Moduł 12
Metody statystyczne_L1

iBigWorld:
Innovations for Big Data in a Real World

Zepół UBB

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Wprowadzenie

- Statystyka - nauka zajmująca się sposobami pozyskiwania i prezentowania, a przede wszystkim analizowania danych opisujących zjawiska, w tym zjawiska masowe.

- Duża część nauki zajmuje się obserwacją otaczającego świata lub wykorzystuje eksperyment do potwierdzania swoich teorii. Badania takie zazwyczaj przebiegają według schematu: zbieranie danych, ich analiza i interpretacja.

Badania statystyczne zwykle podążają za wzorcem

Badacz potrzebuje zestawu narzędzi – sprawdzonych metod, które pozwolą mu operować na dużych zbiorach danych. Statystyka zajmuje się tworzeniem i rozwijaniem takich przydatnych narzędzi.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Metody statystyczne

- ## Gdzie i do czego wykorzystywane są metody statystyczne?

W badaniach naukowych

*poznać prawidłowości zjawisk masowych – gdzie badane są problemy demograficzne, ekonomiczne, socjologiczne

*do opisu zjawisk masowych (te zbiory danych mają postać tabel, z których można wyliczyć procent danego zjawiska

*określić zmienność zjawisk masowych, tendencje ich przemian w czasie

Na koniec rysuje się wykresy ilustrujące krzywą rozwoju danego zjawiska lub jaką część większej całości ono stanowi.

- ## Jaki jest cel stosowania metod statystycznych?
- Celem analizy statystycznej jest wydobycie jak największej wiedzy z uzyskanych danych

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Metody statystyczne

- Co należy zrobić, aby zbiór danych był dobrą podstawą do analizy statystycznej?

1. Uzgodnij, jaką wiedzę o badanym zjawisku mają dostarczyć dane

2. Zaplanuj badanie

3. Podsumuj zestaw danych obserwacyjnych, podkreślając trendy, ale rezygnując ze szczegółów

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Opis pojedynczego zbioru danych

Cel: Przeprowadź analizę opisową dla następującego przykładu.

Przykład: ***Dzięki poczcie pantoflowej i dobrej reklamie sieć społecznościowa DataSciencester rozrosła się do kilkudziesięciu użytkowników, a dyrektor ds. pozyskiwania funduszy prosi o przeanalizowanie, ilu znajomych mają użytkownicy sieci, aby mógł uwzględnić te dane w swoim „wyciągu na boisku”.***

Każdy zestaw danych charakteryzuje się sam:

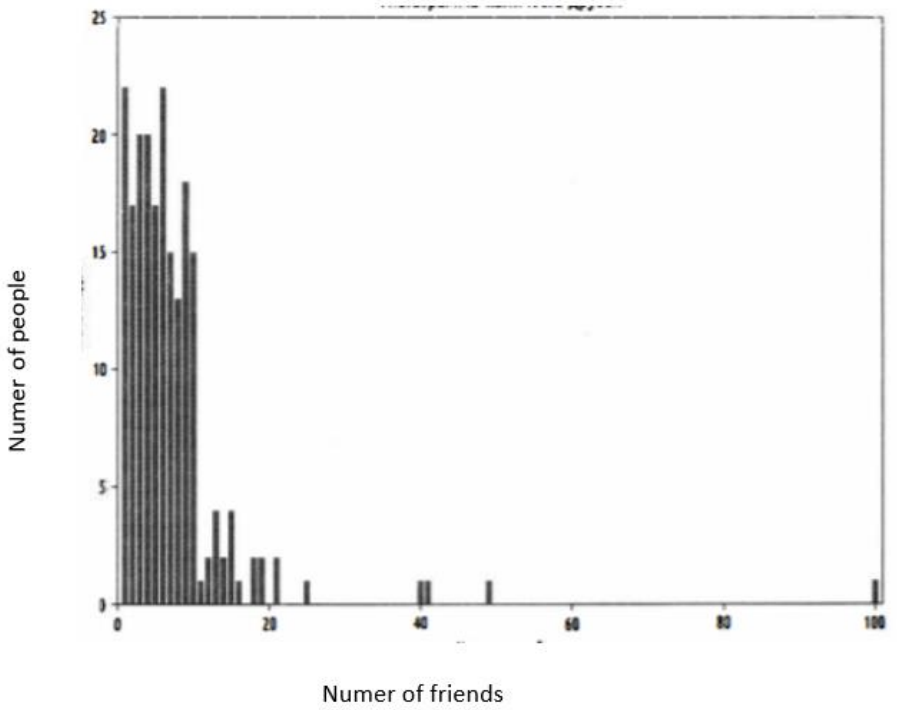
#liczba przyjaciół

nurn friends [100, 49, 41, 40, 25,

... Oraz inne

Aby umieścić liczbę przyjaciół w histogramie, użyj słownika zaliczeń i metod plt .bar.

```
friend_counts = Counter(num_friends)

xs range ( 101) # maximum equals 100

ys = (friend_counts(x] for x in xs] # height is number of friends

plt.bar(xs, ys)

plt.axis([O, 101, O, 25))

plt. title ("Histogram of the number of friends") Number of friends

plt.xlabel("Number of Friends")

plt.ylabel("Number of Friends” Number of friends

plt.show ()
```



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Opis pojedynczego zbioru danych. Generowanie statystyk.

Najprostsza statystyka – liczba punktów danych

Przejdźmy do generowania statystyk. Najprostsza statystyka to liczba punktów danych:

num_points len(num_friends) # liczba punktów to 204

Interesujące mogą być również wartości największe i najmniejsze; biggest_value = max(num_friends) # największa wartość to 100

smallest_value = min(num_friends) # najmniejsza wartość to 1

które są szczególnymi przypadkami, w których chcemy poznać wartości na określonych pozycjach:

sorted_values = sorted(num_friends)

smallest_value = sorted_values[0]

second_smallest_value = sorted_values[l]

second_largest_value = sorted_values[-2]

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Terminy i przykłady używane w statystykach

Opis pojedynczego zbioru danych

Tendencje centralne

a) Średnia wartość

b) Mediana

c) Kwantyl

d) Zmienność

c) Wariancja

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Tendencje centralne. Średnia wartość

Dla dwóch punktów średnia to punkt leżący pomiędzy nimi.

liczba różnych wartości cechy

liczba wystąpień wartości xi

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n}$$

```
# średnia wartość
dsf mean(xs: List[float]) -> float:
return sum(x) / len(x)
mean(num_friends) # 7.333333
```

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Tendencje centralne. Mediana

- Czasami interesuje nas również **mediana**, która jest najbliżej środka (jeśli liczba punktów danych jest nieparzysta) lub średnia arytmetyczna, przyjmowana jako połowa sumy dwóch wartości najbliższych centrum (jeśli liczba punktów jest parzysta)

Zwróć uwagę, że mediana – w przeciwieństwie do średniej – jest niezależna od żadnej wartość w zbiorze danych.

```
Napiszemy różne funkcje dla parzystych i nieparzystych przypadków i połączymy
je:
# Podkreślenia wskazują, że te funkcje są
# "prywatne", ponieważ mają się nazywać
# od naszej funkcji,
# a nie korzystać z dostępnej biblioteki statycznej
def _median_odd(xs: List[float]) -> float:

"""If len(xs) is odd,
then the median is the middle element"""
retuzn sorted(xs) [len(xs) // 2]
def _median_even(xs: List[float]) -> float:
"""If len(xs) is even, then it is the average of
of the two middle elements""
sorted_xs = sorted(xs)
hi_midpoint = len(xs) // 2 #e.g. length 4 => hi_midpoint 2
return (sorted_xs[hi_midpoint - 1] + sorted_xs[hi_midpoint]) / 2
def median(v: List[float]) -> float:
"""Finds the 'closest to the middle' value of v""
return _median_even(v) if len(v) % 2 == O else median_odd(v)
assert median( (1, 10, 2, 9, 5]) 5
assert median([l, 9, 2, 10]) == (2 + 9) / 2
And now we can calculate the median of the number of friends:
print(median(num_friends)) # 6
```

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Tendencje centralne. Kwantyl

Uogólnieniem mediany jest **Kwantyl**, który reprezentuje wartość poniżej którego znajduje się pewien percentyl danych (mediana reprezentuje wartość poniżej której leży 50% danych).

```
def quantile(xs: List[float], p: float) -> float:
'''Returns the value of the fifth percentile in xs'''.
p_index = int(p * len(x)) #converts% to list index
retuzn sorted(x) [p_index]
assert quantile(num_friends, O .10) 1
assert quantile(num_friends, 0.25) 3 # Lower quantile
assertile(num_friends, O .75) 9 # Upper quantile
assert quantile(num_friends, O. 90) 13
Less frequently, you may want to use mode, which is the
value or values that occur most frequently:
def mode(x: List[float]) -> List[float]:
"""Returns a list, since there may be more than one mod.""
counts = counter(x)
max_count = max(counts.values())
retuzn [x_i for x_i, count in counts.items()
if count == max_count]
assert set(mode(num_friends)) == {1, 6}
More often, however, we will simply use the average value.
```

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Tendencje centralne. Zmienność

Wariancja służy jako miara zmienności naszych danych.
Z reguły są to wskaźniki statystyczne, gdzie wartości bliskie zeru oznaczają brak zmienności, a duże wartości (cokolwiek to znaczy) bardzo dużą zmienność.
Na przykład najprostszym wskaźnikiem jest zakres, który jest zdefiniowany jako różnica między maksymalną i minimalną wartością danych:

```
# Słowo kluczowe „range" w Pythonie ma
# już ma swoje znaczenie, więc używamy innego
def data_range(xs: List[float]) -> float:
retuzn max(x) - min(x)
assert data_range(num_friends) == 99
```

Zakres wynosi zero, gdy max i min są równoważne, co zdarza się tylko gdy wszystkie elementy x są równe, a zatem rozproszenie danych jest nieobecne. I odwrotnie, gdy zakres jest szeroki, maksimum jest znacznie większe niż minimum, a rozrzut danych jest duży. Podobnie jak mediana, rozrzut nie jest szczególnie zależny od całego zestawu danych.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Tendencje centralne. Wariancja

Dokładniejszą miarą zmienności jest wariancja obliczona jako :
from scratch.linear_algebra i.mport surn_of_squares

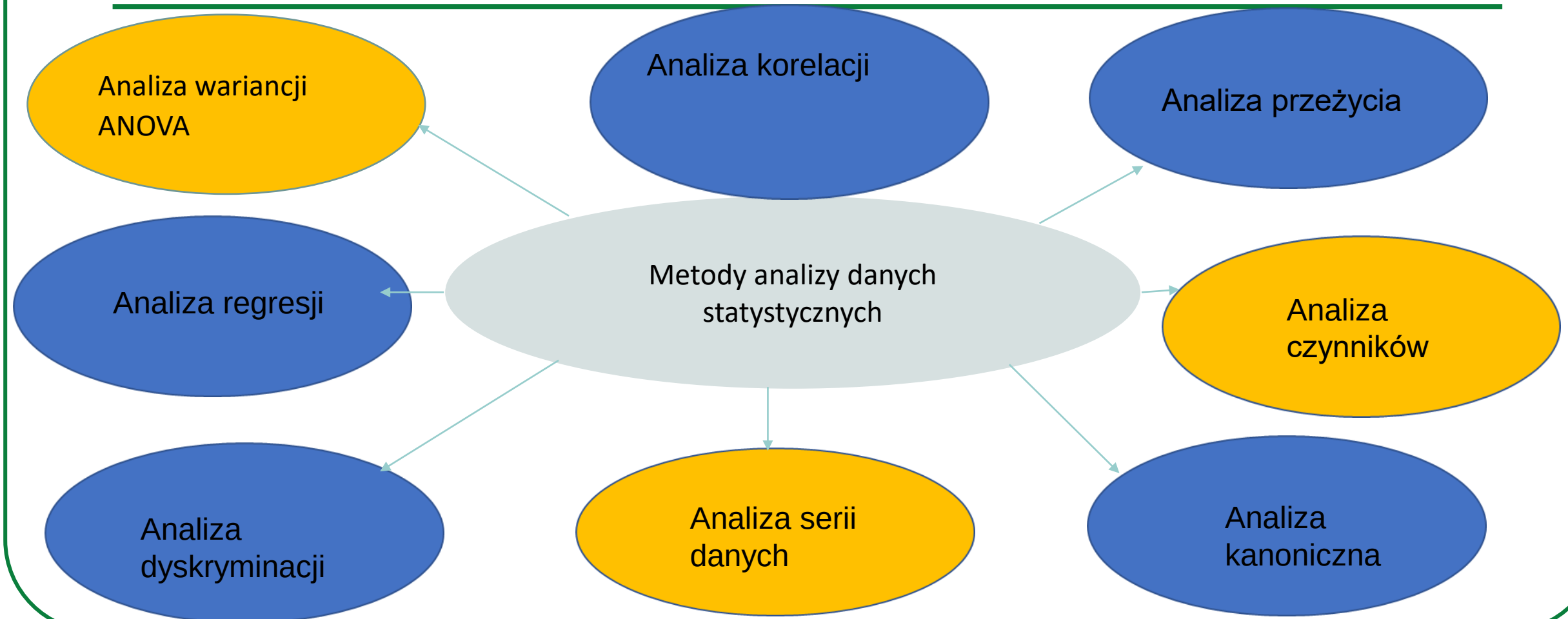
```
clef de_mean(xs: List[float]) -> List[float]:
"""Translate xs by subtracting its mean
(the result has zero mean)""
x_bar = mean(xs)
return [x - x_bar for x in xs]
"""clef variance(xs: List[float]) -> float:
"" "xs: "almost mean(xs)""
assert len(xs) >= 2, "variance requires at least two elements"
n = len(xs)
variance = de_mean(xs)
return surn_of_squares(variance) / (n - 1)
assert 81.54 < variance(nurn_friends) < 81.55
```

""The standard deviation is the square root of the variance""

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Metody analizy danych statystycznych



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza wariancji

- Analiza wariancji, ANOVA (z analizy wariancji) – metoda statystyczna służąca do badania obserwacji zależnych od jednego lub większej liczby jednocześnie działających czynników. Metoda wyjaśnia, z jakim prawdopodobieństwem wyodrębnione czynniki mogą powodować różnice między obserwowanymi średnimi grupowymi

Analizę modeli wariancji można podzielić na:

jednowymiarowa analiza wariancji
(wpływ każdego czynnika jest rozpatrywany osobno)

wielowymiarowa analiza wariancji
(wpływ różnych czynników jest rozpatrywany łącznie)

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza regresji

- Regresja liniowa - w modelowaniu statystycznym metody oparte na liniowych kombinacjach zmiennych i parametrów dopasowujących model do danych. Dopasowana linia lub krzywa regresji reprezentuje szacowaną oczekiwaną wartość zmiennej {displaystyle y} y przy określonych wartościach innej zmiennej lub zmiennych {displaystyle x}x.
- W najprostszym przypadku dopasowywana jest stała lub funkcja liniowa, zmienna y jest tradycyjnie nazywana zmienną objaśniającą lub zależną. Zmienne x nazywane są zmiennymi objaśniającymi lub zmiennymi niezależnymi. Zarówno zmienne objaśniające, jak i objaśniające mogą być wielkościami skalarnymi lub wektorami.
- Generalnie regresja jest warunkowym problemem szacowania wartości oczekiwanej. Regresję liniową nazywamy liniową, ponieważ założonym modelem relacji między zmienną zależną i niezależną jest przekształcenie liniowe (afiniczne) względem parametrów reprezentowanych w przypadku wielowymiarowym przez macierz.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza regresji - artykuł

An Introduction to Logistic Regression Analysis and Reporting

CHAO-YING JOANNE PENG
KUK LIDA LEE
GARY M. INGERSOLL
Indiana University-Bloomington

ABSTRACT **The purpose of this article is to provide researchers, editors, and readers with a set of guidelines for what to expect in an article using logistic regression techniques. Tables, figures, and charts that should be included to comprehensively assess the results and assumptions to be verified are discussed. This article demonstrates the preferred pattern for the application of logistic methods with an illustration of logistic regression applied to a data set in testing a research hypothesis. Recommendations are also offered for appropriate reporting formats of logistic regression results and the minimum observation-to-predictor ratio. The authors evaluated the use and interpretation of logistic regression presented in 8 articles published in *The Journal of Educational Research* between 1990 and 2000. They found that all 8 studies met or exceeded recommended criteria.**

Key words: binary data analysis, categorical variables, dichotomous outcome, logistic modeling, logistic regression

Many educational research problems call for the analysis and prediction of a dichotomous outcome: & Kravitz, 1994; Tolman & Weisz, 1995) and in educational research—especially in higher education (Austin, Yaffee, & Hinkle, 1992; Cabrera, 1994; Peng & So, 2002a; Peng, So, Stage, & St. John, 2002. With the wide availability of sophisticated statistical software for high-speed computers, the use of logistic regression is increasing. This expanded use demands that researchers, editors, and readers be attuned to what to expect in an article that uses logistic regression techniques. What tables, figures, or charts should be included to comprehensibly assess the results? What assumptions should be verified? In this article, we address these questions with an illustration of logistic regression applied to a data set in testing a research hypothesis. Recommendations are also offered for appropriate reporting formats of logistic regression results and the minimum observation-to-predictor ratio. The remainder of this article is divided into five sections: (1) Logistic Regression Models, (2) Illustration of Logistic Regression Analysis and Reporting, (3) Guidelines and Recommendations, (4) Evaluations of Eight Articles Using Logistic Regression, and (5) Summary

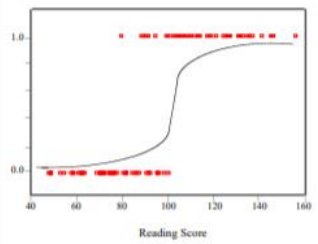
a boy's odds of being recommended for remedial reading instruction relative to a girl's odds. The result is an odds ratio of 2.33, which suggests that boys are 2.33 times more likely, than not, to be recommended for remedial reading classes compared with girls. The odds ratio is derived from two odds (73/23 for boys and 15/11 for girls); its natural logarithm [i.e., ln(2.33)] is a logit, which equals 0.85. The value of 0.85 would be the regression coefficient of the gender predictor if logistic regression were used to model the two outcomes of a remedial recommendation as it relates to gender.

Generally, logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous predictor variables. In the simplest case of linear regression for one continuous predictor X (a child's reading score on a standardized test) and one dichotomous outcome variable Y (the child being recommended for remedial reading classes), the plot of such data results in two parallel lines, each corresponding to a value of the dichotomous outcome (Figure 1). Because the two parallel lines are difficult to be described with an ordinary least squares regression equation due to the dichotomy of outcomes, one may instead create categories for the predictor and compute the mean of the outcome variable for the respective categories. The resultant plot of categories' means will appear linear in the middle, much like what one would expect to see on an ordinary scatter plot, but curved at the ends (Figure 1, the S-shaped curve). Such a shape, often referred to as sigmoidal or S-shaped, is difficult to describe with a linear equation for two reasons. First, the extremes do not follow a linear trend. Second, the errors are neither normally distributed nor constant across the entire range of data (Peng, Manz, & Keck, 2001). Logistic regression solves these problems by applying the logit transformation to the dependent variable. In essence, the logistic model predicts the logit of Y from X. As stated earlier, the logit is the natural logarithm (ln) of odds of Y, and odds are ratios of probabilities (π) of Y happening (i.e., a student is recommended for remedial reading instruction) to probabilities ($1 - \pi$) of Y not happening (i.e., a student is not recommended for remedial reading instruction). Although logistic regression can accommodate categorical outcomes that are polytomous, in this article we focus on dichotomous outcomes only. The illustration presented in this article can be extended easily to polytomous variables with ordered (i.e., ordinal-scaled) or unordered (i.e., nominal-scaled) outcomes.

Table 1.—Sample Data for Gender and Recommendation for Remedial Reading Instruction

Remedial reading instruction	Gender: Boys	Gender: Girls	Total
Recommended (coded as 1)	73	15	88
Not recommended (coded as 0)	23	11	34
Total	96	26	122

Figure 1. Relationship of a Dichotomous Outcome Variable, Y (1 = Remedial Reading Recommended, 0 = Remedial Reading Not Recommended) With a Continuous Predictor, Reading Scores



The simple logistic model has the form

$$\text{logit}(Y) = \textit{natural}\log(\textit{odds}) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X. \quad (1)$$

For the data in Table 1, the regression coefficient (β) is the logit (0.85) previously explained. Taking the antilog of Equation 1 on both sides, one derives an equation to predict the probability of the occurrence of the outcome of interest as follows:

$$\pi = \textit{Probability}(Y = \text{outcome of interest} \mid X = x,$$

$$\textit{a specific value of } X) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}, \quad (2)$$

where π is the probability of the outcome of interest or "event," such as a child's referral for remedial reading classes, α is the Y intercept, β is the regression coefficient, and $e = 2.71828$ is the base of the system of natural logarithms. X can be categorical or continuous, but Y is always categorical. According to Equation 1, the relationship between logit (Y) and X is linear. Yet, according to Equation 2, the relationship between the probability of Y and X is nonlinear. For this reason, the natural log transformation of the odds in Equation 1 is necessary to make the relationship between a categorical outcome variable and its predictor(s) linear.

The value of the coefficient β determines the direction of the relationship between X and the logit of Y. When β is greater than zero, larger (or smaller) X values are associated with larger (or smaller) logits of Y. Conversely, if β is less than zero, larger (or smaller) X values are associated with smaller (or larger) logits of Y. Within the framework of inferential statistics, the null hypothesis states that β equals zero, or there is no linear relationship in the population. Rejecting such a null hypothesis implies that a linear relationship exists between X and the logit of Y. If a predictor is binary, as in the Table 1 example, then the odds ratio is equal to e, the natural logarithm base, raised to the exponent of the slope B (e^B).

The Journal of Educational Research

September/October 2002 [Vol. 96(No. 1)]

Extending the logic of the simple logistic regression to multiple predictors (say X_1 = reading score and X_2 = gender), one can construct a complex logistic regression for Y (recommendation for remedial reading programs) as follows:

$$\text{logit}(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2. \quad (3)$$

Therefore,

$$\pi = \text{Probability}(Y = \textit{outcome of interest} \mid X_1 = x_1, X_2 = x_2 = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2}}{1+e^{\alpha+\beta_1 X_1+\beta_2 X_2}}, \quad (4)$$

where π is once again the probability of the event, α is the Y intercept, βs are regression coefficients, and Xs are a set of predictors. α and βs are typically estimated by the maximum likelihood (ML) method, which is preferred over the weighted least squares approach by several authors, such as Haberman (1978) and Schlesselman (1982). The ML method is designed to maximize the likelihood of reproducing the data given the parameter estimates. Data are entered into the analysis as 0 or 1 coding for the dichotomous outcome, continuous values for continuous predictors, and dummy codings (e.g., 0 or 1) for categorical predictors.

The null hypothesis underlying the overall model states that all βs equal zero. A rejection of this null hypothesis implies that at least one β does not equal zero in the population, which means that the logistic regression equation predicts the probability of the outcome better than the mean of the dependent variable Y. The interpretation of results is rendered using the odds ratio for both categorical and continuous predictors.

Illustration of Logistic Regression Analysis and Reporting

For the sake of illustration, we constructed a hypothetical data set to which logistic regression was applied, and we interpreted its results. The hypothetical data consisted of reading scores and genders of 189 inner city school children (Appendix A). Of these children, 59 (31.22%) were recommended for remedial reading classes and 130 (68.78%) were not. A legitimate research hypothesis posed to the data was that "the likelihood that an inner city school child is recommended for remedial reading instruction is related to both his/her reading score and gender." Thus, the outcome variable, remedial, was students being recommended for remedial reading instruction (1 = yes, 0 = no), and the two predictors were students' reading score on a standardized test (X_1 = the reading variable) and gender (X_2 = gender). The reading scores ranged from 40 to 125 points, with a mean of 64.91 points and standard deviation of 15.29 points (Table 2). The gender predictor was coded as 1 = boy and 0 = girl. The gender distribution was nearly even with 49.21% (n = 93) boys and 50.79% (n = 96) girls.

Table 2.—Description of a Hypothetical Data Set for Logistic Regression

Remedial reading recommended?	Total sample (N)	Boys (n_1)	Girls (n_2)	Reading score: M	Reading score: SD
Yes	59	36	23	61.07	13.28
No	130	57	73	66.65	15.86
Summary	189	93	96	64.91	15.29

Logistic Regression Analysis

A two-predictor logistic model was fitted to the data to test the research hypothesis regarding the relationship between the likelihood that an inner city child is recommended for remedial reading instruction and his or her reading score and gender. The logistic regression analysis was carried out by the Logistic procedure in SAS® version 8 (SAS Institute Inc., 1999) in the Windows 2000 environment (SAS programming codes are found in Table 3). The result showed that

$$\text{Predicted logit of (REMEDIAL)} = 0.5340 + (-0.0261)\text{*READING} + (0.6477)\text{*GENDER}. \quad (5)$$

According to the model, the log of the odds of a child being recommended for remedial reading instruction was negatively related to reading scores ($p < .05$) and positively related to gender ($p < .05$; Table 3). In other words, the higher the reading score, the less likely it is that a child would be recommended for remedial reading classes. Given the same reading score, boys were more likely to be recommended for remedial reading classes than girls because boys were coded to be 1 and girls 0. In fact, the odds of a boy being recommended for remedial reading programs were 1.9111 ($= e^{0.6477}$; Table 3) times greater than the odds for a girl.

The differences between boys and girls are depicted in Figure 2, in which predicted probabilities of recommendations are plotted for each gender group against various reading scores. From this figure, it may be inferred that for a given score on the reading test (e.g., 60 points), the probability of a boy being recommended for remedial reading programs is higher than that of a girl. This statement is also confirmed by the positive coefficient (0.6477) associated with the gender predictor.

Evaluations of the Logistic Regression Model

How effective is the model expressed in Equation 5? How can an educational researcher assess the soundness of a logistic regression model? To answer these questions, one must attend to (a) overall model evaluation, (b) statistical tests of individual predictors, (c) goodness-of-fit statistics, and (d) validations of predicted probabilities. These evaluations are illustrated below for the model based on Equation 5, also referred to as Model 5.

Overall model evaluation. A logistic model is said to provide a better fit to the data if it demonstrates an improvement over the intercept-only model (also called the null model). An

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Dziękuję bardzo

Thank You very much

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union