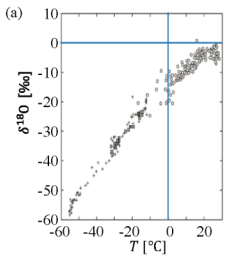
Relacja liniowa między temperaturą powierzchni Grenlandii a δ^{18}O

Joanna Michalska

15 listopada 2024

1 Wstęp



Rysunek 1: liniowa zależność stężeń δ^{18}O od średniej rocznej temperatury powierzchni Grenlandii. Źródło: link

W pracy chciałabym pokazać zależność liniową pomiędzy występowaniem stabilnych izotopów tlenu O^{18} oraz O^{16} a temperaturą powierzchni Grenlandii. Jest to sensowne zagadnienie, które może posłużyć oszacowaniu wieku lodu na Grenlandii. Zatem temat jest bardzo ważny.

2 Podstawowe charakterystyki danych

2.1 Źródło danych

Zagadnienie pochodzi z 44-rtej Międzynarodowej Olimpiady Fizycznej w Danii w 2013 roku i znajduje się pod linkiem: link do zadania; link do rozwiązania Oprócz tego dane pochodzą z NASA i mapka ilustrująca, gdzie wykonano pomiary jest pod linkiem (może chwilkę zająć ładownie strony- UWAGA), dane tabelaryczne są pod linkiem

2.2 Opisy podstawowych zmiennych

- Zmienne:

1. temperatura na Grenlandii, która jest jednocześnie zmienną objaśnianą, ponieważ przypuszczamy liniowy model: gdy temperatura jest wysoka np 10 stopni, wtedy nie ma opadów śniegu oraz ujawnia się starsza pokrywa lodowa wskutek topnienia lodowca. Wtedy docieramy do małych pokładów izotopów (wskutek zachodzenia połowicznego rozpadu, gdy ujawnia się "starszy" lód, wtedy jest mniej izotopów), w przeciwieństwie do niskich temperatur, gdy nie docieramy do tak głębokich pokładów izotopów. Starszy lód nie topnieje przy niskich temperaturach, natomiast możliwym jest, że przybywa świeżego śniegu. Wtedy też, przy niskich temperaturach powinniśmy otrzymywać wysokie odczyty poziomu izotopów. Potwierdzają to dane.
2. miara stabilnych tlenów $\delta^{18}O$, która jest jednocześnie zmienną objaśniającą

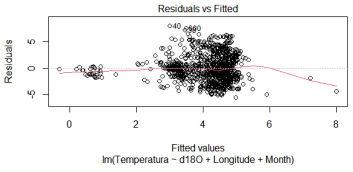
$$\delta^{18}O = \frac{R_{ice} - R_{ref}}{R_{ref}} 1000‰ \quad (1)$$

3. zasolenie, które może wpłynąć na odczyt stężeń
4. miesiąc pobrania próbki
5. szerokość geograficzna

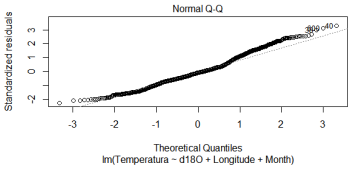
- Statystyki
  Model składa się z 1088 obserwacji, każda po 5 zmiennych: temperatury oraz stężeń $\delta^{18}O$, misiąca pobrania próbki, szerokości geograficznej i zasolenia.

1. co do mierzonej temperatury: min: -1.01 pierwszy quantil: 2,558, mediana: 3.650, średnia: 4,121 trzeci quantill: 5.705 , maximum: 11.02
2. co do mierzonego stężenia izotopów: min: 0 pierwszy quantil: 0,19 , mediana: 0,25 , średnia: 0,2417 , trzeci quantill: 0,3 , maximum: 1.03 pośród 568 danych.
3. co do mierzonej szerokości geograficznej: min: -75,92 pierwszy quantil: -42,1 , mediana: -40 , średnia: -37,49 , trzeci quantill: -30,68 , maximum: -11,01 pośród 568 danych.
4. R rozpoznaje zasolenie jako "char" jednak nie powinien. Nie mam danych więc na temat zasolenia i mimo chęci nie włączę go do modelu.
5. miesiące: próbki pobierane były w miesiącach 3-10, próba mieisęcy jest bardzo scentrowana wokół miesięcy 8-9.

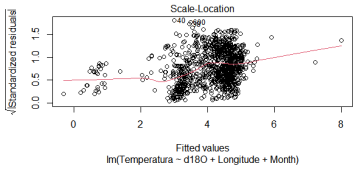
3 Model liniowy w oparciu o dane



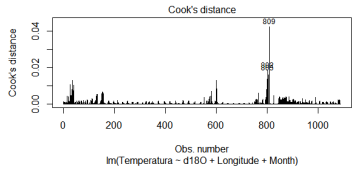
a



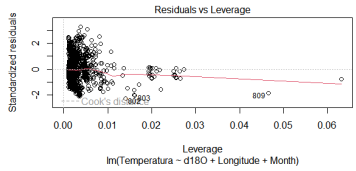
b



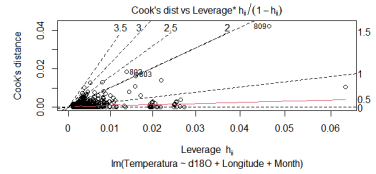
c



d



e



f

Komentarze:

- a) wpływowe obserwacje są zaznaczone poza dystansem Cooka zaznaczonego na czerwono. Widzimy wyraźnie wpływ odstających obserwacji oraz, że są niepotrzebne.
- b) Q-Q sprawdzenie normalności i empirycznej dystrybuanty, która okazuje się być bliska dystrybuancie rozkładu normalnego.
- c) sprawdzenie homeoskatyczności: gdy czerwona linia jest ustalona horyzontalnie możemy mówić o równej wariancji. Z rysunku wynika, że mają na model ogromny wpływ wartości odstające.
- d) Cook's distance a więc badanie wpływu poszczególnych obserwacji i ich istotności. Jest kilka wysokich dźwigni, ale ogólnie są sumarycznie niskie.
- e) Residua vs dźwignie: badanie wpływu poszczególnych obserwacji na model. Widzimy wyraźny wpływ wartości odstających. Chcielibyśmy, aby linia czerwona była horyzontalna.
- f) zbiorczy wykres dystansu Cook'a a dźwigni

4 Istotność całego modelu

Wedle obliczeń przeprowadzonych w R na podstawie analizy residuów przez test R^2 stwierdzamy na poziomie istotności 0.05, że nie możemy odrzucić hipotezy zerowej o niedopasowaniu modelu do danych. P-value na statystyce testowej wyniosło $2 * 10^{-16}$, podczas gdy statystyka testu F o 1084 stopniach swobody wyniosła 2,428.

Model Akaike Information Criterion oraz Bayesian Information Criterion dla modelu liniowego wcześniej zdefiniowanego, dla model2: liniowej zależności między Temperaturą a stężeniami delatO18 i szerokością, model3 dla liniowej zależności między temperaturą a stężeniami deltaO18 i miesiącem oraz model4: dla liniowej zależności między temperaturą a deltaO18. Nie sprawdzam liniowej zależności pomiędzy temperaturą a miesiącem, bo jest oczywista oraz nie sprawdzam liniowej lub innej zależności między temperaturą a szerokością geograficzną bo mija się z celem.

Wedle wyników z obu kryteriów AIC oraz BIC najniższą wartość początkową miał model wyjściowy, zatem ostatecznie uznajemy, że jest on istotny statystycznie.

5 Krótka interpretacja niematematyczna

Gdy temperatura jest wysoka np 10 stopni, wtedy nie ma opadów śniegu oraz ujawnia się starsza pokrywa lodowa wskutek topnienia lodowca. Wtedy docieramy do małych pokładów izotopów (wskutek zachodzenia połowicznego rozpadu, gdy ujawnia się "starszy" lód, wtedy jest mniej izotopów), w przeciwieństwie do niskich temperatur, gdy nie docieramy do tak głębokich pokładów izotopów. Starszy lód nie topnieje, natomiast możliwym jest, że przybywa świeżego śniegu. Wtedy też, przy niskich temperaturach powinniśmy otrzymywać wysokie odczyty poziomu izotopów, bo mamy do czynienia z "młodym" śniegiem, w którym nie zaszły na dużą skalępołowiczne rozpady. Potwierdza to model liniowy.

6 Diagnostyka

- Test normalności reszt.

  Jeżeli chodzi natomiast o współczynnik RSS, nie widać wyraźnego trendu pośród sumy kwadratów reszt modelu- zatem możemy wnioskować o homoskedastyczności. Jeśli chodzi o autokorelację reszt, na poziomie istotności 0.05 możemy odrzucić hipotezę zerową o autokorelacji reszt

- Testowanie poprawności formy funkcyjnej modelu metodą RESET.

  Na poziomie istotności p-value= 0.05 odrzucamy hipotezę zerową wobec hipotezy alternatywnej, że nasz model jest postaci $y = X\beta + Z\gamma + \psi$ lub $y = X\beta + Z\gamma + Z\gamma^2 + \psi$ lub $y = X\beta + Z\gamma + Z\gamma^2 + Z\gamma^3 + \psi$ a parametr $\gamma = 0$ na poziomie istotności p-value=0.05, otrzymane p-value= 0.10141 przy wartości statystyki testowej równej 3.7037.

  Wobec poprawności Testu RESET czyli Regression Equation Specification Error Test nie powinnam wykonywać przekształcenia Boxa-Coxa.

  Ponieważ zasadniczo w próbce nie występują skośności, nie musimy rozważać logarytmicznej funkcji modelu (patrz. podstawowe statystyki).

- Residua i czynnik losowy czynnik lodowy ma zerową wartość oczekiwaną, a jak pokażemy później, jest także liniowo niezależny.

  Jednakże ze względu na brak widocznego wzorca w kształtowaniu się residuów (rys.(a)) w modelu oraz przypuszczalnego braku dowodu dla odrzucenia hipotezy o występowaniu homoskedastyczności o którym najpewniej orzeknie test Breuscha-Pagana (bptest(bez_lm)), zamierzam również przyjrzeć się współczynnikowi RSS, aby ocenić homoskedastyczność i liniową niezależność błędów.

- Na poziomie istotności p-value=0.05 brak dowodu dla odrzucenia hipotezy o występowaniu homoskedastyczności wedle testu Breuscha-Pagana. Otrzymane p-value= 0,256 przy wartości statystyki testowej równej 1.2903.

- brak współliniowości: metodą VIF z biblioteki car orzekam, że ponieważ współczynniki poszczególnych zmiennych są bliskie 1, nie występuje żadna współliniowość.

7 Wnioski z modelu a

Na podstawie analiz, stwierdzam, że problemem mogą być wartości odstające. Postaram się je usunąć tak, aby stężenia deltaO18 nie były w przybliżeniu 1 i dla tak ustalonych danych przeprowadzę dalszą diagnostykę. Usuwam je, ponieważ na 1088 zmiennych jedna jest bliska 1. Wpływa istotnie na model i uważam, że jest to błądp omiaru ze względu na ogromne odstawanie od danych, których statystyki bez tej zmiennej przedstawię poniżej.

8 model b

8.1 opisy podstawowych zmiennych

– Zmienne:

1. temperatura na Grenlandii, która jest jednocześnie zmienną objaśnianą,
2. miara stabilnych tlenów $\delta^{18}O$, która jest jednocześnie zmienną objaśniającą

$$\delta^{18}O = \frac{R_{ice} - R_{ref}}{R_{ref}} 1000‰ \quad (2)$$

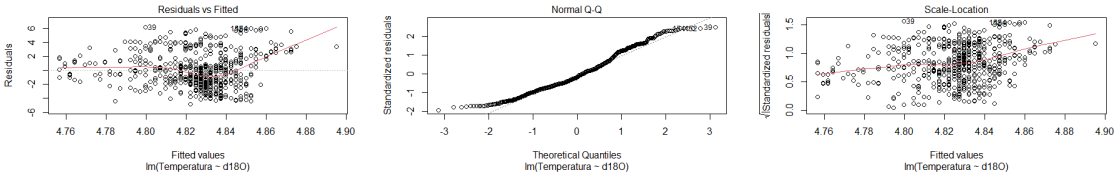
Chciałam dodać zmienne objaśniane "longitude" oraz "month" ale plik R nie współpracował.

– Statystyki
Model składa się z 568 obserwacji, każda po 2 zmienne: temperatury oraz stężeń δ^{18}O.

1. co do mierzonej temperatury: min: 0,03 pierwszy quantil: 3, mediana: 4,425, średnia: 4,825 trzeci quantill: 6,4 , maximum: 10,95
2. co do mierzonego stężenia izotopów: min: 0 pierwszy quantil: 0,24 , mediana: 0,28 , średnia: 0,27 , trzeci quantill: 0,33 , maximum: 0,55 pośród 568 danych.

8.2 Model liniowy w oparciu o dane

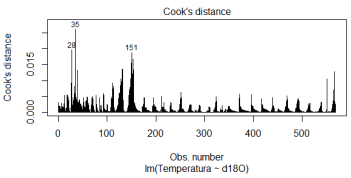
9 Model liniowy w oparciu o dane



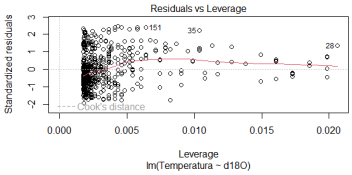
a b c

Komentarze:

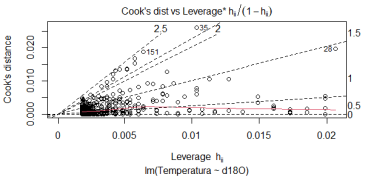
* a) wpływowe obserwacje są zaznaczone poza dystansem Cooka zaznaczonego na czerwono. Widzimy wyraźnie, że im większe wartości, tym większy mają wpływ na model.
* b) Q-Q sprawdzenie normalności i empirycznej dystrybuanty, która okazuje się być jeszcze bliższa dystrybuancie rozkładu normalnego.



d



e



f

* c) sprawdzenie homeoskatyczności: gdy czerwona linia jest ustalona horyzontalnie możemy mówić o równej wariancji. Z rysunku wynika, że mają na model ogromny wpływ wartości większe.
* d) Cook's distance a więc badanie wpływu poszczególnych obserwacji i ich istotności. Jest kilka niezbyt wysokich dźwigni i ogólnie dźwignie nie są wysokie.
* e) Residua vs dźwignie: badanie wpływu poszczególnych obserwacji na model. Widzimy wyraźny wpływ wartości odstających.Czerwona linia jest niemal horyzontalna.
* f) zbiorczy wykres dystansu Cook'a a dźwigni. WIdzimy, że najlepszy model uzyskalibyśmy dla wartości bliskich 0.

10 Istotność modelu

Współczynnik R^2: dla statystyki F o 566 stopniach swobody p-value wyniosło 0.8392, stąd na poziomie istotności 0.05 odrzucamy hipotezę zerową o dobrym dopasowaniu danych do modelu przy jednocześnie niskiej bo równej $7,281 * 10^{-5}$ wartości R^2. Dodatkowo, model AIC i BIC wskazuje jednak niższą wartość dla modelu b, jednakże z metodologi testów AIC i BIC wiadomo, że może to być spowodowane bezpośrednio tym, że model b ma mniej parametrów niż model pierwszy. Stąd też najbardziej martwiącym jest niska miara dopasowania współczynnika R^2.

11 Diagnostyka

1. Ze względu na brak widocznego wzorca w kształtowaniu się residuów (rys.(a)) w modelu oraz przypuszczalnego braku dowodu dla odrzucenia hipotezy o występowaniu homoskedastyczności o którym najpewniej orzeknie test Breuscha-Pagana (bptest(bez_lm)), zamierzam również przyjrzeć się współczynnikowi RSS, aby ocenić homoskedastyczność i liniową niezależność błędów.
2. Na poziomie istotności p-value=0.05 odrzucamy hipotezę zerową o występowaniu homoskedastyczności wedle testu Breuscha-Pagana. Otrzymane p-value= $1,865 * 10^{-7}$.
3. Testowanie poprawności formy funkcyjnej modelu metodą RESET.
   Na poziomie istotności p-value= 0.05 odrzucamy hipotezę zerową wobec hipotezy alternatywnej, że nasz model jest postaci $y = X\beta + Z\gamma + \psi$ lub $y = X\beta + Z\gamma + Z\gamma^2 + \psi$ lub $y = X\beta + Z\gamma + Z\gamma^2 + Z\gamma^3 + \psi$ a parametr $\gamma = 0$ na poziomie istotności p-value=0.05, otrzymane p-value= 0.10141 przy wartości statystyki testowej równej 3.7037.
   Wobec poprawności Testu RESET czyli Regression Equation Specification Error Test nie powinnam wykonywać przekształcenia Boxa-Coxa.
   Ponieważ zasadniczo w próbce nie występują skośności, nie musimy rozważać logarytmicznej funkcji modelu (patrz. statystyki).
4. występowanie autokorelacji oraz orzekanie o zerowej wariancji reszt
   Wedle testu Breusch-Godfrey'a na poziomie istotności 0.05 możemy odrzucić hipotezę zerową o występowaniu autokorelacji reszt, ponieważ p-value wyniosło $2,2 * 10^{-16}$
5. Mamy do czynienia zatem z modelem liniowym.

12 Porównanie modeli a i b

1. analiza współczynników BIC i AIC: jak pisałam wcześniej, analiza tych dwóch czynników wskazuje za model b jako lepszy. Może być to związane z tym, że model b ma mniej czynników. Jednakże związane testy metodologicznie czyli wswpółczynniki R^2 wskaują model a za lepszy. Dodatkowo, stosowałam metodę usuwania zmiennych objaśniających aby sprawdzić, czy polepszy się współczynnik AIC lub BIC. Niestety, model a był modelem lepszym.
2. Porcedury pojedynczego doboru zmiennych w modelach AIC i BIC również wskazywały na to, że model a ma najmniejszą możliwą liczbę zmiennych.
3. Obydwa modele mająwysokie prawdopodobieństwo bycia modelami liniowymi, o czym mówi test RESET.
4. Skorygowany współczynnik R^2 modelu b jest bliski 0, podczas gdy skorygowany współczynnik modelu a jest bliski 1/2.

13 Wnioski

Po przedstawieniu wszystkich tych powodów takich jak analiza AIC, BIC, skorygowany wspólczynnik R^2, procedury pojedynczego doboru zmiennych oraz metod RESET i po całych diagnostykach modeli, stwierdzam z całą pewnością że model pierwszy jest lepszy od modelu b.