

Programowanie w języku R. Dane dwuwymiarowe

Aleksander Denisiuk
Uniwersytet Warmińsko-Mazurski
Olsztyn, ul. Słoneczna 54
denisjuk@matman.uwm.edu.pl

18 marca 2025

Dane dwuwymiarowe

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

Najnowsza wersja tego dokumentu dostępna jest pod adresem

<http://wmii.uwm.edu.pl/~denisjuk/uwm>

Testowanie dwóch prób

- ❖ Zagadnienie
- ❖ Test Studenta
- ❖ Test Wilkoxona
- ❖ Test znaków

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

Testowanie dwóch prób

Zagadnienie

Testowanie dwóch prób

❖ Zagadnienie

❖ Test Studenta

❖ Test Wilkoxy

❖ Test znaków

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Dane są dwie próby
- Dwie hipotezy:
 - ◆ H_0 — próby pochodzą z tej samej populacji generalnej
 - brak różnicy
 - ◆ H_1 — próby pochodzą z różnych populacji
 - różnica istotna
- Na podstawie średnich
- Zakładamy, że średnie odchylenia są równe
 - ◆ te dwie próby się nie różnią:
 - $C(1, 2, 3, 4, 5, 6, 7, 8, 9)$
 - $C(5, 5, 5, 5, 5, 5, 5, 5, 5)$

Test Studenta

Testowanie dwóch prób

❖ Zagadnienie

❖ Test Studenta

❖ Test Wilkoxy

❖ Test znaków

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Test parametryczny
- Różne obiekty: test dla prób niezależnych
 - ◆ `t.test(x1, x2)`
- Dane pochodzą z tych samych obiektów: test dla prób zależnych
 - ◆ `t.test(x1, x2, paired=TRUE)`

Przykład

Testowanie dwóch prób

❖ Zagadnienie

❖ Test Studenta

❖ Test Wilcoxona

❖ Test znaków

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- William Sealy Gosset, dwa środki nasenne, 1908

- Dane: sleep

- Wykres:

```
plot(extra ~ group, data = sleep)
```

- test:

```
with(sleep, t.test(extra[group == 1],  
                    extra[group == 2], var.equal = FALSE))
```

- ◆ var.equal = **TRUE** założenie równości wariancji
- ◆ var.equal = **FALSE** modyfikacja Welcha

- Czy zaakceptować hipotezę zerową?
- Czy któryś z leków jest skuteczniejszy?

Nieparametryczny test Wilcoxona

Testowanie dwóch prób

❖ Zagadnienie

❖ Test Studenta

❖ Test Wilcoxona

❖ Test znaków

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Dane `airquality` o jakości powietrza w Nowym Yorku od maja do września 1973
- Czy koncentracja ozonu w maju i sierpniu była taka sama?

✦ czemu test nieparametryczny?

```
wilcox.test(Ozone ~ Month, data = airquality,  
            subset = Month %in% c(5, 8))
```

- Wykres:

```
boxplot(Ozone ~ Month, data = airquality,  
        subset = Month %in% c(5, 8))
```

Test znaków

Testowanie dwóch prób

- ❖ Zagadnienie
- ❖ Test Studenta
- ❖ Test Wilkoxy

❖ Test znaków

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Dane są dwie próby
 - ◆ Obliczyć różnice między parami odpowiednich elementów
 - ◆ Jeżeli próby pochodzą z tej samej populacji generalnej, ilość dodatnich różnic jest około 50%
 - można zweryfikować za pomocą testu binominalnego

Testowanie dwóch prób

Analiza tabel

❖ Tabele krzyżowe

❖ χ^2

❖ Przykład

❖ Kappa Cohena

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

Analiza tabel

Tabele krzyżowe

Testowanie dwóch prób

Analiza tabel

❖ Tabele krzyżowe

❖ χ^2

❖ Przykład

❖ Kappa Cohena

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Porównanie dwóch prób nominalnych

- ❖ tabele krzyżowe (rozdzielcze, kontyngencji, dwudzielcze)

```
with(airquality,  
      table(cut(Temp, quantile(Temp)), Month))
```

- Więcej niż dwie próby

- ❖ tabele wielowymiarowe

```
Titanic
```

- ❖ tabele „spłaszczone”

```
ftable(Titanic, row.vars = 1:3)  
ftable(Titanic, col.vars = 1:3)
```

- Obliczenie częstości funkcją `table()`

```
table(factor(rep(c("A", "B"), 10),  
           levels=c("A", "B", "C")))
```

Wykresy tabel dwudzielczych

Testowanie dwóch prób

Analiza tabel

❖ Tabele krzyżowe

❖ χ^2

❖ Przykład

❖ Kappa Cohena

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

```
titanic <- apply(Titanic, c(1, 4), sum)
mosaicplot(titanic, col = c("red", "green"),
            cex.axis=1)
```

Test zgodności χ^2

Testowanie dwóch prób

Analiza tabel

❖ Tabele krzyżowe

❖ χ^2

❖ Przykład

❖ Kappa Cohena

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Niezależność koloru włosów i oczu
 - ✦ dane `HairEyeColor`
- Sumujemy ilości dla płci (trzeci indeks w `HairEyeColor`)

```
x <- margin.table(HairEyeColor, c(1, 2))
```
- Test χ^2
 - ✦ hipoteza zerowa: cechy są niezależne

```
x <- margin.table(HairEyeColor, c(1, 2))  
chisq.test(x)
```
- Na wykresie:

```
assocplot(x)
```

Case study

Testowanie dwóch prób

Analiza tabel

❖ Tabele krzyżowe

❖ χ^2

❖ Przykład

❖ Kappa Cohena

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- 9 marca 2009 roku, z okazji Dnia Statystyki Polskiej odbył się bankiet. Następnego wielu uczestników bankietu miało objawy zatrucia pokarmowego. Aby ustalić przyczynę, zebrano informację, kto co jadł.

◆ dane w pliku `data/zatrucie.txt`

```
tox <- read.table("data/zatrucie.txt", h=TRUE)
head(tox)
sapply(tox[, -1], function(x) {
  tmp <- chisq.test(tox$ILL, x)
  print(tmp$p.value)
})
assocplot(table(ILL=tox$ILL,
                CAESAR=tox$CAESAR))
```

- Co jest przyczyną?

Kappa Cohena

Testowanie dwóch prób

Analiza tabel

❖ Tabele krzyżowe

❖ χ^2

❖ Przykład

❖ Kappa Cohena

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Stopień zgodności dwukrotnych pomiarów tej samej zmiennej w różnych warunkach (zgodność ocen ekspertów)
- Skala nominalna
 - ✦ współczynnik od -1 do 1
 - $\kappa \approx 1$ — oceny są zgodne
 - $\kappa \approx 0$ — zgodność na poziomie *ocen losowych*
 - $\kappa < 0$ — jeszcze gorsza zgodność
 - ✦ wartość p hipotezy zerowej ($\kappa = 0$)

```
install.packages("irr")  
library(irr)  
kappa2
```

Przykład. Rzucanie monetą

Testowanie dwóch prób

Analiza tabel

❖ Tabele krzyżowe

❖ χ^2

❖ Przykład

❖ Kappa Cohena

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

```
flips1 <- rbinom(n = 40, size = 1, prob = 0.5)
flips2 <- rbinom(n = 40, size = 1, prob = 0.5)
coins<-cbind(flips1, flips2)
```

```
agree(coins)
kappa2(coins)
```

Przykład. Badanie roślin na wyspie Kij

Testowanie dwóch prób

Analiza tabel

❖ Tabele krzyżowe

❖ χ^2

❖ Przykład

❖ Kappa Cohena

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Dwie niezależne grupy badawcze robili spis gatunków roślin na wyspie Kij na morzu Białym

```
pok <- read.table("data/pokorm_03.dat",  
                  h=TRUE, sep=";")  
agree(pok)  
kappa2(pok)
```


Przykład. Ocena wiarygodności skali

Testowanie dwóch prób

Analiza tabel

❖ Tabele krzyżowe

❖ χ^2

❖ Przykład

❖ Kappa Cohena

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Przykład z bloga Sary Locatelli
- Na potrzeby badań zostały opracowane kryteria ocen materiałów
- Dane testowe zostały oszacowane przez dwóch ekspertów

```
kappa2 (numstudies[, 2:3])
```

```
## Cohen's Kappa for 2 Raters (Weights: unweighted)
##
## Subjects = 62
## Raters = 2
## Kappa = 0.521
##
## z = 6.22
## p-value = 5.12e-10
```

Ocena wiarygodności skali (c.d.)

Testowanie dwóch prób

Analiza tabel

❖ Tabele krzyżowe

❖ χ^2

❖ Przykład

❖ Kappa Cohena

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Po konsultacjach kryteria zostały doprecyzowane:

```
kappa2 (CGinstruct[, 2:3])
```

```
## Cohen's Kappa for 2 Raters (Weights: unweighted)
##
## Subjects = 97
## Raters = 2
## Kappa = 0.954
##
## z = 14.5
```

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

❖ Korelacja
Pearsona

❖ Korelacja
Spearmana

❖ Wykresy

❖ Istotność korelacji

Analiza regresji

Regresja logistyczna

ANOVA

Analiza korelacji

Współczynnik korelacji Pearsona

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

❖ Korelacja Pearsona

❖ Korelacja Spearmana

❖ Wykresy

❖ Istotność korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Miara liniowej zależności cech $r \in [-1, 0]$
 - ◆ $r \approx 0$ — brak zależności *liniowej*
 - ◆ $r \approx 1$ — silna *liniowa* zależność dodatnia
 - ◆ $r \approx -1$ — silna *liniowa* zależność ujemna
- Liniowość zależności można sprawdzić na wykresie `plot()`
- Współczynnik determinacji $s = r^2 \in [0, 1]$
 - ◆ jaki procent jednej zmiennej wyjaśnia zmienność drugiej zmiennej
- Charakter zależności
 - ◆ A zależy od B
 - ◆ B zależy od A
 - ◆ A i B zależą jedno od drugiego
 - ◆ A i B zależą od trzeciej cechy

Obliczenie współczynnika korelacji

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

❖ Korelacja Pearsona

❖ Korelacja Spearmana

❖ Wykresy

❖ Istotność korelacji

Analiza regresji

Regresja logistyczna

ANOVA

```
cor(5:15, 7:17)
```

```
cor(5:15, c(7:16, 23))
```

Macierz korelacji

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

❖ Korelacja Pearsona

❖ Korelacja Spearmana

❖ Wykresy

❖ Istotność korelacji

Analiza regresji

Regresja logistyczna

ANOVA

`cor(trees)`

- Brakujące dane:

- ◆ `use=all.obs` — komunikat o błędzie
- ◆ `use=complete.obs` — usuwane wszystkie wiersze
- ◆ `use=pairwise.complete.obs` — usuwane wiersze tylko dla par kolumn
 - różne ilości obserwacji

Współczynnik korelacji Spearmana

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

❖ Korelacja Pearsona

❖ Korelacja Spearmana

❖ Wykresy

❖ Istotność korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Dane nieparametryczne
- Współczynnik korelacji rang

```
x <- rexp(50)
lx <- log(x)
plot(x, lx)
cor(x, lx)
cor(x, lx, method="spearman")
```

Tekstowa ilustracja korelacji

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

❖ Korelacja Pearsona

❖ Korelacja Spearmana

❖ Wykresy

❖ Istotność korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Zestaw danych `longley`

```
cor.l <- cor(longley)
```

- funkcja `symnum()`

```
symnum(cor.l)
```


Graficzna ilustracja korelacji

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

❖ Korelacja Pearsona

❖ Korelacja Spearmana

❖ Wykresy

❖ Istotność korelacji

Analiza regresji

Regresja logistyczna

ANOVA

● wykres `image()`

```
image(1:ncol(cor.l), 1:nrow(cor.l), cor.l,
      col=heat.colors(22), axes=FALSE,
      xlab="", ylab="")
axis(1, at=1:ncol(cor.l),
      labels=abbreviate(colnames(cor.l)))
axis(2, at=1:nrow(cor.l),
      labels=abbreviate(rownames(cor.l)),
      las = 2)
```

● wykres `plotcorr()`

```
library(ellipse)
colnames(cor.l) <- abbreviate(colnames(cor.l))
rownames(cor.l) <- abbreviate(rownames(cor.l))
plotcorr(cor.l, type="lower", mar=c(0,0,0,0))
```

Graficzna ilustracja korelacji

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

❖ Korelacja

Pearsona

❖ Korelacja

Spearmana

❖ Wykresy

❖ Istotność korelacji

Analiza regresji

Regresja logistyczna

ANOVA

● wykres `corrplot()`

```
library(corrplot)
corrplot(cor.l, method="circle")
corrplot(cor.l, method="pie")
corrplot(cor.l, method="color")
corrplot(cor.l, method="number")
corrplot(cor.l, type="upper")
corrplot(cor.l, type="lower")
corrplot(cor.l, type="upper", order="hclust")
col <- colorRampPalette(
  c("#FFFFFF", "#EE9988"))
corrplot(cor.l, col=col(20),
  method="color", addCoef.col = "black",
  tl.col = "black", number.cex=0.75
)
```

Istotność statystyczna korelacji

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

❖ Korelacja

Pearsona

❖ Korelacja

Spearmana

❖ Wykresy

❖ Istotność korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- Hipoteza zerowa $r = 0$
- Jeżeli H_0 jest odrzucana, to korelacja jest istotna

```
with(trees, cor.test(Girth, Height))
```

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet
Anscombe'a

❖ Porównywanie
regresji

❖ Regresja
nieliniowa

Regresja logistyczna

ANOVA

Analiza regresji

Model regresji

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

- Regresja liniowa

$$m = b_0 + b_1 x$$

- ❖ m jest wartością oczekiwaną zmiennej y przy znanej wartości x , b_0 , b_1 — parametry regresji

- Błąd regresji (reszty)

$$E = y - m$$

- Metoda najmniejszych kwadratów

$$\sum (y_i - (b_0 + b_1 x_i))^2 \rightarrow \min$$

Oszacowanie dopasowania modelu

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

- W przypadku idealnym reszty mają rozkład normalny $N(0, \sigma)$, σ nie zależy od x, y
 - ◆ jeżeli średnia reszt zależy od x , to istnieje zależność nieliniowa
 - zawsze warto zobaczyć na wykresie

- Współczynnik zbieżności (R-squared)

$$\phi^2 = 1 - \frac{\sigma_m^2}{\sigma_y^2}$$

- Oszacowanie dopasowania modelu do danych: test F Fishera
 - ◆ hipoteza zerowa: model jest dopasowany (średnia reszt jest zerowa)

Przykład

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

- Dane `women` wzrost (w calach) i waga kobiet (w funtach)

- ◆ przeliczamy w metry i kilogramy

```
women.metr <- women
women.metr$height <-
  0.0254*women.metr$height
women.metr$weight <-
  0.45359237*women.metr$weight
```

- ◆ obliczenie parametrów regresji

(zmienna zależna ~ predyktor)

```
lm.women <- lm(formula = weight ~ height,
  data = women.metr)
lm.women$coefficients
```

- ◆ wyniki dopasowania

```
summary(lm.women)
```

- kwadrat wzrostu

Przykład. Wykres

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet

Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

```
plot(women.metr$height, women.metr$weight,  
     main="", xlab="wzrost (m)", ylab="waga (kg)")
```

```
b0 <- lm.women$coefficient[1]
```

```
b1 <- lm.women$coefficient[2]
```

```
x1 <- min(women.metr$height)
```

```
x2 <- max(women.metr$height)
```

```
x <- c(x1, x2)
```

```
y <- b0 + b1*x;
```

```
lines(x, y, col="red")
```

```
grid()
```


Kwartet Anscombe'a

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

- cztery zestawy danych o identycznych cechach
 - ◆ średnia arytmetyczna, wariancja
 - ◆ współczynnik korelacji
 - ◆ równanie regresji liniowej
- różne przedstawienia graficzne.

```
ab <- read.table("data/anscombe.csv", sep=";",  
                 header = TRUE)
```

```
attach(ab)  
mean(ay1); mean(ay2); mean(ay3); mean(ay4)  
sd(ay1); sd(ay2); sd(ay3); sd(ay4)  
cov(ax1, ay1); cov(ax2, ay2)  
    cov(ax3, ay3); cov(ax4, ay4)  
lm(ay1 ~ ax1); lm(ay2 ~ ax2)  
    lm(ay3 ~ ax3); lm(ay4 ~ ax4)  
detach(ab)
```

Wykresy

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

```
old.par <- par(mfrow=c(2,2))
for (i in 1:4) {
  lms <- lm(ab[[2*i]] ~ ab[[2*i-1]])
  plot(ab[[2*i-1]], ab[[2*i]],
       xlim = c(0,20), ylim = c(4, 13),
       xlab='x', ylab='y')
  abline(lms, col='red')
}
par(old.par)
```

- `abline(a, b)` dodaje do wykresu linię $y = a + bx$

Skrypt *anscombe.r*

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

- Wewnątrz R

```
source("functions/anscombe.r")
```

- W konsoli

```
$ Rscript functions/anscombe.r
```

- ✦ nie zobaczymy wykresu :(
 - jakieś pomysły?

Porównywanie regresji

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet

Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

- Porównywanie regresji opartych na tych samych danych
- Przykład: wybory
 - ◆ Procent wyborców zagłosowawszych na kandydata zależy od procentu zagłosowawszych wyborców
 - zależność jest inna dla różnych kandydatów
 - ◆ charakteryzuje wyborców
 - ◆ Dane są w pliku `data/wybory.txt`
 - ◆ Wczytywanie i weryfikacja

```
wybory <- read.table("data/wybory.txt",  
                      h=TRUE)
```

```
str(wybory)
```

```
head(wybory)
```

```
attach(wybory)
```

Obliczenie korelacji

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

```
cand.percent <- cbind(KAND.1, KAND.2, KAND.3) /  
  WYBORCÓW  
votes.percent <- (WAŻNE + NIEWAŻNE) / WYBORCÓW  
cor(cand.percent, votes.percent)
```

- Wyborcy którego kandydata się różnią?

Obliczenie regresji

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

```
lm.1 <- lm(cand.percent[,1] ~ votes.percent)
lm.2 <- lm(cand.percent[,2] ~ votes.percent)
lm.3 <- lm(cand.percent[,3] ~ votes.percent)
lapply(list(lm.1, lm.2, lm.3), summary)
```

Wykres

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet

Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

```
plot(cand.percent[,3] ~ votes.percent,
     xlim=c(0,1), ylim=c(0,1),
     xlab="Zagłosowało",
     ylab="Odsetek zagłosowawszych na kandydata")
points(cand.percent[,1] ~ votes.percent, pch=2)
points(cand.percent[,2] ~ votes.percent, pch=3)
abline(lm.3)
abline(lm.2, lty=2)
abline(lm.1, lty=3)
legend("topleft", lty=c(3,2,1),
       legend=c("Kandydat 1", "Kandydat 2",
                "Kandydat 3"))
detach(wybory)
```

Regresja nieliniowa

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

- Przedstawienie graficzne
- Przykład: kolor pierwiosnków
 - ◆ w Chinach wzdłuż drogi nr. G318 rosną pierwiosniki
 - im bliżej Himalajów, tym mniej kwiatów białych i żółtych i więcej różowych, czerwonych i fioletowych
- *W jaki sposób zmienia się kolor?*
 - ◆ dane są w pliku `data/primula.txt`
 - ◆ brawa kwiatów zakodowana jest proporcją — `yfrac`
 - ◆ odległość — `nwse`

Wizualizacja

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

❖ Model regresji

❖ Przykład

❖ Kwartet Anscombe'a

❖ Porównywanie regresji

❖ Regresja nieliniowa

Regresja logistyczna

ANOVA

```
prp <- read.table("data/primula.txt",
  h=TRUE)
plot(yfrac ~ nwse, data=prp,
  xlab="Odległość od Himalajów, km",
  ylab="Proporcja kwiatów jasnych, %")
rect(129, -10, 189, 110, col=gray(.8), border=NA)
box()
mtext("129", at=128, side=3, line=0, cex=.8)
mtext("189", at=189, side=3, line=0, cex=.8)
points(yfrac ~ nwse, data=prp)
abline(lm(yfrac ~ nwse, data=prp), lty=2)
lines(loess.smooth(prp$nwse,
  prp$yfrac), lty=1)
```

- `loess.smooth(x, y)` — lokalne wygładzanie łamanej przechodzącej przez punkty (x, y)
- Wniosek: regresja liniowa nie odzwierciadla zależności

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

❖ Zmienna
dychotomiczna

❖ AIC

ANOVA

Regresja logistyczna

Zmienna dychotomiczna

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

❖ Zmienna dychotomiczna

❖ AIC

ANOVA

- Zmienna niezależna jest interwałowa, zmienna objaśniana — dychotomiczna (sukces/niepowodzenie)
- Prawdopodobieństwo sukcesu
- Przykład: wpływ stażu pracy na jakość napisania kodu
 - ◆ programista pisze kod w ciągu 20 minut
 - ◆ testowanie: działa/nie działa

```
l <- read.table("data/logit.txt")  
head(l)
```

```
l.logit <- glm(formula=V2 ~ V1,  
               family=binomial, data=l)  
summary(l.logit)
```

Kryterium informacyjne Akaikego

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

❖ Zmienna dychotomiczna

❖ AIC

ANOVA

- Jeden ze wskaźników dopasowania modelu
- Im mniejsze, tym lepiej dopasowanie
- Przykład: zatrucie pokarmowe
 - ◆ które z dań było zepsute?
 - operator `I()`, aby minus nie był traktowany jako część formuły
 - funkcja `update()` przelicza model
 - ◆ funkcja *generyczna*

```
tox.logit <- glm(formula=I(2-ILL) ~  
  CAESAR + TOMATO,  
  family=binomial, data=tox)  
tox.logit2 <- update(tox.logit, . ~ . - TOMATO)  
tox.logit3 <- update(tox.logit, . ~ . - CAESAR)  
tox.logit$aic; tox.logit2$aic; tox.logit3$aic
```

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

- ❖ Wiele prób
- ❖ Przykład
- ❖ Test nieparametryczny

ANOVA

Porównywanie wielu prób

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

❖ Wiele prób

❖ Przykład

❖ Test
nieparametryczny

- t-test Studenta
 - ◆ kumulacja błędu I rodzaju
- Analiza wariancji
 - ◆ H_0 : próby się nie różnią
 - ◆ jaka H_1 ?
- Funkcja `anova(object, ...)`
 - ◆ `object` zawiera wyniki dopasowania modelu
 - na przykład wynik funkcji `lm()` bądź `glm()`
 - przykładowo: `lm(zależna ~ predyktor)`
 - ◆ test F Fishera
 - ◆ prawdopodobieństwo błędu I rodzaju: $\Pr(>F)$
- Test parametryczny

Przykład

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

❖ Wiele prób

❖ Przykład

❖ Test
nieparametryczny

- W miasteczku (200 mieszkańców) badano wpływ diety na rzut serca
 - ◆ Wybrano losowo 28 mieszkańców
 - ◆ Podzielono na 4 grupy po 7 osób
 - Pierwsza grupa nie zmieniała diety
 - Druga jadła same makarony
 - Trzecia — mięso
 - Czwarta — owoce
 - ◆ po miesiącu zmierzono rzut serca
- Czy dieta ma wpływ na rzut serca?
- Hipoteza zerowa: nie ma wpływu.

Dane

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

❖ Wiele prób

❖ Przykład

❖ Test nieparametryczny

```
cardiac <-  
  data.frame (  
output=  
c(4.6, 4.7, 4.9, 4.6, 5.4, 5.1, 5.3, # kontrolna  
4.6, 5.0, 5.2, 5.2, 5.6, 5.5, 5.5, # makarony  
5.6, 5.3, 4.9, 5.2, 4.9, 4.4, 4.3, # mięso  
4.9, 4.4, 4.8, 4.5, 4.9, 4.8, 5.6), # owoce  
diet = rep( c("Grupa kontrolna", "Makarony",  
"Mięso", "Owoce"), c(7, 7, 7, 7) )  
)
```

- Plik data/diet.r

Test

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

❖ Wiele prób

❖ Przykład

❖ Test
nieparametryczny

```
boxplot(output ~ diet, data=cardiac)  
anova(lm(output ~ diet, data=cardiac))
```

- Czy dieta ma wpływ na rzut serca?
- Test studenta dla par (z poprawką dla porównywań wielokrotnych):

```
pairwise.t.test(cardiac$output, cardiac$diet)
```

Test Kruskala-Wallis

Testowanie dwóch prób

Analiza tabel

Analiza korelacji

Analiza regresji

Regresja logistyczna

ANOVA

❖ Wiele prób

❖ Przykład

❖ Test
nieparametryczny

```
kruskal.test(cardiac$output ~ cardiac$diet)
```