Bootcamp Data Science
Zajęcia 3

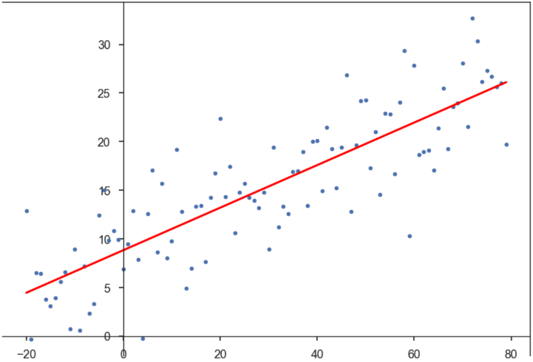
Przemysław Spurek

Regresja

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

Ogólny model regresji liniowej

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



Ogólny model regresji liniowej

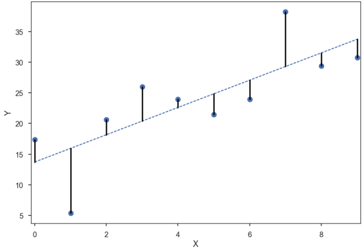
Na przykład, gdy szukamy linii najlepiej dopasowanej do danego zbioru danych:

$$(x_i, y_i)$$

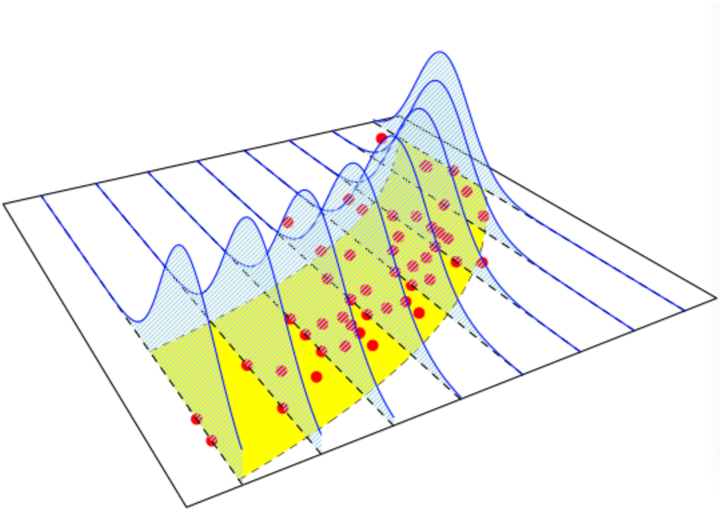
to tak naprawdę szukamy parametrów (a,b) które minimalizują błąd kwadratowy (squared residuals) ϵ_i w modelu:

$$y_i = a \cdot x_i + b + \epsilon_i$$

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



Ogólny model regresji liniowej



Ogólny model regresji liniowej

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

Ogólny model regresji liniowej

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

Zauważmy, że w przeciwieństwie do korelacji związek między x i y nie jest symetryczny: zakłada się, że wartości x są dokładnie znane, a zmienna y jest tylko przybliżeniem.

Simple Linear Regression

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

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

Taki model można zapisać w postaci

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

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

Quadratic Fit

Kwadratowe dopasowanie do danych jest dane modelem:

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

W postaci macierzowej mamy:

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

Quadratic Fit

Kwadratowe dopasowanie do danych jest dane modelem:

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

W postaci macierzowej mamy:

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

Uwaga

Zauważ, że nieznane parametry β_i pojawiają się liniowo, a składniki macierzy pojawiają się z kwadratami.

Quadratic Fit

- Zestaw danych ma wartości y_i, z których każda ma skojarzoną wartość modelową f_i (czasami również oznaczaną $\hat{y}_i$).
- 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.

Quadratic Fit

"Zmienność" zbioru możemy mierzyć różnymi miarami:

- 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 (równoważna wariancji próbki pomnożonej przez $(n-1)$).

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

Quadratic Fit

Dla modelu regresji liniowej mamy:

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

Przy powyższych oznaczeniach **współczynnik determinacji** *coefficient of determination* oznaczmy R^2:

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

Uwaga

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

Quadratic Fit

Dla regresji liniowej, współczynnik determinacji jest kwadratem współczynnika korelacji R. 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.

Uwaga

Zauważmy, że dla modeli ogólnych często pisze się R^2, podczas gdy dla prostej regresji liniowej r^2.

Oznaczenia

Jeśli mamy zmienną y i chcemy ją opisać za pomocą x, to możemy po prostu napisać:

$$y \sim x$$

Bardziej złożona sytuacja jest wtedy gdy y zależy od zmiennych x, a, b oraz $a \cdot b$:

$$y \sim x + a + b + a : b$$

Operator	Meaning
$\sim$	Separates the left-hand side from the right-hand side. If omitted, a formula is assumed right-hand side only
+	Combines terms on either side (set union)
−	Removes terms on the right from set of terms on the left (set difference)
*	$a * b$ is shorthand for the expansion $a + b + a : b$
/	a/b is shorthand for the expansion $a + a : b$. It is used when b is nested within a (e.g., states and counties)
:	Computes the interaction between terms on the left and right
**	Takes a set of terms on the left and an integer n on the right and computes the * of that set of terms with itself n times

Zapis macierzowy

Bardzo ogólna definicja modelu regresji jest następująca:

$$y = f(x, \epsilon).$$

W przypadku modelu regresji liniowej model może zostać zapisany jako:

$$y = X\beta + \epsilon$$

Zapis macierzowy

Dla danych w postaci:

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

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

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

gdzie ^{T} oznacza transpozycję, a $\mathbf{x}_i^T \beta$ oznacza iloczyn skalarny.
W notacji macierzowej:

$$y = X\beta + \epsilon.$$

Zapis macierzowy

W notacji macierzowej:

$$y = X\beta + \epsilon.$$

gdzie:

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

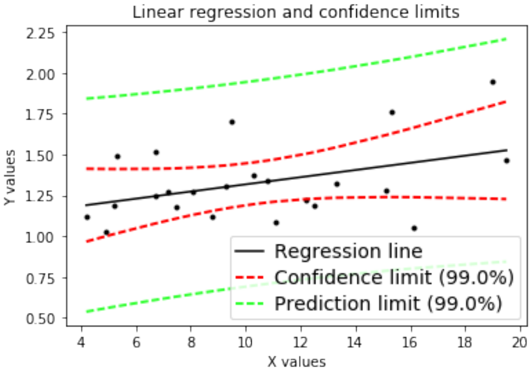
Przykład 1

W przypadku jednowymiarowych rozkładów, przedział ufności:

- oparty na odchyleniu standardowym wskazuje przedział, który powinien zawierać 95% danych,
- a błąd standardowy wskazuje przedział, który zawiera prawdziwą średnią z prawdopodobieństwem 95%.

Mamy więc dwa typy przedziałów ufności (jeden dla danych, a drugi dla odpowiednich parametrów) dla dopasowanej linii.

Przykład 1



Regresja, ze średnimi przedziałami ufności, jak również przewidywanymi danymi. Czerwona przerywana linia pokazuje ufności dla średniej, a zielona kropkowana linia przedział ufności dla przewidywanych danych.

Przykład 2

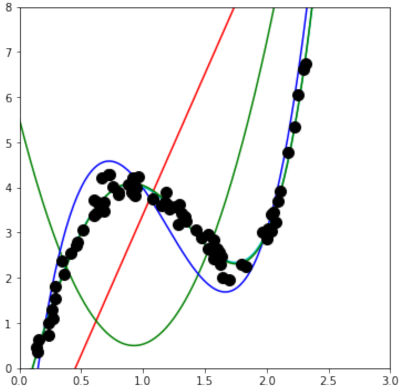
https://github.com/przem85/bootcamp/blob/master/statistics/D10_Z01.ipynb

Aby zobaczyć, jak różne modele mogą być użyte do oceny danego zbioru danych, spójrzmy na prosty przykład dopasowując:

- prostą,
- parabolę,
- $y = a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3 + b$,
- $y = a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3 + a_4 \cdot x^4 + b$,
- $y = a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3 + a_4 \cdot x^4 + a_5 \cdot x^5 + b$.

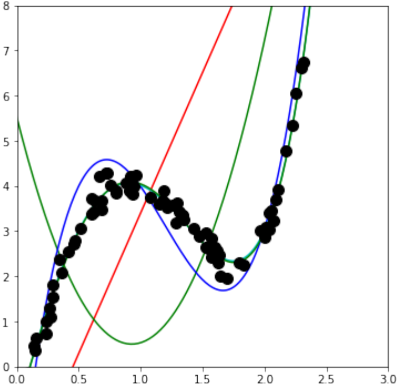
Przykład 2

Jak widzimy zarówno krzywa stopnia trzeciego, czwartego i piątego znajdują dobre dopasowanie.



Przykład 2

Jak widzimy zarówno krzywa stopnia trzeciego, czwartego i piątego znajdują dobre dopasowanie.



Które dopasowanie jest lepsze?

Przykład 2

- 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.

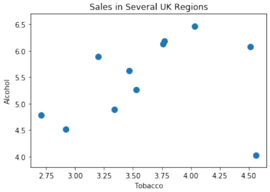
```
                 Results: Ordinary least squares
=================================================================
Model:              OLS              Adj. R-squared:     0.983
Dependent Variable: y                AIC:                909.6344
Date:               2015-06-27 13:50 BIC:                914.8447
No. Observations:   100              Log-Likelihood:     -452.82
Df Model:           1                F-statistic:        5818.
Df Residuals:       98               Prob (F-statistic): 4.46e-89
R-squared:          0.983            Scale:              512.18
-------------------------------------------------------------------
         Coef.     Std.Err.      t       P>|t|     [0.025    0.975]
-------------------------------------------------------------------
const   100.4163     4.4925   22.3519   0.0000   91.5010   109.3316
x1        5.9802     0.0784   76.2769   0.0000    5.8246     6.1358
-------------------------------------------------------------------
Omnibus:               10.925       Durbin-Watson:          0.131
Prob(Omnibus):         0.004        Jarque-Bera (JB):       6.718
Skew:                  0.476        Prob(JB):               0.035
Kurtosis:              2.160        Condition No.:          114
=================================================================
```

Przykład 3

https://github.com/przem85/bootcamp/blob/master/statistics/D10_Z02.ipynb

Najpierw zajmiemy się małym zbiorem danych z biblioteki DASL dotyczące korelacji między zakupem wyrobów tytoniowych i alkoholu w różnych regionach Wielkiej Brytanii.
Użyjemy `df[:-1]` aby usunąć ostatni element, który możemy traktować jako element odstający.

```
result = sm.ols('Alcohol ~ Tobacco', df[:-1]).fit()
print(result.summary())
```



Przykład 3

- 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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3

- 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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3

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:

- (Corrected) Model Degrees $DF_{mod} = k - 1$
- Residuals Degrees of Freedom $DF_{res} = n - k$
- Total Degrees of Freedom $(DF_{mod} + DF_{res} = DF_{tot})$
  $DF_{tot} = n - 1$
- Model Mean of Squares

$$MS_{mod} = SS_{mod}/DF_{mod}$$

- Residuals Mean of Squares (jest estymatorem nieobciążonym σ^2)

$$MS_{res} = SS_{res}/DF_{res}$$

- Total Mean of Squares

$$MS_{tot} = SS_{tot}/DF_{tot}$$

Przykład 3

```
                            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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3 – The R2 Value

The R^2 Value wyraża się wzorem:

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

The Adjusted $\bar{R}^2$ Value jest modyfikacją R^2 biorącą pod uwagę karę za dużą liczbę parametrów w modelu p:

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

gdzie (Sample) Residual Variance to:

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

(Sample) Total Variance to:

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

Przykład 3

$$\bar{R}^2 - 1 = \frac{SS_{res}}{SS_{tot}}\frac{n-1}{n-k} = 1 - (1 - R^2)\frac{n-1}{n-k}$$

Przykład 3

```
                            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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3 – The F-Test for regression.

W przypadku modelu regresji:

$$Y_i = \alpha + \beta_1 X_{1j} + \ldots + \beta_n X_{nj} + \epsilon_i = \alpha + \sum_{i=1}^{n} \beta_i X_{ij} + \epsilon_j.$$

Chcemy przetestować hipotezę:

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_n = 0$$

$$H_1 : \beta_j \neq 0 \text{ dla co najmniej jednego } j$$

Przykład 3 – The F-Test for regression.

Pamiętamy, że jeżeli zmienne losowe $t_1, t_2, \ldots, t_m$ są niezależne o rozkładzie normalnym $N(0, \sigma^2)$, to:

$$\sum_{i=1}^{m} \frac{t_i^2}{\sigma^2}$$

ma rozkład chi kwadrat z m stopniami swobody.
W konsekwencji, jeżeli hipoteza zerowa jest prawdziwa, to:

- SS_{res}/σ^2 ma rokład χ^2 z DF_{res} stopniami swobody,
- SS_{mod}/σ^2 ma rokład χ^2 z DF_{mod} stopniami swobody,
- SS_{res} oraz SS_{mod} są niezależne.

Przykład 3 – The F-Test for regression.

Jeżeli zmienna losowa U ma rozkład χ^2 z n stopniami swobody oraz V jest zmienną losową o rozkładzie χ^2 z m stopniami swobody, to:

$$F = \frac{U/n}{V/m}$$

ma rozkład F z (n, m) stopniami swobody.
Jeżeli hipoteza H_0 jest prawdziwa, to:

$$F = \frac{(SS_{mod}/\sigma^2)/DF_{mod}}{(SS_{res}/\sigma^2)/DF_{res}} = \frac{SS_{mod}/DF_{mod}}{SS_{res}/DF_{res}}$$

ma rozkład z (DF_{mod}, DF_{res}) stopniami swobody i jest niezależna od σ.

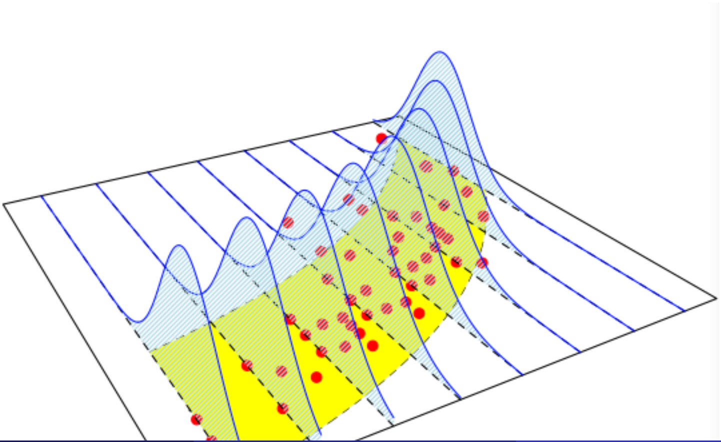
Przykład 3

```
                            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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3 – Log-Likelihood Function

Dla klasycznej regresji liniowej mamy:

$$\epsilon = y_i - \sum_{k=1}^{n} \beta_k x_{ik} = y_i - \hat{y}_i \sim N(0, \sigma)$$



Przykład 3 – Log-Likelihood Function

W konsekwencji wiemy, że:

$$p(\epsilon_i) = f(\frac{y_i - \hat{y}_i}{\sigma})$$

gdzie f jest gęstością standardowego rozkładu normalnego.
Zakładając niezależność między błędami mamy funkcję wiarygodności:

$$l_{total} = \prod_{i=1}^{n} p(\epsilon_i).$$

Logarytmiczną funkcją wiarygodności (Log Likelihood function) nazywamy:

$$L = \ln(l) = \ln\left(\prod_{i=1}^{n} f(\frac{y_i - \hat{y}_i}{\sigma})\right).$$

Przykład 3 – Log-Likelihood Function

Czyli mamy:

$$L = \ln(l) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \hat{y}_i)^2}{2\sigma^2}\right)\right) =$$

$$= \sum_{i=1}^{n} \left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \left(\frac{(y_i - \hat{y}_i)^2}{2\sigma^2}\right)\right) =$$

Można pokazać, że estymatorem największej wiarygodności wariancji jest:

$$\sigma^2 = \frac{SS_{res}}{n}.$$

Przykład 3 – AIC and BIC

Aby ocenić jakość modelu, najpierw należy wizualnie sprawdzić błędy. Ponadto można również skorzystać z kilku liczbowych kryteriów oceny jakości modelu statystycznego. Te kryteria reprezentują różne podejścia do oceny modelu.

AIC and BIC działają podobnie do R i $\bar{R}$.

Przykład 3 – AIC and BIC

Innymi powszechnie spotykanymi kryteriami jest Akaike Information Criterion (AIC) oraz Bayesian Information Criterion (BIC), które opierają się na funkcji wiarygodności.

Uwaga

Obie miary wprowadzają kary za złożoność modelu, ale AIC kara mniej za złożoność niż BIC.

Przykład 3 – AIC and BIC

Kryterium Informacyjne Akaike (AIC):

$$AIC = 2 \cdot k - 2 \cdot \ln(L)$$

Kryterium Informacyjne Bayesian (BIC):

$$BIC = k \cdot \ln(N) - 2 \cdot \ln(L)$$

gdzie, N jest liczbą obserwacji, k jest liczbą parametrów, a L jest funkcją wiarygodności.

Uwaga

Powinniśmy wybrać model o niższej wartości AIC lub BIC.

Przykład 3

```
                            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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3 – błąd standardowy

Parametr β możemy łatwo otrzymać wyznaczając macierz odwrotną do X

$$\beta = (X^T X)^{-1} X^T y.$$

W celu uzyskania odchylenia standardowego współczynników obliczymy macierz kowariancji dla β:

$$C = \sigma^2 (X^T X)^{-1}, \text{ gdzie } \sigma^2 \text{ jest wariancją } \hat{y}_i.$$

Błąd standardowy jest dany przez pierwiastki wartości na diagonali macierzy kowariancji.

Przykład 3 – t-Statistic

- Używamy testu t-Studenta, aby przetestować hipotezę zerową mówiącą, że: *współczynnik wynosi zero, co sugeruje, że dany predyktor nie ma znaczącego wpływu na zmienną objaśnianą.*
- Alternatywna hipoteza mówi, że współczynnik predykcyjny ma wpływ na zmienną objaśnianą.
- Podczas testowania ustalamy pewien próg $\alpha = 0.05$ lub $\alpha = 0.001$.
- Gdy $P(T \geq |y|) < \alpha$, wtedy odrzucamy hipotezę zerową.
- Test t-Studenta zazwyczaj pozwala nam ocenić znaczenie różnych predyktorów, zakładając, że *błąd modelu opisywany jest rozkładem normalnym (wokół zera)*.
- Jeśli błąd nie zachowuje się w ten sposób, to najlepiej byłoby spróbować zmodyfikować model.

Przykład 3 – t-Statistic

Statystyka t jest dana wzorem:

$$t_i = \beta / SE_{i,i}.$$

Gdy mamy statystykę t, możemy obliczyć p-value.

Przykład 3 – przedział ufności

- Przedział ufności jest zbudowany za pomocą standardowego błędu, p-value oraz testu t-Studenta z $N - k$ stopniami swobody, gdzie N jest liczbą obserwacji, k jest liczbą parametrów modelu (to znaczy liczbą zmiennych objaśniających).
- Przedział ufności, to zakres wartości, w jakich spodziewamy się znaleźć parametr.
- Mniejszy przedział ufności wskazuje, że jesteśmy pewni co do wartości szacowanego współczynnika.
- Większy przedział ufności wskazuje na większą niepewność.

Przykład 3 – przedział ufności

Przedział ufności dany jest wzorem:

$$CI = \beta_i \pm z \cdot SE_{i,i}.$$

Ponieważ, β jest jednym z estymowanych współczynników, to z jest krytyczną wartością dla której statystyka t-Studenta przyjmuje wartość mniejszą niż zadany poziom, a $SE_{i,i}$ jest standardowym błędem. Wartość krytyczna jest obliczana przy użyciu odwrotnej funkcji do dystrybuanty.

Przykład 3

```
                            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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3 – Skewness and Kurtosis

Skośność i kurtoza odnoszą się do kształtu rozkładu. Skośność jest miarą asymetrii rozkładu, a kurtoza jest miarą jego krzywizny (grube ogony):

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)^{3/2}}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)^{2}}$$

Kurtozę definiuje się jako $K - 3$ gdy rozkłady normalne mają kurtozę równą 3.

Przykład 3 – Skewness and Kurtosis

```
d = Y - result.fittedvalues
S = np.mean( d**3.0 ) / np.mean( d**2.0 )**(3.0/2.0)
# equivalent to:
# S = stats.skew(result.resid, bias=True)
K = np.mean( d**4.0 ) / np.mean( d**2.0 )**(4.0/2.0)
# equivalent to:
# K = stats.kurtosis(result.resid, fisher=False,
# bias=True)
print('Skewness: {:.3f}, Kurtosis: {:.3f}'.format(S,K))
```

```
Skewness:  -0.014, Kurtosis:  1.527
```

Przykład 3

```
                            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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3 – Omnibus Test

Omnibus Test wykorzystuje skośność i kurtozę, aby przetestować hipotezę zerową mówiącą, że rozkład błędów (residuals) jest normalny.
Jeśli otrzymamy bardzo małą p-value dla Omnibus Test, wówczas błędy nie pochodzą z rozkładu normalnego.

```
(K2, p) = stats.normaltest(result.resid)
print('Omnibus: {0:.3f}, p = {1:.3f}'.format(K2, p))
```

```
Omnibus:  2.542, p = 0.281
```

Przykład 3

```
                            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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3 – Durbin–Watson

Durbin-Watson jest testem używanym do wykrywania obecności autokorelacji (relacji pomiędzy wartościami oddzielonymi od siebie określonym czasem opóźnienia) w błędach. U nas opóźnienie jest jedno:

$$DW = \frac{\sum_{i=1}^{N}((y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1}))^2}{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

```
DW = np.sum( np.diff( result.resid.values )**2.0 )
        / result.ssr
print('Durbin-Watson: {:.5f}'.format( DW ))
```

```
Durbin-Watson:  1.97535
```

Jeśli statystyka Durbin-Watson jest znacznie mniejsza od 2, to dane są skorelowane dodatnio. W zasadzie, jeśli statystyka Durbin-Watsona jest mniejsza niż 1.0, to należy zastanowić się nad zmianą modelu.

Przykład 3

```
                            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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3 – Jarque–Bera Test

Test Jarque-Bera to kolejny test, który uwzględnia skośność (S) i kurtozę (K). Hipoteza zerowa mówi, że rozkład jest normalny w sensie zerowej skośności i kurtozy.
Niestety, przy małych próbkach, test Jarque-Bera jest podatny na odrzucenie hipotezy zerowej (że rozkład jest normalny) gdy nie powinien.

$$JB = \frac{N}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

Statystyka Jarque–Bera ma rozkład chi kwadrat z dwoma stopniami swobody.

Przykład 3

```
                            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, 28 Apr 2017   Prob (F-statistic):            0.00723
Time:                        15:22:23   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|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
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
==============================================================================
```

Przykład 3 – Condition Number

Condition Number określa czułość wyjścia funkcji na jego wejście. Gdy dwie zmienne objaśniające są wysoce skorelowane mała zmiana w danych lub modelu drastycznie zmienia wyniki. W idealnej sytuacji podobne modele powinny dawać podobne wyniki.
Condition Number obliczamy wyznaczając wartości własne $X^T X$ (w tym wektora stałych), a następnie biorąc pierwiastek ze stosunku największej wartości własnej do najmniejszej.
Jeśli Condition Number przekracza 30, to model regresji powinien zostać zmieniony.

```
X = np.matrix(X)
EV = np.linalg.eig( X * X.T )
CN = np.sqrt( EV[0].max() / EV[0].min() )
print('Condition No.: {:.5f}'.format( CN ))
```

```
Condition No.:  27.22887
```

Przykład 4 – Wartości odstające

Proszę wykonać regresją na całym zbiorze danych:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Alcohol   R-squared:                       0.050
Model:                            OLS   Adj. R-squared:                 -0.056
Method:                 Least Squares   F-statistic:                    0.4735
Date:                Sat, 29 Apr 2017   Prob (F-statistic):              0.509
Time:                        09:47:30   Log-Likelihood:                -12.317
No. Observations:                  11   AIC:                             28.63
Df Residuals:                       9   BIC:                             29.43
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Tobacco        0.3019      0.439      0.688      0.509        -0.691     1.295
Ones           4.3512      1.607      2.708      0.024         0.717     7.986
==============================================================================
Omnibus:                        3.123   Durbin-Watson:                   1.655
Prob(Omnibus):                  0.210   Jarque-Bera (JB):                1.397
Skew:                          -0.873   Prob(JB):                        0.497
Kurtosis:                       3.022   Cond. No.                         25.5
==============================================================================
```

Założenia regresji liniowej

Założenia regresji liniowej:

- zależność jest liniowa,
- brak znaczących obserwacji odstających,
- homoscedastyczność – wariancja reszt składnika losowego jest taka sama dla wszystkich obserwacji,
- reszty mają rozkład zbliżony do rozkładu normalnego.
  Regresja wielokrotna:
  - liczba obserwacji musi być większa, bądź równa liczbie parametrów,
  - brak współliniowości parametrów,
  - nie występuje autokorelacja reszt.

Zadania

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z16.ipynb

Zadania

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z16.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z15.ipynb

Zadania

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z16.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z15.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z21.ipynb

Zadania

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z16.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z15.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z21.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z23.ipynb

Zadania

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z16.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z15.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z21.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z23.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D11_Z24.ipynb