a4

Carlos A. García

May 26, 2019

1. Anàlisi descriptiva i visualització

1.1 Distribució de la qualitat del son

Mostrar en un diagrama de caixa la distribució de la qualitat del son (variable psqi) de la mostra.

Llegim el fitxer proporcionat a la pràctica

```
CAD <- read.csv2("RawDataUnbalanced.csv", header = TRUE, sep = " ", dec = ".")
names(CAD)[names(CAD) == "martial_status"] <- "marital_status"
attach(CAD)

summary(CAD)
```

```
##  sex                 educ_level  monthly_income  marital_status
##  F:475   College or Higher:323   Inc1:243       Divorced:234
##  M:477   HighSchool       :318   Inc2:230       Married :242
##          Primary          :311   Inc3:240       Single  :239
##                                  Inc4:239       Widowed :237
##
##
##  residence        psqi       income_monthly_c sleep_duration_avg
##  Rural:469   Min.   : 4.00   Min.   :3724     Min.   : 3.700
##  Urban:483   1st Qu.:15.00   1st Qu.:4358     1st Qu.: 6.900
##              Median :20.00   Median :4558     Median : 8.000
##              Mean   :20.05   Mean   :4559     Mean   : 8.052
##              3rd Qu.:24.25   3rd Qu.:4772     3rd Qu.: 9.200
##              Max.   :39.00   Max.   :5499     Max.   :13.700
```

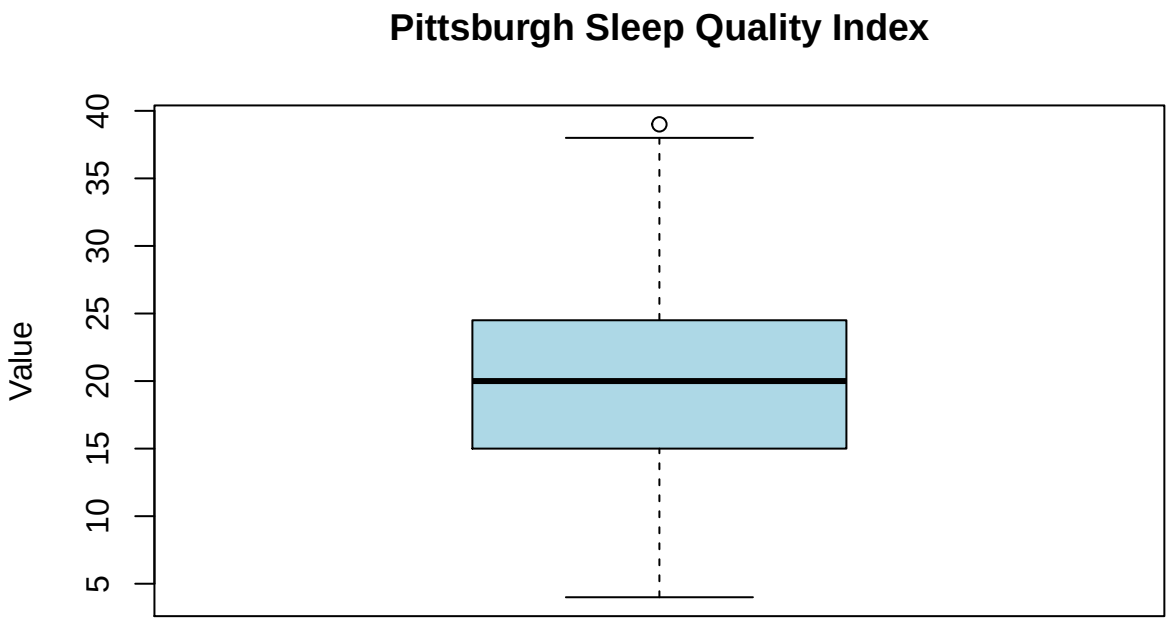
```
dim( CAD )
```

```
## [1] 952   8
```

```
str(CAD)
```

```
## 'data.frame':    952 obs. of  8 variables:
##  $ sex               : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ educ_level        : Factor w/ 3 levels "College or Higher",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ monthly_income    : Factor w/ 4 levels "Inc1","Inc2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ marital_status    : Factor w/ 4 levels "Divorced","Married",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ residence         : Factor w/ 2 levels "Rural","Urban": 2 2 2 2 1 1 1 1 1 1 ...
##  $ psqi              : int  19 18 21 19 17 18 16 15 14 14 ...
##  $ income_monthly_c  : num  4622 4765 4228 4576 4510 ...
##  $ sleep_duration_avg: num  7.1 8.1 7.2 8.1 7.6 7.9 6.9 8.4 7.5 7.1 ...
```

```
boxplot(psqi, main="Pittsburgh Sleep Quality Index",
   xlab="", ylab="Value", col="#ADD8E6")
```

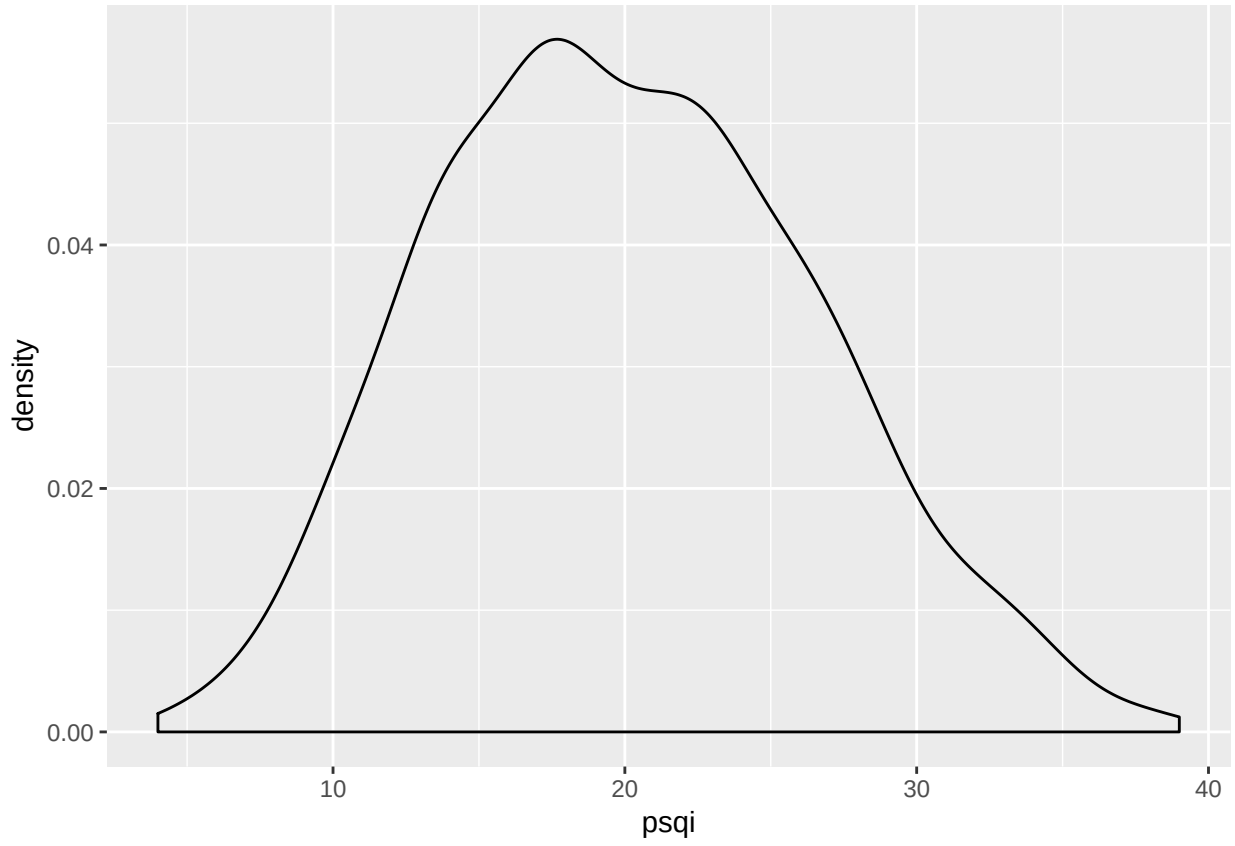



```
boxplot.stats(psqi)$out
```

```
## [1] 39
```

Podem veure que només hi ha un valor outlier. Tots els altres es mantenen dins del rang definit. Mostram la distribució de les dades:

```
ggplot(CAD, aes(psqi)) +
  geom_density(alpha=0.15)
```

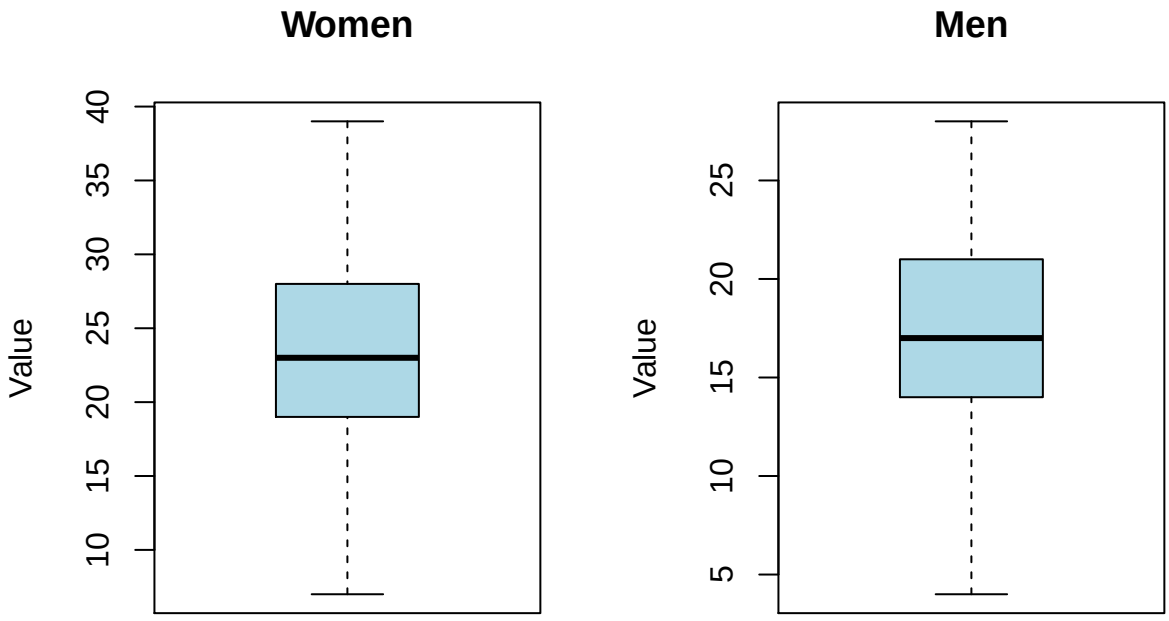


1.2 Diagrames de caixa

Mostrar en varis diagrames de caixa la distribució de la variable psqi segons el sexe, segons el nivell educatiu, segons el nivell d'ingressos, segons l'estat civil i segons la residència.

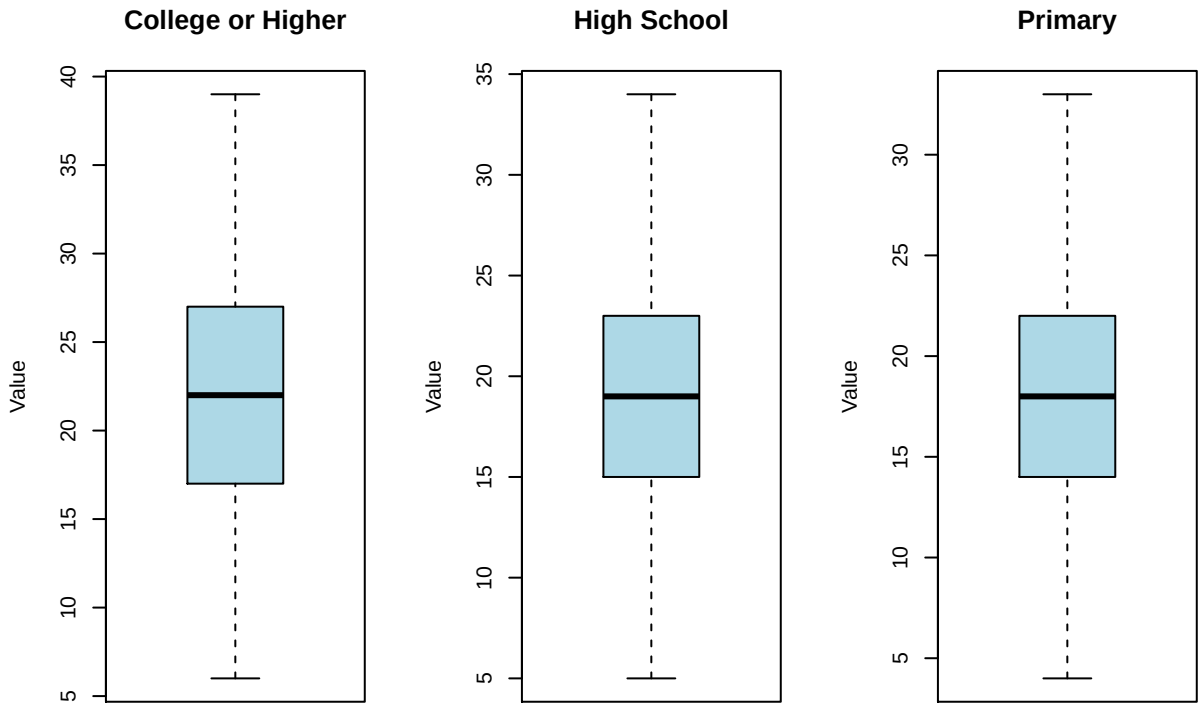
PSQI by Sex

```
par(mfrow=c(1,2))
CAD.Sex.F <- CAD[sex == 'F', ]
boxplot(CAD.Sex.F$psqi, main="Women",
   xlab="", ylab="Value", col="#ADD8E6")
CAD.Sex.M <- CAD[sex == 'M', ]
boxplot(CAD.Sex.M$psqi, main="Men",
   xlab="", ylab="Value", col="#ADD8E6")
```



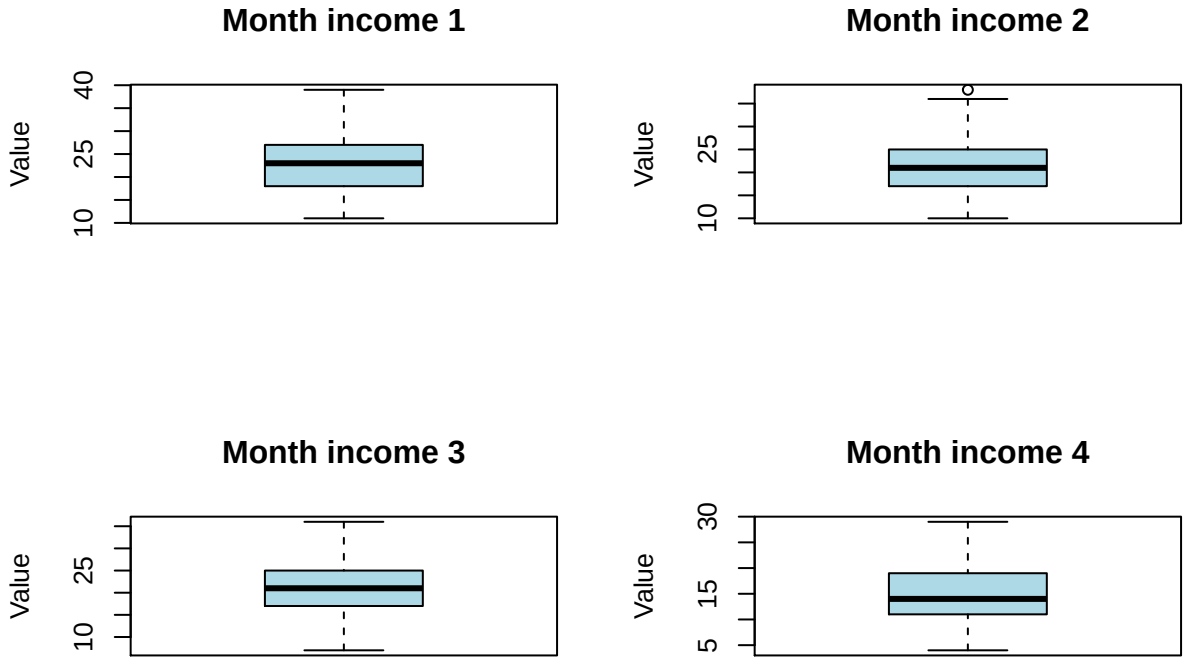
PSQI by Education Level

```
par(mfrow=c(1,3))
educ_level_fac <- factor(educ_level,
                         levels=c("College or Higher","HighSchool", "Primary"),
                         labels=c("CH", "HS", "P"))
CAD.educ_level_fac.CH <- CAD[educ_level_fac == 'CH', ]
boxplot(CAD.educ_level_fac.CH$psqi, main="College or Higher",
   xlab="", ylab="Value", col="#ADD8E6")
CAD.educ_level_fac.HS <- CAD[educ_level_fac == 'HS', ]
boxplot(CAD.educ_level_fac.HS$psqi, main="High School",
   xlab="", ylab="Value", col="#ADD8E6")
CAD.educ_level_fac.P <- CAD[educ_level_fac == 'P', ]
boxplot(CAD.educ_level_fac.P$psqi, main="Primary",
   xlab="", ylab="Value", col="#ADD8E6")
```



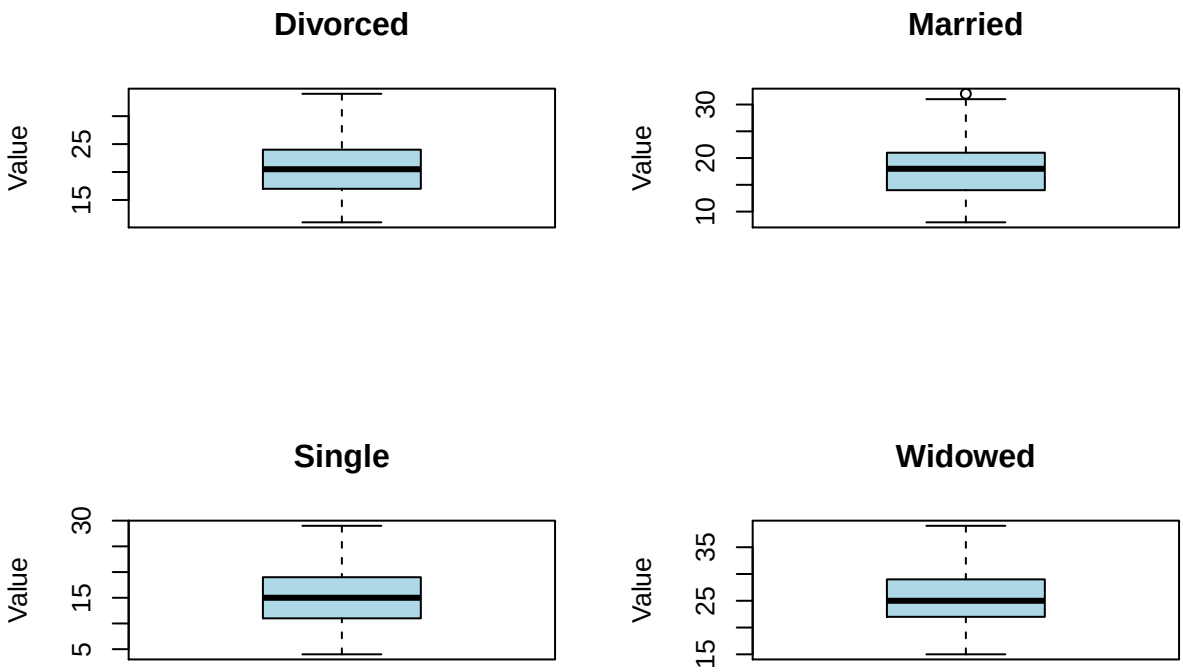
PSQI by Monthly Income

```
par(mfrow=c(2,2))
CAD.monthly_income.1 <- CAD[monthly_income == 'Inc1', ]
boxplot(CAD.monthly_income.1$psqi, main="Month income 1",
   xlab="", ylab="Value", col="#ADD8E6")
CAD.monthly_income.2 <- CAD[monthly_income == 'Inc2', ]
boxplot(CAD.monthly_income.2$psqi, main="Month income 2",
   xlab="", ylab="Value", col="#ADD8E6")
CAD.monthly_income.3 <- CAD[monthly_income == 'Inc3', ]
boxplot(CAD.monthly_income.3$psqi, main="Month income 3",
   xlab="", ylab="Value", col="#ADD8E6")
CAD.monthly_income.4 <- CAD[monthly_income == 'Inc4', ]
boxplot(CAD.monthly_income.4$psqi, main="Month income 4",
   xlab="", ylab="Value", col="#ADD8E6")
```



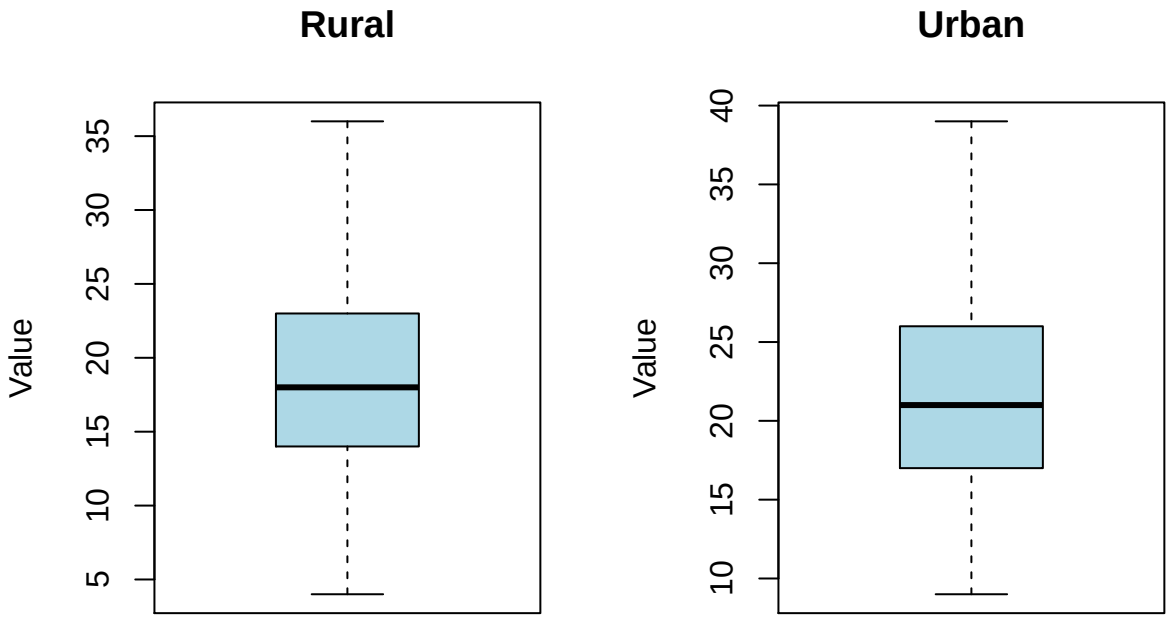
PSQI by Marital Status

```
par(mfrow=c(2,2))
CAD.marital_status.Divorced <- CAD[marital_status == 'Divorced', ]
boxplot(CAD.marital_status.Divorced$psqi, main="Divorced",
   xlab="", ylab="Value", col="#ADD8E6")
CAD.marital_status.Married <- CAD[marital_status == 'Married', ]
boxplot(CAD.marital_status.Married$psqi, main="Married",
   xlab="", ylab="Value", col="#ADD8E6")
CAD.marital_status.Single <- CAD[marital_status == 'Single', ]
boxplot(CAD.marital_status.Single$psqi, main="Single",
   xlab="", ylab="Value", col="#ADD8E6")
CAD.marital_status.Widowed <- CAD[marital_status == 'Widowed', ]
boxplot(CAD.marital_status.Widowed$psqi, main="Widowed",
   xlab="", ylab="Value", col="#ADD8E6")
```



PSQI by Residence

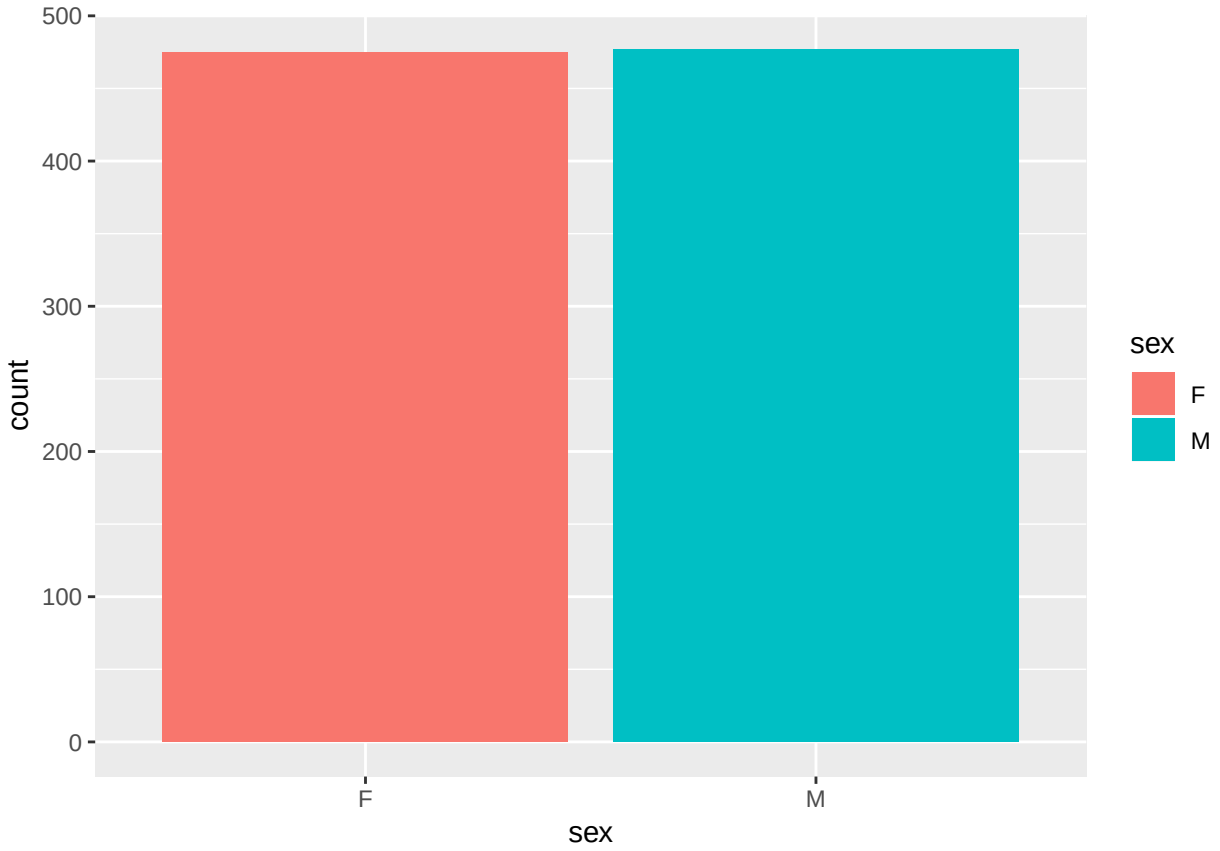
```
par(mfrow=c(1,2))
CAD.residence.Rural <- CAD[residence == 'Rural', ]
boxplot(CAD.residence.Rural$psqi, main="Rural",
   xlab="", ylab="Value", col="#ADD8E6")
CAD.residence.Urban <- CAD[residence == 'Urban', ]
boxplot(CAD.residence.Urban$psqi, main="Urban",
   xlab="", ylab="Value", col="#ADD8E6")
```



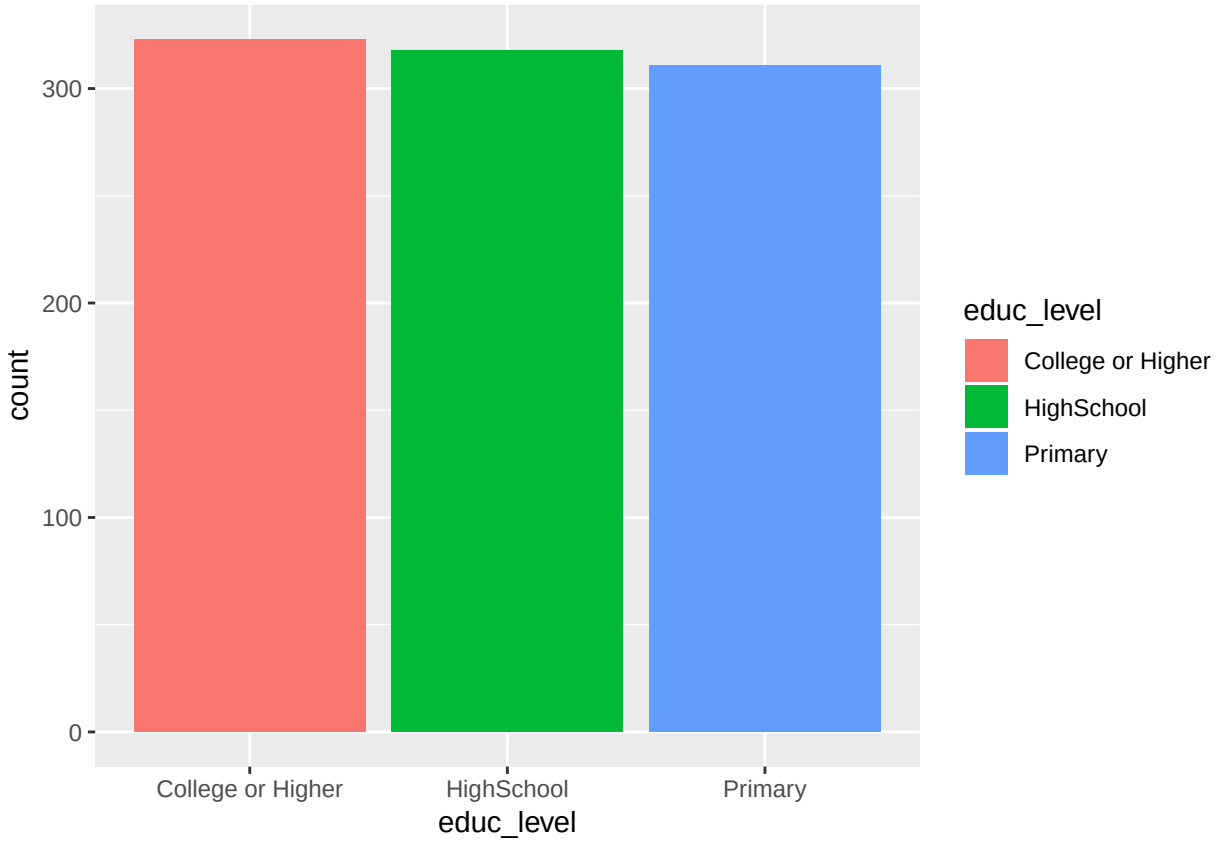
1.3 Variables categòriques

Mostreu en diagrames circulars o en diagrames de barres la distribució de valors de les variables categòriques de la mostra

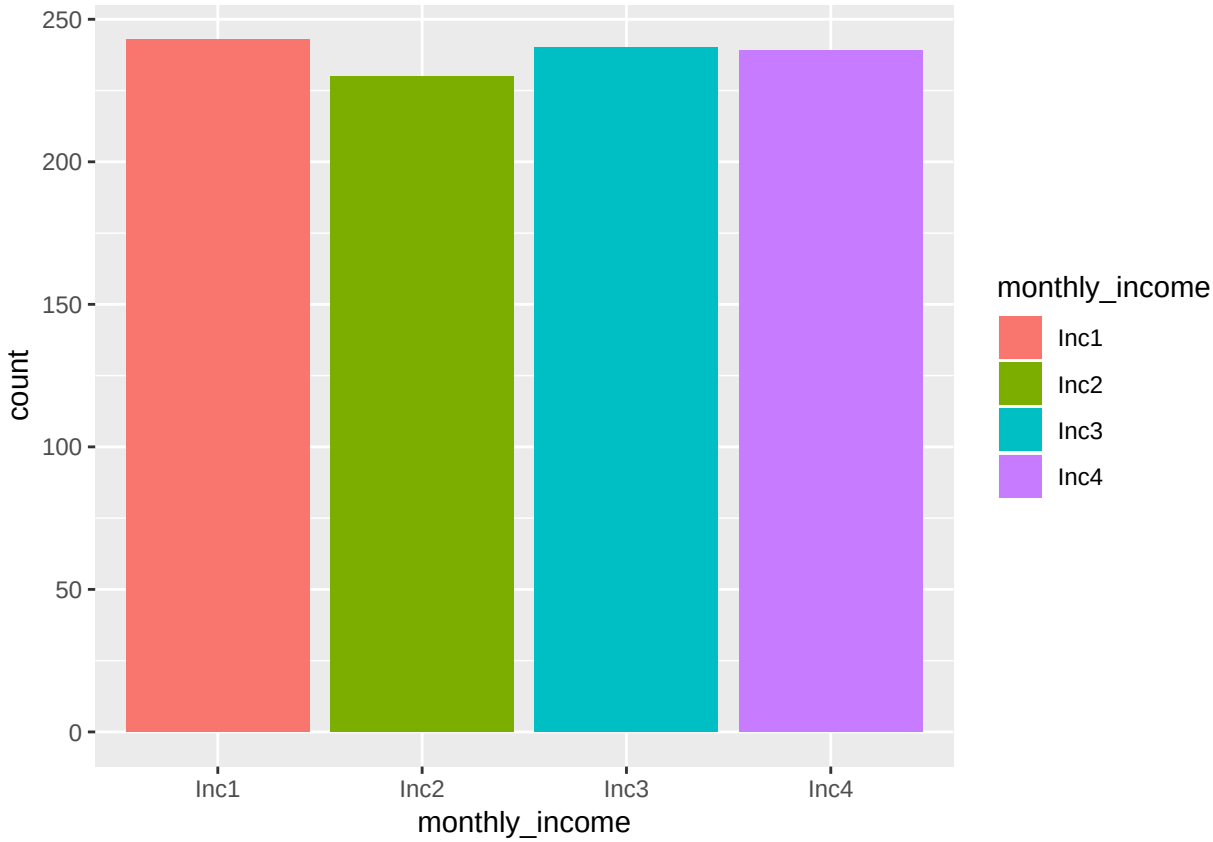
```
ggplot(CAD, aes(x=sex, fill=sex ))  + geom_bar( )
```



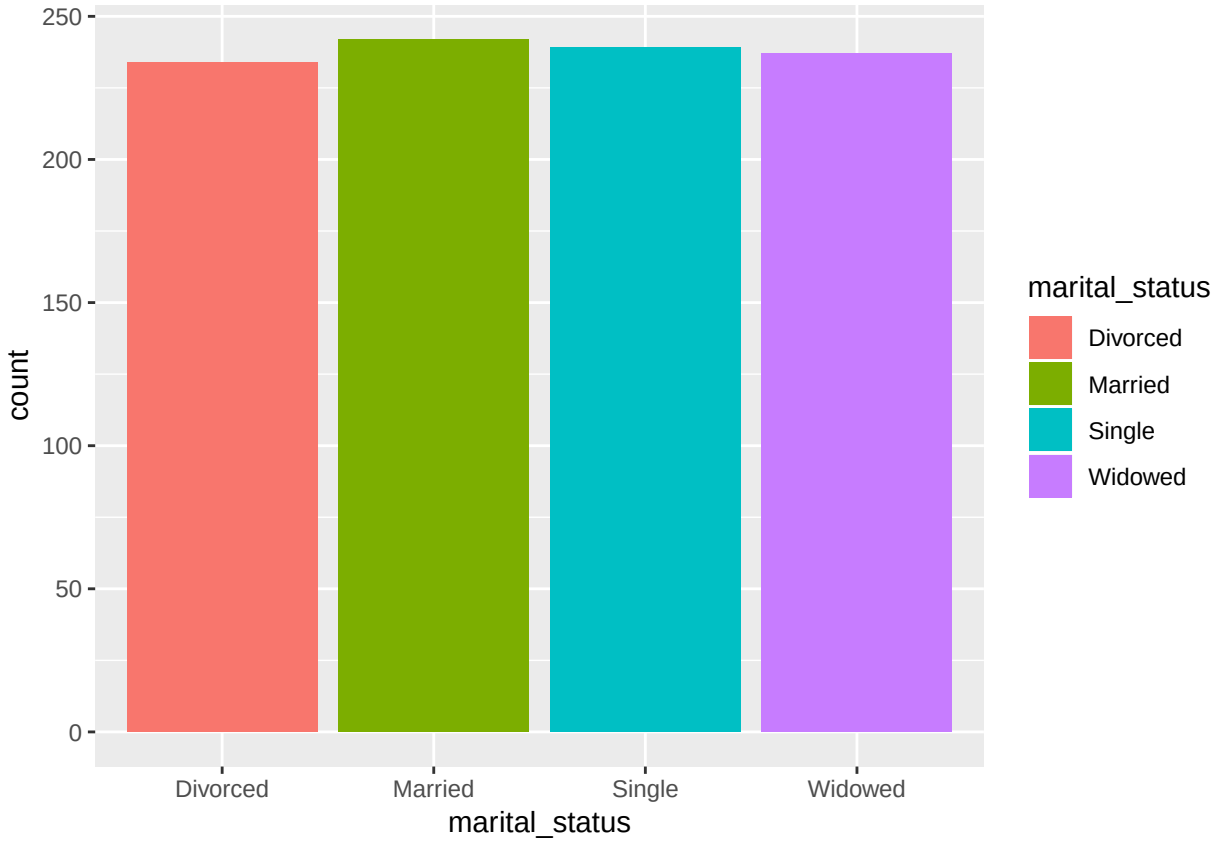

```
ggplot(CAD, aes(x=educ_level, fill=educ_level ))  + geom_bar( )
```



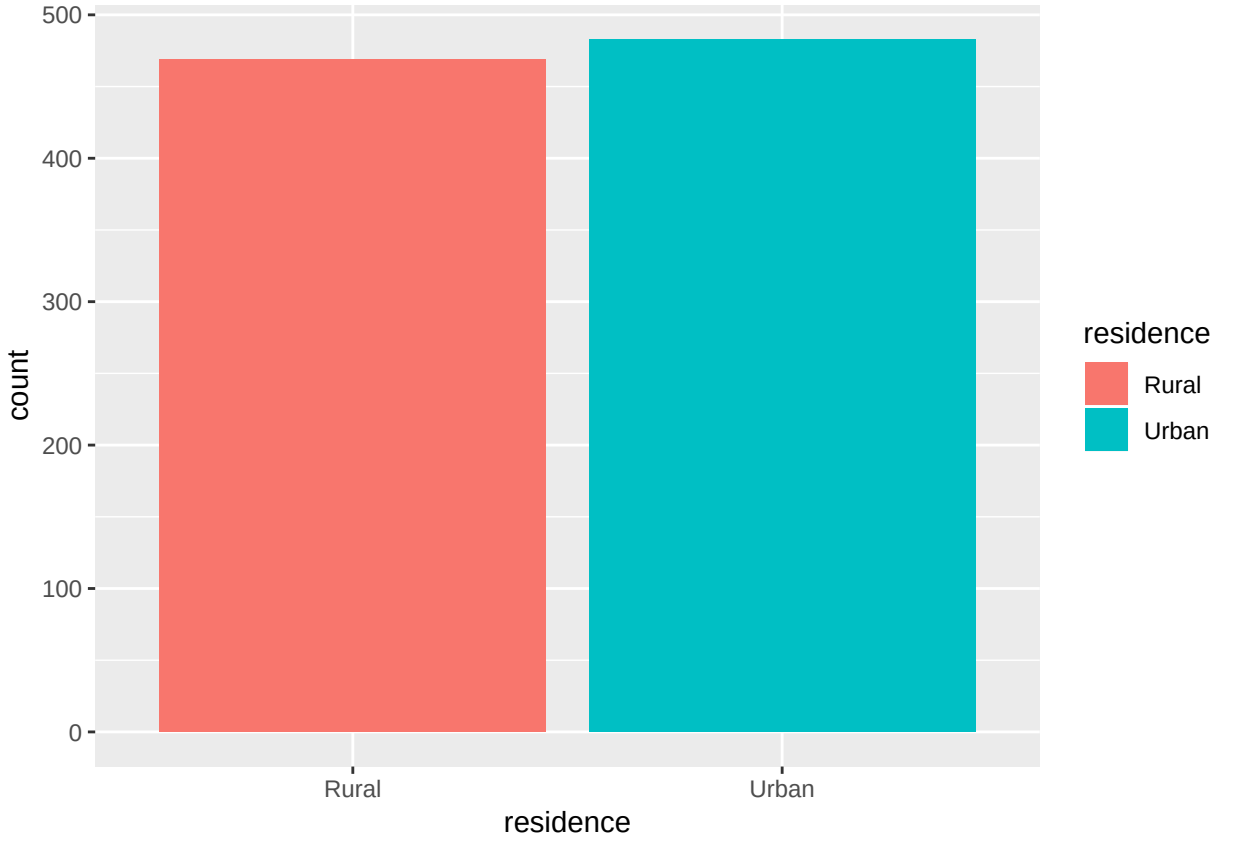

```
ggplot(CAD, aes(x=monthly_income, fill=monthly_income ))  + geom_bar( )
```




```
ggplot(CAD, aes(x=marital_status, fill=marital_status ))  + geom_bar( )
```




```
ggplot(CAD, aes(x=residence, fill=residence ))  + geom_bar( )
```



1.4 Interpretació

Interpreteu els gràfics breument.

La distribució de les variables està prou equilibrada. No existeix cap variable amb valors infra o sobre representats. Respecte a la variable psqi:

1. Sembla que depen del sexe. Els valors amb sexe F sóm més elevats que amb sexe M.
2. Sembla que depen del nivell d'educació. Les persones amb un nivell elevat d'educació tenen un valor més alt que les altres.
3. Sembla que depen del ingresos. Els valors associats als ingresos 1 i 3 semblen més elevats que el 2 i el 4.
4. Sembla que depen de l'estat marital. Els solters, per exemple, tenen valors molt més baixos que els altres estats.

2 Estadística inferencial

2.1 Interval de confiança de la qualitat del son

Calcular l'interval de confiança al 95% de la variable psqi dels pacients. A partir del valor obtingut, expliqueu com s'interpreta el resultat de l'interval de confiança.

Si assumim (com diu l'enunciat) que la variable psqi segueix una distribució normal:

```
#Mitjana, desviació típica i nombre d'elements
mean(psqi)
```

```
## [1] 20.04622
```

```
sd(psqi)
```

```
## [1] 6.457982
```

```
n <- nrow( CAD )
n
```

```
## [1] 952
```

```
#Càlcul INTERVAL DE CONFIANÇA
alpha <- 1 - 0.95
#Error típic
errorTipic <- sd(psqi) / sqrt( n )
errorTipic
```

```
## [1] 0.2093044
```

```
#Valor z
z <- qnorm( 1-alpha/2 )
z
```

```
## [1] 1.959964
```

```
error <- z * errorTipic
error
```

```
## [1] 0.4102291
```

```
#Interval
c( mean(psqi) - error, mean(psqi) + error )
```

```
## [1] 19.63599 20.45645
```

```
#Comprovació amb t Student.
#No dóna exactament igual, però el resultat és molt semblant
t.test( psqi, conf.level = 0.95 )
```

```
##
##  One Sample t-test
##
## data:  psqi
## t = 95.775, df = 951, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  19.63547 20.45697
## sample estimates:
## mean of x
##  20.04622
```

El nivell de confiança del 95% representa que el 95% d'intervals que es podrien construir a partir d'infinites mostres de la població contindrien el valor real del psqi promig.

2.2 Test de dues mostres: de la qualitat del son en funció del tipus de residència

Es pot afirmar que la qualitat del son (variable psqi) en zones rurals és millor que la qualitat del son en zones urbanes? Calcular-ho per un nivell de confiança del 95% i 97%.

2.2.1 Escriure la hipòtesi nul · la i alternativa

L'indicador PSQI valora la qualitat del son. Com més elevat és aquest índex, pitjor és la qualitat. Es vol testejar si el nivell de qualitat de són és superior en entorns rurals que en entorns urbans. Escribim:

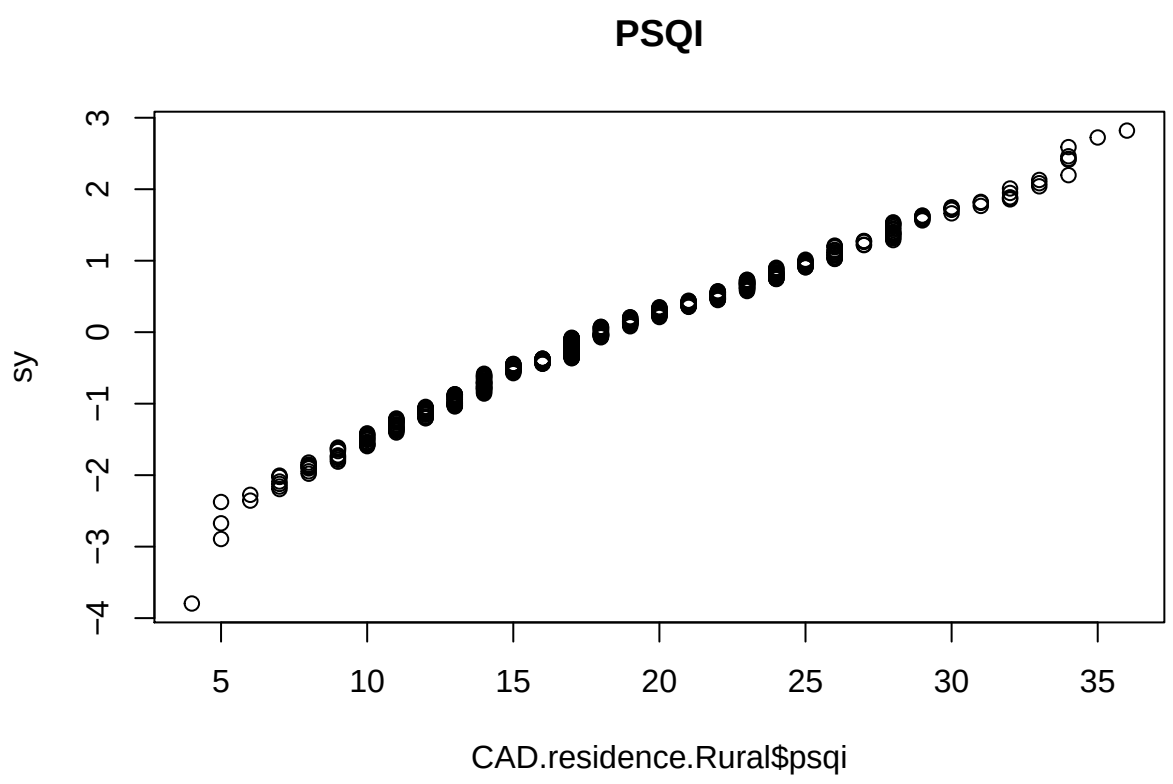
$$H_0 : \mu_R >= \mu_U$$

$$H_1 : \mu_R < \mu_U$$

2.2.2 Justifiqueu quin mètode aplicareu

Primer validem si les variables segueixen una distribució normal

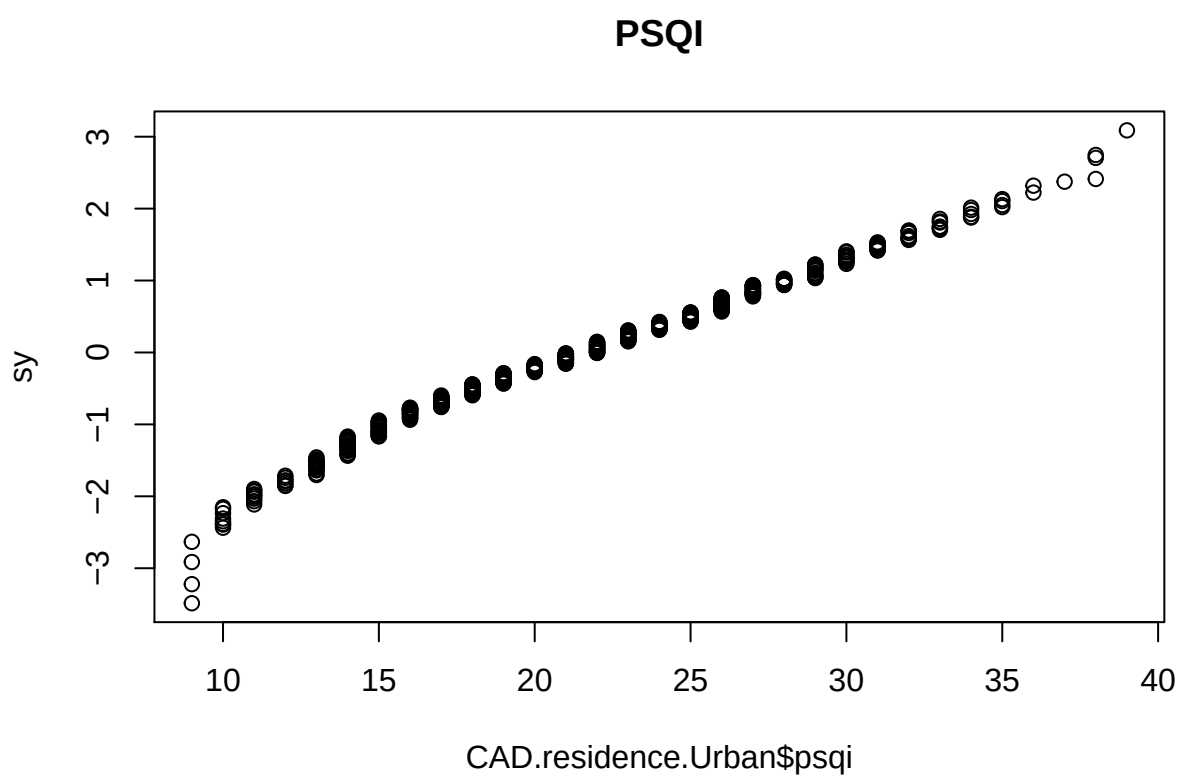
```
qqplot(CAD.residence.Rural$psqi, rnorm(nrow(CAD)), main="PSQI", ylab = NULL)
```




```
shapiro.test(CAD.residence.Rural$psqi)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CAD.residence.Rural$psqi
## W = 0.98738, p-value = 0.0004385
```

```
qqplot(CAD.residence.Urban$psqi, rnorm(nrow(CAD)), main="PSQI", ylab = NULL)
```




```
shapiro.test(CAD.residence.Urban$psqi)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CAD.residence.Urban$psqi
## W = 0.98689, p-value = 0.000245
```

Com es pot veure, que els valors de p obtinguts pel test Shapiro siguin tan baixos (inferiors a 0.05) significa que el valor NO segueix una distribució normal.

En aquest cas, podem fer servir el test de Mann-Whitney-Wilcoxon. Ens permet comparar les dues mitjanes sense asumir que les distribucions són normals.

En qualsevol cas, tal i com diu l'enunciat, assumirem que segueixen una distribució normal.

2.2.3 Realitzeu els càlculs de l'estadístic de contrast, valor crític i valor p, al 95% i 97% de nivell de confiança

Assumim que les distribucions són normals, utilitzarem un test paramètric per tal d'obtenir els valors crítics

```
# Rural stats
nRural <- nrow(CAD.residence.Rural)
meanRural <- mean(CAD.residence.Rural$psqi)
sdRural <- sd(CAD.residence.Rural$psqi)
# Urban stats
nUrban <- nrow(CAD.residence.Urban)
meanUrban <- mean(CAD.residence.Urban$psqi)
sdUrban <- sd(CAD.residence.Urban$psqi)
```

```
# Calculate value t
s <-sqrt( ((nRural-1)*sdRural^2 + (nUrban-1)*sdUrban^2 )/(nRural+nUrban-2) )
Sb <- s*sqrt(1/nRural + 1/nUrban)
t <- (meanRural-meanUrban)/ Sb
t
```

```
## [1] -7.712048
```

```
alfa <- (1-0.95)
tcritical <- qt( alfa/2, df=nRural+nUrban-2, lower.tail=FALSE )

pvalue<-pt( abs(t), df=nRural+nUrban-2, lower.tail=FALSE )*2

info<-data.frame(nRural,meanRural,sdRural,nUrban,meanUrban,sdUrban,t,tcritical,pvalue)
info %>% kable() %>% kable_styling()
```

nRural	meanRural	sdRural	nUrban	meanUrban	sdUrban	t	tcritical	pvalue
469	18.45629	6.322588	483	21.59006	6.214832	-7.712048	1.962464	0

```
#Comprovació t-test:
t.test( CAD.residence.Rural$psqi, CAD.residence.Urban$psqi, conf.level=0.95, paired=FALSE,
var.equal=TRUE, alternative="two.sided" )
```

```
##
##  Two Sample t-test
##
## data:  CAD.residence.Rural$psqi and CAD.residence.Urban$psqi
## t = -7.712, df = 950, p-value = 3.119e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.931215 -2.336330
## sample estimates:
## mean of x mean of y
##  18.45629  21.59006
```

Com podem veure, el valor de pvalue és molt propera a 0. És inferior al 5%, així que acceptem l'hipòtesi alternativa; hi ha diferències significatives entre les dues poblacions (rural i urbana)

```
alfa <- (1-0.97)
tcritical <- qt( alfa/2, df=nRural+nUrban-2, lower.tail=FALSE )

pvalue<-pt( abs(t), df=nRural+nUrban-2, lower.tail=FALSE )*2

info<-data.frame(nRural,meanRural,sdRural,nUrban,meanUrban,sdUrban,t,tcritical,pvalue)
info %>% kable() %>% kable_styling()
```

nRural	meanRural	sdRural	nUrban	meanUrban	sdUrban	t	tcritical	pvalue
469	18.45629	6.322588	483	21.59006	6.214832	-7.712048	2.173356	0

```
#Comprovació t-test:
t.test( CAD.residence.Rural$psqi, CAD.residence.Urban$psqi, conf.level=0.97, paired=FALSE,
var.equal=TRUE, alternative="two.sided" )
```

```
##
```

```
##  Two Sample t-test
##
## data:  CAD.residence.Rural$psqi and CAD.residence.Urban$psqi
## t = -7.712, df = 950, p-value = 3.119e-14
## alternative hypothesis: true difference in means is not equal to 0
## 97 percent confidence interval:
##  -4.016910 -2.250634
## sample estimates:
## mean of x mean of y
##  18.45629  21.59006
```

Al 97% els resultats són pràcticament els mateixos. Podem afirmar que les dues poblacions són diferents.

3. Regressió

3.1 Model de regressió

Apliqueu un model de regressió lineal múltiple que usi com a variables explicatives: les hores promig de son, els ingressos mensuals, el sexe, el nivell d'educació i el tipus de residència, i com a variable dependent la qualitat del son. Especifiqueu en el nivell base (en el relevel): per a la variable sexe, la categoria "M", per al nivell d'educació, la categoria "Primary", i per la variable tipus de residència, la categoria "Rural".

```r
#Creació de les variables
CAD$sexR <- relevel(CAD$sex, ref = "M")
CAD$educ_levelR <- relevel(educ_level, ref = "Primary")
CAD$residenceR <- relevel(residence, ref = "Rural")
#Regressió lineal
lmFrame <- lm(formula = psqi ~ sexR + educ_levelR + residenceR
              +  sleep_duration_avg + income_monthly_c, data = CAD)
summary(lmFrame)
```

```
##
## Call:
## lm(formula = psqi ~ sexR + educ_levelR + residenceR + sleep_duration_avg +
##     income_monthly_c, data = CAD)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.982 -1.946  0.071  1.941  8.566
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  47.2035793  2.4617690  19.175  < 2e-16 ***
## sexRF                         1.7216810  0.2108866   8.164 1.03e-15 ***
## educ_levelRCollege or Higher  1.3218808  0.2413838   5.476 5.57e-08 ***
## educ_levelRHighSchool         0.2777128  0.2368860   1.172    0.241
## residenceRUrban               1.0098237  0.1978550   5.104 4.02e-07 ***
## sleep_duration_avg            1.7639080  0.0770745  22.886  < 2e-16 ***
## income_monthly_c             -0.0094924  0.0004473 -21.220  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.959 on 945 degrees of freedom
```

```
## Multiple R-squared:  0.7914, Adjusted R-squared:  0.7901
## F-statistic: 597.6 on 6 and 945 DF,  p-value: < 2.2e-16
```

Com es pot veure, el valor de R^2 és molt més proper a 1 (adjust perfecte) que a 0 (cap adjust). Totes les variables són significatives (inferiors a 5%) excepte per al valor educ_levelRHighSchool.

3.2 Interpretació

Interpreteu el model de regressió resultant, indicant quins regressors són significatius i expliqueu si existeixen diferències degut al nivell educatiu i si hi són, en quin sentit. De manera anàloga, repetiu la mateixa discussió per a la resta de variables regressores en el model.

El valor de R^2 és correcte. No explica del tot bé el resultat, però s'acosta més a 1 que a 0. Totes les variables són significatives (valor de p inferior al 5%) excepte per al valor educ_levelRHighSchool. Respecte a les variables (tenint en compte que, com major és el valor de psqi, pitjor és la qualitat de la són):

1. Sexe. Els homes (valor M) dormen mijor que les dones (valor F).
2. Nivell educatiu. Un nivell superior d'estudis (educ_levelRCollege or Higher) implica una pitjor qualitat de són respecte a les persones amb nivell d'estudis primaris. Passa el mateix amb el nivell High School, però en aquest cas els valors no són significatius.
3. Residència. Les persones que viuen a un entorn urbà tenen pitjor qualitat de són que les persones que viuen a un entorn rural.
4. Temps mitjà de són. Dorm pitjor qui més hores dorm.
5. Ingressos mensuals. Dorm millor qui més doblers guanya.He agafat la variable income_monthly_c perque és numèrica i permet l'operació que se'ns demana després. Entenc que seria més correcte agafar el valor monthly_income.

3.3 Predicció

Apliqueu el model de regressió per a predir psqi d'una dona, amb ingressos mensuals de 3600 dòlars, que dorm en promig 6 hores, de nivell d'estudis primaris i residència de tipus rural

```
# dona 3600 6 primari rural
salariVar <- 3600
horesVar <- 6
educ_leveVar <- "Primary"
sexVar <- "F"
qdpr <- lmFrame$coefficients[1] + salariVar * lmFrame$coefficients[7] +
  horesVar * lmFrame$coefficients[6] +
  lmFrame$coefficients[2]
qdpr
```

```
## (Intercept)
##    25.33604
```

```
# dona 3600 6 primari urbà
qdpu <- lmFrame$coefficients[1] + salariVar * lmFrame$coefficients[7] +
  horesVar * lmFrame$coefficients[6] +
  lmFrame$coefficients[2] + lmFrame$coefficients[5]
qdpu
```

```
## (Intercept)
##    26.34586
```

```r
# home 3600 6 primari rural
qhpr <- lmFrame$coefficients[1] + salariVar * lmFrame$coefficients[7] +
  horesVar * lmFrame$coefficients[6]
qhpr
```

```
## (Intercept)
##    23.61436
```

```r
# Com es pot veure, qui millor dorm és l'home que viu a un ambient rural.
# Després la dona a ambient rural. En darrer lloc, la dona en ambient urbà
# Això és degut que el coeficient "dona" es:
lmFrame$coefficients[2]
```

```
##    sexRF
## 1.721681
```

```r
# I el coeficient urbà és:
lmFrame$coefficients[6]
```

```
## sleep_duration_avg
##           1.763908
```

És a dir, la "penalització" per dormir a un entorn urbà és lleugerament superior a la de ser dona.

3.4 Intervals de predicció

Calculeu els intervals de predicció dels tres pacients al 98%

```r
# dona, amb ingressos mensuals de 3600 dòlars, que dorm en promig 6 hores, de nivell
# d'estudis primaris i residència de tipus rural
newdata = data.frame(sexR = sexVar, educ_levelR = educ_leveVar,
                     income_monthly_c = salariVar,
                     sleep_duration_avg = horesVar, residenceR = "Rural")
predict.lm(lmFrame, newdata, interval = "prediction", level = 0.98)
```

```
##        fit      lwr      upr
## 1 25.33604 18.31521 32.35688
```

```r
# dona, amb ingressos mensuals de 3600 dòlars, que dorm en promig 6 hores, de nivell
# d'estudis primaris i residència de tipus urbà
newdata = data.frame(sexR = sexVar, educ_levelR = educ_leveVar,
                     income_monthly_c = salariVar,
                     sleep_duration_avg = horesVar, residenceR = "Urban")
predict.lm(lmFrame, newdata, interval = "prediction", level = 0.98)
```

```
##        fit      lwr      upr
## 1 26.34586 19.32811 33.36362
```

```r
# home, amb ingressos mensuals de 3600 dòlars, que dorm en promig 6 hores, de nivell
# d'estudis primaris i residència de tipus rural
newdata = data.frame(sexR = "M", educ_levelR = educ_leveVar,
                     income_monthly_c = salariVar,
                     sleep_duration_avg = horesVar, residenceR = "Rural")
predict.lm(lmFrame, newdata, interval = "prediction", level = 0.98)
```

```
##        fit      lwr      upr
## 1 23.61436 16.58133 30.64739
```

Als intervals es pot veure que es manté la tendència que ja hem vist: el són de les dones és pitjor que el dels homes i el són a un entorn rural és millor que un entorn urbà. Els rangs són amplis: està clar que les variables afecten, però també és veritat que es pot trobar qualsevol valor independentment de les variables

També obtingueu l'interval de confiança al 98% per la resposta mitjana

```
t.test( CAD$psqi, conf.level = 0.98 )
```

```
##
##  One Sample t-test
##
## data:  CAD$psqi
## t = 95.775, df = 951, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 98 percent confidence interval:
##  19.55848 20.53396
## sample estimates:
## mean of x
##  20.04622
```

Personalment, me crida l'atenció que, amb tanta variabilitat en les dades, la mitjana estigui acotada a un rang tan reduït.

3.5 Interpretació de la predicció

Interpreteu els resultats obtinguts a l'apartat anterior. Concretament, consideraríeu que el model és precís, tenint en compte els valors d'R2 i p-value obtinguts?

Sí, el model és precís. El p-value és pràcticament 0, el que vol dir que les dades són significatives. El valor de R^2 és proper al 80%, el que vol dir que s'ajusta bé a les dades reals.

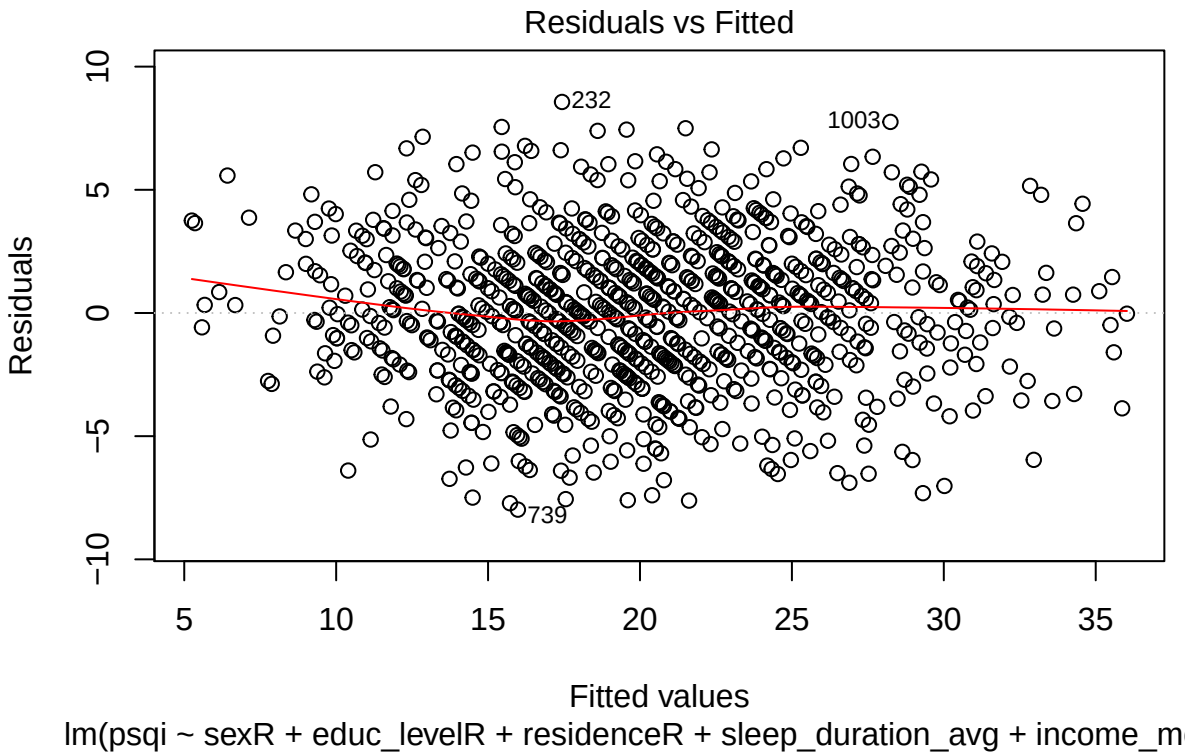
3.6 Ajust del model

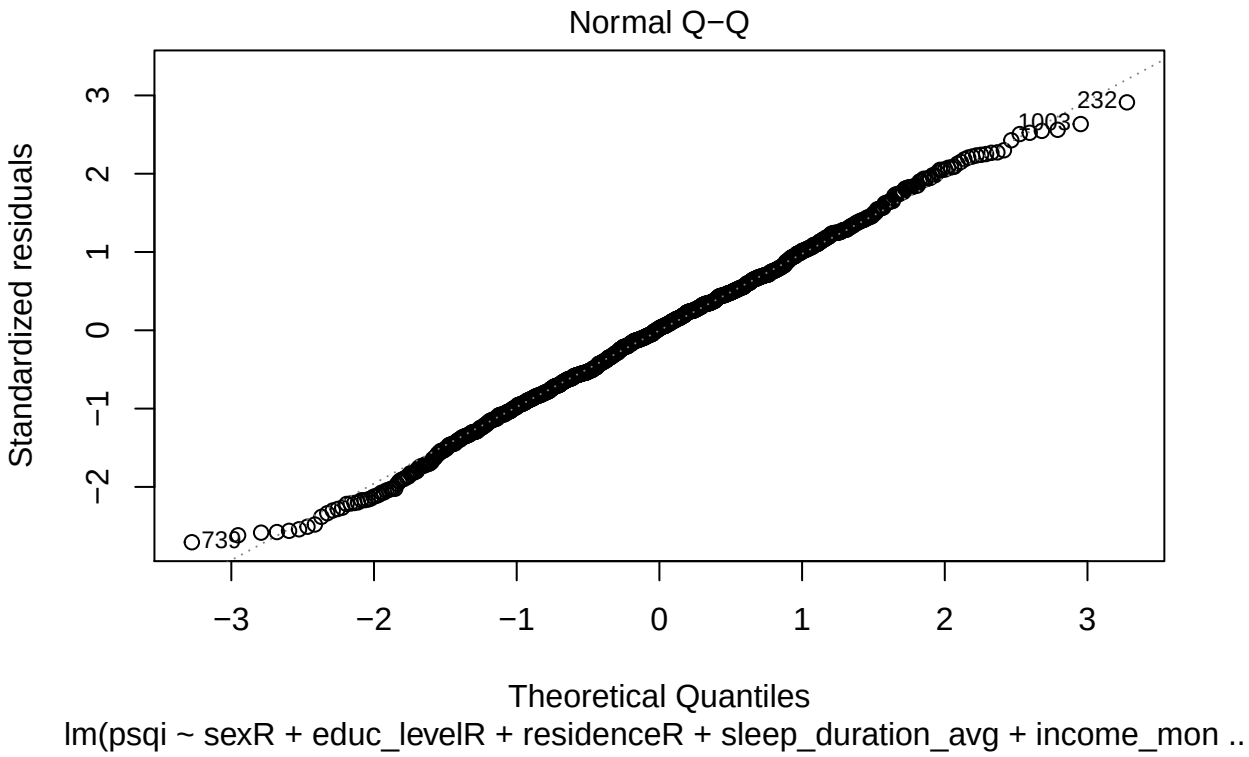
Analitzeu l'adequació del model a partir de l'anàlisi gràfic de residus. A tal efecte, es pot usar la instrucció plot() passant com a paràmetre el model de regressió lineal. Per saber com interpretar els gràfics de residus, podeu consultar l'enllaç següent: http://data.library.virginia.edu/diagnostic-plots/

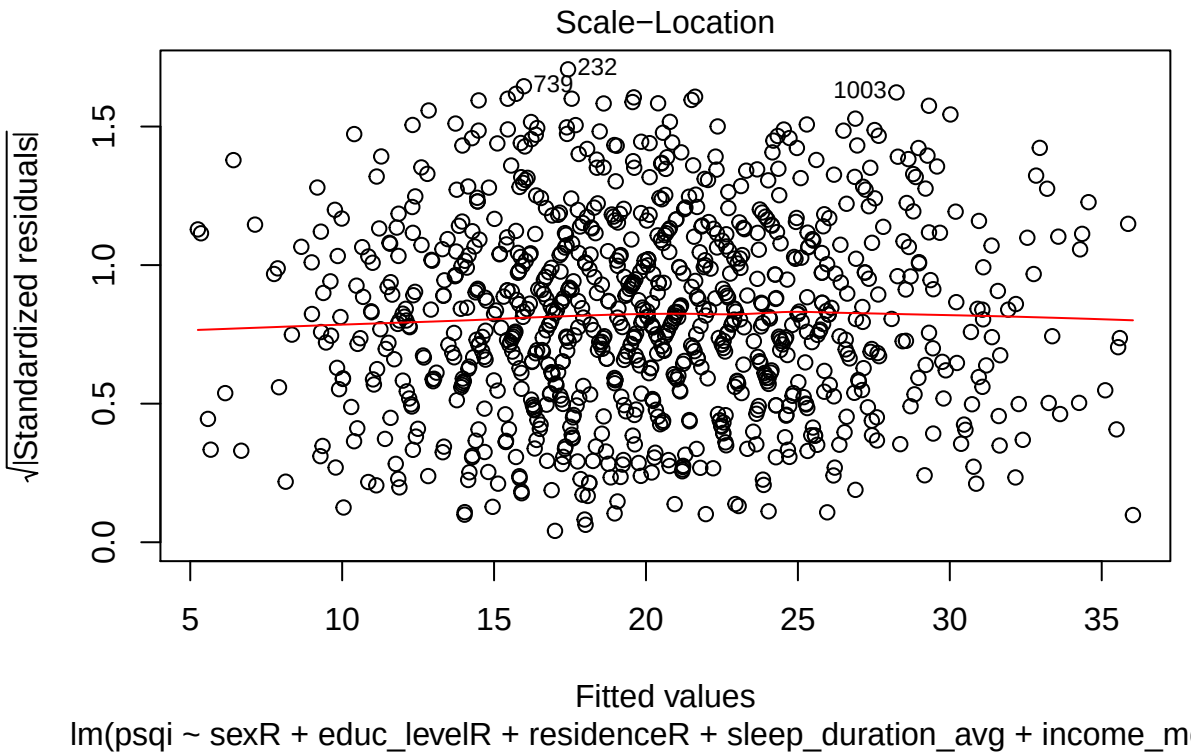
Segons l'anàlisi gràfic de residus, el model s'adequa prou bé.

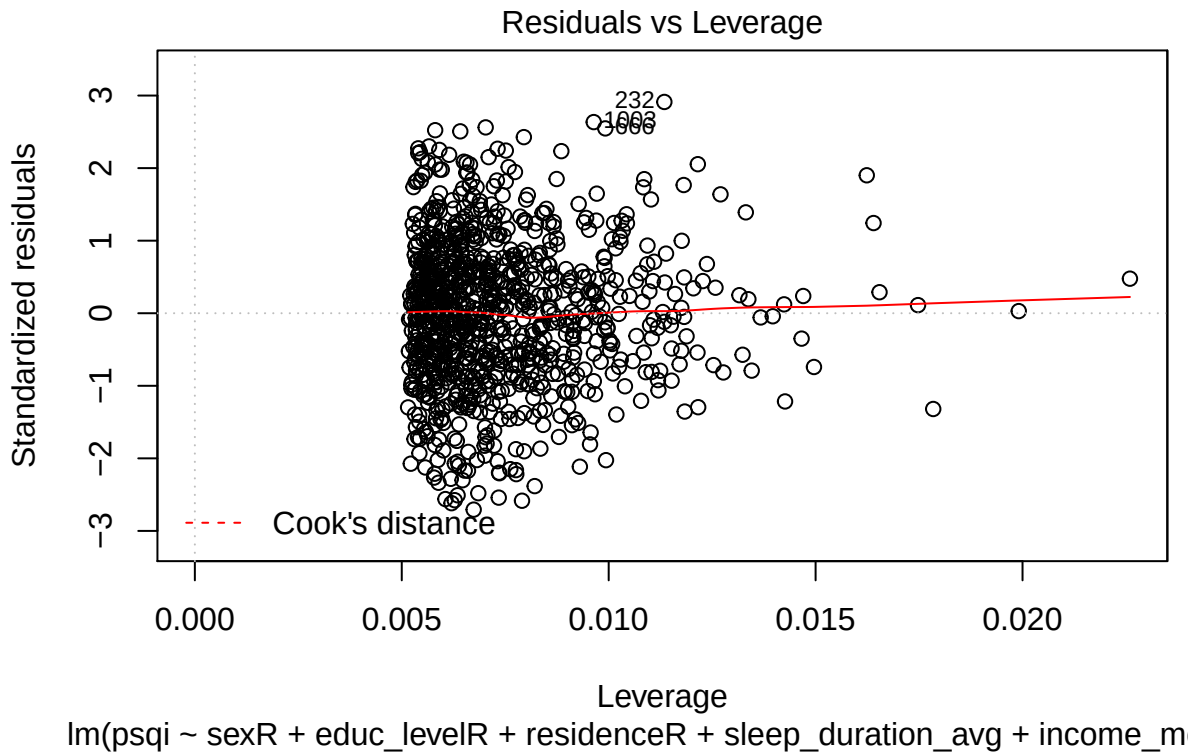
1. La gràfica "Residuals vs Fitted" és molt horizontal, separant l'espai en dues meitats molt similars. Això significa que els residuals segueixen un patró lineal.
2. Normal Q-Q. La gràfica mostra pràcticament una recta. Això vol dir que els residus estàn distribuits de forma normal.
3. Scale-Location. La línia és pràcticament horitzontal. Mostra que els residus estàn distribuits de forma similar a tot el rang.
4. Residuals vs Leverage. No tenim valors a les cantonades de la dreta. Això vol dir que no hi ha outliers.

```
plot(lmFrame)
```


Residuals vs Fitted
Residuals
Fitted values
lm(psqi ~ sexR + educ_levelR + residenceR + sleep_duration_avg + income_mon ..
232
1003
739
10
5
0
−5
−10
5
10
15
20
25
30
35


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(psqi ~ sexR + educ_levelR + residenceR + sleep_duration_avg + income_mon ..
232
1003
739
−3
−2
−1
0
1
2
3


Scale−Location
232
739
1003
√|Standardized residuals|
0.0
0.5
1.0
1.5
5
10
15
20
25
30
35
Fitted values
lm(psqi ~ sexR + educ_levelR + residenceR + sleep_duration_avg + income_mon ..



4 Anàlisi de variança unifactorial

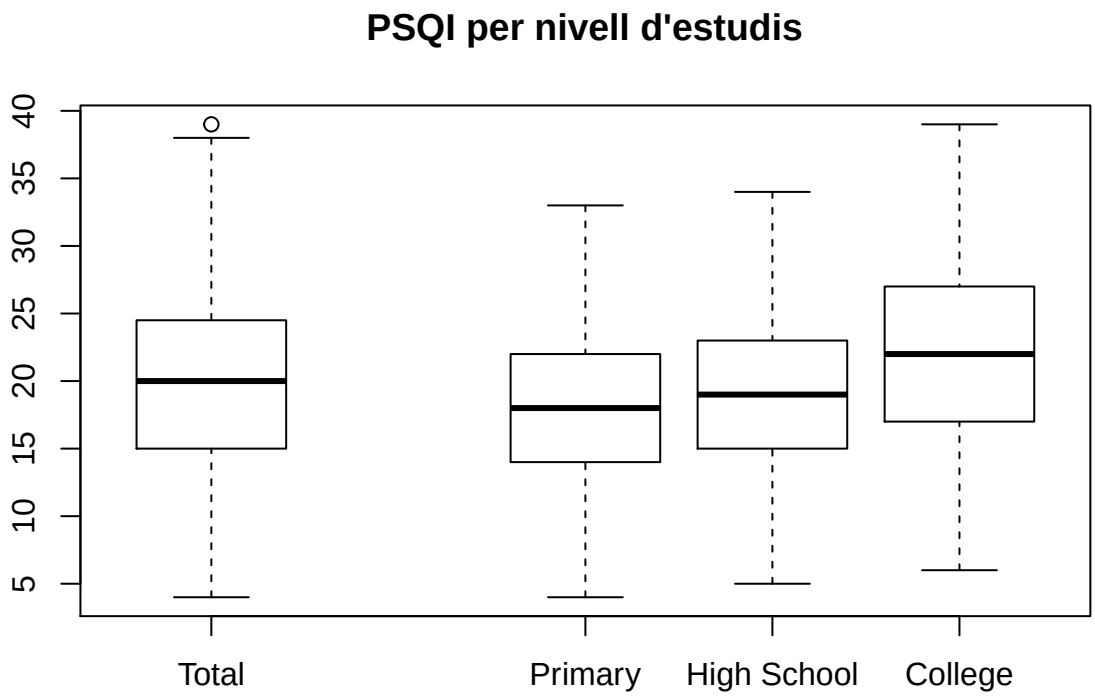
4.1 Hipòtesis nul · la i alternativa

L'indicador PSQI valora la qualitat del son. Com més elevat és aquest índex, pitjor és la qualitat. Es vol testejar si el nivell d'educació afecta a aquest índex. La nostra hipòtesi és que el nivell d'estudis és determinant. Escribim (els subíndexos es corresponen amb primària (p), high school (h) i college o superior(c)):

$$H_0(les\ mitjanes\ són|iguals) : \mu_p = \mu_h = \mu_c$$

$$H_1 : les\ mitjanes\ no\ són iguals$$

```
boxplot(psqi, CAD.educ_level_fac.P$psqi, CAD.educ_level_fac.HS$psqi, CAD.educ_level_fac.CH$psqi,
        main = "PSQI per nivell d'estudis",
        at = c(1,3,4,5),
        names = c("Total", "Primary", "High School", "College")
        )
```




```
# PSQI en funció del nivell d'educació
anova_educ_level <- aov(psqi ~ educ_level)
# mostrem les dades
summary(anova_educ_level)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## educ_level    2   2655    1328   34.04 5.26e-15 ***
## Residuals   949  37007      39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Com es pot veure, el p value (Pr(>F)) dóna com a resultat un valor molt proper a 0, clarament inferior a 0.05. Així doncs, rebutgem l'hipòtesi nul · la. Podem dir que el nivell d'estudis afecta a la qualitat de la són.

Per una altra banda, el sumatori "Sum Sq" de "educ_level" és el sumatori dins del grup. El de "Residuals" se correspon amb el sumatori entre grups. La distància entre grups és elevada de mitja. Això vol dir que hi ha diferència entre les mitges dels diferents grups; per tant, és un indicador de que hem de rebutjar l'hipòtei nul · la.

4.3 Càlculs

```
# mitja general (tots els valors) i nombre d'elements
m_cad <- mean(psqi)
n_cad <- nrow(CAD)
# mitja de primària, variança i nombre d'elements
m_p <- mean(CAD.educ_level_fac.P$psqi)
v_p <- sd(CAD.educ_level_fac.P$psqi)^2
```

```r
n_p <- nrow(CAD.educ_level_fac.P)
# mitja de high school, variança i nombre d'elements
m_h <- mean(CAD.educ_level_fac.HS$psqi)
v_h <- sd(CAD.educ_level_fac.HS$psqi)^2
n_h <- nrow(CAD.educ_level_fac.HS)
# mitja de college, variança i nombre d'elements
m_c <- mean(CAD.educ_level_fac.CH$psqi)
v_c <- sd(CAD.educ_level_fac.CH$psqi)^2
n_c <- nrow(CAD.educ_level_fac.CH)
# Entre grups (SQE)
SQE <- n_p * ((m_p - m_cad)^2) + n_h * ((m_h - m_cad)^2) + n_c * ((m_c - m_cad)^2)
SQE
```

```
## [1] 2655.123
```

```r
# SQE promig
k <- length(unique(educ_level_fac))
freedom_degrees <- k - 1
SQE_promig <- SQE /freedom_degrees
# SQE_pro
SQE_promig
```

```
## [1] 1327.561
```

```r
# SQD
SQD <- (n_p - 1) * v_p + (n_h - 1) * v_h + (n_c - 1) * v_c
SQD
```

```
## [1] 37006.84
```

```r
# SQD promig
freedom_degrees <- n_cad - k
SQD_promig <- SQD / freedom_degrees
SQD_promig
```

```
## [1] 38.99562
```

```r
# SQT
SQT <- SQD + SQE
SQT
```

```
## [1] 39661.97
```

```r
# Estadístic de prova F
f <- (SQE / (k-1)) / (SQD / (n_cad - k))
f
```

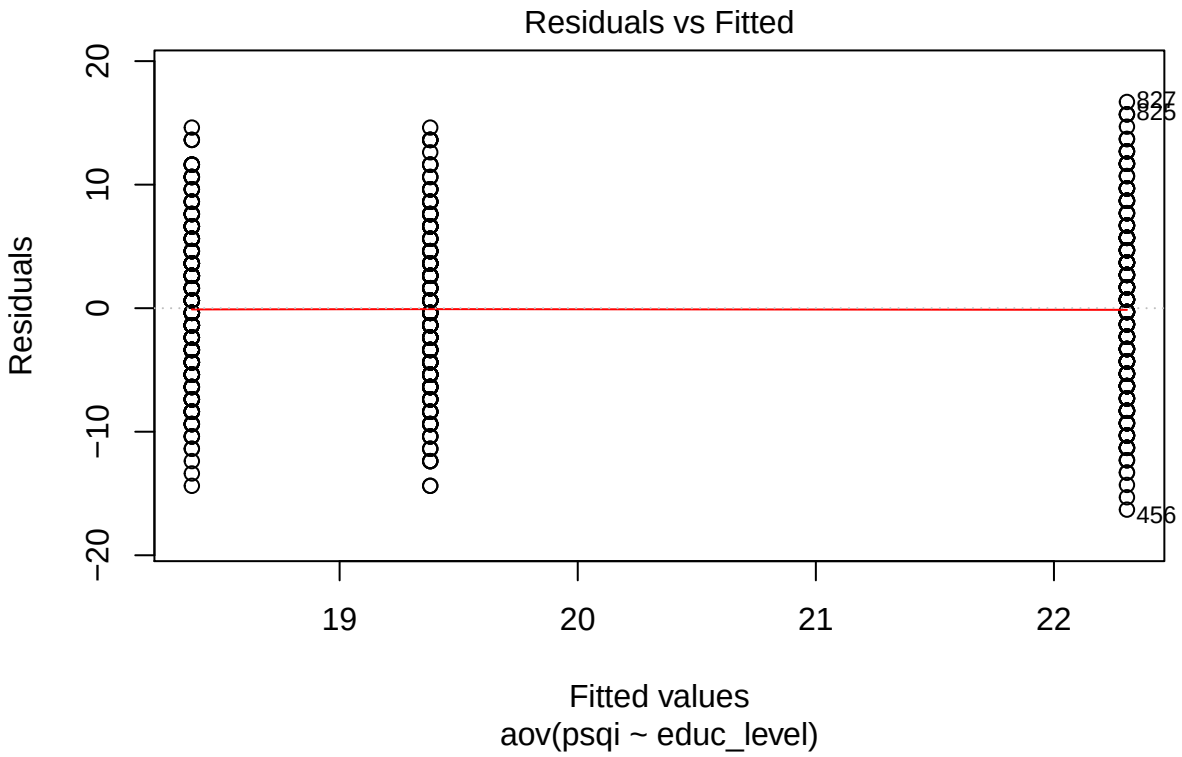
```
## [1] 34.04386
```

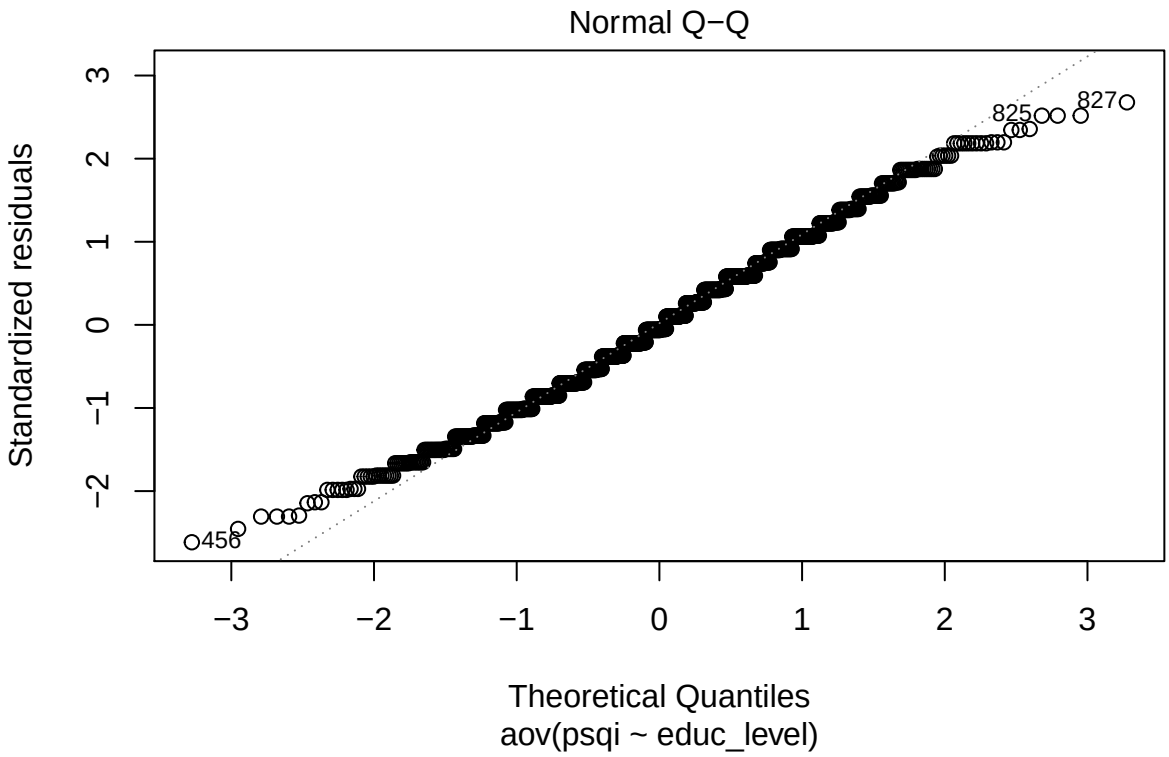
Com es pot veure, els resultats manuals quadren amb els que hem obtingut amb la funció.

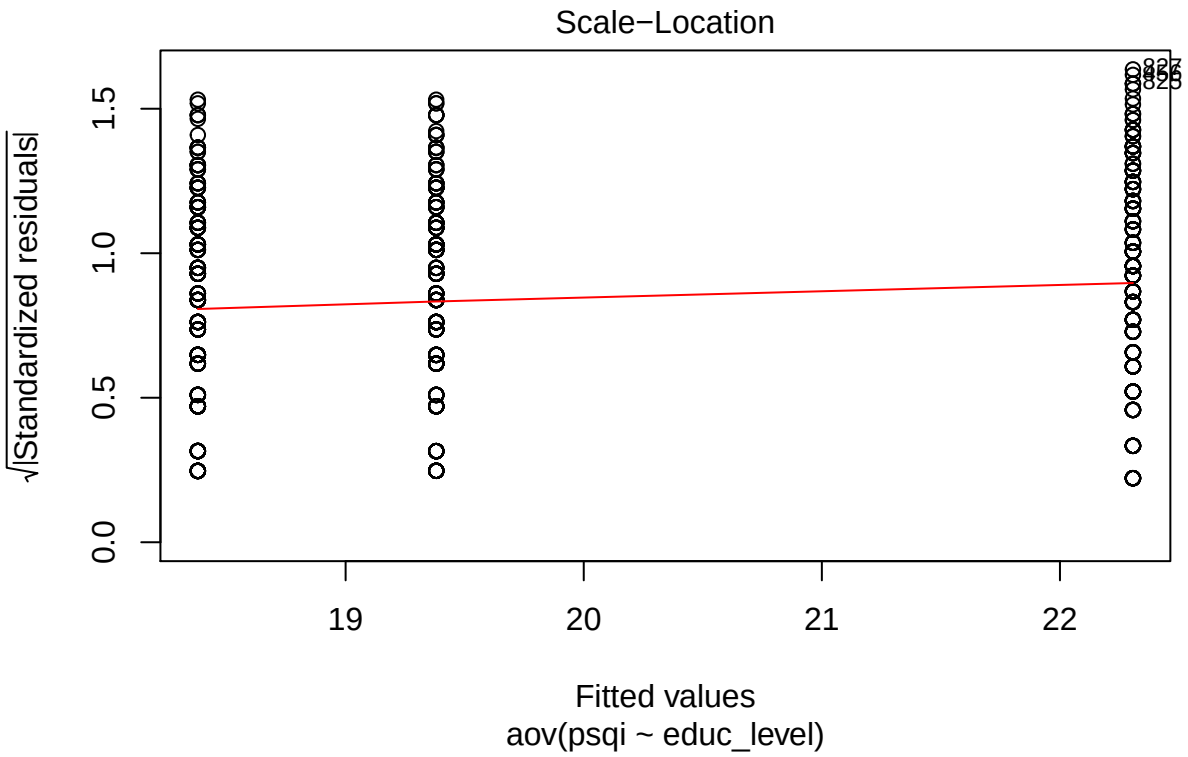
5 Adequació del model ANOVA

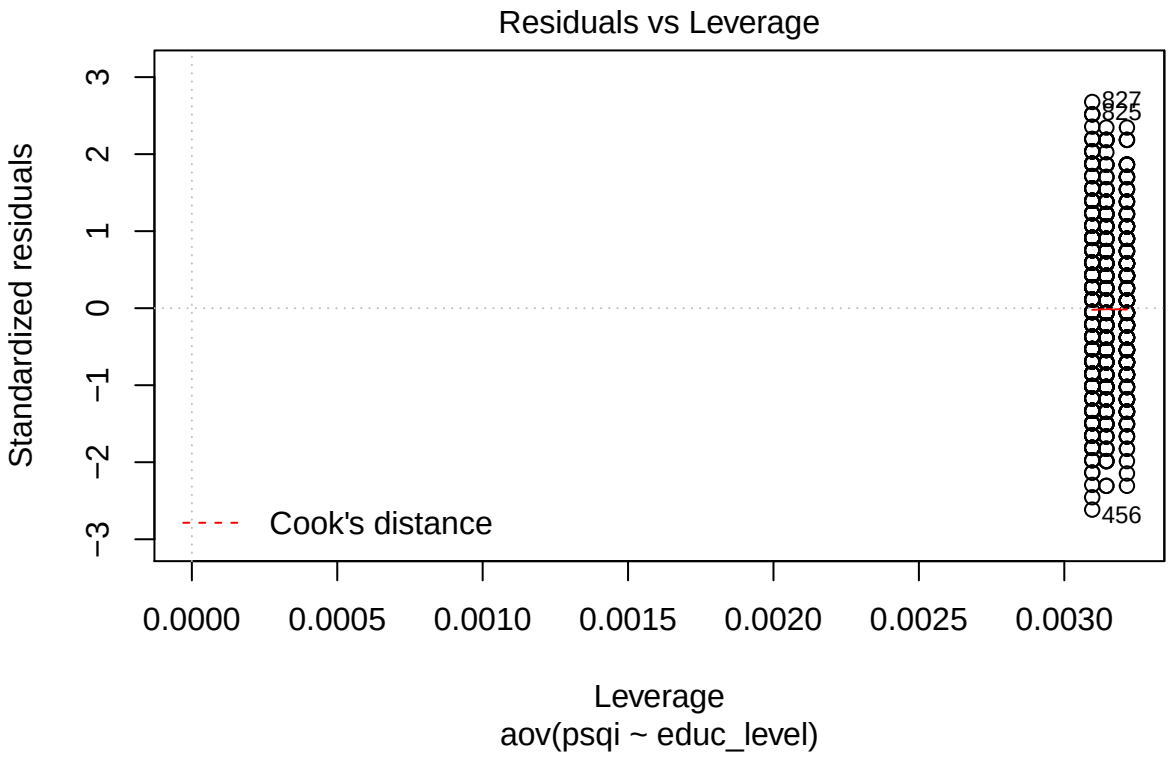
5.1 Visualització de l'adequació del model

```r
plot(anova_educ_level)
```


Residuals vs Fitted
Residuals
20
10
0
−10
−20
827
825
456
19
20
21
22
Fitted values
aov(psqi ~ educ_level)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(psqi ~ educ_level)
456
825
827
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
19
20
21
22
Fitted values
aov(psqi ~ educ_level)



5.2 Normalitat dels residus

Interpreteu la normalitat dels residus a partir del gràfic Normal Q-Q que es mostra en l'apartat anterior a partir del plot

Els residus segueixen una distribució molt propera a la normal. La diagonal del gràfic representa la distribució normal. Com més propera és la nostra distribució a la diagonal, més semblant és a la normal. Com es pot veure, la distribució s'allunya molt poc de la diagonal.

5.3 Homocedasticitat dels residus

Els gràfics Residuals vs Fitted i Scale-Location i Residuals vs Factor levels donen informació sobre la homocedasticitat dels residus. Interpreteu aquests gràfics

El gràfic residuals versus fits dibuixa els residuals a l'eix y i els valors correctes a l'eix x. Per a afirmar que la variància dels residus és constant (homocedasticitat), els valors han d'estar igualment distribuïts a ambdues bandes de l'eix de les abcisses. Visualment es veu que es prou similar. A la gràfica Scale Location es veu que la línea no és totalment horitzontal i que els valors no estan exactament igual distruïts. Sembla que hi ha més concentració de punts adalt que abaix. Això vol dir que l'homocedasticitat no és perfecta. La gràfica Residuals vs Leverage ens mostra que hi ha 3 valors outliers.

```
bptest(lmFrame)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lmFrame
## BP = 13.079, df = 6, p-value = 0.0418
```

5.4 ANOVA no paramètric

5.4.1 Kruskal-Wallis

```
# Executem el test
kruskal.test(psqi ~ educ_level, data = CAD)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  psqi by educ_level
## Kruskal-Wallis chi-squared = 54.563, df = 2, p-value = 1.418e-12
```

```
# Podem veure que p-value < 0.05, el que vol dir que hi ha, al menys,
#  dues distribución diferents
```

5.4.2 Interpretació

Expliqueu en què es diferencia el càlcul d'ANOVA del càlcul del test Kruskal-Wallis.

Pel que he pogut llegir[1] [2], tècnicament el test de Krustal-Wallis no compara valors, sino rànkings; recordem que s'ordenen els valors i es puntuen segons la posició que ocupen a aquesta llista. El test d'ANOVA és més robust (sí compara valors).
Dit això, el test de Krustal-Wallis dóna molt bon resultats i es prou robust.

6 ANOVA multifactorial

6.1 Factors: sexe i nivell educatiu

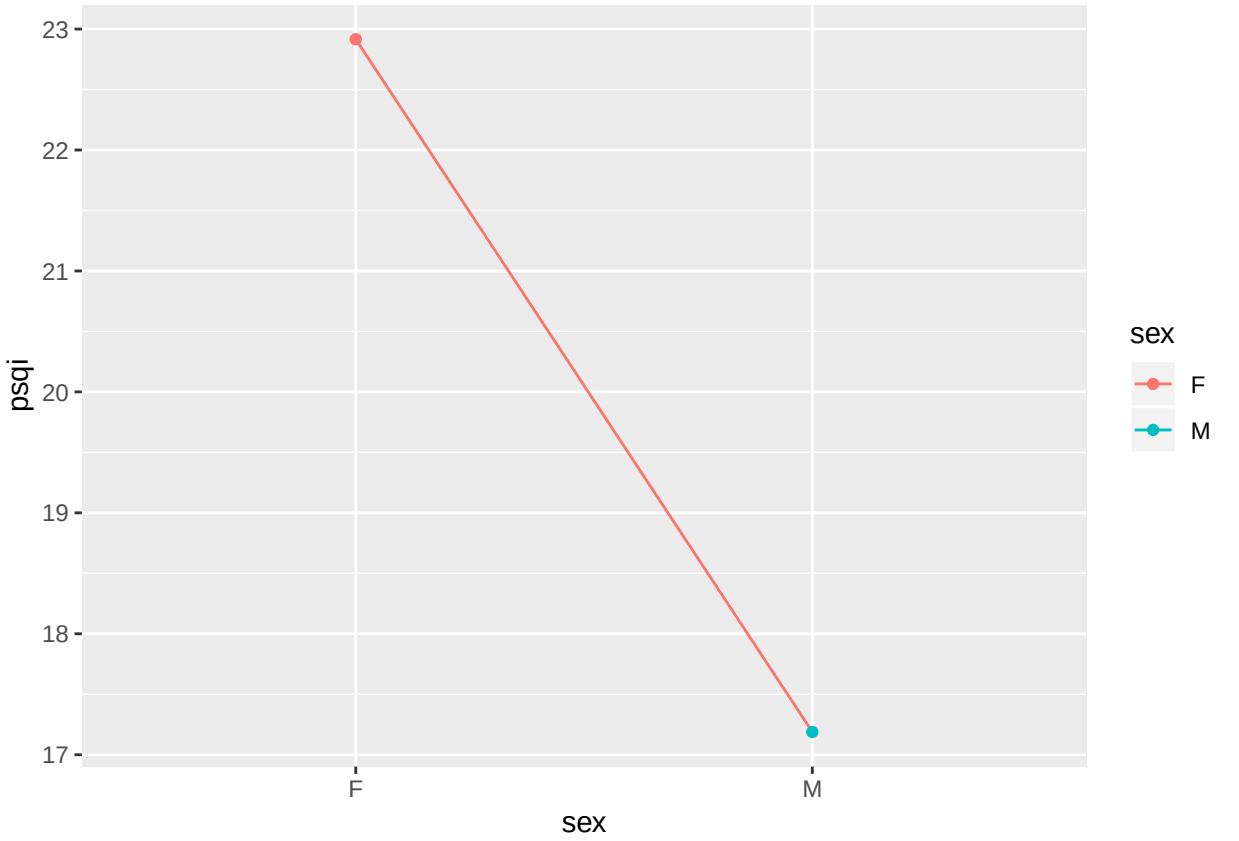
Anàlisi visual dels efectes principals i possibles interaccions

```
# Si agrupem per sexe, es pot veure com hi ha una diferència molt evident
group_sex <- group_by(CAD, sex)
result_sex <- as.data.frame(summarise(group_sex, psqi = mean(psqi)))
kable(result_sex)
```

sex	psqi
F	22.91579
M	17.18868

```
ggplot(result_sex, aes(x = sex, y = psqi, colour = sex, group = 1)) +
  geom_line() +
  geom_point()
```

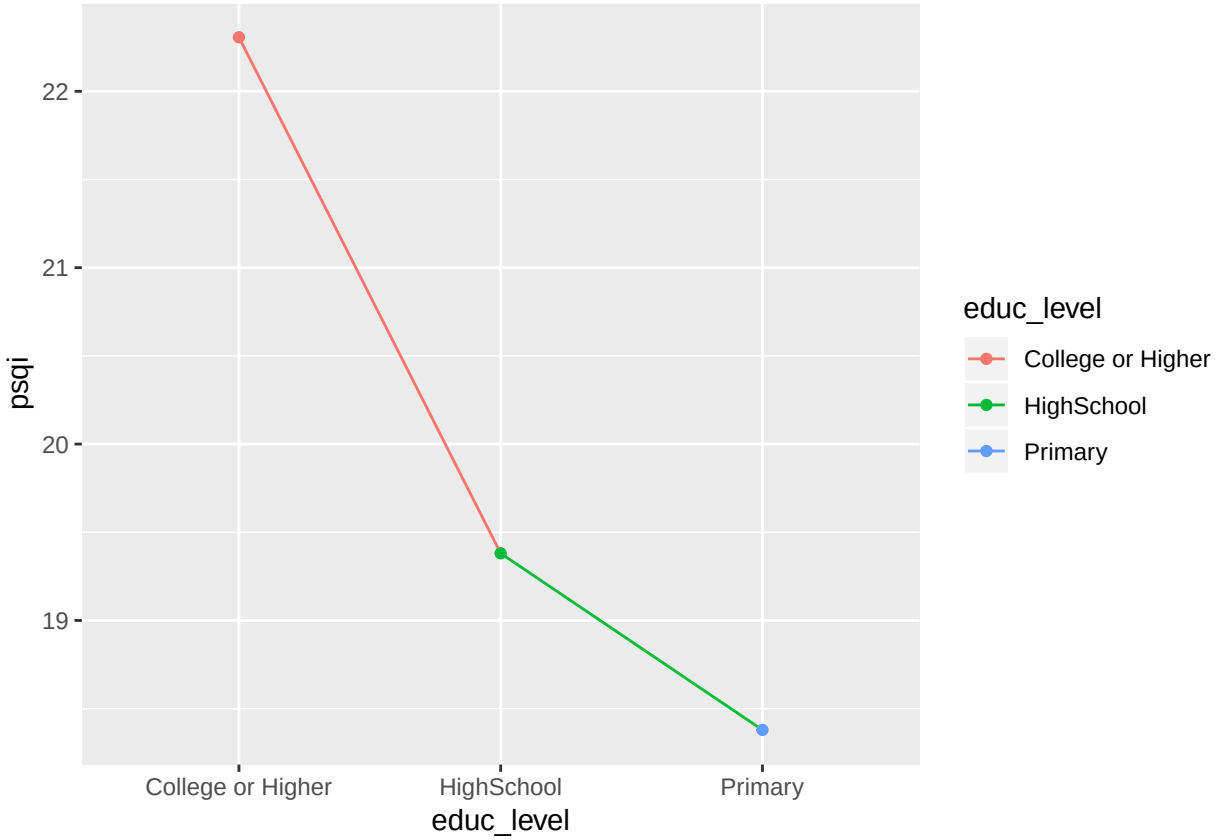
[1] stats.stackexchange.com/a/76080
[2] www.uclm.es/profesorado/mdsalvador/58109/teoria/anova_un_factor-lectura.pdf




```
# Si agrupem per nivell educatiu,
# es pot veure com hi ha una diferència molt evident en el cas
# de nivell d'estudis "College or Higher"
group_educ <- group_by(CAD, educ_level)
result_educ <- as.data.frame(summarise(group_educ, psqi = mean(psqi)))
kable(result_educ)
```

educ_level	psqi
College or Higher	22.30650
HighSchool	19.38050
Primary	18.37942

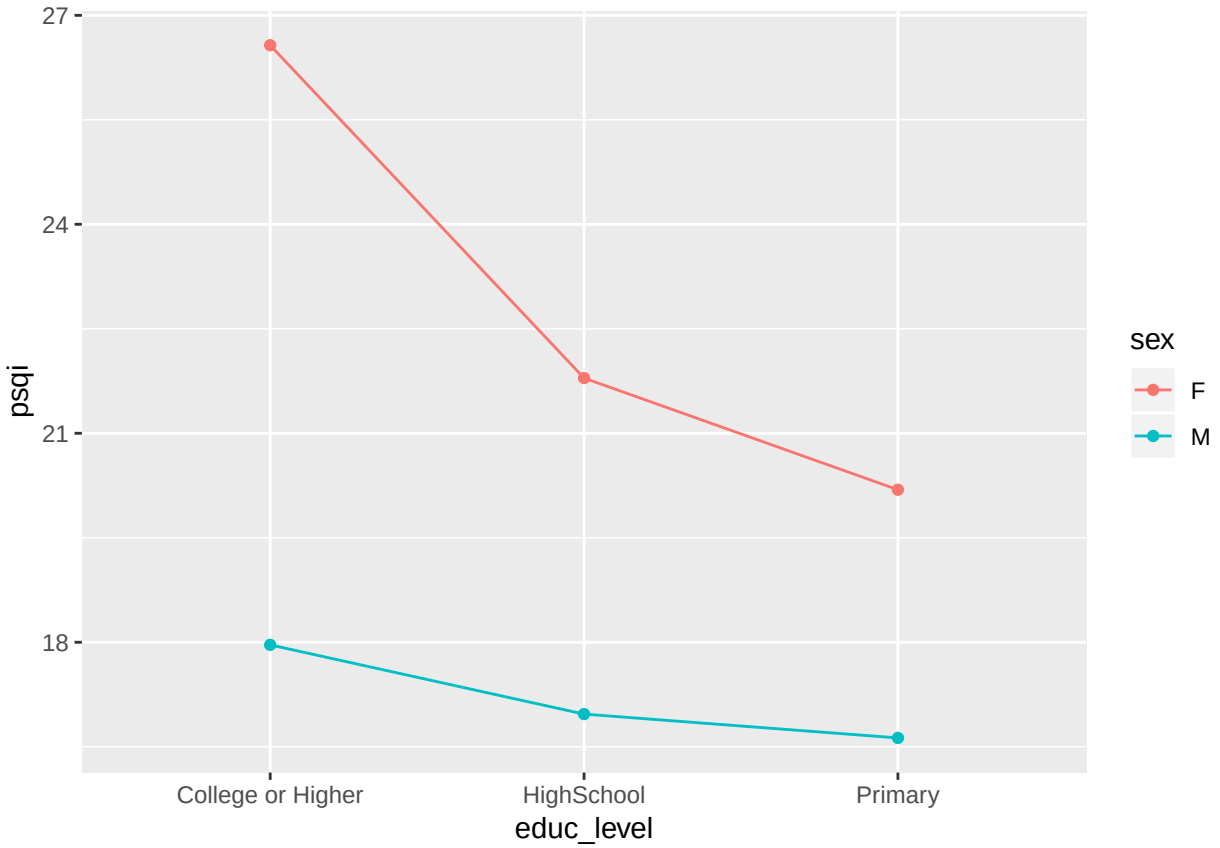
```
ggplot(result_educ, aes(educ_level, psqi, colour = educ_level, group = 1)) +
  geom_line() +
  geom_point()
```



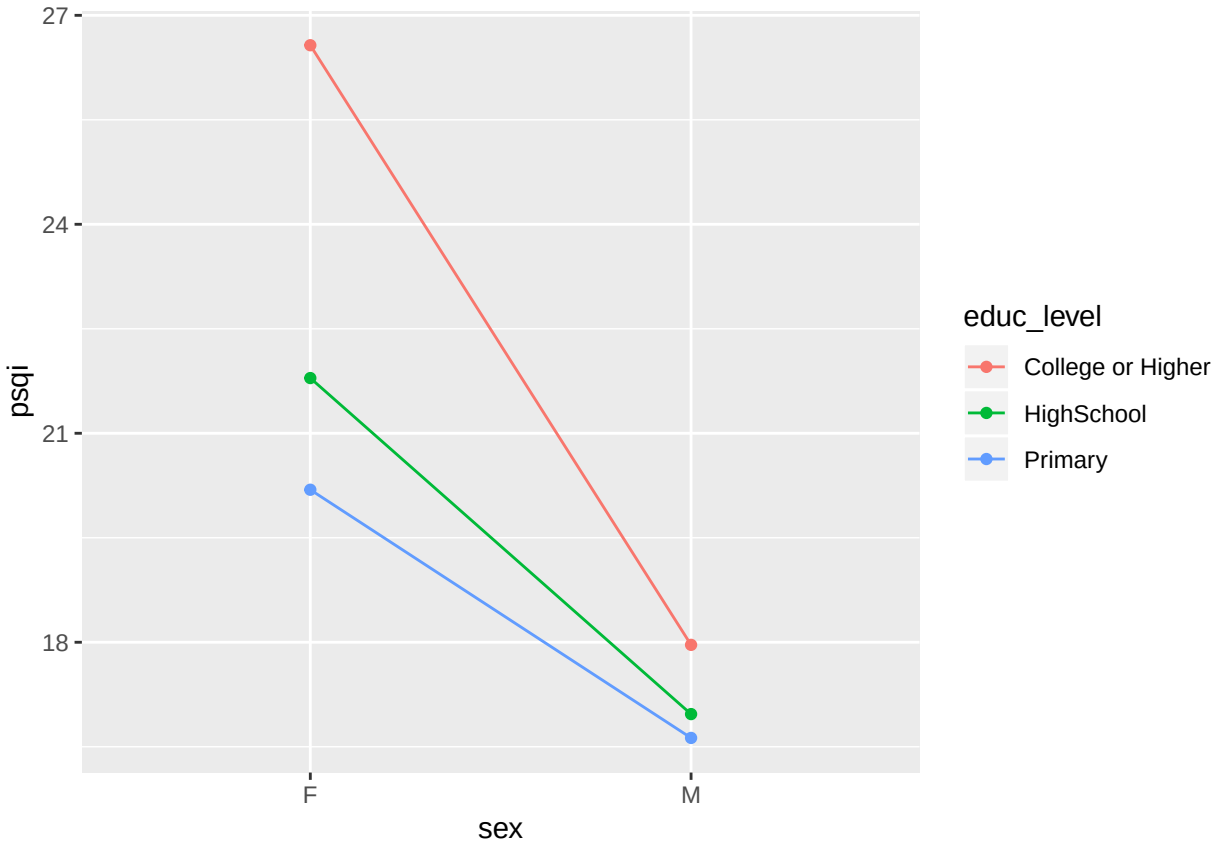

```
# Agrupeu el conjunt de dades por sexe i per nivell educatiu
group_sex_educ <- group_by(CAD, sex, educ_level)
result_sex_educ <- as.data.frame(summarise(group_sex_educ, psqi = mean(psqi)))
# Mostreu el conjunt de dades en forma de taula
kable(result_sex_educ)
```

sex	educ_level	psqi
F	College or Higher	26.57055
F	HighSchool	21.79245
F	Primary	20.18954
M	College or Higher	17.96250
M	HighSchool	16.96855
M	Primary	16.62658

```
# Mostreu en un gràfic el valor mig de la variable psqi per a cada sexe i educació.
ggplot(result_sex_educ, aes(educ_level, psqi, colour=sex, group = sex)) +
    geom_line() +
    geom_point()
```




```
ggplot(result_sex_educ, aes(sex, psqi, colour=educ_level, group = educ_level)) +
    geom_line() +
    geom_point()
```



Interpreteu el resultat sobre si existeixen efectes principals o hi ha interacció entre els factors

Com més horitzontals siguin les línies, menor és l'efecte de la variable de l'eix x (abscisses). Com major és la separació entre les línies, major és l'efecte de la variable que marquem amb colors. Podem veure que:

1. El sexe és determinant. És molt superior el valor de del dones (F) que dels homes (M).
2. El nivell d'estudis és determinant en el cas de les dones. La gràfica és molt més plana en el cas dels homes.
3. Podem veure la major pendent en cas de dones amb un elevat nivell d'estudis. La combinació d'ambdós valors és determinant.

6.1.2 ANOVA multifactorial

Realitzeu l'anàlisi ANOVA amb els factors sexe i nivell educatiu i la seva interacció. Interpreteu el resultat del model i expliqueu si els factors (i la interacció entre factors) són significatius per a modelar la variable psqi.

```
# Com ja hem calculat l'ANOVA de nivell educatiu,
#  no repetirem els càlculs

# ANOVA amb el factor nivell educatiu
summary(anova_educ_level)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## educ_level    2   2655    1328   34.04 5.26e-15 ***
## Residuals   949  37007      39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# ANOVA amb el factor sexe
anova_sex <- aov(psqi ~ sex)
summary(anova_sex)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## sex           1   7806    7806   232.8 <2e-16 ***
## Residuals   950  31856      34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# ANOVA amb el factor sexe i nivell educatiu
anova_sex_educ <- aov(psqi ~ educ_level + sex)
summary(anova_sex_educ)
```

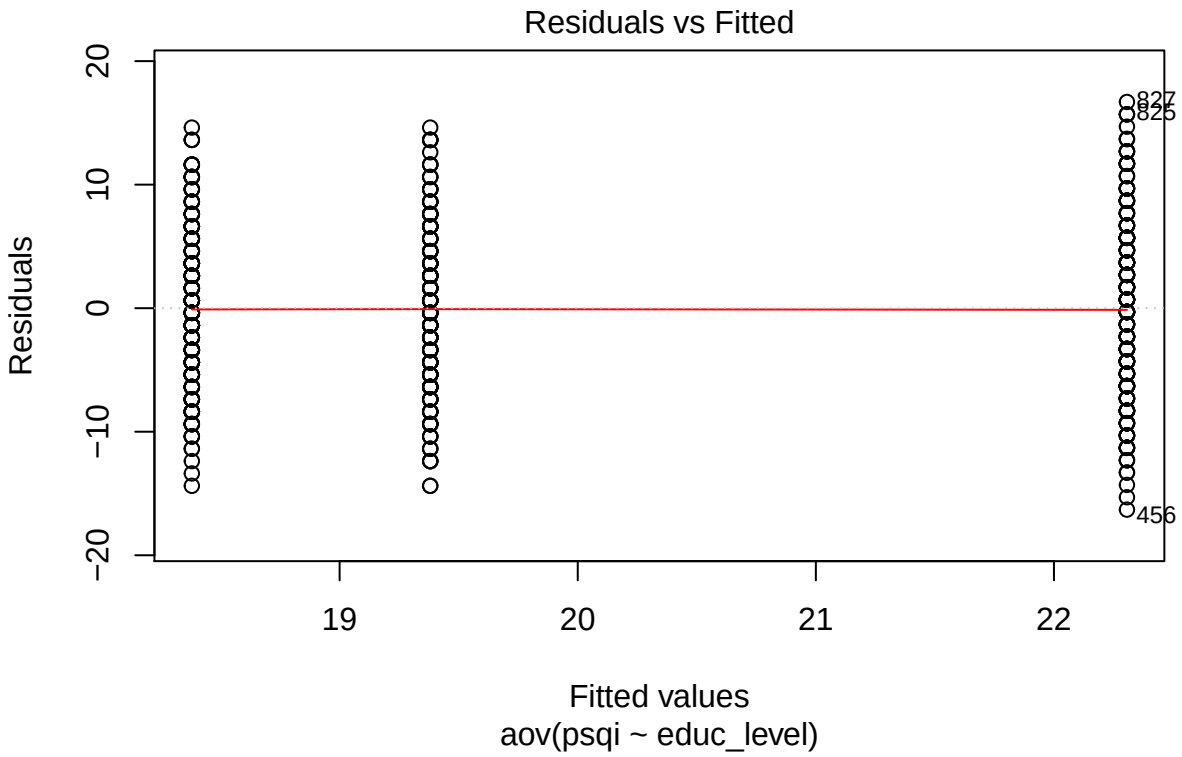
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## educ_level    2   2655    1328   42.97 <2e-16 ***
## sex           1   7721    7721  249.93 <2e-16 ***
## Residuals   948  29286      31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Com es pout veure, els factors sex i educ_level són significatius, tant individualment com de forma conjunta. El p value és sempre molt proper a 0 en tots els casos. Si mirem l'error mitjà, veurem que l'anàlisi que combina ambdós factors és lleugerament millor.

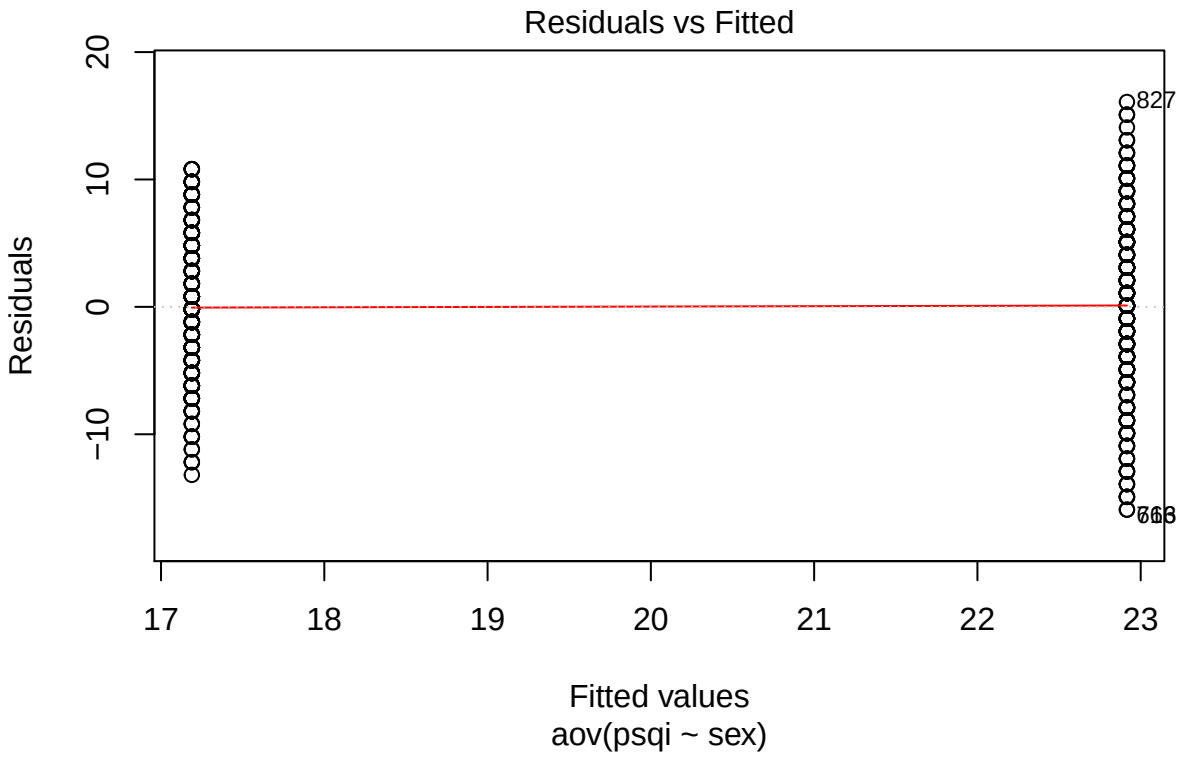
6.1.3 Adequació del model

Interpreteu l'adequació del model ANOVA obtingut usant els gràfics de residus.

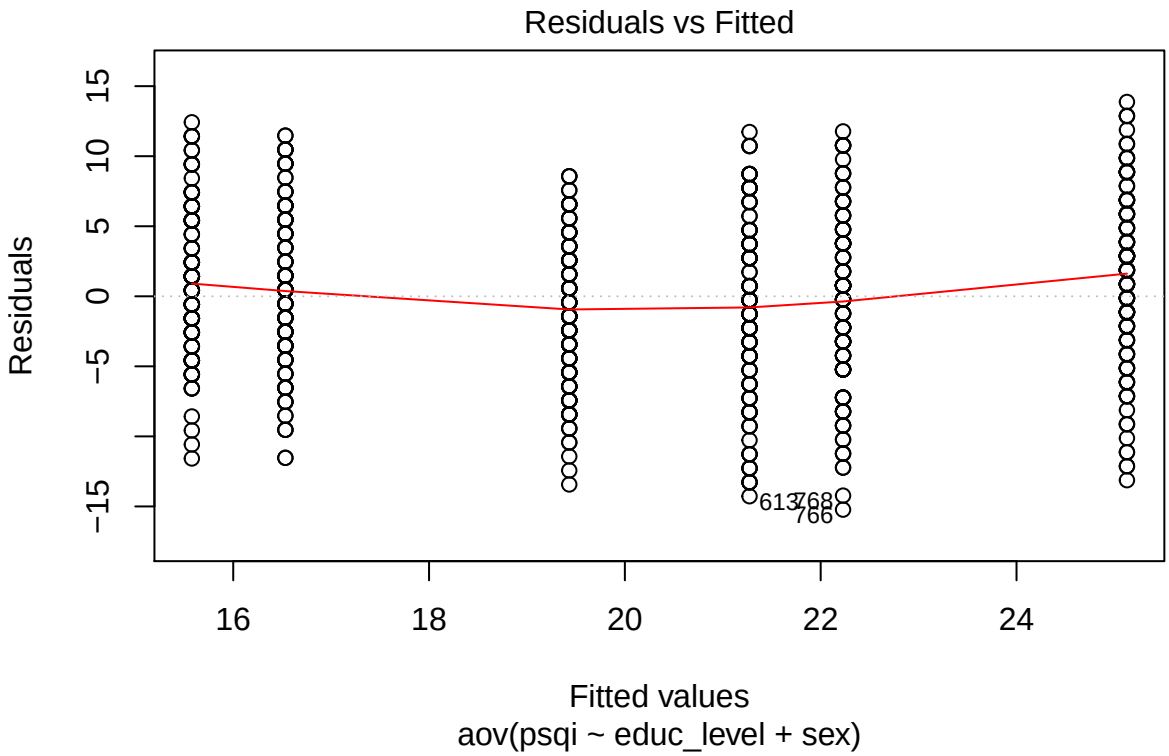
```r
# Gràfics de residus nivell educatiu
plot(anova_educ_level, 1)
```




```
# Gràfics de residus sexe
plot(anova_sex, 1)
```




```
# Gràfics de residus nivell educatiu i sexe
plot(anova_sex_educ, 1)
```



Visualment es pot veure que el model que es comporta pitjor (o que no té els residus igualment distribuits) és el que combina ambdós factors. La recta horitzontal fa un poc de corva cap abaix on sembla que hi ha més valors (i on són els outliers).

Les dades estan perfectament distruides en el cas anova per sexe (malgrat hi ha ouliers)

6.2 Factors: sexe i ingressos mensuals

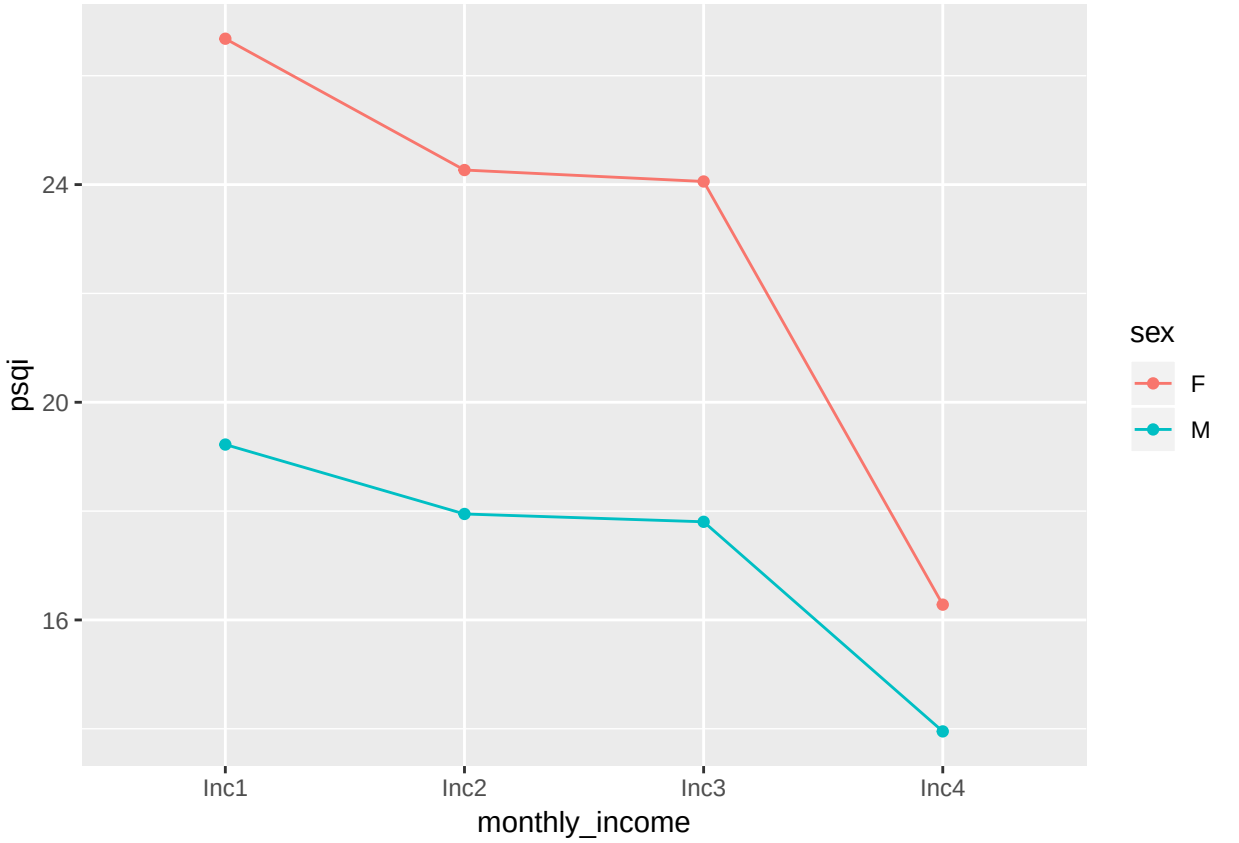
6.2.1 Anàlisi visual dels efectes principals i possibles interaccions

```
group_sex_ing <- group_by(CAD, sex, monthly_income)
result_sex_ing <- as.data.frame(summarise(group_sex_ing, psqi = mean(psqi)))
# Conjunt de dades en forma de taula
kable((result_sex_ing))
```

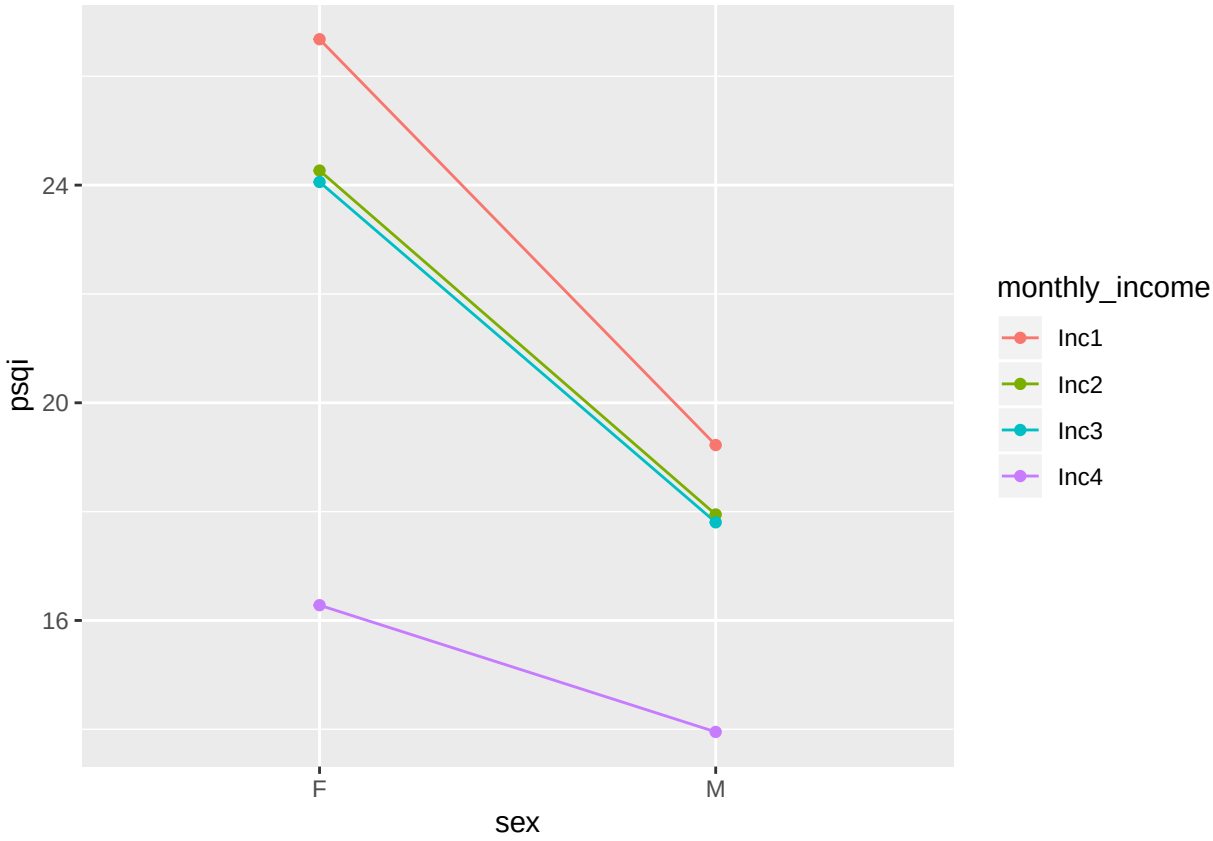
sex	monthly_income	psqi
F	Inc1	26.68033
F	Inc2	24.26724
F	Inc3	24.05691
F	Inc4	16.28070
M	Inc1	19.22314
M	Inc2	17.94737
M	Inc3	17.80342
M	Inc4	13.95200

```
# A la següent gràfica es pot veure com el sexe afecta en tots els casos
#  (en el cas d'ingressos Inc4 un poc menys que als altres casos)
# I que el nivell d'ingressos afecta, excepte en el pas Inc2 -> Inc3, que són
```

```r
#  molt similars
ggplot(result_sex_ing, aes(monthly_income, psqi, colour = sex, group = sex)) +
  geom_line() +
  geom_point()
```



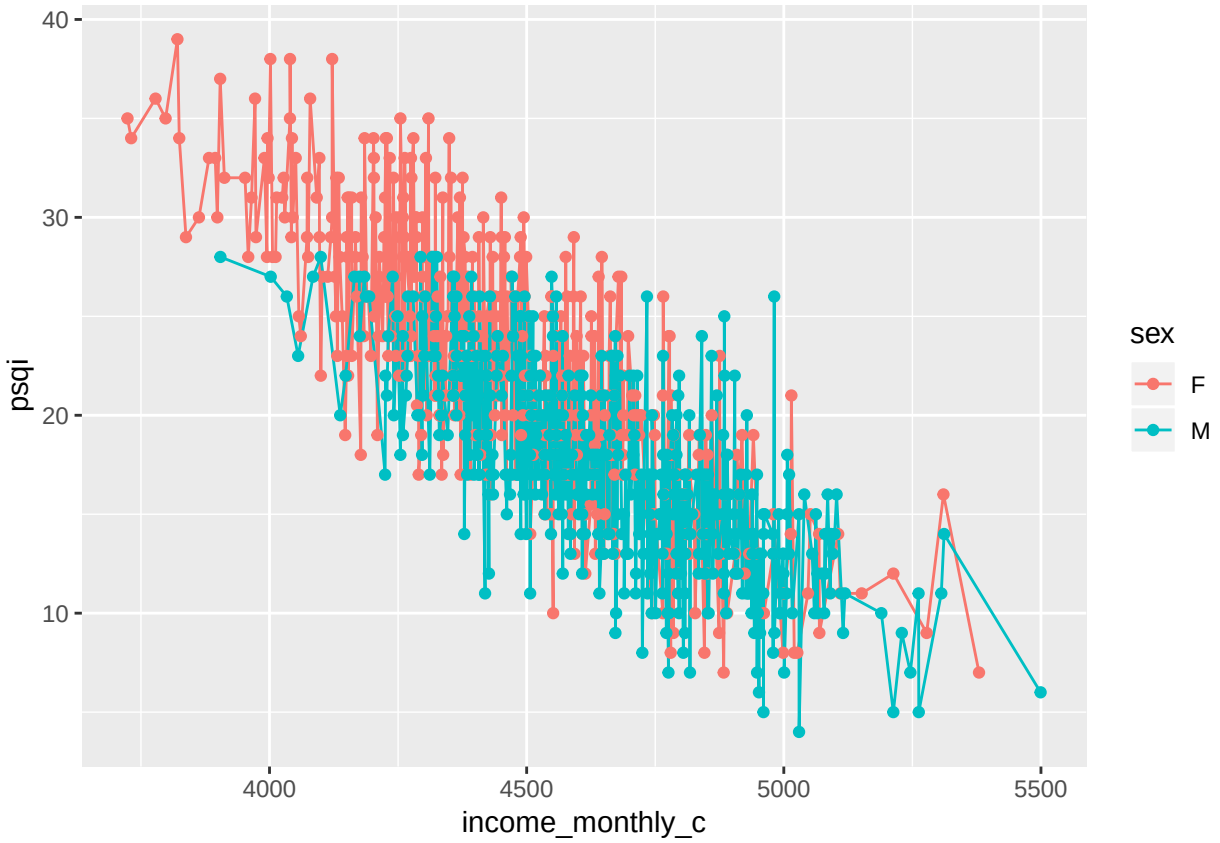

```r
# Els resultats són molt similars si feim l'anàlisi girant les variables:
#  El sexe és sempre rellevant
#  El nivell d'ingressos no ho és en el cas Inc2 -> Inc3; sí ho és en
#  qualsevol altre cas
ggplot(result_sex_ing, aes(sex, psqi, colour = monthly_income, group = monthly_income)) +
  geom_line() +
  geom_point()
```




```
# Com a afegit, he agafat l'altra variable d'ingressos (income_monthly_c) per
#  a veure el seu comportament.
#  Efectivament: qui  dorm pitjor són les dones amb baix nivell d'ingressos.
group_sex_inc <- group_by(CAD, sex, income_monthly_c)
result_sex_inc <- as.data.frame(summarise(group_sex_inc, psqi = mean(psqi)))
kable(head(result_sex_inc))
```

sex	income_monthly_c	psqi
F	3723.878	35
F	3730.983	34
F	3778.425	36
F	3797.938	35
F	3820.950	39
F	3824.520	34

```
ggplot(result_sex_inc, aes(x = income_monthly_c, y = psqi, colour = sex, group = sex)) +
  geom_line() +
  geom_point()
```



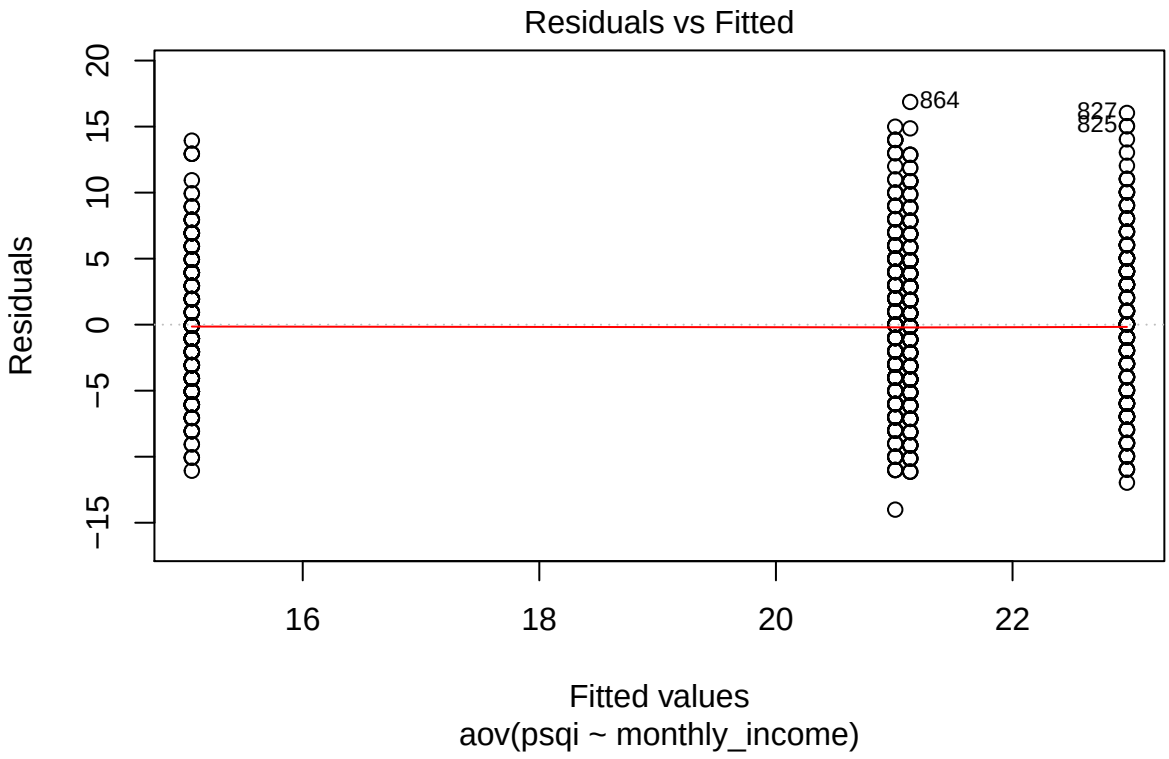
6.2.2 ANOVA multifactorial

Realitzeu l'anàlisi ANOVA amb els factors sexe i ingressos mensuals, i la interacció. Interpreteu el resultat del model i expliqueu si els factors (i la interacció entre factors) són significatius per a modelar la variable psqi.

```
anova_income <- aov(psqi ~ monthly_income)
summary(anova_income)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## monthly_income   3   8503  2834.5   86.24 <2e-16 ***
## Residuals      948  31159    32.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

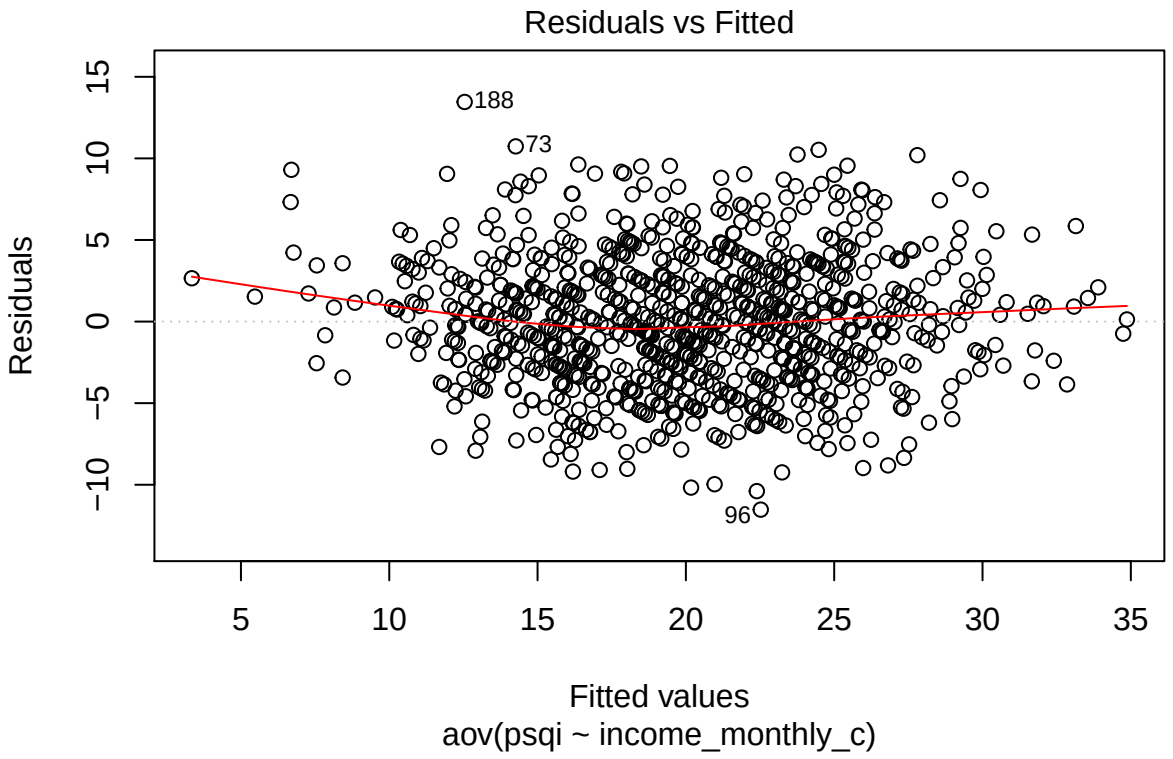
```
plot(anova_income, 1)
```




```r
anova_income_c <- aov(psqi ~ income_monthly_c)
summary(anova_income_c)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## income_monthly_c   1  24489   24489    1533 <2e-16 ***
## Residuals        950  15173      16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

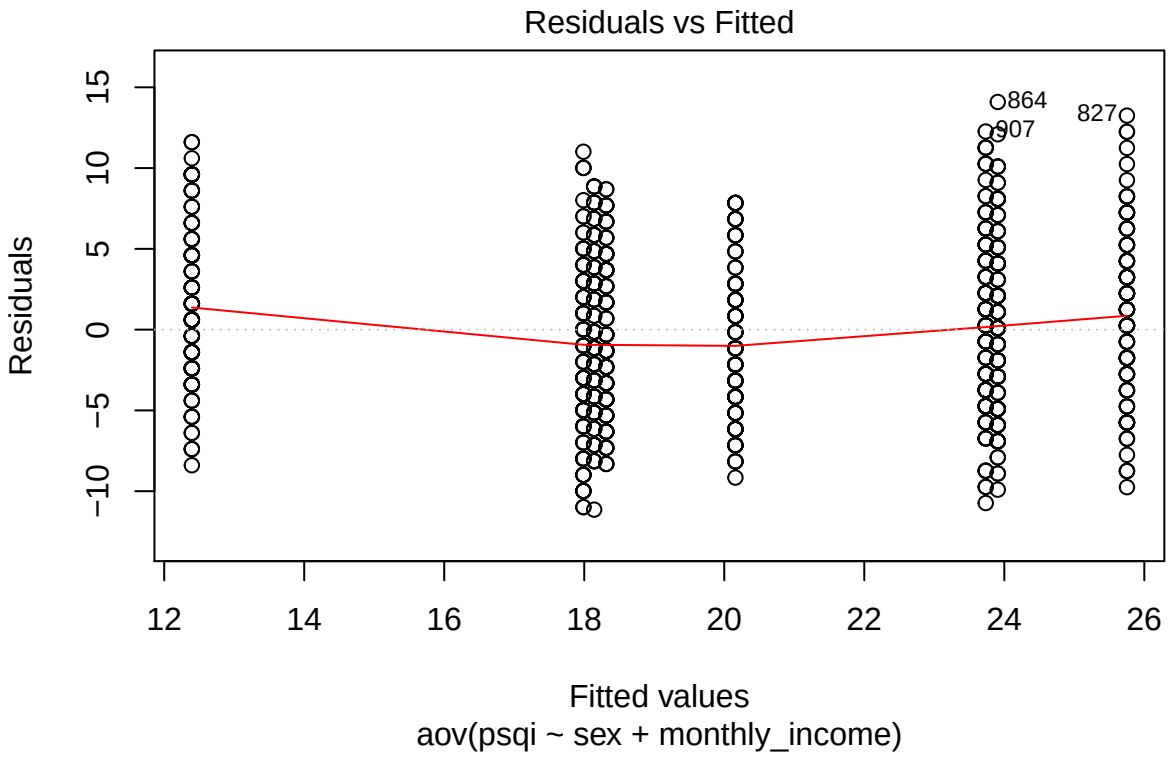
```r
plot(anova_income_c, 1)
```




```
anova_sex_income <- aov(psqi ~ sex+monthly_income)
summary(anova_sex_income)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## sex              1   7806    7806   311.7 <2e-16 ***
## monthly_income   3   8137    2712   108.3 <2e-16 ***
## Residuals      947  23719      25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

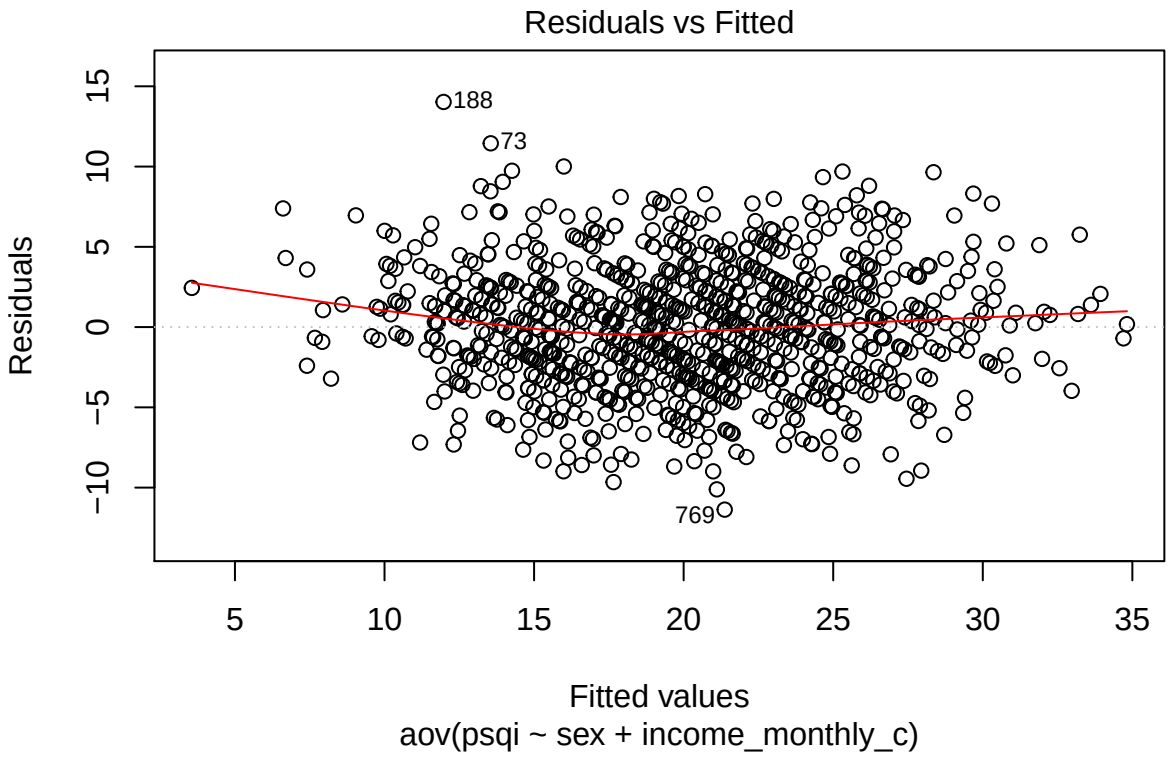
```
plot(anova_sex_income, 1)
```




```
anova_sex_income_c <- aov(psqi ~ sex+income_monthly_c)
summary(anova_sex_income_c)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## sex                1   7806    7806   530.5 <2e-16 ***
## income_monthly_c   1  17892   17892  1216.0 <2e-16 ***
## Residuals        949  13963      15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(anova_sex_income_c, 1)
```



7 Comparacions múltiples

Prenent com a referència el model ANOVA multifactorial, amb els factores sexe i nivell educatiu, apliqueu el test de comparació múltiple Scheffé. Interpreteu el resultat del test i indiqueu quins grups són diferents significativament entre si.

El test de Scheffé es una prova que s'aplica per fer comparacions múltiples de las mitges de grups. El seu ús està relacionat amb la prova de l'anàlisi de la variança, i se inclou dins de les denominadoes com a proves de comparacions múltiples[3].

```
STA <- ScheffeTest(anova_sex_income)
STA
```

```
##
##   Posthoc multiple comparisons of means : Scheffe Test
##     95% family-wise confidence level
##
## $sex
##         diff    lwr.ci    upr.ci   pval
## M-F -5.72711 -6.728314 -4.725906 <2e-16 ***
##
## $monthly_income
##                 diff    lwr.ci     upr.ci    pval
## Inc2-Inc1 -1.8454119 -3.266345 -0.4244785 0.00309 **
## Inc3-Inc1 -2.0185496 -3.424192 -0.6129067 0.00064 ***
```

[3] https://es.wikipedia.org/wiki/Prueba_de_Scheff%C3%A9

```
## Inc4-Inc1 -7.7607371 -9.167859 -6.3536155 < 2e-16 ***
## Inc3-Inc2 -0.1731377 -1.598383  1.2521075 0.99764
## Inc4-Inc2 -5.9153252 -7.342029 -4.4886216 < 2e-16 ***
## Inc4-Inc3 -5.7421875 -7.153663 -4.3307119 < 2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#scheffe_income <- scheffe.test(anova_sex_income, "monthly_income")
#scheffe_income
#ScheffeTest(anova_sex_income, g=sex)
```

El resultats són els ja esperats (sempre tenint en compte el p_value):

- El sexe és determinant. Que el p_value sigui tan petit vol dir que hi ha una diferència significativa entre els dos sexes.
- El nivell d'ingressos és determinant excepte en el cas Inc3-Inc2. El p_value és inferior al 5% en la resta de casos. En els parells Inc4-Inc1, Inc4-Inc2 i Inc4-Inc3 és pràcticament 0. Això indica que el grup Inc4 és significtivament diferent que la resta. Mirant la resta d'estadístiques, podem afirmar que dormen millor.

8 Conclusions

8.1. Escriure les conclusions finals de l'estudi en relació als objectius de la investigació.

Analitzant les variables, hem pogut veure que:

- És rellevant on es viu, si a l'entorn rural o urbà. Hem demostrat que és un valor rellevant i que es dorm millor a un entorn rural que a un urbà.
- És rellevan el sexe de la persona. Hem pogut provar que els homes dormen millor que les dones.
- És rellevant el nivell educatiu, especialment si es tenen estudis superiors. Les persones que tenen estudis superiors dormen significativament pitjor que la resta (especialment en el cas de les dones). Pràcticament no hi ha diferència entre institut i primària en el cas dels homes.
- És rellevant els nivell d'ingressos, excepte en el rang mig. Els que guanyen més, dormen mijor. Els que guanyen menys, dormen pitjor. En el rang intermig no hi ha grans diferències. Una altra vegada més, aqueste diferències són molt més significatives en el cas de les dones.

En resum: qui dorm pitjor és una dona amb un baix nivell d'ingressos, estudis superiors i que viu a la ciutat. Qui dorm millor és un home amb alts nivells d'ingressos, nivell d'estudis de primària i que viu a un entorn rural.