Proyecto 2021

Estadísticas Ciencias de la Computación

Orientaciones metodológicas:

Fase II

Estudiantes David Guaty, Rodrigo Pino y Adrian Portales

Ejercicios

Ejercicio 1 *Realice un estudio de sus datos usando las técnicas de regresión, reducción de dimensión y de ANOVA.*

a. *Escoja las variables a las cuales les aplicara cada técnica y explique por qué.*

b. *En las técnicas que lo requieran realice el análisis de los supuestos y explique si es válida la aplicación de la técnica en esa variable.*

Objetivos

Introducción

Técnicas de Regresión

(código referente al análisis de regresión lineal en lm.R)

Por las características de las variables de los datos con los que estamos trabajando, poder predecir el *income* de un individuo a partir del resto de las variables resulta de gran interés. Por este motivo decidimos hacer una análisis y tratar de obtener un modelo de regresión lineal que nos permita hacer predicciones sobre los ingresos anuales de una persona a partir de un grupo de datos del sujeto.

Como primer paso antes de realizar cualquier análisi con los datos se eliminaron todas aquellas entradas que poseían valores faltantes, ya que al final la muestra era bastante grande y eliminar dichos *records* no representaba una pérdida considerable de los datos. De este modo evitábamos perturbaciones en los cáculos y análisis posteriores.

Antes de hacer la regresión lineal nos interesaba analizar si existía algún tipo de relación entre los datos (ya sea lineal ono). La gran mayoría de las variables son categóricas y contábamos con un muestra bastante grande (alrededor de 32000) por lo que decidimos hacer pruebas de Chi-Cuadrado de Independencia entre pares de variables categóricas, sobre todo *income* con algunas otras variables. En el caso de la educación y el sexo se obtuvo que las variables no eran independientes, lo cual tiene cierto sentido, pues se sabe que un mayor grado de educación implica mayores salarios y por el lado del sexo es bien sabido que en muchos lugares las mujeres suelen ser discriminadas y tienen salarios más bajos en comparación con los hombres en puestos equivalentes. Para el resto de combinaciones que se probaron con *income* no se pudo obtener un resultado conciso, pues se emitía una advertencia por parte de R que decía que los resultados podían no ser correctos.

Una vez hecho este pequeño análisis pasamos a la confección de los modelos de regresión lineal. Como habíamos mencinoado anteriormente la variable que resulta de mayor interés para predecir es el *income*, por lo que se intentó obtener un modelo donde dicha variable fuera la dependiente. Cuando un modelo de regresión lineal se está confeccionando, la opinión de un experto en el campo de estudio tiene muchísimo peso sobre la selección de las variables. Dado que no somos expertos decidimos usar un enfoque *back-*

ward, es decir, planteamos un modelo con todas las variables. Un punto importante a señalar aqués que muchas de las variables son categóricas y para poder hacer la regresión lineal sobre dicho conjunto de datos, estas pasan por una transformación donde se convierten en variables binarias, en dependencia del nivel de cada categoría. Lamentablemente este modelo no tuvo buenos resultados como se muestra a continuación:

```
Residual standard error: 0.3451 on
    30086 degrees of freedom
Multiple R-squared:  0.3644,
    Adjusted R-squared:  0.3629
F-statistic:   230 on 75 and 30086 DF,
     p-value: < 2.2e-16
```

Como podemos apreciar el *R-squared* es bastante bajo, por lo que incluso si este modelo cumpliera los supuestos no sería muy bueno (está muy por debajo de 0.70). Curiosamente el *p-value* es significativo, por lo que al menos podemos decir que el modelo está haciendo algo. Después de obtener dicho resultado, pasamos analizar los supuestos del modelo:

En el caso de la media y la suma de los errores, las mismas toman valores muy cercanos a 0. Sin embargo, al realizar el test de Breusch-Pagan se obtuvo que no cumplía la homocedasticidad. Por lo que al no cumplir uno de los supuestos nuestro modelo dejó de ser factible.

El siguiente paso fue quitar variables, en dependencia de la signicación de cada una, un proceso bastante engorroso dado que las variables categóricas pasaban a convertirse en variables binarias en dependencia del nivel de cada categoría. Dado que lo anterior también resultaba un poco complicado, en algunos pasos se decidió usar un criterio intuitivo, y en ocasiones se probó elimiando una variable y despues volviendola a añadir para eliminar otra en su lugar.

Siguiendo el criterio de la significación de cada variable se decidió eliminar como primera variable *native.country*. El modelo resultante obtuvo el siguiente resultado:

```
Residual standard error: 0.3452 on
    30126 degrees of freedom
Multiple R-squared:  0.3634,
    Adjusted R-squared:  0.3626
F-statistic: 491.3 on 35 and 30126 DF,
     p-value: < 2.2e-16
```

Como se puede apreciar el *R-squred* disminuyó un poco, pero no de manera significativa. Aunque, por supuesto, el mismo todavía nos indicaba la presencia de un modelo bastante malo, pero quizá cumplía con los supuestos. Sin embargo, pasaba lo mismo que el modelo inicial: no se cumplía la homocedasticidad.

En este punto decidimos seguir con el proceso de eliminación de variables. La siguiente variable fue *rece*. Los resultados fueron los siguientes:

```
Residual standard error: 0.3453 on
    30130 degrees of freedom
Multiple R-squared:  0.363,
    Adjusted R-squared:  0.3623
F-statistic: 553.8 on 31 and 30130 DF,
     p-value: < 2.2e-16
```

Se puede ver que el *R-squared* sigue disminuyendo, aunque no de manera significativa. Lamentablemente, este modelo tampoco cumplía los supuestos: fallaba en el test de homocedasticidad.

La búsqueda de un conjunto de variables independendientes continuó, pero no fuimos capaces de un encontrar un modelo útil (quizá nuestra búsqueda no fue lo suficientemente exahustiva), pues en primer lugar la tendencia a la dismución del *R-squared* seguía y ya desde el principio el valor del mismo nos indicaba que el modelo no era bueno.

Por el análisis hecho con la prueba de independencia de Chi-cuadrado, pensamos que sí podría ser posible encontrar un modelo para predecir el *income* de los individuos, sin embargo con los resultados obtenidos a través de la búsqueda del mismo a través de regresión lineal podemos concluir que si existe realmente dicha relación, la misma no debe ser lineal y sería necesario el uso de técnicas más sofisticadas.

Reducción de dimensión

TODO

ANOVA

(código referente al análisis anova en anova.R)

En esta sección se buscarán diferencias significativas entre grupos de una variable.

Por ejemplo, nos podríamos preguntar cual *workclass* tiene mayor *income*, o cual de las *ocuppation* tienen un mayor promedio de edad. Estos análisis nos

sirvirían para diferenciar características entre grupos de personas.

Una de las variables más importantes a analizar es el *income*, ya que es la variable que se quiere predecir dados los demás datos. Pero tenemos que la variable *income* es categórica(con categorías $\leq 50K$ y $> 50K$). Por lo que convertimos la variable *income* a una variable numérica donde se le asigna el valor 1 a la categoría $\leq 50K$ y el valor 2 a la categoría $> 50K$. Con esta asignación es posible realizar un anova en el cual la variable dependiente sea el *income*. Esto tiene sentido porque no nos interesa el significado exacto de la media de la variable numérica *income*, sino la diferencias significativas entre las medias de los distintos grupos analizados. Así por ejemplo podemos analizar cual *workclass* posee un mayor *income*.

Para la realización de los análisis anova tomando como variable dependiente el *income* se escogieron las variables: *education*, *ocupation*. En su mayor parte, un mayor nivel de *education* se correlaciona con un mayor porcentaje de individuos con $> 50k$ de *income*. El salario de una persona depende fuertemente de su profesión, existen profesiones que tienen un mayor porcentaje de individuos con $> 50K$ de *income*. La variable *sex* es otro buen predictor del *income*, pero la variable *sex* al tener solo dos categorías sería mejor realizar un análisis mediante t-student y no anova, ya que solo que compararían dos medias, parecido a como se hizo en la primera fase del proyecto.

Las variables escogidas son categóricas y se puede analizar que categoría presenta una diferencia significativa respecto al *income*.

Para realizar los análisis anova en R, se implementó una función auxiliar que puede imprimir dos cosas: de cumplirse los supuestos del modelo se imprime el *summary* del resultado de la función *aov*, en caso de no cumplirse los supuestos se imprime el mensaje "assumptions not fulfilled", dando a entender que el modelo no funciona.

A continuación se explica por partes el código de la función:

Funcion do_anova Parte 1

```
do_anova <- function(independent,
    dependent,name_of_independent, name_
    of_dependent){
independent <- sample(independent,
    1000)
dependent <- sample(dependent, 1000)
anova_data <- data.frame(independent,
    dependent)
anova_data <- anova_data[order(anova_
    data$independent),]
plot(dependent ~ independent, data =
    anova_data, ylab = name_of_dependent
    , xlab= name_of_independent)
```

En esta parte se leen dos vectores: independent y dependent. Se toma una muestra de 1000 elementos de ambos. Se conforma un data frame y además se ordena el data frame por la variable independendiente categórica, así el data frame queda estructurado como fue visto en conferencia para la correcta utilización del anova. Además, se grafican las distintas categorías en un gráfico de caja para analizar gráficamente si existen diferencias.

Funcion do_anova Parte 2

```
result <- aov(dependent ~ independent,
    data = anova_data)
res <- result$residuals
```

En esta parte se lleva a cabo el anova, se guardan en la variable *res* los residuos para comprobar los supuestos.

Funcion do_anova Parte 3

```
is_model_ok = TRUE
stest <- shapiro.test(res)
if(stest$p.value < 0.05){
        is_model_ok = FALSE
}

btest <- bartlett.test(res, anova_data$
    independent)
if(btest$p.value < 0.05){
        is_model_ok = FALSE
}
dtest <- dwtest(result)
if(dtest$p.value < 0.05){
        is_model_ok = FALSE
}
if(is_model_ok){
        summary(result)
}
else{
        print("assumptions not
            fulfilled")
```

```
}
```

En esta parte se hacen todas las pruebas para los supuestos. Si alguna prueba falla entonces no se cumplen los supuestos del modelo y se imprime el mensaje "assumptions not fulfilled". En caso de cumplirse los supuestos se imprime un resumen del anova.

Income ∼ education

```
> do_anova(data$education, data$income,
    "education", "income")
[1] "residuals do not have a normal
    distribution"
[1] "residuals are not homogeneous"
[1] "assumptions not fulfilled"
```

Por lo que el modelo no cumple los supuestos.

Income ∼ occupation

```
> do_anova(data$occupation, data$income
    , "occupation", "income")
[1] "residuals do not have a normal
    distribution"
[1] "assumptions not fulfilled"
```

Por lo que el modelo no cumple los supuestos.

En las variables escogidas no se cumple el modelo. Veamos otras variables.

Income ∼ workclass

```
> do_anova(data$workclass, data$age, "
    workclass", "age")
[1] "residuals do not have a normal
    distribution"
[1] "assumptions not fulfilled"
```

Por lo que el modelo no cumple los supuestos.

Income ∼ marital.status

```
> do_anova(data$marital.status, data$
    income, "marital.status", "income")
[1] "residuals do not have a normal
    distribution"
[1] "assumptions not fulfilled"
```

Por lo que el modelo no cumple los supuestos.

Por lo que no se puede aplicar Anova, al menos con la variable *income*.

Pasemos a analizar otros pares de variables como *workclass* y *age*, podríamos ver cual es el grupo de trabajadores con mayor o menor edad.

Age ∼ workclass

```
> do_anova(data$workclass, data$age, "
    workclass", "age")
[1] "residuals do not have a normal
    distribution"
[1] "assumptions not fulfilled"
```

Por lo que el modelo no cumple los supuestos.

Tratemos otro modelo con age. Seleccionamos la variable *occupation*. Puede ser interesante saber la ocupación que tiene más empleados de una edad mayor.

Age ∼ occupation

```
> do_anova(data$occupation, data$age, "
    occupation", "age")
[1] "residuals do not have a normal
    distribution"
[1] "assumptions not fulfilled"
```

Por lo que el modelo no cumple los supuestos.

Tratemos de analizar si existe alguna diferencia entre el promedio de la ganancia capital entre los individuos de diferentes ocupaciones.

capital-gain ∼ occupation

```
> do_anova(data$occupation, data$
    capital.gain, "occupation", "capital
    .gain")
[1] "residuals do not have a normal
    distribution"
[1] "residuals are not homogeneous"
[1] "assumptions not fulfilled"
```

Por lo que el modelo no cumple los supuestos.

Otro caso de interés es saber si existe diferencia significativa entre el promedio de las horas de trabajo semanales entre los individuos de diferentes profesiones.

hours-per-week ∼ occupation

```
> do_anova(data$occupation, data$hours.
    per.week, "occupation", "capital.
    gain")
[1] "residuals do not have a normal
    distribution"
[1] "residuals are not homogeneous"
[1] "assumptions not fulfilled"
```

Por lo que el modelo no cumple los supuestos.

Para no hacer más extensa la sección, se mencionará que también se comprobó el anova con otros pares de variables. Digase: native-countre∼education.num, native-countre∼hours-per-week, race∼hours-per-week, entre otras. En todas ellas los modelos no cumplían los supuestos.

Por lo que se puede afirmar con cierta seguridad que no se puede aplicar técnicas de anova a los datos. Al menos no se encontró una forma de aplicarlas :-(. Los investigadores están tristes.

Conclusiones

TODO

Contribuciones de cada integrante

TODO