Instituto Tecnológico Autónomo de México

Econometría Aplicada I

Tarea 2

Profesor: Arturo A. Aguilar Esteva

Alumno: Marco Antonio Ramos Juárez

142244

Ciudad de México 2020

Índice

I.

El primer paso es conocer los datos. Por ello, en primer lugar, presentamos una tabla de resumen general. En segundo lugar, para contestar el inciso, con base en la información del primer cuadro, generamos una tabla con la media, desviación estandar, valor mínimo y máximo de las variables númericas (sin la variable *gdp_pc_er*).

Cuadro 1. Características de los datos

Variable	Clase	NAs
iso_code	character	0
country	character	0
continent	character	0
confirmed	numeric	0
confirmed_per_mil	numeric	0
deaths	numeric	0
deaths_per_mil	numeric	0
tests_performed	numeric	0
gdp_pc	numeric	3
median_age	numeric	0
aged_65_older	numeric	1
diab_prev	numeric	1
cardio_dr	numeric	0
hosp_beds_per_thou	numeric	6
hdi	numeric	1
overwgh_prev	numeric	1
gdp_pc_er	numeric	3

Cuadro 2. Características de las variables numéricas

Variable	Media	DE	Max	Min
confirmed	292100.63	962088.16	6519979.00	32.00
confirmed_per_mil	5401.69	6930.77	42255.28	20.91
deaths	8923.70	27185.10	194079.00	2.00
deaths_per_mil	139.99	195.39	856.37	0.29
tests_performed	4225910.28	12550338.74	96786798.00	9076.00
gdp_pc	29491.88	24022.83	114481.53	1059.72

Variable	Media	DE	Max	Min
median_age	34.13	8.84	48.20	16.40
aged_65_older	11.65	7.00	28.00	1.16
diab_prev	7.37	3.54	19.90	1.80
cardio_dr	220.10	105.61	539.85	79.37
hosp_beds_per_thou	3.49	2.73	13.05	0.30
hdi	0.78	0.13	0.95	0.47
overwgh_prev	51.68	17.33	72.10	18.10

II.

Antes que nada, creo la variable *test_per_mil* con el siguiente código:

```
main_data<- mutate(main_data, test_per_mil=
main_data$tests_performed*main_data$confirmed_per_mil/main_data$confirmed)
```

(a)

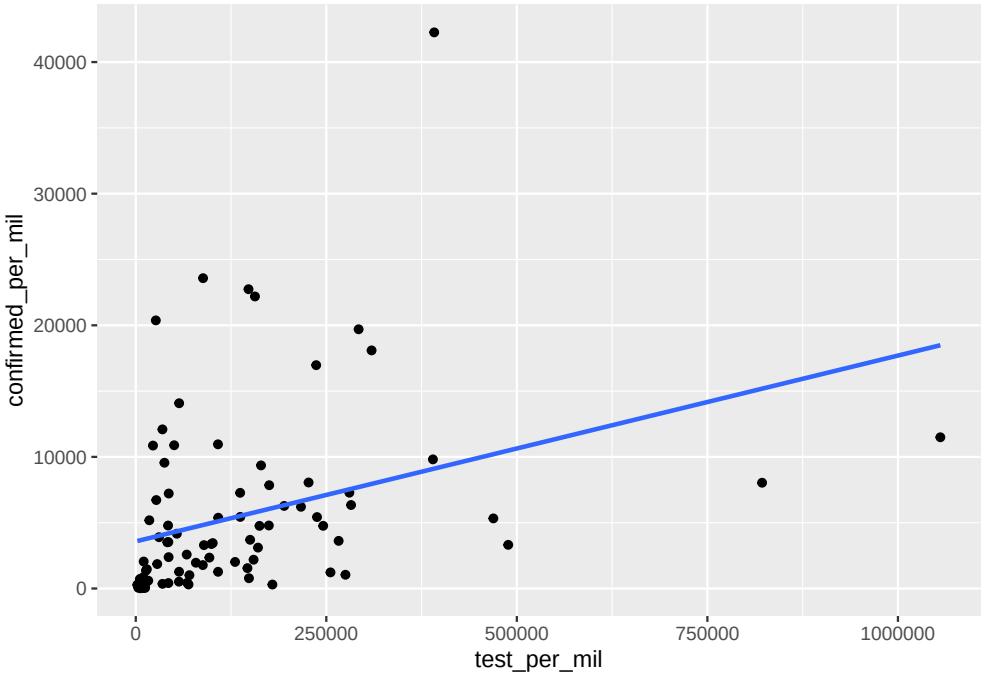

Figura 1. Relación entre pruebas y casos confirmados por millón

Ecuación de regresión:

$$\hat{confirmed\_per\_mil} = 3579.55 + 0.0141 * test\_per\_mil$$

.....(880.80)...(0.00419).....

El coeficiente de la regresión *confirmed_per_mil* ~ *test_per_mil* es igual a .0141. Esto lo podemos interpretar como un aumento de uno en el número de pruebas por cada millón habitantes tiene un impacto positivo de .0141 en los casos confirmados por cada millón.

(b)

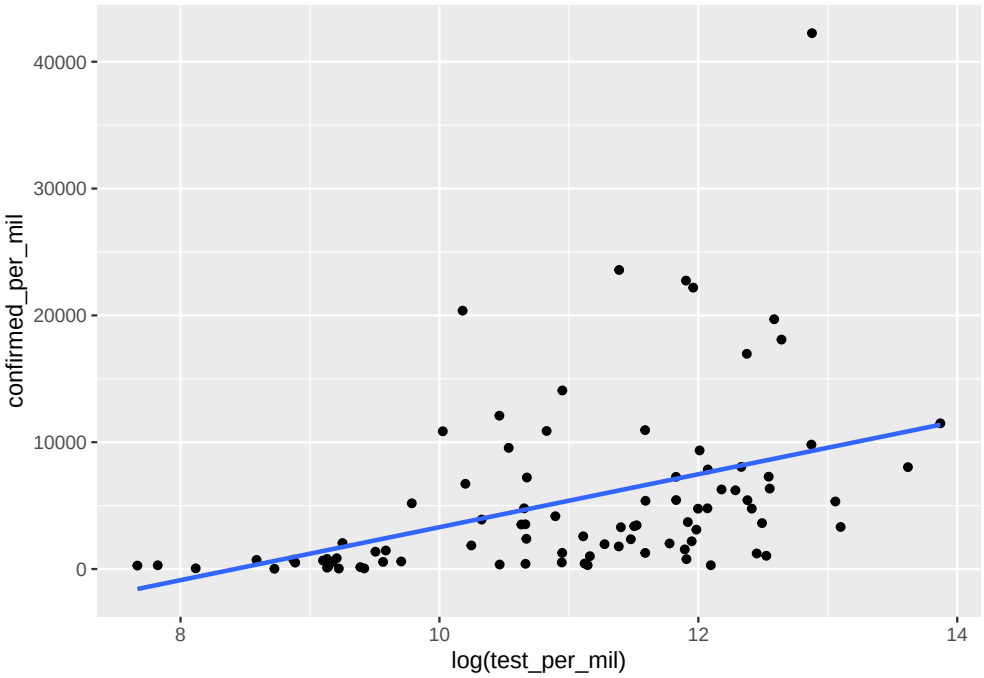

Figura 2. Relación entre pruebas (log) y casos confirmados por millón

Ecuación de regresión:

$$\hat{confirmed\_per\_mil} = -17582 + 2088 * log(test\_per\_mil)$$

.....(5391)...(486).....

En este caso, un aumento de 1 en el log(*test_per_mil*) tiene un impacto positivo de 2088 en *confirmed_per_mil*. Esto lo podemos interpretar como un cambio del 1 % en las pruebas por millón tiene un impacto positivo de 20.88 en los casos confirmados por cada millón (o bien).

(c)

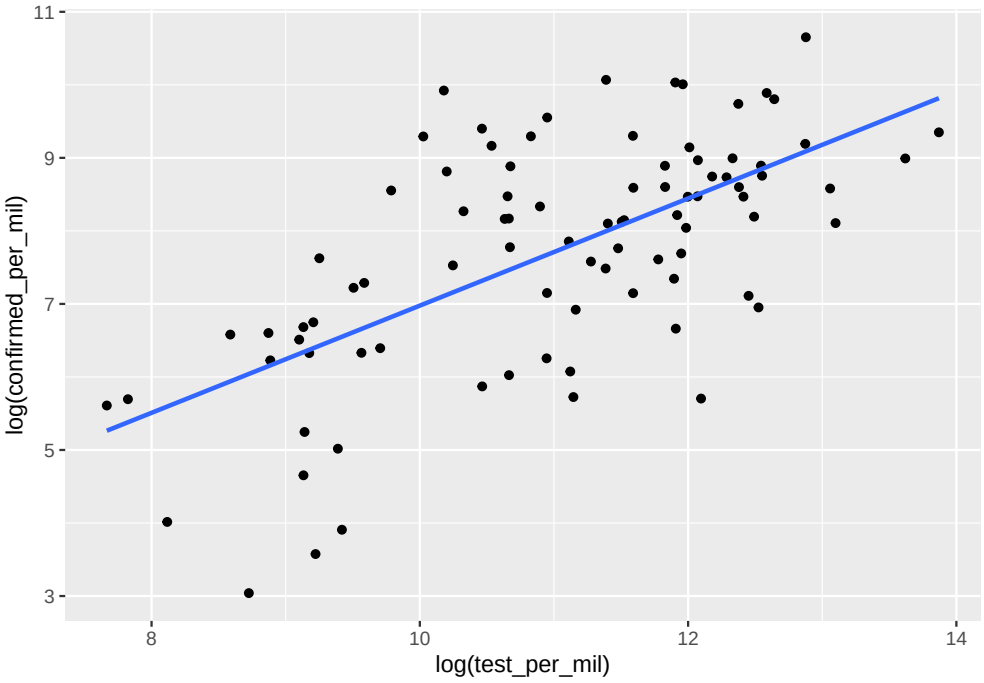

Figura 3. Relación entre pruebas (log) y casos confirmados (log) por millón

Ecuación de regresión:

$$\hat{log(confirmed\_per\_mil)} = -0.363 + .734 * log(test\_per\_mil)$$

$$.....(1.04)...(0.0943).....$$

En este caso, un aumento de 1 en el log(*test_per_mil*) tiene un impacto positivo de .734 en el log(*confirmed_per_mil*). Esto lo podemos interpretar como un cambio del 1 % en las pruebas por millón tiene un impacto positivo del .734 % en los casos confirmados por cada millón.

(d)

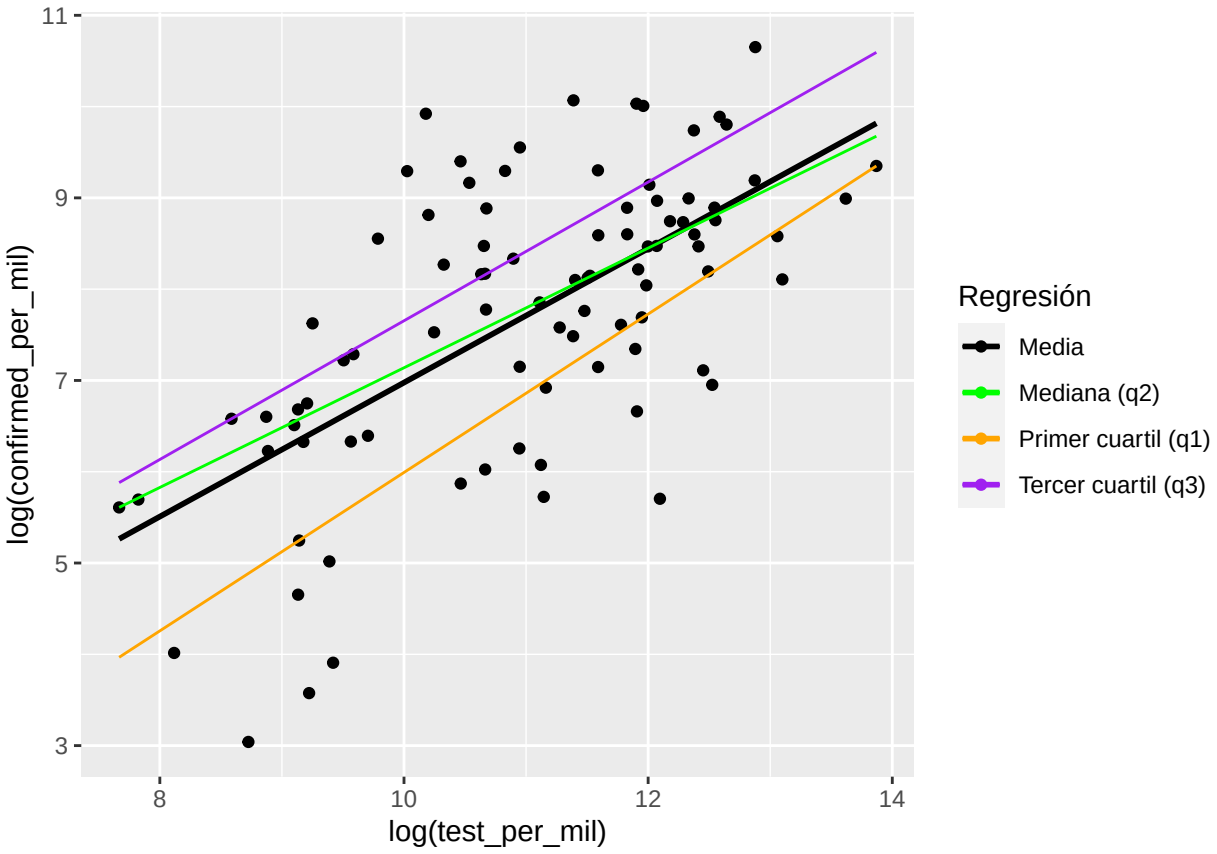

Figura 4. Regresiones cuantílicas

Ecuaciones de regresión:

$$\hat{log(confirmed\_per\_mil)}_{q1} = -2.68 + 0.87 * log(test\_per\_mil)$$

$$.....(2.20)...(0.183).....$$

$$\hat{log(confirmed\_per\_mil)}_{q2} = 0.58 + 0.66 * log(test\_per\_mil)$$

$$.....(1.071)...(0.097).....$$

$$\hat{log(confirmed\_per\_mil)}_{q3} = 0.056 + 0.76 * log(test\_per\_mil)$$

$$.....(1.734)...(0.156).....$$

De la gráfica anterior se puede concluir que las regresiones lineales te dan la flexibilidad de tener un pivote alrededor de distintos cuantiles lo cual puede ser provechoso para analizar distintas situaciones (por ejemplo de efectos diferenciados condicionales). En este caso notamos que las

regresiones del primer y tercer cuartil son casi paralelas entre ellas; que la regresión media y la mediana son muy similares aunque en los puntos iniciales y en los finales se alejan; y finalmente, que la distancia entre todas las rectas disminuye conforme aumenta log(*test_per_mil*).

Asimismo, lo más interesante es la regresión de la mediana que tiene una pendiente más pequeña que las demás y en un lapso menor intersecta a las demás líneas, lo cuál indica la distribución particual de nuestras variables: condicionadas a valores de ln(x) bajos, la mediana de ln(y)|ln(x) es mayor a la media, lo cual indica que incialmente la distribución de ln(y)|ln(x) está sesgada hacia la izquierda; por el conrario, condicionadas a valores altos de ln(x) ocurre lo opuesto, la media de ln(y)|ln(x) es mayor a la mediana, lo que indica que la distribución de ln(y)|ln(x) está sesgada a la derecha.

III.

(a)

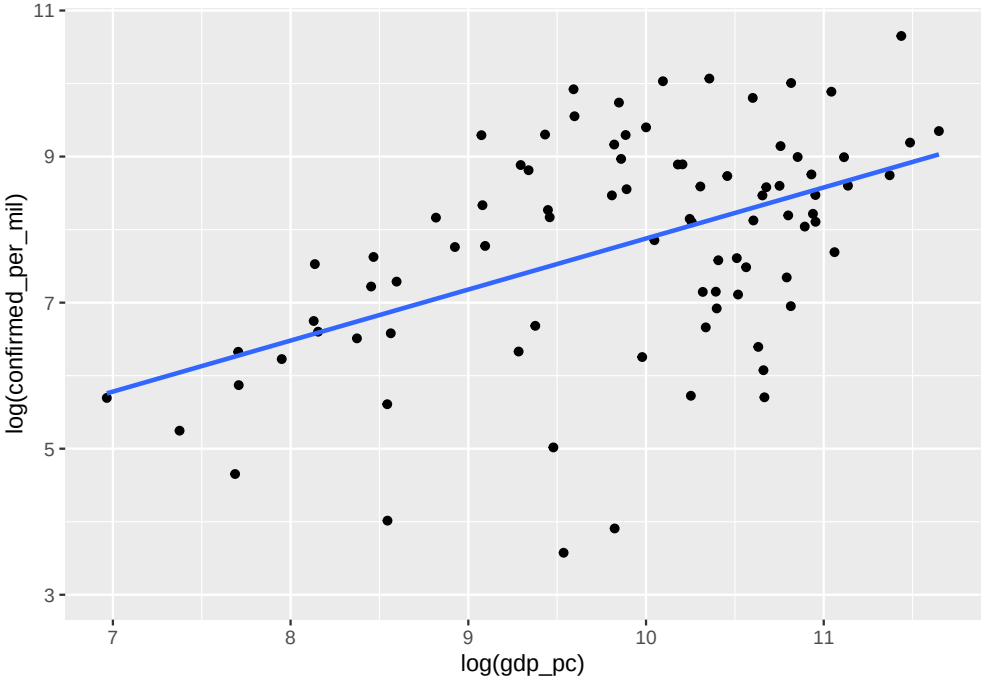

Figura 5. Relación entre pruebas (log) y casos confirmados (log) por millón

Ecuación de regresión:

$$\hat{log(confirmed\_per\_mil)} = 0.889 + 0.698 * log(gdp\_pc)$$

.....(1.345)...(0.136).....

El coeficiente de la regresión *ln(confirmed_per_mil) ~ ln(gdp_pc)* es igual a .698. Esto lo podemos interpretar como un aumento del 1 % en el pib per capita tiene un impacto positivo de .698 % en los casos confirmados por cada millón.

(b)

(i)

$$\beta_1 = \frac{cov(x,y)}{var(x)} = \frac{E(xy) - \mu_x \mu_y}{var(x)}$$

Sin realizar ningún cálculo (solo observando la definición de β_1), el $\hat{\beta}_1$ disminuye pues el termino $1000 * v$ afecta solamente (o al menos de mayor manera) a la varianza.

(ii)

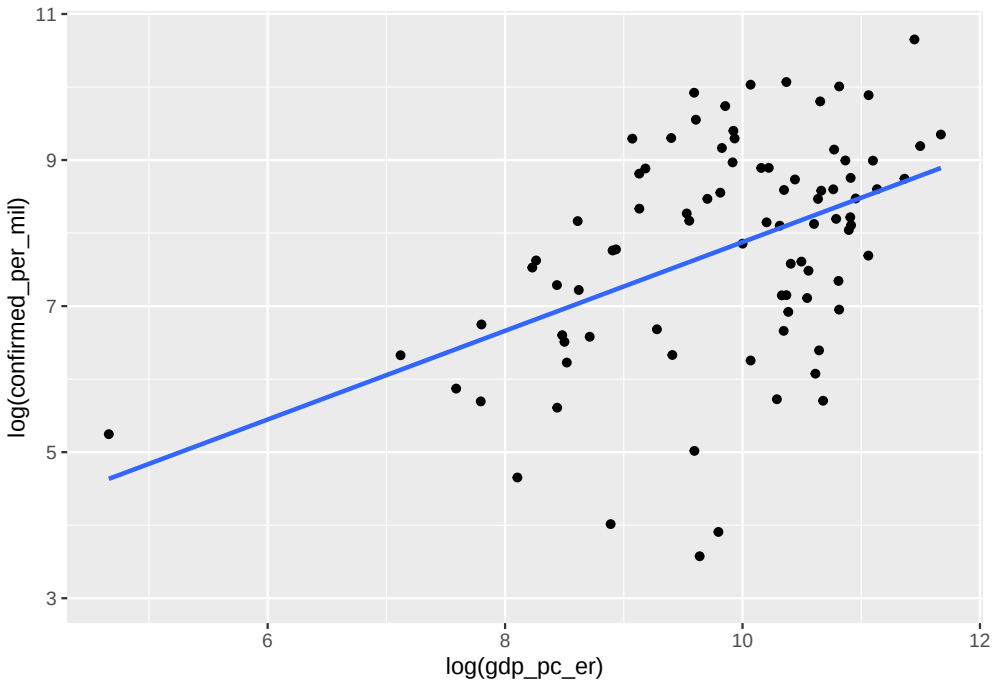

Figura 6. Relación entre pruebas (log) y casos confirmados (log) por millón

Ecuación de regresión:

$$\hat{log(confirmed\_per\_mil)} = 1.803 + 0.607 * log(gdp\_pc\_er)$$

$$.....(1.262)...(0.127).....$$

El sesgo se llama error de medición. Esta aumentando el ruido blanco en nuestra regresión por lo que, debido al aumento en la varianza, la relación entre nuestras variables se hace menos clara. En este sentido, se confirma la teoría revisada en clase.

(c)

Sí existe un sesgo pues las tres variables tienen un efecto en la misma dirección. En este caso se corre el riesgo de que *gdp_pc* este capturando el efecto de *test_per_mil*. En este sentido, el sesgo es positivo, el $\hat{\beta}_{gdp\_pc}$ muestra un efecto muy optimista. Habría que controlar por ambas variables para corregir el sesgo de variable omitida.

(d)

(i)

Ecuación de regresión:

$$\hat{log(gdp\_pc)} = 2.788 + 0.641 * log(test\_per\_mil)$$

$$.....(0.5154)..(0.0463).....$$

(ii)

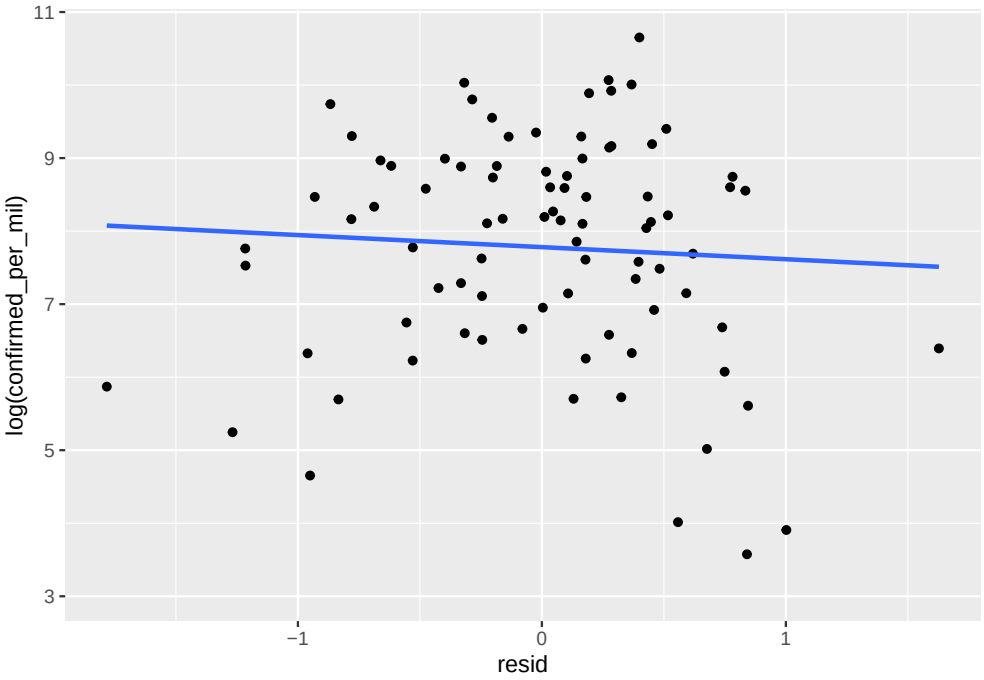

Figura 7. Partial out del pib per capita (log) y tests por millón de habitantes (log)

Ecuación de regresión auxiliar:

$$\hat{log(confirmed\_per\_mil)} = 7.781 - 0.165 * residuales$$

$$.....(0.165).(0.281).....$$

(e)

Ecuación de regresión:

$$\hat{log(confirmed\_per\_mil)} = 0.615 - 0.165 * log(gdp\_pc) + .797 * log(test\_per\_mil)$$

$$.....(1.202)...(0.219).....(0.168).....$$

El efecto de *test_per_mil* es tan fuerte que incluso volteo el signo de $\hat{\beta}_{gdp\_pc}$. Esto quiere decir que sí había sesgo de variable omitida en la pregunta anterior. Asimismo, esta regresion también nos indica que aunque la cantidad de tests esta correlacionado con el pib per capita, si controlamos por la cantidad de tests (ceteris paribus), notamos que hay un efecto negativo de la riqueza en la cantidad de confirmados: es decir, para un mismo nivel en tests por millón, los países más pobres presentan más casos confirmados.

IV.

En primer lugar, creo las variables que se van a necesitar con el siguiete código:

```
main_data<-main_data %>% mutate(cfr=deaths/confirmed,
eur = ifelse (continent == "EU",1,0), asia = ifelse (continent == "AS",1,0),
nam = ifelse (continent == "NAM",1,0),
std_hdi = (hdi - mean(hdi, na.rm = T))/sd(hdi,na.rm = T),
hm_cfr = ifelse (cfr > .019, 1, 0),
std_hdi_2= std_hdi^2)
```

Posteriormente, elaboro una tabla con las regresiones solicitadas (con o sin errores robustos):

Cuadro 3. Modelos con errores homocedásticos

	Dependent variable:			
	log(deaths)	deaths_per_mil	cfr	hm_cfr
	(1)	(2)	(3)	(4)
overwgh_prev	0.045*	5.100**	0.001***	0.015***
	(0.024)	(2.000)	(0.0002)	(0.005)
log(cardio_dr)	−0.780	−149.000***	−0.012*	−0.240*
	(0.650)	(48.000)	(0.007)	(0.140)
diab_prev	0.120	3.000	−0.001	0.012
	(0.075)	(7.200)	(0.001)	(0.017)
log(hosp_beds_per_thou)				0.033
				(0.110)
log(test_per_mil)	0.015	−6.700	−0.008**	−0.100
	(0.350)	(23.000)	(0.003)	(0.078)
aged_65_older				0.034***
				(0.013)
log(gdp_pc)	−0.600			−0.260*
	(0.540)			(0.140)
eur		8.500		
		(75.000)		
asia		−62.000		
		(62.000)		
std_hdi			−0.0002	
			(0.008)	
std_hdi_2			0.001	
			(0.003)	
median_age		−5.700	0.0003	
		(4.100)	(0.001)	
nam				0.020
				(0.190)
Constant	13.000**	926.000**	0.130**	4.200***
	(5.900)	(387.000)	(0.049)	(1.400)
Observations	86	88	88	81
R^2	0.085	0.280	0.220	0.290
Adjusted R^2	0.027	0.220	0.150	0.210
Residual Std. Error	2.400 (df = 80)	173.000 (df = 80)	0.023 (df = 80)	0.440 (df = 72)
F Statistic	1.500 (df = 5; 80)	4.500*** (df = 7; 80)	3.200*** (df = 7; 80)	3.700*** (df = 8; 72)
Note:				*p<0.1; **p<0.05; ***p<0.01

Cuadro 4. Modelos con errores robustos

	Dependent variable:			
	log(deaths)	deaths_per_mil	cfr	hm_cfr
	(1)	(2)	(3)	(4)
overwgh_prev	0.045	5.100***	0.001***	0.015***
	(0.028)	(1.500)	(0.0002)	(0.004)
log(cardio_dr)	−0.780	−149.000***	−0.012	−0.240
	(0.650)	(44.000)	(0.007)	(0.150)
diab_prev	0.120	3.000	−0.001	0.012
	(0.081)	(5.600)	(0.001)	(0.019)
log(hosp_beds_per_thou)				0.033
				(0.120)
log(test_per_mil)	0.015	−6.700	−0.008**	−0.100
	(0.450)	(18.000)	(0.003)	(0.092)
aged_65_older				0.034***
				(0.011)
log(gdp_pc)	−0.600			−0.260
	(0.530)			(0.160)
eur		8.500		
		(86.000)		
asia		−62.000		
		(56.000)		
std_hdi			−0.0002	
			(0.008)	
std_hdi_2			0.001	
			(0.003)	
median_age		−5.700	0.0003	
		(3.700)	(0.001)	
nam				0.020
				(0.200)
Constant	13.000**	926.000***	0.130***	4.200***
	(5.600)	(349.000)	(0.047)	(1.400)
Observations	86	88	88	81
R^2	0.085	0.280	0.220	0.290
Adjusted R^2	0.027	0.220	0.150	0.210
Residual Std. Error	2.400 (df = 80)	173.000 (df = 80)	0.023 (df = 80)	0.440 (df = 72)

Note: *p<0.1; **p<0.05; ***p<0.01

V.

(a)

Controlando por *cardio_dr, diab_prev, log(test_per_mil) y log(gdp_pc)*, un aumento en una unidad en la prevalencia de obesidad aumenta el log(deaths) en .045. Esto quiere decir que, *ceteris paribus*, un aumento de uno en la prevalencia de obesidad tiene un impacto positivo de 4.5 % en la cantidad de decesos.

(b)

Controlando por *overwgh_prev, cardio_dr, diab_prev, y log(test_per_mil)* , un aumento de uno en el ln(*gdp_pc*) disminuye en .600 el log(*deaths*). Esto lo podemos interpretar como, ceteris paribus, un aumento del 1 % en el pib per capita tiene un impacto negativo del .6 % en la canitdad de decesos.

(c)

Un aumento de una unidad en log(*test_per_mil*) tiene un impacto negativo de 6.7 en *deaths_per_mil*. Esto lo interpretamos como, controlando por *overwgh_prev, log(cardio_dr), diab_prev, eur, asia* y *median_age*, un aumento del 1 % en los tests por millón disminuye los decesos por millón en .067.

(d)

Un cambio en una unidad de eur tiene un impacto de 8.5 en *deaths_per_mil*. Esto lo interpretamos como, controlando por *overwgh_prev,log(cardio_dr), diab_prev, asia, median_age*, los países de europa tienen en promedio 8.5 decesos más por cada milloón habitantes.

(e)

Un cambio en una unidad de *median_age* tiene un impacto de -5.7 en *deaths_per_mil*. Esto lo interpretamos como, controlando por *overwgh_prev,log(cardio_dr), diab_prev, log(test_per_mil), eur* y *asia*, un aumento de un año en la mediana de la edad disminye los decesos por millón en 5.7. Esto parece contraintuitivo pero lo que pasa es que las variables de diabetes, sobrepeso y *cardio_dr* están absorviendo el efecto de las comorbidades que podría tener la mediana de edad; asimismo, como la variable *gdp_pc* está omitida, muy probablemente la edad está capturando el efecto de

la riqueza, que como vimos en ejercicios anteriores, el efecto es negativo en los decesos siempre y cuando controlemos por las variables adecuadas.

(f)

Controlando por las demás variables (*overwgh_prev, log(cardio_dr), log(test_per_mil), std_hdi, std_hdi_2* y *median_age*), un aumento de una unidad en la prevalencia de diabetes, tienen un impacto de -.001 unidades en la tasa de fatalidad.

(g)

overwgh_prev, log(cardio_dr), diab_prev, log(test_per_mil), std_hdi, std_hdi_2, median_age

Controlando por las demás variables (*overwgh_prev, log(cardio_dr), diab_prev, std_hdi, std_hdi_2* y *median_age*), un aumento del 1 % en los tests por millón de habitantes, disminuye en .008 la tasa de fatalidad.

(h)

Controlando por las demás variables (*overwgh_prev, log(cardio_dr), log(test_per_mil), diab_prev, std_hdi* y *median_age*), un aumento de una unidad en std_hdi^2 aumenta la tasa de fatalidad en .001. Este efecto es adicional al efecto lineal de *std_hdi*, es decir, es el término del efecto cuadrático. Por cada cambio en una unidad de *std_hdq*, la tasa de fatalidad va a disminuir en 0.0002 unidades pero al mismo tiempo aumentará $2 * .001$ unidades.

(i)

Controlando por las demás variables (*log(cardio_dr), log(hosp_beds_per_thou), log(test_per_mil), aged_65_older, log(gdp_pc)* y *nam*), un aumento de una unidad en la prevalencia de obesidad aumenta la probabilidad de tener una tasa de fatalidad acumulada del país mayor a 0.019 en 1.2 puntos porcentuales.

(j)

Controlando por las demás variables (*overwgh_prev,log(cardio_dr), diab_prev, log(test_per_mil), aged_65_older, log(gdp_pc)* y *nam*), un aumento de 1 % en la cantidad de camas por cada millón de habitantes aumenta la probabilidad de tener una tasa de fatalidad acumulada del país mayor a 0.019 en .033 %.

(k)

Controlando por las demás variables (*overwgh_prev,log(cardio_dr), log(hosp_beds_per_thou), log(test_per_mil), aged_65_older* y *log(gdp_pc)*), el pertenecer a los países de América del Norteaumenta la probabilidad de tener una tasa de fatalidad acumulada del país mayor a 0.019 en 2 puntos porcentuales.

VI.

(a)

$$\text{Predicción para México} = 331$$

$$\text{Intervalo de confianza para México} : 186 \leq 311 \leq 436$$

$$\text{Predicción puntual para México} : -56 \leq 311 \leq 678$$

$$\text{Valor real de México} = 549.29$$

Las predicciones y los intervalos se realizaron con el 95 % de confianza.

La predicción media subestima la cantidad de muertes, además de que el valor real queda por fuera del intervalo de confianza. Asimismo, aunque el valor real se encuentre dentro del intervalo de la predicción puntual, este se encuentra muy cerca de el límite superior.

(b)

Recordemos que este tipo de cálculos lo podemos realizar mediante una estimación exacta o una estimación aproximada.

Estimación exacta :

$$\Delta cfr = \Delta(\beta_{std\_hdi} + 2 * \beta_{std\_hdi\_2} * std\_hdi + \beta_{std\_hdi\_2})$$

Estimación aproximada :

$$\Delta cfr = \Delta(\beta_{std\_hdi} + 2 * \beta_{std\_hdi\_2} * std\_hdi)$$

Pero recordemos que el promedio de *std_hdi* es cero por definición (debido a que está estandarizada). Por lo que las ecuaciones de Δcfr quedan de la siguiente manera:

$$\text{Estimación exacta :}$$
$$\Delta cfr = \Delta * (\beta_{std\_hdi} + \beta_{std\_hdi\_2})$$

$$\text{Estimación aproximada :}$$
$$\Delta cfr = \Delta * \beta_{std\_hdi}$$

Ahora simplemente calculamos un intevalo de confianza (al 95 %) para cada estimación.

$$\text{Estimación exacta :}$$
$$-0.0018 \leq \Delta cfr \leq 0.002$$
$$\text{Valor estimado } \Delta cfr = .1 * (\beta_{std\_hdi} + \beta_{std\_hdi\_2}) = 0.000055$$
$$\text{Estimación aproximada :}$$
$$-0.0017 \leq \Delta cfr \leq 0.0016$$
$$\text{Valor estimado } \Delta cfr = .1 * \beta_{std\_hdi} = -0.0017$$

(c)

$$-4.2\,\% \leq \Delta P(X > .19) \leq 1.3\,\%$$
$$E(\Delta P(X > .19)) = -1.5\,\%$$

El intervalo de confianza al 95 % del posible efecto es de -4.2 % hasta 1.3 %. Es decir, es ambiguo si el efecto aumentaría o disminuiría la probabilidad de tener una tasa de fatalidad por arriba de la mediana.

Asimismo, en cuanto a la prueba de hipotesis:

$$h_o : 1.5 * \beta_{overwgh\_prev} + 2 * \beta_{diab\_prev} + 15 * \beta_{log(test\_per\_mil)} = 0$$
$$h_1 : 1.5 * \beta_{overwgh\_prev} + 2 * \beta_{diab\_prev} + 15 * \beta_{log(test\_per\_mil)} \neq 0$$

No se rechaza h0 con valor p de .3 y una valor F de 1.1. El efecto es ambiguo, lo cual es sensato debido a que el efecto va en las dos direcciones y además, aunque, segmentáramos los efectos en dos (uno positivo y uno negativo), el signo de cada uno de estos sigue siendo ambiguo por separado.

VII.

(a)

(i)

Para este ejercicio, el país promedio se podía calcular de distintas maneras: ya sea omitiendo las filas con *missing values*; imputando los valores faltantes; realizando los logaritmos después o antes de promediar, etc. En esta tarea se optó simplemente por calcular la media de todas las observaciones que cumplen el criterio solicitado, de la siguiente manera:

```
data_eu_prom<-summarise_all(subset(main_data, continent == 'EU'),
              funs(if(is.numeric(.)) mean(.,na.rm = T) else "EU"))

data_as_prom<-summarise_all(subset(main_data, continent == 'AS'),
                funs(if(is.numeric(.)) mean(.,na.rm = T) else "AS"))
```

Intervalo de confianza para país promedio de EU : $73 \leq 138 \leq 203$

Intervalo de predicción para país promedio de EU $- 213 \leq 138 \leq 489$

Intervalo de confianza para país promedio de AS : $-44 \leq 40 \leq 124$

Intervalo de predicción para país promedio de AS : $-315 \leq 40 \leq 395$

$$h_o : \sum(\overline{x_{i,eu}} - \overline{x_{i,as}}) * \beta_i = 0$$

$$h_1 : \sum(\overline{x_{i,eu}} - \overline{x_{i,as}}) * \beta_i \neq 0$$

Con un valor de F de 5.71 y un valor p de .019 rechazamos la hipotesis nula. La diferencia entre ambas predicciones es estadísticamente significante al 98 %.

(ii)

Ahora volvemos a realizar la estimación pero solamente con los datos de países europeos y asiaticos. Lo que podemos notar esque los valores de los parámetros han cambiado, algunos de manera drástica como la prevalencia de diabetes que cambio de signo. Cabe destacar que se tuvo que omitir la variable "AS" en el modelo de submuestra pues Asia ahora es representada por nuetro modelo base.

Cuadro 5. Modelo B estimado con diferentes muestras

	Dependent variable:	
	deaths_per_mil	
	Modelo B	Modelo B (estimado de submuestra)
overwgh_prev	5.100***	3.200***
	(1.500)	(1.200)
log(cardio_dr)	−149.000***	−111.000**
	(44.000)	(48.000)
diab_prev	3.000	−7.700*
	(5.600)	(4.700)
log(test_per_mil)	−6.700	−0.570
	(18.000)	(12.000)
eur	8.500	98.000
	(86.000)	(69.000)
asia	−62.000	
	(56.000)	
median_age	−5.700	−9.000**
	(3.700)	(4.400)
Constant	926.000***	890.000**
	(349.000)	(394.000)
Observations	88	54
R^2	0.280	0.270
Adjusted R^2	0.220	0.170
Residual Std. Error	173.000 (df = 80)	150.000 (df = 47)
Note:		*p<0.1; **p<0.05; ***p<0.01

Intervalo de confianza para país promedio de EU : $87 \leq 145 \leq 203$

Intervalo de predicción para país promedio de EU $- 162 \leq 145 \leq 451$

Intervalo de confianza para país promedio de AS : $-31 \leq 47 \leq 126$

Intervalo de predicción para país promedio de AS : $-366 \leq 47 \leq 358$

Como podemos notar, los intervalos se redujeron aunque por muy poco. Para probar si la diferencia sigue siento significativa planteamos una prueba de hipótesis.

$$h_o : \sum(\overline{x_{i,eu}} - \overline{x_{i,as}}) * \beta_i = 0$$

$$h_1 : \sum(\overline{x_{i,eu}} - \overline{x_{i,as}}) * \beta_i \neq 0$$

Los resultados fueron un valor F de 6.94 y un valor p de 0.011 lo que quiere decir que la diferencia sigue siendo significativa pero ahora con un 99 % de confianza.

(iii)

En primer lugar, con el siguiente código, realizó la selección aleatoria y con repetición de países que cumplen la condición de ser europeos o asiaticos.

```
boots_data<-sample_n(main_data_subset, 200, replace = T)
```

En segundo lugar, en el siguiente código, realizo la estimación del número de muertes de cada modelo y de una vez calculo los residuales.

```
residuals_IVb<-as.data.frame(unlist(
  boots_data$deaths_per_mil-predict(IVb, boots_data)))

residuals_IVb_subset<-as.data.frame(
  unlist(boots_data$deaths_per_mil-predict(IVb_subset, boots_data)))
```

Finalmente, realizo un histograma para visualizar el desempeño de cada modelo:

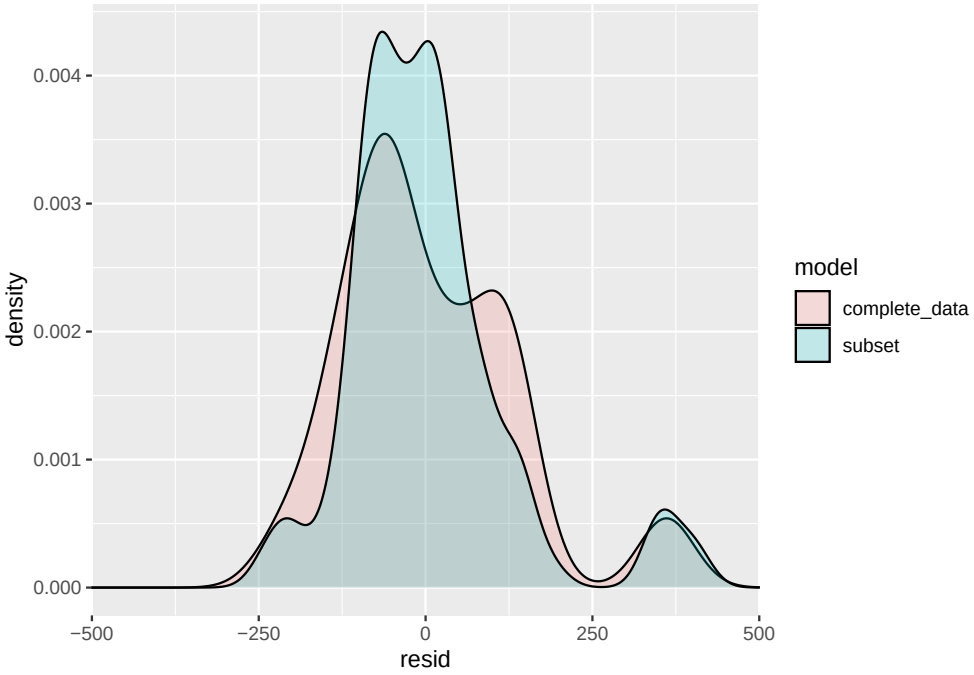

Figura 8. Comparación del desempeño de los modelos B

(iiii)

Para la tarea específica de predecir los decesos por cada millón de habitantes para un país de Europa o Asia el modelo hecho con el subset funciona mejor. Como podemos ver en el histógrama, la varianza es menor y el modelo subset está más concentrado en la media 0 de los residuales. Esto debido a que sus parámetros están hechos "a la medida" pues se estimaron con base en sus propios valores. Sin embargo, esto se logro a costa de sacrificar la capacidad del modelo para predecir los decesos de los demás países. De cierta manera, el modelo estimado con el subset está sobreajustado.

(b)

(i)

Ahora específicamos solo deaths_per_mil ~ median_age y n-1 variables *dummy* para representar los n contintentes del modelo 2. En esta especificación $\beta_1 : \beta_5$ se refieren a las ordenadas al origen diferenciadas de los n-1 continentes, en donde la ordenada al origen del continente base se representa con la β_0.

Ecuación de regresión:

$$\hat{log(deaths\_per\_mil)} = +\beta_0 + \sum_{i=1}^{5} Continente_i * \beta_i + \beta_6 * median\_age$$

Cuadro 6. Decesos por millón de habitantes

	Dependent variable:
	deaths_per_mil
median_age	−2.000
	(3.100)
factor(continent)AS	54.000
	(46.000)
factor(continent)EU	175.000**
	(81.000)
factor(continent)NAM	286.000***
	(87.000)
factor(continent)OC	12.000
	(51.000)
factor(continent)SAM	402.000***
	(97.000)
Constant	70.000
	(63.000)
Observations	89
R^2	0.350
Adjusted R^2	0.300
Residual Std. Error	163.000 (df = 82)
Note:	*p<0.1; **p<0.05; ***p<0.01

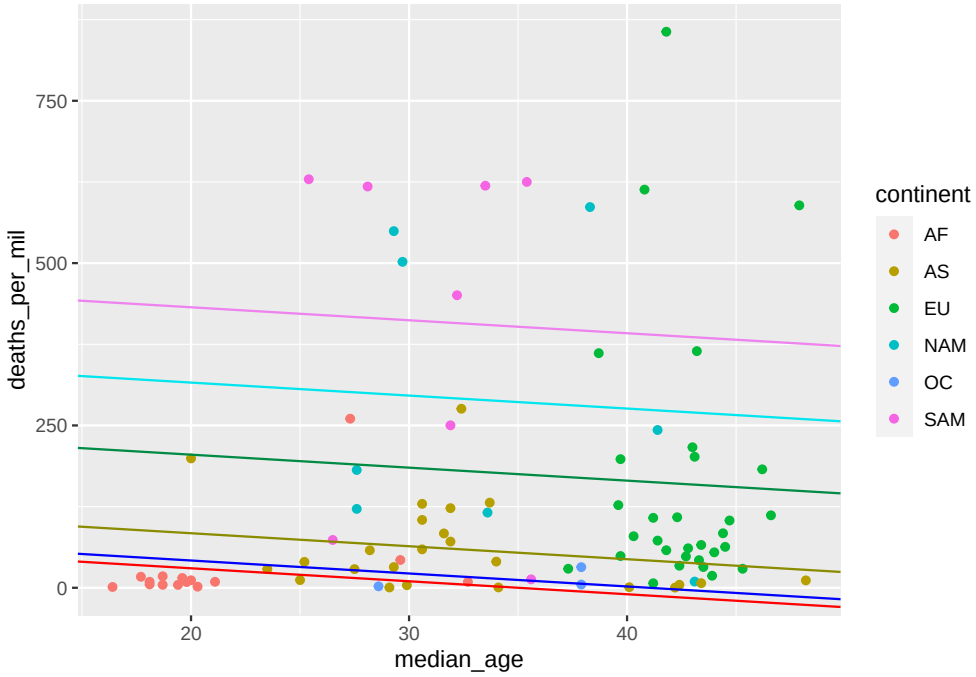

Figura 9. Modelo 2 con diferentes interceptos por continente

(ii)

La nueva especificación ahora con variables de interacción sería de la siguiente manera:

Ecuación de regresión:

$$\hat{log(deaths\_per\_mil)} = +\beta_0 + \sum_{i=1}^{5} Continente_i * \beta_i + \beta_6 * median\_age + \sum_{i=1}^{5} Continente_i * \beta_{i+6} * median\_age$$

De esta manera, estimo el nuevo modelo y lo comparo con el modelo anterior . Los resultados se presentan en el cuadro 7.

Para probar si son o no paralelas las rectas realizamos una prueba de hipótesis sobre los terminos de interacción. Si todos los términos son cero de manera conjunta entonces podemos afirmar que las rectas son paralelas. Por el contrario, si almenos una es estadisticamente distinta a cero entonces no podemos sostener la afirmación.

$$h_o : \beta_7 = 0; \beta_8 = 0; \beta_9 = 0; \beta_{10} = 0; \beta_{11} = 0$$

$$h_1 : \text{cualquier i e [7-11]} : \beta_i \neq 0$$

La prueba nos arroja una F de .25 y un valor p de .94 por lo cual, no puedo rechazar la hipotesis nula, no hay evidencia para rechazar que las lineas sean paralelas.

Cuadro 7. Decesos por millón de habitantes

	Dependent variable:	
	deaths_per_mil	
	Modelo sin interacciones	Modelo con interacciones
median_age	−2.000	5.800
	(3.100)	(5.400)
factor(continent)AS	54.000	266.000**
	(46.000)	(116.000)
factor(continent)EU	175.000**	142.000
	(81.000)	(759.000)
factor(continent)NAM	286.000***	616.000
	(87.000)	(435.000)
factor(continent)OC	12.000	48.000
	(51.000)	(104.000)
factor(continent)SAM	402.000***	724.000
	(97.000)	(892.000)
median_age:factor(continent)AS		−9.200
		(5.600)
median_age:factor(continent)EU		−3.100
		(18.000)
median_age:factor(continent)NAM		−13.000
		(14.000)
median_age:factor(continent)OC		−4.100
		(5.500)
median_age:factor(continent)SAM		−13.000
		(29.000)
Constant	70.000	−95.000
	(63.000)	(99.000)
Observations	89	89
R^2	0.350	0.360
Adjusted R^2	0.300	0.270
Residual Std. Error	163.000 (df = 82)	167.000 (df = 77)
Note:		*p<0.1; **p<0.05; ***p<0.01

(iii)

Ahora solo especifico el modelo con América del Norte como *dummy* pues me interesa analizar específicamente contra esa variable.

Ecuación de regresión:

$$\hat{log(deaths\_per\_mil)} = \beta_o + \beta_{nam} + \beta_1 * median\_age + \beta_{nam} * median\_age$$

Los resultados de la estimación se presentan en el cuadro 8.

Ahora en el planteamiento de la prueba de hipótesis simplemente comparamos Norte América frente al resto de los países. La hipotesis nula es si los dos terminos de *nam* son iguales a cero; en contra de si al menos alguno es distinto de cero.

$$h_o : \beta_{nam} = 0; \beta_{nam*median\_age} = 0$$

$$h_1 : \text{para cualquier i entre las anteirores } \beta_i \neq 0$$

Sin embargo, en este caso, al menos una es distinta de cero (probablemente la β sin interacción). Con un valor F de 3.09 y un valor p de .051, rechazamos h0 con el 95 % de confianza, las linea de NAM es diferente a las del resto de los continentes.

Finalmente, realizamos una gráfica para observar la curva.

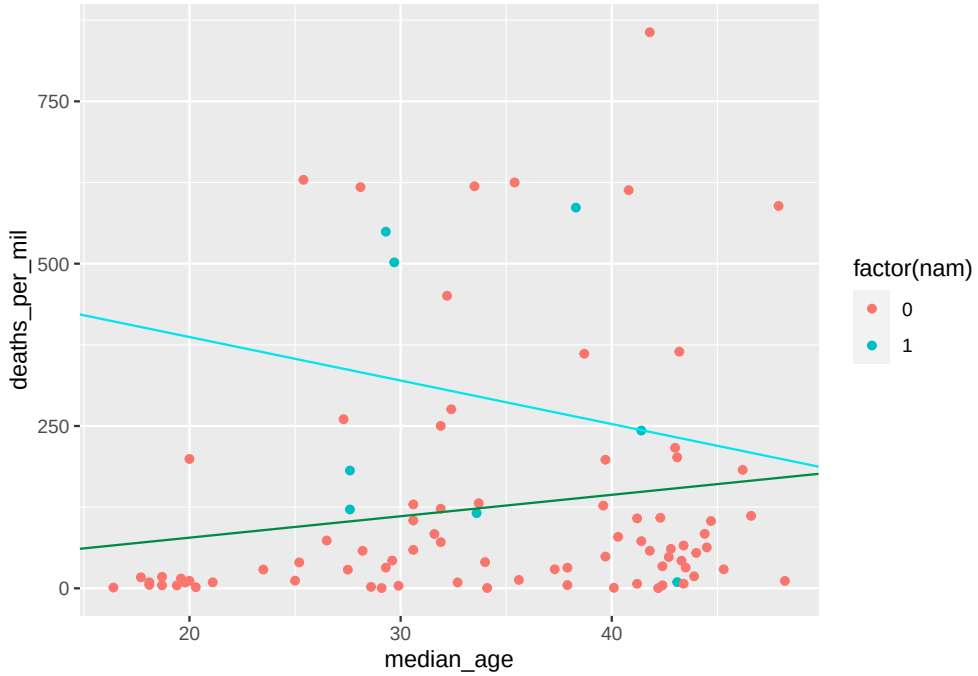

Figura 10. Modelo con América del Norte frente a los demás continentes

Cuadro 8. Decesos por millón de habitantes

	Dependent variable:		
	deaths_per_mil		
	Modelo sin interacciones	Modelo con interacciones	Modelo NAM
median_age	−2.000	5.800	3.300
	(3.100)	(5.400)	(5.400)
factor(continent)AS	54.000	266.000**	
	(46.000)	(116.000)	
factor(continent)EU	175.000**	142.000	
	(81.000)	(759.000)	
factor(continent)NAM	286.000***	616.000	
	(87.000)	(435.000)	
factor(continent)OC	12.000	48.000	
	(51.000)	(104.000)	
factor(continent)SAM	402.000***	724.000	
	(97.000)	(892.000)	
median_age:factor(continent)AS		−9.200	
		(5.600)	
median_age:factor(continent)EU		−3.100	
		(18.000)	
median_age:factor(continent)NAM		−13.000	
		(14.000)	
median_age:factor(continent)OC		−4.100	
		(5.500)	
median_age:factor(continent)SAM		−13.000	
		(29.000)	
nam			509.000
median_age:nam			−10.000
Constant	70.000	−95.000	12.000
	(63.000)	(99.000)	(99.000)
Observations	89	89	89
R^2	0.350	0.360	0.083
Adjusted R^2	0.300	0.270	0.051
Residual Std. Error	163.000 (df = 82)	167.000 (df = 77)	190.000 (df = 85)
Note:			*p<0.1; **p<0.05; ***p<0.01

Como podemos notar, a pesar de que el termino de interacción tiene un valor grande relativamente en la grafica observamos que la relación no es tan clara (esto se traduce en un estadístico t estadñisticamente insignificante para dicha β). Por el contrario, si podemos notar en la grafica de manera más clara que el intercepto de la linea de America del Norte es diferente al del resto de los continentes.