演習問題

復習問題

1　総積記号

次の内容を総積記号を使わずに表現しなさい。

(1) $\prod_{i=1}^{3} x_i$
(2) $\prod_{i=1}^{3} p(x_i)$
(3) $\log[\prod_{i=1}^{3} p(x_i)]$
(4) $\prod_{i=1}^{3} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$
(5) $\log\left[\prod_{i=1}^{3} \pi_i^{y_i}(1-\pi_i)^{1-y_i}\right]$

基礎問題

2　ロジスティック回帰モデルの使用

次の(1)から(5)に挙げた状況の中で、二項ロジスティック回帰モデルを利用して分析をすることが最も望ましいものには「A」と、多項ロジスティックモデルを利用して分析することが最も望ましいものには「B」と、どちらの使用にも妥当性がないものには「C」と記しなさい。

(1) 都市部にあるのか地方にあるのか、偏差値がどのくらいなのかという二つの独立変数から、各大学の 1 年生に占める女子学生の割合をモデル化したい。
(2) 動詞が過去形なのか現在形なのか、直接目的語の長さが 4 単語以上なのか、という構文から「give 間接目的語　直接目的語」という語順を取るのか、「give 直接目的語 to 間接目的語」という語順を取るのかを予測したい。
(3) 話者の社会階層と、聞き手の社会階層から、各話者が語末の[r]という音をどのくらいの割合で発音するのか予測できるかどうか知りたい。
(4) ある大学では、理系学生の入試に「地理」「世界史」「日本史」のどれをいずれかを一つ選択して回答するように課している。各学生が私立高校の生徒か、国公立高校の生徒かからどの科目を選択するのかを予測したい。
(5) ある大学の学生 30 人に、ある文の容認度を 1，2，3，4，5 の中から一つを選んで回答してもらった。その文に言語パターン A が含まれているか否かで、各学生がどの回答を選択したのかを予測したい。

3　重回帰とロジスティック回帰の比較 I

問１ 復習（重回帰モデル）

次の[A]に示された状況を重回帰分析のモデルで分析しようと、ある二人の学生が、i 番目の従属変数 y の値を予測するモデルとして[B]と[C]の数理的構造を提案した。それぞれのモデルについて、この数学的定式化が正しいものか判断し、誤りがあれば指摘し、正しく修正しなさい。

[A] 全部で n 人の大学 1 年生の学生の身長を測定し、それを従属変数として分析対象に据えた。これらの学生の身長を、性別と右手中指の長さという二つの独立変数から構造的に予測でき、各学生の身長のその予測値からのずれ（すなわち誤差）は、標準偏差がσ の正規分布に従うと仮定する。

[B]

$$y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

[C]

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

問２ （ロジスティック回帰モデル）

次の[D]の状況を前に、母集団にロジスティック回帰モデルを想定して分析をすることにした。どのような数理的モデルを提案することが適切か、問 1 の構造を参考に、必要な関係を表したモデルを式で表現しなさい。

[D] 全部で n 個の単数の可算名詞を対象に調査を行った。i 番目の事例における従属変数としてその名詞が a/an $(y_i = 0)$を取るか the $(y_i = 1)$を取るかを、i 番目の事例において直前の文で同じ可算名詞が用いられているかを独立変数x_iとして表記し、後者で前者を予測したい。

[E] 全部で n 個の異なる種類の可算名詞を対象に調査を行った。i 番目の種類の可算名詞がそのコーパスで観察された頻度をn_i回、そのうち the を伴った頻度をy_i回と表記する。その可算名詞が物理的に手で触れるものを指すか否かを表す独立変数x_iからこのy_iを予測したい。

4　重回帰とロジスティック回帰の比較Ⅱ

次の文章を読んながら、問いに答えなさい。

とくがわ　もー、やんなった！

とよとみ　わぁ、とくがわちゃん、どうしたの…！

とくがわ　わたしは、前期の授業、相当がんばったのだ！

とよとみ　お、よよ…？

とくがわ　重回帰分析のこと！頑張って覚えた！大変だった！でもめげなかった！独立変数と従属変数の関係を表す式も書けるようになった！

とよとみ　いや、分かるよ。悪態付きながらも、私のノート借り続けて、統計学んでたよね。最初から授業中に自分でノート取ればいいのになって、思ってたけど。

とくがわ　にもかかわらず！後期になって、ロジスティック回帰ってのはじまって、重回帰の拡張だっていうから、私はもう楽ができるって思ってたの！そしたら、もー全然似ても似つかない式が出てきて、私は悟ったの！きっと、私と統計学はきっと前々前世で敵同士だったんだよ！

とよとみ　それってどんな輪廻観？！

とくがわ　だから、とよとみちゃん！たすけてよー！教えてよー！

とよとみ　や、もちろん、別にいいけど、具体的に、こう、どこがこう分からないとかって、あるといいんだけど。

とくがわ　これです！これ！重回帰分析の式とロジスティック回帰の式のつながり！！！徹頭徹尾、とっても似てない！！

重回帰式のモデル（1）	ロジスティック回帰のモデル
y_i = [　A　] + [　B　]	①　$y_i \sim Binom($ [D], [E] $)$
[　B　] ~ [　C　]	②　[E] = [　F　]
	③　[G] = [　A　]

問１　空欄 A から G に入る数式、母集団パラメータや確率分布を記しなさい。ただし、複数の変数変換が想定できる時には、一般的な関数として、F という記号を用いなさい。

とよとみ　あー。言ってることは、なんとなく分かる。でも、実はこれ、左の重回帰モデルの方を少しいじると、おんなじ形になるんだよ。

とくがわ　そんなばかな！

とよとみ　同じモデルでも、別の視点から眺めると異なる定式化ができるっていうことは結構あって、じゃあ、ちょっと一緒にやってみようよ。

とくがわ　わたしにもできる？

とよとみ　できるできる！まずね、この重回帰のもともとの式を考えてみて。これは、どういう意味だったか覚えてる？

とくがわ　それはわかる！i 番目の y の値は[　A　]で表される直線でまず中心が決まって、そこから、ランダムに[　B　]で表される誤差項で上へ下へ値がぶれるっていうことでしょ！その誤差項は、[　C　]で定められた分布に従うってこと！

とよとみ　おー。あってる、あってる。じゃあ、ここで視点を変えたいんだけど、いま、分布に従っているのは、誤差項だったじゃん。じゃあ、y_iは、どういう分布に従っているって言えるか分かる？

とくがわ　えーと、さっきは[　B　]の後ろに、にょろって「~」マークがあったけど、[　B　]ではなく、y_iの後ろに「~」を続ける標記にしろってこと？

とよとみ　そう。それを書いてみると、こんな感じになるの。

重回帰式のモデル（２）	
$y_i \sim N([\ H\ ],[\ I\ ])$	

とくがわ　あー、えー、簡潔になって、短くなったけど、パーツは（１）と似てるとこもあるって感じ？

とよとみ　そだね。この（２）のモデルは、「i 番目の y はどこを中心に正規分布をするかというと、それは[　H　]。その中心からの分散は[　I　]です」、そういう書き方になってるの。

問２ 空欄 H と I に入る数式を記しなさい。

とくがわ　モデル（１）とは違うの？

とよとみ　モデル（1）もモデル（2）も同じことを表現しているけど、違う視点から記述してるの。モデル（1）は、「i 番目の y は、[　A　]という構造的な部分と、[　B　]というランダムな部分から予測できます。そして、ランダムなパートは、平均が[　C　]で、分散が[　D　]です」っていう書き方。

とくがわ　むむむ…。そうか、同じことを意味しているのかー！

とよとみ　また別の書き方もできるよ。次の（3）を見て。これは、[　H　]っていうのを、そのまま、N のパラメータの位置に書くと長くなるから、とりあえず、それをμ_iと置いて、それがいったい何なのかを、別途下に書き加えただけ。

重回帰式のモデル（2）	重回帰式のモデル（3）
$y_i \sim N([\ H\ ],[\ I\ ])$	$y_i \sim N(\mu_i,[\ I\ ])$ $\mu_i = [\ H\ ]$

とくがわ　これが、ロジスティック回帰のモデルと対応するの？

とよとみ　あともう 1 ステップあるんだけど、もうだいぶロジスティック回帰との共通点が見え始めてるよ。ロジスティック回帰のモデルの①番目の式は、「y_iが確率分布に従っています。そのパラメータはこれこれです」っていう情報でしょ。それは、（3）の式の一行目と同じでしょ。

とくがわ　なるほど！そして、（3）の式の二行目は、③番の式に対応しているってこと！？一番最後が[　A　]と[　H　]で終わってるから！

とよとみ　ざっくり言うと、そんな感じ。でも、正確に言うと、（3）の式の二行目は②番目の式と③番目の式を合体した情報をいっぺんに表現しているの。ほら、②と③の式を併せると、母集団のパラメータと[　E　]と独立変数の[　J　]との関係を表しているでしょ。$\mu_i = [\ H\ ]$っていうのも、まさにそうでしょ。

問３ 空欄 J には、偏回帰係数と独立変数の積たちと切片との和を表す用語が入る。これが何か、漢字四文字で答えなさい。

とくがわ　わかる…！けどー、ロジスティック回帰が重回帰の発展って言われるのはー、このあたりの②と③の複雑さがあるから？なんか、ちょいとわかったような、そうでないような！

とよとみ　あ、じゃあ、こうしてみるね。ロジスティック回帰と重回帰を比較するために、最後にもう一度（３）の式を変換してみて、対応番号を振ってみたよ。

重回帰式のモデル（４）	ロジスティック回帰のモデル
① $y_i \sim N(\mu_i, [\ I\ ])$	① $y_i \sim Binom($ [D], [E] $)$
② $\mu_i \quad = I([\ \mathrm{K}\ ])$	② [　E　] = [　F　]
③ [　K　] = [　L　]	③ [　G　] = [　A　]

とくがわ　！！！何ということでしょうー！何一つ共通点がなさそうに見えたあの二ページ前のモデル式たちが、各行に細やかな対応関係を描き出す親しみ溢れるお姿に！

とよとみ　①番目が「~」の式で、③番目が独立変数の[　J　]に関する式。②も左が母集団パラメータで、右が[　K　]や[　G　]に関数がくっついているっていう点で同じ。

とくがわ　この重回帰の方の②の式のＩって、何？

とよとみ　これはね、[　M　]って言って、[　N　]っていう関数なの。これは、単に、ロジスティック回帰との平行性を明示的に表すためにあえて、無意味に見えるけど差し挟んでるの。

問４　空欄ＫとＬに、適切な数学的表現を入れなさい。
問５　空欄Ｍには、このＩという関数の名称が入る。これが何か答えたうえで、その定義として適切な説明をＮに入れなさい。

とくがわ　ふむふむ…！あーとさ！（４）の式では、②番でμ_iに対しての式が作られてるけど、ロジスティック回帰だと[　E　]じゃん！μ_iの方はいいの！「平均をモデル化してます」って分かりがいいから！でも、「[　E　]をモデル化します」って、なんか、こう、すっとね、入ってこない！…やんなっちゃう！

とよとみ　うーんと、そこにも別に実は大きな乖離はなくて…、ベルヌーイ分布$Bern($ [Ｅ] $)$の[　O　]ってさ、[　P　]だったじゃん。だから、ロジスティック回帰の②の式も、本質的には、重回帰と同じで[　O　]をモデル化してるんだよ。

とくがわ　はー！ほー！こういうの考えついた人は、これを勉強してるわたしと同じくらい偉いのかもしれない！٩(。·ᴗ·。)۶

とよとみ　…それをとくがわちゃんに説明してる私は？ (´-﹏-`;)

問6 空欄 O には、確率分布の性質を示すある特徴量が入る。これが何か解答した上で、空欄 P に、ベルヌーイ分布$Bern($[E])におけるその値を書きなさい。

5 点推定値の解釈 I：二項ロジスティック回帰の場合

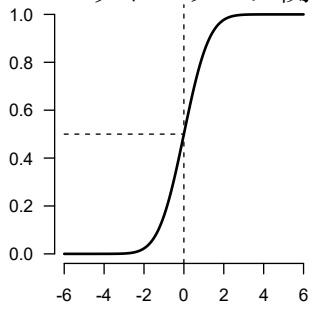
0 か 1 の値を取る従属変数 y を複数回観察し、それを独立変数の値たちから予測する二項ロジスティック回帰を構築した。モデル比較の結果、独立変数x_4を一つだけ含むモデルが複数の情報量基準の観点から「ベスト」なモデルだという結論を得た。

この「ベスト」なモデルにおける点推定の結果とそのワルド検定の結果を表の形で表したものが次のテーブルである。

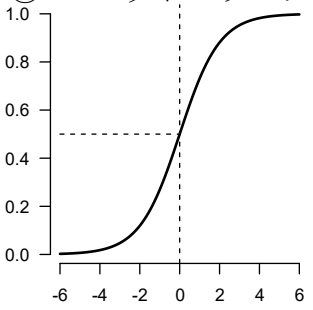
	点推定値	標準誤差	z 値	p 値
切片$\hat{\beta}_0$	0.9163	0.7738	1.184	0.2364
係数$\hat{\beta}_4$	1.7136	0.8871	1.932	0.0534

なお、モデル比較の過程では、独立変数の数だけではなく、下記の図に示される四つのリンク関数を試しており、ベストなモデルに利用されていたのは、「clog-log リンク関数」であった。

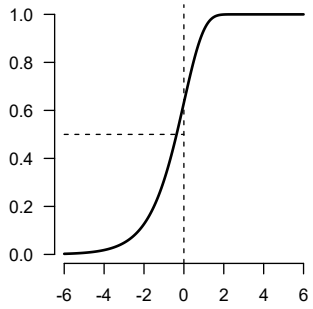
① プロビット・リンク関数



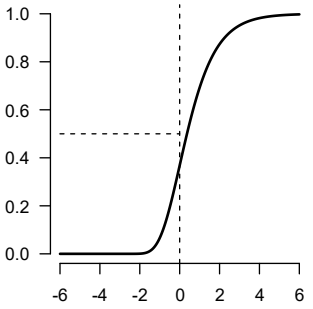
② ロジット・リンク関数



③ clog-log 関数



④ log-log 関数



これらの情報を踏まえて、次の問いに答えなさい。

問１ 次の解釈うち、正しいものには〇を、誤っているものには×を付しなさい。

(1) 切片の p 値は非常に大きい。これは、得られたデータが、帰無仮説「切片は 0 である」という命題を否定しうるだけの根拠を持っていないことを示している。
(2) cloglog リンク関数で最もパフォーマンスがよかったということは、「独立変数の線形結合の値が上がり 1 を取り始めるときには徐々に徐々に 1 を取り始めるようになる緩やかな変化が存在する一方で、さらに独立変数の線形結合の値が上がると、突然ほぼすべてが 1 を取るようになるという急激な変化に傾向が変わる」性質がある、と言えそうである。

問２ この推定結果をもとにすると、独立変数x_4が-2という値をとるとき 1 を取る確率はどのくらいになると考えられるか。最も近いものを次の中から選び記号で答えなさい。

①ほぼ 0%　②約 25%　③約 33%　④約 50%
⑤約 66%　⑥約 75%　⑦ほぼ 100%