Analiza Danych - Podstawy Statystyczne

Analiza wariancji i metody regresji

Marek Rupniewski

25 maja 2019

Analiza wariancji (ANOVA)

Sformułowanie problemu

Mamy do dyspozycji I prób (każda z pewnego rozkładu)

$$
\begin{aligned}
&Y_{11}, \ldots, Y_{1J_1}, \\
&Y_{21}, \ldots, Y_{2J_2}, \\
&\ldots, \ldots, \ldots, \\
&Y_{I1}, \ldots, Y_{IJ_I}.
\end{aligned}
$$

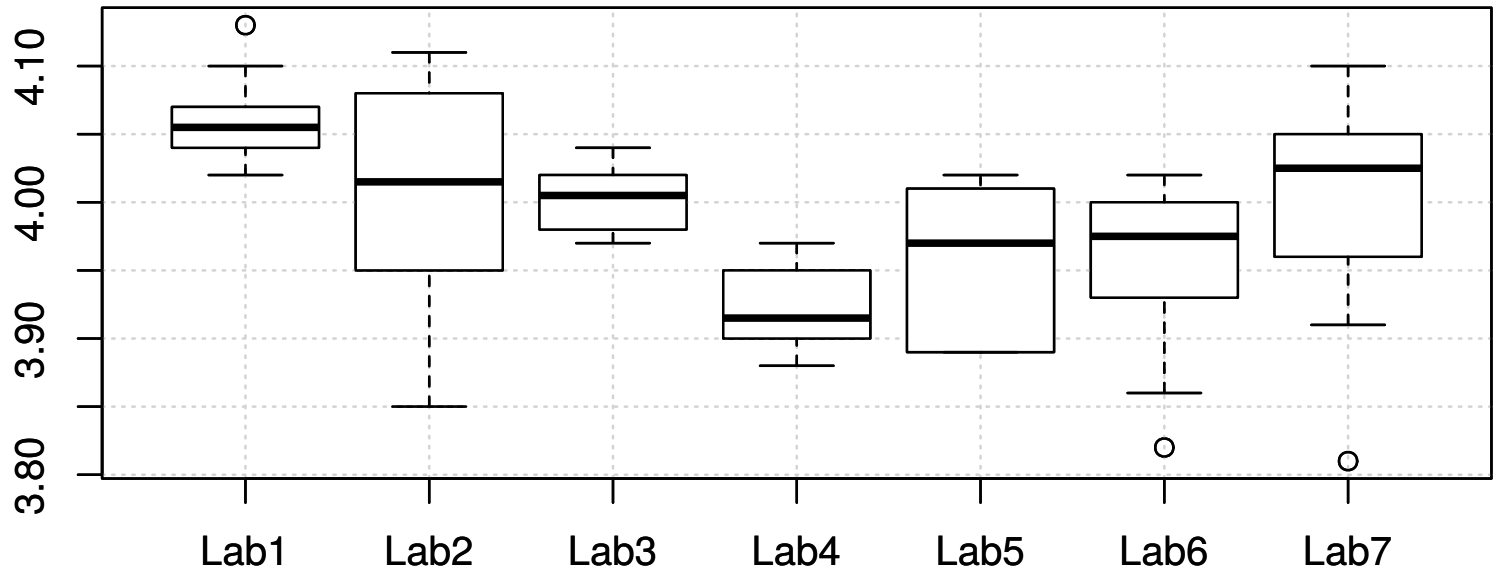
Chcemy sprawdzić, czy wartości średnie tych rozkładów są wszystkie równe.

Przykład (Kirchhoefer (1979): Semiautomated method for analysis...)

W 7 laboratoriach badano zawartość maleinianu chlorfeniraminy w pewnych tabletkach (nominalna zawartość: 4 mg).

Lab1	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7
4.13	3.86	4.00	3.88	4.02	4.02	4.00
4.07	3.85	4.02	3.88	3.95	3.86	4.02
4.04	4.08	4.01	3.91	4.02	3.96	4.03
4.07	4.11	4.01	3.95	3.89	3.97	4.04
4.05	4.08	4.04	3.92	3.91	4.00	4.10
4.04	4.01	3.99	3.97	4.01	3.82	3.81
4.02	4.02	4.03	3.92	3.89	3.98	3.91
4.06	4.04	3.97	3.90	3.89	3.99	3.96
4.10	3.97	3.98	3.97	3.99	4.02	4.05
4.04	3.95	3.98	3.90	4.00	3.93	4.06

Przykład (c.d.)



Oznaczenia

Y_{ij} — j-ty elementy i-tej próby.

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

μ średni poziom, α_i — różnica specyficzna dla i-tej próby (metody, kuracji), przy czym

$$\sum_{i=1}^{I} \alpha_i = 0,$$

ϵ_{ij} ciąg niezależnych zmiennych losowych o rozkładzie $N(0, \sigma^2)$.

$$\mathbb{E} Y_{ij} = \mu + \alpha_i.$$

Przypadek równolicznych grup

Analiza wariancji opiera się na następującej zależności

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y})^2}_{\text{suma odchyleń, } S_T^2} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i\bigstar})^2}_{\text{suma odch. wewnątrzgrup. } S_W^2} + \underbrace{J\sum_{i=1}^{I}(\overline{Y}_{i\bigstar}-\overline{Y})^2}_{\text{suma odch. międzygrup. } S_M^2},$$

gdzie

$$\overline{Y}_{i\bigstar} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}, \quad \overline{Y} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij} = \overline{\overline{Y}_{i\bigstar}}.$$

$$\mathbb{E}S_W^2 = \sigma^2(J-1)I, \quad \mathbb{E}S_M^2 = \sigma^2(I-1) + J\sum_{i=1}^{I}\alpha_i^2,$$

Jeśli $\alpha_1 = \cdots = \alpha_I = 0$, to $\frac{1}{I(J-1)}\mathbb{E}S_W^2 = \frac{1}{I-1}\mathbb{E}S_M^2$.

Jeśli $\alpha_1 = \cdots = \alpha_I = 0$, to

$$\frac{1}{I(J-1)}\mathbb{E}S_W^2 = \frac{1}{I-1}\mathbb{E}S_M^2.$$

Innymi słowy jeśli $\alpha_1 = \cdots = \alpha_I = 0$, to należy spodziewać się, że

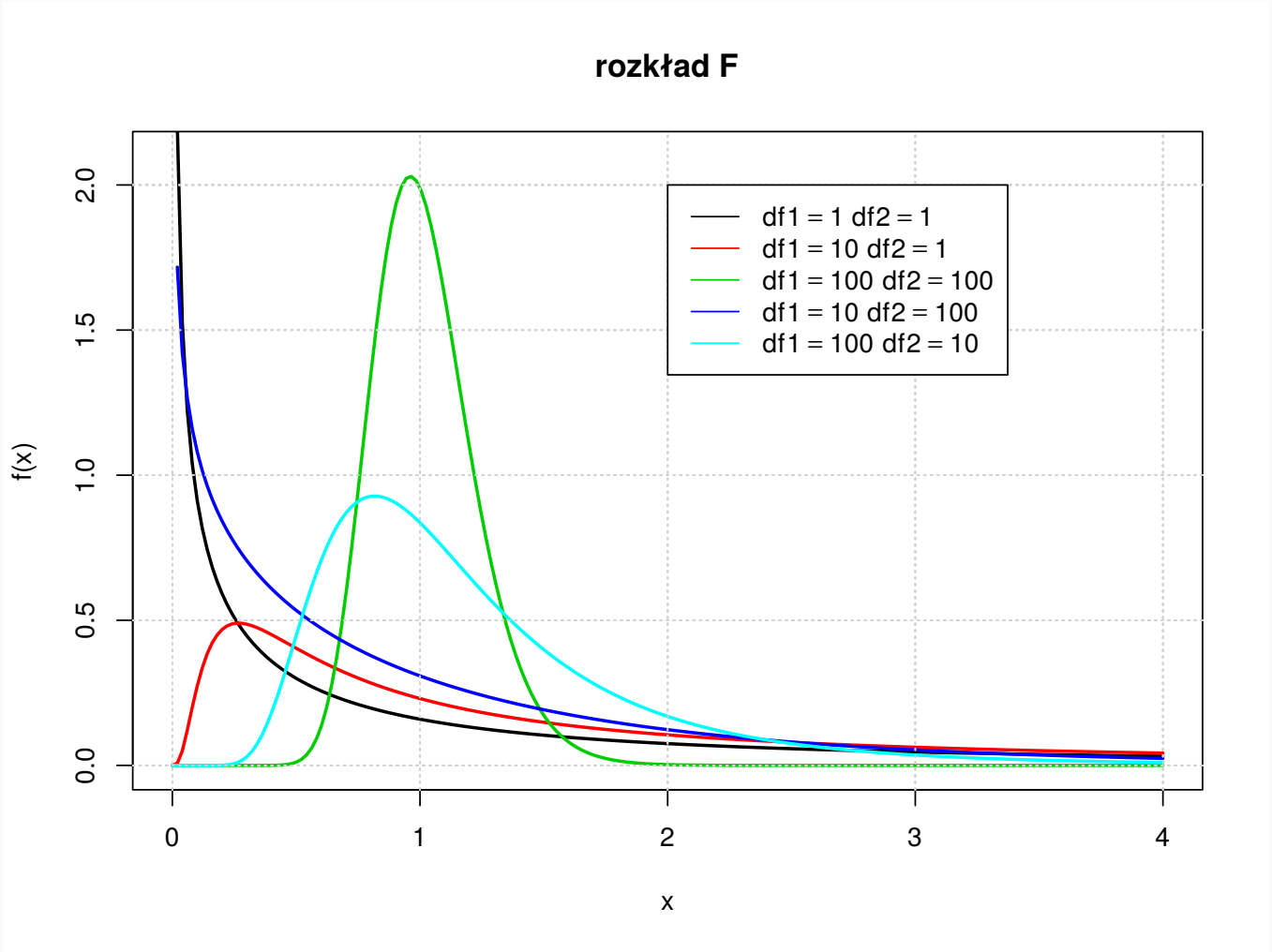
$$T = \frac{J\frac{1}{I-1}\sum_{i=1}^{I}(\overline{Y}_{i\bigstar} - \overline{Y})^2}{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{J-1}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{i\bigstar})^2}$$

przyjmuje wartości zbliżone do 1. Okazuje się, że przy przyjętym modelu zmienna T ma rozkład F Snedecora o $d_1 = I - 1$ oraz $d_2 = I(J-1)$ stopniach swobody.

$$f_{F_{d_1,d_2}}(x) = \frac{\sqrt{\frac{(d_1x)^{d_1}d_2^{d_2}}{(d_1x+d_2)^{d_1+d_2}}}}{xB(\frac{d_1}{2},\frac{d_2}{2})}, \quad x \geq 0, \qquad B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Rozkład F

$$\mathbb{E}F = \frac{d_2}{d_2 - 2}, \qquad \text{moda} : \frac{d_1 - 2}{d_1}\frac{d_2}{d_2 + 2}.$$



ANOVA — równoliczne grupy

1. Obliczamy globalną średnią $\overline{Y}$ oraz średnie wewnątrzgrupowe $\overline{Y}_{i\bigstar}$, $i = 1, \ldots, I$,
2. Wyznaczamy stosunek T wyestymowanych wariancji (międzygrupowej i wewnątrzgrupowych),
3. Jeśli $T > c$, to odrzucamy hipotezę o równości średnich

$$c = F^{-1}_{F_{d_1,d_2}}(1-\alpha).$$

Przypadek prób różnej wielkości

Było

$$T = \frac{J\frac{1}{I-1}\sum_{i=1}^{I}(\overline{Y}_{i\bigstar} - \overline{Y})^2}{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{J-1}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{i\bigstar})^2}$$

Teraz

$$T = \frac{\frac{1}{I-1}\sum_{i=1}^{I} J_i(\overline{Y}_{i\bigstar} - \overline{Y})^2}{\frac{1}{\sum_{i=1}^{I}(J_i-1)}\sum_{i=1}^{I}(J_i-1)\frac{1}{J_i-1}\sum_{j=1}^{J_i}(Y_{ij} - \overline{Y}_{i\bigstar})^2}$$

Jeśli zachodzi hipoteza zerowa, to T ma rozkład F z liczbami stopni swobody $d_1 = I - 1$ oraz $d_2 = \sum_{i=1}^{I} J_i - I$.

Przykład obliczeń w R

```
> tab
   Lab1 Lab2 Lab3 Lab4 Lab5 Lab6 Lab7
1  4.13 3.86 4.00 3.88 4.02 4.02 4.00
2  4.07 3.85 4.02 3.88 3.95 3.86 4.02
3  4.04 4.08 4.01 3.91 4.02 3.96 4.03
4  4.07 4.11 4.01 3.95 3.89 3.97 4.04
5  4.05 4.08 4.04 3.92 3.91 4.00 4.10
6  4.04 4.01 3.99 3.97 4.01 3.82 3.81
7  4.02 4.02 4.03 3.92 3.89 3.98 3.91
8  4.06 4.04 3.97 3.90 3.89 3.99 3.96
9  4.10 3.97 3.98 3.97 3.99 4.02 4.05
10 4.04 3.95 3.98 3.90 4.00 3.93 4.06
> stack(tab)->tablica;
> names(tablica)<-c('poziom','lab.')
> summary(aov(poziom~lab.,tablica))
             Df Sum Sq  Mean Sq F value   Pr(>F)
lab.          6 0.1247 0.020790    5.66 9.45e-05 ***
Residuals    63 0.2314 0.003673
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Analiza nieparametryczna

Mamy do dyspozycji I prób (każda z pewnego rozkładu, tym razem nie zakładamy normalności rozkładów)

$$X_{11}, \ldots, X_{1J_1},$$
$$X_{21}, \ldots, X_{2J_2},$$
$$\ldots, \ldots, \ldots,$$
$$X_{I1}, \ldots, X_{IJ_I}.$$

Chcemy sprawdzić (hipoteza zerowa), czy dane we wszystkich próbach są z tego samego rozkładu (zajmowaliśmy się już tym problemem w przypadku 2 grup! — test Manna-Whitneya).

Oznaczenia

$$N = \sum_{i=1}^{I} J_i.$$

R_{ij} ranga (pozycja) X_{ij} obserwacji w uporządkowanym rosnąco ciągu ($R_{ij} \in [1, 2, \ldots, N]$).

$$\overline{R} = \frac{1}{\sum_{i=1}^{I} J_i} \sum_{i=1}^{I} \sum_{j=1}^{J_i} R_{ij} = \frac{N+1}{2}, \quad \overline{R}_{i\bigstar} = \frac{1}{J_i} \sum_{j=1}^{J_i} R_{ij}.$$

Okazuje się, że statystyka

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{I} J_i (\overline{R}_{i\bigstar} - \overline{R})^2$$

ma rozkład zbliżony do χ^2_{I-1} (jeśli H_0 prawdziwa).

Test Kruskala-Wallisa (uogólnienie testu Manna-Whitneya)

- Szeregujemy wszystkie próbki w rosnącej kolejności
- Obliczamy $\overline{R}$, $\overline{R}_{i\bigstar}$
- Obliczamy statystykę K i odrzucamy hipotezę zerową jeśli $K > c$, gdzie c jest wartością krytyczną odpowiadającą zadanemu poziomowi istotności α (można wyznaczyć dokładną wartość symulacyjnie, lub przybliżyć z wykorzystaniem rozkładu χ^2).

Przykład obliczeń w R

```
> kruskal.test(poziom~lab.,tablica)

        Kruskal-Wallis rank sum test

data:  poziom by lab.
Kruskal-Wallis chi-squared = 29.606, df = 6, p-value = 4.67e-05
```

Istotne odstępstwa od średniej

Wiemy już, że w rozważanym przykładzie wyniki poszczególnych laboratoriów nie mają (raczej) takich samych średnich. Jak obiektywnie ocenić, które z tych średnich odstają od pozostałych?

Metoda Tukeya

Jeśli błędy ε_{ij} są niezależne z rozkładu $N(0, \sigma^2)$, a próby są równoliczne, to zmienne $\overline{Y}_{i\bigstar} - \alpha_i$ są niezależne i mają rozkład $N(\mu, \sigma^2/J)$. Ich wariancja może być estymowana jako

$$\frac{s^2}{J} = \frac{S_W^2}{IJ(J-1)}.$$

Zmienna losowa

$$\max_{i_1, i_2} \frac{|(\overline{Y}_{i_1\bigstar} - \alpha_{i_1}) - (\overline{Y}_{i_2\bigstar} - \alpha_{i_2})|}{s/\sqrt{J}}$$

ma rozkład nazywany studentyzowanym rozkładem rozstępu (ang. studentized range distribution) z parametrami I (liczba prób) oraz $I(J-1)$ (liczba stopni swobody "w" s^2) (patrz funkcje `ptukey` oraz `qtukey` pakietu R).

Metoda Tukeya, c.d.

Niech $q_{I,I(J-1)}$ oznacza funkcję kwantylową dla studentyzowanego rozkładu rozstępu.

$$\gamma = \mathbb{P}\left(\max_{i_1,i_2} |(\overline{Y}_{i_1\bigstar} - \alpha_{i_1}) - (\overline{Y}_{i_2\bigstar} - \alpha_{i_2})| \leq q_{I,I(J-1)}(\gamma) s/\sqrt{J}\right).$$

Zatem przedziałami ufności dla różnic $\alpha_{i_1} - \alpha_{i_2}$ na poziomie ufności γ (jednocześnie dla wszystkich różnic) są przedziały o końcach

$$\overline{Y}_{i_1\bigstar} - \overline{Y}_{i_2\bigstar} \pm q_{I,I(J-1)}(\gamma)\frac{s}{\sqrt{J}}.$$

Metoda Tukeya, przykład

i	1	2	3	4	5	6	7
$\overline{Y}_{i\bigstar}$	4.062	3.997	4.003	3.920	3.957	3.955	3.998

$$s = 0.06$$

Załóżmy poziom ufności $\gamma = 95\%$. Wówczas

$$q_{I,I(J-1)}(\gamma)\frac{s}{\sqrt{J}} = q_{7,63}(0.95)\frac{0.06}{\sqrt{10}} = 0.083.$$

Przy tym poziomie ufności przedziały dla różnic między średnimi z poszczególnych laboratoriów nie zawierają 0 tylko dla par laboratoriów: (1,4); (1,5); (1,6) oraz (3,4).

Metoda Bonferroniego

Mamy I prób. Gdybyśmy dla każdej różnicy między średnimi grupowymi konstruowali test, na poziomie istotności α, sprawdzający czy ta średnia jest zerowa, to łączny poziom istotności (prawdopodobieństwo tego, że przy prawdziwych hipotezach zerowych któryś test hipotezę zerową odrzuci) szacuje się jedynie przez

$$1 - (1-\alpha)^{\binom{I}{2}} \overset{\alpha \ll 1}{\approx} \binom{I}{2}\alpha = \alpha I(I-1)/2.$$

Pomysł Bonferroniego polega na tym, aby w takim przypadku testować każdą z $\binom{I}{2}$ hipotez na poziomie istotności

$$\alpha_B = \alpha / \binom{I}{2}.$$

Metoda Bonferroniego, przykład

Dla założonego łącznego poziomu istotności 0.05 oraz $I = 7$ mamy

$$\alpha_B = \frac{0.05}{21}.$$

Test na poziomie istotności α_B dla hipotezy zerowej $\alpha_{i_1} - \alpha_{i_2} = 0$ wobec alternatywy „$\neq 0$” ma obszar akceptacji dla statystyki $\overline{Y}_{i_1} - \overline{Y}_{i_2}$ postaci odcinka o końcach:

$$\pm F_{t_{63}}^{-1}\left(1 - \frac{\alpha_B}{2}\right)\frac{s}{\sqrt{5}} \approx 0.086.$$

Do takich obszarów akceptacji nie wpadają różnice między średnimi z prób dla par laboratoriów (1,4); (1,5); (1,6).

Metoda Tukeya daje węższe przedziały ufności, ale za to wymaga (w przeciwności do metody Bonferroniego) równoliczności poszczególnych prób.

Regresja liniowa

Metoda najmniejszych kwadratów

Mamy do dyspozycji n par liczb:

$$(x_1, y_1), \ldots, (x_n, y_n).$$

Na podstawie tych par chcemy wyznaczyć zależność

$$y = f(x),$$

która w możliwie najlepszy sposób opisuje nasz zestaw danych. Dalej będziemy rozważać przede wszystkim przypadek liniowej zależności:

$$f(x) = \beta_0 + \beta_1 x.$$

Metoda najmniejszych kwadratów

W metodzie najmniejszych kwadratów minimalizujemy błąd średniokwadratowy, tzn. minimalizujemy ze względu na parametry β_0 oraz β_1 wyrażenie

$$L = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

Równania $0 = \frac{\partial L}{\partial \beta_0}$, $0 = \frac{\partial L}{\partial \beta_1}$ sprowadzają się do:

$$\beta_0 + \beta_1 \overline{x} = \overline{y}, \quad \beta_0 \overline{x} + \beta_1 \overline{x^2} = \overline{xy}.$$

$$\beta_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}, \quad \beta_0 = \overline{y} - \beta_1 \overline{x}.$$

Regresja liniowa

Mamy do dyspozycji próbę:

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

gdzie X_i oraz Y_i są pewnymi zmiennymi losowymi, przy czym

$$\mathbb{E}(Y_i | X_i = x) = \beta_0 + \beta_1 x + \varepsilon_i,$$

gdzie $\varepsilon_1, \ldots, \varepsilon_n$ są niezależnymi zmiennymi losowymi o rozkładzie $N(0, \sigma^2)$.

Na podstawie posiadanej próby chcemy wyznaczyć:

$$\beta_0,\ \beta_1 \operatorname{oraz} \sigma^2.$$

Przy ustalonych wartościach $X_1, \ldots, X_n$ funkcja wiarygodności (dla obserwacji $Y_1, \ldots, Y_n$) ma postać:

$$\mathcal{L}(Y_1, \ldots, Y_n; \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2}.$$

Metoda największej wiarygodności daje:

$$\hat{\beta}_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2}, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Ponadto, przy ustalonych $X_1, \ldots, X_n$ mamy:

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma_1^2\right), \quad \hat{\beta}_0 \sim N\left(\beta_0, \sigma_1^2 \overline{X^2}\right), \quad \sigma_1^2 = \frac{\sigma^2}{n(\overline{X^2} - \overline{X}^2)}.$$

Mamy

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma_1^2\right), \quad \hat{\beta}_0 \sim N\left(\beta_0, \sigma_1^2 \overline{X^2}\right), \quad \sigma_1^2 = \frac{\sigma^2}{n(\overline{X^2} - \overline{X}^2)}.$$

Zmienne $\hat{\beta}_0$ oraz $\hat{\beta}_1$ są zależne!:

$$\mathbb{C}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma_1^2 \overline{X}.$$

Parametry estymatora $\hat{\sigma}^2$

Okazuje się, że zmienna losowa $\hat{\sigma}^2$ jest niezależna od $\hat{\beta}_0$ i $\hat{\beta}_1$. Ponadto

$$\mathbb{E}\hat{\sigma}^2 = \frac{n-2}{n}\sigma^2$$

oraz

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

W szczególności nieobciążonym estymatorem σ^2 jest

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Przedział ufności dla β_1

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{n(\overline{X^2} - \overline{X}^2)}\right).$$

Niech $s^2_{\beta_1} = \frac{s^2}{n(\overline{X^2} - \overline{X}^2)}$. Wówczas

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\beta_1}} \sim t_{n-2}$$

Jeśli $c = F^{-1}_{t_{n-2}}(1 - \frac{\alpha}{2})$, to przedziałem ufności dla $\hat{\beta}_1$ na poziomie ufności $1 - \alpha$ jest

$$\left[\hat{\beta}_1 - cs_{\beta_1}, \hat{\beta}_1 + cs_{\beta_1}\right].$$

Przedział ufności dla β_0

Podobnie jak dla β_1 uzyskuje się przedział ufności dla β_0 na poziomie ufności $1 - \alpha$:

$$\left[\hat{\beta}_0 - cs_{\beta_0}, \hat{\beta}_0 + cs_{\beta_0}\right],$$

gdzie

$$s^2_{\beta_0} = s^2 \left(\frac{\overline{X^2}}{n(\overline{X^2} - \overline{X}^2)} \right) = s^2_{\beta_1} \overline{X^2}$$

oraz

$$c = F^{-1}_{t_{n-2}}(1 - \frac{\alpha}{2}).$$

Przedział ufności dla σ^2

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

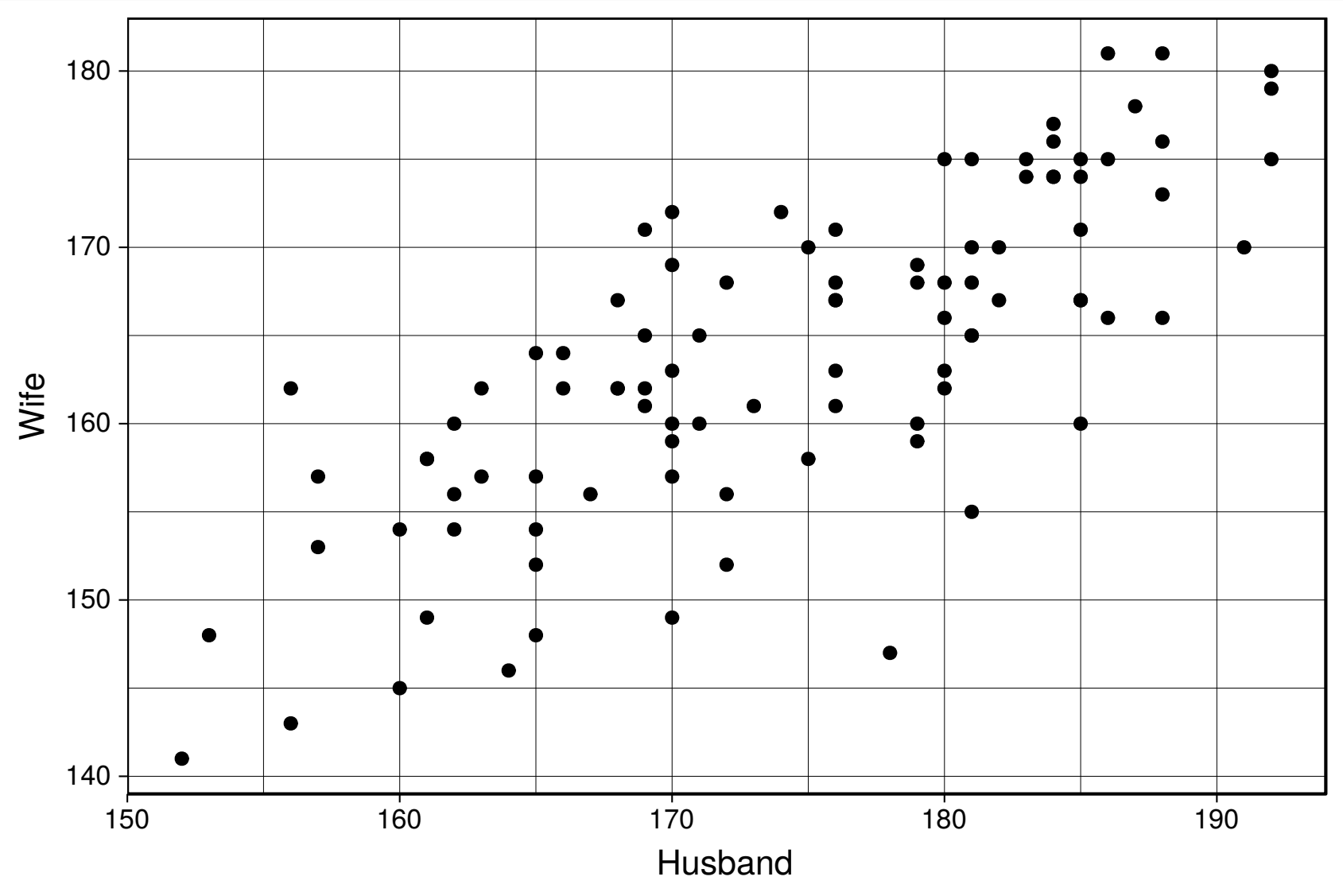
Zatem przedziałem ufności dla σ^2 na poziomie ufności $1-\alpha$ jest :

$$\left[\frac{n\hat{\sigma}^2}{c_2}, \frac{n\hat{\sigma}^2}{c_1}\right],$$

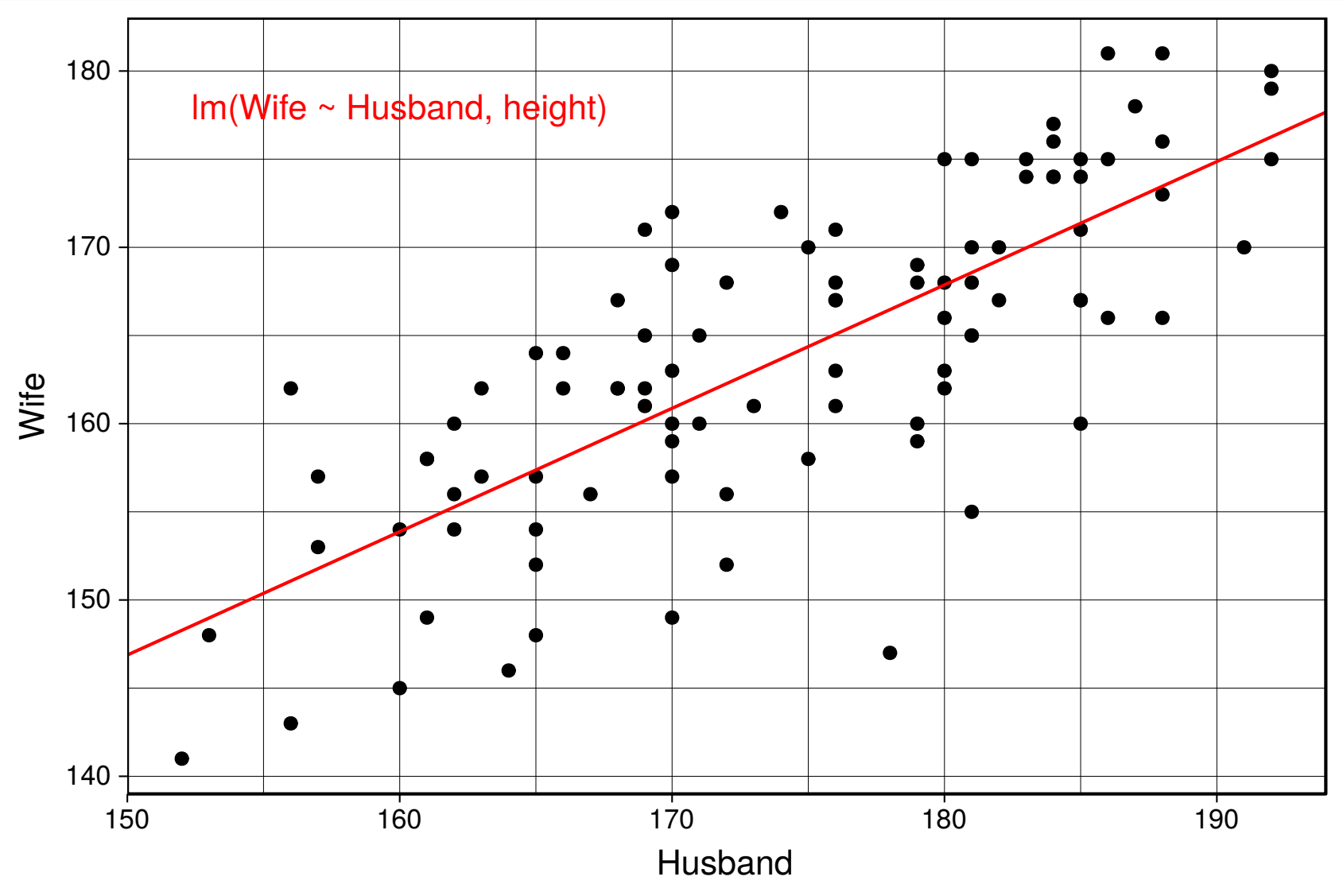
gdzie $F_{\chi^2_{n-2}}(c_2) - F_{\chi^2_{n-2}}(c_1) = 1-\alpha$. Np.

$$c_2 = F^{-1}_{\chi^2_{n-2}}(1-\frac{\alpha}{2}), \quad c_1 = F^{-1}_{\chi^2_{n-2}}(\frac{\alpha}{2}).$$

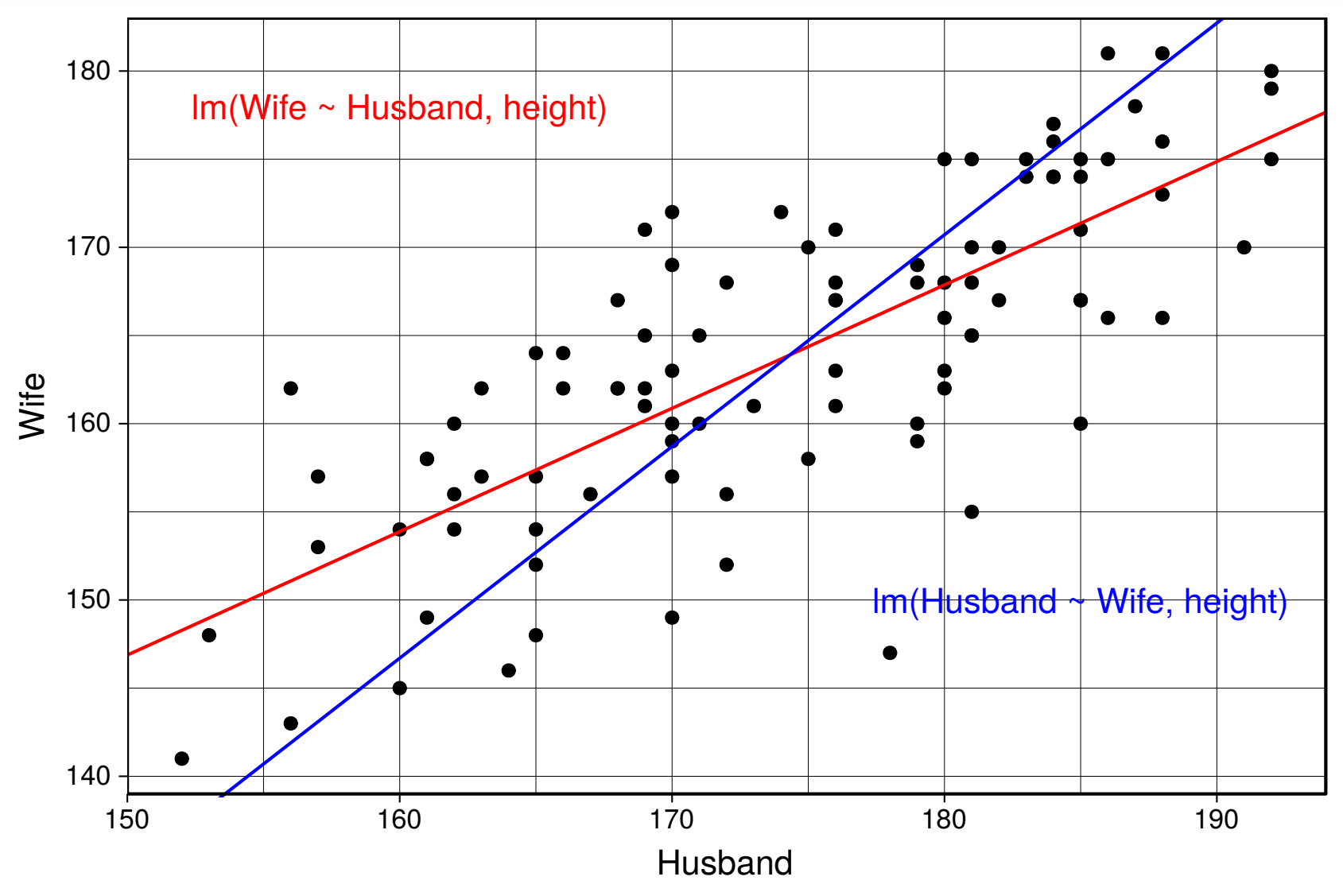
Przykład — wzrost w małżeństwie



Przykład — wzrost w małżeństwie



Przykład — wzrost w małżeństwie



Przykład — wzrost w małżeństwie

```
> summary(model1)

Call:
lm(formula = Wife ~ Husband, data = height)

Residuals:
     Min       1Q   Median       3Q      Max
-19.4685  -3.9208   0.8301   3.9538  11.1287

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.93015   10.66162   3.933 0.000161 ***
Husband      0.69965    0.06106  11.458  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.928 on 94 degrees of freedom
Multiple R-squared:  0.5828,    Adjusted R-squared:  0.5783
F-statistic: 131.3 on 1 and 94 DF,  p-value: < 2.2e-16
```

Współczynnik determinacji

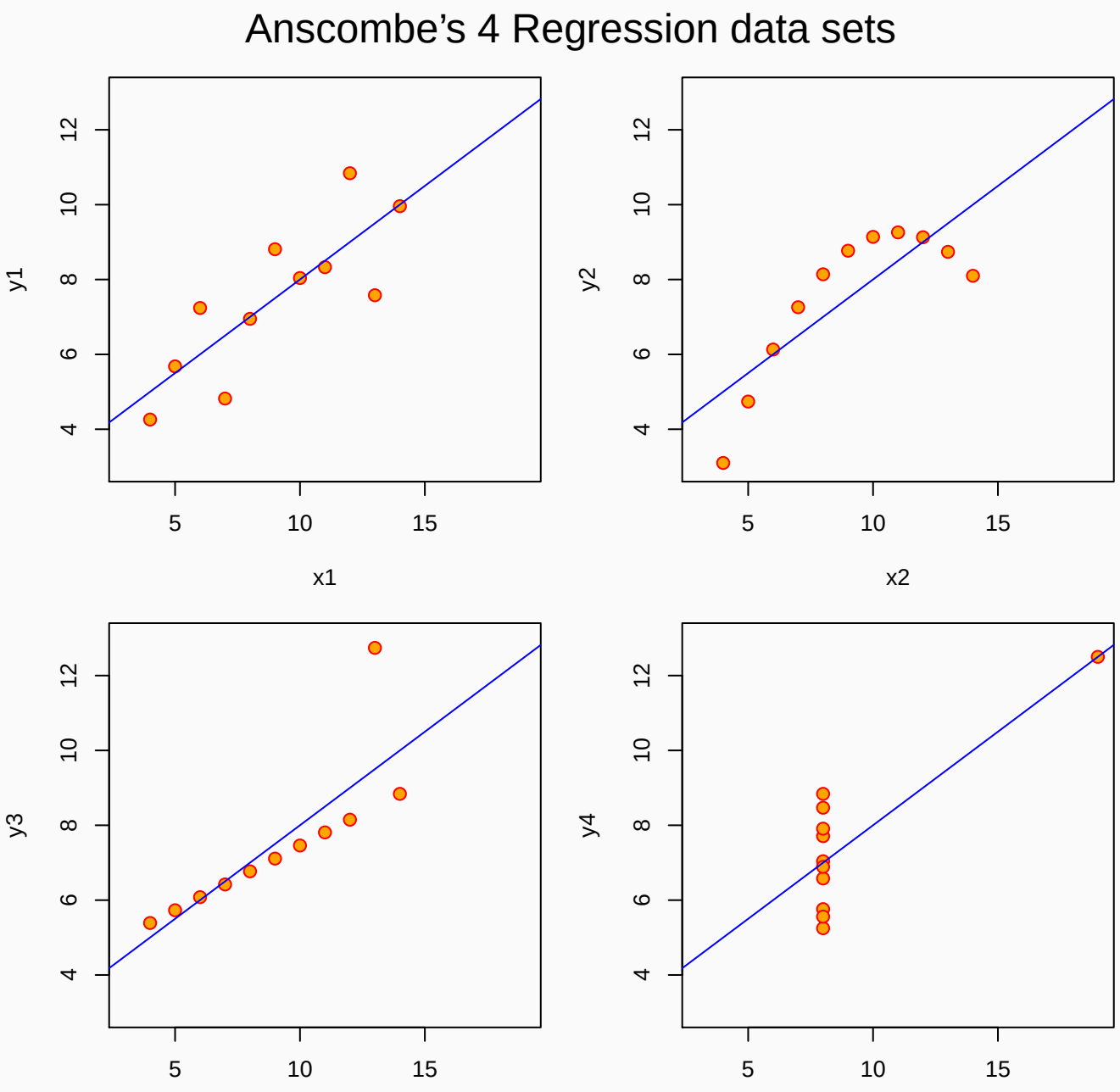
Współczynnik determinacji R^2 to jedna z podstawowych miar jakości dopasowania modelu do danych.

$$R^2 = 1 - \frac{s^2_{\hat{\varepsilon}\hat{\varepsilon}}}{s^2_{YY}} = \frac{s^2_{\hat{Y}\hat{Y}}}{s^2_{YY}} \in [0, 1].$$

Dla klasycznego modelu z jedną zmienną objaśniającą $R^2 = r^2_{XY}$ (współczynnik determinacji jest kwadratem estymatora współczynnika korelacji).

Dopasowanie modelu jest tym lepsze im większy jest współczynnik R^2, ale warto przyjrzeć się jeszcze innymi miarom dopasowania, a przede wszystkim zilustrować dopasowanie graficznie.

Graficzna weryfikacja modelu



Przykład Anscombe-a — przypadek 1

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,  Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

Przykład Anscombe-a — przypadek 2

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x2             0.500      0.118   4.239  0.00218 **
---

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6662,  Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

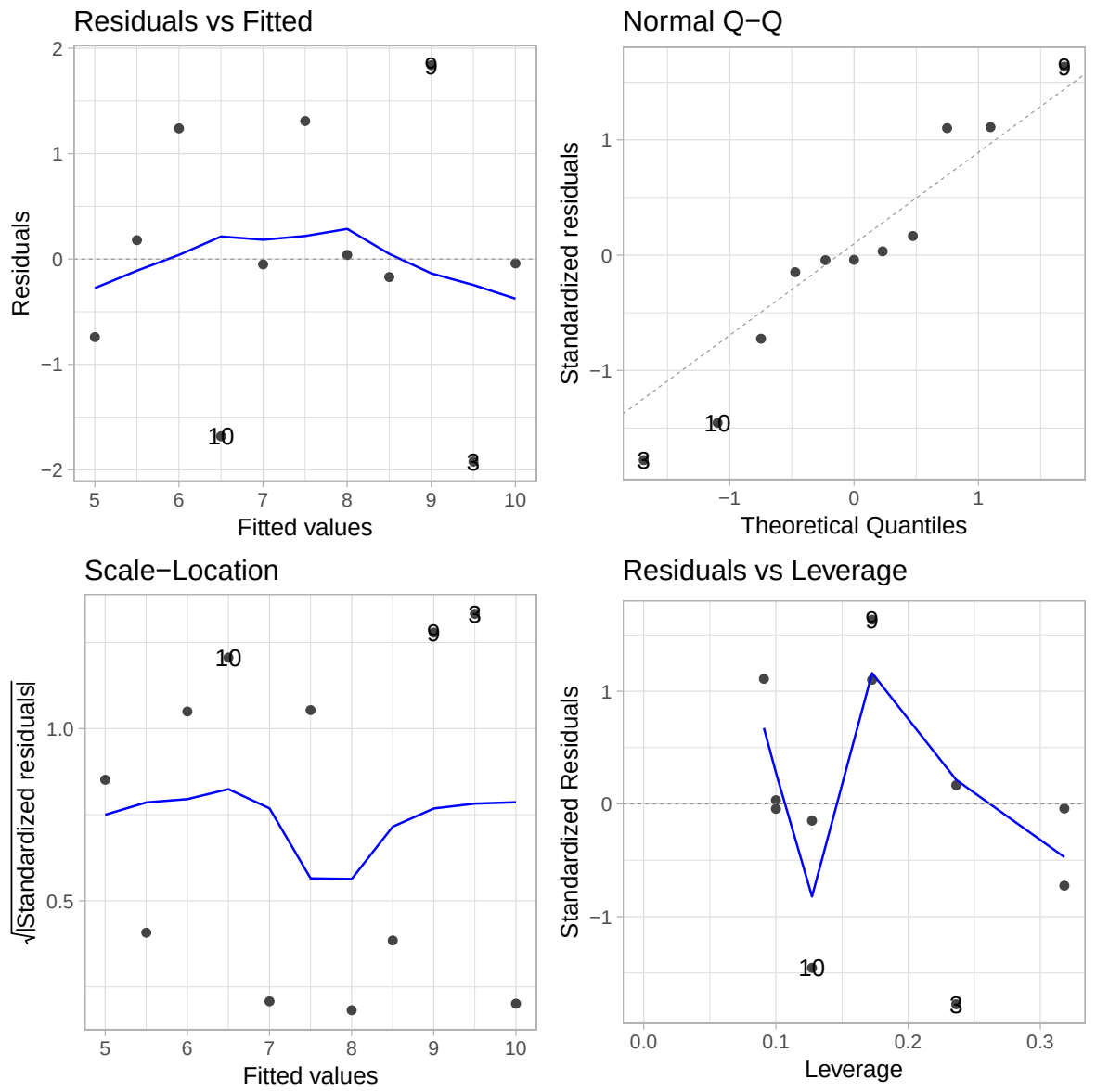
Przykład Anscombe-a — przypadek 3

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
x3            0.4997     0.1179   4.239  0.00218 **
---

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6663,  Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

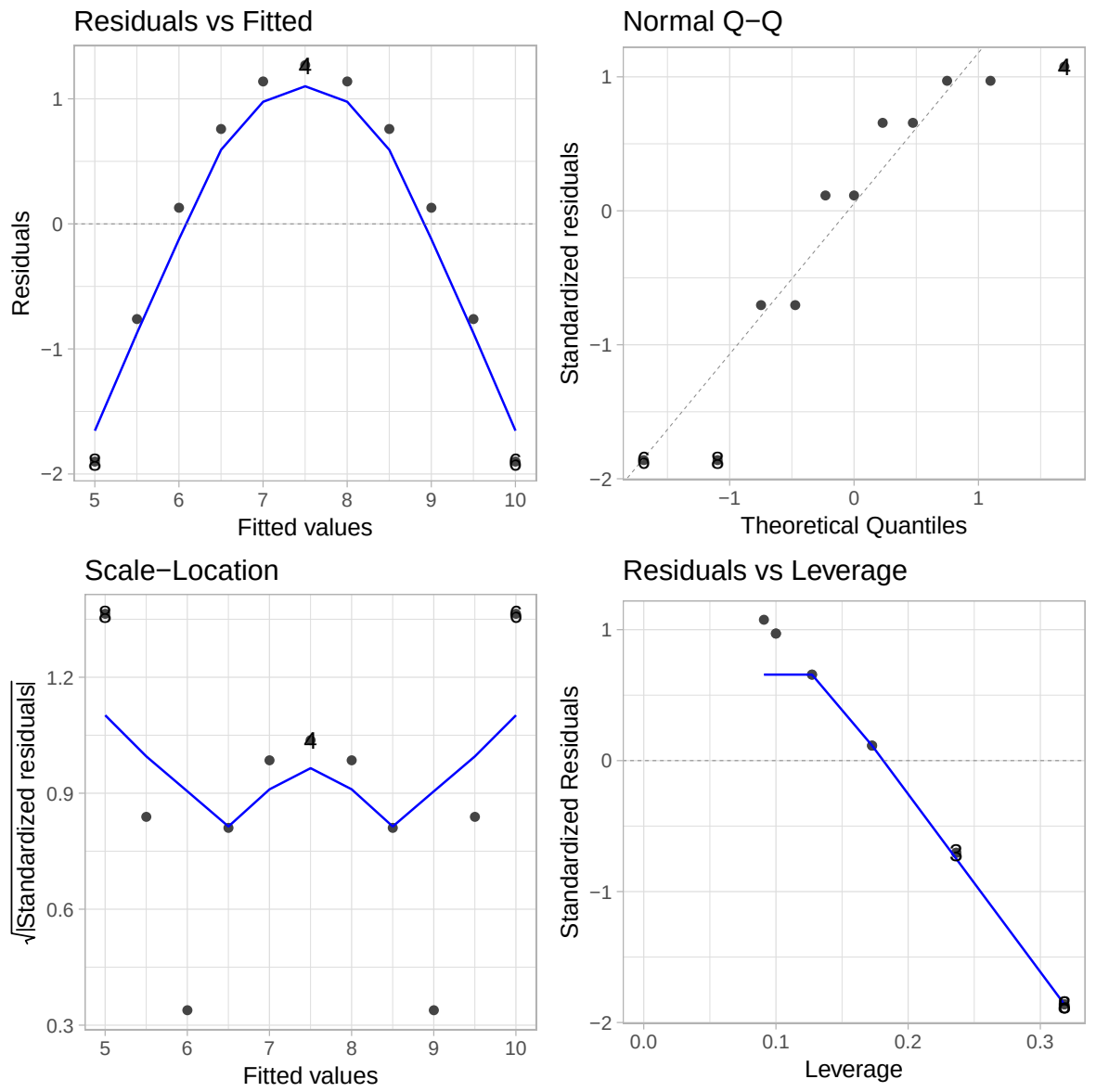
Przykład Anscombe-a — przypadek 4

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **
---

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667,  Adjusted R-squared:  0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

Przykład Anscombe-a — przypadek 1 — diagnostyka



Przykład Anscombe-a — przypadek 2 — diagnostyka



Przedział ufności dla predykcji

Przypuśćmy, że otrzymaliśmy nową wartość X_{n+1} i chcemy przewidzieć wartość Y_{n+1} (przy założeniach regresji liniowej), tzn. chcemy wyznaczyć estymator $\hat{Y}_{n+1}$.

W naturalny sposób bierzemy $\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1}$. Mamy wówczas:

$$\mathbb{E}\hat{Y}_{n+1} = \mathbb{E}Y_{n+1}$$

oraz:

$$\mathbb{V}\hat{Y}_{n+1} = \sigma^2 \left(\frac{1}{n} + \frac{(\overline{X} - X_{n+1})^2}{n(\overline{X^2} - \overline{X}^2)} \right).$$

Zatem:

$$\hat{Y}_{n+1} - Y_{n+1} \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(\overline{X} - X_{n+1})^2}{n(\overline{X^2} - \overline{X}^2)} \right)\right).$$

Przedział ufności dla predykcji

$$\hat{Y}_{n+1} - Y_{n+1} \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(\overline{X} - X_{n+1})^2}{n(\overline{X^2} - \overline{X}^2)}\right)\right).$$

Zatem

$$\frac{\hat{Y}_{n+1} - Y_{n+1}}{s_*} \sim t_{n-2},$$

gdzie

$$s_*^2 = s^2\left(1 + \frac{1}{n} + \frac{(\overline{X} - X_{n+1})^2}{n(\overline{X^2} - \overline{X}^2)}\right).$$

Jeśli $c = F^{-1}_{t_{n-2}}(1 - \frac{\alpha}{2})$, to końcami przedziału ufności dla $\hat{Y}_{n+1}$ na poziomie ufności $1 - \alpha$ są wartości $\hat{Y}_{n+1} \pm cs_*$.

Dlaczego właściwie regresja

$$\hat{\beta}_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2} = \frac{c_{XY}}{s^2_{XX}}, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

Zatem:

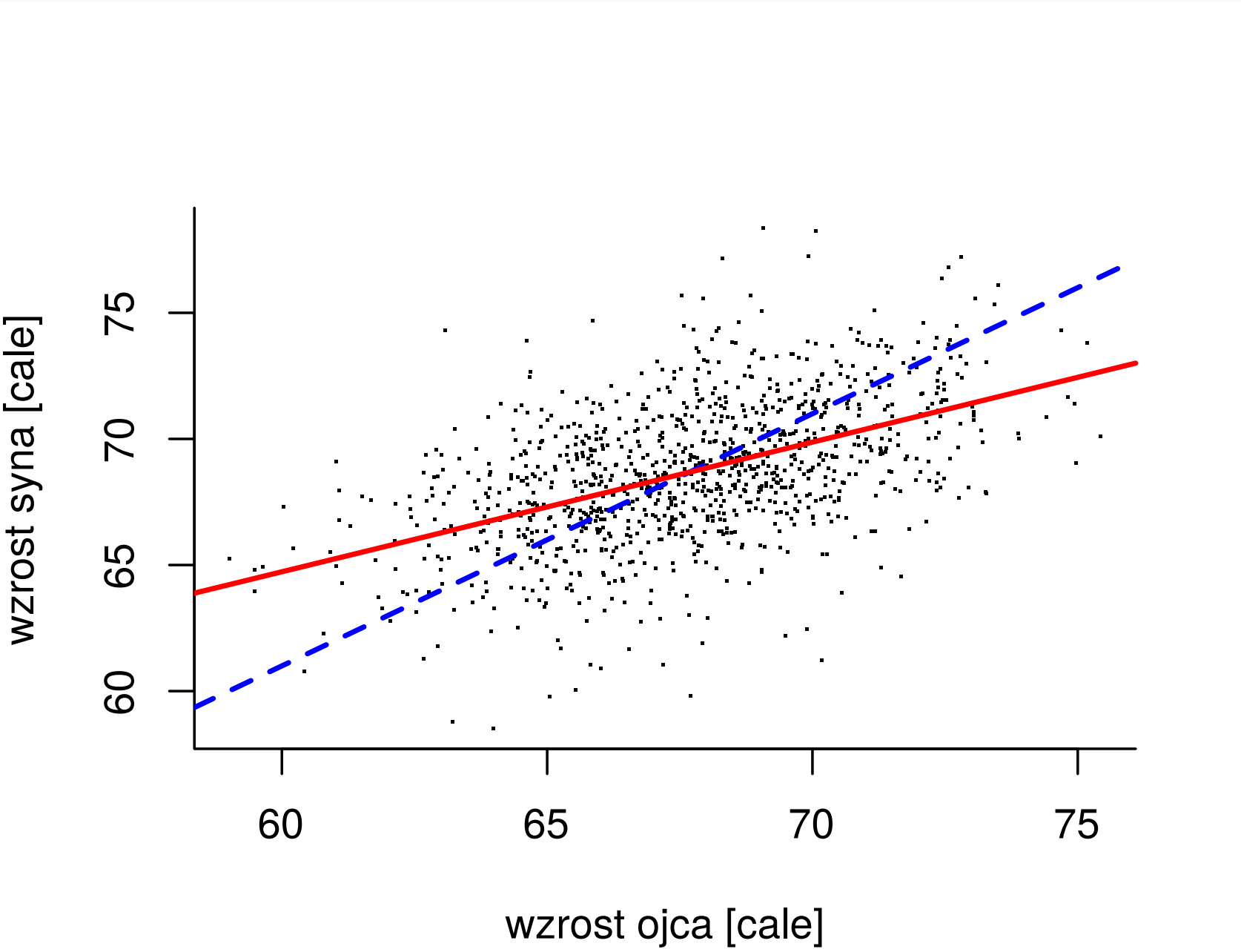
$$\frac{\hat{y} - \overline{Y}}{s_{YY}} = \underbrace{\left(\frac{c_{XY}}{s_{XX} s_{YY}} \right)}_{r} \frac{x - \overline{X}}{s_{XX}},$$

r to wyestymowany współczynnik korelacji $r \in [-1, 1]$.

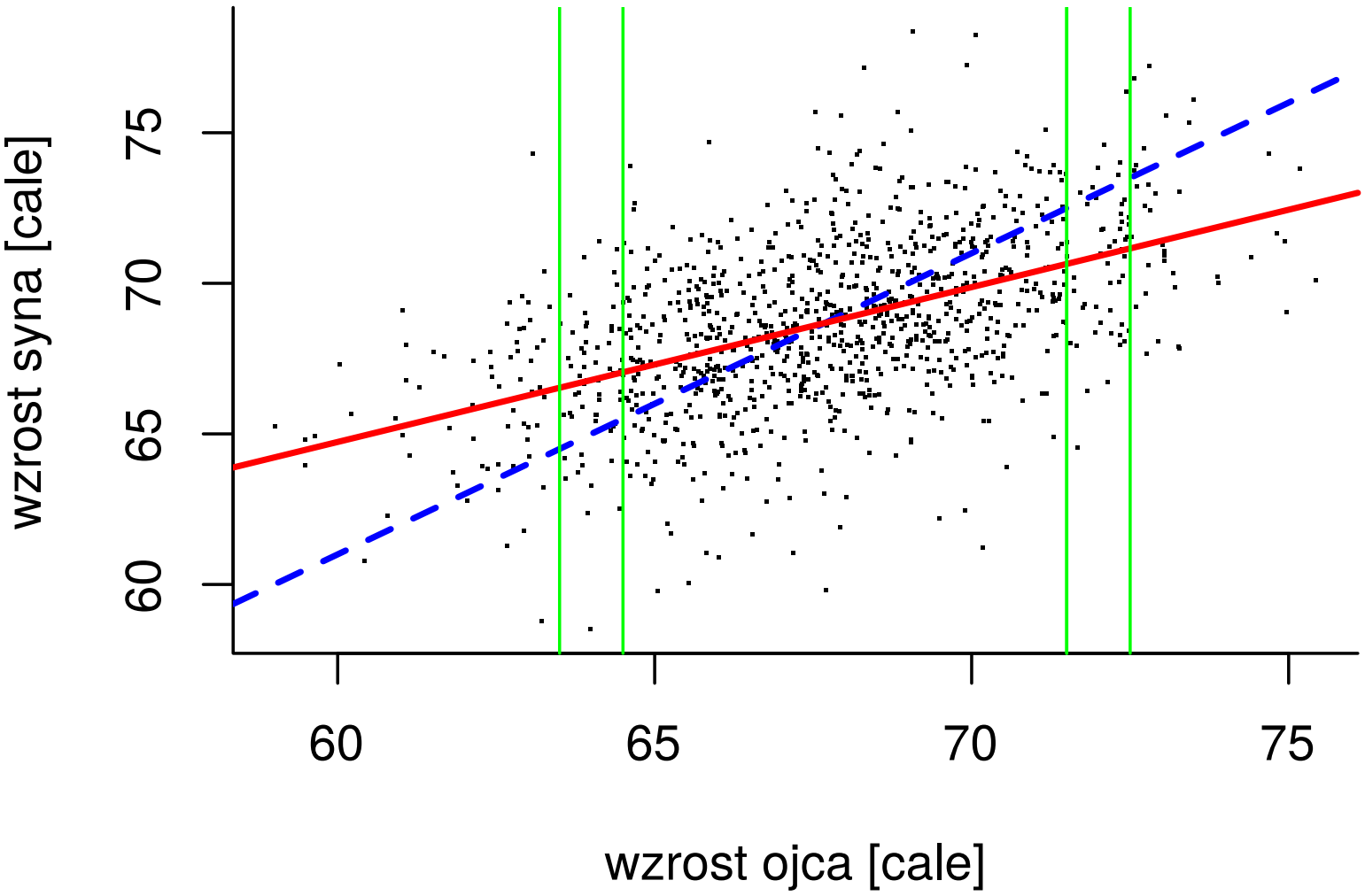
Załóżmy, że $1 > r > 0$ wówczas powyższe równanie można interpretować następująco: odstępstwo od średniej wartości zmiennej objaśnianej y liczone w odchyleniach standardowych σ_Y jest mniejsze $(0 < r < 1)$ niż odchylenie od średniej wartości zmiennej objaśniającej x liczone w odch. standardowych σ_X.

F. Galton nazywał to regresją w kierunku przeciętności.

Przykład



Przykład



Regresja liniowa — przypadek wielu zmiennych

$$Y_k = \beta_{p-1} X_{k,p-1} + \beta_{p-2} X_{k,p-2} + \cdots + \beta_1 X_{k,1} + \beta_0 \underbrace{1}_{x_{k,0}} + \varepsilon_k$$

$X_{i,j}$ traktujemy jako dane, $\varepsilon_k \sim N(0, \sigma^2)$ niezależne.

Dane

$$y_k,\ x_{k,0},\ \ldots,\ x_{k,p-1}, \qquad k = 1, \ldots, n.$$

Oznaczenia

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$$\mathbf{X} = \begin{pmatrix} x_{1,0} & x_{1,1} & \dots & x_{1,p-1} \\ x_{2,0} & x_{2,1} & \dots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,0} & x_{n,1} & \dots & x_{n,p-1} \end{pmatrix}.$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}\boldsymbol{\varepsilon} = 0, \quad \Sigma_{\varepsilon\varepsilon} = \sigma^2 I.$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}\boldsymbol{\varepsilon} = 0, \quad \Sigma_{\varepsilon\varepsilon} = \sigma^2 I.$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \quad \mathbb{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}.$$

$$\Sigma_{\hat{\beta}\hat{\beta}} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Nieobciążonym estymatorem wariancji σ^2 jest

$$s^2 = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2}{n - p}.$$

Inne modele

Zdarza się, że nieliniowe modele dają się sprowadzić do liniowych.

Przykład 1

Niech

$$y_i = A\cos(\omega t_i + \phi) + B + \varepsilon_i,$$

gdzie nieznanymi parametrami są amplituda A, faza ϕ oraz składowa stała B. Pulsacja ω jest znana. Posiadamy zestaw par (t_i, y_i).

$$y = \underbrace{A\cos(\phi)}_{\beta_2}\underbrace{\cos(\omega t)}_{x_2} + \underbrace{A\sin(-\phi)}_{\beta_1}\underbrace{\sin(\omega t)}_{x_1} + \underbrace{B}_{\beta_0} + \varepsilon.$$

W ramach wstępnego przetwarzania danych z par (t_i, y_i) tworzymy czwórki $(\cos(\omega t_i), \sin(\omega t_i), 1, y_i)$. W ramach przetwarzania końcowego, na podstawie estymat $\hat{\beta}_1$ oraz $\hat{\beta}_2$ odtwarzamy:

$$\hat{A} = \sqrt{\hat{\beta}_1{}^2 + \hat{\beta}_2{}^2}, \quad \hat{\phi} \text{ takie, że } \cos\hat{\phi} = \hat{\beta}_2/\hat{A}, \sin\hat{\phi} = -\hat{\beta}_1/\hat{A}.$$

Przykład 2

Niech

$$y_i = Ax^2 + Bx + C + \varepsilon_i.$$

Dysponujemy parami (x, y).

$$y_i = \underbrace{A}_{\beta_2}\underbrace{x^2}_{x_2} + \underbrace{B}_{\beta_1}\underbrace{x}_{x_1} + \underbrace{C}_{\beta_0} + \varepsilon_i.$$

Model heteroskedastyczny

$$y_k = \sum_{i=1}^{n} \beta_i x_{i,k} + \beta_0 + \varepsilon_k,$$

gdzie $\varepsilon_k \sim N(0, \rho_k^2 \sigma^2)$ i znamy ρ_k, $k = 1, \ldots, n$.

Można sprowadzić do modelu standardowego

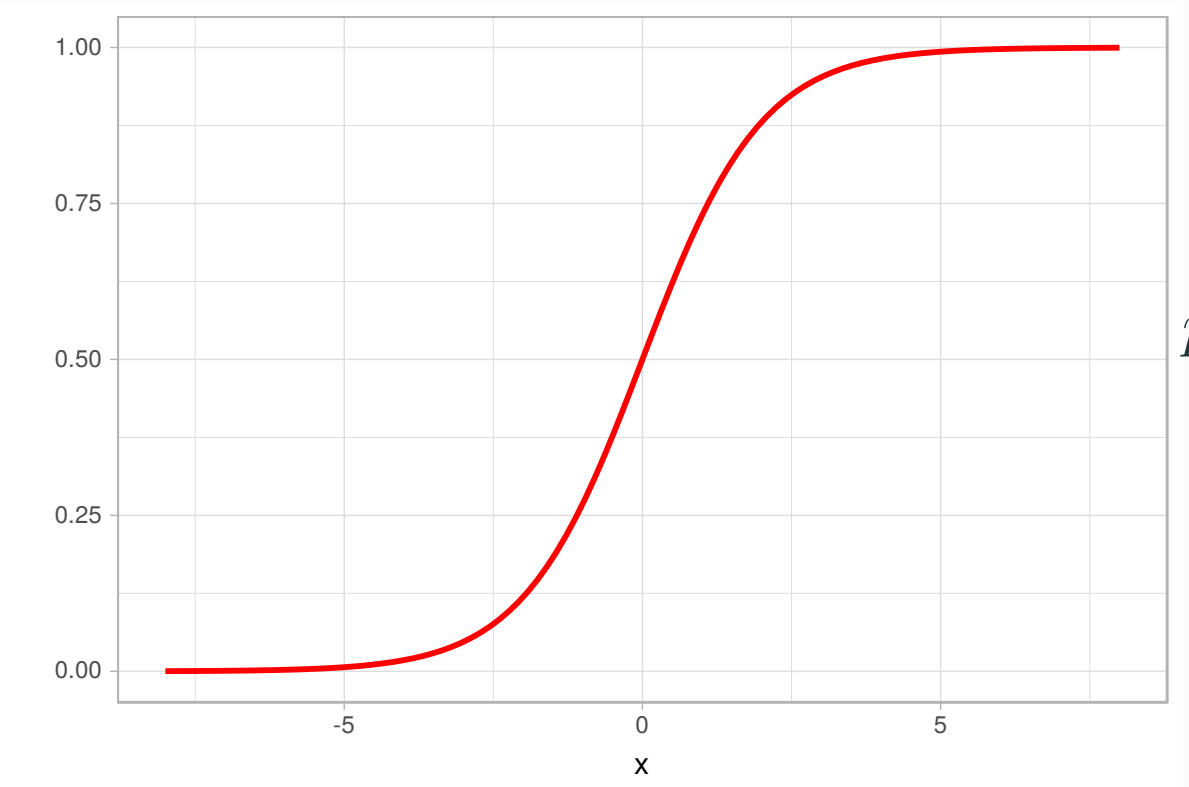
$$\underbrace{\frac{y_k}{\rho_k}}_{y'_k} = \sum_{i=1}^{n} \beta_i \underbrace{\frac{x_{i,k}}{\rho_k}}_{x'_{k,i}} + \beta_0 \underbrace{\frac{1}{\rho_k}}_{x'_{k,0}} + \underbrace{\frac{\varepsilon_k}{\rho_k}}_{\sim N(0,\sigma^2)}.$$

Regresja logistyczna

Metodę regresji można stosować również w kontekście binarnej klasyfikacji, tzn. w przypadku gdy zmienna objaśniana przyjmuje wartości $0, 1$. Model statystyczny:

$$\mathbb{P}(Y_i = 1 | X = \mathbf{x}) = \frac{\mathrm{e}^{\mathbf{x}\boldsymbol{\beta}}}{1 + \mathrm{e}^{\mathbf{x}\boldsymbol{\beta}}} = p(\mathbf{x}\boldsymbol{\beta}),$$

gdzie p to tzw. funkcja logistyczna



$$p(u) = \frac{\mathrm{e}^{u}}{1+\mathrm{e}^{u}}$$