統計学第12/13講

明治大学情報ｺﾐｭﾆｹｰｼｮﾝ学部
後藤　晶
akiragoto@meiji.ac.jp

今日のお話

前回の復習

回帰分析

ダミー回帰分析とt検定

t検定

1要因分散分析

演習問題

2要因分散分析モデル（交互作用あり）

前回の復習

概要

一般線形モデルとは，統計学の中でも，以下の数式（モデル式）を元に考えていくモデルです．

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...\alpha + \epsilon_i$$

さて，何か複雑そうなモデル式が出てきてしまいましたが，恐れることはありません．少し，簡単な形にしてあげましょう．そうすると，こんな感じに書くことが出来ます．

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

このモデル式，何だか見覚えのあるグラフとそっくりだと思います．中学校の時に”一次関数”というのを教わったのを覚えていますでしょうか？一次関数ではこんな数式を使いました．

$$Y = \beta X + \alpha$$

この数式を元に，グラフを書く，ということもやったかと思います．この時，β を傾き，α を切片という呼び方をしていました．ちなみに，この数式で直線のグラフを書く時には，Xに0を代入した時のポイント$(0, \alpha)$とXに1を代入したときのポイント$(1, \beta + \alpha)$を結ぶ直線を引いてあげれば，グラフを作成することができます．

一般線形モデルの一番理解しやすい最初の考え方は，「実際に観察されたデータを元にして，一次関数のような直線を引いてあげよう！」という発想です．ただし，一次関数とちょっと違うのは「全ての点を通らなくてよい」ということです．

誤差

一次関数の場合はその直線上にある全ての点を通ることが前提となっていました．しかし，実際には直線であるので，直線上の2点を通れば，全てその条件を満たす直線を引くことが出来ます．

しかし，一般線形モデルの場合は常に全ての点を通るとは限りません．ベストは全ての点を通ることではありますが，実際にはデータには「誤差」というものが存在します．これは本来得られるべき結果と実際に得られた結果にずれがあることを示しています．

この誤差には大きく分けて次の3種類あります．

3種類の誤差

- 測定誤差：実際に何かを計測する時に生じる誤差．中でも以下の2種類がある．
  - 系統誤差（システマティック）：何らかの要因により，常に生じてしまう誤差．例えば，自動車で運転者が40km/hで走っているつもりであっても，外部から正確なスピードメーターによって調べると38km/hしか出ていない，など．これはメーターが原因で生じる系統（システマティック）誤差である．
  - 偶然誤差：何らかの要因により，偶然生じてしまう誤差．例えば，ブレーキをかけたときに60mで普段止まるが，偶然入ったホコリや水分などによって70mで止まってしまうかもしれない．これは偶然入ったホコリや水分による偶然誤差である．

- 計算誤差：数値をどこかで四捨五入したことによって生じる誤差．例えば，1/3を0.333にして計算することによって計算誤差が生じる．
- 統計誤差（標準誤差）：母集団からある一部の集団を取り出す時，選ぶ集団によってどの程度数値が異なり得るのかを調べたもの．統計的に異なり得る範囲を推測することができる，

本題に戻って

さて，少し本題に戻りましょう．ちょっと一般線形モデルのモデル式を考えたいと思います．

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

改めて，このモデル式を説明したいと思います．ここで，“Y_i”のことを”応答変数”，“X_1”のことを”説明変数”と呼びましょう．

文字についている”$_i$“は各データによって異なる！という区別をするためについています．ちなみに，”Y_i“は他にも，被説明変数と呼ばれたりします．

また，”β_1” は係数，“α” は切片と呼ばれます．そして，“ϵ_i”が一番問題となる誤差です．この誤差は予測されたモデル式である”$Y_i = \beta_1 X_1 + \alpha$“からどれだけそのデータの値が離れているかを示しています．

と，言ってもなかなか理解し難いと思うので，一つ試しにやってみましょう．ここでは，「回帰分析」という方法と「t検定」という方法についてお話をしたいと思います．

検定名	応答変数	説明変数
回帰分析	数値データ	数値データ(順序データ)
t検定	数値データ	因子データ(ダミー変数,1, 0)

回帰分析

回帰分析とは

　回帰分析とは，応答変数が数値データであり，説明変数も数値データである場合に用いる方法です．例えば，「身長」と「体重」の間の相関関係について分析をする際にも用います．ここでは，今まで授業で使ってきた「主観的幸福度」と「生活満足度」の間に相関関係があるかどうか，以下の順番に沿って考えてみましょう．

この関係はモデル式で表すと，このような形になります．

$$(SH) = \beta_1(LS) + \alpha + \epsilon_i$$

この時，切片であるαは生活満足度が0であった時に対応する主観的幸福度を示しています．

仮説を立てる

何はともあれ，統計分析をするときには仮説を立ててあげる必要があります．仮説を立てるときには，「帰無仮説」と「対立仮説」の2つを考える必要があります． 対立仮説は「イイタイコト」，帰無仮説は「イイタイコトではないこと」でした．

ここで主観的幸福度と生活満足度の関係ですので，以下のように設定できます．

- 対立仮説：生活満足度が変化するにつれて，主観的幸福度も変化する．
- 帰無仮説：生活満足度が変化するにつれて，主観的幸福度も変化するとはいえない．

特に，以下では応答変数を主観的幸福度，説明変数を生活満足度とします．

散布図をプロットする

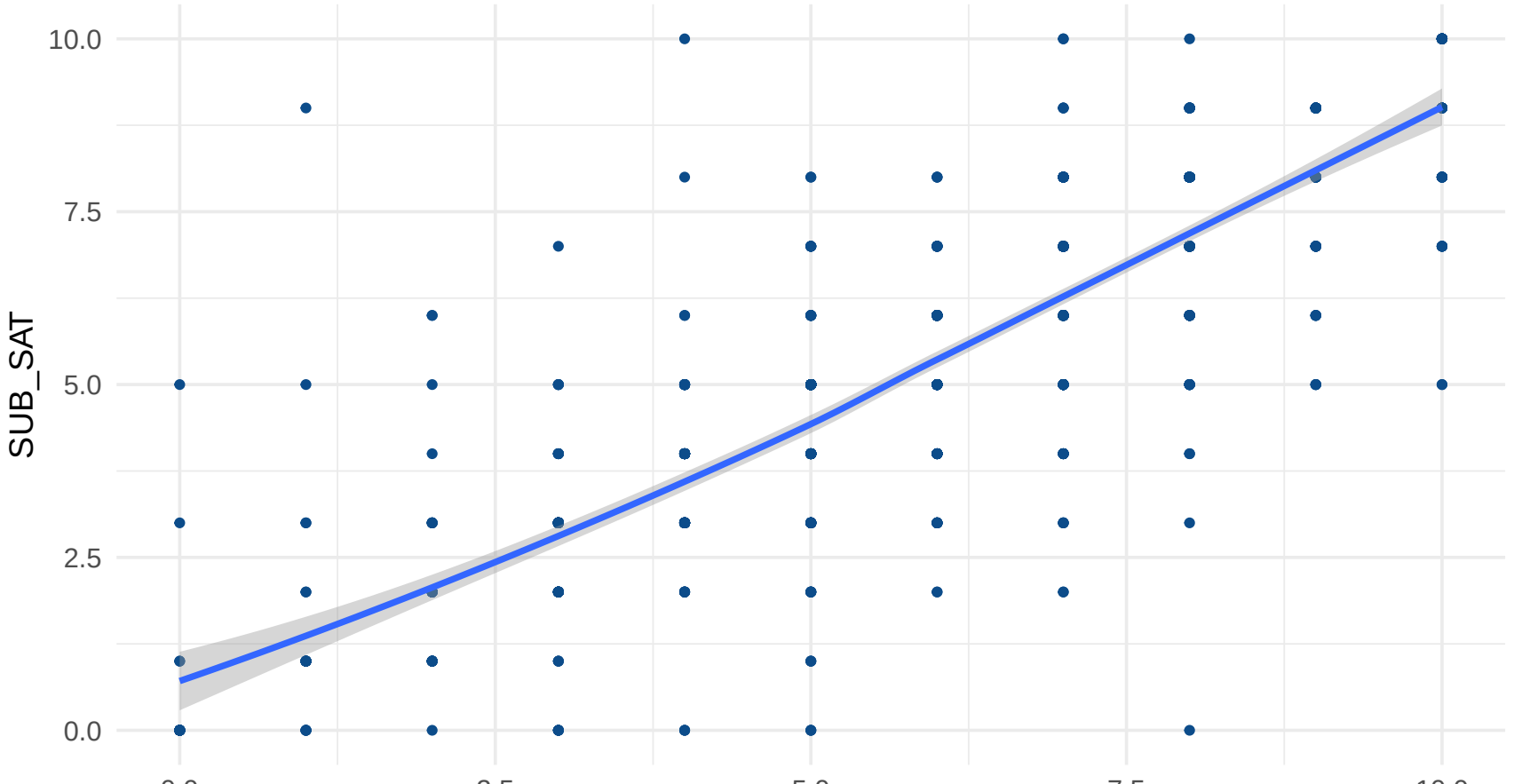
はじめに，分析対象となるデータを読み込んでおきましょう．
* もちろん，既に読み込んである場合は飛ばしてもらって構いません．

散布図のプロットは他の機能から持ってきてもよいのですが，今回はRStudio上でクリックだけで入れられる方法を紹介します．

その上で，コードを貼り付けて出力することにしましょう．

```
library(ggplot2)

ggplot(exdataset) +
 aes(x = SUB_HAP, y = SUB_SAT) + geom_point(size = 1L, colo
 geom_smooth(span = 1L) + theme_minimal()
```



どうもグラフを見ている限りだと，この2変数間には正の相関関係，すなわち「生活満足度が高ければ高いほど，主観的幸福度が高くなる」という傾向にはありそうです．

ただし，今はグラフを見ているだけなので，果たしてこの傾向が本当にあるのかどうかがわかりません．今度はこの傾向が科学的に認められるのかどうかを考えてみましょう．

回帰分析をやってみる.

さて，今度はRで分析してみましょう．ここでは，2行ほどのコードを書いてもらいます．

```
hapsat_model<-lm(SUB_HAP~SUB_SAT, data = exdataset)
summary(hapsat_model)
```

```
##
## Call:
## lm(formula = SUB_HAP ~ SUB_SAT, data = exdataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8918 -0.6503 -0.0814  0.7289  6.4015
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.59853    0.10176   15.71   <2e-16 ***
```

出力結果について説明しましょう.

```
## Call:
## lm(formula = SUB_HAP ~ SUB_SAT, data = exdataset)
```

この行では，分析したモデル式について示しています．簡単に言うと，「生活満足度によって，主観的幸福度は説明できるかどうか試してます．．．」ということを示しています.

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8918 -0.6503 -0.0814  0.7289  6.4015
```

ここでは，モデル式からのズレ(ϵ_i)である誤差がどの程度あるのかを示しています．ここでは誤差の最小値，第1四分位点，中央値，第3四分位点，最大値を示しています．一般線形モデルではこの誤差が正規分布になっていることを仮定しています.

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.59853    0.10176   15.71   <2e-16 ***
## SUB_SAT      0.81036    0.01711   47.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

- ここではその分析結果について示しています．第一に注目すべきはこの項目です．
- “Intercept”は切片を示しています．先程のモデル式でいうと，α にあたる部分です．
- 加えて，“SUB_SAT”は生活満足度です．先程のモデル式でいうと，β_1にあたる部分です．“Estimate”は推定値を示しています．“Intercept”と交わるところではαに入る具体的な数字を示しています．また，“SUB_SAT”と交わるところではβ_1に当てはまる数字が入ります．

したがって，この結果はモデル式で書くと，以下のように示すことが出来ます．

$$(SH) = 0.81036 \times (LS) + 1.59853 + \epsilon_i$$

このモデル式は生活満足度が1あがると，主観的幸福度が0.8106ポイント増加すること，そして生活満足度が0である人の主観的幸福度は1.59853であることが推定されています．

ここに出てくるt valueはt値を，Pr(>|t|)はp値を示しています．そして，最後のsign.if. codesでは，どのような基準で*をつけているかを説明しています．この場合，p値が1-0.1の場合は無印，0.1-0.05の場合は”.”，0.05-0.01の場合は”*”，0.01-0.001の場合は”**”，0.001-0の場合は”***”，としてつけている，ということが示されています．

統計学の基本的な考え方ではp値が0.05以下，すなわち5%以下である場合には対立仮説を採択することがお約束となっています．．．が，単純に5%以下であることによって対立仮説を採択することがあってはいけません．

それは以下の理由によります．

- 分野によって10%以上でも有意差を認めることがある．
- 統計的な有意性はデータの量にも依拠するため，単純に評価してよいかどうかは課題がある．
  - 心理学系だと「効果量」という議論がある．

```
## Multiple R-squared:  0.7002, Adjusted R-squared:  0.6999
## F-statistic:  2244 on 1 and 961 DF,  p-value: < 2.2e-16
```

続いて，確認したいのはこの2行です．“Multiple R-squared”はR2乗(あーるにじょう)値を示しています．ただし，このR2値は決定係数と呼ばれており，回帰式の当てはまり具合を示しています．寄与率とも呼ばれて，この値が1に近ければ近いほどよく説明できているモデル式であると言われます．ただし，R2乗値はこのモデルに組み込まれる説明変数が増えれば増えるほど，より良くなっていきます．そうするといくらでも興味のない変数を入れて重回帰分析（後日説明します）．．．．と，なると決して意味があるモデル式になるとは言えません．

そこで，たくさん変数を入れたことに対するペナルティを加えたのが”Adjusted R-squared”，調整済みR2乗値と呼ばれるものです．こちらを報告してあげると良いかと思います．

最後の”F-statistic”はF検定と呼ばれるものの結果です．2つの群の「標準偏差」が等しいかどうか，を示しているものであり，「等分散性の分析」に用いられているものです．この結果は，主観的幸福度と生活満足度では分散，すなわちばらつき方が異なっている，ということを示しています．

結果の表記例.

- 生活満足度1が改善すると，主観的幸福度が0.81改善することが，0.1%水準で示された．（一緒に表を見せると良い．）
- 生活満足度1が改善すると，主観的幸福度が0.81改善することが示された（t(961)=47.37, p=.001）.
- $(SH) = 0.81036(t = 47.37) \times (LS) + 1.59853 + \epsilon_i$

結果をきれいに表記しよう.

- パッケージpanderの中にある関数panderを使うと，結果がわかりやすく表示されます.

```
library(pander)
pander(hapsat_model)
```

- 私のはCSSをいじっているので少し色が変わっています.

- 他にもパッケージhuxtableの中にhuxregという関数があります.

```
library(huxtable)
huxreg(hapsat_model)
```

- パッケージstargazerの中にあるstargazerという関数を使うとxls形式で出力できます.

```
library(stargazer)
stargazer(hapsat_model, type = "html", align=TRUE,
          title = "   ", out = "hapsatmodel.xls")
```

- 作業フォルダの中に”hapsatmodel.xls”というファイルができていますので，そちらを開いてください.
  - 開く際に注意画面が出てきますが，「気にせずに開く」を選んでください.

t値とは？

$$t-value = \\ (Expected\,Value)-(Average)/(Standard\,Deviation)$$

t値はこんな数式から算出されます．

標準誤差は(標準偏差)/(データ数の平方根)によって計算できることを思い出しておいて下さい． t値は分子が大きければ，平均値との差が大きいことを示しており，分母が大きければ，標準偏差（分散）が小さく，データ数が十分にあることを示しています．このt値が大きければ大きいほど，帰無仮説を棄却して対立仮説を採択できることを示しています．

一方，p値は帰無仮説が成立していることを前提として，0.05，すなわち5%未満であれば，帰無仮説を棄却するための基準となります．実際に確率的に示すことによって，得られた差異がどの程度珍しいのか，ということを示しています．例えば，p値が0.03，すなわち3%であれば，帰無仮説が正しいとした時に今得られた結果は3%でしか観察できないような珍しいことが起こっていることを示しています．こんなに珍しいことが起こったのは，その帰無仮説が正しくないからであり対立仮説を選ぼう！という論理のもとに対立仮説を採択することになります．

ここでは，t値とp値の計算方法については別書に譲ることとして，ざっくりとした理解で先に行きましょう．

ダミー回帰分析とt検定

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．

平均値をプロットする

はじめに，分析対象となるデータを読み込んでおきましょう．

- `ggplotgui`を使ったプロットの方法についても紹介したいと思います．

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

そうすると新しいウィンドウが開きます.

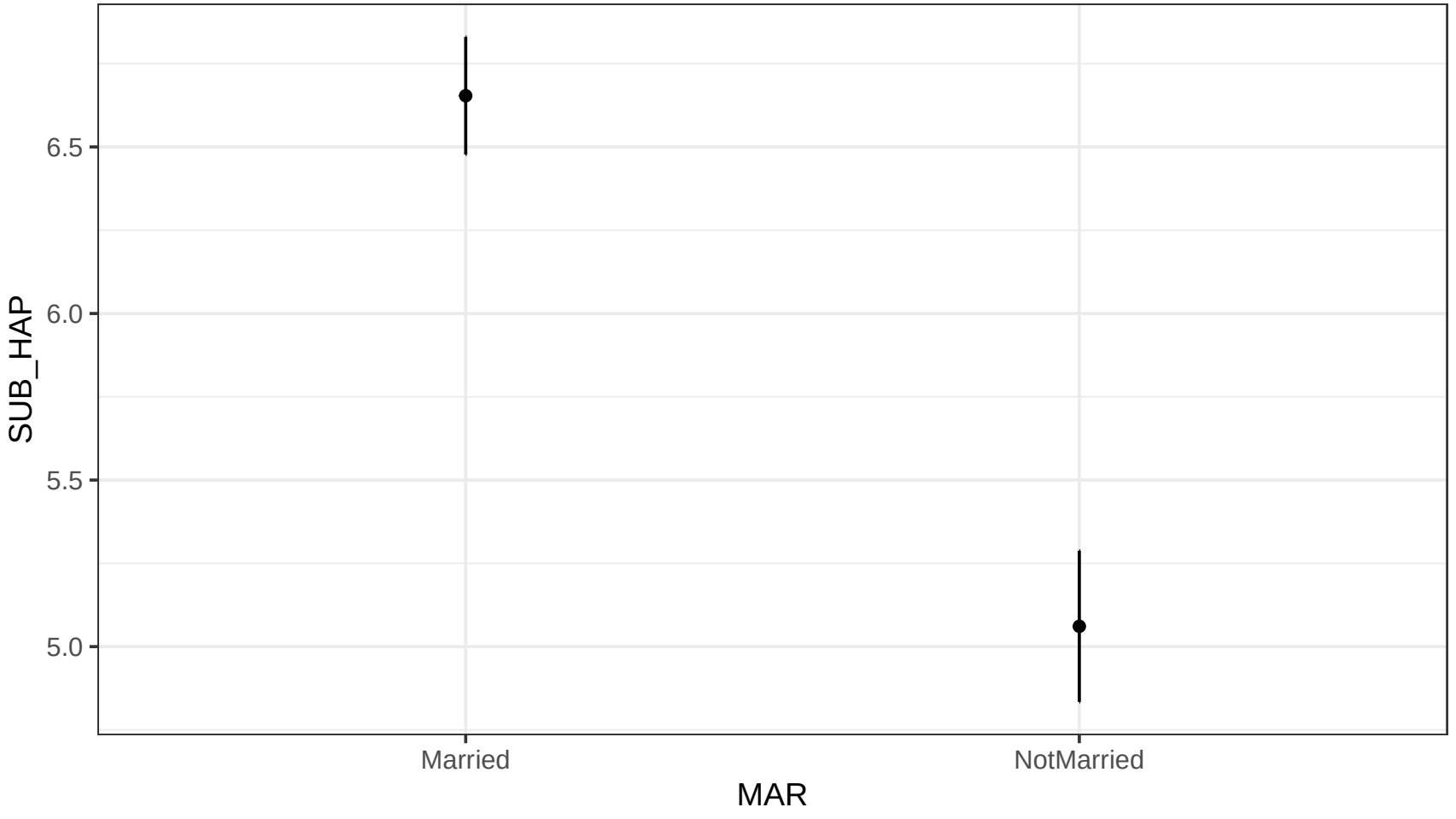
以下の通りの作業をしましょう.

- ggplotタブへ
- “Type of graph:”は”Dot + Error”，Y-variableは”SUB_HAP”，X-variableは”MAR”を設定
- “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dataset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)) +
  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
```



```
# 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)
```

ダミー回帰をやってみる

- 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 ***
## ---
```

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
```

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）．
- $(SH) = 1.59291(t = 10.99) \times (MAR\_dum) + 5.0609 + \epsilon_i$

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 betwee
## 95 percent confidence interval:
##  1.30479 1.88094
## sample estimates:
##    mean in group Married mean in group NotMarried
##                 6.653779                 5.060914
```

参考：スチューデントのt検定

```
student_t.testmodel<-t.test(SUB_HAP ~ MAR,
                            data = exdataset,
                            var.equal = T)
student_t.testmodel
```

```
##
##  Two Sample t-test
##
## data:  SUB_HAP by MAR
## t = 10.986, df = 961, p-value < 2.2e-16
## alternative hypothesis: true difference in means betwee
## 95 percent confidence interval:
##  1.308323 1.877406
## sample estimates:
##    mean in group Married mean in group NotMarried
##                 6.653779                 5.060914
```

ちなみに，スチューデントのt検定と一般線形モデルにおけるダミー回帰モデルは結果が一致します．これは一般線形モデルが等分散性を仮定していることによります．

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

$$Y_i = \alpha + \epsilon_i$$

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．

$$(SH) = \beta_1(Hokkaido\_dum) + \beta_2(Tohoku\_dum)+$$

$$\beta_3(Chubu\_dum)+\beta_4(Kinki\_dum)+\beta_5(Chugoku\_dum)+$$

$$\beta_6(Shikoku\_dum) + \beta_7(Kyushu\_dum) + \alpha + \epsilon_i$$

- 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．

平均値をプロットする

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()
```

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.

- “**Data upload**”をクリック
- datasetをコピーする
- “**Paste Data**”にペーストをする
- ggplotタブへ
- “**Type of graph:**”は”**Dot + Error**”，Y-variable は”**SUB_HAP**”，X-variableは”**ARE**”を設定
- “**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.

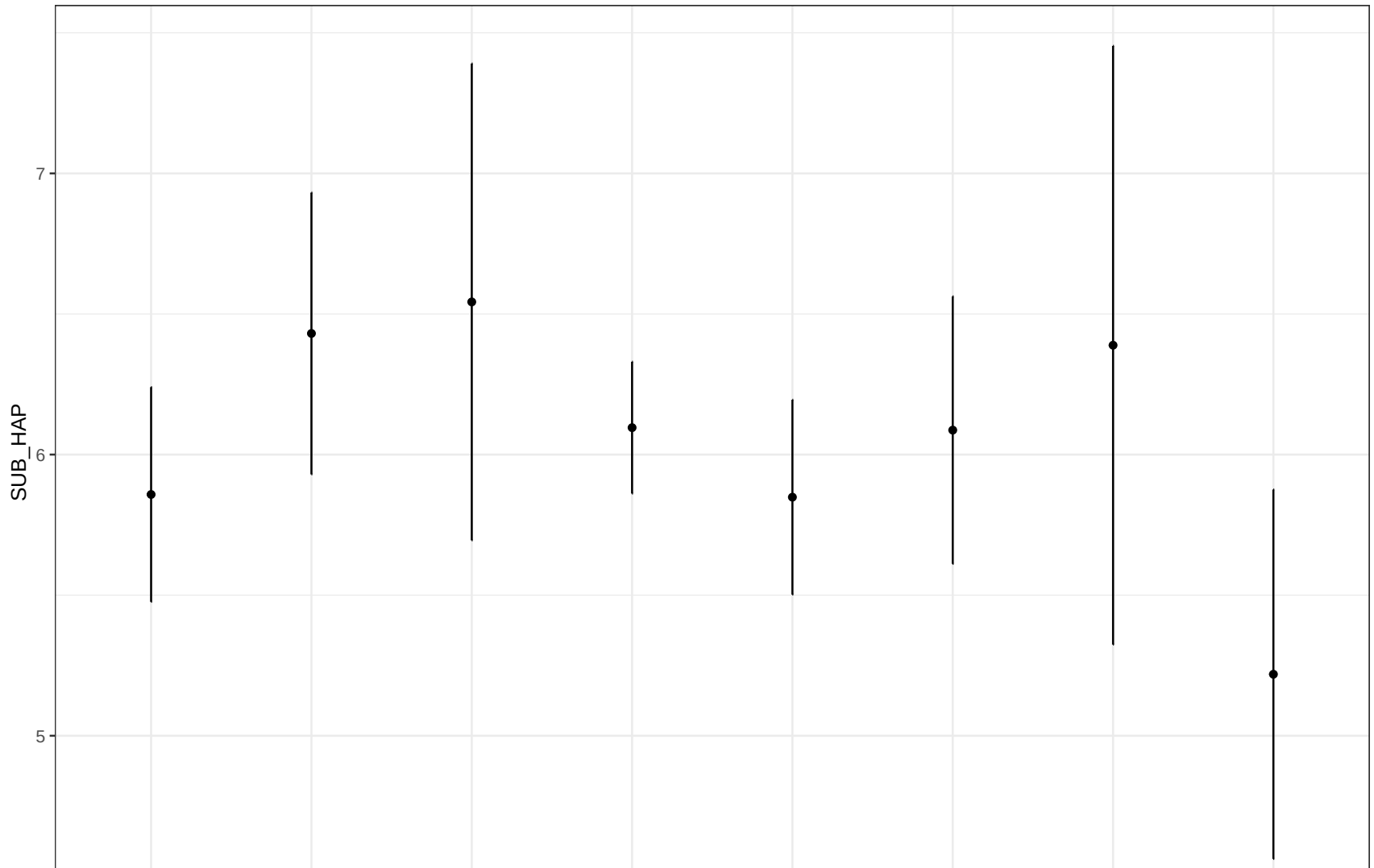
- こんな感じのコードができます.

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

# The code below will generate the graph:
graph <- ggplot(exdataset, aes(x = ARE, y = SUB_HAP)) +
  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()
```

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"
arehap_model<-lm(SUB_HAP ~ ARE, data = exdataset)
```

```
#
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
```

- 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.

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．

$$(SH) = 0.447112*(Hokkaido\_dum) - 0.876995*(Tohoku\_d$$

$$0.237637 * (Chubu\_dum) - 0.247260 * (Kinki\_dum) +$$

$$0.335025*(Chugoku\_dum) + 0.293144*(Shikoku\_dum) -$$

$$0.008788 * (Kyushu\_dum) + 6.095745 + \epsilon_i$$

- 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

```
#
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
```

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 ‘
```

- 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．

結果をきれいに表記しよう.

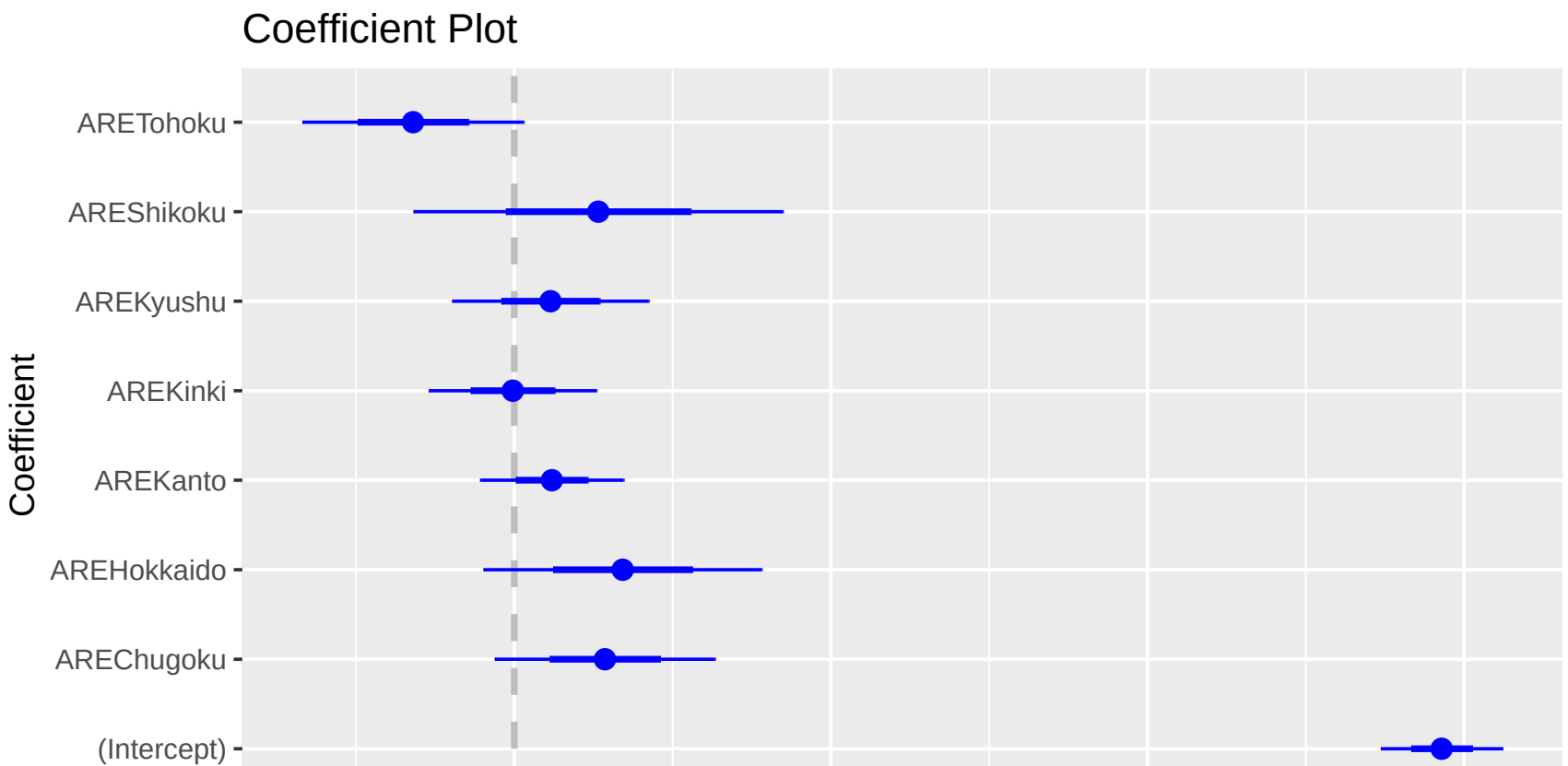
- パッケージhuxtableの中にhuxregという関数があります.

```
library(huxtable)
huxreg(arehap_model)
```

結果をきれいに表記しよう.

- パッケージcoefplotを使って各係数の大きさをグラフで示す.

```
library(coefplot)
coefplot(arehap_model)
```



結果をきれいに表記しよう.

- パッケージstargazerの中にあるstargazerという関数を使うとxls形式で出力できます.

```
library(stargazer)
stargazer(arehap_model, type = "html", align=TRUE, title =
  ", out = "arehap_model.xls")
```

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.

```
   <-lm(     ~     ,
                data =       )
       t
summary(    )

anova(    )
```

演習問題

演習問題1

“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，地域を表す”ARE”を説明変数として，以下の2つの分析を実施せよ．

- 生活満足度の地域差を分析せよ．
- 睡眠満足度の地域差を分析せよ．

演習問題2

“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，年代を表す”GEN”を説明変数として，以下の2つの分析を実施せよ．

- 主観的幸福度の年代差を分析せよ．
- 生活満足度の年代差を分析せよ．
- 睡眠満足度の年代差を分析せよ．

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

- $X_1 = 1$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

- $X_1 = 1$と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グッと幸せ

- $X_1 = 0$と$X_2 = 0$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．

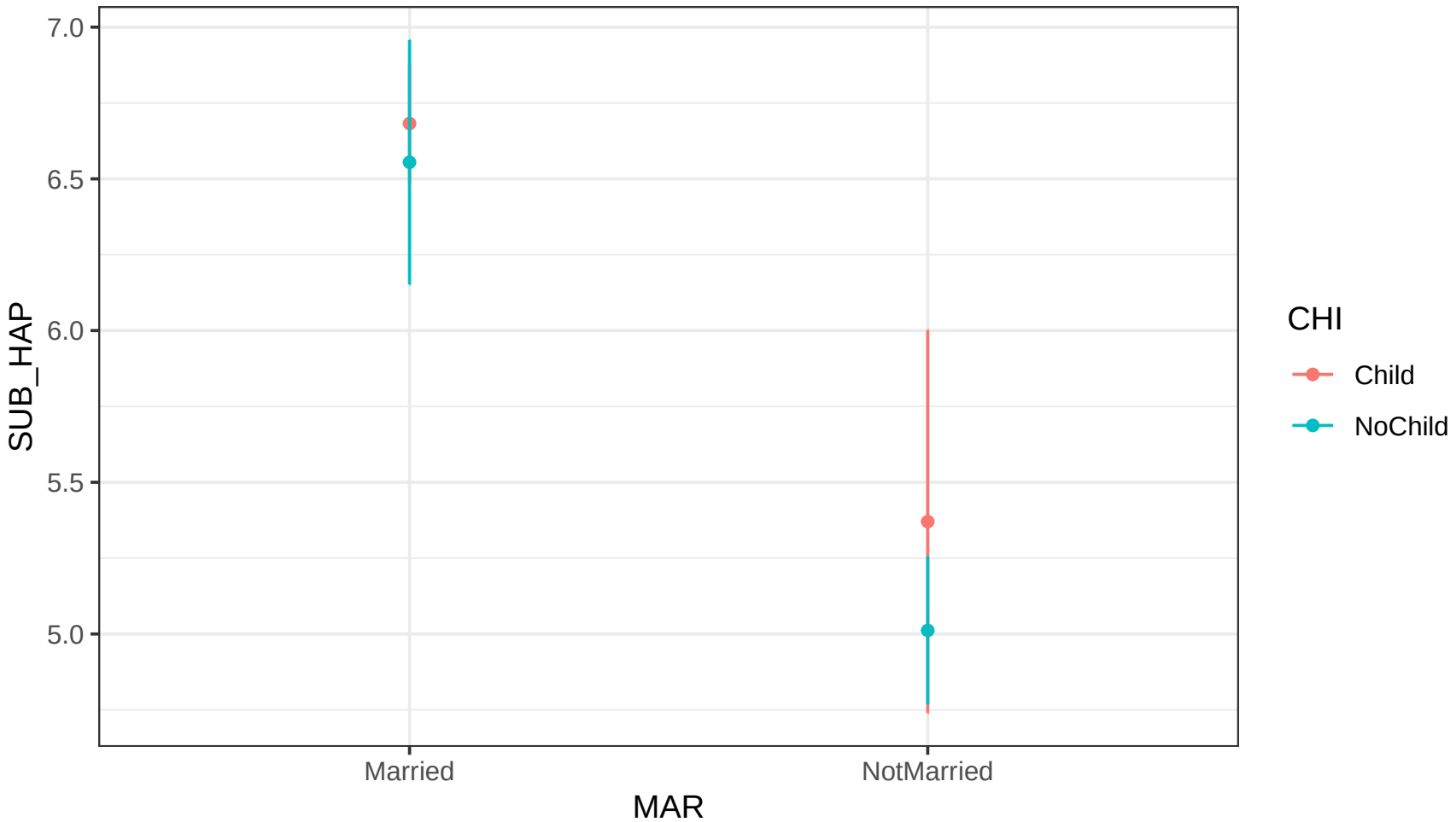
以下の通りの作業をしましょう.

- 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, colou
  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
```


7.0
6.5
6.0
5.5
5.0
SUB_HAP
Married
NotMarried
MAR
CHI
Child
NoChild

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)
#   MARCHIHAP_model
```

```
#
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  <
## MARNotMarried             -1.3122     0.3190  -4.113 4.2
## CHINoChild                -0.1279     0.2222  -0.575
## MARNotMarried:CHINoChild  -0.2308     0.3930  -0.587
```

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.

$$(SH) = 1.543(Married\_dum) + 0.359(Child\_dum) -$$
$$0.231(Married\_dum * Child\_dum) + 5.012$$

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」

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

$$(SH) = 5.012$$

- 「未婚者かつ子あり」

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

$$(SH) = 0.359 + 5.012 = 5.371$$

- 「既婚者かつ子なし」

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

$$(SH) = 1.543 + 5.012 = 6.555$$

- 「既婚者かつ子あり」

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

$$(SH) = 1.543 + 0.359 - 0.231 + 5.012 = 6.683$$

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)