Trabalho de grupo
Econometria dos Mercados Financeiros

Regressão

Ficheiro "employees.csv"

Este ficheiro continha 1000 observações para cada uma das variáveis seguintes:

- "*First Name*"
- "*Gender"*
- "*Start Date"*
- "*Last Login Time"*
- "*Salary"*
- "*Bonus %"*
- "*Senior Management"*
- "*Team"*

Da análise aos dados verificámos que:

- Tínhamos dois tipos de variáveis no ficheiro, quantitativas ("Salary"; "Bonus %";"Start Date") e qualitativas (as restantes variáveis);
- Existiam dados em falta para algumas variáveis

Questões:

1. **Analisar gráficos e estatísticas descritivas das variáveis**

Para tentar fazer esta questão adotámos vários procedimentos, descritos abaixo:

- Começámos por eliminar duas colunas do nosso ficheiro, "First Name" e "Last Login Time", por considerámos não serem variáveis relevantes para o estudo que queríamos fazer;
- Para substituir a variável "Start Date", que representa o ano em que o colaborador entrou na empresa, criámos uma nova variável "*Experience"*:

$$Experience = 2019 - StartDate$$

  A variável "*Experience*" representa o número de anos de experiência na empresa a que pertencem atualmente os colaboradores.

- Eliminámos as linhas com dados nulos, "NaN".

Análise dos dados obtidos

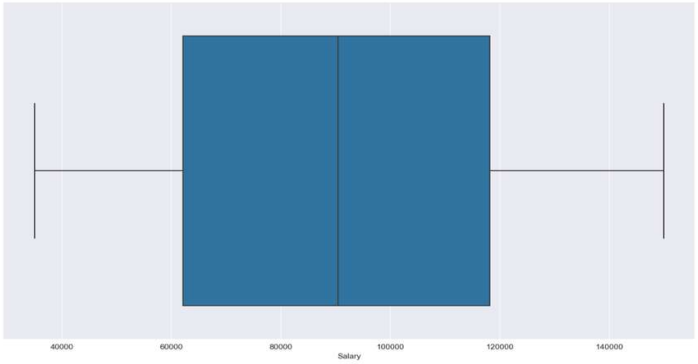
Começámos por fazer um estudo das variáveis numéricas. Refira-se que, após a eliminação das linhas nulas, passámos a ter 764 observações para todas as variáveis.

	Salary	Bonus %	Experience
count	764.000000	764.000000	764.000000
mean	90433.196335	10.148041	20.316754
std	32864.665282	5.608733	10.292788
min	35013.000000	1.015000	3.000000
25%	62071.750000	5.193250	11.000000
50%	90428.000000	9.658500	20.000000
75%	118075.250000	14.965000	29.000000
max	149908.000000	19.944000	39.000000

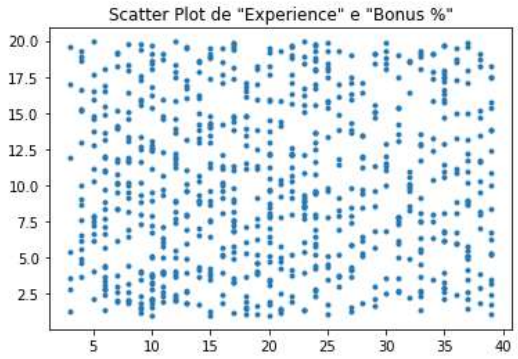
Da análise dos resultados obtidos podemos verificar que:

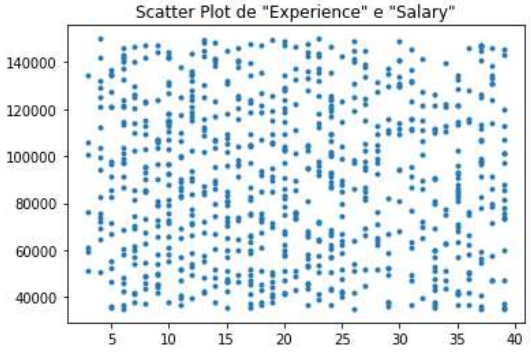
- A média dos salários ronda 90.433. No entanto, como a média é superior à mediana, 90.428, assume-se que a distribuição dos salários é assimétrica positiva. Para confirmar este facto, foi calculado o coeficiente de assimetria, que resultou em 0,0469, ou seja, podemos concluir que a distribuição é assimétrica positiva. No entanto, como o valor da média está próximo da mediana, verifica-se que esta assimetria não é muito significativa.
- O valor médio dos bónus recebidos ronda os 10% e todos os funcionários recebem bónus sobre o salário, entre 1% e 19,9%.
- A média dos anos de experiência é cerca de 20 anos (considerando como ano de análise: 2019) e os trabalhadores com menos experiência estão na empresa há 3 anos.
- Observando os dados percebe-se também que 50% dos trabalhadores, 382, têm mais de 20 anos de experiência e que, destes, 25% (191) têm mais de 29 anos de experiência.
- Foi calculado o excesso de curtose, que resultou em -1,1618, o que evidencia que a distribuição dos salários é platicúrtica (pois o valor é inferior a 3).

  Foi obtido o "box plot" dos salários da amostra, que permite verificar o concluído anteriormente de que, apesar de os salários apresentarem uma distribuição assimétrica positiva, esta assimetria não é significativa.



De seguida fizemos os seguintes "*scatter plot*":





Os gráficos acima representam a relação entre o salário e i) anos de experiência na empresa e ii) bónus, em %. Da análise dos mesmos, verifica-se que a relação entre as variáveis é muito dispersa. Não se consegue concluir por uma relação linear entre as variáveis.

De seguida, olhámos para as estatísticas descritivas das variáveis quantitativas separadas para trabalhadores masculinos e femininos.

	Salary	Bonus %	Experience
count	371.000000	371.000000	371.000000
mean	91170.851752	10.346571	19.743935
std	32049.177253	5.687854	10.523163
min	35013.000000	1.015000	3.000000
25%	64267.500000	5.231500	10.000000
50%	91124.000000	10.169000	19.000000
75%	118452.500000	15.404500	28.000000
max	148985.000000	19.944000	39.000000

Trabalhadores masculinos

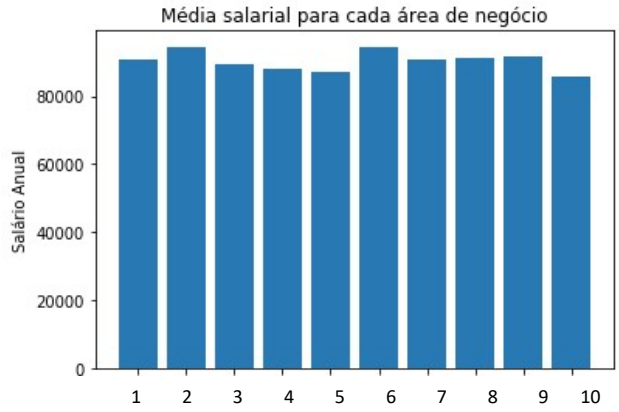
	Salary	Bonus %	Experience
count	393.000000	393.000000	393.000000
mean	89736.834606	9.960623	20.857506
std	33642.323667	5.533714	10.053939
min	35381.000000	1.027000	3.000000
25%	59070.000000	5.146000	13.000000
50%	89780.000000	9.375000	20.000000
75%	118037.000000	14.249000	29.000000
max	149908.000000	19.850000	39.000000

Trabalhadores femininos

A análise destes dados permite-nos concluir que:

- A média de salários é superior para os homens, assim como a percentagem dos bónus recebidos, embora, em média, sejam as mulheres que apresentam maior experiência nesta empresa.

Foi ainda obtido o gráfico com a média salarial pelo departamento:



1. Marketing
2. Finance
3. Client Services
4. Legal
5. Product
6. Engeneering
7. Business Development
8. Human Resources
9. Sales
10. Distribution

A elaboração do gráfico de barras evidencia alguma diferença entre as médias salariais por departamento. Verificou-se que o departamento de Finanças é onde a média salarial é mais elevada. Em contrapartida, é no departamento de Distribuição que, em média, se ganha menos.

Por último, foi obtido o gráfico com a média salarial entre colaboradores que são Senior Managers (SM) ou não (nSM). Verifica-se que em média, os colaboradores que têm cargos de SM ganham ligeiramente acima dos restantes.

De forma a analisar a correlação entre cada uma das variáveis da regressão, foi ainda obtido o "heatmap" da correlação:

Da análise deste mapa verifica-se que não existe uma correlação significativa entre as variáveis consideradas nesta amostra, pois todos os coeficientes de correlação apresentam-se abaixo de 0,1 em valor absoluto. Os gráficos de correlação das variáveis, duas a duas, podem ser consultado no anexo a este documento.

2. **<u>Analisar a estacionariedade das variáveis (quando se trata de séries temporais).</u>**

Apenas aplicável a time series.

3. **Descobrir o melhor modelo que se ajusta aos dados e justificar a escolha do respetivo modelo**

De forma a encontrar o melhor modelo de regressão linear foram realizados alguns ajustamentos à amostra disponibilizada, nomeadamente:

a. Colunas que não iriam ser incluídas na regressão ("*First Name*"; "*Start Date*"; "*Last Login Time*") foram removidas

b. Transformação das variáveis qualitativas em "dummy":

- ‘‘Gender" com 2 categorias: Male == 1 e Female == 0;
- ‘‘Senior Management" com 2 categorias: True == 1 e False == 0;
- ‘‘Team" com 10 categorias:
  - Marketing == 1;
  - Finance == 2;
  - Legal == 3;
  - Product == 4;
  - Client Services == 5;
  - Engineering == 6;
  - Business Development == 7;
  - Human Resources == 8;
  - Sales == 9;
  - Distribution == 10.

Adicionalmente, para as variáveis "dummy", foi sempre eliminada uma das categorias de cada variável, de forma a evitar o risco de obter multicolinearidade entre as variáveis.

Após estes ajustamentos, com recurso aos diversos pacotes de tratamento estatísticos disponíveis para Python, foi definida a regressão linear utilizando o método dos mínimos quadrados.

Este modelo define como:

- Variável dependente: "*Salary*";

- Variáveis independentes: "*Gender*", "*Bonus %*", "*Senior Management*", "*Team*" e "*Experience (years)*".

A sua forma geral é:

$$Salary_i = \beta_1 Gender_1 + \beta_2 Bonus_2 + \beta_3 SeniorManagement_3 + \beta_4 Team_4 + \beta_5 Experience_5 + \varepsilon$$

onde β_1= 1,615e+04; β_2= 1982,3496; β_3= 1,638e+04; β_4 = 4332,0347; β_5 = 1325,4885

Conclusões obtidas com este modelo:

Este modelo foi aquele que apresentou um coeficiente de determinação mais elevado, de 0,812, ou seja, 81,2% da variação do salário é explicado pela variação das variáveis independentes consideradas neste modelo. Este modelo assume que:

a. Em média, os homens ganham mais 16.150 que as mulheres.
b. Os colaboradores em funções de SM ganham em média mais 16.380 face aos restantes colaboradores.

De forma a testar utilidade do modelo foi utilizada a **estatística teste F (global)**. Considerando o teste de hipótese que define como hipótese nula que todos os coeficientes sejam iguais a zero versus a hipótese alternativa que assume que pelo menos um dos coeficientes é diferente de zero.

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs. H_1: pelo menos um β_i diferente de zero.

Foi obtido o p-value de 1,81e-272 (≈0), claramente inferior ao valor definido para nível de significância (alfa = 0,05), o que significa a rejeição da hipótese nula. Podemos então concluir que pelo menos um dos coeficiente β_i é diferente de zero.

Adicionalmente, foi ainda verificada a relevância estatística do coeficiente de cada variável independente do modelo, definindo como hipótese nula o coeficiente i seja igual a zero:

H_0: $\beta_i = 0$ vs. H_1: $\beta_i \neq$ de zero

Os p-value obtidos para cada coeficiente foram:

β_1 == 0,00	β_2 == 0,00	β_3 == 0,00	β_4 == 0,00	β_5 == 0,00

Conclusão: Todos os p-value são inferiores a 0,05, o que implica rejeitar a hipótese nula. Assim, todos os coeficientes são estatisticamente significantes e devem ser incluídos na regressão linear.

Apesar de com este modelo se obter um R-square relativamente elevado e os testes estatísticos levarem a concluir pela significância do modelo, quando se analisou a linearidade dos dados verificou-se que a relação entre estes não é linear, pois os gráficos em baixo não demonstram padrões de linearidade. Assim, sabemos que a regressão linear não é aquela que melhor reflete a relação entre as variáveis. A solução para esta situação passaria por recorrer a modelos de relação não linear (ver Anexo).

4. **<u>Analisar as propriedades dos resíduos dos modelos estimados (do melhor modelo de cada base de dados)</u>**

Foram analisados os resíduos do modelo considerado, de forma a verificar se estes cumprem com os pressupostos exigidos:

Pressuposto 1: <u>(os erros têm média zero)</u>

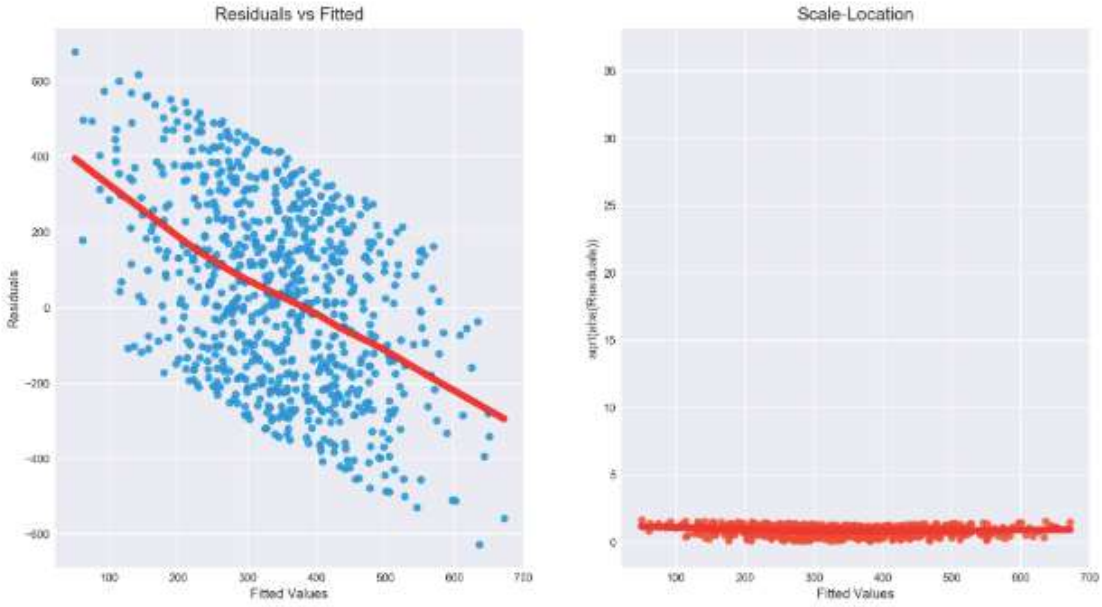
A média dos resíduos do modelo é de 7.317 (significativamente diferente de zero), o que leva a que não esteja cumprido este pressuposto para os resíduos.

Neste modelo está definido que o termo constante é nulo (β_0= 0).

A solução poderia passar por definir um novo modelo com o termo constante diferente de zero. Esta situação foi testada, tendo sido obtido um coeficiente de interceção de β_0= 91.117. No entanto, neste modelo, o coeficiente de determinação era de 0,003, e o p-value obtido para a estatística F era acima de 0,05, o que levaria à não rejeição da hipótese nula de que todos os coeficientes do modelo são iguais a zero. Ou seja: este modelo não seria adequado para estabelecer uma regressão linear para os dados em análise, pelo que se optou por considerar o modelo definido na questão 3.

Pressuposto 2: <u>(a variância dos erros é constante e finita) – homocedasticidade</u>

O objetivo da representação gráfica dos resíduos versus uma ou mais variáveis independentes serve para, de uma forma superficial, mostrar que a variância não é constante. Dito de outra forma: os gráficos não são em forma de funil. Esta situação não se verifica para os gráficos apresentados em baixo, no entanto, esta verificação é não formal.



Assim, de forma a testar a homocedasticidade dos resíduos, foi realizado o teste **Breusch-Pagan** em que foi definida com hipótese nula H_0: erros homocedásticos (variância dos erros constante) versus a hipótese alternativa H_1: erros heterocedásticos. Foi obtido um p-value de 1,104143e-53 (≈0) o que leva à rejeição da hipótese nula, ou seja, o modelo não cumpre com o pressuposto de que a variância dos erros é constante.

Por outro lado, o teste **Goldfeld-Quandt** leva a concluir de forma oposta ao teste de **Breusch-Pagan**, pois definindo como hipótese nula que os resíduos são homocedásticos, obteve-se um p-value de 0.649930, que leva à não rejeição da hipótese nula, ou seja, a variância dos erros é constante.

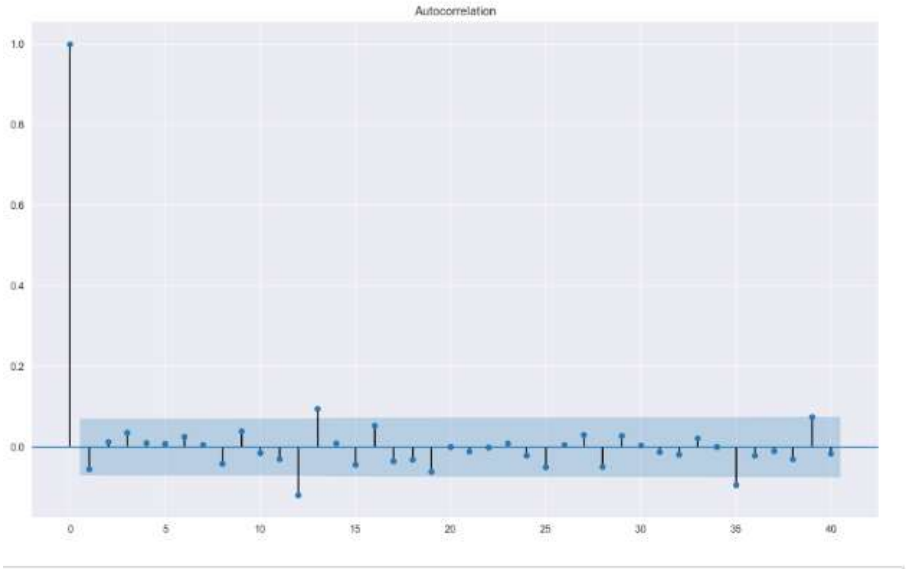
Pressuposto 3: <u>(os erros são linearmente independentes)</u>

Para testar se os erros são linearmente independentes foi realizado o teste **Durbin-Watson**, onde foi definido como hipótese nula que H_0: a correlação é igual a zero (os erros são independentes, não existindo correlação entre eles) versus a hipótese alternativa H_1: a correlação ser diferente de zero.

Os resultados para este teste foram: 2,045, que podemos assumir, por arredondamento, que é igual a 2, o que indica que não existe correlação entre os resíduos.

No entanto, neste caso temos β_0= 0 , e uma das condições para a aplicação do teste de Durbin-Watson é que o valor do intercept seja diferente de zero.

O gráfico da autocorrelação parece indicar que existe correlação entre os resíduos no desfasamento 12 e 13.



Pressuposto 4: (os erros são normalmente distribuídos)

A hipótese de que os resíduos são normalmente distribuídos é muito importante para a realização de testes de hipóteses e intervalos de confiança sobre os parâmetros.

Para testar essa condição recorremos ao teste **Jarque-Bera** onde é definido como hipótese nula, H_0: os dados seguem uma distribuição normal versus a hipótese alternativa (H_1) de que os erros não seguem uma distribuição normal. Para o modelo definido foi obtido um p-value de 0,031, ou seja, rejeitamos a hipótese nula e, consequentemente, os dados não seguem uma distribuição normal.

Para confirmar essa conclusão, foram ainda efetuados os testes **Shapiro-Wilk** e **Kolmogorov-Smirnov**. Em ambos os testes define-se como hipótese nula, H_0: resíduos seguem uma distribuição normal versus H_1: resíduos não seguem uma distribuição normal. Os p-value obtidos foram: i) Shapiro-Wilk: p-value = 0,0318 e ii) Kolmogorov-Smirnov: p-value = 0,000. Assim, para ambos, a conclusão é a mesma, considerando um alfa de 0,05, o que conduz à rejeição da hipótese nula. Logo não é possível assumir que os resíduos assumem uma distribuição normal.

Esta conclusão sobre a não normalidade dos resíduos, tal como referido, pode levar a que os testes de hipótese e intervalos de confiança utilizados sobre os parâmetros do modelo não estejam corretos.

Pressuposto 5: (a variável explicativa é independente dos resíduos)

Para verificar este pressuposto foram utilizados os coeficientes de correlação de Pearson (que testa a correlação entre os resíduos e as variáveis independentes).

Para todas as variáveis independentes foi obtido neste teste um p-value de 0, ou seja, rejeitamos a hipótese nula (não existência de correlação) para cada par, logo não se pode afirmar que este pressuposto esteja cumprido.

Em suma, verifica-se que este modelo não cumpre com o conjunto de pressupostos assumidos para os resíduos. Isto apesar de o modelo definido na questão 3 apresentar um R^2 elevado e cada uma das variáveis independentes ser estatisticamente significativa para explicar o salário auferido pelo trabalhador.

5. **Fazer a previsão out-of-sample de cada uma das variáveis de interesse, um ponto no futuro**

Tendo em consideração o modelo selecionado, para efeitos de previsão foram considerados os seguintes casos:

i. Um colaborador do sexo masculino, SM, com 26 anos de experiência, da equipa de marketing e com um bónus de 6,945% ganhará 42.173;

ii. Um colaborador do sexo feminino, e com as restantes condições exatamente iguais às referidas na i) ganhará 40.558;

iii. Um colaborador do sexo feminino, nSM, com 26 anos de experiência, da equipa de marketing e com um bónus de 6,945%, ganhará 72.500;

iv. Um colaborador do sexo masculino, SM, com 26 anos de experiência, da equipa de cliente services e com um bónus de 6,945% ganhará 59.501.

Apesar de se usar o modelo definido em 3 para estimar o salário nesta questão, é importante ressaltar que, estamos conscientes da limitação do modelo (relação não linear entre as variáveis, falha nos pressupostos para os resíduos).

Trabalho de grupo
Econometria dos Mercados Financeiros

Séries Temporais

Ficheiro "MerkDaily"

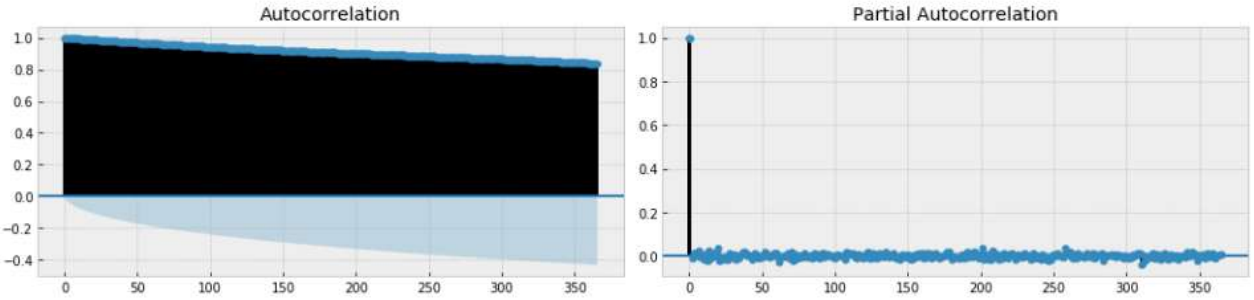
Questões:

1. **<u>Gráficos e estatísticas descritivas</u>**

A base de dados disponibilizada contém a cotação diária da Merk (empresa farmacêutica) desde 02/01/1970 até 23/01/2019. Para efeitos da análise que se segue foi considerada a informação que consta na coluna "adj Close".

O gráfico que representa a evolução das cotações da Mark apresenta-se de seguida:

E as funções de autocorrelação e de autocorrelação parcial apresentam-se:



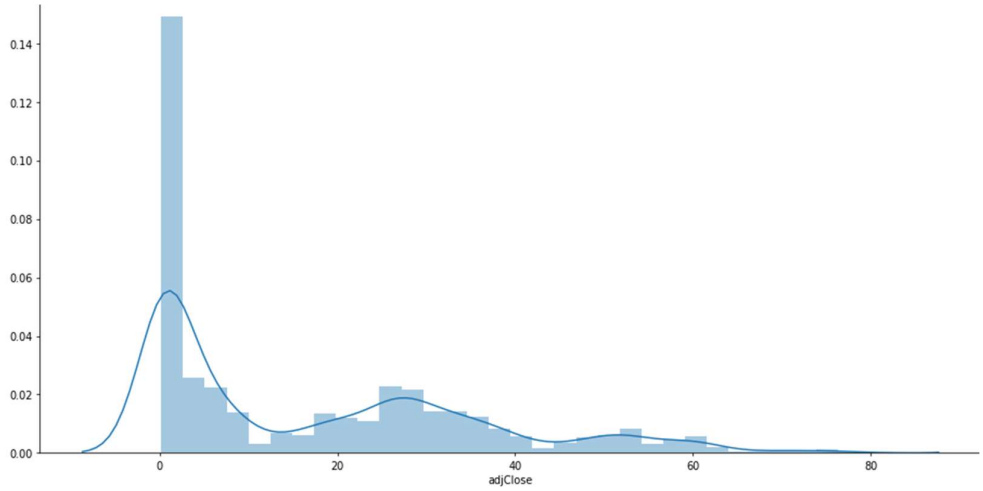
Quer o gráfico do comportamento do preço, quer os correlogramas ACF e PACF mostram que:

A série parece apresentar uma tendência positiva (existe uma correlação quase perfeita entre o valor num período e no seu período anterior) e ser cíclica.

A série não parece conter sazonalidade (não existe uma correlação significativa para certos intervalos de tempo (5 dias, 1 mês, 3 meses, etc))

O gráfico que representa a frequência do preço ajustado de fecho é apresentado em baixo. Este mostra que que as cotações dos títulos da Merk não seguem uma distribuição normal.

A série a trabalhar apresenta um total de 12.376 dados, uma média de 17,078927, um desvio padrão de 18,361550. A distribuição é enviesada à direita (Skewness = 0,9033677539952046) e é platicúrtica (excesso de Curtose = -0,1302936188109336).



Adicionalmente foi realizado um teste de hipótese (Jarque-Bera) onde é definida como hipótese nula (H_0): a distribuição das cotações segue uma distribuição normal. Foi obtido um p-value de 0 o que leva à rejeição da H_0, logo a série não segue uma distribuição normal.

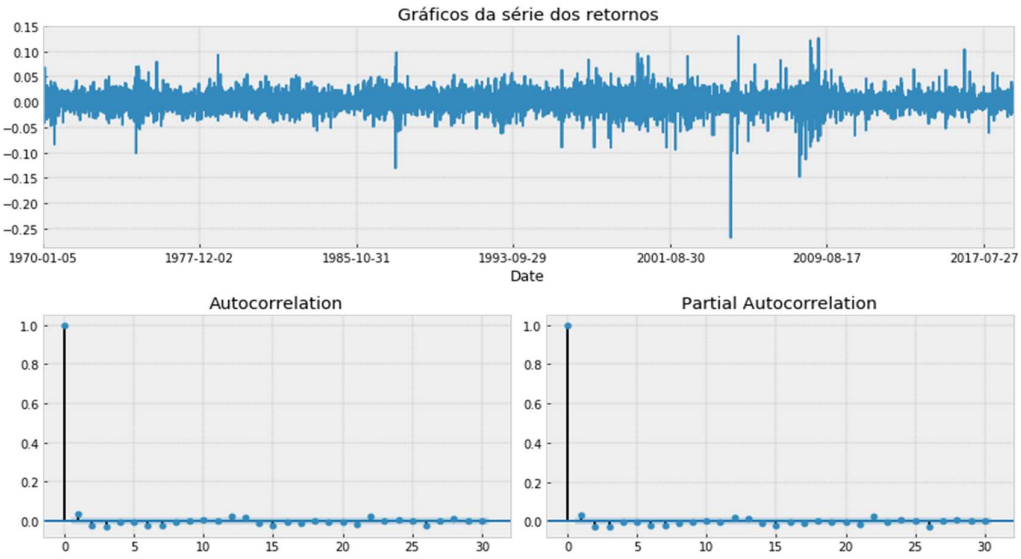
2. Estudar a estacionariedade

Como é visível no gráfico acima, com a evolução das cotações deste título, a sua média não é constante, logo esta série não é estacionária. O gráfico da função ACF também confirma a não estacionariedade (os valores passados influenciam a série hoje, criando tendência na série)

De forma a ter uma séria estacionária, foi definida uma nova variável: os retornos simples das cotações de fecho ajustadas.

$$Retornos\ Simples = \frac{y_t - y_{t-1}}{y_{t-1}}\ onde\ y_t = valor\ da\ ação\ no\ período\ t$$

Com essa série foram obtidos os seguintes resultados (que parecem confirmar a estacionaridade da série):



Para confirmar os resultados obtidos através da representação gráfica foi realizado o teste de Dickey-Fuller Aumentado (ADF), onde é definido como hipótese nula a existência de não estacionariedade. Formalmente, H_0: rho - 1 = 0 e H_1: rho – 1 < 0. Para a série das cotações de fecho ajustadas foi obtido um p-value para o ADF de 0,934 (não se rejeita H_0, o que significa que a série não é estacionária). Para a série dos retornos das cotações foi obtido um p-value para o ADF de 0 (rejeita-se H_0, o que significa que a série é estacionária).

Foram também utilizados os testes Phillipe-Perron (PP) e Kwiatkowski-Phillips-Schmidt-Shin (KPSS). O primeiro (PP) assume como H_0: série é não estacionária. O segundo assume como H_0: série é estacionária. Os p-values obtidos confirmaram as conclusões anteriores: série em níveis é não estacionária e série de retornos é estacionária.

Todos os testes confirmam que a série em níveis é não estacionária, mas a série da primeira diferença é estacionária. Logo a série é integrada de ordem 1 ou I(1).

3. <u>Definir o modelo que melhor se ajusta aos dados</u>

Depois de um processo de tentativa e erro concluiu-se que um modelo EGARCH é o que melhor se ajusta aos dados. A forma final do modelo é:

$$y_t = 0{,}0005 + 0{,}039 \times y_{t-1} - 0{,}0225 \times y_{t-2} + u_t, u_t \sim N(0, \sigma_t^2)\ ln(\sigma_t^2)$$
$$= -0{,}1474 + 0{,}0706 \times \left(\left| \frac{u_{t-1}}{\sigma_{t-1}} \right| - \sqrt{\frac{2}{\pi}} \right) - 0{,}0548 \times \frac{u_{t-1}}{\sigma_{t-1}}$$
$$+ 0{,}9819 \times ln(\sigma_{t-1}^2)$$

Começámos por tentar adaptar um modelo ARIMA à nossa série em níveis, pois esta é I(1). Utilizando a função pmdarima.auto_arima (AA), o modelo recomendado foi um ARIMA(2,1,1), pois era o que apresentava o menor AIC. Contudo, ao realizar a análise dos resíduos (mais concretamente, no teste para "efeitos ARCH"), verificou-se que os resíduos passados influenciavam os resíduos futuros (rejeitava-se a hipótese nula). Isso indica que a variância não é constante ao longo do tempo, logo um modelo da família ARCH seria mais apropriado para a série em análise. Tentou-se ainda reduzir a amostra para utilizar o modelo ARIMA para efeitos de previsão, mas a função AA não encontrou uma especificação do modelo que contivesse "lags" quer AR, quer MA.

Dentro dos modelos da família ARCH, tentaram-se dois modelos. Um modelo AR(2) + GJR-GARCH(1,1), pois este captaria o efeito de alavancagem, um facto estilizado de séries financeiras. O outro modelo testado foi AR(2) + EGARCH(1,1), pois captaria o facto estilizado de choques de preço positivos e negativos terem efeitos diferentes na volatilidade do ativo. A utilização de 2 "lags" AR deveu-se ao facto do modelo ARIMA "ótimo" utilizar esse número.

O modelo GJR-GARCH foi abandonado pois um dos coeficientes que multiplica o quadrado do resíduo do momento temporal anterior não era estatisticamente significativo. Além disso o Akaike Information Criterion (AIC) era superior (-54.400,3) ao obtido com o modelo EGARCH (-68.694,1).

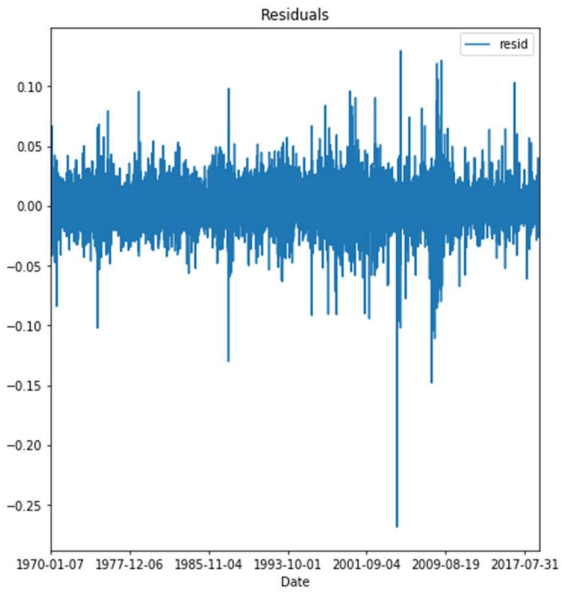
4. Estudar resíduos do modelo estimado

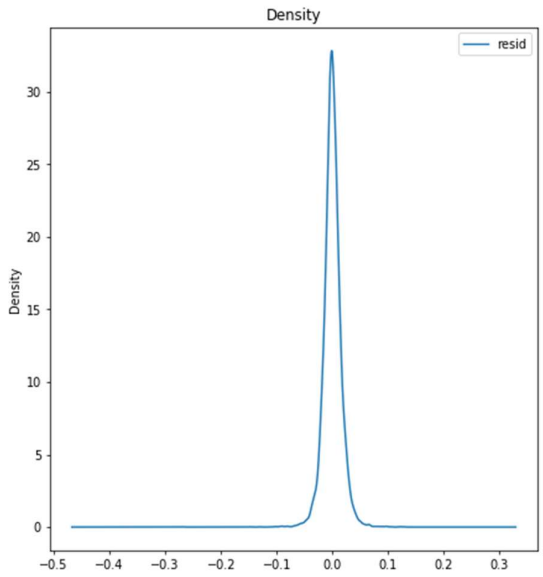
No que diz respeito à análise dos resíduos do modelo escolhido, devemos verificar três pressupostos:

- O valor esperado da série deve ser 0;
- A sua variância deve ser constante;
- Os erros não devem apresentar correlação serial.

Pressuposto 1 (O valor esperado da série deve ser 0)

Os gráficos que se seguem parecem mostrar que a média é algum valor perto de 0.





Também se calculou a média dos resíduos em -0,000124, valor perto de 0.

Pressuposto 2 (A sua variância deve ser constante)

A variância obviamente não é constante e foi uma das razões para utilizarmos um modelo da família ARCH.

Pressuposto 3 (Os erros não devem apresentar correlação serial)

No que diz respeito à autocorrelação entre os erros foi efetuado o teste de Ljung-Box (LB) para 10 "lags". Este mostrou autocorrelação para alguns "lags" de resíduos e portanto a condição iii dos resíduos é violada.

5. Fazer previsão "out-of-sample", um ponto no futuro

Para finalizar, realizámos uma previsão do retorno das ações da Merk para uma unidade temporal no futuro (dia útil). Essa previsão foi de 0,0315%. Ou seja, com base no modelo definido, espera-se que o retorno no dia seguinte seja de 3,15% para estas ações.

Anexos

Anexo 1- Gráfico da correlação entre cada par de variáveis do ficheiro employees.csv:

Anexo 2 – Modelo não linear para regressão

Em alternativa ao modelo de regressão linear que define os salários foi ainda considerado outro modelo que estabelece a seguinte relação entre as variáveis.

A sua forma geral é:

$$\ln(Salary_{)i} = \beta_1 Gender_1 + \beta_2 \ln(Bonus_2) + \beta_3 SeniorManagement_3 + \beta_4 Team_4 + \beta_5 \ln(Experience_5) + \varepsilon$$

onde β_1= 0,8330; β_2= 1,1889; β_3= 0,7215; β_4 = 0,1819; β_5 = 2,4221

Conclusões obtidas com este modelo:

Este modelo foi aquele que apresentou um coeficiente de determinação mais elevado, de 0,97, ou seja, 97% da variação do salário é explicado pela variação das variáveis independentes consideradas neste modelo. Este modelo assume que:

a. Em média, os homens ganham que as mulheres (pois o coeficiente associado ao Gender (1==male) é positivo.
b. Os colaboradores em funções de SM ganham em média mais face aos restantes colaboradores, pois o coeficiente desta variável é positiva.

Foi testada a significância deste modelo (estatística F), tendo sido obtido o p-value de 0, inferior ao nível de significância definido de 0,05, o que significa a rejeição da hipótese nula. Podemos então concluir que pelo menos um dos coeficiente β_i é diferente de zero.

Relativamente aos pressupostos dos resíduos verificou-se que, tal como o modelo linear, este modelo também falhava no cumprimento desses pressupostos, o único pressuposto que apresentava melhores resultado era no valor esperado dos resíduos que era mais próximo zero. E por essa razão, apesar de este modelo apresentar um R-squared superior ao linear, foi descartado.

Anexo 3 - Gráfico com a representação da distribuição dos resíduos do modelo de regressão linear para a base de dados employees.csv: