

# Statystyka w analizie i planowaniu eksperymentu

Wykład 7

Uzupełnienie informacji o regresji i zagadnienia  
pokrewne

Przemysław Biecek

Dla 1 roku studentów Biotechnologii

## Estymator

Estymator to funkcja próby.

Estymatory używane są aby na podstawie próby wyestymować (ocenić) wartość pewnego, nieznanego, parametru populacji.

Poznaliśmy już między innymi następujące estymatory:

- średniej  $\hat{\mu}$  oznaczany najczęściej przez  $\bar{x}$ ,
- częstości  $\hat{p}$ ,
- wariancji  $\hat{\sigma}^2$  oznaczany najczęściej przez  $S^2$ ,
- współczynników modelu regresji  $\hat{\beta}_1, \hat{\beta}_0$ ,
- korelacji  $\hat{\rho}$ .

Ta sama wartość może być oceniana na różne sposoby.

Przykładowo do tej pory poznaliśmy dwa estymatory wariancji,

- $S_1^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$ ,
- $S_2^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ .

Gdy estymatorów jest więcej niż jeden, naturalnymi pytaniami są:

- Który z nich jest lepszy?
- Jak porównywać te estymatory?

Dwa popularne kryteria to obciążoność i wariancja.  
Obciążenie estymatora parametru  $\theta$  wyznacza się następująco

$$bias_{\theta} = \theta E_{\theta}(\hat{\theta})$$

gdzie  $\hat{\theta}$  oznacza estymator parametru  $\theta$  (estymator a nie ocenę!).

- Estymator jest nieobciążony, jeżeli jego obciążenie wynosi 0.
- Estymator jest obciążony jeżeli jego obciążenie jest różne od zera.
- Estymator jest asymptotycznie nieobciążony, jeżeli jego obciążenie maleje do zera wraz z rozmiarem próby.

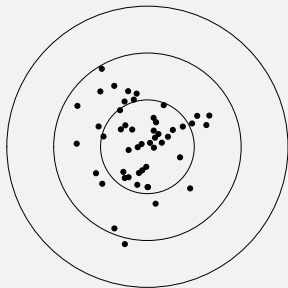
Wariancja estymatora opisuje jak duży jest rozrzut estymatora.  
Wariancje estymatora można wyznaczyć ze wzoru

$$\text{var}_\theta = E_\theta(\theta - \hat{\theta})^2.$$

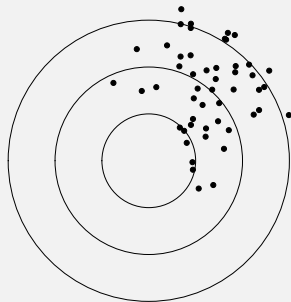
gdzie  $\hat{\theta}$  oznacza estymator parametru  $\theta$  (estymator a nie ocenę!).  
Jeżeli jakiś estymator ma mniejszą wariancję niż każdy inny estymator (dla każdego  $\theta$ ) to mówimy, że jest on estymatorem o minimalnej wariancji.

# Obciążenie estymatora

Który z tych estymatorów jest nieobciążony?



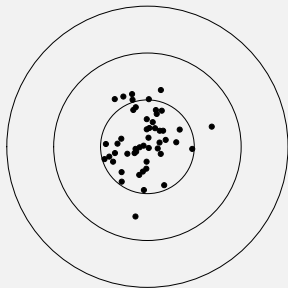
estymator A



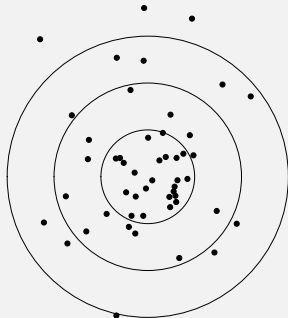
estymator B

# Wariancja estymatora

Który z tych estymatorów ma mniejszą wariancję?



estymator A

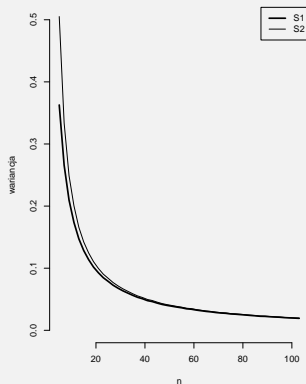
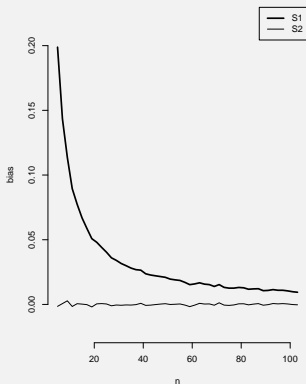


estymator B

# Porównanie estymatorów wariancji

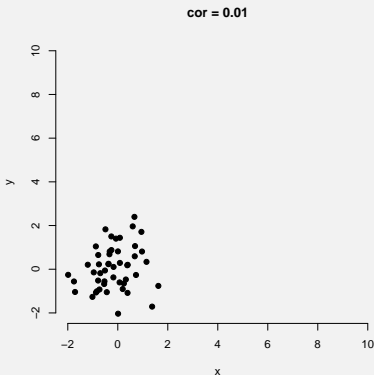
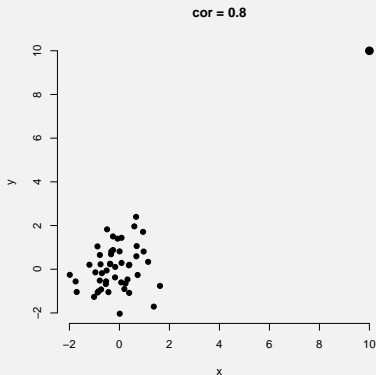
$$S_1^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$$

$$S_2^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$



# Współczynnik korelacji Spearmana

Co robić, gdy dane nie mają rozkładu normalnego lub obecne są wartości odstające?



W tej sytuacji nie należy używać współczynnika Pearsona. Używa się współczynnika Spearmana, który korelacje pomiędzy obserwacjami wyznacza na podstawie rang tych obserwacji

$$\rho_{Spearmana} = cor(r(X), r(Y))$$

Ranga obserwacji  $x_i$  odpowiada indeksowi tej obserwacji w uporządkowanej próbie.

Mamy próbę  $X$

$$X = (-53, 124, -19, -46, 87, 16, -13, 6, -68, -97)$$

i próbę  $Y$

$$Y = (-3, 117, -38, 105, 244, 115, 102, -31, -10, -136)$$

Wyznaczamy rangi dla tych elementów

$$r(X) = (3, 10, 5, 4, 9, 8, 6, 7, 2, 1)$$

$$r(Y) = (5, 9, 2, 7, 10, 8, 6, 3, 4, 1)$$

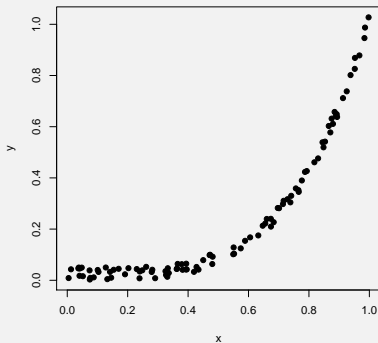
I liczymy korelację rang.

# Współczynnik korelacji Spearmana

```
> x
[1] -53 124 -19 -46 87 16 -13 6 -68 -97
> rank(x)
[1] 3 10 5 4 9 8 6 7 2 1
> cor(x,y)
[1] 0.7261815
> cor(x,y, method='pearson')
[1] 0.7261815
> cor(x,y, method='spearman')
[1] 0.7333333
> cor.test(x,y, method='spearman')
      Spearman's rank correlation rho
data: x and y
S = 44, p-value = 0.01976
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7333333
```

# Współczynnik korelacji Spearmana

Ile wynosi korelacja w takim przypadku?



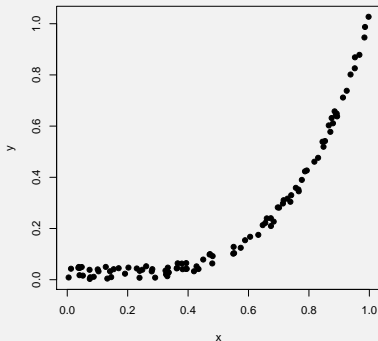
które  $\rho = 0.85$  a które  $\rho = 0.95$ ?

Wartości współczynników korelacji Pearsona i Spearmana mogą się różnić. Ale dla dużych prób, gdy rozkład obserwacji jest normalny a relacja pomiędzy zmiennymi ma charakter liniowy, to współczynniki te przyjmują podobne wartości.

Współczynnik Spearmana powinno się stosować, gdy:

- rozkład danych nie jest normalny,
- obecne są obserwacje odstające,
- nie jest dla nas ważna liniowość relacji pomiędzy zmiennymi.

Co zrobić gdy zależność pomiędzy zmiennymi wygląda na nieliniową?

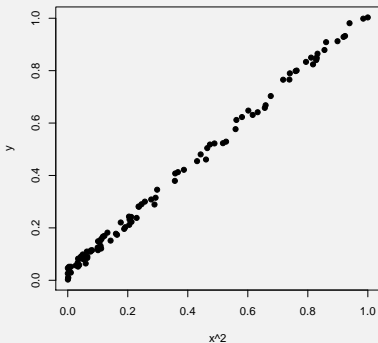
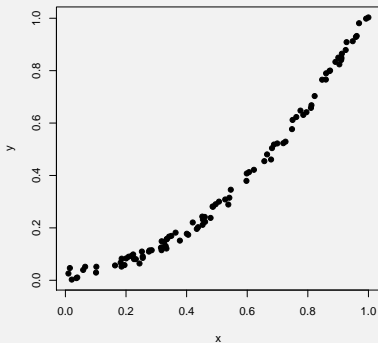


Mamy kilka możliwości:

- wykonać transformacje zmiennych objaśniających,
- wykonać transformacje zmiennych objaśnianych,
- przybliżyć zależność za pomocą prostych na fragmentach dziedziny  $x$ ,
- wybrać inny model regresji (nieliniowy).

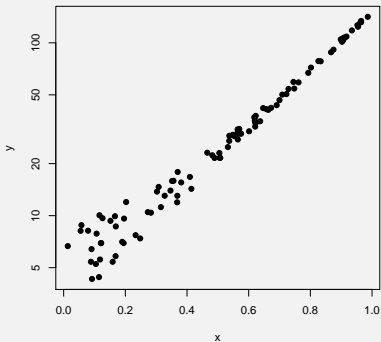
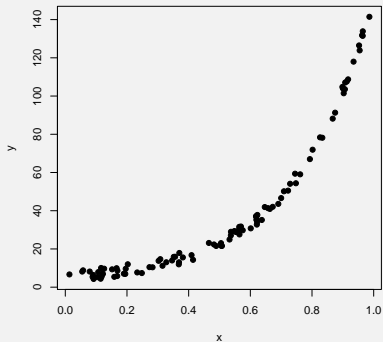
Co dają nam transformacje wielomianowe?

$$x' = x^2$$



Co daje nam logarytmowanie?

$$y' = \log(y)$$



Ogólny model gaussowskiej regresji nieliniowej

$$Y = f(X, \theta) + \varepsilon,$$

gdzie

$$\varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Znalezienie ocen parametrów w takim modelu nie jest łatwe, rozwiązania wyznaczane są korzystając z metod numerycznych.

Ale w R to jest proste!

```
> x = runif(100)*3
> y = x2.5 - 5+rnorm(100,0,3)
> model <- nls(y ~ xa - b, start = list(a = 2, b=2))
> summary(model)
```

Formula:  $y \sim x^a - b$

Parameters:

	Estimate	Std. Error	t value	Pr(> t )
a	2.4281	0.0722	33.6	<2e-16 ***
b	5.1934	0.3729	13.9	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

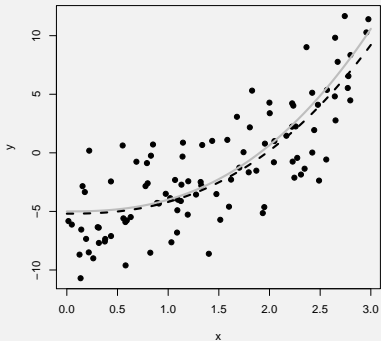
Residual standard error: 2.99 on 98 degrees of freedom

Number of iterations to convergence: 4

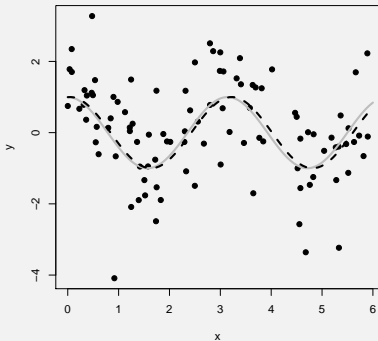
Achieved convergence tolerance: 3.42e-07

## Bardziej skomplikowane zależności

$$y = x^{2.5} - 5 + \epsilon$$



$$y = \sin(\pi/2 + 2.5x) + \epsilon$$



Regresja liniowa służyła nam do opisywania zależności pomiędzy zmienną ilościową a inną zmienną ilościową, lub zbiorem zmiennych ilościowych (regresja wieloraka).

W sytuacji gdy zmienną objaśnianą jest zmienna binarna to regresji liniowej użyć nie możemy.

Zmienne binarne są bardzo częste w rzeczywistych analizach. Przykładowo gdy interesuje nas odpowiedź myszy na podanie pewnej ilości inhibitora genu BRCA1. Jeżeli mysz może być tylko w dwóch stanach, np: żywa/zdechła, chora/zdrowa, normalna/zmieniona to jak stan myszy opisać ilością podanego inhibitora?

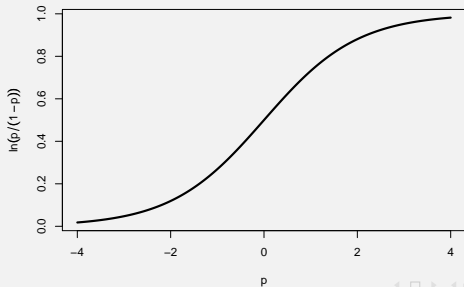
# Regresja logistyczna dla jednej zmiennej

W regresji logistycznej modeluje się prawdopodobieństwo wystąpienia określonego zjawiska za pomocą rozkładu dwumianowego

$$Y = \mathcal{B}(1, p)$$

gdzie prawdopodobieństwo  $p$  sukcesu określa się następująco

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$



# Regresja logistyczna

```
> z
[1] nie zyje nie zyje zyje zyje nie zyje zyje zyje zyje ....
Levels: zyje nie zyje
> x
[1] 4.405359 4.002627 2.417993 2.492061 4.244590 2.798724 ....

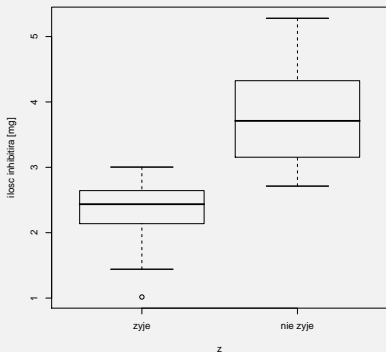
> modelRL <- glm(z~x, family="binomial")
> summary(modelRL)
Call:
glm(formula = z ~ x, family = "binomial")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.37241  -0.24212  -0.03202   0.01613   1.92766

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -21.604     11.221  -1.925   0.0542 .
y              7.342       3.872   1.896   0.0579 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 26.920 on 19 degrees of freedom
```

# Kocowy model regresji logistycznej?

Na podstawie tych wyników, możemy podać jawną postać modelu.

$$Pr(z = 'nie zyje') = \frac{\exp(7.3 * x - 21.6)}{1 + \exp(7.3 * x - 21.6)}$$



Do tych chwili mówiliśmy o zależnością pomiędzy parą zmiennych, ale zmiennych objaśniających może być więcej.

Gdy wiemy, że na jakąś cechę wpływa wiele parametrów i te parametry możemy kontrolować, wtedy możemy zbudować model uwzględniający wszystkie parametry.

Przykładowo na obfitość plonów może wpływać nasłonecznienie, wilgotność, stopień nawożenia. Zamiast budować model dla każdej z tych zmiennych osobno, możemy zbudować jeden model całościowo uwzględniający wpływ tych zmiennych.

```
> modelPP <- lm(cena~powierzchnia+pokoi, data = mieszkania)
> summary(modelPP) wyświetlamy podsumowanie modelu liniowego
Call:
lm(formula = cena ~ powierzchnia + pokoi, data = mieszkania)
Residuals:
      Min       1Q   Median       3Q      Max
-39705.0 -9386.1 -863.5  9454.3 35097.5

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   82407.1     2569.9   32.066 <2e-16 ***
powierzchnia  2070.9       149.2   13.883 <2e-16 ***
pokoi          840.1       2765.1   -0.304  0.762
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 14110 on 197 degrees of freedom
Multiple R-Squared:  0.8937, Adjusted R-squared:  0.8926
```

# Co trzeba zapamiętać?

- Jakie właściwości mogą mają estymatory, jakimi kryteriami kierować się przy wyborze estymatora?
- Jak na ocenę korelacji wpływa niespełnienie jej założeń i co robić gdy zmienne nie mają rozkładu normalnego lub zależność nie jest zależnością liniową.
- Jakie rodzaje regresji poznaliśmy, od czego zależy które metody regresji będziemy stosować?