5.

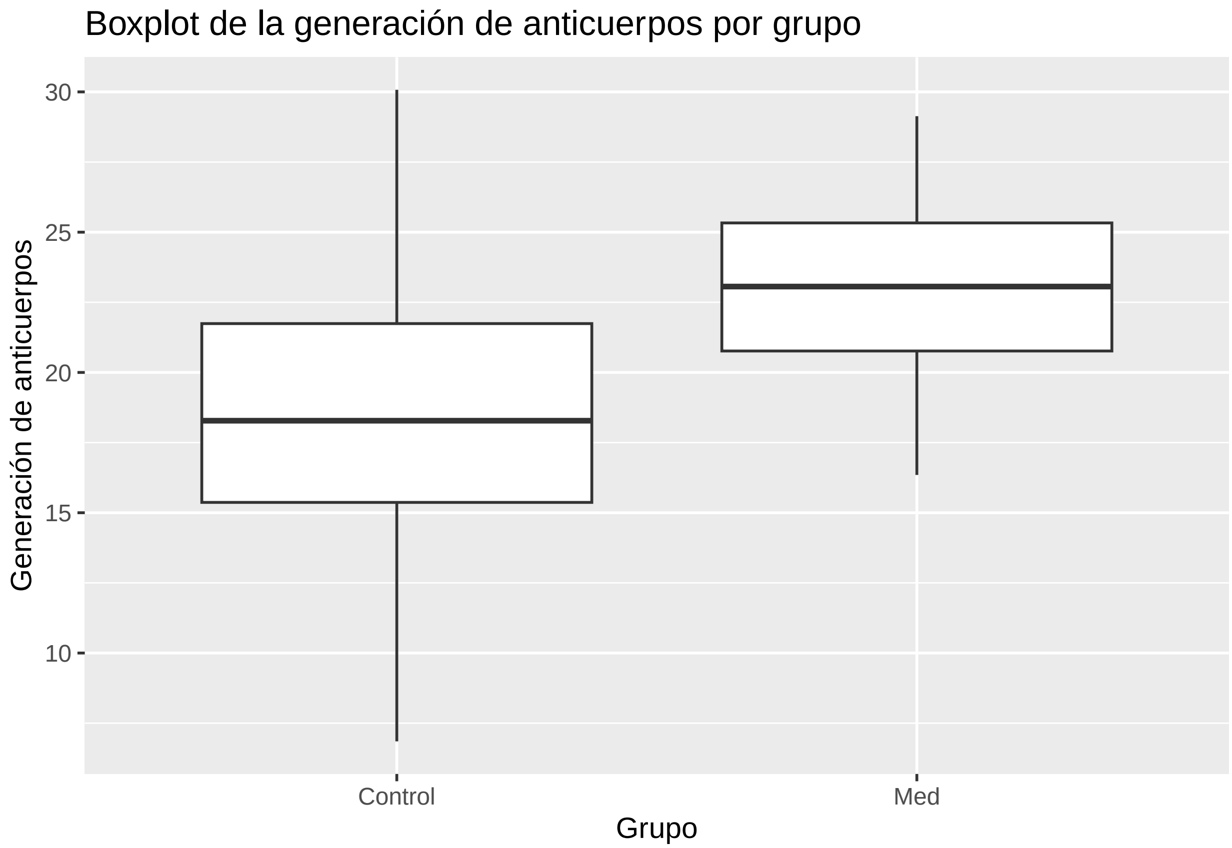
Suponga que una empresa farmacéutica está ofreciendo al gobierno un nuevo medicamento para tratar a pacientes con la enfermedad Covid-19. El costo del medicamento es considerable y para tomar una buena decisión se han acercado a usted para analizar los datos que ha compartido la empresa farmacéutica. El archivo Ex5.csv contiene la información: Ant es el numero total de anticuerpos, $Trat$ es una variable con dos niveles dependiendo si se aplicó o no el medicamento. Se sabe que tener mayores anticuerpos evita que se desarrolle una versión grave de la enfermedad y la empresa afirma que eso se logra al aplicar el medicamento, pues los pacientes que recibieron el medicamento tienen más anticuerpos que los que solo recibieron placebo. También se sabe que la generación de anticuerpos es diferente dependiendo de la edad de los individuos y se sospecha que eso también podría afectar la efectividad del medicamento, así que al diseñar el experimento se seleccionaron al azar 100 personas de 200 que presentaban síntomas leves al iniciar el cuadro de la enfermedad a los que se les administro el medicamento, al resto se les dio solo seguimiento. En todos los pacientes se capturo la edad y se procuro tener pacientes en el rango entre 16 y 60 años en ambos grupos. No se sospecha de otro aspecto que pudiera modificar la evaluación del medicamento. (Para este ejercicio no se requiere verificar supuestos del modelo, asuma que se cumplen)

I. Realice un análisis descriptivo de los datos considerando tanto la información de la edad como de la administración o no del medicamento. Entonces primero realizaremos un box-plot para comparar la población que fue medicada y la población control.

```r
data <- read_csv("Ex5.csv")
```

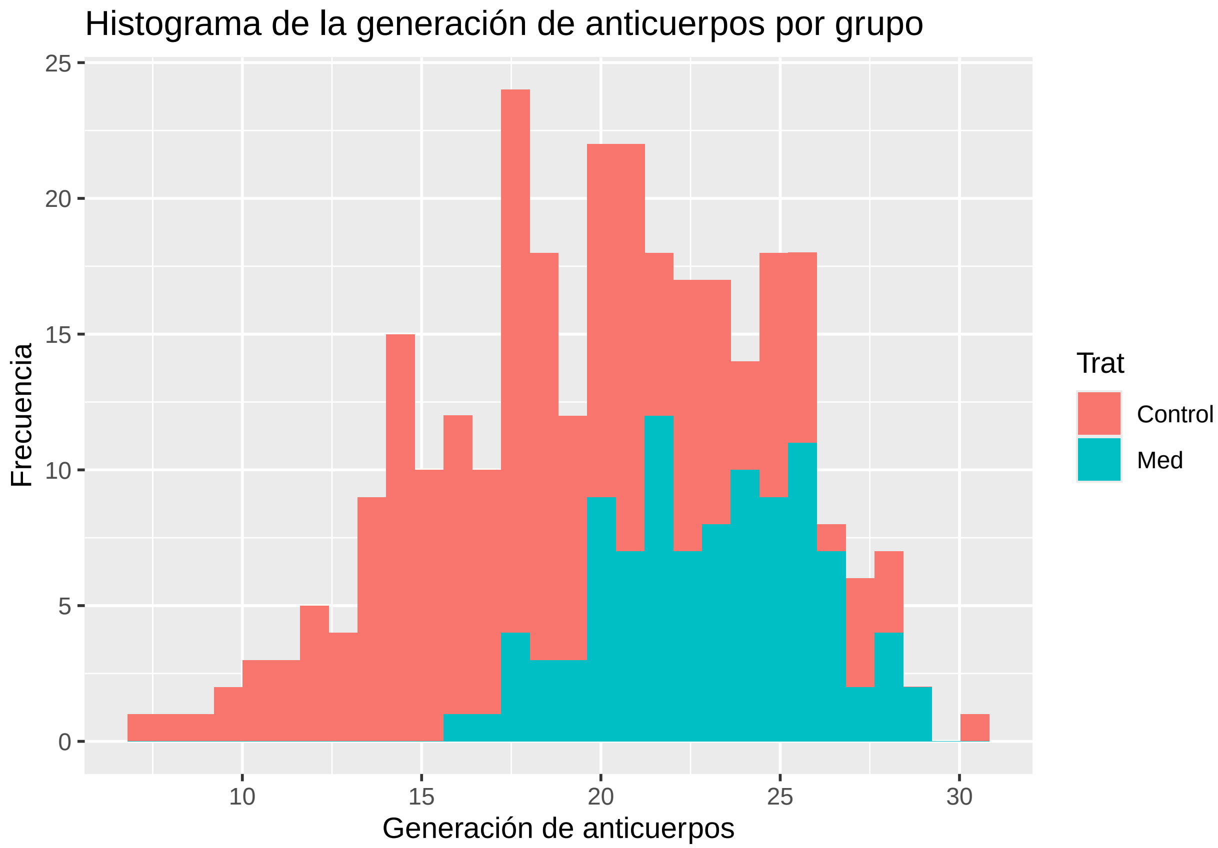
```
## Rows: 300 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Trat
## dbl (2): Ant, Edad
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
data$Trat <- factor(data$Trat)
ggplot() + geom_boxplot(data = data, aes(x = Trat, y = Ant)) +
          labs(title = "Boxplot de la generación de anticuerpos por grupo",
               x = "Grupo", y = "Generación de anticuerpos")
```



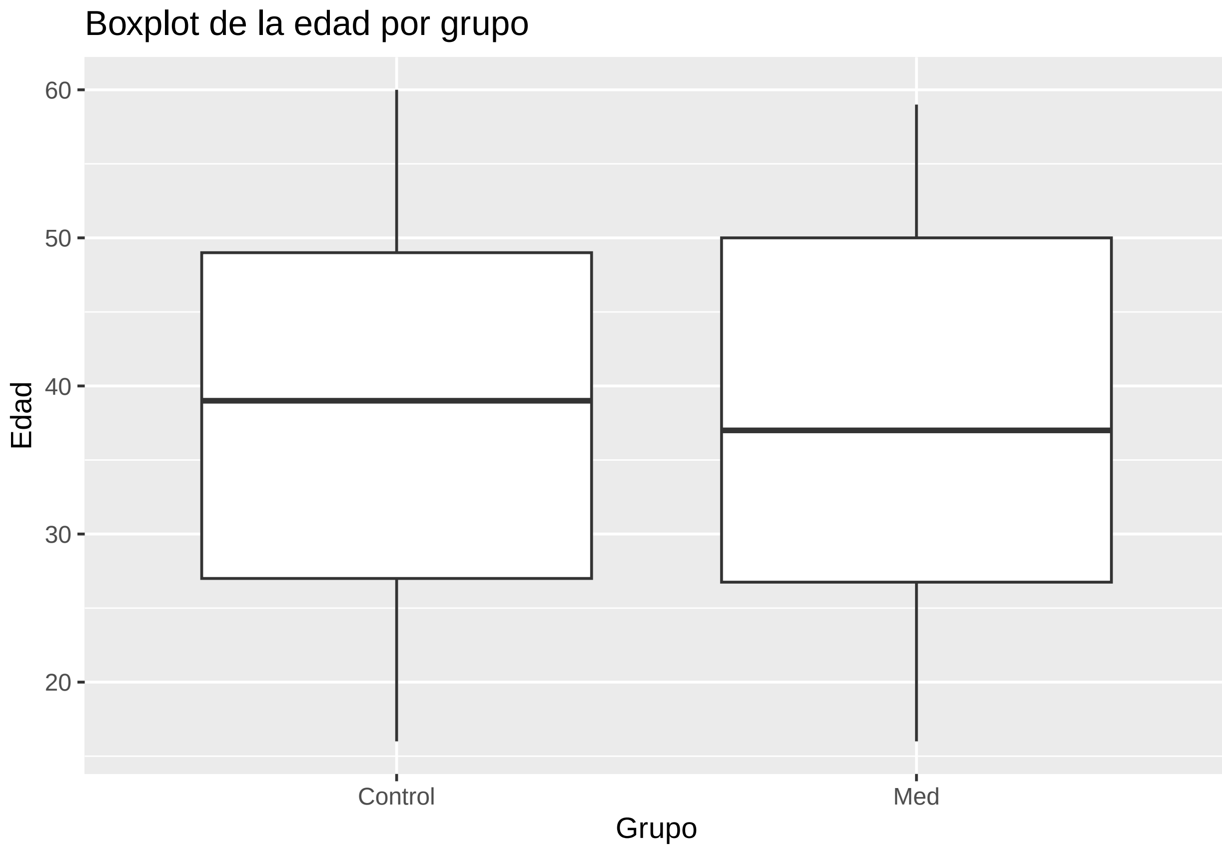
De este box-plot podemos rescatar varias cosas importantes, en primer lugar notemos como tanto el mínimo y el máximo de anticuerpos en el grupo medicado es menor que en el grupo control, lo que nos podría indicar que el tratamiento está funcionando y consolida mas este rango intercuartil, y que el paciente con menores anticuerpos en el grupo medicado, posee mas anticuerpos que el 25% de los pacientes del grupo control, además ambas medianas difieren bastante y presentarían una pendiente positiva, y en general el 50% de los medicados poseen una mayor generación de anticuerpos a comparación del 75% de los pacientes del grupo control. Veamos ahora un histograma.

```
ggplot() + geom_histogram(data = data, aes(x = Ant, fill = Trat), bins = 30) +
           labs(title = "Histograma de la generación de anticuerpos por grupo",
                x = "Generación de anticuerpos", y = "Frecuencia")
```



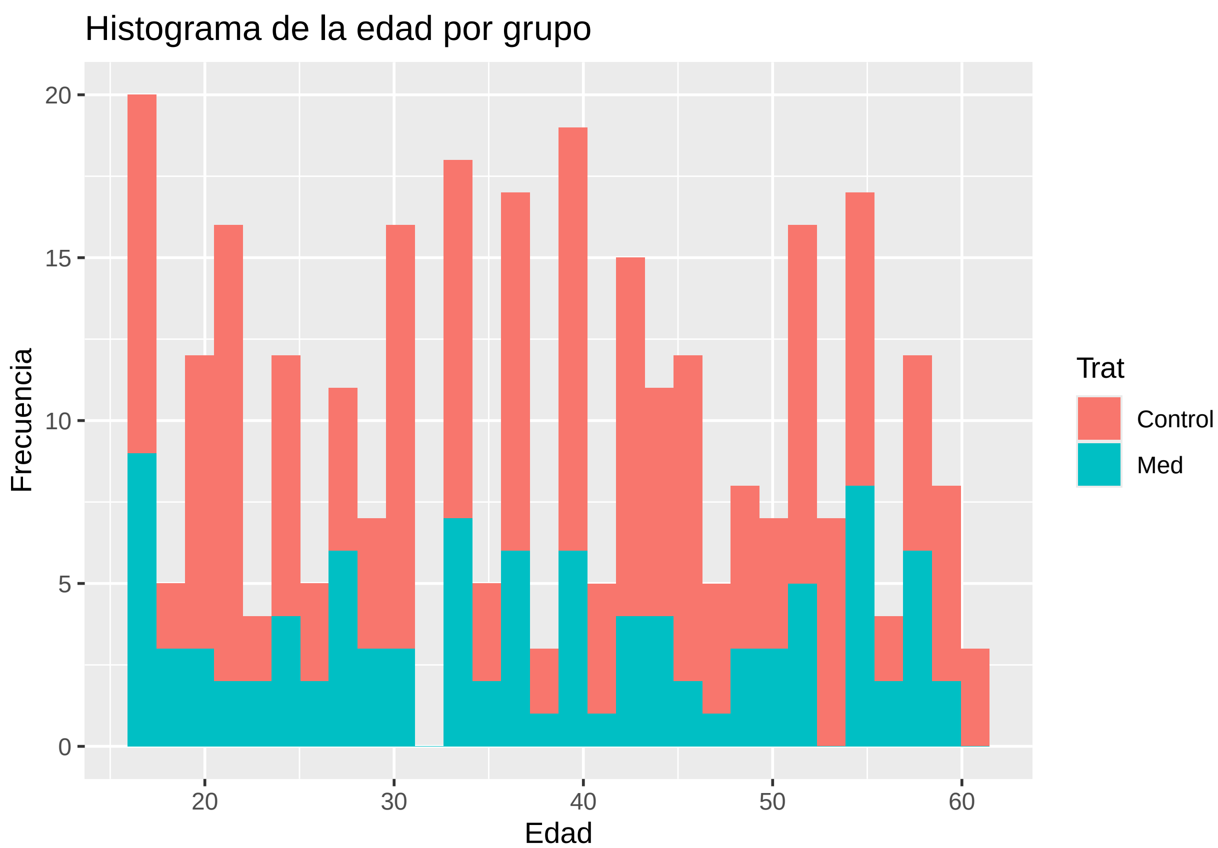
De aquí podemos notar fácilmente la diferencia de la generación de anticuerpos entre ambos grupos, donde notamos que el grupo de medicados tiene una media mayor con respecto a la media del grupo control, además de que la distribución de los anticuerpos en el grupo medicado es mas dispersa que en el grupo control. Lo cual nos da fuertes sospechas positivas sobre la efectividad del medicamento. Ahora veamos un box-plot de la edad.

```
ggplot() + geom_boxplot(data = data, aes(x = Trat, y = Edad)) +
  labs(title = "Boxplot de la edad por grupo",
       x = "Grupo", y = "Edad")
```



De este box-plot podemos notar que la edad en ambos grupos es muy similar, donde la mediana de ambos grupos es de 38 años, y la dispersión de la edad en ambos grupos es muy similar, lo cual nos da una buena señal de que la edad no afectará en la evaluación del medicamento. Ahora veamos un histograma de la edad.

```
ggplot() + geom_histogram(data = data, aes(x = Edad, fill = Trat), bins = 30) +
           labs(title = "Histograma de la edad por grupo",
                x = "Edad", y = "Frecuencia")
```



De este histograma podemos notar que la edad en ambos grupos es muy similar, donde la edad en ambos grupos se concentra en el rango de 30 a 50 años, y la distribución de la edad en ambos grupos es muy similar, lo cual nos da una buena señal de que la edad no afectará en la evaluación del medicamento.

II. Ajuste un modelo adecuado para evaluar la efectividad del medicamento ajustando por la edad de los pacientes. Es decir, un modelo que incluya como explicativas las variables edad, la binaria asociada a la administración del medicamento y la interacción obtenida como el producto de estas dos.\ Entonces queremos un modelo de la forma

$$\mathbb{E}(Ant; Edad, Trat) = b_0 + b_1 Edad + b_2 Trat + b_3 Edad \cdot Trat$$

```
fit <- lm(Ant ~ Edad*Trat, data = data)
summary(fit)
```

```
##
## Call:
## lm(formula = Ant ~ Edad * Trat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6211 -1.9539  0.0277  1.6018  7.0063
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.34298    0.57573  50.966  < 2e-16 ***
## Edad         -0.28290    0.01440 -19.645  < 2e-16 ***
## TratMed      -2.25730    0.96763  -2.333   0.0203 *
## Edad:TratMed  0.17307    0.02437   7.101 9.21e-12 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.61 on 296 degrees of freedom
## Multiple R-squared:  0.6709, Adjusted R-squared:  0.6676
## F-statistic: 201.2 on 3 and 296 DF,  p-value: < 2.2e-16
```

III. De acuerdo con el modelo ajustado, indique las expresiones asociadas a la relación de la generación promedio de anticuerpos con la edad en a) el grupo control y b) en el grupo que recibe el medicamento.

Entonces veamos la expresión para el grupo control:

$$\begin{aligned}\mathbb{E}(Ant; Edad, Trat = Control = 0) &= b_0 + b_1 Edad + b_2 \cdot 0 + b_3 Edad \cdot 0 \\ &= b_0 + b_1 Edad \\ &= 29.3 - 0.3 Edad\end{aligned}$$

Por otro lado para el grupo que recibe el medicamento tenemos que:

$$\begin{aligned}\mathbb{E}(Ant; Edad, Trat = Medicado = 1) &= b_0 + b_1 Edad + b_2 \cdot 1 + b_3 Edad \cdot 1 \\ &= (b_0 + b_2) + (b_1 + b_3) Edad \\ &= 27 - 0.1 Edad\end{aligned}$$

IV. ¿Se puede decir que la edad afecta de la misma forma la generación de anticuerpos en el grupo control que en el grupo que recibe el medicamento? Realice una prueba de hipótesis apropiada e interprete.

Entonces para realizar esta prueba de hipótesis, podemos comparar las medias obtenidas anteriormente, para comparar si existe un cambio en la generación de anticuerpos con respecto a la edad, considerando solamente a los parámetros que están asociados a la edad, así nuestra prueba de hipótesis queda como:

$$H_0 : b_3 = 0 \quad vs \quad H_1 : b_3 \neq 0$$

```r
K=matrix(c(0,0,0,1), ncol=4, nrow=1, byrow=TRUE)
m=c(0)
summary(glht(fit, linfct=K, rhs=m))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Ant ~ Edad * Trat, data = data)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0  0.17307    0.02437   7.101 9.21e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Así como rechazamos la hipótesis nula, podemos afirmar que la edad no afecta de la misma forma la generación de anticuerpos en el grupo control que en el grupo que recibe el medicamento. Pasemos a ver la prueba de hipótesis simultánea.

```r
K=matrix(c(0,1,0,0,
           0,1,0,1), ncol=4, nrow=2, byrow=TRUE)
m=c(0,0)
summary(glht(fit, linfct=K, rhs=m))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Ant ~ Edad * Trat, data = data)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0 -0.28290    0.01440 -19.645  < 1e-10 ***
## 2 == 0 -0.10983    0.01966  -5.586 1.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Entonces con la prueba hipótesis de prueba simultanea podemos rechazar todavía mas que ambas rectas puedan tener pendiente igual a 0, por lo que no podemos reducir el modelo.

V. Comente sobre el ajuste del modelo incluyendo la interpretación de cada uno de los coeficientes.

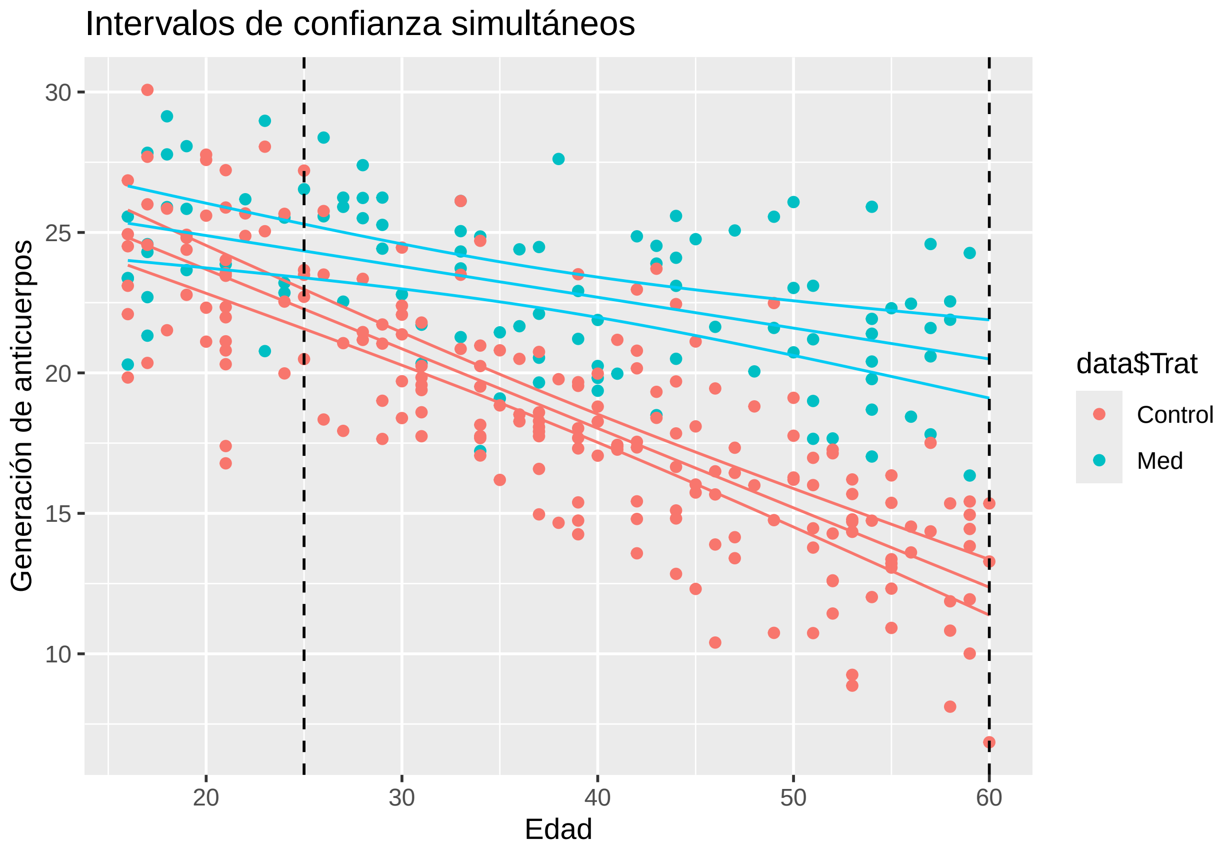
```r
summary(fit)
```

```
##
## Call:
## lm(formula = Ant ~ Edad * Trat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6211 -1.9539  0.0277  1.6018  7.0063
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.34298    0.57573  50.966  < 2e-16 ***
## Edad         -0.28290    0.01440 -19.645  < 2e-16 ***
## TratMed      -2.25730    0.96763  -2.333   0.0203 *
## Edad:TratMed  0.17307    0.02437   7.101 9.21e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.61 on 296 degrees of freedom
## Multiple R-squared:  0.6709, Adjusted R-squared:  0.6676
## F-statistic: 201.2 on 3 and 296 DF,  p-value: < 2.2e-16
```

Entonces regresando a nuestro modelo podemos ver que con una significancia de 0.05 podemos rechazar la hipótesis nula de que todas las betas son iguales a 0, además también podemos ver que con la misma significancia no podemos reducir nuestro modelo quitando alguna de las variables, pues no tenemos información suficiente para decir que dado que las demás variables están en el modelo podemos retirar alguna, por ultimo podemos ver que el modelo explica alrededor del 67% de nuestros datos. Pasemos a la interpretación de cada uno de los coeficientes: - b_0: Es el intercepto de nuestro modelo, es decir el valor esperado de la generación de anticuerpos cuando la edad es 0 y no se aplica el medicamento, en este caso es de 29.3. - b_1: Es el coeficiente asociado a la edad, es decir por cada año que aumenta la edad disminuye en 0.3 la generación de anticuerpos. - b_2: Es el coeficiente asociado a la administración del medicamento, es decir por aplicar el medicamento disminuye en 2.3 la generación de anticuerpos. - b_3: Es el coeficiente asociado a la interacción entre la edad y la administración del medicamento, es decir por cada año que aumenta la edad y se aplica el medicamento aumenta en 0.2 la generación de anticuerpos.

VI. Argumente en contra o a favor de la afirmación: "El medicamento funciona aumentando el número de anticuerpos para todos los pacientes entre 25 y 60 años". Se puede apoyar de pruebas de hipótesis o intervalos de confianza simultáneos.\ Entonces nos apoyaremos de intervalos de confianza simultáneos:

```
age <- seq(from = 16, to = 60, by = .5)

#Bajar? la confianza a 90% pues ser?n intervalos simult?neos

#Para una banda para la recta del tratamiento A
#E(Y;TRT=A, X2)= b0 + b1 age
KA <- cbind(1, age, 0, 0)
#Para una banda para la recta del tratamiento B
#E(Y;TRT=B, X2)=  b0 + b1 age +b2 + b4 age = (b0 + b2) + (b1 + b4) age
KB <- cbind(1, age, 1, age)

K=rbind(KA, KB)

fitE <- glht(fit, linfct = K)
fitci <- confint(fitE, level = 0.95)

ggplot() + geom_point(mapping = aes(x = data$Edad, y = data$Ant, color = data$Trat)) +
           geom_line(mapping = aes(x = age, y = coef(fitE)[1:89]),color = "#f8766dff") +
           geom_line(mapping = aes(x = age, y = fitci$confint[1:89,"upr"]),
                     color = "#f8766dff") +
           geom_line(mapping = aes(x = age, y = fitci$confint[1:89,"lwr"]),
                     color = "#f8766dff" ) +
           geom_line(mapping = aes(x = age, y = coef(fitE)[90:178]),
                     color = "#00cbf4") +
           geom_line(mapping = aes(x = age, y = fitci$confint[90:178,"upr"]),
                     color = "#00cbf4") +
           geom_line(mapping = aes(x = age, y = fitci$confint[90:178,"lwr"]),
                     color = "#00cbf4") +
           labs(title = "Intervalos de confianza simultáneos",
                x = "Edad", y = "Generación de anticuerpos") +
           geom_vline(xintercept = 25, linetype = "dashed") +
           geom_vline(xintercept = 60, linetype = "dashed")
```



De donde podemos concluir con el intervalo de confianza simultáneo que el medicamento funciona aumentando el número de anticuerpos para todos los pacientes entre 25 y 60 años.