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

- TODO

Introducción

TODO

Técnicas de Regresión

TODO

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