UNIVERSIDAD NACIONAL DE COLOMBIA

REGRESIÓN LINEAL MULTIPLE PARTE 2

Autor:

Daniela Pico

Jhonatan Smith

Profesor:

Isabel Cristina Ramirez

2021-01

```
data=read.table(file.choose(),header=T,sep=";",dec=",",
colClasses=c(rep("numeric",7),"factor",rep("numeric",3),"factor"))
datar=data[-c(19:20),]
```

Sin considerar las observaciones con ID=47 e ID=112 de la base de datos asignada y usando variables indicadoras R1,R2,R3 para las regiones 1,2 y 3 respectivamente, suponga inicialmente que las rectas de regresión de DPERM VS. PDP en cada región no son iguales (que difieren tanto en el intercepto como en las pedientes) realice lo siguiente:

1) Plantee el modelo de regresión apropiado si se espera una diferencia entre las rectas de DPERM VS PDP que corresponden a las cuatro regiones.
2) Ajuste el modelo general (muestre la tabla de parámetros estimados) y halle las ecuaciones ajustadas de las rectas en cada región.
3) Analice supuestos de normalidad y varianza constante mediante los residuales, para el modelo general (residuales estudentizados vs. valores ajustados y vs. PDP). Identifique en estos gráficos las observaciones según la región a la cual pertenecen.
4) Determine si existe diferencia entre las ordenadas en el origen de las rectas correspondientes a las regiones.
5) Determinee si existe diferencia en las pendientes de las rectas correspondientes a las regiones. Interprete a la luz de los datos.
6) Si se quiere probar que la recta de DPERM VS PDP es diferente para cada región, plantee la hipoteis a probar, el estadítico de prueba y región crítica a nivel de 0.05, realice la prueba y concluya.
7) Determine si el efecto medio de PDP sobre DPERM es igual a las cuatro regiones (no depende de la región).

Nota: Aquí el estadístico de prueba se calcula recordando la siguiente expresión.

$$F = \frac{SSR(ModeloCompleto) - SSR(Modeloreducido)}{glssr(ModeloCompleto) - glssr(Modeloreducido)} \div MSE(ModeloCompleto)$$

Donde el modelo completo es el modelo 1) y el modelo reducido es el resultante de aplicar lo que dice H_0 acerca de las pendientes en las regiones.

H_0= pendiente región 1=pendiente región 2=pendiente región 3=pendiente región 4.

H_1= Alguna de las pendientes es distinta.

Traduzca estas hipotesis en términos de los parámetros apropiados en el mmodelo. Ajuste el modelo reducido, muestre la tabla de párametros ajustados y escriba las ecuaciones de ajuste para cada región. Interprete los resultados a la luz de los datos.

Resultados

1) Se desea modelar la relación lineal de DPERM VS PDP (variable predictoria cuantativa), en presencia de REGION, esta última variable cualitativa con 4 categorias (R1,R2,R3 y R4), Para este caso usaremos las indicadoras de las primeras 3 categorias, con:

- DPERM:Longitud de permanencia.
- REGION:Región.
- PDP: Censo promedio diario.
- R1= Región 1.
- R2= Región 2.
- R3= Región 3.

CASO 1: El efecto promedio del censo promedio diario sobre la respuesta de la longitud de permanencia cambia según la categoria en que la régión es observada, se plantea el siguiente modelo y usaremos como referencia la categoría Región 4 (R4)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \beta_4 I_{i3} + \beta_{1,1} X_{i1} I_{i1} + \beta_{1,2} X_{i1} I_{i2} + \beta_{1,3} X_{i1} I_{i3} + ei$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

Esta ecuación define 4 rectas de regresión simple de DPERM VS PDP

- Si I_1=1 entonces I_2=0,I_3=0

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i_1} + \beta_2 + \beta_{1,1} X_{i1} + ei \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1}) X_{i1} \\ &\varepsilon_i \sim N(0, \sigma^2) \end{aligned}$$

- Si I_2=1 entonces I_1=0,I_3=0

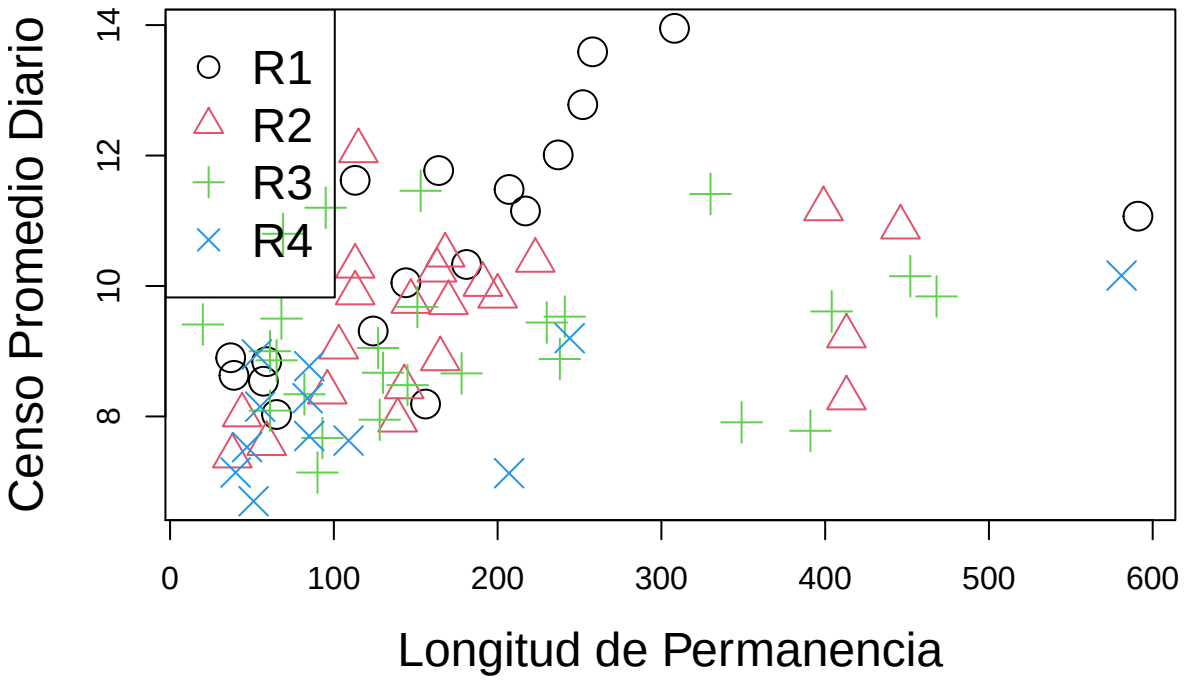
$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_3 + \beta_{1,2} X_{i1} + ei \\ &= (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2}) X_{i1} \\ &\varepsilon_i \sim N(0, \sigma^2) \end{aligned}$$

* Si I_3=1 entonces I_2=0,I_1=0

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_4 + \beta_{1,3} X_{i1} + ei \\ &\varepsilon_i \sim N(0, \sigma^2) \\ &= (\beta_0 + \beta_4) + (\beta_1 + \beta_{1,3}) X_{i1} \\ &\varepsilon_i \sim N(0, \sigma^2) \end{aligned}$$

- Si I_1=0,I_2=0,I_3=0

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + ei \\ &\varepsilon_i \sim N(0, \sigma^2) \end{aligned}$$



Un primer analisis descriptivo del grafico anterior podria indicar una relacion lineal con respecto a R1, ya que si la longitud de permanencia aumenta, el censo promedio diario tambien lo hace en relacion a la region 1.

Se podria pensar que la region 2 (representada por los triangulos) tambien tiene una relacion lineal pues se cunmple lo mencionado en las lineas anterioires, sin embargo se obvserva que sus datos se encuentran mas dispersos y esto finalmente, podria afectar la pendiente del modelo.

Las regiones 3 y 4 (simbolos "+" en verde y simbolos "x" en azul) en un principio se podria pensar que "no tienen pendiente" y esto se interpreta como una carencia de relacion entre las variables epxlicatorias y las regiones 3 y 4.

En resumen; a primera vista se podría sospechar que la variable longitud de permanencia y censo promedio diario tienen una relacion lineal entre las regiones 1 y 2. Para las regiones 3 y 4 no se puede dar indicios claros y se sospecha que no son relevantes.

Por esto se podria pensar que se ha de tener en cuenta que la longitud de permanencia se ve afectada por la variable PDP dadas las regiones 1 y 2.

2) A continuación se presenta la tabla de parámetros estimados

```
##
## Call:
## lm(formula = DPERM ~ PDP * REGION)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.97056 -0.79358 -0.02592  0.65684  2.86859
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  7.5129900  0.4830879  15.552   <2e-16 ***
## PDP          0.0043779  0.0024015   1.823   0.0726 .
## REGION1      1.5562908  0.6923635   2.248   0.0277 *
## REGION2      1.2701019  0.6822809   1.862   0.0669 .
## REGION3      1.4118256  0.6380484   2.213   0.0302 *
## PDP:REGION1  0.0040337  0.0032976   1.223   0.2253
## PDP:REGION2 -0.0007404  0.0032513  -0.228   0.8205
## PDP:REGION3 -0.0030364  0.0030229  -1.004   0.3186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.227 on 70 degrees of freedom
## Multiple R-squared:  0.4212, Adjusted R-squared:  0.3633
## F-statistic: 7.277 on 7 and 70 DF,  p-value: 1.523e-06
```

Se Ajusto el siguiente modelo teniendo en como región de referencia R4 (Región 4) y se llegó al siguiente modelo ajustado:

$$\hat{y_i} = 7.5129900 + 0.0043779X_{i1} + 1.5562908I_{i1}$$
$$+1.2701019I_{i2} + 1.4118256I_{i3} + 0.0040337X_{i1}I_{i1} - 0.0007404X_{i1}I_{i2} - 0.0030364X_{i1}I_{i3}$$

Cuando I_{ii} hace referencia a: I_{i1}=R1,I_{i2}=R2,I_{i3}=R3

- Recta ajustada para la Región 1, Si I_1=1 entonces I_2=0,I_3=0

$$\begin{aligned}\hat{y_i} &= 7.5129900 + 0.0043779X_{i_1} + 1.5562908 + 0.0040337X_{i1} \\ &= (7.5129900 + 1.5562908) + (0.0043779 + 0.0040337)X_{i1}\end{aligned}$$

* Recta ajustada para la Región 2, Si I_2=1 entonces I_1=0,I_3=0

$$\begin{aligned}\hat{y_i} &= 7.5129900 + 0.0043779X_{i1} + 1.2701019 - 0.0007404X_{i1} \\ &= (7.5129900 + 1.2701019) + (0.0043779 - 0.0007404)X_{i1} \\ &\varepsilon_i \sim N(0, \sigma^2)\end{aligned}$$

- Recta ajustada para la Región 3, Si I_1=0,I_2=0,I_3=1

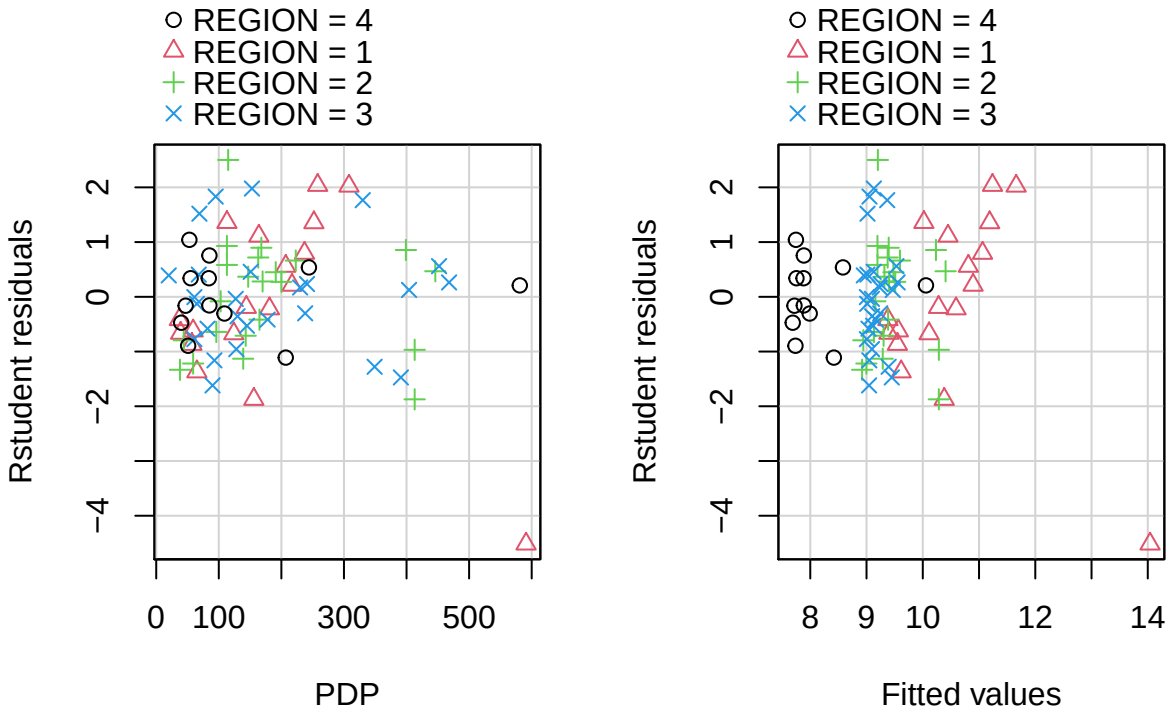
$$\begin{aligned}\hat{y_i} &= 7.5129900 + 0.0043779X_{i1} + 1.4118256 - 0.0030364X_{i1} \\ &= (7.5129900 + 1.4118256) + (0.0043779 - 0.0030364)X_{i1}\end{aligned}$$

- Recta ajustada para la Región 4, Si I_1=0,I_2=0,I_3=0

$$\hat{y_i} = 7.5129900 + 0.0043779X_{i1}$$

3)

Gráfico de Residuales vs. Valores Ajustados y vs.PDP

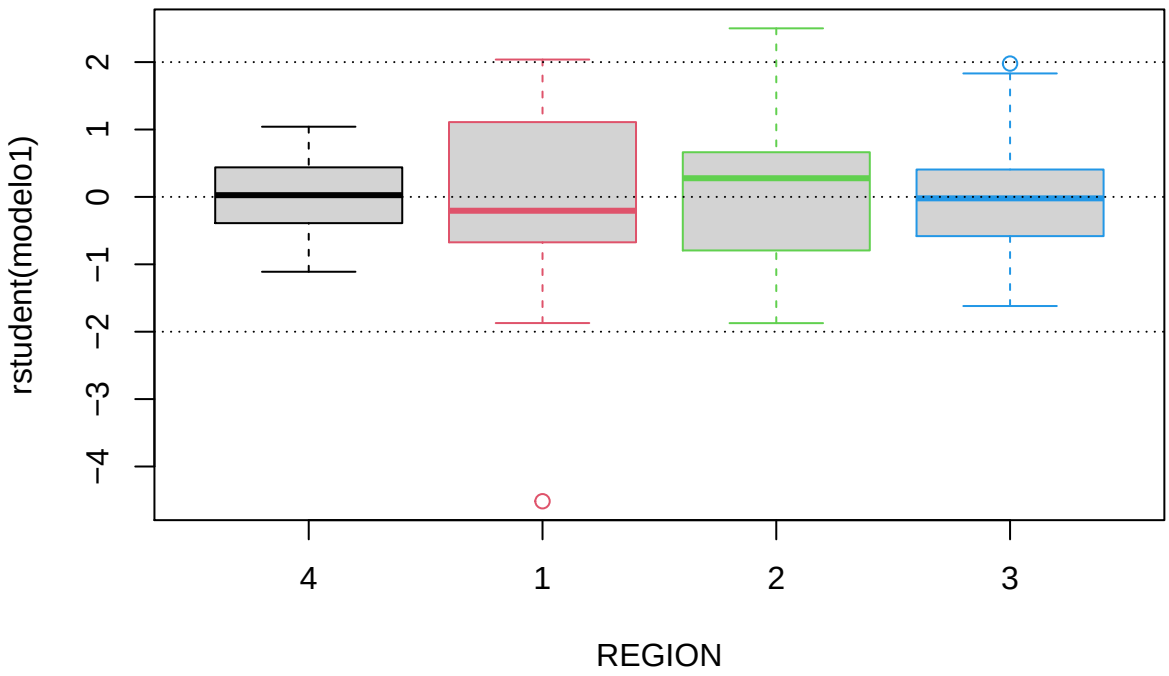


Los residuales estudentizados entre aproximadamente 0 y 150 (valores de PDP) se mueven mas o menos entre los mismos valores; entre -2 y 2, alrededor del cero.

Sin embargo, a partir de poco mas 300 dichos valores se mueven mas o menos entre -2 y 1. Si bien estos datos mas a la derecha no representan una gran parte de los datos, si se puede llegar a pensar que por esto hay problemas de varianza constante.

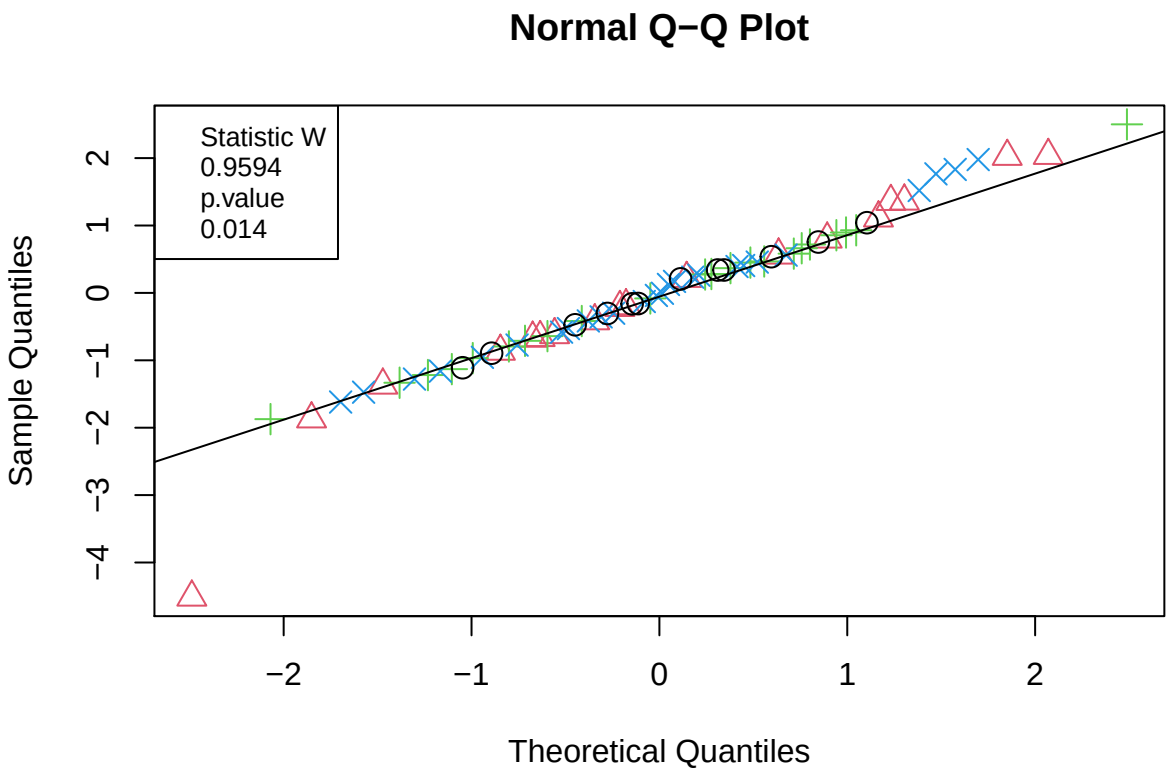
Al ver los residuales estudentizados vs los valores ajustados vemos como al inicio los valores estan mas concentrados (mas concentrados alrededor del cero) pero a medida que aumenta el valor ajustado, los valores se hacen mas amplios en el eje y. Esto nos ayuda a confirmar que hay problemas de varianza constante.

```
boxplot(rstudent(modelo1)~REGION,border=1:4)
abline(h=c(-2,0,2),lty=3)
```



Los colores siguen representando las regiones dadas en el grafico de dispersion. Claramente el boxplot color negro representa una region con datos mas concentrados (Region 1) mientras que la region 2 (boxplot rojo) claramente tiene una variabilidad mas alta. Se concluye que hay problemas de varianza constante.

Gráfico de Normalidad



En este caso, el p-value= $0.014 < \alpha = 0.05$ por tanto NO se rechaza la hipotesis nula y se concluye que los residuales NO distribuyen normal.

Este modelo no cumple los supuestos de varianza constante y normalidad.

4) Si los interceptos de las rectas para la R1,R2,R3 son igueles entonces:

$$\beta_0 = \beta_0 + \beta_2 = \beta_0 + \beta_3 = \beta_0 + \beta_4$$

Se plantea el siguiente juego de Hipotesis:

$$H_0 = \beta_2 = \beta_3 = \beta_4 = 0$$

vs $H_a : Algún \beta_j \neq 0$

```
## Linear hypothesis test
##
## Hypothesis:
## REGION1 = 0
## REGION2 = 0
## REGION3 = 0
##
## Model 1: restricted model
## Model 2: DPERM ~ PDP * REGION
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     73 115.54
## 2     70 105.44  3    10.106 2.2365 0.09156 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como P-valor=0.09156>0.05 entonces no se rechaza la hipotesis nula y se concluye que los interceptos de las rectas para la región 1, región 2 y región 3 son iguales

5) Para determinar si existe diferencia entre las pendientes de las rectas, se plantea el siguiente juego de hipotesis partiendo de que:

$\beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2} = \beta_1 + \beta_{1,3} = \beta_1$ Dadas estas pendientes, se desea probar que:

$H_0 : \beta_{1,1} = \beta_{1,2} = \beta_{1,3}$ vs H_1 : Algún $\beta_j \neq 0$.

```
names(coef(modelo1))
```

```
## [1] "(Intercept)" "PDP"         "REGION1"     "REGION2"     "REGION3"
## [6] "PDP:REGION1" "PDP:REGION2" "PDP:REGION3"
```

```
linearHypothesis(modelo1,c("PDP:REGION1=0","PDP:REGION2= 0","PDP:REGION3=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## PDP:REGION1 = 0
## PDP:REGION2 = 0
## PDP:REGION3 = 0
##
## Model 1: restricted model
## Model 2: DPERM ~ PDP * REGION
##
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     73 114.41
## 2     70 105.44  3    8.9697 1.985 0.1241
```

Segun esto, las pendientes son iguales

$H_0 : \beta_{1,1} = \beta_{1,2}$ vs H_1 :Algún $\beta_j \neq 0$.

```
linearHypothesis(modelo1,c("PDP:REGION1-PDP:REGION2= 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## PDP:REGION1 - PDP:REGION2 = 0
##
## Model 1: restricted model
## Model 2: DPERM ~ PDP * REGION
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     71 108.90
## 2     70 105.44  1    3.4642 2.2999 0.1339
```

Como el P-value $= 0.1339 > \alpha = 0.05$ no se rechaza la hipotesis nula y se concluye que la incidencia de la region 1 y 2 es la misma pues la pendiente de sus respectivas rectas es igual.

```
linearHypothesis(modelo1,c("PDP:REGION1-PDP:REGION3= 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
```

```
## PDP:REGION1 - PDP:REGION3 = 0
##
## Model 1: restricted model
## Model 2: DPERM ~ PDP * REGION
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     71 114.32
## 2     70 105.44  1    8.8812 5.8962 0.01775 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como el P-value $= 0.01775 > \alpha = 0.05$ no se rechaza la hipotesis nula y se concluye que la incidencia de la region 1 y 3 es la misma pues la pendiente de sus respectivas rectas es igual.

```
linearHypothesis(modelo1,c("PDP:REGION2-PDP:REGION3= 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## PDP:REGION2 - PDP:REGION3 = 0
##
## Model 1: restricted model
## Model 2: DPERM ~ PDP * REGION
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     71 106.41
## 2     70 105.44  1   0.97141 0.6449 0.4247
```

```
# p=0.5493 > alfa no se rechaza por tanto son iguales
```

Como el P-value $= 0.4247 > \alpha = 0.05$ se rechaza la hipotesis nula y se concluye que la incidencia de la region 2 y 3 es diferente pues las pendientes de cada recta es diferente.

Asi, la pendiente de REGION 1 y REGION 2 es la misma pero la recta correspondiente a la region 3 es diferente. Esto es concistente con el primer analisis descriptivo.

6) Se desea probar que la recta de regresion DPERM vs PDP es diferente en cada region.

Para probar que cada recta sea diferente, se debe verificar que tenga pendiente e intercepto diferente. Para ello, se plantea el siguiente juego de hipotesis.

$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_{1,1} = \beta_{1,2} = \beta_{1,3} = 0$ vs H_1 :Algún parametro anterior diferente de cero.

```
names(coef(modelo1))
```

```
## [1] "(Intercept)" "PDP"         "REGION1"     "REGION2"     "REGION3"
## [6] "PDP:REGION1" "PDP:REGION2" "PDP:REGION3"
```

```
linearHypothesis(modelo1,c("PDP:REGION2=0","PDP:REGION3= 0","PDP:REGION1=0","REGION1=0","REGION2=0","RE
```

```
## Linear hypothesis test
##
## Hypothesis:
## PDP:REGION2 = 0
## PDP:REGION3 = 0
## PDP:REGION1 = 0
## REGION1 = 0
## REGION2 = 0
## REGION3 = 0
##
```

```
## Model 1: restricted model
## Model 2: DPERM ~ PDP * REGION
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     76 155.51
## 2     70 105.44  6    50.072 5.5404 9.619e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El valor del p-value es casi cero, p-value$>\alpha = 0.05$ por tanto no se rechaza la hipotesis nula y se concluye que las rectas son iguales y no existe un cambio diferente por pendiente e intercepto.

7)

```
matriz_diseño = as.data.frame(model.matrix(modelo1))
```

```
REGION1 = matriz_diseño$REGION1
REGION2 = matriz_diseño$REGION2
REGION3 = matriz_diseño$REGION3
```

Ajuste modelo reducido: