Taller de Consultoria - TP4

Gonzalo Barrera Borla

13/10/2019

Setup

```
library(tidyverse) # manipulación de datos en general, graficos
library(broom) # limpieza y estructuracion de resultados de regresiones
library(PMCMRplus) # comparaciones múltiples para tests de rangos
library(caret) # Validación cruzada para múltiples modelos
library(gridExtra) # Paginado de gráficos
```

Ejercicio 1

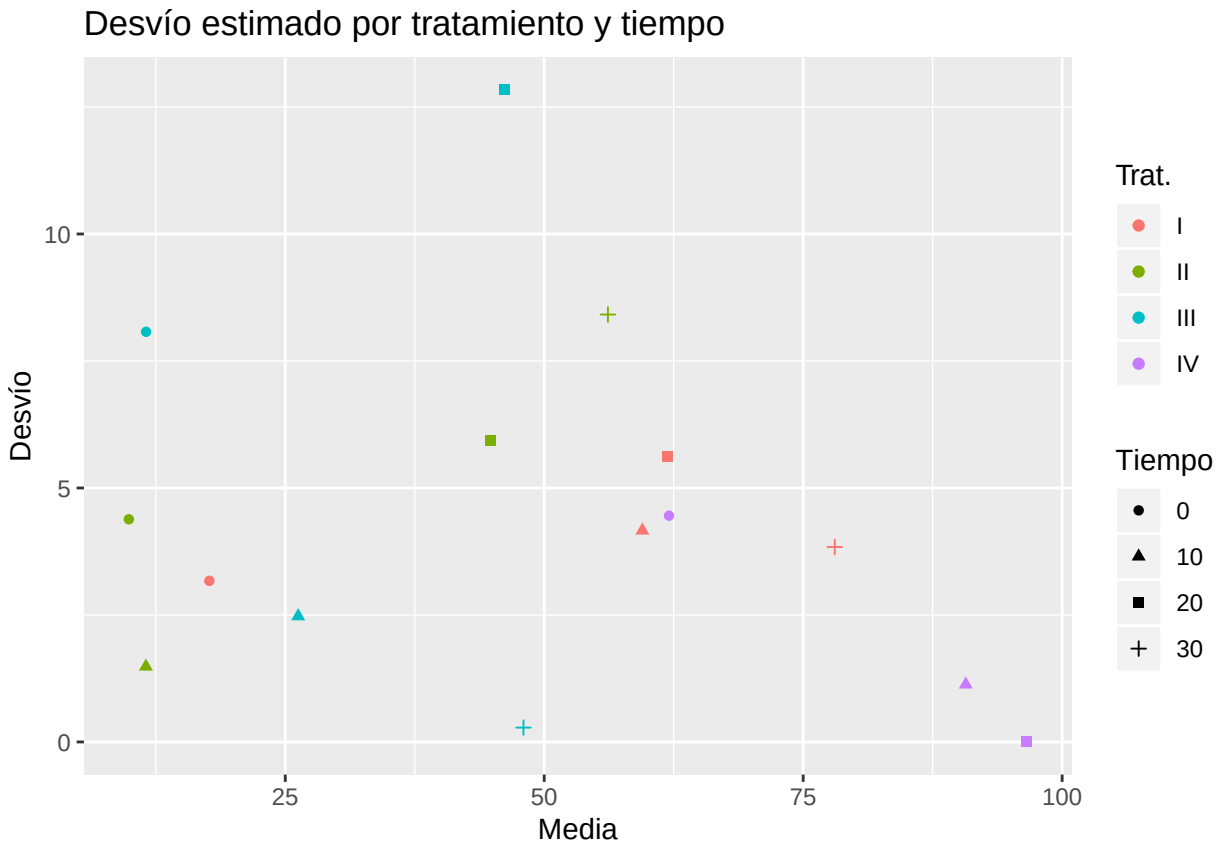
> Se midió el daño al ADN (porcentaje de ADN que se separa durante la electroforesis) en raíces de habas sometidas a 4 tratamientos, aplicados durante diferentes tiempos. Interesa decidir si algún tratamiento es demostrablemente más dañino. Los "-" indican datos no medidos. El tiempo cero para cada uno de los tratamientos puede interpretarse como la cantidad de ADN que está inicialmente dañado, o sea que no es equivalente a "ningún tratamiento". Notar que los resultados son porcentajes, lo que implica heteroscedasticidad.

Comenzamos por analizar la relación entre media y varianza para cada tiempo y tratamiento. Como se ve a continuación, los resultados no son muy alentadores:

Trat.	Tiempo	Media	Desvío
III	20	46.183	12.835
II	30	56.150	8.415
III	0	11.567	8.075
II	20	44.800	5.940
I	20	61.925	5.625
IV	0	62.050	4.455
II	0	9.900	4.384
I	10	59.475	4.163
I	30	78.050	3.839
I	0	17.675	3.172
III	10	26.250	2.475
II	10	11.550	1.485
IV	10	90.700	1.131
III	30	48.000	0.283
IV	20	96.600	0.000

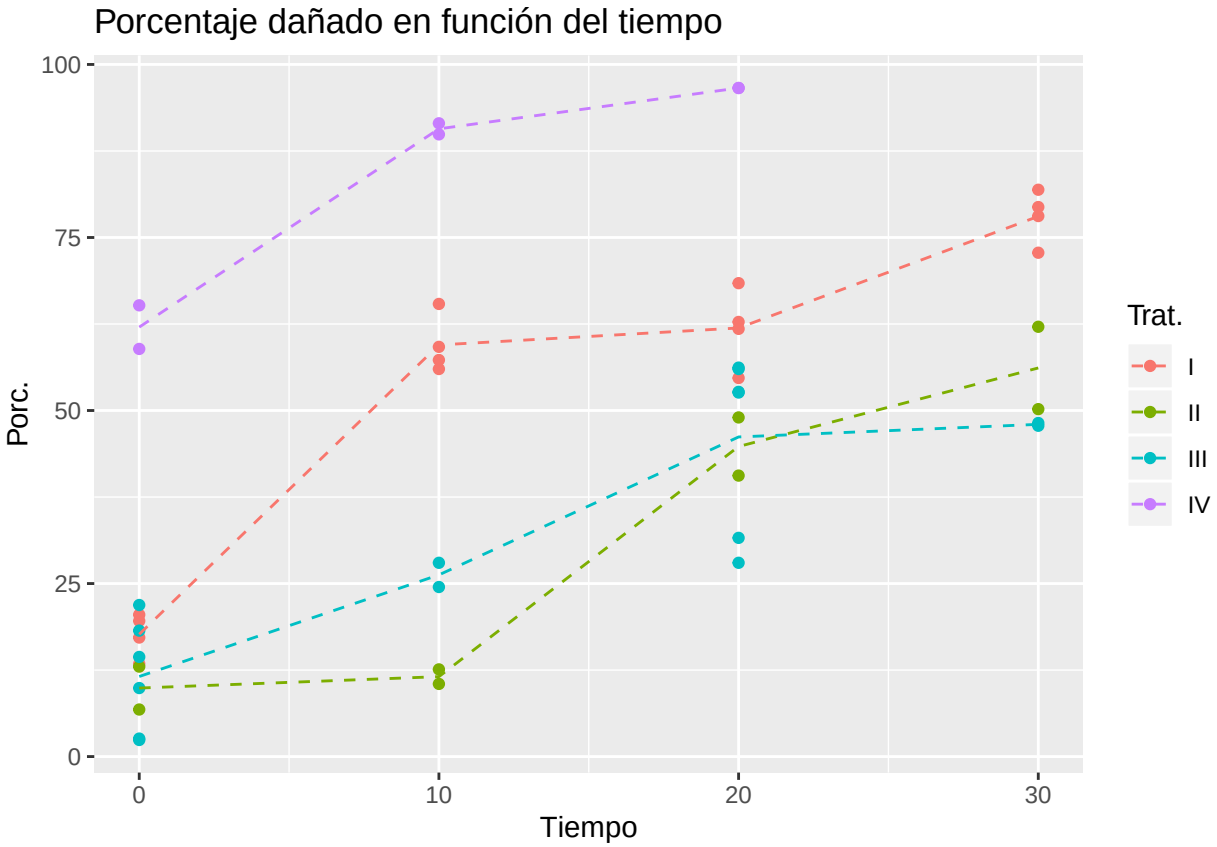
En primer lugar, para el tratamiento IV y $t = 20$, la varianza estimada es exactamente 0, por lo que la razón entre la máxima y mínima varianza para todas las combinaciones de bloque y tratamiento está indefinida, y la regla heurística que plantea Seber [1977, p.195] y utilizáramos en el TP3 no se puede uar. Aún descartando esta última observación, para el tratamiento III la varianza estimada es 12.84 en $t = 20$ y sólo 0.28 en $t = 30$, que arroja una altísima razón de 43.35.

Aún peor, cuando graficamos la varianza estimada en función de la media estimada, no se evidencia ninguna relación, por lo que ni siquiera sabríamos cómo morigerar la heterocedasticidad existente:



Por el momento, nos bastará con haber notado esta complicación, y avanzaremos con el análisis.

¿Qué modelo puede llegar a representar adecuadamente estos datos? Para contestarnos, comenzamos graficando el porcentaje de ADN dañado en función de t para cada observación, coloreada según el tratamiento. De fondo, agregamos en línea punteada el "perfil" del porcentaje promedio de ADN dañado por tratamiento y tiempo, para distinguir tendencias:



Pareciera que al menos el tratamiento IV se distingue desde el comienzo de los demás, y que el porcentaje de daño es siempre creciente en t, a una tasa constante o *tal vez* decreciente.

Aunque esté medida en pocos valores únicos, la covariable `t` (tiempo) es cuantitativa, y por ende entendemos que es más razonable plantear lo que en la literatura se conoce como un *análisis de la covarianza* (ANCOVA) de un factor (el tratamiento) con 4 niveles, antes que un modelo de análisis de la covarianza de 2 factores.

Combinando estas observaciones, planteamos como modelos:

- (de referencia) un ANOVA de 2 factores,
- (y como candidatos) 6 modelos "ANCOVA" con un factor (`tratamiento`) sobre
  - polinomios de grado 1, 2 y 3 en `t`,
  - con y sin interacciones

Elegiremos polinomios "crudos" en lugar de los ortogonales por defecto de R para facilitar la comprensión de los coeficientes resultantes.

Si uno ajusta los modelos sobre los datos completos, no importa la métrica elegida, los modelos con más parámetros (ANOVA 2 factores y el cúbico multiplicativo) tienden a ser los que mejor ajustan: con sólo 46 datos, ajustar 16 coeficientes es casi una garantía de éxito.

Para evitar este sesgo hacia modelos más complejos, utilizamos el muy recomendable paquete `caret`, que nos permite ajustar modelos con distintas técnicas de resampleo. Para ete ejercicio, optamos por el clásico split entre datos de entrenamiento (70%) y testeo (30%), y repetimos el proceso 100 veces. A continuacion, incluimos el error cuadrático medio estimado para cada modelo, y su desvío estándar estimado, para poder seleccionar el "mejor" modelo según la ya discutida regla de un desvío estándar:

Modelo	ECM	Desvio
p ~ factor(t) * trt	9.12	2.80
p ~ poly(t, 2, raw = TRUE) + trt	9.83	1.36

Modelo	ECM	Desvio
p ~ poly(t, 1, raw = TRUE) + trt	9.88	1.23
p ~ poly(t, 3, raw = TRUE) + trt	10.07	1.72
p ~ poly(t, 2, raw = TRUE) * trt	10.45	4.17
p ~ poly(t, 1, raw = TRUE) * trt	10.68	1.92
p ~ poly(t, 3, raw = TRUE) * trt	14.32	11.41

Como es de esperar, el modelo cúbico multiplicativo no resiste la validación cruzada. Sin embargo, el modelo que mejor ajusta, aún usando técnicas de validación cruzada, sigue siendo el ANOVA de 2 factores, que cuenta con la mayor libertad posible para adecuarse a los datos (tiene la misma cantidad de coeficientes que el cúbico multiplicativo, pero sin correlación alguna entre las covariables predictoras, que sí tienen t, t^2, t^3). Sin embargo, se observa que el desvío estimado para el modelo ANOVA de 2 factores es bastante alto, lo suficiente como para que la regla de 1 SD contenga cómodamente a todos los modelos salvo, justamente el cúbico multiplicativo. En este contexto, nos inclinamos por el modelo más sencillo posible, que es el lineal aditivo. Ajustamos entonces

$$p_{ij} = \mu_i + \beta \times t + \epsilon_{ij} \quad \text{con } \epsilon_{ij} \sim N(0, \sigma^2),\ i \in \text{I, II, III, IV}, 1 \leq j \leq n_i$$

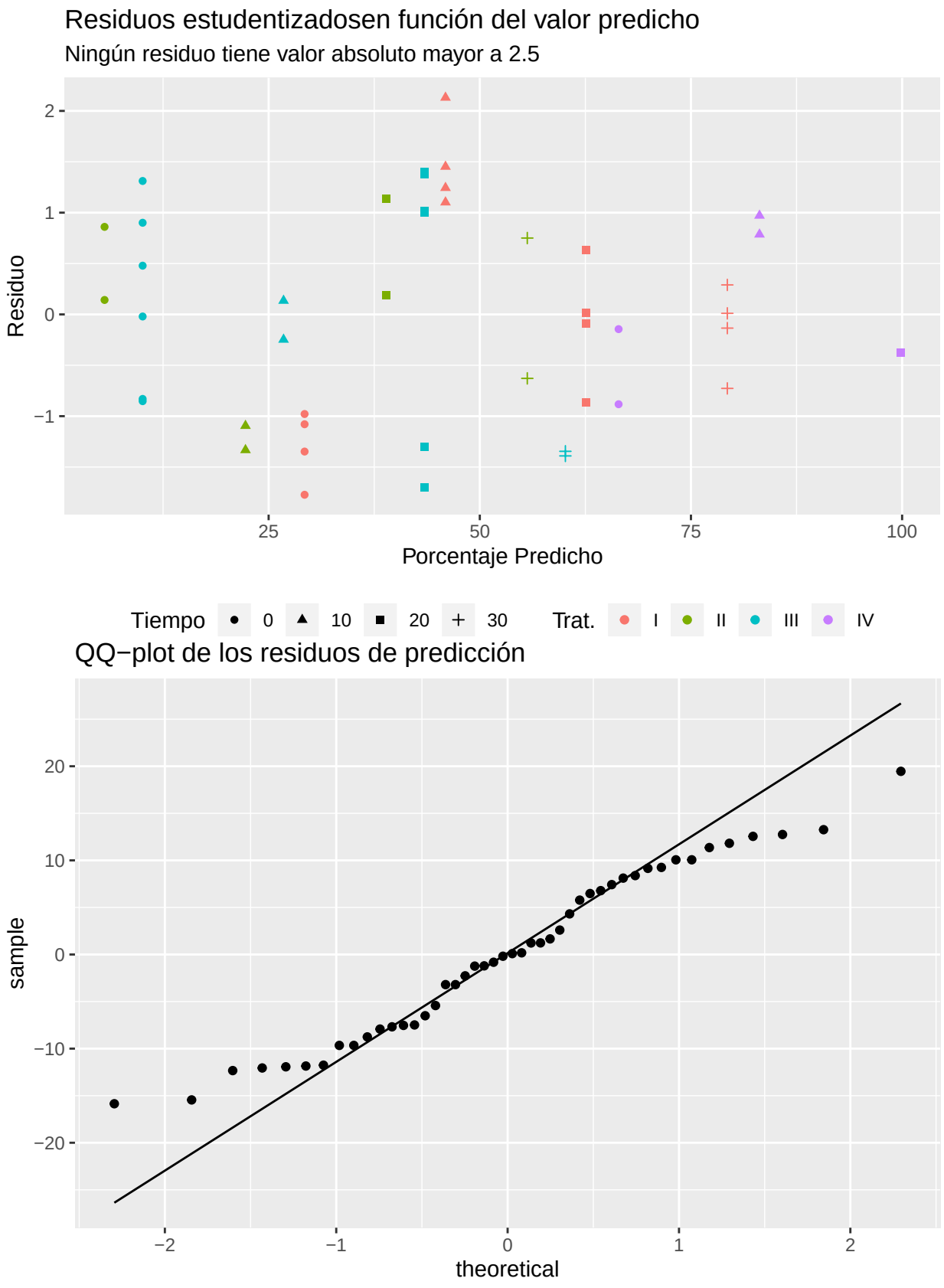
donde μ_i es la ordenada correspondiente al tratamiento i, t es el tiempo de medición y n_i la cantidad de observaciones del i-ésimo tratamiento, *a sabiendas* de que no hemos podido garantizar todavía la homocedasticidad de los ϵ_{ij}. Los principales estadísticos del modelo, efectos principales y coeficientes del modelo *ajustado con todos los datos* resultan:

R^2 aj.	F obs.	P-valor
0.879	82.9	4.36e-19

Coefs	GL	C. Medios	F obs.	P-valor
trt	3	55.5	55.5	1.71e-14
t	1	165	165	5.86e-16
Residuals	41	NA	NA	NA

Coef.	Estimado	P-valor
(Intercept)	29.3	5.57e-12
trtII	-23.7	8.79e-07
trtIII	-19.2	1.13e-06
trtIV	37.2	4.41e-10
t	1.67	5.86e-16

Podemos observar que tanto (i) el modelo completo, (ii) los efectos principales del tratamiento y el tiempo y (iii) cada coeficiente individual son significativos con bajísimo p-valor. Revisamos los residuos estudentizados de las predicciones, para estimar si hemos capturado satisfactoriamente la estructura de los datos, y no observamos nada particularmente alarmante:



Lamentablemente, el gráfico cuantil-cuantil de los residuos de predicción revela que el supuesto de normalidad

en los ϵ_{ij} no es realmente sostenible. Un test de normalidad tradicional como el de Shapiro-Wilks arroja un p-valor de 0.18, pero un buen gráfico vale más que mil tests. Intentamos remover iterativamente las observaciones con residuos más extremos y volver a ajustar el cojunto de modelos planteados, pero recién después de remover 8 observaciones conseguimos un ajuste razonable del QQ-plot. En todo ese proceso, $p \sim tratamiento + t$ siguió siendo siempre el mejor modelo según la R1DE, por lo que preferimos mantener el dataset completo y reportar estadísticos y coeficientes del ajuste con todos los datos.

A pesar de la evidente heterocedasticidad original de las observaciones, hemos obtenido un ajuste bastante satisfactorio para los datos, con una directa interpretación física. Sabemos que cuando asumimos erróneamente una matriz de covarianza $\sigma^2 I_n$ en lugar de la verdadera V, los estimadores de los parámetros del modelo $\hat{\beta}$ son insesgados, pero no los de menor varianza. Así y todo, hemos obtenido unos estimadores indudablemente distintos a cero, con lo cual podemos en principio confiar en ellos, ya que de usar la verdadera matriz de covarianza su significación debería mejorar.

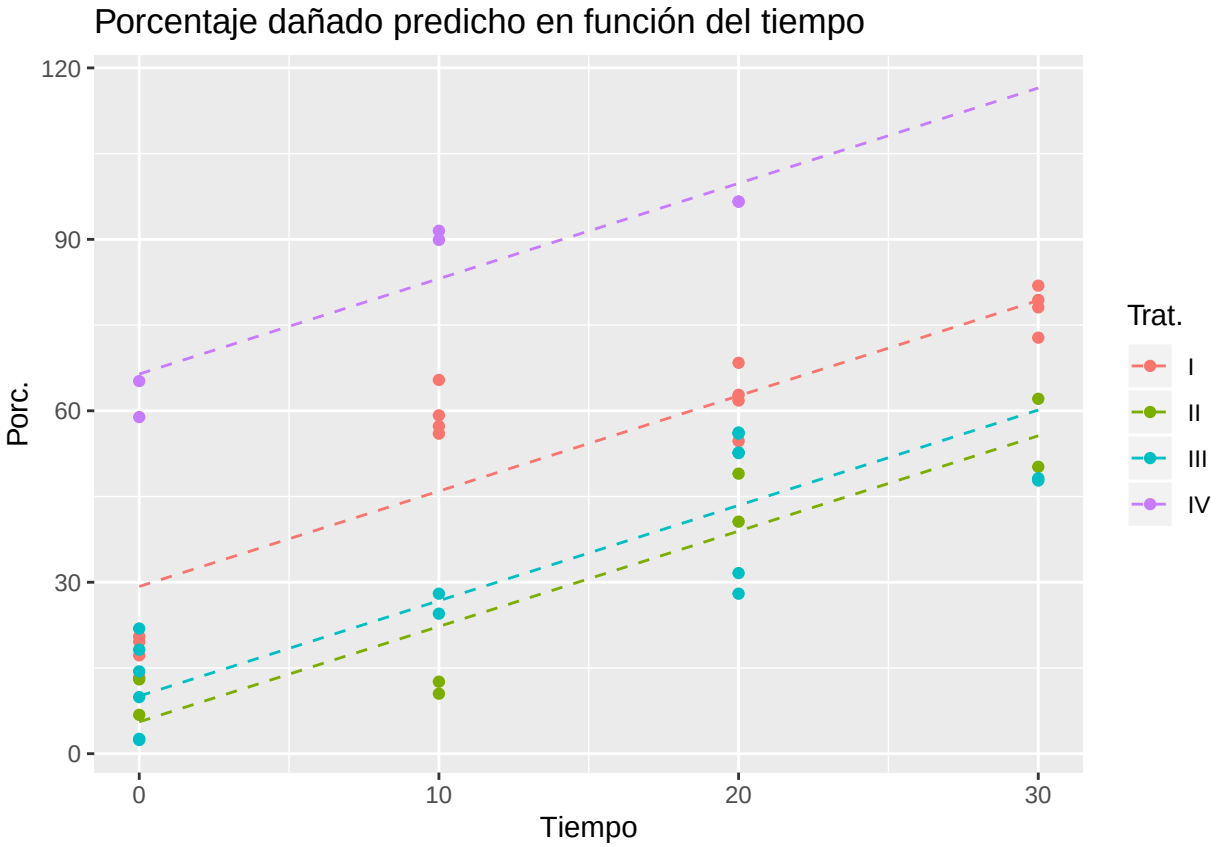
Finalmente, podemos afirmar que

- el porcentaje inicial de ADN dañado varía con el método,
- el avance del daño es aproximadamente lineal en el tiempo, y
- el avance del daño en el tiempo no varía significativamente con el método.

Los modelos ajustados son:

$$\begin{aligned} p_{I,j} &= 29.3 + 1.67 \times t + \epsilon_{I,j} \\ p_{II,j} &= 5.6 + 1.67 \times t + \epsilon_{II,j} \\ p_{III,j} &= 10.1 + 1.67 \times t + \epsilon_{III,j} \\ p_{IV,j} &= 66.5 + 1.67 \times t + \epsilon_{IV,j} \end{aligned}$$

Y los tratamientos, del más al menos dañino, son IV > I > III > II. Concluimos con un gráfico de las funciones ajustadas sobre los datos originales:



Dado que las varianzas entre diferentes poblaciones son tan heterogéneas, no podemos justificar el uso de un procedimiento de comparaciones múltiples como el de Tukey así que no presentaremos la tabla de resultados completa, pero mencionaremos que devuelve básicamente lo que uno intuitivamente observa de analizar los coeficientes y el último gráfico. Del menos al más dañino, se ordenan $II < III < I < IV$, y son todos significativamente diferentes entre sí, salvo por II y III.

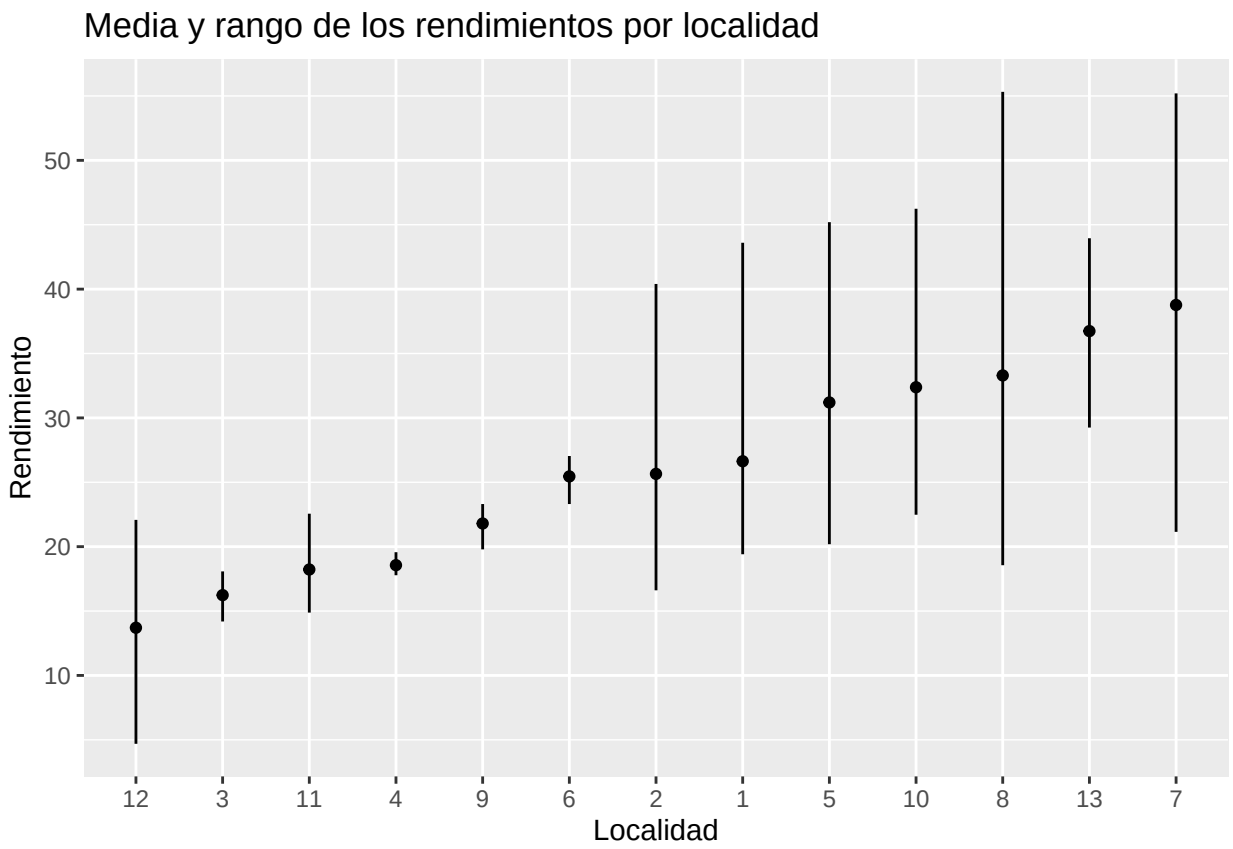
Ejercicio 2

> Se dan los rendimientos de 4 variedades de trigo en 13 localidades de Oklahoma. Interesa determinar qué variedades son recomendables.

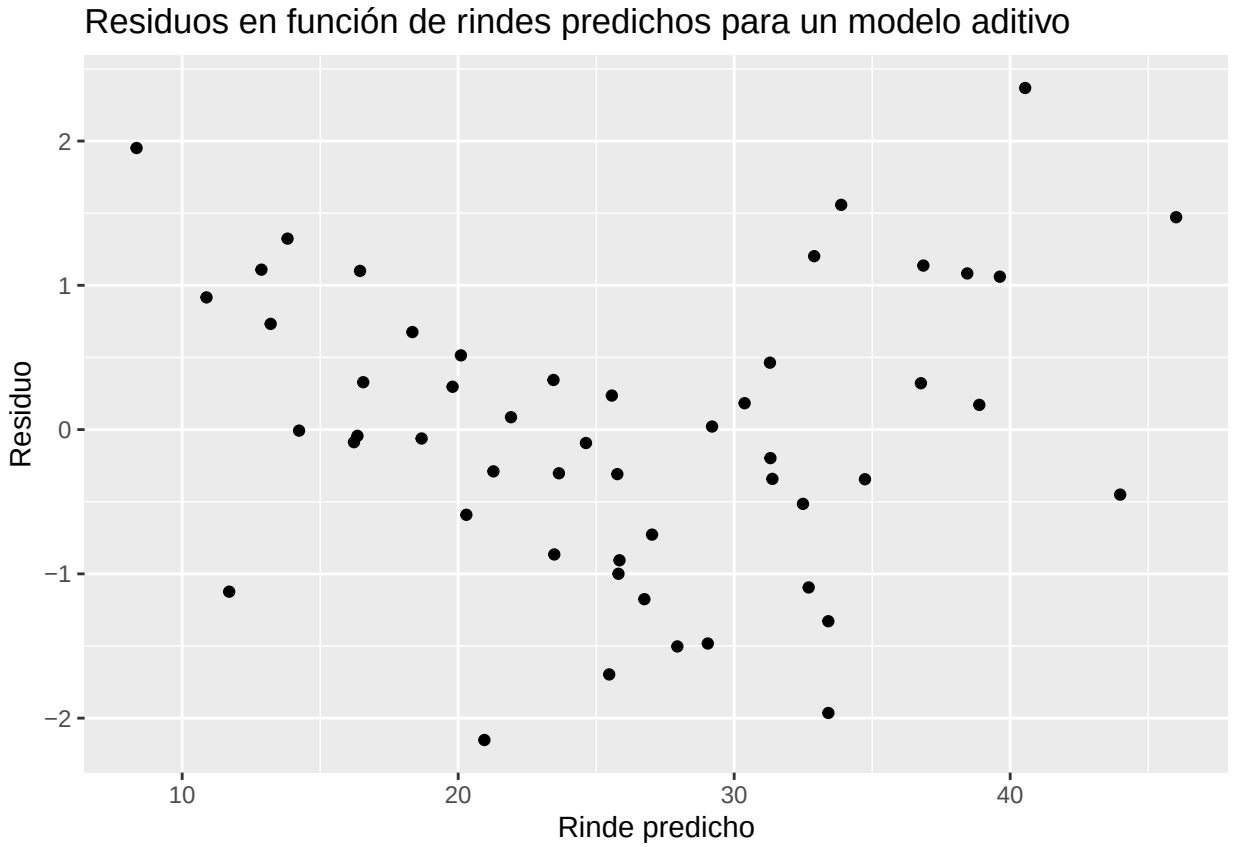
Inmediatamente se observa que no sólo la media de los rendimientos varía considerablemente entre localidades, sino que el rango de rendimientos para cada localidad es sumamente variable:

Loc.	Rango	Media
8	36.76	33.30
7	34.05	38.77
5	25.01	31.20
1	24.19	26.63
2	23.79	25.65
10	23.76	32.38
12	17.39	13.70
13	14.70	36.74
11	7.68	18.23
3	3.89	16.24
6	3.73	25.45
9	3.52	21.80

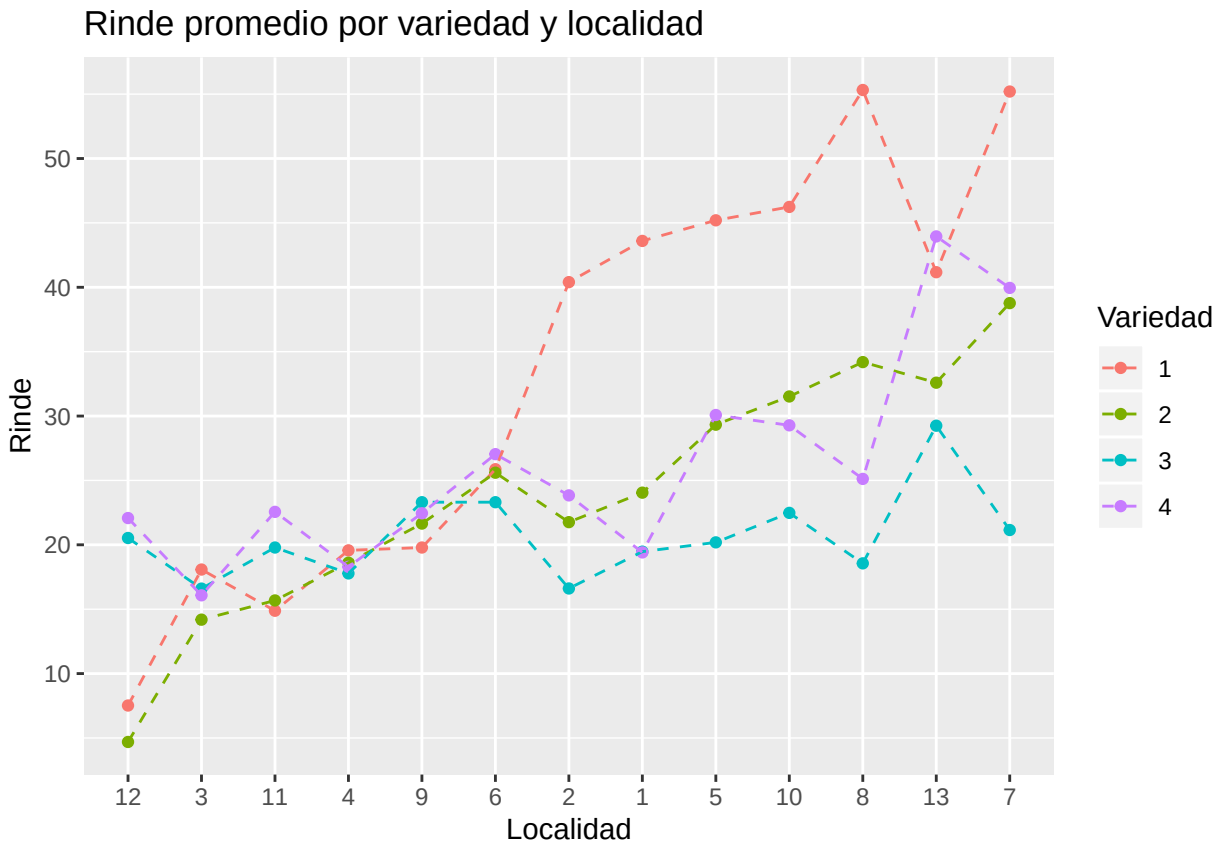
Loc.	Rango	Media
4	1.79	18.56



De lo expuesto resulta improbable que un análisis de la varianza de dos factores "directo" nos provea mucha información útil. En principio, al tener sólo una observación por "celda" (combinación de bloque y tratamiento), sólo podemos plantear un modelo aditivo o con alguna interacción limitada como la que propone Tukey con su invento de "1 grado de libertad para la no-aditividad". En efecto, si ajustamos un modelo del estilo $rinde \sim variedad + localidad$, tanto la localidad como las variedades parecen sumamente significativas, pero el gráfico de los residuos versus predichos tiene una clara estructura "parabólica":



Si graficamos los perfiles de rendimiento en función de la localidad, con las localidades ordenadas crecientemente en rendimiento promedio para mejroar la visibilidad, vemos que las variedades de maíces se entremezclan en las localidades de menor rinde, y se separan un poco mejor en las de mayor rinde, aunque todavía con cruzamientos significativos para las variedades 1, 2 y 4.



Aquí se nos plantea una disyuntiva: por un lado, podríamos considerar por separado los dos regímenes de ciudades antes mencionados:

- las de "bajo rinde" (12, 3, 11, 4, 9, y 6) con perfiles muy superpuestos y rangos de valores más "apretados", y
- las de "alto rinde" (2, 1, 5, 10, 8, 13 y 7), con perfiles mejor separados, y mayores rangos de valores.

Por otra parte, podríamos mantener todas las observaciones juntas, y considerar algún test no paramétrico con pocos supuestos sobre los datos, para ver si podemos obtener alguna conclusión global. En esta línea, surgen naturalmente dos tests para diseños en bloque completos sin replicación: el test de Friedman (análogo al test del signo para varias muestras), y el test de Quade (análogo al test de Wilcoxon de rangos signados).

Siguiendo a Conover [1999, pp. 369-375], los supuestos necesarios para estos tests no paramétricos son mínimos. Ambos requieren que (i) los bloques sean independientes entre sí y que (ii) las observaciones en cada bloque sean ordenables según algún criterio. Para el test de Quade, se pide además que (iii) los bloques se puedan ordenar por su rango muestral. En estos datos los tres supuestos se cumplen, pero nos inclinaremos por el test de Quade, que entendemos aprovechará mejor la información pertinente a la amplitud de rangos intra-bloques.

En ambos casos, asumimos que si X_{ij} es el rinde de la j-ésima variedad en la i-ésima localidad, $X_{ij} \sim F_i(x - \theta_j), F_i \in \Omega_0 \, \forall \, i \in \{1, \ldots, 13\}$, $j \in \{1, \ldots, 4\}$ son todas VA independientes. Es decir que F_i es la distribución de las observaciones del i-ésimo bloque (localidad), y dentro del bloque θ_j es la mediana del j-ésimo tratamiento (variedad). La hipótesis nula será $H_0 : \theta_i = \theta_j \forall i, j$, y la alternativa su complemento (i.e., al menos dos medianas son distintas entre sí).

Sin demasiada confianza en poder obtener resultados muy significativos, elegimos a priori un nivel de significación $\alpha = 0.1$. El test de Quade arroja un p-valor de 0.00236, que nos lleva a rechazar la hipótesis nula con seguridad. De aplicar el test de Friedman, obtenemos un p-valor de 0.0169, más moderado pero aún contundente. Resulta tranquilizador saber que aún sin aprovechar la información sobre el rango de los bloques, hubiésemos rechazado la hipótesis nula.

Hasta aquí sabemos que no todas las variedades tienen el mismo rendimiento, pero aún nos resta saber cuáles son significativamente distintas de las demás. En principio, los rendimientos promedios resultan ser

3	2	4	1
20.69	24.05	26.16	33.3

Siguiendo a Conover, al haber rechazado la hipótesis nula, podemos realizar un test de comparaciones múltiples, manteniendo el $\alpha = 0.1$ original. Acudimos al paquete `PMCMRplus`, que implementa comparaciones múltiples "post-hoc" para una amplia variedad de tests basados en la suma de rangos medios, y obtenemos los siguientes p-valores:

	1	2	3
2	0.064	NA	NA
3	0.002	0.346	NA
4	0.346	0.346	0.078

En otras palabras, a nivel $\alpha = 0.1$,

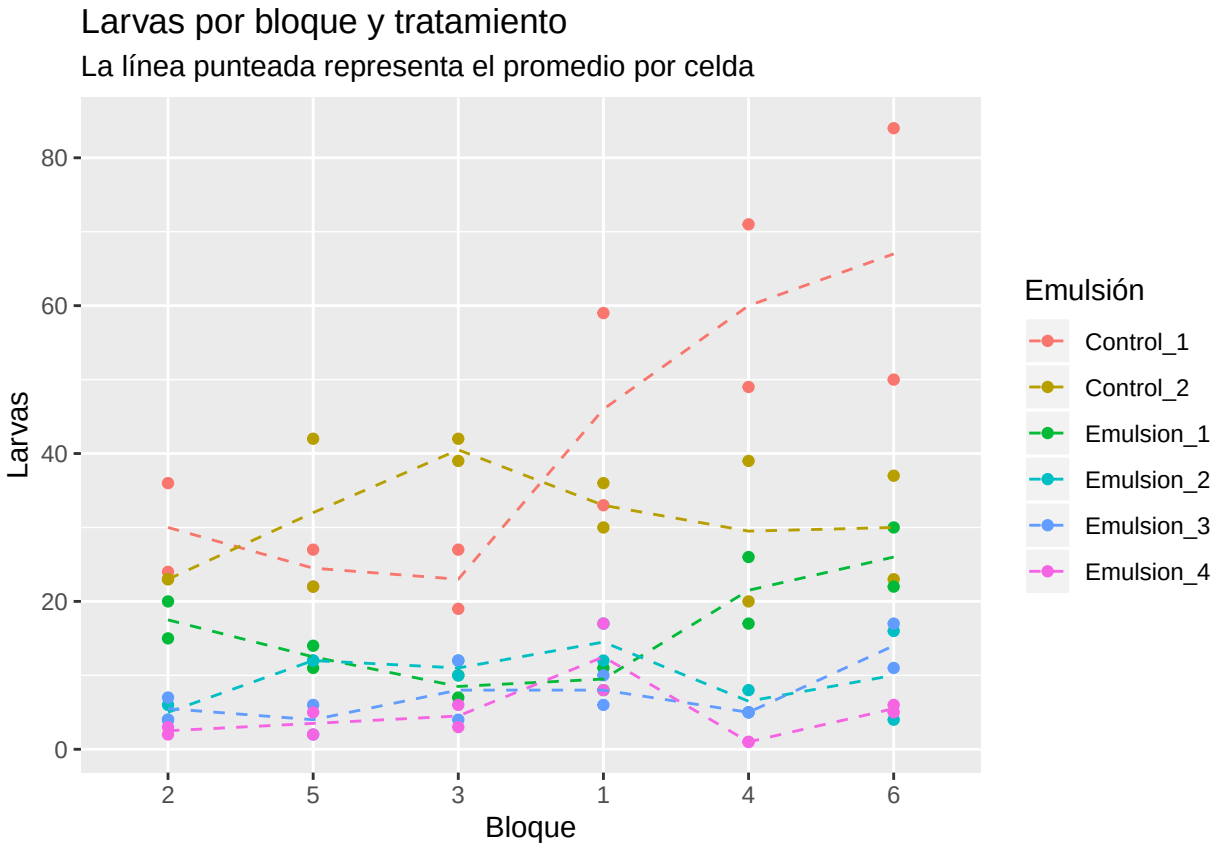
- la variedad 1 es mejor que la 2 y 3, pero no se peude distinguir de la 4,
- la variedad 4 es mejor que la 3, pero no se puede distinguir de la 2,
- las variedades 2 y 3 no se pueden distinguir.

De tener que recomendar una sola variedad, y asumiendo que no hay ningún otro factor decisivo (como ser el precio del grano, la dificultad de su cultivo, las leyes locales, et cetera), nos inclinaríamos por la 1. Nótese que de haber elegido un criterio sólo un poco más restrictivo, como ser $\alpha = 0.5$, sólo "la mejor" y "la peor" variedad (1 y 3) se llegan a distinguir entre sí.

Ejercicio 3

> Se desea comparar los efectos de 4 emulsiones tóxicas para el control de larvas de típula. En cada uno de 6 lotes ("bloques") se marcaron 6 cuadrados de 90 cm. de lado; en 4 se aplicaron las emulsiones, y 2 quedaron sin tratamiento (para control). En cada bloque, la asignación se hizo al azar. Después de algunos días, se registró la cantidad de larvas sobrevivientes. Cada cuadrado fue dividido en 9 "cuadraditos" de 30 cm. de lado, y se contaron las larvas en 2 de ellos. El objetivo es determinar cuáles emulsiones son efectivas, y si hay alguna demostrablemente mejor (para eliminar las larvas, no para engordarlas).

Para fijar el diseño experimental, necesitamos decidir (a) qué hacer con los dos tratamientos de control, y (b) cómo tratar a los dos "subbloques" (cuadraditos de 30 cm. de lado) por bloque (cuadrado de 90 cm. de lado) y tratamiento. Comenzamos por graficar los perfiles en orden creciente de larvas promedio por bloque:



Una primera aproximación *inercial* (en tanto está influida por la metodología del ejercicio anterior), nos lleva a

- como no contamos con información extra sobre la existencia o no de diferencias entre los controles, y los perfiles no son muy similares, *considerar los dos controles por separado* y
- asumiendo que la decisión de contar en sólo 2 de los 9 cuadraditos por bloque fue para ahorrar trabajo, *sumar los dos subbloques* por celda y considerar una única observación.

Y sin hacer más supuestos sobre las distribuciones, intentamos aplicar un test no paramétrico para la diferencia de medianas. Si agrupamos los subbloques y consideramos los rangos y promedios por bloque,

Blq.	Rango	Media
6	123	50.83
4	118	41.17
1	76	41.17
3	72	31.83
5	57	29.50
2	55	27.83

observamos que (a) la razón del mayor al menor rango es escasamente mayor a 2 y (b) el rango por bloque aumenta junto con la media. Además, Conover [1977, p. 368] menciona que para seis o más tratamientos, el test de Friedman tiende a tener mayor potencia que Quade, circunstancias por las cuales nos inclinamos a favor del test de Friedman por sobre el de Quade. Aplicado a nuestros datos, arroja un p-valor de 1.2×10^{-4}, suficiente para rechazar la hipótesis nula de la igualdad de medianas, a nivel, por ejemplo, $\alpha = 0.1$. Si luego realizamos un test "post-hoc" de comparaciones múltiples significativas con el mismo α, obtenemos:

	Control_1	Control_2	Emulsion_1	Emulsion_2	Emulsion_3
Control_2	1.000	NA	NA	NA	NA
Emulsion_1	0.394	0.586	NA	NA	NA
Emulsion_2	0.173	0.309	0.998	NA	NA
Emulsion_3	0.025	0.058	0.848	0.978	NA
Emulsion_4	0.002	0.006	0.394	0.683	0.978

Es decir, que sólo las emulsiones 3 y 4 son significativamente distintas de los controles, y a su vez ninguna de las emulsiones es distinguible de las demás. Recordando el gráfico de perfiles original, estos resultados nos resultan más bien extraños: puede que las emulsiones no se distingan claramente entre sí, pero ¿cómo puede la emulsión 3 ser significativamente distinta de ambos controles y la 2 no, si sus perfiles están básicamente superpuestos?

Pensándolo bien, creeríamos que el problema es autoimpuesto, al usar un test no-paramétrico, para comparar cantidades que están muy cerca entre sí (# de larvas para las emulsiones) contra otras que están muy lejos (los controles). Diferencias ínfimas entre emulsiones y considerables respecto al control, al pasarlas a la escala de rangos, quedan "emparejadas" en distancias unitarias, y si por azar una emulsión mata aunque sólo sea una larva menos que otra, el p-valor en sus tests de comparaciones múltiples sufrirá más de lo justificado. Para comprobar esta hipótesis, realizamos el mismo test de comparaciones múltiples de antes, pero sobre datos "sintéticos" en los cuales los tratamientos se ordenan siempre igual:

Control 1 > Control 2 > Emulsión 1 > Emulsión 2 > Emulsión 3 > Emulsión 4.

Los p-valores resultan:

	Control_1	Control_2	Emulsion_1	Emulsion_2	Emulsion_3
Control_2	0.951	NA	NA	NA	NA
Emulsion_1	0.488	0.951	NA	NA	NA
Emulsion_2	0.086	0.488	0.951	NA	NA
Emulsion_3	0.006	0.086	0.488	0.951	NA
Emulsion_4	0.000	0.006	0.086	0.488	0.951

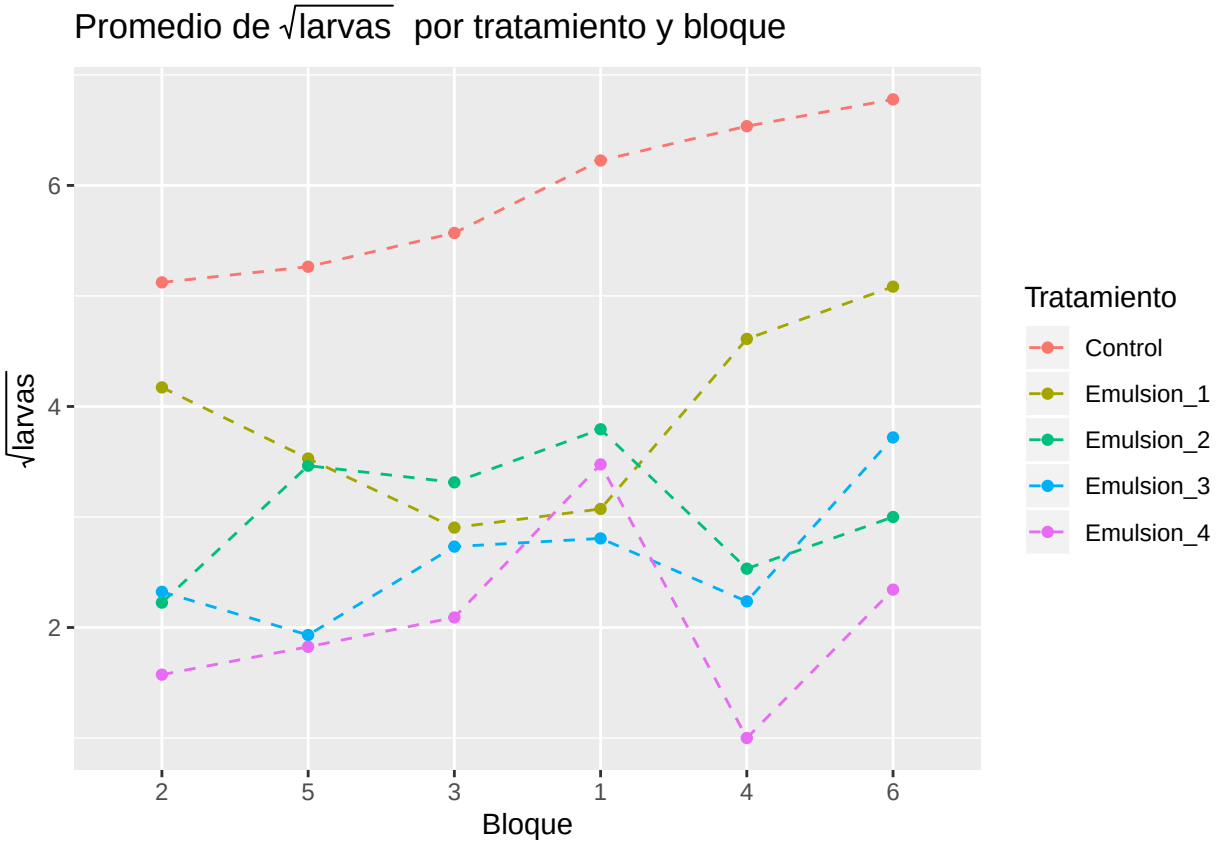
¡Prácticamente iguales a nuestros resultados! Las dos "mejores" emulsiones (3 y 4) son significativamente distintas de los controles (la emulsión 2 resulta significativamente distinta de un sólo control), y ninguna emulsión es distinguible entre sí. En esta línea, para asegurarse que la emulsión 1 sea significativamente distinta al Control 2, ¡se requerirían 57 bloques perfectamente ordenados! Ante la evidente inadecuación del método planteado a la situación, volvemos a empezar.

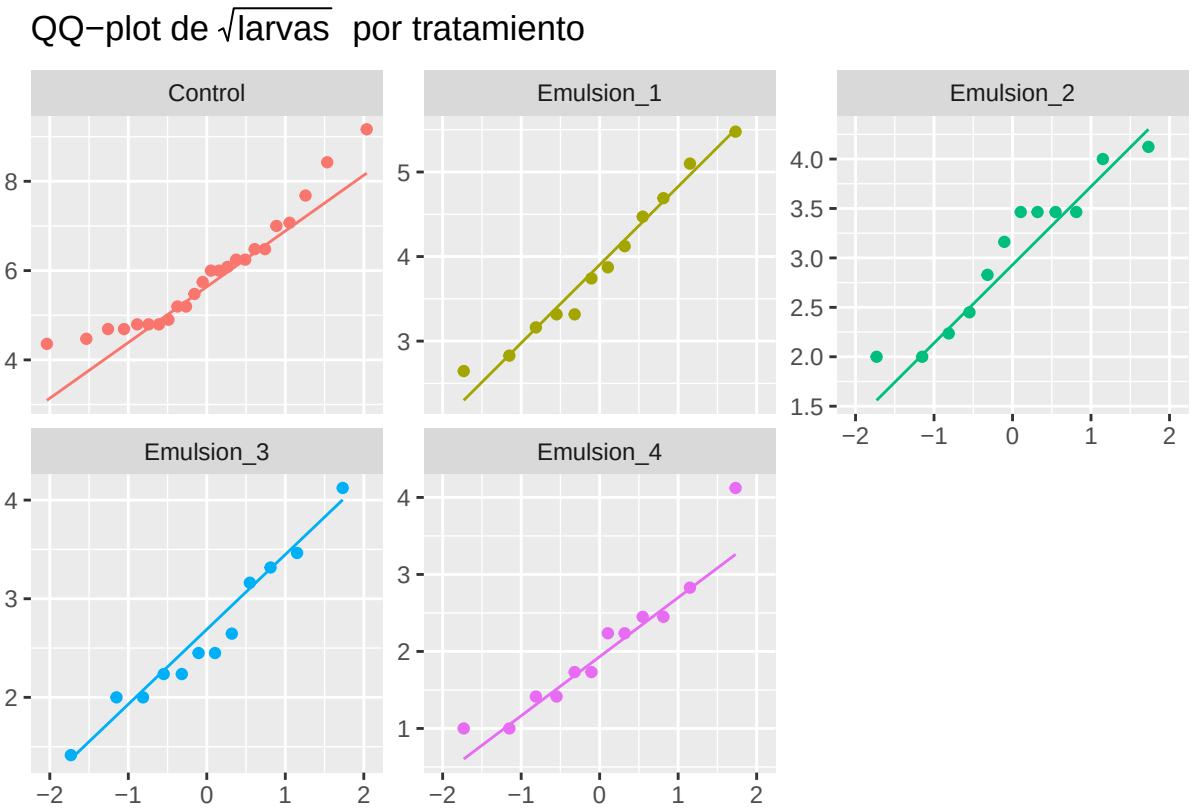
Para disminuir la cantidad de comparaciones múltiples a realizar, preferimos fusionar los dos controles en un único tratamiento. Esto pasa de 15 a 10 la cantidad de comparaciones, y además nos permite estimar mejor la varianza del control y "apretar" los intervalos de confianza que lo consideren. Para hacer uso de la información cuantitativa (y no sólo la ordinal como hasta ahora) con un tradicional diseño de ANOVA con 2 factores, debemos suponer normalidad en los datos, y una misma varianza para todas las poblaciones. Comenzamos por revisar la homocedasticidad:

Trat.	Media	Desvio
Control	36.54	16.55
Emulsion_1	15.92	7.29
Emulsion_2	9.83	4.45
Emulsion_3	7.42	4.27
Emulsion_4	4.92	4.40

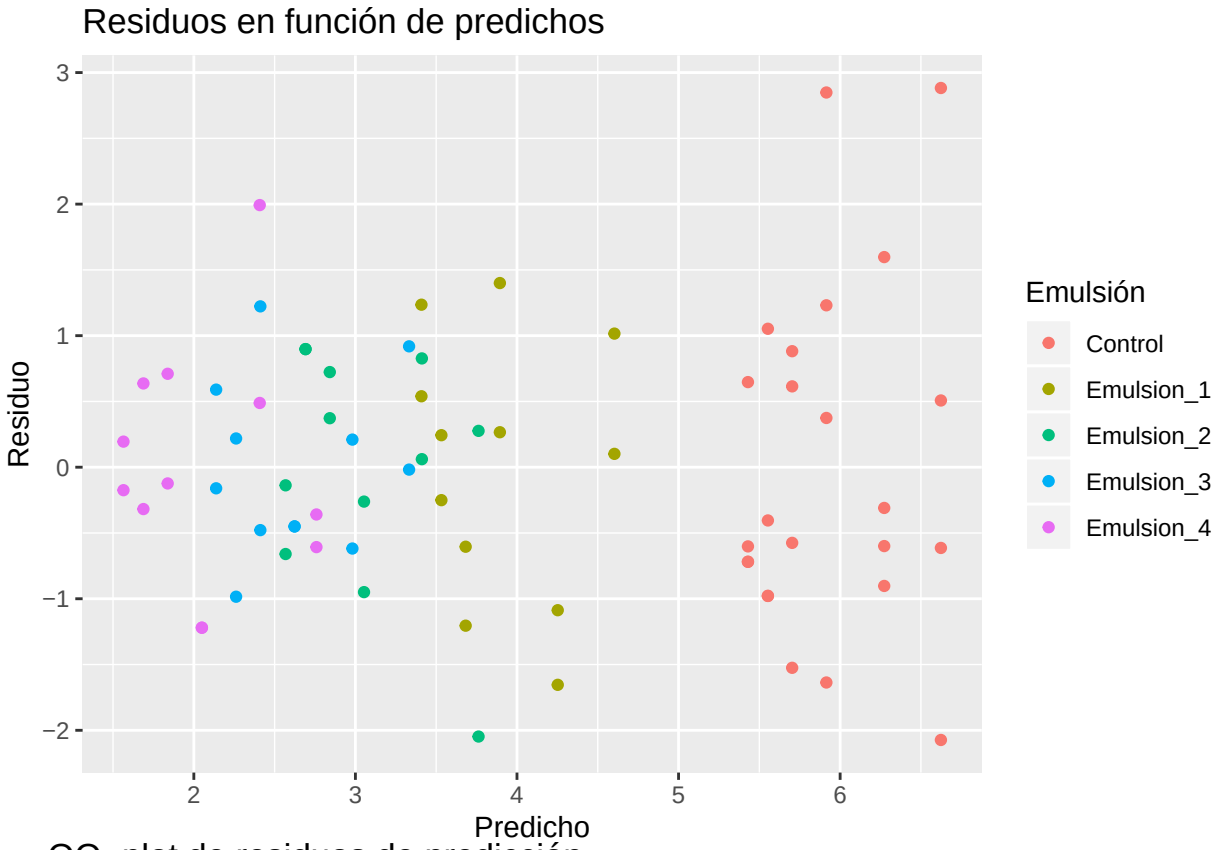
Los desvíos son mas bien disímiles, con lo cual nos convendrá plantear alguna transformación estabilizadora de la varianza. Es razonable aproximar los datos con una expresión de la forma $k - X$, donde k es la cantidad de larvas iniciales, y $X \sim \text{Poisson}(\lambda)$, lo que nos sugiere $g(x) = \sqrt{x}$ como transformación estabilizadora. Despues de esta operación, la tasa del mayor (Control, 1.27) al menor (Emulsión 2, 0.74) desvío resulta ser ≈ 1.72, lo cual nos parece tolerable. Graficamos los perfiles de los datos transformados, junto con los QQ-plot por tratamiento:

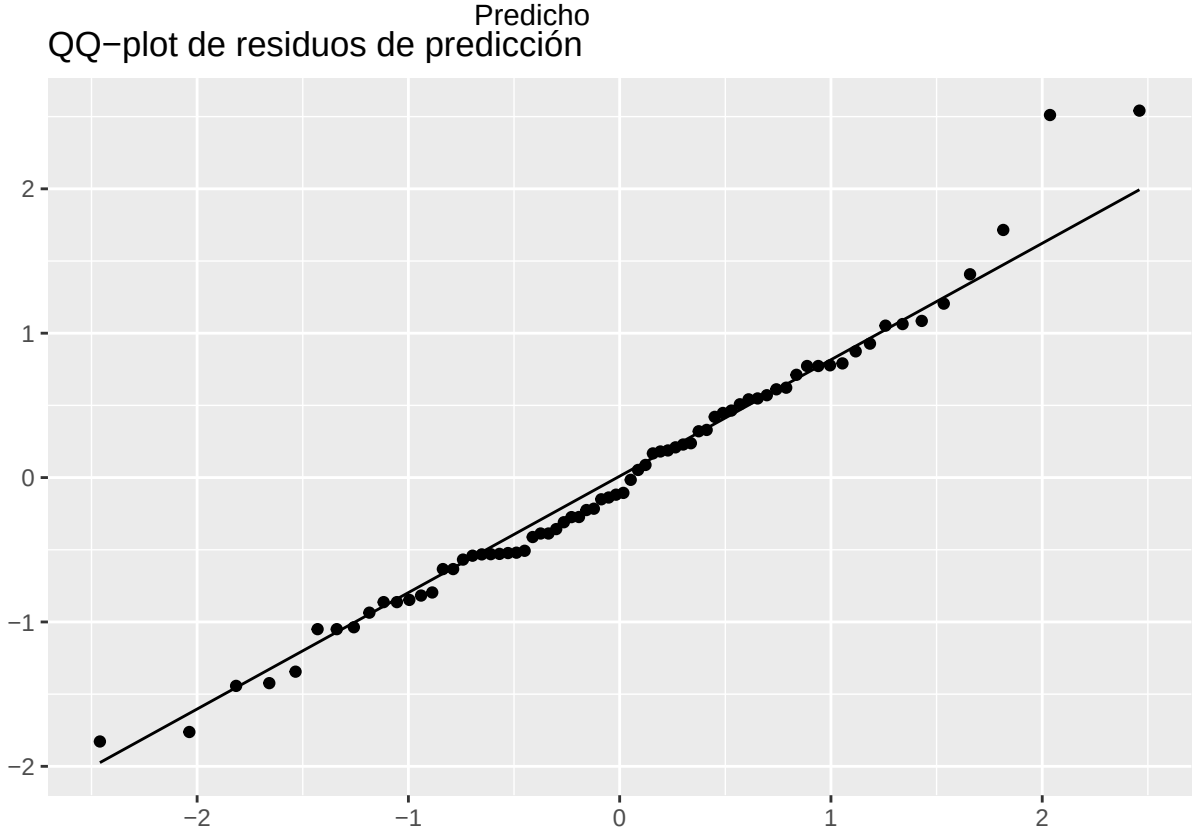
Trat.	Media	Desvio
Control	5.92	1.27
Emulsion_1	3.90	0.90
Emulsion_2	3.05	0.74
Emulsion_3	2.62	0.76
Emulsion_4	2.05	0.88





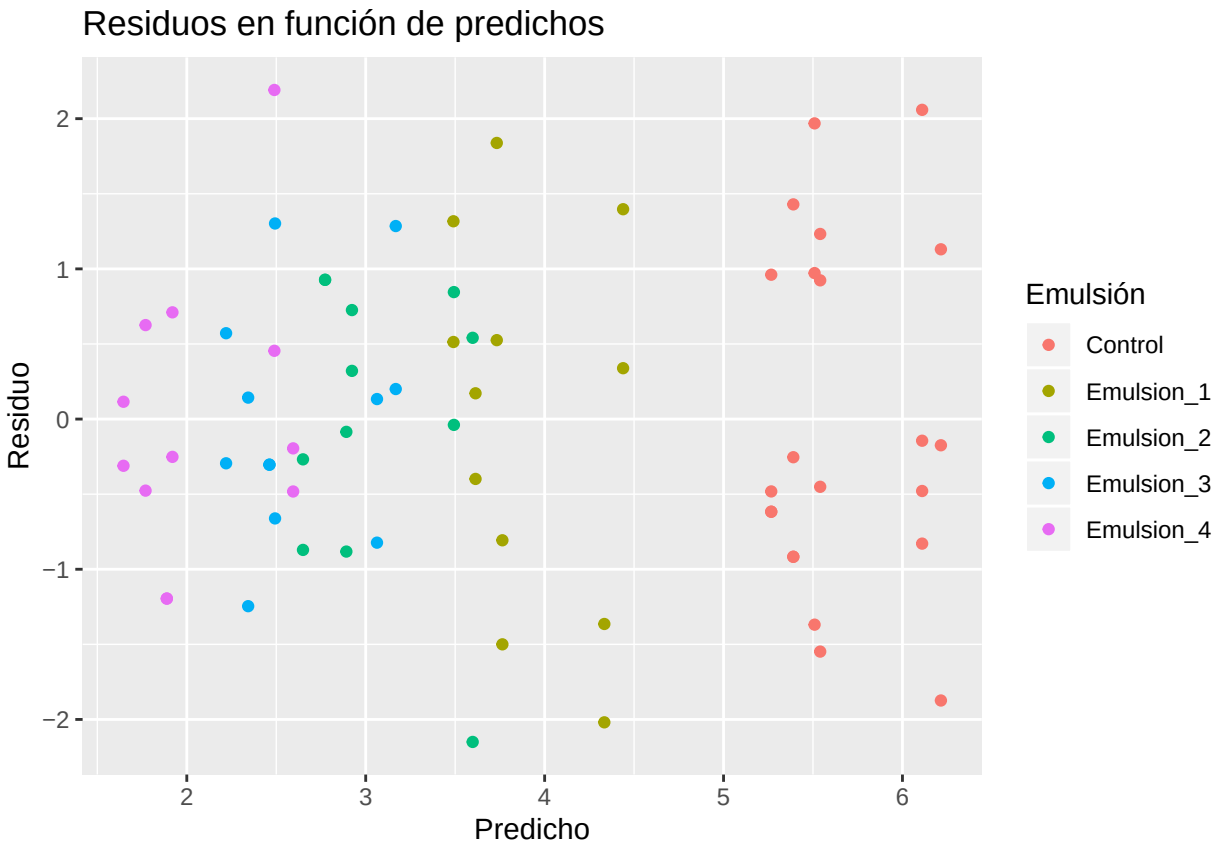
Salvo por una observación atípica para la emulsión 4, los datos correspondientes a las 4 emulsiones tiene una distribución aceptablemente similiar a una normal. Para los controles combinados, una distribución normal no parece demasiado adecuada en los extremos, pero tampoco es aberrante, así que por el momento consideramos que los supuestos de normalidad y homocedasticidad se sostienen razonablemente. Ajustamos un modelo de efectos aditivos del estilo $\sqrt{larvas} \sim bloque + emulsión$, y examinamos los residuos de las predicciones:

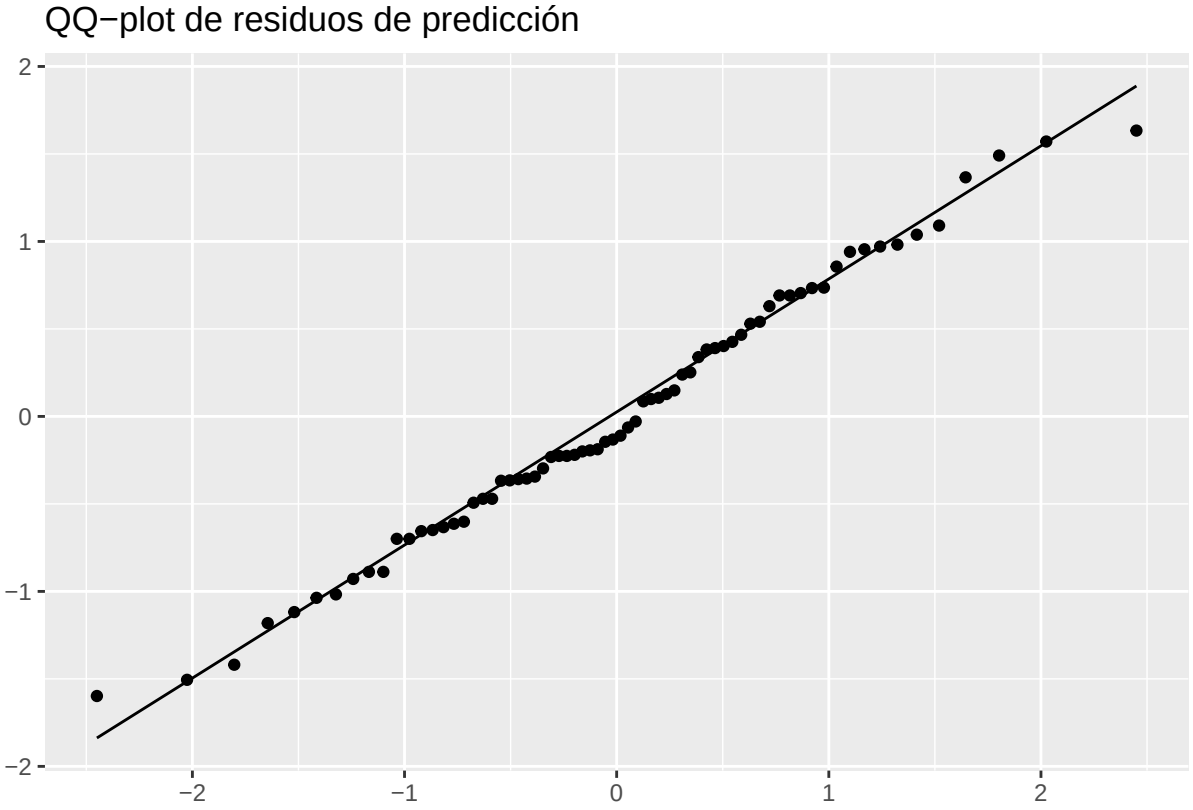




No nos encontramos con ninguna estructura obvia, pero sí observamos un par de *outliers* con residuo mayor a 3 correspondientes a las observaciones con 84 y 71 larvas del Control 1 en los bloques 6 y 4, respectivamente.

Esto era de esperar, ya que como vimos en los QQ-plots por población, el control se alejaba de la normalidad en los extremos. Gracias a haber fusionado los dos controles en uno sólo, podemos descartar ambas observaciones sin dejar de tener un estimador de la varianza para en control en dichos bloques, así que las abandonamos. El ajuste resultante luce fundamentalmente creíble:





R^2 aj.	F obs.	P-valor
0.754	24.5	5.98e-17

Coefs	GL	C. Medios	F obs.	P-valor
blq	5	2.42	2.42	0.0457
trt	4	52	52	7.53e-19
Residuals	60	NA	NA	NA

Coef.	Estimado	P-valor
(Intercept)	5.27	1.8e-27
blq5	0.123	0.71
blq3	0.274	0.411
blq1	0.843	0.0133
blq4	0.242	0.477
blq6	0.948	0.00682
trtEmulsion_1	-1.78	8.25e-08
trtEmulsion_2	-2.62	1.02e-12
trtEmulsion_3	-3.05	3.65e-15
trtEmulsion_4	-3.62	2.92e-18

Se observa inmediatamente que el modelo produce un buen ajuste, y que el efecto de los tratamientos es indudablemente significativo, tanto conjuntamente como individualmente. El efecto de los bloques es significativo al tradicional nivel del 5%, por lo que hemos decidido conservarlo en el modelo, pero obviar los

bloques podría ser una decisión razonable también para mejorar la interpretabilidad de los resultados.

Luego de fusionar los dos controles y eliminar los valores extremos de larvas, los desvíos estimados por tratamiento son fundamentalmente idénticos, lo cual nos permitirá aplicar un test de comparaciones múltiples de Tukey con confianza:

Trat.	Media	Desvio
Control	5.65	0.94
Emulsion_1	3.90	0.90
Emulsion_2	3.05	0.74
Emulsion_3	2.62	0.76
Emulsion_4	2.05	0.88

	Estimado	Lím. inf. IC	Lím. sup. IC	P-valor
Emulsion_1-Control	-1.766	-1.977	-1.555	0.000
Emulsion_2-Control	-2.607	-2.818	-2.396	0.000
Emulsion_3-Control	-3.037	-3.248	-2.826	0.000
Emulsion_4-Control	-3.610	-3.821	-3.400	0.000
Emulsion_2-Emulsion_1	-0.841	-1.081	-0.601	0.095
Emulsion_3-Emulsion_1	-1.271	-1.510	-1.031	0.003
Emulsion_4-Emulsion_1	-1.844	-2.084	-1.605	0.000
Emulsion_3-Emulsion_2	-0.430	-0.670	-0.190	0.692
Emulsion_4-Emulsion_2	-1.003	-1.243	-0.764	0.028
Emulsion_4-Emulsion_3	-0.574	-0.813	-0.334	0.421

A nivel 5%, podemos concluir que

- el tratamiento con cualquier emulsión es efectivo para eliminar larvas,
- la emulsión 1 es significativamete más efectiva que las 3 y 4, pero no que la 2, y
- la emulsión 2 es significativamnte más efectiva que 4, pero no que la 3, y
- las emulsiones 3 y 4 no se distinguen entre sí.

Mas allá de la significatividad estadística, de tener que decantarse por una única opción, nuestra sugerencia sería usar la emulsión 1.