Ejercicio 1 de la Evaluación continua

Pérez Efremova, Daniel

Noviembre 2020

Índice

Índice de Códigos

Ejercicio 1

apartado a

Se plantea la creación de dos variables aleatorias relacionadas linealmente. En primer lugar una variable independiente $x \sim Uniforme(0, 10)$ y una variable dependiente de x que es $y = \beta_0 + \beta_1 x + u$, donde consideramos además, $u \sim N(0, 15)$ que introduce cierta perturbación en la relación lineal.

Para los valores $\beta_0 = 2$ y $\beta_1 = 4$ la creación de 60 observaciones de tales variables pueden generarse en R con las siguientes instrucciones.

Código 1: Generador de las variables aleatorias

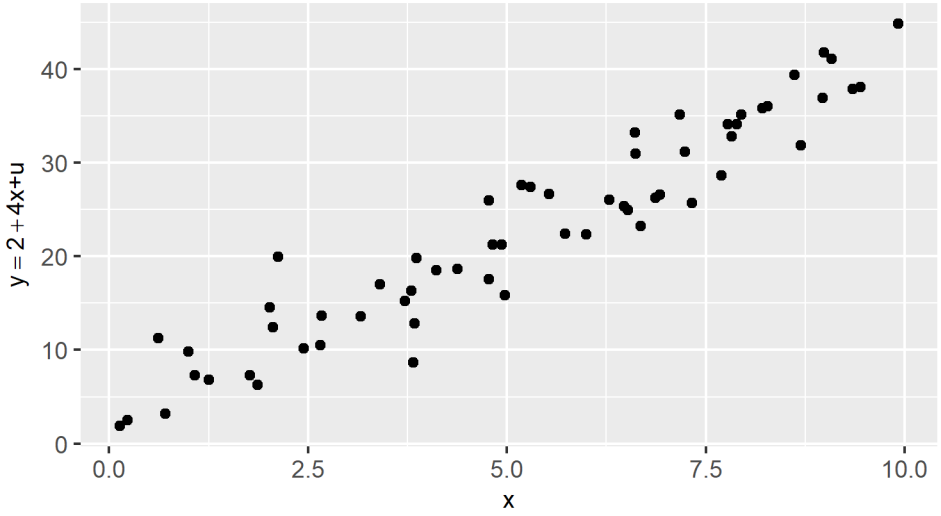
```
set.seed(0)
x <- runif(60, 0, 10)
u <- rnorm(60, mean=0, sd=sqrt(15))
y <- 2 + 4*x + u
```

Para finalizar, creamos un diagrama de dispersión para visualizar nuestras observaciones

Código 2: Gráfico de dispersión de las variables x e y

```
data = data.frame(x,y)

plot = ggplot(data=data, aes(x,y)) + geom_point() +
labs(x ="x", y = TeX('$y = 2 + 4x+u$')) +
theme(plot.title = element_text(color='black',size=9, hjust=0.5),
                                axis.title.x = element_text(size=9),
                                axis.title.y = element_text(size=9))
```

Figura 1: Gráfico de dispersión



En la Figura 1 observamos el resultado del gráfico de dispersión, donde vemos clara relación lineal, pero a causa de la perturbación u introducida, los puntos no están alineados en una recta.

apartado b

Primeramente, comprobaremos que las variables definidas en el experimento del apartado a, cumplen las hipótesis de un modelo de regresión lineal simple (RLS):

1. La perturbación tiene esperanza nula. Por definición $u \sim N(0, 15)$, luego $E[u] = 0$, y la hipótesis se sumple.

2. *homocedasticidad*. Por definición, la variable u es independiente de x y con varianza constante, es decir, $\sigma^2 = 15$.

3. La perturbación u tiene distribución normal. Resulta trivial por la definición de $u \sim N(0, 15)$.

4. Las perturbaciones u_i son independientes. También trivial por definición, ya que en cada observación i de las 60 generadas, u_i se genera de forma independiente al resto.

Vamos ahora a simular 1000 experimentos independientes del tipo $\{(x_i, y_i)\}_{i=1}^{60}$ y para cada experimento j, estimaremos el coeficiente de regresión de y sobre x que es

$$\hat{\beta}_1^{(j)} = \frac{cov(x^{(j)}, y^{(j)})}{s^2_{x^{(j)}}}$$

teniendo en cuenta que dicha estimación es adecuada si se cumplen las hipótesis que hemos comprobado al inicio del apartado.

Teóricamente se cumple que

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{n s_x^2}) \tag{1}$$

así que la estimación es centrada en el verdadero valor de la pendiente $\beta_1 = 4$, y deberemos observar un valor medio de los $\hat{\beta}_1^{(j)}$ muy próximo a 4. Por otro lado, en el histograma de los $\hat{\beta}_1^{(j)}$ deberemos observar una distribución muy próxima a una normal, con colas poco pesadas, y gran concentración de valores alrededor de 4. Aplicaremos el test Kolmogorv-Smirnov para probar tal normalidad.

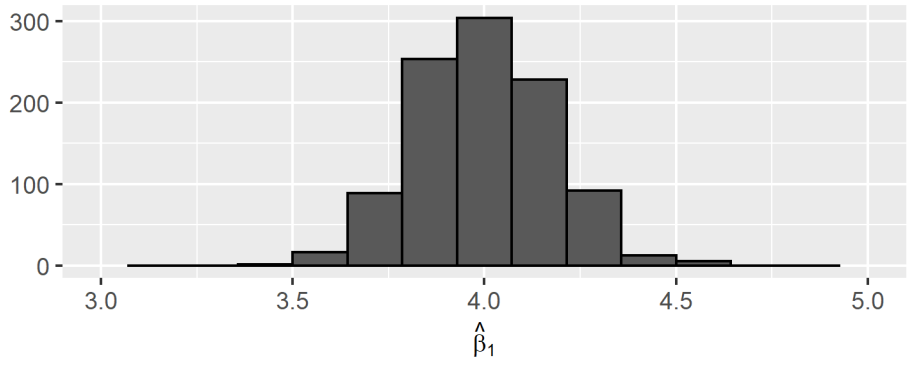
El código para simular los experimentos es similar al del primer apartado, pero incluyendo un bucle durante 1000 ocasiones.

Código 3: Repetición del experimento 1000 veces

```
betas = c() # vector que contiene los coeficientes de regresion de cada
    experimento
```

```
S_x = (10**2)/12  # varianza de la distribucion uniforme unif(0,10)
sigma2 = 15  # varianza de la perturbacion

for(i in 1:1000)  #simulacion de 1000 observaciones de la poblacion
{
    set.seed(i)

    x <- runif(60, 0, 10)
    u <- rnorm(60, mean=0, sd=sqrt(15))
    y <- 2 + 4*x + u
    n <- length(x)

    # R por defecto quita un grado de liberta, lo deshacemos
    s_x = var(x)*((n-1)/n)
    cov_xy = cov(x,y)*((n-1)/n)

    beta_1 = cov_xy/s_x  # calculo del coeficiente de regresion linal
        simple

    betas <- append(betas,beta_1)
}

# la escala del eje x la fijamos entre 3 y 5
hist = qplot(betas, bins=15, col=I('black'), xlab=TeX('$\\hat{\\beta}_1$
    '), xlim = c(3,5))+
theme(axis.title.x = element_text(size=9))

print(mean(betas))

[1] 3.997411

# test kolmogorov smirnov para la distribucion normal teorica de beta
print(ks.test(x = betas,"pnorm", mean(betas),
sqrt((sigma2)/(60*S_x)), alternative='two.side'))

        One-sample Kolmogorov-Smirnov test

data:  betas
D = 0.020519, p-value = 0.7938
alternative hypothesis: two-sided
```

Figura 2: Histograma de las estimaciones $\hat{\beta}_1^{(j)}$



En la Figura 2 vemos que el histograma es característico de una distribución normal centrada en 4. El test arroja un p-valor $\approx 0{,}79$, y concluimos que no

podemos rechazar el pronóstico teórico de (1).

apartado c

Si consideramos que u no sigue una normal, estamos violando una de las especificaciones del modelo RLS, de modo que deberíamos observar algunos cambios importantes en el histograma de la Figura 2. Veamos esto con varios ejemplos.

Simularemos los mismos experimentos pero con tres distribuciones distintas para las perturbaciones u: la uniforme, la exponencial y la de Cauchy. Después, para observar los posibles cambios generaremos los histogramas de las estimaciones en el intervalo $[3, 5]$, para ver si efectivamente ha cambiado su distribución respecto al apartado b.

Código 4: Simulación de los experimentos para distintas distribuciones de la perturbación.

```
# vectores que contienen los coeficientes de regresion de cada
    experimento
betas_1 = c()
betas_2 = c()
betas_3 = c()

for(i in 1:1000)  #simulacion de 1000 observaciones de la poblacion
{

    set.seed(i) # fijamos la semilla de cada experimento para que los
        resultados sean reproducibles

    x <- runif(60, 0, 10)

    u1 <- runif(60, -3*sqrt(5), 3*sqrt(5))
    u2 <- rexp(60, rate=1/sqrt(15))
    u3 <- rcauchy(60, location = 0, scale = sqrt(15))

    y1 <- 2 + 4*x + u1
    y2 <- 2 + 4*x + u2
    y3 <- 2 + 4*x + u3

    n <- length(x)

    # R por defecto quita un grado de liberta, lo deshacemos
    s_x = var(x)*((n-1)/n)

    cov_xy_1 = cov(x,y1)*((n-1)/n)
    cov_xy_2 = cov(x,y2)*((n-1)/n)
    cov_xy_3 = cov(x,y3)*((n-1)/n)

    beta_1_1 = cov_xy_1/s_x  # calculo del coeficiente 1
    beta_1_2 = cov_xy_2/s_x  # calculo del coeficiente 1
    beta_1_3 = cov_xy_3/s_x  # calculo del coeficiente 1

    betas_1 <- append(betas_1,beta_1_1)
    betas_2 <- append(betas_2,beta_1_2)
    betas_3 <- append(betas_3,beta_1_3)
}

hist1 = qplot(betas_1, bins=15, col=I('black'),
xlab=TeX('$\\hat{\\beta}_1$'), xlim = c(3,5))+
labs(title=TeX('$u_\\sim unif(-3/\\sqrt{15}, -3/\\sqrt{15})$'))+
```

```
theme(axis.title.x = element_text(size=9), title = element_text(size=9))
scatter = qplot(x, y1) + labs(x='', y='') + labs(title='ejemplo de nube
    de puntos')+
theme(axis.title.x = element_text(size=9), title = element_text(size=9))
plot1 = ggarrange(hist1, scatter, ncol=2, nrow=1)

hist2 = qplot(betas_2, bins=15, col=I('black'),
xlab=TeX('$\\hat{\\beta}_1$'), xlim = c(3,5))+
labs(title=TeX('$u \\sim exp(1/\\sqrt{15})$'))+
theme(axis.title.x = element_text(size=9), title = element_text(size=9))
scatter = qplot(x, y2) + labs(x='', y='')+
labs(title='ejemplo de nube de puntos')+
theme(axis.title.x = element_text(size=9), title = element_text(size=9))
plot2 = ggarrange(hist2, scatter, ncol=2, nrow=1)

hist3 = qplot(betas_3, bins=15, col=I('black'),
xlab=TeX('$\\hat{\\beta}_1$'), xlim = c(3,5))+
labs(title=TeX('$u \\sim Cauchy(0, \\sqrt{15})$'))+
theme(axis.title.x = element_text(size=9), title = element_text(size=9))
scatter = qplot(x, y3) + labs(x='', y='')+
labs(title='ejemplo de nube de puntos') +
theme(axis.title.x = element_text(size=9), title = element_text(size=9))
plot3 = ggarrange(hist3, scatter, ncol=2, nrow=1) + labs(title='fr')

resultado = ggarrange(plot1, plot2, plot3, nrow=3, ncol=1)

print(c('media beta_1, u uniforme:', mean(betas_1)))
print(c('media beta_1, u exponencial:', mean(betas_2)))
print(c('media beta_1, u Cauchy:', mean(betas_3)))


[1] "media beta_1, u uniforme:" "3.99181045687601"
[1] "media beta_1, u exponencial:" "4.00357537683733"
[1] "media beta_1, u Cauchy:" "11.9698329507627"
```

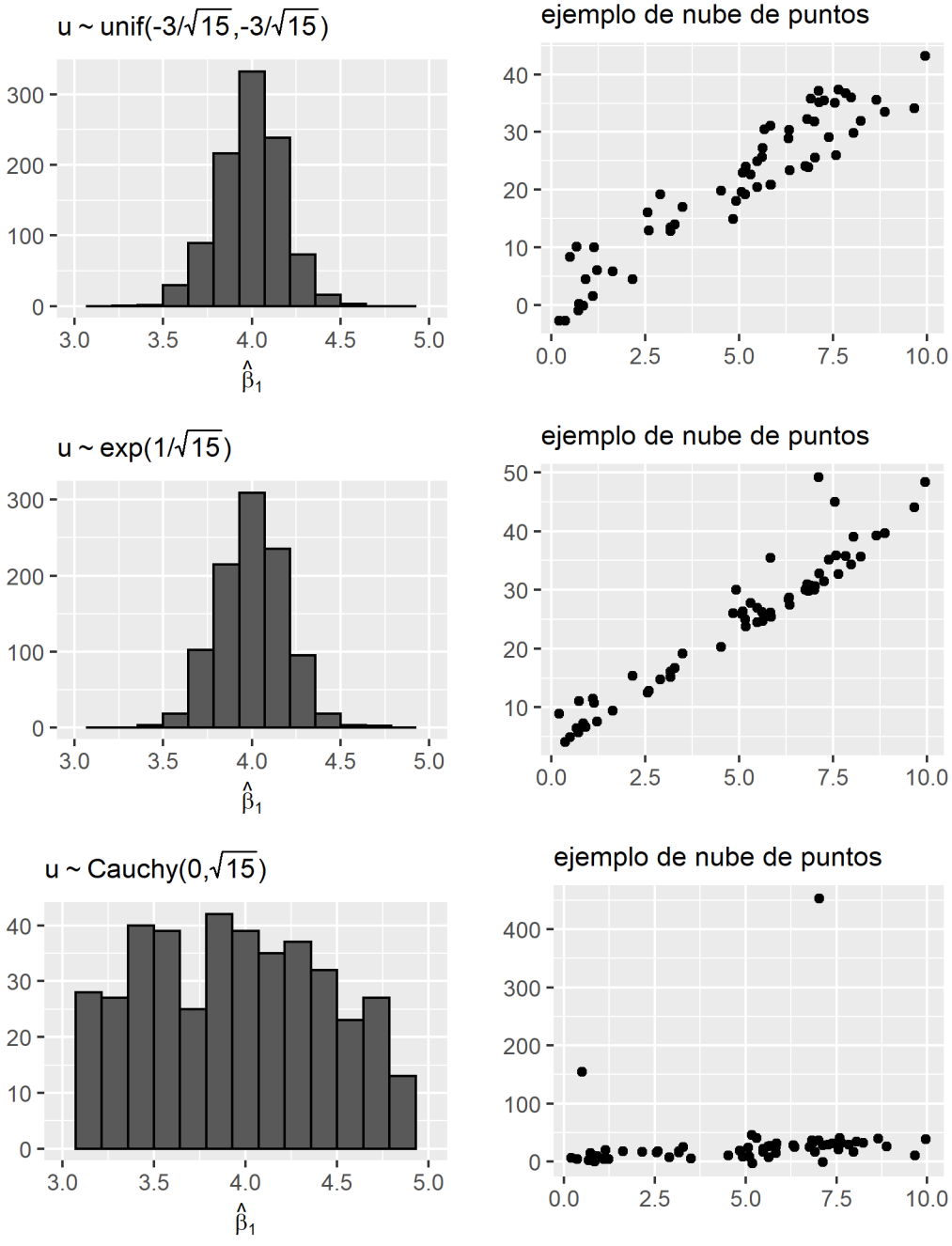
Vemos en la Figura 3 que la distribución muestral de $\hat{\beta}_1$ se ha mantenido para las distribuciones exponencial y uniforme, sin embargo, para la distribución de Cauchy observamos cambios importantes.

Una justificación sencilla en el caso de la distribución uniforme es que dicha distribución no genera valores fuera del intervalo $[a, b]$ donde está definida, por tanto, no es posible que genere grandes perturbaciones impredecibles (outliers), manteniendo casi inalterada la distribución de $\hat{\beta}_1$ respecto al apartado b.

Por otro lado, la distribución exponencial no tiene una cola pesada, concentrando gran parte de su densidad en valores cercanos a cero, luego, esta distribución no genera grandes perturbaciones, haciendo que la distribución de $\hat{\beta}_1$ quede apenas intacta también.

Finalmente, la distribución de Cauchy tiene unas colas extremadamente pesadas en comparación con las distribuciones anteriores, esto aumenta la probabilidad de generar valores altos respecto a la media de la distribución, apareciendo, como vemos en el ejemplo de nube de puntos, valores muy separados del resto, que actúan como puntos palanca, resultando en cada experimento, valores muy dispares de la estimación de $\hat{\beta}_1$ que dependen de tales puntos palanca (que poco o nada tienen que ver con el verdadero valor $\beta_1 = 4$), esto se puede ver en que la media es muy distinta respecto al apartado b.

Figura 3: Histograma de las estimaciones según distintas distribuciones de la perturbación u.



Ejercicio 2

Selección de la variable explicativa

Considerando la base de datos *werner.txt*, pretendemos construir un modelo RLS para explicar las variaciones que presenta la variable *Cholesterol*, que es, a partir de ahora, nuestra variable respuesta y.

Para seleccionar, una variable explicativa x, miraremos los gráficos de dispersión para identificar posibles candidatos, excluyendo la variable *Pill* por ser cualitativa.

Código 5: Generación de todos los gráficos de dispersión respecto de *Cholesterol* en una misma ventana.

```
p1 = qplot(datos$Age, datos$Cholesterol) +
labs(title='', x='Age', y='Cholesterol')

p2 = qplot(datos$Height, datos$Cholesterol) +
labs(title='', x='Height', y='')

p3 = qplot(datos$Weight, datos$Cholesterol) +
labs(title='', x='Weight', y='Cholesterol')

p5 = qplot(datos$Albumin, datos$Cholesterol)+
labs(title='', x='Albumin', y='')

p6 = qplot(datos$Calcium, datos$Cholesterol)+
labs(title='', x='Calcium', y='Cholesterol')

p7 = qplot(datos$Uric, datos$Cholesterol)+
labs(title='', x='Uric', y='')

plot = ggarrange(p1, p2, p3, p5, p6, p7, nrow=3, ncol=2)
```

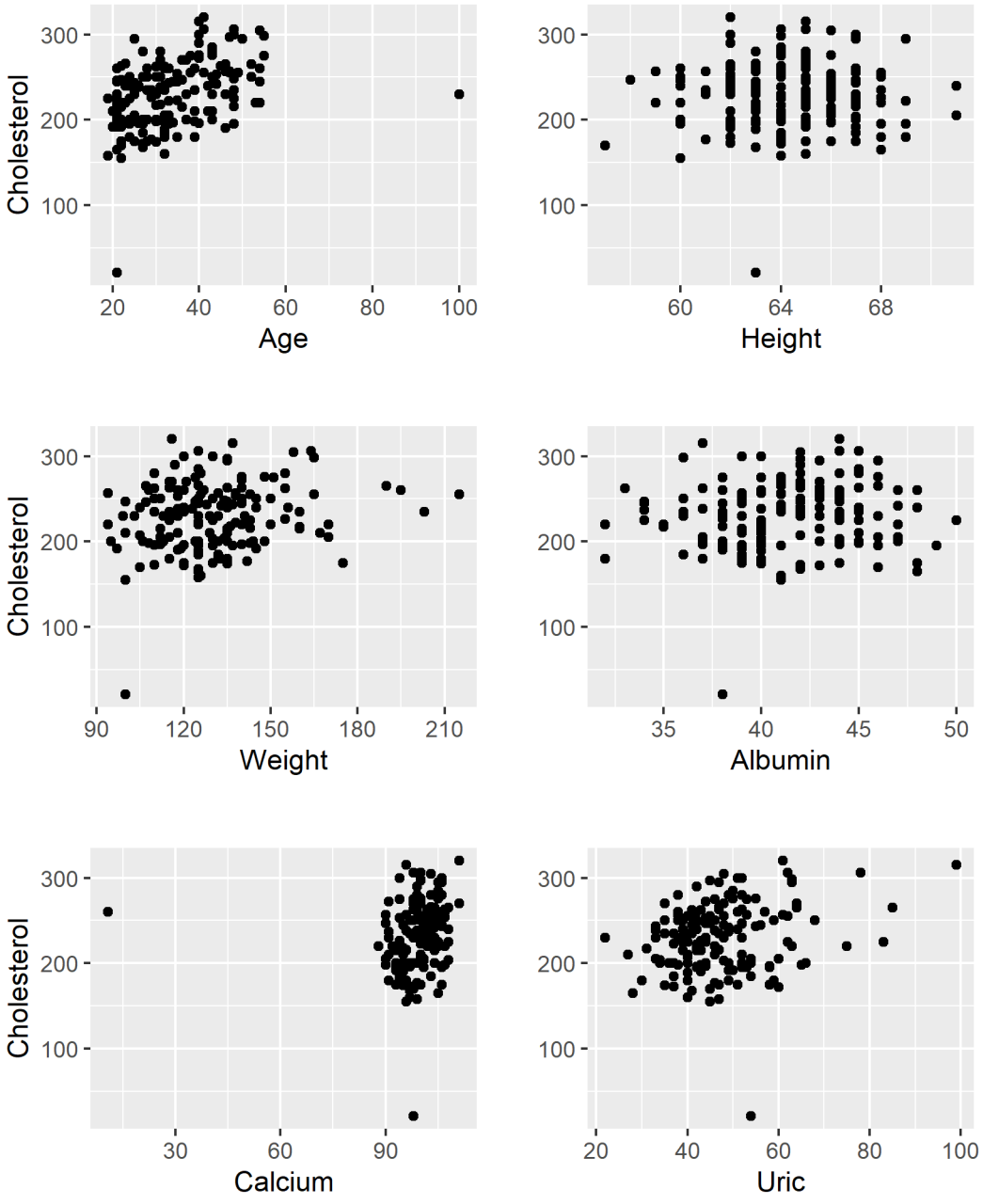
En la Figura 4 podemos observar los gráficos de dispersión de y respecto de las otras variables.

Apreciamos que para las variables *Albumin*, *Weight*, *Calcium* y *Height* no parece haber relación lineal no constante, es decir, para distintos valores de dichas variables, tenemos el mismo rango de valores para y. Esto equivaldría a que la esperanza de y condicionada por cualquiera de las variables es constante $E[y_i|x_i] = cte$, y por tanto, en un hipotético modelo RLS ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$) tendríamos que el coeficiente de regresión sería nulo. Tenemos entonces dos candidatos, la variable *Age* y la variable *Uric*. Para complementar nuestra decisión, nos apoyaremos en el coeficiente de correlación lineal de Pearson, para ver qué variable muestra mayor correlación lineal con y.

```
cor(datos)[, 'Cholesterol']

 [1] Age         Height      Weight         Pill  Cholesterol     Albumin
0.39002316  0.01379705  0.18933366  0.03368792  1.00000000  0.08527338
    Calcium        Uric
0.12538369  0.22994455
```

Confirmamos que la relación lineal mas fuerte en la figura 4 se corresponde con la variable *Age*, que, aun que se observan claramente distintos *outliers* en todas

Figura 4: Diagramas de dispersión con *Cholesterol* como variable respuesta



las variables, esta correlación es suficientemente fuerte en comparación al resto, como para descartar el resto de variables.

A partir de ahora usaremos *Age* como variable explicativa y la denotaremos simplemente como x.

En particular, si eliminamos el outlier de *Age*, que muestra una observación para una persona de 100 años de edad, vemos que la correlación lineal entre ambas variables aumenta

```
print(cor(datos$Age[datos$Age<100],datos$Cholesterol[datos$Age<100]))
[1] 0.4407396
```

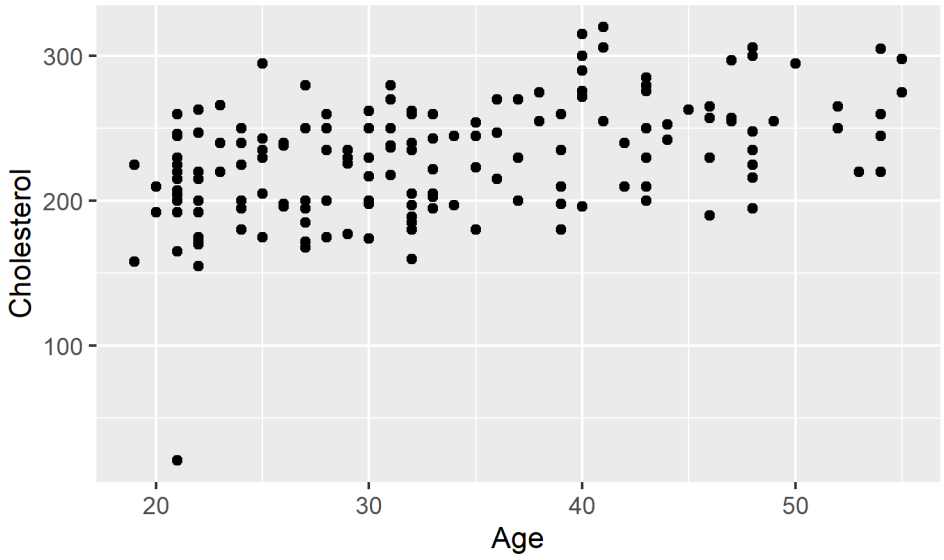
lo cual refuerza aun más nuestra decisión de tomar *Age* frente a *Uric*.

```
print(cor(datos$Uric[datos$Age<90],datos$Cholesterol[datos$Age<90]))
[1] 0.2299439
```

Observamos en la Figura 5 el gráfico de dispersión sin el outlier, donde se aprecia mejor la presunta relación lineal.

```
plot = qplot(datos$Age[datos$Age<100],
datos$Cholesterol[datos$Age<100]) +
labs(x='Age', y='Cholesterol')
```

Figura 5: Diagrama de dispersión sin el outlier en *Age*



Formulación del modelo

Por las conclusiones del apartado anterior, debido a que la correlación entre x e y es positiva, cuando aumente el valor de x, cabe esperar un aumento de

y, por tanto, para explicar las variaciones de y en función de x nos basaremos en la distribución de y condicionada por x, planteando un modelo RLS con las siguientes especificaciones:

1. Hipótesis de linealidad: estimaremos la esperanza de y condicionada por x mediante un modelo lineal, es decir, asumiremos que $E[y|x] = \beta_1 + \beta_0 x$.
2. Hipótesis de homocedasticidad: asumiremos que $V[y|x] = \sigma^2 = cte$.
3. Hipótesis de Normalidad: la distribución de $y|x$ es normal.
4. Hipótesis de independencia: las observaciones de cada y_i son independientes.

Estas hipótesis nos llevan a proponer una relación lineal entre las observaciones x_i e y_i del tipo:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + u_i$$

donde u_i es la perturbación del modelo debida a los errores de estimación de las esperanzas $E[y|x]$.

Comprobamos ahora las distintas especificaciones, puesto que para un mismo valor de x tenemos varios valores de y.

Hipótesis de Normalidad

Para esta hipótesis, vamos a usar el test de Shapiro-Wilks sobre los valores y_{ij} observados para cada x_j, siempre que haya almenos 5 valores de y_{ij}, para que el test sea relativamente fiable.

Código 6: Comprobación de la hipótesis de normalidad para cada valor de x.

```
valores_x = c()
pvalores = c()
cantidad_valores = c()
valores_x = c()

tests = c()

for(x in (unique(datos$Age))){

    valores_y = datos$Cholesterol[datos$Age == x]

    if (length(valores_y)>=5) {

        result = shapiro.test(valores_y)
        pvalores = append(pvalores, result$p.value)
        cantidad_valores = append(cantidad_valores, length(valores_y))
        valores_x = append(valores_x, x)
    }

}

print(data.frame(valores_x, pvalores, cantidad_valores))

[1] valores_x     pvalores       cantidad_valores
    22            0.3361973132          14
    25            0.8623191745           6
```

```
    21          0.0003627803        14
    24          0.6795858651         6
    27          0.1547350151         7
    30          0.9107667396         7
    31          0.8050598325         6
    32          0.4057686240        10
    33          0.4462040084         6
    39          0.8084913246         6
    40          0.1362547906         6
    43          0.3168065059         7
    48          0.3629610123         7
```

Aun que los test se realizan sobre tamaños muestrales pequeños, vemos que todos los *p-valores* son altos a excepción de uno de ellos ($x = 21$). En general, no podemos afirmar que la hipótesis de normalidad sea falsa, ya que un único p-valor bajo puede haberse debido a factores ambientales si lo comparamos con que ninguno de los otros *p-valores* son significativos para rechazar la hipótesis de normalidad, aun que repito, el test se ha aplicado sobre muestras pequeñas y debería comprobarse tal hipótesis cuando se recaben más datos.

Concluimos que no hay evidencia suficiente para falsar la hipótesis de normalidad en la distribución de y para cada x.

Hipótesis de Homocedasticidad

Para esta hipótesis nos basaremos en un procedimiento análogo al anterior pero aplicando el test de Barlett, contrastando si los conjuntos de valores $\{y_{ij}\}$ para cada x_j pertenecen a la misma población con varianza constante, imponiendo nuevamente que almenos haya 5 valores y_{ij} para x_j.

Código 7: Comprobación de la hipótesis de homocedasticidad para cada valor de x

```
valores = c()

for (x in unique(datos$Age)){

    valores_y = datos$Cholesterol[datos$Age==x]

    if(length(valores_y) > 5){
        valores[[length(valores)+1]] = c(valores_y)
    }

}
print(bartlett.test(valores))

[1]      Bartlett test of homogeneity of variances

data:  valores
Bartlett's K-squared = 11.28, df = 12, p-value = 0.5051
```

Obtenemos un p-valor muy superior a cualquier nivel de significación α habitual, y concluimos nuevamente que no hay evidencia suficiente para rechazar que todos los conjuntos de valores para x fijo pertenezcan a una población con varianza σ^2, almenos, para los valores de x para los que hay valores de y suficientes. Como en el caso anterior, sería recomendable recabar más datos, para poder incluir los demás valores de x y volver a comprobar la hipótesis.

En general, es probable que la muestra no se haya tomado bajo criterios de edad, si no atendiendo a algún criterio clínico, y por tanto, es esperable que aparezcan en la muestra mas individuos de una determinada edad que de otra, de manera que no hemos obtenido suficientes valores de y para determinados valores de x que nos permitan aplicar los contrastes anteriores al resto de edades, pero dada la poca evidencia para falsar ambas hipótesis con los datos disponibles, nada nos hace pensar que tales hipótesis no debieran cumplirse en los valores que no alcanzan el mínimo de 5 valores para aplicarse el contraste de normalidad u homocedasticidad.

Hipótesis de independencia

Para esta hipótesis asumimos que las observaciones y_i se han tomado de forma independiente durante el experimento, ya que al tratarse de datos sobre salud, no tendría sentido la dependencia entre las observaciones del colesterol de varios pacientes sea cual sea su edad, ya que una relación del tipo *si mi vecino tiene el colesterol alto, es probable que yo lo tenga alto/bajo* es poco razonable si los datos se han recabado bajo estrictos procedimientos de muestreo, seleccionando distintos pacientes y en el mismo momento, sin arbitrariedad.

Hipótesis de Linealidad

Ahora, comprobaremos si cabe esperar (poblacionalmente) una relación lineal entre la esperanza de y condicionada por x y los valores de x, a través de una estimación muestral, en particular, para cada valor de x_j cacularemos la media muestral

$$\bar{y}_i^{(j)} = \frac{1}{J}\sum_{j=1}^{J} y_{ij}$$

y las representaremos en el gráfico de dispersión para ver si cabe esperar una relación lineal del tipo $E[y|x] = \beta_0 + \beta_1 x$. Deberíamos observar, en la nube de puntos, una línea similar a una recta en caso de que tal relación exista. Por coherencia con las pruebas anteriores, usaremos solamente valores de x con al menos 5 observaciones de y para comprobar la linealidad.

Código 8: Comprobación de la hipótesis de linealidad

```
valores_x = c()
esperanzas = c()
count = 0

for (x in unique(datos$Age)){

    valores_y = c(datos$Cholesterol[datos$Age==x])

    if(length(valores_y)>5){

        esperanzas = append(esperanzas, mean(valores_y))
        valores_x = append(valores_x, x)
    }
```

```
}

esperanzas_cond = data.frame(valores_x, esperanzas)
names(esperanzas_cond) = c('Age_unique', 'esperanzas')

plot = ggplot(data = datos, aes(Age, Cholesterol)) +
geom_point() +
geom_point(data = esperanzas_cond,
aes(Age_unique, esperanzas), col='red', shape=15,size=2)+
geom_line(data = esperanzas_cond,
aes(Age_unique, esperanzas), col='red')
```

En la Figura 6 apreciamos cierta tendencia al alza que, efectivamente, confirma que cuanto mayor es una persona mayores son los niveles de colesterol, sin embargo, no se aprecia una tendencia lineal en las esperanzas condicionadas muestrales. La evidencia sería suficientemente fuerte como para rechazar, con

Figura 6: Diagrama de dispersión con la línea generada al unir las medias condicionadas muestrales

los datos actuales, la relación lineal en que se basa el modelo propuesto. Sin embargo, teniendo en cuenta que hemos despreciado muchos valores de x para los que no hay al menos 5 cinco valores de y, es muy posible que esto se deba a que la muestra no es suficientemente grande, y que al aumentar el tamaño muestral, efectivamente, exista tal relación, ya que la línea trazada, al menos, tiene tendencia creciente.

Pondremos esta hipótesis en entredicho y continuaremos adelante con el proceso, asumiendo linealidad, pero teniendo en cuenta que es posible que esta hipótesis no sea realmente cierta, y lo decidiremos finalmente comprobando los residuos del modelo ajustado.

Estimación

En esta parte, procedemos a la estimación de los parámetros del modelo β_1 y β_0, que segun las hipótesis anteriores, son:

$$\hat{\beta}_1 = \frac{cov(x,y)}{s_x^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Con el siguiente código obtenemos los parámetros del modelo y las sumas de cuadrados correspondientes, a saber, la varianza total VT, varianza explicada VE y varianza no explicada VNE.

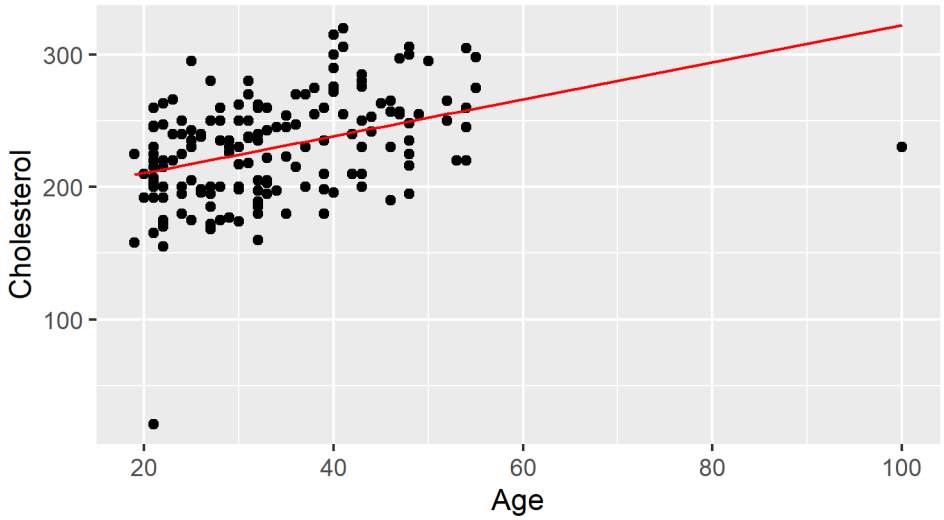
Código 9: Ajuste del modelo con outliers

```
x = datos$Age
y = datos$Cholesterol
n = length(y)

s_x = var(x)*((n-1)/n)
cov_xy = cov(x,y)*((n-1)/n)

beta_1 = cov_xy/s_x
beta_0 = mean(y) - beta_1*mean(x)

datos$ajustados = beta_0 + beta_1*datos$Age
datos$resid = datos$Cholesterol - datos$ajustados

VT = sum((y - mean(y))**2)
VE = sum((datos$ajustados - mean(y))**2)
VNE = sum((y - datos$ajustados)**2)

varianza_residual = VNE/(n-2)  # varianza residual

R_sq = VE/VT
S_r = sqrt(varianza_residual)  # error de estimacion de las esperanzas
    cond.

# se puede comprobar que VT = VNE + VE

plot = ggplot(datos, aes(Age, Cholesterol, ajustados)) + geom_point() +
    geom_line(aes(Age, ajustados), color='red')

print(c('beta_0:', beta_0))
print(c('beta_1:', beta_1))
print(c('R_sq:', VE/VT))
print(c('S_r:', S_r))

[1] "beta_0:"          "182.731532435508"
[1] "beta_1:"          "1.39369399357575"
[1] "R_sq:"            "0.152118064001415"
[1] "S_r:"             "36.8297225022208"
```

Vemos en la Figura 7 el resultado del ajuste.

Por otro lado, como la hipótesis de linealidad está en entredicho, no podemos asumir que el coeficiente de regresión tenga una distribución normal, y por tanto, las inferencias sobre él deben leerse con cuidado. De ahora en adelante siempre que hagamos referencia a inferencias sobre el verdadero valor de β_1,

Figura 7: Resultado del ajuste del modelo



vaya por delante la aserción *si la hipótesis de linealidad es cierta....* Realizamos ahora el contraste de regresión

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

cuyo estadístico de contraste es:

$$F^* = \frac{\hat{\beta}_1 n s_x^2}{\hat{s}_R^2} = \frac{\hat{\beta}_1 n s_x^2}{VNE/(n-2)} \sim F_{(1,n-2)}$$

de manera que el p-valor del test es:

$$P\{F_{(1,n-2)} > F^*\} = \alpha_C$$

Ejecutando el siguiente código obtenemos el resultado del test.

Código 10: Contraste de regresión para el modelo ajustado con outliers

```
s_x = var(x)*((n-1)/n)
s_r = VNE/(n-2)

F = n*(beta_1**2)*var(x)/s_r
print(c('F:', F))
print(c('pvalor:', pf(F, 1, n-2, lower.tail=FALSE)))

[1] "F:"                "29.242645868346"
[1] "pvalor:"           "2.25443021857548e-07"
```

Como se puede ver, el p-valor es suficientemente pequeño como para rechazar la hipótesis nula, y concluimos que $\beta_1 \neq 0$. Esto implica que el modelo, asumiendo $E[y|x] = \beta_0 + \beta_1 x$ explica mejor las observaciones que asumiendo $E[y|x] = \beta_0$.

Diagnóstico

A la vista de los resultados anteriores, estudiaremos los residuos (estandarizados por sencillez en la interpretación) y el coeficiente de determinación para juzgar la bondad del modelo.

Primeramente, cabe adelantar que el coeficiente de determinación $R^2 = \frac{VE}{VT}$ tiene un valor mas cercano a cero que a uno, por lo que no se puede concluir que el modelo explique adecuadamente las variaciones de y a través de los valores que toma x, lo vemos en el siguiente código.

Código 11: Cálculo del R^2 para el modelo ajustado con outliers.

```
VNE = sum((datos$ajustados - datos$Cholesterol)**2)
VT = sum((datos$Cholesterol - mean(datos$Cholesterol))**2)

R = (VT-VNE)/VT
print(R)

[1] 0.1521181
```

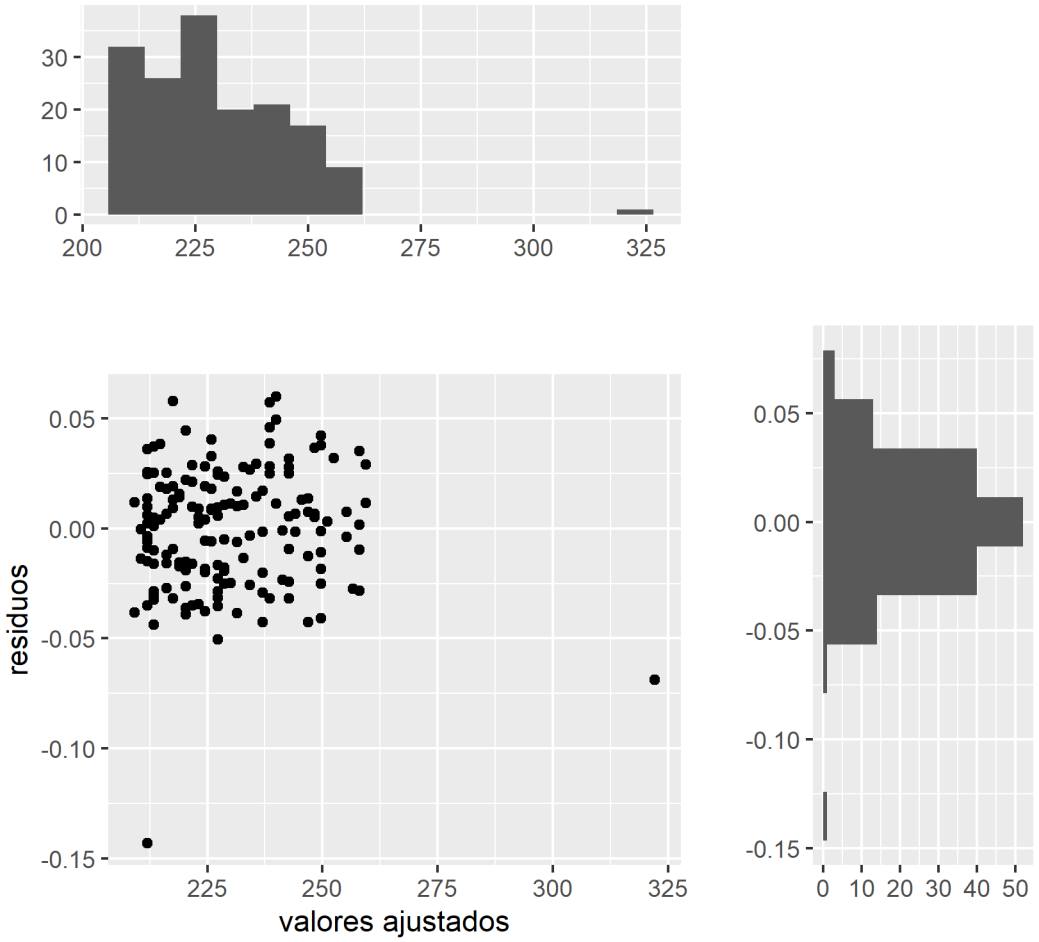
Veamos qué información aportan los residuos para comprender por qué el modelo no ha resultado ser adecuado. Para ello generamos un gráfico de los valores ajustados frente a los residuos y examinamos sendos histogramas en los margenes para visualizar ambas distribuciones.

Código 12: Cálculo del efecto palanca y residuos estandarizados

```
efecto_palanca = c()

for(x in datos$Age){

    h = 1/n + ((x - mean(datos$Age))**2)/(n*s_x)
    efecto_palanca = append(efecto_palanca, h)
}

datos$efecto_palanca = efecto_palanca
datos$residuos_estandarizados = datos$resid/(s_r*sqrt(1-h))


hist_top = ggplot() + geom_histogram(aes(datos$pred), bins=15) + labs(x=
    '', y='')
hist_right = ggplot() + geom_histogram(aes(datos$residuos_estandarizados
    ), bins=10) + coord_flip() + labs(x='', y='')
empty <- ggplot()+geom_point(aes(1,1), colour="white")+
         theme(axis.ticks=element_blank(),
               panel.background=element_blank(),
               axis.text.x=element_blank(), axis.text.y=element_blank(),
               axis.title.x=element_blank(), axis.title.y=element_blank
                   ())
scatter = ggplot() + geom_point(aes(datos$ajustados, datos$residuos_
    estandarizados)) + geom_abline() +
labs(x='valores_ajustados', y='residuos', title='')

plot = ggarrange(hist_top, empty, scatter, hist_right, ncol=2, nrow=2,
    widths=c(2, 1), heights=c(1, 2))
```

En la Figura 8 observamos en el margen derecho que los residuos son normales, cosa que cabía esperar por las hipótesis de normalidad impuestas sobre la distribución condicionada $y|x$.

Figura 8: Residuos estandarizados del modelo frente a los valores ajustados junto con las distribuciones marginales y outliers incluidos.



Además en el margen superior del histograma se puede deducir que los residuos están distribuidos uniformemente a lo largo de todos los valores ajustados.

Si que conviene destacar que los residuos no están muy pegados a cero, lo que indica una alta varianza residual $\sigma^2 = V[u] = V[y|x]$ y debe hacernos sospechar que el modelo no explica bien las medias de y condicionadas por x.

Cabe destacar dos outliers muy marcados que debemos examinar y plantearnos que es posible que hayan comprometido los resultados. Para ver qué efecto

tractor tienen estos outliers sobre la recta de regresión, usamos el efecto palanca

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}$$

Examinando en el Código 12 el efecto palanca de las observaciones, vemos que solamente existe un punto con $h_i > \frac{6}{n} \approx 0{,}041$, que es el paciente con 100 años de edad (el otro outlier no parece tener un efecto palanca reseñable).

Código 13: Comprobación de las observaciones con alto efecto palanca

```
filter(datos, datos$efecto_palanca>6/n)

Age Height Weight Pill Cholesterol Albumin Calcium Uric
100     62    125    0         230      46     104   48
ajustados    resid efecto_palanca residuos_estandarizados
322.1009 -92.10093 0.2231809              -0.06881957
```

Como a día de hoy resulta extraño encontrar personas de esa edad, vamos a prescindir de este paciente para un nuevo modelo y recalcular los coefcientes de la recta y el R^2, para ver si hay diferencias significativas.

Código 14: Ajuste de un segundo modelo eliminando las observaciones con alto efecto palanca

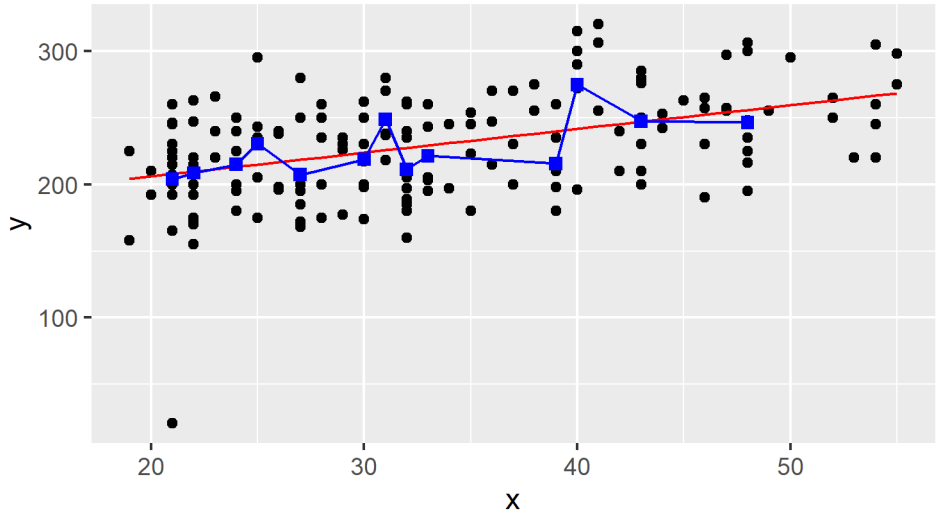
```
x = datos$Age[datos$Age < 100]  # nuevos x filtrando por Age < 100
y = datos$Cholesterol[datos$Age < 100] # nuevos y filtrando por Age <
    100
datos_modelo2 = data.frame(x,y)

n = length(y)

s_x = var(x)*((n-1)/n)
cov_xy = cov(x,y)*((n-1)/n)

beta_1 = cov_xy/s_x
beta_0 = mean(y) - beta_1*mean(x)

datos_modelo2$ajustados = beta_0 + beta_1*datos_modelo2$x
datos_modelo2$resid = datos_modelo2$y - datos_modelo2$ajustados

VT = sum((y - mean(y))**2)
VE = sum((datos_modelo2$ajustados - mean(y))**2)
VNE = sum((y - datos_modelo2$ajustados)**2)

R_sq = VE/VT
varianza_residual = VNE/(n-2)
S_r = sqrt(varianza_residual)

# se puede comprobar que VT = VNE + VE

plot = ggplot(datos_modelo2, aes(x, y, ajustados), fill=x) +
geom_point(color='black') +
geom_line(aes(x, ajustados), color='red')+
geom_point(data = esperanzas_cond, aes(Age_unique, esperanzas), col='
    blue', shape=15, size=2)+
geom_line(data = esperanzas_cond, aes(Age_unique, esperanzas), col='blue
    ')

print(c('beta_0:', beta_0))
print(c('beta_1:', beta_1))
print(c('R_sq:', VE/VT))
print(c('S_r:', S_r))
```

```
[1] "beta_0:"          "170.417727172176"
[1] "beta_1:"          "1.78143894919381"
[1] "R_sq:"            "0.194251436007729"
[1] "S_r:"             "36.0142979357107"
```

Se puede ver que efectivamente se producen cambios significativos en la pendiente, la ordenada en el origen y en el R^2, resultando un modelo totalmente distinto, pero con mayor proporción de varianza explicada R^2, en la Figura 9 se puede ver el nuevo ajuste. Esta vez, el modelo parece más razonable, ya que:

- El coeficiente de regresión ha aumentado, pasando a tener mayor pendiente, lo que se traduce en que la media de colesterol condicionada por la edad aumenta más que con el modelo anterior.
- La ordenada en el origen ha disminuido, pero carece interpretación en nuestro caso, puesto que una persona de cero años queda fuera del rango de edad durante el experimento $x \in [19, 100]$.
- El coeficiente de determinación ha aumentado, por lo que el modelo explica mayor proporción de varianza, sin embargo el valor $R^2 \approx 0{,}2$ continua sin ser significativo puesto que continua siendo más cercano a 0 que a 1.
- La varianza residual $\hat{s}_R$ apenas ha variado, esto nos da a entender que apenas hemos reducido el error de estimación de las medias condicionadas, esto puede ser efecto de la falta de linealidad en ambos modelos como vemos en la Figura 9.

Figura 9: segundo modelo (sin el paciente de 100 años) ajustado y con la línea de medias de y condicionas a los valores de x



Conclusiones

En teoría deberíamos interpretar el segundo modelo (el mejor) de la siguiente manera: cuando aumenta la edad de una persona (*Age*) en una unidad, el colesterol (*Cholesterol*) aumenta en $1.78 \approx 2$ unidades.

Aunque a la vista de los resultados obtenidos, y a pesar de que se cumplen las especificaciones de un modelo RLS (salvo la linealidad que está en duda por factores ambientales), el modelo no es capaz de explicar adecuadamente la variabilidad de *Cholesterol* únicamente a través de los valores de *Age*, ya que apenas explica, en el mejor de los casos, el 20 % de la varianza de *Cholesterol* ($R^2 \approx 0{,}2$).

Además, la estimación de la desviación típica residual es muy alta, $\hat{s}_R \approx 36$, lo que sugiere errores demasiado altos al estimar las medias condicionadas (a pesar de las unidades de medida).

Por otro lado, el contraste de regresión nos indica sin lugar a dudas que el coeficiente $\hat{\beta}_1$ no es nulo, por lo que *Age* es una variable que puede formar parte de un modelo más general. Dada la posible relación lineal que muestran otras variables con *Cholesterol*, la ineficacia del modelo puede justificarse con que deberíamos tener en cuenta otras variables que, aun que en el gráfico de la figura 4 haga sospechar que en un modelo simple fuera $\beta_1 \approx 0$, en espacios de mayor dimensión es muy posible que esto no sea así, de manera que un hipotético modelo del tipo

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

podría explicar mejor la variabilidad de y.