第 5 章　R を使用した解説・補論

5-1　上記 2 変数の標準偏差、共分散、相関係数を求めましょう。

【解説】

5 章の解答ファイルでは以下の計算結果を示しました。

基本統計量

総人口2019		県民所得2019	
平均	2,692,661.2340	平均	3,016.7991
標準誤差	407,116.0741	標準誤差	71.7159
中央値（メジアン）	1,601,865.0000	中央値（メジアン）	2,972.6710
最頻値（モード）	#N/A	最頻値（モード）	#N/A
標準偏差	2,791,047	標準偏差	491.6597
分散	7,789,944,397,431	分散	241,729.2542
尖度	5.4354	尖度	21.0163
歪度	2.2559	歪度	3.8224
範囲	13,449,694.0000	範囲	3,360.4391
最小	557,370.0000	最小	2,396.3448
最大	14,007,064.0000	最大	5,756.7839
合計	126,555,078.0000	合計	141,789.5557
データの個数	47	データの個数	47

共分散

	総人口2019	県民所得2019
総人口2019	7,624,200,899,614	
県民所得2019	881,695,487	236,586

相関係数

	総人口2019	県民所得2019
総人口2019	1	
県民所得2019	0.656488102	1

同様のことは RStudio でも行うことができます。データの読み込みやパッケージのインストールについては第 1 章の練習問題 1-5 の解説に書きましたので、そちらを参照してください。ここでは、データのファイル名を「Data5.xlsx」としました。

標準偏差を含む基本統計量を計算するためには、psych パッケージの describe 関数がよいでしょう。

describe(データ名)

を実行すれば、以下のように平均や標準偏差を計算してくれます。有効桁数の違いがありますが、平均値や標準偏差が（ほぼ）同じ値になっていることが分かります。

```
> describe(Data5_123)
             vars  n      mean         sd     median    trimmed       mad       min         max
都道府県*       1 47      24.0      13.71      24.00      24.00     17.79      1.00       47.00
総人口2019      2 47 2692661.2 2791047.19 1601865.00 2160589.26 933594.70 557370.00 14007064.00
県民所得2019    3 47    3016.8     491.66    2972.67    2969.07    283.83   2396.34     5756.78
                  range skew kurtosis        se
都道府県*          46.00 0.00    -1.28      2.00
総人口2019   13449694.00 2.11     4.42 407116.07
県民所得2019     3360.44 3.58    17.81     71.72
```

次に共分散と相関係数について計算しましょう。R では共分散に cov 関数、相関係数に cor 関数を用います。以下のように入力して実行すると、

cov(Data5$総人口 2019,Data5$県民所得 2019)

cor(Data5$総人口 2019,Data5$県民所得 2019)

共分散は「900862780」、相関係数は「0.6564881」となります。Excel のアドイン「データ分析」の計算結果と比べると、相関係数は同じですが、共分散が異なるはずです。「総人口」の標準偏差と「1 人当たり県民所得」の標準偏差を掛け合わせた値で共分散を割ると相関係数が得られるはずなので、計算してみると、

$$\frac{900862780}{2791047 \times 491.6597} = 0.6564881\cdots$$

です。他方、練習問題解答の p.2 下部で触れたように、Excel のアドイン「データ分析」の共分散を利用すると、下のようになり、上で得られた相関係数と一致しません。

$$\frac{881695487}{2791047 \times 491.6597} = 0.642520\cdots$$

なぜ異なる 2 つの共分散が存在するのでしょうか。答えは算式が異なることにあります。「900862780」は次の式で計算されています。

$$Cov_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

「881695487」は次の式で計算されています。

$$Cov_{XY} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

違いは「n-1」（ここでは 46）で割るか、「n」（ここでは 47）で割るかにあります。実は、相関係数を計算する際には「n-1」で割っています（標準偏差については本書第 3 章）。したがって、標準偏差に合わせて、共分散についても「n-1」で割るというわけです。しかし、Excel のアドイン「データ分析」では、共分散の計算で「n-1」ではなく「n」を用いているため、相関係数との整合性が取れないのです。なお、Excel でも COVARIANCE.S 関数を用いれば、「n-1」を用いた共分散を計算することができます。これに対して、「n」を用いた共分散を計算する関数は COVARIANCE.P 関数です。

5-2 「総人口」を横軸、「1 人当たり県民所得」を縦軸にして散布図を作成しましょう。

【解説】

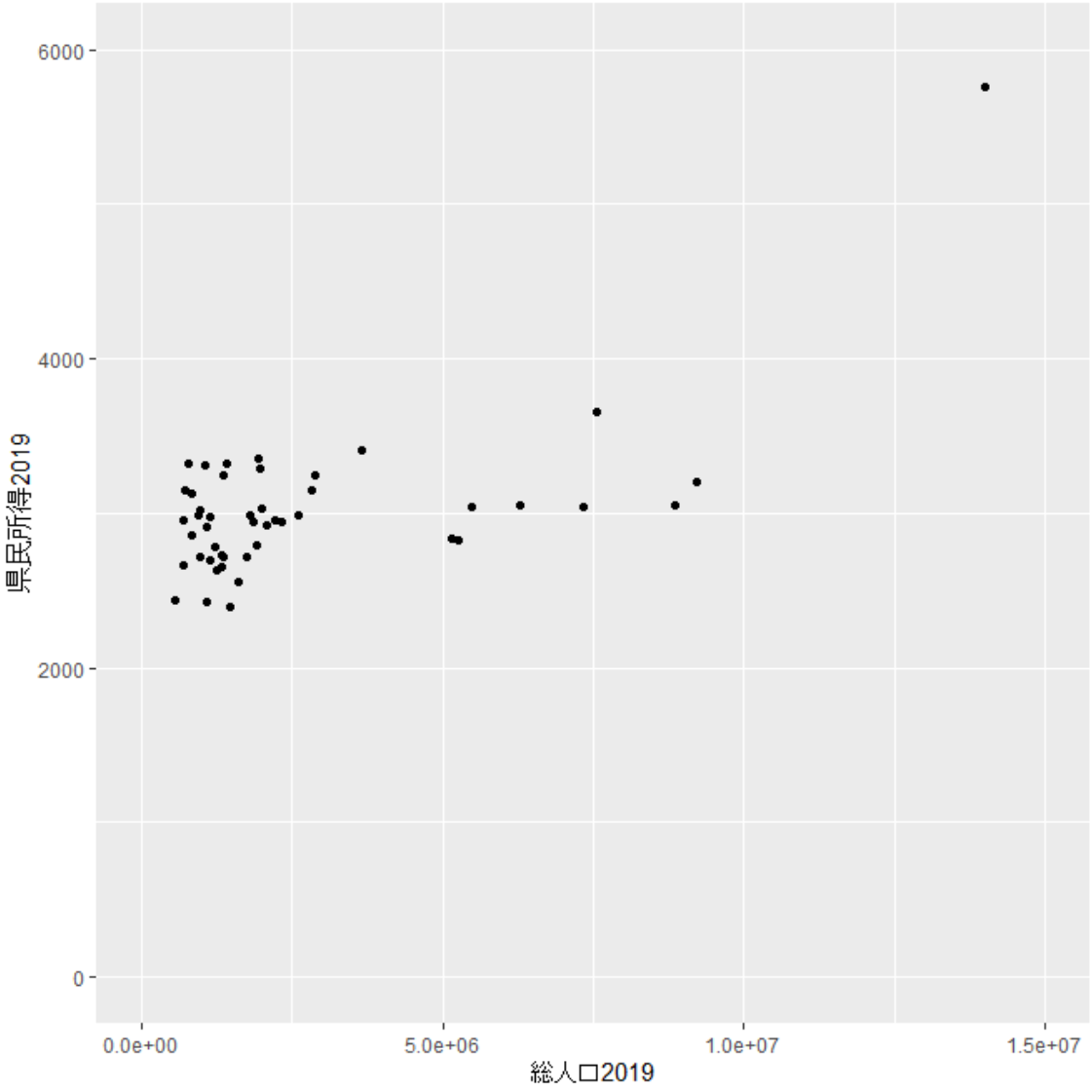
まず、ggplot2 パッケージを準備します（第 1 章解答で解説）。

ggplot (Data5) + geom_point (aes (x=総人口 2019, y=県民所得 2019))

を実行すると散布図を描くことができます。ただし、軸の目盛りが 0 から始まっていないなどの問題が生じるでしょう。そこで、次のように軸目盛についてのオプション（2 行目）付きで実行してみると下の図を描くことができます。ここでは、コマンド後半の

scale_x_continuous()で横軸目盛、scale_y_continuous()で縦軸目盛の設定を行っています。最初の数字（ここでは横・縦とも 0）が下限、後の数字が（横 15000000、縦 6000）上限を意味しています。

```
ggplot ( Data5 ) + geom_point ( aes ( x=総人口 2019, y=県民所得 2019))
+scale_x_continuous ( limits = c(0,15000000))
+scale_y_continuous ( limits = c(0,6000))
```



5-3　上記 5-2 の図を対数目盛に変更した散布図を作成しましょう。ただし、横軸のみ対数目盛、縦軸のみ対数目盛、横軸・縦軸の両方とも対数目盛の 3 種類を作成します。

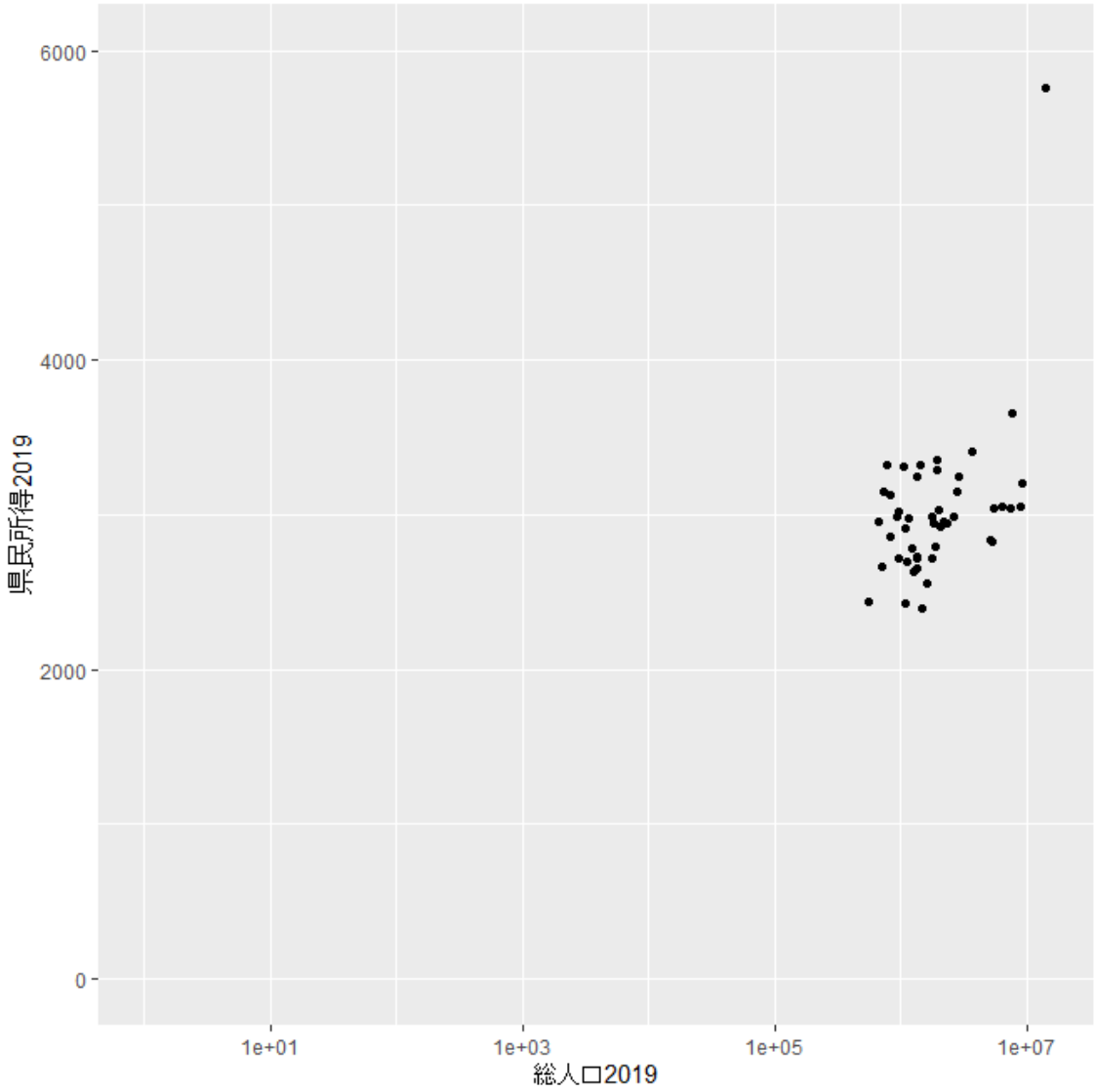
【解説】

横軸、縦軸の設定を対数変換するために、下限・上限の設定に加えて「trans = 'log10'」

を加えてみましょう。「log10」は常用対数に変換することを意味します。以下では、見やすさを優先して常用対数変換としましたが、「trans = 'log10'」を「trans = 'log'」とすれば自然対数変換も可能です。同様に「trans='log2'」とすれば、2 を庭とした対数目盛に変換することができます。

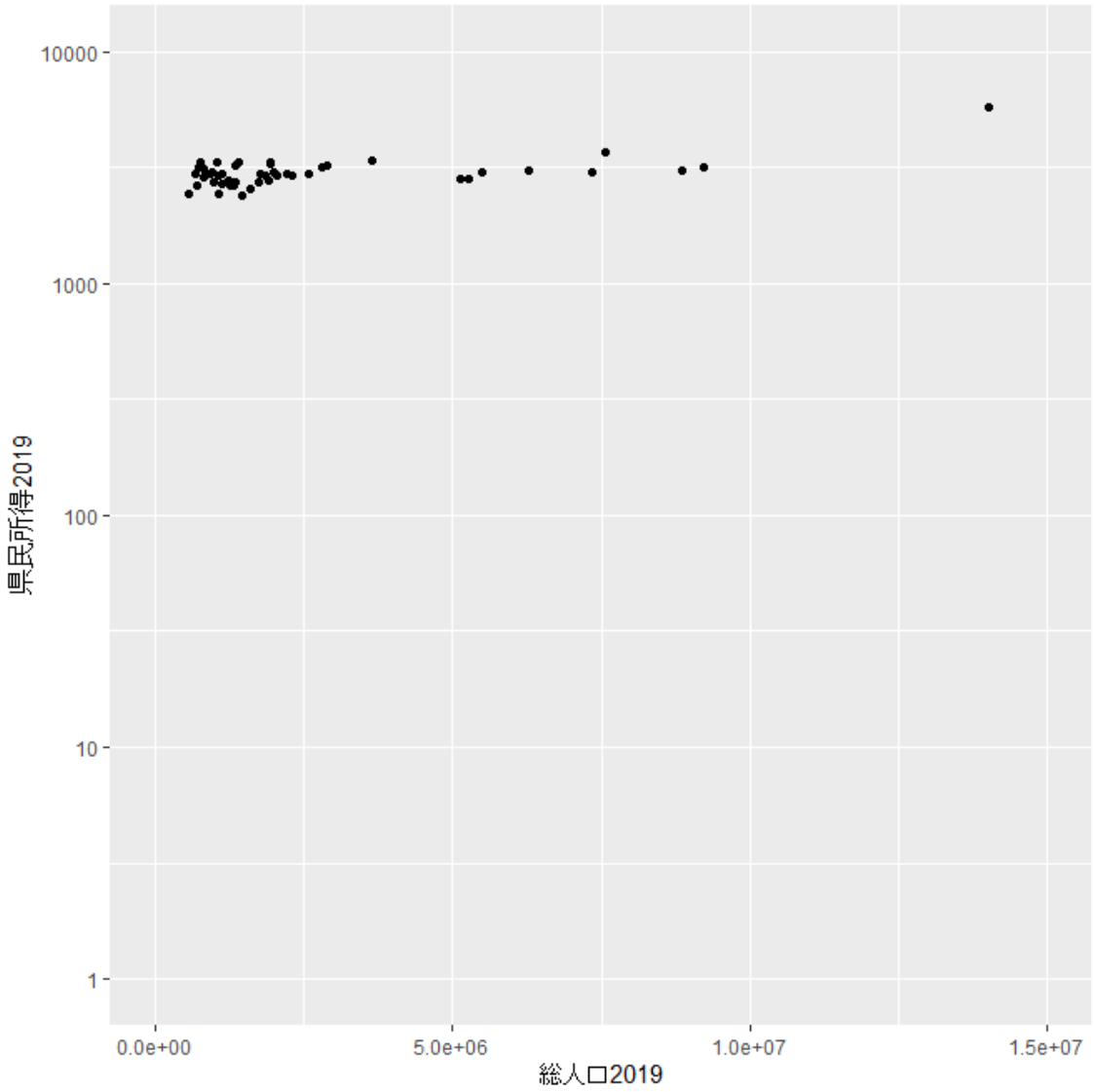
横軸のみ常用対数目盛にするため、,以下を実行すると下図が得られます。ただし、0 を対数変換することはできないので、下限を 1 に変更しています。

```
ggplot ( Data5 ) + geom_point ( aes ( x=総人口 2019, y=県民所得 2019))
+scale_x_continuous ( trans = 'log10', limits = c(1,15000000))
+scale_y_continuous ( limits = c(0,6000))
```



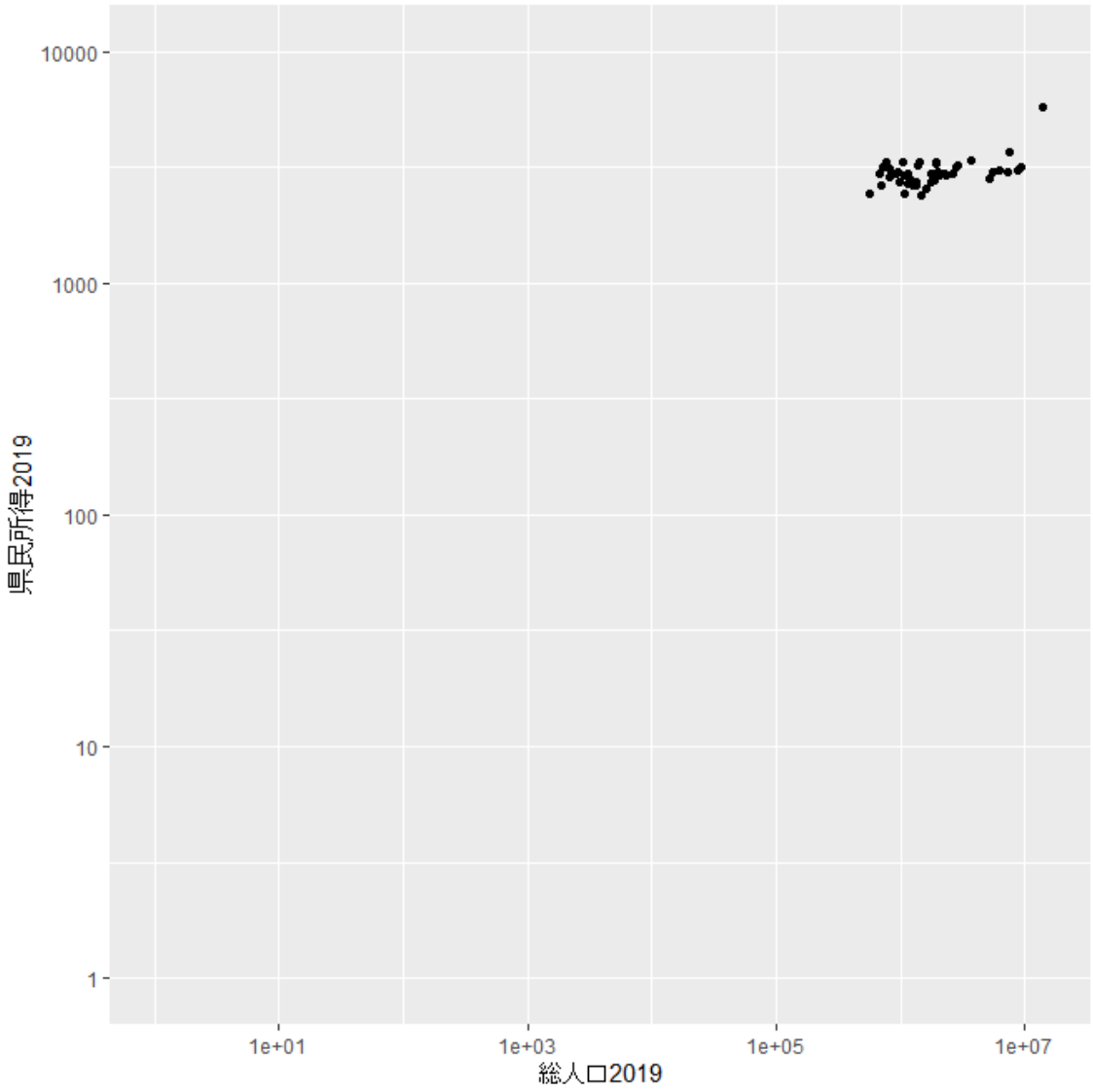
縦軸のみ対数目盛にするためには以下を実行します。こちらも下限を 1 に変更し、上限は 10000（10 の 4 乗）としています。

```
ggplot ( Data5 ) + geom_point ( aes ( x=総人口 2019, y=県民所得 2019))
+scale_x_continuous ( limits = c(0,15000000))
+scale_y_continuous ( trans = 'log10', limits = c(1,10000))
```



最後に、横軸・縦軸ともに対数目盛にする場合です。

```
ggplot ( Data5 ) + geom_point ( aes(x=総人口 2019, y=県民所得 2019))
+scale_x_continuous ( trans = 'log10', limits=c(1,15000000))
+scale_y_continuous ( trans = 'log10', limits=c(1,10000))
```



5-4　本書のウェブサポートページにある都道府県別スターバックス店舗数のデータを被説明変数とし、「総人口」と「1 人当たり県民所得」を説明変数として回帰分析を実行しましょう。また、本文中の「総人口」のみを説明変数とした回帰分析と比較しましょう。ただし、すべての変数は自然対数に変換してください。

【解説】

スターバックス店舗数データはウェブサポートページの「Data5_4.xlsx」です。このデータは、2011 年 5 月 13 日から 2022 年 6 月 30 日までの不定期なデータとなっています。本書では、2021 年 6 月のスーバックス店舗数、2021 年 1 月の住民基本台帳人口を使用しました。ここで手元にある県民経済計算のデータは 2019 年度のものなので、店舗数として最も時点の近い 2020 年 7 月 16 日のデータを使用することにします。

まず、2020 年 7 月 16 日のスーバックス店舗数を RStudio に読み込んで（import して）ください。2020 年 7 月 16 日のスーバックス店舗数を「STARBUCKS2020」とし、このファイルを「Data5_4.xlsx」で保存したものとします。次に、「総人口 2019」と「県民所得 2019」を含む「Data5」に、「STARBUCKS2020」を含む「Data5_4」を結合します。

```
Data5 <- cbind ( Data5, Data5_4)
```

Excel 上で「STARBUCKS2020」の列を追加してから RStudio に読み込んでも構いません。

線形モデルの回帰分析は lm(被説明変数 ~ 説明変数のリスト, データ名)という関数で実行することができます。まず、「STARBUCKS2020」を自然対数変換した「log(STARBUCKS2020)」を被説明変数、「総人口 2019」を自然対数変換した「log(総人口 2019)」を説明変数として回帰分析を実行してみます。

```
lm ( log(STARBUCKS2020) ~ log(総人口 2019), data=Data5)
```

以下のように、定数項と「log(総人口 2019)」の係数が表示されます。

```
Coefficients:
    (Intercept)  log(総人口 2019)
        -14.809            1.224
```

もう少し詳しい回帰分析の結果が欲しいと思いますので、summary()関数を使ってみましょう。lm()関数に summary()関数を組み合わせて、以下を実行してみます。

```
summary ( lm ( log(STARBUCKS2020) ~ log(総人口 2019), data=Data5))
```

下のような出力が得られます。5 章解答で示した Excel の計算結果と（桁数の精度を除けば）同じであることを確認してください。

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.59765 -0.17502  0.00129  0.17553  0.63471

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -14.80871    0.79210  -18.70   <2e-16 ***
log(総人口 2019)   1.22392    0.05474   22.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2939 on 45 degrees of freedom
Multiple R-squared:  0.9174,    Adjusted R-squared:  0.9156
F-statistic:   500 on 1 and 45 DF,  p-value: < 2.2e-16
```

次に、説明変数を「log(総人口 2019)」と「log(県民所得 2019)」の 2 つにしてみましょう。2 つの説明変数を「+」でつなげて、以下のように記述します。

```
summary ( lm ( log(STARBUCKS2020) ~ log(総人口 2019) + log(県民所得 2019),
data=Data5))
```

実行すると、以下の出力が得られます。これも Excel の計算結果と同じになっています。

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.63599 -0.14140  0.01991  0.11537  0.80426

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -21.32077    2.45341  -8.690 4.21e-11 ***
log(総人口 2019)    1.14122    0.05906  19.322  < 2e-16 ***
log(県民所得 2019)  0.96315    0.34603   2.783   0.0079 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.274 on 44 degrees of freedom
Multiple R-squared:  0.9298,   Adjusted R-squared:  0.9266
F-statistic: 291.3 on 2 and 44 DF,  p-value: < 2.2e-16
```

「log(県民所得 2019)」だけを説明変数とするケースも次のように分析することができるはずです。5 章解答で、これについても Excel の計算結果と同じになっていることを確認できます。。

```
summary ( lm ( log(STARBUCKS2020) ~ log(県民所得 2019), data=Data5))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.7239 -0.4987 -0.1165  0.3273  1.8131

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -31.7455     7.2887  -4.355 7.58e-05 ***
log(県民所得 2019)   4.3267     0.9107   4.751 2.10e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8345 on 45 degrees of freedom
Multiple R-squared:  0.334,    Adjusted R-squared:  0.3192
F-statistic: 22.57 on 1 and 45 DF,  p-value: 2.099e-05
```

最後に、5 章解答に倣って、「県民所得計」の対数を説明変数とした場合も分析してみましょう。「県民所得計」は次のように計算できますから、

県民所得計 = 総人口 2019 × 県民所得 2019

説明変数の log の中を「総人口 2019)」と「県民所得 2019」の掛け算にすればよいのです。以下を実行すると、下の出力が得られます。

```
summary ( lm ( log(STARBUCKS2020) ~ log (総人口 2019*県民所得 2019),
data=Data5))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.65514 -0.14655  0.01884  0.11063  0.83538

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -22.36282    1.03746  -21.55   <2e-16 ***
log(総人口 2019 * 県民所得 2019)   1.12417    0.04618   24.35   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2716 on 45 degrees of freedom
Multiple R-squared:  0.9294,    Adjusted R-squared:  0.9279
F-statistic: 592.7 on 1 and 45 DF,  p-value: < 2.2e-16
```