Trabajo Práctico: Regresión Lineal

Bodean-Ojeda

10 de junio de 2019

Enunciado del Problema:

El conjunto de datos 'Telefonía 2019' tiene datos de clientes que contrataron un servicio de telefonía fija. Nos contrataron para analizar si es posible estimar adecuadamente el ingreso del grupo familiar al que se presta el servicio, a partir de las variables disponibles. Esto permitiría a la empresa de telefonía, gran accionista de cierto banco, realizar una campaña entre potenciales clientes. El objetivo final de este trabajo es entregar un modelo que sea capaz de realizar la predicción del ingreso.

0. Carga de datos

```
Telefonia_2019 <- read_excel("TP regresion/Telefonia 2019.xls")
kable(head(Telefonia_2019))
```

id	ingresos	retenc	edad	educ	empleo	personas	internet	gastoNac	gastoIN
200	360	72	66	3	41	7	1	90.20521	49.382007
278	218	61	46	3	18	7	0	47.61513	22.438351
216	200	62	46	3	9	7	0	50.28847	6.839271
466	199	71	46	4	7	7	1	51.49550	10.315871
326	245	72	54	1	32	6	0	37.85180	40.696107
677	194	10	42	3	9	2	1	45.13971	2.384162

1. Realice previamente los análisis univariados/bivariados que crea conveniente.

```
summary(Telefonia_2019)
```

```
##        id            ingresos          retenc           edad
##  Min.   :   1.0   Min.   :  9.00   Min.   : 1.00   Min.   : 6.00
##  1st Qu.: 250.8   1st Qu.: 30.00   1st Qu.:17.00   1st Qu.:32.00
##  Median : 500.5   Median : 48.00   Median :34.00   Median :40.00
##  Mean   : 500.5   Mean   : 65.17   Mean   :35.53   Mean   :41.63
##  3rd Qu.: 750.2   3rd Qu.: 86.00   3rd Qu.:54.00   3rd Qu.:51.00
##  Max.   :1000.0   Max.   :360.00   Max.   :72.00   Max.   :77.00
##       educ           empleo         personas        internet
##  Min.   :1.000   Min.   : 0.00   Min.   :0.000   Min.   :0.000
##  1st Qu.:2.000   1st Qu.: 3.00   1st Qu.:5.000   1st Qu.:0.000
##  Median :3.000   Median : 8.00   Median :6.000   Median :0.000
##  Mean   :2.371   Mean   :10.99   Mean   :5.669   Mean   :0.418
##  3rd Qu.:3.000   3rd Qu.:17.00   3rd Qu.:7.000   3rd Qu.:1.000
##  Max.   :4.000   Max.   :47.00   Max.   :7.000   Max.   :1.000
##     gastoNac        gastoIN
##  Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 3.14   1st Qu.: 4.113
```

```
##  Median :11.00   Median :10.586
##  Mean   :13.87   Mean   :13.105
##  3rd Qu.:20.40   3rd Qu.:19.923
##  Max.   :90.21   Max.   :59.058
```

```
str(Telefonia_2019)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1000 obs. of  10 variables:
##  $ id      : num  200 278 216 466 326 677 785 696 902 155 ...
##  $ ingresos: num  360 218 200 199 245 194 49 93 207 242 ...
##  $ retenc  : num  72 61 62 71 72 10 67 72 58 38 ...
##  $ edad    : num  66 46 46 46 54 42 77 75 53 53 ...
##  $ educ    : num  3 3 3 4 1 3 1 1 3 3 ...
##  $ empleo  : num  41 18 9 7 32 9 45 44 12 12 ...
##  $ personas: num  7 7 7 7 6 2 7 7 7 7 ...
##  $ internet: num  1 0 0 1 0 1 0 0 1 0 ...
##  $ gastoNac: num  90.2 47.6 50.3 51.5 37.9 ...
##  $ gastoIN : num  49.38 22.44 6.84 10.32 40.7 ...
```

- Detalle de las variables:

Nombre	Tipo	Descripción
id		Identificación
retenc	Continua	Meses dentro del servicio
edad	Continua	Edad del cliente que contrató el servicio
ingresos	Continua	Ingresos del grupo familiar en el último mes en miles
educ	Ordinal	Máximo nivel educativo alcanzado
empleo	Continua	Años en el actual empleo
personas	Discreta	Cantidad de personas en el hogar
gastoIN	Continua	Consumo internacional último mes
gastoNac	Continua	Consumo larga distancia nacional último mes
internet	Nominal	Si contrató el paquete con servicio de Internet

Codificación Educación:

#	Descripción
1	Secundario incompleto
2	Secundario completo
3	Terciario/universitario
4	Posgrado

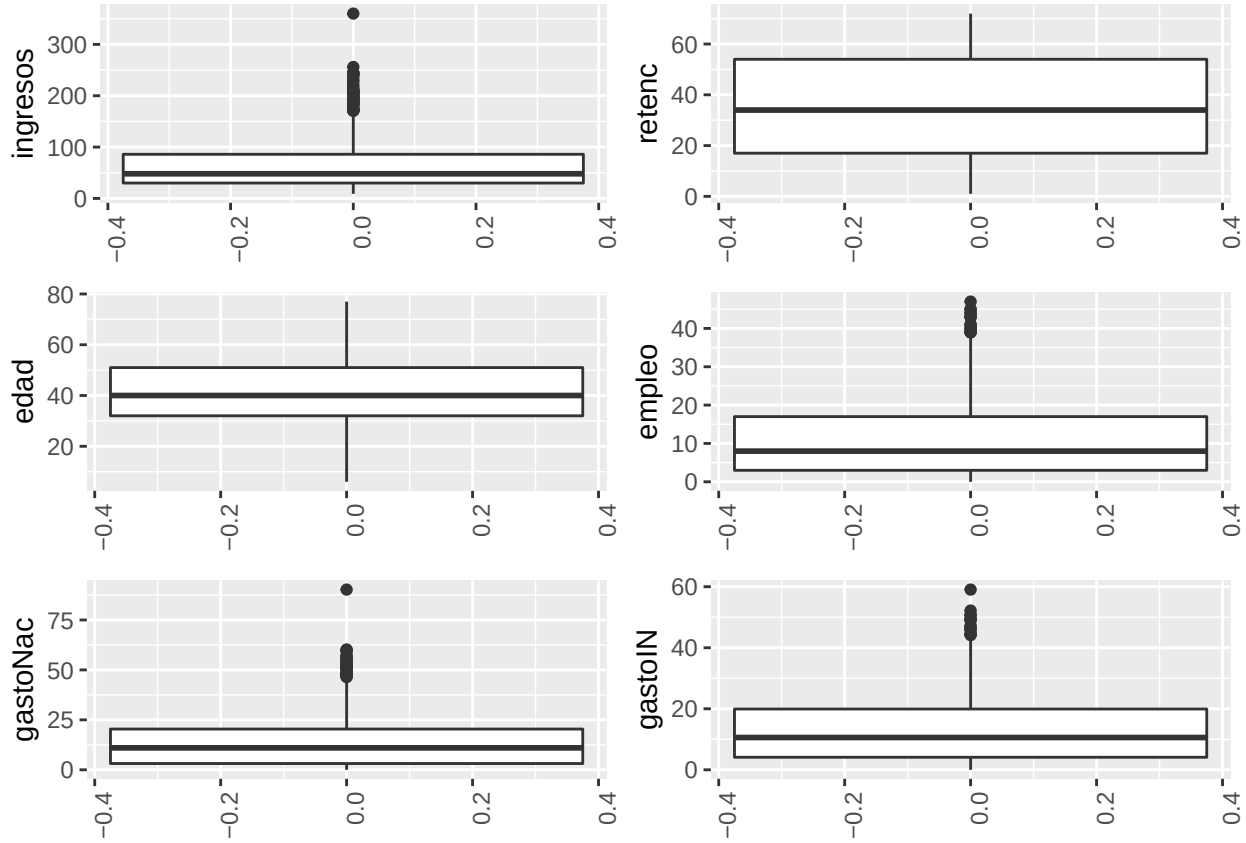
Codificación Internet:

#	Descripción
0	No

Observamos que los tipos de variables en R no corresponen con el enunciado del problema, deberan ser cambiadas antes de usarlas.

- Realizamos BoxPlot de las variables continuas:

```
plot_1 <- ggplot(Telefonia_2019,aes(y=ingresos))+geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
plot_2 <- ggplot(Telefonia_2019,aes(y=retenc))+geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
plot_3 <- ggplot(Telefonia_2019,aes(y=edad))+geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
plot_4 <- ggplot(Telefonia_2019,aes(y=empleo))+geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
plot_5 <- ggplot(Telefonia_2019,aes(y=gastoNac))+geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
plot_6 <- ggplot(Telefonia_2019,aes(y=gastoIN))+geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

gridExtra::grid.arrange(plot_1, plot_2, plot_3, plot_4, plot_5, plot_6,
                        ncol=2, nrow=3)
```



Podemos observar que algunas variables tienen valores atípicos, esto puede traer problema para la regresión. También observamos que las distribuciones no parecen ser normales.

- Realizamos tablas de frecuencia de las variables no continuas:

```
kable(table(Telefonia_2019$internet), caption = "Tabla de Frecuencia - Internet") %>%
  kable_styling(latex_options =c("hold_position"))
```

Table 4: Tabla de Frecuencia - Internet

Var1	Freq
0	582
1	418

```
kable(table(Telefonia_2019$educ), caption = "Tabla de Frecuencia - Educación") %>%
  kable_styling(latex_options =c("hold_position"))
```

Table 5: Tabla de Frecuencia - Educación

Var1	Freq
1	204
2	287
3	443
4	66

```
kable(table(Telefonia_2019$personas), caption = "Tabla de Frecuencia - Personas") %>%
  kable_styling(latex_options =c("hold_position"))
```

Table 6: Tabla de Frecuencia - Personas

Var1	Freq
0	2
1	4
2	29
3	60
4	120
5	138
6	272
7	375

Podemos observar que los datos son todos correctos, no poseemos datos nulos.

- Correlación entre las variables:

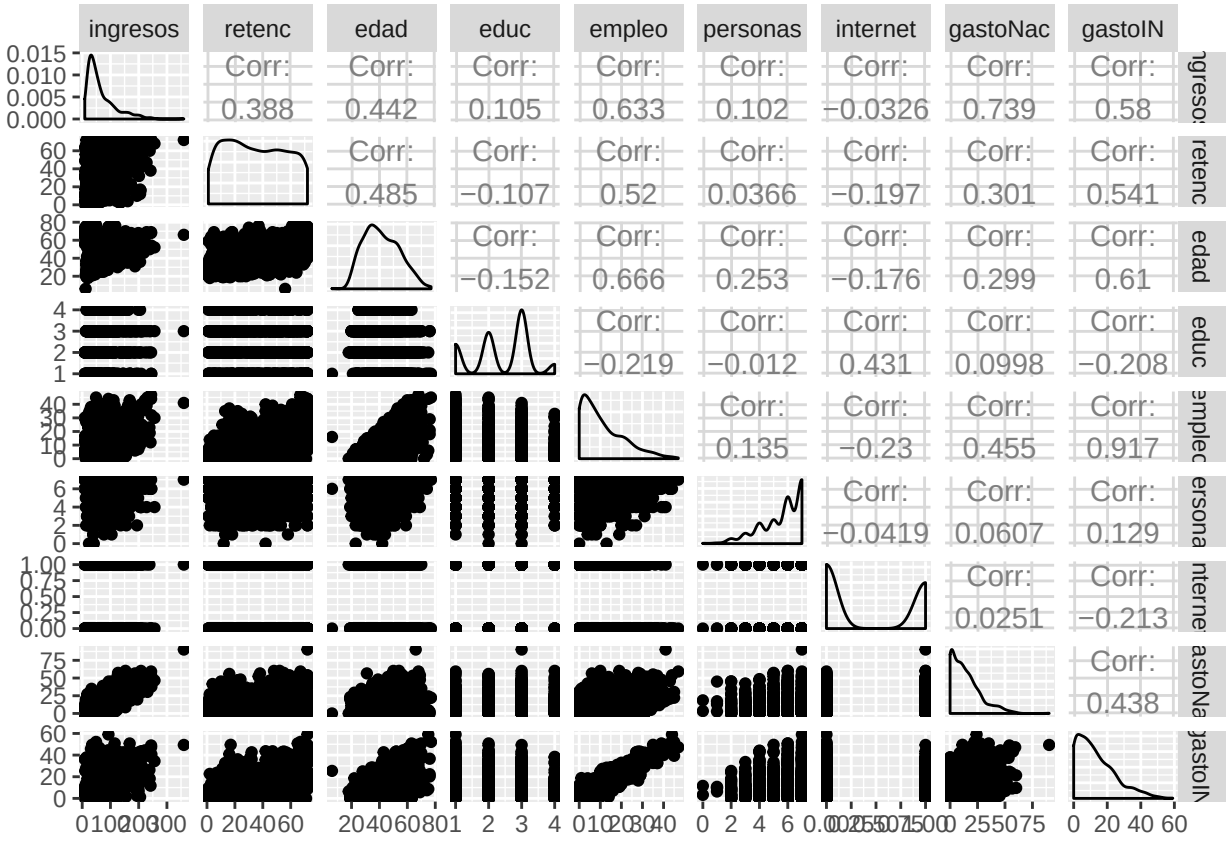
```
kable(round(cor(x=Telefonia_2019[-1], method = "pearson"), 3), caption = "Matriz de Correlación") %>%
```

Table 7: Matriz de Correlación

	ingresos	retenc	edad	educ	empleo	personas	internet	gastoNac	gastoIN
ingresos	1.000	0.388	0.442	0.105	0.633	0.102	-0.033	0.739	0.580
retenc	0.388	1.000	0.485	-0.107	0.520	0.037	-0.197	0.301	0.541
edad	0.442	0.485	1.000	-0.152	0.666	0.253	-0.176	0.299	0.610
educ	0.105	-0.107	-0.152	1.000	-0.219	-0.012	0.431	0.100	-0.208
empleo	0.633	0.520	0.666	-0.219	1.000	0.135	-0.230	0.455	0.917
personas	0.102	0.037	0.253	-0.012	0.135	1.000	-0.042	0.061	0.129
internet	-0.033	-0.197	-0.176	0.431	-0.230	-0.042	1.000	0.025	-0.213
gastoNac	0.739	0.301	0.299	0.100	0.455	0.061	0.025	1.000	0.438
gastoIN	0.580	0.541	0.610	-0.208	0.917	0.129	-0.213	0.438	1.000

Realizamos una gráfica tipo Pairs, que muestra la correlación y las gráficas de dispersión entre variables.

```
ggpairs(Telefonia_2019[,-1])
```



De las gráficas se observa que podría haber una relación lineal entre los pares de variables ingresos-gastoNac y empleo-gastoIN. Además su valor de correlación son los mas altos.

- Hacemos un test para probar si hay correlación lineal entre las variables

```
cor.test(x = Telefonia_2019$empleo,   y = Telefonia_2019$ingresos, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Telefonia_2019$empleo and Telefonia_2019$ingresos
## t = 25.817, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5941016 0.6685555
## sample estimates:
##       cor
## 0.6327889
```

```
cor.test(x = Telefonia_2019$gastoNac, y = Telefonia_2019$ingresos, method = "pearson")
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  Telefonia_2019$gastoNac and Telefonia_2019$ingresos
## t = 34.688, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7098851 0.7662156
## sample estimates:
##       cor
## 0.7393412
```

En los dos casos obtenemos un p-valor ($< 2.2e\text{-}16$) menor a 0.05, tendría sentido intentar generar un modelo de regresión lineal con alguna de estas variables.

2. Desarrolle un modelo de regresión que relacione el ingreso del grupo familiar con la variable continua que esté más correlacionada con esta.

Del análisis del punto anterior podemos ver que la variable más correlacionada con ingresos es gastoNac, 0.739.

```
modelo_gastoNac <- lm(ingresos~gastoNac,data = Telefonia_2019)
summary(modelo_gastoNac)
```

```
##
## Call:
## lm(formula = ingresos ~ gastoNac, data = Telefonia_2019)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.422 -21.708  -2.554  17.086 119.773
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.55405    1.53320   17.32   <2e-16 ***
## gastoNac     2.78380    0.08025   34.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.33 on 998 degrees of freedom
## Multiple R-squared:  0.5466, Adjusted R-squared:  0.5462
## F-statistic:  1203 on 1 and 998 DF,  p-value: < 2.2e-16
```

En la salida podemos ver lo siguiente:

- El valor estimado para los dos parámetros de la ecuación del modelo lineal (Beta0 y Beta1) que equivalen a la ordenada en el origen y la pendiente (Columna "Estimate").
- Para el modelo generado, tanto la ordenada en el origen como la pendiente son significativas (p-values < 0.05), tienen importancia en el modelo.
- El valor de R2 indica que el modelo calculado explica el 54.66% de la variabilidad presente en la variable respuesta (ingresos) mediante la variable independiente (gastoNac).

- El p-value obtenido en el test F (< 2.2e-16) determinaría que el modelo es significativo y por lo tanto se podría aceptar si se cumplen con los supuestos de la regrasión lineal.

Solo para comparar, volvemos a generar el modelo sin el intercepto.

```
modelo_gastoNac2 <- lm(ingresos~gastoNac-1,data = Telefonia_2019)
summary(modelo_gastoNac2)
```

```
##
## Call:
## lm(formula = ingresos ~ gastoNac - 1, data = Telefonia_2019)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -98.99 -10.62  14.77  34.42 120.46
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## gastoNac  3.79312    0.06289   60.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.99 on 999 degrees of freedom
## Multiple R-squared:  0.7845, Adjusted R-squared:  0.7843
## F-statistic:  3638 on 1 and 999 DF,  p-value: < 2.2e-16
```

Metricas de comparación de modelos:

```
glance(modelo_gastoNac) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC   p.value
##           <dbl> <dbl> <dbl> <dbl>     <dbl>
## 1         0.546  33.3 9855. 9870. 1.27e-173
```

```
glance(modelo_gastoNac2) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma    AIC    BIC p.value
##           <dbl> <dbl>  <dbl>  <dbl>   <dbl>
## 1         0.784  38.0 10116. 10126.       0
```

Teniendo en cuenta las métricas consideradas anterior el *adj.r.squared* mejora quitando el intercepto, por lo tanto se explica mejor el la variación de la salida con respecto a la variable predictora ajustada al nro de variables. Sin embargo las métricas raiz cuadrada de varianza en los residuos (σ) como los criterios de informacion tanto de Akaike como Bayesiano (AIC, BIC) empeoran.

a. ¿Cuál es la ecuación resultante?

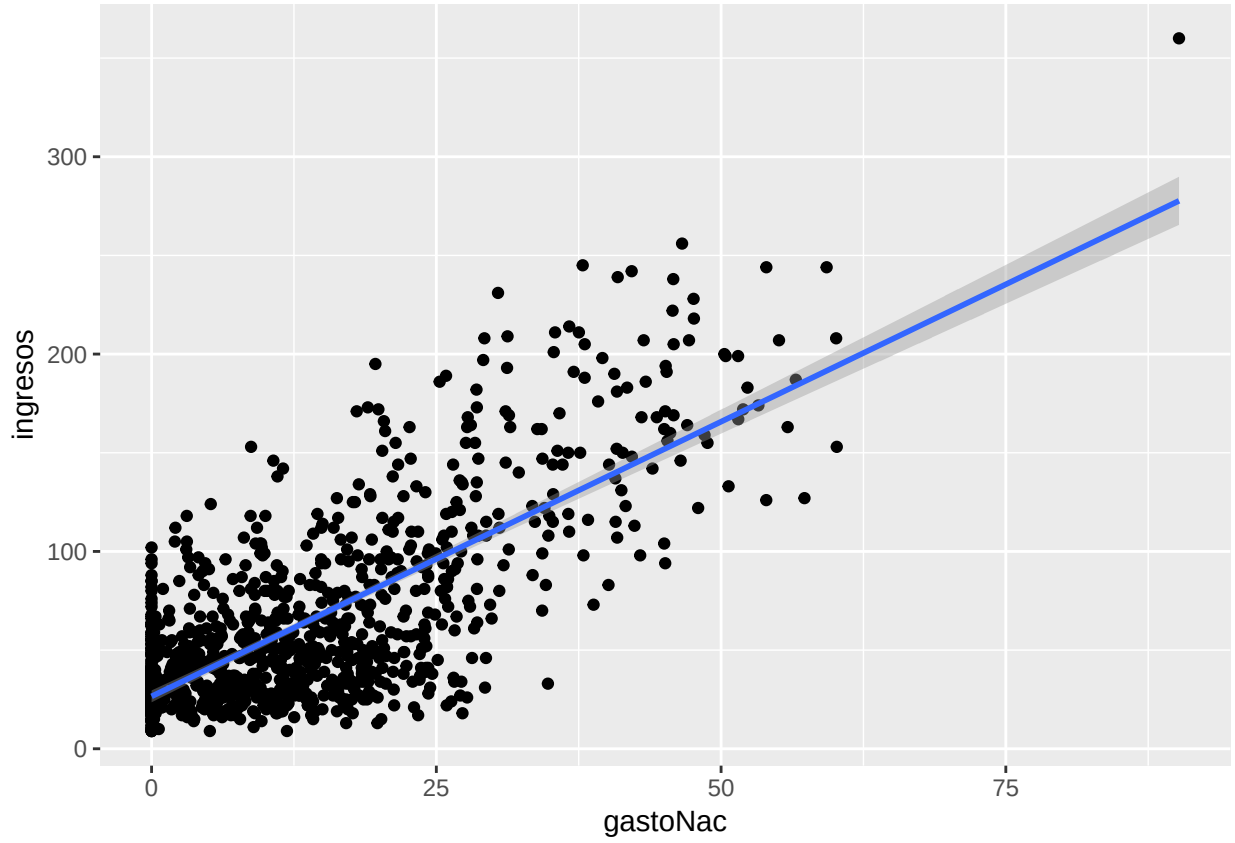
Ecuación resultante del primer modelo con el intercepto.

$$ingresos = 26.5540498 + 2.7838024 * gastoNac$$

Podemos decir que en promedio, cada incremento en una unidad de gastoNac, corresponde a un incremento de 2.78 del ingreso.

b. Diseñe un gráfico que incluya la recta y el intervalo de confianza del 95% para la media (con línea punteada o similar).

```
Telefonia_2019 %>%
  ggplot(aes(x=gastoNac, y=ingresos)) +
  geom_point() +
  geom_smooth(method='lm')
```

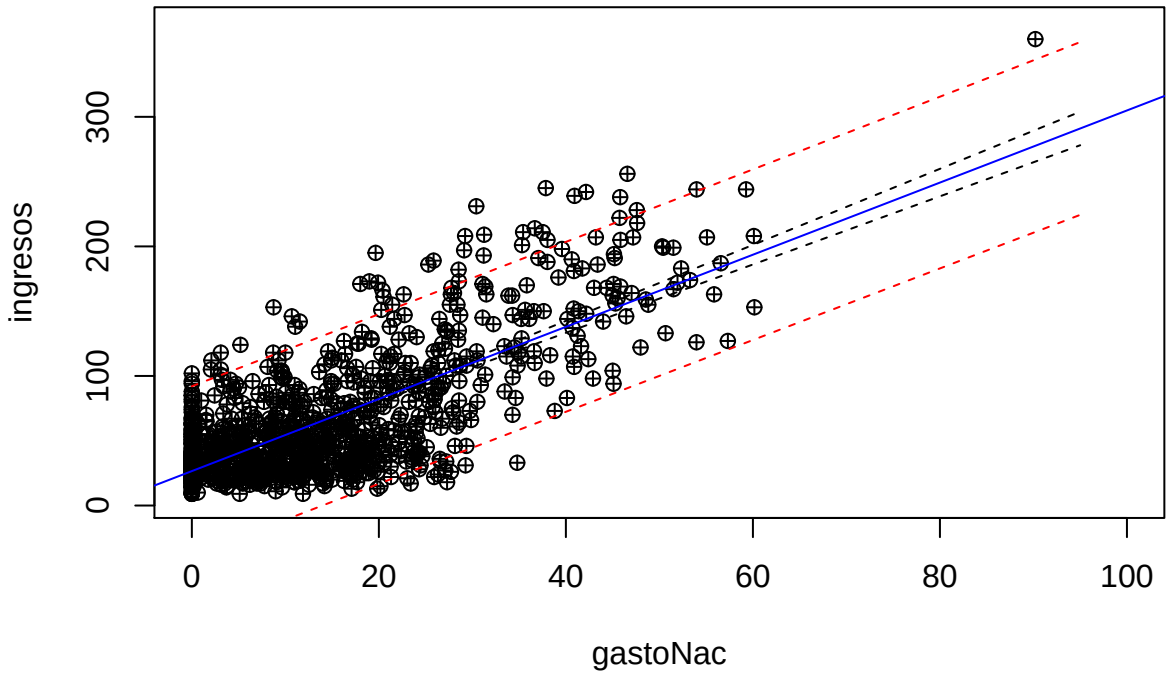


En el grafico podemos ver un punto aislado, más remoto en las x. Esto puede afectar el ajuste por mínimos cuadrados al generar el modelo

Realizamos una gráfica agregando el intervalo de predicción.

```
#I confianza para la media y de predicción para la respuesta
res.pred1 <- predict(modelo_gastoNac,
                     list(gastoNac= c(0:95)),
                     interval="confidence")
```

```
res.pred2 <- predict(modelo_gastoNac,
                     list(gastoNac= c(0:95)),
                     interval="prediction")
#Graficando bandas de confianza y bandas de predicción
par(mfrow=c(1,1))
plot(ingresos~gastoNac, data=Telefonia_2019, xlim=c(0,100),
     ylim=c(5,370), pch=10)
abline(modelo_gastoNac,col="blue")
lines(c(0:95), res.pred1[, 2], lty = 2)
lines(c(0:95), res.pred1[, 3], lty = 2)
lines(c(0:95), res.pred2[, 2], lty = 2, col = "red")
lines(c(0:95), res.pred2[, 3], lty = 2, col = "red")
```



Se puede observar, como indica la teoria, que el intervalo de predición es más ancho que el intervalo de confianza de la media, y que incluye a l mayoría de la observaciones dentro del intervalo.

3. Desarrolle ahora un modelo de regresión que relacione la variable ingreso con todas las variables continuas y también la de cantidad de personas en el hogar.

Generamos dos modelos, uno tomando la variable personas como factor (varible dummie) y otro tomando la variable como continua.

```
modelo_3 = lm( ingresos ~ retenc + edad + empleo + factor(personas)
                     + gastoNac + gastoIN, data = Telefonia_2019)
summary(modelo_3)
```

```
##
## Call:
## lm(formula = ingresos ~ retenc + edad + empleo + factor(personas) +
##     gastoNac + gastoIN, data = Telefonia_2019)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.621 -18.772  -1.722  14.850 110.748
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.27996   20.63962  -0.353   0.7244
## retenc             0.07421    0.05311   1.397   0.1626
## edad               0.11828    0.10323   1.146   0.2521
## empleo             2.00463    0.24528   8.173 9.19e-16 ***
## factor(personas)1  8.40358   25.04148   0.336   0.7373
## factor(personas)2 16.40275   21.13430   0.776   0.4379
## factor(personas)3 16.26334   20.77677   0.783   0.4340
## factor(personas)4 21.69762   20.61236   1.053   0.2928
## factor(personas)5 12.52798   20.59450   0.608   0.5431
## factor(personas)6 17.90050   20.53291   0.872   0.3835
## factor(personas)7 18.79522   20.51107   0.916   0.3597
## gastoNac           2.15576    0.07875  27.376  < 2e-16 ***
## gastoIN           -0.36422    0.20732  -1.757   0.0793 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.9 on 987 degrees of freedom
## Multiple R-squared:  0.6629, Adjusted R-squared:  0.6588
## F-statistic: 161.8 on 12 and 987 DF,  p-value: < 2.2e-16
```

```
modelo_3_2 = lm( ingresos ~ retenc + edad + empleo + personas
                    + gastoNac + gastoIN, data = Telefonia_2019)
summary(modelo_3_2)
```

```
##
## Call:
## lm(formula = ingresos ~ retenc + edad + empleo + personas + gastoNac +
##     gastoIN, data = Telefonia_2019)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.496 -18.940  -1.636  14.753 111.536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.92919    4.58839   1.728   0.0843 .
## retenc       0.07782    0.05281   1.474   0.1409
## edad         0.11027    0.10286   1.072   0.2840
## empleo       2.05042    0.24420   8.396   <2e-16 ***
## personas     0.48333    0.66295   0.729   0.4661
## gastoNac     2.14268    0.07859  27.265   <2e-16 ***
## gastoIN     -0.38945    0.20612  -1.889   0.0591 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.93 on 993 degrees of freedom
## Multiple R-squared:  0.6602, Adjusted R-squared:  0.6581
## F-statistic: 321.5 on 6 and 993 DF,  p-value: < 2.2e-16
```

a. ¿Es el modelo significativo?

Podemos observar que ambos modelos tiene un R2 mayora 0.65, y los p-valor obtenidos en el test F (< 2.2e-16) determinarían que los modelos son significativos y por lo tanto se podría aceptar si se cumplen con los supuestos de la regrasión lineal.

Podemos ver por los p-valor que las variables empleo y gastosNac son las más significativas. Tienen p-valor menor a 0.05.

Hacemos tambien una prueba de significancia de la regresión con ANOVA.

```
anova(modelo_3)
```

```
## Analysis of Variance Table
##
## Response: ingresos
##                   Df Sum Sq Mean Sq  F value  Pr(>F)
## retenc             1 368025  368025 440.5734 < 2e-16 ***
## edad               1 205944  205944 246.5424 < 2e-16 ***
## empleo             1 417344  417344 499.6154 < 2e-16 ***
## factor(personas)   7   3719     531   0.6360 0.72637
## gastoNac           1 623769  623769 746.7332 < 2e-16 ***
## gastoIN            1   2578    2578   3.0862 0.07927 .
## Residuals        987 824471     835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modelo_3_2)
```

```
## Analysis of Variance Table
##
## Response: ingresos
##            Df Sum Sq Mean Sq  F value  Pr(>F)
## retenc      1 368025  368025 439.6823 < 2e-16 ***
## edad        1 205944  205944 246.0438 < 2e-16 ***
## empleo      1 417344  417344 498.6049 < 2e-16 ***
## personas    1    687     687   0.8206 0.36524
## gastoNac    1 619698  619698 740.3584 < 2e-16 ***
## gastoIN     1   2988    2988   3.5702 0.05912 .
## Residuals 993 831165     837
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En cuatro variables obtuvimos un valor grande de F, y un p-valor muy chico(< 2e-16) indicando que la regresión es significativa. Solo nos falta validar los supuestos.

b. ¿Qué variables son significativas al 5%?

Filtramos las Variables con p-valor menor a 0.05:

```
kable(summary(modelo_3)$coefficients[
  which(summary(modelo_3)$coefficients[,4]<0.05),4],
  col.names = "p-valor", digits = 16)
```

	p-valor
empleo	9e-16
gastoNac	0e+00

Observamos también si los intervalos de confianza de los coeficients contienen al 0.

```
confint(modelo_3_2)
```

```
##                   2.5 %      97.5 %
## (Intercept) -1.07485540 16.93323644
## retenc      -0.02580354  0.18144194
## edad        -0.09158141  0.31212920
## empleo       1.57120494  2.52963386
## personas    -0.81760928  1.78426100
## gastoNac     1.98846311  2.29689756
## gastoIN     -0.79392514  0.01501913
```

Observamos que las dos variables que no incluyen al cero en su intervalo de confianza son empleo y gastoNac, las mismas que filtramos por p-valor.

c. ¿Cómo compara el impacto de las variables en el modelo? ¿Cuál es la variable 'más importante' en el modelo? ¿Y la 'menos importante'?

Podemos comparar el impacto de las variables por su significancia, observando el p-valor, por el peso de su coeficiente beta, y observando si el intervalo de confianza el coeficiente beta incluye al cero como se ve en el punto anterior.

- Variable menos importante:

```
which.max(summary(modelo_3)$coefficients[,4])
```

```
## factor(personas)1
##                 5
```

```
summary(modelo_3)$coefficients[which.max(summary(modelo_3)$coefficients[,4]),4]
```

```
## [1] 0.7372541
```

```
which.max(summary(modelo_3_2)$coefficients[,4])
```

```
## personas
##        5
```

```r
summary(modelo_3_2)$coefficients[which.max(summary(modelo_3_2)$coefficients[,4]),4]
```

```
## [1] 0.466138
```

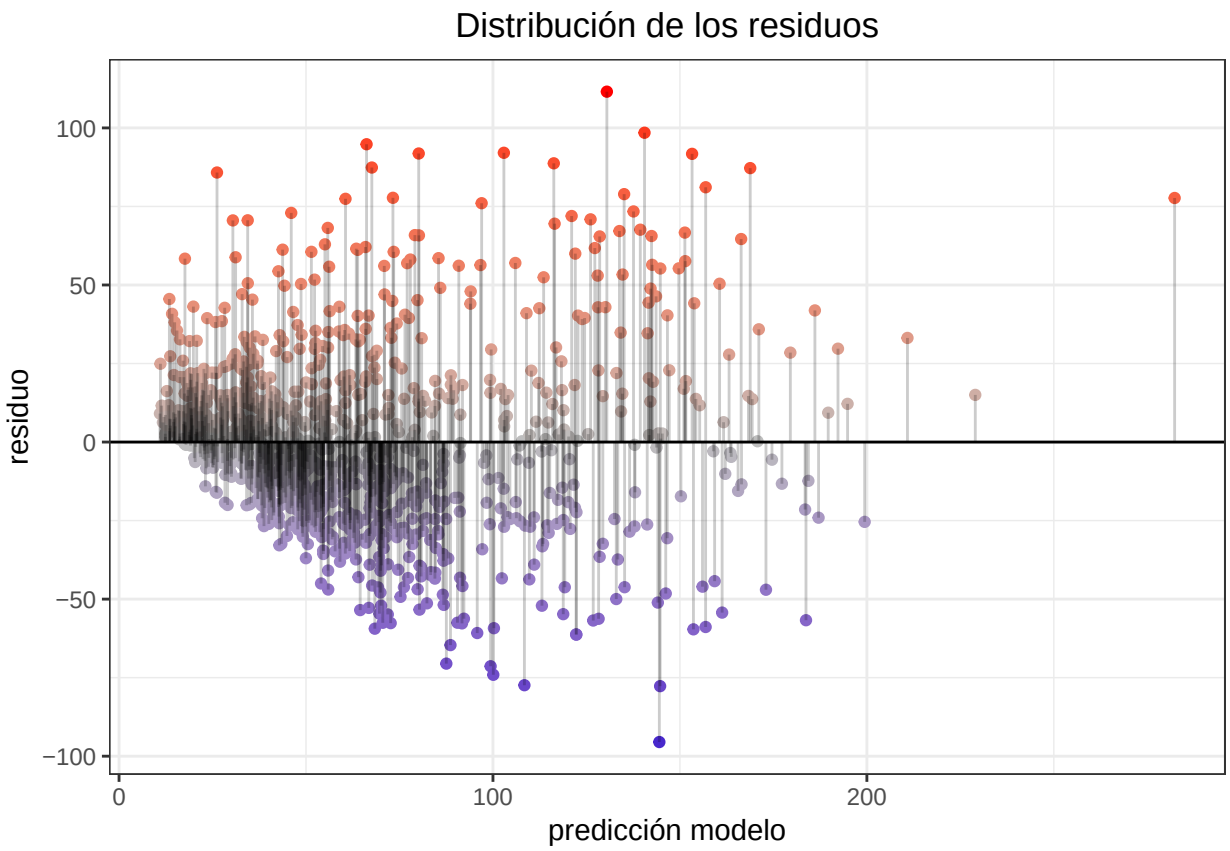
La variable con el p-valor más grande es $factor(personas)1$, o *personas* si tomamos el modelo sin factor.

d. Verifique los supuestos necesarios.

- Relación lineal entre variable dependiente e independiente:

Se calculan los residuos para cada observación y se grafican.

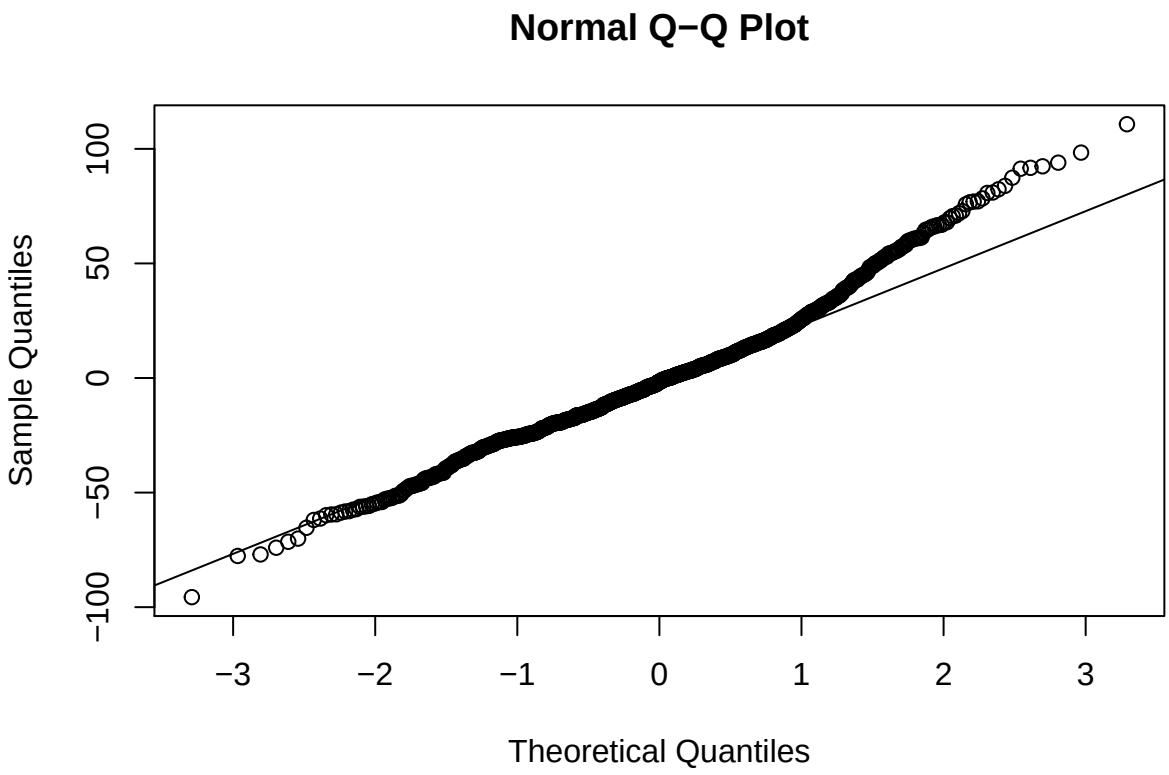
```r
Telefonia_2019$prediccion <- modelo_3_2$fitted.values
Telefonia_2019$residuos   <- modelo_3_2$residuals

ggplot(data = Telefonia_2019, aes(x = prediccion, y = residuos)) +
  geom_point(aes(color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
  geom_hline(yintercept = 0) +
  geom_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +
  labs(title = "Distribución de los residuos", x = "predicción modelo",
       y = "residuo") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
```



Los residuos no ven distribuidos aleatoriamente entorno al valor 0, por lo que no se acepta la linealidad.

- Distribución normal de los residuos:

```
## Grafica qqplot para analizar normalidad:
qqnorm(modelo_3$residuals)
qqline(modelo_3$residuals)
```




```
## Teste de Shapiro para normalidad
shapiro.test(modelo_3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_3$residuals
## W = 0.9808, p-value = 3.292e-10
```

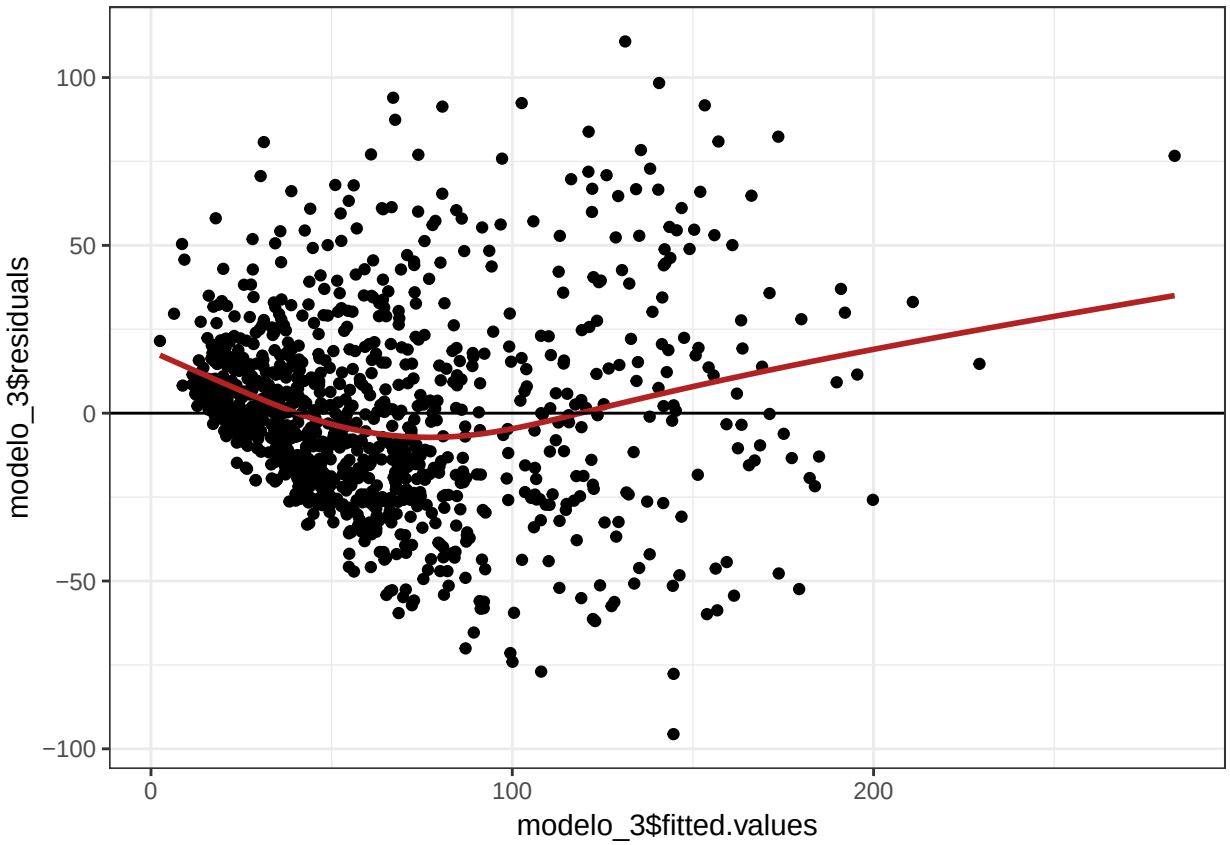
De la gráfica qqplot podemos observar que los puntos no están todos bien cercanos a la linea, esto indica que no se cumple el supuesto de normalidad. Verificamos lo que observamos en la gráfica con el test de Shapiro, concluimos que no se cumple el supuesto de normalidad.

- Supuesto de homocedasticidad (Variabilidad constante de los residuos):

Representamos los residuos frente a los valores ajustados por el modelo.

```
ggplot(data = Telefonia_2019, aes(modelo_3$fitted.values, modelo_3$residuals)) +
  geom_point() +
  geom_smooth(color = "firebrick", se = FALSE) +
  geom_hline(yintercept = 0) +
  theme_bw()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



No se ve una distribución aleatoria de los puntos, hay mayor concentración de puntos a la izquierda. Por esto suponemos que no hay homocedasticidad.

Realizamos el test de Breusch-Pagan para verificar lo observado en la gráfica anterior.

```
bptest(modelo_3)
```

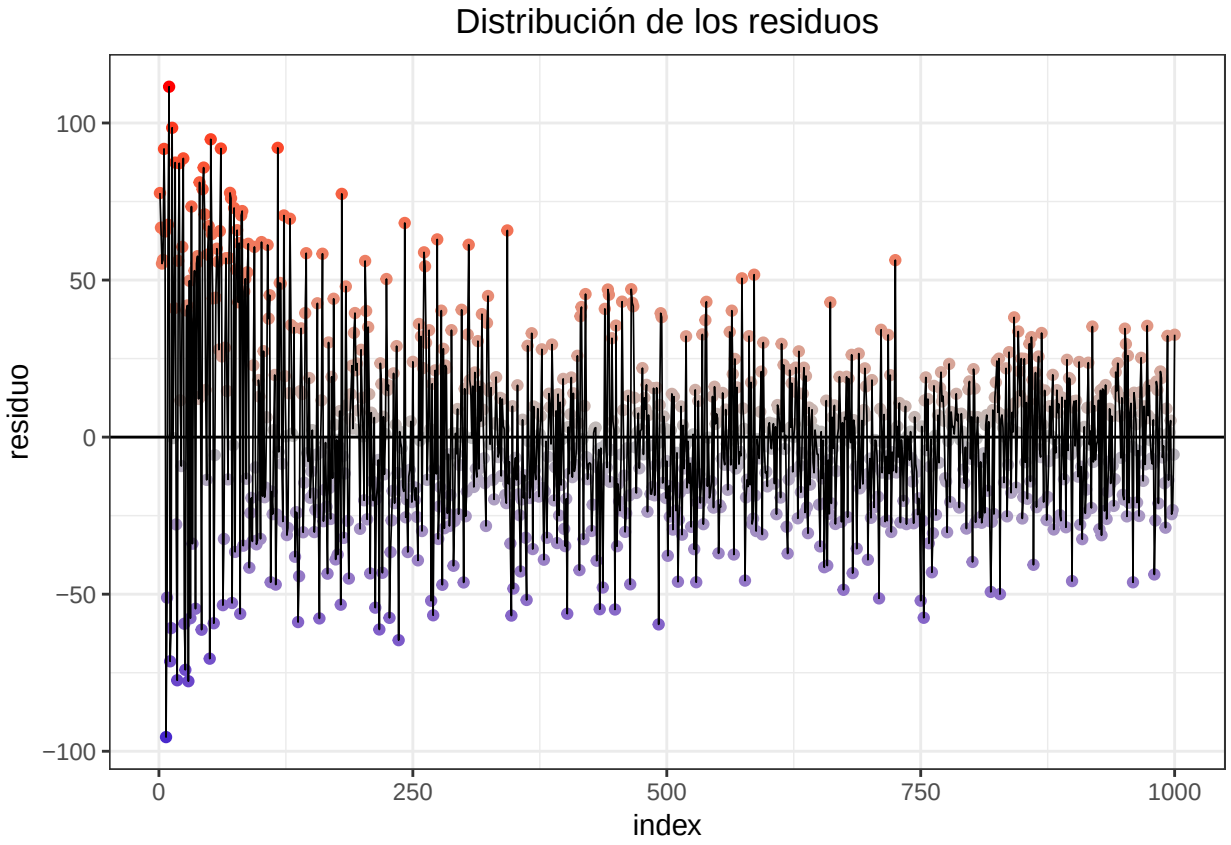
```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_3
## BP = 143.17, df = 12, p-value < 2.2e-16
```

Se rechaza H0, concluimos que no hay homocedasticidad.

- Autocorrelación de residuos:

Se observa si hay patrones en la distribución de los residuos.

```
ggplot(data = Telefonia_2019, aes(x = seq_along(residuos), y = residuos)) +
  geom_point(aes(color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
  geom_line(size = 0.3) +
  labs(title = "Distribución de los residuos", x = "index", y = "residuo") +
  geom_hline(yintercept = 0) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
```



Solo se observa una pequeña tendencia a ser mayores a la izquierda del gráfico.

Procedemos ahora a tratar de buscar indicios de multicolinealidad, para lo cual utilizaremos la métrica Variance Inflation Factor (VIf):

```
car::vif(modelo_3)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## retenc            1.538909  1        1.240528
## edad              2.021505  1        1.421797
## empleo            7.313866  1        2.704416
## factor(personas) 1.139468  7        1.009369
## gastoNac          1.280729  1        1.131693
## gastoIN           6.620211  1        2.572977
```

En este caso solo empleo tiene un valor mayor a 5 y puede significar que dicha variable puede expresarse como combinación lineal de las otras. De todas maneras, el inidicio fuerte estaría dado con un valor superior a 10. Podríamos proceder a quitar esta variable y reevaluar el modelo

Sabiendo que no se verifican los supuestos, concluimos que los modelos obtenidos no son significativos.

e. ¿Cómo interpreta el coeficiente de la variable "personas"? es significativa esta variable en el modelo?

La variable tiene un p-valor alto que india que no es siginificativa para el modelo. Tmabién observamos que el intervalo de confianza del coeficiente incluía al cero, esto nos muestra que la varible no influye de forma significativa en el modelo.

Por el coeficiente obtenido, 0.48333, podemos decir que en promedio por cada incremento de una unidad en la varible personas se corresponde en un incremento de 0.48333 del ingreso si todas las otras variables permanecen constantes.

4. Proponga ahora un modelo de regresión seleccionando las variables con backward y forward. ¿Conducen al mismo modelo ambos métodos?

- Selección con forward

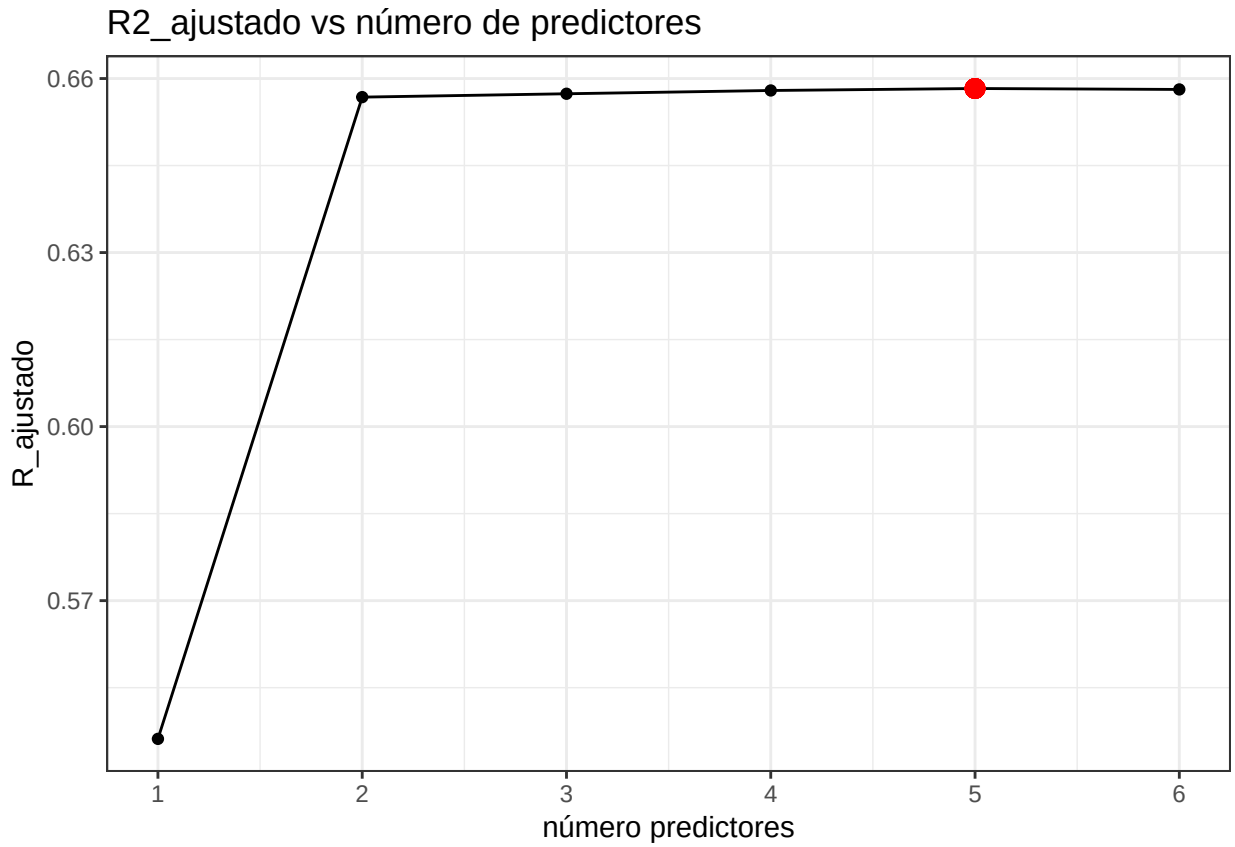
```
forwRM <- regsubsets( ingresos ~ retenc + edad + empleo + personas
                      + gastoNac + gastoIN, data = Telefonia_2019, method="forward", nvmax = 6)
summary(forwRM)
```

```
## Subset selection object
## Call: regsubsets.formula(ingresos ~ retenc + edad + empleo + personas +
##     gastoNac + gastoIN, data = Telefonia_2019, method = "forward",
##     nvmax = 6)
## 6 Variables  (and intercept)
##          Forced in Forced out
## retenc       FALSE      FALSE
## edad         FALSE      FALSE
## empleo       FALSE      FALSE
## personas     FALSE      FALSE
## gastoNac     FALSE      FALSE
## gastoIN      FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: forward
##          retenc edad empleo personas gastoNac gastoIN
## 1  ( 1 ) " "    " "  " "    " "      "*"      " "
## 2  ( 1 ) " "    " "  "*"    " "      "*"      " "
## 3  ( 1 ) " "    "*"  "*"    " "      "*"      " "
## 4  ( 1 ) " "    "*"  "*"    " "      "*"      "*"
## 5  ( 1 ) "*"    "*"  "*"    " "      "*"      "*"
## 6  ( 1 ) "*"    "*"  "*"    "*"      "*"      "*"
```

Hacemos un grafico comparativo de los modelos generados con forward.

```
p <- ggplot(data = data.frame(n_predictores = 1:6,
                              R_ajustado = summary(forwRM)$adjr2),
            aes(x = n_predictores, y = R_ajustado)) +
```

```
    geom_line() +
    geom_point()

#Se identifica en rojo el máximo
p <- p + geom_point(aes(
                    x = n_predictores[which.max(summary(forwRM)$adjr2)],
                    y = R_ajustado[which.max(summary(forwRM)$adjr2)]),
                    colour = "red", size = 3)
p <- p +  scale_x_continuous(breaks = c(0:6)) +
          theme_bw() +
          labs(title = 'R2_ajustado vs número de predictores',
               x =  'número predictores')
p
```



Seleccionando el modelo por R2, obtenemos un modelo con 5 variables. Se puede ver en la salida del summary anterior cuales son las primeras 5 (gastoNac, empleo, edad, gastoIN y retenc)

- Selección con backward

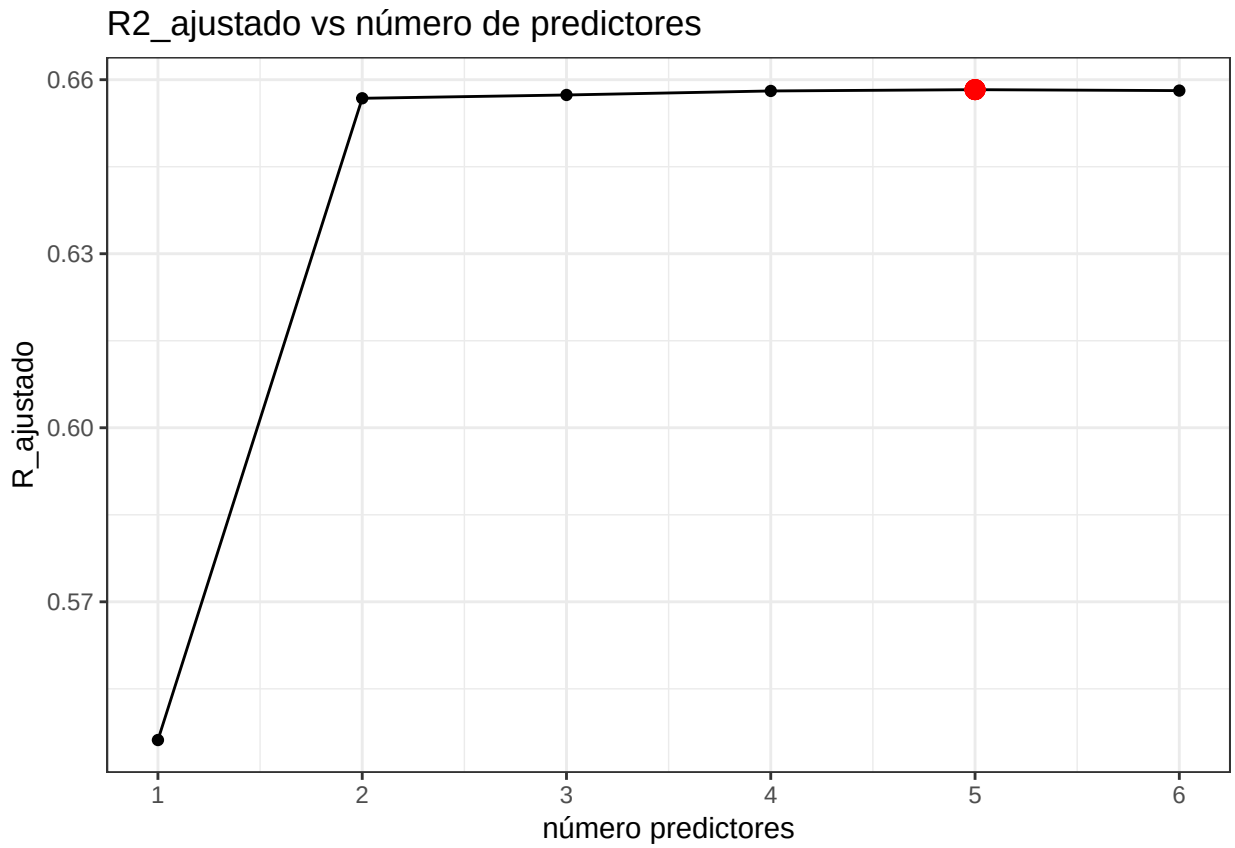
```
backRM <- regsubsets( ingresos ~ retenc + edad + empleo + personas
                      + gastoNac + gastoIN, data = Telefonia_2019, method="backward", nvmax = 6)
summary(backRM)
```

```
## Subset selection object
## Call: regsubsets.formula(ingresos ~ retenc + edad + empleo + personas +
```

```
##     gastoNac + gastoIN, data = Telefonia_2019, method = "backward",
##     nvmax = 6)
## 6 Variables  (and intercept)
##          Forced in Forced out
## retenc       FALSE      FALSE
## edad         FALSE      FALSE
## empleo       FALSE      FALSE
## personas     FALSE      FALSE
## gastoNac     FALSE      FALSE
## gastoIN      FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: backward
##          retenc edad empleo personas gastoNac gastoIN
## 1  ( 1 ) " "    " "  " "    " "      "*"      " "
## 2  ( 1 ) " "    " "  "*"    " "      "*"      " "
## 3  ( 1 ) " "    " "  "*"    " "      "*"      "*"
## 4  ( 1 ) "*"    " "  "*"    " "      "*"      "*"
## 5  ( 1 ) "*"    "*"  "*"    " "      "*"      "*"
## 6  ( 1 ) "*"    "*"  "*"    "*"      "*"      "*"
```

Hacemos un grafico comparativo de los modelos generados con backward.

```
p <- ggplot(data = data.frame(n_predictores = 1:6,
                              R_ajustado = summary(backRM)$adjr2),
            aes(x = n_predictores, y = R_ajustado)) +
    geom_line() +
    geom_point()

#Se identifica en rojo el máximo
p <- p + geom_point(aes(
                    x = n_predictores[which.max(summary(backRM)$adjr2)],
                    y = R_ajustado[which.max(summary(backRM)$adjr2)]),
                    colour = "red", size = 3)
p <- p +  scale_x_continuous(breaks = c(0:6)) +
          theme_bw() +
          labs(title = 'R2_ajustado vs número de predictores',
               x =  'número predictores')
p
```



Seleccionando el modelo por R2, obtenemos un modelo con 5 variables. Se puede ver en la salida del summary anterior cuales son las primeras 5 (gastoNac, empleo, gastoIN, retenc y edad).

Los modelos generados con backward y forward nos dieron igules pero en diferentes orden. Estos son los dos valores de R2 obtenidos.

```
# R2 en Forward
summary(forwRM)$adjr2[5]
```

```
## [1] 0.6582814
```

```
# R2 en Backward
summary(backRM)$adjr2[5]
```

```
## [1] 0.6582814
```

5. Ajuste ahora un modelo tomando en cuenta las variables regresoras elegidas en el ítem anterior, más la variable Internet agregada adecuadamente.

```
modelo_5 <- lm(formula = ingresos ~ gastoNac + empleo + gastoIN +  retenc + edad +
                 factor(internet), data = Telefonia_2019)
summary(modelo_5)
```

```
##
## Call:
## lm(formula = ingresos ~ gastoNac + empleo + gastoIN + retenc +
##     edad + factor(internet), data = Telefonia_2019)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.452 -18.226  -1.793  14.949 114.656
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.43024    3.69070   2.013   0.0444 *
## gastoNac           2.11308    0.07938  26.619   <2e-16 ***
## empleo             2.10058    0.24456   8.589   <2e-16 ***
## gastoIN           -0.38702    0.20547  -1.884   0.0599 .
## retenc             0.08687    0.05269   1.649   0.0995 .
## edad               0.12982    0.09967   1.302   0.1931
## factor(internet)1  4.62091    1.93399   2.389   0.0171 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.86 on 993 degrees of freedom
## Multiple R-squared:  0.6619, Adjusted R-squared:  0.6599
## F-statistic: 324.1 on 6 and 993 DF,  p-value: < 2.2e-16
```

Vemos en este caso que la mayoría de las variables son significativas y las que no lo son tienen p-valor no tan alejados del 0.05.

6. Finalmente, compare adecuadamente todos los modelos ajustados y elija el modelo ganador.

La comparación de modelos puede hacerce desde dos perspectivas, uno por el poder *explicativo* de los modelos y otras por el poder *predictivo*, en cada caso, un conjunto específico de métricas deben seleccionarse.

Poder explicativo

Comparación por R2:

```
glance(modelo_3) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC   p.value
##           <dbl> <dbl> <dbl> <dbl>     <dbl>
## 1         0.659  28.9 9581. 9649. 2.83e-223
```

```
glance(modelo_5) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC   p.value
##           <dbl> <dbl> <dbl> <dbl>     <dbl>
## 1         0.660  28.9 9572. 9611. 7.66e-230
```

Elegimos el modelo_5 como mejor modelo por tener un R2 mayor y un AIC menor.

Poder predictivo

```
MSE(Telefonia_2019$ingresos, predict(modelo_3,Telefonia_2019 ))
```

```
## [1] 824.4714
```

```
MSE(Telefonia_2019$ingresos, predict(modelo_5,Telefonia_2019 ))
```

```
## [1] 826.8559
```

En este caso la diferencia es muy pequeña en favor al modelo_3, de todas formas es conveniente evaluar poder predictivo en un dataset de validación que no ha sido utilizado en entrenamiento o con técnicas más avanzadas como cross validation.