Statystyczna analiza danych biomedycznych

1 Podstawowe typy planów eksperymentów biomedycznych

Zaczniemy od wprowadzenia pewnych podziałów dla różnych typów badań:

1. Obserwacyjne czy eksperymentalne?

   Badanie obserwacyjne polega na analizie zebranych danych bez możliwości wpływania na nie.

   Badanie eksperymentalne to takie gdzie w kontrolowanych warunkach możemy wpływać na wyniki przez tak zwane interwencje.

2. Prospektywne, retrospektywne, czy krzyżowe?

   Badanie prospektywne to takie w którym dane zbierane są po rozpoczęciu eksperymentu. Zwykle wybieramy dwie grupy różniące się czynnikami ryzyka a następnie patrzymy czy w pewnym przedziale czasowym zajdzie interesujące nas zdarzenie czy nie (zwykle choroba lub zgon). Tak więc w planie prospektywnym znamy początkowe liczności dla grup ryzyka i obserwujemy liczności wystąpień interesujących nas zdarzeń.

   Badanie retrospektywne to takie w którym dysponujemy danymi o interesującym nas zdarzeniu z przeszłości i poszukujemy przyczyn ich zajścia badając w której grupie ryzyka znajdowały się poszczególne osoby. Tak więc w planie retrospektywnym znamy liczbę wystąpień interesujących nas zdarzeń i obserwujemy liczności w poszczególnych grupach ryzyka.

   Badanie krzyżowe to jednoczesna obserwacja czynników ryzyka i wystąpień interesującego nas zdarzenia. Tak więc w planie krzyżowym znamy jednie początkową wielkość próby (zwykle liczbę osób biorących w eksperymencie) a następnie obserwujemy liczności w podgrupach utworzonych przez podział na czynniki ryzyka i wystąpienie interesującego nas zdarzenia.

2 Test medyczny i jego charakterystyki

Rozważmy dwie dychotomiczne zmienne losowe D i T. Przyjmiemy konwencję biomedyczną i potraktujemy zmienną D jako indykator choroby. Zdarzenie $\{D = 1\}$ oznaczane także jako D^+ oznacza, że badana osoba jest chora. Podobnie zdarzenie $\{D = 0\}$ oznaczane jako D^- oznacza, że badana osoba jest zdrowa.

Uwaga 1 *Stwierdzenie, że badana osoba jest zdrowa oznacza tylko, że nie stwierdzono u niej choroby, która jest przedmiotem naszego zainteresowania.*

Drugą zmienną dychotomiczną T interpretować będziemy jako test medyczny, będący podstawą diagnozy. Zdarzenie $\{T = 1\}$ oznaczane także jako T^+ oznacza, że stosując daną procedurę testową uznajemy badaną osobę jako chorą. Podobnie zdarzenie $\{T = 0\}$ oznaczane jako T^- oznacza, że test nie wykrył choroby u badanej osoby.

Uwaga 2 *Sam test medyczny może mieć złożoną strukturę logiczną i wykorzystywać różne szczegółowe parametry kliniczne. W naukach biomedycznych zmienną T interpretuje się jako czynnik ryzyka. Osoba u której test medyczny wykrył chorobę wcale nie musi być chora ale ma większe prawdopodobieństwo że jest chora niż osoba u której test nie wykrył choroby.*

Przyjmiemy następujące oznaczenia zdarzeń dotyczących wskazań testu T:

$$\begin{aligned} \textit{prawdziwie dodatnie - true positive} \quad & TP = D^+ \cap T^+ \\ \textit{prawdziwie ujemne - true negative} \quad & TN = D^- \cap T^- \\ \textit{fałszywie ujemne - false negative} \quad & FN = D^+ \cap T^- \\ \textit{fałszywie dodatnie - false positive} \quad & FP = D^- \cap T^+, \end{aligned}$$

oraz ich prawdopodobieństw *prawdziwie dodatnich - true positive fraction*

$$TPF = P(D^+ \cap T^+)$$

prawdziwie ujemnych - true negative fraction

$$TNF = P(D^- \cap T^-)$$

fałszywie ujemnych - false negative fraction

$$FNF = P(D^+ \cap T^-)$$

fałszywie dodatnich - false positive fraction

$$FPF = P(D^- \cap T^+).$$

Podstawowe charakterystyki testu medycznego T to:

- *Czułość testu - Sens* (sensitivity), - prawdopodobieństwo wykrycia testem T choroby u osoby rzeczywiście chorej. Wyraża się ono wzorem

$$Sens = P(T^+|D^+) = \frac{TPF}{TPF + FNF}.$$

- *Specyficzność testu - Spec* (specifity), - prawdopodobieństwo negatywnego wyniku testu T (nie wykrycia testem choroby) u osoby faktycznie zdrowej określone wzorem

$$Spec = P(T^-|D^-) = \frac{TNF}{TNF + FPF}.$$

- *Skuteczność testu - ACC* (accuracy) - prawdopodobieństwo prawidłowego wskazania testu dane wzorem

$$ACC = P((T^+ \cap D^+) \cup (T^- \cap D^-)) = TPF + TNF.$$

- *Wartość rozpoznawcza potwierdzająca testu - PPV*(positive predictive value), to prawdopodobieństwo zdarzenia, że badana osoba, u której test wskazuje na chorobę, jest faktycznie chora. Jest ona definiowana wzorem

$$PPV = P(D^+|T^+) = \frac{TPF}{TPF + FPF}.$$

- *Wartość rozpoznawcza wykluczająca testu - NPV* (negative predictive value), to prawdopodobieństwo zdarzenia, że badana osoba, u której test nie wskazuje na chorobę, jest faktycznie zdrowa. Określa ją wzór

$$NPV = P(D^-|T^-) = \frac{TNF}{TNF + FNF}.$$

Powyższe charakterystyki powinny być oczywiście jak najwyższe. Zdefiniujmy jeszcze:

- *LR - iloraz wiarygodności - likellihood ratio*

$$LR = \frac{P(T^+|D^+)}{P(T^+|D^-)}$$

Iloraz wiarygodności LR informuje nas, ile razy częściej dodatni wynik testu pojawia się w przypadku populacji chorych w stosunku do populacji zdrowych. Czułość, specyficzność i iloraz wiarygodności są cechami charakterystycznymi testu. Pozostałe charakterystyki zależą od testu i populacji, w której stosujemy test. Jeżeli ten sam test stosujemy w różnych populacjach różniących się między sobą prawdopodobieństwem $P(D^+)$, to wartości predykcyjne danego testu będą różne w tych populacjach. Test charakteryzujący się wysoką czułością i specyficznością stosowany jako test kliniczny może mieć bardzo wysokie wartości predykcyjne potwierdzające i wykluczające. Ten sam test stosowany jako test przesiewowy w populacji z rzadką chorobą będzie miał oczywiście wartość predykcyjną potwierdzającą tym niższą im niższe jest prawdopodobieństwo $P(D^+)$.

3 Miary zależności pomiędzy dychotomicznymi zmiennymi losowymi

Omówimy różne miary powiązań T i D w populacji oraz ich estymatory w różnych schematach próbkowania. Skupimy się tylko na tych miarach, które traktują zmienną D jako zmienną objaśnianą, a T jako zmienną objaśniającą. Wybór miary zależy od celu badania. Najczęściej stosowane miary to **iloraz szans**, **ryzyko względne** oraz **różnica ryzyk**. Aby zbadać wpływ zmiennej objaśniającej T na zmienną objaśnianą D, możemy porównać prawdopodobieństwa D^+ przy różnych stanach zmiennej T, czyli $P(D^+|T^+)$ i $P(D^+|T^-)$. Różnica tych prawdopodobieństw jest nazywana różnicą ryzyk (**Risk Difference**) i oznaczana

$$RD = P(D^+|T^+) - P(D^+|T^-).$$

Jeżeli zmienną T będziemy interpretowali jako czynnik ryzyka, którego obecność zwiększa prawdopodobieństwo choroby, to mnożąc RD przez liczność badanej populacji uzyskujemy oczekiwaną liczbę osób danej populacji, które nie zachorują na rozważaną chorobę, gdy czynnik ryzyka zostanie wyeliminowany. W przypadku, gdy rozważamy rzadką chorobę, to oba prawdopodobieństwa warunkowe $P(D^+|T^+)$ i $P(D^+|T^-)$ mogą być małe, co prowadzi do małej wartości RD. W takiej sytuacji bardziej sugestywną miarą wpływu T na D jest iloraz

$$RR = \frac{P(D^+|T^+)}{P(D^+|T^-)}$$

zwany ryzykiem względnym (**Relative Risk**). W przypadku, gdy $P(D^+|T^+) = 0.03$, a $P(D^+|T^-) = 0.01$, otrzymujemy $RD = 0.02$, co może sugerować brak wpływu zmiennej T na D, natomiast $RR = 3$ wskazuje raczej na istotny wpływ zmiennej T na zmienną D. Zamiast prawdopodobieństw sukcesu można analizować tzw. szansę sukcesu definiowaną jako stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki, czyli

$$\Omega = \frac{p}{1-p}.$$

Iloraz szans OR (**Odds Ratio**) wyraża się wzorem

$$OR = \frac{P(D^+|T^+)/P(D^-|T^+)}{P(D^+|T^-)/P(D^-|T^-)}.$$

W zależności od interpretacji zmiennej zależnej T iloraz szans jest równy ilorazowi szansy wystąpienia choroby wśród osób z pozytywnym wynikiem testu bądź przy obecności czynnika ryzyka przez szansę

wystąpienia choroby wśród osób z negatywnym wynikiem testu bądź przy braku czynnika ryzyka. Łatwo pokazać, że iloraz szans OR można zdefiniować także wzorem

$$OR = \frac{P(T^+|D^+)/P(T^-|D^+)}{P(T^+|D^-)/P(T^-|D^-)}.$$

Można go więc interpretować jako iloraz szansy dodatniego wyniku testu wśród osób chorych przez szansę dodatniego wyniku testu wśród osób zdrowych. Wynika stąd, że w definicji ilorazu szans rozróżnienie pomiędzy zmienną objaśnianą a objaśniającą nie jest konieczne co jest bardzo wygodne w statystycznej analizie danych. Iloraz szans można wyrazić także wzorem

$$OR = \frac{P(D^+ \cap T^+)P(D^- \cap T^-)}{P(D^- \cap T^+)P(D^+ \cap T^-)}.$$

Należy zwrócić uwagę, że jeżeli iloraz szans jest równy 1, to nie ma związku pomiędzy T a D, natomiast, jeżeli nie jest on równy 1, to taki związek istnieje.

4 Estymacja charakterystyk

Rozważmy eksperyment ze zmiennymi T i D, którego wyniki przedstawimy w formie tabeli.

	D^+	D^-
T^+	N_{11}	N_{10}
T^-	N_{01}	N_{00}

- **W eksperymencie prospektywnym** traktujemy liczności brzegowe $N_{11} + N_{10} = n_{1\cdot}$ i $N_{01} + N_{00} = n_{0\cdot}$ dla wierszy jako ustalone. Rozkład liczności rozważanej tablicy czteropolowej jest w tym przypadku produktem dwóch rozkładów dwumianowych $B(n_{1\cdot}, p_{1\cdot})$ i $B(n_{0\cdot}, p_{0\cdot})$.
- **W eksperymencie retrospektywnym** traktujemy liczności brzegowe $N_{11} + N_{01} = n_{\cdot 1}$ i $N_{10} + N_{00} = n_{\cdot 0}$ dla kolumn jako ustalone. Rozkład liczności rozważanej tablicy czteropolowej jest w tym przypadku produktem dwóch rozkładów dwumianowych $B(n_{\cdot 1}, p_{\cdot 1})$ i $B(n_{\cdot 0}, p_{\cdot 0})$.
- **W eksperymencie krzyżowym** traktujemy sumę wszystkich liczności $N_{11} + N_{10} + N_{01} + N_{00} = n$ jako ustaloną. Rozkład liczności komórek to rozkład wielomianowy $W(n, p_{11}, p_{10}, p_{01}, p_{00})$. W eksperymencie krzyżowym możemy również obserwować rozkłady warunkowe traktując liczności dla wierszy jako ustalone lub liczności dla kolumn jako ustalone. Tak więc wszystkie techniki wnioskowania statystycznego dostępne w eksperymencie prospektywnym lub retrospektywnym są dostępne w eksperymencie krzyżowym.

W eksperymencie krzyżowym jesteśmy w stanie estymować odpowiednimi częstościami (czyli estymatorami największej wiarygodności) prawdopodobieństwa dla poszczególnych pól tablicy czteropolowej. W eksperymencie prospektywnym lub retrospektywnym możemy stosownymi częstościami estymować jedynie niektóre prawdopodobieństwa warunkowe.

- **W badaniu prospektywnym** jesteśmy w stanie estymować prawdopodobieństwa:

$$\begin{aligned} \widehat{PPV} = \hat{P}(D^+|T^+) &= \frac{N_{11}}{n_{1\cdot}}, \\ \widehat{NPV} = \hat{P}(D^-|T^-) &= \frac{N_{00}}{n_{0\cdot}}. \end{aligned}$$

  Korzystając z faktu $B(n,p) \sim AN(np, np(1-p))$ oraz z lematu Słuckiego otrzymujemy

$$P\left(PPV \in \left[\widehat{PPV} - \hat{\sigma}_{PPV} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right), \widehat{PPV} + \hat{\sigma}_{PPV} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right]\right) \to 1-\alpha,$$

$$P\left(NPV \in \left[\widehat{NPV} - \hat{\sigma}_{NPV} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right), \widehat{NPV} + \hat{\sigma}_{NPV} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right]\right) \to 1-\alpha,$$

gdzie

$$\hat{\sigma}_{PPV} = \sqrt{\frac{1}{n_{1\cdot}}\left(\widehat{PPV}(1-\widehat{PPV})\right)}, \quad \hat{\sigma}_{NPV} = \sqrt{\frac{1}{n_{0\cdot}}\left(\widehat{NPV}(1-\widehat{NPV})\right)}$$

Uwaga: Można również obliczyć dokładne przedziały ufności korzystając z dystrybuanty rozkładu dwumianowego tzn, Jeśli N jest liczbą zaobserwowanych sukcesów z rozkładu $B(n,p)$ z nieznanym prawdopodobieństwem sukcesu p to dokładnym przedziałem ufności na poziomie $1-\alpha$ dla parametru p jest przedział $[L(N,\alpha), U(N,\alpha)]$, gdzie

$$L(N,\alpha) = \inf\left\{p \in [0,1] : 1 - F_{B(n,p)}(N-1) \geq \frac{\alpha}{2}\right\}, \quad R(N,\alpha) = \sup\left\{\{p \in [0,1] : F_{B(n,p)}(N) \geq \frac{\alpha}{2}\right\}.$$

Tego typu przedziały ufności nazywamy przedziałami Cloppera-Pearsona i należy stosować gdy $N < 5$ lub $n - N < 5$ bo wtedy asymptotyczne przedziały ufności są zbyt niedokładne.

- **W badaniu retrospektywnym** dysponujemy estymatorami

$$\begin{aligned}
\widehat{Sens} = \hat{P}(T^+|D^+) &= \frac{N_{11}}{n_{\cdot 1}}, \\
\widehat{Spec} = \hat{P}(T^-|D^-) &= \frac{N_{00}}{n_{\cdot 0}}, \\
\widehat{LR} = \frac{\hat{P}(T^+|D^+)}{\hat{P}(T^+|D^-)} &= \frac{\widehat{Sens}}{1-\widehat{Spec}}.
\end{aligned}$$

Asymptotyczne przedziały ufności dla czułości i specyficzności konstruujemy analogicznie

$$P\left(Sens \in \left[\widehat{Sens} - \hat{\sigma}_{Sens} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right), \widehat{Sens} + \hat{\sigma}_{Sens} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right]\right) \to 1-\alpha,$$

$$P\left(Spec \in \left[\widehat{Spec} - \hat{\sigma}_{Spec} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right), \widehat{Spec} + \hat{\sigma}_{Spec} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right]\right) \to 1-\alpha,$$

gdzie

$$\hat{\sigma}_{Sens} = \sqrt{\frac{1}{n_{\cdot 1}}\left(\widehat{Sens}(1-\widehat{Sens})\right)}, \quad \hat{\sigma}_{Spec} = \sqrt{\frac{1}{n_{\cdot 0}}\left(\widehat{Spec}(1-\widehat{Spec})\right)}$$

Aby wyznaczyć przedział ufności dla ilorazu wiarygodności obłożymy $\widehat{LR}$ logarytmem i skorzystamy metody delta z której wynika, że $\log(B(n,p)/n) \sim AN(\log p, (1-p)/(np))$ tak więc

$$\log(\widehat{LR}) \sim AN(\log Sens/(1-Spec), \hat{\sigma}^2_{\log LR}),$$

gdzie

$$\hat{\sigma}_{\log LR} = \sqrt{\hat{\sigma}^2_{Sens}/\widehat{Sens}^2 + \hat{\sigma}^2_{Spec}/(1-\widehat{Spec})^2)} = \sqrt{\frac{1-\widehat{Sens}}{N_{11}} + \frac{\widehat{Spec}}{N_{10}}}$$

Czyli

$$P\left(LR \in \left[\widehat{LR}\exp\left(-\hat{\sigma}_{\log LR} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right), \widehat{LR}\exp\left(\hat{\sigma}_{\log LR} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right)\right]\right) \to 1-\alpha$$

W badaniu retrospektywnym nie możemy estymować PPV i NPV bez dodatkowej informacji o $P(D^+) = p$ czyli o częstości chorych w populacji. jeśli jednak mamy taką dodatkową informację (np WHO) to ze wzoru Bayesa

$$PPV = \frac{Sens\ p}{Sens\ p + (1-Spec)(1-p)}$$

$$NPV = \frac{Spec\ (1-p)}{Spec\ (1-p) + (1-Sens)p}$$

Estymujemy więc powyższe wielkości:

$$\widehat{\widetilde{PPV}} = \frac{\widehat{Sens}\ p}{\widehat{Sens}\ p + (1-\widehat{Spec})(1-p)}$$

$$\widehat{\widetilde{NPV}} = \frac{\widehat{Spec}\ (1-p)}{\widehat{Spec}\ (1-p) + (1-\widehat{Sens})p}$$

Przedział ufności możemy otrzymać podobnie jak dla LR wykorzystując tym razem transformację logitową

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right).$$

Tak więc

$$\text{logit}(\widehat{\widetilde{PPV}}) = \log\left(\frac{p}{1-p}\right) + \log(\widehat{Sens}) - \log(1-\widehat{Spec})$$

$$\text{logit}(\widehat{\widetilde{NPV}}) = \log\left(\frac{1-p}{p}\right) + \log(\widehat{Spec}) - \log(1-\widehat{Sens})$$

Asymptotyczna odchylenia standardowe obliczone z metody delta to

$$\hat{\sigma}_{\text{logit}(PPV)} = \sqrt{\frac{1}{n_{\cdot 1}}\frac{1-\widehat{Sens}}{\widehat{Sens}} + \frac{1}{n_{\cdot 0}}\frac{\widehat{Spec}}{1-\widehat{Spec}}}$$

$$\hat{\sigma}_{\text{logit}(NPV)} = \sqrt{\frac{1}{n_{\cdot 1}}\frac{\widehat{Sens}}{1-\widehat{Sens}} + \frac{1}{n_{\cdot 0}}\frac{1-\widehat{Spec}}{\widehat{Spec}}}.$$

Czyli

$$P\left(PPV \in \left[\frac{e^{\text{logit}(\widehat{\widetilde{PPV}}) - \hat{\sigma}_{\text{logit}(PPV)}F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}}{1+e^{\text{logit}(\widehat{\widetilde{PPV}}) - \hat{\sigma}_{\text{logit}(PPV)}F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}}, \frac{e^{\text{logit}(\widehat{\widetilde{PPV}}) + \hat{\sigma}_{\text{logit}(PPV)}F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}}{1+e^{\text{logit}(\widehat{\widetilde{PPV}}) + \hat{\sigma}_{\text{logit}(PPV)}F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}}\right]\right) \to 1-\alpha$$

$$P\left(NPV \in \left[\frac{e^{\text{logit}(\widehat{\widetilde{NPV}}) - \hat{\sigma}_{\text{logit}(NPV)}F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}}{1+e^{\text{logit}(\widehat{\widetilde{NPV}}) - \hat{\sigma}_{\text{logit}(NPV)}F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}}, \frac{e^{\text{logit}(\widehat{\widetilde{NPV}}) + \hat{\sigma}_{\text{logit}(NPV)}F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}}{1+e^{\text{logit}(\widehat{\widetilde{NPV}}) + \hat{\sigma}_{\text{logit}(NPV)}F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}}\right]\right) \to 1-\alpha$$

- **W badaniu krzyżowym** dysponujemy wszystkimi estymatorami dostępnymi w eksperymencie prospektywnym i retrospektywnym. Ponadto możemy estymować prawdopodobieństwa

$$\begin{aligned}
\widehat{TPF} = \hat{P}(D^+ \cap T^+) &= \frac{N_{11}}{n},\\
\widehat{FNF} = \hat{P}(D^+ \cap T^-) &= \frac{N_{01}}{n},\\
\widehat{FPF} = \hat{P}(D^- \cap T^+) &= \frac{N_{10}}{n},\\
\widehat{TNF} = \hat{P}(D^- \cap T^-) &= \frac{N_{00}}{n},\\
\widehat{ACC} = \widehat{TPF} + \widehat{TNF} &= \frac{N_{11}+N_{00}}{n}.
\end{aligned}$$

Przedziały ufności otrzymujemy z asymptotyki rozkładu dwumianowego

$$P\left(X \in \left[\widehat{X} - \hat{\sigma}_X F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right), \widehat{X} + \hat{\sigma}_X F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right]\right) \to 1-\alpha,$$

gdzie

$$\hat{\sigma}_X = \sqrt{\frac{1}{n}\left(\widehat{X}(1-\widehat{X})\right)}$$

podstawiając za X odpowiednio TPF, FNF, FPF, TNF, ACC.

5 Estymacja różnicy ryzyk i ryzyka względnego

Różnicę ryzyk oraz ryzyko względne możemy estymować w eksperymencie prospektywnym lub krzyżowym. W eksperymencie retrospektywnym mając wiedzę o $P(D^+)$ da się teoretycznie estymować te wielkości jednak w praktyce się tego nie robi. Stosuje się po prosu iloraz szans który jest odporny na rodzaj eksperymentu. Tak więc w eksperymencie prospektywnym i krzyżowym mamy

$$\widehat{RD} = \widehat{PPV} + \widehat{NPV} - 1,$$

$$\widehat{RR} = \frac{\widehat{PPV}}{1-\widehat{NPV}},$$

$$P\left(RD \in \left[\widehat{RD} - \hat{\sigma}_{RD} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right), \widehat{RD} + \hat{\sigma}_{RD} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right]\right) \to 1-\alpha,$$

$$P\left(RR \in \left[\widehat{RR}\exp\left(-\hat{\sigma}_{\log RR} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right), \widehat{RR}\exp\left(\hat{\sigma}_{\log RR} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right)\right]\right) \to 1-\alpha,$$

gdzie

$$\hat{\sigma}_{RD} = \sqrt{\hat{\sigma}_{PPV} + \hat{\sigma}_{NPV}}$$

oraz analogicznie jak dla LR

$$\hat{\sigma}_{\log RR} = \sqrt{\frac{1-\widehat{PPV}}{N_{11}} - \frac{\widehat{NPV}}{N_{01}}}.$$

6 Estymacja ilorazu szans

- **W badaniu prospektywnym**

$$\widehat{OR} = \frac{\hat{P}(D^+|T^+)/\hat{P}(D^-|T^+)}{\hat{P}(D^+|T^-)/\hat{P}(D^-|T^-)} = \frac{N_{11}N_{00}}{N_{01}N_{10}}.$$

- **W badaniu retrospektywnym**

$$\widehat{OR} = \frac{\hat{P}(T^+|D^+)/\hat{P}(T^-|D^+)}{\hat{P}(T^+|D^-)/\hat{P}(T^-|D^-)} = \frac{N_{11}N_{00}}{N_{01}N_{10}}.$$

- **W badaniu przekrojowym**

$$\widehat{OR} = \frac{\hat{P}(D^+ \cap T^+)\hat{P}(D^- \cap T^-)}{\hat{P}(D^- \cap T^+)\hat{P}(D^+ \cap T^-)} = \frac{N_{11}N_{00}}{N_{01}N_{10}}.$$

Widać, że estymator ilorazu szans OR jest niezmienniczy ze względu na plan eksperymentu. Uzyskanie przedziału ufności wymaga zastosowania wielowymiarowej metody delta dla asymptotyki rozkładu wielomianowego. Wprowadźmy oznaczenia:

$$\hat{\mathbf{p}} = [\hat{p}_{11}, \hat{p}_{01}, \hat{p}_{10}, \hat{p}_{00}]^T$$

$$\mathbf{p} = [p_{11}, p_{01}, p_{10}, p_{00}]^T$$

Z CTG mamy

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \sim AN(\mathbf{0}, \mathbf{\Sigma}),$$

gdzie

$$\mathbf{\Sigma} = \begin{bmatrix} p_{11}(1-p_{11}) & -p_{11}p_{01} & -p_{11}p_{10} & -p_{11}p_{00} \\ -p_{11}p_{01} & p_{01}(1-p_{01}) & -p_{01}p_{10} & -p_{01}p_{00} \\ -p_{11}p_{10} & -p_{01}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{00} \\ -p_{11}p_{00} & -p_{01}p_{00} & -p_{10}p_{00} & p_{00}(1-p_{00}) \end{bmatrix}.$$

Iloraz szans określony możemy zapisać w postaci

$$\widehat{OR} = \frac{\hat{p}_{11}\hat{p}_{00}}{\hat{p}_{01}\hat{p}_{10}}.$$

Obkładając logarytmem definiujemy funkcję

$$\varphi(\mathbf{p}) = \log OR = \log p_{11} - \log p_{01} - \log p_{10} + \log p_{00}.$$

Pochodna mocna funkcji $\varphi(\mathbf{p})$ w punkcie $\mathbf{p}$ jest odwzorowaniem liniowym reprezentowanym przez macierz

$$\varphi^{'}(\mathbf{p}) = \begin{bmatrix} \frac{1}{p_{11}} & \frac{-1}{p_{01}} & \frac{-1}{p_{10}} & \frac{1}{p_{00}} \end{bmatrix}$$

Stosując *metodę delta* dla funkcji $\varphi(\mathbf{p})$ otrzymujemy

$$\varphi(\hat{\mathbf{p}}) \sim AN(\varphi(\mathbf{p}), \sigma^2_{\log OR}),$$

gdzie

$$\sigma^2_{\log OR} = \frac{1}{n}[\varphi'(\mathbf{p})]\mathbf{\Sigma}[\varphi'(\mathbf{p})]^T = \frac{1}{n}\left(\frac{1}{p_{11}} + \frac{1}{p_{01}} + \frac{1}{p_{10}} + \frac{1}{p_{00}}\right).$$

Z mocnej zgodności estymatora $\hat{\mathbf{p}}$ wektora $\mathbf{p}$ i z lematu Słuckiego otrzymujemy

$$\log \widehat{OR} \sim AN(\log OR, \hat{\sigma}^2_{\log OR}),$$

gdzie

$$\hat{\sigma}_{\log OR} = \sqrt{\frac{1}{N_{11}} + \frac{1}{N_{01}} + \frac{1}{N_{10}} + \frac{1}{N_{00}}}.$$

Tak więc

$$P\left(OR \in \left[\widehat{OR}\exp\left(-\hat{\sigma}_{\log OR}F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right), \widehat{OR}\exp\left(\hat{\sigma}_{\log OR}F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)\right)\right]\right) \to 1-\alpha$$

Bardzo często przeprowadza się test statystyczny ilorazu szans:

$$H_0 : OR = 1, \quad H_1 : OR \neq 1$$

którego interpretacja jest oczywista. Hipoteza mówi że czynnik ryzyka czy też pozytywny wynik testu nie wpływa na szanse wystąpienia choroby przeciwko alternatywie że wpływa. Korzystając z asymptotycznej normalności $\log \widehat{OR}$ możemy łatwo obliczyć p-wartość takiego testu

$$p_{\text{value}} = 2\left(1 - F_{N(0,1)}\left(\frac{|\log \widehat{OR}|}{\hat{\sigma}_{\log OR}}\right)\right).$$

7 Porównanie wartości diagnostycznych dwóch testów medycznych

W badaniach medycznych bardzo często chcemy porównać ze sobą dwa testy medyczne zwykle po to żeby wybrać lepszy chodź nie zawsze jest to możliwe. Testy medyczne mogą różnić się w sposób istotny charakterystykami w taki sposób, że nie można uznać jednego z nich za lepszy na przykład jeden test medyczny może mieć większą czułość ale za to mniejszą specyficzność od drugiego testu medycznego.

7.1 Porównanie wartości diagnostycznych dla niezależnych testów medycznych

Rozważmy 2 niezależne eksperymenty medyczne przeprowadzone na rozłącznych grupach pacjentów A i B (rozłączne grupy pacjentów gwarantują niezależność wyników). Wyniki obu eksperymentów przedstawiamy w 2 tabelach:

	D_A^+	D_A^-
T_A^+	M_{11}	M_{10}
T_A^-	M_{01}	M_{00}

	D_B^+	D_B^-
T_B^+	N_{11}	N_{10}
T_B^-	N_{01}	N_{00}

Niech $\widehat{CA}$ i $\widehat{BA}$ oznaczają estymatory jakichś charakterystyk CA i CB testów T_A i T_B (np. czułość).

- jeśli $\widehat{CA} \sim AN(CA, \hat{\sigma}^2_{CA})$ i $\widehat{CB} \sim AN(CB, \hat{\sigma}^2_{CB})$ to z niezależności mamy

  $$\widehat{CA} - \widehat{CB} \sim AN(CA - CB, \hat{\sigma}^2_{CA} + \hat{\sigma}^2_{CB}).$$

  Tak więc dla testu statystycznego

  $$H_0 : CA = CB, \quad H_1 : CA \neq CB$$

  możemy obliczyć p-wartość ze wzoru:

  $$p_{\text{value}} = 2\left(1 - F_{N(0,1)}\left(\frac{|\widehat{CA} - \widehat{CB}|}{\sqrt{\hat{\sigma}^2_{CA} + \hat{\sigma}^2_{CB}}}\right)\right)$$

- jeśli dla pewnej funkcji ściśle rosnącej ϕ, $\phi(\widehat{CA}) \sim AN(\phi(CA), \hat{\sigma}^2_{\phi(CA)})$ i $\phi(\widehat{CB}) \sim AN(\phi(CB), \hat{\sigma}^2_{\phi(CB)})$ to z niezależności mamy

  $$\phi(\widehat{CA}) - \phi(\widehat{CB}) \sim AN(\phi(CA) - \phi(CB), \hat{\sigma}^2_{\phi(CA)} + \hat{\sigma}^2_{\phi(CB)}).$$

  Tak więc dla testu statystycznego

  $$H_0 : CA = CB, \quad H_1 : CA \neq CB$$

  możemy obliczyć p-wartość ze wzoru:

  $$p_{\text{value}} = 2\left(1 - F_{N(0,1)}\left(\frac{|\phi(\widehat{CA}) - \phi(\widehat{CB})|}{\sqrt{\hat{\sigma}^2_{\phi(CA)} + \hat{\sigma}^2_{\phi(CB)}}}\right)\right)$$

Na przykład dla ilorazu szans p-wartość testu statystycznego

$$H_0 : OR_A = OR_B, \quad H_1 : OR_A \neq OR_B$$

obliczamy ze wzoru:

$$p_{\text{value}} = 2 - 2F_{N(0,1)}\left(\frac{|\log(M_{11}) - \log(M_{01}) - \log(M_{10}) + \log(M_{00}) - \log(N_{11}) + \log(N_{01}) + \log(N_{10}) - \log(N_{00})|}{\sqrt{\frac{1}{M_{11}} + \frac{1}{M_{01}} + \frac{1}{M_{10}} + \frac{1}{M_{00}} + \frac{1}{N_{11}} + \frac{1}{N_{01}} + \frac{1}{N_{10}} + \frac{1}{N_{00}}}}\right)$$

7.2 Porównanie wartości diagnostycznych dla zależnych testów medycznych

Teraz rozważymy sytuację w której na n pacjentach wykonuje się 2 testy medyczne T_A i T_B. Wyniki przedstawiamy w tabeli

	D^+	D^-
$T_A^+\ T_B^+$	N_{111}	N_{110}
$T_A^+\ T_B^-$	N_{101}	N_{100}
$T_A^-\ T_B^+$	N_{011}	N_{010}
$T_A^-\ T_B^-$	N_{001}	N_{000}

Rozważymy tu od razu plan krzyżowy. Przez p_{ijk} oznaczmy prawdopodobieństwo $p_{ijk} = P(T_A = i, T_B = j, D = k)$. Ponieważ wektor $(N_{111}, ..., N_{000})$ ma rozkład wielomianowy $W(n, p_{111}, ..., p_{000})$ to wektor frakcji

$$\hat{\mathbf{p}} = (\hat{p}_{111}, ..., \hat{p}_{000}) = \left(\frac{N_{111}}{n}, ..., \frac{N_{000}}{n}\right)$$

jest asymptotycznie normalny, a konkretnie

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \sim AN(\mathbf{0}, \mathbf{\Sigma}(\mathbf{p})),$$

gdzie

$$\mathbf{\Sigma}(\mathbf{p}) = \begin{bmatrix} p_{111}(1-p_{111}) & \cdots & -p_{111}p_{000} \\ \vdots & & \vdots \\ -p_{111}p_{000} & \cdots & p_{000}(1-p_{000}) \end{bmatrix}.$$

Charakterystyki Testu A i B mogą być estymowane za pomocą wektora $\hat{\mathbf{p}}$. Na przykład:

$$\widehat{PPV_A} = \hat{P}(D^+|T_A^+) = \frac{\hat{p}_{111} + \hat{p}_{101}}{\hat{p}_{111} + \hat{p}_{110} + \hat{p}_{101} + \hat{p}_{100}}$$

$$\widehat{PPV_B} = \hat{P}(D^+|T_B^+) = \frac{\hat{p}_{111} + \hat{p}_{011}}{\hat{p}_{111} + \hat{p}_{110} + \hat{p}_{011} + \hat{p}_{010}}.$$

Estymator logarytmu każdej charakterystyki jest gładką funkcją wektora $\hat{\mathbf{p}}$, tzn:

$$\log(\widehat{CA}) = \phi_A(\hat{\mathbf{p}}), \quad \log(\widehat{CB}) = \phi_B(\hat{\mathbf{p}}).$$

Różnica

$$\phi(\hat{\mathbf{p}}) = \phi_A(\hat{\mathbf{p}}) - \phi_B(\hat{\mathbf{p}})$$

jest więc asymptotycznie normalna tzn,

$$\log(\widehat{CA}) - \log(\widehat{CB}) \sim AN(\log(CA) - \log(CB), \hat{\sigma}^2_{\log(CA)-\log(CB)}),$$

gdzie

$$\hat{\sigma}_{\log(CA)-\log(CB)} = \sqrt{\frac{1}{n}\phi'(\hat{\mathbf{p}})\mathbf{\Sigma}(\hat{\mathbf{p}})\phi'(\hat{\mathbf{p}})^T}, \qquad \phi'(\mathbf{p}) = \left(\frac{\partial\phi}{\partial p_{111}}(\mathbf{p}), ..., \frac{\partial\phi}{\partial p_{000}}(\mathbf{p})\right).$$

Tak więc dla testu statystycznego

$$H_0 : CA = CB, \quad H_1 : CA \neq CB$$

możemy obliczyć p-wartość ze wzoru:

$$p_{\text{value}} = 2\left(1 - F_{N(0,1)}\left(\frac{|\log(\widehat{CA}) - \log(\widehat{CB})|}{\hat{\sigma}_{\log(CA)-\log(CB)}}\right)\right)$$

8 Krzywa ROC

A co jeśli wynik testu medycznego nie jest zero-jedynkowy? Niech wynik testu medycznego T będzie zmienną losową na podstawie której chcemy podjąć decyzję czy wynik jest pozytywny czy negatywny. Załóżmy że małe wartości T świadczą przeciwko chorobie a duże wartości T za chorobą. Jeśli więc ustalimy pewien próg t to decyzja o wyniku testu wygląda tak:

$$T^- = \{T \leq t\}, \quad T^+ = \{T > t\}.$$

Zauważmy teraz że czułość i specyficzność testu T zależy od progu t poprzez dystrybuanty rozkładów warunkowych $T|D^+$ oraz $T|D^-$:

$$Sens(t) = P(T > t|D^+) = 1 - F_{T|D^+}(t), \quad Spec(t) = P(T \leq t|D^-) = F_{T|D^-}(t)$$

Dobry test medyczny to taki którego czułość i specyficzność jest duża. Ponieważ każda dystrybuanta jest niemalejąca to łatwo widać, że małe t skutkują dużą czułością i małą specyficznością a duże t skutkują małą czułością i dużą specyficznością. Aby zobrazować jak test medyczny reaguje na zmianę progu wymyślono tzw. krzywą ROC (Receiver operating characteristic). Jest to krzywa parametryczna:

$$(x(t), y(t)) = (1 - Spec(t), Sens(t)) = (P(T > t|D^-), P(T > t|D^+)), \quad t \in \mathbb{R}$$

Zauważmy że

$$(x(t), y(t)) \in [0,1]^2, \quad \lim_{t\to-\infty}(x(t), y(t)) = (1,1) \qquad \lim_{t\to\infty}(x(t), y(t)) = (0,0)$$

oraz $x(t)$ i $y(t)$ są nierosnącymi funkcjami t. Dobry test medyczny powinien przyjmować stochastycznie większe wartości T w przypadku choroby niż w przypadku braku choroby co można zapisać w języku stochastycznej dominacji:

$$\forall t \in \mathbb{R} \quad P(T > t|D^+)) \geq P(T > t|D^-))$$

co dla krzywej ROC oznacza, że:

$$(x(t), y(t)) \in \{(x, y) : 0 \leq x \leq y \leq 1\}$$

Tak więc idealny test medyczny to taki dla którego $(x(t), y(t)) = (0, 1)$ (czułość i specyficzność równa 1) natomiast najgorszy test medyczny to taki którego wynik nie ma związku z chorobą czyli T i D są niezależne. W takim przypadku

$$T, D \ - \ \text{niezal.} \Rightarrow P(T > t|D^-) = P(T > t) = P(T > t|D^+) \Rightarrow (x(t), y(t)) \in \{(x, y) : 0 \leq x = y \leq 1\}$$

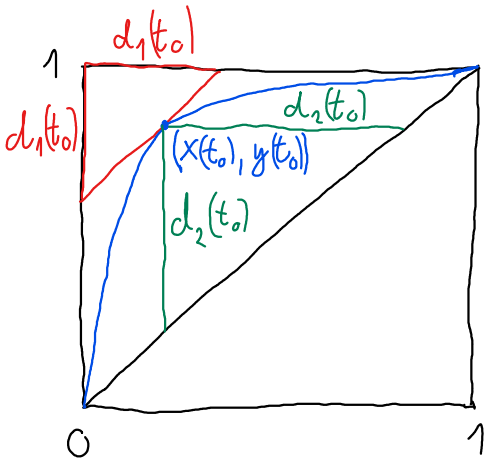
czyli krzywa ROC jest na przekątnej kwadratu $[0, 1]^2$.

8.1 Wybór progu i pole pod krzywą ROC

Jaki próg t daje najlepszą klasyfikację? To zależy od kryterium. Najczęściej stosowanym kryterium to maksimum sumy czułości i specyficzności. Ponieważ $y(t) \geq x(t)$ to $1 \leq Sens(t) + Spec(t) \leq 2$. W związku z tym definiujemy tzw. indeks Youdena

$$YI = \max_{t\in\mathbb{R}} \ Sens(t) + Spec(t) - 1,$$

który przyjmuje wartości z przedziału $[0, 1]$. Optymalny próg według tego kryterium to takie t_0 dla którego $Sens(t_0) + Spec(t_0) - 1 = YI$. Wybór tego progu obrazuje rysunek:



$t_0 = \underset{t}{\operatorname{argmin}}\, d_1(t) = \underset{t}{\operatorname{argmax}}\, d_2(t)$

$d_1(t) = |x(t) - 0| + |y(t) - 1| = 1 + x(t) - y(t) \to \min$

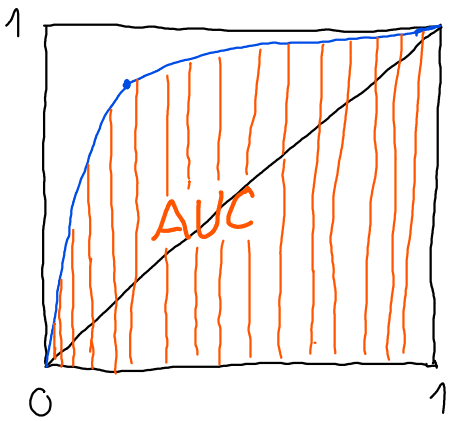
$d_2(t) = y(t) - x(t) = Sens(t) + Spec(t) - 1 \to \max$

$d_2(t_0)$ – Youden's index

Jeśli natomiast chcemy zbadać wartość prognostyczną testu T bez ustalania progu to możemy ją mierzyć za pomocą prawdopodobieństwa zdarzenia, że wartość T bedzie większa dla losowo wybranej osoby chorej niż dla losowo wybranej osoby zdrowej:

$$AUC = \int_{\mathbb{R}} \int_{\mathbb{R}} I(t_1 > t_2) dP^{T|D^+}(t_1) dP^{T|D^-}(t_2)$$

Okazuje się że takie prawdopodobieństwo ma bardzo ładną interpretację graficzną jako pole pod krzywą ROC (AUC - Area under curve) wynika to z prostego przekształcenia całki



$$AUC = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} I(t_1 > t_2) dP^{T|D^+}(t_1) \right) dP^{T|D^-}(t_2) = \int_{\mathbb{R}} Sens(t_2) dP^{T|D^-}(t_2) = \int_{\mathbb{R}} y(t) dP^{T|D^-}(t)$$

8.2 Estymacja krzywej ROC, progu klasyfikacji i AUC

Niech $(T_1, D_1), ..., (T_n, D_n)$ będzie próbą prostą gdzie (T_i, D_i) oznacza wynik testu medycznego a D_i jest dychotomiczna zmienną oznaczającą chorobę bądź jej brak. Ponieważ krzywa ROC to $(x(t), y(t)) = (1 - F_{T|D^-}(t), 1 - F_{T|D^+}(t))$ to naturalnym estymatorem krzywej ROC jest

$$(\hat{x}(t), \hat{y}(t)) = (1 - \hat{F}_{T|D^-}(t), 1 - \hat{F}_{T|D^+}(t)),$$

gdzie

$$\hat{F}_{T|D^-}(t) = \frac{\sum_{i=1}^{n} I(T_i \leq t) I(D_i = 0)}{\sum_{i=1}^{n} I(D_i = 0)}, \qquad \hat{F}_{T|D^+}(t) = \frac{\sum_{i=1}^{n} I(T_i \leq t) I(D_i = 1)}{\sum_{i=1}^{n} I(D_i = 1)}$$

to dystrybuanty empiryczne rozkładów warunkowych. Ponieważ $\hat{x}(t)$ i ($\hat{y}(t)$ są kawałkami stałe to wystarczy obliczyć estymator dla $t \in \{T_1, ..., T_n\}$.

Optymalny próg t_0 ze względu na kryterium indeksu Youdena możemy łatwo estymować jako:

$$\hat{t}_0 = \underset{t \in \{T_1, ..., T_n\}}{\arg\max} = \hat{y}(t) - \hat{x}(t)$$

Ponieważ AUC jest prawdopodobieństwem pewnego zdarzenia estymujemy je jako frakcję:

$$\widehat{AUC} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I(T_i > T_j) I(D_i = 1) I(D_j = 0)}{\sum_{i=1}^{n} \sum_{j=1}^{n} I(D_i = 1) I(D_j = 0)}$$

8.3 Przedział ufności dla AUC

Pomimo że AUC jest prawdopodobieństwem estymowanym przez frakcję to nie łatwo wyznaczyć dla niego przedział ufności. Jeśli frakcja jest liczona jako suma indykatorów niezależnych zdarzeń o tym samym prawdopodobieństwie to korzystając z asymptotyki rozkładu dwumianowego możemy wszystko przybliżać rozkładem normalnym. Jednak tutaj zdarzenia $\{T_i > T_j\}$ są zależne. Istnieją metody kombinatoryczne wyznaczania przedziałów ufności dla AUC jednak są one na tyle skomplikowane, że w praktyce się ich raczej nie stosuje. Najczęściej w praktyce stosuje się przedziały oparte na nieparametrycznym bootstrapie. Przekształćmy nasze oryginalne dane $(T_1, D_1), ..., (T_n, D_n)$ na:

$$(T_{1,0}, ..., T_{n_0,0}), \qquad (T_{1,1}, ..., T_{n_1,1}), \qquad n_0 + n_1 = n,$$

gdzie $(T_{1,0}, ..., T_{n_0,0})$ to wszystkie wyniki testu medycznego dla osób zdrowych a $(T_{1,1}, ..., T_{n_1,1})$ to wszystkie wyniki testu medycznego dla osób chorych. Z tych danych obliczamy estymator $\widehat{AUC}$. Następnie z ciągu $(T_{1,0}, ..., T_{n_0,0})$ losujemy niezależnie z rozkładu jednostajnego (ze zwracaniem) n_0 wartości a z ciągu $(T_{1,1}, ..., T_{n_1,1})$ losujemy niezależnie z rozkładu jednostajnego (ze zwracaniem) n_1 wartości. Otrzymaliśmy w ten sposób jedną replikacje boostrapową:

$$(T^1_{1,0}, ..., T^1_{n_0,0}), \qquad (T^1_{1,1}, ..., T^1_{n_1,1}), \qquad n_0 + n_1 = n,$$

z której obliczamy $\widehat{AUC}_1$. Powtarzamy replikacje niezależnie MC razy otrzymując $\widehat{AUC}_1, ..., \widehat{AUC}_{MC}$. Można udowodnić, że jeśli T jest ograniczoną zmienną losową oraz $\lim_{n\to\infty} n_1/(n_1 + n_0) = c \in (0, 1)$ to

$$\lim_{n\to\infty} \lim_{MC\to\infty} P\left(AUC \in \left[\widehat{AUC}_{(\lfloor MC\ \alpha/2 \rfloor)}, \widehat{AUC}_{(\lfloor MC\ (1-\alpha/2) \rfloor)}\right]\right) = 1 - \alpha.$$

8.4 Test statystyczny dla wartości prognostycznej testu medycznego

Jak sprawdzić czy test medyczny ma jakakolwiek wartość prognostyczną dla choroby. Jeśli nie prezentuje takiej wartości to krzywa ROC leży na przekątnej lub równoważnie (przy założeniu że $y(t) \geq x(t)$) $AUC = 0.5$ lub równoważnie $YI = 0$. Zauważmy że jeśli chcemy testować hipotezę że test medyczny nie

ma wartości prognostycznej przeciwko alternatywie że ją ma możemy to matematycznie zapisać w sposób następujący. Niech

$$F_{T|D^+}(t) = F_{T|D^-}(t) - \Delta(t),$$

gdzie $\Delta(t)$ jest funkcją rzeczywistą nieujemną. Testujemy

$$H_0 : \Delta(t) = 0, \qquad H_1 : \Delta \geq 0 \ \wedge \ \Delta \neq 0$$

Prowadzi to do jednostronnego testu Manna-Whitneya. P-wartość tego testu obliczamy ze wzoru:

$$p = 1 - F_{N(0,1)}(W),$$

gdzie

$$W = \frac{\bar{R}^+ - \frac{(n+1)}{2}}{\sqrt{\frac{M(n+1)}{12N}}}, \quad M = \sum_{i=1}^{n} I(D_i = 0), \quad N = \sum_{i=1}^{n} I(D_i = 1), \quad \bar{R}^+ = \frac{1}{N}\sum_{i=1}^{n} R_i I(D_i = 1)$$

oraz $(R_1, ..., R_n)$ jest wektorem rang dla $(T_1, ..., T_n)$.

9 Regresja logistyczna

A co w sytuacji gdy wynik testu medycznego składa się z wielu pomiarów których wyniki mogą być ilościowe lub jakościowe? Wtedy możemy zamodelować wpływ obserwacji (tzw. predyktorów) na prawdopodobieństwo wystąpienia choroby w następujący sposób:

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{k} \beta_i X_i$$

gdzie $X_1, ..., X_k$ to predyktory ciągłe lub dychotomiczne predyktory grupujące (oznaczające przynależność do danej grupy ryzyka), a nieznane parametry $\beta_0, ..., \beta_k$ modelują liniowy wpływ predyktorów na zlogarytmowaną szansę wystąpienia choroby ($\text{logit}(p) = \log \frac{p}{1-p}$).

Uwaga: Jeśli pacjent należy do jednej z j grup to kodujemy za pomocą $j-1$ zmiennych dychotomicznych $(X_1, ..., X_{j-1})$ gdzie $(0, ..., 0)$ oznacza przynależność do pierwszej grupy (zwykle grupa kontrolna) a $(0, .., 0, 1, 0...,)$ z jedynką na i-tym miejscu oznacza przynależność do $i+1$ grupy.

Zauważmy, że $\text{logit}(p)$ jest ściśle rosnącą funkcją p taką, że:

$$\lim_{p\to 0} \text{logit}(p) = -\infty, \quad \text{logit}\left(\frac{1}{2}\right) = 0, \quad \lim_{p\to 1} \text{logit}(p) = \infty$$

Funkcję odwrotną do funkcji logit oznaczamy przez

$$\text{logistic}(y) = \frac{1}{1+e^{-y}}$$

tak więc prawdopodobieństwo wystąpienia choroby zależy w tym modelu o predyktorów w następujący sposób

$$p = p(\beta_0, ..., \beta_k, X_1, ..., X_k) = \text{logistic}\left(\beta_0 + \sum_{i=1}^{k} \beta_i X_i\right)$$

Niech $D_1, ..., D_n$ będą niezależnymi zmiennymi losowymi zero-jedynkowymi, takimi że

$$D_i \sim B(1, p(\beta_0, ..., \beta_k, X_{1i}, ..., X_{ki}),$$

gdzie $X_{1i}, ..., X_{ki}$ to wyniki pomiarów dla i-tego pacjenta. Mamy tu do czynienia z $k+1$-wymiarowym modelem parametrycznym z nieznanym parametrem $\beta = (\beta_0, ..., \beta_k)$. Parametr ten możemy estymować metodą największej wiarygodności

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{k+1}}{\arg\max} \sum_{i=1}^{n} [D_i \log(p(\beta_0, ..., \beta_k, X_{1i}, ..., X_{ki})) + (1 - D_i) \log(1 - p(\beta_0, ..., \beta_k, X_{1i}, ..., X_{ki}))]$$

9.1 Test istotności parametrów

Część pomiarów może nie mieć związku z chorobą. Zakładając poprawność modelu logistycznego niezależność wyniku danego pomiaru X_i jest równoważna zerowaniu się parametru β_i. Zajmiemy się teraz testem hipotezy:

$$H_0 : \beta_{k_1}, ..., \beta_{k_l} = 0,$$

($(k_1, ..., k_l)$ jest rosnącym podciągiem ciągu $(0, ..., k)$) przeciwko alternatywie że tak nie jest. Wykorzystamy tu twierdzenie o asymptotycznym rozkładzie ilorazu wiarygodności (patrz wykład z Testowania Hipotez Statystycznych). Oznaczmy:

$$L(\beta, X) = \prod_{i=1}^{n} p(\beta_0, ..., \beta_k, X_{1i}, ..., X_{ki})^{D_i} (1 - p(\beta_0, ..., \beta_k, X_{1i}, ..., X_{ki}))^{1-D_i},$$

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^{k+1},\ \beta_{k_1}, ..., \beta_{k_l} = 0}{\arg\max} \sum_{i=1}^{n} [D_i \log(p(\beta_0, ..., \beta_k, X_{1i}, ..., X_{ki})) + (1 - D_i) \log(1 - p(\beta_0, ..., \beta_k, X_{1i}, ..., X_{ki}))]$$

$$T = 2 \log \left(\frac{L(\hat{\beta}, X)}{L(\tilde{\beta}, X)} \right)$$

Przy prawdziwości H_0 statystyka T ma asymptotyczny rozkład χ^2_l. Tak więc p-wartość powyższego testu obliczamy ze wzoru:

$$p_{\text{value}} = 1 - F_{\chi^2_l}(T)$$

9.2 Iloraz szans

Trudno jest ocenić siłę wpływu danego pomiaru X_i na prawdopodobieństwo choroby na podstawie wyestymowanej wartości parametru β_i. P-wartość testu informuje nas o istotności lub jej braku ale nawet istotny wpływ może być słaby (słabszy niż wpływ innego pomiaru). Do oceny siły wpływu pomiaru X_i używa się ilorazu szans. W zależności czy pomiar jest predyktorem ciągłym czy dychotomiczną zmienną grupującą iloraz szans ma różną interpretację.

9.2.1 Predyktor jakościowy

Zajmiemy się najpierw przypadkiem gdy X_i jest zero-jedynkową zmienną grupującą. Oznaczmy

$$X^{0i} = (X_1, ..., X_{i-1}, 0, X_{i+1}, ..., X_k), \quad X^{1i} = (X_1, ..., X_{i-1}, 1, X_{i+1}, ..., X_k).$$

Iloraz szans dla β_i wyraża się wzorem:

$$OR(\hat{\beta}_i) = \frac{\frac{p(\hat{\beta}, X^{1i})}{1 - p(\hat{\beta}, X^{1i})}}{\frac{p(\hat{\beta}, X^{0i})}{1 - p(\hat{\beta}, X^{0i})}} = \frac{e^{\text{logit}(p(\hat{\beta}, X^{1i}))}}{e^{\text{logit}(p(\hat{\beta}, X^{0i}))}} = e^{\hat{\beta}_0 + \sum_{j \in \{1, ...k\} \setminus \{i\}} \hat{\beta}_j X_j + \hat{\beta}_i - \hat{\beta}_0 - \sum_{j \in \{1, ...k\} \setminus \{i\}} \hat{\beta}_j X_j} = e^{\hat{\beta}_i}.$$

Tak więc $OR(\hat{\beta}_i) = e^{\hat{\beta}_i}$ informuje nas ilokrotnie jest większa szansa na wystąpienie choroby w grupie $X_i = 1$ od szansy w grupie $X_i = 0$. Zauważmy że szansa ta nie zależy od pozostałych parametrów co jest bardzo wygodne.

9.2.2 Predyktor ilościowy

Teraz weźmy przypadek gry X_i jest predyktorem ciągłym. Oznaczmy

$$X^{0i} = (X_1, ..., X_{i-1}, X_i, X_{i+1}, ..., X_k), \qquad X^{1i} = (X_1, ..., X_{i-1}, X_i + 1, X_{i+1}, ..., X_k).$$

Iloraz szans dla β_i wyraża się wzorem:

$$OR(\hat{\beta}_i) = \frac{\frac{p(\hat{\beta}, X^{1i})}{1-p(\hat{\beta}, X^{1i})}}{\frac{p(\hat{\beta}, X^{0i})}{1-p(\hat{\beta}, X^{0i})}} = \frac{e^{\text{logit}(p(\hat{\beta}, X^{1i}))}}{e^{\text{logit}(p(\hat{\beta}, X^{0i}))}} = e^{\hat{\beta}_0 + \sum\limits_{j \in \{1,...k\}} \hat{\beta}_j X_j + \hat{\beta}_i - \hat{\beta}_0 - \sum\limits_{j \in \{1,...k\}} \hat{\beta}_j X_j} = e^{\hat{\beta}_i}.$$

Tak więc $OR(\hat{\beta}_i) = e^{\hat{\beta}_i}$ informuje nas ilokrotnie wzrośnie szansa na wystąpienia choroby gdy zwiększymy X_i o jedną jednostkę. Szansa ta nie zależy od pozostałych parametrów tak jak dla pedyktora jakościowego.

Uwaga: Iloraz szans dla predyktora ciągłego zależy od skali. Porównywanie ilorazów szans predyktorów ciagłych jest więc nieco utrudnione bo należy zadbać aby skale były porównywalne.

9.2.3 Przedział ufności dla ilorazu szans

Możemy obliczyć asymptotyczny przedział ufności ilorazu szans korzystając z twierdzenia o asymptotycznym rozkładzie estymatora największej wiarygodności. Tak więc

$$\hat{\beta} \sim AS(\beta, I^{-1}(\hat{\beta})),$$

gdzie $I(\beta)$ jest macierzą informacji Fishera wymiaru $(k+1) \times (k+1)$:

$$I(\beta)_{ab} = -E_\beta \left[\frac{\partial^2}{\partial \beta_a \partial \beta_b} \sum_{i=1}^{n} (D_i \log(p(\beta_0, ..., \beta_k, X_{1i}, ..., X_{ki})) + (1 - D_i) \log(1 - p(\beta_0, ..., \beta_k, X_{1i}, ..., X_{ki}))) \right]$$

$$= \sum_{i=1}^{n} \frac{X_{ai} X_{bi} e^{2\beta_0 + 2\sum\limits_{j=1}^{k} \beta_j X_{ji}}}{\left(1 + e^{\beta_0 + \sum\limits_{j=1}^{k} \beta_j X_{ji}}\right)^3} + \sum_{i=1}^{n} \frac{X_{ai} X_{bi} e^{-2\beta_0 - 2\sum\limits_{j=1}^{k} \beta_j X_{ji}}}{\left(1 + e^{-\beta_0 - \sum\limits_{j=1}^{k} \beta_j X_{ji}}\right)^3},$$

gdzie $X_{0i} = 1$ $(a, b \in \{0, 1..., k\})$. Mamy wiec ładny wzór analityczny na macierz $I(\beta)$ jednak potrzebujemy obliczyć $I(\hat{\beta})$ gdzie $\hat{\beta}$ jest wyznaczone numerycznie. Tak więc macierz $I(\hat{\beta})$ obliczamy numerycznie wstawiając do powyższego wzoru analitycznego numerycznie obliczone $\hat{\beta}$. Dodatkowo musimy tę macierz odwrócić co również zwykle robi się przy pomocy komputera (szczególnie dla dużych k). Po obliczeniu $I^{-1}(\hat{\beta})$ odczytujemy odpowiedni wyraz na przekątnej:

$$\hat{\sigma}_{\beta_j} = \sqrt{I^{-1}(\hat{\beta})_{jj}}$$

Otrzymujemy:

$$P\left(\beta_j \in \left[\hat{\beta}_j - \hat{\sigma}_{\beta_j} F^{-1}_{N(0,1)}\left(1 - \frac{\alpha}{2}\right), \hat{\beta}_j + \hat{\sigma}_{\beta_j} F^{-1}_{N(0,1)}\left(1 - \frac{\alpha}{2}\right)\right]\right) \to 1 - \alpha$$

a z tego wynika że

$$P\left(OR(\beta_j) \in \left[e^{\hat{\beta}_j - \hat{\sigma}_{\beta_j} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}, e^{\hat{\beta}_j + \hat{\sigma}_{\beta_j} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}\right]\right) \to 1 - \alpha$$

9.3 Prognoza i wartość prognostyczna

Zauważmy, że estymator z modelu logitowego

$$\hat{p}_i = p(\hat{\beta}_0, ..., \hat{\beta}_k, X_{1i}, ..., X_{ki})$$

jest jednowymiarową statystyką zamieniającą nam zestaw pomiarów dla i-tego pacjenta na liczbę która im jest wyższa tym większe prawdopodobieństwo wystąpienia choroby. Możemy więc narysować wykres empirycznej krzywej ROC:

$$(\hat{x}(p), \hat{y}(p)) = \left(1 - \frac{\sum_{i=1}^{n} I(\hat{p}_i \leq p) I(D_i = 0)}{\sum_{i=1}^{n} I(D_i = 0)}, 1 - \frac{\sum_{i=1}^{n} I(\hat{p}_i \leq p) I(D_i = 1)}{\sum_{i=1}^{n} I(D_i = 1)} \right)$$

Krzywa ROC dobrze obrazuje wartość prognostyczną pomiarów $X_1, ..., X_k$. W celu uzyskania najlepszej prognozy według kryterium indeksu Youdena obliczamy

$$\hat{p}_0 = \underset{p \in \{\hat{p}_1, ..., \hat{p}_n\}}{\arg\max} = \hat{y}(p) - \hat{x}(p).$$

Prognoza dla i-tego pacjenta na podstawie pomiarów $X_{1i}, ..., X_{ki}$ będzie postaci

$$T_i = I(\hat{p}_i > \hat{p}_0)$$

10 Analiza przeżycia

Zajmiemy się teraz badaniem zależności pomiędzy zmiennymi objaśniającymi (predyktorami) a czasem życia pacjenta. Czasu życia nie należy tu traktować dosłownie. Jest to czas do wystąpienia interesującego nas zdarzenia (np powikłań dla osób chorych). Niech T będzie nieujemną, absolutnie ciągłą względem pewnej miary dominującej μ (zwykle miara Lebesgue'a lub miara licząca) zmienną losową oznaczającą czas życia. Zdefiniujmy podstawowe pojęcia dla zmiennej losowej T. Niech F oznacza dystrybuantę rozkładu T a f gęstość tego rozkładu względem miary μ. Zdefiniujmy podstawowe pojęcia dla analizy przeżycia.

- *Funkcja przeżycia*: $S(t) = P(T \geq t) = 1 - F(t^-)$
- *Hazard*: $h(t) = f_{T|T \geq t}(t) = \frac{f(t)}{S(t)}$
- *Skumulowany hazard*: $H(t) = \int_0^t h(x) d\mu(x)$

Warto również zauważyć, że

$$E(T) = \int_0^\infty S(t) dt$$

10.1 Przypadek rozkładu absolutnie ciągłego względem miary Lebesgue'a

W przypadku rozkładu absolutnie ciągłego względem miary Lebesgue'a $f(t) = -S'(t)$ tak więc

$$h(t) = -\frac{d}{dt} log(S(t))$$

a z tego wynika że

$$S(t) = \exp(-H(t)) = \exp\left(- \int_0^t h(x) dx \right)$$

10.2 Przypadek rozkładu dyskretnego

Rozważmy teraz rozkład dyskretny na zbiorze $\{t_1, t_2, ...\}$ gdzie $0 < t_1 < t_2 < ....$ Wtedy

$$f(t) = \sum_{j=1}^{\infty} P(T = t_j) I(t = t_j) = P(T = t)$$

$$S(t) = \sum_{j|t_j \geq t} f(t_j)$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{P(T = t)}{S(t)} = P(T = t|T \geq t)$$

Pokażemy teraz bardzo ważny związek funkcji przeżycia z hazardem, który wykorzystamy później przy estymacji nieparametrycznej. Wprowadźmy oznaczenia:

$$t_0 = 0, \lambda_k = h(t_k), \quad k \in \mathbb{N}_0$$

Zauważmy, że $\lambda_0 = 0$ bo $P(T = 0) = 0$. Niech

$$t_k(t) = \min\{t_j : j \in \mathbb{N}_0, t_j \geq t\}.$$

Wtedy:

$$S(t) = S(t_k(t)) = \frac{S(t_k(t))}{S(t_0)} = \prod_{j|t_j<t} \frac{S(t_{j+1})}{S(t_j)} = \prod_{j|t_j<t} \frac{S(t_j) - f(t_j)}{S(t_j)} = \prod_{j|t_j<t} (1 - \lambda_j)$$

Uwaga: Dystrybuanta rozkładu dyskretnego jest sumą gęstości. Rozkład ciągły można przybliżać rozkładem dyskretnym zagęszczając nośnik. W granicy suma zmienia się w całkę Lebesgue'a. Jeśli teraz zastosujemy to samo rozumowanie do funkcji przeżycia która jest iloczynem funkcji $1 - h$ to w granicy dostaniemy tak zwaną całkę iloczynową. Nie będziemy jednak rozwijać tej teorii na tym wykładzie.

10.3 Przykłady modeli parametrycznych

W praktyce zdarza się że zakładamy o zmiennej losowej T że pochodzi z jakiejś klasy rozkładów. Najprostsza klasa rozkładów dla T to klasa rozkładów wykładniczych ponieważ hazard jest funkcją stałą:

$$f_\lambda(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}} \Rightarrow h_\lambda(t) = \frac{1}{\lambda}$$

Jest raczej mało realistyczny model w praktyce oznaczający, że ryzyko śmierci nie zmienia się w czasie. Do modelowania czasu życia ludzkiego używa się raczej rozkładów dla których hazard rośnie przynajmniej od pewnego momentu czasowego. Najpopularniejszą klasą rozkładów o rosnącym hazardzie którą modeluje się w pratyce czas życia jest klasa rozkładów Weibulla w której skumulowany hazard oraz zwykły hazard rosną potęgowo:

$$H_{\alpha,\lambda}(t) = \left(\frac{t}{\lambda}\right)^\alpha \Rightarrow h_{\alpha,\lambda}(t) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \Rightarrow S_{\alpha,\lambda}(t) = e^{-\left(\frac{t}{\lambda}\right)^\alpha} \Rightarrow f_{\alpha,\lambda}(t) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{t}{\lambda}\right)^\alpha}$$

10.4 Parametryczna estymacja funkcji przeżycia dla danych cenzurowanych

Zwykle w praktyce mamy do czynienia z danymi cenzurowanymi. Obserwujemy każdego pacjenta przez jakiś okres czasu. Jeśli dla k-ty pacjent w tym okresie czasu zmarł (lub doszło do interesującego nas zdarzenia) to mamy zaobserwowaną długość jego życia od momentu rozpoczęcia obserwacji. Jeśli nie to odnotowujemy, że długość życia pacjenta od momentu rozpoczęcia obserwacji jest większa niż czas obserwacji. Niech $T_1, ..., T_n$ oraz $V_1, ..., V_n$ będą niezależnymi próbami prostymi, gdzie (T_k, V_k) oznaczają

czas życia od momentu rozpoczęcia obserwacji oraz czas obserwacji k-tego pacjenta. Obserwujemy próbę prostą $(T_1^*, D_1), ..., (T_n^*, D_n)$, gdzie:

$$T_k^* = \min\{T_k, V_k\}, \qquad D_k = I(T_k \leq V_k).$$

$(T_k^*, 1)$ oznacza że czas życia od momentu rozpoczęcia obserwacji k-tego pacjenta to T_k^* (brak cenzurowania czyli $T_k = T_k^*$) natomiast $(T_k^*, 0)$ oznacza że czas życia od momentu rozpoczęcia obserwacji k-tego pacjenta jest większy niż T_k^* (cenzurowanie czyli $T_k > T_k^*$). Parametr θ (może być wielowymiarowy jak w modelu Weibulla) założonej klasy rozkładów T estymujemy metodą największej wiarygodności:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} [D_i \log(f_\theta(T_i^*)) + (1 - D_i) \log(S_\theta(T_i^*))] = \arg\max_{\theta} \sum_{i=1}^{n} [D_i \log(h_\theta(T_i^*)) + \log(S_\theta(T_i^*))]$$

W przypadku rozkładu wykładniczego prosty rachunek daje nam

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} T_i^*}{\sum_{i=1}^{n} D_i}$$

Korzystając z asymptotycznej normalności estymatora największej wiarygodności oraz z faktu, że asymptotyczna wariancja to odwrotność informacji Fishera dostajemy:

$$\hat{\sigma}_\lambda = \frac{\sum_{i=1}^{n} T_i^*}{\left(\sum_{i=1}^{n} D_i\right)^{\frac{3}{2}}}$$

co daje nam możliwość skonstruowania przedziału ufności dla parametru co przekłada się na obszar ufności dla funkcji przeżycia. Dla innych bardziej skomplikowanych klas rozkładów estymatory największej wiarygodności parametrów oblicza się numerycznie.

10.5 Model nieparametryczny bez cenzurowania

Zacznijmy najpierw od prostego przypadku danych niecenzurowanych, czyli obserwujemy $T_1, ..., T_n$. Mocno zgodny estymator funkcji przeżycia S dostajemy z twierdzenia Gliwienki-Cantellego:

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} I(t_i \geq t).$$

Niech $U_1 < U_2 < ... < U_k$ będzie zbiorem wartości czasów życia (ponieważ obserwacje mogą się powtarzać to $k \leq n$). Ponieważ $\hat{S}$ jest funkcją przeżycia rozkładu dyskretnego na $U_1, ..., U_k$ to możemy ją zapisać w postaci:

$$\hat{S}(t) = \prod_{j|U_j<t} (1 - \hat{\lambda}_j),$$

gdzie

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{n} I(T_i = U_j)}{\sum_{i=1}^{n} I(T_i \geq U_j)}, \qquad j \in \{1, ..., k\},$$

tak więc

$$\hat{S}(t) = \prod_{j|U_j<t} \frac{\sum_{i=1}^{n} I(T_i > U_j)}{\sum_{i=1}^{n} I(T_i \geq U_j)},$$

Zauważmy teraz że $\hat{S}(t)$ jest rozwiązaniem problemu maksymalizacji tak zwanej uogólnionej wiarygodności:

$$L(S) = \prod_{i=1}^{n} (S(T_i) - S(T_i^+))$$

Interpretacja jest taka: szukamy rozkładu dla którego prawdopodobieństwo otrzymania $T_1, ..., T_n$ w próbie prostej jest największe. Jest to oczywiście rozkład empiryczny. Tak więc

$$\hat{S} = \arg\max_S L(S).$$

10.6 Model nieparametryczny z danymi cenzurowanymi

Obserwujemy próbę prostą $(T_1^*, D_1), ..., (T_n^*, D_n)$. Aby na podstawie takiej próby wyestymować funkcję przeżycia zmaksymalizujemy uogólnioną wiarygodność, czyli

$$\hat{S} = \arg\max_S L^*(S) = \arg\max_S \prod_{i=1}^{n} (S(T_i^*) - S(T_i^{*+}))^{D_i} S(T_i^{*+})^{1-D_i},$$

Zauważmy że aby funkcja L^* była niezerowa to:

$$\forall\ j \quad D_j = 1 \quad \Rightarrow \quad S(T_i^*) > S(T_i^{*+}).$$

Tak więc w obserwacjach niecenzurowanych muszą być skoki. Jeśli ustalimy te skoki to druga część wzoru na uogólnioną wiarygodność powoduje że najlepiej aby funkcja S przyjmowała (po za punktami skoku) tak duże wartości jak to tylko możliwe. Ponieważ S jest nierosnąca to implikuje, że po za punktami skoku funkcja S powinna być stała. Tak więc problem redukuje się do poszukiwania wielkości skoków w obserwacjach niecenzurowanych.

Wprowadźmy następujące oznaczenia: Niech $U_1 < U_2 < ... < U_k$ będzie zbiorem wartości niecenzurowanych czasów życia. Niech $C_1, .., C_{k+1}$ będą liczbami cenzurowanych obserwacji takimi że: C_1 jest liczbą ocenzurowanych obserwacji w przedziale $[0, U_1)$, C_j dla $j = 2, .., k$ jest liczbą ocenzurowanych obserwacji w przedziale $[U_{j-1}, U_j)$ oraz C_{k+1} jest liczbą ocenzurowanych obserwacji w przedziale $[U_k, \infty)$. Niech

$$n_j = \sum_{i=1}^{n} I(T_i^* = U_j) I(D_i = 1).$$

Jeśli teraz zapiszemy uogólnioną wiarygodność dla kawałkami stałych funkcji S o skokach w $U_1, ..., U_k$ to wygląda ona tak:

$$L^*(S) = S(U_k^+)^{C_{k+1}} \prod_{j=1}^{k} (S(U_j) - S(U_j^+))^{n_j} S(U_j)^{C_j}$$

Ponieważ szukamy funkcji przeżycia rozkładu dyskretnego na $U_1, ..., U_k$ to możemy ją zapisać za pomocą hazardu w następujący sposób:

$$S(U_j) = \prod_{i=1}^{j-1} (1 - \lambda_i), \qquad S(U_j^+) = \prod_{i=1}^{j} (1 - \lambda_i)$$

Otrzymujemy więc

$$L^*(S) = \prod_{i=1}^{k}(1-\lambda_i)^{C_{k+1}} \cdot \prod_{j=1}^{k}\left[\left(\prod_{i=1}^{j-1}(1-\lambda_i) - \prod_{i=1}^{j}(1-\lambda_i)\right)^{n_j} \prod_{i=1}^{j-1}(1-\lambda_i)^{C_j}\right]$$

$$= \prod_{i=1}^{k}(1-\lambda_i)^{C_{k+1}} \cdot \prod_{j=1}^{k}\left[\lambda_j^{n_j}\prod_{i=1}^{j-1}(1-\lambda_i)^{C_j+n_j}\right]$$

$$= (1-\lambda_1)^{\sum_{j=2}^{k+1}C_j+\sum_{j=2}^{k}n_j}\lambda_1^{n_1} \cdot (1-\lambda_2)^{\sum_{j=3}^{k+1}C_j+\sum_{j=3}^{k}n_j}\lambda_2^{n_2} \cdot \ldots \cdot (1-\lambda_k)^{C_{k+1}}\lambda_k^{n_k}$$

Po zlogarytmowaniu tej funkcji i przyrównaniu pochodnych cząstkowych do zera otrzymujemy:

$$\hat{\lambda}_l = \frac{n_l}{\sum_{j=l+1}^{k+1}C_j + \sum_{j=l}^{k}n_j} = \frac{n_l}{\sum_{i=1}^{n}I(T_i^* \geq U_l)}, \quad l \in 1, \ldots, k$$

Ostatecznie otrzymujemy estymator postaci

$$\hat{S}(t) = \prod_{j|U_j<t}(1-\hat{\lambda}_j) = \prod_{j|U_j<t}\left(1 - \frac{\sum_{i=1}^{n}I(T_i^* = U_j)I(D_i = 1)}{\sum_{i=1}^{n}I(T_i^* \geq U_j)}\right)$$

zwany estymatorem Kaplana-Meiera.

10.7 Fomuła Greenwooda i punktowe przedziały ufności dla estymatora Kaplana-Meiera

Jest wiele różnych metod estymacji wariancji $\hat{S}(t)$. Najpopularniejsza to metoda Greenwooda polegająca na oszacowaniu wariancji $\ln(\hat{S}(t))$ wykorzystując rozwinięcie logarytmu w szereg Taylora. Pomijając szczegóły rachunkowe otrzymujemy:

$$\hat{V}(\ln(\hat{S}(t))) = \sum_{j|U_j<t}\frac{\sum_{i=1}^{n}I(T_i^* = U_j)I(D_i = 1)}{\left(\sum_{i=1}^{n}I(T_i^* \geq U_j)\right)\left(\sum_{i=1}^{n}I(T_i^* \geq U_j) - I(T_i^* = U_j)I(D_i = 1)\right)}$$

Korzystając z metody delta oraz z asymptotycznej nieobciążoności $\hat{S}(t)$ dostajemy

$$\hat{V}(\hat{S}(t)) = (\hat{S}(t))^2\sum_{j|U_j<t}\frac{\sum_{i=1}^{n}I(T_i^* = U_j)I(D_i = 1)}{\left(\sum_{i=1}^{n}I(T_i^* \geq U_j)\right)\left(\sum_{i=1}^{n}I(T_i^* \geq U_j) - I(T_i^* = U_j)I(D_i = 1)\right)}$$

Asymptotyczny przedział ufności dla $S(t)$ wykorzystujący asymptotyczną normalność estymatora $\hat{S}(t)$ można wyrazić następująco:

$$P\left(S(t) \in \left[\hat{S}(t) - F_{N(0,1)}^{-1}(1-\alpha/2)\sqrt{\hat{V}(\hat{S}(t))}, \hat{S}(t) + F_{N(0,1)}^{-1}(1-\alpha/2)\sqrt{\hat{V}(\hat{S}(t))}\right]\right) \to 1-\alpha.$$

Powyższy przedział może jednak wychodzić po za przedział $[0,1]$. Dlatego najczęściej stosuje się dodatkowe techniczne przekształcenie. Korzystając z techniki Greenwooda i metody delta obliczamy wariancję dla $\ln(-\ln(\hat{S}(t)))$. W ten sposób otrzymujemy

$$\sigma^* = \sqrt{\hat{V}(\ln(-\ln(\hat{S}(t))))} = \sqrt{\frac{\hat{V}(\hat{S}(t))}{(\hat{S}(t)\ln(\hat{S}(t)))^2}}$$

Obkładając 2 razy funkcją exp przedział ufności dla $\ln(-\ln(S(t)))$ otrzymujemy:

$$P\left(S(t) \in \left[\hat{S}(t)^{\exp(-F^{-1}_{N(0,1)}(1-\alpha/2)\sigma^*)}, \hat{S}(t)^{\exp(F^{-1}_{N(0,1)}(1-\alpha/2)\sigma^*)}\right]\right) \to 1-\alpha.$$

Taki przedział ufności zawsze zawiera się w przedziale $[0,1]$. Analizy numeryczne wykazały, że powyższy asymptotyczny przedział można już stosować w praktyce gdy $n \geq 25$ oraz liczba obserwacji cenzurowanych nie przekracza połowy wszystkich obserwacji.

10.8 Punktowy test dla dwóch funkcji przeżycia

Obserwujemy dwie niezależne próby proste $(T^*_{1,A}, D_{1,A}), ..., (T^*_{m,A}, D_{m,A})$ oraz $(T^*_{1,B}, D_{1,B}), ..., (T^*_{n,B}, D_{n,B})$. Rozważmy test

$$H_0 : S_A(t) = S_B(t), \quad H_1 : S_A(t) \neq S_B(t)$$

lub test

$$H_0 : S_A(t) \leq S_B(t), \quad H_1 : S_A(t) > S_B(t).$$

Naturalną statystyką testową jest:

$$W(t) = \frac{\hat{S}_A(t) - \hat{S}_B(t)}{\sqrt{\hat{V}(\hat{S}_A(t)) + \hat{V}(\hat{S}_B(t))}}.$$

p-wartość dla pierwszego testu obliczamy ze wzoru: która dla $S_A(t) = S_B(t)$ ma asymptotyczny rozkład $N(0,1)$. Tak więc p-wartość dla pierwszego testu obliczamy ze wzoru:

$$p(t) = 2(1 - F_{N(0,1)}(|W(t)|)),$$

a dla drugiego testu

$$p(t) = 1 - F_{N(0,1)}(W(t)).$$

10.9 Globalny test dla dwóch funkcji przeżycia przy założeniu stochastycznej dominacji

Testowanie stochastycznej dominacji dla danych cenzurowanych jest możliwe ale to dość skomplikowane. W praktyce zazwyczaj zakłada się, że krzywe przeżycia się nie przecinają (tak samo jak w teście Manna-Whitneya). Będziemy rozważali następujące problemy testowania:

$$H_0 : \forall\ t \geq 0\ \ S_A(t) = S_B(t), \quad H_1 : (\forall\ t \geq 0\ S_A(t) \leq S_B(t) \vee \forall\ t \geq 0\ S_A(t) \geq S_B(t))\ \wedge \exists\, t \geq 0\ \ S_A(t) \neq S_B(t),$$

lub

$$H_0 : \forall\ \ t \geq 0\ \ S_A(t) \leq S_B(t), \qquad H_1 : \forall\ \ t \geq 0\ \ S_A(t) \geq S_B(t)\ \ \wedge\ \ \exists\ \ t \geq 0\ \ S_A(t) > S_B(t).$$

Najpopularniejszym testem w tych problemach testowania jest tzw. test Logrank. Niech $U_1 < ... < U_k$ będzie zbiorem wartości niecenzurowanych dla obu grup. Wprowadźmy następujące oznaczenia:

$$n_{jA} = \sum_{i=1}^{m} I(T^*_{iA} = U_j) I(D_{iA} = 1), \quad n_{jB} = \sum_{i=1}^{n} I(T^*_{iB} = U_j) I(D_{iB} = 1)$$

$$N_{jA} = \sum_{i=1}^{m} I(T^*_{iA} \geq U_j), \qquad N_{jB} = \sum_{i=1}^{n} I(T^*_{iB} \geq U_j).$$

Zauważmy że

$$\hat{\lambda}_{jA} = \frac{n_{jA}}{N_{jA}}, \quad \hat{\lambda}_{jB} = \frac{n_{jB}}{N_{jB}}$$

są estymatorami hazardu który przy założeniu $S_A = S_B$ jest taki sam w obu grupach. Tak więc przy założeniu $S_A = S_B$ rozkład warunkowy $n_{jB}|n_{jA} + n_{jB}$ jest rozkładem hipergeometrycznym z parametrami $N_{jA} + N_{jB}, N_{jB}, n_{jA} + n_{jB}$.

Uwaga: dla dwóch niezależnych zmiennych o rozkładzie dwumianowym i tym samym prawdopodobieństwie sukcesu rozkład hipergeometryczny jest rozkładem warunkowym pierwszej zmiennej przy ustalonej sumie.

Wartość oczekiwana rozkładu hipergeometrycznego z parametrami $N_{jA} + N_{jB}, N_{jB}, n_{jA} + n_{jB}$ to

$$E_{jB} = \frac{N_{jB}(n_{jA} + n_{jB})}{N_{jA} + N_{jB}},$$

natomiast wariancja to

$$V_{jB} = \frac{N_{jA}N_{jB}(n_{jA} + n_{jB})(N_{jA} + N_{jB} - n_{jA} - n_{jB})}{(N_{jA} + N_{jB})^2(N_{jA} + N_{jB} - 1)}$$

Statystyka testowa testu Logrank jest postaci:

$$Z = \frac{\sum_{j=1}^{k}(n_{jB} - E_{jB})}{\sqrt{\sum_{j=1}^{k} V_{jB}}}$$

która przy założeniu $S_A = S_B$ ma asymptotyczny rozkład $N(0, 1)$. Tak więc p-wartość dla pierwszego testu obliczamy ze wzoru:

$$p = 2(1 - F_{N(0,1)}(|Z|)),$$

a dla drugiego testu (duże dodatnie wartości świadczą przeciwko hipotezie)

$$p = 1 - F_{N(0,1)}(Z).$$

Uwaga: Tak samo jak test Manna-Whitneya test Lograng jest niezmienniczy ze względu na przekształcenia ciągłe i ściśle rosnące. Jego zgodności dowodzi się podobnie, przekształcając oryginale obserwacje na rangi jednak sama wartość statystyki testowej może być obliczona bez rangowania i tak zwykle się ją przedstawia.

10.10 Model proporcjonalnego hazardu Coxa

Podobnie jak w modelu regresji logistycznej będziemy teraz badali wpływ zmiennych towarzyszących ale tym razem nie na pojedyncze prawdopodobieństwo a na całą funkcję przeżycia. Oprócz niezależnych par $(T_1^*, D_1), ..., (T_n^*, D_n)$ (tu nie tworzą one próby prostej bo różni pacjenci mają różne rozkłady czasu życia) obserwujemy również dla każdego pacjenta zestaw zmiennych towarzyszących

$$X_{1i}, ..., X_{di}, \quad i \in 1, ..., n$$

zakodowanych tak samo jak dla regresji logistycznej. Zakładamy następujący wpływ zmiennych towarzyszących na funkcje hazardu:

$$h_i(t) = h_0(t) \exp\left(\sum_{j=1}^{d} \beta_j X_{ji}\right),$$

gdzie h_0 jest nieznaną funkcją hazardu bazowego a $\beta = (\beta_1, ..., \beta_d)$ są nieznanymi współczynnikami modelu. Zakładając, że czas życia jest zmienną losową absolutnie ciągłą względem miary Lebesgue'a otrzymujemy

$$S_i(t) = [S_0(t)]^{\exp\left(\sum\limits_{j=1}^{d} \beta_j X_{ji}\right)},$$

gdzie

$$S_0(t) = \exp\left(-\int\limits_0^t h_0(x)dx\right).$$

Zauważmy teraz, że jeśli oznaczymy

$$X^{0l} = (X_1, ..., X_{l-1}, 0, X_{l+1}, ..., X_d), \quad X^{1l} = (X_1, ..., X_{l-1}, 1, X_{l+1}, ..., X_d),$$

w przypadku zmiennej grupującej X_l oraz

$$X^{0l} = (X_1, ..., X_{l-1}, X_l, X_{l+1}, ..., X_d), \quad X^{1l} = (X_1, ..., X_{l-1}, X_l + 1, X_{l+1}, ..., X_d),$$

w przypadku zmiennej ciągłej X_l to

$$S(t, X^{1l}) = [S_0(t)]^{\exp\left(\beta \circ X^{1l}\right)} = [S_0(t)]^{\exp\left(\beta \circ X^{0l}\right)\frac{\exp\left(\beta \circ X^{1l}\right)}{\exp\left(\beta \circ X^{0l}\right)}} = [S(t, X^{0l})]^{\exp(\beta_l)}.$$

Wyrażenie

$$HR(\beta_l) = \exp(\beta_l) = \frac{h(t, X^{1l})}{h(t, X^{0l})} = \frac{h_0(t)\exp\left(\beta \circ X^{1l}\right)}{h_0(t)\exp\left(\beta \circ X^{0l}\right)}$$

zależy tylko od β_l i nazywa się ilorazem hazardu (hazard ratio). Jeśli iloraz hazardu jest równy 1 oznacza to, że $\beta_l = 0$ czyli zmienna towarzysząca X_l nie wpływa na funkcję przeżycia. Iloraz hazardu dla zmiennych objaśniających ma taka samą interpretację jak iloraz szans w modelu regresji logistycznej.

Wiarygodność parametrów β dla i-tej osoby u której zaobserwowaliśmy niecenzurowany czas życia $(T_i^*, 1)$ to

$$L_i(\beta) = \frac{h_i(T_i^*)}{\sum\limits_{k|T_k^* \geq T_i^*} h_k(T_i^*)} = \frac{h_0(T_i^*)\exp\left(\sum\limits_{j=1}^{d} \beta_j X_{ji}\right)}{\sum\limits_{k|T_k^* \geq T_i^*} h_0(T_i^*)\exp\left(\sum\limits_{j=1}^{d} \beta_j X_{jk}\right)} = \frac{\exp\left(\sum\limits_{j=1}^{d} \beta_j X_{ji}\right)}{\sum\limits_{k|T_k^* \geq T_i^*} \exp\left(\sum\limits_{j=1}^{d} \beta_j X_{jk}\right)}$$

Tak więc estymator β to

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^d}{\arg\max} \sum_{i=1,...,n|D_i=1} \log(L_i(\beta)) = \underset{\beta \in \mathbb{R}^d}{\arg\max} \sum_{i=1,...,n|D_i=1} \left(\sum_{j=1}^{d} \beta_j X_{ji} - \log\left(\sum_{k|T_k^* \geq T_i^*} \exp\left(\sum_{j=1}^{d} \beta_j X_{jk}\right)\right)\right)$$

Testując hipotezę

$$H_0 : \beta_l = 0, \quad H_1 : \beta_l \neq 0$$

estymujemy dodatkowo β przy założeniu H_0

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^d,\ \beta_l = 0}{\arg\max} \sum_{i=1,...,n|D_i=1} \left(\sum_{j=1}^{d} \beta_j X_{ji} - \log\left(\sum_{k|T_k^* \geq T_i^*} \exp\left(\sum_{j=1}^{d} \beta_j X_{jk}\right)\right)\right)$$

i obliczamy wartość statystyki testowej

$$T = 2\left(\sum_{i=1,...,n|D_i=1} \log(L_i(\hat{\beta})) - \log(L_i(\tilde{\beta}))\right),$$

która przy H_0 ma asymptotyczny rozkład χ^2_1 tak więc p-wartość powyższego testu obliczamy ze wzoru:

$$p = 1 - F_{\chi^2_1}(T).$$

Macierz informacji Fishera dla wyestymowanego parametru $I(\hat{\beta})$ wyraża się wzorem:

$$I(\hat{\beta}) = \sum_{i=1,...,n|D_i=1} \left(\frac{\sum\limits_{k|T_k^* \geq T_i^*} \hat{\theta}_k \mathbf{X}(k)^T \mathbf{X}(k)}{\sum\limits_{k|T_k^* \geq T_i^*} \hat{\theta}_k} - \frac{\left(\sum\limits_{k|T_k^* \geq T_i^*} \hat{\theta}_k \mathbf{X}(k)^T \right) \left(\sum\limits_{k|T_k^* \geq T_i^*} \hat{\theta}_k \mathbf{X}(k) \right)}{\left(\sum\limits_{k|T_k^* \geq T_i^*} \hat{\theta}_k \right)^2} \right),$$

gdzie

$$\hat{\theta}_k = \exp\left(\sum_{j=1}^{d} \hat{\beta}_j X_{jk} \right), \quad \mathbf{X}(k) = (X_{1k}, ..., X_{dk}).$$

Po odwróceniu tej macierzy i odczytaniu odpowiedniej wartości z przekątnej mamy asymptotyczne odchylenie standardowe dla estymatora parametru β_l:

$$\hat{\sigma}_{\beta_l} = \sqrt{I^{-1}(\hat{\beta})_{ll}}$$

Otrzymujemy:

$$P\left(\beta_l \in \left[\hat{\beta}_l - \hat{\sigma}_{\beta_l} F^{-1}_{N(0,1)}\left(1 - \frac{\alpha}{2}\right), \hat{\beta}_l + \hat{\sigma}_{\beta_l} F^{-1}_{N(0,1)}\left(1 - \frac{\alpha}{2}\right) \right] \right) \to 1 - \alpha$$

a z tego wynika że

$$P\left(HR(\beta_l) \in \left[e^{\hat{\beta}_l - \hat{\sigma}_{\beta_l} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)}, e^{\hat{\beta}_l + \hat{\sigma}_{\beta_l} F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)} \right] \right) \to 1 - \alpha$$

Możemy jeszcze prognozować krzywą przeżycia dla dowolnego ustalonego wektora zmiennych towarzyszących $X' = (X'_1, ..., X'_d)$. Wstawiając wzór $S(T_i^*) = [S_0(T_i^*)]^{\exp\left(\sum\limits_{j=1}^{d} \hat{\beta}_j X_{ji} \right)}$ do uogólnionej wiarygodności i rozwiązując problem maksymalizacji ze względu na S_0 identycznie jak przy estymatorze Kalpana-Meiera otrzymujemy

$$\hat{S}_0(t) = \prod_{i=1,...,n \ | \ D_i=1 \ \wedge \ T_i^* < t} \left(1 - \frac{\exp\left(\sum\limits_{j=1}^{d} \hat{\beta}_j X_{ji} \right)}{\sum\limits_{k|T_k^* \geq T_i^*} \exp\left(\sum\limits_{j=1}^{d} \hat{\beta}_j X_{jk} \right)} \right)^{\exp\left(- \sum\limits_{j=1}^{d} \hat{\beta}_j X_{ji}) \right)}.$$

W modelu proporcjonalnego hazardu $S(t, X') = [S_0(t)]^{\exp\left(\sum\limits_{j=1}^{d} \beta_j X'_j \right)}$ więc

$$\hat{S}(t, X') = \prod_{i=1,...,n \ | \ D_i=1 \ \wedge \ T_i^* < t} \left(1 - \frac{\exp\left(\sum\limits_{j=1}^{d} \hat{\beta}_j X_{ji} \right)}{\sum\limits_{k|T_k^* \geq T_i^*} \exp\left(\sum\limits_{j=1}^{d} \hat{\beta}_j X_{jk} \right)} \right)^{\exp\left(\sum\limits_{j=1}^{d} \hat{\beta}_j (X'_j - X_{ji}) \right)}$$

Uwaga: W modelu proporcjonalnego hazardu zakładamy wszędzie, że czas życia jest zmienną absolutnie ciągłą względem miary Lebesgue'a. W praktyce jednak dane są zapisane z pewną dokładnością (np. kilka miejsc po przecinku) i mogą się powtarzać. Najprostszym rozwiązaniem w takiej sytuacji jest dodać do danych niezależne zmienne z rozkładu jednostajnego na przedziale $[0, r]$ gdzie r jest o kilka rzędów wielkości mniejsze niż zadana dokładność. Taki zabieg nie zaburzy numerycznie wektora obserwacji ale spowoduje, że z prawdopodobieństwem jeden obserwacje będą różne.