回帰分析I

5. 確率の基礎

イントロダクション

- これから確率論を扱います。
- なぜ確率論を？
- 回帰分析では、観測されるデータの振る舞いやバラつきを確率的な事象として取り扱うから。
- データは観測前には不確定なもの。
  - サイコロ投げる前
- 観測して初めて確定、って考え方。
  - サイコロ投げた後

イントロダクション

- 特に大切となる概念は以下のものです（次回以降に出てきます）：
  - 確率変数
  - 期待値
  - 分散（標準偏差）
  - 共分散
  - 相関
  - 条件付き分布
  - 条件付き期待値

この講義ノートの内容で最も重要なものは「条件付き確率」です。

「条件付き期待値」を理解するために必要となります。

- 回帰分析では、変数は（基本的には）確率変数として取り扱います。
- 回帰分析の重要な仮定は、期待値、分散、共分散（または相関）、条件付き期待値を使って表現されます。
- このような確率の概念それ自体の学習はどちらかというと退屈ですが、回帰分析をよりよく理解するためには必要ですので頑張っていきましょう。

標本空間と事象

- 結果（outcome）を前もって予測することのできない実験（experiment）を行なうとする。
- 結果を前もって知ることは出来ないが、起こりうる結果がどのようなものかはすべて知っているとする。
- この起こりうるすべての結果を「**標本空間**（sample space）」という。
  - 以下では標本空間をSで表す。

標本空間と事象

- 例1：もし実験が「コイン投げ（flip a coin）」だとすると、

$$S = \{H, T\}$$

Hはコイン投げの結果が表、Tはコイン投げの結果が裏であることを意味する。

- 例2：もし実験が「さいころ投げ（toss a dice）」だとすると、

$$S = \{1, 2, 3, 4, 5, 6\}$$

結果iは i がさいころの目としてでたことを表す。

標本空間と事象

- 例3：実験が二枚のコインを投げるものだとする。このとき標本空間は以下の4つの点からなる：

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

(H,H) は最初のコインが表、2枚目のコインも表
(H,T) は最初のコインが表、2枚目のコインが裏

以下同様

- 例4：「ある車が故障するまでの時間」を計測する実験を考えよう。このとき

$$S = [0, \infty)$$

標本空間と事象

- 標本空間Sの部分集合を「**事象**（event）」という。
  - 以下では事象をEやFで表す（AやBを使うこともある）。
- 例1を考えよう。もし$E = \{H\}$なら、E は「コインを投げたときに表が出る」という事象。
- もし$E = \{T\}$なら、E は「コインを投げたときに裏が出る」という事象。

標本空間と事象

- 例2を考えよう。もし$E=\{1\}$なら、Eは「さいころを投げたとき1がでる」という事象。
- もし$E=\{2,4,6\}$なら、Eは「さいころを投げたとき偶数（even number）がでる」という事象。
- 例3を考えよう。もし$E = \{(H,H),(H,T)\}$なら、Eは「1枚目のコインが表」という事象。
- 例4を考えよう。もし$E = (2,6)$なら、Eは「車が2年から6年の間に故障する」という事象。

標本空間と事象

- 標本空間Sのいかなる2つの事象EとFについて、新しい事象

$$E \cup F$$

を次のように定義する：

「E またはFに属するすべての点」（E<u>または</u>F）

- すなわち、もし事象E か事象Fのどちらかが起きたときに、事象$E \cup F$ が起きる（**和事象**と呼ぶ）。
- 例2を考えよう。もし$E = \{1,3,5\}$で$F = \{1,2,3\}$なら、

$$E \cup F = \{1, 2, 3, 5\}$$

- すなわち、もしさいころ投げの結果が1,2,3,5のいずれかであれば、事象 $E \cup F$ が起きたという。

標本空間と事象

- 標本空間Sのいかなる2つの事象EとFについて、新しい事象

$$E \cap F$$

を次のように定義する：

「E とFの両方に属するすべての点」（E<u>かつ</u>F）

- すなわち、もし事象E と事象Fの両方が起きたときに、事象$E \cap F$ が起きる（**積事象**と呼ぶ）。
- 例2を考えよう。もし$E = \{1,3,5\}$で$F = \{1,2,3\}$なら、

$$E \cap F = \{1, 3\}$$

- すなわち、もしさいころ投げの結果が1か3のいずれかであれば、事象 $E \cap F$ が起きたという。

標本空間と事象

- 例1を考えよう。もし$E = \{H\}$で$F = \{T\}$なら、事象 $E \cap F$ はいかなる点も含まない（従ってこの事象は起きない）。
- このような事象は「**空事象**（null event）」と呼び、ϕと表す。
- $E \cap F = \phi$ のとき、事象EとFは「**排反**（mutually exclusive）」であるという。
  - どっちかが起きたらもう一方は起きないケース

標本空間と事象

- いかなる事象Eについて、事象E^Cを次のように定義する：

「事象Eに含まれ<u>ない</u>標本空間Sのすべての点」

- これを事象Eの「**余事象**（complement of E）」と呼ぶ。
- 例2を考えよう。もし$E = \{1,3\}$なら、$E^C=\{2,4,5,6\}$。
- 実験した場合何らかの結果が起きるので、標本空間Sの余事象は空事象：

$$S^C = \phi$$

確率

- 標本空間がSである実験を考えよう。
- 標本空間Sにおけるそれぞれの事象Eに、1つの数$P(E)$を与え、その数が次の3つの条件を満たすとする。

（1）$0 \le P(E) \le 1$
（2）$P(S) = 1$
（3） 標本空間Sにおける2つの事象E_iとE_jが排反なら（すなわち $E_i \cap E_j = \phi$ なら）、$P(E_i \cup E_j) = P(E_i) + P(E_j)$

このとき$P(E)$を事象Eの「**確率**（probability）」という。

確率

- 例1を考えよう。もし表が裏が同じように出やすいとすれば、

$$P(\{H\}) = P(\{T\}) = 1/2$$

- コインがバイアスしていて表が裏より2倍出やすいとすれば、

$$P(\{H\}) = 2/3;\ P(\{T\}) = 1/3$$

- 例2を考えよう。それぞれの目が同じようにでやすいとすれば、

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = 1/6$$

- 条件3より、偶数が出る確率は

$$P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = 1/2$$

確率

- ノート：ここでは確率を「標本空間の事象の上に定義された関数」とした。
- これはフォーマルな定義。
- これらの確率は「直観的」な特性を持つ。
  - 実験を何回も（とても多くの数）繰り返した場合に、事象Eの起きる回数の割合は$P(E)$になる。

確率

- 次にいくつかの重要な式について説明する。
- 事象Eと事象E^Cはつねに排反、また $E \cup E^C = S$ なので、条件（2）と（3）より

$$1 = P(S) = P(E \cup E^C) = P(E) + P(E^C)$$

または

$$\boxed{P(E^C) = 1 - P(E)}$$

- これは「ある事象が起こらない確率」は「1マイナスその事象が起こる確率」であることを示している（当然、、、）。

確率

- 次に$P(E \cup F)$の式がどうなるか考えよう。
  - 事象E またはFが起こる確率
- まず$P(E)$ + $P(F)$を考えよう。
  - $P(E)$: 事象Eのすべての点の確率
  - $P(F)$: 事象Fのすべての点の確率
- $P(E)$ + $P(F)$では、事象Eと事象Fの両方に属している点は2回数えられている。
- 一方 $P(E \cup F)$ では、事象Eと事象Fの両方に属している点は1回だけ数えられている。

17

確率

- よって次の関係が成り立つ。

$$P(E)+P(F)=P(E \cup F)+P(E \cap F)$$

- 同じことだが、

$$\boxed{P(E \cup F)=P(E)+P(F)-P(E \cap F)}$$

- もし事象EとFが排反（すなわち$E \cap F=\phi$）なら、

$$P(E \cup F)=P(E)+P(F)-P(\phi)=P(E)+P(F)$$

- この結果は条件（3）からも得られる。

18

確率

- Q: 2枚のコインを投げるとする。このとき標本空間は

$$S=\{(H,H),(H,T),(T,H),(T,T)\}$$

これら4点が等しく起こりやすいとする（よってそれぞれの確率は1/4）。

次の2つの事象EとFを考えよう。

$$E=\{(H,H),(H,T)\}; \quad F=\{(H,H),(T,H)\}$$

このとき $P(E \cup F)$ は？

19

確率

- A: $P(E \cup F)=P(E)+P(F)-P(E \cap F)$ を使う。

$$\begin{aligned}P(E \cup F)&=P(E)+P(F)-P(E \cap F)\\&=\frac{1}{2}+\frac{1}{2}-P(\{H,H\})\\&=1-\frac{1}{4}=\frac{3}{4}\end{aligned}$$

- この確率はこの式を使わずに直接的にも求められる。

$$P(E \cup F)=P(\{(H,H),(H,T),(T,H)\})=\frac{3}{4}$$

20

条件付き確率

- 2つのさいころを投げる実験を考えよう。
- 起こりうるすべての結果、すなわち標本空間は

$$S = \left\{ \begin{matrix} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6) \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6) \\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6) \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6) \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6) \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \end{matrix} \right\}$$

- それぞれの点が等しく起こりやすいとする（よってそれぞれの確率は1/36）。

21

条件付き確率

- ここで次のような状況を考えよう。

「1個目のさいころを投げたときに4が出た」

- この情報が与えられらとき、2つのさいころの目の和が6になる確率は？
- この確率を計算するために次のように考えることが出来る。
- 1個目のさいころが4だと、起こりうる結果は次の通り:

$$(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)$$

22

条件付き確率

- それぞれの結果は同じように起こりやすい。
- すなわち、最初のさいころの目が4の場合、

$$(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)$$

のそれぞれの（条件付き）確率は1/6。標本空間にある他の30の点の（条件付き）確率は0。

- したがって2つのさいころの目の和が6になる確率は1/6。

23

条件付き確率

- もしEを「2つのさいころの目の和が6」という事象、Fを「最初のさいころの目が4」という事象とすると、今求めた確率は

「Fが起きたという条件の下でEが起きる確率」

または

「Fが起きた場合にEが起きる**条件付き確率**」

という。

- これを

$$\boxed{P(E \mid F)}$$

と表す。

24

条件付き確率

- $P(E \mid F)$ の確率は

$$\boxed{P(E \mid F) = \frac{P(E \cap F)}{P(F)}}$$

で与えられる。ただしこの確率は<u>$P(F) > 0$のときにだけ定義される</u>。

- 上の式は例と同じようなロジックで導出できる。
- 事象Fが起きたとする。
- このとき、事象Eが起きるためには、結果は事象EとFの両方に属する点でなければならない。

条件付き確率

- すなわち、その点は事象 $E \cap F$ に属する点でなければならない。
- また、事象Fが起きたことを知っているので、事象Fが新しい標本空間になる。
- 従って、Fが起きた条件の下で $E \cap F$ が起こる確率は、$E \cap F$ が起きる確率をFが起こる確率で割ったものになる。
- 上の式は変形して次の形でも良く使う:

$$P(E \cap F) = P(E \mid F)P(F)$$

条件付き確率

- Q:1から10までの番号がつけられた10枚のカードが帽子の中に入っている。そこから1枚のカードが引かれるとする。
- もし引かれたカードに書いてある数が少なくとも5だとすれば、そのカードの数が10である(条件付き)確率は?
- A:Eを「引かれたカードに書いてある数が10である」という事象、Fを「引かれたカードに書いてある数が少なくとも5である」という事象だとする。
- このとき求める確率は$P(E|F)$。
- よって

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E)}{P(F)} = \frac{1/10}{6/10} = \frac{1}{6}$$

条件付き確率

- Q:箱に黒いボールが7個、白いボールが5個入っているとする。この箱から2つのボールを引くとする(取ったボールは箱の中にもどさない:without replacement)。それぞれのボールが引かれる確率は同じとする。
- このとき引かれたボールが両方とも黒の確率は?
- A: Fを「1番目に引かれたボールが黒」という事象、Eを「2番目に引かれたボールが黒」という事象だとする。
- 最初に引かれたボールが黒なら、箱の残りは、黒いボールが6個、白いボールが5個。よって$P(E|F) = 6/11$。
- $P(F) = 7/12$は明らか。
- 従って

$$P(E \cap F) = P(E \mid F)P(F) = \frac{6}{11}\frac{7}{12} = \frac{42}{132}$$

（確率的）独立

2つの事象EとFについて、もし

$$P(E \cap F) = P(E)P(F)$$

ならEとFは「**独立**（independent）」という。

- 条件付き確率の式を思い出そう:

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}$$

- もしEとFが独立なら

$$P(E \mid F) = \frac{P(E)P(F)}{P(F)} = P(E)$$

（確率的）独立

もしEとFが独立なら、

$$P(E \mid F) = P(E)$$

が成り立つ。

「EとFが独立」

//

Fが起きたこと（または起きなかったこと）は、Eが起こる確率とは何の関係もなし。

- EとFが独立でないとき、EとFは「**従属**（dependent）」という。

（確率的）独立

- 例：2つのさいころを投げるとする。E_1を「2つのさいころの目の和が6」という事象、E_2を「最初のさいころの目が4」という事象とする。

- このとき

$$P(E_1 \cap E_2) = P(\{4,2\}) = \frac{1}{36}$$

- 一方

$$P(E_1)P(E_2) = \frac{5}{36}\frac{1}{36} = \frac{5}{216}$$

- 従ってこの2つの事象は独立ではない。

（確率的）独立

- 直感的な理解：
  - 2つのさいころの目の和が6になると勝ちというゲームをあなたがやっているとする。
  - 最初のさいころの目が4だったら、まだあなたにはチャンスあり。
  - 最初のさいころの目が6だったら、もうチャンスなし。
  - 別の言い方をすれば、さいころの目の和が6になるという事象は、最初のさいころの目の結果に依存している。
  - したがって、E_1とE_2は独立ではない。