

# Using an Autocoder to Code Industry and Occupation in the American Community Survey

# Why Autocode Industry and Occupation?

- In 2010, workload of 2.48 million manually coded records per year
- Clerical - very labor intensive
  - Cost
  - Time
- Concern about accuracy and consistency
  - Keying and other errors
  - Variation in coding clerk's interpretations of the respondent's write-in
- Future increase in workload (e.g., sample expansion, bridge coding, competing surveys)

# American Community Survey Variables Autocoded

## Type and Method of Coding

Race.....	} Backcoding, automated with clerical follow-up
Hispanic Origin .....	
Ancestry .....	
Language .....	
Health Insurance.....	
Field of Degree.....	
Computer Types.....	
Internet Service.....	
<b>Industry .....</b>	- <b>Industry, automated with clerical follow-up</b>
<b>Occupation.....</b>	- <b>Occupation, automated with clerical follow-up</b>
Place of Birth.....	} Geocoding Automated with clerical follow-up
Migration .....	
Place of Work.....	

# Implementation

## ACS Data Year 2014

### Autocoded

Industry	55%
Occupation	43%
Joint I&O	29%

# \$avings

<b>Year</b>	<b>Number of Records</b>	<b>Autocoded Records</b>	<b>Records Sent to Clerical Coding</b>	<b>Joint Coding Rate</b>
2012	3,016,617	887,750	2,128,867	29.4
2013	2,823,924	824,122	1,999,802	29.2
2014	2,808,920	816,452	1,992,468	29.1

# Autocoder

- All records with I&O write-in information go through autocoder
- Designed to replicate clerical coding
- Industry and occupation coded separately
- Assigned codes with quality scores below cutoff go to clerical coding
- Result:
  - 30% both codes assigned (no clerical coding)
  - 40% have one code assigned (partially coded )
  - 30% no codes assigned

**Goes to  
clerical  
coding**

# ACS I & O Data Process

**Questionnaire completion (collection)**

- Paper
- Internet
- CATI
- CAPI

**Data capture**

- Keyed from image (KFI), truncated to 60 characters
- Data capture file

**Coding**

- Coding file for I&O
- Autocoder
- Clerical coding

**Edits**

# ACS Questionnaire Industry Items

## 42 For whom did this person work?

*If now on active duty in the Armed Forces, mark (X) this box →  and print the branch of the Armed Forces.*

Name of company, business, or other employer

## 43 What kind of business or industry was this?

*Describe the activity at the location where employed. (For example: hospital, newspaper publishing, mail order house, auto engine manufacturing, bank)*

- Industry data describe the kind of business conducted by a person's employing organization
- 3 industry questions
  - 2 write-ins
  - 1 checkbox

# ACS Questionnaire Occupation Items

**45** What kind of work was this person doing?  
*(For example: registered nurse, personnel manager, supervisor of order department, secretary, accountant)*

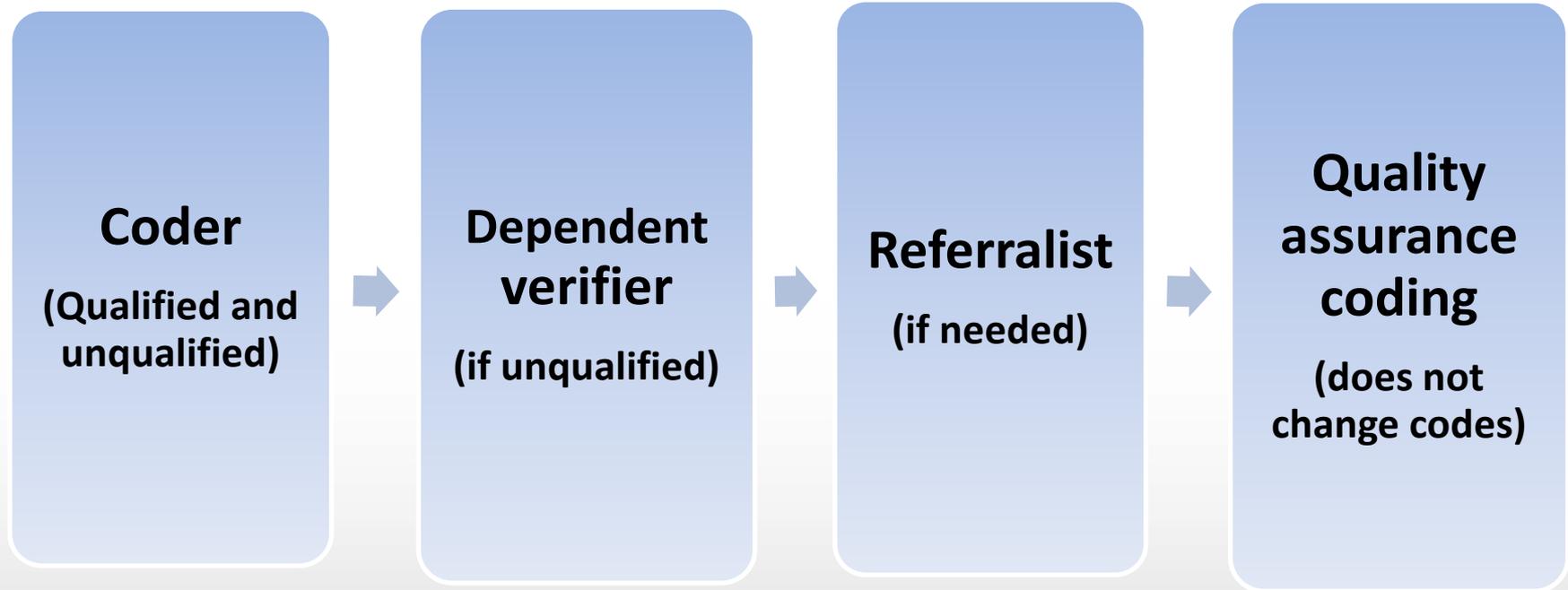
**46** What were this person's most important activities or duties?  
*(For example: patient care, directing hiring policies, supervising order clerks, typing and filing, reconciling financial records)*

- Occupation describes the kind of work a person does on the job
- 2 occupation questions

# Example of Write-ins

Industry Write-in (INW3)	Occupation Write-in (OCW1)
BANK	ASSISTANT VP LOAN DEPARTMENT
CONSTRUCTION COMPANY	PROJECT MANAGER
CONSTRUCTION CONTRACTER	BRIDGE CARPENTER
CROP CATTLE FARM	ALL JOBS THAT NEEDED DOING
UNIVERSITY HIGHER EDUCATION	ASSISTANT DEAN/ADMINISTRATION
CHILD CARE	BOSS
PROPERTY MANAGEMENT	PROPERTY MANAGER
MEAT PROCESSING	MEAT PROCESSING
BOWLING ALLEY- SNACK BAR	NIGHT SUPERVISOR FOR SNACK BAR
STEEL BRIDGE FABRICATOR	MACHINE PROGRAMMER-LAYOUT
HIGH SCHOOL	COUNSELOR
NON-PROFIT LAW FIRM	ATTORNEY
UNIVERSIDAD	AYUDANTE DE PROGRAMA P E A N
EDUCATION	INSTRUCTOR OF ENGLISH
RESTAURANT	COOK
PIZZA	HE MAKES PIZZAS
BURYING PEOPLE	I BURY THE DEAD
MILITARY	RADIO MAINTENANCE
GOVERNMENT	D

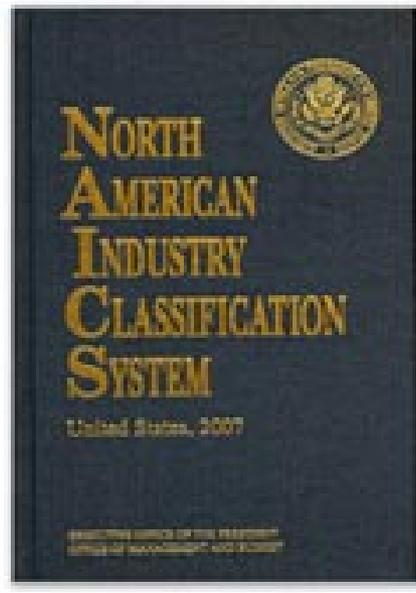
# Clerical Coding Process



# Variables Coders Use to Code I&O

- Age
- Sex
- Date of birth
- Educational attainment
- Residence county, state
- Active duty Armed Forces checkbox
- Class of worker checkbox
- Employer name write-in
- Kind of business write-in
- Industry type checkbox
- Kind of work write-in
- Job duties write-in

# Industry and Occupation Codes

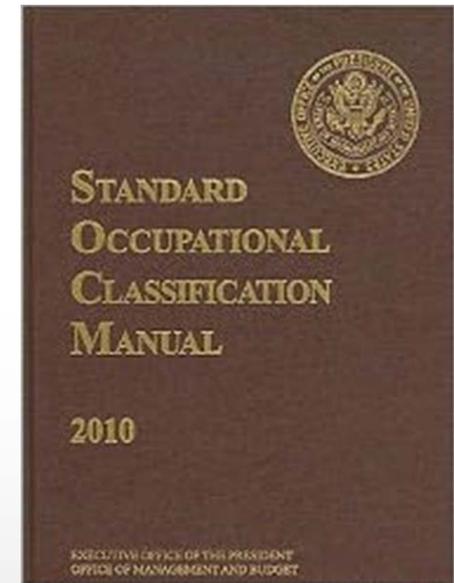


## Census INDUSTRY Codes

- Covers all 20 sectors
- Classified based on NAICS two-digit through six-digit codes
- 269 Census industry codes (4 digits)

## Census OCCUPATION Codes

- Covers all 23 major occupation groups
- Based on SOC two-digit through six-digit codes
- 539 Census occupation codes (4 digits)



# Coding Indexes

OCCUPATION TITLE	NAICS RESTRICTION	SOC CODE	INDUSTRY RESTRICTION	OCCUPATION CODE
------------------	-------------------	----------	----------------------	-----------------

Teacher, elementary school	6111	<a href="#">25-2021</a>	7860	<a href="#">2310</a>
Teacher, french	6112, 6113	<a href="#">25-1124</a>	7870	<a href="#">2200</a>

Teacher \ n.s.	51331	<a href="#">43-2011</a>	6680	<a href="#">5010</a>
Teacher \ n.s.	6112, 6113	<a href="#">25-1199</a>	7870	<a href="#">2200</a>
Teacher \ n.s.	524, 8121, 6114, 6115	<a href="#">25-1194</a>	6990, 7880, 8870, 8880, 8970	<a href="#">2200</a>
Teacher \ n.s.	Bible school 611699	<a href="#">21-2099</a>	Bible school 7890	<a href="#">2060</a>
Teacher \ n.s.	Elementary school 6111	<a href="#">25-2021</a>	Elementary school 7860	<a href="#">2310</a>

# I&O Coding System

2010 Industry Reference No Exact Match! Cou/ST: HUDSON, NJ

Description	Code	Occupation
▶ Clothing patterns and plans publishing	6480	
Clothing store (ret)	5170	
Clothing used (ret)	5490	
Cloud seeding	7490	
Clubhouse membership	9170	
Clubhouse recreational	8590	
Clubhouse residential	8670	
Clutches auto (mfg)	3570	
Clutches exc auto (mfg)	3180	

Search For:  Prime Words:  Special Codes:

Batch: 003

PSU: 34017 User ID: 042933971003 17 F GRADE 10

Employer:

Industry:

Industry Type:

Code Def. :

Occupation:

Duties:

Code Def. :

C.O.W. :

Ind Code:

Occ Code:

Orig:  Class of Worker Code:

4 of 10

Industry,  
occupation,  
or military  
Index

Industry code

Occupation  
code

# Autocoder Process

Industry write-ins (INW2, INW3), Occupation write-ins (OCW1, OCW2)

Text parsed from uncoded record

Word-string matched to dictionaries, One-word, two-word and full string

Multiple matches or no matches possible

Logistic regression model to select best possible code

# 12 Autocoder Dictionaries

- 1 million coded records from 2010 ACS used to build dictionaries – random selection
- 3 dictionaries for each write-in variable – 6 for industry and 6 for occupation
- Cross-coding in dictionaries to assign industry from occupation entries and vice versa

**Autocoder Dictionaries**

		1 word bit	2 word bits	full write in
Industry	INW1	1	2	3
	INW2	4	5	6
Occupation	OCW1	7	8	9
	OCW2	10	11	12



# Autocoder Dictionaries Criteria

Criteria for inclusion:

- 30 occurrences for 1-word, 2-word
- 15 occurrences for full-string
- 50% of occurrences map to one code
- 75% of occurrences map to one code for cross-codes

Dictionaries include word-bit, code, total frequency and frequency percentage of the code

# Examples from Dictionaries for Occupation Write-in

Occupation (OCW1)	Word Occupation Count	Occupation Code	Occupation Count	Frequency Percent	Industry Code	Industry Count	Industry Percent
<b><u>One-word:</u></b>							
PICKING	43	6050	25	58.1%		0	0.0%
<b><u>Two-word:</u></b>							
PICKING FRUIT	22	6050	20	90.9%		0	0.0%
PICKING GRAPES	33	6050	33	100.0%	0170	26	78.8%
<b><u>Full-String:</u></b>							
PICKING GRAPES	31	6050	31	100.0%	0170	24	77.4%
PICKING UP TRASH	17	9720	11	64.7%	7790	14	82.4%

# Selecting the Best Code: Logistic Regression Model

- 2 models: industry and occupation
- SAS Proc Logistic with Stepwise
- Independent variables:
  - Variables used in clerical coding
  - Plus Coding dictionary (frequency percentage, total frequency)
  - 111 variables in industry model; 91 variables in occupation model
- Dependent variable
  - 1 if dictionary match agrees with assigned code
  - 0 if code does not agree
- Model estimates the probability the code agrees with what a clerk would assign



# Hardcodes

- Industry or occupation code reassignment based on additional text information
- Example: elementary and secondary education industry reassigned to colleges and universities based on 'University'
- Corrects most common errors

# Ensuring Quality

## Disagreement Rates Between Autocoder and Clerical Coding\*

Coding Rate	Industry	Occupation
30%	1.01%	4.49%
40%	2.12%	5.30%
50%	3.96%	8.35%
60%	5.39%	13.43%

\*From expert coding test with 2,000 records

# Cutoff Level

- Cutoffs based on model score
  - Keep code if score > cutoff
  - Separate cutoffs for industry and occupation
  - Set cutoffs to generate similar quality as 100% clerical coding
- Records with scores below cutoff go to clerical coding

	Industry	Occupation
Error rate	4.5%	5.9%
Acceptable Codes	56%	43%



# Production QC Process

- 3 times annually
- Sample by code categories
- Clerical coding done by referralists
- Track error rates
- Evaluate problem wordbits
- Test change in autocoder
- Outcomes- change codes, frequency percentages, dictionary entries

# QC Update- Occupation Example

- Autocoder assigned **5420** (43-4199)  
“Information and Record Clerks, All Other”
- Expert coders assigned **3510** (29-2071)  
“Medical Records and Health Information Technicians”
- Wordbits occurring in 86% of cases:  
“coder”, “coding”, “medical coder”, “medical coding”
- “Medical coding”  
Two-word job duties dictionary- changed to 87 percent frequency  
Full-string occupation dictionary- changed code and frequency
- “Medical Coder”  
Two-word job duties dictionary- added with 100 percent frequency  
Two-word occupation dictionary- changed code and frequency
- Clerk coding index – added entries

# Updating Codes

- Constructed using 2010 I&O coding
- Updates based on I&O Indexes, QC, and coding changes
  - 2010 data year– occupation updates consistent with 2010 SOC changes
  - 2013 data year– industry updates consistent with 2012 NAICS update

# Industries with Highest Frequency of Autocoding

Industry	% autocoded	Industry	% autocoded
Postal service	95	Legal services	83
Elementary and secondary schools	94	Veterinary services	82
Beauty salons	92	Funeral homes	81
Offices of dentists	90	Rail transportation	80
Restaurants	90	Landscaping services	77
Religious organizations	86	Libraries	77
Hospitals	86	Barber shops	76
Banking	84	Grocery stores	76
Insurance carriers	84	Accounting services	75
Coast guard	84	Bowling centers	75
Justice, public order, and safety	84	Nursing care	74

Half of autocoded records to these industries

# Occupations with Highest Frequency of Autocoding

Occupation	% autocoded	Occupation	% autocoded
Hairdressers	95	Clergy	85
Speech-lang pathologists	91	Dentists	84
Massage therapists	88	Dishwashers	83
Postal service mail carriers	88	Bartenders	82
Dental hygienists	88	Hosts	81
Flight attendants	87	Truck drivers	81
Elementary and middle school teachers	86	Pilots	80
Firefighters	86	Cashiers	79
Lawyers	86	Dental assistants	78
Waiters	85	Nurse practitioners	78
Respiratory therapists	85	Misc personal appearance	78

One-quarter of autocoded records to these occupations

# Autocoding Rates for Largest Industries

Industry	% autocoded
Elementary and secondary schools	94
Construction	72
Restaurants	90
Hospitals	86
Colleges	67
Grocery stores	76
Justice, public order, and safety	84
Department stores and discount stores	42
Insurance carriers	84

These industries total one-third of I&O records across all industries.

# Autocoding Rates for Largest Occupations

Occupation	% autocoded
Secretaries	66
Cashiers	79
Elementary and middle school teachers	86
Retail salespersons	24
Truck drivers	81
Managers, all other	3
Janitors and building cleaners	38
Registered nurses	72
First-line supervisors of retail sales workers	13
Nursing aides	57
Customer service	52
Cooks	77
Laborers and freight, stock, and material movers, hand	24
Waiters	85
Accountants	52
Construction laborers	8
Stock clerks	54

These occupations total one-third of records for all occupations

# Codes Not Autocoded

Industry ~26 codes

- 10 percent of codes
- 1.3 percent of records

Occupation ~ 100 codes

- 20 percent of codes
- 1.6 percent of records

# Partially Coded Records

Clerical coders can change the autocoded code

Industry:	1.8 percent of industry partial codes
Occupation:	1.5 percent of occupation partial codes

# Evaluating the 2012 Autocoder

*(in progress)*

- **Multiple meaning for key words**
  - Manager, dealer, editor
- **Too much weight given to particular words**
  - E.g., “X-ray installer”
- **Phrasing of activities**
  - “engineering manager” vs. “engineer, managing a team”
- **Spelling - not all variations included**
- **Occupation types difficult to autocode** - “all other”, managers, engineers, teachers, laborers, designers, entertainers, editors, cafeteria workers

# Unexpected Consequences

Class of worker:

- Can be changed by ACS clerical coders
- Not changed by autocoder
- Not changed in CPS clerical coding

Distributions vary, particularly when crossed by occupation

41

**Was this person -**  
*Mark (X) ONE box.*

- an employee of a PRIVATE FOR-PROFIT company or business, or of an individual, for wages, salary, or commissions?
- an employee of a PRIVATE NOT-FOR-PROFIT, tax-exempt, or charitable organization?
- a local GOVERNMENT employee (city, county, etc.)?
- a state GOVERNMENT employee?
- a Federal GOVERNMENT employee?
- SELF-EMPLOYED in own NOT INCORPORATED business, professional practice, or farm?
- SELF-EMPLOYED in own INCORPORATED business, professional practice, or farm?
- working WITHOUT PAY in family business or farm?



# More Information

---

Census Industry and Occupation website:

<http://www.census.gov/people/io/>

Industry and Occupation Statistics Branch

Telephone: 301-763-3239

[Jennifer.Cheeseman.Day@Census.gov](mailto:Jennifer.Cheeseman.Day@Census.gov)

[Melissa.C.Chiu@Census.gov](mailto:Melissa.C.Chiu@Census.gov)

[Ana.J.Montalvo@Census.gov](mailto:Ana.J.Montalvo@Census.gov)

# The American Community Survey Industry & Occupation Autocoder

## Additional Technical Points

Matthew Thompson, U.S. Census Bureau

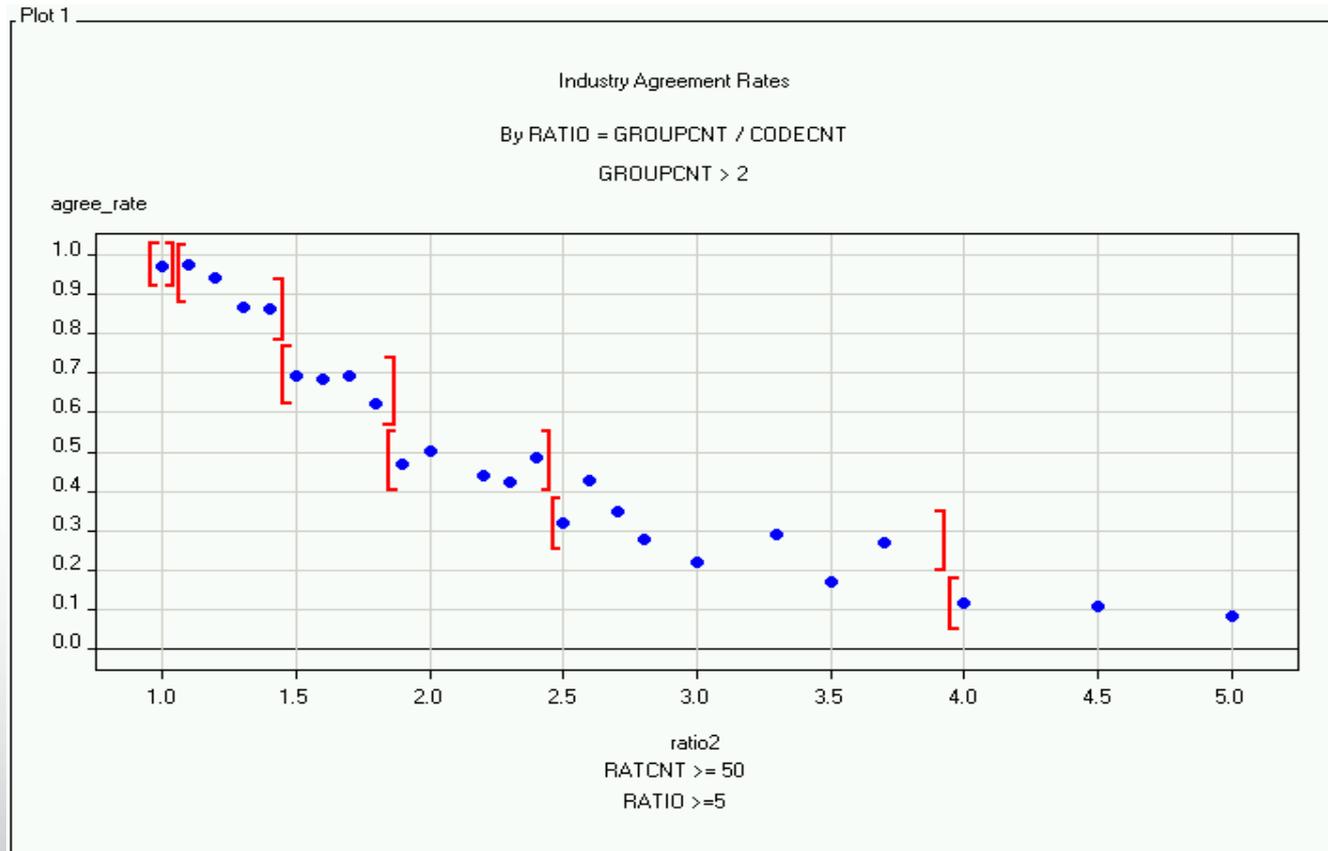
# Model Variables

- Variables Used by Clerical Coders
  - Age, Sex, Educational Attainment, etc...
  - Included to mimic the clerical coding process
- Variables from the Data Dictionaries
  - Word Frequency, Code Frequency, dictionary matched
- Calculated
  - # of words in write-in fields, etc...

# Model Variables

- Continuous variables
  - Only frequency of code
  - Shown to be most informative piece of information from Form SS-4 auto-coding
- All other data in model represented by binary variables
  - e.g. If SEX = M then  $x_1 = 1$ , else  $x_1 = 0$   
If SEX = F then  $x_2 = 1$ , else  $x_2 = 0$

# Grouping into Binary Variables



# Hard-Codes

- Ran Chi-Squared goodness-of-fit test
  - Model fit was acceptable
- Used test statistic to identify problem industries and occupations

IND	Clerical Count	Auto Count	Test Statistic Contribution
7890	3,688	3,133	83.52
7860	50,006	51,893	71.21
9480	4,301	3,883	40.62
4380	1,299	1,071	40.02
7880	530	395	34.39
6690	1,633	1,433	24.49
9950	1,324	1,144	24.47
7380	7,690	8,120	24.04
2190	2,444	2,668	20.53
8680	38,969	39,829	18.98

OCC	Clerical Count	Auto Count	Test Statistic Contribution
2310	6,732	7,824	177.13
2320	1,561	1,187	89.61
2300	1,044	800	57.03
5840	463	311	49.90
2200	2,420	2,090	45.00
9130	5,396	5,866	40.94
5510	483	344	40.00
7720	296	209	25.57
5240	3,680	3,986	25.45
4320	201	133	23.01

# Creating Hard-codes

1. Identify problem Industry/Occupation
2. Create list of most frequent disagreements between clerk and autocoder
3. Perform case-by-case review
4. Establish rule for correcting common errors, i.e. hard-codes

# Questions?