

# 統計学第10/11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

$\chi^2$ 検定をしましょう

一般線形モデルとは

回帰分析

演習問題

ダミー回帰分析とt検定

## 前回の復習

## クロス集計：

- ▶ **クロス集計表**：複数の質問項目を組み合わせて集計する方法
  - ex. 朝食を食べているか否か×深夜アルバイトしているか否かなど.
  - 企業の中でも基本的な統計手法としてよく用いられている.
  - 2つの質的変数間の関連性である「連関」を示す.

## 組み合わせの数をカウントする.

- ▶ 使用するパッケージ

```
# install.packages("dplyr")  
library(dplyr)
```

```
##  
##           : 'dplyr'  
  
##           'package:stats'           :  
##  
##           filter, lag  
  
##           'package:base'           :  
##  
##           intersect, setdiff, setequal, union
```

```
# install.packages("tidyr")  
library(tidyr)
```

## 地域ごとに子どもがいる人の数を数える.

### ▶ 2つの手法

- dplyrのgroup\_by関数を使う方法
- table 関数を使う方法
- 現在では前者がメインの手法だが, 念のために後者の方法についても紹介する.
- 今の御時世の最先端の関数を使っている

## dplyrのgroup\_by関数を使う方法

```
tablea <- exdataset %>%
  group_by(ARE, CHI) %>% #
  tally() %>% #
  spread(ARE, n) #
tablea
```

```
## # A tibble: 2 x 9
##   CHI      Chubu Chugoku Hokkaido Kanto Kinki Kyushu Shikoku
##   <chr>    <int>    <int>    <int> <int> <int> <int> <int>
## 1 Child      80      34      21    184     86     53
## 2 NoChild    68      31      14    192     79     39
```

## dplyrのcount関数を使う方法

```
tableb <- exdataset %>%
  count(ARE, CHI) %>%
  spread(ARE, n)
tableb
```

```
## # A tibble: 2 x 9
##   CHI      Chubu Chugoku Hokkaido Kanto Kinki Kyushu Shikoku
##   <chr>   <int>   <int>   <int> <int> <int> <int> <int>
## 1 Child      80      34      21    184     86     53
## 2 NoChild    68      31      14    192     79     39
```



## 参考：ggplot2で可視化する方法

```
library(ggplot2)
exdataset%>%
  ggplot(aes(x=ARE, fill=CHI), stat="count")+geom_bar()
```



## xtabs 関数を使う方法

```
tablec<-xtabs(~ CHI, exdataset)
tablec
```

```
## CHI
##      Child NoChild
##      495      468
```

```
tabled<- xtabs(~ ARE, exdataset)
tabled
```

```
## ARE
##      Chubu  Chugoku Hokkaido      Kanto      Kinki      Kyushu S
##      148      65      35      376      165      92
```

## table 関数を使う方法

```
tablee<-xtabs(~ ARE + CHI, exdataset)
tablee
```

```
##           CHI
## ARE      Child NoChild
## Chubu      80      68
## Chugoku     34      31
## Hokkaido    21      14
## Kanto     184     192
## Kinki       86      79
## Kyushu      53      39
## Shikoku      8      10
## Tohoku      29      35
```

## ▶ 行のパーセント表示

```
tableg<-prop.table(tablee, 1)  
tableg
```

| ##          | CHI       |           |  |
|-------------|-----------|-----------|--|
| ## ARE      | Child     | NoChild   |  |
| ## Chubu    | 0.5405405 | 0.4594595 |  |
| ## Chugoku  | 0.5230769 | 0.4769231 |  |
| ## Hokkaido | 0.6000000 | 0.4000000 |  |
| ## Kanto    | 0.4893617 | 0.5106383 |  |
| ## Kinki    | 0.5212121 | 0.4787879 |  |
| ## Kyushu   | 0.5760870 | 0.4239130 |  |
| ## Shikoku  | 0.4444444 | 0.5555556 |  |
| ## Tohoku   | 0.4531250 | 0.5468750 |  |

## ▶ 列のパーセント表示

```
tableh <- prop.table(tablee, 2)
tableh
```

| ## |          | CHI        |            |
|----|----------|------------|------------|
| ## | ARE      | Child      | NoChild    |
| ## | Chubu    | 0.16161616 | 0.14529915 |
| ## | Chugoku  | 0.06868687 | 0.06623932 |
| ## | Hokkaido | 0.04242424 | 0.02991453 |
| ## | Kanto    | 0.37171717 | 0.41025641 |
| ## | Kinki    | 0.17373737 | 0.16880342 |
| ## | Kyushu   | 0.10707071 | 0.08333333 |
| ## | Shikoku  | 0.01616162 | 0.02136752 |
| ## | Tohoku   | 0.05858586 | 0.07478632 |

## もっと細かいクロス集計表を出してみよう

```
xtabs(~ CHI + ARE +F_SEX, exdataset)
```

```
## , , F_SEX = female
```

```
##
```

```
##          ARE
```

```
## CHI          Chubu Chugoku Hokkaido Kanto Kinki Kyushu Shikoku
```

```
##   Child         49        24         11    117     56        33
```

```
##   NoChild       30        17          7    103     39        22
```

```
##
```

```
## , , F_SEX = male
```

```
##
```

```
##          ARE
```

```
## CHI          Chubu Chugoku Hokkaido Kanto Kinki Kyushu Shikoku
```

```
##   Child         31         10         10     65     30        20
```

```
##   NoChild       37         14          7     88     39        16
```

```
##
```

```
##          F_SEX
```

## 連関係数を出力しよう

- ▶ 連関係数：クラメール連関係数V
  - 下限が0, 上限が1で完全な連関に近づくにつれて1に近い値を取る.

```
#install.packages('vcd', dependencies = T)  
library(vcd)
```

```
##          grid
```

```
assocstats(tablee)
```

```
##                X^2 df P(> X^2)
## Likelihood Ratio 5.1570  7  0.64082
## Pearson          5.1408  7  0.64278
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.073
## Cramer's V        : 0.073
```

- ▶ 今回の場合は、地域と子供の有無にほとんど連関は認められませんでした。



$\chi^2$  検定をしましょう

## 2種類の $\chi^2$ 検定

- ▶  $\chi^2$  検定はその目的に応じて2種類ある
- ▶ 適合度検定：観測度数が理論比率にもとづいて得られるかどうかを検証する仮説検定
- ▶ 独立性検定：複数の特性の間に関連があるかどうかを調べる仮説検定

## 適合度検定：

- ▶ 普通のサイコロを振ったときに、各目が等しい確率で出る.
- ▶ あるサイコロを振ったとき、以下のような結果が得られた. このサイコロは「普通のサイコロ」であろうか？それとも、「普通ではないサイコロ」ではないだろうか？
- ▶ 1:40
- ▶ 2:21
- ▶ 3:40
- ▶ 4:90
- ▶ 5:50
- ▶ 6:70

## 適合度検定

- ▶ 対立仮説：観測された頻度分布と期待される頻度分布に差がある.
- ▶ 帰無仮説：観測された頻度分布と期待される頻度分布に差があるとは言えない.

```
psy <- c(40, 21, 40, 90, 50, 70)
```

- ▶ サイコロの出た目

```
the_psy <- c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
```

- ▶ サイコロの目の理論値

## 適合度検定の実施

```
chisq.test(psy, p = the_psy)
```

```
##
```

```
## Chi-squared test for given probabilities
```

```
##
```

```
## data:  psy
```

```
## X-squared = 58.28, df = 5, p-value = 2.754e-11
```

- ▶ p値は有意水準を大きく下回るために帰無仮説を棄却し、対立仮説を採択する。
- ▶ このサイコロは「普通ではないサイコロ」である。

## 適合度検定の実施その2

- ▶ 普通のサイコロを振ったときに，各目が等しい確率で出る．
- ▶ 1-6の目が出るサイコロをシミュレーションで10000回振った．このサイコロは「普通のサイコロ」であろうか？それとも「普通ではないサイコロ」ではないだろうか？

```
roll <- function(){  
  die <- 1:6  
  a_die <- sample(x = die, size = 1 , replace = TRUE)  
}
```

- ▶ size=1に注意

## 適合度検定の実施その2

```
rolls <- replicate(10000, roll())
```

- ▶ サイコロを10000回振ります.

## 適合度検定の実施その2

```
rolls_table <- table(rolls)
rolls_table
```

```
## rolls
##      1      2      3      4      5      6
## 1669 1654 1648 1714 1667 1648
```

▶ 度数を出力しておこう.



## 適合度検定の実施その2

```
chisq.test(rolls_table, p = the_psy)
```

```
##  
## Chi-squared test for given probabilities  
##  
## data: rolls_table  
## X-squared = 1.862, df = 5, p-value = 0.8679
```

- ▶ 適合度検定を実施する.
  - 「大数の法則」により統計的には有意差は認められないはず.
  - 全てが同じ確率で発生する確率分布を「一様分布」という.

## 独立性の検定

- ▶ 性別と旅行の好みについて、以下のクロス表が得られた場合の変数AおよびBの独立性の検定を行う。

|    | 旅行好き | どちらともいえない | 旅行嫌い |
|----|------|-----------|------|
| 男性 | 70   | 50        | 60   |
| 女性 | 40   | 30        | 20   |

## 独立性検定

- ▶ 対立仮説：性別と旅行の好みに関連性がある
- ▶ 帰無仮説：性別と旅行の好みに関連性があるとは言えない  
(独立である)

## 独立性検定

```
ryoko_seibetsu <- matrix(c(70, 50, 60, 40, 30, 20),  
                           nrow = 2, byrow = T)
```

- ▶ 行列をオブジェクトにしまう.

## 独立性検定

```
chisq.test(ryoko_seibetsu)
```

```
##
```

```
##  Pearson's Chi-squared test
```

```
##
```

```
## data:  ryoko_seibetsu
```

```
## X-squared = 3.5795, df = 2, p-value = 0.167
```

- ▶ p値は有意水準より大きいために帰無仮説を採択する.
- ▶ 性別と旅行の好みに関連性があるとは言えない（独立である）

## $\chi^2$ 検定

- ▶ 対立仮説：居住地域と子供の有無は独立ではない（連関がある）
- ▶ 帰無仮説：居住地域と子供の有無は独立である（連関があるとは言えない）

```
chitest.tablee<-chisq.test(tablee)
chitest.tablee
```

```
##
##  Pearson's Chi-squared test
##
## data:  tablee
## X-squared = 5.1408, df = 7, p-value = 0.6428
```

- ▶ 検定の結果，p値が.05以上なので，対立仮説を採択できず，帰無仮説を採択する．

## $\chi^2$ 検定

- ▶ レポートにまとめる時には，こんな書き方をします.  
 $\chi^2$  検定を行った結果，居住地域と子供の有無は独立であることがわかった( $\chi=5.1408$ ,  $df=7$ ,  $p=.64$ ).

▶ もし、 $\chi^2$  検定でp値が.05以下であった場合、残差分析を行います。

- どのセルで有意な逸脱が生じたのかを検討する。
- 標準化残差が1.96以上であれば、5%水準で有意な逸脱があったと評価する。

```
chitest.tablee$stdres
```

```
##          CHI
## ARE          Child    NoChild
## Chubu    0.7017295 -0.7017295
## Chugoku   0.1513129 -0.1513129
## Hokkaido  1.0367594 -1.0367594
## Kanto     -1.2252616  1.2252616
## Kinki      0.2030909 -0.2030909
## Kyushu     1.2524729 -1.2524729
## Shikoku   -0.5961874  0.5961874
## Tohoku    -1.0087797  1.0087797
```



▶ もしくは、以下の計算でp値を算出しても良い.

```
pnorm(abs(chitest.tablee$stdres), lower.tail = FALSE) * 2
```

```
##          CHI
## ARE          Child    NoChild
## Chubu    0.4828479 0.4828479
## Chugoku   0.8797289 0.8797289
## Hokkaido  0.2998480 0.2998480
## Kanto     0.2204767 0.2204767
## Kinki     0.8390640 0.8390640
## Kyushu    0.2103976 0.2103976
## Shikoku   0.5510501 0.5510501
## Tohoku    0.3130803 0.3130803
```

# 一般線形モデルとは

## 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．

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

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できます．

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

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、 $\beta$  を傾き、 $\alpha$  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できます。

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しています。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、“ $\beta_1$ ”は係数、“ $\alpha$ ”は切片と呼ばれます。そして、“ $\epsilon_i$ ”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 $\alpha$ 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

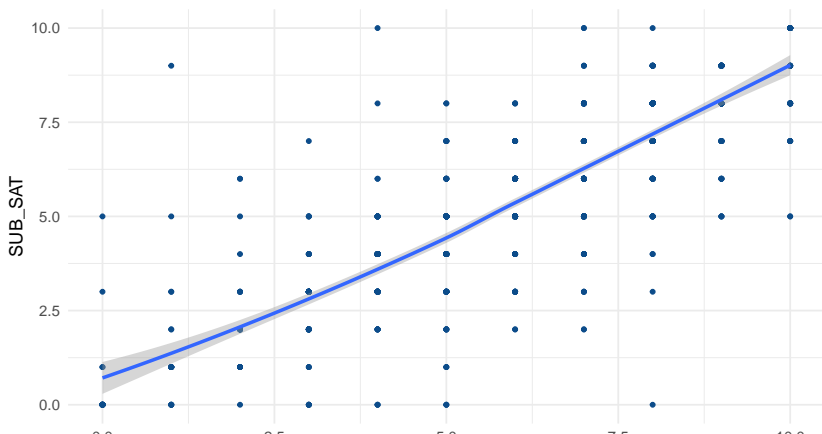
はじめに、分析対象となるデータを読み込んでおきましょう。  
\* もちろん、既に読み込んである場合は飛ばしてもらって構いません。

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, color = "blue") +  
  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分かりません。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( $\epsilon_i$ )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、 $\alpha$ 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、 $\beta_1$ にあたる部分です。“Estimate”は推定値を示しています。“Intercept”と交わる場所では $\alpha$ に入る具体的な数字を示しています。また、“SUB\_SAT”と交わる場所では $\beta_1$ 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？

$$t - value = \frac{(Expected Value) - (Average)}{(Standard Deviation)}$$

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。

## 演習問題

## 問題

- ▶ “SUB\_SLP”は睡眠満足度として、以下の質問項目を尋ねたものである。  
これについて、以下の2つの分析を実施してください。
  - 主観的幸福度を応答変数、睡眠満足度を説明変数とした回帰分析を行ってください。
  - 生活満足度を応答変数、睡眠満足度を説明変数とした回帰分析を行ってください。

## 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 $X_1$ に入り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 $\beta_1$ 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に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、ということの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.

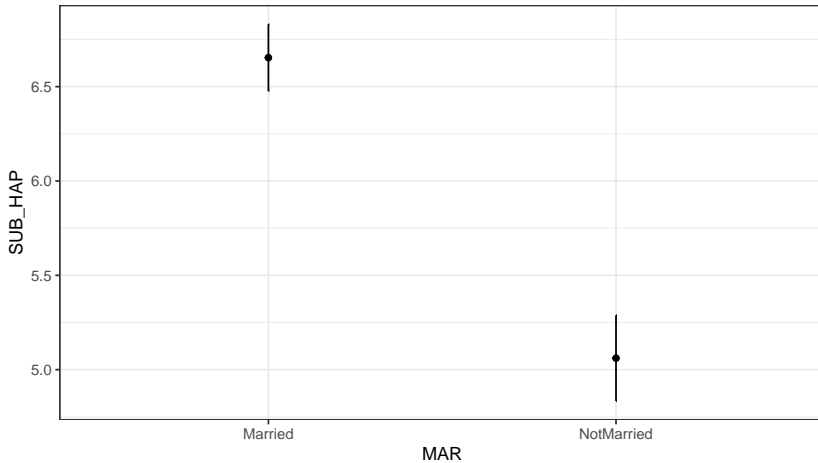
*# You need the following package(s):*

```
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にです。
- ▶ 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 $\alpha$  が5.0609,  $\beta$  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(marhap_model, type = "html", align=TRUE, title =
  "",
out = "marhap_model.xls")
```

- ▶ 作業フォルダの中に"marhap\_model.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.
  - 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.

## 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.

- ▶ 最近には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, という指摘もあつたりします.
- ▶ 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 between
```

```
## 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 between
```

```
## 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．



## 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 ~ 説明変数, 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

## 演習問題

## 演習問題

“CHI”は子どもの有無を尋ねる項目である。

これについて，以下の3つの分析を実施せよ．また，それぞれについてグラフも作成せよ．

- ▶ 子の有無による主観的幸福度の差を分析せよ．
- ▶ 子の有無による生活満足度の差を分析せよ．
- ▶ 子の有無による睡眠満足度の差を分析せよ．