Regional differences in sex and origin, on work discrimination in Argentina 2019

Anexo: sintaxis de operaciones

Eduardo Bologna

06 octubre, 2020

Los datos

Lectura de la base, se toman todas las variables como factores para conservar los cinco dígitos en CNO

```
ecetss<-read.table("ECETSS_ocupados.csv",
                   header = TRUE, sep= ",",
                   colClasses="factor")
```

Componentes de los índices:

Seguridad en el empleo

seguridad_normalizada Opción para el índice_2: temporalidad + obra social: C1P2.6 1 Permanente, tiene trabajo durante todo el año o de manera continua 2 De temporada o estacional 3 Intermitente (no de temporada o estacionario)

```
table(ecetss$C1P2.6)
```

```
##
##    1    2    3
## 7112  739 1115
```

Cruzada con la combinación de estas dos:
C2P4.2 ¿Usted tiene obra social? "Asalariados (cat_ocup = 3)"
1 Sí
2 No
99 Ns./Nc.

C2BP4.1
¿Usted tiene obra social? "Independientes (cat_ocup = 1 o 2)"
1 Sí
2 No
99 Ns./Nc.

```
addmargins(table(ecetss$C2P4.2))
```

```
##
##         1    2   99  Sum
## 2548 4619 1798    1 8966
```

```
levels(ecetss$C2P4.2)<-c(NA, 1, 2, NA)
ecetss$C2P4.2<-factor(ecetss$C2P4.2)

levels(ecetss$C2BP4.1)<-c(NA, 1, 2, NA)
ecetss$C2BP4.1<-factor(ecetss$C2BP4.1)
addmargins(table(ecetss$C2BP4.1))
```

```
##
##    1    2  Sum
## 1072 1474 2546
```

```
ecetss$obra_social<-ifelse(ecetss$cat_ocup==3,
                           ecetss$C2P4.2, ecetss$C2BP4.1)

kable(addmargins(table(ecetss$obra_social, ecetss$C2P4.2)))
```

	1	2	Sum
1	4619	0	4619
2	0	1798	1798
Sum	4619	1798	6417

```
kable(addmargins(table(ecetss$obra_social, ecetss$C2BP4.1)))
```

	1	2	Sum
1	1072	0	1072
2	0	1474	1474
Sum	1072	1474	2546

```
kable(addmargins(table(ecetss$C1P2.6, ecetss$C2BP4.1)))
```

	1	2	Sum
1	831	811	1642
2	72	192	264
3	169	471	640
Sum	1072	1474	2546

La variable seguridad tiene seis categorías que van de desde 1= intermitente sin obra social, hasta 6=estable con obra social

```
table(ecetss$C1P2.6,ecetss$obra_social)
```

```
##
##        1    2
##   1 5178 1933
##   2  244  494
##   3  269  845
```

```r
ecetss$seguridad<-ifelse(
  ecetss$C1P2.6==1 & ecetss$obra_social==1, 6, ifelse(
    ecetss$C1P2.6==1 & ecetss$obra_social==2, 5, ifelse(
      ecetss$C1P2.6==2 & ecetss$obra_social==1, 4,ifelse(
        ecetss$C1P2.6==2 & ecetss$obra_social==2, 3, ifelse(
          ecetss$C1P2.6==3 & ecetss$obra_social==1, 2, 1
          )
        )
      )
    ))

summary(ecetss$seguridad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   5.000   6.000   4.973   6.000   6.000       3
```

```r
# verificación
table(ecetss$obra_social, ecetss$seguridad, ecetss$C1P2.6)
```

```
## , ,  = 1
##
##
##        1    2    3    4    5    6
##   1    0    0    0    0    0 5178
##   2    0    0    0    0 1933    0
##
## , ,  = 2
##
##
##        1    2    3    4    5    6
##   1    0    0    0  244    0    0
##   2    0    0  494    0    0    0
##
## , ,  = 3
##
##
##        1    2    3    4    5    6
##   1    0  269    0    0    0    0
##   2  845    0    0    0    0    0
```

```r
# de los permanentes, hay con seguridad 6 (tienen obra social) y 5 (no la tienen y así los demás)
```

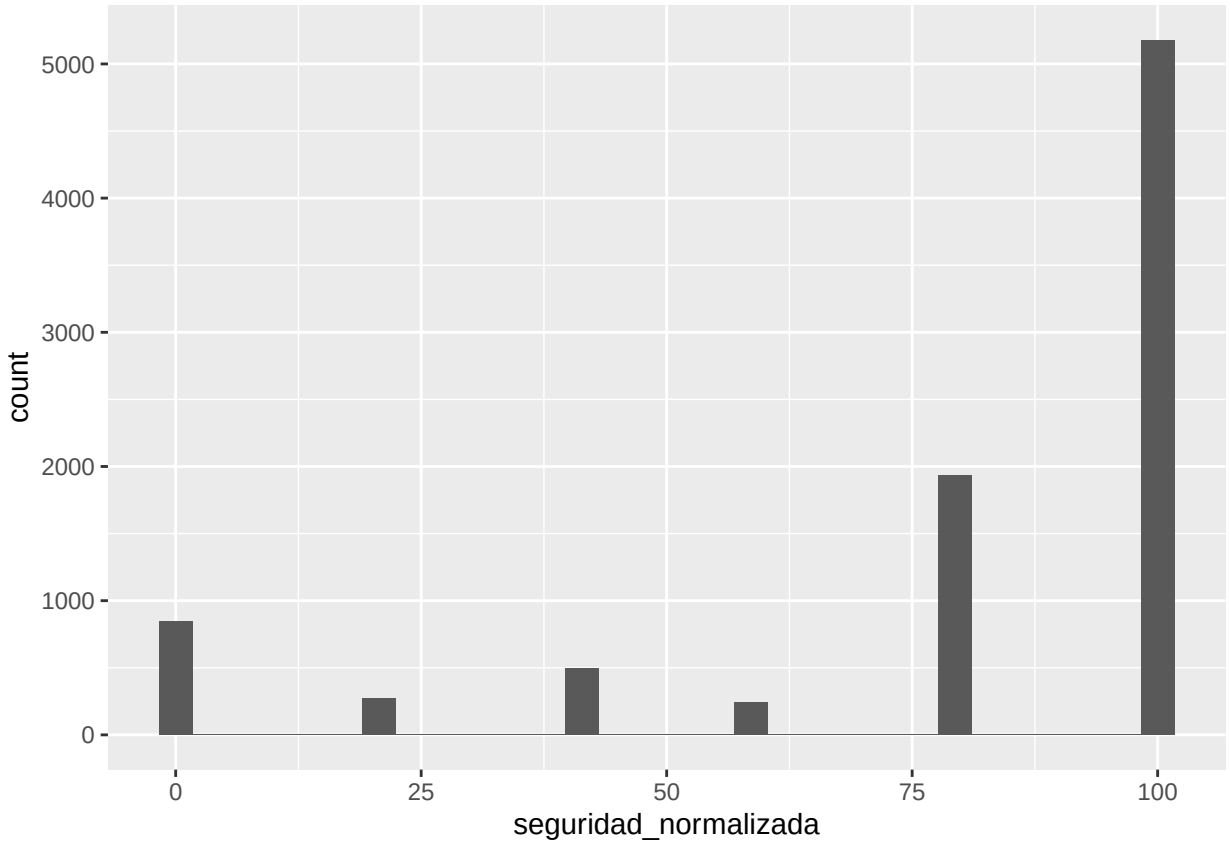
Se estandariza:

```r
ecetss$seguridad_normalizada<-
  100*(ecetss$seguridad-min(ecetss$seguridad, na.rm = TRUE))/(
    max(ecetss$seguridad, na.rm = TRUE)-min(ecetss$seguridad, na.rm = TRUE))

summary(ecetss$seguridad_normalizada)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   80.00  100.00   79.46  100.00  100.00       3
```

```r
ggplot(ecetss)+geom_histogram(aes(seguridad_normalizada))
```



Consistencia educación - calificación

Calificación ocupacional

Se extrae el quinto dígito de CNO, se eliminan los casos no válidos, se rotula e invierte su codificación. Luego se lo trata como numérico.

```r
ecetss$calif.ocup= substr(ecetss$ocupa_cno, 5,5)

ecetss$calif.ocup[ecetss$calif.ocup==8]<-NA
ecetss$calif.ocup[ecetss$calif.ocup==9]<-NA
ecetss$calif.ocup<-factor(ecetss$calif.ocup)

levels(ecetss$calif.ocup)=c(
  "profesional", "técnica", "operativa", "no calificada")

ecetss$calif.ocup=factor(ecetss$calif.ocup,
                         levels(
                           factor(
                             ecetss$calif.ocup))[c(4,3,2,1)])

ecetss$calif.ocup_num<-as.numeric(ecetss$calif.ocup)
```

```r
# Verificación
table(ecetss$calif.ocup, ecetss$calif.ocup_num)
```

```
##
##                    1    2    3    4
##   no calificada 1897    0    0    0
##   operativa        0 4460    0    0
##   técnica          0    0 1581    0
##   profesional      0    0    0 1019
```

Educación

Se eliminan los valores perdidos, se trata como numérica

```r
ecetss$nivel_ed[ecetss$nivel_ed==99]<-NA
ecetss$nivel_ed<-factor(ecetss$nivel_ed)

ecetss$nivel_ed_num<-as.numeric(as.character(ecetss$nivel_ed))

# verificación
table(ecetss$nivel_ed, ecetss$nivel_ed_num)
```

```
##
##         0    1    2    3    4    5    6    7    8    9   10
##   0    36    0    0    0    0    0    0    0    0    0    0
##   1     0  404    0    0    0    0    0    0    0    0    0
##   10    0    0    0    0    0    0    0    0    0    0  194
##   2     0    0 1155    0    0    0    0    0    0    0    0
##   3     0    0    0 1571    0    0    0    0    0    0    0
##   4     0    0    0    0 2353    0    0    0    0    0    0
##   5     0    0    0    0    0  455    0    0    0    0    0
##   6     0    0    0    0    0    0 1049    0    0    0    0
##   7     0    0    0    0    0    0    0  687    0    0    0
##   8     0    0    0    0    0    0    0    0 1002    0    0
##   9     0    0    0    0    0    0    0    0    0   48    0
```

Inconsistencia

inconsistencia: más alto más inconsistencia

```r
ecetss$inconsistencia<-ecetss$nivel_ed_num/ecetss$calif.ocup_num
summary(ecetss$inconsistencia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.500   2.000   2.236   3.000   8.000      21
```

Se ajusta el signo: *consistencia_normalizado*

```
ecetss$consistencia_normalizado<-100*(ecetss$inconsistencia-max(ecetss$inconsistencia, na.rm = TRUE))/(
```

Ingresos - hora

Ingresos

Se lo trata como numérico y se retienen de la base solo los casos con ingreso mayor a cero y menor al percentil 99

```
ecetss$ingreso_op_num<-as.numeric(as.character(ecetss$ingreso_op))
summary(ecetss$ingreso_op_num)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     -99    4000   12000   15689   22000  600000
```

```
ecetss<-subset(ecetss, ecetss$ingreso_op_num>0 &
  ecetss$ingreso_op_num<quantile(ecetss$ingreso_op_num, .99))
```

Horas

Se eliminan dos casos con 24/24, 7/7 = 168 horas

```
ecetss$horas_ocup_ppal<-as.numeric(as.character(ecetss$hs_sem_ref))
table(ecetss$horas_ocup_ppal)
```

```
##
## -99   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##  36 104  13  37  45  62  44  80  28  93  55  76  26 138  21  54 125  73  24  73
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38
##  20 444  39  45  19 240 242  20  29  83  32 555  26 107  29  62 319 161  34  38
##  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
##  35 958  26 100  36 419 397  75  33 507  55 158  35  68  30 101  59 133   9  24
##  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78
##  13 183   7  10  24  29  19  31   3  20   9  44   6  75   3   3   4  10  13  16
##  79  80  81  82  83  84  85  86  87  88  90  91  92  93  94  96  98 102 104 105
##   3  10   2   5   1  79   2   2   1   3   4   6   1   1   2  10   7   1   1   9
## 108 109 112 114 116 118 120 128 132 135 144 168
##   1   1   7   1   1   1   1   1   1   1   1   2
```

```
summary(ecetss$horas_ocup_ppal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -99.0    24.0    40.0    35.9    46.0   168.0
```

```
ecetss<-subset(ecetss, ecetss$horas_ocup_ppal>0 &
                 ecetss$horas_ocup_ppal<168)
```

Ingresos por hora semanal

```r
ecetss$ing_hora<-ecetss$ingreso_op_num/(4*ecetss$horas_ocup_ppal)
summary(ecetss$ing_hora)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.581   59.028  100.000  127.393  163.880 3750.000
```

```r
ecetss$ing_hora_bruto_normalizado<-
  100*(
    ecetss$ing_hora-min(ecetss$ing_hora, na.rm = TRUE))/(
      max(ecetss$ing_hora, na.rm = TRUE)-min(
        ecetss$ing_hora, na.rm = TRUE))

ecetss$ing_hora_bruto_normalizado_log<-
  100*(
    log(ecetss$ing_hora)-log(min(
      ecetss$ing_hora, na.rm = TRUE)))/(log(max(
        ecetss$ing_hora, na.rm = TRUE))-log(min(
          ecetss$ing_hora, na.rm = TRUE)))

summary(ecetss$ing_hora_bruto_normalizado)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.559   2.652   3.382   4.355 100.000
```

Ajuste ingresos

Considerando que el ingreso tiene un valor relativo al lugar de residencia, se transforman los ingresos/hora de la ocupación principal en puntajes z, con las medias y desviaciones estándar de cada región.

1. Se construyen vectores que contienen medias y desviaciones estándar por región.

```r
regiones<-c(10, 40:44)
ingresos_hora_medios_region<-vector(length = 6)

for (j in 1:6) {
  ingresos_hora_medios_region[[j]]<-
    mean(subset(ecetss, ecetss$region==regiones[j])$ing_hora)
}

desviaciones_ingresos_hora_region<-vector(length = 6)

for (j in 1:6) {
  desviaciones_ingresos_hora_region[[j]]<-
    sd(subset(ecetss, ecetss$region==regiones[j])$ing_hora)
}
```

2. Se estandarizan los ingresos hora en torno a la media y desviación propias de cada región.

```
ecetss$z_ingreso_hora<-
  ifelse(
    ecetss$region==10,
    (ecetss$ing_hora-ingresos_hora_medios_region[1])/
      desviaciones_ingresos_hora_region[1],
    ifelse(
      ecetss$region==40,
      (ecetss$ing_hora-ingresos_hora_medios_region[2])/
        desviaciones_ingresos_hora_region[2],
      ifelse(
        ecetss$region==41,
        (ecetss$ing_hora-ingresos_hora_medios_region[3])/
          desviaciones_ingresos_hora_region[3],
        ifelse(ecetss$region==42,
               (ecetss$ing_hora-ingresos_hora_medios_region[4])/
                 desviaciones_ingresos_hora_region[4],
          ifelse(
            ecetss$region==43,
            (ecetss$ing_hora-ingresos_hora_medios_region[5])/
              desviaciones_ingresos_hora_region[5],
            (ecetss$ing_hora-ingresos_hora_medios_region[6])/
              desviaciones_ingresos_hora_region[6])))))
```

3. Se normaliza

ingreso_hora_normalizado

```
ecetss$ingreso_hora_normalizado<-
  100*(ecetss$z_ingreso_hora-min(
    ecetss$z_ingreso_hora, na.rm = TRUE))/
  (max(ecetss$z_ingreso_hora, na.rm = TRUE)-min(
    ecetss$z_ingreso_hora, na.rm = TRUE))

summary(ecetss$ingreso_hora_normalizado)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.542   3.756   4.561   5.552 100.000
```

Autonomía

Se usan variables de ECETSS que no están en EPH: aut_org, aut_metod, aut_ritmo, aut_pausas, aut_cantt Con categorías: La categorización de cada una es: 1 Siempre 2 Muchas veces 3 Algunas veces 4 Solo alguna vez 5 Nunca 99 ns/nc

Se eliminan los 99, se la trata como numérica y se define el índice como suma simple

```
levels(ecetss$aut_org)<-c(1,2,3,4,5,NA)
levels(ecetss$aut_metod)<-c(1,2,3,4,5,NA)
levels(ecetss$aut_ritmo)<-c(1,2,3,4,5,NA)
levels(ecetss$aut_pausas)<-c(1,2,3,4,5,NA)
levels(ecetss$aut_cantt)<-c(1,2,3,4,5,NA)
```

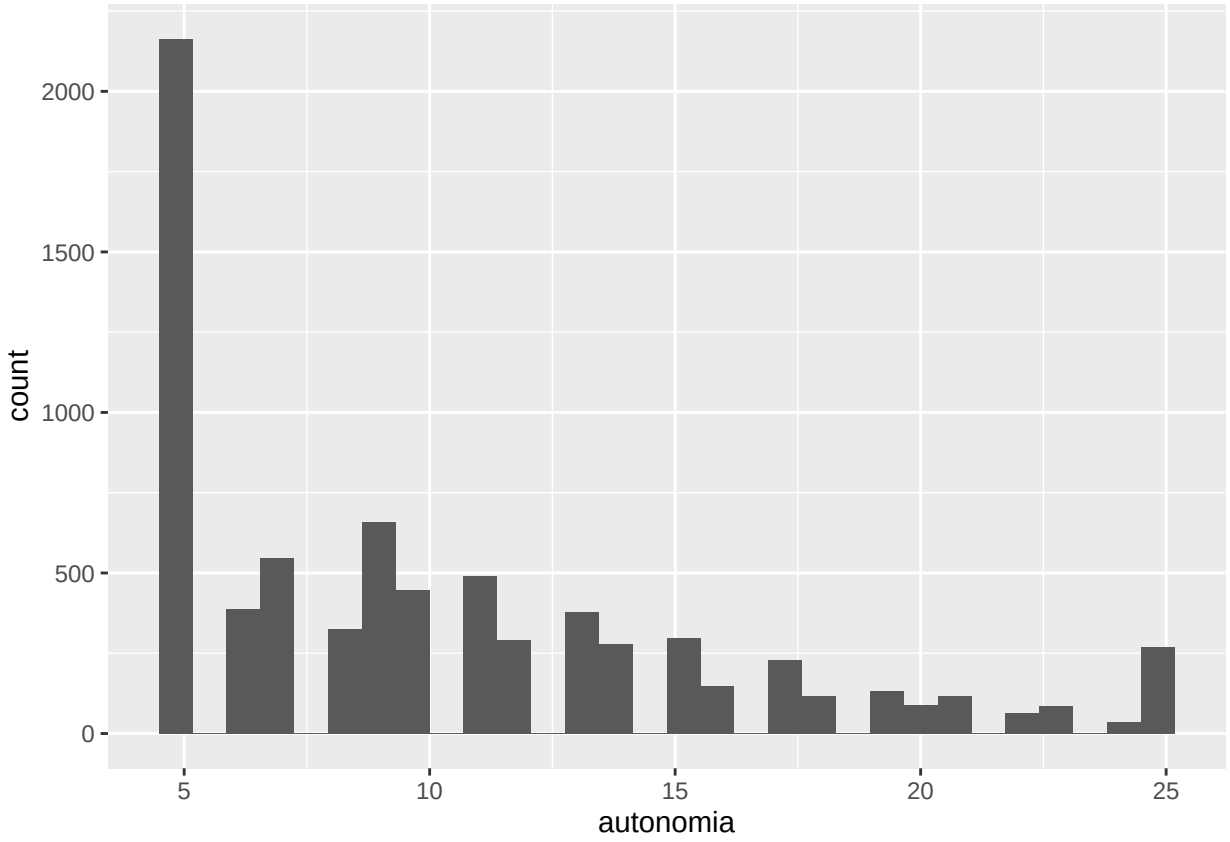
```r
ecetss$aut_org_num<-as.numeric(as.character(ecetss$aut_org))
ecetss$aut_metod_num<-as.numeric(as.character(ecetss$aut_metod))
ecetss$aut_ritmo_num<-as.numeric(as.character(ecetss$aut_ritmo))
ecetss$aut_pausas_num<-as.numeric(as.character(ecetss$aut_pausas))
ecetss$aut_cantt_num<-as.numeric(as.character(ecetss$aut_cantt))


ecetss$autonomia<-
  ecetss$aut_org_num+ecetss$aut_metod_num+
  ecetss$aut_ritmo_num+
  ecetss$aut_pausas_num+
  ecetss$aut_cantt_num

summary(ecetss$autonomia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    5.00    5.00    9.00   10.35   13.00   25.00       6
```

```r
ggplot(ecetss)+geom_histogram(aes(autonomia))
```



Se normaliza con el orden invertido para que los números más altos correspondan a mayor autonomía:

```r
ecetss$autonomia_normalizada<-
  100*(
    ecetss$autonomia-max(ecetss$autonomia, na.rm = TRUE))/
```

```r
  (-max(ecetss$autonomia, na.rm = TRUE)+
     min(ecetss$autonomia, na.rm = TRUE))
# verificación
table(ecetss$autonomia_normalizada, ecetss$autonomia)
```

```
##
##          5    6    7    8    9   10   11   12   13   14   15   16   17   18
##   0      0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   5      0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   10     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   15     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   20     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   25     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   30     0    0    0    0    0    0    0    0    0    0    0    0    0    0
##   35     0    0    0    0    0    0    0    0    0    0    0    0    0  118
##   40     0    0    0    0    0    0    0    0    0    0    0    0  228    0
##   45     0    0    0    0    0    0    0    0    0    0    0  149    0    0
##   50     0    0    0    0    0    0    0    0    0    0  297    0    0    0
##   55     0    0    0    0    0    0    0    0    0  278    0    0    0    0
##   60     0    0    0    0    0    0    0    0  380    0    0    0    0    0
##   65     0    0    0    0    0    0    0  291    0    0    0    0    0    0
##   70     0    0    0    0    0    0  489    0    0    0    0    0    0    0
##   75     0    0    0    0    0  448    0    0    0    0    0    0    0    0
##   80     0    0    0    0  659    0    0    0    0    0    0    0    0    0
##   85     0    0    0  324    0    0    0    0    0    0    0    0    0    0
##   90     0    0  548    0    0    0    0    0    0    0    0    0    0    0
##   95     0  389    0    0    0    0    0    0    0    0    0    0    0    0
##   100 2164    0    0    0    0    0    0    0    0    0    0    0    0    0
##
##          19   20   21   22   23   24   25
##   0      0    0    0    0    0    0  270
##   5      0    0    0    0    0   35    0
##   10     0    0    0    0   85    0    0
##   15     0    0    0   63    0    0    0
##   20     0    0  116    0    0    0    0
##   25     0   89    0    0    0    0    0
##   30   132    0    0    0    0    0    0
##   35     0    0    0    0    0    0    0
##   40     0    0    0    0    0    0    0
##   45     0    0    0    0    0    0    0
##   50     0    0    0    0    0    0    0
##   55     0    0    0    0    0    0    0
##   60     0    0    0    0    0    0    0
##   65     0    0    0    0    0    0    0
##   70     0    0    0    0    0    0    0
##   75     0    0    0    0    0    0    0
##   80     0    0    0    0    0    0    0
##   85     0    0    0    0    0    0    0
##   90     0    0    0    0    0    0    0
##   95     0    0    0    0    0    0    0
##   100    0    0    0    0    0    0    0
```

Primer índice de calidad (comparable con datos EPH)

```
ecetss$IC<-(ecetss$seguridad_normalizada+
              ecetss$consistencia_normalizado+
              ecetss$ingreso_hora_normalizado)/3
ecetss<-subset(ecetss, is.na(ecetss$IC)==FALSE)
```

Segundo índice de calidad (agrega autonomía)

```
ecetss$IC_2<-(ecetss$seguridad_normalizada+
                ecetss$consistencia_normalizado+
                ecetss$ingreso_hora_normalizado+
                ecetss$autonomia_normalizada)/4

ecetss<-subset(ecetss, is.na(ecetss$IC_2)==FALSE)
```

Análisis de los componentes de los índices

Coeficientes de Spearman y significación

```
para_correlaciones<-ecetss[,c(381,388,375,394,396,397)]
v<-rcorr(as.matrix(para_correlaciones), type = "spearman")
v_r<-as.data.frame(v[1])
names(v_r)<-c("r.consistencia", "r.ingr_hora", "r.seguridad","r.autonomia", "r.IC", "r.IC_2" )
kable(v_r)
```

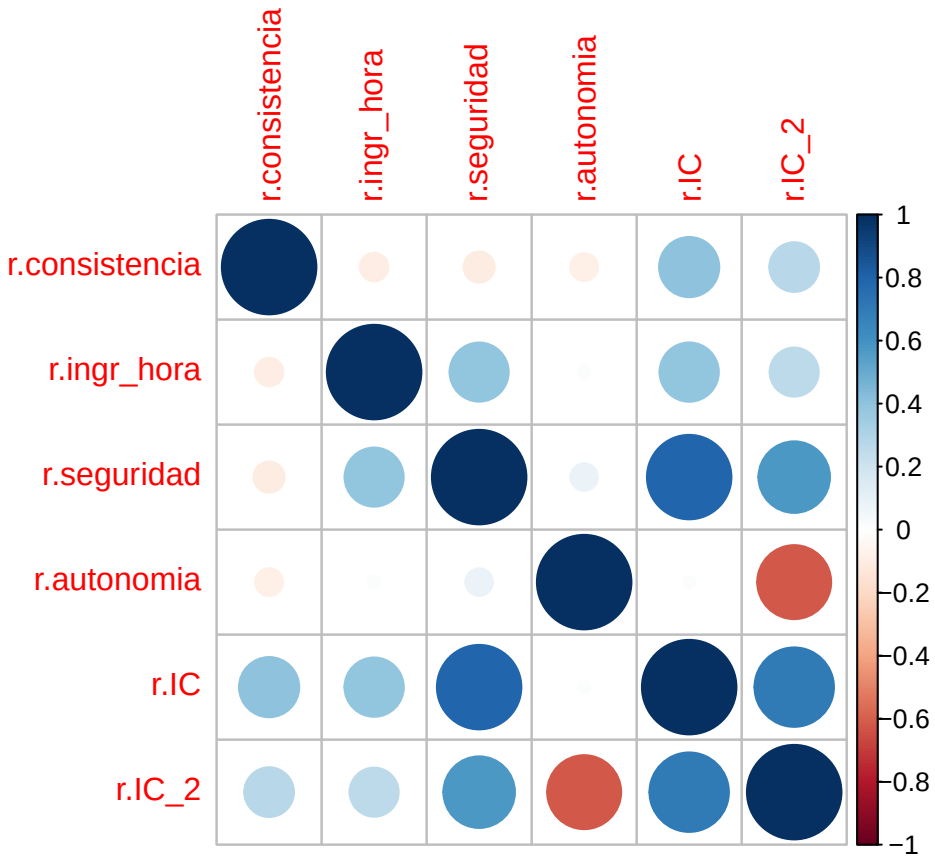
	r.consistencia	r.ingr_hora	r.seguridad	r.autonomia	r.IC	r.IC_2
consistencia_normalizado	1.0000000	-0.0909495	-0.1081307	-0.0870264	0.4059908	0.2770842
ingreso_hora_normalizado	-0.0909495	1.0000000	0.3911605	0.0188683	0.3940396	0.2699768
seguridad	-0.1081307	0.3911605	1.0000000	0.0849394	0.7985265	0.5734446
autonomia	-0.0870264	0.0188683	0.0849394	1.0000000	0.0168068	-0.6124136
IC	0.4059908	0.3940396	0.7985265	0.0168068	1.0000000	0.7063789
IC_2	0.2770842	0.2699768	0.5734446	-0.6124136	0.7063789	1.0000000

```
v_sig<-as.data.frame(v[3])
names(v_sig)<-c("p.consistencia", "p.ingr_hora", "p.seguridad","p.autonomia", "p.IC", "p.IC_2" )

kable(v_sig)
```

	p.consistencia	p.ingr_hora	p.seguridad	p.autonomia	p.IC	p.IC_2
consistencia_normalizado	NA	0.0000000	0	0.0000000	0.0000000	0
ingreso_hora_normalizado	0	NA	0	0.1014557	0.0000000	0
seguridad	0	0.0000000	NA	0.0000000	0.0000000	0
autonomia	0	0.1014557	0	NA	0.1446022	0
IC	0	0.0000000	0	0.1446022	NA	0
IC_2	0	0.0000000	0	0.0000000	0.0000000	NA

```
dimnames(v$r)<-list(names(v_r), names(v_r))
corrplot(v$r)
```



Se retiene IC

Porque correlaciona mejor con las componentes, la correlación entre ellos es alta y permitirá comparar con EPH

Descripción del índice

```
summary(ecetss$IC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.432  47.865  56.044  51.926  60.031  87.703
```

```
skewness(ecetss$IC)
```

```
## [1] -1.285026
```

```r
kurtosis(ecetss$IC)
```

```
## [1] 3.86755
```

```r
skewness(ecetss$IC^3.19)
```

```
## [1] -0.3648484
```

```r
(summary(ecetss$IC^3.19))^(1/3.19)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.432  47.865  56.044  54.382  60.031  87.703
```

```r
(abs(skewness(ecetss$IC^3.19)))^(1/3.19)
```

```
## [1] 0.7290061
```

```r
(abs(kurtosis(ecetss$IC^3.19)))^(1/3.19)
```

```
## [1] 1.383171
```

```r
lillie.test(x = ecetss$IC)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ecetss$IC
## D = 0.16598, p-value < 2.2e-16
```

Variables explicativas

```r
ecetss$sexo<-ecetss$C3P16.1
levels(ecetss$sexo)<-c("varones", "mujeres")

ecetss$origen<-ecetss$C3P16.6
levels(ecetss$origen)<-c("natives", "extranjeres", NA)

ecetss$edad<-as.numeric(as.character(ecetss$C0P10.3))
```

Comparaciones de las distribuciones por sexos y orígenes

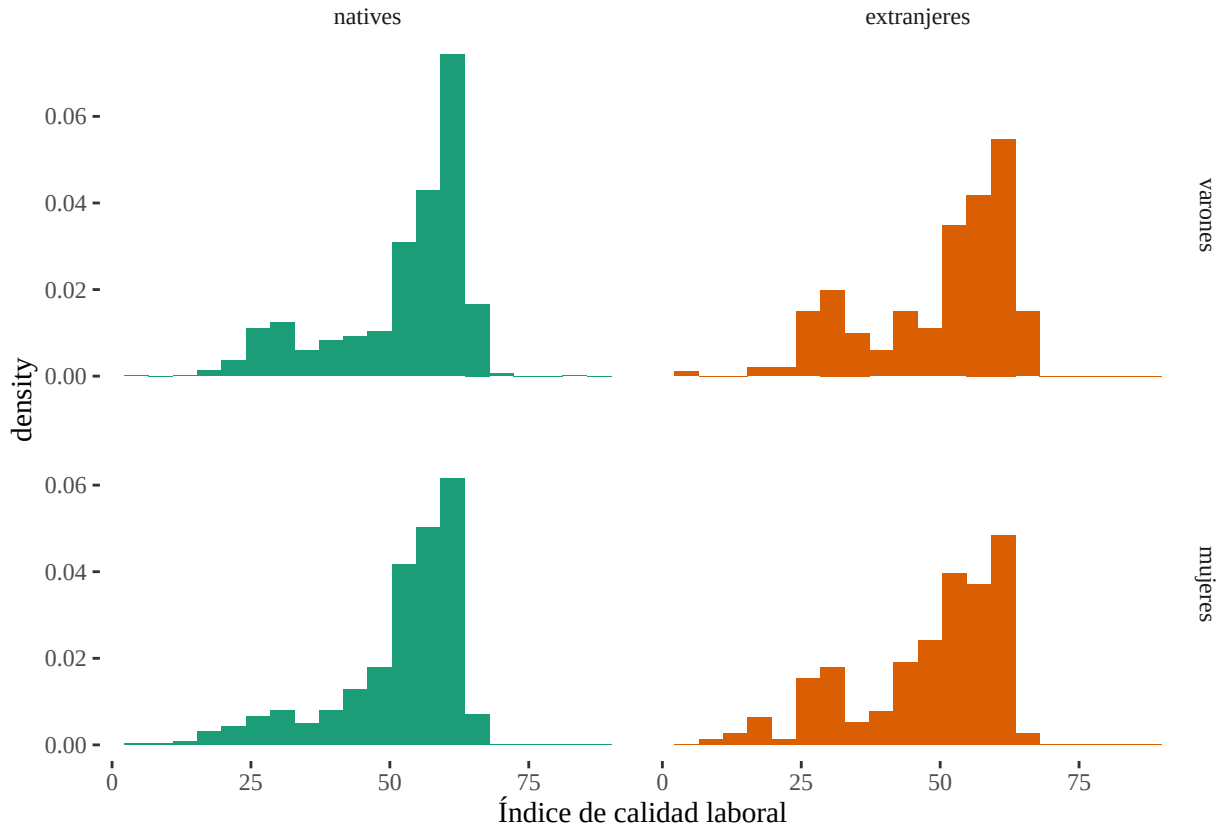
```r
t.test(IC~sexo, data = ecetss)
```

```
##
##  Welch Two Sample t-test
##
## data:  IC by sexo
## t = 3.3619, df = 7171.1, p-value = 0.0007782
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3803372 1.4442525
## sample estimates:
## mean in group varones mean in group mujeres
##              52.32217              51.40988
```

```
t.test(IC~origen, data = ecetss)
```

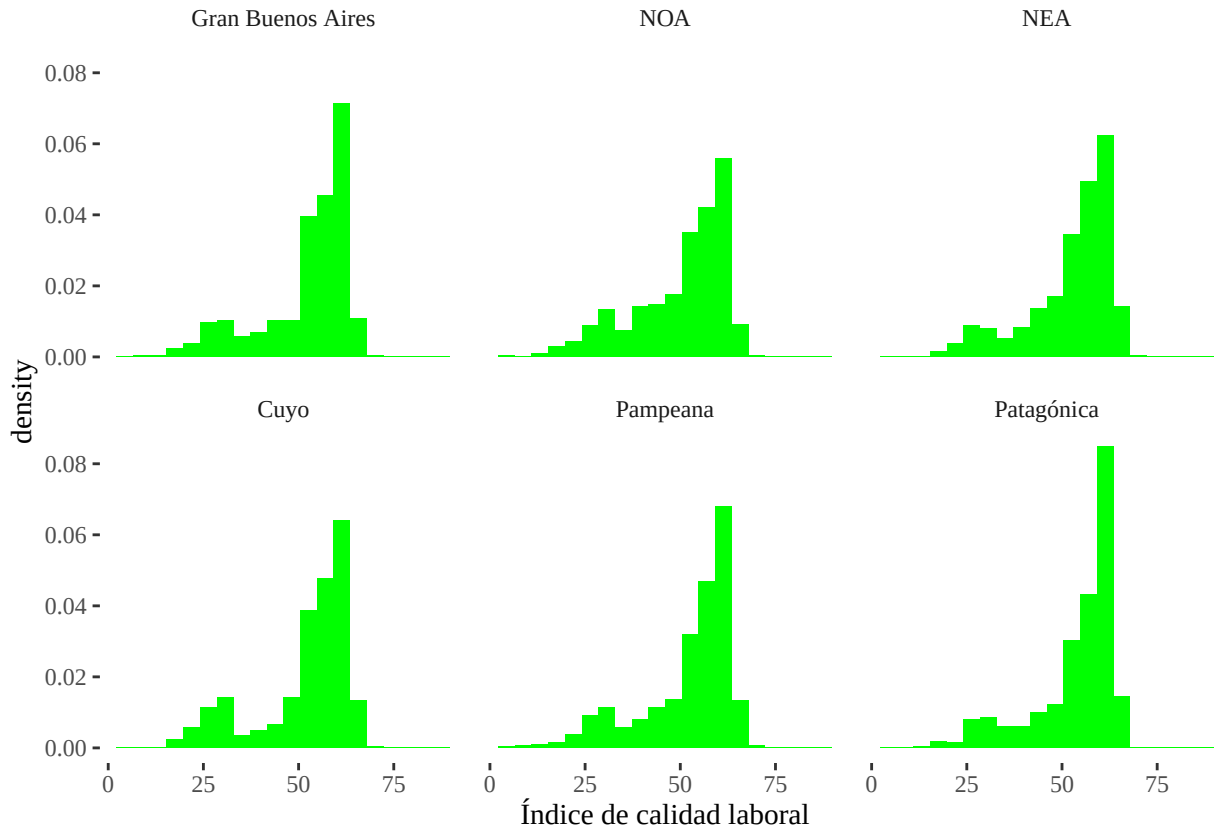
```
##
##  Welch Two Sample t-test
##
## data:  IC by origen
## t = 4.6826, df = 445.24, p-value = 3.768e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.787928 4.374239
## sample estimates:
##     mean in group natives mean in group extranjeres
##                  52.09251                  49.01143
```

```
ggplot(ecetss)+geom_histogram(aes(IC, y=..density.., fill=origen), bins = 20)+
  xlab("Índice de calidad laboral")+theme_tufte()+facet_grid(sexo~origen)+
  scale_fill_brewer(palette="Dark2")+ theme(legend.position = "none")
```




```
ecetss$region_cod<-ecetss$region
levels(ecetss$region_cod)<-c("Gran Buenos Aires", "NOA", "NEA", "Cuyo", "Pampeana", "Patagónica")

ggplot(ecetss)+geom_histogram(aes(IC, y=..density..), bins = 20, fill="green")+
  xlab("Índice de calidad laboral")+theme_tufte()+facet_wrap(ecetss$region_cod)+
  scale_fill_brewer(palette="Dark2")+ theme(legend.position = "none")
```



Comparaciones de las medias por sexos y orígenes

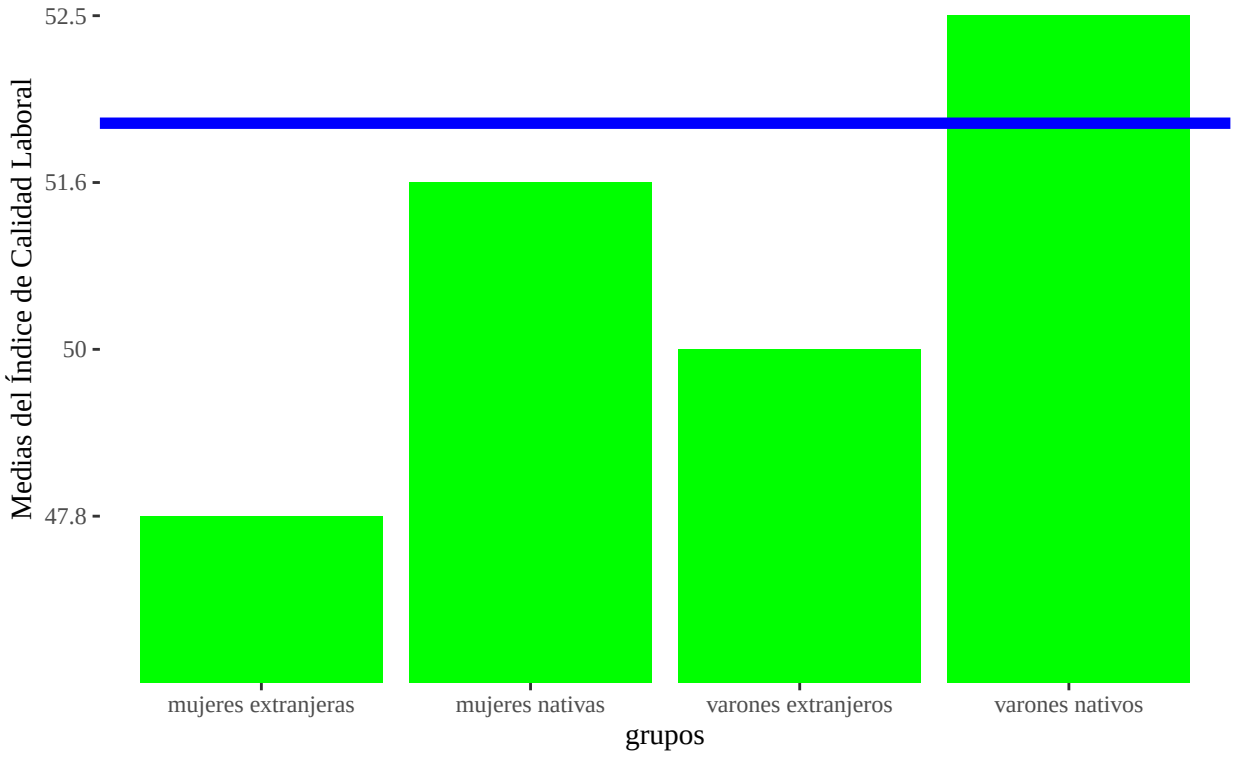
```
grupos<-c("varones nativos", "varones extranjeros",
          "mujeres nativas", "mujeres extranjeras", "total")

medias<-
  round(c(mean(ecetss[ecetss$sexo=="varones" & ecetss$origen=="natives",]$IC, na.rm = TRUE),
          mean(ecetss[ecetss$sexo=="varones" & ecetss$origen=="extranjeres",]$IC, na.rm = TRUE),
          mean(ecetss[ecetss$sexo=="mujeres"& ecetss$origen=="natives",]$IC, na.rm = TRUE),
          mean(ecetss[ecetss$sexo=="mujeres"& ecetss$origen=="extranjeres",]$IC, na.rm = TRUE),
          mean(ecetss$IC, na.rm = TRUE)),1)

medias_IC<-data.frame(cbind(medias, grupos))
mean(ecetss$IC, na.rm = TRUE)
```

```
## [1] 51.9257
```

```
ggplot(medias_IC)+
  geom_bar(aes(grupos,medias), stat = "identity", fill="green",
           data = medias_IC[-5,])+
           geom_hline(yintercept = 3+.32/.9, col="blue", size=2)+
  ylab("Medias del Índice de Calidad Laboral")+
  theme_tufte()
```

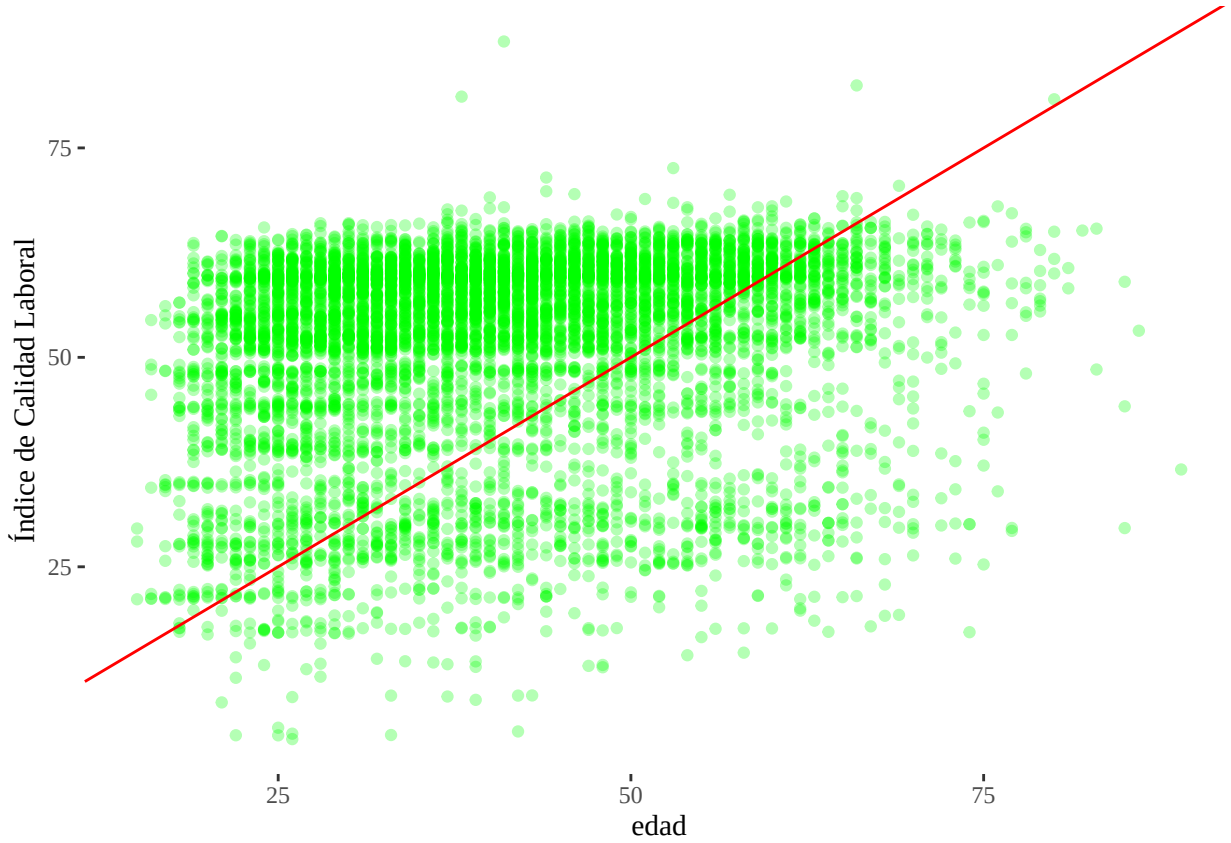


Índice según edad

```
cor.test(ecetss$IC,ecetss$edad)
```

```
##
##  Pearson's product-moment correlation
##
## data:  ecetss$IC and ecetss$edad
## t = 15.269, df = 7534, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1512674 0.1950691
## sample estimates:
##       cor
## 0.1732539
```

```
ggplot(ecetss)+geom_point(aes(edad, IC),col="green", alpha=0.3)+
  geom_abline(col="red")+ylab("Índice de Calidad Laboral")+
theme_tufte()
```



Modelo lineal

Directo con IC

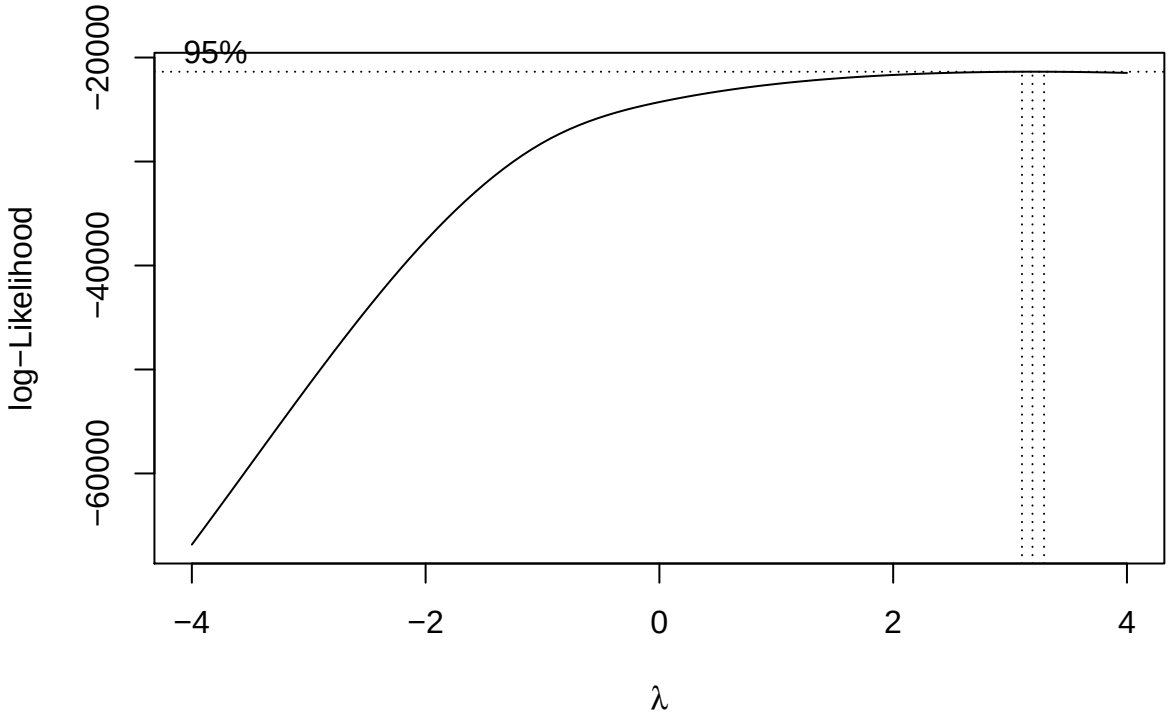
```
modelo.1<-lm(IC~origen+sexo+region_cod+edad, data = ecetss)
summary(modelo.1)
```

```
##
## Call:
## lm(formula = IC ~ origen + sexo + region_cod + edad, data = ecetss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.928  -3.782   3.939   7.945  37.991
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          46.51291    0.51432  90.436  < 2e-16 ***
## origenextranjeres    -3.92329    0.60437  -6.492 9.04e-11 ***
## sexomujeres          -1.00665    0.26714  -3.768 0.000166 ***
## region_codNOA        -2.25710    0.42760  -5.279 1.34e-07 ***
## region_codNEA        -0.10338    0.43340  -0.239 0.811469
## region_codCuyo       -0.61765    0.43749  -1.412 0.158049
```

```
## region_codPampeana   -0.75951    0.45351  -1.675 0.094030 .
## region_codPatagónica  1.35721    0.44057   3.081 0.002073 **
## edad                  0.15762    0.01012  15.575  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.49 on 7527 degrees of freedom
## Multiple R-squared:  0.04376,    Adjusted R-squared:  0.04274
## F-statistic: 43.06 on 8 and 7527 DF,  p-value: < 2.2e-16
```

Transformación Box-Cox

```
para.ajuste<-boxcox(modelo.1, lambda = seq(-4,4))
```




```
para.ajuste$x[which(para.ajuste$y==max(para.ajuste$y))]
```

```
## [1] 3.191919
```

```
ecetss$ICbc<-ecetss$IC^3.1919
ecetss$ICnuevo<-100*(ecetss$ICbc-min(ecetss$ICbc))/(max(ecetss$ICbc)-min(ecetss$ICbc))
```

- Modelo corregido Box-Cox

```r
modelo.2<-lm(ICnuevo~origen+sexo+region_cod+edad, data = ecetss)
summary(modelo.2)
```

```
##
## Call:
## lm(formula = ICnuevo ~ origen + sexo + region_cod + edad, data = ecetss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.790  -6.803   2.139   8.080  80.758
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           16.063023   0.469425  34.218  < 2e-16 ***
## origenextranjeres     -3.639546   0.551615  -6.598 4.45e-11 ***
## sexomujeres           -1.498655   0.243818  -6.147 8.32e-10 ***
## region_codNOA         -2.201861   0.390272  -5.642 1.74e-08 ***
## region_codNEA         -0.234287   0.395567  -0.592 0.553680
## region_codCuyo        -0.390479   0.399300  -0.978 0.328150
## region_codPampeana    -0.553144   0.413926  -1.336 0.181480
## region_codPatagónica   1.383628   0.402114   3.441 0.000583 ***
## edad                   0.167791   0.009237  18.166  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.49 on 7527 degrees of freedom
## Multiple R-squared:  0.05789,    Adjusted R-squared:  0.05689
## F-statistic: 57.81 on 8 and 7527 DF,  p-value: < 2.2e-16
```

Con PCA

```r
# se retienen solo las tres componentes de IC
solo_componentes_indice<-ecetss[, c(376,381,389)]
# se ejecuta pca
pca<-prcomp(solo_componentes_indice)
# se agregan a la base estas tres columnas
ecetss<-data.frame(cbind(ecetss,pca$x))
# para facilitar la comparación se cambia el signo
# a PC1 y se llama QI
ecetss$QI<- -ecetss$PC1
```

- Modelo corregido con PCA (conservando la primera componente, con signo cambiado)

```r
modelo.3<-lm(QI~origen+sexo+region_cod+edad, data = ecetss)
summary(modelo.3)
```

```
##
## Call:
## lm(formula = QI ~ origen + sexo + region_cod + edad, data = ecetss)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.415  -2.488  14.756  20.666  36.275
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -6.04160    1.43630  -4.206 2.63e-05 ***
## origenextranjeres     -10.46275    1.68778  -6.199 5.98e-10 ***
## sexomujeres             4.07978    0.74601   5.469 4.68e-08 ***
## region_codNOA          -7.48946    1.19411  -6.272 3.76e-10 ***
## region_codNEA          -2.95845    1.21032  -2.444  0.01453 *
## region_codCuyo         -3.07602    1.22174  -2.518  0.01183 *
## region_codPampeana     -3.97418    1.26649  -3.138  0.00171 **
## region_codPatagónica    3.00349    1.23035   2.441  0.01466 *
## edad                    0.17304    0.02826   6.123 9.65e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.09 on 7527 degrees of freedom
## Multiple R-squared:  0.02165,    Adjusted R-squared:  0.02061
## F-statistic: 20.82 on 8 and 7527 DF,  p-value: < 2.2e-16
```

Resumen de la comparación de los modelos

```
coeficientes<-data.frame(modelo.1$coefficients, modelo.2$coefficients,
                         modelo.3$coefficients)
kable(round(coeficientes,3), col.names = c("modelo 1", "modelo 2", "modelo 3"))
```

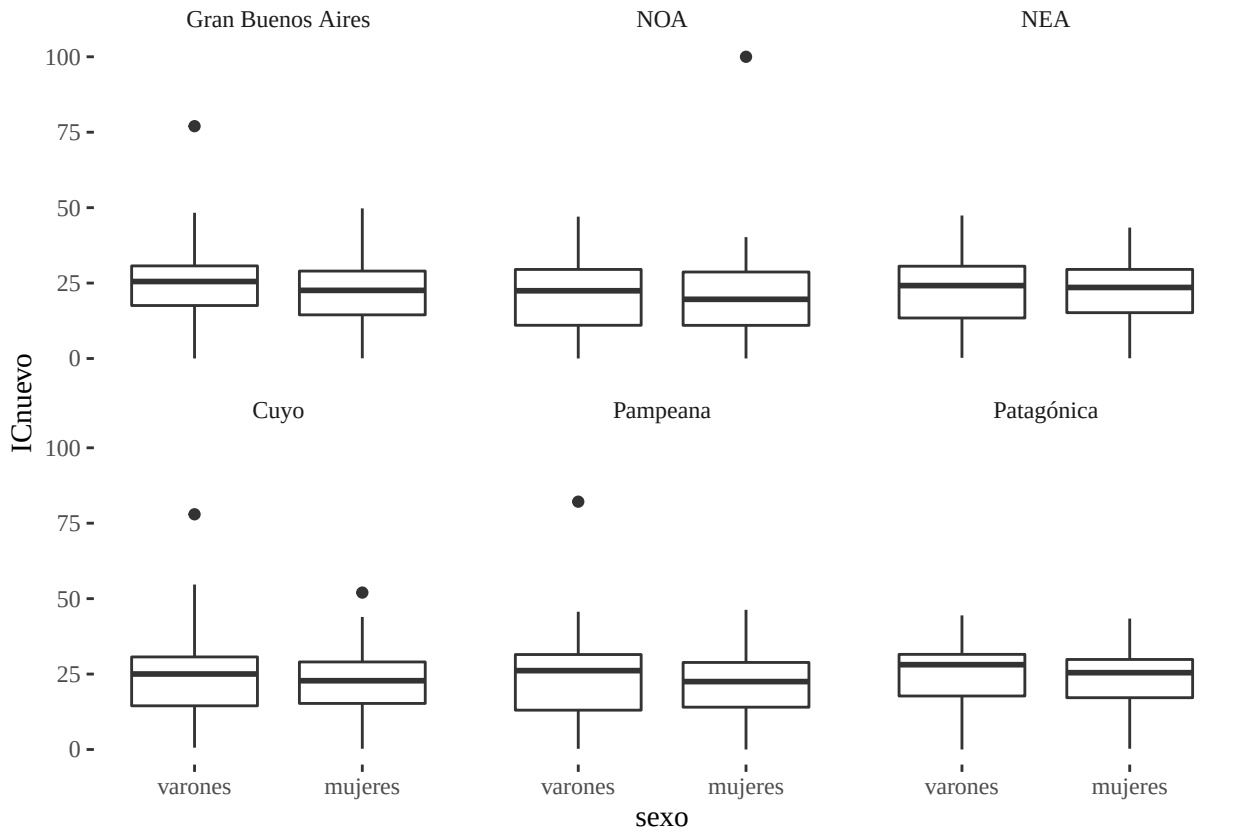
	modelo 1	modelo 2	modelo 3
(Intercept)	46.513	16.063	-6.042
origenextranjeres	-3.923	-3.640	-10.463
sexomujeres	-1.007	-1.499	4.080
region_codNOA	-2.257	-2.202	-7.489
region_codNEA	-0.103	-0.234	-2.958
region_codCuyo	-0.618	-0.390	-3.076
region_codPampeana	-0.760	-0.553	-3.974
region_codPatagónica	1.357	1.384	3.003
edad	0.158	0.168	0.173

```
R_cuadrado<- data.frame(summary(modelo.1)$r.squared,
                        summary(modelo.2)$r.squared,
                        summary(modelo.3)$r.squared)
kable(round(R_cuadrado,3), col.names = c("modelo 1", "modelo 2", "modelo 3"))
```

modelo 1	modelo 2	modelo 3
0.044	0.058	0.022

Otras comparaciones

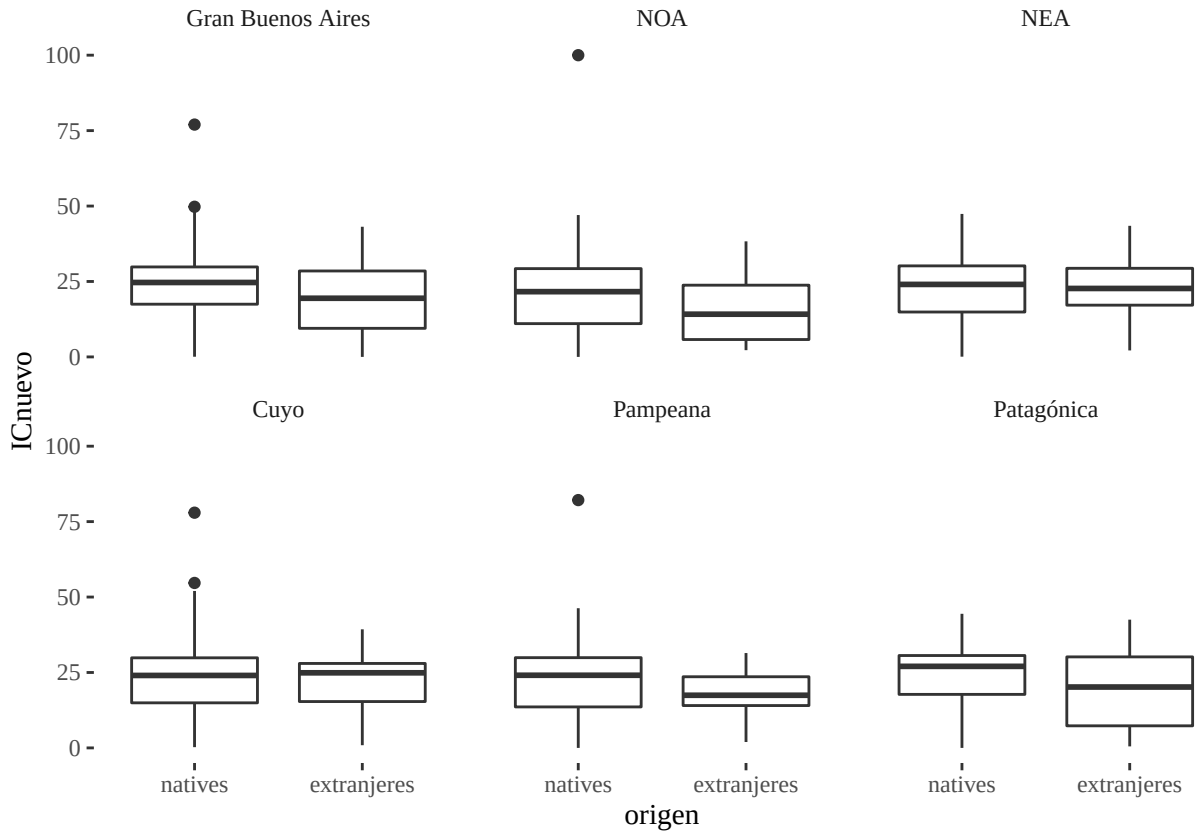
```
ggplot(ecetss)+geom_boxplot(aes(y=ICnuevo, x=sexo))+
  facet_wrap(ecetss$region_cod)+theme_tufte()
```




```
wilcox.test(ecetss$ICnuevo ~ ecetss$sexo)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  ecetss$ICnuevo by ecetss$sexo
## W = 7704466, p-value = 8.085e-15
## alternative hypothesis: true location shift is not equal to 0
```

```
ggplot(ecetss)+geom_boxplot(aes(y=ICnuevo, x=origen))+
  facet_wrap(ecetss$region_cod)+theme_tufte()
```




```
wilcox.test(ecetss$ICnuevo ~ ecetss$origen)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  ecetss$ICnuevo by ecetss$origen
## W = 1670567, p-value = 4.093e-07
## alternative hypothesis: true location shift is not equal to 0
```