

**PÓS-GRADUAÇÃO IESB**

Curso de Pós-Graduação em Ciência de Dado

*Métodos Estatísticos de Apoio à Decisão II*

*Prof: Aline Riquetti*

*Equipe:*

**Trabalho de Métodos Estatísticos de Apoio à Decisão II**

**1. Introdução**

Analisar os gastos dos parlamentares brasileiros, do Senado Federal, com foco na CEAPS - Cota para o Exercício da Atividade Parlamentar dos Senadores -, sua lógica de justificação, a distribuição por Unidades da Federação, por Partidos Políticos, a identificação de aparentes desvios de finalidade e, também, sua interação com outras informações como participação em comissões, relatorias e matérias apresentadas.

O tema se torna particularmente relevante uma vez que pesquisas apontam o Parlamento Brasileiro como um dos mais caros do mundo, ao mesmo tempo que a sociedade brasileira se torna sequiosa de informações, exigindo transparência por parte do poder público. Portanto, entender como a utilização da CEAPS se dá, qual a sua motivação, e como a sociedade civil pode fiscalizar e estabelecer limites para esses gastos corrobora com o desejo da sociedade.

Qual técnica foi escolhida;

**2. Objetivos**

Em termos de objetivo geral, o propósito é o montante dos gastos dos parlamentares brasileiros, mais especificamente os senadores.

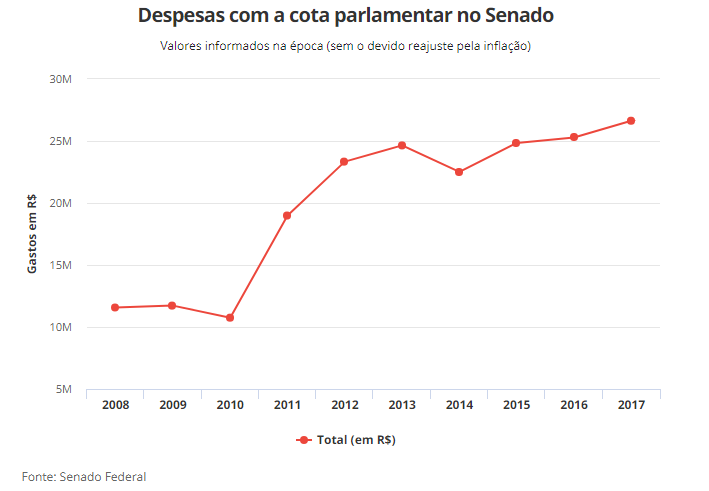
Quanto aos objetivos específicos, pretendemos analisar, confrontando esses gastos com a estrutura partidária, a estrutura federativa, bem como medir a produtividade a partir de critérios definidos considerando-se variáveis como a participação em comissões, matérias apresentadas e o exercício da relatoria.

Geral: Critérios de sucesso

Específicos: subetapas para atingir o objetivo final

**3. Referencial teórico:**

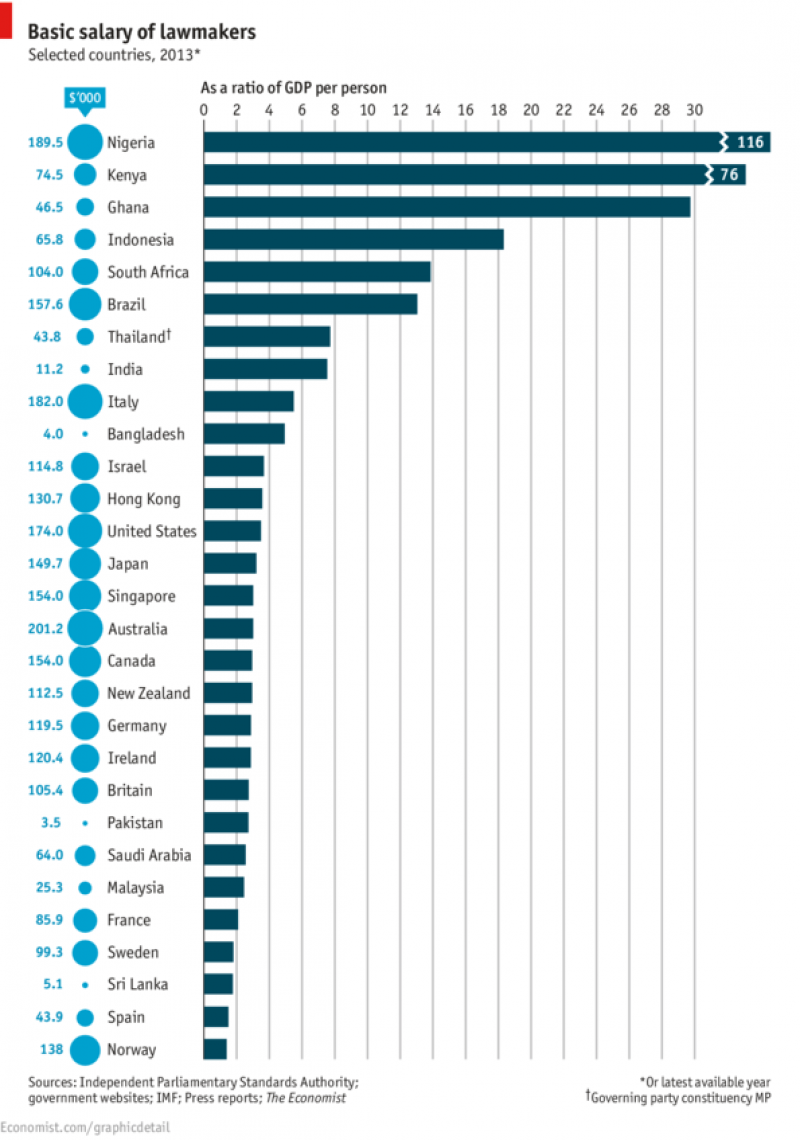
Os gastos do Senado Federal com a chamada cota parlamentar cresceram sensivelmente ao longo dos anos. As diversas mídias noticiam com frequência que viagens de jatinhos, flats de luxo, refeições a preços exorbitantes e outras despesas têm colocado nossos senadores como um dos parlamentares mais caros do mundo.



**Figura** 1 **- Despesas com a cota parlamentar no Senado.**   
Fonte: [https://g1.globo.com/politica/noticia/viagens-de-jatinho-flat-de-luxo-refeicao-de-r-1000-gastos-de-senadores-com-cota-parlamentar-chegam-a-r-266-milhoes-em-2017.ghtml. Acessado em 16/05/2018](https://g1.globo.com/politica/noticia/viagens-de-jatinho-flat-de-luxo-refeicao-de-r-1000-gastos-de-senadores-com-cota-parlamentar-chegam-a-r-266-milhoes-em-2017.ghtml.%20Acessado%20em%2016/05/2018).

Segundo DOMINGUES (2015), um estudo da Organização Transparência Brasil, publicado em 2011, revelou que os parlamentares brasileiros estão entre os mais caros do mundo, estimando um custo anual de aproximadamente R$ 33 milhões para cada senador.

A revista “The Economist” (The Economist, 2013) publicou um estudo comparativo da remuneração dos legisladores, em um grupo de países, que colocava o congressista brasileiro como 6º colocado em um ranking que mede a relação entre a remuneração do parlamentar e o PIB[[1]](#footnote-1) per capita. No caso brasileiro, o parlamentar recebia uma salário básico de aproximadamente 13 vezes o PIB per capita do país. (figura 2)



**Figura 2 - Comparativo remuneração parlamentares**Fonte: <https://www.economist.com/graphic-detail/2013/07/15/rewarding-work>. Acessado em 16/09/2018.

O termo GDP, empregado na figura 2, significa “Gross Domestic Product”, equivalente a PIB - Produto Interno Bruto - que é um indicador econômico da riqueza gerada por um país em um determinado período de tempo.

Parte significativa desses gastos está associado à CEAPS, regulamentada por ato da Comissão Diretora do Senado Federal, que instituiu o instrumento, em 2003, a fim de prover parlamentares com verba indenizatória pelo exercício da atividade parlamentar, destinada ao pagamento de despesas mensais realizadas pelo senador com o aluguel – de imóvel, de veículos ou de equipamentos – com material de expediente para escritório, com locomoção e com outras despesas diretas e exclusivamente relacionadas ao exercício da função parlamentar. (SECRETARIA DE GESTÃO DE INFORMAÇÃO E DOCUMENTAÇÃO DO SENADO FEDERAL, 2003)

A partir da sua constituição, a regulamentação da verba indenizatória sofreu uma série de alterações ajustando-a à necessidade da atividade parlamentar, com revisão dos valores remuneratórios, classificação das despesas, disciplinamento da concessão de passagens aéreas (locomoção) e adequação dos procedimentos a serem observados para a administração, controle e ressarcimento das despesas realizadas à conta da cota.

1. O que a literatura diz sobre a técnica aplicada

A análise exploratória de dados é uma filosofia que consiste

[...] no estudo dos dados a partir de todas as perspectivas e com todas as ferramentas possíveis, incluindo as já existentes. O propósito é extrair toda a informação possível, gerar novas hipóteses no sentido de construir conjecturas sobre as observações que dispomos. (BATANERO, ESTEPA e GODINO, 1991).

A finalidade da Análise Exploratória de Dados é examinar os dados previamente à aplicação de qualquer técnica estatística. Desta forma se consegue um entendimento básico de seus dados e das relações existentes entre as variáveis analisadas.

A Análise Exploratória de Dados, também conhecida como Estatística Descritiva, permite ao pesquisador compreender os dados, organizá-los e sintetizá-los de forma a obter as informações necessárias do conjunto de dados para responder as questões que estão sendo estudadas.

Segundo Reis (2002) a descrição dos dados também tem como objetivo identificar anomalias, até mesmo resultante do registro incorreto de valores, e dados dispersos, aqueles que não seguem a tendência geral do restante do conjunto.

População é o conjunto constituído por todos os indivíduos que representam pelo menos uma característica comum, cujo comportamento interessa analisar, ao passo que amostra pode ser definida como um subconjunto da população objeto de estudo, devendo preservar as principais características da população para ser considerada uma amostra representativa.

Variável é a característica de interesse que é medida em cada indivíduo da amostra ou população. Como o nome diz, seus valores variam de indivíduo para indivíduo. As variáveis podem ter valores numéricos ou não numéricos. (**Unsupported source type (DocumentFromInternetSite) for source Rei02.**).

No processo de entendimento e exploração dos dados fez-se necessário a criação de um indicador de produtividade, denominado **Índice de Produtividade (VL\_IPD)**. Segundo (YEUNG, DANTAS e LAZZARINI, 2013), a construção de indicadores de desempenho do poder legislativo é um desafio que invariavelmente provoca questionamentos pelas escolhas realizadas. Essas escolhas não denotam desconhecimento da complexidade do trabalho legislativo, mas muitas vezes uma opção por medidas simples de fácil compreensão e replicação.

A [análise de regressão](https://www.escolaedti.com.br/o-que-e-regressao-linear-entenda-aqui/?utm_source=blog&utm_campaign=rc_blogpost) procura explicar o comportamento entre variáveis que se relacionam e se dá por meio de modelos matemáticos, equações, que correlacionam a variável dependente com as independentes.

O modelo de regressão linear simples, com uma variável explicativa, aplica-se a várias situações, todavia quando envolvemos duas ou mais variáveis explicativas influenciando o comportamento da variável resposta (dependente), utilizamos o modelo de regressão linear múltipla. Por exemplo, o desempenho de um parlamentar (variável resposta) pode ser explicado por muitas variáveis (explicativas), como: participação em comissões, presença nas sessões, número de autorias, número de relatorias entre outras.

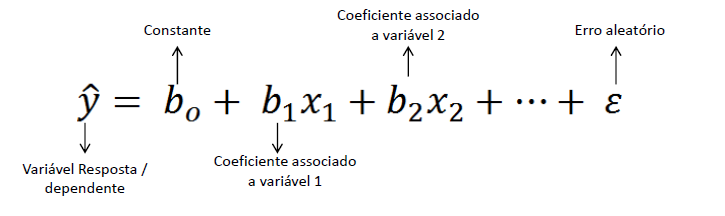


Figura 3- Equação da Regressão Linear Múltipla

Esta equação representa um modelo aditivo, no qual as variáveis preditoras (explicativas) somam-se na explicação de uma variável predita (resposta), possibilitando a melhor predição de uma variável dependentes. Para se chegar a esta reta, alguns métodos são aplicados como por exemplo: o mínimo quadrados, nele a melhor reta é aquela que deixa o menor valor de erro quadrado .

As variáveis quantitativas (numéricas de -∞ a +∞), podem ser discretas (inteiros) ou contínuas (decimais) e são trabalhadas diretamente no modelo da regressão. Já, as variáveis qualitativas não são representadas numericamente e são categorizadas em ordinais e nominais. As ordinais apresentam uma ordenação natural, por exemplo, fases do desenvolvimento de uma pessoa: infância, adolescência e adulta. As nominais não obedecem a um ordenamento natural, por exemplo, sexo: feminino e masculino. As variáveis nominais são substituídas nos modelos por variáveis ***dummie*** que assumem apenas dois níveis (0,1).

Variáveis nominais com apenas dois valores, assumem automaticamente os valores 0 e 1, por exemplo: sexo feminino = 0 e sexo masculino = 1. Todavia, as variáveis com mais valores, devemos criar uma variável para cada nível, indicando por 0 ou 1 se aquela observação pertence àquele nível.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Partido A | Partido B | Partido C | Partido E | Partido F |
| Senador 1 | 0 | 0 | 1 | 0 | 0 |
| Senador 2 | 1 | 0 | 0 | 0 | 0 |
| Senador 3 | 0 | 1 | 0 | 0 | 0 |
| Senador 4 | 0 | 0 | 0 | 0 | 1 |
| Senador 5 | 0 | 0 | 0 | 1 | 0 |

Para que o uso desta equação seja eficaz na predição da variável dependente em estudo, o pesquisador deve examinar previamente os pressupostos da RM, bem como identificar as consequências da sua violação. E esses pressupostos são: relação linear entre variáveis explicativas e dependentes (linearidade); erros provêm de distribuição normal (normalidade); os erros têm média 0 (zero); **homocedasticidade** (variância constante – homogeneidade nas variâncias); o erro de uma observação não é relacionado com o de outra; e que as variáveis explicativas sejam ortogonais – não sejam relacionadas entre si.

Devemos verificar se há alguma correlação muito forte entre variáveis explicativas porque, se houver, as inferências baseadas no modelo de regressão podem ser errôneas ou pouco confiáveis. Dois termos são empregados para definir essas correlações: **colinearidade** – relação linear entre duas variáveis explicativas; **multicolinearidade** – relação linear entre uma variável explicativa e as demais.

O VIF é a medida que representa o fator de inflação da variância e mede o quanto a variância do coeficiente é inflacionada pela colinearidade, e indica problemas quando seu valor é maior que 10.

A ausência de normalidade compromete os intervalos de confiança e os testes de hipótese, provocando o efeito denominado de **heterocedasticidade** quandoas probabilidades calculadas serão diferentes daquelas dadas nas tabelas de significância.

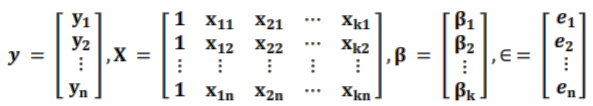
Quando a distribuição de variáveis apresenta uma viés, medidas como correlação ou regressão podem ser bastante influenciadas pelo pico da distribuição, como outliers, em situações como essas, devemos empregar uma das funções de transformações de variáveis: logaritma, raiz quadrada, inversa, quadrática, exponencial ou box-cox (potência). No processo de transformação, devemos atenção com as seguintes questões: o valor predito será dado em função da variável resposta transformada; os coeficiente ajustados acabam tendo uma interpretação indireta, um tanto mais complexa; e qualquer interpretação deverá ser feita após realizar a conta inversa.

A normalização das variáveis explicativas é comum ​quando o modelo de regressão contém múltiplas variáveis explicativas ou interação. A propósito da normalização é minimizar os problemas oriundos do uso de unidades e dispersões distintas entre as variáveis. Elas podem ser normalizadas segundo a amplitude ou segundo a distribuição.

Segundo (PETENATE, 2017), se o número de variáveis independentes for grande, “será necessário utilizar técnicas de subconjuntos visando remover preditores que não estejam associados às respostas, ajustando o modelo de regressão.” Deve-se decidir pelo subconjunto de variáveis explicativas que melhor explique a variável resposta.

Para lidar com o problema de regressão linear múltipla, é mais apropriado empregar a notação matricial, obtendo-se uma apresentação mais compacta dos dados, do modelo, bem como, dos resultados, sendo dada por:

Em que:

**

Lembrando que esse modelo de regressão deve obedecer aos seis pressupostos relacionados anteriormente.

Para estimar parâmetros no modelo de regressão múltiplo, podemos empregar o método dos mínimos quadrados onde uma reta encontrada reduz as distâncias entre os pontos observados e a reta, fazendo, em média, a soma dos desvios quadráticos ser igual a zero.

Obtido o estimador para o vetor de parâmetros, deve-se observar a significância dos resultados alcançados por meio da definição das somas dos quadrados: quadrados de resíduos, quadrados total, quadrados da regressão. Uma vez definidos, podemos escrever que:

Todas essas somas de quadrados segue uma distribuição de qui-quadrado, sendo que SQRes, SQReg e SQTot têm n-p, p-1 e n-1 graus de liberdade, respectivamente. A partir destes resultados é possível construir o quadro da análise de variância da regressão, que é usada para calcular a estatística F, utilizada para verificar a validade do modelo estimado.

Através da estatística F pode-se testar as hipóteses:

Ou seja, definimos o teste F para verificar se realmente existe uma relação linear entre Y e uma ou mais variáveis . (Morais, 2010).

O coeficiente de determinação é uma estatística usada para medir a proporção da soma de quadrados que é explicada pela regressão linear múltipla. Este coeficiente pode ser obtido pela expressão:

Contudo, essa medida tem uma série falha: na medida que se inclui mais variáveis explicativas no modelo, aumenta. Para solucionar este problema, ao se comparar diferentes modelos, deve-se utilizar o coeficiente de determinação ajustado, pois ele ajusta o valor de com base no número da amostra.

,

onde n = tamanho da amostra e k = nº de variáveis

Testes de hipóteses são empregados para se verificar se existe regressão linear entre as variáveis do modelo, um teste consagrado para esse fim é o teste F. No teste F para significância da equação de regressão testa-se a existência da regressão linear do modelo, por meio da fórmula:

Utilizando para testar as hipóteses a seguir:

Se for maior que , onde F tem distribuição de Snedecor com p-1 e n-p graus de liberdade, então rejeita-se e pode-se afirmar que, ao nível α de significância, pelo menos um , portanto pode-se dizer que existe regressão linear entre variáveis do modelo.

A influência de uma variável explicativa no modelo pode ser determinada pelo teste F parcial. Dessa forma, avalia-se a contribuição de uma variável explicativa para a soma dos quadrados devido à regressão, depois que todas as outras variáveis independentes foram incluídas no modelo. Assim, a influência desta variável do modelo para a soma de quadrados da regressão será estimada pela diferença dada por:

As hipóteses que serão testadas são:

Uma forma que nos permite mostrar que este teste é o dado pela expressão.

Dos vários métodos de seleção de variáveis explicativas, quatro se destacam: All Regression, Forward, Backward e Stepwise.

O teste mais utilizado para medir a significância individual das variáveis do modelo é o teste **t** que nos mostra se existe alguma variável significativa no modelo. Assim a quantidade a ser testada para cada será dada pela fórmula:

Em que representa o estimador da variância de , e segue a distribuição **t** de Student com n-p graus de liberdade e que é usado para verificar as seguintes hipóteses:

Portanto, se || > , então, rejeita-se e conclui-se que, no nível

α de significância, , j = 1,2...n e assim portanto pode-se dizer afirmar que esta variável é importante para explicar a regressão linear. Caso contrário, esta variável não tem influência na regressão linear.

Para confirmação a respeito das suposições de uma regressão linear é fundamental fazer uma investigação no conjunto de dados para verificar a condição de normalidade através dos gráficos de probabilidade normal que são o PP-plot (probabilidade acumulada esperada para a distribuição normal, em função da probabilidade observada acumuladas dos resíduos) e o QQ-plot ( Quantil e probabilidade esperado para a distribuição normal, em função dos resíduos).

Graficamente podemos verificar que, se os erros possuírem distribuição normal, os pontos devem estar mais ou menos alinhados sobre uma reta, senão, os dados não apresentam indícios de normalidade.

**4.Aplicação da metodologia**

Análise exploratório do banco

|  |  |  |
| --- | --- | --- |
| **Variável** | **Descrição da Variável** | **Detalhamento** |
| CD\_Senador | Identificados único do senador |  |
| AN\_Referencia | Ano de referência |  |
| NM\_Senador | Nome do senador |  |
| QT\_Tipo\_Despesa | Quantidade ????? |  |
| QT\_Fornecedor | Quantidade ????? |  |
| NM\_Uf | Nome da Unidade da Federação representada pelo senador |  |
| SG\_Partido | Sigla do partido político o qual faz parte o senador |  |
| DT\_Nasc | Data de Nascimento do parlamentar |  |
| QT\_Anos\_Idade | Idade do parlamentar em anos |  |
| CS\_Sexo | Classificador da variável Sexo | CS\_SEXO = “F” ou “M” |
| VL\_Reembolsado | Valor do reembolso efetuado pelo Senado Federal |  |
| VL\_Ipd | Índice de produtividade |  |

Aplicação da técnica

**Índice de participação em comissões (VL\_IPC)**

As comissões no Senado se dividem em permanentes, temporárias e CPI[[2]](#footnote-2). Atualmente há 14 comissões permanentes, nove comissões temporárias e outras quatro destinadas à CPI. (Atividade Legislativa, 2018)

Esse indicador é calculado a partir do somatório dos dias em que cada senador participou de comissões, não importando se como titular ou suplente. Ao senador com maior nº de dias de participações, atribuímos o valor 1 ao índice. Aos demais, o índice foi definido como a relação entre o nº de dias de participações do senador pelo nº dias de participações do senador mais participativo.

**Índice de autorias (VL\_IAU)**

Autoria, no contexto do Senado Federal, corresponde ao ato de formular avisos, atos, consultas, denúncias, grupos de trabalho, indicações, mensagens, moções, medidas provisórias, ofícios, pareceres, petições, projetos, portarias, questões de ordem, recursos, representações, requerimentos, substitutivos, sugestões e vetos. Para medir esta atividade, criamos o **índice de autorias (VL\_IAU).**

Este indicador trata da quantidade de formulações de autoria de um determinado senador. Para fins de cálculo será considerado o ano de apresentação do projeto.

Ao senador com maior nº de autorias, atribuímos o valor 1 ao índice. Aos demais, o índice foi definido com a relação entre o nº de autorias do senador pelo nº autorias do senador mais participativo.

**Índice de relatoria (VL\_IRL)**

Relator é o parlamentar designado pelo presidente da comissão para apresentar parecer sobre matéria de competência do colegiado, sendo escolhido de acordo com a proporção das bancadas partidárias ou blocos. (Senado Notícias, 2018)

Este indicador trata da quantidade de relatorias realizadas por um determinado senador.

Para fins de cálculo deste indicador, consideramos o exercício mínimo de 30 dias na relatoria por ano.

Ao senador com maior nº de relatorias, atribuímos o valor 1 ao índice. Aos demais, o índice foi definido com a relação entre o nº de relatorias do senador pelo nº relatorias do senador mais participativo.

**Índice de Produtividade (VL\_IPD)**

Este indicador é calculado como o somatório dos índices de relatoria (VL\_IRL), de autoria (VL\_IAU) e de participação em comissões (VL\_IPC).

**5.Conclusões**

Do ponto de vista do negócio

Do ponto de vista da técnica

**6.Bibliografia**

**Unsupported source type (DocumentFromInternetSite) for source Rei02.**

**Unsupported source type (ElectronicSource) for source Dom15.**

ATIVIDADE Legislativa. **Senado Federal**, 18 set. 2018. Disponivel em: <https://legis.senado.leg.br/comissoes/;jsessionid=6CDAB5FA2AB58BE227328C0993C09742?0#>.

BATANERO, C.; ESTEPA, A.; GODINO, J. D. Análisis exploratorio de datos: sus posibilidades en la enseñanza secundaria. **Suma**, p. 25-31, 1991.

BRASIL, S. F. D. Senado Federal. **Senado Federal**, 2005. Disponivel em: <https://www12.senado.leg.br/transparencia/leg/pdf/normas/decisao%20da%20comissao%20diretora%20no%203%20de%202005.pdf>. Acesso em: 30 jun. 2018.

BRASILEIRO, S. Transparência Senado. **Senado Federal**, 2009. Disponivel em: <https://www12.senado.leg.br/transparencia/leg/pdf/normas/ATO%20DA%20COMISSAO%20DIRETORA>. Acesso em: 28 jun. 2018.

FEDERAL, C. D. D. S. Senado Federal. **Senado Federal**, 2007. Disponivel em: <https://adm.senado.gov.br/normas/ui/pub/normaConsultada?20&idNorma=206676>. Acesso em: 30 jun. 2018.

FEDERAL, C. D. D. S. Senado Federal. **Senado Federal**, 2009. ISSN HTTPS://ADM.SENADO.GOV.BR/NORMAS/UI/PUB/NORMACONSULTADA?10&IDNORMA=259330. Disponivel em: <NID=6AAA25BC27787F58449D85DBE39B3C5E?0&IDNORMA=259714>. Acesso em: 30 jun. 2018.

FEDERAL, P. D. S. Senado Federal. **https:** //adm.senado.gov.br/normas, Brasilia, 2003. Disponivel em: <https://adm.senado.gov.br/normas/ui/pub/normaConsultada?6&idNorma=236036>. Acesso em: 29 jun. 2018.

FEDERAL, S. Senado Federal. **Transparência Senado Federal**, 2005. Disponivel em: <https://www12.senado.leg.br/transparencia/leg/pdf/normas/DECISAO%20DA%20COMISSAO%20DIRETORA%20No%203%20DE%202005.pdf>. Acesso em: 29 jun. 2018.

SECRETARIA DE GESTÃO DE INFORMAÇÃO E DOCUMENTAÇÃO DO SENADO FEDERAL. Transparência. **Senado Federal**, 2003. Disponivel em: <https://adm.senado.gov.br/normas/ui/pub/normaConsultada;jsessionid=C247F601642D366934A51E6DA3D0150F?0&idNorma=236035>. Acesso em: 17 jun. 2018.

SECRETARIO, P. Senado Federal. **Senado Federal**, 2011. Disponivel em: <https://adm.senado.gov.br/normas/ui/pub/normaConsultada?8&idNorma=209408>. Acesso em: 30 jun. 2018.

SECRETÁRIO, P. Senado Federal. **Senado Federal**, 2014. Disponivel em: <https://adm.senado.gov.br/normas/ui/pub/normaConsultada?2&idNorma=203003>. Acesso em: 30 jun. 2018.

SENADO FEDERAL. Transparência. **Senado Federal**. Disponivel em: <https://www12.senado.leg.br/transparencia/leg/pdf/CotaExercicioAtivParlamSenadores.pdf>. Acesso em: 17 jun. 2018.

SENADO Notícias. **Glossário Legislativo**, 2018. Disponivel em: <https://www12.senado.leg.br/noticias/glossario-legislativo/relator>.

SENADO, C. D. D. Senado Federal. **adm.senado.gov.br**, 2009. Disponivel em: <https://adm.senado.gov.br/normas/ui/pub/normaConsultada?8&idNorma=209408>. Acesso em: 30 jun. 2018.

SENADO, C. D. D. Senado Federal. **ADM.SENADO.GOV.BR**, 2011. Disponivel em: <HTTPS://ADM.SENADO.GOV.BR/NORMAS/UI/PUB/NORMACONSULTADA?10&IDNORMA=259330>. Acesso em: 30 jun. 2018.

THE Economist. **The Economist**, 15 jul. 2013. Disponivel em: <https://www.economist.com/graphic-detail/2013/07/15/rewarding-work.>.

YEUNG, L.; DANTAS, H.; LAZZARINI, S. Medindo o desempenho de um parlamento brasileiro: o caso da Câmara Municipal de São Paulo. **Insper**, 2013. 4.

**7.Anexos**

Imagens

Códigos

1. O Produto Interno Bruto é um indicador econômico que, basicamente, define o valor de toda a riqueza gerada em um determinado país ou região (daí o índice ser considerado também geográfico) em um determinado período de tempo. [↑](#footnote-ref-1)
2. CPI - Comissão Parlamentar de Inquérito [↑](#footnote-ref-2)