Analysis of Home Ownership in Kentucky

Blake Weimer, Guysnove Lutumba, and Matt Henn

Our Interests

- Owning and renting are both considerable financial choices.
- There are many factors that influence home ownership status:
  - Marriage status, location, age, income, family size, education, poverty
- Understanding these factors gives an insight to people's financial choices.

Research Goals

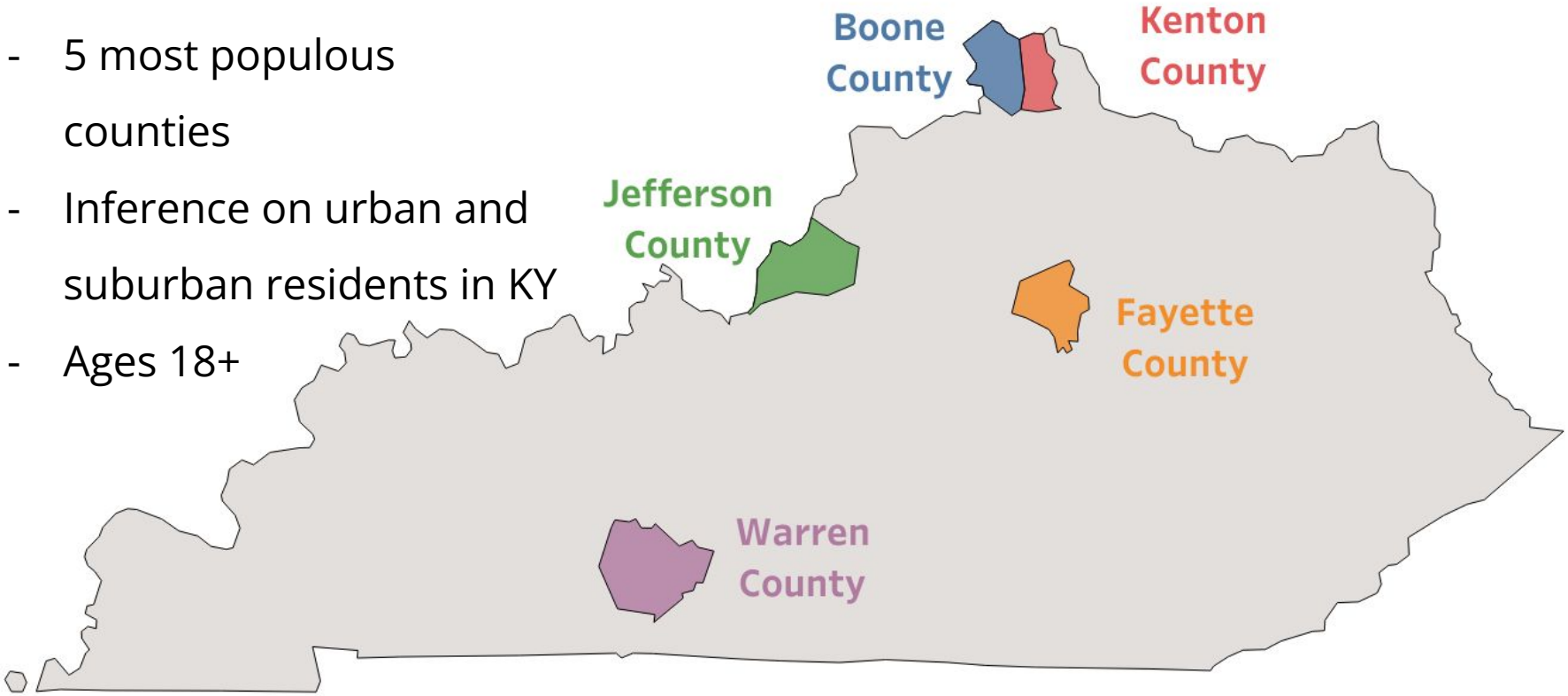
- Interested in Kentuckian home ownership status.
- Research Question 1
  - Is there a relationship between certain variables and home ownership status?
  - What is the full relationship between our variables?
  - How are our variables related? Is there any collapsibility?
- Research Question 2
  - Can we predict the probability of success of owning a home using one of our variables?
  - Using all of our explanatory variables, can we build the best and most appropriate model using logistic regression?

American Community Survey Data

- Data downloaded from the Census Bureau site
- Collected 56,000 observations from Kentucky from the past 5 years
- Randomly selected 10,000 observations without replacement using R `sample` function

Population

- 5 most populous counties
- Inference on urban and suburban residents in KY
- Ages 18+



Variables in the Dataset

Categorical

Ownership - Own / Rent

Marriage Status - Married / Not Married

County - 5 KY counties

In Poverty - Yes / No

Sex - Male / Female

College Degree - Yes / No

Quantitative

Age - 18 to 95

Household Income - -7,900 to 1,287,500

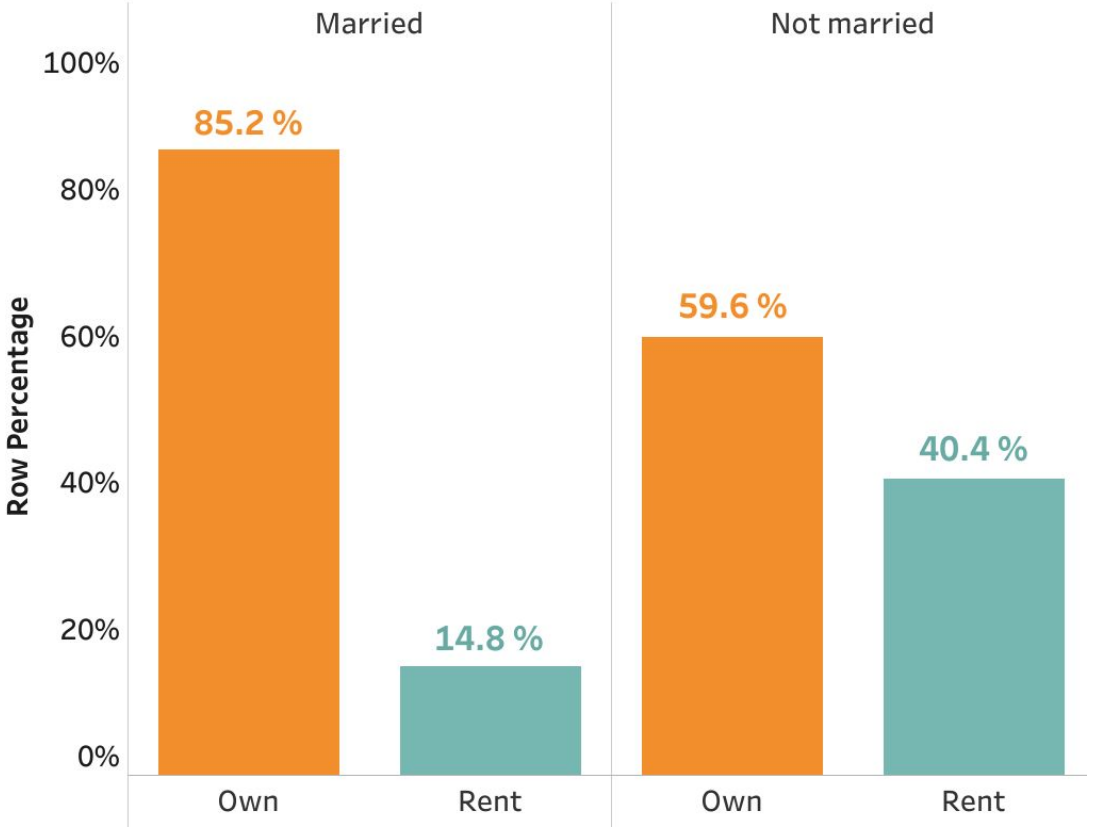
Family Size - 1 to 14

Research Question 1 Goals

- Does your marriage status (X) influence your home ownership status (Y) ?
  - Perform Chi-Square test for Independence
  - Create & Interpret a C.I. for the Relative Risk
- What is the full relationship of our variables when accounting for county?
  - Check Simpson’s Paradox, CMH Test, Breslow-Day Test, and Collapsibility
- Fully investigate and interpret results

2x2 Table

Marriage Status	Ownership Status: Own	Ownership Status: Rent
Married	4,579	795
Not Married	2,759	1,867



Test for Independence

H_0= Ownership is independent of marriage status

H_A= Ownership is dependent of marriage status

RR: Reject H_0 if $X^2 > 5.9915$ `qchisq(0.95, 2)`

Test Statistic: $X^2 = 836.8 \Rightarrow$ in RR, reject H_0

X^2 Contributions:

Ownership Status

Marriage Status	Own	Rent
Married	102.4	287.4
Not Married	119.0	328.0

Interpretation:

At the $\alpha = 0.05$ level, there is statistically significant evidence to support that home ownership is dependent on marriage status

Follow-up:

What is the risk of renting for those who are not married to those who are married?

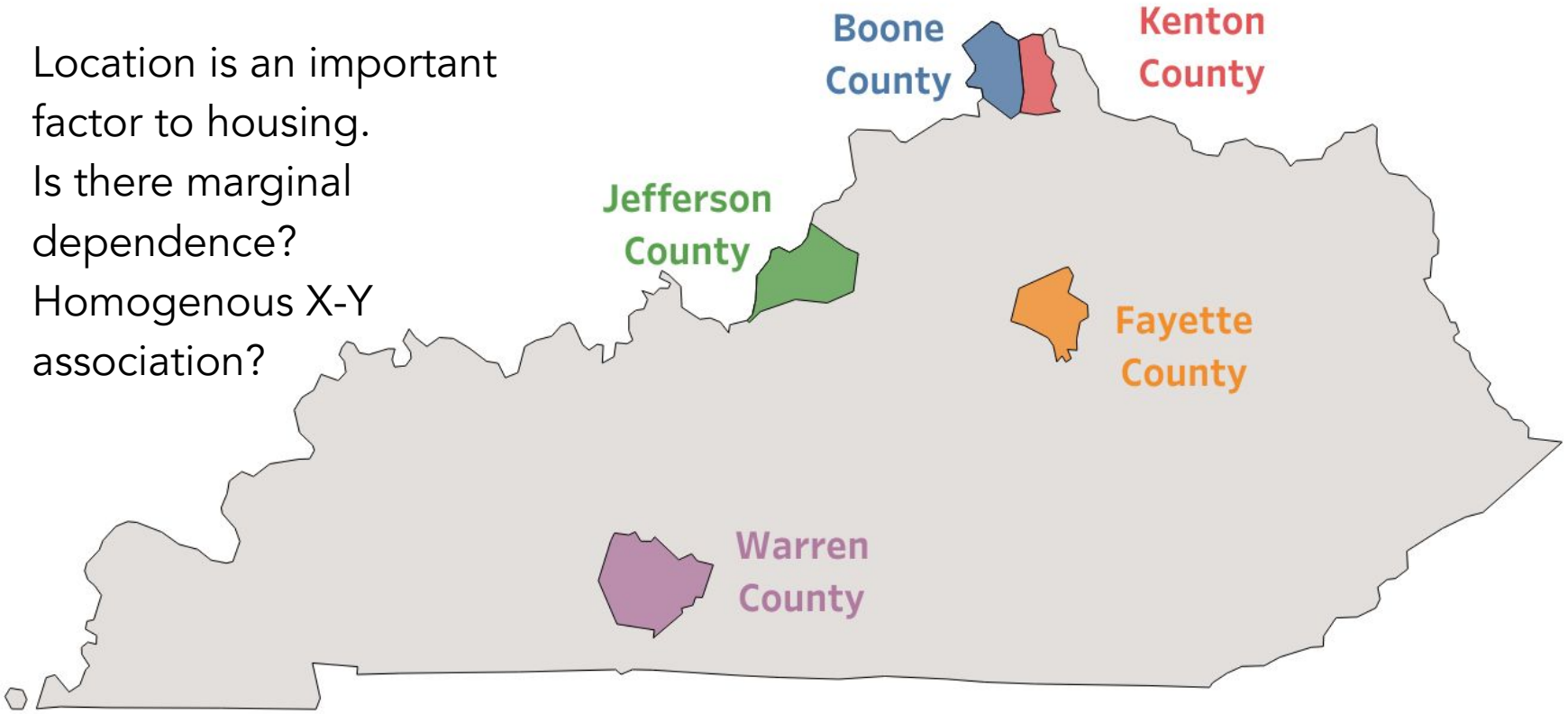
Relative Risk 95 % Confidence Interval

Risk Married	Risk Not Married	Relative Risk (Not Married / Married)	Confidence Interval
0.1479	0.4036	2.7282	(2.5358, 2.9351)

We are 95% confident that the risk of renting is between 2.5358 and 2.9351 times as likely for those who are not married than those who are married.

County as an Indicator of Ownership

- Location is an important factor to housing.
- Is there marginal dependence?
- Homogenous X-Y association?



2x2x5 Table - Accounting for County

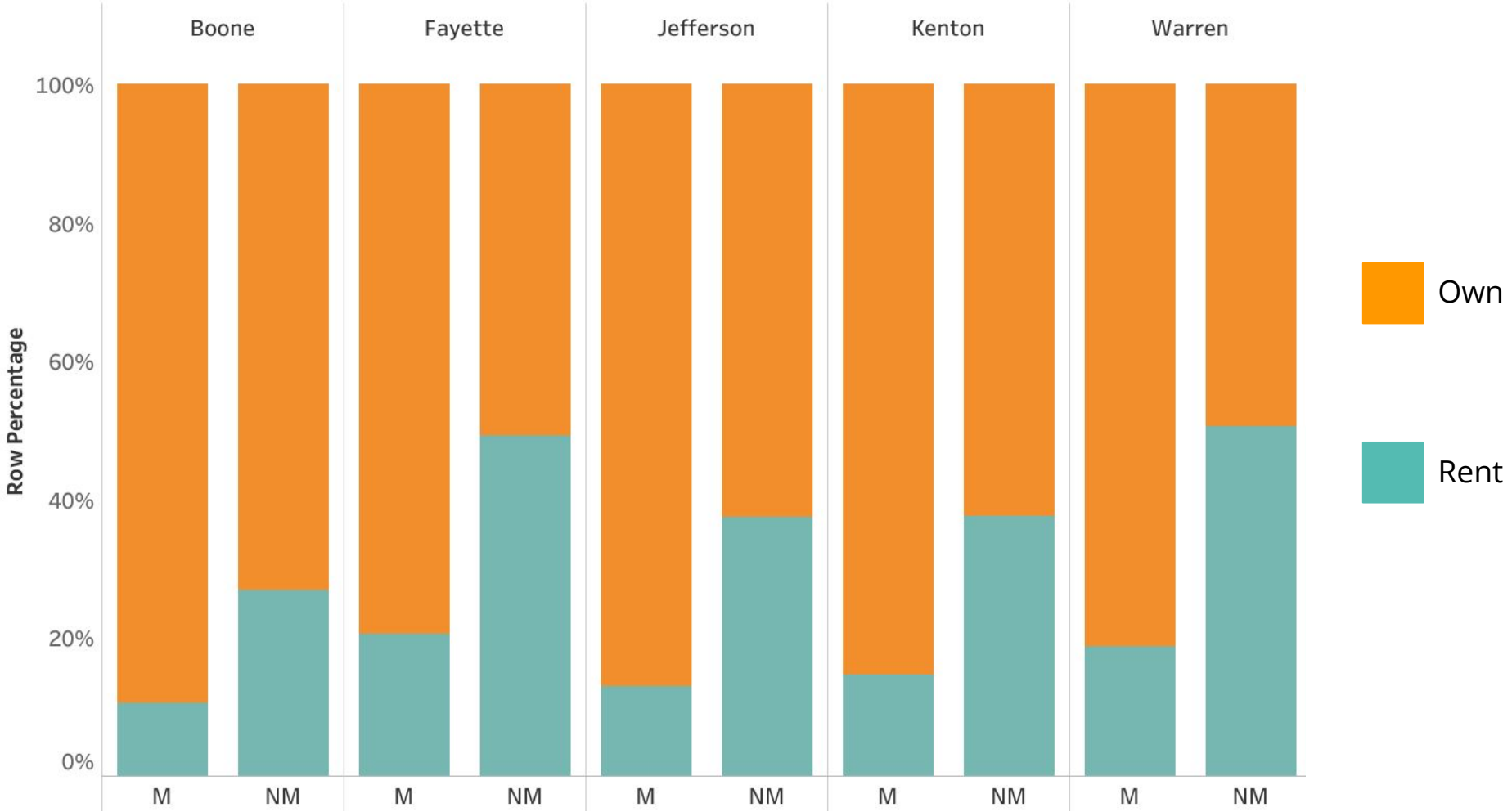
Boone	Own	Rent
Married	358	42
Not Married	150	55

Fayette	Own	Rent
Married	807	209
Not Married	502	486

Jefferson	Own	Rent
Married	2,657	398
Not Married	1628	486

Kenton	Own	Rent
Married	479	82
Not Married	327	197

Warren	Own	Rent
Married	278	64
Not Married	152	155


Boone
Fayette
Jefferson
Kenton
Warren
100%
80%
60%
40%
20%
0%
Row Percentage
M
NM
Own
Rent

Checking Simpson's Paradox

Table	Odds Ratio **θ**
Combined / Marginal	3.8976
Boone	3.1254
Fayette	3.7382
Jefferson	3.8976
Kenton	3.5192
Warren	4.4295

- Odds **θ** =

$$\frac{\text{(own / rent)}}{\text{(married / not married)}}$$

- All θ's > 1 (i.e same direction of 1)
- Simpson's Paradox is not an issue.

CMH Test

H_0= Conditional independence of $\boldsymbol{\theta}$'s

H_A= At least one $\boldsymbol{\theta}_{xy} \neq 1$

RR: Reject H_0 if $X^2 > 3.8415$ `qchisq(0.95, 1)`

Conditions:

All expected counts ≥ 5 (smallest = 32.87 Boone, Rent, Not Married)

All $\boldsymbol{\theta}$'s are in same direction as 1

Test Statistic:

$X^2 = 809.6 \Rightarrow$ in rejection region, reject H_0

Interpretation:

At the $\alpha = 0.05$ level, there is statistically significant evidence to support that at least one of the $\boldsymbol{\theta}_{xy(\ )}$ for counties is not equal to 1.

Follow-up:

Although all $\boldsymbol{\theta}_{xy}$ are not 1, are they equal? Is there homogenous X-Y association?

Breslow-Day Test

$H_0 = \boldsymbol{\theta}_B = \boldsymbol{\theta}_F = \boldsymbol{\theta}_J = \boldsymbol{\theta}_K = \boldsymbol{\theta}_W$

H_A = At least one $\boldsymbol{\theta}_{xy}$ is different

P-value: Reject H_0 if $p < \alpha = 0.05$

Conditions:

All expected counts ≥ 5 (smallest = 42 Boone, Rent, Married)

All table totals ≥ 50 (smallest = 605 Boone)

Test Statistic: `breslowday.test`

HBDT = 2.1957

P-value = 0.6998 ⇒ above α, fail to reject

Interpretation:

At the α = 0.05 level, there is not statistically significant evidence to support that the **θ**'s (odds ratios) for counties are different.

Follow-up:

Use CMH common odds ratio estimate for a CI. Check collapsibility.

CMH Common Odds and Collapsibility

CMH Common Odds θ	95 % Confidence Interval
3.8666	(3.5121, 4.2569)

We are 95% confident that the odds of owning to renting for those who are married are between 3.5121 and 4.2569 times as large than those who are not married.

Since Simpson's Paradox is not a concern and we failed to reject homogenous X-Y association, we may collapse over county.

Summary of Q1 Results

- Evidence that ownership status is dependent on marriage status
  - 95% confident the risk of renting is 2.5358 and 2.9351 times as likely for those who are not married than those who are married.
- Simpson's Paradox is not a concern
  - all **θ**'s in the same direction of 1.
- Evidence that there is some sort of conditional dependence
  - not all **θ**'s equal 1.
- No evidence to reject that there is homogenous XY association
  - all **θ**'s are equal.
  - CMH Common Odds confidence interval
- We may collapse over county.

Research Question 2 Goals

- Predict the probability of owning a home using the most appropriate quantitative variable and determine which link function is best.
  - Variable Options: Household Income, Age, Family Size
  - Link Function Options: Logit, Probit, Cloglog
- Predict the probability of owning a home using all explanatory variables and build the best and most appropriate model using logistic regression.
  - Calculate confidence intervals for odds ratios **θ's**
  - Find appropriate predicted probabilities

1 Quantitative Variable Model

Dependent Variable

indownership - 0 if rent, 1 if own

Possible Independent Variables

hhincome - household income of individual

agequant - age of individual

famsize - family size of individual

Model and AIC Value

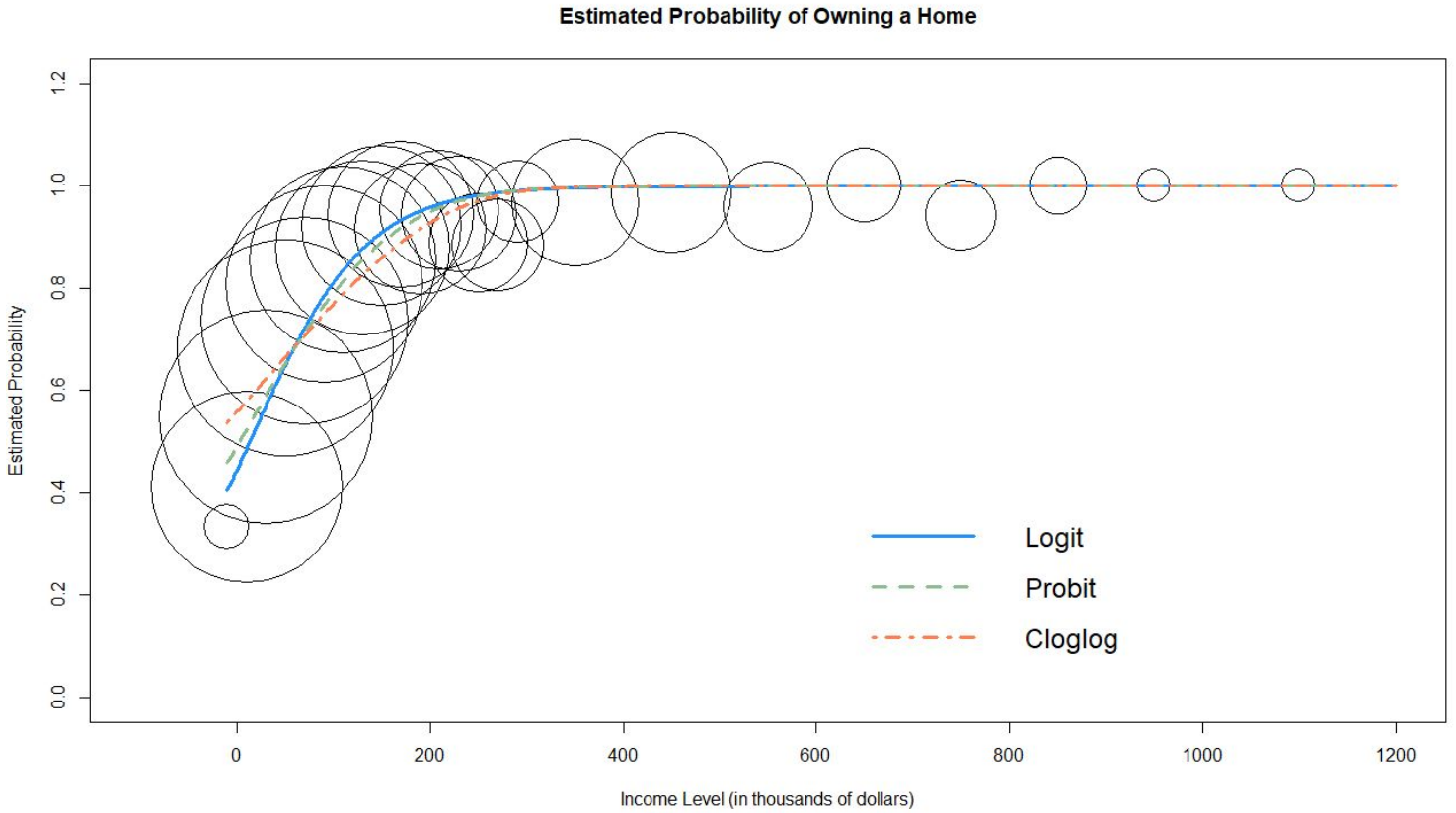
Variable	Logit	Probit	Cloglog
hhincome	10,220	10,330	10,540
agequant	10,850	10,870	10,890
famsize	11,480	11,480	11,500

Model Fitting - Using x = hhincome

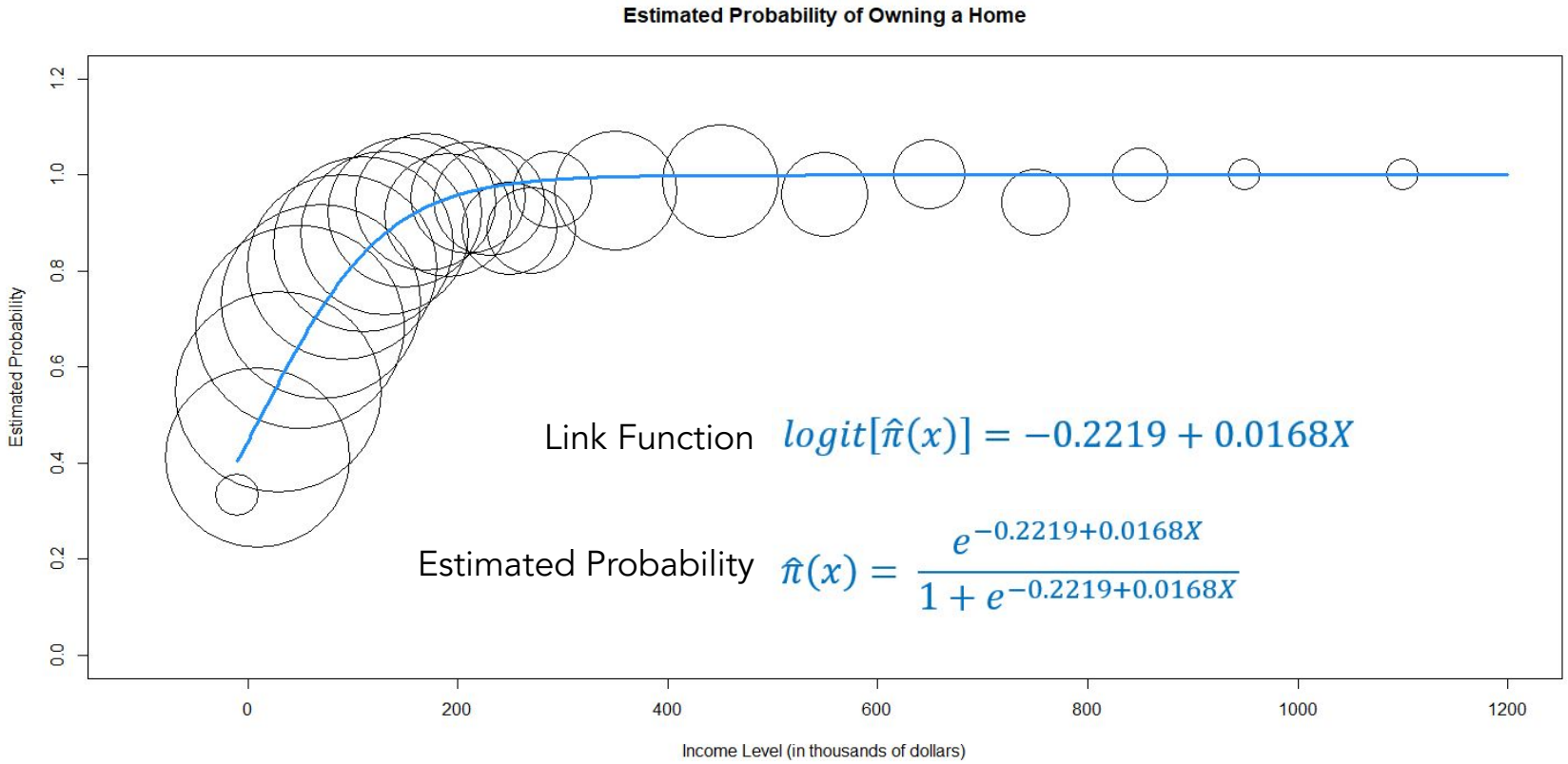
	Estimated Probability Equation	AIC
Logit	$\hat{\pi}(x) = \frac{e^{-0.2219+0.0168X}}{1+e^{-0.2219+0.0168X}}$	10,220
Probit	$\hat{\pi}(x) = \Phi\left(-0.0215 + 0.0083X\right)$	10,330
Cloglog	$\hat{\pi}(x) = 1 - e^{-e^{-0.2029+0.0059X}}$	10,540

The Logit model has the lowest AIC.

Model Fitting - Bubble Plot



Model Fitting - Logit Model



Link Function $logit[\hat{\pi}(x)] = -0.2219 + 0.0168X$

Estimated Probability $\hat{\pi}(x) = \frac{e^{-0.2219+0.0168X}}{1+e^{-0.2219+0.0168X}}$

Full Model Variables

Dependent

indownership - 0 if rent, 1 if own

Indicator Explanatory

indmarrst - 0 if not married, 1 if married

indpov - 0 if not in poverty, 1 if in poverty

indsex - 0 if female, 1 if male

inddeg - 0 if no degree, 1 if degree

Quantitative Explanatory

hhincome - household income of individual

age- age of individual

famsize - family size of individual

Fitting Variables Separately

Variables	P-value
hhincome	0
agequant	0
famsize	0
indmarrst	0
indpov	0
indsex	0.2105
inddeg	0

- α to leave is 0.20
- indsex is not significant on its own, so it will not be included during backwards elimination.

Backwards Elimination

	Estimate	Standard error	Z-value	P-value
Intercept	-2.4706	0.1150	-21.48	2e-16 ***
hhincome	0.0145	0.0007	20.84	2e-16 **
age	0.0402	0.0016	25.17	2e-16 ***
famsize	0.0966	0.0213	4.68	2.92e-06 ***
indmarrst	0.5742	0.0588	9.76	2e-16 ***
indpov	-0.6120	0.0835	-7.33	2.36e-13 ***
inddeg	0.1771	0.0564	3.14	0.0017 **

Significance codes: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1,

All variables are significant in the full model without the indicator for sex.

Model Equation

$$logit[\hat{\pi}(x)] = -2.4706 + 0.0145(hhincome) + 0.0402(age) + 0.0996(famsize) + 0.5742(indmarst) - 0.6120(indpov) + 0.1771(inddeg)$$

$$\hat{\pi}(x) = \frac{e^{-2.4706+0.0145(hhincome)+0.0402(age)+0.0996(famsize)+0.5742(indmarst)-0.6120(indpov)+0.1771(inddeg)}}{1 + e^{-2.4706+0.0145(hhincome)+0.0402(age)+0.0996(famsize)+0.5742(indmarst)-0.6120(indpov)+0.1771(inddeg)}}$$

Odds Ratios 95 % Confidence Intervals

Variable	Constant	Odds Ratio	Formula	Confidence Interval
hhincome	10 (thousands)	1.1558	$e^{10*0.0145\pm1.96*10*0.0007}$	(1.1402, 1.1717)
indmarrst	1	1.7758	$e^{0.5742\pm1.96*0.0588}$	(1.5823, 1.9928)
indpov	1 (decrease)	1.8441	$\frac{1}{e^{-0.6120\pm1.96*0.0835}}$	(1.5656,2.1721)

Interpretations

Variable	Confidence Interval	Assuming all other variables remain constant…
hhincome	(1.1402, 1.1717)	We are 95% confident that the odds of owning your home to renting are between 1.1402 and 1.1717 times as large when household income is increased by 10,000 dollars.
indmarrst	(1.5823, 1.9928)	We are 95% confident that the odds of owning your home to renting are between 1.5823 and 1.9928 times as large when you are married than when you are not.
indpov	(1.5656,2.1721)	We are 95% confident that the odds of owning your home to renting are between 1.5656 and 2.1721 times as large if you are not in poverty than if you are in poverty.

Predicted Probabilities - Recent Stats Grad

Variable	Value
hhincome	50
age	23
famsize	1
indmarrst	0
indpov	0
inddeg	1

The estimated probability of owning a home for an individual with a household income of 50,000, age 23, family size of 1, not married, not in poverty, and holding a college degree is approximately 0.3670.

$$\hat{\pi}(x) = \frac{e^{-2.4706+0.0145(50)+0.0402(23)+0.0996(1)+0.5742(0)-0.6120(0)+0.1771(1)}}{1+e^{-2.4706+0.0145(50)+0.0402(23)+0.0996(1)+0.5742(0)-0.6120(0)+0.1771(1)}} = 0.3670$$

Predicted Probabilities - Avg Kenton Resident

Variable	Value
hhincome	78
age	40
famsize	4
indmarrst	1
indpov	0
inddeg	1

The estimated probability of owning a home for an individual with a household income of 78,000, age 40, family size of 4, married, not in poverty, and holding a college degree is approximately 0.8049.

$$\hat{\pi}(x) = \frac{e^{-2.4706+0.0145(78)+0.0402(40)+0.0996(4)+0.5742(1)-0.6120(0)+0.1771(1)}}{1+e^{-2.4706+0.0145(78)+0.0402(40)+0.0996(4)+0.5742(1)-0.6120(0)+0.1771(1)}} = 0.8049$$

Summary of Q2 Results

- Income best quantitative variable to model the probability of owning a home.
- Logit most appropriate link function to model the probability of owning a home.
- Sex was not significant in modeling the probability of owning a home
- Income, age, family size, marriage status, and college degree had all a positive influence on the probability of owning a home.
- Poverty had a negative influence on the probability of owning a home.
- Among variables that positively impacted the probability of owning a home, marriage status had a huge influence.

Thank you for listening!