Universitat Oberta de Catalunya

Patricia Lázaro Tello
plazarotello@uoc.edu

A4: Análisis de la varianza y repaso del curso

Patricia Lázaro Tello

Índice general

Patricia Lázaro Tello
plazarotello@uoc.edu

1 Introducción

En este trabajo se va a analizar estadísticamente el conjunto de datos `CensusIncomedata.txt`, un *dataset* que contiene información de individuos tal que s sabe su edad, sector laboral, profesión, estado civil, ocupación, género, horas semanales trabajadas y estudios, así como los ingresos que percibe anualmente, expresados en miles de euros.

Se desea analizar especialmente si existe sesgo de género y raza, y el impacto de tales sesgos. También se busca analizar el impacto de otras variables en los ingresos percibidos, como puede ser el sector en que trabaja, el trabajo que desempeña, su edad o su nivel educativo.

---

2 Lectura y preparación de los datos

Se procede en primer lugar a cargar en memoria el archivo `CensusIncomedata.txt` y comprobar sus variables:

```
adult <- read.csv('CensusIncomedata.txt', sep=' ')
adult.dim <- dim(adult)

head(adult, n=3)
```

```
##   age      workclass education_num marital_status    occupation   race   sex
## 1  50  Self-Employed            13        Married  White-Collar  White  Male
## 2  38        Private             9       Divorced   Blue-Collar  White  Male
## 3  53        Private             7        Married   Blue-Collar  Black  Male
##   hours_per_week income
## 1             13     54
## 2             40     52
## 3             40     50
```

```
colSums(is.na(adult))
```

```
##            age      workclass  education_num marital_status     occupation
##              0              0              0              0              0
##           race            sex hours_per_week         income
##              0              0              0              0
```

El *dataset* contiene 32.560 observaciones con 9 variables en cada una de ellas. No existen valores nulos en ninguno de los atributos del conjunto de datos.

2.1 Preparación de los datos

A continuación se procede a limpiar los datos para que puedan ser tratados de manera más sencilla en pasos posteriores del trabajo. En primer lugar se quitarán los espacios en blanco que pueda haber en las variables `workclass`, `marital_status`, `occupation`, `race` y `sex`:

```
adult <- adult %>% dplyr::mutate(workclass=trimws(workclass, 'b'),
                                 marital_status=trimws(marital_status, 'b'),
                                 occupation=trimws(occupation, 'b'),
                                 race=trimws(race, 'b'),
                                 sex=trimws(sex, 'b'))
```

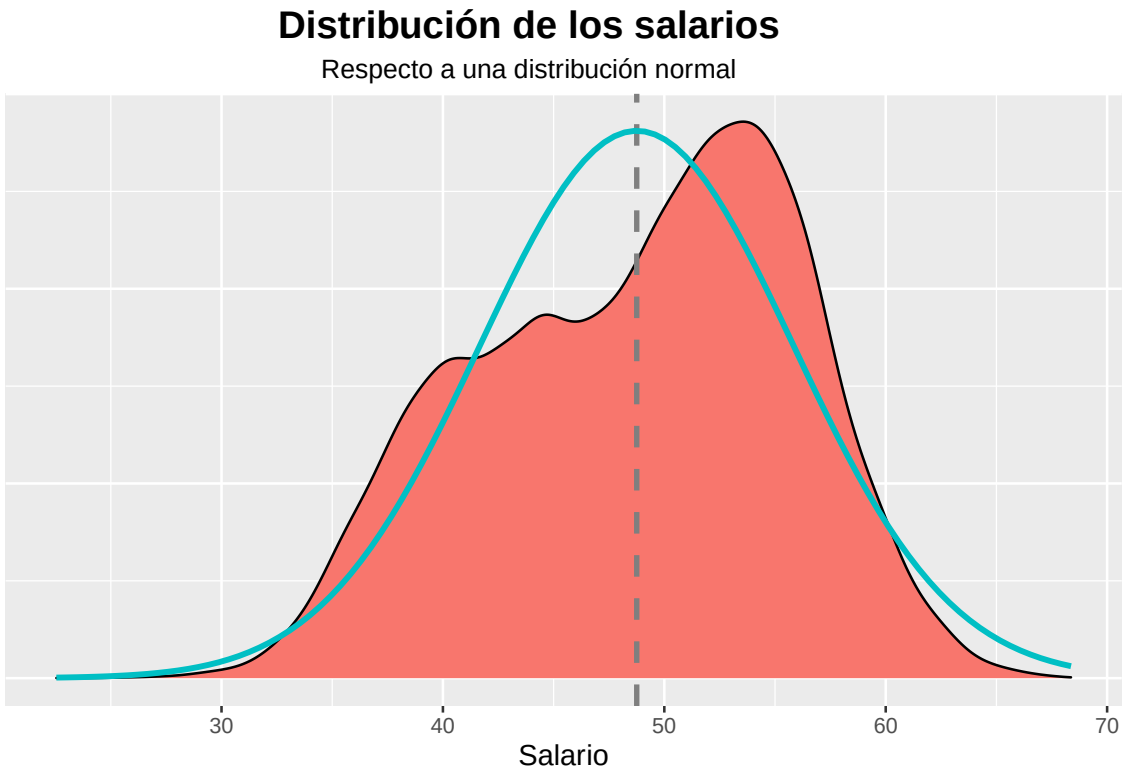
La variable `sex` se refiere al rol social o percepción individual del género propia del individuo y no a su sexo. Se procede a cambiar el nombre de la variable para reflejar su significado real:

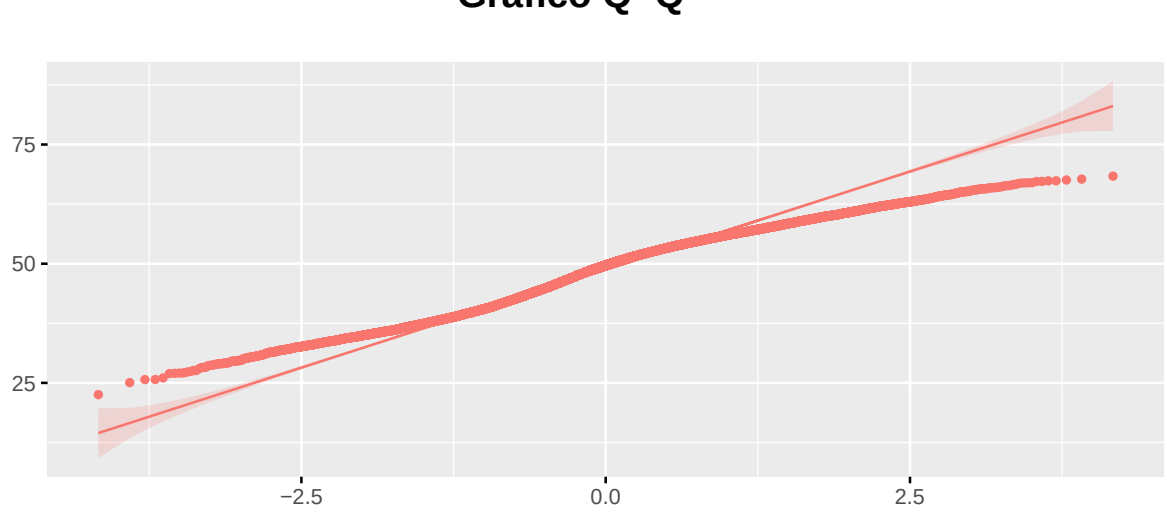
```
adult <- adult %>% dplyr::mutate(gender=sex, sex=NULL)
```

Seguidamente se procede a comprobar la normalidad de la variable `salario` mediante inspección visual y el test de normalidad de Lilliefors (Kolmogorov-Smirnov):

```
income.n <- length(adult$income)
income.mean <- mean(adult$income)
income.sd <- sd(adult$income)

ggarrange(nrow=2, ncol=1, align='hv', heights=c(1, 0.75),
          ggplot(adult, aes(x=income)) +
            geom_density(mapping=aes(y=..density..), fill=default.color.main) +
            geom_vline(xintercept=income.mean, size=1.05,
                       linetype='dashed', color='gray50') +
            stat_function(fun=dnorm, args=c(mean=income.mean,
                                            sd=income.sd),
                          color=default.color.secondary, size=1.15) +
            no.axis.y + xlab('Salario') + ylab('') + title.centered +
            ggtitle('Distribución de los salarios',
                    subtitle='Respecto a una distribución normal'),

          ggqqplot(adult$income, color=default.color.main,
                   ggtheme = theme_gray(), xlab='Cuantiles teóricos',
                   ylab='Cuantiles de la muestra', title='Gráfico Q-Q',
                   shape=16) + title.centered + xlab('') + ylab('')
```

```
)
```






```
lillie.test(adult$income)
```

```
##
```

```
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  adult$income
## D = 0.06, p-value <0.0000000000000002
```

El test de normalidad de Lilliefors confirma que la distribución de la variable `income` no es normal (p-value $< \alpha = 0.05$), como se puede comprobar en la figura anterior.

La mayor concentración de valores se da en torno a $[53, 56]$ miles de euros, mientras que la media se sitúa a la izquierda, en 49 miles de euros. Se observa también que las colas son más amplias que las de una distribución normal, aunque decaen antes.

A continuación se procede a crear una nueva variable `Less50` para clasificar los salarios dependiendo de si el ingreso es menor que 50k euros o no:

```
adult <- adult %>% dplyr::mutate(Less50=(income < 50.0))

adult %>% dplyr::select(income, Less50) %>% head(n=5)
```

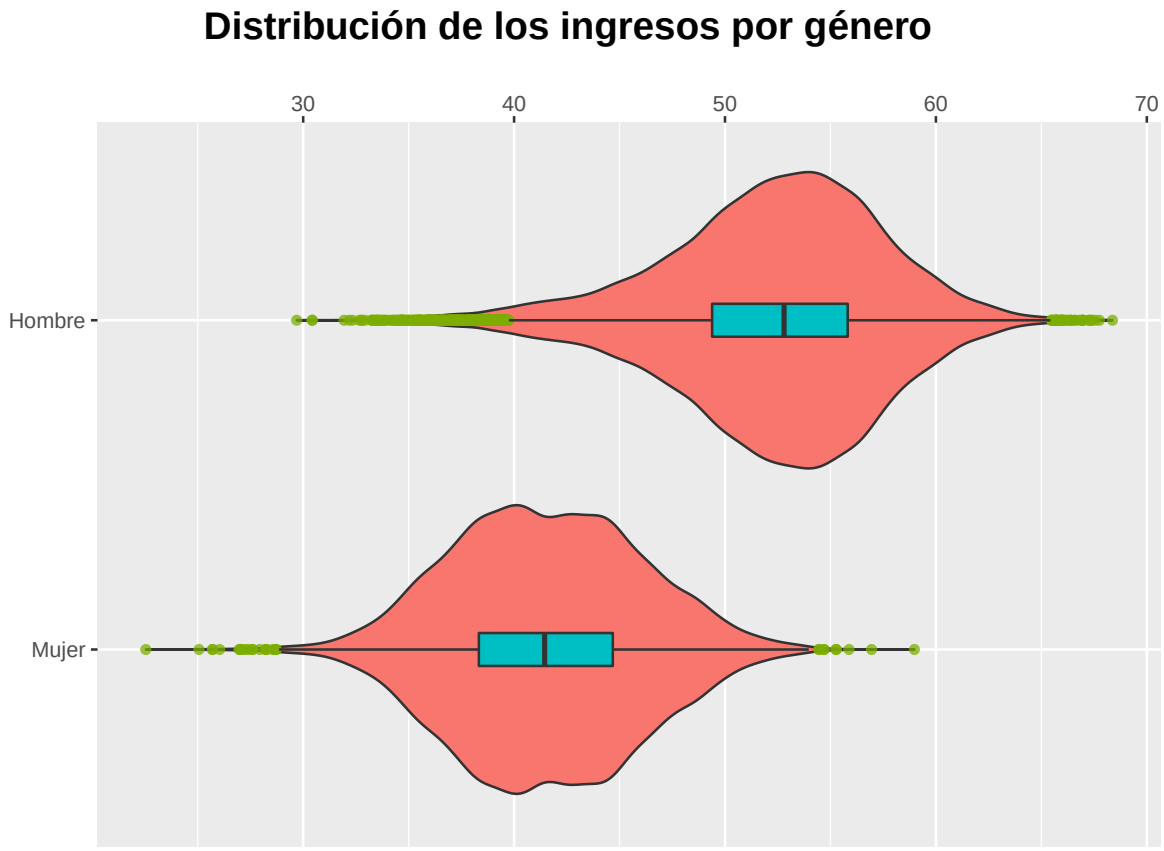
```
##   income Less50
## 1     54  FALSE
## 2     52  FALSE
## 3     50  FALSE
## 4     44   TRUE
## 5     49   TRUE
```

2.2 Análisis visual

Se procede a analizar la distribución de los ingresos según los distintos atributos que hay en el conjunto de datos. Un análisis preliminar podría sugerir la existencia de relaciones entre los ingresos y alguna de las variables.

```
ggplot(data=adult, mapping=aes(x=gender, y=income)) +
  geom_violin(fill=default.color.main) +
  scale_y_continuous(position='right') +
  scale_x_discrete(labels=as_labeller(c(`Male`='Hombre', `Female`='Mujer'))) +
  geom_boxplot(width=0.1, outlier.color=default.color.terciary,
               fill=default.color.secondary, outlier.alpha=0.75) +
  coord_flip() + title.centered + xlab('') + ylab('') +
  ggtitle('Distribución de los ingresos por género')
```

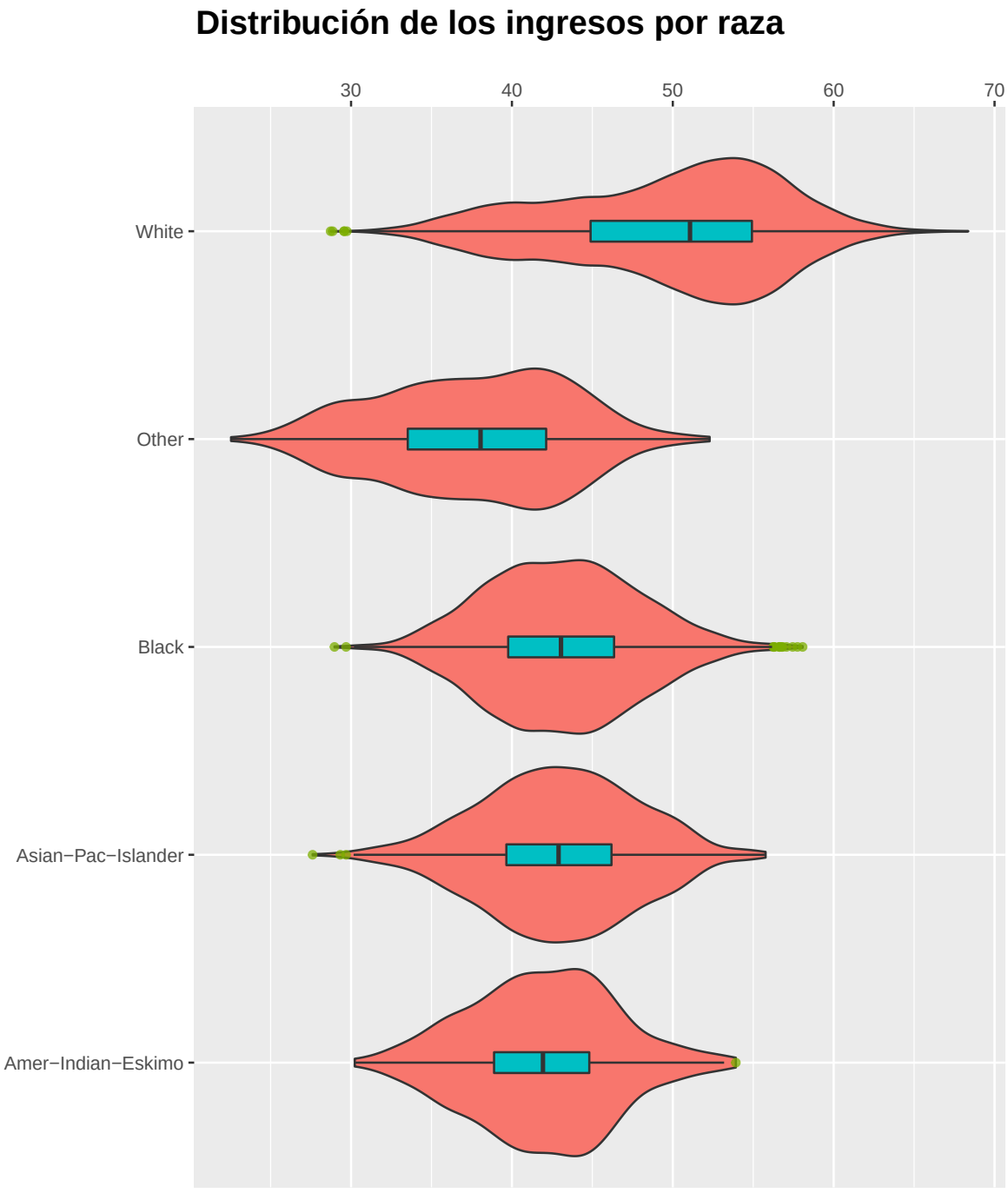
Patricia Lázaro Tello
plazarotello@uoc.edu



Se observa que de media las mujeres cobran menos que los hombres. A priori se puede hablar de sesgo de género, aunque será necesario analizar más a fondo los datos para confirmar su existencia.

```
ggplot(data=adult, mapping=aes(x=race, y=income)) +
  geom_violin(fill=default.color.main) +
  scale_y_continuous(position='right') +
  geom_boxplot(width=0.1, outlier.color=default.color.terciary,
               fill=default.color.secondary, outlier.alpha=0.75) +
  coord_flip() + xlab('') + ylab('') + title.centered +
  ggtitle('Distribución de los ingresos por raza')
```

Patricia Lázaro Tello
plazarotello@uoc.edu



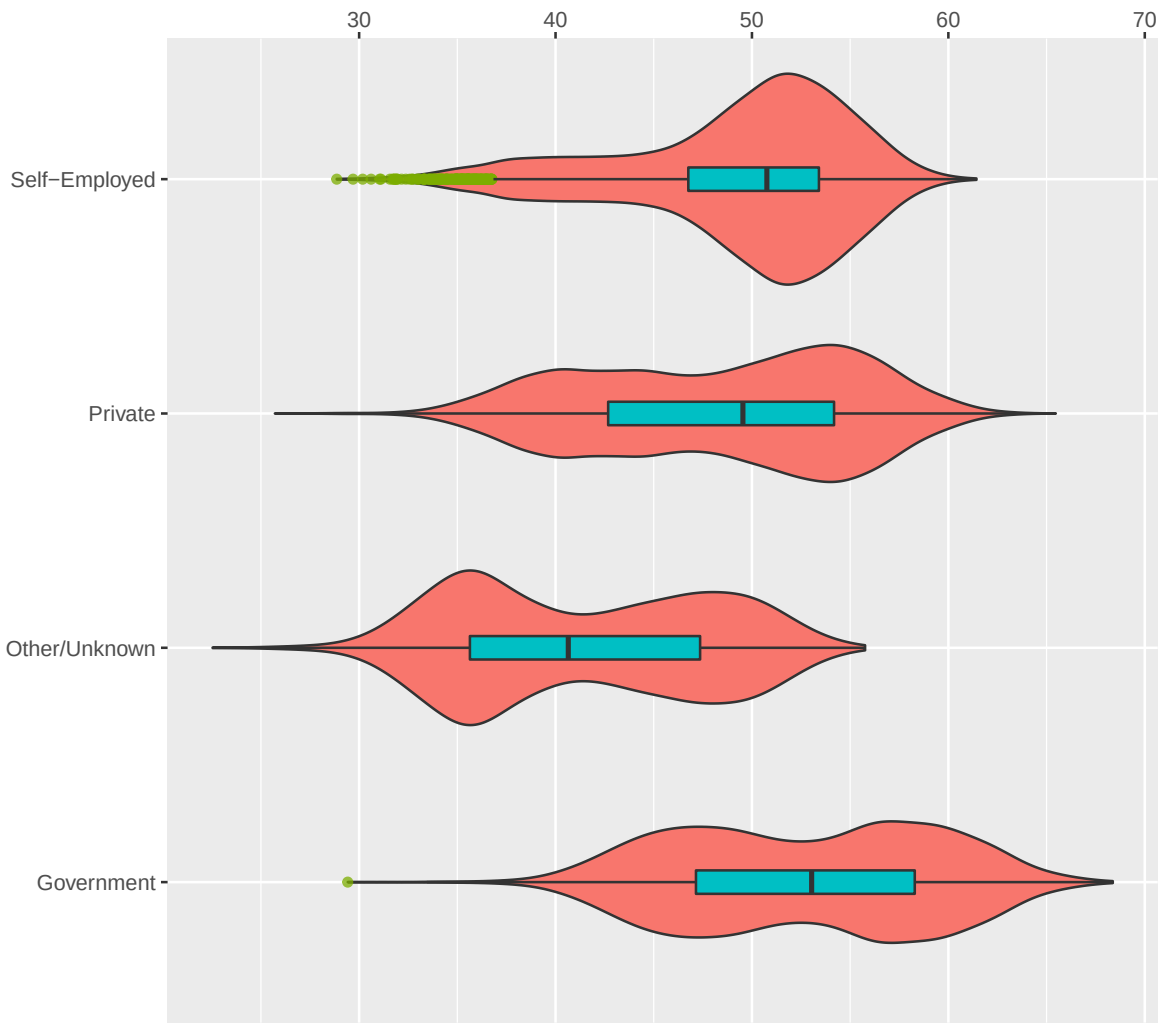
Se observa que también existe sesgo de raza: las personas blancas cobran de media más que el resto de razas. Las personas de raza negra, asiática e india cobran aproximadamente lo mismo, mientras que los identificados como de otra raza distinta a las nombradas cobran menos que los demás.

De las 3 razas con media de ingresos similares, las personas indias tienen menos

Patricia Lázaro Tello
plazarotello@uoc.edu

posibilidades de cobrar más de 45k euros que el resto. A su vez, observando las colas de los *violin plots*, la raza asiática es la que más posibilidades tiene de cobrar más de 45k euros.

```
ggplot(data=adult, mapping=aes(x=workclass, y=income)) +
  geom_violin(fill=default.color.main) +
  scale_y_continuous(position='right') +
  geom_boxplot(width=0.1, outlier.color=default.color.terciary,
               fill=default.color.secondary, outlier.alpha=0.75) +
  coord_flip() + xlab('') + ylab('') + title.centered +
  ggtitle('Distribución de los ingresos por tipo de trabajo')
```



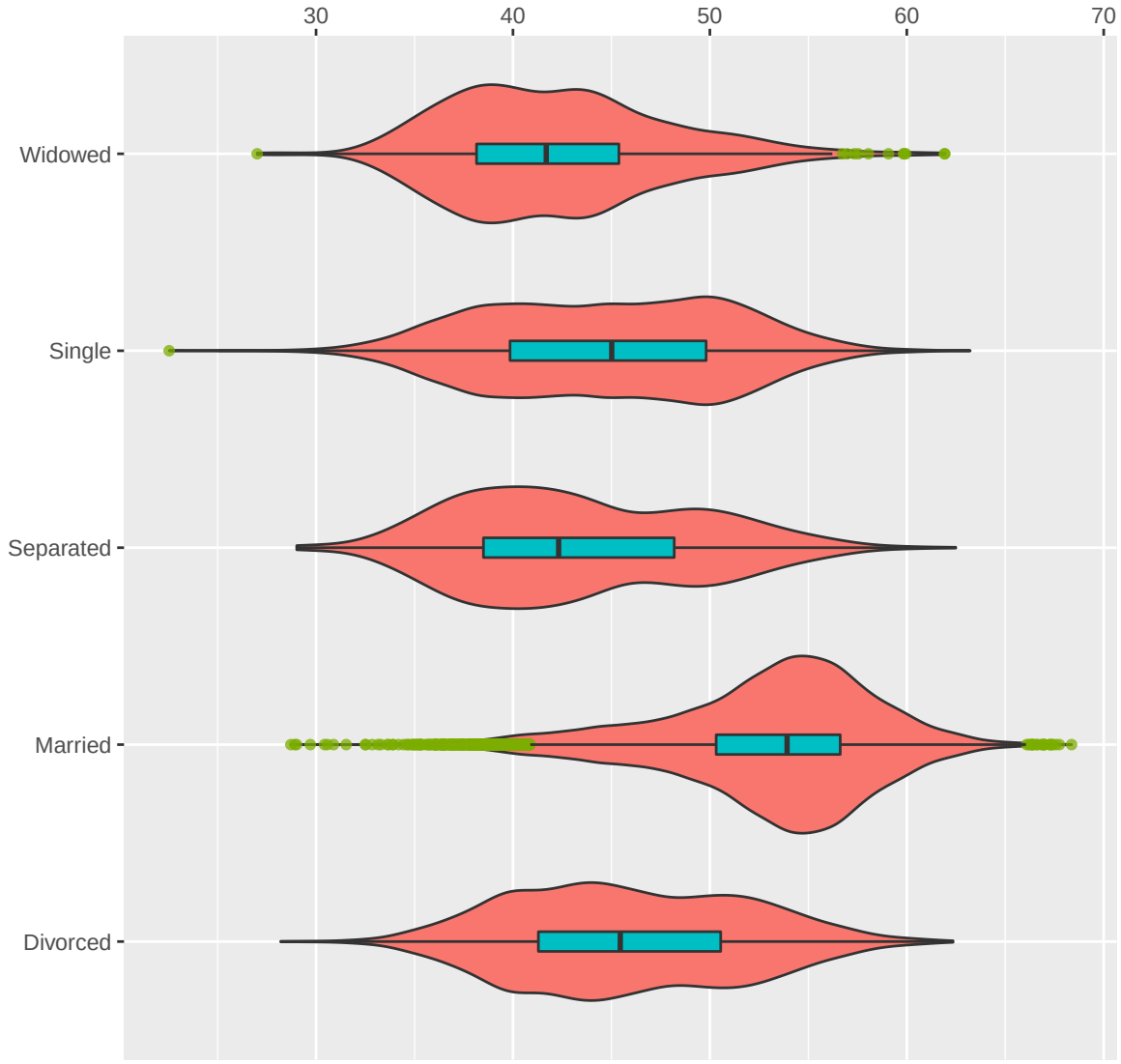
Patricia Lázaro Tello
plazarotello@uoc.edu

Las personas que trabajan para el gobierno (funcionarios públicos) son las que perciben de media los ingresos más altos. Los trabajadores del sector privado son los que tienen unos ingresos más dispersos, mientras que los autónomos suelen cobrar de media en torno a 50k euros.

Las personas cuyo trabajo se desconoce o no se tienen datos suficientes en el *dataset* son las que menos cobran, casi 10k euros menos que el resto.

```
ggplot(data=adult, mapping=aes(x=marital_status, y=income)) +
  geom_violin(fill=default.color.main) +
  scale_y_continuous(position='right') +
  geom_boxplot(width=0.1, outlier.color=default.color.terciary,
               fill=default.color.secondary, outlier.alpha=0.75) +
  coord_flip() + xlab('') + ylab('') + title.centered +
  ggtitle('Distribución de los ingresos por estado civil')
```

Distribución de los ingresos por estado civil

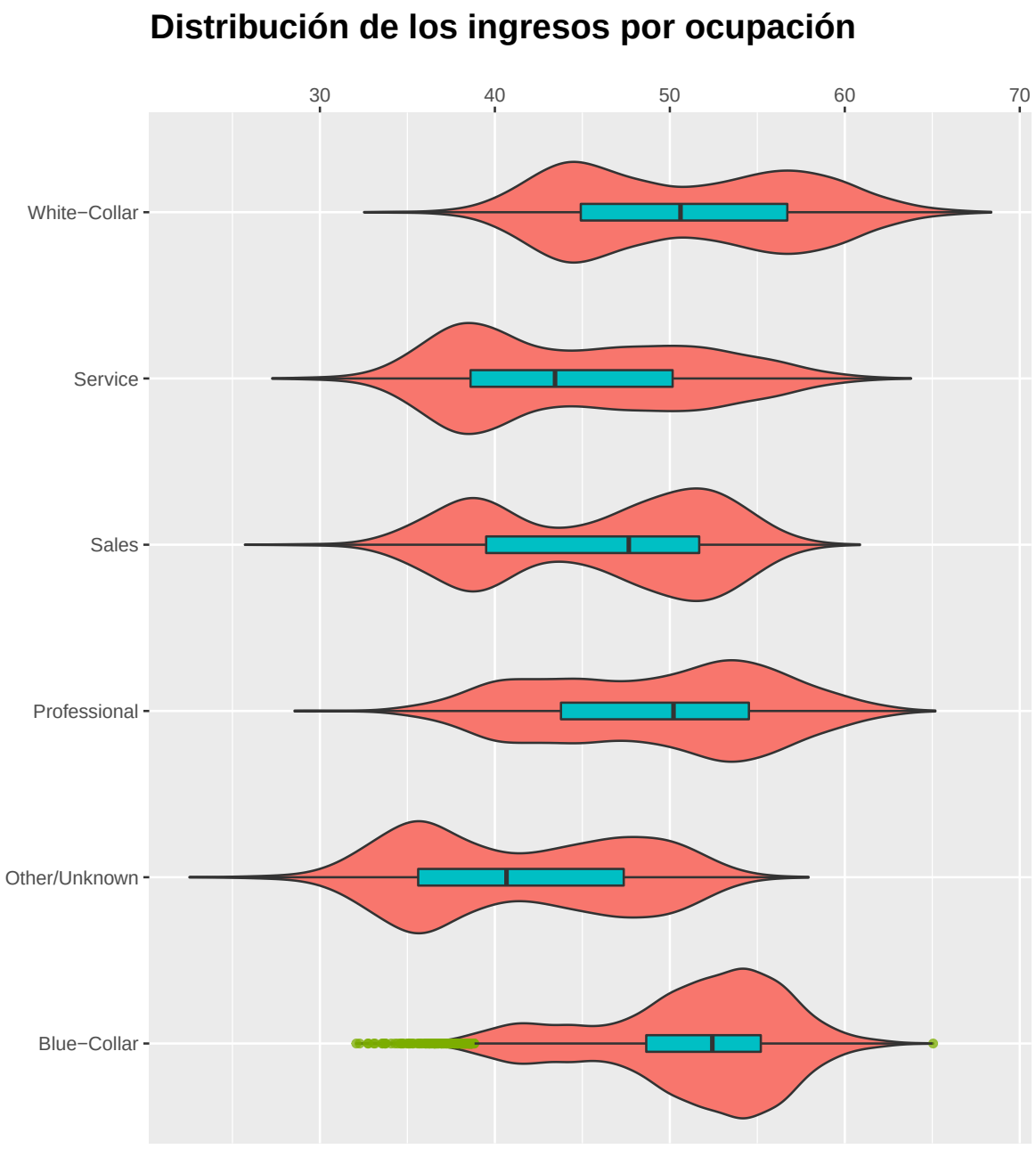


Las personas casadas son sin lugar a duda las que cobran más de media. Los divorciados y solteros perciben menos ingresos, casi 10k euros menos de media; y las personas separadas o viudas obtienen alrededor de 15k euros menos que los casados.

```
ggplot(data=adult, mapping=aes(x=occupation, y=income)) +
  geom_violin(fill=default.color.main) +
  scale_y_continuous(position='right') +
  geom_boxplot(width=0.1, outlier.color=default.color.terciary,
               fill=default.color.secondary, outlier.alpha=0.75) +
  coord_flip() + xlab('') + ylab('') + title.centered +
```

Patricia Lázaro Tello
plazarotello@uoc.edu

```
  ggtitle('Distribución de los ingresos por ocupación')
```



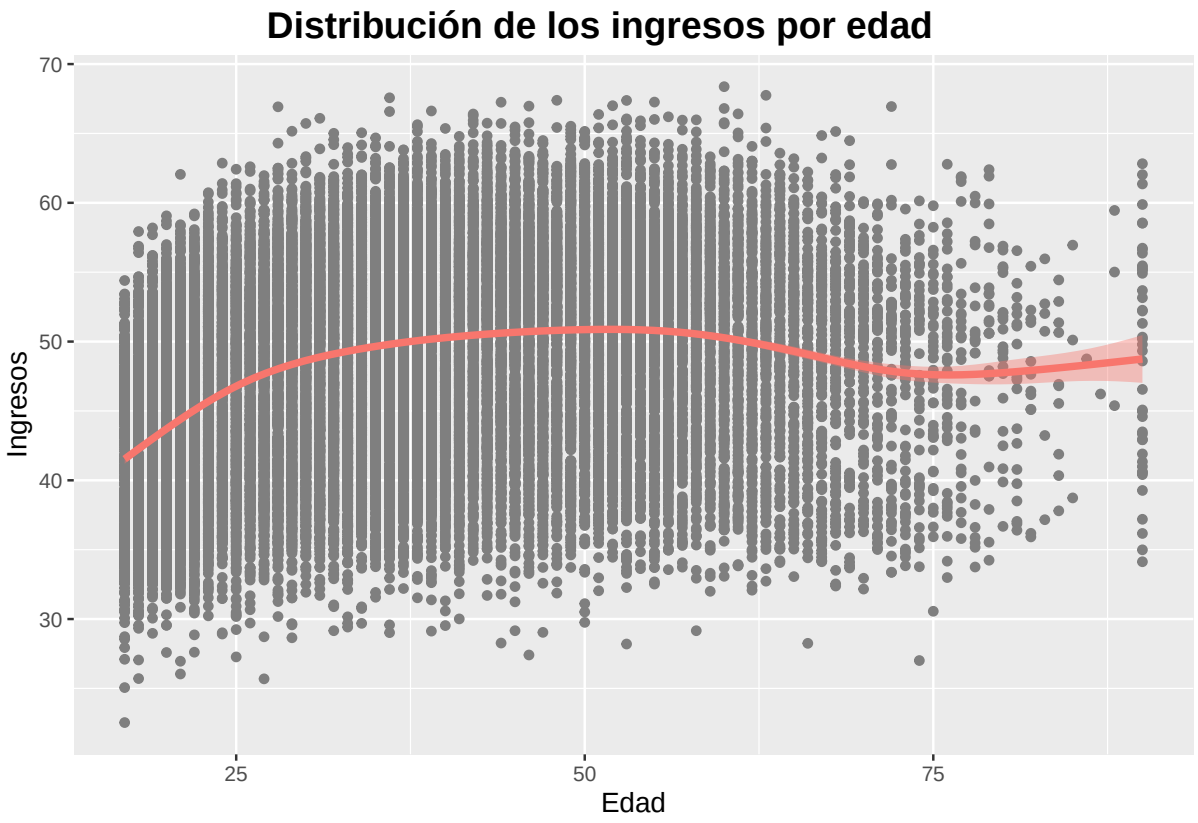
De media, los trabajadores de cuello azul (*blue collar*, hace referencia a obreros y los que hacen trabajos manuales) son los que más cobran, pero también son los que perciben salarios más dispares.

Los trabajadores de cuello blanco (*white collar*, hace referencia a secretarios, adminis-

trativos, etc) y los profesionales son los siguientes que más ingresos perciben, siendo los trabajadores de cuello blanco los que perciben salarios más dispares de los dos.

Los empleados de ventas obtienen ingresos dispares, mientras que los empleados de servicio cobran salarios generalmente más bajos que su propia media. Las observaciones con trabajo desconocido son las que menos cobran.

```
ggplot(data=adult, mapping=aes(x=age, y=income)) +
  xlab('Edad') + ylab('Ingresos') + geom_point(color='gray50') +
  geom_smooth(fill=default.color.main, color=default.color.main, size=1.5) +
  ggtitle('Distribución de los ingresos por edad') + title.centered
```



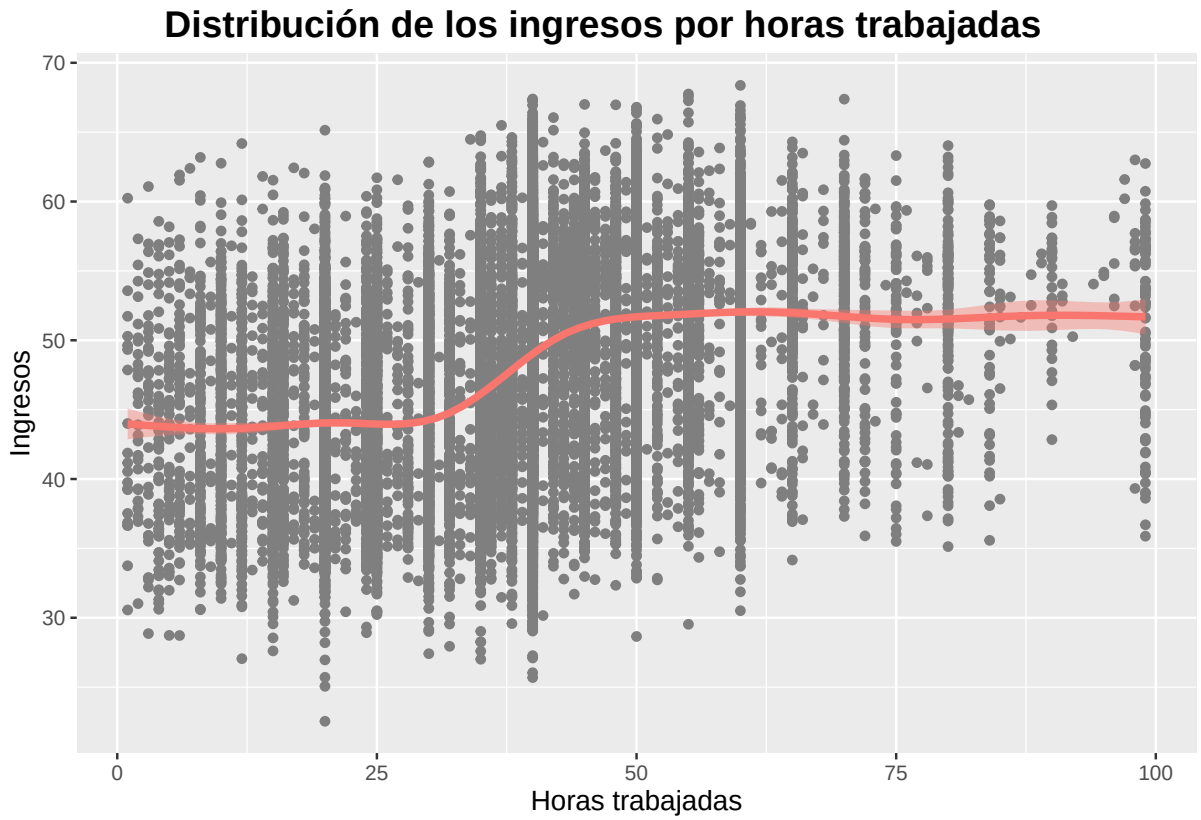
Se observa que hasta los 55 años aproximadamente se confirma la regla de 'a más edad, más salario'. Esto se debe a que, como norma general, la edad supone que la persona lleva ciertos años trabajando y por tanto tiene más experiencia laboral, que es recompensada mediante un salario mayor.

A partir de los 55 años la tendencia baja ligeramente debido a que sobre esa franja de edades comienza el periodo de prejubilaciones, y más adelante el de jubilaciones. Las

Patricia Lázaro Tello
plazarotello@uoc.edu

personas con menores ingresos deben trabajar más años que las que tienen salarios más altos, siendo más fácil por tanto para los que reciben un salario alto el jubilarse y despedirse de la vida laboral.

```
ggplot(data=adult, mapping=aes(x=hours_per_week, y=income)) +
  xlab('Horas trabajadas') + ylab('Ingresos') + geom_point(color='gray50') +
  geom_smooth(fill=default.color.main, color=default.color.main, size=1.5) +
  ggtitle('Distribución de los ingresos por horas trabajadas') + title.centered
```

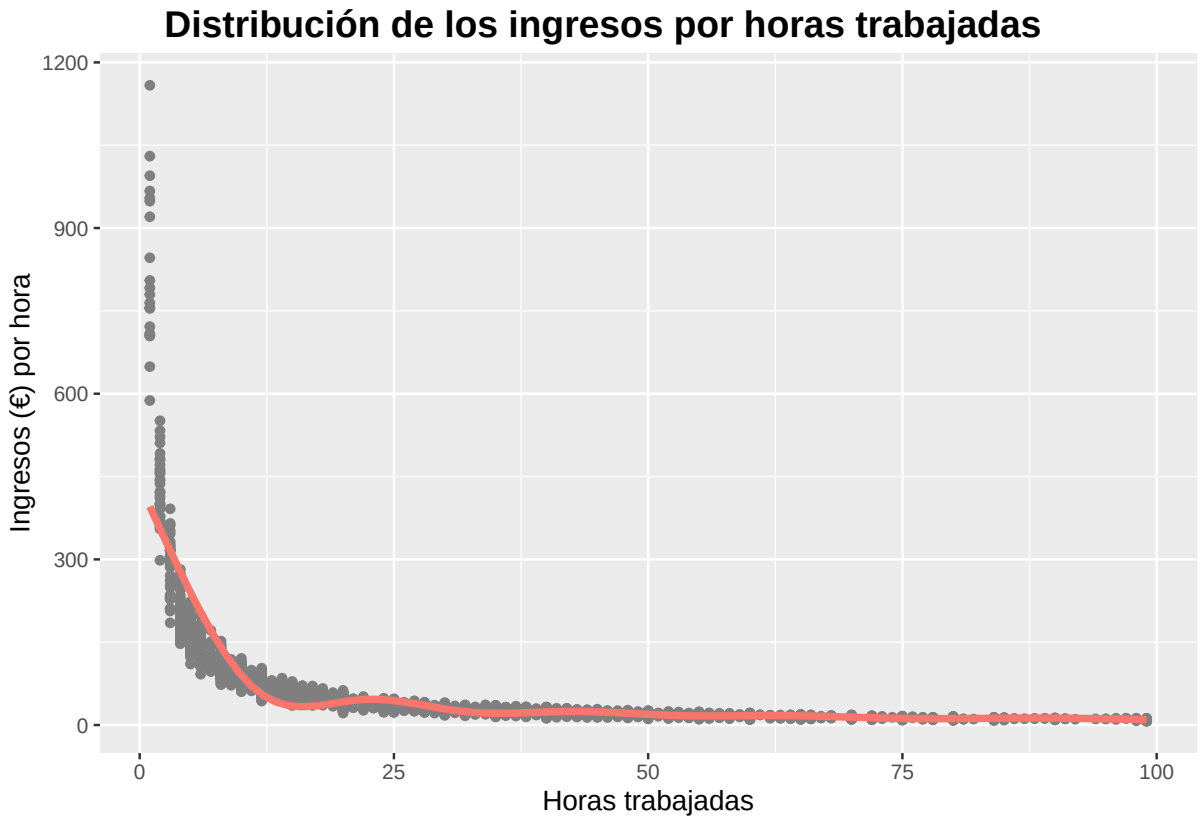


El salario percibido es mayor cuantas más horas de trabajo se invierten. Esta relación era esperable, ya que en general sesuele mantener estable el dinero percibido por hora trabajada y no el dinero percibido por semana. Se procede a comprobar si la asunción de que el dinero percibido por hora trabajada es estable se cumple:

```
ggplot(data=adult, mapping=aes(x=hours_per_week,
                               y=(income*1000)/(hours_per_week*52))) +
  xlab('Horas trabajadas') + ylab('Ingresos (€) por hora') +
  geom_point(color='gray50') +
  geom_smooth(fill=default.color.main, color=default.color.main, size=1.5) +
```

Patricia Lázaro Tello
plazarotello@uoc.edu

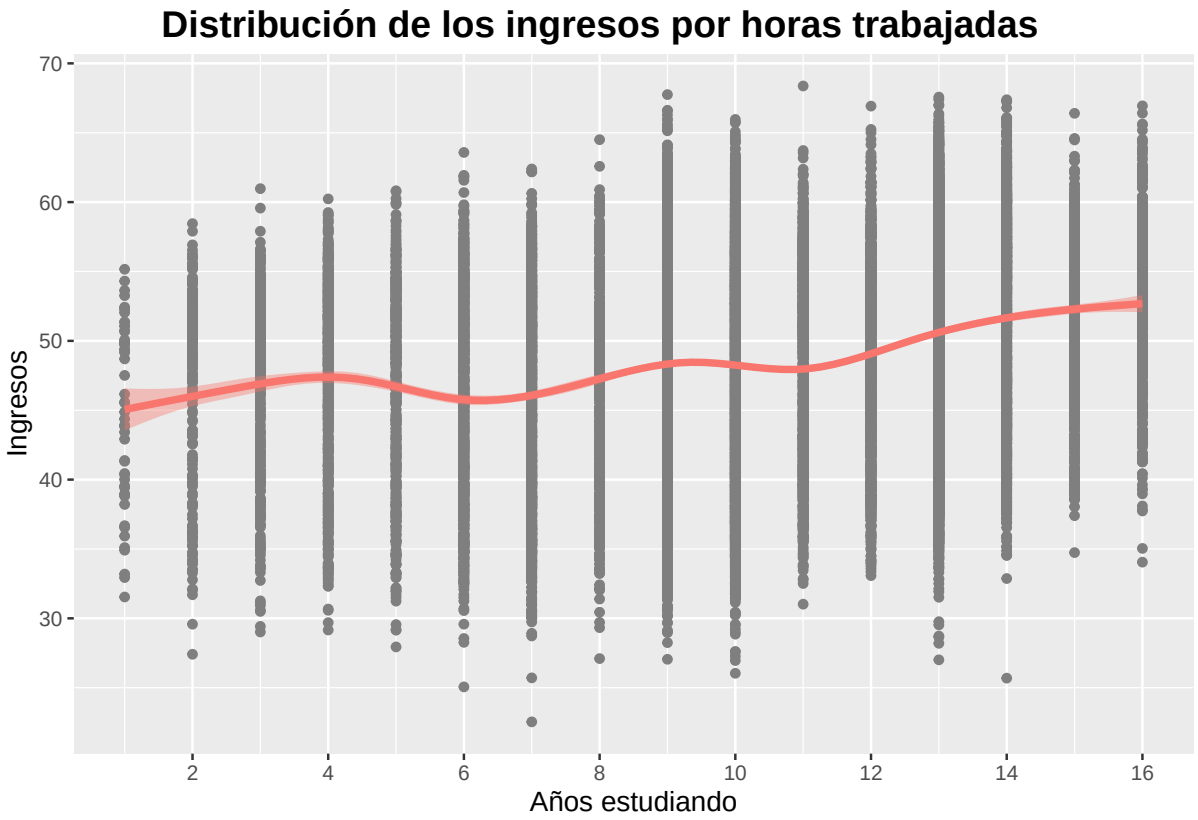
```
  ggtitle('Distribución de los ingresos por horas trabajadas') + title.centered
```



Existe un porcentaje de la población que cobra salarios muy altos y trabaja muy pocas horas semanalmente. Las personas que trabajan menos de 30 horas semanales cobran de media más dinero por hora que las que trabajan a jornada completa. Sobre las 30 horas semanales, los salarios por hora trabajada se vuelven más regulares, en torno a 25-30€/hora. La fórmula se ha calculado como sigue:

$$€/hora = \frac{k\ €/año}{horas/semana * 52\ semanas}$$

```
ggplot(data=adult, mapping=aes(x=education_num, y=income)) +
  xlab('Años estudiando') + ylab('Ingresos') + geom_point(color='gray50') +
  geom_smooth(fill=default.color.main, color=default.color.main, size=1.5) +
  scale_x_continuous(n.breaks=9) + title.centered +
  ggtitle('Distribución de los ingresos por horas trabajadas')
```



Se observa que a más años estudiando más salario percibe la persona. Con 6 y 11 años de estudio, sin embargo, se obtienen menos ingresos que con menos años de estudio.

3 Estadística inferencial

A continuación se plantean varios contrastes de hipótesis a los que hay que dar respuesta utilizando el conjunto de datos propuesto en el trabajo. El *dataset* representa una muestra poblacional; por lo tanto no se conoce la varianza de la población en ninguno de los contrastes de hipótesis.

Para realizar el cálculo manual del contraste de hipótesis se utilizarán las funciones ya creadas en el trabajo `A2: Analítica descriptiva e inferencial`.

Mostrar área de confianza

```
confidence.plot <- function(type){
  ic.x <- seq(-4, 4, by=0.01)
  ic.y <- dnorm(ic.x, mean=0, sd=1)

  ic.xmin <- -2.0; ic.xmax <- 2.0
  ic.ymin <- ic.y[which(ic.x == ic.xmin)]
  ic.ymax <- ic.y[which(ic.x == ic.xmax)]
  ic.factual.x <- ifelse(switch(type, left = ic.x>ic.xmin,
                          two.sided = ic.x>ic.xmin & ic.x<ic.xmax,
                          right = ic.x<ic.xmax), ic.x, NA)

  ic.plot <- ggplot(mapping=aes(x=ic.x, y=ic.y)) +
    geom_area(mapping = aes(x = ic.factual.x), fill = default.color.main,
              alpha=0.75) + geom_line(size=1.05) +
    annotate('text', label='Área de \nconfianza', x=0,
             y=ic.y[which(ic.x==0)]/2, size=5, color='black') +
    no.axis.y + ylab('') + xlab('') + xlim(c(-3.5, 3.5))

  if(type %in% c('two.sided', 'left')){
    ic.plot <- ic.plot + geom_segment(aes(x=ic.xmin, y=0, xend=ic.xmin,
                                          yend=ic.ymin), size=1.05) +
      annotate('text', x=ic.xmin, y=ic.ymin + 0.05, parse=TRUE, size=4,
               label=switch(type, two.sided='frac(~alpha,2)', left='~alpha'))
  }
  if(type %in% c('two.sided', 'right')){
    ic.plot <- ic.plot + geom_segment(aes(x=ic.xmax, y=0, xend=ic.xmax,
                                          yend=ic.ymax), size=1.05) +
      annotate('text', x=ic.xmax, y=ic.ymax + 0.05, parse=TRUE, size=4,
```

```
                label=switch(type, two.sided='1-frac(~alpha,2)',
                             right='1-~alpha'))
  }
  return(ic.plot)
}
```

Cálculo de intervalos de confianza

```
confidence.interval.2.means <- function(nc, mean1, mean2, se1, se2, v, type){
  alpha <- 1.0 - nc
  val <- switch(type, two.sided=alpha/2, right=1-alpha, left=alpha)
  z <- qt(val, df=v, lower.tail=FALSE)

  ic.1 <- (mean1-mean2) - z*sqrt(se1+se2)
  ic.2 <- (mean1-mean2) + z*sqrt(se1+se2)
  if(ic.1 > ic.2){ tmp.ic <- ic.2; ic.2 <- ic.1; ic.1 <- tmp.ic; }

  return(switch(type, two.sided=c(ic.1, ic.2),
                right=c(ic.1, Inf), left=c(-Inf, ic.2)))
}
```

Contraste de hipótesis de igualdad de varianzas

```
variance.equals.2.samples <- function(nc, dist1, dist2){
  alpha <- 1-nc

  n1 <- length(dist1); n2 <- length(dist2)
  sd1 <- sd(dist1); sd2 <- sd(dist2)
  fobs <- sd1^2 / sd2^2

  fcrit.lower <- qf(alpha/2, df1=n1-1, df2=n2-1)
  fcrit.upper <- qf(1-alpha/2, df1=n1-1, df2=n2-1)

  pvalue <- min(pf(fobs, df1=n1-1, df2=n2-1, lower.tail=FALSE),
                pf(fobs, df1=n1-1, df2=n2-1, lower.tail=TRUE))*2

  return(list(fobs=fobs, fcrit=c(fcrit.lower, fcrit.upper), pvalue=pvalue))
}
```

Contraste de hipótesis de dos muestras independientes sobre la media con varianzas desconocidas diferentes

```
means.contrast.2.samples.vars.unknown.dif <- function(x1, x2, dif, nc, type){
  x1.media <- mean(x1); x1.n <- length(x1); x1.sd <- sd(x1);
  x1.se <- x1.sd^2/x1.n

  x2.media <- mean(x2); x2.n <- length(x2); x2.sd <- sd(x2);
  x2.se <- x2.sd^2/x2.n

  alfa <- 1-nc
  t <- (x1.media - x2.media - dif) / sqrt(x1.se + x2.se)
  v <- ceiling((x1.se + x2.se)^2 / ((x1.se)^2/(x1.n-1) + (x2.se)^2/(x2.n-1)))
  pvalue <- pt(t, df=v, lower.tail=switch(type, two.sided=FALSE,
                                          right=FALSE, left=TRUE))
  tcrit <- qt(switch(type, two.sided=alfa/2, right=1-alfa, left=alfa), v)
  ic <- confidence.interval.2.means(nc, x1.media, x2.media, x1.se, x2.se,
                                    v, type)

  return(list(t=t, v=v, p=pvalue, tcrit=tcrit, ic=ic))
}
```

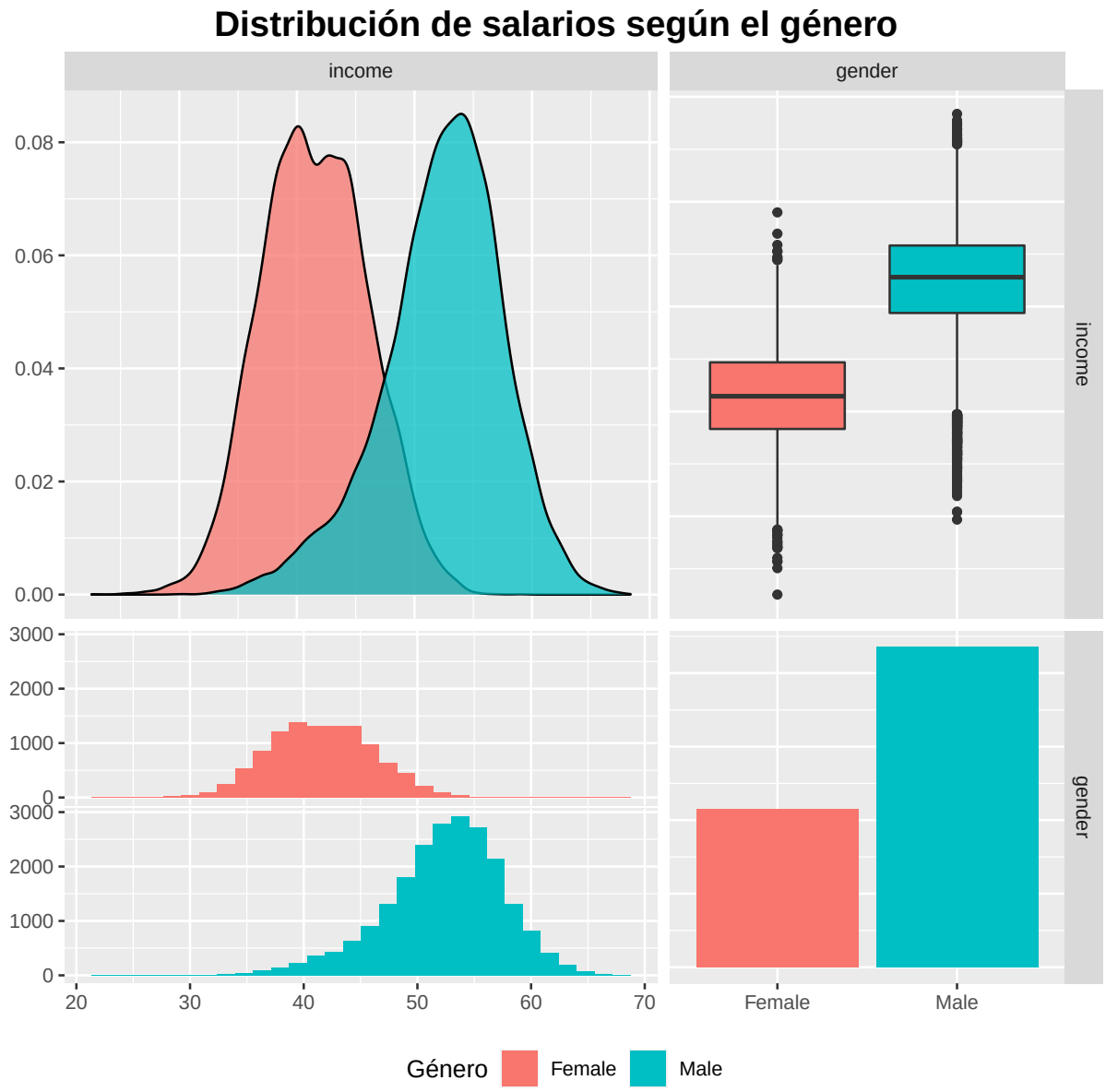
3.1 Diferencia de salario entre hombres y mujeres

La pregunta a la que se ha de responder es '*¿Cobran los hombres más que las mujeres?*'. Se utilizará un nivel de confianza del 95%, o dicho de otra manera, un nivel de significancia $\alpha = 0.05$.

En primer lugar se procede a mostrar los datos:

```
income.female <- adult$income[adult$gender == 'Female']
income.male <- adult$income[adult$gender == 'Male']

GGally::ggpairs(dplyr::select(adult, income, gender),
                mapping=aes(fill=factor(gender)),
                columns=c('income', 'gender'), proportions=c(1.5,1),
                title='Distribución de salarios según el género', legend=4,
                diag=list(continuous = wrap("densityDiag", alpha = 0.75))) +
  scale_colour_manual(name='Género', values=palette) +
```

```
  scale_fill_manual(name='Género', values=palette) +
  theme(legend.position='bottom') + title.centered
```



Existen 10.771 observaciones de mujeres y 21.789 observaciones de hombres en el conjunto de datos. El histograma, el gráfico de densidad y el *boxplot* muestran que los hombres obtienen de media más ingresos que las mujeres, como se había observado en el análisis visual.

A la hora de elegir el contraste de hipótesis que responda a la pregunta, se han de

Patricia Lázaro Tello
plazarotello@uoc.edu

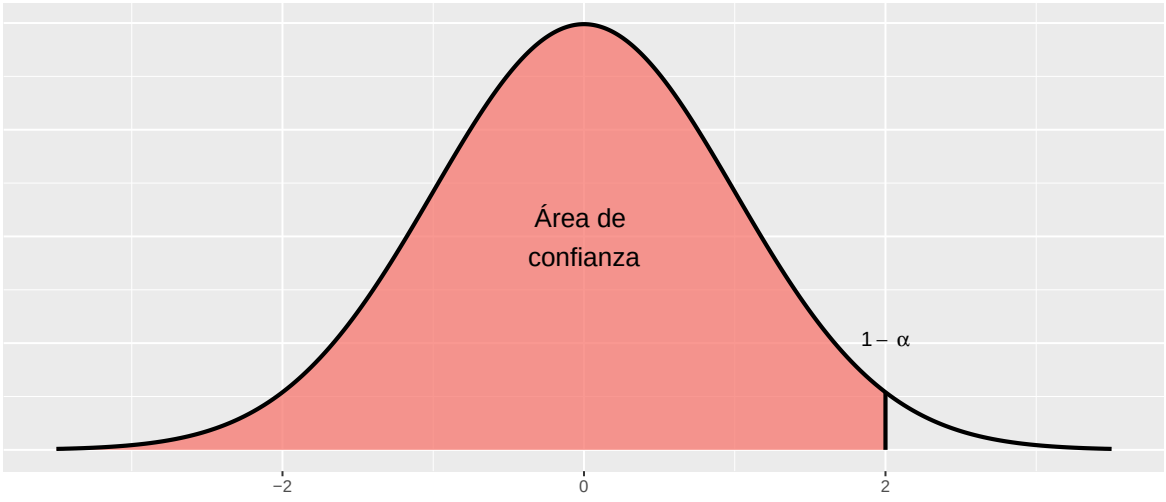
tener en cuenta las siguientes consideraciones:

- Se hace referencia a la variable `income`, pero al ser filtrada por género se obtienen dos variables distintas. Por tanto, se trata de un **contraste de dos muestras**.
- Los ingresos del hombre no dependen de los ingresos de la mujer ni viceversa: son **muestras independientes**.
- Se busca saber si los hombres cobran de media más que las mujeres; por tanto, es un **contraste de hipótesis sobre las medias de las muestras**.
- Se desea saber si los hombres cobran más que las mujeres: esto indica que se trata de un **contraste unilateral por la derecha**.

A continuación se plantea la hipótesis nula y la hipótesis alternativa del **contraste de dos muestras independientes sobre la media con varianzas desconocidas**:

$$H_0 : \mu_{hombre} = \mu_{mujer}$$
$$H_1 : \mu_{hombre} > \mu_{mujer}$$

Por tanto el intervalo de confianza quedará:



Para elegir el estadístico a utilizar, se ha de comprobar si las varianzas de las muestras son iguales o no; es decir, hay que realizar un test de homocedasticidad, que se plantea a continuación:

$$H_0 : s^2_{hombre} = s^2_{mujer}$$
$$H_1 : s^2_{hombre} \neq s^2_{mujer}$$

El test estadístico es:

$$f_{obs} = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1}$$

```
gender.var <- variance.equals.2.samples(0.95, income.male, income.female)
```

No se puede asumir que las varianzas de los ingresos de hombres y mujeres sean iguales, dado que p-value $= 0 \ll \alpha = 0.05$. Analizando $F_{obs} = 1,3$, se observa que la varianza muestral de los hombres es superior a la de las mujeres.

Se comprueba con la función del test estadístico de R:

```
var.test(income.male, income.female, conf.level=0.95)
```

```
##
##  F test to compare two variances
##
## data:  income.male and income.female
## F = 1, num df = 21788, denom df = 10770, p-value <0.0000000000000002
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.3 1.4
## sample estimates:
## ratio of variances
##                1.3
```

Por tanto, para el **contraste de dos muestras independientes sobre la media con varianzas desconocidas *diferentes*** se utilizará el siguiente estadístico:

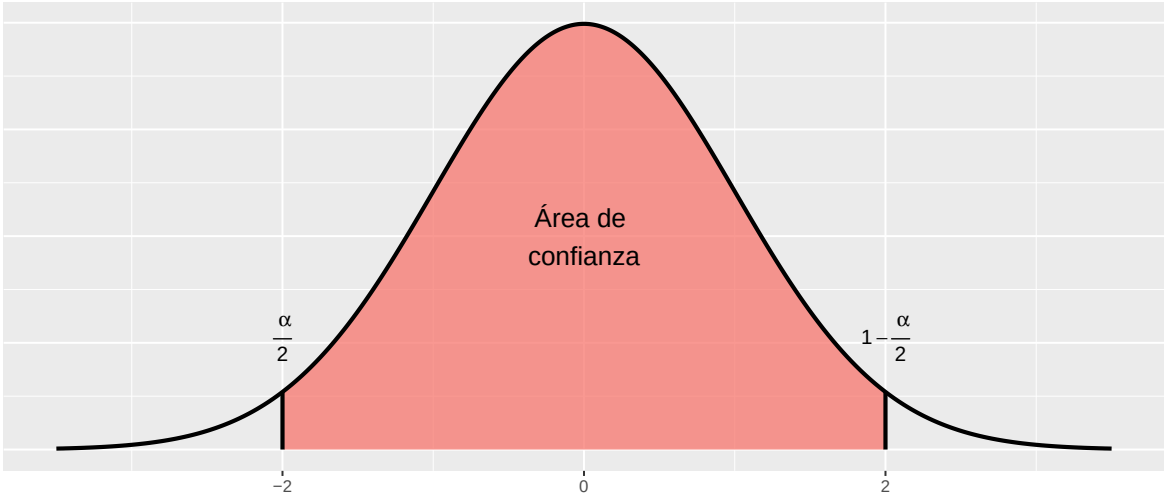
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_v$$

siendo t_v una distribución *t* de Student con v grados de libertad, calculados según:

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{(n_1-1)} + \frac{\left(s_2^2/n_2\right)^2}{(n_2-1)}}$$

Y el intervalo de confianza queda:

$$\left[(\overline{X}_1 - \overline{X}_2) - t_{1-\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\overline{X}_1 - \overline{X}_2) + t_{1-\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$




```
income.medias.iguales <- means.contrast.2.samples.vars.unknown.dif(
  income.male, income.female, 0, 0.95, 'right')
```

De los cálculos anteriores se obtiene que $t_v = 194$ y el valor p es p-value $= 0e + 00 \ll 0.05$; por tanto la hipótesis nula (H_0) se rechaza.

La media del salario de los hombres es mayor que la media del salario de las mujeres con un nivel de confianza del 95%. Analizando el intervalo de confianza, se puede afirmar que, con ese nivel de significancia, los hombres cobran al menos 11k € más que las mujeres.

	Valor
Estadístico	$t_{obs} = 194$
Grados de libertad v	$v = 24.387$
Valor crítico	$t_v = 1,6$

	Valor
Valor p	p-value $= 0e+00$
Intervalo de confianza	$[11, Inf)$

Se comprueban los resultados con los que ofrece la función `t.test` de R:

```
t.test(income.male, income.female, alternative='greater', var.equal=FALSE,
       conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  income.male and income.female
## t = 194, df = 24387, p-value <0.0000000000000002
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##   11 Inf
## sample estimates:
## mean of x mean of y
##        52        42
```

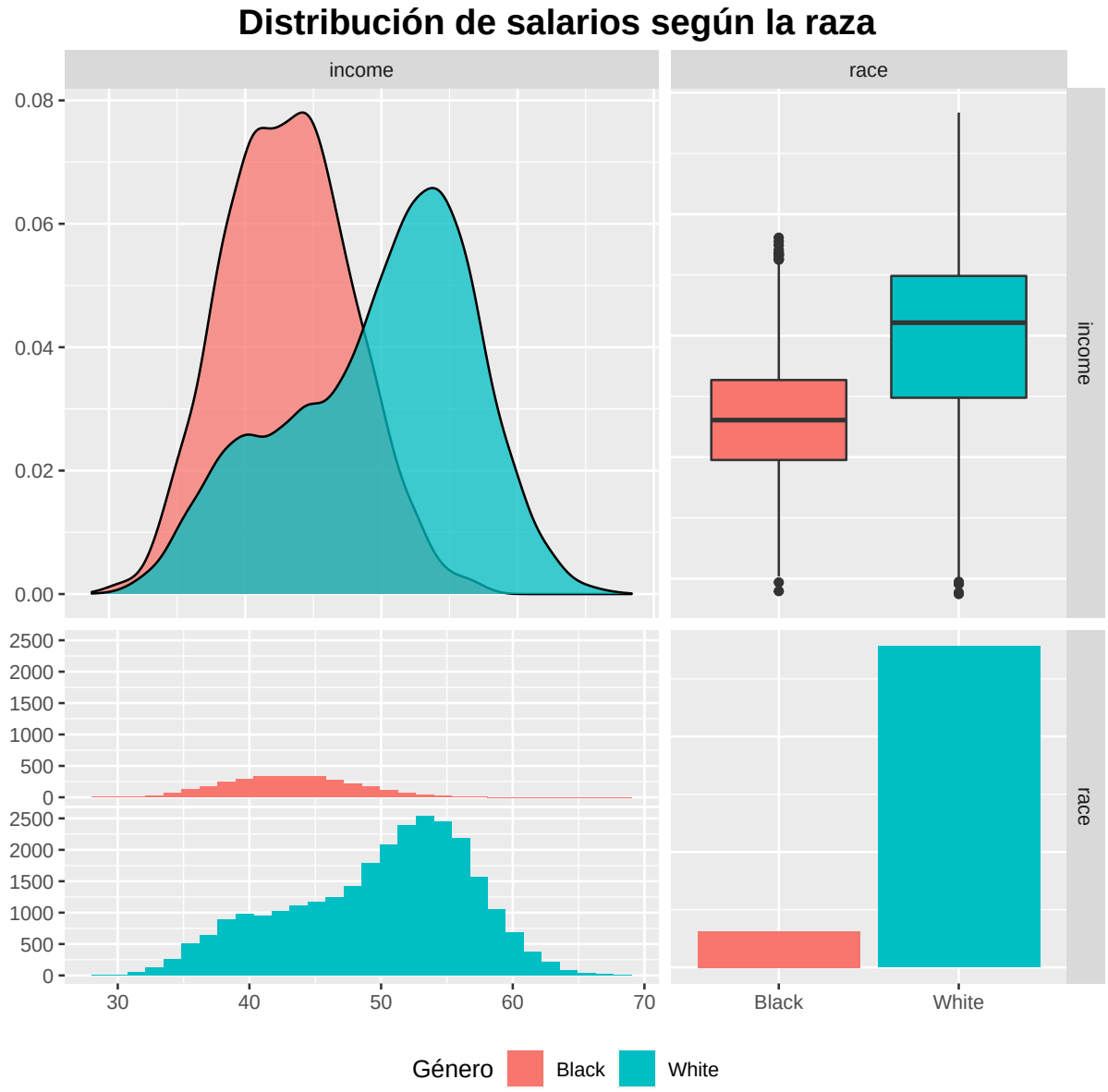
3.2 Diferencia de salario por raza blanca o negra

La pregunta a la que se ha de responder es '*¿Cobra la gente blanca 6450€ más al año que la gente negra?*'. Se utilizará un nivel de confianza del 95%, o dicho de otra manera, un nivel de significancia $\alpha = 0.05$.

En primer lugar se procede a mostrar los datos:

```
income.white <- adult$income[adult$race == 'White']
income.black <- adult$income[adult$race == 'Black']

GGally::ggpairs(dplyr::select(adult, income, race) %>%
                  dplyr::filter(race %in% c('White', 'Black')),
                mapping=aes(fill=factor(race)),
                columns=c('income', 'race'), proportions=c(1.5,1),
                title='Distribución de salarios según la raza', legend=4,
                diag=list(continuous = wrap("densityDiag", alpha = 0.75))) +
  scale_colour_manual(name='Género', values=palette) +
  scale_fill_manual(name='Género', values=palette) +
```

```
theme(legend.position='bottom') + title.centered
```



Existen 27.815 observaciones de personas de raza blanca y 3.124 observaciones de personas de raza negra en el conjunto de datos. El histograma, el gráfico de densidad y el *boxplot* muestran que los blancos obtienen de media más ingresos que los negros, como se había observado en el análisis visual.

A la hora de elegir el contraste de hipótesis que responda a la pregunta, se han de tener en cuenta las siguientes consideraciones:

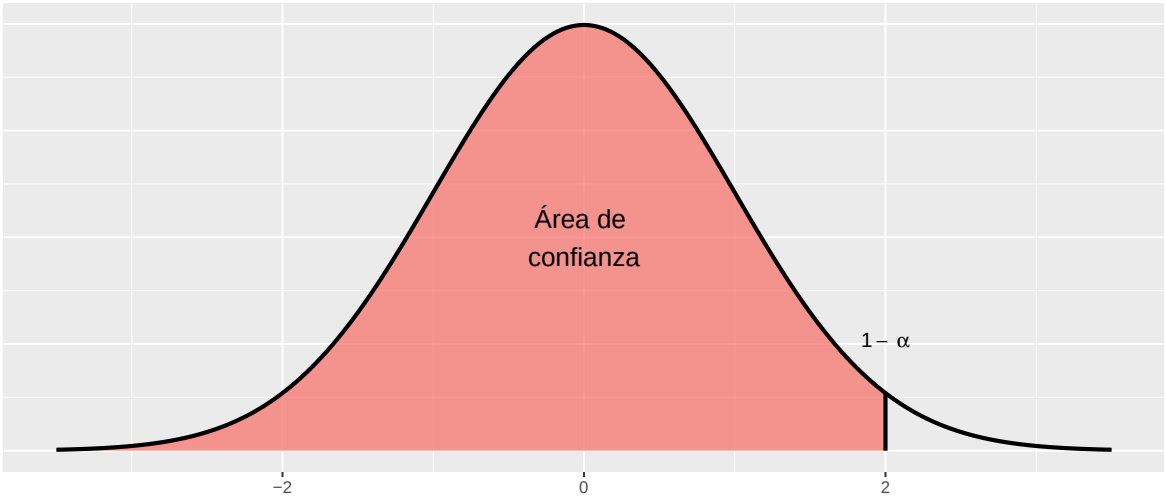
Patricia Lázaro Tello
plazarotello@uoc.edu

- Se hace referencia a la variable `income`, pero al ser filtrada por raza se obtienen dos variables distintas. Por tanto, se trata de un **contraste de dos muestras**.
- Los ingresos de las personas de raza blanca no dependen de los ingresos de las personas de raza negra ni viceversa: son **muestras independientes**.
- Se busca saber si los blancos cobran de media 6450€ más que los negros; por tanto, es un **contraste de hipótesis sobre las medias de las muestras**.
- Se desea saber si los blancos cobran 6450€ más que los negros: esto indica que se trata de un **contraste unilateral por la derecha**.

A continuación se plantea la hipótesis nula y la hipótesis alternativa del **contraste de dos muestras independientes sobre la media con varianzas desconocidas**:

$$H_0 : \mu_{blanco} - \mu_{negro} = 6450$$
$$H_1 : \mu_{blanco} - \mu_{negro} > 6450$$

Por tanto el intervalo de confianza quedará:



Para elegir el estadístico a utilizar, se ha de comprobar si las varianzas de las muestras son iguales o no; es decir, hay que realizar un test de homocedasticidad, que se plantea a continuación:

$$H_0 : s^2_{blanco} = s^2_{negro}$$

$$H_1 : s^2_{blanco} \neq s^2_{negro}$$

El test estadístico es:

$$f_{obs} = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1}$$

```
race.var <- variance.equals.2.samples(0.95, income.white, income.black)
```

No se puede asumir que las varianzas de los ingresos de blancos y negros sean iguales, dado que p-value $= 0 \ll \alpha = 0.05$. Analizando $F_{obs} = 2,1$, se observa que la varianza muestral de los blancos es muy superior a la de los negros.

Se comprueba con la función del test estadístico de R:

```
var.test(income.white, income.black, conf.level=0.95)
```

```
##
##  F test to compare two variances
##
## data:  income.white and income.black
## F = 2, num df = 27814, denom df = 3123, p-value <0.0000000000000002
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.0 2.3
## sample estimates:
## ratio of variances
##                2.2
```

Por tanto, para el **contraste de dos muestras independientes sobre la media con varianzas desconocidas *diferentes*** se utilizará el siguiente estadístico:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_v$$

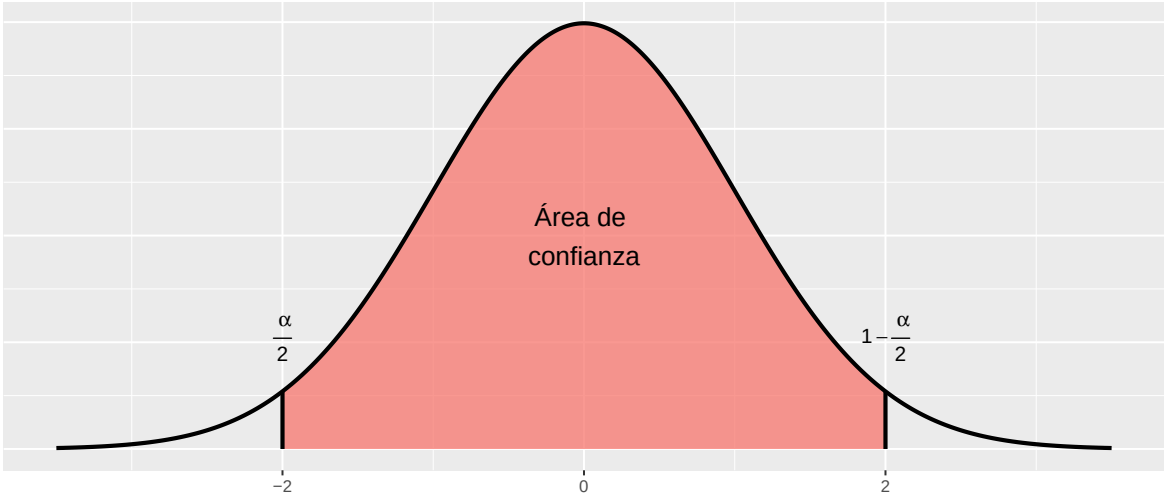
siendo t_v una distribución *t* de Student con v grados de libertad, calculados según:

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{(n_1-1)} + \frac{\left(s_2^2/n_2\right)^2}{(n_2-1)}}$$

Patricia Lázaro Tello
plazarotello@uoc.edu

Y el intervalo de confianza queda:

$$\left[(\overline{X}_1 - \overline{X}_2) - t_{1-\alpha/2,\upsilon}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\overline{X}_1 - \overline{X}_2) + t_{1-\alpha/2,\upsilon}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$




```
income.raza.medias.iguales <- means.contrast.2.samples.vars.unknown.dif(
  income.white, income.black, 6.450, 0.95, 'right')
```

De los cálculos anteriores se obtiene que $t_\upsilon = 2$ y el valor p es p-value $= 2.1e{-}02 < 0.05$; por tanto la hipótesis nula (H_0) se rechaza.

La media del salario de los blancos es 6450€ mayor que la media del salario de los negros con un nivel de confianza del 95%. Analizando el intervalo de confianza, se puede afirmar que, con ese nivel de significancia, los blancos cobran al menos 6,5k € más que los negross.

	Valor
Estadístico	$t_{obs} = 2$
Grados de libertad υ	$\upsilon = 4.784$
Valor crítico	$t_\upsilon = 1,6$
Valor p	p-value $= 2.1e-02$
Intervalo de confianza	$[6,5, Inf)$

Se comprueban los resultados con los que ofrece la función `t.test` de R:

Patricia Lázaro Tello
plazarotello@uoc.edu

```
t.test(income.white, income.black, alternative='greater', mu=6.45,
       var.equal=FALSE, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  income.white and income.black
## t = 2, df = 4783, p-value = 0.02
## alternative hypothesis: true difference in means is greater than 6.4
## 95 percent confidence interval:
##  6.5 Inf
## sample estimates:
## mean of x mean of y
##        50        43
```

Patricia Lázaro Tello
plazarotello@uoc.edu

4 Modelo de regresión lineal

A continuación se procede a crear modelos de regresión lineal para estimar los ingresos que tendrá una persona dada su edad, educación, horas trabajadas por semana y género. Luego, se añadirá al modelo la variable de raza, y se comprobará si la inclusión de este atributo ha supuesto una mejora en el modelo.

4.1 Creación de modelos

Se van a estimar dos modelos, para los que se utilizará la siguiente notación:

$$\begin{aligned} modelo_1: \ Ingresos =& \beta_0 + \beta_1\, edad + \beta_2\, educacion + \\ & \beta_3\, horas + \beta_4\, genero \\ modelo_2: \ Ingresos =& \beta_0 + \beta_1\, edad + \beta_2\, educacion + \\ & \beta_3\, horas + \beta_4\, genero + \beta_5\, raza \end{aligned}$$

Algunas de estas variables se han de transformar en factores para ser procesadas adecuadamente. Se toman como niveles de referencia el hombre en la variable género y la raza blanca en la variable raza.

```
adult.factors <- adult %>%
  dplyr::mutate(gender=relevel(as.factor(gender), ref='Male'),
                race=relevel(as.factor(race), ref='White'))
```

Se procede a crear los dos modelos explicados anteriormente e interpretar sus coeficientes:

```
lm.1 <- lm(income ~ age + education_num + hours_per_week + gender,
           data=adult.factors)
summary(lm.1)
```

```
##
## Call:
## lm(formula = income ~ age + education_num + hours_per_week +
##     gender, data = adult.factors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.432  -2.799   0.286   3.114  17.072
##
## Coefficients:
```

```
##                 Estimate Std. Error t value            Pr(>|t|)
## (Intercept)     41.42418    0.14416   287.3 <0.0000000000000002 ***
## age              0.08251    0.00186    44.4 <0.0000000000000002 ***
## education_num    0.44682    0.00991    45.1 <0.0000000000000002 ***
## hours_per_week   0.07383    0.00212    34.8 <0.0000000000000002 ***
## genderFemale   -10.10840    0.05521  -183.1 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.5 on 32555 degrees of freedom
## Multiple R-squared:  0.589,  Adjusted R-squared:  0.589
## F-statistic: 1.17e+04 on 4 and 32555 DF,  p-value: <0.0000000000000002
```

La regresión lineal es estadísticamente significativa con un nivel de confianza del 95%.

Todas las variables escogidas en el modelo 1 son estadísticamente significativas. Analizando sus coeficientes, se observa que la edad, los años de educación y las horas que trabaja la persona a la semana afectan de forma positiva al salario; es decir, cuanto más mayor es la persona, o cuantos más años ha estudiado, o cuantas más horas trabaja a la semana, sus ingresos serán mayores.

De las variables continuas explicadas, el nivel de estudios es la que más afecta a los ingresos que percibe una persona. La edad afecta ligeramente más que las horas trabajadas; por tanto, esta última variable es la que menos variabilidad genera en el salario.

Respecto al género, al tratarse de una variable categórica se ha de analizar cada valor que puede tomar respecto al nivel de referencia, el hombre. En este caso, si es una mujer cobrará 10k euros menos.

```
lm.2 <- lm(income ~ age + education_num + hours_per_week + gender + race,
           data=adult.factors)
summary(lm.2)
```

```
##
## Call:
## lm(formula = income ~ age + education_num + hours_per_week +
##     gender + race, data = adult.factors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.264  -2.817   0.008   2.775  16.073
##
## Coefficients:
```

```
##                          Estimate Std. Error t value            Pr(>|t|)
## (Intercept)              42.64574    0.13080   326.0 <0.0000000000000002 ***
## age                       0.07825    0.00167    46.8 <0.0000000000000002 ***
## education_num             0.41904    0.00897    46.7 <0.0000000000000002 ***
## hours_per_week            0.07127    0.00191    37.3 <0.0000000000000002 ***
## genderFemale             -9.78094    0.04998  -195.7 <0.0000000000000002 ***
## raceAmer-Indian-Eskimo   -6.68149    0.23349   -28.6 <0.0000000000000002 ***
## raceAsian-Pac-Islander   -6.89601    0.12954   -53.2 <0.0000000000000002 ***
## raceBlack                -4.29917    0.07797   -55.1 <0.0000000000000002 ***
## raceOther               -10.24416    0.25023   -40.9 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.1 on 32551 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.667
## F-statistic: 8.17e+03 on 8 and 32551 DF,  p-value: <0.0000000000000002
```

La regresión lineal es estadísticamente significativa con un nivel de confianza del 95%.

Todos los atributos escogidos en el modelo 2 son estadísticamente significativos. Al analizar sus coeficientes y compararlos con los del modelo 1, se observa que la inclusión de la variable raza hace que el resto de las variables afecten algo menos a los ingresos percibidos.

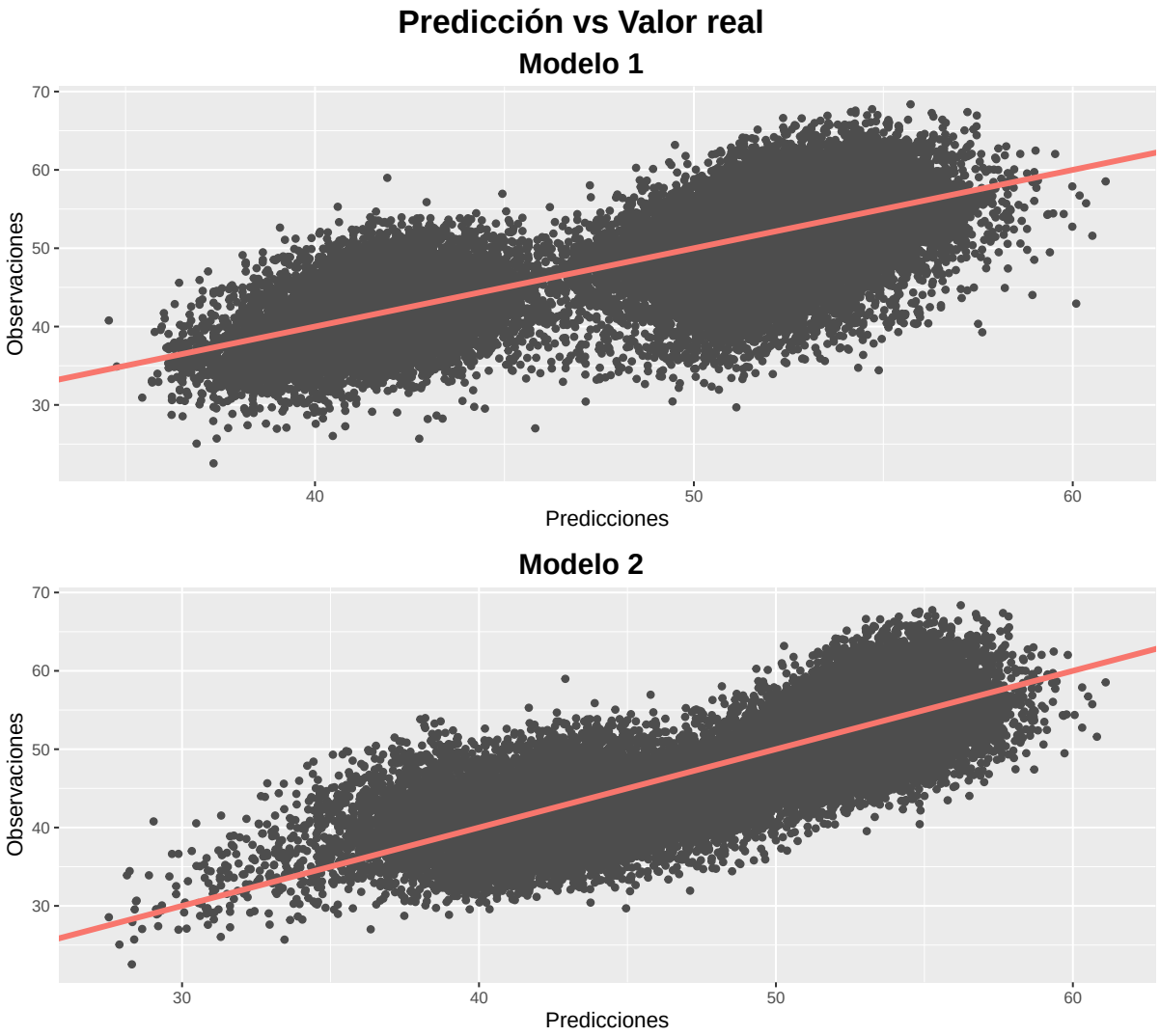
Respecto a la raza, tomando como nivel de referencia la raza blanca, se observa que esta es la más beneficiada en ingresos. Las personas de raza negra cobran 4k euros menos que los blancos, mientras que los indios y asiáticos cobran casi 7k euros menos que los blancos. La raza más perjudicada es la raza minoritaria (otras razas), que percibe más de 10k euros menos que los blancos.

Se quiere comparar los dos modelos creados: el modelo 1 hace referencia al modelo reducido, mientras que el modelo 2 será el modelo completo. Para saber si un modelo es mejor que otro, se han de comparar los valores de R^2 ajustado, siendo un valor más alto un modelo mejor ajustado.

En este caso, $R^2_{M1} = 0.589 < R^2_{M2} = 0.667$; por tanto el modelo completo o extendido se ajusta mejor a los datos que el modelo reducido. El modelo 2 explica un 13% más de la variación de los ingresos que el modelo 1.

$$\hat{ingresos} = 42.6456 + 0.078 * edad + 0.419 * educacion + 0.071 * horas + \left\{ \begin{array}{ll} 0 & \text{si es hombre} \\ -9.781 & \text{si es mujer} \end{array} \right\} + \left\{ \begin{array}{ll} 0 & \text{si es blanco} \\ -6.681 & \text{si es indio} \\ -6.89 & \text{si es asiático} \\ -4.299 & \text{si es negro} \\ -10.244 & \text{si es de otra raza} \end{array} \right.$$

```
m1.pred <- predict(lm.1, adult.factors)
m2.pred <- predict(lm.2, adult.factors)

ggarrange(nrow=2, ncol=1, align='hv',
  ggplot(adult.factors, aes(x=m1.pred, y=income)) +
    geom_point(color='gray30') +
    geom_abline(intercept=0, slope=1, size=1.5, color=default.color.main) +
    labs(x='Predicciones', y='Observaciones') +
    ggtitle('Modelo 1') + title.centered,
  ggplot(adult.factors, aes(x=m2.pred, y=income)) +
    geom_point(color='gray30') +
    geom_abline(intercept=0, slope=1, size=1.5, color=default.color.main) +
    labs(x='Predicciones', y='Observaciones') +
    ggtitle('Modelo 2') + title.centered) %>%
  annotate_figure(., top=text_grob('Predicción vs Valor real',
                                   face='bold', size=18))
```



Se observa que los valores de las predicciones se ajustan mejor en el segundo modelo (modelo 2, modelo completo o modelo extendido).

4.2 Análisis de residuos

Se procede a analizar la calidad del ajuste del modelo 2. En primer lugar se va a observar la distribución de los residuos en el resumen del objeto:

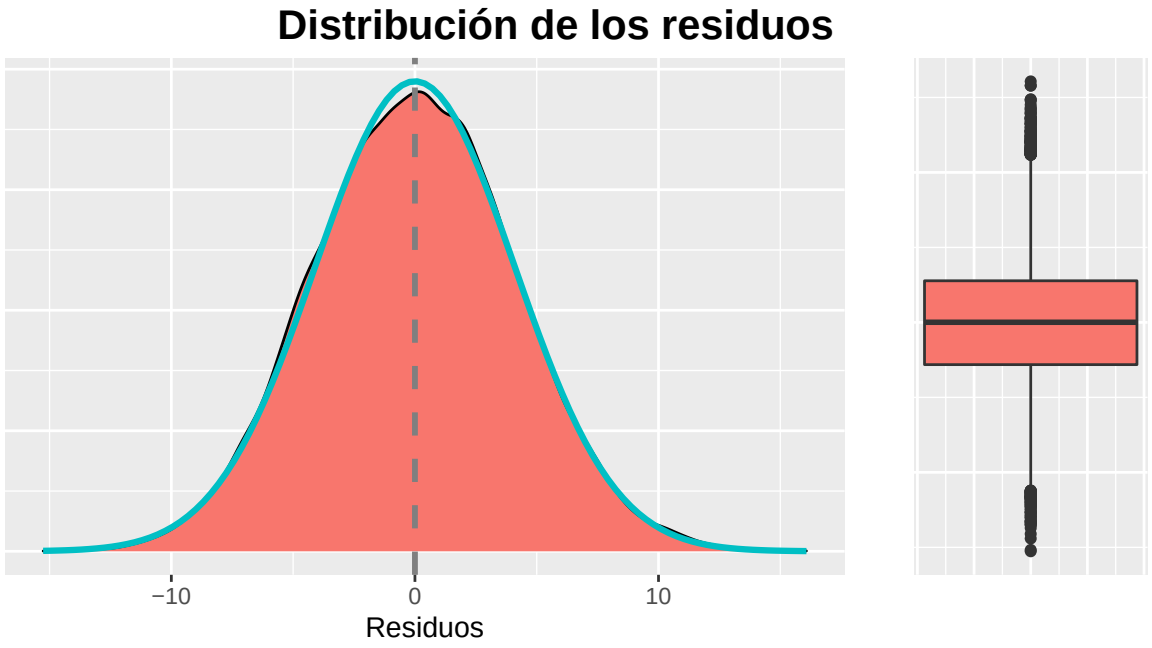
```
summary(lm.2)
```

```
##
## Call:
```

```
## lm(formula = income ~ age + education_num + hours_per_week +
##     gender + race, data = adult.factors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.264  -2.817   0.008   2.775  16.073
##
## Coefficients:
##                          Estimate Std. Error t value            Pr(>|t|)
## (Intercept)              42.64574    0.13080   326.0 <0.0000000000000002 ***
## age                       0.07825    0.00167    46.8 <0.0000000000000002 ***
## education_num             0.41904    0.00897    46.7 <0.0000000000000002 ***
## hours_per_week            0.07127    0.00191    37.3 <0.0000000000000002 ***
## genderFemale             -9.78094    0.04998  -195.7 <0.0000000000000002 ***
## raceAmer-Indian-Eskimo   -6.68149    0.23349   -28.6 <0.0000000000000002 ***
## raceAsian-Pac-Islander   -6.89601    0.12954   -53.2 <0.0000000000000002 ***
## raceBlack                -4.29917    0.07797   -55.1 <0.0000000000000002 ***
## raceOther               -10.24416    0.25023   -40.9 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.1 on 32551 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.667
## F-statistic: 8.17e+03 on 8 and 32551 DF,  p-value: <0.0000000000000002
```

Los residuos siguen una distribución centrada en 0 y a priori están normalmente distribuidos, alcanzando unos valores mínimo y máximo con valor absoluto similar. Se procede a mostrar la distribución de los residuos y comprovar su normalidad:

```
lm.2.fortified <- fortify(lm.2)
ggarrange(ncol=2, nrow=1, widths=c(3,1), align='hv',
          ggplot(lm.2.fortified, aes(x=.resid)) +
            geom_density(mapping=aes(y=..density..), fill=default.color.main) +
            geom_vline(xintercept=mean(lm.2.fortified$.stdresid), size=1.05,
                       linetype='dashed', color='gray50') +
            stat_function(fun=dnorm, args=c(mean=mean(lm.2.fortified$.resid),
                                            sd=sd(lm.2.fortified$.resid)),
                          color=default.color.secondary, size=1.15) +
            no.axis.y + xlab('Residuos') + ylab(''),
          ggplot(lm.2.fortified, aes(x=.resid)) + coord_flip() +
            geom_boxplot(fill=default.color.main) + no.axis.x + no.axis.y +
            xlab('') + ylab('')) +
  title.centered + ggtitle('Distribución de los residuos')
```




```
lillie.test(lm.2.fortified$.resid)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  lm.2.fortified$.resid
## D = 0.006, p-value = 0.02
```

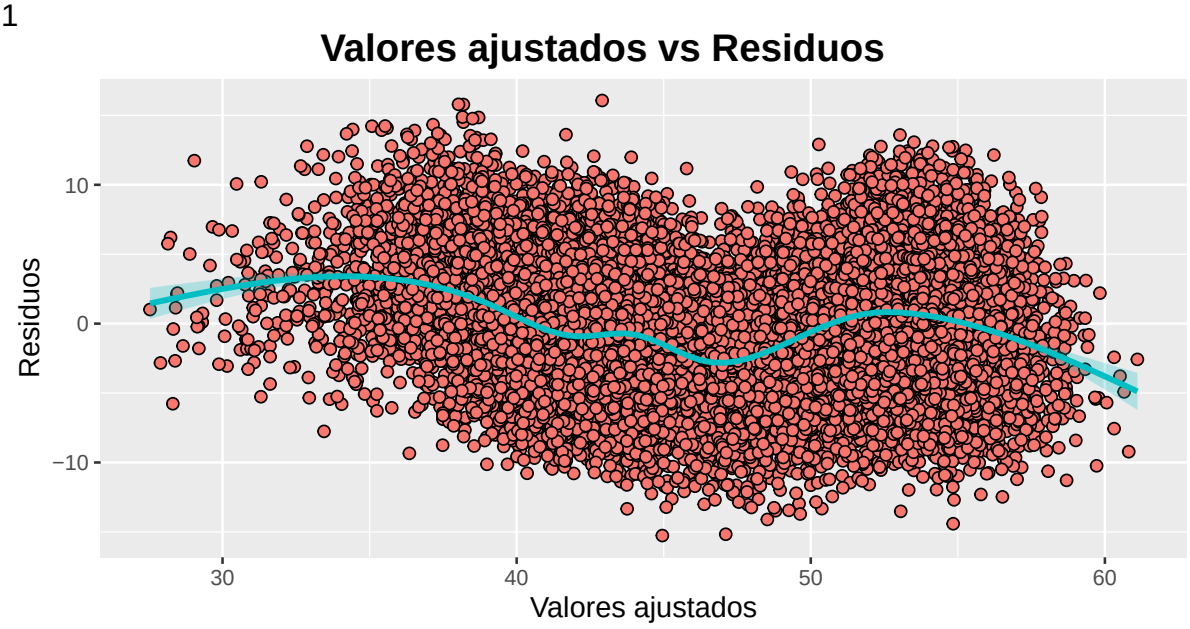
Aunque a primera vista los residuos parecen seguir una distribución normal, al realizar el test de Kolmogorov-Smirnov se observa que no es así: el p-value $< \alpha = 0.05$, por tanto con un nivel de significancia del 95% se puede afirmar que **los residuos no siguen una distribución normal**.

Se procede a comprobar la **homocedasticidad** de los residuos, es decir, si tienen varianza constante:

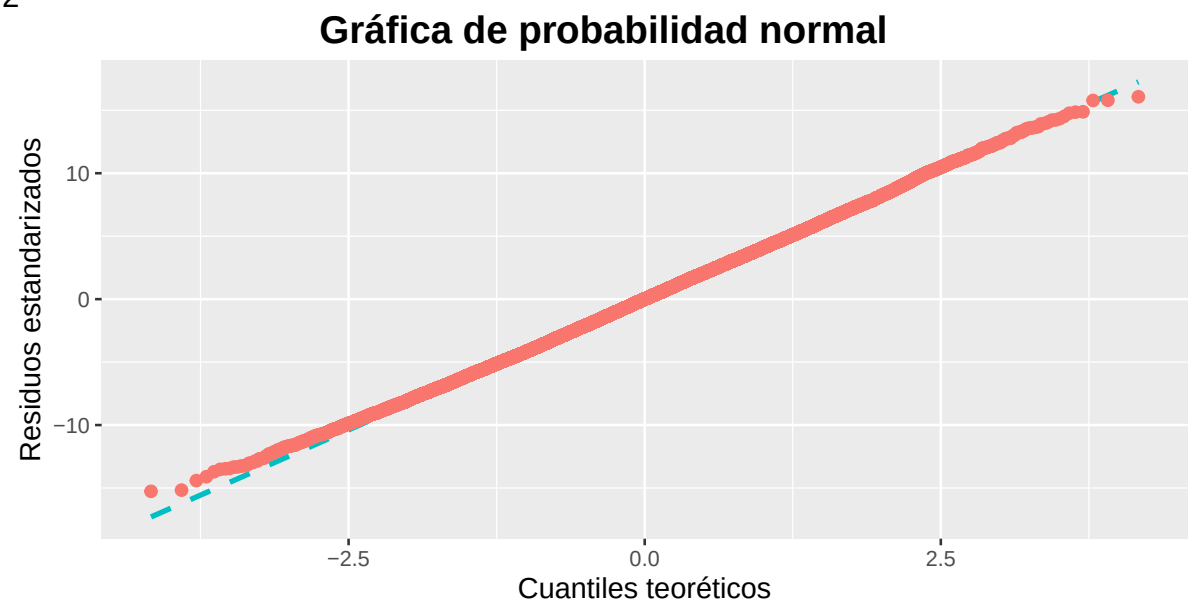
```
ggarrange(nrow=2, ncol=1,
          ggplot(lm.2.fortified, aes(x=.fitted, y=.resid)) +
            geom_point(size=2, shape=21, fill=default.color.main) +
            geom_smooth(color=default.color.secondary, size=1.15,
                        fill=default.color.secondary, alpha=0.25) +
            title.centered + ggtitle('Valores ajustados vs Residuos') +
            xlab('Valores ajustados') + ylab('Residuos') + labs(tag='1'),
          ggplot(lm.2.fortified, aes(sample=.resid)) +
            geom_qq_line(linetype='dashed', size=1.05,
```

```
                    color=default.color.secondary) +
          geom_qq(size=2, color=default.color.main) + title.centered +
          ggtitle('Gráfica de probabilidad normal') + labs(tag='2') +
          xlab('Cuantiles teoréticos') + ylab('Residuos estandarizados'))
```

1



2



En la gráfica QQ se observa que las colas de los residuos estandarizados son más

pesadas que la de la distribución normal. En el gráfico comparativo de valores ajustados frente a los residuos, se observa un patrón reconocible: los datos hacen una curva parecida a la de una sonrisa. Se puede afirmar que la **varianza del error no es constante**.

Además, como se había observado en el resumen del modelo al crearlo, $R^2_{M2} = 0.667$, lo que quiere decir que el modelo solo explica el 66.7% de la varianza observada.

Por tales razones, se determina que **el modelo no termina de ajustar bien a los datos**. Además, **no se cumplen las suposiciones de homocedasticidad y normalidad de residuos**.

4.3 Predicción

Se desea predecir los ingresos que percibiría una mujer negra de 24 años que trabaja 40 horas a la semana y ha estudiado durante 4 años:

```
lm.2.pred.in <- data.frame(age=24, education_num=4, gender='Female',
                           race='Black', hours_per_week=40)
lm.2.pred.out <- predict(lm.2, lm.2.pred.in, interval='confidence', level=0.95)
```

Una persona con esas características cobraría 35k euros al año. Con un nivel de confianza del 95%, se puede afirmar que los ingresos percibidos se encontrarían en el rango $[35, 35]$.

5 Regresión logística

Se desea predecir la **probabilidad de tener unos ingresos menores de 50k euros**. Para crear el modelo de regresión logística, se utilizarán todas las variables del conjunto de datos, que se dividirá en un conjunto de entrenamiento y test para poder evaluar la precisión del modelo más objetivamente.

En primer lugar, se procede a modificar el conjunto de datos para transformar las cadenas de texto a factores y eliminar la variable `income`, que no se necesitará en este modelo:

```
adult.logi <- adult %>% dplyr::select(-income) %>%
  dplyr::mutate(workclass = as.factor(workclass),
                marital_status=as.factor(marital_status),
                occupation=as.factor(occupation),
                race=as.factor(race), gender=as.factor(gender))
head(adult.logi, n=3)
```

```
##   age     workclass education_num marital_status   occupation  race
## 1  50 Self-Employed            13        Married White-Collar White
## 2  38       Private             9       Divorced  Blue-Collar White
## 3  53       Private             7        Married  Blue-Collar Black
##   hours_per_week gender Less50
## 1             13   Male  FALSE
## 2             40   Male  FALSE
## 3             40   Male  FALSE
```

Para las variables categóricas se procede a reordenar los niveles de tal forma que el nivel de referencia sea el que tenga más valores:

```
summary(adult.logi$workclass)         # Private
```

```
##    Government Other/Unknown       Private Self-Employed
##          4350          1857         22696          3657
```

```
summary(adult.logi$marital_status)    # Married
```

```
##  Divorced   Married Separated    Single   Widowed
##      4443     15417      1025     10682       993
```

```
summary(adult.logi$occupation)        # Blue-Collar
```

```
##   Blue-Collar Other/Unknown  Professional         Sales       Service
##         10062          1852          4140          3650          5021
##  White-Collar
```

```
##          7835
```

```
summary(adult.logi$race)              # White
```

```
## Amer-Indian-Eskimo Asian-Pac-Islander              Black              Other
##                311               1039               3124                271
##              White
##              27815
```

```
summary(adult.logi$gender)            # Male
```

```
## Female   Male
##  10771  21789
```

```
adult.logi <- adult.logi %>%
  dplyr::mutate(workclass=relevel(workclass, ref='Private'),
                marital_status=relevel(marital_status, ref='Married'),
                occupation=relevel(occupation, ref='Blue-Collar'),
                race=relevel(race, ref='White'),
                gender=relevel(gender, ref='Male'))
```

5.1 Creación de los conjuntos de *train* y *test*

A continuación se procede a dividir el *dataset* en un conjunto de entrenamiento y otro de test. El conjunto de entrenamiento tendrá el 80% de las observaciones, mientras que el conjunto de test tendrá el 20%.

```
set.seed(seed)
sample_train <- sample(1:nrow(adult.logi), replace=FALSE,
                           size=floor(0.8*nrow(adult.logi)))
adult.logi.train <- adult.logi[sample_train,]
adult.logi.test <- adult.logi[-sample_train,]
```

Se muestra el resultado de la selección de entrenamiento y test, comparando sus tamaños con el conjunto de datos completo:

	Train	*Test*	Total
Nº observaciones	26.048	6.512	32.560
% observaciones	80%	20%	100%

5.2 Modelo predictivo

Se procede a crear el modelo de regresión logística. Para su entrenamiento solo se utilizará el conjunto de datos de entrenamiento, reservando el conjunto de datos de test para evaluar la calidad del modelo.

```
logi.model <- glm(Less50 ~ ., data=adult.logi.train, family='binomial')
summary(logi.model)
```

```
##
## Call:
## glm(formula = Less50 ~ ., family = "binomial", data = adult.logi.train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -4.435  -0.192   0.000   0.083   3.099
##
## Coefficients:
##                           Estimate Std. Error z value             Pr(>|z|)
## (Intercept)                0.13098    0.17815    0.74               0.4622
## age                       -0.02578    0.00250  -10.30 <0.0000000000000002 ***
## workclassGovernment       -3.46783    0.11773  -29.46 <0.0000000000000002 ***
## workclassOther/Unknown     2.27840    0.95280    2.39               0.0168 *
## workclassSelf-Employed     2.51059    0.08406   29.87 <0.0000000000000002 ***
## education_num             -0.28073    0.01220  -23.01 <0.0000000000000002 ***
## marital_statusDivorced     2.58624    0.09065   28.53 <0.0000000000000002 ***
## marital_statusSeparated    3.30438    0.17413   18.98 <0.0000000000000002 ***
## marital_statusSingle       3.45941    0.08244   41.96 <0.0000000000000002 ***
## marital_statusWidowed      2.81995    0.22630   12.46 <0.0000000000000002 ***
## occupationOther/Unknown    2.99614    0.95025    3.15               0.0016 **
## occupationProfessional     1.79823    0.10763   16.71 <0.0000000000000002 ***
## occupationSales            2.53575    0.08984   28.23 <0.0000000000000002 ***
## occupationService          2.56114    0.09145   28.01 <0.0000000000000002 ***
## occupationWhite-Collar    -1.73384    0.11214  -15.46 <0.0000000000000002 ***
## raceAmer-Indian-Eskimo     7.77704    0.46562   16.70 <0.0000000000000002 ***
## raceAsian-Pac-Islander     7.79530    0.21343   36.52 <0.0000000000000002 ***
## raceBlack                  6.28974    0.14898   42.22 <0.0000000000000002 ***
## raceOther                 20.90859  170.18323    0.12               0.9022
## hours_per_week            -0.02525    0.00224  -11.28 <0.0000000000000002 ***
## genderFemale               8.44393    0.15843   53.30 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 36095.7  on 26047  degrees of freedom
```

```
## Residual deviance:  9373.8  on 26027  degrees of freedom
## AIC: 9416
##
## Number of Fisher Scoring iterations: 16
```

Todas las variables explicativas utilizadas son estadísticamente significativas con un nivel de significancia del 95%. Los residuos parecen seguir una distribución normal.

El modelo de regresión logística creado tiene como objetivo detectar y predecir qué personas, por sus características, no tienen ingresos de 50k euros o superiores. Se procede a explicar la contribución de las variables:

- **Age**: la relación entre la edad y la probabilidad de percibir ingresos de 50k euros o superiores es inversamente proporcional. Cuanta más edad tenga el sujeto, menos probabilidad hay de cobrar menos de 50k euros.
- **WorkClass**: el tipo de trabajo es una variable categórica; por tanto es necesario analizar cada nivel por separado, teniendo en cuenta que se utiliza el sector privado como nivel de referencia. Todos los niveles de esta variable son significativos.

  A este respecto, los empleados gubernamentales tienen menos probabilidad de percibir ingresos de menos de 50k euros que los empleados del sector privado; mientras que los autónomos y aquellos para los que no se tienen datos de su trabajo (Otros/Desconocido) tienen más probabilidad que los del sector privado.
- **Education Num**: el número de años dedicados al estudio también afecta a la probabilidad de percibir ingresos de menos de 50k euros. Cuantos más años se haya estudiado, menos probabilidades hay de que el sujeto perciba menos de 50k euros de ingresos. Cabe destacar que la influencia de esta variable es superior a la influencia de la edad.
- **Marital Status**: el estado civil de la persona afecta a la probabilidad de percibir ingresos de menos de 50k euros. Las personas casadas (nivel de referencia) son las que más probabilidades tienen de cobrar 50k euros o más, mientras que las personas separadas tienen más posibilidades de percibir menos de 50k euros de ingresos.

  Los sujetos divorciados y viudos son los que, tras los casados, tienen menos probabilidades de percibir menos de 50k euros. Los solteros tienen casi tantas probabilidades de cobrar menos de 50k euros que los separados, siempre respecto a las personas casadas.

- **Occupation**: el trabajo que desempeña la persona es una variable categórica cuyo nivel de referencia es el personal de mantenimiento y trabajos manuales (*blue collar*).

  La probabilidad de cobrar menos de 50k euros es mayor para los sujetos que trabajan en ventas, servicios o son profesionales. Para los que tienen otra ocupación o esta es desconocida, existe la mayor probabilidad de cobrar menos de 50k euros respecto a los trabajadores de cuello azul. Los trabajadores de cuello blanco son los que menos probabilidades tienen de cobrar menos de 50k euros.

- **Race**: la raza también es un factor determinante a la hora de predecir si una persona cobrará menos de 50k euros. En este caso, las personas de raza blanca (nivel de referencia) son las que menos posibilidades tienen de percibir menos de 50k euros. Las personas de raza negra tienen más probabilidades, que aumentan todavía más si se es de raza india o asiática; se obtienen las probabilidades más altas para aquellos sujetos de otra raza o raza desconocida.

- **Hours per Week**: cuantas más horas a la semana se trabaja, menos probabilidad existe de cobrar menos de 50k euros. Esto quiere decir que las personas que trabajan más horas semanales tienen más posibilidades de cobrar 50k euros o más.

- **Gender**: en este caso, siendo el nivel de referencia el hombre, se observa que ser mujer es un factor de riesgo para cobrar menos de 50k euros.

5.3 Calidad del modelo

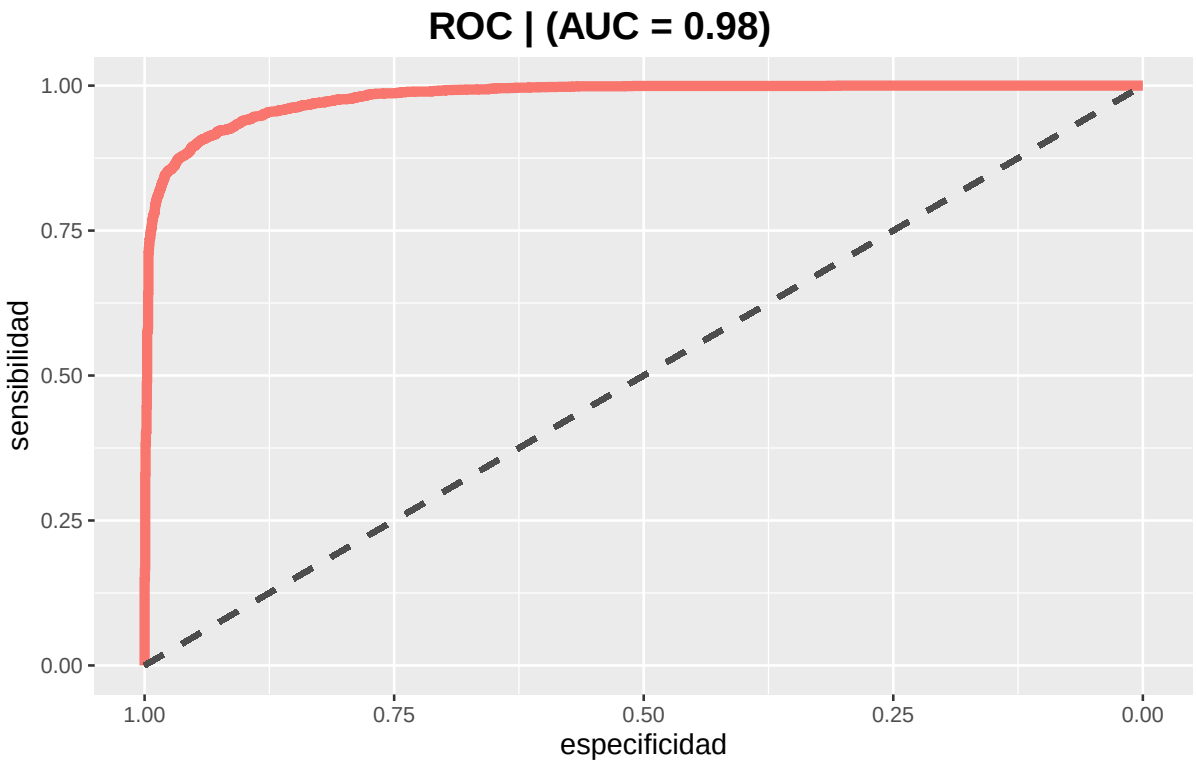
Para calcular la bondad del ajuste del modelo de regresión logística se realizará un análisis ROC del modelo y una matriz de confusión para los datos de test:

```
test.prob <- predict(logi.model, newdata=adult.logi.test, type='response')
conf.matrix <- confusionMatrix(data=as.factor(test.prob>=0.5),
                               reference=as.factor(adult.logi.test$Less50),
                               positive='TRUE')
conf.matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
```

```
##      FALSE  2929  315
##      TRUE    181 3087
##
##                Accuracy : 0.924
##                  95% CI : (0.917, 0.93)
##     No Information Rate : 0.522
##     P-Value [Acc > NIR] : < 0.0000000000000002
##
##                   Kappa : 0.848
##
##  Mcnemar's Test P-Value : 0.00000000235
##
##             Sensitivity : 0.907
##             Specificity : 0.942
##          Pos Pred Value : 0.945
##          Neg Pred Value : 0.903
##              Prevalence : 0.522
##          Detection Rate : 0.474
##    Detection Prevalence : 0.502
##       Balanced Accuracy : 0.925
##
##        'Positive' Class : TRUE
##
```

```
r <- roc(adult.logi.test$Less50, test.prob, data=adult.logi.test)
ggroc(r, color=default.color.main, size=2) +
  geom_segment(aes(x=1, xend=0, y=0, yend=1), color='grey30',
               linetype='dashed', size=1.05) +
  title.centered + ggtitle(paste0('ROC | (AUC = ', round(auc(r),2), ')')) +
  xlab('especificidad') + ylab('sensibilidad')
```



El área bajo la curva es grande ($AUC = 0,98 \geq 0.9$), por tanto se puede afirmar que el modelo discrimina de modo excepcional. Se observan valores muy altos de precisión y *recall*, así como una sensibilidad y especificidad también muy altas. Todo esto sugiere que el modelo está bien ajustado a los datos y, al haber separado los datos en *train* y *test*, se puede comprobar que no se ha producido *overfitting*.

Precisión	*Recall*	Sensibilidad	Especificidad
0,94	0,91	0,91	0,94

5.4 Predicción

Se procede por último a llevar a cabo predicciones sobre el modelo de regresión logística creado. Estas predicciones se van a realizar manualmente, utilizando `predict(...)` para validar los cálculos manuales.

En primer lugar se desea saber la probabilidad de que el salario de un individuo sea menor a 50k euros anuales para un hombre blanco de 20 años, autónomo, con 3 años de estudios, soltero, trabajando en el sector profesional y haciendo 25 horas semanales.

```
logi.coeffs <- logi.model$coefficients
logi.pred1.vars <- data.frame(age=20, workclass='Self-Employed',
                              education_num=3, marital_status='Single',
                              occupation='Professional', race='White',
                              hours_per_week=25, gender='Male')

# intercept + edad*estimate_age + estimate_self-employed +
#   educación*estimate_education_num + estimate_single +
#   estimate_professional + horas*estimate_hours_per_week
logi.pred1.res.link <- logi.coeffs[['(Intercept)']] +
  logi.coeffs[['age']]*logi.pred1.vars$age +
  logi.coeffs[['workclassSelf-Employed']] +
  logi.coeffs[['education_num']]*logi.pred1.vars$education_num +
  logi.coeffs[['marital_statusSingle']] +
  logi.coeffs[['occupationProfessional']] +
  logi.coeffs[['hours_per_week']]*logi.pred1.vars$hours_per_week

logi.pred1.response <- exp(logi.pred1.res.link)/(1+exp(logi.pred1.res.link))

logi.pred1.model.link <- predict(logi.model, logi.pred1.vars, type='link')
logi.pred1.model.resp <- predict(logi.model, logi.pred1.vars, type='response')
```

	Link	Response
Cálculo manual	5,9	1
`predict(...)`	5,9	1

Un hombre blanco de 20 años, autónomo, con 3 años de estudios, soltero y trabajando 25 horas semanales en el sector profesional cobrará menos de 50k euros anuales con un 100% de probabilidad.

También se desea saber la probabilidad de que el salario de un individuo sea menor a 50k euros para un hombre negro de 60 años con trabajo gubernamental, 15 años de estudios, casado, trabajando como collar blanco haciendo 35 horas semanales.

```
logi.coeffs <- logi.model$coefficients
logi.pred2.vars <- data.frame(age=60, workclass='Government',
                              education_num=15, marital_status='Married',
                              occupation='White-Collar', race='Black',
                              hours_per_week=35, gender='Male')
```

Patricia Lázaro Tello
plazarotello@uoc.edu

```
logi.pred2.res.link <- logi.coeffs[['(Intercept)']] +
  logi.coeffs[['age']]*logi.pred2.vars$age +
  logi.coeffs[['workclassGovernment']] +
  logi.coeffs[['education_num']]*logi.pred2.vars$education_num +
  logi.coeffs[['occupationWhite-Collar']] +
  logi.coeffs[['raceBlack']] +
  logi.coeffs[['hours_per_week']]*logi.pred2.vars$hours_per_week
logi.pred2.response <- exp(logi.pred2.res.link)/(1+exp(logi.pred2.res.link))

logi.pred2.model.link <- predict(logi.model, logi.pred2.vars, type='link')
logi.pred2.model.resp <- predict(logi.model, logi.pred2.vars, type='response')
```

	Link	Response
Cálculo manual	-5,4	0
`predict(...)`	-5,4	0

Un hombre negro de 60 años, con 15 años de estudios, casado, trabajando 35 horas semanales como collar blanco en un trabajo gubernamental cobrará menos de 50k euros anuales con un 0% de probabilidad; es decir, cobrará al menos 50k euros anuales con un 100% de probabilidad.

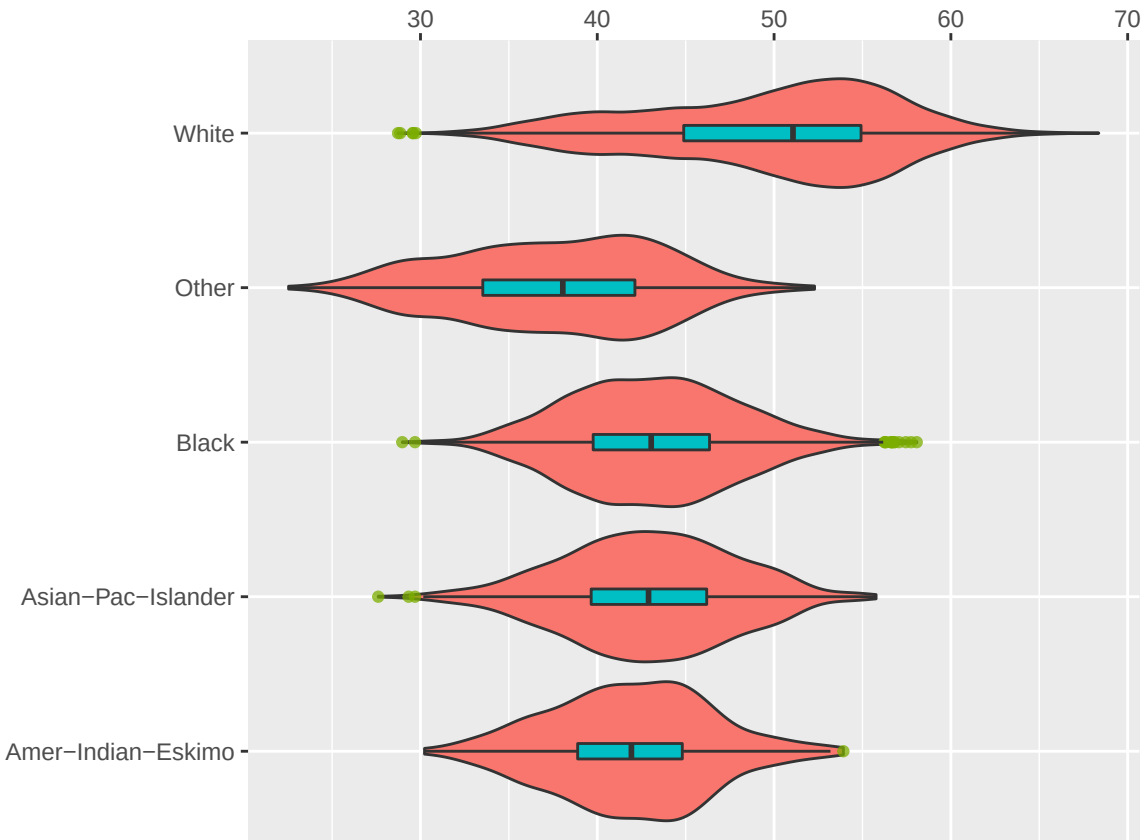
6 Análisis de la varianza de un factor

Se procede a analizar la existencia de diferencias significativas de ingresos (`income`) entre los diferentes grupos raciales.

6.1 Visualización

En primer lugar se va a visualizar la distribución de ingresos por raza:

```
ggplot(data=adult, mapping=aes(x=race, y=income)) +
  geom_violin(fill=default.color.main) +
  scale_y_continuous(position='right') +
  geom_boxplot(width=0.1, outlier.color=default.color.terciary,
               fill=default.color.secondary, outlier.alpha=0.75) +
  ggtitle('Distribución de los ingresos por raza') +
  xlab('') + ylab('') + coord_flip() + title.centered
```



Los grupos de personas negras, asiáticas e indias poseen una distribución de los ingresos más compacta que las personas de otra raza o las personas blancas. También cambia significativamente la media de las distribuciones entre negros/asiáticos/indios y blancos o de otra raza.

6.2 Modelo ANOVA

El modelo ANOVA es un test utilizado para comprobar si existen diferencias estadísticamente significativas entre las medias de 3 o más grupos.

En el caso presentado en este trabajo, se desea determinar si existen diferencias entre las medias de los ingresos por raza con un nivel de significación del 5% ($\alpha = 0.05$). El modelo corresponde con:

$$modelo_{ri} : income = \beta_0 + \beta_1 \, race$$

```
race.income.model <- lm(income ~ race, data=adult.factors)
summary(race.income.model)
```

```
##
## Call:
## lm(formula = income ~ race, data = adult.factors)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.06  -4.54   1.04   4.89  18.58
##
## Coefficients:
##                         Estimate Std. Error t value            Pr(>|t|)
## (Intercept)              49.7875     0.0397  1254.1 <0.0000000000000002 ***
## raceAmer-Indian-Eskimo   -7.9007     0.3775   -20.9 <0.0000000000000002 ***
## raceAsian-Pac-Islander   -6.8885     0.2092   -32.9 <0.0000000000000002 ***
## raceBlack                -6.6407     0.1249   -53.1 <0.0000000000000002 ***
## raceOther               -12.1841     0.4042   -30.1 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.6 on 32555 degrees of freedom
## Multiple R-squared:  0.129,  Adjusted R-squared:  0.129
## F-statistic: 1.21e+03 on 4 and 32555 DF,  p-value: <0.0000000000000002
```

Todos los coeficientes del modelo son estadísticamente significativos. El modelo explica un 12.9% de la varianza de los ingresos percibidos.

Respecto a los niveles del factor, se observa que las razas india, asiática, negra y otras razas minoritarias juegan un papel importante a la hora de calcular los ingresos del individuo. No ser una persona blanca perjudica a los ingresos percibidos.

El test ANOVA utilizará el siguiente contraste de hipótesis:

$$H_0: \mu_{White} = \mu_{Black} = \mu_{Asian} = \mu_{Indian} = \mu_{Other}$$
$$H_1: \exists\, i, j \in \{White,\ Black,\ Asian,\ Indian,\ Other\}: \mu_i \neq \mu_j$$

Se procede a calcular la variabilidad explicada por la variable raza sobre los ingresos:

```
race.income.anova <- anova(race.income.model)
race.income.anova
```

```
## Analysis of Variance Table
##
## Response: income
##              Df  Sum Sq Mean Sq F value              Pr(>F)
## race          4  211909   52977    1208 <0.0000000000000002 ***
## Residuals 32555 1427206      44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La variable raza es estadísticamente significativa con un nivel de significación del 5%; se acepta la hipótesis alternativa. El grupo racial al que pertenece el individuo explica el 13% de la varianza de los ingresos percibidos, que concuerda con el valor de R^2 obtenido del resumen del modelo.

Para estimar los efectos de los niveles del factor, se procede a calcular la media global de ingresos y restarla a la media de cada grupo.

```
income.mean <- mean(adult.factors$income)
group.means <- adult.factors %>% dplyr::select(income, race) %>%
  dplyr::group_by(race) %>% dplyr::summarise(avg=mean(income)) %>%
  dplyr::mutate(avg=avg-income.mean)

group.means
```

```
## # A tibble: 5 x 2
##   race                 avg
```

```
##   <fct>               <dbl>
## 1 White                1.03
## 2 Amer-Indian-Eskimo  -6.87
## 3 Asian-Pac-Islander  -5.85
## 4 Black               -5.61
## 5 Other              -11.2
```

Los valores son muy parecidos a los ofrecidos por el modelo de regresión lineal que se ha creado anteriormente. Ser de raza blanca supone una ventaja a la hora de percibir mayores ingresos, mientras que las razas minoritarias (otras razas) son las más afectadas negativamente.

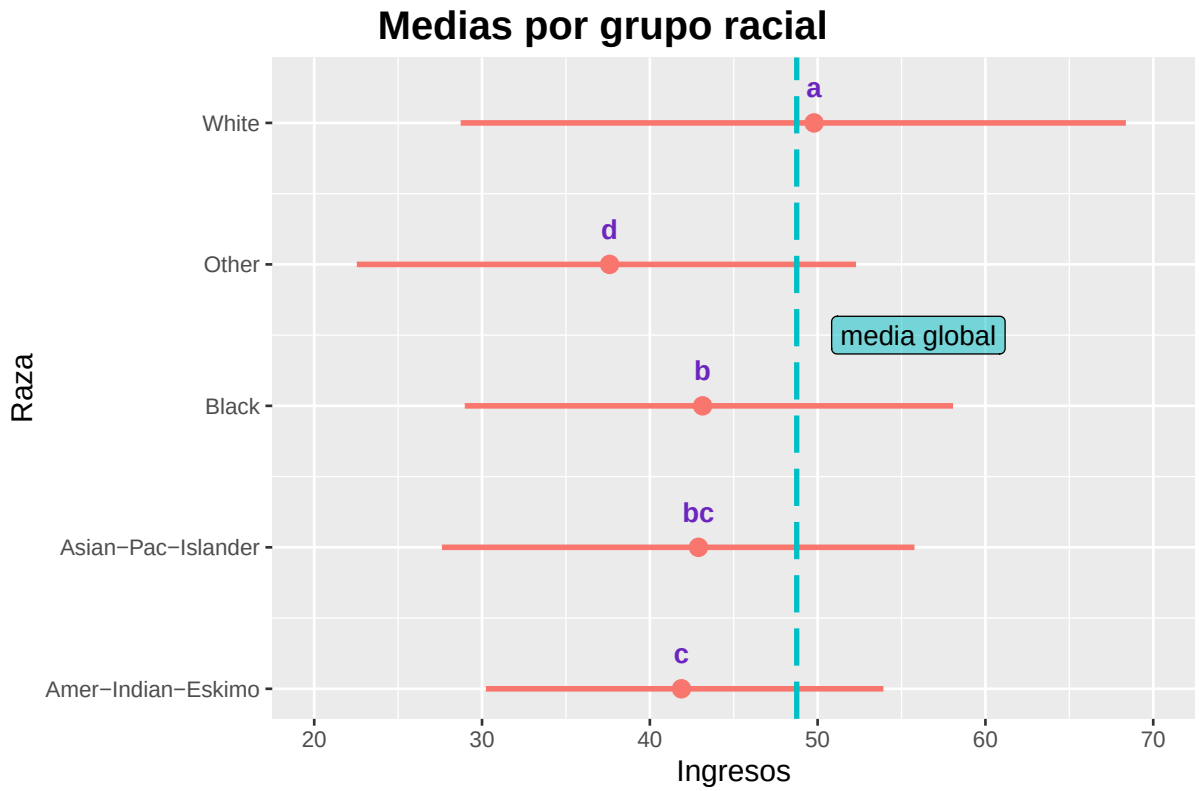
Se procede a realizar un contraste dos a dos para observar las diferencias entre los niveles:

```
tukey <- HSD.test(race.income.model, 'race', alpha=0.05, console=TRUE,
                  group=TRUE)
```

```
##
## Study: race.income.model ~ "race"
##
## HSD Test for income
##
## Mean Square Error:  44
##
## race,  means
##
##                     income std     r Min Max
## Amer-Indian-Eskimo      42 4.5   311  30  54
## Asian-Pac-Islander      43 4.8  1039  28  56
## Black                   43 4.7  3124  29  58
## Other                   38 5.8   271  23  52
## White                   50 6.9 27815  29  68
##
## Alpha: 0.05 ; DF Error: 32555
## Critical Value of Studentized Range: 3.9
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Treatments with the same letter are not significantly different.
##
##                     income groups
## White                   50      a
## Black                   43      b
## Asian-Pac-Islander      43     bc
```

```
## Amer-Indian-Eskimo     42      c
## Other                  38      d
```

```
tukey.plot <- tukey$means %>% dplyr::mutate(race=rownames(.)) %>%
  right_join(., tukey$groups, by='income') %>%
  dplyr::mutate(index=as.numeric(rownames(.)))

ggplot(data=tukey.plot) +
  geom_segment(mapping=aes(x=Min, xend=Max, y=index, yend=index),
               color=default.color.main, size=1.05) +
  geom_point(aes(x=income, y=index), color=default.color.main, size=3) +
  geom_vline(mapping=aes(xintercept=income.mean), linetype='longdash',
             color=default.color.secondary, size=1) +
  annotate('label', x=income.mean+7.25, y=3.5, label='media global',
           fill=default.color.secondary, alpha=0.5) +
  geom_text(mapping=aes(x=income, y=index, label=groups, fontface='bold'),
            nudge_y = 0.25, color=default.color.quat) +
  scale_y_continuous(labels=tukey.plot$race) + xlim(20, 70) + title.centered +
  ggtitle('Medias por grupo racial') + xlab('Ingresos') + ylab('Raza')
```

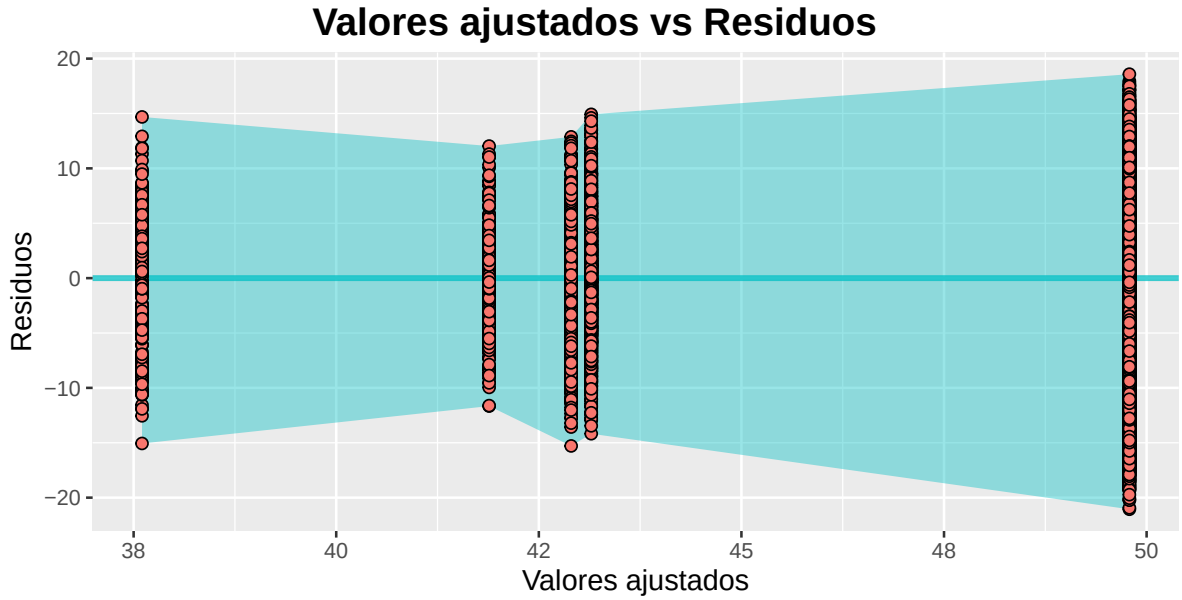


Se observa que la media global equivale a 49k euros. La raza blanca tiene la media más similar a la global, y conforma su propio grupo. La raza negra, asiática e india tienen medias parecidas, pero no iguales: negros y asiáticos podrían ser el mismo grupo, mientras que asiáticos e indios también podrían conformar un solo grupo. Las otras razas (`Other`) conforman su propio grupo con la media más baja de todas las razas.

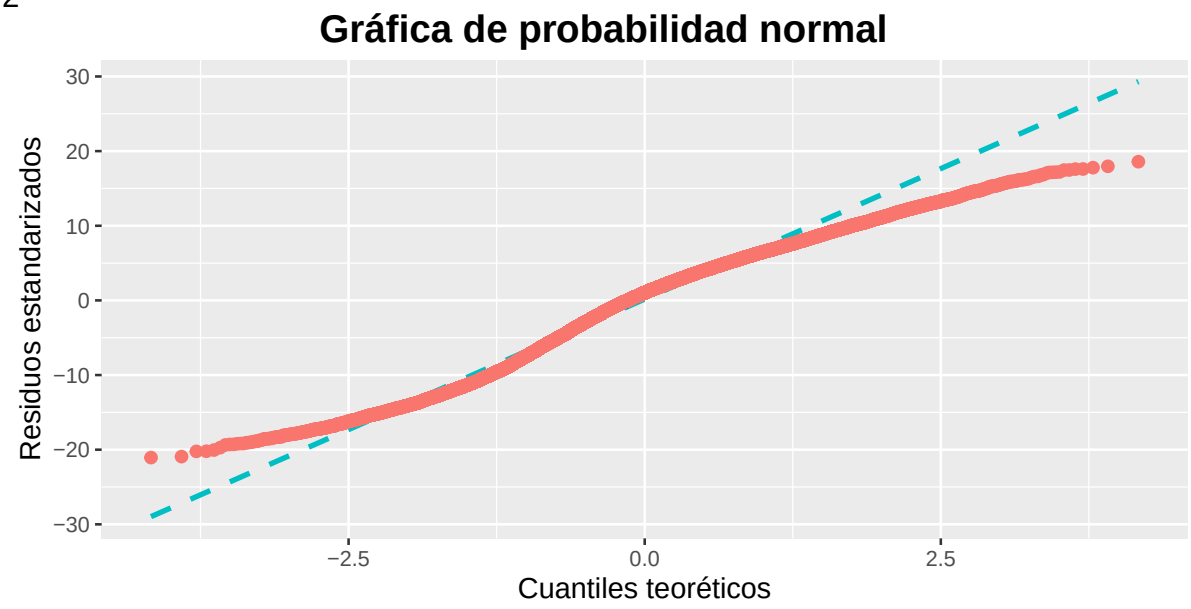
6.3 Adecuación del modelo ANOVA

```
race.income.model.fortified <- fortify(race.income.model)
race.income.model.minmax <- race.income.model.fortified %>%
  dplyr::group_by(race) %>% dplyr::summarise(max=max(.resid), min=min(.resid))
race.income.model.fortified <- race.income.model.fortified %>%
  dplyr::left_join(., race.income.model.minmax, by='race')

ggarrange(nrow=2, ncol=1,
          ggplot(race.income.model.fortified, aes(x=.fitted, y=.resid)) +
            geom_ribbon(mapping=aes(ymin=min, ymax=max), alpha=0.4,
                        fill=default.color.secondary) +
            geom_hline(color=default.color.secondary, size=1.15,
                       alpha=0.75, mapping=aes(yintercept=0)) +
            geom_point(size=2, shape=21, fill=default.color.main) +
            title.centered + ggtitle('Valores ajustados vs Residuos') +
            xlab('Valores ajustados') + ylab('Residuos') + labs(tag='1'),
          ggplot(race.income.model.fortified, aes(sample=.resid)) +
            geom_qq_line(linetype='dashed', size=1.05,
                         color=default.color.secondary) +
            geom_qq(size=2, color=default.color.main) + title.centered +
            ggtitle('Gráfica de probabilidad normal') + labs(tag='2') +
            xlab('Cuantiles teoréticos') + ylab('Residuos estandarizados'))
```

1



2



Los residuos no siguen una distribución normal: las colas son más pesadas de lo que deberían ser para tratarse de una distribución normal. Además, los residuos no se distribuyen homogéneamente, dado que en los extremos los residuos están más dispersos y en el centro, donde están las razas negra, asiática e india, se ajustan mejor.

Se comprueba que los residuos no siguen una distribución normal mediante el test de Kolmogorov-Smirnov:

```
lillie.test(race.income.model.fortified$.resid)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  race.income.model.fortified$.resid
## D = 0.06, p-value <0.0000000000000002
```

El test ofrece un p-value $\ll \alpha = 0.05$, por lo tanto con un nivel de significancia del 95% **los residuos no siguen una distribución normal**. Se procede a comprobar la homocedasticidad de los residuos mediante el test de NVC (*Non-COnstant Variance score*):

```
car::ncvTest(race.income.model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 592, Df = 1, p = <0.0000000000000002
```

El p-value $\ll \alpha = 0.05$, se rechaza la hipótesis nula de que los residuos son homocedásticos y se acepta la hipótesis alternativa de que **los residuos son heterocedásticos**.

Se procede a utilizar el test no paramétrico de Kruskal-Wallis para comprobar el resultado del test ANOVA:

```
kruskal.test(income~race, data=adult.factors)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  income by race
## Kruskal-Wallis chi-squared = 4210, df = 4, p-value <0.0000000000000002
```

Se obtiene un p-value $\ll \alpha = 0.05$; por tanto, se rechaza la hipótesis nula (todas las muestras provienen de la misma distribución) y se acepta la hipótesis alternativa (existe al menos una muestra que proviene de una población con una distribución distinta).

Se concluye que existe al menos una raza cuya media no es igual que la media del resto de razas en el conjunto de datos.

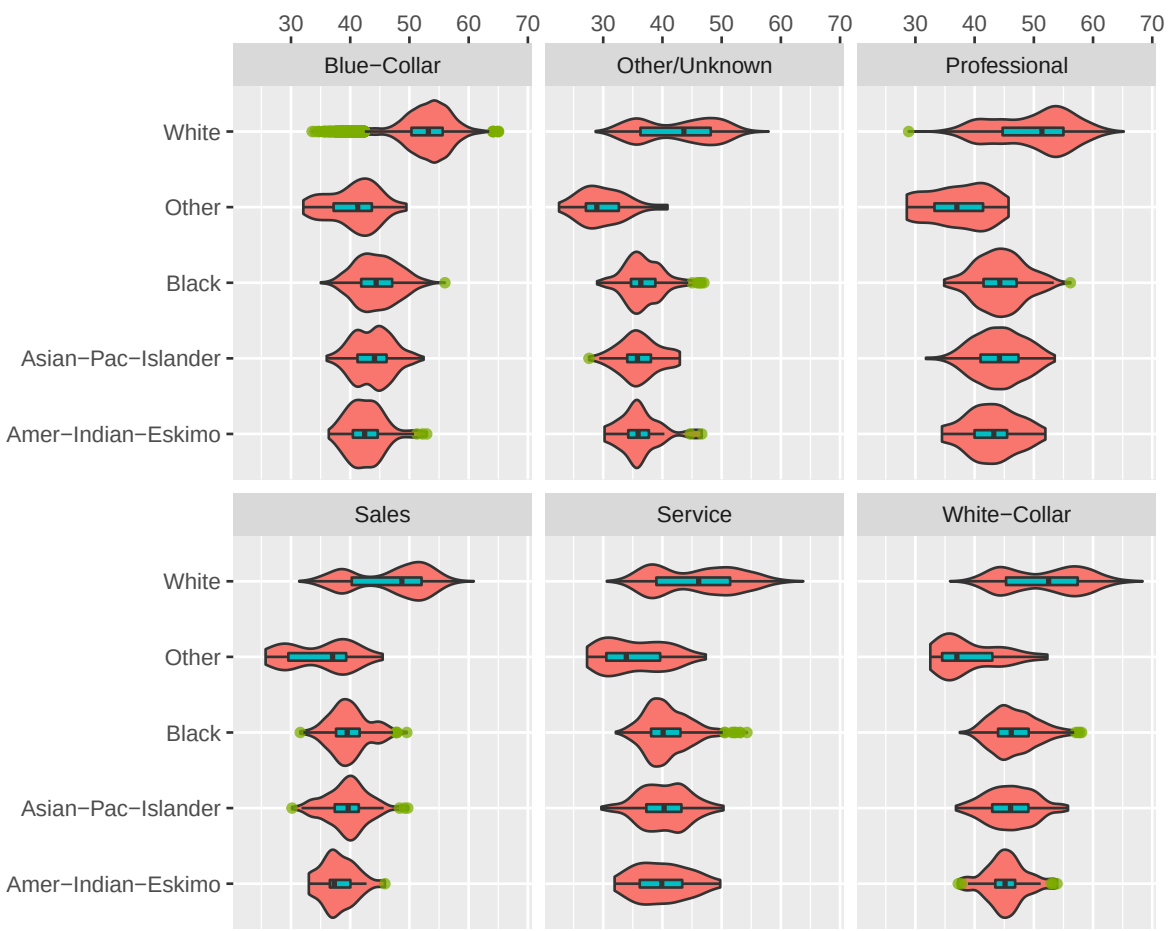
Patricia Lázaro Tello
plazarotello@uoc.edu

7 ANOVA multifactorial

Se procede a analizar la existencia de diferencias significativas de ingresos (`income`) entre los diferentes grupos raciales, añadiendo la ocupación del invididuo a la ecuación.

En primer lugar se va a visualizar la distribución de ingresos por raza y empleo:

```
ggplot(data=adult, mapping=aes(x=race, y=income)) +
  geom_violin(fill=default.color.main) +
  scale_y_continuous(position='right') + facet_wrap(.~occupation) +
  geom_boxplot(width=0.1, outlier.color=default.color.terciary,
               fill=default.color.secondary, outlier.alpha=0.75) +
  ggtitle('Distribución de los ingresos por raza y empleo') +
  xlab('') + ylab('') + coord_flip() + title.centered
```



En todos los campos profesionales la raza blanca es la que más ingresos obtiene. Se ovserva además que los collares blancos son los que pueden acceder a mayores ingresos para todas las razas.

Se procede a comprobar que se trata de un escenario balanceado, es decir, que existe un número similar de observaciones para cada grupo. Para ello se va a calcular el número de muestras por cada condición:

```
table(adult.factors$race, adult.factors$occupation)
```

```
##
##                      Blue-Collar Other/Unknown Professional Sales Service
##   White                     8714          1522         3651  3237    3962
##   Amer-Indian-Eskimo         120            26           33    26      45
##   Asian-Pac-Islander         215            65          186   108     191
##   Black                      909           216          239   254     772
##   Other                      104            23           31    25      51
##
##                      White-Collar
##   White                      6729
##   Amer-Indian-Eskimo           61
##   Asian-Pac-Islander          274
##   Black                       734
##   Other                        37
```

Se trata de un **escenario desbalanceado**, donde una raza (la raza blanca) acumula la mayoría de observaciones. Si se observa la tabla atendiendo al tipo de trabajo, los collares blancos y collares azules son los que acumulan la mayoría de las observaciones.

Realizar una modelización basada en ANOVA sin comprobar antes si se trata de un escenario balanceado o desbalanceado puede conllevar algunos inconvenientes si finalmente se trata de un caso desbalanceado:

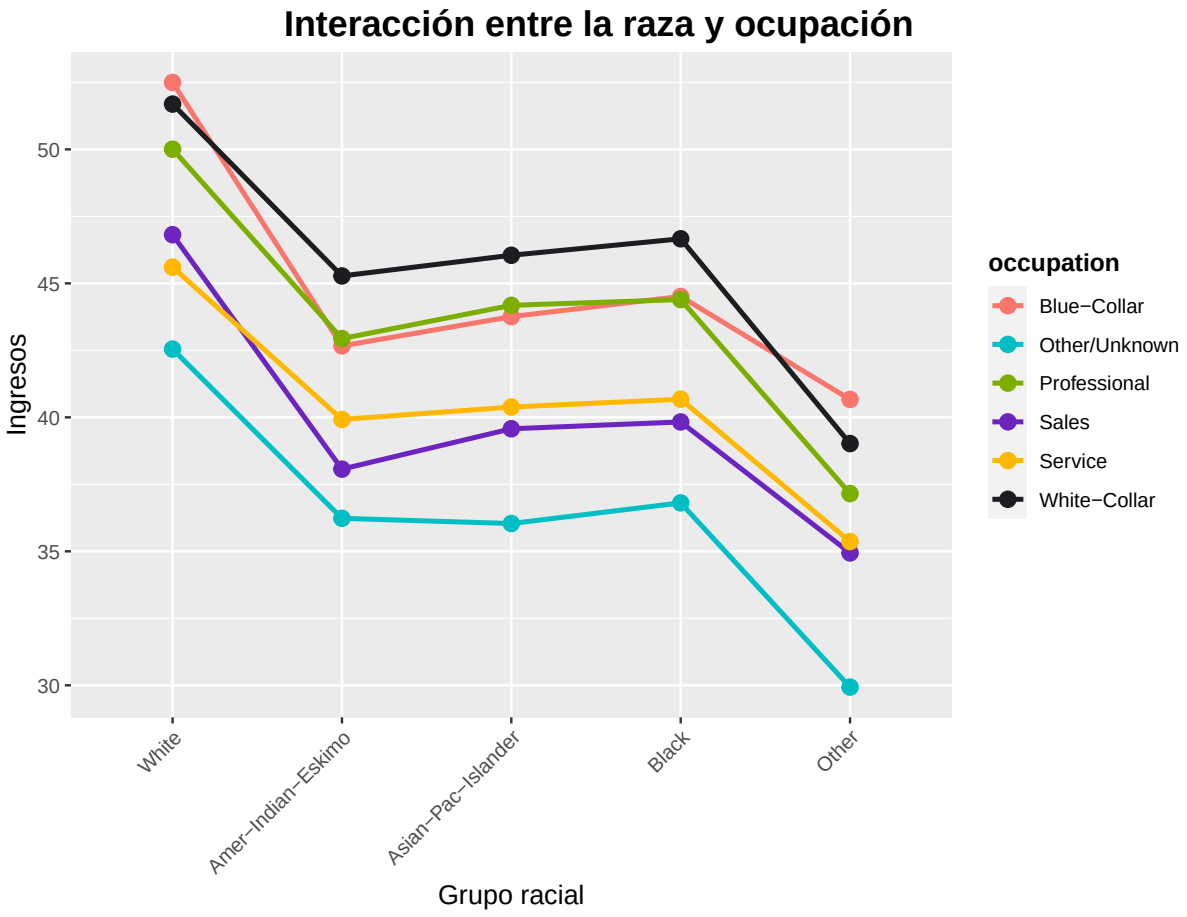
- El test es menos significativo, es decir, existe una probabilidad menor de detectar diferencias entre las medias de los grupos.
- El test es menos robusto cuando las varianzas entre los grupos son distintas. Esto, añadido a que el test es de por sí menos significativo, hace que no sea nada recomendable utilizar ANOVA si el tamaño de cada grupo y su varianza son distintos.

Patricia Lázaro Tello
plazarotello@uoc.edu

7.1 Estudio visual de la interacción

Se procede a observar la media de los ingresos por raza y ocupación para determinar si existe interacción entre los grupos:

```
ggplot(data=adult.factors, mapping=aes(x=race, color=occupation,
                                       group=occupation, y=income)) +
  stat_summary(fun='mean', geom='line', size=1.05) +
  stat_summary(fun='mean', geom='point', size=3) +
  scale_color_manual(values=palette) + title.centered +
  xlab('Grupo racial') + ylab('Ingresos') +
  theme(legend.title=element_text(face='bold')) +
  ggtitle('Interacción entre la raza y ocupación') +
  theme(axis.text.x=element_text(angle=45, hjust=1.0))
```



Se aprecia lo ya constatado en secciones anteriores: dada una profesión, la raza blanca percibe más ingresos que las demás razas. También se observa que dentro de una

raza, los individuos que realizan otro tipo de trabajo o cuya profesión es desconocida perciben en todos los casos ingresos menores al resto de profesiones.

Existe confusión entre los dos factores, dado que las diferentes líneas, aunque mayormente paralelas, se cruzan en algunos puntos:

- Las personas blancas cobran más como collares azules que como collares blancos, pero los indios cobran más como collares blancos que como collares azules.
- Las personas blancas cobran más como collares azules que como profesionales, pero los indios y asiáticos cobran más como profesionales que como collares azules.
- Los blancos cobran más como personal de ventas que de servicios, pero el resto de razas cobra más como personal de servicios que de ventas.

Se procede a comprobar si esta interacción es significativa mediante el modelo ANOVA:

```
ro.income.model <- lm(income~race*occupation, data=adult.factors)
summary(ro.income.model)
```

```
##
## Call:
## lm(formula = income ~ race * occupation, data = adult.factors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.146  -4.248   0.661   4.273  18.184
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                      52.4974     0.0629  835.16
## raceAmer-Indian-Eskimo                           -9.8322     0.5393  -18.23
## raceAsian-Pac-Islander                           -8.7403     0.4051  -21.58
## raceBlack                                        -7.9856     0.2045  -39.05
## raceOther                                       -11.8301     0.5788  -20.44
## occupationOther/Unknown                          -9.9508     0.1630  -61.04
## occupationProfessional                           -2.4915     0.1157  -21.54
## occupationSales                                  -5.6822     0.1208  -47.05
## occupationService                                -6.8910     0.1124  -61.29
## occupationWhite-Collar                           -0.8081     0.0952   -8.49
## raceAmer-Indian-Eskimo:occupationOther/Unknown    3.5168     1.2798    2.75
## raceAsian-Pac-Islander:occupationOther/Unknown    2.2309     0.8464    2.64
## raceBlack:occupationOther/Unknown                 2.2470     0.4731    4.75
```

```
## raceOther:occupationOther/Unknown                -0.7838     1.3619   -0.58
## raceAmer-Indian-Eskimo:occupationProfessional     2.7645     1.1592    2.38
## raceAsian-Pac-Islander:occupationProfessional     2.9143     0.5989    4.87
## raceBlack:occupationProfessional                  2.3732     0.4420    5.37
## raceOther:occupationProfessional                 -1.0212     1.2063   -0.85
## raceAmer-Indian-Eskimo:occupationSales            1.0840     1.2751    0.85
## raceAsian-Pac-Islander:occupationSales            1.4986     0.7025    2.13
## raceBlack:occupationSales                         0.9982     0.4336    2.30
## raceOther:occupationSales                        -0.0450     1.3126   -0.03
## raceAmer-Indian-Eskimo:occupationService          4.1465     1.0318    4.02
## raceAsian-Pac-Islander:occupationService          3.5175     0.5942    5.92
## raceBlack:occupationService                       3.0563     0.3084    9.91
## raceOther:occupationService                       1.5885     1.0094    1.57
## raceAmer-Indian-Eskimo:occupationWhite-Collar     3.4190     0.9276    3.69
## raceAsian-Pac-Islander:occupationWhite-Collar     3.0992     0.5430    5.71
## raceBlack:occupationWhite-Collar                  2.9586     0.3064    9.66
## raceOther:occupationWhite-Collar                 -0.8378     1.1273   -0.74
##                                                          Pr(>|t|)
## (Intercept)                                  < 0.0000000000000002 ***
## raceAmer-Indian-Eskimo                       < 0.0000000000000002 ***
## raceAsian-Pac-Islander                       < 0.0000000000000002 ***
## raceBlack                                    < 0.0000000000000002 ***
## raceOther                                    < 0.0000000000000002 ***
## occupationOther/Unknown                      < 0.0000000000000002 ***
## occupationProfessional                       < 0.0000000000000002 ***
## occupationSales                              < 0.0000000000000002 ***
## occupationService                            < 0.0000000000000002 ***
## occupationWhite-Collar                       < 0.0000000000000002 ***
## raceAmer-Indian-Eskimo:occupationOther/Unknown             0.00600 **
## raceAsian-Pac-Islander:occupationOther/Unknown             0.00840 **
## raceBlack:occupationOther/Unknown                    0.0000020512 ***
## raceOther:occupationOther/Unknown                         0.56492
## raceAmer-Indian-Eskimo:occupationProfessional             0.01709 *
## raceAsian-Pac-Islander:occupationProfessional        0.0000011422 ***
## raceBlack:occupationProfessional                     0.0000000793 ***
## raceOther:occupationProfessional                          0.39725
## raceAmer-Indian-Eskimo:occupationSales                    0.39526
## raceAsian-Pac-Islander:occupationSales                    0.03292 *
## raceBlack:occupationSales                                 0.02134 *
## raceOther:occupationSales                                 0.97265
## raceAmer-Indian-Eskimo:occupationService             0.0000586968 ***
## raceAsian-Pac-Islander:occupationService             0.0000000033 ***
## raceBlack:occupationService                  < 0.0000000000000002 ***
## raceOther:occupationService                               0.11556
## raceAmer-Indian-Eskimo:occupationWhite-Collar             0.00023 ***
## raceAsian-Pac-Islander:occupationWhite-Collar        0.0000000116 ***
```

```
## raceBlack:occupationWhite-Collar               < 0.0000000000000002 ***
## raceOther:occupationWhite-Collar                             0.45737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.9 on 32530 degrees of freedom
## Multiple R-squared:  0.317,  Adjusted R-squared:  0.316
## F-statistic:  520 on 29 and 32530 DF,  p-value: <0.0000000000000002
```

El modelo de regresión lineal es estadísticamente significativo y la interacción también lo es, dado que existen combinaciones de niveles de los factores estadísticamente significativos. El modelo explica el 32% de la variabilidad de los datos.

```
roi.anova <- anova(ro.income.model)
roi.anova
```

```
## Analysis of Variance Table
##
## Response: income
##                    Df  Sum Sq Mean Sq F value              Pr(>F)
## race                4  211909   52977  1538.7 <0.0000000000000002 ***
## occupation          5  300041   60008  1742.8 <0.0000000000000002 ***
## race:occupation    20    7125     356    10.3 <0.0000000000000002 ***
## Residuals       32530 1120040      34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La raza y la ocupación son estadísticamente significativas; el término que resulta de la combinación de ambos factores también es estadísticamente significaivo. Esto quiere decir que la interacción entre raza y ocupación es estadísticamente significativa.

	Varianza explicada
Raza	13%
Ocupación	18%
Interacción	0,43%

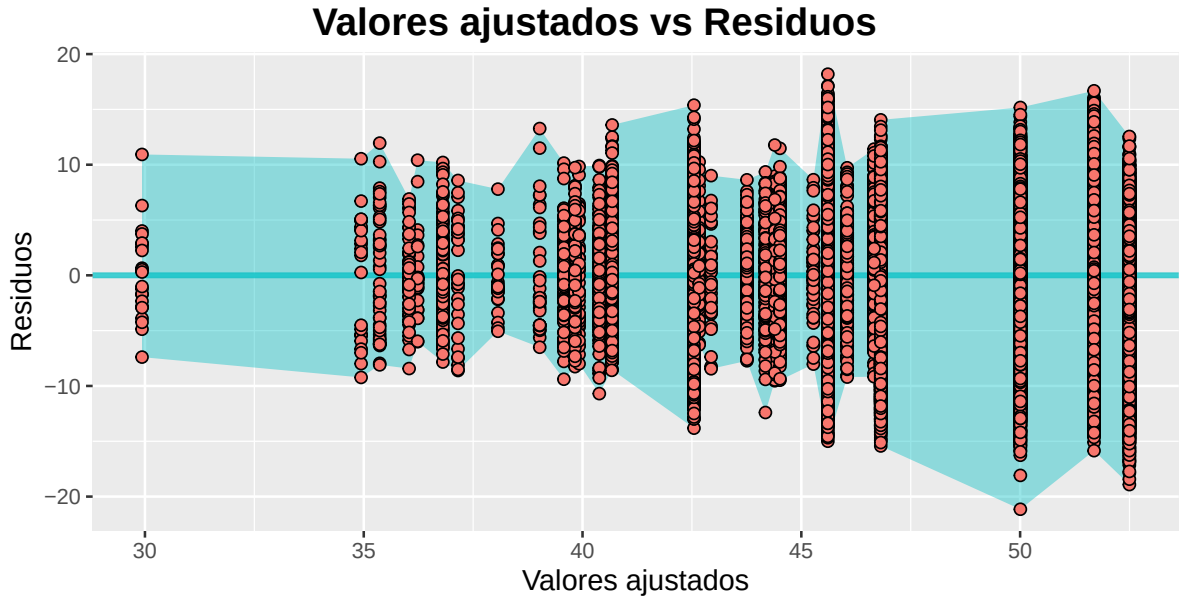
Para comprobar la adecuación del modelo, se procede a analizar el gráfico de residuos y el gráfico QQ de residuos para comprobar que se cumplen los supuesto de homocedasticidad y normalidad.

```
ro.income.model.fortified <- fortify(ro.income.model)
ro.income.model.minmax <- ro.income.model.fortified %>%
  dplyr::group_by(race, occupation) %>%
```

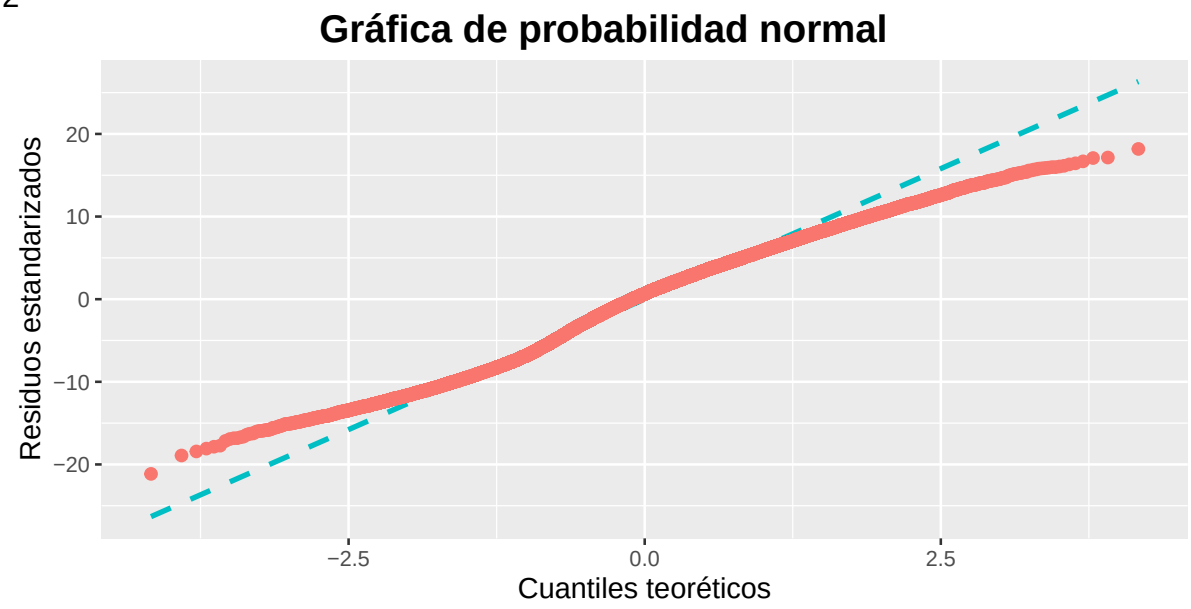
```
  dplyr::summarise(max=max(.resid), min=min(.resid), .groups='keep')
ro.income.model.fortified <- ro.income.model.fortified %>%
  dplyr::left_join(., ro.income.model.minmax, by=c('race', 'occupation'))

ggarrange(nrow=2, ncol=1,
          ggplot(ro.income.model.fortified, aes(x=.fitted, y=.resid)) +
            geom_ribbon(mapping=aes(ymin=min, ymax=max), alpha=0.4,
                        fill=default.color.secondary) +
            geom_hline(color=default.color.secondary, size=1.15,
                       alpha=0.75, mapping=aes(yintercept=0)) +
            geom_point(size=2, shape=21, fill=default.color.main) +
            title.centered + ggtitle('Valores ajustados vs Residuos') +
            xlab('Valores ajustados') + ylab('Residuos') + labs(tag='1'),
          ggplot(ro.income.model.fortified, aes(sample=.resid)) +
            geom_qq_line(linetype='dashed', size=1.05,
                         color=default.color.secondary) +
            geom_qq(size=2, color=default.color.main) + title.centered +
            ggtitle('Gráfica de probabilidad normal') + labs(tag='2') +
            xlab('Cuantiles teoréticos') + ylab('Residuos estandarizados'))
```

Patricia Lázaro Tello
plazarotello@uoc.edu

1



2



Los residuos parecen más dispersos en el extremo derecho, mientras que en el extremo izquierdo son más compactos y en el centro parecen más aleatorios; **no se cumple el supuesto de homocedasticidad de residuos**. En las gráfica de probabilidad normal se observan colas mucho más pesadas de lo que deberían para tratarse de una distribución normal; **no se cumple el supuesto de normalidad de residuos**.

Patricia Lázaro Tello
plazarotello@uoc.edu

8 Conclusiones

- El *dataset* contiene 32.560 observaciones con 9 atributos en cada una de ellas, sin valores nulos que tratar.
- Se observan sesgos de género y raza en los ingresos percibidos, siendo el hombre blanco el beneficiado y el resto de grupos los perjudicados por el sesgo.
- Se ha comprobado el sesgo de género mediante un contraste de hipótesis, del que se concluye que de media los hombres perciben más ingresos que las mujeres.
- Se ha comprobado el sesgo de raza mediante un contraste de hipótesis, del que se concluye que de media las personas de raza blanca perciben al menos 6.450 euros más que las personas de raza negra.
- Se ha creado un modelo de regresión lineal utilizando las variables edad, educación, horas trabajadas semanalmente y género. La inclusión de la variable raza supone una mejora significativa en el modelo, que no termina de ajustar bien a los datos.
- Se ha creado un modelo de regresión logística utilizando todas las variables del *dataset* con el objetivo de determinar la probabilidad de que un individuo de ciertas características perciba ingresos menores de 50.000 euros. El modelo ajusta de forma excelente a los datos.
- Se ha realizado un análisis de la varianza del factor raza, del que se concluye que las medias de los ingresos por raza son distintas: Las razas blanca, negra, india y otras minoritarias pertenecen a grupos distintos; los asiáticos podrían pertenecer indistintamente al grupo de raza negra o raza india. El modelo no se ajusta correctamente a los datos.
- Se ha realizado un análisis de la varianza del factor raza y ocupación, del que se concluye que existe una interacción estadísticamente significativa entre ambos factores. Además, se trata de un escenario desbalanceado y de un modelo que no ajusta bien a los datos.