SPSS分析マニュアル

田中愛治ゼミ*

Ver. October 15, 2014

Contents

*この資料は早稲田大学で、学部3－4年生を対象に開講された「田中愛治ゼミ」のために作成されたものです。不正確な部分がある可能性もありますので、内容に関して著者は責任を負いません。再配布などを希望する場合は、加藤言人<gento.badger@gmail.com>まで一報いただければと思います。

1　SPSSシンタックスを使ってみる *Use SPSS Syntax*

この節では、SPSSをシンタックスから使用する方法について説明します。この方法を使うと、自分がやった作業の実行プロセスを保存できるので、同じ分析を簡単に何度も繰り返すことができます。また、変数の作成に関しては、コードだけ保存することで、元のデータセットからどのように変数をつくりかえたかを記録することができます。コードだけを保存すれば、新しいデータセットを何度も保存し直す必要もありません。もちろん全ての作業はマウスをクリックすることによっても行うことができます。マウスで作業したい人は、各節にある「メニューから」項目を参照してください。また、マウスで操作した作業はシンタックスとしてアウトプットに表示されるので、それをコピーアンドペーストして繰り返すこともできます。

1.1　便利な事前設定　Pre-Settings

メニューから：＜編集＞⇒＜オプション＞

SPSSを使いやすく（見やすく）するためにデフォルト設定を変更します。

①＜全般＞変数リストを「名前を表示」「アルファベット順」に設定。※言語英語もここ。
②＜出力ラベル＞アウトラインのラベル付けを「名前とラベル」「値とラベル」に設定。
③＜出力ラベル＞ピボットテーブルのラベル付けも「名前とラベル」「値とラベル」。

1.2　データセットを開く Open Datasets

メニューから：＜ファイル＞⇒＜開く＞⇒＜データ＞

本演習で使用するデータは次の3つです。

①GLOPE0507　※ 2 (変数記述)/ 3 (変数作成)/ 6 (ロジット)/7.1(サブセット)で使用。

概要：
2005年と2007年に早稲田大学によって行われた世論調査のデータ。正式名は「21 世紀日本人の社会・政治意識に関する調査」。ケース単位は日本全国の個人（1397ケース）。衆院選2ヶ月後の2005年11月に調査（第1波）を行った後、2007年2月に第1波の対象者に対して再び調査を行っている（第2波）。
フォルダ内容：
データセット⇒ GL05&07_Ver1.sav
調査概要⇒ GL05&07-調査概要_Ver2.pdf　（配布資料参照）
第1波コードブック⇒ GL05&07-05Nov調査コードブック.pdf（配布資料参照）
第2波コードブック⇒ GL05&07-07Feb調査コードブック_Ver1.pdf

②dcd09　※4(二変数分析)/5(重回帰分析)/7.2(交互作用項)で使用。

概要：
2009年に存在した国々を対象とした各国比較のデータ。正式名は「Democracy Crossnational Data,Release 3.0 Spring 2009」。Pippa Norris Data Shared datasets <http://www.hks.harvard.edu/fs/pnorris/Data/Data.htm>から入手できる。ケース単位は国(191ケース)。変数によっては過去の複数の年におけるデータを含んでいる。
フォルダ内容：
データセット⇒　dcd09.sav
データ概要⇒　dcd09_description.txt
コードブック⇒　dcd09_codebook.pdf （配布資料参照）
SPSS参考資料⇒　Introductory Guide to using SPSS.pdf

③prefpol　※8 (時系列データ分析)で使用。

概要：
1969年から2011年までの都道府県知事・県議会に関するデータ。砂原庸介によるデータ<http://www.geocities.jp/yosuke_sunahara/data/data.html>を基にしている。ケース単位は都道府県×年（47都道府県×43年＝2021ケース、ただし選挙が無い年やデータ欠損などもあるので、変数ごとにケース数は一致しない）。年ごとの経済指数や人口構成なども含まれている。
フォルダ内容：
データセット　⇒　prefpol.sav
コードブック　⇒　prefpol_codebook.pdf （配布資料参照）
参考資料（知事出身コード）⇒ governer_list.pdf
元データ（Excel）⇒ prefpol_data.xlsx

シンタックスから（例）：

```
GET FILE = 'E:GLOPE0507.sav'
```

(' 'の中に、ファイルの場所と名前を入力する。）

1.3　シンタックスを開く Open Syntax File

メニューから：＜ファイル＞⇒＜新規作成＞⇒＜シンタックス＞ もしくは、
メニューから：＜ファイル＞⇒＜開く＞⇒＜シンタックス＞ を選択します。

1.4　適用するデータセットを指定する Define Active Dataset

データセットを指定しない場合、基本的にシンタックスファイルはその直前に開いていたデータセットに対して適用されます。どのデータセットを使って分析をしているかを確認・指定するには、シンタックス画面の右上にある「アクティブ：データセット＃」となっている欄が自分の使用したいデータセットと一致しているかを確認します。

シンタックスの中で指定する場合は、以下のコマンドを実行してから分析を始めます。（「データセット＃」については前段も同じですが、データセットファイル名ではなく、SPSSによってその時点で定義されているデータセット番号を入力します。）

DATASET ACTIVATE データセット＃.

1.5　コマンドを実行する Execute Commands

コマンドを実行するには、大まかに分けて次の３つの方法があります。

①実行したい範囲を選択してから、画面の上にある緑色の矢印（実行ボタン）をクリック。
②実行したい範囲を選択してから、Ctrl+Rを押す。（Windowsの場合）
③シンタックスファイルを全て実行したい場合は、＜実行＞⇒＜すべて実行＞をクリック。

データセットの変更などを行った場合、その変更がすぐにデータセットに反映されない場合があります。 その様な時は、シンタックスの最後に「Execute.」と入力しておくと、変更がすぐにデータセットに反映されます。

1.6　基本的な注意事項（ＳＰＳＳ特有の性質）Tips for SPSS Syntax

- コマンド途中での改行は自由。（ただし単語の途中では改行しないこと）
- 大文字・小文字は関係なし。
- 各コマンドの後にピリオド「．」をつける。（コマンド完了の印）
- コマンドの途中に全角スペースなどが入ると、エラーが起きる。
- コマンドと関係ないメモをする場合は、その行の最初に「*」「*/」「/*」を入力。

  (例) */これはメモです。．

- SPSSの出力画面では、クリックして分析を行った場合でも、実行されたコマンドがシンタックス形式で出力されます。分析結果の直前にある「ログ」という項目の中に出力されますので、この内容をシンタックスにコピー＆ペーストしてカスタマイズするというのも、シンタックスを学ぶ方法の一つでしょう。

2　変数を記述する *Describe Variables*

※このセクションの例には全て「GLOPE0507」を使用しています。

2.1　度数分布 *Frequencies (Freq)*

メニューから：＜分析＞⇒＜記述統計＞⇒＜度数分布表＞

　Frequencies (Freq)コマンドは、変数の度数分布を記述するのに使います。特に名義変数や順序変数の分布を確認するのに有用です。各変数を分析に投入する前に必ず変数の分布を確認するようにしましょう。想定した様な分布を持った変数になっていますか？

```
Frequencies variables = FA62 A17
/BARCHART FREQ
/ORDER=ANALYSIS.

Freq FA62 A17.
```

記入箇所（上段の場合）

variables = **［出力したい変数のリスト］**
/BARCHART FREQ
/ORDER=ANALYSIS の2行は、度数分布の棒グラフを出力したい時のみ追加。
（/BARCHARTの後を PERCENTにすると棒グラフの縦軸がパーセントになる）

Figure 1: 度数分布表

FA62 性別

		度数	パーセント	有効パーセント	累積パーセント
有効	1 男性	672	48.1	48.1	48.1
	2 女性	725	51.9	51.9	100.0
	合計	1397	100.0	100.0	

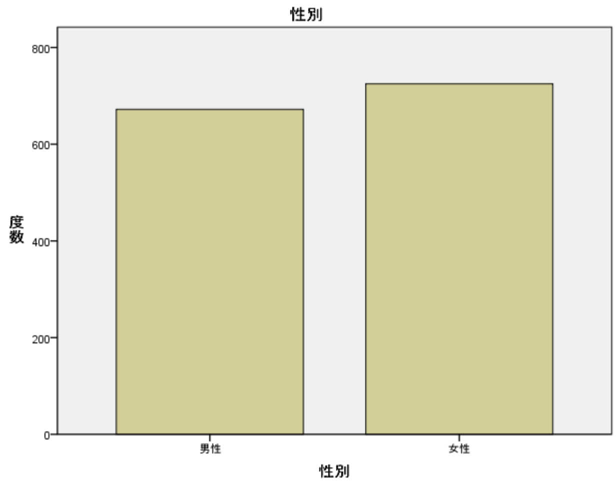
(a) FA62（性別）

A17 保守-革新

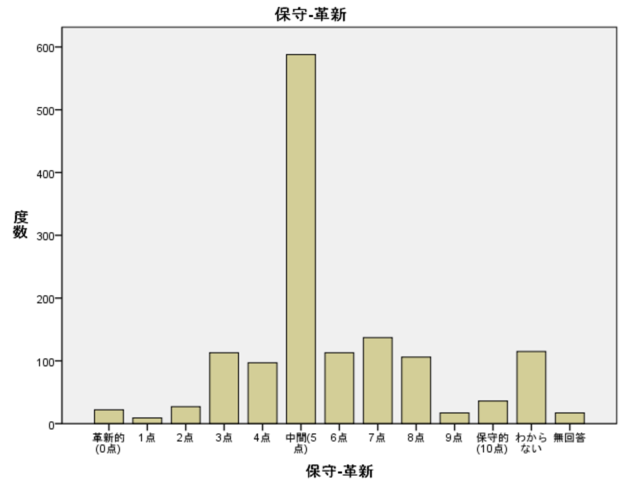
		度数	パーセント	有効パーセント	累積パーセント
有効	0 革新的(0点)	22	1.6	1.6	1.6
	1 1点	9	.6	.6	2.2
	2 2点	27	1.9	1.9	4.2
	3 3点	113	8.1	8.1	12.2
	4 4点	97	6.9	6.9	19.2
	5 中間(5点)	588	42.1	42.1	61.3
	6 6点	113	8.1	8.1	69.4
	7 7点	137	9.8	9.8	79.2
	8 8点	106	7.6	7.6	86.8
	9 9点	17	1.2	1.2	88.0
	10 保守的(10点)	36	2.6	2.6	90.6
	11 わからない	115	8.2	8.2	98.8
	12 無回答	17	1.2	1.2	100.0
	合計	1397	100.0	100.0	

(b) A17 保革イデオロギー

Figure 2: 度数分布棒グラフ



(a) FA62（性別）



(b) A17 保革イデオロギー

2.2　記述統計 *Descriptives (Descr)　/　Examine*

メニューから：＜分析＞⇒＜記述統計＞⇒＜記述統計＞

　Descriptives (Descr)コマンドでは、変数の記述統計を出力します。特に、連続変数の分布を確認するのに有効です。ヒストグラムや分位点を含めた詳細結果を出力したい場合には、Examineコマンドを使います。

```
Descriptives variables = FA63AGE.
Descr FA63AGE.

Examine FA63AGE
/plot histogram
/percentiles(5,10,25,50,75,90,95).
```

記入箇所

variables = **［出力したい変数のリスト］**　　（for Descriptives)
/plot ⇒ histogram　に加えて、boxplotなども出力できる。 (for Examine)
/percentile ⇒ 出力したい分位点を任意で指定 。(for Examine)

Figure 3: Descriptivesコマンドによる出力結果

記述統計量

	度数	最小値	最大値	平均値	標準偏差
FA63AGE 生年月日　満年齢	1397	20	88	54.80	16.286
有効なｹｰｽの数 (ﾘｽﾄごと)	1397				

Figure 4: Examineコマンドによる出力結果（順不同）

処理したｹｰｽの要約

	ｹｰｽ					
	有効		欠損値		合計	
	度数	ﾊﾟｰｾﾝﾄ	度数	ﾊﾟｰｾﾝﾄ	度数	ﾊﾟｰｾﾝﾄ
FA63AGE 生年月日　満年齢	1397	100.0%	0	.0%	1397	100.0%

(a) ケース数

ﾊﾟｰｾﾝﾀｲﾙ

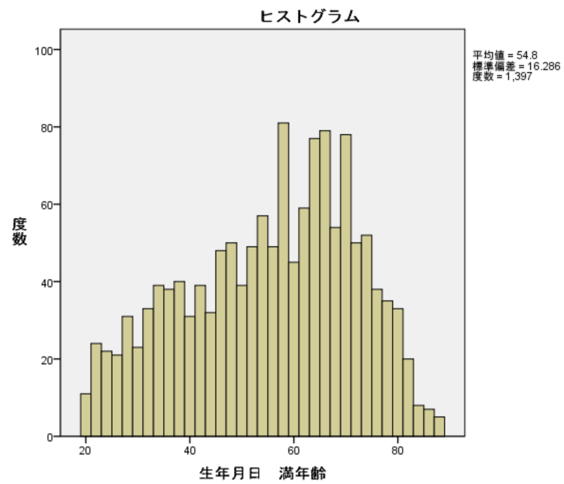
		ﾊﾟｰｾﾝﾀｲﾙ						
		5	10	25	50	75	90	95
重み付き平均(定義 1)	FA63AGE 生年月日　満年齢	26.00	31.00	42.00	57.00	68.00	75.00	79.00
Tukey のﾋﾝｼﾞ	FA63AGE 生年月日　満年齢			42.00	57.00	68.00		

(b) 分位点

記述統計

			統計量	標準誤差
FA63AGE 生年月日　満年齢	平均値		54.80	.436
	平均値の 95% 信頼区間	下限	53.94	
		上限	55.65	
	5%ﾄﾘﾑ平均		55.08	
	中央値		57.00	
	分散		265.218	
	標準偏差		16.286	
	最小値		20	
	最大値		88	
	範囲		68	
	4分位範囲		26	
	歪度		-.306	.065
	尖度		-.831	.131

(c) 詳細統計量



(d) ヒストグラム

3　変数を作る *Recode/Create Variables*

※このセクションの例には全て「GLOPE0507」を使用しています。

3.1　既にある変数の値を書き換える　*Recode* / *Missing Values*

メニューから：<変換>⇒<他の変数への値の再割り当て>

Recodeコマンドでは、既にある変数の値を書き換えます。コマンドの最後に「into」をつけることで、書き換えた後の値を新しい変数に書き込むことができます。(つけない場合は同じ変数が書き換わるので注意）また以下のような関数を使うことも可能です。

「A thru B」=「値Ａから値Bまで」
「lowest」＝最小の値
「highest」＝最大の値

「else」＝その他の値
「sysmis」＝システム欠損値（値自体を削除）
「missing」=ユーザー欠損値（値は残るが、分析には投入しない）
「else=copy」=その他の値については、元の値をコピーする

Missing values コマンドは、既にある変数の値の内一部を、ユーザー欠損値に指定するのに使います。「値は消したくないが、この分析では投入したくない」といった時に便利です。

```
*/Recode 2005 voting direction into LDP=1 DPJ=0 Others=missing.
Freq A28S2.
Recode A28S2 (1=1)(2=0)(3 thru highest = 9) into voteLDP.
Missing values voteLDP (9).
Freq voteLDP.

*/Recode A21A2, just elminate "Don't know" and "No Answer".
Freq A21A2.
Recode A21A2 (7 thru 8 = sysmis)(else = copy) into postatt.
Freq postatt.
```

3.2　計算式を書く *Compute*

メニューから：＜変換＞⇒＜変数の計算＞

Computeコマンドは、計算式を書くことで変数の作成を行うコマンドです。数式の書き方はほとんど直観に従って大丈夫です。足し算は「+」引き算は「-」掛け算は「*」割り算は「/」で行います。(　)で数式をくくることも可能です。

```
*/The identical var. as A17, just define "DK" and "NA" as missing.
Freq A17.
Compute ideology = A17.
Missing Values ideology (11, 12).
Freq ideology.

*/Compute the variable "koizumisup", with higer value means higher support.
 Freq A7.
 Compute koizumisup = 5-A7.
 Recode koizumisup (lowest thru 0 = sysmis) (else = copy).
 Freq koizumisup.

 */ Create 4 cats LDP support variable.
 Freq A6 A6S1 A6S3.
 Recode A6 (1=2)(else=0) into A6LDP.
 Recode A6S1 (1=1)(else=0) into A6S1x.
```

```
Recode A6S3 (1=1)(else=0) into A6S3x.
Compute LDPsup = A6LDP + A6S1x + A6S3x.
Freq LDPsup.
```

3.3　条件を指定する *If*/*Do If*

メニューから：※この機能は恐らくシンタックスからのみです。

　Ifコマンドは、条件を指定した後に、変更を適用します。

```
*/Create Knowledge for LDP and DPJ's Party Position.
Compute postknow = 0.
If (A21LDP2 < 7 and A21DPJ2 < 7 and A21LDP2 > A21DPJ2) postknow = 1.
```

　Do if コマンドは、間に関数を含むことができます。Else if　は「直前までに指定した範囲以外で、次の様に条件付けして適用」という意味で、何度使うこともできます。Elseは「上記全体で指定されなかったところ全てに適用」という意味です。全ての条件付けが終わったら、End if で閉じる必要があります。

```
*/Create 4 cats variable for DPJ support.
 Freq A6 A6S1 A6S3.
 Do if (A6=2).
 Recode A6S1 (1=3)(2 thru 4 = 2) into DPJsup.
 Else if (A6=7).
 Recode A6S3 (2=1)(else=0) into DPJsup.
 Else.
 Compute DPJsup=0.
 End if.
 Freq DPJsup.
```

3.4　ラベルをつける　*Variable*/*Value Labels*

メニューから：※変数ビューにしてから、「ラベル」「値」の欄に直接入力します。

　Variable Labels　コマンドは、変数ラベルを設定するのに使います。Value Labels コマンドは、変数の値にラベルを設定するのに使います。どちらも別に設定しなくても大丈夫ですが、 設定しておくと、出力などが見やすくなります。複数の変数に適用したい場合は「/」でつなぎます。

```
Variable Labels
    voteLDP "Voting Direction 2005 HoR PR" /
```

```
    postatt "Postal Privatization Attitude" /
    ideology "R's ideological position" /
    koizumisup "Koizumi Cabinet Support" /
    LDPsup "LDP supprt strength" /
    DPJsup "DPJ supprt strength" /
    postknow "Party Policy Knowledge - Postal Privatization".

Value Labels
    voteLDP 1 "Vote for LDP" 0 "Vote for DPJ" /
    postatt 0 "Negative" 3 "Neutral" 6 "Positive" /
    ideology 0 "Progressive" 5 "Neutral" 10 "Conservative" /
    koizumisup 1 "Don't Support" 2 "Rather Don't support"
               3 "Somewhat Supprt" 4 "Support" /
    LDPsup DPJsup 3 "Strong supporter" 2 "Not so strong supporter"
                  1 "Leaner" 0 "Non supporter" /
    postknow 1 "Informed" 0 "Uninformed" .
```

4　二変数間の関係を説明する *Explain Bivariate Rel.*

分析の手始めとして、興味のある2つの変数間の関係を探ることは、とても重要です。後々、様々な統制変数などを投入しながら、分析を精緻化していく訳ですが、まずは予測された関係が２つの変数だけ見ても現れているか確認するのは肝心なことです。（※今後、有意確率のしきい値は基本的に$p<0.05$として扱いますが、この値は絶対ではありません。所によっては適宜、別の基準を使用する場合があります。）

4.1　クロス表分析 *Crosstab*

メニューから：＜分析＞⇒＜記述統計＞⇒＜クロス集計表＞

　クロス表分析は、**名義変数・順序変数間の関係を見る**のに有効です。基本的な出力方法は次の通りです。さらに三重、四重にクロスをかけて表を拡張することもできます。

4.1.1　結果を出力してみる

※dcd09を使用した分析例:
　各国におけるFreedom of Speechの度合い（変数「Speech」）と、
　Extrajudicial Killing が発生している頻度（変数「Kill」）の関係が知りたい。

```
CROSSTABS
 /TABLES=  Kill BY Speech
 /CELLS=COUNT ROW COLUMN
 /STATISTICS=CHISQ PHI GAMMA
```

記入箇所（その他はSPSSのオプション参照）

/TABLES= **［列に入れる変数⇒独立変数？］**BY **［行に入れる変数⇒従属変数？］**
BY ［さらに表を分割⇒統制変数］
/CELLS= **［各セルに表示させたい数値］**
⇒COUNT（各セルに入る実数）
⇒ROW (各行の合計を100%とした場合の割合)
⇒COLUMN（各列の合計を100%とした場合の割合）
⇒TOTAL（対象サンプル全体を100%とした場合の割合）
⇒EXPECTED （各セルの期待値）etc...
/STATISTICS= **［出力するSummary Statistics］**
⇒CHISQ （カイ二乗検定 *期待値とどれだけ外れているかを見る）
⇒PHI（Phi 及び Cramer's V *名義変数を含む変数同士の関連を見る）
⇒GAMMA（Gamma *順序変数同士の関連を見る）etc...
⇒ALL（出力可能なSummary Statistics全て）

Figure 5: クロス表分析出力結果

Kill Extrajudicial killing (CIRI) と Speech Freedom of Speech (CIRI) のクロス表

			Speech Freedom of Speech (CIRI)			合計
			0	1	2	
Kill Extrajudicial killing (CIRI)	0 Practiced frequently	度数	15	18	5	38
		Kill Extrajudicial killing (CIRI) の %	39.5%	47.4%	13.2%	100.0%
		Speech Freedom of Speech (CIRI) の %	42.9%	27.3%	9.6%	24.8%
	1 Practiced infrequently	度数	15	26	11	52
		Kill Extrajudicial killing (CIRI) の %	28.8%	50.0%	21.2%	100.0%
		Speech Freedom of Speech (CIRI) の %	42.9%	39.4%	21.2%	34.0%
	2 Have not occurred	度数	5	22	36	63
		Kill Extrajudicial killing (CIRI) の %	7.9%	34.9%	57.1%	100.0%
		Speech Freedom of Speech (CIRI) の %	14.3%	33.3%	69.2%	41.2%
合計		度数	35	66	52	153
		Kill Extrajudicial killing (CIRI) の %	22.9%	43.1%	34.0%	100.0%
		Speech Freedom of Speech (CIRI) の %	100.0%	100.0%	100.0%	100.0%

Figure 6: クロス表分析検定量

カイ 2 乗検定

	値	自由度	漸近有意確率 (両側)
Pearson のカイ 2 乗	30.554[a]	4	.000
尤度比	31.974	4	.000
線型と線型による連関	26.921	1	.000
有効なケースの数	153		

a. 0 セル (.0%) は期待度数が 5 未満です。最小期待度数は 8.69 です。

(a) カイ二乗検定量

対称性による類似度

		値	漸近標準誤差[a]	近似 T 値[b]	近似有意確率
名義と名義	ファイ	.447			.000
	Cramer の V	.316			.000
順序と順序	ガンマ	.567	.082	6.205	.000
有効なケースの数		153			

a. 帰無仮説を仮定しません。
b. 帰無仮説を仮定して漸近標準誤差を使用します。

(b) その他の検定量

4.1.2　結果を解釈する

表の解釈の仕方は、見た通りだとは思いますが、Summary Statistics について少しだけ説明します。まず、「カイ二乗検定」表の中の、**「Pearsonのカイ二乗」**行における「漸近有意確率」を確認します。この値が<0.05であれば、掛け合わせた変数によって分布に「差がない」という帰無仮説が棄却されて、「差があるだろう」という結連に至る訳です。（ただしカイ二乗検定は、ケース数が５を下回るセルが多い場合、結果が信頼できないので注意）

　他の指標についても色々ありますが、ここでは２つを紹介します。１つ目は、**「クラメールのV(Cramer's V)」で、特に名義変数を含む2つの変数の関係の深さを判定する時に使用されます。**値は、２つの変数の関係の深さを０から１の間で表します。１は完全な連関関係、０は全く連関がない状態を示します。カイ二乗値を利用した指標なので、有意確率はカイ二乗と同じになります。

　2つ目の指標は**「ガンマ（グッドマン・クラスカルのガンマ）」で、順序がある変数同士の関係の深さを、方向性も含めて判定する時に使用されます。**値は、１が完全な正の連関、０が連関なし、-1が完全な負の連関を示します。（検定方法がカイ二乗とは異なるので、有意確率はカイ二乗と同じにはなりません）

　分析例では、どちらも3つの値を持つ順序変数である、「Speech」と「Kill」が分析に投入されていますので、「ガンマ」を使用することができそうです。ガンマの値も0.567で、$p<0.05$で有意ですね。よって、Speechの値が上がるほど、Killの値が高い場合が多くなる、という正の連関が見られる、といえそうです。これらの検定量は、最終的に表にした場合には、表の下に加えて表記するようにしてください。(例：Table1)他にも様々な指標がありますが、結果を最終的に示す際には、**分析に投入している変数の性質を考えて適切な指標だけを表記する**ようにしましょう。

4.2　平均値の差の検定 *T-test*

メニューから：＜分析＞⇒＜平均の比較＞⇒＜独立したサンプルのＴ検定＞

　平均値の差の検定は、**連続変数の平均値を、２つのグループごとに比べて、差があるかどうかを調べたい**場合に使います。ここでは２つのグループの差だけですが、応用として3つ以上のグループの差を検定する方法もあります。（ここでは述べません）

4.2.1　結果を出力してみる

※dcd09を使用した分析例:

　各国における、安全な水にアクセス出来る人の割合（変数「Water2000」）と、
　一人当たりのGDP（変数「GDP2000」）を、
　民主国家と独裁国家の政治体制別（変数「Cheibub2Type」）に比べたい。

```
T-TEST GROUPS=Cheibub2Type (0,1)
/MISSING=ANALYSIS
/VARIABLES= Water2000 GDP2000
/CRITERIA=CI(.95).
```

記入箇所（その他はSPSSのオプション参照）

T-TEST GROUPS= **［グループ(G)分け変数］**（**［G１の値］，［G２の値］**）
/VARIABLES= **［平均値を比較する変数（連続変数）のリスト］**

Figure 7: 平均値の差の検定

グループ統計量

	Cheibub2Type Cheibub type of regime 2000	N	平均値	標準偏差	平均値の標準誤差
Water2000 Population with sustainable access to an improved water source (%) 2000 (UNDP 2004)	0 Democracy	85	84.6588	15.87042	1.72139
	1 Dictatorship	63	71.1905	23.48703	2.95909
GDP2000 GDP per capita ppp 2000 (world bank 2002)	0 Democracy	94	$10,927.03	$10,267.663	$1,059.029
	1 Dictatorship	47	$4,234.30	$4,457.797	$650.236

(a) 実際の平均値の差

独立サンプルの検定

		等分散性のための Levene の検定		2つの母平均の差の検定					差の 95% 信頼区間	
		F 値	有意確率	t 値	自由度	有意確率 (両側)	平均値の差	差の標準誤差	下限	上限
Water2000 Population with sustainable access to an improved water source (%) 2000 (UNDP 2004)	等分散を仮定する。	16.279	.000	4.161	146	.000	13.46835	3.23719	7.07054	19.86615
	等分散を仮定しない。			3.934	102.407	.000	13.46835	3.42336	6.67846	20.25824
GDP2000 GDP per capita ppp 2000 (world bank 2002)	等分散を仮定する。	40.720	.000	4.266	139	.000	$6,692.734	$1,568.769	$3,591.000	$9,794.468
	等分散を仮定しない。			5.386	136.979	.000	$6,692.734	$1,242.718	$4,235.340	$9,150.128

(b) 平均値の差の検定

4.2.2　結果を解釈する

結果の解釈の仕方を説明します。まず、２つ表が出力されることが分かります。1つ目の表は、実際の平均値の値と差を示し、2つ目の表はその差が統計的に有意かどうかを検定しています。最終的な結果を示す際には、実際の平均値をもちろん示しますが、ここでは2つ目の表に注目します。

①**ルビーン（Levene）の等分散検定**：ここでは、平均値の差の検定を見る前に、2つのグループの間で分散が等しいかどうかを検定します。分散が違う場合、平均値の差の検定の公式が変わってしまうからです。等分散検定の帰無仮説は「2つのグループの間で分散は等しい」です。表の「等分散性のためのLeveneの検定」の有意確率を見てください。

（１）**p>0.05の場合**：「2つのグループの間での分散が等しいという帰無仮説は、否定できない」となるので、等分散を仮定してよいことになります。よって、**平均値の差の検定では「等分散を仮定する。」上の行の結果を見ます**。

（２）**p<0.05の場合**：帰無仮説が棄却され、「２つのグループの間での分散が等しいことは仮定できない」となります。よって、**平均値の差の検定では「等分散を仮定しない。」下の行の結果を見ます**。

②**平均の差のＴ検定**：等分散検定の結果に従って、適当な方の「２つの母平均の差の検定」の有意確率を見ます。ここでの帰無仮説は「2つのグループ間での平均値には差が無い（同

じである）」ということになります。よって、**p<0.05であれば帰無仮説を棄却し、「2つのグループ間の平均値には差がある」と結論付ける**ことになります。

使用した例では、Water2000について見ると、等分散検定の有意確率が.000で、等分散は仮定できないので、「等分散を仮定しない」下の行の結果を見ます。すると、平均値の差の検定の有意確率は.000となっています。よって、「安全な水にアクセスできる人の割合は、民主国家の方が独裁国家よりも平均して高い」ということができるでしょう。
ただし、この結果は2変数だけを観察して得られた結果であることには注意してください。もう１つの結果を見てみると、1人当たりGDPの平均も、統計的に有意に民主国家の方が独裁国家よりも高くなっていることが分かります。ここで、1人当たりGDPが高い様な豊かな国において、安全な水にアクセスできる人が多くなることは想像に難くありません。GDPが持つインパクトを取り去っても、国家体制は安全な水へのアクセスに影響を与えるのでしょうか？この疑問に答えるのが、複数の独立変数を考慮した重回帰分析です。

4.2.3　平均値の差を図示する―標準誤差付き棒グラフ　*Error Bar Plot*

メニューから：＜グラフ＞⇒＜レガシーダイアログ＞⇒＜棒＞

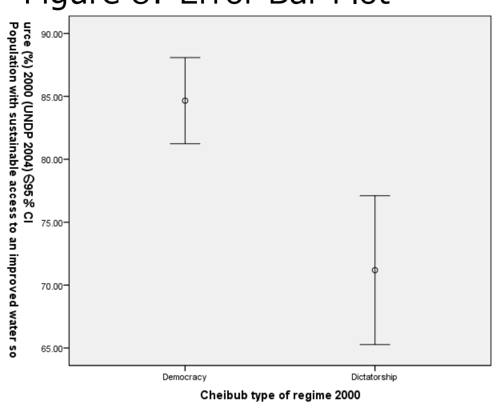
グループごとの平均値の差は、標準誤差付き棒グラフで簡単に図示することができます。この図は、グループごとの平均値を棒グラフにし、その平均値に標準誤差を示す線を加えたものです。基本的には、標準誤差の範囲が重複していなければ、差が有意になるはずです（等分散が仮定できない場合はその限りではないと思われますが）。Ｔ検定の結果(有意確率など）と合わせて示すと有効かもしれません。以下、Water2000の平均値をCheibub2Typeのカテゴリーごとに比べた例です。

```
GRAPH
/ERRORBAR(CI 95)=Water2000 BY Cheibub2Type.
```

記入箇所（その他はSPSSのオプション参照）

/ERRORBAR(CI 95)=**［平均値を比較する変数］**BY **［グループ分け変数］**

Figure 8: Error Bar Plot



4.3 相関 *Correlation*

メニューから：<分析>⇒<相関>⇒＜２変量>

　相関分析は、２つの連続変数間の関係の深さを分析する手法です。とても手軽で便利ですが、解釈にはやや注意が必要です。

4.3.1 結果を出力してみる

※dcd09を使用した分析例:
　各国における、安全な水にアクセス出来る人の割合（変数「Water2000」）が、
　一人当たりのGDP（変数「GDP2000」）と共に変化するかを調べたい。

```
CORRELATIONS
 /VARIABLES= Water2000 GDP2000
 /PRINT=TWOTAIL NOSIG
 /MISSING=PAIRWISE.
```

記入箇所（他はSPSSのオプションを参照してください）

variables = **［相関をとりたい変数のリスト］**
注：相関分析では変数の順番は関係ありません。

Figure 9: GDP2000とWater2000の相関

相関係数

		Water2000 Population with sustainable access to an improved water source (%) 2000 (UNDP 2004)	GDP2000 GDP per capita ppp 2000 (world bank 2002)
Water2000 Population with sustainable access to an improved water source (%) 2000 (UNDP 2004)	Pearson の相関係数	1	.559**
	有意確率 (両側)		.000
	N	150	116
GDP2000 GDP per capita ppp 2000 (world bank 2002)	Pearson の相関係数	.559**	1
	有意確率 (両側)	.000	
	N	116	141

**. 相関係数は 1% 水準で有意 (両側) です。

4.3.2 結果を解釈する

表中に示された相関係数は、**２つの連続変数間に線型関係を想定した場合に、その線にどれだけ分布が近いか**を表します。-1から１の値をとり、-１ならば完全な負の線型関係（あてはめた直線の上に全ての分布が存在している）、０ならば全く線型関係が無い（二つの

変数の間では、分布がランダム）、１ならば完全な正の線型関係ということになります。必ずしも、**線の傾き（関係の強さ）を表している指標ではない**ことに注意してください。有意確率は、これまでと似たように、帰無仮説「2つの変数の間には線型関係が無い」が真である確率を示しています。p<0.05で帰無仮説が棄却され「何らかの線型関係がある」という結論に至る訳です。

使用した例の結果を見ると、相関係数が0.559で有意確率が0.000でなので、GDP2000はWater2000と統計的に有意な、かなり深い線型関係にあることが分かります。一人当たりGDPの値が上昇するほど、水にアクセスできる人の割合も上昇するということです。

4.3.3　散布図を出力する *Scatterplots*

メニューから：<グラフ>⇒<レガシーダイアログ>⇒<散布図/ドット>

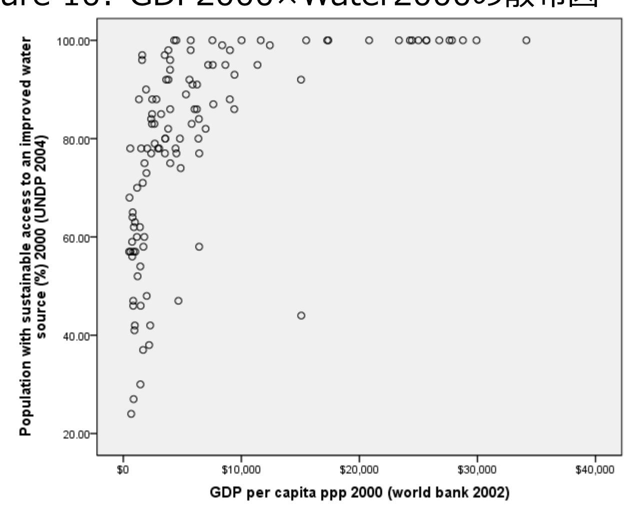
前節で0.559という相関係数が出力されましたが、これは具体的にどれくらい線型であることを指しているのでしょうか？この疑問にビジュアルで答えてくれるのが、散布図です。SPSSでは以下の様に簡単に散布図を出力することができます。

```
GRAPH
/SCATTERPLOT(BIVAR)=GDP2000 with Water2000
/MISSING=LISTWISE.
```

記入箇所（他はSPSSのオプションを参照してください）

/SCATTERPLOT(BIVAR)= **［独立変数］** with **［従属変数］**

Figure 10: GDP2000×Water2000の散布図



散布図を見て、いかがでしょうか。確かにGDP2000が上昇するに従って、Water2000が上昇するという関係が見られますね。一方で、この関係は線型だと言えるでしょうか？よく図を見ると、GDP2000の値が大きくなるに従って、「GDP2000の値が上昇すると、

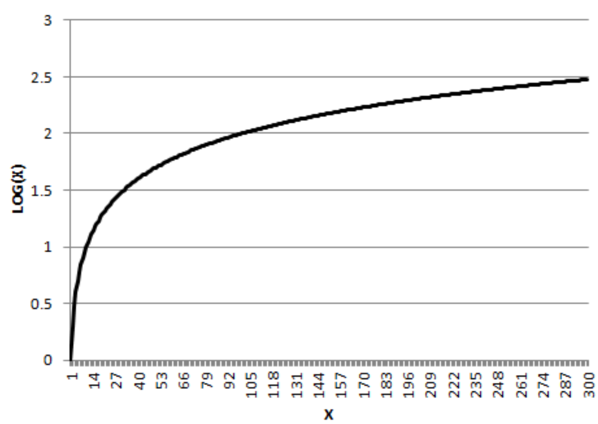
Water2000が上昇する」速度が段々遅くなっていることが分かります。２つの変数の関係は、直線というよりかは、曲線ですね。次節では、この問題を解消する1つの方法について説明します。

4.3.4　変数を変換する *Transforming Variables*

前節でみられたような曲線関係は、経済指数を独立変数として扱った分析を行う場合によく見られます。「経済成長率」などと言う通り、経済指数は現在の値に対する割合で変化することが多く、実数に直すと加速度的に変化をしているためです。しかし、次節で述べる重回帰分析では相関分析と同じように独立変数と従属変数の線型関係を想定しているので、曲線関係が発生してしまっては、正確な分析ができません。

　そこで、独立変数Ｘの値を関数を使って変換して、従属変数Ｙと線型関係になりそうな値に変換してしまおう、というアイデアがあります。変換したＸを分析に投入すれば、Ｙとの間に線型関係が想定できるからです。ここで、Ｘの自然対数(Log)という関数があります。Figure11を見てください。このグラフはＸが大きくなればなるほど「ＸがLOG(X)と持つ関係の傾き」が緩くなりますね。GDP2000とWater2000の関係にとてもよく似ています。それでは、この関数でGDP2000を変換して再分析してみましょう。

Figure 11: Xの自然対数 Log をとる




```
Compute logGDP2000 = Ln(GDP2000).

CORRELATIONS
/VARIABLES= Water2000 logGDP2000
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.

GRAPH
/SCATTERPLOT(BIVAR)=logGDP2000 with Water2000
/MISSING=LISTWISE.
```

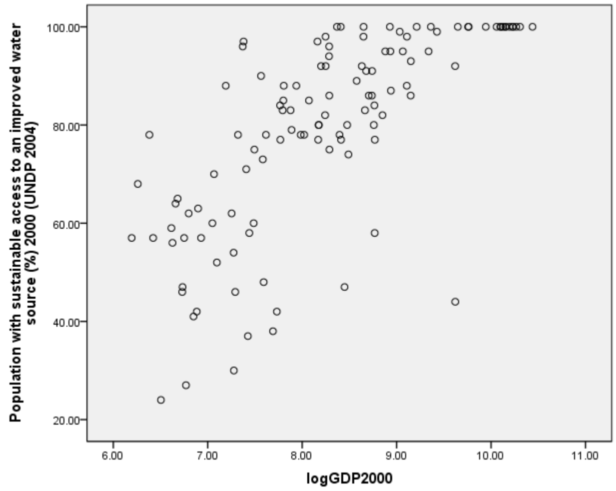
Figure 12: 変数を変換するとより関係が線型になる

相関係数

		Water2000 Population with sustainable access to an improved water source (%) 2000 (UNDP 2004)	logGDP2000
Water2000 Population with sustainable access to an improved water source (%) 2000 (UNDP 2004)	Pearson の相関係数	1	.712**
	有意確率 (両側)		.000
	N	150	116
logGDP2000	Pearson の相関係数	.712**	1
	有意確率 (両側)	.000	
	N	116	141

**. 相関係数は 1% 水準で有意 (両側) です。

(a) logGDP2000とWater2000の相関



(b) logGDP2000×Water2000の散布図

結果を見ていただくと、相関係数が0.712と、大幅に上昇しています。そして、散布図でも、XとＹの関係がより線型に近づいていますね。Log変換をしたことで、XとＹの線型関係をより予測できるようになったといえるでしょう。（もちろん、いつも、どの変数も変換すればいいというものではありません。散布図などをみながら、変換が必要かどうかについては慎重に検討しましょう。）

5　連続変数を説明する―重回帰分析 *OLS Regression*

メニューから：<分析>⇒<回帰>⇒<線型>

最小二乗法を使用した重回帰分析（Ordinally Least Square OLS Regression)は、連続変数の従属変数を説明する上で、最もポピュラーな統計手法です。その仕組みや求め方についてはここでは深く触れませんが、ぜひメカニズムを理解した上で分析に臨んでください。以下ではシンタックスによる出力の仕方や、結果の解釈方法について説明します。

5.1　結果を出力してみる

※dcd09を使用した分析例（*Recode_dcd09.sps*を実行する）:
各国における、民主化度（変数「fhmean1990_1999rev」）と、
一人当たりのGDP（変数「logGDP2000」）が、
安全な水にアクセスできる人の割合（変数「Water2000」）に、
どのくらいインパクトを持っているかを調べたい。
仮説１：民主化度が高い国ほど、安全な水にアクセスできる人の割合が増える。
仮説２：経済的に豊かな国ほど、安全な水にアクセスできる人の割合が増える。

```
*/Reverse the score of freedomhouse scale.
Compute fhmean1990_1999rev = 8 - fhmean1990_1999.

REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT Water2000
/METHOD=ENTER fhmean1990_1999rev logGDP2000.
```

記入箇所（他はSPSSのオプションを参照してください）

/DEPENDENT= **［従属変数］**
/METHOD=ENTER **［独立変数・統制変数のリスト］**

Figure 13: fhmean1990_1999とGDP2000でWater2000を予測する重回帰分析

モデル集計

モデル	R	R2 乗	調整済み R2 乗	推定値の標準誤差
1	.733[a]	.538	.530	13.69127

a. 予測値: (定数)、logGDP2000, fhmean1990_1999rev。

分散分析[b]

モデル		平方和 (分散成分)	自由度	平均平方	F 値	有意確率
1	回帰	24661.116	2	12330.558	65.780	.000[a]
	残差 (分散分析)	21181.944	113	187.451		
	合計 (ピボットテーブル)	45843.060	115			

a. 予測値: (定数)、logGDP2000, fhmean1990_1999rev。
b. 従属変数 Water2000 Population with sustainable access to an improved water source (%) 2000 (UNDP 2004)

係数[a]

モデル		標準化されていない係数		標準化係数	t 値	有意確率
		B	標準誤差	ベータ		
1	(定数)	-19.069	10.221		-1.866	.065
	fhmean1990_1999rev	2.487	.908	.219	2.738	.007
	logGDP2000	10.572	1.453	.581	7.274	.000

a. 従属変数 Water2000 Population with sustainable access to an improved water source (%) 2000 (UNDP 2004)

5.2　結果を解釈する

すでに学んだこととは思いますが、重回帰分析の予測値を求める式は、次の様に書くことができます。(変数の数=*a*の時。予測値なので、残差は式に入っていません。)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_a X_{ai} \tag{1}$$

SPSSの「係数」の表にある、「標準化されていない係数」の「Ｂ」は、上記の式(1)の$\hat{\beta}_1$から$\hat{\beta}_a$を表している訳です。（（定数）と書いてあるところの数値は$\hat{\beta}_0$です。）式の中の他の値が全て固定されていると仮定すると、$\hat{\beta}_1$の値が、X_{1i}が１上昇した時の、$\hat{Y}_i$の変化量を示しているということは、知っての通りです。有意確率が$p<0.05$ならば、その変数が統計的に有意なインパクトを従属変数に持っているというのも知っていると思います。

　ここで、例の分析結果を解釈すると、まず、投入したfhmean1990_1999とGDP2000のどちらもが統計的に有意なインパクトをWater2000に持っていることが分かります。具体的には、まずfhmean1990_1999revが１上昇すると、安全な水にアクセスできる人の割合が2.49%上昇するという結果になっています。fhmean1990_1999revは、１がNot-Free、７がFreeとなっている連続変数なので、「民主化が進んでいる国ほど、安全な水にアクセスできる人の割合が高くなる」という結果になり、仮説１に整合的です。
　logGDP2000に関しては、値が１上昇するごとに、安全な水にアクセスできる人の割合が10.57%上昇するということが分かります。この関係も、仮説２に整合的です。（ただし、これはlogGDP2000が持つインパクトですので、直接GDP2000が持つインパクトを述べたい場合には、Xの値を指数変換し直して解釈する必要があります。）

　最後に、SPSSだとデフォルトで出力される、**標準化係数の解釈**について少し触れておきます。標準化係数は、Ｘが「１標準偏差」分上昇した時に、Ｙが「何標準偏差」分変化するかを示した指標として理解されます。そして、一般的には「１標準偏差」としてＸの上昇量を標準化しているので、スケールの数が違う変数同士のインパクトをモデル内で比べるのに有効であるとされています。一方で、「標準偏差」は、分析対象のサンプルに深く依存している指標であるため、違うサンプル間で標準化係数を比べることは不適切だと言えるでしょう。更に、「標準偏差」が、本当にXの上昇量を比較可能にしているかどうかには批判もあります。例えば、教育年数の標準偏差と収入の標準偏差が同じ意味を持っていると誰が言えるでしょうか？それならば、標準化係数を比べるよりも、教育年数が1年長くなった場合と、収入が100万円増加した場合の非標準化係数を比べた方がずっと比較可能な気がするでしょう。もちろん、標準化係数を使用して結果を解釈するかどうかは個人の自由ですが、最近は上記の様な問題点を含めて、**使われることが少なくなってきている**ようです。

5.3　モデル全体を評価する　*Evaluate the Model*

分析の係数の表に載っているもの以外で、モデル全体を評価するものとして、最終的な結果に含めた方が良い（かもしれない）指標は以下の通りです。

- **R２乗（R-Squared）**：「モデル集計」表の中にあります。当該モデルがＹを予測するのに、Ｙ平均値から予測する（Null Model）のに比べて、Fitted Line(Plane, Hyperplane)でＹを予測するとどのくらい予測が改善するかを示す指標です。０から１の値を取ります。「Yの真の分布」が完全に直線であるとは限らないので、値が高いことが必ずしも「当てはまりの良さ」を意味するわけではありませんが、0.1を大きく下回っている場合は少し心配した方が良いかもしれません。また、変数が増えるに従って自動的に値が増えるので、基本的に変数の数が違うモデル間の比較は不可能です。

- **推定値の標準誤差（Standard Error of Estimation）**：「モデル集計」表の中にあります。当該モデルのFitted LineによるＹの推定値と実際のＹの値の距離を自由度を考慮して標準化したものです。モデル全体の「当てはまりの良さ」の指標としては、R２乗よりもこちらの方が適切です。値が小さいほど、当てはまりが良くなります。また、値の単位がＹの単位と一致しているので、ＹとＹの推定値の間の標準誤差として、実は直接解釈が可能です。

- **調整済みR2乗（Adjusted R-Squared）**：「モデル集計」表の中にあります。R２乗の値を、自由度を考慮して調整したものです。R２乗の弱点の1つである「変数の数と共に自動的に値が増える」ことが修正されています。ただし、０から１の間の値を完全にはとらなくなり、厳密に何の値を示しているのかは謎です。一方で「推定値の標準誤差」と共変するので、一般的な解釈としてはモデル全体の「当てはまりの良さ」の指標として使われることが多いです。

- **F検定**：「分散分析」表の「Ｆ値」とその横にある「有意確率」をみます。これは、モデルに含まれている変数のうち「最低1つの変数」の係数が0よりも大きいかどうかを検定しています。$p>0.05$だった場合は、モデルに含まれている「全ての変数」の係数が０（全て意味が無い）可能性が否定できません。ほとんどの場合は$p<0.05$になるので心配しなくても大丈夫です。特に表に載せる必要もありません。

- **ケース数**：これが意外と分かりにくいのですが、「分散分析」の表の「合計」行の「自由度」列を見ます。分散分析の自由度は「N-1」なので、その数値に１を足したものが重回帰分析で扱われているケース数になります。

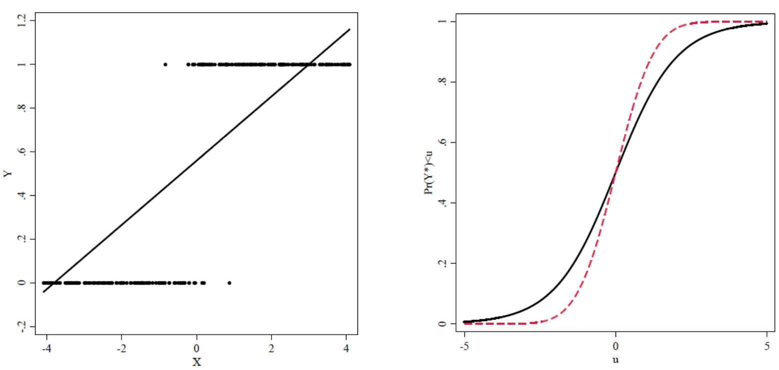
6　二値変数を説明する─ロジット分析 *Logit*

ロジット分析は、説明したい従属変数が２つの値しか持たない離散型変数の場合に使う分析です。分析上、従属変数Yは「1」と「0」の２つの値を持つ変数として扱われます。よって、Yの値を直接予測するのではなく、Yの値が「1」になる確率を求めるモデルであると理解すれば良いと思います。（ロジット分析の応用として３つ以上の値をもつ離散型変数を従属変数とする分析もありますが、ここでは扱いません）

6.1　何故、OLSではなくロジットなのか

Figure14を見てください。左側が、無理やり二値変数に対してOLSを行った場合のfitted-lineです。予測値が、0と1の範囲を飛び出してしまっています。不可能な値を予測してしまっては適切な分析とはいえません。よって、ロジット分析では、図右側のような曲線（実線の方）を予測にあてはめます。そうすることで、０から１の範囲内で予測値が返せるようになります。また、０%や１００%のような極端な値になるほど予測が難しくなる、というのは直観的にも正しいといえるでしょう。

Figure 14: なぜ二値の従属変数をOLSであてはめてはいけないのか



(a) Linear Prob. Model　　(b) Logit (Probit) Model

6.2　結果を出力してみる

メニューから：<分析>⇒<回帰>⇒<二項ロジスティック>

　ロジット分析の結果を出力する方法は、重回帰分析とほぼ同じです。クリックで行うこともできますし、シンタックスコマンドを使用して行うこともできます。

※GLOPE0507を使った例

■RQ：2005年の衆議院総選挙で自民党に投票した人の投票決定要因は何か。

　仮説1: 郵政民営化に強く賛成している人ほど、自民党に投票した

　仮説2:　イデオロギー的に保守である人ほど、自民党に投票した

■ 使う変数（*Recode_GL0507.sps*を実行する）：

- voteLDP ⇒　自民党投票=1 / 民主党投票 = 0 （比例区。その他の政党は欠損）
- postatt ⇒　郵政民営化に積極的=6から消極的=0まで7カテゴリー（DKは欠損）

- ideology ⇒ 保守的=10から革新的=0まで11カテゴリー (DKは欠損)
- LDPsup ⇒ 自民党支持。3=強い支持/2=弱い支持/1=好ましい/0=支持しない
- DPJsup ⇒ 民主党支持。3=強い支持/2=弱い支持/1=好ましい/0=支持しない
- koizumisup ⇒ 小泉内閣支持。4=支持/3=ある程度支持/2=あまり支持しない/1=支持しない（DKは欠損）

```
LOGISTIC REGRESSION
 VARIABLES voteLDP
/METHOD=ENTER postatt ideology LDPsup DPJsup koizumisup
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

記入箇所（他はSPSSのオプションを参照してください）

/VARIABLES= **［従属変数］**
/METHOD=ENTER **［独立変数・統制変数のリスト］**

6.3 結果を解釈する

6.3.1 係数はOLSと同じようには解釈できない

SPSSは、ロジスティック回帰分析の結果として複数の表を出力してくれますが、基本的に重要となるのは、一番最後に出力される「方程式中の変数」の表のみです。（モデル全体を評価する指標などについては後ほど述べます。）その表を見てもらうと、OLSで出力したのと似たような結果が並んでいますね。（ただし、定数項の係数が、一番上ではなく、一番下に出力されていることには注意してください。）基本的な結果を発表する際は「Ｂ」「標準誤差」「有意確率（または*で示す）」そして好みによっては、「Exp(B)」を載せると良いでしょう。この中で「標準誤差」や「有意確率」の理解は、OLSとほぼ同じで大丈夫です。すなわち、標準誤差は係数の値の標準偏差ですし、有意確率$p<0.05$であれば「帰無仮説が真である確率が５％未満である」と解釈し、その変数が有意な関係をＹと持っていると理解してよいということです。

　一方で、「Ｂ」を単純に「Ｘが１大きくなった時のＹの変化量」として解釈することはできません（解釈できるのは、プラスマイナスの記号までです）。なぜなら、この分析が予測しているのは曲線的な関係であり、Ｘの値によって、ＸがＹに与えるインパクトは変化してしまうからです。この理由はFigure14を見ても明らかでしょう。

　よって、ここまでの知識でも（例）のモデルでは投入した全ての変数が有意な関係を投票方向と持っていることが分かります。また、係数の方向も*DPJsup*以外はプラスの方向を示していて、これも予想通りです。仮説は支持されたといえるでしょう。ただし、係数の大きさや直接の解釈は、ここではできません。次の節以降では、「Ｂ」の解釈や、新しく出てきた「Exp(B)」の解釈、そして、予測確率の計算・示し方について述べます。

Figure 15: voteLDPを予測するロジット分析結果（一部省略）

ケース処理の要約

重み付きのないケース[a]		N	パーセント
選択されたケース	分析で使用	709	50.8
	欠損ケース	688	49.2
	合計	1397	100.0
選択されなかったケース		0	.0
合計		1397	100.0

a. 重み付けが有効な場合には、ケースの総数について分類表を参照してください。

ブロック 1: 方法 = 強制投入法

モデル係数のオムニバス検定

		カイ 2 乗	自由度	有意確率
ステップ 1	ステップ	453.446	5	.000
	ブロック	453.446	5	.000
	モデル	453.446	5	.000

モデル集計

ステップ	-2 対数尤度	Cox-Snell R2 乗	Nagelkerke R2 乗
1	494.381[a]	.472	.641

a. パラメータ推定値の変化が .001 未満であるため、反復回数 6 で推定が打ち切られました。

分類テーブル[a]

観測			予測: voteLDP Voting Direction 2005 HoR PR .00 Vote for DPJ	予測: voteLDP Voting Direction 2005 HoR PR 1.00 Vote for LDP	正解の割合
ステップ 1	voteLDP Voting Direction 2005 HoR PR	.00 Vote for DPJ	218	58	79.0
		1.00 Vote for LDP	34	399	92.1
	全体のパーセント				87.0

a. 分類値は .500 です

方程式中の変数

		B	標準誤差	Wald	自由度	有意確率	Exp(B)
ステップ 1[a]	postatt	.162	.076	4.522	1	.033	1.176
	ideology	.206	.073	7.871	1	.005	1.229
	LDPsup	.623	.137	20.714	1	.000	1.865
	DPJsup	-1.149	.164	49.037	1	.000	.317
	koizumisup	.895	.147	36.943	1	.000	2.448
	定数	-3.765	.577	42.525	1	.000	.023

a. ステップ 1: 投入された変数 postatt, ideology, LDPsup, DPJsup, koizumisup

6.3.2　ロジスティック回帰式の仕組みを理解する　*Understand Logit Function*

予測値を求める方法（最尤法）の理解についてはここでは置いておいて、分析結果の解釈についてここでは述べます。ロジスティック回帰分析が予測した曲線の式は、以下の様に書けます。(変数の数=*a*の時。予測値なので、残差は式に入っていません。)

$$Pr(\hat{Y}_i = 1) = \Lambda(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_a X_{ai}) \qquad (2)$$

見た目重回帰分析とほとんど同じなのですが、予測されているのはΛで表現されているロジット関数であり、直線ではないので、係数をそのまま解釈することは出来ません。ここで、$\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_a X_{ai}$の部分をまとめて$X_i\hat{\beta}$と置きます。すると、次の様に式が書きかえられます。

$$Pr(\hat{Y}_i = 1) = \Lambda(X_i\hat{\beta}) = \frac{exp(X_i\hat{\beta})}{1 + exp(X_i\hat{\beta})} \qquad (3)$$

ここで、一番知りたい（と思われる）$Pr(\hat{Y}_i = 1)$の値は、「$X_i\hat{\beta}$の指数（exponential）を取った数」を、「1＋$X_i\hat{\beta}$の指数を取った数」で割ると求められます。これがFigure14右側にある曲線（実線）を作り出しているロジット関数です。

6.3.3　オッズ比Exp(B)を解釈する *Interpret Odds Ratio*

SPSSでロジット分析を行うと、係数の横に Exp(B)という値を、デフォルト出力してくれます。これがオッズ比です。このオッズ比の意味合いは、式(3)の変形の、以下の式(4)から導けます。

$$\begin{aligned}
Pr(\hat{Y}_i = 1) &= \frac{exp(X_i\hat{\beta})}{1 + exp(X_i\hat{\beta})} \\
Pr(\hat{Y}_i = 1) * [1 + exp(X_i\hat{\beta})] &= exp(X_i\hat{\beta}) \\
Pr(\hat{Y}_i = 1) + Pr(\hat{Y}_i = 1) * exp(X_i\hat{\beta}) &= exp(X_i\hat{\beta}) \\
Pr(\hat{Y}_i = 1) &= exp(X_i\hat{\beta}) - Pr(\hat{Y}_i = 1) * exp(X_i\hat{\beta}) \\
Pr(\hat{Y}_i = 1) &= exp(X_i\hat{\beta}) * [1 - Pr(\hat{Y}_i = 1)] \\
\frac{Pr(\hat{Y}_i = 1)}{1 - Pr(\hat{Y}_i = 1)} &= exp(X_i\hat{\beta}) \qquad (4)
\end{aligned}$$

ここで、$X_i\hat{\beta}$は、$\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+...+\hat{\beta}_aX_{ai}$のことを示しているのでした。よって、**他の値は全て変化しないとして、例えばX_{1i}の値だけが１大きくなると、$X_i\hat{\beta}$は、$\hat{\beta}_1$分だけ大きく（$\hat{\beta}_1$の値が負の場合は小さく）なる**ことが分かります。

また、**オッズ比が意味するのは、Yが0になる確率に対する、Yが１になる確率の比率**です。値が１よりも大きくなれば、より$Y=1$になりやすいことを意味しますし、値が１よりも小さければ、よりY=0になりやすいことを意味します。例えば、オッズ比が1.5であれば「Ｙが１である確率は、Ｙが０である確率の1.5倍」という解釈になります。

上記をまとめると、$exp(\hat{\beta}_1)$**は、他の値は全て変化しないとして、X_{1i}が１大きくなった際の「Yが0になる確率に対するYが１になる確率の比率」の変化を表している**訳です。よって、例のモデルでは、*postatt*の$exp(\hat{\beta})$が1.17ですので、「*postatt*の値が１大きくなる（郵政民営化に積極的になる）ごとに、回答者は、1.17倍（もしくは117%）、民主党に比べて自民党に投票しやすくなる」ということを示しています。民主党支持については、「*DPJsup*の値が１大きくなる（民主党支持が強くなる）ごとに、回答者が民主党に対して自民党に投票する可能性は0.31倍（31%)になる」という理解になるでしょう。（もし、発病などを１と定義するならば、「リスク」という表現を使う場合もあります。）

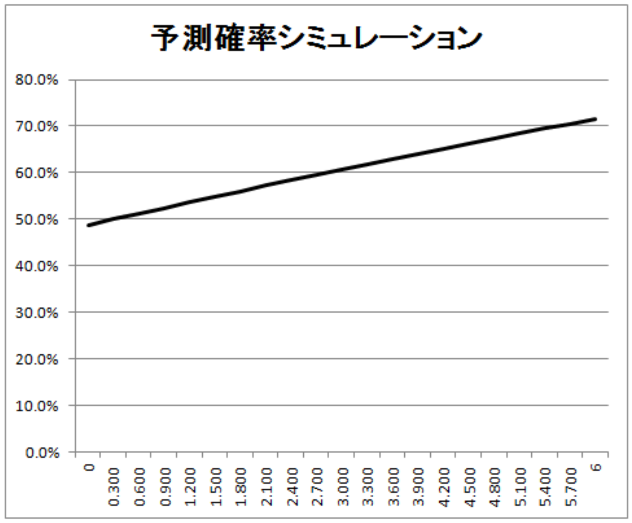
6.3.4　予測確率の計算 *Calculate Predicted Probabilities*

「オッズ比では、なかなか直観的に結果が理解できない」という人もいると思います。そんな人は、式(3)に従って、各Xの値におけるＹの予測確率を計算してしまいましょう。SPSSにはそのようなオプションがない（はず）なので、手動で簡単に計算を行えるファイルをExcelで用意しました。「ロジット予測確率計算 140724.xls」を開いてください。

まずは、ロジットの予測確率を計算してみます。「予測確率」シートを開いてください。表の指示に従って、係数を入力していってください。値については、まず「postatt=3,ideology=5,LDPsup=0,DPJsup=0,koizumisup=3」で固定してみます。すると、予測確率が60.8%と出ます。この数値は、式(3)に従って、予測値「X Beta」をロジット変換したものです。すなわち、郵政民営化に対する意見とイデオロギーが中立で、自民も民主も支持しておらず、小泉内閣を「ある程度支持」している人の、予測自民党投票確率は、60.8%だということです（ただし、分析上自民・民主以外の投票先は考慮していません）。固定する値を入れ替えることで、あらゆる条件で予測をすることができます。

更に、今回の独立変数の1つである、郵政民営化に対する意見が自民党投票に与えたインパクトをシミュレーションしてみます。「予測確率シミュレーション」シートを開いてください。「予測確率」シートとほぼ同じなのですが、X1の値を変化させて、予測確率をその都度計算できるようにしてあります。まずは、「予測確率」と同じように数値を入力します（固定する値も同じです）。入力できたら、インパクトを見たい変数「X1」を変化させる範囲を指定します。現在は、X1がpostattになっていますね。postattは０から６の値をとる変数なので、最小値を0、最大値を６に指定してみましょう。すると、各Ｘ1の値における自民党投票の予測確率が表示されます。表を見ると、郵政民営化に対する意見が「０＝消去的」から「6＝積極的」に変化すると、自民党投票の予測確率が48.8%から71.6%まで上昇することが分かります。かなり大きな変化ですね。

Figure 16: postattの値ごとのvoteLDP予測確率シミュレーション



最終的に予測確率の結果を示す方法としては、大まかに2通りが考えられます。１つ目は、**予測確率のシミュレーション結果を、グラフにして示す**方法です。「予測確率シミュレーション」シートに用意したグラフ(Figure16)はその一例です。一目で見て、どのくらい変化するのかが分かりやすいですね。一方で、結果中の複数の変数に関心がある場合には、グラフの数が増えてしまい、見にくくなることもありそうです。よって2つ目は、**Xの値を特定範囲動かした時のＹ変化量の絶対値を示す**方法です。今回のExcelでは「First Difference」と書いてある場所に、その数値が示されています。この数値を使えば独立変数が従属変数に与える効果を1つの数字でまとめることができて便利です。ただし、数値を算出する時に、どの範囲を選ぶかで、変化量は異なる値を示します（ロジット関数は曲線ですからね）。First Differenceを並列で並べて、インパクトを論じる際には、各変数で選択した変化範囲がなぜ比較可能なのかについて、正当化する必要があるでしょう。(結果を示した例：Table6)

6.3.5　モデル全体を評価する　*Evaluate the Model*

分析の係数の表に載っているもの以外で、モデル全体を評価するものとして最終的な結果に含めた方が良い（かもしれない）指標は以下の通りです。

- **カイ2乗検定**:「モデル係数のオムニバス検定」を参照。p<0.05であれば、意味のあるモデル（意味のないモデルとはいえない）ということです。Ｆ検定みたいなものです。
- **-2*対数尤度(Log-Likelihood)**：「モデル集計」の中にあります。モデルのあてはまりのよさを示す指標の一つです。これといって基準はありませんが大きい方が当てはまりがよくなります。
- **R2乗**：「モデル集計」の中にあります。様々な求め方があり、SPSSでは2つ出力されます。かなり数値が違うこともあるので、結構指標としては怪しいです。また、

重回帰分析に比べて非常に低い数値が出ることが多いですが、あまり気にしなくても大丈夫です。まあ、とりあえず表には載せることが多いです。

- **正確に予測された%（ % Correctly Predicted）**：ブロック１の方にある「分類テーブル」の「全体パーセント」の行の「正解の割合」の列を見ます。この値は、予測確率を50%で分割して、50%以上を１、50%未満を０としてケースの値を予測した時に、その予測値が実際の値とどのくらい一致していたかを示しています。ブロック0の「分類テーブル」上にある「全体パーセント」は、Ｙの平均値で予測した場合の正解率だということに注意してください。
- **ケースの数**：最初に出力される「ケース処理の要約」と書いてある表の「選択されたケース　分析で使用」を参照します。

6.3.6 プロビット分析？*What about Probit ?*

ロジット分析の仲間にプロビット分析というものがあります。この分析はロジット分析であてはめたロジスティック分布ではなく、正規分布を使って曲線をあてはめます。正規分布を使うという意味では、より重回帰分析などにコンセプトは近いといえるでしょう。この分析であてはめた曲線（実はFigure14右側の点線と同じです）の式は、以下の様になります。

$$Pr(\hat{Y}_i = 1) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_a X_{ai}) \tag{5}$$

$$Pr(\hat{Y}_i = 1) = \Phi(X_i\hat{\beta}) = \int_{-\infty}^{X_i\hat{\beta}} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{(X_i\hat{\beta})^2}{2}\right) dX_i\hat{\beta} \tag{6}$$

関数の式がやや複雑になっていますが、式(2),(3)と見比べると、基本的な考え方はロジット分析と同じだということが分かると思います。ただし、関数が違うのでオッズ比の解釈は使えませんし、出てくる係数の値も違います。一方で、予測確率を計算すると、ほぼ同じ値が出るはずです（興味がある方は、今回お渡しした「ロジット予測確率計算 140724.xls」の関数を書き換えてみてください。$= NORMSDIST(X_i\hat{\beta})$と入力すると簡単に結果が出ます）。ロジットとプロビットの間には、実質上優劣がありませんので、どちらを使うかは研究者の好みです。

7　条件付け効果を分析する　*Conditional Effects*

研究の仮説を立てる中で、**XがYに与えるインパクトが「ある条件Zがある時のみ」または「ある条件Zが満たされている程度に従って」変わる**、というロジックを考えることがあると思います。このようなロジックを条件付け効果（Conditional Effect）と呼びます。この節は、そんな条件付け効果仮説を検証するための方法をいくつか紹介します。

7.1　サブセットに分けて分析 *Subsetting Samples*

条件付け効果を考慮した仮説の、一番基本的な検証方法は、条件によってサンプルを分割して分析することです。以下、ロジット分析の例の続きでやってみましょう。

※GLOPE0507を使った例

■理論：2005年の衆院選は、郵政選挙といわれ、伝統的なイデオロギーを無視して郵政民営化政策だけで投票先が決まったかのように思える。しかし、それは、自民党と民主党の間の郵政民営化政策に対する立場の違いを認識していた人のみ取れた行動なのではないか。もし立場の違いという手掛かりを認識できていなかったならば、伝統的な価値観を手掛かりにして投票を決めるしかなかったのではないか。

仮説1.5: 自民党と民主党の郵政民営化に対する立場の違いを認識していた人は、郵政民営化に強く賛成している人ほど自民党に投票したが、自民党と民主党の郵政民営化に対する立場の違いを認識していなかった人は、必ずしも郵政民営化に強く賛成している人ほど自民党に投票した訳ではない。

仮説2.5: 自民党と民主党の郵政民営化に対する立場の違いを認識していた人は、伝統的なイデオロギーに頼らず投票先を決めたが、自民党と民主党の郵政民営化に対する立場の違いを認識していなかった人は、イデオロギー的に保守である人ほど、自民党に投票した。

■条件付けに使う変数（*Recode_GL0507.sps*で作成済み）：
postknow⇒
自民党が民主党よりも郵政民営化に積極的であると認識している=1、認識していない=0

7.1.1　結果を出力してみる

メニューから：＜データ＞⇒＜ケースの選択＞
各分析を行う前に、条件付け変数によって分析を適用するサンプルを指定します。
分析結果の出し方は、普通の分析を行う時とほぼ同じです。
分析が終わったら、全てのケースを対象に戻しておくことを忘れずに。

```
Use all.
Compute Filter_x = (postknow = 1).
Filter BY Filter_x.
```

```
LOGISTIC REGRESSION
VARIABLES voteLDP
/METHOD=ENTER postatt ideology LDPsup DPJsup koizumisup
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).

Use all.
Compute Filter_x = (postknow = 0).
Filter BY Filter_x.

LOGISTIC REGRESSION
VARIABLES voteLDP
/METHOD=ENTER postatt ideology LDPsup DPJsup koizumisup
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).

Use all.
```

新しいコマンド

Use all （全てのケースを分析に含める）
Filter (BY の後に入力された変数が 1 のケースのみ分析に含める)

Figure 17: voteLDPを予測するロジット分析結果（係数表以外省略）

方程式中の変数

		B	標準誤差	Wald	自由度	有意確率	Exp(B)
ステップ 1[a]	postatt	.198	.092	4.679	1	.031	1.219
	ideology	.154	.086	3.212	1	.073	1.167
	LDPsup	.711	.169	17.732	1	.000	2.037
	DPJsup	-1.116	.195	32.654	1	.000	.328
	koizumisup	.850	.177	22.958	1	.000	2.340
	定数	-3.652	.689	28.063	1	.000	.026

a. ステップ 1: 投入された変数 postatt, ideology, LDPsup, DPJsup, koizumisup

(a) postknow=1の時の結果

方程式中の変数

		B	標準誤差	Wald	自由度	有意確率	Exp(B)
ステップ 1[a]	postatt	.058	.145	.161	1	.688	1.060
	ideology	.389	.155	6.297	1	.012	1.476
	LDPsup	.460	.241	3.651	1	.056	1.583
	DPJsup	-1.218	.315	14.983	1	.000	.296
	koizumisup	.988	.278	12.641	1	.000	2.686
	定数	-4.266	1.139	14.035	1	.000	.014

a. ステップ 1: 投入された変数 postatt, ideology, LDPsup, DPJsup, koizumisup

(b) postknow=0の時の結果

7.1.2　結果を解釈する

サブセットに分けた場合、結果の解釈は基本的に全体で分析した時と同じです。上記の結果では、仮説通り、postattはpostknow=1の時だけ、ideologyはpostknow=0の時だけ、有意（p<.0.05)に自民党投票確率にインパクトを持っていることが分かります。全体の分析に加えて、新しい知見が提供できていますね。興味のある方は、予測確率シミュレーションも行ってみてください。サブセットによって、曲線の形や位置が変わることに気づくはずです。

　サンプルを分割する方法は、シンプルで有効です。ただし、いくつかの問題点があります。一つは、分割によってサンプルの数が減ってしまうことです。サンプル数が小さくなりすぎてしまうと、結果が不安定になってしまいます。二つは、条件付け効果が1つの分析で確認できないことでしょう。三つは、前の2つとも関連しますが「ある条件が満たされている程度に従って」変わるといったような、連続変数による条件付け効果の分析がとても難しいことです。値がたくさんあるからといってサンプルを無限に分割してしまっては、分析ができません。そのような仮説を持っている場合のために、1つの分析で条件付け効果が分析できるような方法は、次の節で紹介します。

7.1.3　サブセット間の構造変化の有無を検定する（OLSのみ）*Chow Test*

２つのサブセットで分析を行ったら、「全体の分析に対して、２つのサブセットに分割したことが本当に予測を改善するのか」という問いが生じることでしょう。この問いに統計的に答えるのがChow Testです。この検定は、F検定の応用となっており、モデル間で残差の平方和の大きさを比べます。この検定の帰無仮説は「全体サンプルのモデルと、サブセットに分けたモデル群の間で、予測力に違いは無い」ということになります。サブセットに分けると予測力は必ず改善しますから、**有意確率p<0.05で帰無仮説が棄却されれば「サブセットに分けて分析した方が、全体で分析するより、モデルの予測力が高くなる」といえます**。統計量の計算式は次の通りで、これをＦ分布にあてはめて検定を行います。（SSE＝各モデルの残差の平方和、ｇ＝サブセットの数、ｋ＝独立変数＋統制変数の数、ｎ＝全体のケース数）

$$F^{(g-1)(k+1),n-g(k+1)} = \frac{\frac{SSE_{full}-\sum_{i=1}^{g} SSE_{subseti}}{(g-1)(k+1)}}{\frac{\sum_{i=1}^{g} SSE_{subseti}}{n-g(k+1)}} \qquad (7)$$

　Excelの「条件付け効果の算出 140727.xls」に、Chow-Testの結果が出せるシートを用意したので、興味のある方は、試してみてください。ただし、**この検定が適用できるのはOLSの分析結果のみです。二値変数を従属変数とした分析では、あてはめる曲線（ロジット・プロビット）によってSSEの大きさが変わってしまうなどの問題があるので、この検定を使用するのは不適切**だといわれています。もちろん二値変数を含むモデル同士でも検定方法はある様ですが、（まだ私が完全に理解できていない部分もあるので）ここでは紹介しません。

7.2　交互作用項を使う *Use Interactions*

交互作用項は、独立変数と条件付け変数を掛け合わせた変数です。例えば、Ａという独立変数のインパクトが、Ｂという条件付け変数の値によって変化するという仮説がたてられるのであれば、「Ａ＊Ｂ」という交互作用項を分析に含めるわけです。この方法を使えば、1つの分析の中で条件付け効果が分析できますし、サンプルを分ける必要が無いので、連続的な変数による条件付け効果を分析することも可能です。ただし、解釈がやや面倒くさいので、以下では、単純な重回帰分析の式を使いながら、交互作用項の仕組みについて説明します。（ロジット分析でも仕組みは同じなのですが、ロジット変換を使うために、直接的な解釈はさらに難しくなります。）

7.2.1　交互作用項の仕組みについて理解する *Understand the role of Interactions*

式(8)は、一番単純な交互作用項モデルとして、独立変数Ａと条件付け変数Ｂ、そして交互作用項ＡＢを含めた重回帰分析モデルの予測式です（よって残差などは含まない）。もちろん、加えて幾つか統制変数を含んでいても、基本的な解釈は変わりません。

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_A X_{Ai} + \hat{\beta}_B X_{Bi} + \hat{\beta}_{AB} X_{Ai} X_{Bi} \tag{8}$$

このモデルでは、交互作用項なしのモデルの様に、単純に係数の意味を解釈することはできません。関心のあるX_Aは、2つの変数の中に存在しているので、1つの変数だけでは「全体として」X_AがＹに与えるインパクトを表現できないからです。よって、式(8)を次の様に変形します。

$$\hat{Y}_i = \hat{\beta}_0 + (\hat{\beta}_A + \hat{\beta}_{AB} X_{Bi}) X_{Ai} + \hat{\beta}_B X_{Bi} \tag{9}$$

ここで、(　)内の値が「全体として」X_AがYに与えるインパクトの大きさになる訳です。
X_A**の「全体としての」インパクトの大きさが、**X_B**の値の大きさに「条件付け」されて変化する**ことが分かると思います。そして同時に、$\hat{\beta}_A$の値は、「X_Bの値が0である」という特別な条件下でのX_Aのインパクトの大きさであることが分かるでしょう。よって、**一番興味があるのは、（　）内の係数が有意にＹに影響を与えているかどうかであって、一つ一つの**$\hat{\beta}_A$**や**$\hat{\beta}_{AB}$**の係数は二次的な情報でしかない**ことが分かります。（ただし、式を見れば分かりますが、X_AやX_Bが二値変数である場合にはそのまま解釈できる場合もあります）この「全体としての」係数の有意性の求め方については後ほど説明します。

話題は変わりますが、「交互作用項を入れたなら、単独のX_AやX_Bは分析に含める必要はあるのか」という疑問が出てくると思います。ここで、「分析に入れないことが持つ意味」について考えてみます。まず、X_Aについては、前段でも述べたことを考慮すると、分析に入れないということは、「X_Bの値が0である時には、X_AがＹに与えるインパクトは０である」という非常に強い前提を置いていることになります。100%の確信をもって、

この前提が真であるということがいえますか？もし「理論的に」上記の前提が成立しないのであれば、含めておいた方が良いでしょう。X_Bについては、以下の式を見てください。

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_A X_{Ai} + (\hat{\beta}_B + \hat{\beta}_{AB} X_{Ai}) X_{Bi} \tag{10}$$

この式(10)は、式(9)と同じく、式(8)の変形です。交互作用項は「二重の役割を担っている」ことが分かります。ここでの$(\hat{\beta}_B$が持つ意味は、式(9)で$(\hat{\beta}_A$が持つ意味とパラレルです。すなわち、X_Bを分析に含まないということは、「X_Aの値が0である時には、X_BがＹに与えるインパクトは０である」という、これまた強い前提を置いていることになる訳です。以上からまとめると、**単独のX_AやX_Bは、「含めるべきではない」という強い理論が無い限り、分析に含めておくことが妥当**、といえるでしょう。

7.2.2　結果を出力してみる

※dcd09を使用した分析例（基本の枠組みは5.1の重回帰分析の時と同じ）：

仮説1.5：民主化度が高い国ほど、安全な水にアクセスできる人の割合が増えるという関係は、経済的に豊かな国ほど弱くなる（経済的に貧しい国ほど強くなる）。

```
*/Create Interaction Variable.
Compute fhrev_logGDP = fhmean1990_1999rev*logGDP2000.

REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA BCOV
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT Water2000
/METHOD=ENTER fhmean1990_1999rev logGDP2000 fhrev_logGDP.
```

追加記入箇所（その他は重回帰分析と一緒）

/STATISTICS　に　BCOV（係数の分散・共分散マトリックス）を追加

7.2.3　結果を解釈する

7.2.1でも述べた通り、交互作用項を入れたモデルでは、係数をそのまま解釈することが基本的には出来ません。（ただし、交互作用項に０と１の値を持つ二値変数を含むモデルでは、そのまま解釈できる場合もあります。）そこで、式(9)に書かれている様に、X_{Ai}の総合効果を理解するには、$\hat{\beta}_A + \hat{\beta}_{AB} X_{Bi}$を計算する必要がある訳です。今回もSPSSには直接この数値を計算する機能がない（or面倒くさい）ので、Excelを使って示したいと思います。条件付け効果の算出 *140727.xls*を開いてください。（今後表記が入り混じりますが、X_AとX1、X_BとX2、X_{AB}とX1*2は同じものだと考えてください。）

Figure 18: 交互作用項を含めてWater2000を予測する（一部省略）

モデル集計

モデル	R	R2 乗	調整済み R2 乗	推定値の標準誤差
1	.744[a]	.553	.541	13.52951

a. 予測値: (定数)、fhrev_logGDP, logGDP2000, fhmean1990_1999rev。

分散分析[b]

モデル		平方和 (分散成分)	自由度	平均平方	F 値	有意確率
1	回帰	25341.721	3	8447.240	46.148	.000[a]
	残差 (分散分析)	20501.339	112	183.048		
	合計 (ピボットテーブル)	45843.060	115			

a. 予測値: (定数)、fhrev_logGDP, logGDP2000, fhmean1990_1999rev。
b. 従属変数 Water2000 Population with sustainable access to an improved water source (%) 2000 (UNDP 2004)

係数[a]

モデル		標準化されていない係数		標準化係数	t 値	有意確率
		B	標準誤差	ベータ		
1	(定数)	-68.172	27.395		-2.488	.014
	fhmean1990_1999rev	13.764	5.916	1.210	2.326	.022
	logGDP2000	16.507	3.397	.907	4.860	.000
	fhrev_logGDP	-1.320	.685	-1.222	-1.928	.056

a. 従属変数 Water2000 Population with sustainable access to an improved water source (%) 2000 (UNDP 2004)

まず、今回のモデルにおいて興味がある単独のX_Aは、fhmean1990_1999revですから、その係数である13.764をX1の欄に入力します。そして、交互作用項の係数である-1.320をX1*2の欄に入力します。ここで、X1の総合効果はX2に条件付けされて変化する訳ですから、X2であるlogGDP2000を動かす範囲を設定します。その準備としてまずはlogGDP2000の記述統計を確認します。

```
Descr logGDP2000.
```

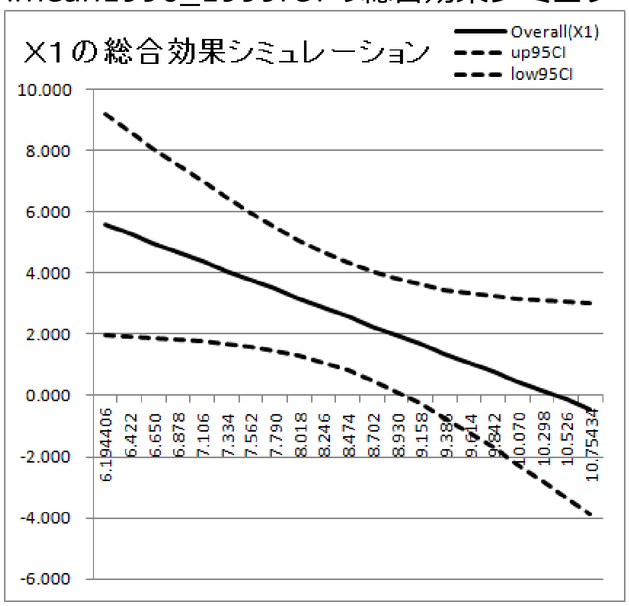
そうすると、logGDP2000の最小値は6.194、最大値は10.754であることが分かりますから、とりあえずその2つの値をX2の最小値と最大値として入力します。すると、Excelが勝手に計算をして$\hat{\beta}_A+\hat{\beta}_{AB}X_{Bi}$の結果が表とグラフに現れるはずです。これが、logGDP 2000に条件付けされた、fhmean1990_1999revがWater2000に対して持つインパクトです。logGDP2000の値が上昇するに従って、係数が小さくなることが明らかに見てとれると思います。

それでは、どの時点でfhmean1990_1999revがWater2000に与える効果は統計的に有意でなくなるのでしょうか。これを知るためには、総合効果の標準誤差を計算する必要があります。計算式は次の通りです。

$$SE(\hat{\beta}_A + \hat{\beta}_{AB}X_{Bi}) = \sqrt{Var(\hat{\beta}_A) + {X_B}^2 * Var(\hat{\beta}_{AB}) + 2X_B * Cov(\hat{\beta}_A, \hat{\beta}_{AB})} \qquad (11)$$

この式を計算するためには、係数の値の分散・共分散マトリックスが必要です。SPSSで今回新しく出力した「相関係数」表から、必要な分散と共分散の値を取り出して、Excelファイルの適当な欄に貼り付けてください。すると、式(11)に従って標準誤差が計算され、標準誤差×(両側から2．5%のt値 ≈ 1.96)の範囲で95%信頼区間が計算されて図にプロットされます。結果をFigure19に示しました。fhmean1990_1999revがWater2000に与える効果は、logGDP2000が上がるにつれて小さくなり、大体logGDP2000が9を超えるあたりで、統計的に有意でなくなることが分かりますね。具体的に言うと、一番貧しい国では民主化度が1から7まで上昇することによって安全な水にアクセスできる人の割合が30%以上も上昇するのに対して、一番豊かな国では、ほぼその効果が0%になるということが分かります。相互作用項自体は、$0.05<p<0.10$ではありますが、仮説1.5は、おおむね確認されたと言えそうです。

Figure 19: fhmean1990_1999revの総合効果シミュレーション



最終的に相互作用項を含むモデルの結果をレポートする際には、重回帰分析の表に加えて、上記の様な総合効果のシミュレーショングラフや、X2を動かした時の係数と標準誤差の変化をまとめた表（例：Table7）などが用意できると、より内容が伝わりやすくなりますし、解釈もしやすくなります。

8　時系列データを分析する *Time Series*

分析に関する、本演習最後のトピックは時系列データの分析です。ここで時系列データとは、同じ人・国・対象物に関して、複数の時点(t)のケースを含んでいる様な構造をしたデータを指します。今回でいうと、「prefpol」がそれにあたります。

この様なデータを扱う場合には、今までの分析に加えて自己相関(autocorrelation)と呼ばれる問題が生じます。これは、隣り合っている（または近くにある）ケース同士の値が関係しあってしまう、という現象です。もっと具体的に言えば、prefpolのデータにおいて、ある都道府県におけるt時点の値が、同じ都道府県のt-x時点（例えばt-1時点）の

値と関連を持ってしまっているということです。**自己相関を無視して分析を進めると、分析結果に深刻なバイアスをもたらす恐れがあります**。例えば、人口と平均収入の間に強い相関関係を見つけたとしても、そのどちらもが、時系列的な人口増加と収入の増加の影響による偽相関である可能性なども考えられるからです。

自己相関の問題を解決するために分析方法論の最先端で現在様々な手法が議論されていますが、ここではその効果を統制するための非常に基本的な２つの方法を紹介します。興味がある方は、更に自分で学びを深めてみてください。

8.1　固定効果モデル

固定効果モデルは、時系列データの問題点を統制する、一番初歩的な手法です。このモデルでは、分析に入っている全ての時点t-1個と、分析に入っている全ての対象（人・国・都道府県など）m-1個のダミー変数を統制変数として分析に投入してしまうことで、各時点における特徴および各対象固有の特徴を全て取り去ってしまおうという手法です。このモデルで、**分析対象として残るのは各対象の中における、各時点の特徴に左右されない様な変化だけ**になります。対象ごと、年ごとの特徴を全てダミーで拾ってしまうため、分析としてはかなり厳しいテストを行っているといっても良いと思います。このモデルで効果が出るのであれば、かなり深い関係を持っていると言えるのではないでしょうか。ただし、対象間での差を分析できなくさせてしまうことや、時間を流れではなく各時点でコントロールしてしまうことなど、モデルとしてはいくつかの問題点を抱えています。

結果の解釈については、重回帰分析と変わりませんので、ここでは割愛します。表として結果を発表する際には、時点ダミーや対象ダミーの部分を取り除いて、「（以下各時点ダミー）」などと表記してもらえれば大丈夫かと思います。（例：Table5）

8.1.1　結果を出力してみる

※prefpolを使った例

■RQ：地方議会における自民党議員の多さはどのような背景によって説明されるか。
仮説：農林水産業に従事する人が多い地域と年である程、地方議会において自民党議員が多い。

■ 使う変数：
ldpprop ⇒　t時点の各都道府県議会における自民党所属議員の割合
pop_indus01 ⇒ t時点の各都道府県における第1次産業人口割合
popdens ⇒　t時点の各都道府県における人口密度
year* ⇒　年ダミー 1976－2006（1975年が参照カテゴリー）
pref* ⇒　都道府県ダミー（北海道が参照カテゴリー）

```
Compute ldpprop = asb_ldp / asb_sosu.
Compute popdens = pop_total/geo_size.
```

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT ldpprop
 /METHOD=ENTER pop_indus01 popdens
 year1976 year1977 year1978 year1979 year1980 year1981
 year1982 year1983 year1984 year1985 year1986 year1987 year1988
 year1989 year1990 year1991 year1992 year1993 year1994 year1995
 year1996 year1997 year1998 year1999 year2000 year2001 year2002
 year2003 year2004 year2005 year2006
 pref02 pref03 pref04 pref05 pref06 pref07 pref08 pref09 pref10
 pref11 pref12 pref13 pref14 pref15 pref16 pref17 pref18 pref19 pref20
 pref21 pref22 pref23 pref24 pref25 pref26 pref27 pref28 pref29 pref30
 pref31 pref32 pref33 pref34 pref35 pref36 pref37 pref38 pref39 pref40
 pref41 pref42 pref43 pref44 pref45 pref46 pref47.
```

Figure 20: 固定効果を含めてldppropを予測する（係数表の固定効果変数を省略）

モデル集計

モデル	R	R2 乗	調整済み R2 乗	推定値の標準誤差
1	.841	.707	.691	.07587

分散分析[b]

モデル		平方和 (分散成分)	自由度	平均平方	F 値	有意確率
1	回帰	19.821	79	.251	43.591	.000[a]
	残差 (分散分析)	8.196	1424	.006		
	合計 (ピボットテーブル)	28.018	1503			

a.b. 従属変数 ldpprop

係数[a]

モデル		標準化されていない係数		標準化係数	t 値	有意確率
		B	標準誤差	ベータ		
1	(定数)	.434	.030		14.556	.000
	pop_indus01 産業別人口割合：第1次産業　(単位%)	.004	.001	.179	2.760	.006
	popdens	.000	.000	.927	2.882	.004
	year1976	.023	.016	.030	1.488	.137
	⋮	⋮	⋮	⋮	⋮	⋮
	year2006	-.073	.024	-.093	-3.069	.002
	pref02	.065	.022	.069	2.932	.003
	⋮	⋮	⋮	⋮	⋮	⋮
	pref47	-.144	.028	-.152	-5.168	.000

a. 従属変数 ldpprop

8.2　ラグ従属変数モデル

ラグ従属変数モデルでは、時点ダミーや対象ダミーを投入する代わりに、t-x時点（ここではt-1時点）の従属変数の値を分析に投入します。同じ都道府県のすぐ前の時点の従属変数の値が統制されるわけですから、残りの**説明対象は各t時点とt-1時点の間の差（残差）となる**訳です。具体的には次の図を見てください。

■ 新たに使う変数： lag1ldpprop ⇒　t-1の年における地方議会の自民党所属議員割合

```
Compute lag1ldpprop = Lag(ldpprop)

GRAPH
/SCATTERPLOT(BIVAR)= lag1ldpprop with ldpprop
/MISSING=LISTWISE.
```

新しいコマンド

Lag ( )内の変数において1つ前のケースに入っている値を取得する。

Figure 21: lag1ldppropでldppropを予測した残差が分析対象

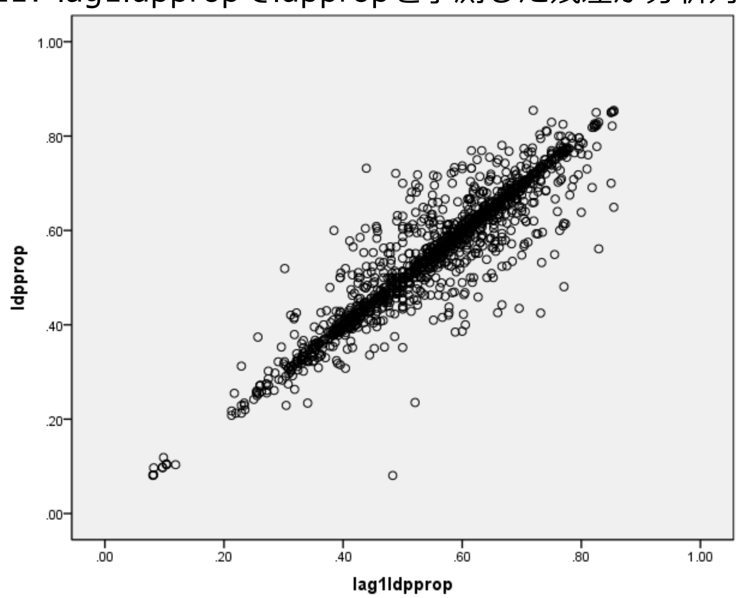


Figure21を見ると、lag1ldppropはldppropと非常に強い線型関係にありますね。一方で、2つの変数はお互いの全てを説明し合っている訳ではなく、中心の線からある程度のばらつき（残差）が存在することが分かります。このバラつきを自分が投入した独立変数によって説明しよう、というのがラグ従属変数モデルの考え方です。

このモデルは、自己相関を統制するという意味では理にかなったモデルになっていますが、課題もあります。１つ紹介すると、t時点の値を説明してしまうのはt-1時点の値だけなのか、という点があります。加えてt-2時点、t-3時点などの値も効果を与えていると

すれば投入する必要があるでしょう。また、例えば、月次データで「ショートケーキの売上」といったような従属変数を扱っている場合に、毎年12月だけ売り上げが激増する場合があるとすれば、t-12のラグ変数を投入する必要があると考えられます。また、時点によって与える影響が違うのであれば、ラグ変数同士に重み付けをするという考え方もあります。最小二乗法で推定すると適切な結果が出力されない、という話もあります。この分野はまだまだ奥深く発展していっていますが、結局重要なのは、自分の想定している理論がどのような関係を予測しているのかということなのは、変わりません。

これも、解釈については、重回帰分析とほぼ一緒なので説明を割愛します。

8.2.1　結果を出力してみる

※prefpolを使った例　(仮説は8.1.1と同じ)

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT ldpprop
/METHOD=ENTER pop_indus01 popdens lag1ldpprop.
```

Figure 22: ラグldppropを含めてldppropを予測する

モデル集計

モデル	R	R2 乗	調整済み R2 乗	推定値の標準誤差
1	.924[a]	.853	.853	.05238

a.

分散分析[b]

モデル		平方和 (分散成分)	自由度	平均平方	F 値	有意確率
1	回帰	23.141	3	7.714	2811.525	.000[a]
	残差 (分散分析)	3.986	1453	.003		
	合計 (ピボットテーブル)	27.128	1456			

a.b. 従属変数 ldpprop

係数[a]

モデル		標準化されていない係数		標準化係数	t 値	有意確率
		B	標準誤差	ベータ		
1	(定数)	.039	.007		5.892	.000
	pop_indus01 産業別人口割合：第1次産業　(単位%)	.001	.000	.031	2.551	.011
	popdens	-1.198E-6	.000	-.009	-.794	.427
	lag1ldpprop	.911	.011	.907	80.631	.000

a. 従属変数 ldpprop

※prefpol.savを使ってラグ変数を作る際に、もしasbcandから始まる選挙候補者データを使う場合は、「前回選挙」というラグを使いたいこともあると思います。その場合は、まず、asbcandデータの欠損ケースをデータセットから削除してしまいましょう。Select if (asbcand_data=1).と入力するとasbcandデータが欠損しているケースをデータから削除できます。（ただし、一回削除すると元には戻せないので注意してください。）削除後にラグ変数を作成すれば、「前回選挙」のラグになるはずです。

9　分析結果を保存・発表する *Save and Present Results*

9.1　分析結果を保存する *Save Results*

9.1.1　データセットの変更を保存する

元のデータセットに対して変更を行ったら、**こまめに保存**します。また、元のデータセットとは**別の名前で保存**するのも重要です。元のデータを残しておくことで、どのように変更したかをたどれるようにします。保存方法としては主に以下の２つがあります。

- **データセット本体を保存する**：データセットを表示させた画面から＜ファイル＞⇒＜名前をつけて保存＞で、好きな場所に保存します。手軽な方法ではありますが、何回も保存すると重くなってしまうことや、どのように変数を変更したのかという経緯までは保存できないのが欠点です。
- **データ変更に使用したシンタックスを保存する**：変更したデータセットは保存せずに、データを変更する時に使ったシンタックスのみを保存します。次の回からは、元のデータセットを開いて、データ変更シンタックスを実行してから分析を再開します。シンタックスはただのテキストファイルなので軽いですし、どのように変数に変更を加えたのかが一目で分かるのがメリットです。

9.1.2　出力をExcel形式で保存する

SPSSのアウトプット画面は、そのままでも保存できますが、そのままの形式だと、SPSSでしか開くことができません。Excel形式で結果を保存するには、次の手順で行います。

- 保存したい部分の図表をCtrlまたはShiftを押しながら複数選択する。
- ＜ファイル>⇒＜エクスポート＞で、エクスポート画面を開く。
- 「エクスポートするオブジェクト」チェックボックスで、「選択のみ」にチェック。
- ドキュメントの型で「.xls」または「.xlsx」を選ぶ。
- 表についている脚注と解説を含めたくなければ、「オプション変更」から「脚注と解説を含む」のチェックを外す。
- 「ファイル名」欄で、保存場所と新しいファイル名を指定。
- 「OK」を押すと、エクスポートが始まる。

9.2 分析結果を発表する *Present Results*

分析が終わり、結果が出力できたら、その結果をレポートやパワーポイントに載せます。SPSSの表（や今回使ったシミュレーションの図など）はそのまま資料に貼り付けてしまうと少し不格好ですね。せっかく良い結果が出ているのならば、見やすい表や図を作って発表しましょう。見やすい表や図を作るためには、以下のような点を心掛けられると良いと思います。

- 一目で内容（言いたいこと）が分かる題名をつける。
- 全体で統一した形式（フォント・変数の示し方・左右位置の揃え方）を保つ。
- 小数点以下の桁数を（できるだけ）揃える。
- 必要な情報だけ載せる。
- あまりゴテゴテと装飾をつけすぎない。

以下では、最終的に結果を発表する際に使う表の例を載せます。参考にしてみてください。(例ももちろん完璧なものではありません。所々英語になっていたり、フォーマットが揃っていなかったりします。参考にしながら、更に良いものを作ってみてください。)

- **クロス表**：クロス表の結果を示す時は、どの数値を載せるべきかを慎重に吟味しましょう。以下の示し方は、Freedom of Speech（独立変数）の値が上がるほど、ExtraJudicial Killingが頻繁に起こる割合（従属変数）が減っていくことが言いたい場合に有効でしょう。別の目的があれば、実数や行%を示しても良いでしょう（ただし数値を示し過ぎると、逆に分かりにくくなります。）また、この分析は順序変数同士なので、ガンマは示しても、Cramer's Vなどは示す必要は無いですね。

Table 1: Freedom of Speech と Extrajudicial Killing のクロス表

	Freedom of Speech (CIRI)			
Extrajudicial killing (CIRI)	0	1	2	合計
0 Practiced frequently	42.9%	27.3%	9.6%	24.8%
1 Practiced infrequently	42.9%	39.4%	21.2%	34.0%
2 Have not occurred	14.3%	33.3%	69.2%	41.2%
合計	100.0%	100.0%	100.0%	100.0%

Pearsonのカイ二乗 30.554^{**}
グッドマン・クラスカルのガンマ 0.567^{**}
ケース数 153
$p^{*} < 0.05$ $p^{**} < 0.01$

- **平均値の差の検定**：Table2は、示し方の一例です。平均値の差の検定を示した表を、実は私はあまり見たことがありません。結果はシンプル（差が有意か否か）なので、表として示さずに、文章中に書いてしまうというのも1つの手段でしょう。

Table 2: 民主国家・独裁国家別の平均値の差の検定

	安全な水にアクセスできる人口割合 (%)	一人あたりGDP (US$)
0 民主国家	84.7	10927.0
1 独裁国家	71.2	4234.3
ルビーンの等分散検定	16.279^{**}	40.720^{**}
T値（等分散を仮定しない）	3.934^{**}	5.386^{**}
ケース数	148	141

$p^{**} < 0.01$

- **相関**：1つの分析で出力される相関の数値は1つなので、それだけ表になっているものはほぼ見ません。一方で、例えば二変数を掛け合わせた散布図の上に二変数の相関係数を示すことは、読者にとって情報量が増えるので有効かもしれません。
- **多変量解析（主要な表）**：Table3・4・5は示し方の1例です。もし２つ以上モデルがある場合は、比較ができるようなモデル同士を横に並べると、伝えたいことが分かりやすくなるでしょう。

Table 3: 交互作用項を加えることによる効果（重回帰分析）

	安全な水にアクセスできる人口%	
	基本モデル	交互作用項モデル
定数項	-19.069	-68.172^{*}
	(10.221)	(27.395)
民主化度	2.487^{**}	13.764^{*}
	(0.908)	(5.916)
一人当たりGDP(対数)	10.572^{**}	16.507^{**}
	(1.453)	(3.397)
民主化度*GDP(対数)		$-1.320^{\dagger}$
		(0.685)
R^2	0.538	0.553
調整済み R^2	0.530	0.541
推定値の標準誤差	13.69	13.53
ケース数	116	116

$p^{\dagger} < 0.10$ $p^{*} < 0.05$ $p^{**} < 0.01$ （ ）内は標準誤差

横に並べるモデルはTable3・5の様に「同じ従属変数だけれども、一部の独立変数が入れ替わったり増えたりする」モデルや、「同じ独立変数だけれども、従属変数が入れ替わる」モデル、そしてTable4の様に「投入する変数は同じでも、サンプル

が異なる」モデルなどが考えられます。比較不可能なモデルを同じ表に押し込めても結果が分かりやすくなる訳ではないので、注意してください。

係数の小数点以下の桁数については（できるだけ）合わせます。もちろんケース数などの整数に小数点以下を付ける必要はありません。また、例えばTable5の様に、係数や標準誤差の小数点以下で定めたところが0.000になってしまう様な場合は、1以上の数が出てくるまで桁数を広げた方が理解し易いでしょう。

また、表示するモデル全体の統計量についてですが、出力された結果全てを表示する必要は必ずしもありません。例えば、Table3では参考に全て示していますが、調整済みR^2と推定値の標準誤差は同じような内容を表している訳ですから、どちらかを載せるだけでも大丈夫でしょう。Table4についても同じです。R^2や対数尤度は省略してしまっても良いかもしれません。

Table 4: サンプルをサブセットに分割することによる効果 (ロジット分析)

	自民投票=1 民主投票=0 (2005比例区)		
	基本モデル	郵政知識あり	郵政知識なし
定数項	-3.765^{**}	-3.652^{**}	-4.266^{**}
	(0.577)	(0.689)	(1.139)
郵政民営化積極度	0.162^{*}	0.198^{*}	0.058
	(0.076)	(0.092)	(0.145)
イデオロギー保守度	0.206^{**}	0.154	0.389^{*}
	(0.073)	(0.086)	(0.155)
自民支持強度	0.623^{**}	0.711^{**}	0.460
	(0.137)	(0.169)	(0.241)
民主支持強度	-1.149^{**}	-1.116^{**}	-1.218^{**}
	(0.164)	(0.195)	(0.315)
小泉支持強度	0.895^{**}	0.850^{**}	0.988^{**}
	(0.147)	(0.177)	(0.278)
カイ２乗	453.446^{**}	348.565^{**}	107.463^{**}
-2 対数尤度	494.381	359.118	131.826
Cox-Snell R^2	0.472	0.485	0.444
Nagelkerke R^2	0.641	0.655	0.609
正確に予測された割合	0.870	0.884	0.847
ケース数	709	526	183

$p^{*} < 0.05$ $p^{**} < 0.01$ （ ）内は標準誤差

余談ではありますが、近年、表の上で有意確率を*などで表すことはあまり好まれない傾向にある様です。p=0.05やp=0.10といったようなしきい値は、それ自体に意味がある訳ではないので、わずかに超えたり超えなかったりすることで解釈が激変

してしまうことには批判があるからです。よって、表上でp値の絶対値や、95%信頼区間を示すことも増えています。逆に、標準誤差さえあれば t 値は計算できる訳ですから、何も示さないという人もいます。

Table 5: 時系列データの重回帰分析における 2 つの対処法の比較

	自民党議席割合	
	固定効果モデル	ラグ従属変数モデル
(定数項)	0.434^{**}	0.039^{**}
	(0.030)	(0.007)
第1次産業従事人口比率	0.004^{**}	0.0007^{*}
	(0.001)	(0.0003)
人口密度	0.0001^{**}	-0.000001
	(0.00004)	(0.000002)
都道府県ダミー	.....	
	(.....)	
年ダミー	.....	
	(.....)	
自民党議席割合（t-1）		0.911^{**}
		(0.011)
R^2	0.707	0.853
調整済み R^2	0.691	0.853
ケース数	1504	1457

$p^{*} < 0.05$ $p^{**} < 0.01$ ( )内は標準誤差
固定効果モデルの都道府県・年ダミーの係数は省略

- **多変量解析（特殊な表）**：ロジット分析や相互作用項の分析結果は、Figure16やFigure19にある様なシミュレーションを行った図を（整えてから）載せることで、とても分かりやすく表現できると思いますが、中には厳密に数値で表現したいという人もいると思います。そこで以下では、ロジット分析と相互作用項の分析結果を表で示す例を紹介します。

  Table 6は、ロジット分析によって、各独立変数の値がある一定範囲変化した場合の従属変数の変化量を示しています。一方で、Table7は交互作用項の一方の変数の値を変化させた時に、他方の変数の総合効果どのように変化するかを表で示したものです。どちらも、シミュレーションの図がある場合は必ずしも必要な表とは言えませんが、数値で表現できるので、より厳密な結果を提示したい場合には便利です。

Table 6: 各独立変数の値の変化に従った自民党投票確率の変化量 (ロジット分析)

	自民党投票確率の変化量（%）		
	基本モデル	郵政知識あり	郵政知識なし
郵政民営化積極度			
消極的（０）⇒積極的（６）	22.8	28.4	7.4
保革イデオロギー			
革新（０）⇒保守（１０）	45.6	36.2	69.6
自民党支持			
無支持（０）⇒強い支持（３）	30.2	35.1	20.6
民主党支持			
無支持（０）⇒強い支持（３）	-56.1	-52.2	-63.8
小泉内閣支持			
不支持（１）⇒支持（４）	58.6	56.1	62.0

変化させる以外の変数は、次の値で固定：
郵政（３）・保革（５）・自民支持（０）・民主支持（０）・小泉支持（３）

Table 7: 民主化度が水にアクセスできる割合に与える効果は豊かになるほど小さい

一人当たりＧＤＰ（相対的位置）	民主化度の総合効果	標準誤差	有意確率（両側）	一人当たりＧＤＰ(対数：参考)
最小値	5.584	1.840	0.003	6.194
25パーセンタイル	3.763	1.115	0.001	7.574
中央値	2.608	0.900	0.005	8.449
75パーセンタイル	1.361	1.071	0.207	9.393
最大値	-0.437	1.762	0.805	10.754