Hanna Dudek[a], Monika Dybciak[b]
[a]Katedra Ekonometrii i Informatyki SGGW
[b]studentka Międzywydziałowego Studium Informatyki i Ekonometrii
e-mail: hdudek@mors.sggw.waw.pl

ZASTOSOWANIE MODELU LOGITOWEGO DO ANALIZY WYNIKÓW EGZAMINU

Streszczenie: W pracy podjęto próbę wyjaśnienia przyczyn wysokiego udziału ocen negatywnych uzyskanych z egzaminu z ekonometrii na Międzywydziałowym Studium Informatyki i Ekonometrii SGGW. W celu określenia wielkości wpływu czynników objaśniających wynik egzaminu zastosowano model logitowy. Model ten stanowi jeden z rodzajów modeli dwumianowych, w których zmienna objaśniana jest zmienną zerojedynkową.
Przedstawiono metody estymacji i weryfikacji modelu logitowego. Podano sposób interpretacji otrzymanych wyników. Na podstawie oszacowanego modelu stwierdzono, że systematyczna praca w ciągu semestru oraz dobry wypoczynek bezpośrednio przed pisaniem pracy egzaminacyjnej zwiększały prawdopodobieństwo zdania otrzymania oceny pozytywnej.

Słowa kluczowe: model logitowy, zmienna zerojedynkowa, wyniki egzaminu.

WSTĘP

Uzyskanie oceny negatywnej z egzaminu jest przykrym doświadczeniem. Oznacza konieczność ponownej weryfikacji wiedzy z danego przedmiotu. W niektórych wyższych uczelniach niezdanie egzaminu w pierwszym terminie wiąże się z utratą otrzymywania stypendium naukowego w następnym roku akademickim. Regulamin Szkoły Wyższej Gospodarstwa Wiejskiego nie narzuca takich sankcji. Jednakże negatywna ocena uzyskana ponownie na egzaminie poprawkowym przekreśla zwykle możliwość ubiegania się o takie stypendium. Zdecydowanie poważniejszą konsekwencją jest zasadnicza trudność w kontynuowaniu studiów. Negatywna ocena uzyskana na egzaminie może mieć także znaczenie psychologiczne. Student czasem zastanawia się nad tym, czy wybrał właściwy kierunek studiów oraz czy w ogóle powinien studiować. Rozważania takie rzadko bywają konstruktywne. Dlatego też powinno się minimalizować ryzyko niezdania egzaminu.

W pracy tej podjęto próbę analizy zależności wyniku egzaminu od różnych czynników na podstawie zbudowanego modelu ekonometrycznego. Oszacowany model może pomóc odpowiedzieć na pytanie jak zwiększyć prawdopodobieństwo uzyskania oceny pozytywnej.

DANE EMPIRYCZNE

Analizowane w pracy dane dotyczą studentów trzeciego roku pięcioletnich dziennych studiów magisterskich Międzywydziałowego Studium Informatyki i Ekonometrii SGGW w roku akademickim 2004/2005. Studenci w semestrze zimowym uczęszczali na zajęcia z „Ekonometrii" w wymiarze 30 godzin wykładów i 30 godzin ćwiczeń. Ekonometria była na kierunku studiów „Informatyka i ekonometria" przedmiotem kierunkowym kończącym się egzaminem z liczbą punktów ECTS wynoszącą 5.

W roku akademickim 2004/2005 na trzecim roku pięcioletnich studiów dziennych magisterskich Międzywydziałowego Studium Informatyki i Ekonometrii SGGW zarejestrowanych było 82 osoby. Do egzaminu z przedmiotu „Ekonometria" w pierwszym terminie przystąpiło 77 studentów. 5 osób nie zdawało wtedy egzaminu z powodu braku zaliczenia ćwiczeń bądź z powodu choroby. Ocenę pozytywną otrzymało 45 studentów, co stanowiło 58,44% wszystkich zdających.

Po konsultacjach ze studentami ustalono zestaw czynników wpływających na wyniki egzaminu. Na podstawie tych informacji sporządzono anonimową ankietę internetową. Odpowiedziało na nią 39 studentów, wśród których 23 zdały egzamin, co stanowiło 58,97% wszystkich osób udzielających odpowiedzi. Procentowy udział liczby prac egzaminacyjnych z oceną pozytywną był zatem zbliżony do analogicznego udziału wśród wszystkich osób przystępujących do egzaminu. Dzięki ankiecie internetowej otrzymano informacje, na podstawie których określono następujące zmienne:

Y przyjmującą wartość 1 jeśli student zdał egzamin z ekonometrii w pierwszym terminie oraz 0 w przypadku otrzymania oceny negatywnej,

X_1 określającą liczbę opuszczonych godzin wykładów i ćwiczeń z ekonometrii,

X_2 odnoszącą się do liczby godzin poświęconych na naukę indywidualną tego przedmiotu przed egzaminem,

X_3 oznaczającą ocenę uzyskaną na zaliczenie ćwiczeń z ekonometrii,

X_4 określających liczbę godzin snu danego studenta w ostatniej dobie przed egzaminem.

Poniżej przedstawiono krótką charakterystykę zmiennych X_1, X_2, X_3 i X_4.

Tabela 1. Miary położenia i zróżnicowania zmiennych

Miary:	Zmienne:			
	X_1	X_2	X_3	X_4
Średnia arytmetyczna	3,41	33,77	3,85	7,06
Mediana	3,00	25,00	4,00	7,00
Odchylenie standardowe	3,15	27,38	0,67	1,20
Minimum	0	2	3	4
Maksimum	11	100	5	10

Źródło: obliczenia własne.

Największym zróżnicowaniem cechuje się X_1 - współczynnik zmienności wynosi 89,74%. Bardzo budujące są informacje dotyczące mediany rozkładu tej zmiennej - absencja co najmniej 50% studentów nie przekraczała 3 godzin. Wśród badanych 39 osób byli studenci, którzy nie opuścili ani jednych zajęć, „rekordziści" zaś nie byli obecni na 11 godzinach. Współczynniki zmienności dla X_2 jest równy 81,08%, co świadczy o znacznym zróżnicowaniu liczby godzin poświęconej na naukę indywidualną przed egzaminem. Średnio do egzaminu przygotowywano się ok. 30 godzin, zdarzały się przy tym osoby uczące się jedynie 2 godziny jak i takie, które poświęcały 100 godzin. Zmienne X_3 i X_4 charakteryzują się stosunkowo niewielkim zróżnicowaniem, współczynniki zmienności są zbliżone: dla X_3 - 17,40% oraz dla X_4 - 17,00%. Zwraca tu uwagę stosunkowo wysoka średnia uzyskana na zaliczenie ćwiczeń oraz fakt, że co najmniej połowa ankietowanych studentów spała przed egzaminem nie mniej niż 7 godzin.

Dla powyższych zmiennych zbudowano model, w którym zmienną objaśnianą jest Y a zestaw potencjalnych zmiennych objaśniających stanowią X_1, X_2, X_3 i X_4.

MODELE ZMIENNYCH JAKOŚCIOWYCH

Modele dwumianowe (dychotomiczne) są najprostszymi i najpopularniejszymi modelami, w których zmienna objaśniana jest zmienną jakościową. W modelach tych zmienna objaśniana jest kwantyfikowana za pomocą zmiennej zerojedynkowej. Niech y_i oznacza i-tą realizację zmiennej zerojedynkowej Y. Zmienna y_i ma rozkład Bernoulliego. Przyjmuje wartość 1 z prawdopodobieństwem P_i oraz wartość 0 z prawdopodobieństwem $1-P_i$.
Wartość oczekiwana zmiennej y_i wynosi:

$$E(y_i) = 1 \cdot P_i + 0 \cdot (1 - P_i) = P_i \tag{1}$$

W modelach dwumianowych zakłada się, że P_i jest funkcją wektora wartości zmiennych objaśniających $\mathbf{x_i}$ dla i-tego obiektu oraz wektora parametrów $\boldsymbol{\beta}$:

$$P_i = P(y_i = 1) = F(\mathbf{x_i^T}\boldsymbol{\beta}) \tag{2}$$

W zależności od typu funkcji F wyróżnia się różne rodzaje modeli [Judge i in. 1985]. Do najbardziej znanych należą:

- liniowy model prawdopodobieństwa, którym $P_i = F(\mathbf{x_i^T}\boldsymbol{\beta}) = \mathbf{x_i^T}\boldsymbol{\beta}$, (3)
- model probitowy, gdzie $P_i = F(\mathbf{x_i^T}\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x_i^T}\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$, (4)
- model logitowy, dla którego $P_i = F(\mathbf{x_i^T}\boldsymbol{\beta}) = \frac{1}{1 + \exp\left(-\mathbf{x_i^T}\boldsymbol{\beta}\right)}$. (5)

Zastosowanie najprostszego z przedstawionych modeli - liniowego modelu prawdopodobieństwa ma wiele negatywnych konsekwencji [Gruszczyński 2002, Maddala 2002].

1. Składnik losowy modelu $y_i = \mathbf{x_i^T}\boldsymbol{\beta} + \varepsilon_i$ jest heteroskedastyczny, gdyż $Var(\varepsilon_i) = P_i(1-P_i)$.
2. Składnik losowy modelu $y_i = \mathbf{x_i^T}\boldsymbol{\beta} + \varepsilon_i$ nie ma rozkładu normalnego, co powoduje trudności w zastosowaniu testów istotności.
3. Wartości $\hat{y}_i = \mathbf{x_i^T b}$ mogą wykraczać poza przedział [0, 1] (przez **b** oznaczono wektor ocen wektora parametrów **β)**.
4. Współczynnik determinacji R^2 w modelu LMP przyjmuje zwykle bardzo niskie wartości.

Ponadto, jak wskazuje Gujarati [Gujarati 2003], fundamentalny problem w stosowaniu LMP polega na przyjęciu założenia, że prawdopodobieństwo w sposób liniowy zależy od zmiennych objaśniających, co jest równoznaczne z założeniem, że krańcowy efekt jest stały. W większości problemów praktycznych zależność prawdopodobieństwa od zmiennych objaśniających jest nieliniowa.

Jak wskazują niektórzy autorzy modele probitowe i logitowe są podobne do siebie i w praktyce wykorzystuje się jeden z nich [Judge i in. 1985].

MODEL LOGITOWY

Wartość funkcji odwrotnej do F, określonej wzorem (5), czyli

$$F^{-1}(P_i) = \ln\frac{P_i}{1-P_i} \tag{6}$$

nazywa się logitem. Stąd dla modelu (5) przyjęło się w literaturze przedmiotu określenie „model logitowy".

Na podstawie tego modelu można określić marginalny przyrost prawdopodobieństwa:

$$\frac{\partial P_i}{\partial x_{ji}} = \beta_j \frac{\exp\left(-\mathbf{x_i^T}\boldsymbol{\beta}\right)}{[1+\exp\left(-\mathbf{x_i^T}\boldsymbol{\beta}\right)]^2} = \beta_j P_i(1-P_i)\,. \tag{7}$$

Ponieważ $P_i(1-P_i) > 0$, (8)

to znak parametru stojącego przy zmiennej X_j określa kierunek wpływu X_j na Y:

- dodatniemu β_j odpowiada wzrost prawdopodobieństwa tego, że Y=1, jeśli X_j zwiększa się,
- ujemnemu β_j towarzyszy spadek prawdopodobieństwa tego, że Y=1, jeśli X_j zwiększa się, przy założeniu, że pozostałe zmienne objaśniające pozostają bez zmian.

Do interpretacji oszacowanego modelu logitowego wykorzystuje się również wyrażenie $\frac{P_i}{1-P_i}$ nazywane ilorazem szans. Iloraz szans określa zatem stosunek prawdopodobieństwa, że Y=1 do prawdopodobieństwa, że Y=0. Ponieważ

$\frac{P_i}{1-P_i} = \exp(\mathbf{x_i^T \beta})$, zatem $\exp(\beta_j)$ informuje ile razy zwiększa się iloraz szans jeśli zmienna X_j wzrasta o jednostkę, ceteris paribus.

Na podstawie oszacowanego modelu $\hat{P}_i = \frac{1}{1+\exp(-b_0 - b_1 x_{1i} - ... - b_k x_{ki})}$ można określić prognozy:

$\hat{y}_i = 1$, jeśli $\hat{P}_i > p^*$ oraz

$\hat{y}_i = 0$, jeśli $\hat{P}_i \leq p^*$

Zwykle przyjmuje się wartość odcinającą $p^* = 0,5$. Jednakże niektórzy autorzy [Judge i in. 1985] proponują ustalić tę wartość w taki sposób, aby uwzględnić fakt niezbilansowania próby. Przez próbę niezbilansowaną uważa się próbę, gdzie n_1 różni się od n_0,. gdzie n_1 i n_0 -liczba przypadków, dla których odpowiednio Y przyjmuje wartość 1 oraz 0. W takiej sytuacji proponuje się $p^* = \frac{n_1}{n}$, gdzie $n = n_0 + n_1$.

ESTYMACJA PARAMETRÓW MODELU LOGITOWEGO

Do estymacji parametrów β_0, β_1, ...β_k stosuje się zwykle metodę największej wiarygodności. Jeśli dysponuje się n-elementową próbą y_1, y_2,...,y_n, gdzie każda z y_i przyjmuje wartość 1 z prawdopodobieństwem P_i określonym jako (5), to funkcja wiarygodności ma postać: $L = \prod_{i=1}^{n} P_i^{y_i} (1-P_i)^{1-y_i}$, (9)

stąd logarytm tej funkcji można zapisać jako:

$$\ln L = \sum_{i=1}^{n} [y_i \ln P_i + (1-y_i)\ln(1-P_i)] = \sum_{i=1}^{n} y_i \ln\left(\frac{P_i}{1-P_i}\right) + \sum_{i=1}^{n} \ln(1-P_i) \quad (10)$$

Wykorzystując zależność (5) otrzymuje się:

$$\ln L = \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}) + \sum_{i=1}^{n} \ln\left(\frac{1}{1+\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}\right) \quad (11)$$

Należy zatem przy znanych wartościach y_i, x_{1i}, ...,x_{ki}, i=1, 2, ...,n, oszacować tak parametry β_0, β_1, ...β_k, by zapewniały maksymalną wartość logarytmu funkcji wiarygodności. W tym celu należy obliczyć pochodne cząstkowe pierwszego rzędu funkcji lnL i przyrównać je do zera. Otrzymuje się wówczas k+1 równań nieliniowych;

$$\frac{\partial \ln L}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \frac{\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}{1+\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}) = 0 \quad (12)$$

$$\frac{\partial \ln L}{\partial \beta_1} = \sum_{i=1}^{n} (y_i - \frac{\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}) x_{1i} = 0 \quad (13)$$

$$\vdots$$

$$\frac{\partial \ln L}{\partial \beta_k} = \sum_{i=1}^{n} (y_i - \frac{\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}) x_{ki} = 0 \quad (14)$$

Powyższe równania są nieliniowe ze względu na parametry β_0, β_1, ...β_k. Nie można podać analitycznych wzorów określających estymatory tych parametrów. Dlatego też do poszukiwania maksimum logarytmu funkcji wiarygodności należy zastosować procedury iteracyjne. Macierz drugich pochodnych jest ujemnie określona dla wszystkich wartości parametrów β_0, β_1, ...β_k [Gujarati 2003]. To oznacza, że logarytm funkcji wiarygodności jest funkcją wklęsłą, zatem maksimum lokalne jest maksimum globalnym. Zwykle więc osiągana jest zbieżność w procesie iteracyjnym.

TESTY ISTOTNOŚCI PARAMETRÓW

Estymatory parametrów uzyskane metodą największej wiarygodności mają asymptotyczny rozkład normalny i są asymptotycznie najefektywniejsze. Zatem dla dostatecznie dużych prób do testowania statystycznej istotności parametrów można wykorzystać asymptotyczny test t Studenta.
Do weryfikacji hipotezy zerowej:

$$H_0: \beta_j = 0, \quad (15)$$

wobec hipotezy alternatywnej: $H_1: \beta_j \neq 0$, =1, 2,...,k,
wykorzystuje się także statystykę ilorazu wiarygodności [Greene 2000]:

$$LR_j = -2(\ln \hat{L}_{Rj} - \ln \hat{L}_{UR}) \quad (16)$$

gdzie $\ln\hat{L}_{Rj}$ jest wartością maksymalną logarytmu funkcji wiarygodności dla modelu z wyrazem wolnym zawierającego zmienne X_1,...X_{j-1}, X_{j+1},...,X_k, (tj. bez zmiennej X_j),

$\ln\hat{L}_{UR}$ - wartość wartością maksymalną logarytmu funkcji wiarygodności dla pełnego modelu (tj. ze zmiennymi X_1,...X_{j-1}, X_j, X_{j+1},...,X_k).
Statystyka LR_j ma dla dużych prób rozkład χ^2 z 1 stopniem swobody.
Test ilorazu wiarygodności stosuje się także do weryfikacji hipotezy o braku statystycznej istotności wszystkich parametrów przy zmiennych objaśniających Hipoteza zerowa ma wtedy postać:

$$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0, \quad (17)$$

a hipotezę alternatywną można sformułować w następujący sposób:

H_1: co najmniej jeden parametr $\beta_j \neq 0$, =1, 2,...,k.

Wtedy statystyka ilorazu wiarygodności może być zapisana jako:

$$LR = -2(\ln \hat{L}_R - \ln \hat{L}_{UR}), \tag{18}$$

gdzie $\ln\hat{L}_R$ jest maksymalną wartością logarytmu funkcji wiarygodności dla modelu zawierającego jedynie wyraz wolny,

$\ln\hat{L}_{UR}$ - wartością maksymalną logarytmu funkcji wiarygodności dla pełnego modelu.

Statystyka LR ma dla dużych prób rozkład χ^2 z k stopniami swobody.

OCENA ZGODNOŚCI MODELU Z DANYMI EMPIRYCZNYMI

Dla modeli binarnych stosuje się różne miary oceniające zgodność modelu z danymi empirycznymi. Wiele z tych miar konstruuje się na zasadzie odpowiedników klasycznego współczynnika determinacji dla modelu liniowego szacowanego metodą najmniejszych kwadratów.

Najprostszą miarą jest kwadrat współczynnika korelacji między wartościami empirycznymi zmiennej objaśnianej a wartościami wyznaczonymi z modelu:

$$R^2 = [r(y, \hat{P})]^2 . \tag{19}$$

Propozycja Efrona zaś opiera się na pomyśle, aby we wzorze na klasyczny współczynnik determinacji : $R^2 = 1 - \dfrac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$ zamiast sumy kwadratów reszt podstawić $\sum_{i=1}^{n} (y_i - \hat{P}_i)^2$ [Maddala 2002].:

$$R^2_{Efron} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{P}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{20}$$

Ponieważ

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = n_1 - n\left(\frac{n_1}{n}\right)^2 = \frac{n_1 n_0}{n}, \tag{21}$$

to współczynnik Efrona można zapisać jako:

$$R^2_{Efron} = 1 - \frac{n}{n_1 n_0} \sum_{i=1}^{n} (y_i - \hat{P}_i)^2 \tag{22}$$

gdzie n_1 i n_0 -liczba przypadków, dla których odpowiednio Y przyjmuje wartość 1 oraz 0.

Kolejna miara zaproponowana przez McFaddena dotyczy modelu szacowanego za pomocą metody największej wiarygodności:

$$R^2_{McFadden} = 1 - \frac{\ln \hat{L}_{UR}}{\ln \hat{L}_R} \tag{23}$$

gdzie $\ln\hat{L}_R$ jest wartością maksymalną logarytmu funkcji wiarygodności dla modelu zawierającego jedynie wyraz wolny,

$\ln\hat{L}_{UR}$ - wartość maksymalna logarytmu funkcji wiarygodności dla pełnego modelu.

Ten współczynnik jest w tym sensie odpowiednikiem klasycznego współczynnika determinacji, że przyjmuje wartość 0 jeśli $b_1 = b_2 = ...b_k = 0$ oraz wartość 1 w przypadku idealnego dopasowania, gdy $\hat{p}_i = y_i$ dla każdego i =1, 2, ...,n [Greene 2000, Gujarati 2003].

Do określenia zgodności modelu z danymi wykorzystuje się także mierniki dokładności prognoz [Judge i in. 1985]. Wielu praktyków uważa bowiem, że o jakości modelu decyduje trafność prognoz uzyskiwanych na jego podstawie. Najczęściej wykorzystuje się tu miarę podaną przez Maddalę nazywaną przez Gruszczyńskiego zliczeniowym R^2:

$$Zliczeniowy\ R^2 = \frac{n_{00} + n_{11}}{n} \tag{24}$$

gdzie n_{00} - liczba obserwacji, dla których $\hat{y}_i = y_i = 0$, n_{11} - liczba obserwacji, dla których $\hat{y}_i = y_i = 1$.

Zliczeniowy R^2 określa zatem udział poprawnie prognozowanych przypadków w łącznej liczbie przypadków. Wszystkie podane tu miary zgodności przyjmują wartości z przedziału [0, 1]. Wartości 0 odpowiada brak dopasowania. Im bliższe 1 jest R^2, tym większa zgodność modelu z danymi empirycznymi.

WYNIKI

Oszacowany model ma postać:

$$\hat{P}_i = \frac{1}{1 + \exp[-(-57,34 - 1,06x_{1i} + 0,27x_{2i} + 2,77x_{3i} + 6,16x_{4i})]} \tag{25}$$

W tabeli 2 przedstawiono wyniki badania statystycznej istotności parametrów.

Tabela 2. Wyniki testu ilorazu wiarygodności

Parametr	Wartość LR	Stopnie swobody	Wartość krytyczna testu χ^2 dla poziomu istotności 0,05
β_1	4,1126	1	3,841
β_2	6,3239	1	3,841
β_3	4,3656	1	3,841
β_4	18,5274	1	3,841
Łącznie β_1, β_2, β_3, β_4	43,0769	4	7,779

Źródło: Obliczenia własne wykonane przy pomocy programu Statgraphics

Na podstawie testu ilorazu wiarygodności można sądzić, że parametry β_1, β_2, β_3, β_4 są statystycznie istotne. Należy jednak wyniki uzyskane na podstawie tego testu przyjąć z ostrożnością z powodu niewielkiej liczebności próby[1].

Dla oszacowanego modelu określono wartości miar zgodności z danymi empirycznymi. Ponieważ współczynnik korelacji między wartościami zmiennej Y a wartościami prawdopodobieństwa wyznaczonymi z modelu wyniósł 0,9126, to $R^2 = [r(y, \hat{P})]^2 = 0{,}8328$. Zbliżone wartości otrzymano na podstawie miar $R^2_{Efron} = 0{,}8349$ oraz $R^2_{McFadden} = 0{,}8158$. Ponieważ na 39 badanych osób 23 zdały egzamin, to wartość odcinającą ustalono na poziomie $p^* = \frac{23}{39} = 0{,}5897$. Prognozy wyznaczano zatem w następujący sposób:

$$\hat{y}_i = \begin{cases} 1 & gdy \quad \hat{p}_i > 0{,}5897 \\ 0 & gdy \quad \hat{p}_i \leq 0{,}5897 \end{cases} \tag{26}$$

Na tej podstawie obliczono liczbę poprawnie prognozowanych przypadków.

Tabela 3. Klasyfikacja przypadków

Liczba obserwacji, dla których:			
	$\hat{y}_i = 1$	$\hat{y}_i = 0$	Razem
y_i= 1	22	1	23
y_i= 0	1	15	16
Razem	23	16	39

Źródło: obliczenia własne

Zliczeniowy $R^2 = \frac{n_{00} + n_{11}}{n} = \frac{22+15}{39} = 0{,}9487$, stąd niemal w 95% klasyfikacja przypadków okazała się prawidłowa.

Weryfikacja statystyczna modelu polegająca na określeniu stopnia dopasowania modelu do danych oraz na badaniu statystycznej istotności parametrów przebiegła pozytywnie, można zatem przejść do etapu interpretacji modelu.

Na podstawie znaku oceny parametru stojącego przy zmiennej X_j można określić kierunek wpływu zmiennej objaśniającej na prawdopodobieństwo zdania egzaminu. Zatem ponieważ

- $b_1<0$, to zwiększenie liczby opuszczonych zajęć zmniejszało prawdopodobieństwo uzyskania oceny pozytywnej,
- $b_2>0$, więc wzrost czasu nauki indywidualnej powodowało zwiększanie szansy zdania egzaminu,

[1] Dotyczy to zwłaszcza parametrów β_1 i β_3.

- $b_3>0$, to poprawa oceny z zaliczenia ćwiczeń oznaczała wzrost prawdopodobieństwa zdania egzaminu,
- $b_4>0$, więc poświęcenie więcej czasu na sen w ostatniej dobie przed egzaminem poprawiało szansę otrzymania oceny pozytywnej, ceteris paribus.

Interpretując ilorazy szans dla poszczególnych zmiennych (zakładając, że pozostałe zmienne uwzględnione w modelu pozostawały bez zmian) uzyskuje się informację:

- zwiększenie absencji na zajęciach o 1 godzinę powodowało spadek ilorazu szans o 65,45%,
- wzrost czasu nauki indywidualnej zwiększało ten iloraz o 30,59%,
- poprawa oceny z zaliczenia ćwiczeń o 1 stopień łączyła się z 16-to krotnym wzrostem ilorazu szans,
- wydłużenie snu w ostatniej dobie przed egzaminem o 1 godzinę powiększało 477 razy iloraz szans.

Zdumiewać może tak duży wpływ snu na wynik egzaminu. Być może osoby, które pozwoliły sobie na długi sen, to studenci bardzo uzdolnieni, nie obawiający się o wynik egzaminu. Zmienna ukryta odnosząca się do zdolności byłaby wtedy reprezentowana przez zmienną określającą liczbę godzin snu w ostatniej dobie przed egzaminem. Zjawisko to daje się także wytłumaczyć możliwą nierzetelnością udzielanych przez studentów informacji, na podstawie których oszacowano parametry modelu.

Ponieważ marginalny przyrost prawdopodobieństwa zależy od wartości zmiennych objaśniających, poniżej podano przykładowo wartości tych przyrostów dla wybranego studenta. Osoba ta opuściła 3 jednostki lekcyjne zajęć, poświęciła 10 godzin na przygotowanie się do egzaminu, z ćwiczeń na zaliczenie uzyskała ocenę 3 oraz ostatniej doby przed egzaminem poświęciła 7 godzin na sen.

Tabela 4. Wartości przyrostów marginalnych prawdopodobieństwa dla wybranych wartości zmiennych objaśniających.

Wartości	X_1	X_2	X_3	X_4
Zmiennych objaśniających	3	10	3	7
Przyrostów marginalnych prawdopodobieństwa	-0,0018	0,0004	0,0046	0,0001

Źródło: obliczenia własne

- Opuszczenie 1 jednostki lekcyjnej więcej zajęć spowodowałoby zmniejszenie się prawdopodobieństwa zdania egzaminu o 0,0018, ceteris paribus.
- Poświęcenie o 1 godzinę więcej na naukę indywidualną poprawiłoby prawdopodobieństwo uzyskania oceny pozytywnej z egzaminu o 0,0004, przy założeniu, że pozostałe zmienne pozostawałyby bez zmian.
- Uzyskanie o 1 stopień wyższej oceny na zaliczenie zwiększyłoby prawdopodobieństwo zdania egzaminu o 0,0046, ceteris paribus.

- Zwiększenie długości snu o 1 godzinę poprawiłoby prawdopodobieństwo uzyskania oceny pozytywnej z egzaminu o 0,0001, zakładając, że pozostałe zmienne pozostawałyby bez zmian.

Należy w tym miejscu podkreślić, że rozważana osoba otrzymała ocenę 3.0 z zaliczenia ćwiczeń oraz przygotowywała się do egzaminu przez jedynie 10 godzin, stąd niewielkie wartości przyrostów marginalnych prawdopodobieństw. Dla studenta słabego i leniwego jednostkowe zwiększenie danej zmiennej objaśniającej, ceteris paribus, nie przyczynia się w znaczący sposób do zmiany prawdopodobieństwa zdania egzaminu. Poniżej przedstawiono dwie przykładowe prognozy uzyskane na podstawie oszacowanego modelu logitowego.

Tabela 5. Prognozy zdania egzaminu

Numer prognozy	X_1	X_2	X_3	X_4	$\hat{P}$	$\hat{y}$
1	5	10	4	8	0,6033	1
2	1	40	3,5	6	0,2607	0

Źródło: obliczenia własne

Na podstawie informacji przedstawionych w tabeli 5 można sądzić, że osoba, która opuściłaby 5 jednostek lekcyjnych zajęć, przeznaczyłaby 10 godzin na przygotowanie się do egzaminu, z ćwiczeń na zaliczenie uzyskałaby ocenę 4 oraz w ostatniej dobie przed egzaminem poświęciła 8 godzin na sen, zdałaby egzamin. Natomiast student z oceną 3,5 zaliczającą ćwiczenia, który nie był obecny na jednej godzinie lekcyjnej, uczący się do egzaminu 40 godzin i śpiący jedynie 6 godzin w ciągu doby bezpośrednio przed egzaminem nie uzyskałby oceny pozytywnej z egzaminu.

PODSUMOWANIE

W pracy wykorzystano dane pozyskane na podstawie przeprowadzonej wśród studentów anonimowej ankiety internetowej. Z powodu braku informacji na temat rzetelności udzielanych odpowiedzi, trudno jest ocenić wiarygodność otrzymanych tu wyników. Zakładając jednak, że studenci udzielali prawdziwych informacji, podjęto próbę wyjaśnienia przyczyn wysokiego udziału ocen negatywnych uzyskanych z egzaminu z ekonometrii na Międzywydziałowym Studium Informatyki i Ekonometrii SGGW. W celu określenia wielkości wpływu czynników objaśniających rezultat egzaminu wykorzystano wyniki otrzymane na podstawie oszacowanego modelu logitowego.

Interpretacja ocen parametrów modelu prowadzi do wniosku, że systematyczna nauka w ciągu całego semestru (mała liczba opuszczonych zajęć i wysoka ocena na zaliczenie ćwiczeń) zwiększała prawdopodobieństwo zdania egzaminu. Bezpośrednio przed egzaminem należało powtórzyć materiał,

ewentualnie uzupełnić luki w wiedzy i dobrze wyspać się. Zdumiewająco duży wpływ na wynik egzaminu miała długość snu studentów. Wyrażono przypuszczenie, że osoby które pozwoliły sobie na długi sen mogły być bardzo uzdolnionymi studentami, nie obawiającymi się o wynik egzaminu.

LITERATURA

Greene W.H. (2000) Econometric Analysis. Prentice Hall Inc., Upper Saddle River, New Jersey.
Gruszczyński M. (2002) Modele i prognozy zmiennych jakościowych w finansach i bankowości. Oficyna Wydawnicza SGH. Warszawa.
Gujarati D. N. (2003) Basic Econometrics. McGraw Hill.
Judge G. G., Hill C., Griffiths W. E., Lütkepohl H., Lee T. (1985) The Theory and Practice of Econometrics, John Wiley&Sons. New York.
Maddala C. S. (2002) Introduction to Econometrics. John Wiley&Sons. New York.

Application of Logit Model to Analysis of Examination Results

Summary: The aim of this paper is application of logit model to explanation of examination results in econometrics on Interfaculty Studies in Computer Sciences and Econometrics. Logit model is the type of the binary choice models, where explained variable is dummy. Methods of estimation and measurement of goodness of fit are presented in the paper. Moreover interpretation of the estimated logit model is described. On the basis of estimated model it is found that systematic learning during semester and good rest immediately before writing examination paper increased probability of passing the examination.

Key words: logit model, dummy variable, result of examination.