Correlation and Regression

Esther Yang

CS51 Spring 2020

Introduction

This article assesses how ages different in IBM are related to a different monthly salary. The sample dataset (n = 1470) is from IBM HR Analytics Employee Attrition & Performance from Kaggle. We also consider the influence of different factors on monthly income, such as job level and working years at the company. Since different firms have different settings, cultures, and characteristics, the population we want to predict will only be a firm-level analysis, limited explicitly to IBM. This report will provide a model to estimate monthly income and including the validity of the model.

Dataset

The dataset includes 35 factors related to job performance and features of employees such as department, performance rating, years in the current role. Lack of country information, our prediction for IBM employees, can be biased in this sample dataset. IBM employees in different countries might vary in salaries and expectations of performance due to different cultural, social, and economic structures. Therefore, we might need to restrict the implication only in the U.S. [1]

Our response variable is the monthly salary, which ranges from 1009 USD to 19999 USD. With a median of 4919 USD and an average of 6503 USD, we can see our distribution skew to the

[1] **#induction:** The 4 assumptions(the last paragraph) acknowledge that the model is not totally reliable. In this part, we focus on the limitation of sample to stronger the strength to the specific population.

right in our distribution, indicating a small number of employees in IBM receive a higher salary than most other employees.

Our main predict variable is age. It is a discrete variable, but we can see it normally distributed from 18 to 60, with a majority of employees in the thirties. The percentage of different age groups is about 1% in their tenish, 21% in their twenties, 42% in their thirties, 24% in their forties, and 12% in their forties. [2]

	mean	median	standard deviation	percentage
<20	1813.06	1675	548.33	1.16%
20-30	3795.74	3102	2034.55	21.02%
30-40	5599.25	4983	3128.11	42.31%
40-50	8537.96	6377	5621.88	23.74%
50-60	10942.91	10725	6005.30	11.77%

Table 1 : Summary statistics of different ages(Appendix A)

[2] **#variables:** I describe the dependent and independent variables and provide their summary statistics.

To investigate the relationship between age and salaries, we group employees of different ages and calculate their average salary. First, we observe their relationship with a scatterplot in Figure 1. We can see a different pattern of income before and after 40-year-old. After 40-year-old, the top income hit the ceiling of 20000 USD per month.

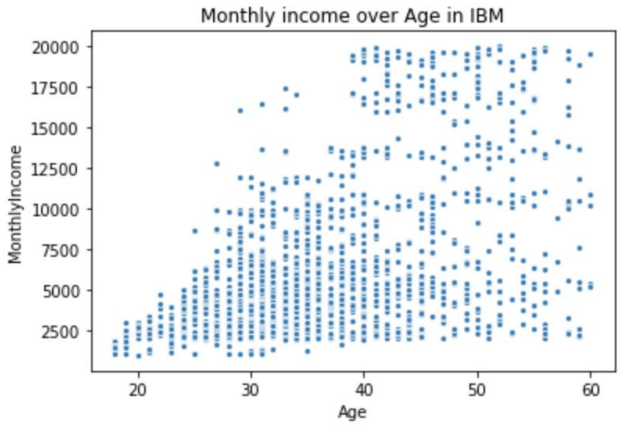


Figure1: Scatterplot for Age and Monthly Income(Appendix B)

Drilling down to the effect of ages, we observe the average income in different age groups in table 1 and construct boxplots in different groups.

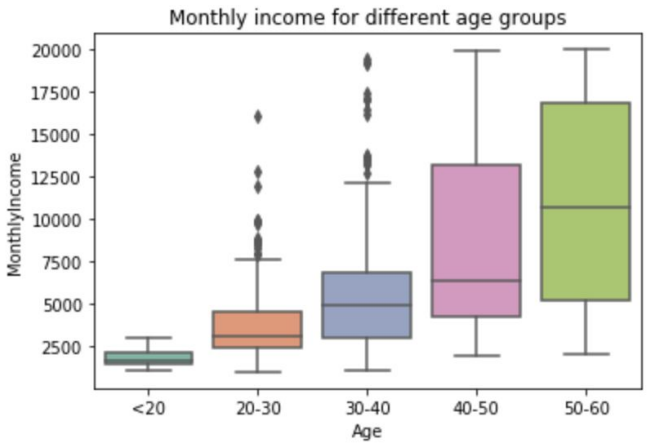


Figure 2: Monthly income different age groups(Appendix C)[3]

From Table 1 and Figure 2, it shows that older age groups have a higher average and median salary, but older age groups also vary more in ages, meaning that the income gap between the forties and fifties among individuals increases.

We can, therefore, conclude that age might have a positive relationship with monthly income, but the older an employee is, the more unpredictable it is of his or her salary. To measure the strength of this relationship, we perform a Pearson's correlation coefficient between age and income, which is 0.497. It means that age and income are moderately positive relationship, meaning that age and income increase and decrease in the same direction.[4]

Find other correlation factors

Besides ages, we want to examine what other factors might contribute to monthly income, so we

[3] **#dataviz:** I visualize the monthly salary of different age groups to see the variability and gap within each group.
[4] **#correlation:** I explain the relationship between age and monthly income and create correlation plot to look for other independent variables.

construct a correlation table. From Table 2, we can see that job level, years with current managers, total working years, years at the company, years since last promotion also associates with monthly income. We select age, job level, and total working years in our multiple regression models by forward selection approach.

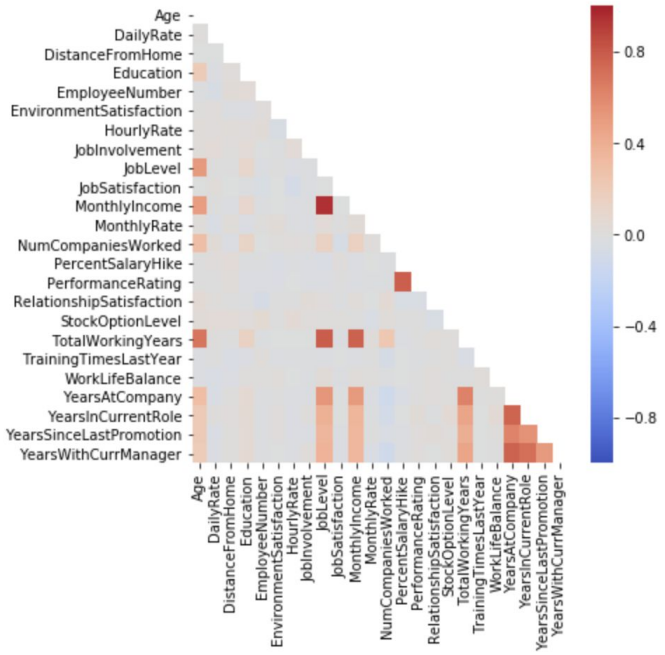


Figure 3: heatmap of correlation among factors(Appendix D)

Dep. Variable:	MonthlyIncome	**R-squared:**	0.905
Model:	OLS	**Adj. R-squared:**	0.905
Method:	Least Squares	**F-statistic:**	4679.
Date:	Fri, 24 Jan 2020	**Prob (F-statistic):**	0.00
Time:	19:50:26	**Log-Likelihood:**	-12784.
No. Observations:	1470	**AIC:**	2.558e+04
Df Residuals:	1466	**BIC:**	2.560e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-1621.3075	178.919	-9.062	0.000	-1972.272	-1270.343
Age	-7.5809	5.655	-1.341	0.180	-18.673	3.511
TotalWorkingYears	52.5431	9.169	5.731	0.000	34.558	70.529
JobLevel	3784.7365	54.895	68.945	0.000	3677.055	3892.418

Table 3 & Table 4: Result of multiple regression(Appendix E)

Assessing r-squared is an approach to evaluate how well our model can explain and predict outcomes. Though adding more predicted variables will create a more favorable prediction for the response variable for our multiple regression, it might lead to overfitting, which does not apply to the population. To penalize for the number of variables in the model, we assess adjusted r-squared. The value of adjusted r-squared is 0.905, indicating that 90.5% of the variation in monthly income can be explained by the combination of age, total working years, and job level. Our population regression model is

$$y_{income} = \beta_0 + \beta_1 x_{age} + \beta_2 x_{work\ year} + \beta_3 x_{job\ level} + \epsilon$$

From Table 3 and Table 4, we can construct the formula of our sample regression model

$$\hat{y}_{income} = -1621.3 - 7.58\, x_{age} + 52.54\, x_{work\ year} + 3784.74\, x_{job\ level}$$ [5]

In the model, Y hat is our predicted value of monthly income, and the regression coefficient of the total working year and job level are positive while age is negative. The coefficient of age represents that for every additional year in age, one can expect monthly income to decrease an average of 7.58 USD while holding total working year and job level constant. Though age has a positive correlation with monthly income and has a positive coefficient in single linear regression in Appendix F, it has a negative effect in this model. It can be due to the multicollinearity with job level and the total working year, which has a 0.51 and 0.68 correlation with age, respectively.

Statistical significance testing for age

We assume that there is a relationship between age and monthly income for our population, and the alternative hypothesis is that the two are not related to each other. Our two-tailed hypothesis for the slope of age is

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

[5] **#regression:** I calculate the population and sample regression model, adjusted R square, and the meaning of coefficients.

We set out significance level α as 0.05 and perform T test from Table 4

$$T = \frac{point\ estimate - 0}{SE} = \frac{-7.5809 - 0}{5.655} = 1.34$$ [6]

The corresponding p value 0.18 is when T is equal to 1.34 with degree of freedom 1466. Since p value > 0.05 we do not reject our hypothesis. The impact of age can be 0, which is different from our bivariate model. To understand the change, we check multicollinearity.

[6] **#significance:** I calculate its p value from the Table 3&4's T value.

Multicollinearity

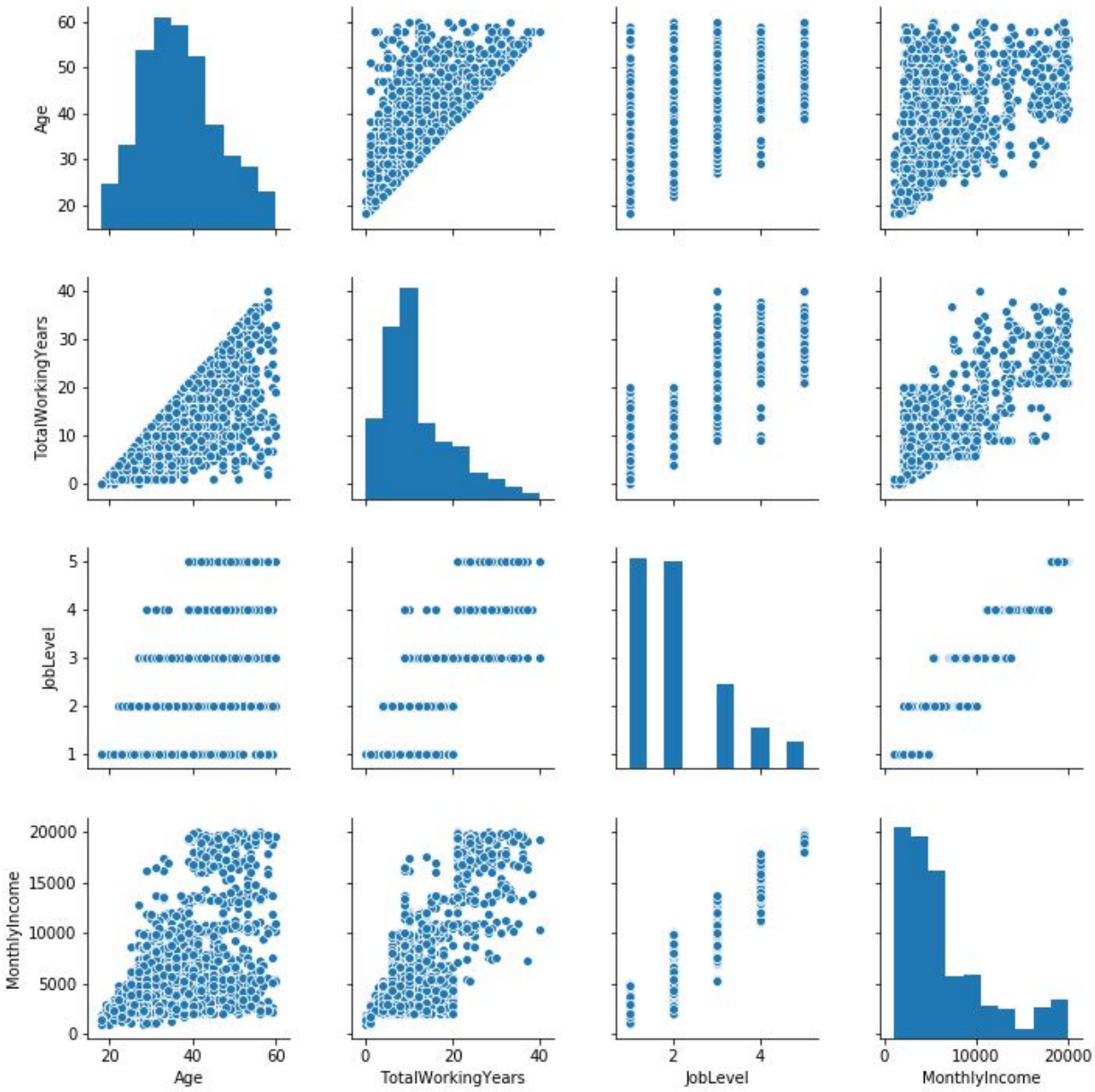


Figure 4 Pairplot of monthly income, age, and total working years(Appendix G)

We use a pair plot to check the multicollinearity for the selected variables. As mentioned before, we need to acknowledge that there is a pattern for the total working year and ages since it is

impossible for one's working year to exceed one's age. The correlation coefficient is 0.68. Another finding is that 20 years seems to be a threshold for monthly income. Most people's monthly salary is more than 5000 USD after 20 total working years. It is also possible to earn 20000 USD per month after working for 20 years.

We assess the quality of the model by the following four assumptions(Appendix H)**.**

- Linearity. The expected value of monthly income should be a linear function of age, job level, and total working years. We can see in general, and our predicted variables are linearly correlated with monthly income.

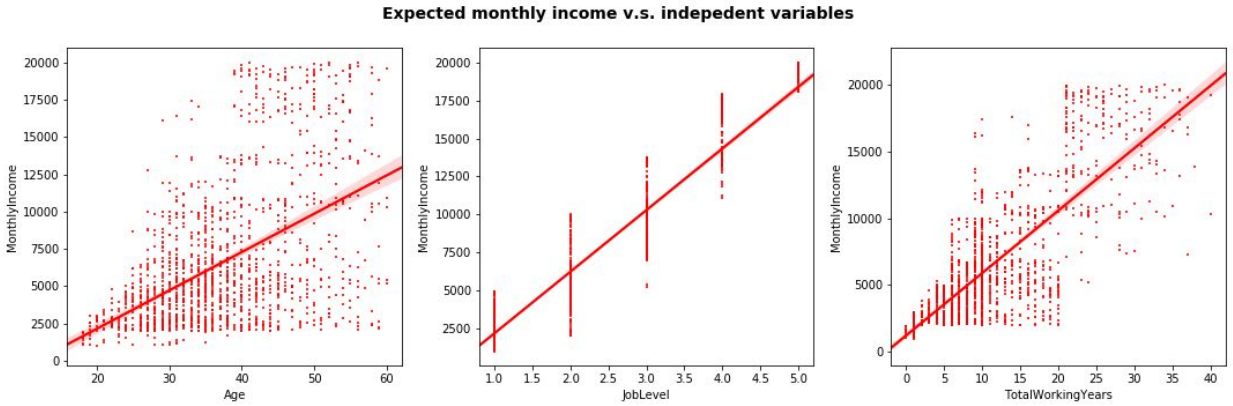


Figure 5 The regression and scatter plot for all independent variables(Age, Job level, Total Working Years) corresponds to dependent variable(Monthly income)

- Constant variability. The variance of the residuals is constant with our expected value. Though the variability of residuals is constant, we acknowledge 5 clusters might affect the quality of our model.

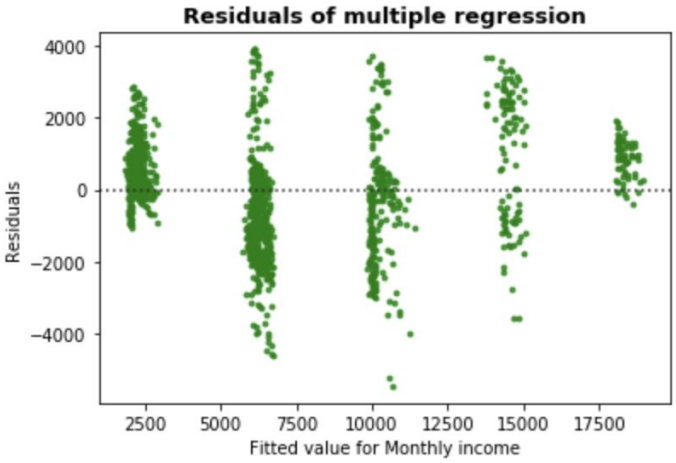


Figure 6 The residuals plot over fitted value for the multiple regression

- Nearly normal residuals: Residuals are distributed normally.

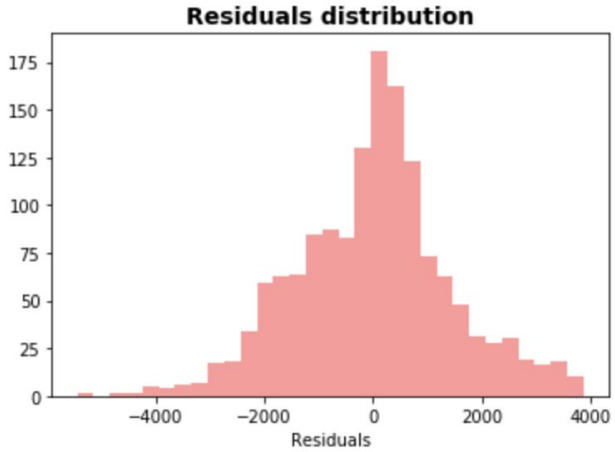


Figure 7[7] The residuals distribution for residuals in the multiple regression model

[7] **#distribution:** I plot the residual samples and find it is normally distributed, so I can draw inference to my population.

- Independent observations. The residuals are independent of our predicting variables. For the age and total working years, residuals are distributed uniformly randomly without specific clusters, so they are independent.

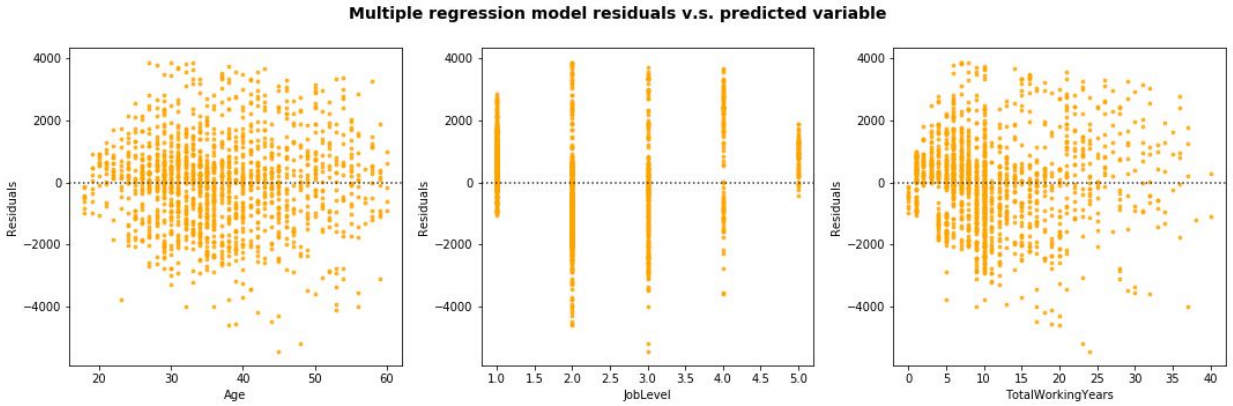


Figure 8 The residuals of the multiple regression model over all independent variables(Age, Job level, Total Working Years)

Conclusion

Though 90% of the variability of monthly income can be explained by the age, job level, and total working years. To conclude, first, we acknowledge the multicollinearity between the age and total working year will influence the quality of our model. Second, we would like to include measurement for categorical variables such as job level. Third, since we observed clusters for residuals, we expect there is another extraneous variable that affects our model. Including dummy variables for each cluster is the next approach to adopt.

WORD COUNT: 1180 words(not including the description of tables and figures)

Appendix

Appendix A

In [110]:
```
#group Age into 5 groups in range
num_data['Age_range'] = num_data['Age'].apply(lambda x: '<20'if x < 20
                                              else '20-30' if x>=20 and x<30
                                              else '30-40' if 40>x and x>=30
                                              else '40-50' if 50>x and x>=40
                                              else'50-60' if 60>=x and x>=50 else x)
num_data.sort_values('Age', inplace = True)
demical_per = num_data['Age_range'].value_counts(normalize = True) #normalize Age_range to demical points
percentage = demical_per.astype(float).map(lambda x: '{:.2%}'.format(x)) #transfer
percentage
```

Out[110]:
```
30-40    42.31%
40-50    23.74%
20-30    21.02%
50-60    11.77%
<20       1.16%
Name: Age_range, dtype: object
```

In [111]:
```
#calculate mean, median, and standard deviation and sort them
mean = num_data.groupby(['Age_range'])['MonthlyIncome'].mean().sort_values()
median = num_data.groupby(['Age_range'])['MonthlyIncome'].median().sort_values()
std = num_data.groupby(['Age_range'])['MonthlyIncome'].std().sort_values()

#rename for the table
mean.rename('mean', inplace = True)
median.rename('median', inplace = True)
std.rename('standard deviation', inplace = True)
percentage.rename('percentage', inplace = True)

table = pd.concat([round(mean,2), median, round(std,2)], axis=1) #connect all series and round mean and median to 2 digi
table.merge(percentage, left_index = True, right_index = True) #Add percentage into the table too
```

Out[111]:

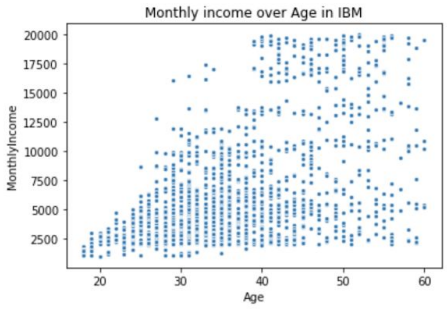
	mean	median	standard deviation	percentage
<20	1813.06	1675	548.33	1.16%
20-30	3795.74	3102	2034.55	21.02%
30-40	5599.25	4983	3128.11	42.31%
40-50	8537.96	6377	5621.88	23.74%
50-60	10942.91	10725	6005.30	11.77%

Appendix B

Relatioship between age and monthly income

```
In [15]: sns.scatterplot(x= 'Age', y = 'MonthlyIncome',data = num_data, s = 15)
         plt.title('Monthly income over Age in IBM')
```

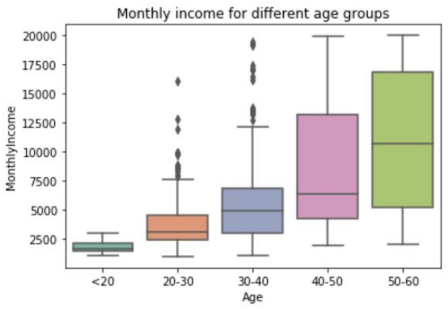
```
Out[15]: Text(0.5,1,'Monthly income over Age in IBM')
```



Appendix C

```
In [130]: #Create boxplot for different age group
          sns.boxplot(x= 'Age_range', y = 'MonthlyIncome',data = num_data, palette="Set2")
          plt.title('Monthly income for different age groups')
          plt.xlabel('Age')
```

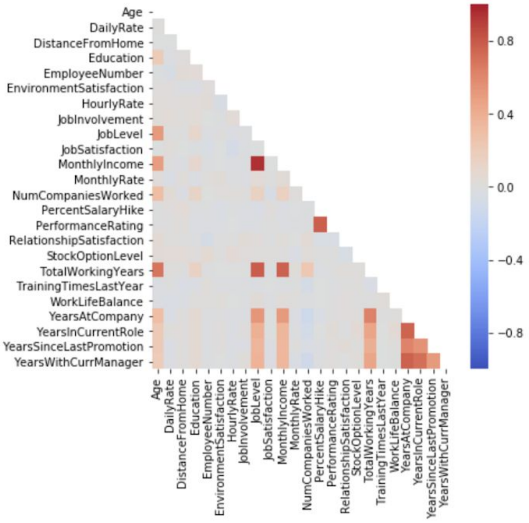
```
Out[130]: Text(0.5,0,'Age')
```



Appendix D

In [106]:
```
fig, ax = plt.subplots(figsize=(6,6))
mask = np.zeros_like(corr) #Return an array of zeros with the same shape and type as a given array
mask[np.triu_indices_from(mask)] = True #triu_indices: Return the indices for the upper-triangle of an (n, m) array.
sns.heatmap(corr, vmin = -1, vmax = 1, mask = mask, cmap = 'coolwarm')
```

Out[106]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x127c47f60>
```



Appendix E

Multiple regression statistics

119]:
```
x = num_data[['Age','TotalWorkingYears', 'JobLevel']]
x = sm.add_constant(x)
y = num_data['MonthlyIncome']
multimodel = sm.OLS(y,x).fit()
multimodel.summary()
```

```
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site-packages/numpy/core/fromnumeric.py:2389: FutureW
arning: Method .ptp is deprecated and will be removed in a future version. Use numpy.ptp instead.
  return ptp(axis=axis, out=out, **kwargs)
```

Appendix F: Result of Bivariate regression

Bivariate regression Statistics

```
[21]: x = num_data['Age']
      x = sm.add_constant(x)
      y = num_data['MonthlyIncome']
      model = sm.OLS(y,x).fit()
      model.summary()
```

```
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site-packages/numpy/core/fromnumeric.py:2389: FutureW
arning: Method .ptp is deprecated and will be removed in a future version. Use numpy.ptp instead.
  return ptp(axis=axis, out=out, **kwargs)
```

OLS Regression Results

Dep. Variable:	MonthlyIncome	**R-squared:**	0.248
Model:	OLS	**Adj. R-squared:**	0.247
Method:	Least Squares	**F-statistic:**	483.8
Date:	Mon, 27 Jan 2020	**Prob (F-statistic):**	6.67e-93
Time:	18:16:17	**Log-Likelihood:**	-14308.
No. Observations:	1470	**AIC:**	2.862e+04
Df Residuals:	1468	**BIC:**	2.863e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-2970.6712	443.702	-6.695	0.000	-3841.030	-2100.313
Age	256.5716	11.665	21.995	0.000	233.689	279.454

Omnibus:	140.178	**Durbin-Watson:**	1.963
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	182.119
Skew:	0.799	**Prob(JB):**	2.84e-40
Kurtosis:	3.649	**Cond. No.**	159.

Appendix G

Check multicollinearity

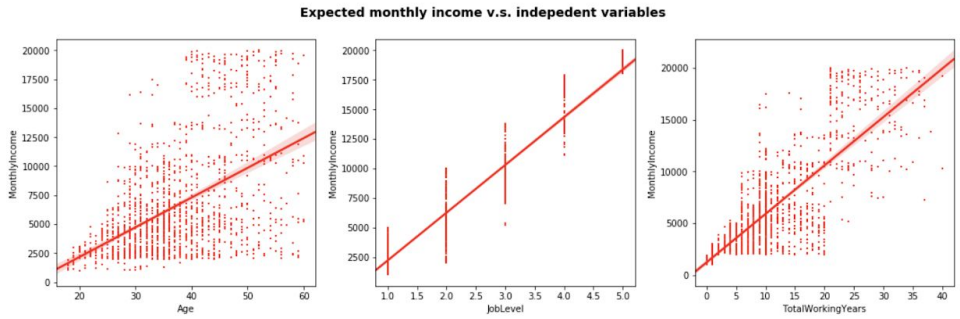
```
126]: column_x = ['Age','TotalWorkingYears', 'JobLevel']
      column_y = ['MonthlyIncome']
      columnstoplot = column_x + column_y

      sns.pairplot(data[columnstoplot])
      plt.tight_layout()
      plt.savefig('pairplot.png')
      #Total working year cannot exceed age
```

Appendix H

linearity

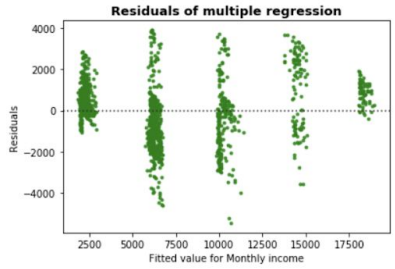
```
123]: fig, (ax1, ax2, ax3) = plt.subplots(1, 3, figsize=(15,5))
      sns.regplot(num_data['Age'], num_data['MonthlyIncome'], fit_reg = True, scatter_kws ={'s': 5}, marker = '+', color = 'r', ax = ax1)
      sns.regplot(num_data['JobLevel'], num_data['MonthlyIncome'], fit_reg = True, scatter_kws ={'s': 5}, marker = '+', color = 'r', ax = ax2)
      sns.regplot(num_data['TotalWorkingYears'], num_data['MonthlyIncome'], fit_reg = True, scatter_kws ={'s': 5}, marker = '+', color = 'r', ax = ax3)
      plt.tight_layout()
      fig.suptitle('Expected monthly income v.s. indepedent variables', fontweight='bold',fontsize=14)
      plt.subplots_adjust(top=0.88)
      plt.savefig('linearity.png')
```



Constant variability

```
[58]: sns.residplot(x= multimodel.predict() , y = multimodel.resid ,data = num_data, scatter_kws={'s':8}, color = 'g')
      plt.title('Residuals of multiple regression', fontweight = 'bold', fontsize = 13)
      plt.ylabel('Residuals')
      plt.xlabel('Fitted value for Monthly income')
```

```
[58]: Text(0.5,0,'Fitted value for Monthly income')
```

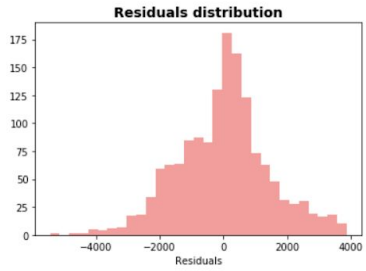


Normality

121]:
```
distplot = sns.distplot(multimodel.resid, kde=False, axlabel='Residuals', color='red') #The code is referred from the class
distplot.set_title('Residuals distribution',fontweight='bold',fontsize=14)

qqplot = sm.qqplot(multimodel.resid,fit=True,line='45')
qqplot.suptitle("Normal Probability (\"QQ\") Plot for Residuals",fontweight='bold',fontsize=14)
```

121]:
```
Text(0.5,0.98,'Normal Probability ("QQ") Plot for Residuals')
```



Indepedence

[98]:
```
fig, (ax1, ax2, ax3) = plt.subplots(1, 3, figsize=(15,5))
sns.residplot(x= num_data.Age , y = multimodel.resid ,data = num_data, scatter_kws={'s':8}, color = 'orange', ax = ax1)
sns.residplot(x= num_data.JobLevel , y = multimodel.resid ,data = num_data, scatter_kws={'s':8}, color = 'orange', ax = ax2)
sns.residplot(x= num_data.TotalWorkingYears , y = multimodel.resid ,data = num_data, scatter_kws={'s':8}, color = 'orange', ax = ax3)

ax1.set_ylabel("Residuals")
ax2.set_ylabel("Residuals")
ax3.set_ylabel("Residuals")
fig.suptitle('Multiple regression model residuals v.s. predicted variable', fontweight='bold',fontsize=14)


plt.tight_layout()
plt.subplots_adjust(top=0.88)
plt.savefig('independece.png')
```

Reference

Kalleberg, A. L., & Loscocco, K. A. (1983). Aging, Values, and Rewards: Explaining Age Differences in Job Satisfaction. American Sociological Review, 48(1), 78. doi: 10.2307/2095146

Koehrsen, W. (2018). Visualizing data with pairs plots in python. Retrieved from https://towardsdatascience.com/visualizing-data-with-pair-plots-in-python-f228cf529166

Moreno, A. I. (2019). Simple and multiple linear regression with python. Retrieved from https://towardsdatascience.com/simple-and-multiple-linear-regression-with-python-c9ab422ec29c

Sarkar, T. (2019). How do you check the quality of your regression model in python? Retrieved from https://towardsdatascience.com/how-do-you-check-the-quality-of-your-regression-model-in-python-fa61759ff685