M5_AI1_CANOJORGE

JORGE CANO

2024-10-04

EJERCICIOS

1. Propón la regresión para explicar el salario a través de los años de servicio y los años desde el doctorado. Justifica si era lo esperado o no y si difiere justificar la razón de dicho diferimiento. Obtén la suma de residuos al cuadrado, el coeficiente de determinación y el coeficiente de determinación corregido del modelo.

#**COMETARIO**: Una vez propuesta la progresión, observo que el resultado no era el esperado, pues nos encontramos antes una relación negativa entre el salario y los años de servicios, lo cual sorprende, pues la lógica implicaría que a mayor años de servicio, tendrías un mayor salario. En cuánto a los años de doctorado si que implica relación positiva, lo cual parece lógico. Obtenemos tanto un R2 y R2 ajustado muy bajos, lo cual implica que el modelos no está correctamente representado, por lo que debemos considerar agregar más variables o realizar transformaciónes para mejorar la capacidad predictiva del modelo

```
## salary ~ yrs.service + yrs.since.phd
```

```
##
## Call:
## lm(formula = formula, data = bbdd_salarios)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79735 -19823  -2617  15149 106149
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    89912.2     2843.6  31.620  < 2e-16 ***
## yrs.service     -629.1      254.5  -2.472   0.0138 *
## yrs.since.phd   1562.9      256.8   6.086 2.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27360 on 394 degrees of freedom
## Multiple R-squared:  0.1883, Adjusted R-squared:  0.1842
## F-statistic: 45.71 on 2 and 394 DF,  p-value: < 2.2e-16
```

```
##
```

```
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 salary
## ------------------------------------------------
## yrs.service                   -629.101**
##                                (254.469)
##
## yrs.since.phd                1,562.889***
##                                (256.820)
##
## Constant                     89,912.180***
##                                (2,843.560)
##
## ------------------------------------------------
## Observations                      397
## R2                               0.188
## Adjusted R2                      0.184
## Residual Std. Error    27,357.140 (df = 394)
## F Statistic           45.714*** (df = 2; 394)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

2. Incluye el género en el modelo. Valora la nueva suma de residuos al cuadrado.

#**COMETARIO**:Si valoramos la RSS del nuevo modelo, vemos que no mejora respecto la propuesta anterior. Para determinar esto, hemos comparado el test F dónde para un mismo nivel de significación <0.01, el primero modelo presenta una F más alta, explicando así una mayor proporción en la variabilidad en los datos

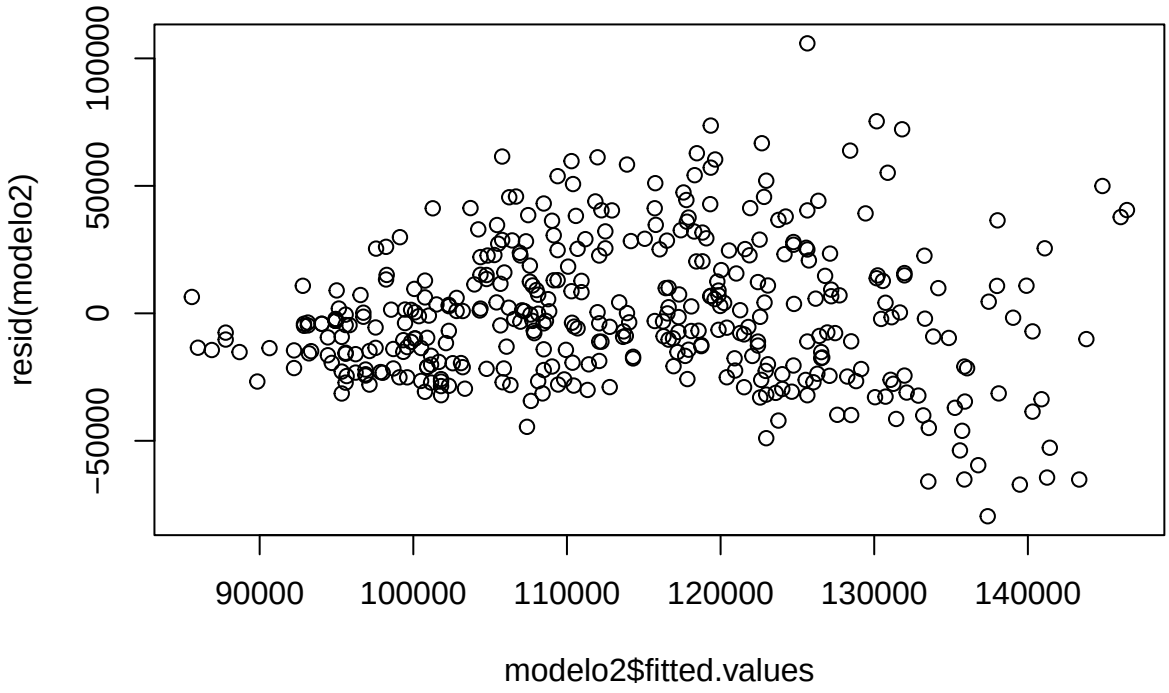
```
## salary ~ yrs.service + yrs.since.phd + sex

##
## Call:
## lm(formula = formula2, data = bbdd_salarios)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79586 -19564  -3018  15071 105898
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    82875.9     4800.6  17.264  < 2e-16 ***
## yrs.service     -649.8      254.0  -2.558   0.0109 *
## yrs.since.phd   1552.8      256.1   6.062 3.15e-09 ***
## sexMale         8457.1     4656.1   1.816   0.0701 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27280 on 393 degrees of freedom
## Multiple R-squared:  0.1951, Adjusted R-squared:  0.189
```

```
## F-statistic: 31.75 on 3 and 393 DF,  p-value: < 2.2e-16
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 salary
## ------------------------------------------------
## yrs.service                   -649.761**
##                                (253.985)
##
## yrs.since.phd                1,552.757***
##                                (256.134)
##
## sexMale                       8,457.065*
##                               (4,656.137)
##
## Constant                     82,875.910***
##                               (4,800.630)
##
## ------------------------------------------------
## Observations                     397
## R2                              0.195
## Adjusted R2                     0.189
## Residual Std. Error    27,277.670 (df = 393)
## F Statistic           31.754*** (df = 3; 393)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
## [1] "Suma de errores al cuadrado"
```

```
## [1] 9242.393
```



3. Justifica, a través del coeficiente de determinación corregido, si el género es una variable a tener en cuenta para mejorar el modelo de predicción del salario.

#**COMETARIO**: Si incluímos en el modelo la variable género, mejoramos muy residualmente la predicción del salario, pues la mejor es muy residual, pasando de un r2_ajustado en el modelo 1 de 0.184 a 0.189 en el modelo 2. Cómo consecuencia estamos añadiendo al modelo una variable adicional que no es determinaste y puede generar problemas de consistencia y explicabilidad

```
## [1] "R2_adjusted_1"
```

```
## [1] 0.1842252
```

```
## [1] "R2_adjusted_2"
```

```
## [1] 0.1889577
```

4. Indica cómo incrementa el salario ante una variación en los años de servicio.

#**COMETARIO**:Cómo hemos podido contrastar, la variable años de servicio, tiene una relación negativa, lo que supone que al incrementar los años de servicio, el salario decrece

5. Indica cómo afecta a las betas del modelo si dividimos el salario por mil para expresarlo en miles.

#**COMETARIO**: Las betas se ven divididas por el mismo número, en este caso 1000

```
##        rank discipline yrs.since.phd yrs.service  sex salary salary_thousands
## 1      Prof          B            19          18 Male 139750           139.75
## 2      Prof          B            20          16 Male 173200           173.20
## 3  AsstProf          B             4           3 Male  79750            79.75
## 4      Prof          B            45          39 Male 115000           115.00
## 5      Prof          B            40          41 Male 141500           141.50
## 6 AssocProf          B             6           6 Male  97000            97.00
```

```
##
## Call:
## lm(formula = formula3, data = bbdd_salarios_ajus)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.586 -19.564  -3.018  15.071 105.898
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    82.8759     4.8006  17.264  < 2e-16 ***
## yrs.service    -0.6498     0.2540  -2.558   0.0109 *
## yrs.since.phd   1.5528     0.2561   6.062 3.15e-09 ***
## sexMale         8.4571     4.6561   1.816   0.0701 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.28 on 393 degrees of freedom
## Multiple R-squared:  0.1951, Adjusted R-squared:  0.189
## F-statistic: 31.75 on 3 and 393 DF,  p-value: < 2.2e-16
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                           salary_thousands
## ------------------------------------------------
## yrs.service                   -0.650**
##                                (0.254)
##
## yrs.since.phd                 1.553***
##                                (0.256)
##
## sexMale                        8.457*
##                                (4.656)
##
## Constant                      82.876***
##                                (4.801)
##
## ------------------------------------------------
```

```
## Observations                    397
## R2                             0.195
## Adjusted R2                    0.189
## Residual Std. Error      27.278 (df = 393)
## F Statistic           31.754*** (df = 3; 393)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

6.Con el modelo anterior, teniendo en cuenta años de servicio y años desde el doctorado, realiza el mismo modelo, pero con el logaritmo neperiano del salario. Indica si se mantienen los signos de las betas obtenidas.

#**COMETARIO**: Aplicando el logaritmo neperiano sobre la variable respuesta, se mantiene los signos, pero tiene menor impacto en el modelo los años de servicio

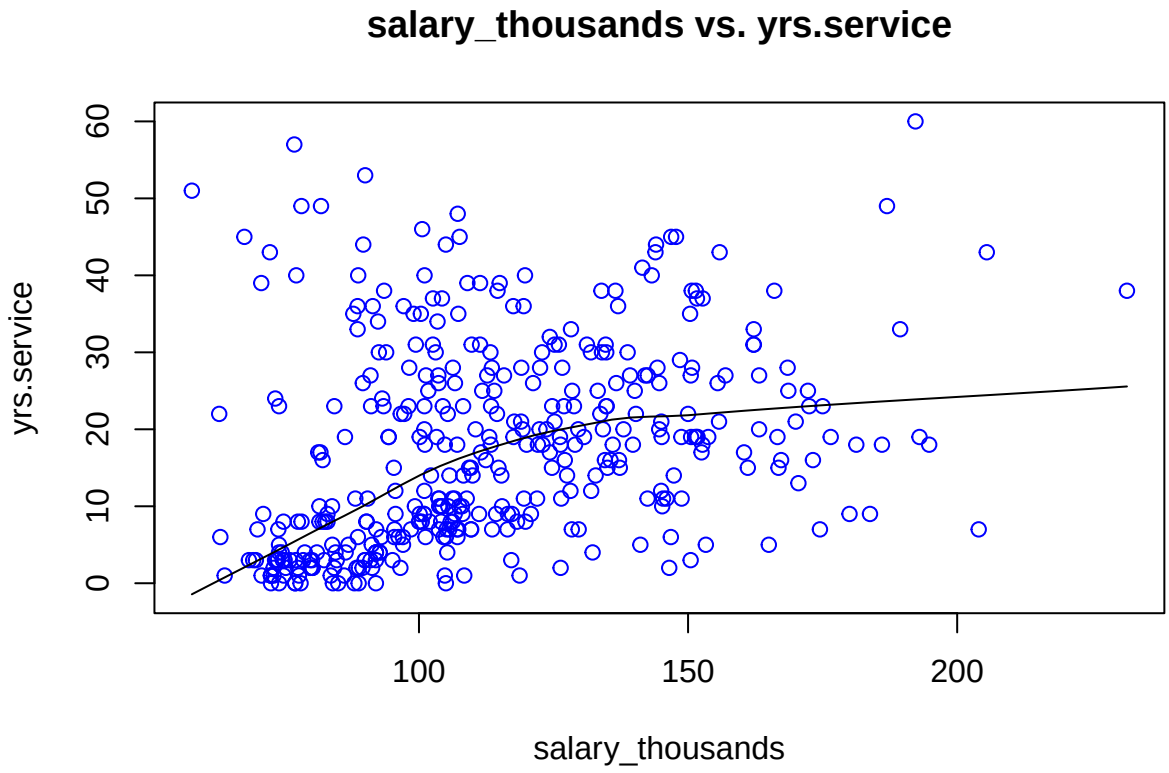
```
##
## Call:
## lm(formula = formula4, data = bbdd_salarios_ajus)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85152 -0.17076 -0.00342  0.15606  0.64229
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.492590   0.024352 184.486  < 2e-16 ***
## yrs.service   -0.005316   0.002179  -2.439   0.0152 *
## yrs.since.phd  0.013471   0.002199   6.125  2.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2343 on 394 degrees of freedom
## Multiple R-squared:  0.1932, Adjusted R-squared:  0.1892
## F-statistic: 47.19 on 2 and 394 DF,  p-value: < 2.2e-16
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                        log(salary_thousands)
## ------------------------------------------------
## yrs.service                   -0.005**
##                                (0.002)
##
## yrs.since.phd                 0.013***
##                                (0.002)
##
## Constant                      4.493***
##                                (0.024)
##
```

```
## ------------------------------------------------
## Observations                      397
## R2                               0.193
## Adjusted R2                      0.189
## Residual Std. Error       0.234 (df = 394)
## F Statistic           47.189*** (df = 2; 394)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

7. Indica cómo incrementa el salario ante una variación, en los años de servicio en este nuevo modelo.

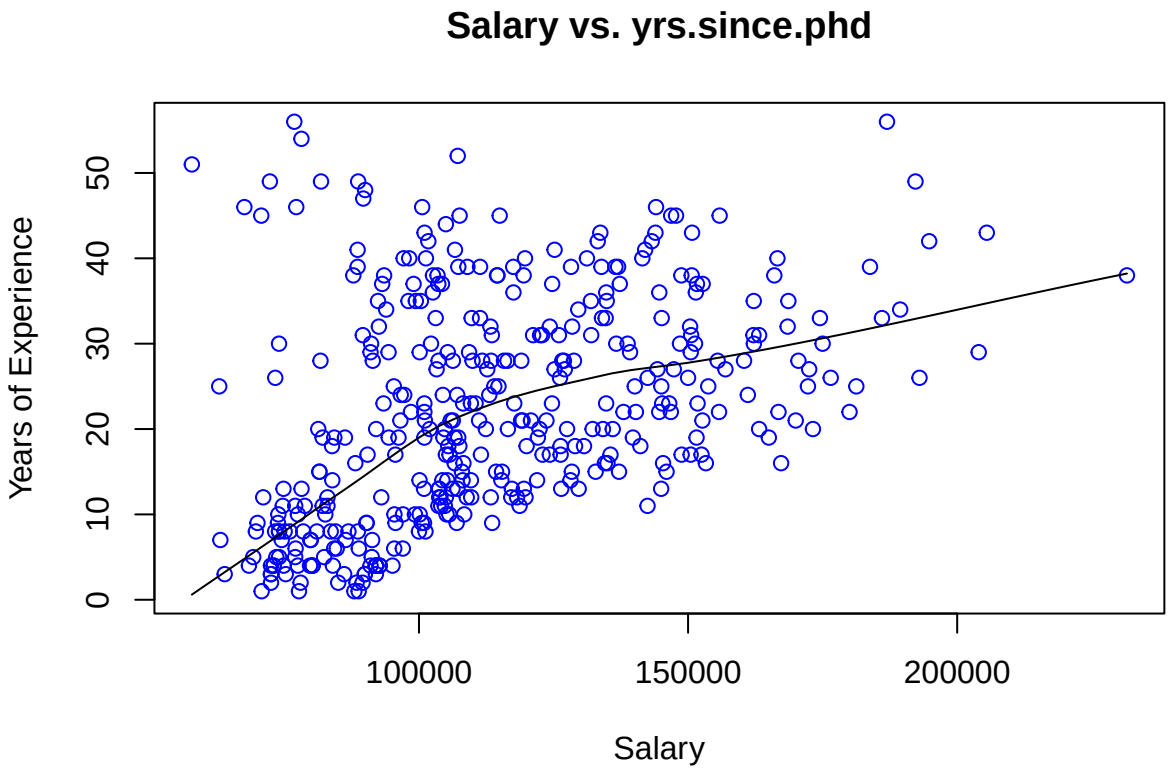
#**COMETARIO**: Los años de servicio continua teniendo una beta negativa, lo que supone que a mayor años de servicio, menor salario. Por tanto no podemos observar coherencia en esta variable. Para tratar de añalizar la variabilidad de los años de servicio, representaré un gráfico donde se relacione con el salario

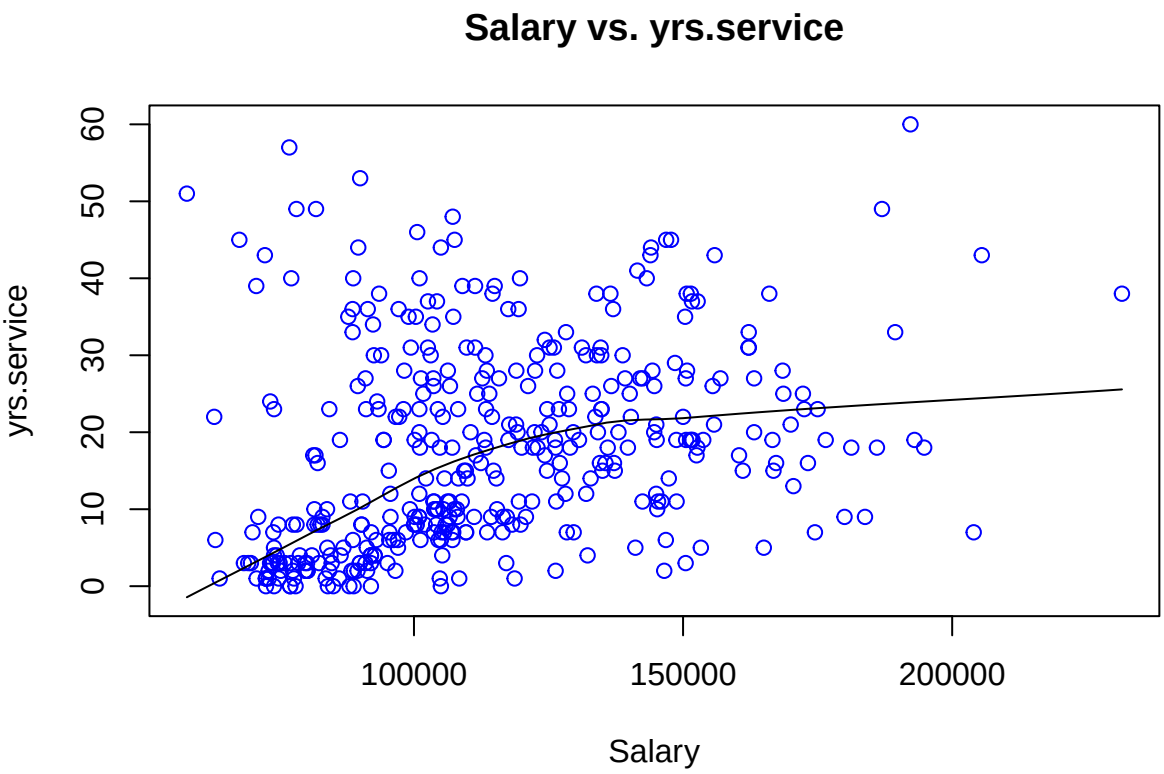


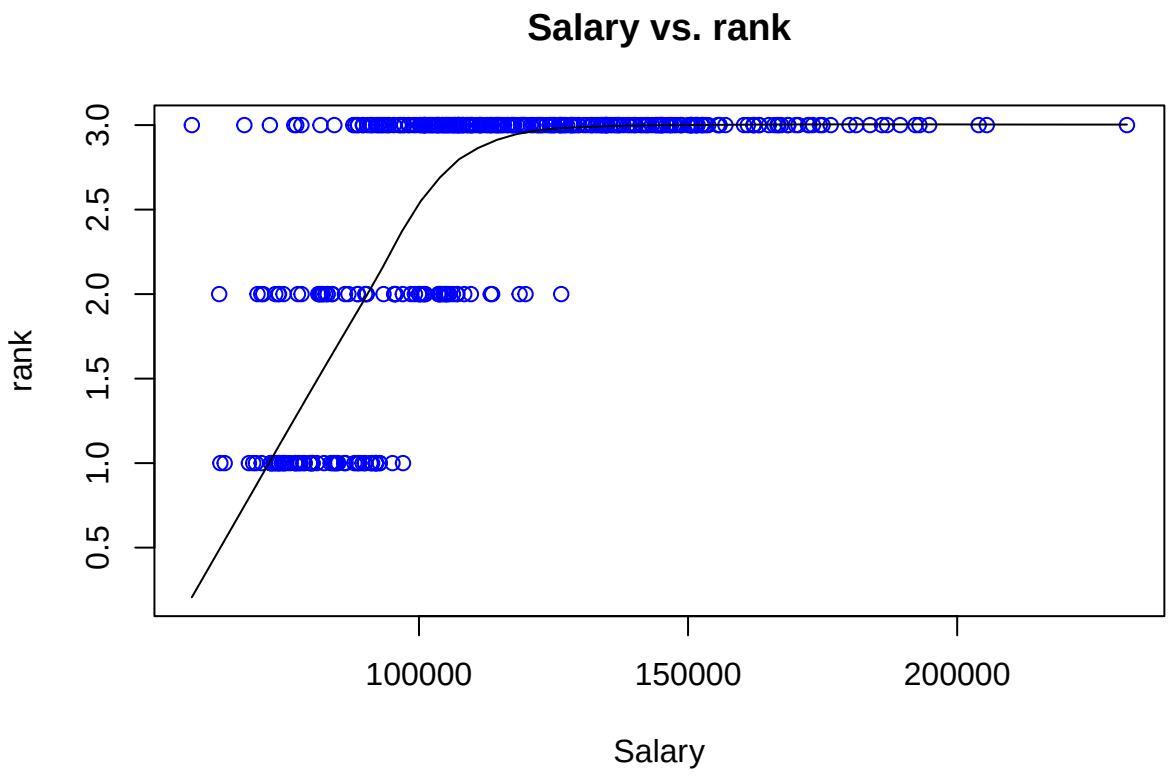
8. Utilizando un modelo de regresión lineal (lm), realiza una modelización correcta del salario (utilizando las variables que desees de la base de datos) y presenta los resultados argumentando, desde tu conocimiento, las razones por las que eliges dicho modelo.

#**COMETARIO**: En primer lugar, voy a representar gráficamente la relación entre las variables, para determinar cuales pueden aportar más al modelo de forma visual

#**JUSTIFICACIÓN**: De cara a seleccionar mi modelo, en base el análisis visual de los gráficos de dispersión, generaría mi modelo con las variables años desde obtención del doctorado, el rank y la disciplina, siendo el rank la variable que mejor explica el modelo. En cuando la variable salario, aplicaré el logaritmo neperiano, pues en el caso propuesto incremento el r2, mejorando la explicabildiad de las variables.




Salary vs. yrs.service
yrs.service
0
10
20
30
40
50
60
100000
150000
200000
Salary




```
##
## Call:
## lm(formula = formula5, data = bbdd_salarios_ajus)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68747 -0.11150 -0.00558  0.09355  0.57779
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.3059175  0.0263059 163.686  < 2e-16 ***
## yrs.since.phd -0.0003998  0.0010165  -0.393    0.694
## rankAssocProf  0.1546442  0.0334462   4.624 5.13e-06 ***
## rankProf       0.4589318  0.0340107  13.494  < 2e-16 ***
## disciplineB    0.1289090  0.0188240   6.848 2.90e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1819 on 392 degrees of freedom
## Multiple R-squared:  0.5161, Adjusted R-squared:  0.5111
## F-statistic: 104.5 on 4 and 392 DF,  p-value: < 2.2e-16

##
## ================================================
##                         Dependent variable:
```

```
##                       ---------------------------
##                          log(salary_thousands)
## -----------------------------------------------
## yrs.since.phd                  -0.0004
##                                (0.001)
##
## rankAssocProf                 0.155***
##                                (0.033)
##
## rankProf                      0.459***
##                                (0.034)
##
## disciplineB                   0.129***
##                                (0.019)
##
## Constant                      4.306***
##                                (0.026)
##
## -----------------------------------------------
## Observations                     397
## R2                              0.516
## Adjusted R2                     0.511
## Residual Std. Error       0.182 (df = 392)
## F Statistic           104.509*** (df = 4; 392)
## ===============================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```