Corpus Statistics
2008/11/18
Baayen (2008: pp. 91-97)

Toru Hitomi

4.3.2.5 Problems and pitfalls of linear regression

このセクションでは、線形回帰における問題や不測の事態とその対処法を 2 つ例示する。

❖ e.g.①データセット(ratings)での名詞の単数形の使用頻度から複数形の使用頻度を予測する。

→Figure 4.11 left panel のような分布に非常に偏りがある場合

➢ 問題点:lm()という関数を使用してしまうと、**外れ値(outliers)**が回帰の傾きに大きく影響してしまう。

① 散布図を作成し、回帰線を引いてみる。

```
> plot(ratings$FreqSingular, ratings$FreqPlural)
> abline(lm(FreqPlural~FreqSingular, data=ratings), lty=1)
```

cf.

```
> lm(FreqPlural~FreqSingular, data=ratings)
```

→

```
Coefficients:
(Intercept)  FreqSingular
    31.6276        0.6018
```

② 外れ値の値を除いた回帰線を引いてみる。

```
> abline(lm(FreqPlural~FreqSingular,
+ data=ratings[ratings$FreqSingular < 500, ][T1],), lty=2)
```

```
> lm(FreqPlural~FreqSingular, data=ratings[ratings$FreqSingular < 500, ])
```

→

```
Coefficients:
(Intercept)  FreqSingular
     9.8377        0.8922
```

③ 外れ値に対して頑健であるとされる関数 lmsreg()[T2]を用いて回帰線を引いてみる。

```
> abline(lmsreg(FreqPlural~FreqSingular, data=ratings), lty=3)
```

→

```
Coefficients:
(Intercept)  FreqSi  ngular
    10.0346        0.5435
Scale estimates 27.75 28.14
```

➢ 解決策:対数化し、外れ値を他の値の群へ組み込むことで、重大な歪度を矯正する。

① それぞれの頻度を対数化し、散布図を作成する。

```
> singular.log = log(ratings$FreqSingular)
> plural.log = log(ratings$FreqPlural)

> plot(singular.log, plural.log)
```

❖ e.g.② RQ:あるものがどれほど重いかという認識(meanSizeRating)を、その言語の中でモノの名前がどれほどの頻度かという認識(meanFamiliarity)から予測することが可能か?

① lm()を用いて、傾きと切片を算出する。

```
ratings.lm = lm(meanSizeRating ~ meanFamiliarity, data = ratings)
```

② summary()を用いて、基礎統計量を算出する。

```
> round(summary(ratings.lm)$coef, 4)[T3]
                 Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)        3.7104      0.4143   8.9549    0.0000
meanFfamiliarity  -0.2066      0.1032  -2.0014    0.0488
```

③ Figure 4.12 の散布図を作成する。(別紙参照)

④ plants の名詞だけで傾きと切片を算出する。

```
> plants.lm = lm(meanSizeRating~meanFamiliarity + I(meanFamiliarity^2), data=plants)
> summary(plants.lm)$coef
                    Estimate  Std. Error  t value     Pr(>|t|)
(Intercept)       5.1902476  1.28517759 4.038545 0.0003142449
meanFamiliarity  -1.6717053  0.59334724 -2.817415 0.0082290129
I(meanFamiliarity^2) 0.2030369  0.06659252 3.048944 0.0045826280
```

Toru Hitomi

$$meanSizeRating = 5.19 - 1.67 * meanFamiliarity + 0.20 * meanFamiliarity^2$$

- POINT! “linear”：重み付けされた予測（独立）変数の総和によって従属変数が表現される。
  ＝従属変数は、予測変数の **linear combination（線形結合）**である。
- まとめ
  - 視覚化！
  - 外れ値に注意！
  - 直線を押し付けるな！
  - モデルはシンプルに！