Exploración de wage1

Carlos Ortiz

Table of contents

1 Análisis del dataset `wage1`

Este documento explora el dataset `wage1` del paquete de `wooldridge`. Lo primero que debemos hacer es instalar las librerías que vamos a utilizar y después importarlas.

```
`modelsummary` has built-in support to draw text-only (markdown) tables.
  To generate tables in other formats, you must install one or more of
  these libraries:
```

```
install.packages(c(
    "kableExtra",
    "gt",
    "flextable",

  "huxtable",
    "DT"
))
```

```
  Alternatively, you can set markdown as the default table format to
  silence this alert:
```

```
config_modelsummary(factory_default = "markdown")
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
Loading required package: MASS


Attaching package: 'MASS'


The following object is masked from 'package:dplyr':

    select


The following object is masked from 'package:wooldridge':

    cement


Loading required package: msm
```

```
Loading required package: polycor

Loading required package: carData


Attaching package: 'car'


The following object is masked from 'package:dplyr':

    recode


The following object is masked from 'package:purrr':

    some
```

1.1 Exploración inicial de datos

Veamos los primeros datos del dataset

```r
wage <- wage1
head(wage) # primeros cinco registros
```

```
  wage educ exper tenure nonwhite female married numdep smsa northcen south
1 3.10   11     2      0        0      1       0      2    1        0     0
2 3.24   12    22      2        0      1       1      3    1        0     0
3 3.00   11     2      0        0      0       0      2    0        0     0
4 6.00    8    44     28        0      0       1      0    1        0     0
5 5.30   12     7      2        0      0       1      1    0        0     0
6 8.75   16     9      8        0      0       1      0    1        0     0
  west construc ndurman trcommpu trade services profserv profocc clerocc
1    1        0       0        0     0        0        0       0       0
2    1        0       0        0     0        1        0       0       0
3    1        0       0        0     1        0        0       0       0
4    1        0       0        0     0        0        0       0       1
5    1        0       0        0     0        0        0       0       0
6    1        0       0        0     0        0        1       1       0
  servocc    lwage expersq tenursq
1       0 1.131402       4       0
2       1 1.175573     484       4
```

```
3       0 1.098612       4       0
4       0 1.791759    1936     784
5       0 1.667707      49       4
6       0 2.169054      81      64
```

Veamos los nombres de las columnas para revisar si el formato es el adecuado

```
names(wage)
```

```
 [1] "wage"     "educ"     "exper"    "tenure"   "nonwhite" "female"
 [7] "married"  "numdep"   "smsa"     "northcen" "south"    "west"
[13] "construc" "ndurman"  "trcommpu" "trade"    "services" "profserv"
[19] "profocc"  "clerocc"  "servocc"  "lwage"    "expersq"  "tenursq"
```

Nuestro dataset tiene muchas columnas que no utilizaremos, adicionalmente algunas ya vienen transformadas y su análisis no será tan intuitivo con este formato. Vamos a modificarlas haciendo uso de las librerías del `tidyverse`.

```
wage <- wage |>
  dplyr::select(c(wage, lwage, educ, exper, expersq, female, married, tenure)) |>
  mutate(gender = ifelse(female == 1, "female", "male"),
         fam.status = ifelse(married == 1, "married", "single")) |>
  dplyr::select(-c(female, married))

wage[, "gender"] <- as.factor(wage[, "gender"])
wage[, "fam.status"] <- as.factor(wage[, "fam.status"])
```

Veamos los datos nuevamente

```
head(wage)
```

```
  wage    lwage educ exper expersq tenure gender fam.status
1 3.10 1.131402   11     2       4      0 female     single
2 3.24 1.175573   12    22     484      2 female    married
3 3.00 1.098612   11     2       4      0   male     single
4 6.00 1.791759    8    44    1936     28   male    married
5 5.30 1.667707   12     7      49      2   male    married
6 8.75 2.169054   16     9      81      8   male    married
```

Con los datos listos para analizar vamos mostrar un resumen de los estadísticos más importantes:

```
summary(wage)
```

```
      wage            lwage              educ           exper      
 Min.   : 0.530   Min.   :-0.6349   Min.   : 0.00   Min.   : 1.00  
 1st Qu.: 3.330   1st Qu.: 1.2030   1st Qu.:12.00   1st Qu.: 5.00  
 Median : 4.650   Median : 1.5369   Median :12.00   Median :13.50  
 Mean   : 5.896   Mean   : 1.6233   Mean   :12.56   Mean   :17.02  
 3rd Qu.: 6.880   3rd Qu.: 1.9286   3rd Qu.:14.00   3rd Qu.:26.00  
 Max.   :24.980   Max.   : 3.2181   Max.   :18.00   Max.   :51.00  
    expersq          tenure          gender      fam.status 
 Min.   :   1.0   Min.   : 0.000   female:252   married:320  
 1st Qu.:  25.0   1st Qu.: 0.000   male  :274   single :206  
 Median : 182.5   Median : 2.000                             
 Mean   : 473.4   Mean   : 5.105                             
 3rd Qu.: 676.0   3rd Qu.: 7.000                             
 Max.   :2601.0   Max.   :44.000                             
```

Podemos continuar con el análisis gráfico.

1.2 Análisis gráfico

Los gráficos son más elocuentes que la descripción estadística. Con ayuda de estos veremos la distribución de las variables tanto discretas y continuas como categóricas; y la relación de cada una con `wage` que para nuestros intereses es la variable objetivo.

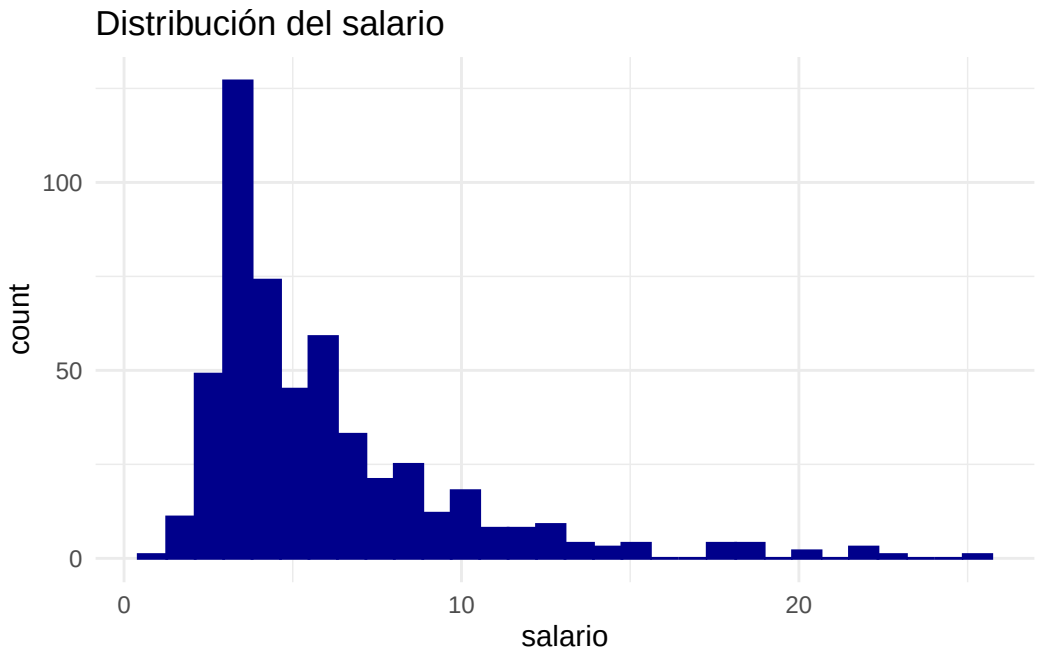
1.2.1 Análisis univariado

El análisis univariado podemos simplificarlo con ayuda de gráficos de barras para variables categóricas y discretas (si el rango no es muy amplio) e histogramas para las variables continuas.

1.2.1.1 `wage`

```
ggplot(data = wage) +
  geom_histogram(mapping = aes(x = wage), color = 'darkblue', fill='darkblue') +
  labs(x = 'salario',
       title = 'Distribución del salario') +
  theme_minimal()
```

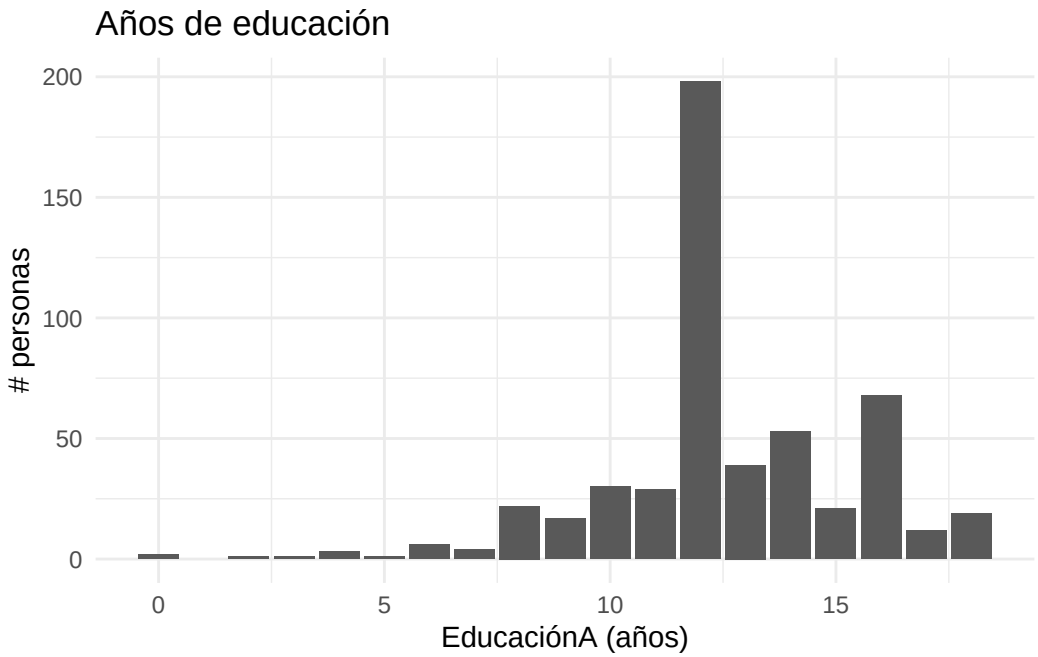
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



El gráfico nos muestra una asimetría positiva (sesgo a la derecha) de la variable `wage`, esto es común en esta variable dada la desigualdad en la distribución del ingreso (unos pocos ganan mucho más que la mayoría).

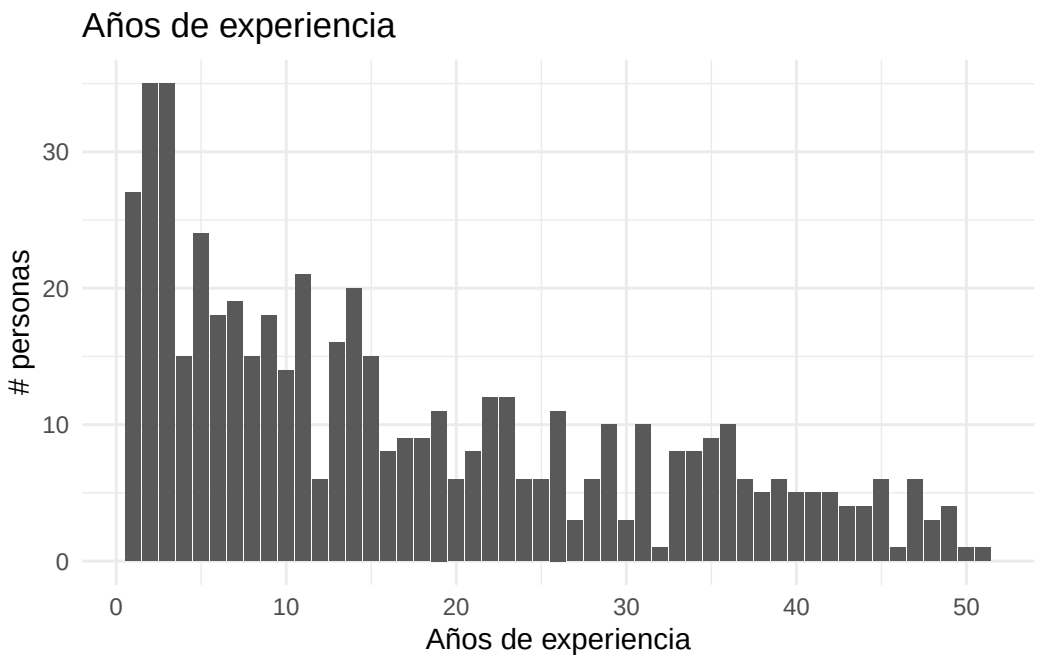
1.2.1.2 `educ`

```
ggplot(data = wage) +
  geom_bar(mapping = aes(x = educ)) +
  labs(x = "EducaciónA (años)",
       y = "# personas",
       title = "Años de educación") +
  theme_minimal()
```



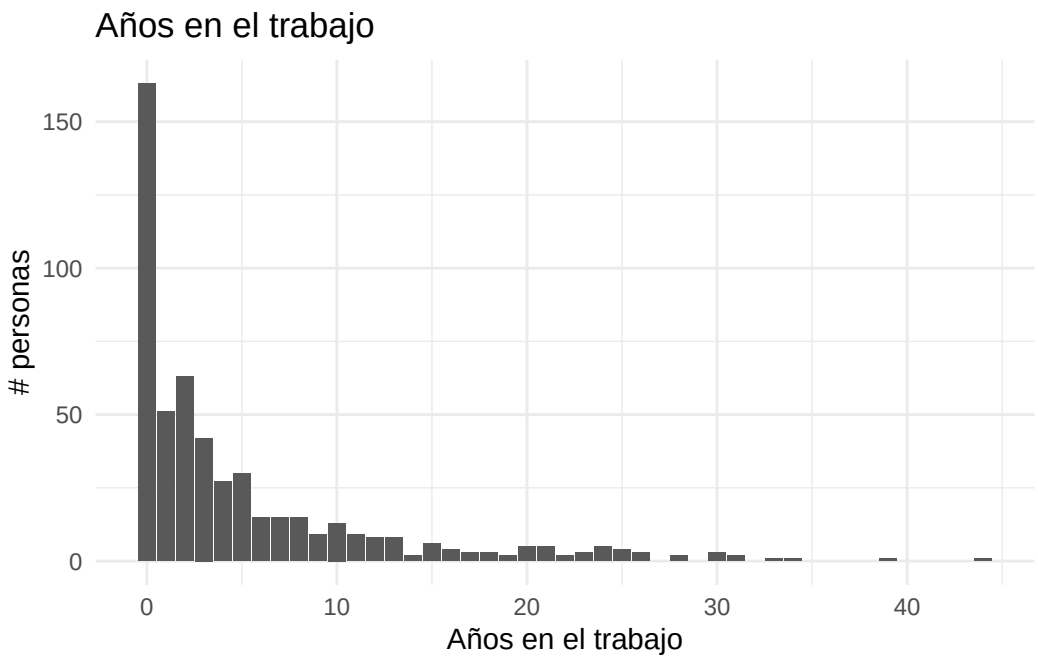
1.2.1.3 `exper`

```
ggplot(data = wage) +
  geom_bar(mapping = aes(x = exper)) +
  labs(x = "Años de experiencia",
       y = "# personas",
       title = "Años de experiencia") +
  theme_minimal()
```



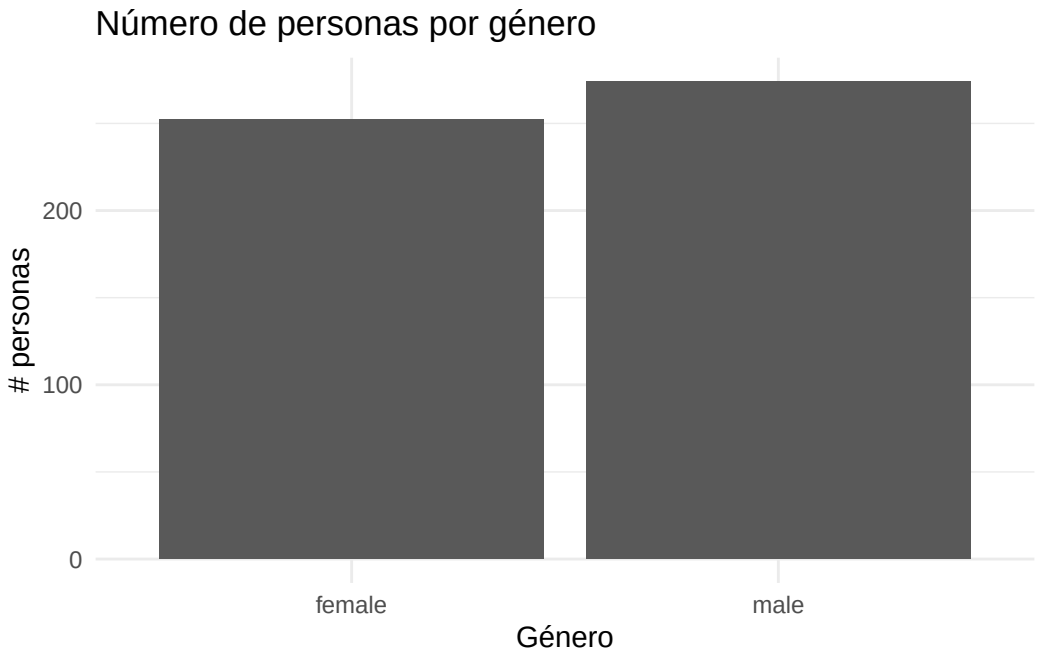
1.2.1.4 `tenure`

```
ggplot(data = wage) +
  geom_bar(mapping = aes(x = tenure)) +
  labs(x = "Años en el trabajo",
       y = "# personas",
       title = "Años en el trabajo") +
  theme_minimal()
```



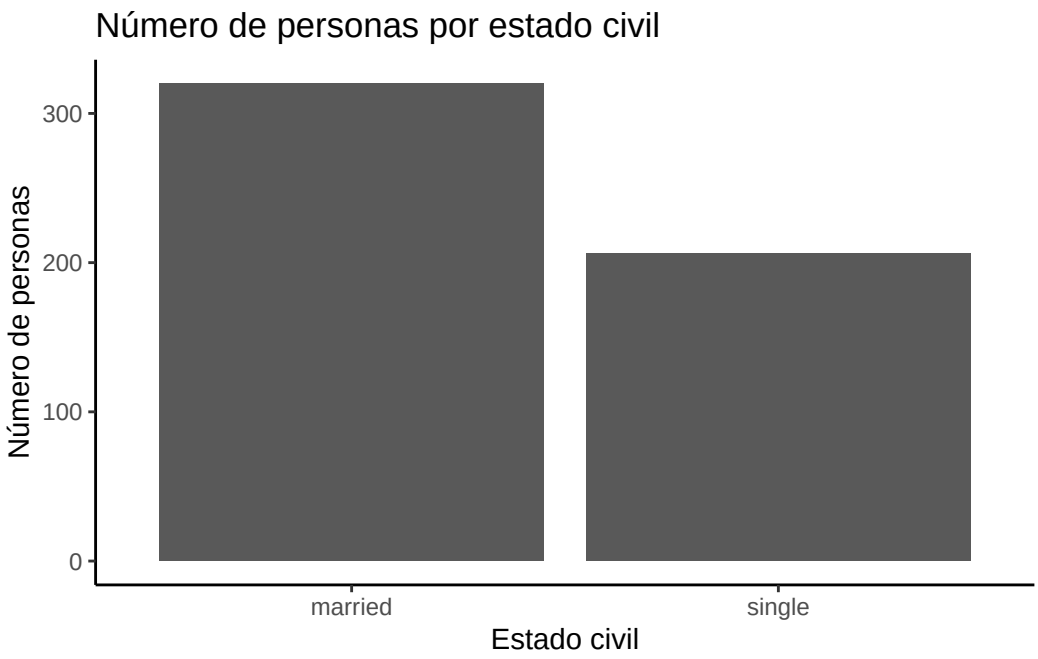
1.2.1.5 `gender`

```
ggplot(data = wage) +
  geom_bar(mapping = aes(x = gender)) +
  labs(x = "Género",
       y = "# personas",
       title = "Número de personas por género") +
  theme_minimal()
```



1.2.1.6 `fam.status`

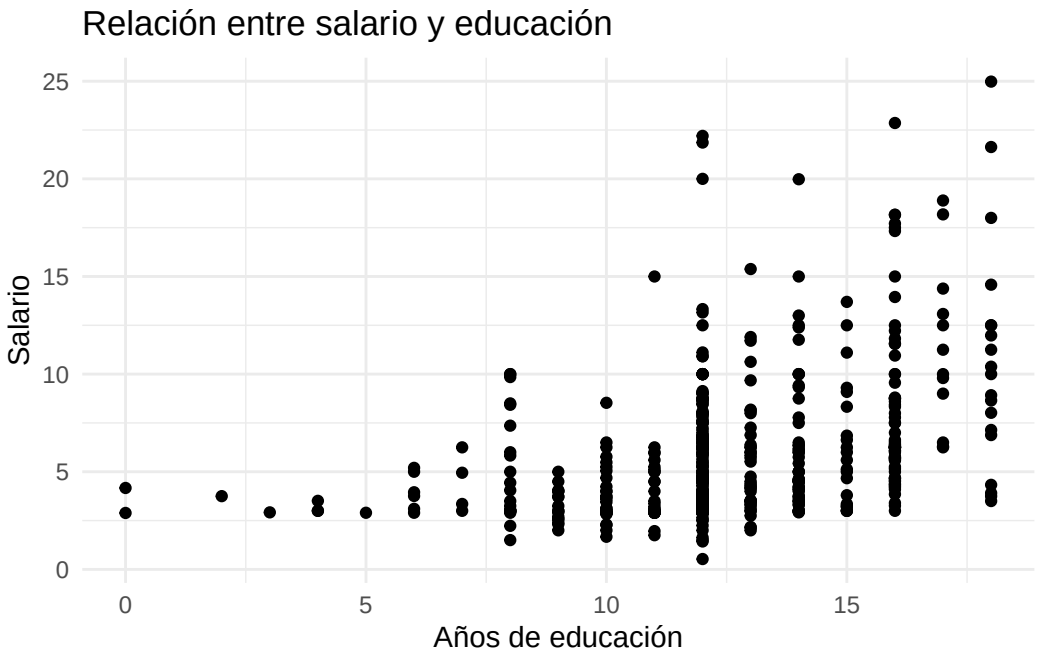
```
ggplot(data = wage) +
  geom_bar(mapping = aes(x = fam.status)) +
  labs(x = "Estado civil",
       y = "Número de personas",
       title = "Número de personas por estado civil") +
  theme_classic()
```



1.2.2 Análisis bivariado

1.2.2.1 `wage vs. educ`

```
ggplot(data = wage) +
  geom_point(mapping = aes(x = educ, y = wage)) +
  labs(x = "Años de educación",
       y = "Salario",
       title = "Relación entre salario y educación") +
  theme_minimal()
```

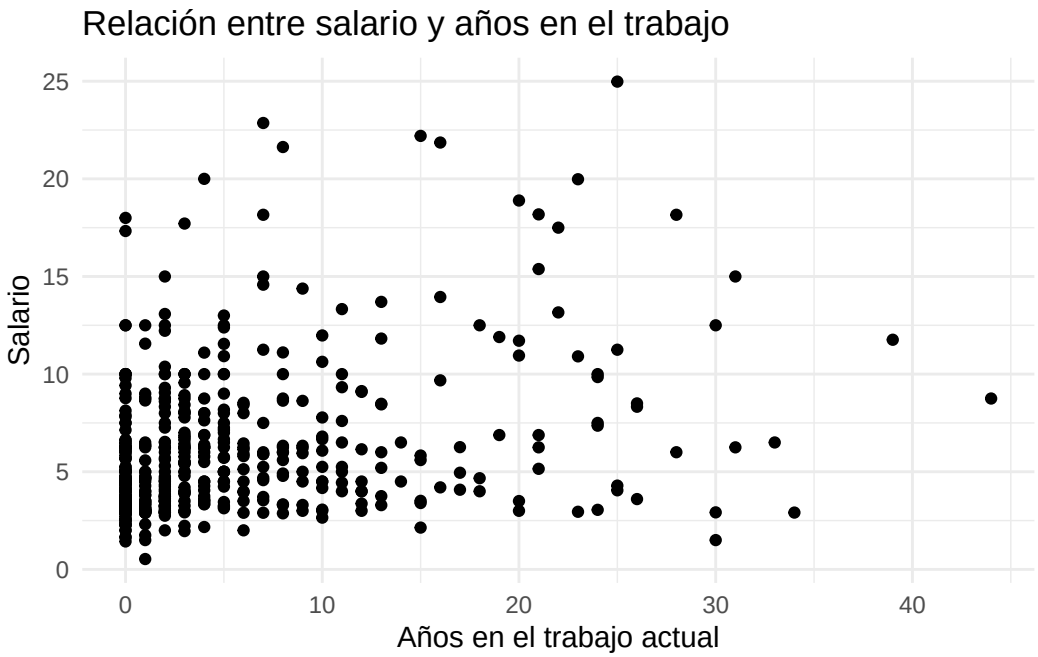


1.2.2.2 `wage vs. exper`

```
ggplot(data = wage) +
  geom_point(mapping = aes(x = exper, y = wage)) +
  labs(x = "Años de experiencia",
       y = "Salario",
       title = "Relación entre salario y experiencia") +
  theme_minimal()
```

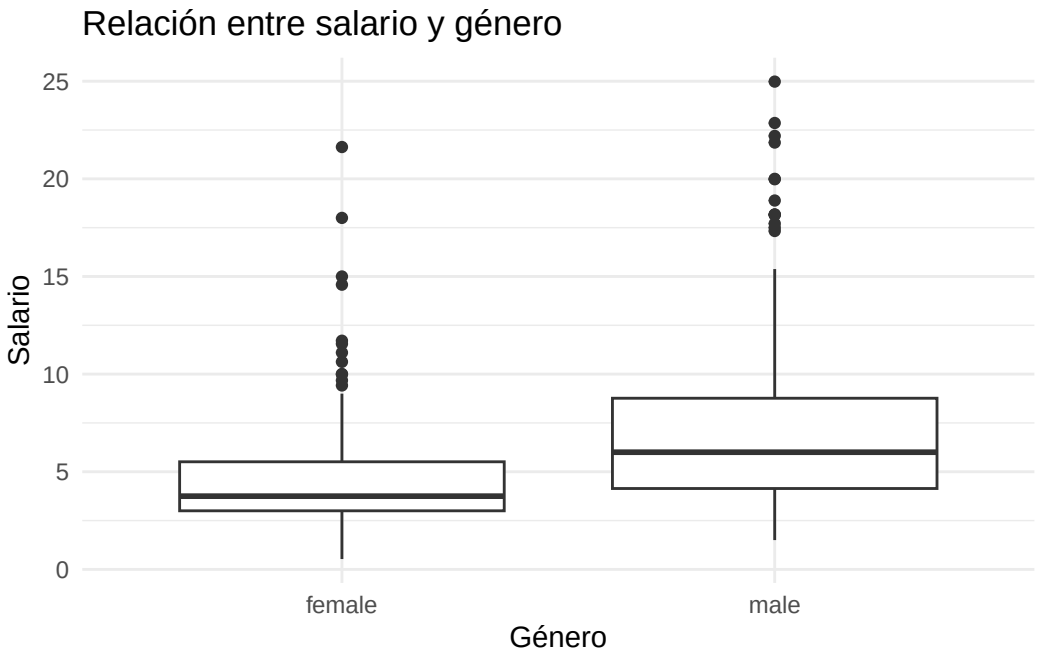
1.2.2.3 `wage vs. tenure`

```
ggplot(data = wage) +
  geom_point(mapping = aes(x = tenure, y = wage)) +
  labs(x = "Años en el trabajo actual",
       y = "Salario",
       title = "Relación entre salario y años en el trabajo") +
  theme_minimal()
```



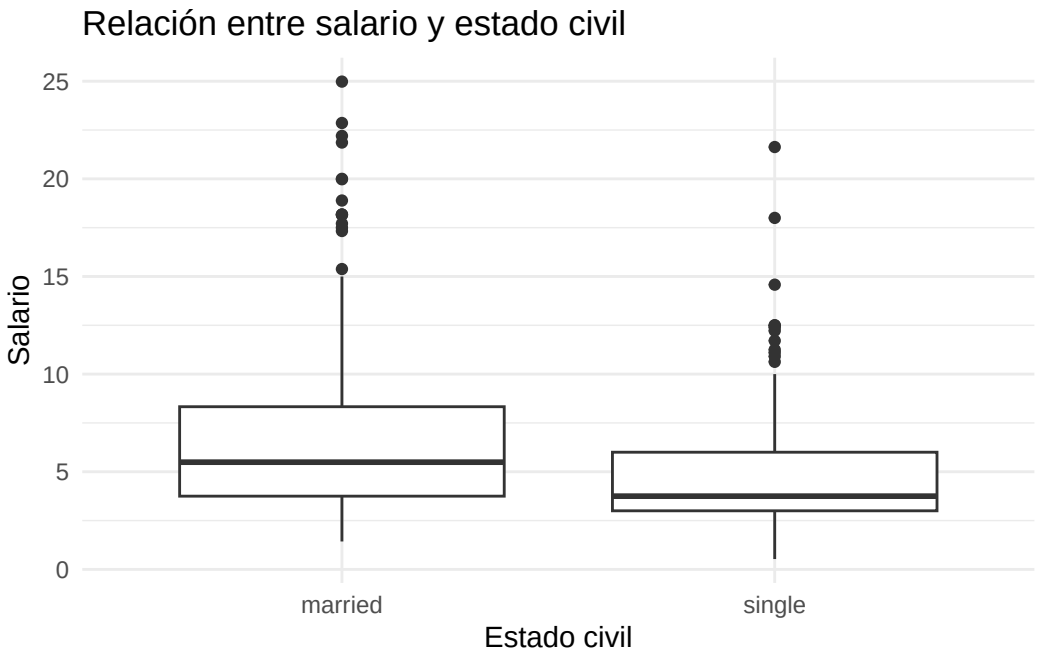
1.2.2.4 `wage vs. gender`

```
ggplot(data = wage) +
  geom_boxplot(mapping = aes(x = gender, y = wage)) +
  labs(x = "Género",
       y = "Salario",
       title = "Relación entre salario y género") +
  theme_minimal()
```



1.2.2.5 `wage vs. fam.status`

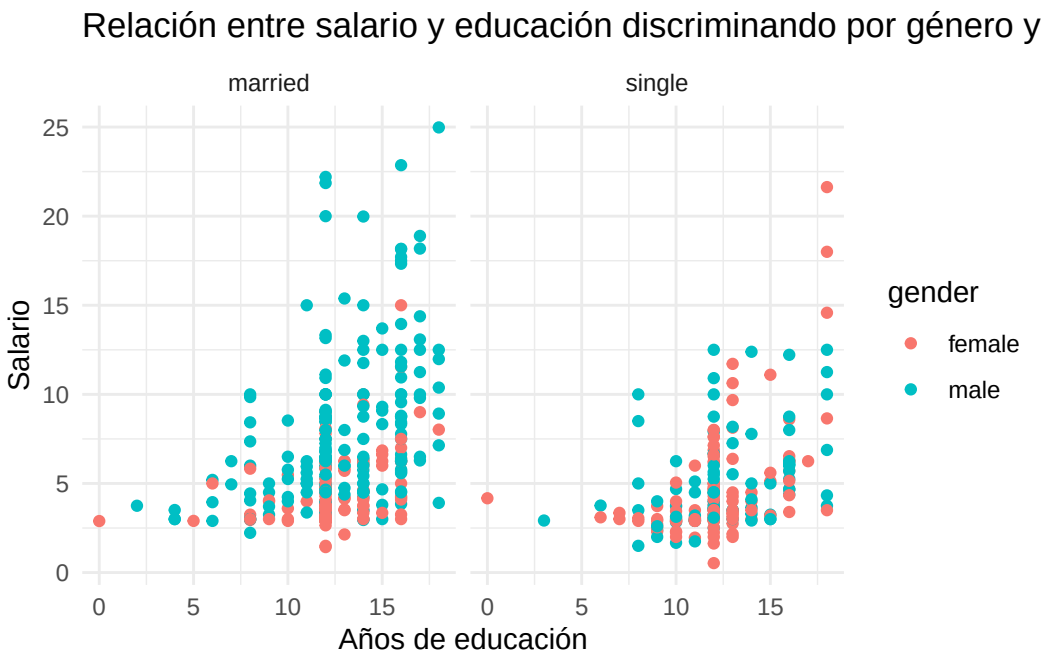
```
ggplot(data = wage) +
  geom_boxplot(mapping = aes(x = fam.status, y = wage)) +
  labs(x = "Estado civil",
       y = "Salario",
       title = "Relación entre salario y estado civil") +
  theme_minimal()
```



1.2.3 Análisis multivariado

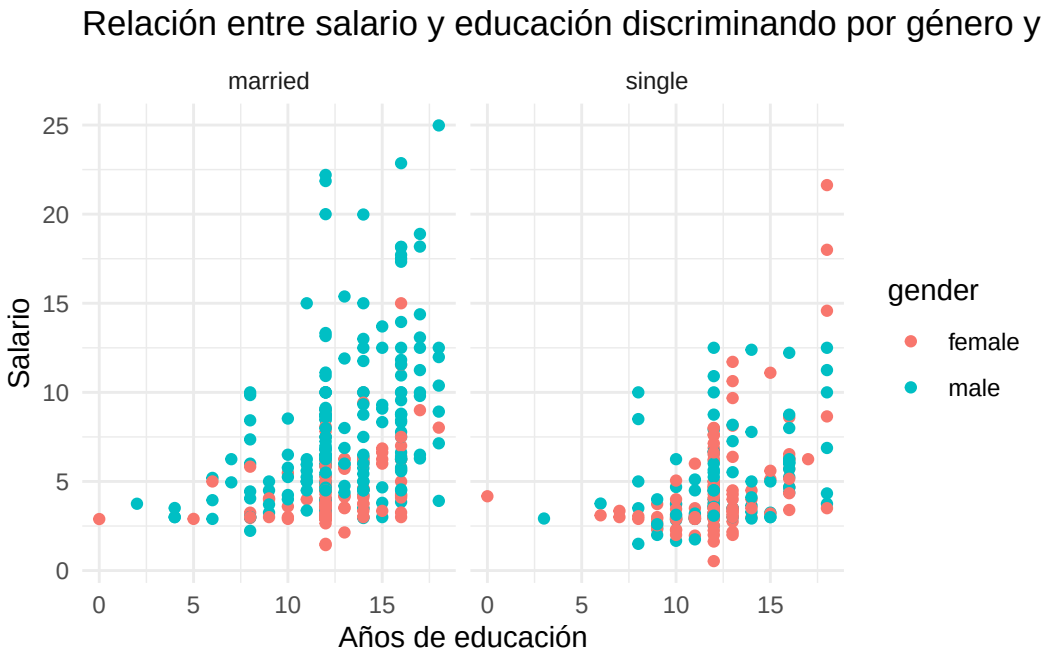
1.2.3.1 `wage vs. educ vs. gender vs. fam.status`

```
ggplot(data = wage) +
  geom_point(mapping = aes(x = educ, y = wage, color = gender)) +
  facet_grid(cols=vars(fam.status)) +
  labs(x = "Años de educación",
       y = "Salario",
       title = "Relación entre salario y educación discriminando por género y estado civil") +
  theme_minimal()
```



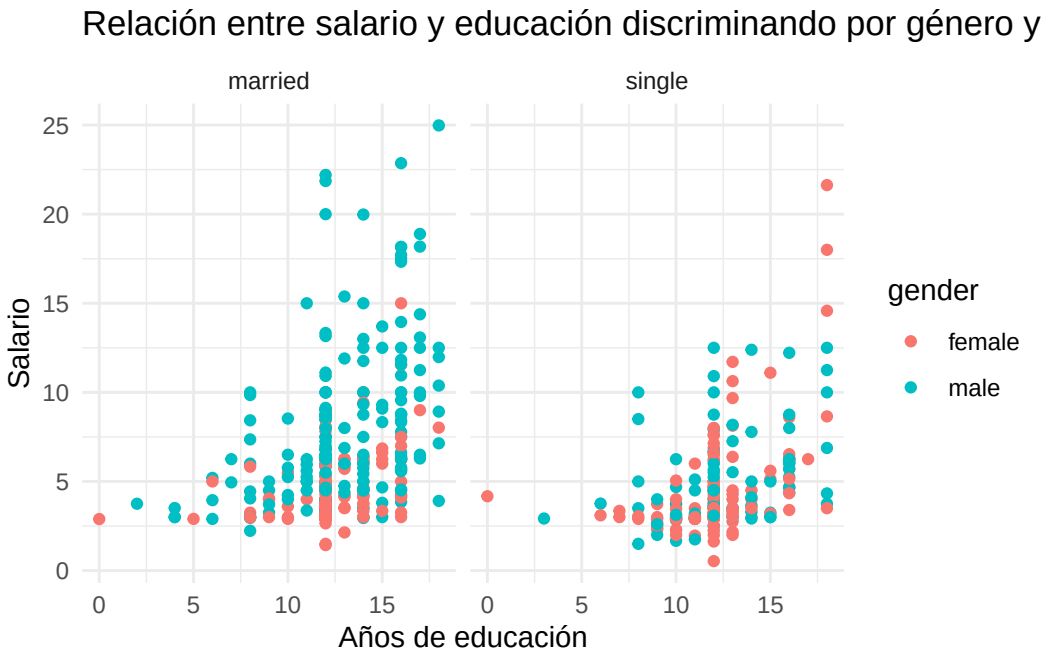
1.2.3.2 `wage vs. exper vs. gender vs. fam.status`

```
ggplot(data = wage) +
  geom_point(mapping = aes(x = educ, y = wage, color = gender)) +
  facet_grid(cols=vars(fam.status)) +
  labs(x = "Años de educación",
       y = "Salario",
       title = "Relación entre salario y educación discriminando por género y estado civil") +
  theme_minimal()
```



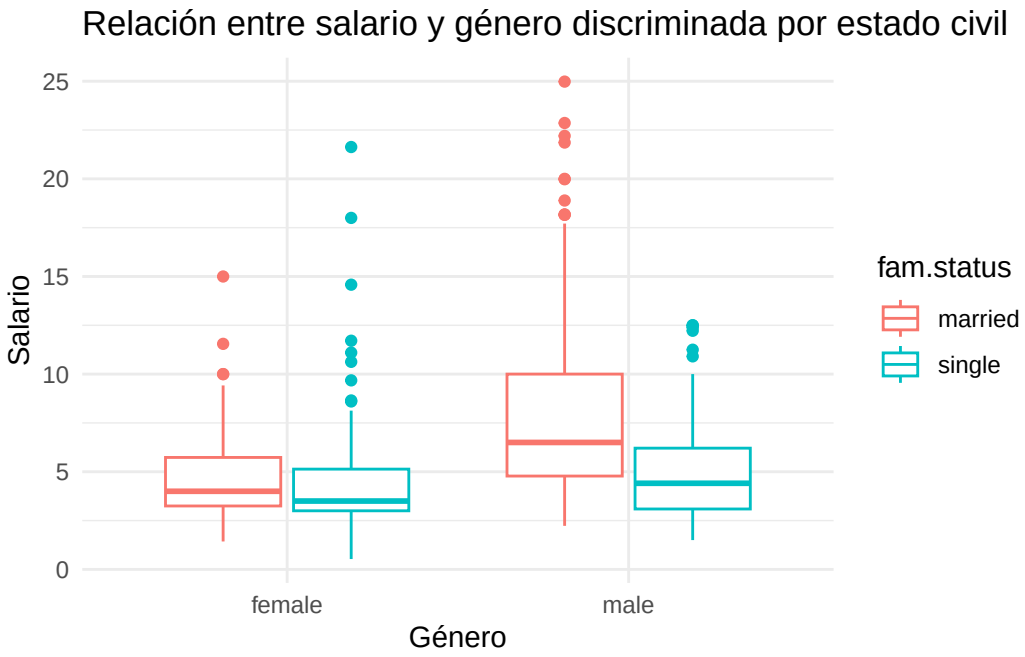
1.2.3.3 `wage vs. teure vs. gender vs. fam.status`

```
ggplot(data = wage) +
  geom_point(mapping = aes(x = educ, y = wage, color = gender)) +
  facet_grid(cols=vars(fam.status)) +
  labs(x = "Años de educación",
       y = "Salario",
       title = "Relación entre salario y educación discriminando por género y estado civil") +
  theme_minimal()
```



1.2.3.4 `wage vs. gender vs. fam.status`

```
ggplot(data = wage) +
  geom_boxplot(mapping = aes(x = gender, y = wage, color = fam.status)) +
  labs(x = "Género",
       y = "Salario",
       title = "Relación entre salario y género discriminada por estado civil") +
  theme_minimal()
```



1.3 Análisis estadístico

1.3.1 Análisis de correlación

1.3.1.1 Correlación de Pearson

```
numeric <- wage |>
  dplyr::select(-c(gender, fam.status))
cor(numeric)
```

```
              wage      lwage        educ      exper     expersq      tenure
wage     1.00000000 0.93706171  0.40590333  0.1129034  0.03023781  0.34688957
lwage    0.93706171 1.00000000  0.43105276  0.1113729  0.02329833  0.32553794
educ     0.40590333 0.43105276  1.00000000 -0.2995418 -0.33125594 -0.05617257
exper    0.11290344 0.11137287 -0.29954184  1.0000000  0.96097091  0.49929145
expersq  0.03023781 0.02329833 -0.33125594  0.9609709  1.00000000  0.45922323
tenure   0.34688957 0.32553794 -0.05617257  0.4992914  0.45922323  1.00000000
```

1.3.1.2 Correlación de Spearman

```
numeric <- wage |>
  dplyr::select(-c(gender, fam.status))
cor(numeric, method = "spearman")
```

```
            wage     lwage        educ      exper    expersq     tenure
wage    1.0000000 1.0000000  0.45776942  0.1744161  0.1744161 0.38078668
```

```
lwage   1.0000000 1.0000000  0.45776942  0.1744161  0.1744161 0.38078668
educ    0.4577694 0.4577694  1.00000000 -0.1989940 -0.1989940 0.04847676
exper   0.1744161 0.1744161 -0.19899396  1.0000000  1.0000000 0.48724650
expersq 0.1744161 0.1744161 -0.19899396  1.0000000  1.0000000 0.48724650
tenure  0.3807867 0.3807867  0.04847676  0.4872465  0.4872465 1.00000000
```

1.3.1.3 Correlación biserial puntual

```
biserial.cor(x = wage$wage, y = wage$gender)
```

```
[1] -0.3400979
```

```
biserial.cor(x = wage$wage, y = wage$fam.status)
```

```
[1] 0.2288172
```

1.4 Análisis estadístico

Vamos a identificar si existen diferencias significativas entre el ingreso de hombres y mujeres, si existen diferencias significativas entre el ingreso de hombres casados y hombres solteros, y de mujeres casadas y mujeres solteras. Por último, analizaremos si existe alguna relación entre el estado civil y el género. Para llevar a cabo este análisis, seguiremos los pasos listados a continuación:

1. Construir los vectores a comparar.
2. Verificar si cada uno sigue una distribución normal o no.
3. Si sigue una distribución normal, realizaremos una comparación de varianzas y después una comparación de medias.
4. Si no sigue una distribución normal, llevaremos a cabo una prueba `Mann-Whitney`.
5. Después construiremos una tabla de contingencia con las variables categóricas y después aplicaremos una prueba χ^2.

```
filters <- list(
"wage.male" = wage[wage["gender"] == "male", "wage"],

"wage.female" = wage[wage["gender"] == "female", "wage"],

"w.fmarried"= wage[wage["gender"] == "female" & wage["fam.status"] == "married", "wage"],
"w.fsingle" = wage[wage["gender"] == "female" & wage["fam.status"] == "single", "wage"],

"w.mmarried" = wage[wage["gender"] == "male" & wage["fam.status"] == "married", "wage"],
"w.msingle" = wage[wage["gender"] == "female" & wage["fam.status"] == "single", "wage"]
)
cont.table <- table(c(wage$gender, wage$fam.status))
```

1.4.1 Prueba de normalidad

```
for (df in filters){
  print(shapiro.test(df))
  print("**********")
```

```
  }
```

```
    Shapiro-Wilk normality test

data:  df
W = 0.8605, p-value = 4.989e-15

[1] "**********"

    Shapiro-Wilk normality test

data:  df
W = 0.74449, p-value < 2.2e-16

[1] "**********"

    Shapiro-Wilk normality test

data:  df
W = 0.81754, p-value = 1.627e-11

[1] "**********"

    Shapiro-Wilk normality test

data:  df
W = 0.7078, p-value = 3.864e-14

[1] "**********"

    Shapiro-Wilk normality test

data:  df
W = 0.8659, p-value = 7.476e-12

[1] "**********"

    Shapiro-Wilk normality test

data:  df
W = 0.7078, p-value = 3.864e-14

[1] "**********"
```

Parece que ninguno proviene de una distribución normal según la prueba de `shapiro`. Sin embargo, llevaremos a cabo una prueba t para ejemplificar (el t-test es igualmente robusto cuando la distribución no es normal).

1.4.2 Prueba de varianzas

```
  leveneTest(wage$wage, group = wage$gender)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   1  35.427 4.85e-09 ***
```

```
      524
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Según la prueba de Levene, las varianzas son diferentes

1.4.3 t-test

```
t.test(filters$wage.male, filters$wage.female, var.equal = FALSE)
```

```
    Welch Two Sample t-test

data:  filters$wage.male and filters$wage.female
t = 8.44, df = 456.33, p-value = 4.243e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.926971 3.096690
sample estimates:
mean of x mean of y
 7.099489  4.587659
```

1.4.4 Prueba de Mann-Whitney

```
wilcox.test(filters$wage.male, filters$wage.female)
```

```
    Wilcoxon rank sum test with continuity correction

data:  filters$wage.male and filters$wage.female
W = 49798, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

1.4.5 Prueba χ^2

```
chisq.test(cont.table)
```

```
    Chi-squared test for given probabilities

data:  cont.table
X-squared = 25.627, df = 3, p-value = 1.141e-05
```

Según la prueba χ^2 las variables no son independientes.

1.5 Análisis de regresión

1.5.1 Modelos

Modelo 1: $wage = \beta_0 + \beta_1 educ + u$

Modelo 2: $wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$

Modelo 3: $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 expersq + u$

Modelo 4: $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 expersq + \beta_4 tenure + u$

Modelo 5: $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 expersq + \beta_4 tenure + \beta_5 male + u$

Modelo 6: $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 expersq + \beta_4 tenure + \beta_5 male + \beta_6 single + u$

1.5.2 Regresión

```
wage.model <- dummy_cols(wage, remove_first_dummy = TRUE)
```

```
lista.modelos <- list(
  "model_1" = lm(wage ~ educ, data = wage.model),
  "model_2" = lm(wage ~ educ + exper, data = wage.model),
  "model_3" = lm(wage ~ educ + exper + expersq, data = wage.model),
  "model_4" = lm(wage ~ educ + exper + expersq + tenure, data = wage.model),
  "model_5" = lm(wage ~ educ + exper + expersq + tenure + gender_male, data = wage.model),
  "model_6" = lm(wage ~ educ + exper + expersq + tenure + gender_male + fam.status_single, data = wage.model)
)
```

1.5.3 Resultados

```
modelsummary(lista.modelos, stars=TRUE)
```

	model_1	model_2	model_3	model_4	model_5	model_6
(Intercept)	-0.905	-3.391***	-3.965***	-3.420***	-3.910***	-3.805***
	(0.685)	(0.767)	(0.752)	(0.718)	(0.691)	(0.768)
educ	0.541***	0.644***	0.595***	0.556***	0.530***	0.528***
	(0.053)	(0.054)	(0.053)	(0.051)	(0.049)	(0.049)
exper		0.070***	0.268***	0.205***	0.205***	0.200***
		(0.011)	(0.037)	(0.036)	(0.034)	(0.037)
expersq			-0.005***	-0.004***	-0.004***	-0.004***
			(0.001)	(0.001)	(0.001)	(0.001)
tenure				0.161***	0.134***	0.133***
				(0.021)	(0.021)	(0.021)

	model_1	model_2	model_3	model_4	model_5	model_6
gender_male					1.790***	1.779***
					(0.258)	(0.260)
fam.status_single						-0.092
						(0.294)
:————	———-:	———-:	———-:	———-:	———-:	———-:
Num.Obs.	526	526	526	526	526	526
R2	0.165	0.225	0.269	0.343	0.399	0.399
R2 Adj.	0.163	0.222	0.265	0.338	0.393	0.392
AIC	2777.4	2739.9	2711.1	2657.3	2612.6	2614.5
BIC	2790.2	2757.0	2732.5	2682.9	2642.5	2648.6
Log.Lik.	-1385.712	-1365.969	-1350.565	-1322.646	-1299.304	-1299.253
RMSE	3.37	3.25	3.15	2.99	2.86	2.86

Note: ^^ + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001