Regresja, Anova

Stanisław Jaworski

UM
Zakład Profilaktyki....

Przykład

W pewnej klinice badano związek między aktywnością enzymów aminotransferazy a stężeniem amoniaku we krwi u chorych z ostrą niewydolnością wątroby. Pobrano losową próbę 10 pacjentów i otrzymano następujące wyniki:

aktywność	430	470	520	570	630	690	740	770	800	780
stężenie	31	33	36	39	42	47	51	54	55	57

Pytania

1. Czy poziom aktywności enzymów zależy od stężenia amoniaku we krwi u dziesięciu badanych pacjentów?
2. Czy w przypadku istnienia takiej zależności, można ją przedstawić w zwartej formie, za pomocą funkcji?

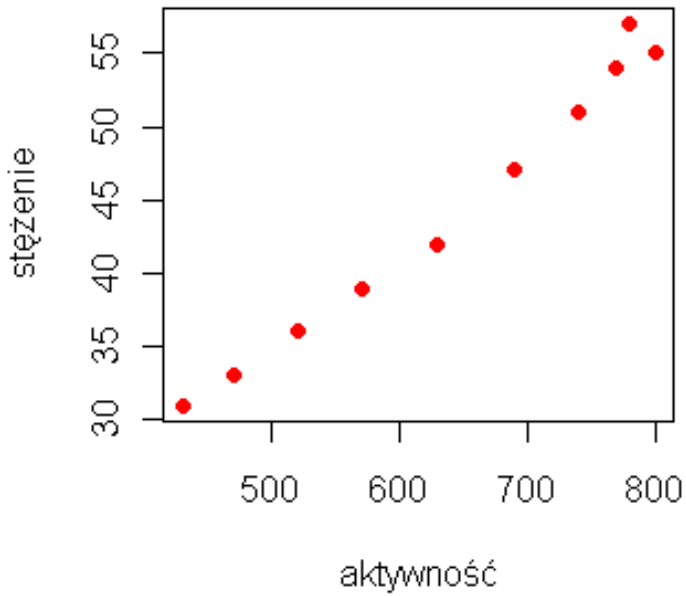
Wprowadzenie danych

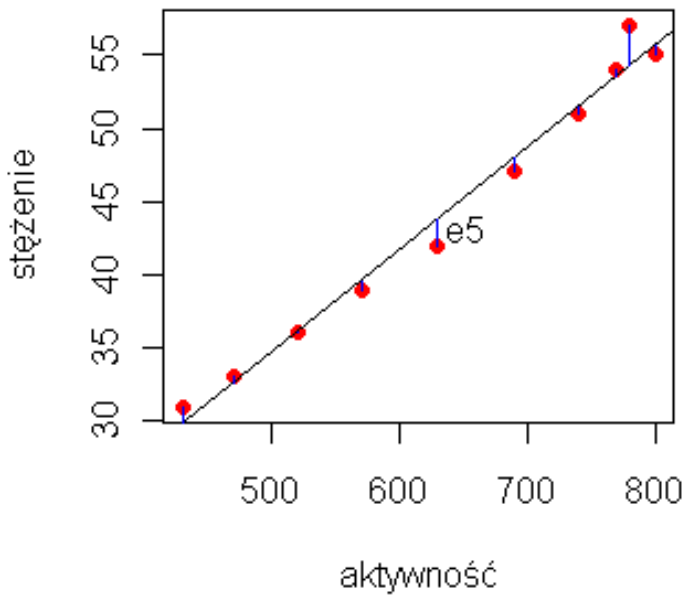
```
> aktywność< −c(430,470,520,570,630,690,740,770,800,780)
> stężenie< −c(31,33,36,39,42,47,51,54,55,57)
```

Utworzenie wykresu

```
> plot(aktywność,stężenie,pch=19,col="red")
```

Punkty na wykresie są wypełnione: pch=19.
Punkty są koloru czerwonego: col="red".





Ocena wykresu daje pewne podstawy do przyjęcia następującej zależności:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, n,$$

gdzie
y_i – stężenie amoniaku we krwi i – tego pacjenta
x_i – aktywność enzymu w organiźmie i – tego pacjenta
e_i – błąd dopasowania

Metoda najmniejszych kwadratów

Parametry β_0 oraz β_1 dobieramy tak, aby średniokwadratowy błąd dopasowania, mianowicie $\sum_i e_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$, był minimalny. W ten sposób dobrane parametry oznaczamy przez $\hat{\beta}_0$ oraz $\hat{\beta}_1$. Wyrażają się one wzorami

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x}_i)^2}, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

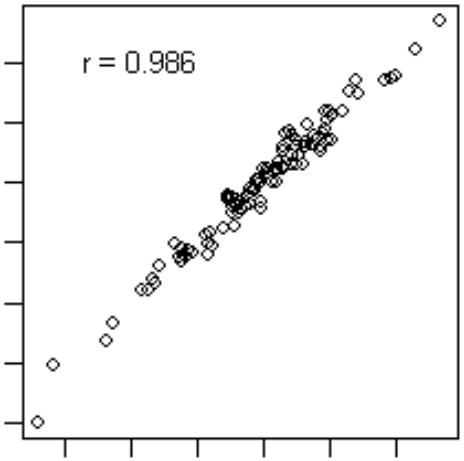
Wówczas

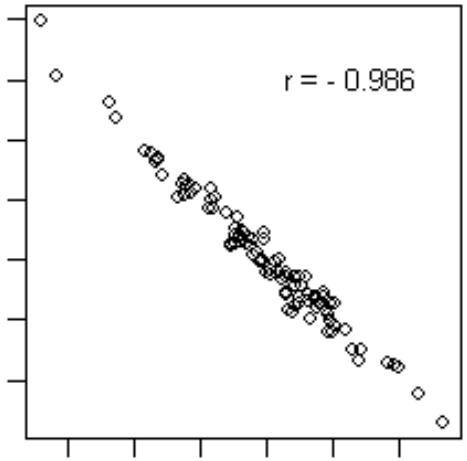
$$\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = (1 - r^2) \sum_i (y_i - \bar{y})^2,$$

gdzie

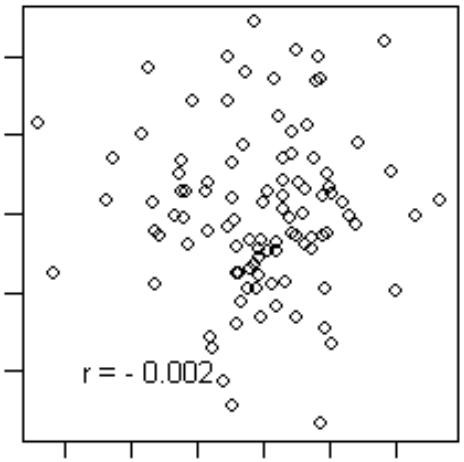
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (y_i - \bar{y}_i)^2 \sum_i (x_i - \bar{x})^2}}$$

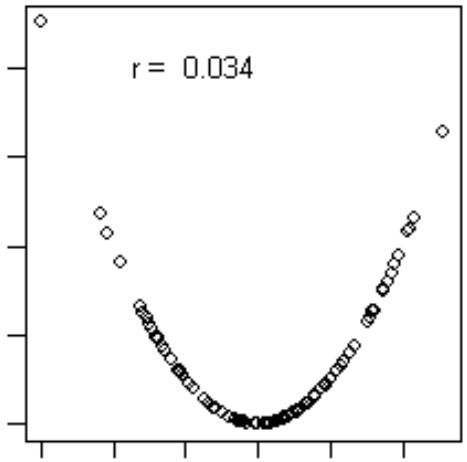
Współczynnik r jest miernikiem zależności liniowej.





Wartość r jest zawsze z przedziału <-1,1>





Pytania

1. Czy poziom aktywności enzymów zależy od stężenia amoniaku we krwi u wszytkich pacjentów w zbiorowości, z której wylosowano 10 pacjentów do badań?

2. W jaki sposób wyznaczyć ewentualną zależność liniową między aktywnością enzymów a stężeniem amoniaku? W jaki sposób sprawdzić wiarygodność tej zależności?

Aby odpowiedzieć na postawione pytania, musimy przyjąć model statystyczny, który pozwoli nam na uogólnienie wniosków z próby 10 pacjentów na całą populację pacjentów.

Model

Przyjmujemy następujący model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

gdzie $\varepsilon_i, \quad i = 1, 2, \ldots, n$, są niezależnymi zmiennymi losowymi o tym samym rozkładzie $N(0, \sigma^2)$.

Uwagi

- $Y_1, Y_2, \ldots, Y_n$, są zmiennymi losowymi, a $y_1, y_2, \ldots, y_n$ są ich realizacjami.
- Model dotyczy rozkładu warunkowego $Y|X = x$.

Model

Przyjmujemy następujący model:

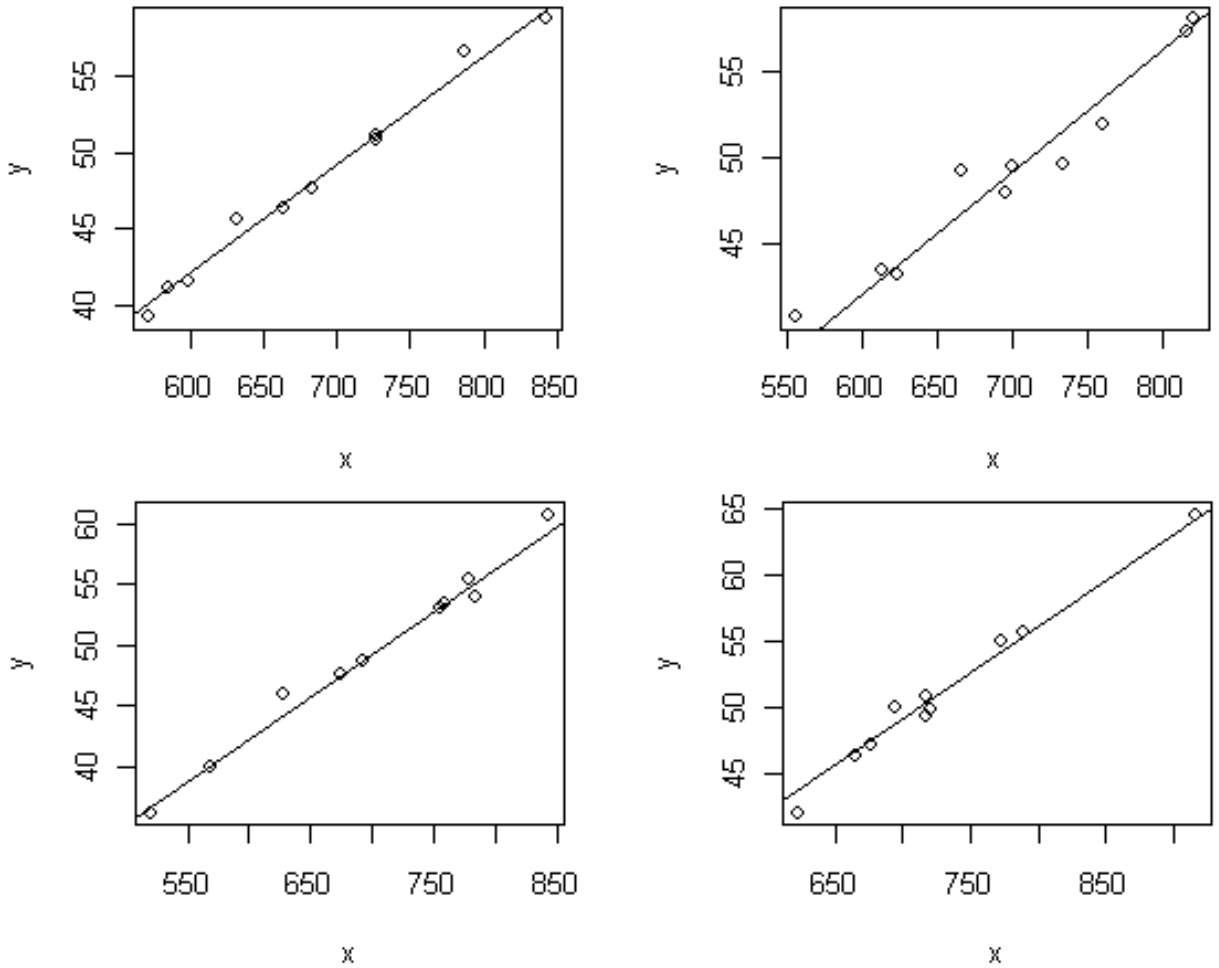
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

gdzie $\varepsilon_i, \quad i = 1, 2, \ldots, n$, są niezależnymi zmiennymi losowymi o tym samym rozkładzie $N(0, \sigma^2)$.

Uwagi

- $Y_1, Y_2, \ldots, Y_n$, są zmiennymi losowymi, a $y_1, y_2, \ldots, y_n$ są ich realizacjami.
- Model dotyczy rozkładu warunkowego $Y|X = x$.

Cztery hipotetyczne realizacje doświadczenia według modelu. Takich realizacji jest nieskończenie wiele. Wśród nich znajduje się rzeczywiste doświadczenie.



Estymacja

- $\hat{\beta}_0$, $\hat{\beta}_1$ są oszacowaniami punktowymi parametrów β_0 oraz β_1.
- Oszacowania przedziałowe dla β_0 oraz β_1 są postaci

$$\beta_1 \in (\hat{\beta}_1 - t(\alpha; n-2)S_{\beta_1}, \hat{\beta}_1 + t(\alpha; n-2)S_{\beta_1})$$

$$\beta_0 \in (\hat{\beta}_0 - t(\alpha; n-2)S_{\beta_0}, \hat{\beta}_0 + t(\alpha; n-2)S_{\beta_0})$$

gdzie

$$S^2_{\beta_1} = \frac{S^2}{\mathrm{var}x}, \quad S^2_{\beta_0} = \frac{S^2}{\mathrm{var}x}\left(\frac{\mathrm{var}x}{n} + \bar{x}^2\right)$$

$$S^2 = \frac{\mathrm{var}y - \hat{\beta}_1 \mathrm{cov}(x,y)}{n-2} = \frac{\mathrm{var}y(1-r^2)}{n-2}$$

Estymacja

- $\hat{\beta}_0$, $\hat{\beta}_1$ są oszacowaniami punktowymi parametrów β_0 oraz β_1.
- Oszacowania przedziałowe dla β_0 oraz β_1 są postaci

$$\beta_1 \in (\hat{\beta}_1 - t(\alpha; n-2)S_{\beta_1}, \hat{\beta}_1 + t(\alpha; n-2)S_{\beta_1})$$

$$\beta_0 \in (\hat{\beta}_0 - t(\alpha; n-2)S_{\beta_0}, \hat{\beta}_0 + t(\alpha; n-2)S_{\beta_0})$$

gdzie

$$S^2_{\beta_1} = \frac{S^2}{\mathrm{var}x}, \quad S^2_{\beta_0} = \frac{S^2}{\mathrm{var}x}\left(\frac{\mathrm{var}x}{n} + \bar{x}^2\right)$$

$$S^2 = \frac{\mathrm{var}y - \hat{\beta}_1 \mathrm{cov}(x,y)}{n-2} = \frac{\mathrm{var}y(1-r^2)}{n-2}$$

Weryfikacja hipotez

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Statystyka testowa

$$F_{\text{emp}} = \frac{\hat{\beta}_1^2}{S_{\beta_1}^2} = \frac{\hat{\beta}_1 \text{cov}(x, y)}{S^2}$$

Hipotezę odrzucamy, jeżeli $F_{\text{emp}} > F(\alpha; 1, n-2)$.

$F(\alpha; 1, n-2)$ jest wartością krytyczną rozkładu F.

Weryfikacja hipotez

$$H_0 : \beta_1 = a$$
$$H_1 : \beta_1 \neq a$$

Statystyka testowa

$$t_{\text{emp}} = \frac{\hat{\beta}_1 - a}{S_{\beta_1}}$$

Hipotezę odrzucamy, jeżeli $|t_{\text{emp}}| > t(\alpha; n-2)$.

$t(\alpha; n-2)$ jest wartością krytyczną rozkładu $t-$Studenta.

Obliczenia w R dla podanego przykładu

```
> model<-lm(stężenie~aktywność)
> summary(model)

Call:
lm(formula = stężenie ~ aktywność)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8018 -0.6566 -0.3018  0.4099  2.7251

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.185030   2.143159  -0.086    0.933
aktywność    0.069820   0.003282  21.271 2.51e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.341 on 8 degrees of freedom
Multiple R-squared: 0.9826,     Adjusted R-squared: 0.9805
F-statistic: 452.5 on 1 and 8 DF,  p-value: 2.509e-08
```

Obliczenia w R dla podanego przykładu

```
> confint(model,level=0.95)
                    2.5 %     97.5 %
(Intercept) -5.12716355 4.75710367
aktywność    0.06225104 0.07738968
```

Zgodnie z oznaczeniami oraz obliczeniami w R mamy:

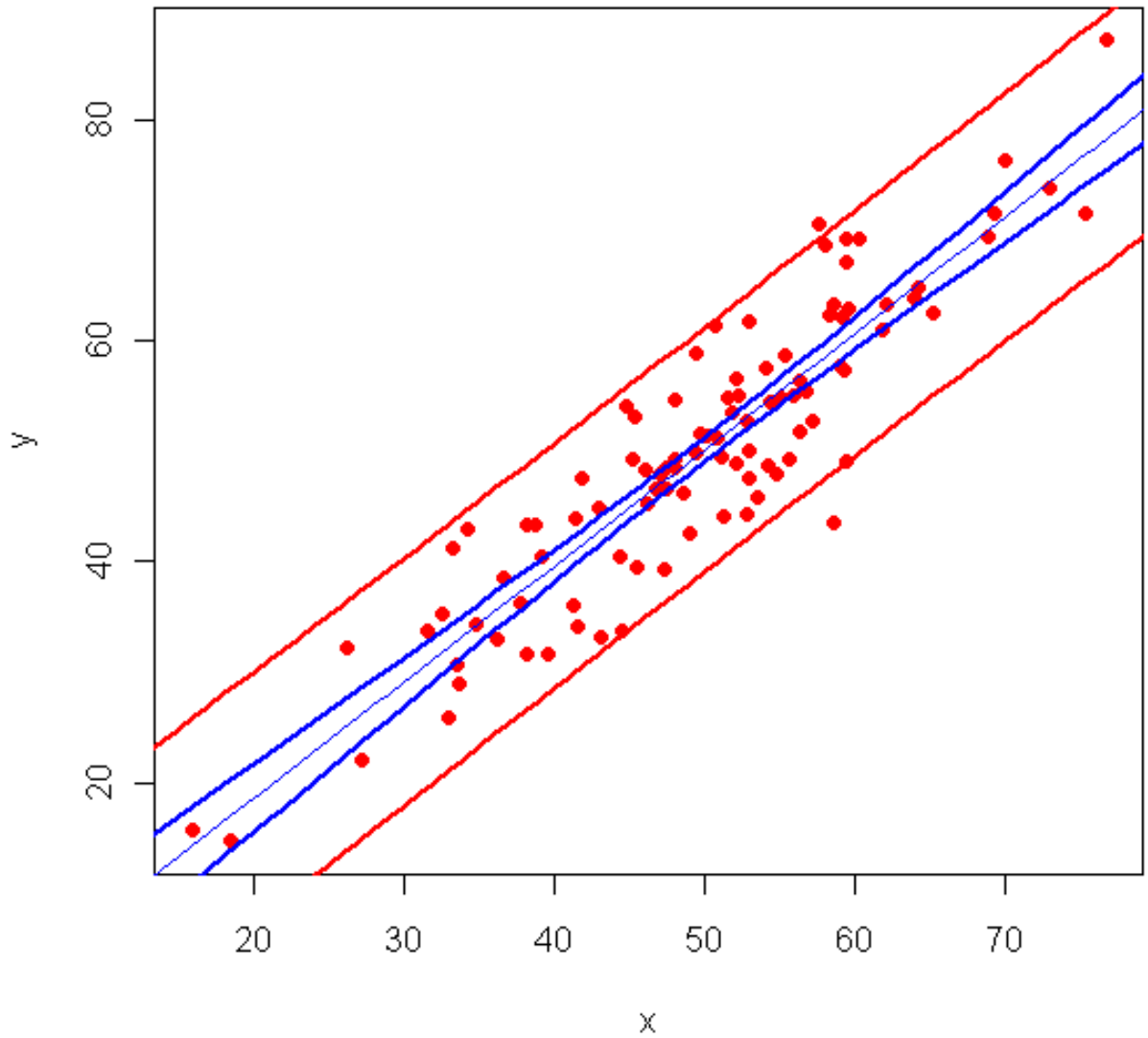
$$r^2 = 0.9826,\ F_{\text{emp}} = 452.5,\ t_{\text{emp}} = 21.271 \text{ (przy } a = 0)$$

Przedział ufności dla β_1 (na poziomie ufności 0.95):

$$(0.06225104,\ 0.07738968)$$

Przedział ufności dla β_0 (na poziomie ufności 0.95):

$$(-5.12716355, 4.75710367)$$


y
20
40
60
80
20
30
40
50
60
70
x

Obszar ufności dla prostej regresji

Obszar ufności dla prostej regresji umożliwia nam wnioskowanie o wartościach średnich zmiennej Y jednocześnie dla wielu wybranych wartości zmiennej X.

$$f(x) \in (\hat{f}(x) - t(\alpha; n-2)S_Y; \hat{f}(x) + t(\alpha; n-2)S_Y)$$

gdzie

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$S_Y^2 = S^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\text{var}x} \right)$$

Obszar predykcji

Obszar predykcji umożliwia nam wnioskowanie o wartościach zmiennej Y jednocześnie dla wielu wybranych wartości zmiennej X.

$$Y(x) \in (\hat{f}(x) - t(\alpha; n-2)S_{Y(x)}; \hat{f}(x) + t(\alpha; n-2)S_{Y(x)})$$

gdzie $Y(x)$ oznacza wartość zmiennej Y dla wybranej wartości x zmiennej X oraz

$$S^2_{Y(x)} = S^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\text{var}x}\right)$$

Obliczenia w R dla podanego przykładu

```
> predict(model,data.frame(aktywność=c(635)),level=0.95,interval="confidence")
      fit      lwr      upr
1 44.1509 43.17199 45.1298
> predict(model,data.frame(aktywność=c(635)),level=0.95,interval="prediction")
      fit      lwr      upr
1 44.1509 40.90645 47.39535
```

Zgodnie z oznaczeniami oraz obliczeniami w R mamy (na poziomie ufności 0.95):
Przedział ufności dla $f(635)$: (43.17199, 45.1298)
Przedział ufności dla $Y(635)$: (40.90645, 47.39535)

Model

Przyjmujemy następujący model:

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

gdzie $\varepsilon_i, \quad i = 1, 2, \ldots, n$, są niezależnymi zmiennymi losowymi o tym samym rozkładzie $N(0, \sigma^2)$.

Uwagi

- $Y_1, Y_2, \ldots, Y_n$, są zmiennymi losowymi, a $y_1, y_2, \ldots, y_n$ są ich realizacjami.
- Model dotyczy rozkładu warunkowego $Y|X = x$.

Model

Przyjmujemy następujący model:

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

gdzie $\varepsilon_i, \quad i = 1, 2, \ldots, n$, są niezależnymi zmiennymi losowymi o tym samym rozkładzie $N(0, \sigma^2)$.

Uwagi

- $Y_1, Y_2, \ldots, Y_n$, są zmiennymi losowymi, a $y_1, y_2, \ldots, y_n$ są ich realizacjami.
- Model dotyczy rozkładu warunkowego $Y|X = x$.

Przykład

- Y (Poziom) – liczba urodzin w przeliczeniu na 1000 osób
- X_1 (Status) – status socjoekonomiczny (wypadkowa poziomu wykształcenia, oczekiwań życiowych, śmiertelności niemowląt, frakcji kobiet w wieku 15-64 zatrudnionych poza rolnictwem, produktu krajowego brutto, frakcji populacji żyjącej w miastach) par
- X_2 (Planowanie) – indeks planowania rodziny (odzwierciedla poziom zaangażowania państwa w planowanie struktury rodziny)

Kraj	Status	Planowanie	Poziom
Boliwia	46	0	1
Brazylia	74	0	10
Chile	89	16	29
Kolumbia	77	16	25
Kostaryka	84	21	29
Kuba	89	15	40
Dominikana	68	14	21
Ekwador	70	6	0
Salwador	60	13	13
Gwatemala	55	9	4
Haiti	35	3	0
Honduras	51	7	7
Jamajka	87	23	21
Meksyk	83	4	9
Nikaragua	68	0	7
Panama	84	19	22
Paragwaj	74	3	6
Peru	73	0	2
Trynidad i Tobago	84	15	29
Wenezuela	91	7	11

Przypuśćmy, że mamy dane wprowdzone do R pod zmienną *dane*:

```
> head(dane)
       Kraj Status Planowanie Poziom
1   Boliwia     46          0      1
2  Brazylia     74          0     10
3     Chile     89         16     29
4  Kolumbia     77         16     25
5 Kostaryka     84         21     29
6      Kuba     89         15     40
>
```

Chcemy za pomocą metody najmniejszych kwadratów oszacować współczynniki funkcji regresji. Przyjmujemy model regresji:

$$\text{Poziom}_i = \beta_0 + \beta_1 \text{Status}_i + \beta_2 \text{Planowanie}_i + \varepsilon_i$$

gdzie $i = 1, \ldots, 20$.

```
> model<-lm(Poziom~Status+Planowanie,data=dane)
> summary(model)

Call:
lm(formula = Poziom ~ Status + Planowanie, data = dane)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3475  -3.6426   0.6384   3.2250  15.8530

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.4511     7.0938  -2.037 0.057516 .
Status        0.2706     0.1079   2.507 0.022629 *
Planowanie    0.9677     0.2250   4.301 0.000484 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.389 on 17 degrees of freedom
Multiple R-squared: 0.7381,     Adjusted R-squared: 0.7073
F-statistic: 23.96 on 2 and 17 DF,  p-value: 1.132e-05
```

Oszacowane współczynniki:
$\hat{\beta}_0 = -14.4510978, \quad \hat{\beta}_1 = 0.2705885, \quad \hat{\beta}_2 = 0.9677137$

Zanim przejdziemy do zinterpretowania tych współczynników, należy sprawdzić, czy różnią się istotnie od zera.
W pierwszej kolejności należy zweryfikować hipotezę

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

Oznacza ona, że **Status** oraz **Planowanie** nie wyjaśniają różnic w **Poziomie** urodzeń w poszczególnych krajach.

Do weryfikacji tej hipotezy stosujemy test F. Wartość tej statystyki wynosi $F_{\text{emp}} = 23.96$. Hipotezę odrzucamy (na poziomie istotności α), jeżeli $F_{\text{emp}} > F(\alpha, p, n - p - 1)$.

$F(0.05, 2, 17) = 4.618874$

Ponieważ $F_{\text{emp}} = 23.96 > F(0.05, 2, 17)$ (równoważnie $p - value = 1.132e - 05 = 0.00001132 < \alpha = 0.05$), hipotezę odrzucamy. Wniosek jest taki, że przynajmniej jedna cecha objaśniająca (Status, Planowanie) wyjaśnia poziom urodzeń w poszczególnych krajach. Następnym krokiem jest weryfikacja hipotez cząstkowych. Postać $i-$tej hipotezy cząskowej jest następująca:

$$H_0 : \ \beta_i = 0$$

Na poziomie istotności $\alpha = 0.05$ weryfikujemy tę hipotezę za pomocą testu t-Studenta. Statystyka testowa ma postać:

$$t_{\text{emp}} = \hat{\beta}_i / S_{\beta_i}$$

Jeżeli $t_{\text{emp}} > t(\alpha, n - p - 1)$, to hipotezę odrzucamy.

W podanym przykładzie możemy na przykład zweryfikować hipotezę

$$H_0 : \ \beta_1 = 0$$

mówiącą, że poziom urodzeń nie zależy od statusu socjoekonomicznego. Statystyka testowa dla tej hipotezy wynosi

$$t_{\text{emp}} = 0.2706/0.1079 = 2.507,$$

a wartość krytyczna $t(0.05, 17) = 2.109816$. Hipotezę zatem odrzucamy na poziomie istotności 0.05.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.4511     7.0938  -2.037 0.057516 .
Status        0.2706     0.1079   2.507 0.022629 *
Planowanie    0.9677     0.2250   4.301 0.000484 ***
```

Podobnie odrzucamy hipotezę

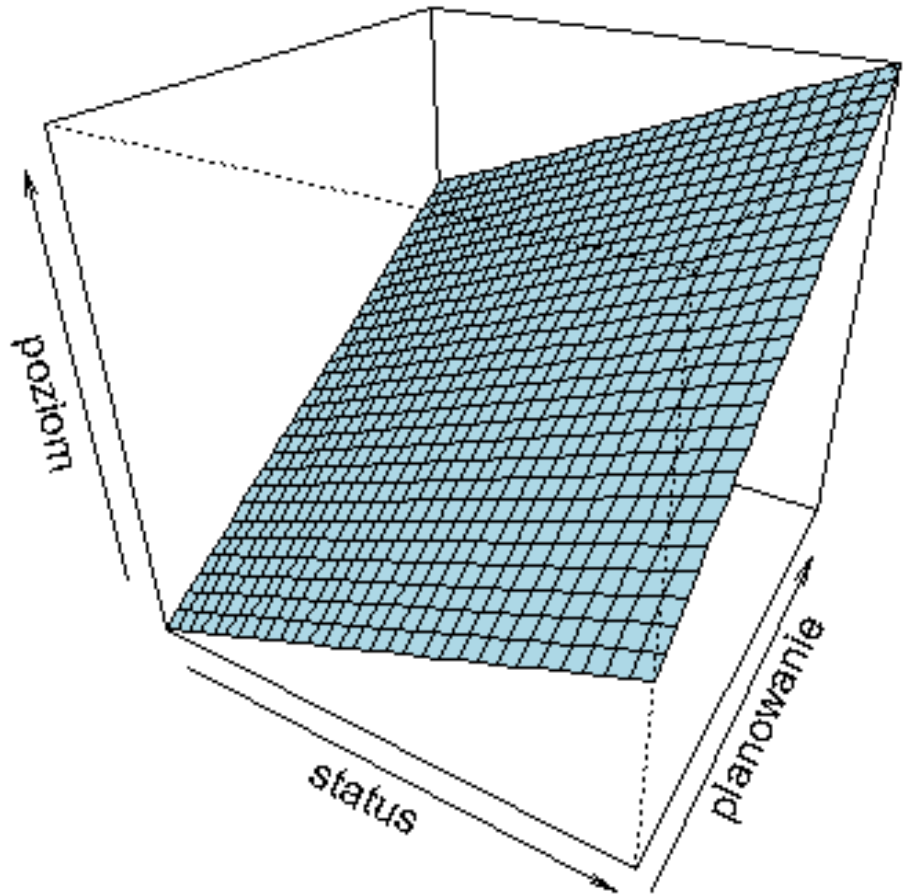
$$H_0 : \ \beta_2 = 0$$

Wniosek końcowy jest taki, że za pomocą cech Status oraz Planowanie możemy lepiej wyjaśnić zmienność liczby urodzeń w badanych krajach, niż byśmy mogli to zrobić bez tych cech. Pewnym miernikiem dopasowania modelu do danych jest współczynik korelacji wielokrotnej R, który jest liczbą z przedziału $(0, 1)$. W naszym przykładzie $R^2 = 0.7381$

Możemy teraz zinterpretować współczynniki funkcji regresji. Oszacowanie tej funkcji ma postać:

$$\text{Poziom} = \hat{\beta}_0 + \hat{\beta}_1 \text{Status} + \hat{\beta}_2 \text{Planowanie}$$

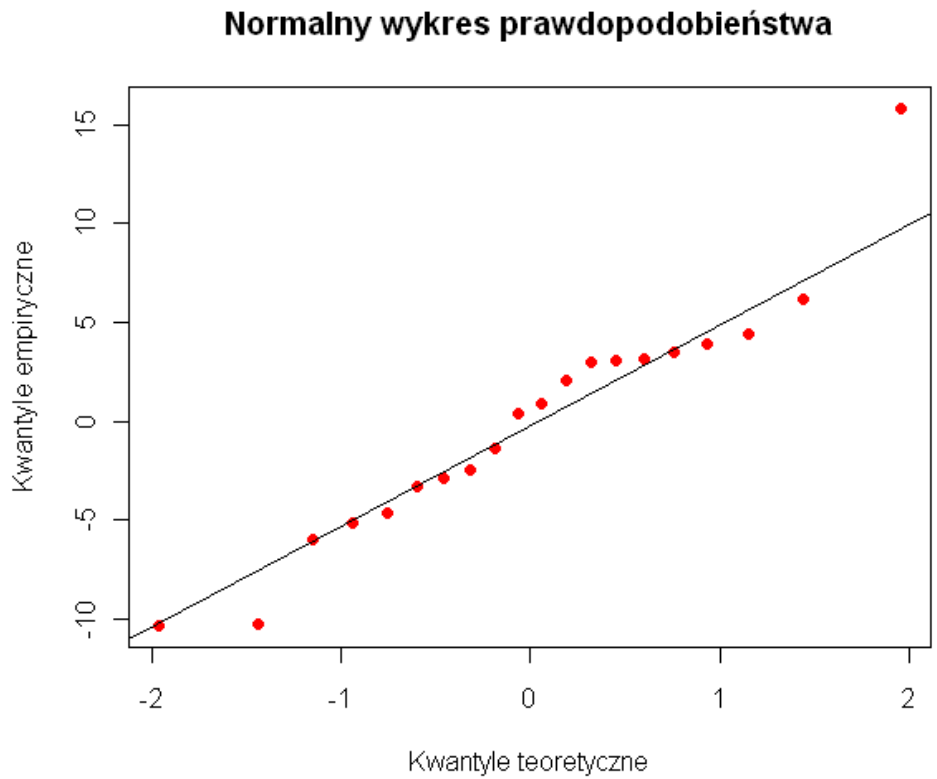
$$\text{Poziom} = -14.45 + 0.27\ \text{Status} + 0.97\ \text{Planowanie}$$



Weryfikacja modelu za pomocą analizy reszt

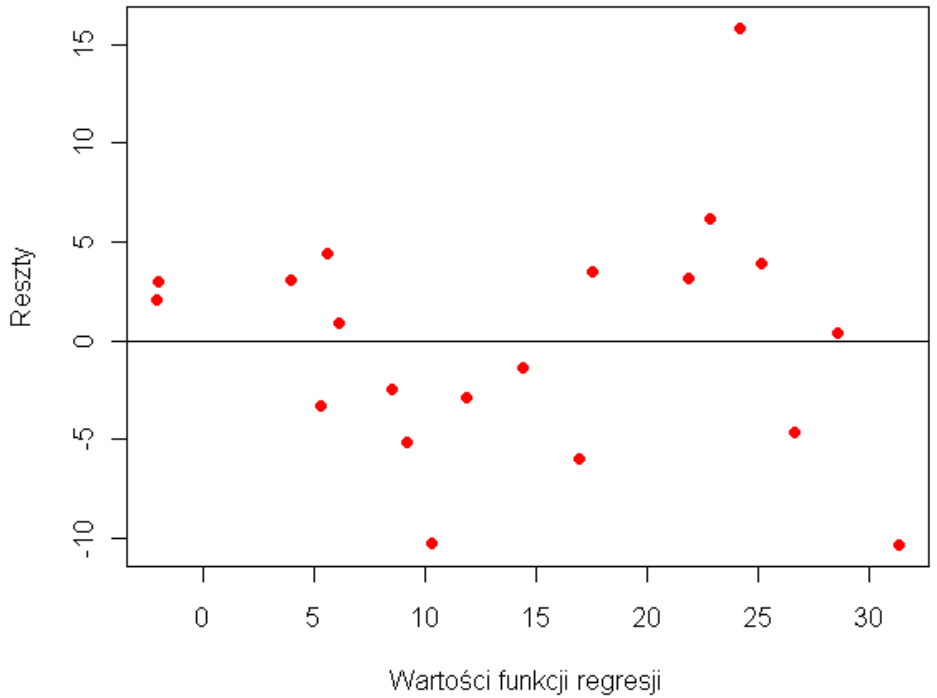
qqnorm(model$res,xlab="Kwantyle teoretyczne",ylab="Kwantyle empiryczne",main="Normalny wykres prawdopodobieństwa")

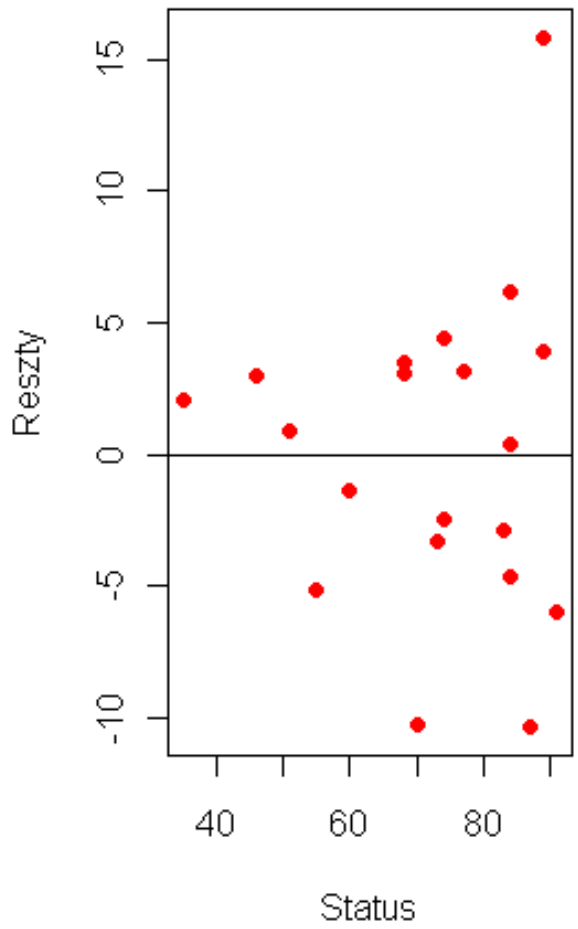
qqline(model$res)

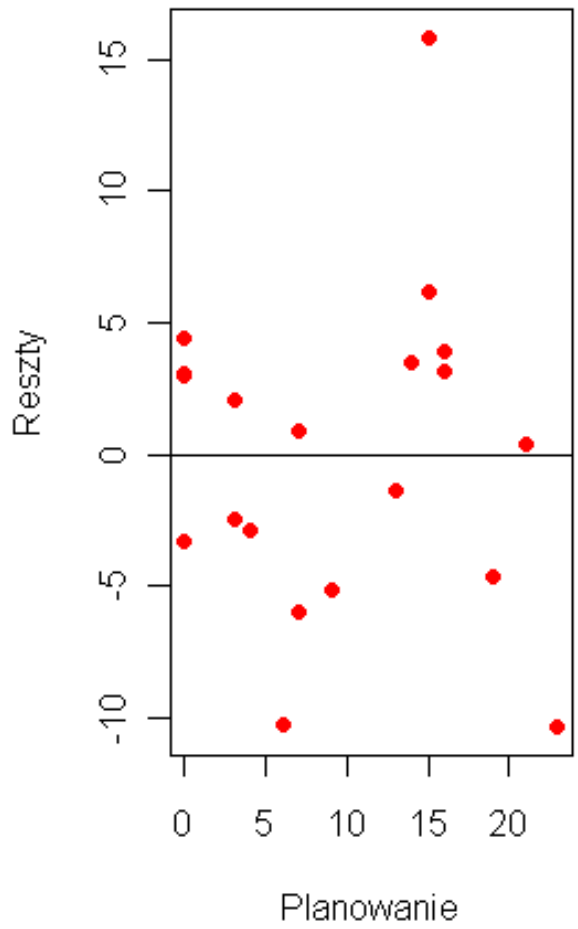


plot(model$fitted.values,model$res,xlab="Wartości funkcji regresji",ylab="Reszty",col="red",pch=19)

abline(h=0) qqline(model$res)




Reszty
15
10
5
0
-5
-10
40
60
80
Status


Reszty
15
10
5
0
-5
-10
0
5
10
15
20
Planowanie

- Testy analizy wariancji są podstawowym narzędziem statystyki eksperymentalnej.
- Służą do sprawdzenia, czy czynniki, których poziomy możemy z góry ustalić przed wykonaniem eksperymentu, wywierają istotny wpływ na kształtowanie średniej wartości badanej cechy mierzalnej.
- Czynniki te nazywamy *esperymentalnymi*. Badanym obietom (jednostkom eksperymentalnym) poziomy tych czynników przydzielane są losowo.
- Są czynniki, których poziomów nie możemy przydzielić losowo do obiektów, ponieważ stanowią ich integralną charakterystykę. Nazywamy je *czynnikami klasyfikującymi*.
- Poziomy każdego czynnika mogą być ustalone. Mówimy wtedy stałych czynnikach lub efektach. Poziomy czynnika mogą też być wylosowane z populacjii wszystkich możliwych poziomów. Wtedy są to czynniki losowe

- Testy analizy wariancji są podstawowym narzędziem statystyki eksperymentalnej.
- Służą do sprawdzenia, czy czynniki, których poziomy możemy z góry ustalić przed wykonaniem eksperymentu, wywierają istotny wpływ na kształtowanie średniej wartości badanej cechy mierzalnej.
- Czynniki te nazywamy *esperymentalnymi*. Badanym obietom (jednostkom eksperymentalnym) poziomy tych czynników przydzielane są losowo.
- Są czynniki, których poziomów nie możemy przydzielić losowo do obiektów, ponieważ stanowią ich integralną charakterystykę. Nazywamy je *czynnikami klasyfikującymi*.
- Poziomy każdego czynnika mogą być ustalone. Mówimy wtedy stałych czynnikach lub efektach. Poziomy czynnika mogą też być wylosowane z populacjii wszystkich możliwych poziomów. Wtedy są to czynniki losowe

- Testy analizy wariancji są podstawowym narzędziem statystyki eksperymentalnej.
- Służą do sprawdzenia, czy czynniki, których poziomy możemy z góry ustalić przed wykonaniem eksperymentu, wywierają istotny wpływ na kształtowanie średniej wartości badanej cechy mierzalnej.
- Czynniki te nazywamy *esperymentalnymi*. Badanym obietom (jednostkom eksperymentalnym) poziomy tych czynników przydzielane są losowo.
- Są czynniki, których poziomów nie możemy przydzielić losowo do obiektów, ponieważ stanowią ich integralną charakterystykę. Nazywamy je *czynnikami klasyfikującymi*.
- Poziomy każdego czynnika mogą być ustalone. Mówimy wtedy stałych czynnikach lub efektach. Poziomy czynnika mogą też być wylosowane z populacjii wszystkich możliwych poziomów. Wtedy są to czynniki losowe

- Testy analizy wariancji są podstawowym narzędziem statystyki eksperymentalnej.
- Służą do sprawdzenia, czy czynniki, których poziomy możemy z góry ustalić przed wykonaniem eksperymentu, wywierają istotny wpływ na kształtowanie średniej wartości badanej cechy mierzalnej.
- Czynniki te nazywamy *esperymentalnymi*. Badanym obietom (jednostkom eksperymentalnym) poziomy tych czynników przydzielane są losowo.
- Są czynniki, których poziomów nie możemy przydzielić losowo do obiektów, ponieważ stanowią ich integralną charakterystykę. Nazywamy je *czynnikami klasyfikującymi*.
- Poziomy każdego czynnika mogą być ustalone. Mówimy wtedy stałych czynnikach lub efektach. Poziomy czynnika mogą też być wylosowane z populacjii wszystkich możliwych poziomów. Wtedy są to czynniki losowe

- Testy analizy wariancji są podstawowym narzędziem statystyki eksperymentalnej.
- Służą do sprawdzenia, czy czynniki, których poziomy możemy z góry ustalić przed wykonaniem eksperymentu, wywierają istotny wpływ na kształtowanie średniej wartości badanej cechy mierzalnej.
- Czynniki te nazywamy *esperymentalnymi*. Badanym obietom (jednostkom eksperymentalnym) poziomy tych czynników przydzielane są losowo.
- Są czynniki, których poziomów nie możemy przydzielić losowo do obiektów, ponieważ stanowią ich integralną charakterystykę. Nazywamy je *czynnikami klasyfikującymi*.
- Poziomy każdego czynnika mogą być ustalone. Mówimy wtedy stałych czynnikach lub efektach. Poziomy czynnika mogą też być wylosowane z populacjii wszystkich możliwych poziomów. Wtedy są to czynniki losowe

Przykład

- Mamy trzy ośrodki szkolenia. W każdym ośrodku przeprowadzamy dwa rodzaje szkolenia. Rodzaj szkolenia przydzielany jest losowo do każdego uczestnika programu (eksperymentu). Mamy zatem dwa czynniki: *rodzaj szkolenia* oraz *ośrodek szkolenia*.
- Jeżeli ten sam rodzaj szkolenia jest lepszy w każdym ośrodku, wniosek nasuwa się sam, ponieważ rodzaj szkolenia przydzielony został losowo do każdego uczestnika.
- W przypadku wykrycia różnic między ośrodkami, wniosek nie jest już tak oczywisty. Nie wiemy, czy uczestnicy programu są lepiej szkoleni w danym ośrodku, ponieważ
  - jest on lepiej wyposażony,
  - posiada lepszych wykładowców,
  - uczestnicy są lepiej przygotowani, (czynnik edukacji możnaby wyeliminować poprzez rozlosowanie uczestników do ośrodków)
- *Rodzaj szkolenia* jest czynnikiem eksperymentalnym, a *ośrodek szkolenia* klasyfikującym.

Przykład

- Mamy trzy ośrodki szkolenia. W każdym ośrodku przeprowadzamy dwa rodzaje szkolenia. Rodzaj szkolenia przydzielany jest losowo do każdego uczestnika programu (eksperymentu). Mamy zatem dwa czynniki: *rodzaj szkolenia* oraz *ośrodek szkolenia*.
- Jeżeli ten sam rodzaj szkolenia jest lepszy w każdym ośrodku, wniosek nasuwa się sam, ponieważ rodzaj szkolenia przydzielony został losowo do każdego uczestnika.
- W przypadku wykrycia różnic między ośrodkami, wniosek nie jest już tak oczywisty. Nie wiemy, czy uczestnicy programu są lepiej szkoleni w danym ośrodku, ponieważ
  - jest on lepiej wyposażony,
  - posiada lepszych wykładowców,
  - uczestnicy są lepiej przygotowani, (czynnik edukacji możnaby wyeliminować poprzez rozlosowanie uczestników do ośrodków)
- *Rodzaj szkolenia* jest czynnikiem eksperymentalnym, a *ośrodek szkolenia* klasyfikującym.

Przykład

- Mamy trzy ośrodki szkolenia. W każdym ośrodku przeprowadzamy dwa rodzaje szkolenia. Rodzaj szkolenia przydzielany jest losowo do każdego uczestnika programu (eksperymentu). Mamy zatem dwa czynniki: *rodzaj szkolenia* oraz *ośrodek szkolenia*.
- Jeżeli ten sam rodzaj szkolenia jest lepszy w każdym ośrodku, wniosek nasuwa się sam, ponieważ rodzaj szkolenia przydzielony został losowo do każdego uczestnika.
- W przypadku wykrycia różnic między ośrodkami, wniosek nie jest już tak oczywisty. Nie wiemy, czy uczestnicy programu są lepiej szkoleni w danym ośrodku, ponieważ
  - jest on lepiej wyposażony,
  - posiada lepszych wykładowców,
  - uczestnicy są lepiej przygotowani, (czynnik edukacji możnaby wyeliminować poprzez rozlosowanie uczestników do ośrodków)
- *Rodzaj szkolenia* jest czynnikiem eksperymentalnym, a *ośrodek szkolenia* klasyfikującym.

Przykład

- Mamy trzy ośrodki szkolenia. W każdym ośrodku przeprowadzamy dwa rodzaje szkolenia. Rodzaj szkolenia przydzielany jest losowo do każdego uczestnika programu (eksperymentu). Mamy zatem dwa czynniki: *rodzaj szkolenia* oraz *ośrodek szkolenia*.
- Jeżeli ten sam rodzaj szkolenia jest lepszy w każdym ośrodku, wniosek nasuwa się sam, ponieważ rodzaj szkolenia przydzielony został losowo do każdego uczestnika.
- W przypadku wykrycia różnic między ośrodkami, wniosek nie jest już tak oczywisty. Nie wiemy, czy uczestnicy programu są lepiej szkoleni w danym ośrodku, ponieważ
  - jest on lepiej wyposażony,
  - posiada lepszych wykładowców,
  - uczestnicy są lepiej przygotowani, (czynnik edukacji możnaby wyeliminować poprzez rozlosowanie uczestników do ośrodków)
- *Rodzaj szkolenia* jest czynnikiem eksperymentalnym, a *ośrodek szkolenia* klasyfikującym.

Przykład

- Mamy trzy ośrodki szkolenia. W każdym ośrodku przeprowadzamy dwa rodzaje szkolenia. Rodzaj szkolenia przydzielany jest losowo do każdego uczestnika programu (eksperymentu). Mamy zatem dwa czynniki: *rodzaj szkolenia* oraz *ośrodek szkolenia*.
- Jeżeli ten sam rodzaj szkolenia jest lepszy w każdym ośrodku, wniosek nasuwa się sam, ponieważ rodzaj szkolenia przydzielony został losowo do każdego uczestnika.
- W przypadku wykrycia różnic między ośrodkami, wniosek nie jest już tak oczywisty. Nie wiemy, czy uczestnicy programu są lepiej szkoleni w danym ośrodku, ponieważ
  - jest on lepiej wyposażony,
  - posiada lepszych wykładowców,
  - uczestnicy są lepiej przygotowani, (czynnik edukacji możnaby wyeliminować poprzez rozlosowanie uczestników do ośrodków)
- *Rodzaj szkolenia* jest czynnikiem eksperymentalnym, a *ośrodek szkolenia* klasyfikującym.

Przykład

- Mamy trzy ośrodki szkolenia. W każdym ośrodku przeprowadzamy dwa rodzaje szkolenia. Rodzaj szkolenia przydzielany jest losowo do każdego uczestnika programu (eksperymentu). Mamy zatem dwa czynniki: *rodzaj szkolenia* oraz *ośrodek szkolenia*.
- Jeżeli ten sam rodzaj szkolenia jest lepszy w każdym ośrodku, wniosek nasuwa się sam, ponieważ rodzaj szkolenia przydzielony został losowo do każdego uczestnika.
- W przypadku wykrycia różnic między ośrodkami, wniosek nie jest już tak oczywisty. Nie wiemy, czy uczestnicy programu są lepiej szkoleni w danym ośrodku, ponieważ
  - jest on lepiej wyposażony,
  - posiada lepszych wykładowców,
  - uczestnicy są lepiej przygotowani, (czynnik edukacji możnaby wyeliminować poprzez rozlosowanie uczestników do ośrodków)
- *Rodzaj szkolenia* jest czynnikiem eksperymentalnym, a *ośrodek szkolenia* klasyfikującym.

Przykład

- Mamy trzy ośrodki szkolenia. W każdym ośrodku przeprowadzamy dwa rodzaje szkolenia. Rodzaj szkolenia przydzielany jest losowo do każdego uczestnika programu (eksperymentu). Mamy zatem dwa czynniki: *rodzaj szkolenia* oraz *ośrodek szkolenia*.
- Jeżeli ten sam rodzaj szkolenia jest lepszy w każdym ośrodku, wniosek nasuwa się sam, ponieważ rodzaj szkolenia przydzielony został losowo do każdego uczestnika.
- W przypadku wykrycia różnic między ośrodkami, wniosek nie jest już tak oczywisty. Nie wiemy, czy uczestnicy programu są lepiej szkoleni w danym ośrodku, ponieważ
  - jest on lepiej wyposażony,
  - posiada lepszych wykładowców,
  - uczestnicy są lepiej przygotowani, (czynnik edukacji możnaby wyeliminować poprzez rozlosowanie uczestników do ośrodków)
- *Rodzaj szkolenia* jest czynnikiem eksperymentalnym, a *ośrodek szkolenia* klasyfikującym.

Jednoczynnikowa analiza wariancji

- **Założenia**

  1. Na każdym poziomie czynnika rozkład prawdopodobieństwa badanej cechy mierzalnej
     a) jest normalny
     b) ma tę samą wariancję
  2. Obserwacje są niezależne

- **Cel**

  1. Zweryfikować, czy średnia wartość badanej cechy zależy od czynnika
  2. Jeżeli średnia zależy od czynnika, zbadać w jaki sposób i jakie są tego konsekwencje

Jednoczynnikowa analiza wariancji

- **Założenia**

  1. Na każdym poziomie czynnika rozkład prawdopodobieństwa badanej cechy mierzalnej
     a) jest normalny
     b) ma tę samą wariancję
  2. Obserwacje są niezależne

- **Cel**

  1. Zweryfikować, czy średnia wartość badanej cechy zależy od czynnika
  2. Jeżeli średnia zależy od czynnika, zbadać w jaki sposób i jakie są tego konsekwencje

Jednoczynnikowa analiza wariancji

- Model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

gdzie

- i jest indeksem określający poziom czynnika ($i = 1, \ldots, k$; $k-$liczba poziomów czynnika)
- j jest indeksem określającym numer powtórzenia obserwacji przy ustalonym poziomie czynnika ($j = 1, \ldots, n_i$; n_i-liczba powtórzeń przy $i-$tym poziomie czynnika)
- Y_{ij} jest zmienną losową; przy i-tym poziomie czynnika oznacza j-tą wartość obserwowanej cechy
- μ_i jest wartością oczekiwaną zmiennej Y_{ij}; oznacza średnią wartość obserwowanej cechy przy i-tym poziomie czynnika
- ε_{ij} jest zmienną losową o rozkładzie $N(0, \sigma^2)$. O zmiennych losowych ε_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, zakładamy, że są niezależne

Jednoczynnikowa analiza wariancji

- Model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

gdzie

- i jest indeksem określający poziom czynnika ($i = 1, \ldots, k$; $k-$liczba poziomów czynnika)
- j jest indeksem określającym numer powtórzenia obserwacji przy ustalonym poziomie czynnika ($j = 1, \ldots, n_i$; n_i-liczba powtórzeń przy $i-$tym poziomie czynnika)
- Y_{ij} jest zmienną losową; przy i-tym poziomie czynnika oznacza j-tą wartość obserwowanej cechy
- μ_i jest wartością oczekiwaną zmiennej Y_{ij}; oznacza średnią wartość obserwowanej cechy przy i-tym poziomie czynnika
- ε_{ij} jest zmienną losową o rozkładzie $N(0, \sigma^2)$. O zmiennych losowych ε_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, zakładamy, że są niezależne

Jednoczynnikowa analiza wariancji

- Model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

gdzie

- i jest indeksem określający poziom czynnika ($i = 1, \ldots, k$; $k-$liczba poziomów czynnika)
- j jest indeksem określającym numer powtórzenia obserwacji przy ustalonym poziomie czynnika ($j = 1, \ldots, n_i$; n_i-liczba powtórzeń przy $i-$tym poziomie czynnika)
- Y_{ij} jest zmienną losową; przy i-tym poziomie czynnika oznacza j-tą wartość obserwowanej cechy
- μ_i jest wartością oczekiwaną zmiennej Y_{ij}; oznacza średnią wartość obserwowanej cechy przy i-tym poziomie czynnika
- ε_{ij} jest zmienną losową o rozkładzie $N(0, \sigma^2)$. O zmiennych losowych ε_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, zakładamy, że są niezależne

Jednoczynnikowa analiza wariancji

- Model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

gdzie

- i jest indeksem określający poziom czynnika ($i = 1, \ldots, k$; $k-$liczba poziomów czynnika)
- j jest indeksem określającym numer powtórzenia obserwacji przy ustalonym poziomie czynnika ($j = 1, \ldots, n_i$; n_i-liczba powtórzeń przy $i-$tym poziomie czynnika)
- Y_{ij} jest zmienną losową; przy i-tym poziomie czynnika oznacza j-tą wartość obserwowanej cechy
- μ_i jest wartością oczekiwaną zmiennej Y_{ij}; oznacza średnią wartość obserwowanej cechy przy i-tym poziomie czynnika
- ε_{ij} jest zmienną losową o rozkładzie $N(0, \sigma^2)$. O zmiennych losowych ε_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, zakładamy, że są niezależne

Jednoczynnikowa analiza wariancji

- Model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

gdzie

- i jest indeksem określający poziom czynnika ($i = 1, \ldots, k$; $k-$liczba poziomów czynnika)
- j jest indeksem określającym numer powtórzenia obserwacji przy ustalonym poziomie czynnika ($j = 1, \ldots, n_i$; n_i-liczba powtórzeń przy $i-$tym poziomie czynnika)
- Y_{ij} jest zmienną losową; przy i-tym poziomie czynnika oznacza j-tą wartość obserwowanej cechy
- μ_i jest wartością oczekiwaną zmiennej Y_{ij}; oznacza średnią wartość obserwowanej cechy przy i-tym poziomie czynnika
- ε_{ij} jest zmienną losową o rozkładzie $N(0, \sigma^2)$. O zmiennych losowych ε_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, zakładamy, że są niezależne

Jednoczynnikowa analiza wariancji

- Model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

gdzie

- i jest indeksem określający poziom czynnika ($i = 1, \ldots, k$; $k-$liczba poziomów czynnika)
- j jest indeksem określającym numer powtórzenia obserwacji przy ustalonym poziomie czynnika ($j = 1, \ldots, n_i$; n_i-liczba powtórzeń przy $i-$tym poziomie czynnika)
- Y_{ij} jest zmienną losową; przy i-tym poziomie czynnika oznacza j-tą wartość obserwowanej cechy
- μ_i jest wartością oczekiwaną zmiennej Y_{ij}; oznacza średnią wartość obserwowanej cechy przy i-tym poziomie czynnika
- ε_{ij} jest zmienną losową o rozkładzie $N(0, \sigma^2)$. O zmiennych losowych ε_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, zakładamy, że są niezależne

Zatem

$Y_{ij} \sim N(\mu_i, \sigma^2),\ i = 1, \ldots, k,\ j = 1, \ldots, n_i$, są niezależnymi zmiennymi losowymi

Porównanie wartości średnich

$$H_0 :\ \mu_1 = \ldots = \mu_k$$

Test F (poziom istotności α)
Statystyka testowa

$$F_{\text{emp}} = \frac{S_a^2}{S_e^2}$$

Jeżeli $F_{\text{emp}} > F(\alpha; k-1, N-k)$,
to hipotezę $H_0 : \mu_1 = \cdots = \mu_k$ odrzucamy.

Wniosek praktyczny:
przynajmniej jedna ze średnich $\mu_1, \ldots, \mu_k$ jest inna od pozostałych

Podział całkowitej sumy kwadratów

$$\underbrace{Y_{ij} - \bar{Y}_{..}}_{\text{zmienność całkowita}} = \underbrace{\bar{Y}_{i.} - \bar{Y}_{..}}_{\text{zmienność średnich}} + \underbrace{Y_{ij} - \bar{Y}_{i.}}_{\text{zmienność wokół średniej}}$$

$$\underbrace{\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2}_{SST} = \underbrace{\sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2}_{SSTR} + \underbrace{\sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2}_{SSE}$$

SST – całkowita suma kwadratów

$SSTR$ – suma kwadratów dla czynnika

SSE – suma kwadratów reszt

$S_a^2 = SSTR/(k-1), \quad S_e^2 = SSE/(N-k), \quad N = \sum_i n_i$

Grupy jednorodne — podzbiory średnich, które można uznać za takie same

Procedury porównań wielokrotnych — postępowanie statystyczne zmierzające do podzielenia zbioru średnich na grupy jednorodne

Procedury: Tukeya, Scheffégo, Bonfferroniego, Duncana, Newmana–Kuelsa i inne. Ogólna idea procedur porównań wielokrotnych

$(n_1 = \cdots = n_k)$

NIR — najmniejsza istotna różnica

Jeżeli $|\bar{Y}_i - \bar{Y}_j| < NIR$, to uznajemy, że $\mu_i = \mu_j$. Jeżeli

$|\bar{Y}_i - \bar{Y}_j| < NIR$
$|\bar{Y}_i - \bar{Y}_l| < NIR$
$|\bar{Y}_l - \bar{Y}_j| < NIR$,
to uznajemy, że $\mu_i = \mu_j = \mu_l$.

Badając w ten sposób wszystkie pary średnich próbkowych otrzymujemy podział zbioru średnich na grupy jednorodne.

Procedura Tukeya

Założenie: $n_1 = \cdots = n_k = n$

$$NIR = t(\alpha; k, N-k) S_e \sqrt{\frac{1}{n}}$$

$t(\alpha; k, N-k)$ — wartość krytyczna studentyzowanego rozstępu

Przypadek nierównolicznych prób

Jedna z modyfikacji procedury Tukeya

$$NIR_{ij} = t(\alpha; k, N-k) S_e \sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Regresja logistyczna

- Rozważmy dwuwartościową zmienną: 1 – choroba wystąpiła, 0 – choroba nie wystąpiła. Oznaczmy ją przez D (na przykład: choroba wieńcowa).
- Przedmiotem zainteresowania jest dwuwartościowa zmienna, któr określa grupę ryzyka: 1 – jest w grupie ryzyka, 0 – nie ma w grupie ryzyka. Oznaczmy ją przez E (na przykład: palenie papierosów).
- Zauważmy, że aby zbadać związek grupy ryzyka (palaczy) z chorobą (chorobą wieńcową), często musimy brać pod uwagę dodatkowe charakterystyki takie, jak wiek C_1, rasa C_2 czy płeć C_3. Te charakterystyki nie interesują nas w sposób bezpośredni. Są to tak zwane zmienne kontrolne.
- Zmienną E (ekspozycja) razem ze zmiennymi kontrolnymi $C_1,\ C_2,\ C_3$ nazywamy zmiennymi NIEZALEŻNYMI, a zmienną D nazywamy zmienną ZALEŻNĄ.

$$\underbrace{(E,\ C_1,\ C_2,\ C_3)}_{\text{niezależne}} \longrightarrow D$$

Regresja logistyczna

- Rozważmy dwuwartościową zmienną: 1 – choroba wystąpiła, 0 – choroba nie wystąpiła. Oznaczmy ją przez D (na przykład: choroba wieńcowa).
- Przedmiotem zainteresowania jest dwuwartościowa zmienna, któr określa grupę ryzyka: 1 – jest w grupie ryzyka, 0 – nie ma w grupie ryzyka. Oznaczmy ją przez E (na przykład: palenie papierosów).
- Zauważmy, że aby zbadać związek grupy ryzyka (palaczy) z chorobą (chorobą wieńcową), często musimy brać pod uwagę dodatkowe charakterystyki takie, jak wiek C_1, rasa C_2 czy płeć C_3. Te charakterystyki nie interesują nas w sposób bezpośredni. Są to tak zwane zmienne kontrolne.
- Zmienną E (ekspozycja) razem ze zmiennymi kontrolnymi $C_1,\ C_2,\ C_3$ nazywamy zmiennymi NIEZALEŻNYMI, a zmienną D nazywamy zmienną ZALEŻNĄ.

$$\underbrace{(E,\ C_1,\ C_2,\ C_3)}_{\text{niezależne}} \longrightarrow D$$

Regresja logistyczna

- Rozważmy dwuwartościową zmienną: 1 – choroba wystąpiła, 0 – choroba nie wystąpiła. Oznaczmy ją przez D (na przykład: choroba wieńcowa).
- Przedmiotem zainteresowania jest dwuwartościowa zmienna, któr określa grupę ryzyka: 1 – jest w grupie ryzyka, 0 – nie ma w grupie ryzyka. Oznaczmy ją przez E (na przykład: palenie papierosów).
- Zauważmy, że aby zbadać związek grupy ryzyka (palaczy) z chorobą (chorobą wieńcową), często musimy brać pod uwagę dodatkowe charakterystyki takie, jak wiek C_1, rasa C_2 czy płeć C_3. Te charakterystyki nie interesują nas w sposób bezpośredni. Są to tak zwane zmienne kontrolne.
- Zmienną E (ekspozycja) razem ze zmiennymi kontrolnymi $C_1,\ C_2,\ C_3$ nazywamy zmiennymi NIEZALEŻNYMI, a zmienną D nazywamy zmienną ZALEŻNĄ.

$$\underbrace{(E,\ C_1,\ C_2,\ C_3)}_{\text{niezależne}} \longrightarrow D$$

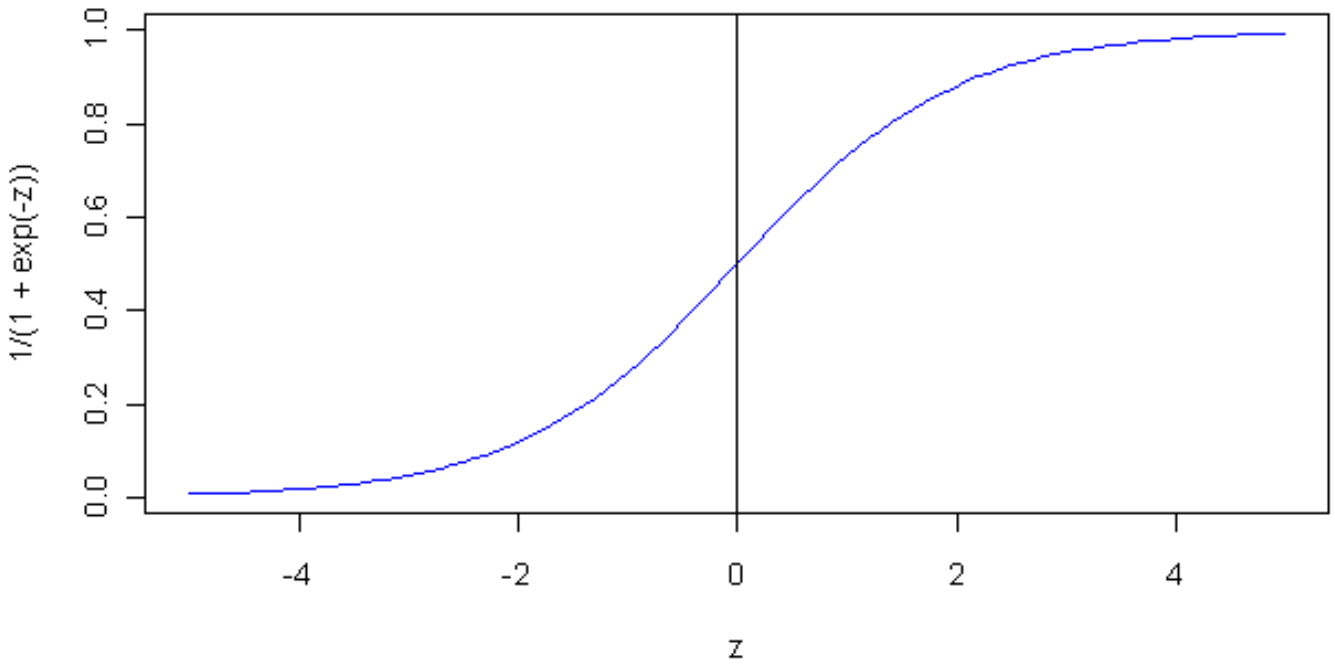
Regresja logistyczna

- Rozważmy dwuwartościową zmienną: 1 – choroba wystąpiła, 0 – choroba nie wystąpiła. Oznaczmy ją przez D (na przykład: choroba wieńcowa).
- Przedmiotem zainteresowania jest dwuwartościowa zmienna, któr określa grupę ryzyka: 1 – jest w grupie ryzyka, 0 – nie ma w grupie ryzyka. Oznaczmy ją przez E (na przykład: palenie papierosów).
- Zauważmy, że aby zbadać związek grupy ryzyka (palaczy) z chorobą (chorobą wieńcową), często musimy brać pod uwagę dodatkowe charakterystyki takie, jak wiek C_1, rasa C_2 czy płeć C_3. Te charakterystyki nie interesują nas w sposób bezpośredni. Są to tak zwane zmienne kontrolne.
- Zmienną E (ekspozycja) razem ze zmiennymi kontrolnymi $C_1,\ C_2,\ C_3$ nazywamy zmiennymi NIEZALEŻNYMI, a zmienną D nazywamy zmienną ZALEŻNĄ.

$$\underbrace{(E,\ C_1,\ C_2,\ C_3)}_{\text{niezależne}} \longrightarrow D$$

Funkcja logistyczna $f(z) = \frac{1}{1+e^{-z}}$

Ryzyko wystąpienia choroby chcemy wyrazić za pomocą prawdopodobieństwa. Prawdopodobieństwo jest wartością z przedziału (0,1). W naszym przypadku chcemy uzależnić je od zmiennych niezależnych. Argument z reprezentuje mi zmienne niezależne, a $f(z)$ ryzyko wystąpienia choroby.



Model

Podstawiamy

$$z = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 E$$

Otrzymujemy

$$P(D = 1 | C_1,\ C_2,\ C_3,\ E) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 E)}}$$

Przykład 1

D = wystąpienie choroby wieńcowej (0 lub 1)
E = poziom katecholamin (0 - niski, 1 - wysoki)
C_1 = wiek (ciągła)
C_2 = wynik elektrokardiogramu (0 - w normie, 1 - poza normą)
$n = 609$ – liczba badanych kobiet
Czas obserwacji = 9 lat

$$(E,\ C_1,\ C_2) \longrightarrow D$$

$$\text{Czas: } T_0 \longrightarrow T_1$$

Wynik estymacji ($z = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_3 E$):

$$\hat{\beta}_0 = -3.911,\ \hat{\beta}_1 = 0.029,\ \hat{\beta}_2 = 0.342,\ \hat{\beta}_3 = 0.652$$

Przykład, cd

Chcemy wykorzystać wyniki estymacji w celu oszacowania ryzyka wystąpienia choroby wieńcowej u osób:
a) o wysokim poziomie katecholamin w wieku 40 lat, których elektrokardiogram nie wykazał odstępstw od normy

$$E = 1,\ C_1 = 40,\ C_2 = 0$$

b) o niskim poziomie katecholamin w wieku 40 lat, których elektrokardiogram nie wykazał odstępstw od normy

$$E = 0,\ C_1 = 40,\ C_2 = 0$$

Przykład, cd

ad a)

$$\hat{P}(D = 1|\ C_1 = 40, C_2 = 0,\ E = 1) = \frac{1}{1 + e^{-(-3.911+0.029\cdot(40)+0.342\cdot(0)+0.652\cdot(1))}} = 0.109$$

ad b)

$$\hat{P}(D = 1|\ C_1 = 40, C_2 = 0,\ E = 0) = \frac{1}{1 + e^{-(-3.911+0.029\cdot(40)+0.342\cdot(0)+0.652\cdot(0))}} = 0.060$$

$$RR = \frac{\hat{P}(D = 1|\ C_1 = 40, C_2 = 0,\ E = 1)}{\hat{P}(D = 1|\ C_1 = 40, C_2 = 0, E = 0)} = 1.82$$

Przykład 2

W badaniach retrospektywnych nad przypuszczalnym znaczeniem grupy krwi w chorobie wrzodowej układu pokarmowego zebrano dane:

	Choroba wrzodowa		Grupa kontrolna	
	Grupa O	Grupa A	Grupa O	Grupa A
Londyn	911	579	45878	4219
Manchester	361	246	4532	3775
Newcastle	396	219	6598	5261

Celem badań jest sprawdzenie, czy ryzyko choroby wrzodowej zależy od grupy krwi. Zatem za zmienną zależną powinniśmy przyjąć występowanie choroby wrzodowej, a za niezależną grupę krwi. Oznacza to, że zbierając dane, powinniśmy wylosować osoby z każdą grupą krwi, a następnie po ustalonym czasie zliczyć osoby z chorobą i bez choroby wrzodowej:

$$\left(\text{Miasto}, \boxed{\text{Grupa krwi}}\right) \longrightarrow \textit{Choroba}$$

$$\text{Czas: } T_0 \longrightarrow T_1$$

Przykład, cd.

- Poręczniej jest zebrać dane odwrotnie:

$$\textit{Choroba} \longrightarrow \left(\text{Miasto}, \boxed{\text{Grupa krwi}}\right)$$

najpierw wybieramy osoby spośród chorych oraz zdrowych, a następnie zliczamy osoby z grupą krwi O oraz A. Jeżeli nie kontrolujemy, miejsca zamieszkania osoby, zmienna Miasto ma charakter losowy i mamy powyższy schemat.

- Przy wyborze osób z konkretnych miast mamy schemat:

$$\text{Miasto, Choroba} \longrightarrow \boxed{\text{Grupa krwi}} \quad (!)$$

- Nie możemy oszacować w sposób bezpośredni ryzyka choroby wrzodowej, ponieważ liczba chorych w stosunku do zdrowych jest z góry ustalona i nie odzwierciedla rozkładu liczby chorych osób w populacji osób chorych i zdrowych. Zatem nie możemy bezpośrednio oszacowć prawdopodobieństwa

$$P(\text{Choroba}|\text{Grupa krwi, Miasto})$$

.

Przykład, cd.

Możemy za to oszacować

$$P(\text{Grupa krwi}|\text{Choroba, Miasto}),$$

Przyjmijmy

Nazwa cechy	Przyjmowane wartości
Grupa krwi	A, O
Choroba	Tak, Nie
Miasto	Londyn, Manchester, Newcastle

Przykłady zapisu:

$$P(\text{A} \mid \text{Tak, Londyn}),\ P(\text{O} \mid \text{Tak, Londyn}),\ P(\text{A} \mid \text{Nie, Newcastle})$$

Przyjmujemy model

$$P(\text{A}|\text{Choroba, Miasto}) = \frac{1}{1+e^{-z}}$$

$$\begin{aligned} z = \beta_0 + \beta_1 L(\mathit{Choroba} = \mathit{Tak}) \quad &+ \quad \beta_2 L(\mathit{Miasto} = \mathit{Manchester}) + \\ &+ \quad \beta_3 L(\mathit{Miasto} = \mathit{Newcastle}) \end{aligned}$$

$$\boxed{L(\text{PRAWDA}) = 1 = 1 - L(\text{FAŁSZ})}$$

Przykłady

$$P(\text{A}|\text{Tak, Londyn}) = \frac{1}{1+e^{-(\beta_0+\beta_1)}}$$

$$P(\text{A}|\text{Nie, Manchester}) = \frac{1}{1+e^{-(\beta_0+\beta_2)}}$$

$$P(\text{O}|\text{Tak, Londyn}) = 1 - P(\text{A}|\text{Tak, Londyn}) = \frac{1}{1+e^{(\beta_0+\beta_1)}}$$

Obliczenia w R:

```
> dane
      Miasto Grupa.O Grupa.A Choroba
1     Londyn     911     576     _Tak
2 Manchester     361     246     _Tak
3  Newcastle     396     219     _Tak
4     Londyn    4578    4219     _Nie
5 Manchester    4532    3775     _Nie
6  Newcastle    6598    5261     _Nie
> attach(dane)
> model=glm(cbind(Grupa.A,Grupa.O)~Choroba+Miasto,family=binomial)
> summary(model)
```

Obliczenia w R:
$\hat{\beta}_0 = -0.08776,\ \hat{\beta}_1 = -0.33292,\ \hat{\beta}_2 = -0.08635,\ \hat{\beta}_3 = -0.14021$

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.08776    0.02064  -4.253 2.11e-05 ***
Choroba_Tak       -0.33292    0.04170  -7.983 1.42e-15 ***
MiastoManchester  -0.08635    0.02925  -2.952  0.00316 **
MiastoNewcastle   -0.14021    0.02707  -5.179 2.23e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 84.5078  on 5  degrees of freedom
Residual deviance:  3.0992  on 2  degrees of freedom
AIC: 61.236
```

Znaczenie oszacowanych parametrów.

Można pokazać, że (dla każdego Miasta)

$$\ln\left(\frac{P(A|Tak, Miasto)}{P(O|Tak, Miasto)} \Big/ \frac{P(A|Nie, Miasto)}{P(O|Nie, Miasto)}\right) = \beta_1$$

Zatem mamy następujący iloraz szans:

$$\text{OR} = \frac{P(A|Tak, Miasto)}{P(O|Tak, Miasto)} \Big/ \frac{P(A|Nie, Miasto)}{P(O|Nie, Miasto)} = e^{\beta_1}$$

Można pokazać:

$$\begin{aligned} \text{OR} &= \frac{P(A|Tak, Miasto)}{P(O|Tak, Miasto)} \Big/ \frac{P(A|Nie, Miasto)}{P(O|Nie, Miasto)} = \\ &= \frac{P(Tak|A, Miasto)}{P(Tak|O, Miasto)} \Big/ \frac{P(Nie|A, Miasto)}{P(Nie|O, Miasto)} = \\ &= \frac{P(Tak|A, Miasto)}{P(Nie|A, Miasto)} \Big/ \frac{P(Tak|O, Miasto)}{P(Nie|O, Miasto)} \end{aligned}$$

Oznacza to, że iloraz szans w badaniu retrospektywnym jest identyczny z ilorazem szans w badaniu prospektywnym.
Przyjmując

$$RR = \frac{P(Tak|A, Miasto)}{P(Tak|O, Miasto)} \text{ oraz } \frac{P(Nie|A, Miasto)}{P(Nie|O, Miasto)} \approx 1$$

otrzymujemy

$$RR \approx OR$$

Otrzymane przybliżenia jest realne, jeżeli procentowo jest bardzo mało ludzi chorych na wrzody układu pokarmowego.

Obliczenia:

$$e^{\hat{\beta}_1} = e^{-0.33292} \approx 0.72$$

Wnioski:
1. Ryzyko choroby wrzodowej jest większe u osób z grupą krwi O.
2. Ryzyko choroby wrzodowej u osoby z grupą A jest o ok. 30% mniejsze niż u osoby z grupą O.