Projekt1BognaPawlus

Bogna Pawlus

2022-05-01

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##    wiek waga wzrost plec stan_cywilny liczba_dzieci       budynek wydatki
## 1    25 61.7 121.12 <NA>        FALSE             2          loft 1662.91
## 2    37 63.9 145.00    M         TRUE             6  wielka_plyta 4041.86
## 3    41 50.2 145.03    K         TRUE             2    apartament 3853.45
## 4    43 72.4 179.90    M        FALSE             1  wielka_plyta 2398.88
## 5    26 78.4 163.91    M        FALSE             1    apartament 2344.45
## 6    49 59.4 151.86    K         TRUE             2          loft 1967.87
## 7    27 67.5 169.31    M        FALSE             1 jednorodzinny 2249.04
## 8    49 82.3 179.17    K        FALSE             0  wielka_plyta 1775.58
## 9    38 64.1 138.37    K         TRUE             5  wielka_plyta 3382.54
## 10   33 77.4 193.44    M        FALSE             2    apartament 2761.55
## 11   44 73.1 183.18    M        FALSE             3    apartament 3623.29
## 12   27 84.4 192.40    M         TRUE             4  wielka_plyta 1875.45
##    oszczednosci
## 1         23.44
## 2         96.84
## 3        312.68
## 4        447.43
## 5        -78.23
## 6       1241.98
## 7       -211.43
## 8        793.16
## 9        486.20
## 10       288.45
## 11       636.00
## 12        38.54
```

1., 2. Podsumowanie danych

W danych mamy 6 zmiennych ilościowych i 3 jakościowe, zawierające 500 obserwacji. Nie mamy danej płci dla 38 obserwacji

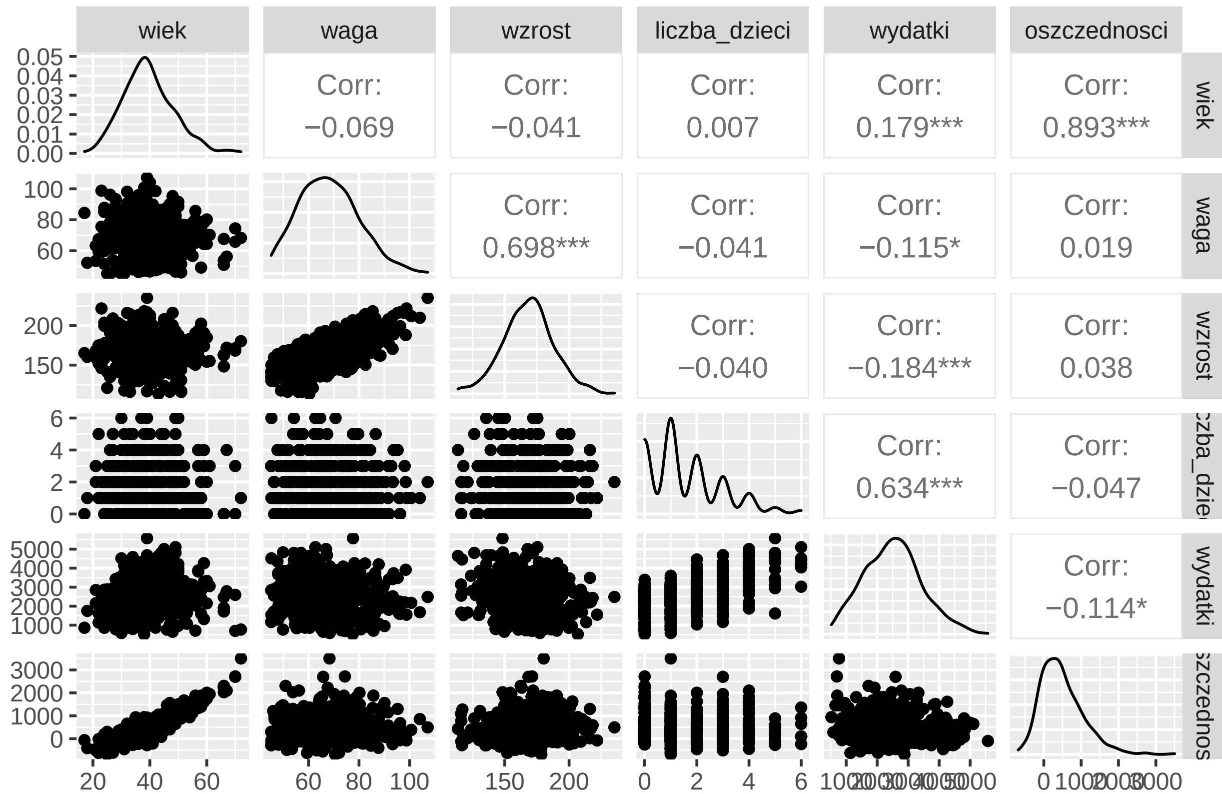
```
sum(is.na(people.tab$plec) == TRUE)
```

```
## [1] 38
```

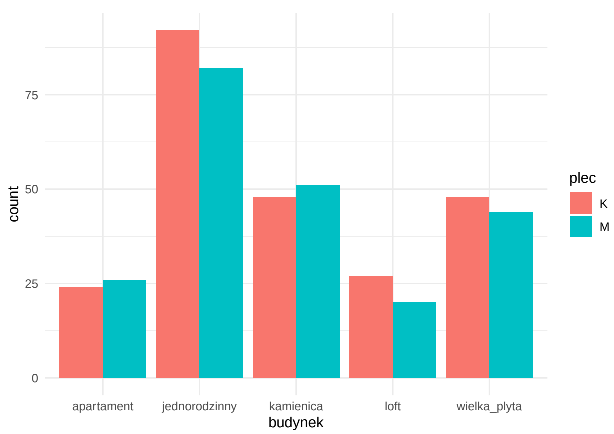
Z wykresu 1. widzimy, że bardzo wysoka korelacja występuje między zmiennymi ‘oszczednosci’ i ‘wiek’, a wysoka korelacja występuje pomiędzy zmiennymi ‘wzrost’ i ‘waga’, ‘wydatki’ i ‘liczba dzieci’.

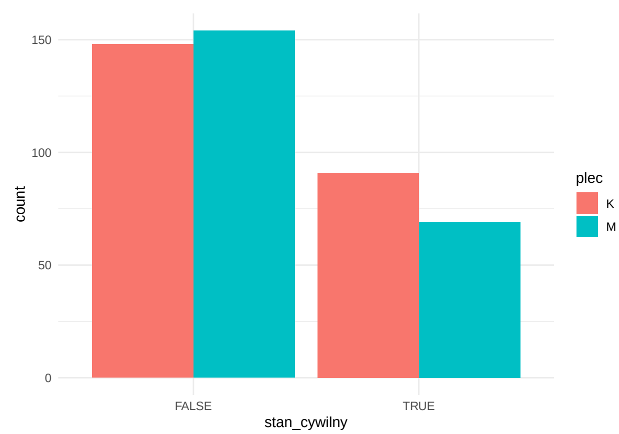
Zależności pomiędzy zmiennymi ilościowymi objaśniającymi:

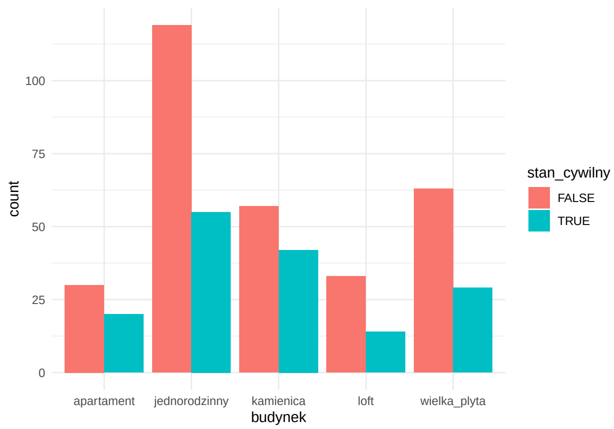
WYKRES 1: scatterplot dla zmiennych ilościowych i objaśnianej



Zależności pomiędzy zmiennymi jakościowymi możemy mniej więcej ocenić z poniższych wykresów

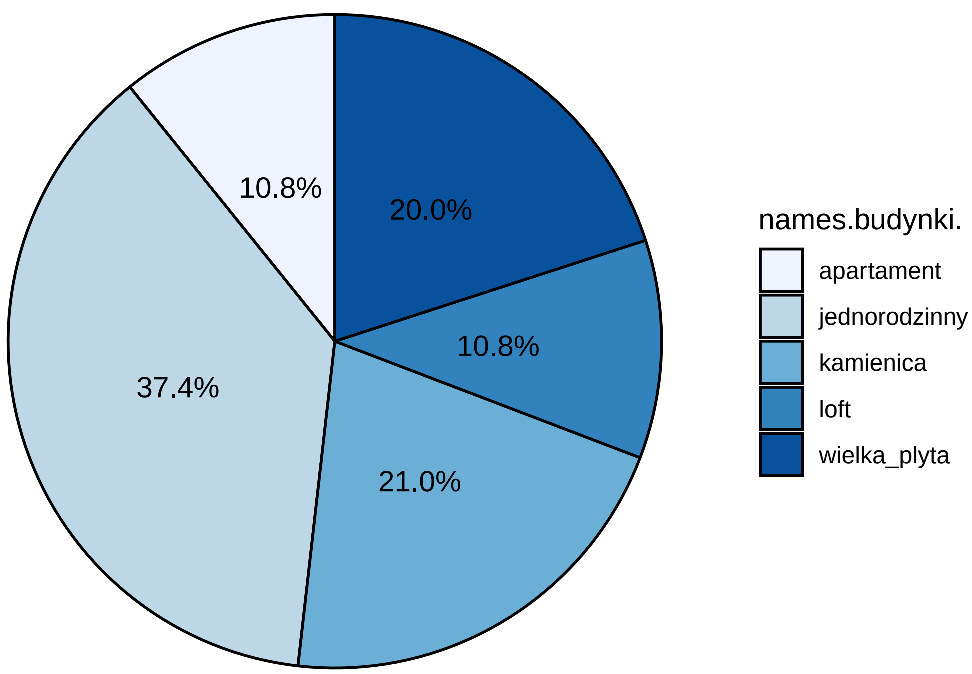


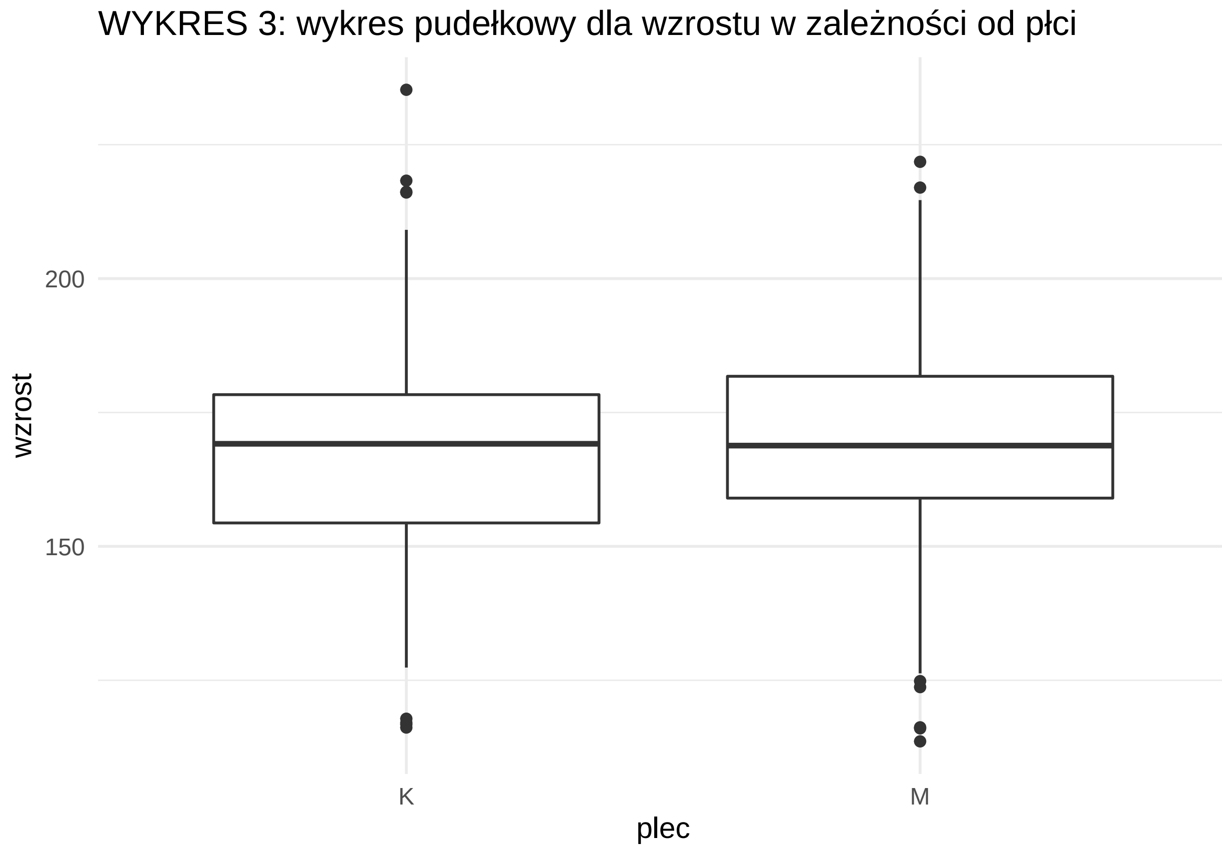




z których mniej więcej widać, że budynek jest zależny od stanu cywilnego (duża różnica w domach jednorodzinnych), a płeć od stanu cywilnego i płeć od budynku są mniej zależne.

WYKRES 2: piechart dla liczby mieszkańców w budynkach





3. p-wartości dla wzrostu

Rozważamy hipotezy dla zmiennej $X =$'wzrost'.

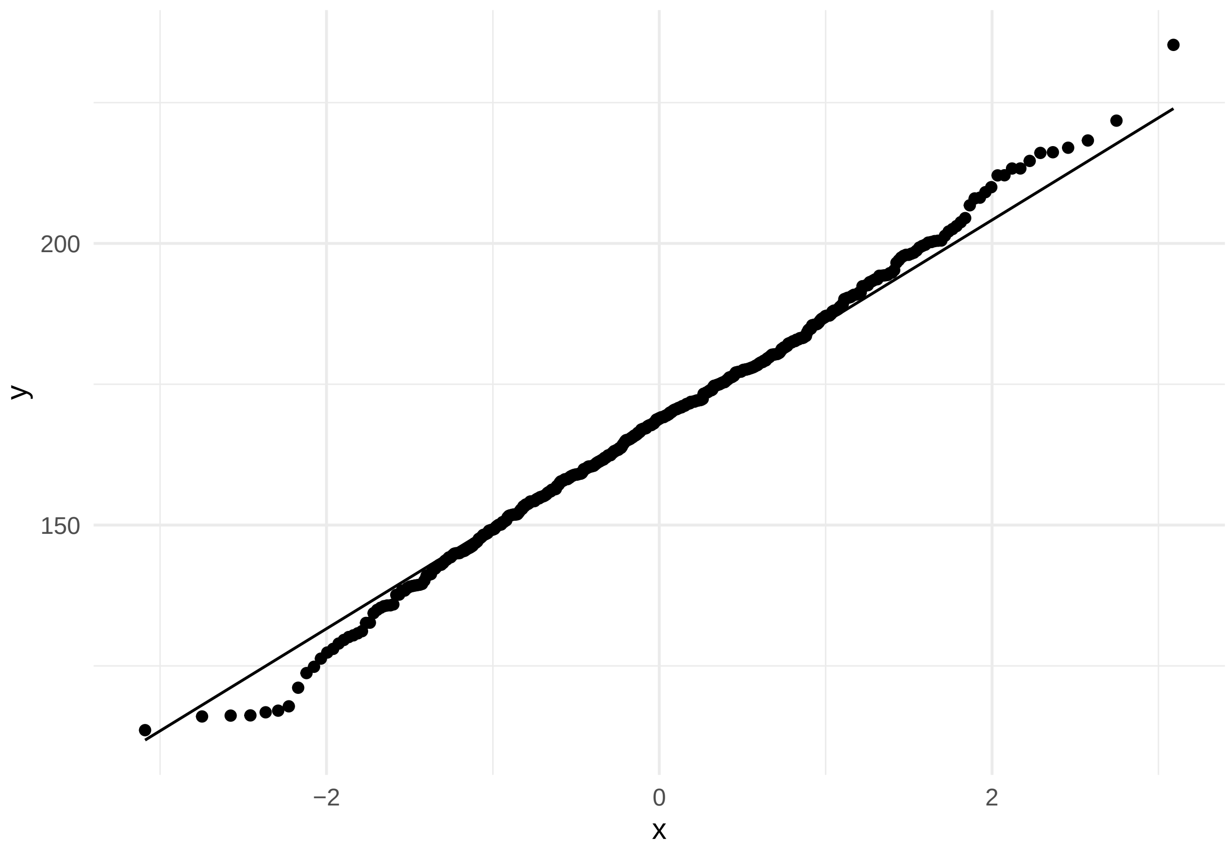
$$H_0 : m = 170 = m_0$$
$$H_1 : m < 170 = m_0$$

Na podstawie wykresu kwantylowego widzimy, że rozkład wzrostu niewiele odbiega od rozkładu normalnego. Ponieważ nie znamy wariancji wzrostu, więc do testu użyjemy statystyki $\frac{\overline{X}-m_0}{S_n}\sqrt{n-1}$ o rozkładzie $t(n-1)$

```
summary(wzrost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   113.6   155.6   169.0   168.2   180.1   235.2
```

```
ggplot() + stat_qq(aes(sample = wzrost)) + stat_qq_line(aes(sample = wzrost)) + theme_minimal()
```



Na podstawie kodu

```
sn = sqrt(1/500*sum((wzrost - mean(wzrost))^2))
stat = (mean(wzrost) - 170)/(sn)*sqrt(499)
p.wart = pt(stat, 499)
```

lub automatycznie

```
t.test(wzrost, alternative = "less", mu=170)
```

```
##
##  One Sample t-test
##
## data:  wzrost
## t = -2.0699, df = 499, p-value = 0.01949
## alternative hypothesis: true mean is less than 170
## 95 percent confidence interval:
##     -Inf 169.629
## sample estimates:
## mean of x
##  168.1804
```

wnioskujemy, że $p-$wartość dla tego testu jest mała (wynosi 0.01948), więc mamy podstawy do odrzucenia hipotezy H_0.

Rozważmy drugą hipotezę dla zmiennej $X =$'wzrost', dla mediany X. Korzystając z założenia, że rozkład wzrostu jest rozkładem normalnym, wiemy że mediana wzrostu jest równa wartości średniej. Badamy więc hipotezę

$$H_0 : m = 165 = m_0$$

$$H_1 : m < 165 = m_0$$

tym samym testem co poprzednio:

```
t.test(wzrost, alternative = "less", mu=165)
```

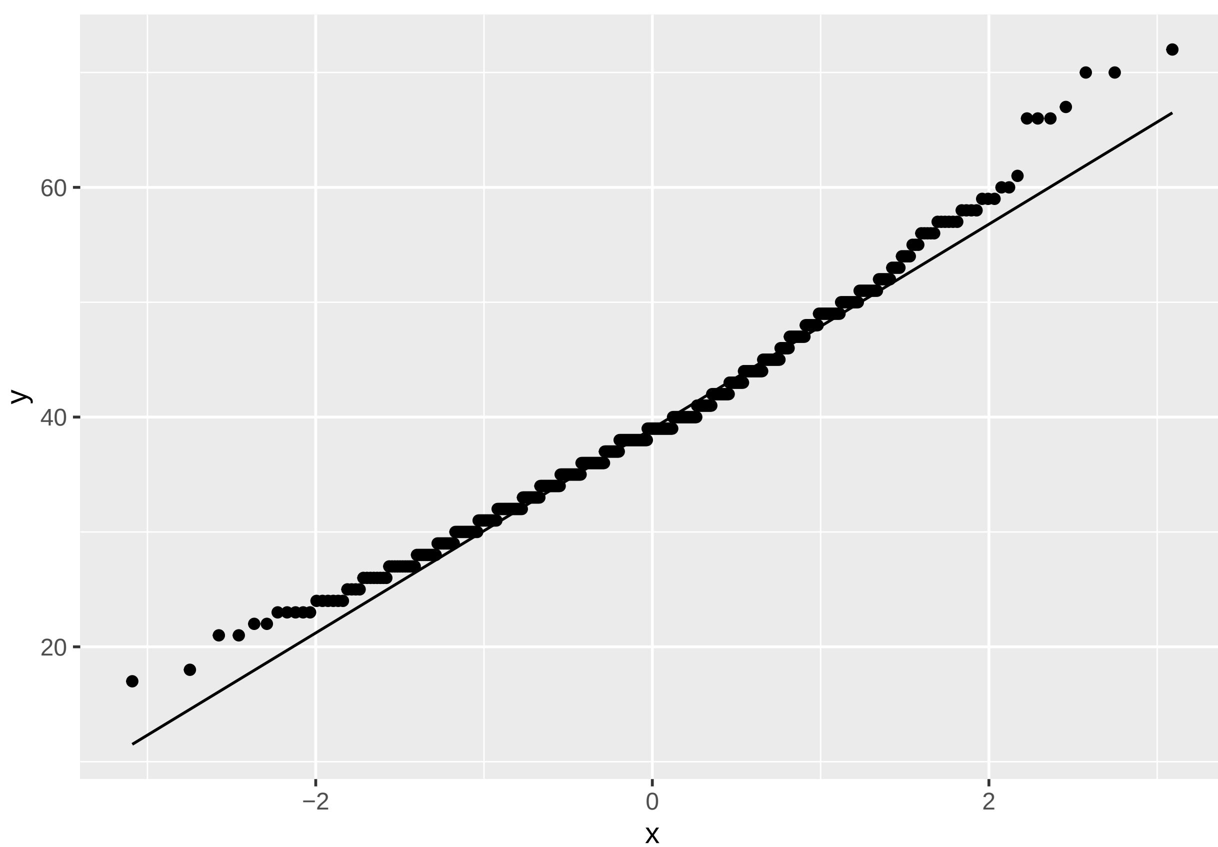
```
##
##  One Sample t-test
##
## data:  wzrost
## t = 3.6178, df = 499, p-value = 0.9998
## alternative hypothesis: true mean is less than 165
## 95 percent confidence interval:
##     -Inf 169.629
## sample estimates:
## mean of x
##  168.1804
```

otrzymyjemy bardzo dużą p-wartość 0.9998, zatem nie mamy podstaw do odrzucenia hipotezy H_0.

4. Przedziały ufności dla wieku

Rozważamy zmienną $X =$'wiek'. Z wykresu kwantylowego widzimy, że rozkład wieku jest zbliżony do rozkładu normalnego

```
ggplot() + stat_qq(aes(sample = wiek)) + stat_qq_line(aes(sample = wiek))
```



Zatem licząc przedziały ufności dla μ i σ^2 możemy skorzystać ze wzorów

$$\mu \in \left(\overline{X} - t_{0.995,499} \cdot \frac{S_n}{\sqrt{n-1}}, \overline{X} + t_{0.995,499} \frac{S_n}{\sqrt{n-1}}\right)$$

$$\sigma^2 \in \left(\frac{n \cdot S_n^2}{\chi^2_{0.995,499}}, \frac{n \cdot S_n^2}{\chi^2_{0.005,499}}\right)$$

gdzie $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})$

Z kodu

```
mean(wiek) #m
```

```
## [1] 39.484
```

```
sn2 =  sqrt(1/500*sum((wiek - mean(wiek))^2))
qt(0.995,499)*sqrt(sn2/(499)) #qt
```

```
## [1] 0.3466278
```

```
500*sn2/qchisq(0.995, 499) #sig1
```

```
## [1] 7.675899
```

```
500*sn2/qchisq(0.005, 499) #sig2
```

```
## [1] 10.64036
```

otrzymujemy, że przedział ufności na poziomie 99% dla μ to (m-qt, m+qt) $= (39.1374, 39.8306)$ a dla σ to (sig1, sig2) $= (7.6759, 10.6404)$

Znajdziemy przedział ufności dla kwantyla x_p rzędu p dla wieku. Niech $X = (x_{(1)}, x_{(2)}, ..., x_{(500)})$ oznacza posortowany niemalejąco wiek

```
x = sort(wiek)
```

znajdziemy takie $x_{(d)} \leq x_p \leq x_{(e)}$, że $P(x_{(d)} \leq x_p \leq x_{(e)}) \geq 0.99$. Skorzystamy ze wzorów z artykułu, że

$$d = \arg\max P(x_{(r)} \leq x_p) \geq 0.995$$

$$e = \arg\min P(x_p \geq x_{(e)}) \geq 0.995$$

Wykonując funkcje

```
qbinom(0.005, size = 500, prob=0.25)
```

```
## [1] 101
```

```
qbinom(0.995, size = 500, prob=0.25)+1
```

```
## [1] 151
```

```
qbinom(0.005, size = 500, prob=0.5)
```

```
## [1] 221
```

```
qbinom(0.995, size = 500, prob=0.5)+1
```

```
## [1] 280
```

```
qbinom(0.005, size = 500, prob=0.75)
```

```
## [1] 350
```

```
qbinom(0.995, size = 500, prob=0.75)+1
```

```
## [1] 400
```

otrzymujemy przedziały ufności dla kwantyli $x_{0.25}, x_{0.5}, x_{0.75}$ na poziomie 99%.

Dla $x_{0.25}$ mamy przedział $[x_{101}, x_{151}] = [32, 35]$

Dla $x_{0.5}$ mamy przedział $[x_{221}, x_{280}] = [38, 40]$

Dla $x_{0.75}$ mamy przedział $[x_{350}, x_{400}] = [43, 47]$

5. Testowanie hipotez

a) Sprawdzamy hipotezę, czy średni wiek osób zamężnych 'people.tab.zm' jest równy średniemu wiekowi osób niezamężnych 'people.tab.nzm'.

```
people.tab.zm = people.tab[people.tab$stan_cywilny == "TRUE",]
people.tab.nzm = people.tab[people.tab$stan_cywilny =="FALSE",]
```

Najpierw przetestujemy hipotezę, czy wariancja wieku osób niezamężnych jest równa wariancji wieku osób zamężnych, to znaczy

$H_0 : \sigma^2_{nzm} = \sigma^2_{zm}$ $H_1 : \sigma^2_{nzm} \neq \sigma^2_{zm}$

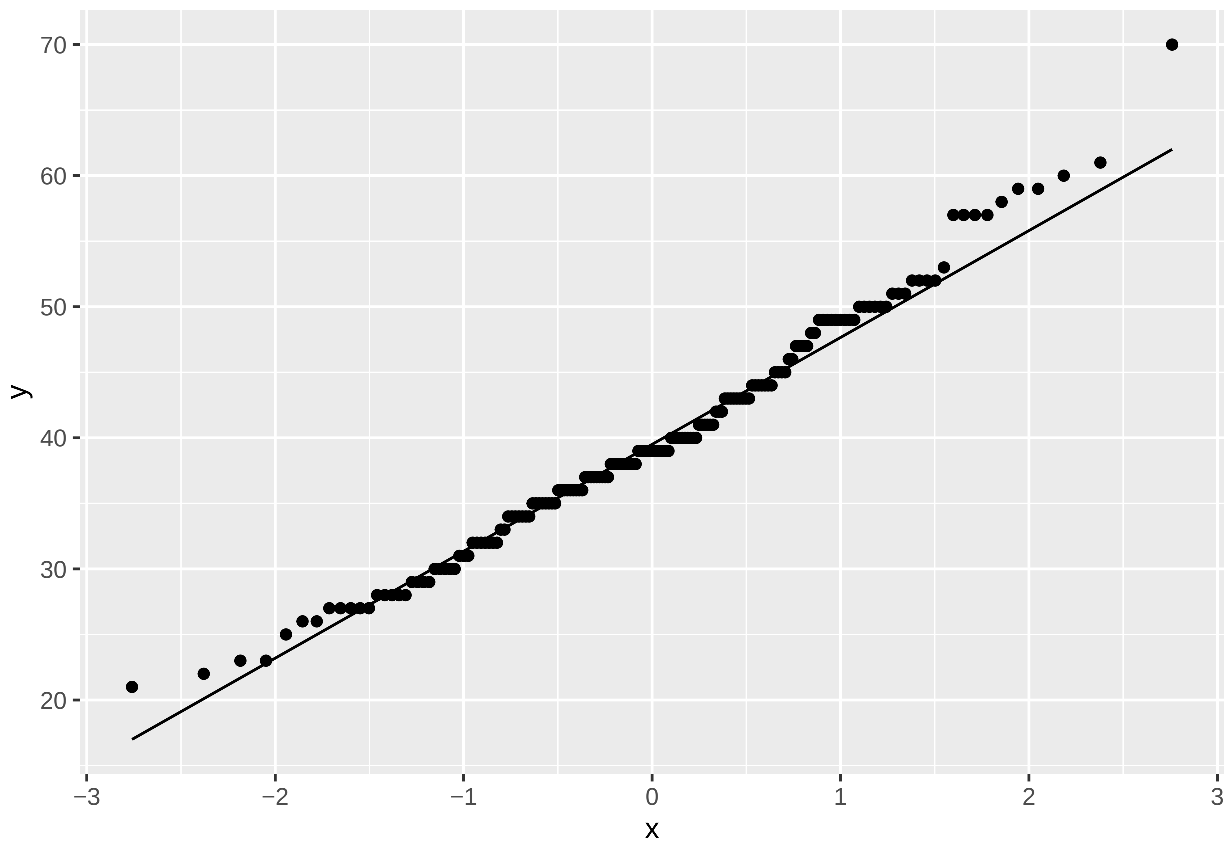
Wykorzystamy F-test dla dwóch wariancji o statystyce $t = \frac{S^2_{zm}}{S^2_{nzm}}$, gdzie S_{zm}, S_{nzm} to wariancje próbkowe. otrzymujemy

```
var.test(people.tab.zm$wiek, people.tab.nzm$wiek, alternative = "two.sided", conf.level = 0.99)
```

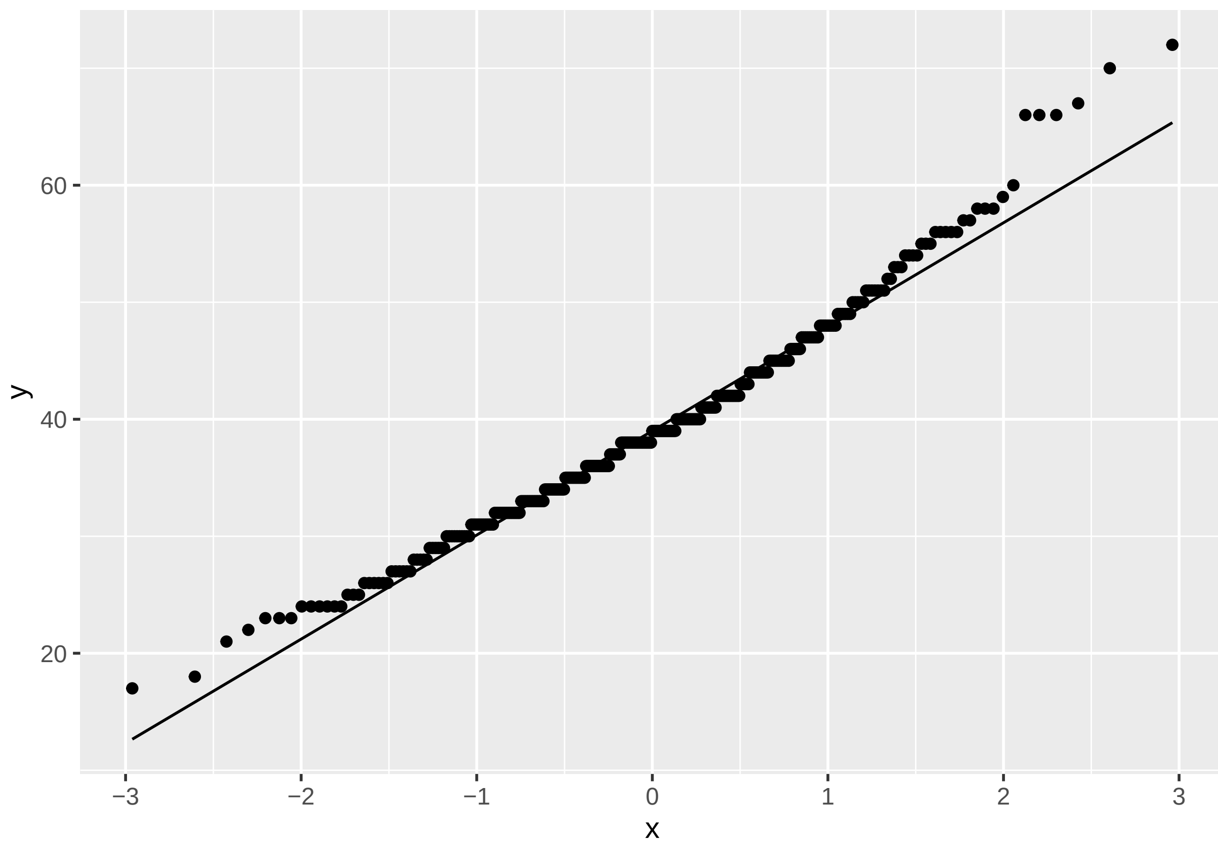
```
##
##  F test to compare two variances
##
## data:  people.tab.zm$wiek and people.tab.nzm$wiek
## F = 0.90935, num df = 172, denom df = 326, p-value = 0.4871
## alternative hypothesis: true ratio of variances is not equal to 1
## 99 percent confidence interval:
##  0.649425 1.293729
## sample estimates:
## ratio of variances
##          0.909355
```

Ponieważ wartość statystyki F nie należy do przedziału krytycznego, nie mamy podstaw, by odrzucić H_0. Załóżmy zatem, że $\sigma^2_{nzm} = \sigma^2_{zm}$. Będziemy również zakładać, że wiek zarówno w grupie zamężnych jak i niezamężnych ma rozkład normalny (na podstawie poniższych wykresów kwantylowych)

```
ggplot() + stat_qq(aes(sample = people.tab.zm$wiek)) + stat_qq_line(aes(sample = people.tab.zm$wiek))
```




```
ggplot() + stat_qq(aes(sample = people.tab.nzm$wiek)) + stat_qq_line(aes(sample = people.tab.nzm$wiek))
```



Teraz przejdźmy do głównej hipotezy

$H_0 : \mu_{zm} = \mu_{nzm}$ tzn. średni wiek w tych grupach jest równy

$H_1 : \mu_{zm} \neq \mu_{nzm}$

Zastosujemy test istotności dla dwóch średnich dla prób z rozkładu normalnego o równych wariancjach

```
#ggplot(people.tab.nzm) + geom_histogram(aes(x=wiek), bins=15) + ggtitle("histogram wieku niezamężnych")
#ggplot(people.tab.zm) + geom_histogram(aes(x=wiek), bins=15) + ggtitle("histogram wieku zamężnych")
t.test(people.tab.zm$wiek, people.tab.nzm$wiek, var.equal=TRUE, conf.level = 0.99)
```

```
##
##  Two Sample t-test
##
## data:  people.tab.zm$wiek and people.tab.nzm$wiek
## t = 0.57846, df = 498, p-value = 0.5632
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -1.695061  2.672028
## sample estimates:
## mean of x mean of y
##  39.80347  39.31498
```

Statystyka $t = 0.578$ nie należy do obszeru krytycznego $(-\infty, -1.695) \cup (2.672, \infty)$, zatem nie mamy podstaw do odrzucenia hipotezy, że $\mu_{zm} = \mu_{nzm}$

b) Zmienne jakościowe. Przetestujemy, czy płeć i budynek są niezależne.

Na początku usuniemy te obserwacje, dla których nie znamy płci i stworzymy tabelę kondygnacji.

```
new.people.tab = people.tab[is.na(people.tab$plec) == FALSE,] #obserwacje z płcią
people.k = new.people.tab[new.people.tab$plec == 'K',]
people.m = new.people.tab[new.people.tab$plec == 'M',]

budynki.k = unname(summary(people.k$budynek))
budynki.m = unname(summary(people.m$budynek))
cont.tab = matrix(c(budynki.k, budynki.m), nrow = 5)
cont.tab
```

```
##      [,1] [,2]
## [1,]   24   26
## [2,]   92   82
## [3,]   48   51
## [4,]   27   20
## [5,]   48   44
```

Liczby w tabeli kondygnacji są odpowiednio duże (większe od 20), żeby przyjąć, że statystyka χ^2 ma w przybliżeniu rozkład $\chi^2((5-1)(2-1))$.

Opis testu

H_0 : budynki i płeć są niezależne

H_1 : budynki i płeć są zależne

Robimy test niezaleznosci χ^2-Pearsona:

```
chisq.test(cont.tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont.tab
## X-squared = 1.4097, df = 4, p-value = 0.8425
```

Skąd mamy, że p-wartość statystki X-squared jest większa niż 0.01, zatem nie mamy podstaw do odrzucenia hipotezy, że budynki i płeć są niezależne.

c) Zmienne ilościowe. Czy waga i wydatki są niezależne?

H_0 : waga i wydatki są niezależne

H_1 : istnieje jakikolwiek rodzaj zależności między wagą a wydatkami

Zrobimy test p-Spearmana

```
cor.test(waga, wydatki,  method = "spearman")
```

```
## Warning in cor.test.default(waga, wydatki, method = "spearman"): Cannot compute
## exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  waga and wydatki
## S = 23232437, p-value = 0.00996
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
```

```
## -0.1151614
```

p-wartość wyszła mniejsza, niż 0.01, zatem odrzucamy H_0 i zakładamy, że waga i wydatki są zależne

d) Sprawdzimy hipotezę czy wydatki mają rozkład wykładniczy z parametrem λ, tzn. o gęstości $f(x) = \lambda e^{\lambda x}$. Na początku znajdziemy parametr λ metodą największej wiarygodności. $X = (x_1, ..., x_{500})$ oznacza obserwacje wydatków

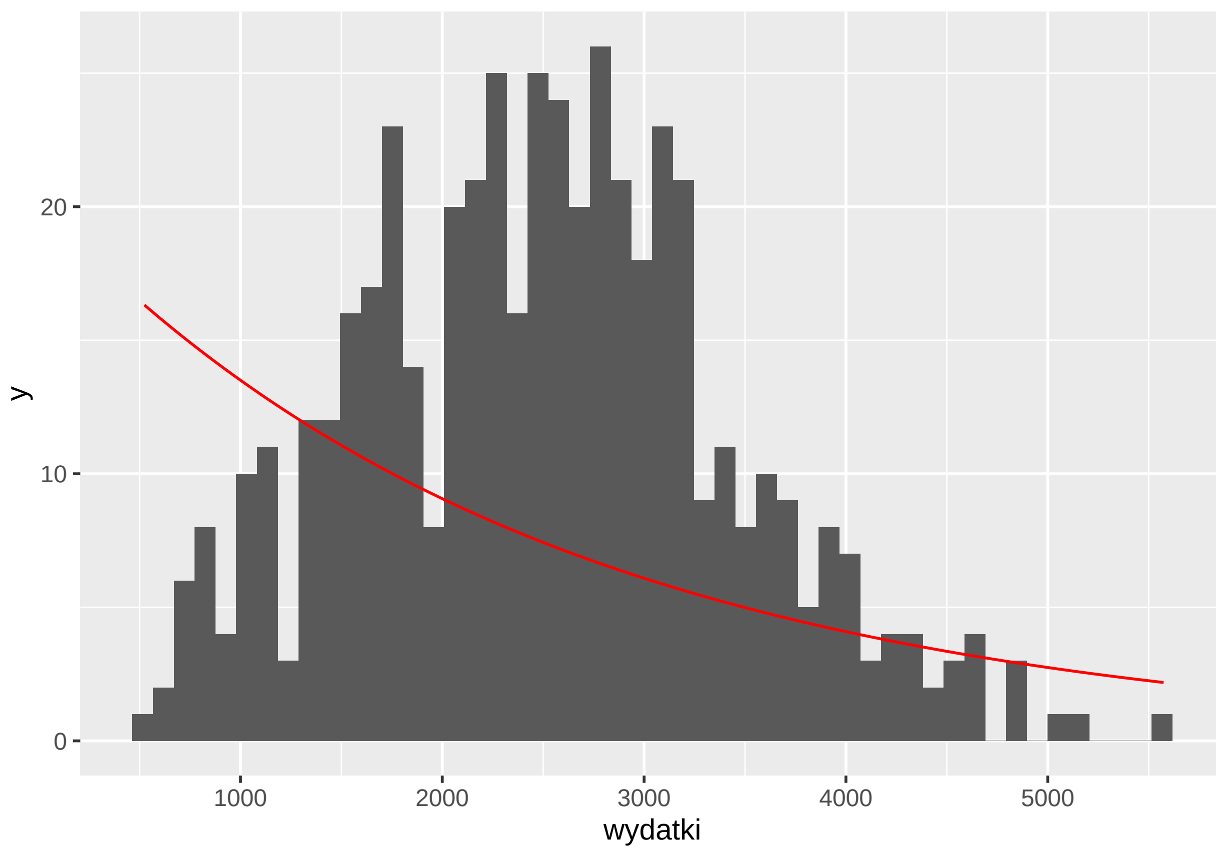
$$L_\lambda(wydatki) = \Pi_i \; \lambda e^{-\lambda x_i} = \lambda^{500} e^{-\lambda \sum x_i} = \lambda^{500} e^{-\lambda \cdot 1255836}$$

Szukamy λ, które zmaksymalizuje powyższą wartość, czyli takiego, które zmaksymalizuje

$$\ln(L_\lambda(wydatki)) = 500 \ln(\lambda) - 1255836\lambda$$

Po rozwiązaniu równania $\frac{d}{d\lambda} \ln(L_\lambda(wydatki)) = 0 \Leftrightarrow \frac{500}{\lambda} - 1255836 = 0$ otrzymujemy, że $\lambda = \frac{125}{313959}$.

```
x = 524:5574
liczba_obserwacji = 500
szerokosc_kubelka = (5574-524)/50
y = liczba_obserwacji*szerokosc_kubelka*125/313959*exp(-125/313959*x)
d = data.frame(x, y)
ggplot(people.tab) + geom_histogram(aes(x=wydatki), bins=50) + geom_line(aes(x=x, y=y), data=d, color =
```



H_0 : wydatki mają rozkład wykładniczy z parametrem $\lambda = \frac{125}{313959}$

H_1 : wydatki mają inny rozkład

Do przetestowania użyjemy testu Kołmogorowa-Smirnowa

```r
ks.test(x, "pexp", rate = 125/313959)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.25907, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Otrzymaliśmy p-wartość istotnie mniejszą niż 0.01, zatem odrzucamy hipotezą, że wydatki mają rozkład wykładniczy z parametrem $\lambda = \frac{125}{313959}$

Regresja liniowa

Oszacujemy model regresji liniowej dla people.tab z usuniętymi obserwacjami, dla których nie znamy płci. Patrząc na p-wartości widzimy dwie duże, przy płci(0.886) i stanie cywilnym(0.721). W pełnym modelu $R^2 = 0.9665$ i $RSE = \sqrt{\frac{RSS}{df}} = 0.4760952$

```r
model1 = lm(oszczednosci ~., new.people.tab)
summary(model1)
```

```
##
## Call:
## lm(formula = oszczednosci ~ ., data = new.people.tab)
##
## 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
```

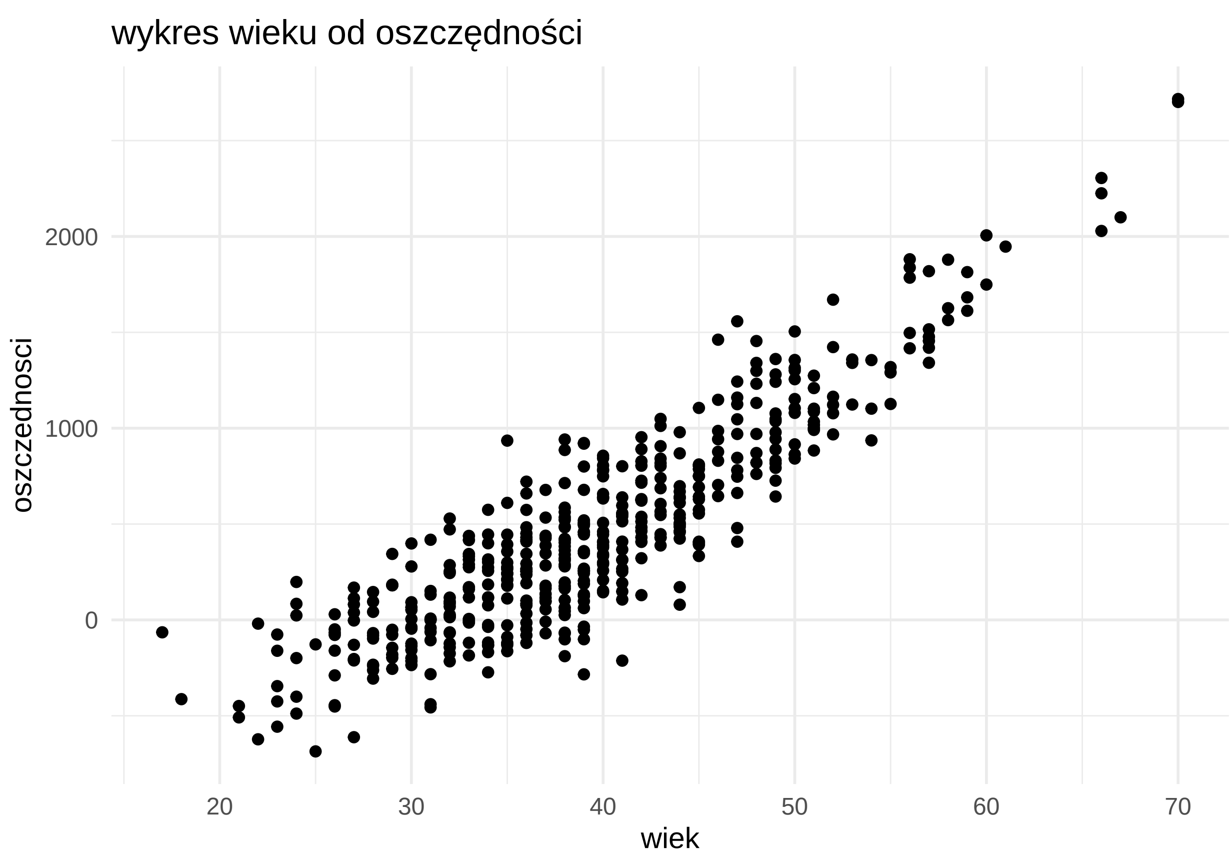
Przedziały na poszczególne współczynniki to

```r
confint(model1)
```

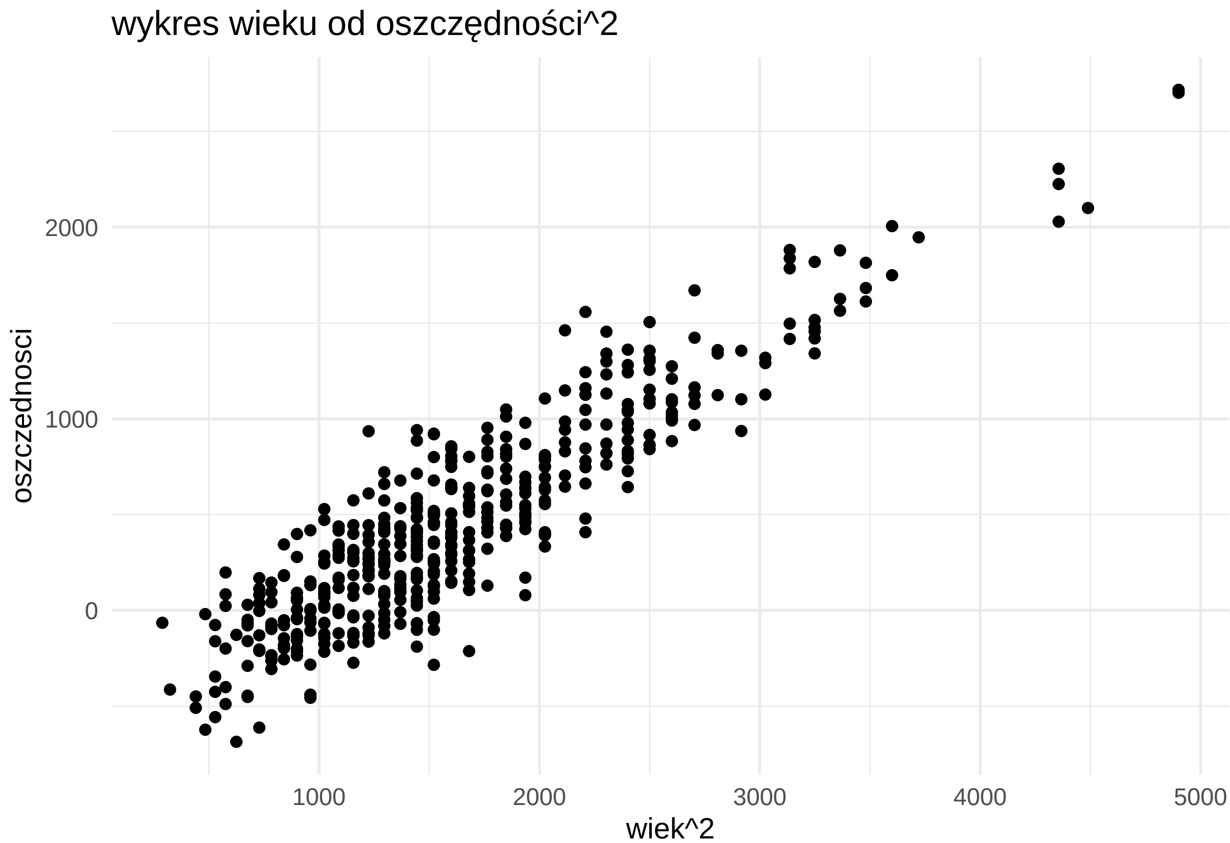
```
##                              2.5 %       97.5 %
## (Intercept)             -989.071057 -758.1110708
## wiek                      62.828047   65.0571085
## waga                       2.825166    5.0630144
## wzrost                    -3.076484   -1.6927966
## plecM                    -17.548186   20.3095560
## stan_cywilnyTRUE         -29.987562   20.7625166
## liczba_dzieci            139.503762  163.7033356
## budynekjednorodzinny    -214.378846 -149.7617725
## budynekkamienica        -340.790148 -270.4727413
## budynekloft             -387.877909 -289.0621031
## budynekwielka_plyta     -604.729067 -523.7912246
## wydatki                   -0.416700   -0.3751522
```

Patrząc na wykres ze strony 2 możemy przypuszczać, że gdy podniesiemy zmienną wiek do kwadratu, to zależność między oszczędnościami a wiekiem stanie się bardziej liniowa:

```r
ggplot(new.people.tab) + geom_point(aes(x=wiek, y=oszczednosci)) +theme_minimal() +ggtitle("wykres wieku
```

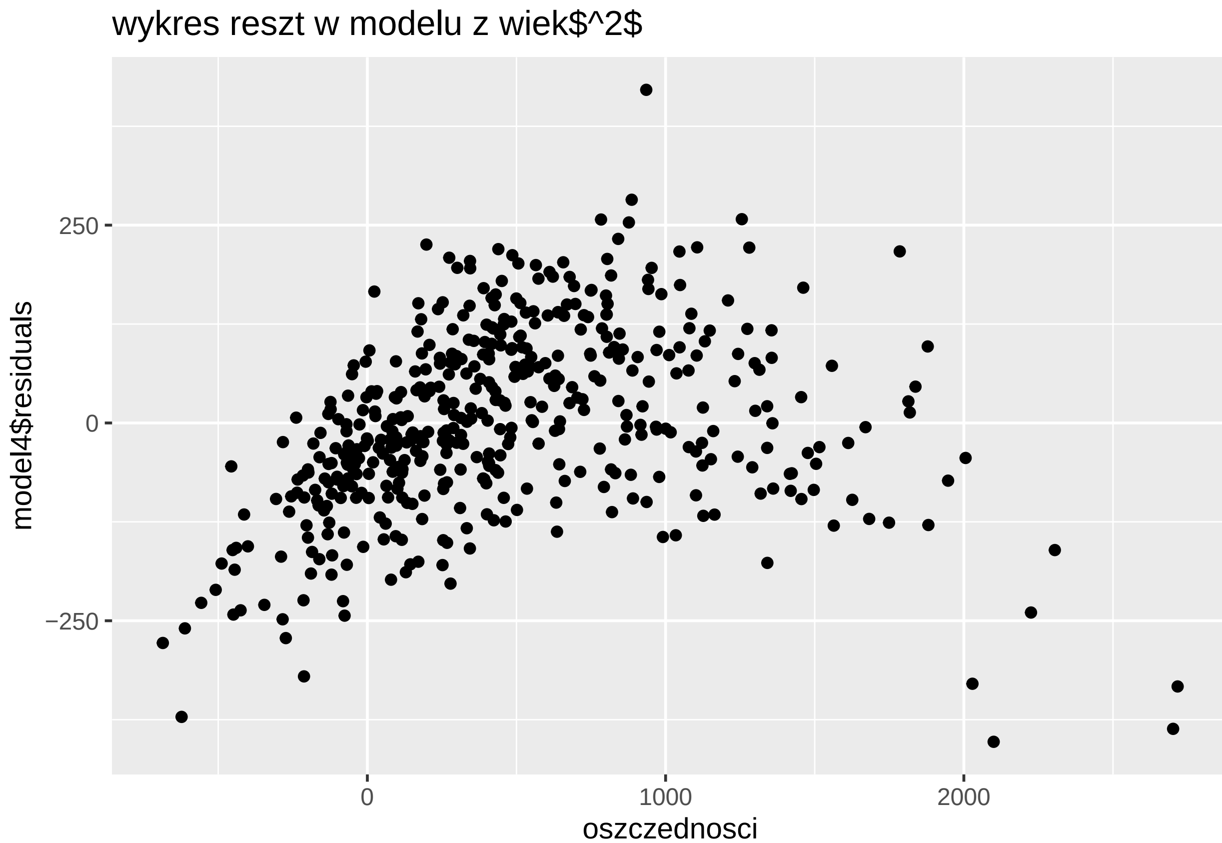



```r
ggplot(new.people.tab) + geom_point(aes(x=wiek^2, y=oszczednosci)) +theme_minimal() + ggtitle("wykres wi
```



Okazuje się jednak, że lepiej nie rozpatrzać w modelu zmiennej wiek^2, ponieważ na poniższym wykresie reszty nie układają się symetrycznie względem prostej y=0.

```
new.people.tab2 = new.people.tab
new.people.tab2$wiek = new.people.tab2$wiek^2
model4 = lm(oszczednosci ~ ., new.people.tab2)
ggplot(new.people.tab, aes(x=oszczednosci, y=model4$residuals)) + geom_point() +ggtitle("wykres reszt w
```



Spróbujemy usunąć jedną ze zmiennych płeć lub stan cywilny (ponieważ tylko przy tych zmiennych mamy duże p-wartości). Z poniższych modeli widzimy, że po usunięciu z modelu płci i stanu cywilnego R^2 prawie się nie zmienia (ciągle jest równe 0.9665), natomiast w obu modelach 'model2' i 'model3' $RSE = \sqrt{\frac{RSS}{df}} = 0.4753339$, czyli zwiększa się o 0.0007613.

```
model2 = lm(oszczednosci ~.-plec, new.people.tab)
summary(model2)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - plec, data = new.people.tab)
##
## 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 ***
```

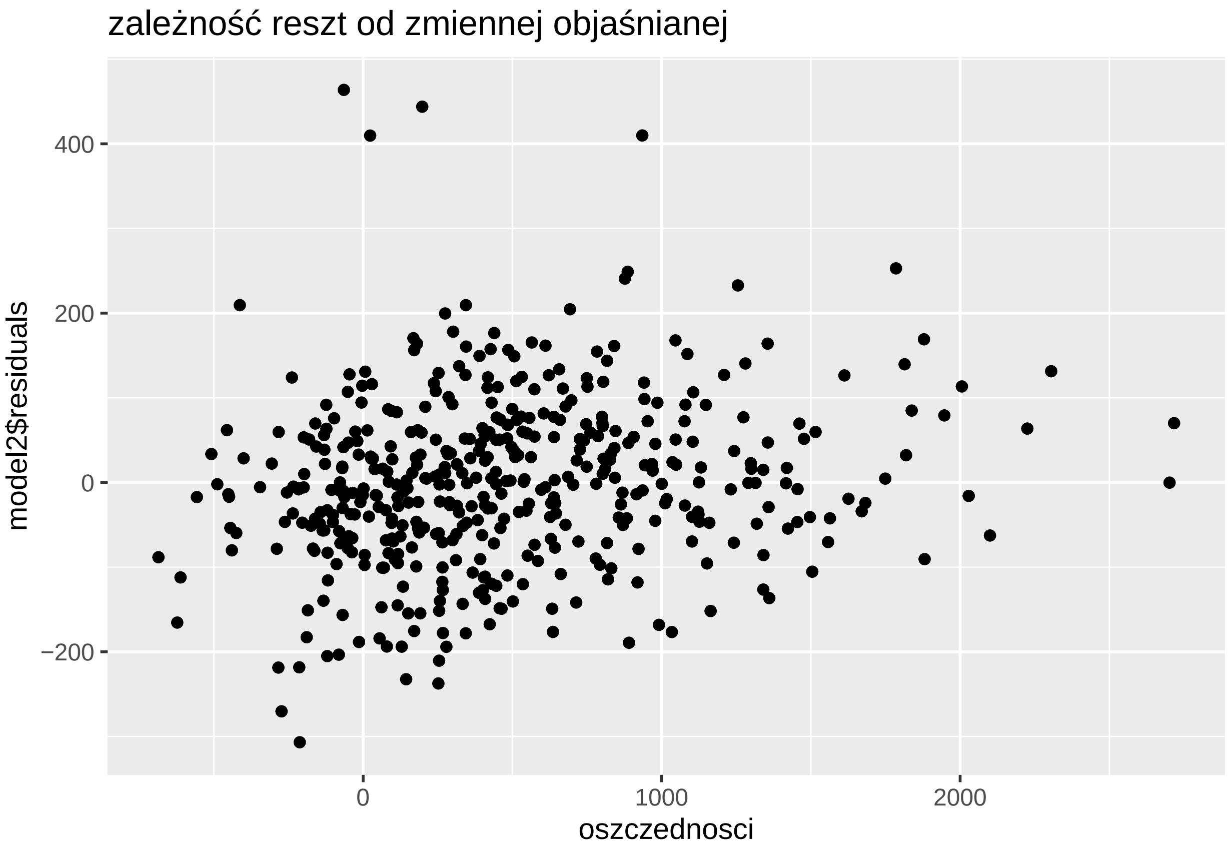
```
## budynekkamienica     -305.65338   17.87011 -17.104  < 2e-16 ***
## budynekloft          -338.69765   25.06331 -13.514  < 2e-16 ***
## budynekwielka_plyta  -564.40762   20.54419 -27.473  < 2e-16 ***
## wydatki                -0.39600    0.01055 -37.548  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 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
```

```
model3 = lm(oszczednosci ~.-stan_cywilny, new.people.tab)
summary(model3)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - stan_cywilny, data = new.people.tab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -307.76  -59.75   -1.38   57.75  462.87
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -873.53974   58.70394 -14.880  < 2e-16 ***
## wiek                   63.93900    0.56648 112.870  < 2e-16 ***
## waga                    3.93725    0.56848   6.926 1.50e-11 ***
## wzrost                 -2.38392    0.35169  -6.778 3.82e-11 ***
## plecM                   1.79131    9.55370   0.187    0.851
## liczba_dzieci         150.65024    5.54311  27.178  < 2e-16 ***
## budynekjednorodzinny -181.67008   16.38582 -11.087  < 2e-16 ***
## budynekkamienica     -305.59019   17.87251 -17.098  < 2e-16 ***
## budynekloft          -338.23676   25.10798 -13.471  < 2e-16 ***
## budynekwielka_plyta  -563.72863   20.51856 -27.474  < 2e-16 ***
## wydatki                -0.39599    0.01056 -37.504  < 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
```

Wybierzmy zatem do odrzucenia zmienną płeć, ze względu na większą p-wartość w wyjściowym modelu, czyli od tego momentu rozważamy model2.

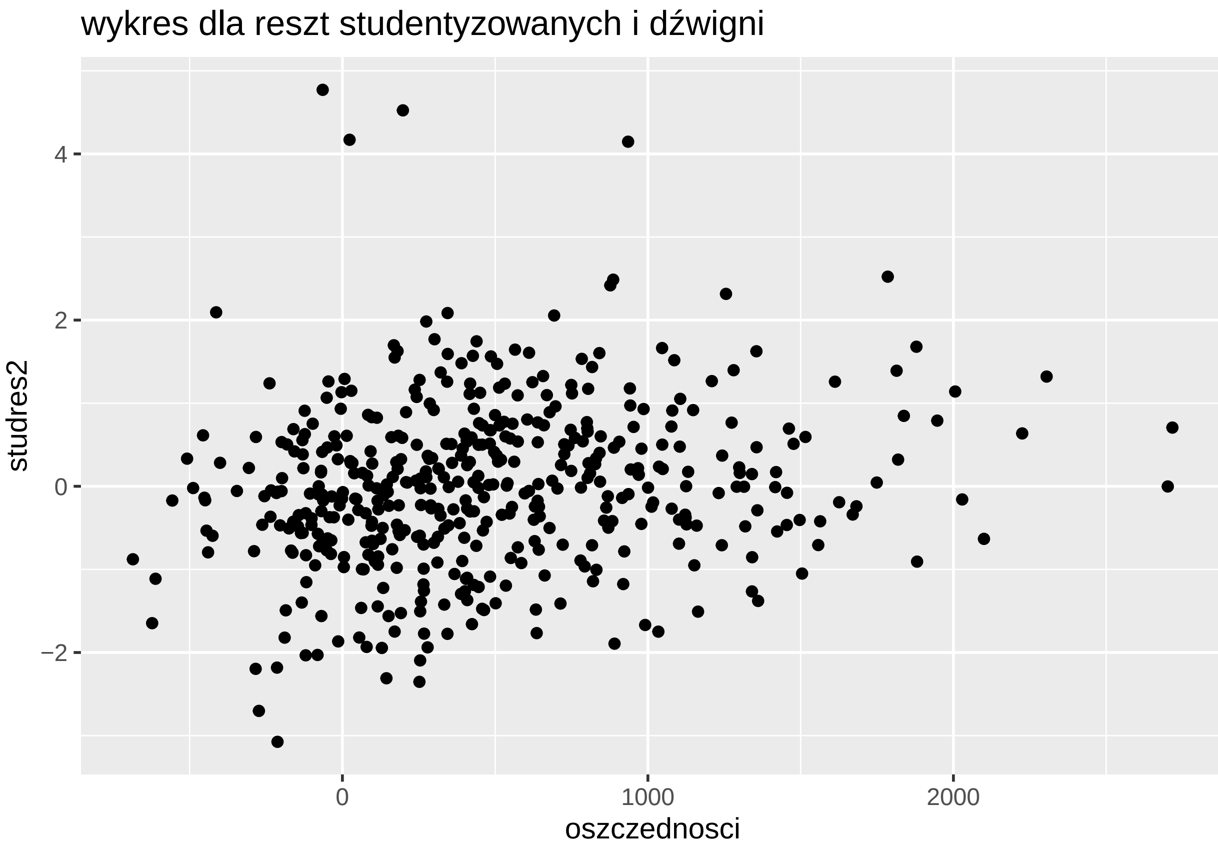
```
ggplot(new.people.tab, aes(x=oszczednosci, y=model2$residuals)) + geom_point() + ggtitle("zależność res
```



Z powyższego wykresu widać mniej więcej, że reszty są symetrycznie rozłożone względem prostej y=0 i (zatem wartość oczekiwana reszt to mniej więcej 0), więc możemy przyjąć, że reszty są niezależne od jakiejkolwiek zmiennej objaśniającej i rozkład reszt jest normalny. Ponadto wariancja błędów jest mniej więcej taka sama dla różnych poziomów 'oszczednosci', oraz z powyższej funkcji summary widać, że trend jest mniej więcej liniowy (to znaczy, nie musimy tansformować żadnej zmiennej). Można więc przyjąć, że założenia modelu liniowego są spełnione.

WYKRES RESZT STUDENTYZOWANYCH

```
library(MASS)
studres2 <- studres(model2)
ggplot(new.people.tab, aes(x=oszczednosci, y=studres2)) + geom_point() + ggtitle("wykres dla reszt stud
```



Z powyższego wykresu widzimy, że mamy 4 obserwacje wysokiej dźwigni (o resztach studentyzowanych większych niż 3):

```
head(studres2, 4)
```

```
##          2          3          4          5 
## -0.4286974 -0.6076688  0.7598235 -0.0968573
```

Ponadto mamy 2 obserwacje odstające (dla oszczędności większych od 2500)

```
sort(new.people.tab$oszczednosci, decreasing = TRUE)[1]
```

```
## [1] 2716.79
```

```
sort(new.people.tab$oszczednosci, decreasing = TRUE)[2]
```

```
## [1] 2701.63
```