Regresja liniowa i trend liniowy

Tomasz Przechlewski

Czerwiec 2016

1 Regresja liniowa

$$Y = \alpha_0 + \alpha_1 X + e \tag{1}$$

gdzie e oznacza tzw. **składnik losowy**.

Na podstawie i-elementowej próby szacujemy **linię regresji**:

$$\hat{y}_i = a_0 + a_1 x_i \tag{2}$$

gdzie $i = 1, ..., n$

Metoda najmniejszych kwadratów (MNK) polega na takim oszacowaniu parametrów α_0 i α_1, aby na podstawie n-elementowej próby wyrażenie

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \rightarrow \min \tag{3}$$

Powyższy warunek prowadzi do następujących wzorów na parametry a_0 i a_1:

$$a_0 = \bar{y} - a_1\bar{x}; \tag{4}$$

$$a_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

gdzie $\bar{x}$, $\bar{y}$ to średnie wartości x oraz y.

Ocenę a_1 nazywamy **współczynnikiem regresji liniowej**. Interpretacja: jaki jest **przeciętny wzrost/spadek wartości zmiennej zależnej (Y), przy zmianie wartości zmiennej niezależnej o jednostkę**.

Resztami nazywamy różnice pomiędzy wartościami empirycznymi a teoretycznymi:

$$u_i = y_i - \hat{y}_i \qquad i = 1, ..., n \tag{6}$$

Dla parametrów oszacowanych MNK zawsze jest spełniony warunek:

$$\sum_{i=1}^{n} u_i = 0 \tag{7}$$

Dlatego miarą dokładności jest suma kwadratów reszt (**wariancja resztowa/wariancja składnika resztowego**) lub **pierwiastek kwadratowy wariancji resztowej** zwany **odchyleniem standardowym składnika resztowego** lub **błędem standardowym reszt**:

$$s_e^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

gdzie k to liczba szacowanych parametrów równania regresji (czyli 2).

W miarę wzrostu liczbowej wartości odchylenia standardowego składnika resztowego „dobroć" dopasowania funkcji regresji do danych empirycznych maleje.

1.1 Współczynniki zbieżności ϕ^2 i determinacji R^2

Wartość empiryczna zmiennej objaśnianej może zostać zapisana jako **wartość teoretyczna** plus reszta:

$$y = \hat{y} + e \tag{9}$$

odejmując $\bar{y}$ od obu stron równania, podnosząc obie strony do kwadratu i sumując po $i = 1, ..., n$:

$$\sum_{i=1}^{n}(y - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 + \sum_{i=1}^{n} e^2 \tag{10}$$

Zwróćmy uwagę, że:

$\sum_{i=1}^{n}(y-\bar{y})^2$ – to suma kwadratów odchyleń zmiennej Y od jej średniej (OSK – ogólna suma kwadratów)

$\sum_{i=1}^{n}(\hat{y} - \bar{y})^2$ – to suma kwadratów wartości teoretycznych od średniej (WSK – wyjaśniona suma kwadratów)

$\sum_{i=1}^{n} e^2$ – to suma kwadratów odchyleń pomiędzy wartościami empirycznymi a teoretycznymi (RSK – resztowa suma kwadratów)

$$\text{OSK} = \text{WSK} + \text{RSK} \tag{11}$$

Współczynnik zbieżności ϕ^2

$$\phi^2 = \frac{\text{RSK}}{\text{OSK}} \times 100 \tag{12}$$

Interpretacja: udział zmienności resztowej w całkowitej zmienności, tj. % zmienności zmiennej objaśnianej nie wyjaśniona przez model regresji liniowej. Wartość współczynnika zawiera się w przedziale 0–100, im mniej tym lepiej.

Współczynnik determinacji R^2

$$R^2 = \frac{\text{WSK}}{\text{OSK}} \times 100 \tag{13}$$

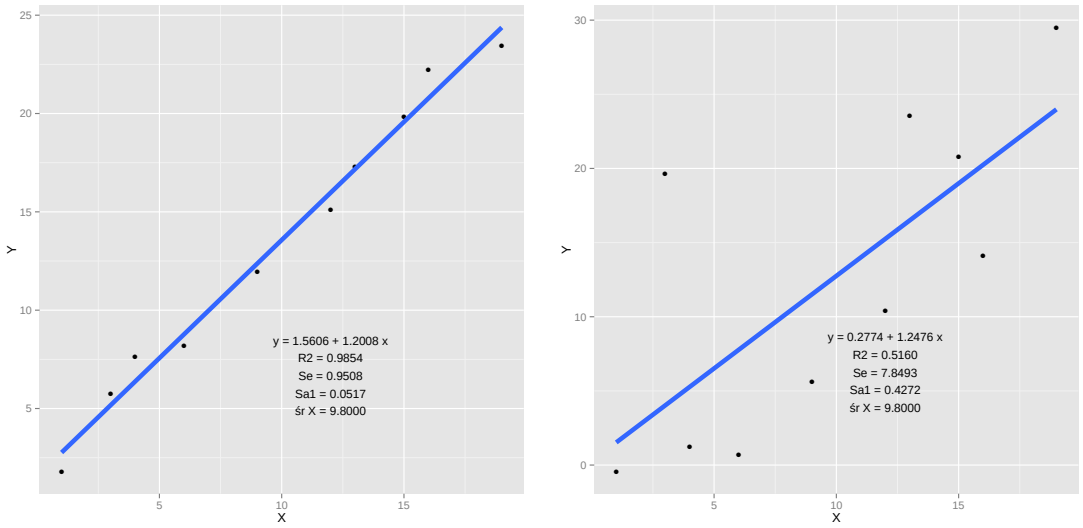
Interpretacja: udział zmienności wyjaśnionej w całkowitej zmienności, tj. % zmienności zmiennej objaśnianej wyjaśniona przez model regresji liniowej. Wartość współczynnika zawiera się w przedziale 0–100, im więcej tym lepiej.

Współczynnik zmienności losowej

$$V_e = \frac{S_e}{\bar{y}} \times 100 \tag{14}$$

Interpretacja: ile procent przeciętnego poziomu zmiennej objaśnianej stanowi przeciętne odchylenie wartości empirycznych od wartości teoretycznych.

Ilustracja graficzna (por. rys. 1.1): składnik losowy e ma rozkład normalny o wartości średniej 0 oraz odchyleniu standardowym 1 (lewy wykres) oraz 6 (prawy wykres).



Zgodnie z założeniami mniejsze rozproszenie składnika losowego przekłada się na lepsze dopasowanie linii regresji, większą wartość R^2 oraz mniejsze wartości S_e i S_{a_1}.

1.2 Badanie istotności współczynnika regresji

Ponieważ o wartościach α_0 i α_1 wnioskujemy na podstawie próby a model zawiera składnik losowy, to parametry a_0 oraz a_1 są zmiennymi losowymi.

Przy przyjęciu pewnych założeń odnośnie składnika losowego można wywieść, że zmienne a_0 oraz a_1 mają wartości oczekiwane (średnie) równe prawdziwym wartościom, tj. $\bar{a}_0 = \alpha_0$ oraz $\bar{a}_1 = \alpha_1$. Odchylenia standardowe S_{a_1} oraz S_{a_0} (zwane średnimi błędami szacunku parametrów modelu) określają precyzję oszacowania parametrów.

1.3 Istotność współczynnika regresji

Statystyczną oceną jakości modelu regresji liniowej jest zweryfikowanie hipotezy o istotności współczynnika regresji, tj:

$H_0 : \alpha_1 = 0$, wobec $H_1 : \alpha_1 \neq 0$ (można też stosować hipotezy $H_1 : \alpha_1 < 0$ lub $H_1 : \alpha_1 > 0$ jeżeli dysponujemy stosownym informacjami pozastatystycznymi wskazującymi iż alternatywą dla wartości zerowej jest mniejsza/większa wartość współczynnika)

Można udowodnić, że statystyka:

$$t = \frac{a_1}{S_{a_1}} \tag{15}$$

ma rozkład T-Studenta z $n - k$ stopniami swobody.

Reguła decyzyjna:

Mając obliczone t porównujemy je dla określonego poziomu istotności oraz $n - k$ stopni swobody z tzw. wartością krytyczną, którą można obliczyć przykładowo korzystając z funkcji `ROZKŁAD.T` w programie OOCalc/Excel (niżej dokładniej opisanej).

Jeżeli $t <$ wartość krytyczna nie ma podstaw do odrzucenia H_0. W przypadku przeciwnym H_0 należy odrzucić.

Uwaga:

W przypadku $H_1 : \alpha_1 \neq 0$ mówimy o **dwustronnym obszarze krytycznym** – duże odchylenia na $\pm$ świadczą przeciwko H_0; w przypadku

$H_1 : \alpha_1 < 0$ mówimy o **lewostronnym obszarze krytycznym** – tylko duże odchylenia na minus świadczą przeciwko H_0 (te odchylenia na minus są *po lewej* stronie w kartezjańskim układzie współrzędnych.

$H_1 : \alpha_1 > 0$ mówimy o **prawostronnym obszarze krytycznym** – tylko duże odchylenia na plus świadczą przeciwko H_0 (te odchylenia na pplus są *po prawej* stronie w kartezjańskim układzie współrzędnych.

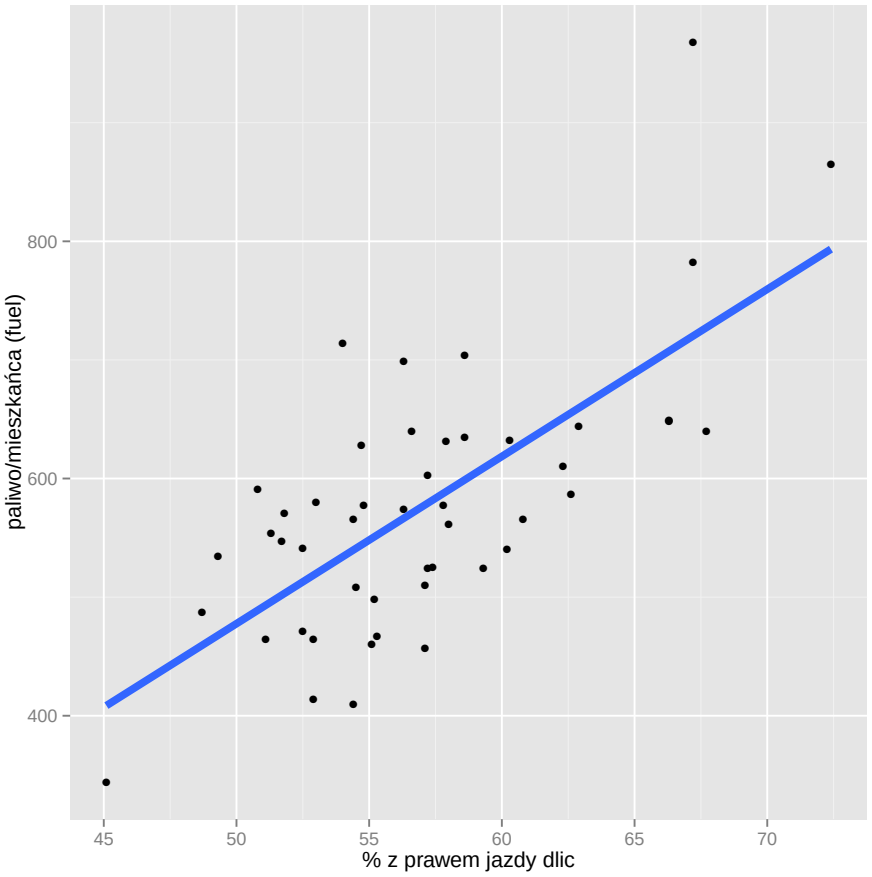
Alternatywna reguła decyzyjna:

Programy statystyczne często podają prawdopodobieństwo $P(|T|) > t$, tj prawdopodobieństwo że zmienna T przyjmie (co do wartości bezwzględnej) wartość t lub większą. Znając wartość prawdopodobieństwa P wystarczy porównać ją z poziomem istotności:

jeżeli $P \geqslant \alpha$ nie ma podstaw do odrzucenia H_0 jeżeli $P < \alpha$ H_0 należy odrzucić

1.4 Przykład

Plik `fueldata.ods` zawiera dane dotyczące m.in zużycia paliwa na mieszkańca (FUEL) oraz odsetek mieszkańców posiadających prawo jazdy (DLIC) dla 48 stanów w USA (por. rys. 1.4).



Zależność pomiędzy przeciętnym zużyciem benzyny na mieszkańca a odsetkiem kie-

rowców w stanie można zapisać jako:

$$\text{FUEL} = \alpha_0 + \alpha_1 \text{DLIC} + e \tag{16}$$

Oszacowana linia regresji dana jest równaniem:

$$\text{FUEL} = 14,01\text{DLIC} - 227,31 \tag{17}$$

Interpretacja: Zwiększenie o 1% odsetka mieszkańców posiadających prawo jazdy przeciętnie zwiększy zużycie na głowę o 14,01 galona.

Uwaga: do obliczenia linii regresji w programie OOCalc/Excel służy funkcja:

```
REGLINP{y;x;1;1}
```

Jest to funkcja tablicowa (tj zwracająca obszar a nie pojedynczą wartość), co oznacza że należy ją zatwierdzić naciskając Ctr-Shift-Enter (zamiast zwykłego Enter). Funkcja `REGLINP` (w wersji OpenOffice) zwraca obszar o wielkości 5 wierszy na 2 kolumny. Poszczególne komórki tego obszaru zawierają co następuje:

$$\begin{array}{cc} a_1 & a_0 \\ S_{a_1} & S_{a_0} \\ R^2 & S_e \end{array}$$

Zawartość wierszy 4 i 5 nie interesuje nas.... Użytkownicy Excela proszę sprawdzić w dokumentacji jak interpretować działanie funkcji `REGLINP` w Excelu.

Odsyłam do pliku `fueldata.ods` w celu przećwiczenia praktycznego wyznaczania linii regresji za pomocą `REGLINP`.

Ocena dopasowania:

$S_e = 80,88$, Przeciętne odchylenie wartości teoretycznych od empirycznych wynosi 80.88 galona (S_e jest zawsze mianowane w jednostkach zmiennej Y).

Współczynnik zbieżności $R^2 = 48.9\%$ oznacza, że 48,9% zmienności zużycia benzyny na głowę jest objaśnione przez model regresji liniowej pomiędzy zużyciem benzyny na głowę a odsetkiem mieszkańców posiadających prawo jazdy.

Zatem $\phi^2 = 100 - R^2 = 100 - 48,9 = 51,1$. 51,1% zmienności zużycia benzyny na głowę **nie jest wyjaśniana** przez model regresji liniowej.

Odchylenie ocen parametrów: $S_{a_1} = 2,13$ oraz $S_{a_0} = 121,9$.

Już na pierwszy rzut oka widać, że parametr a_1 jest istotny ($H_0 : \alpha = 0$ należy odrzucić) ponieważ przeciętny błąd S_{a_1} stanowi zaledwie 15% oceny parametru.

Sprawdzanie istotności parametru a_1:

Obliczenie t i porównanie jej z wartością krytyczną

Do wyznaczanie wartości zmiennej t odpowiadającej określonemu prawdopodobieństwu służy:

```
ROZKŁ.T.ODWR.DS(poziom-istotności;stopnie-swobody)
```

Zatem `ROZKŁ.T.ODWR.DS(0,05;46)` = 2,013.

Wartością krytyczną (obszar dwustronny) jest zatem 2,013.

Ponieważ $t_{n-2} = 6,577 > 2,013$ (wartość krytyczna na poziomie istotności $\alpha = 0.05$ dla $48 - 2 = 46$ stopni swobody) to H_0 należy odrzucić.

Obliczanie prawdopodobieństwa dla t i porównanie z poziomem istotności α:

Wykonujemy funkcję `ROZKŁAD.T(wartość-zmiennej;stopnie-swobody;tryb)`.

Wartość zmiennej to oczywiście a_1/S_{a_1};

Stopnie-swobody to oczywiście liczba stopni swobody $(n-k)$;

Tryb określa czy testujemy hipotezę alternatywną postaci $H_1 : \alpha_1 \neq 0$ (dwustronny obszar krytyczny; wtedy tryb = 2) czy też $H_1 : \alpha_1 < 0$ / $H_1 : \alpha_1 > 0$ (prawo/lewo-stronny obszar krytyczny ; wtedy tryb = 1).

Użytkownicy Excela proszę sprawdzić w dokumentacji jak interpretować działanie funkcji `ROZKŁAD.T` w Excelu.

Stosując funkcję `ROZKŁAD.T` otrzymujemy (dla dwustronnego obszaru krytycznego):

```
ROZKŁAD.T(6,577;46;2) = 3,289E-008
```

Ponieważ $3,289\text{E}-008 \approx 0,000000003 < 0,05$ (porównujemy z zakładanym poziomem istotności $\alpha = 0,05$) H_0 należy odrzucić – przy założeniu prawdziwości H_0 wartość $t = 6,577$ lub większa co do wartości bezwzględnej zdarza się 3 razy na 100000000.

2 Liniowa funkcja trendu

Liniowa funkcja trendu ma postać:

$$Y = \alpha_0 + \alpha_1 t + e \tag{18}$$

gdzie e oznacza tzw. składnik losowy a t – zmienna czasowa (np. $t = 1, ..., n$)

Od strony rachunkowej ocena jakości dopasowania oraz ocena istotność parametrów strukturalnych jest wykonywana **identycznie jak w metodzie regresji** tylko należy zmodyfikować interpretacje:

Współczynnik zbieżności – procent zmienności zmiennej y nie objaśniony przez liniową funkcję trendu.

Współczynnik determinacji $R^2 = 100 - \phi^2$ interpretuje się jako „procent zmienności zmiennej y objaśniony przez liniową funkcję trendu."

2.1 Przykład

Plik `Cena_paliw_sprzedaz_i_samochody_2.ods`.