統計学第 9 講/第 10 講

後藤　晶　　akiragoto@meiji.ac.jp

目次

前回の復習 1

t 検定 5

1 要因分散分析 7

2 要因分散分析モデル (交互作用あり) 13

2 要因分散分析モデル (交互作用なし) 18

演習問題 26

目次

前回の復習

ダミー回帰分析

t 検定とは 2 群の「平均値」を比較する方法です．しかし，実はこれも一般線形モデルの枠組みの中で考えることが出来ます．ここではその考え方について説明します．そこには「ダミー変数」という考え方が必要になります．

ダミー変数とは

一般線形モデルではこんなモデル式から考える，というような話をしたかと思います．

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

- 回帰分析では Y_i と X_1 が数値データだった場合を示していました．しかし，例えば X_1 に入れたいのが未婚者か既婚者，という因子データだったとします．

- この場合は，未婚者に対して 0，既婚者に対して 1 という数字を割り当てると次のように理解することができます．

0 を割り振られた未婚者の場合

数式の X_1 に 0 を代入しましょう．

$$Y_i = \alpha + \epsilon_i$$

- 係数がなくなってしまいました．したがって，切片のみになります．

1 を割り振られた既婚者の場合

数式の X_1 に 1 を代入しましょう．

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

- X_1 の係数のみが増えています．したがって，0 を代入した未婚者に比べて，既婚者の方が β_1 の分だけ変化していることがわかります．

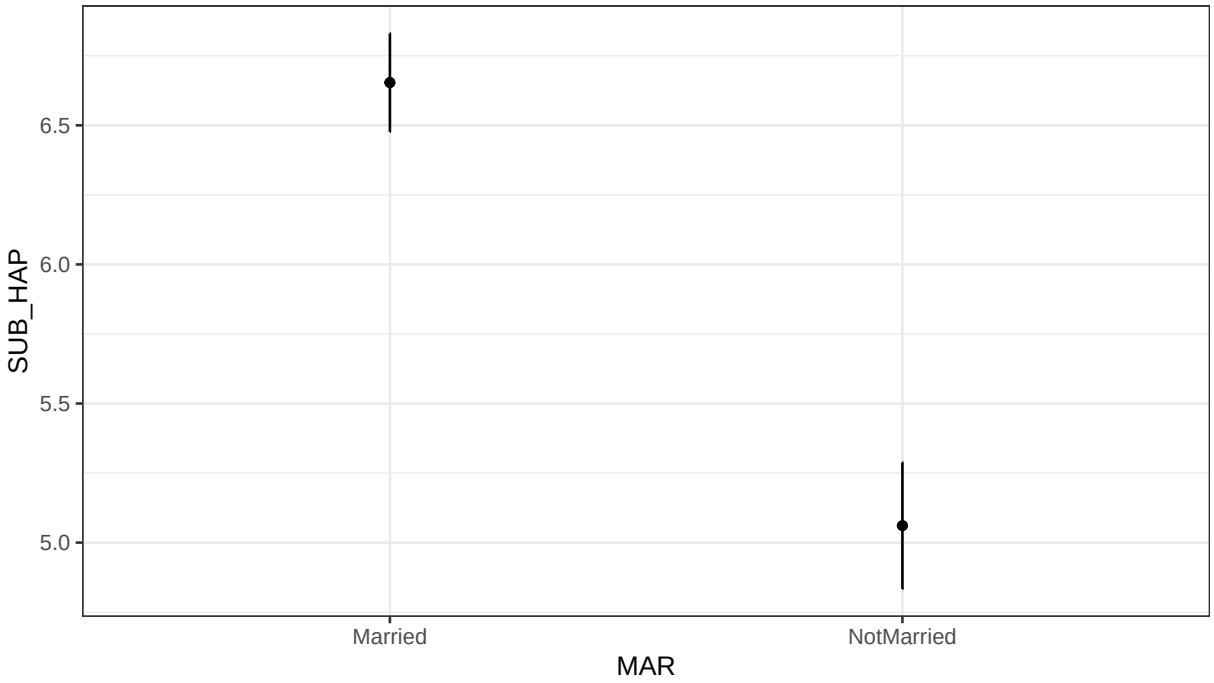
このように，0 か 1 の数字を入れてあげると 0 を入れられたグループと 1 を割り振られたグループでどれだけ差があるのか，ということを評価することができます．
そして，その「差」がどの程度あるのかも比較することができます．ここでは，主観的幸福度に未婚者と既婚者の間に差があるのか否かを，先ほどと同じような流れで考えていきましょう．

仮説を立てる

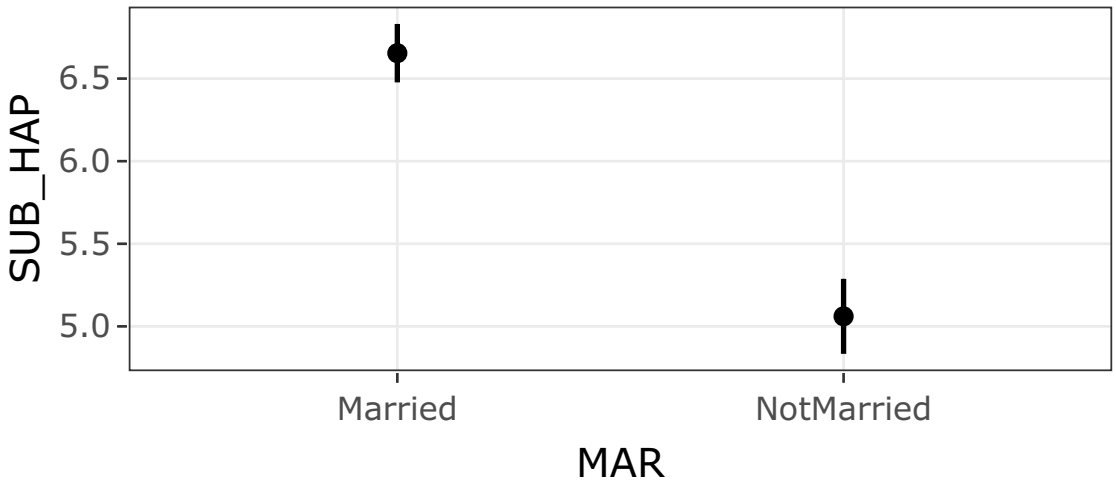
t 検定に当たるのは 2 つの群に差があるのか否か，です．「差がある」を対立仮説，「差があるとはいえない」を帰無仮説とします．したがって，以下のような仮説を立てることが出来ます．

- 対立仮説：未婚者と既婚者の主観的幸福度に差がある．
- 帰無仮説：未婚者と既婚者の主観的幸福度に差があるとはいえない．

```
graph
```




```
# If you want the plot to be interactive,
# you need the following package(s):
library("plotly")
ggplotly(graph)
```



- 0 は未婚者を，1 は既婚者を示しています．

これも同様に，本当に差があるのかどうかは，感覚的には明らかになっても科学的な根拠がありません．同じように検定をして確かめてみましょう．

ダミー回帰をやってみる

```
#"hapsat_model"というオブジェクトに，分析モデルを代入する．
marhap_model<-lm(SUB_HAP ~ MAR, data = exdataset)
```

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

```
##
## Call:
## lm(formula = SUB_HAP ~ MAR, data = exdataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6538 -1.6538  0.3462  1.3462  4.9391
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.65378    0.09274   71.74   <2e-16 ***
## MARNotMarried -1.59286    0.14499  -10.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.212 on 961 degrees of freedom
## Multiple R-squared:  0.1116, Adjusted R-squared:  0.1106
## F-statistic: 120.7 on 1 and 961 DF,  p-value: < 2.2e-16
```

分析結果の見方

- さて，この分析結果の見方は基本的なところは回帰分析と一緒です．
- 特に着目すべきは Coefficients のところなので，こちらについて説明します．

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0609     0.1115   45.41   <2e-16 ***
## MAR           1.5929     0.1450   10.99   <2e-16 ***
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

この結果について，またモデル式と共に説明します．この結果は α が 5.0609，β が 1.5929 という結果でした．したがって，モデル式は以下のように示すことができます．

$$Y_i = 1.59291X_1 + 5.0609 + \epsilon_i$$

まずは係数について説明します．これは未婚者の場合と既婚者の場合について考えてみましょう．

未婚者の場合

未婚者の場合は X_1 が 0 でした．したがって，以下のように示されます．

$$Y_i = 5.0609 + \epsilon_i$$

- すなわち，未婚者の平均値の予測は 5.0609 であると推定されます．

既婚者の場合

既婚者の場合は X_1 が 1 でした．したがって，以下のように示されます．

$$Y_i = 1.59291 + 5.0609 + \epsilon_i$$

- したがって，平均値は 6.65381 であると推定されます．
- また，これらの推定値の妥当性は p 値によって推定されます．
- いずれの結果についても 0.001% 以下であるためにこの結果は統計的にも明らかな差があると理解できます．
- したがって，未婚者に比べて，既婚者の主観的幸福度は明らかに高いと理解することができます．この結果を簡単にまとめましょう．

結果の表記例.

- ダミー回帰分析モデルによって未婚者に比べて，既婚者の方が主観的幸福度が 1.59 高いこと 0.001% 水準で示された．（一緒に表を見せると良い．）
- ダミー回帰分析モデルによって未婚者に比べて，既婚者の方が主観的幸福度が 1.59 高いことが示された．(t(961)=10.99, p<.001)．
- $(\quad) = 1.59291(t = 10.99) \times (\quad)$
  $+ 5.0609 + \epsilon_i$

t 検定

t 検定

今までは一般線形モデルの枠組みから t 検定の紹介を，すなわちダミー回帰分析の 1 つとしての t 検定を紹介しました．一方で，普通の t 検定は以下のように行うことができます．

ここだけの話.

- 最近は t 検定にもいろいろな方法が提案されています．従来は等分散性を検定する F 検定を実施し後に，等分散を仮定したスチューデント (Student) の t 検定を行ったり，不等分散を仮定したウェルチ (Welch) の t 検定を実施する，ということが行われてきました．
- しかしながら，2 回検定を行うことは「検定の多重性」の観点から問題ではないか，という指摘もあったりします．
- そこで，最近では F 検定を実施せずにいきなりウェルチの t 検定を行うことが多くなっています．

ウェルチの t 検定

```
welch_t.testmodel<-t.test(SUB_HAP ~ MAR, data = exdataset)
welch_t.testmodel
```

```
##
##  Welch Two Sample t-test
##
## data:  SUB_HAP by MAR
## t = 10.854, df = 808.29, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.30479 1.88094
## sample estimates:
##    mean in group Married mean in group NotMarried
##                 6.653779                 5.060914
```

今日の要約

- ダミー回帰分析：
  –

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

– をモデル式として，“Y_i” が数値データ，“X_1” が 1or0 の場合に用いる．
  * 対立仮説：“説明変数” の有無に応じて，“応答変数” が影響を受ける
  * 帰無仮説：“説明変数” の有無に応じて，“応答変数” が影響を受けるとは言えない．
– R の関数では次の形式を用いる．

```
# モデルを作る
オブジェクト <- lm(応答変数 ~ ダミー変数,
                   data = データセットの名前)

# 結果を出力する
summary(オブジェクト)
```

今日の要約

- t 検定
  – 平均値の差の検定としての t 検定は以下のように検定する．

```
オブジェクト <- t.test(応答変数(数量データ) ~
                         説明変数(2つで分けられるもの),
                       data = データセットの名前)
オブジェクト
```

1 要因分散分析

分散分析とは

分散分析とは，「3 群以上の分散に差があるかどうか」を比較・分析するための方法です．その後「多重比較」という手法を用いて，「3 群以上の平均値の差があるかどうか」を明らかにします．この授業では「1 元配置分散分析」および「2 元配置分散分析」というものについて説明します．いずれについても，説明変数が因子データ，応答変数が数値データとなります．

- 1 元配置分散分析：「地域によって，主観的幸福度の分散・平均値が異なる」などのような，1 つの要因によって影響を受けるかどうかを分析する手法です．
- 2 元配置分散分析：「地域と未婚・既婚によって分散・平均値が主観的幸福度が異なる」，「地域と子の有無によって主観的幸福度が異なる」などのような，2 つの要因によって影響を受けるかどうかを分析する手法です．

分散分析を一般線形モデルの枠組みで説明すると，平均値の推定がベースとなりますが，以下のように理解することができます．ここでは，「3 つの群の影響を受ける」場合について，モデル式を元に説明します．また，以下では「分散分析モデル」という表現をします．

- 個人的には一般線形モデルの枠組みの方が理解しやすいと思っています..

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

このモデルでは X_1 と X_2 はそれぞれ $(1, 0)$ の値を取る「ダミー変数」です．しかし，これでは β が 2 つしかありません．しかし，これだけで 3 つの群を表すことができます．以下には 3 つの条件についてモデル式を書き入れてあげたいと思います．

- $X_1 = 1$ と $X_2 = 0$ の場合
$$Y_i = \beta_1 + \alpha + \epsilon_i$$
  - 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．
- $X_1 = 0$ と $X_2 = 1$ の場合
$$Y_i = \beta_2 + \alpha + \epsilon_i$$
  - 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．
- $X_1 = 0$ と $X_2 = 0$ の場合
$$Y_i = \alpha + \epsilon_i$$
  - 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．

このモデルについて，平均値が異なるかどうかを調べます．特に，分散分析の場合は「分散分析表」と呼ばれるものを出して評価してあげます．

分散分析モデルの例

- テストの点数がクラスによって異なる．

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

- $X_1 = 1$ と $X_2 = 0$: B クラス
$$Y_i = \beta_1 + \alpha + \epsilon_i$$

- $X_1 = 0$ と $X_2 = 1$: C クラス
$$Y_i = \beta_2 + \alpha + \epsilon_i$$

- $X_1 = 0$ と $X_2 = 0$: A クラス
$$Y_i = \alpha + \epsilon_i$$

- このモデル式からわかること：A クラスに比べて B クラス／ C クラスの得点が高いか低いか

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) が地域 (SUB_ARE) によって異なる」かどうかを分析します．一要因分散分析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地域によって異なる．
- 帰無仮説：主観的幸福度の平均値は地域によって異なるとは言えない．

この 2 つの仮説のもとに分析を行ないます．

分析のモデル式

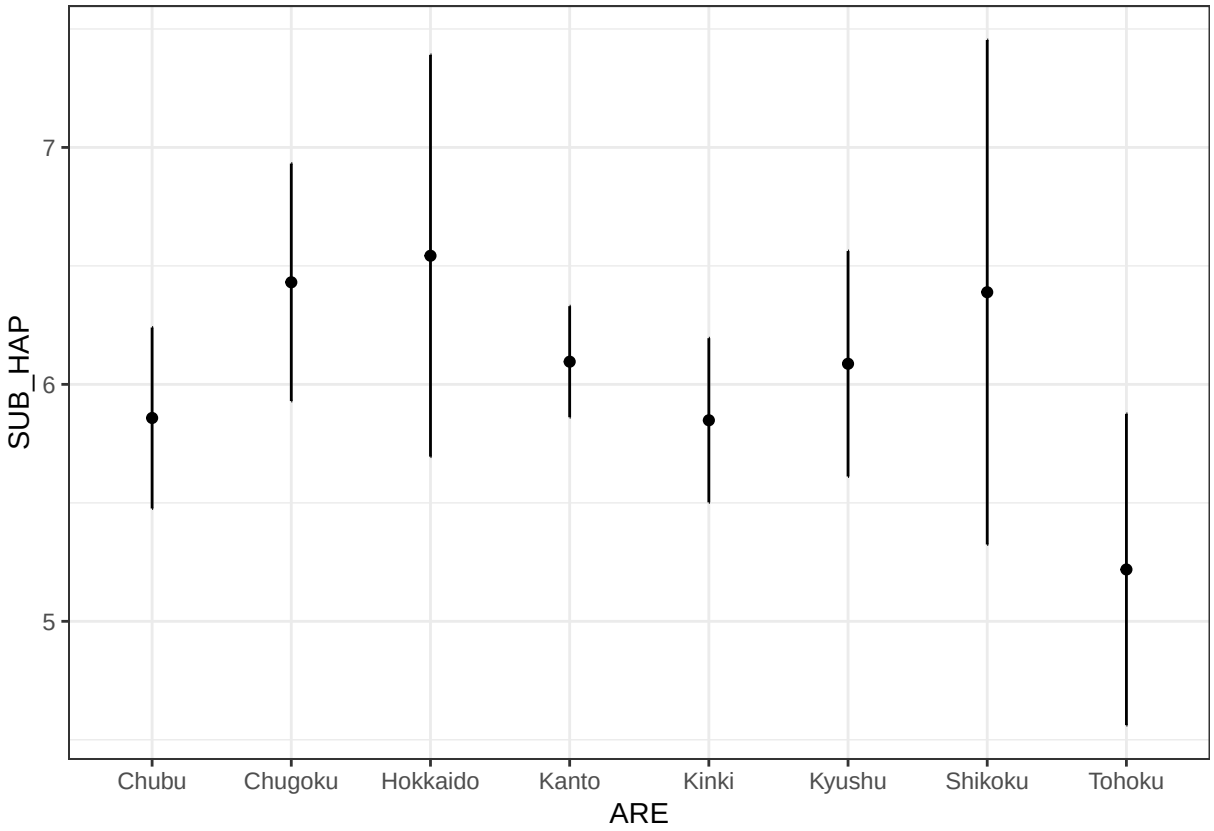
今回の分析には，以下のモデルを前提とします．

$$(\quad) = \beta_1(\ \ ) + \beta_2(\ \ ) +$$
$$\beta_3(\ \ ) + \beta_4(\ \ ) + \beta_5(\ \ ) +$$
$$\beta_6(\ \ ) + \beta_7(\ \ ) + \alpha + \epsilon_i$$

- なお，このモデルではそれぞれの値は 1 か 0 の値しか取りません．
- ex. 東北地方のデータである場合には，東北ダミーが 1，それ以外のダミー変数は 0 を取ります．
- また，すべてのダミー変数が 0 の場合はコントロール群となる関東地方の値を示しています．

そうすると，こんなグラフが算出されます．

```
graph
```



このグラフを見る限り，地域ごとに差があるかどうかはわかりません．以前，平均値を算出してみたことがありましたが，今回はそれぞれが「統計的に差がある」ということが言えるかどうかを考えたいと思います．

分析をやってみる

さて，分散分析モデルを作成してみましょう．

```
#"arehap_model"というオブジェクトに，分析モデルを代入する．
arehap_model<-lm(SUB_HAP ~ ARE, data = exdataset)
```

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

```
##
## Call:
## lm(formula = SUB_HAP ~ ARE, data = exdataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5429 -1.4308  0.1515  1.9043  4.7813
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.858108   0.192158  30.486   <2e-16 ***
## AREChugoku   0.572661   0.347849   1.646   0.1000
## AREHokkaido  0.684749   0.439390   1.558   0.1195
## AREKanto     0.237637   0.226845   1.048   0.2951
## AREKinki    -0.009623   0.264660  -0.036   0.9710
## AREKyushu    0.228848   0.310363   0.737   0.4611
## AREShikoku   0.530781   0.583547   0.910   0.3633
## ARETohoku   -0.639358   0.349733  -1.828   0.0678 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.338 on 955 degrees of freedom
## Multiple R-squared:  0.01418,    Adjusted R-squared:  0.006954
## F-statistic: 1.962 on 7 and 955 DF,  p-value: 0.05729
```

- 出力結果が入り切らないので Coefficients だけ示します.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.095745   0.120558  50.563  < 2e-16 ***
AREHokkaido  0.447112   0.413125   1.082  0.27941
ARETohoku   -0.876995   0.316105  -2.774  0.00564 **
AREChubu    -0.237637   0.226845  -1.048  0.29510
AREKinki    -0.247260   0.218299  -1.133  0.25764
AREChugoku   0.335025   0.314020   1.067  0.28629
AREShikoku   0.293144   0.564036   0.520  0.60338
AREKyushu   -0.008788   0.271909  -0.032  0.97422
```

- α は 6.095745 である.
- 関東地方と比べて，東北地方の主観的幸福度が低い.
  - 実は昔から言われている結果.
  - 東日本大震災の影響？という声もあったが逆で，東日本大震災によって幸福度が改善したとも言われている.
- その他の地域は影響が認められなかった.

分析結果の解釈

- さらに，モデル式による分析結果を出力しました．この結果が示しているのは以下のようなことです.

$$
\begin{aligned}
(\quad) = 0.447112 * (\quad) - 0.876995 * (\quad) - \\
0.237637 * (\quad) - 0.247260 * (\quad) + \\
0.335025 * (\quad) + 0.293144 * (\quad) - \\
0.008788 * (\quad) + 6.095745 + \epsilon_i
\end{aligned}
$$

- 今度はモデル式についても同じように出力してあげましょう.
- 回帰分析や t 検定と同じです.

分散分析表の出力

```
# 分散分析表
anova(arehap_model)
```

```
## Analysis of Variance Table
##
## Response: SUB_HAP
##            Df Sum Sq Mean Sq F value  Pr(>F)
## ARE         7   75.1 10.7238  1.9623 0.05729 .
## Residuals 955 5218.9  5.4648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

分散分析表の読み方：

`Analysis of Variance Table`

- 分散分析表です. 分散分析の結果を示しています.

```
Response: SUB_HAP
           Df Sum Sq Mean Sq F value  Pr(>F)
ARE         7   75.1 10.7238  1.9623 0.05729 .
Residuals 955 5218.9  5.4648
```

- `Df` は自由度を示しています.
- `Sum Sq` は平方和
- `Mean Sq` は平均平方
- `F value` は F 値
- `Pr(>F)` は p 値を示しています.

`Response: SUB_HAP`

```
           Df Sum Sq Mean Sq F value  Pr(>F)
ARE         7   75.1 10.7238  1.9623 0.05729 .
Residuals 955 5218.9  5.4648
```

- 応答変数は SUB_HAP です.
- ARE の自由度（分子自由度）は 7：全部で 8 地域ある →N-1 が自由度
  - モデル式の β（パラメータ）の数と一致している.
  - DF は Degree of Freedom
- ARE の F 値は 1.9623，P 値は 0.05729
- Residuals の自由度（分母自由度）は 955：全部で 963 個のデータがあり，モデル式の β（パラメータ）で 7 つ，さらにもう 1 地域（$=\alpha$ で使われる）を引いたもの.

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- p 値の大きさを示す記号.
- 0 -0.001 では *** で表される.
- 0.001-0.01 では ** で表される.
- 0.01-0.05 では * で表される.
- 0.05-0.1 では . で表される.
- 0.1-1 では何にもありません.

書き方

- 主観的幸福度は地域によって異なるかを分析した．その結果，F(7, 955)=1.9623(p< .10) であり，10% 水準で有意にあることが示されている．したがって，主観的幸福度は居住地域によって異なる傾向にあることが示されている.
  - 分散分析表を合わせて示してあげましょう.
  - ちなみに，心理学などでは有意水準を 5% に設定されることが多い.
  - 経済学系では 10% 水準を採用することもある.
  - いずれにしろ，分析の前に有意水準を設定する必要がある.

自由度とは

- 自由度＝ n-p
  - n：標本の大きさ
  - p：推定されたパラメータの数
- 自由度＝ n-q-1
  - n：標本の大きさ
  - q：モデル式で推定されたパラメータ（β）の数
  - 1 は（α）の分

要約

- 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.

```
オブジェクト<-lm(応答変数　~　説明変数,
                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 (交互作用あり)

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.

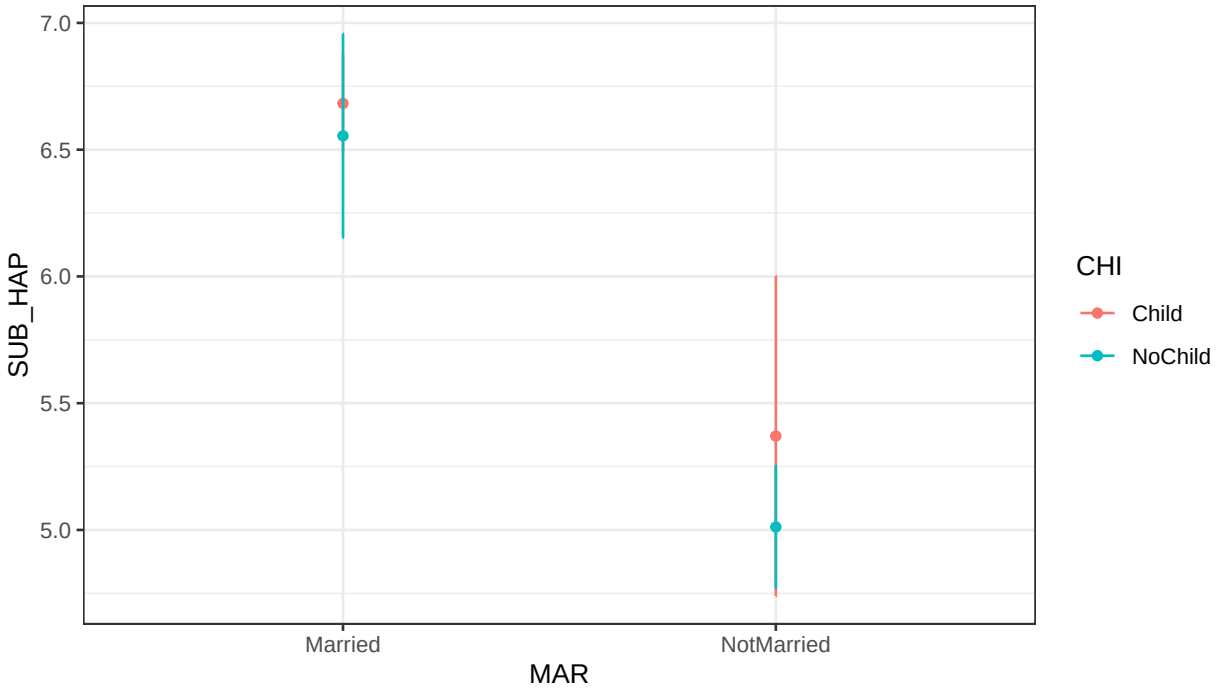
以下の通りの作業をしましょう.

- 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．

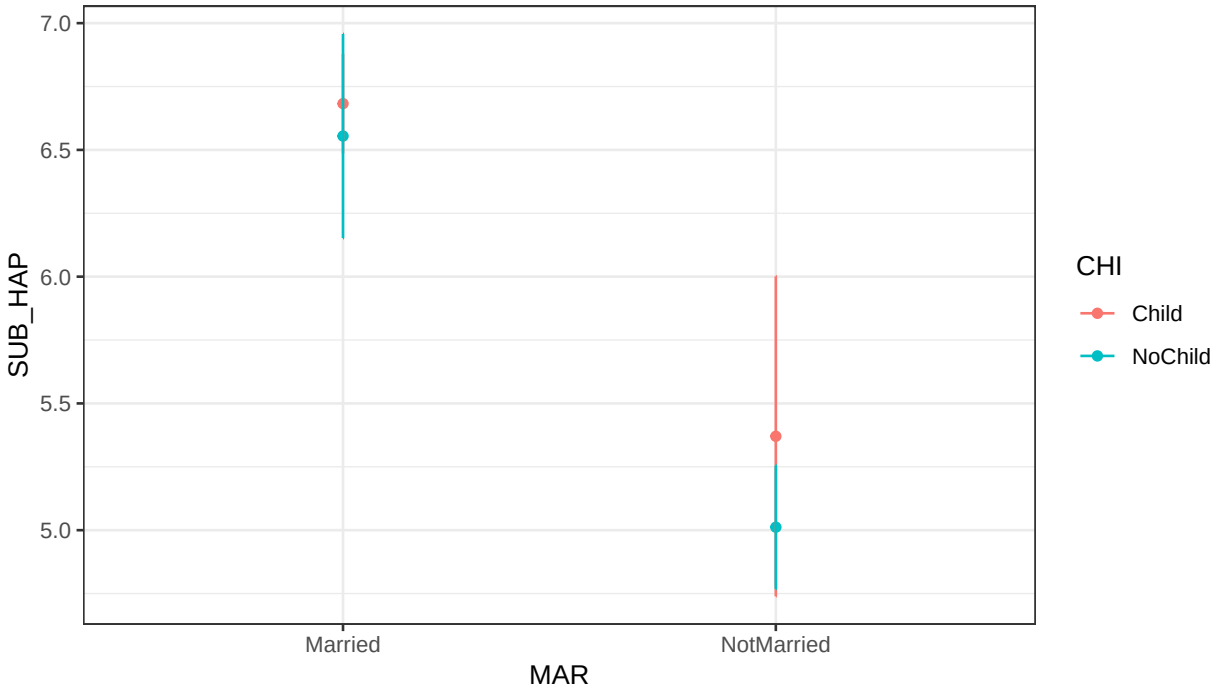
以下の通りの作業をしましょう．

- 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．

```r
# 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()
```

```r
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．

```r
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.
```

```r
#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.

今日の Take Home Messages

- 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特別なパワーパップ or パワーダウンあり！
- 交互作用なしモデル：
  - 組み合わせによって特別なパワーパップ or パワーダウンしない.

```
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)
```

演習問題

演習問題 1

“SUB_HAP” は主観的幸福度，“SUB_SAT” は生活満足度，“SUB_SLP” は睡眠満足度に関するデータであった（各 10 点尺度）．これらを応答変数，性別を表す “F_SEX” 及び結婚を示す “MAR”(0: 未婚，1: 既婚) を説明変数として，以下の 3 つの「交互作用あり」と「講義作用なし」の分析を実施せよ．それぞれについてグラフ，分散分析表とモデルの結果を出力すること．

- 主観的幸福度の性別差（男女その他）および既婚・未婚の影響を分析せよ．
- 生活満足度の性別差（男女その他）および既婚・未婚の影響を分析せよ．
- 睡眠満足度の性別差（男女その他）および既婚・未婚の影響を分析せよ．

演習問題 2

同様に，可能であればこちらについても挑戦すること．子の有無は “CHI”(0: 子なし，1: 子あり) で示されている．

- 主観的幸福度の性別差（男女その他）および既婚・未婚と子の有無の影響を分析せよ．
- 生活満足度の性別差（男女その他）および既婚・未婚と子の有無の影響を分析せよ．
- 睡眠満足度の性別差（男女その他）および既婚・未婚と子の有無の影響を分析せよ．