統計学第 15 講

後藤　晶　　akiragoto@meiji.ac.jp

目次

前回の復習 1

2 要因分散分析モデル (交互作用なし) 6

目次

前回の復習

2 要因分散分析 (交互作用あり)

- 続いて，2 要因分散分析に進みたいと思います．2 要因分散分析とは，複数の要因による影響を分析するものです．例えば，主観的幸福度は子の有無 (1,0 のダミー変数) だけでなく，結婚しているか否か (1, 0 のダミー変数) によっても影響を受ける可能性があります．これを用いると「子がいない未婚者」「子がいない既婚者」「子がいる未婚者」「子がいる既婚者」の計 4 つの状態があります．
- したがって，これらが影響を与えているかどうかを明らかにするために，いずれの要因についても投入したモデル式について考えたいと思います．ここでは，次のようなモデル式を考えたいと思います．

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2 + \alpha + \epsilon_i$$

このモデル式によって，「4 つの状態」を分析することができます．一度整理してみましょう．

- $X_1 = 1$ と $X_2 = 0$ の場合

$$Y_i = \beta_1 * 1 + \beta_2 * 0 + \beta_3 * 1 * 0 + \alpha + \epsilon_i$$

$$Y_i = \beta_1 + \alpha + \epsilon_i$$

  - この場合，ある因子 X_1 によって，傾きが変化することを示しています．
  - ex. 子がいない独身者よりも，子がいる独身者の方が幸せとか

- $X_1 = 0$ と $X_2 = 1$ の場合

$$Y_i = \beta_1 * 0 + \beta_2 * 1 + \beta_3 * 0 * 1 + \alpha + \epsilon_i$$
$$Y_i = \beta_2 + \alpha + \epsilon_i$$

  - この場合，ある因子 X_2 によって，傾きが変化することを示しています．
  - ex. 子がいない未婚者よりも，子がいない既婚者の方が幸せとか

- $X_1 = 1 X_2 = 1$ の場合

$$Y_i = \beta_1 * 1 + \beta_2 * 1 + \beta_3 * 1 * 1 + \alpha + \epsilon_i$$
$$Y_i = \beta_1 + \beta_2 + \beta_3 + \alpha + \epsilon_i$$

  - この場合，X_1 と X_2 が影響する場合の値を示していることになります．特に，$X_1 * X_2$ の係数が有意になる場合は単純に X_1 と X_2 が同じように影響を与えているだけでなく，組み合わさることによって効果が強まることを示しています．
  -「組み合わさることにより効果が変化する」ことを「交互作用」といいます．
  - ex. 子がいない未婚者よりも，子がいる既婚者の方が幸せ
  - 子どもがいることによる幸福度の改善と，結婚していることによる幸福度の改善から予想できないくらいグッと幸せ．

- $X_1 = 0 X_2 = 0$ の場合

$$Y_i = \beta_1 * 0 + \beta_2 * 0 + \beta_3 * 0 * 0 + \alpha + \epsilon_i$$
$$Y_i = \alpha + \epsilon_i$$

- この場合，全ての要因が影響しない場合（何らかの基準となる点）の値を示していることになります．
  - ex. 子がいない未婚者の幸福度の推定値

仮説を立てる

さて，それでは仮説を立ててみましょう．今回分析するテーマは「主観的幸福度 (SUB_HAP) が子の有無 (CHI) と結婚 (MAR) によって異なる」かどうかを分析します．二要因分散分析（交互作用有り）の場合は以下のような仮説を立てます．

* 対立仮説：主観的幸福度の平均値は結婚かつ子の有無によって異なる．
* 帰無仮説：主観的幸福度の平均値は結婚かつ子の有無によって異なるとは言えない．

この 6 つの仮説のもとに分析を行ないます．

平均値をプロットする

さて，最初のお約束です．平均値をプロットしましょう．まずは各自でやってみましょう．

さて，例によって ggplotgui を使いましょう．

以下のコードは Console（コンソール）に直接打ち込みます．

```
library(ggplotgui)
ggplot_shiny(exdataset)
```

そうすると新しいウィンドウが開きます．

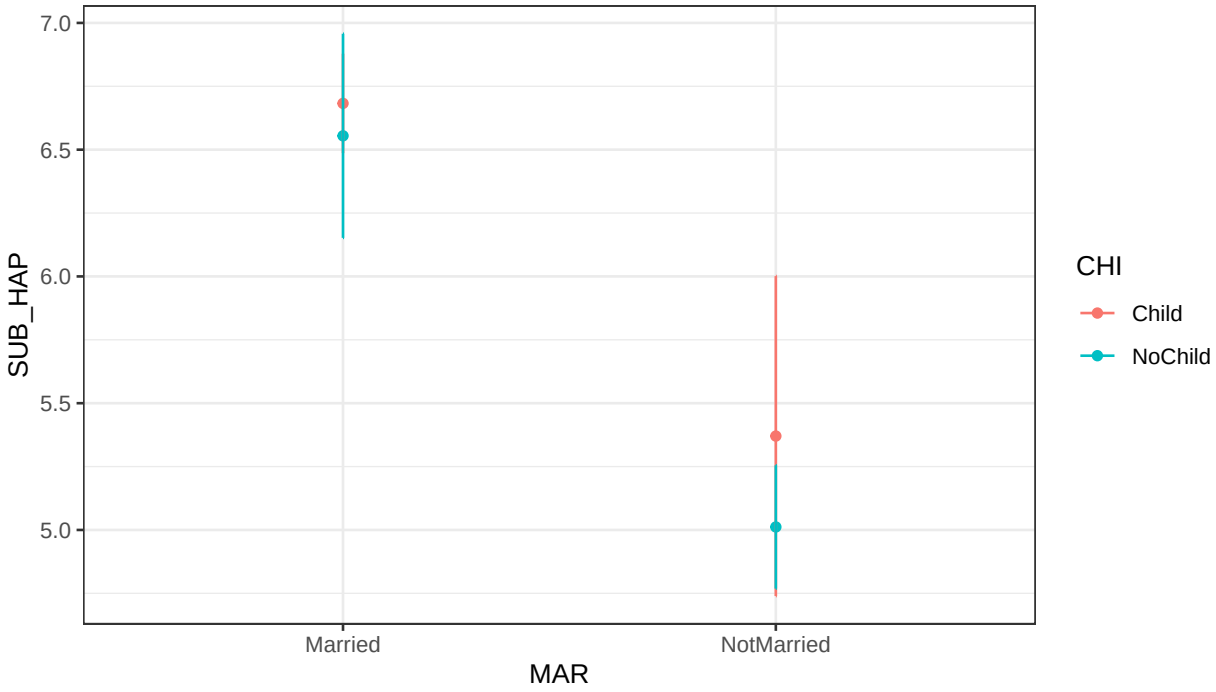
以下の通りの作業をしましょう．

- ggplot タブへ
- “Type of graph:” は “Dot + Error”，Y-variable は “SUB_HAP”，X-variable は “MAR” を設定
- “Group(or colour)” を CHI に変更
- “Confidence Interval:” を 95% にする．
- R-code タブへ行って，以下のコードのうち，真ん中のみを以下にする．-また，コード内の df を exdataset に変える．

```
# You need the following package(s):
library("ggplot2")

# The code below will generate the graph:
graph <- ggplot(exdataset, aes(x = MAR, y = SUB_HAP, colour = CHI)) +
  geom_point(stat = 'summary', fun.y = 'mean') +
  geom_errorbar(stat = 'summary', fun.data = 'mean_se',
                width=0, fun.args = list(mult = 1.96)) +
  theme_bw()
```

```
graph
```



このグラフを見る限り，未婚者に比べて既婚者の方が主観的幸福度が高そうですが，子の有無の影響はありそうな気もしますし，なさそうな気もしますし何とも言えません．したがって，この点についても統計的に差があるのかどうかを明らかにしましょう．

2 要因分散分析（交互作用あり）のモデル式

```
marchihap_model <-lm(SUB_HAP ~ MAR*CHI, data = exdataset)
# モデル式を MARCHIHAP_model というオブジェクトにしまいます.
```

```
# 分析結果の要約を出力する
summary(marchihap_model)
```

```
##
## Call:
## lm(formula = SUB_HAP ~ MAR * CHI, data = exdataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6825 -1.6825  0.3175  1.3175  4.9882
##
```

```
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  6.6825     0.1054  63.419  < 2e-16 ***
## MARNotMarried               -1.3122     0.3190  -4.113 4.24e-05 ***
## CHINoChild                  -0.1279     0.2222  -0.575    0.565
## MARNotMarried:CHINoChild    -0.2308     0.3930  -0.587    0.557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.213 on 959 degrees of freedom
## Multiple R-squared:  0.113,  Adjusted R-squared:  0.1102
## F-statistic: 40.73 on 3 and 959 DF,  p-value: < 2.2e-16
```

結果の書き方

この分散分析表の結果より以下のように結果を導き出すことが出来ます．交互作用のある分散分析により，主観的幸福度は結婚および子の有無によって異なるかを分析した．その結果，結婚については F(1, 959)=120.63(p< .001) であり，結婚が主観的幸福度に対して有意に影響を与えていることが明らかとなった．一方，子の有無については F(1, 959)=1.2102(p> .05)，結婚と子の有無の交互作用については F(1, 959)=0.3448(p> .05) であり，有意差は認められなかった．

結果の解釈

この結果は以下のように解釈することが出来ます．

$$(\quad) = 1.543(\ ) + 0.359(\ ) - 0.231(\ \times\ ) + 5.012$$

ただし，以下のように変数を割り振っています．

- 結婚：未婚 →0，既婚 →1
- 子ども：子なし →0，子あり →1

したがって，「未婚者かつ子なし」「未婚者かつ子あり」「既婚者かつ子なし」「既婚者かつ子あり」という 4 つのありえる状態について，次のように主観的幸福度を推定することが出来ます．

結果の解釈

- 「未婚者かつ子なし」

$$(\quad) = 1.543 \times 0 + 0.359 \times 0 - 0.231(0 \times 0) + 5.012$$

$$(\quad) = 5.012$$

- 「未婚者かつ子あり」

$$(\quad) = 1.543 \times 0 + 0.359 \times 1 - 0.231(0 \times 1) + 5.012$$

$$(\quad) = 0.359 + 5.012 = 5.371$$

- 「既婚者かつ子なし」

$$(\quad) = 1.543 \times 1 + 0.359 \times 0 - 0.231(1 \times 0) + 5.012$$

$$(\quad) = 1.543 + 5.012 = 6.555$$

- 「既婚者かつ子あり」

$$(\quad) = 1.543 \times 1 + 0.359 \times 1 - 0.231(1 \times 1) + 5.012$$

$$(\quad) = 1.543 + 0.359 - 0.231 + 5.012 = 6.683$$

ここから，未婚者に比べて既婚者の主観的幸福度が高いことはわかりますが，子の有無は主観的幸福度に対して影響をどうも与えなそうです.

分散分析（一般線形モデルによる分散分析モデルによる分析）

- 一般線形モデルによる分散分析モデル
  - ダミー変数が複数あるような状況を前提とする.
- 交互作用ありモデル：
  - 組み合わせによってパワーパップ or パワーダウン...

```
オブジェクト<-lm(応答変数　<-　説明変数1 * 説明変数2,
            data = データセットの名前)
これについて，回帰分析／t検定の時は以下のコードを使っています.
summary(オブジェクト)
これについて，分散分析の時は以下のコードを使っています.
anova(オブジェクト)
```

2 要因分散分析モデル (交互作用なし)

2 要因分散分析 (交互作用なし)

今までの例題，分散分析表からは「結婚」が主観的幸福度に影響を与えることは明らかになりましたが,「子の有無」や「結婚と子の有無の交互作用」は認められませんでした．したがって，結婚をしているかどうかで主

観的幸福度が高くなることは明らかとなりましたが，子がいるかどうかが主観的幸福度に影響を与えるとはいえないこと，さらに結婚しているかどうか，かつ子がいるかどうかという両者の影響が組み合わさっても影響がないことが明らかとなりました．

この結果は「未婚者かつ子なし」「未婚者かつ子あり」「既婚者かつ子なし」「既婚者かつ子あり」という 4 つの状態がありえました．

「未婚者」に比べて，「既婚者」の主観的幸福度が高いことがわかりましたが，子の有無が与える影響と，「結婚していることかつ子の有無が与える影響」についてはあるとは言えない結果が得られました．先程の「交互作用」は「結婚していることかつ子の有無が与える影響」を示しています．

しかし，この「交互作用」が認められなかった場合は「結婚が影響しているのか?」「子の有無が影響しているのか?」のみを検討する必要があります．すなわち，「交互作用」がない場合についても検討する必要があります．そのために，「交互作用なし」の分散分析をする必要があります

> ただし，いきなり「交互作用なし」の分析，すなわち「結婚していることかつ子の有無が与える影響」はないものとして検討することもあります．これについては研究領域の違いがあるので，その領域の慣習に従ってください．

言い換えると，交互作用なしの分析では「結婚していることかつ子の有無が与える影響」という組み合わせによる特別な影響はないことを前提とした分析ということになります．

モデル式で考えると，こんな感じです．

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \alpha + \epsilon_i$$

このモデル式によって以下の 4 つの状態を考えることができます．

- $X_1 = 1 X_2 = 0$ の場合

$$Y_i = \beta_1 * 1 + \beta_2 * 0 + \alpha + \epsilon_i$$

$$Y_i = \beta_1 + \alpha + \epsilon_i$$

  - この場合，ある因子 X_1 によって，傾きが変化することを示しています．
  - ex. 既婚で，子どもがいない人の幸福度がわかる．

- $X_1 = 0 X_2 = 1$ の場合

$$Y_i = \beta_1 * 0 + \beta_2 * 1 + \alpha + \epsilon_i$$
$$Y_i = \beta_2 + \alpha + \epsilon_i$$

  - この場合，ある因子 X_2 によって，傾きが変化することを示しています．
  - ex. 未婚で，子どもがいる人の幸福度がわかる．

- $X_1 = 1 X_2 = 1$ の場合

$$Y_i = \beta_1 * 1 + \beta_2 * 1 + \alpha + \epsilon_i$$
$$Y_i = \beta_1 + \beta_2 + \alpha + \epsilon_i$$

  - この場合，X_1 と X_2 が影響する場合の値を示していることになります．
    - ex. 既婚で，子どもがいる人の幸福度がわかる．

- $X_1 = 0 X_2 = 0$ の場合

$$Y_i = \beta_1 * 0 + \beta_2 * 0 + \alpha + \epsilon_i$$
$$Y_i = \alpha + \epsilon_i$$

  - この場合，全ての要因が影響しない場合（何らかの基準となる点）の値を示していることになります．
  - ex. 未婚で，子どもがいない人の幸福度がわかる．

仮説を立てる

さて，それでは仮説を立ててみましょう．今回分析するテーマは「主観的幸福度 (SUB_HAP) が子の有無 (CHI) と結婚 (MAR) によって異なる」かどうかを分析します．二要因分散分析（交互作用なし）の場合は以下のような仮説を立てます．

- 対立仮説 1：主観的幸福度の平均値は結婚によって異なる
- 対立仮説 2：主観的幸福度の平均値は子どもの有無によって異なる
- 帰無仮説 1：主観的幸福度の平均値は結婚によって異なるとはいえない
- 帰無仮説 2：主観的幸福度の平均値は子どもの有無によって異なるとはいえない

これらの仮説のもとに分析を行ないます．

平均値をプロットする

さて，最初のお約束です．平均値をプロットしましょう．まずは各自でやってみましょう．

さて，例によって ggplotgui を使いましょう．

以下のコードは Console（コンソール）に直接打ち込みます．

```
library(ggplotgui)
ggplot_shiny(exdataset)
```

そうすると新しいウィンドウが開きます．

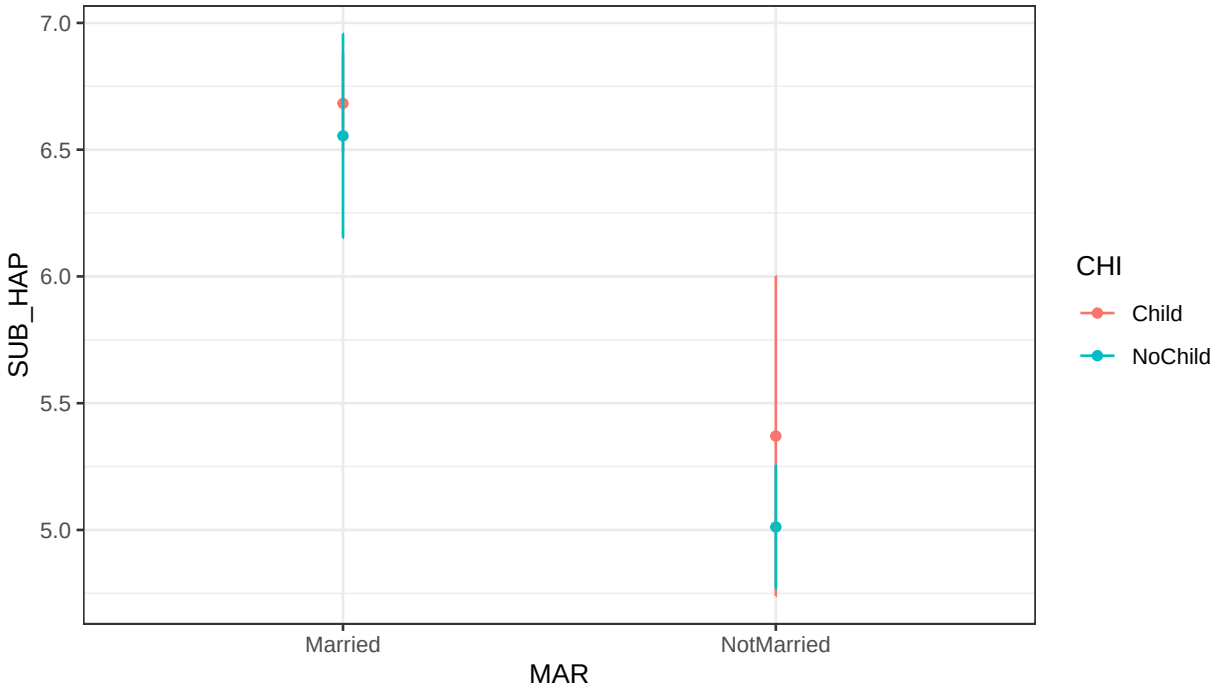
以下の通りの作業をしましょう．

- ggplot タブへ
- “Type of graph:” は “Dot + Error”，Y-variable は “SUB_HAP”，X-variable は “MAR” を設定
- “Group(or colour)” を CHI に変更
- “Confidence Interval:” を 95% にする．
- R-code タブへ行って，以下のコードのうち，真ん中のみを以下にする．-また，コード内の df を exdataset に変える．

```
# You need the following package(s):
library("ggplot2")

# The code below will generate the graph:
graph <- ggplot(exdataset, aes(x = MAR, y = SUB_HAP, colour = CHI)) +
  geom_point(stat = 'summary', fun.y = 'mean') +
  geom_errorbar(stat = 'summary', fun.data = 'mean_se',
                width=0, fun.args = list(mult = 1.96)) +
  theme_bw()
```

```
graph
```



このグラフを見る限り，未婚者に比べて既婚者の方が主観的幸福度が高そうですが，子の有無の影響はありそうな気もしますし，なさそうな気もしますし何とも言えません．したがって，この点についても統計的に差があるのかどうかを明らかにしましょう．

```
marchihap_model_noint <-lm(SUB_HAP ~ MAR+CHI, data = exdataset)
# モデル式を marchihap_model_noint というオブジェクトにしまいます．
```

```
# 分析結果の要約を出力する
summary(marchihap_model_noint)
```

```
##
## Call:
## lm(formula = SUB_HAP ~ MAR + CHI, data = exdataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6991 -1.6991  0.3009  1.3009  4.9667
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.6991     0.1015  66.015  < 2e-16 ***
```

```
## MARNotMarried  -1.4642     0.1862  -7.863 1.01e-14 ***
## CHINoChild     -0.2016     0.1832  -1.100    0.271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.212 on 960 degrees of freedom
## Multiple R-squared:  0.1127, Adjusted R-squared:  0.1108
## F-statistic: 60.96 on 2 and 960 DF,  p-value: < 2.2e-16
```

結果の書き方

交互作用のない分散分析により，主観的幸福度は結婚しているか否か，および子どもがいるか否かによって異なるかを分析した．その結果，結婚の影響は $F(1, 960)=120.71(p<.001)$ であり，結婚は主観的幸福度に対して有意に影響を与えることが明らかとなった．一方，子の有無の影響は $F(1, 960)=11.21(p>.05)$ であり，有意な影響は認められなかった．

結果の解釈

この結果は以下のように解釈することが出来ます．

$$(\quad) = 1.464(\ ) + 0.202(\ ) + 5.033$$

ただし，以下のように変数を割り振っています．

結婚：未婚 $\rightarrow$0，既婚 $\rightarrow$1

子ども：子なし $\rightarrow$0，子あり $\rightarrow$1

したがって，結婚と子の有無の影響は以下のように表すことができます．

結果の解釈

- 「未婚者かつ子なし」

$$(\quad) = 1.464 \times 0 + 0.202 \times 0 + 5.033$$

$$(\quad) = 5.033$$

- 「未婚者かつ子あり」

$$(\quad) = 1.464 \times 0 + 0.202 \times 1 + 5.033$$

$$(\quad) = 0.202 + 5.033 = 5.235$$

- 「既婚者かつ子なし」

$$(\quad) = 1.464 \times 1 + 0.202 \times 0 + 5.033$$

$$(\quad) = 1.464 + 5.033 = 6.497$$

- 「既婚者かつ子あり」

$$(\quad) = 1.464 \times 1 + 0.202 \times 1 + 5.033$$

$$(\quad) = 1.464 + 0.202 + 5.033 = 6.699$$

ここから，未婚者に比べて既婚者の主観的幸福度が高いことはわかりますが，子の有無は主観的幸福度に対して影響をどうも与えなそうです．

モデル選択：

モデル選択とは，複数の統計モデルを比較する時に用いる手法です．ここでは，モデル選択の手法として，分散分析によるモデル選択と AIC に基づくモデル選択を紹介します．

尤度比検定によるモデル選択：

分散分析に基づいた近似計算とは，２つのモデル式をもとにして，分散分析を用いることでモデル選択をすることができます．

ここでは，主観的幸福度を応答変数として，説明変数として未既婚と子どもの有無を設定したモデルについて，交互作用ありとなしの 2 つを比較します．

```
anova(marchihap_model,marchihap_model_noint)
```

```
## Analysis of Variance Table
##
## Model 1: SUB_HAP ~ MAR * CHI
## Model 2: SUB_HAP ~ MAR + CHI
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    959 4695.7
## 2    960 4697.4 -1   -1.6883 0.3448 0.5572
```

この結果はモデル 1 である交互作用ありモデルと，モデル 2 である交互作用なしモデルを比較すると，どちらのモデルも差があるとはいえない確率が 55% もあるということを示しております．

この場合は，より単純なモデルとして交互作用のないモデルを選択します．

ちなみに，「分散分析」は実は作成したモデルと説明変数の入っていない「ヌルモデル」を比較しているものと同値になります．

AIC：

AIC とは Akaike's Information Criterion（赤池情報量規準）と呼ばれるものであり，モデル評価の規準の一つです．

$$AIC = -2log(\ \ ) + k \times (\ )$$

として算出され，この値が最小になるモデルを採択します．特に，2 つのモデルを選択する時には 2 つのモデルについて AIC の差分が 2 以上あるとそのモデルを選択することができます．

まずは AIC を算出してみましょう．

```
AIC(marchihap_model,marchihap_model_noint)
```

```
##                         df      AIC
## marchihap_model          5 4268.608
## marchihap_model_noint    4 4266.954
```

ここでは交互作用なしのモデルの方が小さい値を示しています．2 つのモデルの AIC の差分が 2 はギリギリありませんが，この場合は交互作用なしのモデルを採択してもよいかと思います．

AIC：

実際には，こんな感じで複数モデルを並列して，比較できる形で示すことが多いです．

```
library(huxtable)
huxreg(marchihap_model,marchihap_model_noint)
```

ここでは交互作用なしのモデルの方が小さい値を示しています．2 つのモデルの AIC の差分が 2 はギリギリありませんが，この場合は交互作用なしのモデルを採択してもよいかと思います．

	(1)	(2)
(Intercept)	6.683 ***	6.699 ***
	(0.105)	(0.101)
MARNotMarried	-1.312 ***	-1.464 ***
	(0.319)	(0.186)
CHINoChild	-0.128	-0.202
	(0.222)	(0.183)
MARNotMarried:CHINoChild	-0.231	
	(0.393)	
N	963	963
R2	0.113	0.113
logLik	-2129.304	-2129.477
AIC	4268.608	4266.954

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.