Zaawansowane modele liniowe - Lista 3

Klaudia Weigel

1 Zadanie 1

Wygenerujemy losową macierz $X \in \mathbb{R}_{1000\times 2}$, taką że $X_{ij} \sim^{iid} N(0, \sigma = 1/\sqrt{1000})$, oraz ciąg predyktorów liniowych $\eta = X\beta$, gdzie $\beta = (3,3)'$. Na ich podstawie wygenerujemy 10000 niezależnych replikacji wektora odpowiedzi y z rozkładu Poissona. Dla każdej replikacji wektora odpowiedzi y dopasujemy model regresji ujemnej dwumianowej i regresji Poissona i wyznaczymy na ich podstawie ciąg $\hat{\alpha}$ oraz statystyk z problemu testowania:

$$H_0 : \text{dane pochodzą z rozkładu Poissona}$$

$$przeciwko$$

$$H_1 : \text{dane pochodzą z rozkładu ujemnego dwumianowego.}$$

Statystyka testowa ma postać

$$\chi^2 = D(M_1) - D(M_2)$$

Przy H_0 statystyka ma asymptotycznie rozkład będący mieszanką, rozkładu skoncentrowanego w 0 oraz rozkładu χ^2 z jednym stopniem swobody.

$$\chi^2 \sim 0.5F_0 + 0.5F_{\chi^2(df=1)}.$$

Na poziomie istotności q hipotezę zerową odrzucamy dla wartości statystyki χ^2 większych od kwantyla rzędu $1-2q$ z rozkładu χ^2 z 1 stopniem swobody.

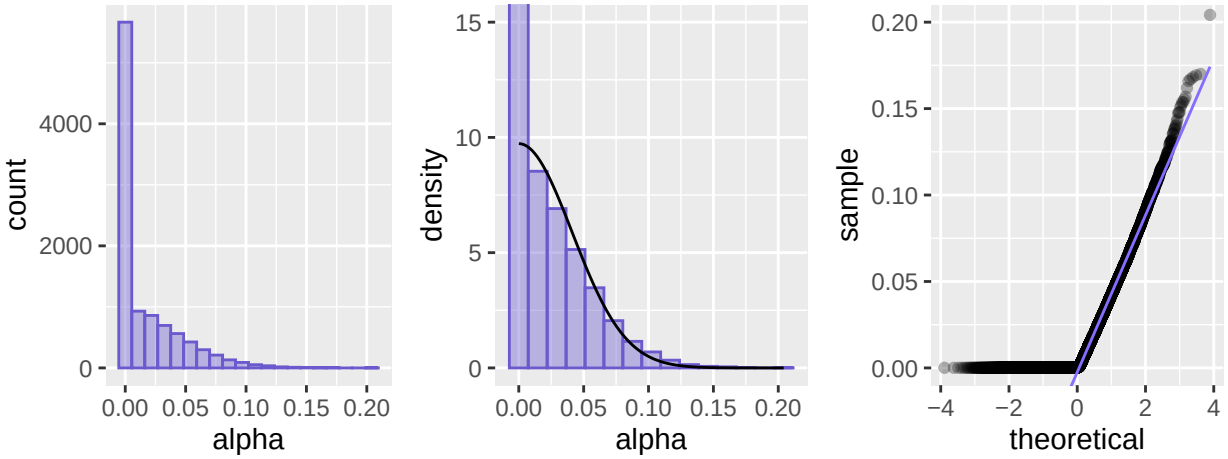
```
x = matrix(rnorm(2000, mean = 0, sd = 1/sqrt(1000)), nrow = 1000, ncol = 2)

lambda_i = exp(x%*%c(3,3)); reps = 10000
Y = replicate(reps, rpois(1000, lambda = lambda_i))

ex_1 = function(y) {
  mod1 = glm(y~x-1, family = poisson())
  mod2 = glm.nb(y~x-1)
  alpha = 1/mod2$theta
  chi2 = 2*(logLik(mod2) - logLik(mod1))

  return(list(alpha = alpha, chi2 = chi2))
}

mydist = function(x) { 0.5*dchisq(x, df=1) }

res = apply(Y, 2, function(y) ex_1(y))
alpha = sapply(1:reps, function(i) res[[i]]$alpha)
chi2 = sapply(1:reps, function(i) res[[i]]$chi2)
```

1.1 Estymator $\hat{\alpha}$

Poniżej przedstawiony jest histogram oraz wykres kwantylowo-kwantylowy dla $\hat{\alpha}$. Histogram porównujemy z rozkładem normalnym ze średnią 0 oraz odchyleniem standardowym, które można przybliżyć przez:

$$\hat{\sigma} \approx \frac{F^{-1}(0.75)}{\Phi^{-1}(0.75)},$$

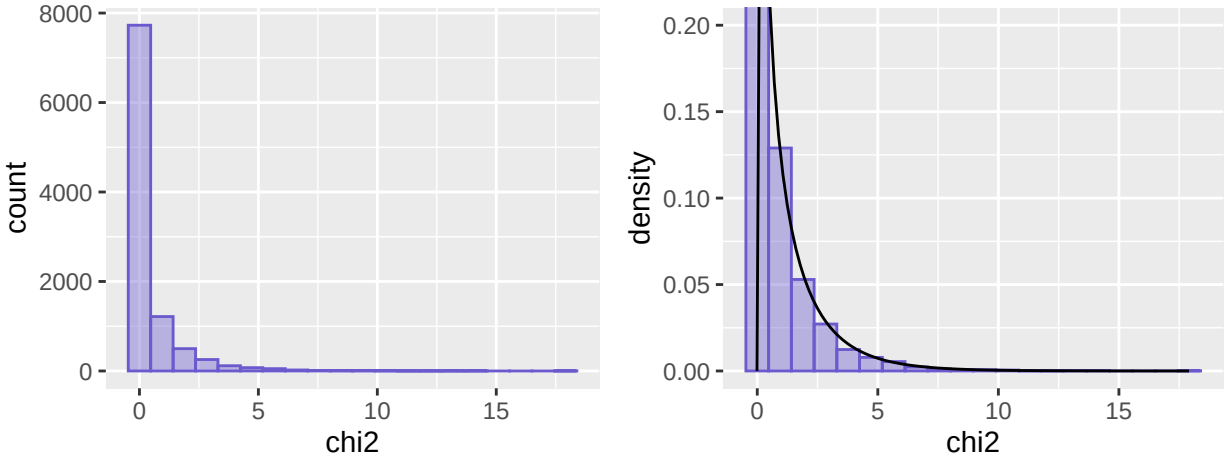
gdzie $F^{-1}(0.75)$ jest kwantylem próbkowym rzędu 0.75.



Bardzo dużo wartości estymatora jest bliskich zeru. Dla wartości na prawo od zera rozkład estymatora jest bliski rozkładowi normalnemu, co potwierdza także wykres kwantylowo-kwantylowy. Wyniki są zgodne z teorią przedstawioną na wykładzie.

1.2 χ^2

Poniżej mamy histogram dla replikacji statystyki χ^2 wraz z przeskalowaną gęstością $(0.5 * F_{\chi^2(df=1)})$.

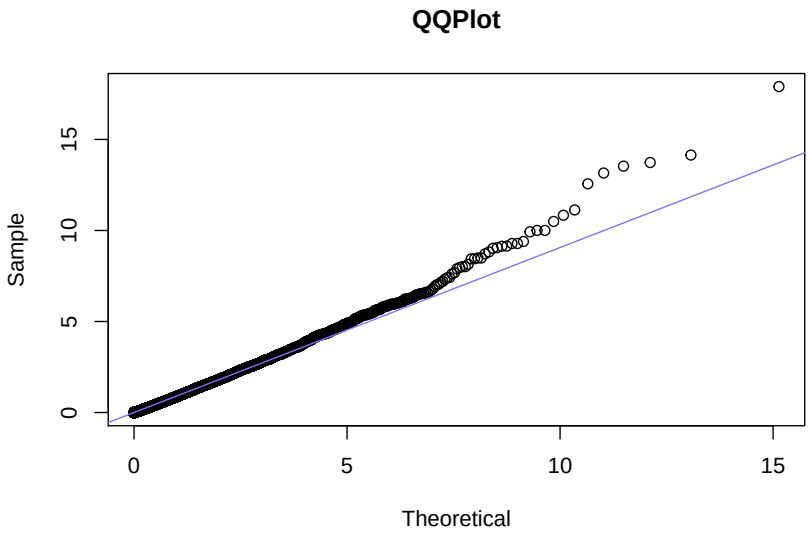


Wyznaczmy teraz funkcję kwantylową dla $F = 0.5F_0 + 0.5F_{\chi^2(df=1)}$. Dla $q < 0.5$, $F^{-1} = 0$, jako że oba składniki mieszanki przyjmują wartości nieujemne, a połowa masy znajduje się w zerze. Jeśli chcemy znaleźć kwantyl rzędu $q = 0.5 + q_0$, wystarczy znaleźć kwantyl rzędu $q_0 = q - 0.5$ z rozkładu $0.5F_{\chi^2(df=1)}$.

$$0.5F_{\chi^2(df=1)}(x) = q - 0.5 \iff F_{\chi^2(df=1)}(x) = \frac{q-0.5}{0.5} \iff x = F^{-1}_{\chi^2(df=1)}\left(\frac{q-0.5}{0.5}\right).$$

Czyli musimy wyznaczyć kwantyl rzędu $\frac{q-0.5}{0.5}$ z rozkładu $F_{\chi^2(df=1)}$.

```
myquantile = function(q) {
  q = sapply(q, function(qi) max(0, (qi-0.5)/0.5))
  qchisq(q, df=1)
}
qqplot(myquantile(ppoints(reps)), chi2, main = "QQPlot", xlab = "Theoretical", ylab = "Sample")
qqline(chi2, distribution = myquantile, probs = c(0.5, 0.9), col = 'slateblue1')
```



2 Zadanie 2

Chcemy zbadać związek pomiędzy liczbą wizyt w gabinecie lekarskim (zmienna zależna, kolumna "ofp") i zmiennymi niezależnymi opisującymi pacjenta:

- "hosp" – liczba pobytów w szpitalu,
- "health" – zmienna opisująca subiektywny odczucie pacjenta o jego zdrowiu,
- "numchron" – liczba przewlekłych stanów chorobowych,
- "gender" – płeć,
- "school" – liczba lat edukacji,
- "privins" – indykator opisujący to czy pacjent ma dodatkowe prywatne ubezpieczenie zdrowotne.

```
debtrivedi =  read.csv("DebTrivedi.csv")[ ,c('ofp', 'hosp', 'health', 'numchron',
                                             'gender', 'school', 'privins')]
head(debtrivedi)
```

```
##   ofp hosp  health numchron gender school privins
## 1   5    1 average        2   male      6     yes
## 2   1    0 average        2 female     10     yes
## 3  13    3    poor        4 female     10      no
## 4  16    1    poor        2   male      3     yes
## 5   3    0 average        2 female      6     yes
## 6  17    0    poor        5 female      7      no
```

3 Zadanie 3

Wykonamy wstępną analizę danych z zadania 2.

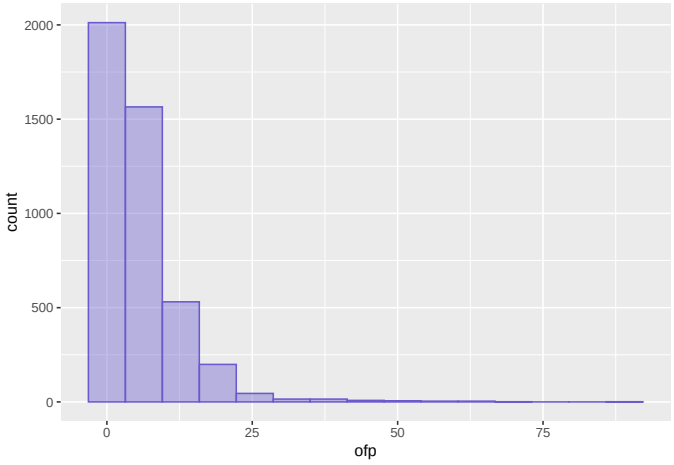


Figure 1: Histogram zmiennej extitofp.

Widzimy, że bardzo dużo wartości jest równych zero, mamy więc do czynienia z inflacją w zerze.

```
sum(debtrivedi$ofp ==0)
```

```
## [1] 683
```

```
mean(debtrivedi$ofp); var(debtrivedi$ofp)
```

```
## [1] 5.774399
```

```
## [1] 45.68712
```

Wariancja zmiennej objaśnianej jest dużo wyższa niż średnia, mamy do czynienia ze zjawiskiem nadmiernej dyspersji.

Ze względu na dużą ilość zerowych wartości zmiennej ofp, wprowadzimy teraz nową zmienną $f(ofp)$, taką że

$$f(ofp) = \log(ofp + 0.5)$$

Na potrzeby wykonania wykresów, pogrupujemy wartości danego regresora tam gdzie jest mało obserwacji: dla zmiennych *hosp* oraz *numchron*.

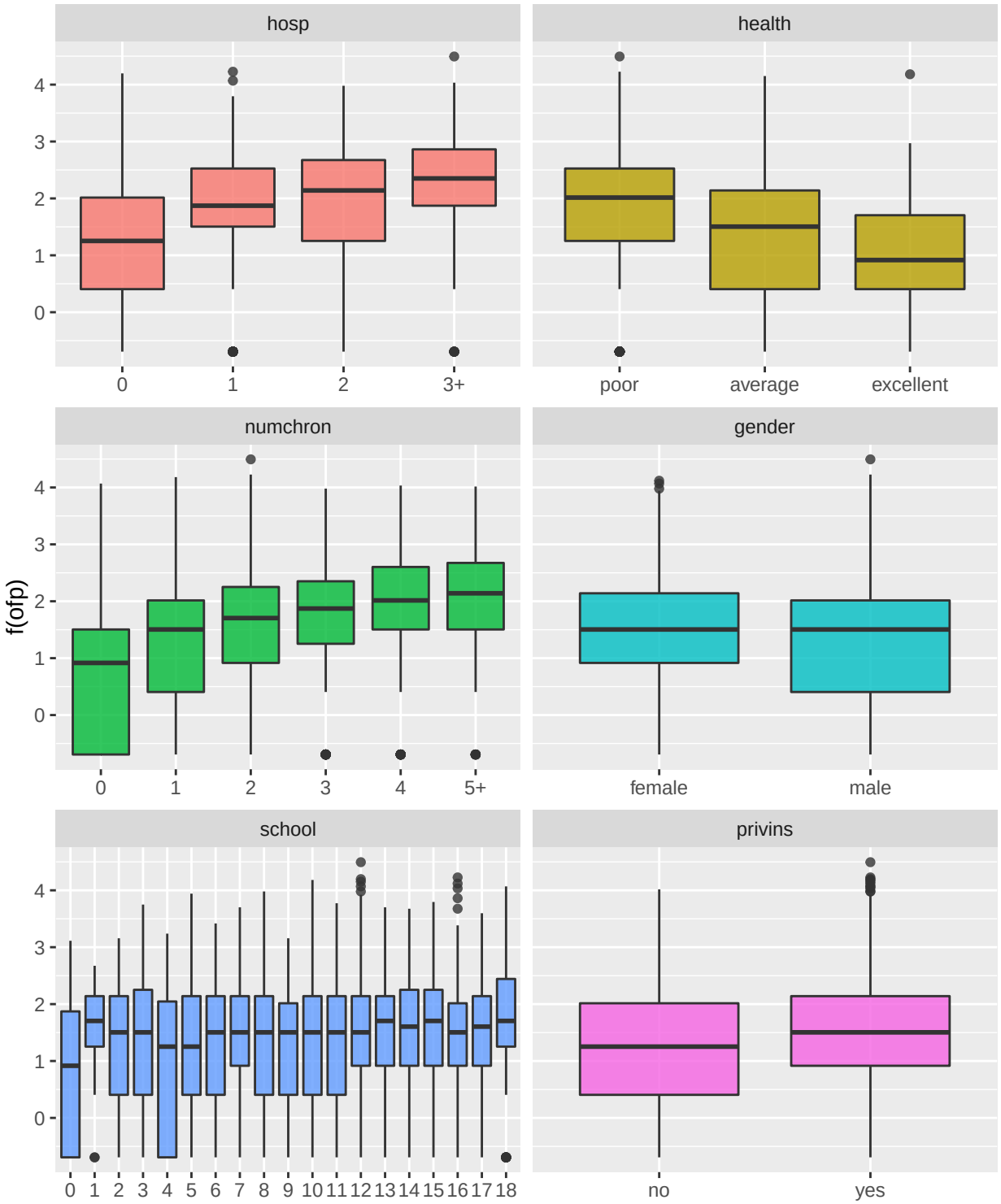
```
table(debtrivedi$hosp)
```

```
##
##    0    1    2    3    4    5    6    7    8
## 3541  599  176   48   20   12    5    1    4
```

```
table(debtrivedi$numchron)
```

```
##
##    0    1    2    3    4    5    6    7    8
## 1025 1498  968  525  220  127   34    6    3
```

Możemy pogrupować liczbę pobytów w szpitalu na “1”, “2”, “3 i więcej”, natomiast ilość chorób przewlekłych na “1”, “2”, “3”, “4”, “5 i więcej”.



Ilość wizyt w gabinecie lekarskim rośnie ze względu na ilość pobytów szpitalu (*hosp*) oraz maleje wraz z poprawą subiektywnego odczucia pacjenta o jego zdrowiu (*health*). Obserwujemy także wzrost w ilości wizyt w zależności od ilości chorób przewlekłych (*numchrom*). W grupie mężczyzn więcej obserwacji leży poniżej mediany, niż w grupie kobiet, można więc przypuszczać, że kobiety częściej odwiedzają gabinet lekarski. Dla edukacji dłuższej niż 11 lat, rozkład danych staje się bardziej symetryczny i mniej elementów jest poniżej mediany. Możemy oczekiwać, że dłuższa edukacja będzie zwiększać wartość średniej liczby wizyt. Osoby posiadające prywatne ubezpieczenie medyczne częściej odwiedzają gabinet lekarski, niż osoby nie mający takiego ubezpieczenia.

4 Zadanie 4

4.1 Regresja Poissona

Regresja Poissona zakłada, że zmienne objaśniane y_i są realizacjami niezależnych zmiennych losowych z rozkładu Poissona ze średnią λ_i. Dla każdego i związek między parametrem λ_i a $(x_{i,1}, \ldots, x_{i,p-1})'$ ma postać:

$$\log(\lambda_i) = \beta_0 + x_{i,1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1},$$

gdzie $\beta \in \mathbb{R}^p$ to nieznany wektor parametrów.

Pierwotne modele w tym i następnych punktach będziemy tworzyć w oparciu o wszystkie zmienne: 3 kategoryczne (*health*, *gender* i *privins*) oraz 3 ciągłe (*hosp*, *numchron* i *school*).

```
mod_pois = glm(ofp~ . , data = debtrivedi, family = poisson())
summary(mod_pois)
```

```
##
## Call:
## glm(formula = ofp ~ ., family = poisson(), data = debtrivedi)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4055  -1.9962  -0.6737   0.7049  16.3620
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.028874   0.023785  43.258   <2e-16 ***
## hosp              0.164797   0.005997  27.478   <2e-16 ***
## healthexcellent -0.361993   0.030304 -11.945   <2e-16 ***
## healthpoor        0.248307   0.017845  13.915   <2e-16 ***
## numchron          0.146639   0.004580  32.020   <2e-16 ***
## gendermale       -0.112320   0.012945  -8.677   <2e-16 ***
## school            0.026143   0.001843  14.182   <2e-16 ***
## privinsyes        0.201687   0.016860  11.963   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 23168  on 4398  degrees of freedom
## AIC: 35959
##
## Number of Fisher Scoring iterations: 5
```

Wszystkie zmienne są mocno istotne.

Sprawdźmy ile wartości zerowych zmiennej objaśnianej ofp przewiduje ten model:

```
mu = predict(mod_pois, type = "response")
sum(dpois(0, mu))
```

```
## [1] 46.71402
```

Przewidywana ilość jest znacznie mniejsza niż prawdziwa ilość zer.

4.2 Regresja ujemna dwumianowa

Zmienna losowa Y ma rozkład ujemny dwumianowy $NB(\mu, \alpha)$ z parametrami $\mu > 0$ i $\alpha > 0$, gdy przyjmuje wartości ze zbioru $\{0, 1, 2 \dots\}$ z prawdopodobieństwem:

$$P(Y = y) = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(y + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu}\right)^{y}.$$

Regresja ujemna dwumianowa zakłada, że zmienne objaśniane y_i są realizacjami niezależnych zmiennych losowych z rozkładu ujemnego dwumianowego $NB(\mu_i, \alpha)$. Dla każdego i związek między parametrem μ_i, a $(x_{i,1}, \dots, x_{i,p-1})'$ ma postać:

$$\log(\mu_i) = \beta_0 + x_{i,1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1},$$

gdzie $\beta \in \mathbb{R}^p$ to nieznany wektor parametrów.

```
mod_nb = glm.nb(ofp~., data = debtrivedi)
summary(mod_nb)
```

```
##
## Call:
## glm.nb(formula = ofp ~ ., data = debtrivedi, init.theta = 1.206603534,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0469  -0.9955  -0.2948   0.2961   5.8185
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.929257   0.054591  17.022  < 2e-16 ***
## hosp             0.217772   0.020176  10.793  < 2e-16 ***
## healthexcellent -0.341807   0.060924  -5.610 2.02e-08 ***
## healthpoor       0.305013   0.048511   6.288 3.23e-10 ***
## numchron         0.174916   0.012092  14.466  < 2e-16 ***
## gendermale      -0.126488   0.031216  -4.052 5.08e-05 ***
## school           0.026815   0.004394   6.103 1.04e-09 ***
## privinsyes       0.224402   0.039464   5.686 1.30e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.2066) family taken to be 1)
##
##     Null deviance: 5743.7  on 4405  degrees of freedom
## Residual deviance: 5044.5  on 4398  degrees of freedom
## AIC: 24359
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.2066
##           Std. Err.:  0.0336
##
##  2 x log-likelihood:  -24341.1070
```

Jak poprzednio wszystkie zmienne są istotne.

4.3 Regresja Poissona z inflacją w zerze (ZIPR)

W modelu regresji Poissona z inflacją w zerze zakładamy że zmienne objaśniane $y_1, \ldots, y_n$ są realizacjami niezależnych zmiennych losowych $Y_1, \ldots, Y_n$ które pochodzą z mieszanki rozkładu dwupunktowego i Poissona:

$$P(Y_i = k) = \begin{cases} \pi_i + (1-\pi_i)e^{\mu_i} & k = 0 \\ (1-\pi_i)\frac{\mu_i^k}{k!}e^{-\mu_i} & k = 1, 2, \ldots \end{cases}$$

gdzie $\pi_i \in [0, 1]$ oraz $\mu_i > 0$.
Związek między parametrami μ_i oraz π_i, a $(x_{i,1}, \ldots, x_{i,p-1})'$ oraz $(z_{i,1}, \ldots, z_{i,m-1})'$ ma postać:

$$\log(\mu_i) = \beta_0 + x_{i,1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1},$$
$$\text{logit}(\pi_i) = \gamma_0 + z_{i,1}\gamma_1 + \cdots + z_{i,m-1}\gamma_{m-1},$$

gdzie $\beta \in \mathbb{R}^p$ i $\gamma \in \mathbb{R}^m$ to nieznane wektory parametrów.

```
mod_zipr = zeroinfl(ofp~., data = debtrivedi)
summary(mod_zipr)
```

```
##
## Call:
## zeroinfl(formula = ofp ~ ., data = debtrivedi)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -5.4092 -1.1579 -0.4769  0.5435 25.0380
##
## Count model coefficients (poisson with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.405812   0.024175  58.152  < 2e-16 ***
## hosp              0.159011   0.006060  26.239  < 2e-16 ***
## healthexcellent -0.304134   0.031151  -9.763  < 2e-16 ***
## healthpoor        0.253454   0.017705  14.315  < 2e-16 ***
## numchron          0.101836   0.004721  21.571  < 2e-16 ***
## gendermale       -0.062332   0.013054  -4.775 1.80e-06 ***
## school            0.019144   0.001873  10.221  < 2e-16 ***
## privinsyes        0.080557   0.017145   4.699 2.62e-06 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.08102    0.14233  -0.569 0.569202
## hosp            -0.30330    0.09158  -3.312 0.000927 ***
## healthexcellent  0.23785    0.14990   1.587 0.112563
## healthpoor       0.02166    0.16170   0.134 0.893419
## numchron        -0.53117    0.04601 -11.545  < 2e-16 ***
## gendermale       0.41527    0.08919   4.656 3.22e-06 ***
## school          -0.05677    0.01223  -4.640 3.49e-06 ***
## privinsyes      -0.75294    0.10257  -7.341 2.12e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 20
## Log-likelihood: -1.613e+04 on 16 Df
```

Obserwujemy, że w modelu inflacji w zerze (zero-inflation model) nieistotna jest zmienna *health*. Porównajmy zatem model oryginalny do modelu bez tej zmiennej. Użyjemy testu opartego na statystyce deviance:

$$\chi^2 = D(M_0) - D(M_A) = 2(\text{logLik}(M_A) - \text{logLik}(M_0)),$$

gdzie M_0 odpowiada uproszczonemu modelowi regresji (hipoteza zerowa), natomiast M_A odpowiada modelowi dla hipotezy alternatywnej. Hipotezę zerową odrzucamy gdy statystyka jest większa od kwantyla z rozkładu chi-kwadrat z ilością stopni swobody równą liczbie wyzerowanych parametrów w M_0.

```
mod_zipr2 = zeroinfl(ofp~.|hosp + numchron + gender + school + privins, data = debtrivedi)
#Test ilorazu wiarogodności
chi_sq = 2*(logLik(mod_zipr) - logLik(mod_zipr2))   # statystka
pchisq(as.numeric(chi_sq), df = 2, lower.tail = F)   #p-wartość
```

```
## [1] 0.2965172
```

Na poziomie istotności 0.05 przyjmujemy hipotezę zerową.

Czyli ostatecznie nasz model to

$$\begin{aligned} log(\mu_i) = \beta_0 + \beta_1 * hosp_i + \beta_2 * \mathbb{I}\{health_i = \text{excellent}\} + \beta_3 * \mathbb{I}\{health_i = \text{poor}\} \\ + \beta_4 * numchron_i + \beta_5 * \mathbb{I}\{gender_i = \text{male}\} + \beta_6 * school_i + \beta_7 * \mathbb{I}\{privins_i = \text{yes}\} \end{aligned}$$

$$logit(\pi_i) = \gamma_0 + \gamma_1 * hosp_i + \gamma_2 * numchron_i + \gamma_3 * \mathbb{I}\{gender_i = \text{male}\} + \gamma_4 * school_i + \gamma_5 * \mathbb{I}\{privins_i = \text{yes}\}$$

4.4 Regresja ujemna dwumianowa z inflacją w zerze (ZINBR)

W modelu regresji ujemnej dwumianowej z inflacją w zerze zakładamy że zmienne objaśniane $y_1, \ldots, y_n$ są realizacjami niezależnych zmiennych losowych $Y_1, \ldots, Y_n$ które pochodzą z mieszanki rozkładu dwupunktowego i ujemnego dwumianowego:

$$P(Y_i = k) = \begin{cases} \pi_i + (1-\pi_i) f(0; \mu_i, \phi) & k = 0 \\ (1-\pi_i) f(k; \mu_i, \phi) & k = 1, 2, \ldots \end{cases}$$

gdzie $\pi_i \in [0, 1]$ oraz $\mu_i > 0$, a f jest funkcją rozkładu prawdopodobieństwa $NB(\mu_i, \phi)$.

Związek między parametrami μ_i oraz π_i a $(x_{i,1}, \ldots, x_{i,p-1})'$ oraz $(z_{i,1}, \ldots, z_{i,m-1})'$ ma postać:

$$\begin{aligned} \log(\mu_i) &= \beta_0 + x_{i,1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1}, \\ \text{logit}(\pi_i) &= \gamma_0 + z_{i,1}\gamma_1 + \cdots + z_{i,m-1}\gamma_{m-1}, \end{aligned}$$

gdzie $\beta \in \mathbb{R}^p$ i $\gamma \in \mathbb{R}^m$ to nieznane wektory parametrów.

```
mod_zinbr = zeroinfl(ofp~., data = debtrivedi, dist = "negbin")
```

Korzystając z wyników `summary` otrzymujemy, że zmienna *health* jest nieistotna w modelu dla inflacji w zerze.

```
mod_zinbr2 = zeroinfl(ofp~.|hosp + numchron + gender + school + privins, data = debtrivedi,
                      dist = "negbin")
chi_sq = 2*(logLik(mod_zipr) - logLik(mod_zipr2))
pchisq(as.numeric(chi_sq), df = 2, lower.tail = F)
```

```
## [1] 0.2965172
```

Na poziomie istotności 0.05 przyjmujemy hipotezę zerową. Ostateczny model ma taką sama postać jak w poprzednim punkcie.

4.5 Model regresji z barierą

W modelu z barierą zakładamy, że zmienne objaśniane y_i są realizacjami niezależnych zmiennych losowych z mieszanki dwóch rozkładów

$$P(Y_i = k) = \begin{cases} f_{zero}(0) & k = 0 \\ (1 - f_{zero}(0)) \frac{f_{count}(k)}{1 - f_{count}(k)} & k = 1, 2, \ldots \end{cases}$$

gdzie $f_{zero}(0)$ opisuje rozkład wartości 0, a $f_{count}(k)$ rozkład zmiennej zliczającej.

Dla każdego i związek między parametrami μ_i^{count} oraz μ_i^{zero} a regresorami ma postać:

$$g_1(\mu_i^{count}) = \beta_0 + x_{i,1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1},$$
$$g_2(\mu_i^{zero}) = \gamma_0 + z_{i,1}\gamma_1 + \cdots + z_{i,m-1}\gamma_{m-1},$$

gdzie $\beta \in \mathbb{R}^p$ i $\gamma \in \mathbb{R}^m$ to nieznane wektory parametrów.

4.5.1 Regresja Poissona z barierą

W tym przypadku f_{count} to rozkład Poissona.

```
mod_hurdlep = hurdle(ofp~., data = debtrivedi)
summary(mod_hurdlep)
```

```
##
## Call:
## hurdle(formula = ofp ~ ., data = debtrivedi)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -5.4144 -1.1565 -0.4770  0.5432 25.0228
##
## Count model coefficients (truncated poisson with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.406459   0.024180  58.167  < 2e-16 ***
## hosp              0.158967   0.006061  26.228  < 2e-16 ***
## healthexcellent -0.303677   0.031150  -9.749  < 2e-16 ***
## healthpoor        0.253521   0.017708  14.317  < 2e-16 ***
## numchron          0.101720   0.004719  21.557  < 2e-16 ***
## gendermale       -0.062247   0.013055  -4.768 1.86e-06 ***
## school            0.019078   0.001872  10.194  < 2e-16 ***
## privinsyes        0.080879   0.017139   4.719 2.37e-06 ***
## Zero hurdle model coefficients (binomial with logit link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.043147   0.139852   0.309 0.757688
## hosp              0.312449   0.091437   3.417 0.000633 ***
## healthexcellent -0.289570   0.142682  -2.029 0.042409 *
## healthpoor       -0.008716   0.161024  -0.054 0.956833
## numchron          0.535213   0.045378  11.794  < 2e-16 ***
## gendermale       -0.415658   0.087608  -4.745 2.09e-06 ***
## school            0.058541   0.011989   4.883 1.05e-06 ***
## privinsyes        0.747120   0.100880   7.406 1.30e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 14
## Log-likelihood: -1.613e+04 on 16 Df
```

Usuwamy zmienną *health* z modelu dla inflacji w zerze.

```
mod_hurdlep2 = hurdle(ofp~.|hosp + numchron + gender + school + privins, data = debtrivedi)
chi_sq = 2*(logLik(mod_hurdlep) - logLik(mod_hurdlep2))
pchisq(as.numeric(chi_sq), df = 2, lower.tail = F)
```

```
## [1] 0.1361861
```

Na poziomie istotności 0.05 nie mamy podstaw do odrzucenia hipotezy zerowej. Przyjmujemy model bez *health*.

4.5.2 Regresja ujemna dwumianowa z barierą

Rozkład f_{count} to rozkład ujemny dwumianowy.

```
mod_hurdlenb = hurdle(ofp~., data = debtrivedi, dist = "negbin")
```

Wyniki `summary` sugerują, że ponownie możemy usunąć zmienną *health* z modelu inflacji w zerze.

```
mod_hurdlenb2 = hurdle(ofp~.|hosp + numchron + gender + school + privins, data = debtrivedi,
                       dist = "negbin")
chi_sq = 2*(logLik(mod_hurdlenb) - logLik(mod_hurdlenb2))
pchisq(as.numeric(chi_sq), df = 2, lower.tail = F)
```

```
## [1] 0.1361861
```

Na poziomie istotności 0.05 przyjmujemy hipotezę zerową, zakładającą model inflacji bez zmiennej *health*.

4.6 Porównanie wyników

Porównamy teraz wyniki zwracane przez wszystkie wyżej wymienione modele.

	GLM		Inflacja w zerze			
	Poisson	NB	ZIPR	ZINBR	Hurdle (Pois)	Hurdle (NB)
Model dla średniej						
(Intercept)	1.0289	0.9293	1.4056	1.1937	1.4065	1.1977
hosp	0.1648	0.2178	0.1590	0.2015	0.1590	0.2119
healthexcellent	-0.3620	-0.3418	-0.3074	-0.3193	-0.3037	-0.3319
healthpoor	0.2483	0.3050	0.2534	0.2851	0.2535	0.3160
numchron	0.1466	0.1749	0.1018	0.1290	0.1017	0.1264
gendermale	-0.1123	-0.1265	-0.0624	-0.0803	-0.0622	-0.0683
school	0.0261	0.0268	0.0192	0.0214	0.0191	0.0207
privinsyes	0.2017	0.2244	0.0805	0.1259	0.0809	0.1002
Model dla inflacji w zerze						
(Intercept)	-	-	-0.0594	-0.0469	0.0159	0.0159
hosp	-	-	-0.3067	-0.8005	0.3184	0.3184
numchron	-	-	-0.5397	-1.2479	0.5478	0.5478
gendermale	-	-	0.4181	0.6477	-0.4191	-0.4191
school	-	-	-0.0556	-0.0838	0.0571	0.0571
privinsyes	-	-	-0.7537	-1.1756	0.7457	0.7457
Parametr dyspersji						
Theta (α^{-1})	-	1.2066	-	1.4831	-	1.3955
Kryteria						
AIC	35959.2256	24359.1072	32298.4871	24211.4440	32300.8825	24210.1432
BIC	36010.3514	24416.6237	32387.9572	24307.3049	32390.3526	24306.0040
Funkcja log-wiarogodności						
logL	-17971.6128	-12170.5536	-16135.2435	-12090.7220	-16136.4412	-12090.0716
Ilość parametrów						
params	8.0000	9.0000	14.0000	15.0000	14.0000	15.0000
Oczekiwana liczba zer						
E(#Zeros)	46.7140	608.0085	682.2980	708.8533	683.0000	683.0000

Estymatory parametrów regresji w modelu dla średniej są podobne dla każdego przypadku. Z kryterium AIC oraz BIC wynika, że najlepszy model to model ujemny dwumianowy z barierą, a zaraz po nim jest model ujemny dwumianowy z inflacją w zerze. Najsłabiej radzą sobie modele wykorzystujące rozkład Poissona. Odzwierciedla to fakt, że nadmierna dyspersja w danych jest lepiej wychwytywana przez modele oparte na rozkładzie ujemnym dwumianowym. Oprócz modelu GLM Poissona, wszystkie modele przewidują ilość zer dość dokładnie, zwłaszcza modele z barierą, gdzie oczekiwana ilość zer jest równa rzeczywistej wartości.