Estadística avançada
A4 Anàlisic estadística avançada

Noemi Lorente Torrelles

31 de diciembre, 2018

Contents

Introducció

En aquesta activitat es realitzarà una anàlisi sobre la satisfacció laboral dels treballadors d'una empresa.

En una primera fase de la investigació, s'escullen dades de 38 treballadors. L'objectiu de l'anàlisi és investigar si la satisfacció laboral dels treballadors de l'empresa està relacionada amb la qualificació del treball i amb el nivell d'estudis.

La mostra de 38 persones presenta una mitjana d'edat de 35.2 ś 1 anys, i una antiguitat a l'empresa de 8.34 ś 1.75 anys.

Les dades recollides per cada treballador són:

- el nivell d'estudis (1: sense estudis, 2: estudis primaris, 3: educació secundària o educació professional, 4: universitaris),
- el tipus de treball que realitza (Q: qualificat, PQ: poc qualificat),
- les hores que treballa a la setmana en promig,
- el sexe,
- i la satisfacció laboral del treballador (compresa entre 0 i 10) recollida a través d'un formulari.

La mostra està recollida en el fitxer adjunt amb l'activitat: "sat02.csv".

Per a realitzar aquesta anàlisi, se seguiran els passos següents:

- Es començarà l'estudi per una anàlisi descriptiva senzilla, juntament amb gràfics que poden donar una primera idea de les variables que influeixen en la satisfacció laboral.
- En els apartats 2 i 3, s'analitzarà si la satisfacció laboral dels treballadors està influïda pel tipus del treball (qualificat/poc qualificat). Per a fer-ho, s'aplicaran tests d'hipòtesis de dues mostres i anàlisi de regressió, revisant els conceptes que s'han tractat al llarg del curs.
- A continuació (apartats 4 i 5), es realitzarà una anàlisi ANOVA unifactorial, tenint en compte el nivell d'estudis com a factor que pot influir en la satisfacció laboral.
- A l'apartat 6, s'aplicarà l'anàlisi de variànçia tenint en compte dos factors (tipus de treball i nivell d'estudis; tipus de treball i sexe). S'estudiaran els efectes principals i les interaccions entre els factors.
- Finalment, s'aplicarà un test de comparació múltiple per a investigar quins grups de treballadors tenen una satisfacció laboral significativament diferent a la resta.

Notes importants a tenir en compte pel lliurament de l'activitat:

- És necessari lliurar el fixer Rmd i el fixer de sortida (PDF o html). El fitxer de sortida ha d'incloure el codi i el resultat de la seva execució (pas a pas). S'ha d'incloure en el document: el nom complet, el títol de l'activitat, i l'índex o taula de continguts. S'ha de respectar la numeració dels apartats de l'enunciat.
- No realitzeu llistats dels conjunts de dades, ja que aquests poden ocupar vàries pàgines. Si voleu comprovar l'efecte d'una instrucció sobre un conjunt de dades podeu usar la funció 'head' que mostra les primeres 10 files del conjunt de dades.

1 Anàlisi descriptiva i visualització

a) Realitzeu una primera anàlisi descriptiva de les dades de la mostra.

Abans de carregar el fitxer, he visualitzat el seu contingut amb *gedit* i he pogut comprobar que s'utilitza la coma "," com a separador de camps. Així m'asseguro que la lectura del fitxer es realitza de forma correcta.

```
# carrego el fitxer amb read.table, separador=',', decimal='.'
satlab <- read.table("sat02.csv", header=TRUE, sep=",",
                         na.strings="NA", dec=".", strip.white=TRUE,
                         stringsAsFactors = FALSE)
```

Amb la funció **str** puc veure l'estructura interna del data frame *satlab*. Veig que té 38 observacions, 5 variables i es mostra el nom de les variables:

- *Wtype*: el tipus de treball que realitza (Q: qualificat, PQ: poc qualificat),
- *EType*: el nivell d'estudis (1: sense estudis, 2: estudis primaris, 3: educació secundària o educació professional, 4: universitaris),
- *S*: satisfacció laboral del treballador (compresa entre 0 i 10),
- *Sex*: gènere,
- *H*: hores que treballa a la setmana en promig.

```
# la funció str mostra l'estructura interna del data frame satlab
str(satlab)
```

```
## 'data.frame':    38 obs. of  5 variables:
##  $ Wtype: chr  "PQ" "PQ" "PQ" "PQ" ...
##  $ Etype: int  1 3 2 1 2 1 1 2 4 1 ...
##  $ S    : num  1.32 2.44 7.75 4.61 8.7 ...
##  $ Sex  : chr  "F" "F" "F" "F" ...
##  $ H    : num  38 26.8 31.5 23.8 27.7 ...
```

Com R no ha assignat correctament el tipus apropiat a les variables qualitatives nominals *Sex*, *Etype* i *Wtype*, cal fer la conversió de caràcter a factor.

```
# Canvio a variable factor la variable Sex
satlab$Sex <- as.factor(satlab$Sex)
```

```
# Canvio a variable factor la variable Etype
satlab$Etype <- as.factor(satlab$Etype)
```

```
# Canvio a variable factor la variable Wtype
satlab$Wtype <- as.factor(satlab$Wtype)
```

Per conèixer el tipus de variable utilitzo la funció **class**. A continuació, mostro el tipus de variable de cada variable. Per mostrar els tipus de les variables en una taula, utilitzo la funció **kable** .

```
# Recupero el tipus de variable
tipus <- sapply(satlab,class)
kable(data.frame(Variable=names(tipus),Classe=as.vector(tipus)), align='l',
      caption="Tipus de les variables")
```

Table 1: Tipus de les variables

Variable	Classe
Wtype	factor
Etype	factor
S	numeric
Sex	factor
H	numeric

b) Mostreu en un diagrama de caixa la distribució de la satisfacció laboral de la mostra.

```
# Genero boxplot de la variable S, satisfacció laboral
boxplot(satlab$S, main="Satisfacció laboral", horizontal=TRUE, col="gold")
```

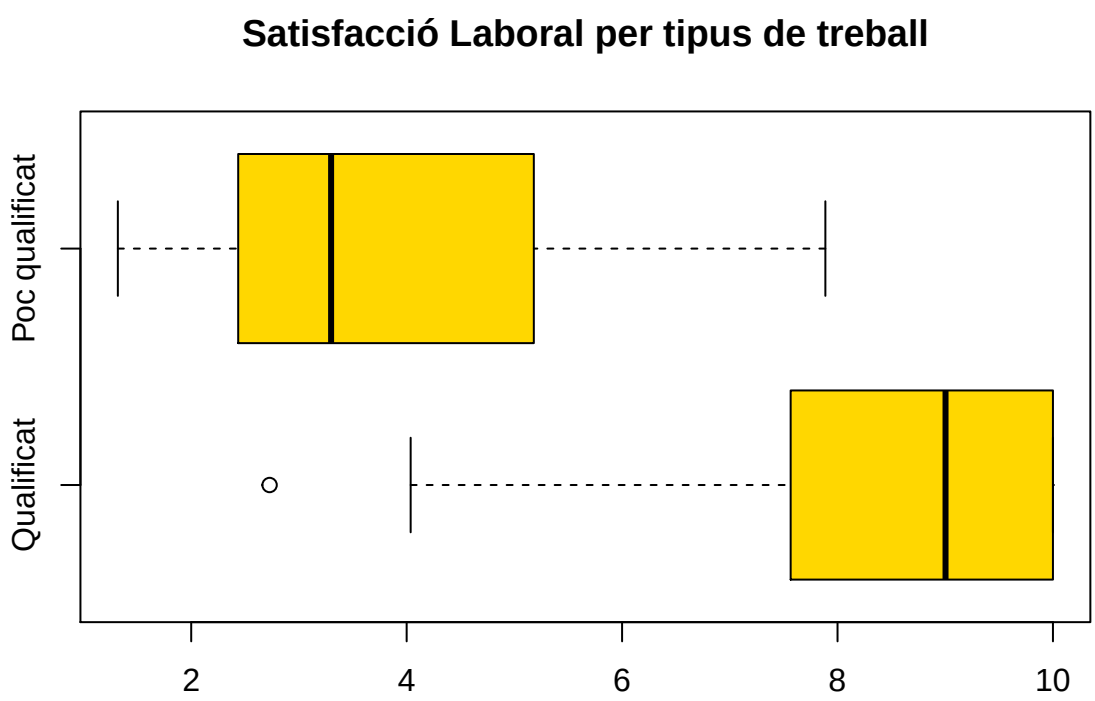
c) Mostreu en varis diagrames de caixa la distribució de la satisfacció laboral segons el tipus de treball, segons el nivell educatiu i segons el sexe, respectivament. Interpreteu els gràfics breuement.

La satisfacció laboral dels treballadors amb tipus de treball 'Qualificat' és molt més alta (*mediana* = 9) que per als treballadors amb tipus de treball 'Poc qualificat' (*mediana* = 3.3).

El rang interquartílic de tipus treball 'Qualificat' és igual a 2.18 (del $Q_1 = 7.82$ al $Q_3 = 10$) sent el valor mínim igual 2.72 El rang interquartílic de tipus treball 'Poc qualificat' és igual a 2.55 (del $Q_1 = 2.49$ al $Q_3 = 5.04$) sent el valor mínim igual 1.32 i el valor màxim igual a 7.89

Observo un outlier en el boxplot del tipus de treball 'Qualificat'.

```
#  Genero boxplot de la variable S en funció del tipus de treball, Wtype
par(mfrow=c(1,1))
idxQ <- satlab[satlab$Wtype=='Q',]
idxPQ <- satlab[satlab$Wtype=='PQ',]
boxplot(idxQ$S,idxPQ$S, main="Satisfacció Laboral per tipus de treball",
         names=c("Qualificat","Poc qualificat"), horizontal=TRUE, col="gold")
```



```r
# Executo summary i calculo el rang interquartílic de S per tipus treball 'Qualificat'
summary(idxQ$S)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.728   7.821   9.003   8.203  10.000  10.000
```

```r
IQR(idxQ$S)
```

```
## [1] 2.179331
```

```r
# Executo summary i calculo el rang interquartílic de S per tipus treball 'Poc qualificat'
summary(idxPQ$S)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.319   2.485   3.299   3.928   5.038   7.887
```

```r
IQR(idxPQ$S)
```

```
## [1] 2.553332
```

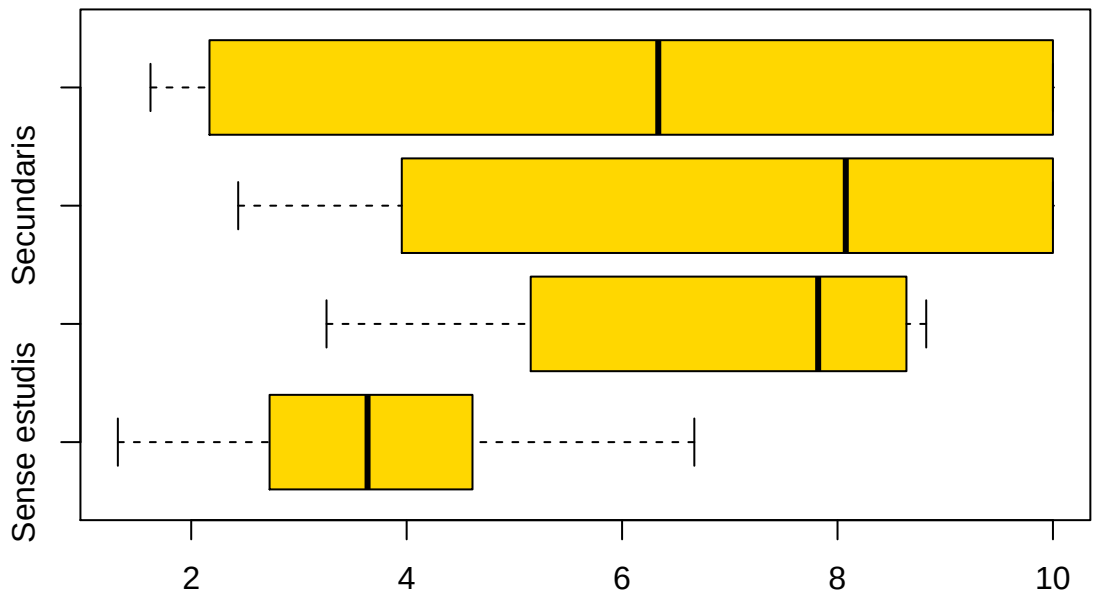
Per altra banda, observo que la satisfacció laboral de les persones amb nivell educatiu 'Sense estudis' és molt menor ($mediana = 3.64$) que les persones amb nivell educatiu 'Estudis primaris' i 'Estudis secundaris' ($mediana_{primaris} = 7.82$ i $mediana_{secundaris} = 8.08$).

I les persones amb nivell educatiu 'Universitaris' tenen una mediana de satisfacció menor que els 'primaris' i 'secundaris' ($mediana_{universitaris} = 6.08$).

La satisfacció de les persones amb nivell educatiu 'Sense estudis' està més concentrada, Q1=2.73 i Q3=4.61. En canvi, la satisfacció de les persones amb nivell educatiu 'Universitaris' és més variada, ja que es veu una major distància del rang inerquartílic, Q1=2.28 i Q3=10.

No observo outliers en cap dels boxplots que mostren la satisfacció laboral per nivell educatiu.

```r
#  Genero boxplot de la variable S separant les dades per nivell educatiu
par(mfrow=c(1,1))
idxE1 <- satlab[satlab$Etype=='1',]
idxE2 <- satlab[satlab$Etype=='2',]
idxE3 <- satlab[satlab$Etype=='3',]
idxE4 <- satlab[satlab$Etype=='4',]
boxplot(idxE1$S,idxE2$S,idxE3$S,idxE4$S, main="Satisfacció Laboral per nivell educatiu",
         names=c("Sense estudis","Primaris","Secundaris", "Universitaris"),
        horizontal=TRUE, col="gold")
```




```
# Executo summary de S per nivell educatiu
summary(idxE1$S)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.319   2.728   3.637   3.877   4.611   6.671
```

```
summary(idxE2$S)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.256   6.074   7.820   6.913   8.611   8.824
```

```
summary(idxE3$S)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.436   3.955   8.077   7.037  10.000  10.000
```

```
summary(idxE4$S)
```

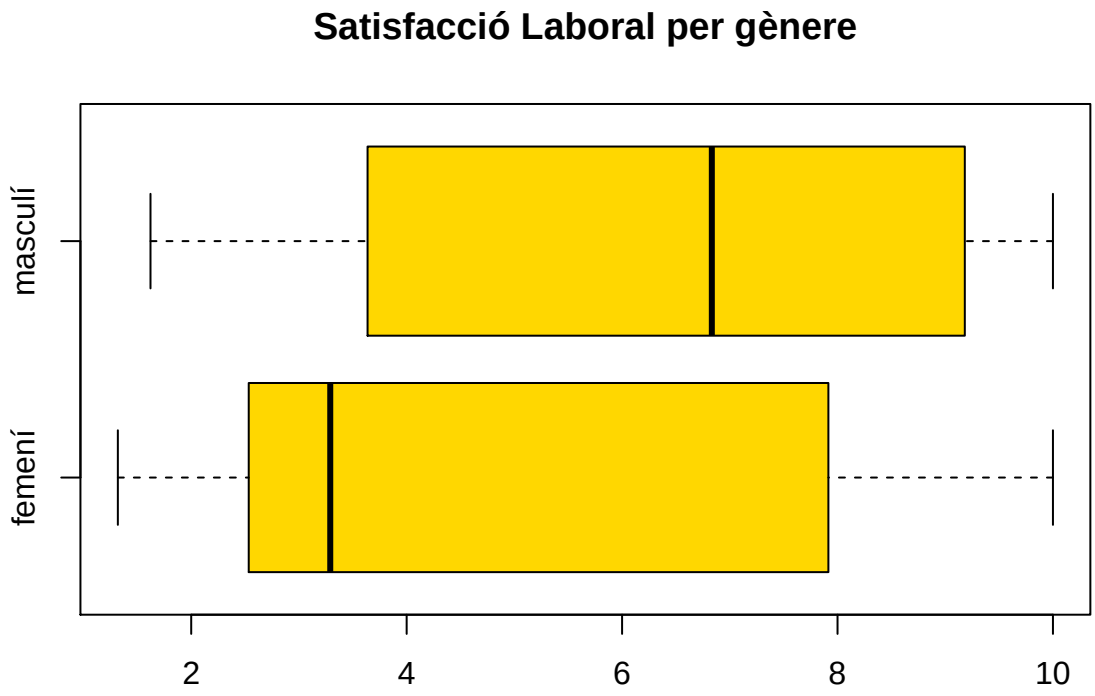
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.622   2.277   6.336   6.079  10.000  10.000
```

A més, la satisfacció laboral dels homes és molt més alta ($mediana_M = 6.83$) que les dones ($mediana_F = 3.29$).

La satisfacció laboral dels homes està més equilibrada, la mitjana ($\hat{x}_M = 6.44$) és molt similar a la mediana. Observo que hi ha la mateixa distancia entre Q1 i la mediana, que entre la mediana i Q3.

En canvi, la mitjana de la satisfacció laboral de les dones ($\hat{x}_F = 4.76$) és molt diferent a la mediana. Observo que la distancia entre Q1 i la mediana és molt diferent a la distancia entre la mediana i Q3.

```
#  Genero boxplot de la variable S separant les dades per gènere
par(mfrow=c(1,1))
idxF <- satlab[satlab$Sex=='F',]
idxM <- satlab[satlab$Sex=='M',]
boxplot(idxF$S,idxM$S, main="Satisfacció Laboral per gènere",
         names=c("femení","masculí"), horizontal=TRUE, col="gold")
```




```
# Executo summary i calculo el rang interquartílic de S per gènere femení
summary(idxF$S)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.319   2.583   3.291   4.755   7.834  10.000
```

```
IQR(idxF$S)
```

```
## [1] 5.251509
```

```r
# Executo summary i calculo el rang interquartílic de S per gènere masculí
summary(idxM$S)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.622   3.737   6.833   6.436   9.092  10.000
```

```r
IQR(idxM$S)
```

```
## [1] 5.355204
```

2 Estadística inferencial

2.1 Interval de confiança del nivell de satisfacció laboral

Calcular l'interval de confiança al 97% de la satisfacció laboral dels treballadors. A partir del valor obtingut, expliqueu com s'interpreta el resultat de l'interval de confiança.

Nota: Cal realitzar els càlculs manualment. No és vàlid aplicar funcions del tipus t.test o similars que donen els càlculs fets. Sí que podeu usar funcions del tipus qt, pt, qnorm, pnorm.

Tenint en compte que un interval de confiança d'un cert paràmetre amb un nivell de confiança del 97% és un interval calculat a partir d'una mostra de manera que el procediment de càlcul garanteix que el 97% de les mostres donen lloc a un interval que conté el valor real del paràmetre.

Suposem que la població és normal i desconeixem la desviació típica poblacional, per tant, caldrà estimar la desviació típica usant els valors mostrals i treballar amb la distribució de la mitjana mostral $\overline{x}$, ja que, amb un procediment semblant a la estandardització, es pot relacionar amb una altra variable que segueix una distribució de Student.

Els resultats son vàlids sempre que la mostra sigui major que trenta (tenim 38 observacions).

Per calcular l'interval de confiança del 97% faig el següent:

1. Fixo el nivell de confiança, s'escriu com $p = (1 - \alpha)\ \% = 97\ \%$

Un nivell de confiança del 97% ($p = 0.97$), implica que el nivell de significació és del 3% ($\alpha = 1 - p = 0.03$).

```
alpha<- 1-0.97
```

2. Calculo la desviació típica mostral: $s = \frac{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}}{n-1}$ Utilitzo la funció *sd* perquè ja divideix entre n-1.

```
desviacio_mostral <- sd(satlab$S)
desviacio_mostral
```

```
## [1] 3.051463
```

3. Calculo l'**error estàndard** de la mitjana: $s_{\overline{x}} = \frac{s}{\sqrt{n}}$

```
n <- length(satlab$S)
error_std <- desviacio_mostral / sqrt(n)
error_std
```

```
## [1] 0.4950126
```

4. Calculo el **valor crític**. És aquell punt $t_{\frac{\alpha}{2},n-1}$ tal que $P(t_{n-1} >= t_{\frac{\alpha}{2},n-1}) = \frac{\alpha}{2}$, on t_{n-1} és una variable Student amb n-1 graus de llibertat.

La funció *qt* retorna el quantil del valor donat en una distribució t-Student.

$z_{\alpha/2} = z_{0.03/2}$

```
valor_critic <- qt(1-alpha/2, df=(n-1))
valor_critic
```

```
## [1] 2.25702
```

5. Calculo el **marge d'error**, és igual al producte de l'error estandard i el valor crític: $t_{\frac{\alpha}{2},n-1} * \frac{s}{\sqrt{n}}$

```
marge_error <- valor_critic * error_std
marge_error
```

```
## [1] 1.117253
```

6. L'**interval de confiança** és el següent: (4.611032, 6.845539)

```
lim_inf <- mean(satlab$S) - marge_error
lim_inf
```

```
## [1] 4.611032
```

```
lim_sup <- mean(satlab$S) + marge_error
lim_sup
```

```
## [1] 6.845539
```

7. Comprovo que el resultat de la funció *t.test* és igual a l'interval de confiança calculat manualment.

```
t.test(satlab$S,conf.level = 0.97)
```

```
##
##  One Sample t-test
##
## data:  satlab$S
## t = 11.572, df = 37, p-value = 7.419e-14
## alternative hypothesis: true mean is not equal to 0
## 97 percent confidence interval:
##  4.611032 6.845539
## sample estimates:
## mean of x
##  5.728286
```

Per tant, l'**interval de confiança** calculat (4.611032, 6.845539) amb un nivell de confiança del 97% vol dir que el nivell de satisfacció laboral es troba dins d'aquest interval i que aquest interval ha estat calculat amb el 97% de les mostres.

2.2 Test de dues mostres: satisfacció laboral en funció del tipus de treball

Hi ha diferències significatives en la satisfacció laboral dels treballadors que ocupen un lloc de treball qualificat i els que estan en un lloc de treball poc qualificat? Calcular-ho per a un nivell de confiança del 90% i 95%.

Nota: Cal realitzar els càlculs manualment. No és vàlid aplicar funcions del tipus t.test o similars que donen els càlculs fets. Sí que podeu usar funcions del tipus qt, pt, qnorm, pnorm.

Seguiu els passos que es detallen a continuació:

2.2.1 Escriure la hipòtesi nul · la i alternativa

- La hipòtesi nul · la és: els valors de les mitjanes de totes dues mostres són iguals, és a dir, la satisfacció laboral dels treballadors que ocupen un lloc de treball qualificat (μ_Q) és igual a la satisfacció laboral dels treballadors que ocupen un lloc de treball poc qualificat (μ_{PQ}).

  $H_0 : \mu_Q = \mu_{PQ} => \mu_Q - \mu_{PQ} = 0$

- La hipòtesi alternativa és: els valors de les mitjanes de totes dues mostres són diferents, és a dir, la satisfacció laboral dels treballadors que ocupen un lloc de treball qualificat (μ_Q) és diferent a la satisfacció laboral dels treballadors que ocupen un lloc de treball poc qualificat (μ_{PQ}).

  $H_1 : \mu_Q <> \mu_{PQ} => \mu_Q - \mu_{PQ} <> 0$

2.2.2 Justifiqueu quin mètode aplicareu

Amb el contrat d'hipòtesis sobre diferències de mitjanes poblacionals, es consideren dues mostres d'observacions i es comparen les mitjanes contrastant hipòtesis sobre la diferència i construïnt intervals de confiança per a aquesta diferència.

Les variàncies poblacionals són desconegudes però iguals a un cert valor σ^2, és a dir, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ amb σ desconeguda.

Aquesta desviació típica comuna σ es pot estimar per mitjà de la fórmula:$s = \frac{\sqrt{\sum_{i=1}^{n}(x_{i1}-\overline{x_1})^2+\sum_{i=1}^{n}(x_{i2}-\overline{x_2})^2}}{n_1+n_2-2}$

O el que és el mateix: $s = \sqrt{\frac{((n_Q-1)*s_Q^2)+((n_{PQ}-1)*s_{PQ}^2)}{n_Q+n_{PQ}-2}}$

I l'estadístic de contrast correspon a una observació d'una distribució t de Student amb $n_1 + n_2 - 2$ graus de llibertat: $t = \frac{(\overline{x_1}-\overline{x_2})}{s*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

L'error estandard és igual a $s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Com la hipòtesi alternativa és bilateral, on la mitjana d'una població és superior o inferior a l'altra poblacio i es comparen en les dues direcciones ($H_1 : \mu_Q <> \mu_{PQ}$), i es correspon a una

distribució t de *student*, el p-valor corresponent a l'estadístic de contrast es igual a $P(t_{n_1+n_2-2} > |t|) = 2(1 - P(t_{n_1+n_2-2} <= t))$.

Segons quin sigui el p-valor, la regla de decisió és la següent:

* Acceptem H_0 si el *p-valor* és >= al nivell de significació α

* Rebutjarem H_0 si el *p-valor* és < al nivell de significació α

2.2.3 Realitzeu els càlculs de l'estadístic de contrast, valor crític i valor p, al 90% i 95% de nivell de confiança

1. Calculo la mitjana de la satisfacció laboral dels treballadors que ocupen un lloc de treball qualificat ($n_Q = 16, \mu_Q = 8.203051$) i la mitjana de la satisfacció laboral dels treballadors que ocupen un lloc de treball poc qualificat ($n_{PQ} = 22, \mu_{PQ} = 3.928457$):

```
# Calculo la mitjana PC dels fumadors i dels no fumadors
satlabQ <- satlab[satlab$Wtype=='Q',]$S
nsatlabQ <- length(satlabQ)
nsatlabQ
```

```
## [1] 16
```

```
satlabPQ <- satlab[satlab$Wtype=='PQ',]$S
nsatlabPQ <-length(satlabPQ)
nsatlabPQ
```

```
## [1] 22
```

```
meanQ <- mean(satlabQ)
meanQ
```

```
## [1] 8.203051
```

```
meanPQ <- mean(satlabPQ)
meanPQ
```

```
## [1] 3.928457
```

2. Calculo la desviació típica comuna: $s = \sqrt{\frac{((n_Q-1)*s_Q^2)+((n_{PQ}-1)*s_{PQ}^2)}{n_Q+n_{PQ}-2}} = 2.206461$

```
desviacio_comuna <- sqrt((((16-1)*sd(satlabQ)^2)
                          +((22-1)*sd(satlabPQ)^2))
                         /(16+22-2))
desviacio_comuna
```

```
## [1] 2.206461
```

3. Calculo l'**error estàndard** de la mitjana: $s * \sqrt{\frac{1}{n_Q} + \frac{1}{n_{PQ}}} = 0.7249643$

```
error_std <- desviacio_comuna * sqrt(((1/16)+ (1/22)))
error_std
```

```
## [1] 0.7249643
```

4. Calculo l'**estadístic de contrast**: $t = \frac{\overline{x_Q} - \overline{x_{PQ}}}{s * \sqrt{\frac{1}{n_Q} + \frac{1}{n_{PQ}}}} = -23.5391$

```
t <- (meanQ - meanPQ) / error_std
t
```

```
## [1] 5.896282
```

5. Calculo el **p-valor**: $2P(t_{n_Q+n_{PQ}-2} > |t|) = 2(1 - P(t_{36} <= t)) = 2 * (1 - 0.9999995) \equiv 0.$

Utilitzo la funció *pt* per calcular la probabilitat que li correspon a t amb 36 graus de llibertat.

```
# Calculo el p-valor
p_valor <- 2 * (1 - pt(t, df=(16+22-2)))
p_valor
```

```
## [1] 9.561786e-07
```

2.2.4 Interpreteu el resultat i doneu resposta a la pregunta plantejada

- Fixo el nivell de confiança del 90%, s'escriu com $p = (1 - \alpha)\ \% = 90\ \%$

Un nivell de confiança del 90% ($p = 0.90$), implica que el nivell de significació és del 10% ($\alpha = 1 - p = 0.10$).

```
alpha<- 1-0.90
```

Com p-valor equival a 0 i és menor que el nivell de significació $\alpha = 0.10$, rebutjo la hipòtesi nul · la H_0 en favor de la hipòtesi alternativa, és a dir, es pot afirmar que els valors de les mitjanes de totes dues mostres són diferents, és a dir, la satisfacció laboral dels treballadors que ocupen un lloc de treball qualificat (μ_Q) és diferent a la satisfacció laboral dels treballadors que ocupen un lloc de treball poc qualificat (μ_{PQ}) amb un nivell de confiança del 90%.

- Fixo el nivell de confiança del 95%, s'escriu com $p = (1 - \alpha)\ \% = 95\ \%$

Un nivell de confiança del 95% ($p = 0.95$), implica que el nivell de significació és del 5% ($\alpha = 1 - p = 0.05$).

```
alpha<- 1-0.95
```

Com p-valor equival a 0 i també és menor que el nivell de significació $\alpha = 0.05$, rebutjo la hipòtesi nul · la H_0 en favor de la hipòtesi alternativa, és a dir, es pot afirmar que els valors de les mitjanes

de totes dues mostres són diferents, és a dir, la satisfacció laboral dels treballadors que ocupen un lloc de treball qualificat (μ_Q) és diferent a la satisfacció laboral dels treballadors que ocupen un lloc de treball poc qualificat (μ_{PQ}) amb un nivell de confiança del 95%.

2.3 Test de dues mostres: satisfacció laboral en funció del sexe

Es pot afirmar que les dones tenen una satisfacció laboral inferior a la dels homes? Calcular-ho per a un nivell de confiança del 90% i 95%.

Nota: Cal realitzar els càlculs manualment. No és vàlid aplicar funcions del tipus t.test o similars que donen els càlculs fets. Sí que podeu usar funcions del tipus qt, pt, qnorm, pnorm.

2.3.1 Escriure la hipòtesi nul · la i alternativa

- La hipòtesi nul · la és: els valors de les mitjanes de totes dues mostres són iguals, és a dir, la satisfacció laboral de les dones (μ_D) és igual a la satisfacció laboral dels homes (μ_H).

  $H_0 : \mu_D = \mu_H => \mu_D - \mu_H = 0$

- La hipòtesi alternativa és: la satisfacció laboral de les dones (μ_D) és menor que la satisfacció laboral dels homes (μ_H).

  $H_1 : \mu_D < \mu_H => \mu_D - \mu_H < 0$

2.3.2 Justifiqueu quin mètode aplicareu

Amb el contrat d'hipòtesis sobre diferències de mitjanes poblacionals, es consideren dues mostres d'observacions i es comparen les mitjanes contrastant hipòtesis sobre la diferència i construïnt intervals de confiança per a aquesta diferència.

Les variàncies poblacionals són desconegudes però iguals a un cert valor σ^2, és a dir, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ amb σ desconeguda.

Aquesta desviació típica comuna σ es pot estimar per mitjà de la fórmula:$s = \frac{\sqrt{\sum_{i=1}^{n}(x_{i1}-\overline{x_1})^2+\sum_{i=1}^{n}(x_{i2}-\overline{x_2})^2}}{n_1+n_2-2}$

O el que és el mateix: $s = \sqrt{\frac{((n_Q-1)*s_Q^2)+((n_{PQ}-1)*s_{PQ}^2)}{n_Q+n_{PQ}-2}}$

I l'estadístic de contrast correspon a una observació d'una distribució t de Student amb $n_1 + n_2 - 2$ graus de llibertat: $t = \frac{(\overline{x_1}-\overline{x_2})}{s*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

L'error estandard és igual a $s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Com la hipòtesi alternativa és unilateral, només es compara en una direcció ($H_1 : \mu_D < \mu_H$), el p-valor corresponent a l'estadístic de contrast es igual a $P(t_{n_1+n_2-2} < t)$.

Segons quin sigui el p-valor, la regla de decisió és la següent:

* Acceptem H_0 si el *p-valor* és >= al nivell de significació α

* Rebutjarem H_0 si el *p-valor* és < al nivell de significació α

2.3.3 Realitzeu els càlculs de l'estadístic de contrast, valor crític i valor p, al 90% i 95% de nivell de confiança

1. Calculo la mitjana de la satisfacció laboral de les dones ($n_D = 16, \mu_D = 4.75508$) i la mitjana de la satisfacció laboral dels homes ($n_H = 22, \mu_H = 6.436071$):

```
# Calculo la mitjana PC dels fumadors i dels no fumadors
satlabD <- satlab[satlab$Sex=='F',]$S
nsatlabD <- length(satlabD)
nsatlabD
```

```
## [1] 16
```

```
satlabH <- satlab[satlab$Sex=='M',]$S
nsatlabH <-length(satlabH)
nsatlabH
```

```
## [1] 22
```

```
meanD <- mean(satlabD)
meanD
```

```
## [1] 4.75508
```

```
meanH <- mean(satlabH)
meanH
```

```
## [1] 6.436071
```

2. Calculo la desviació típica comuna: $s = \sqrt{\frac{((n_D-1)*s_D^2)+((n_H-1)*s_H^2)}{n_D+n_H-2}} = 2.973716$

```
desviacio_comuna <- sqrt((((16-1)*sd(satlabD)^2)
                          +((22-1)*sd(satlabH)^2))
                         /(16+22-2))
desviacio_comuna
```

```
## [1] 2.973716
```

3. Calculo l'**error estàndard** de la mitjana: $s * \sqrt{\frac{1}{n_D} + \frac{1}{n_H}} = 0.9770571$

```
error_std <- desviacio_comuna * sqrt(((1/16)+ (1/22)))
error_std
```

```
## [1] 0.9770571
```

4. Calculo l'**estadístic de contrast**: $t = \frac{\overline{x_D} - \overline{x_H}}{s * \sqrt{\frac{1}{n_D} + \frac{1}{n_H}}} = -1.720463$

```
t <- (meanD - meanH) / error_std
t
```

```
## [1] -1.720463
```

5. Calculo el **p-valor**: $P(t_{n_D+n_H-2} < t) = 0.04696726$.

Utilitzo la funció *pt* per calcular la probabilitat que li correspon, amb *lower.tail=TRUE* per a determinar el càlcul de la probabilitat de la cua de l'esquerra.

```
# Calculo el p-valor
p_valor <- pt(t, df=(16+22-2) , lower.tail=TRUE)
p_valor
```

```
## [1] 0.04696726
```

2.3.4 Interpreteu el resultat i doneu resposta a la pregunta plantejada

- Fixo el nivell de confiança del 90%, s'escriu com $p = (1 - \alpha)\ \% = 90\ \%$

Un nivell de confiança del 90% ($p = 0.90$), implica que el nivell de significació és del 10% ($\alpha = 1 - p = 0.10$).

```
alpha<- 1-0.90
```

Com p-valor=0.0470 i és menor que el nivell de significació $\alpha = 0.10$, rebutjo la hipòtesi nul · la H_0 en favor de la hipòtesi alternativa, és a dir, es pot afirmar que la satisfacció laboral de les dones (μ_D) és menor que la satisfacció laboral dels homes (μ_H) amb un nivell de confiança del 90%.

- Fixo el nivell de confiança del 95%, s'escriu com $p = (1 - \alpha)\ \% = 95\ \%$

Un nivell de confiança del 95% ($p = 0.95$), implica que el nivell de significació és del 5% ($\alpha = 1 - p = 0.05$).

```
alpha<- 1-0.95
```

Com p-valor=0.0470 i també és menor que el nivell de significació $\alpha = 0.05$, rebutjo la hipòtesi nul · la H_0 en favor de la hipòtesi alternativa, és a dir, es pot afirmar que la satisfacció laboral de les dones (μ_D) és menor que la satisfacció laboral dels homes (μ_H) amb un nivell de confiança del 95%.

3 Regressió

3.1 Model de regressió

Apliqueu un model de regressió lineal múltiple que usi com a variables explicatives el nombre d'hores, el sexe, el nivell d'educació i el tipus de treball i com a variable dependent la satisfacció laboral. Especifiqueu en el nivell base (en el relevel): per la variable sexe, la categoria 'F', per la variable educació, la categoria '1', i per la variable tipus de treball, la categoria 'PQ'.

Amb el model de regressió lineal multiple busco explicar la variable dependent o explicada S amb les variables independents o explicatives H, Sex, $Etype$ i $Wtype$, mitjançant la expressió següent:

$PC = \beta_0 + \beta_1 H + \beta_2 Sex + \beta_3 Etype + \beta_4 Wtype + e$

Amb la funció *lm* obtinc els components principals de la regressió:

- *Residuals*: mostra el mínim, màxim i quartils dels residus de la regressió, els quals proporcionen informació sobre la seva distribució.
- *Coefficients*: informació de l'estimació dels paràmetres (o coeficients) estimats.
- *Estimate*: estimació de cada paràmetre (intercept significa constant).
- *Std.Error*: desviació (o error) estàndard de cada paràmetre estimat.
- *t value*: estadístic t de cada paràmetre estimat, otingut dividint l'estimació del paràmetre entre la seva desviació estàndard. Aquest estadístic és el que s'utilitza per a fer el contrast de significació individual dels paràmetres estimats.
- *Pr(>|t|)*: *p-valor* del contrast de significació individual de cada paràmetre estimat, el qual indica la seva significació estadística.
- *Signif. codes*: mostra, amb asteriscos i punts, per a quins nivells de significació els coeficients estimats són significatius o no.
- *Residual standard error*: desviació (o error) estàndard dels residus.
- *Multiple R-squared*: coeficient de determinació.
- *Ajusted R-squared*: coeficient de determinació ajustat.
- *F - statistic*: estadístic F per al contrast de la significació global o conjunta dels paràmetres estimats del model.
- *p-value*: p-valor associat al constrast anterior. En aquest model de regressió, veig que el conjunt de paràmetres estimats és significatiu amb un nivell de significació del 1% (p-valor<0.01).

```r
# aplico el model lineal amb la funció lm
model <- lm(data = satlab, formula = S ~ H+Sex+Etype+Wtype)
summary(model)
```

```
##
## Call:
## lm(formula = S ~ H + Sex + Etype + Wtype, data = satlab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1736 -1.5668  0.1855  1.7080  3.2844
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.39778    2.49245   1.764   0.0875 .
## H           -0.05711    0.07770  -0.735   0.4679
## SexM         0.63932    0.74585   0.857   0.3979
## Etype2       2.50859    0.95900   2.616   0.0136 *
## Etype3       1.34802    1.03121   1.307   0.2008
## Etype4       1.17412    0.99334   1.182   0.2462
## WtypeQ       3.78734    0.79076   4.789 3.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.119 on 31 degrees of freedom
## Multiple R-squared:  0.596,  Adjusted R-squared:  0.5178
## F-statistic: 7.621 on 6 and 31 DF,  p-value: 4.416e-05
```

Observo que la funció *lm* afegeix automàticament una variable diferent per cada una de les categories de les variables qualitatives, menys una.

Però com vull determinar quina és la categoria base que cal considerar en la reordenació, utilitzo la funció *relevel*.

Afegeixo la variable **SexR** tenint en compte que la categoria de referència de la variable **Sex** és 'F'.

```r
# afegeixo la variable SexRM amb la funció relevel
satlab$SexR <- relevel( satlab$Sex, ref='F')
```

També afegeixo la variable **EtypeR** tenint en compte que la categoria de referència de la variable **Etype** és '1' (sense estudis). Recordo que el significat dels possibles valors de Etype: '2' (estudis primaris), '3' (estudis secundaris) i '4' (universitaris).

```r
# afegeixo la variable SportR amb la funció relevel
satlab$EtypeR <- relevel( satlab$Etype, ref='1')
```

I també afegeixo la variable **WtypeR** tenint en compte que la categoria de referència de la variable **Wtype** és 'PQ' (poc qualificat). L'altre valor de la variable WtypeR és 'Q' (qualificat).

```r
# afegeixo la variable SportR amb la funció relevel
satlab$WtypeR <- relevel( satlab$Wtype, ref='PQ')
```

I torno a generar el model de regressió lineal multiple però amb les noves variables afegides **SexR**, **EtypeR** i **WtypeR**.

```r
# aplico el model lineal amb la funció lm
model <- lm(data = satlab, formula = S ~ H+SexR+EtypeR+WtypeR)
summary(model)
```

```
##
## Call:
## lm(formula = S ~ H + SexR + EtypeR + WtypeR, data = satlab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1736 -1.5668  0.1855  1.7080  3.2844
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.39778    2.49245   1.764   0.0875 .
## H           -0.05711    0.07770  -0.735   0.4679
## SexRM        0.63932    0.74585   0.857   0.3979
## EtypeR2      2.50859    0.95900   2.616   0.0136 *
## EtypeR3      1.34802    1.03121   1.307   0.2008
## EtypeR4      1.17412    0.99334   1.182   0.2462
## WtypeRQ      3.78734    0.79076   4.789 3.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.119 on 31 degrees of freedom
## Multiple R-squared:  0.596,  Adjusted R-squared:  0.5178
## F-statistic: 7.621 on 6 and 31 DF,  p-value: 4.416e-05
```

Per tant, el model de regressió lineal múltiple té la forma següent:

$$S = \beta_0 - \beta_1 H + \beta_2 SexRM + \beta_3 EtypeR2 + \beta_4 EtypeR3 + \beta_5 EtypeR4 + \beta_6 WtypeRQ$$

$$S = 4.39778 - 0.05711H + 0.63932SexRM + 2.50859EtypeR2 + 1.34802EtypeR3 + 1.17412EtypeR4 + 3.78734WtypeRQ$$

3.2 Interpretació

Interpreteu el model de regressió resultant, indicant quins regressors són significatius. Expliqueu si hi ha diferències significatives degudes al sexe i si hi són, en quin sentit. Feu el mateix per la resta de variables.

A la taula Coefficients puc veure els valors estimats dels paràmetres. També es mostra l'error estàndard, el valor de l'estadístic t-Student i un p-valor que serveix per a contrastar el nivell de significació del paràmetre.

Amb el contrast pretenc determinar si els efectes de la constant i de les variables independents són realment importants per a explicar la variable dependent o bé els efectes es poden considerar nuls.

Per a cada i, considero el contrast d'hipòtesi següent:

$H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$

Com el p-valor de la variable *EtypeR2* (0.0136) i de la variable *WtypeRQ* (3.92e-05) és menor que 0.05, ens indica que cal rebutjar la hipòtesi nul.la, per tant, β_3 i β_6 sempre tindran valors significativament diferents a 0.

De manera que els regressors de *WtypeRQ* (tipus de treball qualificat) i, en menor mesura, *EtypeR2* (estudis primaris) són més rellevants en el model de regressió i influiran en el nivell de satisfacció de l'individu.

En canvi, com el p-valor de la resta de variables són majors que 0.05, ens indica que no es rebutja la hipòtesi nul.la, per tant, podem suposar que els valors β_0, β_1, β_2, β_4 i β_5 són iguals a 0 i no es consideren significatives.

3.3 Predicció

Apliqueu el model de regressió per a predir la satisfacció laboral d'un home, que treballa 40h setmanals, de nivell d'estudis universitaris i amb un treball qualificat.

Per realitzar la predicció, inicialitzo les variables amb els valors següents:

SexRM=1, H=50, EtypeR2=0, EtypeR3=0, EtypeR4=1, WtypeRQ=1

```
# inicialitzo les variables
H <- 40
SexRM <- 1
EtypeR2 <- 0
EtypeR3 <- 0
EtypeR4 <- 1
WtypeRQ <- 1
```

Per calcular manualment el valor estimat de **S**, substitueixo els valors de les variables en el model de regressió lineal multiple:

$S = 4.39778 - 0.05711H + 0.63932SexRM + 2.50859EtypeR2 + 1.34802EtypeR3 + 1.17412EtypeR4 + 3.78734WtypeRQ = 7.71416$

```r
# calculo manualment el valor de S amb el model de regressió
S_manual <- 4.39778 - (0.05711 * H) + (0.63932 * SexRM) + (2.50859 * EtypeR2) + (1.34802 * Ety
S_manual
```

```
## [1] 7.71416
```

També calculo el valor de la predicció amb la funció *predict*. Amb el paràmetre $se.fit = TRUE$ retorna més components com:

- fit: valor predit
- se.fit: error estandard de la predicció de les mitjanes
- residual.scale: desviació estandar dels residus
- df: graus de llibertat dels residus

```r
# inicialitzo les variables
H <- 40
SexR <- 'M'
EtypeR <- '4'
WtypeR <- 'Q'
```

```r
# calculo la predicció del valor de S
S_predict <- predict (model, newdata = data.frame(H, SexR, EtypeR, WtypeR),
                      se.fit = TRUE)
S_predict
```

```
## $fit
##        1
## 7.714188
##
## $se.fit
## [1] 1.131771
##
## $df
## [1] 31
##
## $residual.scale
## [1] 2.11904
```

La funció *predict* retorna el mateix valor 7.143096.

Compareu el resultat amb el d'un home, que treballa 40 h/s, de nivell d'estudis universitaris, i treball poc qualificat.

A continuació, inicialitzo les variables per al nou cas plantejat:

```
# inicialitzo les variables
H <- 40
SexR <- 'M'
EtypeR <- '4'
WtypeR <- 'PQ'
```

Calculo la predicció:

```
# calculo la predicció del valor de S
S_predict <- predict (model, newdata = data.frame(H, SexR, EtypeR, WtypeR),
                      se.fit = TRUE)
S_predict
```

```
## $fit
##        1
## 3.926849
##
## $se.fit
## [1] 1.259294
##
## $df
## [1] 31
##
## $residual.scale
## [1] 2.11904
```

Les prediccions realitzades indiquen que la satisfacció laboral és major en l'individu amb tipus de treball qualificat (7.71) que en l'individu amb tipus de treball poc qualificat (3.93).

Com el regressor *WtypeRQ* (tipus de treball qualificat) és una variable significativa en el model, influeix en el nivell de satisfacció predit per a l'individu.

3.4 Interpretació de la predicció

Interpreteu els resultats obtinguts en l'apartat anterior. Concretament, comenteu els aspectes següents:

a) Considereu que el model serà precís, tenint en compte els valors de R^2 i p-value obtinguts?

El coeficient de determinació R^2 indica el grau d'ajust de la recta de regressió als valors de la mostra, i es defineix com la proporció de variància explicada per la recta de regressió, és a dir: $R^2 = \frac{VariànciaExplicadaPelModel}{VariànciaTotalMostra}$

La bondat d'ajust no sembla massa bona, $R^2 = 0.596$, ja que no s'apropa massa a 1, i indica que el model de regressió lineal només explica el 59.6% de la variància de les observacions.

El $R^2 ajustat = 0.5178$ no és molt proper a R^2, és a dir, el model està penalitzat pel nombre de variables i que algunes d'elles no siguin significatives.

Tot i així, el test F mostra un p-value=4.416e-05, per tant, el model en conjunt és suficientment significatiu i es corrobora amb els p-values que he interpretat en l'apartat 3.2.

Tot i que en l'enunciat no es demana, si el coeficient d'un regresor no és significatiu, es podria treure el regressor del model de regressió i es podria tornar a calcular de nou el model de regressió sense usar aquestes variables.

Habitualment el que es cerca és un model de regressió que relacioni correctament les variables explicatives amb la variable dependent (és a dir, que expliqui la variabilitat de la variable dependent en funció de les variables regressores) i a la vegada, que sigui senzill (que contingui el mínim nombre de regressors). Per tant, si es veu que hi ha regresssors que no influeixen significativament en el model, es poden eliminar del model final.

b) Com es pot interpretar la diferència de satisfacció laboral entre els dos individus, a partir dels coeficients del model de regressió?

La descripció dels dos individus és igual excepte que un individu té un tipus de treball 'qualificat' i l'altre 'poc qualificat'.

El coeficient estimat β_6 de la variable *WtypeRQ*, que identifica el tipus de treball, és igual 3.78734, per tant, una observació amb tipus de treball 'qualificat' (WtypeRQ=1) tindrá 3.78734 unitats més de nivell de satisfacció que una observació amb tipus de treball 'poc qualificat' (WtypeRQ=0).

3.5 Intervals de predicció

Calculeu els intervals de predicció dels dos individus al 95%. Per a fer-ho, podeu afegir a la funció predict el paràmetre interval="prediction". Per especificar el nivell de confiança podeu usar el paràmetre level (per defecte és 0.95). Interpreteu els resultats.

Els *intervals de predicció* acompanyen al valor predit per indicar l'espai en el que es distribueixen les dades amb una probabilitat donada.

Els *intervals de confiança* representen l'espai on podem trobar un paràmetre estadístic concret amb una probabilitat concreta, com pot ser la mitjana, desviació,... En canvi, l'*interval de predicció* no es refereix a cap paràmetre estadístic concret, si no a la distribució dels valors en el seu espai, representant la probabilitat de trobar una observació dins d'aquest interval.

Ara que és estrany que els intervals de predicció calculats retornen valors inferiors a 0 i majors que 10!!!!

1. Interval de predicció al 95% de l'individu amb tipus treball 'Qualificat': (2.81, 12.61)

```
# inicialitzo les variables
H <- 40
SexR <- 'M'
EtypeR <- '4'
WtypeR <- 'Q'
```

```
# calculo l'interval de predicció
S_predict <- predict (model, newdata = data.frame(H, SexR, EtypeR, WtypeR)
                      , interval = "prediction"
                      , level = 0.95
                      )
S_predict
```

```
##        fit      lwr      upr
## 1 7.714188 2.814585 12.61379
```

2. Interval de predicció al 95% de l'individu amb tipus treball 'Poc qualificat': (-1.10, 8.95)

```
# inicialitzo les variables
H <- 40
SexR <- 'M'
EtypeR <- '4'
WtypeR <- 'PQ'
```

```
# calculo l'interval de predicció
S_predict <- predict (model, newdata = data.frame(H, SexR, EtypeR, WtypeR)
                      , interval = "prediction"
                      , level = 0.95
```

```
                      )
S_predict
```

```
##        fit       lwr      upr
## 1 3.926849 -1.100521 8.954218
```

3.6 Ajust del model

Analitzar l'adequació del model a partir de l'anàlisi gràfic de residus. Per això, podeu usar la instrucció plot passant com a paràmetre el model de regressió lineal. Per a saber com interpretar els gràfics de residus, consulteu l'enllaç següent: http://data.library.virginia.edu/diagnostic-plots/

Si a la funció *plot* li passes com a paràmetre el model de regressió,retorna 4 gràfiques:

- Amb la gràfica *'Residuals vs Fitted'* es visualitza si els residus tenen patrons no lineals, és a dir, podria haver-hi una relació no lineal entre les variables predictores i la variable resultat i el patró d'aquesta relació es visualitza en aquesta gràfica si el model no es capaç de capturar la relació no lineal. Si els residus estan distribuits al voltant d'una línia horitzonal sense patrons observables és un bon indici que no existeix relació no lineal.

En aquest cas no s'observa cap patró perquè els residus estan distribuis al voltant d'una línia hortizontal sense patrons observables.

- Amb la gràfica *'Normal Q-Q'* es visualitza si els residus es distribueixen normalment.

Observo que els residus es distribueixen normalment perquè estan bastant alineats en la línia recta discontinua. Tot i que també observo que en els extrems hi ha alguns punts allunyats de la línia.

- Amb la gràfica *'Scale-Location'* es visualitza si els residus es reparteixen equitativament al llarg dels rangs predictors. Així es pot verificar l'assumpció de variancies iguals (homocedasticitat).

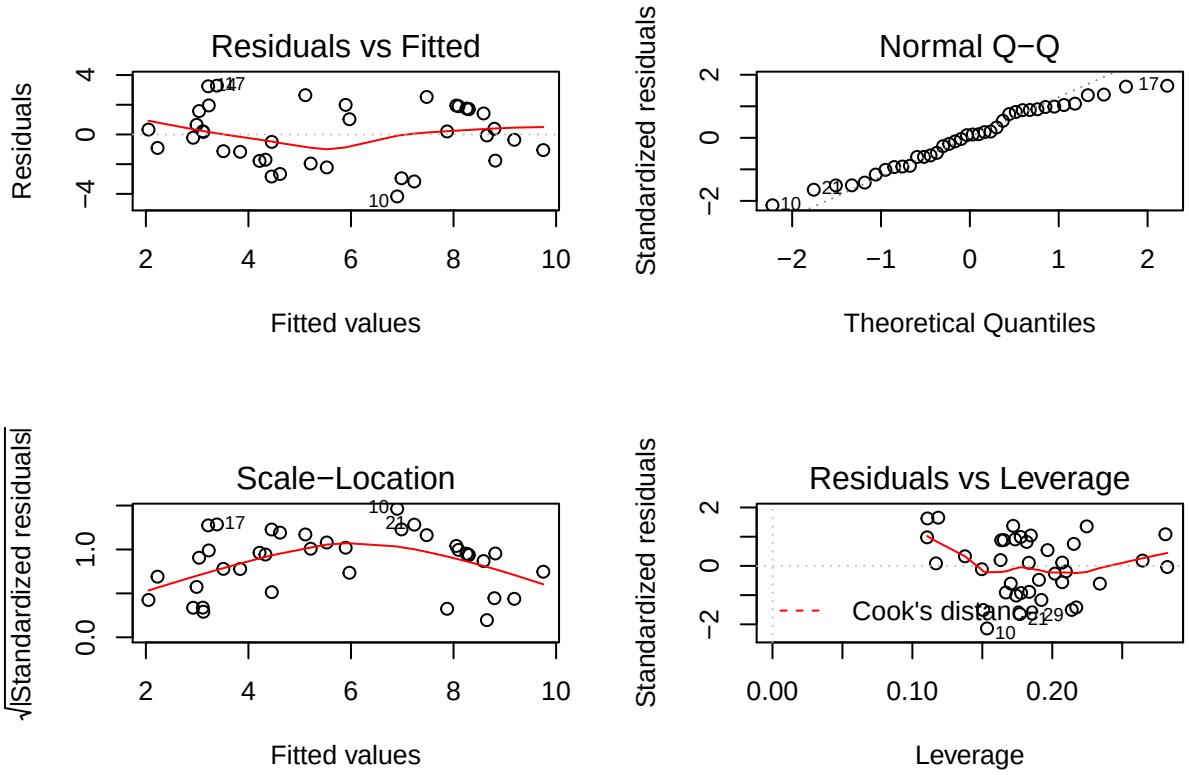
Observo que la línia no és massa horitzontal, ja que presenta una lleugera corba degut a que els residus s'extenen quan passen per X=6.

- Amb la gràfica *'Residuals vs Leverage'* es visualitza si hi ha valors atípics que són influents en l'anàlisi de la regressió lineal. Els valors extrems poden no influir en la recta de regressió i això significa que el resultat no seria molt diferent si s'inclouen o s'exclouen de l'anàlisi. Per altra banda, alguns casos podrien ser molt influents i poden alterar el resultat de la regressió si s'exclouen de l'anàlisi. En el gràfic, cal observar la part superior dreta i la part inferior esquerra, ja que aquests punts són els que poden influir en una línia de regressió. Si hi ha observacions fora de la distancia de Cook, és a dir tenen valors alts de distancia de Cook, els valors atípics són influents en els resultats de la regressió.

En aquest cas no observo cap valor fora de la distancia de Cook.

També observo que els 4 gràfics destaquen gairebé els mateixos casos: 10,14 i 17 en Fited, 10,17 i 21 en Q-Q, 10,17 i21 en Scale i 10,21 i29 en Leverage. De moment, considero que no és necessari eliminar aquests casos de l'anàlisi.

```
# visualitzo el model amb la funció plot
par(mfrow=c(2,2))
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

4 Anàlisi de variància unifactorial

A continuació, ens preguntem si el nivell de satisfacció laboral està influït pel nivell d'estudis. Donat que aquesta variable té quatre nivells, s'aplicarà anàlisi de variància.

L'anàlisi de la variància (ANOVA) d'un conjunt de mostres consisteix a contrastar la hipòtesi nul · la 'totes les mitjanes poblacionals d'on provenen les mostres són iguals', contra la hipòtesi alternativa 'no totes les mitjanes són iguals' amb un nivell de significació α prefixat.

Per a poder fer una anàlisi de la variància (ANOVA), cal tenir les hipòtesis següents sobre les dades:

- les k mostres han de ser aleatòries i independents entre si
- les poblacions han de ser normals
- les variàncies de les k poblacions han de ser idèntiques: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$

Sota aquestes hipòtesis i quan es compleix $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$, és a dir, si les mitjanes poblacionals són totes iguals, les sumes dels quadrats, entre mostres (SQE) i dintre les mostres (SQD), es distribueixen segons distribucions X^2 amb (k-1) i (n-k) graus de llibertat, respectivament.

Com les mostres són independents, una conseqüència important és que el quocient entre aquests estadístics:

$f = \frac{SQE/(k-1)}{SQD/n-k}$

es distribueix segons una distribució f de Snedecor amb (k-1) graus de llibertat al numerador i (n-k) al denominador, on k és el nombre de mostres i n és el nombre d'observacions.

4.1 Hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i l'alternativa per a l'ANOVA.

- La hipòtesi nul · la és: els valors de les mitjanes de totes les mostres són iguals, és a dir, la satisfacció laboral dels treballadors sense estudis (μ_{sense}) és igual a la satisfacció laboral dels treballadors amb estudis primaris (μ_{pri}) i és igual a la satisfacció laboral dels treballadors amb estudis secundaris (μ_{sec}) i també és igual a la satisfacció laboral dels treballadors amb estudis universitàris (μ_{uni}).

  $$H_0 : \mu_{sense} = \mu_{pri} = \mu_{sec} = \mu_{uni} = \mu$$

- La hipòtesi alternativa és: no tots els valors de les mitjanes de les mostres són iguals, és a dir, pot haver-hi una o més mitjanes de satisfacció laboral en funció del nivell d'estudis que sigui diferent a la resta de mitjanes.

  $$H_1 : \mu_Q <> \mu_{PQ} => \mu_Q - \mu_{PQ} <> 0$$

4.2 Model

Calculeu l'anàlisi de variància, usant la funció aov. Interpretreu el resultat de l'anàlisi, tenint en compte els valors Sum Sq, Mean SQ, F i Pr(>F).

1. Creo 4 vectors amb els valors de la satisfacció laboral en funció del nivell d'estudis:

'1'- 'sense estudis': 13 registres

'2'- 'estudis primaris': 8 registres

'3'- 'estudis secundaris': 9 registres

'4'- 'estudis unviersitaris': 8 registres

```
# calculo l'anàlisi de variància amb la funció aov
SE1 <- satlab[satlab$Etype=='1',]$S
SE2 <- satlab[satlab$Etype=='2',]$S
SE3 <- satlab[satlab$Etype=='3',]$S
SE4 <- satlab[satlab$Etype=='4',]$S
```

2. Calculo el nombre d'observacions de cada vector

```
# calculo el nombre d'observacions de cada vector
nSE1 <- length(SE1)
nSE1
```

```
## [1] 13
```

```
nSE2 <- length(SE2)
nSE2
```

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

```
nSE3 <- length(SE3)
nSE3
```

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

```
nSE4 <- length(SE4)
nSE4
```

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

3. Creo un únic vector amb totes les observacions

```
# calculo el nombre d'observacions de cada vector
SEstudis <- c (SE1, SE2, SE3, SE4)
```

4. Genero un vector per identificar el casos que corresponen a cada nivell d'estudis

```
# La variable factors identifica els casos que corresponen a cada nivell d'estudis
factors <- factor(c(rep("1",nSE1),rep("2",nSE2),rep("3",nSE3),rep("4",nSE4)))
factors
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4 4
## [36] 4 4 4
## Levels: 1 2 3 4
```

5. Calculo l'anàlisi de variància amb la funció aov

```
# calculo l'anàlisi de variància amb la funció aov
anova <- aov( lm(SEstudis ~ factors))
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## factors      3  72.17   24.05   3.003 0.0439 *
## Residuals   34 272.36    8.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El summary de l'ANOVA retorna:

- *Df* són els graus de llibertat del factor (3) i els graus de llibertat residuals (34). Representa la quantitat d'informació disponible en les dades.
- *Sum Sq* és la suma de quadrats dels grups (72.17) i la suma de quadrats dels errors (272.36)
- *Mean Sq* és la mitjana corresponent de la suma de quadrats dels grups (24.05) i dels errors (8.01)
- *F* és el valor del estadístic F (3.003)
- *p-value* és el nivell de significació de l'estadístic F (0.0439) és inferior a 0.05, per tant, rebutjo la hipòtesi nul.la i puc afirmar que hi ha diferències significatives entre els valors de les mitjanes de les mostres, és a dir, com a mínim una de les mitjanes de satisfacció laboral en funció del nivell d'estudis és diferent a la resta de mitjanes.

4.3 Càlculs

Per tal d'aprofundir en la comprensió del model ANOVA, calculeu manualment la suma de quadrats intra i la suma de quadrats entre grups. Els resultats han de coincidir amb el resultat del model ANOVA. Podeu fixar-vos en les fórmules de López-Roldán i Fachelli (2015), pàgines 29-33.

La variabilitat total de les dades està desglosada en dos components per poder comparar-los entre ells: la variabilitat dins de cada mostra i entre les mostres. Així es pot demostrar la existencia

d'associació o de determinació de la variable independent sobre la variable depenent.

De manera que, com més gran sigui la variabilitat entre els grups, més importants són les diferencies de les mitjanes de cada grup en relació a la mitjana total i, per tant, més homogenis són els grups, és a dir, menor serà la variabilitat interna i les diferencies en relació a la mitjana dins de cada grup no són importants, sent indicatiu de que les mitjanes són diferents.

El cas contrari, com menor sigui la variabilitat entre els grups, menys importancia tenen l'existencia dels grups, tendiran a ser grups heterogenis, amb una variabilitat interna alta, amb mitjanes que no difereixen entre els diferents grups i tendeixen a ser iguals.

En aquest cas, la variabilitat entre les mostres és bastant superior a la variabilitat dins les mostres, per tant, és un indicatiu que les mitjanes són diferents.

1. Calculo la mitjana de la mostra global.

$\overline{x} = \frac{\overline{x_1}n_1+\overline{x_2}n_2+\overline{x_3}n_3+\overline{x_4}n_4}{n_1+n_2+n_3+n_4} = 5.728286$

```
# calculo la mitjana de cada mostra
meanS <- mean(satlab$S)
meanS
```

```
## [1] 5.728286
```

2. Calculo la mitjana de cada mostra.

```
# calculo la mitjana de cada mostra
meanSE1 <- mean(SE1)
meanSE1
```

```
## [1] 3.877328
```

```
meanSE2 <- mean(SE2)
meanSE2
```

```
## [1] 6.912927
```

```
meanSE3 <- mean(SE3)
meanSE3
```

```
## [1] 7.037076
```

```
meanSE4 <- mean(SE4)
meanSE4
```

```
## [1] 6.079061
```

3. Calculo la variància de cada mostra.

```
# calculo la mitjana de cada mostra
sdSE1 <- sd(SE1)
sdSE1
```

```
## [1] 1.553916
```

```
sdSE2 <- sd(SE2)
sdSE2
```

```
## [1] 2.319241
```

```
sdSE3 <- sd(SE3)
sdSE3
```

```
## [1] 3.203398
```

```
sdSE4 <- sd(SE4)
sdSE4
```

```
## [1] 4.202624
```

4. Calculo la suma dels quadrats entre mostres: $SQE = \sum_i n_i(\overline{x_i} - \overline{x})^2 = 72.17$

El resultat és igual al que s'obté amb el paràmetre *Sum Sq* de la funció *aov*.

```
# calculo la suma dels quadrats entre mostres
SQE <- (nSE1 * (meanSE1-meanS)^2) + (nSE2 * (meanSE2-meanS)^2) + (nSE3 * (meanSE3-meanS)^2) +
SQE
```

```
## [1] 72.16631
```

Calculo la mitjana de quadrats entre mostres: $ SQE / (k-1) = 72.17 / (4-1) = 24.06$

```
# inicialitzo el nombre de mostres
k<-4
# calculo la mitjana de quadrats entre mostres
mSQE <- SQE / (k-1)
mSQE
```

```
## [1] 24.05544
```

5. Calculo la suma dels quadrats dins les mostres (residus): $SQD = \sum_j \sum_i n_j(\overline{x_{ij}} - \overline{x_j})^2 = \sum_{j=1}^{K}(n_j - 1)s_j^2 = 272.36$

```
# calculo la suma dels quadrats dins les mostres
SQD <- ((nSE1-1) * sdSE1^2) + ((nSE2-1) * sdSE2^2) + ((nSE3-1) * sdSE3^2) + ((nSE4-1) * sdSE4^
SQD
```

```
## [1] 272.3564
```

Calculo la mitjana de quadrats dins les mostres: $ SQD / (n-k) = 272.36 / (38-4) = 8.01$

```r
# calculo la mitjana de quadrats dins mostres
mSQD <- SQD / (n-k)
mSQD
```

```
## [1] 8.010482
```

6. Calculo la suma de quadrats total: SQT = SQD + SQE = 72.17 + 272.35 = 344.52 amb 37 (n-1) graus de lliberat

```r
# calculo la suma dels quadrats total
SQT <- SQD + SQE
SQT
```

```
## [1] 344.5227
```

7. Calculo l'estadístic de contrast: $f = \frac{SQE/(k-1)}{SQD/(n-k)} = 3.003$

```r
# calculo la suma dels quadrats total
f <- (SQE / (k-1)) / (SQD / (n-k))
f
```

```
## [1] 3.002995
```

5 Adequació del model

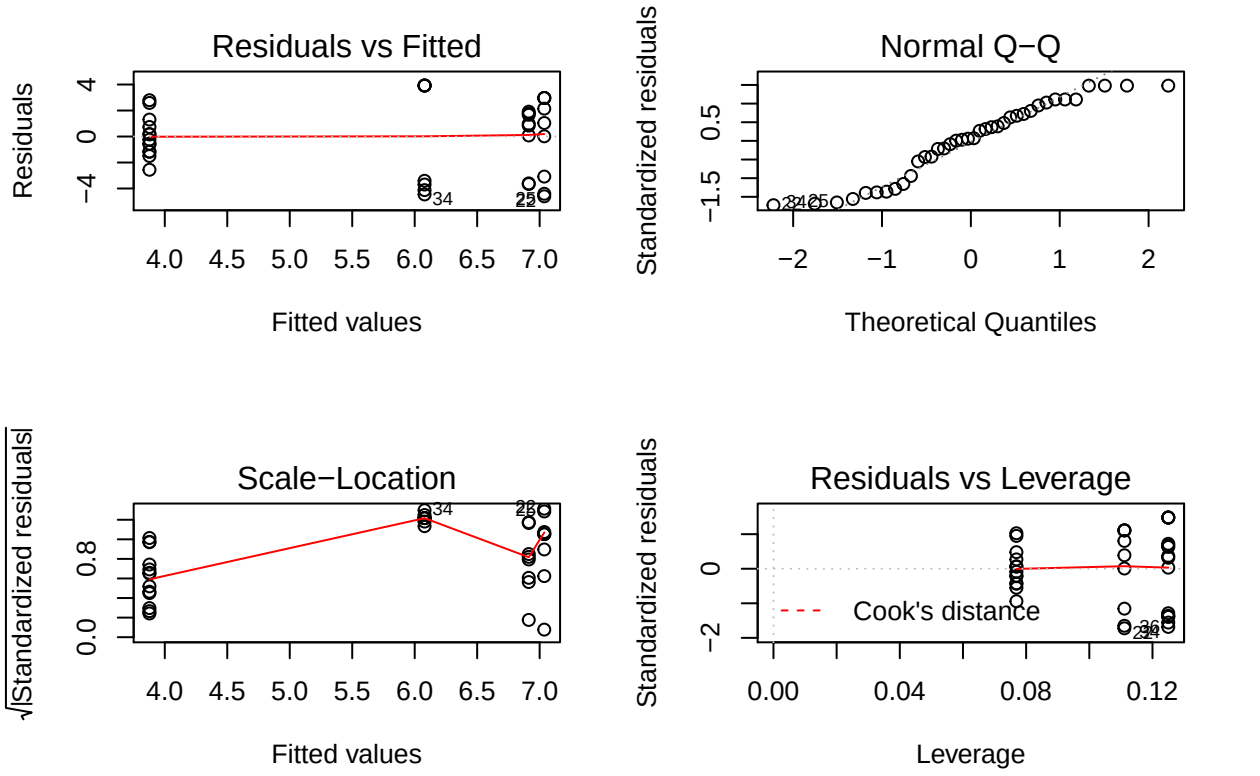
Es validarà l'adequació del model ANOVA. Podeu consultar López-Roldán i Fachelli (2015), gràfic III.8.6, pàgina 25.

5.1 Visualització de l'adequació del model

Mostreu visualment l'adequació del model ANOVA. Podeu usar plot sobre el model ANOVA resultant. En els apartats següents us demanem la interpretació d'aquests gràfics.

Si a la funció *plot* li passo per paràmetre el model ANOVA, retorna 4 gràfics que permeten visualitzar l'adeqüació del model.

```
# visualitzo el model ANOVA
par(mfrow=c(2,2))
plot (anova)
```



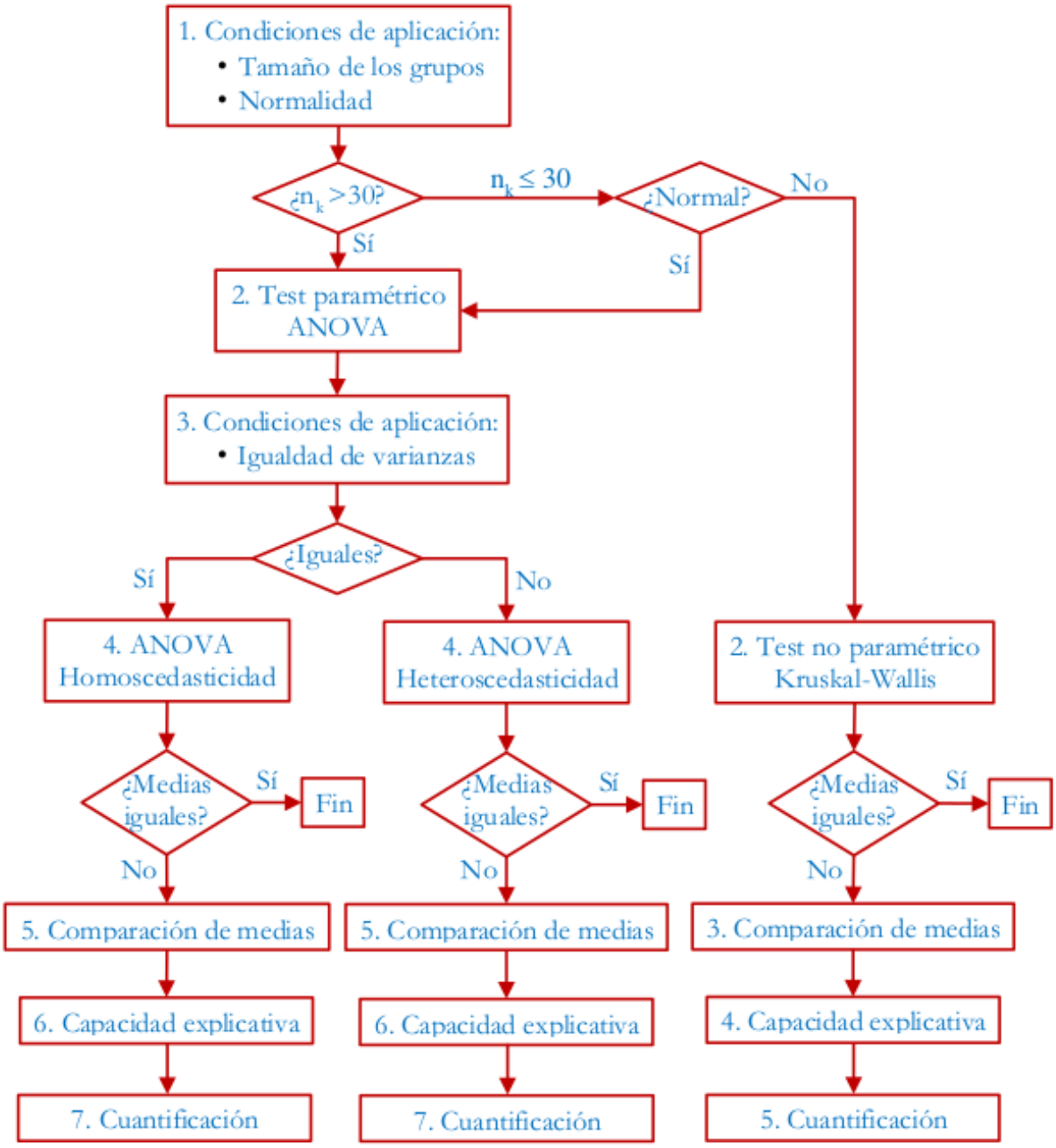


Figure 1: Procés d'anàlisi de ANOVA (gràfic III.8.6 de López-Roldán i Fachelli (2015))

5.2 Normalitat dels residus

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

Amb la gràfica *'Normal Q-Q'* es visualitza si els residus es distribueixen normalment.

Observo que els residus no es distribueixen normalment al llarg de la línia recta discontinua, dibuixa una corva S i hi ha alguns punts allunyats de la línia que indiquen que és una distribució amb outliers.

Si s'observa que els residus no es distribueixen normalment, potser que els intervals de predicció no siguin acurats.

5.3 Homoscedasticitat dels residus

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

Podeu consultar informació complementària sobre com interpretar aquests gràfics a l'enllaç següent:

- https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/anova/how-to/one-way-anova/interpret-the-results/all-statistics-and-graphs/#normal-probability-plot-of-the
- https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/anova/how-to/two-way-anova/interpret-the-results/all-statistics-and-graphs/#residuals-versus-fits

El gràfic *Residuals vs Fitted* permet verificar la suposició que els residus es distribueixen al atzar i tenen una variació constant. Idealment, els punts han de caure a l'atzar a ambdos costats del 0, sense patrons identificats en els punts.

Si s'observa algun punt allunyat, indica que és un valor atípic. Si hi ha massa valors atípics, el model pot no ser acceptable, per tant, cal identificar la causa de qualsevol valor atípic i corregir qualsevol error en l'entrada de dades o medició o bé considerar l'eliminació dels valors atípics i tornar a realitzar l'anàlisi.

Observo que els punts 22,25 i 34 es detecten com a outliers, fet que pot afectar greument la normalitat i la homogeneitat de la variància. Potser serà útil eliminar els valors atípics per complir amb els supòsits del model d'anàlisi.

A més a més, si s'observa una distribució desigual de residus a través dels valors ajustats, indica una variància no constant.

Observo que a mesura que augmenta el valor dels *fitted*, la dispersió entre els residus s'eixampla. Aquest patró indica que la variància dels residus no és constant, per tant, no es compleix la igualtat de variància.

5.4 ANOVA no paramètric

5.4.1. Kruskal-Wallis

Si les condicions d'aplicació d'ANOVA no es compleixen, llavors se sol aplicar un test no paràmetric. El test Kruskal-Wallis és l'equivalent no paramètric d'ANOVA. Apliqueu el test usant la funció kruskal.test i interpreteu el resultat.

Podeu consultar l'enllaç següent:

- https://www.sheffield.ac.uk/polopoly_fs/1.714570!/file/stcp-karadimitriou-KW.pdf

Com les variàncies de les mostres no són idéntiques, provem un ANOVA no paramètric com el *Kruskal-Wallis.*

```
# aplico el test Kruskall-Wallis
kw <- kruskal.test (SEstudis ~ factors)
kw
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SEstudis by factors
## Kruskal-Wallis chi-squared = 5.5492, df = 3, p-value = 0.1357
```

5.4.2 Interpretació

Explicar en què es diferencia el càlcul d'ANOVA del càlcul del test de Kruskal-Wallis. Si el test de Kruskal-Wallis no té les mateixes restriccions d'aplicació que el test ANOVA, per què creus que es continua aplicant ANOVA, si les condicions de satisfacció ho permeten, enlloc de Kruskal-Wallis?

La principal diferencia entre el test *ANOVA* i el test de *Kruskal-Wallis* és que el test *Kruskal-Wallis* s'ha d'aplicar quan les dades no compleixen les assumpcions necessàries per aplicar el test *ANOVA*. Tot i que cal tenir en compte que els tests no paramètrics no tenen tanta potència o poder estadístic, és a dir, augmenta la probabilitat de cometre un error del tipus II o fals negatiu (no rebutjar la hipòtesi nul · la, quan aquesta és falsa).

A més, el test *ANOVA* serveix per analitzar les diferencies entre les mitjanes dels grups, en canvi, el test Kruskal-Wallis s'utilitza per comparar les medianes de 3 o més grups independents, és a dir, es basa en rangs, en lloc de valors concrets, per tant, és un bona opció per a distribucions sesgades o amb valors atípics.

Per tant, com el test *Kruskal-Wallis* s'utilitza per comprovar si les medianes són idèntiques, podem establir les hipòtesis següents:

H_0: totes les medianes són iguals

H_1: alguna de les medianes és diferent

Del resultat del test *Kruskal-Wallis* obtinc un p-value (0.1357) superior a 0.05, per tant, no hi ha evidència suficient per a rebutjar la hipòtesi nul · la i podem concloure, amb un nivell de significació del 5%, que totes les medianes dels grups de satisfacció laboral en funció del nivell d'estudis són iguals.

Amb la funció *median* calculo les medianes de cada un dels grups i comprovo que no són iguals ($me_1=3.64, me_2=7.82, me_3=8.08, me_4=6.34).

Per tant, considero que és millor continuar amb l'anàlisi *ANOVA* perquè el test *Kruskal-Wallis* no rebutja la hipòtesi nul · la, quan aquesta és falsa.

```
# calculo la mediana de cada mostra
mdSE1 <- median(SE1)
mdSE1
```

```
## [1] 3.636904
```

```
mdSE2 <- median(SE2)
mdSE2
```

```
## [1] 7.820422
```

```
mdSE3 <- median(SE3)
mdSE3
```

```
## [1] 8.076564
```

```
mdSE4 <- median(SE4)
mdSE4
```

```
## [1] 6.335843
```

6 ANOVA multifactorial

A continuació, es vol avaluar l'efecte de la qualificació del treball combinat amb un altre factor. Primer es realitzarà l'anàlisi amb el factor nivell d'estudis i posteriorment, amb el factor sexe.

6.1 Factors: tipus de treball i nivell educatiu

6.1.1 Anàlisi visual dels efectes principals i possibles interaccions

Dibuixeu en un gràfic la satisfacció laboral en funció del tipus de treball i en funció del nivell educatiu. El gràfic ha de permetre avaluar si hi ha interacció entre els dos factors.

Per això, us recomanem que seguiu els passos següents:

1. Agrupeu el conjunt de dades per tipus de treball i per nivell d'estudis. Calculeu la mitjana de satisfacció laboral per cada grup. Podeu fer-ho amb les funcions group_by i summarise de la llibreria dplyr per a realitzar aquest procés.
2. Mostreu el conjunt de dades en forma de taula, on es mostri la mitjana de cada grup segons el tipus de treball i el nivell d'estudis.
3. Mostrar en un gràfic el valor S promig per cada tipus de treball i educació. Us podeu inspirar en els gràfics de López-Roldán y Fachelli (2015), p.38. Podeu realitzar aquest tipus de gràfic usant la funció ggplot de la llibreria ggplot2.
4. Interpreteu el resultat sobre si sols existeixen efectes principals o hi ha interacció entre els factors. Si hi ha interacció, expliqueu com s'observa aquesta interacció en el gràfic.

6.1.2 ANOVA multifactorial

A continuació, realitzeu l'anàlisi ANOVA usant els factors i si s'escau, la interacció entre els mateixos.

Interpreteu el resultat del model i expliqueu si els factors (i la interacció entre factors) són significatius per a explicar la satisfacció laboral.

6.1.3 Adequació del model

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

6.2 Factors: tipus de treball i sexe

6.2.1 Anàlisi visual dels efectes principals i possibles interaccions

Realitzeu l'anàlisi visual dels factors tipus de treball i sexe, de manera anàloga al cas anterior.

Interpreteu el resultat del gràfic en relació als efectes principals i possibles interaccions.

6.2.2 ANOVA multifactorial

Realitzeu l'anàlisi ANOVA amb els factors tipus de treball, sexe i interacció, si s'escau.

Interpreteu el resultat.

7 Comparacions múltiples

Prenent com a referència el model ANOVA multifactorial, amb els factors tipus de treball i nivell d'estudis, apliqueu el test de comparació múltiple Scheffé. Interpreteu el resultat del test i indiqueu quins grups són diferents significativament entre si.

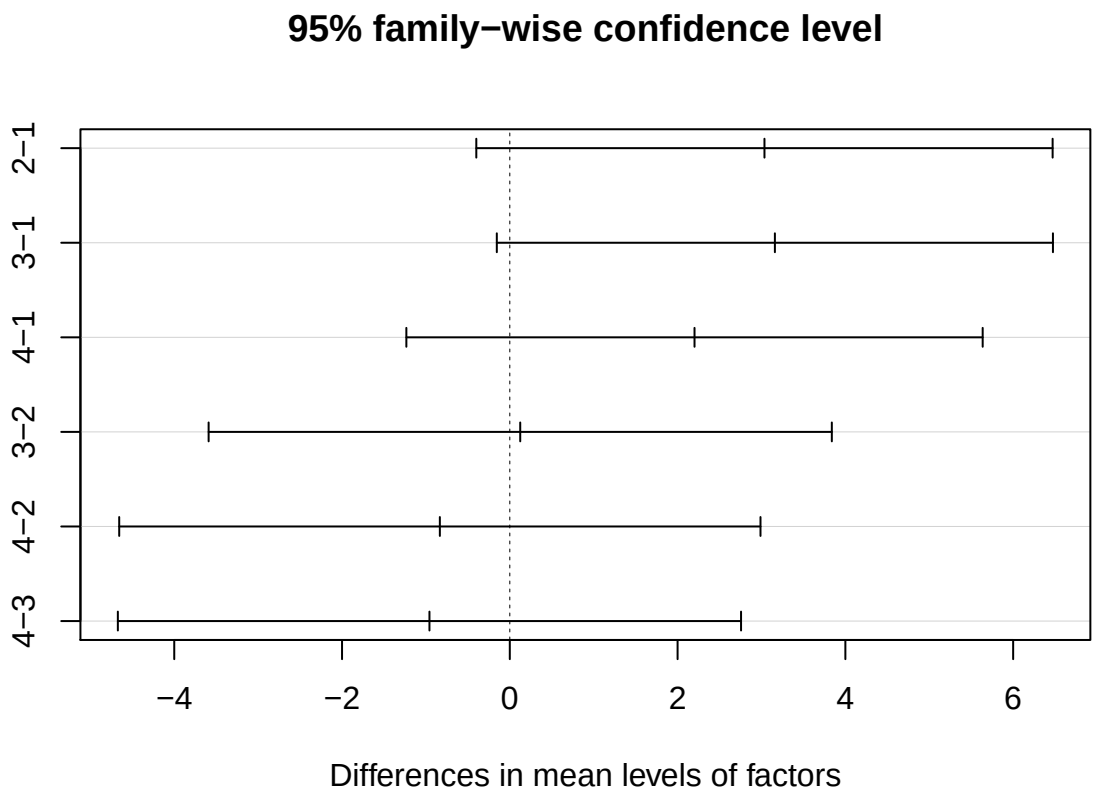
8 Conclusions

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

```
intervals <- TukeyHSD(anova)
intervals
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(SEstudis ~ factors))
##
## $factors
##           diff        lwr      upr     p adj
## 2-1  3.0355987 -0.3993095 6.470507 0.0989172
## 3-1  3.1597482 -0.1549239 6.474420 0.0662894
## 4-1  2.2017331 -1.2331751 5.636641 0.3237938
## 3-2  0.1241495 -3.5901796 3.838479 0.9997309
## 4-2 -0.8338656 -4.6558787 2.988147 0.9346475
## 4-3 -0.9580151 -4.6723442 2.756314 0.8976708
```

```
plot(intervals)
```



```r
shapiro.test(anova$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova$residuals
## W = 0.93377, p-value = 0.02641
```

9 Referències

- Rmarkdown cheat sheet
  https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf
- Rmarkdown: The Definitive Guide
  https://bookdown.org/yihui/rmarkdown/pdf-document.html
- L. Kocbach, LaTeX Math Symbols
  http://web.ift.uib.no/Teori/KURS/WRK/TeX/symALL.html

López-Roldán, P., i S. Fachelli. 2015. «Capítulo III.8 Análisis de varianza». En Metodología de la investigación social cuantitativa, editat per P. López-Roldán i S. Fachelli. Barcelona: UAB.

https://rpubs.com/Joaquin_AR/226291

https://biocosas.github.io/R/050_anova.html

http://www.diegocalvo.es/anova-en-r/

https://stats.stackexchange.com/questions/76059/difference-between-anova-and-kruskal-wallis-test

https://www.slideshare.net/doncua1/power-study-anova-vs-kruskall-wallis

- Joaquín Amat Rodrigo
  Regresión logística simple y múltiple
  https://rpubs.com/Joaquin_AR/229736
- Generalized Linear Models
  https://www.statmethods.net/advstats/glm.html
- Arpan Gupta (Indian Institute of Technology,Roorkee)
  Logistic Regression output interpretation in R
  https://analyticsdataexploration.com/logistic-regression-output-interpretation-in-r/
- R documentation - glm - Fitting Generalized Linear Models
  https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/glm
- Jaime Ashander Github
  So, you did some GLMs & compared with AIC. Congrats!
  http://www.ashander.info/posts/2015/10/model-selection-glms-aic-what-to-report/
- Wangzhefeng
  pROC-Package
  https://rpubs.com/Wangzf/pROC
- R documentation - roc - Build a ROC curve
  https://www.rdocumentation.org/packages/pROC/versions/1.13.0/topics/roc
- R documentation - plot.roc - Plot a ROC curve
  https://www.rdocumentation.org/packages/pROC/versions/1.13.0/topics/plot.roc