Proyecto Estadística Descriptiva

Fase 2

Integrantes:

C-412 Richard García De la Osa. richard.garcia@estudiantes.matcom.uh.cu

C-412 Andy A. Castañeda Guerra. andy.castaneda@estudiantes.matcom.uh.cu

C-411 Adrián Hernández Pérez. a.hernandez3@estudiantes.matcom.uh.cu

Github repo

Abstract:

Un breve análisis estadístico sobre los jugadores del draft de la NBA del año 2014.

Partiendo de los datos de las mediciones físicas de los jugadores del draft de la NBA, se realizaron varios análisis sobre los mismos, en aras de caracterizar y hallar relaciones entre las mismas. Se consideraron técnicas de reducción de dimensión, análisis de varianza y correlación entre los datos.

Introducción

La ***National Basketball Association***(NBA) cuenta cada año con discímiles aspirantes a unirse a uno de los equipos y probar su capacidad y talento a nivel profesional. Para ello se realiza anualmente hace más de medio siglo un evento conocido como el *NBA draft pick*, donde dichos equipos escogen a qué jugadores incorporar a sus bancas de todos aquellos posibles candidatos. El análisis que proponemos a continuación está basado en los datos recogidos de dicho *draft pick* en el año 2014. Los datos consisten en las aptitudes físicas cuantificadas para cada uno de los candidatos considerados, por ejemplo su altura, su peso y grasa corporal, altura de salto, envergadura de las extremidades, entre otros. Primeramente haremos un análisis de la relación entre las variables con las que se trabaja mediante regresión lineal múltiple. También incurrimos en un análisis de varianza(ANOVA) entre los datos recogidos en el 2012, 2014 y 2015, para contrastar las variaciones de las capacidades de los atletas. Además, se hará una clasificación de los jugadores por categorías para reducir la dimensión de los datos.

Regresión

Partiendo de los datos del draft de los jugadores de la NBA del año 2014, se desea hacer el análisis de regresión sobre los datos de estudio.

La variable a la que le aplicaremos esta técnica será Vertical (Max Reach), pues consideramos que es una de las más importantes entre los datos con los que se trabaja. El objetivo es encontrar un modelo el cual represente la mejor combinación para explicar el comportamiento de la variable dependiente.

Tabla de correlación de las variables.

	Height..No.Shoes.	Height..With.Shoes.	Wingspan	Standing.reach	Vertical..Max.	Vertical..Max.Reach.	Vertical..No.Step.	Vertical..No.Step.Reach.	Weight	Body.Fat	Hand..Length.	Hand..Width.	Agility	Sprint
Height..No.Shoes.	1.0000000	0.9967596	0.8594758	0.9211098	-0.3487695	0.63919274	-0.22004461	0.76440865	0.7536527	0.21199550	0.6545991	0.37891929	0.19204525	0.32743329
Height..With.Shoes.	0.9967596	1.0000000	0.8564495	0.9152647	-0.3479920	0.64679720	-0.22203273	0.77011342	0.7512308	0.19532901	0.6655314	0.38498203	0.18353903	0.31319118
Wingspan	0.8594758	0.8564495	1.0000000	0.9342444	-0.2486753	0.68669549	-0.14660877	0.78006960	0.7288458	0.25824392	0.7693679	0.45867688	0.27289104	0.28263476
Standing.reach	0.9211098	0.9152647	0.9342444	1.0000000	-0.3778135	0.64960162	-0.29413267	0.75171561	0.6902313	0.27867157	0.6584751	0.34853947	0.20816116	0.34441650
Vertical..Max.	-0.3487695	-0.3479920	-0.2486753	-0.3778135	1.0000000	0.36950865	0.82901627	0.10632088	-0.2643241	-0.37753017	-0.2193689	0.13446914	-0.11936538	-0.65826806
Vertical..Max.Reach.	0.6391927	0.6467972	0.6866955	0.6496016	0.3695086	1.00000000	0.32804923	0.90533084	0.4909208	-0.04001031	0.5109241	0.49386099	0.13440511	-0.13703061
Vertical..No.Step.	-0.2200446	-0.2220327	-0.1466088	-0.2941327	0.8290163	0.32804923	1.00000000	0.30348279	-0.1350599	-0.37833373	-0.1037953	0.24736164	-0.02516673	-0.60533227
Vertical..No.Step.Reach.	0.7644087	0.7701134	0.7800696	0.7517156	0.1063209	0.90533084	0.30348279	1.00000000	0.6059673	0.01787001	0.6098301	0.53713703	0.20932490	-0.00219432
Weight	0.7536527	0.7512308	0.7288458	0.6902313	-0.2643241	0.49092076	-0.13505987	0.60596729	1.0000000	0.52014832	0.7100076	0.45196160	0.19218759	0.29952791
Body.Fat	0.2119955	0.1953290	0.2582439	0.2786716	-0.3775382	-0.04001031	-0.37833373	0.01787001	0.5201483	1.00000000	0.1341388	-0.04379619	0.06366859	0.38711048
Hand..Length.	0.6545991	0.6655314	0.7693679	0.6584751	-0.2193689	0.51092409	-0.10379525	0.60983012	0.7100076	0.13413884	1.0000000	0.59013565	0.24880902	0.24219838
Hand..Width.	0.3789193	0.3849820	0.4586769	0.3485395	0.1344691	0.49386099	0.24736164	0.53713703	0.4519616	-0.04379619	0.5901356	1.00000000	0.17010569	0.02263921
Agility	0.1920453	0.1835390	0.2728910	0.2081612	-0.1193654	0.13440511	-0.02516673	0.20932490	0.1921876	0.06366859	0.2488090	0.17010569	1.00000000	0.13351775
Sprint	0.3274333	0.3131912	0.2826348	0.3444165	-0.6582681	-0.13703061	-0.60533227	-0.00219432	0.2995279	0.38711048	0.2421984	-0.02263921	0.13351775	1.00000000

Dado que se aprecia que existen valores bastante altos en algunos de los pares de variables, podemos decir que existe dependencia lineal entre estas variables.

Modelo:

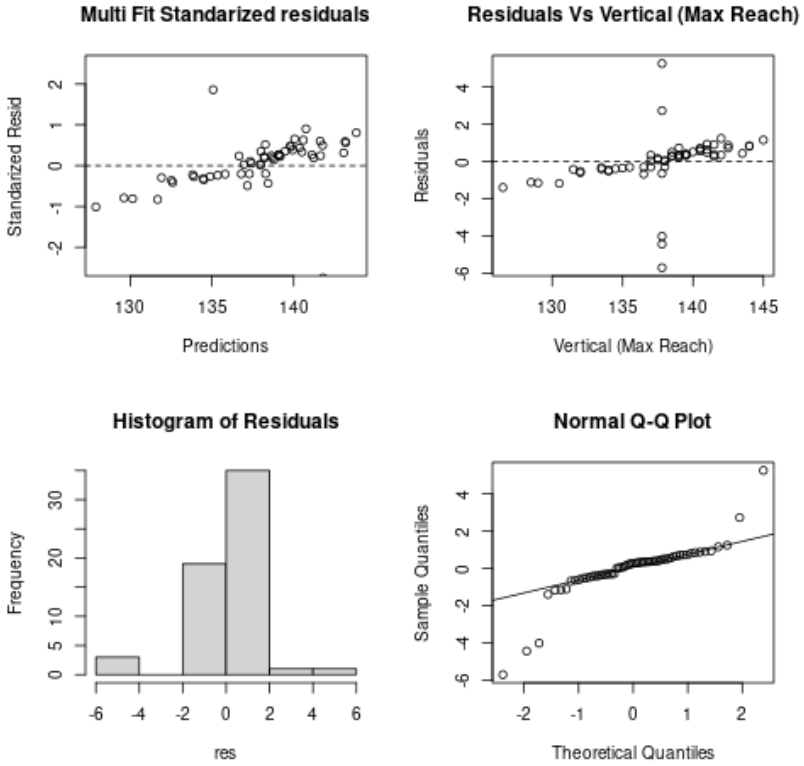
$$Vertical\ (Max\ Reach) = \beta_0 + \beta_1 Standing\ reach + \beta_2 Vertical\ (Max) + e.$$

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.6919 -0.3929  0.2808  0.5386  5.2541

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       18.98752    6.33591   2.997  0.00406 **
data$Standing.reach  0.84200    0.04884  17.239  < 2e-16 ***
data$Vertical..Max.  0.91670    0.06824  13.433  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.488 on 56 degrees of freedom
Multiple R-squared:  0.8631,    Adjusted R-squared:  0.8582
F-statistic: 176.5 on 2 and 56 DF,  p-value: < 2.2e-16
```

Podemos ver que el valor del intercepto es alto, esto indica que gran parte del Vertical (Max Reach) no está muy bien explicada a partir de las variables independientes, lo cual no es deseable. Además notamos que los valores de Pr(>|t|) para todas son menores que 0.05.

De estos valores podemos ver que el R-squared es 0.8582, que está por encima de 0.70 y como el p-value es menor que 0.05 todo parece indicar que nuestro modelo está bien. Dado que el valor del intercepto es alto se podría decir que necesitamos considerar otros factores en el análisis.

Analizando los Residuos:




```
> mean(multi.fit$residuals)
[1] -7.573979e-17
> sum(multi.fit$residuals)
[1] -4.468648e-15
```

La media de los errores es cero y la suma de los errores es cero.

```
> shapiro.test(multi.fit$residuals)

        Shapiro-Wilk normality test

data:  multi.fit$residuals
W = 0.76937, p-value = 3.045e-08
```

El valor de p-value es 3.045e-08 < 0.05 por lo que no podemos decir que los errores siguen una distribución normal(Falla).

```
> dwtest(multi.fit)

        Durbin-Watson test

data:  multi.fit
DW = 2.2893, p-value = 0.8654
alternative hypothesis: true autocorrelation is greater than 0
```

El p-value es 0.8654 > 0.05 por lo que podemos afirmar que los errores son independientes.

```
> bptest(multi.fit)

        studentized Breusch-Pagan test

data:  multi.fit
BP = 0.97391, df = 2, p-value = 0.6145
```

Como el p-value es 0.6145 > 0.05 no podemos rechazar la heterocedasticidad. Por lo que el supuesto de Homocedasticidad se mantiene.

No se cumplen todos los supuestos del modelo.

Modelo:

$Vertical\ (Max\ Reach) = \beta_0 + \beta_1 Height\ (No\ Shoes) + \beta_2 Height..(With\ Shoes) + \beta_3 Standing\ reach + \beta_4 Vertical\ (Max)$

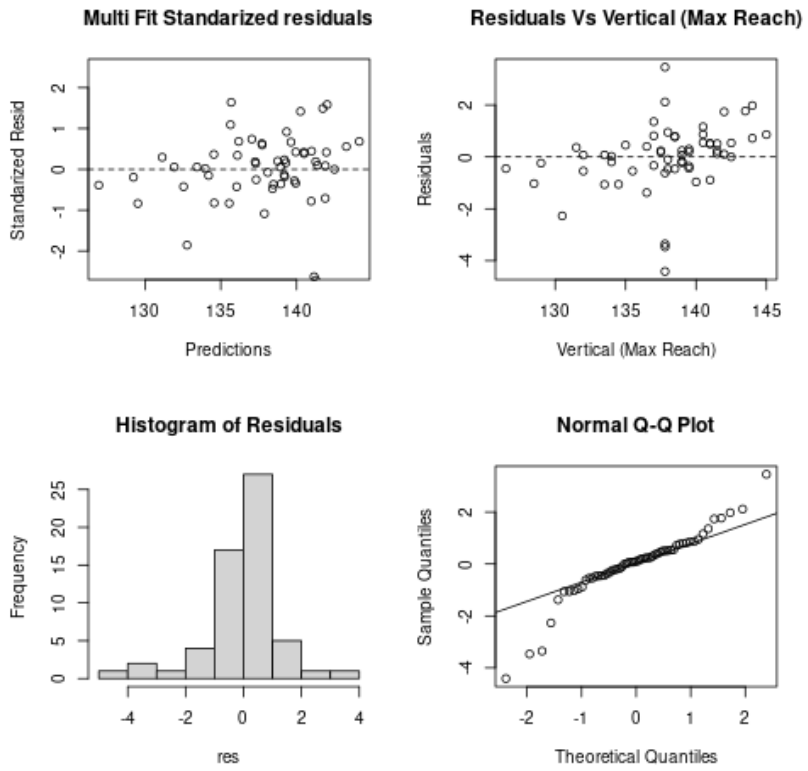
$+\beta_5 Vertical\ (No\ Step) + e.$

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.4142 -0.4556  0.1043  0.5447  3.4573

Coefficients:
mean(multi.fit$residuals)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            7.28071    6.43046   1.132  0.26264
data$Height..No.Shoes.  -1.95544    0.70891  -2.758  0.00795 **
data$Height..With.Shoes.  2.33245    0.70217   3.322  0.00163 **
data$Standing.reach     0.64127    0.10800   5.937 2.28e-07 ***
data$Vertical..Max.     0.93528    0.10659   8.775 6.61e-12 ***
data$Vertical..No.Step.  -0.02138    0.13006  -0.164  0.87005
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.331 on 53 degrees of freedom
Multiple R-squared:  0.8963,    Adjusted R-squared:  0.8865
F-statistic: 91.63 on 5 and 53 DF,  p-value: < 2.2e-16
```

Podemos apreciar que el valor del intercepto es alto, esto indica que parte del Vertical (Max Reach) no está bien explicada a partir de las variables independientes y esto no es lo ideal. Podemos notar que los valores de Pr(>|t|) para algunos son menores que 0.05 y para otros mayores.

En esta ocación tenemos el R-squared con valor 0.8865, este es mayor que 0.70 y como el p-value < 0.05 parece indicar que nuestro modelo está bien. Dado que el valor del intercepto es alto podríamos decir que necesitamos considerar otros factores para este análisis.

Analizando los Residuos:




```
> mean(multi.fit$residuals)
[1] -2.070114e-17
> sum(multi.fit$residuals)
[1] -1.221245e-15
```

La media de los errores es cero y la suma de los errores es cero.

```
> shapiro.test(multi.fit$residuals)

        Shapiro-Wilk normality test

data:  multi.fit$residuals
W = 0.90584, p-value = 0.0002441
```

El valor de p-value es 0.0002441 < 0.05 por lo que no podemos decir que los errores siguen una distribución normal(Falla).

```
> dwtest(multi.fit)

        Durbin-Watson test

data:  multi.fit
DW = 2.0875, p-value = 0.6313
alternative hypothesis: true autocorrelation is greater than 0
```

El p-value es 0.6313 > 0.05 por lo que podemos afirmar que los errores son independientes.

```
> bptest(multi.fit)

        studentized Breusch-Pagan test

data:  multi.fit
BP = 3.5969, df = 5, p-value = 0.6088
```

Como el p-value es 0.6088 > 0.05 no podemos rechazar la heterocedasticidad. Por lo que el supuesto de Homocedasticidad se mantiene.

No se cumplen todos los supuestos del modelo.

Modelo:

$$Vertical\ (Max\ Reach) = \beta_0 + \beta_1 Height\ (No\ Shoes) + \beta_2 Height..(With\ Shoes) + \beta_3 Standing\ reach + \beta_4 Vertical\ (Max)$$

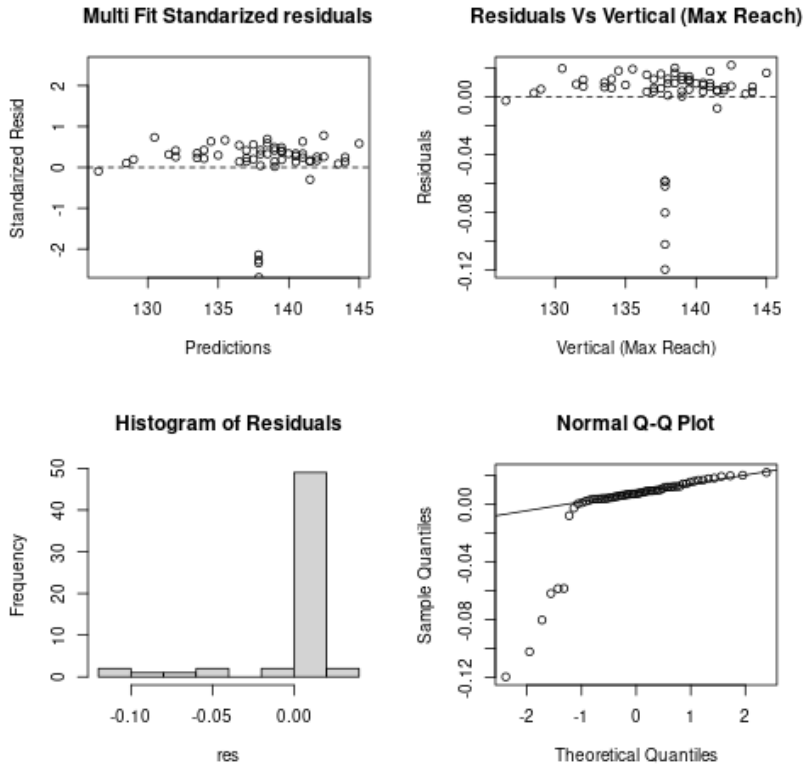
$$+\beta_5 Vertical\ (No\ Step) + \beta_6 Vertical\ (No\ Step\ Reach) + e.$$

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.119516  0.003864  0.007401  0.012190  0.022115

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 0.052442   0.148581    0.353   0.7256
data$Height..No.Shoes.      0.024701   0.017339    1.425   0.1603
data$Height..With.Shoes.   -0.023091   0.017655   -1.308   0.1967
data$Standing.reach        -0.007731   0.003198   -2.417   0.0192 *
data$Vertical..Max.         1.000364   0.002443  409.564   <2e-16 ***
data$Vertical..No.Step.    -1.006692   0.004287 -234.830   <2e-16 ***
data$Vertical..No.Step.Reach.  1.006227   0.003157  318.730   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03039 on 52 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 1.634e+05 on 6 and 52 DF,  p-value: < 2.2e-16
```

Podemos ver que el valor del intercepto es bastante bajo, esto indica que gran parte del Vertical (Max Reach) está bien explicado a partir de las variables independientes que empleamos, esto resulta ideal. Además notamos que los valores de Pr(>|t|) para algunos son menores que 0.05 y para otros mayores.

Tenemos que el R-squared tiene valor 0.9999, este es mayor que 0.70 y es casi 1. A lo que podemos añadir que el p-value < 0.05, lo que nos deja claro que nuestro modelo es muy bueno.

Analizando los Residuos:




```
> mean(multi.fit$residuals)
[1] 1.06037e-19
> sum(multi.fit$residuals)
[1] 6.288373e-18
```

La media de los errores es cero y la suma de los errores es cero.

```
> shapiro.test(multi.fit$residuals)

        Shapiro-Wilk normality test

data:  multi.fit$residuals
W = 0.55483, p-value = 4.178e-12
```

El valor de p-value es 4.178e-12 < 0.05 por lo que no podemos decir que los errores siguen una distribución normal(Falla).

```
> dwtest(multi.fit)

        Durbin-Watson test

data:  multi.fit
DW = 2.2232, p-value = 0.8062
alternative hypothesis: true autocorrelation is greater than 0
```

El p-value es 0.8062 > 0.05 por lo que podemos afirmar que los errores son independientes.

```
> bptest(multi.fit)

        studentized Breusch-Pagan test

data:  multi.fit
BP = 7.5075, df = 6, p-value = 0.2764
```

Como el p-value es 0.2764 > 0.05 no podemos rechazar la heterocedasticidad. Por lo que el supuesto de Homocedasticidad se mantiene.

No se cumplen todos los supuestos del modelo.

Modelo:

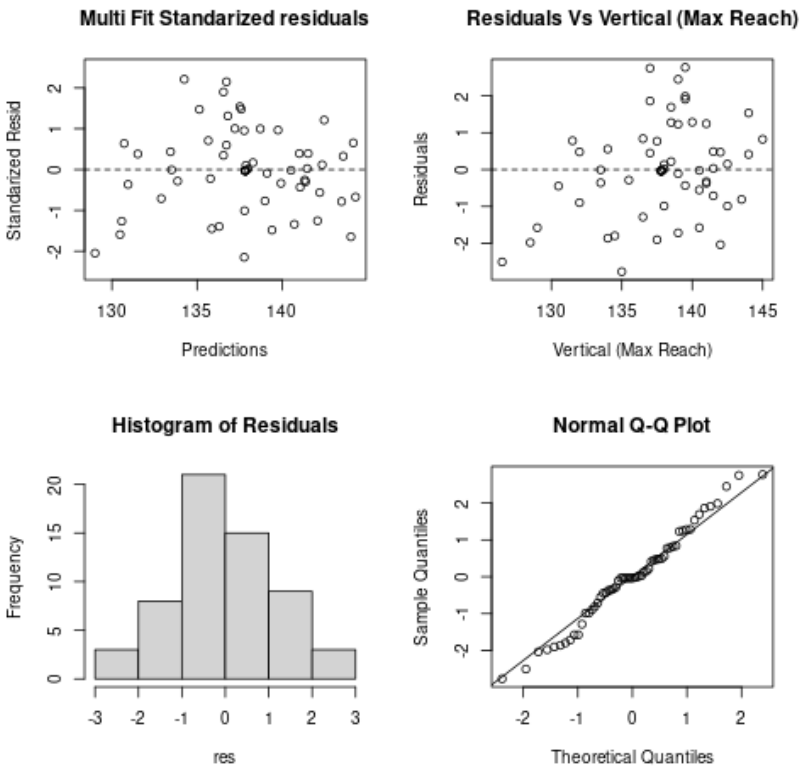
$$Vertical\ (Max\ Reach) = \beta_0 + \beta_1 Vertical(No\ Step\ Reach) + \beta_2 Vertical\ (Max) + e.$$

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.7711 -0.7617 -0.0381  0.7762  2.7705

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 10.88442    5.81857   1.871   0.0666 .
data$Vertical..No.Step.Reach.  0.86494    0.04297  20.128  < 2e-16 ***
data$Vertical..Max.            0.35319    0.05561   6.351 4.06e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.302 on 56 degrees of freedom
Multiple R-squared:  0.8951,    Adjusted R-squared:  0.8914
F-statistic:   239 on 2 and 56 DF,  p-value: < 2.2e-16
```

Notamos que el valor del intercepto no es muy bajo, esto que indica que gran parte del Vertical (Max Reach) puede no estar bien explicada a partir de las variables independientes elegidas, esto no es lo ideal. Además notamos que los valores de Pr(>|t|) para unos es menor que 0.05 y para el otro es mayor.

Obtenemos un R-squared de valor 0.8914 que está por encima de 0.70. También podemos notar que el p-value es menor que 0.05 lo que indica que nuestro modelo es bueno.

Analizando los Residuos:




```
> mean(multi.fit$residuals)
[1] 6.066433e-17
> sum(multi.fit$residuals)
[1] 3.580469e-15
```

La media de los errores es cero y la suma de los errores es cero.

```
> shapiro.test(multi.fit$residuals)

        Shapiro-Wilk normality test

data:  multi.fit$residuals
W = 0.98196, p-value = 0.5286
```

El valor de p-value es 0.5286 > 0.05 por lo que podemos decir que los errores siguen una distribución normal.

```
> dwtest(multi.fit)

        Durbin-Watson test

data:  multi.fit
DW = 2.1121, p-value = 0.6693
alternative hypothesis: true autocorrelation is greater than 0
```

El p-value es 0.6693 > 0.05 por lo que podemos afirmar que los errores son independientes.

```
> bptest(multi.fit)

        studentized Breusch-Pagan test

data:  multi.fit
BP = 3.5369, df = 2, p-value = 0.1706
```

Como el p-value es 0.1706 > 0.05 no podemos rechazar la heterocedasticidad. Por lo que el supuesto de Homocedasticidad se mantiene.

Se cumplen todos los supuestos del modelo.

Modelos analizados:

$-Vertical\ (Max\ Reach) = \beta_0 + \beta_1 Standing\ reach + \beta_2 Vertical\ (Max) + e$. No cumple los supuestos, Adjusted R-squared: 0.8582.

$-Vertical\ (Max\ Reach) = \beta_0 + \beta_1 Height\ (No\ Shoes) + \beta_2 Height\ (With\ Shoes) + \beta_3 Standing\ reach + \beta_4 Vertical\ (Max)$

$+\beta_5 Vertical\ (No\ Step) + e$. No cumple los supuestos, Adjusted R-squared: 0.8865.

$-Vertical\ (Max\ Reach) = \beta_0 + \beta_1 Height\ (No\ Shoes) + \beta_2 Height..(With\ Shoes) + \beta_3 Standing\ reach + \beta_4 Vertical\ (Max)$

$+\beta_5 Vertical\ (No\ Step) + \beta_6 Vertical\ (No\ Step\ Reach) + e$. No se cumplen los supuestos,
Adjusted R-squared: 0.9999.

$-Vertical\ (Max\ Reach) = \beta_0 + \beta_1 Vertical(No\ Step\ Reach) + \beta_2 Vertical\ (Max) + e$. Se cumplen los supuestos,
Adjusted R-squared: 0.8914.

El modelo con mejor R-squared ajustado es el de:

$Vertical\ (Max\ Reach) = \beta_0 + \beta_1 Vertical(No\ Step\ Reach) + \beta_2 Vertical\ (Max) + e.$

Quedaría $Vertical\ Max\ Reach) = 10.88442 + 0.86494 * Vertical(No\ Step\ Reach) + 0.35319 * Vertical\ (Max).$

Reducción de dimensión

La técnica de reducción de dimesión empleada para clasificar los datos de los jugadores de la NBA en el año 2014 fue la **ténica jerárquica** de **clústers**.

Se consideraron todas aquellas mediciones cuantitativas que los jugadores tenían, además se omitió el análisis de las columnas vacías o parcialmente incompletas. Posteriormente, se ejecutó el algoritmo k-means con distintos valores para la cantidad de clústers creados, comenzando en 2, y analizando hasta para 5 clústers.

De todos los resultados obtenidos, elegimos el más razonable, tanto por el porcentaje de similitud de los jugadores que pertenecían a las mismas categorías, como para claridad en cuanto a la interpretación que se le puede dar a los posibles *outliers* que detectamos. A continuación, la tabla:

Cat	Height (NoShoes)	Height (WithShoes)	Wingspan	Standing Reach	Vertical (Max)	Vertical (MaxReach)	Vertical (NoStep)
1	80.4	81.72	86.16	106.27	34.50	140.77	29.72
2	74.3	75.78	79.57	99	36.47	135.47	30.60
3	78.4	79.80	82.88	103.73	35.21	138.95	29.59

Cat	Vertical (NoStepReach)	Weight	Body Fat	Hand Length	Hand Width	Agility	Sprint
1	136.00	251.22	9.70	9.19	9.86	11.58	3.37
2	129.60	188.63	5.71	8.42	8.93	11.33	3.26
3	133.33	218.52	7.21	8.82	9.11	11.34	3.31

Con un porcentaje de similitud del 79.0%

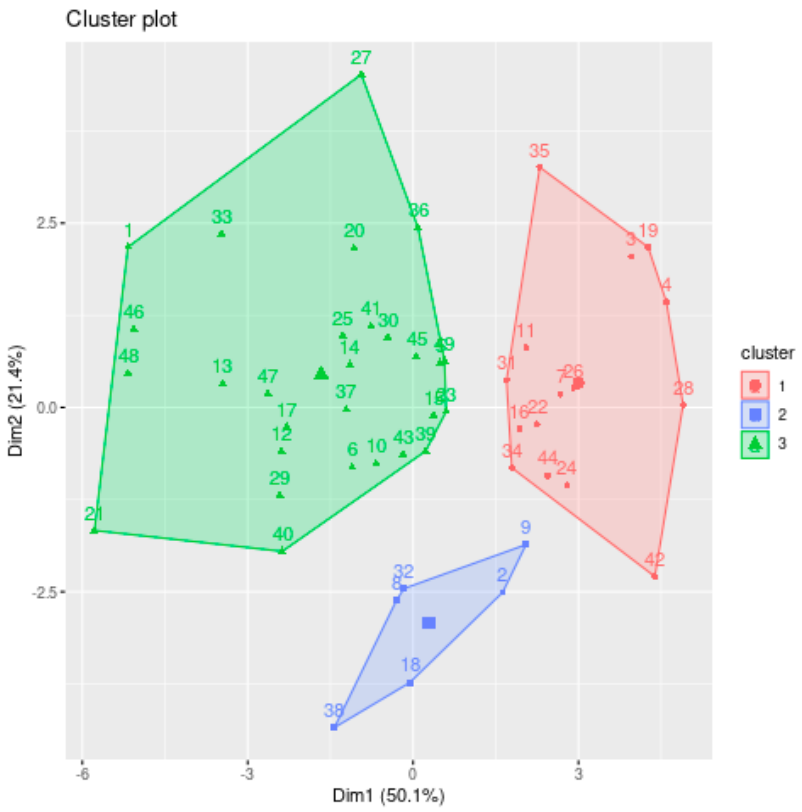
Acá se observa la clara división entre los jugadores en tres categorías. La categoría **No.1**, se trata de los jugadores más corpulentos en genral, aquellos de mayor altura, mayor envergadura de los brazos, superior alcance estando de pie, mayor peso, manos más extensas, y mayor grasa corporal. La contraparte de estos, serían los jugadores pertenecientes a la categoría **No.2**, comprendiendo aquellos más pequeños(aun así son considerablemente más altos que cualquier persona promedio), de menor extensión entre brazo y brazo, menos pesados y de menor grasa corporal, además con las manos notablemente más pequeñas, y a la vez menor alcance vertical estando de pie. Finalmente en la categoría **No.3** restante, quedarían aquellos jugadores "promedio", cuyas mediciones son intermedias, su altura no es suficiente como para estar considerados entre los más altos, pero tampoco es tan pequeña como para caer en la categoría **No.2**, igualmente con su peso, su alcance vertical sin saltar, la evergadura de sus brazos, etc.

De estos datos, cabe resaltar algunas observaciones suficientemente llamativas:

- En primer lugar, los jugadores de menor estatura son aquellos que mayor alcance vertical(Vertical Max) poseen al saltar sin contar la altura extra que le confieren sus brazos, es decir, que al saltar, son los que más elevan sus pies del suelo. Aún así, esto no es suficiente como para ser los de mayor alcance vertical al saltar si se mide la altura a

la que se extienden sus brazos(Vertical Max Reach), siendo nuevamente los más modestos en este campo. Esto puede resultar intuitivo si tenemos en cuenta que estos jugadores son aquellos de menor masa y grasa corporal.

- En segundo lugar, las medidas de agilidad y velocidad de los jugadores más corpulentos y de mayor peso y grasa corporal son ligeramente superiores a las de los demás jugadores. Uno esperaría que dado que son más pesados, quizás sean también más "torpes" o lentos, sin embargo, estas medidas en los datos estudiados se mantienen bastante similares para todos los jugadores estudiados, y curiosamente, los jugadores más altos son los más ágiles y veloces, mientras que los más chicos son los menos ágiles y más lentos.



En la figura se observa la representación gráfica de los datos estandarizados, por categorías, reflejándose que la categoría **1**, de los jugadores corpulentos, es menor en tamaño que la de los jugadores tamaño promedio pertenecientes a la **3**, y que los jugadores más pequeños son realmente la minoría de entre la población analizada, siendo el reducido grupo de la categoría **2**. Esto coincide con los resultados que obtuvimos al escribir en la consola la cardinalidad de cada uno de los conjuntos disjuntos creados, donde la categoría **3** comprendía 25 jugadores, la **2** con 21 jugadores un poco por detrás, y finalmente la categoría **3** con 12 jugadores solamente.

ANOVA

Partiendo de los datos del draft de los jugadores de la NBA del año 2014, se desea hacer un análisis de varianza sobre los datos de estudio. Para ello, el planteamiento fue el siguiente:

Dado que en un inicio solamente se tenían datos para un año específico, teníamos dos opciones, o considerar la varianza de los datos para cada jugador(pero para cada jugador solamente teniamos un único dato, de manera que esto carecía de sentido), o buscar más datos de años anteriores y/o posteriores, y analizar las varianzas de ciertas características(altura, peso, agilidad, etc) por años.

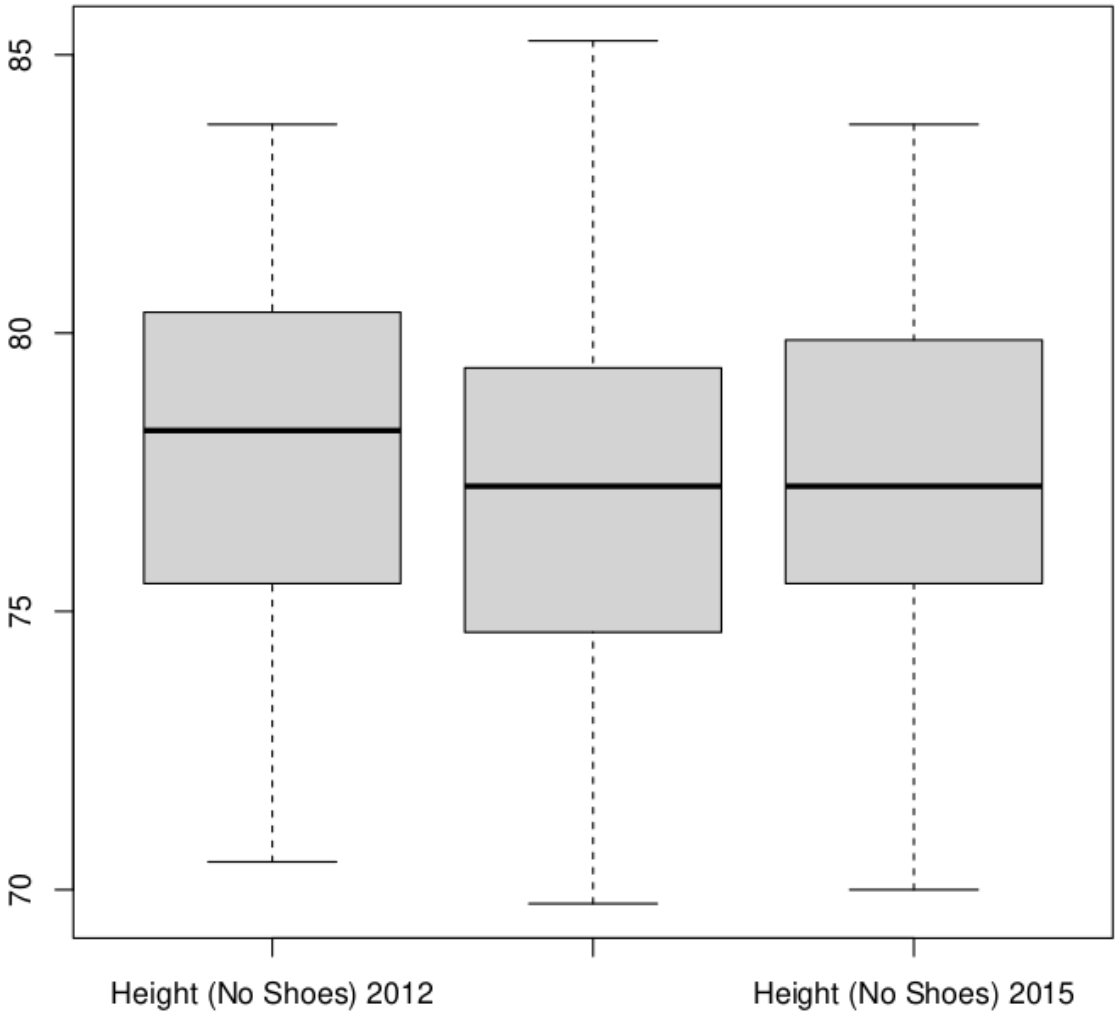
Height (No Shoes)

De modo que, se desea saber si existen diferencias entre las alturas de los jugadores de la nba de los años 2012, **2014** y 2015

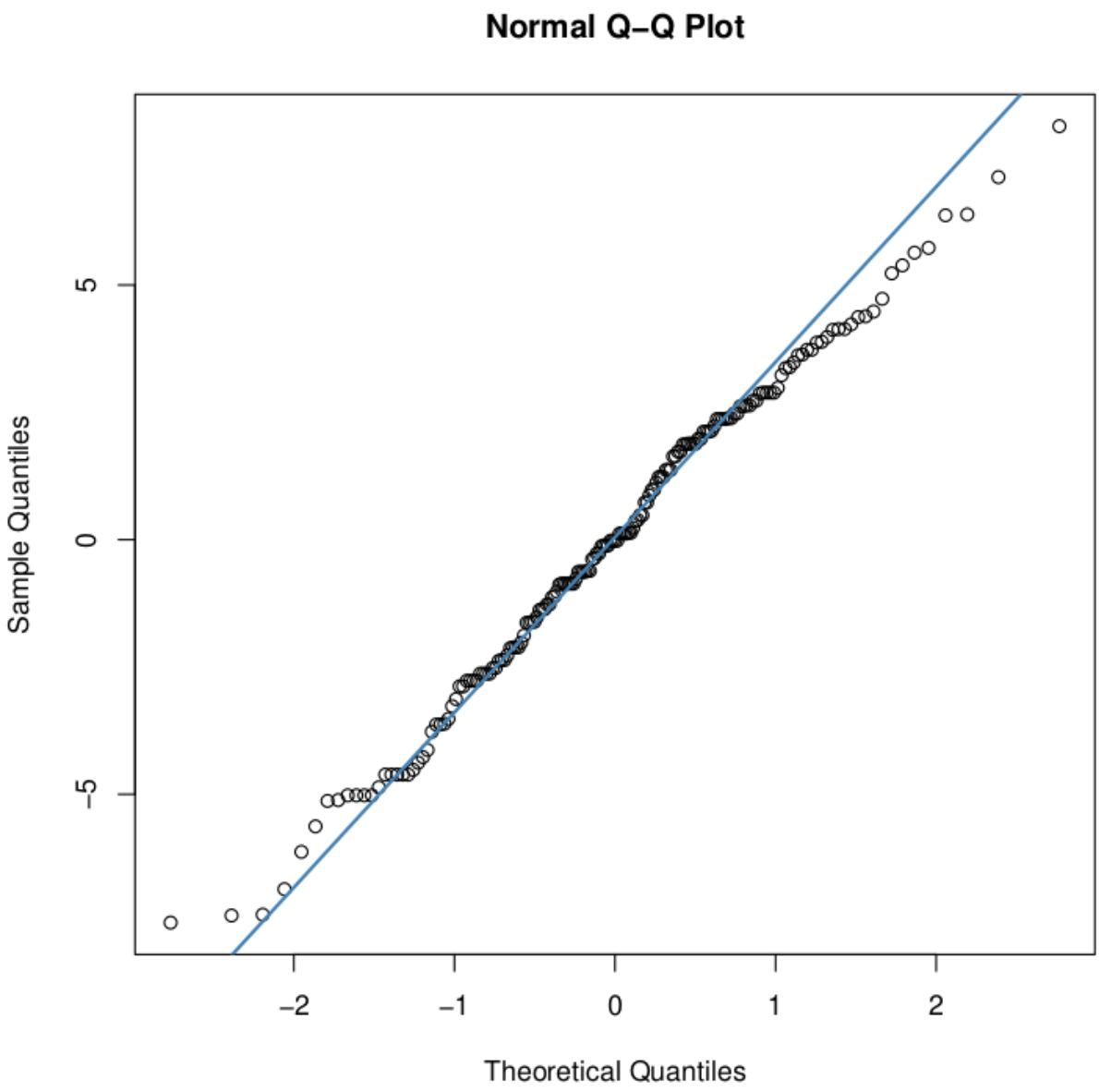
$H_0 : \mu_{2012} = \mu_{2014} = \mu_{2015}$

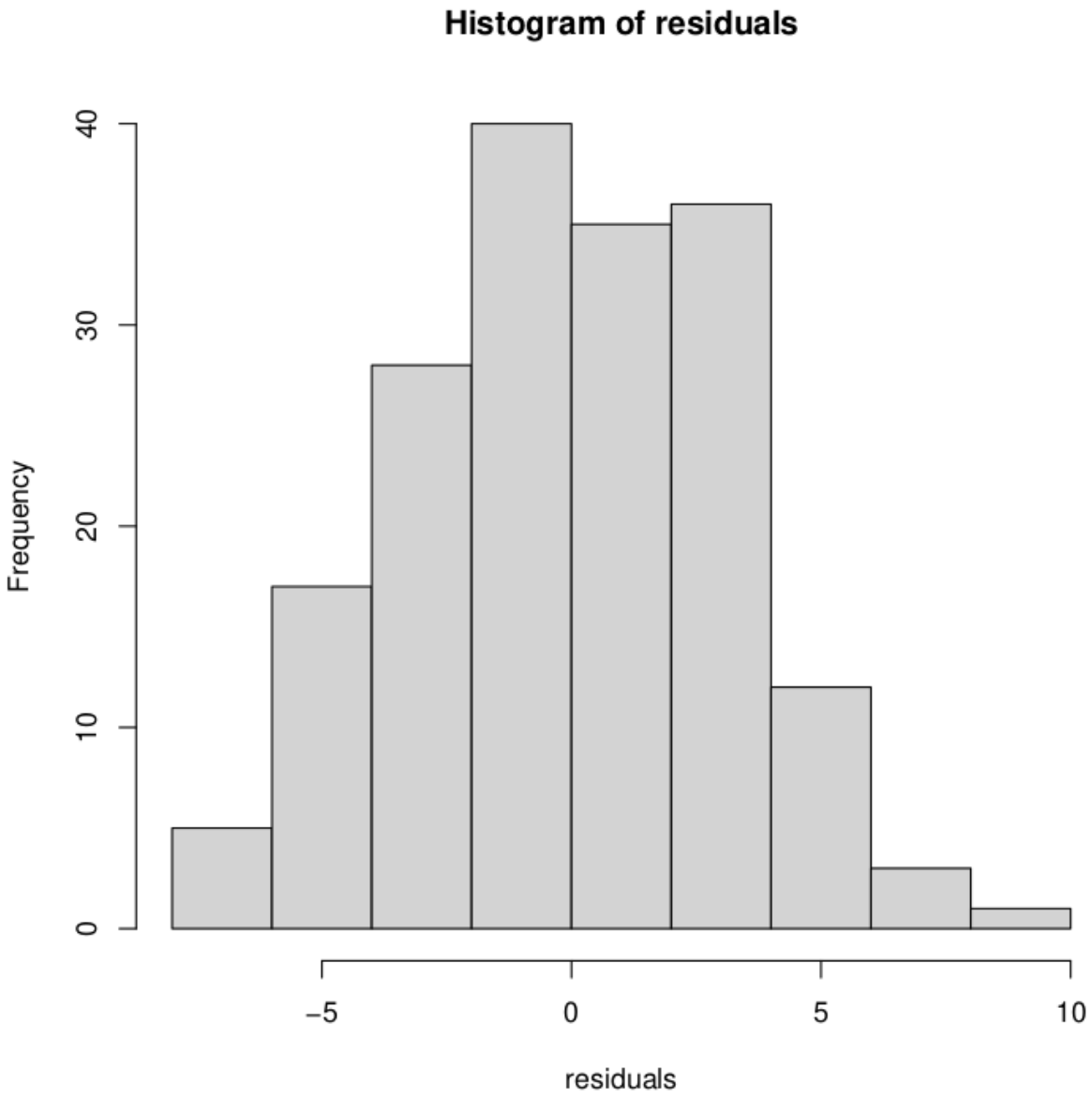
$H_a : \mu_i \neq \mu_j, \ \ i \neq j$

Apoyándonos de la representación gráfica de las medias de las alturas, tenemos:



A simple vista, parecen ser considerablemente parecidos los datos, de modo que, según cual sea la tolerancia con la cual trabajemos, se rechazará o no nuestra hipótesis nula. Al realizar el **análisis de varianza**, tenemos que el p-value 0.28 > 0.05, aceptando entonces H_0, la altura de los jugadores de la nba en los años 2012, 2014 y 2015 son muy similares.


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
5
0
−5
−2
−1
0
1
2



Verificando finalmente los supuestos del modelo realizando las pruebas de hipótesis de Shapiro, Durbin-Watson y Bartlett respectivamente:

1) los e_{ij} siguen una distribución normal: p-value = 0.43 $\gg$ 0.05, podemos decir entonces que están normalmente distribuidos.

2) los e_{ij} son independientes: p-value = 0.14 > 0.05, por lo que los errores son independientes.

3) los e_{ij} tienen la misma varianza: p-value = 0.92 $\gg$ 0.05, las varianzas de los residuos son homogéneas.

Vertical (Max Reach)

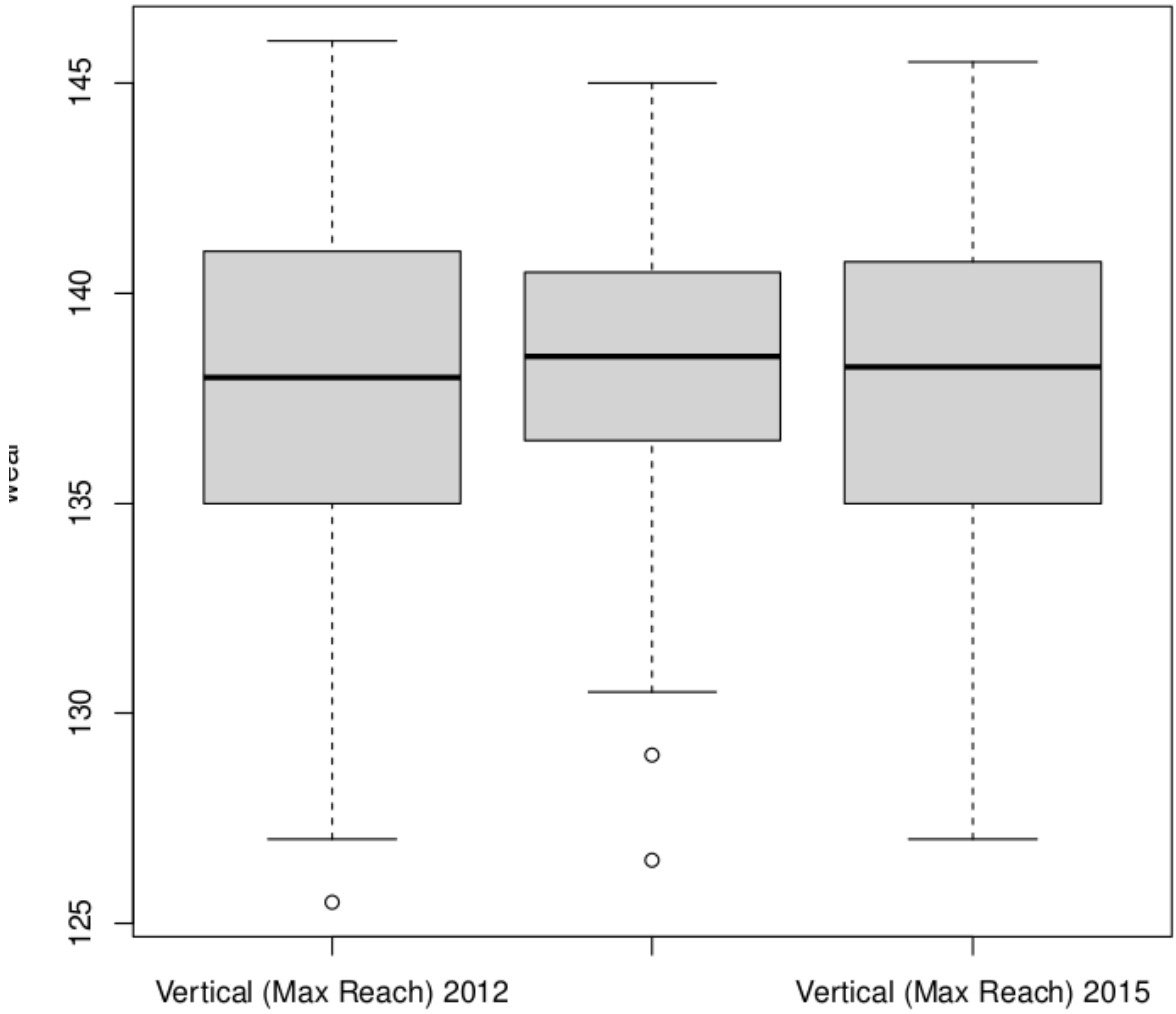
A continuación, analizamos el alcance vertical máximo.

Se desea entonces saber si existen diferencias entre el alcance máximo de los jugadores de la nba en los años 2012, **2014** y 2015

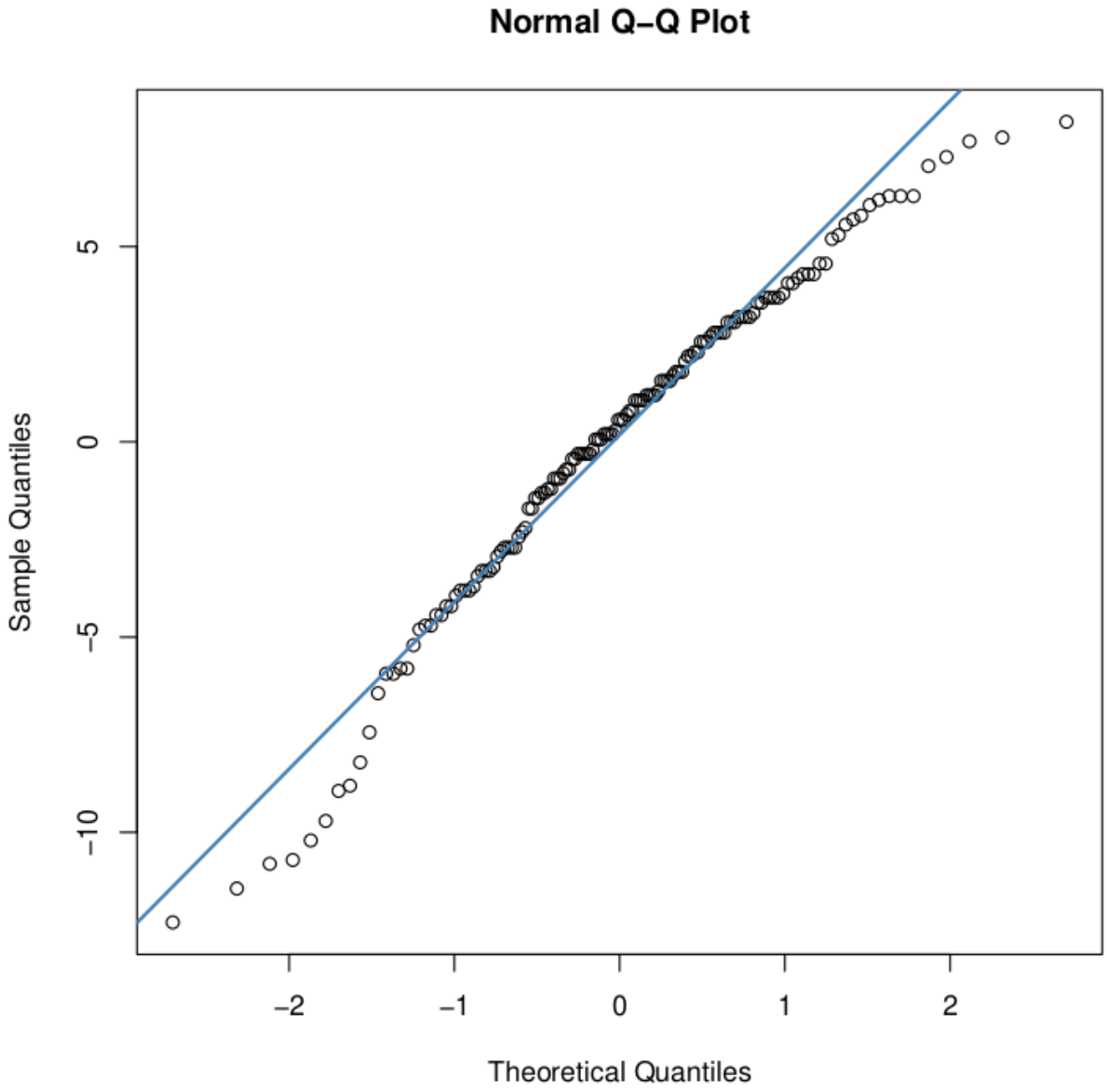
$H_0 : \mu_{2012} = \mu_{2014} = \mu_{2015}$

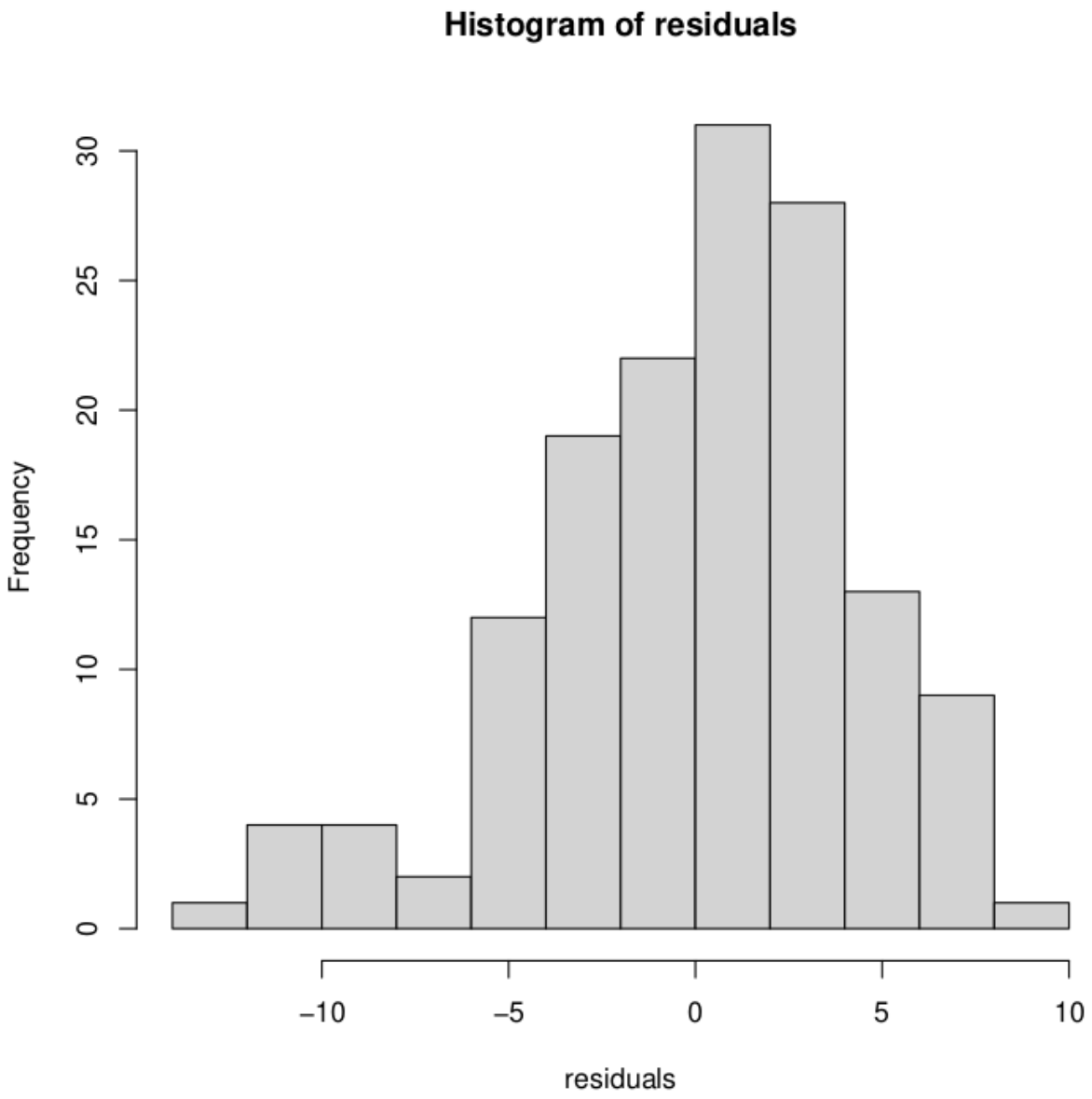
$H_a : \mu_i \neq \mu_j, \ \ i \neq j$

Al graficar los resultados podemos observar:



De manera parecida al caso anterior, los datos se asemejan bastante. Luego de hacer el **análisis de varianza**, tenemos que el p-value 0.96 $\gg$ 0.05, aceptando entonces H_0, el alcance vertical máximo de los jugadores de la nba en los años 2012, 2014 y 2015 es casi idéntico.


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
5
0
−5
−10
−2
−1
0
1
2



Al observar las gráficas anteriores, vemos que la normalidad de los residuos en esta ocasión es bastante discutible, aún así, realizaremos las tres pruebas de hipótesis como ya es costumbre(aunque baste con que una de ellas falle):

1) los e_{ij} **no** siguen una distribución normal: p-value = 0.003 $\ll$ 0.05, luego los residuos no están normalmente distribuidos.

2) los e_{ij} son independientes: p-value = 0.49 $\gg$ 0.05, los errores son independientes.

3) los e_{ij} tienen la misma varianza: p-value = 0.48 $\gg$ 0.05, las varianzas de los residuos son homogéneas.

Weight

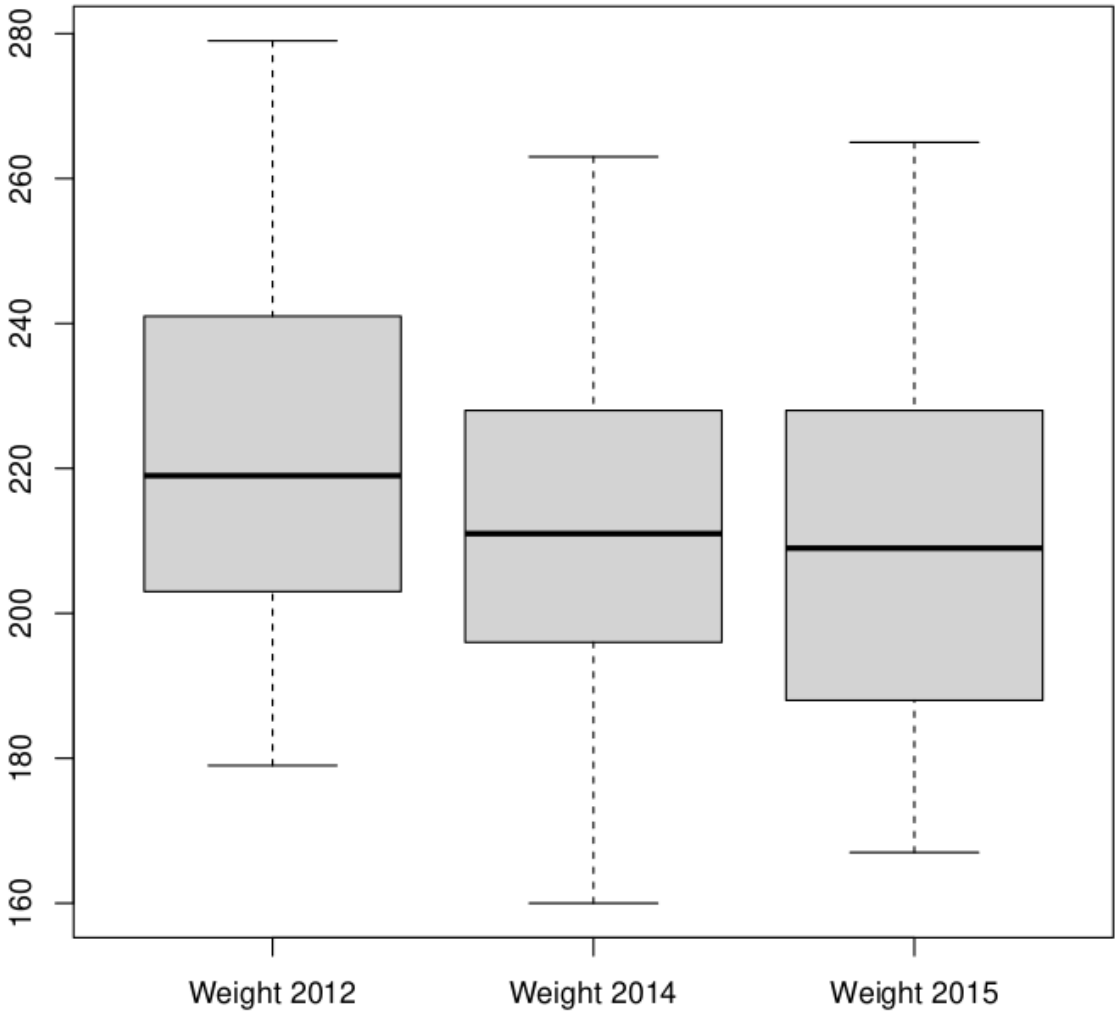
Analizemos en esta ocasión el peso de los jugadores.

Nuestra misión ahora será saber si existen diferencias entre el peso de los jugadores de la nba en los años 2012, **2014** y 2015

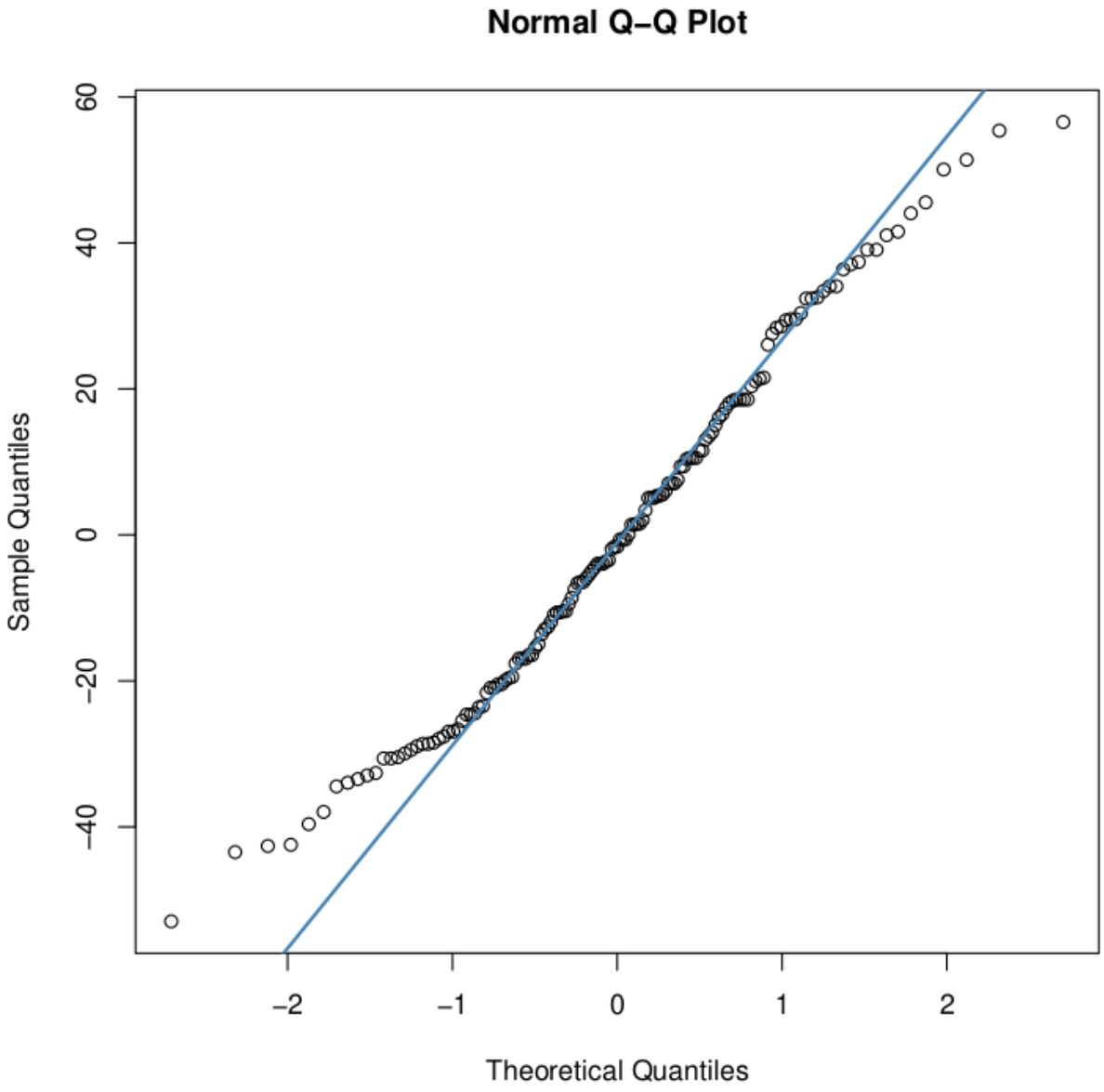
$H_0 : \mu_{2012} = \mu_{2014} = \mu_{2015}$

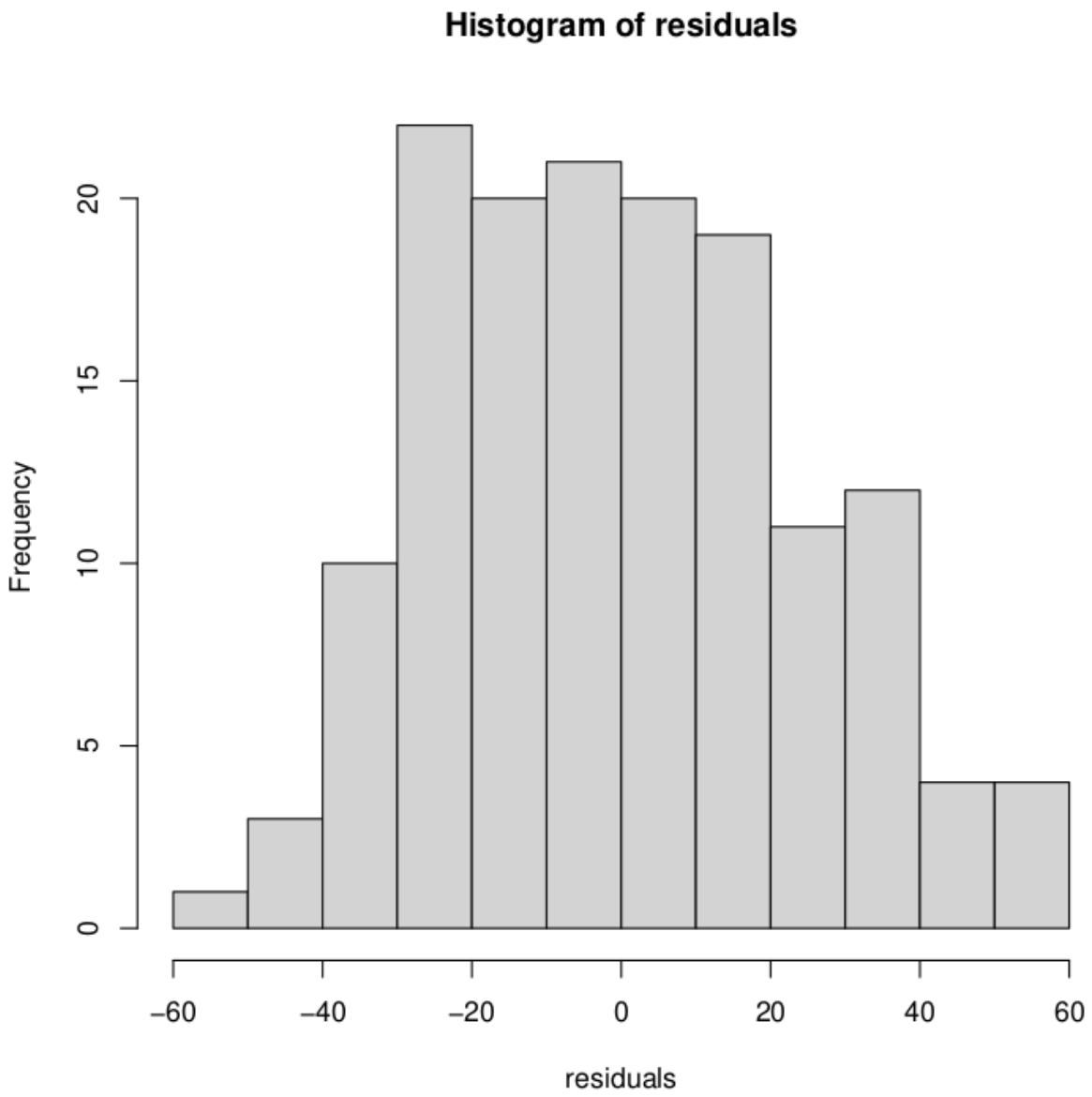
$H_a : \mu_i \neq \mu_j, \ \ i \neq j$

Como es costumbre, veamos la gráfica de los datos



En esta ocasión, vemos una diferencia ligera, pero notable, de los datos del 2012 respecto a los demás años. Luego de hacer el **análisis de varianza**, tenemos que el p-value 0.02 < 0.05, rechazando H_0, luego el peso de los jugadores de la **NBA** en los años estudiados ha cambiado.


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
60
40
20
0
−20
−40
−2
−1
0
1
2



Viendo las gráficas, nuevamente la normalidad de los residuos es dudosa. Realizando las pruebas de hipótesis de Shapiro, Durbin-Watson y Bartlett respectivamente tenemos:

1) los e_{ij} siguen una distribución normal: p-value = 0.054 > 0.050, podemos decir entonces que están normalmente distribuidos para la tolerancia del 5%.

2) los e_{ij} son independientes: p-value = 0.19 > 0.05, por lo que los errores son independientes.

3) los e_{ij} tienen la misma varianza: p-value = 0.97 $\gg$ 0.05, las varianzas de los residuos son homogéneas.

Sprint

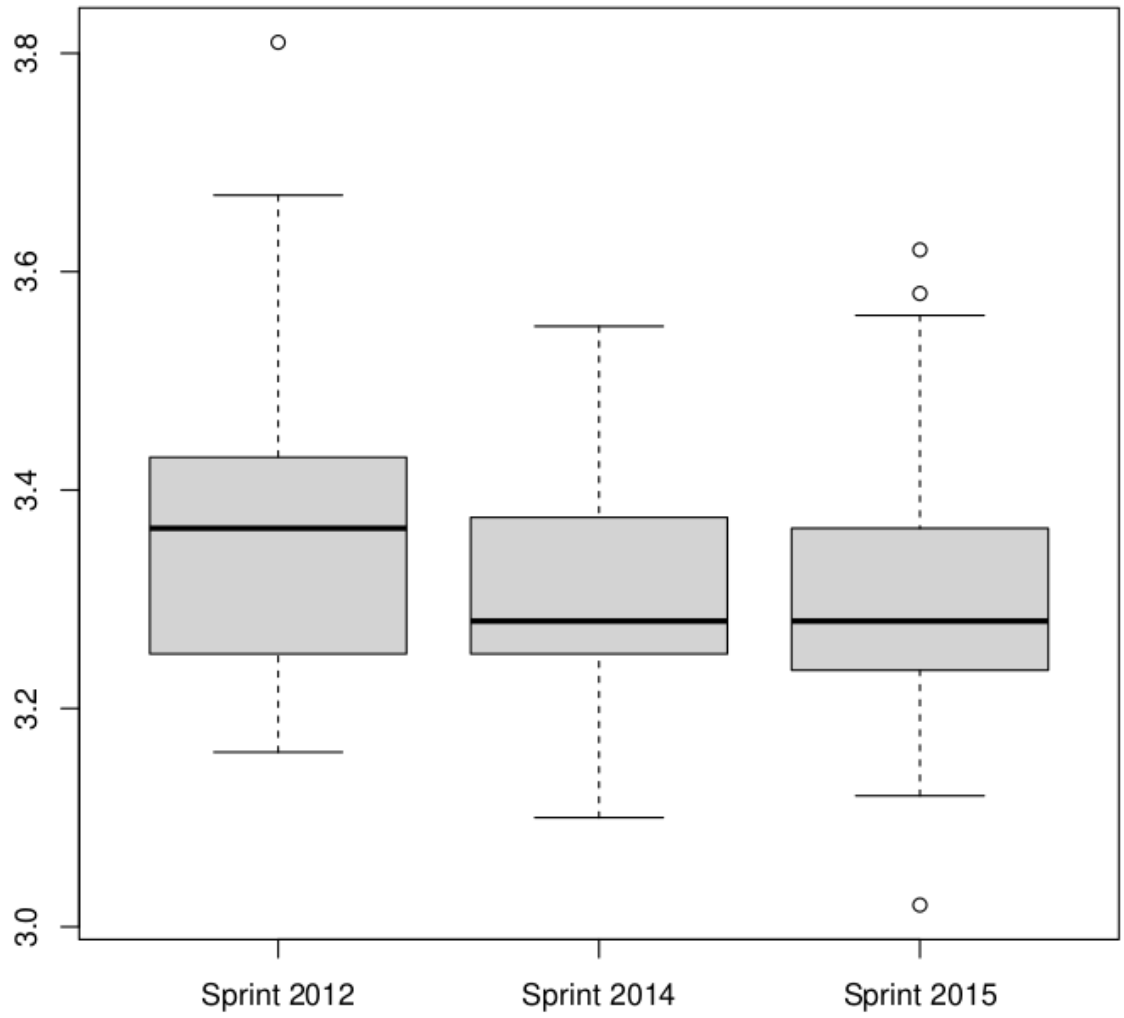
Por último, hagamos un último análisis de varianza, en este caso para la velocidad de los jugadores.

Deseamos saber si la velocidad de los jugadores de la nba ha cambiado en los años 2012, **2014** y 2015

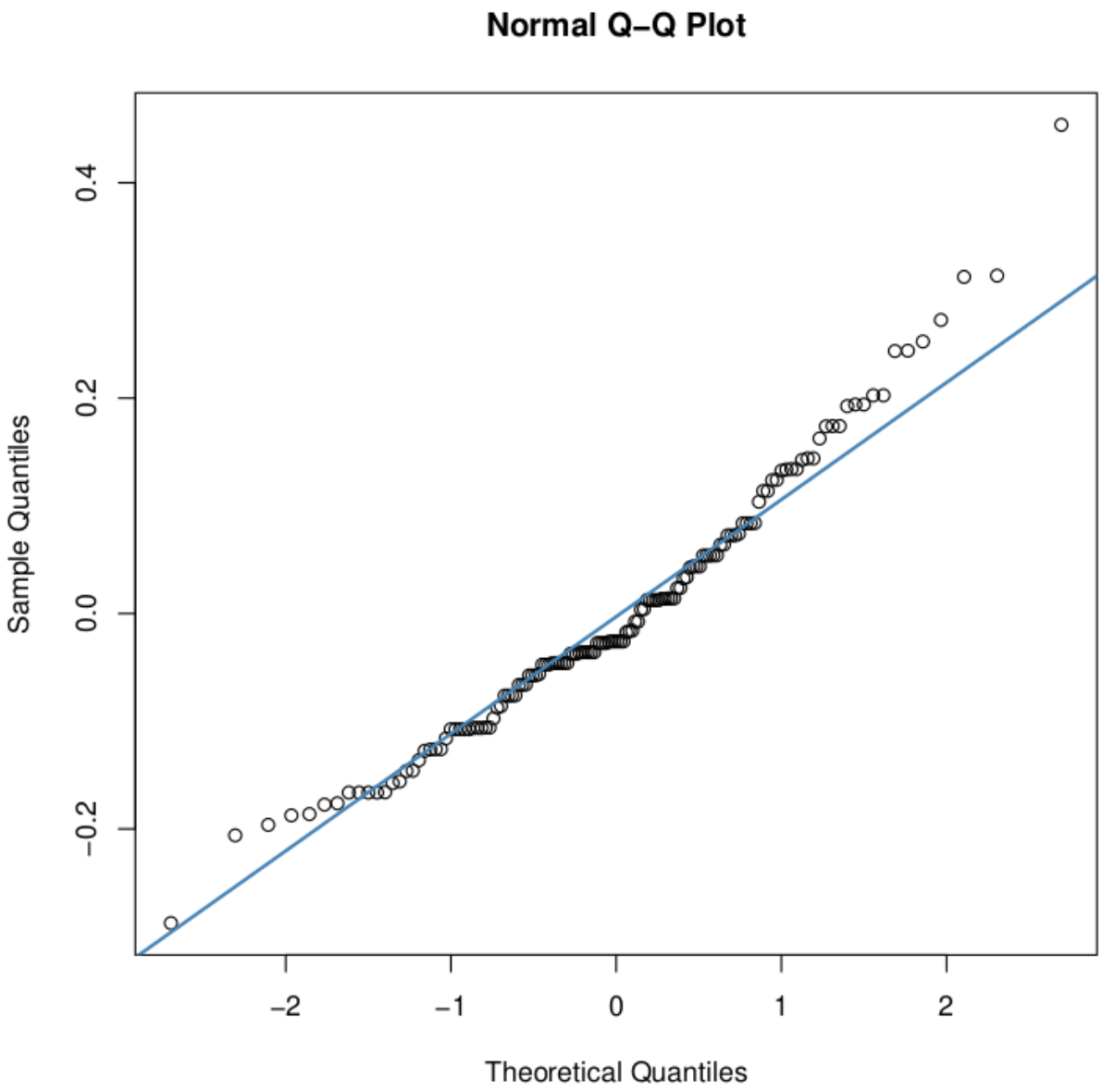
$H_0 : \mu_{2012} = \mu_{2014} = \mu_{2015}$

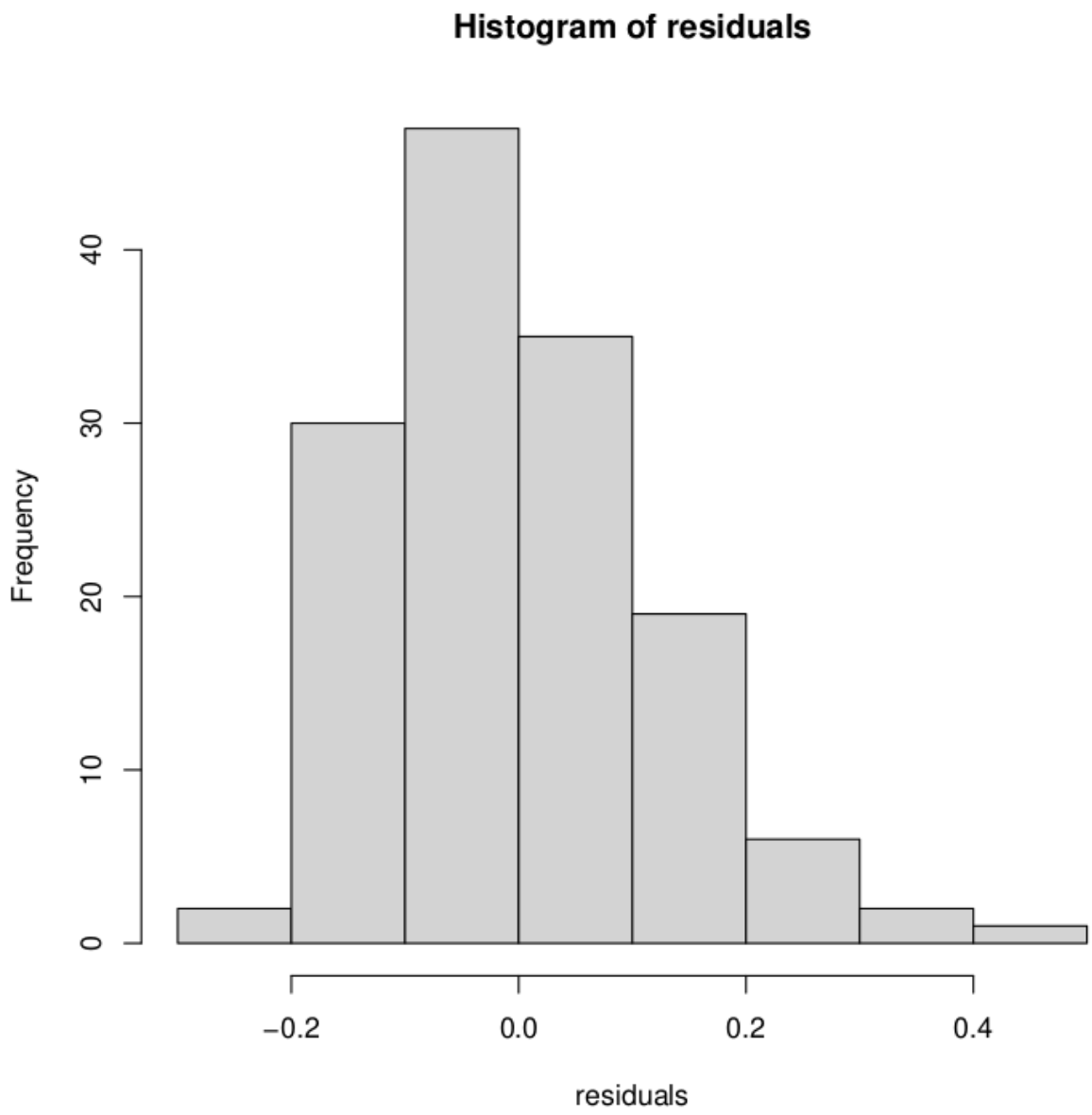
$H_a : \mu_i \neq \mu_j, \ \ i \neq j$

Viendo la gráfica de los datos, tenemos



Nuevamente, vemos una leve diferencia de los datos del 2012 respecto al resto. Haciendo entonces el **análisis de varianza**, tenemos que el p-value 0.03 < 0.05, rechazando H_0, luego las velocidades de los jugadores de la **NBA** difieren en los años estudiados.


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
0.4
0.2
0.0
−0.2
−2
−1
0
1
2



En este último análisis, nos volvemos a enfrentar con la dubitativa de la normalidad de los residuos, si bien en esta ocasión la gráfica habla por si sola. Aún así, haremos las pruebas de hipótesis de Shapiro, Durbin-Watson y Bartlett respectivamente tenemos:

1) los e_{ij} **no** siguen una distribución normal: p-value = 0.0001 $\ll$ 0.05, luego los residuos no están normalmente distribuidos

2) los e_{ij} son independientes: p-value = 0.39 $\gg$ 0.05, por lo que los errores son independientes.

3) los e_{ij} tienen la misma varianza: p-value = 0.28 $\gg$ 0.05, las varianzas de los residuos son homogéneas.

Conclusión

Luego de este extenso análisis de los datos con los que contábamos, en el cual profundizamos los conocimientos adquiridos a lo largo del curso para el tratamiento de los datos, logramos obtener **información**, encontrar conexiones invisibles entre lo que a simple vista son números, y en general lograr un mejor entendimiento del objeto de estudio. Se analizó la correlación(directa, inversa o inexistente) entre distintas medidas físicas de los aspirantes a incorparse a los equipos profesionales de la NBA en el año 2014. Igualmente, se aplicaron técnicas de clasificación para reducción de dimensión, clasificando a los jugadores en categorías similares según los datos, y se culminó realizando un análisis a mayor escala, considerando datos de años anteriores y posteriores con tal de contrastar y realizar un análisis de varianza(ANOVA) efectivo sobre distintas categorías como la agilidad, el peso, la altura y la altura de salto de los deportistas.