2025/02/19

藪友良

付録 B
エクセルの使い方

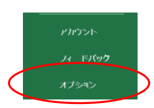
統計学を勉強しても、やはり実際に自分で使ってみないと理解は十分ではありません。ここでは、実際に統計分析を使う方法のひとつとして、エクセルの使い方を解説します。本稿は、藪友良『入門　実践する統計学』(2012 年、東洋経済新報社)の付録 B をアップデートしたものです。データは、本書ウェブサイトからファイル appendix_b_data.xlsx をダウンロードしてください。

B.1　分析ツール

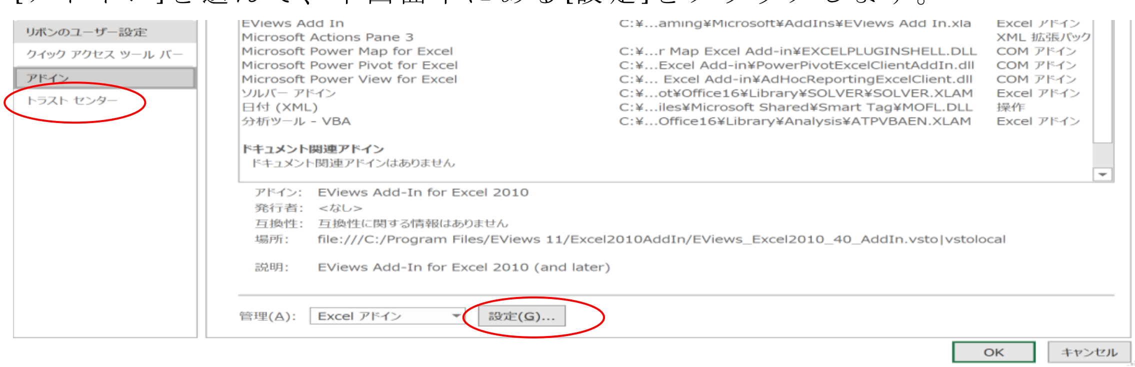
エクセルについている分析ツールという機能を使えば、さまざまな統計分析が可能です。まず、この機能を使えるように設定します。もし[データ]タブに[データ分析]という項目があれば、本節は無視してください。[データ分析]がなければ、以下の手順に従って設定します[1]。

エクセル画面左上の[ファイル]タブをクリックします(下画面参照)。

そして左列の一番下にある[オプション]をクリック(下画面参照)。

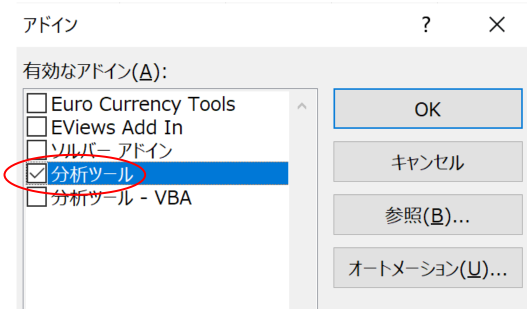


そうすると、[Excel のオプション]ウィンドウ画面がでますから、下画面から[アドイン]を選んで、下画面下にある[設定]をクリックします。



[1] ここではエクセル 2007 に基づいて説明をしています。本節のやり方で設定できないときは、ヘルプ機能で分析ツールを検索し、そこで示されている方法に従って設定してください。

それから下画面で[分析ツール]をチェックして OK とします。

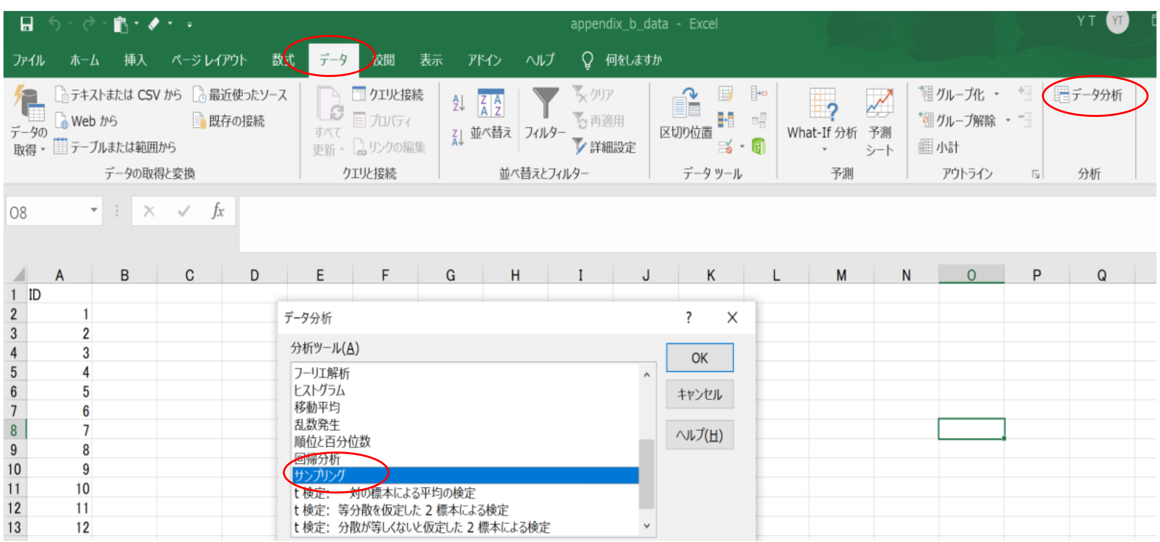


これで[データ]タブの一番右に[データ分析]という項目が加わります。

B.2　無作為抽出

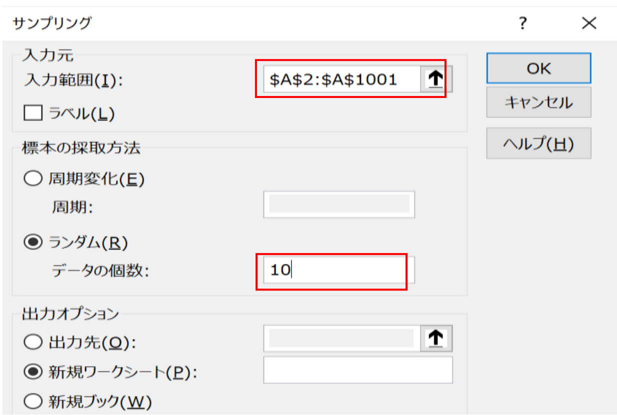
1 章では無作為抽出の方法として、10 面体のサイコロを用いる方法を説明しました。しかし、エクセルを使えば無作為抽出を容易に行うことができます(エクセルのシート「ID」参照)。

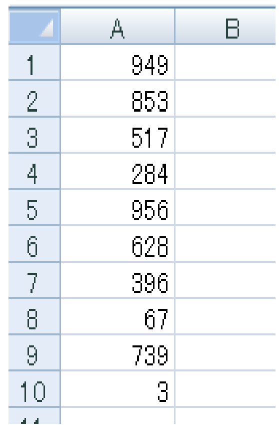
たとえば、1000 人の学生から 10 人を無作為抽出するときは、まず全員に 1 から 1000 までの番号（ID）を付けます。次に、[データ]タブの[データ分析]をクリックします。そして、[サンプリング]を選んでから OK をクリックします。



すると、左下の画面が出てきます。ここで□に情報を入力します。A 列の 2 行から 1001 行まで番号が記録されていますから、入力範囲として、`$A$2:$A$1001` を入力します。入力が面倒であれば、ボタンを押して、データの入力範囲を画面上で範囲指定することもできます。また、10 人を無作為

抽出しますから、データの個数に 10 を入力します。そして OK を押すと右下の画面となり、乱数が抽出されます。この場合、選ばれた乱数は 949、853、517、284、956、628、396、67、739、3 です。この番号に対応した 10 人の学生を選べば無作為抽出は完了です[2]。無作為抽出ですから、みなさんは異なる番号が抽出されたと思います。





B.3　特性値の計算方法

エクセルを使った特性値の計算方法を説明します。エクセルにはさまざまな特性値を計算するため、多くの関数が定義されています。たとえば、平均であれば=AVERAGE(配列)とすれば、配列指定したデータを使って、平均を計算してくれます（配列はデータ入力範囲）。代表的関数には

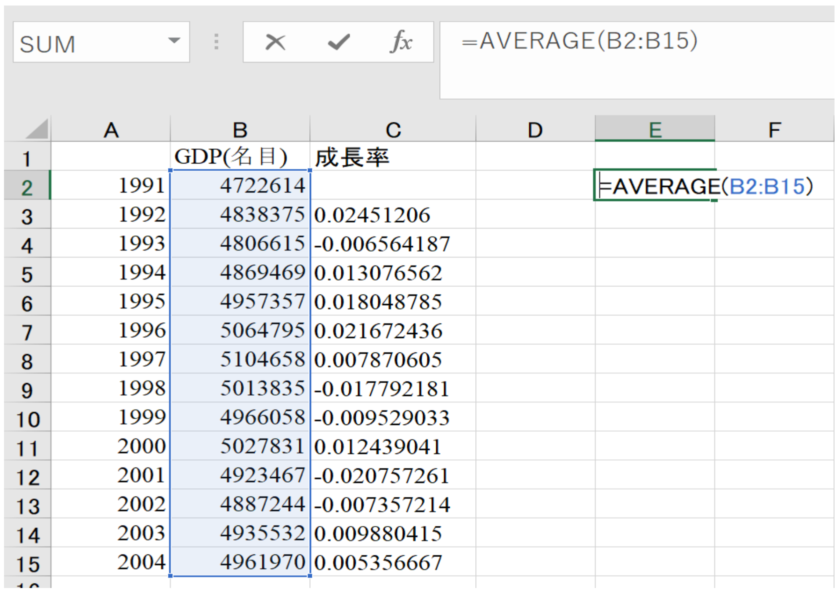
平均=AVERAGE（配列）、中央値=MEDIAN（配列）、最頻値=MODE（配列）、標本分散=VAR（配列）、標本標準偏差=STDEV（配列）、標本共分散=COVAR（配列 1，配列 2）、標本相関係数=CORREL（配列 1，配列 2）

などがあります。標本共分散と標本相関係数は 2 変数の関係をとらえる指標ですから、配列を 2 つ指定します。

下画面は GDP を 1991～2004 年まで記録したものです(エクセルのシート「GDP」参照)。A 列には時間（年）、B 列には GDP が記録され、C 列では

[2]無作為抽出すると、同じ番号が重複することがあります。番号の重複を認める方法を**復元抽出**、認めない方法を**非復元抽出**といいます。たとえば、番号 1～100 から 3 つを抽出したとき、乱数として 91、7、91 が抽出されたとしましょう。復元抽出なら、選ばれた乱数は 91、7、91 ですが、非復元抽出なら（もう一度抽出したところ乱数 80 が得られたとする）、選ばれた乱数は 91、7、80 です。非復元抽出は同一母集団からの抽出とならず、i.i.d.の仮定は満たされません。ただし、母集団規模が十分に大きければどちらでも変わりません。

GDP 成長率（変化率）を計算しています。たとえば、C 列の 3 行目は 0.0245 となっていますが、これは=(B3-B2)/B2 として計算できます[3]。そして、このセル（C3）をコピーして、C 列の 4～15 行までに貼り付ければ変化率を全て計算できます。ここで GDP の平均は=AVERAGE(B2:B15)、GDP と成長率の相関は=CORREL(B3:B15, C3:C15) として計算できます（計算すると、それぞれ 4934272 と 0.1009)。

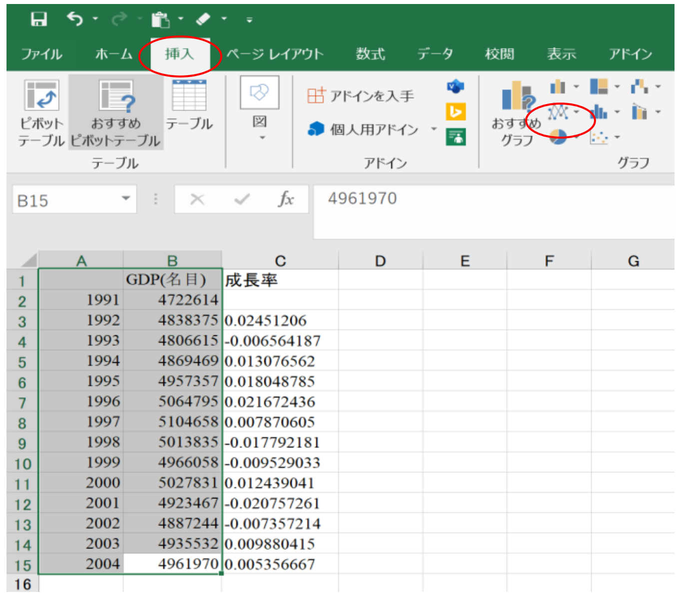


	A	B	C	D	E	F
1		GDP(名目)	成長率			
2	1991	4722614			=AVERAGE(B2:B15)	
3	1992	4838375	0.02451206			
4	1993	4806615	-0.006564187			
5	1994	4869469	0.013076562			
6	1995	4957357	0.018048785			
7	1996	5064795	0.021672436			
8	1997	5104658	0.007870605			
9	1998	5013835	-0.017792181			
10	1999	4966058	-0.009529033			
11	2000	5027831	0.012439041			
12	2001	4923467	-0.020757261			
13	2002	4887244	-0.007357214			
14	2003	4935532	0.009880415			
15	2004	4961970	0.005356667			

B.4. 図の描き方

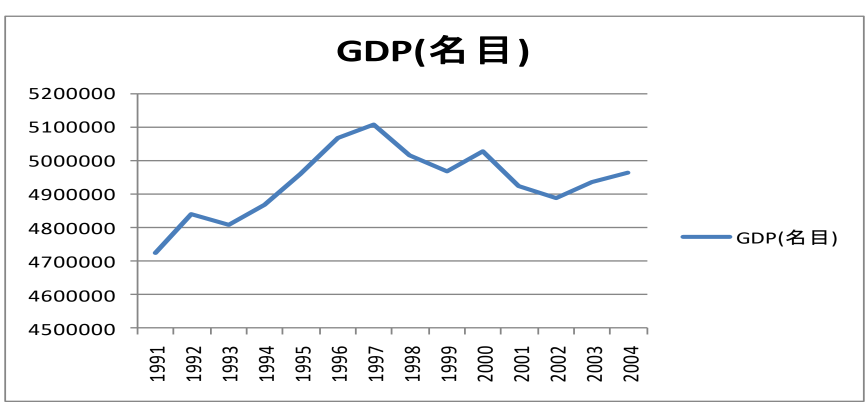
ここでは図の書き方を説明します。下画面では、1991～2004 年までの GDP が記録されています。図を描くためには、まず、下画面のようにデータ範囲を指定します。指定が終わってから、[挿入]タブをクリックし、折れ線マークをクリックします（折れ線以外に、散布図、縦棒、横棒、円グラフなどもあります)。

[3] 1992 年の変化率は、1991～92 年の変化分（B3-B2）を 1991 年の値(B2)で割ったものです。



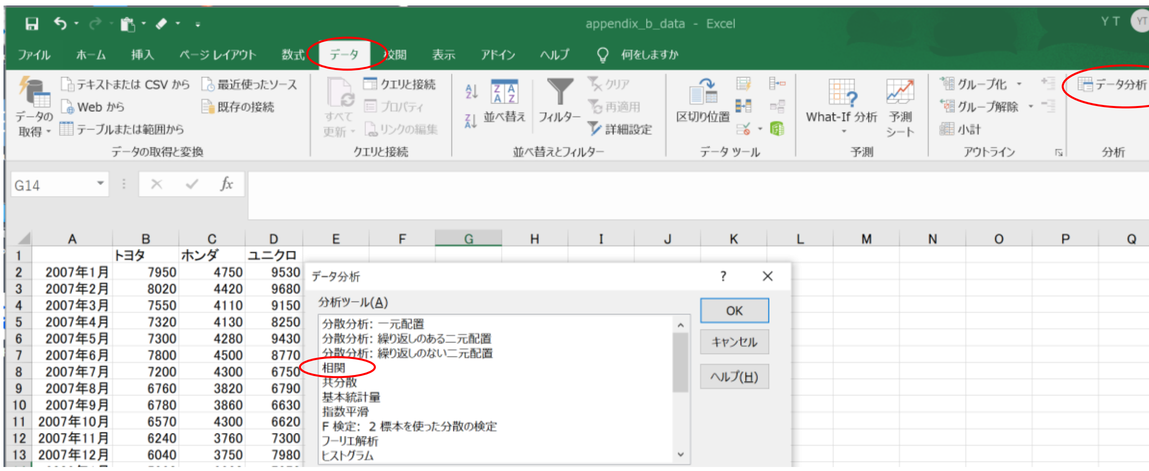
そして折れ線の中で、自分が描きたい図の種類を選択します。ここでは、左上の図を選びましょう。

そうすると、下図が出てきます。横軸は年、縦軸は GDP の規模となっています。このままではあまりきれいな図ではありませんから、微調整して図を見やすくする必要があります。作成された図をクリックして、[デザイン]タブを選ぶと、さまざまな微調整ができるようになります。また、図の軸をクリックし、書式設定を選ぶことでも調整ができます。

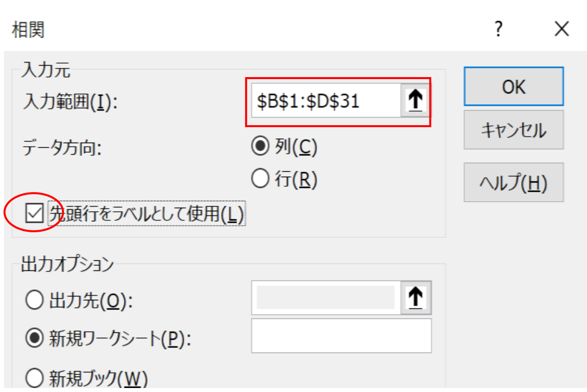


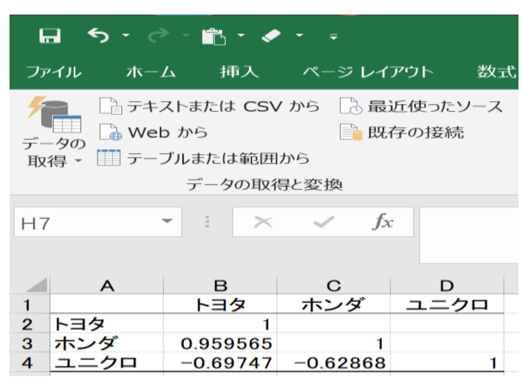
B.5　標本相関係数の計算

エクセルを使って標本相関係数を計算しましょう。5 章ではトヨタ、ホンダ、ユニクロの株価の動きを紹介しました。ここでは、これら 3 系列の標本相関係数を計算します(エクセルのシート「株価」参照)。まず、[データ]タブの[データ分析]をクリックします。そして、[相関]をクリックして OK します。



すると、左下の画面が出てきます。ここで□に情報を入力します。B～D 列の 1～31 行にデータが記録されていますから、入力範囲として、$B$1:$D$31 を代入します。ただし、最初のセルはデータ名ですから、[先頭行をラベルとして使用]をチェックしてください。そうすると、最初のセルをラベルとして認識してくれます。そして OK すると、右下の画面が出てきます。同じ系列同士では相関係数は 1 となりますから、対角要素は全て 1 です。トヨタとホンダの相関係数は 0.960、トヨタとユニクロの相関係数は－0.697、ホンダとユニクロの相関係数は－0.629 となっています。





B.6　確率の計算

ここでは確率変数の確率を計算する方法を説明します。

Z は標準正規確率変数とします。確率 $P\{Z<z\}$ を求めたい場合、=NORMSDIST(z)と入力します。たとえば、$P\{Z<1.96\}$ を求めるには＝NORMSDIST(1.96)と入力すれば 0.9750 が得られます（エクセルのシート「確率、臨界値の計算」参照）。

W は自由度 n の χ^2 確率変数とします。このとき、$P\{\chi^2_{n,\alpha}<W\}=\alpha$ となる $\chi^2_{n,\alpha}$ の値を計算するには、=CHIINV(α,n)と入力します。たとえば、$\chi^2_{99,0.025}$ を求めるときは、=CHIINV(0.025,99)と入力すれば 128.42 が得られます。また、$\chi^2_{99,0.975}$ を求めたい場合は、=CHIINV(0.975,99)と入力すれば 73.36 が得られます。

U は自由度 n の t 確率変数とします。このとき、$P\{t_{n,\alpha}<|U|\}=\alpha$ が成立する $t_{n,\alpha}$ を計算するには、=TINV(α,n)と入力します。たとえば、$t_{1,0.1}$ は、=TINV(0.1,1)と入力すると、6.314 が得られます。

V は第 1 自由度 m、第 2 自由度 n の F 確率変数とします。$P\{F_{m,n}<V\}=\alpha$ となる $F_{m,n}$ の値を計算するには=FINV(α,m,n)と入力します。$\alpha=0.05$、$m=n=2499$ であれば、=FINV(0.05, 2499, 2499)として 1.068 が得られます。

B.7. 回帰分析

回帰分析の仕方とその解釈を説明します。2010/1/1～2010/12/31 の円ドルレートを使って、為替レートが予測できるのかを調べます（エクセルのシート「円ドル」参照）。t 日の為替レートを S_t とし、t 日の変化率を

$$dS_t = \frac{S_t - S_{t-1}}{S_{t-1}}$$

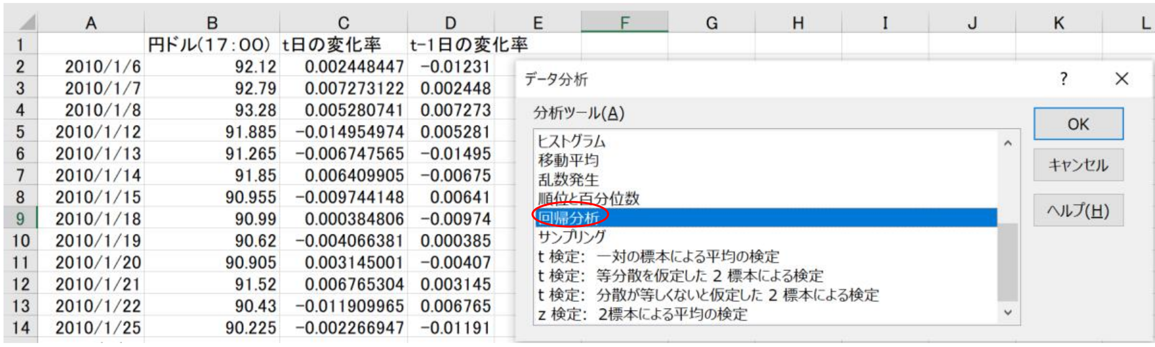
と表します。

ここでは被説明変数を t 日の変化率(dS_t)、説明変数を $t-1$ 日の変化率（dS_{t-1}）　とします。

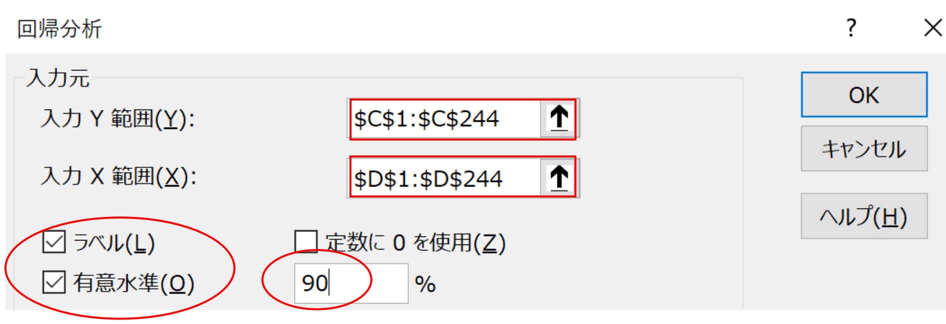
$$dS_t = \alpha + \beta dS_{t-1} + u_t$$

β が 0 から有意に異ならなければ為替の予測はできませんし、β が有意に 0 から異なれば為替の予測が可能だといえます。これは自己回帰（AR）モデルといわれ、時系列分析で重要なモデルのひとつです（詳しくは、参考文献で紹介されている本を読んでください）。

[データ]タブの[データ分析]をクリックします。そして、[回帰分析]を選択して OK をクリックします。



すると下画面が表示されます。ここで被説明変数（Y）は C の列(t 日の変化率、つまり dS_t)、説明変数（X）は D の列($t-1$ 日の変化率、つまり dS_{t-1})となります。したがって、入力 Y 範囲は C の列$C$1:$C$244 とし、入力 X 範囲は D の例$D$1:$D$244 と範囲指定します[4]。Y と X の入力範囲は 1 行目から始めていますが、1 行目はデータの名前（ラベル）で、データそのものではありません。このため、ラベルにチェックを入れて、データではないことを明示しておきます。また、ここでは有意水準にチェックし、横の空欄に 90 という値を入れておきます。こうすることで信頼区間 90%を計算してくれます。



最後に OK をクリックすると、以下の画面が表示されます。ここでは重要な情報とその意味を紹介していきます。

[4] 入力 X 範囲は 1 変数ではなく、複数の変数を指定することもできます。その場合は多重回帰分析となります。

回帰統計	
重相関 R	0.111537
重決定 R2	0.012441
補正 R2	0.008343
標準誤差	0.006376
観測数	243

分散分析表

	自由度	変動	分散	された分散	有意 F
回帰	1	0.000123	0.000123	3.035945	0.082715
残差	241	0.009796	4.06E-05		
合計	242	0.009919			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 90.0%	上限 90.0%
切片	-0.00053	0.00041	-1.28724	0.199245	-0.00134	0.00028	-0.00121	0.000149
t-1日の変	-0.11097	0.06369	-1.7424	0.082715	-0.23643	0.014487	-0.21614	-0.00581

最初のブロックには、当てはまりの尺度である決定係数などがまとめられています。重決定 R2 は決定係数であり、0.0124 とあまり当てはまりはよくありません。また、補正 R2 は自由度調整済み決定係数であり、0.0083 と低い値ですから、やはり当てはまりはよくありません。観測数はサンプルサイズであり、この場合は 243 となります。

最後のブロック（分散分析表）に、母数（α、β）の推定値、標準誤差、95%と 90%信頼区間がまとめられています（有意水準をチェックし、90 と入力したため、90%信頼区間も表示）。たとえば、α は－0.00053、β は－0.11097 と推定されています。それぞれの標準誤差は 0.00041 と 0.06369 です。t 値は推定値を標準誤差で割ったもので、それぞれ－1.28724（=－0.00053/0.00041）、－1.7424（=－0.11097/0.06369）となります。この場合、α に関する t 値は 0 に近いため、帰無仮説（H_0：$\alpha=0$）は採択されます。これに対し、β に関する t 値は-1.7424 と 0 から乖離し、有意水準 10%で帰無仮説（H_0：$\beta=0$）は棄却され、対立仮説が支持されます。90%信頼区間は、母数が 90%の確率でその範囲に含まれることを意味します。たとえば、β は 90%の確率で－0.21614 から－0.00581 の範囲に収まり、信頼区間内に 0 を含みませんから有意な結果といえます。

最後に、p 値は「H_0 が正しいにもかかわらず、t 統計量の絶対値が t 値（t*）の絶対値より大きな値をとる確率（$P\{|t_{\hat{\beta}}|>|t^*|\}$）」です。たとえば、この表から、$\beta$ に関する t 値は $t^*=-1.7424$ で、p 値は 0.082715 となっていま

す。p 値は、H_0 が正しいもとで、t 統計量の絶対値($|t_{\hat{\beta}}|$)が t 値の絶対値($|t^*|=1.7424$)より大きくなる確率ですから、$P\{|t_{\hat{\beta}}|>1.7424\}=0.082715$ となっているのです。p 値を見れば何%の有意水準で帰無仮説を棄却できるかが分かります。かりに p 値が 1%を下回っていれば、1%の有意水準でも帰無仮説を棄却できます。この場合、β の p 値 0.082715 は 0.1 を下回りますから、有意水準10%で帰無仮説は棄却されます。