Zjazd 7

Maciej Rosoł

Agenda

- Regresja
- Metoda Najmniejszych Kwadratów
- Współczynnik determinacji R^2

Regresja

Regresja – metoda statystyczna pozwalająca na opisanie współzmienności kilku zmiennych przez dopasowanie do nich funkcji. Umożliwia przewidywanie nieznanych wartości jednych wielkości na podstawie znanych wartości innych.

Czyli znamy przykładowe wartości (x_i, y_i), ktoś nam podaje nowy punkt x_0 i chcemy przewidzieć wartość y_0.

Metoda Najmniejszych Kwadratów

W roku 1801 astronomowie zgubili z oczu asteroidę, i chodziło o to by odszukać ją z powrotem na niebie. Gauss stworzył **Metodę Najmniejszych Kwadratów** właśnie w celu by ja odszukać, co mu się udało – znalazła się dokładnie tam, gdzie Gauss przewidział, że będzie.

Metoda Najmniejszych Kwadratów

Dokonaliśmy pomiarów pewnej funkcji:

x_i	1	2	4
y_i	1	2	3

Podejrzewamy, że dane mogą być dobrze przybliżone za pomocą funkcji liniowej

$$y = ax + b$$

W związku z tym szukamy takich parametrów a, b aby przybliżenie

$$y_i = ax_i + b$$

dla i = 1, .., n (gdzie w naszym przypadku n = 3) było **optymalne**.

Metoda Najmniejszych Kwadratów

Inaczej mówiąc szukamy takich a, b

$$\begin{cases} 1a + b \approx 1 \\ 2a + b \approx 2 \\ 4a + b \approx 3 \end{cases}$$

co w zapisie macierzowym możemy przedstawić następująco:

$$\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 4 & 1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} \approx \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Metoda Najmniejszych Kwadratów

Pojawia się problem co to znaczy optymalne, oraz (jak to już doprecyzujemy) jak to optymalne znaleźć. Precyzyjniej mówiąc potrzebujemy dookreślić jaką **funkcję kosztu** tu dopasujemy, która będzie nam mówiła ile „kosztuje” nas dany błąd (w zależności od doboru parametrów a, b).

Metoda Najmniejszych Kwadratów

Naturalnym wydawałoby się posumowanie modułów błędów:

$$\mathrm{error}(a,b) = \sum_i |y_i - (ax_i + b)|$$

Tak się czasami robi, ale takie podejście ma wadę, bo nie da się tych współczynników wyliczyć jawnym wzorem. W związku z tym, dokonamy naturalnej modyfikacji, zastępując moduł kwadratem (Square Error):

$$se(a,b) = \sum_i (y_i - (ax_i + b))^2$$

Metoda Najmniejszych Kwadratów

Można pokazać, że funkcja kosztu $se(a, b)$ przyjmuje minimum w:

$$a = \frac{\sum y_i x_i - b \sum x_i}{\sum x_i^2}$$

$$b = \frac{\sum y_i - a \sum x_i}{n}$$

W naszym przypadku otrzymujemy układ:

$$a = \frac{17 - 7b}{21}, \quad b = \frac{6 - 7b}{3}$$

i wynik: $a = 9/10,\ b = 1/2$

Metoda Najmniejszych Kwadratów

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} \approx \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Można ten problem rozwiązać przy użyciu macierzy

Aby znaleźć optymalne przybliżenie, mnożymy obie strony przez A^T:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Rozwiązujemy prosty układ równań:

$$\begin{bmatrix} 3 & 7 \\ 7 & 21 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 6 \\ 17 \end{bmatrix}$$

I otrzymujemy $a = 9/10,\ b = 1/2$

Regresja liniowa

Możemy zastosować metodę regresji liniowej, gdy chcemy przewidzieć wartość jednej zmiennej na podstawie innych zmiennych. Np.:

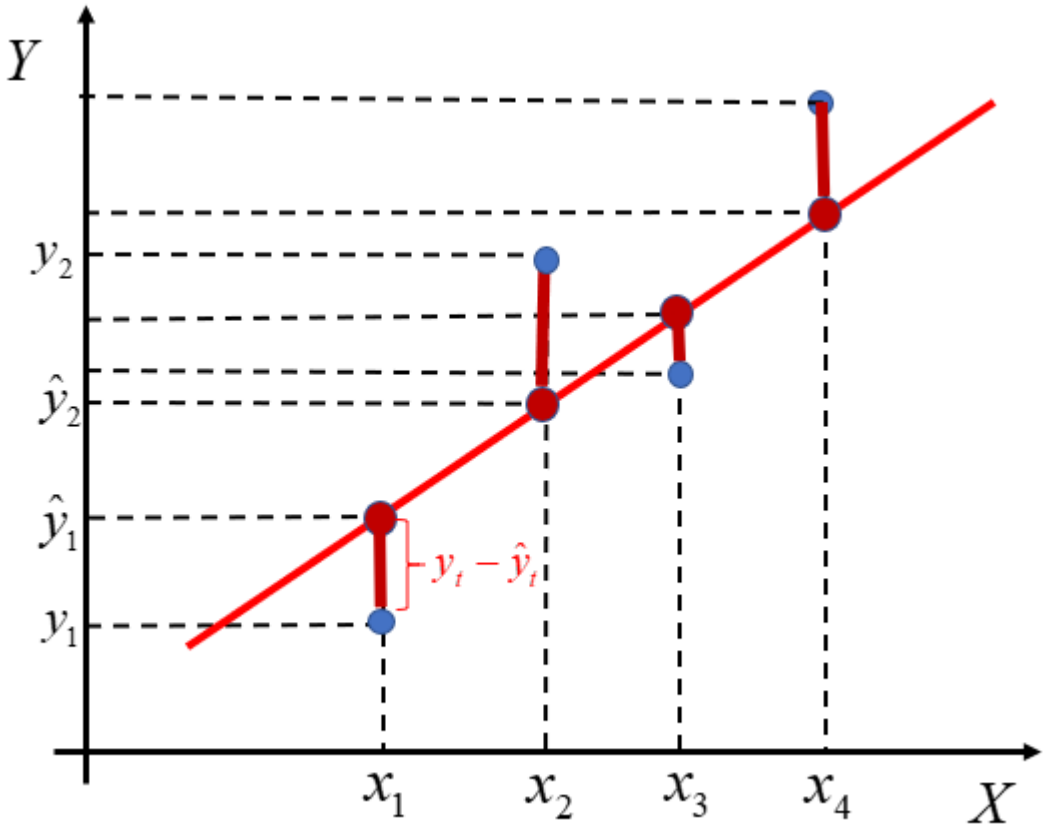
- Kaloryczność pokarmu na podstawie ilości węglowodanów,
- Zużycie energii elektrycznej w gospodarstwie domowym na podstawie liczby i wieku domowników,
- Wartość nieruchomości na podstawie powierzchni i odległości od centrum,
- Liczbę zachorowań na COVID na podstawie liczby osób w transporcie publicznym tydzień temu.

Regresja liniowa

Szukamy parametrów (a, b) które **minimalizują błąd kwadratowy** (squared residuals) ε_iw modelu:

$$y_i = ax_i + b + \varepsilon_i$$

gdzie a jest nachyleniem linii, b przesunięciem, ε_i (residua) są różnicami między zaobserwowanymi wartościami, a przewidywanymi wartościami.



https://blog.etrapez.pl/ekonometria/o-regresji-i-metodzie-najmniejszych-kwadratow-czyli-skad-wziely-sie-oszacowania-parametrow-modelu/

Regresja liniowa

Ponieważ równanie regresji liniowej jest stworzone w celu zminimalizowania sumy kwadratowej reszt (residua), regresja liniowa czasami nazywana jest **Ordinary Least-Squares** (**OLS**) Regression

Regresja liniowa

Załóżmy, że mamy kilka punktów (x_i, y_i), gdzie i = 1, 2, . . . , 7. Wtedy najprostszy model regresji liniowej ma postać:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Taki model można też zapisać w postaci macierzowej:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \\ 1 & x_7 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{bmatrix}$$

gdzie pierwsza kolumna w macierzy reprezentuje przesunięcie, a druga kolumna to wartości x_i odpowiada nachyleniu.

Regresja liniowa

Możemy zastosować regresję liniową również do bardziej złożonych modeli, jak na przykład:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

W postaci macierzowej:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \\ 1 & x_6 & x_6^2 \\ 1 & x_7 & x_7^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{bmatrix}$$

Jest to również przypadek regresji liniowej, ponieważ nieznane parametry βi pojawiają się liniowo, a składniki macierzy pojawiają się z kwadratami.

Regresja liniowa

- Zestaw danych zawiera wartości y_i, z których każda ma skojarzoną wartość modelową $\hat{y}_i$(czasami również oznaczaną f_i).
- Wartości y_i nazywane są wartościami zaobserwowanymi - observed values,
- Wartości modelowe f_i lub $\hat{y}_i$ wartościami przewidywanymi -predicted values ,
- Wartość $\bar{y}$ jest średnią z zaobserwowanych danych:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

gdzie n oznacza liczbę obserwacji.

Sum of squares

Sum of squares (SS) to miara, która służy do opisu „zmienność” danych i tego jak dobrze model dopasowuje się do naszych danych. Miary z wykorzystaniem SS to:

- Model Sum of Squares (Explained Sum of Squares)

$$SS_{mod} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Residuals Sum of Squares (sum of squares for the errors)

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Total Sum of Squares (równoważne wariancji próbki pomnożonej przez (n − 1)).

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Współczynnik determinacji

Dla modelu regresji liniowej mamy:

$$SS_{mod} + SS_{res} = SS_{tot}$$

Przy powyższych oznaczeniach **współczynnik determinacji** (coeficient of determination) oznaczamy R^2:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{model}}{SS_{tot}}$$

Współczynnik determinacji, to stosunek sumy kwadratów odległości zmiennej wyjaśnianej przez model do całkowitej sumy kwadratów.

Współczynnik determinacji

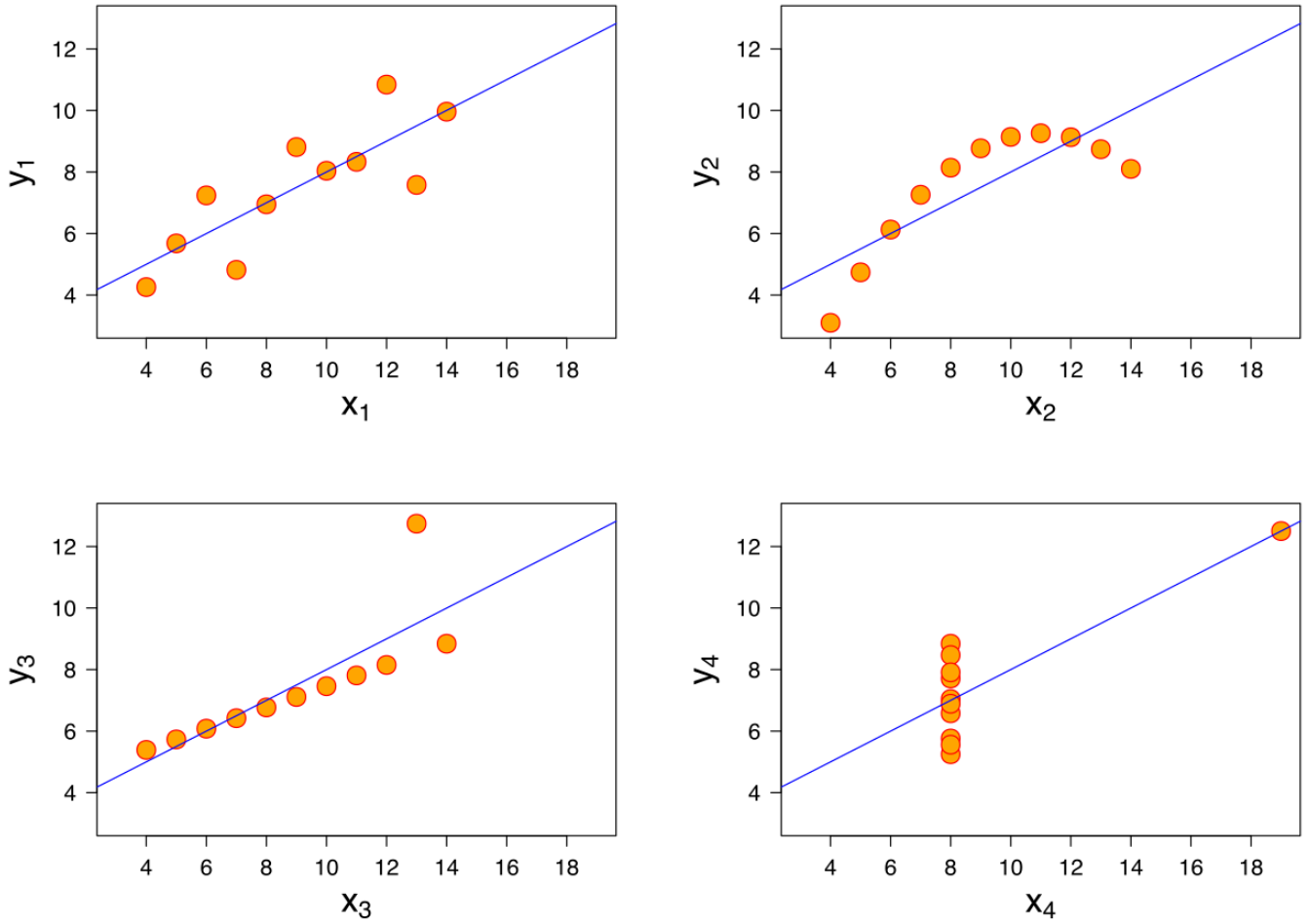
R^2 jest jedną z miar jakości dopasowania modelu do danych uczących. Wartości R^2 zbliżone do 1 odpowiada ścisłej korelacji, wartości zbliżone do 0 odpowiada słabej:

- 0,0 - 0,5 - dopasowanie niezadowalające,
- 0,5 - 0,6 - dopasowanie słabe,
- 0,6 - 0,8 - dopasowanie zadowalające,
- 0,8 - 0,9 - dopasowanie dobre,
- 0,9 - 1,0 - dopasowanie bardzo dobre.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Regresja liniowa

Kwartet Anscombe'a to zestaw czterech zestawów danych o identycznych cechach statystycznych, takich jak średnia arytmetyczna, wariancja, współczynnik korelacji, współczynnik determinacji R^2 czy równanie regresji liniowej, jednocześnie wyglądających zgoła różnie przy przedstawieniu graficznym.



https://pl.wikipedia.org/wiki/Kwartet_Anscombe%E2%80%99a

Regresja liniowa

Bardzo ogólna definicja modelu regresji jest następująca:

$$y = f(x, \varepsilon)$$

W przypadku prostego modelu regresji liniowej model może zostać zapisany jako:

$$y = X\beta + \varepsilon$$

Regresja liniowa

Dla danych w postaci:

$$\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n}$$

mówimy, że y_i jest zmienną objaśnianą, a $x_{i1}, \dots, x_{ip}$ są zmiennymi objaśniającymi, a model regresji ma postać:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i$$

gdzie ^{T} oznacza transpozycję, a $x_i^T \beta$ oznacza iloczyn skalarny.

Regresja liniowa

W notacji macierzowej:

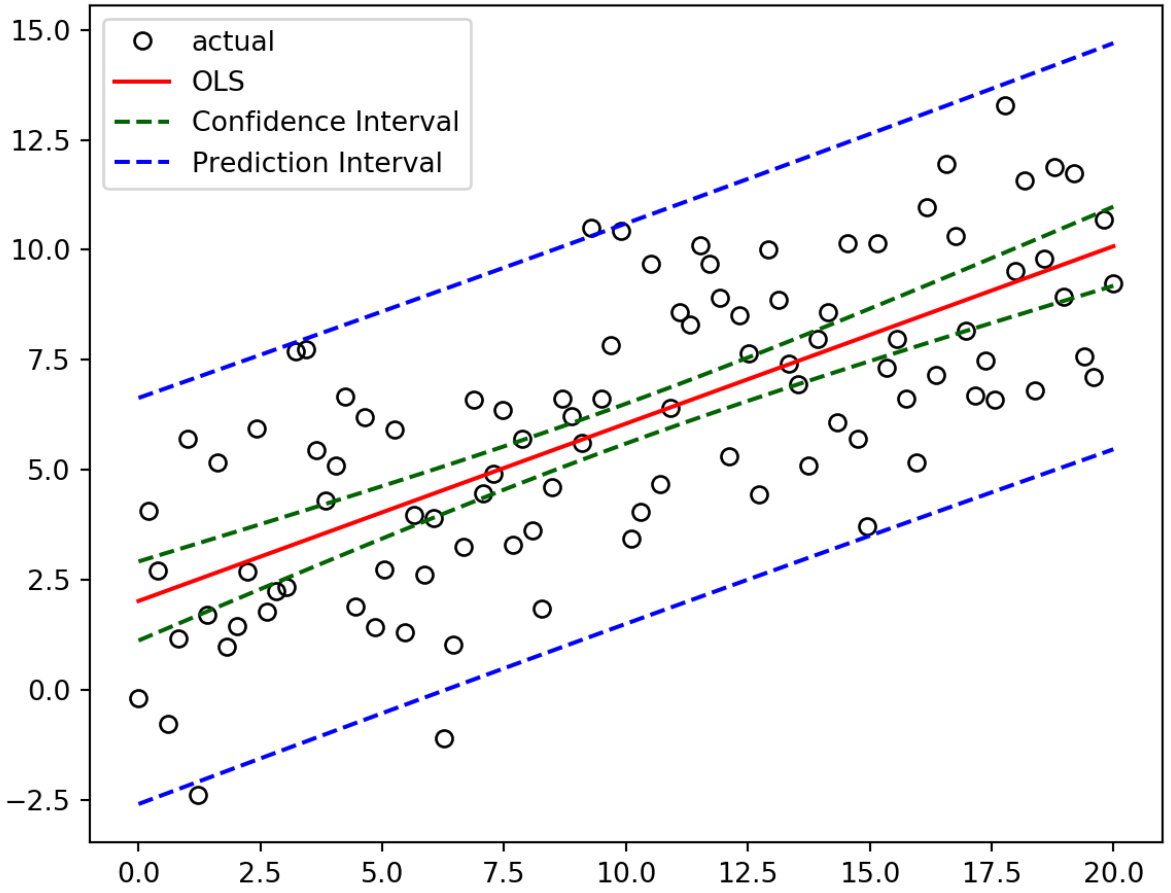
$$y = X\beta + \varepsilon$$

Gdzie:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, x = \begin{bmatrix} \mathrm{x}_1^T \\ \mathrm{x}_2^T \\ \vdots \\ \mathrm{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

Regresja liniowa

Dla regresji liniowej można wyznaczyć przedział ufności dla parametrów modelu (**confidence interval**) oraz dla predykowanych wartości (**prediction interval**)



http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/3-7-UnivarPredict.html

Notebook – D10_Z01

Regresja liniowa

- W następnej części wyjaśnimy wszystkie parametry.
- Na razie zwróćmy uwagę na Kryterium Akaike Information Criterion (**AIC**), które można wykorzystać do oceny jakości modelu.
- Im **niższa** jest wartość AIC, tym lepszy model.

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.470
Dependent Variable:	y	AIC:	415.2971
Date:	2020-11-22 11:18	BIC:	420.5075
No. Observations:	100	Log-Likelihood:	-205.65
Df Model:	1	F-statistic:	88.70
Df Residuals:	98	Prob (F-statistic):	2.22e-15
R-squared:	0.475	Scale:	3.6521

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	2.9869	0.3877	7.7044	0.0000	2.2175	3.7562
x1	2.1818	0.2317	9.4181	0.0000	1.7221	2.6415

Omnibus:	22.265	Durbin-Watson:	1.818
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31.869
Skew:	1.040	Prob(JB):	0.000
Kurtosis:	4.822	Condition No.:	4

Notebook – D10_Z02

Regresja liniowa

Lewa kolumna przeważnie zawiera informacje dotyczące użytej metody.

OLS Regression Results

Dep. Variable:	Alcohol	R-squared:	0.615
Model:	OLS	Adj. R-squared:	0.567
Method:	Least Squares	F-statistic:	12.78
Date:	Fri, 15 Jan 2021	Prob (F-statistic):	0.00723
Time:	11:28:07	Log-Likelihood:	-4.9998
No. Observations:	10	AIC:	14.00
Df Residuals:	8	BIC:	14.60
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.0412	1.001	2.038	0.076	-0.268	4.350
Tobacco	1.0059	0.281	3.576	0.007	0.357	1.655

Omnibus:	2.542	Durbin-Watson:	1.975
Prob(Omnibus):	0.281	Jarque-Bera (JB):	0.904
Skew:	-0.014	Prob(JB):	0.636
Kurtosis:	1.527	Cond. No.	27.2

Regresja liniowa

Df Model - oznacza stopnie swobody modelu czyli liczbę predyktorów (zmiennych objaśniających). Df Residuals - oznacza liczbę obserwacji pomniejszoną o stopnie swobody modelu minus jeden (dla przesunięcia).

OLS Regression Results

Dep. Variable:	Alcohol	R-squared:	0.615
Model:	OLS	Adj. R-squared:	0.567
Method:	Least Squares	F-statistic:	12.78
Date:	Fri, 15 Jan 2021	Prob (F-statistic):	0.00723
Time:	11:28:07	Log-Likelihood:	-4.9998
No. Observations:	10	AIC:	14.00
Df Residuals:	8	BIC:	14.60
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.0412	1.001	2.038	0.076	-0.268	4.350
Tobacco	1.0059	0.281	3.576	0.007	0.357	1.655

Omnibus:	2.542	Durbin-Watson:	1.975
Prob(Omnibus):	0.281	Jarque-Bera (JB):	0.904
Skew:	-0.014	Prob(JB):	0.636
Kurtosis:	1.527	Cond. No.	27.2

Regresja liniowa

Jeżeli oznaczymy przez n liczbę obserwacji, a k liczbę parametrów regresji/modelu (np. dla modelu liniowego z przykładu mamy k = 2), a $\hat{y}$ przewidywaną wartość modelu oraz $\bar{y}$ średnią z zaobserwowanych wartości, to:

- Model Degrees

$$DF_{mod} = k - 1$$

- Residuals Degrees of Freedom

$$DF_{res} = n - k$$

- Total Degrees of Freedom

$$DF_{tot} = DF_{tot} + DF_{res} = n - 1$$

Regresja liniowa

Współczynnik determinacji R^2 wyraża się wzorem:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{model}}{SS_{tot}}$$

The Adjusted R^2 Value jest modyfikacją R^2 biorącą pod uwagę karę za dużą liczbę parametrów w modelu.

OLS Regression Results

Dep. Variable:	Alcohol	R-squared:	0.615
Model:	OLS	Adj. R-squared:	0.567
Method:	Least Squares	F-statistic:	12.78
Date:	Fri, 15 Jan 2021	Prob (F-statistic):	0.00723
Time:	11:28:07	Log-Likelihood:	-4.9998
No. Observations:	10	AIC:	14.00
Df Residuals:	8	BIC:	14.60
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.0412	1.001	2.038	0.076	-0.268	4.350
Tobacco	1.0059	0.281	3.576	0.007	0.357	1.655

Omnibus:	2.542	Durbin-Watson:	1.975
Prob(Omnibus):	0.281	Jarque-Bera (JB):	0.904
Skew:	-0.014	Prob(JB):	0.636
Kurtosis:	1.527	Cond. No.	27.2

Regresja liniowa

The Adjusted R^2 Value określona jest wzorem:

$$1 - \bar{R}^2 = \frac{ResidualVariance}{TotalVariance}$$

gdzie Residual Variance to:

$$ResidualVariance = SS_{res}/DF_{res} = SS_{res}/(n-k)$$

Total Variance to:

$$TotalVariance = SS_{tot}/DF_{tot} = SS_{tot}/(n-1)$$

Regresja liniowa

F test dla regresji – sprawdza, czy chociaż jeden współczynnik modelu jest statystycznie istotnie różny od zera.

OLS Regression Results

Dep. Variable:	Alcohol	R-squared:	0.615
Model:	OLS	Adj. R-squared:	0.567
Method:	Least Squares	F-statistic:	12.78
Date:	Fri, 15 Jan 2021	Prob (F-statistic):	0.00723
Time:	11:28:07	Log-Likelihood:	-4.9998
No. Observations:	10	AIC:	14.00
Df Residuals:	8	BIC:	14.60
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.0412	1.001	2.038	0.076	-0.268	4.350
Tobacco	1.0059	0.281	3.576	0.007	0.357	1.655

Omnibus:	2.542	Durbin-Watson:	1.975
Prob(Omnibus):	0.281	Jarque-Bera (JB):	0.904
Skew:	-0.014	Prob(JB):	0.636
Kurtosis:	1.527	Cond. No.	27.2

Regresja liniowa

Dla modelu:

$$y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

Chcemy przetestować hipotezę:

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_n = 0$$
$$H_1\colon \beta_j \neq 0$$

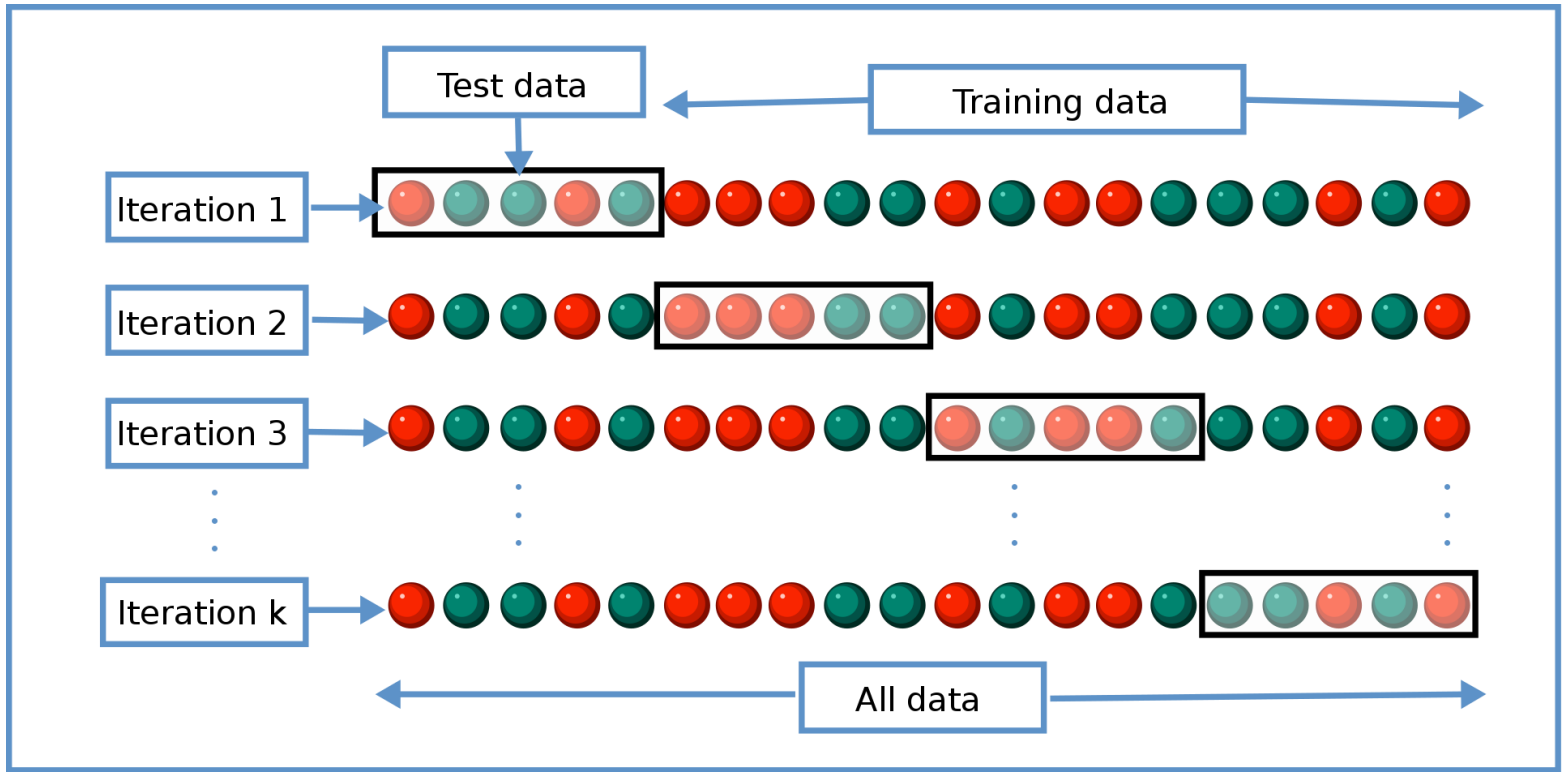
Dla co najmniej jednego j

Notebook – Regresja

Notebook – D11_Z01

Notebook – D11_Z02

K-fold cross validation

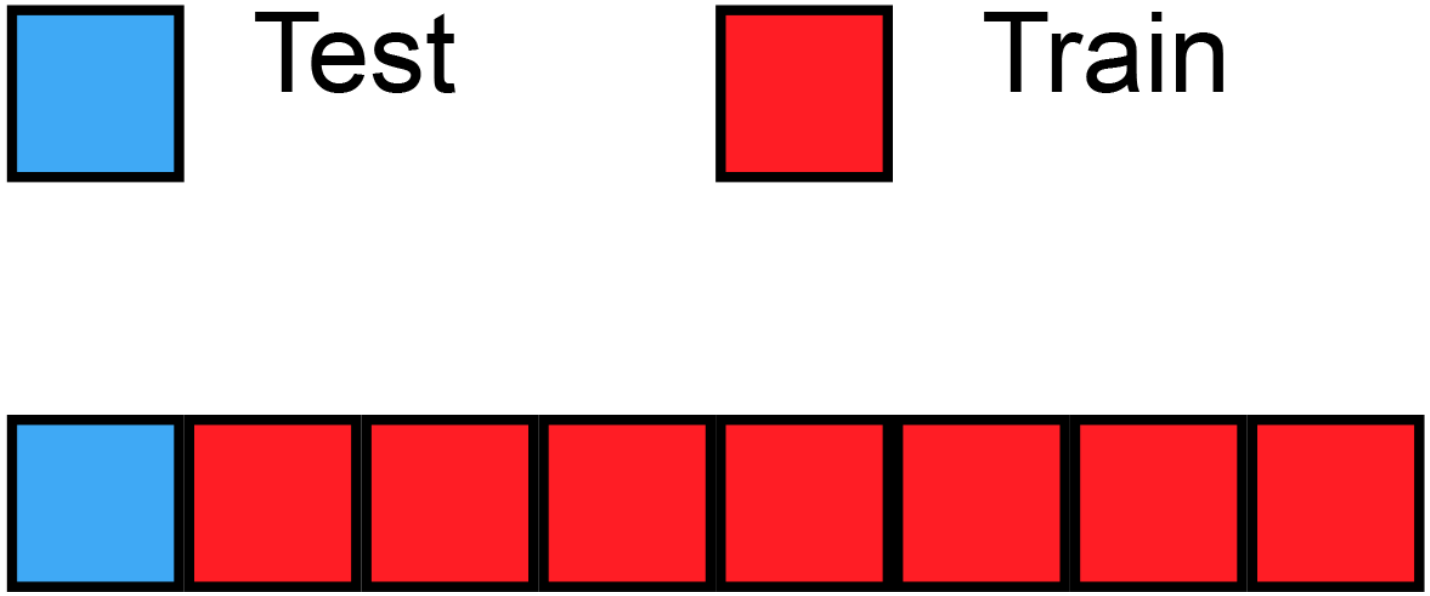



```
from sklearn.model_selection import KFold
kf = KFold(n_splits=10)
for train, test in kf.split(X):
    ...
```

https://en.wikipedia.org/wiki/Cross-validation_(statistics)#/media/File:K-fold_cross_validation_EN.svg

Leave-one-out

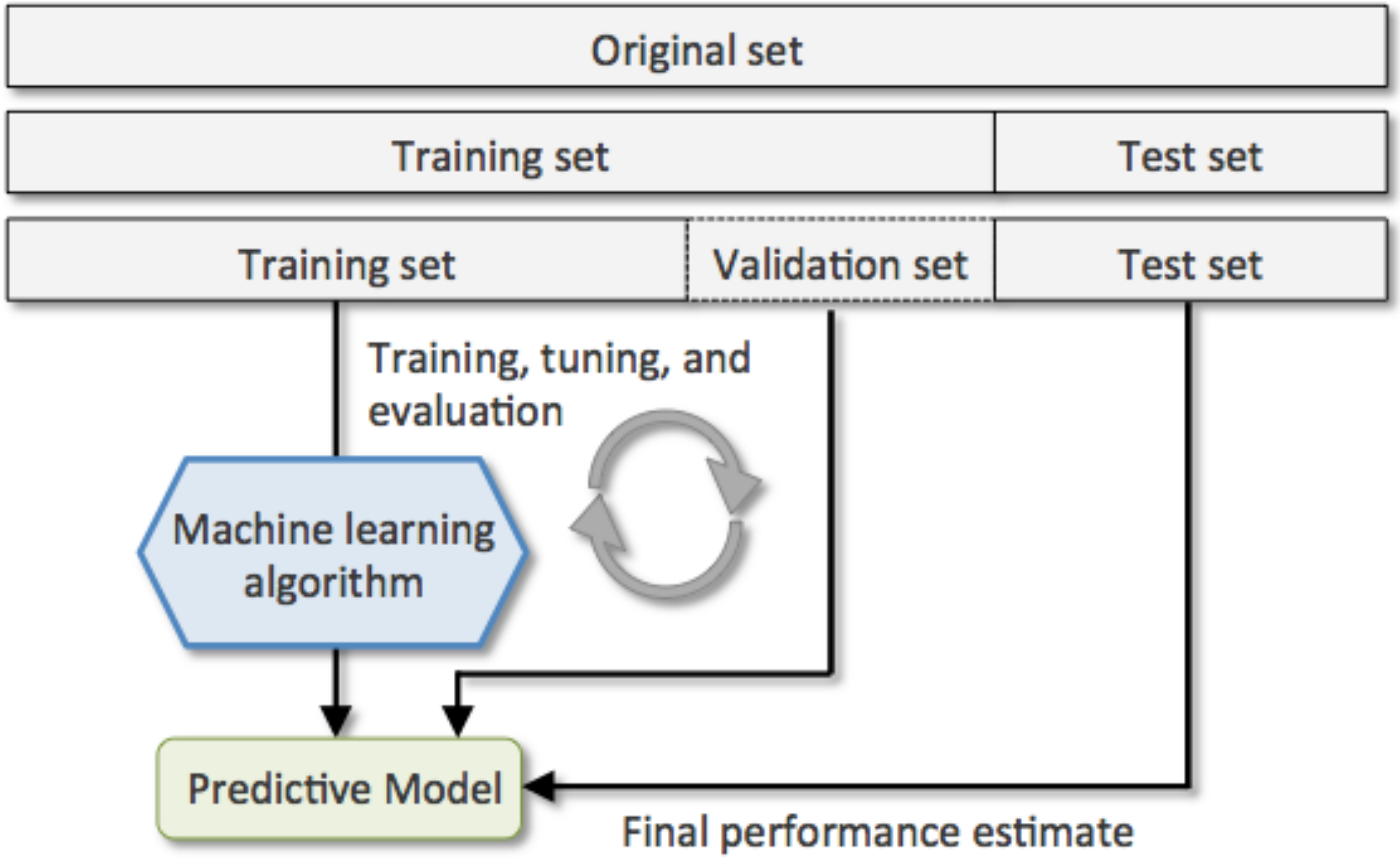
***n* = 8**




```
from sklearn.model_selection import LeaveOneOut
loo = LeaveOneOut()
for train_index, test_index in loo.split(X):
    ...
```

https://en.wikipedia.org/wiki/Cross-validation_(statistics)#/media/File:LOOCV.gif

Schemat testowania modeli



https://www.kaggle.com/robinteuwens/anomaly-detection-with-auto-encoders

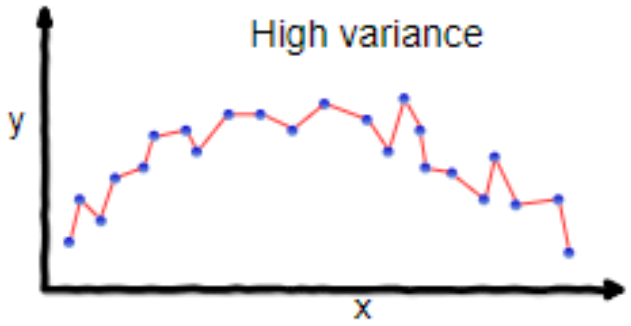
Bias-variance tradeoff

Na błąd modelu składają się trzy czynniki:

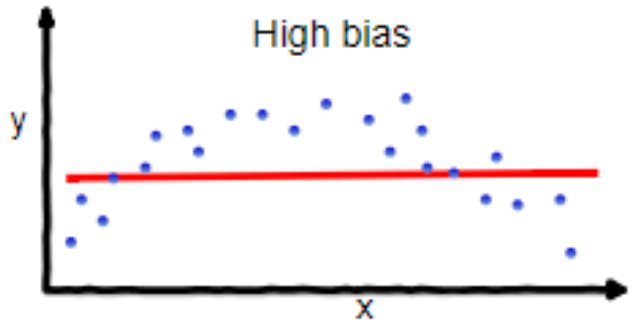
- **Bias** - wynika z błędnych założeń algorytmu uczenia się. Wysoki bias może spowodować, że algorytm przeoczy odpowiednie relacje między cechami a docelowymi wynikami (niedopasowanie - underfitting).
- **Variance** – wynika z wrażliwości na małe wahania w zbiorze uczącym. Wysoka wariancja może spowodować, że algorytm będzie modelował losowy szum w danych uczących, zamiast zamierzonych wyników (nadmierne dopasowanie - overfitting).
- **Noise** – szum zawarty w danych. Różnica między y przewidzianym a y prawdziwym.

$$\boldsymbol{E = E_b + E_v + E_n}$$

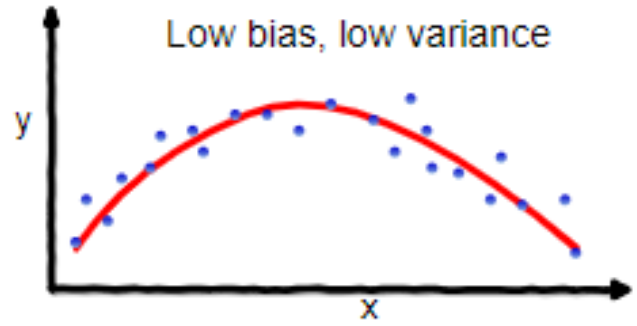
Bias-variance tradeoff



overfitting



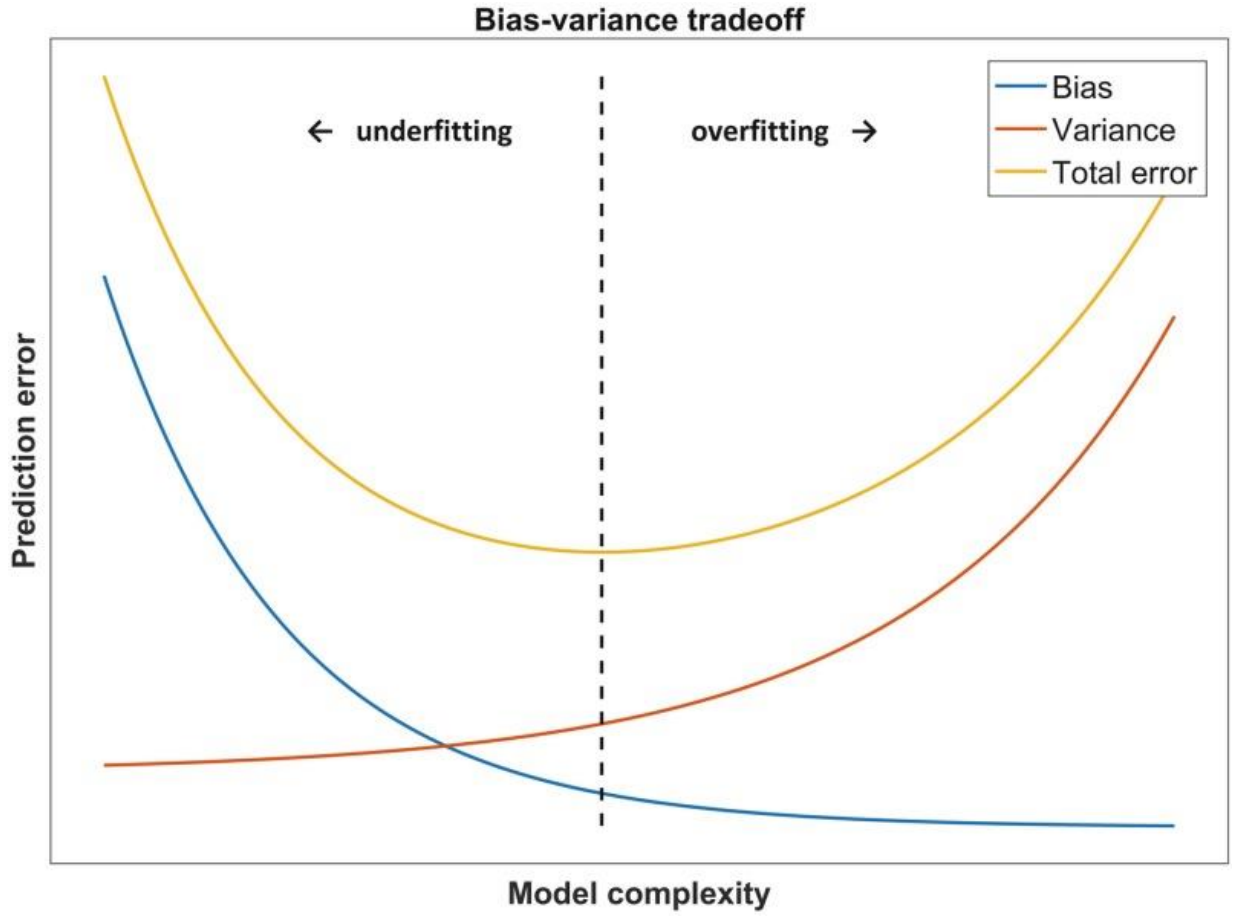
underfitting



Good balance

https://storybydata.com/wp-content/uploads/2019/02/1_9hPX9pAO3jqLrzt0IE3JzA.png

Bias-variance tradeoff



https://www.ncbi.nlm.nih.gov/books/NBK543534/figure/ch8.Fig3/

Notebook - Crossvalidation

Notebook – of_uf_train_test