Projekt Zaliczeniowy nr 1

Mateusz Kapusta

2022-05-13

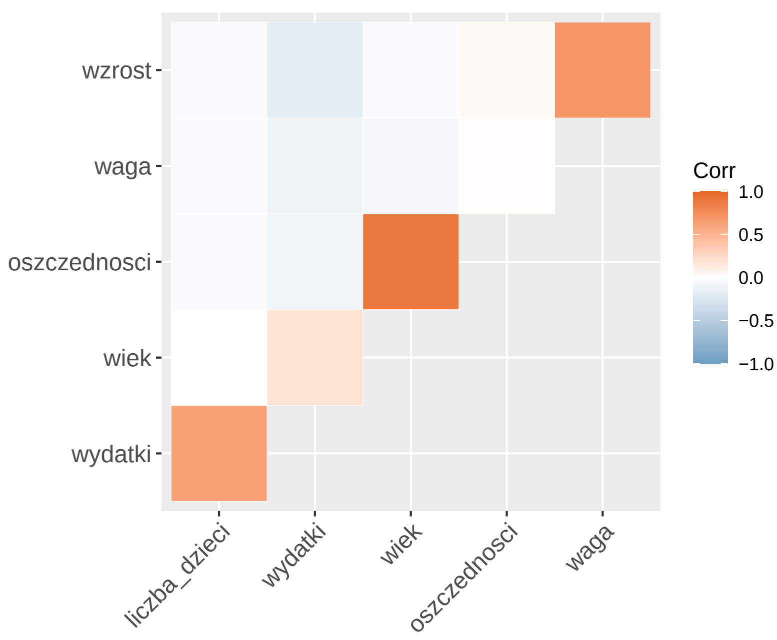
1

Na samym początku wczytujemy dane, które posłużą nam do wykonania modelu.

```
data<-read.csv("people_tab.csv",sep="\t")
num<-lapply(data,is.numeric)
```

Dane składają się z 500 obserwacji natomiast każda obserwacja liczy sobie 9 parametrów z czego 3 to parametry jakościowe. W celu zbadaniu korelacji pomiędzy zmiennymi znajdujemy macierz korelacji metodą Pearsona.

```
data_ilo<-data[unlist(num)]
cor_matrix<-cor(data_ilo)
ggcorrplot(cor_matrix, hc.order = TRUE, type = "upper",
   outline.col = "white",
   ggtheme = ggplot2::theme_grey,
   colors = c("#6D9EC1", "white", "#E46726"),
   insig = "blank")
```



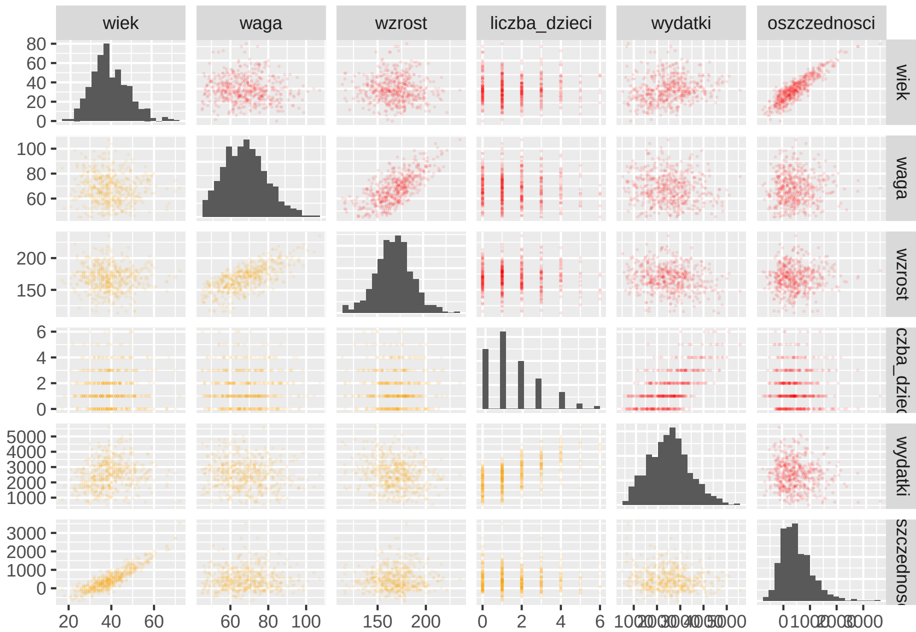
Widzimy, że największe dodatnie korelacje zachodzą pomiędzy wiekiem a oszczędnościami, wydatkami a liczbą dzieci oraz wzrostem a wagą. Brakuje natomiast nam silnych ujemnych korelacji pomiędzy danymi. Teraz zbadajmy korelacje pomiędzy zmiennymi jakościowymi. Do tego celu wykorzystamy korelację polichoryczną z pakietu polycor.

```
library(polycor)
data_jako<-na.omit(data[!unlist(num)]) #usuwamy pola bez wartości
c1<-polychor(data_jako$budynek,data_jako$plec)
c2<-polychor(data_jako$budynek,data_jako$stan_cywilny)
c3<-polychor(data_jako$stan_cywilny,data_jako$plec)
```

Korelacje pomiędzy parami zmiennych budynek-płeć, budynek-stan cywilny, stan cywilny-płeć wynoszą kolejno -0.0226709, -0.035394, -0.1210071. W przypadku zmiennej płeć mamy braki w danych, które na samym początku usuwamy.

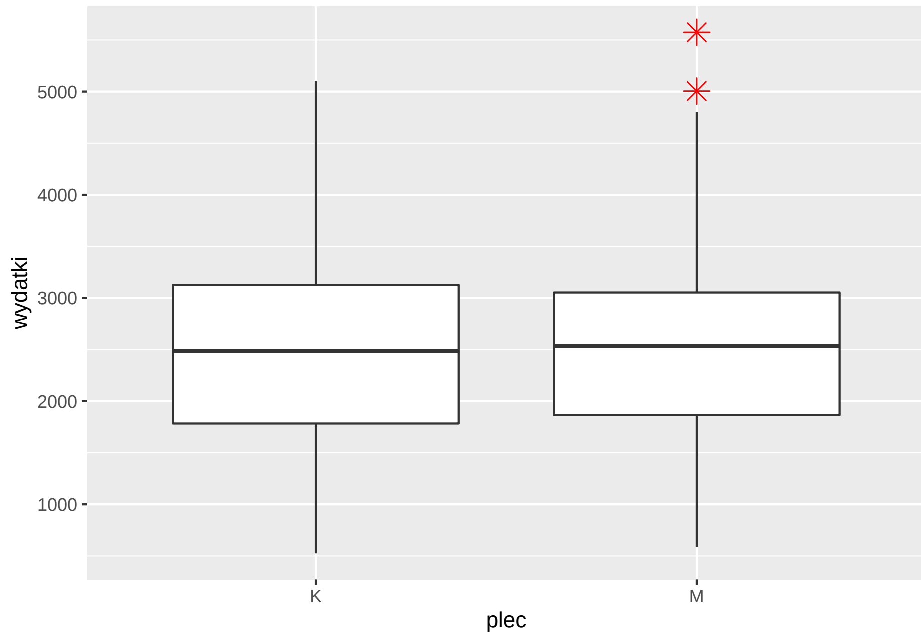
2

```
ggpairs(data_ilo,
        upper= list(continuous = wrap("points",color="red",size=0.1,alpha=1/10), combo = "box_no_facet")
        lower=list(continuous=wrap("points",color="orange",size=0.1,alpha=1/10)),
        diag=list(continuous=wrap("barDiag",bins=20))
        )
```



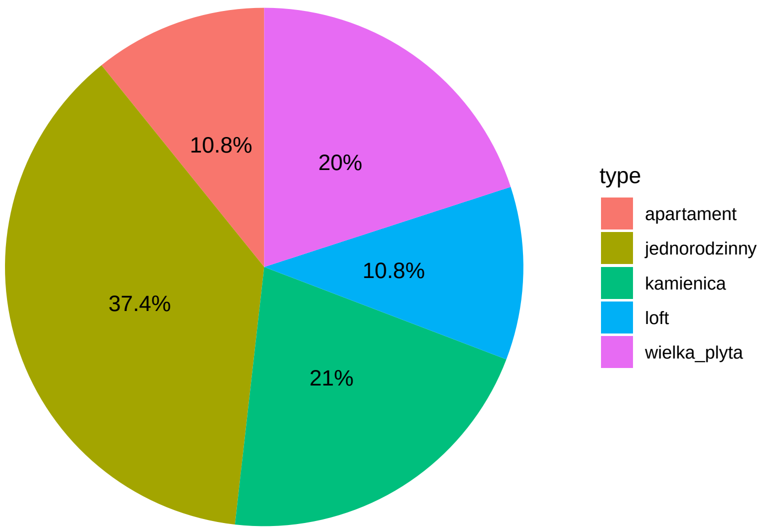
Wykres przedstawiający wydatki respondentów ze względu na płeć:

```
ggplot(na.omit(data))+geom_boxplot(aes(x=plec,y=wydatki),outlier.colour="red", outlier.shape=8,
                outlier.size=4)
```



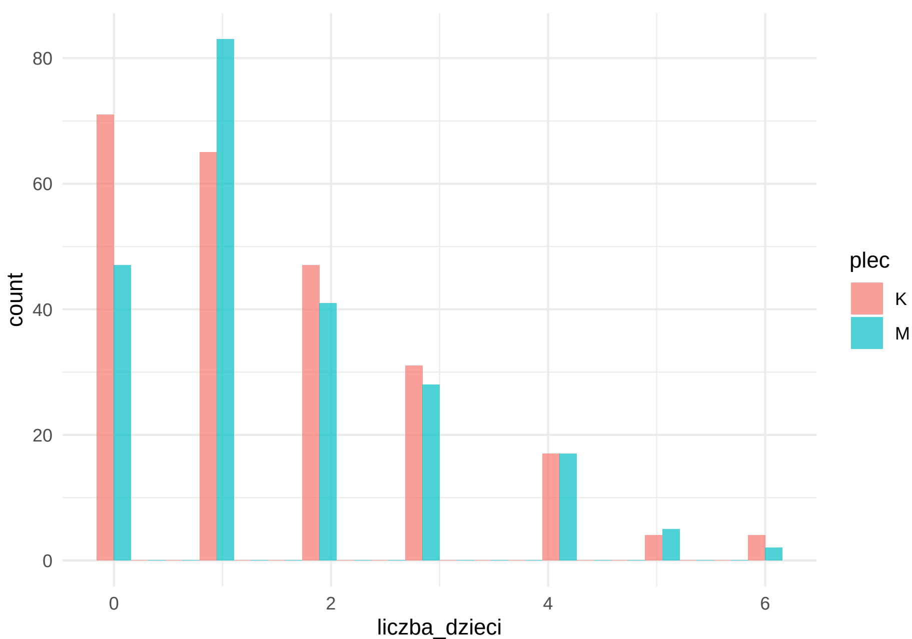
Widzimy, że męszczyźni średnio wydają więcej jednakże u kobiet mamy do czynienia z większym rozrzutem danych. Wykres kołowy przedstawiający rozkład osób miekszających w różnego typu budynkach:

```
data$budynek=as.factor(data$budynek)
temp=count(data$budynek) #korzystamy z bublioteki plyr
d=data.frame("val"=temp$freq,"type"=temp$x,"labels"=paste0(temp$freq/sum(temp$freq)*100,"%"))
ggplot(d,aes(x="",y=val,fill=type))+
  geom_bar(stat="identity",width=1)+
  coord_polar("y",start=0)+
  geom_text(aes(label = labels),
            position = position_stack(vjust = 0.5)) +
  theme_void()
```



Na koniec sprawdźmy jak rozkłada się liczba dzieci pośród naszych respondentów z podziałem na płeć.

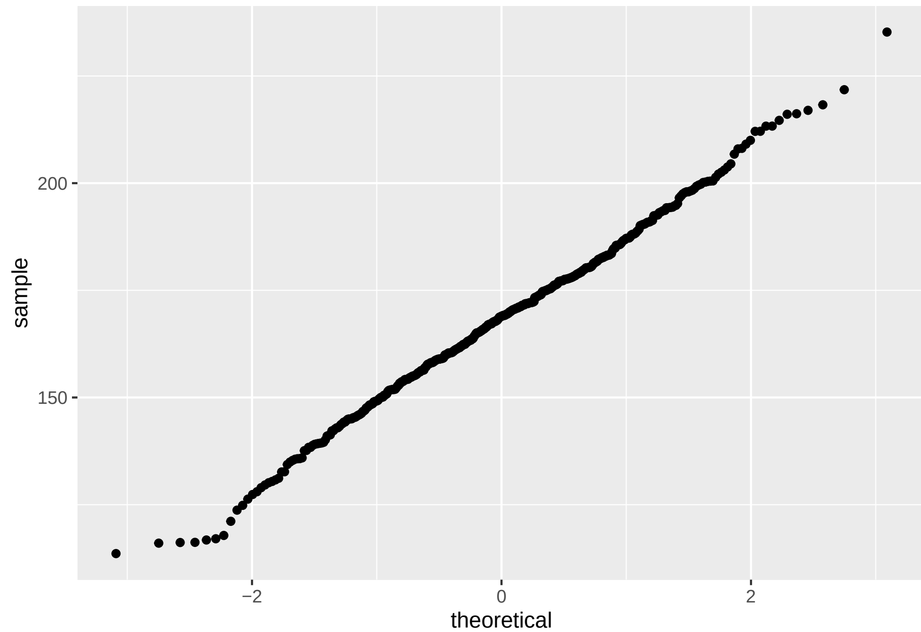
```
ggplot(na.omit(data), aes(x = liczba_dzieci, fill = plec)) +
  geom_histogram(position = "dodge", alpha = 0.7, bins = 20)+
  theme_minimal()
```



3

Rozważmy teraz jaka jest p-wartość dla hipotezy, że średnia wzrostu to $m = 170$ cm. Wpierw zobaczymy jak rozkłada się wzrost wśród danych przy pomocy wykresu kwantylowego.

```
ggplot(data)+stat_qq(aes(sample=wzrost))+theme_grey()
```



Widzimy więc, że z bardzo dobrym przybliżeniem dane pochodzą z rozkładu normalnego. Do sprawdzenia hipotezy zerowej wystarczy wykorzystać test t-studenta. Hipotezą zerową jest to, że dane pochodzą z rozkładu normalnego o średniej 170 cm natomiast hipotezą alternatywną że średnia jest mniejsza.

```
mu_hip<-170 #średnia wartość wzrostu według hipotezy zerowej
med_hip<-165 #mediana wzorsty według hipotezy zerowej
x<-t.test(data$wzrost,mu=mu_hip,alternative="less")
```

Widzimy, że p-wartość wynosi 0.019487 a więc na poziomie istotności $0,05$ hipotezę zerową należy odrzucić. Aby przetestować medianę wykrozystamy test jednopopulacyjny Wilcoxona. Niech na moment alternatywą będzie to, że mediana jest inna. Wtedy

```
y<-wilcox.test(data$wzrost,mu=med_hip)
```

.Odpowiadająca testow p-wartość to 2.6477538×10^{-4} a więc na poziomie istotności $0,05$ należy odrzucić hipotezę zerową. Jeżeli chcemy sprawdzić czy mediana jest mniejsza od 165 skorzystamy z testu jednostronnego.

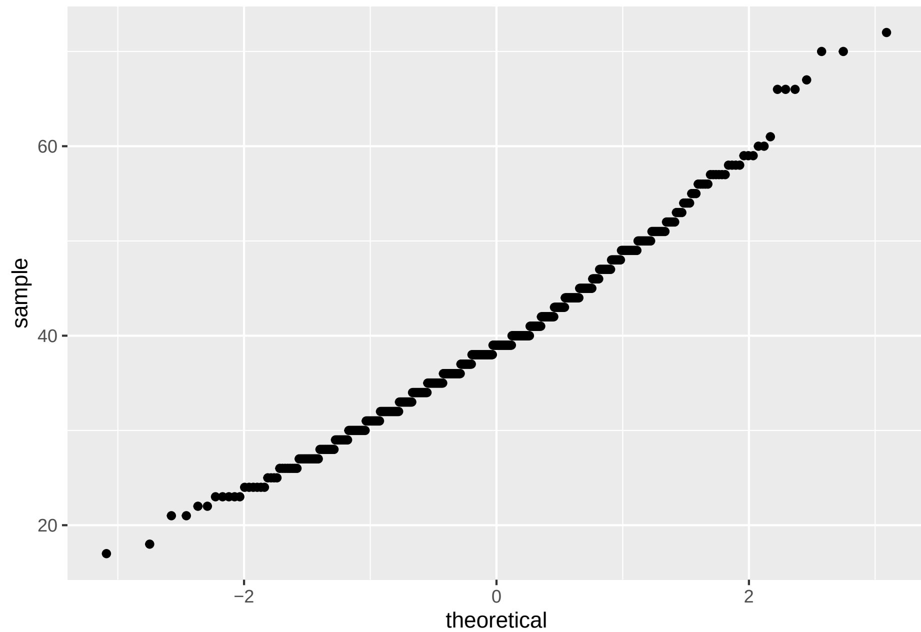
```
y<-wilcox.test(data$wzrost,mu=med_hip,alternative="less")
```

W tym przypadku nasza p-wartośc to 0.9998678 a więc hipotezy nie należy odrzucić. W rzeczywistości widzimy, że mediana rozkładu jest zdecydowanie większa niż 165.

4

Przejdźmy teraz do obliczenia przedziałów ufności dla parametrów na poziomie $0,99$. Zanim przejdziemy do wzorów szybko rzućmy okiem na rozkład kwantylowy danych.

```
ggplot(data)+stat_qq(aes(sample=wiek))+theme_grey()
```



Dane pochodzą z grubsza z rozkładu normalnego. W przypadku średniej i danych z rozkładu normalnego wiemy, że

$$T = \frac{X - \mu}{S}\sqrt{N} \tag{1}$$

ma rozkład t-studenta (X oznacza średnią populacji a S odchylenie standardowe uzyksane estymatorem nieobciążonym). Chcemy zbadać, jaki jest przedział ufności dla statystyki T. wykorzystujac funkcje R mamy, że

```
a<-0.01
c<-qt(1-a/2,df=length(data$wiek)-1)
```

Jeżeli T mieści się pomiędzy c a $-c$ to μ musi się mieścić pomiędzy $X - \frac{cS}{\sqrt{N}}$ oraz $X + \frac{cS}{\sqrt{N}}$.

```
up<-mean(data$wiek)+c*sd(data$wiek)/sqrt(length(data$wiek))
down<-mean(data$wiek)-c*sd(data$wiek)/sqrt(length(data$wiek))
```

Stąd przedział ufności dla μ to 40.5219967 do 38.4460033. W celu wyznaczenia przedziałów ufności dla wariancji wykorzystamy podobną metodę z tą różnicą, że zamiast wykorzystywać statystykę t studenta wykorzystamy statystykę χ^2. Wiemy albowiem, że statystyka

$$\frac{(N-1)S}{\sigma^2} \tag{2}$$

ma rozkład χ^2 o $N - 1$ stopniach swobody. Stąd analogicznie znajdujemy wartości przedziałów dla statystyki

```
p<-qchisq(1-a/2,df=length(data$wiek)-1)
l<-qchisq(a/2,df=length(data$wiek)-1)
```

i po transformacjach znajdujemy jakie są przedziały ufności dla wariancji:

```
N<-length(data$wiek)
lv<-(N-1)/p*var(data$wiek)
pv<-(N-1)/l*var(data$wiek)
```

.Przedział ufności dla odchylenia standardowego to pierwiastek z tych granic a więc rozprzestrzenia się od 8.2965437 do 9.7681131. Aby zbadać przedziały ufności dla kwantyli wykorzystajmy metodę dokładną przeszukującą wektor obserwacji nie zakładając symetryczności rozkładu zaimplementowaną w bibliotecie MKmisc (polecaną przez autorów artukułu podlinkowanego w treści zadania zaliczeniowego).

```
library(MKmisc)
kwant<-c(1/4,1/2,3/4)
przed<-sapply(kwant,quantileCI,x=data$wiek, conf.level = 1-a, method = "exact",minLength = TRUE)
przed<-sapply(0:2,\(x) przed[[3*x+2]])
```

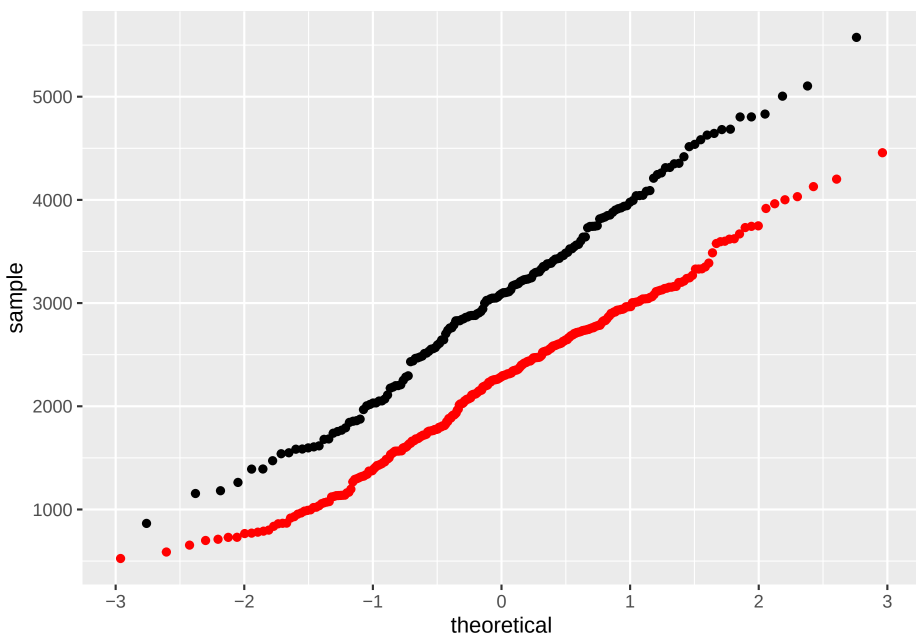
Otrzymane przedziały dla naszych kwantyli to kolejno 32-35, 38-40, 43-47. Obliczenie przedziałów ufności dla średniej oraz odchylenia standardowego został obliczony przy założeniu, że zmienna wiek ma rozkład normalny co jak widać z wykresu kwantylowego jest sensownym założeniem.

5

1

Sprawdźmy, czy różnica pomiędzy wydatkami osób w związku małżeńskim a singlami jest statystycznie różna. Wpierw przygotujmy dane i sprawdzimy czy pochodzą one z rozkładu normalnego wykorzystując wykres kwantylowy.

```
a<-0.01
marriage<-data$wydatki[data$stan_cywilny]
single<-data$wydatki[!data$stan_cywilny]
ggplot()+stat_qq(aes(sample=marriage),data=data.frame("marriage"=marriage))+
  stat_qq(aes(sample=single),color="red",data=data.frame("single"=single))+theme_grey()
```

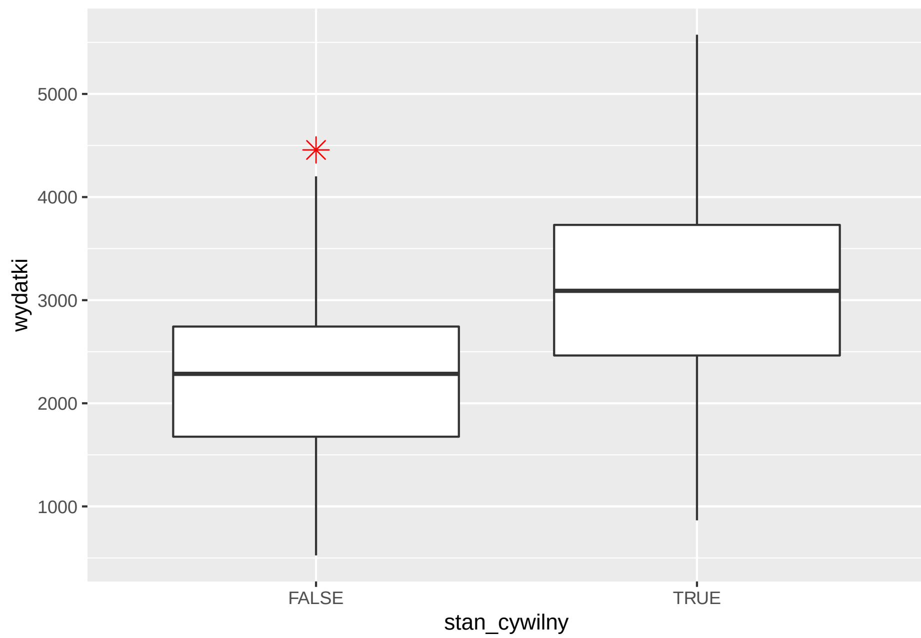


Ponieważ obie zmienne w przybliżeniu pochodzą z rokładu normalnego co objawia się ładną zależnością liniową danych na wykresie kwantylowym to wykorzystamy test t studenta dla dwóch populacji o różnej wariancji (test Welsha). Według hipotezy zerowej zarobki w obu populacjach są identyczne.

```
test_t<-t.test(marriage,single,var.equal=FALSE)
```

p-wartość dla naszego testu to 1.165043×10^{-20} co przy poziomie istotności 0.01 sugeruje że istnieje znaczna różnica pomiędzy danymi i hipotezę zerową należy odrzucić. Możemy zwizualizować nasze dane:

```
ggplot(data)+geom_boxplot(aes(x=stan_cywilny,y=wydatki),outlier.colour="red", outlier.shape=8,
                outlier.size=4)
```



Jedyny wykorzystanym założeniem jest normalność obu populacji co jak widzimy jest dobrze spełnione.

2

Zastanówmy się, czy istnieje zależność pomiędzy wydatkami oraz oszczędnościami. W tym celu skorzystamy z testu ρ Spearmana. Policzmy współczynnik korelacji. Według hipotezy zerowej zmienne te są nieskorelowane.

```
p<-cor.test(data$wydatki,data$oszczednosci,method="spearman",exact=FALSE)
```

p-wartość odpowiadająca naszemu testowi to 0.0775552 więc na poziomie $0,05$ nie możemy odrzucić hipotezy zerowej. Korelacja ρ Spearmana mierzy nam czy pomiędzy zmiennymi istnieje korelacja monotoniczna (w jedną stronę lub nie) w przeciwieństwie do korelacji Pearsona mierzacej liniowość relacji. Pakiet R wykorzystuje asymptotyczną zbieżność do rozkładu t odpowiedniej statystyki co daje tym lepsze wyniki im więcej danych wykorzystujemy co przy 500 obserwacjach powinno dać dobre wyniki.

3

Zbadajmy, czy stan cywilny jest niezależny od płci, według hipotezy zerowej zmienne te są niezależne. W tym celu wykorzystamy dokładny test Fishera, który nie wymaga od danych rzadnych dodatkowych założeń. Wpierw musimy znaleźć macierz mówiącą nam, ile razy sklasyfikowane zostały poszczególne obserwacje.

```
a<-nrow(data[data$stan_cywilny==TRUE & data$plec=="M",]) # żonaci mężczyźni
b<-nrow(data[data$stan_cywilny==FALSE & data$plec=="M",]) #mężczyźni singlowie
c<-nrow(data[data$stan_cywilny==FALSE & data$plec=="K",]) # samotne kobiety
d<-nrow(data[data$stan_cywilny==TRUE & data$plec=="K",]) # zamężne kobiety
mat<-matrix(c(a,b,d,c),ncol=2,byrow=TRUE)
x<-fisher.test(mat)
```

p-wartość naszego testu to 0.1471494 a więc na żądanym poziomie istotności nie można stwierdzić, że istnieje zależność pomiędzy płcią a stanem cywilnym.

4

Sprawdźmy, czy prawdą jest że liczba dzieci w parach z przynajmniej jednym dzieckiem pochodzi z rozkładu geometrycznego o parametrze prawdopodobieństwa $p = 0.4428755$ który jest przycięty od jeden do sześciu (prawadopodobieństwo proporcjonalne do $(1-p)^{(x-1)}$). Liczba ta jest nieprzypadkowa i bierze się ona z faktu, że gdyby dane pochodziły ze zwykłego rozkładu geometrycznego to odwrotność wartości oczekiwanej równa się p a więc estymujemy p jako odwrotność średniej z danych (ponieważ rozkład jest ucięty zmniejszamy tą wartość o 8%). Wpierw napiszmy funkcję która pozwoli na odpowiednie samplowanie oraz zwróci gęstość prawdopodobieństwa.

```r
dtran_geom<-function(x,p,dol,up)
{
  suma<-sum(sapply(dol:up,\(c) (1-p)^(c-1)))
  (1-p)^(x-1)/suma
}

rtran_geom<-function(n,p,dol,up)
{
  out<-rep(0,n)
  i<-1
  while (i<=n)
  {
    t<-rgeom(1,p)
    if (t>=dol & t<=up)
    {
      out[i]<-t
      i<-i+1
    }
  }
  out
}
```

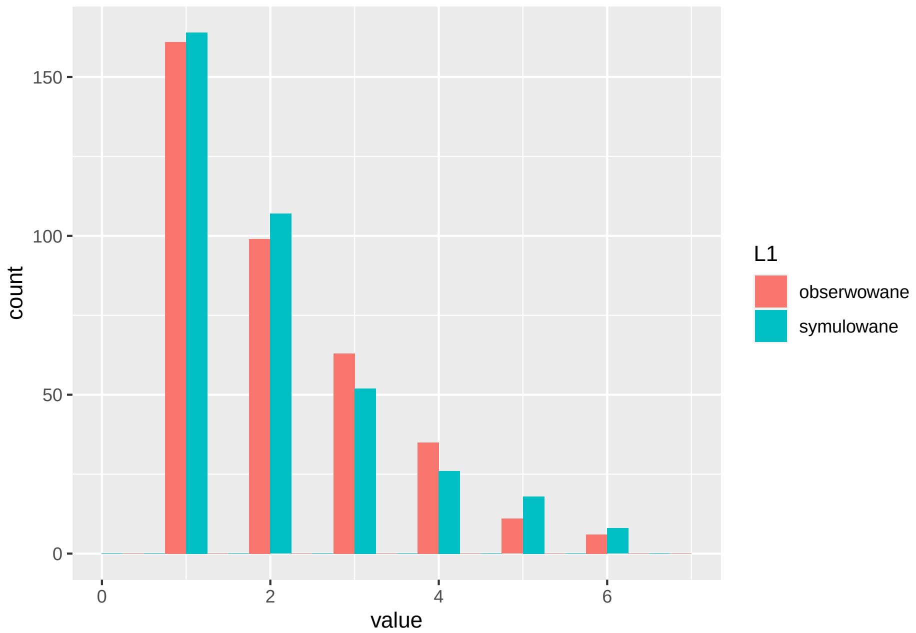
Kiedy mamy zdefiniowane nasze rozkłady to możemy wykorzystać test χ^2 zgodności z rozkładem. Mamy 6 kategorii i stąd

```r
p<-1/mean(data$liczba_dzieci[data$liczba_dzieci>0])*0.92
liczba<-length(data$liczba_dzieci[data$liczba_dzieci>0])
observed<-sapply(1:6,\(x) length(data$liczba_dzieci[data$liczba_dzieci==x]))
theoretical<-sapply(1:6,\(x) dtran_geom(x,p,1,6))
test<-chisq.test(x=observed,
                 p=theoretical,
                 correct=TRUE)
```

. Otrzymana tą drogą p-wartość to 0.2609816 co sugeruje, iż hipotezy zerowej nie powinniśmy odrzucić na rządanym poziomie istotności. Na koniec zobaczmy jak wygląda histogram liczby dzieci wraz z porównaniem z rozkładem prawdopodobieństwa.

```r
library(reshape2)
simple<-list("symulowane"=rtran_geom(liczba,p,1,6),"obserwowane"=data$liczba_dzieci)
ggplot(melt(simple), aes(value, fill = L1)) +
geom_histogram(position = "dodge", bins=15) +
  xlim(0,7)
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Widzimy więc, że model przeszacowuje liczbę dzieci dla małżeństw z jednym dzieckiem natomiast w pozostałych przypadkach zachowuje się z grubsza dobrze.

6

Stwórzmy podstawowy model, wykorzystujący wszystkie zmienne do objaśnienia oszczędności.

```
data<-na.omit(data)
model<-lm(oszczednosci~.,data=data)
su<-summary(model)
su
```

```
##
## Call:
## lm(formula = oszczednosci ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -307.64  -60.13   -1.69   58.06  462.98
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -873.59106   58.76098 -14.867  < 2e-16 ***
## wiek                    63.94258    0.56712 112.750  < 2e-16 ***
## waga                     3.94409    0.56935   6.927 1.49e-11 ***
## wzrost                  -2.38464    0.35204  -6.774 3.94e-11 ***
## plecM                    1.38069    9.63179   0.143    0.886
## stan_cywilnyTRUE        -4.61252   12.91187  -0.357    0.721
## liczba_dzieci          151.60355    6.15687  24.623  < 2e-16 ***
## budynekjednorodzinny -182.07031   16.43991 -11.075  < 2e-16 ***
## budynekkamienica     -305.63144   17.89020 -17.084  < 2e-16 ***
## budynekloft          -338.47001   25.14078 -13.463  < 2e-16 ***
```

```
## budynekwielka_plyta  -564.26015   20.59225 -27.402  < 2e-16 ***
## wydatki                -0.39593    0.01057 -37.455  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102 on 450 degrees of freedom
## Multiple R-squared:  0.9673, Adjusted R-squared:  0.9665
## F-statistic:  1209 on 11 and 450 DF,  p-value: < 2.2e-16
```

Oszacowanie współczynników możemy odczytać z podsumowania modelu, wartość R^2 w naszym modelu to 0.9672621 natomiast RSS wynosi 4.6813339×10^6. Widzimy, że p-wartości odpowiadające poszczególnym parametrom są bardzo małe za wyjątkiem współczynników odpowiadających płci oraz stanu cywilnemu. Skonstruujmy nowy model w którym wykorzystamy wszystkie współczynniki poza jednym i zbadamy jak zmieniają się parametry. Poniższy fragment kodu kolejno wyświetla nazwę zmiennej, jaki jest parametr R^2 bez tej zmiennej, jaki jest RSS bez niej oraz jaka jest o-wartość danego współczynnika.

```
for (name in colnames(data)[1:8])
{
  modelp<-lm(paste("oszczednosci~.-",name),data=data)
  print(c(name, sprintf("%.4f",summary(modelp)$r.squared),
          sprintf("%.f",deviance(modelp)),su$coefficients[match(name,colnames(data)),4]))
}
```

```
## [1] "wiek"                 "0.0424"               "136929242"
## [4] "5.90395298760922e-41"
## [1] "waga"    "0.9638"  "5180546" "0"
## [1] "wzrost"               "0.9639"               "5158668"
## [4] "1.49011162168797e-11"
## [1] "plec"                 "0.9673"               "4681548"
## [4] "3.93728564678186e-11"
## [1] "stan_cywilny"      "0.9673"            "4682661"
## [4] "0.886080559289676"
## [1] "liczba_dzieci"     "0.9232"            "10988814"
## [4] "0.721086110052039"
## [1] "budynek"              "0.9045"              "13661076"
## [4] "2.0613873679291e-85"
## [1] "wydatki"              "0.8652"               "19275685"
## [4] "2.25639059750048e-25"
```

Widzimy wyraźnie, że wyeliminowanie płci lub stanu cywilnego daje najmniejszą zmianę współczynników (oraz odpowiadają im największe p-wartości). Dlatego też podjęto dezycję o wyeliminowaniu płci. W sten sposób otrzymujemy nowy model.

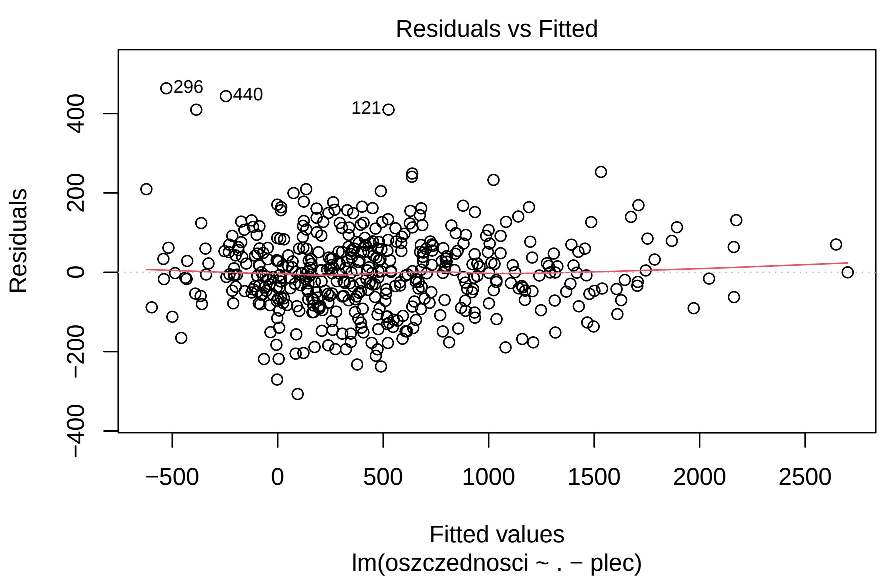
```
modelprim<-lm(oszczednosci~.-plec,data=data)
summary(modelprim)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - plec, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -306.83  -60.62   -1.79   58.67  463.67
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          -872.88762   58.49209 -14.923  < 2e-16 ***
## wiek                   63.94122    0.56642 112.886  < 2e-16 ***
## waga                    3.94887    0.56776   6.955 1.24e-11 ***
## wzrost                 -2.38524    0.35163  -6.783 3.70e-11 ***
## stan_cywilnyTRUE       -4.83341   12.80566  -0.377    0.706
## liczba_dzieci         151.69318    6.11838  24.793  < 2e-16 ***
## budynekjednorodzinny -182.14265   16.41431 -11.097  < 2e-16 ***
## budynekkamienica     -305.65338   17.87011 -17.104  < 2e-16 ***
## budynekloft          -338.69765   25.06331 -13.514  < 2e-16 ***
## budynekwielka_plyta  -564.40762   20.54419 -27.473  < 2e-16 ***
## wydatki                -0.39600    0.01055 -37.548  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101.9 on 451 degrees of freedom
## Multiple R-squared:  0.9673, Adjusted R-squared:  0.9665
## F-statistic:  1332 on 10 and 451 DF,  p-value: < 2.2e-16
```

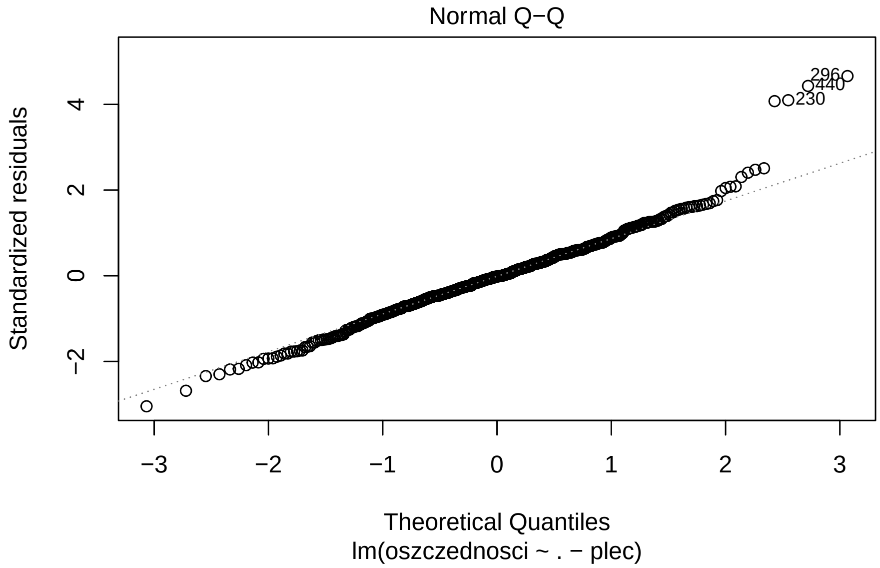
W celu zbadania założeń LINE zbadajmy wykresy diagnostyczne.

```
plot(modelprim, which=1)
```



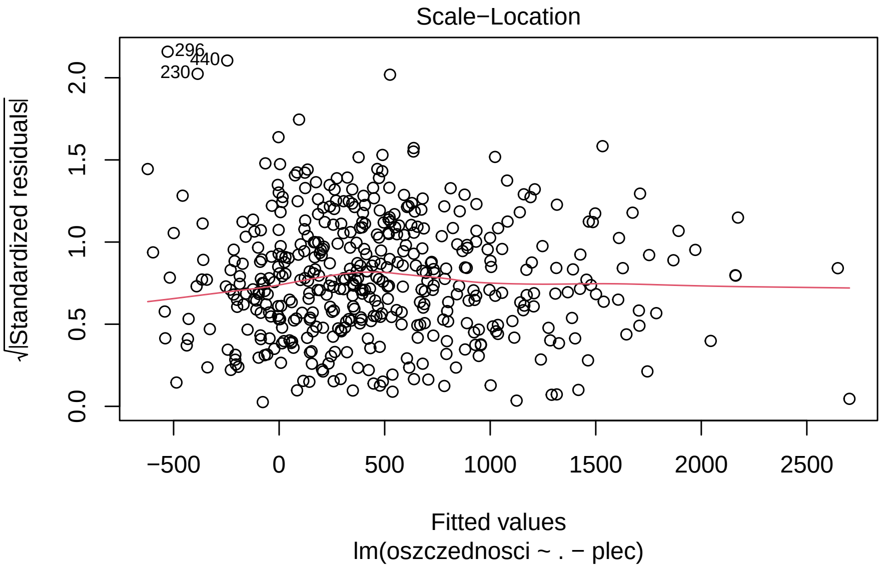
Wykres residuów w zależności od dopasowywanej wartości pokazuje nam, że trend z bardzo dużą dokładnością jest liniowy i nie potrzebuje on jakichkolwiek przekształceń aby był doprze opisywany przez zależność liniową.

```
plot(modelprim, which=2)
```



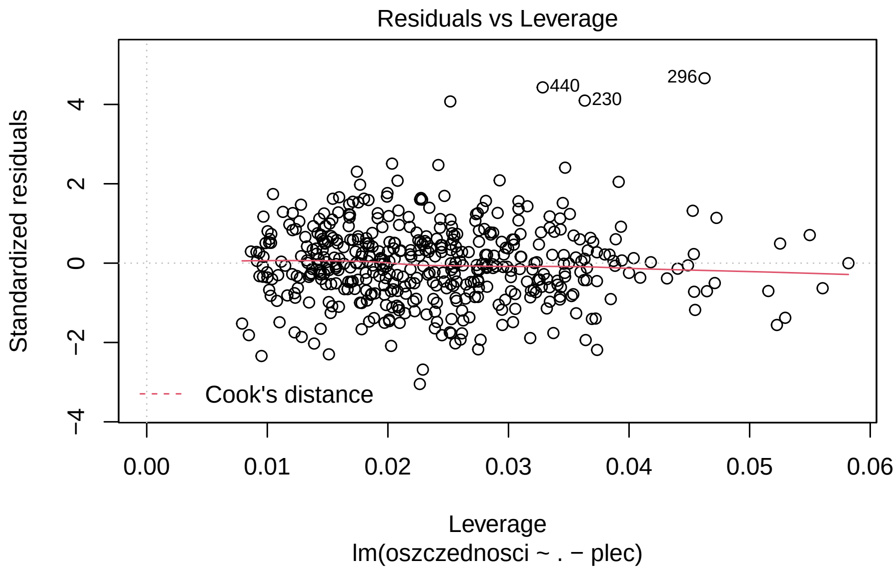
Wykres kwantylowy dla naszych residuów także pokazuje, że z bardzo dużą dokładnością nasze residua pochodzą z rozkładu normalnego.

```
plot(modelprim,which=3)
```



Wykres zależności pierwiasta z standaryzowanych residuuów w zależności od predykowanej wartości mówi nam, że błędy jakie popełniamy są hemoskedastyczne.

```r
plot(modelprim,which=5)
```



Ostatni wykres pozwala zidentyfikować obserwacje o dużej dźwigni. Wykres pozwala ustalić, że pomiary o numerach 440, 230 oraz 296 mają nadzwyczejnie duże odchylenie i można rozważyć ich usunięcie albowiem mają dośc duży wpływ na współczynniki. Widzimy więc że dane dobrze opisywane są przez model liniowy oraz potrzebne założenia LINE są spełnione.