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

金田　拓

4 Basic statistical methods

✓ **4.3 Paired vectors　pp. 86~91　　(in pp. 82~101)**

4.3.2.2 Estimating slope and intercept　pp. 86~87

- 線形モデルの切片と傾きを、lm( )を使って求めることができる。
- lm( )には 2 つの引数が必要で、従属変数と予測変数で 1 つ、それにデータがもう 1 つの引数となる。

```
> ratings.lm = lm(meanSizeRating ~ meanWeightRating, data = ratings)
```

- 上のコマンドで、meanSizeRating（従属変数）は、meanWeightRating（予測変数）によって決定されることを入力した。

- ratings.lm とコマンド入力することで、代入した値の最小二乗回帰線の係数を見ることができる。（最小二乗法は、個々の点と線との、縦座標の差の 2 乗を最小にすることで切片と傾きを求める方法である）

```
> ratings.lm
```

- coef( )で、傾きと切片のみ取り出すことも可能。

4.3.2.3 Correlation　pp.87~89

・ Figure 4.10 (p.88)のような図を 2 変量正規分布という。
・ Figure4.10 では、点線が Y=X の線、実線が回帰線である。
・ 回帰線の周りにどれくらい値があるかという指標を相関と呼び、相関係数で表す。
・ 母集団の相関係数が ρ 、母集団から取り出したサンプルの相関を r とする。相関係数は-1 から+1 までの値をとり得る。
・ 相関係数の値が高いほど、回帰線の周りに値が集まる。
・ 最終的に予測変数から従属変数を導くのがモデルの目的であるため、相関係数が高いほど、予測変数から正確に予想が可能という意味で良いモデルになる。

- Figure4.10(p. 88)のような散布図は、mvrnormplot.fnc()で描くことができる。

```
> mvrnormplot.fnc(r = 0.9)
```

r の値を変えることで、相関の強さに従って散布図がどのように変わるかを観察できる。

4.3.2.4 Summarizing a linear model object

- summary()でモデルの説明。これに R^2 に関する情報が含まれている

```
> summary(ratings.lm)
```

引数に格納されているデータの要旨が最初に出てくる。

- 係数の載っているところを見ると、切片が 0.527、傾きが 0.926 となっている。
- 標準誤差、 *t* 値、 *p* 値も出力される。
- summary(ratings.lm)に、$をつけて列の値を個々に計算することもできる。

```
> summary(ratings.lm)$coef
```

names(summary(ratings.lm))で、列の名前を確認できる。

- これは行列(matrix)なので、 *t* 値や相関係数を求めたりすることができる

```
> summary(ratings.lm)$coef[ ,3]
> summary(ratings.lm)$coef[ ,1]
```

- いちいち計算せずに、$で取り出せるようにするためには、データフレームに変換してやる必要がある。

```
> data.frame(summary(ratings.lm)$coef)$Estimate
```

- summary の説明に戻ると、残差標準誤差(Residual Standard Error)はモデルがどれくらい不適合かという指標。モデルが良ければ良いほど値は小さくなる。
- その次の R=0.9986 とは、*r* の 2 乗のこと。cor()で求めることができる。

```
> cor(ratings$meanSizeRating, ratings$meanWeightRating)
```

- cor.test()でピアソンの積率相関係数の検定もできる

```
> cor.test(ratings$meanSizeRating, ratings$meanWeightRating)
```

- method="spearman"のオプションを追加すると、ノンパラメトリック検定のスピアマンの順位相関係数で計算可能。

```
> cor.test(ratings$meanSizeRating, ratings$meanWeightRating, method = "spearman")
```

・ 例によってタイについて警告が出る(jitter で回避できる）が、p 値が低ければ大した問題ではないといえる。
・ スピアマンの相関係数は rs という形で表わされることが多い。
・ ratings.lm の値のうち、次章で F 値を扱う。通常 F 値が大きいほど p 値が小さくなる。