ALGORYTMICZNA I STATYSTYCZNA ANALIZA DANYCH

20/11/2014

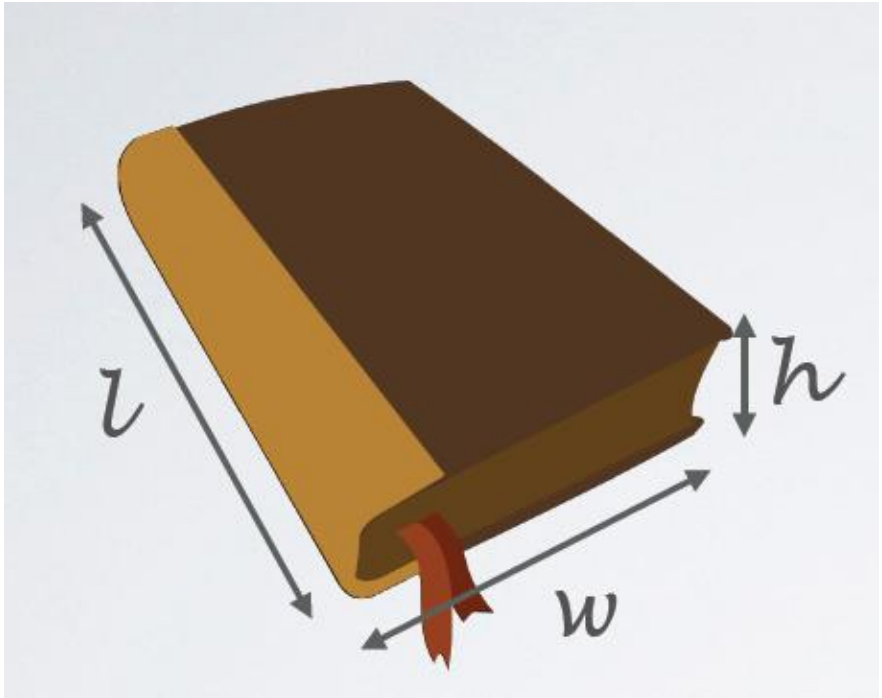
WFAiS UJ, Informatyka Stosowana
II stopień studiów

Regresja liniowa w wielu wymiarach

Model w oparciu o wiele zmiennych
Selekcja zmiennych
Przewidywanie w oparciu o model
Wnioskowanie w oparciu o model
Diagnostyka (ocena poprawności) modelu

Przykład

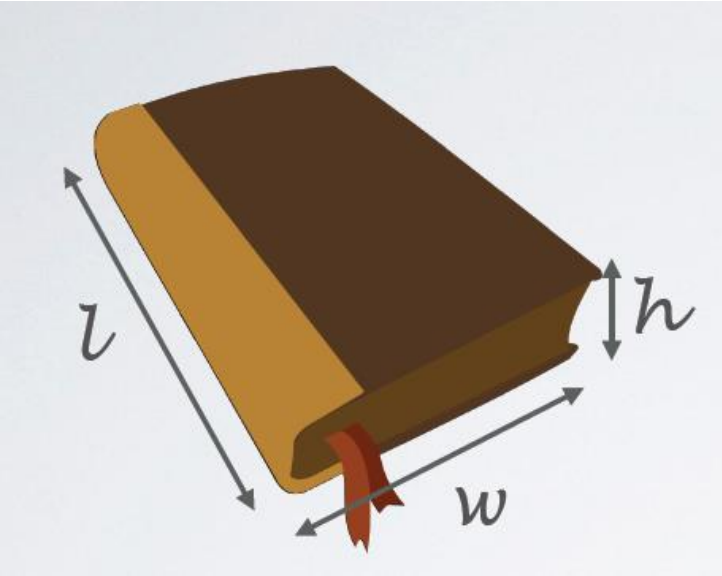
- Czy można przewidzieć relację pomiędzy wielkością książki i jej ciężarem w zależności jaka jest okładka (twarda czy miękka?)
- Na ogół książki w papierowych okładkach są lżejsze

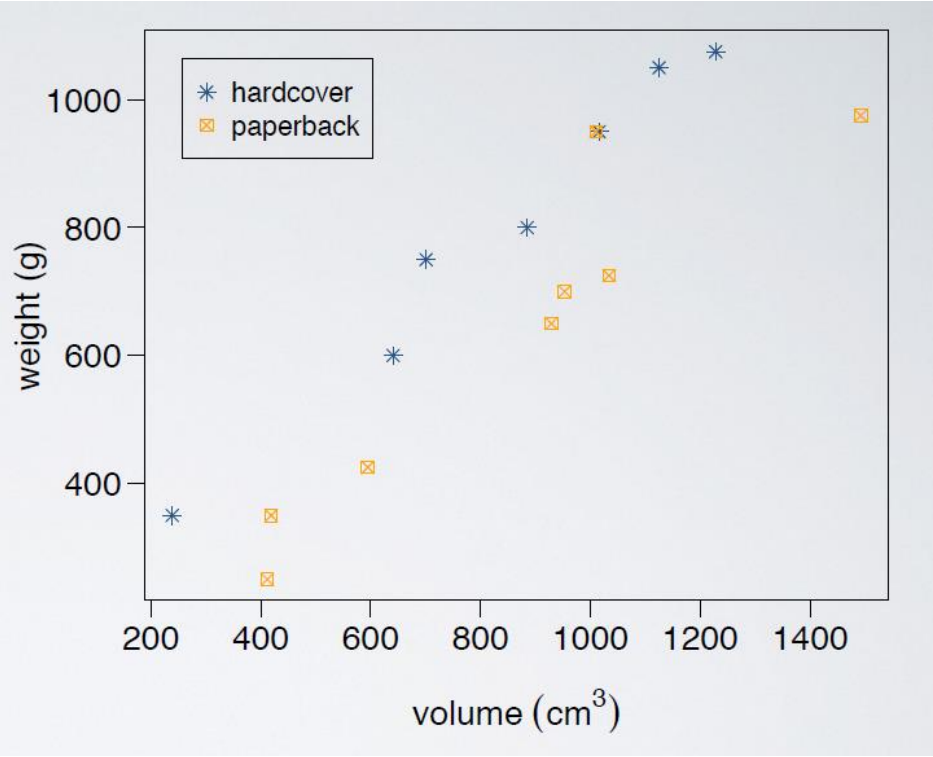


	weight (g)	volume (cm	cover
1	800	885	hb
2	950	1016	hb
3	1050	1125	hb
4	350	239	hb
5	750	701	hb
6	600	641	hb
7	1075	1228	hb
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

Przykład

- Czy można przewidzieć relację pomiędzy wielkością książki i jej ciężarem w zależności jaka jest okładka (twarda czy miękka) ?
- Na ogół książki w papierowych okładkach są lżejsze.





Przykład

R

```
# load data
> library(DAAG)
> data(allbacks)

# fit model
> book_mlr = lm(weight ~ volume + cover, data = allbacks)
> summary(book_mlr)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***

Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared:  0.9275, Adjusted R-squared:  0.9154
F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

Przykład

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

- Dla twardej okładki: cover = 0

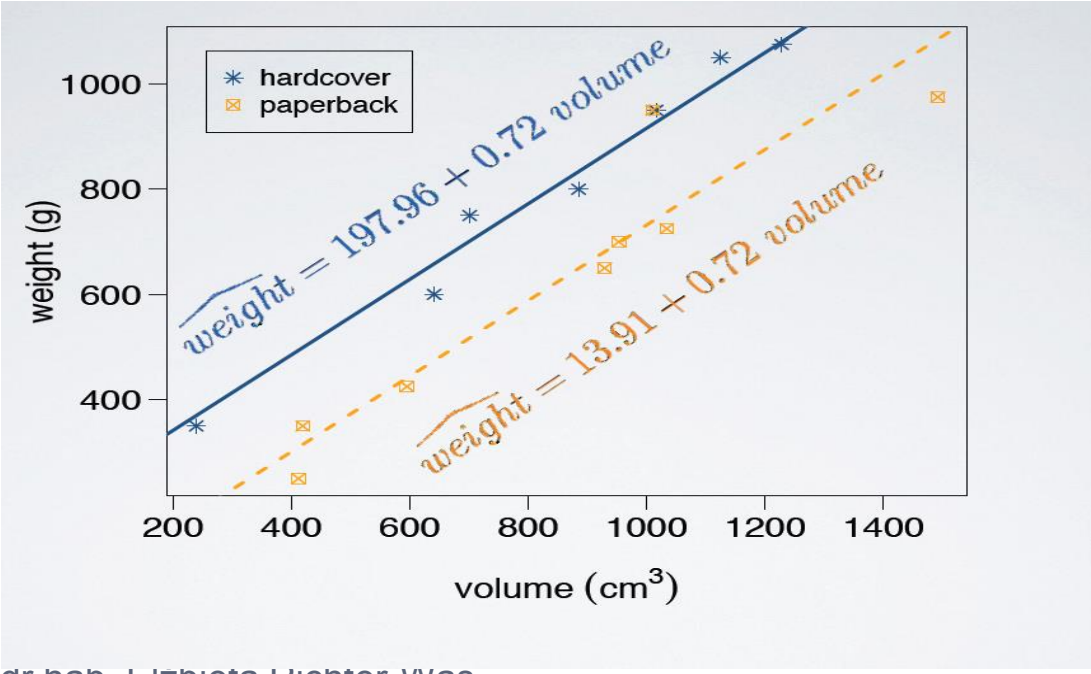
$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 0$$
$$= 197.96 + 0.72\ volume$$

- Dla miękkiej okładki

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 1$$
$$= 13.91 + 0.72\ volume$$

Przykład

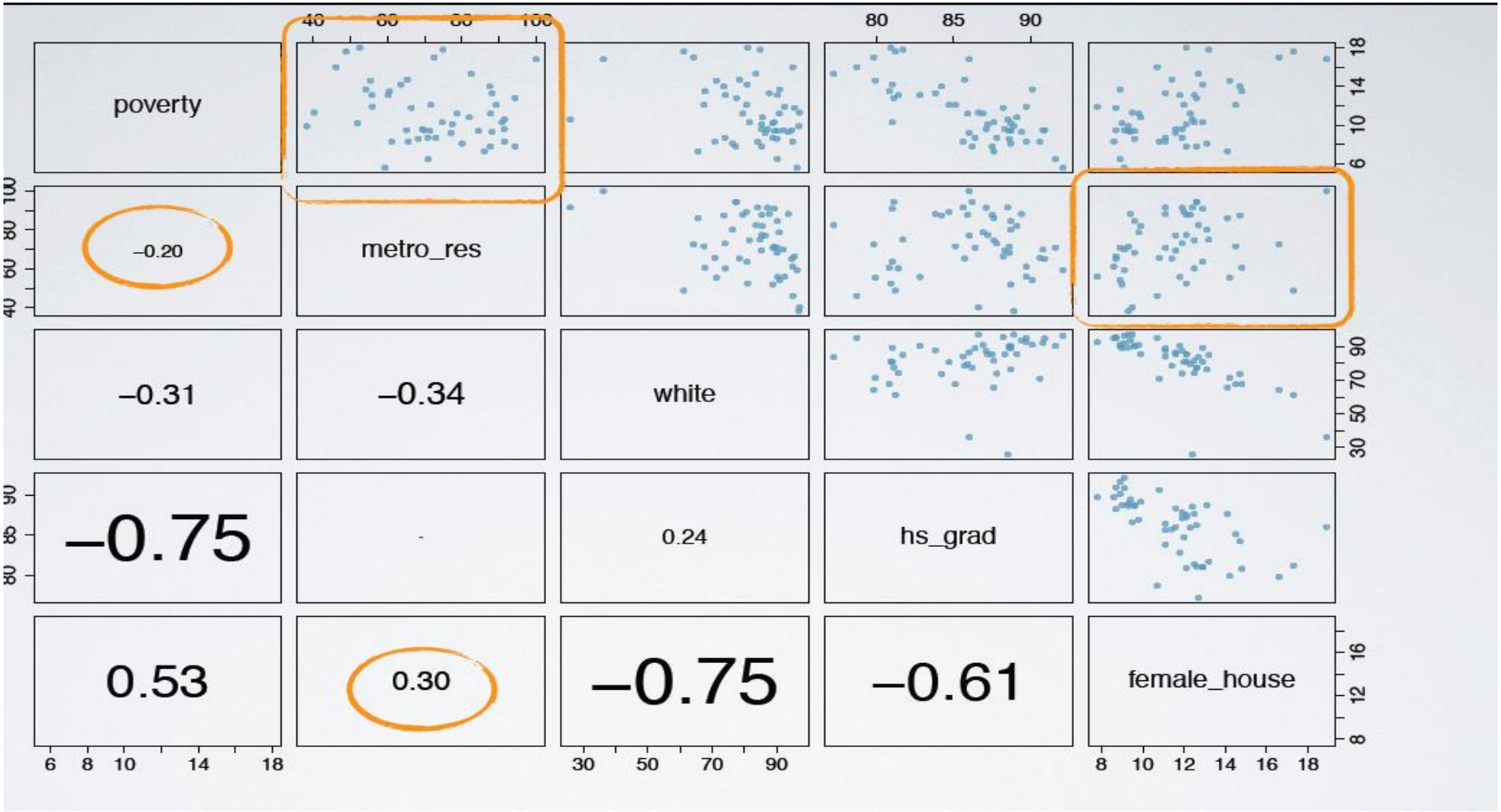
- **Interpretacja parametrów:** współczynniki dla zmiennej „cover” i „volume”
  - dla zwiększenia objętości o 1cm2 waga wzrośnie o 0.72g;
  - średnio w okładce papierowej książka lżejsza o 184.5g
- **Przewidywania w oparciu o model:** ile będzie ważyć książka o objętości 600cm^2



Przykład

- Model zakłada ten sam współczynnik dla zmiennej „volume” niezależnie od rodzaju okładki
- Jeżeli założenie nie jest realistyczne należy wprowadzić dodatkowy parametr.

Scatter & Correlation plot



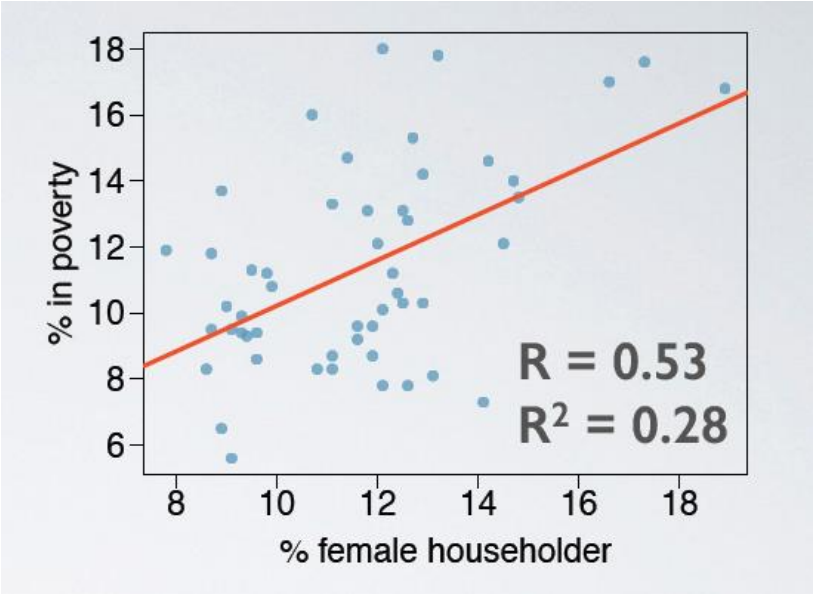
Przykład

R

```
# load data
> states = read.csv("http://bit.ly/dasi_states")

# fit model
> pov_slr = lm(poverty ~ female_house, data = states)
> summary(pov_slr)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.3094     1.8970   1.745   0.0873 .
female_house   0.6911     0.1599   4.322 7.53e-05 ***

Residual standard error: 2.664 on 49 degrees of freedom
Multiple R-squared:  0.276,  Adjusted R-squared:  0.2613
F-statistic: 18.68 on 1 and 49 DF,  p-value: 7.534e-05
```

Przewidywania w oparciu o model

- Przewidywanie % ubóstwa z względu na % rodzin utrzymywanych przez kobiety



Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00

Analiza wariancji

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57}{480.25} = 0.28$$

Dodajemy jest %białych

R

```
> pov_mlr = lm(poverty ~ female_house + white, data = states)
> summary(pov_mlr)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

R

```
> anova(pov_mlr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2 = \frac{132.57 + 8.21}{480.25} = 0.29$$

adjusted R²: $R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1} \right)$ k : number of predictors

Skalowane R^2

adjusted R^2: $R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1} \right)$ k : number of predictors

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1} \right)$$

$$= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) = 0.26$$

$n = 51$

R2 vs skalowane R^2

- Jeżeli dodajemy nową zmienna to R^2 wzrasta
- Ale jeżeli zmienna nie wprowadza nowej informacji to skalowane R^2 się nie zmienia

	R	adjusted R
Model 1 (poverty vs. female_house)	0.28	0.26
Model 2 (poverty vs. female_house + white)	0.29	0.26

Własności skalowanego R^2

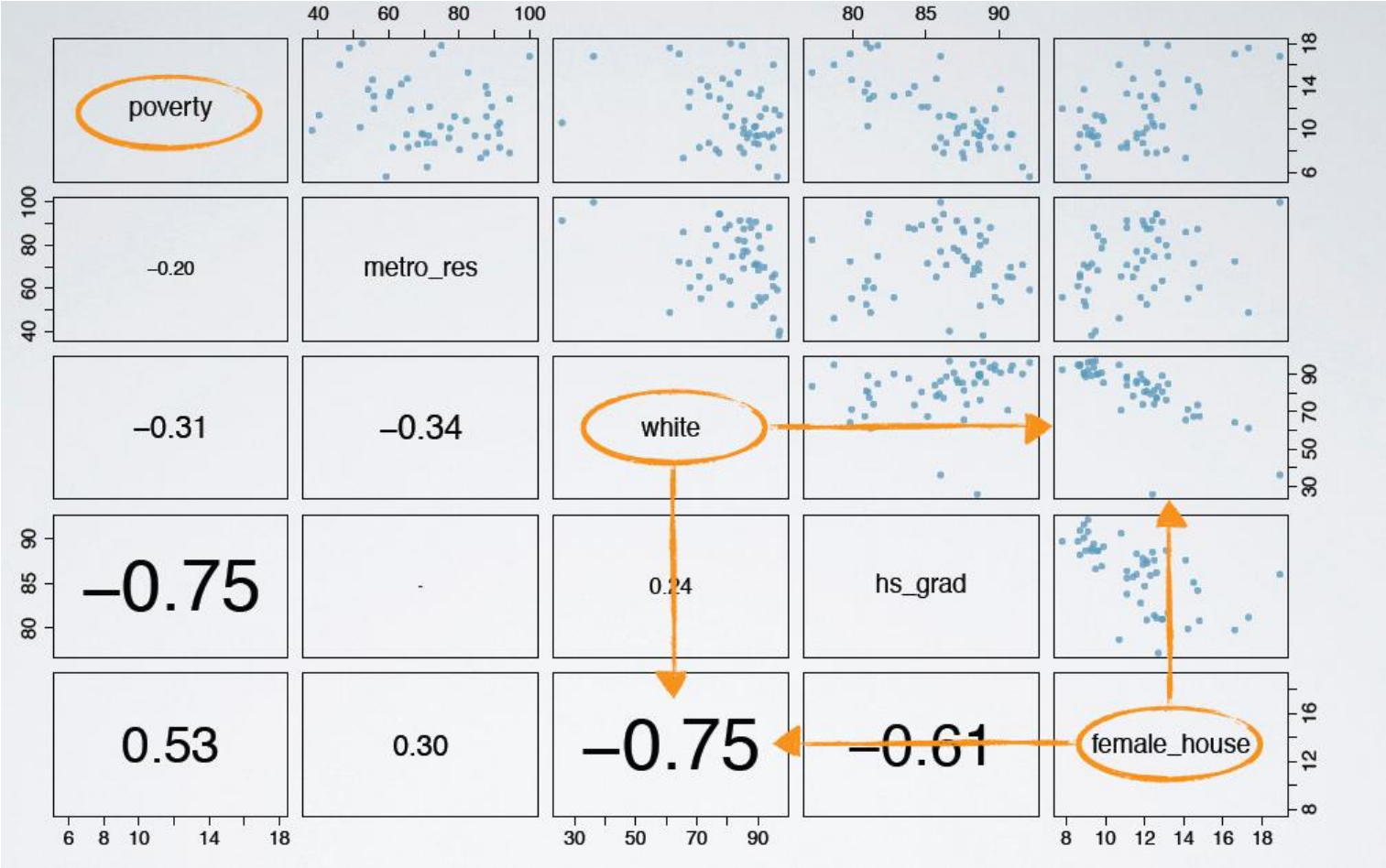
$$R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1} \right)$$

- k jest zawsze dodatnie: skalowane $R^2 < R$
- Skalowane R^2 uwzględnia czynnik „karzący” za dodawanie nowego parametru
- Wybieramy zawsze model o największej wartości skalowanego R^2

Korelacja i minimalny model

- Dwie zmienne x1, x2 nazywamy kolinearnymi jeżeli są ze sobą skorelowane
  - Zasada jest że powinniśmy tego unikać
  - Jeżeli jest taka sytuacja, to wymaga specjalnego traktowania przy budowaniu modelu
- Staramy się budować model z użyciem minimalnej ilości zmiennych x (predyktorów)
  - Wprowadzenie kolinearnych predyktorów może prowadzić do biasowania rozwiazania

Zmienne kolinearne



Testy poznawcze wśród dzieci (cognitive tests scores)

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
…	…	…	…	…	…
6	98	no	107.90	no	18
…	…	…	…	…	…
434	70	yes	91.25	yes	25

Pełny model

R

```
# load data
> cognitive = read.csv("http://bit.ly/dasi_cognitive")

# full model
> cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive)
> summary(cog_full)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341 *
mom_hs:yes   5.09482    2.31450   2.201   0.0282 *
mom_iq       0.56147    0.06064   9.259   <2e-16 ***
mom_work:yes 2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
Multiple R-squared:  0.2171, Adjusted R-squared:  0.2098
F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```

Wnioskowanie na podstawie modelu

$H_0: \beta_1 = \beta_1 = \ldots = \beta_k = 0$

H_A: At least one β_i is different than 0

```
F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```

- Ponieważ p-wartość < 0.05 model jest znaczący
  - F-statystyka: jeżeli wynik znaczący to wskazuje że choć jeden parameter β jest niezerowy
  - F-statystyka: jeżeli wynik nie znaczący to wskazuje to niekoniecznie znaczy że zmienne x nie są dobrymi predykatorami dla y, to tylko znaczy że model nie jest dobry

Testowanie hipotezy

- Czy wykształcenie matki ma wpływ na wynik testów dzieci? Tak jest znaczącą zmienną

$$H_0: \beta_1 = 0,$$
$$H_A: \beta_1 \neq 0$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hs:yes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_work:yes	2.53718	2.35067	1.079	0.2810
mom_age	0.21802	0.33074	0.659	0.5101

t-zmienna i p-wartość dla współczynnika przy mom_hs

t-statistic for the slope:

$$T = \frac{b_1 - 0}{SE_{b_1}} \qquad df = n - k - 1$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hs:yes   5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_work:yes 2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
```

$$T = \frac{5.095 - 0}{2.315} = 2.201$$

$$df = n - k - 1 = 434 - 4 - 1 = 429$$

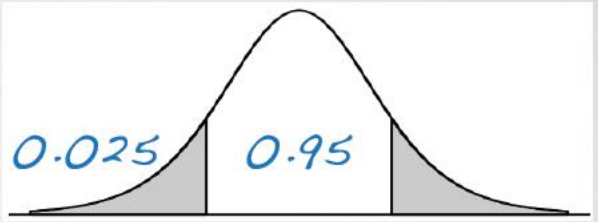
R

```
> pt(2.201,df = 429, lower.tail = FALSE) * 2
[1] 0.0282
```

95% przedział ufności dla współczynnika przy zmiennej = mom_work

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hs:yes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_work:yes	2.53718	2.35067	1.079	0.2810
mom_age	0.21802	0.33074	0.659	0.5101

Residual standard error: 18.14 on 429 degrees of freedom



R

```
> qt(0.025, df = 429)
[1] -1.97
```

$df = 434 - 4 - 1 = 429$

$t^*_{429} = 1.97$

$2.54 \pm 1.97 \times 2.35 \approx (-2.09, 7.17)$

We are 95% confident that, all else being equal, the model predicts that children whose moms worked during the first three years of their lives score 2.09 points lower to 7.17 points higher than those whose moms did not work.

Budowanie modelu

- **Backward selekcja**: rozpocznij z pełnego modelu (uwzględniając wszystkie zmienne), redukuj po jednej zmiennej do osiągnięcia minimalnego modelu
- **Forward selekcja**: rozpocznij od pustego modelu, dodawaj po jednej zmiennej do osiągnięcia minimalnego modelu
- **Kryteria:**
  - p-wartość, skalowane R^2
  - AIC, BIC, DIC, Bayes factor, itd..

Backward eliminacja: skalowane R^2

- Wystartuj z pełnego modelu
- Opuszczaj jedną zmienną na raz, oblicz skalowane R^2
- Wybierz model przy którym największe zwiększenie skalowanego R^2
- Powtarzaj dopóki skalowane R^2 przestaje rosnąć

Backward eliminacja: skalowane R^2

step	variables included	removed	adjusted R
FULL	kid_score ~ mom_hs + mom_iq + mom_work + mom_age		0.2098
STEP 1	kid_score ~ mom_iq + mom_work + mom_age	[-mom_hs]	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	[-mom_iq]	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	[-mom_work]	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	[-mom_age]	0.2109
STEP 2	kid_score ~ mom_iq + mom_work	[-mom_hs]	0.2024
	kid_score ~ mom_hs + mom_work	[-mom_iq]	0.0546
	kid_score ~ mom_hs + mom_iq	[-mom_work]	0.2105

Backward eliminacja: p-wartość

- Wystartuj z pełnego modelu
- Opuszczaj zmienną o najwyższej p-wartości
- Powtarzaj dopóki wszystkie zmienne są znaczące

Backward eliminacja: p-wartość

FULL	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.5924	9.2191	2.13	0.0341
mom_hs:yes	5.0948	2.3145	2.20	0.0282
mom_iq	0.5615	0.0606	9.26	0.0000
mom_work:yes	2.5372	2.3507	1.08	0.2810
mom_age	0.2180	0.3307	0.66	0.5101

STEP 1	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.1794	6.0432	4.00	0.0001
mom_hs:yes	5.3823	2.2716	2.37	0.0183
mom_iq	0.5628	0.0606	9.29	0.0000
mom_work:yes	2.5664	2.3487	1.09	0.2751

STEP 2	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.7315	5.8752	4.38	0.0000
mom_hsyes	5.9501	2.2118	2.69	0.0074
mom_iq	0.5639	0.0606	9.31	0.0000

Skalowane R^2 vs p-wartość

- p-wartość: znaczący parametr
  - Zależy od założonego poziomu ufności np. 5%
  - Różne modele dla różnych poziomów ufności
  - Chętnie używane ponieważ na ogół wymaga mniej iteracji
- skalowane R^2: bardziej wiarygodny parametr

Forward selekcja: skalowane R^2

- Wystartuj z jedna zmienną y dla każdego x
- Wybierz model o najwyższym R^2
- Dodawaj po jednej zmiennej, za każdym razem wybierając konfigurację z największym skalowanym R^2
- Powtarzaj dopóki skalowane R^2 rośnie

Forward selekcja: skalowane R^2

step	variables included	adjusted R
STEP 1	kid_score ~ mom_hs	0.0539
	kid_score ~ mom_work	0.0097
	kid_score ~ mom_age	0.0062
	kid_score ~ mom_iq	0.1991
STEP 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_iq + mom_age	0.1999
	kid_score ~ mom_iq + mom_hs	0.2105
STEP 3	kid_score ~ mom_iq + mom_hs + mom_age	0.2095
	kid_score ~ mom_iq + mom_hs + mom_work	0.2109
STEP 4	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	0.2098

Forward selekcja: p-wartość

- Wystartuj z jedna zmienna y dla każdego x
- Wybierz model o najmniejszej p-wartości
- Dodawaj po jednej zmiennej, za każdym razem wybierając konfigurację z najmniejszej p-wartości
- Powtarzaj dopóki p-wartość nie będzie się już zmniejszać

Końcowy model

R

```
> cog_final = lm(kid_score ~ mom_hs + mom_iq + mom_work, data = cognitive)
> summary(cog_final)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.17944    6.04319   4.001 7.42e-05 ***
mom_hsyes    5.38225    2.27156   2.369   0.0183 *
mom_iq       0.56278    0.06057   9.291  < 2e-16 ***
mom_workyes  2.56640    2.34871   1.093   0.2751

Residual standard error: 18.13 on 430 degrees of freedom
Multiple R-squared:  0.2163, Adjusted R-squared:  0.2109
F-statistic: 39.57 on 3 and 430 DF,  p-value: < 2.2e-16
```

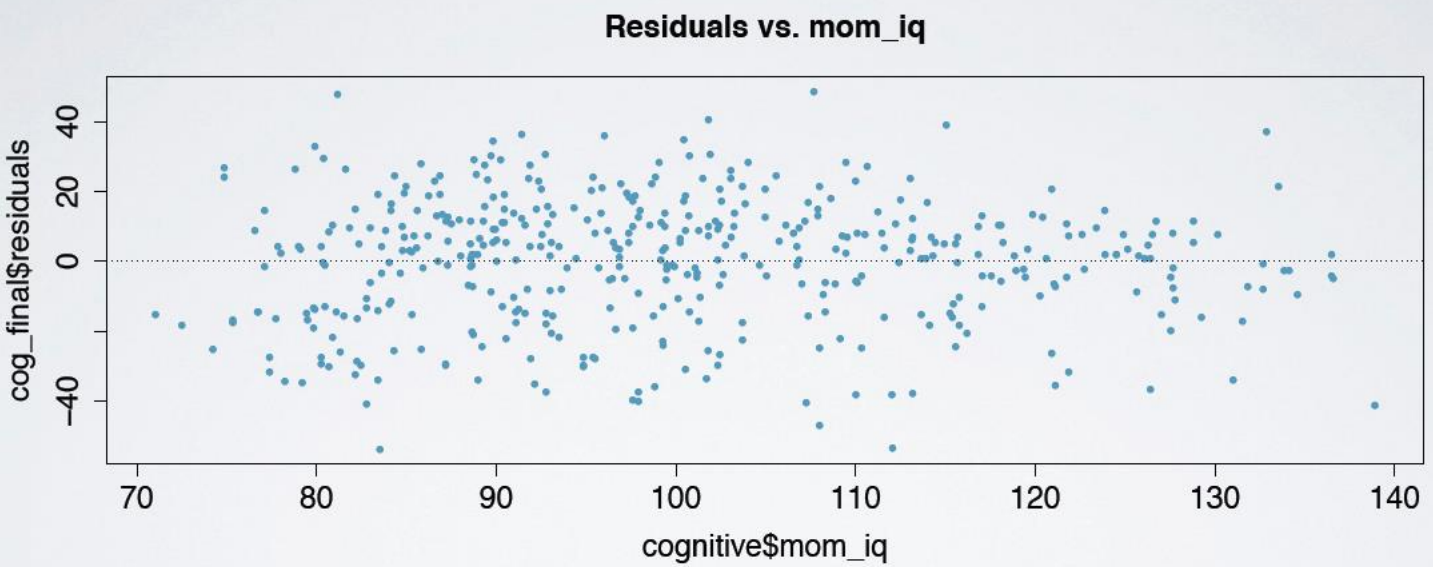
Diagnostyka modelu

- Liniowy związek pomiędzy x i y
- Prawie normalny rozkład residuals wokół zera
- Stała zmienność dla residuals
- Niezależność od okresu zbierania danych dla residuals

Liniowy związek pomiędzy x i y

R

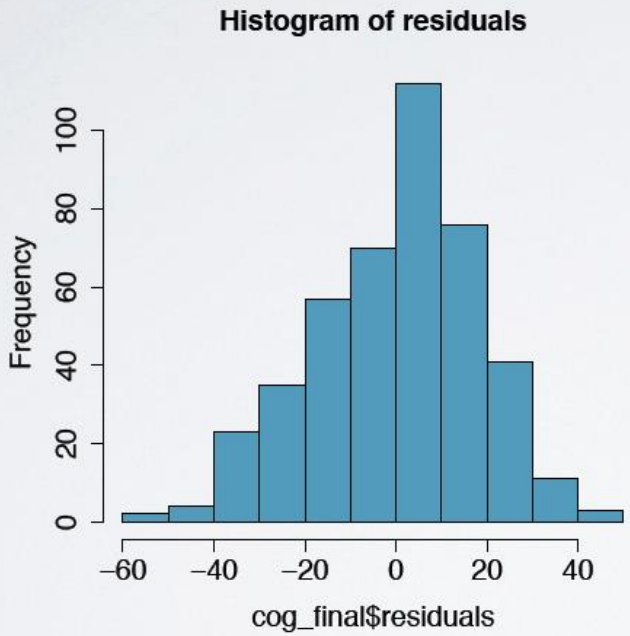
```
> cog_final = lm(kid_score ~ mom_hs + mom_iq + mom_work, data = cognitive)
> plot(cog_final$residuals ~ cognitive$mom_iq)
```

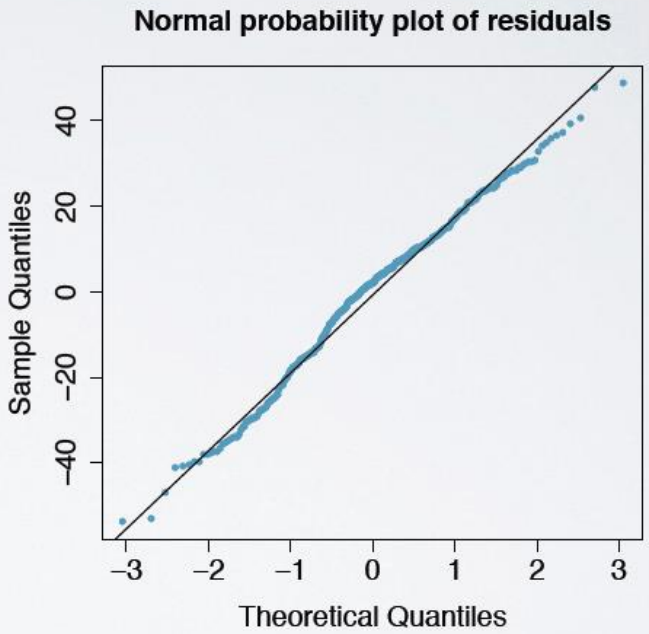


Prawie normalny rozkład residuals

R

```
> hist(cog_final$residuals)
> qqnorm(cog_final$residuals)
> qqline(cog_final$residuals)
```

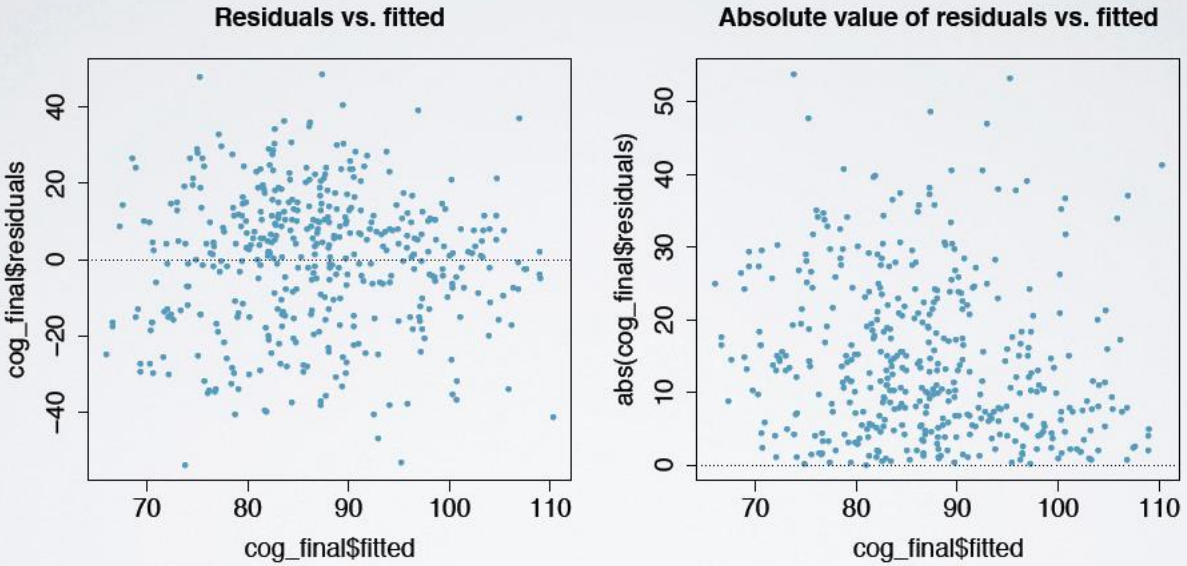




Stały rozrzut residuals

R

```
> plot(cog_final$residuals ~ cog_final$fitted)
> plot(abs(cog_final$residuals) ~ cog_final$fitted)
```



Niezależność residuals

R

```
> plot(cog_final$residuals)
```