統計学第 11 講/第 12 講

後藤　晶　　akiragoto@meiji.ac.jp

目次

目次

前回の復習

概要

相関係数とは，数値データ同士の関連性を探る指標です．相関係数の絶対値が 0 に近いと 2 つの変数同士には線形関係がないことを示します．

- $|r|=1.00$：完全に相関がある
- $0.70<|r|<1.00$：高い相関がある
- $0.40<|r|<0.70$：中程度の相関がある
- $0.20<|r|<0.40$：低い相関がある
- $0.00<|r|<0.20$：ほとんど相関がない
- $|r|=0.00$：完全に無相関である．

概要

ちなみに，この「相関の強さ」について分野によって評価が異なります．例えば，社会科学研究では高い相関が認められることは少ないです．今回の基準で中程度の相関や低い相関で議論をすることもあります．

この辺は分野によって異なりますので，ご承知おきください．

- 次のスライドからは同じ記述統計量の散布図を見てもらって，相関係数を確認することの重要性を感じてもらいます．

相関係数で比較をしてみる．

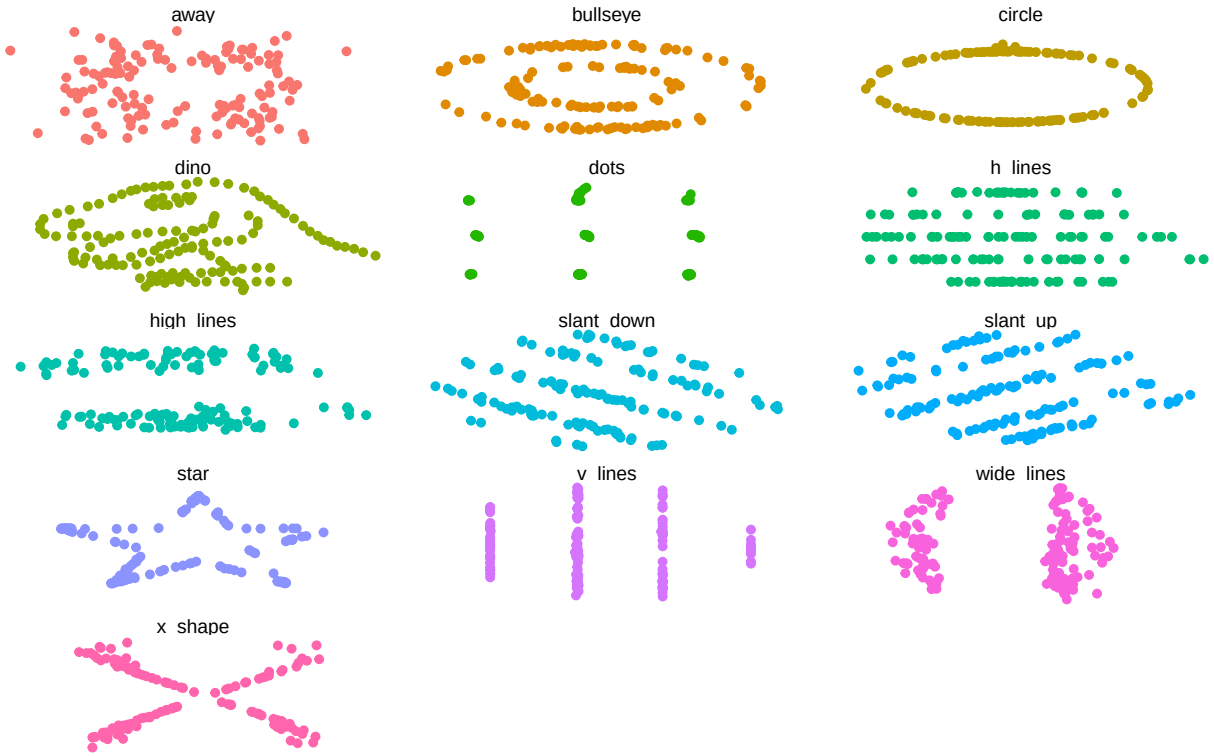
dataset	平均値	標準偏差	標本数	標準誤差
away	54.27	16.77	142	1.407
bullseye	54.27	16.77	142	1.407
circle	54.27	16.76	142	1.406
dino	54.26	16.77	142	1.407
dots	54.26	16.77	142	1.407
h_lines	54.26	16.77	142	1.407
high_lines	54.27	16.77	142	1.407
slant_down	54.27	16.77	142	1.407
slant_up	54.27	16.77	142	1.407
star	54.27	16.77	142	1.407
v_lines	54.27	16.77	142	1.407
wide_lines	54.27	16.77	142	1.407
x_shape	54.26	16.77	142	1.407

相関係数で比較をしてみる．

```
datasaurus<-datasaurus_dozen %>%
             ggplot(aes(x=x, y=y, colour=dataset))+
               geom_point()+
               theme_void()+
               theme(legend.position = "none")+
               facet_wrap(~dataset, ncol=3)
```

相関係数で比較をしてみる.

```
datasaurus
```



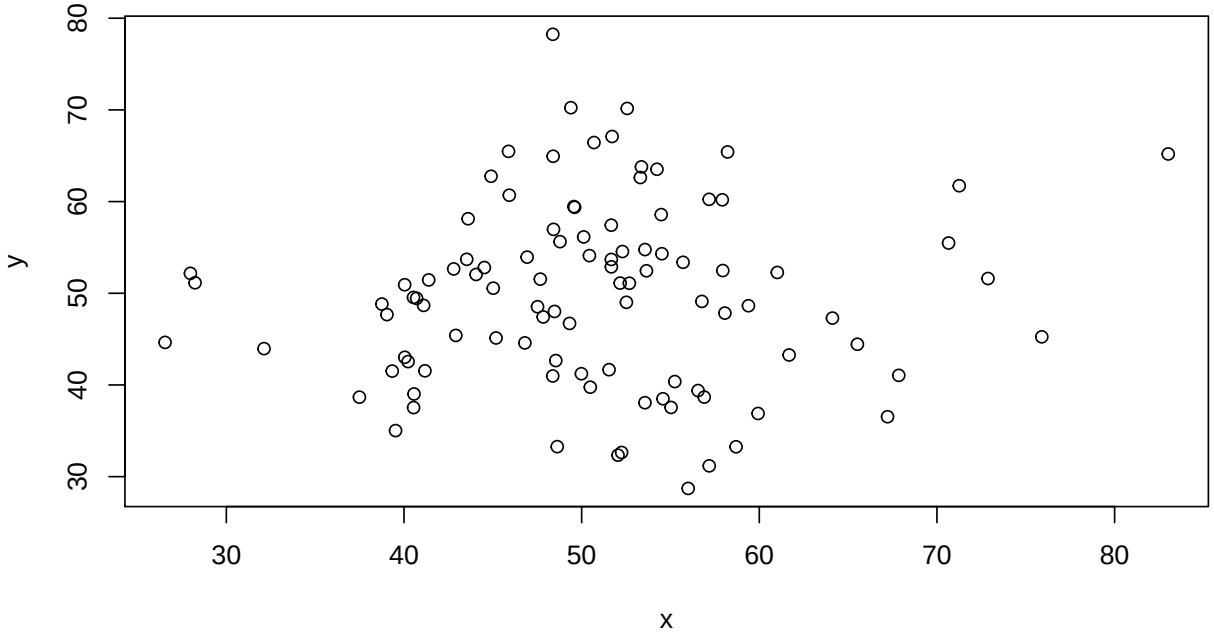
相関係数を出してみる

- 乱数で比較をしてみましょう.
  - 平均 50，標準偏差 10 のデータを 100 個 ×2 を作ります.
  - さらに，x と y を足して 2 で割ります.

```
x <- rnorm(100, 50, 10)
y <- rnorm(100, 50, 10)
z <- (x+y) / 2
```

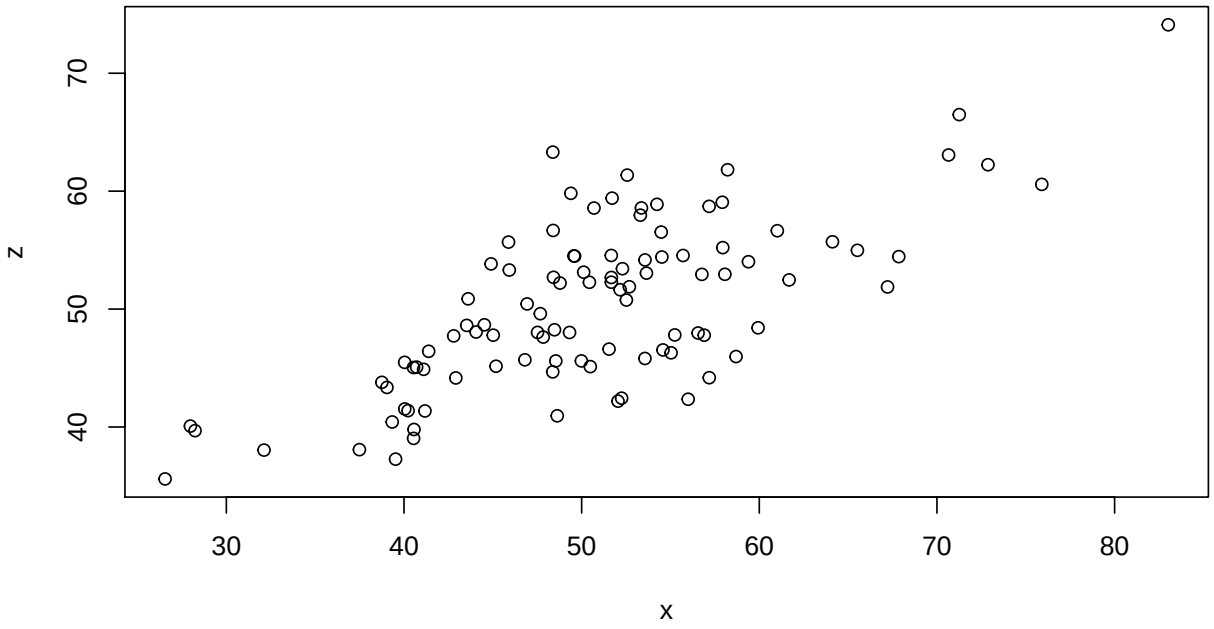
相関係数を出してみる

```
plot(x, y)
```



相関係数を出してみる

```
plot(x, z)
```



（不偏）共分散

- 2 種類のデータの関係を示す指標であり，2 つの変数の偏差の積の平均を計算する．
- 共分散が大きいほど関係性が強い,,, と言えるが，ちょっと不安がある．

- 「2 つの変数の関係の強さ」と「単位」の影響を受けてしまうため，標準偏差の積で割ってあげる必要がある．
- N-1 で割ると不偏共分散，N で割ると標本共分散

$$s_{xy} = \Sigma((x\ \ ) * (y\ \ ))/(\ \ -1)$$

不偏共分散を算出する

```
x_hensa <- x-mean(x)
y_hensa <- y-mean(y)
goukeixy <- sum(x_hensa * y_hensa)
kyobunsanxy <- goukeixy/(length(x)-1)
kyobunsanxy
```

```
## [1] 6.232461
```

- 演習問題
  - x と z について，不偏共分散を算出してみよう．

関数で不偏共分散を求める

```
cov(x, y)
```

```
## [1] 6.232461
```

```
cov(x, z)
```

```
## [1] 49.85821
```

相関係数を出してみる

- x と y の相関係数：

$$r = \frac{s_{xy}}{s_x s_y}$$

- s_{xy}：x と y の共分散
- s_x：x の標準偏差
- s_y：y の標準偏差

相関係数を算出する

```
soukanxy <- kyobunsanxy/(sd(x)*sd(y))
soukanxy
```

```
## [1] 0.06510957
```

- 演習問題
  - x と z について，相関係数を算出してみよう．

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．

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

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．

$$(\quad) = \beta_1(\quad) + \alpha + \epsilon_i$$

この時，切片である　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．

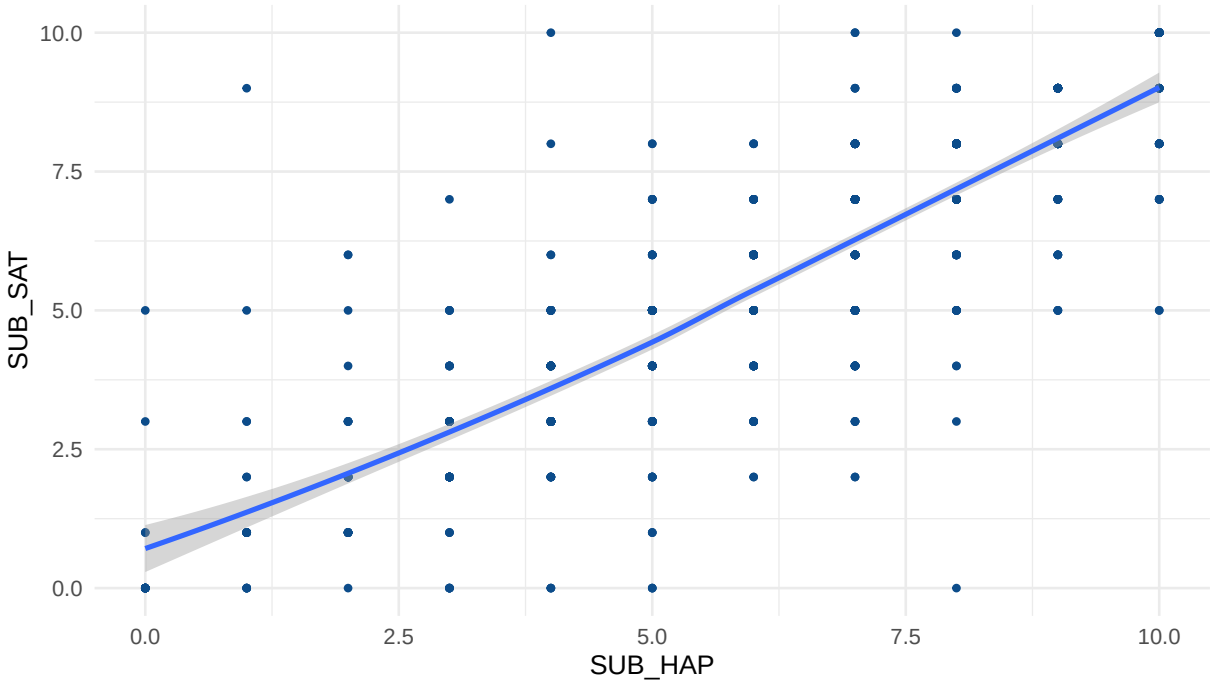
散布図のプロットは他の機能から持ってきてもよいのですが，今回は RStudio 上でクリックだけで入れられる方法を紹介します．

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．

- 以前紹介した “esquisse” を使います．動画をご確認ください．

```
library(ggplot2)

ggplot(exdataset) +
 aes(x = SUB_HAP, y = SUB_SAT) + geom_point(size = 1L, colour = "#0c4c8a") +
 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 ***
## SUB_SAT      0.81036    0.01711   47.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.285 on 961 degrees of freedom
## Multiple R-squared:  0.7002, Adjusted R-squared:  0.6999
## F-statistic:  2244 on 1 and 961 DF,  p-value: < 2.2e-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 ' ' 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．

$$(\quad) = 0.81036 \times (\quad) + 1.59853 + \epsilon_i$$

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）.
- $$(\quad) = 0.81036(t = 47.37) \times (\quad) + 1.59853 + \epsilon_i$$

結果をきれいに表記しよう.

- 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)
```

	(1)
(Intercept)	1.599 ***
	(0.102)
SUB_SAT	0.810 ***
	(0.017)
N	963
R2	0.700
logLik	-1607.061
AIC	3220.121

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

- 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 = "分析結果", out = "hapsatmodel.xls")
```

```
##
## <table style="text-align:center"><caption><strong>分析結果</strong></caption>
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:le
## <tr><td></td><td colspan="1" style="border-bottom: 1px solid black"></td></tr>
## <tr><td style="text-align:left"></td><td>SUB_HAP</td></tr>
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:le
## <tr><td style="text-align:left"></td><td>(0.017)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">Constant</td><td>1.599<sup>***</sup></td></tr>
```

```
## <tr><td style="text-align:left"></td><td>(0.102)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:le
## <tr><td style="text-align:left">R<sup>2</sup></td><td>0.700</td></tr>
## <tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>0.700</td></tr>
## <tr><td style="text-align:left">Residual Std. Error</td><td>1.285 (df = 961)</td></tr>
## <tr><td style="text-align:left">F Statistic</td><td>2,244.149<sup>***</sup> (df = 1; 961)</td></t
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:le
## </table>
```

- 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？

$$t = (\quad) - (\quad)/(\quad)$$

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．

- 以前紹介した “esquisse” を使っていただいて構いません．動画をご確認ください．
- また，別の方法として `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．

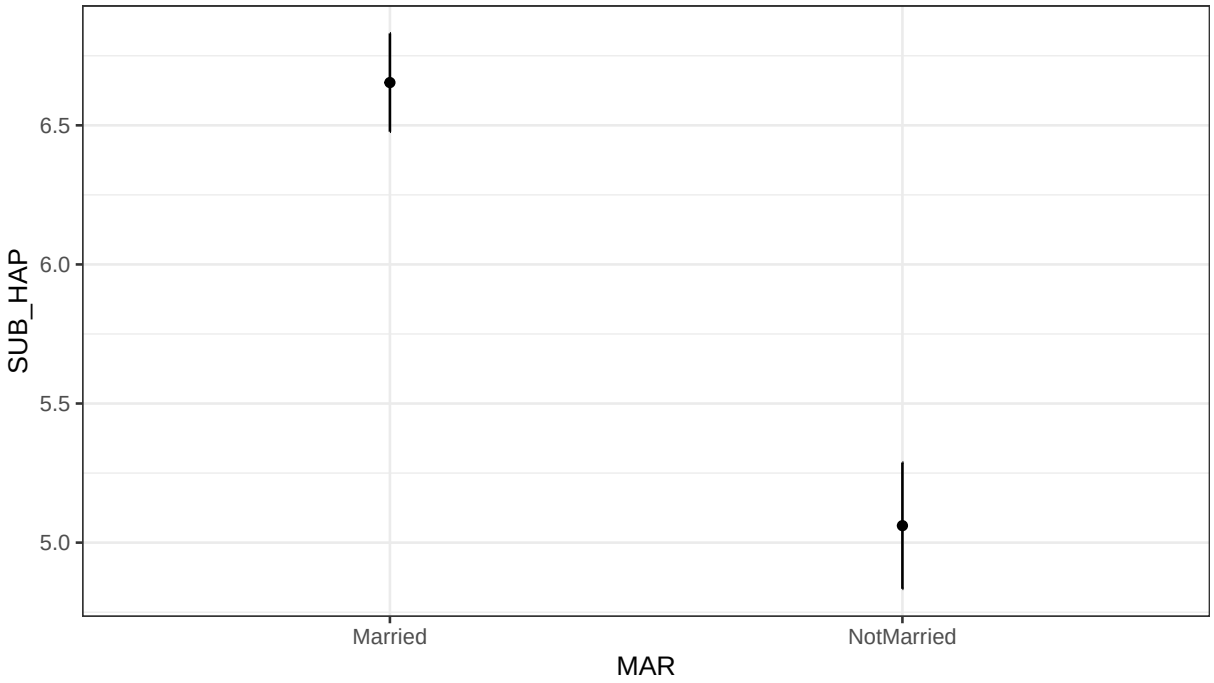
以下の通りの作業をしましょう．

- 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．

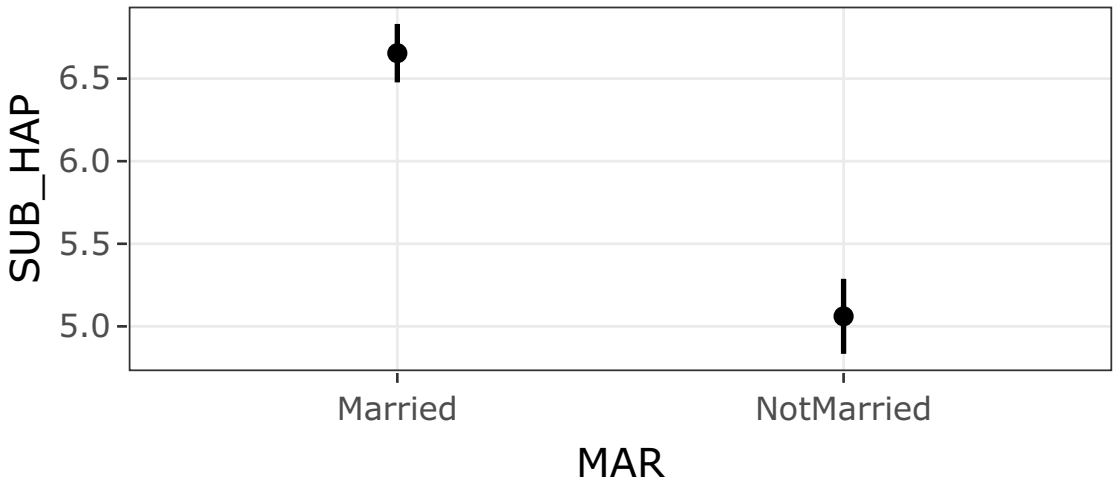
```
# You need the following package(s):
library("ggplot2")

# 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)
```

```
#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．

結果をきれいに表記しよう．

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

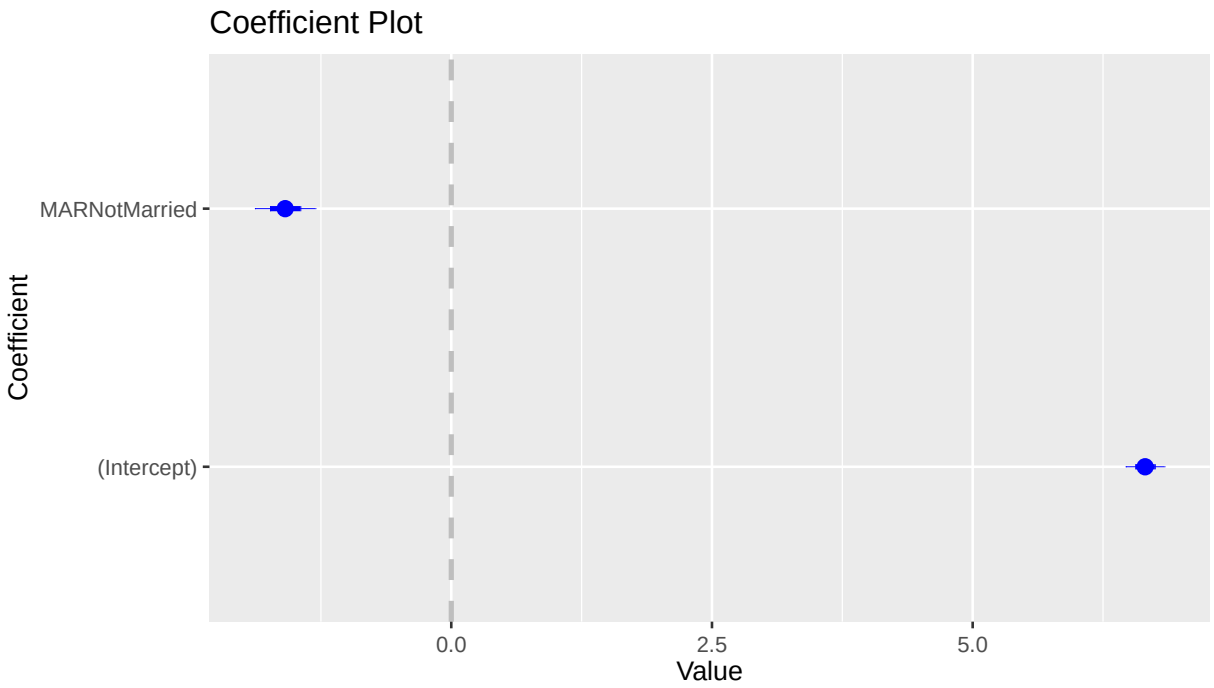
```
library(huxtable)
huxreg(marhap_model)
```

- パッケージ coefplot を使って，各係数の大きさをグラフで示しておこう．
  - 特にこれから重回帰分析などを学ぶ上で知っておくと便利です．

	(1)
(Intercept)	6.654 ***
	(0.093)
MARNotMarried	-1.593 ***
	(0.145)
N	963
R2	0.112
logLik	-2130.084
AIC	4266.168

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

```
# 初めて使う時は install.packages("coefplot") が必要です.
library(coefplot)
coefplot(marhap_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(marhap_model, type = "html", align=TRUE, title = "分析結果",
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)．
- $(\quad) = 1.59291(t = 10.99) \times (\ )$
  $+ 5.0609 + \epsilon_i$

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
```

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 is not equal to 0
## 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.

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．

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

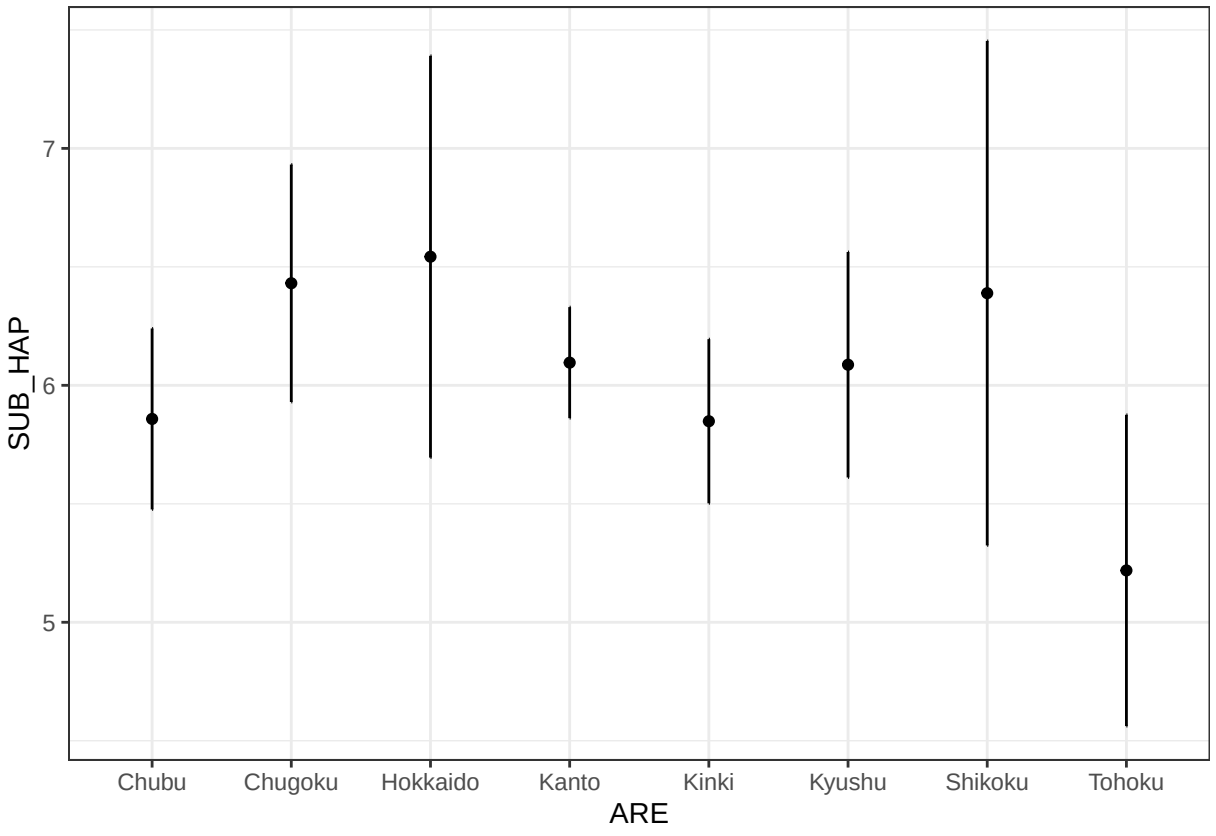
- “**Type of graph:**” は “**Dot + Error**”，Y-variable は “**SUB_HAP**”，X-variable は “**ARE**” を設定
- “**Confidence Interval:**” を 95% にする．
- R-code タブへ行って，以下のコードのうち，真ん中のみを以下にする．-また，コード内の **df** を **exdataset** に変える．
- こんな感じのコードができます．

```
# 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"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

```r
# 分散分析表
anova(arehap_model)
```

Df	Sum Sq	Mean Sq	F value	Pr(>F)
7	75.1	10.7	1.96	0.0573
955	5.22e+03	5.46		

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．

結果をきれいに表記しよう．

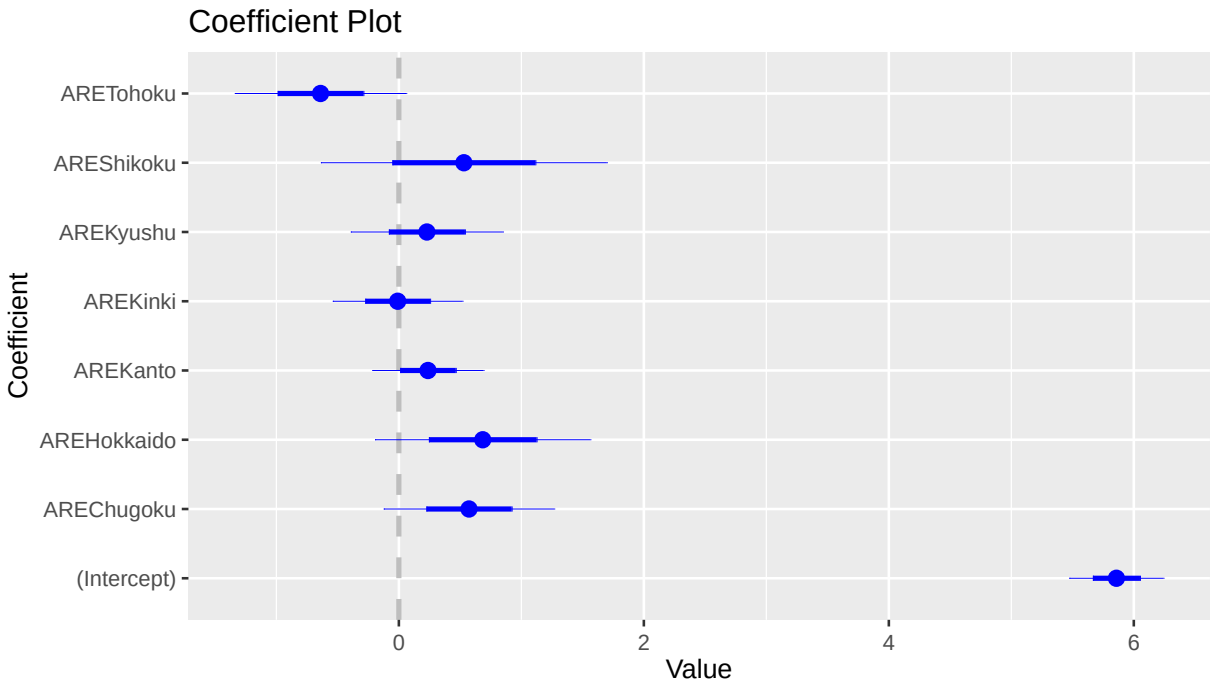
- パッケージ huxtable の中に huxreg という関数があります．

```
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)
```



	(1)
(Intercept)	5.858 ***
	(0.192)
AREChugoku	0.573
	(0.348)
AREHokkaido	0.685
	(0.439)
AREKanto	0.238
	(0.227)
AREKinki	-0.010
	(0.265)
AREKyushu	0.229
	(0.310)
AREShikoku	0.531
	(0.584)
ARETohoku	-0.639
	(0.350)
N	963
R2	0.014
logLik	-2180.170
AIC	4378.340

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

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 = "分析結果", 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．

```
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)
```

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．