala je pobačaj na nacionalnoj razini.

Otprilike 18 godina nakon presude Roe protiv Wade, stopa kriminala iznenada je počela opadati. Od 1991. do 1997. godine, nasilni kriminal smanjio se za 20%, imovinski kriminal za 16%, a ubojstva za 30%. Između 1997. i 2014. godine, nasilni i imovinski kriminal po glavi stanovnika smanjili su se za 40%, a ubojstva za 35%.

Put do legalizacije nije stvorio uvjerljive nasumične varijacije u izloženosti pobačaju. Iako bi rana legalizacija pobačaja u pet država mogla poslužiti kao prirodni eksperiment, te su države jasni izuzeci. Čak i nakon što su stope pobačaja dosegle stabilno stanje tijekom 1980-ih, efektivna stopa pobačaja povezana s nasilnim kriminalom u državama koje su rano legalizirale pobačaj bila je gotovo dvostruko veća od one u ostatku zemlje. Efektivna stopa pobačaja na 1000 živorođenih predstavlja metriku koja izražava očekivani utjecaj legalizacije na stope kriminala, po danoj godini i saveznoj državi. Definirana je kao ponderirani prosjek stopa pobačaja za rođene kohorte unutar države, pri čemu su ponderi određeni udjelom ukupnih uhićenja na nacionalnoj razini za određenu kategoriju zločina iz 1985. godine za pojedince te dobi. Formalno:

$$\text{Efektivni pobačaj}_t = \sum_a \text{Pobačaj}_{t-a} \cdot \frac{\text{Uhićenja}_a}{\text{Uhićenja}_{\text{ukupno}}}$$

gdje t označava godine, a a označava dob kohorte. Pobačaj je broj pobačaja na 1.000 živorođenih, a omjer iz 1985. godine predstavlja udio uhićenja za određeni zločin koji uključuje osobe s dobi a.

3.3 Rezultati koji otkrivaju povezanost između pobačaja i kriminala

Kriminal je ranije i više opao u pet država koje su ranije legalizirale pobačaj

U ovom dijelu analiziramo obrasce kriminala u pet država (Aljaska, Kalifornija, Havaji, New York i Washington) koje su legalizirale ili polulegalizirale pobačaj oko 1970. godine, u usporedbi s ostatkom zemlje gdje pobačaj nije postao legalan do odluke Vrhovnog suda u slučaju Roe protiv Wade iz siječnja 1973. godine. U tablici 3.1 i 3.2 prikazani su ažurirani podaci iz Donohuea i Levitta (2001), koji pokazuju postotne promjene u kriminalu između 1976. i 1982., između 1982. i 1997., te između 1997. i 2014. za rane legalizatore i ostatak zemlje. Također prikazujemo razlike u tim postotnim promjenama kriminala između ranih legalizatora i ostatka nacije. Prva dva stupca odnose se na podatke dostupne u Donohueu i Levittu (2001), dok treći i četvrti stupac, koji prikazuju promjene kriminala od 1997. do 2014. te kumulativno između 1982. i 2014., predstavljaju nove podatke.

Grupa legalizacije	1976–82	1982–97	1997–2014	Kumulativno (1982–2014)
Nasilni kriminal				
Rani legalizatori	15.8	-12.9	-61.7	-74.7
Ostatak SAD-a	20.9	14.5	-41.7	-27.1
Razlika	-5.1	-27.5	-20.1	-47.5
SE	5.1	7.3	8.6	11.8
P-vrijednost	0.3	0.0	0.0	0.0
Imovinski kriminal				
Rani legalizatori	0.8	-44.3	-54.4	-98.6
Ostatak SAD-a	5.2	-9.5	-52.3	-61.8
Razlika	-4.3	-34.7	-2.1	-36.8
SE	2.7	5.7	4.8	8.8
P-vrijednost	0.1	0.0	0.7	0.0

Tablica 3.1: Trendovi nasilnog kriminala za države koje su rano legalizirale pobačaj naspram ostatka SAD-a

Grupa legalizacije	**1976–82**	**1982–97**	**1997–2014**	**Kumulativno (1982–2014)**
Ubojstva (UCR)				
Rani legalizatori	5,4	-40,8	-62,4	-103,3
Ostatak SAD-a	0,2	-24,7	-33,1	-57,7
Razlika	5,3	-16,2	-29,3	-45,5
SE	7,3	10,7	6,9	11,4
P-vrijednost	0,5	0,1	0,0	0,0
Ubojstva (VS)				
Rani legalizatori	8,4	-38,3	-58,3	-96,7
Ostatak SAD-a	4,2	-24,6	-27,3	-51,9
Razlika	4,2	-13,7	-31,1	-44,8
SE	6,1	9,9	6,1	10,4
P-vrijednost	0,5	0,2	0,0	0,0
Efektivna stopa pobačaja na kraju razdoblja				
Rani legalizatori	1,6	281,0	514,4	514,4
Ostatak SAD-a	0,1	139,4	294,6	294,6
Razlika	1,5	141,6	219,8	219,8

Tablica 3.2: Trendovi ubojstava za države koje su rano legalizirale pobačaj naspram ostatka SAD-a

Pet država koje su ranije legalizirale pobačaj ne samo da su održale više stope pobačaja, već su također doživjele rastuću razliku u efektivnim stopama pobačaja vezanim uz nasilni kriminal u usporedbi s drugim državama. Do 1997. razlika je iznosila 141.6, a povećala se na 219.8 do 2014. Naša teorija predviđa da nije bilo značajnih razlika u obrascima kriminala između ranih legalizatora i drugih država prije 1982., ali predviđa veće smanjenje stope kriminala nakon toga. Podaci iz tablica podržavaju ovo predviđanje, pokazujući da nije bilo statistički značajnih trendova kriminala prije 1982. Od 1982. do 1997., nasilni kriminal smanjio se za 27.5% u državama koje su ranije legalizirale pobačaj u odnosu na druge, dok su se imovinski kriminal i ubojstva također značajno smanjili. Između 1997. i 2014., nasilni kriminal je pao dodatnih 20.1%, dok imovinski kriminal nije pokazao značajnu promjenu, ali su stope ubojstava pale za oko 30%. Kumulativne razlike u stopama kriminala tijekom cijelog razdoblja bile su statistički značajne.

Kriminal se više smanjio u državama s visokim stopama pobačaja nego u onima s niskim stopama

Drugi izvor varijacije za identifikaciju veze između pobačaja i kriminala je usporedba obrazaca kriminala među državama s različitim razinama korištenja pobačaja nakon legalizacije. Slijedeći Donohuea i Levitta (2001), rangiramo države prema njihovim efektivnim stopama pobačaja za nasilni kriminal iz 1997. i dijelimo ih u tri kategorije: niska, srednja i visoka. Gornja tri dijela Tablice 3.3 prikazuju postotne promjene u državama s visokim, srednjim i niskim stopama pobačaja za nasilni kriminal, imovinski kriminal i ubojstva tijekom razdoblja 1973–85, 1985–97 i 1997–2014. Posljednji dio tablice prikazuje prosječnu efektivnu stopu pobačaja na kraju relevantnog razdoblja za tri grupe država.

Trebalo bi biti malo ili nimalo utjecaja pobačaja na kriminal prije 1985., jer su efektivne stope pobačaja izuzetno niske te godine, čak i u državama s visokim stopama pobačaja. Rezultati u prvom stupcu za razdoblje 1973–85 su u skladu s tom pretpostavkom. Obrasci stopa nasilnog kriminala su vrlo slični među državama s niskim, srednjim i visokim stopama pobačaja. Imovinski kriminal raste manje u državama s visokim stopama pobačaja nego u onima s niskim, ali je suprotan obrazac za ubojstva, gdje su smanjenja kriminala najmanja u državama s visokim stopama pobačaja.

Promjene u kriminalu između 1985. i 1997. otkrivaju vrlo drugačiji obrazac. U svakoj kategoriji kriminala, države s visokim stopama pobačaja doživljavaju povoljnije trendove kriminala od država sa srednjim stopama, dok države s niskim stopama prolaze najgore. Za sve tri kategorije kriminala, razlika između država s visokim i niskim stopama pobačaja iznosi približno 30 postotnih bodova, a nešto je veća za ubojstva prema podacima o nasilju.

Učestalost pobačaja (1997)	**1973–85**	**1985–97**	**1997–2014**	**Kumulativno (1985–2014)**
Nasilni kriminal				
Najniže	32.9	26.3	-23.3	3.1
Srednje	28.5	20.6	-36.2	-15.6
Najviše	28.6	-1.5	-60.7	-62.2

Učestalost pobačaja (1997)	**1973–85**	**1985–97**	**1997–2014**	**Kumulativno (1985–2014)**
Imovinski kriminal				
Najniže	33.6	10.8	-42.1	-31.2
Srednje	27.4	2.9	-45.7	-42.8
Najviše	13.2	-22.2	-61.5	-83.7
Ubojstvo (UCR)				
Najniže	-23.5	7.4	-32.3	-24.9
Srednje	-20.8	-12.7	-32.4	-45.2
Najviše	-11.9	-25.3	-46.7	-71.9
Ubojstvo (VS)				
Najniže	-17.0	13.7	-27.0	-13.3
Srednje	-22.8	-12.3	-25.5	-37.7
Najviše	-7.4	-23.6	-42.3	-65.9
Efektivni pobačaj na 1,000 na kraju razdoblja				
Najniže	0.8	77.0	179.8	179.8
Srednje	1.4	125.6	265.6	265.6
Najviše	5.4	232.6	450.8	450.8

Tablica 3.3: Promjene u kriminalu 1985.–2014. u državama s niskim, srednjim i visokim stopama pobačaja

U trećem stupcu prikazani su rezultati za razdoblje nakon objave originalnog rada. U sve tri kategorije kriminala, smanjenje kriminala je znatno veće u državama s visokim stopama pobačaja nego u onima s niskim, što teorija predviđa. Razlike su značajne: nasilni kriminal pao je dodatnih 35 postotnih bodova od 1997. godine u državama s visokim stopama pobačaja u odnosu na niske. Za imovinski kriminal ta razlika iznosi više od 19 postotnih bodova, a za ubojstva približno 15 postotnih bodova. Kada se agregiraju podaci za cijelo razdoblje od 1985. do 2014. (četvrti stupac), države s visokim stopama pobačaja doživjele su smanjenje kriminala u odnosu na niske države od -65.3, -52.5 i -47.0 postotnih bodova za nasilni kriminal (UCR), imovinski kriminal i ubojstva, redom. Prema podacima Vital Statistics, razlika u ubojstvima iznosi -52.6 postotnih bodova.

Pobačaj je izuzetno značajan u objašnjenju smanjenja kriminala u panel podacima

Treći izvor varijacije dolazi iz analize panel podataka koja nam omogućuje kontrolu za druge faktore, uz stope pobačaja, koji utječu na kriminal. Procjena ima oblik:

$$\ln(\text{Kriminal}_{st}) = \beta_1 \cdot \text{Efektivni pobačaj}_{st} + X_{st} \cdot \Theta + \gamma_s + \lambda_t + \epsilon_{st}$$

Zavisna varijabla predstavlja logaritmiranu stopu kriminala po glavi stanovnika u saveznoj državi s u trenutku t. Naša glavna nezavisna varijabla je efektivna stopa pobačaja za određenu državu, godinu i kategoriju kriminala. X je vektor kontrolnih varijabli na razini države, uključujući broj zatvorenika i policajaca po glavi stanovnika, skup varijabli koje prikazuju ekonomske uvjete u državi, kašnjenje državne socijalne pomoći, indikator prisutnosti zakona o skrivenom nošenju pištolja i potrošnju piva po glavi stanovnika. Uključeni su fiksni efekti za državu i godinu, označeni kao γ_s i λ_t. Sve regresije su ponderirane prema populaciji države i prilagođene za serijsku korelaciju koristeći metodu koju su opisali Bhargava i sur. (1982) u [3]. Sažeti statistički podaci za cijeli uzorak procjene prikazani su u Tablici 3.4. Prikazujemo kako ukupne tako i unutar-državne standardne devijacije, što su relevantnije mjere kada su uključeni fiksni efekti za države. Efektivna stopa pobačaja razlikuje se među kategorijama kriminala zbog razlike u dobnoj distribuciji uhićenja.

Rezultati regresije prikazani su u Tablici 3.5. Zavisna varijabla u stupcima 1 i 2 je (logaritmirani) nasilni kriminal, dok stupci 3 i 4 prikazuju (logaritmirani) imovinski kriminal. Stupci 5 i 6 odražavaju (logaritmirana) ubojstva prema UCR-u, a stupci 7 i 8 prikazuju (logaritmirana) ubojstva prema Vital Statistics. Za svaku od četiri mjere kriminala prikazane su dvije različite specifikacije: neparni stupci bez kontrolnih varijabli, a parni s punim skupom kontrola.

Svi koeficijenti za pobačaj u tablici 3.5 su negativni, što implicira da su više stope pobačaja povezane s nižim kriminalom. Ovi učinci su statistički značajni, pri čemu je jedanaest od dvanaest procjena značajno na razini od 0.01. Koeficijenti za pobačaj u razdoblju nakon našeg prvog istraživanja veći su u svih osam specifikacija tablice, što sugerira da su rezultati izvan uzorka jači nego u originalnom radu.

Povećanje efektivne stope pobačaja od 100 na 1000 živorođenih povezano je sa smanjenjem kriminala od otprilike 10–20%.

Dosljedna snažna negativna povezanost između pobačaja i kriminala je upečatljiva. Dodavanje kontrola modelima panel podataka ima mali utjecaj na procjenu učinka pobačaja za nasilni kriminal, dok povećava učinak za imovinski kriminal i mjere ubojstava. Rezultati Tablice 3.5 ostali su izuzetno robusni čak i kada su korištene različite specifikacije i dodatne kontrole.

Varijabla	**Prosjek**	**Standardna devijacija (Ukupno)**	**Standardna devijacija (Unutar države)**
Nasilni zločin na 100,000 stanovnika	540.93	238.43	156.97
Imovinski zločin na 100,000 stanovnika	3,882.96	1,215.86	968.50
Ubojstvo na 100,000 stanovnika (UCR)	6.59	3.60	2.33
Ubojstvo na 100,000 stanovnika (VS)	7.03	3.53	2.25
EAR: Nasilni zločin	203.64	152.38	128.47
EAR: Imovinski zločin	240.93	151.27	117.39
EAR: Ubojstvo	179.33	148.15	128.97
Zatvorenici na 1,000 stanovnika (t-1)	3.83	1.61	1.05
Policija na 1,000 stanovnika (t-1)	3.08	0.71	0.37
Realni osobni dohodak po glavi stanovnika	17,045.93	2,914.40	1,942.68
Realna AFDC socijalna pomoć po obitelji primatelja/1,000 (t-15)	3.76	1.73	1.05
Stopa nezaposlenosti u državi (%)	6.20	1.92	1.73
Potrošnja piva po glavi stanovnika (Galoni etanola)	1.22	0.19	0.09
Stopa siromaštva	13.47	3.24	1.77

Tablica 3.4: Sažetak Statistika, 1985–2014

Ovisna varijabla:	**Nasilni kriminal**		**Imovinski kriminal**		**UCR Ubojstva**		**VS Ubojstva**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Stope pobačaja 1985–1997	-0,184**	-0,178**	-0,138**	-0,152**	-0,087**	-0,100**	-0,098**	-0,116**
	(0,022)	(0,022)	(0,017)	(0,016)	(0,038)	(0,040)	(0,034)	(0,036)
Stope pobačaja 1998–2014	-0,192**	-0,189**	-0,149**	-0,168**	-0,131**	-0,152**	-0,144**	-0,164**
	(0,019)	(0,019)	(0,016)	(0,015)	(0,017)	(0,021)	(0,016)	(0,019)
ln (zatvorenici po glavi stanovnika, zaostalo)		0,007		-0,111**		-0,021*		-0,133**
		(0,037)		(0,034)		(0,056)		(0,051)
ln (policija po glavi stanovnika, zaostalo)		-0,015		-0,027		-0,137*		-0,186**
		(0,015)		(0,014)		(0,053)		(0,049)
Stopa nezaposlenosti		-0,027		0,536		1,212		1,084
		(0,356)		(0,314)		(0,716)		(0,657)
ln (realni dohodak po glavi stanovnika)		0,003		-0,076		0,329		0,172
		(0,129)		(0,114)		(0,224)		(0,203)
Stopa siromaštva		-0,002		-0,001		-0,003		-0,001
		(0,001)		(0,001)		(0,003)		(0,003)
Realni AFDC		0,004		-0,001		-0,008		-0,003
		(0,003)		(0,003)		(0,008)		(0,007)
Zakon o skrivenom oružju		0,014		0,019		-0,042		-0,023
		(0,015)		(0,011)		(0,022)		(0,022)
Potrošnja piva po glavi stanovnika		0,077		0,026		0,286**		0,286**
		(0,050)		(0,044)		(0,106)		(0,096)
Godine FE	Da	Da	Da	Da	Da	Da	Da	Da
Države FE	Da	Da	Da	Da	Da	Da	Da	Da
Broj promatranja	1.530	1.517	1.530	1.517	1.530	1.517	1.512	1.499
R^2	0,815	0,844	0,974	0,978	0,908	0,853	0,877	0,887
Prilagođeni R^2	0,805	0,834	0,973	0,978	0,845	0,857	0,870	0,880

Tablica 3.5: Panelni podaci: Procjena odnosa između stopa pobačaja i kriminala, 1985–2014

Napomena: * $P < 0,05$, ** $P < 0,01$.
Ovisna varijabla je log prirodnog broja kriminala po glavi stanovnika prema vrsti kriminala. Drugi stupac uključuje dodatne varijable s fiksnim efektima za države i godine. Rezultati se temelje na ponderiranim procjenama najmanjih kvadrata. Standardne greške prikazane su u zagradama.

3.4 Povezivanje stopa pobačaja s uhićenjima prema dobi

Prednosti prelaska s podataka o kriminalu na podatke o stopi uhićenja

Dosadašnja analiza pokazala je povezanost između porasta učinkovitih stopa pobačaja i smanjenja nasilnih zločina, imovinskih zločina i ubojstava. Stope kriminala su prirodan i logičan ishod za proučavanje, ali imaju ograničenje: ne omogućuju uvid u dob počinitelja jer počinitelj često ostaje nepoznat. To znači da se analize kriminaliteta moraju provoditi na razini saveznih država i godina, što onemogućuje detaljno testiranje hipoteze o povezanosti pobačaja i kriminala. Ta hipoteza predviđa da bi obrasci kriminala trebali varirati među generacijama unutar iste države i godine, ovisno o stopi pobačaja u trenutku kada je određena generacija bila u maternici.

Kako nije moguće dobiti precizne podatke o stopama kriminala prema dobi počinitelja, oslanjamo se na podatke o uhićenjima. Kod zločina gdje postoji uhićenje, dostupni su podaci o dobi uhićenih, što omogućuje analizu na razini države, godine i pojedinačne dobi. Ovo nam omogućuje uključivanje varijabli za kombinacije država, godina i dobnih skupina u panel analizu, čime se hipoteza testira s razinom preciznosti koja nije moguća s agregiranim podacima o kriminalu. Preciznije:

$$\ln(\text{Uhićenja}_{sta}) = \beta_1 \cdot \text{Efektivni pobačaj}_{sta} + \gamma_{sa} + \lambda_{at} + \Theta_{st} + \epsilon_{sta} \tag{3.1}$$

gdje su s, t i a indeksi za državu, godinu i dob, respektivno. Varijabla Uhićenja predstavlja broj uhićenja za određeni zločin. Kao mjeru stope pobačaja za specifičnu kohortu koristimo stopu pobačaja u državi gdje je uhićenje izvršeno, u kalendarskoj godini koja je najvjerojatnije prethodila rođenju uhićenika. Dummy varijable za državu $\times$ dob, dob $\times$ godina i država $\times$ godina apsorbiraju varijaciju duž tih različitih dimenzija. Sva varijacija u kovarijatima korištenim u prethodnim panelnim regresijama kriminala nalazi se na razini država $\times$ godina, tako da u ovim specifikacijama ne ostaje varijacija u tim kovarijatima. Podaci po pojedinačnim godinama dobi dostupni su samo za uzrast 15–24 (za starije i mlađe uzraste podaci su grupirani, obično u petogodišnje dobne skupine), stoga ograničavamo naš uzorak na to dobno razdoblje.

Tablica 3.6 prikazuje rezultate. Ovisna varijabla je prirodni logaritam broja uhićenja za kategoriju zločina navedenu na vrhu stupca. U skladu s izvornim radom, predstavljamo rezultate za nasilni kriminal (stupci 1–3) i imovinski kriminal (stupci 4–6), ali ne i za ubojstvo, jer je ubojstvo dovoljno rijetko (s još rjeđim uhićenjima) tako da mnoge ćelije za državu-godinu-dob ostanu prazne. Skup kovarijata se povećava od lijeva prema desno unutar svake kategorije zločina, što je naznačeno u donjem dijelu tablice. Gornje dvije vrste prikazuju koeficijent za stopu pobačaja u razdoblju obuhvaćenom izvornim podacima (prva vrsta) i u kasnijim godinama (druga vrsta). U tablici je prikazan samo koeficijent za stopu pobačaja. Kao i kod naše regresije kriminala, procijenili smo i regresiju uhićenja koristeći specifikaciju s jednom varijablom pobačaja, koju prikazujemo u trećem retku Tablice 3.6.

	ln (Nasilna uhićenja)			*ln* (Imovinska uhićenja)		
	(1)	(2)	(3)	(4)	(5)	(6)
Stopa pobačaja '85.–'97. (×100)	-0,033	-0,056	-0,031	-0,048	-0,032	-0,029
	(0,006)**	(0,008)**	(0,006)**	(0,007)**	(0,004)**	(0,004)**
	[0,012]**	[0,024]*	[0,014]*	[0,017]*	[0,012]*	[0,010]
Stopa pobačaja '98.–'14. (×100)	-0,042	-0,049	-0,057	-0,086	-0,080	-0,082
	(0,006)**	(0,006)**	(0,007)**	(0,005)**	(0,005)**	(0,005)**
	[0,024]	[0,026]	[0,017]**	[0,015]*	[0,012]*	[0,011]*
Stopa pobačaja '85.–'14. (×100)	-0,039	-0,051	-0,038	-0,074	-0,067	-0,033
	(0,005)**	(0,006)**	(0,005)**	(0,005)**	(0,005)**	(0,005)**
	[0,019]*	[0,025]*	[0,013]**	[0,012]*	[0,012]*	[0,010]
Godina × Dob?	Da	Da	Da	Da	Da	Da
Fiksni učinci države?	Da	Implied	Implied	Da	Implied	Implied
Država × Dob?	Ne	Da	Da	Ne	Da	Da
Država × Godina?	Ne	Ne	Da	Ne	Ne	Da
Broj promatranja	13,765	13,765	13,765	13,770	13,770	13,770

Tablica 3.6: Povezanost između stopa pobačaja i uhićenja prema dobi, 1985.–2014.

Napomena: * $p < 0,05$; ** $p < 0,01$. Rezultati u tablici predstavljaju koeficijente regresijskih procjena jednadžbe (3.1). Jedinica promatranja u regresijama je godišnji broj uhićenja prema pojedinačnoj dobi unutar države. Uzorak pokriva razdoblje od 1985. do 2014. za dob od 15 do 24 godine. Gornji dio tablice procjenjuje učinak pobačaja na početni period (1985.–1997.) i za ostatak perioda (1998.–2014.). Donji dio procjenjuje jednu varijablu za stopu pobačaja kroz cijeli period. Pogreške standardizirane po klasterima prikazane su u uglatim zagradama.

Tablica 3.6 prikazuje da su koeficijenti pobačaja negativni i visoko značajni za obje kategorije zločina i sve specifikacije. Uključivanje dodatnih kovarijata ne utječe očigledno na veličinu koeficijenata pobačaja. U skladu s rezultatima regresije iz Tablice 3.5, procjene koeficijenata pobačaja su veće u kasnijem razdoblju nego u početnom uzorku u pet od šest stupaca u Tablici 3.6, a u nekim specifikacijama te razlike su statistički značajne.

Treći redak Tablice 3.6 pokazuje učinke pobačaja na kriminal procijenjene tijekom cijelog razdoblja od 1985. do 2014. godine. Primijetimo da stupci 3 i 6, koji procjenjuju učinak pobačaja na uhićenja s punim skupom fiksnih efekata, pokazuju ukupni učinak od -0.038 za nasilni kriminal i -0.033 za imovinski kriminal.

Oba su visoko statistički značajna koristeći bilo koju od dviju prikazanih procjena standardne greške: grupirane prema rodu i državi (u zagradama) i prema državi (u uglatim zagradama).

Poboljšanje preciznosti mjera za pobačaj

Dosadašnji rezultati izravno su slijedili specifikacije i definicije podataka Donohuea i Levitta (2001) kako bi se usporedba novih rezultata s izvornim učinila što jasnijom. Primjetna iznimka bila je korištenje boljih podataka o pobačaju prema državi prebivališta, koji su im dostavljeni nakon prvotne objave i na koje se u ovom članku oslanjaju kao na glavnu mjeru pobačaja.

Od objavljivanja tog prvog rada, pokušali su riješiti problem mjernih pogrešaka u varijabli pobačaja na tri načina: s dva poboljšanja u konstrukciji varijable koja bolje povezuju te varijable s teorijom i korištenjem instrumentalne varijable za mjeru pobačaja prema prebivalištu. Prvo, konstruirali su mjeru pobačaja koja bolje odgovara stvarnom mjesecu i godini rođenja pojedinca. Drugo, prilagodili su mjeru pobačaja uzimajući u obzir mobilnost između država od rođenja do adolescencije. Treće, prepoznajući šum u proxy varijabli pobačaja (temeljenoj na podacima Alan Guttmacher Instituta), koristili su drugu neovisno generiranu procjenu stope pobačaja (iz Centara za kontrolu bolesti) kao instrumentalnu varijablu.

Tablica 3.7 ilustrira učinak provođenja istih regresija iz prethodne tablice koristeći ova dva prilagođavanja naše mjere pobačaja i instrumentiranje kako bismo adresirali utjecaj mjernih pogrešaka. Prikazujemo naše procjene instrumentalnih varijabli (IV) na dva načina: prvo razdvajanjem učinka pobačaja na 1985–1997 i 1998–2014 u prva dva reda, a zatim procjenjujući jedinstvenu varijablu pobačaja za 1985–2014, koja je prikazana u trećem redu. U usporedbi s Tablicama 3.5 i 3.6, vidi se da su procijenjeni učinci pobačaja na kriminal veći kada se koristi bolja mjera pobačaja i instrumentiranje—često udvostručujući veličinu—u svim slučajevima osim učinka na imovinska uhićenja u drugom razdoblju s interakcijom država $\times$ godina (stupac 6), koja je praktički nula. Sve ukupne periodične procjene u redu 3 su statistički značajne za oba skupa procjena standardne greške i znatno veće od odgovarajućih vrijednosti iz Tablice 3.6.

Premještanje iz Tablice 3.6 u Tablicu 3.7 osvjetljava neke aspekte naših podataka i važan društveni fenomen: opadajuću međudržavnu mobilnost od ranih 1970-ih. Svi procijenjeni učinci u prvom razdoblju su se barem udvostručili uvođenjem ispravki, dok nijedan od procjena u drugom razdoblju nije. Fokusirajući se na procjene nasilnog kriminala, tri procjene iz prvog razdoblja porasle su za najmanje 100%, dok su povećanja u drugom razdoblju bila između 30% i 60%. To nije iznenađujuće s obzirom na to da su naši podaci o pobačaju očito slabije mjereni u

Tablica 3.7: Procijenjeni učinci pobačaja na kriminal s prilagodbama za pogreške u mjerenju, 1985.–2014.

	ln (Nasilna uhićenja)			*ln* (Imovinska uhićenja)		
	(1)	(2)	(3)	(4)	(5)	(6)
IV učinak pobačaja ’85.–’97. (×100)	-0,065	-0,114	-0,115	-0,100	-0,074	-0,080
	(0,013)**	(0,018)**	(0,026)**	(0,013)**	(0,014)**	(0,017)**
	[0,017]**	[0,025]**	[0,029]**	[0,026]**	[0,025]**	[0,018]**
IV učinak pobačaja ’98.–’14. (×100)	-0,067	-0,068	-0,074	-0,166	-0,154	-0,136
	(0,010)**	(0,012)**	(0,021)**	(0,014)**	(0,016)**	(0,022)**
	[0,032]*	[0,034]*	[0,034]*	[0,056]*	[0,062]*	[0,028]*
IV učinak pobačaja ’85.–’14. (×100)	-0,066	-0,087	-0,108	-0,135	-0,121	-0,065
	(0,010)**	(0,013)**	(0,022)**	(0,013)**	(0,015)**	(0,014)**
	[0,021]**	[0,029]**	[0,030]**	[0,041]**	[0,052]**	[0,017]**
Godina × Dob?	Da	Da	Da	Da	Da	Da
Fiksni učinci države?	Da	Implicitni	Implicitni	Da	Implicitni	Implicitni
Država × Dob?	Ne	Da	Da	Ne	Da	Da
Država × Godina?	Ne	Ne	Da	Ne	Ne	Da
Broj promatranja	13,765	13,765	13,765	13,770	13,770	13,770

Napomena: * $p < 0,05$; ** $p < 0,01$. Tablica prilagođava mjeru stope pobačaja kako bi bolje povezala vrijeme pobačaja s odgovarajućom dobnom kohortom te kako bi odražavala međudržavno kretanje od države rođenja do države prebivališta u trenutku mjerenja. Naše instrumentalne varijable (IV) koriste mjeru CDC-a kao instrument za AGI mjeru pobačaja. Gornji dio tablice procjenjuje učinak pobačaja za početni period (1985.–1997.) i za ostatak perioda (1998.–2014.). Donji dio tablice procjenjuje jednu varijablu za stopu pobačaja kroz cijelo razdoblje 1985.–2014.

ranim danima legalizacije, pa stoga očekujemo da će instrumentiranje za poboljšanje točnosti podataka o pobačaju imati veći učinak na naše procjene iz prvog razdoblja. Slično tome, prilagodbe migracije iz Tablice 3.7 su značajnije u prvom razdoblju kada je međudržavna migracija bila mnogo veća. Kao rezultat, naš pokušaj povezivanja stope pobačaja za određenu kohortu s državom u kojoj članovi kohorte konačno žive poboljšava kvalitetu naših mjera, povećavajući time veličinu procijenjenog učinka pobačaja na kriminal.

Foote i Goetz (2008) u svom radu [7] tvrdili su da naša analiza panel podataka o stopama kriminala (Tablica 3.6 u ovom radu) i regresije koje objašnjavaju uhićenja prema dobi možda ne uspostavljaju potpuno vezu između pobačaja i kriminala koju smo postavili. Srž njihove kritike bila je da stopa pobačaja može predstavljati neku specifičnu varijablu koja je izostavljena na razini države, te da bi stoga "ključni" test trebao eliminirati "potencijalnu pristranost zbog izostavljenih varijabli" na razini država i godina kroz regresiju stopa uhićenja po glavi stanovnika koja uključuje fiksne efekte država i godina. Ova regresija kontrolira čimbenike koji utječu na kriminal u određenoj državi i godini te određuje hoće li stopa pobačaja u vrijeme rođenja bilo koje kohorte korelirati sa stopom uhićenja te kohorte tijekom godina kada su imali između 15 i 24 godine (za koje imamo specifične stope uhićenja).

U Tablici 3.8 rješavamo ove probleme, nastavljajući praksu prikazivanja dva skupa standardnih grešaka u svim regresijama stopa uhićenja, uključujući njihovo preferirano grupiranje prema državi. Svaka regresija u Tablici 3.8 također uključuje "ključne" fiksne efekte država i godina. Prema tome, prikazujemo u stupcima 2 i 4 upravo ono što su Foote i Goetz naveli kao sredstvo za uspostavljanje ili opovrgavanje veze između pobačaja i kriminala. Kratak odgovor je da, koristeći preciznu varijablu stope uhićenja po glavi stanovnika i prilagodbu grupiranja koju Foote i Goetz preporučuju, učinak pobačaja na kriminal koji smo identificirali 2001. ostaje snažan i statistički značajan na razini od 0,05 za nasilni kriminal (vidi vrijednost -0,05 u redu 4, stupac 2) te je negativan ali ne statistički značajan za imovinski kriminal (vrijednost -0,007 u redu 4, stupac 4) kada se procjenjuje za razdoblje od 1985. do 2014.

	Zavisna varijabla:			
	ln(Nasil. uhićenja) (1)	ln(Nasil. uhićenja po stan.) (2)	ln(Imov. uhićenja) (3)	ln(Imov. uhićenja po stan.) (4)
IV efekt pobačaja ’85–’97 (×100)	-0.065	-0.041	-0.044	-0.006
	(0.025)**	(0.021)*	(0.017)**	(0.013)
	[0.028]*	[0.029]	[0.019]*	[0.020]
IV efekt pobačaja ’98–’14 (×100)	-0.084	-0.089	-0.004	-0.011
	(0.021)**	(0.021)**	(0.021)	(0.022)
	[0.029]**	[0.029]**	[0.031]	[0.037]
ln(SEER populacija)	0.680		0.486	
	(0.066)**		(0.053)**	
	[0.125]**		[0.134]**	
IV efekt pobačaja ’85–’14 (×100)	-0.069	-0.050	-0.036	-0.007
	(0.021)**	(0.018)**	(0.014)**	(0.014)
	[0.023]**	[0.023]*	[0.019]	[0.020]
ln(SEER populacija)	0.672		0.503	
	(0.061)**		(0.051)**	
	[0.124]**		[0.137]**	
Godina * Dob?	Da	Da	Da	Da
Državni fiksni efekti?	Implicitno	Implicitno	Implicitno	Implicitno
Država * Dob?	Da	Da	Da	Da
Država * Godina?	Da	Da	Da	Da
ln(Populacija)?	Da	Ne	Da	Ne
Broj opažanja	13,765	13,765	13,770	13,770

Tablica 3.8: Razlikovanje između kanala kroz koje pobačaj utječe na kriminal, 1985–2014

Napomena: $^*p < 0.05$; $^{**}p < 0.01$
Tablica mijenja stupce 3 i 6 na dva načina kako bi uklonila efekt veličine kohorte na uhićenja po pojedinoj godini dobi za dob 15–24. Stupci 1 i 3 jednostavno dodaju kontrolu za populaciju svake države po pojedinoj godini dobi. Činjenica da su procijenjene vrijednosti za ovu kontrolu populacije znatno ispod 1 (vidi redove 3 i 5) ilustrira prisutnost pogreške mjerenja u varijabli populacije. Stupci 2 i 4 kontroliraju za populaciju mijenjanjem zavisne varijable u ln(stopa uhićenja po stanovniku) po pojedinoj godini dobi, a te procjene će patiti od pristranosti omjera zbog opažene pogreške mjerenja u varijabli populacije koja se pojavljuje u nazivnicima obje zavisne varijable i nezavisne varijable pobačaja. Standardne pogreške grupirane po kohorti godine rođenja i države navedene su u zagradama, dok su standardne pogreške grupirane po državi navedene u uglatim zagradama odmah ispod.

Iako su dodatni podaci iz 17 godina ojačali dokaze u korist hipoteze o pobačaju i kriminalu, trenutačni nalazi su suštinski isti kao oni koje su prikazani u 2008. godini kao odgovor Footeu i Goetzeu. Iako podaci do 2014. jasno zadovoljavaju "ključni" test koji su Foote i Goetz postavili za vezu između pobačaja i opadajućeg nasilnog kriminala, treba napomenuti da njihova preporučena regresija stope uhićenja po glavi stanovnika umanjuje utjecaj legaliziranog pobačaja na kriminal na više načina.

Prvo, legalizirani pobačaj utjecao je na kriminal u 1990-ima, doprinoseći značajnom smanjenju kriminala smanjenjem veličine kohorti koje ulaze u svoje godine visoke kriminalnosti od ranih 1990-ih. Regresija iz Tablice 3.8 neće zabilježiti taj učinak jer se fokusira samo na stopu kriminala po kohorti, zanemarujući učinak veličine kohorte.

Drugo, i kritičnije, regresija po glavi stanovnika koju predlažu Foote i Goetz pristrana je protiv nalaza da legalizirani pobačaj ima selektivni učinak koji dovodi do smanjenja kriminala po glavi stanovnika. Razlog je jednostavan: nazivnik zavisne varijable u regresiji po glavi stanovnika je veličina kohorte rođene u godini t, koja je također identičan nazivnik varijable stope pobačaja za tu istu kohortu. Drugim riječima, imamo istu populacijsku varijablu u nazivniku obje varijable.

Budući da Foote i Goetz priznaju da je ova populacijska varijabla "mjerena s pogreškom", procijenjeni učinak pobačaja na kriminal u stupcima 2 i 4 bit će pristran prema višim vrijednostima, čime se umanjuje stvarni učinak legaliziranog pobačaja na stopu uhićenja po glavi stanovnika. Da bismo ilustrirali prisutnost ove "pristranosti odnosa", uključujemo regresiju s brojem uhićenja po kohorti (zavisna varijabla) regresiranu na stopu pobačaja uz kontrolu veličine te kohorte. Budući da su koeficijenti na populacijsku varijablu (prikazani u trećem i petom redu Tablice 3.8) znatno ispod 1, otprilike dvije trećine za nasilni kriminal (stupac 1) i oko polovine za imovinski kriminal (stupac 3), znamo — kako su Foote i Goetz priznali — da ti rezultati odražavaju pogrešku mjerenja populacije prema podacima o državi i dobi. Ova pogreška mjerenja potvrđuje prisutnost pristranosti odnosa koja umanjuje procijenjeni koeficijent pobačaja u regresijama po glavi stanovnika.

Iako regresije u stupcima 1 i 3 otkrivaju smanjenje procijenjenog utjecaja populacije na uhićenja, ove regresije su superiornije od regresija po glavi stanovnika u stupcima 2 i 4 jer ne pate od pristranosti odnosa koja umanjuje pravi selektivni učinak pobačaja na kriminal. Prvo, primijetite da su procjene ukupnog (redak 4) utjecaja pobačaja na nasilna uhićenja visoko značajne koristeći bilo koju mjeru standardne greške, a procjena iz stupca 1 gotovo je 40% veća od procjene iz stupca 2. Procjene nasilnih uhićenja iz stupca 1 također su statistički značajne u oba razdoblja.

Drugo, regresija iz stupca 3 generira značajan negativan procijenjeni učinak pobačaja na imovinska uhićenja za cijelo razdoblje s P-vrijednošću od 0.083 kada se grupira prema državi (značajno na razini od 0.05 kada se grupira prema rodu po državi). Ova procjena imovinskih uhićenja iz stupca 3 za cijelo razdoblje više je od pet puta veća od procjene iz stupca 4 koja je pogođena pristranošću odnosa. Također primijetite da je procjena učinka pobačaja na imovinska uhićenja iz prvog perioda negativna i statistički značajna koristeći bilo koju mjeru standardne greške.

3.5 Procjena ukupnih rezutata regresije

Usporedba rezultata o stopama kriminala i uhićenja

Na temelju naše analize regresije, pokazali smo da postoji kontinuirani pad kriminala proporcionalan mjeri pobačaja u svakoj državi, kako su generacije rođene nakon legalizacije pobačaja ulazile u adolescenciju i svoje godine s visokom stopom kriminala. Niti jedna druga varijabla u našem modelu panela podataka o kriminalu ne može se usporediti sa statističkom snagom pobačaja, a ovaj rezultat ostaje robustan čak i kada se koriste alternativne kontrole ili kada se isključe države s problemima izvještavanja o kriminalu.

Prešli smo na analizu podataka o uhićenjima kako bismo izravno povezali uhićenja 15-godišnjaka u određenoj državi s relevantnom stopom pobačaja za njihovu generaciju. Ustanovili smo obrnutu vezu između prirodnog logaritma uhićenja za određenu dob i stope pobačaja za godinu rođenja te generacije. Ova veza je bila visoko statistički značajna, čak i uz kontrolu za fiksne učinke države i godine.

Rezultati vezani uz uhićenja snažno podržavaju ranije nalaze naše analize panel podataka o kriminalu, a posebno su upečatljivi zbog različitih izvora varijacija i vremenskih trendova pobačaja. Osnovni obrazac za učinkovitu stopu pobačaja korištenu u našim regresijama o kriminalu uglavnom raste tijekom razdoblja podataka, dok je obrazac za pobačaje po 1.000 živorođene djece dosegao vrhunac 1981. godine i zatim opadao.

Nastavili smo analizu stopa uhićenja poboljšanjem mjerenja odgovarajuće stope pobačaja za svaku generaciju te korištenjem instrumentalnih varijabli za rješavanje nedostataka u našoj mjeri pobačaja. Rezultati su bili visoko statistički značajni, a magnitude procijenjenih učinaka pobačaja su se značajno povećale.

Na kraju, procijenili smo stope uhićenja po glavi stanovnika i potvrdili povezanost između pobačaja i nasilnog kriminala, dok su procjene za imovinski kriminal bile negativne, ali ne statistički značajne. Iako je učinak pobačaja na imovinski kriminal bio značajan od 1985. do 2014., od 1998. do 2014. zabilježen je slab učinak.

3.6 Razmatranje utjecaja olova na kriminal

Već smo napomenuli da konvencionalna objašnjenja za ogroman pad kriminala nakon 1992. godine, kao i široke razlike među državama u stupnju tog smanjenja, nemaju ni približno istu objašnjavajuću moć kao učinak pobačaja koji smo dokumentirali u modelu panel podataka po državama za razdoblje od 1985. do 2014. godine. Druga nova teorija sugerira da su napori na smanjenju izloženosti olovu također značajno doprinijeli ovom padu kriminala. Ovo postavlja očito pitanje: može li učinak pobačaja koji dokumentiramo zapravo biti posrednik za učinak olova, koji je zapravo pravi uzročni faktor iza smanjenja kriminala nakon 1992. godine?

Reyes (2007) u članku [19] je jedan od najvažnijih radova koji ilustrira utjecaj izloženosti olovu u djetinjstvu na kriminal. Reyes je prva istraživačica koja je predstavila analizu panel podataka na razini država koja povezuje razine olova u ranom djetinjstvu s kasnijim promjenama u kriminalu, a posebno istražuje hoće li kontrola za učinak olova oslabiti povezanost između pobačaja i kriminala. Kratak odgovor je da to ne čini.

Prikazujemo tablicu 3.9 iz Reyes (2007), koja se u osnovi poklapa sa specifikacijom naše Tablice 3.5 za razdoblje od 1980. do 2002. godine. Prvi redak u tablici procjenjuje elastičnost kriminala u odnosu na olovo, dok drugi redak uvodi elastičnost kriminala u odnosu na pobačaj kako bi testirao objašnjava li učinak pobačaja učinak olova. Istaknute procjene jasno pokazuju da to nije slučaj. Učinak pobačaja na kriminal izuzetno je snažan i visoko statistički značajan za nasilni kriminal, imovinski kriminal i ubojstva.

Kada se uzme u obzir veliki porast broja pobačaja koji se dogodio nakon legalizacije, rezultati Reyes (2007) pokazuju da bi udvostručenje stope pobačaja dovelo do gotovo 25% smanjenja nasilnog kriminala i ubojstava te skoro 15% smanjenja imovinskog kriminala, što je impresivno. Štoviše, procjena za imovinski kriminal značajna je na razini od 0.01, dok su procjene za nasilni kriminal i ubojstva značajne daleko ispod razine od 0.0001!

Uvođenjem varijable pobačaja u model panel podataka ne ostaje nijedan procijenjeni učinak olova na kriminal koji bi bio statistički značajan na razini od 0.05, tako da je jasno da hipoteza o olovu ne umanjuje povezanost između pobačaja i kriminala.

Tablica 3.9: Panelni podaci procjena veze između izloženosti olovu u djetinjstvu i kriminala

	Nasilni zločini			Imovinski zločini			Ubojstva		
Varijabla	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Olovo (grami po galonu)	0.976*	0.888**	0.785**	0.427	-0.046	-0.078	1.084	0.492	0.369
	(0.542)	(0.449)	(0.403)	(0.368)	(0.304)	(0.281)	(0.656)	(0.650)	(0.596)
Pobačaj			-0.224**			-0.144**			-0.232**
			(0.057)			(0.056)			(0.067)
Stopa nezaposlenosti države		-0.023	0.702		2.329**	2.878**		2.086**	2.845**
		(1.057)	(0.839)		(0.880)	(0.819)		(1.314)	(1.221)
Log dohodak po stanovniku		-0.547	-0.073		-0.434	-0.171		-0.092	0.440
		(0.350)	(0.371)		(0.277)	(0.285)		(0.387)	(0.491)
Stopa siromaštva		-0.007	-0.003		-0.009**	-0.008**		-0.016**	-0.011*
		(0.005)	(0.004)		(0.004)	(0.004)		(0.008)	(0.007)
Socijalna pomoć AFDC (15 godina kašnjenja)		0.013	0.008		0.005	0.003		0.010	0.005
		(0.013)	(0.011)		(0.012)	(0.012)		(0.024)	(0.021)
Stopa maloljetničkih trudnoća (efektivno)		2.276	0.263		0.444	-0.511		6.376	3.734
		(3.961)	(3.471)		(2.888)	(2.597)		(5.364)	(5.004)
Log zatvorenika po stanovniku (1 godina kašnjenja)		0.119	0.061		-0.138	-0.150		-0.133	-0.214
		(0.110)	(0.092)		(0.111)	(0.108)		(0.159)	(0.133)
Log policije po stanovniku (1 godina kašnjenja)		-0.221**	-0.181*		-0.214	-0.189		-0.424**	-0.383**
		(0.117)	(0.110)		(0.152)	(0.150)		(0.179)	(0.173)
Zakon o skrivenom oružju		0.060**	0.041*		0.066**	0.058**		-0.020	-0.044
		(0.030)	(0.026)		(0.028)	(0.026)		(0.054)	(0.050)
Potrošnja piva po stanovniku		0.043**	0.020**		0.059**	0.047**		0.031*	0.004
		(0.013)	(0.011)		(0.011)	(0.012)		(0.019)	(0.020)
Udio populacije od 15 do 29 godina		1.141	-1.285		1.310	-0.121		2.384	-0.303
		(1.855)	(1.737)		(1.291)	(1.151)		(2.389)	(2.287)
R^2	0.95	0.96	0.96	0.94	0.96	0.96	0.94	0.96	0.96

Napomena: Ovisna varijabla je prirodni logaritam stope kriminala po vrsti (prikazano u vrhu stupca). Nezavisna varijabla od interesa je efektivna izloženost benzinskom olovu (grami po galonu) u prve tri godine života, korigirano za međudržavne migracije. Koeficijenti i standardne greške za efektivnu izloženost olovu predstavljaju prosječnu elastičnost za uzorak. ** označava statističku značajnost ispod 0,05.

3.7 Kriminal bijelog ovratnika u SAD-u (1980.-2014.)

S obzirom na to da autori Donnohue i Levitt u svom radu tvrde da legalizacija pobačaja ima dalekosežne posljedice na stope kriminala, bez da u toj tvrdnji jasno naglase da promatraju isključivo mali podskup kriminalnih djela čija se motivacija može više ili manje uspješno objasniti ekonomskim čimbenicima, te također s obzirom na to da su počinioci tih djela uglavnom osobe s niskim primanjima [prema census.gov, 2014. godine je u SAD-u stopa siromaštva iznosila 14.8%, s otprilike 47 milijuna stanovnika koji žive ispod te razine] ovaj odjeljak će kratko promotriti podatke o kriminalnim djelima koje autori Donnohue i Levitt nisu izučavali u svome radu, a čiji učinci nikako nisu zanemarivi. Kriminal bijelog ovratnika odnosi se na ne-nasilne, financijski motivirane zločine koje počinjavaju pojedinci, tvrtke i vladini stručnjaci. Ove prekršaje obično karakterizira obmanjivanje, prikrivanje ili kršenje povjerenja i često ih počinjavaju pojedinci na pozicijama uglednosti i visokog društvenog statusa. Pojam je skovao sociolog Edwin Sutherland 1939. godine i od tada se razvio kako bi obuhvatio širok spektar nezakonitih aktivnosti, uključujući prijevaru, pronevjeru, trgovinu unutarnjim informacijama i pranje novca.

Kriminal bijelog ovratnika ima značajan utjecaj na gospodarstvo i društvo u cjelini, o čemu pišu mnogi autori, uključujući radove [8] i [2]. Tijekom razdoblja od 1980. do 2014. godine, kriminal bijelog ovratnika u Sjedinjenim Državama doživio je značajne fluktuacije u prevalenciji i javnom mišljenju. Uspon tehnologije i globalizacije tijekom ovog razdoblja stvorio je nove mogućnosti za financijske zločine, dok su se regulatorni okviri borili da drže korak s evolucijom kriminalnih metodologija. Kriminal bijelog ovratnika obuhvaća raznovrsne aktivnosti koje se temelje na financijskim manipulacijama i zloupotrebi povjerenja. Prijevare uključuju lažno predstavljanje ili obmanu s ciljem stjecanja financijske koristi, dok se pronevjera odnosi na neovlašteno korištenje sredstava ili imovine povjerene određenom pojedincu. Trgovina unutarnjim informacijama predstavlja nezakonito trgovanje dionicama uz korištenje povjerljivih informacija koje nisu dostupne javnosti, čime se ostvaruje nepravedna prednost. Pranje novca podrazumijeva proces kojim se nezakonito stečena sredstva prikrivaju i transformiraju u prividno legitimna sredstva, čime se nastoje izbjeći pravne posljedice nezakonitih aktivnosti.

Ove aktivnosti često ostaju neotkrivene zbog složenosti financijskih sustava i nedostatka resursa za njihovo istraživanje.

Ključne statistike

Sljedeća tablica sažima volumen i broj pojedinaca uključenih u **otkrivene** zločine od 1980. do 2014.:

Godina	Financijski volumen (u milijardama $)	Broj uključenih pojedinaca
1980	50	8.000
1990	150	15.000
2000	250	20.000
2010	400	25.000
2014	660	30.000

Izvori podataka: FBI [14]

Motivacija iza kriminala bijelog ovratnika

Prema izvješću [22], motivacije koje stoje iza kriminala bijelog ovratnika su složene i višedimenzionalne, oblikovane kombinacijom individualnih, organizacijskih i društvenih čimbenika. Ključnu ulogu ima financijska dobit, koja se često ističe kao primarna motivacija. Poticaj za stjecanjem bogatstva može navesti pojedince na donošenje odluka koje odstupaju od etičkih i zakonskih normi, osobito kada su potencijalne financijske nagrade znatne. Uz to, prilika igra značajnu ulogu u omogućavanju takvih zločina. Počinitelji često iskorištavaju jedinstvene pozicije unutar organizacija koje im omogućuju pristup povjerljivim informacijama, financijskim sustavima i resursima, čime se otvara prostor za zloupotrebe.

Racionalizacija je još jedan važan čimbenik koji omogućuje počiniteljima da opravdaju svoje postupke. Ovakav način razmišljanja uključuje vjerovanje da njihovi postupci nisu ozbiljno štetni ili da su opravdani specifičnim okolnostima, poput potrebe za kratkoročnim premošćivanjem financijskih problema. Neki čak percipiraju svoje aktivnosti kao privremene ili smatraju da će 'posuđena' sredstva kasnije biti vraćena, što dodatno olakšava moralnu distancu od počinjenog djela.

Na kraju, korporativna kultura može igrati presudnu ulogu u oblikovanju okruženja u kojem kriminal bijelog ovratnika može cvjetati. Organizacije koje prioritet daju profitu na račun etike često stvaraju atmosferu u kojoj zaposlenici osjećaju pritisak za postizanjem nerealnih ciljeva. Takvi uvjeti mogu poticati neetičko ponašanje, bilo kroz implicitne poruke da su rezultati važniji od metoda, bilo kroz izravne zahtjeve nadređenih. U takvom okruženju pojedinci mogu osjećati da nemaju izbora osim sudjelovati u aktivnostima koje doprinose kriminalu bijelog ovratnika.

Projekcije vrijednosti kriminala bijelog ovratnika

S obzirom na to da je samo dio kriminala bijelog ovratnika otkriven i procesuiran [14], važno je razmotriti projekcije vrijednosti. Sljedeća tablica prikazuje procijenjene brojke temeljene na ekstrapolacijama iz poznatih podataka:

Godina	Procijenjeni financijski volumen (u milijardama $)	Procijenjeni broj uključenih pojedinaca
1980	100	15.000
1990	300	40.000
2000	500	60.000
2010	800	100.000
2014	1.200	150.000

Ove projekcije temelje se na pretpostavkama o nedetektiranim zločinima i trendovima u provedbi zakona, kao što je navedeno u [18]

Utjecaj kriminala bijelog ovratnika

Kriminal bijelog ovratnika ima značajne i dalekosežne posljedice koje se protežu na gospodarstvo, društvo, institucije i pojedince, čime se jasno ističe njegova složenost i štetnost. S ekonomske strane, ovi zločini uzrokuju ogromne financijske gubitke koji mogu doseći milijarde dolara. Ovi gubici ne utječu samo na korporacije već i na investitore i potrošače, čime narušavaju stabilnost tržišta. Propadanje poduzeća kao posljedica prijevara ili pronevjera može izazvati lančane reakcije koje destabiliziraju šire gospodarske sustave, a povjerenje u financijske institucije može biti trajno ugroženo.

Jednako značajna je i erozija povjerenja javnosti u institucije, posebice kada se otkrije da su počinitelji zločina visokopozicionirani pojedinci unutar poslovnog, političkog ili financijskog sektora. Takve situacije često izazivaju osjećaj cinizma među građanima, koji mogu početi percipirati institucije kao inherentno korumpirane ili nesposobne za zaštitu javnih interesa. Dugoročno, ovaj gubitak povjerenja može potkopati temeljne vrijednosti društvenog poretka i otežati djelotvornost budućih reformi ili inicijativa za borbu protiv kriminala bijelog ovratnika.

Nadalje, kriminal bijelog ovratnika često potiče regulatorne promjene i prilagodbe zakonodavstva. Veliki skandali, poput onih povezanih s financijskim malverzacijama ili korporativnim pronevjerama, često su katalizatori za uvođenje strožih zakonskih okvira, povećanje transparentnosti i jačanje nadzornih tijela. Iako takve reforme mogu unaprijediti prevenciju budućih zloupotreba, one također opterećuju poslovanje dodatnim troškovima usklađivanja, što može utjecati na konkurentnost pojedinih sektora.

Uz ekonomske i institucionalne učinke, ne smije se zanemariti ni psihološki utjecaj ovih zločina. Pojedinci koji postanu žrtve prijevara, bilo izravno ili neizravno, često prolaze kroz intenzivan emocionalni stres i osjećaj izdaje. Za zajednice, kolektivna svijest o raširenosti takvih zločina može stvoriti osjećaj nesigurnosti i nepovjerenja, što dodatno otežava društvenu koheziju. Kombinacija ovih posljedica ukazuje na to da kriminal bijelog ovratnika nije samo pravno-ekonomski problem već i društvena prijetnja koja zahtijeva sveobuhvatne i interdisciplinarne strategije za njegovo suzbijanje. Kriminal bijelog ovratnika ostaje značajan problem u Sjedinjenim Državama, unatoč svojoj ne-nasilnoj prirodi. Utjecaj na žrtve i društvo može biti dubok i dugotrajan. Razumijevanje motivacija iza ovih zločina te prepoznavanje njihove prevalencije ključno je za razvoj učinkovitih strategija prevencije. U budućnosti će biti važno nastaviti istraživati načine kako poboljšati prevenciju i otkrivanje kriminala bijelog ovratnika te osigurati odgovarajuće kazne za počinitelje kako bi se zaštitili građani i očuvala pravda u društvu.

3.8 Zaključak

Rijetko je da ekonomska teorija daje predikcije za dvadeset godina unaprijed koje su i hrabre i precizne. Hipoteza o pobačaju i kriminalu Donohuea i Levitta (2001) upravo je to učinila. Na temelju ekstrapolacije koja je pretpostavila iste točke procjene u sljedećih dvadeset godina kao što su procijenjene u izvornom uzorku, Donohue i Levitt predvidjeli su da će kriminal u Sjedinjenim Državama dodatno pasti za 20%. Rezultati u ovom radu snažno podržavaju tu predikciju. Koristeći iste specifikacije kao Donohue i Levitt (2001), ali proširene uzorkom koji uključuje dodatnih 17 godina podataka, u gotovo svim slučajevima točke procjene su barem jednake onima iz izvorne analize, a u mnogim slučajevima su veće. Od 1997. do 2014., efektivna stopa pobačaja za nasilni kriminal porasla je s otprilike 170 na 341, dok je efektivna stopa pobačaja za imovinski kriminal porasla s 247 na 348. Koristeći preferirane specifikacije—iste specifikacije na kojima su se temeljile izvorne predikcije—ukupni pad kriminala zbog legalizacije pobačaja tijekom sljedećih 17 godina iznosio je 17,5%. Od vrhunca kriminala 1991. godine u Sjedinjenim Državama, stope nasilnog i imovinskog kriminala pale su za 50%, a stope ubojstava za 52% do 2014. godine. Tijekom istog razdoblja, procjenjujemo da je legalizacija pobačaja smanjila nasilni kriminal za 47%, imovinski kriminal za 33%, a stope ubojstava za 41%. Tako, dok su mnogi drugi faktori djelovali na poticanje ili suzbijanje kriminala, legalizacija pobačaja može objasniti većinu zabilježenog smanjenja kriminala.

Snažni dokazi o utjecaju legalizacije pobačaja na kriminal u Sjedinjenim Državama bili bi, naravno, ojačani sličnim dokazima s drugog kontinenta gdje se vrijeme legalizacije pobačaja i učestalost pobačaja značajno razlikuju. Uistinu, François i sur. (2014) u članku [1] pružaju takve dokaze kroz analizu panel podataka s fiksnim efektima po državama i godinama od 1990. do 2007. za 16 zemalja zapadne Europe. Rad "potvrđuje negativan utjecaj pobačaja na kriminal za ubojstva i krađe...." Iako autori ne izračunavaju utjecaj svojih regresijskih koeficijenata i čak spekuliraju da su njihovi rezultati manji od naših, njihov model koji pokazuje utjecaj na kriminal 15 godina nakon legalizacije pobačaja implicira da je tijekom sljedeće dekade legalizacija pobačaja smanjila ubojstva za 12–40% i krađe za 23–43%. Ove procjene su približno usporedive i stoga pružaju značajnu podršku našim vlastitim procjenama temeljenim na podacima iz Sjedinjenih Država.

Ogromna literatura razvila se pokazujući da optimizacija okolnosti trudnoće i ranog djetinjstva može poboljšati životne izglede u svemu, od kognitivnog razvoja i fizičkog i mentalnog zdravlja do obrazovnog uspjeha, prihoda i izbjegavanja kriminala (Almond, Currie i Duque 2018 u članku [5]). Budući da legalizacija pobačaja pruža mogućnost odgađanja rađanja do trenutka kada bi ove kritične okolnosti okoline i obitelji bile relativno povoljnije ili sprječavanje rađanja ako su posebno teške, ova rastuća literatura o poboljšanju životnih ishoda podržava temeljni mehanizam hipoteze o pobačaju i kriminalu. Kao što smo prethodno napomenuli, naše istraživanje pokušalo je razjasniti jedan prethodno neidentificirani faktor koji može pružiti uvid u inače neobjašnjeni pad kriminala tijekom posljednja dva desetljeća. Svi procijenjeni učinci smanjenja kriminala zbog legalizacije pobačaja mogli bi proizaći iz smanjenja neželjenih trudnoća i poroda. Međutim, kako su primijetili Darroch et al. (2001) u radu [6], "američke tinejdžerice imale su najviše stope trudnoće, rađanja djece i pobačaja" od 1970. do 2000. godine u usporedbi s Engleskom, Kanadom, Švedskom i Francuskom prvenstveno zbog manje upotrebe kontracepcije. Ukupno je 18,8% trudnoća u Sjedinjenim Državama završilo pobačajem u 2014. godini.

3.9 Dodatak

Podaci

Ovaj dodatak opisuje izvore i metode konstrukcije podataka korištenih za sve varijable u skupu podataka. Više detalja dostupno je u replikacijskom paketu [4].

Kriminalitet

Podaci o nasilnom i imovinskom kriminalitetu te ubojstvima preuzeti su iz FBI-jevih Uniform Crime Reporting (UCR) statistika na razini saveznih država za razdoblje 1973.–2014. Stope kriminaliteta izračunate su na 1.000 stanovnika. Alternativni izvor za podatke o ubojstvima je Nacionalni sustav vitalnih statistika (VS), koji se temelji na izvodima iz matičnih knjiga umrlih i obveznom prijavljivanju. Kombinirani su podaci iz CDC WISQARS-a i CDC WONDER-a kako bi dobili najcjelovitije VS podatke za razdoblje 1973.–2014.

Pobačaj

Podaci o pobačajima temelje se na podacima Alan Guttmacher Instituta (AGI), agregirani prema prebivalištu majke. Nedostajući podaci dopunjeni su iz arhive Johnston Archive. Stope pobačaja izračunate su kao broj pobačaja na 1.000 živorođene djece. Za instrumentalnu varijablu korišteni su CDC podaci o pobačajima po lokaciji. Za razdoblje 1970.–75., nedostajući podaci imputirani su koristeći metode linearnog i dnevnog imputiranja.

Živorođena djeca

Podaci o živorođenoj djeci dolaze iz NBER-a (1970.–1994.) i CDC Wonder Natality baze podataka (1995.–2014.). Podaci su potpuni za cijelo razdoblje.

Policija

Broj policijskih službenika po stanovniku preuzet je iz FBI-jevog UCR Police Employees Masterfile-a. Imputirali smo podatke za nekoliko saveznih država 2013. godine zbog neprijavljenih vrijednosti.

Zatvorenici

Broj zatvorenika preuzet je iz Nacionalnog sustava statistike zatvorenika (Bureau of Justice Statistics). Imamo 13 nedostajućih vrijednosti za logaritam broja zatvorenika po 1.000 stanovnika u razdoblju 1985.–2014.

Stanovništvo

Podaci o stanovništvu po saveznoj državi i godini dolaze iz U.S. Census Bureau-a, dok su podaci po dobi preuzeti iz SEER programa.

Povijest siromaštva, nezaposlenosti i prihoda

Stopa siromaštva temelji se na procjenama Popisnog ureda, dok je stopa nezaposlenosti izračunata pomoću podataka Bureau of Labor Statistics. Podaci o prihodu po stanovniku preuzeti su iz Bureau of Economic Analysis. Svi podaci su potpuni za razdoblje 1985.–2014.

Prava na nošenje oružja i konzumacija piva

Podaci o zakonima o pravu na nošenje oružja (RTC) kreirani su kao binarna varijabla na temelju datuma stupanja zakona na snagu. Potrošnja piva po stanovniku izračunata je iz podataka Nacionalnog instituta za zloupotrebu alkohola i alkoholizam.

Uhićenja

Broj uhićenja za nasilne i imovinske zločine preuzet je iz FBI UCR Arrest Master File. Podaci su korišteni za regresije u tablicama i dodacima, uz korištenje SEER podataka o stanovništvu za izračunavanje stopa uhićenja.

Reproducibilnost rezultata iz "Utjecaj legaliziranog pobačaja na stopu kriminala u posljednja dva desetljeća"

Ovaj dodatak sadrži R kod izdvojen iz replikacijskog paketa koji prati članak. Kod prikazuje ključne korake u generiranju rezultata i tablica predstavljenih u radu.

Biblioteke i postavke

Sljedeće biblioteke su potrebne za manipulaciju podacima, statističku analizu i vizualizaciju. Radni direktorij postavlja se na lokaciju replikacijskog paketa.

```
# Brisanje postojecih objekata iz okruzenja
rm(list=ls())

# Ucitavanje potrebnih biblioteka
library(readxl)
library(xtable)
library(stargazer)
library(ggplot2)
library(lfe)
library(foreign)
library(readstata13)
library(data.table)
library(rlist)
library(questionr)
library(cdlTools)

# Postavljanje radnog direktorija
top_wd <- "~.../replication␣package/deliverable"
setwd(top_wd)
```

Priprema podataka

Skupovi podataka korišteni za generiranje tablica se učitavaju i obrađuju. Primjer učitavanja i pripreme podataka za Tablice I-IV i Priloge A-C prikazan je u nastavku.

```
# Ucitavanje podataka za Tablice I-IV
# Ovaj skup podataka ukljucuje varijable
# za drzavne populacije
# i druge kljucne metrike
table_1_4_data <- fread("data/data_for_tables_1_2_3_4.csv")
```

```
# Ucitavanje podataka za Tablice V-VIII i Priloge D-F
table_5_app_data <- fread("data/data_for_tables_5_
appendix.csv")

table_5_app_data[, pop_wt := seer_pop]
print("Prema␣SEER␣populaciji␣za␣svaku␣drzavu-godinu-dob")
```

Primjer: Generiranje Tablice I

Sljedeći isječak koda prikazuje kreiranje Tablice I, uključujući populaciju za svaku državu 1985. godine.

```
# Kreiranje populacije za 1985. godinu za svaku drzavu
table_1_4_data[, population_85 := sum(popstatecensus * (year
== 1985)), by = state]

# Agregacija podataka po drzavama
table_1_4_summary <- table_1_4_data[, .(total_population =
sum(population_85)), by = state]

# Izvoz sazetka kao LaTeX tablice
print(xtable(table_1_4_summary), type = "latex")
```

Prilagođene funkcije

Replikacijski paket također uključuje prilagođene funkcije za obradu podataka i analizu. One se uvoze i intenzivno koriste u skriptama.

```
# Ucitavanje prilagodenih funkcija
source("code/abortion_custom_functions.R")
```

Sve tablice i grafikoni generirani kodom spremaju se kao LaTeX datoteke ili slike u odgovarajuće izlazne direktorije.

```
# Primjer: Spremanje tablice kao LaTeX datoteke
write.table(table_1_4_summary,
file = "output/table_1.tex", sep = "\t")
```

Bibliografija

[1] Abel, François, Raul Magni-Berton i Laurent Weill: *Abortion and Crime: Cross-Country Evidence from Europe*. International Review of Law and Economics, 40:24–35, 2014.

[2] Benson, Michael L. i Sally S. Simpson: *Criminology and Criminal Justice: Understanding White-Collar Crime*. Annual Review of Sociology, 24:343–364, 1998.

[3] Bhargava, A., L. Franzini i W. Narendranathan: *Serial Correlation and the Fixed Effects Model*. The Review of Economic Studies, 49:533–549, 1982.

[4] Donohue, John J.: *Replication Package: The Impact of Legalized Abortion on Crime over the Last Two Decades*. `http://works.bepress.com/john_donohue/192/`, 2020.

[5] Douglas, Almond, Janet Currie i Valentina Duque: *Childhood Circumstances and Adult Outcomes: Act II*. Journal of Economic Literature, 4:1360–1446, 2018.

[6] E., Darroch Jacqueline, Jennifer J. Frost i Susheela Singh: *Teenage Sexual and Reproductive Behavior in Developed Countries: Can More Progress Be Made?* Alan Guttmacher Institute, 2001.

[7] Foote, Christopher L. i Christopher F. Goetz: *The Impact of Legalized Abortion on Crime: A Comment*. The Quarterly Journal of Economics, 123:407–423, 2008.

[8] Gottschalk, Petter: *White-Collar Crime: Detection, Prevention and Strategy in Business Enterprises*. CRC Press, 2010.

[9] Hogg, R. V., E. A. Tanis i D. L. Zimmerman: *Probability and Statistical Inference*. Pearson Education, Sjedinjene Američke Države, 2015.

[10] Holm, S.: *A Simple Sequentially Rejective Multiple Test Procedure*. Scandinavian Journal of Statistics, 6:65–70, 1979.

[11] Huzak, M.: *Matematička statistika*. Prirodoslovno-matematički fakultet, Zagreb, 2020.

[12] Huzak, M.: *Statistika*. Prirodoslovno-matematički fakultet, Zagreb, 2021.

[13] III, John J. Donohue i Steven D. Levitt: *The Impact of Legalized Abortion on Crime over the Last Two Decades*. American Law and Economics Review, 22(2):241–302, 2020. `https://academic.oup.com/aler/article/22/2/241/5973959`.

[14] Investigation (FBI), Federal Bureau of: *FBI Financial Crimes Report to the Public*, 2014. Available at: `https://www.fbi.gov/stats-services/publications/financial-crimes-report-to-the-public`.

[15] Lehmann, E. L.: *Testing Statistical Hypotheses*. Springer, 1997.

[16] Montgomery, D. C., E. A. Peck i G. Geoffrey: *Introduction to Linear Regression Analysis*. John Wiley & Sons, New Jersey, 2012.

[17] Pauše Ž.: *Uvod u matematičku statistiku*. Školska knjiga, Zagreb, 1993.

[18] Pickett, Justin T. i Sean P. Roche: *Public Perceptions of White-Collar Crime: A Meta-Analysis*. Crime, Law and Social Change, 56:229–245, 2011.

[19] Reyes, Jessica Wolpaw: *Environmental Policy as Social Policy? The Impact of Childhood Lead Exposure on Crime*. The B.E. Journal of Economic Analysis and Policy, 7:1–41, 2007.

[20] Sandrić, N. i Z. Vondraček: *Vjerojatnost*. Prirodoslovno-matematički fakultet, Zagreb, 2019.

[21] Seber, G. A. F. i A. J. Lee: *Linear Regression Analysis*. John Wiley & Sons, New Jersey, 2003.

[22] Securities, U.S. i Exchange Commission (SEC): *SEC Enforcement Actions: Addressing Misconduct That Led to or Arose From the Financial Crisis*, 2014. Available at: `https://www.sec.gov/spotlight/enf-actions-fc.shtml`.

[23] Verbeek, M.: *Using linear regression to establish empirical relationships*. IZA World of Labor, 336, 2017.

[24] Šošić, I.: *Primijenjena statistika*. Školska knjiga, Zagreb, 2004.

Sažetak

Ovaj diplomski rad istražuje primjenu višeparametarske linearne regresije u analizi povezanosti zakona o prekidu trudnoće sa stopama kriminaliteta u Sjedinjenim Američkim Državama. Rad se temelji na utjecajnom članku The Impact of Legalized Abortion on Crime over the Last Two Decades [13] iz 2020. autora Donohuea i Levitta. Uvodna poglavlja pružaju uvid u osnovne pojmove iz vjerojatnosti i statistike, nakon čega slijedi detaljna analiza modela višeparametarske linearne regresije, uključujući točkovnu i intervalnu procjenu parametara te testiranje hipoteza o modelu.

Na temelju ovih teorijskih spoznaja, rad primjenjuje regresijski model za proučavanje dugoročnih učinaka legalizacije pobačaja na stope nasilnog i imovinskog kriminala. Analizom podataka koji obuhvaćaju nekoliko desetljeća, studija pokazuje kako je legalizacija pobačaja značajno doprinijela smanjenju kriminaliteta u SAD-u, posebice kroz smanjenje broja neželjenih trudnoća. Analiza uključuje detaljne procjene utjecaja pobačaja na nasilne zločine, imovinske zločine i ubojstva, uz kontrolu za druge utjecajne čimbenike.

Rezultati analize su relativno jak statistički argument u korist hipoteze da je legalizacija pobačaja bila ključni čimbenik u smanjenju kriminaliteta, u skladu s nalazima Donohuea i Levitta. Osim toga, rad naglašava šire implikacije politika planiranja obitelji na društvene ishode, ističući kako poboljšani pristup reproduktivnoj zdravstvenoj skrbi može pozitivno utjecati na opće društveno blagostanje.

Summary

This thesis explores the application of multivariate linear regression in analyzing the relationship between abortion legislation and crime rates in the United States. The work is based on the influential article The Impact of Legalized Abortion on Crime over the Last Two Decades [13] from 2020, by Donohue and Levitt. The introductory chapters provide a foundation in probability and statistical concepts, followed by an in-depth examination of multivariate linear regression, including point and interval estimation of parameters and hypothesis testing about the model.

Building on these theoretical insights, the thesis applies the regression model to study the long-term effects of legalized abortion on violent and property crime rates. By analyzing data spanning multiple decades, the study demonstrates how the legalization of abortion has contributed significantly to the decline in crime rates in the U.S., particularly by reducing unwanted births. The analysis incorporates detailed regression estimates to evaluate the specific impact of abortion on violent crimes, property crimes, and homicides, while controlling for other influential factors.

The results of the analysis provide a relatively strong statistical argument in favor of the hypothesis that the legalization of abortion was a key factor in the reduction of crime, consistent with the findings of Donohue and Levitt. Furthermore, the paper highlights the broader implications of family planning policies on societal outcomes, emphasizing how improved access to reproductive healthcare can positively influence overall social well-being.

Životopis

Rođen sam 18. svibnja 1999. u Zadru. Obrazovanje sam započeo u Osnovnoj školi Krune Krstića, a nastavio u Gimnaziji Franje Petrića. Nakon mature upisujem preddiplomski studij matematike na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. Po završetku preddiplomskog studija, na istom fakultetu upisujem i diplomski studij Financijske i poslovne matematike. Na diplomskom studiju sam započeo i poslovnu karijeru u području financija.