Permutacyjna metoda oceny istotności regresji

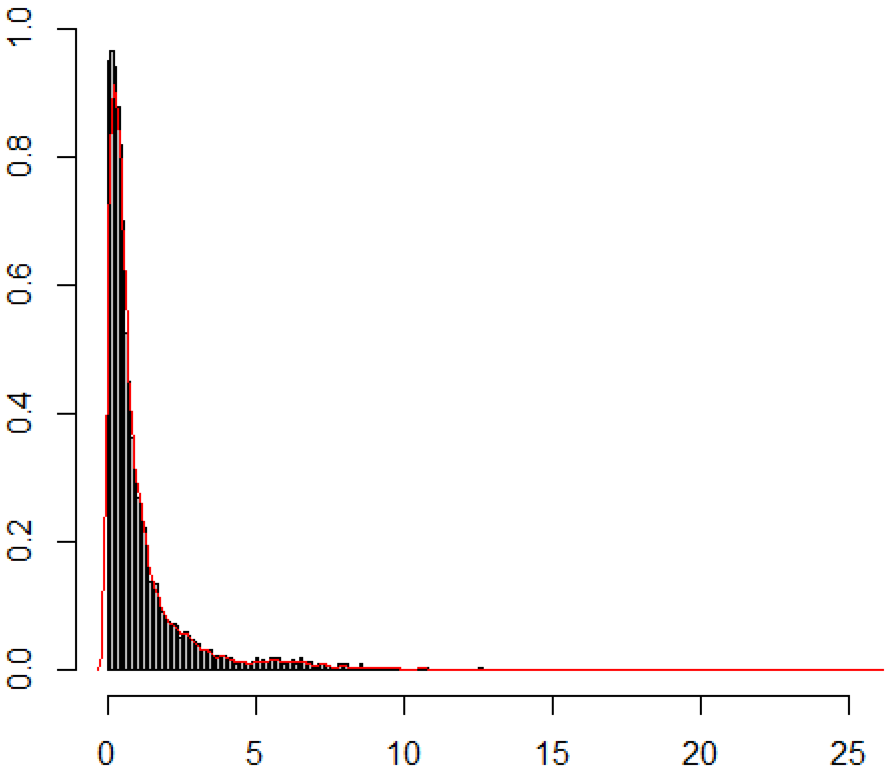
(bez założenia normalności)

```
f <-numeric(10000)
for (i in 1:10000){
  g <- lm(sample(gala$Species)~elevation+adjacent,data=gala.eaa)
  f[i] <- summary(g)$fstatistic[1]
}
length(f>37.94)/10000
quantile(f,probs = c(0.95,0.99,0.999))
histdens(f)
```

```
> length(f>37.94)/10000
[1] 1
> quantile(f,probs = c(0.95,0.99,0.999))
      95%       99%     99.9%
 4.691446  8.742880 14.156021
```

Wyniki regresji

```
> g2 <- lm(species~elevation+adjacent,data=gala.eaa)
> anova(g2)
Analysis of Variance Table

Response: species
          Df Sum Sq Mean Sq F value    Pr(>F)
elevation  1 207828  207828  56.112 4.662e-08 ***
adjacent   1  73251   73251  19.777 0.0001344 ***
Residuals 27 100003    3704
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> 207828+73251
[1] 281079

> 281079/2
[1] 140539.5

> 1-pf(140539/3704,2,27)
[1] 1.434361e-08
```



Niech

$$\varepsilon \sim N(0, \sigma^2 I)$$

$$y \sim N(X\beta, \sigma^2 I)$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \sim N(\beta, (X^T X)^{-1} \sigma^2)$$

Uwagi
W wielu przypadkach, wiemy, że punktowa hipoteza zerowa jest fałszywa, nawet nie patrząc na dane. Ponadto wiemy, że im więcej danych, tym większa moc testów.
Nawet niewielkie różnice od zera zostaną wykryte przy dużej próbce.
Zgodnie z tym poglądem, test hipotezy prostu staje się testem wielkości próby.
Z tego powodu lepszym argumentem jest przedział ufności.

Confidence Intervals for β

We start with the simultaneous regions. Some results from multivariate analysis show that

$$\frac{(\hat\beta-\beta)^T X^T X(\hat\beta-\beta)}{\sigma^2} \sim \chi^2_p$$

and

$$\frac{(n-p)\hat\sigma^2}{\sigma^2} \sim \chi^2_{n-p}$$

and these two quantities are independent. Hence

$$\frac{(\hat\beta-\beta)^T X^T X(\hat\beta-\beta)}{p\hat\sigma^2} \sim \frac{\chi^2_p/p}{\chi^2_{n-p}/(n-p)} \equiv F_{p,n-p}$$

So to form a $100(1-\alpha)$ % confidence region for β, take β such that

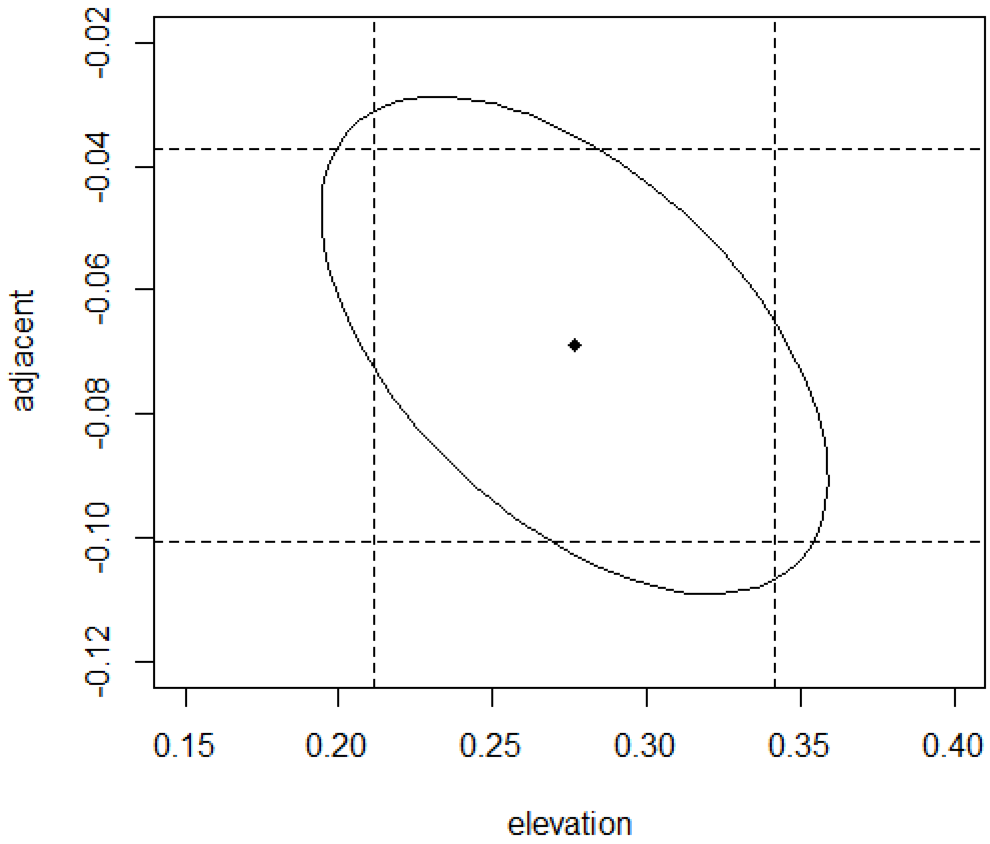
$$(\hat\beta-\beta)^T X^T X(\hat\beta-\beta) \le p\hat\sigma^2 F^{(\alpha)}_{p,n-p}$$

dla pojedynczych parametrów

$$\hat\beta_i \pm t^{(\alpha/2)}_{n-p}\hat\sigma\sqrt{(X^TX)^{-1}_{ii}}$$

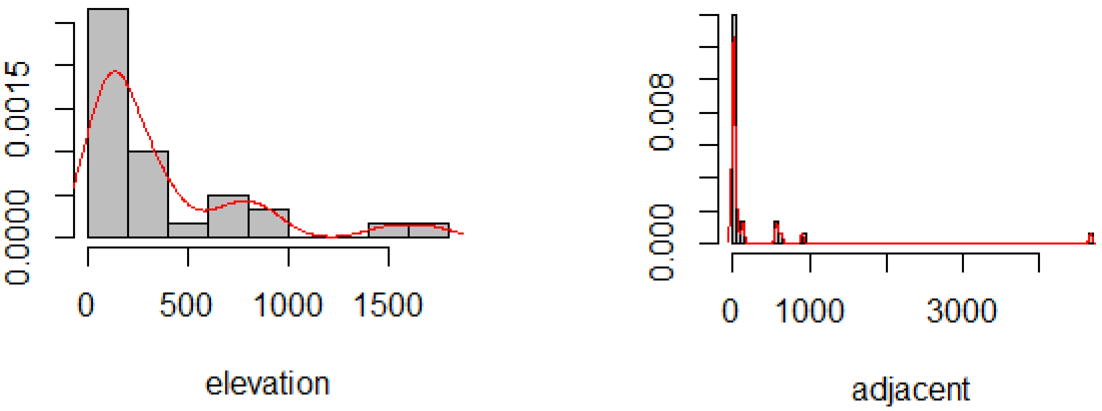
```
> g2 <- lm(species~elevation+adjacent,data=gala.eaa)
> summary(g2)

Call:
lm(formula = species ~ elevation + adjacent, data = gala.eaa)

Residuals:
    Min      1Q  Median      3Q     Max 
-103.41  -34.33  -11.43   22.57  203.65 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.43287   15.02469   0.095 0.924727    
elevation    0.27657    0.03176   8.707 2.53e-09 ***
adjacent    -0.06889    0.01549  -4.447 0.000134 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 60.86 on 27 degrees of freedom
Multiple R-squared:  0.7376,    Adjusted R-squared:  0.7181 
F-statistic: 37.94 on 2 and 27 DF,  p-value: 1.434e-08

> qt(0.975,27)

[1] 2.051831
> c(0.27657-2.051831*0.03176,0.27657+2.051831*0.03176)#elevation
```

```
[1] 0.2114038 0.3417362
> c( -0.06889-2.051831*0.01549, -0.06889+2.051831*0.01549)#adjacent
[1] -0.10067286 -0.03710714

> plot(ellipse(g2,c(2,3)),type="l",xlim=c(0.15,0.4),ylim=c(-0.12,-0.02))

> points(g2$coef[2],g2$coef[3],pch=18)

> abline(v=c(0.27657-2.051831*0.03176,0.27657+2.051831*0.03176),lty=2)
> abline(h=c( -0.06889-2.051831*0.01549, -0.06889+2.051831*0.01549),lty=2)
```




```
Dlaczego przecinają elipsę ufności?
> cor(gala.eaa$elevation,gala.eaa$adjacent)
[1] 0.5364578
im wyższa wartość elevation tym wyższa wartość adjacent a więc współczynniki są
ujemnie skorelowane
```

Przedziały ufności dla prognoz




```
> x0 <- c(1,200,100)
> y0 <- sum(x0*g2$coef)
> y0
[1] 49.85798
```

dla wartości oczekiwanej:

```
> x <- cbind(1,gala.eaa[,3:4])
> x <- as.matrix(x)
> xtxi <- solve(t(x) %*% x)
> bm <- sqrt(x0 %*% xtxi %*% x0) *2.051831 * 60.86
> bm
         [,1]
[1,] 24.61241
> c(y0-bm,y0+bm)
[1] 25.24557 74.47039
```

dla indywidualnej wartości

```
> bm <- sqrt(1+x0 %*% xtxi %*% x0) *2.051831 * 60.86
> bm
         [,1]
[1,] 127.2768
> c(y0-bm,y0+bm)
[1] -77.41886 177.13483
```

INACZEJ

```
> predict(g2,data.frame(elevation=200,adjacent=100),se=T)
$fit
       1
49.85798

$se.fit
[1] 11.99514

$df
```

```
[1] 27

$residual.scale
[1] 60.85898

> sqrt(x0 %*% xtxi %*% x0) * 60.85898
         [,1]
[1,] 11.99514
```

Interpretacja parametrów

```
> head(savings)
             sr pop15 pop75     dpi ddpi
Australia 11.43 29.35  2.87 2329.68 2.87
Austria   12.07 23.32  4.41 1507.99 3.93
Belgium   13.17 23.80  4.43 2108.47 3.82
Bolivia    5.75 41.89  1.67  189.13 0.22
Brazil    12.88 42.19  0.83  728.47 4.56
Canada     8.79 31.72  2.85 2982.88 2.43
…
```

sr - savings rate - personal saving divided by disposable income

pop15 - percent population under age of 15

pop75 - percent population over age of 75

dpi - per-capita disposable income in dollars

ddpi - percent growth rate of dpi

```
> g <- lm(sr~pop15 + pop75 + dpi + ddpi, data=savings)
> summary(g)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,    Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904

Co można powiedzieć o istotności pop75?
> cor(savings$pop15,savings$pop75)
[1] -0.9084787

> g2 <- lm(sr ˜ pop75 + dpi + ddpi, data=savings)
> summary(g2)
```

```
Call:
lm(formula = sr ~ pop75 + dpi + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.0577 -3.2144  0.1687  2.4260 10.0763 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.4874944  1.4276619   3.844  0.00037 ***
pop75       0.9528574  0.7637455   1.248  0.21849    
dpi         0.0001972  0.0010030   0.197  0.84499    
ddpi        0.4737951  0.2137272   2.217  0.03162 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 4.164 on 46 degrees of freedom
Multiple R-squared:  0.189,	Adjusted R-squared:  0.1361 
F-statistic: 3.573 on 3 and 46 DF,  p-value: 0.02093
```

Te zmienne są nieistotne.

```
> cor(savings$pop75,savings$dpi)
[1] 0.7869995
```

Usuńmy dpi

```
> g3 <- lm(sr ~ pop75 + ddpi, data=savings)
> summary(g3)

Call:
lm(formula = sr ~ pop75 + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.0223 -3.2949  0.0889  2.4570 10.1069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.4695     1.4101   3.879 0.000325 ***
pop75         1.0726     0.4563   2.351 0.022992 *  
ddpi          0.4636     0.2052   2.259 0.028562 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 4.122 on 47 degrees of freedom
Multiple R-squared:  0.1883,	Adjusted R-squared:  0.1538 
F-statistic: 5.452 on 2 and 47 DF,  p-value: 0.007423

> cor(savings$pop75,savings$ddpi)
[1] 0.02532138

> g4 <- lm(sr ~ pop75, data=savings)

> summary(g4)

Call:
lm(formula = sr ~ pop75, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.2657 -3.2295  0.0543  2.3336 11.8498 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.1517     1.2475   5.733  6.4e-07 ***
pop75         1.0987     0.4753   2.312   0.0251 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 4.294 on 48 degrees of freedom
Multiple R-squared:  0.1002,    Adjusted R-squared:  0.08144
F-statistic: 5.344 on 1 and 48 DF,  p-value: 0.02513
```

współczynnik przy pop75 niewiele się zmienił, bo pop75 i ddpi są prawie ortogonalne

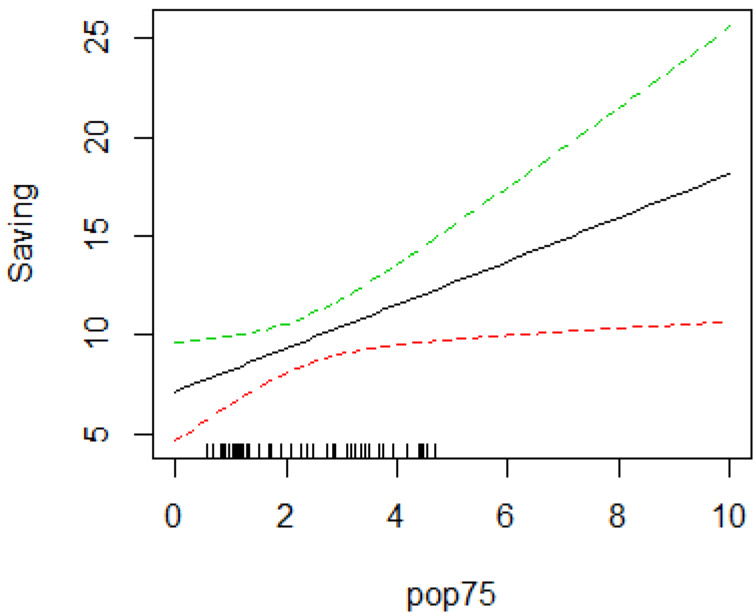
Predykcje przy różnych modelach:

```
> x0 <- data.frame(pop15=32,pop75=3,dpi=700,ddpi=3)
> predict(g,x0)
       1
9.726666
> predict(g2,x0)
       1
9.905503
> predict(g3,x0)
       1
10.07808
> predict(g4,x0)
       1
10.44777
```

Predykcje są bardziej stabilne niż estymatory przy pop75:

```
-1.70   0.95 1.07   1.10
```

```
grid <- seq(0,10,0.1)
p <- predict(g4,data.frame(pop75=grid),se=T)
cv <- qt(0.975,48)
matplot(grid,cbind(p$fit,p$fit-cv*p$se,p$fit+cv*p$se),lty=c(1,2,2),
          type="l",xlab="pop75",ylab="Saving")
rug(savings$pop75)
```



Przedział(y ufności nierealistyczne poza zakresem rzeczywistych danych