**Akademia Górniczo-Hutnicza IM. Stanisława Staszica**

**w Krakowie Wydział Zarządzania**

# Analiza wpływu czynników na wydatki na gry komputerowe

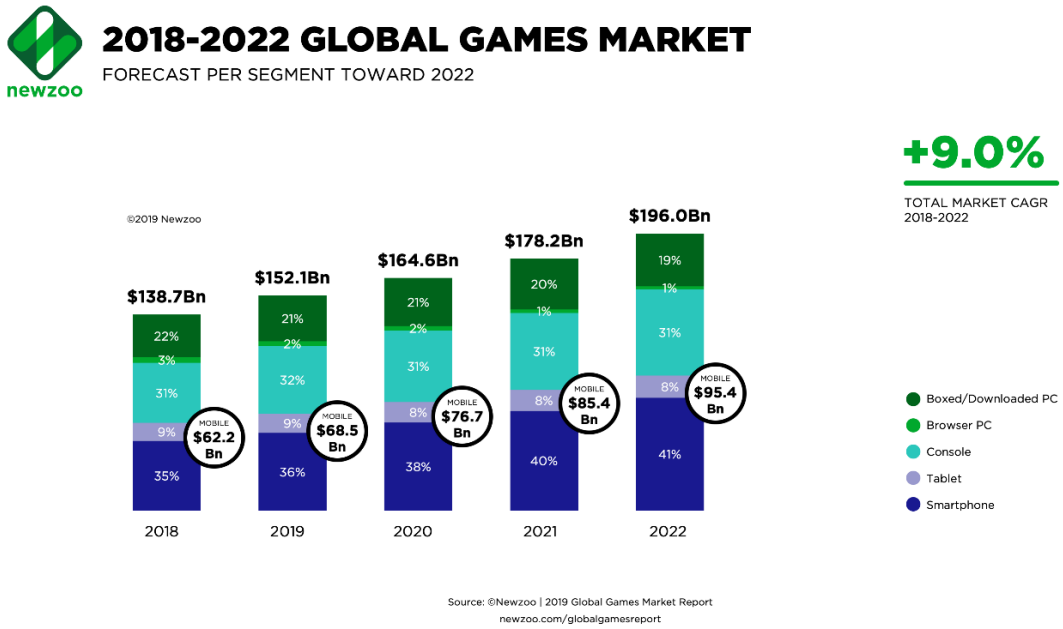
Opracował: Marcin Klimczak, 06.2021 r.

# Spis treści

# Wprowadzenie do tematyki projektu

## Wstęp

Projekt dotyczy badania rynku gier komputerowych i mobilnych, czyli tematu, który nie jest często przytaczany. Rynek ten jest jednak bardzo ważny i warty uwagi, ze względu na jego ciągłe światowe wzrosty, które można zobaczyć na Rysunku 1:




*Rysunek 1 - Wielkość rynku gier w czasie w miliardach USD*
*źródło: newzoo.com – strona poświęcona danym dotyczącym rynku gier*

Jest to również rynek, na którym dochodzi do codziennej kradzieży i łamania prawa, w postaci pobierania gier nielegalnie. Już te dwa powody wystarczają, aby zainteresować się tematem wydatków na gry, stąd też pomysł na projekt.

## Tematyka, cel projektu i hipotezy badawcze

Celem projektu jest analiza wpływu wybranych czynników na wydatki roczne na gry komputerowe lub mobilne.

W projekcie chciałbym sprawdzić, jakiego rodzaju dobrem są gry komputerowe w badanej populacji – dobrem niższego rzędu, dobrem normalnym czy dobrem luksusowym. Zatem pierwszą hipotezą badawczą będzie:

$$H_0: Dochód\ gospodarstwa\ domowego\ nie\ wpływa\ istotnie\ na\ wydatki\ na\ gry$$
$$H_1: Dochód\ gospodarstwa\ domowego\ wpływa\ istotnie\ na\ wydatki\ na\ nie \quad (1.1)$$

A hipotezą bezpośrednio dotyczącą problemu:

$$H_0: Zmiana\ dochodu\ gospodarstwa\ domowego\ wpływa\ na\ spadek\ wydatków\ na\ gry$$
$$H_1: Zmiana\ dochodu\ gospodarstwa\ domowego\ nie\ wpływa\ na\ spadek\ wydatków\ na\ gry \quad (1.2)$$

Można chcieć także sprawdzić, czy osoby spędzające więcej czasu na graniu, wydają na nie więcej w skali roku. W tym celu sformułowano hipotezę pomocniczą:

$$H_0: Tygodniowy\ czas\ spędzony\ na\ graniu\ nie\ wpływa\ istotnie\ na\ wydatki\ na\ gry,$$
$$H_1: Tygodniowy\ czas\ spędzony\ na\ graniu\ wpływa\ istotnie\ na\ wydatki\ na\ gry. \quad (1.3)$$

Oraz hipotezę główną:

$$H_0 Tygodniowy\ czas\ spędzony\ na\ graniu\ wpływa\ dodatnio\ na\ wydatki\ na\ gry,$$
$$H_1: Tygodniowy\ czas\ spędzony\ na\ graniu\ wpływa\ ujemnie\ na\ wydatki\ na\ gry. \quad (1.4)$$

Dysponując danymi o preferencjach gatunku najczęściej grywanej gry, chcąc w projekcie również sprawdzić, czy typ gry wpływa na wydatki na nie, tj. Czy istnieje rodzaj gier, w które regularne granie powoduje istotne zwiększenie wydatków rocznych na gry sformułowano hipotezę pomocniczą:

$$H_0: Żaden\ rodzaj\ gry\ nie\ wpływa\ istotnie\ na\ wydatki\ na\ gry,$$
$$H_1: Conajmniej\ jeden\ rodzaj\ gry\ wpływa\ istotnie\ na\ wydatki\ na\ gry. \quad (1.5)$$

I hipotezę właściwą:

$$H_0: Conajmniej\ jeden\ rodzaj\ gry\ charakteryzują\ wyższe\ wydatki\ roczne$$
$$H_1: Wszystkie\ rodzaje\ gier\ charakteryzują\ się\ równymi\ wydatkami \quad (1.6)$$

Kolejnym celem projektu jest wyznaczenie jak najdokładniejszej prognozy wydatków rocznych na gry dla osoby z badanej populacji, o której specyfikacji można przeczytać w podrozdziale Specyfikacja Zbioru Danych.

# Wstęp teoretyczny i wybrane narzędzia

## Klasyczna Metoda Najmniejszych Kwadratów

Jest w modelu regresji wielorakiej zdefiniowana następująco [Maddala, 2006]:

Szacując model regresji z $k$ zmiennymi objaśniającymi postaci:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + e_i, \quad \text{dla i = 1, 2,..., n.}$$

Możemy zapisać go w postaci macierzowej jako:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \tag{2.1}$$

czyli:

$$y = X\beta + u,$$

gdzie:

- y – wektor obserwacji zmiennej objaśnianej o wymiarach $n \times 1$;
- X – macierz obserwacji zmiennych objaśniających o wymiarach $n \times k$;
- $\beta$ – wektor szacowanych parametrów o wymiarach $k \times 1$;
- u – wektor składników losowych o wymiarach $n \times 1$.

Następnie zakładam, że:

- Składniki losowe mają identyczne i niezależne rozkłady o wartości oczekiwanej 0 i wariancji $\sigma^2$,
- Zmienne $x$ są nielosowe i niezależne od składnika losowego,
- Zmienne $x$ są liniowo niezależne, zatem rząd $X^T X =$ rząd $X = k$. Przez to $(X^T X)^{-1}$ istnieje.

Przy powyższych założeniach, nieobciążony liniowy i o najmniejszej wariancji estymator $\beta$ otrzymamy minimalizując resztową sumę kwadratów (dowód twierdzenia Gaussa-Markowa):

$$Q = u^T u = (y - X\beta)^T (y - X\beta), \tag{2.2}$$

wymnażając równanie (2.2) otrzymujemy:

$$Q = y^T y - 2\beta^T X^T y + \beta^T X^T X \beta. \tag{2.3}$$

Następnie, aby zminimalizować Q, liczymy jego pochodną względem $\beta$ i przyrównujemy do zera otrzymując:

$$-2X^T y + 2X^T X\beta = 0, czyli\ \beta = (X^T X)^{-1} X^T y. \quad (2.4)$$

## Test Shapiro-Wilka

Niech $Y_1, \dots, Y_n$ będą niezależnymi obserwacjami definiowanymi przez dystrybuantę $F(\frac{y-\mu}{\sigma})$, gdzie $F, \mu \in \mathbb{R}$, a $\sigma > 0$ jest nieznana. Chcemy sprawdzić, czy dystrybuanta F, jest dystrybuantą rozkładu normalnego, zatem hipotezami testu Shapiro-Wilka są:

$$H_0: F = \Phi, \quad (2.5)$$

$$H_1: F \neq \Phi, \quad (2.6)$$

gdzie $\Phi$ to dystrybuanta rozkładu normalnego.

Podejście zaproponowane przez autorów testu [Shapiro, Wilk, 1965] opiera się na teście zgodności dwóch estymatorów $\sigma$:

$L_n = \sum_{i=1}^{n} a_{ni} Y_{n:i}$ – najlepszego liniowego estymatora,

$\hat{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ – najlepszego estymatora uzyskanego metodą największej wiarygodności.

gdzie:

$Y_{n:1} \leq \cdots \leq Y_{n:n}$ są ułożone rosnąco, a $a_n = (a_{n1}, \dots, a_{nn})' = \frac{V_n^{-1} M_n}{M_n^T V_n^{-1} M_n}$ i $\sum_{i=1}^{n} a_{ni} = 0$,

Gdzie $M_n = M$ jest wektorem wartości oczekiwanych uporządkowanych statystyk wielkości $n$ z standardowego rozkładu normalnego, a $V_n = V$, jest macierzy kowariancji. Z tego względu zaproponowano następujące przybliżenie $L_n$:

$$L_{n0} = \sum_{i=1}^{n} a_{ni,0} Y_{n:i} \approx \frac{1}{n} \sum_{i=1}^{n} \Phi^{-1}\left(\frac{i}{n+i}\right) Y_{n:i},$$

gdzie $a_{n0} = \frac{V^{-1} M}{(M^T V^{-1} V^{-1} M)^{\frac{1}{2}}}$.

Wtedy statystyka testowa jest postaci:

$$W_n = n(1 - \frac{L_{n0}^2}{\hat{\sigma}_n}), \quad (2.7)$$

i na jej podstawie rozstrzyga się hipotezy (2.5) i (2.6).

## Wartość p i poziom istotności

Jest to prawdopodobieństwo obliczone przy założeniu prawdziwości hipotezy zerowej testu, informujące o tym z jakim prawdopodobieństwem statystyka testowa $Z^*$, w zależności

od sformułowanych hipotez i rodzaju przeprowadzanego testu (jedno i dwustronny), jest większa, mniejsza bądź na moduł większa od statystyki $Z$, odczytanej dla odpowiedniej ilości stopni swobody i poziomu istotności wybranego rozkładu [Dodge, 2008]:

$$p = (Z \geq Z^*)$$

$$p = (Z \leq Z^*)$$

$$p = (Z \leq |Z^*|)$$

W projekcie przyjęto poziom istotności $\boldsymbol{\alpha = 0.05}$ i dla takiej wartości będą podejmowane decyzje na temat postawionych hipotez wykonywanych testów.

### Testy parametrów t-Studenta

Test t-Studenta równości parametrów zdefiniowany jest następująco [Dodge, 2008].

Spójrzmy na prosty model regresji liniowej:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{2.8}$$

Jeśli $\varepsilon_i$, to błąd losowy, mający rozkład normalny ze średnią $m = 0$ i wariancją $\sigma^2$. Wtedy estymator $\hat{\beta}_1$ zgodnie z właśnościami estymatrów KMNK, ma rozkład normalny ze średnią $m = \beta_1$ i wariancją $Var(\hat{\beta}_1) = \frac{\sigma^2}{RSS_i}$, o rozkładzie $\chi^2$ z $n-2$ $(n-k-1)$ stopniami swodoby. Wtedy staytstyka:

$$t = \frac{\hat{\beta}_1 - \beta_1}{Var(\hat{\beta}_1)}\sqrt{n}, \tag{2.9}$$

ma rozkład t-Studenta z $n-2$ $(n-k-1)$ stopniami swobody.

Test t-Studenta polega na obliczeniu wartości t z równania (2.9), i porównaniu jej z wartością teoretyczną zmiennej z tablic rozkładu t-Studenta.

Hipotezą zerową testu istotności parametrów t-Studenta jest:

$$H_0: \hat{\beta}_i = 0, \tag{2.10}$$

co oznacza, że $\hat{\beta}_i$ nie różni się istotnie od 0, a hipotezą alternatywną jest:

$$H_1: \hat{\beta}_i \neq 0, \tag{2.11}$$

co oznacza, że $\hat{\beta}_i$ jest istotnie różna od 0. Wtedy w miejsce $\beta_1$ w statystyce (2.9) wstawiamy 0. Hipotezy dla testu równości są analogiczne jak (2.10) i (2.11), tylko w miejsce zera wstawiamy $\beta_1$.

## Metoda Hellwiga

Metoda ta [Hellwig, 1969] mówi, że nośnikiem informacji o zmiennej objaśnianej jest potencjalna zmienna objaśniająca. Pojemnością takiego nośnika nazywamy:

$$h_{ij} = \frac{r_j^2}{1+\sum_{\substack{i=1 \\ i\neq j}}^{m_k} |r_{ij}|}, \text{ dla } k = 1, \dots, 2^n - 1, \text{ oraz } j = 1, \dots, m_k, \tag{2.12}$$

gdzie:

$r_j$ – to współczynnik korelacji liniowej między zmienną endogeniczną a j-tą zmienną objaśniającą,

$r_{ij}$ – to współczynnik korelacji między $i$-tą, a $j$-tą zmienna objaśniającą,

$k$ – to numer kombinacji zmiennych objaśniających $2^n - 1$,

$m_k$ – liczność podzbioru zmiennych objaśniających $k$-tej kombinacji.

Wtedy pojemność integralna kombinacji $k$ wyrażona jest jako:

$$H_k = \sum_{j-1}^{m_k} h_{kj}\,, k = 1, \dots, 2^m - 1.$$

Dla tak obliczonego $H_k$ dla każdej z kombinacji zmiennych objasniających wybieramy $H_k$ o wartości maksymalnej i do naszego modelu dodajemy zmienne z $k$-tej kombinacji.

## Metoda krokowa wsteczna

Polega na stopniowym usuwaniu nieistotnych statystycznie zmiennych z utworzonego modelu zgodnie z poniższym algorytmem [Vu, Muttaqi, Agalgaonkar, 2015]:

1. Wykonuję pojedyncze testy t-Studenta istotności parametrów w modelu.
2. Zgodnie z hipotezą (2.2) im większa wartość p dla tego testu, tym większe prawdopodobieństwo, że parametr jest równy zero. W związku z tym spośród wszystkich parametrów modelu, wybieram ten, który najmniej istotnie różni się od zera, czyli ma największą wartość p.
3. Usuwam ten parametr z modelu i wracam do kroku 1.

Adnotacja: Należy również pamiętać, aby sprawdzać założenie dotyczące testu t-Studenta w każdym przejściu regresji krokowej, tylko przy założeniu normalności reszt modelu można stosować statystykę testową (2.9).

### Współczynnik korelacji liniowej Pearsona

Korelacja pomiędzy zmiennymi $X$ oraz $Y$ wyrażona współczynnikiem Pearsona dana jest wzorem [Buda, Jarynowski, 2010]:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

czyli:

$$r_{XY} = \frac{cov(X,Y)}{\sigma_Y \sigma_X}, \tag{2.13}$$

gdzie $\sigma_X$ to odchylenie standardowe $X$, $\sigma_{XY}$ to odchylenie standardowe $Y$, a $cov(X,Y)$ to kowariancja $X$ i $Y$.

Można również przeprowadzić test istotności tego współczynnika z hipotezami [Cohen, 1988]:

$$H_0: \rho = 0, \tag{2.14}$$

$$H_0: \rho \neq 0. \tag{2.15}$$

Statystyka t-Studenta ma postać:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \tag{2.16}$$

I rozkład t-Studenta o $n-2$ stopniach swobody. Na jej podstawie rozstrzygamy hipotezy (2.14) i (2.15), porównując ją z wartościami krytycznymi dla rozkładu t-Studenta.

### Współczynnik determinacji $R^2$ i $\bar{R}^2$

Jest to miara jakości dopasowania modelu do danych, czyli w jakim stopniu nasz model opisuje zmienną objaśnianą [Glantz, Stanton, 1990]. Dany jest on wzorem:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(\bar{y}-y_i)^2}, \qquad R^2 = 1 - \frac{ESS}{TSS}, \tag{2.17}$$

gdzie:

- $y_i$ to $i$-ta wartość empiryczna zmiennej $y$,
- $\hat{y}_i$ to $i$-ta wartość teorytyczna zmiennej $y$,
- $\bar{y}$ to średnia arytmetyczna wartości empirycznych zmiennej $y$.
- $ESS$ to suma kwadratów błędów (error sum of squares),
- $TSS$ to całkowita suma kwadratów (total sum of squares).

Oprócz zwykłego współczynnika determinacji, w projekcie również wystąpi wartość $\bar{R}^2$. Ponieważ zwykły $R^2$ nie uwzględnia stopni swobody występujących w modelu, porównanie go z $\bar{R}^2$, daje informacje o możliwym uwzględnieniu za dużej ilości zmiennych objaśniających w porównaniu do ilości obserwacji dla modelu. Współczynnik $\bar{R}^2$ dany jest wzorem:

$$\bar{R}^2 = R^2 - \frac{k}{n-k-1}(1 - R^2), \tag{2.17.1}$$

gdzie $k$ to ilość zmiennych objaśniających w modelu, a $n$ to ilość obserwacji.

### Uogólniony Test Walda

Wykonujemy, gdy chcemy sprawdzić hipotezę o istotności wielu parametrów, przy założeniu, że reszty modelu mają rozkład normalny:

$$H_0: \alpha_0 = \cdots = \alpha_k = 0, \tag{2.18}$$

$$H_1: \alpha_0 \neq \cdots \neq \alpha_k \neq 0, \tag{2.19}$$

gdzie hipoteza (2.19) mówi o tym, że co najmniej jeden z testowanych parametrów, jest różny od 0. Statystyka testowa dana jest wzorem [Maddala, 2006]:

$$F = \frac{R^2(n-k-1)}{(1-R^2)k}, \tag{2.20}$$

gdzie:

- $R^2$ – współczynnik determinacji modelu,
- $k$ – ilość zmiennych w modelu
- $n$ – ilość obserwacji.

Statystyka ta ma rozkład F-Snedecora z $k$ i $n - k - 1$ stopniami swobody i na jej podstawie wnioskujemy o prawdziwości hipotez (2.18) i (2.19).

Uogólnionym testem Walda można także stwierdzić czy dodanie zmiennych do modelu jest istotne statystycznie, niech model podstawowy będzie dany jako:

$$y = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_k x_k + \varepsilon, \tag{2.17.1}$$

a model z rozszerzony jako:

$$y = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_k x_k + \alpha_{k+1} x_{k+1} + \cdots + \alpha_{k+m} x_{k+m} + \varepsilon. \tag{2.17.2}$$

Wtedy hipotezami testu będą:

$$H_0: \alpha_{k+1} = \cdots = \alpha_{k+m} = 0, \tag{2.18.1}$$

$$H_1: \textit{co najmniej jeden z dodawanych parametrów jest istotnie różny od } 0. \tag{2.19.1}$$

Statystyka testowa $F$ jest postaci:

$$F = \frac{PESS-RESS}{RESS} * \frac{n-k-1-m}{m}, \tag{2.20.1}$$

i ma rozkład F-Snedecora z $m$ i $n-k-m-1$ stopniami swobody, gdzie:

- $m$ to ilość zmiennych dodanych do modelu,
- $PESS$ to suma kwadratów błędów modelu podstawowego,
- $RESS$ to suma kwadratów błędów modelu z dodanymi zmiennymi.

Na jej podstawie wnioskujemy w sprawie hipotez (2.18.1) i (2.19.1).

### Współczynnik VIF jako miara współliniowości

Jest on bezpośrednio związany z współczynnikiem determinacji, bowiem:

$$VIF = \frac{1}{1-R_i^2}, \tag{2.21}$$

gdzie $R_i^2$ to współczynnik determinacji obliczony dla modelu pomocniczego, objaśniającego $i$-tą zmienną objasniającą z głównego modelu, przy pomocy pozostałych zmiennych objaśniających z modelu głównego.

W literaturze stosuje się interpretacje wartości $VIF > 10$ jako współliniowość, bądź [Sheather, 2009] $VIF > 5$ jako oznakę współliniowości. Ja w mojej pracy będę używał **pierwszego z tych warunków**.

### Efekt katalizy

Informacja uzyskana dzięki współczynnikowi determinacji $R^2$ może być fałszywa, jeśli w modelu występują zmienne nazywane katalizatorami. Efektem katalizy nazywamy więc uzyskanie wysokiej wartości współczynnika determinacji, pomimo tego, że powiązanie zmiennej objaśnianej ze zmiennymi objaśniającymi nie uzasadniają takiego wyniku [Gruszczyński, Kuszewski, Podgórska, 2009].

Załóżmy, że $R$ i $R_0$ to regularna para korelacyjna postaci [Hellwig 1976]:

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix} \text{ i } R_0 = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix},$$

gdzie:

- $r_{ij}$ – współczynnik korelacji pomiędzy zmiennymi $X_i$ i $X_j$,
- $r_j$ – współczynnik korelacji między $X_j$, a $Y$.

Para ta musi spełniać warunek:

$$0 \leq r_1 \leq r_2 \leq \ldots \leq r_k,$$

Wtedy spośród pary zmiennych $X_i$ i $X_j$, katalizatorem jest $X_i$ $(i < j)$ jeżeli:

$$r_{ij} < 0 \; lub \; r_{ij} > \frac{r_i}{r_j}. \tag{2.22}$$

Natężenie katalizy w modelu dane jest wzorem:

$$\eta = R^2 - H, \tag{2.23}$$

gdzie $R^2$ to współczynnik determinacji modelu, a H to integralna pojemnośc podzbioru zmiennych objasniających w sensie Hellwiga.

## Przedziały ufności

Koncepcja przedziału ufności to po prostu stworzenie przedziału przyjmowanych wartości dla estymatora z danym prawdopodobieństwem. Przedział ten zawierać będzie wartość rzeczywistą parametru w opisie populacji [Dodge, 2008].

Aby skonstruować przedział ufności parametru $\beta$ z danym prawdopodobieństwem, trzeba rozwiązać następujące równanie:

$$P(L_i \leq \beta \leq L_s) = 1 - \alpha, \tag{2.24}$$

gdzie:

- $\alpha$ to przyjęty poziom istotności, a $1 - \alpha$ to poziom ufności,
- $\beta$ to estymowany parametr,
- $L_i$ to dolny kraniec przedziału,
- $L_s$ to górny kraniec przedziału.

Aby rozwiązać to równanie, należy skonstruować funkcję $f(t, \beta)$, wtedy równanie (2.24) można zapisać jako:

$$P(k_1 \leq f(t, \beta) \leq k_2) = 1 - \alpha, \tag{2.25}$$

gdzie $k_1$ i $k_2$ są dane przez dystrybuantę funkcji $f(t, \beta)$.

Przedziały ufności dla parametrów szacowanych Klasyczną metoda najmniejszych kwadratów można obliczyć jako [Maddala, 2006]:

Używając statystyki (2.9) tworze równanie (2.25) korzystając z wartości krytycznych dla rozkładu t-Studenta dla przykładu zmiennej $t$ z 8 stopniami swobody na poziomie istotności $\alpha = 0{,}05$:

$$P\left(-2{,}306 \leq \frac{\hat{\beta}_1-\beta_1}{Var(\hat{\beta}_1)}\sqrt{n} \leq 2{,}306\right) = 0{,}95.$$

Ponieważ w teście istotności parametru $\beta_1 = 0$, a $SE(\hat{\beta}_1)$ - błąd standardowy $\hat{\beta}_1 = \frac{Var(\hat{\beta}_1)}{\sqrt{n}}$, zatem po przekształceniach przedział ufności parametru $\hat{\beta}_1$ dany jest jako:

$$P\left(-2{,}306 * SE(\hat{\beta}_1) \leq \hat{\beta}_1 \leq 2{,}306 * SE(\hat{\beta}_1)\right) = 0{,}95,$$

zatem z 95% prawdopodobieństwem można powiedzieć, że rzeczywista wartość:

$$\hat{\beta}_1 \in [-2{,}306 * SE(\hat{\beta}_1)\ ;\ 2{,}306 * SE(\hat{\beta}_1)].$$

### Kryterium informacyjne Bayesa-Shwarza

Jest to kryterium umożliwiające porównywanie modeli i wybór najlepszego spośród nich. Najlepszym modelem jest ten o najmniejszej wartości BIC, przy:

$$BIC = -2\ln(L) + kln(n), \tag{2.26}$$

gdzie $L$ to zmaksymalizowana wartość funkcji wiarydogności estymowanego modelu.

### Testowanie stabilności modelu

Test Ramseya RESET:

Do testowania stabilności postaci analitycznej modelu będę wykorzystywał test Ramseya, zwany także testem RESET (Regression Specification Error Test). Polega on na oszacowaniu modelu regresji błędów $y_i$ względem wszystkich zmiennych, dodając $\hat{y}_i^2, \hat{y}_i^3$, itd. Następnie testowaniu, czy parametry tego modelu są istotnie różne od zera wykorzystując przy tym wspomniany wcześniej test istotności parametrów Walda [Maddala 2006], czyli sprawdzając czy przyrost $R^2$ jest istotny.

Hipotezami testu RESET są:

$$H_0: Postać\ modelu\ jest\ dobrze\ dobrana, \tag{2.27}$$

$$H_1: Postać\ modelu\ jest\ źle\ dobrana. \tag{2.28}$$

I wnioskujemy w ich sprawie na podstawie statystyki F dla testu Walda (2.20).

Test Chowa [Maddala 2006]:

Jest to predykcyjny test stabilności. Mianowicie, korzystamy z $n_1$ pierwszych obserwacji do oszacowania modelu regresji i tak oszacowany model używamy, aby sporządzić prognozy dla pozostałych $n_2$ obserwacji. Hipotezy testu Chowa:

$$H_0: Postać\ modelu\ jest\ stabilna, \tag{2.29}$$

$$H_1: Postać\ modelu\ nie\ jest\ stabilna. \tag{2.29.1}$$

Wniosek podejmujemy na podstawie statystyki testowej $F$:

$$F = \frac{\frac{RESS-UESS}{k+1}}{\frac{UESS}{n_1-n_2-2k-2}}, \tag{2.30}$$

gdzie:

- RESS – to suma kwadratów błędów w modelu oszacowanym z wykorzystaniem wszystkich obserwacji$(n_1 + n_2)$,
- UESS – to suma sum kwadratów błędów w modelach oszacowanym tylko dla $n_1$ i oszacowanym tylko dla $n_2$ obserwacji,
- $k$ – ilośc zmiennych objaśniających w modelu.

Test serii:

Nazywany także testem Walda-Wolfowitza, polega na podziale reszt modelu na serie wartości dodatnich, i wartości ujemnych, po uporządkowaniu ich według rosnących wartości wybranej zmiennej objaśniającej. Przykładowo, jeśli błędami modelu są:

$$0{,}5;\ 0{,}7; -0{,}6; 0{,}7; 0{,}8; -0{,}2; -1{,}2;$$

i założymy, że wartość wybranej zmiennej objaśniającej odpowiadającej tym resztą rośnie, to znaki reszt wyglądają następująco: $++-+++--$ czyli mamy 4 serie, 2 dodatnie i 2 ujemne. Hipotezami testu serii są:

$$H_0: Postać\ modelu\ jest\ dobrze\ dobrana, \tag{2.31}$$

$$H_1: Postać\ modelu\ jest\ źle\ dobrana. \tag{2.31.1}$$

Wniosek podejmujemy na podstawie liczby wszystkich serii, liczby serii dodatnich i liczby serii ujemnych, odczytując wartości z tablic rozkładu liczby serii.

## Testowanie heteroskedastyczności

W celu wykrywania heteroskedastyczności w modelu wykorzystam test White'a. Polega on na oszacowaniu modelu regresji kwadratów błędów w wyjściowym modelu względem wszystkich zmiennych objaśniających, ich kwadratów i iloczynów [Maddala 2008]. Na przykład dla modelu wyjściowego:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + u.$$

Szacowany model pomocniczy będzie więc postaci:

$$\hat{u}_i^2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2. \quad (2.32)$$

Hipotezy testu White'a:

$$H_0: W\ modelu\ \ występuje\ homoskedastyczność\ składnika\ losowego, \quad (2.33)$$

$$H_1: W\ modelu\ występuje\ heteroskedastyczność\ składnika\ losowego \quad (2.33.1)$$

Następnie obliczam statystykę $nR^2$ dla modelu (2.32) mającą rozkład $\chi^2$ z $\frac{k(k+1)}{2}$ stopniami swobody. I na jej podstawie wnioskuję w sprawie hipotez (2.33) o (2.33.1).

## Graficzne przedstawienie danych

Zostało wykonane przy użyciu wbudowanych funkcji programu gretl, oraz przy użyciu języka R, z wykorzystaniem bibliotek ggplot, ggpubr i psych

## Stosowanie wybranych metod statystycznych

Zostało wykonane przy użyciu skryptów programu gretl.

# Specyfikacja zbioru danych

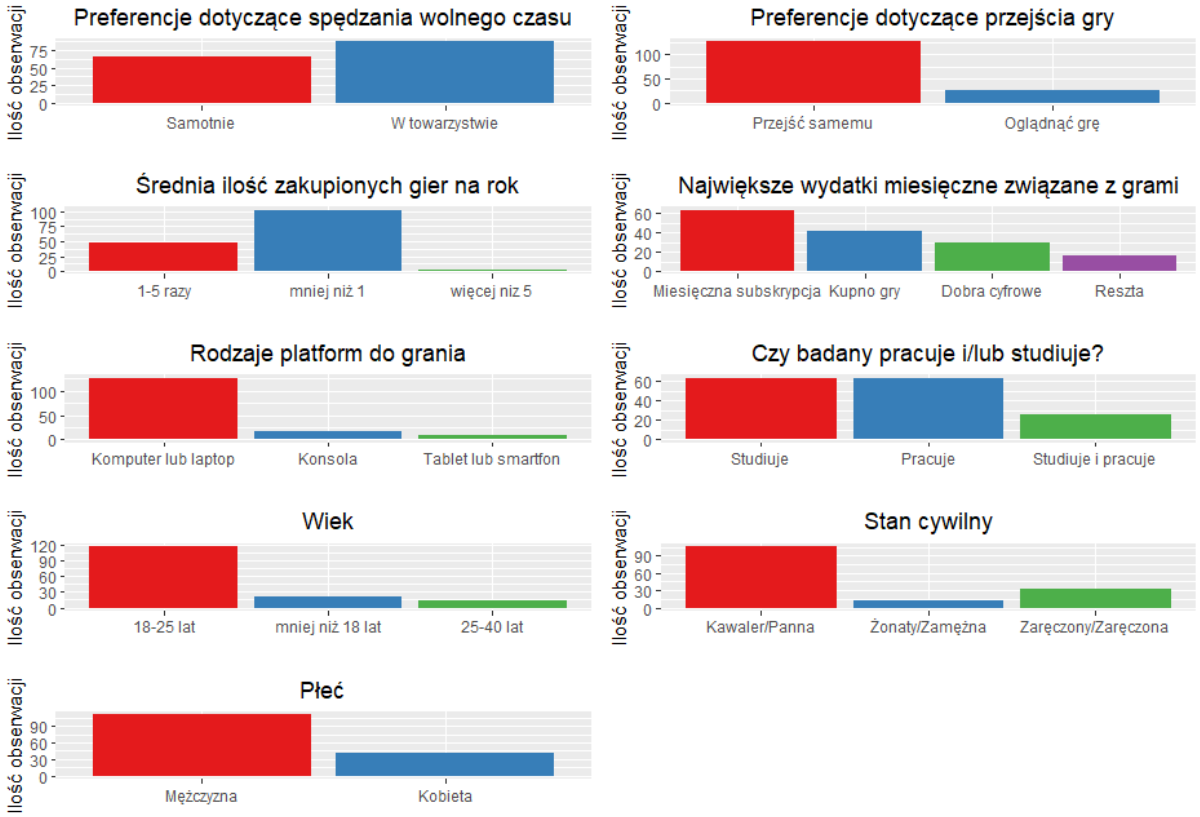
## Wstępne przedstawienie danych

Dane zostały zebrane samodzielnie i pochodzą z ankiety, dotyczącej wydatków na gry, opublikowanej na dwóch grupach internetowych: World of Warcraft Polska oraz Dota 2 Polska. Badaną populacją są zatem uczestnicy powyższych forów internetowych. W ankiecie znajdowało się 14 pytań mających na celu zebrać cechy charakteryzujące tą populację. Próbą, którą udało się uzyskać, są odpowiedzi od 152 osób, które ze względu na losową technikę wyboru osób do ankiety, można uznać za próbę reprezentatywną badanej populacji [Szreder, 2010]. W wynikach ankiety zebrano następujące zmienne dyskretne:

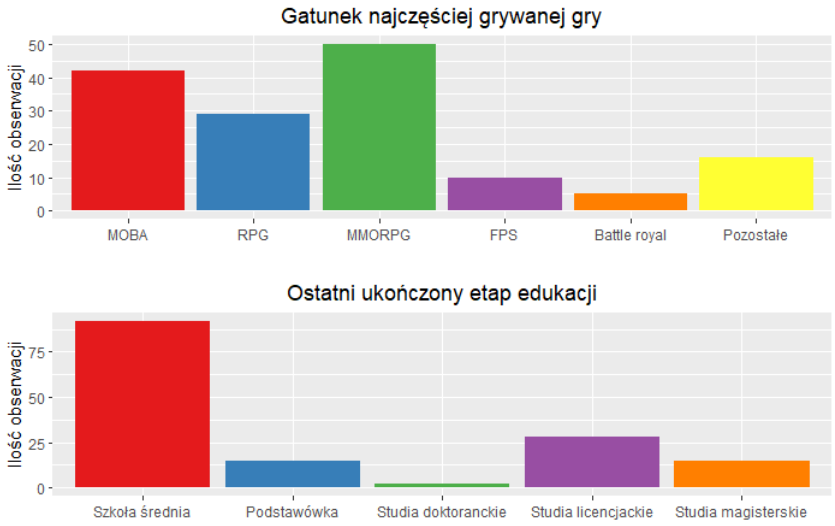
- KM – zmienna informująca o płci w grupach: Kobieta, Mężczyzna,
- SC – zmienna informująca o stanie cywilnym w grupach: kawaler/panna, żonaty/zamężna, zaręczony/zaręczona,
- WIEK – zmienna informująca o wieku w postaci przedziałów, mniej niż 18 lat, od 18-tu do 25-ciu lat, od 25-ciu do 40-tu lat oraz więcej niż 40 lat,
- EDU – zmienna informująca o ostatnim ukończonym etapie edukacji, w kategoriach: szkoła podstawowa, szkoła średnia, studia licencjackie/inżynierskie, studia magisterskie oraz studia doktoranckie,
- STPR - Informacja o osobach uczących się i pracujących w postaci rozdziału na pracujących, uczących się oraz pracujących i jednocześnie uczących się,
- GRA – Informacja o gatunku gry, w który najczęściej gra badana osoba z podgrupami: MOBA, RPG, MMORPG, FPS, Battle royal oraz pozostałymi gatunkami,
- PCS – informacja o platformie, na której najczęściej grają badani w podziale na: komputer lub laptop, konsolę, oraz smartfon i tablet,
- MAXW – zmienna informująca o przedmiocie największych średnich miesięcznych wydatków dotyczących gier w kategoriach: miesięcznej subskrypcji, czyli cyklicznej płatności co miesiąc, na przykład za możliwość grania, kupna gry, dóbr cyfrowych, czyli przedmiotów wirtualnych, które po zapłacie otrzymujemy w grze, oraz innych płatności niewymienionych powyżej,
- KUPNO – zmienna informująca o tym jak często badani decydują się na zakup gry podczas jednego roku kalendarzowego, w kategoriach: rzadziej niż 1 raz w roku, od 1-go do 5-ciu razy w roku oraz częściej niż 5 razy w roku,
- SAM – zmienna informująca o preferencjach badanego dotyczących gry w rozdziale na kategorie: wolę przejść grę sam oraz wolę oglądnąć grę w serwisie streamingowym,

- SAMOTNI – zmienna informująca o preferencjach badanego dotyczących spędzania wolnego czasu w kategoriach: samotnie oraz w towarzystwie.

Na Rysunkach 2 i 3, można zobaczyć wykresy słupkowe pokazujące jak ilościowo prezentują się odpowiedzi w poszczególnych kategoriach.




*Rysunek 2 - Wykresy kolumnowe dla odpowiedzi uzyskanych w ankiecie, opracowanie własne*




*Rysunek 3 - Wykresy kolumnowe dla odpowiedzi uzyskanych w ankiecie cd. opracowanie własne*

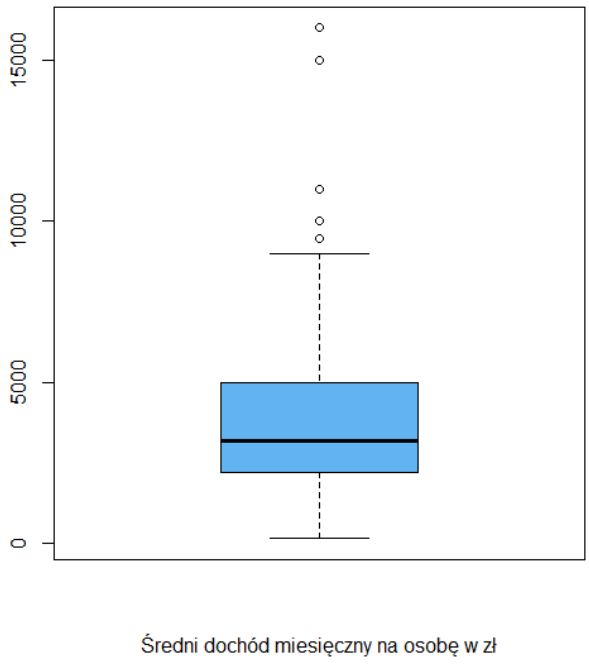
Z wykresów na rysunkach 2 i 3, można odczytać, że w badanej próbie najwięcej jest osób, które są mężczyznami i wolą spędzać czas w towarzystwie, przechodzić gry samemu, kupują gry mniej niż raz w roku, najwięcej wydają na płatności cykliczne typu abonament i subskrypcja oraz grają na komputerze bądź laptopie. Znaczna większość badanych nie jest zaręczona ani w związku, jest w wieku 18 do 25 lat, ukończyła szkołę średnią, a ich ulubionym gatunkiem gier jest MMORPG. Osób pracujących w próbie jest mniej więcej tyle samo co studiujących lub uczących się, a osób, które na raz studiują i pracują jest znacznie mniej.

W ramach ankiety zebrano także zmienne, które można potraktować jako ciągłe ze względu na to, że mają ponad 10 możliwych odpowiedzi, można je przedstawić na skali ilościowej oraz przedstawienie ich w postaci ułamków ma sens [Mishra i Pandey, 2018 Oct-Dec] – na przykład średni czas spędzany tygodniowo na graniu o wartości 20,5 godziny. Są to zmienne:

- CZAS – zmienna informująca o średnim tygodniowym czasie spędzonym na graniu wyrażonym w godzinach,
- DOCHOD – zmienna informująca o średnim dochodzie na osobę w gospodarstwie domowym wyrażonym w zł,
- WYDATKI – zmienna informująca o średnich rocznych wydatkach na gry wyrażonych w zł.

| | |
|---|---|
| Średnia | 3994,90 |
| Mediana | 3200,00 |
| Wartość minimalna | 150,00 |
| Wartość maksymalna | 16000,00 |
| Odchylenie standardowe | 2577,60 |
| Współczynnik zmienności | 0,65 |
| Skośność | 2,00 |
| Kurtoza | 5,33 |
| Percentyl 5% | 1360,00 |
| Percentyl 95% | 9322,00 |
| Zakres Q3-Q1 | 2800,00 |
| Brakujące obserwacje | 27 |

*Tabela 1 - Statystyki opisowe zmiennej DOCHOD, opracowanie własne*



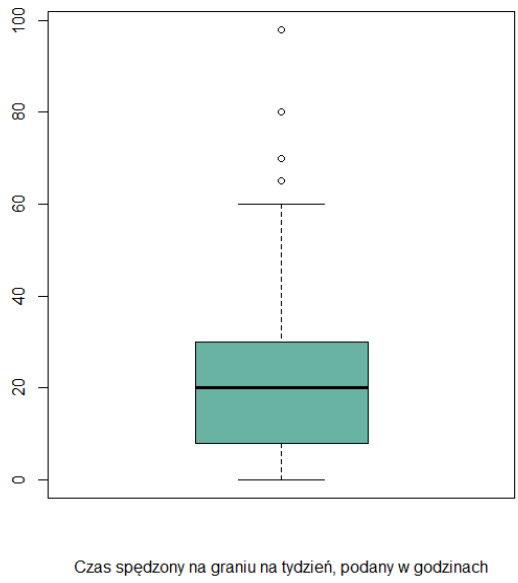

*Wykres 1 - wykres pudełkowy z wąsem zmiennej DOCHOD, opracowanie własne*

Na podstawie Tabeli 1 i Wykresu 1, można więc stwierdzić, że średni dochód na osobę w gospodarstwie domowym badanych wynosi 3994,90 zł. Co najmniej 50% badanych prezentuje średni dochód na osobę wyższy niż 3200,00 zł, a co najwyżej 5% badanych ma odpowiednio: średni dochód na osobę wyższy niż 9322,00 zł i mniejszy niż 1360,00 zł. W skrajnych przypadkach, badani należą do gospodarstw domowych o dochodach 150,00 zł oraz 16000,00 zł na osobę. Wartości zmiennej są oddalone od średniej o średnio 2577,60 zł i zmienną cechuje silna zmienność. Skośność mówi o tym, że rozkład zmiennej jest prawostronnie skośny, co można zobaczyć na Rysunek 4. Wartości średniego dochodu na głowę są także mocno skoncentrowane wokół średniej o czym świadczy wysoka kurtoza. Ponieważ pytanie o dochód w ankiecie było oznaczone jako nieobowiązkowe, występuje tutaj 27 brakujących danych.

Statystyki opisowe dla zmiennej CZAS:

| | |
|---|---|
| Średnia | 21,82 |
| Mediana | 20,00 |
| Wartość minimalna | 0,00 |
| Wartość maksymalna | 98,00 |
| Odchylenie standardowe | 17,92 |
| Współczynnik zmienności | 0,82 |
| Skośność | 1,32 |
| Kurtoza | 2,23 |
| Percentyl 5% | 1,65 |
| Percentyl 95% | 55,35 |
| Zakres Q3-Q1 | 22,00 |
| Brakujące obserwacje | 0 |

*Tabela 2 - Statystyki opisowe dla zmiennej CZAS, opracowanie własne*




*Wykres 2 - wykres pudełkowy z wąsem dla zmiennej CZAS, opracowanie własne*

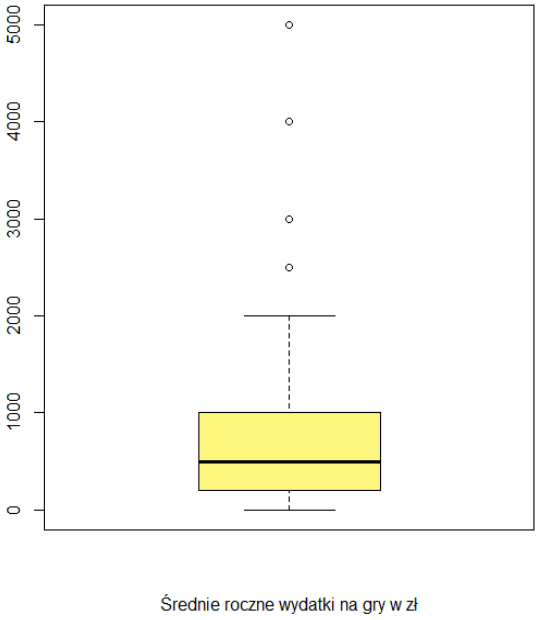
Na podstawie Tabeli 2 i Wykresu 2, można więc stwierdzić, że średni czas poświęcany tygodniowo na granie w badanej próbie wynosi 21,82 godziny. Co najmniej 50% badanych poświęca na granie tygodniowo więcej niż 20,00 godzin, a co najwyżej 5% badanych poświęca na granie tygodniowo odpowiednio: więcej niż 55,35 godzin i mniej niż 1,65 godziny. W skrajnych przypadkach, badani w ogóle nie przeznaczają czasu na granie w tygodniu lub grają nawet 98,00 godzin. Wartości zmiennej są oddalone od średniej arytmetycznej o średnio 17,92 godziny, a zmienną cechuje silna zmienność. Skośność mówi o tym, że rozkład zmiennej jest prawostronnie skośny, co można zobaczyć na Rysunek 4. Wartości średniego dochodu

na głowę są także skoncentrowane wokół średniej o czym świadczy dodatnia kurtoza. Nie występują braki w danych.

Statystyki opisowe zmiennej WYDATKI:

| | |
|---|---|
| Średnia | 695,00 |
| Mediana | 500,00 |
| Wartość minimalna | 0,00 |
| Wartość maksymalna | 5000,00 |
| Odchylenie standardowe | 796,04 |
| Współczynnik zmienności | 1,15 |
| Skośność | 2,92 |
| Kurtoza | 11,50 |
| Percentyl 5% | 0,00 |
| Percentyl 95% | 2000,00 |
| Zakres Q3-Q1 | 800,00 |
| Brakujące obserwacje | 0 |

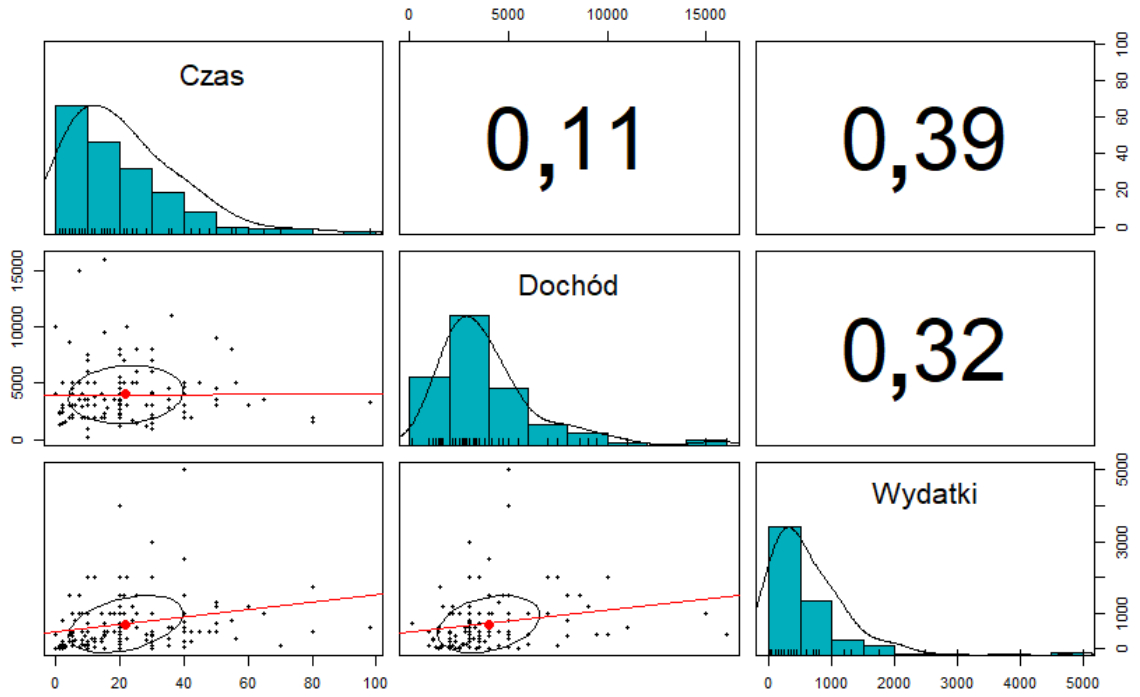
*Tabela 3 - statystyki opisowe dla zmiennej WYDATKI, opracowanie własne*




*Wykres 3 - wykres pudełkowy z wąsem dla zmiennej WYDATKI, opracowanie własne*

Na podstawie Tabeli 3 i Wykresu 3, można więc stwierdzić, że średnie roczne wydatki na gry w badanej próbie wynoszą 695,00 zł. Co najmniej 50% badanych wydaje rocznie na gry więcej niż 500,00 zł, a co najwyżej 5% badanych przeznacza rocznie na gry odpowiednio: więcej niż 2000,00 zł i nie przeznacza nic. W skrajnych przypadkach, badani w ogóle nie przeznaczają pieniędzy na gry w skali roku lub wydają nawet 5000,00 zł. Wartości zmiennej są oddalone od średniej arytmetycznej o średnio 796,04 zł, a zmienną cechuje bardzo silna zmienność. Skośność mówi o tym, że rozkład zmiennej jest prawostronnie skośny, co można zobaczyć na Rysunek 4. Wartości średniego dochodu na głowę są także bardzo mocno skoncentrowane wokół średniej o czym świadczy dodatnia kurtoza. Nie występują braki w danych.

Popatrzmy teraz na wybrane wykresy i zależności między zmiennymi ciągłymi, czyli na Rysunek 4.




*Rysunek 4 - wybrane wykresy i cechy zmiennych CZAS, DOCHOD i WYDATKI*

Na przekątnej rysunku, znajdują się histogramy zmiennych wraz z dopasowanymi funkcjami gęstości. Można na nich zobaczyć wspomniane wcześniej prawostronne skośności rozkładów zmiennych DOCHOD, CZAS i WYDATKI.

Pod przekątną, znajdują się wykresy rozrzutu zmiennych, z dopasowanymi liniami regresji. Można już teraz postawić hipotezę o tym, że spłaszczenie tych rozrzutów, na przykład dodając logarytmy zmiennych ciągłych, da nam lepsze wyniki w przypadku dopasowania modelu klasyczną metodą najmniejszych kwadratów oraz własności prognostycznych. W dodatku, przy założeniu, że zmienne mają dwuwymiarowy rozkład normalny, na wykresie zaznaczono także 95% elipsy ufności, oznaczające, że obserwacje z populacji będą znajdywać się w tym obszarze z prawdopodobieństwem 95%.

Nad przekątną, znajdują się współczynniki korelacji liniowej Pearsona między zmiennymi leżącymi na przecięciu rzędu i kolumny – to znaczy na przykład, że 0,39 jest współczynnikiem korelacji liniowej Pearsona dla zmiennych WYDATKI i CZAS, czyli zmienne są ze sobą słabo skorelowane. Z kolei prawie korelacji między zmiennymi DOCHOD I CZAS.

## Przygotowanie danych do badania

Ponieważ większa część danych jest kategoryczna, to aby włączyć je do modelu, potrzebuję rozdzielić je na zmienne binarne. Będę przy tym pamiętać, aby jeśli rozdzielam zmienną SAMOTNI, na zmienne binarne postaci:

- WSAM – 1 tam, gdzie SAMOTNI przyjmuje wartość „samotnie”, 0 dla innych obserwacji,
- WTOW – 1 tam, gdzie SAMOTNI przyjmuje wartość „w towarzystwie”, 0 dla innych obserwacji.

To do modelu włączę jedynie jedną z tych zmiennych, ze względu na współliniowość zmiennych WSAM i WTOW ($WSAM + WTOW = 1$). Poniżej przedstawiam zatem tabelę przekształconych odpowiednio zmiennych:

| | Wartość 1, gdy |
|---|---|
| KM | Mężczyzna |
| W18_25 | Wiek w przedziale 18-25 lat |
| W25_40 | Wiek w przedziale 25-40 lat |
| KAWALER | Stan cywilny to kawaler/panna |
| ZARECZONY | Stan cywilny to zaręczony/zaręczona |
| SS | Ostatni ukończony etap edukacji to szkoła średnia |
| P | Ostatni ukończony etap edukacji to podstawówka |
| SL | Ostatni ukończony etap edukacji to studia licencjackie/inżynierskie |
| SM | Ostatni ukończony etap edukacji to studia magisterskie |
| US | Badany uczy się lub studiuje |
| P | Badany pracuje |
| RPG | Najczęściej grywany gatunek gier to RPG |
| MOBA | Najczęściej grywany gatunek gier to MOBA |
| FPS | Najczęściej grywany gatunek gier to FPS |
| MMORPG | Najczęściej grywany gatunek gier to MMORPG |
| PC | Platforma, na której gra badany najczęściej to komputer |
| KONSOLA | Platforma, na której gra badany najczęściej to konsola |
| KUPNO | Badany najwięcej wydatków związanych z grami przeznacza na zakup gry |
| MIKRO | Badany najwięcej wydatków związanych z grami przeznacza na mikropłatności w grach |
| SUB | Badany najwięcej wydatków związanych z grami przeznacza na miesięczną subskrypcję |
| SAM | Badany woli sam przejść grę |
| K1_5 | Badany kupuje gry od 1 do 5 razy w roku |
| K_1 | Badany kupuje mniej niż jedną grę rocznie |
| WSAM | Badany woli spędzać czas samotnie |

*Tabela 4 - Przekształcenia zmiennych w celu implementacji w modelu*

| | Wartość 0, gdy |
|---|---|
| KM | Kobieta |
| W18_25 | Wiek w innych przedziałach |
| W25_40 | Wiek w innych przedziałach |
| KAWALER | Stan cywilny jest inny |
| ZARECZONY | Stan cywilny jest inny |
| SS | Ostatni ukończony etap edukacji to nie szkoła średnia |
| P | Ostatni ukończony etap edukacji to nie podstawówka |
| SL | Ostatni ukończony etap edukacji to nie studia licencjackie/inżynierskie |
| SM | Ostatni ukończony etap edukacji to nie studia magisterskie |
| US | Badany nie tylko uczy się lub studiuje, bądź nie uczy się i nie studiuje |
| P | Badany nie pracuje bądź nie tylko pracuje |
| RPG | Najczęściej grywany gatunek gier jest inny |
| MOBA | Najczęściej grywany gatunek gier jest inny |
| FPS | Najczęściej grywany gatunek gier jest inny |
| MMORPG | Najczęściej grywany gatunek gier jest inny |
| PC | Platforma, na której gra badany najczęściej jest inna |
| KONSOLA | Platforma, na której gra badany najczęściej jest inna |
| KUPNO | Badany najwięcej wydatków związanych z grami przeznacza na co innego |
| MIKRO | Badany najwięcej wydatków związanych z grami przeznacza na co innego |
| SUB | Badany najwięcej wydatków związanych z grami przeznacza na co innego |
| SAM | Badany woli obejrzeć grę w serwisie streamingowym |
| K1_5 | Badany kupuje gry częściej lub rzadziej |
| K_1 | Badany kupuje więcej niż jedną grę rocznie |
| WSAM | Badany woli spędzać czas w towarzystwie |

*Tabela 5 - Przekształcenia zmiennych w celu implementacji w modelu c.d.*

Zmienne ciągłe DOCHOD, CZAS i WYDATKI można by bez większych zmian włączyć do modelu, usuwając wybrakowane obserwacje przy zmiennej dochód, jednak patrząc na rysunek 4, na histogramy i funkcje gęstości powyższych zmiennych można zauważyć, że są one bardzo podobne do rozkładu log-normalnego. Zatem już teraz można wysunąć hipotezę, że logarytmy tych zmiennych, przy założeniu prawdziwości poprzedniego zdania, będą miały w przybliżeniu niezależne rozkłady normalne, co może wpłynąć na normalność reszt modelu. Dodaję więc do zbioru danych sztuczne zmienne LDOCHOD, LCZAS i LWYDATKI, będące logarytmami naturalnymi zmiennych DOCHOD, CZAS i WYDATKI.

Patrząc jednak na wykres 2 i 3 i sąsiadujące z nimi tabele, widać, że zmienne CZAS i WYDATKI przyjmują zerowe wartości, co po zlogarytmowaniu da wartość $-\infty$. Takich wartości nie można szacować KMNK ze względu na jej założenia, zatem zawężam populację podaną we wstępie, tylko do osób z forów internetowych World of Warcraft Polska i Dota2 Polska, które poświęcają tygodniowo jakikolwiek różny od zera czas na granie, i których wydatki roczne na gry są również większe niż zero.

Zdecydowałem się także na pominięcie obserwacji z wybrakowanymi danymi dotyczącymi dochodu, gdyż uważam, że inne podejście, np. prognozowanie wartości tej zmiennej, czy użycie średniej ruchomej mogłoby zakłócić sens wprowadzania tej zmiennej do modelu.

Ostatecznie zatem przyjmując, że dane są ułożone losowo, bo ankiety zostały opublikowane na obu grupach w tym samym czasie, zaczynam tworzyć model na podstawie około 90% dostępnych obserwacji (105), mając do dyspozycji 30 zmiennych.

# Wstępna postać modelu

## Model początkowy

Zmienną zależną jest zmienna WYDATKI. W związku z tym na podstawie równania (2.4) wyznaczam klasyczną metoda najmniejszych kwadratów parametry modelu uwzględniając 26 zmiennych objaśniających (bez logarytmów), oraz stałą, Wyniki oszacowań zamieszczono w tabeli 6.

| zmienna | współczynnik | błąd standardowy |
|---|---|---|
| stała | −397,12 | 932,48 |
| KM | −53,73 | 176,03 |
| W18_25 | 62,12 | 338,41 |
| W25_40 | −78,95 | 432,02 |
| KAWALER | 161,87 | 215,33 |
| ZARECZONY | 244,94 | 285,74 |
| SS | −92,71 | 629,42 |
| P | −207,28 | 607,75 |
| SL | −89,80 | 636,92 |
| SM | 89,03 | 639,1 |
| US | −103,32 | 209,4 |
| P | 125,78 | 250,24 |
| DOCHOD | 0,02 | 0,03 |
| CZAS | 6,38 | 4,57 |
| RPG | 364,16 | 288,98 |
| MOBA | 367,21 | 265,85 |
| FPS | 373,29 | 398,94 |
| MMORPG | 598,27 | 292,85 |
| PC | −1,05 | 479,63 |
| KONSOLA | 203,55 | 517,64 |
| KUPNO | 93,03 | 309,54 |
| MIKRO | 69,79 | 300,51 |
| SUB | 115,88 | 272,6 |
| SAM | 401,59 | 193,35 |
| K1_5 | 229,14 | 524,16 |
| K_1 | −95,38 | 530,84 |
| WSAM | −37,28 | 149,94 |

*Tabela 6 - wyniki estymacji parametrów metoda KMNK dla początkowego modelu*

| | |
|---|---|
| błąd standardowy reszt | 645,16 |
| ESS | 32466069 |
| $R^2$ | 0,32 |
| Wartość F | 1,41 |
| $\bar{R}^2$ | 0,09 |
| Wartość p test F | 0,12 |

*Tabela 7 - podstawowe statystyki dla modelu szacowanego KMNK z tabeli 6*

Dla tak skonstruowanego modelu można obliczyć podstawowe informacje na jego temat wypisane w tabeli 7, czyli współczynniki $R^2$(2.17) i $\bar{R}^2$ (2.17.1), $ESS$, czyli sumę kwadratów błędów modelu, oraz wartość statystyki F (2.20) i wartość p dla uogólnionego testu Walda.

Na poziomie istotności $\alpha = 0{,}05$ nie można więc na podstawie uogólnionego testu Walda i przy użyciu centralnego twierdzenia granicznego, czyli założeniu o asymptotycznym rozkładzie normalnym reszt ze względu na dużą liczność próby, stwierdzić, że co najmniej jeden z parametrów modelu jest istotny statystycznie.

Patrząc na wykresy znajdujące się na rysunkach 2 i 3, widzimy, że w niektórych przypadkach odpowiedzi z danej kategorii jest bardzo mało. Przykładem niech będzie oryginalna zmienna EDU, której jedynie wartość „Studia doktoranckie" nie została bezpośrednio włączona do modelu. Osób, które zaznaczyły taką odpowiedz jest bardzo mało, więc można wysunąć hipotezę o potencjalnej współliniowości pozostałych zmiennych dotyczących pierwotnej zmiennej EDU. W podobnej sytuacji jest między innymi wykres zmiennej KUPNO i wartość „Więcej niż 5 razy w roku". Stosując zatem współczynnik $VIF$ (2.21) sprawdzam czy w początkowym modelu występuje współliniowość. Wyniki współczynników $VIF$ zamieszczone są w tabeli 8.

| zmienna | VIF | zmienna | VIF |
|---|---|---|---|
| KM | 1,46 | MIKRO | 3,37 |
| W18_25 | 7,08 | SUB | 3,14 |
| W25_40 | 10,61 | SAM | 1,82 |
| KAWALER | 2,34 | K1_5 | 5,10 |
| ZARECZONY | 1,93 | K_1 | 6,71 |
| SS | 23,33 | WSAM | 6,34 |
| P | 5,02 | MOBA | 4,93 |
| SL | 13,89 | FPS | 3,51 |
| SM | 11,91 | MMORPG | 4,64 |
| US | 2,52 | PC | 1,45 |
| PR | 3,92 | KONSOLA | 15,18 |
| DOCHOD | 1,63 | KUPNO | 16,02 |
| CZAS | 1,55 | RPG | 1,36 |

*Tabela 8 - Wartości współczynników VIF dla pierwotnego modelu*

Na podstawie przyjętego założenia zgodnie z przewidywaniami, współliniowość wykazują zmienne należące do grupy pierwotnej zmiennej EDU i KUPNO. Po usunięciu z modelu zmiennej SS oraz zmiennej K_1 nowe współczynniki VIF można zobaczyć w tabeli 9. Współczynniki te usuwałem pojedynczo, aby upewnić się czy współliniowość nie była przypadkiem związana tylko z jedną z tych zmiennych, jednak ze względu na powtarzalność reszty wyników prezentuję efekt końcowy w tabeli 9. Na podstawie tej tabeli zgodnie z moim założeniem dotyczącym współliniowości, gdy współczynnik $VIF$ wynosi więcej niż 10, w modelu nie występuje już współliniowość, wartości parametrów jak i podstawowe statystyki dla modelu bez współliniowości można zobaczyć w tabelach 10 i 12.

| zmienna | VIF | zmienna | VIF |
|---|---|---|---|
| KM | 1,44 | CZAS | 1,50 |
| W18_25 | 5,35 | RPG | 3,34 |
| W25_40 | 7,71 | MOBA | 3,07 |
| KAWALER | 2,24 | FPS | 1,82 |
| ZARECZONY | 1,90 | MMORPG | 5,03 |
| P | 1,48 | PC | 6,65 |
| SL | 1,70 | KONSOLA | 6,32 |
| SM | 1,74 | KUPNO | 4,73 |
| US | 2,52 | MIKRO | 3,51 |
| P | 3,73 | SUB | 4,63 |
| DOCHOD | 1,56 | SAM | 1,44 |
| WSAM | 1,34 | K1_5 | 1,50 |

*Tabela 9 - Wartości współczynników VIF po usunięciu zmiennej SS*

| zmienna | współczynnik | SE | statystyka t | wartość p |
|---|---|---|---|---|
| stała | −557,66 | 618,45 | −0,90 | 0,37 |
| KM | −49,46 | 172,72 | −0,29 | 0,78 |
| W18_25 | 36,27 | 290,71 | 0,12 | 0,90 |
| W25_40 | −118,15 | 363,88 | −0,32 | 0,75 |
| KAWALER | 154,32 | 208,29 | 0,74 | 0,46 |
| ZARECZONY | 238,36 | 279,86 | 0,85 | 0,40 |
| P | −131,80 | 325,49 | −0,40 | 0,69 |
| SL | 6,90 | 220,12 | 0,03 | 0,98 |
| SM | 183,97 | 241,35 | 0,76 | 0,45 |
| US | −104,33 | 206,70 | −0,50 | 0,62 |
| P | 137,82 | 241,05 | 0,57 | 0,57 |
| DOCHOD | 0,02 | 0,03 | 0,83 | 0,41 |
| CZAS | 6,46 | 4,45 | 1,45 | 0,15 |
| RPG | 362,00 | 284,24 | 1,27 | 0,21 |
| MOBA | 375,61 | 259,59 | 1,45 | 0,15 |
| FPS | 374,89 | 393,96 | 0,95 | 0,34 |
| MMORPG | 605,73 | 287,29 | 2,11 | 0,04 |
| PC | −10,17 | 471,90 | −0,02 | 0,98 |
| KONSOLA | 198,27 | 510,42 | 0,39 | 0,70 |
| KUPNO | 101,72 | 299,45 | 0,34 | 0,74 |
| MIKRO | 72,54 | 296,59 | 0,24 | 0,81 |
| SUB | 117,67 | 269,13 | 0,44 | 0,66 |
| SAM | 405,44 | 190,26 | 2,13 | 0,04 |
| K1_5 | 320,03 | 162,84 | 1,97 | 0,05 |
| WSAM | −41,69 | 146,66 | −0,28 | 0,78 |

*Tabela 10 - Wyniki oszacowania parametrów metoda KMNK dla modelu bez współliniowości*

| błąd standardowy reszt | 637,26 |
|---|---|
| ESS | 32488543 |
| $R^2$ | 0,32 |
| Wartość F | 1,57 |
| $\bar{R}^2$ | 0,12 |
| Wartość p test F | 0,07 |

*Tabela 11 - Podstawowe statystyki dla modelu z tabeli 10*

Mimo usunięcia współliniowości, model nie poprawił się wyraźnie. $R^2$ nadal różni się znacznie od $\bar{R}^2$, co świadczy o tym, że model jest mocno przeparametryzowany. W dodatku wartość p dla testu F nadal jest niższa niż $0{,}05$ co świadczy o tym, że nie ma podstaw do odrzucenia

hipotezy zerowej o nieistotności każdego z parametrów modelu. Nie chcąc jednak jeszcze rezygnować z modelu ściśle liniowego więc wykonam teraz test normalności rozkładu reszt, aby mieć podstawę do używania wartości p testu t-Studenta obliczonej na podstawie statystyki (2.9) z $\beta_1 = 0$, istotności pojedynczych zmiennych, która potrzebna jest do metody krokowej wstecznej. Póki co zdecydowana większość zmiennych, opierając się na powyższej wartości p jest nieistotna statystycznie, co podobnie jak $\bar{R}^2$ sugeruje, aby pominąć część zmiennych objaśniających.

Używając testu Shapiro-Wilka na podstawie wartości $p = 0{,}00$ obliczonej przy użyciu statystyki $(2.7) = 0.85$, należy odrzucić hipotezę (2.5) na rzecz hipotezy (2.6) czyli nie można potwierdzić normalności reszt testem. Dla celów użycia metody krokowej wstecznej jednak znów można podeprzeć się centralnym twierdzeniem granicznym i założyć asymptotyczną normalność rozkładu reszt ze względu na dużą liczność próby.

## Metoda krokowa wsteczna

Zgodnie z algorytmem przedstawionym we wstępie teoretycznym, z modelu będą kolejno odrzucane zmienne o największej wartości p testu t-Studenta. Zgodnie z tabelą 10, największą wartość p ma zmienna PC. W pierwszym kroku więc oszacowano model przy użyciu klasycznej metody najmniejszych kwadratów z pominiętą zmienną PC i ten proces powtarzano, aż do momentu uzyskania wszystkich zmiennych istotnych na poziomie $\alpha = 0{,}05$. Ze względu na objętość pokazywania wyników w każdym z kroków, zaprezentowano już efekt końcowy, do zobaczenia w tabelach 12 i 13. Zmienne usuwane były w następującej kolejności: PC, SL, W18_25, MIKRO, KUPNO, WSAM, KM, SUB, P, PR, W25_40, KAWALER, SM, ZARECZONY, KONSOLA, DOCHOD, FPS, MOBA, RPG, US, a ostatnią zmienną była zmienna CZAS.

| zmienna | współczynnik | błąd standardowy | statystyka t | wartość p |
|---|---|---|---|---|
| stała | 144,05 | 147,17 | 0,98 | 0,33 |
| MMORPG | 386,59 | 122,35 | 3,16 | 0 |
| SAM | 362,57 | 152,05 | 2,38 | 0,02 |
| K1_5 | 456,55 | 128,00 | 3,57 | 0 |

*Tabela 12 - wyniki estymacji parametrów modelu KMNK po użyciu metody krokowej wstecznej*

| | |
|---|---|
| błąd standardowy reszt | 606,74 |
| ESS | 37181867 |
| $R^2$ | 0,22 |
| Wartość F | 9,63 |
| $\bar{R}^2$ | 0,19 |
| Wartość p test F | 0 |
| Kryt. Bayes. Schwarza | 1658,22 |

*Tabela 13 - Podstawowe statystyki dla modelu z tabeli 12*

Jak widać model uległ znacznemu polepszeniu, tzn. Mimo spadku $R^2$, współczynnik $\bar{\bar{R}}^2$ nie różni się już tak bardzo od zwykłego współczynnika determinacji, co nakazuje sądzić, że model nie jest przeparametryzowany. Wartość p dla testu Walda jest już na poziomie 0,00 co sprawia, że można odrzucić hipotezę o braku istotności wszystkich zmiennych w modelu, na rzecz hipotezy o tym, że co najmniej jedna z nich jest istotna. Dla tego modelu obliczono także współczynnik informacyjny $BIC$ (2.26) w celu późniejszego porównania tego modelu z innymi. Mimo polepszenia niektórych elementów modelu $R^2$ jest na bardzo niskim poziomie, co nakazuje twierdzić, że stała oraz zmienne MMORPG, SAM i K1_5 jedynie w 22% wyjaśniają roczne wydatki na gry komputerowe. Wszystkie zmienne w modelu, na podstawie wartości p (ostatnia kolumna tabeli 12) testu istotności parametrów t-Studenta (2.9) są istotne oprócz stałej. Zamierzam jednak zostawić ją w modelu jako wpływ pominiętych, nieobecnych w dostępnym zbiorze danych zmiennych. Poszukam teraz modeli konkurencyjnych dla modelu uzyskanego metodą krokową wsteczną.

### Metoda Hellwiga

Dla kontrprzykładu według metody Hellwiga największą integralną pojemnością informacyjną podzbioru zmiennych ZARECZONY, P, SM, US, DOCHOD, CZAS, MMORPG, SAM i K1_5 jest $H = 0.23$ i jest to największa wartość $H$ dla wszystkich podzbiorów zmiennych objaśniających. Zatem według metody Hellwiga model z powyższymi zmiennymi jest najlepszy. Wyniki dla tego modelu można zobaczyć w tabeli 14.

| zmienna | współczynnik | błąd standardowy | statystyka t | wartość p |
|---|---|---|---|---|
| stała | 335,43 | 232,62 | 1,44 | 0,15 |
| ZARECZONY | 151,55 | 202,03 | 0,75 | 0,46 |
| P | −207,58 | 261,38 | −0,79 | 0,43 |
| SM | 103,68 | 183,33 | 0,57 | 0,57 |
| US | −115,03 | 135,87 | −0,85 | 0,40 |
| DOCHOD | 0,03 | 0,02 | 1,15 | 0,25 |
| CZAS | 6,71 | 3,71 | 1,81 | 0,07 |
| MMORPG | 253,81 | 132,59 | 1,91 | 0,06 |
| SAM | 366,08 | 159,96 | 2,29 | 0,02 |
| K1_5 | −344,99 | 134,85 | −2,56 | 0,01 |

*Tabela 14 - Wyniki oszacowania parametrów modelu dla zmiennych wybranych metodą Hellwiga*

| błąd standardowy reszt | 602,60 |
|---|---|
| ESS | 34497505 |
| $R^2$ | 0,28 |
| Wartość F | 4,07 |
| $\bar{R}^2$ | 0,21 |
| Wartość p test F | 0,00 |
| Kryt. Bayes. Schwarza | 1678,27 |

*Tabela 15 - Podstawowe statystyki dla modelu z tabeli 14*

Na podstawie tabeli 15 i 13 używając kryterium Bayesa Shwarza można stwierdzić, że model zbudowany przy użyciu metody Hellwiga jest gorszy od końcowego modelu uzyskanego metoda krokową wsteczną. W dodatku na podstawie statystyki t-Studenta i wartości p dla testu istotności parametrów można powiedzieć, że jedynie zmienne SAM, K1_5 i CZAS są statystycznie istotne. $R^2$ nie różni się zbytnio od $\bar{R}^2$, ale oba współczynniki sa niskie, więc mozna powiedzieć, że zmienne najlepsze w sensie Hellwiga tylko w 28% opisują wydatki roczne na gry komputerowe.

Ponieważ i metoda Hellwiga i metoda krokowa wsteczna nie przyniosły zadowalających rezultatów, ze względu brak potwierdzenia normalności błędów testem, w następnym rozdziale pokazano efekt włączenia zlogarytmizowanych zmiennych do modelu.

## Modele z logarytmami

Ze względu na brak normalności rozkładu reszt potwierdzonych testem w poprzednich modelach, jak i sugerując się bardzo dużymi wartościami odchyleń standardowych i sum kwadratów błędów w poprzednich modelach, wprowadzono teraz do modelu w miejsce zmiennej objaśnianej jej logarytm – zmienną LWYDATKI, a zmienne ciągłe CZAS i DOCHOD zamieniono ich logarytmami – odpowiednio LCZAS i LDOCHOD. Wyniki oszacowania modelu z użyciem logarytmów zmiennych przedstawiam w tabelach 16 i 17.

Patrząc na wartości statystyk z tabeli 17, można stwierdzić, że już na początku model uległ znacznej poprawie, bowiem porównując wyniki z tabelami 6 i 7, widać znaczny wzrost $R^2$ z poziomu $0{,}34$ do poziomu $0{,}51$. $R^2$ dotyczy jednak objaśniania zmiennej $LWYDATKI$, a nie wydatki, więc licząc współczynnik determinacji dla zmiennej objaśnianej WYDATKI uzyskano spadek $R^2 = 0{,}16$. Można zatem powiedzieć, że wszystkie modele w małym stopniu opisują zmienną WYDATKI. Wartość p dla testu Walda nakazuje sądzić, że co najmniej jeden z parametrów modelu jest niezerowy. Co prawda wartość $\bar{R}^2$ dość rózni się od $R^2$ co wskazuje na za dużą ilość zmiennych w modelu, jednak nie wszystkie z nich są istotne, co można zobaczyć w kolumnie wartości p dla testu t-Studenta (wartości większe niż $0{,}05$ oznaczają, że korespondujące zmienne mogą być nieistotne statystycznie). Aby móc użyć metody krokowej wstecznej w celu redukcji zmiennych, wykonano test normalności błędów.

Używając testu Shapiro-Wilka na podstawie wartości $\mathrm{p} = 0{,}54$ obliczonej przy użyciu statystyki $(2.7) = 0{,}99$; można stwierdzić, że nie ma podstaw do odrzucenia hipotezy zerowej (2.5) o normalności rozkładu błędów modelu. Z taką informacją można przystąpić więc do wykonania metody krokowej wstecznej dla modelu z tabeli 16. Będę postępować analogicznie jak w przypadku modelu ściśle liniowego, z tym, że co odrzucenie zmiennych będę sprawdzać normalność rozkładu błędów zmienionego modelu. Ponownie ze względu na powtarzalność wyników i dużą ich objętość, zaprezentuję tylko końcowy model zawarty w tabeli 18.

Tabele 16, 17 i 18 można zobaczyć na następnej stronie.

| zmienna | współczynnik | błąd standardowy | statystyka t | wartość p |
|---|---|---|---|---|
| stała | 2,94 | 1,48 | 1,99 | 0,05 |
| KM | 0,11 | 0,23 | 0,50 | 0,61 |
| W18_25 | −0,13 | 0,39 | −0,33 | 0,74 |
| W25_40 | −0,30 | 0,47 | −0,64 | 0,52 |
| KAWALER | 0,11 | 0,27 | 0,41 | 0,67 |
| ZARECZONY | −0,12 | 0,37 | −0,33 | 0,74 |
| P | −0,10 | 0,42 | −0,24 | 0,80 |
| SL | −0,12 | 0,28 | −0,43 | 0,66 |
| SM | 0,15 | 0,31 | 0,47 | 0,63 |
| US | −0,49 | 0,27 | −1,84 | 0,06 |
| PR | 0,16 | 0,31 | 0,51 | 0,61 |
| LDOCHOD | 0,12 | 0,16 | 0,73 | 0,46 |
| LCZAS | 0,39 | 0,11 | 3,69 | 0,00 |
| RPG | 0,77 | 0,38 | 2,01 | 0,04 |
| MOBA | 0,63 | 0,34 | 1,85 | 0,06 |
| FPS | 1,14 | 0,53 | 2,16 | 0,03 |
| MMORPG | 1,15 | 0,37 | 3,07 | 0,00 |
| PC | −0,12 | 0,63 | −0,18 | 0,85 |
| KONSOLA | 0,20 | 0,68 | 0,29 | 0,77 |
| KUPNO | 0,06 | 0,39 | 0,14 | 0,88 |
| MIKRO | −0,04 | 0,39 | −0,09 | 0,92 |
| SUB | −0,04 | 0,35 | −0,10 | 0,91 |
| SAM | 0,69 | 0,25 | 2,81 | 0,00 |
| K1_5 | 0,42 | 0,21 | 2,04 | 0,04 |
| WSAM | −0,22 | 0,19 | −1,14 | 0,25 |

*Tabela 16 - Wyniki oszacowania modelu KMNK z użyciem logarytmów zmiennych*

| | |
|---|---|
| błąd standardowy reszt | 0,83 |
| ESS | 55,59 |
| $R^2$ | 0,51 |
| Wartość F | 3,51 |
| $\bar{R}^2$ | 0,36 |
| Wartość p test F | 0,00 |

*Tabela 17 - Podstawowe statystyki dla modelu z tabeli 16*

| zmienna | współczynnik | błąd standardowy | statystyka t | wartość p |
|---|---|---|---|---|
| stała | 3,85 | 0,38 | 10,19 | 0,00 |
| US | −0,54 | 0,17 | −3,17 | 0,00 |
| LCZAS | 0,39 | 0,09 | 4,20 | 0,00 |
| RPG | 0,71 | 0,27 | 2,62 | 0,01 |
| MOBA | 0,54 | 0,27 | 2,01 | 0,05 |
| FPS | 1,21 | 0,44 | 2,75 | 0,01 |
| MMORPG | 0,89 | 0,25 | 3,56 | 0,00 |
| SAM | 0,75 | 0,21 | 3,60 | 0,00 |
| K1_5 | 0,49 | 0,17 | 2,78 | 0,00 |

*Tabela 18 - Wyniki modelu z logarytmami zmiennych szacowanego KMNK po użyciu metody krokowej wstecznej*

| | |
|---|---|
| błąd standardowy reszt | 0,79 |
| ESS | 60,24 |
| $R^2$ | 0,47 |
| Wartość F | 10,73 |
| $\bar{R}^2$ | 0,42 |
| Wartość p test F | 0,00 |
| Kryt. Bayes. Schwarza | 281,52 |

*Tabela 19 - Podstawowe statystyki dla modelu z tabeli 18*

Po metodzie krokowej wstecznej $R^2$ utrzymało się na wysokim poziomie, a $R^2$ dla wyjaśnienia zmiennej WYDATKI wzrosło do $0{,}19$. $R^2$ dla zmiennej LWYDATKI nie różni się bardzo wartością od $\bar{R}^2$, zatem nie ma podstaw sądzić, że w modelu zachwiany jest stosunek ilości parametrów do ilości obserwacji. Wartość p testu Walda pozostała bez zmian od modelu z tabeli 15. W dodatku zachowano normalność błędów, bo nadal nie ma podstaw, aby odrzucić hipotezę o ich normalności, przy wartości $\mathrm{p} = 0{,}26$ uzyskanej w teście Shapiro-Wilka.

W tabeli 18 można także zobaczyć obliczone statystyki t-Studenta (2.9) i wartości p dla testu istotności kolejnych parametrów. Na podstawie tych wartości p, mogę stwierdzić, że nie mam podstaw sądzić, aby którakolwiek zmienna była nieistotna na poziomie $\alpha = 0{,}05$.

We wstępnej analizie pozostało jeszcze sprawdzić występowanie współliniowości w modelu. W tabeli 20 zaprezentowano zatem wartości współczynnika VIF dla zmiennych w modelu z tabeli 18:

| **zmienna** | **VIF** | **zmienna** | **VIF** |
|---|---|---|---|
| US | 1,12 | FPS | 1,48 |
| LCZAS | 1,25 | MMORPG | 2,48 |
| RPG | 1,98 | SAM | 1,12 |
| MOBA | 2,10 | K1_5 | 1,12 |

*Tabela 20 - Wartości współczynników VIF dla zmiennych z modelu z tabeli 18*

Na ich podstawie mogę sądzić, że w modelu nie występuje współliniowość.

Według kryterium informacyjnego Bayessa Shwarza, oraz ze względu na normalność błędów modelu z tabeli 18, można przyjąć, że końcowy model ściśle liniowy jest gorszy od modelu oszacowanego przy użyciu logarytmów zmiennych. Do dalszego badania wybierano więc model z logarytmami.

Mając na uwadze cel pracy, pomimo wyboru modelu z logarytmami, przetestowane zostanie również jak model ściśle liniowy będzie prognozować i dopiero na tej podstawie wybrany zostanie najlepszy model.

# Testowanie wybranych własności ostatecznego modelu

## Analiza własności modelu

Wybrany model ma postać:

$$\begin{aligned} LWYDATKI = {} & 3{,}85 - 0{,}54US + 0{,}39LCZAS + 0{,}71RPG \\ & + 0{,}54MOBA + 1{,}21FPS + 0{,}89MMORPG \\ & + 0{,}75SAM + 0{,}49K1\_5 \end{aligned} \tag{3.1}$$

Wiedząc, że wybrany model nie zawiera zmiennych współliniowych, oraz jego błędy mają rozkład normalny sprawdzono teraz ostatnie z założeń metody najmniejszych kwadratów – homoskedastyczność składnika losowego.

W tym celu wykonano test White'a i otrzymano statystykę $\chi^2 = 29{,}78$ z 31 stopniami swobody wraz z wartością p $= 0{,}53$. Na jej podstawie można stwierdzić, że nie ma podstaw do odrzucenia hipotezy zerowej (2.33) o homoskedastyczności składnika losowego. Zatem wszystkie założenia metody najmniejszych kwadratów są spełnione, a co za tym idzie na mocy twierdzenia Gaussa-Markowa, można powiedzieć, że estymatory parametrów modelu są nieobciążone i najlepsze w klasie estymatorów liniowych.

Na podstawie wartości p testu t-Studenta z tabeli 18, wyznaczono teraz 95% przedziały ufności (2.25) dla parametrów modelu. W tym celu odczytano wartość krytyczną dla statystyki $t$ z 96 stopniami swobody na poziomie istotności $\alpha = 0{,}05$ i otrzymano $t^* = 1{,}98$. Zatem przedziały ufności dla parametrów będą postaci: $[\hat{\beta}_i - 1{,}98 * SE(\hat{\beta}_i);\ \hat{\beta}_i + 1.98 * SE(\hat{\beta}_i)]$. Wyniki można zobaczyć w tabeli 21.

| zmienna | współczynnik | błąd standardowy | Przedział ufności |
|---|---|---|---|
| stała | 3,85 | 0,38 | [0,95; 6,75] |
| US | −0,54 | 0,17 | [-0,36; -0,72] |
| LCZAS | 0,39 | 0,09 | [0,32; 0,46] |
| RPG | 0,71 | 0,27 | [0,33; 1,09] |
| MOBA | 0,54 | 0,27 | [0,25; 0,83] |
| FPS | 1,21 | 0,44 | [0,16; 2,26] |
| MMORPG | 0,89 | 0,25 | [0,45; 1,33] |
| SAM | 0,75 | 0,21 | [0,44; 1,06] |
| K1_5 | 0,49 | 0,17 | [0,33; 0,65] |

*Tabela 21 - Przedziały ufności dla parametrów ostatecznego modelu*

Tak więc można powiedzieć, że z 95% prawdopodobieństwem wartość parametru przy zmiennej US, opisująca całą badaną populację leży w przedziale $[-0{,}36;\ -0{,}72]$. Analogicznie dla reszty zmiennych.

Następnie chcąc przetestować istotność parametrów zawartych w modelu, można spojrzeć na wartości p dla testu t-Studenta, zawarte w tabeli 18. Na ich podstawie wyciągnięto wniosek, że na poziomie istotności $\alpha = 0{,}05$ należy przyjąć hipotezę alternatywną o istotnej różnicy tych parametrów od 0.

Chcąc upewnić się, że nie pominięto żadnej statystycznie istotnej zmiennej na drodze doboru zmiennych do modelu, wykonano uogólniony test Walda dla pominiętych zmiennych KM, W18_25, W25_40, KAWALER, ZARĘCZONY, SS, P, SL, SM, PR, DOCHOD, PC, KONSOLA, KUPNO, MIKRO, SUB, K1_5 I K1. I na podstawie wartości p dla uogólnionego testu Walda $= 0{,}97$ można wnioskować, że nie ma podstaw sądzić, że któraś z pominiętych zmiennych jest istotna.

W celu zweryfikowania, czy interpretacja $R^2$ zawartego w tabeli 19 jest poprawna należy sprawdzić, czy w modelu występują katalizatory. Jeśli tak, to interpretacja współczynnika determinacji może być zakłamana. Po zweryfikowaniu warunków (2.22) występowania katalizatorów otrzymano wynik, że katalizatorami w modelu są zmienne RPG, MOBA, FPS, MMORPG i SAM. Zgodnie z równaniem (2.23) zostanie teraz wyznaczone natężenie katalizy w modelu. Po dodaniu pojemności nośników informacyjnych zmiennych w kombinacji (2.12) pojemność integracyjna podzbioru zmiennych objaśniających jest równa $H = 0.19$, zatem natężenie katalizy w modelu jest równe $\eta = 0{,}26$. W rzeczywistości, wartość współczynnika determinacji $R^2$ może nie być tak wysoka, co za tym idzie, nie można stwierdzić jak dokładnie model ten opisuje zmienną objaśnianą WYDATKI. W celu porównania jak katalizatory wpłyną na prognozowanie modelu podjęto decyzję o tym, żeby zostawić w modelu katalizatory. Decyzja o ich ewentualnym usunięciu zostanie podejęta porównując prognozy dla modelu bez katalizatorów i modelu z katalizatorami.

Aby sprawdzić czy postać modelu w rzeczywistości jest liniowa, czyli czy model jest analitycznie stabilny i czy można przy jego pomocy wykonać prognozy, wykonano teraz kilka testów pozwalających udzielić odpowiedzi na zadane pytania.

Pierwszym z nich będzie test Ramseya RESET. Postępując zgodnie z algorytmem dla tego testu utworzono model pomocniczy i obliczono statystykę testową $F = 0{,}23$. Następnie wykonano uogólniony test Walda dla modelu pomocniczego otrzymując wartość p testu $= 0{,}79$. Na jej podstawie, i podstawie hipotez (2.27 i 2.28) można stwierdzić, że na mocy testu Ramseya nie ma podstaw do odrzucenia hipotezy o dobrze dobranej liniowej postaci modelu.

Następnie wykonano test Chowa, a ponieważ ciężko stwierdzić na podstawie posiadanych danych w jaki sposób należałoby je rozdzielić, wykonano test dzieląc obserwacje w połowie.

W wyniku otrzymano statystykę $F$ o wartości $0{,}42$ z wartością p dla testu równą $0{,}91$. Zatem test Chowa pozwala na nieodrzucenie hipotezy głównej (2.29) o stabilności postaci modelu. Co za tym idzie, można stwierdzić, że model nadaje się do prognozowania.

W celu ostatecznego upewnienia na temat postaci modelu wykonano także nieparametryczny test serii porządkując obserwacje według wartości zmiennej LCZAS. Dla tak posortowanych danych obliczono reszty modelu, a następnie wykonano test i otrzymano wartość $\mathrm{p} = 0{,}77$. Na jej podstawie również nie można odrzucić hipotezy o stabilności modelu.

Na podstawie i mocy powyższych testów można więc stwierdzić, że liniowa postać modelu jest dobrze dobrana i stabilna analitycznie.

Na koniec sprawdzania własności modelu należy sprawdzić, czy model jest koincydentny. W tabeli 22 w rzędzie $r_j^2$, zawarto współczynniki korelacji zmiennej $X_i$ ze zmienną LWYDATKI, a parametry modelu stojące przy zmiennej $X_i$ w rzędzie $\alpha$.

| | **US** | **LCZAS** | **RPG** | **MOBA** | **FPS** | **MMORPG** | **SAM** | **K1_5** |
|---|---|---|---|---|---|---|---|---|
| $r_j^2$ | -0,35 | 0,41 | 0,06 | -0,07 | -0,13 | 0,28 | 0,28 | 0,34 |
| $\alpha$ | −0,54 | 0,39 | 0,71 | 0,53 | 1,21 | 0,89 | 0,74 | 0,49 |

*Tabela 22 - Współczynniki korelacji liniowej Pearsona oraz parametry modelu*

Zatem na podstawie przeciwnych znaków przy współczynniku korelacji liniowej Pearsona i przy parametrze dla zmiennych FPS i MOBA można stwierdzić, że model nie jest koincydentny. W tym miejscu warto by było zaznaczyć, że usuwając katalizatory uzyskano by model koincydentny ze względu na to, że zmienne odpowiedzialne za jej brak są katalizatorami.

## Prognozowanie

W tym podrozdziale sprawdzono jak wybrane modele prognozują. W tym celu do otrzymanego wzoru modelu, podstawiam wartości zmiennych i obliczam $\hat{y}$. W tabeli 23 można zobaczyć, jak prezentują się wartości rzeczywiste $y$, PU, czyli 95% przedziały ufności dla prognoz i błędy ex post, czyli $y - \hat{y}$. Średni błąd bezwzględny prognozy wyniósł $765{,}68$ zł. Błąd ten może być zawyżony w stosunku do błędu uzyskanego w populacji, gdyż warto zwrócić uwagę na to, że na 10 obserwacji znajduje się tu najbardziej odstającą obserwację z całego zbioru. Porównajmy otrzymane wyniki z prognozami dla modelu bez katalizatorów. Zobaczmy na tabelę 24, gdzie średni błąd bezwzględny prognozy wyniósł mniej, bo $679{,}26$ zł. Na następnej stronie w tabelach 25 i 26 przedstawiam wyniki modelu bez katalizatorów użytego do prognoz z tabeli 24:

| $y$ | $\hat{y}$ | PU | Błąd ex post |
|---|---|---|---|
| 150 | 340,42 | [47,76,1238,51] | 190,41 |
| 1500 | 1185,64 | [172,83;4245,10] | -314,36 |
| 1000 | 494,32 | [67,77;1815,80] | -505,68 |
| 300 | 975,60 | [109,67;3875,07] | 675,60 |
| 1000 | 428,31 | [60,27;1556,38] | -571,69 |
| 1200 | 2141,01 | [296,49;7831,94] | 941,01 |
| 100 | 214,30 | [28,74;794,36] | 114,30 |
| 500 | 1310,73 | [140,63;5297,62] | 810,73 |
| 40 | 368,33 | [50,50;1352,99] | 328,33 |
| 5000 | 1429,38 | [206,49;5137,42] | -3570,60 |
| 350 | 678,93 | [79,33;2657,33] | 328,93 |
| 500 | 1336,53 | [182,55;4916,88] | 836,53 |

*Tabela 23 - Wyniki prognozowania modelu (3.1)*

| $y$ | $\hat{y}$ | PU | Błąd ex post |
|---|---|---|---|
| 150 | 321,60 | [37,87;1254,15] | 171,60 |
| 1500 | 877,04 | [104,18;3408,75] | -622,97 |
| 1000 | 877,04 | [104,18;3408,75] | -122,96 |
| 300 | 435,74 | [51,096;1702,01] | 135,74 |
| 1000 | 403,12 | [47,40;1572,93] | -596,88 |
| 1200 | 2023,55 | [234,80;7935,85] | 823,55 |
| 100 | 474,74 | [55,73;1853,49] | 374,74 |
| 500 | 576,75 | [66,36;2269,14] | 76,75 |
| 40 | 392,55 | [45,22;1543,75] | 352,55 |
| 5000 | 1053,69 | [123,87;4111,72] | -3946,31 |
| 350 | 308,94 | [36,35;1205,12] | -41,06 |
| 500 | 1385,98 | [158,65;5463,62] | 885,98 |

*Tabela 24 - Wyniki prognozowania modelu (3.1) bez katalizatorów*

| zmienna | współczynnik | błąd standardowy | statystyka t | wartość p |
|---|---|---|---|---|
| stała | 5,16 | 0,28 | 18,54 | 0,00 |
| US | −0,65 | 0,18 | −3,66 | 0,00 |
| LCZAS | 0,38 | 0,09 | 4,11 | 0,00 |
| K1_5 | 0,56 | 0,18 | 3,05 | 0,00 |

*Tabela 25 - Wyniki oszacowania parametrów modelu (3.1) bez katalizatorów KMNK*

| | |
|---|---|
| błąd standardowy reszt | 0,87 |
| ESS | 76,26 |
| $R^2$ | 0,33 |
| Wartość F | 16,73 |
| $\bar{R}^2$ | 0,31 |
| Wartość p test F | 0,00 |
| Kryt. Bayes. Schwarza | 283,02 |

*Tabela 26 - Podstawowe statystyki dla modelu z tabeli 25*

Pomimo tego, że model prognozuje lepiej niż poprzedni, to według kryterium informacyjnego Bayessa Shwarza, oraz na podstawie $R^2$ dla zmiennej WYDATKI (a nie LWYDATKI, które widnieje w tabeli) na poziomie $0{,}06$ stwierdzam jednak, że dopasowanie prognoz jest zbiegiem przypadku, i uważam, że model poprzedni byłby lepszy.

Pozostało jeszcze sprawdzić, jak prognozuje model ściśle liniowy i wyniki są zaskakujące:

| $y$ | $\hat{y}$ | PU | Błąd ex post |
|---|---|---|---|
| 150 | 506,62 | [-711,34; 1724,58] | 356,62 |
| 1500 | 893,21 | [-329,28; 2115,71] | -606,79 |
| 1000 | 506,62 | [-711,34; 1724,58] | -493,38 |
| 300 | 506,62 | [-711,34; 1724,58] | 206,62 |
| 1000 | 506,62 | [-711,34; 1724,58] | -493,38 |
| 1200 | 963,17 | [-261,04; 2187,37] | -236,83 |
| 100 | 144,05 | [-1094,46; 1382,57] | 44,05 |
| 500 | 506,62 | [-711,34; 1724,58] | 6,62 |
| 40 | 506,62 | [-711,34; 1724,58] | 466,62 |
| 5000 | 893,21 | [-329,28; 2115,71] | -4106,79 |
| 350 | 506,62 | [-711,34; 1724,58] | 156,62 |
| 500 | 506,62 | [-711,34; 1724,58] | 6,62 |

*Tabela 27 - wyniki prognoz dla modelu ściśle liniowego szacowanego dla 105 obserwacji model można zobaczyć w tabeli 28*

Na podstawie tabeli 27 średni błąd bezwzględny prognozy wyniósł najmniej, bo $598{,}41$ zł, czyli aż o $\sim 80$ zł lepiej niż model bez katalizatorów. Daje to podstawę sądzić, że w wyniku założenia, że normalność błędów modelu da lepszy rezultat, pominięto najlepiej prognozujący model.

## Dodatkowe badanie modelu liniowego bez logarytmów

W związku z hipotezą z poprzedniego podrozdziału dodatkowo przetestowany zostanie model z tabeli 12.

Ponieważ wiadomo, że model ten nie posiada normalności rozkładu błędów potwierdzonej testem, to należy założyć, że na centralnego mocy twierdzenia granicznego błędy te posiadają asymptotyczny rozkład normalny.

Jako że model zawiera same zmienne binarne, chcąc upewnić się, że nie pominięto żadnej zmiennej wykonano uogólniony test Walda dla pominiętych zmiennych ciągłych: CZAS, LCZAS, DOCHOD, LDOCHOD, i tylko dla zmiennej LCZAS otrzymano wartość p w teście $= 0{,}04$. Co daje powód sądzić, że dodanie do tego modelu zmiennej LCZAS, spowoduje istotny wzrost $R^2$. Model po dodaniu zmiennej LCZAS można zobaczyć w tabeli 28:

| zmienna | współczynnik | błąd standardowy | statystyka t | wartość p |
|---|---|---|---|---|
| stała | −182,48 | 218,21 | −0,84 | 0,41 |
| LCZAS | 131,53 | 65,67 | 2 | 0,05 |
| MMORPG | 330,34 | 123,79 | 2,67 | 0,01 |
| SAM | 370,63 | 149,89 | 2,47 | 0,02 |
| K1_5 | 406,52 | 128,59 | 3,16 | 0 |

*Tabela 28 - Oszacowania parametrów modelu z tabeli 29 po dodaniu zmiennej LCZAS*

| | |
|---|---|
| błąd standardowy reszt | 678.14 |
| ESS | 35748022 |
| $R^2$ | 0.25 |
| Wartość F | 8.44 |
| $\bar{R}^2$ | 0.22 |
| Wartość p test F | 0 |

*Tabela 29 - podstawowe statystyki dla z modelu z tabeli 28*

W celu porównania modelu z modelem z logarytmami, przeprowadzone zostaną wybrane testy, lecz większą uwagę skupiono na samych wynikach i wnioskach z nich płynących.

- Test heteroskedastyczności White'a z wartością $p = 0{,}74$ pozwala stwierdzić, że nie ma podstaw do odrzucenia hipotezy o homoskedastyczności błędów.
- Test stabilności Chowa przy podziale w połowie z wartością $p = 0{,}37$ pozwala stwierdzić, że nie ma podstaw do odrzucenia hipotezy zerowej o stabilności analitycznej modelu.
- Test stabilności modelu RESET z wartością $p = 0{,}26$ potwierdza, że nie ma podstaw do odrzucenia hipotezy o poprawnie dobranej liniowej postaci modelu.
- Na podstawie $R^2$ mogę powiedzieć, że zmienne LCZAS, MMORPG, SAM i K1_5 w 25% opisują zmienną WYDATKI.
- Wartość p dla testu F wskazuje na odrzucenie hipotezy zerowej o nieistotności wszystkich parametrów, co potwierdzają wartości p testu t-Studenta, czyniąc każdą ze zmiennych (wyłączając stałą) istotna statystycznie.

Na podstawie tabeli 28 i 22, można stwierdzić, że model jest koincydentny (ponieważ współczynnik korelacji CZAS z WYDATKI, będzie tego samego znaku co LCZAS z WYDATKI). Po obliczeniu katalizatorów okazuje się, że są nimi zmienne SAM i MMORPG, a natężenie katalizy w modelu wynosi $\eta = 0{,}25 - 0{,}22 = 0{,}03$, czyli jest bardzo niskie.

Pozostało zatem sprawdzić, czy dodanie zmiennej LCZAS do modelu poprawi jego prognozowanie. W tym celu wyznaczono prognozy, które można zobaczyć w tabeli 30:

| $y$ | $\hat{y}$ | PU | Błąd ex post |
|---|---|---|---|
| 150 | 491,00 | [-709,44; 1691,45] | -341,00 |
| 1500 | 941,86 | [-263,93; 2147,64] | 558,14 |
| 1000 | 611,52 | [-593,32; 1816,35] | 388,48 |
| 300 | 594,70 | [-608,81; 1798,22] | -294,70 |
| 1000 | 568,31 | [-633,59; 1770,21] | 431,69 |
| 1200 | 1109,21 | [-105,94; 2324,35] | 90,79 |
| 100 | 29,20 | [-1196,69; 1255,10] | 70,80 |
| 500 | 688,83 | [-525,02; 1902,67] | -188,83 |
| 40 | 332,65 | [-880,01; 1545,30] | -292,65 |
| 5000 | 1003,67 | [-206,10; 2213,45] | 3996,33 |
| 350 | 477,15 | [-723,56; 1677,85] | -127,15 |
| 500 | 764,50 | [-462,73; 1991,73] | -264,50 |

*Tabela 30 - Prognozy modelu liniowego po dodaniu zmiennej LCZAS*

Tak więc na podstawie średniego błędu bezwzględnego równego 587,08 można stwierdzić, że jest to najlepiej prognozujący model.

## Ostateczny wybór modelu i wnioski

Model z tabeli 30 ma postać:

$$WYDATKI = -182{,}48 + 131{,}53LCZAS + 330{,}34MMORPG + 370{,}63SAM + 406{,}52K1\_5 \quad (3.2)$$

Model (3.1) różni się od modelu (3.2) tym, że w pierwszym z nich potwierdzono normalność błędów testem Shapiro-Wilka. Z kolei oba modele mają poprawną i stabilną postać, w obu występują katalizatory, z tym, że względne natężenie katalizy w modelu (3.1) wynosi 55%, a w modelu (3.2) tylko 12%. Model (3.2) jest koincydentny, a jego średni bezwzględny błąd prognozy jest aż o ~180 zł dokładniejszy w porównaniu z modelem (3.1). Mimo wszystkich przytoczonych argumentów można zauważyć, że przez to, że większość testów opiera się na normalności rozkładu błędów, otrzymane wyniki mogą być niewiarygodne, a lepsze prognozy przypadkowe. W związku z tym jako najlepszy model opisujący i prognozujący wydatki roczne na gry komputerowe dla populacji, którą jest społeczność grup World of Warcraft Polska i Dota2 Polska, przy założeniu niezerowych wartości cech: dochód, czas spędzony na graniu tygodniowo oraz wydatki na gry, wybrano model (3.1).

## Podsumowanie i rozstrzygnięcie hipotez badawczych

Najlepszym uzyskanym modelem jest więc model (3.1) i dla niego podano następującą interpretacje parametrów, a także na jego podstawie rozstrzygnięto postawione hipotezy badawcze:

- Jeśli najczęściej grywanym gatunkiem gier jest MMORPG, to przy pozostałych czynnikach niezmiennych, roczne wydatki rosną o 89% zł.
- Jeśli najczęściej grywanym gatunkiem gier jest RPG, to przy pozostałych czynnikach niezmiennych, roczne wydatki rosną o 71% zł.
- Jeśli najczęściej grywanym gatunkiem gier jest MOBA, to przy pozostałych czynnikach niezmiennych, roczne wydatki rosną o 54% zł.
- Jeśli najczęściej grywanym gatunkiem gier jest FPS, to przy pozostałych czynnikach niezmiennych, roczne wydatki rosną o 121% zł.
- Jeśli badany woli sam przejść grę, a nie oglądnąć ją w serwisie streamingowym to wartość rocznych wydatków na gry rośnie o 75% zł.
- Jeśli badany kupuje gry od 1-go do 5-ciu razy w roku, to roczne wydatki na gry rosną o 49% zł.
- Jeśli tygodniowy czas spędzony na graniu w gry wzrośnie o 1%, to w okolicach średnich roczne wydatki na gry wzrosną o 0,39% zł.
- Jeśli badany tylko uczy się, a nie pracuje to wydatki spadają o 54%.

Na podstawie tych interpretacji oraz na podstawie wartości p testu t-Studenta z tabeli 28, można stwierdzić, że w badanej populacji dochód gospodarstwa domowego w przeliczeniu na osobę nie wpływa istotnie na roczne wydatki na gry, co powoduje niemożliwość rozstrzygnięcia problemu (1.1) jakiego rodzaju dobrem są gry komputerowe dla tej populacji.

Rozstrzygnięto hipotezy (1.3) i (1.4) i na mocy testu istotności parametrów t-Studenta poprzez stwierdzenie, że wpływ tygodniowego czasu na granie na roczne wydatki na gry jest istotny. A przez dodatni znak parametru stojącego przy zmiennej LCZAS w modelu (3.1) można powiedzieć, że im więcej czasu osoby w badanej populacji poświęcają na grę, tym więcej wydają na gry w skali roku.

Na mocy testu t-Studenta parametrów można także stwierdzić, że należy odrzucić $H_0$ (1.5) na rzecz hipotezy o wpływie co najmniej jednego rodzaju gry na wydatki roczne. Istotny i powodujący różnice jest podział na gatunek MMORPG, FPS, RPG, MOBA i wszystkie pozostałe, czyli taki w jakiej postaci został włączony do modelu. Zatem przez znak i wartość parametru przy zmiennej FPS w modelu (3.1) mogę stwierdzić, że w badanej populacji osoby grające w FPS wydają rocznie najwięcej w stosunku do osób grających w pozostałe gry.

# Spis tabel i rysunków

# Bibliografia


Szreder M. „Metody i techniki sondażowych badań opinii”, 2010 PWN Warszawa,

Mishra P, Pandey CM, Singh U, Gupta A. “Scales of measurement and presentation of statistical data” Ann Card Anaesth. 2018 Oct-Dec; 21(4):419-422,

Newzoo.com – strona poświęcona danym dla rynku gier – odwiedzona w dniu 21.05.2021 r. https://newzoo.com/insights/articles/the-global-games-market-will-generate-152-1-billion-in-2019-as-the-u-s-overtakes-china-as-the-biggest-market/ ,

Dodge Y. „The Concise Encyclopedia of Statistics”, 2008,

Hellwig Z. *Problem optymalnego wyboru predykant*. „Przegląd Statystyczny” 1969 nr 3-4.

D.H. Vu, K.M. Muttaqi, A.P. Agalgaonkar, *A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables*, “Applied Energy”, Tom 140, Strony 385-394, 2015,

S.S. Shapiro, M.B. Wilk, *An analysis of variance for normality (complete samples),*

“Biometrika”, Tom 52 (1965), Strony 591-611,

A. Buda, A. Jarynowski (2010), *Life-time of correlations and its applications* tom 1, Wydawnictwo Niezależne: 5–21, Grudnia 2010, ISBN 978-83-915272-9-0.

Cohen, J. (1988). “Statistical power analysis for the behavioral sciences” (drugie wydanie).

Glantz, Stanton A.; Slinker, B. K. (1990). “Primer of Applied Regression and Analysis of Variance”. Wydawnictwo McGraw-Hill. ISBN 978-0-07-023407-9.

Maddala G. S. (2006) “Ekonometria”, Wydawnictwo Naukowe PWN SA. ISBN 978-83-01-14638-2,

Sheather, Simon (2009). *A modern approach to regression with R.* wydawnictwo New York, NY: Springer. ISBN 978-0-387-09607-0,

Gruszczyński M., Kuszewski T., Podgórska M. (2009). Ekonometria i badania operacyjne Wydawnictwo Naukowe PWN,

Hellwig Z. *Przechodniość relacji skorelowania zmiennych losowych i płynące stąd wnioski ekonometryczne* „Przegląd Statystyczny” 1976, nr 1,

Gruszczyński M., Kuszewski T., Podgórska M. (2009). Ekonometria i badania operacyjne Wydawnictwo Naukowe PWN.