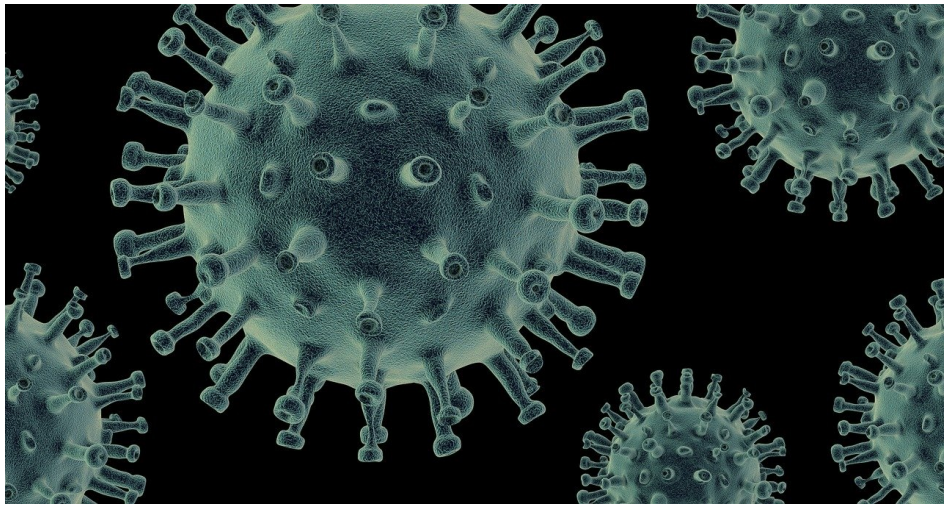
Model regresji wielorakiej

10 listopada 2022



Model regresji wielorakiej - sformułowanie

Rozważamy model

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \ldots + \beta_k X_{t,k} + \epsilon_t,$$

dla $t = 1, 2, \ldots, n$, gdzie

- $X_{t,1}, X_{t,2}, \ldots, X_{t,k}$ - zaobserwowane wartości zmiennych objaśniających
- Y_t - zaobserwowane wartości zmiennych objaśnianych
- ϵ_t - nieznane wartości składnika losowego
- $\beta_0, \beta_1, \ldots, \beta_k$ - nieznane wartości parametrów modelu.

Model można wyrazi, jako

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \ldots + \beta_k X_{t,k} + \epsilon_t,$$

lub w formie układu równań liniowych:

$$\begin{cases} Y_1 &= \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{1,2} + \ldots + \beta_k X_{1,k} + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{2,1} + \beta_2 X_{2,2} + \ldots + \beta_k X_{2,k} + \epsilon_2 \\ \vdots & \vdots \qquad\qquad\qquad\qquad \vdots \\ Y_n &= \beta_0 + \beta_1 X_{n,1} + \beta_2 X_{n,2} + \ldots + \beta_k X_{n,k} + \epsilon_n \end{cases}$$

który można wyrazić w formie macierzowej.

Model liniowy - forma macierzowa

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\beta} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\epsilon}.$$

Innymi słowy, model wyrażamy w formie:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

Estymator, reszty i prognozy

- Estymator parametru wektorowego β ma formę

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

- Sformułujemy definicję reszt i prognoz.

Definicja 1 (Prognozy i reszty)

Prognozami *wartości Y nazywamy wielkość*

$$\hat{\mathbf{Y}} = X\hat{\beta}.$$

Resztami *modelu nazywamy*

$$\hat{\epsilon} := \mathbf{Y} - \mathbf{X}\hat{\beta}.$$

Uwaga 1

Prognozy $\hat{Y}$ można traktować jako estymator parametru Y. Reszty $\hat{\epsilon}$ można traktować jako naturalny estymator składnika losowego ϵ.

Uwagi o prognozach na przyszłość

- Definicja $\hat{Y}$ wskazuje, że $\hat{Y}_t$ jest to prognoza obecnych i znanych wartości $Y_1, Y_2, ..., Y_n$ na podstawie znanych wartości:

$$\begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,k} \\ X_{2,1} & X_{2,2} & \dots & X_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,k} \end{bmatrix}.$$

- Aby opracować prognozy $\hat{Y}_{n+1}, ..., \hat{Y}_{n+\tau}$ przyszłych wartości $Y_{n+1}, ..., Y_{n+\tau}$ trzeba byłoby znać przyszłe wartości wartości objaśniających:

$$\begin{bmatrix} X_{n+1,1} & X_{n+1,2} & \dots & X_{n+1,k} \\ X_{n+2,1} & X_{n+2,2} & \dots & X_{n+2,k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n+\tau,1} & X_{n+\tau,2} & \dots & X_{n+\tau,k} \end{bmatrix}. \quad \text{(VERTE)}$$

Uwagi o prognozach na przyszłość

- Wtedy

$$\begin{aligned}\hat{Y}_{n+j} &= [1\, X_{n+j,1}\, X_{n+j,2}\, \ldots\, X_{n+j,k}] \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \\ &= \hat{\beta}_0 + \hat{\beta}_1 X_{n+j,1} + \ldots + \hat{\beta}_k X_{n+j,k}.\end{aligned}$$

- W modelach gdzie $X_{t,i}$ występuje jako funkcja t, przyszłe wartości X mamy "za darmo" np. w modelu

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_k t^k + \epsilon_t,$$

gorzej gdy $X_{t,i}$ to "suche" liczby. Wtedy trzeba czekać, aż pojawią się nowe wartości $X_{n+j,i}$.

Współczynnik determinacji

Szukamy zagregowanego współczynnika, który będzie mierzył odległość $\hat{\mathbf{Y}}$ od $\mathbf{Y}$.

- Współczynnik ten ma podsumować różnice między wektorami $\hat{\mathbf{Y}}$ oraz $\mathbf{Y}$ uwzględniając wszystkie współrzędne.
- Wspłczynnik ma za zadanie zmierzyć jak bardzo prognozy $\hat{Y}$ odzwierciedlają prawdziwe wartości Y, **w stosunku do faktycznej zmienności zmiennych objaśnianych** (inna zmienność jest przy pomiarze odległości planet w Układzie Słonecznych, a inna jest przy pomiarze odległości między atomami i cząsteczkami).

Współczynnik determinacji - konstrukcja R^2

Podsumowanie:

- dla modelu

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

- estymator β jest postaci

$$\hat{\beta} = (\mathbf{X}^T * \mathbf{X})^{-1} * \mathbf{X}^T * \mathbf{Y},$$

Współczynnik determinacji - konstrukcja R^2

- prognoza ma postać

$$\hat{Y} = \mathbf{X}\hat{\beta} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

- średnia arytmetyczna zaobserwowanych wartości Y_t:

$$\bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t.$$

Współczynnik determinacji - konstrukcja R^2

Wychodzimy z oczywistej równości:

$$Y_t - \bar{y} = (Y_t - \hat{Y}_t) + (\hat{Y}_t - \bar{Y}).$$

Mniej oczywista ale prawdziwa jest równość:

$$\sum_{t=1}^{n}(Y_t - \bar{Y})^2 = \sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 + \sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2$$

oraz

$$\bar{\hat{Y}} = \bar{Y}.$$

Współczynnik determinacji - konstrukcja R^2

Zmienność zmiennej objaśnianej:

$$\underbrace{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}_{\text{zmieność całkowita}} = \underbrace{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}_{\text{zmienność niewyjaśniona przez model}}$$

$$+ \underbrace{\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2}_{\text{zmienność wyjaśniona przez model}} .$$

Współczynnik determinacji - konstrukcja R^2

Definicja 2 (R^2 - współczynnik determinacji)

Współczynnik determinacji R^2 *jest zdefiniowany jako:*

$$R^2 = \frac{\sum_{t=1}^{n} (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2}.$$

Uwaga 2

Współczynnik determinacji mierzy proporcję między ***zmiennością wyjaśnioną przez model****, a* ***całkowitą zmiennoścnią zmiennej objaśnianej****.*

Własności R^2

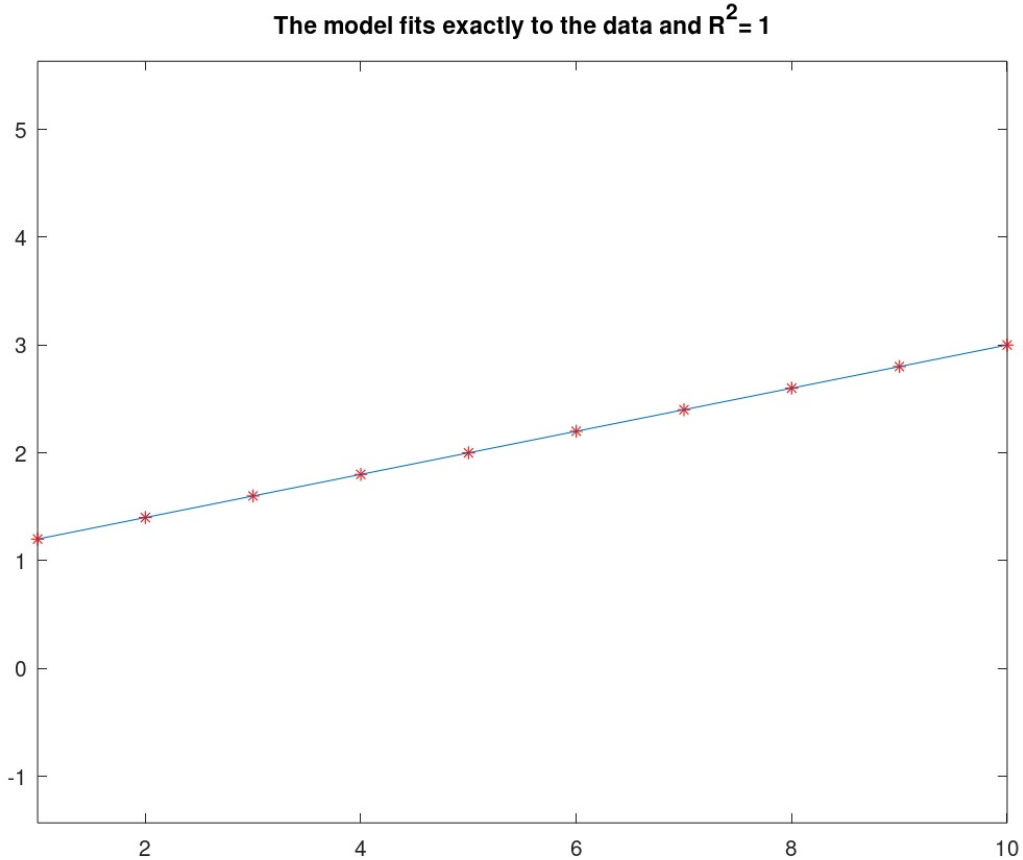
- $R^2 \in [0,1]$:

$$R^2 = \frac{\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} = \frac{\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 + \sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2}.$$

- Jeśli $R^2 = 1$ wtedy $\hat{Y}_t = Y_t$ dla wszystkich t, prognozy pokrywają się z danymi;
- Jeśli $R^2 = 0$ wtedy trajektoria prognoz pokrywa się z linią poziomą $y = \bar{Y}$ (tradycyjny układ (x, y)), tzn. $\hat{Y}_t = \bar{Y}$ dla $t = 1, 2, ..., n$.

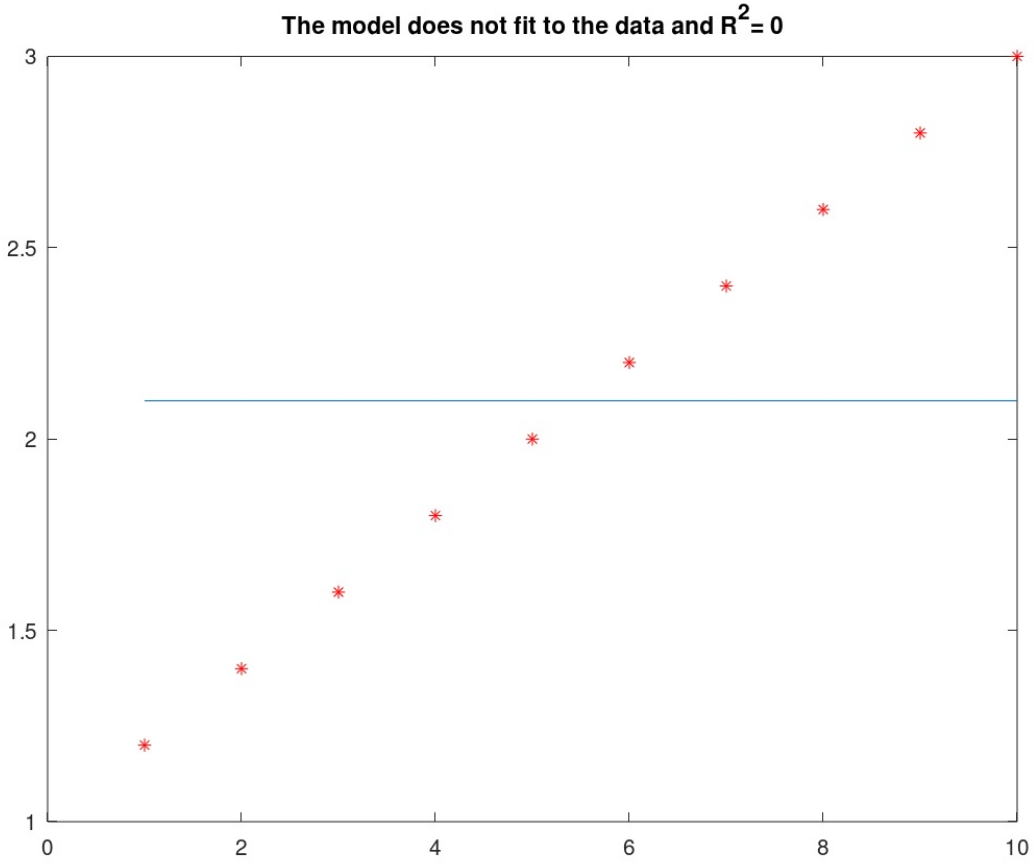
Własności R^2 - ilustracja z $R^2 = 1$

Rozważmy $Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$. Trajektoria $(X_t, Y_t)_{t=1,\ldots,n}$ leży w całości na trajektorii prognoz (niebieska linia przechodzi przez $(X_t, \hat{Y}_t)_{t=1}^{\infty}$).



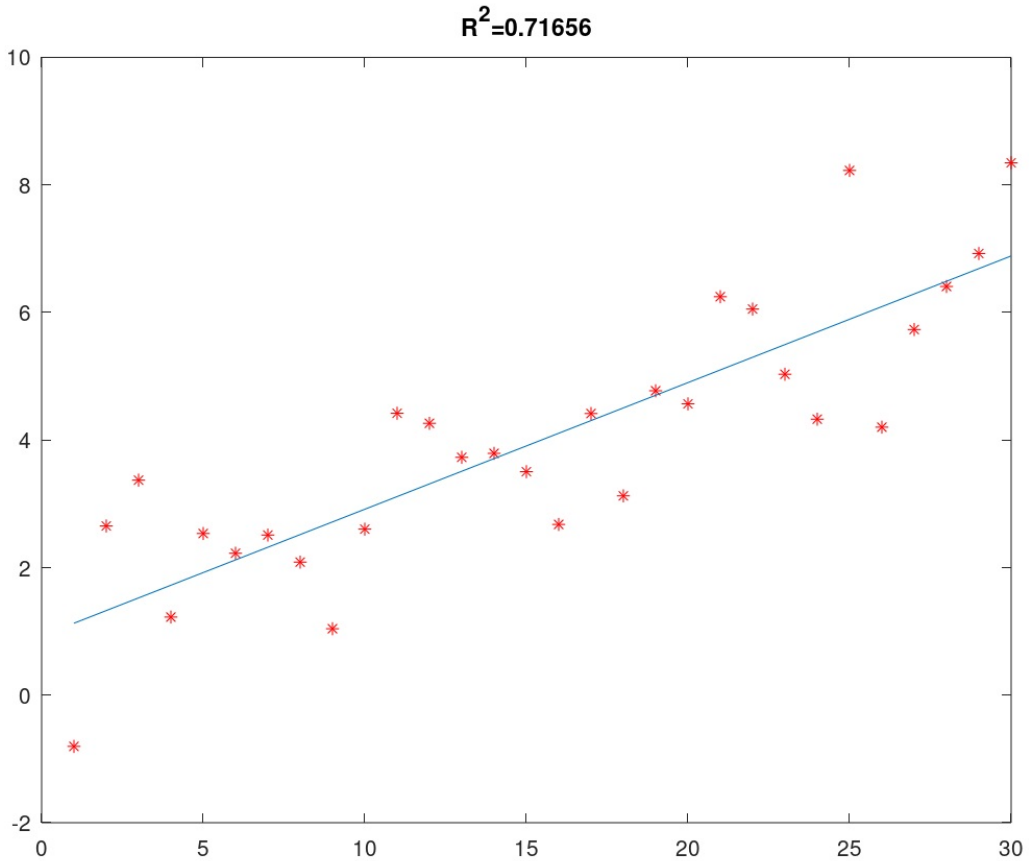
Własności R^2 - ilustracja z $R^2 = 0$

Trajektoria $(X_t, \hat{Y}_t)_{t=1,\ldots,n}$ leży w całości na prostej pionowej.



Własności R^2 - ilustracja z $R^2 \in (0, 1)$

Trajektoria $(X_t, Y_t)_{t=1,\ldots,n}$ oplata linię regresji (na rysunku niebieska linia przechodzi przez $(X_t, \hat{Y}_t)_{t=1}^{\infty}$).



Alternatywne wzory na R^2

Poniżej wprowadzam alternatywne sformułowania na R^2:

$$\begin{aligned} R^2 &= \frac{\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \\ &= 1 - \frac{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \\ &= 1 - \frac{\hat{\epsilon}^T * \hat{\epsilon}}{\mathbf{Y}^T * \mathbf{Y} - n\bar{Y}^2}. \end{aligned}$$

Skorygowany współczynnik determinacji

Definicja 3 (Skorygowany współczynnik determinacji)

Skorygowany współczynnik determinacji R^2 *definiujemy jako*

$$\tilde{R}^2 = R^2 - \frac{k}{n-k-1}(1-R^2).$$

Uwaga 3

Współczynnik $\tilde{R}^2$ *jest użyteczny gdy liczba parametrów w modelu* $k+1$ *jest niewiele mniejsza niż liczba obserwacji* n*:*

- *czasami* $\tilde{R}^2$ *jest liczbą ujemną;*
- *zawsze* $\tilde{R}^2 \leq R^2$ *oraz* $\tilde{R}^2 = R^2$ *wtedy i tylko wtedy* $R^2 = 1$*;*

Współczynnik $\tilde{R}^2$ *jest stosowany w celu* ***obniżenia rangi modelu****, który zawiera zbyt dużo parametrów. Model ze zbyt wieloma parametrami nie jest zbyt przejrzysty.*

Przykład

W tym przykładzie analizujemy miesięczną temperaturę w Poznaniu w latach 1999-2004. Dane miesięcznej temperatury dla miasta Poznań podsumowuje poniższa tabela.

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1999	1,3	-0,6	5	9,5	13,7	16,4	20,5	17,9	16,7	8,4	2,7	1,6
2000	-0,2	3,2	4	11,9	16	17,8	16,3	18,2	12,6	11,9	6,4	2,3
2001	0	0,3	2,4	8,1	14,9	15,1	20,1	19,6	12	11,9	3,1	-1,6
2002	0,7	3,9	4,5	8,8	17	17,9	20,5	21,1	13,7	7,2	4,2	-3,6
2003	-2	-3,6	2,5	8,2	15,9	19,4	19,7	19,8	14,2	5,3	5,5	1,7
2004	-4	1,4	4,4	9,4	12,8	16,2	17,9	19,8	13,9	10	4,1	1,7

Przykład

- Tu $Y_1 = 1.3,\ Y_2 = -0.6,\ ..., Y_{12} = 1.6,\ Y_{13} = -0.2,\ ...,$ $Y_{72} = 1.7$ oznacza zmienną objaśnianą (temperatura w Poznaniu);
- Tworzymy zmienne objaśniające jako $X_{t,1} = t$, $X_{t,2} = \cos\left(\frac{\pi}{6}t\right)$, $X_{t,3} = \cos\left(\frac{\pi}{6}t\right)$;
- Model ma następującą formułę:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 \cos\left(\frac{\pi}{6}t\right) + \beta_3 \sin\left(\frac{\pi}{6}t\right) + \epsilon_t.$$

- Za pomocą metody najmniejszych kwadratów mamy $\hat{\beta}_0 = 9.736$, $\hat{\beta}_1 = -0.012$, $\hat{\beta}_2 = -8.77$, $\hat{\beta}_3 = -5.14$ oraz

$$\hat{Y}_t = 9.736 - 0.012 * t - 8.77 \cos\left(\frac{\pi}{6}t\right) - 5.14 \sin\left(\frac{\pi}{6}t\right).$$

Przykład

Ponieważ współczynnik $\hat{\beta}_1 = -0.012$ jest jedynym nieistotnym parametrem (wg obliczeń w Gretlu) modyfikujemy model:

$$Y_t = \beta_0 + \beta_1 \cos\left(\frac{\pi}{6}t\right) + \beta_2 \sin\left(\frac{\pi}{6}t\right) + \epsilon_t.$$

Tu $X_{t,1} = \cos\left(\frac{\pi}{6}t\right)$, $X_{t,2} = \sin\left(\frac{\pi}{6}t\right)$, natomiast Y_t jest jak poprzednio. Nowe obliczenia dają $\hat{\beta}_0 = 9.29861$, $\hat{\beta}_1 = -8.78099$, $\hat{\beta}_2 = -5.09812$. Stąd prognozy mają formułę

$$\hat{Y}_t = 9.29861 - 8.78099 \cos\left(\frac{\pi}{6}t\right) - 5.09812 \sin\left(\frac{\pi}{6}t\right)$$

dla $t = 1, 2, \ldots, 72$.

Przykład -C.D.

Prognozy $\hat{Y}_t$ dla miesięcznej temperatury w Poznaniu w latach 1999-2004 podsumowuje tabela:

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1999	-0.86	0.49	4.2	9.27	14.35	18.08	19.45	18.1	14.4	9.32	4.24	0.52
2000	-0.86	0.49	4.2	9.27	14.35	18.08	19.45	18.1	14.4	9.32	4.24	0.52
2001	-0.86	0.49	4.2	9.27	14.35	18.08	19.45	18.1	14.4	9.32	4.24	0.52
2002	-0.86	0.49	4.2	9.27	14.35	18.08	19.45	18.1	14.4	9.32	4.24	0.52
2003	-0.86	0.49	4.2	9.27	14.35	18.08	19.45	18.1	14.4	9.32	4.24	0.52
2004	-0.86	0.49	4.2	9.27	14.35	18.08	19.45	18.1	14.4	9.32	4.24	0.52

Wstawiając liczby Y_t (z przedostatniej tabeli) oraz $\hat{Y}_t$ (z ostatniej tabeli) do wzorów na współczynnik determinacji otrzymamy

$$R^2 = 0.942351 \quad \text{and} \quad \tilde{R}^2 = 0.940680.$$

Podsumowanie

Na podstawie współczynników determinacji wnioskujemy, że około 95% zmienności wartości temperatury w Poznaniu jest wyjaśnione przez model.

Dlaczego przeparametryzowanie modelu pogarsza przejrzystość?

Rozważam prosty model wygenerowany przez *Octave*. "Prawdziwy" model jest postaci

$$Y_t = 0.2. * t + 1 + \epsilon_t,$$

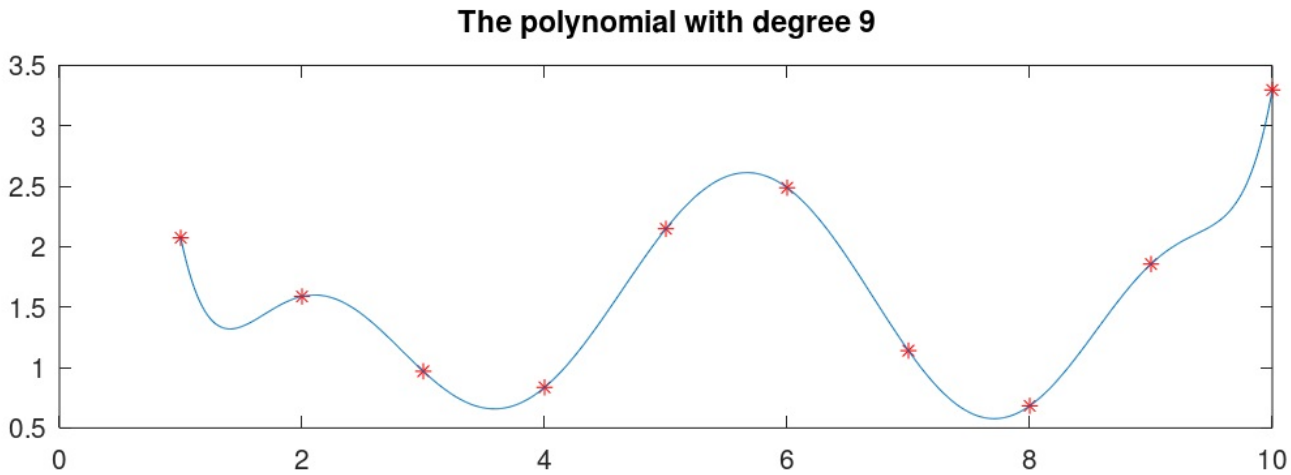
gdzie $t = 1, 2, ..., 10$ gdzie ϵ_t ma standardowy rozkład normalny $\mathcal{N}(0, 1)$. Dopasowujemy dwa modele:

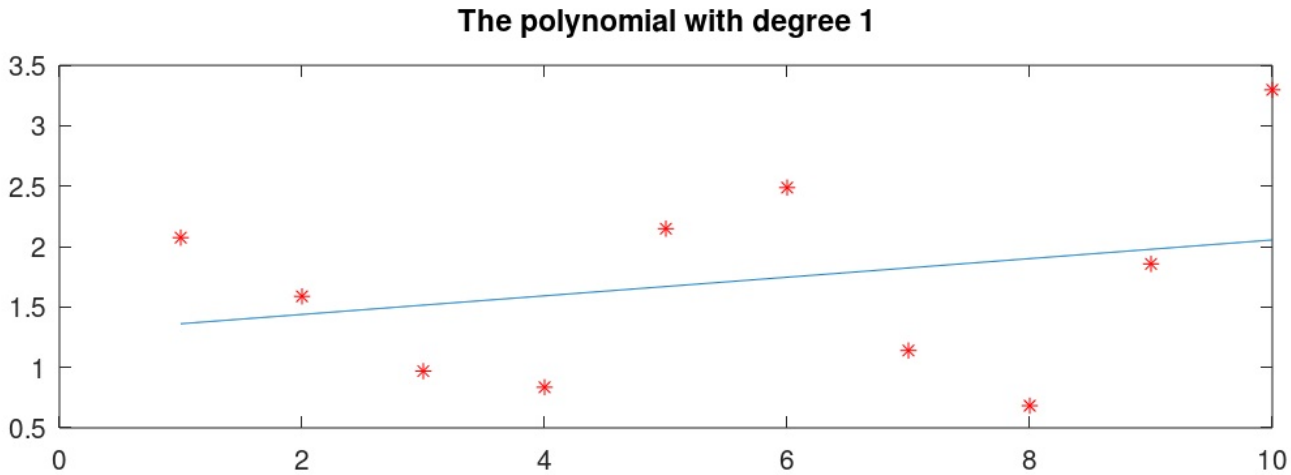
- wielomanowy stopnia 9 do danych (t, Y_t) (najwyższy "dozwolony" stopień p. założenie Z2 wykład 15.03.2021, slajd 13),
- liniowy model

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t.$$

Drugi wydaje się być bardziej przejrzysty mimo iż, pierwsza regresja "pasuje" idealnie (patrz rysunek)

Dlaczego przeparametryzowanie modelu pogarsza przejrzystość?





Dlaczego przeparametryzowanie modelu pogarsza przejrzystość?

- Gdy dopasujemy model wielomianowy stopnia 9 do danych (t, Y_t) i dopasujemy

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2 + \ldots + \hat{\beta}_9 t^9,$$

wielomian przechodzi przez (t, Y_t) (wielomian interpolacyjny), ale nic nie wnosi poza "naśladowaniem" danych. W szczególności model jest bezużyteczny dla *ekstrapolacji* i prognozowania. Daje to jednak $R^2 = 1$. Liczba parametrów jest jednak dużo wyższa niż w modelu liniowym.

Dlaczego przeparametryzowanie modelu pogarsza przejrzystość?

- Jeśli dopasujemy model

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$

do danych (t, Y_t), to model wykryje liniowy trend

$$\hat{Y}_t = 0.077023 * t + 1.284503$$

($\hat{\beta}_0 = 1.284503, \hat{\beta}_1 = 0.077023$) i jest lokalną aproksymacją trendu $y_t = 0.2. * t + 1$ (dwie różne linie proste w nieskończoności i tak znacznie się rozchodzą). Ponadto linia może prognozować **najbliższe** przyszłe wartości $t > 10$. Tu $R^2 = 0.578$.

Dlaczego przeparametryzowanie modelu pogarsza przejrzystość??

- Wg R^2 model wielomianowy ma większą rangę niż model liniowy.
- Powyższy przykład pokazuje, że R^2 jest bezużyteczny gdy porównujemy modele z dużą liczbą parametrów.