Taller de Consultoria - TP4

Gonzalo Barrera Borla

6/10/2019

Setup

Ejercicio 1

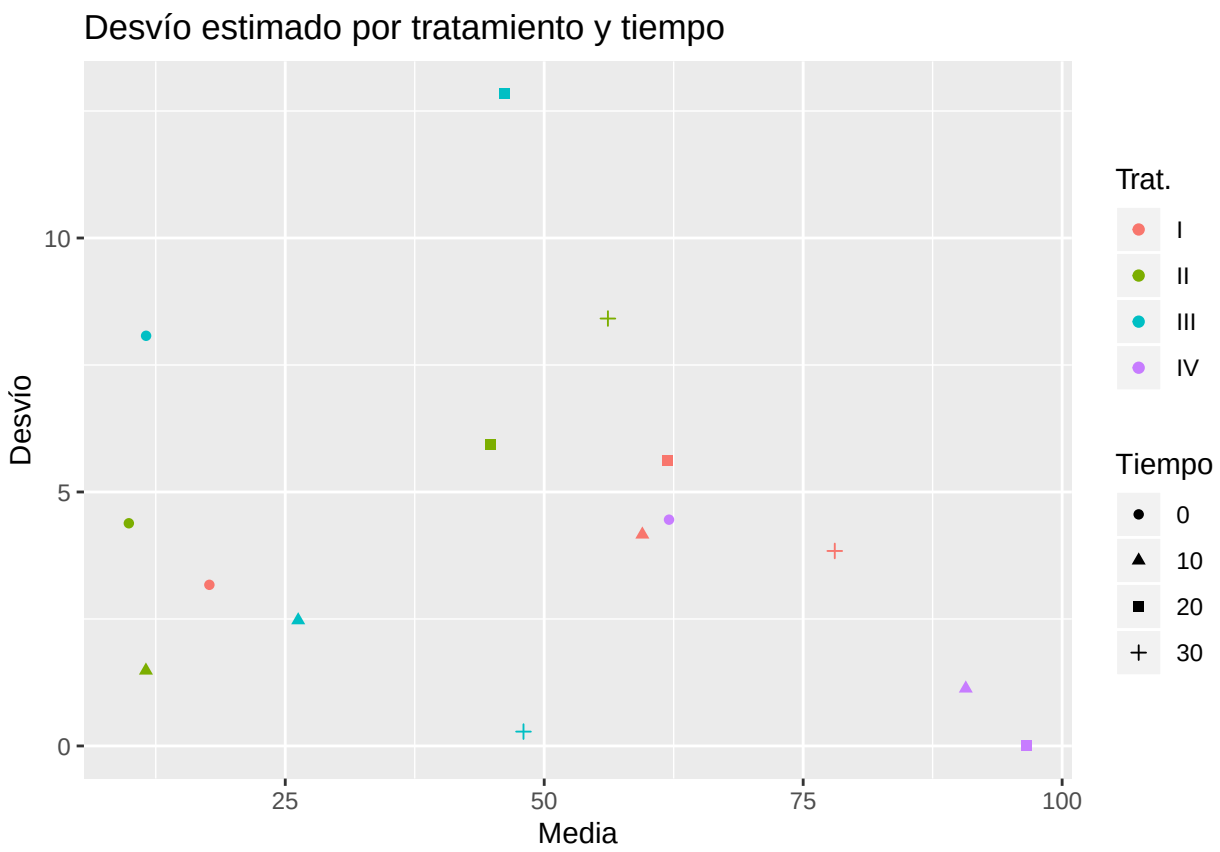
> Se midió el daño al ADN (porcentaje de ADN que se separa durante la electroforesis) en raíces de habas sometidas a 4 tratamientos, aplicados durante diferentes tiempos. Interesa decidir si algún tratamiento es demostrablemente más dañino. Los "-" indican datos no medidos. El tiempo cero para cada uno 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:

trt	t	mean	sd
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

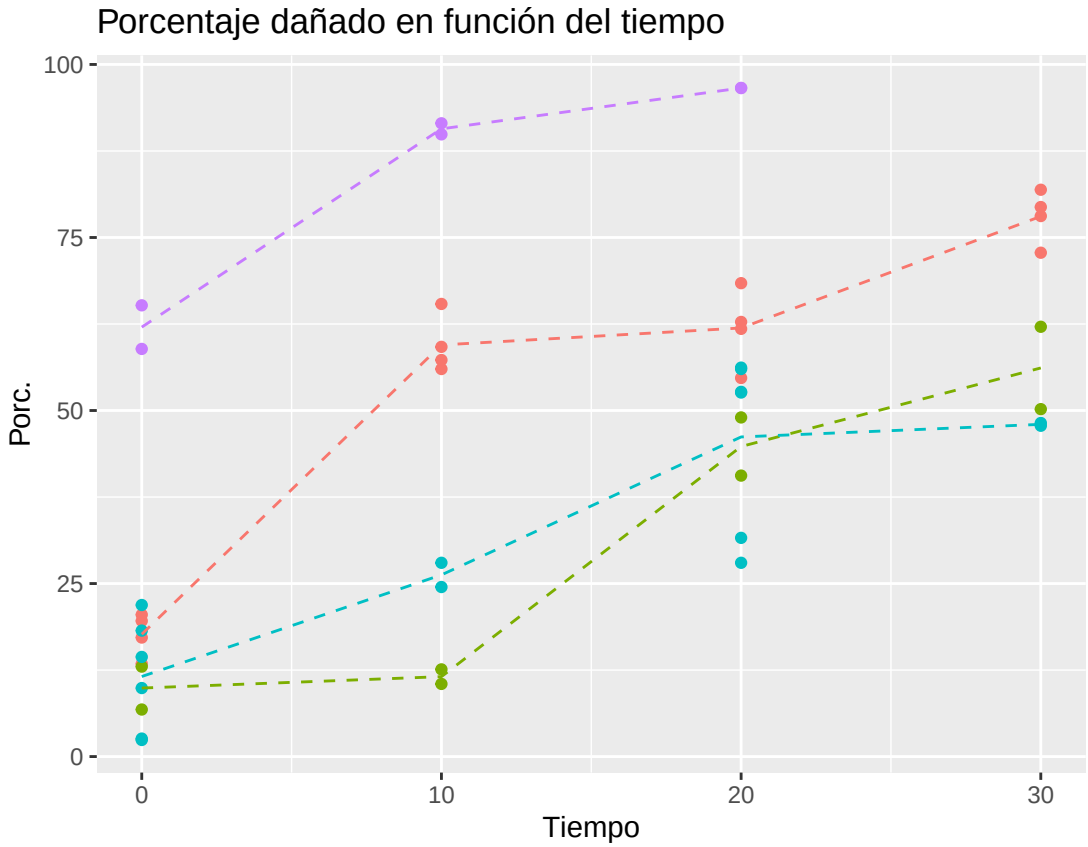
En primer lugar, para el tratamiento IV y $t = 20$, la varianza estimada es exactamente 0, por lo que la razón entre la máxima y mínima varianza para todas las combinaciones de bloque y tratamiento está indefinida, y la regla heurística que plantea Seber [1977, p.195] y utilizaramos en el TP3 no se puede 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 modelo de referencia - (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 mdoelos más complejos, utilizamos el muy recomendable paquete `caret`, que nos permite ajustar modelso con distintas técnicas de resampleo. Para ete ejercicio, optamos por el clásico split entre datos de entrenamiento (70%) y teste (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
p ~ poly(t, 1, raw = TRUE) + trt	9.88	1.23
p ~ poly(t, 3, raw = TRUE) + trt	10.07	1.72

Modelo	ECM	Desvio
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 cuando 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

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

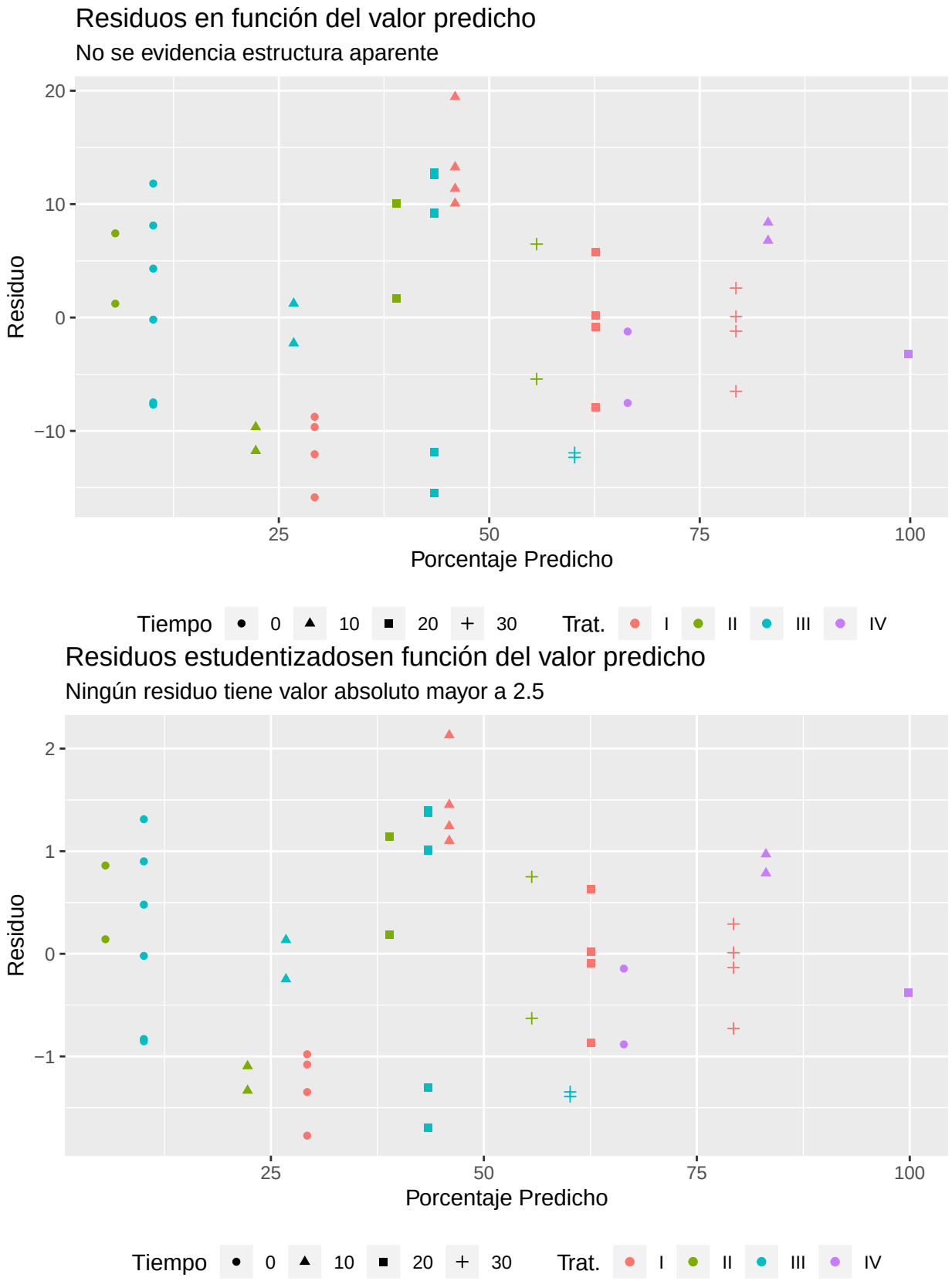
donde μ_i es la ordenada correspondiente al tratamiento i, t es el tiempo de medición y n_i la cantidad de observaciones del i-'esimo tratamiento, *a sabiendas* de que no hemos podido garantizar todavía la homocedasticidad de los ϵ_{ij}. Los principales estadísticos del modelo, efectos principales y 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 (absolutos y estudentizados) de las predicciones, para estimar si hemos capturado satisfactoriamente la estructura de los datos, y no observamos nada particularmente alarmante:



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.

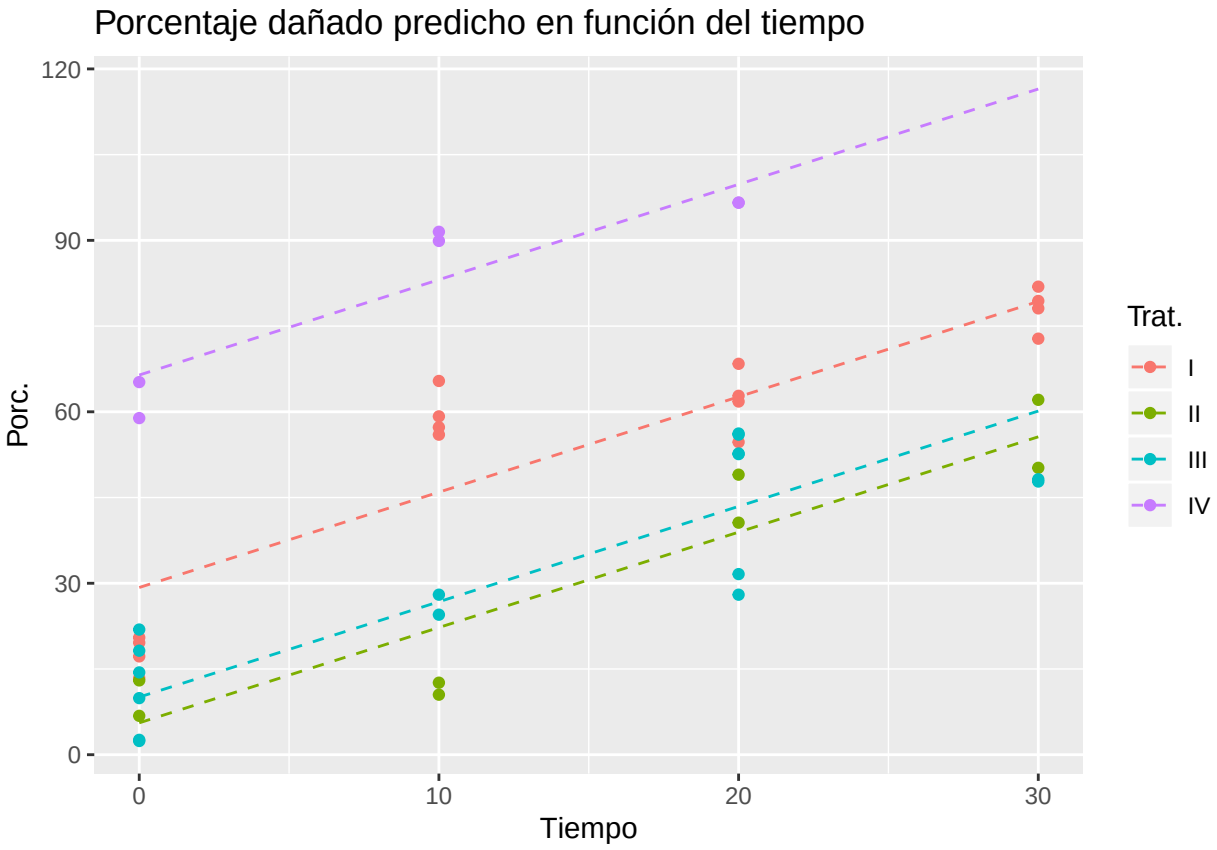
Finalmente, podemos asegurar que

- el porcentaje inicial de ADN dañado varía con el método,
- pero el avance del daño en el tiempo no varía 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:

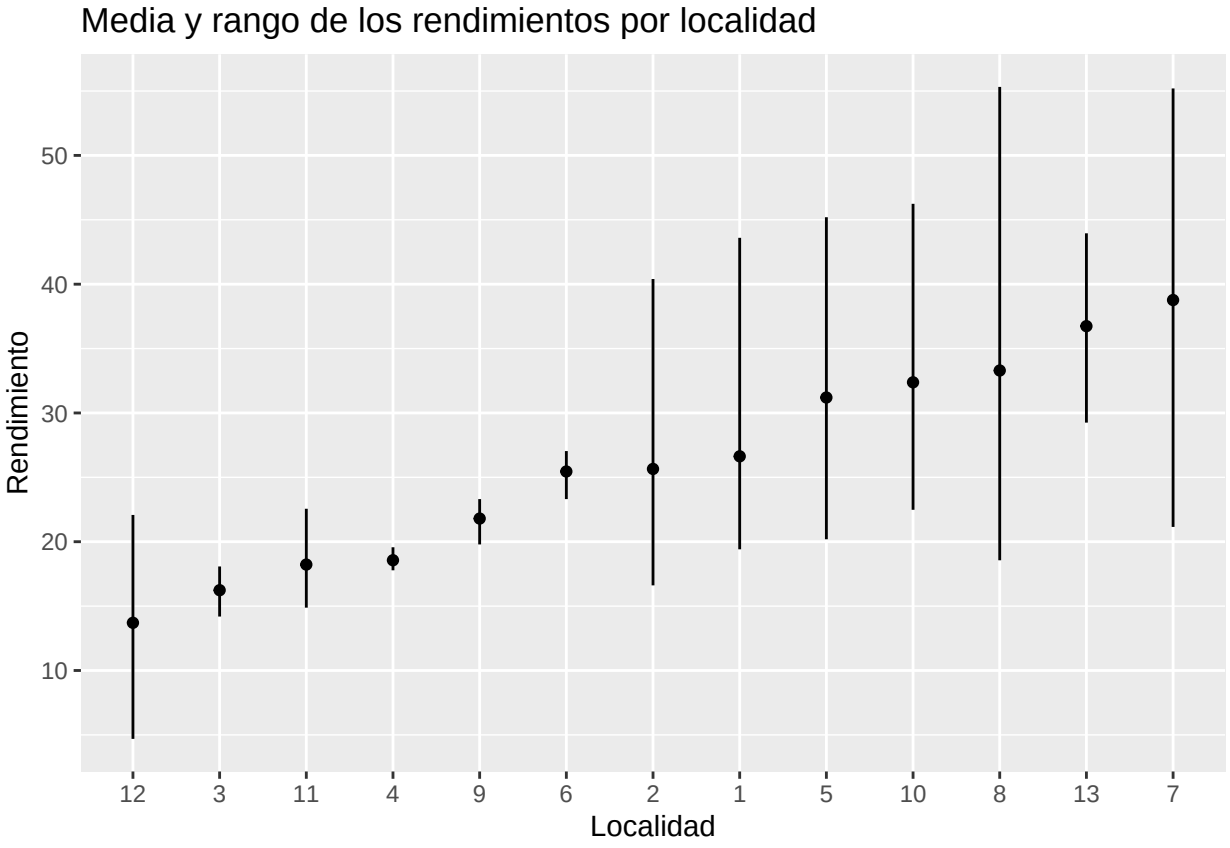


Ejercicio 2

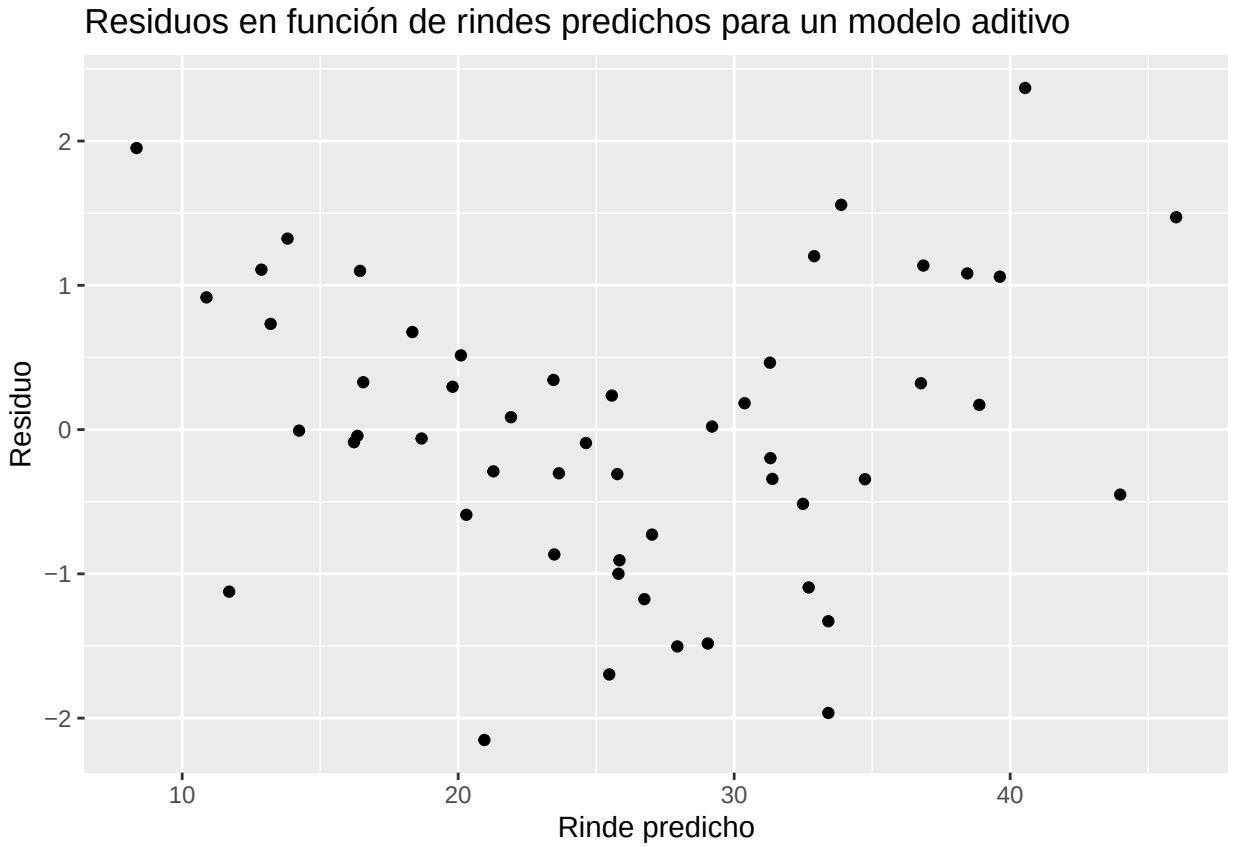
Se dan los rendimientos de 4 variedades de trigo en 13 localidades de Oklahoma. Interesa determinar qué variedades son recomendables.

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:

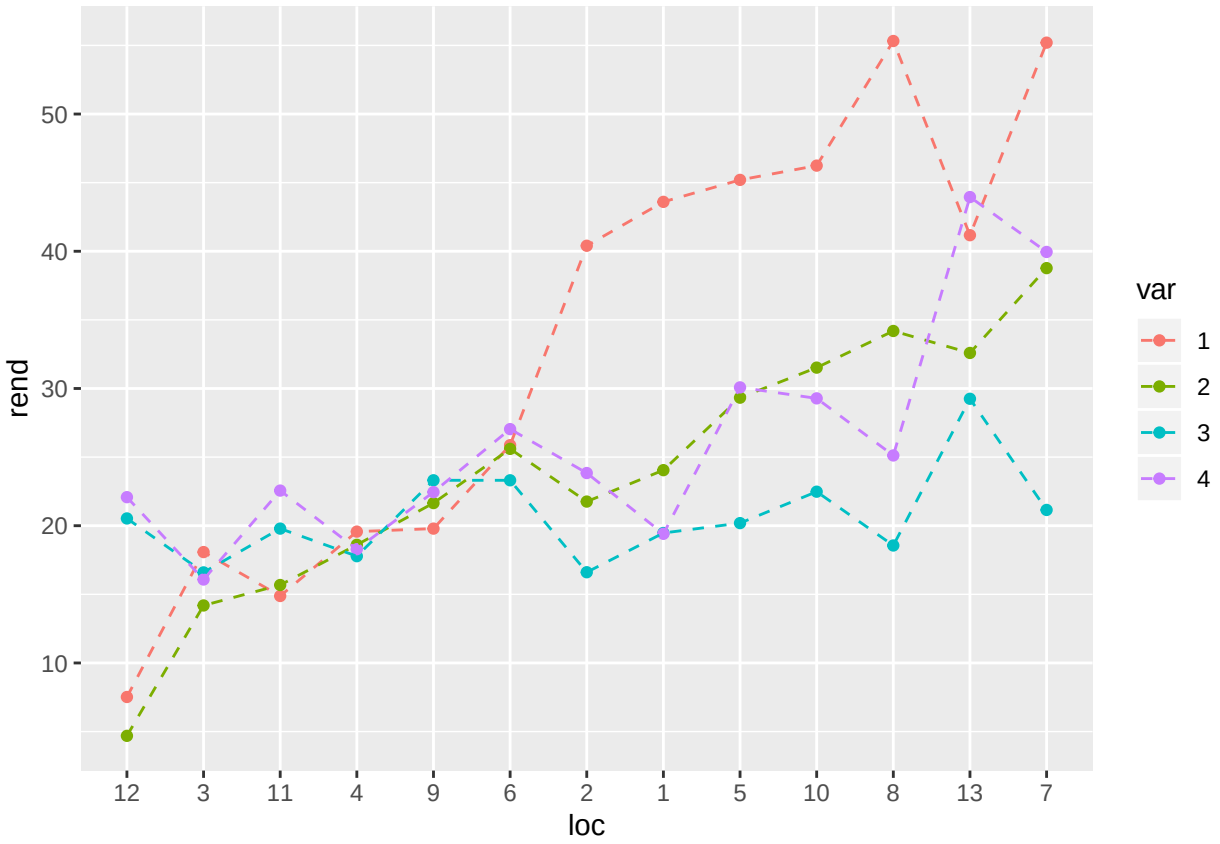
Loc.	Rango	Media
8	36.76	33.30
7	34.05	38.77
5	25.01	31.20
1	24.19	26.63
2	23.79	25.65
10	23.76	32.38
12	17.39	13.70
13	14.70	36.74
11	7.68	18.23
3	3.89	16.24
6	3.73	25.45
9	3.52	21.80
4	1.79	18.56



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 $rinde \sim variedad \mid localidad$, 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 son mínimos. Para ambos se requiere 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, y dentro del bloque θ_j es la mediana del j-ésimo tratamiento. La hipótesis nula será $H_0 : \theta_i = \theta_j \forall i, j$, y la alternativa su complemente (que al menos dos medianas son distintas).

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.