Instituto
Politécnico de Setúbal
Escola Superior de
Tecnologia de Setúbal

Métodos
Estatísctícos

Licenciatura em Engenharia Informática
Ano letivo 2021-2022

Unidade Curricular Métodos Estatísticos

Docente José Palma

Ema Barão | 201400238
Nuno Reis | 202000753
Bernardo Teixeira | 201801954

Conteúdo

Índice de gráficos

Índice de tabelas

Índice de figuras

Resumo

No âmbito da disciplina de métodos estatísticos, após o trabalho 1 que se tratou da análise de um conjunto de dados na esfera da estatística descritiva, iremos dar seguimento com o mesmo conjunto de dados relativo a um estudo conduzido pela Coreia do Sul, de 2005 a 2018, que recolheu várias informações sobre os seus cidadãos, particularmente sobre o rendimento das famílias.

Agora, conforme solicitado no enunciado do trabalho 2, iremos apenas trabalhar os dados na esfera de ação da regressão linear. Esta fase deverá consistir na elaboração de um modelo de regressão linear resultante da relação de duas variáveis quantitativas, na análise dos seus resíduos, tal como, analisar a possibilidade de efetuar previsões plausíveis com este modelo.

Durante a análise irá ser produzido um script de R e todos os materiais daí resultantes serão aqui devidamente apresentados.

Palavras-chave: Regressão Linear; Coeficiente de correlação linear de Pearson; Diagrama de Dispersão; Análise de resíduos;

Ema Barão | **201400238**
Nuno Reis | **202000753**
Bernardo Teixeira | **201801954**

Introdução

Este trabalho foi nos solicitado no âmbito da Unidade Curricular de Métodos Estatísticos e tem como principal objetivo estudar a relação de 2 variáveis quantitativas presentes no nosso conjunto de dados.

Selecionamos as seguintes variáveis aleatórias quantitativas:

- Tamanho da companhia (no script de R denominada como company_size);
- Ano de nascimento(no script de R denominada como year_born);

Adicionamos a análise destes dados por camada da variável região (no script de R denominada como region);

Para o estudo da relação das variáveis quantitativas apresentamos:

- Qual o modelo obtido com indicação da variável independente e dependente escolhidas;
- Caso o modelo seja uma regressão linear, quantificamos a força da correlação linear:
  - Através do diagrama de dispersão:
  - Coeficiente de correlação linear de Pearson;
- Efetuamos previsões com o modelo produzido, tal como, a destrinça de previsões absurdas;
- Analisamos os resíduos obtidos pelo modelo;
- Apresentamos o estudo com os dados separados pelos níveis da variável qualitativa region;

Na conclusão deste processo será possível afirmar se o modelo de regressão linear será aplicável a este conjunto de dados e caso seja em que contornos.

Dados fornecidos e tratamento dos dados

O conjunto de dados Korea Income and Welfare apresentava as seguintes características:

1. Representa os dados que caracterizam o rendimento das famílias em determinada área geográfica e em determinado período de tempo;
2. Têm a dimensão de 92857 linhas (observações) e 14 colunas (variáveis aleatórias);
3. É composto pelas seguintes variáveis aleatórias:
    a. id;
    b. year;
    c. wave;
    d. region;
    e. income;
    f. family_member;
    g. gender;
    h. year_born;
    i. education_level;
    j. marriage;
    k. religion;
    l. occupation;
    m. company_size;
    n. reason_none_worker

Para este trabalho e apesar de trabalhar com todo o conjunto de dados vamos focar-nos na relação entre duas variáveis quantitativas são elas:

```
# variável independente -> company_size = X

# variável dependente -> year_born = Y
```

Apesar de se vir a demonstrar que têm uma fraca correlação linear, foi a melhor correlação encontrada de entre as variáveis aleatórias quantitativas constantes no conjunto de dados, para cumprir com os objetivos do trabalho fomos instruídos a perseguir este modelo.

E iremos modelar os dados com o uso da variável qualitativa region para criar níveis diferentes de analise.

No conjunto de dados Korea Income and Welfare, procedemos à remoção dos valores a Null, tal como, removemos os outliers conforme se pode observar no script de R constante na entrega do projeto.

Apresentamos de seguida a analise estatística descritiva das variáveis de maior foco neste estudo, apenas para conhecimento transversal do conjunto de dados.

Variável region- análise estatística descritiva

É uma variável qualitativa nominal, representa a área geográfica onde o estudo foi efetuado. Apresenta se originalmente de uma forma discreta como forma de codificação dos valores observados e assume a seguinte forma:

1) Seoul 2) Kyeong-gi 3) Kyoung-nam 4) Kyoung-buk 5) Chung-nam 6) Gang-won &. Chung-buk 7) Jeolla & Jeju

Apresenta se na seguinte tabela de frequências.

i	xi	ni	fi	Ni	Fi
1	Seoul	14437	0.15547562	14437	0.1554756
2	Kyeong-gi	19353	0.20841724	33790	0.3638929
3	Kyoung-nam	16154	0.17396642	49944	0.5378593
4	Kyoung-buk	12205	0.13143866	62149	0.6692980
5	Chung-nam	7843	0.08446321	69992	0.7537612
6	Gang-won & Chung-buk	6927	0.07459858	76919	0.8283597
7	Jeolla & Jeju	15938	0.17164026	92857	1.0000000

Tabela 1 - Tabela de frequências da variável region

Podemos visualizar as frequências quer relativas quer absolutas da variável nos gráficos 3,4.

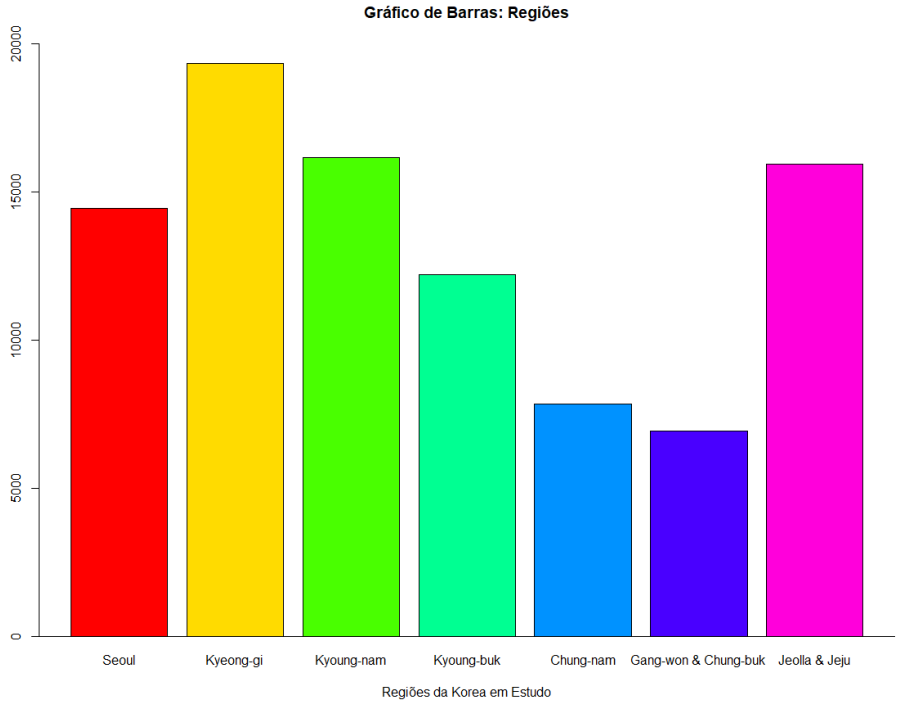


Gráfico 1-Gráfico de barras das frequências absolutas das regiões onde o estudo decorreu

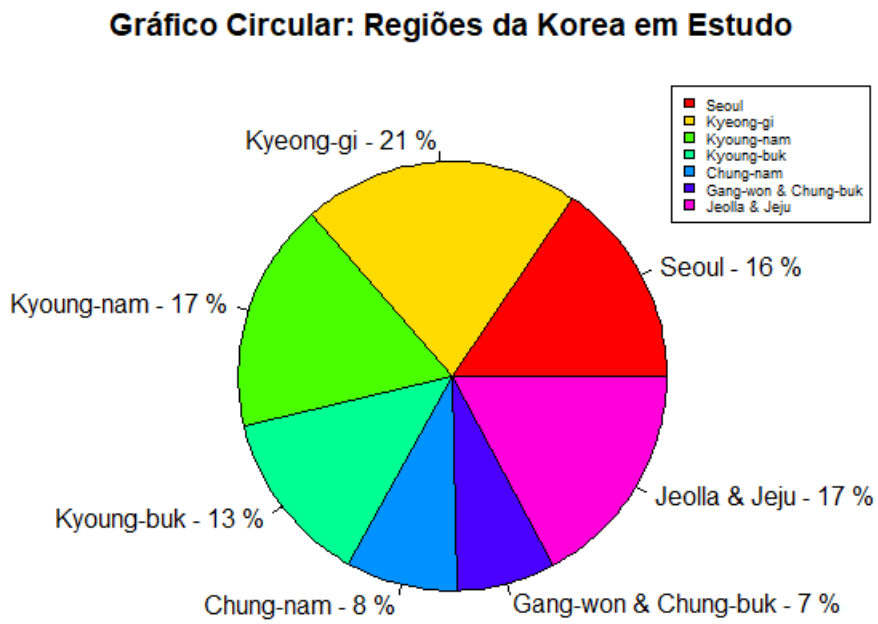


Gráfico 2 - Gráfico circular das frequências relativas das regiões onde decorreu o estudo

Podemos visualizar como se comportam as medidas de localização e de dispersão respetivamente nas tabelas 6 e 7.

Medidas de Localização					
Moda	2 ⇔ "Kyeong-gi"	Quartis		Decis	
Média	Não aplicável[1]				
Mediana	Não aplicável[2]				
		25%	Não aplicável[2]		
		50%	Não aplicável[2]		
		75%	Não aplicável[2]		
				10%	Não aplicável[2]
				20%	Não aplicável[2]
				30%	Não aplicável[2]
				40%	Não aplicável[2]
				50%	Não aplicável[2]
				60%	Não aplicável[2]
				70%	Não aplicável[2]
				80%	Não aplicável[2]
				90%	Não aplicável[2]

Tabela 2 - Medidas de localização da variável region

[1] Não aplicável quando a variável é qualitativa.

Medidas de dispersão	
Variância	Não aplicável[2]
Desvio Padrão	Não aplicável[3]
Amplitude Total	Não aplicável[3]
Amplitude Interquartil	Não aplicável[3]

Tabela 3 - Medidas de dispersão da variável region

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1= 0.3701752). Podemos assim afirmar que a assimetria é positiva pois, b1>0.

Relativamente aos valores de curtose, podemos afirmar que a variável apresenta uma curva platicúrtica ou achatada dado que b2<0, (b2= -1.172695).

[2] Não aplicável quando a variável é qualitativa.

Variável year_born - análise estatística descritiva

É uma variável quantitativa discreta, representa o ano do nascimento do individuo observado. Por apresentar tantos níveis(90), foi agrupada em classes.

Aplicando se a regra de Sturges consegui-o apurar 17 classes (K=17), as classes são fechadas à direita e a amplitude de cada classe é de 5.411765, pois h= 5.411765. Apresenta a seguinte tabela de frequências.

Apresenta se na seguinte tabela de frequências.

	classes	ni	fi	Ni	Fi
1	[1910,1915]	37	0.0004	37	0.0004
2	(1915,1921]	261	0.0028	298	0.0032
3	(1921,1926]	1766	0.0190	2064	0.0222
4	(1926,1932]	4647	0.0500	6711	0.0723
5	(1932,1937]	11895	0.1281	18606	0.2004
6	(1937,1942]	12397	0.1335	31003	0.3339
7	(1942,1948]	9058	0.0975	40061	0.4314
8	(1948,1953]	8541	0.0920	48602	0.5234
9	(1953,1959]	7807	0.0841	56409	0.6075
10	(1959,1964]	10657	0.1148	67066	0.7223
11	(1964,1970]	7972	0.0859	75038	0.8081
12	(1970,1975]	8218	0.0885	83256	0.8966
13	(1975,1980]	6225	0.0670	89481	0.9636
14	(1980,1986]	2236	0.0241	91717	0.9877
15	(1986,1991]	873	0.0094	92590	0.9971
16	(1991,1997]	244	0.0026	92834	0.9998
17	(1997,2002]	23	0.0002	92857	1.0000

Tabela 4- Tabela de frequências da variável year_born

No histograma do gráfico 11 podemos visualizar a distribuição das observações desta variável.

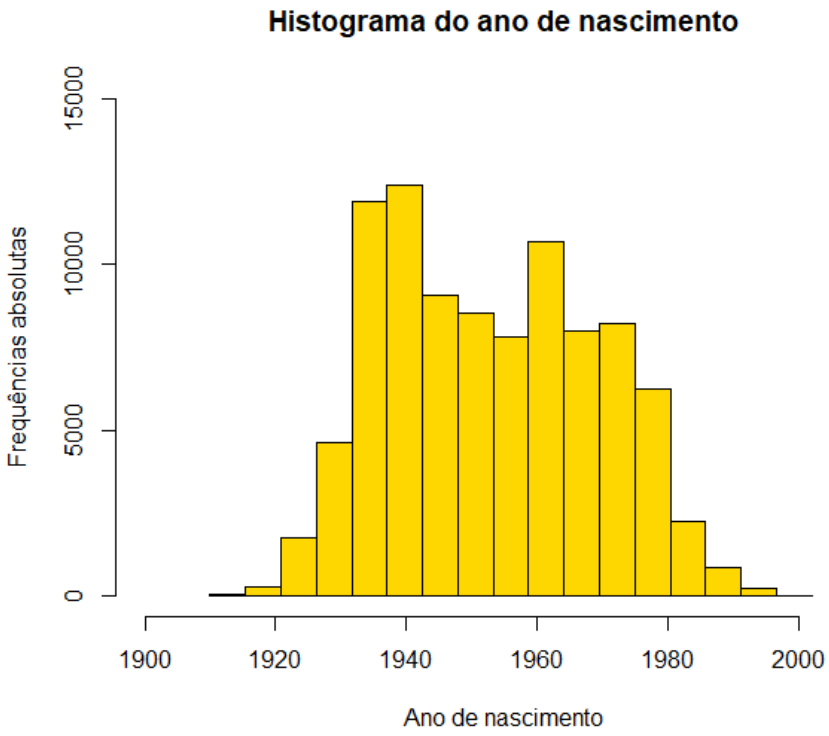


Gráfico 3- Histograma das frequências absolutas da variável year_born

Podemos visualizar como se comportam as medidas de localização e de dispersão respetivamente nas tabelas 18 e 19.

Medidas de Localização					
Moda	1942	Quartis		Decis	
Média	1952.957				
Mediana	1952				
		25%	1939		
		50%	1952		
		75%	1966		
				10%	1933
				20%	1937
				30%	1941
				40%	1946
				50%	1952
				60%	1958
				70%	1963
				80%	1969
				90%	1975

Tabela 5 - Medidas de localização da variável year_born

Medidas de dispersão	
Variância	256.1941
Desvio Padrão	16.00607
Amplitude Total	92
Amplitude Interquartil	27

Tabela 6 - Medidas de dispersão da variável year_born

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1= 0. 0.1768795). Podemos agora afirmar que a assimetria é positiva pois, b1>0.

Relativamente aos valores de curtose, podemos afirmar que a variável apresenta uma curva platicúrtica ou achatada dado que b2<0, (b2= -0.9478836).

Podemos verificar que os quartis da caixa de bigodes tem uma concentração de dados muito uniforme.

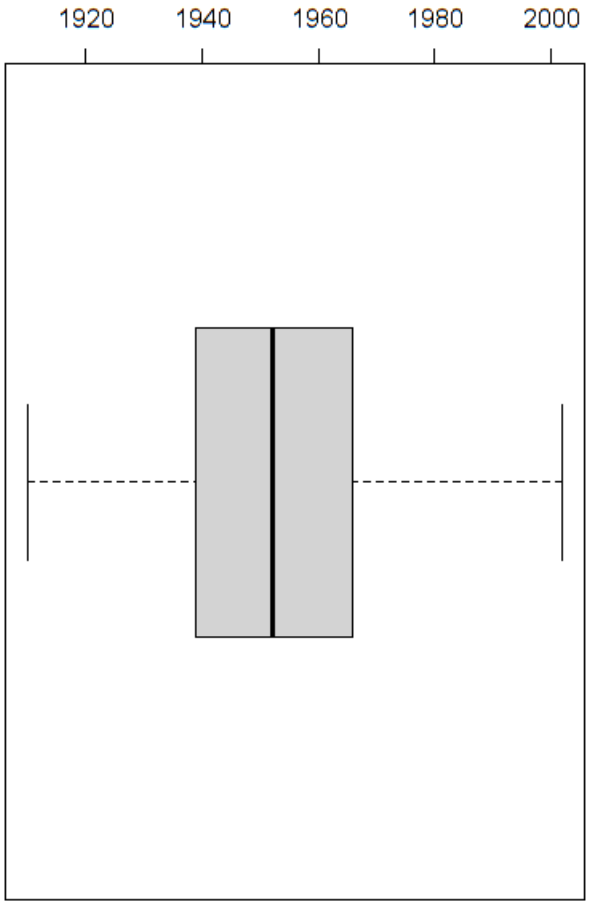


Figura 1- Caixa de bigodes da variável year_born

Variável company_size - análise estatística descritiva

É uma variável quantitativa discreta, representa o tamanho de uma companhia através do número de funcionários. O estudo incidiu em empresas com um intervalo do número de funcionários de 1 a 99.

Apresenta a seguinte tabela de frequências.

i	xi	ni	fi	Ni	Fi
1	1	28319	0.478304930	28319	0.4783049
2	2	5612	0.094786089	33931	0.5730910
3	3	6497	0.109733646	40428	0.6828247
4	4	2669	0.045079129	43097	0.7279038
5	5	1860	0.031415204	44957	0.7593190
6	6	1346	0.022733798	46303	0.7820528
7	7	3478	0.058743054	49781	0.8407959
8	8	1031	0.017413482	50812	0.8582093
9	9	1097	0.018528215	51909	0.8767375
10	10	6905	0.116624723	58814	0.9933623
11	11	393	0.006637729	59207	1.0000000

Tabela 7- Tabela de frequências da variável company_size

Podemos visualizar as frequências quer relativas quer absolutas da variável nos gráficos 14 e 15.

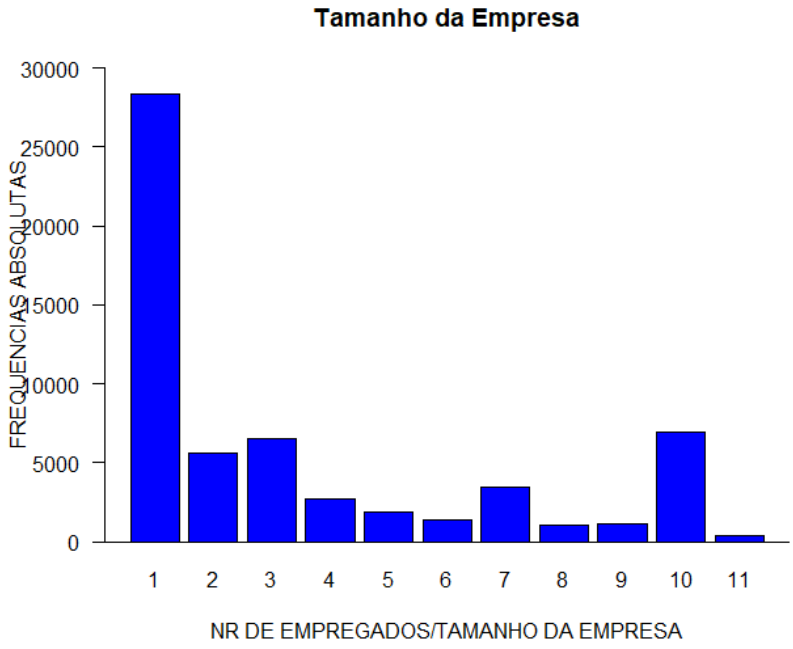


Gráfico 4- Gráfico de barras das frequências absolutas da variável company_size

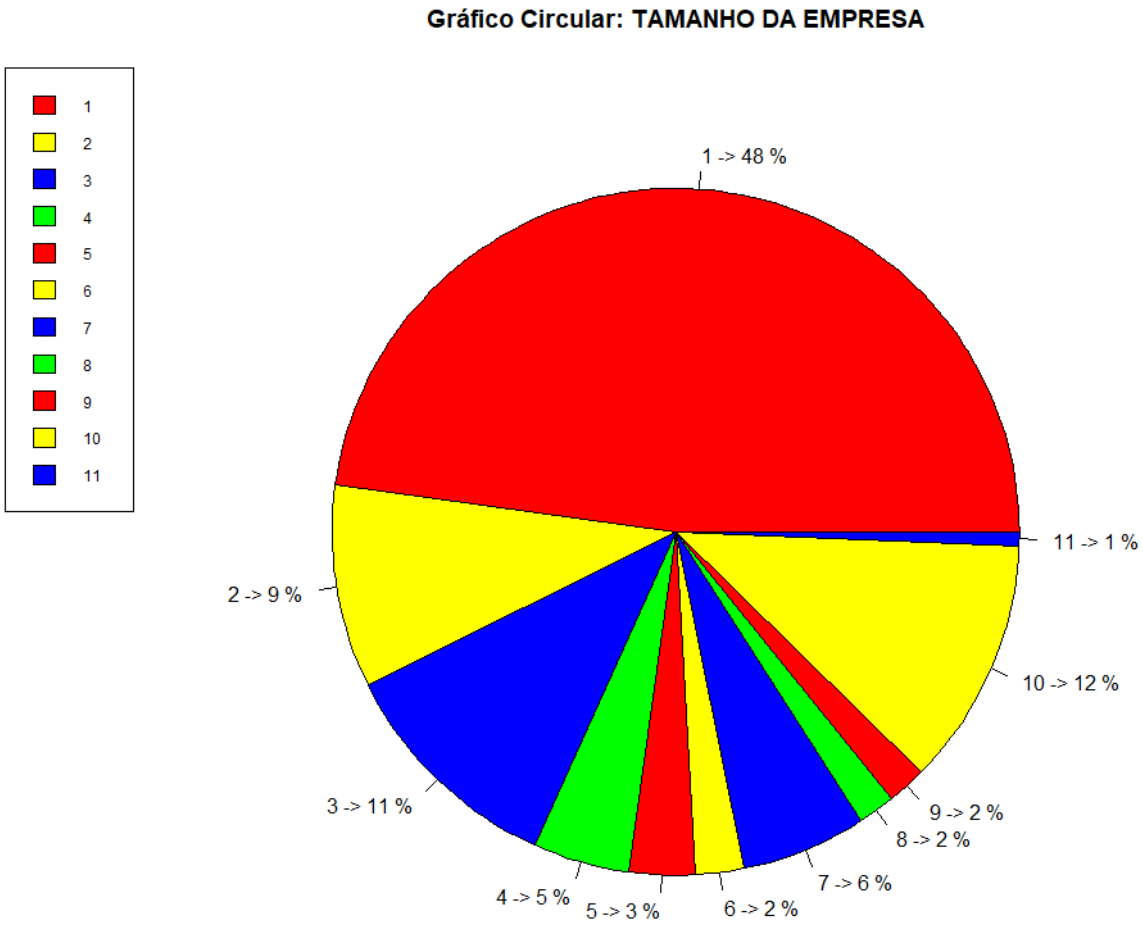


Gráfico 5-Gráfico circular das frequências relativas da variável company_size

Podemos visualizar como se comportam as medidas de dispersão e de localização respetivamente nas tabelas 24 e 25.

Medidas de dispersão	
Variância	10.38252
Desvio Padrão	3.222192
Amplitude Total	10
Amplitude Interquartil	4

Tabela 8- Medidas de dispersão da variável company_size

Medidas de Localização					
Moda	1	Quartis		Decis	
Média	3.427399				
Mediana	2				
		25%	1		
		50%	2		
		75%	5		
				10%	1
				20%	1
				30%	1
				40%	1
				50%	2
				60%	3
				70%	4
				80%	7
				90%	10

Tabela 9 - Medidas de localização da variável company_size

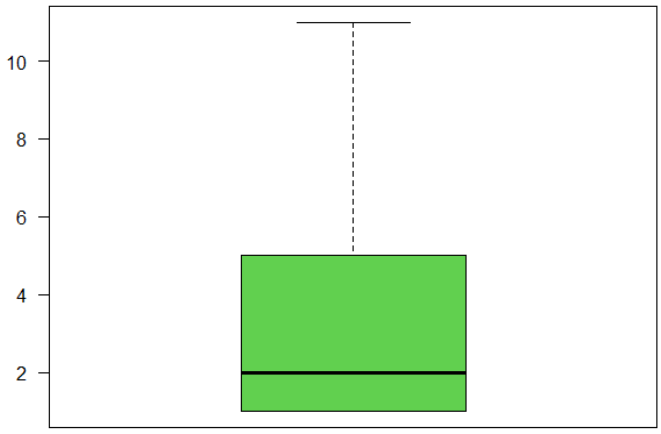


Figura 2-Caixa de bigodes variável company_size

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1= 1.104122). Podemos assim afirmar que a assimetria é positiva pois, b1>0.

Relativamente aos valores de curtose verificamos que a variável apresenta uma curva platicúrtica ou achatada dado que b2<0, (b2= -0.284578).

Na figura 6 podemos visualizar a caixa de bigodes da variável company_size e verificamos o estudo focou se em empresas de pequena dimensão pois a mediana situa se em empresas com 2 funcionários .

Regressão Linear

Para atingir este objetivo vamos investigar a presença ou ausência de relação linear entre as duas variáveis com todo o conjunto de dados.

Após diversos ensaios, onde foram exploradas as combinações entre 5 variáveis quantitativas tomadas de 2 a 2, para verificar qual seria a melhor combinação para uma correlação linear mais forte, chegou se há conclusão de que as melhores variáveis aleatórias quantitativas em estudo são:

variável independente -> company_size = X

variável dependente -> year_born = Y

Esta relação estuda se o do tamanho da companhia está a ser alterada pelo ano de nascimento dos funcionários.

Diagrama de dispersão

Pela análise do diagrama de dispersão não se vê uma relação linear entre as variáveis, pois não é possível imaginar uma reta nem com declive negativo nem com declive positivo a passar pela nuvem de pontos

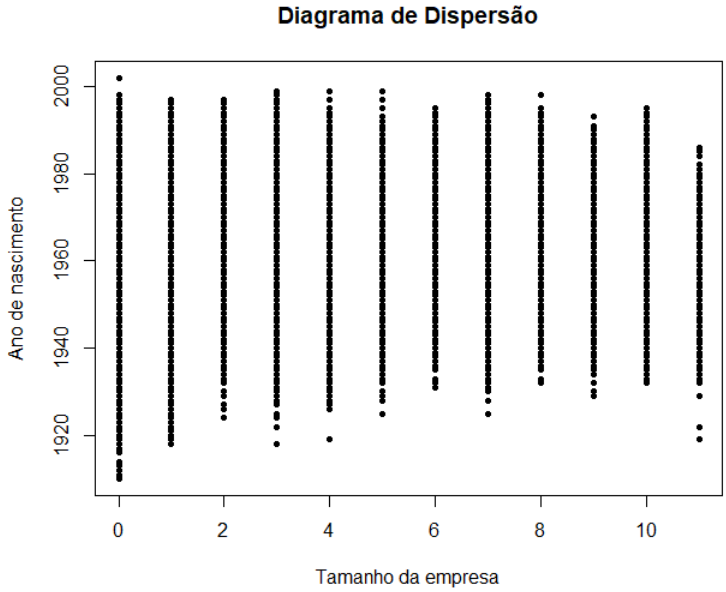


Gráfico 6- Diagrama de dispersão com todo o conjunto de dados

Coeficiente de correlação linear de Pearson

Coeficiente confirma o que vimos no diagrama de dispersão, a correlação linear é muito fraca.

O rxy= 0.5082532 não se encontra entre -1<rxy<-0.8 nem entre 0.8<rxy<1, onde poderíamos considerar um coeficiente de correlação linear muito forte.

Sabemos que seria a melhor opção abandonar este modelo, contudo no âmbito deste trabalho vamos mantê-lo e considera lo como válido para cumprir os objetivos que nos foram propostos.

Reta de regressão linear

Quando a correlação linear é forte, podemos inferir o valor de uma se conhecer mos a outra. A reta que atravessa a nuvem de pontos conforme podemos verificar no gráfico X divide o diagrama de dispersão em dois grupos idênticos.

A reta de regressão passa pelo ponto cujas coordenadas são, respetivamente, as médias das variáveis em estudo, ou seja, o centro de gravidade da nuvem de pontos (ponto de coordenadas (x, y)).

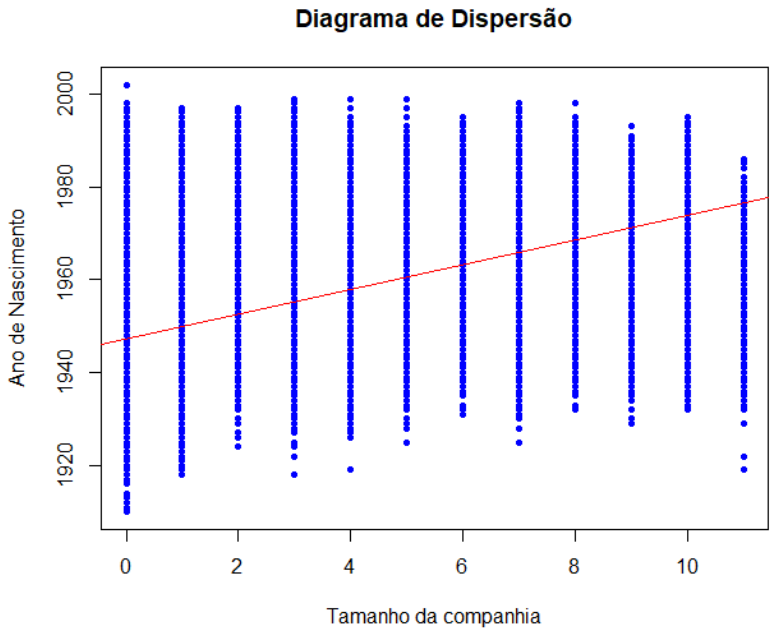


Gráfico 7- Diagrama de dispersão com a reta da regressão linear

No gráfico 10 podemos visualizar a reta regressão linear que apresenta os valores de interceção para a variável independente, company_size ⇔ X=2.663 e para a variável dependente, year_born ⇔ Y=1947.138.

Resíduos

Ao analisarmos os resíduos podemos concluir a qualidade do nosso modelo. Vamos analisar a diferença entre os valores observados e os valores ajustados.

Começamos por analisar o diagrama de dispersão dos resíduos no gráfico 8.

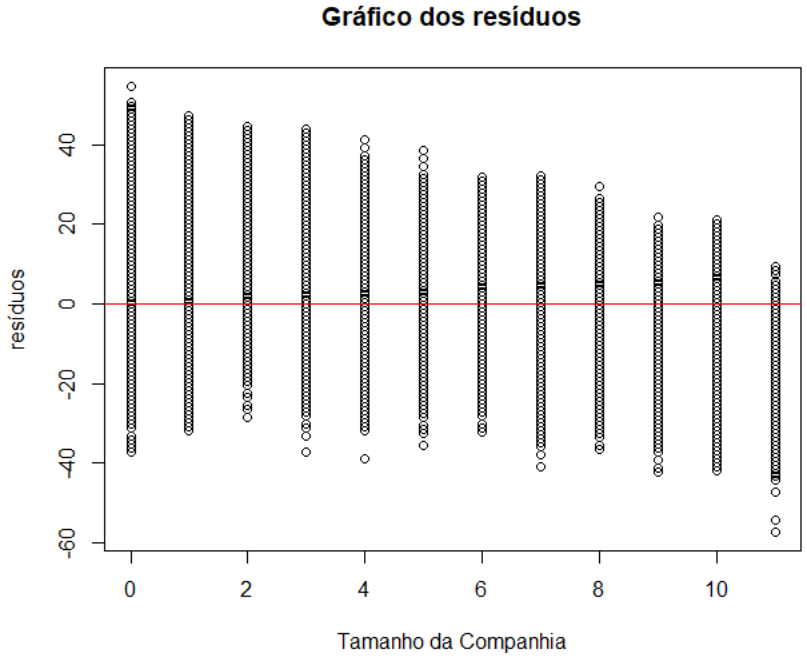


Gráfico 8 - Gráfico de resíduos

Os resíduos são muito grandes (entre mais 40 e -40), têm um padrão bem definido e constante e isto é sintoma que o modelo ajustado não é bom. Este padrão indica que os resíduos não são independentes.

Isto apenas vêm reforçar o que já havíamos concluído anteriormente, que este modelo não têm uma correlação linear forte e apenas não abandonamos e escolhemos outro modelo mais ajustado para correta conclusão dos objetivos propostos neste trabalho.

Previsões do modelo

O modelo construído permite previsão para y com base em valores de x dentro do intervalo analisado ou para valores muito próximos, se não se aplicar este caso assumimos que as previsões são absurdas, pois não temos garantia que a relação linear se mantém.

Vamos prever o quando o tamanho da companhia é de 8 funcionários e de 99 funcionários, iremos obter a previsão do ano de nascimento.

Ano de nascimento (year_born) =Y -> prever -> variável dependente

Tamanho da companhia(company_size) =X -> variável independente

Na previsão para 8 funcionários, o resultado da previsão do ano de nascimento é igual a 1968.439. O valor de 8 encontra-se no intervalo estudado pelo modelo, contudo a correlação linear de Pearson é fraca e não nos apresenta garantias de um modelo bem ajustado. Apenas por este motivo não consideramos esta previsão válida.

Na previsão para 99 funcionários, o resultado da previsão do ano de nascimento é igual a 2210.745. O valor de 99 encontra-se no muito longe do intervalo estudado pelo modelo, logo é absurdo e não pode ser considerado.

Modelação por níveis da variável region

Iremos agora proceder à modelação de dados mediante a divisão de níveis da variável qualitativa nominal region.

Esta representa a área geográfica onde o estudo foi efetuado, vamos verificar se o nosso estudo apresenta variações mediante a área geográfica onde incide.

Apresenta se originalmente de uma forma discreta como forma de codificação dos valores observados e assume a seguinte forma:

1) Seoul 2) Kyeong-gi 3) Kyoung-nam 4) Kyoung-buk 5) Chung-nam 6) Gang-won &. Chung-buk 7) Jeolla & Jeju

Tabela com os dados de estatística descritiva da variável region(localização)

Code	Region	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	Seoul	0.000	0.000	1.000	2.454	3.000	11.000
2	Kyeong-gi	0.000	0.000	1.000	2.456	3.000	11.000
3	Kyoung-nam	0.000	0.000	1.000	2.153	3.000	11.000
4	Kyoung-buk	0.000	0.000	1.000	1.705	1.000	11.000
5	Chung-nam	0.000	0.000	1.000	2.347	3.000	11.000
6	Gang-won &. Chung-buk	0.000	0.000	1.000	2.287	3.000	11.000
7	Jeolla & Jeju	0.000	0.000	1.000	1.89	2.000	11.000

Tabela 10- Tabela com os dados de estatística descritiva da variável region(localização)

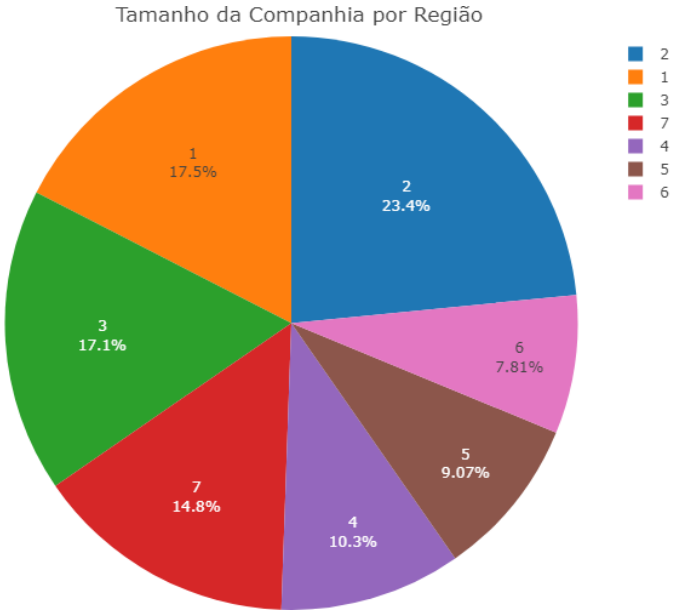


Gráfico 9- Gráfico circular sobre a informação dos níveis da variável por região

1) Seoul

A modelação apenas para os dados de Seoul para as variáveis quantiavas em estudo apresenta o seguinte panorama.

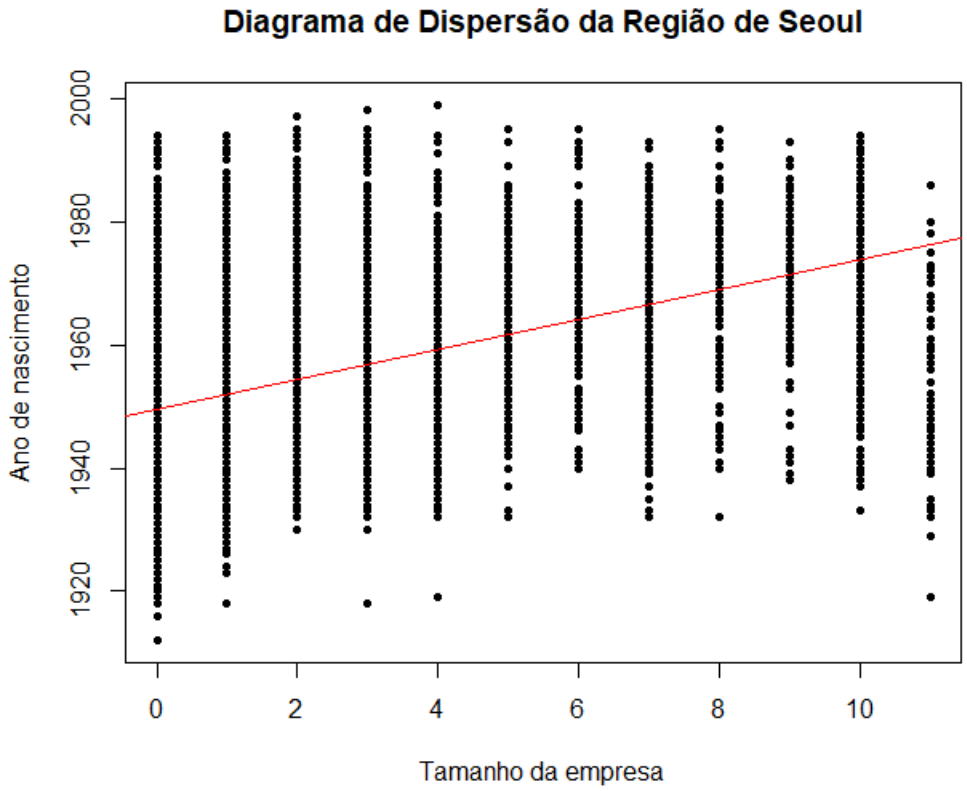


Gráfico 10- - Diagrama de Dispersão para a Região de Seoul

No diagrama 10 podemos visualizar a reta regressão linear que apresenta os valores de interceção para a variável independente, company_size ⇔ X=2.663 e para a variável dependente, year_born ⇔ Y=1947.138.

O coeficiente linear de Pearson confirma o que vimos no diagrama de dispersão, a correlação linear é muito fraca , não se verifica de forma clara a reta que atravessa a nuvem de pontos.

O rxy= 0.4936752 não se encontra entre -1<rxy<-0.8 nem entre 0.8<rxy<1, onde poderíamos considerar um coeficiente de correlação linear muito forte.

O modelo não é adequado. Contudo iremos dar continuidade ao modelo para o cálculo de resíduos.

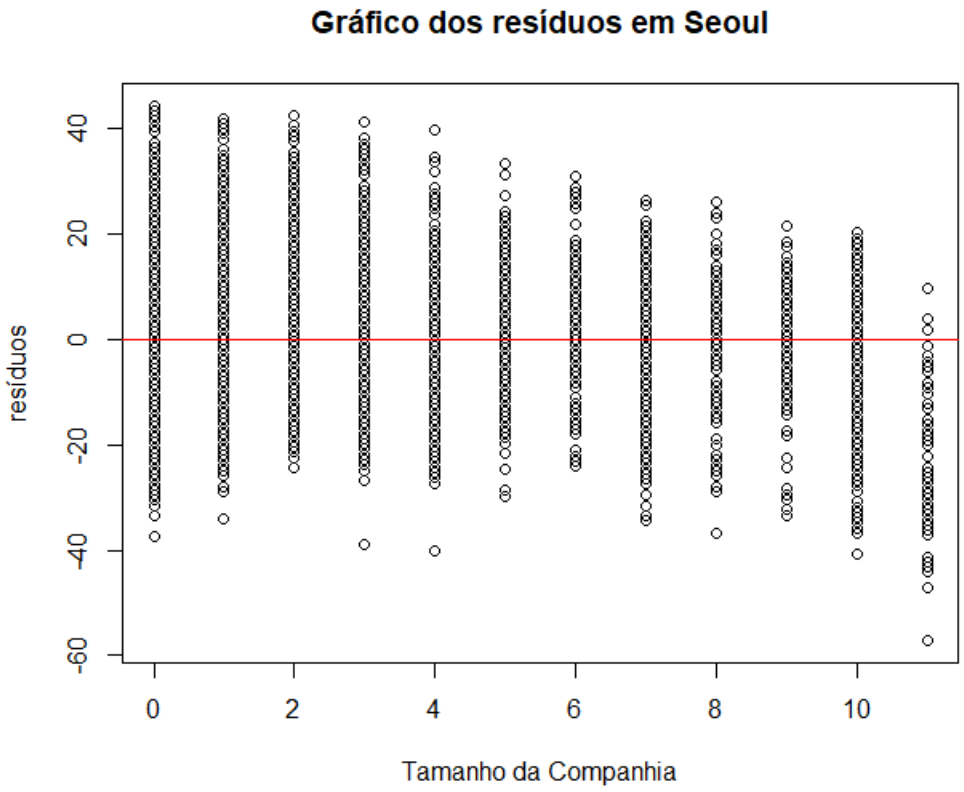


Gráfico 11-Resíduos em Seoul

Os resíduos são muito grandes(varia entre -40 e +40 aproximadamente) o padrão não é aleatório está muito bem definido e constante e isso é uma evidência de que o modelo ajustado não é bom. Este padrão indica que os resíduos não são independentes

2) Kyeong-gi

A modelação apenas para os dados de Kyeong-gi para as variáveis quantiavas em estudo apresenta o seguinte panorama.

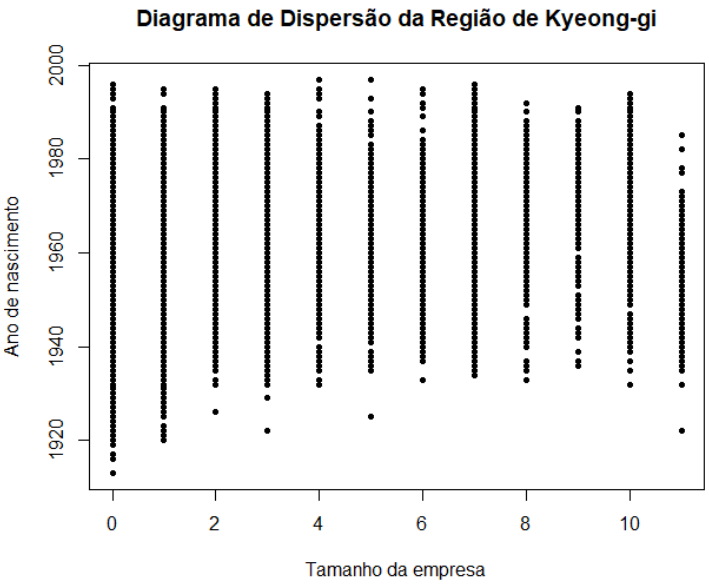


Gráfico 12- Diagrama de dispersão da region de Kyeong-gi

O coeficiente linear de Pearson confirma o que vimos no diagrama de dispersão, a correlação linear é muito fraca , não se verifica de forma clara a reta que atravessa a nuvem de pontos.

O rxy= 0.468182 não se encontra entre -1<rxy<-0.8 nem entre 0.8<rxy<1, onde poderíamos considerar um coeficiente de correlação linear muito forte.

O modelo não é adequado. Contudo iremos dar continuidade ao modelo para o cálculo de resíduos.

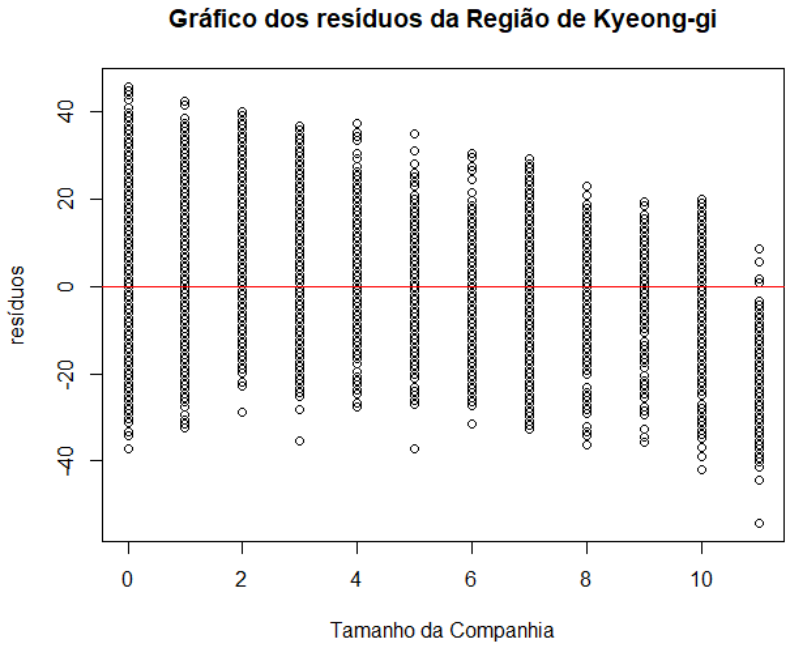


Gráfico 13- Resíduos da Região de Kyeong-gi

Os resíduos são muito grandes(variam entre -40 e +40 aproximadamente) o padrão não é aleatório está muito bem definido e constante e isso é uma evidência de que o modelo ajustado não é bom. Este padrão indica que os resíduos não são independentes

3) Kyoung-nam

A modelação apenas para os dados de Kyoung-nam para as variáveis quantiavas em estudo apresenta o seguinte panorama.

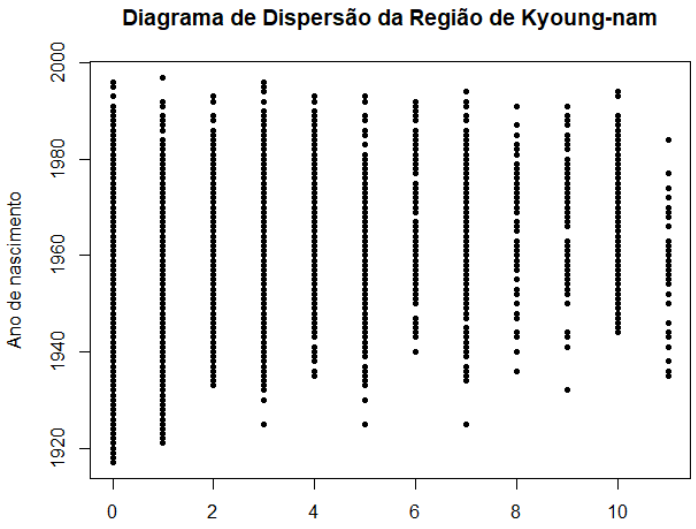


Gráfico 14 - - Diagrama de dispersão da região de Kyoung-nam

O coeficiente linear de Pearson confirma o que vimos no diagrama de dispersão, a correlação linear é muito fraca , não se verifica de forma clara a reta que atravessa a nuvem de pontos.

O rxy= 0.5277147 não se encontra entre -1<rxy<-0.8 nem entre 0.8<rxy<1, onde poderíamos considerar um coeficiente de correlação linear muito forte.

O modelo não é adequado. Contudo iremos dar continuidade ao modelo para o cálculo de resíduos.

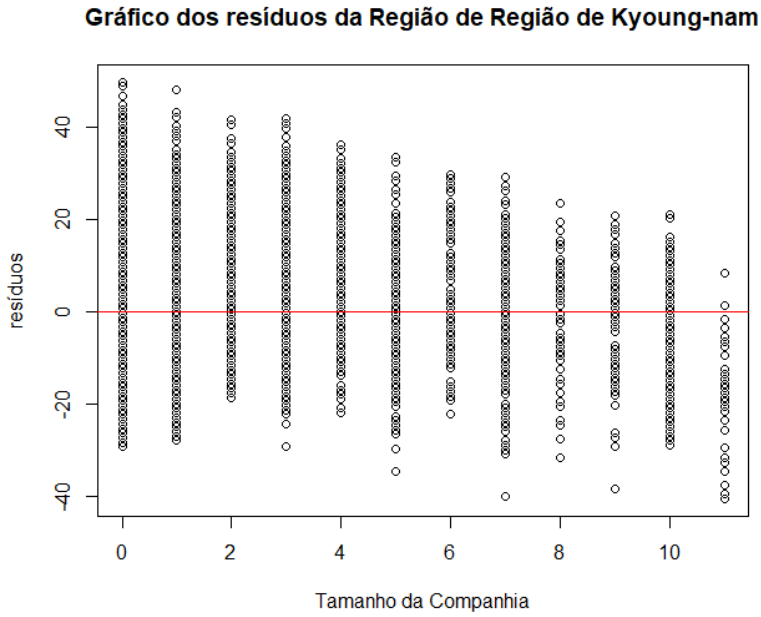


Gráfico 15 - Resíduos da Região de Kyoung-nam

Os resíduos são muito grandes (variam entre -40 e +40 aproximadamente) o padrão não é aleatório está muito bem definido e constante e isso é uma evidência de que o modelo ajustado não é bom. Este padrão indica que os resíduos não são independentes

4) Kyoung-buk

A modelação apenas para os dados de Kyoung-buk para as variáveis quantiavas em estudo apresenta o seguinte panorama.

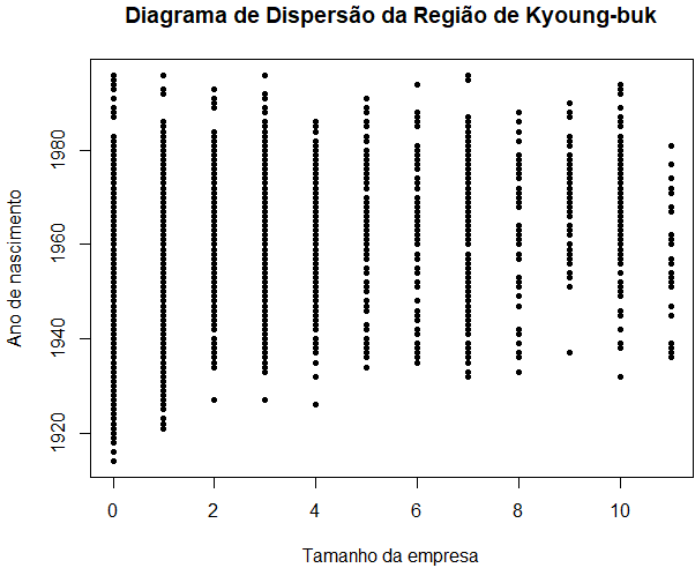


Gráfico 16- Diagrama de dispersão da região de Kyoung-buk

O coeficiente linear de Pearson confirma o que vimos no diagrama de dispersão, a correlação linear é muito fraca , não se verifica de forma clara a reta que atravessa a nuvem de pontos.

O rxy= 0.5146754 não se encontra entre -1<rxy<-0.8 nem entre 0.8<rxy<1, onde poderíamos considerar um coeficiente de correlação linear muito forte.

O modelo não é adequado. Contudo iremos dar continuidade ao modelo para o cálculo de resíduos.

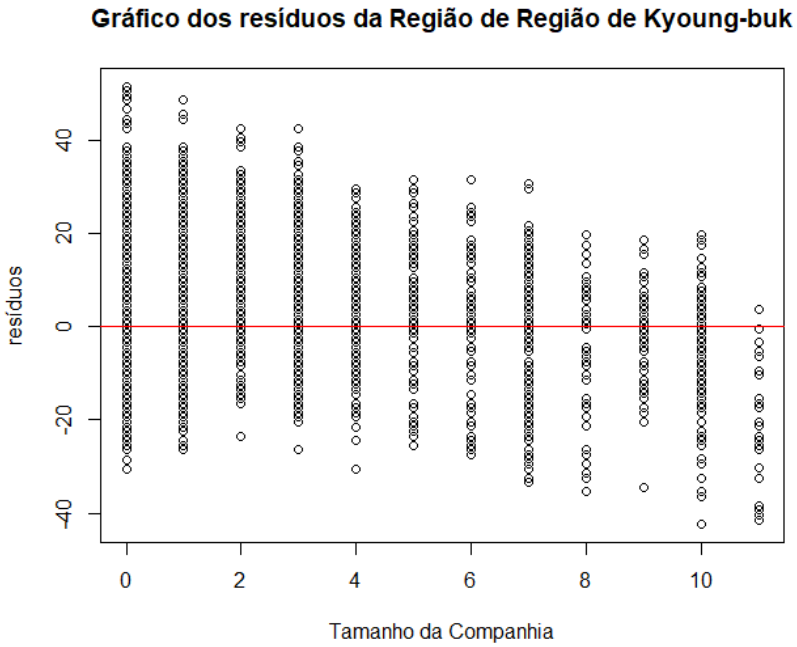


Gráfico 17- Resíduos da região de Kyoung-buk

Os resíduos são muito grandes(variam entre -40 e +40 aproximadamente) o padrão não é aleatório está muito bem definido e constante e isso é uma evidência de que o modelo ajustado não é bom. Este padrão indica que os resíduos não são independentes

5) Chung-nam

A modelação apenas para os dados de Chung-nam para as variáveis quantiavas em estudo apresenta o seguinte panorama.

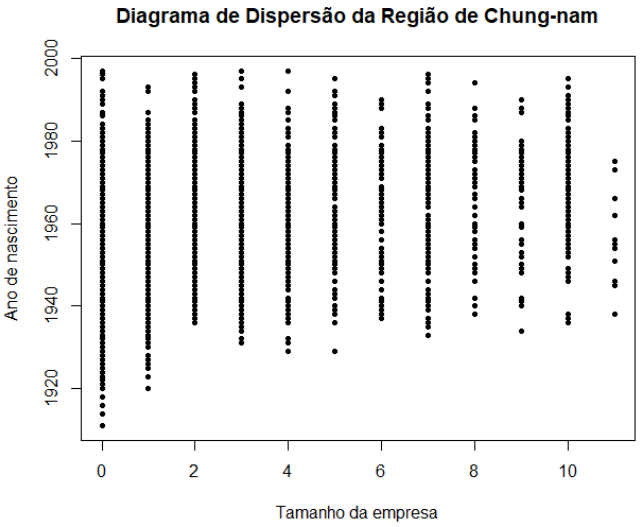


Gráfico 18 - Diagrama de dispersão da região de Chung-nam

O coeficiente linear de Pearson confirma o que vimos no diagrama de dispersão, a correlação linear é muito fraca , não se verifica de forma clara a reta que atravessa a nuvem de pontos.

O rxy= 0.5222542 não se encontra entre -1<rxy<-0.8 nem entre 0.8<rxy<1, onde poderíamos considerar um coeficiente de correlação linear muito forte.

O modelo não é adequado. Contudo iremos dar continuidade ao modelo para o cálculo de resíduos.

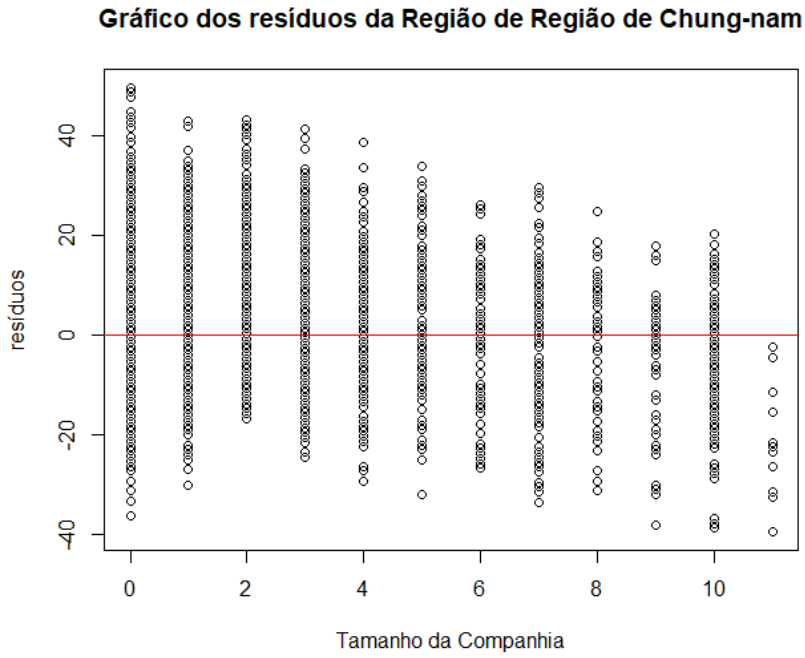


Gráfico 19 - Residuos da região de Chung-nam

Os resíduos são muito grandes(variam entre -40 e +40 aproximadamente) o padrão não é aleatório está muito bem definido e constante e isso é uma evidência de que o modelo ajustado não é bom. Este padrão indica que os resíduos não são independentes

6) Gang-won &. Chung-buk

A modelação apenas para os dados de Gang-won &. Chung-buk para as variáveis quantiavas em estudo apresenta o seguinte panorama.

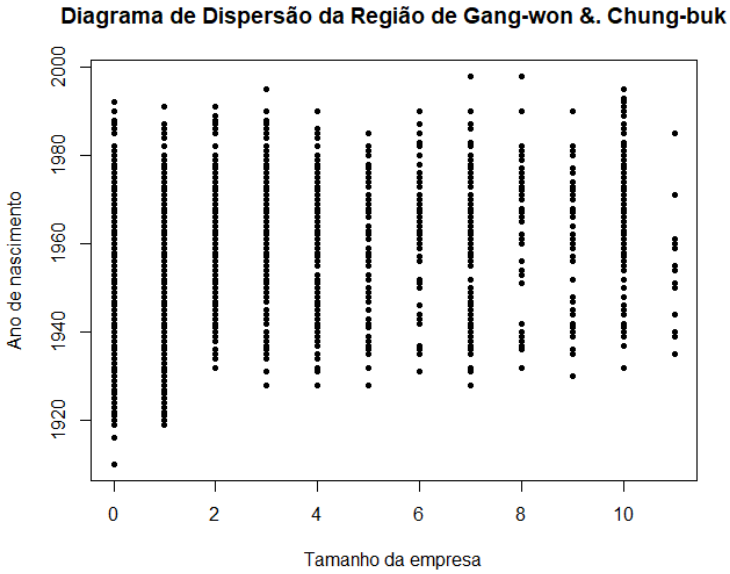


Gráfico 20 - Diagrama de dispersão da região de Gang-won &. Chung-buk

O coeficiente linear de Pearson confirma o que vimos no diagrama de dispersão, a correlação linear é muito fraca , não se verifica de forma clara a reta que atravessa a nuvem de pontos.

O rxy= 0.5228782 não se encontra entre -1<rxy<-0.8 nem entre 0.8<rxy<1, onde poderíamos considerar um coeficiente de correlação linear muito forte.

O não é adequado. Contudo iremos dar continuidade ao modelo para o cálculo de resíduos.

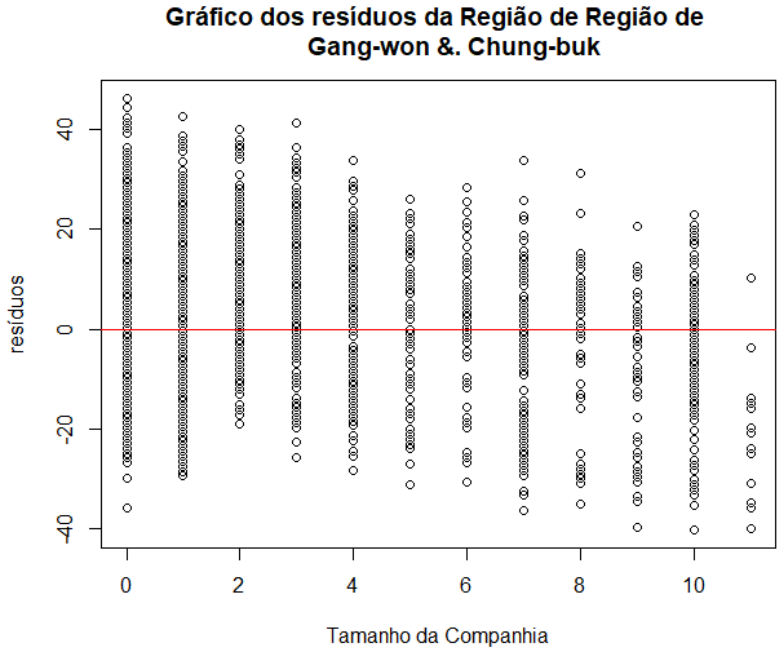


Gráfico 21 - Residuos da região de Gang-won &. Chung-buk

Os resíduos são muito grandes(variam entre -40 e +40 aproximadamente) o padrão não é aleatório está muito bem definido e constante e isso é uma evidência de que o modelo ajustado não é bom. Este padrão indica que os resíduos não são independentes

7) Jeolla & Jeju

A modelação apenas para os dados de Jeolla & Jeju para as variáveis quantiavas em estudo apresenta o seguinte panorama.

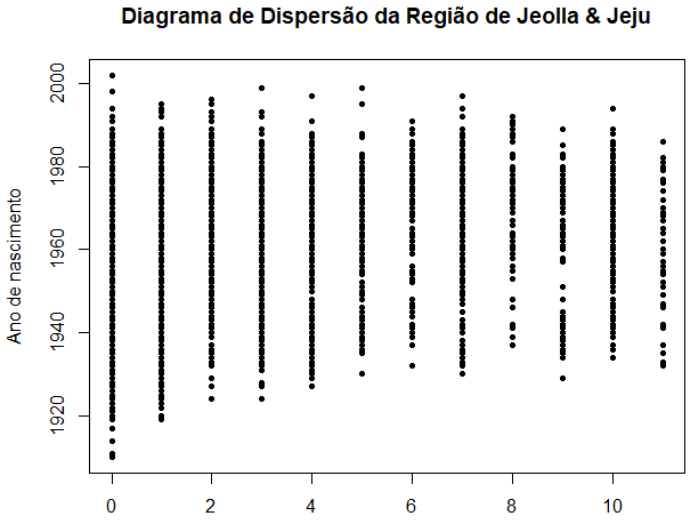
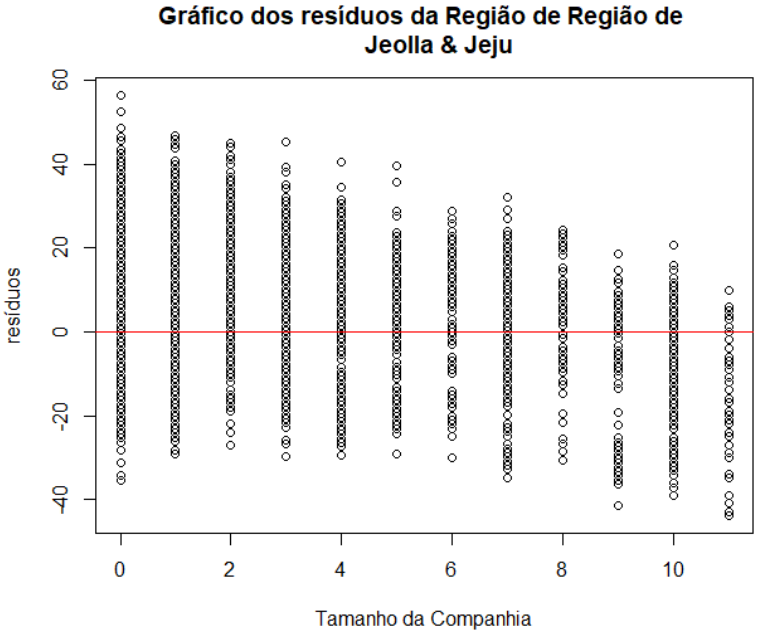


Gráfico 22- Diagrama de dispersão para a região de Jeolla & Jeju

O coeficiente linear de Pearson confirma o que vimos no diagrama de dispersão, a correlação linear é muito fraca , não se verifica de forma clara a reta que atravessa a nuvem de pontos.

O rxy= 0.5082045 não se encontra entre -1<rxy<-0.8 nem entre 0.8<rxy<1, onde poderíamos considerar um coeficiente de correlação linear muito forte.

O modelo não é adequado. Contudo iremos dar continuidade ao modelo para o cálculo de resíduos.



Os resíduos são muito grandes(variam entre -40 e +40 aproximadamente) o padrão não é aleatório está muito bem definido e constante e isso é uma evidência de que o modelo ajustado não é bom. Este padrão indica que os resíduos não são independentes

Conclusões

Neste projeto abordamos o conjunto de dados intitulado '*Korea Income and Welfare'*, com o objetivo de utilizar a regressão linear simples para modelar uma variável em função de outra.

Com o auxílio do RStudio e através da linguagem de R produzimos um script que nos permitiu modelar os dados.

Cumprimos todos os objetivos a que nos tínhamos proposto, mas infelizmente não foi possível trabalhar com um modelo com uma correlação linear forte. Nesse caso teríamos obtido resultados mais motivadores. No entanto criamos um modelo de regressão linear, efetuamos a analise da relação entre as variáveis em estudo, analisamos os resíduos, efetuamos previsões e elaboramos a documentação de apoio na medida que o estudo foi sendo elaborado.

Este projeto teve uma importância valiosa na aquisição de conhecimentos na esfera dos modelos de regressão linear e da linguagem de R, pois obrigou todos os elementos deste grupo a pesquisar e analisar e aperfeiçoar técnicas fundamentais nesta área.

Referências bibliográficas

- Departamento de Matemática Escola Superior de Tecnologia de Setúbal. Capítulo 1 – Estatística Descritiva. 2021-2022. Materiais de apoio. Disponível em: <https://moodle.ips.pt/2122/mod/resource/view.php?id=3386 >. Acesso em: 10/04/2022
- Departamento de Matemática Escola Superior de Tecnologia de Setúbal. Capítulo 2 – Regressão Linear Simples. 2021-2022. Materiais de apoio. Disponível em: <https://moodle.ips.pt/2122/mod/resource/view.php?id=3386 >. Acesso em: 07/05/2022