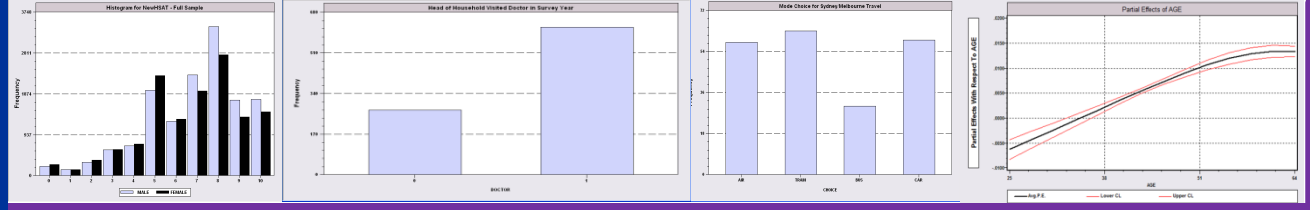
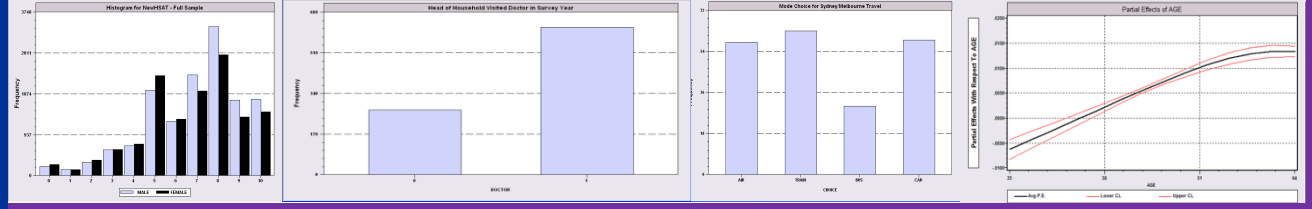


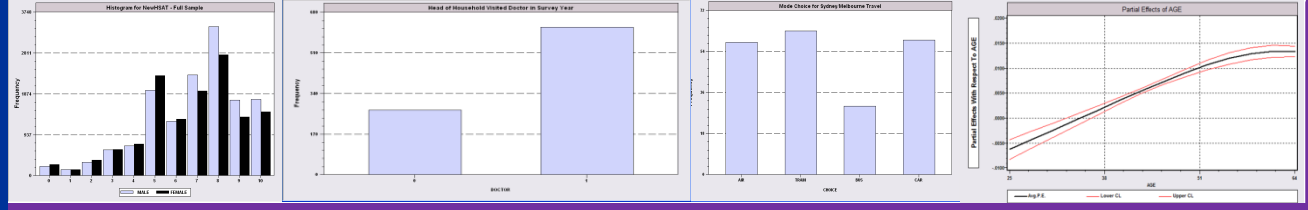
Discrete Choice Modeling

William Greene
Stern School of Business
New York University



Objectives in Model Building

- **Specification**: guided by underlying theory
  - Modeling framework
  - Functional forms
- **Estimation**: coefficients, partial effects, model implications
- **Statistical inference**: hypothesis testing
- **Prediction**: individual and aggregate
- **Model assessment** (fit, adequacy) and evaluation
- **Model extensions**
  - Interdependencies, multiple part models
  - Heterogeneity
  - Endogeneity and “causal inference”
- **Exploration**: Estimation and inference methods



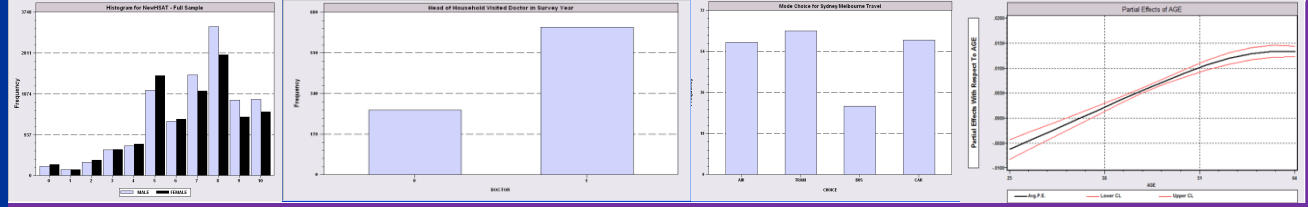
Regression Basics

The “MODEL”

- Modeling the conditional mean – Regression

Other features of interest

- Modeling quantiles
- Conditional variances or covariances
- Modeling probabilities for discrete choice
- Modeling other features of the population



Application: Health Care Usage

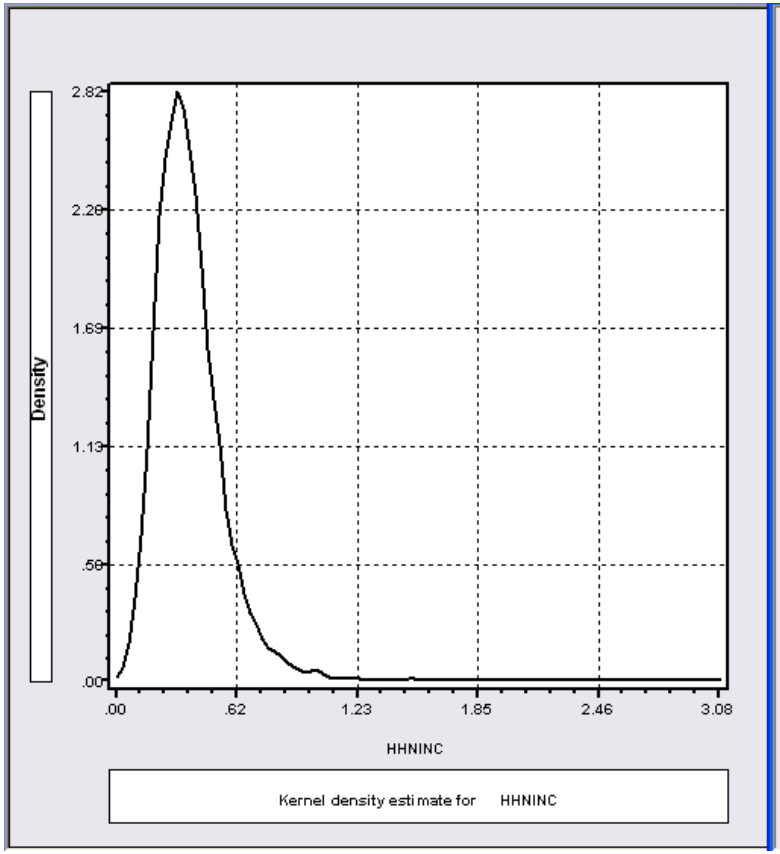
German Health Care Usage Data, 7,293 Individuals, Varying Numbers of Periods

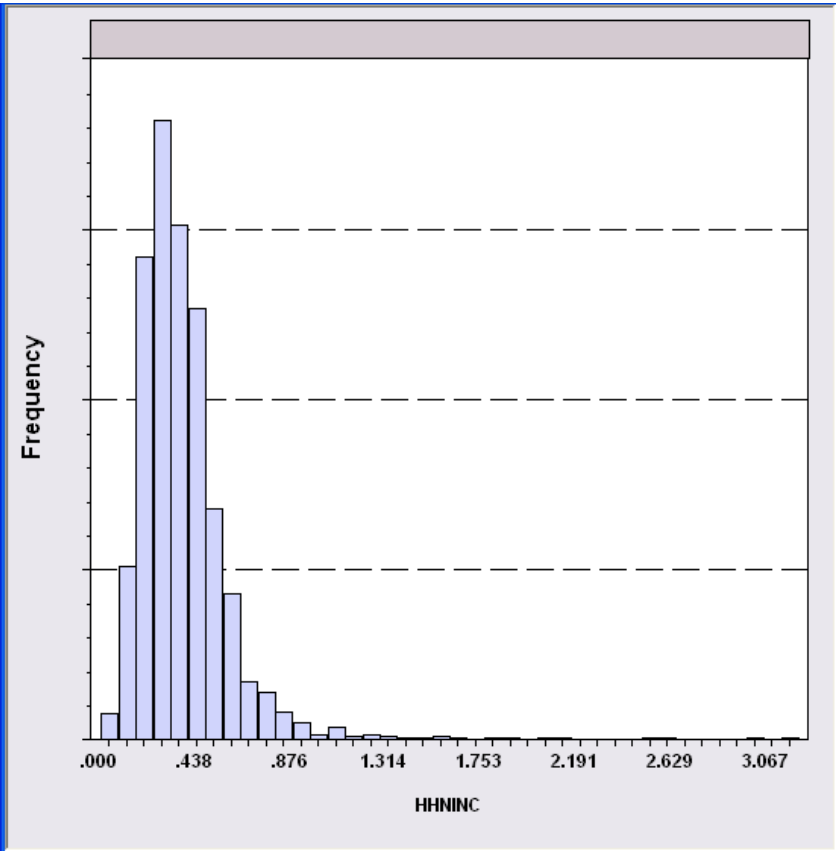
Data downloaded from Journal of Applied Econometrics Archive. This is an unbalanced panel with 7,293 individuals. They can be used for regression, count models, binary choice, ordered choice, and bivariate binary choice. This is a large data set. There are altogether 27,326 observations. The number of observations ranges from 1 to 7. (Frequencies are: 1=1525, 2=2158, 3=825, 4=926, 5=1051, 6=1000, 7=987).
(Downloaded from the JAE Archive)

Variables in the file are

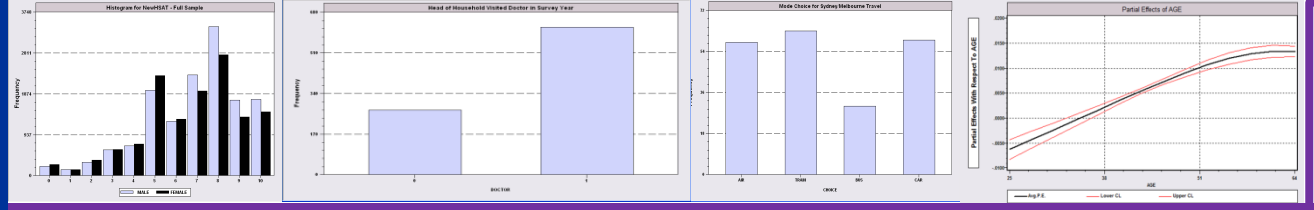
DOCTOR	=	**1(Number of doctor visits > 0)**
HOSPITAL	=	**1(Number of hospital visits > 0)**
HSAT	=	**health satisfaction, coded 0 (low) - 10 (high)**
DOCVIS	=	**number of doctor visits in last three months**
HOSPVIS	=	**number of hospital visits in last calendar year**
PUBLIC	=	**insured in public health insurance = 1; otherwise = 0**
ADDON	=	**insured by add-on insurance = 1; otherswise = 0**
HHNINC	=	**household nominal monthly net income in German marks / 10000**. (4 observations with income=0 were dropped)
HHKIDS	=	**children under age 16 in the household = 1; otherwise = 0**
EDUC	=	**years of schooling**
AGE	=	**age in years**
MARRIED	=	**marital status**

Household Income





Kernel Density Estimator **Histogram**

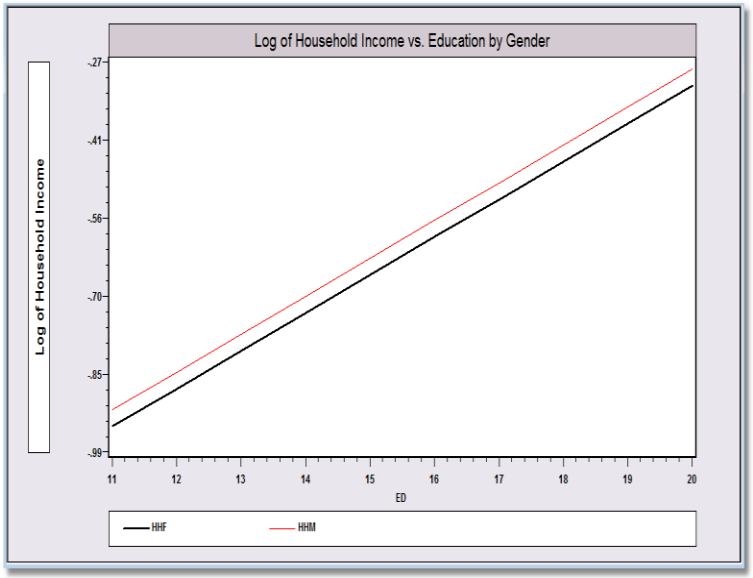


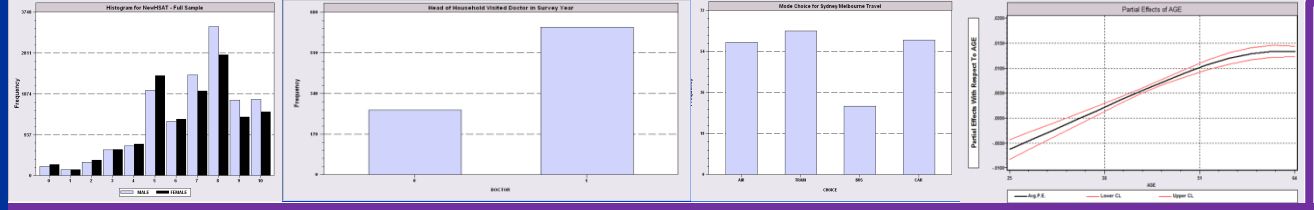
Regression – Income on Education

```
---------------------------------------------------------------------
Ordinary      least squares regression .............
LHS=LOGINC    Mean                     =         -.92882
              Standard deviation       =          .47948
              Number of observs.       =             887
Model size    Parameters               =               2
              Degrees of freedom       =             885
Residuals     Sum of squares           =      183.19359
              Standard error of e      =          .45497
Fit           R-squared                =          .10064
              Adjusted R-squared       =          .09962
Model test    F[  1,   885] (prob)     =    99.0(.0000)
Diagnostic    Log likelihood           =     -559.06527
              Restricted(b=0)          =     -606.10609
              Chi-sq [  1]   (prob)    =    94.1(.0000)
Info criter.  LogAmemiya Prd. Crt.     =       -1.57279
--------+------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+------------------------------------------------------------
Constant|   -1.71604***        .08057      -21.299   .0000
    EDUC|     .07176***        .00721        9.951   .0000       10.9707
--------+------------------------------------------------------------
Note: ***, **, * = Significance at 1%, 5%, 10% level.
---------------------------------------------------------------------
```

Specification and Functional Form

```
---------------------------------------------------------------------
Ordinary      least squares regression ............
LHS=LOGINC    Mean                    =          -.92882
              Standard deviation      =           .47948
              Number of observs.      =              887
Model size    Parameters              =                3
              Degrees of freedom      =              884
Residuals     Sum of squares          =       183.00347
              Standard error of e     =           .45499
Fit           R-squared               =           .10157
              Adjusted R-squared      =           .09954
Model test    F[  2,   884] (prob)    =     50.0(.0000)
Diagnostic    Log likelihood          =      -558.60477
              Restricted(b=0)         =      -606.10609
              Chi-sq [  2]   (prob)   =     95.0(.0000)
Info criter.  LogAmemiya Prd. Crt.    =        -1.57158
--------+------------------------------------------------------------
Variable| Coefficient     Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+------------------------------------------------------------
Constant|   -1.68303***        .08763      -19.207   .0000
    EDUC|     .06993***        .00746        9.375   .0000      10.9707
  FEMALE|    -.03065           .03199        -.958   .3379       .42277
--------+------------------------------------------------------------
```





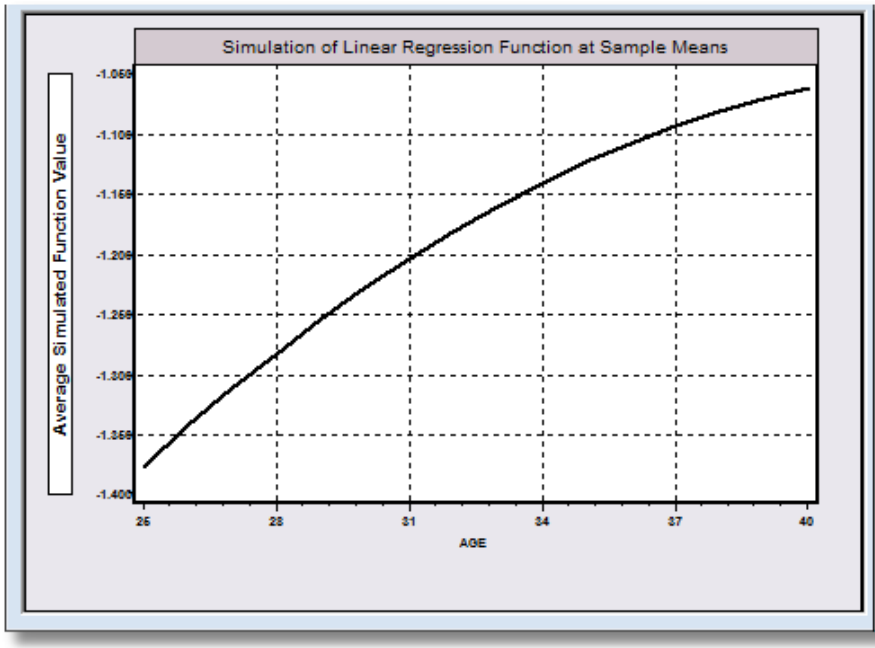
Interesting Partial Effects

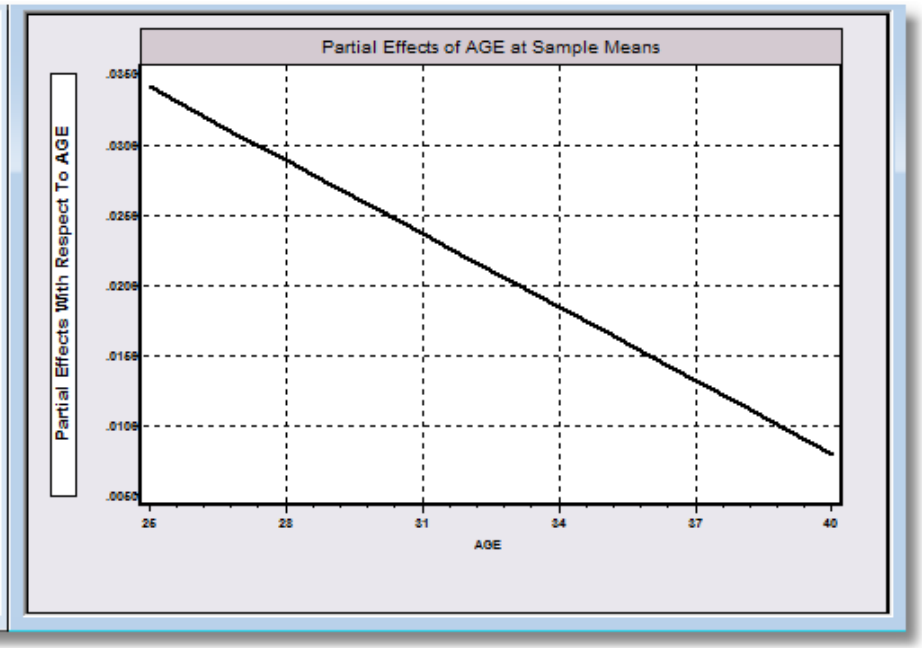
```
---------------------------------------------------------------------
Ordinary      least squares regression ............
LHS=LOGINC    Mean                    =         -.92882
              Standard deviation      =          .47948
              Number of observs.      =             887
Model size    Parameters              =               5
              Degrees of freedom      =             882
Residuals     Sum of squares          =      171.87964
              Standard error of e     =          .44145
Fit           R-squared               =          .15618
              Adjusted R-squared      =          .15235
Model test    F[  4,   882] (prob)    =    40.8(.0000)
Diagnostic    Log likelihood          =     -530.79258
              Restricted(b=0)         =     -606.10609
              Chi-sq [  4]  (prob)    =   150.6(.0000)
Info criter.  LogAmemiya Prd. Crt.    =       -1.62978
--------+------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+------------------------------------------------------------
Constant|   -5.26676***       .56499      -9.322   .0000
    EDUC|     .06469***       .00730       8.860   .0000      10.9707
  FEMALE|    -.03683          .03134      -1.175   .2399       .42277
     AGE|     .15567***       .02297       6.777   .0000      50.4780
   AGE^2 |    -.00161***       .00023      -7.014   .0000      2620.79
--------+------------------------------------------------------------
```

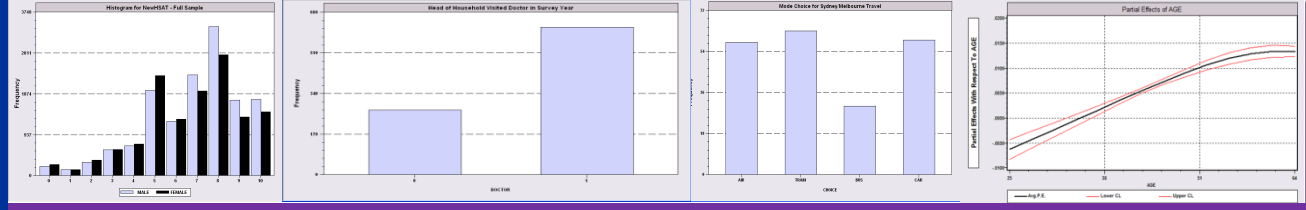
$$\frac{\partial E[Income\,|\,\mathbf{x}]}{\partial Age} = \beta_{Age} + 2\,Age\,\beta_{Age^2}$$

Function: Log Income | Age

Partial Effect wrt Age

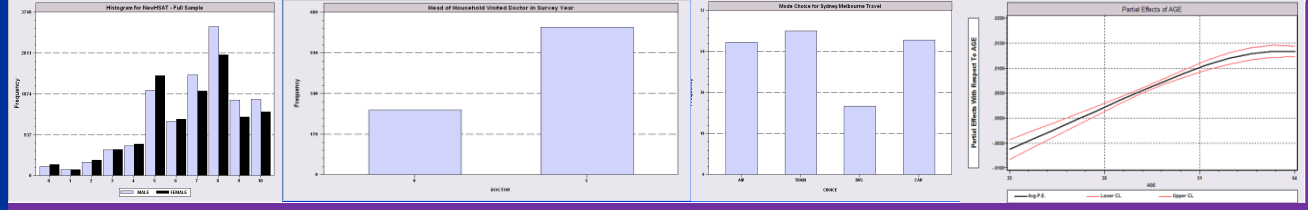






Modeling Categorical Variables

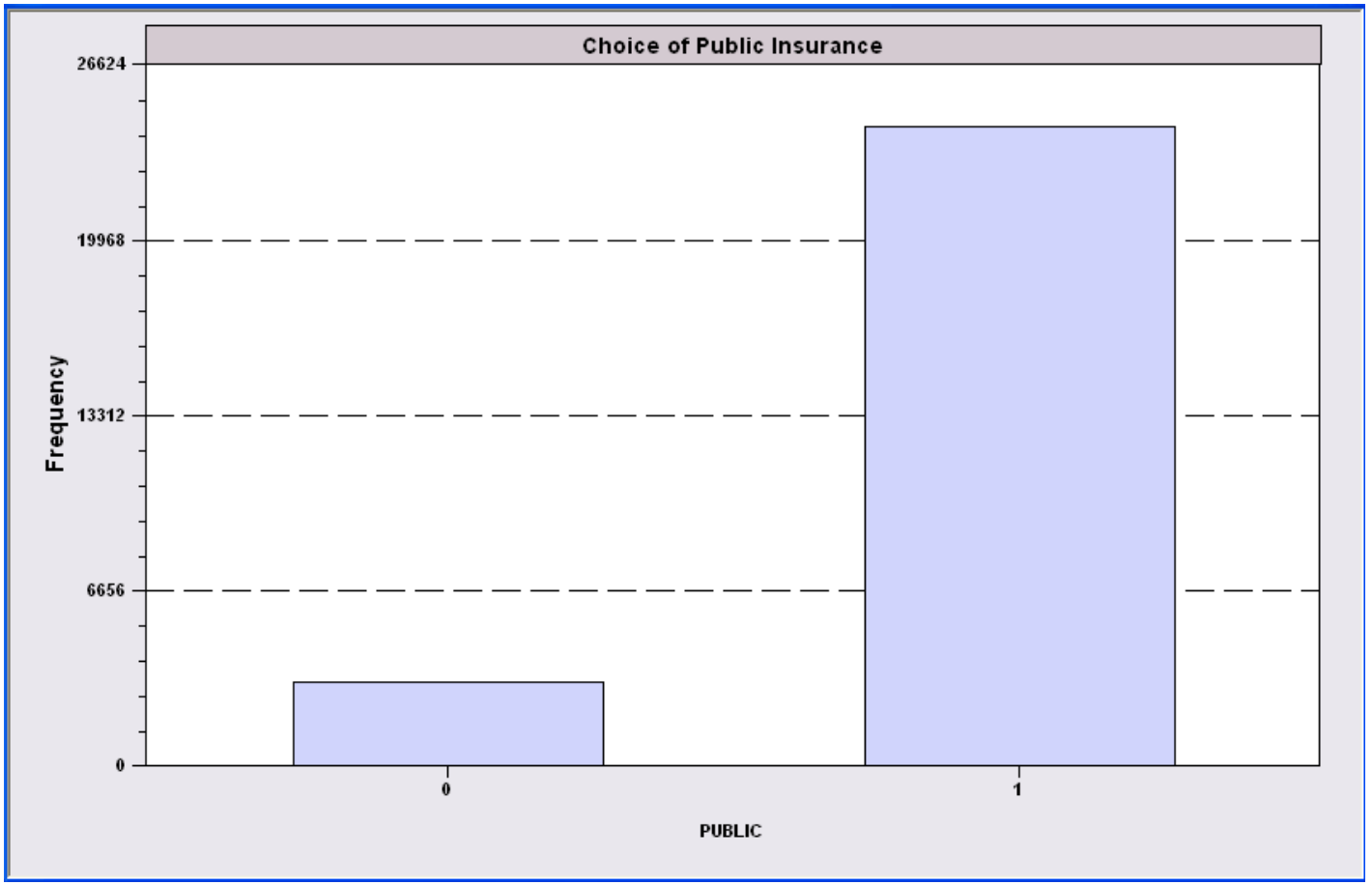
- Theoretical foundations
- Econometric methodology
  - Models
  - Statistical bases
  - Econometric methods
- Applications



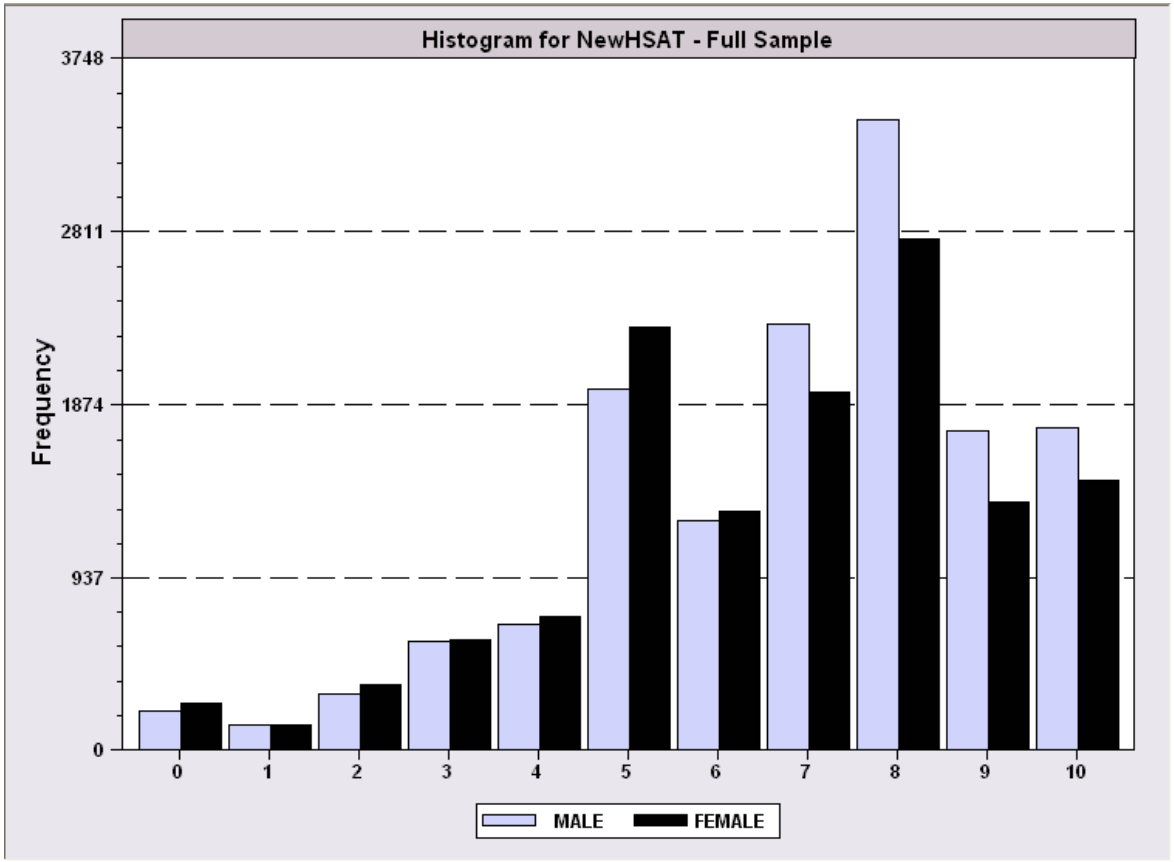
Categorical Variables

- Observed outcomes
  - Inherently discrete: number of occurrences, e.g., family size
  - Multinomial: The observed outcome indexes a set of unordered labeled choices.
  - Implicitly continuous: The observed data are discrete by construction, e.g., revealed preferences; our main subject
  - Discrete, cardinal: Counts of occurrences
- Implications
  - For model building
  - For analysis and prediction of behavior

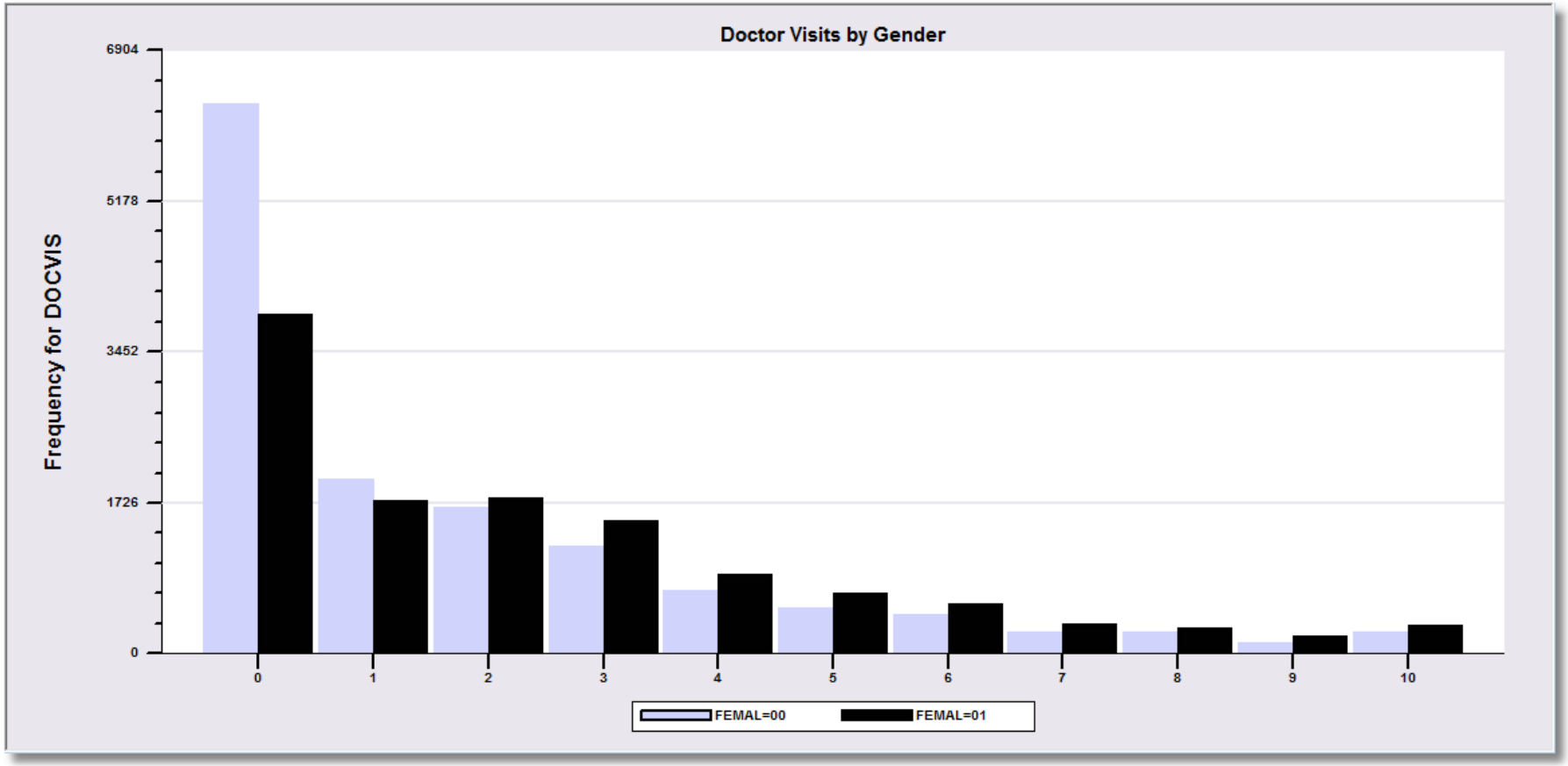
Simple Binary Choice: Insurance



Ordered Outcome
Self Reported Health Satisfaction



Counts of Occurrences



Multinomial Unordered Choice