1 Współliniowość

Intuicyjnie, ze współliniowością mamy do czynienia w sytuacji, gdy jeden predyktor x_i jest mocno skorelowany z predyktorem x_j lub pewną kombinacją liniową podzbioru predyktorów x_{-i}. W jaki sposób wykryć współliniowość?

- duże zmiany w oszacowaniach współczynników gdy zmienna jest dodana/usuwana,
- nieistotne statystycznie t-testy dla ważnych (istotnych teoretycznie) predyktorów,
- współczynniki o znakach przeciwnych, niż oczekiwane na podstawie wiedzy/teorii,
- duże wartości korelacji między predyktorami,
- szerokie przedziały ufności (duże błędy standardowe) współczynników.

Jaki jest wpływ współliniowości na współczynniki MNK? Przede wszystkim: **obecność współliniowości nie powoduje obciążenia estymatorów MNK**. Problematyczna natomiast staje się macierz kowariancji

$$\Sigma_{\hat{\beta}} = \sigma^2 (X^T X)^{-1}$$

Dlaczego? W sytuacji współliniowości predyktorów macierz $X^T X$ jest bliska nieodwracalnej, zatem posiada wyznacznik który jest bliski wartości 0. Rozpatrzmy rozkład spektralny dla macierzy $X^T X$

$$X^T X = V \Lambda V^T$$
$$(X^T X)^{-1} = V \Lambda^{-1} V^T$$

macierz Λ jest macierzą diagonalną o wartościach własnych $X^T X$. W przypadku, gdy $\det(X^T X) \approx 0$, to wiemy, że przynajmniej jedna z wartości własnych t_i jest bliska zeru, zatem wartości na diagonali macierzy Λ^{-1} (i tym samym elementy macierzy kowariancji) są bardzo duże (por. Transza 2, zadanie 1). W przypadku współliniowości mamy więc potencjalnie problem z dużymi błędami standardowymi estymatorów. Jest to problematyczne w sytuacji, gdy zależy nam na wyjaśnialności, bo wnioskowanie statystyczne na temat istotności parametrów populacyjnych jest utrudnione. W przypadku predykcji współliniowość zazwyczaj nie jest tak problematyczna, o ile dane testowe i treningowe pochodzą z tego samego DGP (data generating process).

Rozważmy prosty przypadek 2 predyktorów x_1, x_2. Jeśli nie ma współliniowości/jest ona niewielka, wówczas dopasowujemy płaszczyznę do danych (y jest trzecim wymiarem) i często istnieje bardzo wyraźna *najlepsza* płaszczyzna. Jednak w przypadku współliniowości związek jest tak naprawdę linią przechodzącą przez trójwymiarową przestrzeń z rozproszonymi wokół niej danymi. MNK próbuje dopasować płaszczyznę do linii, więc istnieje nieskończona liczba płaszczyzn, które idealnie przecinają się z tą linią, a to, która płaszczyzna zostanie wybrana, zależy od wpływowych punktów w danych: niewielka zmiana w pojedynczym punkcie danych może gwałtownie zmienić oszacowania współczynnika.

1.1 Współczynnik korelacji wielokrotnej

Miara tego, jak dobrze dany predyktor może być przewidziany za pomocą liniowej kombinacji pozostałych zmiennych. Jest to maksymalny współczynnik korelacji między predyktorem a kombinacją liniową pozostałych predyktorów.

$$r_i = \frac{\mathrm{Cov}(x_i, \hat{x}_i)}{\sqrt{\mathrm{Var}(x_i)\,\mathrm{Var}(\hat{x}_i}}$$

gdzie $\hat{x}_i$ są wartościami dopasowanymi w modelu $x_i \sim x_{-i}$. Współczynnik determinacji jest jego kwadratem

$$R_i^2 = r_i^2$$

1.2 Współczynnik podbicia wariancji

VIF jest współczynnikiem, o który wariancja $j-$tego predyktora jest *zawyżona* przez istnienie korelacji między zmiennymi predykcyjnymi w modelu (czyli w porównaniu do sytuacji, gdy zmienne są nieskorelowne).

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

Pierwiastek z VIF wskazuje, o ile wzrasta błąd standardowy w porównaniu do sytuacji, w której zmienna miała zerową korelację z innymi zmiennymi predykcyjnymi w modelu.

Problem: VIF nie mówi, z którymi innymi predyktorami nasz predyktor jest skorelowany; możemy mieć jedną (lub więcej) grupę skorelowanych zmiennych. Zmienna j jest **potencjalnie** problematyczna, gdy

$$\mathrm{VIF}_j > 5, \text{ mała próba}$$
$$\mathrm{VIF}_j > 10, \text{ duża próba}$$

Co ciekawe, duże wartości VIF mogą występować niezależnie od tego, jak dokładnie oszacowano parametry regresji.[1]

[1] O'Brien, R. M. (2007). *A Caution Regarding Rules of Thumb for Variance Inflation Factors*. Quality & Quantity. 41 (5): 673–690.

1.3 Współliniowość – co zrobić?

Tak, jak opisaliśmy wcześniej, współliniowość nie zawsze jest problematyczna. Podstawą każdej analizy powinna być próba zrozumienia danych i tego, dlaczego współliniowość występuje i o czym ona świadczy. **Automatyczne** pozbywanie się np. jednej ze skorelowanych zmiennych rzadko kiedy jest dobrą praktyką (chyba, że mamy idealną współliniowość) i może prowadzić do problemu tzw. zmiennej zakłócającej (confounding bias, a także omitted variable bias)[2]. Dużą popularnością cieszą się tzw. metody regularyzacji.

2 Ridge

W regresji grzbietowej nakładamy karę na wielkość współczynników w postaci normy L_2:

$$\left(\hat{\beta}_0^r, \hat{\beta}_{-0}^r\right) = \text{argmin}_{(\beta_0,\beta_{-0})} \left[SSE(\beta_0, \beta_{-0}) + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right]$$

Inaczej

$$\left(\hat{\beta}_0^r, \hat{\beta}_{-0}^r\right) = \text{argmin}_{(\beta_0,\beta_{-0})} \, SSE(\beta_0, \beta_{-0})$$

przy warunku

$$\sum_{j=1}^{p} \beta_j^2 \leq t \iff ||\beta_{-0}||_2^2 \leq t$$

Oryginalna koncepcja polegała na poprawieniu uwarunkowania macierzy X^TX (**regularyzacja Tichonowa**):

$$\hat{\beta}^r = (X^TX + \lambda I)^{-1} X^T y$$

Dlaczego poprawia to uwarunkowanie? Otóż, X^TX jest macierzą symetryczną o wartościach rzeczywistych, więc jest diagonalizowalna rzeczywistymi wartościami własnymi $t_1, ..., t_p$. Szukamy ich w następujący sposób

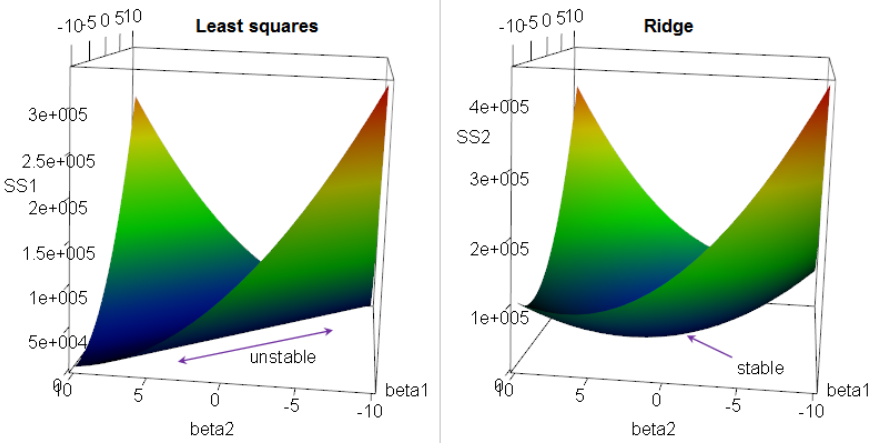
$$\det\left(X^TX - t_i I\right) = 0 \tag{1}$$

Spójrzmy teraz, co dzieje się w przypadku ridge

$$\begin{aligned} \det\left(X^TX + \lambda I - t_i I\right) &= 0 \\ \det\left(X^TX - (t - \lambda) I\right) &= 0 \end{aligned} \tag{2}$$

Możemy rozwiązać to ze względu na $(t - \lambda)$ i dostać takie same wartości własne, jak w (1). Załóżmy, że w (1) mamy pewną wartość własną t_i. Wtedy wartość własna w (2) jest równa $t_i + \lambda$. Oczywiście w przypadku macierzy pół-dodatnio określonej mamy tylko $t_i \geq 0$. Geometrycznie dodanie stałej do diagonali jest tożsame z dodaniem paraboli obrotowej w początku układu współrzędnych. Jest to o tyle istotne, że w przypadku współliniowości predyktorów SSE wygląda właśnie jak odwrócony grzbiet (ang. *ridge*)/dolina. W takiej sytuacji minimum SSE nie jest dobrze określone (Rys. 1).

Okazuje się, że taki sposób penalizacji prowadzi do ściągania współczynników do 0. Rozważmy rozkład SVD



Rysunek 1: Ridge

$$X = UDV^T$$

[2]Giles, Dave (15 September 2011). *Econometrics Beat: Dave Giles' Blog: Micronumerosity*. Econometrics Beat. Retrieved 3 September 2023.

gdzie U oraz V są ortogonalnymi macierzami, kolumny U rozpinają przestrzeń kolumnową X (wektory własne XX^T), a kolumny V rozpinają przestrzeń wierszową (wektory własne X^TX). D jest macierzą diagonalną o wartościach osobliwych X.

$$\begin{aligned}\hat{\beta}_{OLS} &= (X^TX)^{-1}X^Ty \\ &= ((UDV^T)^T(UDV^T))^{-1}(UDV^T)^Ty \\ &= (VDU^TUDV^T)^{-1}VDU^Ty \\ &= VD^{-2}V^TVDU^Ty \\ &= VD^{-1}U^Ty\end{aligned}$$

$$\begin{aligned}\hat{\beta}^r &= (X^TX+\lambda)^{-1}X^Ty \\ &= (VD^2V^T+\lambda 1_p)^{-1}VDU^Ty \\ &= (VD^2V^T+\lambda VV^T)^{-1}VDU^Ty \\ &= (V(D^2+\lambda)V^T)^{-1}VDU^Ty \\ &= V(D^2+\lambda)^{-1}V^TVDU^Ty \\ &= V(D^2+\lambda)^{-1}DU^Ty\end{aligned}$$

czyli wyrażenie D^{-1} w OLS mnożymy przez $(D^2+\lambda)^{-2}D^2$, zatem dla $\lambda \to +\infty$ mamy $\hat{\beta}^r \to 0$, czyli rzeczywiście ściąganie współczynników do zera.

Ważne własności

(a) musimy wystandaryzować nasze dane

(b) nie penalizujemy interceptu

Ad (a): estymator MNK ma własność skalowania: $c \cdot x_i \implies \hat{\beta}_i/c$, zatem wpływ $\hat{\beta}_i x_i$ nie zależy od jednostek (np. waga w kg vs lbs), które mamy (i bardzo dobrze). To jednak nie jest prawda dla regresji grzbietowej: współczynniki zależą od skali, w której jest wyrażona dana zmienna, więc aby penalizowanie było *fair*, wszystkie współczynniki muszą być na tej samej skali.

Ad (b): jeśli nie penalizujemy interceptu, to dodanie c do wszystkich y powoduje wzrost β_0 również o c i tym samym wzrost $\hat{y}$ również o c. Jeśli penalizowalibyśmy intercept, to wtedy β_0 musiałoby wzrosnąć o mniej, niż c. Poza tym, nie penalizujac interceptu mamy zachowane przydatne własności, takie jak $R^2 = \text{corr}^2(\hat{y}, y)$.

3 LASSO

W LASSO nakładamy karę na wielkość współczynników w postaci normy L_1:

$$\hat{\beta}^r = \text{argmin}_\beta \left[SSE(\beta) + 2\lambda \sum_{j=1}^{p-1} |\beta_j| \right]$$

Okazuje się, że taka kara prowadzi do zerowania współczynników: LASSO działa jako metoda selekcji zmiennych. Wraz ze zmniejszaniem λ włączają się do modelu kolejne zmienne, jednak rangowanie zmiennych na podstawie kolejności ich włączania się do modelu (przy zmniejszającym się λ) może być mylące. Trzeba pamiętać, że LASSO **nie zawsze** zeruje współczynniki: optymalizuje funkcję straty, a szczególna struktura tego problemu optymalizacyjnego sprawia, że rozwiązanie prawdopodobnie znajduje się w wierzchołku kwadratu (punkcie ekstremalnym kuli w normie ℓ_1).

Niestety, analityczne rozwiązania istnieją jedynie w przypadkach: jednowymiarowym oraz ortogonalnych kolumn X. Istnieją jednak skuteczne algorytmy, takie jak least-angle regression (LAR) oraz coordinate descent.

Za pomocą symulacji, w których mamy: małą liczbę prawdziwych, związanych z odpowiedzią predyktorów oraz dużą liczbę niezwiązanych, można pokazać, że LASSO faktycznie znajduje *prawdziwe* predyktory. Jednak w rzeczywistych przypadkach, przy skorelowanych zmiennych, wybór *ważnych* predyktorów może różnić się między próbkami (w takiej sytuacji algorytm rzadki nie może być stabilny[3]). Nawet, jeśli LASSO zwraca wartość 0 dla współczynnika predyktora (zgodnie z jego przeznaczeniem), nie oznacza to, że jest on *nieistotny*; oznacza to tylko, że nie dodał wystarczająco dużo do modelu, aby mieć znaczenie dla konkretnej próby i jej wielkości. Wybrane predyktory mogą być ważne w ramach konkretnej próbki danych, ale nie oznacza to, że są one najważniejsze w jakimkolwiek fundamentalnym sensie w całej populacji i z pewnością nie można ich interpretować jako mających wpływ przyczynowy na wynik. Można próbować tworzyć ranking zmiennych, przeprowadzając kroswalidację/bootstrap (*stability selection*) i sprawdzać, ile razy jaka zmienna okazała się *ważna*. Ważne uwagi

- w LASSO zaczynamy od dużego λ i stopniowo *łagodzimy* karę. W rezultacie **zmienne wprowadzane są pojedynczo**, przy czym w każdym punkcie relaksacji podejmowana jest decyzja czy bardziej wartościowe jest zwiększenie współczynników zmiennych już znajdujących się w regresji, czy dodanie też dodanie kolejnej zmiennej,

[3] H. Xu, C. Caramanis and S. Mannor, *Sparse Algorithms Are Not Stable: A No-Free-Lunch Theorem*, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 1, pp. 187-193, Jan. 2012, doi: 10.1109/TPAMI.2011.177.

- **wczesny wybór jednej zmiennej może wpłynąć na to, kiedy inne zmienne skorelowane z nią wejdą później w proces relaksacji**. Możliwe jest również wczesne wprowadzenie zmiennej, a następnie spadek jej współczynnika LASSO w miarę wprowadzania innych skorelowanych zmiennych.
- w przypadku użycia algorytmu LARS, gdy występuje współliniowość, LASSO wybiera jedną zmienną, która ma największą korelację z odpowiedzią.

4 LASSO a ridge

Załóżmy model $y \sim x + z$, gdzie x i z są wysoce współliniowe, więc x i z mniej więcej mogą się wzajemnie zastępować w przewidywaniu y. Zatem wiele liniowych kombinacji x, z, w których po prostu zastępujemy część x przez z, będzie działać bardzo podobnie jako predyktory, na przykład

$$0.2x + 0.8z$$
$$0.3x + 0.7z$$
$$0.5x + 0.5z$$

będą mniej więcej równie *dobre* jako predyktory. Spójrzmy teraz na te trzy przykłady:

- kara LASSO we wszystkich trzech przypadkach jest równa i wynosi 1
- kara ridge różni się, wynosi odpowiednio $0.68, 0.58, 0.5$

zatem ridge będzie preferować równe ważenie zmiennych współliniowych, podczas gdy LASSO może nie być w stanie jednoznacznie wybrać. Jest to jeden z powodów, dla których ridge może działać lepiej ze współliniowymi predyktorami: gdy dane dają niewielki powód do wyboru między różnymi liniowymi kombinacjami predyktorów współliniowych, LASSO może być *niezdecydowane*, podczas gdy ridge ma tendencję do wybierania równych wag. Jeśli interesuje nas moc predykcyjna, to być może ridge będzie lepszym wyborem na przyszłym zbiorze danych: może być lepiej zachować wszystkie predyktory z równą wagą (ridge), niż wybrać losowy podzbiór (lasso). Ponieważ **podzbiór jest losowy**, nie ma gwarancji, że jest to właściwy podzbiór *poza próbką*. Uśrednienie wszystkich cech nadal nie jest tak dobre, jak poznanie prawdziwego podzbioru, ale może być bliższe niż *pechowy* losowy podzbiór.

4.1 Elastic Net

Oba algorytmy mają swoje wady i zalety. Pewną modyfikacją (albo uogólnieniem) jest tzw. Elastic Net, gdzie mamy kombinację wypukłą norm L_1 oraz L_2

$$\lambda P_\alpha(\beta) = \lambda \frac{1-\alpha}{2} ||\beta||_2^2 + \lambda\alpha||\beta||_1 = \lambda \sum_{j=1}^{p} \left[\frac{1-\alpha}{2}\beta_j^2 + \alpha|\beta| \right]$$

Na przykład w przypadku $p \gg n$, LASSO wybiera co najwyżej n zmiennych. Ponadto, jeśli istnieje grupa silnie skorelowanych zmiennych, LASSO ma tendencję do wybierania jednej zmiennej z grupy i ignorowania pozostałych. Kara L_2 z kolei sprawia, że funkcja straty jest wypukła, a zatem ma unikalne minimum.

5 Podejście bayesowskie

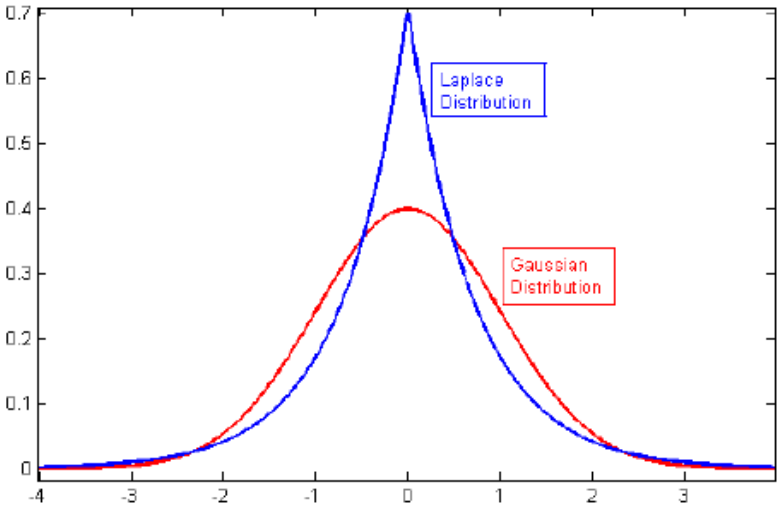
W dużym uproszczeniu, we wnioskowaniu bayesowskim uwzględniamy pewną dodatkową wiedzę o naszych parametrach (np. wiedzę ekspercką, doświadczenie) za pomocą rozkładu a priori. Otrzymujemy (w bardzo dużym uproszczeniu) rozkład a posteriori

$$\text{a posteriori} \propto \text{wiarogodność} \times \text{a priori}$$

Okazuje się, że estymator MAP (maximum a posterior) postaci

$$\hat{\theta}_{\text{MAP}} = \arg\max_\theta \log P(y|\theta) + \log P(\theta)$$

jest tożsamy z estymatorem metody największej wiarogodności, gdy rozkładem a priori parametrów jest rozkład jednostajny. W przypadku, gdy rozkładem a priori jest rozkład normalny, to jest to równoważne z regularyzacją L_2, natomiast gdy a priori jest rozkładem Laplace'a (podwójnie wykładniczy), dostajemy regularyzację L_1 (Rys. 2).



Rysunek 2: Regularyzacja jako rozkład apriori