Linear Regression Activity

Isaí Ambrocio

Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import seaborn as sns
from scipy import stats
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
sns.set_theme()

df = pd.read_csv("/content/ds_salaries.csv")
```

Exploratory Data Analysis (EDA)

```
df
```

```
     Unnamed: 0  work_year experience_level employment_type  \
0             0       2020               MI              FT
1             1       2020               SE              FT
2             2       2020               SE              FT
3             3       2020               MI              FT
4             4       2020               SE              FT
..          ...        ...              ...             ...
602         602       2022               SE              FT
603         603       2022               SE              FT
604         604       2022               SE              FT
605         605       2022               SE              FT
606         606       2022               MI              FT

                      job_title  salary salary_currency  salary_in_usd
\
0                 Data Scientist   70000             EUR          79833

1    Machine Learning Scientist  260000             USD         260000

2              Big Data Engineer   85000             GBP         109024

3           Product Data Analyst   20000             USD          20000

4     Machine Learning Engineer  150000             USD         150000
```

```
..                          ...     ...              ...             ...
602                Data Engineer  154000              USD          154000
603                Data Engineer  126000              USD          126000
604                 Data Analyst  129000              USD          129000
605                 Data Analyst  150000              USD          150000
606                 AI Scientist  200000              USD          200000


    employee_residence  remote_ratio company_location company_size
0                   DE             0               DE            L
1                   JP             0               JP            S
2                   GB            50               GB            M
3                   HN             0               HN            S
4                   US            50               US            L
..                 ...           ...              ...          ...
602                 US           100               US            M
603                 US           100               US            M
604                 US             0               US            M
605                 US           100               US            M
606                 IN           100               US            L

[607 rows x 12 columns]
```

We check that there is no null data.

```
df.isnull().sum()

Unnamed: 0            0
work_year             0
experience_level      0
employment_type       0
job_title             0
salary                0
salary_currency       0
salary_in_usd         0
employee_residence    0
remote_ratio          0
company_location      0
company_size          0
dtype: int64

df1 = df.drop("Unnamed: 0", axis=1)

df1
```

```
     work_year experience_level employment_type
job_title  \
0         2020               MI              FT              Data
Scientist
1         2020               SE              FT  Machine Learning
Scientist
2         2020               SE              FT           Big Data
Engineer
3         2020               MI              FT        Product Data
Analyst
4         2020               SE              FT   Machine Learning
Engineer
..         ...              ...             ...
...
602       2022               SE              FT               Data
Engineer
603       2022               SE              FT               Data
Engineer
604       2022               SE              FT                Data
Analyst
605       2022               SE              FT                Data
Analyst
606       2022               MI              FT                AI
Scientist

     salary salary_currency  salary_in_usd employee_residence
remote_ratio  \
0     70000             EUR          79833                  DE
0
1    260000             USD         260000                  JP
0
2     85000             GBP         109024                  GB
50
3     20000             USD          20000                  HN
0
4    150000             USD         150000                  US
50
..      ...             ...            ...                 ...
...
602  154000             USD         154000                  US
100
603  126000             USD         126000                  US
100
604  129000             USD         129000                  US
0
605  150000             USD         150000                  US
100
606  200000             USD         200000                  IN
100
```

```
    company_location company_size
0                 DE            L
1                 JP            S
2                 GB            M
3                 HN            S
4                 US            L
..               ...          ...
602               US            M
603               US            M
604               US            M
605               US            M
606               US            L

[607 rows x 11 columns]
```

```
df1["experience_level"].unique()
```

```
array(['MI', 'SE', 'EN', 'EX'], dtype=object)
```

```
dummies=pd.get_dummies(df1["experience_level"],prefix="experience_leve
l")
```

Now, we use dummy.

```
dummies
```

```
     experience_level_EN  experience_level_EX  experience_level_MI  \
0                      0                    0                    1
1                      0                    0                    0
2                      0                    0                    0
3                      0                    0                    1
4                      0                    0                    0
..                   ...                  ...                  ...
602                    0                    0                    0
603                    0                    0                    0
604                    0                    0                    0
605                    0                    0                    0
606                    0                    0                    1

     experience_level_SE
0                      0
1                      1
2                      1
3                      0
4                      1
..                   ...
602                    1
603                    1
604                    1
```

```
605                    1
606                    0

[607 rows x 4 columns]
```

```
df1 = pd.concat([df1,dummies],axis=1)
```

```
df1
```

```
     work_year experience_level employment_type
job_title  \
0         2020               MI              FT              Data
Scientist
1         2020               SE              FT  Machine Learning
Scientist
2         2020               SE              FT           Big Data
Engineer
3         2020               MI              FT        Product Data
Analyst
4         2020               SE              FT   Machine Learning
Engineer
..         ...              ...             ...
...
602       2022               SE              FT               Data
Engineer
603       2022               SE              FT               Data
Engineer
604       2022               SE              FT                Data
Analyst
605       2022               SE              FT                Data
Analyst
606       2022               MI              FT                AI
Scientist

     salary salary_currency  salary_in_usd employee_residence
remote_ratio  \
0     70000             EUR          79833                 DE
0
1    260000             USD         260000                 JP
0
2     85000             GBP         109024                 GB
50
3     20000             USD          20000                 HN
0
4    150000             USD         150000                 US
50
..      ...             ...            ...                ...
...
602  154000             USD         154000                 US
100
```

```
603  126000             USD         126000                  US
100
604  129000             USD         129000                  US
0
605  150000             USD         150000                  US
100
606  200000             USD         200000                  IN
100

    company_location company_size  experience_level_EN
experience_level_EX  \
0                  DE            L                     0
0
1                  JP            S                     0
0
2                  GB            M                     0
0
3                  HN            S                     0
0
4                  US            L                     0
0
..                ...          ...                   ...
...
602                US            M                     0
0
603                US            M                     0
0
604                US            M                     0
0
605                US            M                     0
0
606                US            L                     0
0

     experience_level_MI  experience_level_SE
0                      1                    0
1                      0                    1
2                      0                    1
3                      1                    0
4                      0                    1
..                   ...                  ...
602                    0                    1
603                    0                    1
604                    0                    1
605                    0                    1
606                    1                    0

[607 rows x 15 columns]
```

```
df1["employment_type"].unique()
```

```
array(['FT', 'CT', 'PT', 'FL'], dtype=object)
```

```
dummies=pd.get_dummies(df1["employment_type"],prefix="employment_type"
)

df1 = pd.concat([df1,dummies],axis=1)

df1
```

```
     work_year experience_level employment_type
job_title  \
0         2020               MI              FT               Data
Scientist
1         2020               SE              FT  Machine Learning
Scientist
2         2020               SE              FT            Big Data
Engineer
3         2020               MI              FT        Product Data
Analyst
4         2020               SE              FT   Machine Learning
Engineer
..         ...              ...             ...
...
602       2022               SE              FT               Data
Engineer
603       2022               SE              FT               Data
Engineer
604       2022               SE              FT                Data
Analyst
605       2022               SE              FT                Data
Analyst
606       2022               MI              FT                AI
Scientist

     salary salary_currency  salary_in_usd employee_residence
remote_ratio  \
0     70000             EUR          79833                 DE
0
1    260000             USD         260000                 JP
0
2     85000             GBP         109024                 GB
50
3     20000             USD          20000                 HN
0
4    150000             USD         150000                 US
50
..      ...             ...            ...                ...
...
602  154000             USD         154000                 US
100
```

```
603  126000             USD         126000                  US
100
604  129000             USD         129000                  US
0
605  150000             USD         150000                  US
100
606  200000             USD         200000                  IN
100

    company_location company_size  experience_level_EN
experience_level_EX  \
0                  DE            L                     0
0
1                  JP            S                     0
0
2                  GB            M                     0
0
3                  HN            S                     0
0
4                  US            L                     0
0
..                ...          ...                   ...
...
602                US            M                     0
0
603                US            M                     0
0
604                US            M                     0
0
605                US            M                     0
0
606                US            L                     0
0

     experience_level_MI  experience_level_SE  employment_type_CT  \
0                      1                    0                   0
1                      0                    1                   0
2                      0                    1                   0
3                      1                    0                   0
4                      0                    1                   0
..                   ...                  ...                 ...
602                    0                    1                   0
603                    0                    1                   0
604                    0                    1                   0
605                    0                    1                   0
606                    1                    0                   0

     employment_type_FL  employment_type_FT  employment_type_PT
0                     0                   1                   0
1                     0                   1                   0
```

```
2                     0                     1                     0
3                     0                     1                     0
4                     0                     1                     0
..                  ...                   ...                   ...
602                   0                     1                     0
603                   0                     1                     0
604                   0                     1                     0
605                   0                     1                     0
606                   0                     1                     0

[607 rows x 19 columns]
```

```
df1.columns
```

```
Index(['work_year', 'experience_level', 'employment_type',
'job_title',
       'salary', 'salary_currency', 'salary_in_usd',
'employee_residence',
       'remote_ratio', 'company_location', 'company_size',
       'experience_level_EN', 'experience_level_EX',
'experience_level_MI',
       'experience_level_SE', 'employment_type_CT',
'employment_type_FL',
       'employment_type_FT', 'employment_type_PT'],
      dtype='object')
```

```
correlacion=df1.corr()
```

```
<ipython-input-50-b5adb217c2e4>:1: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it
will default to False. Select only valid columns or specify the value
of numeric_only to silence this warning.
  correlacion=df1.corr()
```

```
correlacion
```

```
                     work_year    salary  salary_in_usd  remote_ratio
\
work_year             1.000000 -0.087577       0.170493      0.076314

salary               -0.087577  1.000000      -0.083906     -0.014608

salary_in_usd         0.170493 -0.083906       1.000000      0.132122

remote_ratio          0.076314 -0.014608       0.132122      1.000000

experience_level_EN  -0.234542 -0.015845      -0.294196     -0.010490

experience_level_EX   0.005446  0.014130       0.259866      0.041208

experience_level_MI  -0.136382  0.074626      -0.252024     -0.127850
```

```
experience_level_SE   0.294008 -0.065995       0.343513      0.113071

employment_type_CT   -0.053407 -0.008268       0.092907      0.065149

employment_type_FL   -0.047729 -0.014568      -0.073863     -0.016865

employment_type_FT    0.105342  0.025685       0.091819     -0.023834

employment_type_PT   -0.075845 -0.020006      -0.144627     -0.002935


                     experience_level_EN  experience_level_EX  \
work_year                      -0.234542             0.005446
salary                         -0.015845             0.014130
salary_in_usd                  -0.294196             0.259866
remote_ratio                   -0.010490             0.041208
experience_level_EN             1.000000            -0.087108
experience_level_EX            -0.087108             1.000000
experience_level_MI            -0.302761            -0.155539
experience_level_SE            -0.381033            -0.195751
employment_type_CT              0.066013             0.070739
employment_type_FL             -0.033537            -0.017229
employment_type_FT             -0.167828            -0.008698
employment_type_PT              0.204028            -0.027379

                     experience_level_MI  experience_level_SE  \
work_year                      -0.136382             0.294008
salary                          0.074626            -0.065995
salary_in_usd                  -0.252024             0.343513
remote_ratio                   -0.127850             0.113071
experience_level_EN            -0.302761            -0.381033
experience_level_EX            -0.155539            -0.195751
experience_level_MI             1.000000            -0.680373
experience_level_SE            -0.680373             1.000000
employment_type_CT             -0.028817            -0.047768
employment_type_FL              0.068108            -0.034520
employment_type_FT             -0.006597             0.128381
employment_type_PT             -0.013805            -0.119762

                     employment_type_CT  employment_type_FL  \
work_year                     -0.053407           -0.047729
salary                        -0.008268           -0.014568
salary_in_usd                  0.092907           -0.073863
remote_ratio                   0.065149           -0.016865
experience_level_EN            0.066013           -0.033537
experience_level_EX            0.070739           -0.017229
experience_level_MI           -0.028817            0.068108
experience_level_SE           -0.047768           -0.034520
employment_type_CT             1.000000           -0.007423
```

```
employment_type_FL            -0.007423            1.000000
employment_type_FT            -0.506989           -0.453089
employment_type_PT            -0.011795           -0.010541

                     employment_type_FT  employment_type_PT
work_year                      0.105342           -0.075845
salary                         0.025685           -0.020006
salary_in_usd                  0.091819           -0.144627
remote_ratio                  -0.023834           -0.002935
experience_level_EN           -0.167828            0.204028
experience_level_EX           -0.008698           -0.027379
experience_level_MI           -0.006597           -0.013805
experience_level_SE            0.128381           -0.119762
employment_type_CT            -0.506989           -0.011795
employment_type_FL            -0.453089           -0.010541
employment_type_FT             1.000000           -0.719987
employment_type_PT            -0.719987            1.000000
```

```
alta_corr = np.where((correlacion > 0.95) & (correlacion < 1))
```

```
baja_corr = np.where((correlacion < -0.95) & (correlacion > -1))
```

```
correlacion
```

```
                     work_year    salary  salary_in_usd  remote_ratio
\
work_year             1.000000 -0.087577       0.170493      0.076314

salary               -0.087577  1.000000      -0.083906     -0.014608

salary_in_usd         0.170493 -0.083906       1.000000      0.132122

remote_ratio          0.076314 -0.014608       0.132122      1.000000

experience_level_EN  -0.234542 -0.015845      -0.294196     -0.010490

experience_level_EX   0.005446  0.014130       0.259866      0.041208

experience_level_MI  -0.136382  0.074626      -0.252024     -0.127850

experience_level_SE   0.294008 -0.065995       0.343513      0.113071

employment_type_CT   -0.053407 -0.008268       0.092907      0.065149

employment_type_FL   -0.047729 -0.014568      -0.073863     -0.016865

employment_type_FT    0.105342  0.025685       0.091819     -0.023834

employment_type_PT   -0.075845 -0.020006      -0.144627     -0.002935
```

	experience_level_EN	experience_level_EX	\
work_year	-0.234542	0.005446	
salary	-0.015845	0.014130	
salary_in_usd	-0.294196	0.259866	
remote_ratio	-0.010490	0.041208	
experience_level_EN	1.000000	-0.087108	
experience_level_EX	-0.087108	1.000000	
experience_level_MI	-0.302761	-0.155539	
experience_level_SE	-0.381033	-0.195751	
employment_type_CT	0.066013	0.070739	
employment_type_FL	-0.033537	-0.017229	
employment_type_FT	-0.167828	-0.008698	
employment_type_PT	0.204028	-0.027379	

	experience_level_MI	experience_level_SE	\
work_year	-0.136382	0.294008	
salary	0.074626	-0.065995	
salary_in_usd	-0.252024	0.343513	
remote_ratio	-0.127850	0.113071	
experience_level_EN	-0.302761	-0.381033	
experience_level_EX	-0.155539	-0.195751	
experience_level_MI	1.000000	-0.680373	
experience_level_SE	-0.680373	1.000000	
employment_type_CT	-0.028817	-0.047768	
employment_type_FL	0.068108	-0.034520	
employment_type_FT	-0.006597	0.128381	
employment_type_PT	-0.013805	-0.119762	

	employment_type_CT	employment_type_FL	\
work_year	-0.053407	-0.047729	
salary	-0.008268	-0.014568	
salary_in_usd	0.092907	-0.073863	
remote_ratio	0.065149	-0.016865	
experience_level_EN	0.066013	-0.033537	
experience_level_EX	0.070739	-0.017229	
experience_level_MI	-0.028817	0.068108	
experience_level_SE	-0.047768	-0.034520	
employment_type_CT	1.000000	-0.007423	
employment_type_FL	-0.007423	1.000000	
employment_type_FT	-0.506989	-0.453089	
employment_type_PT	-0.011795	-0.010541	

	employment_type_FT	employment_type_PT
work_year	0.105342	-0.075845
salary	0.025685	-0.020006
salary_in_usd	0.091819	-0.144627
remote_ratio	-0.023834	-0.002935
experience_level_EN	-0.167828	0.204028
experience_level_EX	-0.008698	-0.027379
experience_level_MI	-0.006597	-0.013805

```
experience_level_SE            0.128381           -0.119762
employment_type_CT            -0.506989           -0.011795
employment_type_FL            -0.453089           -0.010541
employment_type_FT             1.000000           -0.719987
employment_type_PT            -0.719987            1.000000
```

```
alta_corr
```

```
(array([], dtype=int64), array([], dtype=int64))
```

```
baja_corr
```

```
(array([], dtype=int64), array([], dtype=int64))
```

```
entrenamiento,
prueba=train_test_split(df1,test_size=0.20,random_state=42)
```

```
entrenamiento
```

```
     work_year experience_level employment_type  \
9         2020               SE              FT
227       2021               MI              FT
591       2022               SE              FT
516       2022               SE              FT
132       2021               MI              FT
..         ...              ...             ...
71        2020               MI              FT
106       2021               MI              FT
270       2021               EN              FT
435       2022               MI              FT
102       2021               MI              FT

                              job_title    salary salary_currency  \
9                   Lead Data Engineer    125000             USD
227                     Data Scientist     75000             EUR
591                     Data Architect    144854             USD
516                Data Science Manager    152500             USD
132  Applied Machine Learning Scientist     38400             USD
..                                  ...       ...             ...
71                      Data Scientist     37000             EUR
106                 Research Scientist    235000             CAD
270                      Data Engineer     72500             USD
435                      Data Engineer     70000             GBP
102                   BI Data Analyst  11000000             HUF

     salary_in_usd employee_residence  remote_ratio
company_location  \
9           125000                 NZ            50                NZ

227          88654                 DE            50                DE
```

```
591         144854                  US           100                US
516         152500                  US           100                US
132          38400                  VN           100                US
..             ...                 ...           ...               ...
71           42197                  FR            50                FR
106         187442                  CA           100                CA
270          72500                  US           100                US
435          91614                  GB           100                GB
102          36259                  HU            50                US


    company_size  experience_level_EN  experience_level_EX  \
9              S                    0                    0
227            L                    0                    0
591            M                    0                    0
516            M                    0                    0
132            M                    0                    0
..           ...                  ...                  ...
71             S                    0                    0
106            L                    0                    0
270            L                    1                    0
435            M                    0                    0
102            L                    0                    0

     experience_level_MI  experience_level_SE  employment_type_CT  \
9                      0                    1                   0
227                    1                    0                   0
591                    0                    1                   0
516                    0                    1                   0
132                    1                    0                   0
..                   ...                  ...                 ...
71                     1                    0                   0
106                    1                    0                   0
270                    0                    0                   0
435                    1                    0                   0
102                    1                    0                   0

     employment_type_FL  employment_type_FT  employment_type_PT
9                     0                   1                   0
227                   0                   1                   0
591                   0                   1                   0
516                   0                   1                   0
```

```
132                  0                   1                   0
..                 ...                 ...                 ...
71                   0                   1                   0
106                  0                   1                   0
270                  0                   1                   0
435                  0                   1                   0
102                  0                   1                   0

[485 rows x 19 columns]
```

Model

$$Y = \beta X + \epsilon$$

$$\hat{Y} = \beta^{i} X$$

$$\beta X = \beta_0(1) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

```
modelo=smf.ols(formula="salary_in_usd~salary+remote_ratio+experience_level_EN+experience_level_EX+experience_level_MI+employment_type_CT+employment_type_FL+employment_type_FT",data=entrenamiento)
modelo=modelo.fit()
print(modelo.summary())
```

```
                            OLS Regression Results

==============================================================================
Dep. Variable:          salary_in_usd   R-squared:
0.264
Model:                            OLS   Adj. R-squared:
0.252
Method:                 Least Squares   F-statistic:
21.37
Date:                Fri, 18 Aug 2023   Prob (F-statistic):
8.41e-28
Time:                        04:48:29   Log-Likelihood:
-6044.0
No. Observations:                 485   AIC:
1.211e+04
Df Residuals:                     476   BIC:
1.214e+04
Df Model:                           8

Covariance Type:            nonrobust

==============================================================================
                 coef    std err          t      P>|t|
```

```
[0.025      0.975]
------------------------------------------------------------------------------
-----------------
Intercept             9.787e+04   2.44e+04      4.015      0.000
5e+04    1.46e+05
salary                  -0.0067      0.003     -2.251      0.025      -
0.013      -0.001
remote_ratio           103.1885     70.546      1.463      0.144     -
35.431     241.809
experience_level_EN -7.485e+04   8886.547     -8.423      0.000   -
9.23e+04   -5.74e+04
experience_level_EX  6.662e+04   1.42e+04      4.698      0.000
3.88e+04    9.45e+04
experience_level_MI  -4.79e+04   6485.002     -7.387      0.000   -
6.06e+04   -3.52e+04
employment_type_CT   7.688e+04   3.89e+04      1.974      0.049
367.309    1.53e+05
employment_type_FL  -3707.2560   5.04e+04     -0.073      0.941   -
1.03e+05    9.54e+04
employment_type_FT   3.578e+04   2.34e+04      1.531      0.126   -
1.01e+04    8.17e+04
==============================================================================
========
Omnibus:                      242.000   Durbin-Watson:
1.979
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
2005.484
Skew:                           2.000   Prob(JB):
0.00
Kurtosis:                      12.123   Cond. No.
2.01e+07
==============================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 2.01e+07. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

Enhacing Model

Fitting a new linear regression model without `remote_ratio` to see if we get better performance.

```
df2 = df1.drop(["remote_ratio"], axis = 1)
```

```
modelo1=smf.ols(formula="salary_in_usd~salary+experience_level_EN+expe
rience_level_EX+experience_level_MI+employment_type_CT+employment_type
_FL+employment_type_FT",data=entrenamiento)
modelo1=modelo1.fit()
print(modelo1.summary())
```

```
                            OLS Regression Results

==============================================================================
========
Dep. Variable:          salary_in_usd   R-squared:
0.261
Model:                            OLS   Adj. R-squared:
0.250
Method:                 Least Squares   F-statistic:
24.06
Date:                Fri, 18 Aug 2023   Prob (F-statistic):
4.61e-28
Time:                        03:01:01   Log-Likelihood:
-6045.1
No. Observations:                 485   AIC:
1.211e+04
Df Residuals:                     477   BIC:
1.214e+04
Df Model:                           7
Covariance Type:            nonrobust

==============================================================================
=================
                             coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
-----------------
Intercept            1.061e+05   2.37e+04      4.469      0.000
5.95e+04    1.53e+05
salary                 -0.0069      0.003     -2.341      0.020      -
0.013      -0.001
experience_level_EN -7.531e+04   8891.646     -8.470      0.000   -
9.28e+04   -5.78e+04
experience_level_EX   6.72e+04   1.42e+04      4.735      0.000
3.93e+04    9.51e+04
experience_level_MI -4.862e+04   6473.969     -7.511      0.000   -
6.13e+04   -3.59e+04
employment_type_CT   7.909e+04    3.9e+04      2.030      0.043
2536.143    1.56e+05
employment_type_FL  -1249.6913   5.05e+04     -0.025      0.980      -
1e+05    9.79e+04
employment_type_FT   3.516e+04   2.34e+04      1.503      0.133   -
1.08e+04    8.11e+04
```

```
==============================================================================
========
Omnibus:                      240.438   Durbin-Watson:
1.986
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
1967.629
Skew:                           1.989   Prob(JB):
0.00
Kurtosis:                      12.031   Cond. No.
2.01e+07
==============================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 2.01e+07. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

Removing Salary

Fitting a new linear regression model without `salary` to see if we get better performance.

```
df3 = df1.drop(["salary"], axis = 1)

modelo2=smf.ols(formula="salary_in_usd~remote_ratio+experience_level_E
N+experience_level_EX+experience_level_MI+employment_type_CT+employmen
t_type_FL+employment_type_FT",data=entrenamiento)
modelo2=modelo2.fit()
print(modelo2.summary())
```

```
                            OLS Regression Results

==============================================================================
========
Dep. Variable:          salary_in_usd   R-squared:
0.256
Model:                            OLS   Adj. R-squared:
0.246
Method:                 Least Squares   F-statistic:
23.50
Date:                Fri, 18 Aug 2023   Prob (F-statistic):
1.89e-27
Time:                        03:01:13   Log-Likelihood:
-6046.6
No. Observations:                 485   AIC:
1.211e+04
Df Residuals:                     477   BIC:
```

```
1.214e+04
Df Model:                           7

Covariance Type:            nonrobust

==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      9.718e+04   2.45e+04      3.970      0.000    4.91e+04    1.45e+05
remote_ratio    112.7886     70.716      1.595      0.111     -26.165     251.743
experience_level_EN -7.517e+04   8923.178     -8.425      0.000   -9.27e+04   -5.76e+04
experience_level_EX   6.49e+04   1.42e+04      4.564      0.000    3.7e+04    9.28e+04
experience_level_MI -4.854e+04   6506.373     -7.460      0.000   -6.13e+04   -3.58e+04
employment_type_CT   7.612e+04   3.91e+04      1.947      0.052     -714.894    1.53e+05
employment_type_FL  -4186.9851   5.07e+04     -0.083      0.934   -1.04e+05    9.54e+04
employment_type_FT   3.419e+04   2.34e+04      1.458      0.145   -1.19e+04    8.03e+04
==============================================================================
Omnibus:                      235.750   Durbin-Watson:                   1.984
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1878.276
Skew:                           1.949   Prob(JB):                         0.00
Kurtosis:                      11.817   Cond. No.                     1.63e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.63e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Removing Remote Ratio & Employment Type

Fitting a new linear regression model without `remote_ratio` and `employment_type` to see if we get better performance.

```
df4 = df1.drop(["employment_type_CT", "employment_type_FL",
"employment_type_FT","remote_ratio"], axis = 1)

modelo3=smf.ols(formula="salary_in_usd~salary+experience_level_EN+expe
rience_level_EX+experience_level_MI",data=entrenamiento)
modelo3=modelo3.fit()
print(modelo3.summary())
```

```
                            OLS Regression Results

==============================================================================
========
Dep. Variable:          salary_in_usd   R-squared:
0.253
Model:                            OLS   Adj. R-squared:
0.247
Method:                 Least Squares   F-statistic:
40.70
Date:                Fri, 18 Aug 2023   Prob (F-statistic):
2.24e-29
Time:                        05:16:42   Log-Likelihood:
-6047.6
No. Observations:                 485   AIC:
1.211e+04
Df Residuals:                     480   BIC:
1.213e+04
Df Model:                           4

Covariance Type:            nonrobust

==============================================================================
=================
                            coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
-----------------
Intercept             1.413e+05   4276.518     33.037      0.000
1.33e+05     1.5e+05
salary                  -0.0068      0.003     -2.299      0.022      -
0.013      -0.001
experience_level_EN -7.754e+04   8570.685     -9.048      0.000   -
9.44e+04   -6.07e+04
experience_level_EX  6.913e+04   1.42e+04      4.883      0.000
4.13e+04    9.69e+04
experience_level_MI  -4.91e+04   6483.621     -7.573      0.000   -
```

```
6.18e+04   -3.64e+04
==============================================================================
========
Omnibus:                      239.858   Durbin-Watson:
1.990
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
1925.137
Skew:                           1.990   Prob(JB):
0.00
Kurtosis:                      11.912   Cond. No.
5.07e+06
==============================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 5.07e+06. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

Selected Model

Based on the previous models, we consider that the 4th model explains the best the dependent variable. We can clearly see that it does **not** out-performs the rest of the models, the coefficients make the most sense of them all.

Leaving us with the columns:

- `salary`
- `experience_level_EN`
- `experience_level_EX`
- `experience_level_MI`

And our model:

$$\hat{Y} = \beta_0 + \beta_1(\text{salary}) + \beta_2(\text{experience\_level\_EN}) + \beta_3(\text{experience\_level\_EX}) + \beta_4(\text{experience\_level\_MI})$$

$$\hat{Y} = 141300.0 - 0.0068(\text{salary}) - 77540.0(\text{experience\_level\_EN}) + 69130.0(\text{experience\_level\_EX}) - 49100.0(\text{exp}$$

```
modelo3.params

Intercept              141282.300520
salary                     -0.006834
experience_level_EN    -77543.574397
experience_level_EX     69131.013881
experience_level_MI    -49099.920182
dtype: float64
```

```
Intercept              141282.300520
salary                     -0.006834
experience_level_EN    -77543.574397
experience_level_EX     69131.013881
experience_level_MI    -49099.920182
dtype: float64
```

```
y_aprox = 141282.300520 - 0.006834 * prueba["salary"] - 77543.574397 *
prueba["experience_level_EN"] + \
            69131.013881 * prueba["experience_level_EX"] -
49099.920182 * prueba["experience_level_MI"]

print(y_aprox)
```

```
563    140323.832020
289    140359.710520
76      91498.980338
78      90337.200338
182     92032.032338
           ...
249    140120.520520
365    140335.108120
453     91362.300338
548    140605.392820
235     91430.640338
Length: 122, dtype: float64
```

```
modelo3.predict(prueba.loc[:, ["salary", "experience_level_EN",
"experience_level_EX", "experience_level_MI"]])
```

```
563    140323.877371
289    140359.754174
76      91499.012674
78      90337.287645
182     92032.039451
           ...
249    140120.575491
365    140335.152938
453     91362.339141
548    140605.424849
235     91430.675907
Length: 122, dtype: float64
```

```
table =
pd.DataFrame({"Real":prueba["salary_in_usd"],"Prediccion":y_aprox,"Err
ores":prueba["salary_in_usd"]-y_aprox})

table
```

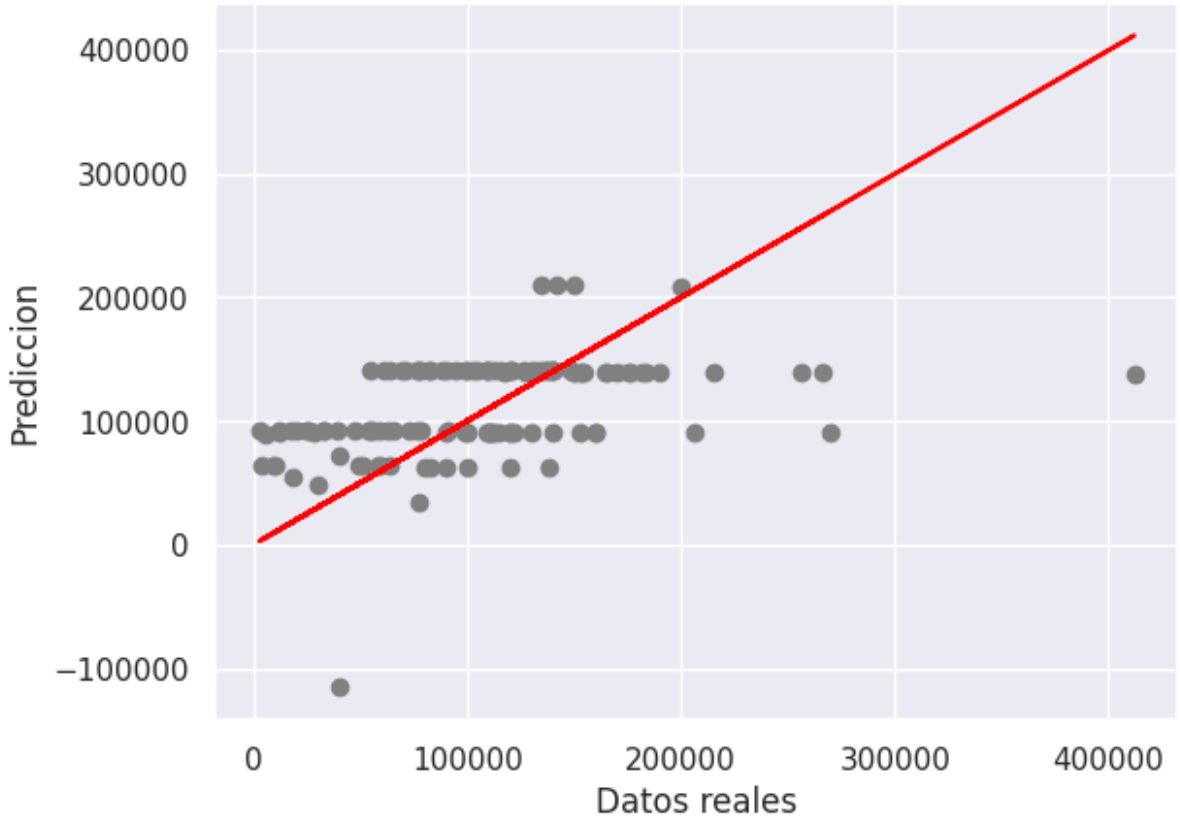
```
       Real     Prediccion        Errores
563  140250  140323.832020     -73.832020
```

```
289  135000  140359.710520   -5359.710520
76   100000   91498.980338    8501.019662
78   270000   90337.200338  179662.799662
182   26005   92032.032338  -66027.032338
..      ...            ...            ...
249  170000  140120.520520   29879.479480
365  138600  140335.108120   -1735.108120
453  120000   91362.300338   28637.699662
548   99050  140605.392820  -41555.392820
235  110000   91430.640338   18569.359662

[122 rows x 3 columns]
```

```
plt.scatter(prueba["salary_in_usd"],y_aprox,color="gray")
plt.plot(prueba['salary_in_usd'],prueba['salary_in_usd'],color='red')
plt.xlabel("Datos reales")
plt.ylabel("Prediccion")
```

```
Text(0, 0.5, 'Prediccion')
```

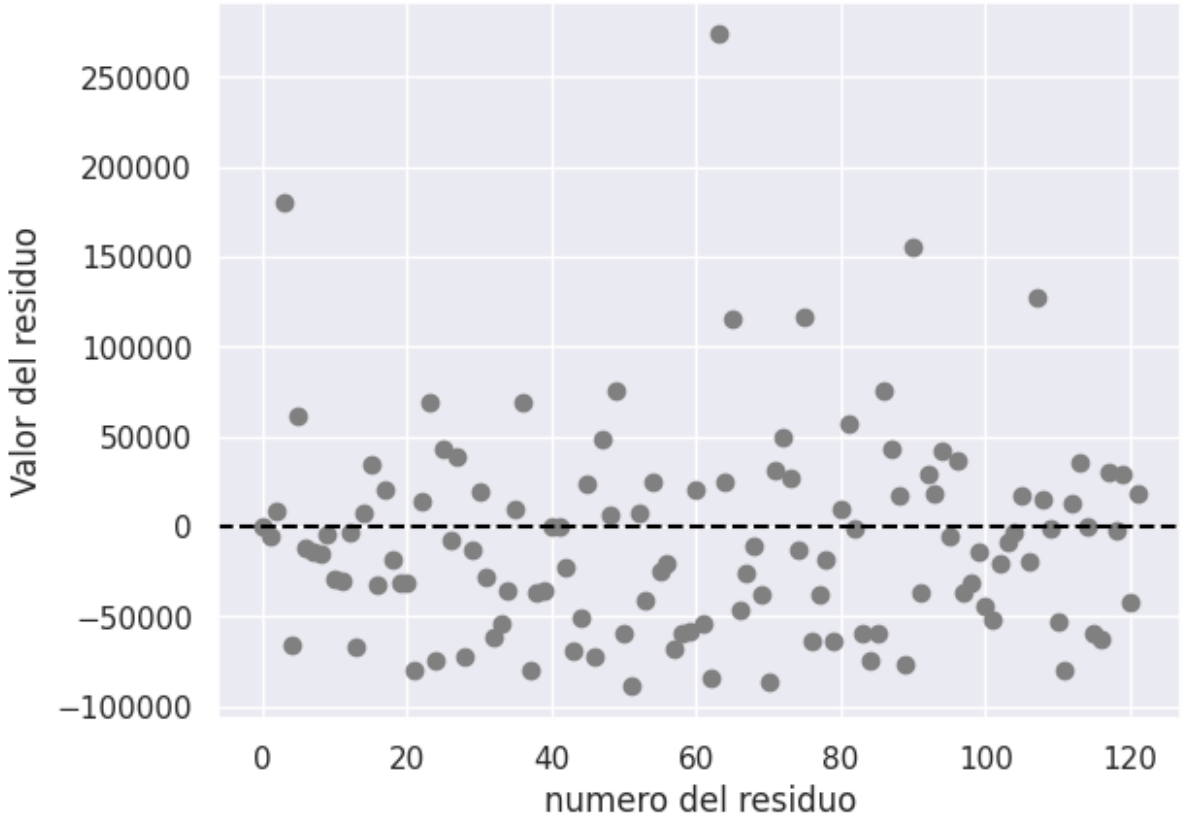



```
l_residuos=len(table["Errores"])

plt.scatter(range(l_residuos),table["Errores"],color="gray")
plt.axhline(y=0,linestyle="--",color="black")
```

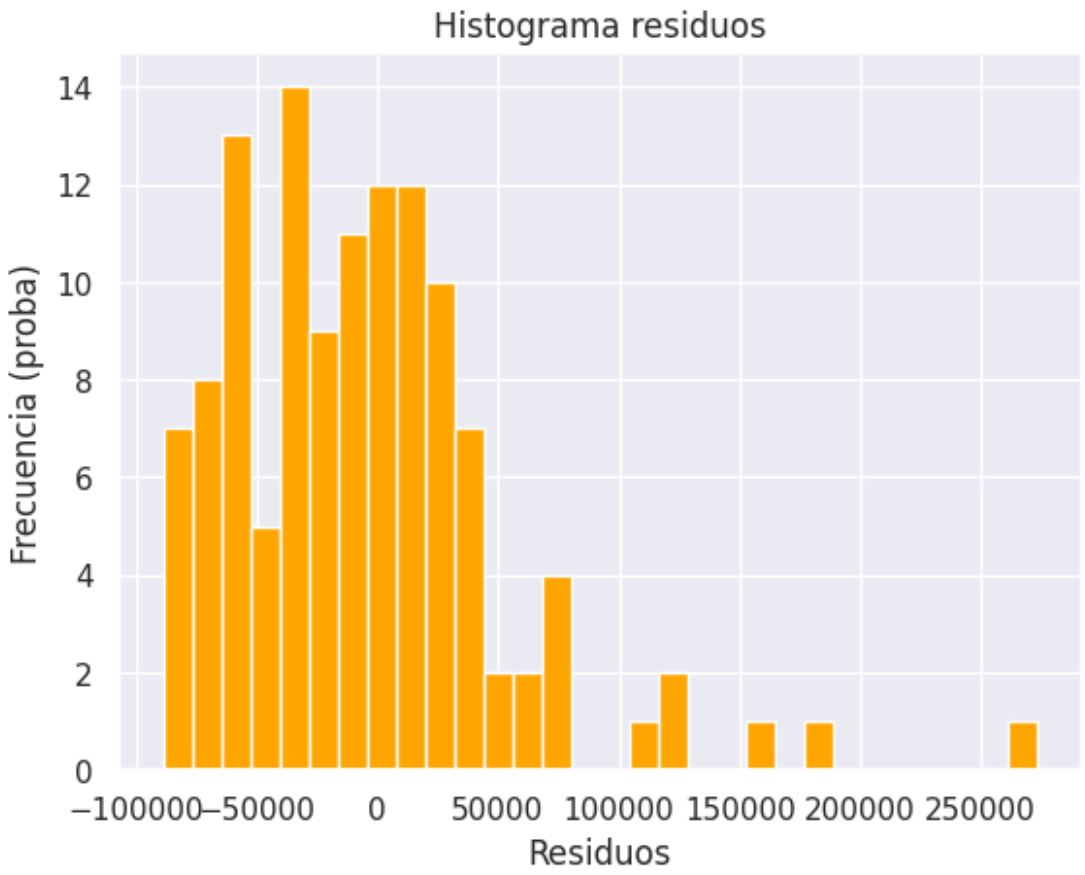
```
plt.xlabel("numero del residuo")
plt.ylabel("Valor del residuo")
```

```
Text(0, 0.5, 'Valor del residuo')
```




```
plt.hist(x=table["Errores"],color="orange", bins=30)
plt.title("Histograma residuos")
plt.xlabel("Residuos")
plt.ylabel("Frecuencia (proba)")
```

```
Text(0, 0.5, 'Frecuencia (proba)')
```




```
media=table["Errores"].mean()
std=table["Errores"].std()
Errores_est=(table["Errores"]-media)/std

stats.kstest(Errores_est,"norm")

KstestResult(statistic=0.08390914139054195,
pvalue=0.33786192266394016, statistic_location=0.11868717093119897,
statistic_sign=1)
```

The fourth model was selected for simplicity rather than perfomance (Principle of parsimony).

We have a base salary of our implicit rank (SE). From there, we see that someone with an entry rank (EN) is deducted a higher amount compared to the MID, while the Expert is given an increase. In the case of salary the coefficient is so small that it is not relevant.

Isai Ambrocio - A01625101