

ロジスティック回帰分析

米倉 佑貴

聖路加国際大学大学院

看護学研究科看護情報学分野

今回の内容

- ロジスティック回帰分析
 - なぜ必要か
 - 変数間の関連性の見方
 - 結果の論文での示し方の例
 - モデルの当てはまりの指標
 - 変数の選択の仕方
 - 必要なサンプルサイズ
 - ロジスティック回帰分析での3カテゴリ以上の目的変数の扱い

ロジスティック回帰分析の概要

- 二値データや確率などのデータ(変数)を目的変数とした回帰分析
 - 疾患の有無
 - 要介護認定の有無
 - 有病率
- 目的変数が二値データ・確率データなだけで、考え方は重回帰分析などの一般線形モデルと同じ
 - 説明変数の選択, 投入の仕方や回帰係数の解釈の考え方も共通
 - 回帰係数の解釈
 - 回帰係数の値は他の変数の値を固定して1つの変数だけ変化させたときの目的変数との関連性の強さを表す
 - ただし, ロジスティック回帰分析では回帰係数の値は説明変数1単位あたりの目的変数の変化量ではなく, 対数オッズの変化やオッズ比として解釈する
 - 質的変数を説明変数として使うときはダミー変数を使う

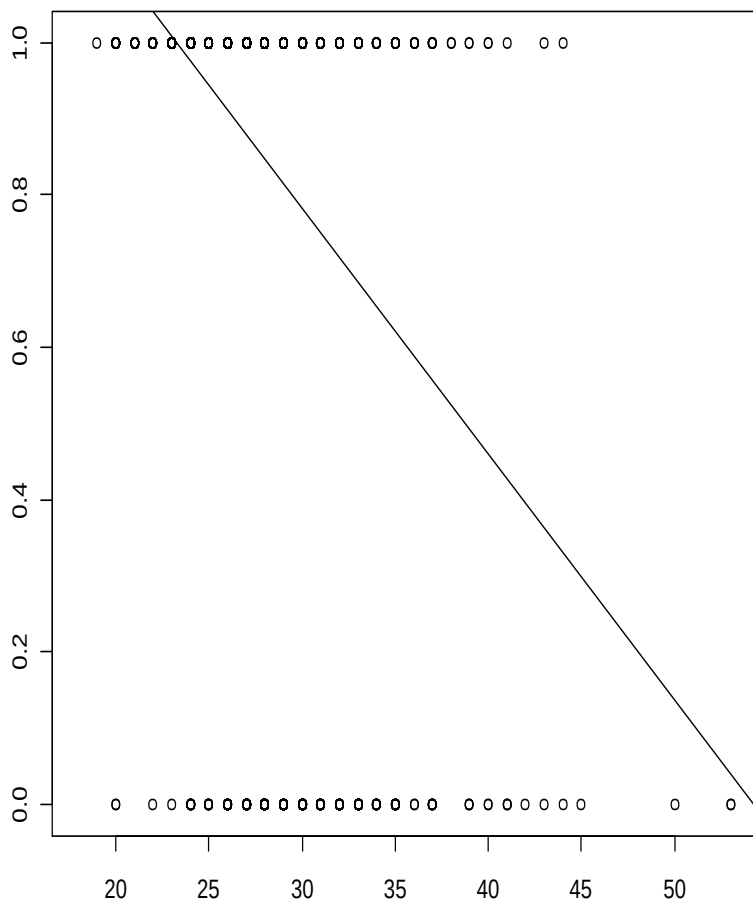
ロジスティック回帰分析はなぜ必要か

- 重回帰分析など一般線形モデルでは残差に以下の仮定をおいている
 1. 正規分布に従う
 2. 互いに独立
 3. グループ間（回帰ならすべての区間）で分散が等しい
 - 制約が満たされていないと
 - 誤った結果
 - P値が不正確に
 - 検出力の低下
 - モデルで予測する値がありえない範囲になる
 - **確率, 比率のデータがマイナスになったり1を超えたりする**
- 制約を満たさないデータは一般線形モデルでは扱えない

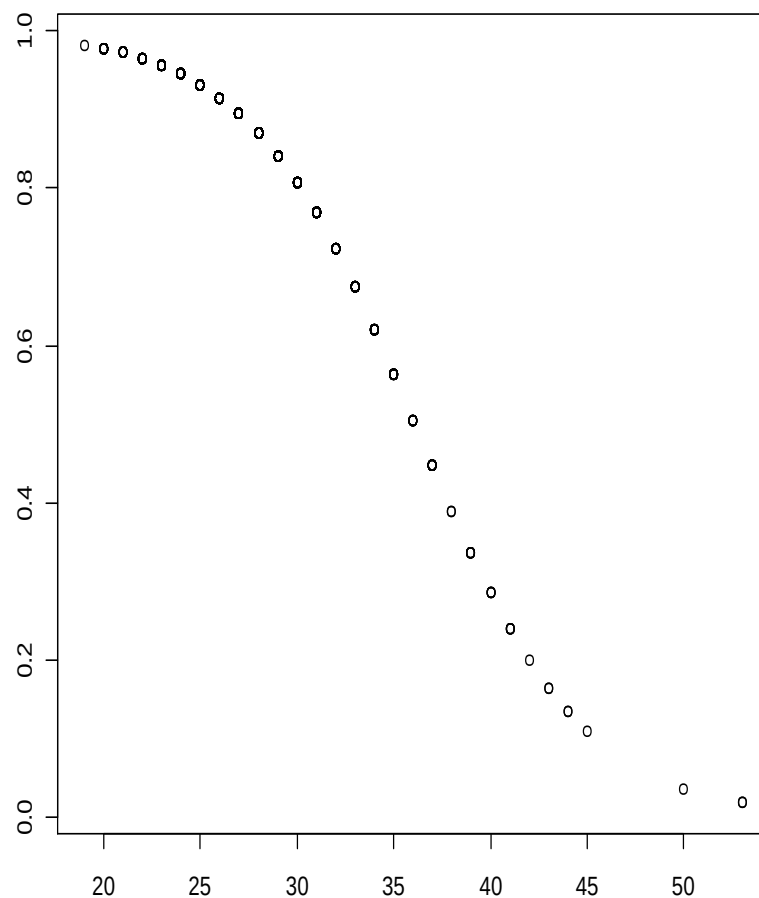
二値データの扱い

- 二値データの例
 - 疾患の有無
 - 基準を超えるか超えないか（カットオフ）
 - 一定期間内のイベントの発生(要介護認定, 心筋梗塞の発症など)
 - そのままでは2値しかとらないので扱いにくい
- 二値のうちどちらになるかを決める「確率」をモデルにあてはめる
- 「確率」が一定の基準を超えれば「有り」超えなければ「なし」が観測される(と考える)
- 確率のモデル化の仕方はいろいろある
 - ロジスティック回帰分析は確率と説明変数の関係をロジスティック関数でモデル化している
 - 他にも正規分布の累積分布関数でモデル化したプロビットモデルなどがある

0,1のまま



「確率」に置き換える



ロジスティック回帰分析(二項ロジットモデル)

イベントが起こる確率をロジスティック関数でモデル化

- ロジスティック関数は x が $-\infty$ から $+\infty$ の範囲で y が0から1の範囲をとるので確率のモデル化にちょうどいい

$$\Pr(Y = 1) = p = \frac{\exp(\beta_0 + \beta X)}{1 + \exp(\beta_0 + \beta X)}$$

$\Pr(Y = 1)$: Y が1になる確率, β_0 : 切片,
 β : 偏回帰係数, X : 説明変数

変形すると



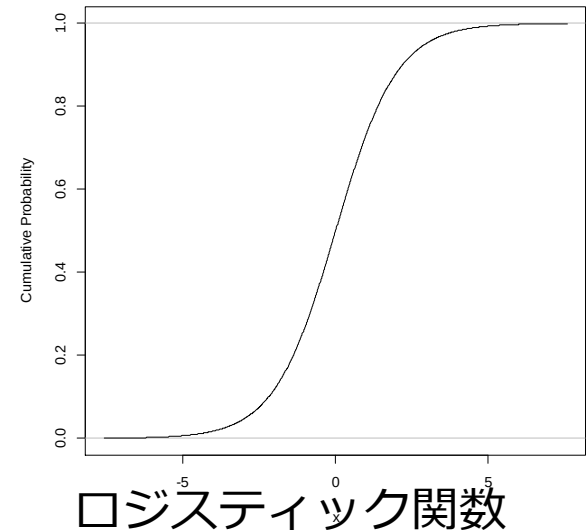
$$p + p \times \exp(\beta_0 + \beta X) = \exp(\beta_0 + \beta X)$$

$$p = (1 - p)\exp(\beta_0 + \beta X)$$

$$\frac{p}{1 - p} = \exp(\beta_0 + \beta X)$$

$$\ln\left(\frac{p}{1 - p}\right) = \beta_0 + \beta X$$

対数オッズを目的変数とした重回帰分析のように扱える



ロジスティック回帰分析の偏回帰係数，オッズ比

- ロジスティック回帰では，目的変数と説明変数の関連の強さはオッズ比を用いて解釈することが多い

- オッズ

- あるイベントが起こる確率と起こらない確率の比

$$\text{オッズ} = \frac{\text{イベントが発生した人の数(割合)}}{\text{イベントが発生しなかった人の数(割合)}}$$

- オッズ比の解釈

- 量的変数：説明変数の値が1増えるとオッズが何倍になるか
 - 質的変数：参照カテゴリに比べてオッズが何倍か

ロジスティック回帰分析の回帰係数からのオッズ比の導出

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta x, \ln\left(\frac{p_{x+1}}{1-p_{x+1}}\right) = \beta_0 + \beta(x+1)$$

$$\beta = \ln\left(\frac{p_{x+1}}{1-p_{x+1}}\right) - \ln\left(\frac{p_x}{1-p_x}\right)$$

$$\beta = \ln\left(\frac{p_{x+1}}{1-p_{x+1}} \div \frac{p_x}{1-p_x}\right)$$

$$\exp(\beta) = \frac{p_{x+1}}{1-p_{x+1}} \div \frac{p_x}{1-p_x}$$

ロジスティック回帰分析の結果の例

表3 男性における震災2年後の精神健康不良の関連要因

	オッズ比	95%信頼区間		有意確率
		下限	上限	
年齢	0.99	0.98	1.00	0.034
婚姻状況 (ref=既婚)				
未婚	1.09	0.75	1.58	0.648
離婚	1.07	0.59	1.92	0.829
死別	1.29	0.87	1.90	0.208
仕事の状況 (ref=仕事をしている)				0.002
求職中	0.99	0.50	1.95	0.979
仕事をしていない	1.57	1.22	2.02	0.000
暮らし向き (ref=普通)				
苦しい	2.44	1.96	3.03	0.000
疾患の有無 (ref=疾患なし)				
疾患あり	1.39	1.07	1.80	0.014
2011年度K6得点 (ref=4点以下)				
5点以上	6.20	4.99	7.70	0.000
同居者の死亡・行方不明 (ref=なし)				
あり	1.04	0.74	1.46	0.812
家屋被害 (ref=被害なし)				
被害あり	1.15	0.91	1.44	0.237
失業 (ref=失業なし)				
あり	1.10	0.84	1.43	0.501

注 ref: 参照カテゴリ

- 東日本大震災による家屋被害, 同居者の死亡・行方不明, 失業の状況と被災2年後の被災者の精神健康の状態の関連性を分析
- 男性では震災による家屋被害や同居者の死亡・行方不明, 失業は震災2年後の精神健康とは関連が認められなかったが, 現在の社会経済的状況との関連が認められた

分析結果のチェック

- 多重共線性 (multicollinearity) のチェック
 - 変数間の相関やVIFなどを確認する
 - SPSSではロジスティック回帰のオプションに多重共線性の診断指標を出力する機能はないが、線形回帰のオプションで出力すればよい
- オッズ比(または対数オッズ比)の信頼区間のチェック
 - 信頼区間が広すぎないか
 - 信頼区間の上限が大きすぎないか
 - 大きすぎたり広すぎる場合は質的変数の参照カテゴリのサンプル数が少なかったり、イベント数が少ない場合が多い
 - 参照カテゴリをサンプル数が多いカテゴリに変更したり、カテゴリを併合したりすると改善する場合もある

ロジスティック回帰分析の適合度指標

- 疑似決定係数(pseudo R^2)
 - 一般線形モデルの R^2 と同じようなもの。
 - Nagelkerkeの R^2 や, Cox-Snellの R^2 , McFaddenの R^2 がある
 - NagelkerkeはCox-Snellを0から1に収まるように補正したもの
 - **絶対的な基準はない**ので, モデルを比較するときなどに使う
- 情報量基準(Information Criterion)
 - AIC(Akaike Information Criterion), BIC(Bayesian Information Criterion)など
 - モデルの相対比較に使う
 - 値が低いほど少ない変数で精度よく予測できていることを表す
- AUC(Area Under Curve)/C統計量(C statistics)
 - 観測値, モデルによって予測された確率を元にROC曲線を書いた時の曲線の下での面積
 - 予測能の指標. 0.5~1の間を取り, 0.5~0.7は不良, 0.7~0.9は適度, 0.9以上は良好な当てはまりであることを表す
- 適合度の検定
 - Hosmer-Lemeshowの検定(Hosmer-Lemeshow test)
 - 帰無仮説は「モデルによる予測が観測と一致している」
→有意でなければ適合していると考え

説明変数の選択

- 研究の仮説に従って選ぶのが原則
 - 先行研究のレビューが重要
 - 因果推論を行う際には非巡回有向グラフ(DAG; Directed Acyclic Graph)による視覚化と整理が変数選択に有効
 - 回帰係数が有意ではない変数でも仮説モデルにあったら残す
 - 仮説が妥当なものであれば有意でない(関連がない)ことも重要な知見となる
- ステップワイズ法(変数増加法, 変数減少法, 変数増減法, 変数減増法)は使わない
 - ステップワイズ法はP値を基準に変数を選択する方法
 - ステップワイズ法の欠点
 - 有意な変数が残って有意でない変数は残らない
 - 自分が注目していた変数や本当は関連がある変数が残らないこともある
 - 本当は関係ない変数なのに残ることも(α エラー)
 - オーバーフィッティングが起こる
 - 解析データへの当てはまりはいいが, 別のデータには当てはまらないモデルになってしまう

Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, 5(1), 32. <https://doi.org/10.1186/s40537-018-0143-6>

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour?: Stepwise modelling in ecology and behaviour. *Journal of Animal Ecology*, 75(5), 1182–1189. <https://doi.org/10.1111/j.1365-2656.2006.01141.x>

サンプルサイズを決める

- EPV(Event per variable)を元に決める
 - EPVは説明変数1つあたりのイベント数
 - EPVが10以上になるようにデータを集める(Peduzzi, et. al., 1996)
 - 例: 説明変数を10個使いたければ $10 \times 10 = 100$ ケースのイベントが必要となる. イベント発生確率が20%とすると, 全体で $100 \div 0.2 = 500$ ケース集める必要がある.
 - EPV10以上というのは目安
 - EPV以外の条件次第で, 10より小さくてもよいとするもの (Vittinghoff & McCulloch, 2007), 10では不十分とするもの (Austin & Steyerberg, 2017)もある
 - 予測モデルの構築の場合はEPVだけでは不十分とするものもある (van Smeden et. al., 2019; Riley et. al., 2019)
 - 探索的な研究の場合はこの方法をベースに決めればよい
- 検出力を元に決める
 - 有意にしたい回帰係数の大きさなどから決める
 - 臨床試験等の検証的な研究の場合はこちらの方法で決める
 - パワーアナリシス用のソフトで計算できる
 - G*Power: <http://www.pscho.uni-duesseldorf.de/abteilungen/aap/gpower3/>

ロジスティック回帰分析の拡張

多項・順序ロジスティック回帰分析

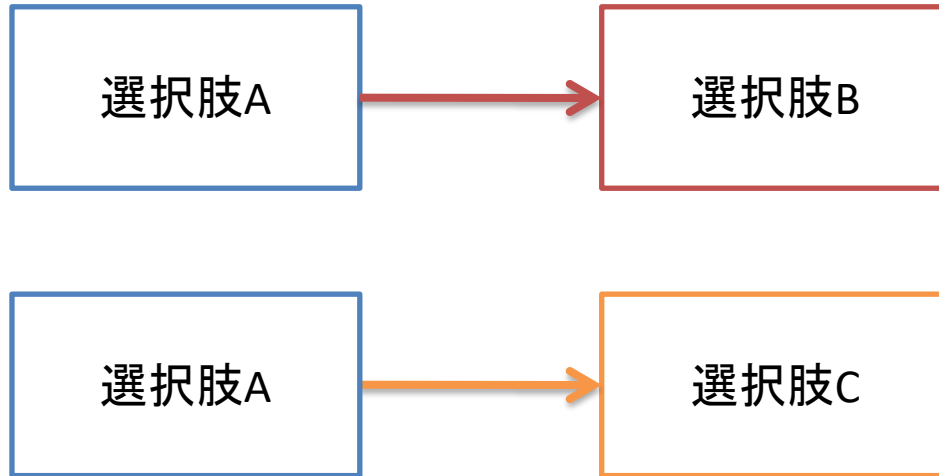
- 二項ロジスティック回帰分析の拡張
 - 3値以上の選択肢のデータも扱うことができる
- 多項ロジスティック回帰分析(多項ロジットモデル)
 - 目的変数の選択肢(カテゴリ)に順序をつけられない変数(名義変数)
 - 職種(技術 vs 事務 vs 営業)
 - インターネットの回線(光ファイバー vs ADSL vs CATV)
- 順序ロジスティック回帰分析(順序ロジットモデル)
 - 目的変数の選択肢に順序を付けられる変数(順序変数)
 - 健康状態の自己評価
 - 年収をカテゴリーにしたもの
 - 要介護度
 - カテゴリが多い場合は量的変数とみなして重回帰分析をすることもあ

考え方

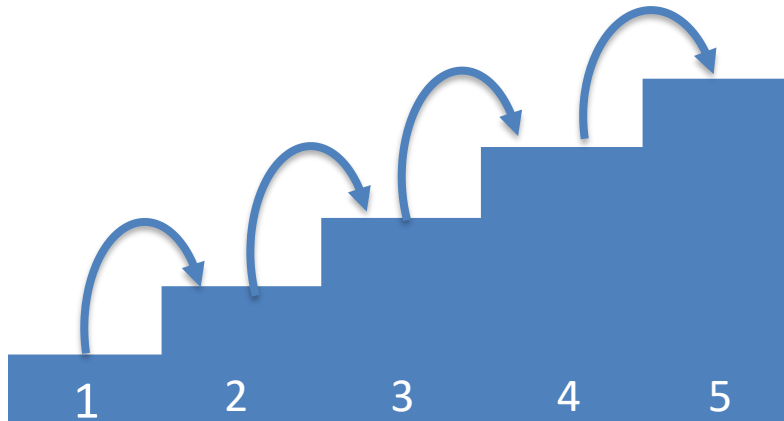
- 基本的には2項ロジスティック回帰分析と同じ
 - どの選択肢を選ぶかを決める潜在的な「確率」を使ってモデル化する
- 多項ロジスティック回帰分析(多項ロジットモデル)
 - (目的変数の)参照カテゴリと注目するカテゴリの選択を決める「確率」を比較
 - 説明変数の値によって、参照カテゴリと注目するカテゴリのどちらが選ばれやすいかがわかる
 - その「確率」に説明変数を与える影響をみる
 - 目的変数の参照カテゴリの選び方で比べるものが変わる
 - 説明変数の回帰係数(オッズ比)は説明変数の値が1大きくなると、参照カテゴリと比べて注目するカテゴリがどの程度選ばれやすくなるかを表す
- 順序ロジスティック回帰分析(順序ロジットモデル)
 - 選択肢全体に共通する特性に説明変数を与える影響をみる
 - 説明変数の回帰係数(オッズ比)は説明変数の値が1大きくなると、どの程度、順序が上のカテゴリが選ばれやすくなるかを表す

多項ロジットと順序ロジットのイメージ

- 多項ロジット:ペアごとに比較



- 順序ロジット:カテゴリー群全体で比較



多項ロジスティック回帰分析の例

Table 3. Summary of the associations of psychological distress with participation status among survivors

Variable	Psychological distress (K6)			<i>P</i> for interaction
	None (0–4)	Mild (5–12)	Severe (13–24)	
Male (<i>n</i> = 3,031)				
Number of participants	2,352	610	69	
Nonresponders	468	195	27	
Age-adjusted	Reference	1.85 (1.51–2.26)	2.41 (1.45–4.00)	
Multivariate-adjusted^a	Reference	1.69 (1.26–2.26)	1.66 (0.76–3.62)	
<i>P</i> for trend		0.0019		
By age				0.608
<65 years (<i>n</i> = 998)	Reference	1.31 (0.81–2.10)	1.20 (0.49–2.94)	
≥65 years (<i>n</i> = 2,033)	Reference	1.91 (1.31–2.79)	1.99 (0.58–6.88)	
Female (<i>n</i> = 4,989)				
Number of participants	2,864	1,009	120	
Nonresponders	623	311	62	
Age-adjusted	Reference	1.43 (1.23–1.67)	2.32 (1.68–3.22)	
Multivariate-adjusted^a	Reference	1.46 (1.17–1.84)	2.30 (1.45–3.64)	
<i>P</i> for trend		<0.0001		
By age				0.051
<65 years (<i>n</i> = 2,026)	Reference	1.42 (1.03–1.94)	1.88 (0.92–3.81)	
≥65 years (<i>n</i> = 2,963)	Reference	1.53 (1.10–2.12)	2.68 (1.34–5.34)	

*Adjusted for age plus variables with age-adjusted $P \leq 0.20$, and the independence of the variables relative to each of the other variables in the prediction psychological distress risk was tested.

- 東日本大震災の被災地域住民を対象とした調査で，精神健康度を「問題なし」「軽度の異常」「重度」の異常の3段階に分類し，健診受診の状況との関連をみたもの
- 男女別に年齢調整のみ行った結果，多変量調整を行った結果が示されており，健診受診状況以外の変数と精神健康度の関連性は表では省略されている

順序ロジスティック回帰分析の例

表3 健康相談の必要性の認識^{*1}と対象の特徴（基本属性、生活状況、医療介護状況）との関連：順序ロジスティック回帰分析^{*2}の結果

変数	オッズ比（95%信頼区間）	P 値
〈基本属性〉		
年収（200万円未満）	1.01(0.71-1.43)	0.97
〈生活状況〉		
日常生活動作（ADL）（自立していない）	1.46(1.03-2.07)	0.03*
健康感（とても健康である）	0.71(0.38-1.30)	0.26
生活満足度（とても満足している）	1.07(0.67-1.70)	0.78
家族関係にストレス（あり）	1.80(1.23-2.63)	0.003**
自らの健康にストレス（あり）	1.13(0.78-1.63)	0.53
〈医療介護状況〉		
がん医療通院中	1.98(0.99-3.76)	0.05
気軽に相談できる医療者として看護師がいる	2.26(1.01-5.12)	0.04*
健康について相談したい医療者（必ず医師がいい）	0.53(0.38-0.74)	<.001***
医師に質問・疑問を相談できるか（遠慮せずにできる）	0.61(0.43-0.88)	0.007**
嫌な医療経験（あり）	1.16(0.82-1.63)	0.40
地元の健康相談支援：暮らしの保健室（知っている）	1.46(1.06-2.00)	0.02*

^{*1}：5段階評価（5：とても必要である，4：まあ必要である，3：どちらともいえない，2：あまり必要でない，1：全く必要ない）にて測定した値（範囲 1-5）

^{*2}：単変量解析（ウィルコクソンの順位検定もしくはスピアマンの順位相関係数）にて，健康相談の必要性の認識の程度と $P<.10$ の関連の認められた因子を投入

* $P<0.05$, ** $P<0.01$, *** $P<0.001$

- 健康相談の必要性の認識（「5とても必要」～「1全く必要ない」）と対象者の特徴との関連を分析したもの
- オッズ比が1を超えていれば必要性の認識が強くなり，1未満であれば必要性の認識が弱くなる方向に関連している

引用&参考文献

- 生態学のデータ解析
 - <http://hosho.ees.hokudai.ac.jp/~kubo/ce/FrontPage.html>
- Alan Grafen, Rosie Hails. Modern Statistics for the Life Sciences. Oxford University Press.
- Michael J. Crawley. Statistics: An Introduction using R. John Wiley & Sons, Ltd.
- Julian J Faraway. Extending the Linear Model with R . Taylor & Hall/CRC
- UCLA Academic Technology Services
 - http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm

参考文献

- Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, 5(1), 32. <https://doi.org/10.1186/s40537-018-0143-6>
- Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour?: Stepwise modelling in ecology and behaviour. *Journal of Animal Ecology*, 75(5), 1182–1189. <https://doi.org/10.1111/j.1365-2656.2006.01141.x>
- Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373-1379. doi:10.1016/s0895-4356(96)00236-3
- Vittinghoff E, McCulloch CE. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*. 2007;165(6):710-718. doi:10.1093/aje/kwk052
- Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res*. 2017;26(2):796-808. doi:10.1177/0962280214558972
- van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat Methods Med Res*. 2019;28(8):2455-2474. doi:10.1177/0962280218784726
- Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes [published correction appears in *Stat Med*. 2019 Dec 30;38(30):5672]. *Stat Med*. 2019;38(7):1276-1296. doi:10.1002/sim.7992