# Model regresji liniowej

Martyna Śpiewak
Bootcamp Data Science

$Y$ - odpowiedź, zmienna zależna, zmienna wyjaśniana;

$X_1, X_2, \ldots, X_p$ - predyktory, zmienne niezależne, zmienne wyjaśniające, cechy;

$$Y = f(X) + \varepsilon,$$

gdzie $f$ jest nieznaną funkcją predyktorów $X_1, X_2, \ldots, X_p$ oraz $\varepsilon$ to błąd losowy.

Istnieją dwa główne powody, dla których chcemy oszacować postać $f$:

- **predykcja** (ang. *prediction*), oraz
- **wnioskowanie** (ang. *inference*).

# Predykcja

Niech $\hat{Y}$ oznacza przewidzianą wartość zmiennej zależnej $Y$, zaś $\hat{f}$, to założona/wyestymowana postać nieznanej funkcji $f$:

$$\hat{Y} = \hat{f}(X).$$

Dokładność prognozy $\hat{Y}$ zależy od dwóch wielkości:

- **błędu redukowalnego** (postać $\hat{f}$ zazwyczaj nie jest idealnym oszacowaniem funkcji $f$, ta niedokładność wprowadza pewien błąd do predykcji);
- **błędu nieredukowalnego** (wartość $Y$ zależy od pewnego nieznanego błędu $\varepsilon$).

$$\begin{aligned}\mathbb{E}(Y - \hat{Y})^2 &= \mathbb{E}[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{błąd redukowalny}} + \underbrace{\mathrm{Var}(\varepsilon)}_{\text{błąd nieredukowalny}}.\end{aligned}$$

# Wnioskowanie

Często oprócz predykcji nieznanych wartości *Y*, jesteśmy również zainteresowani zrozumieniem relacji pomiędzy zmienną zależną a zmiennymi niezależnymi.

Wówczas jesteśmy zainteresowani odpowiedzią na następujące pytania:

- **Które zmienne niezależne są powiązane ze zmienną odpowiedzi?** Ważny elementem analizy jest określenie podzbioru dostępnych predyktorów, które istotnie są związane ze zmienną odpowiedzi *Y*.

# Wnioskowanie

Często oprócz predykcji nieznanych wartości *Y*, jesteśmy również zainteresowani zrozumieniem relacji pomiędzy zmienną zależną a zmiennymi niezależnymi.

Wówczas jesteśmy zainteresowani odpowiedzią na następujące pytania:

- **Które zmienne niezależne są powiązane ze zmienną odpowiedzi?** Ważny elementem analizy jest określenie podzbioru dostępnych predyktorów, które istotnie są związane ze zmienną odpowiedzi *Y*.
- **Jaki jest związek między zmienną odpowiedzi a każdym predyktorem?** Niektóre zmienne niezależne mogą mieć pozytywny wpływ na zmienną odpowiedzi (np. wzrost wartości zmiennej *X* wiąże się ze wzrostem wartości zmiennej *Y*), inne predyktory mogą mieć odwrotny związek.

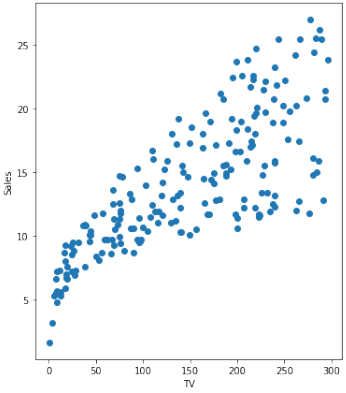
- **Czy związek między zależną $Y$ a każdą zmienną odpowiedzi $X_i$ można przedstawić jako liniowy? Czy też związek jest bardziej złożony?** Historycznie wiele metod zakładała liniową postać funkcji $f$. W niektórych sytuacjach takie założenie jest uzasadnione. Jednakże, w większości przypadków prawdziwy związek jest bardziej skomplikowany, wówczas model liniowy może być niewystarczający do odpowiedniego określenia związku między zmiennymi wejściowymi i wyjściowymi.

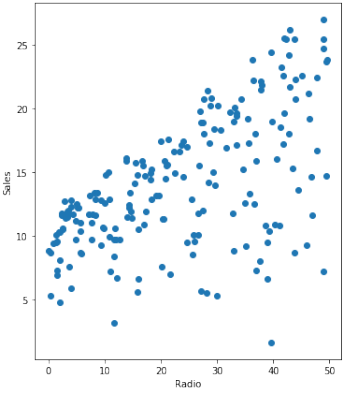
# Zbiór danych Advertising

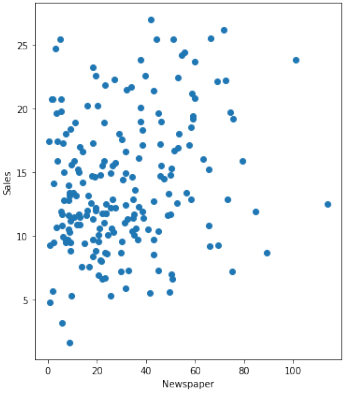
Zbiór danych **Advertising** zawiera dane ze sprzedaży reklamy pewnego produktu na 200 różnych rynkach wraz z budżetami reklamowymi tego produktu w każdym z nich dla trzech różnych typów mediów: telewizji, radia i gazety.

| | TV | Radio | Newspaper | Sales |
|---|---|---|---|---|
| 0 | 230.1 | 37.8 | 69.2 | 22.1 |
| 1 | 44.5 | 39.3 | 45.1 | 10.4 |
| 2 | 17.2 | 45.9 | 69.3 | 9.3 |
| 3 | 151.5 | 41.3 | 58.5 | 18.5 |
| 4 | 180.8 | 10.8 | 58.4 | 12.9 |

# Zbiór danych Advertising



# Zbiór danych Advertising

1. Czy istnieje związek między budżetem przeznaczonym na reklamę a wielkością sprzedaży?
2. Jak silny jest związek między budżetem a wielkością sprzedaży?
3. Które rodzaje mediów wpływają na sprzedaż?
4. Jak dokładnie możemy przewidzieć wpływ każdej formy mediów na wielkość sprzedaży?
5. Jak dokładnie możemy przewidzieć wielkość przyszłej sprzedaży?
6. Czy związek między zmienną odpowiedzi a zmiennymi niezależnymi jest liniowy?

# Model regresji prostej

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

gdzie $\varepsilon \sim \mathcal{N}(0, \sigma)$.

W praktyce, współczynniki $\beta_0$ i $\beta_1$ są nieznane.

**Cel:** Przy użyciu par $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ odpowiadającym pomiarom, odpowiednio, zmiennej $X$ i $Y$, wyznaczyć współczynniki $b_0, b_1$ tak, aby

$$y_i \approx b_0 + b_1 x_i$$

**Zapis:**

$\hat{y}_i = b_0 + b_1 x_i$ – wartość prognozowana $Y$ na podstawie $i$-tej wartości $X$

$e_i = y_i - \hat{y}_i$ – $i$-te rezyduum (wartość resztowa)

# Model regresji prostej — zbiór `Advertising`

Niech

- $X$ będzie zmienną niezależną opisującą wysokość budżetu przeznaczonego na reklamę w telewizji (TV),
- $Y$ będzie zmienną zależną opisującą wysokość sprzedaży pewnego produktu (`Sales`).

Zakładamy, że prawdziwy jest związek:

$$\texttt{Sales} \approx b_0 + b_1 \cdot \text{TV}.$$

# Jak wyznaczyć $b_0$ i $b_1$?

**Suma kwadratów błędów** (ang. *residual sum of squares*):

$$\begin{aligned}\text{RSS} &= e_1^2 + e_2^2 + \ldots + e_n^2 \\ &= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2\end{aligned}$$

**Funkcja kryterialna:**

$$\begin{aligned}(b_0, b_1) &= \arg\min_{(b_0,b_1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \arg\min_{(b_0,b_1)} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2\end{aligned}$$

**Metoda najmniejszych kwadratów** wyznacza $b_0, b_1$ minimalizując RSS, tj.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$
$$b_0 = \overline{y} - b_1\overline{x}$$

gdzie $\overline{x} = \frac{1}{n}x_i$ oraz $\overline{y} = \frac{1}{n}y_i$

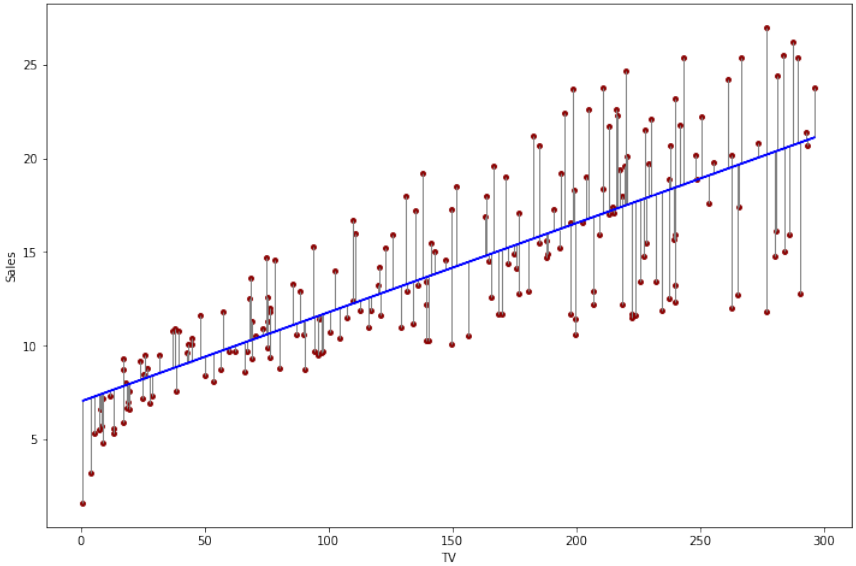
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.612 |
| Model: | OLS | Adj. R-squared: | 0.610 |
| Method: | Least Squares | F-statistic: | 312.1 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.47e-42 |
| Time: | 11:50:53 | Log-Likelihood: | -519.05 |
| No. Observations: | 200 | AIC: | 1042. |
| Df Residuals: | 198 | BIC: | 1049. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 7.0326 | 0.458 | 15.360 | 0.000 | 6.130 | 7.935 |
| TV | 0.0475 | 0.003 | 17.668 | 0.000 | 0.042 | 0.053 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.531 | Durbin-Watson: | 1.935 |
| Prob(Omnibus): | 0.767 | Jarque-Bera (JB): | 0.669 |
| Skew: | -0.089 | Prob(JB): | 0.716 |
| Kurtosis: | 2.779 | Cond. No. | 338. |

# Model regresji prostej — zbiór `Advertising`



# Stopnie swobody

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.612 |
| Model: | OLS | Adj. R-squared: | 0.610 |
| Method: | Least Squares | F-statistic: | 312.1 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.47e-42 |
| Time: | 11:50:53 | Log-Likelihood: | -519.05 |
| No. Observations: | 200 | AIC: | 1042. |
| Df Residuals: | 198 | BIC: | 1049. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 7.0326 | 0.458 | 15.360 | 0.000 | 6.130 | 7.935 |
| TV | 0.0475 | 0.003 | 17.668 | 0.000 | 0.042 | 0.053 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.531 | Durbin-Watson: | 1.935 |
| Prob(Omnibus): | 0.767 | Jarque-Bera (JB): | 0.669 |
| Skew: | -0.089 | Prob(JB): | 0.716 |
| Kurtosis: | 2.779 | Cond. No. | 338. |

## Stopnie swobody

`Df Model` — liczba stopni swobody modelu, wyrażona jako

$$p,$$

`Df Residuals` — liczba stopni swobody, wyrażona jako

$$n - p - 1$$

gdzie $n$ oznacza liczbę obserwacji w modelu, zaś $p$ to liczba predyktorów w modelu.

# Współczynnik determinacji $R^2$

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.612 |
| Model: | OLS | Adj. R-squared: | 0.610 |
| Method: | Least Squares | F-statistic: | 312.1 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.47e-42 |
| Time: | 11:50:53 | Log-Likelihood: | -519.05 |
| No. Observations: | 200 | AIC: | 1042. |
| Df Residuals: | 198 | BIC: | 1049. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 7.0326 | 0.458 | 15.360 | 0.000 | 6.130 | 7.935 |
| TV | 0.0475 | 0.003 | 17.668 | 0.000 | 0.042 | 0.053 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.531 | Durbin-Watson: | 1.935 |
| Prob(Omnibus): | 0.767 | Jarque-Bera (JB): | 0.669 |
| Skew: | -0.089 | Prob(JB): | 0.716 |
| Kurtosis: | 2.779 | Cond. No. | 338. |

## Współczynnik determinacji $R^2$

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

gdzie

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2,$$

to **całkowita suma kwadratów** (ang. *total sum of squares*).

1. $0 \leq R^2 \leq 1$
2. TSS — mierzy zmienność zmiennej $Y$ przed zastosowaniem regresji
3. RSS — mierzy wielkość zmienności, która jest niewyjaśniana przez model regresji
4. TSS − RSS — mierzy wielkość zmienności, która jest wyjaśniana przez model regresji
5. $R^2$ — mierzy stosunek zmienności $Y$, która może być wyjaśniona przez $X$

# Funkcja wiarogodności

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.612 |
| Model: | OLS | Adj. R-squared: | 0.610 |
| Method: | Least Squares | F-statistic: | 312.1 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.47e-42 |
| Time: | 11:50:53 | Log-Likelihood: | -519.05 |
| No. Observations: | 200 | AIC: | 1042. |
| Df Residuals: | 198 | BIC: | 1049. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 7.0326 | 0.458 | 15.360 | 0.000 | 6.130 | 7.935 |
| TV | 0.0475 | 0.003 | 17.668 | 0.000 | 0.042 | 0.053 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.531 | Durbin-Watson: | 1.935 |
| Prob(Omnibus): | 0.767 | Jarque-Bera (JB): | 0.669 |
| Skew: | -0.089 | Prob(JB): | 0.716 |
| Kurtosis: | 2.779 | Cond. No. | 338. |

# Funkcja wiarogodności

Zgodnie z założenia regresji

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma).$$

wówczas funkcja wiarogodności ma postać

$$L(y_1, y_2 \ldots, y_n, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\}.$$

Mamy

$$l(y_1, y_2, \ldots, y_n, \sigma) = \ln L = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

# Funkcja wiarogodności

Można pokazać, że wartości $b_0$ i $b_1$ maksymalizujące funkcję $l$, to:

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad \hat{\beta}_0 = b_0 = \overline{y} - b_1\overline{x}.$$

Wówczas szukając postaci ENW dla $\sigma^2$ postępujemy następująco:

1. $\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$;
2. $\frac{\partial l}{\partial \sigma} = 0 \implies \sigma_0^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
3. $\frac{\partial^2 l}{\partial \sigma^2}\Big|_{\sigma^2=\sigma_0^2} = -\frac{2n}{\sigma_0^2} < 0$.

Estymatorem największej wiarogodności wariancji $\sigma^2$ jest:

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n}.$$

# AIC i BIC

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.612 |
| Model: | OLS | Adj. R-squared: | 0.610 |
| Method: | Least Squares | F-statistic: | 312.1 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.47e-42 |
| Time: | 11:50:53 | Log-Likelihood: | -519.05 |
| No. Observations: | 200 | AIC: | 1042. |
| Df Residuals: | 198 | BIC: | 1049. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 7.0326 | 0.458 | 15.360 | 0.000 | 6.130 | 7.935 |
| TV | 0.0475 | 0.003 | 17.668 | 0.000 | 0.042 | 0.053 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.531 | Durbin-Watson: | 1.935 |
| Prob(Omnibus): | 0.767 | Jarque-Bera (JB): | 0.669 |
| Skew: | -0.089 | Prob(JB): | 0.716 |
| Kurtosis: | 2.779 | Cond. No. | 338. |

**Kryterium Akaike** (ang. *Akaike Information Criterion*):

$$\mathrm{AIC} = -2\ln L + 2p$$

**Kryterium Schwarza** (ang. *Bayesian Information Criterion*):

$$\mathrm{BIC} = -2\ln L + \ln(n)p,$$

gdzie $p$ jest liczbą parametrów w modelu, $n$ jest liczbą obserwacji, a $L$ jest funkcją wiarygodności.

# Współczynniki regresji

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.612 |
| Model: | OLS | Adj. R-squared: | 0.610 |
| Method: | Least Squares | F-statistic: | 312.1 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.47e-42 |
| Time: | 11:50:53 | Log-Likelihood: | -519.05 |
| No. Observations: | 200 | AIC: | 1042. |
| Df Residuals: | 198 | BIC: | 1049. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 7.0326 | 0.458 | 15.360 | 0.000 | 6.130 | 7.935 |
| TV | 0.0475 | 0.003 | 17.668 | 0.000 | 0.042 | 0.053 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.531 | Durbin-Watson: | 1.935 |
| Prob(Omnibus): | 0.767 | Jarque-Bera (JB): | 0.669 |
| Skew: | -0.089 | Prob(JB): | 0.716 |
| Kurtosis: | 2.779 | Cond. No. | 338. |

# Odchylenie standardowe estymatorów współczynników $\beta_0$ i $\beta_1$

$$\mathrm{SE}(\hat{\beta}_0) = \sigma^2 \Big[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \Big],$$

$$\mathrm{SE}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

gdzie $\sigma^2 = \mathrm{Var}(\varepsilon)$.

95% przedziały ufności dla parametrów regresji:

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1), \quad \hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0).$$

# Istotność współczynników regresji

$$H_0 : \beta_1 = 0,$$
$$H_1 : \beta_1 \neq 0.$$

W przypadku prostej regresji liniowej, hipoteza służy do oceny, czy istnieje związek między zmienną $X$ i $Y$.

Jeśli $\beta_1 = 0$ wówczas model regresji redukuje się do $Y = \beta_0 + \varepsilon$ i $X$ nie jest związana ze zmienną $Y$.

Będziemy używać statystyki testowej:

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}.$$

Przy prawdziwości hipotezy zerowej statystyka $t$ ma rozkład $t$-Studenta z $n - 2$ stopniami swobody.

# Testy normalności — Omnibus

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.612 |
| Model: | OLS | Adj. R-squared: | 0.610 |
| Method: | Least Squares | F-statistic: | 312.1 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.47e-42 |
| Time: | 11:50:53 | Log-Likelihood: | -519.05 |
| No. Observations: | 200 | AIC: | 1042. |
| Df Residuals: | 198 | BIC: | 1049. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 7.0326 | 0.458 | 15.360 | 0.000 | 6.130 | 7.935 |
| TV | 0.0475 | 0.003 | 17.668 | 0.000 | 0.042 | 0.053 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.531 | Durbin-Watson: | 1.935 |
| Prob(Omnibus): | 0.767 | Jarque-Bera (JB): | 0.669 |
| Skew: | -0.089 | Prob(JB): | 0.716 |
| Kurtosis: | 2.779 | Cond. No. | 338. |

# Test normalności — Omnibus

Test typu **omnibus D'Agostino-Pearsona** oparty o kurtozę i skośność.

Łącząc dwa testy otrzymuje się test czuły na odstępstwa od normalności zarówno w postaci niezerowej skośności jak i kurtozy istotnie różniej od 3.

Statystyką testową jest

$$K^2 = (Z(\sqrt{b_1}))^2 + (Z(b_2))^2,$$

gdzie $Z(\sqrt{b_1})$ to statystyka testowa testu opartego o skośność a $Z(b_2)$ to statystyka testowa testu opartego o kurtozę.

Asymptotyczny rozkład tej statystyki to rozkład $\chi^2$.

Ponadto: $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$, $\sqrt{b_1} = \frac{m_3}{m_2^{3/2}}$, $b_2 = \frac{m_4}{m_2^2} - 3$.

# Test normalności — Jarque-Bera

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.612 |
| Model: | OLS | Adj. R-squared: | 0.610 |
| Method: | Least Squares | F-statistic: | 312.1 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.47e-42 |
| Time: | 11:50:53 | Log-Likelihood: | -519.05 |
| No. Observations: | 200 | AIC: | 1042. |
| Df Residuals: | 198 | BIC: | 1049. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 7.0326 | 0.458 | 15.360 | 0.000 | 6.130 | 7.935 |
| TV | 0.0475 | 0.003 | 17.668 | 0.000 | 0.042 | 0.053 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.531 | Durbin-Watson: | 1.935 |
| Prob(Omnibus): | 0.767 | Jarque-Bera (JB): | 0.669 |
| Skew: | -0.089 | Prob(JB): | 0.716 |
| Kurtosis: | 2.779 | Cond. No. | 338. |

# Test normalności — Jarque-Bera

Innym testem opartym o kurtozę i skośność jest **test Jarque-Bera**. Statystyka testowa w przypadku tego testu ma łatwiejszą postać niż dla testu D'Agostino-Pearsona. Traci się jednak na niedokładnym oszacowaniu wartości krytycznych przy niewielkich wielkościach próby. Asymptotycznie ten test jest tak samo mocny jak test D'Agostino-Pearsona, ale na asymptotykę można liczyć jedynie w przypadku dużych prób.

Statystyka testowa ma postać:

$$\mathrm{JB} = \frac{n}{6}\left((\sqrt{b_1})^2 + \frac{1}{4}(b_2 - 3)^2\right).$$

# Statystyka Durbina-Watsona

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.612 |
| Model: | OLS | Adj. R-squared: | 0.610 |
| Method: | Least Squares | F-statistic: | 312.1 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.47e-42 |
| Time: | 11:50:53 | Log-Likelihood: | -519.05 |
| No. Observations: | 200 | AIC: | 1042. |
| Df Residuals: | 198 | BIC: | 1049. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 7.0326 | 0.458 | 15.360 | 0.000 | 6.130 | 7.935 |
| TV | 0.0475 | 0.003 | 17.668 | 0.000 | 0.042 | 0.053 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.531 | Durbin-Watson: | 1.935 |
| Prob(Omnibus): | 0.767 | Jarque-Bera (JB): | 0.669 |
| Skew: | -0.089 | Prob(JB): | 0.716 |
| Kurtosis: | 2.779 | Cond. No. | 338. |

**Test Durbina-Watsona** (statystyka) służy do oceny występowania korelacji pomiędzy resztami. Wzór na statystykę testu Durbina-Watsona ma postać:

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

Jeżeli statystyka DW $\approx 2$, możemy uznać **brak autokorelacji** pomiędzy resztami w modelu.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon,$$

gdzie $\varepsilon_i \sim \mathcal{N}(0, \sigma)$.

Zapis macierzowy:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

gdzie

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ 1 & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix},$$

$\beta^T = (\beta_0, \beta_1, \ldots, \beta_p)^T$ oraz $\varepsilon^T = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$.

Funkcja kryterialna:

$$\begin{aligned} \mathsf{b} &= \underset{\mathsf{b}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \underset{\mathsf{b}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathsf{X}_i\beta)^2, \end{aligned}$$

wówczas

$$\mathsf{b} = (\mathsf{X}'\mathsf{X})^{-1}\mathsf{X}'\mathsf{y}$$

# Wielowymiarowy model regresji liniowej — zbiór Advertising

Dopasowujemy model

$$\texttt{Sales} \approx b_0 + b_1 \cdot \texttt{TV} + b_2 \cdot \texttt{Radio} + b_3 \cdot \texttt{Newspaper}.$$

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.897 |
| Model: | OLS | Adj. R-squared: | 0.896 |
| Method: | Least Squares | F-statistic: | 570.3 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.58e-96 |
| Time: | 12:24:33 | Log-Likelihood: | -386.18 |
| No. Observations: | 200 | AIC: | 780.4 |
| Df Residuals: | 196 | BIC: | 793.6 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 2.9389 | 0.312 | 9.422 | 0.000 | 2.324 | 3.554 |
| TV | 0.0458 | 0.001 | 32.809 | 0.000 | 0.043 | 0.049 |
| Radio | 0.1885 | 0.009 | 21.893 | 0.000 | 0.172 | 0.206 |
| Newspaper | -0.0010 | 0.006 | -0.177 | 0.860 | -0.013 | 0.011 |

| | | | |
|---|---|---|---|
| Omnibus: | 60.414 | Durbin-Watson: | 2.084 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 151.241 |
| Skew: | -1.327 | Prob(JB): | 1.44e-33 |
| Kurtosis: | 6.332 | Cond. No. | 454. |

Dopasowując model regresji wielokrotnej ważnym jest odpowiedzenie na następujące pytania:

1. **Czy przynajmniej jeden z predyktorów $X_1, \ldots, X_p$ jest przydatny w przewidywaniu zmiennej odpowiedzi $Y$?**

# Test *F*

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.897 |
| Model: | OLS | Adj. R-squared: | 0.896 |
| Method: | Least Squares | F-statistic: | 570.3 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.58e-96 |
| Time: | 12:24:33 | Log-Likelihood: | -386.18 |
| No. Observations: | 200 | AIC: | 780.4 |
| Df Residuals: | 196 | BIC: | 793.6 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 2.9389 | 0.312 | 9.422 | 0.000 | 2.324 | 3.554 |
| TV | 0.0458 | 0.001 | 32.809 | 0.000 | 0.043 | 0.049 |
| Radio | 0.1885 | 0.009 | 21.893 | 0.000 | 0.172 | 0.206 |
| Newspaper | -0.0010 | 0.006 | -0.177 | 0.860 | -0.013 | 0.011 |

| | | | |
|---|---|---|---|
| Omnibus: | 60.414 | Durbin-Watson: | 2.084 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 151.241 |
| Skew: | -1.327 | Prob(JB): | 1.44e-33 |
| Kurtosis: | 6.332 | Cond. No. | 454. |

# Test $F$

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

$H_1$ : co najmniej jeden współczynnik $\beta_j$ jest niezerowy

Do weryfikacji hipotezy korzystamy ze statystyki

$$F = \frac{\text{TSS} - \text{RSS}}{p} : \frac{\text{RSS}}{n - p - 1},$$

która przy prawdziwości hipotezy zerowej ma rozkład $F$-Snedecora z $(p, n - p - 1)$ stopniami swobody.

Dopasowując model regresji wielokrotnej ważnym jest odpowiedzenie na następujące pytania:

1. Czy przynajmniej jeden z predyktorów $X_1, \ldots, X_p$ jest przydatny w przewidywaniu zmiennej odpowiedzi $Y$?
2. **Które zmienne niezależne $X_1, \ldots, X_p$ są istotne w wyjaśnianiu zmiennej odpowiedzi $Y$?**

# Istotność zmiennych niezależnych

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.897 |
| Model: | OLS | Adj. R-squared: | 0.896 |
| Method: | Least Squares | F-statistic: | 570.3 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.58e-96 |
| Time: | 12:24:33 | Log-Likelihood: | -386.18 |
| No. Observations: | 200 | AIC: | 780.4 |
| Df Residuals: | 196 | BIC: | 793.6 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 2.9389 | 0.312 | 9.422 | 0.000 | 2.324 | 3.554 |
| TV | 0.0458 | 0.001 | 32.809 | 0.000 | 0.043 | 0.049 |
| Radio | 0.1885 | 0.009 | 21.893 | 0.000 | 0.172 | 0.206 |
| Newspaper | -0.0010 | 0.006 | -0.177 | 0.860 | -0.013 | 0.011 |

| | | | |
|---|---|---|---|
| Omnibus: | 60.414 | Durbin-Watson: | 2.084 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 151.241 |
| Skew: | -1.327 | Prob(JB): | 1.44e-33 |
| Kurtosis: | 6.332 | Cond. No. | 454. |

# Istotność współczynników regresji

$$H_0 : \beta_i = 0,$$
$$H_1 : \beta_i \neq 0,$$

dla $i = 1, \ldots, p$.

Używamy statystyki testowej:

$$t = \frac{\hat{\beta}_i - 0}{\mathrm{SE}(\hat{\beta}_i)},$$

która przy prawdziwości hipotezy zerowej statystyka $t$ ma rozkład $t$-Studenta z $n - 2$ stopniami swobody.

Dopasowując model regresji wielokrotnej ważnym jest odpowiedzenie na następujące pytania:

1. Czy przynajmniej jeden z predyktorów $X_1, \ldots, X_p$ jest przydatny w przewidywaniu zmiennej odpowiedzi $Y$?
2. Które zmienne niezależne $X_1, \ldots, X_p$ są istotne w wyjaśnianiu zmiennej odpowiedzi $Y$?
3. **Jak dobrze model jest dopasowany do danych?**

# Dopasowanie modelu

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Sales | R-squared: | 0.897 |
| Model: | OLS | Adj. R-squared: | 0.896 |
| Method: | Least Squares | F-statistic: | 570.3 |
| Date: | Sun, 16 Feb 2020 | Prob (F-statistic): | 1.58e-96 |
| Time: | 12:24:33 | Log-Likelihood: | -386.18 |
| No. Observations: | 200 | AIC: | 780.4 |
| Df Residuals: | 196 | BIC: | 793.6 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 2.9389 | 0.312 | 9.422 | 0.000 | 2.324 | 3.554 |
| TV | 0.0458 | 0.001 | 32.809 | 0.000 | 0.043 | 0.049 |
| Radio | 0.1885 | 0.009 | 21.893 | 0.000 | 0.172 | 0.206 |
| Newspaper | -0.0010 | 0.006 | -0.177 | 0.860 | -0.013 | 0.011 |

| | | | |
|---|---|---|---|
| Omnibus: | 60.414 | Durbin-Watson: | 2.084 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 151.241 |
| Skew: | -1.327 | Prob(JB): | 1.44e-33 |
| Kurtosis: | 6.332 | Cond. No. | 454. |

# Materiały na podstawie

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, An Introduction to Statistical Learning with Applications in R , Springer 2014.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning, Springer 2009.