PD2 Daniel Ponikowski

Daniel Ponikowski

22 marca 2019

Wybrane zmienne :

+ ppwork - aktualny status zatrudnienia
+ w6_q20 - czy obecnie mieszkasz z partnerem?
+ Q21A_Year - w ktorym roku pierwszy raz spotkałes partnera?
+ ppage - wiek

Odtworzenie modelu

```
df <- data[,c("S1","ppwork","w6_q19","Q21A_Year","ppage")]


df <- df %>% mutate(Q21A_Year = as.numeric(as.character(Q21A_Year))
                    ,ppwork = factor(ppwork)
                    ,w6_q19 = factor(w6_q19)
                    ,ppage = as.numeric(ppage)
                    ,S1= factor(S1)) %>%
  na.omit() %>% unique() %>% as.data.frame()
```

```
## Warning in evalq(as.numeric(as.character(Q21A_Year)), <environment>):
## pojawiły się wartości NA na skutek przekształcenia
```

```
control <- trainControl(method = "cv", number=10, search = "random")
metric <- "Accuracy"
RF <- train(df[2:5],df$S1 ,method = "rf", metric = metric,
                  trControl = control)
```

Zmienna ppage

```
dane <- df
liczba_grup <- 50

l_wiersz <- dim(dane)[1]
liczn_grup <- floor(l_wiersz/liczba_grup)
grupy <- c(sort(rep(1:liczba_grup,liczn_grup)),
           rep(liczba_grup+1,l_wiersz-liczn_grup*liczba_grup))


dane <- dane %>% arrange(ppage) %>% mutate(grupa = grupy)

granice <- dane %>% group_by(grupa) %>% summarise(minimum = min(ppage)) %>%
  select(minimum) %>%
   add_row(minimum = max(dane[["ppage"]]))

srodki <- cbind(granice,rbind(0,granice[1:(liczba_grup+1),])) %>%
  apply(MARGIN = 1,FUN = mean)
granice <- unlist(granice) %>% unname()
```

```
dane_low <- dane %>% mutate(ppage = granice[grupy])
dane_up <- dane %>% mutate(ppage = granice[grupy+1])


low_pred <- predict(object = RF,dane_low,type = "prob")[,1]
up_pred <- predict(object = RF, dane_up ,type = "prob")[,1]

result <- dane %>% mutate(roznica = up_pred - low_pred) %>%
  group_by(grupa) %>%
    summarise(srednia = mean(roznica)) %>% select(srednia) %>%
  mutate(srodki = srodki[2:length(srodki)])

ggplot(result,aes(srodki,srednia))+geom_point()+geom_line()+
  ggtitle("ALE Plot zmienna ppage")
```

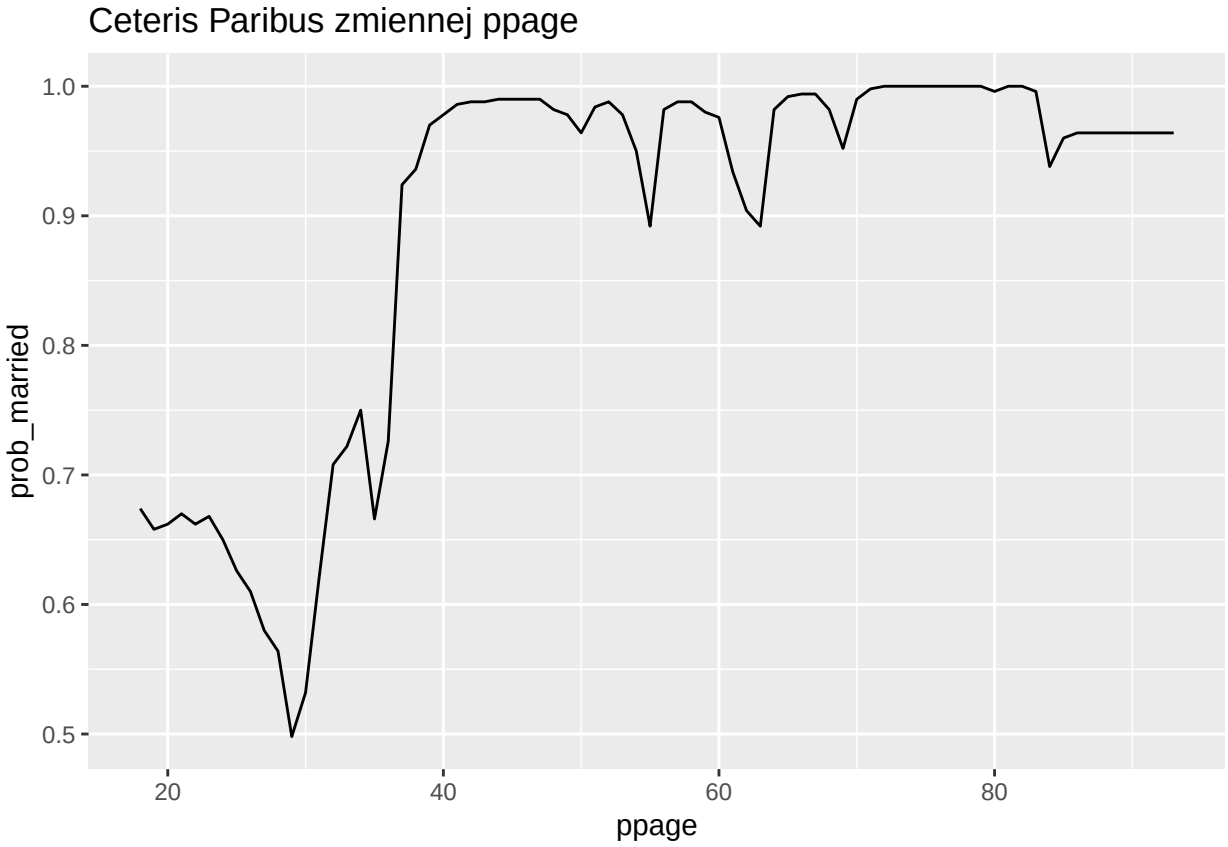
Osoba dla której bedziemy rysowac wykresy Ceteris Paribus

```
(os <- df[sample(1:nrow(df),size = 1),])
```

```
##                      S1                        ppwork w6_q19 Q21A_Year ppage
## 1489 Yes, I am Married Working - as a paid employee    Yes      1966    72
```

```
age <- min(df$ppage):max(df$ppage)
df_ppage <- data.frame(ppwork = rep(os$ppwork,length(age))
                       ,w6_q19 = rep(os$w6_q19,length(age))
                       ,Q21A_Year = rep(os$Q21A_Year,length(age))
                       ,ppage = age )
```

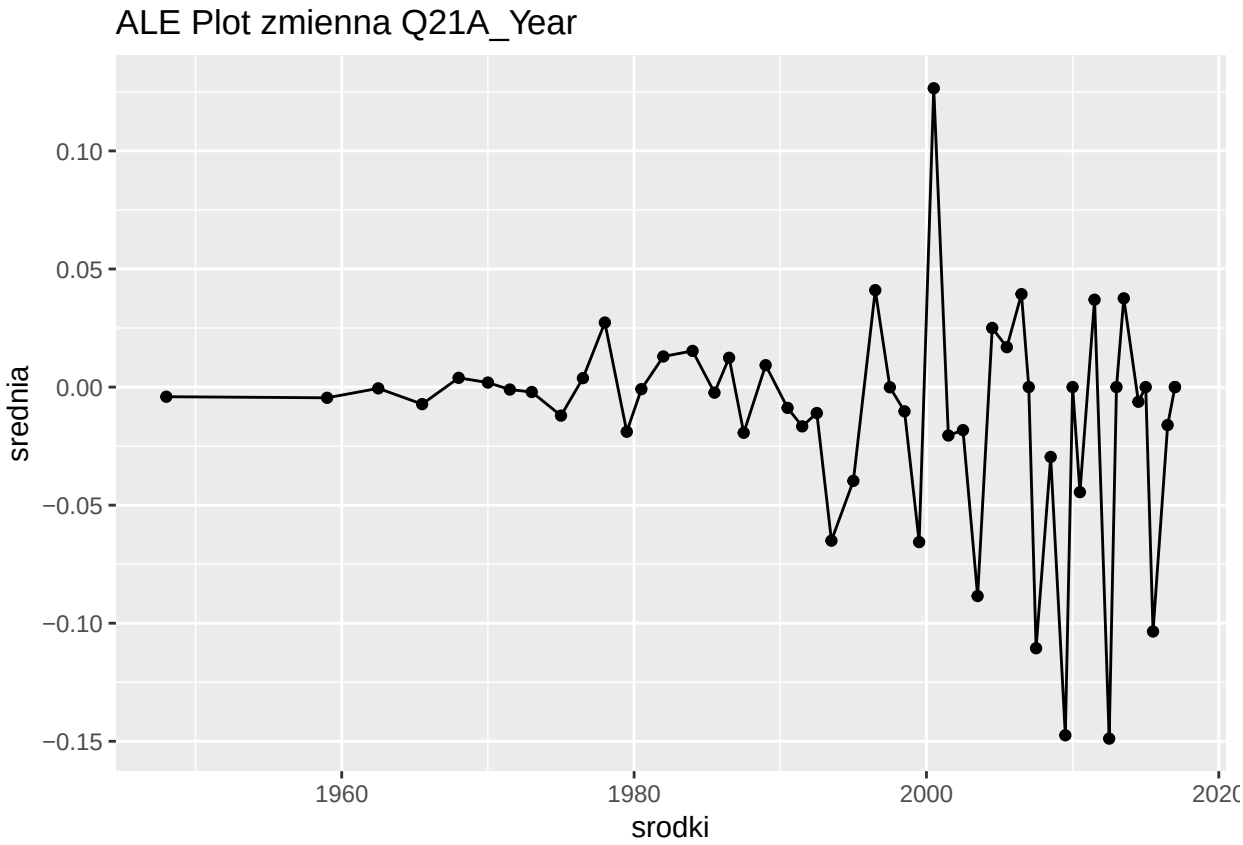
```
df_ppage$prob_married <- predict(RF,df_ppage,type = "prob")[,1]

ggplot(df_ppage,aes(x = ppage,y = prob_married)) + geom_line() +
  ggtitle("Ceteris Paribus zmiennej ppage")
```



Zmienna Q21A_Year

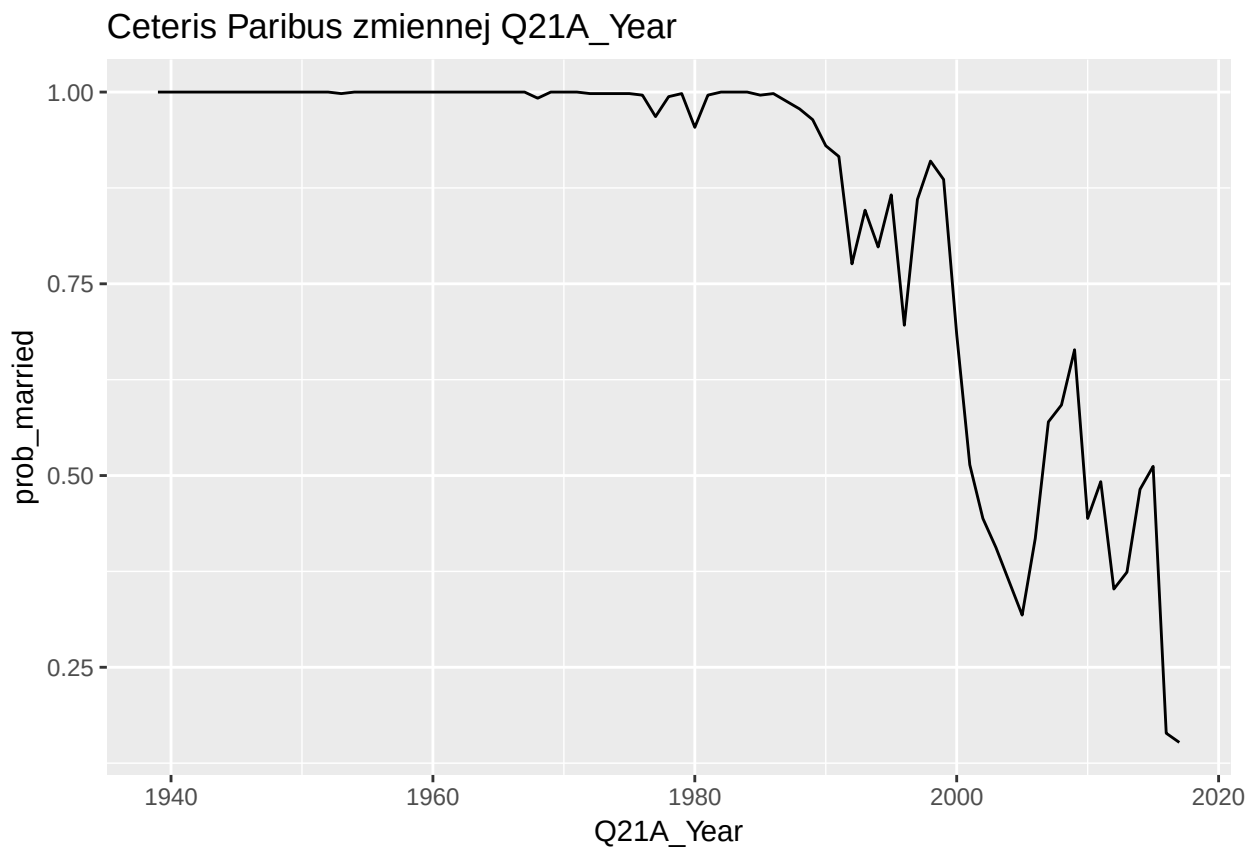
```
dane <- df
liczba_grup <- 50
zmienna <- "Q21A_Year"

l_wiersz <- dim(dane)[1]
liczn_grup <- floor(l_wiersz/liczba_grup)
grupy <- c(sort(rep(1:liczba_grup,liczn_grup)),
           rep(liczba_grup+1,l_wiersz-liczn_grup*liczba_grup))


dane <- dane %>% arrange(Q21A_Year) %>% mutate(grupa = grupy)

granice <- dane %>% group_by(grupa) %>% summarise(minimum = min(Q21A_Year)) %>%
  select(minimum) %>%
   add_row(minimum = max(dane[["Q21A_Year"]]))

srodki <- cbind(granice,rbind(0,granice[1:(liczba_grup+1),])) %>%
  apply(MARGIN = 1,FUN = mean)
```

```r
granice <- unlist(granice) %>% unname()

dane_low <- dane %>% mutate(Q21A_Year = granice[grupy])
dane_up <- dane %>% mutate(Q21A_Year = granice[grupy+1])


low_pred <- predict(object = RF,dane_low,type = "prob")[,1]
up_pred <- predict(object = RF, dane_up ,type = "prob")[,1]

result <- dane %>% mutate(roznica = up_pred - low_pred) %>%
  group_by(grupa) %>%
    summarise(srednia = mean(roznica)) %>% select(srednia) %>%
  mutate(srodki = srodki[2:length(srodki)])

ggplot(result,aes(srodki,srednia))+geom_point()+geom_line()+
  ggtitle("ALE Plot zmienna Q21A_Year")
```




```r
year <- min(df$Q21A_Year):max(df$Q21A_Year)
df_Q21A_Year <- data.frame(ppwork = rep(os$ppwork,length(year))
                           ,w6_q19 = rep(os$w6_q19,length(year))
                           ,Q21A_Year = year
                           ,ppage = rep(os$ppage,length(year)) )

df_Q21A_Year$prob_married <- predict(RF,df_Q21A_Year,type = "prob")[,1]
```

```
ggplot(df_Q21A_Year,aes(x = Q21A_Year,y = prob_married)) + geom_line() +
  ggtitle("Ceteris Paribus zmiennej Q21A_Year")
```



Wnioski:

Zmienna *ppage* bardzo sie zmienia dla malych wartosci, obserwujemy tam najwieksze zmiany Dla wiekszych wartosci zachowuje sie stabilnie (nie ma duzych zmian na malych odcinkach). Najwieksze zmiany na plus obserwujemy na odcinku od 20 lat do 35 lat, czyli w okresie w którym najwiecej osób bierze ślub. Dla wiekszych wartosci zmiennej obserwujemy o wiele mniejsze zmiany w prawdopobienstwie bycia w zwiazku malzenskim. Mozna wysnuc wniosek, ze jezeli nie wezmie sie slubu do pewnego wieku to potem prawdopodbienstwo bycia w zwiazku malzenskim nie zmienia sie.

Zmienna *Q21A_Year* jest bardzo stabilna dla poczatkowych wartosci, tzn ze od pewnej wartosci lat znajomosci z obecnym partnerem prawdopodobienstwo malzenstwa nie zmienia sie. Praktycznie nie ma roznicy czy osoby te znaja sie 60 czy 30 lat. Jednak dla lat mniej odleglych od dnia dzisiejszego obserwujemy duze zmiany w wartosciach prawdopodbienstwa, czyli prawdopodobienstwo malzenstwa maleje wraz ze spadkiem dlugosci znajomosci. Podobny trend obserwujemy w wykresie Ceteris Paribus.