統計学第 9 講/第 10 講

後藤　晶　　akiragoto@meiji.ac.jp

目次

目次

前回の復習

2 種類の χ^2 検定

- χ^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．

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）

クロス表を作る

```
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
```

χ^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．

χ^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．

  χ^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 (=5.1408, df=7, p=.64)．

- もし，χ^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
```

相関係数とは

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

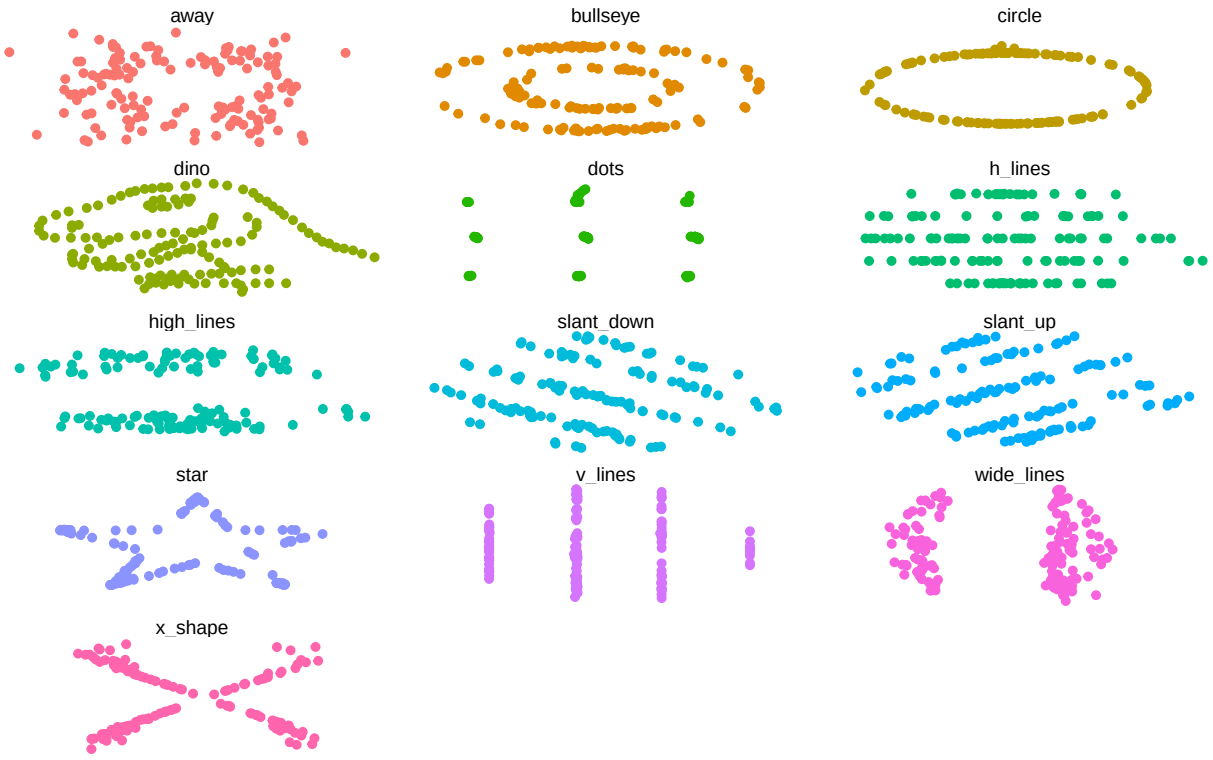
dataset	平均値	標準偏差	標本数	標準誤差
slant_down	54.27	16.77	142	1.407
slant_up	54.27	16.77	142	1.407
star	54.27	16.77	142	1.407
v_lines	54.27	16.77	142	1.407
wide_lines	54.27	16.77	142	1.407
x_shape	54.26	16.77	142	1.407

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
```



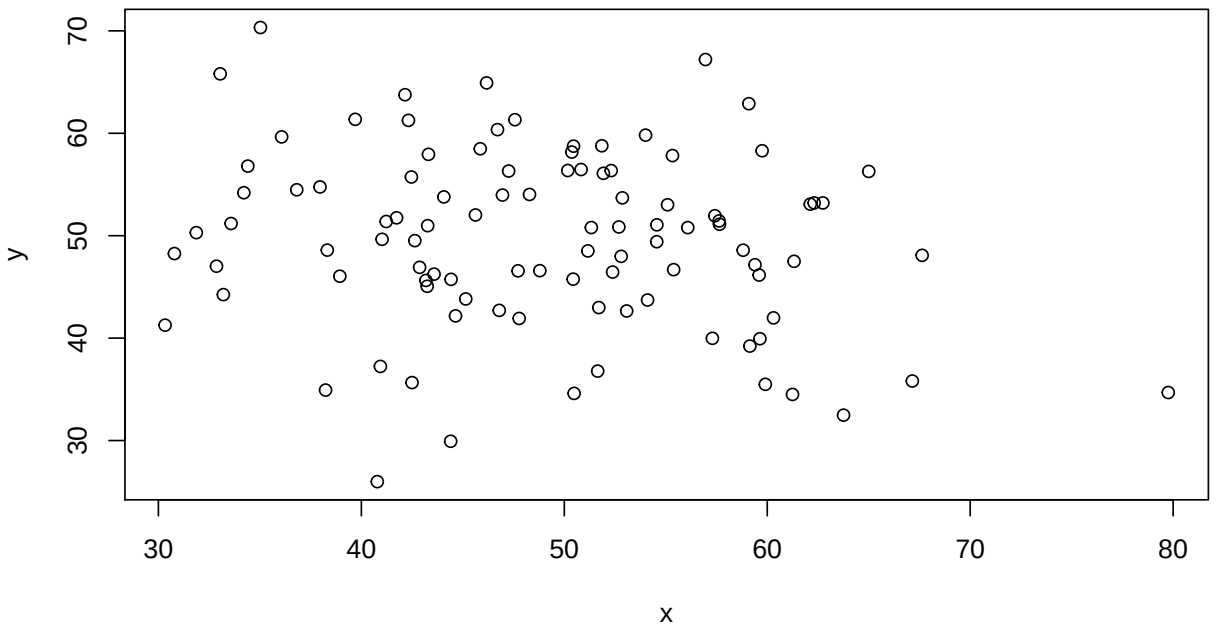
相関係数を出してみる

- 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
```

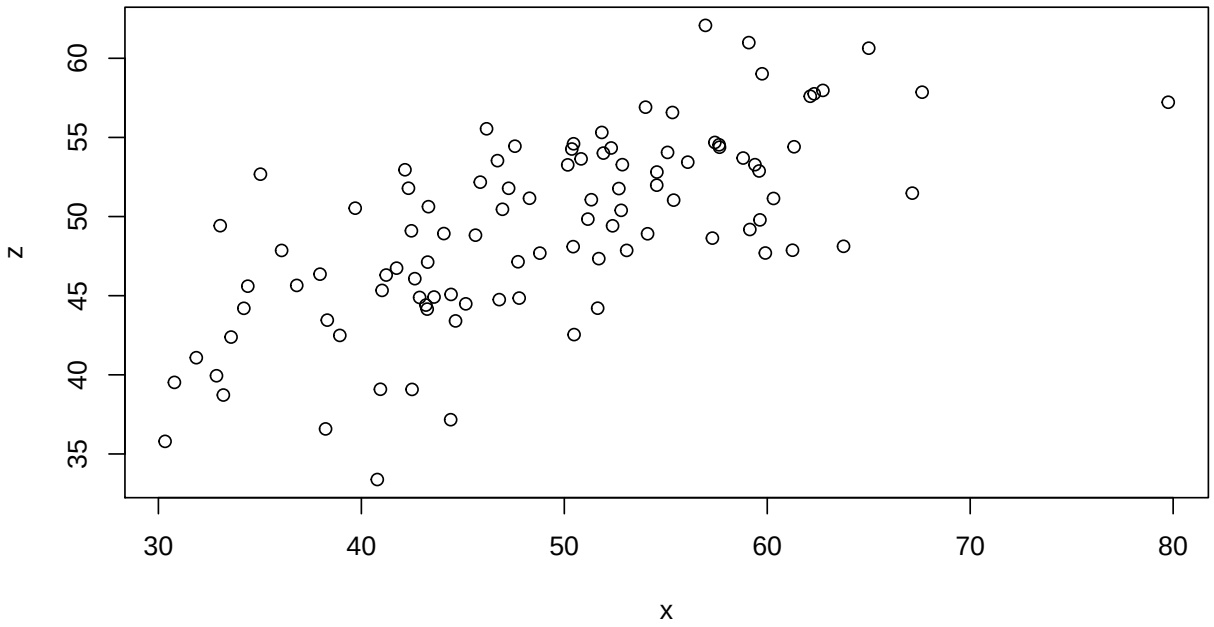
相関係数を出してみる

```
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] 0.5203983
```

- 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] 0.5203983
```

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

```
## [1] 53.69547
```

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] 0.005111368
```

- 演習問題
  - x と z について，相関係数を算出してみよう．

関数で相関係数を求める

```
cor(x, y)
```

```
## [1] 0.005111368
```

```
cor(x, z)
```

```
## [1] 0.7257103
```

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．

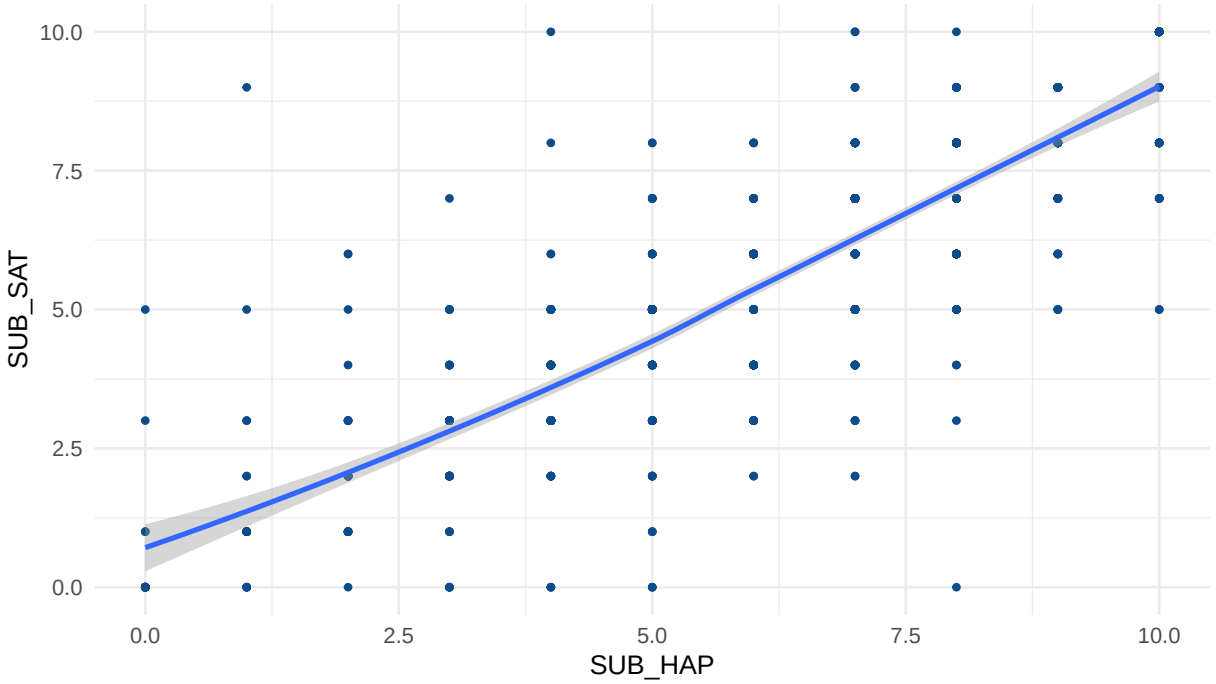
その上で，コードを貼り付けて出力することにしましょう．

- 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.

- 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)
```

表4: Fitting linear model: SUB_HAP ~ SUB_SAT

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.599	0.1018	15.71	1.156e-49
SUB_SAT	0.8104	0.01711	47.37	1.337e-253

- 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)
```

	(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")
```

- 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．

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．