Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

金田　拓

2 Graphical data exploration

✓ 2.3 Visualizing two or more variables　pp. 34~36　(in pp. 32~37)

```
> lines(lowess(ratings$Frequency, ratings$FamilySize), col="darkgrey")
```

このコマンドで、散布図に頻度と相関を示す回帰線を引くことができる。

このような滑らかな曲線を”scatterplot smoother”と呼び、lowess[1]( )関数を用いてデータの X 軸、Y 軸座標を格納し、lines( )関数で実線を引く。

基本的な考え方としては、X 軸の間隔ごとに Y 軸の平均上昇率を求め、それを横につないでいく。ヒストグラムの密度推定の考え方と同様。間隔が不適切と感じた場合、自分で修正することが可能。方法については、オンラインヘルプや Venables and Ripley (2002: 228-232)[2]などを参照すること。

次に、この図を点ではなく、具体的な項目名で表記することが可能である。

Step 1

```
> plot(ratings$Frequency, ratings$FamilySize, type = "n",
+ xlab = "Frequency", ylab = "Family Size")
```

これは、データフレーム ratings の列、Frequency と FamilySize を基にグラフを書くが、実際の値は入力しない。X 軸に「Frequency」、Y 軸に「Family Size」というラベルを貼る、というコマンドである。

Step 2

```
> text(ratings$Frequency, ratings$FamilySize, as.character(ratings$Word),
+ cex = 0.7)
```

これは、text( )関数を用いて先ほどのグラフに値を入力するコマンド。最初の 2 つが X 軸、Y 軸座標情報で、後の 3 つ目が実際に書き込まれる値。as.character( )をつけているのは、ベクトルが数値でなく、確実に「文字列」として扱われるためである。

cex は、こうして記入する文字列のフォントサイズ指定のコマンド。

---

[1] LOWESS とは”LOcally WEighted Scatterplot Smoother”のことらしい
Cleveland, W.S. (1979). Robust locally weighted regression and smoothing. *Journal of the American Statistical Association 74* (368), 829-836.

[2] 教科書の Refernces では 2003 年の出版となっているが、どうやら **2002 年の誤植**である。
Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S-Plus 4th edition*. New York: Springer.