Laboratorio 1 - Estadística Aplicada II

The Powerpuff Girls

05/08/2019

Introducción

El prestigio, del latín “praestigium”, es el realce o el buen crédito de alguien o algo. Es por esto que en sociología entendemos por prestigio social al grado de aceptación (que puede ser bueno o malo) que genera una conducta, actitud o situación entre los miembros de una comunidad. Cuanto mayor es el prestigio social de algo, más grande es el número de personas dispuestas a encontrarse relacionadas o involucradas con ello.

En el sentido ocupacional, el prestiguio se define como:

1. La naturaleza consensual de clasificar una ocupación basándose en la creencia colectiva de su valor.
2. El prestigio como una medida de que tan “deseable” es la ocupación en base a las recompensas socioeconómicas.

A continuación se presenta un análisis que busca concluir la relación entre distintas variables y el prestigio ocupacional.

La base de datos considerada se realizó en 1971 y se llama “Prestigio de las Ocupaciones Canadienses”. Las variables consideradas son (respuesta a la pregunta 2):

1. **Educación** : Nivel educativo promedio de la ocupación, en 1971. Tipo de variable cuantitativa y continua.
2. **Ingreso** : Ingreso promedio de la ocupación,expresado en dólares, en 1971. Tipo de valiable cuantitativa y continua.
3. **Mujeres** : Porcentaje de mujeres por ocupación.Tipo de valiable cuantitativa y continua.
4. **Prestigio** : Calificación del prestigio de la ocupación según Pineo-Porter. Tipo de valiable cuantitativa y continua.
5. **Censo** : Clave de la ocupación en el Censo. Tipo de valiable cuantitativa y discreta (categórica).
6. **Tipo** : Tipo de ocupación: obrero (bc), profesional, gerencial y técnico (prof), oficinista (wc). Tipo de valiable cualitativa, se puede representar como una variable categórica, por ejemplo bc = 1, prof = 2 y wc = 3.

Análisis exploratorio

Resumen estadístico del conjunto de datos

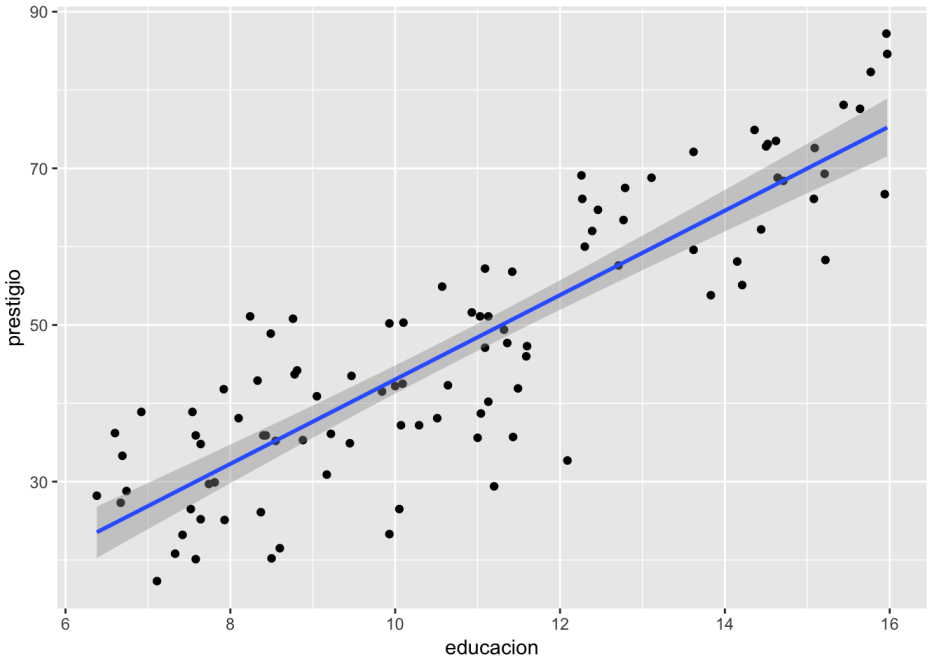
Resalta que existen profesiones con procentaje de mujeres igual a cero (FIREFIGHTERS, ROTARY.WELL.DRILLERS, RAILWAY.SECTIONMEN, TRAIN.ENGINEERS, LONGSHOREMEN). También podemos observar que hay una ocupación con un porcentaje mujeres del 97.51%, el cual es para la ocupación de secretaria.

```
summary(datos)
```

```
##               ocupacion    educacion         ingreso         mujeres
##  ACCOUNTANTS       : 1   Min.   : 6.380   Min.   :  611   Min.   : 0.000
##  AIRCRAFT.REPAIRMEN: 1   1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592
##  AIRCRAFT.WORKERS  : 1   Median :10.540   Median : 5930   Median :13.600
##  ARCHITECTS        : 1   Mean   :10.738   Mean   : 6798   Mean   :28.979
##  ATHLETES          : 1   3rd Qu.:12.648   3rd Qu.: 8187   3rd Qu.:52.203
##  AUTO.REPAIRMEN    : 1   Max.   :15.970   Max.   :25879   Max.   :97.510
##  (Other)           :96
##    prestigio         censo        tipo
##  Min.   :14.80   Min.   :1113   <NA>: 4
##  1st Qu.:35.23   1st Qu.:3120   bc  :44
##  Median :43.60   Median :5135   prof:31
##  Mean   :46.83   Mean   :5402   wc  :23
##  3rd Qu.:59.27   3rd Qu.:8312
##  Max.   :87.20   Max.   :9517
##
```

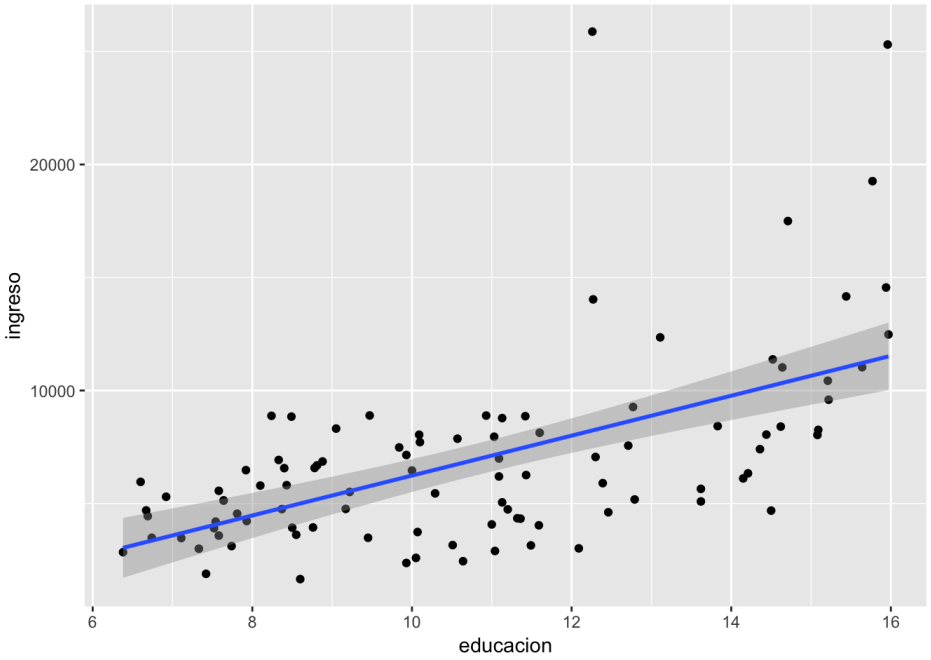
Gráficas entre los datos . La primera gráfica que presentamos nos permite ver la relación entre las variables de educación y prestigio. El segundo nos permite observar la educación con el ingreso y el último el ingreso contra el prestigio.

```
sinNa <- subset (datos, tipo!="<NA>")
ggplot(sinNa, aes(x=educacion, y = prestigio)) +geom_point()+stat_smooth(method = lm)
```

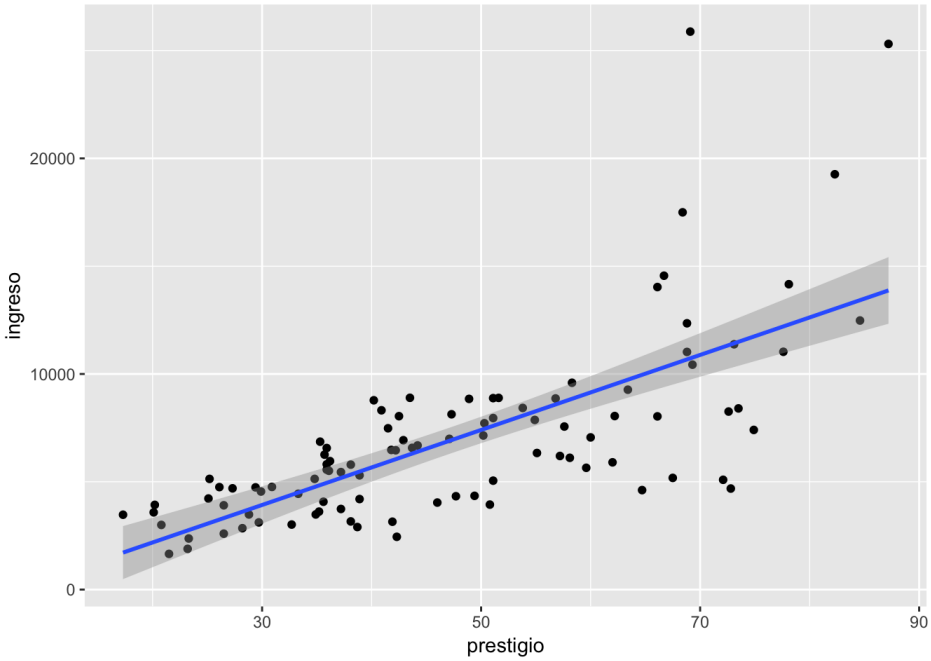



```
#ggplot(sinNa, aes(x=tipo, y = mujeres)) + geom_bar(stat="identity", width=0.5, fill="pink", size=10) + xlab("Tipo de ocupación") + #ylab("Número de mujeres") + ggtitle("obrero (bc)  - profesional, gerencial y técnico (prof) - oficinista (wc)")
```

```
ggplot(sinNa, aes(x=educacion, y = ingreso)) +geom_point()+stat_smooth(method = lm)
```




```
sinNa <- subset (datos, tipo!="<NA>")
ggplot(sinNa, aes(x=prestigio, y = ingreso)) +geom_point()+stat_smooth(method = lm)
```




```
#ggplot(sinNa, aes(x=tipo, y = mujeres)) + geom_bar(stat="identity", width=0.5, fill="pink", size=10) + xlab("Tipo de ocupación") + #ylab("Número de mujeres") + ggtitle("obrero (bc)  - profesional, gerencial y técnico (prof) - oficinista (wc)")
```

Correlaciones entre las variables

Para el cálculo de la correlación lineal entre las variables no se consideraron las variables ocupación ni censo.

Resalta que las variables prestigio y educación están muy relacionadas (linealmente), de igual manera, las variables de prestigio e ingreso muestran relación entre sí.

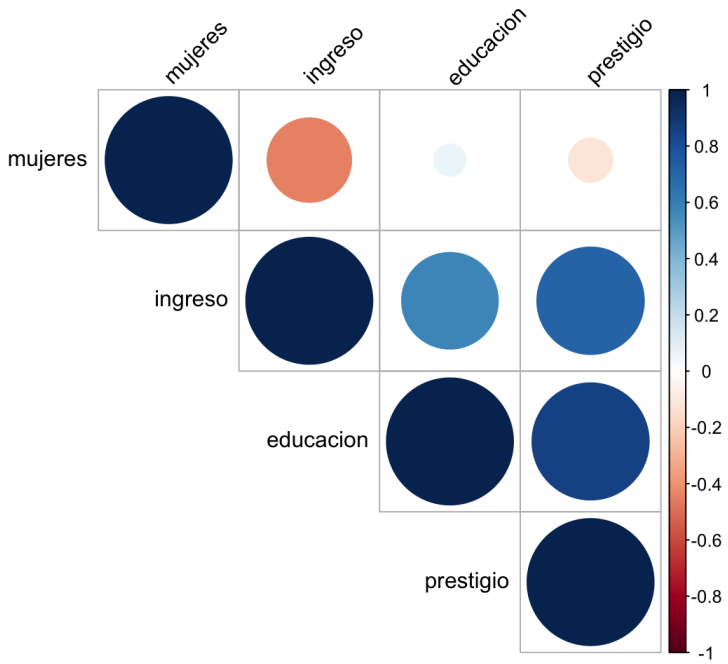
Al parecer no existe relación (lineal) entre las variables educación y mujeres.

Las variables ingreso y mujeres parecen estar relacionadas (negativamente) entre sí.

```
datos_num <- data.frame(data[c(2,3,4,5)])
cor_matrix <- cor(datos_num, method = "pearson", use = "complete.obs")
cor_matrix
```

```
##            educacion    ingreso     mujeres  prestigio
## educacion 1.00000000  0.5775802  0.06185286  0.8501769
## ingreso   0.57758023  1.0000000 -0.44105927  0.7149057
## mujeres   0.06185286 -0.4410593  1.00000000 -0.1183342
## prestigio 0.85017689  0.7149057 -0.11833419  1.0000000
```

```
corrplot(cor_matrix, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```

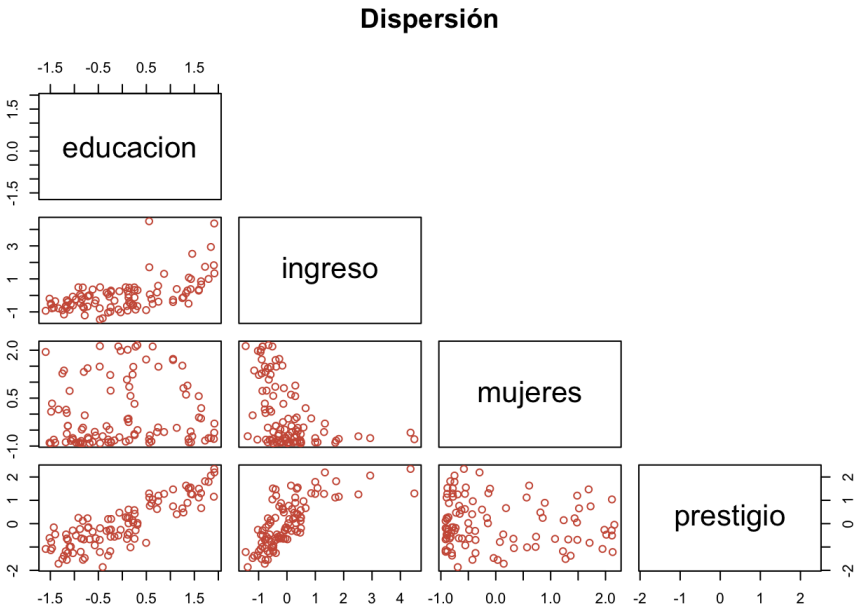


Estandarización de los datos

```
scaled_datos <- scale(datos_num)
```

Gráficas de dispersión de las variables

```
pairs(scaled_datos,main="Dispersión",col="coral3",upper.panel = NULL)
```



Análisis por Componentes Principales

A menudo es útil medir datos en términos de sus componentes principales en lugar de hacerlo en un eje x-y normal. Pero, ¿cuáles son los componentes principales? Son la estructura subyacente en los datos. Son las direcciones donde hay más variación, las direcciones donde los datos están más dispersos.

De acuerdo con wikipedia, el análisis de componentes principales (PCA) es un procedimiento estadístico que utiliza una transformación ortogonal para convertir un conjunto de observaciones de variables posiblemente correlacionadas (entidades que adquieren varios valores numéricos) en un conjunto de valores de variables linealmente no correlacionadas llamadas componentes principales. Esta transformación se define de tal manera que el primer componente principal tiene la mayor varianza posible (es decir, representa la mayor variabilidad posible en los datos), y cada componente subsiguiente a su vez tiene la mayor varianza posible bajo la restricción que es ortogonal a los componentes anteriores. Los vectores resultantes (cada uno de los cuales es una combinación lineal de las variables y contiene n observaciones) son un conjunto de bases ortogonales no correlacionadas.

```
acp<-prcomp(scaled_datos,center = TRUE, scale. = TRUE)
```

Proporción de la varianza y varianza acumulada

Con la primer componente principal se obtiene el 62% de la varianza y con las primeras 2 componentes principales se obtine el 90% de la varianza.

```
summary(acp)
```

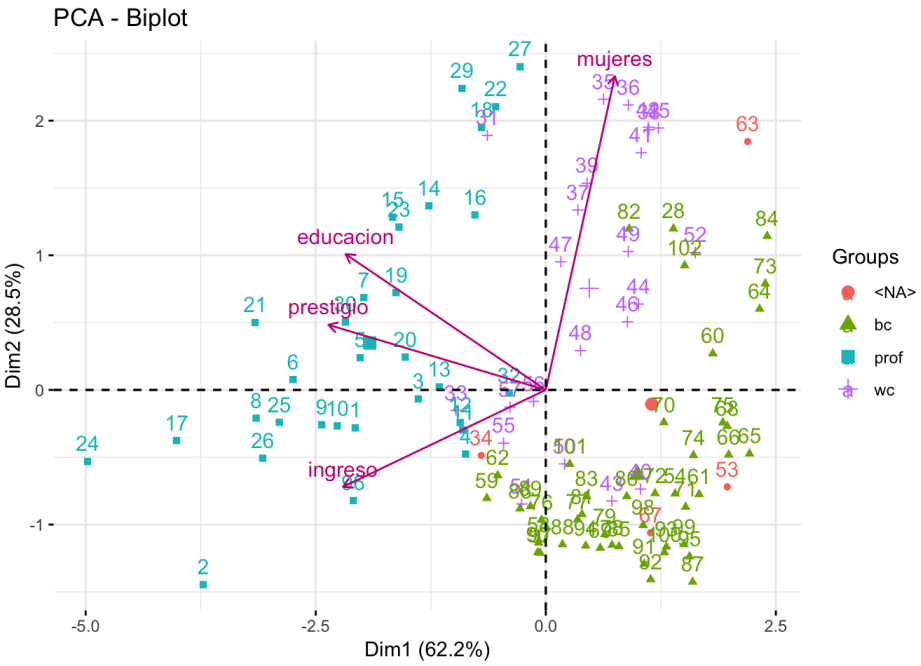
```
## Importance of components:
##                           PC1    PC2    PC3     PC4
## Standard deviation     1.5771 1.0670 0.5000 0.35279
## Proportion of Variance 0.6218 0.2846 0.0625 0.03112
## Cumulative Proportion  0.6218 0.9064 0.9689 1.00000
```

En la siguiente gráfica se muestran los datos graficados en las primeras dos componentes principales (coloreados de acuerdo al tipo), y además, se muestran los vectores de las variables consideradas. Se puede interpretar de la gráfica que:

1. Los profesionistas (pf) tienen más educación, prestigio e ingreso.
2. La mayoría de los obreros (bc) están en sentido opuesto de los vectores educación y prestigio, por lo que son ocupaciones consideradas como poco prestigiosas y para las cuales el número de años de educación es bajo.

Destacan los outliers 2 y 24. General Managers (2): En esta ocupación se tiene el ingreso máximo (25879), la educación está por arriba del promedio (12.26), el prestigio está arriba de tercer cuantil (69.1) y el porcentaje de mujeres está muy por abajo de la media (4.02). Physicians (24): Esta ocupación está muy cerca del ingreso máximo (25308), la educación también está muy cerca de la educación máxima (15.96), el porcentaje de mujeres es cercano a la media y tiene el máximo en prestigio (87.2)

```
fviz_pca_biplot(acp, repel = FALSE, col.var = "mediumvioletred",habillage=data$tipo, col.ind = "#696969")
```



Conclusión (respuestas a las preguntas 1 y 3)

Con base en el análisis anterior, se puede concluir:

1. Pudimos notar la gran relación entre el prestigio, ingreso y educación. Cabe señalar que hay casos con excepciones, donde el ingreso no está del todo relacionado con la educación, como es el caso de los gerentes, los cuales tienen el ingreso más alto, pero un nivel de educación del 12.26%, que está por debajo del tercer cuartil. Así mismo, observamos que los trabajos de obreros están en sentido opuesto a los valores de educación y prestigio.

2. Otro dato interesante es la relación negativa entre mujeres y el ingreso. Podemos notar que donde más trabajaban las mujeres fue como secretarias (97.51%) con un nivel de educación de 11.59 y como enfermeras(96.12), con un nivel de eucación de 12.46. En ambos casos notamos que recibían un salario mucho menor que el de los plomeros que contaban con un nivel de educación de 8.33 e incluso que el doe los carpinteros, con un nivel de educación de 6.92. También notamos que el salario más bajo le corresponde a las niñeras, con un nivel de educación de 9.46, que está por encima del primer cuartil de nivel edicativo.

3. Por último, notamos que el más alto nivel de prestigio lo tienen los físicos, con uno de los niveles educativos más altos y con el segundo ingreso más elevado.