Module 2 – Regression and Prediction

CASE STUDY ACTIVITY - CODEBOOK

CASE STUDY 4 – Predicting Wages 2

PROFESSIONAL EDUCATION
Digital Programs

CASE STUDY ACTIVITY - CODEBOOK

CASE STUDY 4 – Predicting Wages 2

Codebook for Wage Prediction Exercise

Data files and Programs

wage2015.Rdata: This data file contains 60 demographic and job-relevant variables from the CPS March Supplement 2015. For a list of variables included in this data set see below.

Data Cleaning: In order to obtain the relevant sample the data is cleaned by applying the following rules:

1. Drop armed forces and children from the sample. (Drop if popstat equals 2 or 3)
2. Drop individuals who have worked 0 hour last year. (Drop if wkswork1 equals 0)
3. Drop individuals with age less than 16. (Drop if age less than 16)
4. Drop individuals with allocated income and work variables. (Drop if qincwage or qwkswork equal to 1)
5. Keep individuals working for wages or salary. (Drop if classwly equals to (10,13,14,99,29,00))
6. Drop part-time workers. (Drop if uhrsworkly is less than 36)
7. Drop individuals who worked less than 50 weeks last year. (Drop if wkswork1 is less than 50)
8. Keep only never-married individuals (Keep if marst equals 6).
9. Drop individuals with zero annual wage. (Keep if incwage equals 0).
10. Trim %2 from both the top and bottom log wage distribution to drop outlier observations.

After data cleaning, new variables are constructed from the raw CPS variables. Table 1 provides the list of variables and calculation methods.

Note that occupation and industry variables are coded as factors instead of dummy variables. They should be turned into dummy variables before prediction.

Table 1: Constructed Variables in wage2015.Rdata

Characteristics	Variables	Calculation	Type	CPS Variable
Log Hourly Wage	*wage* *lwage*	$\frac{\text{annualincome}}{(\text{weeksworked} \times \text{hoursworked})}$	*Continuous* *Continuous*	*incwage, uhrsworklyt, wkswork1*
Gender	*sex*	*1 if female*	*Dummy*	*sex*
Race	*white* *black*	*1 if white* *1 if black*	*Dummy* *Dummy*	*race*
Hispanic	*hisp*	*1 if hispanic*	*Dummy*	*hispan*
Education	*shs* *hsg* *scl* *clg*	*some high school (years of educ < 12)* *high school graduate (years of educ = 12)* *some collage (12< years of educ < 16)* *collage graduate (15< years of educ < 18)*	*Dummy* *Dummy* *Dummy* *Dummy*	*educ*
Region	*mw* *so* *we*	*1 if living in midwest* *1 if living in south* *1 if living in west*	*Dummy* *Dummy* *Dummy*	*region*
Union Membership	*union*	*1 if covered by union*	*Dummy*	*union*
Veteran Status	*vet*	*1 if veteran*	*Dummy*	*vetstat*
City Status	*cent* *ncent*	*1 if living in central city* *1 if living outside central city*	*Dummy* *Dummy*	*metro*
Family Size	*fam1* *fam2* *fam3*	*1 if family size is 1* *1 if family size is 2* *1 if family size is 3*	*Dummy* *Dummy* *Dummy*	*famsize*
Children	*child*	*1 if individual has children*	*Dummy*	*nchild*
Foreign Born	*fborn*	*1 if born in a foreign country*	*Dummy*	*nativity*
Citizenship	*cit*	*1 if US citizen*	*Dummy*	*citizen*
School Attendance	*sch*	*1 if attending school*	*Dummy*	*schlcoll*
Pension	*pens*	*1 if included in pension plan available at work*	*Dummy*	*pension*
Firm Size	*fsize10* *fsize100*	*1 if number of employees less than 10* *1 if number of employees between 10 and 100*	*Dummy* *Dummy*	*fsize*
Health Status	*health*	*1 if health is very good or perfect*	*Dummy*	*health*

Age	*age*	*age of the worker*	*Discrete*	*age*
Experience	*exp1*	*(age - 1) - years of educ -7*	*Continuous*	*age, educ*
	exp2	*exp1^2=10*	*Continuous*	
	exp3	*exp1^3=100*	*Continuous*	
	exp4	*exp1^4=1000*	*Continuous*	
Occupation level[a]	*occ*	*Occupation Categories in CSP (456 categories)*	*Categorical*	*occly*
	occ2	*Aggregated Occupation Categories (22 categories)*	*Categorical*	
Industry level[b]	*ind*	*Industry Categories in CSP (257 categories)*	*Categorical*	*indly*
	indg2	*Aggregated Industry Categories (23 categories)*	*Categorical*	

[a] *Since the original CPS occupation variable(occ) includes too many categories a second variable named occ2 is constructed by aggregating occ according to aggregation in this webpage. For the aggregated occupation codes see below.*

[b] *Similar to occupation, industry categories are aggregated based on categorization in this webpage. For the aggregated industry codes see below.*

Occupation Codes for ooc2 variable

1 Management occupations
2 Business and financial operations occupations
3 Computer and mathematical occupations
4 Architecture and engineering occupations
5 Life, physical, and social science occupations
6 Community and social service occupations
7 Legal occupations
8 Education, training, and library occupations
9 Arts, design, entertainment, sports, and media occupations
10 Healthcare practitioners and technical occupations
11 Healthcare support occupations
12 Protective service occupations
13 Food preparation and serving related occupations
14 Building and grounds cleaning and maintenance occupations
15 Personal care and service occupations
16 Sales and related occupations
17 Office and administrative support occupations

18 Farming, fishing, and forestry occupations
19 Construction and extraction occupations
20 Installation, maintenance, and repair occupations
21 Production occupations
22 Transportation and material moving occupations

Industry Codes for ind2 variable

1 Agriculture, Forestry, Fishing, and Hunting
2 Mining
3 Utilities
4 Construction
5 Nondurable Goods manufacturing
6 Durable Goods Manufacturing
7 Durable Goods Wholesale
8 Nondurable Goods Wholesale
9 Retail Trade
10 Transportation and Warehousing
11 Information
12 Finance and Insurance
13 Real Estate and Rental and Leasing
14 Professional, Scientific, and Technical Services
15 Management of companies and enterprises
16 Administrative and support and waste management services
17 Educational Services
18 Health Care and Social Assistance
19 Arts, Entertainment, and Recreation
20 Accommodation and Food Services
21 Other Services (Except Public Administration)
22 Public Administration
23 Armed Forces

Downloaded CPS Variables

HWTSUPP	*Household weight*
GQ	*Group Quarters status*
HHINTYPE	*Type of household*
REGION	*Region and division*
STATEFIP	*State (FIPS code)*
ASECFLAG	*Flag for ASEC*
METRO	*Metropolitan central city status*
CBSASZ	*Core-based statistical area size*
HHINCOME	*Total household income*
MONTH	*Month*
PERNUM	*Person number in sample unit*
WTSUPP	*Supplement Weight*
EARNWT	*Earnings weight*
FAMSIZE	*Number of own family members in hh*
NCHILD	*Number of own children in household*
NCHLT5	*Number of own children under age 5 in hh*
RELATE	*Relationship to household head*
AGE	*Age*
SEX	*Sex*
RACE	*Race*
MARST	*Marital status*
POPSTAT	*Adult civilian, armed forces, or child*
CITIZEN	*Citizenship status*
NATIVITY	*Foreign birthplace or parentage*
HISPAN	*Hispanic origin*
EDUC	*Educational attainment recode*
EDUC99	*Educational attainment, 1990*
EMPSTAT	*Employment status*
SCHLCOLL	*School or college attendance*
OCC	*Occupation*
OCCLY	*Occupation last year*
INDLY	*Industry last year*
CLASSWLY	*Class of worker last year*
WKSWORK1	*Weeks worked last year*
WKSWORK2	*Weeks worked last year, intervalled*
UHRSWORK	*Usual hours worked per week (last yr)*
AHRSWORKT	*Hours worked last week*
WKSUNEM1	*Weeks unemployed last year*
HOURWAGE	*Hourly wage*
PENSION	*Pension plan at work*
UNION	*Union membership*
FIRMSIZE	*Number of employees*
FTOTVAL	*Total family income*
INCTOT	*Total personal income*
INCWAGE	*Wage and salary income*
EARNWEEK	*Weekly earnings*
VETSTAT	*Veteran status*
HEALTH	*Health status*

QAGE	*Data quality flag for AGE*
QMARST	*Data quality flag for MARST*
QSEX	*Data quality flag for SEX*
QEDUC	*Data quality flag for EDUC*
QCLASSWL	*Data quality flag for CLASSWLY*
QUHRSWORK	*Data quality flag for UHRSWORKT*
QWKSWORK	*Data quality flag for WKSWORK1 and WKSWORK2*
QEARNWEE	*Data quality flag for EARNWEEK*
QINCWAGE	*Data quality flag for INCWAGE*
MHMARNUM	*Number of times married*