Czynniki deterministyczne

$$m = 1 \qquad p = 2$$

Obserwujemy cechę Y oraz zmienne X_1, X_2
Obiekt $\longrightarrow (X_1, X_2, Y)$

1. Propozycja funkcji regresji f.

2. Dopasowanie zaproponowanej funkcji.

3. Ocena jakości dopasowania.

4. Wnioski.

Założenie:
Cecha Y ma rozkład normalny

Funkcja regresji

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$(Y_1, x_{11}, x_{21}), \ldots, (Y_n, x_{1n}, x_{2n})$ — obserwacje
Model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2_i} + \varepsilon_i, \ i = 1, \ldots, n,$$

ε_i są niezależnymi zmiennymi losowymi o tym samym rozkładzie normalnym $N(0, \sigma^2)$.

Estymacja współczynników metodą najmniejszych kwadratów

Znaleźć takie β_0, β_1 i β_2 by

$$\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2_i}))^2 = \min$$

Rozwiązanie

$$\begin{cases} \widehat{\beta}_1 \mathrm{var} x_1 + \widehat{\beta}_2 \mathrm{cov}(x_1, x_2) = \mathrm{cov}(x_1, x_3) \\ \widehat{\beta}_1 \mathrm{cov}(x_1, x_2) + \widehat{\beta}_2 \mathrm{var} x_2 = \mathrm{cov}(x_2, x_3) \\ \bar{Y} - \widehat{\beta}_1 \bar{x}_1 - \widehat{\beta}_2 \bar{x}_2 = \widehat{\beta}_0 \end{cases}$$

Resztowa suma kwadratów

$$\mathrm{RSS} = \sum_{i=1}^{n} (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}))^2$$

Ocena wariancji σ^2

$$S^2 = \frac{1}{n-3} \left(\mathrm{var} Y - \widehat{\beta}_1 \mathrm{cov}(Y, x_1) - \widehat{\beta}_2 \mathrm{cov}(Y, x_2) \right)$$

Wariancje estymatorów

$$S^2_{\beta_1} = \frac{S^2}{(1 - R^2_{12}) \mathrm{var} x_1} \qquad S^2_{\beta_2} = \frac{S^2}{(1 - R^2_{12}) \mathrm{var} x_2}$$

$$S^2_{\beta_0} = S^2 \left(\frac{1}{n} + \frac{\left(\frac{\bar{x}_1^2}{\mathrm{var} x_1} + \frac{\bar{x}_2^2}{\mathrm{var} x_2} - \frac{R^2_{12}}{\mathrm{cov}(x_1, x_2)} \right)}{1 - R^2_{12}} \right)$$

Istnienie zależności
Weryfikacja hipotezy $H_0 : \beta_1 = \beta_2 = 0$

Źródło zmienności	Suma kwadratów	Stopnie swobody	Średnie kwadraty	F
Regresja	var$R$	2	s_R^2=var$R$	s_R^2/s^2
Błąd	RSS	$n-3$	s^2=RSS/$(n-3)$	
Całkowita	var$Y$	$n-1$		

$$\text{var}R = \widehat{\beta}_1 \text{cov}(Y, x_1) + \widehat{\beta}_2 \text{cov}(Y, x_2)$$

Jeżeli hipoteza $H_0 : \beta_1 = \beta_2 = 0$ jest prawdziwa, to

$$F = \frac{s_R^2}{s^2}$$

ma rozkład F z $(2, n-3)$ stopniami swobody

Hipotezę odrzucamy, jeżeli $F > F(\alpha; 2, n-3)$
$F(\alpha; 2, n-3)$ — wartość krytyczna rozkładu F.

$$H_0 : \beta_1 = 0$$

Test Studenta (poziom istotności α)

Statystyka testowa

$$t_{\text{emp}} = \frac{\widehat{\beta}_1}{S_{\beta_1}}$$

Wartość krytyczna $t(\alpha; n-3)$
Hipotezę odrzucamy, jeżeli $|t_{\text{emp}}| > t(\alpha; n-3)$

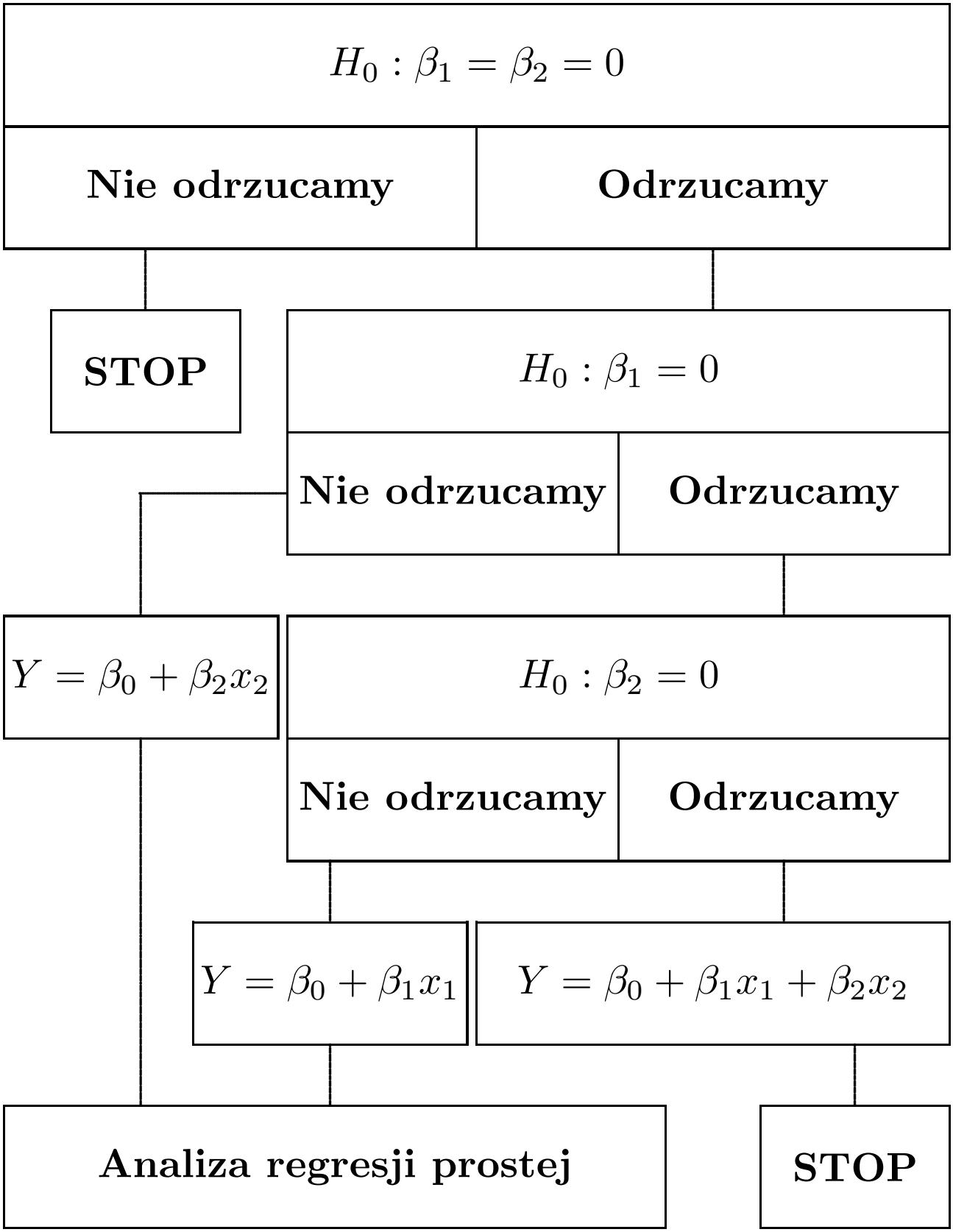
$$H_0 : \beta_2 = 0$$

Test Studenta (poziom istotności α)

Statystyka testowa

$$t_{\text{emp}} = \frac{\widehat{\beta}_2}{S_{\beta_2}}$$

Wartość krytyczna $t(\alpha; n-3)$
Hipotezę odrzucamy, jeżeli $|t_{\text{emp}}| > t(\alpha; n-3)$


$H_0 : \beta_1 = \beta_2 = 0$
Nie odrzucamy
Odrzucamy
STOP
$H_0 : \beta_1 = 0$
Nie odrzucamy
Odrzucamy
$Y = \beta_0 + \beta_2 x_2$
$H_0 : \beta_2 = 0$
Nie odrzucamy
Odrzucamy
$Y = \beta_0 + \beta_1 x_1$
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
Analiza regresji prostej
STOP

Przedział ufności dla β_1
Poziom ufności $1-\alpha$

$$\beta_1 \in (\widehat{\beta}_1 - t(\alpha; n-3)S_{\beta_1}; \widehat{\beta}_1 + t(\alpha; n-3)S_{\beta_1})$$

Interpretacja współczynnika regresji β_1
jeżeli wartość zmiennej niezależnej x_1 wzrośnie o jednostkę zaś zmienna x_2 pozostanie na tym samym poziomie, to średnia wartość cechy Y zmieni się (wzrośnie lub zmaleje) o około $\widehat{\beta}_1$ jednostek, a dokładniej zmieni się o $\widehat{\beta}_1 \pm t(\alpha; n-3)S_{\beta_1}$ jednostek.

Przedział ufności dla β_2: podobnie jak dla β_1

Współczynnik determinacji

Niech

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, i = 1, \ldots, n$$

oraz niech

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}, i = 1, \ldots, n$$

Dla par $(Y_i, \widehat{Y}_i)$ wyznaczamy

$$R = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(\widehat{Y}_i - \bar{\widehat{Y}})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sum_{i=1}^{n}(\widehat{Y}_i - \bar{\widehat{Y}})^2}}.$$

Współczynnik determinacji zmiennej Y przez X

$$D = R^2 \cdot 100\% = \frac{\sum(\widehat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \cdot 100\%.$$

Jest to liczba z przedziału $(0\%, 100\%)$ i dopasowanie funkcji regresji jest tym lepsze, im ten współczynnik jest wyższy.

Obszar ufności dla prostej regresji

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Średnia wartość cechy Y dla ustalonych wartości $X_1 = x_1, X_2 = x_2$

$$\widehat{y}(x_1, x_2) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$$

Obszar ufności (poziom ufności $1 - \alpha$)

$$E(Y|x_1, x_2) \in (\widehat{y}(x_1, x_2) - t(\alpha; n-2) S_Y; \\ \widehat{y}(x_1, x_2) + t(\alpha; n-2) S_Y)$$

$$S_Y^2 = S^2 \cdot [1 \; x_1 \; x_2] \cdot \\ \cdot \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{2i} & \sum x_{1i} x_{2i} & \sum x_{2i}^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

Na podstawie obszaru ufności wnioskujemy o **wartościach średnich** cechy Y jednocześnie dla wielu wybranych wartości cech X_1, X_2

Predykcja wartości zmiennej $Y(x_1, x_2)$

Wartość cechy Y dla ustalonych $X_1 = x_1, X_2 = x_2$

$$\widehat{y}(x_1, x_2) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$$

Obszar predykcji (poziom ufności $1 - \alpha$)

$$Y(x_1, x_2) \in (\widehat{y}(x_1, x_2) - t(\alpha; n-2) S_{y(x_1,x_2)}; \\ \widehat{y}(x_1, x_2) + t(\alpha; n-2) S_{y(x_1,x_2)})$$

$$S^2_{y(x_1,x_2)} = S^2 \cdot \Bigg(1 + [1 \ x_1 \ x_2] \cdot \\ \cdot \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \Bigg)$$

Na podstawie obszaru predykcji wnioskujemy o **wartościach** cechy Y jednocześnie dla wielu wybranych wartości cech X_1, X_2

Przykład. Badano wielkość produkcji (*prod*) pewnego artykułu w zależności od ilości dwóch surowców (sur_1, sur_2) wykorzystywanych w wytwarzaniu tego artykułu. Na podstawie poniższych danych przeprowadzić analizę regresji.

Plan działania

1. Propozycja funkcji regresji

2. Dopasowanie funkcji regresji

3. Istnienie zależności

- **3a.** Badanie globalne
- **3b.** Badanie szczegółowe

4. Jakość dopasowania

5. Wnioski

s_1	s_2	$prod$	s_1	s_2	$prod$	s_1	s_2	$prod$	s_1	s_2	$prod$
0.1	0.1	1.936248	0.6	0.3	4.697887	0.1	0.6	3.776876	0.6	0.8	9.024716
0.2	0.1	2.017051	0.7	0.3	6.212012	0.2	0.6	6.488632	0.7	0.8	7.385809
0.3	0.1	2.547019	0.8	0.3	5.711818	0.3	0.6	5.356985	0.8	0.8	7.319159
0.4	0.1	2.991221	0.9	0.3	6.801152	0.4	0.6	6.040919	0.9	0.8	9.403422
0.5	0.1	3.103400	1.0	0.3	5.130012	0.5	0.6	7.274057	1.0	0.8	9.533901
0.6	0.1	3.395465	0.1	0.4	4.711792	0.6	0.6	7.327822	0.1	0.9	7.462994
0.7	0.1	2.366942	0.2	0.4	3.901310	0.7	0.6	7.871890	0.2	0.9	7.943808
0.8	0.1	2.954253	0.3	0.4	5.246389	0.8	0.6	7.862603	0.3	0.9	8.495195
0.9	0.1	3.454655	0.4	0.4	5.762669	0.9	0.6	6.612192	0.4	0.9	7.988018
1.0	0.1	2.836646	0.5	0.4	6.670547	1.0	0.6	6.928189	0.5	0.9	8.260379
0.1	0.2	2.633539	0.6	0.4	5.662259	0.1	0.7	5.658337	0.6	0.9	8.963044
0.2	0.2	2.737200	0.7	0.4	5.588580	0.2	0.7	6.262777	0.7	0.9	8.811154
0.3	0.2	2.922328	0.8	0.4	6.470962	0.3	0.7	5.986275	0.8	0.9	8.164607
0.4	0.2	3.376518	0.9	0.4	5.960982	0.4	0.7	6.799810	0.9	0.9	7.778411
0.5	0.2	3.603429	1.0	0.4	6.822329	0.5	0.7	7.379986	1.0	0.9	10.306121
0.6	0.2	3.267117	0.1	0.5	3.861578	0.6	0.7	7.987376	0.1	1.0	6.596179
0.7	0.2	3.934322	0.2	0.5	4.708645	0.7	0.7	7.899379	0.2	1.0	7.709768
0.8	0.2	4.107574	0.3	0.5	4.773405	0.8	0.7	7.304735	0.3	1.0	8.029625
0.9	0.2	4.438335	0.4	0.5	5.677243	0.9	0.7	9.891345	0.4	1.0	7.512992
1.0	0.2	4.311634	0.5	0.5	6.135761	1.0	0.7	8.784312	0.5	1.0	9.852992
0.1	0.3	3.344719	0.6	0.5	6.402305	0.1	0.8	5.684369	0.6	1.0	8.752144
0.2	0.3	3.825492	0.7	0.5	6.375133	0.2	0.8	7.043533	0.7	1.0	8.561350
0.3	0.3	4.923739	0.8	0.5	8.421879	0.3	0.8	7.663122	0.8	1.0	8.809613
0.4	0.3	4.521357	0.9	0.5	5.816456	0.4	0.8	6.987355	0.9	1.0	9.380318
0.5	0.3	4.680259	1.0	0.5	6.569014	0.5	0.8	7.099786	1.0	1.0	9.556762

Funkcja regresji
(funkcja produkcji Cobba–Douglasa)

$$prod = a \cdot sur_1^{\alpha_1} \cdot sur_2^{\alpha_2}$$

Model

$$\ln(prod) = \ln(a) + \alpha_1 \ln(sur_1) + \alpha_2 \ln(sur_2)$$

$$Y = \ln(prod) \qquad x_1 = \ln(sur_1) \qquad x_2 = \ln(sur_2)$$

$$\beta_0 = \ln(a) \qquad \beta_1 = \alpha_1 \qquad \beta_2 = \alpha_2$$

Dopasowanie funkcji regresji

$$\widehat{\beta}_0 = 2.307795 \quad \widehat{\beta}_1 = 0.200292 \quad \widehat{\beta}_2 = 0.511396$$

$$s^2 = 0.011964$$

$$s_{\beta_0} = 0.020739 \quad s_{\beta_1} = 0.015729 \quad s_{\beta_2} = 0.015729$$

Niezbędne obliczenia

$$n = 100$$

$$\sum x_{1i} = -79.214384 \qquad \sum x_{2i} = -79.214384$$

$$\sum y_i = 174.40359$$

$$\sum x_{1i}^2 = 111.10834 \qquad \sum x_{2i}^2 = 111.10834$$

$$\sum y_i^2 = 319.91382$$

$$\sum x_{1i}y_i = -128.46678 \qquad \sum x_{2i}y_i = -113.42206$$

$$\sum x_{1i}x_{2i} = 62.749186$$

Istnienie zależności

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

Źródło zmienności	Suma kwadratów	Stopnie swobody	Średnie kwadraty	F
Regresja	14.587176	2	7.293588	609.62
Błąd	1.160518	97	0.011964	
Całkowita	15.747694	99		

Wartość krytyczna

$$F(0.05; 2, 97) = 3.09$$

Wniosek:
zaproponowana funkcja regresji może opisywać zależność między wielkością produkcji a nakładami

$$H_0 : \beta_1 = 0$$

Test Studenta $(\alpha = 0.05)$

Statystyka testowa

$$t_{\text{emp}} = \frac{\widehat{\beta}_1}{S_{\beta_1}} = \frac{0.200292}{0.015729} = 12.733942$$

Wartość krytyczna $t(0.05; 97) = 1.984723$
Hipotezę odrzucamy

$$H_0 : \beta_2 = 0$$

Test Studenta $(\alpha = 0.05)$

Statystyka testowa

$$t_{\text{emp}} = \frac{\widehat{\beta}_2}{S_{\beta_2}} = \frac{0.511396}{0.015729} = 32.512951$$

Wartość krytyczna $t(0.05; 97) = 1.984723$
Hipotezę odrzucamy

Równania regresji

$$Y = 2.307795 + 0.200292x_1 + 0.511396x_2$$

$$prod = 10.052235 \cdot sur_1^{0.200292} \cdot sur_2^{0.511396}$$

Współczynnik determinacji

$$D^2 = 92.63\%$$

Zastosowanie

Przedział ufności dla oczekiwanej produkcji przy nakładach $sur_1 = 0.2$ oraz $sur_2 = 0.4$

dla Y	$(1.483127, 1.550574)$
dla produkcji	$(4.406703, 4.714174)$

. . . . . .

Przedział predykcji dla wielkości produkcji przy nakładach $sur_1 = 0.2$ oraz $sur_2 = 0.4$

dla Y	$(1.297156, 1.736544)$
dla produkcji	$(3.658878, 5.677688)$