

Trabalho individual de Análise de Dados Económicos I

Profª Marta Silva

Mestrado em Economia – Universidade do Porto – 2020/2021

Aluno: Fellipe Mateus Miranda da Silva

Novembro 2020

# **Índice**

[**Índice** 2](#_Toc55249788)

[**Introdução** 3](#_Toc55249789)

[**Justificação da base de dados** 4](#_Toc55249790)

[**Descrição da base de dados** 5](#_Toc55249791)

[**Documentação inicial da base de dados** 6](#_Toc55249792)

[**Depuração da base de dados** 9](#_Toc55249793)

[**Alterações na base de dados** 9](#_Toc55249794)

[**Base de dados final** 9](#_Toc55249795)

[**Descrição das variáveis da base de dados final** 10](#_Toc55249796)

[**Conclusão** 11](#_Toc55249797)

[**Bibliografia** 12](#_Toc55249798)

# **Introdução**

A pesquisa acadêmica é de importância sumária à sociedade ao passo em que ela contribui para a sua evolução ao trazer novos modos de avaliação dos problemas que nos acometem, e também soluções para tais problemas. E no contexto atual de crises econômicas, sociais, ecológicas e também de saúde, o papel da pesquisa acadêmica torna-se ainda mais prominente em meio à grande necessidade de se encontrar ações e ferramentas que permitam que governos e indivíduos deixem para trás os problemas do passado e do presente, rumo a um futuro melhor.

A Economia é a “mãe” das ciências sociais aplicadas justamente por ter um papel especial na identificação de origens dos problemas que acometem indivíduos e consequentemente a sociedade como um todo, e no encontro das soluções que corrigem este curso. É uma ciência valorizada pela excelência de seus métodos, seu rigor científico acoplado a uma sólida base matemática, e a falta de temor em colocar de forma clara os problemas e as soluções que estas exigem, unindo uma visão positiva, normativa e até mesmo política neste processo.

Desse modo, torna-se sumariamente importante ao economista em potencial estudar desde seus primeiros passos na academia os métodos de pesquisa “de ponta” que pesquisadores profissionais utilizam. A conjunção de ferramentas tecnológicas que facilitam o trabalho matemático, em conjunto com a abertura dos grandes bancos de dados para análises a níveis micro e macrossociais, permite que economistas, outros cientistas sociais e até membros da sociedade em outros campos possam utilizar tais dados para realizar suas próprias análises, e desta forma contribuir para o processo de acumulação de conhecimento científico que nos permite continuar a evoluir a sociedade como um todo

Mas tal processo ainda exige um conhecimento metodológico e também processual eficaz para que as pesquisas possam não só serem realizadas, mas também apuradas, repetidas e continuadas por outros cientistas. Por isso a importância de “logs” de sessões de depuração de bancos de dados, compilação de dados iniciais e finais, entre outros processos que permitem a continuidade e reprodução do método científico na economia e além.

# **Justificação da base de dados**

A desigualdade é um problema social que voltou em voga na ciência econômica muito a partir do trabalho seminal de Thomas Piketty no livro *O Capital no Século XXI*, publicado em Português em 2014 pela editora Temas e Debates. Enquanto a ideia central do livro, $$r>g$$ (Piketty, 2013), já passou por enorme escrutínio pela classe acadêmica, o impacto na ciência é claro a partir do aumento de … % em trabalhos acadêmicos com referências a Piketty (referência), e também trabalhos próprios iniciados por outros cientistas sobre o assunto (referência).

O estudo da desigualdade tem importância a partir do degrau em que a mesma dá origem e também explica alguns dos grandes problemas sociais que acometem sociedades em desenvolvimento, e também países já desenvolvidos. Para além do desnível de ganhos em renda e da posse do capital entre classes sociais, existem os problemas políticos de influência de grupos de interesse sobre decisões públicas (referência) e até ecológicos, a partir dos efeitos que geram os problemas de mudança climática partindo de países mais ricos, e afetando principalmente as nações mais pobres (referência).

Por muito tempo, o esforço de estudos sobre a desigualdade era baseado na procura minuciosa em trabalhos acadêmicos especializados de alguns economistas conhecidos por tratarem do assunto mesmo quando o mesmo estava fora de “moda”, como Simon Kuznets e Anthony Atkinson, e dados de censos e outras pesquisas públicas cujo acesso ao público poderia ser aberto, mas não muito facilitado (referência). A partir do trabalho de Piketty em obras anteriores a *O Capital no Século XXI*, viu-se a oportunidade e também a necessidade de se compilar em um só lugar estas séries de dados já reunidos para análise do fenômeno da desigualdade em nossas sociedades.

Em 2011 surge o *The World Top Incomes Database*, cuja meta inicial era a divulgação dos dados compilados por Piketty, Atkinson e Emmanuel Saez. Este evoluiu para o *World Wealth and Income Database* em 2015 com a adição dos trabalhos realizados por Piketty e Gabriel Zucman entre 2013 e 2014; e chegou à sua edição atual, o *World Inequality Database*, em 2017 graças ao trabalho de um time internacional com 132 pesquisadores que hoje analisam os dados de mais de 70 países em 5 continentes diferentes (World Inequality Database, 2020).

# **Descrição da base de dados**

O *WID* é uma base de dados de acesso público e livre, com foco na divulgação de dados sobre a desigualdade de renda e de riqueza, as estatísticas socioeconômicas que dão base a tal análise, e trabalhos sobre desigualdades realizados por pesquisadores da organização (WID, 2020). Os dados são encontrados por meio de mapas mundiais que podem ser divididos entre países e regiões; gráficos que ilustram a evolução dos dados compilados ano a ano para cada país analisado, e tabelas de dados montadas na seção “Data” do website que permitem a escolha de indicadores, países e/ou regiões e o período anual a serem analisados, cujo resultado pode ser descarregado nos formatos .csv, .xls, .xlsx e .sdmx (WID, 2020).

O *WID* funciona a partir da angariação de um conjunto de dados como contas nacionais, dados censitários, dados fiscais e rankings de riqueza para a construção da base de dados divulgada através de seus vários canais. Um dos conceitos centrais por trás dos dados é a noção de renda nacional, que é o Produto Interno Bruno subtraído da depreciação de capital, e a fração da produção interna que é transferida para estrangeiros. Este gera o “Distributional National Account”, traduzido livremente para Conta Nacional Distribucional, cuja metodologia discorre sobre unidades de medida, conceitos de renda e riqueza, e métodos de medição de contas macroeconômicas, séries de distribuição de renda e riqueza, como lidar com países com dados limitados, desigualdades regionais e globais, e medidas que garantem a qualidade dos dados e também a sua transparência (WID, 2020).

A descrição dos dados distrbuídos pela *WID* pode ser encontrada na “Summary table”, encontrada na seção “Data” do banco de dados. Esta seção também possibilita o download de todos os dados disponibilizados no banco de dados em formato .csv, através de um único arquivo .zip (WID, 2020).

O *WID* também possui “packages” para uso externo em softwares focados em análise de dados. Estes servem para programas que usem a linguagem de programação R, como o RStudio; e também para o Stata (WID, 2020).

# **Documentação inicial da base de dados**

Considerando os ensinamentos passados na disciplina de Análise de Dados Econômicos I, apura-se que o Stata é uma excelente ferramenta para a compilação, tratamento e análise de dados. Logo, este será o programa no qual o processo descrito neste relatório será realizado.

Antes de tudo, faz-se bem em ler a documentação disponibilizada na seção “Codes dictionary” do website do *WID*. Este detalha a construção dos códigos a serem utilizados no Stata para a utilização dos indicadores e outros dados disponibilizados pelo *WID* no programa (WID, 2020).

Uma vez que pesquisas são feitas para encontrar respostas, a situação hipotética deste trabalho será a de um pesquisador que deseja identificar tendências na evolução das desigualdades de renda dos Estados Unidos a partir de fatores econômicos e potencialmente políticos. Sua base de análise partirá das proporções de renda disponíveis para os 50% mais pobres do país, e para os 10% mais ricos da nação, entre os anos de 1980 e 2014.

Para isso, o *WID* apresentou-se como a base de dados mais concisa e também mais focada nos dados que ele necessita, tanto os primários – que tratam das proporções de distribuição em si – quanto os secundários – que identificam componentes que podem explicar as razões pelas quais as desigualdades tem sido encontradas em primeira instância.

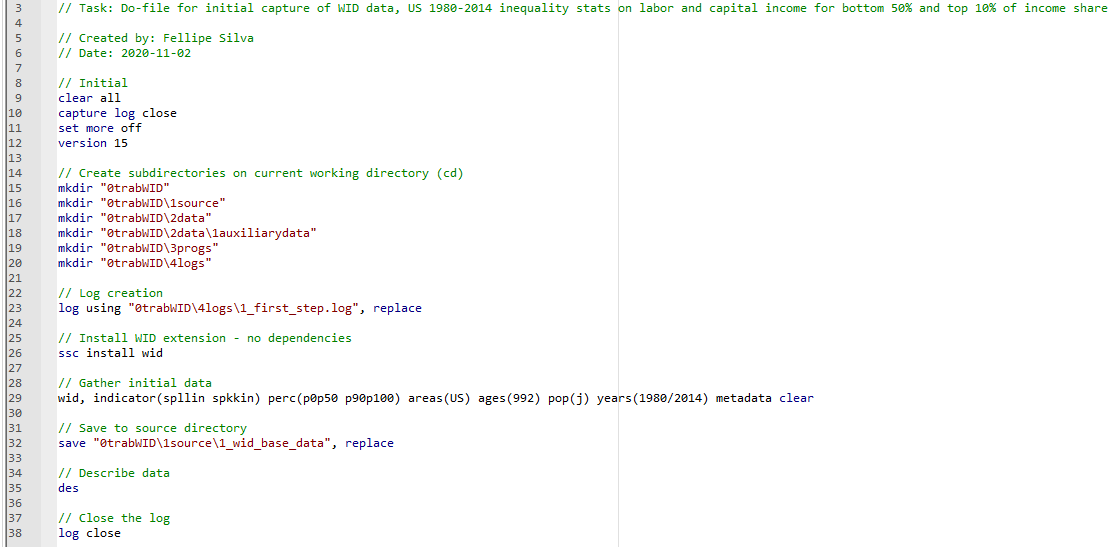
Assim, o objetivo inicial do pesquisador é identificar os componentes de renda de trabalho e de renda de capital, a partir do conceito supramencionado de renda nacional, antes da realização de cobrança de impostos. Estes componentes serão extraídos entre as faixas dos 50% mais pobres e 10% mais ricos dos Estados Unidos, entre indivíduos acima de 20 anos (que fazem assim parte da população economicamente ativa), com partilha idêntica entre os componentes familiares, entre os anos 1980 e 2014.

Antes da execução do código para extração da base de dados inicial, faz-se necessária a organização propícia dos diretórios de trabalho e da documentação auxiliar que guiará o processo. Para este último, foi utilizado o aplicativo de notas Notion – que possui suporte para a linguagem Markdown – para a linearização do processo, e as potenciais etapas que este pode tomar.

A organização dos diretórios de trabalho segue a ordem passada em sala de aula, com um diretório principal para o projeto (“0trabWDI”), subdiretórios para os arquivos do projeto (“1source”, “2data”, “3progs”, “4logs”), e um subdiretório para dados auxiliares (“2data\1auxiliarydata”). A intenção é seguir os princípios de planejamento, organização, documentação e execução conforme as lições do livro-base da disciplina (Long, 2009).

Também conforme as lições de J. Scott Long, que não recomenda o uso do comando “cd” na criação de do-files em projetos a serem distribuídos e/ou compartilhados (página 53), a escolha é pela criação de pastas de trabalho utilizando o comando “mkdir”. Este cria a “árvore” de diretórios necessária para a reprodução do projeto em outras máquinas.

Assim se tem a estrutura do primeiro do-file do projeto:

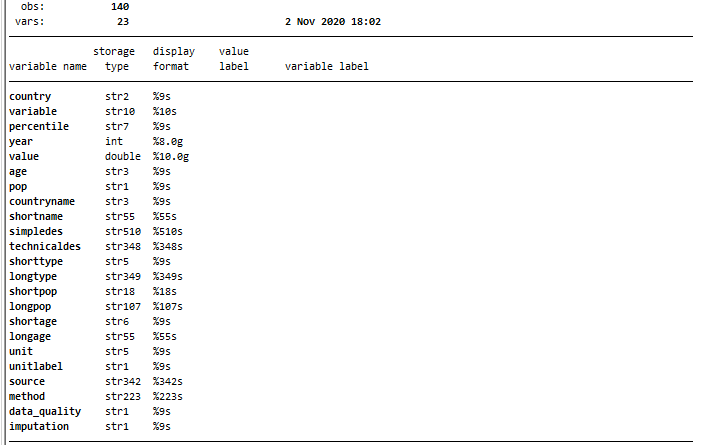


Fonte: o autor (2020)

O comando “log” cria uma espécie de relatório do processo executado via Stata. Uma vez que a extensão de uso do *WID* via Stata não possui dependências (arquivos, extensões e/ou plugins necessários para o seu funcionamento), uma linha de código já serve para que a mesma seja instalada.

O comando “wid” evoca a base de dados do *WID*. A opção “indicator” denota os indicadores a serem trazidos para o programa. “perc” indica as faixas de renda (p0p50 são os 50% mais pobres da população, e p90p100 são os 100% mais ricos da população) a serem analisadas. “areas” são os países a serem analisados, normatizados pelo código ISO 3166-1 alpha-2. “ages” e “pop” referem-se à idade dos indivíduos, e como é feita a partilha de renda em seus domicílios, onde “j” é a partilha idêntica entre casais. “years” é a faixa de anos a ser analisada; e “metadata” descarrega junto com as variáveis anteriores os metadados da base de dados.

O resultado é salvo no diretório “1source”, e os dados são descritos via comando “des” para conferir o nome e a descrição das variáveis, o tipo de armazenamento, seu formato e potenciais registros de valor.



Fonte: o autor (2020)

O processo se encerra com o fechamento do “log” do processo.

# **Depuração da base de dados**

Enquanto os resultados iniciais são satisfatórios, sem quaisquer dados em falta, e todos eles muito bem descritos pelos metadados, será necessário um bom polimento para que os dados angariados possam ser colocados em prova. Vê-se a necessidade de eliminar algumas variáveis de metadata após o uso das mesmas para a construção de descrições e notas das variáveis necessárias, e quiçá transformar a variável “percentile” em duas, uma para os 50% mais pobres e outra para os 10% mais ricos.

Serão necessárias também alterações em formatos, transformando o maior número possível de variáveis “string” (em formato de texto) em forma numérica. Além disso, a adição de mais uma região para comparação com os Estados Unidos seria uma boa maneira de verificar se há discrepâncias entre a trajetória estadunidense e este outro país representativo.

# **Alterações na base de dados**

No dia seguinte, o pesquisador se põe a adicionar à base de dados atual um novo país, a França, que lhe parece representativo da trajetória dentro da União Europeia de evolução das desigualdades de renda e de capital; e também mais uma variável, a riqueza líquida das famílias tanto nos Estados Unidos, quanto na França, para verificar potenciais semelhanças e discrepâncias entre as duas nações.

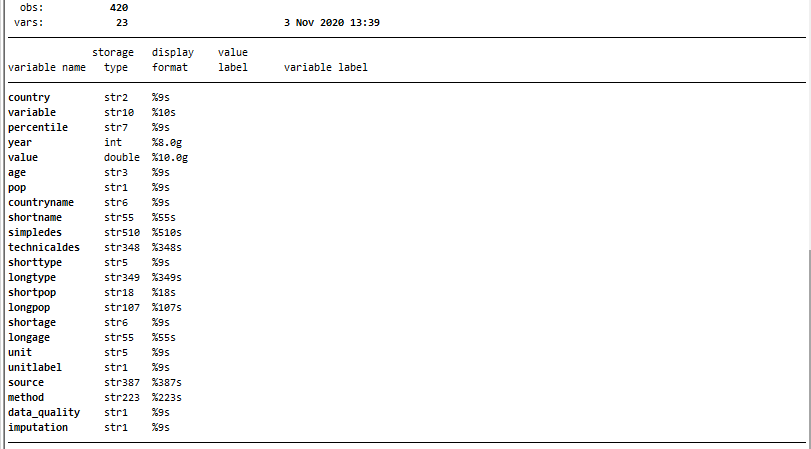
Existem duas maneiras de se fazer tais adições. A primeira é mudar o código de acesso ao *WID* no Stata, adicionando a ela mais elementos a serem descarregados e transformados em uma base de dados. Logo, o código original

wid, indicator(spllin spkkin) perc(p0p50 p90p100) areas(US) ages(992) pop(j) years(1980/2014) metadata clear

Torna-se

wid, indicator(spllin spkkin shweal) perc(p0p50 p90p100) areas(US FR) ages(992) pop(j) years(1980/2014) metadata clear

E o resultado consequente é



Fonte: o autor (2020)

Outra maneira é usar a base de dados original e realizar “appends” com os dados da França, e com os dados de riqueza líquida das famílias. Primeiro evoca-se a base de dados a ser adicionada por meio do comando

wid, indicator(spllin spkkin) perc(p0p50 p90p100) areas(FR) ages(992) pop(j) years(1980/2014) metadata clear

O resultado é salvo no diretório “2data\1auxiliarydata”, subpasta do diretório principal de trabalho “0trabWID”, com o nome “1\_wid\_FR\_income”. Depois, utiliza-se o comando

append using "1source\1\_wid\_base\_data"

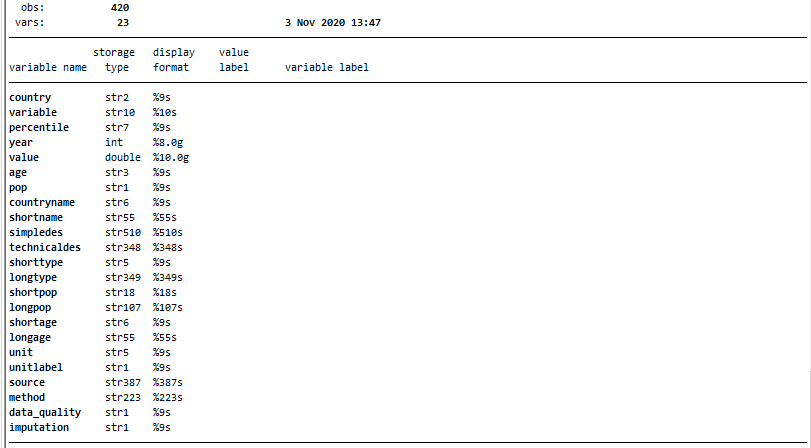
Para “unir” as duas bases de dados em questão. Será necessário um último seguimento de comandos “wid”, “save” e “append” para que o resultado seja o mesmo da utilização de uma só linha de comando, como no exemplo acima. Estes são

save "2data\1auxiliarydata\1\_wid\_USFR\_income", replace

wid, indicator(shweal) perc(p0p50 p90p100) areas(US FR) ages(992) pop(j) years(1980/2014) metadata clear

append using "2data\1auxiliarydata\1\_wid\_USFR\_income"

E o resultado é



Fonte: O autor (2020)

Mas mais uma vez, com as lições de J. Scott Long (2009) em mente, o pesquisador opta pela via de menor resistência. Esta é a primeira opção discutida, modificando o código original de descarregamento do *WID*.

A nova carga de dados exige modificações, em formatação, descrições e estrutura geral da organização das variáveis. Este será o foco a partir deste momento.

Tudo começa com a criação da árvore de diretórios através do comando “capture mkdir” que suprime os erros gerados caso as pastas já existam no diretório de trabalho, seguida pela remoção das variáveis relacionadas aos metadados por meio do comando “drop”. Este é seguido pela criação de uma variável “id” pelo comando “egen”, que usa como referência o grupo de variáveis que permanecerão como referência para as observações da pesquisa (“country”, “year” e “percentile”).

A nova variável é referência para a transformação da base de dados de um formato “long” para “wide” pelo comando “reshape”, com a variável “variable” – que contém o nome dos indicadores descarregados via *WID* – servindo como a “coluna j” do comando. A variável “value” é alocada entre as novas variáveis criadas a partir deste comando, cada uma delas representando a proporção de renda de trabalho, renda de capital e riqueza líquida dos indivíduos.

As variáveis ganham novos nomes (“l\_income”, “k\_income” e “net\_wealth”), descrições e notas detalhando o que cada uma representa. O mesmo é feito para as variáveis restantes (“country”, “year” e “percentile”).

O próximo passo é transformar as variáveis previamente em formato “string” para numéricas, uma vez que isto facilita a realização de testes mais a frente no processo de pesquisa. Isso começa com a criação de uma variável numérica representando o código alfanumérico do padrão ISO 3166 para países, por meio de um “merge” com uma tabela de referência disponibilizada na internet.

O pacote de ferramentas LABUTIL é instalado para o uso do comando “labmask” que em conjunto com o código “numlabel”, transforma a variável “country” em uma variável numérica. O mesmo é feito com a variável “percentile” por meio do comando “encode”.

Mais descrições para variáveis e notas são adicionadas. O próximo processo é usar a mudança da formatação das variáveis “l\_income”, “k\_income” e “net\_wealth” da base de dados, normatizando-as para “%8.0g”.

O processo se encerra com o comando “compress”, cuja única alteração é a mudança da variável “percentile” de long para byte para reduzir o tamanho da base de dados. No fim o resultado do processo é salvo na pasta “2data”, e o “log” da sessão é também salvo na sua pasta de destino “4logs”.

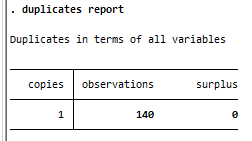
Antes da montagem do do-file, faz-se necessário um processo de “debug” para identificar quaisquer tipos de erros no processo delineado acima. Estes se fazem através dos códigos “codebook” e “duplicates”.

O comando “codebook” associado à opção “problems” reporta quaisquer problemas encontrados no dataset carregado no Stata, como variáveis, valores e descrições faltantes ou incompletas. Para o dataset atual, nenhum problema foi encontrado.



Fonte: o autor (2020)

Já o comando “duplicates”, junto de “report” (sem vírgula) identifica valores duplicados no dataset. Mais uma vez, nenhum problema foi identificado.



Fonte: o autor (2020)

Houve a tentativa de se utilizar o comando “panelstat” para avaliação mais aprofundada do dataset. Entretanto, uma vez que existem duas variáveis de painel (“country” e “percentile”) e não só uma nos dados avaliados, não foi possível obter um resultado de confiança satisfatória no processo de debug.

# **Base de dados final**

A base de dados final deste relatório é resultado do do-file “2\_20201104\_next\_do-file.do”, encontrado no diretório “3progs” da “árvore” principal do projeto. Este reproduz com fidelidade os resultados do processo de alteração descrito acima.

O dataset possui 6 variáveis, “country”, “year”, “percentile”, “l\_income”, “k\_income” e “net\_wealth”; e 140 observações, relacionadas aos países (Estados Unidos e França), anos (1980 a 2014) e percentis (50% mais pobres e 10% mais ricos) analisados.

Alterações futuras seriam a unificação das variáveis “country” e “percentile” em uma só, para que o dataset possa estar em formato de dados em painel; e a adição de novos países à análise. Também poderiam ser adicionados novos indicadores do *WID* que ajudem a identificar as grandes discrepâncias distributivas que existem tanto nos Estados Unidos quanto na França, como consumo e poupança do setor familiar e valor de ativos imobiliários (WID, 2020).

# **Descrição das variáveis da base de dados final**

As descrições são obtidas através dos comandos “summarize”, “describe”, “codebook, compact”, “codebook” e “inspect”. Estes podem ser encontrados no log “7\_var\_description”, no subdiretório “4logs”; e os três primeiros estão também aqui abaixo.

**. summarize**

Variable | Obs Mean Std. Dev. Min Max

-------------+---------------------------------------------------------

country | 140 545 296.0592 250 840

year | 140 1997 10.13577 1980 2014

percentile | 140 1.5 .5017953 1 2

l\_income | 140 .25682 .0734249 .1375 .4047

k\_income | 140 .3500707 .3371299 -.0467 .7548

-------------+---------------------------------------------------------

net\_wealth | 140 .3198171 .2824432 -.0158 .7389

**. describe**

Contains data from 0trabWID\2data\5\_USFR\_income\_wealth.dta

obs: 140

vars: 6 4 Nov 2020 13:19

(\_dta has notes)

-------------------------------------------------------------------------------------------------------------------------------------

storage display value

variable name type format label variable label

-------------------------------------------------------------------------------------------------------------------------------------

country int %8.0g country\_num

\* Country of observation

year int %8.0g Year of observation

percentile byte %8.0g percs \* Income share percentile within the nation

l\_income double %8.0g \* Pre-tax labor income

k\_income double %8.0g \* Pre-tax capital income

net\_wealth double %8.0g \* Net personal wealth

\* indicated variables have notes

**. codebook, compact**

Variable Obs Unique Mean Min Max Label

-------------------------------------------------------------------------------------------------------------------------------------

country 140 2 545 250 840 Country of observation

year 140 35 1997 1980 2014 Year of observation

percentile 140 2 1.5 1 2 Income share percentile within the nation

l\_income 140 135 .25682 .1375 .4047 Pre-tax labor income

k\_income 140 136 .3500707 -.0467 .7548 Pre-tax capital income

net\_wealth 140 136 .3198171 -.0158 .7389 Net personal wealth

# **Conclusão**

# **Bibliografia**