AKADEMIA GÓRNICZO-HUTNICZA

im. Stanisława Staszica w Krakowie

WYDZIAŁ ZARZĄDZANIA

PROJEKT NR 2 - WNIOSKOWANIE BAYESOWSKIE W MODELU REGRESJI WIELORAKIEJ, PROGNOZOWANIE ORAZ WERYFIKACJA HIPOTEZ

METODY BAYESOWSKIE

Autorzy: Krok Justyna, Artur Karamon

Kierunek: Informatyka i Ekonometria

Studia: II Stopnia, Stacjonarne

Kraków – 2020

OPIS BADANIA:

Należy na przykładzie wybranych danych zastosować wnioskowanie bayesowskie w modelu regresji wielorakiej. Badanie składa się z poniższych kroków:

- prezentacja danych oraz rozważanego problemu,
- przygotowanie zbioru testowego (5 danych),
- dobranie odpowiednie parametrów rozkładu a priori i uzasadnienie,
- oszacowanie oraz interpretacja parametrów modelu za pomocą MNK
- badanie autokorelacji, heteroskedastyczności i normalności reszt,
- wyznaczenie parametrów rozkładu a posteriori,
- wyznaczenie rozkładów brzegowych parametrów i porównanie z rozkładami a priori,
- obliczenie bayesowskich estymatorów parametrów modelu, opisanie ich, zbadanie istotności (wyznaczyć HPDI)
- porównanie estymatorów bayesowskich z estymatorami uzyskanymi za pomocą MNK,
- wyznaczenie rozkładów predykcyjnych dla danych testowych, wyznaczenie ich wartości oczekiwanej, HPDI i porównanie wyników z wynikami uzyskanymi w sposób klasyczny,
- porównanie 2 modeli, badanie łączną istotności zmiennych w modelach, porównanie wyników z wartościami testu F.

W rozważanym modelu liniowym przyjąć założenie, że rozkłady a priori wektora β i wariancji σ^2 są niezależne. Stosując algorytm Gibbsa:

- wyznaczyć brzegowe rozkłady a posteriori poszczególnych parametrów modelu,
- oszacować wartości oczekiwane i odchylenia standardowe tych rozkładów (tzn. wyznaczyć „punktowe” oceny parametrów i ich błędy)
- porównać uzyskane wyniki z wynikami uzyskanymi wcześniej (dla rozkładów sprzężonych).

TEORIA:

Regresja wieloraka:

Równanie: $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k} + e_i$ dla $i = 1, ...n$ gdzie $e_i \sim i.i.d.N(0, \sigma^2)$

Nieznane parametry: $\beta = [\beta_0 ... \beta_k]'$, σ^2.

Podejście klasyczne: $\widehat{\beta} = (X'X)^{-1}X'y$, $V(\beta) = \sigma^2(X'X)^{-1}$

Podejście Bayesowskie:

a priori

β_0 - oszacowania parametrów wcześniejszego badania

Σ - macierz $(X'X)^{-1}$ wcześniejszego badania

β ma k-wymiarowy rozkład normalny warunkowy względem σ^2 : $N_k(\beta_0, \sigma^2\Sigma)$

α_0 - liczba stopni swobody wcześniejszego badania $n - k$

δ_0 - suma kwadratów reszt wcześniejszego badania

σ^2 ma rozkład: $IG(\frac{\alpha_0}{2}, \frac{\delta_0}{2})$

a posteriori (rozkłady brzegowe parametrów)

$\beta_1 = \Sigma_1(X'y + \Sigma_0^{-1}\beta_0)$

$\Sigma_1 = (X'X + \Sigma_0^{-1})^{-1}$

β ma k-wymiarowy rozkład t: $t_k(\beta_1, \frac{\delta_1}{\alpha_1}\Sigma_1, \alpha_1)$

β_i ma rozkład t: $t(\beta_{1,i}, \frac{\delta_1}{\alpha_1}\Sigma_{1,ii}, \alpha_1)$

$\alpha_1 = \alpha_0 + n$

$\delta_1 = \delta_0 + y'y - \beta_1'\Sigma_1^{-1}\beta_1 + \beta_0'\Sigma_0^{-1}\beta_0$

σ^2 ma rozkład: $IG(\frac{\alpha_1}{2}, \frac{\delta_1}{2})$

Prognozowanie:

Równanie: $y_i = x_i\beta + e_i$ dla $i = 1, ...n$ gdzie $e_i \sim i.i.d.N(0, \sigma^2)$

Prognoza: y_τ dla $\tau = n + 1, ..., n + m$ przy znanych wartościach x_τ

Podejście klasyczne: $y_\tau = x_\tau\widehat{\beta}$, $S_p^2 = \sigma^2$ (1+ $x_\tau'(X'X)^{-1}x_\tau$)

Podejście Bayesowskie:

rozkład predykcyjny jest m-wymiarowym rozkładem t: $t_m(x_\tau\beta_1, \frac{\delta_1}{\alpha_1}[I_m + x_\tau\Sigma_1x_\tau'], \alpha_1)$

prognoza punktowa: wartość oczekiwana prognoz

prognoza przedziałowa: HPDI

Weryfikacja hipotez:

Porównanie modeli MN (bez restrykcji) oraz MR (z restrykcjami)

iloraz szans a posteriori: $R_{12} = \frac{P(y|MN)}{P(y|MR)} \frac{P(MN)}{P(MR)}$

dla wybranego modelu: $P(y|M) = \frac{1}{\pi^{\frac{n}{2}}} \sqrt{\frac{|\Sigma_1|}{|\Sigma_0|}} \frac{\delta_0^{\alpha_0/2}}{\delta_1^{\alpha_1/2}} \frac{\Gamma(\frac{\alpha_1}{2})}{\Gamma(\frac{\alpha_0}{2})}$

Test Shapiro-Wilka

H_0 : badana cecha ma rozkład normalny

H_1 : badana cecha nie ma rozkładu normalnego

Test Durbina-Watsona

H_0 : brak autokorelacji

H_1 : autokorelacja

Test Breuscha-Pagana

H_0 : składnik losowy jest homoskedastyczny

H_1 : składnik losowy jest heteroskedastyczny

Test F

H_0 : wszystkie zmienne w modelu są nieistotne

H_1 : co najmniej jedna zmienna w modelu jest istotna

Rozkłady brzegowe (algorytm Gibbsa):

$\sigma^2|\beta$ ma rozkład warunkowy odwrotny gamma $(\frac{\alpha_1}{2}, \frac{\delta_1^{(i)}}{2})$.

$\beta|\sigma^2$ ma rozkład warunkowy k wymiarowy $N(\beta_1^{(i)}, \Sigma_1^{(i)})$. gdzie:

$\Sigma_1 = (\sigma^{-2(i-1)} X'X + \Sigma_0^{-1})^{-1}$

$\beta_1 = \Sigma_1 (\sigma^{-2(i-1)} X'y + \Sigma_0^{-1} \beta_0)$

$\delta_1 = \delta_0 + (y - X\beta^{(i)})'(y - X\beta^{(i)})$

DANE:

Do badania wykorzystano dane dla miast na prawach powiatu w Polsce (59 obserwacji). Wybrane zmienne to:

- przeciętne wynagrodzenie
- liczba absolwentów uczelni wyższych
- liczba bezrobotnych (powyżej 1 roku),
- nakłady inwestycyjne przedsiębiorstw,
- stosunek kobiet pracujących do pracowników ogółem,
- liczba pracowników w rozwiniętych sektorach usługowych m.in.: działalność finansowa, ubezpieczeniowa, obsługa rynku nieruchomości.

	Powiat	Wynagrodzenie	Absolwenci	Bezrobotni	Inwestycje	KobietyPrac	PracownicyFUN
1	Powiat m.Jelenia Góra	4546.12	0.0101283342	0.003258681	52905.28	0.5148258	0.007750377
2	Powiat m.Legnica	4259.37	0.0143842444	0.010034887	71053.99	0.5069679	0.012771674
3	Powiat m.Wrocław	5338.47	0.0494530538	0.004245701	92903.07	0.5204789	0.032005095
4	Powiat m.Wałbrzych od 2013	4768.01	0.0074528204	0.006696627	86722.38	0.4877162	0.007238396
5	Powiat m.Bydgoszcz	4481.39	0.0278690919	0.006696594	72599.61	0.5102885	0.021429102
6	Powiat m.Grudziądz	3743.06	0.0012445233	0.015003419	39808.85	0.4973707	0.007975170
7	Powiat m.Toruń	4629.36	0.0364131103	0.008976909	62712.20	0.5045822	0.017642052
8	Powiat m.Włocławek	4166.00	0.0086344496	0.023717983	76101.70	0.4680687	0.012562950
9	Powiat m.Biała Podlaska	3913.47	0.0250931590	0.020313154	34182.43	0.5298408	0.009572465
10	Powiat m.Chełm	3976.28	0.0113314641	0.017408648	46968.95	0.5367785	0.009494176

Rysunek 1: Przykładowe dane.

W związku z faktem, że powyższe zmienne zależą w głównej mierze od liczby mieszkańców, ich wartości zostały podzielone przez liczbę ludności danego miasta (z wyjątkiem stosunku kobiet pracujących do wszystkich oraz średniego wynagrodzenia).

W poniższym badaniu zmienną objaśnianą będzie średnie wynagrodzenie, które za pomocą regresji wielorakiej będzie modelowane przez pozostałe zmienne.

Dobór zmiennych do badania oraz oszacowanie parametrów rozkładu a priori zostało dokonane na podstawie poniższych artykułów:

- https://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5474/11/1/1/regionalne_zroznicowanie_wynagrodzen_1-06-2016.pdf
- http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.desklight-be509e44-8d40-46d6-a0b5-efeffb2ad384/c/13._PRZYCZYNY_ZROZNICOWANIA.pdf

Ze względu na rozbieżność danych, zostały one przeskalowane.

	Powiat	Wynagrodzenie	Absolwenci	Bezrobotni	Inwestycje	KobietyPrac	PracownicyFUN
1	Powiat m.Jelenia Góra	0.001925705	-0.65738007	-0.95529005	-0.46722852	0.141675416	-0.66859136
2	Powiat m.Legnica	-0.477820529	-0.42470061	0.14538995	-0.07108319	-0.071452311	-0.23005072
3	Powiat m.Wrocław	1.327564516	1.49258406	-0.79496536	0.40583281	0.295001422	1.44972204
4	Powiat m.Wałbrzych od 2013	0.373158110	-0.80365596	-0.39685392	0.27092235	-0.593605021	-0.71330581
5	Powiat m.Bydgoszcz	-0.106370628	0.31254402	-0.39685920	-0.03734590	0.018612007	0.52605563
6	Powiat m.Grudziądz	-1.341631439	-1.14307647	0.95244415	-0.75309413	-0.331751329	-0.64895888
7	Powiat m.Toruń	0.141190140	0.77966326	-0.02646055	-0.25316561	-0.136158107	0.19530931
8	Powiat m.Włocławek	-0.634032930	-0.73905384	2.36797779	0.03909691	-1.126494040	-0.24827984
9	Powiat m.Biała Podlaska	-1.056527495	0.16077800	1.81492077	-0.87590603	0.548918574	-0.50945728
10	Powiat m.Chełm	-0.951443411	-0.59160245	1.34313295	-0.59680532	0.737087438	-0.51629474

Rysunek 2: Przykładowe przeskalowane dane.

Dane zostały podzielone na zbiór treningowy (54 danych) oraz testowy (5 danych).

	Powiat	Wynagrodzenie	Absolwenci	Bezrobotni	Inwestycje	KobietyPrac	PracownicyFUN
1	Powiat m.Skierniewice	-0.97269111	-0.7040266	-0.02473953	-0.9213848	0.6191624	-0.52428983
2	Powiat m.Ruda Śląska	-0.77594078	-1.1429185	-1.19858427	-0.9480997	-1.8669985	0.02595114
3	Powiat m.Siedlce	-0.30266924	0.9103565	0.07355852	-0.5620094	0.2300317	-0.56040521
4	Powiat m.Kielce	-0.08667886	1.3789904	0.45854259	-0.2487385	0.4227947	-0.02188012
5	Powiat m.Olsztyn	0.17334610	1.4428848	-0.64304897	0.1579366	0.4320342	0.17563947

Rysunek 3: Zbiór testowy.

BADANIE:

W tym rozdziale przedstawiony został kod w programie R przedstawiający ważniejsze obliczenia programu. Pozostała część kodu została załączona do projektu w postaci pliku programu R. Początkowo obliczono wartości alpha0, delta0, zapisano do wektora B0 ustalone w projekcie parametry a priori, do macierzy X

zmienne objaśniające oraz kolumnę jedynek dla stałej, a do macierzy E diagonalnie dodano ustalone odchylenia parametrów a priori przeskalowane przez sigma2 (wariancję danych). W dalszej części obliczono parametry dla rozkładu a posteriori: alpha1, delta1, oraz macierze parametrów i odchyleń B, E_1. Do obliczeń skorzystano ze wzorów przedstawionych w początkowej części projektu.

```
#apriori
alpha0<-nrow(data_train)-(ncol(data_train)-1)
delta0<-sum(reszty_mnk^2)
B0<-as.matrix(parametry_apriori)
Y<-as.matrix(data_train$Wynagrodzenie)
X<-as.matrix(data.frame("Stala"=rep(1,nrow(data_train)),data_train[,3:ncol(data_train)]))
E<-diag((odchylenia_apriori^2)/sigma2)

#aposteriori
E_1<-solve(t(X)%*%X+solve(E))
B<-E_1%*%(t(X)%*%Y+solve(E)%*%B0)
alpha1<-alpha0 + nrow(data_train)
delta1<-delta0+t(Y)%*%Y-t(B)%*%solve(E_1)%*%B+t(B0)%*%solve(E)%*%B0
```

Rysunek 3: Obliczanie parametrów rozkładów a priori i a posteriori

Kolejną istotną częścią programu jest funkcja t_multi, która oblicza gęstość dla rozkładu wielowymiarowego t studenta z zadaną wartością oczekiwaną i odchyleniem. Parametr n ma domyślną wartość 1 gdyż funkcję użyto w celu obliczenia gęstości rozkładu t studenta 1 wymiarowego.

```
#funkcja tstudent
t_multi<-function(mi,sigma,k,p,n=1){
  res<-c()
  i<-0
  for (ip in p) {
    i<-i+1
    licznik<-gamma((k+n)/2)
    mianownik<-((pi*k)^(n/2))*gamma(k/2)*abs(sigma)^(1/2)
    iloczyn<-(1+(1/k*((ip-mi)^2)*1/sigma))^(-(k+n)/2)
    res[i]<-(licznik/mianownik)*iloczyn
  }
  return(res)
}
```

Rysunek 4: Funkcja obliczająca gęstość dla rozkładu wielowymiarowego t studenta.

Poniższy kod przedstawia obliczenie przeskalowanej macierzy z wariancjami parametrów a posteriori. Została ona przemnożona przez delta1/alpha1. Następnie w pętli, korzystając z funkcji t_multi, obliczono funkcje gęstości wszystkich parametrów i przedstawiono je na wykresach.

```
SIGMA<-as.matrix((as.numeric(delta1)/alpha1)*E_1)
p = seq(-2,2, length=1000)
for (i in 1:6) {
  yt<-t_multi(B[i],SIGMA[i,i],alpha1,p)
  yn<-dnorm(p,parametry_apriori[i],odchylenia_apriori[i])
  y_df<-data.frame("y_apriori"=yn,"y_aposteriori"=yt)
  print(ggplot(y_df,aes(x=p,y=y_aposteriori))+
    geom_line(col=czapo[i])+
    geom_area(fill=czapo[i],alpha=0.1)+
    geom_line(aes(x=p,y=y_apriori),col=czapr[i])+
    geom_area(aes(x=p,y=y_apriori),fill=czapr[i],alpha=0.1)+
    labs(x = "x",y="density",title = paste0("Rozkład a priori
                  oraz a posteriori dla ", rownames(B)[i]))+
    theme_bw())
}
```

Rysunek 5: Obliczanie przeskalowanej macierzy sigma i generowanie rozkładów parametrów

Kolejny fragmentu kodu przedstawia funkcję obliczającą HPDI (highest posterior density interval), która na podstawie funkcji gęstości zwraca przedziały w których występuje największa liczba danych (w projekcie skorzystano z przedziału 95%). Funkcja została wykorzystana w obliczaniu przedziałów dla parametrów oraz prognoz.

```
#funkcja hpdi
hpdi = function(x, x.density, coverage){
  best = 0
  for (ai in 1 : (length(x) - 1))
  {for (bi in (ai + 1) : length(x))
  {mass = sum(diff(x[ai : bi]) * x.density[(ai + 1) : bi])
  if (mass >= coverage && mass / (x[bi] - x[ai]) > best)
  {best = mass / (x[bi] - x[ai])
  ai.best = ai
  bi.best = bi}}}
  c(x[ai.best], x[bi.best])
}
```

Rysunek 6: Funkcja obliczająca przedziały HPDI

Następny fragment kodu służy do obliczenia prognoz. Dzięki funkcji t_multi obliczono rozkład predykcyjny. Na wykresach wyświetlono rozkłady predykcyjne zmiennych oraz obliczone przedziały HPDI. Warto zwrócić uwagę, że obliczenia te były przeprowadzone na zbiorze testowym.

```
#predykcje
data_pred_m<-as.vector(predict(model,data_test))
m_interval<-predict(model,data_test,interval="prediction")[,2:3]
data_pred_c<-as.vector(unlist(as.matrix(data.frame("stala"=rep(1,5),data_test[3:7]))%*%as.matrix(B)))

Xtau<-as.matrix(data.frame("stala"=rep(1,5),data_test[3:7]))
I<-as.matrix((as.numeric(delta1)/alpha1)*diag(rep(1,nrow(Xtau))))

p = seq(-2,2, length=1000)
hpdi_interval<-list()
for (i in 1:nrow(data_test)) {
  r<-t_multi((Xtau%*%B)[i],(I+Xtau%*%SIGMA%*%t(Xtau))[i,i],alpha1,p)
  hpdi_interval[[i]]<-hpdi(p,r,0.95)
  y_df<-data.frame("y_aposteriori"=r)
  print(ggplot(y_df,aes(x=p,y=y_aposteriori))+
          geom_line(col="darkturquoise")+
          geom_area(fill="darkturquoise",alpha=0.1)+
          geom_vline(aes(xintercept=hpdi_interval[[i]][1]),
                     color=chpdi, linetype="longdash", size=1)+
          geom_vline(aes(xintercept=hpdi_interval[[i]][2]),
                     color=chpdi, linetype="longdash", size=1)+
          geom_vline(aes(xintercept=m_interval[i,][1]),
                     color=cmint, linetype="dotted", size=1)+
          geom_vline(aes(xintercept=m_interval[i,][2]),
                     color=cmint, linetype="dotted", size=1)+
          labs(x = "x",y="density",title = paste0("Rozkład predykcyjny z przedziałami ufności i HPDI
                                                  dla obserwacji:\n", data_test$Powiat[i]))+
          theme_bw())
}
```

Rysunek 7: Predykcja wartości na podstawie regresji wielorakiej

Kolejnym fragmentem wartym uwagi jest porównanie 2 modeli: z restrykcjami i bez. Iloraz szans obliczono korzystając ze wzorów przedstawionych w początkowej części projektu. Założono, że pewność modelu bez restrykcji wynosi 0,4, a modelu z restrykcjami 0,6.

```
n<-nrow(data_train)
#czynniki bayesa
PyM<-(1/pi^(n/2))*
  (sqrt(abs(det(E_1)/det(E))))*
  (delta0^(alpha0/2)/delta1^(alpha1/2))*
  (gamma(alpha1/2)/gamma(alpha0/2))

PyMr<-(1/pi^(n/2))*
  (sqrt(abs(det(E_12)/det(E2))))*
  (delta02^(alpha02/2)/delta12^(alpha12/2))*
  (gamma(alpha12/2)/gamma(alpha02/2))

PM<-0.4
PMr<-0.6

iloraz_szans<-(PyM/PyMr)*(PM/PMr)
log_iloraz<-log10(iloraz_szans)
```

Rysunek 8: Porównanie modeli bez restrykcji i z restrykcjami

Ostatnia istotna część programu przedstawia algorytm Gibbsa. W pętli losujemy tysiąc wartości z rozkładu normalnego i odwrotnego gamma o zadanych parametrach obliczonych na podstawie wzorów zmieniających się w każdym przejściu pętli. Losowanie z rozkładu normalnego następuje dla każdej zmiennej.

```
XX<-t(X)%*%X
XY<-t(X)%*%Y
B1_g<-data.frame("Stala"=c(NA),"Absolwenci"=c(NA),"Bezrobotni"=c(NA),
                 "Inwestycje"=c(NA),"KobietyPrac"=c(NA),"PracownicyFUN"=c(NA))
sigma2_g<-c(sigma2)
m<-1000

for (i in 2:m) {
  E_1i<-solve(1/sigma2_g[i-1]*XX+solve(E))
  B1i<-E_1i%*%(1/sigma2_g[i-1]*XY+solve(E)%*%B0)
  delta1i<-delta0+t(Y-X%*%B1i)%*%(Y-X%*%B1i)
  for (j in 1:length(B)) {
    B1_g[i,j]<-rnorm(1,B1i[j],E_1i[j,j])
  }
  sigma2_g[i]<-rinvgamma(1, alpha1/2, delta1i/2)
}
B1_g<-na.omit(B1_g)
sigma2_g<-na.omit(sigma2_g)
```

Rysunek 9: Obliczanie rozkładów a posteriori wg. Algorytmu Gibbsa

---

WYNIKI:

Początkowo należy dobrać odpowiednie parametry rozkładu a priori dla stałej oraz zmiennych:

- stała: $EX = 0\ , \sigma\ = 1$
- absolwenci: $EX = 0,25\ , \sigma\ = 0,1$
- bezrobotni: $EX = -0,15\ , \sigma\ = 0,15$
- inwestycje: $EX = 0,65\ , \sigma\ = 0,25$
- kobietyPract: $EX = -0,15\ , \sigma\ = 0,05$
- pracownicyFUN: $EX = 0,40\ , \sigma\ = 0,5$

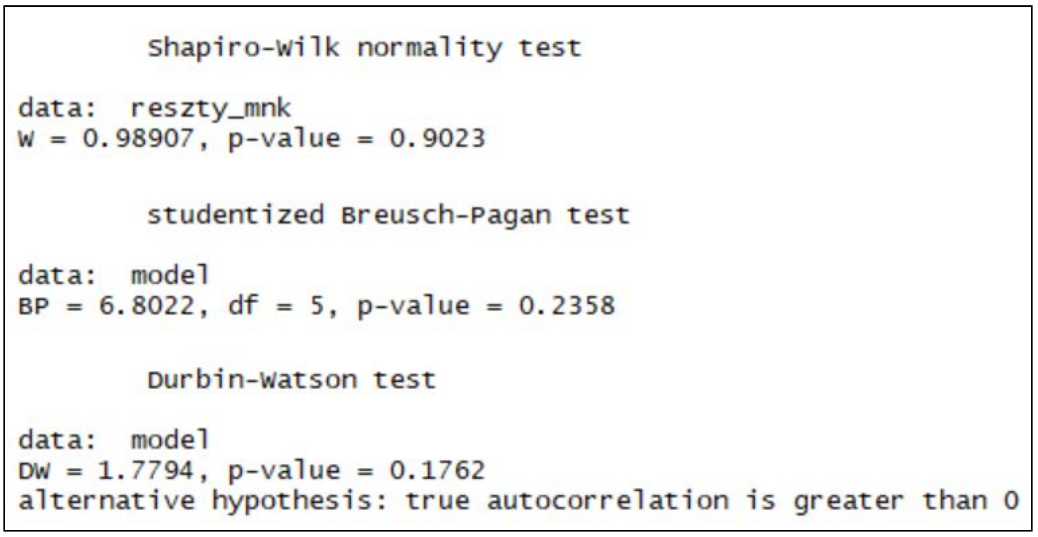
Dobór parametrów przeprowadzono na podstawie przeczytanych artykułów o wpływie różnych zmiennych na przeciętne wynagrodzenie oraz własnych przypuszczeń. W ten sposób uszeregowano zmienne, od tych przypuszczalnie najbardziej istotnych do tych najmniej istotnych i przydzielono im wartości. W modelu a priori wnioskowano, że inwestycje mają największy wpływ, a bezrobocie, stosunek pracujących kobiet oraz stała - najmniejszy. Odchylenie standardowe szacowano na podstawie wiedzy o parametrach, duże odchylenie świadczy o małej wiedzy na

temat wartości parametru. Najmniejsze odchylenie ma zmienna przedstawiająca stosunek kobiet pracujących oraz liczbę absolwentów, a największe ma stała.

```
Call:
lm(formula = Wynagrodzenie ~ Absolwenci + Bezrobotni + Inwestycje +
    KobietyPrac + PracownicyFUN, data = data_train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07674 -0.38808 -0.02781  0.33979  1.26302

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.01473    0.07668   0.192    0.848
Absolwenci    0.12346    0.09287   1.329    0.190
Bezrobotni   -0.11212    0.07802  -1.437    0.157
Inwestycje    0.44470    0.09366   4.748 1.89e-05 ***
KobietyPrac  -0.02345    0.09282  -0.253    0.802
PracownicyFUN 0.47922    0.10140   4.726 2.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5616 on 48 degrees of freedom
Multiple R-squared:  0.7309,    Adjusted R-squared:  0.7028
F-statistic: 26.07 on 5 and 48 DF,  p-value: 1.27e-12
```

Rysunek 10: Model MNK.

Następnym krokiem jest oszacowanie parametrów modelu metodą MNK. Stała ma wartość 0,01, a więc przypuszczenie o jej niskiej wartości były trafne. Największy wpływ na przeciętne wynagrodzenie ma zmienna PracownicyFUN z parametrem 0,48 oraz zmienna Inwestycje z parametrem 0,44. Zmienne te mają istotny wpływ na zmienną objaśnianą - potwierdzają to niskie wartości p-value dla testu t studenta. Średni wpływ na zmienną objaśnianą ma bezrobocie -0,11 oraz absolwenci 0,12 - zmienne te są nieistotne według testu t studenta. Najmniejszą istotność wśród dobranych zmiennych ma stosunek kobiet pracujących, parametr wynosi -0,02. Warto zwrócić uwagę, że znaki wszystkich parametrów są zgodne z oszacowaniami a priori.



Rysunek 11: Testowanie reszt modelu MNK.

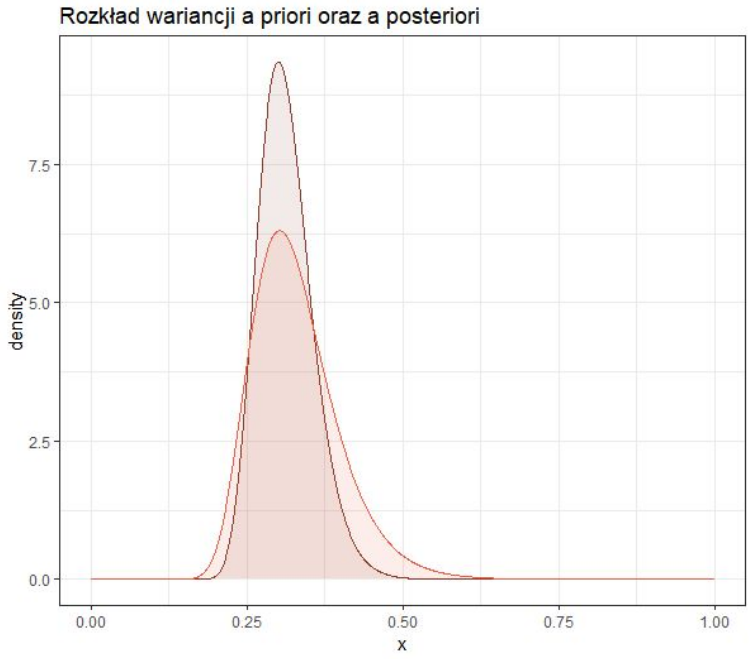
Znając reszty modelu MNK możemy zweryfikować heteroskedastyczność, normalność, oraz występowanie autokorelacji reszt. Do weryfikacji normalności reszt wykorzystano test Shapiro-Wilka, p-value tego testu jest większe niż przyjęty poziom istotności, nie mam podstaw do odrzucenia hipotezy o normalności reszt. W celu weryfikacji heteroskedastyczności wykorzystano test Breuscha-Pagana. P-value jest większe niż przyjęty poziom istotności, a więc nie ma podstaw do odrzucenia hipotezy zerowej - składnik losowy jest homoskedastyczny. Do zbadania autokorelacji wykorzystano test Durbina-Watsona, p-value wynosi więcej niż przyjęty poziom istotności, nie ma podstaw do odrzucenia hipotezy zerowej o braku autokorelacji reszt.

	a priori model I	**a posteriori model I**	**MNK model I**	**a priori model II**	**a posteriori model II**
α	48	102	-	50	104
δ	15.14	31.31	-	15.80	32.58
β	0.00 0.25 -0.15 0.65 -0.15 0.40	0.0181 0.1733 -0.1166 0.4153 -0.1135 0.4811	0.0147 0.1234 -0.1121 0.4447 -0.0234 0.4792	 0.25 0.65 -0.15 0.4	 0.1694 0.4193 -0.1157 0.5112

Tabela 1: Parametry modelu a priori i a posteriori (model zwykły oraz z restrykcjami).

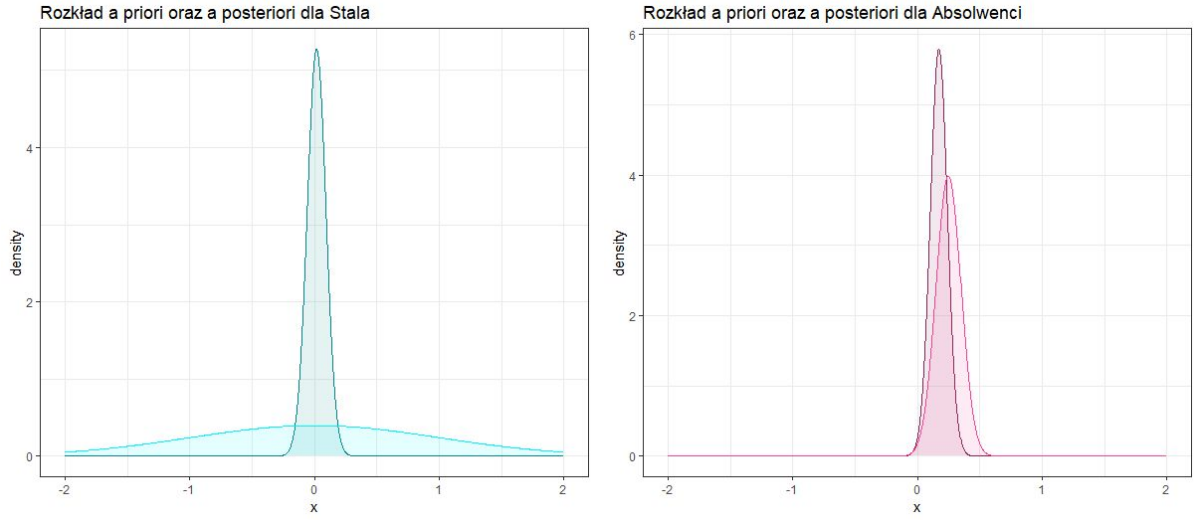
W tabeli nr 1 znajduje się porównanie parametrów a priori i a posteriori (w tym momencie istotne są parametry modelu I). Jak widzimy wartości alfa i delta znacznie wzrosły w rozkładzie a posteriori. Warto skupić się jednak na interpretacji parametrów. Większość z nich została dobrze oszacowana, dla prawie wszystkich zmiennych różnica oszacowań jest poniżej 0,1. Jedynie parametr dla zmiennej inwestycje został znacznie skorygowany, różnica oszacowań to ponad 0,23.

Porównując parametry rozkładu a posteriori z oszacowaniami modelu MNK możemy zauważyć nieznaczne różnice. Maksymalne różnice wynoszą 0,09 dla zmiennej kobietyPrac oraz 0,05 dla zmiennej absolwenci. Pozostałe parametry zostały oszacowane bardzo podobnie obiema metodami.

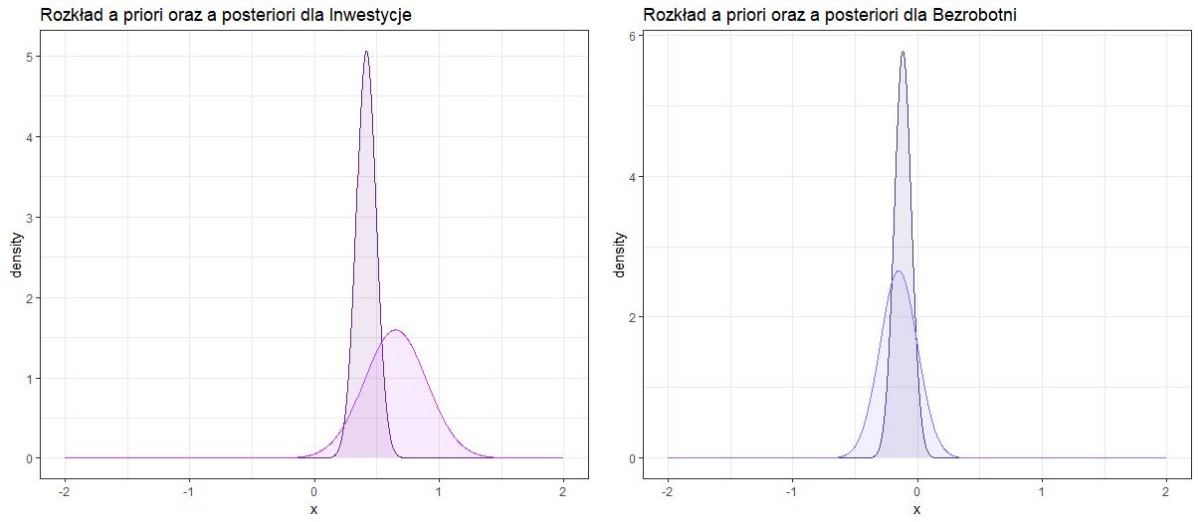


Wykres 1: Wykres rozkładu modelu a priori oraz a posteriori dla σ^2.

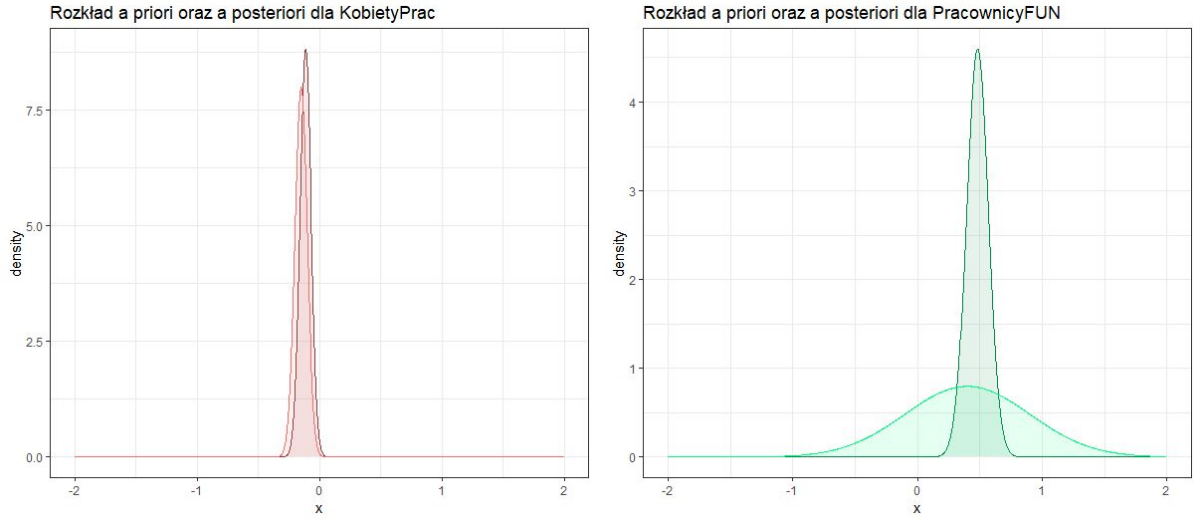
Na wykresie nr 1 można zauważyć porównanie rozkładu a priori (kolor jaśniejszy) oraz a posteriori (kolor ciemniejszy). Rozkład a posteriori *dla* σ^2 znacznie zwiększył swoją wysokość, co świadczy o większej pewności w oszacowaniu parametrów. Posiada również mniejsze ogony (zwłaszcza po prawej stronie) niż rozkład a priori.



Wykres 2: Wykres rozkładu modelu a priori oraz dla parametrów: stała, absolwenci.



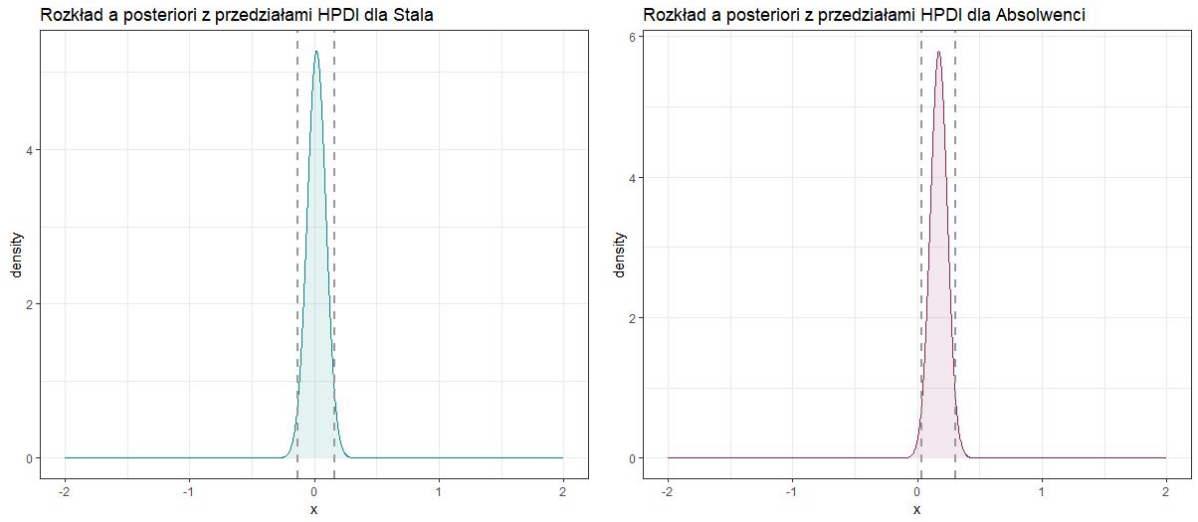
Wykres 3: Wykres rozkładu modelu a priori oraz a posteriori dla parametrów: bezrobotni, inwestycje.



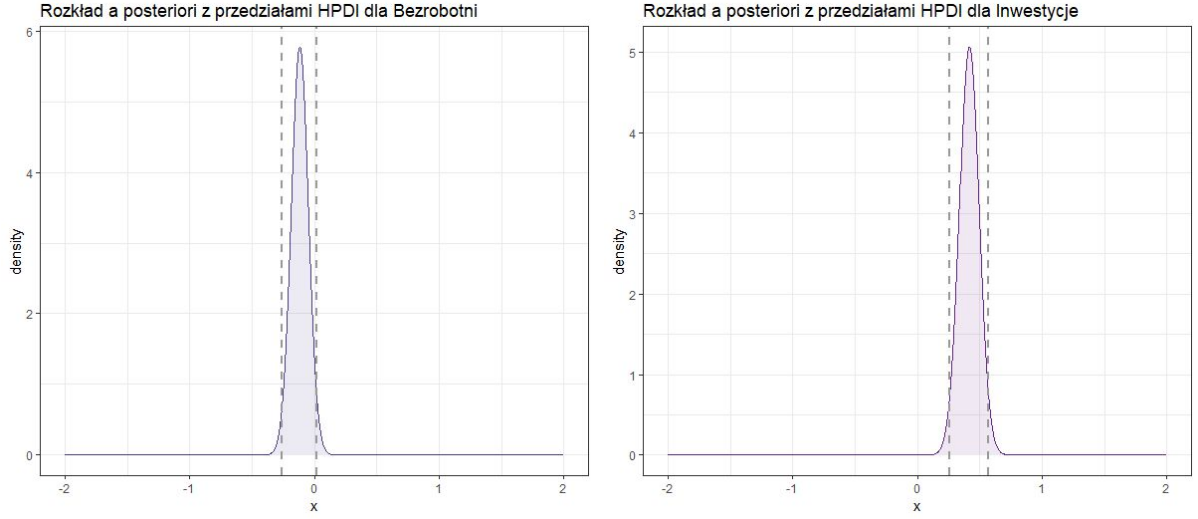
Wykres 4: Wykres rozkładu modelu a priori oraz a posteriori dla parametrów: kobietyPrac, pracownicyFUN.

Powyżej znajdują się wykresy a priori (jaśniejsze) i a posteriori (ciemniejsze) dla każdej zmiennej. Wykres dla stałej (a priori) jest mocno spłaszczony ze względu na małą wiedzę o parametrze. Wartość oczekiwana zmiennej w obu rozkładach jest bliska zeru. Dla zmiennej absolwenci, rozkład a posteriori jest nieznacznie przesunięty w prawo. Tak jak dla wszystkich zmiennych, rozkład a posteriori jest znacznie wyższy. Warto zwrócić uwagę na zmienną inwestycje, gdzie wykresy obu modeli znacznie się różnią, a posteriori wskazuje na mniejszy wpływ inwestycji na wynagrodzenia. Kolejnym ciekawym elementem jest zmienna kobietyPrac, dla której

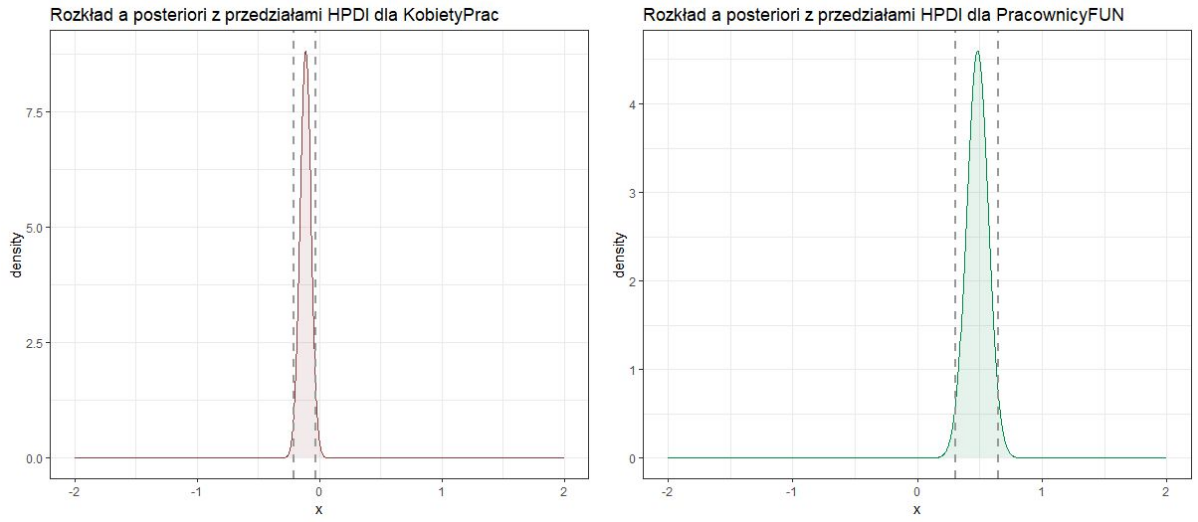
rozkłady prawie się pokrywają, a więc początkowe oszacowanie było bardzo dokładne. Dla zmiennej pracownicyFUN wartość oczekiwana dla obu rozkładów jest podobna, jednak rozkład a priori posiada niską pewność.



Wykres 5: Wykres HPDI dla parametrów: stała, absolwenci.



Wykres 6: Wykres HPDI dla zmiennych: bezrobotni, inwestycje.



Wykres 7: Wykres HPDI dla zmiennych: kobietyPrac, pracownicyFUN.

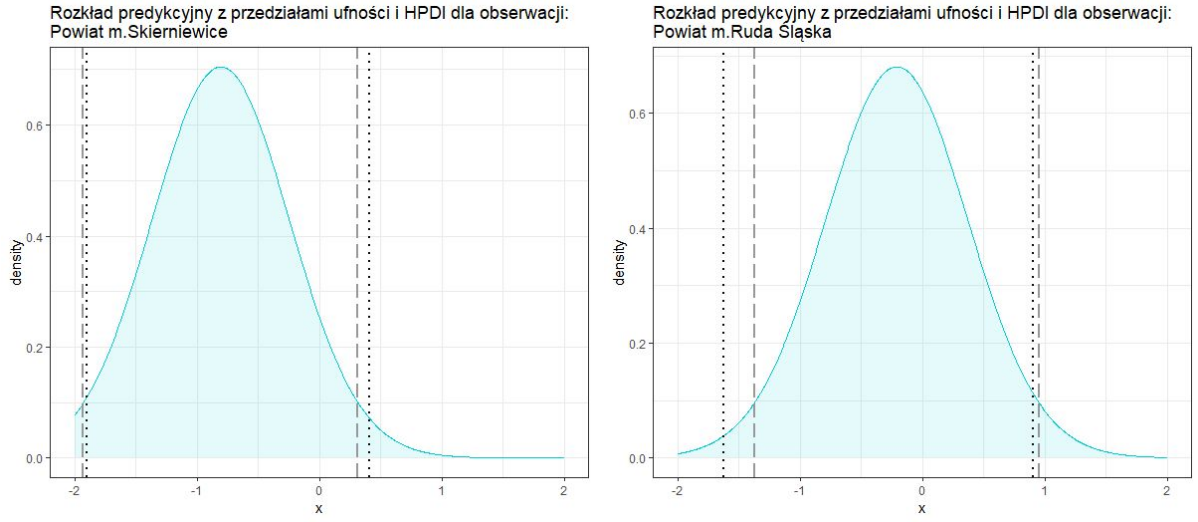
Na powyższych wykresach zostały przedstawione przedziały HPDI dla rozkładów a posteriori wszystkich zmiennych. Na podstawie przedziałów możemy wywnioskować, że stała nie jest istotna. Blisko zera znajdują się również zmienna bezrobotni (przedział HPDI obejmuje wartość zero). Dużą istotność mają zmienne: inwestycje, pracownicyFUN.

Powiat	Wynagrodzenie rzeczywiste	Wynagrodzenie prognoza MNK	Wynagrodzenie prognoza Bayes
Powiat m.Skierniewice	-0.9727	-0.7449	-0.8062
Powiat m.Ruda Śląska	-0.7759	-0.3574	-0.2097
Powiat m.Siedlce	-0.3027	-0.4050	-0.3618
Powiat m.Kielce	-0.0867	0.0025	0.0419
Powiat m.Olsztyn	0.1733	0.4092	0.4442

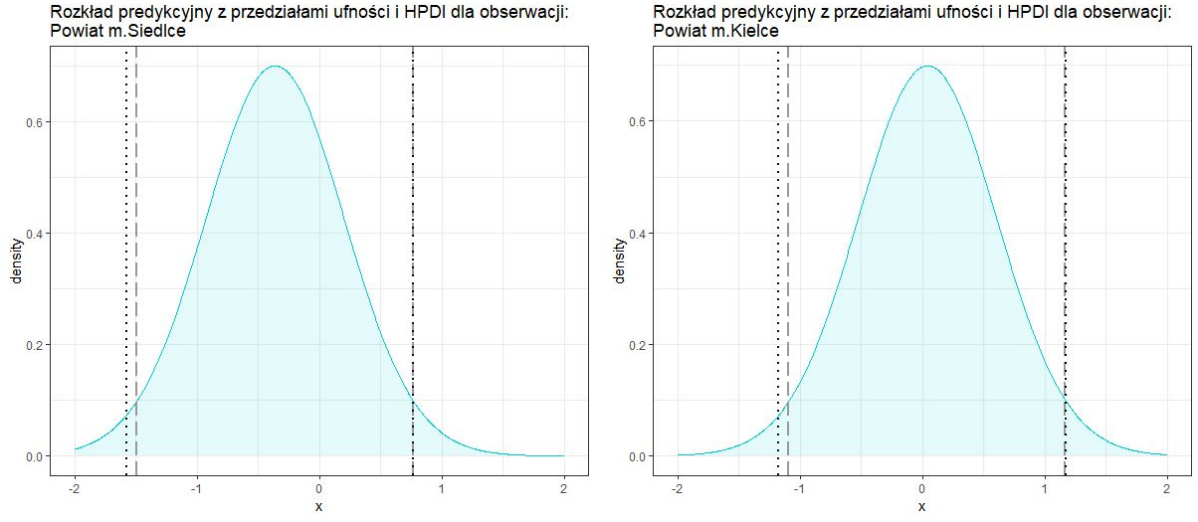
Tabela 2: Porównanie wartości rzeczywistych z oszacowanymi.

Kolejnym etapem jest przeprowadzenie prognoz na zbiorze testowym. Zbiór testowy zawiera dane odnośnie 5 miast na prawie powiatów. W tabeli nr 2 zawarto

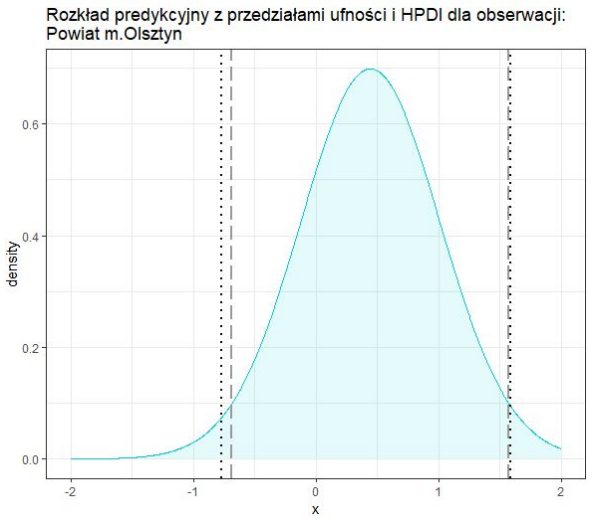
porównanie wartości rzeczywistych dla zmiennej objaśnianej oraz predykcji metodą MNK oraz Bayesa. Średnie różnice między predykcjami wynoszą około 0,05, tylko dla Rudy Śląskiej różnica ta jest większa i wynosi około 0,15. Porównując wartości rzeczywiste z wartościami prognozowanymi widzimy, że rezultaty nie są zbyt dobre. Średnia różnica wynosi 0,22 dla MNK oraz 0,24 dla Bayesa, a więc klasyczne podejście sprawdziło się w tym przypadku lepiej. Największe różnice w oszacowaniach wystąpiły dla Rudy Śląskiej. Metoda Bayesa pozwoliła na lepsze oszacowanie wynagrodzeń dla miast: Skierniewice, Siedlce, MNK dla pozostałych.



Wykres 8: Wykresy HPDI dla prognozy: m. Skierniewice, m. Ruda Śląska



Wykres 9: Wykres HPDI dla prognozy: m. Siedlce, m. Kielce



Wykres 10: Wykres HPDI dla prognozy: m. Olsztyn.

Na powyższych wykresach przedstawiono rozkłady predykcyjne. Pionowymi liniami zaznaczono przedziały dla HPDI (----) oraz przedziały ufności MNK (....). Przedziały te są do siebie bardzo podobne, i w niektórych miejscach pokrywają się. Przedziały dla MNK są szersze. Największe różnice możemy zauważyć dla m. Ruda Śląska.

Nazwa	**Wartość**
PyM	3.514739e-23
PyMr	4.194419e-22
PM	0,4
PMr	0,6
Iloraz szans	0.05586375
Log10 Ilorazu szans	-1.25287

Tabela 3: Porównanie modelu bez restrykcji i z restrykcjami.

Kolejnym etapem projektu jest porównanie modelu z restrykcjami i bez restrykcji. W związku z faktem, że stała oraz zmienna “bezrobotni” wskazują na niską istotność,

zastosowano restrykcje, że zmienne te są nieistotne (mają parametr równy zero). W tabeli nr 1 znajdują się współczynniki dla obu modeli. Współczynniki nie różnią się zbytnio. Mając 2 modele, obliczono dla nich iloraz szans. Logarytm dziesiętny ilorazu szans wynosi -1,25. Ma on wartość negatywną, a więc wspiera model z restrykcjami.

Model bez restrykcji

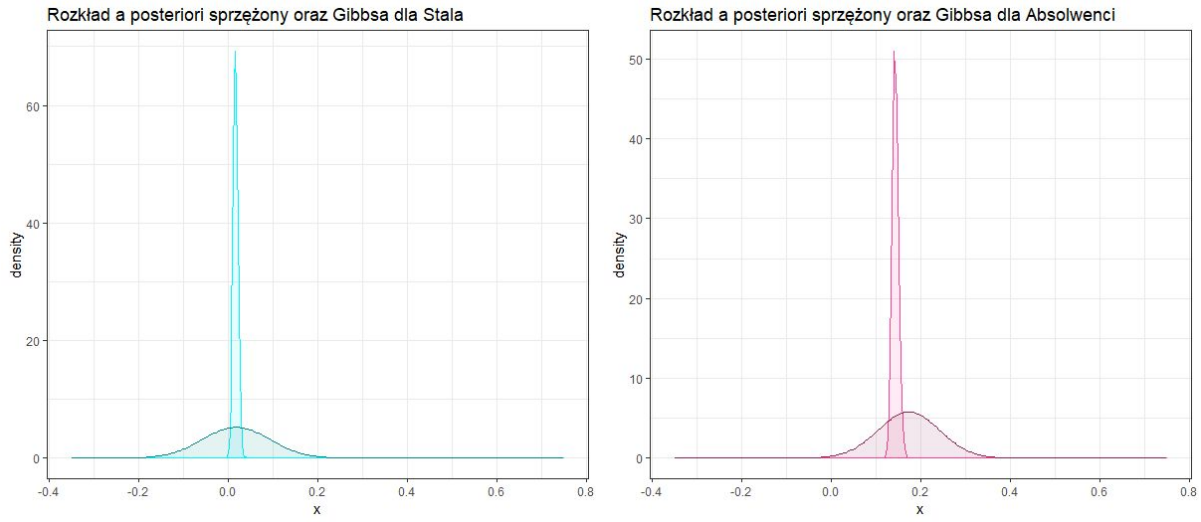
```
F-statistic: 26.07 on 5 and 48 DF,  p-value: 1.27e-12
```

Model z restrykcjami:

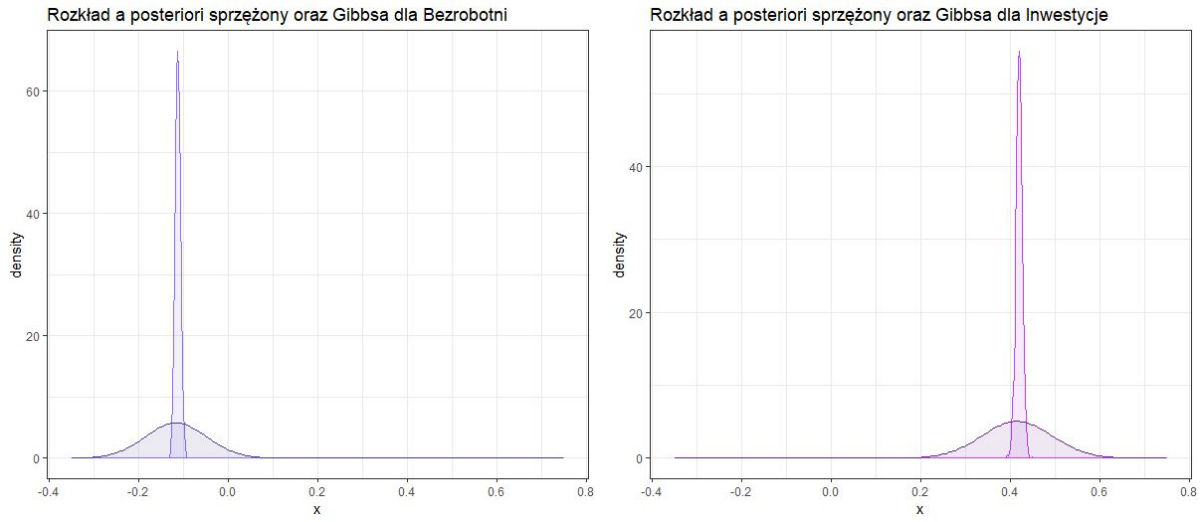
```
F-statistic: 32.07 on 4 and 50 DF,  p-value: 2.989e-13
```

Rysunek 12: Testowanie łącznej istotności zmiennych.

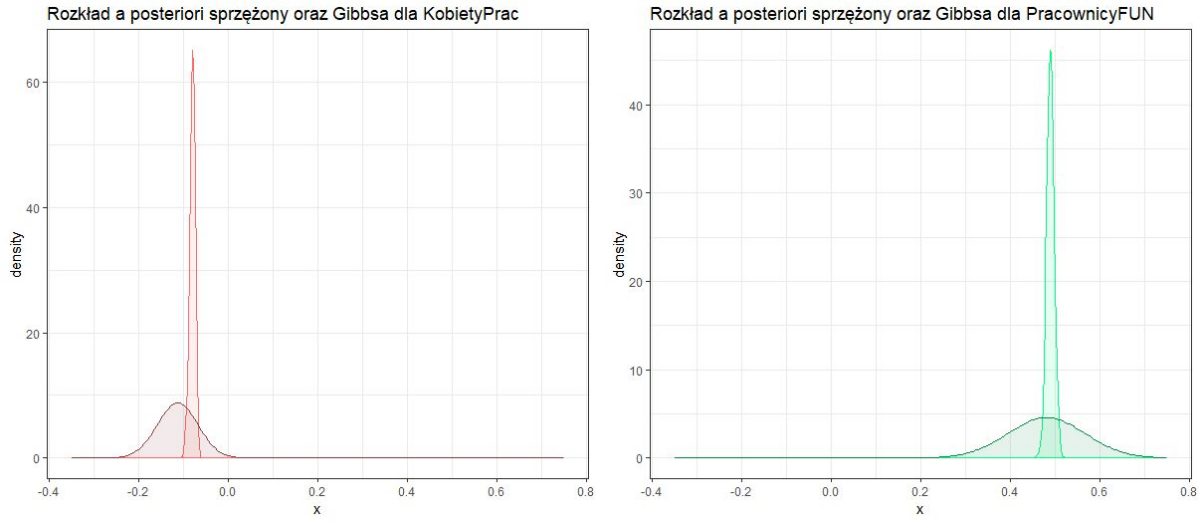
Dla porównania wykonano test F dla modelu z restrykcjami oraz bez restrykcji. Wartości p-value dla obu modeli są niskie, co pokazuje, że parametry są łącznie istotne w każdym z modeli. Wartość p-value dla modelu z restrykcjami jest jednak mniejsza, co potwierdza, że model z restrykcjami jest lepszy.



Wykres 11: Wykres rozkładu modeli a posteriori dla parametrów: stała, absolwenci.



Wykres 12: Wykres rozkładu modeli a posteriori dla parametrów dla parametrów: bezrobotni, inwestycje.



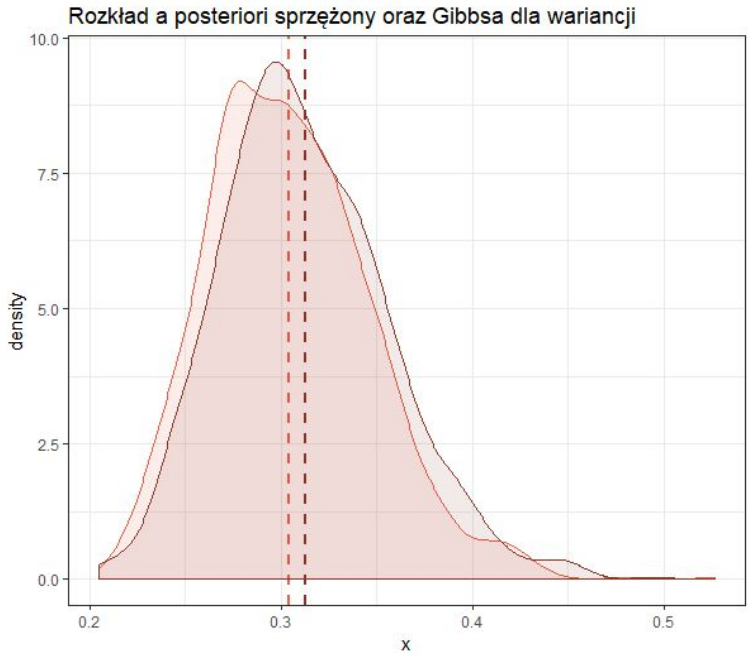
Wykres 13: Wykres rozkładu modeli a posteriori dla parametrów: kobietyPrac, pracownicyFUN.

Powyżej znajdują się wykresy przedstawiające gęstość rozkładów sprzężonych oraz rozkładów uzyskanych algorytmem Gibbsa. Możemy zauważyć, że dla każdej zmiennej wykresy na podstawie algorytmu Gibbsa są znacznie wyższe, co wskazuje na większą dokładność oszacowań tą metodą. Dla zmiennych absolwenci oraz kobietyPrac widzimy, że rozkłady są przesunięte względem siebie. W celu dokładniejszej interpretacji przyjrzyjmy się poniższej tabelce z dokładnymi wartościami oczekiwanymi oraz odchyleniami.

Parametr	Wartość oczekiwana		Odchylenie standardowe	
	Rozkład sprzężony	Algorytm Gibbsa	Rozkład sprzężony	Algorytm Gibbsa
Stała	0,0181	0,0163	0,0754	0,0056
Absolwenci	0,1733	0,1428	0,0686	0,0073
Bezrobotni	-0,1166	-0,1125	0,0689	0,0056
Inwestycje	0,4153	0,4196	0,0785	0,0072
KobietyPrac	-0,1135	-0,0807	0,0451	0,0060
PracownicyFUN	0,4811	0,4894	0,0865	0,0086

Tabela 4: Wartości oczekiwane oraz odchylenia dla rozkładów a posteriori oszacowanych algorytmem Gibbsa

Różnice między wartościami oczekiwanymi nie są duże. Najbardziej różnią się wartości dla zmiennych: absolwenci oraz kobietyPrac. Różnice te wynoszą kolejno 0,03 i 0,02. Analizując odchylenia możemy potwierdzić, że błąd oszacowań dla algorytmu Gibbsa jest znacznie mniejszy (około 10 krotnie), a więc metoda ta sprawdza się lepiej od standardowej.



Wykres 13: Wykres rozkładu modeli a posteriori dla σ^2.

Analizując sprzężony rozkład wariancji oraz rozkład otrzymany algorytmem Gibbsa widzimy, że są one bardzo podobne. Druga metoda pozwoliła na osiągnięcie nieznacznie mniejszej wartości oczekiwanej wariancji. Wartości zmiennej objaśnianej, obliczone przy użyciu parametrów oszacowanych drugą metodą, są dokładniejsze.

WNIOSKI:

Przeprowadzone badanie dostarcza istotnych informacji zarówno na temat wpływu wybranych parametrów na wysokość wynagrodzenia, jak również na temat różnic w ich estymacji przy wykorzystaniu różnych metod.

Po pierwsze, możemy stwierdzić jednoznacznie, że średnia kwota wynagrodzenie w danym mieście na prawach powiatu determinowana jest głównie przez inwestycje prowadzone przez przedsiębiorstwa, a także sektor, w jakim zatrudnieni są pracownicy. Zmienne, opisujące te wartości, obrazują stopień rozwinięcia miejsc

pracy w danym mieście, co bezpośrednio przekłada się na zarobki pracowników. Pozostałe zmienne, takie jak: ilość absolwentów uczelni wyższych w ostatnich latach oraz stosunek kobiet pracujących do wszystkich, mają mniejszy wpływ na wynagrodzenie. Oznacza to, że mimo popularnym poglądom o niższych zarobkach kobiet oraz ludzi mniej wykształconych, te cechy nie decydują tak mocno o wysokości średniego wynagrodzenia w miastach.

Warto zauważyć, że w ogóle nie istotne wydają się być bezrobocie w danym mieście oraz stała. Należy więc interpretować, że kondycja gospodarcza danego regionu, reprezentowana przez ilość osób niepracujących, nie ma tak dużego wpływu na zarobki pozostałej części.

Podsumowując wybrane metody estymacji parametrów modelu można zauważyć, że są one w dużym stopniu zbliżone do siebie. Oszacowania poszczególnych zmiennych mają podobne wartości, mimo usunięcia części z nich. Różnice możemy odnaleźć natomiast w błędach estymacji. Zastosowanie algorytmu Gibbsa wydaje się być pewniejszym rozwiązaniem, pozwalającym na uzyskanie większej pewności wyników. Jego użycie jest najwłaściwszym posunięciem, podobnie jak wszystkich metod Bayesowskich, w przypadku posiadania informacji z innych źródeł, np. poprzednich badań na dany temat.