1　最高気温とアイスティーの売上げの関係に関する回帰分析

ここに最高気温とアイスティーの売上げを列記したデータがある。これについて回帰分析を行う。

最高気温を材料にしてアイスティーの売上げを推定できるか調べる。このような推定を専門的表現で「アイスティーの売上げを最高気温に回帰させる」というらしい。

1.1　相関の確認

まず、回帰分析を行う意味のあるデータなのかどうかを確認する。

２つの項目の相関が認められなければ、回帰分析の結果が実質的な意味を持たない。

「最高気温」と「アイスティーの売上げ」の相関係数を確認すると次のとおり。

r=0.906923, df=12, p-value=7.66141e-06

有意な強い相関が認められる。

ちなみに、散布図は図 1 のようになる。

1.2　回帰分析

強い相関が認められるので、回帰分析を行ってみた。その結果は次のとおり。

- y 切片： -36.36123
- x の傾き： 3.737885

つまり、次のような推定のための計算式 (回帰式) が求められる。

「アイスティーの売上げ」＝-36.36123 ＋ 3.737885 ×「最高気温」

上の回帰式によって描かれる直線が回帰直線である。

散布図に回帰直線を書き入れてみると図 2 のようになる。

個々の実測値と回帰式から得られる値との間には誤差 (残差) がある。すべての実測値についてこの誤差の平方和が最小になるように考え出されたのが上の回帰式である。

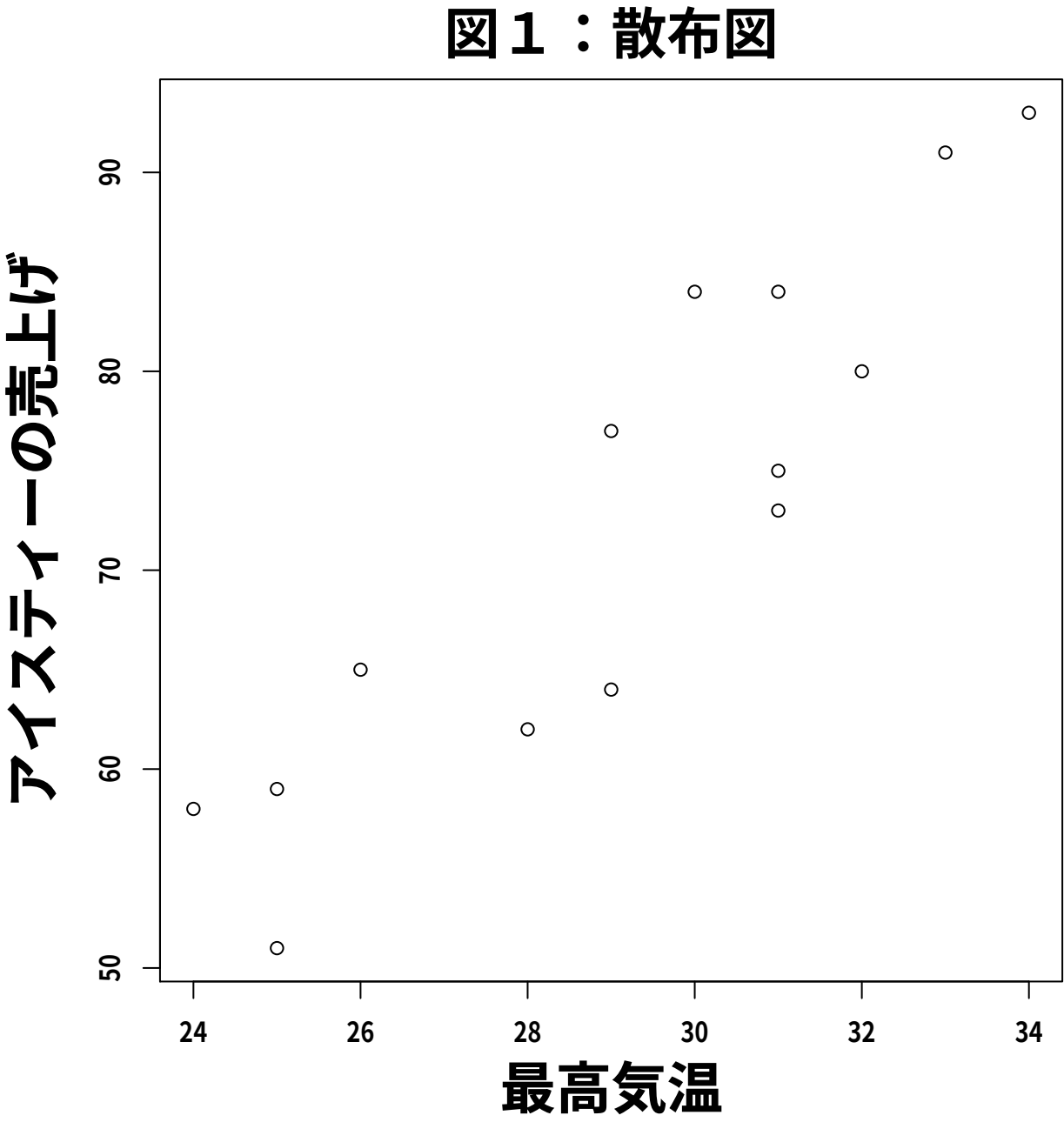

図 1: 図１：散布図

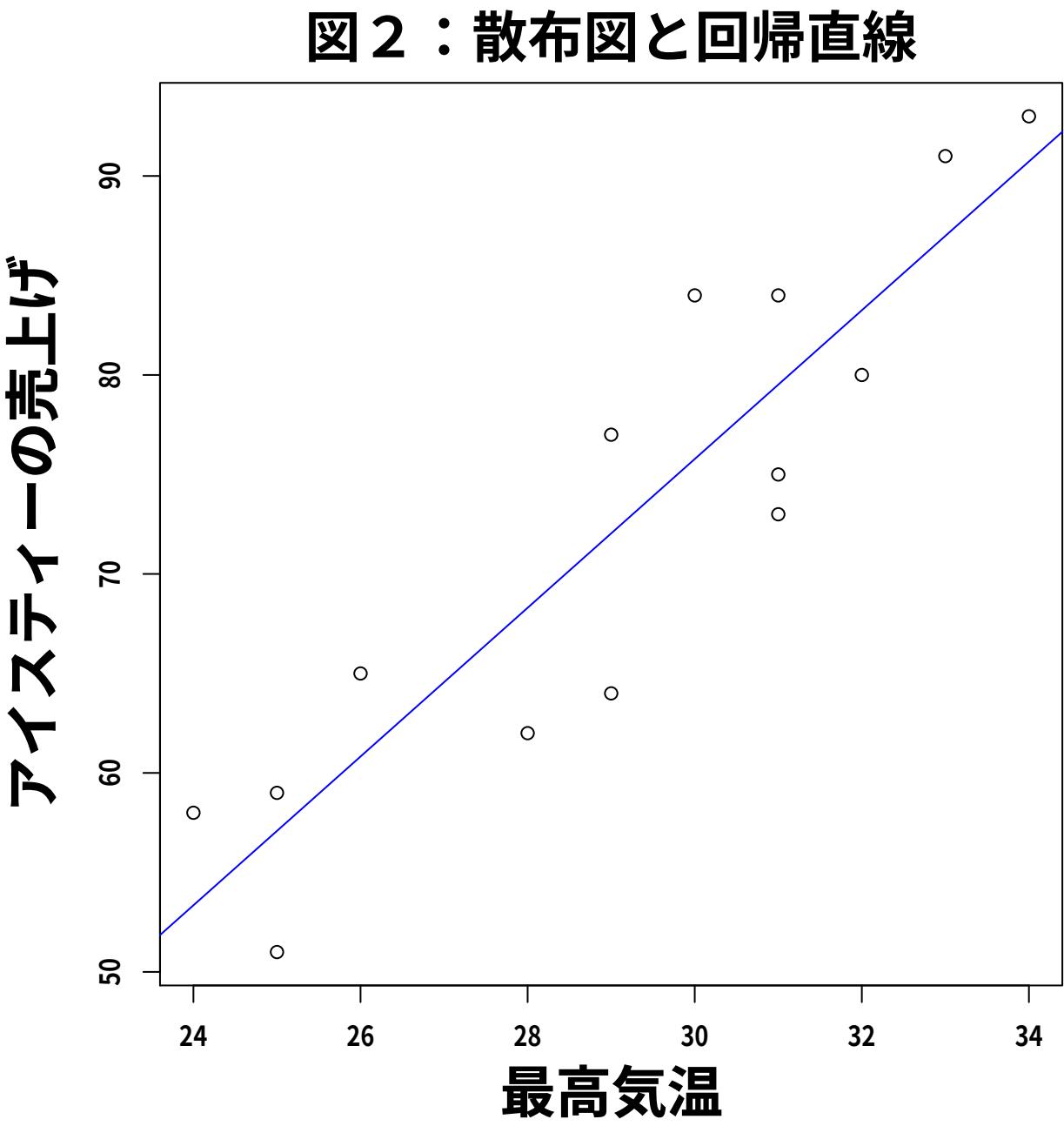

図 2: 図２：散布図と回帰直線

1.3　回帰式の信頼性：決定係数

先の回帰式がどの程度信頼できるかをみる一つの手がかりは、決定係数である。

- 決定係数: 0.8225093
- F 検定の値: 55.60917 (決定係数に関する F 検定量)
- F 検定の p 値: 7.661413e-06

決定係数は、 1 に近いほど信頼性が高い。

今回の 0.8225093 という値は、「アイスティーの売上げ」の変動が「最高気温」の変動によって 82.3 %だけ説明できることを意味している。

残りの 17.7 %は、回帰式に含まれていない別の要素が関係しているとみられる。

決定係数の信頼性は F 検定の結果により判断する。もっともっと幅広く調査したとき、もしかすると決定係数が 0 になるかもしれない。その危険性を示すのが F 検定の p 値である。これが 0.05(5%) とか 0.01、あるいは 0.001 より小さければ、そのレベルに応じた有意性が認められることになる。

この p 値が 0.05 以上だと、「母集団において決定係数が 0 である」という仮説を、母集団に関して棄却できないことになる。

1.4　回帰式の信頼性：回帰係数

回帰係数 (coefficient：y 切片や x の傾きの関連) に関する詳細な情報を掲げると次のとおり。

	見積り	標準誤差	t 検定値	t 検定時の p 値
定数項	-36.3612334801762	14.6872670894724	-2.47569770867988	0.0291872649368303
最高気温	3.73788546255507	0.501248142954886	7.45715573232854	7.66141280445013e-06

この表で、 1 行目の「定数項」の「見積り」は y 切片の値であり、 2 行目の「最高気温」の「見積り」の欄は x の傾きの値である。

重回帰分析の場合は、説明変数の個数に応じて、 3 行目・ 4 行目・……が表示されることになる。

表の右端の「t 検定時の p 値」は、該当の変数が実は影響力を持たないかもしれない危険性 (確率) を示す。

x の傾きが 0 であれば、x の値が何であっても y の値に影響しない。

最高気温の傾きが 3.737885 と見積もられてはいるが、もっともっと調査対象を拡げると、実は 0 ということもあるかもしれない。その「かもしれない」の確率が p 値である。p 値が十分小さければ、有意性が認められること

になる。

　定数項のｐ値については、実際に言及することは少ないと思うが、もっともっと幅広く調査した時にｙ切片が０であるかもしれない確率を示す。