Health Data Sets Accessible Online

(emphasis on ***data sets*** archived in SAS, STATA, and/or delimited [R, excel readable] formats)

pp. 1-4 Introduction to Some Publicly Available Health Data Sets
pp 5-19 Illustrations of downloading Data and Estimating Models

Introduction to Data Set Types/Questionnaires: Advantages/disadvantages
https://www.ahrq.gov/talkingquality/measures/understand/index.html

National Library of Medicine (NIH-National Institutes of Health)
Introductory Page (include common statistics terms in use):
https://www.nlm.nih.gov/nichsr/stats_tutorial/cover.html
glossary summary of data sources:
https://www.nlm.nih.gov/nichsr/stats_tutorial/glossary.html#NHIS

SOME COMMON PORTALS (often you register, but its free):

Bureau of Labor Statistics:
Occupational injures/diseases by industry by year (but not by states)
https://www.bls.gov/iif/oshsum.htm#20Summary_Tables

California Health Interview Survey (CHIS)
http://healthpolicy.ucla.edu/chis/data/Pages/public-use-data.aspx
Comprehensive statewide data files on a variety of topics. Public Use Files (PUFs) enable researchers to customize and run their own data searches. The files are available in a variety of data formats, including SAS, SPSS, and STATA data formats. Free registration is required.

Centers for Medicare and Medicaid Services (CMS)
https://data.cms.gov/
A. Current beneficiary Survey (2016-2019) https://data.cms.gov/medicare-current-beneficiary-survey-mcbs/medicare-current-beneficiary-survey-data
B. Current beneficiary Surveys (year by year) https://www.cms.gov/research-statistics-data-and-systems/downloadable-public-use-files/mcbs-public-use-file

Global Health Data Exchange
http://ghdx.healthdata.org/data-sites-we-love

HealthData.gov (odds/ends of recent data)
https://healthdata.gov/

Inter-university Consortium for Political and Social Research (*ICPSR*, Univ of Michigan)
https://www.icpsr.umich.edu/web/pages/ICPSR/index.html
First time users need to create a MyData account to download data. Its free
A. National Health Interview Surveys [[IPUMS alternative: https://healthsurveys.ipums.org/]]
https://www.icpsr.umich.edu/web/ICPSR/series/40
https://www.icpsr.umich.edu/web/ICPSR/search/studies?start=0&ARCHIVE=ICPSR&PUBLISH_STATUS=PUBLISHED&sort=score%20desc%2CTITLE_SORT%20asc&rows=50&q=health%20interview%20survey

(these include 5100+ general and specialized population surveys on health related topics)
Recent Health Interview Surveys include: … "The Family-level File (Part 2) is made up of reconstructed variables from the person-level data of the basic module and includes information on sex, age, race, marital status, Hispanic origin, education, veteran status, family income, family size, major activities, *health status, activity limits*, and employment status, along with industry and occupation. As part of the basic module, the Person-level File (Part 3) provides information on all family members with respect *to health status, limitation of daily activities, cognitive impairment, and health conditions*. Also included are data on years at current residence, region variables, height, weight, bed days, doctor visits, hospital stays, and health care access and utilization. A randomly-selected adult in each family was interviewed for the Sample Adult File https://www.icpsr.umich.edu/files/ICPSR/images/homepage-march-2021.gif (Part 4) regarding *respiratory conditions, use of nasal spray, renal conditions, AIDS, joint symptoms, health status, limitation of daily activities, and behaviors such as smoking, alcohol consumption, and physical activity*…"
B. Longitudinal Longevity Surveys (repeated surveys of old people and their characteristics; to help ascertain health in old age), including Chinese Surveys:
https://www.icpsr.umich.edu/web/ICPSR/series/487
C. General Social Surveys (smaller samples with specialized issues—a few medical questions)
https://www.icpsr.umich.edu/web/ICPSR/series/28
D. Health Reform Monitoring Survey (HRMS) Series–ACA implementation and trends, survey of nonelderly
https://www.icpsr.umich.edu/web/ICPSR/series/547
E. Irish Longitudinal Study on Aging (health questions, many socio-demographic factors)
https://www.icpsr.umich.edu/web/ICPSR/series/726
F. Linked US Births/Infant Death Data (birth records linked to infant deaths, incl nonresident)
https://www.icpsr.umich.edu/web/ICPSR/series/33
G. Detailed Mortality Files (death certificate data, including cause of death)
https://www.icpsr.umich.edu/web/ICPSR/series/89
H. Multiple Causes of Death Series (criminology interest)
https://www.icpsr.umich.edu/web/ICPSR/series/90
I. National Ambulatory Medical Care Survey Series (physician visits, etc)
https://www.icpsr.umich.edu/web/ICPSR/series/37
J. National Health and Nutrition Examination Survey (NHANES)
https://www.icpsr.umich.edu/web/ICPSR/series/39
K. National Home and Hospice Care Survey Series (on centers, staff, and some patient detail)
https://www.icpsr.umich.edu/web/ICPSR/series/41
L. National Hospital Ambulatory Medical Care Survey Series
https://www.icpsr.umich.edu/web/ICPSR/series/42
M. National Hospital Discharge Survey Series (nonfederal, short-term stay hospital sample of discharges)
https://www.icpsr.umich.edu/web/ICPSR/series/43
N. National Longitudinal Study of Adolescent to Adult Health (Add Health) Series (initial sample of 12-18 year-olds re-interviewed on their health until 24 to 32) *(QUITE POPULAR)*
https://www.icpsr.umich.edu/web/ICPSR/series/1006
O. National Longitudinal Surveys (NLS) (some health information on mothers/children collected, then followed subsequently)

https://www.icpsr.umich.edu/web/ICPSR/series/129
P. National Medical Expenditure Surveys (health expenditures by families and individuals; 1977, 1987) [[also available from IPUMS https://healthsurveys.ipums.org/]]
https://www.icpsr.umich.edu/web/ICPSR/series/45
Q. National Nursing Home Survey (1969 periodically to 2004, nursing homes and residents, cross section samples)
https://www.icpsr.umich.edu/web/ICPSR/series/47
R. National Social Life, Health, and Aging Project (NSHAP) (longitudinal data through NORC (U of Chicago), focus on aging population with face-to-face interviews, biometric info, questionaries… need University faculty support and IRB approval for full access)
https://www.icpsr.umich.edu/web/ICPSR/series/706
S. Newly Licensed Registered Nurse Survey Series (new RNs attitudes about job, intentions, organizational commitment/satisfaction)
https://www.icpsr.umich.edu/web/ICPSR/series/869
T. Panel Study of Income Dynamics (PSID) (world's longest running nationally representative household panel survey, has all the basic individual socio-demographic stuff and occasional surveys also include questions on psychological functioning, fertility-related behavior, computer skills, health and disability)
https://www.icpsr.umich.edu/web/ICPSR/series/131
U. Study of Women's Health Across the Nation (SWAN) (Multi-site, longitudinal, epidemiological study designed to examine middle-aged women's health)
https://www.icpsr.umich.edu/web/ICPSR/series/253
V. Treatment Episode Data Set -- Admissions (TEDS-A) also a Discharge series (National census data system of annual admissions to substance abuse treatment facilities; TEDS collects drug-use admission data from the States on all admissions and discharges aged 12 or older.)
https://www.icpsr.umich.edu/web/ICPSR/series/56 admissions
https://www.icpsr.umich.edu/web/ICPSR/series/238 discharges

Integrated Public Use Microdata (IPUMS-CPS)-Current Population Survey (U of MN)
https://cps.ipums.org/cps-action/samples
Each month the Census Bureau samples the US population for basic socio-demographic, employment (including detailed industries and occupations), and for a quarter of those adults sampled, earnings. Supplemental surveys are regularly attached including fertility and marriage, disability, immunization (older years), and tobacco use. The "March Supplement" (old terminology) or *Annual Social and Economic Supplement (ASEC)* (new terminology) is the largest supplement, and has information on sources of income including receipt of workers compensation payments (from workplace injury or illness) and disability insurance receipts, as well as virtually all sources of income. The first screen shown, list those ASEC (March) supplements by year, and you can download several years of harmonized data just by checking the boxes of those years (separate cross sections of different, sampled individuals) you want to analyze. This is not a longitudinal data set, but pooled cross sections of random surveys. If you check the "BASIC MONTHLY" heading next to the ASEC heading, you can further get data from other monthly CPS surveys. "SUPPLEMENT TOPICS" offers information on the contents of the irregularly scheduled supplements to the monthly CPS survey.

Basic Monthly CPS surveys can also be downloaded from the United States Census Bureau website: https://www.census.gov/data/datasets/time-series/demo/cps/cps-basic.html

The ASEC supplements are also found at https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.html

Integrated Public Use Microdata (IPUMS-I)-International (Univ of Minnesota)
https://www.idhsdata.org/idhs/
(micro-data surveys from around the world, with some—not all—of the surveys containing information on healthcare facility used, disability (including broad groups like independent mobility difficulty or simply a 'work disability' or 'psychological disability'), health coverage, occupation (nurse, doctor), industry (hospital) but not in all data sets. Some search will likely be necessary to get the data you want from the specific country—data availability varies by country. Where relevant, data elements have been harmonized—i.e., they measure the same thing with the same scoring across countries.) See also as an alternative: https://dhsprogram.com/Data/ ; https://dhsprogram.com/data/available-datasets.cfm

Integrated Public Use Microdata (IPUMS-USA)-USA (Univ of Minnesota)
https://usa.ipums.org/usa/
This harmonizes current American Community Survey (ACS) with older Census data. The ACS is the largest household survey given by the Census bureau, gathering information contained previously only in the long form of the decennial census, such as ancestry, citizenship, educational attainment, income, language proficiency, migration, disability, employment, and housing characteristics (it also contains quite detailed socio-demographic information, place of work and travel time to work). It has detailed occupation and industry indicators to identify all types of healthcare personnel, health insurance coverage type, and type of disability if any.

National Sample Survey of Registered Nurses (NSSRN)
https://data.hrsa.gov/topics/health-workforce/nursing-workforce-survey-data#RegisteredNurses
(in 2018, nurse practitioners were also included in the sample)

United States Census Bureau, Statistical Abstract Series (aggregated data useful for time series and macro analyses, instead of the micro analyses-type data listed above: look under the "Births, Deaths" and "Health and Nutrition" sections)
https://www.census.gov/library/publications/time-series/statistical_abstracts.html

Other data bases (fees often charged or otherwise restricted) often available through University Libraries (so you can check yours), such as:

American Hospital Association (AHA) Annual Survey Database
Census of United States hospitals based on the AHA Annual Survey of Hospitals. Download survey data for analysis in Excel, SAS and other statistical software packages

Historical Statistics of the United States

https://www.data-archive.ac.uk/

For more coding examples for these and other statistical models see
Butler, Butler and Wilson, *Advanced Statistics Simplified with Health Examples*

EXAMPLE1: reading the National Registered Nurse (RN) Sample SAS. 1. Click on the link: https://data.hrsa.gov/topics/health-workforce/nursing-workforce-survey-data#RegisteredNurses
2. download the codebook and decide what variables you want to use in your analysis
3. After looking at the codebook, we decide to regress RN job satisfaction (after recode of the PN_SATISFD) with a 4=extremely satisfied, 3=moderately satisfied, 2=moderately unsatisfied, 1=extremely unsatisfied, scale, on numerous independent variables. Those determinates will be a burnout variable ((RE_LVE_BRNOUT) as a reason for contemplating leaving), annual wages (PN_EARN_PUF), telehealth used in the workplace (PN_TELHLTH), labor union (PN_UNION) and sociodemographic variables:

$$jobsat = \beta_0 + \beta_1 burnout + \beta_2 wages + \beta_3 telehealth + \beta_4 union + \beta_5 stuff + \mu$$

3a) estimate by OLS, because it is easy to interpret the coefficients but imposes the restriction that the effect of going from a 1 to 2 level of satisfaction is the same as going from a 3 to 4 level of satisfaction;
3b) also estimate by ordered logistic regression because it doesn't impose the same 'unit' effect restriction as OLS, but the predictor variable response is only interpretable for those going into the 4th level of satisfaction (or leaving the 1st level of satisfaction)
4. Extract the zipped data as necessary to where you want it on your computer/thumbdrive. Open up our version of SAS (PC SAS, UNIX, SASStudio, etc… the code below works), and write your code (its extension will be .sas, as 'RN2018.sas'). Frequently there are 'format' programs to give SAS variables formats (appropriate numeric or character denominations), and easy to identify names. Fill in the necessary requested locations at the beginning of such a format program, and run this file in your version of SAS. Then here is typical SAS code to read the SAS dataset (*.sas7bdat extension) and modify the variables according to our needs (we have a NSSRN_2018_PUF.sas7bdat sas data file in the RN2018 library as indicated):

```
/* following turns off SAS output delivery system */
/*  to get simple printed text in output window  */
ods listing;
ods html close;
ods graphics off;
libname RN2018 "C:\NSSRN2018_SAS_encoded_package";/*sets up RN2018 library*/

data RN; set RN2018.NSSRN_2018_PUF;  /* NSSRN_2018_PUF.sas7bdat is in RN0218 library*/
satisfied=5-PN_SATISFD; *reverses ordering of satisfaction variable;
if RE_LVE_BRNOUT=1 then burnout=1; else if RE_LVE_BRNOUT=2 then burnout=0;
*PN_EARN_PUF annual earnings in principal nursing job;
if PN_TELHLTH=1 then telehealth=1; else if PN_TELHLTH=2 then telehealth=0;
if PN_UNION=1 then union=1; else if PN_UNION=2 then union=0;
if SEX=1 then male=1; else male=0;
if MARITAL=1 then married=1; else married=0;
if MARITAL=2 then notmarried=1; else notmarried=0;
if MARITAL=3 then nevermarried=1; else nevermarried=0;
if DEP_CH6=1 then under6kids=1; else under6kids=0;
*AGE_PUF age of nurse in 2018;
if RAC_ETHN_PUF=1 then Hispanic=1; else Hispanic=0;
if RAC_ETHN_PUF=2 then nonHisp_white=1; else nonHisp_white=0;
```

```
if RAC_ETHN_PUF=3 then black=1; else black=0;
if RAC_ETHN_PUF=4 then Asian=1; else Asian=0;
*PN_LOC_ST_PUF state location of primary nursing position;
run;

title "OLS Satisfaction regression";
title2 "no state of work fixed effects";
proc hpreg data=rn;
model satisfied=burnout pn_earn_puf telehealth union male married notmarried
age_puf under6kids Hispanic nonHisp_white black Asian;
run;

title "OLS Satisfaction regression";
title2 "state of work fixed effects";
proc hpreg data=rn;
class pn_loc_st_puf;
model satisfied=burnout pn_earn_puf telehealth union male married notmarried
age_puf under6kids Hispanic nonHisp_white black Asian pn_loc_st_puf;
run;

title "Ordered Logistic Satisfaction regression";
title2 "no state of work fixed effects";
proc logistic data=rn descending;
model satisfied=burnout pn_earn_puf telehealth union male married notmarried
age_puf under6kids Hispanic nonHisp_white black Asian;
run;

title "Ordered Logistic Satisfaction regression";
title2 "state of work fixed effects";
proc logistic data=rn descending;
class pn_loc_st_puf;
model satisfied=burnout pn_earn_puf telehealth union male married notmarried
age_puf under6kids Hispanic nonHisp_white black Asian pn_loc_st_puf;
run;
```

In general, for the SAS code above: green are nonexecutable (comment) statements; blue are reserved words. Here is a partial listing of the regression output without state FEs:

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.036056	0.033642	90.25	<.0001
burnout	1	-0.289069	0.009057	-31.92	<.0001
PN_EARN_PUF	1	0.000000985	0.000000125	7.88	<.0001
telehealth	1	0.047515	0.009553	4.97	<.0001
union	1	-0.075092	0.013883	-5.41	<.0001
male	1	-0.031740	0.015893	-2.00	0.0458
married	1	0.037884	0.015076	2.51	0.0120
notmarried	1	-0.030815	0.018607	-1.66	0.0977
AGE_PUF	1	0.001747	0.000434	4.03	<.0001
under6kids	1	0.044120	0.013208	3.34	0.0008
HISPANIC	1	0.029023	0.032499	0.89	0.3718
nonHisp_white	1	0.019553	0.023761	0.82	0.4106
black	1	-0.073917	0.030676	-2.41	0.0160
Asian	1	-0.060145	0.033013	-1.82	0.0685

Partial output from the second regression with state fixed effects:

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.999275	0.046978	63.84	<.0001

burnout	1	-0.291734	0.009071	-32.16	<.0001
PN_EARN_PUF	1	0.000000987	0.000000128	7.74	<.0001
telehealth	1	0.046174	0.009619	4.80	<.0001
union	1	-0.084508	0.014666	-5.76	<.0001
male	1	-0.037033	0.015930	-2.32	0.0201
married	1	0.035549	0.015135	2.35	0.0188
notmarried	1	-0.034257	0.018661	-1.84	0.0664
AGE_PUF	1	0.001740	0.000436	3.99	<.0001
under6kids	1	0.042041	0.013213	3.18	0.0015
HISPANIC	1	0.031974	0.032842	0.97	0.3303
nonHisp_white	1	0.024608	0.024001	1.03	0.3052
black	1	-0.059683	0.031168	-1.91	0.0555
Asian	1	-0.053607	0.033272	-1.61	0.1072
PN_LOC_ST_PUF AL	1	0.072445	0.047915	1.51	0.1306
PN_LOC_ST_PUF AR	1	0.007685	0.046402	0.17	0.8685
PN_LOC_ST_PUF AZ	1	0.073146	0.045604	1.60	0.1087
PN_LOC_ST_PUF CA	1	0.041021	0.042859	0.96	0.3385
PN_LOC_ST_PUF CO	1	0.069142	0.045543	1.52	0.1290
PN_LOC_ST_PUF CT	1	0.013723	0.045400	0.30	0.7625
PN_LOC_ST_PUF D1	1	0.034070	0.042014	0.81	0.4174
PN_LOC_ST_PUF D4	1	0.024828	0.041949	0.59	0.5539
PN_LOC_ST_PUF D8	1	0.078047	0.041182	1.90	0.0581
PN_LOC_ST_PUF D9	1	0.051266	0.043071	1.19	0.2340
PN_LOC_ST_PUF DC	1	0.104604	0.047321	2.21	0.0271
PN_LOC_ST_PUF DE	1	0.000856	0.047254	0.02	0.9855
PN_LOC_ST_PUF FL	1	0.019486	0.045939	0.42	0.6714
PN_LOC_ST_PUF GA	1	0.057239	0.044272	1.29	0.1961
PN_LOC_ST_PUF IA	1	0.099136	0.046525	2.13	0.0331
PN_LOC_ST_PUF ID	1	0.105951	0.046924	2.26	0.0240
PN_LOC_ST_PUF IL	1	-0.028576	0.043326	-0.66	0.5095
PN_LOC_ST_PUF IN	1	0.027349	0.044788	0.61	0.5415
PN_LOC_ST_PUF KS	1	0.079389	0.046979	1.69	0.0911
PN_LOC_ST_PUF KY	1	0.001809	0.046618	0.04	0.9690
PN_LOC_ST_PUF LA	1	0.005655	0.048714	0.12	0.9076
PN_LOC_ST_PUF MA	1	0.019731	0.043672	0.45	0.6514
PN_LOC_ST_PUF MD	1	-0.005298	0.042279	-0.13	0.9003
PN_LOC_ST_PUF ME	1	0.006135	0.045609	0.13	0.8930
PN_LOC_ST_PUF MI	1	0.063394	0.043507	1.46	0.1451
PN_LOC_ST_PUF MN	1	0.115755	0.046384	2.50	0.0126
PN_LOC_ST_PUF MO	1	-0.022017	0.046503	-0.47	0.6359
PN_LOC_ST_PUF MS	1	-0.005404	0.046756	-0.12	0.9080
PN_LOC_ST_PUF NC	1	0.034811	0.044585	0.78	0.4349
PN_LOC_ST_PUF NE	1	0.028139	0.047829	0.59	0.5563
PN_LOC_ST_PUF NH	1	0.018345	0.045528	0.40	0.6870
PN_LOC_ST_PUF NJ	1	-0.019733	0.044752	-0.44	0.6593
PN_LOC_ST_PUF NM	1	0.043933	0.043201	1.02	0.3092
PN_LOC_ST_PUF NV	1	0.087387	0.046540	1.88	0.0604
PN_LOC_ST_PUF NY	1	0.013487	0.043555	0.31	0.7568
PN_LOC_ST_PUF OH	1	0.027477	0.044582	0.62	0.5377
PN_LOC_ST_PUF OK	1	0.036352	0.046230	0.79	0.4317
PN_LOC_ST_PUF OR	1	0.124201	0.046354	2.68	0.0074
PN_LOC_ST_PUF PA	1	-0.086156	0.042629	-2.02	0.0433
PN_LOC_ST_PUF SC	1	0.055574	0.048306	1.15	0.2500
PN_LOC_ST_PUF TN	1	0.063933	0.043955	1.45	0.1458
PN_LOC_ST_PUF TX	1	0.009503	0.046456	0.20	0.8379
PN_LOC_ST_PUF UT	1	0.122696	0.043530	2.82	0.0048
PN_LOC_ST_PUF VA	1	0.043976	0.043018	1.02	0.3067
PN_LOC_ST_PUF WA	1	0.070752	0.044244	1.60	0.1098
PN_LOC_ST_PUF WI	1	0.020470	0.044218	0.46	0.6434
PN_LOC_ST_PUF WV	0	0	.	.	.

Partial listing of the ordered logistic results in SAS (with 4 ordered categories there are three intercepts estimated)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard Estimate	Error	Wald Chi-Square	Pr > ChiSq
Intercept 4	1	-1.4153	0.1075	173.4582	<.0001
Intercept 3	1	1.7262	0.1079	255.9138	<.0001
Intercept 2	1	3.8113	0.1182	1039.4637	<.0001
burnout	1	-0.9521	0.0301	1003.0625	<.0001
PN_EARN_PUF	1	3.365E-6	3.96E-7	72.1970	<.0001
telehealth	1	0.1461	0.0303	23.2681	<.0001
union	1	-0.2303	0.0443	27.0804	<.0001
male	1	-0.0815	0.0505	2.6063	0.1064
married	1	0.0996	0.0480	4.3018	0.0381
notmarried	1	-0.1059	0.0592	3.1947	0.0739
AGE_PUF	1	0.00724	0.00138	27.5614	<.0001
under6kids	1	0.1569	0.0420	13.9841	0.0002
HISPANIC	1	0.0663	0.1031	0.4133	0.5203
nonHisp_white	1	0.0470	0.0754	0.3882	0.5332
black	1	-0.2628	0.0976	7.2539	0.0071
Asian	1	-0.2240	0.1051	4.5454	0.0330

Here is a partial listing of the ordered logistic results in SAS, but with the state fixed effects also included (with 4 ordered categories there are three intercepts estimated):

Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard Estimate	Error	Wald Chi-Square	Pr > ChiSq
Intercept 4	1	-1.4170	0.1082	171.5469	<.0001
Intercept 3	1	1.7367	0.1086	255.5428	<.0001
Intercept 2	1	3.8242	0.1189	1034.6591	<.0001
burnout	1	-0.9640	0.0302	1020.8292	<.0001
PN_EARN_PUF	1	3.392E-6	4.049E-7	70.1754	<.0001
telehealth	1	0.1418	0.0306	21.5362	<.0001
union	1	-0.2555	0.0468	29.7707	<.0001
male	1	-0.1004	0.0507	3.9271	0.0475
married	1	0.0916	0.0483	3.5943	0.0580
notmarried	1	-0.1183	0.0595	3.9545	0.0467
AGE_PUF	1	0.00724	0.00139	27.2295	<.0001
under6kids	1	0.1518	0.0421	13.0309	0.0003
HISPANIC	1	0.0737	0.1044	0.4977	0.4805
nonHisp_white	1	0.0642	0.0763	0.7076	0.4003
black	1	-0.2159	0.0993	4.7258	0.0297
Asian	1	-0.1984	0.1061	3.4970	0.0615
PN_LOC_ST_PUF AL	1	0.1068	0.1088	0.9628	0.3265
PN_LOC_ST_PUF AR	1	-0.1061	0.1030	1.0603	0.3032
PN_LOC_ST_PUF AZ	1	0.1247	0.0983	1.6071	0.2049
PN_LOC_ST_PUF CA	1	-0.00637	0.0851	0.0056	0.9404
PN_LOC_ST_PUF CO	1	0.1402	0.0985	2.0250	0.1547
PN_LOC_ST_PUF CT	1	-0.0697	0.0981	0.5053	0.4772
PN_LOC_ST_PUF D1	1	-0.0220	0.0823	0.0713	0.7895
PN_LOC_ST_PUF D4	1	-0.0617	0.0825	0.5599	0.4543
PN_LOC_ST_PUF D8	1	0.1175	0.0781	2.2633	0.1325
PN_LOC_ST_PUF D9	1	0.0525	0.0862	0.3707	0.5426
PN_LOC_ST_PUF DC	1	0.1891	0.1058	3.1948	0.0739
PN_LOC_ST_PUF DE	1	-0.1207	0.1066	1.2829	0.2574
PN_LOC_ST_PUF FL	1	-0.0117	0.1002	0.0137	0.9067
PN_LOC_ST_PUF GA	1	0.0509	0.0927	0.3020	0.5826
PN_LOC_ST_PUF IA	1	0.2025	0.1029	3.8762	0.0490
PN_LOC_ST_PUF ID	1	0.2669	0.1044	6.5298	0.0106
PN_LOC_ST_PUF IL	1	-0.2047	0.0887	5.3269	0.0210
PN_LOC_ST_PUF IN	1	-0.0486	0.0956	0.2579	0.6116
PN_LOC_ST_PUF KS	1	0.1084	0.1047	1.0721	0.3005
PN_LOC_ST_PUF KY	1	-0.0911	0.1037	0.7727	0.3794
PN_LOC_ST_PUF LA	1	-0.1252	0.1128	1.2317	0.2671

PN_LOC_ST_PUF MA	1	-0.0720	0.0901	0.6384	0.4243
PN_LOC_ST_PUF MD	1	-0.1457	0.0835	3.0492	0.0808
PN_LOC_ST_PUF ME	1	-0.1121	0.0996	1.2661	0.2605
PN_LOC_ST_PUF MI	1	0.0547	0.0894	0.3738	0.5409
PN_LOC_ST_PUF MN	1	0.2622	0.1021	6.6022	0.0102
PN_LOC_ST_PUF MO	1	-0.1980	0.1034	3.6678	0.0555
PN_LOC_ST_PUF MS	1	-0.1250	0.1044	1.4321	0.2314
PN_LOC_ST_PUF NC	1	0.0275	0.0944	0.0846	0.7712
PN_LOC_ST_PUF NE	1	-0.0450	0.1090	0.1703	0.6799
PN_LOC_ST_PUF NH	1	-0.0575	0.0989	0.3382	0.5609
PN_LOC_ST_PUF NJ	1	-0.2129	0.0949	5.0273	0.0250
PN_LOC_ST_PUF NM	1	0.00700	0.0877	0.0064	0.9364
PN_LOC_ST_PUF NV	1	0.1295	0.1025	1.5942	0.2067
PN_LOC_ST_PUF NY	1	-0.1023	0.0892	1.3130	0.2519
PN_LOC_ST_PUF OH	1	-0.0138	0.0946	0.0213	0.8839
PN_LOC_ST_PUF OK	1	-0.00685	0.1015	0.0046	0.9462
PN_LOC_ST_PUF OR	1	0.2724	0.1016	7.1884	0.0073
PN_LOC_ST_PUF PA	1	-0.3951	0.0860	21.1253	<.0001
PN_LOC_ST_PUF SC	1	0.0421	0.1106	0.1448	0.7035
PN_LOC_ST_PUF TN	1	0.0935	0.0916	1.0413	0.3075
PN_LOC_ST_PUF TX	1	-0.0861	0.1025	0.7049	0.4011
PN_LOC_ST_PUF UT	1	0.2936	0.0892	10.8340	0.0010
PN_LOC_ST_PUF VA	1	0.0102	0.0869	0.0139	0.9062
PN_LOC_ST_PUF WA	1	0.1125	0.0923	1.4865	0.2228
PN_LOC_ST_PUF WI	1	-0.0727	0.0931	0.6098	0.4348

Output will be saved as a '*.lst' extension file, as in RN2018.lst. All four specifications for this sample indicate that increasing age and earnings are associated with increases in job satisfaction, as does the presence of telehealth, while job burnout and unions are associated with lower levels of job satisfaction. Looking at the coefficients for the state fixed effects (dummy variables for each state) indicates Utah has the most satisfied nurses while Pennsylvania has the least satisfied nurses. For more on generating and interpreting results such as these, see Butler, Butler and Wilson *Advanced Statistics Simplified with Health Examples*.

EXAMPLE2: reading the National Registered Nurse (RN) Sample in STATA. 1. Click link: https://data.hrsa.gov/topics/health-workforce/nursing-workforce-survey-data#RegisteredNurses
2. download the codebook and decide what variables you want to use in your analysis
3. As before with SAS, after looking at the codebook, we decide to regress RN job satisfaction (after recode of the PN_SATISFD) with a 4=extremely satisfied, 3=moderately satisfied, 2=moderately unsatisfied, 1=extremely unsatisfied, scale, on numerous independent variables. Those determinates will be a burnout variable ((RE_LVE_BRNOUT) as a reason for contemplating leaving), annual wages (PN_EARN_PUF), telehealth used in the workplace (PN_TELHLTH), labor union (PN_UNION) and sociodemographic variables:

$$jobsat = \beta_0 + \beta_1 burnout + \beta_2 wages + \beta_3 telehealth + \beta_4 union + \beta_5 stuff + \mu$$

3a) estimate by OLS, because it is easy to interpret the coefficients but imposes the restriction that the effect of going from a 1 to 2 level of satisfaction is the same as going from a 3 to 4 level of satisfaction;
3b) also estimate by ordered logistic regression because it doesn't impose the same 'unit' effect restriction as OLS, but the predictor variable response is only interpretable for those going into the 4th level of satisfaction (or leaving the 1st level of satisfaction)
4. Extract the zipped STATA data necessary to where you want it on your computer or thumbdrive. Create your STATA executable code (in window version, find "window" column,

then go down to the "do-file" editor, and choose "new" if doing the program from scratch, or—since it is a standard text format, it can be done in notepad, notepad++, etc) with executable STATA having an extension of *.do, as in 'RN2018.do'—this extension is provided automatically in STATA. Two notes about STATA code relative to SAS code: 1) STATA is case sensitive so the variable 'AGE' is not equal to the variable 'age', and 2) the '=' sign in math and much of programming serves two functions: a) assignment of value, and b) test of equivalence. In STATA, '=' is the assignment of value, while '==' is test of equivalence. With your executable file in STATA (hit "file" on upper left and then "open" up your executable do file if its not already open) and run it:

```
use "C:\Users\rjb99\Documents\healthtrics\online data\RN data\NSSRN2018_Stata_encoded_package\NSSRN_2018_PUF.dta", clear

# delimit ; /*this makes the ';' the delimiter for each command, like SAS */
/* stata uses double equal signs for equivalence, and single equal sign for assignment*/
gen satisfied=5-PN_SATISFD; *reverses ordering of satisfaction variable;
gen burnout=1 if RE_LVE_BRNOUT==1 ;
replace burnout=0 if RE_LVE_BRNOUT==2 ;
gen male = (SEX==1) & !missing(SEX);
/* could do the same with these three commands, the third to handle missing values */
/*          generate male=1 if A_SEX==1;     */
/*          replace male=0 if A_SEX==2;      */
/*          repace male=. if missing(A_SEX);  */
*PN_EARN_PUF annual earnings in principal nursing job;
gen telehealth=1 if PN_TELHLTH==1;
replace telehealth=0 if PN_TELHLTH==2;
gen union=1 if PN_UNION==1;
replace union=0 if PN_UNION==2;
gen married=(MARITAL==1);
gen notmarried=(MARITAL==2);
gen nevermarried=(MARITAL==3);
gen under6kids= (DEP_CH6==1);
*AGE_PUF age of nurse in 2018;
gen Hispanic=(RAC_ETHN_PUF==1);
gen nonHisp_white= (RAC_ETHN_PUF==2);
gen black=(RAC_ETHN_PUF==3);
gen Asian=(RAC_ETHN_PUF==4);
*PN_LOC_ST_PUF state location of primary nursing position,PN_LOC_CODE_PUF;
/* STATA is case sensitive unlike SAS so age does not equal AGE*/

regress satisfied burnout PN_EARN_PUF telehealth union male married notmarried
AGE_PUF under6kids Hispanic nonHisp_white black Asian;

/* the areg with absorb(*) option does fixed effects with string char*/
areg satisfied burnout PN_EARN_PUF telehealth union male married notmarried
AGE_PUF under6kids Hispanic nonHisp_white black Asian, absorb(PN_LOC_CODE_PUF);

ologit satisfied burnout PN_EARN_PUF telehealth union male married notmarried
AGE_PUF under6kids Hispanic nonHisp_white black Asian;
```

There was no way to use state FEs without reformatted our string state identifiers as numeric variables. The results of the estimated coefficients were the same as the SAS output above.

EXAMPLE3: reading the National RN Sample in R. After reading in the ASCII format from the 2018 RN data site used in examples 1 and 2, we found it did not have the variables attached in the top row, as is usually the case (with ASCII, txt, downloads). So, for the R example, we read in the STATA data set using R, then do the same steps as above after reading the 2018 RN data set, which can be done in many ways, including installing the haven library. We download and install RSTUDIO on our computer (see chapter 1 in Butler, Butler, and Wilson *Advanced*

Statistics Simplified for Health Professionals), and then install a R package that reads STATA into R, like haven (`>install.packages("haven")),` download the STATA data set into R, and then change the variables and run the analysis using one of the fixed effects regression packages `(FEreg <- lm(satisfied ~ burnout, N_EARN_PUF, telehealth, union, male, married, notmarried, AGE_PUF, under6kids, Hispanic, nonHisp_white, black, Asian, factor(PN_LOC_CODE_PUF), data=<<…>>))` and one of the ordered logistics regression package `(FEreg <- polr(satisfied ~ burnout, N_EARN_PUF, telehealth, union, male, married, notmarried, AGE_PUF, under6kids, Hispanic, nonHisp_white, black, Asian, data=<<…>>).`

EAMPLES 4 through 6. In the RN 2018 of examples 1 through 3, there was only the option of downloading all the data. More typically when working with webpages, you can choose what variables to include in the downloaded data sets, and often what observations as well (restrict the sample to certain years, ages of those surveyed, for example, or from certain states). In what follows, we only download data from such (more typical) websites. In EAMPLE 4, we download data from the Current Population Survey using an IPUMS portal.

EXAMPLE4: reading the Annual Social and Economic Supplement (ASEC) in SAS. To get this ASEC annual supplement from the Current Population Survey (CPS), we go the IPUMS portal and click off all but the most recent supplements: https://cps.ipums.org/cps-action/samples

Then we go to the bottom of that page and click "SUBMIT SAMPLE SELECTIONS". We want to look at the some of the socio-demographic determinants of a workers compensation claim (these are associated with workplace injuries or disease), we decide get STATEFIP (state numeric identifiers) from the HOUSEHOLD_geograhpic variables category, and from the PERSON category, in the CORE section, we choose AGE, SEX, RACE, MARST, EDUC; and from the ANNUAL SOCIAL ECONOMIC SUPPLEMENT (ASEC), we click on the INCOME VARIABLES to get INCWKCOM (income from workers' compensation), INCWAGE (individual wage and salary income), and FTOTVAL (total family income). Go with the default extraction options, and submit your data request (registering with your email and password of your choice, if you have not previously done this).

$$WC\_amt = \beta_0 + \beta_1 wage + \beta_2 income + \beta_3 school + \beta_4 age + \beta_5 stuff + \mu$$
$$WC\_YN = G(\alpha_0 + \alpha_1 wage + \alpha_2 income + \alpha_3 school + \alpha_4 age + \alpha_5 stuff)$$

On IPUMS "DOWNLOAD OR REVISE EXTRACTS" page the green-highlighted Download_DAT column is the text file ONCE the data request has been processed, and to its immediate right at the SAS, Stata, and R code to download the data into those data types. First, let's download the "SAS" code that creates and formats the ASCII (or text) file, into a "CPS_00011.sas7bdat" located in the libname "CPS2021". So the SAS code for our workers compensation model is:

```
/* following turns off SAS output delivery system  */
/*for simple printed text in output window instead */
ods listing;
ods html close;
ods graphics off;
libname cps2021 "C:\Users\rjb99\cps2021_WC";

data one; set cps2021.cps_00011;  /*cps_00011.sas7bdat in the cps2021 library*/
```

```
/* from the codebook provided with the data downloads, adjust the following* /
if (18<=age<=65); *only include those in analysis between 18 and 65 years of age;
if ((INCWKCOM=0) or (INCWKCOM=.)) then do; INCWKCOM=.; WC_YN=0;
            end;
else WC_YN=1; *those who get positive income reported work injury recently;
if FTOTVAL=0 then FTOTVAL=.;
if INCWAGE=0 then INCWAGE=.;
if sex=1 then male=1; else male=0;
if sex=9 then male=.;
if race=100 then white=1; else white=0;
if race=200 then black=1; else black=0;
if Marst in (1,2) then married=1; else married=0;
/*alternative: if ((marst=1) or (marst=2)) then married=1; else...*/
if marst=6 then never_married=1; else never_married=0;
if educ in (1,999) then school=.;
if educ=2 then school=0; *recode so school=years of school completed;
if educ in (10,14) then school=4;
if educ=11 then school=1;
if educ=12 then school=2;
if educ=13 then school=3;
if educ=14 then school=4;
if ((educ=21) or (educ=20)) then school=5;
if educ=22 then school=6;
if educ in (30,31) then school=7;
if educ=32 then school=8;
if educ=40 then school=9;
if educ=50 then school=10;
if educ=60 then school=11;
if educ in (70,71,72,73) then school=12;
if educ=80 then school=13;
if educ=81 then school=14;
if educ in (90,91,92) then school=14;
if educ=100 then school=15;
if educ in (110,111) then school=16;
if educ in (120,121) then school=17;
if educ in (122, 123,124) then school=18;
if educ=125 then school=20;
run;

title "amount of WC payments received given you received any";
proc hpreg data=one;
class STATEFIP;
model INCWKCOM=incwage ftotval school age male white black married never_married
statefip;
run;

title "likelihood of receiving WC benefits";
proc logistic data=one;
class statefip;
model WC_YN (event='1')=incwage ftotval school age male white black married never_married statefip;
run;
```

Amount of WC income received among only those receiving such income, OLS partial listing includes:

Parameter	DF	Standard Estimate	Error	t Value	Pr > |t|
Intercept	1	2561.757850	3586.751162	0.71	0.4753
INCWAGE	1	-0.086887	0.014817	-5.86	<.0001
FTOTVAL	1	0.073562	0.006608	11.13	<.0001

school	1	77.832866	154.627103	0.50	0.6148
AGE	1	59.830567	28.489900	2.10	0.0360
male	1	2702.168012	691.641906	3.91	0.0001
white	1	954.971808	1358.363622	0.70	0.4822
black	1	1507.011636	1670.750644	0.90	0.3673
married	1	-2334.297817	888.434947	-2.63	0.0088
never_married	1	-1725.328230	1074.700204	-1.61	0.1088
STATEFIP Alabama	1	-2420.770774	3691.807919	-0.66	0.5122
STATEFIP Alaska	1	-5854.538357	2890.291976	-2.03	0.0431
STATEFIP Arizona	1	-4143.524052	3264.834357	-1.27	0.2048
STATEFIP Arkansas	1	-3037.814096	3451.060169	-0.88	0.3790
STATEFIP California	1	-677.087309	2326.516703	-0.29	0.7711
STATEFIP Colorado	1	-5557.964192	. . . . .		

Receving WC income, yes(=1)/no(=0), the results from a logistic regression:

Parameter		DF	Standard Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.4043	0.2760	383.3479	<.0001
INCWAGE		1	-0.00001	1.465E-6	76.2269	<.0001
FTOTVAL		1	-1.53E-7	5.813E-7	0.0693	0.7924
school		1	-0.0550	0.0126	19.0021	<.0001
AGE		1	0.0184	0.00310	35.0746	<.0001
male		1	0.6781	0.0736	84.9651	<.0001
white		1	0.4885	0.1457	11.2367	0.0008
black		1	0.5186	0.1798	8.3224	0.0039
married		1	-0.1449	0.0956	2.2982	0.1295
never_married		1	-0.5227	0.1235	17.9248	<.0001
STATEFIP	Alabama	1	-0.3021	0.3305	0.8354	0.3607
STATEFIP	Alaska	1	0.7887	0.2199	12.8636	0.0003
STATEFIP	Arizona	1	0.0851	0.2759	0.0951	0.7578
STATEFIP	Arkansas	1	-0.0573	0.2994	0.0366	0.8482
STATEFIP	California	1	0.0777	0.1157	0.4504	0.5021
STATEFIP	Colorado	1	-0.3116	0.2990	1.0861	0.2973
STATEFIP	Connecticut	1	0.4901	0.2188	5.0196	0.0251
STATEFIP	Delaware	1	-0.3626	0.3500	1.0732	0.3002

……..

Higher wages are associated with less benefits received (the OLS results from the hpreg) and a lower likelihood of receiving any benefits (the logistic results) since the opportunity cost of being on a claim increases with wage income; more schooling decreases the likelihood of receiving workers' compensation but has no effect on the amount received, holding wages constant, and older people file more claims and get more money, other things equal.

EXAMPLE5: reading Annual Social and Economic Supplement (ASEC) in STATA. To get this ASEC annual supplement from the Current Population Survey (CPS), we go the IPUMS portal and click off all but the most recent supplements: https://cps.ipums.org/cps-action/samples

Then we go to the bottom of that page and click "SUBMIT SAMPLE SELECTIONS". We want to look at the some of the socio-demographic determinants of a workers compensation claim (these are associated with workplace injuries or disease), we decide get STATEFIP (state numeric identifiers) from the HOUSEHOLD_geograhpic variables category, and from the PERSON category, in the CORE section, we choose AGE, SEX, RACE, MARST, EDUC; and from the ANNUAL SOCIAL ECONOMIC SUPPLEMENT (ASEC), we click on the INCOME VARIABLES to get INCWKCOM (income from workers' compensation), INCWAGE

(individual wage and salary income), and FTOTVAL (total family income). Go with the default extraction options, and submit your data request (registering with your email and password of your choice, if you have not previously done this).

$$WC\_amt = \beta_0 + \beta_1 wage + \beta_2 income + \beta_3 school + \beta_4 age + \beta_5 stuff + \mu$$
$$WC\_YN = G(\alpha_0 + \alpha_1 wage + \alpha_2 income + \alpha_3 school + \alpha_4 age + \alpha_5 stuff)$$

On IPUMS "DOWNLOAD OR REVISE EXTRACTS" page the green-highlighted Download_DAT column is the text file ONCE the data request has been processed, and to its immediate right at the SAS, STATA, and R code to download the data into those data types. Example 2 for nurses had code that created the *.dta STATA file on your specified location, but here the code reads in the ASCII file for our study and formats, but leaves it in STATA memory. So, for these types of files: A) Run the format code provided by IPUMS, and you will have your STATA data set in memory (be sure to indicate where the *.dat file generated by IPUMS code is located), then your data will be in memory. B) So click **file**, then **save** to save the STATA *.dta file somewhere on your drive or other storage device, so you can bring it up, edit the variable values, and then getting your regressions above. C) Finally, Run the STATA program you write below by cut and pasting into do-file editor (find "window" column, then go down to the "do-file editor', and choose "new" if doing the program from scratch, or choose a pre-existing do file to edit—also, since it is a standard text format, it can be done in notepad, notepad++, etc) with executable STATA having an extension of *.do, as in 'CPS2021.do'. So the STATA code for our workers compensation model is:

```
use "C:\Users\rjb99\Documents\healthtrics\online data\CPS data\CPS2021_WC.dta", clear

# delimit ;
* the previous command, delimit, means that all commands end with a semicolon;
* sentences that start like this with a * are non-executable, comment statements;
*  comments can also be set off with the  /*  stuff here  */ type of commenting;
/* this is example of this alternative way to comment */
gen wc_yn = 1 if ((incwkcom > 0) );
replace wc_yn= 0 if incwkcom==0;
keep if age > 17;
keep if age < 66;
drop if ftotval==0;
drop if incwage==0;
replace incwkcom=. if incwkcom==0;
gen male = (sex==1) & !missing(sex);
/* could do the same with these three commands, the third to handle missing values  */
/*          generate male=1 if sex==1;     */
/*          replace male=0 if sex==2;      */
/*          repace male=. if missing(sex);  */
gen white = (race==100) & !missing(race);
gen black = (race==200) & !missing(race);

/* dummy variables for educational attainment follow, based on CPS codes */

**********EDUCATIONAL ATTAINMENT VARIABLES FOLLOW **********;
gen school=0 if (educ==2) & !missing(educ);
replace school=4 if ((educ==10) | (educ==14)) & !missing(educ);
replace school=1 if ((educ==11)) & !missing(educ);
replace school=2 if (educ==34) & !missing(educ);
replace school=3 if (educ==13) & !missing(educ);
```

```
replace school=4 if ((educ==14)) & !missing(educ);
replace school=5 if ((educ==20) | (educ==21)) & !missing(educ);
replace school=6 if (educ==22) & !missing(educ);
replace school=7 if ((educ==30) | (educ==31)) & !missing(educ);
replace school=8 if (educ==32) & !missing(educ);
replace school=9 if (educ==40) & !missing(educ);
replace school=10 if (educ==50) & !missing(educ);
replace school=11 if (educ==60) & !missing(educ);
replace school=12 if ((educ==70) | (educ==71) | (educ==72) | (educ==73)) & !missing(educ);
replace school=13 if (educ==80) & !missing(educ);
replace school=14 if ((educ==81) | (educ==90) | (educ==91) | (educ==92))& !missing(educ);
replace school=15 if (educ==100) & !missing(educ);
replace school=16 if ((educ==110)| (educ==110)) & !missing(educ);
replace school=18 if ((educ==122) | (educ==123) )& !missing(educ);
replace school=20 if (educ==125) & !missing(educ);

gen married= ((marst==1) | (marst==2) ) & !missing(marst);
gen nevermarried=((marst==6) ) & !missing(marst); ;

regress incwkcom incwage ftotval school age male white black married nevermarried
i.statefip;

logit wc_yn incwage ftotval school age male white black married nevermarried
i.statefip;
```

With the output same as SAS….

https://cran.r-project.org/web/packages/ipumsr/vignettes/ipums-cps.html

EXAMPLE6: reading Annual Social and Economic Supplement (ASEC) in R. To get this ASEC annual supplement from the Current Population Survey (CPS), we go the IPUMS portal and click off all but the most recent supplements: https://cps.ipums.org/cps-action/samples

Then we go to the bottom of that page and click "SUBMIT SAMPLE SELECTIONS". We want to look at the some of the socio-demographic determinants of a workers compensation claim (these are associated with workplace injuries or disease), we decide get STATEFIP (state numeric identifiers) from the HOUSEHOLD_geograhpic variables category, and from the PERSON category, in the CORE section, we choose AGE, SEX, RACE, MARST, EDUC; and from the ANNUAL SOCIAL ECONOMIC SUPPLEMENT (ASEC), we click on the INCOME VARIABLES to get INCWKCOM (income from workers' compensation), INCWAGE (individual wage and salary income), and FTOTVAL (total family income). Go with the default extraction options, and submit your data request (registering with your email and password of your choice, if you have not previously done this).

$$WC\_amt = \beta_0 + \beta_1 wage + \beta_2 income + \beta_3 school + \beta_4 age + \beta_5 stuff + \mu$$
$$WC\_YN = G(\alpha_0 + \alpha_1 wage + \alpha_2 income + \alpha_3 school + \alpha_4 age + \alpha_5 stuff)$$

On IPUMS "DOWNLOAD OR REVISE EXTRACTS" page the green-highlighted Download_DAT column that is the text file ONCE the data request has been processed, and to its immediate right at the SAS, STATA, and R code to download the data into those data types.

Suppose that the data is named 'cps_00011.dat' and it is located on the "C:\Users\rjb99\Documents\healthtrics\online data\CPS data\cps_00011.dat" subdirectory; also we will want to download the DDI file, which will be named 'cps_000aa.xml' on the "C:\Users\rjb99\Documents\healthtrics\online data\CPS data" subdirectory.
We will also need to download the ipumsr R-package: install.packages('ipumsr')

```
> install.packages('ipumsr')
> library(ipumsr)
# Change these filepaths to the filepaths of your downloaded extract
>cps_ddi_file <- "<<file_location\cps_00011.xml>>"
>cps_data_file <- "<<file_location\cps_00011.dat>>"
>cps_ddi <- read_ipums_ddi(cps_ddi_file) # Contains metadata, as separate object
>cps2021_WC <- read_ipums_micro(cps_ddi_file, data_file = cps_data_file)
>filter(cps2021_WC, age>17)
>filter(cps2021_WC, age<66)
>cps2021_WC$WC_YN <- ifelse(cps2021_WC$incwkcom>0, 1 , 0)
>cps2021_WC$male <- ifelse(cps2021_WC$sex==1, 1, 0)
>cps2021_WC$male <- if(cps2021_WC$sex==9, NA)
>cps2021_WC$white <- ifelse(cps2021_WC$race==100, 1, 0)
>cps2021_WC$black <- ifelse(cps2021_WC$race==200, 1, 0)
>cps2021_WC$married <- ifelse(((cps2021_WC$marst==1) |(cps2021_WC$marst==2)), 1, 0)
>cps2021_WC$nevermarried <- ifelse(cps2021_WC$marst==6, 1, 0)
>cps2021_WC$school <- if(((cps2021_WC$educ==1) | (cps2021_WC$educ==999), NA)
>cps2021_WC$school <- if(cps2021_WC$educ==2, 0)
>cps2021_WC$school <- if(cps2021_WC$educ==10, 4)
>cps2021_WC$school <- if(cps2021_WC$educ==14, 4)
>cps2021_WC$school <- if(cps2021_WC$educ==11, 1)
>cps2021_WC$school <- if(cps2021_WC$educ==12, 2)
>cps2021_WC$school <- if(cps2021_WC$educ==13, 3)
>cps2021_WC$school <- if(cps2021_WC$educ==14, 4)
>cps2021_WC$school <- if(cps2021_WC$educ==20, 5)
>cps2021_WC$school <- if(cps2021_WC$educ==21, 5)
>cps2021_WC$school <- if(cps2021_WC$educ==22, 6)
>cps2021_WC$school <- if(cps2021_WC$educ==30, 7)
>cps2021_WC$school <- if(cps2021_WC$educ==31, 7)
>cps2021_WC$school <- if(cps2021_WC$educ==32, 8)
>cps2021_WC$school <- if(cps2021_WC$educ==40, 9)
>cps2021_WC$school <- if(cps2021_WC$educ==50, 10)
>cps2021_WC$school <- if(cps2021_WC$educ==60, 11)
>cps2021_WC$school <- if(cps2021_WC$educ==70, 12)
>cps2021_WC$school <- if(cps2021_WC$educ==71, 12)
>cps2021_WC$school <- if(cps2021_WC$educ==72, 12)
>cps2021_WC$school <- if(cps2021_WC$educ==73, 12)
>cps2021_WC$school <- if(cps2021_WC$educ==80, 13)
>cps2021_WC$school <- if(cps2021_WC$educ==81, 14)
>cps2021_WC$school <- if(cps2021_WC$educ==90, 14)
>cps2021_WC$school <- if(cps2021_WC$educ==91, 14)
>cps2021_WC$school <- if(cps2021_WC$educ==92, 14)
>cps2021_WC$school <- if(cps2021_WC$educ==100, 15)
>cps2021_WC$school <- if(cps2021_WC$educ==110, 16)
>cps2021_WC$school <- if(cps2021_WC$educ==111, 16)
>cps2021_WC$school <- if(cps2021_WC$educ==120, 17)
>cps2021_WC$school <- if(cps2021_WC$educ==121, 17)
>cps2021_WC$school <- if(cps2021_WC$educ==122, 18)
>cps2021_WC$school <- if(cps2021_WC$educ==123, 18)
>cps2021_WC$school <- if(cps2021_WC$educ==124, 18)
>cps2021_WC$school <- if(cps2021_WC$educ==111, 16)
```

```
>cps2021_WC$school <- if(cps2021_WC$educ==125, 20)
>
>wc_benefits = lm(incwkkcom ~ incwage + ftotval + school + age + male + white +
black + married + never_married + factor(statefip), data = cps2021_WC)
>#create a linear regression with the rhs variables and state effects
>summary(wc_benefits) #Review the results
>
>WClogit <- glm(wc_YN ~ incwage + ftotval + school + age + male + white +
black + married + never_married + factor(statefip), data = cps2021_WC, family = "binomial")
>summary(WClogit)
```

EXAMPLE 7. Generic ASCII (text) files with the names in the first row of the downloaded data set. Usually you have to enter your email address, pick out the data (including the particular set of data including what variables you want from that data), request the data by clicking something, and then a notice is sent when the data is ready (perhaps an email sent to you, perhaps a notice pops up on your view of the portal).

Let's suppose you choose, and will choose, the following variables: STATE, EDUC, SEX, WSALVAL, and AGE . At this point, you may want to restrict your download to a particular state or a particular age group (say, 18 to 65 if you are analyzing wages of the 'working-age' population. CODEBOOK: Save the variable names with the code values in a codebook available just for the variables and data set you have chosen.

Suppose that the data file you downloaded is an *.asc file (space delimited, say my_data.asc). You may also want to keep a copy of the record layout file, so you remember which variables are in which columns (or just cut and paste them off the *.asc data file). Open the asc data file in a text editor and strip off the top line of data with the variable names into their order, and then paste those variables into your SAS or STATA "infile" statements of your program. Below are SAS and STATA examples of reading the data, which has been downloaded to "D:…" thumbdrive:

STATA

```
# delimit ;
infile state educ sex wsalval age using "D:\my_data.asc", clear;
replace wsalval=. if (wsalval==0);
gen male=(sex==1); *generates a dummy var for male (for women, male=0);
gen white=(race==1); *generates dummy var for white (also includes hispanics);
…..
```

SAS

```
data one;
infile "D:\my_data.asc";
input state educ sex wsalval age;
/*in SAS, data manipulation in the data step, then estimation in the proc step     */
/* can't intermingle like you can in STATA, both data steps/procs end with a 'run;' */
if sex=1 then male =1; else if sex=2 then male=0;
if race=1 then white=1; else if race ne 1 then white=0;
if race=. then white=.;
… …
```