PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

Pre-Entrega Proyecto Modelos Lineales

Felipe López | Camila Straub B.

14 de octubre de 2020

Índice

1. Descripción del problema **1**

2. Análisis descriptivo de los datos **2**

3. Proceso de selección del modelo **11**

Backward 11

Forward 14

Referecnias **14**

1. Descripción del problema

La educación superior en Chile tiene una organización compleja y que ha sufrido múltiples reformas desde la década de los ochenta hasta la actualidad. Probablemente el cambio más importante sea la gratuidad, que se comenzó a implementar desde el 2016.

Una de las aristas desde la que se puede estudiar a las Instituciones de Educación Superior (IES) y el acceso a éstas son los aranceles asociados a estudiar. Las universidades chileneas cobran los aranceles que estiman convenientes por las carreras de pregrado y pueden usar cualquier criterio que les parezca. Cada institución fija, año tras año, y autónomamente el precio que debe pagar el estudiante (Salas & Gómez, 2013; Dooner & Mena, 2006). Según la OCDE, Chile se encuentra en la segunda categoría entre las naciones con mayores aranceles. Sin embargo, basándose en rankings internacionales, Chile no está en el mismo nivel que el resto de los países que pertenecen a esta categoría en cuanto a calidad de la educación impartida (Cancino, 2010). Existe una amplia variación de los aranceles entre las univerisdades y también dentro de cada una, en los precios de los distintos porgramas que ofrecen.

La presente investigción busca indagar en qué factores -y en qué medida - inciden en los precios de los aranceles. Para esto se utiliza información sobre la matrícula de educación superior por estudiante. Esta

información es de carácter público y se encuentra disponible la base de datos y el correspondiente esquema de registro a través del Centro de Estudios MINEDUC, en este enlace.

La base de datos correspondiente al año 2019 contiene 1.268.504 registros y 48 variables que corresponden a los estudiantes que se matricularon en IES ese año. Sin embargo, como la unidad de análisis que interesa son los programas, al agrupar por esta variable se tiene que la base cuenta con información de 14.546 programas, donde cada uno tiene entre uno y 4.873 estudiantes.

La principal variable de interés es el arancel anual de la carrera o programa (`VALOR_ARANCEL`).

Algunas de las variables consideradas para el modelo son:

- Clasificación según categoría de institución que distingue Universidades del Consejo de Rectores y Universidades Privadas que no son del CRUCH (`TIPO_INST_2`)
- Región donde se ubica la sede en la cual se imparte la carrera (`REGIÓN_SEDE`)
- Categorización área OECD de la carrera o programa (`OECD_AREA`)
- Número de años por la que fue acreditada la institución (`ACRE_INST_ANIO`)
- Jornada en que se imparte la carrera o programa (`JORNADA`)
- Clasificación del nivel al que pertenecen los estudios de la carrera o programa: Carreras Técnicas, Carreras Profesionales, Postítulo, Magíster, Doctorado. (`NIVEL_CARRERA_2`)
- Duración teórica o formal en semestres de la carrera o programa, desde el momento en que un estudiante ingresa al primer año de la carrera hasta que obtiene el título o grado terminal de la misma (`DUR_TOTAL_CARRERA`)

2. Análisis descriptivo de los datos

Como se mencionó en el apartado anterior se agrupó la base según programa de estudios y se mantuvieron solo las variables que hicieran relación al programa y no al estudiante. Se creo una variable que contiene el número de estudiantes matriculados en cada programa.

```
# glimpse(base)

base_f <- base %>%
  group_by(codigo_unico) %>%
  summarise(tipo_inst_1 = first(tipo_inst_1),
            tipo_inst_2 = first(tipo_inst_2),
            tipo_inst_3 = first(tipo_inst_3),
            cod_inst = first(cod_inst),
            nomb_inst = first(nomb_inst),
            cod_sede = first(cod_sede),
            nomb_sede = first(nomb_sede),
            cod_carrera = first(cod_carrera),
            nomb_carrera = first(nomb_carrera),
            modalidad = first(modalidad),
            jornada = first(jornada),
            version = first(version),
            tipo_plan_carr = first(tipo_plan_carr),
            dur_estudio_carr = first(dur_estudio_carr),
            dur_proceso_tit = first(dur_proceso_tit),
            dur_total_carr = first(dur_total_carr),
            region_sede = first(region_sede),
            provincia_sede = first(provincia_sede),
            comuna_sede = first(comuna_sede),
            nivel_global = first(nivel_global),
```

```
            nivel_carrera_1 = first(nivel_carrera_1),
            nivel_carrera_2 = first(nivel_carrera_2),
            requisito_ingreso = first(requisito_ingreso),
            vigencia_carrera = first(vigencia_carrera),
            valor_matricula = first(valor_matricula),
            valor_arancel = first(valor_arancel),
            codigo_demre = first(codigo_demre),
            area_conocimiento = first(area_conocimiento),
            oecd_area = first(oecd_area),
            oecd_subarea = first(oecd_subarea),
            area_carrera_generica = first(area_carrera_generica),
            acreditada_carr = first(acreditada_carr),
            acreditada_inst = first(acreditada_inst),
            acre_inst_desde_hasta = first(acre_inst_desde_hasta),
            acre_inst_anio  = first(acre_inst_anio ),
            costo_proceso_titulacion = first(costo_proceso_titulacion),
            costo_obtencion_titulo_diploma = first(costo_obtencion_titulo_diploma),
            forma_ingreso = first(forma_ingreso),
            n_estudiantes = n())
```

Distibución variables de interés:

```
# Estadísticos resumen
summary(base_f$valor_arancel)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0  1433425  1892000  2225992  2701000 25422760
```

Se eliminaron los casos donde el valor del arancel era igual a 0 o mayor a $10.000.000 anuales. También se recodificó la variable región y duración estudio de la carrera.

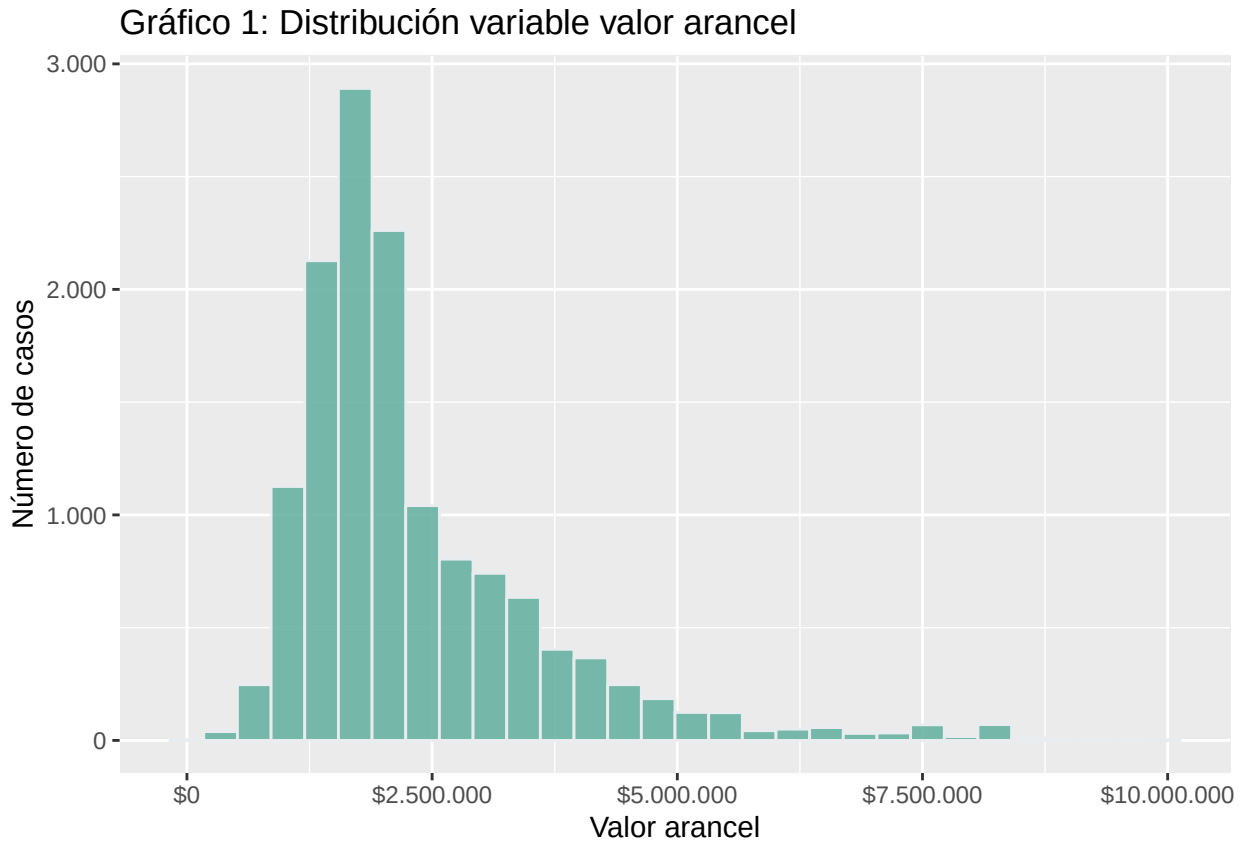
```
# Transformaciones
base_f <- base_f %>% filter(valor_arancel != 0, # filtrar casos = 0
                            valor_arancel <= 10000000) # filtrar casos > 10MM

base_f <- base_f %>%
  mutate(region_sede_d = as_factor(if_else(region_sede == "Metropolitana", 1, 0)),
         dur_estudio_carr_m = as_factor(dur_estudio_carr),
         dur_estudio_carr_m = fct_other(dur_estudio_carr_m,
                                        drop = c(14, 16, 20, 24),
                                        other_level = "Más de 12 semestres"),
         tipo_inst_1 = factor(tipo_inst_1),
         tipo_inst_2 = factor(tipo_inst_2))
```

En el Gráfico 1, se puede observar cómo se distribuye el valor del arancel.

```
# Histograma
ggplot(base_f, aes(x = valor_arancel)) +
  geom_histogram(fill = "#69b3a2", color = "#e9ecef", alpha = 0.9) +
  scale_x_continuous(labels = scales::label_dollar(big.mark = ".", decimal.mark = ",")) +
  scale_y_continuous(labels = scales::label_comma(big.mark = ".", decimal.mark = ",")) +
  labs(title = "Gráfico 1: Distribución variable valor arancel",
```

```
       x = "Valor arancel",
       y = "Número de casos")
```

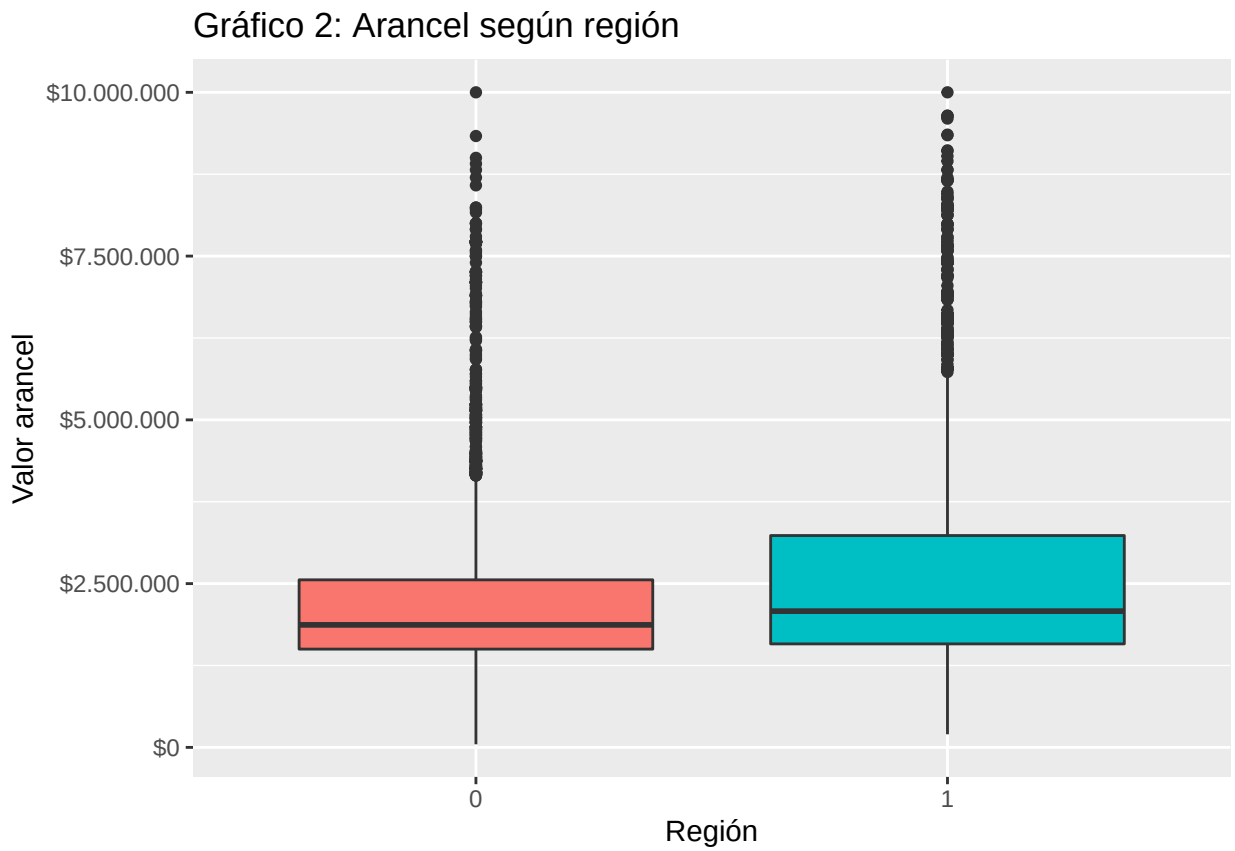



```
#ggplot(base_f, aes(x = log(valor_arancel))) +
#  geom_histogram(fill = "#69b3a2", color = "#e9ecef", alpha = 0.9) +
#  scale_y_continuous(labels = scales::label_comma(big.mark = ".", decimal.mark = ",")) +
#  labs(title = "Distribución variable valor arancel - Logaritmo",
#       x = "Valor arancel",
#       y = "Número de casos")
```

A continuación se muestra la relaicón entre el valor del arancel y algunas variables que podrían potencialmente ser predictoras.

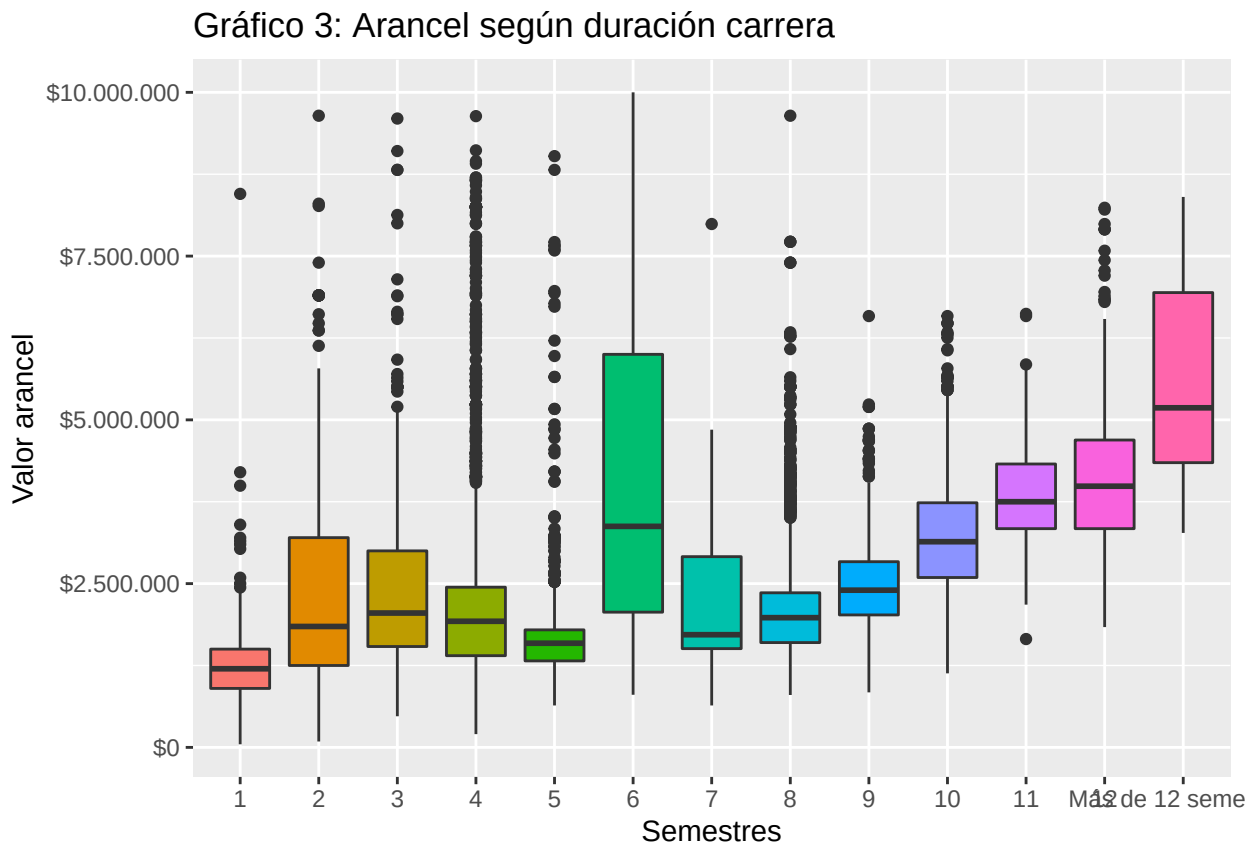
En el Gráfico 2 se observa cómo se distibuye el arancel según si la sede está ubicada en la región metropolitana (1) u en otra región (0).

```
ggplot(base_f, aes(x = region_sede_d, y = valor_arancel, fill= region_sede_d)) +
  geom_boxplot() +
  labs(title = "Gráfico 2: Arancel según región",
       x = "Región",
       y = "Valor arancel") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::label_dollar(big.mark = ".", decimal.mark = ","))
```

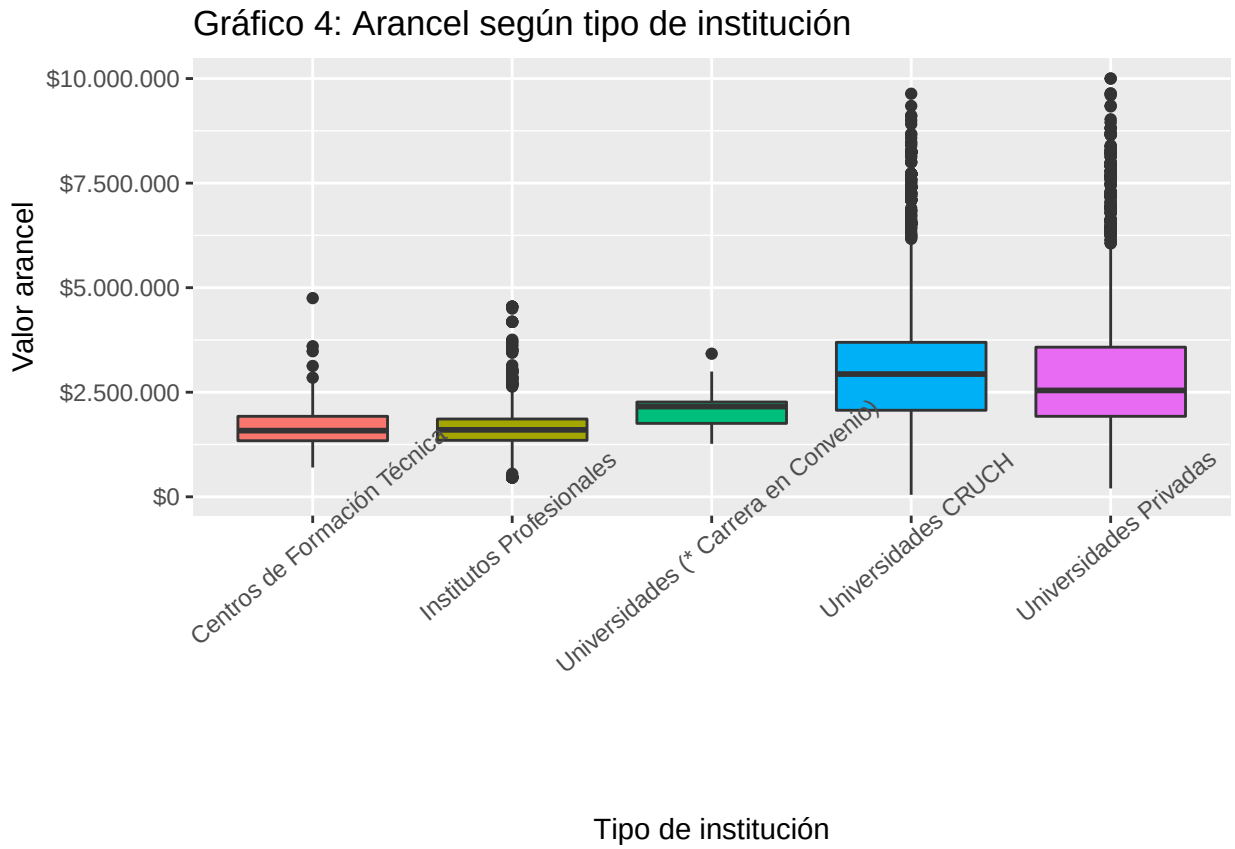


En el Gráfico 3 se observa la relación con la duración de la carrera.

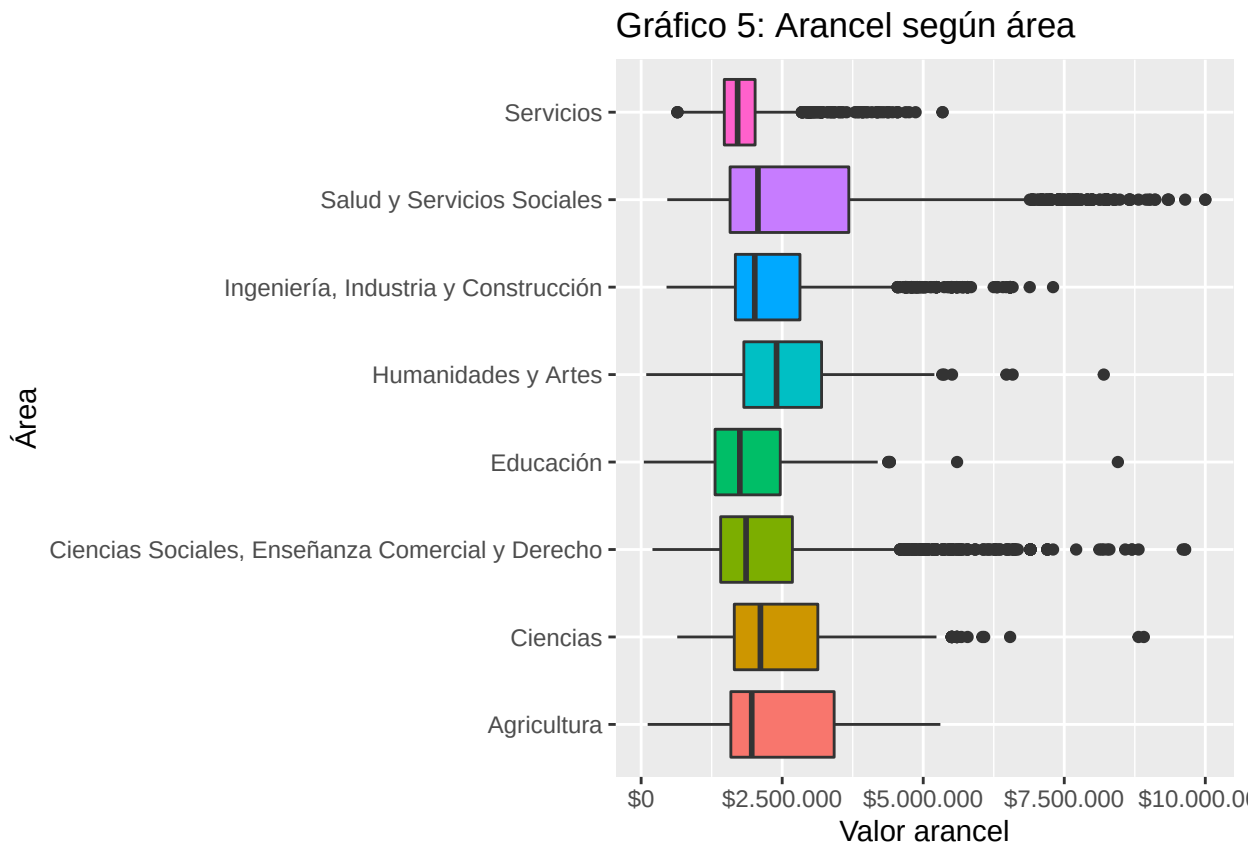
```
ggplot(base_f, aes(x = dur_estudio_carr_m, y = valor_arancel, fill= dur_estudio_carr_m)) +
  geom_boxplot() +
  labs(title = "Gráfico 3: Arancel según duración carrera",
       x = "Semestres",
       y = "Valor arancel") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::label_dollar(big.mark = ".", decimal.mark = ","))
```



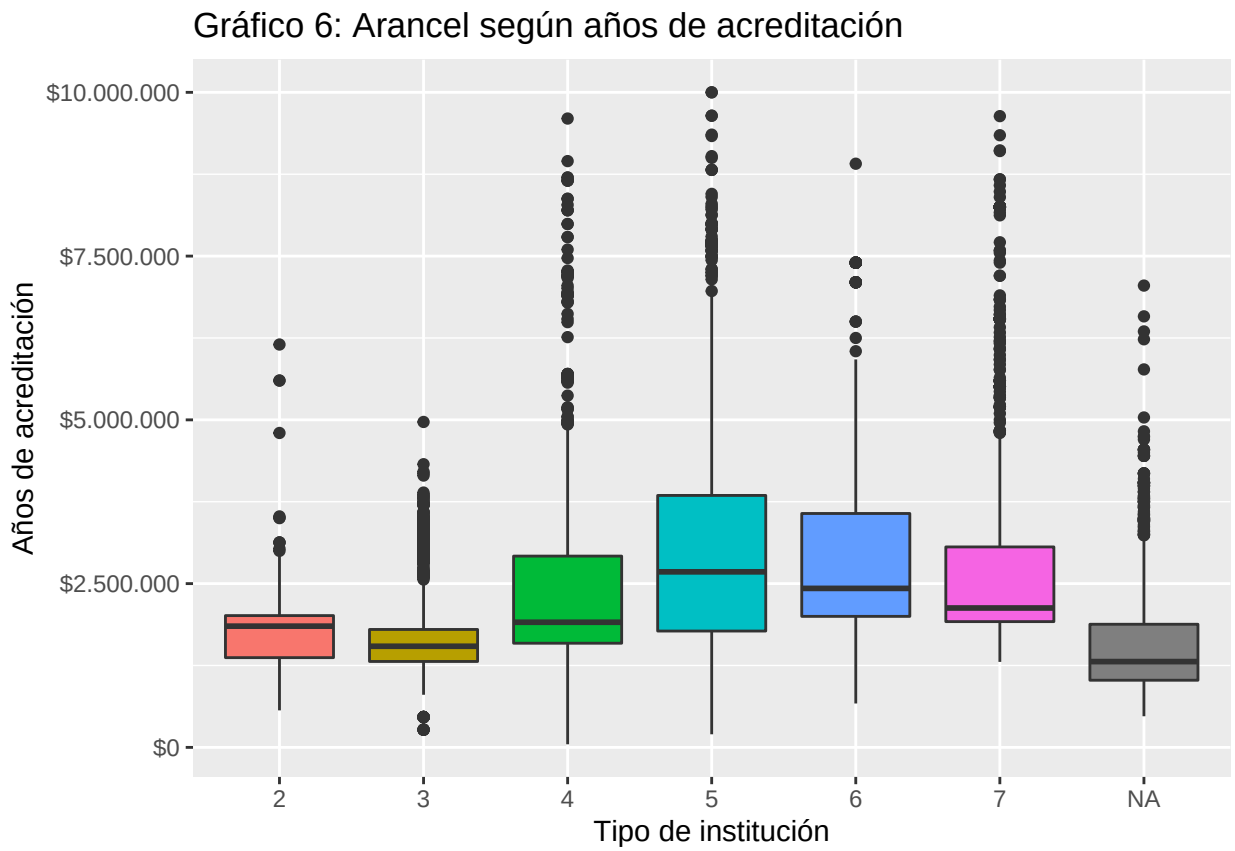

```
ggplot(base_f, aes(x = tipo_inst_2, y = valor_arancel, fill= tipo_inst_2)) +
  geom_boxplot() +
  labs(title = "Gráfico 4: Arancel según tipo de institución",
       x = "Tipo de institución",
       y = "Valor arancel") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::label_dollar(big.mark = ".", decimal.mark = ",")) +
  theme(axis.text.x = element_text(angle = 40))
```



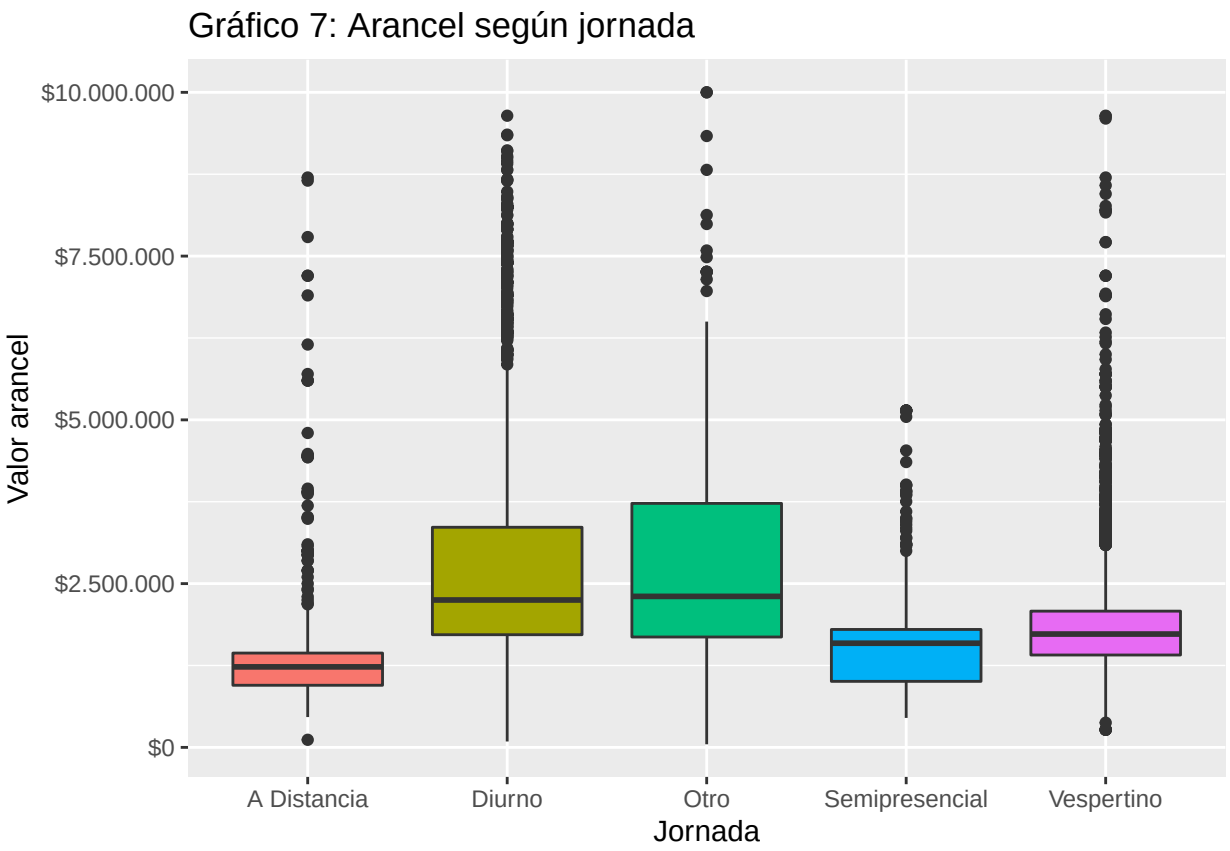

```
ggplot(base_f, aes(x = oecd_area, y = valor_arancel, fill= oecd_area)) +
  geom_boxplot() +
  labs(title = "Gráfico 5: Arancel según área",
       x = "Área",
       y = "Valor arancel") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::label_dollar(big.mark = ".", decimal.mark = ",")) +
  coord_flip()
```



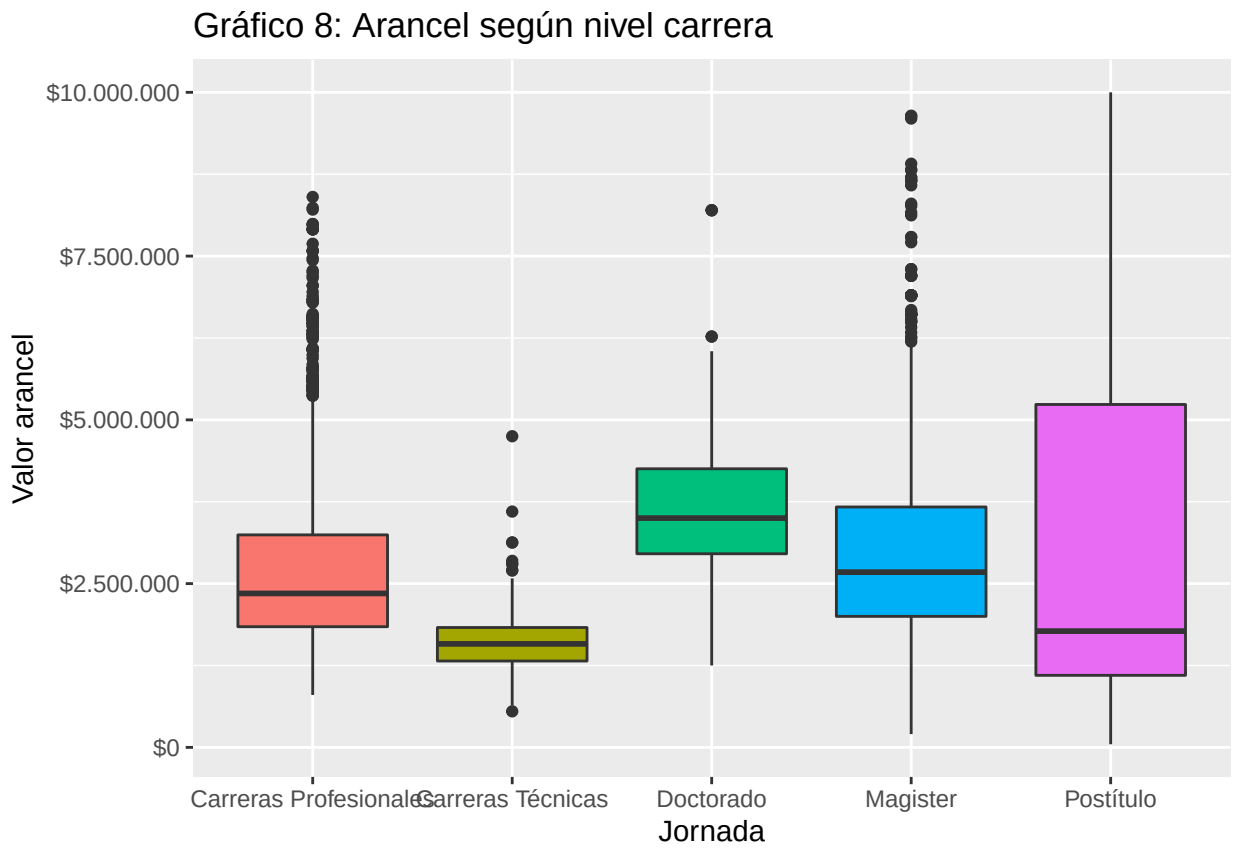

```
ggplot(base_f, aes(x = as.factor(acre_inst_anio), y = valor_arancel, fill = as.factor(acre_inst_anio)))
  geom_boxplot() +
  labs(title = "Gráfico 6: Arancel según años de acreditación",
       x = "Tipo de institución",
       y = "Años de acreditación") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::label_dollar(big.mark = ".", decimal.mark = ","))
```




```
ggplot(base_f, aes(x = jornada, y = valor_arancel, fill = jornada)) +
  geom_boxplot() +
  labs(title = "Gráfico 7: Arancel según jornada",
       x = "Jornada",
       y = "Valor arancel") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::label_dollar(big.mark = ".", decimal.mark = ","))
```




```
ggplot(base_f, aes(x = nivel_carrera_2, y = valor_arancel, fill = nivel_carrera_2)) +
  geom_boxplot() +
  labs(title = "Gráfico 8: Arancel según nivel carrera",
       x = "Jornada",
       y = "Valor arancel") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::label_dollar(big.mark = ".", decimal.mark = ","))
```



3. Proceso de selección del modelo

Backward

```
muestra_base_f <- sample_n(base_f, 500) %>% drop_na()

modelo1 <- lm(log(valor_arancel) ~ region_sede_d + dur_estudio_carr_m + oecd_area + acre_inst_anio +  j
drop1(modelo1, test = "F")
```

```
## Single term deletions
##
## Model:
## log(valor_arancel) ~ region_sede_d + dur_estudio_carr_m + oecd_area +
##     acre_inst_anio + jornada + tipo_inst_2 + nivel_carrera_2
##                    Df Sum of Sq    RSS      AIC F value    Pr(>F)
## <none>                          34.003 -1008.93
## region_sede_d       1    1.3339 35.336  -994.54 15.3775 0.0001039 ***
## dur_estudio_carr_m 12    8.9043 42.907  -933.84  8.5545 1.757e-14 ***
## oecd_area           7    3.4508 37.453  -981.75  5.6833 2.862e-06 ***
## acre_inst_anio      1    6.9393 40.942  -931.81 79.9997 < 2.2e-16 ***
## jornada             4    2.2722 36.275  -989.37  6.5487 4.140e-05 ***
## tipo_inst_2         4    5.8536 39.856  -949.26 16.8709 8.977e-13 ***
```

```
## nivel_carrera_2     4    3.8587 37.861  -971.14 11.1214 1.485e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(update(modelo1, ~ . -dur_total_carr), test = "F")
```

```
## Single term deletions
##
## Model:
## log(valor_arancel) ~ region_sede_d + dur_estudio_carr_m + oecd_area +
##     acre_inst_anio + jornada + tipo_inst_2 + nivel_carrera_2
##                    Df Sum of Sq    RSS      AIC F value    Pr(>F)
## <none>                          34.003 -1008.93
## region_sede_d       1    1.3339 35.336  -994.54 15.3775 0.0001039 ***
## dur_estudio_carr_m 12    8.9043 42.907  -933.84  8.5545 1.757e-14 ***
## oecd_area           7    3.4508 37.453  -981.75  5.6833 2.862e-06 ***
## acre_inst_anio      1    6.9393 40.942  -931.81 79.9997 < 2.2e-16 ***
## jornada             4    2.2722 36.275  -989.37  6.5487 4.140e-05 ***
## tipo_inst_2         4    5.8536 39.856  -949.26 16.8709 8.977e-13 ***
## nivel_carrera_2     4    3.8587 37.861  -971.14 11.1214 1.485e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(update(modelo1, ~ . -dur_total_carr - jornada), test = "F")
```

```
## Single term deletions
##
## Model:
## log(valor_arancel) ~ region_sede_d + dur_estudio_carr_m + oecd_area +
##     acre_inst_anio + tipo_inst_2 + nivel_carrera_2
##                    Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                          36.275 -989.37
## region_sede_d       1    1.0537 37.328 -979.17 11.5030 0.0007648 ***
## dur_estudio_carr_m 12    9.3966 45.671 -915.24  8.5483 1.734e-14 ***
## oecd_area           7    3.8713 40.146 -960.18  6.0374 1.053e-06 ***
## acre_inst_anio      1    8.1896 44.464 -904.65 89.4030 < 2.2e-16 ***
## tipo_inst_2         4    6.9214 43.196 -922.98 18.8896 3.151e-14 ***
## nivel_carrera_2     4    3.5031 39.778 -958.10  9.5606 2.179e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
modelo1 <- lm(log(valor_arancel) ~ region_sede_d + dur_estudio_carr_m + tipo_inst_1 + oecd_area + acre_

summary(modelo1)
```

```
##
## Call:
## lm(formula = log(valor_arancel) ~ region_sede_d + dur_estudio_carr_m +
##     tipo_inst_1 + oecd_area + acre_inst_anio + tipo_inst_2 +
##     nivel_carrera_2, data = muestra_base_f)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.47621 -0.13262  0.00549  0.13513  1.08391
##
## Coefficients: (2 not defined because of singularities)
##                                                      Estimate Std. Error
## (Intercept)                                          13.09968    0.21847
## region_sede_d1                                        0.11373    0.03353
## dur_estudio_carr_m2                                   0.25778    0.12998
## dur_estudio_carr_m3                                   1.13164    0.18670
## dur_estudio_carr_m4                                   0.89687    0.14179
## dur_estudio_carr_m5                                   0.87498    0.14754
## dur_estudio_carr_m6                                   0.99184    0.14404
## dur_estudio_carr_m7                                   1.18652    0.21545
## dur_estudio_carr_m8                                   0.67823    0.14436
## dur_estudio_carr_m9                                   0.77441    0.15978
## dur_estudio_carr_m10                                  0.94061    0.14981
## dur_estudio_carr_m11                                  1.11556    0.17826
## dur_estudio_carr_m12                                  1.12166    0.17187
## dur_estudio_carr_mMás de 12 semestres                 1.28053    0.26028
## tipo_inst_1Institutos Profesionales                  -0.01019    0.06404
## tipo_inst_1Universidades                              0.42672    0.08004
## oecd_areaCiencias                                    -0.00563    0.14730
## oecd_areaCiencias Sociales, Enseñanza Comercial y Derecho  0.07923    0.14179
## oecd_areaEducación                                   -0.11557    0.14707
## oecd_areaHumanidades y Artes                          0.03723    0.15204
## oecd_areaIngeniería, Industria y Construcción         0.09710    0.14283
## oecd_areaSalud y Servicios Sociales                   0.25483    0.14486
## oecd_areaServicios                                    0.15897    0.14813
## acre_inst_anio                                        0.11108    0.01175
## tipo_inst_2Institutos Profesionales                        NA         NA
## tipo_inst_2Universidades (* Carrera en Convenio)     -0.54385    0.18330
## tipo_inst_2Universidades CRUCH                       -0.21421    0.05017
## tipo_inst_2Universidades Privadas                          NA         NA
## nivel_carrera_2Carreras Técnicas                     -0.36943    0.08149
## nivel_carrera_2Doctorado                              0.43418    0.10314
## nivel_carrera_2Magister                              -0.16564    0.08440
## nivel_carrera_2Postítulo                             -0.08191    0.10612
##                                                      t value Pr(>|t|)
## (Intercept)                                           59.962  < 2e-16 ***
## region_sede_d1                                         3.392 0.000765 ***
## dur_estudio_carr_m2                                    1.983 0.048029 *
## dur_estudio_carr_m3                                    6.061 3.15e-09 ***
## dur_estudio_carr_m4                                    6.325 6.84e-10 ***
## dur_estudio_carr_m5                                    5.931 6.58e-09 ***
## dur_estudio_carr_m6                                    6.886 2.26e-11 ***
## dur_estudio_carr_m7                                    5.507 6.57e-08 ***
## dur_estudio_carr_m8                                    4.698 3.62e-06 ***
## dur_estudio_carr_m9                                    4.847 1.81e-06 ***
## dur_estudio_carr_m10                                   6.279 8.99e-10 ***
## dur_estudio_carr_m11                                   6.258 1.01e-09 ***
## dur_estudio_carr_m12                                   6.526 2.07e-10 ***
## dur_estudio_carr_mMás de 12 semestres                  4.920 1.27e-06 ***
## tipo_inst_1Institutos Profesionales                   -0.159 0.873702
## tipo_inst_1Universidades                               5.331 1.64e-07 ***
```

```
## oecd_areaCiencias                                          -0.038 0.969533
## oecd_areaCiencias Sociales, Enseñanza Comercial y Derecho   0.559 0.576646
## oecd_areaEducación                                         -0.786 0.432418
## oecd_areaHumanidades y Artes                                0.245 0.806714
## oecd_areaIngeniería, Industria y Construcción               0.680 0.497023
## oecd_areaSalud y Servicios Sociales                         1.759 0.079329 .
## oecd_areaServicios                                          1.073 0.283855
## acre_inst_anio                                              9.455  < 2e-16 ***
## tipo_inst_2Institutos Profesionales                            NA       NA
## tipo_inst_2Universidades (* Carrera en Convenio)           -2.967 0.003190 **
## tipo_inst_2Universidades CRUCH                             -4.270 2.45e-05 ***
## tipo_inst_2Universidades Privadas                              NA       NA
## nivel_carrera_2Carreras Técnicas                           -4.533 7.70e-06 ***
## nivel_carrera_2Doctorado                                    4.210 3.17e-05 ***
## nivel_carrera_2Magister                                    -1.963 0.050385 .
## nivel_carrera_2Postítulo                                   -0.772 0.440662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3027 on 396 degrees of freedom
## Multiple R-squared:  0.6459, Adjusted R-squared:   0.62
## F-statistic: 24.91 on 29 and 396 DF,  p-value: < 2.2e-16
```

Forward

```
modelo_nulo <- lm(log(valor_arancel)~ 1, data = muestra_base_f)
add1(modelo_nulo, scope = ~ region_sede_d + dur_estudio_carr_m + oecd_area + acre_inst_anio +  jornada +
```

```
## Single term additions
##
## Model:
## log(valor_arancel) ~ 1
##                     Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>                           102.441 -605.12
## region_sede_d        1     0.965 101.476 -607.15  4.0331   0.04525 *
## dur_estudio_carr_m  12    34.995  67.446 -759.16 17.8572 < 2.2e-16 ***
## oecd_area            7     7.717  94.724 -624.48  4.8648 2.726e-05 ***
## acre_inst_anio       1    10.970  91.471 -651.36 50.8482 4.335e-12 ***
## jornada              4    13.126  89.314 -655.53 15.4685 8.149e-12 ***
## tipo_inst_2          4    28.786  73.655 -737.65 41.1339 < 2.2e-16 ***
## nivel_carrera_2      4    25.958  76.483 -721.60 35.7219 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#add1(update(modelo_nulo ~ . + dur_estudio_carr_m), scope = ~ region_sede_d + dur_estudio_carr_m +  + a
```

Referecnias

Cancino, R. (2010). El Modelo Neoliberal y la Educación Universitaria en Latinoamérica. El caso de la universidad chilena. Sociedad y discurso, AAU, (18), 152-167. http://www.discurso.aau.dk/index_16.htm

Mena, Paula & Dooner, Cecilia. (2006). Arancel de referencia v/s arancel real: diagnóstico e interrogantes iniciales. Calidad en la educación, ISSN 0717-4004, Nº. 24, 2006 (Ejemplar dedicado a: La gestión de las instituciones de educación superior), pags. 285-318. 24. 10.31619/caledu.n24.280.

Salas Opazo, V. & Gómez, E. (2014). "Efectos de los Aranceles de Referencia en las Universidades Chilenas," Investigaciones de Economía de la Educación volume 9, in: Adela García Aracil & Isabel Neira Gómez (ed.), Investigaciones de Economía de la Educación 9, edition 1, volume 9, chapter 13, pages 277-292, Asociación de Economía de la Educación.