History 595 Final Examination

Part I (20 points). Below is a partial variable list from the General Social Survey for 1993. This survey of 1500 Americans is collected annually to provide information on a wide variety of attitudes and behaviors of American adults. Identify the variables a nominal, ordinal or interval.

CHILDS Number of Children
AGE Age of Respondent
ZODIAC Respondents Astrological Sign

Value	Label
0	Missing
1	Aries
2	Taurus
3	Gemini
4	Cancer
5	Leo
6	Virgo
7	Libra
8	Scorpio
9	Sagittarius
10	Capricorn
11	Aquarius
12	Pisces
98	Don't Know

EDUC Highest Year of School Completed
DEGREE Respondent's Highest Degree

Value	Label
0	Less than HS
1	High school
2	Junior college
3	Bachelor
4	Graduate

INCOME91 Total Family Income

Value	Label
1	LT $1000
2	$1000-2999
3	$3000-3999
4	$4000-4999
5	$5000-5999
6	$6000-6999
7	$7000-7999
8	$8000-9999
9	$10000-12499
10	$12500-14999
11	$15000-17499
12	$17500-19999
13	$20000-22499
14	$22500-24999
15	$25000-29999
16	$30000-34999
17	$35000-39999

18 $40000-49999
19 $50000-59999
20 $60000-74999
21 $75000+

REGION Region of Interview

Value	Label
0	Not Assigned
1	New England
2	Middle Atlantic
3	E. Nor Central
4	W. Nor Central
5	South Atlantic
6	E. Sou Central
7	W. Sou Central
8	Mountain
9	Pacific

XNORCSIZ Expanded residential Size Code

Value	Label
1	City, GT 250000
2	City, 50-250000
3	Suburb, Lrg City
4	Suburb, Med City
5	UnInc, Lrg City
6	UnInc, Med City
7	City, 10-49999
8	Town, GT 2500
9	Smaller Areas
10	Open Country

PARTYID Political Party Affliation

Value	Label
0	Strong Democrat
1	Not Str Democrat
2	Ind, Near Dem
3	Independent
4	Ind, Near Rep
5	Not Str Republican
6	Strong Republican
7	Other Party

CAPPUN Favor or Oppose Death Penalty for Murder

Value	Label
1	Favor
2	Oppose

GRASS Should Marijuana Be Made Legal

Value	Label
1	Legal
2	Not Legal
8	Don’t know

RELIG Religious Preference

Value	Label
1	Protestant
2	Catholic
3	Jewish

4 None
5 Other
8 Don't know

LIFE Is Life Exciting or Dull

Value Label
1 Dull
2 Routine
3 Exciting

SPANKING Favor Spanking to Discipline Child

Value Label
1 Strongly Agree
2 Agree
3 Disagree
4 Strongly Disagree

NEWS How Often Does R Read Newspaper

Value Label
1 Everyday
2 Few Times a Week
3 Once a Week
4 Less Than Once Wk
5 Never

TVHOURS Hours Per Day Watching TV

ATTSPRTS Attended Sports Event in Last Yr

Value Label
1 Yes
2 No

TVSHOWS How Often R Watches TV Drama or Sitcoms

Value Label
1 Daily
2 Several Days in Week
3 Several Days in Month
4 Rarely
5 Never

PARTNERS How Many Sex Partners Respondent Had in Last Year

Value Label
0 No Partners
1 1 Partner
2 2 Partners
3 3 Partners
4 4 Partners
5 5-10 Partners
6 11-20 Partners
7 21-100 Partners
8 More Than 100 Partners

DWELOWN Homeowner or Renter

1 owns home
2 pays rent
3 other

Part II (10 points) True/False

1. A researcher reports a cross tabulation by sex of a sample of responses on attitudes towards defense spending. The researcher reports a p value for a Chi Square statistic from the table as .01. Such a result means that any differences by sex are likely to be the result of chance.
2. A researcher is trying to test whether a b coefficient in a regression model is statistically significant. She should use a T test and the probability reported for the coefficient.
3. A researcher wants to evaluate the level dispersion in a distribution of reported incomes. A good measure of dispersion is a standard deviation.
4. A researcher wants to evaluate the variability around the point estimate derived from a sample mean. He should calculate a standard error.
5. The adjusted R square from a linear multiple regression model is a measure of the impact of the most important independent variable.
6. Cross tabulations provide a method of examining the relationship between two variables measured at the nominal or small ordinal level.
7. The expected frequency of any cell in a cross tabulation is calculated by multiplying the row marginal total by the column marginal total and dividing by the total N.
8. The expected frequencies are used to calculate the Chi Square statistic.
9. The dependent variable in a regression model should be measured at the interval level.
10. Dummy variables are inappropriate for use in regression models.

Part III (10 points).

A researcher is interested in analyzing the patterns in the General Social Survey. (See Part I above.) She has learned a series of statistical tests and several data analysis techniques in History 595. She now wants to apply her new knowledge to a series of problems. For each situation below, pick the statistical technique or techniques, and the appropriate statistical test, to be used to analyze the problem described. Explain why you would choose the technique and test. Identify which variables are independent variables and which variables are dependent variables for each situation. (There may be more than one correct answer depending on how you set up your analysis.)

1. A researcher wants to know if appreciation of rap music (like it, have mixed feelings, or dislike it) differs by the political outlook of the respondent (liberal, moderate, conservative).
2. A researcher wants to understand if college educated respondents are more liberal than those with less than a college degree, given the household's income.
3. A researcher wants to understand if younger people find life more exciting than older people.
4. A researcher wants to find out if people who report reading newspapers more also spend more time watching TV.
5. A researcher wants to find out what the determinants are of the number of hours per day spent watching TV and if the time spent differs by the sex, age, income, and political attitudes of the respondents.

Statistical Technique:
Univariate Analysis of Sample Data
Cross Tabulation for Two Way or Three Tables
Difference of Two Sample Means
Analysis of Variance of Multiple Sample Means
Linear Regression Model
Logistic Regression Model

Statistical Test: Z Test; T Test; Chi Square Test; F Test

Part IV: (35 points). Using Regression to Understand Household Size

Today and in the past, households are of varying size. At the smallest, a household may contain just one person, for someone living alone. Young couples or "empty nesters" have households of two. At the other end of the spectrum, households can be quite large: extended families, families with servants or boarders, or several families living in one household. We can use regression analysis to study the determinants of household size.

On the following pages are regression models of household size in Milwaukee at the turn of the 20th century, derived from the data collected from the Wisconsin state census by Roger Simon (for 1905), and the 1910 federal census. The Simon data has information from the four wards he studied in his book. The census data provide household information for the entire city. A historian has used the information from both data sources to explore the determinants of household size. There are four different models, two from the Simon data and two from the 1910 census data. Some of the variables are the same in all four models. Some of the variables differ in the four models, and the table will have a blank cell if the variable either was not available, or was not included in the model. There are three tables below. Table IV.1 is the description of the variables. Table IV.2. the determinants of household size, contains the four regression models. Table IV.3 is the descriptives tables with results for the variables in the models.

Answer the questions below using the information in the tables. (2 points each)

1. What was the average household size in 1905 in the four peripheral wards in Milwaukee that Roger Simon studied?
2. What was the average household size in 1910 in Milwaukee according to the federal census?
3. What proportion of households rented in the four peripheral wards in 1905?
4. What proportion of households rented in the city in 1910?
5. What proportion of households in the four peripheral wards in 1905 were of Polish ethnic background?
6. What proportion of households in the city in 1910 were of Polish ethnic background?
7. What proportion of households in the city in 1910 were headed by women?

Because these models were developed using two different data sources and surveyed

different populations, the four models provide analyses that include some common characteristics and patterns and some differing ones. The two data sources have some common variables and some variables that are unique to one source or the other. Regression analysis has the advantage of providing a method for evaluating the impact of any particular variable on a particular model. Keeping in mind the nature of the underlying data, answer the following questions: (3 points each)

8. Identify all the determinants that have a statistically significant affect on household size in any of the models.
9. Identify all the determinants that have a statistically significant affect household size in all of the models.
10. Write a short paragraph for a student who has not taken History 595 explaining whether the ethnic background of the household head had an impact on the size of the household.
11. In early twentieth century Milwaukee, was there a difference in the size of households headed by women compared to those headed by men? Why or why not?
12. Using model 1, estimate the household size for a household headed by an American born skilled male breadwinner in his forties who owned his home and lived with his wife and children. Show the calculations.
13. Using model 2, estimate the household size for a household headed by a 40 year old American born lawyer (professional worker) who owned a house on Milwaukee's East Side. The house was built in 1900, and was valued at $100,000. His wife's younger sister and her husband lived in a carriage house over the garage. Show the calculations.
14. Using model 3, estimate the size of a household headed by a 55 year old German born widow who didn't work and who rented a flat with her two teenage children.

Table IV.1: Variable Descriptions
1. Number of persons in the household
2. Number of families in the household
3. Age Cohort Squared
    Age Cohort:
        29 and under: -2
        30-39 -1
        40-49 0
        50-59 1
        60 and up 2
    Age Cohort Squared range: 0-4
4. Rents. Household rents. (0=No; 1=Yes)
5. Occupational Status of household head

Professional and clerical	1
Proprietor	2
Skilled worker	3
Semiskilled worker	4
Unskilled worker	5

Not in labor force, unemployed, retired 6

6. Polish: Whether household head is of Polish ethnicity (as designated by the 1905 Wisconsin Census or the respondent's father's mother tongue (1910 census) (0=No; 1=Yes)
7. German Whether household head is of German ethnicity (as designated by the 1905 Wisconsin Census or the respondent's father's mother tongue (1910 census) (0=No; 1=Yes)
8. Value: Value of the home in thousands of 2000 dollars (for 1905 data)
9. Year built: Year the house was built: 1887 or earlier=0; 1888=1; 1905=18 (for 1905 data).
10. Female: Whether household head was female (0=No; 1=Yes) (for 1910 data)
11. Peripheral Ward: Whether the household lived in wards 14, 18, 20 or 22.

Table IV.2. Determinants of Household Size in Milwaukee, OLS Regression Coefficients

Variable	Peripheral Wards, 14, 18, 20 and 22, 1905		City, 1910	
	1	2	3	4
Constant	1.659***	2.075***	3.723***	3.687***
Number of Families	3.288***	3.332***	1.000***	1.003***
Age Cohort Squared	-.312***	-.290***	-.364***	.-.364***
Rents	-.197	-.098	-.619**	-.609**
Occupational Status	.208***	.143*	.241**	.244**
Polish	1.770***	1.725***	.228	.210
Value		-.005		
Year built		-.027*		
Female			-1.629***	-1.619***
German			-.187	-.187
Peripheral Ward				.080
N	1039	868	319	319
R Squared	.439	.448	.275	.272

* $p < .05$
** $p < .01$
*** $p < .001$

Source: Simon Data (1905) and1910 IPUMS Data, from the federal population census

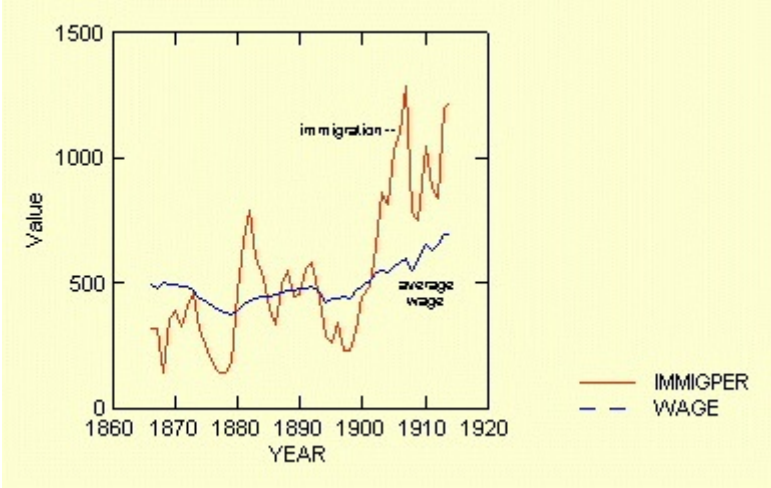
Table IV.3. Descriptive Statistics for Variables in Regression Models

Variable	Peripheral Wards, 14, 18, 20 and 22, 1905		City, 1910	
	Mean	SD	Mean	SD
Number of Persons in the Household	6.418	3.206	4.605	2.204
Number of Families	1.268	.496	1.313	.913
Age Cohort Squared	1.540	2.650	1.586	1.640
Rents	.303	.460	.652	.477
Occupational Status	3.222	1.371	3.348	1.509
Polish	.232	.422	.116	.321
Value	26.914	26.910		
Year built	5.643	6.627		
Female			.113	.317
German	.586	.493	.530	.500
Peripheral Ward			.207	.406

Part V. (15 points). Analyzing Immigration and Wage Levels in the US, 1866-1914

Below is a line graph depicting the pattern of immigration to and average wages in the United States from 1866 to 1914. One line depicts the number of immigrants arriving per year. The scale for immigrants arriving is in thousands of immigrants arriving per year. The other line depicts the average wage paid each year. The data come from Historical Statistics of the United States.

Below are two regression models of the relationship between the number of immigrants arriving and the wage information.



Here are the variables:
Immigper: number of immigrant arriving each year in thousands
Wage: average wage paid per year (adjusted for inflation)
Wagechan: percent change in wages from the previous year
Imperlst: number of immigrants arriving in the previous year in thousands

Model 1:

```
Dep Var: IMMIGPER   N: 49   Multiple R: 0.822   Squared multiple R: 0.676

Adjusted squared multiple R: 0.669   Standard error of estimate: 175.994

Effect         Coefficient    Std Error     Std Coef Tolerance     t   P(2 Tail)

CONSTANT         -1023.544      159.298        0.000      .      -6.425    0.000
WAGE                 3.173        0.320        0.822     1.000    9.904    0.000

Analysis of Variance
Source             Sum-of-Squares   df  Mean-Square     F-ratio       P

Regression           3038493.125     1  3038493.125      98.099       0.000
Residual             1455764.885    47    30973.721
```

Model 2:

```
Dep Var: IMMIGPER   N: 48   Multiple R: 0.933   Squared multiple R: 0.870

Adjusted squared multiple R: 0.861   Standard error of estimate: 114.558

Effect         Coefficient    Std Error     Std Coef Tolerance     t   P(2 Tail)

CONSTANT          -249.840      142.193        0.000      .      -1.757    0.086
IMPERLST             0.719        0.093        0.683     0.377    7.722    0.000
WAGE                 0.813        0.362        0.212     0.332    2.244    0.030
WAGECHAN            19.477        4.711        0.252     0.796    4.134    0.000


Analysis of Variance
Source             Sum-of-Squares   df  Mean-Square     F-ratio       P

Regression           3869246.489     3  1289748.830      98.277       0.000
Residual              577436.121    44    13123.548
```

Answer the following questions:

1. Is there a relationship between the average wage in the United States during these years and the number of immigrants arriving each year? If so, what is it?
2. Explain why the second model is an improvement on the first.
3. For both models, write the equation which predicts the average number of immigrants arriving in 1890. The raw data for the variables are below.
4. Attached is a line graph plotting the dependent variable, estimated y hat from the second regression model, and the residuals [errors] of the model by year. Explain to someone who hasn't taken History 595 what the graph shows.
5. Explain to someone who hasn't taken History 595 (in a short paragraph) what the regression models show about the relationship between wage levels in the U.S. and immigration.

Case number	YEAR	WAGE	WAGECHAN	IMMIGPER	IMPERLST
1	1866.000	489.000	-4.490	318.568	.
2	1867.000	479.000	-2.045	315.722	318.568
3	1868.000	499.000	4.175	138.840	315.722
4	1869.000	496.000	-0.601	352.768	138.840
5	1870.000	489.000	-1.411	387.203	352.768
6	1871.000	482.000	-1.431	321.350	387.203
7	1872.000	486.000	0.830	404.806	321.350
8	1873.000	466.000	-4.115	459.803	404.806
9	1874.000	439.000	-5.794	313.339	459.803
10	1875.000	423.000	-3.645	227.498	313.339
11	1876.000	403.000	-4.728	169.986	227.498
12	1877.000	389.000	-3.474	141.857	169.986
13	1878.000	379.000	-2.571	138.469	141.857
14	1879.000	373.000	-1.583	177.826	138.469
15	1880.000	386.000	3.485	457.257	177.826
16	1881.000	409.000	5.959	669.431	457.257
17	1882.000	428.000	4.645	788.992	669.431
18	1883.000	438.000	2.336	603.322	788.992
19	1884.000	441.000	0.685	518.592	603.322
20	1885.000	446.000	1.134	395.346	518.592
21	1886.000	453.000	1.570	334.203	395.346
22	1887.000	462.000	1.987	490.109	334.203
23	1888.000	466.000	0.866	546.889	490.109
24	1889.000	471.000	1.073	444.427	546.889
25	1890.000	475.000	0.849	455.302	444.427
26	1891.000	480.000	1.053	560.319	455.302
27	1892.000	482.000	0.417	579.663	560.319
28	1893.000	458.000	-4.979	439.730	579.663
29	1894.000	420.000	-8.297	285.631	439.730
30	1895.000	438.000	4.286	258.536	285.631
31	1896.000	439.000	0.228	343.267	258.536
32	1897.000	442.000	0.683	230.832	343.267
33	1898.000	440.000	-0.452	229.299	230.832
34	1899.000	470.000	6.818	311.715	229.299
35	1900.000	487.000	3.617	448.572	311.715
36	1901.000	511.000	4.928	487.918	448.572
37	1902.000	537.000	5.088	648.743	487.918
38	1903.000	548.000	2.048	857.046	648.743
39	1904.000	538.000	-1.825	812.870	857.046
40	1905.000	561.000	4.275	1026.499	812.870
41	1906.000	577.000	2.852	1100.735	1026.499
42	1907.000	598.000	3.640	1285.349	1100.735
43	1908.000	548.000	-8.361	782.870	1285.349

44	1909.000	599.000	9.307	751.786	782.870
45	1910.000	651.000	8.681	1041.570	751.786
46	1911.000	632.000	-2.919	878.587	1041.570
47	1912.000	651.000	3.006	838.172	878.587
48	1913.000	689.000	5.837	1197.892	838.172
49	1914.000	696.000	1.016	1218.480	1197.892

Part VI (10 points). In the last chapter of Simon's study, he summarizes his arguments and adds information about the transformation of the neighborhoods he analyzed in the second half of the twentieth century. He concluded (p. 144) by arguing that "The new neighborhoods on Milwaukee's periphery provided more space for raising children than the older, more densely built-up areas. Further, the opportunity for homeownership was very real and obviously a deeply felt goal for at least part of the population, regardless of whether it was a wise financial investment."

The dataset we have from his study does not provide evidence for these conclusions, since it does not contain information comparing homeownership and the age structure in the other wards in the city, including those that were also at the "periphery," abutted the city limits at the time. The 1910 population census 1.4% sample data file we have, however, does allow us to test Simon's reasoning here, because it contains information on all the wards in the city, and on the ages of everyone living in the city, and the ownership status of the household.

I have combined the 23 wards in the city in 1910 (see p, iv of *The City Building Process*), into 3 categories:

1. The old wards in the city: wards 1-10, 12, and 13.
2. The wards that Simon studied: 14, 18, 20, and 22.
3. The other "peripheral" wards: wards 11, 15-17, 19, 21, and 23.

I have recoded the ages of the residents into 2 categories:

Adults, ages over 18
Children, ages 0 to 18.

1. Attached are cross tabulations of the recoded ages and the recoded wards. Report whether these results support Simon's argument above, including the statistical significance of the results. Report the statistic which supports whether the results are statistically significant.

2. Attached are cross tabulations reporting the proportions of households that own or rent their dwelling by recoded ward. Report whether these results support Simon's argument above, including the statistical significance of the results. Report the statistic which supports whether the results are statistically significant.

3. Attached as well are cross tabulations reporting the numbers and proportions of the homeowners who owned their residences free and clear or whether they had a mortgage of some

sort. Report whether there are differences in the proportions of households with mortgages in the three types of ward, including the statistical significance of the results. Report the statistic which supports whether the results are statistically significant.

4. Since you know that the sample rate for the 1910 census file is 1.4%, calculate an estimate of the total number of homeowners in the city in 1910, and the number in the three categories of wards.