Examen, problema 1

EYP2435 - Análisis Multivariado

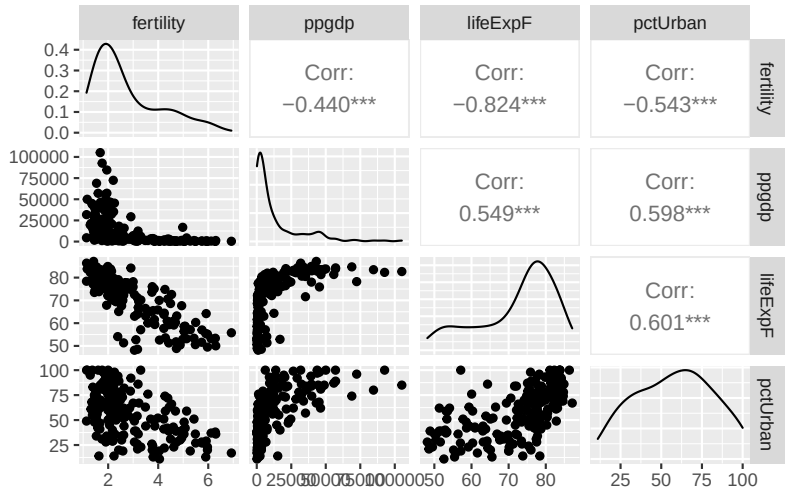
Sebastián Celaya Camila Echeverría

Datos de naciones Unidas, UN. Los datos en el archivo `UNData`, en canvas contiene algunas variables, entre ellas *ppgdp*, el producto nacional bruto por persona de 2009 en dólares estadounidenses, *fertility*, la tasa de natalidad por cada 100 mujeres en la población en el año 2009, *lifeExp* esperanza de vida y *pctUrban* porcentaje de población urbana; además de dos variables cualitativas, región y grupo. Los datos son para 199 localidades, en su mayoría Países miembros de la ONU, pero también para otras áreas como Hong Kong que no son países independientes.

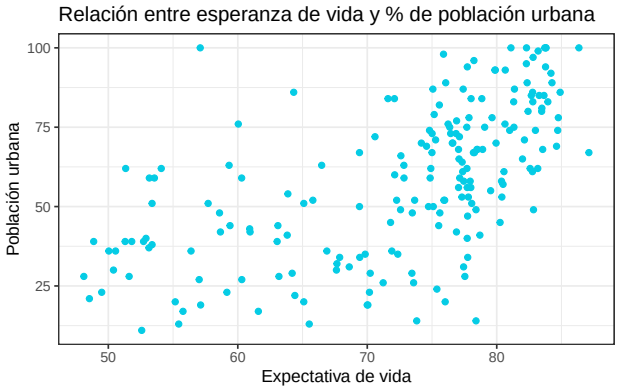
a) Haga un análisis exploratorio de los datos. Escriba sus conclusiones.
b) Suponga que se desea analizar el efecto de las otras variables disponibles en *(fertility, lifeEXP)*. Para esto, proponga un modelo estadístico y verifique si es adecuado. Usando algunas herramientas de inferencia estadística, responda si existe efecto de las otras variables sobre *(fertility, lifeEXP)*. Escriba sus conclusiones.

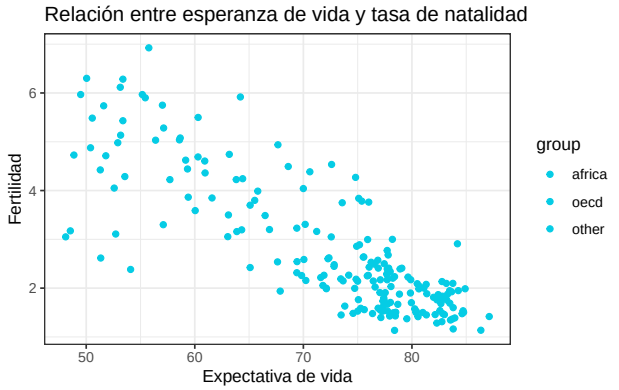
Solución a)

En primer lugar, veremos cómo se comportan las variables continuas entre sí. Para esto, realizaremos un gráfico de correlación.



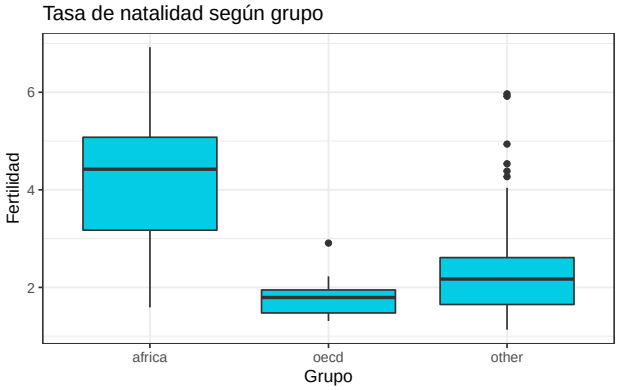
Las dos relaciones más fuertes que encontramos son entre `lifeExpF` y `fertility` y `lifeExpF` y `pctUrban`. En los siguientes gráficos podemos ver más de cerca cómo se relacionan estas variables.

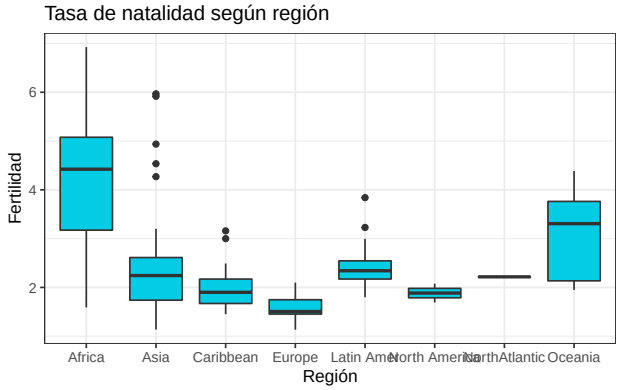




Por un lado, mientras la correlación es de 0.601, el gráfico no muestra una relación completamente lineal entre la esperanza de vida y el porcentaje de población urbana. De hecho, en algún punto hasta pareciera una simple nube de puntos. Sin embargo, la relación con la tasa de natalidad sí se ve bastante fuerte y bastante marcada, lo que era de esperarse de una correlación de -0.824.

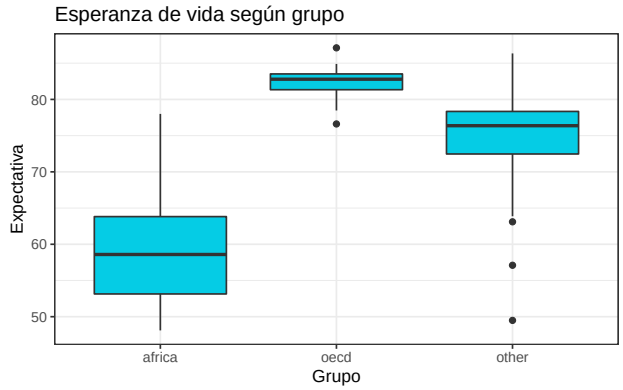
Luego, debemos ver cómo se comportan las variables numéricas si las agrupamos por las variables categóricas. Para esto, comenzaremos realizando boxplots de `fertility`.

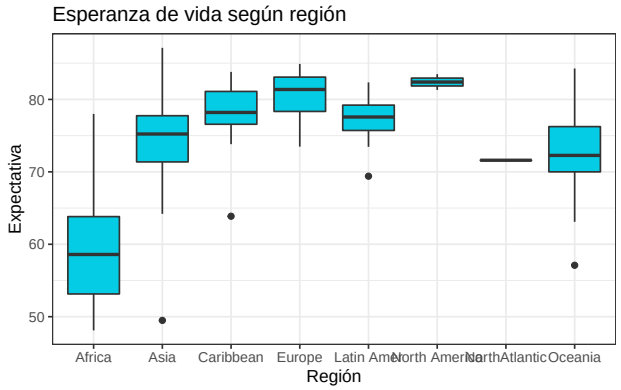




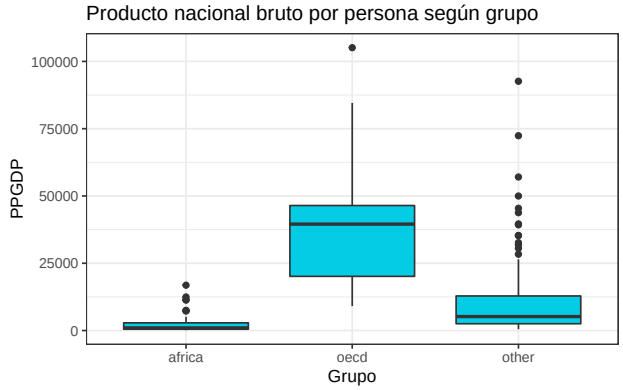
Lo primero que salta a la luz es que África, tanto como región como grupo, posee las tasas de natalidad más altas, siendo su mediana superior al tercer cuartil del resto de los grupos o regiones, aunque presenta un rango intercuartílico bastante amplio. Además las regiones pertenecientes a los grupos `oecd` u otros presentan tasas bastante bajas en comparación.

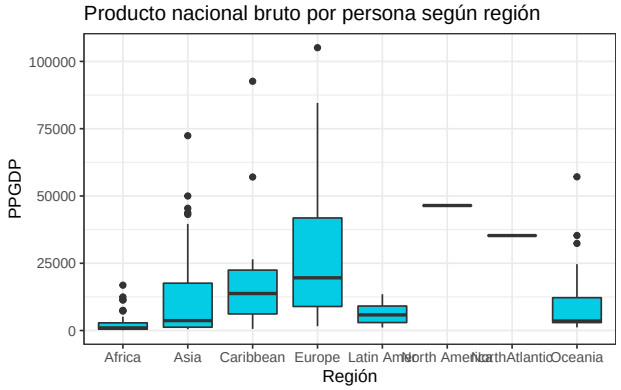
Un detalle que igual podría llamar la atención es la nula dispersión de la fertilidad de la región del Atlántico Norte. Esto se debe a que, según los datos, esta región está compuesta de un único país: Groenlandia.





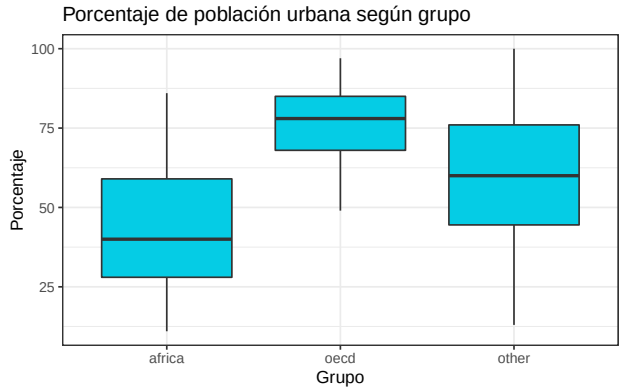
Luego, al realizar los gráficos de `lifeExpF`, podemos ver que la situación se revierte drásticamente. El grupo de la `oecd` es el que tiene mayor esperanza de vida, mientras que África es el que tiene la menor. En cuanto a las regiones, Norte América es la que lidera, seguida desde cerca por Europa y el Caribe.

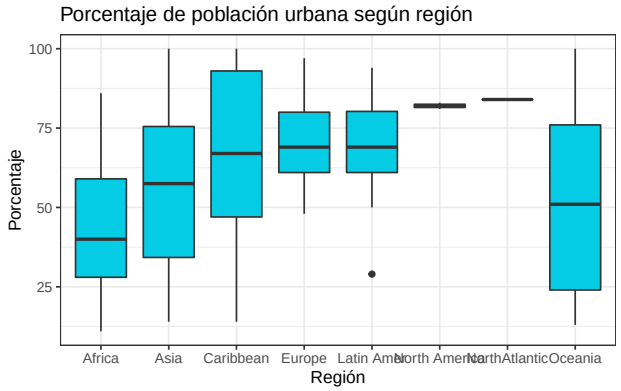




Con respecto al producto nacional, se repite el patrón de la esperanza nde vida: África posee el menor, mientras que el grupo de la oecd posee el mayor. Además, las medianas de América del Norte y Groenlandia son las más altas de entre las regiones.

Finalmente, podemos ver que el patrón se vuelve a repetir para el porcentaje de población urbana.





De esta manera, a priori, pareciera que el grupo y/o región tiene un gran impacto en las distintas variables cuantitativas de la base de datos, lo que, a su vez, determina cómo se comportarán estas.

Solución b)

Para este inciso, generaremos una regresión lineal múltiple donde los vectores de respuesta serán `fertility` y `lifeExpF`. El resto de las variables, menos `localities` que actúa como un identificador, serán los predictores del modelo.

```
modelo <- lm(cbind(fertility, lifeExpF) ~ ., datos[,-1])
```

Ahora que tenemos el modelo, calculamos los estadísticos que nos servirán para probar la hipótesis H_0: $B_2 = 0$. Para esto, utilizamos la función `manova()` de `R`, la que nos permite calcular los distintos estadísticos que fueron mencionados en clases. Estos valores se ilustran en la siguiente tabla:

```
pillai <- summary(manova(modelo), test = "Pillai")[["stats"]][-5,2]
wilks <- summary(manova(modelo), test = "Wilks")[["stats"]][-5,2]
hotelling <- summary(manova(modelo), test = "Hotelling-Lawley")[["stats"]][-5,2]
roy <- summary(manova(modelo), test = "Roy")[["stats"]][-5,2]
```

	Pillai	Wilks	Hotelling-Lawley	Roy
region	0.753	0.283	2.409	2.356
group	0.103	0.897	0.115	0.115
ppgdp	0.132	0.868	0.152	0.152
pctUrban	0.119	0.881	0.135	0.135

De esta manera, podemos ver que todos los estadísticos concuerdan en que la variable `region` es la que más aporta, lo que concuerda con el análisis exploratorio que hicimos anteriormente. Algo similar ocurre con la variable que menos aporta, pues todos señalan a la variable `group`.

```
summary_mod <- summary(modelo)
```

	fertility	lifeExpF
R^2	0.595	0.720
R^2 ajustado	0.574	0.705

Analizando si el modelos es adecuado podemos observar que el R^2 se ajusta correctamente, en específico el R^2 de `lifeExpF` se ajusta de mejor forma que de `fertility`. Sin embargo, un detalle a tener en cuenta son los valores-p y los coeficientes del modelo, que en este caso son los siguientes (redondeados a 3 decimales):

	B1	B2	p1	p2
(Intercept)	4.905	55.249	0.000	0.000
regionAsia	-1.563	12.430	0.000	0.000
regionCaribbean	-1.832	14.674	0.000	0.000
regionEurope	-2.229	14.819	0.000	0.000
regionLatin Amer	-1.388	14.286	0.000	0.000
regionNorth America	-1.769	12.842	0.009	0.003
regionNorthAtlantic	-1.313	5.182	0.145	0.359
regionOceania	-0.985	10.767	0.000	0.000
groupoecd	0.112	2.515	0.649	0.106
ppgdp	0.000	0.000	0.721	0.023
pctUrban	-0.016	0.102	0.000	0.000

De esta manera, podemos ver que tenemos coeficientes que no son significativos, por lo que el modelo no es necesariamente el más adecuado para esta situación. Así, implementaremos un método de selección stepwise que nos ayude a formar un buen modelo para esta situación.

```
update_y.formula <- function(variables, fm) {
  as.formula(paste0(variables, " ~ ", paste(all.vars(fm)[-1], collapse=" + ")))
}

step1 <- function(y, orig_fm){
  fm <- update_y.formula(y, orig_fm)
  step(lm(fm, data=datos[,-1]))
}

Y <- c("fertility", "lifeExpF")
fm <- fertility ~ region + group + ppgdp + pctUrban
modelo2.0 <- lapply(Y, step1, orig_fm=fm)
```

	B1	B2	p1	p2
(Intercept)	4.920	54.784	0.000	0.000
regionAsia	-1.567	0.000	0.000	0.000
regionCaribbean	-1.853	0.000	0.000	0.000
regionEurope	-2.199	0.000	0.000	0.000
regionLatin Amer	-1.371	0.000	0.000	0.000
regionNorth America	-1.720	0.000	0.008	0.000
regionNorthAtlantic	-1.355	0.000	0.128	0.000
regionOceania	-0.985	0.000	0.000	0.000
groupoecd	0.000	16.548	0.000	0.000
groupother	0.000	12.987	0.000	0.000
ppgdp	0.000	0.000	0.000	0.032
pctUrban	-0.016	0.113	0.000	0.000

Donde ahora sí tenemos valores mucho más significativos. Así, podemos ver que efectivamente la región tiene una mayor importancia en la tasa de natalidad, mientras que el grupo es el que más influye en la esperanza de vida.