Informe Treball Final Anàlisi Estadístic

Arnau Busquets Domingo i David Martí Felip

Table of Contents

Introducció al Problema

Aquest treball consisteix en l'aplicació dels conceptes bàsics de Regressió lineal múltiple, Models generalitzats i Dissenys aleatoritzats per blocs complets.

En el nostre cas, la temàtica escollida ha estat Powerlifting, un esport d'eixecament de pesos que es centrar en els tres moviments principals: sentadilla (squat), el press de banca (bench press) i el pes mort (deadlift).

Utilitzarem un dataset tret de OpenPowerlifting Data Service que es pot trobar en la bibliografia.

Aquesta extensa base de dades recopila l'informació de 1219498 atletes (files) amb 41 atributs per atleta (columnes). Però d'aquest conjunt tant gran l'hem simplificat per poder tractar les dades de manera més comode i rendeixi millor depenent de la manera com requerís l'exercici.

Regressió Lineal Múltiple

Introducció

La regressió lineal múltiple és un tipus de model de regressió estadística que utilitza més d'una variable predictora per predir la variable dependent. Una suposició clau d'aquesta regressió és que els errors de la variable dependents es distribueixen normalment.

En aquest apartat s'susarà 4 variables independents explicatives per predir la dependent amb una mostra de 50. Entre les variables independents tenim l'edat de l'atleta (Age), el total de Kg que ha aixecat entre els 3 exercicics (TotalKg) i la nota que li posen els jutges a la seva forma (Goodlift) com a variables continues, el seu sexe (Sex) com a variables categòrica i es prediu els Kg de massa de l'atleta (BodyweightKg) què és una variables continua.

Objectius

1. Preparar el dataset per la regressió
2. Que el model compleixi totes les suposicions

3. Analitzar la significació dels coeficients i així determinar si és òptim remodelitzar el model apartir de validacions del model.
4. Visualitzar les dades per entendre millor el seu desenvolupament en el model.

Resultats

Càrrega de les dades

Aquest bloc carrega les llibreries necessàries i llegeix les dades des d'un arxiu CSV. També transforma la variable Sex en numèrica.

```
library(car)

## Warning: package 'car' was built under R version 4.2.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.2.3

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

library(MASS)
library(lmtest)

## Warning: package 'lmtest' was built under R version 4.2.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.2.3

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

data <- read.csv("combined_data.csv")
data$Sex <- as.numeric(data$Sex == "F")
```

Aquí es seleccionen algunes variables del conjunt de dades per utilitzar-les com a variables independents en l'anàlisi.

```
independent_vars <- data[, c("Age", "Sex", "TotalKg", "Goodlift")]
```

Creació del model

Es crea un model de regressió lineal múltiple amb la variable dependent BodyweightKg i les independents especificades anteriorment.

```
model <- lm(BodyweightKg ~ Age + Sex + TotalKg + Goodlift, data = data)
```

Resum del model

Proporciona un resum estadístic del model de regressió, incloent coeficients, significació estadística, R-squared, entre d'altres. Tamé els intervals de confiança de cada variable

```
summary(model)
```

```
##
## Call:
## lm(formula = BodyweightKg ~ Age + Sex + TotalKg + Goodlift, data =
data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.501  -1.852   0.450   1.701  26.373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.35995    7.68837   7.591 1.37e-09 ***
## Age         -0.06015    0.10661  -0.564    0.575
## Sex         37.61077    2.91822  12.888  < 2e-16 ***
## TotalKg      0.27576    0.01517  18.183  < 2e-16 ***
## Goodlift    -1.64347    0.16858  -9.749 1.15e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.171 on 45 degrees of freedom
## Multiple R-squared:  0.9164, Adjusted R-squared:  0.909
## F-statistic: 123.3 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
confint(model)
```

```
##                   2.5 %     97.5 %
## (Intercept) 42.8747752 73.8451258
## Age         -0.2748695  0.1545780
## Sex         31.7331687 43.4883695
## TotalKg      0.2452147  0.3063068
## Goodlift    -1.9830174 -1.3039246
```

Podem observar els diferents quartils dels residuals, que són la diferència entre els valors predits i els valors reals. D'aqui podem veure com el l'error màxim per sobre és de 26kg i per sota de 12kg, però el 50 per cent de les dades (desde el primer fins al tercer quartil) no s'equivoca en més de 1.701 kg per dalt i 0.45 per sota. Això ens mostra que té una bona capacitat de predicció i que sol equivocar-se a la alta. Cada coeficient representa la contribució d'una variable independent al model. Intercept (Intercepció): L'intercepció és 58.36, la qual cosa significa que quan totes les variables independents són zero, s'espera un pes corporal de 58.36 kg.

Age (Edat): El coeficient d'edat és -0.06, la qual cosa indica que un augment de 1 any en l'edat s'associa amb una disminució de 0.06 kg en el pes corporal, però no és estadísticament significatiu (p = 0.575).L'interval de confiança inclou el zero, el que suggereix que no podem estar segurs que l'edat tingui un efecte significatiu sobre el pes corporal.

Sex (Sexe): El coeficient de sexe és 37.61, la qual cosa significa que si la variable "Sex" és igual a 1, s'espera un augment de 37.61 kg en el pes corporal en comparació amb si fos igual a 0. L'interval de confiança per al sexe no inclou el zero, indicant que estem 95% segurs que el sexe té un efecte significatiu en el pes corporal. El valor del coeficient suggereix que canviar de la categoria de referència de sexe a l'altra augmenta el pes corporal en una quantitat situada dins d'aquest interval.

TotalKg: El coeficient de TotalKg és 0.28, la qual cosa indica que un augment de 1 kg en la variable "TotalKg" s'associa amb un augment de 0.28 kg en el pes corporal.Aquest interval també exclou el zero, suggerint un efecte significatiu de TotalKg sobre BodyweightKg.

Goodlift: El coeficient de Goodlift és -1.64, la qual cosa significa que un augment de 1 en la variable "Goodlift" s'associa amb una disminució de 1.64 kg en el pes corporal.L'interval de confiança per a Goodlift també exclou el zero i és totalment negatiu, indicant que a mesura que Goodlift augmenta, es preveu que BodyweightKg disminueixi dins d'aquest interval.

L'error estàndard dels residus és una mesura de la dispersió dels residus al voltant de la línia de regressió. En aquest cas, és aproximadament de 6.171 kg.

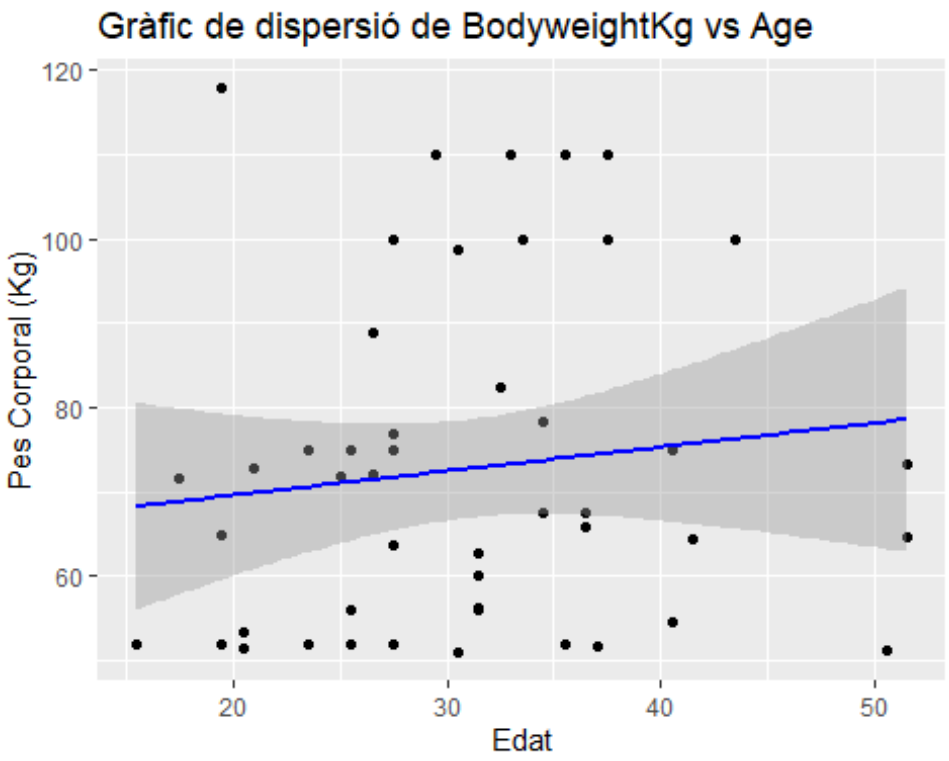
El coeficient de determinació múltiple (R^2) és de 0.9164, la qual cosa indica que aproximadament el 91.64% de la variabilitat en el pes corporal s'explica per les variables independents incloses en el model. És una mesura de quant s'ajusten les dades al model.

L'estadístic F és de 123.3 amb 4 i 45 graus de llibertat. S'utilitza per provar si almenys una de les variables independents és significativa en el model. En aquest cas, el valor de p és extremadament baix (menor que 2.2e-16), la qual cosa indica que almenys una de les variables independents és significativa en la predicció del pes corporal.

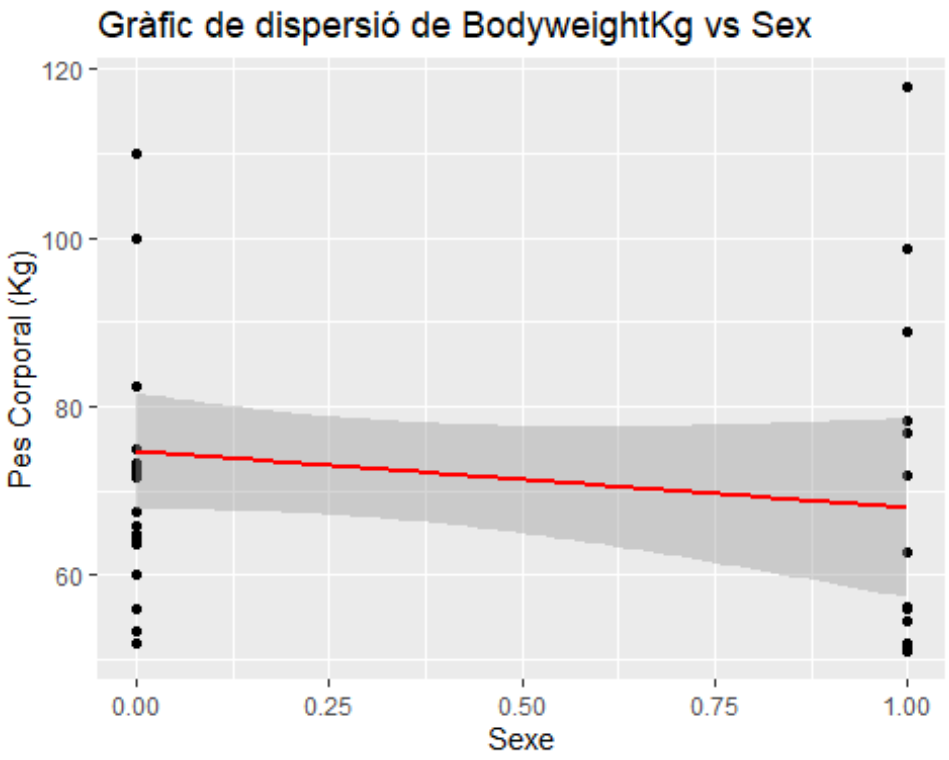
Supòsits del model

Linealitat:

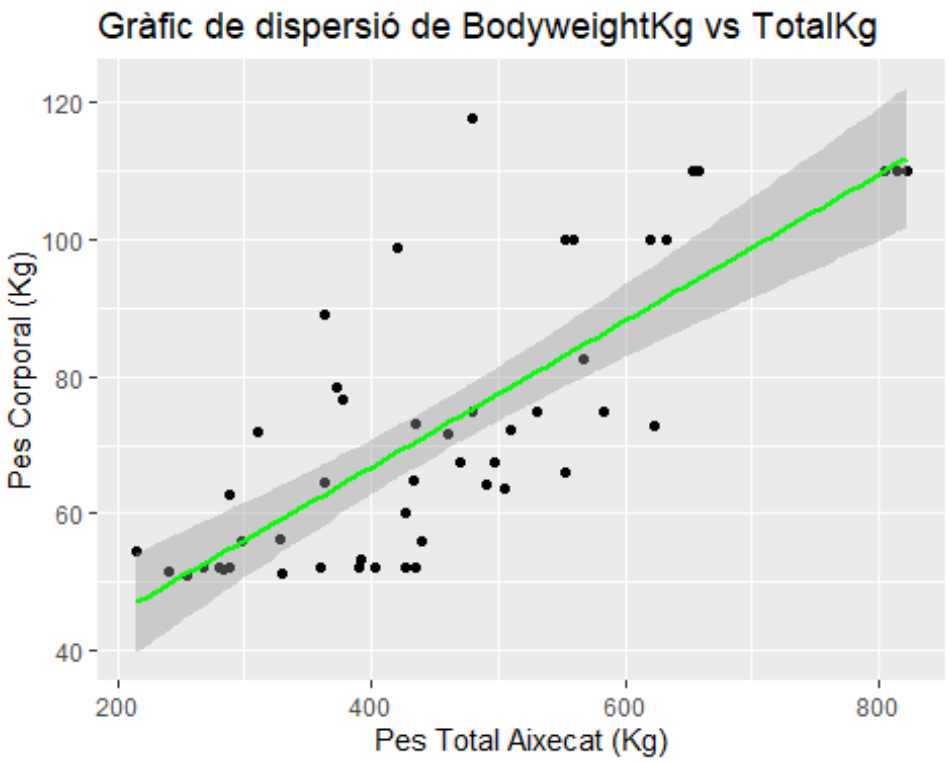
```
## `geom_smooth()` using formula = 'y ~ x'
```



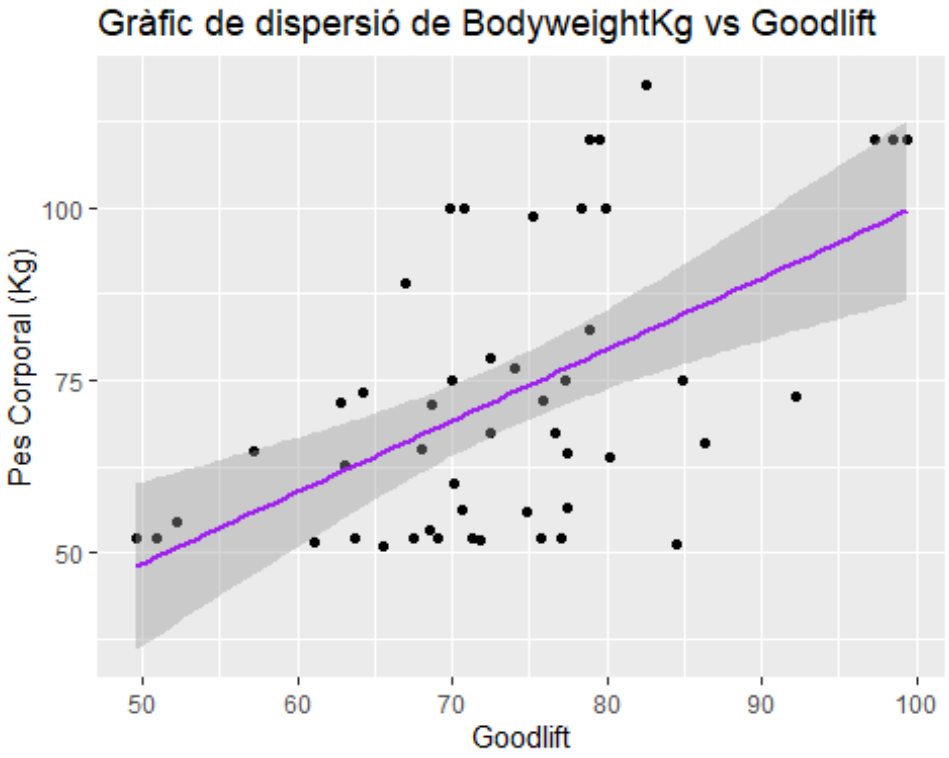

```
## `geom_smooth()` using formula = 'y ~ x'
```




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




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



Multicolinealitat:

Calcula i mostra la matriu de covariancies entre les variables independents, útil per detectar colinealitat.

```
matriu_vcov <- vcov(model)

print(matriu_vcov)
```

```
##             (Intercept)           Age         Sex       TotalKg
Goodlift
## (Intercept) 59.11105804 -0.3017946139  4.21646334  0.0532852776 -
1.0159479415
## Age         -0.30179461  0.0113657090 -0.04278514 -0.0001789913
0.0006320253
## Sex          4.21646334 -0.0427851359  8.51601835  0.0328060138 -
0.2777431734
## TotalKg      0.05328528 -0.0001789913  0.03280601  0.0002300102 -
0.0022112175
## Goodlift    -1.01594794  0.0006320253 -0.27774317 -0.0022112175
0.0284206977
```

Els valors de covariància entre Sex i TotalKg (0.03280601), Sex i Goodlift (-0.2777431734), i TotalKg i Goodlift (-0.0022112175) no semblen excepcionalment grans, el que suggereix que potser no hi ha una multicolinealitat significativa entre aquestes parelles de variables

Avalua la multicolinealitat en el model mitjançant el càlcul dels VIFs de les variables.

```
if(!any(is.na(model$coefficients))){
  vif_model <- vif(model)
  print(vif_model)
}
```

```
##      Age      Sex  TotalKg Goodlift
## 1.030961 2.347866 6.380766 4.230555
```

Age (Edat): 1.030961

Aquest valor és proper a 1, el que indica que hi ha una multicolinealitat mínima o inexistent amb les altres variables independents en el model. Sex (Sexe): 2.347866

Un VIF lleugerament superior a 2 no és preocupant. Generalment, es considera que hi ha un problema de multicolinealitat quan el VIF és superior a 5 (o en alguns casos, superior a 10). TotalKg: 6.380766

Aquest valor supera el llindar comú de 5, indicant que pot haver-hi un problema de multicolinealitat amb aquesta variable. Això significa que "TotalKg" podria estar altament correlacionat amb una o més de les altres variables independents en el model. Goodlift: 4.230555

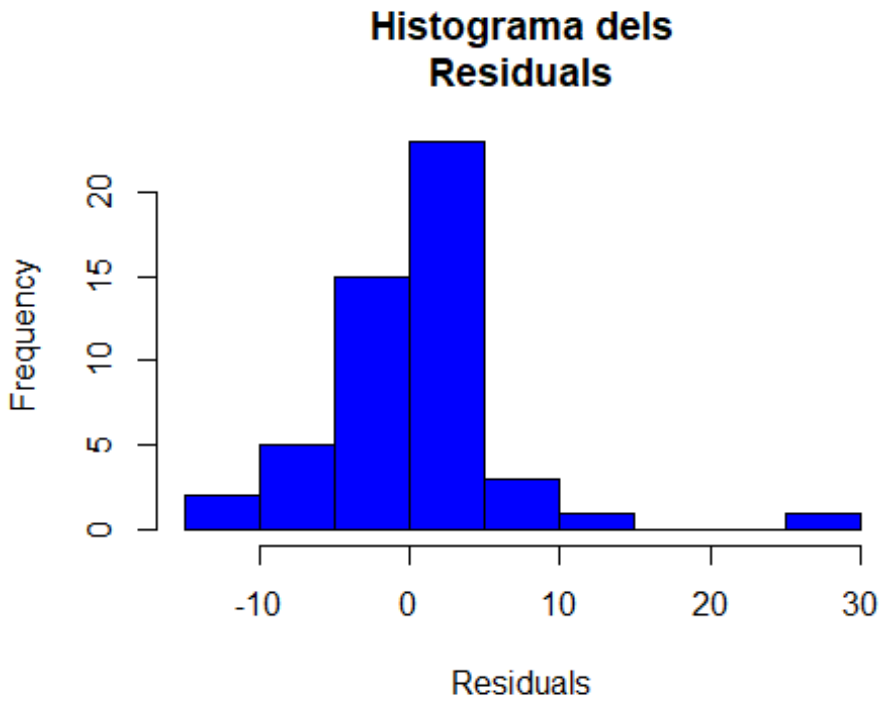
Aquest VIF està per sobre de 4 però per sota del llindar de 5, el que suggeriria una moderada multicolinealitat. Tot i que no és tan alt com per a "TotalKg", caldria considerar aquesta informació en l'interpretació dels resultats.

Normalitat en els errors:

Aplica el test de Shapiro-Wilk als residuals del model per comprovar si segueixen una distribució normal.

```
shapiro_test <- shapiro.test(resid(model))
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.84594, p-value = 1.203e-05
```



W = 0.84594:
Aquest és l'estadístic de Shapiro-Wilk. Els valors més propers a 1 indiquen que les dades s'apropen més a una distribució normal.

p-valor = 1.203e-05: El p-valor menor a 0.05 ens fa pensar que no pertanyen a una distribució normal. Tot i així creiem que això és perquè les dades amb les que contem són molt poques, podem veure que sí que s'hi apropa en l'histograma.

Homocedasticitat:

Realitza el test de Breusch-Pagan per detectar heteroscedasticitat en els errors del model.

```
bptest_model <- bptest(model)
print(bptest_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 12.994, df = 4, p-value = 0.01131
```

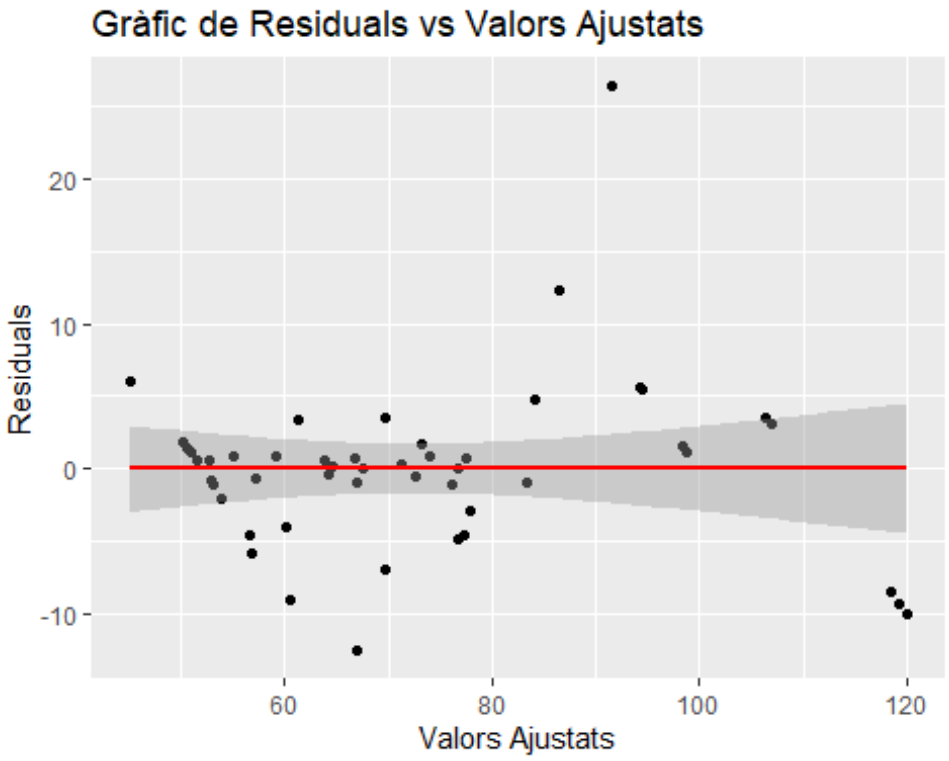
Basant-nos en aquest resultat (p-valor menor a 0.05), podríem concloure que hi ha una presència significativa de heteroscedasticitat en el model. Això pot indicar que el model podria no ser el millor ajust per a les dades. Creiem que això es dona per la poca quantitat de dades que tenim.

Veurem més informació en el següent gràfic.

Aleatorietat (independència en els errors):

Crea un gràfic de dispersió dels residuals versus els valors ajustats del model per visualitzar possibles patrons en els errors.

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



Els residus s'escampen a l'atzar al voltant de la línia de zero residus, sense cap patró clarament discernible o tendència sistemàtica. Això és una indicació que l'aleatorietat dels residus podria ser adequada, la qual cosa és consistent amb el supòsit d'independència dels errors en la regressió lineal.

Sobre el supòsit anterior: Sembla que els residus podrien estar mostrant una lleugera heteroscedasticitat, ja que la dispersió dels residus sembla augmentar lleugerament amb els valors ajustats, però sense una tendència clara o forta.

Validació Global

Per a fer la validació global recuperem aquestes dades del resum del model que hem fet a l'inici:

Multiple R-squared: 0.9164, Adjusted R-squared: 0.909 F-statistic: 123.3 on 4 and 45 DF, p-value: < 2.2e-16

L'estadístic F de 123.3 i el p-valor menor que 2.2e-16 rebutgen la hipòtesi nul·la que les variables independents no aporten informació significativa al model. El coeficient de determinació R-quadrat de 0.9164 mostra que un 91.64% de la variabilitat en el pes corporal és explicada pel model. El R-quadrat ajustat de 0.909 confirma que la inclusió de variables independents és pertinent i no introdueix variabilitat inexplicable. Aquests resultats suporten l'eficàcia del model en capturar la relació entre les variables independents i la variable dependent.

En resum les dades indiquen que el model és estadísticament significatiu en predir el pes corporal (BodyweightKg) a partir de les variables Age, Sex, TotalKg i Goodlift.

Validació Individual

Per a la validació global també hem extret les dades del resum de l'inici i per a cada variable independent fem la següent valoració

Age

- Coeficient: -0.06015
- Error Estàndard: 0.10661
- Valor t: -0.564
- Valor p: 0.575
- Interpretació: L'edat no és estadísticament significativa en aquest model, ja que el seu valor p és molt superior al nivell de significació comú de 0.05.

Sex

- Coeficient: 37.61077
- Error Estàndard: 2.91822
- Valor t: 12.888
- Valor p: < 2e-16
- Interpretació: El sexe és una variable molt significativa, amb un coeficient elevat i un valor p pràcticament nul, indicant que té un impacte important en el pes corporal.

TotalKG

- Coeficient: 0.27576
- Error Estàndard: 0.01517
- Valor t: 18.183
- Valor p: < 2e-16
- Interpretació: El Pes Total Aixecat és altament significatiu, demostrant ser un predictor robust del pes corporal amb un valor p extremadament baix.

Goodlift

- Coeficient: -1.64347
- Error Estàndard: 0.16858
- Valor t: -9.749
- Valor p: 1.15e-12
- Interpretació: La variable Goodlift també és significativa en el model, amb un valor p molt baix, encara que el coeficient negatiu indica que a mesura que Goodlift augmenta, el pes corporal tendeix a disminuir.

En general, la variable `Age` no ha demostrat ser un predictor significatiu del pes corporal en aquest model, mentre que `Sex`, `TotalKg` i `Goodlift` són estadísticament significatius i aporten de manera substancial a la capacitat predictiva del model.

Model Restringit

Seguint les interpretacions anteriors per al model restringit obviarem la variable Age i tornarem a calcular-lo, també tornarem a avaluar-lo globalment per veure si hi ha millores significatives.

```
model_restringit <- lm(BodyweightKg ~ Sex + TotalKg + Goodlift, data =
data)
summary(model_restringit)
```

```
##
## Call:
## lm(formula = BodyweightKg ~ Sex + TotalKg + Goodlift, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.085  -2.184   0.421   1.486  27.202
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 56.76290    7.09508   8.000 2.95e-10 ***
## Sex         37.38436    2.86900  13.030  < 2e-16 ***
## TotalKg      0.27481    0.01496  18.369  < 2e-16 ***
## Goodlift    -1.64013    0.16723  -9.808 7.52e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.125 on 46 degrees of freedom
## Multiple R-squared:  0.9158, Adjusted R-squared:  0.9103
## F-statistic: 166.8 on 3 and 46 DF,  p-value: < 2.2e-16
```

L'error estàndard residual ha disminuït lleugerament de 6.171 a 6.125, indicant un ajust marginalment millor dels residus al voltant de la línia de regressió. El coeficient de determinació R-quadrat ha canviat lleugerament de 0.9164 a 0.9158, la qual cosa significa que el model sense Age explica gairebé la mateixa proporció de la variabilitat en BodyweightKg que el model original. L'R-quadrat ajustat ha incrementat de 0.909 a 0.9103, suggerint que, després d'ajustar pel nombre de predictors, el model restringit potser és lleugerament més adequat. L'estadístic F ha augmentat de 123.3 a 166.8, i el p-valor associat continua sent menor que 2.2e-16, confirmant la significació estadística del model restringit.

En resum,el model restringit que exclou Age manté la seva significació global i fins i tot mostra una lleugera millora alguns punts. Això indica que Age no aportava informació addicional significativa en la predicció del pes corporal en presència de les altres variables.

Exemple de predicció

Per a veure en un exemple clar el funcionament del model farem dues prediccions. Una d'un home que fa un Goodlift de 55 i TotalKg 800 i una altra d'una dona amb un Goodlift de 80 i Totalkg de 255

```
nous_valors <- data.frame(Sex = c(1),  # Suposem que 1 representa el sexe masculí
                          TotalKg = c(800),  # TotalKg suposat per a la predicció
                          Goodlift = c(55))  # Goodlift suposat
prediccio <- predict(model_restringit, newdata = nous_valors)
print(prediccio)
```

```
##        1
## 223.7911
```

```
nous_valors2 <- data.frame(Sex = c(0),  # Suposem que 1 representa el sexe masculí
                          TotalKg = c(255),  # TotalKg suposat per a la predicció
                          Goodlift = c(80))  # Goodlift suposat
prediccio2 <- predict(model_restringit, newdata = nous_valors)
print(prediccio2)
```

```
##        1
## 223.7911
```

##Conclusions

Regressió logística

Introducció

Fins ara hem utilitzat el model de regressió multiple lineal el qual es suposa que els errors tenen una distribució Normal, per així poder fer proves estadístiques més confiables i vàlides, però en aquest exercici, es deixa aquesta suposició enrere, per utilizar Models Lineals Generalitzats com son la regressió logísitca i la de Poisson.

La regressió logísitca és usada per modelar la probabilitat d'un event binari apartir d'un conjunt de variables numèricas o categòricas.

Aamb el dataset que ja hem explicat de powerlifting, hem utilitzat la regressió logísitica per predir si un atleta utilitza o no Equipament (variable dependent) apartir de la seva Edat, com de bo ha fet l'aixecament (Goodlift) i el total en kg que ha aixecat (variables independents).

Objectius

1. Preparar el dataset per la regressió

2. Analitzar quins son els coeficients estadísticament significatius i veure si cal remodelar el model per aconseguir millors prediccions.
3. Veure apartir de la matriu de variàncies i covariàncies quin és l'error de les prediccions , la correlació entre les variables i determinar encara millor si remodelar o no es una bona opció.
4. Representar la logística de manera gràfica per saber com estan representades les prediccions amb el resultat real.

Resultats

Càrrega de les dades

Primerament hem carregat el dataset i l'hem netejat de manera que ens quedi només les 4 columnes necessaries per aquesta regressió. Cal dir que com volem una predicció balancejada ens hem encarregat anteriorment de igualar el nombre de atletas que utilitzen equipament (Raw) amb els que no (Wraps).

```
columnas_lr <- c("Equipment","Age", "Goodlift", "TotalKg")
dataset_lr <- data[,columnas_lr]
dataset_lr$Equipment <- factor(dataset_lr$Equipment, ordered = TRUE)

summary(dataset_lr)
```

```
##  Equipment       Age           Goodlift        TotalKg
##  Raw  :25   Min.   :15.50   Min.   :49.59   Min.   :215.0
##  Wraps:25   1st Qu.:25.50   1st Qu.:68.14   1st Qu.:360.6
##             Median :30.50   Median :72.42   Median :435.0
##             Mean   :30.77   Mean   :73.42   Mean   :455.3
##             3rd Qu.:35.50   3rd Qu.:78.68   3rd Qu.:546.9
##             Max.   :51.50   Max.   :99.40   Max.   :822.5
```

Si analitzem la base de dades podem veure com encara que tinguin valors molt extrems en els mínims i màxims la resta estàn molt poc dispersos i propers a la mitjana.

Creació del model

Ja tenim la base de dades, ara cal crear el model amb totes les variables mencionades, seguint una familia binomial.

```
dataset_lr$Equipment <- ifelse(dataset_lr$Equipment == "Raw", yes=0,
no=1)
logistic <- glm(Equipment ~ ., data=dataset_lr, family="binomial")

summary(logistic)
```

```
##
## Call:
## glm(formula = Equipment ~ ., family = "binomial", data = dataset_lr)
##
```

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.98943  -0.74304   0.09392   0.66503   1.65324
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.112388   3.842793   1.591 0.111697
## Age         -0.059992   0.044908  -1.336 0.181589
## Goodlift    -0.213328   0.079389  -2.687 0.007207 **
## TotalKg      0.024833   0.007372   3.369 0.000756 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.315  on 49  degrees of freedom
## Residual deviance: 44.098  on 46  degrees of freedom
## AIC: 52.098
##
## Number of Fisher Scoring iterations: 5
```

De adalt a abaix, primer ens trobem amb la Deviance Residuals que ens dona una mesura entre el valor predit i el real. Analitzant aquests valors podem veure com la mediana és propera a 0 donant a entendre que la distribució del residuos és simètrica. Seguidament els coeficients veiem que l'Edat i Goodlift son negatius per tant a mesura que augmentin significarà que hi haurà més possibilitat de que NO tingui Equipament. També podem veure com el p-valor de Goodlift i TotalKg son menors que 0.05 i per tant son estadísticament significaius, a diferència de la variables Edat que no ho és i serà la que veurem si empitjora o no el model. La raó per la que l'Edat no es significativa pel model és perquè no està relacionada amb l'absència d'Equipament, cosa que per exemple, si que ve influenciada pels Kg que aixeca (TotalKg) ja que li proporciona més comoditat a l'hora de fer l'aixecament i força menys el cos de l'atleta i per més Kg que aixequi és més possible que estigui utilitzant equipament.

Reajustament del model

Ara cal analitzar si ens dona millors resultats el model remodelat (sense Age) o com l'hem ensenyat fins ara i per això farem una prova de raó de verosimilituds.

```
logistic_remodelat <- glm(Equipment ~ Goodlift + TotalKg,
data=dataset_lr, family="binomial")

library(lmtest)

lrtest(logistic, logistic_remodelat)
```

```
## Likelihood ratio test
##
## Model 1: Equipment ~ Age + Goodlift + TotalKg
## Model 2: Equipment ~ Goodlift + TotalKg
```

```
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -22.049
## 2   3 -22.986 -1 1.8735     0.1711
```

Com es pot veure, el P-valor del test és de major a 0.05, això vol dir no hi ha evidència estadísitica per dir que els dos models son equivalents i per tant ens quedarem amb el remodelat perquè els seus coeficients tenen més significació estadística que la variables exclosa.

Apartir d'ara només utilitzarem el model remodelat.

Matriu de variancies i covariancies i Intervals de confiança

Ara és moment de fixar-nos amb com de bones son les estimacions dels coeficients (matriu de variancies i covariancies), quin error estandard hi ha, els P-valors i els intervals de confiança amb la matriu següent:

```
matriu_cov <- vcov(logistic_remodelat)
matriu_cov
```

```
##              (Intercept)     Goodlift       TotalKg
## (Intercept)  9.157082086 -0.186324069  0.0098834075
## Goodlift    -0.186324069  0.005119025 -0.0004109890
## TotalKg      0.009883408 -0.000410989  0.0000444827
```

```
std.err <- sqrt(diag(matriu_cov))
cbind("Std error" = std.err,
      "Pr(>|z|)" = 2 * pnorm(abs(coef(logistic_remodelat)/std.err),
lower.tail=FALSE),
      LInf = coef(logistic_remodelat) - 1.96 * std.err,
      LSup = coef(logistic_remodelat) + 1.96 * std.err)
```

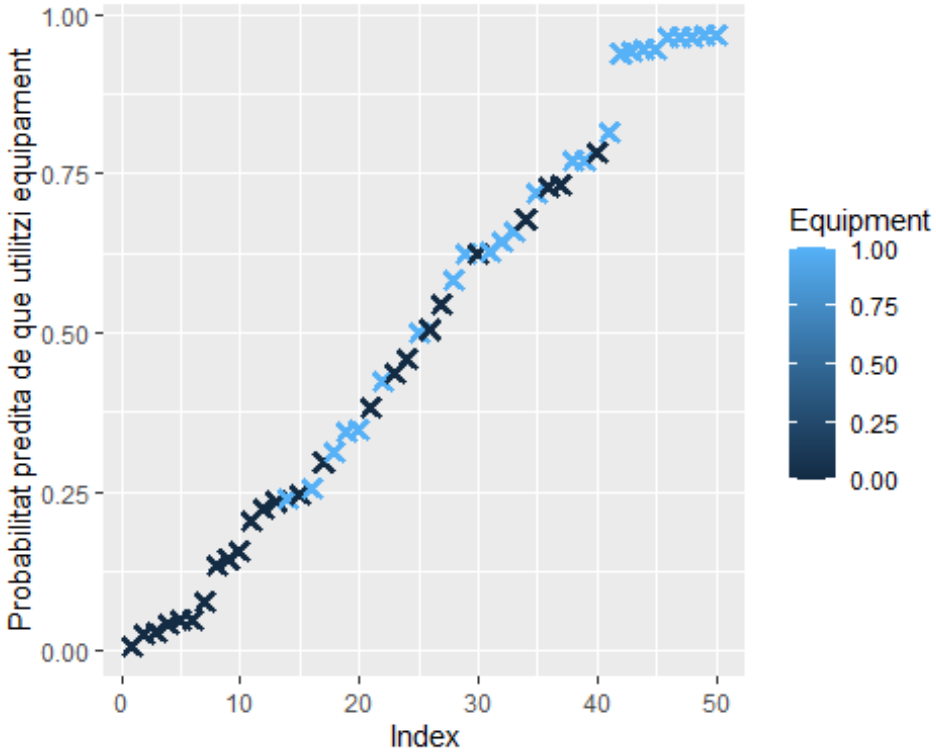
```
##               Std error     Pr(>|z|)         LInf        LSup
## (Intercept) 3.026067099 0.2174951236 -2.199296632  9.66288640
## Goodlift    0.071547362 0.0079399087 -0.330163495 -0.04969784
## TotalKg     0.006669535 0.0007779928  0.009340737  0.03548532
```

En la matriu de variancies ens podem trobar en primer lloc que evidentment la diagonal es tota positiva perquè son les variancies de les variables, i aquests valors son molt petits donant a entendre que l'estimació de la predicció es bona. Després, la resta de valors de la matriu, son les covariancies, les quals podem veure com cap valor és gaire gran i vol dir que no hi ha multicolinealitat entre variables cosa que és també bona per les prediccions.

Respecte la segona matriu, veiem l'error estandard, el p-valor i els límitis inferiors i superiors de cada variable. Una dada a destacar d'aquesta matriu és que entre els límits superiors i inferiors no hi és el 0 per tant aquestes dues variables podem determinar que son estadísticament significatives.

Visualització de la regressió

Ara toca plotejar la logísitica per veure com a predit.



Es pot veure com les prediccions no son les millors degut a la falta de dades que provoca que la generalització no sigui el màxim de precisa possible. Igualment, no dona mals resultats, ja que podem veure com en els dos extres de la regressió els valors estàn molt ben predits.

Conlcusions

1. La variable Edat no és estadísticament significativa pel model i per tant és més òptim remodelar el model sense ella.
2. Les altres dues variables si que ens aportes signficició al model com podem veure en els intervals de confiança, els P-valors, les varancies, els errors...
3. Els residuals no son gaire grans i a més la matriu de covariancies ens ensenya com entre les variables no hi ha multicolinealitat i per tant les prediccions son bones.
4. El model com a tal no prediu perfecte degut a la falta de dades, si el dataset fos major la precisió també ho seria.

Regressió de Poisson

Introducció

Ara és moment de la regressió de Poisson, un tipus de regressió que segueix un model lineal generalitzat com la logística però amb una altre utilitat, modelar les dades per predir una variable de conteig en un interval de temps o espai.

Per fer aquesta regressió en el nostre dataset hem creat una nova columna que és diu "FirstPlace", la qual és la que predirem, que indica el nombre total de cops que ha quedat aquell atleta en primera possició en tota la seva trajectoria (o almenys fins on el nostre datset arribi). Per fer aquesta predicció hem triat com a variables independents la variable continua "TotalKg" i la categòrica "AgeClass".

Objectius

1. Preparar el dataset per la regressió
2. Identificar les variables significatives i si el model cal ser remodelat o no per aconseguir millors prediccions.
3. Analitzar els coeficients per veure com afectan a la regressió.
4. Amb l'ajuda de les prediccions poder analitzar la freqüència en la que s'aconsegueixen primers llocs depenent de la categoria en la que estigui participant aquell atleta i quan aixequi.

Resultats

Càrrega de les dades

En primer lloc cal netejar la base de dades de les columnes no necessaries per aquesta regressió i només quedar-nos amb les 3 desitjades.

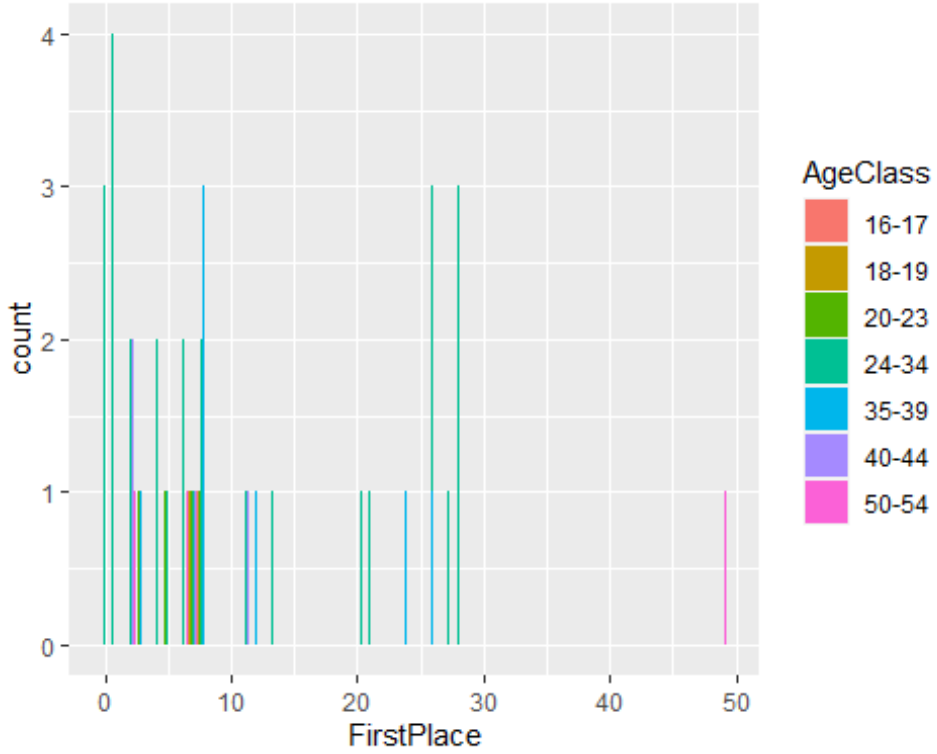
```
columnas_poiss <- c("FirstPlace", "AgeClass", "TotalKg")

dataset_pois <- data[,columnas_poiss]
dataset_pois$AgeClass <- factor(dataset_pois$AgeClass, ordered = TRUE)

summary(dataset_pois)
```

```
##    FirstPlace     AgeClass     TotalKg
##  Min.   : 0.00   16-17: 1   Min.   :215.0
##  1st Qu.: 2.00   18-19: 1   1st Qu.:360.6
##  Median : 7.00   20-23: 6   Median :435.0
##  Mean   :10.36   24-34:28   Mean   :455.3
##  3rd Qu.:12.75   35-39: 7   3rd Qu.:546.9
##  Max.   :49.00   40-44: 4   Max.   :822.5
##                  50-54: 3
```

Analitzant les columnes, es veu com "AgeClass" està clarament desbalancejada i cosa que ens portarà problemes en la regressió. També podem veure com les altres dues

variables tenen outliers com marcats en els màxims que ens aportaràn soroll i empitjoraràn la predicció.

Si fem un visualitzió de com estan repartits el total de primeres posicions per cada "AgeClass" ens trobem la següent gràfica.



Com hem mencionat abans, l'outlier que es troba amb 49 primeres posicions desproporciona la gràfica i fa que les columnes siguin més difícils d'apreciar però si ampliem la imatge podem veure com té sentit que estigui tant apartada perquè és un concursant d'entre 50-54 anys i per tant com porta molt de temps en l'esport té més competicions participades i per tant més primeres posicions. També podem apreciar com el rang de 24-34 que és el grup amb la majoria d'atletes es troba molt repartit i això pot significar que les seves prediccions seràn millors a causa de la gran quantitat d'informació que tenim.

Ja analitzada la base de dades, cal crear el model de la regressió.

```
poisson <- glm(FirstPlace ~ ., data=dataset_pois, family="poisson")
summary(poisson)

##
## Call:
## glm(formula = FirstPlace ~ ., family = "poisson", data = dataset_pois)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8966  -2.4958  -0.9094   1.9201   5.5130
```

```
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.3372574  0.3654455   6.396  1.6e-10 ***
## AgeClass18-19  0.0322073  0.5209517   0.062   0.9507
## AgeClass20-23 -0.5399445  0.4079718  -1.323   0.1857
## AgeClass24-34  0.4896664  0.3624462   1.351   0.1767
## AgeClass35-39  0.6560936  0.3748451   1.750   0.0801 .
## AgeClass40-44 -0.2240175  0.4159281  -0.539   0.5902
## AgeClass50-54  0.9524262  0.3788248   2.514   0.0119 *
## TotalKg       -0.0009208  0.0003302  -2.788   0.0053 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 494.62  on 49  degrees of freedom
## Residual deviance: 422.88  on 42  degrees of freedom
## AIC: 616.02
##
## Number of Fisher Scoring iterations: 5
```

Primerament ens trobem amb la Deviance Residuals la qual la mediana no és gaire propera a 0 i això significa que el model té més error del que hauria. Això, és degut a com es veu a continuació els coeficients, ja que la gran majoria d'ells no tenen un P-valor inferior a 0.05 i per tant estadísticament no significiatius menys el cas del rang d'edat de 50-54 i els TotalKg. Això es fruit de que el dataset és molt petit i els valors de les edats tenen tanta variancia que la predicció es molt poc ajustada als valors reals.

Per aquesta poca signficiació dels resultats, hem decidit provar augmentant el dataset amb 100 dades i analitzar si les prediccions milloren.

```
dataset_pois_100 <- read.csv("combined_data_100.csv")

columnas_poiss <- c("FirstPlace", "AgeClass", "TotalKg")

dataset_pois_100 <- dataset_pois_100[,columnas_poiss]
poisson_100 <- glm(FirstPlace ~ ., data=dataset_pois_100,
family="poisson")
summary(poisson_100)
```

```
##
## Call:
## glm(formula = FirstPlace ~ ., family = "poisson", data =
dataset_pois_100)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.2004  -2.8449  -0.8561   1.2374   8.3422
##
```

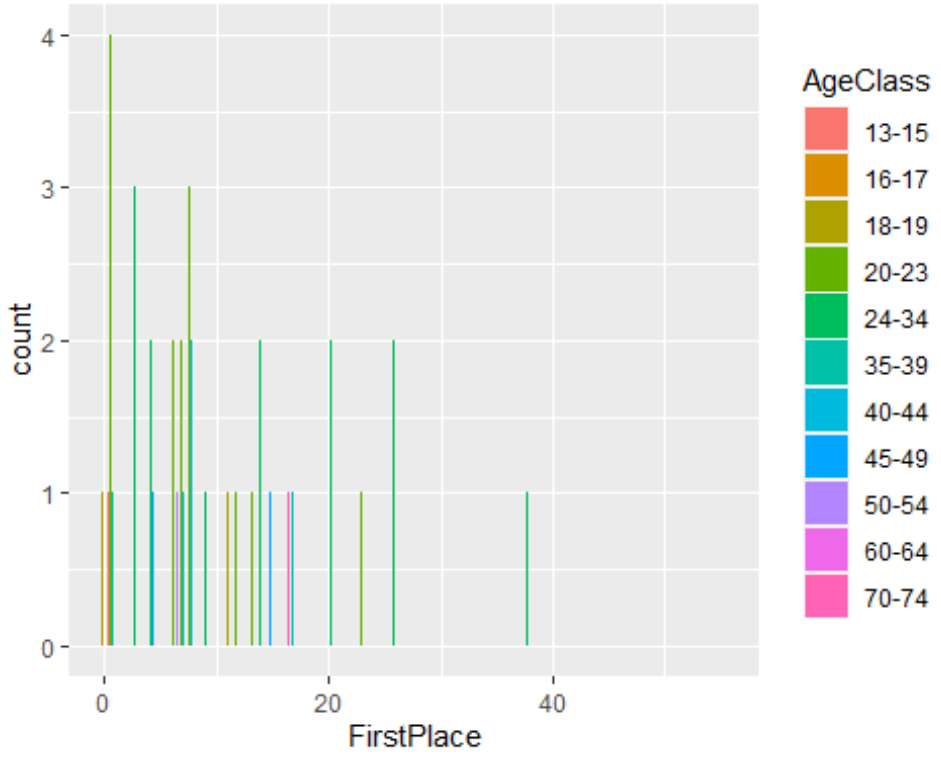
```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.5540456  0.3059505   5.079 3.79e-07 ***
## AgeClass16-17 0.3803971  0.3846615   0.989  0.32271
## AgeClass18-19 0.1793979  0.3554604   0.505  0.61378
## AgeClass20-23 0.3631240  0.3172111   1.145  0.25232
## AgeClass24-34 0.6200872  0.3103889   1.998  0.04574 *
## AgeClass35-39 0.4595582  0.3251969   1.413  0.15761
## AgeClass40-44 0.0592037  0.3293768   0.180  0.85735
## AgeClass45-49 0.5038969  0.3303378   1.525  0.12716
## AgeClass50-54 0.7377318  0.3249165   2.271  0.02318 *
## AgeClass60-64 1.8384534  0.3441001   5.343 9.15e-08 ***
## AgeClass70-74 0.9797123  0.3928352   2.494  0.01263 *
## TotalKg       0.0005755  0.0001978   2.909  0.00363 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 949.68  on 99  degrees of freedom
## Residual deviance: 853.79  on 88  degrees of freedom
## AIC: 1239.1
##
## Number of Fisher Scoring iterations: 5
```

Efectivament, amb més dades hi ha més variables estadísitcament significatives així que utilitzarem aquest dataset per així tenir prediccions més precises i lògiques.

Analitzant els coeficients ens podem fixar com la gran majoria són positius cosa que significa que quan més atletes hi ha en la predicció més possibilitat de que hi hagui més primers llocs. També podem veure com les variables més estadísiticament significatives son les que tenen rangs d'edats més altes i això pot ser degut a que els atletes un cop arriben a una edat si no han tingut una carrera de gaires primers llocs tendeixen més a retirar-se a que si han tingut sobint victòries. Per altre banda, quan son més joves son menys significatives possiblement perquè al portar poc de trajectoria encara els hi quedi molt per entrenar i a més que és on hi ha més participants i per tant hi ha menys possibilitat de guanyar.

Si tornem a fer l'anàlisi de la base de dades veiem el següent:

```
##    FirstPlace       AgeClass     TotalKg
##  Min.   : 0.00   24-34  :39   Min.   :215.0
##  1st Qu.: 3.00   20-23  :19   1st Qu.:405.4
##  Median : 8.00   40-44  :10   Median :564.2
##  Mean   :10.85   35-39  : 9   Mean   :539.7
##  3rd Qu.:14.00   45-49  : 7   3rd Qu.:660.0
##  Max.   :55.00   50-54  : 6   Max.   :890.0
##                  (Other):10
```



Els resultats poden semblar molt semblants perquè el rang d'edat 24-34 segueix desbalancejant molt les classes i la gràfica segueix molt distorsionada però la diferència és que ara aquests outliers tenen menys pes perquè ara hi ha més atletes que balancejan la gràfica i fan que no tingui tanta importància en la regressió.

Ja hem analitzat la base de dades i creat el model, ara cal saber amb la matriu de variancies i covariancies, l'error estandard i els intervals de confiança per així determinar com de significatives son les variables per la regressió.

```
matriu_cov <- vcov(poisson_100)


std.err <- sqrt(diag(matriu_cov))
cbind("Error std" = std.err,
      "Pr(>|z|)" = 2 * pnorm(abs(coef(poisson_100)/std.err),
lower.tail=FALSE),
      LInf = coef(poisson_100) - 1.96 * std.err,
      LSup = coef(poisson_100) + 1.96 * std.err)
```

```
##                 Error std     Pr(>|z|)          LInf         LSup
## (Intercept)   0.305950475 3.786240e-07  0.9543826477 2.1537085109
## AgeClass16-17 0.384661468 3.227052e-01 -0.3735393593 1.1343335957
## AgeClass18-19 0.355460446 6.137755e-01 -0.5173045728 0.8761003772
## AgeClass20-23 0.317211101 2.523172e-01 -0.2586097832 0.9848577311
## AgeClass24-34 0.310388948 4.574108e-02  0.0117248816 1.2284495588
## AgeClass35-39 0.325196856 1.576059e-01 -0.1778276091 1.0969440676
## AgeClass40-44 0.329376821 8.573532e-01 -0.5863748953 0.7047822425
```

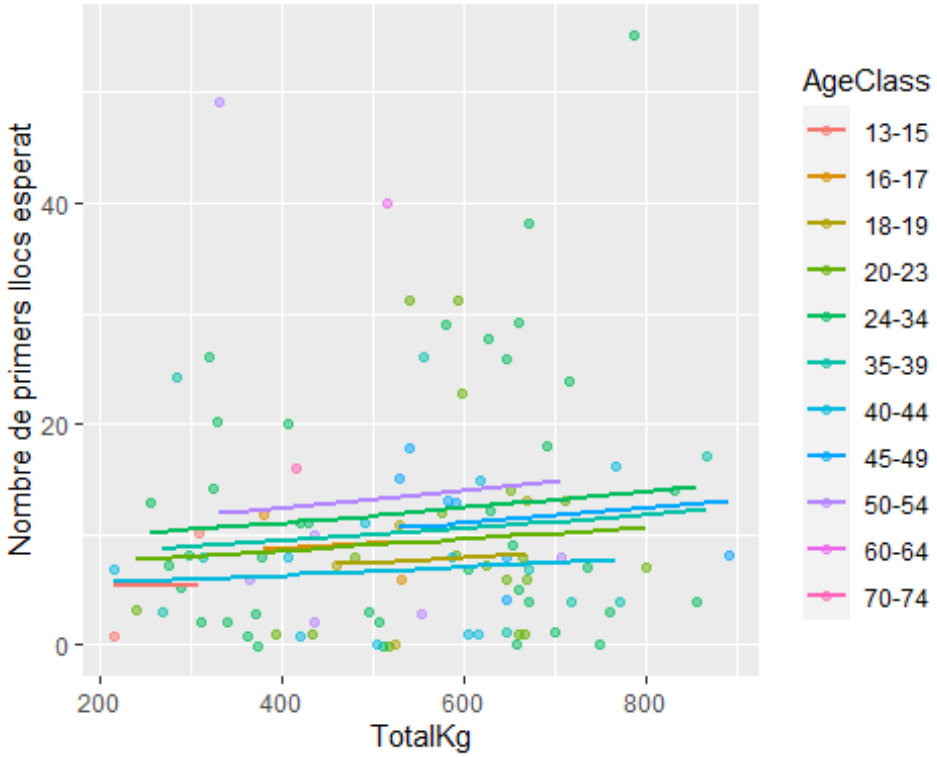
```
## AgeClass45-49 0.330337811 1.271597e-01 -0.1435652552 1.1513589657
## AgeClass50-54 0.324916536 2.317562e-02  0.1008953698 1.3745681891
## AgeClass60-64 0.344100118 9.152859e-08  1.1640172032 2.5128896638
## AgeClass70-74 0.392835220 1.263295e-02  0.2097552699 1.7496693339
## TotalKg       0.000197842 3.627423e-03  0.0001877256 0.0009632664
```

D'aquesta taula el que ens fixem és que els errors estàndards son força baixos destacant TotalKg el qual és el menor de tots i és perquè també és una variable amb un dels P-valor més baixos i per tant amb millors prediccions. També veiem que les variables que anteriorment ens donaven que eren estadísticament significiatives el seu interval de confiança no pasa per 0 i per tant ens reafirma aquesta significiació.

Ja per acabar, cal visualitzar aquesta regressió amb la base de dades de 100.

```
## calculate and store predicted values
dataset_pois_100$prediction <- predict(poisson_100, type="response")

## order by program and then by math
dataset_pois_100 <- dataset_pois_100[with(dataset_pois_100,
order(AgeClass, TotalKg)), ]

## create the plot
ggplot(dataset_pois_100, aes(x = TotalKg, y = prediction, colour =
AgeClass)) +
  geom_point(aes(y = FirstPlace), alpha=.5,
position=position_jitter(h=.2)) +
  geom_line(size = 1) +
  labs(x = "TotalKg", y = "Nombre de primers llocs esperat")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2
3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning
was
## generated.
```



Per la visualització es pot veure com les prediccions a mesura que augmenta la variable TotalKg hi ha més nombre de primers llocs esperats. Respecte les edats la llargada de les corves ve donada per els valors de TotalKg amb els que ha creat el model i podem veure com la corva més allargada és la del rang de a 24-34 anys la qual té sentit perquè és la que hi ha més quantitat de dades i és més possible que estiguin separats entre ells. A més, també veiem que quan la variable TotalKg és menor és en la corva vermella que representa el rang de menor edat i és obvi perquè fins una edat l'atleta encara no s'ha desenvolupat totalment i no pot fer aixecaments de tants Kg.

Conlcusions

1. Els coeficients del model de 50 atletes no eren estadísticamet significatius degut a l'escasetat de dades i val més la pena fer-ho amb 100 i treure millors prediccions on no aporten tant de soroll els outliers,
2. Gràcies a la matriu de variancies i covariancies es poden trobar els errors estàndards de les prediccions i demostrar com de significatives son apartir dels intervals de confiança i el P-valor.
3. Respecte la gràfica, a mesura que els TotalKg augmenten el nombre de primer llocs d'aquell rang d'edat també ho fa. També podem veure la importància de la regressió Poisson veient que al llarg de que els rangs d'edats augmenta, el nombre de primers llocs també.

Bibliografia

https://www.datacamp.com/tutorial/multiple-linear-regression-r-tutorial

https://www.statmethods.net/stats/regression.html

https://stats.oarc.ucla.edu/r/dae/logit-regression/

https://www.youtube.com/watch?v=C4N3_XJJ-jU

https://www.youtube.com/watch?v=yIYKR4sgzI8&list=PLblh5JKOoLUKxzEP5HA2d-Li7IJkHfXSe

https://bookdown.org/roback/bookdown-BeyondMLR/ch-poissonreg.html

https://stats.oarc.ucla.edu/r/dae/poisson-regression/

https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/