Universidad Nacional de Colombia

Facultad de Ciencias

Análisis de Regresión

Trabajo RLM: Parte II

Autores:

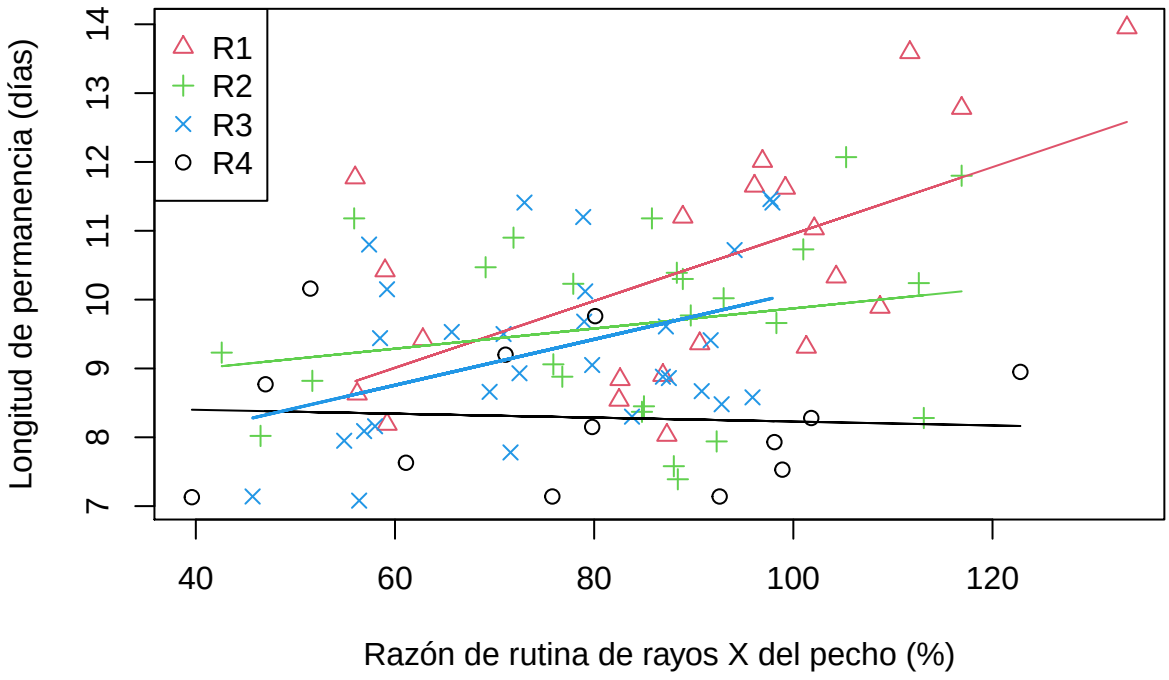
Santiago Franco Valencia
Juan Pablo Martínez Echavarria
Alejandro Salazar Mejía

Agosto 2021

Solución

0. Análisis inicial

Se realiza el gráfico de dispersión Longitud de permanencia promedio vs. Razón de rutina de rayos X del pecho discriminado por cada una de las 4 regiones. Además, por cada región se ajusta una recta a las observaciones respectivas.



En las regiones 1 y 3, parece que la Razón de rutina de rayos X en el pecho y la longitud de permanencia están relacionadas positiva y ligeramente. Para las regiones 2 y 4 no parece que dicha relación exista.

1. Modelo de regresión

Consideremos las siguientes variables:

Y_i : i-ésima observación de la variable respuesta 'Longitud de permanencia' (DPERM).
X_i : i-ésima observación de la variable predictoria 'Razón de rutina de rayos X del pecho' (RRX).
R_{i1} : 1 si la i-ésima observación pertenece a la región 1 = NE, ó 0 en otro caso.
R_{i2} : 1 si la i-ésima observación pertenece a la región 2 = NC, ó 0 en otro caso.
R_{i3} : 1 si la i-ésima observación pertenece a la región 3 = S, ó 0 en otro caso.

Si se espera una diferencia entre las rectas de DPERM VS. RRX que corresponden a las cuatro regiones, el modelo apropiado sería:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 R_{i1} + \beta_3 R_{i2} + \beta_4 R_{i3} + \beta_{1,1} X_i R_{i1} + \beta_{1,2} X_i R_{i2} + \beta_{1,3} X_i R_{i3} + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

2. Modelo Ajustado

En la siguiente tabla se presentan los parámetros estimados, sus errores estándar y los límites inferior y superior de su respectivo I.C. al 95%.

	Estimate	Std. Error	2.5 %	97.5 %
β_0	8.513	1.233	6.059	10.967
β_1	-0.003	0.015	-0.033	0.027
β_2	-2.420	1.736	-5.874	1.034
β_3	-0.104	1.670	-3.428	3.220
β_4	-1.756	1.732	-5.202	1.690
$\beta_{1,1}$	0.051	0.020	0.012	0.091
$\beta_{1,2}$	0.017	0.020	-0.022	0.057
$\beta_{1,3}$	0.036	0.022	-0.007	0.080

La ecuación ajustada sería:

$$\widehat{Y_i} = 8.513363 - 0.002846 X_i - 2.420080 R_{i1} - 0.103785 R_{i2} - 1.755731 R_{i3} + 0.051447 X_i R_{i1} + 0.017477 X_i R_{i2} + 0.036165 X_i R_{i3}$$

Recta para cada región:

- **Región 1:** $R_{i1} = 1, R_{i2} = 0, R_{i3} = 0$
  El modelo sería:
  $$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1}) X_i + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
  Ecuación ajustada:
  $$\widehat{Y_i} = 8.513363 - 0.002846 X_i - 2.420080 + 0.051447 X_i = 6.093283 + 0.048601 X_i$$

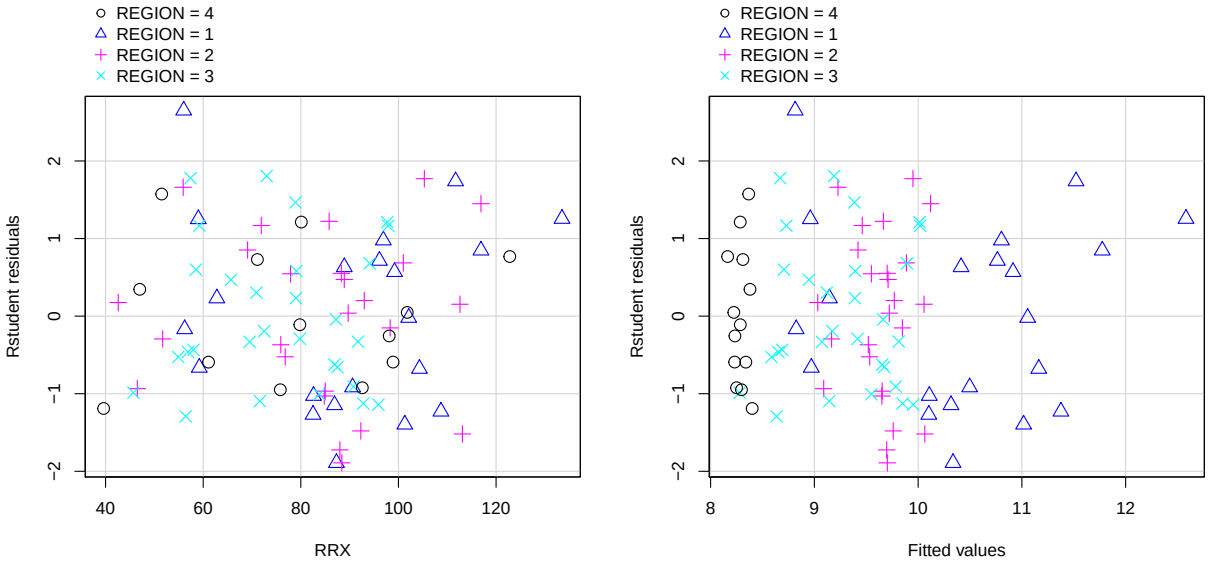
- **Región 2:** $R_{i1} = 0, R_{i2} = 1, R_{i3} = 0$
  El modelo sería:
  $$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2}) X_i + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
  Ecuación ajustada:
  $$\widehat{Y_i} = 8.513363 - 0.002846 X_i - 0.103785 + 0.017477 X_i = 8.409578 + 0.014631 X_i$$

- **Región 3:** $R_{i1} = 0, R_{i2} = 0, R_{i3} = 1$
  El modelo sería:
  $$Y_i = (\beta_0 + \beta_4) + (\beta_1 + \beta_{1,3})X_i + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
  Ecuación ajustada:
  $$\widehat{Y_i} = 8.513363 - 0.002846X_i - 1.755731 + 0.036165X_i = 6.757632 + 0.033319X_i$$

- **Región 4:** $R_{i1} = 0, R_{i2} = 0, R_{i3} = 0$
  El modelo sería:
  $$Y_i = \beta_0 + \beta_1 X_i + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
  Ecuación ajustada:
  $$\widehat{Y_i} = 8.513363 - 0.002846X_i$$

3. Supuestos de normalidad y varianza constante

El supuesto de varianza constante se analiza mediante los residuales para el modelo general. En la siguientes dos gráficas se muestran los residuales estudentizados vs. RRX y vs. valores ajustados.



Observando los puntos conjuntamente, no hay patrones evidentes para creer que no se cumple el supuesto de varianza constante en los errores.

En la Figura 1 podemos observar la distribución de los residuales discriminada por regiones, por medio de diagramas de cajas y pigotes. Se encuentra que la distribución de los residuales estudentizados es similar entre todas las regiones exceptuando la región 1 en donde se encuentran más dispersos.

Se encuentra el estadístico W_0 de la prueba de Shapiro-Wilk asociada a la normalidad de los errores del modelo, y se evalúa la siguiente prueba de hipótesis:

$$H_0 : E_i \sim N(0, \sigma^2) \ \ vs. \ \ H_1 : E_i \nsim N(0, \sigma^2)$$

$$i = 1, ..., 88$$

En la Figura 2 se observa el estadístico W, su valor-p y la gráfica de normalidad. Se puede asumir que los errores son normales, sin embargo es evidente un ligero desajuste al inicio y al final de la recta.

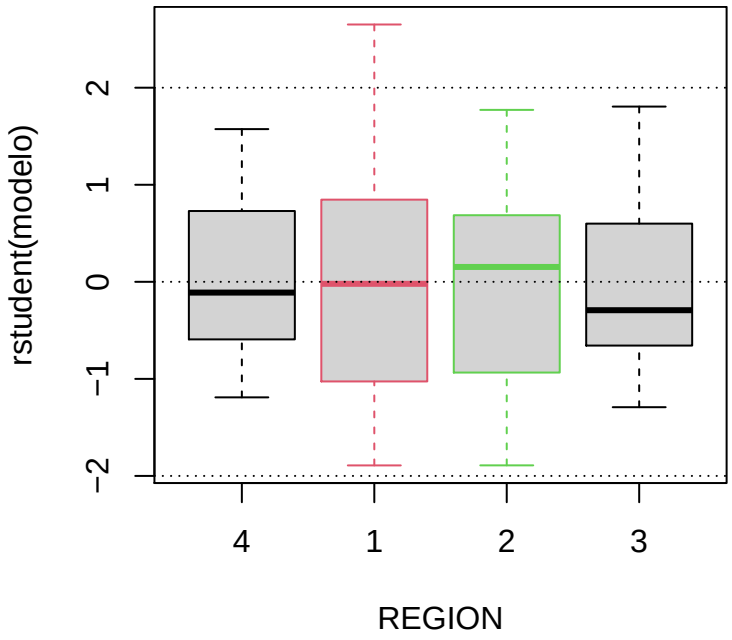


Figure 1: Distribución de residuales por región

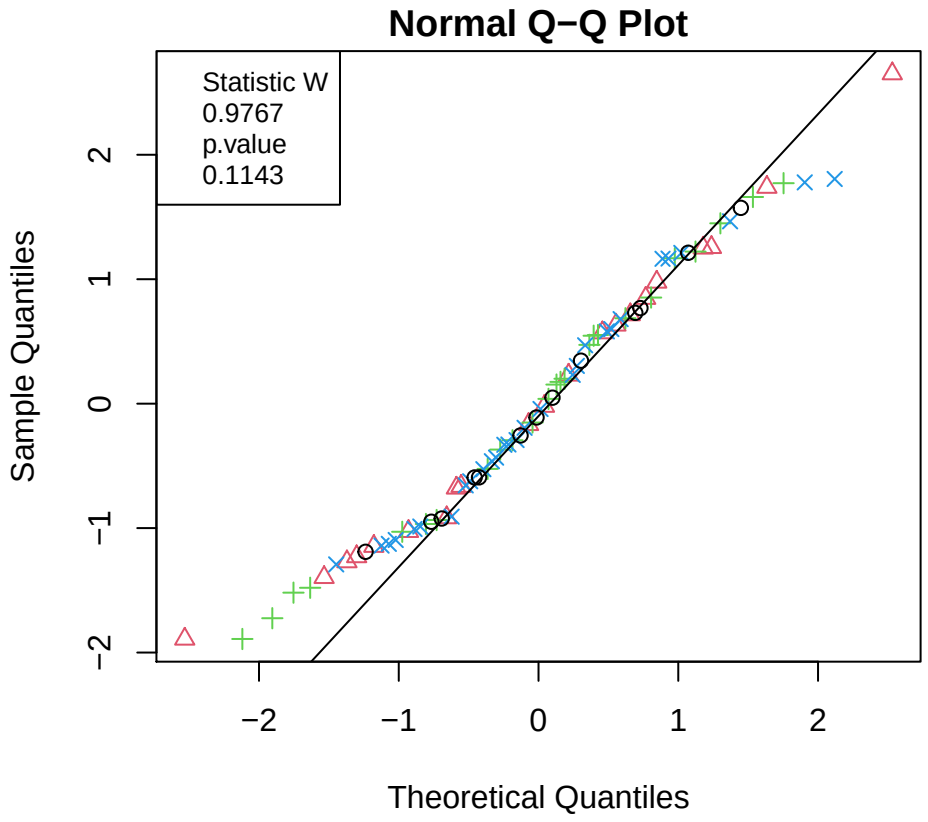


Figure 2: Gráfica de Normalidad

4. Diferencia entre las ordenadas en el origen

Las ordenadas al origen son los interceptos. Dadas las ecuaciones por cada región encontradas en el Punto 2, determinar si existe diferencia entre las ordenadas en el origen de las rectas correspondientes a las regiones es equivalente a probar:

$$\beta_0 + \beta_2 = \beta_0 + \beta_3 = \beta_0 + \beta_4 = \beta_0 \Leftrightarrow \beta_2 = \beta_3 = \beta_4 = 0$$

Para esto basta con realizar una prueba de significancia simultanea para probar $H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \;\; vs. \;\; H_1 : \beta_i \neq 0$, para algún $i = 2, 3, 4$.

Table 2: Tabla Resumen de la Prueba de Hipótesis

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Modelo reducido	83	134				
Modelo completo	80	129	3	4.88	1.01	0.394

Como $V_p > 0.05$, no se rechaza H_0, es decir,no existe diferencia entre las ordenadas en el origen de las rectas correspondientes a las regiones.

5. Diferencia en las pendientes de las rectas.

Dadas las ecuaciones por cada región encontradas en el Punto 2, determinar si existe diferencia en las pendientes de las rectas correspondientes a las regiones es equivalente a probar:

$$\beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2} = \beta_1 + \beta_{1,3} = \beta_1 \Leftrightarrow \beta_{1,1} = \beta_{1,2} = \beta_{1,3} = 0$$

Para esto basta con realizar una prueba de significancia simultanea para probar $H_0 : \beta_{1,1} = \beta_{1,2} = \beta_{1,3} = 0 \;\; vs. \;\; H_1 : \beta_{1,i} \neq 0$, para algún $i = 1, 2, 3$.

Table 3: Tabla Resumen de la Prueba de Hipótesis

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Modelo reducido	83	141				
Modelo completo	80	129	3	12.1	2.49	0.066

Como $V_p > 0.05$, no se rechaza H_0, es decir, no existe diferencia en las pendientes de las rectas correspondientes a las regiones. Por lo que se asume que la taza de cambio en la longitud de permanencia de un paciente dado una unidad de cambio en la razón de rutina de rayos X en el pecho es la misma para cada región.

6. Diferencia entre rectas para cada región

Las rectas serán iguales si coinciden sus interceptos y sus pendientes, entonces se requiere que:
$\beta_0 + \beta_2 = \beta_0 + \beta_3 = \beta_0 + \beta_4 = \beta_0 \Leftrightarrow \beta_2 = \beta_3 = \beta_4 = 0$.
También que: $\beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2} = \beta_1 + \beta_{1,3} = \beta_1 \Leftrightarrow \beta_{1,1} = \beta_{1,2} = \beta_{1,3} = 0$

Luego se debe probar que:

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_{1,1} = \beta_{1,2} = \beta_{1,3} = 0 \;\; vs.$$

$$H_1 : \text{ al menos uno de los parametros } \beta_2, \beta_3, \beta_4, \beta_{1,1}, \beta_{1,2}, \beta_{1,3} \text{ es no nulo}$$

Table 4: Tabla Resumen de la Prueba de Hipótesis

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Modelo reducido	86	169				
Modelo completo	80	129	6	40	4.14	0.001

- Modelo reducido (MR):
$$Y_i = \beta_0 + \beta_1 X_i + E_i \;,\; E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
- Modelo completo (MC):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 R_{i1} + \beta_3 R_{i2} + \beta_4 R_{i3} + \beta_{1,1} X_i R_{i1} + \beta_{1,2} X_i R_{i2} + \beta_{1,3} X_i R_{i3} + E_i \;,\; E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- $SSR(R_{i1}, R_{i2}, R_{i3}, X_i R_{i1}, X_i R_{i2}, X_i R_{i3} | X_i) = SSE(MR) - SSE(MC) = 169.01 - 128.96 = 40.047$
- $gl(SSE(MR)) = n - (k+1) = 88 - (1+1) = 86$
- $gl(SSE(MC)) = n - (k+1) = 88 - (7+1) = 80$
- $gl(SSR(R_{i1}, R_{i2}, R_{i3}, X_i R_{i1}, X_i R_{i2}, X_i R_{i3} | X_i)) = gl(SSE(MR)) - gl(SSE(MC)) = 6$
- $MSE(MC) = \frac{SSE(MC)}{gl(SSE(MC))} = \frac{128.96}{80} = 1.612$
- Estadístico de prueba:

$$F_0 = \frac{SSR(R_{i1}, R_{i2}, R_{i3}, X_i R_{i1}, X_i R_{i2}, X_i R_{i3} | X_i) / [gl(SSE(MR)) - gl(SSE(MC))]}{MSE(MC)}$$

$$= \frac{40.047/6}{1.612} = 4.1404$$

- Región crítica al nivel de 0.05:

Teniendo en cuenta que es una prueba de cola superior, se rechaza H_0 si el estadístico de prueba F_0 está en la región crítica dada por $(F_{0.05,6,80}, +\infty)$, donde $F_{0.05,6,80}$ es el cuantil de una distribución f con 6 grados de libertad en el numerador, 80 en el denominador y que deja un área a la derecha de 0.05.
Haciendo los cálculos respectivos tenemos que:

$$F_0 \in (F_{0.05,6,80}, +\infty) \Longleftrightarrow 4.1404 \in (2.2141, +\infty)$$

Se rechaza H_0 ya que F_0 pertenece a la región critica, concluyendo así que para al menos una región, la recta que modela la longitud de permanencia media con la razón de rutina de rayos X en el pecho es diferente al resto, estadísticamente.

7. Efecto medio de RRX sobre DPERM en las regiones 2, 3, y 4.

Probar si el efecto medio de RRX sobre DPERM es igual en las regiones 2, 3, y 4 implica igualdad en las pendientes de cada recta correspondiente del Punto 2. En términos de los coeficientes esto se traduce a probar si $\beta_1 + \beta_{1,2} = \beta_1 + \beta_{1,3} = \beta_1 \Leftrightarrow \beta_{1,2} = \beta_{1,3} = 0$.
Luego el juego de hipótesis será:

$$H_0 : \beta_{1,2} = \beta_{1,3} = 0 \;\; vs. \;\; H_1 : \beta_{1,i} \neq 0, \text{ para algun } i = 2, 3$$

Table 5: Tabla Resumen de la Prueba de Hipótesis

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Modelo reducido	82	133				
Modelo completo	80	129	2	4.43	1.38	0.259

- Modelo reducido (MR):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 R_{i1} + \beta_3 R_{i2} + \beta_4 R_{i3} + \beta_{1,1} X_i R_{i1} + E_i \;,\; E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Modelo completo (MC):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 R_{i1} + \beta_3 R_{i2} + \beta_4 R_{i3} + \beta_{1,1} X_i R_{i1} + \beta_{1,2} X_i R_{i2} + \beta_{1,3} X_i R_{i3} + E_i \;,\; E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Estadístico de prueba:

$$F_0 = \frac{SSR(X_i R_{i2}, X_i R_{i3} | X_i, R_{i1}, R_{i2}, R_{i3}, X_i R_{i1}) / [gl(SSE(MR)) - gl(SSE(MC))]}{MSE(MC)}$$

$$= \frac{[133.40 - 128.96]/[82 - 80]}{128.96/80} = 1.3755$$

Como $V_p > 0.05$, no rechazamos H_0. Se concluye entonces que no existe evidencia para rechazar que las pendientes asociadas a las rectas de las regiones 2, 3 y 4 son iguales para cada región, es decir, que es posible asumir que el efecto medio de la razón de rutina de rayos X sobre la longitud de permanencia es el mismo para las regiones 2, 3 y 4, estadísticamente.

Ajuste del modelo reducido:

En la siguiente tabla se presentan los parámetros estimados del modelo reducido, sus errores estándar y los límites inferior y superior de su respectivo I.C. al 95%.

	Estimate	Std. Error	2.5 %	97.5 %
β_0	7.152	0.748	5.663	8.641
β_1	0.014	0.008	-0.002	0.031
β_2	-1.059	1.437	-3.918	1.800
β_3	1.269	0.439	0.396	2.141
β_4	1.029	0.426	0.181	1.877
$\beta_{1,1}$	0.034	0.016	0.003	0.065

La ecuación ajustada sería:

$$\widehat{Y_i} = 7.152320 + 0.014497X_i - 1.059037R_{i1} + 1.268506R_{i2} + 1.028843R_{i3} + 0.034104X_iR_{i1}$$

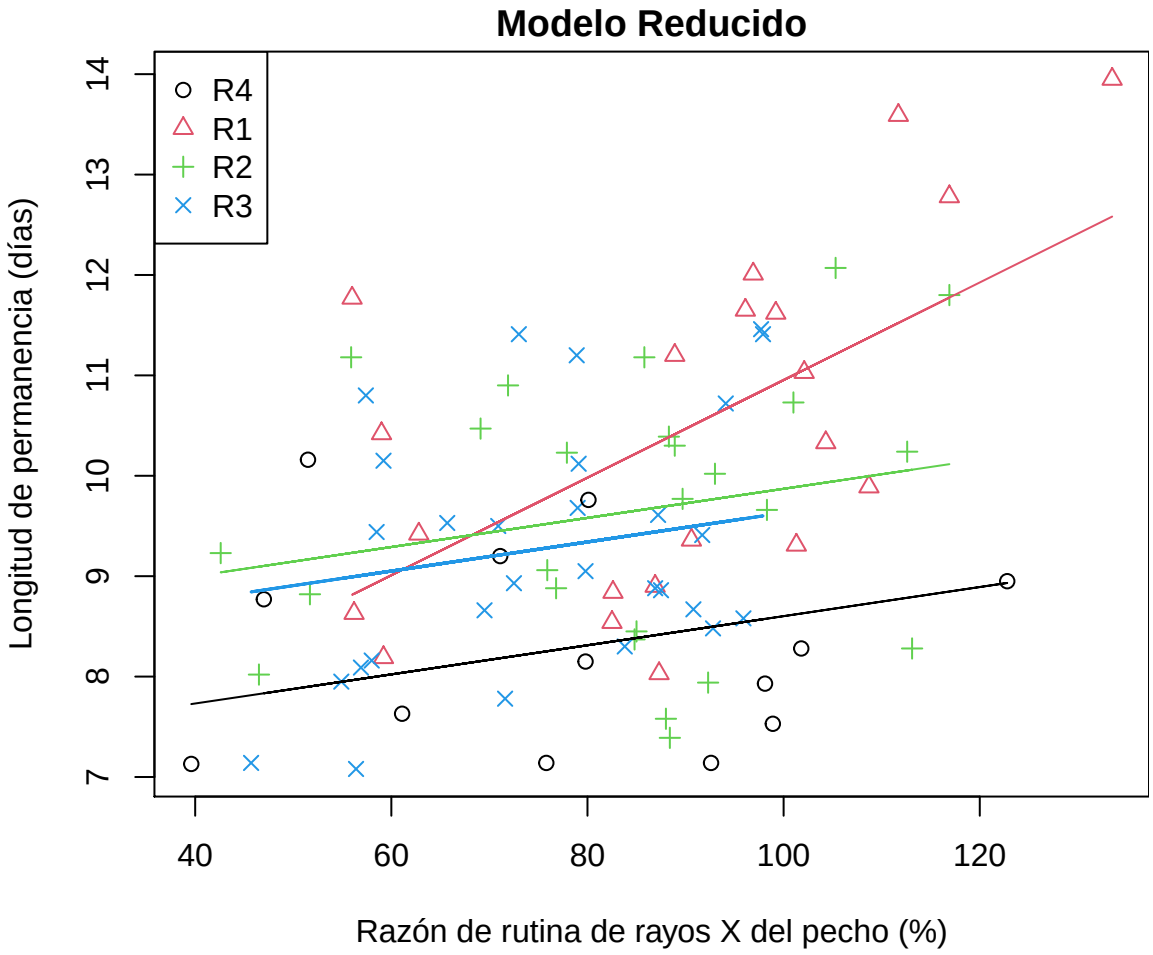
Recta para cada región:

- **Región 1:** $R_{i1} = 1, R_{i2} = 0, R_{i3} = 0$
  El modelo sería:
  $$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1})X_i + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
  Ecuación ajustada:
  $$\widehat{Y_i} = 7.152320 + 0.014497X_i - 1.059037 + 0.034104X_i = 6.093283 + 0.048601X_i$$

- **Región 2:** $R_{i1} = 0, R_{i2} = 1, R_{i3} = 0$
  El modelo sería:
  $$Y_i = (\beta_0 + \beta_3) + \beta_1 X_i + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
  Ecuación ajustada:
  $$\widehat{Y_i} = 7.152320 + 0.014497X_i + 1.268506 = 8.420826 + 0.014497X_i$$

- **Región 3:** $R_{i1} = 0, R_{i2} = 0, R_{i3} = 1$
  El modelo sería:
  $$Y_i = (\beta_0 + \beta_4) + \beta_1 X_i + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
  Ecuación ajustada:
  $$\widehat{Y_i} = 7.152320 + 0.014497X_i + 1.028843 = 8.181163 + 0.014497X_i$$

- **Región 4:** $R_{i1} = 0, R_{i2} = 0, R_{i3} = 0$
  El modelo sería:
  $$Y_i = \beta_0 + \beta_1 X_i + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
  Ecuación ajustada:
  $$\widehat{Y_i} = 7.152320 + 0.014497X_i$$



En esta figura se observa el modelo reducido ajustado, en dónde las pendientes asociadas a las regiones 2, 3 y 4 son iguales, pero sus interceptos son diferentes.