2019/1/16 北大内科 I

観察研究データの解析
～P値からの脱却～

北海道大学　医学統計学
横田　勲

本日の内容

- 分割表に対する検定
  - Fisherの直接確率検定とカイ二乗検定
  - 仮説検定に対する誤解
- 効果の大きさ
  - リスク差、リスク比、オッズ比
  - 信頼区間の推定
- 交絡とロジスティック回帰分析
- （付録）統計解析ソフト: JMPの使用

観察研究データ解析のポイント

- 検定は行わない
  - 比較可能性が乏しい観察研究では、
    有意差があっても解釈困難
  - 症例数を多く集めれば、簡単に有意差がつく
- 効果の大きさを推定
  - 信頼区間を示せば、検定結果もわかる
- 交絡調整のひとつに回帰分析
  - 多変量解析がメインの解析

2つの治療法を比較する
1800名の観察研究

治療法	イベントあり	イベントなし	合計
試験治療	22 (2.2%)	978	1000
標準治療	20 (2.5%)	780	800
合計	42	1758	1800

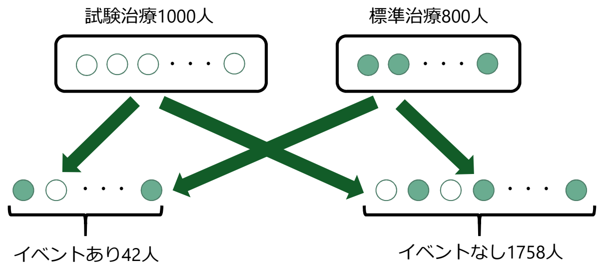
- 試験治療はイベント発生リスクを下げるか？
  - 試験治療での脳卒中リスクは2.2%
  - 標準治療での脳卒中リスクは2.5%
    - 偶然の違いか否か

検定

- 観察された差が、偶然によるものか、系統的な違いがあるかを判断
- 背理法の論理
  - 「群間に差がない」と仮定（帰無仮説）
  - 観察値はどの程度まれな現象であるか、を確率で表現
    - p値：観察値とそれよりまれな現象が起こる確率
  - p値が小さければ(例えば5%)、帰無仮説が誤っていたと判断し、有意差あり

Fisherの直接確率検定

- 試験治療1000人、標準治療800人と割付けたらイベントありが42人含まれていた
  - そのうち試験治療が22人である確率は？



Fisherの直接確率検定

- 試験治療1000人、標準治療800人からイベントあり42人が選ばれた
  - そのうち試験治療が0人である確率は？
  - そのうち試験治療が1人である確率は？

    ⋮
  - そのうち試験治療が22人である確率は？
  - そのうち試験治療が23人である確率は？

    ⋮
  - そのうち試験治療が42人である確率は？

この和の2倍が(両側)p値

カイ二乗検定

- 治療間で等しい確率で脳卒中が発生する
  - 試験治療を1000人、標準治療を800人サンプリングしたら、42人が脳卒中を起こしていた

	イベントあり	イベントなし	合計
試験治療	23.33	976.67	1000
標準治療	18.67	781.33	800
合計	42 (2.33%)	1758	1800

22人 と 20人

- 観察値とのズレを統計モデルを基に、確率で表現

Fisher検定とカイ二乗検定

9

- 幸い、どちらも同じようなp値
  - Fisher検定 p値=0.79>0.05
  - カイ二乗検定 p値=0.68>0.05
- データの得られ方を考えて
  ふさわしい方を採用すべき
  - 目の前の患者1800人をランダムに割付け
    - Fisherの直接確率検定
  - 大勢の患者から1800人ランダムサンプリング
    - カイ二乗検定

10

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

11

Eur J Epidemiol (2016) 31:337–350
DOI 10.1007/s10654-016-0149-3

P値って何だ？

12

以下のうち、正しい説明はどれでしょう？

1. P値は帰無仮説が正しい確率である
2. P=0.05の場合、
   対立仮説が正しい確率が95%である
3. P値は帰無仮説が正しい場合に
   観察された結果が得られる順位である
4. P値が小さいほど、より関連が強い
   （よりリスク因子の影響が強い）
5. P値が1に近ければ、差はない

P値が有意水準を下回った場合

13

- 帰無仮説が誤っていたことは言える
- 「治療によりアウトカムに差があった」と言っていい？
  - 比較可能性があるならば、言ってもよい
    - 治療以外のリスク因子の分布が群間で同じ
      - 年齢、重症度、喫煙歴、、、
      - 測定していない、測定できない因子もすべて
    - 群間差は治療だけから生じる内容、とお膳立て

ランダム化

14

- 様々なリスク因子の分布が完全に同一になることはないが、グループの人数が増えるほど平均的に同一に近づくはず
- 「測定可能な特徴」にもとづいていない
  - 「測定できない特徴」だって平均的に同一になっているに違いない！
  - 比較可能性を担保できる最強の方法

観察研究データであれば

15

- 比較可能性を担保することはできない
- “比較可能”な状態に少しでも近づける
  - 交絡バイアスをデザインや解析で制御
    - 治療にもアウトカムにも影響を与える交絡因子によって引き起こされることがある
    - 回帰分析による制御
- とはいえ、比較可能でない以上、検定を行う意味がない

症例数設計

16

- ランダム化研究では、期待通りの治療効果がある場合、80%や90%の確率で有意差がつくよう設計
- 観察研究（特にレトロ研究）では症例数設計はされないことがほとんど
  - 症例数をたくさん集めてしまえば、わずかな治療効果でも有意差がつく

どっちの分割表がP値は小さい？

17

治療法	治癒	不変	合計
A	12 (60%)	8	20
B	6 (30%)	14	20

治療法	治癒	不変	合計
A	6200 (31%)	13800	20000
B	6000 (30%)	14000	20000

P値のつかいどころ

18

- どちらの治療法（薬）を採用するかの意思決定
  - 比較可能になるようランダム化を行い、適切な症例数を集めるというお膳立てをして、有意差ありとなったら、新治療を採用するということを事前に定めておく場面
- 観察研究においてはP値の計算は無意味
  - Epidemiology（雑誌）では、P値の使用禁止
    - Statistical testingに集められている論文参照
      https://journals.lww.com/epidem/pages/collectiondetails.aspx?TopicalCollectionId=4

どの程度の効果か

19

治療法	イベントあり	イベントなし	合計
試験治療	22 (2.2%)	978	1000
標準治療	20 (2.5%)	780	800
合計	42	1758	1800

- 試験治療はイベント発生リスクをどの程度下げるか？
  - 試験治療での脳卒中リスクは2.2%
  - 標準治療での脳卒中リスクは2.5%
    - 適切な指標で比べてみよう

効果をはかる指標

20

- リスク差：$\frac{A}{A+B} - \frac{C}{C+D}$
- リスク比：$\frac{A}{A+B} / \frac{C}{C+D}$
  - 脳卒中があった割合の群間差、群間比
- オッズ比：$\frac{A}{B} / \frac{C}{D}$

治療法	イベントあり	イベントなし	合計
試験治療	A	B	A+B
標準治療	C	D	C+D
合計			N

効果をはかる指標

21

- リスク差：$\frac{22}{1000} - \frac{20}{800} = -0.003$
- リスク比：$\frac{22}{1000} / \frac{20}{800} = 0.88$
  - 脳卒中があった割合の群間差、群間比
- オッズ比：$\frac{22}{978} / \frac{20}{780} = 0.877 \approx 0.88$

治療法	イベントあり	イベントなし	合計
試験治療	22 (2.2%)	978	1000
標準治療	20 (2.5%)	780	800
合計	42	1758	1800

リスク差/比とオッズ比の違い

22

- リスク差、リスク比
  - 疾病発生割合を比べたものとして解釈
- オッズ比
  - 直接的な解釈は難しい
  - リスク比を近似したものといえる場合がある
  - もともとは「ケースコントロール研究」にて開発された

全員追跡してもオッズ比？

23

- コホート研究ではオッズ比をあえて出す意味はあまりない
  - ロジスティック回帰で推定される
    - 数学的に近似がうまくいきやすい
- 解析ソフトが出してくれる・・
- 疾患分野ごとの慣習・・・・

効果指標の確からしさ

24

- 同じリスク比、オッズ比
  - リスク比：0.88

	イベントあり	イベントなし	合計
試験治療	2200	97800	100000
標準治療	2000	78000	80000
合計	4200	175800	180000

  - このデータから推定したリスク比のほうが確からしそう
    - それらしいリスク比の範囲を示す

1800人集めてリスク比は0.88

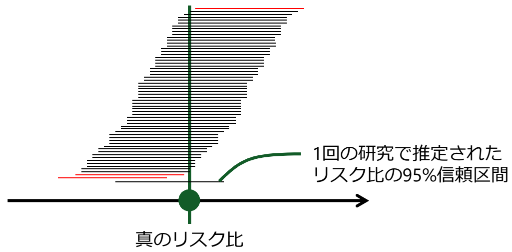
25

- 同じ研究をもう一度行ったら・・？
  - リスク比は0.80かもしれない
  - リスク比は1.00かもしれない
- 研究ごとに推定されるリスク比に、<u>適当な</u>幅をつけてみよう
  - リスク比の95%信頼区間：0.48-1.60

95%信頼区間(Confidence Interval)

26

- 研究ごとに95%信頼区間を計算すれば、100回繰り返したうち、95回は“真の”リスク比を含む



リスク比：0.88 (95%CI 0.48-1.60)

27

- 真のリスク比は0.88だ！
  - そうかもしれない
- 真のリスク比は0.50だ！
  - そうかもしれない
- 真のリスク比は2.00だ！
  - それは違うのでは

信頼区間と検定

28

- 検定で有意とならない値の範囲が信頼区間
- 「P<0.05である」
  ⇔「リスク比の95%CIが1をまたがない」

	あり	なし
試験治療	22	978
標準治療	20	780

95%CI ：0.48-1.60
p=0.68

	あり	なし
試験治療	2200	97800
標準治療	2000	78000

95%CI ：0.83-0.93
p<0.01

2つの治療法を比較する 1800名の観察研究

29

治療法	イベントあり	イベントなし	合計
試験治療	22 (2.2%)	978	1000
標準治療	20 (2.5%)	780	800
合計	42	1758	1800

- 試験治療により、イベントはやや減った
  - リスク比：0.88 (95%CI: 0.48-1.60)
  - オッズ比：0.88 (95%CI: 0.48-1.62)
- 治療効果はもう少し大きいはず・・・？

年齢でサブグループ化

30

- 75歳以上

治療法	イベントあり	イベントなし	合計
試験治療	18 (3.0%)	582	600
標準治療	6 (6.0%)	94	100

- 75歳未満

治療法	イベントあり	イベントなし	合計
試験治療	4 (1.0%)	396	400
標準治療	14 (2.0%)	686	700

- どちらのサブグループでもリスク比0.5

交絡 (confounding)

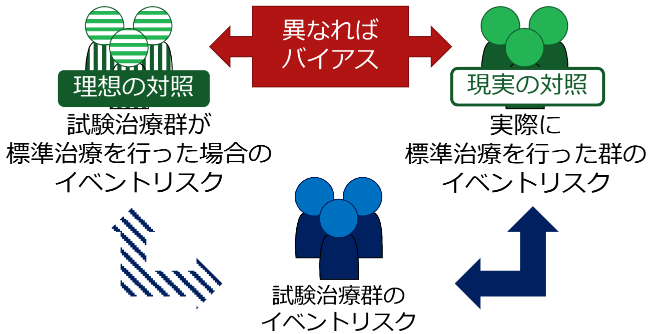
31

- 全体での単純な解析結果と
  サブグループ解析結果が異なる現象
  - 全体でのリスク比：0.88
  - サブグループでのリスク比
    - 75歳以上：0.50
    - 75歳未満：0.50
  - 全体でのリスク比は、
    サブグループ解析の間に入るはずなのに...

治療効果の適切な比較

32

- 試験治療の治療効果を知りたい
  - 比較対照は標準治療



原因を考えてみる

33

- 標準治療(対照)でのイベントリスクが高齢なほど高い
  - 75歳以上： 6 / 100 = 6.0%
  - 75歳未満： 14 / 700 = 2.0%
- 高齢なほど試験治療を受けがち
  - 試験治療での75歳以上：600/1000=60.0%
  - 標準治療での75歳以上：100/800=12.5%

交絡要因の定性的な必要条件

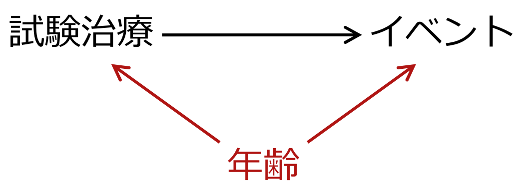
34

1. アウトカムに因果的に影響
   - 高齢なほどイベントリスクは上昇
2. 比較群間で分布が異なる
   - 試験治療群では高齢患者が多い
3. 治療ーアウトカム間の中間変数でない（治療の影響を受けない）
   - 試験治療であれば高齢になりがち でない
   - 高齢であるから試験治療になりがち

年齢による交絡

35

- 試験治療の脳卒中への因果治療効果を知りたい
  - 両者に矢線が入る因子（年齢）を無視すると、歪んだ結果が得られてしまう



交絡の制御や調整

36

- 研究デザインでの制御
  - ランダム化
  - 集団の限定
  - マッチング
- 解析での調整
  （測定できた交絡因子に対して）
  - 層別解析
  - 回帰モデルのあてはめ
    - 例）ロジスティック回帰分析

オッズ比の計算

37

治療法	イベントあり	イベントなし	合計
試験治療	22 (2.2%)	978	1000
標準治療	20 (2.5%)	780	800
合計	42	1758	1800

- $\dfrac{\text{イベントあり}_{in}\text{試験治療} / \text{イベントなし}_{in}\text{試験治療}}{\text{イベントあり}_{in}\text{標準治療} / \text{イベントなし}_{in}\text{標準治療}}$

$$= \frac{22/978}{20/780} \approx 0.88$$

ロジスティック回帰モデル

38

$$\log \frac{(\text{イベントあり})}{(\text{イベントなし})} = \beta_0 + \beta_1 \boxed{\times (\text{試験治療})}$$

オッズ

試験治療であれば1,
標準治療なら0

- 試験治療での対数オッズ：$\beta_0 + \beta_1$
- 標準治療での対数オッズ：β_0

$$\boxed{\text{オッズ比：} \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)}$$

回帰モデルによる交絡調整

39

$\log(\text{オッズ}) = \beta_0 + \beta_1 \times (\text{試験治療}) + \beta_2 \times (75\text{歳以上})$

- 真の治療効果は年齢によらず同じ と仮定

オッズ	75歳未満	75歳以上
試験治療	$\exp(\beta_0 + \beta_1)$	$\exp(\beta_0 + \beta_1 + \beta_2)$
標準治療	$\exp(\beta_0)$	$\exp(\beta_0 + \beta_2)$

75歳未満でのオッズ比

$$\frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \mathbf{exp}(\boldsymbol{\beta_1})$$

75歳以上でのオッズ比

$$\frac{\exp(\beta_0 + \beta_1 + \beta_2)}{\exp(\beta_0 + \beta_2)} = \mathbf{exp}(\boldsymbol{\beta_1})$$

ロジスティック回帰分析結果

40

因子	オッズ比	(95%CI)
試験治療 / 標準治療	0.49	(0.28-1.01)
年齢 (75歳以上 / 未満)	3.10	(1.50-6.49)

- 年齢を調整したところ、
  標準治療に対する試験治療の
  イベント発生は減少する傾向にあった
  （オッズ比:0.49, 95%CI: 0.28-1.01）

単変量解析と多変量解析

41

- 説明変数が1つか2つ以上上か
  - 説明変数が1つであれば、分割表で表現可能
- 観察研究データの解析では
  多変量解析が基本の解析内容
  - 交絡があると考えて、可能な限り調整した
    多変量解析結果をもって議論すべき

因子	オッズ比	(95%CI)
治療	0.88	(0.48-1.62)

単変量解析

因子	オッズ比	(95%CI)
治療	0.49	(0.28-1.01)
年齢	3.10	(1.50-6.49)

多変量解析

結果変数の型による分類

42

目的	分類尺度	連続尺度	時間イベント尺度
分布の記述	ヒストグラム、分割表	ヒストグラム、箱ヒゲ図、散布図	生存曲線 （Kaplan-Meier法）
要約統計量	頻度、一致度、相関係数	平均、分散、中央値、パーセント点、相関係数	x年生存確率、中央生存期間
検定（単純）	χ^2検定、Fisher正確検定	t検定、Welch検定、分散分析、Wilcoxon検定	ログランク検定
検定（層別）	Mantel-Haenszel法	共分散分析	層別ログランク検定
回帰モデル	ロジスティック回帰分析	重回帰分析	Cox回帰分析

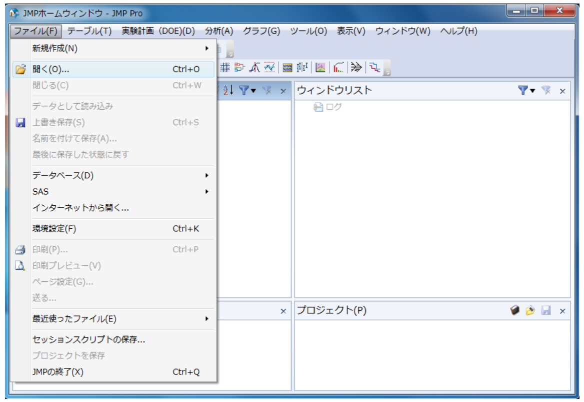
（付録） 統計解析ソフト: JMPの使用

43

- JMP: SAS社が開発
  - マウスで操作して統計解析
  - SASは世界で最も信頼されている解析ソフト
  - JMPの解析結果であれば、どの雑誌もOK(なはず)

CSVファイルを読み込む

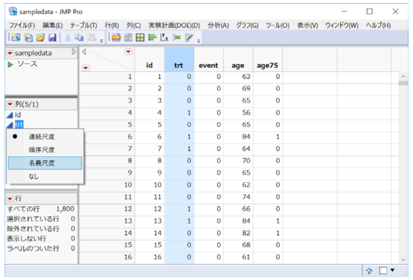
44



正しい尺度に変更

45

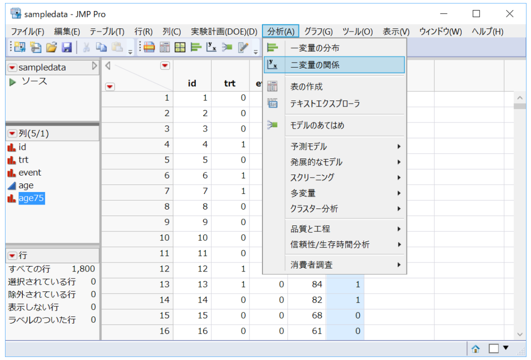
- 連続尺度：年齢、BMI、血圧
- 順序尺度：改善/不変/悪化のような順序性あり
- 名義尺度：あり/なし、術式A/B/C



分割表をかく①

46

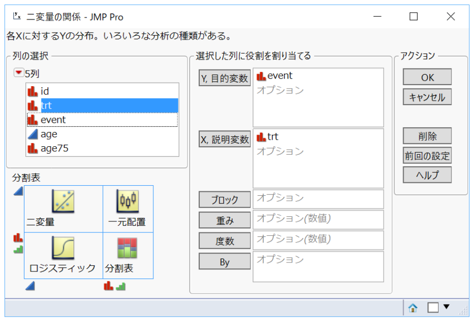
- 分析 > 二変量の関係 を選択



分割表をかく②

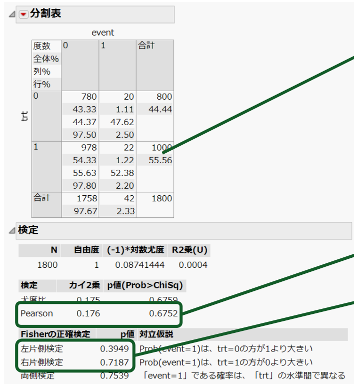
47

- 結果(stroke)を「Y, 目的変数」に
  治療(offp)を「X, 説明変数」に



分割表と検定結果

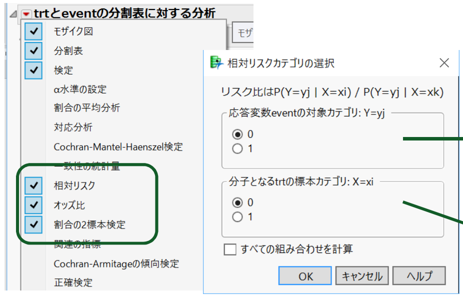
48



- 度数
  列での%
  行での%
  全体での%
  が表示
- カイ二乗p値
- Fisher検定p値
  - 左 or 右片側検定の小さいほうを2倍

リスク差、リスク比、オッズ比

49



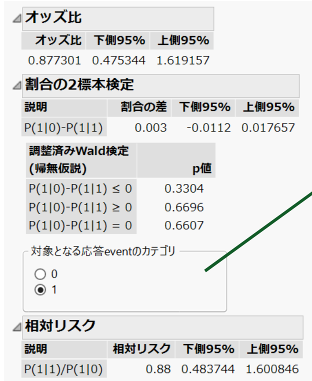
「イベントあり」
の入力内容

分子におく
治療群

- 相対リスク：リスク比
- 割合の2標本検定：リスク差

リスク差、リスク比、オッズ比

50

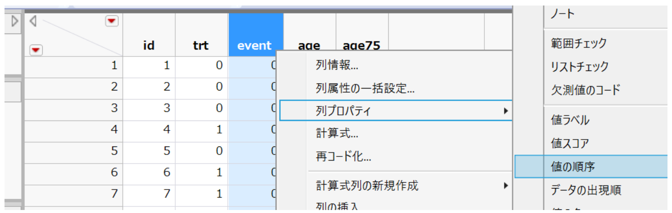


- オッズ比
- リスク差
  - 「イベントあり」
    の入力内容を選ぶ
- リスク比

データの準備

51

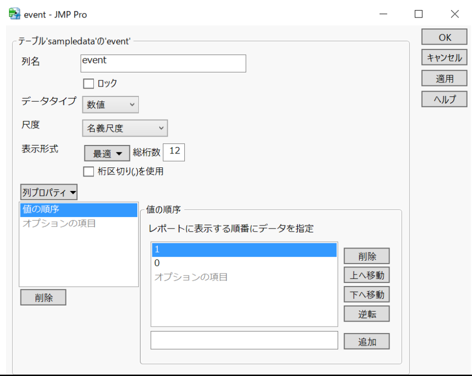
- Stroke を右クリック
  - 列プロパティ ＞ 値の順序 をクリック



イベントあり を一番上にする

52

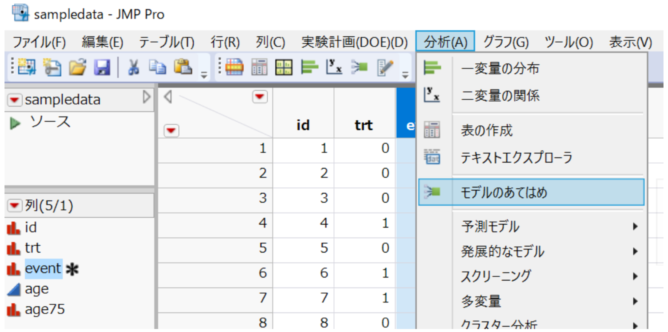
- 値の順序 にて 上へ移動 をクリック



ロジスティック回帰の実行

53

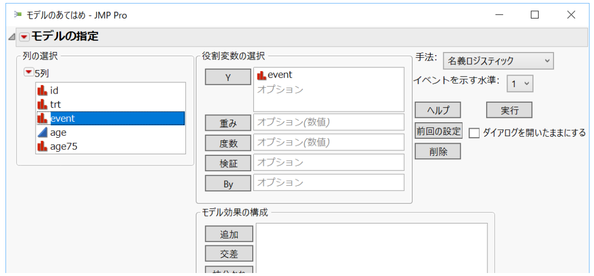
- 分析 > モデルのあてはめ　を選択



アウトカム (stroke) の選択

54

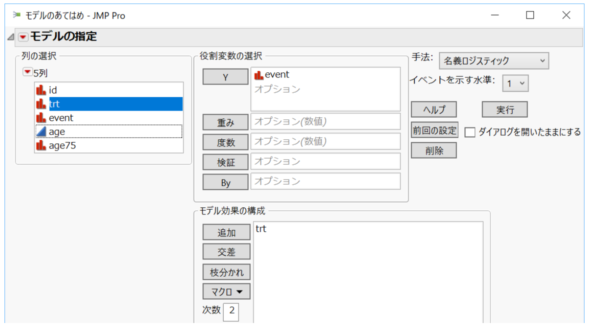
- 列の選択 にある(アウトカム)をクリック
- 役割変数の選択 にある Y をクリック
  - 手法が 名義ロジスティック になったことを確認



オッズ比を調べたい因子を投入

55

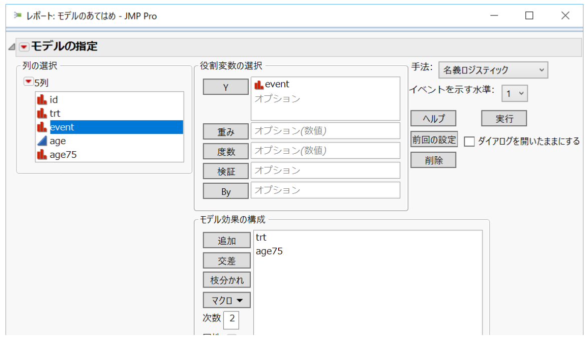
- 列の選択 にある (治療法) を選択
- モデル効果の構成 にある 追加をクリック



調整因子(交絡因子)を投入

56

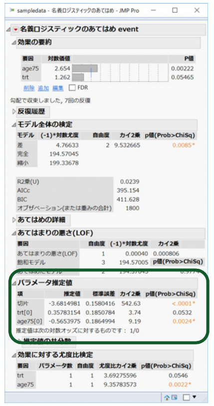
- 前スライドと同じ
  - 興味のある因子も調整因子もモデル上は同じ扱い



解析結果

57

- パラメータ推定値 に表示
  - ？？？？？



パラメータ推定値

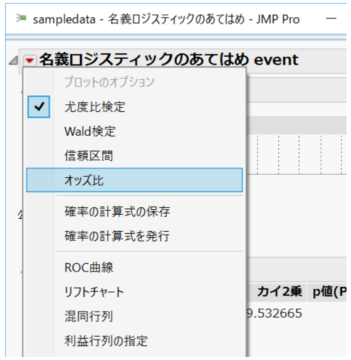
項	推定値	標準誤差	カイ2乗	p値(Prob>ChiSq)
切片	-3.6814981	0.1580416	542.63	<.0001*
trt[0]	0.35783154	0.1850784	3.74	0.0532
age75[0]	-0.5653975	0.1864994	9.19	0.0024*

推定値は次の対数オッズに対するものです： 1/0

オッズ比を表示する

58

- 名義ロジスティックのあてはめ 脇にある ▼をクリック
- オッズ比を クリック



オッズ比 0.49 (95%CI: 0.24-1.01)

59

- 治療群のどちらが分子・分母かに注意

オッズ比

event: 1対0のオッズ比に対して

trtのオッズ比

水準1	/水準2	オッズ比	p値(Prob>ChiSq)	下側95%	上側95%
1	0	0.4888678	0.0532	0.2366538	1.0098793
0	1	2.0455426	0.0532	0.9902173	4.2255819

age75のオッズ比

水準1	/水準2	オッズ比	p値(Prob>ChiSq)	下側95%	上側95%
1	0	3.0981183	0.0024*	1.4914228	6.4356914
0	1	0.3227766	0.0024*	0.1553835	0.6705007

次の信頼限界にはWald近似が使われています: trt age75
オッズ比の検定と信頼区間は、Wald法に基づいて計算されています。

治療群のオッズ比

年齢のオッズ比