Fontecha TP 7 Regresión

María Eugenia Fontecha

02 noviembre, 2020

Regresión lineal simple

Abra y examine las variables del archivo `ingresos.csv`. Corresponde a una base de datos de 40 individuos, para los que se registraron las variables:

- **Id: identificador del número de observación**
- **Educación: codificada de 1 a 10, donde 1 corresponde al menor nivel de educación alcanzado y 10 al mayor**
- **Edad: Edad en años**
- **Salario: salario bruto mensual en dólares.**

```
ingresos <- read.csv('ingresos.csv')
summary(ingresos)
```

```
##        Id          educacion          edad         salario
##  Min.   : 1.00   Min.   : 1.00   Min.   :20.0   Min.   :3039
##  1st Qu.:10.75   1st Qu.: 4.00   1st Qu.:27.5   1st Qu.:4057
##  Median :20.50   Median : 6.00   Median :35.0   Median :4619
##  Mean   :20.50   Mean   : 5.75   Mean   :35.0   Mean   :4683
##  3rd Qu.:30.25   3rd Qu.: 7.25   3rd Qu.:42.5   3rd Qu.:5344
##  Max.   :40.00   Max.   :10.00   Max.   :50.0   Max.   :6149
```

1. **Calcule el coeficiente de correlación entre la variable `educacion` y la variable `salario` para todos los datos e interprete el resultado. Realice el diagrama de dispersión. Describa el tipo de asociación que muestran las variables.**

Para calcular el coeficiente de correlación utilizo el método de Spearman para datos no paramétricos, ya que la variable educación es de tipo categórica: puede tomar como valores enteros del 1 al 10, según el nivel de educación. El test de correlación de Spearman utiliza los rangos de las variables para calcular el coeficiente de correlación y las hipótesis del test son:

H_0 : no hay una asociación monótona entre las variables

H_1 : hay una asociación monótona entre las variables.

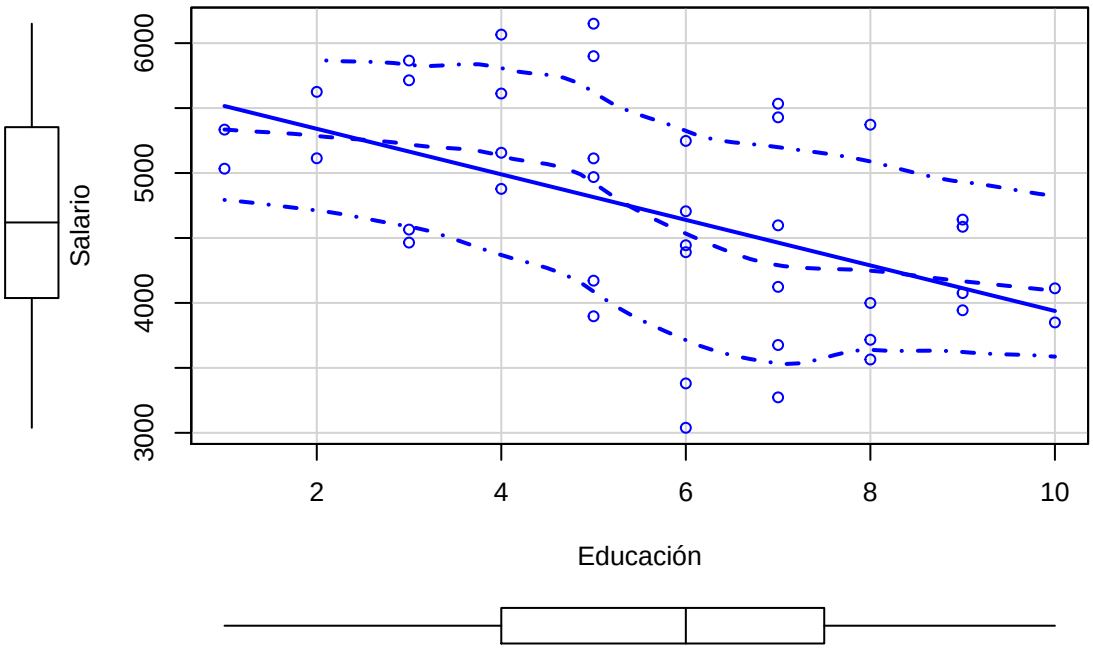
Una asociación monótona implica que cuando una variable se incrementa, la otra también. Una asocación monótona negativa implica que cuando una variable disminuye, la otra aumenta.

```r
cor.test(ingresos$educacion,ingresos$salario, method = 'spearman')
```

```
##
##  Spearman's rank correlation rho
##
## data:  ingresos$educacion and ingresos$salario
## S = 16394, p-value = 0.000344
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.5378934
```

El test arrojó un valor-p menor a 0.05, por lo que se rechaza la hipótesis nula y se concluye que hay una asociación monótona entre las variables `educacion` y `salario`. Además, el coeficiente de correlación dio -0.538, lo cual implica que la asociación es inversa. Entonces, a menor nivel de educación, mayor es el salario mensual.

```r
scatterplot(ingresos$educacion, ingresos$salario,xlab="Educación",ylab="Salario")
```



A partir del scatterplot se puede concluir lo mismo que con el test de correlación, ya que se ve que los datos tienen una tendencia negativa. El salario va disminuyendo a medida que aumenta el nivel de educación.

Estas conclusiones son las que se pueden sacar a partir de los análisis realizados, pero no parece ser lógico, por lo que faltaría analizar otros factores.

2. **Obtenga el coeficiente de correlación y su p-valor entre `educacion` y `salario` para cada edad. Interprete el resultado. Realice los diagramas de dispersión. Describa el tipo de asociación que muestran las variables.**

```r
print(unique(ingresos$edad))
```

```
## [1] 20 30 40 50
```

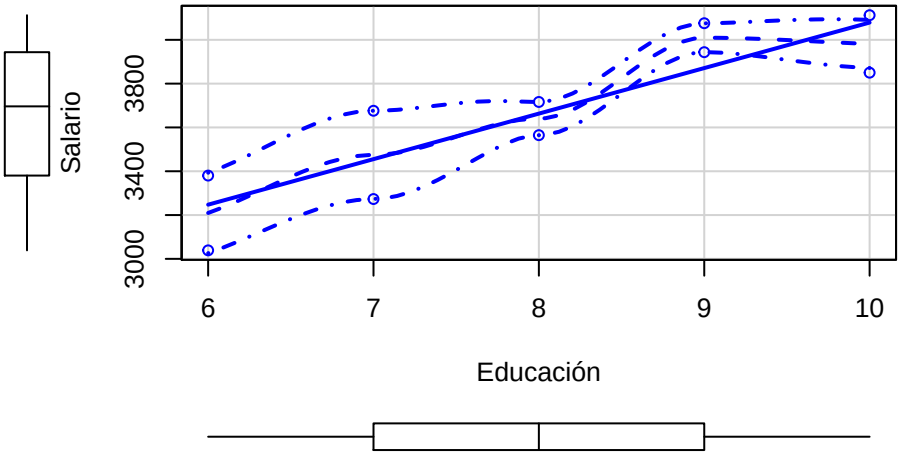
Vemos que hay cuatro grupos de edades, por lo que analizo la correlación entre `educacion` y `salario` para cada uno de ellos. Realizo el test de Spearman para cada grupo.

Edad: 20 años

```r
ingresos20 <- ingresos[ingresos$edad == 20,]
cor.test(ingresos20$educacion,ingresos20$salario, method = 'spearman')
```

```
##
##  Spearman's rank correlation rho
##
## data:  ingresos20$educacion and ingresos20$salario
## S = 18.767, p-value = 0.000637
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8862587
```

```r
scatterplot(ingresos20$educacion, ingresos20$salario,xlab="Educación",ylab="Salario")
```



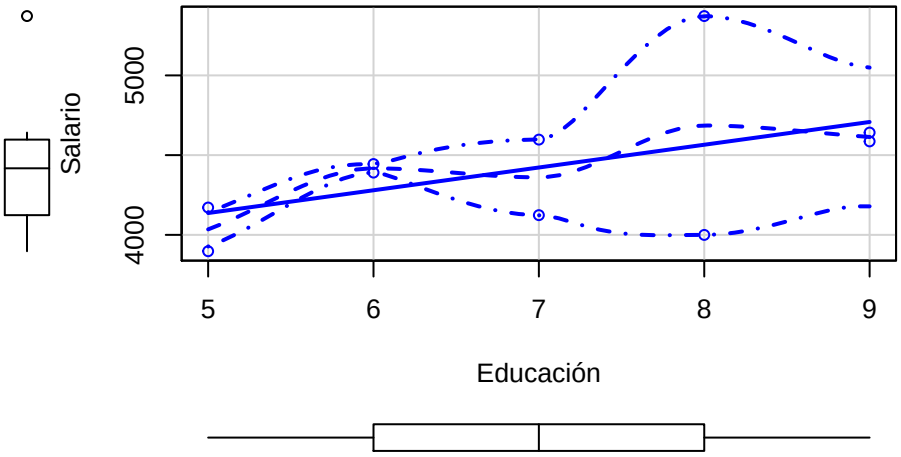
Los resultados del test de correlación indican que hay una asociación monótona entre el salario y el nivel de educación para el grupo de individuos de 20 años, ya que el p-valor es menor a 0.05 y se rechaza la hipótesis nula. Además, tanto con el coeficiente de correlación que dio un valor positivo de 0.886, como con la línea de tendencia del scatterplot que tiene pendiente positiva, vemos que esta relación es monótona positiva. Entonces, a mayor nivel educativo alcanzado, mayor salario mensual.

Edad: 30 años

```
ingresos30 <- ingresos[ingresos$edad == 30,]
cor.test(ingresos30$educacion,ingresos30$salario, method = 'spearman')
```

```
##
##  Spearman's rank correlation rho
##
## data:  ingresos30$educacion and ingresos30$salario
## S = 71.574, p-value = 0.08794
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5662209
```

```
scatterplot(ingresos30$educacion, ingresos30$salario,xlab="Educación",ylab="Salario")
```



Para el grupo de individuos de 30 años no se observa una asociación significativa entre el salario y el nivel de educación alcanzado, ya que el p-valor del test de correlación es mayor a 0.05. El coeficiente de correlación dio de 0.566, lo cual habla de una fuerza de asociación débil. Esto se puede ver en el scatterplot, donde, si bien hay una línea de tendencia con pendiente positiva, esta no es muy pronunciada.

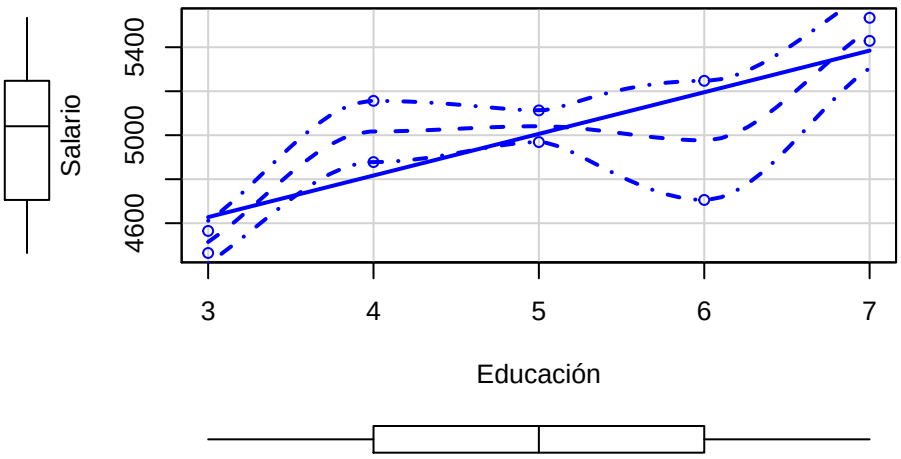
Edad: 40 años

```
ingresos40 <- ingresos[ingresos$edad == 40,]
cor.test(ingresos40$educacion,ingresos40$salario, method = 'spearman')
```

```
##
##  Spearman's rank correlation rho
##
## data:  ingresos40$educacion and ingresos40$salario
```

```
## S = 35.015, p-value = 0.006807
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.7877855
```

```
scatterplot(ingresos40$educacion, ingresos40$salario,xlab="Educación",ylab="Salario")
```



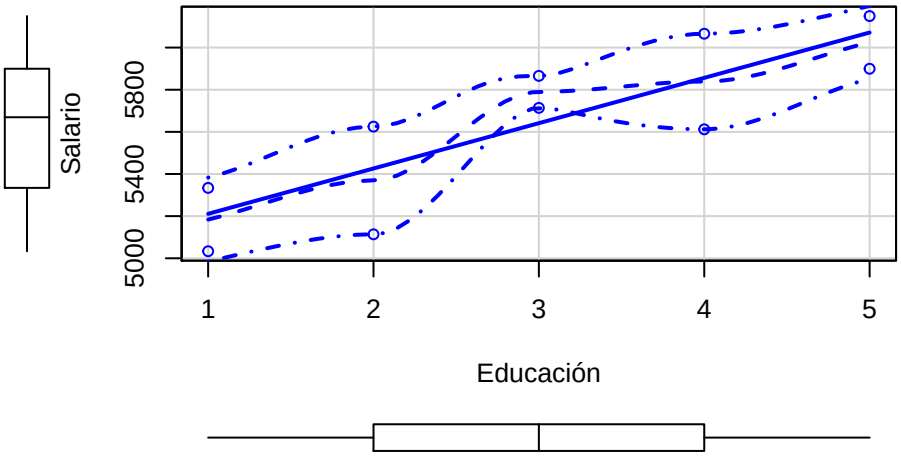
Al igual que en el grupo de 20 años, estos resultados muestran que existe una relación monótona positiva entre el salario y el nivel de educación en el grupo de individuos de 40 años, aunque esta es menor ya que el coeficiente de correlación dio menor. También se puede ver que la pendiente de la recta de tendencia del scatterplot es un poco menos pronunciada que en el caso de indiviudos de 20 años.

Edad: 50 años

```
ingresos50 <- ingresos[ingresos$edad == 50,]
cor.test(ingresos50$educacion,ingresos50$salario, method = 'spearman')
```

```
##
##  Spearman's rank correlation rho
##
## data:  ingresos50$educacion and ingresos50$salario
## S = 26.891, p-value = 0.002523
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8370221
```

```
scatterplot(ingresos50$educacion, ingresos50$salario,xlab="Educación",ylab="Salario")
```

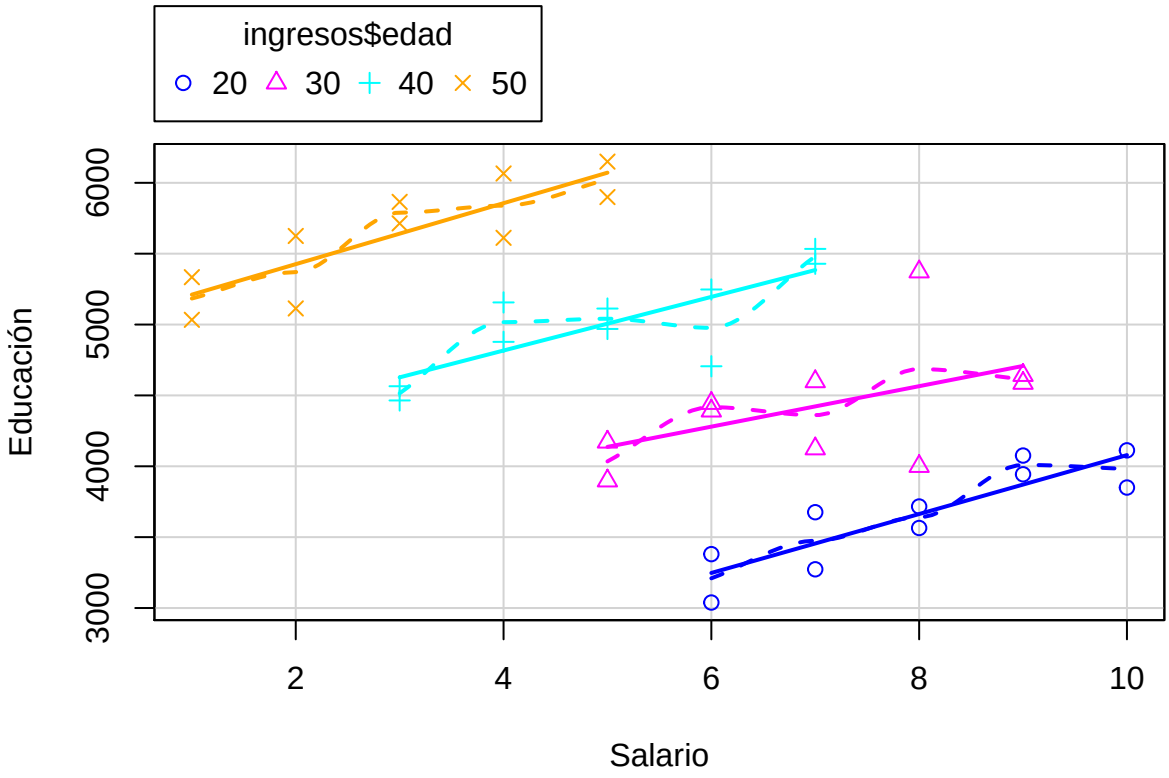


Para el grupo de 50 años, se puede concluir que hay una asociación monótona positiva entre el salario y el nivel de educación alcanzado, ya que el p-valor dio menor a 0.05 y el coeficiente de correlación dio mayor a 0.8. El scatterplot también muestra esta tendencia monótona positiva de los datos. Por lo tanto, en individuos de 50 años de edad se cumple que a mayor nivel educativo, mayor es el salario mensual.

Los resultados obtenidos al dividir por rango etario son distintos de los resultados que se obtuvieron cuando se hizo el análisis global de la relación entre educación y salario. Esto se debe a que los individuos de mayor edad tienen niveles educativos más bajos que los de menor edad (esto se puede ver en los ejes x de los scatterplots), y que a su vez los de mayor edad tienen mayores salarios. Por esta razón, en el primer análisis parecía haber una asociación inversa significativa entre el salario y el nivel educativo, pero al separar por edad, se observó que en realidad la asociacón es positiva, lo cual tiene mayor lógica. Por ende, la edad estaba confundiendo la verdadera asociación entre el salario y el nivel educativo alcanzado.

A continuación se muestra un scatterplot de los datos agrupados por edad, donde podemos ver que a mayor edad, menor nivel educativo y mayor salario mensual, y que a menor edad, el nivel educativo alcanzado es mayor pero los salarios son menores.

```
scatterplot(ingresos$salario~ingresos$educacion | ingresos$edad,
xlab="Salario",ylab="Educación")
```



3. **Ajuste una recta de cuadrados mínimos para la variable respuesta `salario` y la variable explicativa `educación` sin tener en cuenta la variable edad. Describa e interprete cada uno de los resultados. ¿Qué significa el coeficiente de la variable explicativa `educación`?**

Genero el modelo lineal.

```
rg <- lm(salario ~ educacion,data=ingresos)
```

Verifico los supuestos del modelo lineal.

- Distribución normal de los residuos.

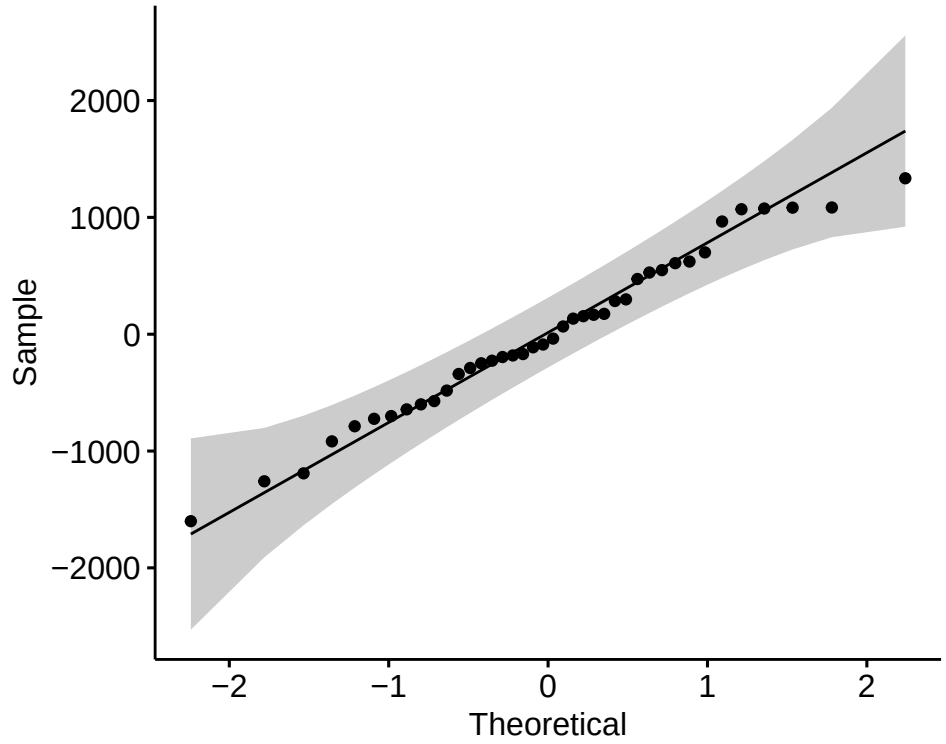
```
print(shapiro.test(rg$residuals))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rg$residuals
## W = 0.9828, p-value = 0.7916
```

```
print(lillie.test(rg$residuals))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rg$residuals
## D = 0.063767, p-value = 0.9499
```

```
ggqqplot(rg$residuals)
```



Ambos test estadísticos arrojaron un p-valor mayor a 0.05, por lo que no se rechaza la hipótesis nula que dice que los residuos tienen distribución normal. Además, vemos en el QQPlot que todos los residuos siguen la línea teórica de normalidad. Por lo tanto, se cumple el supuesto de normalidad de los residuos.

- Los residuos tienen media 0.

```
print(mean(rg$residuals))
```

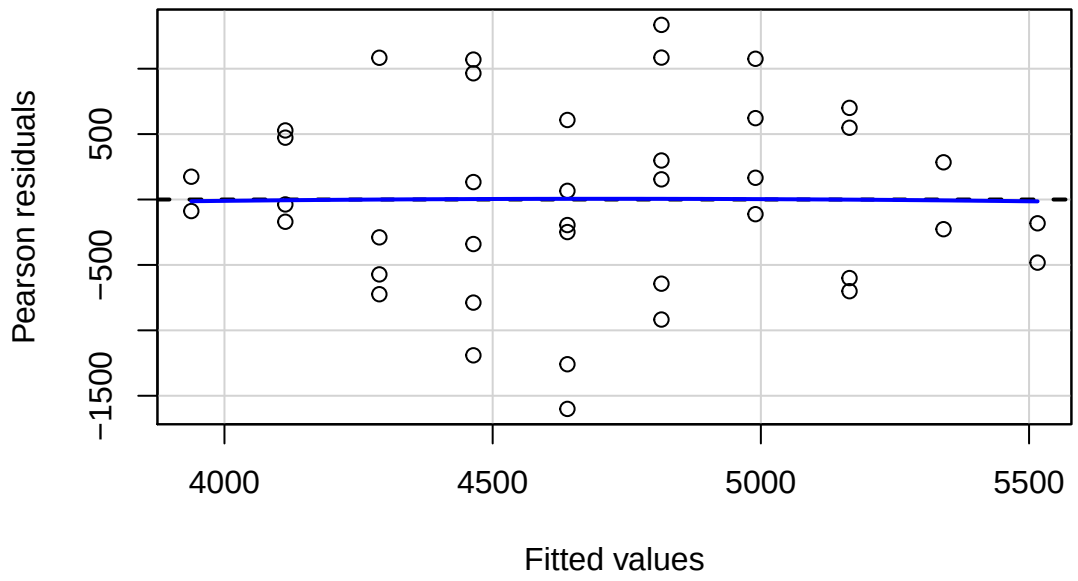
```
## [1] -3.266831e-15
```

Podemos ver que la media no es exactamente 0, pero igual es un valor muy pequeño, del orden de 10^{-15}, por lo que se puede considerar que se cumple el supuesto.

- Los residuos cumplen con el supuesto de homocedasticidad.

Realizo el gráfico de los residuos vs los valores ajustados para ver si se cumple la homocedasticidad

```r
residualPlot(rg)
```



En este gráfico podemos ver que la varianza no es constante, sino que tiene cierto patrón. En los extremos izquierdo y derecho se ve que los desvíos son menores, mientras que en el centro son máximos, formando una especie de rombo. Por lo tanto, no se cumple la homocedasticidad. Entonces, no se puede utilizar este modelo, por no cumplirse el supuesto de varianza constante.

Análisis del modelo lineal generado.

```r
summary(rg)
```

```
##
## Call:
## lm(formula = salario ~ educacion, data = ingresos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1600.58  -504.95   -62.93   533.20  1334.82
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5691.2      295.7  19.249  < 2e-16 ***
## educacion     -175.3       47.5  -3.691 0.000698 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 716.4 on 38 degrees of freedom
## Multiple R-squared:  0.2639, Adjusted R-squared:  0.2445
## F-statistic: 13.62 on 1 and 38 DF,  p-value: 0.0006984
```

Los coeficientes Intercept y educacion son β_0 y β_1, respectivamente. Por lo tanto, la ecuación del modelo quedaría:

- $y = 5691.2 - 175.3x$,

siendo x el nivel de educación e y el salario mensual.

β_0 es el valor que tomaría y si x valiese 0, mientras que β_1 tiene que ver con la relación que hay entre x e y, y con cómo afecta en y la variación de x. Ambos valores se estimaron con un p-valor menor a 0.05, por lo que son estadísticamente significativos. En este test la hipótesis nula es que estos coeficientes son 0.

Además, el modelo tiene su propio valor de significancia, 0.0006984, que en este caso es el mismo que el valor-p de educación ya que es la única variable explicativa. De haber más, sería diferente, ya que la hipótesis nula es que todos los coeficientes que acompañan a las variables independientes son iguales a cero.

Por último, el valor R^2 nos habla de qué tan bien se ajusta este modelo a nuestros datos. En este caso, un valor R^2 de 0.2445 no habla de un buen ajuste. Esto puede tener con la alta variabilidad que tenían los datos que se podía observar en el scatterplot de todos los puntos.

De todas formas, considero que no se cumplieron los supuestos del modelo lineal ya que el gráfico de los residuos vs los valores ajustados no mostraba una distrubución homogénea, sino que percibí cierto patrón.

Resgresión lineal múltiple

Ejercicio 1

Para los datos de niños de bajo peso, se encontró una relación lineal significativa entre la presión sistólica y la edad gestacional. Los datos están el archivo `lowbwt.tsv`.

- **Sbp: Mediciones de presión sistólica**
- **Gestage: correspondientes edades gestacionales**
- **apgar5: score Apgar a los 5 minutos para cada niño recién nacido. (El score Apgar es un indicador del estado general de salud del niño a los 5 minutos de haber nacido; en realidad es una medida ordinal que se suele tomar como si fuera continua).**

```
lowbwt <- read.table('lowbwt.tsv', header = TRUE)
summary(lowbwt)
```

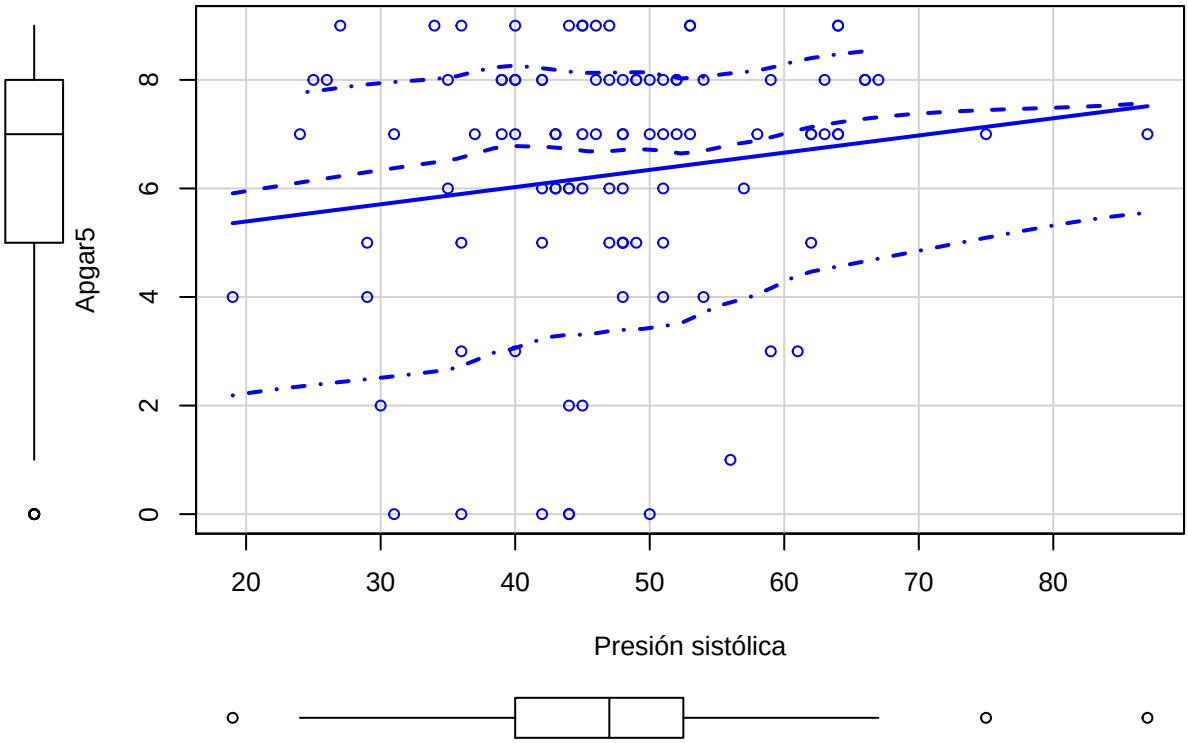
```
##       sbp             sex              tox             grmhem        gestage
##  Min.   :19.00   Min.   :0.00   Min.   :0.00   Min.   :0.00   Min.   :23.00
##  1st Qu.:40.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:27.00
##  Median :47.00   Median :0.00   Median :0.00   Median :0.00   Median :29.00
##  Mean   :47.08   Mean   :0.44   Mean   :0.21   Mean   :0.15   Mean   :28.89
##  3rd Qu.:52.25   3rd Qu.:1.00   3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:31.00
##  Max.   :87.00   Max.   :1.00   Max.   :1.00   Max.   :1.00   Max.   :35.00
##      apgar5
##  Min.   :0.00
##  1st Qu.:5.00
##  Median :7.00
##  Mean   :6.25
##  3rd Qu.:8.00
##  Max.   :9.00
```

```
lowbwt$sex <- as.factor(lowbwt$sex)
str(lowbwt$sex)
```

```
##  Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 1 1 1 ...
```

1. **Realice un diagrama de dispersión de la presión sistólica versus el score Apgar. ¿Parece haber una relación lineal entre estas dos variables?**

```
scatterplot(lowbwt$sbp, lowbwt$apgar5,xlab="Presión sistólica",ylab="Apgar5")
```



A simple vista no pareciera haber una relación entre la presión sistólica y el apgar5. Los datos parecen estar bastante dispersos, sin presentar una clara tendencia. Según la recta de tendencia pareciera haber una pequeña asociación monótona positiva.

2. **Usando la presión sistólica como respuesta y la edad gestacional y el score Apgar como explicativas, ajuste el modelo lineal $E(XY) = \beta_0 + \beta_1 \cdot$ gestage $+ \beta_2 \cdot$ apgar5. Ajuste el modelo e interprete los coeficientes estimados.**

Creo el modelo lineal:

```
ml1 <- lm(sbp ~ gestage + apgar5,data=lowbwt)
```

Verifico los supuestos del modelo lineal.

- Media de los residuos igual a cero.

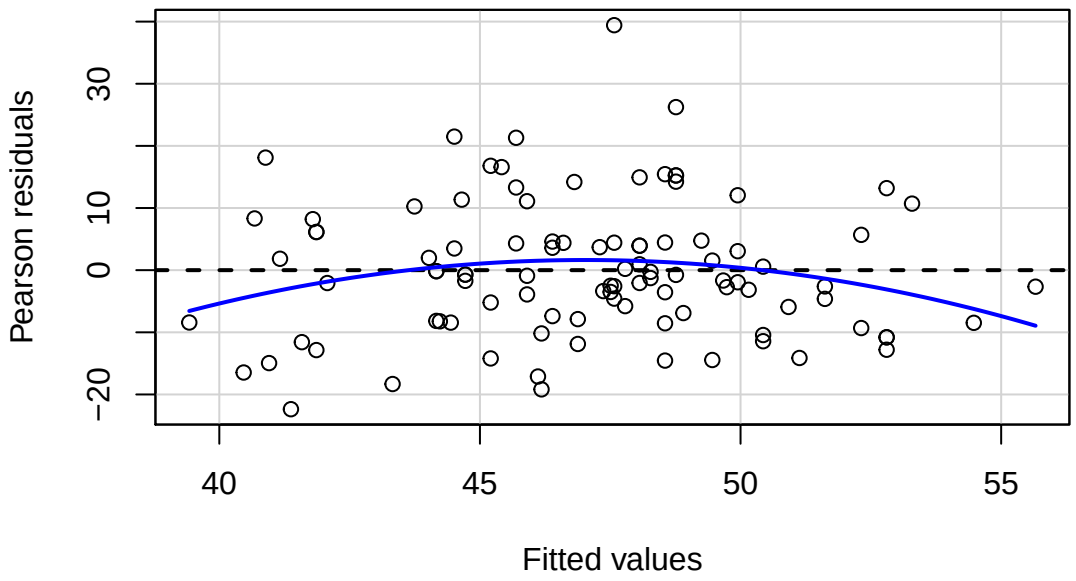
```
mean(ml1$residuals)
```

```
## [1] 2.220446e-16
```

Podemos ver que la media de los residuos es un valor muy chico, cercano a 0, por lo que se comprueba el supuesto de que la media de los residuos sea 0.

- Homocedasticidad de los residuos

```
residualPlot(ml1)
```



El gráfico de los residuos vs los valores ajustados nos muestra una distribución que no sigue un patrón, sino que a lo largo del eje x se ve cierta homogeneidad. Por lo tanto, concluyo que el supuesto de homocedasticidad se cumple.

- Distribución normal de los residuos

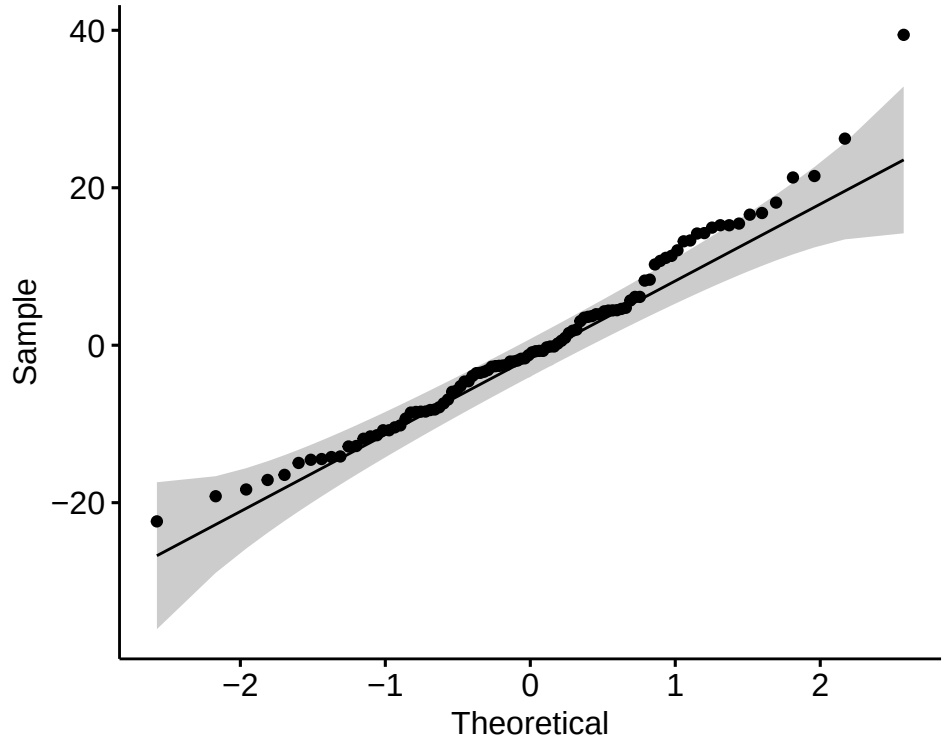
```
print(shapiro.test(ml1$residuals))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ml1$residuals
## W = 0.97422, p-value = 0.04688
```

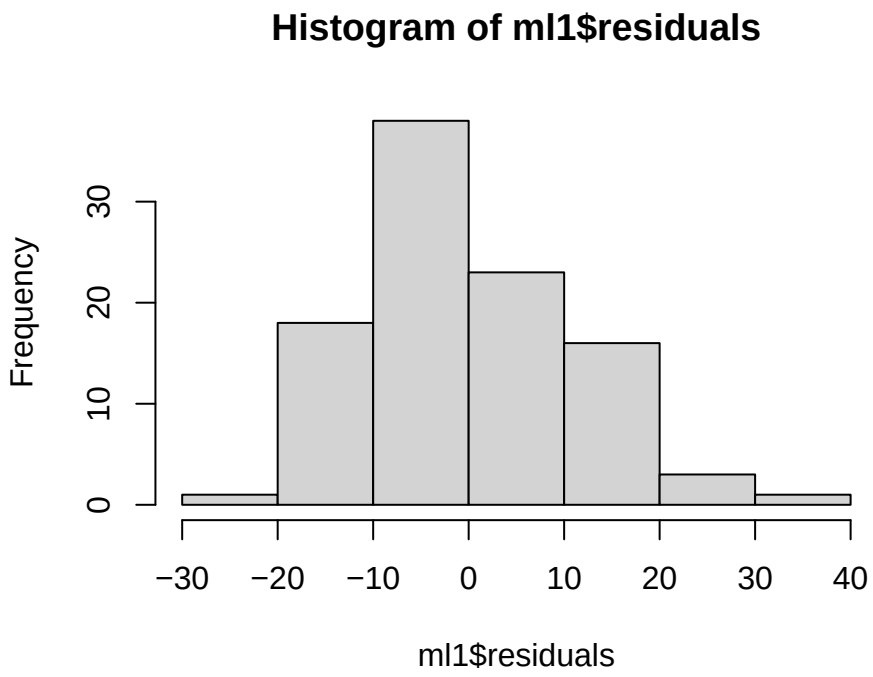
```r
print(lillie.test(ml1$residuals))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ml1$residuals
## D = 0.081171, p-value = 0.1081
```

```r
ggqqplot(ml1$residuals)
```




```r
hist(ml1$residuals)
```



El histograma muestra una distribución que parece ser normal, ya que tiene forma acampanada. Por otro lado, en el QQ-Plot vemos que la mayoría de los datos siguen bastante bien a la recta teórica, aunque hay algunos pocos que se apartan más que los demás. En cuanto a los tests estadísticos, con el de Shapiro-Wilk se conluiría que los residuos no tienen distribución normal ya que el p-valor dio menor a 0.05, pero con el test de Lilliefors se obtiene un p-valor mayor a 0.05, por lo que no se rechazaría la hipótesis nula.

Por lo tanto, a partir de lo visto en los métodos gráficos y dado que el test de Lilliefors es más robusto que el de Shapiro para muestras grandes, concluyo que el supuesto de normalidad de los residuos se cumple.

Habiendo corroborado los supuestos del modelo lineal, analizo los resultados:

```
summary(ml1)
```

```
##
## Call:
## lm(formula = sbp ~ gestage + apgar5, data = lowbwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.374  -8.180  -1.088   4.985  39.424
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.8034    12.6629   0.774   0.4407
## gestage       1.1848     0.4424   2.678   0.0087 **
## apgar5        0.4875     0.4613   1.057   0.2932
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.99 on 97 degrees of freedom
```

```
## Multiple R-squared:  0.08944,    Adjusted R-squared:  0.07066
## F-statistic: 4.764 on 2 and 97 DF,  p-value: 0.01063
```

A partir de los coeficientes estimados y los p-valores de esta estimaciones, se podría concluir que la variable `apgar5` no es estadísticamente significativa para el modelo, ya que con un p-valor mayor a 0.05 no se rechaza la hipótesis nula que dice que la correlación entre esta variable y la presión sistólica es cero. Por ende, podría eliminarse esta variable del modelo. Por otro lado, el p-valor para el coeficiente β_1 es menor a 0.05, lo que indica que existe una correlación estadísticamente significativa entre la variable `gestage` y la variable respuesta. Esta correlación es positiva, ya que el coeficiente dio 1.18, por lo que cuando una aumenta, la otra también. Por último, el coeficiente β_0 dio no significativo, ya que el p-valor es mayor a 0.05, lo que implica que cuando ambas variables valen 0, la variable respuesta no difiere significativamente de 0, aunque este caso en particular no tendría sentido en la práctica.

Por último, el R^2 dio muy bajo, cercano a cero, por lo que el modelo no se está ajustando a los datos. Por lo tanto, este modelo no sería adecuado para predecir la presión sistólica.

3. **¿Cuál es la presión media estimada para la población de niños de bajo peso cuya edad gestacional es 31 semanas y cuyo score Apgar es 7?**

Utilizo el modelo generado para predecir la presión media:

```
predict(ml1, data.frame(gestage = 31, apgar5 = 7))
```

```
##        1
## 49.94562
```

La presión media estimada es 49.94562 mmHg.

4. **Considere un modelo que contiene `gestage` y la variable `sex`. Ajuste el modelo. Comente la significatividad de los parámetros. Dados dos niños con igual edad gestacional, uno varón y otro nena, cual tendrá presión sistólica más alta? ¿Por qué?**

```
ml2 = lm(sbp ~ gestage + sex,data = lowbwt)
```

Verifico los supuestos del modelo lineal.

- Media de los residuos igual a cero.

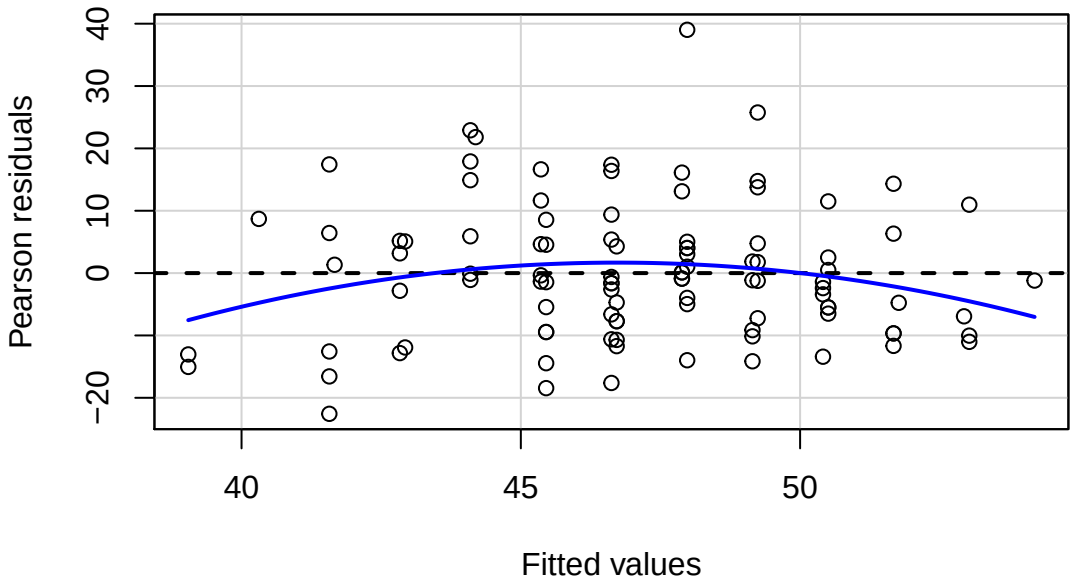
```
mean(ml2$residuals)
```

```
## [1] 1.770806e-16
```

La media de los residuos es muy cercana a 0, por lo que se cumple este supuesto.

- Homocedasticidad de los residuos

```
residualPlot(ml2)
```



La varianza parece ser constante, ya que no se observa ningún patrón en el gráfico que haga que esta varíe a lo largo del eje x, sino que se ve una disribución homogénea. Por lo tanto, se cumple el supuesto de homocedasticidad.

- Distribución normal de los residuos

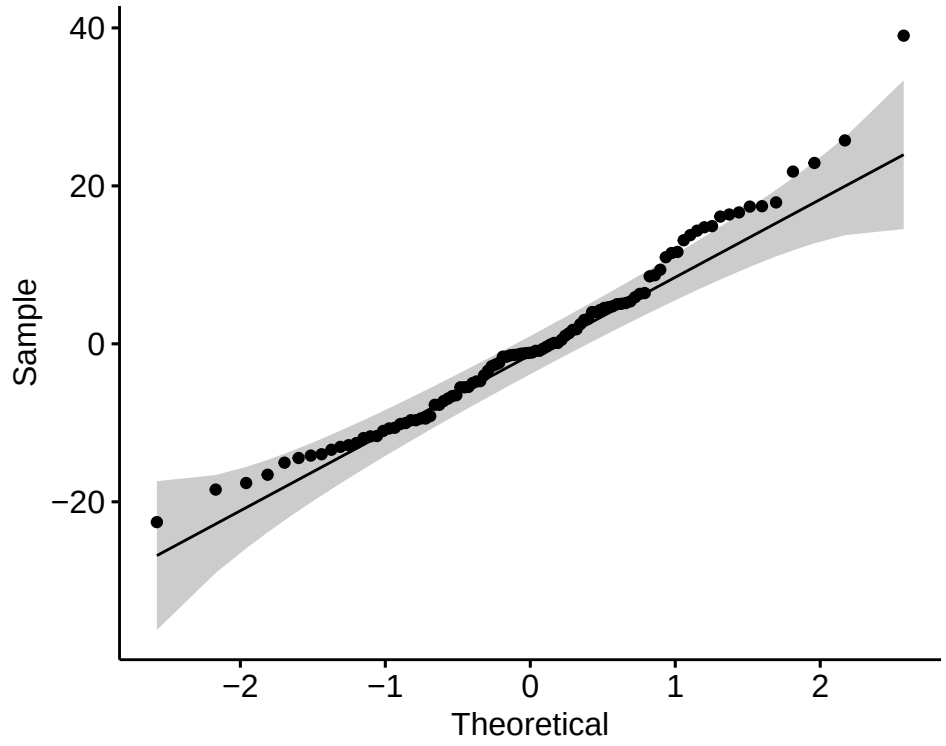
```
print(shapiro.test(ml2$residuals))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ml2$residuals
## W = 0.9712, p-value = 0.02732
```

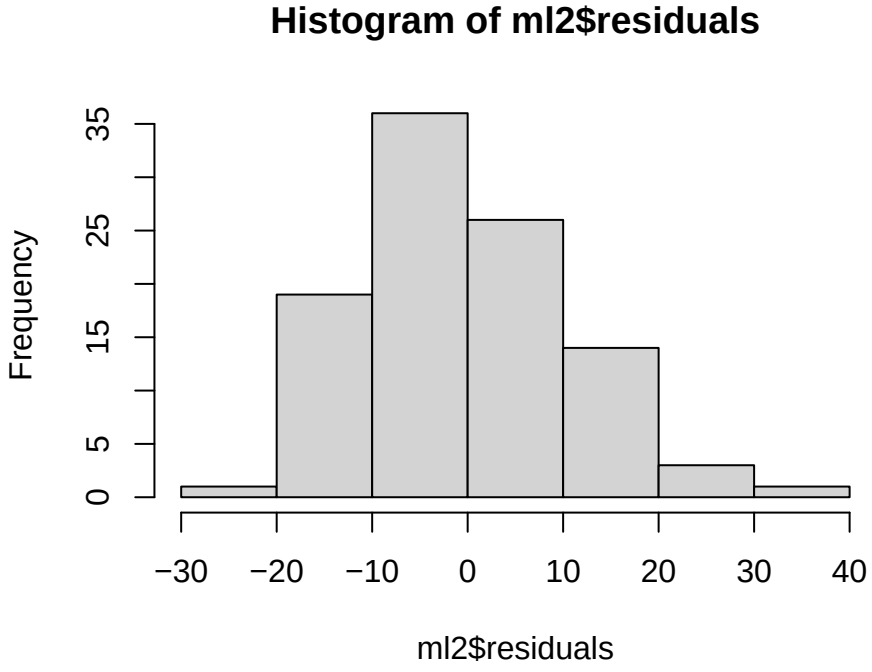
```
print(lillie.test(ml2$residuals))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ml2$residuals
## D = 0.075789, p-value = 0.1695
```

```
ggqqplot(ml2$residuals)
```




```r
hist(ml2$residuals)
```



Los tests de normalidad muestran resultados diferentes, por un lado el p-valor del test de Shapiro es menor a 0.05, lo cual indica no normalidad de lo residuos, y, por otro, el test de Lilliefors dio un p-valor mayor a

0.05, lo cual indica normalidad. En los métodos gráficos vemos que el histograma tiene una forma similar a la de la distribución normal y que el QQ-Plot muestra que la mayoría de los datos se encuentra dentro de los parámetros esperados para una distribución normal, aunque hay algunos que no.

Dado el tamaño de la muestra y que la mayoría de los métodos tiende a indicar normalidad de los residuos, concluyo que el supuesto de normalidad se cumple.

```
summary(ml2)
```

```
##
## Call:
## lm(formula = sbp ~ gestage + sex, data = lowbwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.572  -8.074  -1.122   5.219  39.022
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.0071    12.7227   0.787  0.43346
## gestage       1.2626     0.4376   2.885  0.00482 **
## sex1          1.3563     2.2231   0.610  0.54322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.03 on 97 degrees of freedom
## Multiple R-squared:  0.08248,    Adjusted R-squared:  0.06356
## F-statistic:  4.36 on 2 and 97 DF,  p-value: 0.01538
```

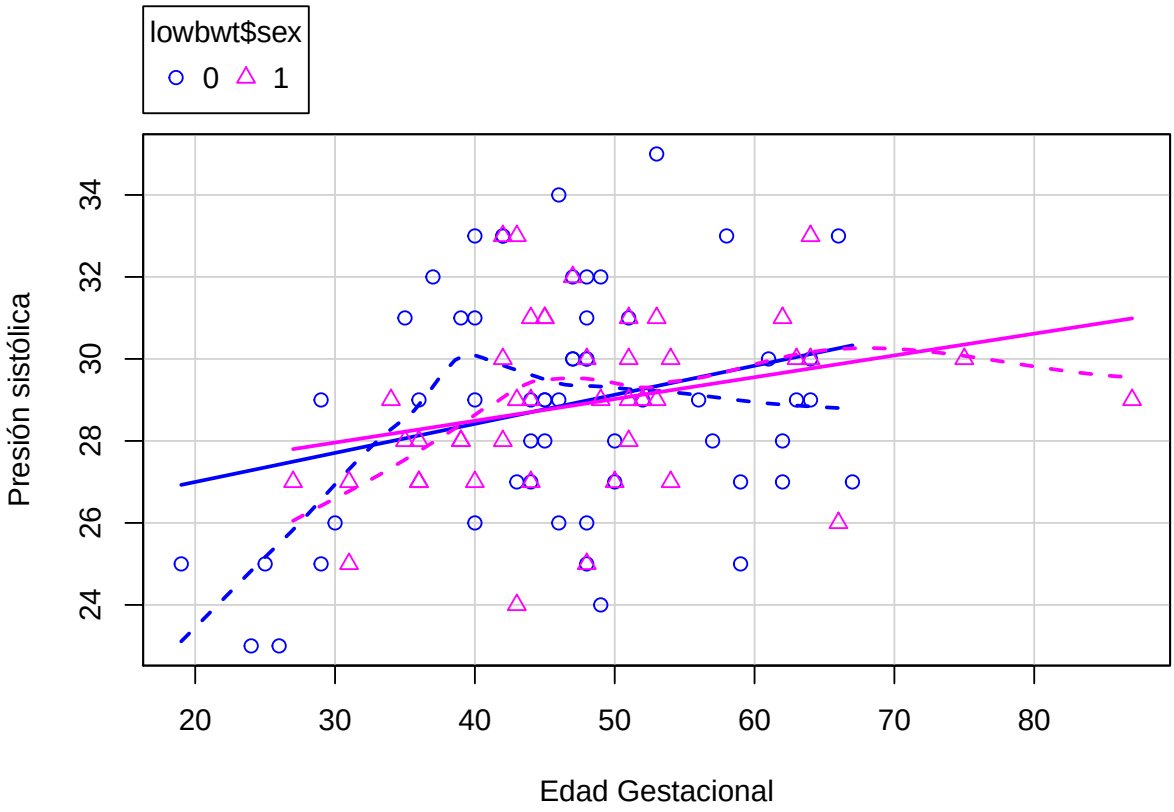
El coeficiente de la variable `gestage` es estadísticamente significativo, ya que su p-valor es menor a 0.05. En cambio, el coeficiente de la variable `sex` no da significativo, ya que el p-valor es mayor a 0.05. Esto indica que no hay una correlación significativa entre el sexo del bebé y la presión sistólica de los bebés con bajo peso al nacer.

Viendo el coeficiente estimado para la variable sex, al utilizar el modelo para predecir la presión sistólica, esta será mayor en el caso en que sex sea 1 ya que el coeficiente es positivo, lo cual produciría un aumento en la presión sistólica. En cambio, en el caso de que sex sea 0, no habría ningún efecto en el valor de la presión. Por lo tanto, suponiendo que sex $= 1$ corresponde a un bebé varón, dados dos niños con igual edad gestacional, uno varón y otro nena, el varón tendrá mayor presión sistólica.

Cabe destacar que el modelo no se ajusta a los datos, ya que R^2 dio muy cercano a 0.

5. **Haga un diagrama de dispersión de presión sistólica versus edad gestacional separando varones de nenas. Superponga las rectas ajustadas**

```
scatterplot(lowbwt$gestage~lowbwt$sbp|lowbwt$sex
,xlab="Edad Gestacional",ylab="Presión sistólica")
```



En este gráfico se ve que no hay una agrupación entre los datos de cada sexo que permita ver una asociación entre el sexo y la presión sistólica, sino que estos están dispersos y mezclados entre sí. Esto muestra por qué la variable explicativa sexo no es significativa.

6. **Agregue la interacción sexo – edad gestacional. Ajuste el modelo. ¿Incluiría al sexo como variable explicativa al modelo que tiene a la edad gestacional? ¿Por qué? ¿Incluiría a la interacción como variable explicativa del modelo? ¿Por qué?**

```
ml3 <- lm(sbp ~ gestage*sex, data = lowbwt)
```

Compruebo los supuestos del modelo de regresión lineal

- Media de los residuos igual a cero.

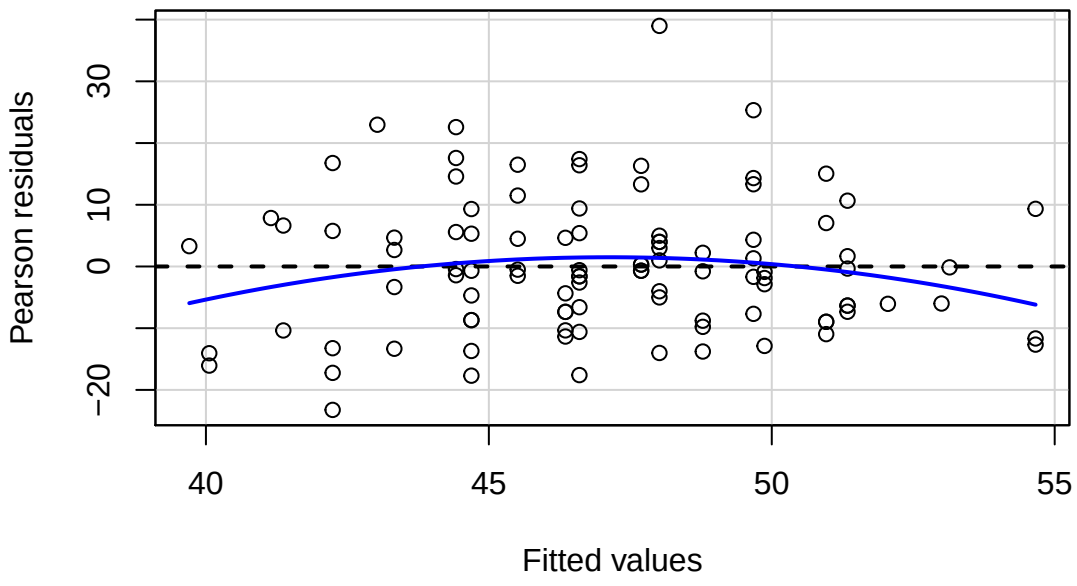
```
mean(ml3$residuals)
```

```
## [1] -2.433817e-16
```

El supuesto de que la media de los residuos sea 0 se cumple, ya que dio un valor muy cercano a 0.

- Homocedasticidad de los residuos

```
residualPlot(ml3)
```



En el gráfico de residuos vs valores ajustados que no se observa ningún patrón que haga pensar que la varianza varía a lo largo del eje x, si no que se ve una disribución homogénea. Por lo tanto, se cumple el supuesto de homocedasticidad.

- Distribución normal de los residuos

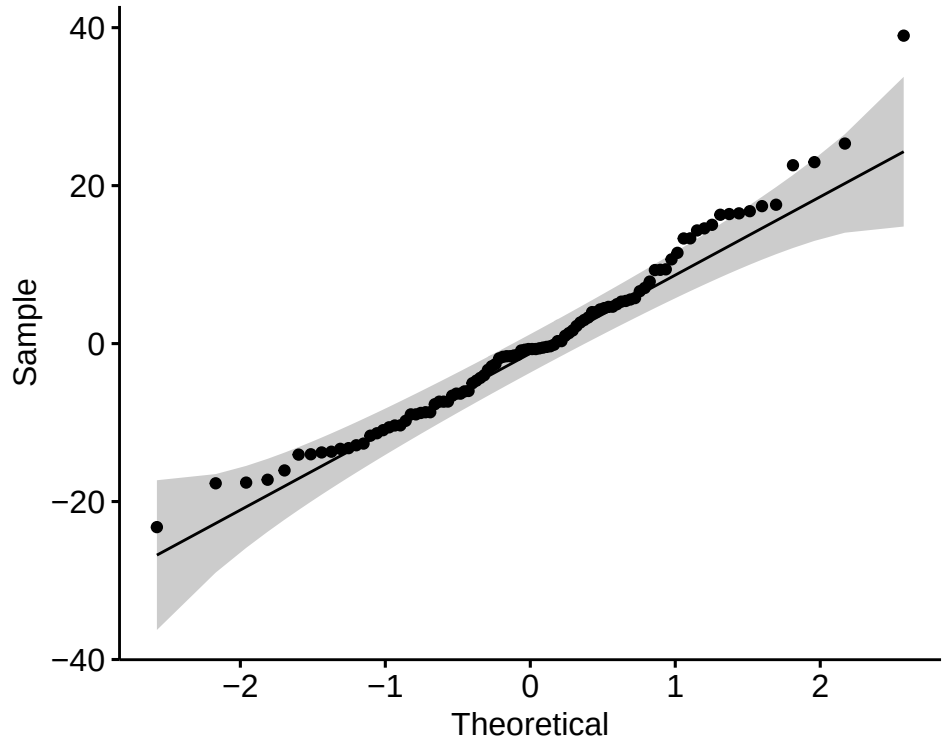
```
print(shapiro.test(ml3$residuals))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ml3$residuals
## W = 0.9727, p-value = 0.03569
```

```
print(lillie.test(ml3$residuals))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ml3$residuals
## D = 0.078692, p-value = 0.1334
```

```
ggqqplot(ml3$residuals)
```




```r
hist(ml3$residuals)
```

De forma similar a los casos anteriores, se puede ver que los métodos gráficos muestran una distribución normal, al igual que el test estadístico de Lilliefors, donde el p-valor es mayor a 0.05, y que el test de Shapiro-

Wilk indica que la distribución de los residuos no es normal. Analizando todos los métodos, concluyo que el supuesto de normalidad de los residuos se cumple.

```
summary(ml3)
```

```
##
## Call:
## lm(formula = sbp ~ gestage * sex, data = lowbwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.239  -7.930  -0.691   5.445  38.985
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.9805    15.2419   0.983   0.3282
## gestage        1.0903     0.5254   2.075   0.0406 *
## sex1         -15.1570    27.7433  -0.546   0.5861
## gestage:sex1   0.5714     0.9569   0.597   0.5518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.07 on 96 degrees of freedom
## Multiple R-squared:  0.08587,    Adjusted R-squared:  0.0573
## F-statistic: 3.006 on 3 and 96 DF,  p-value: 0.03412
```

A partir del p-valor del coeficiente de `sex`, que es mayor a 0.05, se puede ver que la variable `sex` no es significativa para el modelo, por lo que no la incluiría en el modelo que tiene a la edad gestacional.

Además, el coeficiente para la interacción entre `gestage` y `sexo` no da estadísticamente significativo, ya que el p-valor es mayor a 0.05, esto implica que esta variable no es significativa para el modelo. Por lo tanto, tampoco la incluiría en el modelo que tiene a la edad gestacional.

Ejercicio 2

La idea de este ejercicio es discutir qué significan distintos modelos de regresión múltiple. Probaremos distintos modelos en un solo conjunto de datos. Utilizaremos de vuelta el archivo `ingresos.csv`.

1. **En el ejercicio de regresión lineal simple, se propuso un modelo de cuadrados mínimos que llamaremos *Modelo A*. Se propone ajustar el *modelo A* a un modelo lineal múltiple (*Modelo B*), donde se estime el salario a partir de la educación y la edad.(2. Grafique los residuos del *Modelo B*). Interprete los parámetros del modelo.**

Genero el modelo lineal múltiple.

```
ml4 <- lm(salario ~ educacion + edad, data =ingresos )
```

Compruebo los supuestos del modelo de regresión lineal

- Media de los residuos igual a cero.

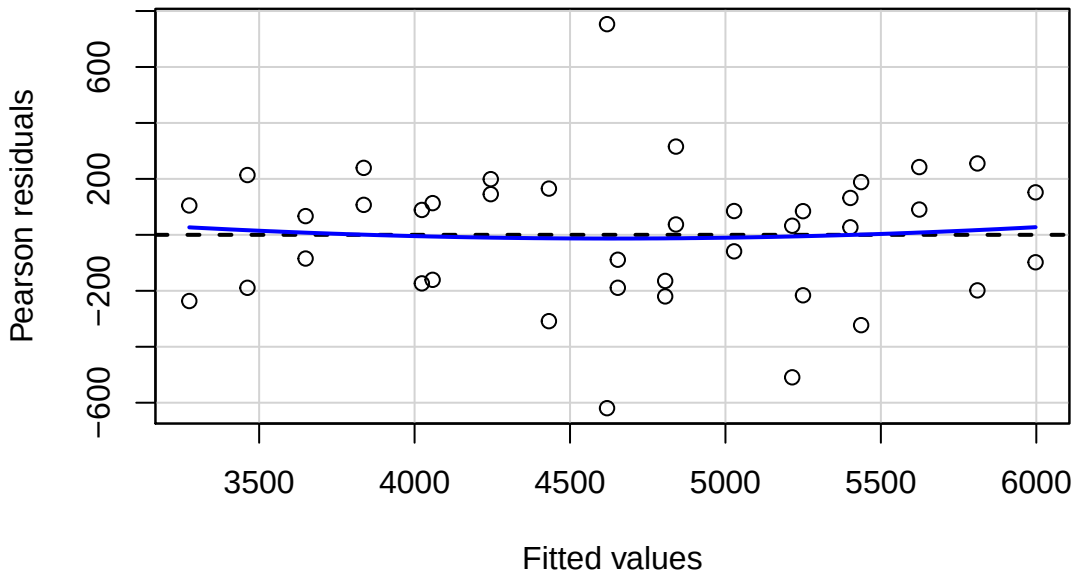
```r
mean(ml4$residuals)
```

```
## [1] -6.926404e-15
```

Se cumple el supuesto de que la media de los residuos sea igual a 0, ya que esta dio del orden de 10^{-15}.

- Homocedasticidad de los residuos

```r
residualPlot(ml4)
```



Se observa una varianza constante a lo largo del eje x, por lo que se cumple el supuesto de homocedasticidad.

- Distribución normal de los residuos

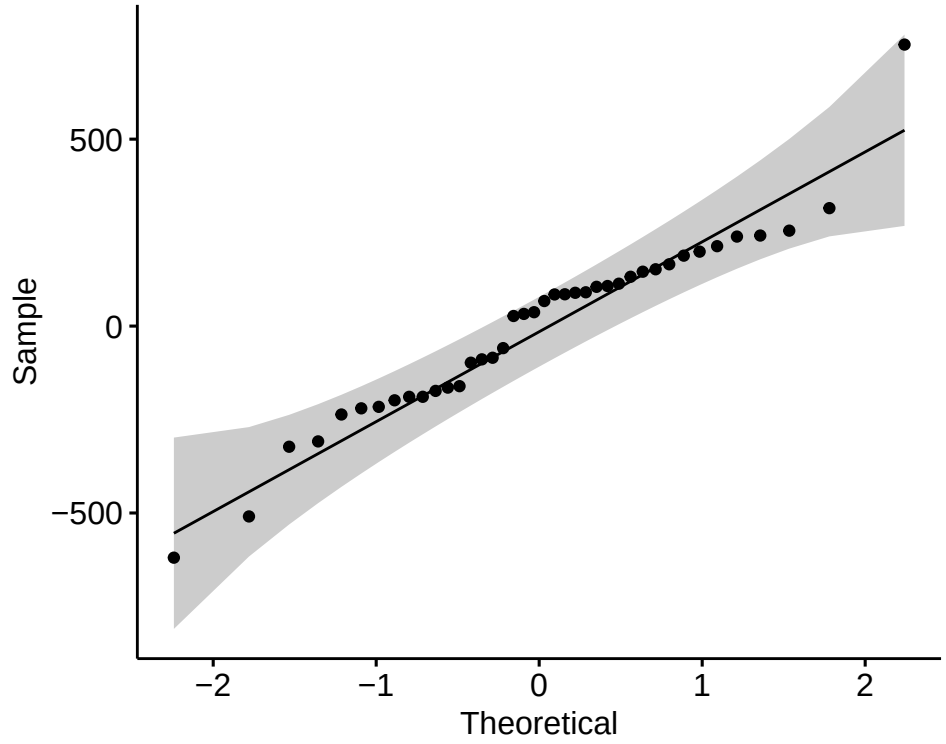
```r
print(shapiro.test(ml4$residuals))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ml4$residuals
## W = 0.95259, p-value = 0.0931
```

```r
print(lillie.test(ml4$residuals))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ml4$residuals
## D = 0.11878, p-value = 0.1661
```

```
ggqqplot(ml4$residuals)
```



El QQ-Plot muestra que todos los datos siguen la recta teórica de normalidad. Además, ambos tests estadísticos dieron un p-valor mayor a 0.05, por lo que no se rechaza la hipótesis nula que dice que los residuos tienen distribubción normal. Por lo tanto, se cumple el supuesto de normalidad de los residuos.

Analizo los parámetros del modelo lineal múltiple.

```
summary(ml4)
```

```
##
## Call:
## lm(formula = salario ~ educacion + edad, data = ingresos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -619.51 -177.35   52.03  147.06  753.03
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  213.588    348.890   0.612    0.544
## educacion    187.031     27.643   6.766 5.82e-08 ***
## edad          96.977      5.896  16.447  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 251.8 on 37 degrees of freedom
## Multiple R-squared:  0.9114, Adjusted R-squared:  0.9066
## F-statistic: 190.4 on 2 and 37 DF,  p-value: < 2.2e-16
```

Los coeficientes de ambas variables dieron significativos, ya que en los dos casos el p-valor es menor a 0.05. Esto significa que las dos variables explicativas son significativas para el modelo.

Este modelo tiene un R^2 mucho más alto que el del **Modelo A** (0.9066 vs 0.2445), por lo tanto, el **Modelo B** ajusta mucho mejor los datos.

3. **Ajustemos ahora un modelo con interacción (*Modelo C*) entre educación y edad ¿Con cuál de los dos modelos (*B* o *C*) se quedaría? ¿Por qué?**

Genero el **Modelo C**.

```
ml5 <- lm(salario ~ educacion*edad, data =ingresos )
```

Compruebo los supuestos del modelo de regresión lineal

- Media de los residuos igual a cero.

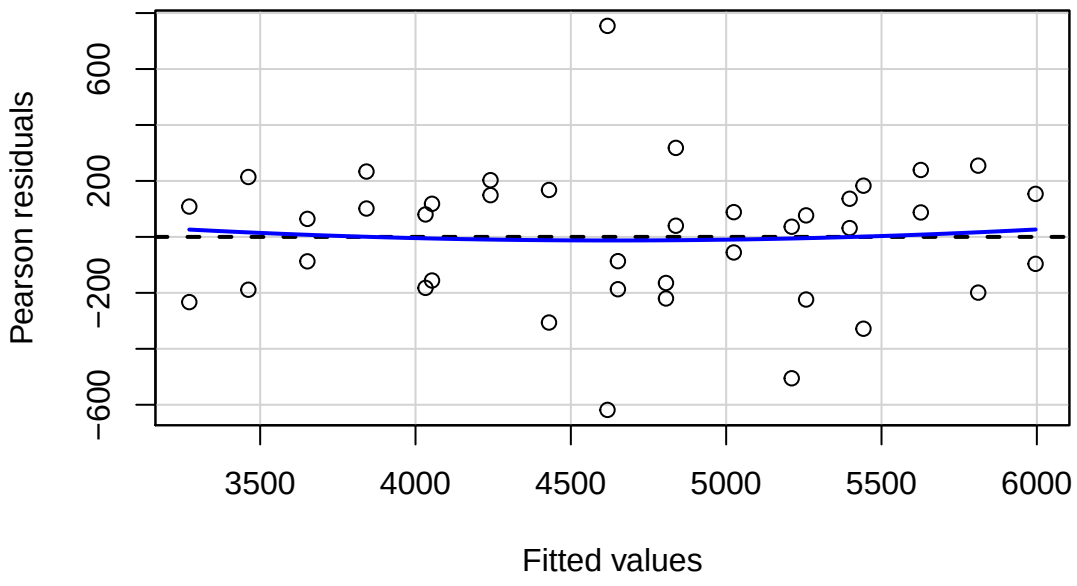
```
mean(ml5$residuals)
```

```
## [1] -4.796857e-15
```

El supuesto se cumple ya que la media da del orden de 10^{-15}.

- Homocedasticidad de los residuos

```
residualPlot(ml5)
```



No se observa ningún patrón en el gráfico de residuos, sino que se ve una varianza constante a lo largo del eje x, por lo que se cumple el supuesto de homocedasticidad.

- Distribución normal de los residuos

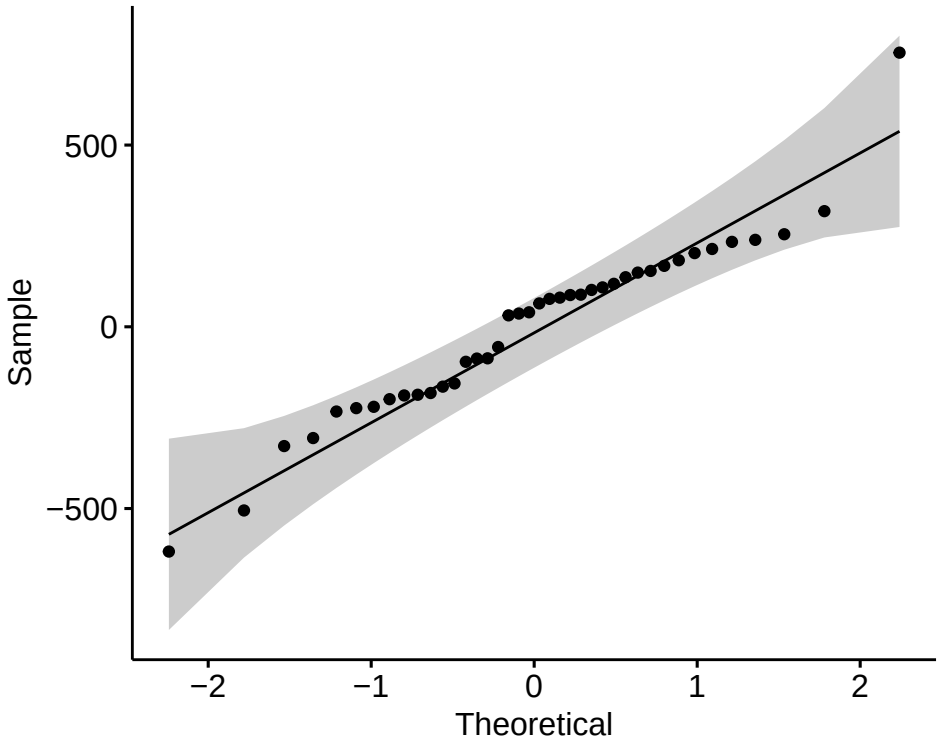
```
print(shapiro.test(ml5$residuals))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ml5$residuals
## W = 0.95299, p-value = 0.09614
```

```
print(lillie.test(ml5$residuals))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ml5$residuals
## D = 0.12599, p-value = 0.1131
```

```
ggqqplot(ml5$residuals)
```



Podemos ver en el QQ-Plot que todos los datos siguen la recta teórica de normalidad. Además, ambos tests estadísticos dieron un p-valor mayor a 0.05, por lo que no se rechaza la hipótesis nula que dice que los residuos tienen distribubción normal. Entonces, se cumple el supuesto de normalidad de los residuos.

Habiendo comprobado los supuestos del modelo lineal, analizo los parámetros del **Modelo C**.

```
summary(ml5)
```

```
##
## Call:
## lm(formula = salario ~ educacion * edad, data = ingresos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -618.35 -183.43   52.05  150.23  754.19
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    170.5476   552.4044   0.309   0.7593
## educacion      193.7167    71.6241   2.705   0.0104 *
## edad            98.0405    12.0737   8.120 1.18e-09 ***
## educacion:edad  -0.1813     1.7879  -0.101   0.9198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 255.3 on 36 degrees of freedom
## Multiple R-squared:  0.9115, Adjusted R-squared:  0.9041
## F-statistic: 123.5 on 3 and 36 DF,  p-value: < 2.2e-16
```

Podemos ver que el coeficiente estimado que acompaña a la covariable `educacion:edad` tiene un p-valor

mayor a 0.05, por lo que esta interacción no es estadísticamente significativa para el modelo. Por esta razón, se puede eliminar.

El R^2 dio muy similar al del **Modelo B**, pero un poco menor. Por esta razón, y habiendo visto que la interacción entre `educacion` y `edad` no es una variable explicativa significativa para el modelo, me quedaría con el **Modelo B**.

Regresión logística

Se desea estudiar la asociación entre la incidencia de cáncer de vejiga con el consumo de café y tabaco. Para ello se tomó una muestra de 200 pacientes y se registraron las siguientes variables en la base de datos `canvej.csv`:

- **Café: 1 (consume café) y 0 (no consume café)**
- **Can_vej: 1 (padece la afección) y 0 (no la padece)**
- **Fuma: 1 (fuma) y 0 (no fuma)**

```
canvej <- read.csv2('canvej.csv', header = TRUE)
summary(canvej)
```

```
##       CAFE         CAN_VEJ         FUMA
##  Min.   :0.00   Min.   :0.0   Min.   :0.00
##  1st Qu.:0.00   1st Qu.:0.0   1st Qu.:1.00
##  Median :1.00   Median :0.5   Median :1.00
##  Mean   :0.58   Mean   :0.5   Mean   :0.77
##  3rd Qu.:1.00   3rd Qu.:1.0   3rd Qu.:1.00
##  Max.   :1.00   Max.   :1.0   Max.   :1.00
```

Paso a factor todas las variables.

```
canvej$CAFE <- as.factor(canvej$CAFE)
str(canvej$CAFE)
```

```
##  Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

```
canvej$CAN_VEJ <- as.factor(canvej$CAN_VEJ)
str(canvej$CAN_VEJ)
```

```
##  Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
canvej$FUMA <- as.factor(canvej$FUMA)
str(canvej$FUMA)
```

```
##  Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

1. **Analice si el consumo de café se asocia con el cáncer de vejiga, mediante el test de Chi cuadrado de Pearson.**

Tabla de frecuencias.

```r
t1 = table(canvej$CAFE,canvej$CAN_VEJ)
t1 = t1[rev(order(rownames(t1))) ,
                              rev(order(colnames(t1)))]
kable(t1, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	1	0
1	66	50
0	34	50

Test de Chi cuadrado de Pearson.

Los supuestos del test son:

- más del 80% de las celdas tienen una frecuencia esperada mayor o igual a 5 y ninguna frecuencia esperada menor a 1
- las frecuencias marginales de la tabla son conocidas.

Como todas las celdas de la tabla de frecuencia son mayores a 5, y bajo el supuesto de que las variables son independientes, las frecuencias esperadas son iguales a las observadas y se conocen las frecuencias marginales, entonces se cumplen los supuestos.

Las hipótesis del test son:

- H_0: el consumo de café y el cáncer de vejiga son variables independientes
- H_1: el consumo de café y el cáncer de vejiga no son variables independientes.

```r
chisq.test(t1,correct=TRUE)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t1
## X-squared = 4.6182, df = 1, p-value = 0.03163
```

Con un p-valor menor a 0.05 se rechaza la hipótesis nula de que las variables son independientes, por lo tanto hay una asociación entre el consumo de café y el cáncer de vejiga.

2. **Estime el odds ratio puntualmente y por intervalo. Indique si le parece que el consumo de café podría considerarse un factor de riesgo para el cáncer de vejiga. Realice el mismo análisis para la relación entre tabaquismo y cáncer de vejiga.**

Café

Odds Ratio

```
epi.2by2(dat = t1, method = "cohort.count", conf.level = 0.95,
units = 100,outcome = "as.columns")
```

```
##              Outcome +    Outcome -      Total        Inc risk *        Odds
## Exposed +           66           50        116              56.9        1.32
## Exposed -           34           50         84              40.5        0.68
## Total              100          100        200              50.0        1.00
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 1.41 (1.04, 1.90)
## Odds ratio                                     1.94 (1.10, 3.43)
## Attrib risk *                                  16.42 (2.59, 30.25)
## Attrib risk in population *                    9.52 (-3.05, 22.10)
## Attrib fraction in exposed (%)                 28.86 (3.60, 47.50)
## Attrib fraction in population (%)              19.05 (0.95, 33.84)
## -------------------------------------------------------------------
##  Test that OR = 1: chi2(1) = 5.255 Pr>chi2 = 0.02
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 population units
```

El consumo de café aumenta en 1.94 veces el riesgo de desarollar cáncer de vejiga, con un intervalo de confianza del 95% entre 1.10 y 3.43. Por lo tanto, a priori, el café sería un factor de riesgo para el cáncer de vejiga.

Tabaquismo

Tabla de frecuencias.

```
t2 = table(canvej$FUMA,canvej$CAN_VEJ)
t2 = t2[rev(order(rownames(t2))) ,
                            rev(order(colnames(t2)))]
kable(t2, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	1	0
1	94	60
0	6	40

Test de Chi cuadrado de Pearson.

Las hipótesis del test son:

- H_0: el tabaquismo y el cáncer de vejiga son variables independientes
- H_1: el tabaquismo y el cáncer de vejiga no son variables independientes.

Todas las celdas de la tabla de frecuencia son mayores a 5, y bajo el supuesto de que las variables son independientes, las frecuencias esperadas son iguales a las observadas y se conocen las frecuencias marginales, por lo tanto se cumplen los supuestos del test. Realizo el test.

```
chisq.test(t2,correct=TRUE)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t2
## X-squared = 30.745, df = 1, p-value = 2.942e-08
```

Como el p-valor es menor a 0.05, las variables tabaquismo y cáncer de vejiga no son independientes y existe una asociación estadísticamente significativa entre ellas.

Odds Ratio

```
epi.2by2(dat = t2, method = "cohort.count", conf.level = 0.95,
units = 100,outcome = "as.columns")
```

```
##              Outcome +    Outcome -      Total        Inc risk *        Odds
## Exposed +           94           60        154                61        1.57
## Exposed -            6           40         46                13        0.15
## Total              100          100        200                50        1.00
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 4.68 (2.20, 9.97)
## Odds ratio                                     10.44 (4.17, 26.13)
## Attrib risk *                                  48.00 (35.58, 60.41)
## Attrib risk in population *                    36.96 (25.01, 48.90)
## Attrib fraction in exposed (%)                 78.63 (54.46, 89.97)
## Attrib fraction in population (%)              73.91 (46.75, 87.22)
## -------------------------------------------------------------------
##  Test that OR = 1: chi2(1) = 32.637 Pr>chi2 = <0.001
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 population units
```

El OR del tabaquismo dio 10.44 por lo que este aumenta en 10.44 veces el riesgo de desarrollar cáncer de vejiga, con un intervalo de confianza del 95% entre 4.17 y 26.13. Por lo tanto, el tabaquismo parece ser un factor de riesgo para el cáncer de vejiga.

3. **Estudie la relación entre el consumo de café y el consumo de tabaco.**

Genero la tabla de frecuencias para el consumo de café y el consumo de tabaco. En las columnas coloco la variable `CAFE` y en las filas la variable `FUMA`.

```
t3 = table(canvej$FUMA,canvej$CAFE)
t3 = t3[rev(order(rownames(t3))) ,
                              rev(order(colnames(t3)))]
kable(t3, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	1	0
1	94	60
0	22	24

Test de Chi cuadrado de Pearson

Las hipótesis del test son:

- H_0: el consumo de café y el tabaquismo son variables independientes
- H_1: el consumo de café y el tabaquismo no son variables independientes.

```
chisq.test(t3,correct=TRUE)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t3
## X-squared = 2.025, df = 1, p-value = 0.1547
```

Como el p-valor dio mayor a 0.05, no se rechaza la hipótesis nula y se concluye que estas variables son independientes, por lo tanto no hay una asociación significativa entre el consumo de café y el consumo de tabaco.

4. **Construya un modelo de regresión logística para estimar la probabilidad de padecer cáncer de vejiga, considerando como variable predictora el consumo de café.**

```
m1 <- glm(CAN_VEJ ~ CAFE, data = canvej, family = "binomial")
summary(m1)
```

```
##
## Call:
## glm(formula = CAN_VEJ ~ CAFE, family = "binomial", data = canvej)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2974  -1.0883   0.0217   1.0620   1.3450
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.3857     0.2223  -1.735   0.0827 .
## CAFE1         0.6633     0.2908   2.281   0.0226 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 277.26  on 199  degrees of freedom
## Residual deviance: 271.98  on 198  degrees of freedom
## AIC: 275.98
##
## Number of Fisher Scoring iterations: 4
```

El coeficiente β dio un p-valor menor a 0.05, por lo que se puede decir que este difiere significativamente de 0 y entonces la variable `CAFE` es estadísticamente significativa para el modelo.

5. **Construya un modelo de regresión logística para estimar la probabilidad de padecer cáncer de vejiga, considerando como variable predictora el consumo de tabaco.**

```
m2 <- glm(CAN_VEJ ~ FUMA, data = canvej, family = 'binomial')
summary(m2)
```

```
##
## Call:
## glm(formula = CAN_VEJ ~ FUMA, family = "binomial", data = canvej)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3730  -1.3730   0.2325   0.9936   2.0184
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.8971     0.4378  -4.333 1.47e-05 ***
## FUMA1         2.3461     0.4679   5.014 5.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 277.26  on 199  degrees of freedom
## Residual deviance: 241.54  on 198  degrees of freedom
## AIC: 245.54
##
## Number of Fisher Scoring iterations: 4
```

El coeficiente β para la variable `FUMA` dio un p-valor menor a 0.05, por lo que este difiere significativamente de 0 y entonces el consumo de tabaco es estadísticamente significativo para el modelo.

6. **Construya un modelo de regresión logística para estimar la probabilidad de padecer cáncer de vejiga, considerando como variables predictoras: el consumo de café y el consumo de tabaco.**

```
m3 <- glm(CAN_VEJ ~ CAFE + FUMA, data = canvej, family = "binomial")
summary(m3)
```

```
##
## Call:
## glm(formula = CAN_VEJ ~ CAFE + FUMA, family = "binomial", data = canvej)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4764  -1.2233   0.2236   0.9054   2.1476
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.2012     0.4770  -4.615 3.93e-06 ***
## CAFE1         0.5727     0.3164   1.810   0.0703 .
## FUMA1         2.3086     0.4704   4.908 9.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 277.26  on 199  degrees of freedom
## Residual deviance: 238.25  on 197  degrees of freedom
## AIC: 244.25
##
## Number of Fisher Scoring iterations: 4
```

El coeficiente estimado de la variable `CAFE` cambió su significación al agregar la variable `FUMA`, ya que antes el p-valor era menor a 0.05 y ahora es mayor. Por lo tanto, hay un efecto confundido en la relación entre `CAFE` Y `FUMA`. Además, la variable `CAFE` dio no significativa para el modelo, por lo que puede sacarse del modelo, mientras que la variable `FUMA` sigue dando significativa, y debe dejarse.

7. **A partir del modelo construido en g indique la probabilidad de padecer cáncer de vejiga para un individuo que:**

a) **No consume café ni tabaco.**

```
predict(m3, data.frame(CAFE = as.factor(0), FUMA = as.factor(0)),type="response")
```

```
##          1
## 0.09964139
```

La probabilidad de padecer cáncer de vejiga para un individuo que no consume café ni tabaco es de 9.964%.

b) **Consume café y tabaco.**

```
predict(m3, data.frame(CAFE = as.factor(1), FUMA = as.factor(1)),type="response")
```

```
##         1
## 0.6637382
```

La probabilidad de padecer cáncer de vejiga para un individuo que consume café y tabaco es de 66.374%.

c) **Consume café pero no consume tabaco.**

```r
predict(m3, data.frame(CAFE = as.factor(1), FUMA = as.factor(0)),type="response")
```

```
##         1
## 0.1640276
```

La probabilidad de padecer cáncer de vejiga para un individuo que consume café y no consume tabaco es de 16.403%.

d) **Consume tabaco pero no café.**

```r
predict(m3, data.frame(CAFE = as.factor(0), FUMA = as.factor(1)),type="response")
```

```
##         1
## 0.5268101
```

La probabilidad de padecer cáncer de vejiga para un individuo que no consume café pero sí consume tabaco es de 52.681%.

8. **¿Cómo interpretaría los coeficientes del modelo? ¿Cuáles resultan significativos? ¿Cuál de los tres modelos planteados presenta un mejor ajuste?**

Los coeficientes del modelo son el $\ln(OR)$ de la variable explicativa correspondiente. Por lo tanto el OR de una variable es e^{β}, siendo β el coeficiente estimado para esa variable.

Los coeficientes significativos para el modelo de regresión múltiple fueron el de `FUMA` y el coeficiente Intercept.

Genero la curva ROC para cada modelo.

- Variable explicativa: `CAFE`

```r
prob=predict(m1,type=c("response"))
canvej$prob = prob
roc1 <- roc(CAN_VEJ ~ prob, data = canvej)
roc1$auc
```

```
## Area under the curve: 0.58
```

- Variable explicativa: `FUMA`

```r
prob=predict(m2,type=c("response"))
canvej$prob = prob
roc2 <- roc(CAN_VEJ ~ prob, data = canvej)
roc2$auc
```

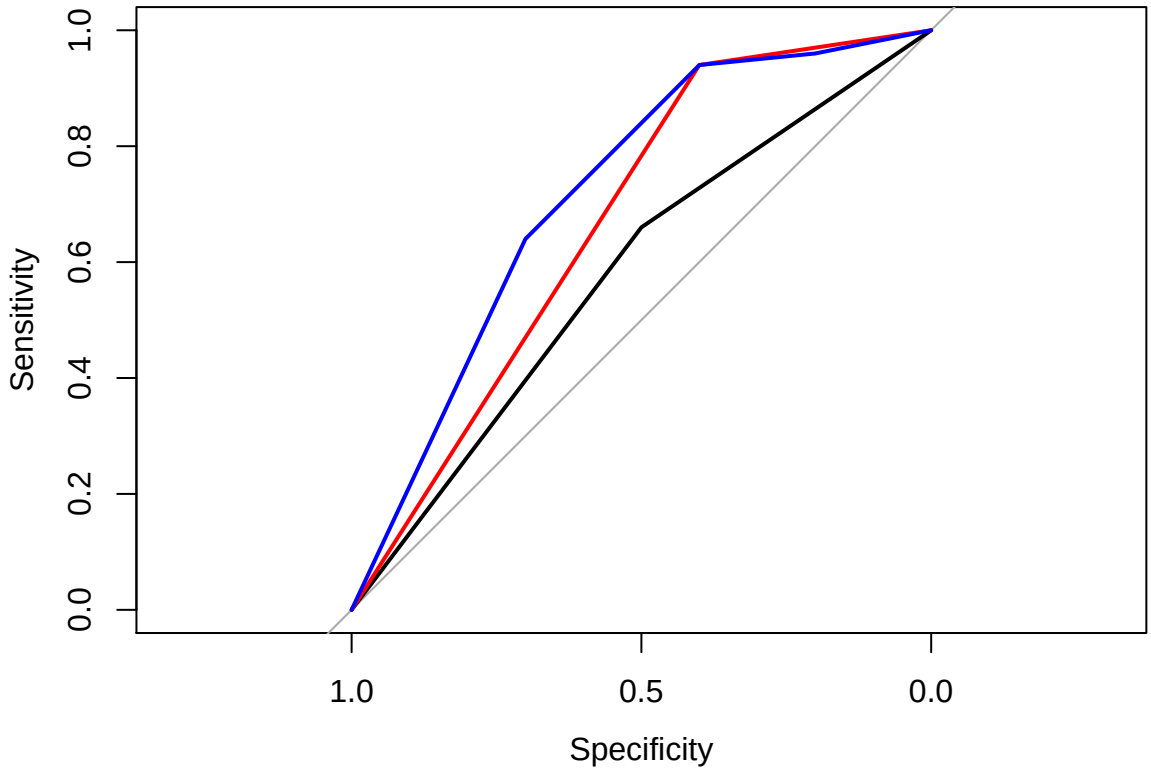
```
## Area under the curve: 0.67
```

- Variables explicativas: `CAFE` y `FUMA`

```
prob=predict(m3,type=c("response"))
canvej$prob = prob
roc3 <- roc(CAN_VEJ ~ prob, data = canvej)
roc3$auc
```

```
## Area under the curve: 0.719
```

```
plot(roc1)
plot(roc2, add=TRUE, col='red')
plot(roc3, add=TRUE, col='blue')
```



- Negro: `CAFÉ`
- Rojo: `FUMA`
- Azul: `CAFÉ` + `FUMA`

A partir de las curvas ROC y del AUC de cada modelo, se puede concluir que el mejor es el que utiliza tanto la variable `FUMA` como la variable `CAFE` como variables explicativas del modelo. En cambio, el modelo que solo utiliza la variable `CAFE` es el que peor rendimiento presenta.