モデルの当てはめ

- 回帰モデル：従属変数のばらつきを独立変数のばらつきで説明。
- 独立変数は複数の変数の線形結合でもOK。
- 独立変数が1つのとき単回帰分析，2つ以上のとき重回帰分析。
- 例えば，収縮期血圧値SBPを，塩分摂取量SALTと年齢AGEで説明するモデルの場合，
  SBP = β0 + β1 SALT + β2 AGE + ε
  という形になる（εは誤差項。β1，β2は偏回帰係数）
- 偏回帰係数の推定の際，多重共線性には注意（例えば，VIFを使ってチェックする）。（EZRは自動的にvifを計算してくれる）
  (cf.) library(fmsb); ?VIF

重回帰モデルの結果の見方

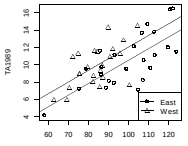
```
Call:
lm(formula = Ozone ~ Wind + Temp)

Residuals:
    Min      1Q  Median      3Q     Max
-41.251 -13.695  -2.856  11.390 100.367

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -71.0332    23.5780  -3.013   0.0032 **
Wind         -3.0555     0.6633  -4.607 1.08e-05 ***
Temp          1.8402     0.2500   7.362 3.15e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.85 on 113 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared: 0.5687,     Adjusted R-squared: 0.5611
F-statistic:  74.5 on 2 and 113 DF,  p-value: < 2.2e-16

> stb 標準化偏回帰係数
(Intercept)        Wind        Temp
  0.0000000  -0.3311197   0.5291333
```

共分散分析

- 量的変数XとY，二値変数Bがあるとき
- 「YにBの2群間で差が無い」を検定したいけれども，XとYに相関があり，Xの大きさを調整した上でBの2群間での比較をしたいとき，
  Y = β0 + β1 X + β2 B + β12 X*B + ε
  というモデルを考える。
- β12がゼロと有意差があるとき→BによってXとYの関係は異なる
- β12が有意でないとき→交互作用は有意でない
  Y = β0 + β1 X + β2 B + ε
  を考える。β2がゼロと有意差があれば最初に立てた帰無仮説は棄却できる。

共分散分析の結果の見方

- summary(east)で，CAR1990の係数は0.1346(p=5.7e-5)
  summary(west)で，CAR1990の係数は0.1352(p=0.0026) ともに5%水準で有意
- summary(lm(TA1989 ~ REGION*CAR1990, data=x))の結果から，REGIONWest:CAR1990の係数が0.00062(p=0.99)なので，REGIONとCAR1990の交互作用効果は5%水準で有意でない。つまりCAR1990のTA1989への影響はREGION間で差が無い→共分散分析へ



- summary(ac)の結果でAdjusted R-squaredが0.4488なので，共分散モデルは，データのばらつきの約45%を説明しているといえる。また

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.95787    2.21349  -1.336   0.1883
REGIONWest   1.52190    0.68689   2.216   0.0319 *
CAR1990      0.13475    0.02177   6.189 1.78e-07 ***
```

5%有意

＊100世帯当たり車保有が1台増えると10万人当たり交通事故が0.135件増え，車保有台数が同じなら西日本は東日本より1.52件事故が多い

- 最後に出てくるEast 9.4446とWest 10.9665が「修正平均」
- EZRでもモデルを保存し，上記acの代わりに保存したモデル名入力でOK

結果の解釈

対数オッズ比

```
Call:
glm(formula=y~ap+hilo+week, family = binomial, data = bacteria)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3763   0.3813   0.5212   0.6576   1.1194

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9278     0.3762   5.124 2.99e-07 ***
ap           -0.8343     0.3816  -2.186  0.02879 *
hilolo       -0.5066     0.3546  -1.428  0.15317
week         -0.1167     0.0443  -2.633  0.00845 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 217.38  on 219  degrees of freedom
Residual deviance: 202.90  on 216  degrees of freedom
AIC: 210.9

Number of Fisher Scoring iterations: 4
```

有意水準5%で有意にオッズ比が1と異なる

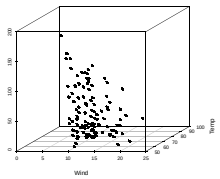
（例）ニューヨーク気象データ：オゾン濃度の分析

- 単回帰分析では

```
attach(airquality)
plot(Ozone ~ Wind)
res1 <- lm(Ozone ~ Wind)
summary(res1)
detach(airquality)
```

としていた。

- 重回帰では風速(Wind)だけでなく，例えば気温(Temp)のオゾン濃度(Ozone)への影響も同時にみることができる




```
attach(airquality)
library(rgl) # 動かせる
plot3d(Wind,Temp,Ozone)
library(scatterplot3d) # きれいな出力
scatterplot3d(Wind,Temp,Ozone,pch=20,angle=30,scale.y=0.5)
res2 <- lm(Ozone ~ Wind + Temp)
summary(res2)
sdd <- sapply(res2$model, sd)
stb <- coef(res2)*sdd/sdd[1]
stb
detach(airquality)
```

AICと尤度比検定

- モデルの当てはまりの悪さの指標AIC
  `> AIC(res1) # 1093.187`となる
  `> AIC(res2) # 1049.741`となる。
  ※EZRでは`extractAIC()`の結果がAICとして表示されるので若干AICの定義が異なる。相対的大小関係はどちらでも変わらない
  http://minato.sip21c.org/msb/medstatbookx.pdf p.208
  重回帰モデルの方が当てはまりは良い→有意性は？
- Ozone = β0 + β1 Wind + ε (1)
  Ozone = β0 + β1 Wind + β2 Temp + ε (2)
  (1)は，(2)でβ2が0という特殊例とみなせる
- 一般性がより低いモデル(1)の最大尤度の，一般性がより高いモデル(2)の最大尤度に対する比の対数をとってマイナス2倍した値が近似的にカイ二乗分布に従うことから，「(1)と(2)で当てはまりに差が無い」帰無仮説を尤度比検定できる

```
LL1 <- logLik(res1)
LL2 <- logLik(res2)
lambda <- -2*(as.numeric(LL1)-as.numeric(LL2))
dff <- attr(LL2,"df")-attr(LL1,"df")
1-pchisq(lambda, dff)
```

共分散分析の例

- 日本の各都道府県(PREF)の，
  1990年の100世帯当たり乗用車台数(CAR1990)，
  1989年の人口10万人当たり交通事故死者数(TA1989)，
  1985年の国勢調査による人口集中地区居住割合(DIDP1985)
  を含むhttp://minato.sip21c.org/grad/sample3.dat
  を読み込み，東日本か西日本か(REGION)間で交通事故死者数を比較する。ただし乗用車台数は交通事故死者数と関連していると思われるので，その影響を共変量として調整する共分散分析実行
- コードは下記（EZRでは統計解析＞連続変数の解析＞共分散分析）

```
x <- read.delim("http://minato.sip21c.org/grad/sample3.dat")
plot(TA1989 ~ CAR1990, pch=as.integer(REGION), data=x)
east <- lm(TA1989 ~ CAR1990, data=subset(x, REGION=="East"))
summary(east); abline(east, lty=1)
west <- lm(TA1989 ~ CAR1990, data=subset(x, REGION=="West"))
summary(west); abline(west, lty=2)
legend("bottomright",pch=1:2,lty=1:2,legend=c("East","West"))
summary(lm(TA1989 ~ REGION*CAR1990, data=x))
ac <- lm(TA1989 ~ REGION+CAR1990, data=x)
summary(ac)
cfs <- dummy.coef(ac)
cfs[[1]] + cfs$CAR1990 * mean(ac$model$CAR1990) + cfs$REGION
```

ロジスティック回帰分析

- 目的変数（応答変数）が2値データ（イベントが起こる／起こらない等）で，正規分布ではなく二項分布に従う回帰モデル。glm()を用いる。
- （例）MASSライブラリのbacteriaデータ（オーストラリア）
  - http://www.menzies.edu.au/publications/anreps/MSHR00.pdf
  - Dr. A. LeachによるRCT。中耳炎の既往のある50人の子供に投薬し，定期的にH. Influenzaeの検出をチェックした結果
  - 変数は，y（菌の有無。nが無し，yが有り），ap（薬の種類。aが実薬，pがプラセボ），hilo（服薬遵守を促す程度，hiかlo），week（研究開始からの週数），ID（個人番号），trt（apとhiloを組み合わせた処理種類，"placebo"，"drug"，"drug+"）
- 投薬と服薬遵守を促す程度と週数が菌の有無に影響するかどうかを調べるデザインなので，目的変数が菌の有無という2値変数であり，ロジスティック回帰分析になる。

```
library(MASS)
res <- glm(y ~ ap+hilo+week, binomial, data=bacteria)
```

作表方法

- ロジスティック回帰分析の結果の表には，対数オッズ比を掲載しても仕方ないので，通常は指数関数を使ってオッズ比に戻す。
  exp(coef(res)) # オッズ比の点推定量を表示
  exp(confint(res)) # 95%信頼区間を表示
- ~~下のように作表し，下にAICやNagelkerkeのR²を付記~~

		95%信頼区間		
説明変数	オッズ比	下限	上限	p値
プラセボ投与	2.30	1.11	5.03	0.029
遵守指導低	0.60	0.29	1.21	0.153
週数	0.89	0.81	0.98	0.008

- プラセボ投与群は治療薬投与群に比べ，菌が存在する確率が2.3倍あり，1週間経つごとに菌が存在する確率が0.89倍になると解釈できる
- EZRでは自動的にここまで出る。AICも出る。
  library(fmsb)； NagelkerkeR2(res)