Projekt - Adam Iwanicki

Wstęp

Poniższy projekt oparłem na danych GSS. W pierwszej części przeprowadzę wstępną analizę dzietności z uwzglednieniem wybranych zmiennych. W drugiej części zaproponuję modele predykcyjne. W trzeciej podsumuję wyniki weryfikacji hipotez w oparciu o najlepsze modele. Sprawdzę czy na podstawie danych zebranych w latach 1972-2018 mozna wysnuć nastepujace wnioski:

1. Wyższe wykształcenie koreluje z mniejszą liczbą potomków.
2. Brak wyznawanej religii koreluje z mniejszą liczbą potomków.
3. Aktywność zawodowa koreluje z mniejszą liczbą potomków.

Pakiety z których korzystam:

```
library(ggplot2)
library(dplyr)
library(Metrics)
library(gridExtra)
```

W celu skompresowania analizy od momentu wprowadzenia zmiennych skategoryzowanych podsumowania modelu ograniczam do dostosowanej statystyki R kwadrat, oraz od początku ograniczam ilość wykresów diagnostycznych model do moim zdaniem niezbędnego minimum.

Projekt wykonałem samodzielnie, ponieważ od osoby z pary otrzymałem odpowiedź: “nie studiuję mmad”.

Analiza

Wgląd w strukturę danych

```
load("GSSdata.Rdata")
GSS.data <- data.frame(GSS.data)
dim(GSS.data)
```

```
## [1] 64814  6108
```

Cały zbiór jest olbrzymi, 64814 wierszy w 6108 kategoriach. Niestety duża część zadawanych pytań zmieniała się na przestrzeni lat, do analizy starałem sie wybrać jak najmniej wybrakowane kolumny. Utworzę nową ramkę danych i zaimportuję jedynie interesujace mnie zmienne, resztę “wyładuję” ze środowiska.

```
df <- data.frame(GSS.data)[c('YEAR', 'SEX', 'AGE', 'SIBS', 'EDUC', 'CHILDS',
                             'DEGREE', 'WRKSTAT', 'RELIG')]
rm(GSS.data)
summary(df)
```

```
##       YEAR          SEX             AGE             SIBS            EDUC
##  Min.   :1972   MALE  :28614   30     : 1450   2      :11796   12     :19663
##  1st Qu.:1984   FEMALE:36200   28     : 1432   1      :10624   16     : 8355
##  Median :1996                  32     : 1431   3      : 9945   14     : 7160
##  Mean   :1995                  34     : 1422   4      : 7268   13     : 5360
```

```
##  3rd Qu.:2006                  27     : 1391   5      : 5242   11     : 3743
##  Max.   :2018                  35     : 1383   6      : 3996   15     : 2910
##                                (Other):56305   (Other):15943   (Other):17623
##      CHILDS                 DEGREE                  WRKSTAT
##  0      :17657   HIGH SCHOOL   :33195   WORKING FULLTIME:31892
##  2      :16072   LT HIGH SCHOOL:13587   KEEPING HOUSE   :10176
##  1      :10304   BACHELOR      : 9475   RETIRED         : 9121
##  3      :10099   GRADUATE      : 4716   WORKING PARTTIME: 6719
##  4      : 5231   JUNIOR COLLEGE: 3668   UNEMPL, LAID OFF: 2179
##  5      : 2398   NA            :  143   SCHOOL          : 1998
##  (Other): 3053   (Other)       :   30   (Other)         : 2729
##         RELIG
##  PROTESTANT:37117
##  CATHOLIC  :15674
##  NONE      : 7797
##  JEWISH    : 1285
##  OTHER     : 1086
##  CHRISTIAN :  791
##  (Other)   : 1064
```

Czyszczenie danych

Następnym krokiem będzie usunięcie wpisów z nieznanymi wartościami w kluczowych polach.

```
df <- df[with(df, ifelse(DEGREE!='NA', TRUE, FALSE)),]
df <- df[with(df, ifelse(DEGREE!='DK', TRUE, FALSE)),]
df <- df[with(df, ifelse(WRKSTAT!='NA', TRUE, FALSE)),]
df <- df[with(df, ifelse(RELIG!='NA', TRUE, FALSE)),]
df <- df[with(df, ifelse(RELIG!='DK', TRUE, FALSE)),]

dim(df)
```

```
## [1] 64372     9
```

```
summary(df)
```

```
##       YEAR          SEX             AGE             SIBS            EDUC
##  Min.   :1972   MALE  :28405   30     : 1444   2      :11717   12     :19560
##  1st Qu.:1984   FEMALE:35967   28     : 1423   1      :10572   16     : 8304
##  Median :1996                  32     : 1421   3      : 9876   14     : 7111
##  Mean   :1995                  34     : 1413   4      : 7222   13     : 5331
##  3rd Qu.:2006                  27     : 1387   5      : 5219   11     : 3700
##  Max.   :2018                  25     : 1375   6      : 3966   15     : 2898
##                                (Other):55909   (Other):15800   (Other):17468
##      CHILDS                 DEGREE                  WRKSTAT
##  0      :17548   HIGH SCHOOL   :33062   WORKING FULLTIME:31692
##  2      :15978   LT HIGH SCHOOL:13548   KEEPING HOUSE   :10111
##  1      :10236   BACHELOR      : 9421   RETIRED         : 9052
##  3      :10040   GRADUATE      : 4696   WORKING PARTTIME: 6684
##  4      : 5198   JUNIOR COLLEGE: 3645   UNEMPL, LAID OFF: 2167
##  5      : 2388   IAP           :    0   SCHOOL          : 1980
##  (Other): 2984   (Other)       :    0   (Other)         : 2686
##         RELIG
##  PROTESTANT:37027
##  CATHOLIC  :15638
##  NONE      : 7780
```

```
##  JEWISH    : 1283
##  OTHER     : 1083
##  CHRISTIAN :  788
##  (Other)   :  773
```

Wszystkie zmienne poza rokiem uzyskania odpowiedzi sa kategoryczne, jednak moim zdaniem istotę części z nich będzie oddawać forma numeryczna. Zmienię: wiek, subiektywne odczucie poziomu zycia, liczby ukonczonych klas, rodzenstwa i dzieci. Dodamy także binarną zmienną religijności.

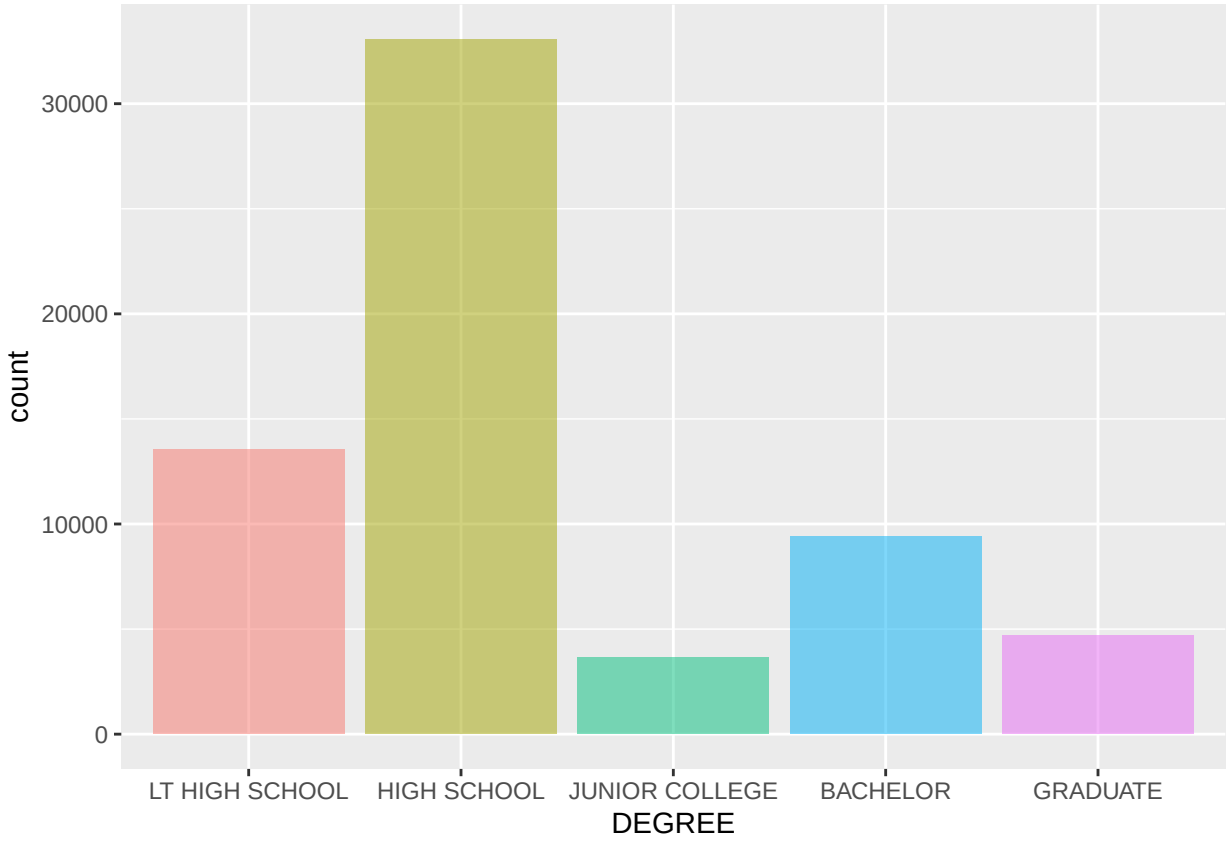
```
df <- df %>%  mutate(  AGE = as.integer(AGE) + 17, # 18 lat ma indeks 1
                       CHILDS = as.integer(CHILDS) - 1, # 0 dzieci ma indeks 1
                       SIBS = as.integer(SIBS) - 1, # 0 rodzeństwa ma indeks 1
                       EDUC = as.integer(EDUC) - 1) # 0 ukończonych klas ma indeks 1
df$religious <- with(df, ifelse(RELIG!='NONE', TRUE, FALSE))
summary(df)
```

```
##       YEAR          SEX             AGE             SIBS            EDUC
##  Min.   :1972   MALE  :28405   Min.   :18.00   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:1984   FEMALE:35967   1st Qu.:32.00   1st Qu.: 3.000   1st Qu.:12.00
##  Median :1996                  Median :44.00   Median : 4.000   Median :12.00
##  Mean   :1995                  Mean   :46.23   Mean   : 4.874   Mean   :12.88
##  3rd Qu.:2006                  3rd Qu.:59.00   3rd Qu.: 6.000   3rd Qu.:15.00
##  Max.   :2018                  Max.   :91.00   Max.   :42.000   Max.   :23.00
##
##      CHILDS                 DEGREE                  WRKSTAT
##  Min.   :0.000   HIGH SCHOOL   :33062   WORKING FULLTIME:31692
##  1st Qu.:0.000   LT HIGH SCHOOL:13548   KEEPING HOUSE   :10111
##  Median :2.000   BACHELOR      : 9421   RETIRED         : 9052
##  Mean   :1.956   GRADUATE      : 4696   WORKING PARTTIME: 6684
##  3rd Qu.:3.000   JUNIOR COLLEGE: 3645   UNEMPL, LAID OFF: 2167
##  Max.   :9.000   IAP           :    0   SCHOOL          : 1980
##                  (Other)       :    0   (Other)         : 2686
##         RELIG       religious
##  PROTESTANT:37027   Mode :logical
##  CATHOLIC  :15638   FALSE:7780
##  NONE      : 7780   TRUE :56592
##  JEWISH    : 1283
##  OTHER     : 1083
##  CHRISTIAN :  788
##  (Other)   :  773
```

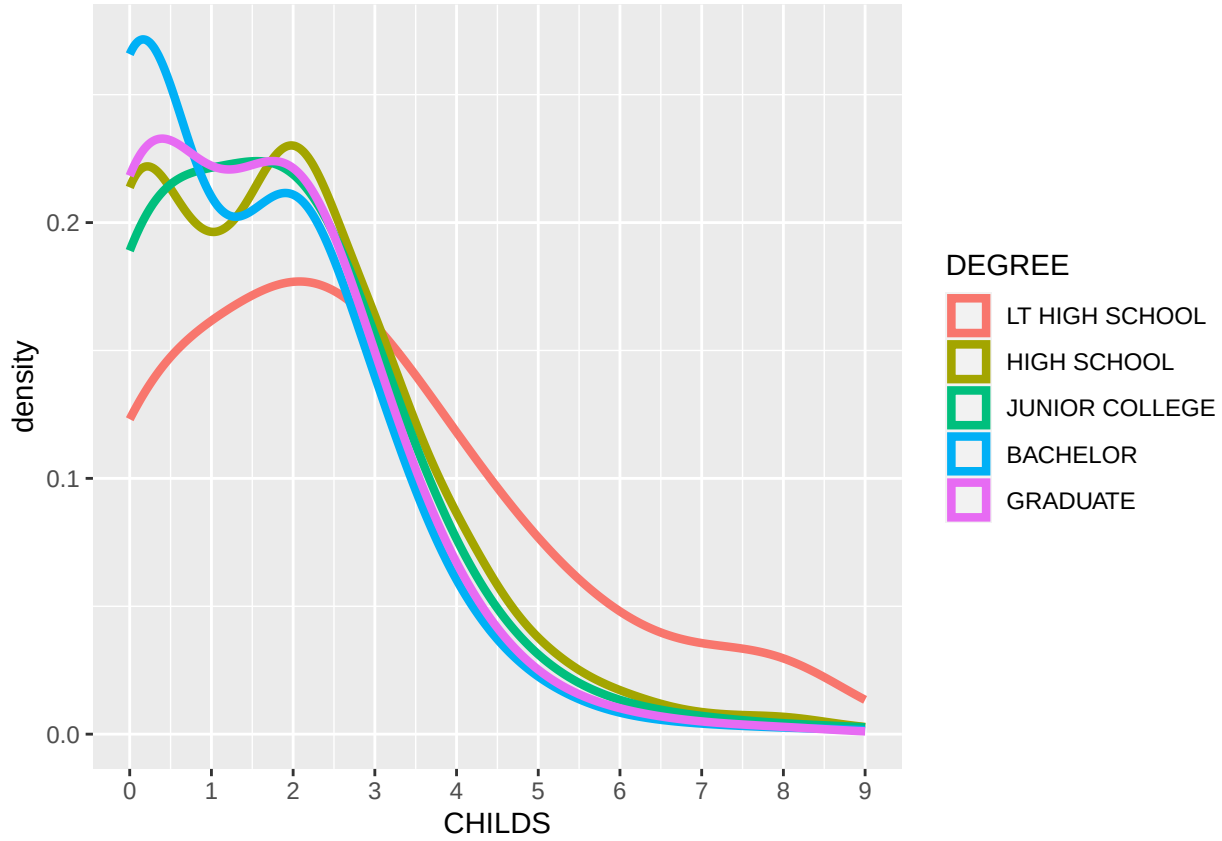
Wgląd w dane

Ad 1.

```
ggplot(df, aes(x = DEGREE, fill = DEGREE)) +
  geom_histogram(stat="count", alpha=0.5) +
  theme(legend.position = 'none')
```

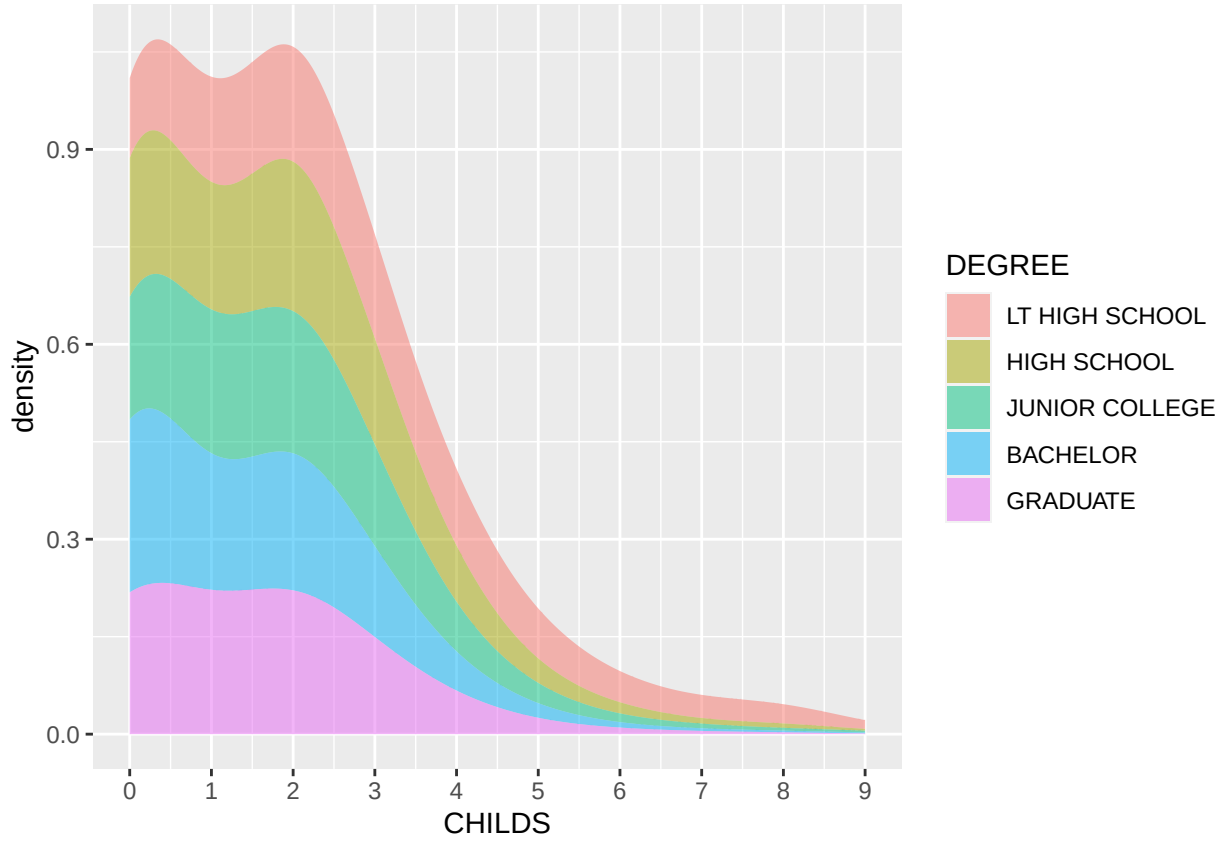



```
ggplot(df, aes(x = CHILDS, color = DEGREE)) +
  geom_density(adjust = 3, size = 1.5) +
  scale_x_continuous(breaks=0:9)
```



Dla lepszego zobrazowania danych (szczególnie większych ilości pociech) zdecydowałem się na umieszczenie także wykresu stosowego:

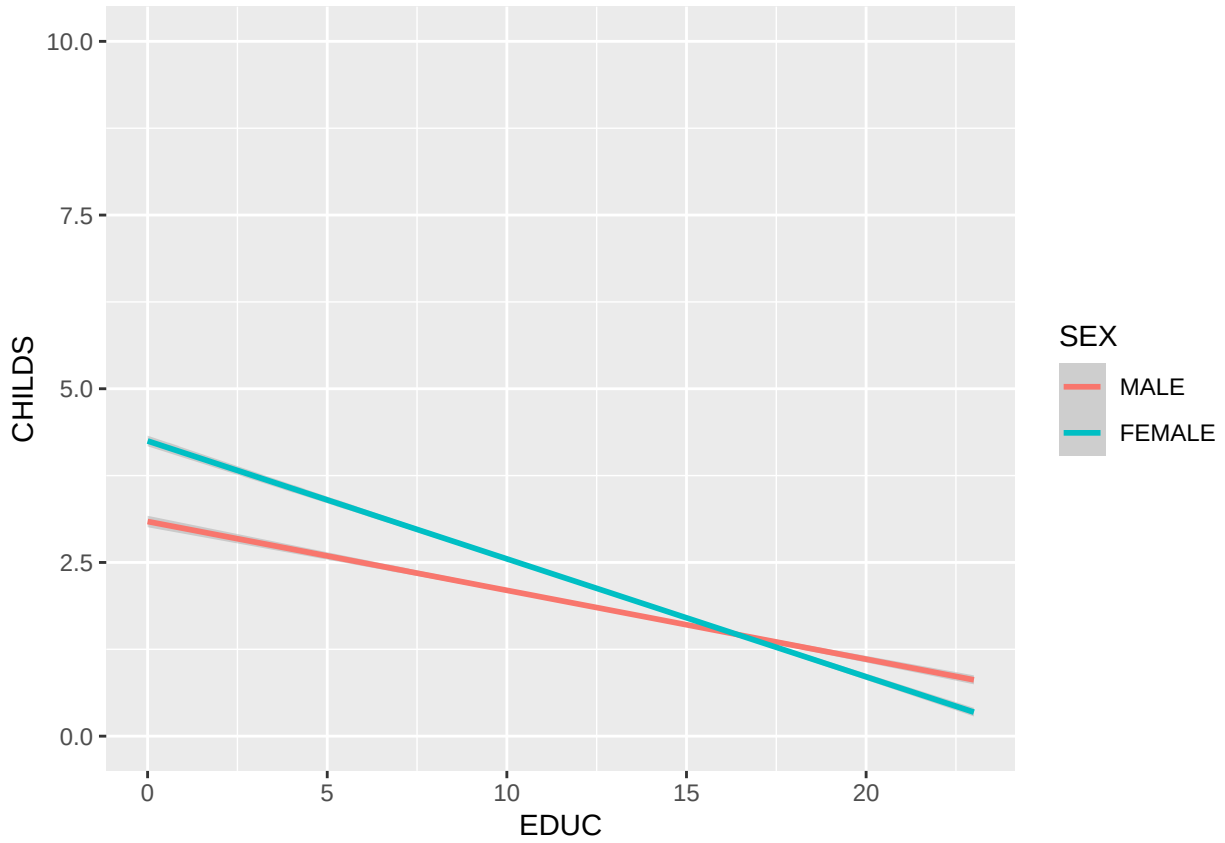
```
ggplot(df, aes(x = CHILDS, fill = DEGREE)) +
  stat_density(adjust = 3, alpha = .5) +
  scale_x_continuous(breaks=0:9)
```



Powyższe wykresy zdają się potwierdzać naszą hipotezę. Na wykresie gęstości zdecydowanie widać, że osoby z wyższym wykształceniem częściej nie posiadają wcale dzieci, oraz istotnie rzadziej decydują się więej niż czworo. Na powyżej pięciorga pociech najczęściej decydują się osoby poniżej średniego wykształcenia (less than high school). Warto zrobić jeszcze wykres uśredniający liczbę dzieci dla ilości ukończonych klas.

```
ggplot(df, aes(x = EDUC, y = CHILDS)) +
  geom_smooth(aes(color = SEX), method='lm') +
  coord_cartesian(ylim = c(0, 10))
```

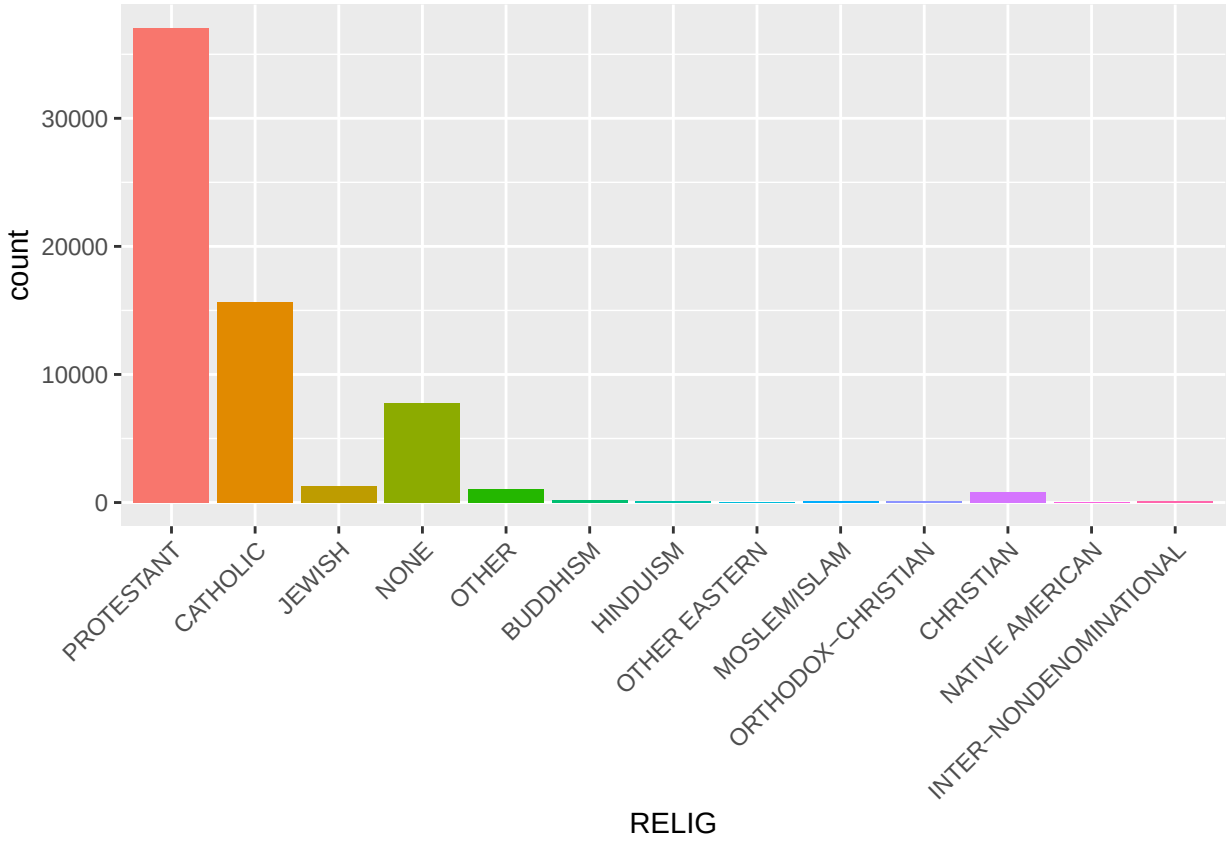
```
## `geom_smooth()` using formula = 'y ~ x'
```



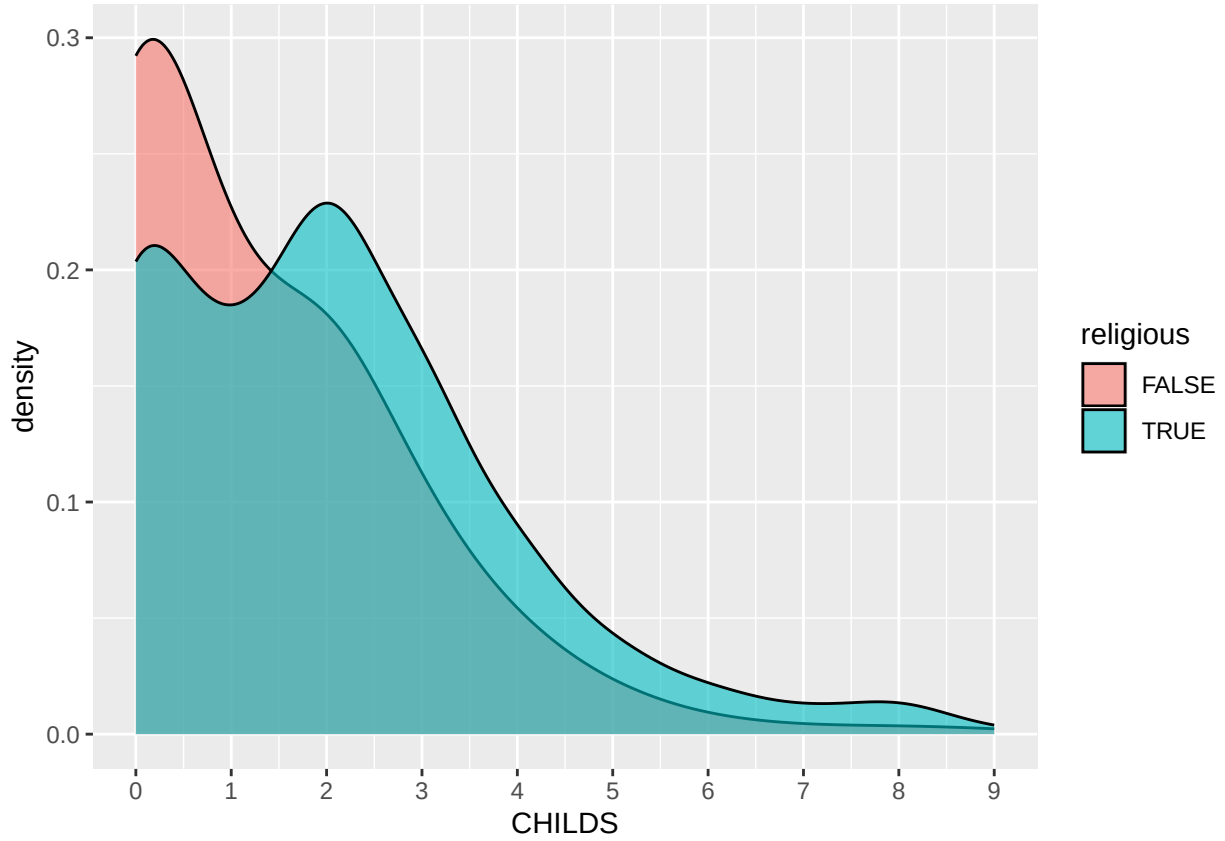
Podsumowując wydaje się, że występuje badana zależność. Efekt wydaje się być silniejszy w przypadku kobiet.

Ad 2.

```
ggplot(df, aes(x = RELIG, fill=RELIG))+
  geom_histogram(stat='count')+
  theme(axis.text.x = element_text(angle=45, vjust=1, hjust=1), legend.position = 'none')
```



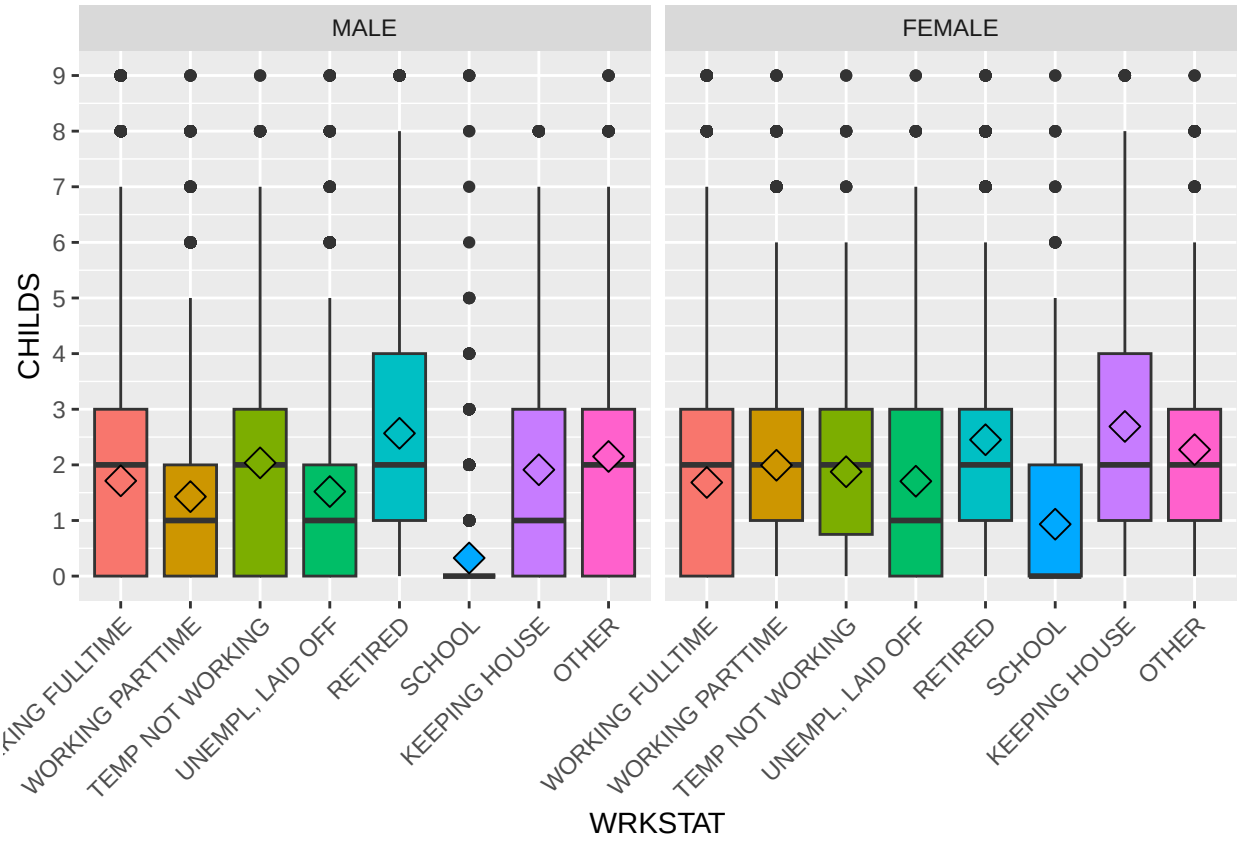

```
ggplot(df, aes(x = CHILDS, fill = religious)) +
  geom_density(adjust = 3, alpha = .6) +
  scale_x_continuous(breaks=0:9)
```



Powyższy wykres zdaje się potwierdzać naszą hipotezę. W naszym zbiorze przeważają protestanci i katolicy, jednak ponad 10% deklaruje brak przynależności do grupy wyznaniowej.

Ad 3.

```
ggplot(df, aes(x = WRKSTAT, y = CHILDS, fill = WRKSTAT)) +
  facet_wrap(~SEX)+
  geom_boxplot() +
  theme(axis.text.x = element_text(angle=45, vjust=1, hjust=1), legend.position = 'none') +
  scale_y_continuous(breaks=0:9)+
  stat_summary(fun.y=mean, geom="point", shape=23, size=4)
```



Na podstawie powyższych wykresów ciężko wysnuć uniwersalne wnioski. Obie płcie pracujące na pełen etat mają minimalnie obniżoną średnią (zaznaczona rombem) liczbę potomstwa. U mężczyzn zdecydowanie status ucznia obniża dzietność, natomiast niepełny etat jak i wstrzymanie się od pracy lekko zbiegają się ze spadkiem. Spodziewałem się większej różnicy. W przypadku kobiet uczących się widzimy delikatny spadek ilości dzieci, jednak nie tak silny jak u mężczyzn. Natomiast pozycja zawodowa zdaje się nie mieć wpływu na dzietność, chociaż decyzja o wstrzymaniu się od pracy na rzecz opiekowania się domem (w tym potomstwem) delikatnie podnosi nam 3. kwantyl jak i średnią.

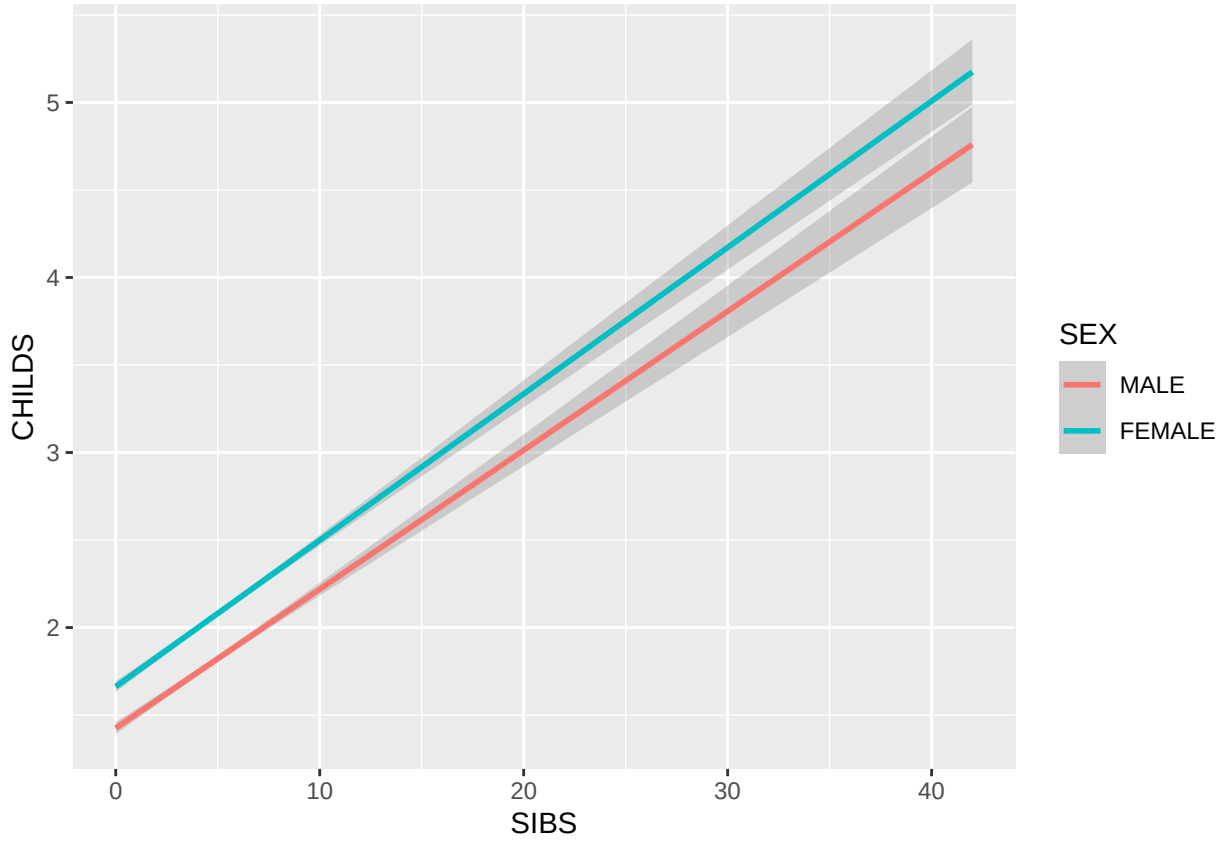
Warto zaznaczyć, że w przypadku badania niektórych statusów zawodowych, są one powiązane z wiekiem. Emeryci i uczniowie (poza wyjątkami) należą do dwóch przeciwlegych grup wiekowych.

Analiza potencjalnych interakcji

Warta zbadania wydaje się uśrednieniona zależność między ilością dzieci i rodzeństwa. Legenda dotyczy się też następnych wykresów.

```
ggplot(df, aes(x = SIBS, y = CHILDS, color = SEX)) +
  geom_smooth(method='lm')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

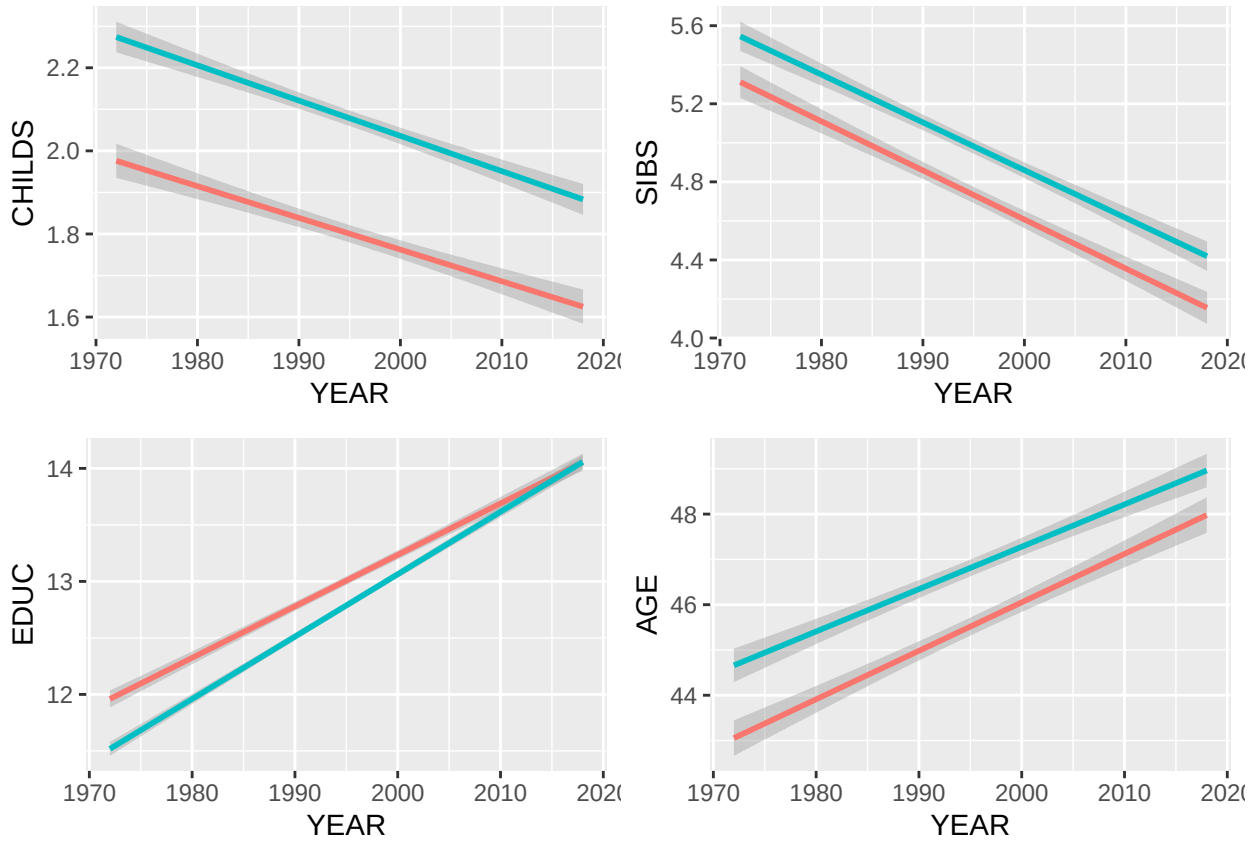


Zależność widoczna gołym okiem, jednak moim zdaniem należy tu wziąć poprawkę na czas. Zbiór danych sięga 50 lat wstecz, gdy większe rodziny były standardem.

Warto także sprawdzić jak na przestrzeni czasu wygląda średni poziom edukacji oraz średni wiek badanych.

```
plot1 <- ggplot(df, aes(x = YEAR, y = CHILDS, color = SEX)) +
  geom_smooth(method='lm') + theme(legend.position = 'none')
plot2 <- ggplot(df, aes(x = YEAR, y = SIBS, color = SEX)) +
  geom_smooth(method='lm') + theme(legend.position = 'none')
plot3 <- ggplot(df, aes(x = YEAR, y = EDUC, color = SEX)) +
  geom_smooth(method='lm') + theme(legend.position = 'none')
plot4 <- ggplot(df, aes(x = YEAR, y = AGE, color = SEX)) +
  geom_smooth(method='lm') + theme(legend.position = 'none')
grid.arrange(plot1, plot2, plot3, plot4, ncol=2, nrow = 2)
```

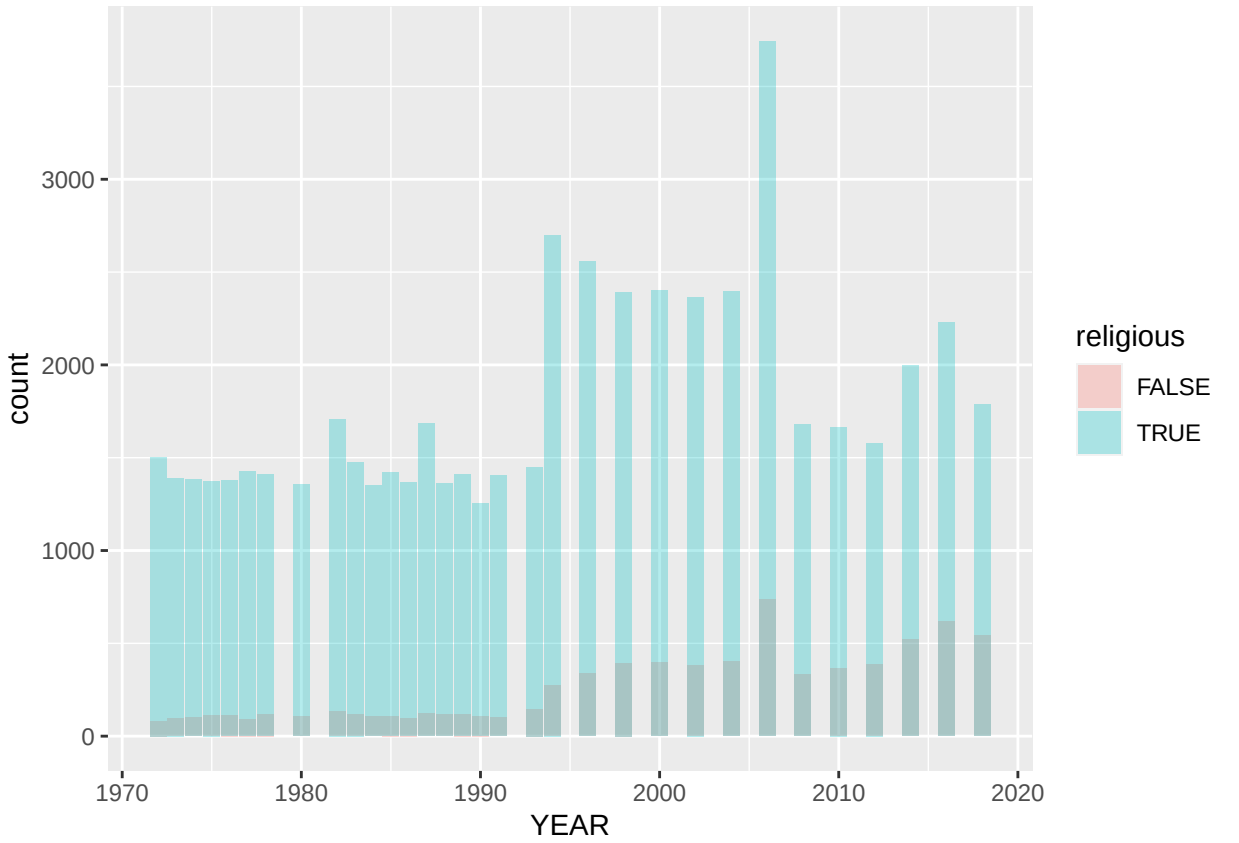
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Rzeczywiście widoczny jest znaczny (mniej więcej o 20%) spadek obu tych wartości. Moim zdaniem zmiana obu wartości jest wynikiem innych czynników i pomimo korelacji nie dopatrywałbym się efektów przyczynowo-skutkowych w żadną stronę.

Poniżej wykres obrazujący religijność społeczeństwa na przestrzeni badanych lat.

```
ggplot(df, aes(x = YEAR, fill = religious))+
  geom_histogram(stat='count', alpha = 0.3, position = 'identity')
```



Jak w przypadku edukacji zmiana jest delikatna, na przestrzeni 50 lat średnia liczba ukończonych klas to 2 u mężczyzn (+17%) i 2.5 u kobiet (+22%). W przypadku osób deklarujących ateizm przyrost jest znaczny. Od wartości marginalnych (~5%) wzrost jest 4.5 krotny do prawie 1/4 badanej populacji.

Modele

Na potrzeby tego rozdziału podzielimy dane na zbiór treningowy i testowy w proporcji 4:1. Testować będziemy jedynie najlepiej sprawdzający się na danych treningowych model.

```
set.seed(7777777)
train_ind <- sample(seq_len(nrow(df)), size = floor(0.8 * nrow(df)))
train <- df[train_ind, ]
test <- df[-train_ind,]
expected_results <- test[,'CHILDS']
test <- test[ , ! names(test) %in% 'CHILDS']
```

Regresja liniowa

Model zerowy, do którego będziemy porównywać badając efektywność, będzie oparty na zmiennej roku uzyskania odpowiedzi.

```
model.year <-lm(formula = CHILDS ~ YEAR, data = train)
summary(model.year)
```

```
##
## Call:
## lm(formula = CHILDS ~ YEAR, data = train)
```

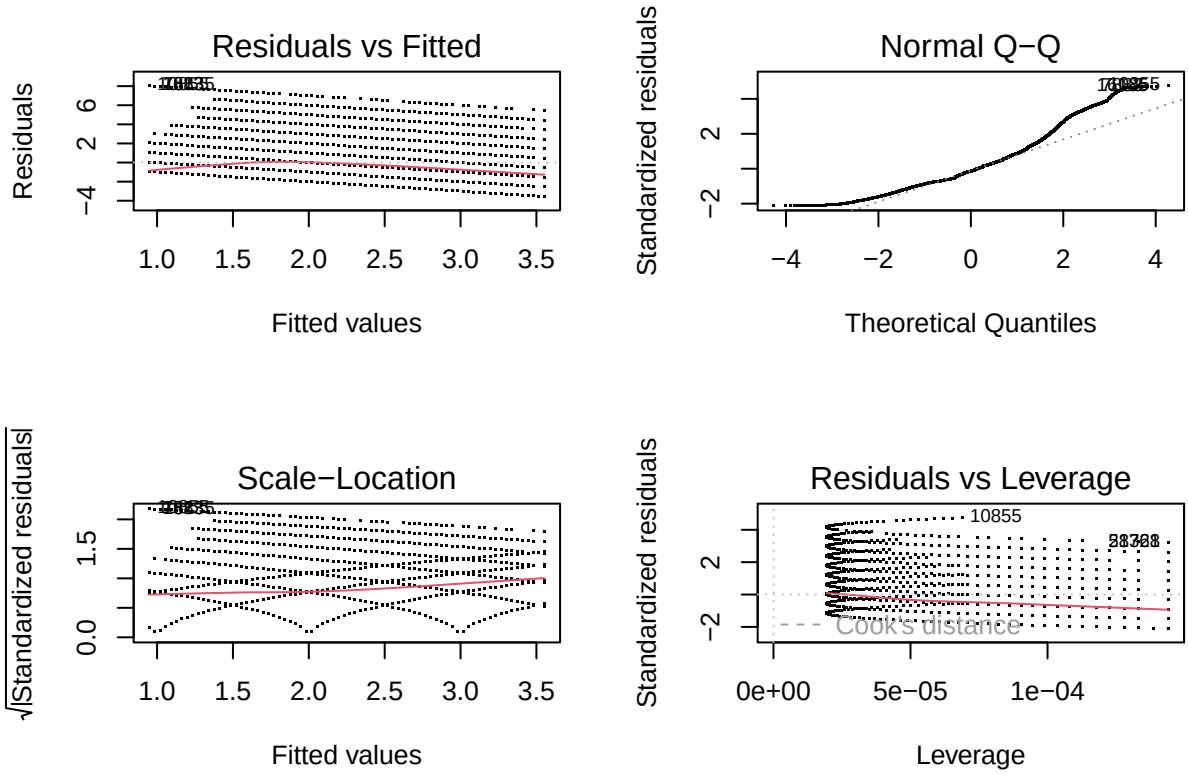
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1401 -1.8014 -0.0272  1.0373  7.2309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.0446189  1.1804082   15.29   <2e-16 ***
## YEAR        -0.0080652  0.0005917  -13.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.805 on 51495 degrees of freedom
## Multiple R-squared:  0.003595,   Adjusted R-squared:  0.003576
## F-statistic: 185.8 on 1 and 51495 DF,  p-value: < 2.2e-16
```

Jak widzimy model sprawuje się słabo, R kwadrat bliskie 0 to bardzo nisko zawieszona poprzeczka. Oglądanie wykresów diagnostycznych mija się z celem. Zobaczmy czy wiek będzie lepszym predyktorem ilości dzieci.

```
model.age <-lm(formula = CHILDS ~ AGE, data = train)
summary(model.age)
```

```
##
## Call:
## lm(formula = CHILDS ~ AGE, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5518 -1.1993 -0.2330  0.8383  8.0503
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.3081342  0.0209038   14.74   <2e-16 ***
## AGE         0.0356448  0.0004225   84.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.695 on 51495 degrees of freedom
## Multiple R-squared:  0.1214, Adjusted R-squared:  0.1214
## F-statistic:  7117 on 1 and 51495 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model.age, pch = '.')
```



Zdecydowanie lepiej, jednak wciąż mało satysfakcjonujące rezultaty. Wydaje się, że predykcje modelu są z grubsza losowe, dodatkowo prawie że nie zwraca wartości między 0 i 1, ani większych od 3.5.

Uwagę przykuwają także wykresy diagnostyczne, linie reszt modelu wynikają z dyskretnego i całościowego faktu posiadania dzieci. W rzeczywistym świecie nie jest możliwe urodzenie połowy dziecka, jednak nasz model będzie zwracać wartości ułamkowe. Należy traktować je jako prawdopodobieństwo posiadania potomka tj 0.2 oznacza 20% na jedno dziecko, 2.8 prawdopodobne dwa + trzecie na 80%. Jeśli chcielibyśmy wynik w wartościach całościowych należałoby go zaokrąglać w funkcji optymalizującej. W pełni mnie satysfakcjonuje jednak wynik probabilistyczny.

Ze zmiennych liczbowych mamy jeszcze ilość ukończonych klas.

```
model.class <-lm(formula = CHILDS ~ EDUC, data = train)
summary(model.class)
```

```
##
## Call:
## lm(formula = CHILDS ~ EDUC, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7373 -1.4920 -0.0769  0.9231  8.4451
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.737263   0.032026  116.70   <2e-16 ***
## EDUC        -0.138363   0.002413  -57.34   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.753 on 51495 degrees of freedom
## Multiple R-squared:  0.06001,    Adjusted R-squared:   0.06
## F-statistic:  3288 on 1 and 51495 DF,  p-value: < 2.2e-16
```

Ten predyktor sprawia się gorzej niż wiek, lepiej niż rok pobrania zmiennej. Można wysnuć wnioski, że taka zależność rzeczywiście istnieje, jednak samodzielna nie jest istotnie znacząca.

```
model.sibs <-lm(formula = CHILDS ~ SIBS, data = train)
summary(model.sibs)
```

```
##
## Call:
## lm(formula = CHILDS ~ SIBS, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0677 -1.7149 -0.0502  1.0337  7.4527
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.547252   0.013095  118.16   <2e-16 ***
## SIBS        0.083820   0.002153   38.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.782 on 51495 degrees of freedom
## Multiple R-squared:  0.0286, Adjusted R-squared:  0.02858
## F-statistic:  1516 on 1 and 51495 DF,  p-value: < 2.2e-16
```

Regresja wieloliniowa

Opierając się na dwóch najlepiej dopasowanych modelach z poprzedniego rozdziału, pierwszym wielorakim modelem regresji liniowej będzie szacujący ilość dzieci na podstawie wieku i ilości ukończonych klas.

```
model.ageclass <-lm(formula = CHILDS ~ AGE+EDUC, data = train)
summary(model.ageclass)
```

```
##
## Call:
## lm(formula = CHILDS ~ AGE + EDUC, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7873 -1.1715 -0.2153  0.8608  8.6571
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.8495224  0.0390403   47.38   <2e-16 ***
## AGE          0.0322828  0.0004203   76.82   <2e-16 ***
## EDUC        -0.1076146  0.0023205  -46.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.661 on 51494 degrees of freedom
```

```
## Multiple R-squared:  0.1567, Adjusted R-squared:  0.1566
## F-statistic:  4783 on 2 and 51494 DF,  p-value: < 2.2e-16
```

Znacznie lepiej niż modele z jedną zmienną. Dodatkowym atutem jest szerszy zakres zwracanych wartości, niestety za tym idzie większy rozrzut reszt naszego modelu. Sprawdźmy czy dodanie zmiennej odpowiedzialnej za rok pochodzenia wpisu poprawi nasze predykcje.

```
model.ageclassyear <- lm(formula = CHILDS ~ AGE+EDUC+YEAR, data = train)
summary(model.ageclassyear)
```

```
##
## Call:
## lm(formula = CHILDS ~ AGE + EDUC + YEAR, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8756 -1.1555 -0.2047  0.8646  8.6852
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.1366024  1.1098193   12.74   <2e-16 ***
## AGE          0.0328401  0.0004228   77.68   <2e-16 ***
## EDUC        -0.1014762  0.0023831  -42.58   <2e-16 ***
## YEAR        -0.0062117  0.0005607  -11.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.659 on 51493 degrees of freedom
## Multiple R-squared:  0.1587, Adjusted R-squared:  0.1586
## F-statistic:  3237 on 3 and 51493 DF,  p-value: < 2.2e-16
```

Statystyki dopasowania modelu lekko się poprawiły, jednak dodanie roku ma minimalny wpływ.

Wiemy że poziom wykształcenia zmieniał się z czasem, tak samo jak średni wiek badanej osoby (malejący przyrost i poprawa świadczeń medycznych). Nie mamy pewności czy istnieje zależność wiek/edukacja, ale wydawałoby się to logiczne. Sprawdźmy jak poradzi sobie potrójny model z uwzględnieniem interakcji.

```
model.interactions <-lm(formula = CHILDS ~ AGE*EDUC*YEAR, data = train)
summary(model.interactions)
```

```
##
## Call:
## lm(formula = CHILDS ~ AGE * EDUC * YEAR, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3142 -1.1118 -0.2031  0.8801  8.9142
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.440e+02  1.389e+01  10.373   <2e-16 ***
## AGE           -2.687e+00  2.491e-01 -10.785   <2e-16 ***
## EDUC          -9.158e+00  1.057e+00  -8.665   <2e-16 ***
## YEAR          -7.077e-02  6.966e-03 -10.160   <2e-16 ***
## AGE:EDUC       1.934e-01  1.936e-02   9.993   <2e-16 ***
## AGE:YEAR       1.353e-03  1.250e-04  10.824   <2e-16 ***
## EDUC:YEAR      4.494e-03  5.300e-04   8.480   <2e-16 ***
```

```
## AGE:EDUC:YEAR -9.610e-05  9.707e-06  -9.901   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.653 on 51489 degrees of freedom
## Multiple R-squared:  0.1645, Adjusted R-squared:  0.1644
## F-statistic:  1448 on 7 and 51489 DF,  p-value: < 2.2e-16
```

Statystyki wyglądają lepiej niż modelu bez interakcji, jednak wciąż jest dalece odbiegający od ideału...

Ponownie ostatnią możliwą zmienną liczbową jest liczba rodzeństwa.

```
model.interactions2 <-lm(formula = CHILDS ~ AGE*EDUC*YEAR*SIBS, data = train)
summary(model.interactions2)
```

```
##
## Call:
## lm(formula = CHILDS ~ AGE * EDUC * YEAR * SIBS, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3563 -1.0985 -0.2013  0.8758  8.9009
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.081e+01  2.339e+01   3.883 0.000103 ***
## AGE                -1.831e+00  4.147e-01  -4.415 1.01e-05 ***
## EDUC               -4.754e+00  1.738e+00  -2.735 0.006246 **
## YEAR               -4.451e-02  1.173e-02  -3.794 0.000148 ***
## SIBS                9.183e+00  3.479e+00   2.639 0.008306 **
## AGE:EDUC            1.296e-01  3.135e-02   4.135 3.55e-05 ***
## AGE:YEAR            9.266e-04  2.081e-04   4.452 8.51e-06 ***
## EDUC:YEAR           2.308e-03  8.717e-04   2.647 0.008117 **
## AGE:SIBS           -1.480e-01  5.812e-02  -2.546 0.010888 *
## EDUC:SIBS          -8.553e-01  2.743e-01  -3.118 0.001819 **
## YEAR:SIBS          -4.538e-03  1.747e-03  -2.597 0.009401 **
## AGE:EDUC:YEAR      -6.426e-05  1.572e-05  -4.087 4.38e-05 ***
## AGE:EDUC:SIBS       1.218e-02  4.690e-03   2.596 0.009427 **
## AGE:YEAR:SIBS       7.370e-05  2.920e-05   2.524 0.011616 *
## EDUC:YEAR:SIBS      4.262e-04  1.377e-04   3.096 0.001963 **
## AGE:EDUC:YEAR:SIBS -6.086e-06  2.355e-06  -2.584 0.009773 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.645 on 51481 degrees of freedom
## Multiple R-squared:  0.173,  Adjusted R-squared:  0.1727
## F-statistic: 717.8 on 15 and 51481 DF,  p-value: < 2.2e-16
```

O dziwo poprawa modelu jest znaczna. Może liczba rodzeństwa ma jednak wpływ na liczbę potomków? Współczynniki modelu sugerują wpływ rzędu 1.6 "dodatkowego" dziecka na dziesięcioro rodzeństwa. Tego typu zależność mogłaby oznaczać, że malejąca dzietność społeczeństwa ma wpływ na dodatkowo malejącą dzietność przyszłych pokoleń. Jednak zagadnienie wydaje się zbyt skomplikowane żeby je wprowadzać jako dodatkowy element.

Analiza wariancji

W tym rozdziale będziemy wyjaśniać dzietność na podstawie zmiennych kategorycznych: płci, religii, statusu pracy oraz uzyskanego stopnia wykształcenia.

Zdecydowałem się użyć funkcji lm() zamiast aov() ze względu na wygodę uzyskania statystyki R kwadrat

```
model.wrk <-lm(formula = CHILDS ~ WRKSTAT, data = train)
summary(model.wrk)
```

```
##
## Call:
## lm(formula = CHILDS ~ WRKSTAT, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6688 -1.6688 -0.5062  1.1756  8.3520
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.69501    0.01098 154.410  < 2e-16 ***
## WRKSTATWORKING PARTTIME    0.12939    0.02632   4.916 8.87e-07 ***
## WRKSTATTEMP NOT WORKING    0.22997    0.05399   4.260 2.05e-05 ***
## WRKSTATUNEMPL, LAID OFF -0.10695    0.04368  -2.449   0.0143 *
## WRKSTATRETIRED             0.81123    0.02332  34.793  < 2e-16 ***
## WRKSTATSCHOOL             -1.04703    0.04498 -23.277  < 2e-16 ***
## WRKSTATKEEPING HOUSE       0.97384    0.02229  43.681  < 2e-16 ***
## WRKSTATOTHER               0.52444    0.05430   9.658  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.748 on 51489 degrees of freedom
## Multiple R-squared:  0.06637,    Adjusted R-squared:  0.06624
## F-statistic: 522.9 on 7 and 51489 DF,  p-value: < 2.2e-16
```

Model źle dopasowany, jednak lepiej niż nasz model zerowy, a nawet opierający się jedynie o ilość ukończonych klas. Pozwolę sobie zakończyć na jednym przykładzie jednoczynnikowe modele AOV.

```
model.aov <-lm(formula = CHILDS ~ WRKSTAT+DEGREE+religious, data = train)
summary(model.aov)$adj.r.squared
```

```
## [1] 0.1092667
```

Jak widać nawet łącząc trzy zmienne nie osiągneliśmy poziomu dopasowania modelu opartego jedynie o wiek. Wiemy jednak, że poziom edukacji jak i status zatrudnienia wpływa nieco inaczej w zależności od płci. Dodajmy tę zmienną i uwzględnijmy interakcje.

```
model.aov.interaction <-lm(formula = CHILDS ~ SEX*WRKSTAT + SEX*DEGREE+religious, data = train)
summary(model.aov.interaction)$adj.r.squared
```

```
## [1] 0.114702
```

Model zyskał nieco na efektywności, jednak kosztem znacznego zwiększenia ilości współczynników. W związku z niewielkim postępem, a coraz bardziej skomplikowanym modelem podarujemy sobie resztę potencjalnych modeli AOV.

Analiza kowariancji

W tej części przeanalizujemy najbardziej obiecujący model regresji wielorakiej i stopniowo będziemy dodawać składowe modeli AOV.

Na pierwszy ogień pójdzie model ze wszystkimi zmiennymi.

```
model.all <-lm(formula = CHILDS ~ ., data = train)
summary(model.all)$adj.r.squared
```

```
## [1] 0.192515
```

Nie jest dobrze, pomimo wrzucenia wszystkich czynników (nie wszystkich możliwych, model jest bez interakcji) osiągamy niską skuteczność predykcji obarczoną olbrzymim błędem.

Wróćmy do najlepszego modelu wielorakiego. Zaczyniemy od dodania do niego predyktora odpowiadającego religijności.

```
model.acov <-lm(formula = CHILDS ~ AGE*YEAR*EDUC*SIBS + religious, data = train)
summary(model.acov)$adj.r.squared
```

```
## [1] 0.1769897
```

Poprawiło to jedynie minimalnie statystykę R kwadrat i błąd standardowy o tysięczne części... Spróbujmy szczęścia dodając interakcję płci ze statusem zatrudnienia i poziomem wykształcenia. Opierając się na końcowych wykresach wglądu w dane dodatkowo zaryzykowałbym dodanie interakcji płci do pierwszej wieloliniowej zależności.

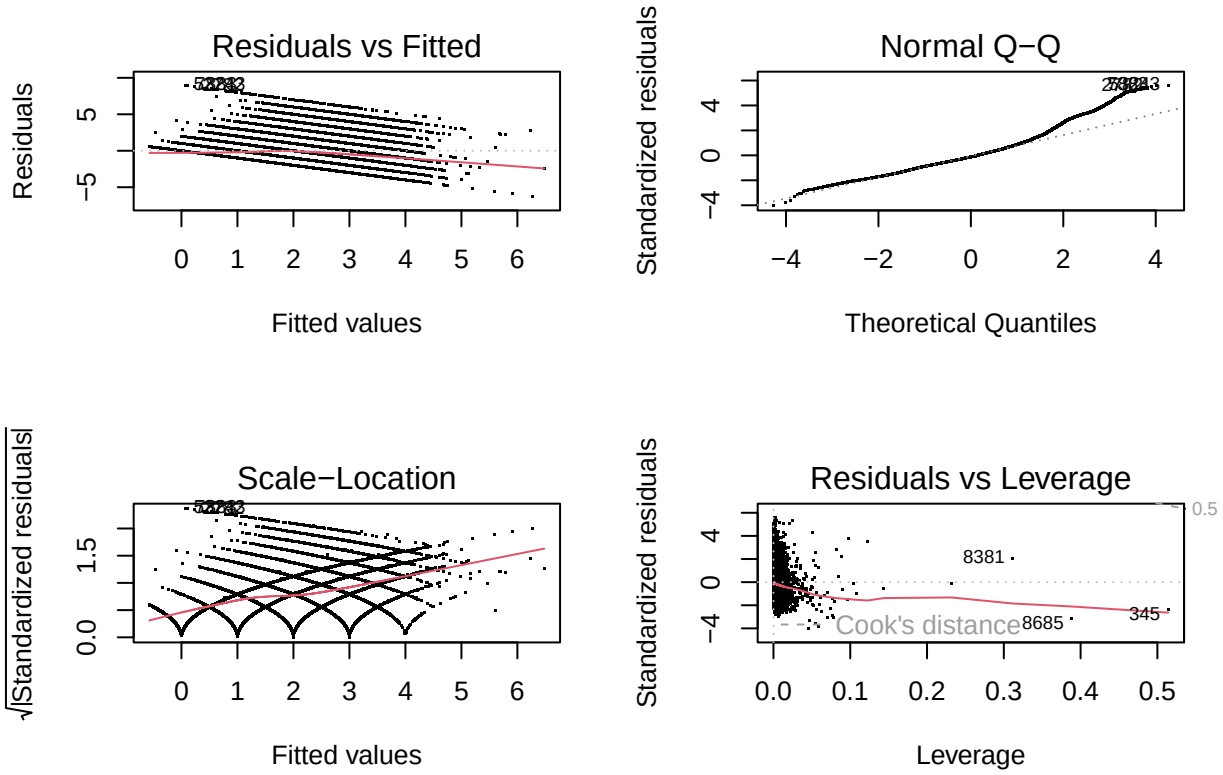
Dodatkowo podejrzewam istnienie zależności: wiek~płeć~status zatrudnienia (dawniej kobiety częściej zajmowały się domem + zależności emerytalne) stopień naukowy~liczba ukończonych klas

W celu skrócenia już przydługiej analizy dodam je wszystkie w jednym kroku.

```
model.acov2 <-lm(formula = CHILDS ~ SEX*AGE*EDUC*YEAR*SIBS + religious +
                   SEX*AGE*WRKSTAT + SEX*AGE*EDUC*DEGREE, data = train)
summary(model.acov2)$adj.r.squared
```

```
## [1] 0.2161819
```

```
par(mfrow=c(2,2))
plot(model.acov2, pch = '.')
```



Udało się pokonać granicę wyznaczaną przez model z pełnym zestawem danych bez interakcji. Dodatkowo wartości zwracane przez model są zróżnicowane (maksymalnie osiągając >6, niestety pojawiły się oczekiwane dzieci "ujemne" co ciężko mi racjonalnie zinterpretować) Jednak spora część współczynników modelu ma wartości t bliskie zeru, o nikłym znaczeniu.

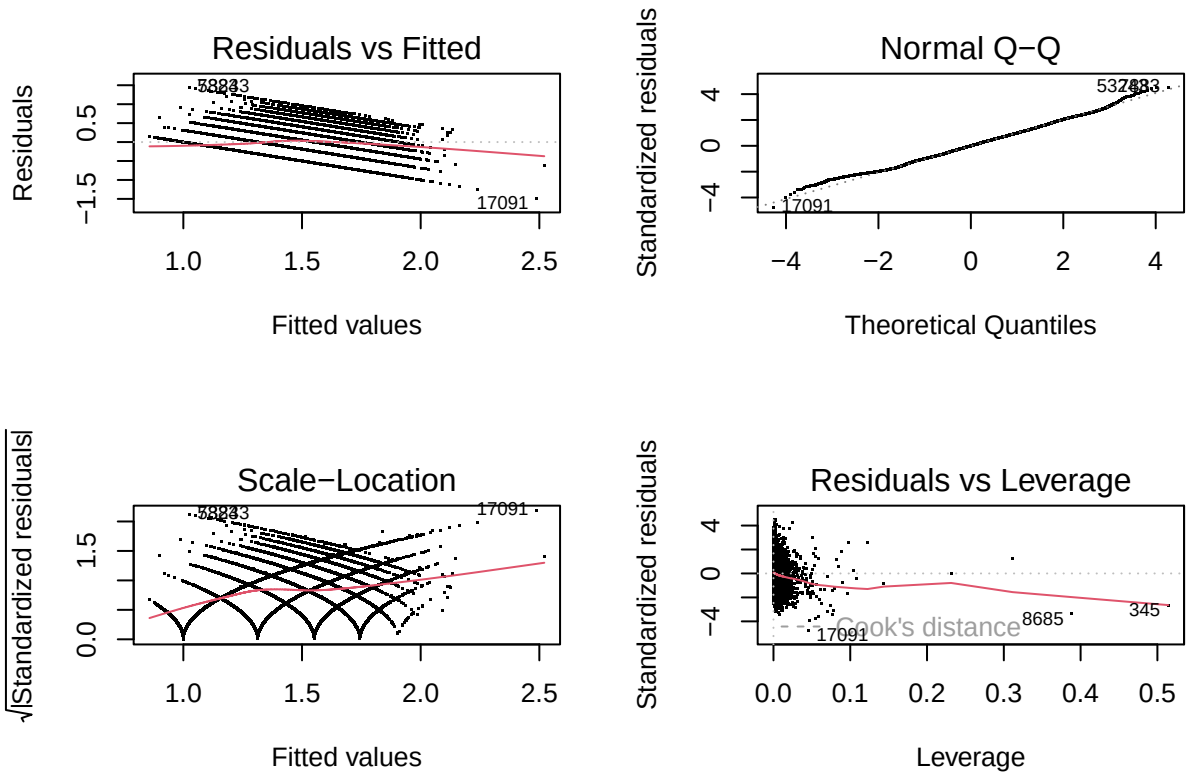
Ostatni model będzie szacował zlogarytmowaną liczbę dzieci powiększoną o liczbę e.

```
train$logCHILDS <- log(train[,'CHILDS']+exp(1))
model.acov3 <-lm(formula = logCHILDS ~ SEX*AGE*EDUC*YEAR*SIBS + religious +
                   SEX*AGE*WRKSTAT + SEX*AGE*EDUC*DEGREE, data = train)

summary(model.acov3)$adj.r.squared
```

```
## [1] 0.2414733
```

```
par(mfrow=c(2,2))
plot(model.acov3, pch = '.')
```



Wydaje się że zlogarytmowanie mogło trochę pomóc, ale zaraz okaże się czy odwracanie tej operacji obarczone błędem implementacji tak skomplikowanych przekształceń nie powiększy i tak niemałych problemów modelu.

Opytmalizacja

Ostatnim krokiem będzie optymalizacja 4 najlepszych modeli za pomocą funkcji step(). W trosce o czytelność pracy zdecydowałem się ukryć kroki optymalizacji.

```
model.opt.all <- step(model.all)
model.opt.acov2 <- step(model.acov2)
model.opt.acov3 <- step(model.acov3)
```

Testy

Test modeli

Najpierw sprawdzimy wartości AIC najlepszych modeli.

```
extractAIC(model.all)[2]
```

```
## [1] 50042.63
```

```
extractAIC(model.opt.all)[2]
```

```
## [1] 50042.63
```

```
extractAIC(model.acov)[2]
```

```
## [1] 51011.35
```

```
extractAIC(model.opt.acov2)[2]
```

```
## [1] 48562.97
```

```
extractAIC(model.acov2)[2]
```

```
## [1] 48574.64
```

```
extractAIC(model.acov3)[2]
```

```
## [1] -117191.9
```

```
extractAIC(model.opt.acov3)[2]
```

```
## [1] -117204.4
```

Rzeczywiście najniższą wartość daje ostatni model, jednak tak duża różnica wydaje się podejrzana. W celu weryfikacji policzymy RMSE dla danych testowych.

```
prediction.all <- as.numeric(predict(model.all, test))
prediction.opt.all <- as.numeric(predict(model.opt.all, test))
prediction.acov <- as.numeric(predict(model.acov, test))
prediction.acov2 <- as.numeric(predict(model.acov2, test))
prediction.opt.acov2 <- as.numeric(predict(model.opt.acov2, test))
prediction.acov3 <- exp(1)^as.numeric(predict(model.acov3, test))-exp(1)
prediction.opt.acov3 <- exp(1)^as.numeric(predict(model.opt.acov3, test))-exp(1)

rmse(prediction.opt.all, expected_results)
```

```
## [1] 1.627192
```

```
rmse(prediction.all, expected_results)
```

```
## [1] 1.627192
```

```
rmse(prediction.acov, expected_results)
```

```
## [1] 1.639767
```

```
rmse(prediction.acov2, expected_results)
```

```
## [1] 1.603901
```

```
rmse(prediction.opt.acov2, expected_results)
```

```
## [1] 1.603777
```

```
rmse(prediction.acov3, expected_results)
```

```
## [1] 1.634304
```

```
rmse(prediction.opt.acov3, expected_results)
```

```
## [1] 1.63408
```

Rzeczywiście ostatni model okazuje się nie być najlepszy, ten sam wzór modelu dla nieprzekształcanej ilości dzieci po zdobywa pierwsze miejsce po optymalizacji, drugie bez optymalizacji funkcją step().

Wnioski:

Pomimo sprawdzenia wielu modeli, predykcje są obarczone dużym błędem. Podjęta próba zlogarytmowania przyniosła przeciwne do oczekiwanych rezultaty.

Wyższe wykształcenie koreluje z mniejszą liczbą potomków.

Uważam, że hipoteza została potwierdzona wykresami analizy, oraz parametrami modeli - odejmując ułamki oczekiwanych dzieci za każdą ukończoną klasę, oraz "premiujących" osoby z niższym wykształceniem.

Brak wyznawanej religii koreluje z mniejszą liczbą potomków.

Uważam, że hipoteza została delikatnie potwierdzona wykresami analizy, oraz parametrami modeli - dodając oczekiwane 0.3 dziecka osobom wyznającym religie.

Aktywność zawodowa koreluje z mniejszą liczbą potomków.

Uważam, że hipotezę należy odrzucić ze względu na brak jednoznacznych wyników analizy danych oraz modeli. Problem jest zbyt skomplikowany i należałoby go rozbić na pomniejsze hipotezy z rozróżnieniem płci, oraz rozbiciem na grupy wiekowe uwzględniające oddzielnie uczniów i emerytów.