Wybór zmiennych
Cele:

1. Prognoza zmiennej zależnej. Zalecana AIC
2. Wybór dla opisu zależności, niekoniecznie dla prognozy Zalecane BIC.
3. Chcemy dobrej estymacji parametrów VIF.

Kryteria GIC

$$GIC(M) = -2\log L(M|y, X) + h|M|$$

gdzie

- M - model regresji
- L - funkcja wiarygodności
- h - współczynnik - im mniejszy tych bogatszy model preferowany. Dla AIC $h = 2$, dla BIC $h = \log(n)$
- $|M|$ - liczba parametrów modelu

Przy $n \to \infty$ w tempie $O(\log(n))$ model M jest zgodny.

PRZYKŁAD

```
metabolizm.full <- lm(m~g*F*A,data=alkohol)

> AIC(metabolizm.full)
[1] 114.1
> AIC(metabolizm.full,k = log(nrow(alkohol)))
[1] 127.3

metabolizm.zero <- lm(m~.,data=alkohol)

> AIC(metabolizm.zero)
[1] 115.9
> AIC(metabolizm.zero,k = log(nrow(alkohol)))
[1] 123.3

> metabolizm.step <- step(metabolizm.full)
Start:  AIC=21.29
m ~ g * F * A
```

```
        Df Sum of Sq    RSS    AIC
- g:F:A  1    0.2515 38.006 19.504
<none>               37.754 21.291

Step:  AIC=19.5
m ~ g + F + A + g:F + g:A + F:A

       Df Sum of Sq    RSS    AIC
- F:A   1    0.3622 38.368 17.807
<none>              38.006 19.504
- g:A   1    2.7817 40.787 19.764
- g:F   1   11.2424 49.248 25.796

Step:  AIC=17.81
m ~ g + F + A + g:F + g:A

       Df Sum of Sq    RSS    AIC
- g:A   1    2.4439 40.812 17.783
<none>              38.368 17.807
- g:F   1   10.9828 49.351 23.863

Step:  AIC=17.78
m ~ g + F + A + g:F

       Df Sum of Sq    RSS    AIC
- A     1     0.001 40.813 15.784
<none>              40.812 17.783
- g:F   1    10.517 51.329 23.120

Step:  AIC=15.78
m ~ g + F + g:F

       Df Sum of Sq    RSS    AIC
<none>              40.813 15.784
- g:F   1    10.587 51.400 21.165

> summary(metabolizm.step)

Call:
lm(formula = m ~ g + F + g:F, data = alkohol)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.4427 -0.6111 -0.0326  0.5436  3.8759

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.1858     0.7117  -1.666   0.1068
g             2.3439     0.2801   8.367 4.22e-09 ***
F             0.9885     1.0724   0.922   0.3645
g:F          -1.5069     0.5591  -2.695   0.0118 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.207 on 28 degrees of freedom
Multiple R-squared: 0.8137,Adjusted R-squared: 0.7938
F-statistic: 40.77 on 3 and 28 DF,  p-value: 2.386e-10
```

Taki sam wynik, gdy zastosuje się kryterium BIC.

```
metabolizm.stepBIC <- step(metabolizm.full,k = log(nrow(alkohol)))
```

Porównanie ostatniego modelu z oszczędniejszym

```
> metabolizm.step0 <- lm(m~g+g:F-1,data=alkohol)
> summary(metabolizm.step0)

Call:
lm(formula = m ~ g + g:F - 1, data = alkohol)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5052 -0.6753 -0.0934  0.3630  4.3959

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
g     1.9278     0.1288  14.971 1.85e-15 ***
g:F  -1.2021     0.2165  -5.553 4.89e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.224 on 30 degrees of freedom
Multiple R-squared: 0.8895,Adjusted R-squared: 0.8821
F-statistic: 120.8 on 2 and 30 DF,  p-value: 4.467e-15

> anova(metabolizm.step,metabolizm.step0)
Analysis of Variance Table

Model 1: m ~ g + F + g:F
Model 2: m ~ g + g:F - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 40.813
2     30 44.947 -2   -4.1345 1.4182  0.259
```

VIF[Jobson, 277]

Estymator współczynników regresji (bez wyrazu wolnego) może być zapisany za pomocą macierzy korelacji:

$$b^* = s_y D^{-\frac{1}{2}} R^{-1} r_{Xy}$$

gdzie r_{Xy} jest p -wymiarowym wektorem korelacji między y a macierzą planu X (bez kolumny jedynek), zaś D jest przekątniową macierzą wariancji zmiennych objaśniających.

Elementy na przekątnej R^{-1} są postaci $V_j = 1/(1-R_j^2)$. Wielkość R_j^2 jest współczynnikiem determinacji w regresji X_j względem pozostałych zmiennych objaśniających. V_j są współczynnikami inflacji wariancji (VIF). Mamy bowiem relacje

$$V(b_j^*) = \frac{\frac{\sigma^2}{n-1}}{(1-R_j^2)s_{x_j}^2} = \frac{\frac{\sigma^2}{n-1}}{s_{x_j}^2} V_j$$

$$d^2(b^*, \beta^*) = \sigma^2 \sum_{j=1}^{p} V_j s_{x_j}^2$$

Dla dużych wartości VIF zarówno wariancja estymatorów współczynników regresji jak i dystans wektora tych współczynników od prawdziwych wartości są duże.

Wartość VIF jest uważana za dużą, gdy przekracza 10. W R wyznacza ją funkcja `vif{car}`.

PRZYKŁAD

```
vif(lm(m~.,data=alkohol))
 g    F    A
```

```
1.31 1.31 1.18

vif(metabolizm.full)
  g     F     A    g:F   g:A   F:A  g:F:A
 2.16  8.89 14.44  6.15 10.84 33.36 30.19

vif(metabolizm.step)
 g    F   g:F
1.55 6.21 5.39
```

Co zrobić, gdy jest duża? [Biecek, 126]:

1. Usuń najbardziej skorelowane zmienne.
2. Przeprowadź regresję grzbietową (w R: `lm.ridge{MASS}`).
3. Wprowadź nowe, nieskorelowane zmienne (składowe główne ze zmiennych objaśniających).