

Regression III

Lecture 1: Preliminary

Dave Armstrong¹

University of Wisconsin – Milwaukee
Department of Political Science

e: armstrod@uwm.edu
w: www.quantoid.net/ICPSR.php

¹These slides have been heavily influenced by the previous instructor of this course - Bob Andersen (U of Toronto)

1 / 43

Outline

Introduction

OLS Basics: Theory and Presentation of Results

Non-Linearity

Outliers and Influential Data

Other Topics

2 / 43

Contact Info

- Instructor: Dave Armstrong
E-mail: armstrod@uwm.edu
Course Website: <http://www.quantoid.net/ICPSR.php>
Office Hours: 2-3PM M-F (or by Appointment)
- Teaching Assistant: Kelly Gleason
E-mail: kgleason@uwm.edu
Website: <http://www.uwm.edu/~kgleason>
Office Hours: 3-5PM M-F (or by Appointment)

3 / 43

On Notecard

1. Name
2. e-mail address
3. Institutional Affiliation and Department
4. Research Interests
5. Other ICPSR Courses
6. Previous Stats Courses
7. Preferred Stats software

4 / 43

Today's Lecture

1. Show some “highlights” of the course:
 - Applied Regression with attention to *modern* extensions.
 - Explore methods for “problem” data.
 - Emphasis on graphical techniques.
 - Use modern methods to overcome problems with regular linear regression.
2. Prerequisites:
 - 2.1 Regression (in matrix form),
 - 2.2 Understanding of Statistical Inference,
 - 2.3 MLE (would be nice, but not a pre-requisite *per se*)
3. Computing...
 - All analysis will be done using **R**.
 - The course on *Computing in R and S* is compulsory (unless you already know **R**).

5 / 43

Course Materials

Suggested Texts:

- Fox, John. (2008) Applied Regression Analysis and Generalized Linear Models, *2nd ed.* Thousand Oaks, CA: Sage Publications, Inc.
- Fox, John. (2011) An R and S-PLUS Companion to Applied Regression, *2nd ed.* Thousand Oaks, CA: Sage Publications, Inc.

A more detailed list is at the back of the course syllabus

- Course materials - lecture slides, syllabus and **R**-scripts will be available at
 - Z:/Armstrong and
 - <http://www.quantoid.net/ICPSR.php>

6 / 43

Introduction

OLS Basics: Theory and Presentation of Results

Non-Linearity

Outliers and Influential Data

Other Topics

7 / 43

Versatility of Linear Regression

- Simple linear regression summarizes the relationship between a quantitative predictor variable and quantitative response variable with a straight line
- The linear model can be extended to handle:
 - Several explanatory variables (multiple regression)
 - Categorical explanatory variables and interactions between explanatory variables
 - Categorical dependent variables
 - Simple and monotone nonlinear relationships
- The linear model is desirable because it is simple to fit and easy to interpret
- The data must satisfy a number of assumptions, however, for the linear model to be the appropriate model
- Looking at the data graphically allows us to assess whether these assumptions are likely to be met

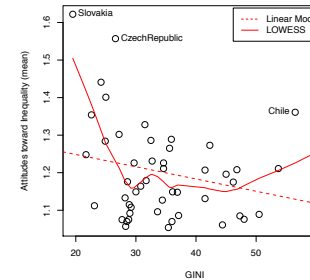
8 / 43

Regression and Causation

- When observational data are used, statistical models are inherently descriptive. . . they are not causal though we often want to describe a causal process
- Criteria for a causal theory
 - Empirical relationship
 - Cause precedes the effect in time
 - Elimination of rival explanations (i.e., no confounding variables)
 - We must control for possible confounding and intervening variables when using observational data
- The goal is to fit the model that tells the right story
 - We must ensure that our models adequately represent the patterns in the data, without over-fitting.
 - When making inferences about populations, we must also ensure that the assumptions of the model are met

9 / 43

Importance of Control Variables



(Intercept)	1.31*
gini	-0.00 (0.00)
N	49
R ²	0.06
adj. R ²	0.04
Resid. sd	0.12

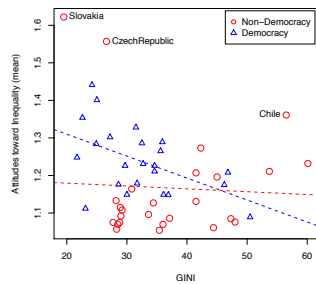
Standard errors in parentheses

* indicates significance at $p < 0.05$

- Relationship is clearly not linear: at high levels of income inequality, attitudes towards inequality are negatively related to the Gini coefficient; at low levels of inequality the trend is in the opposite direction
- There also may be influential outliers
- Next step: Explore outliers, possible control variables and interactions

10 / 43

Control Variables and Interactions



(Intercept)	1.19*
gini	-0.00 (0.09)
democrat	0.23 (0.00)
gini:democrat	-0.01 (0.14)
N	49
R ²	0.14
adj. R ²	0.08
Resid. sd	0.12

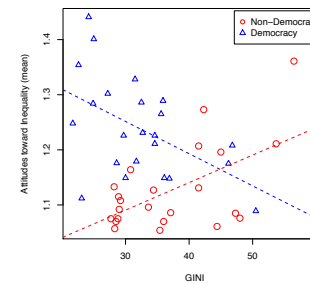
Standard errors in parentheses

* indicates significance at $p < 0.05$

- Addition of democracy and its interaction with Gini coefficient (allowing for different slopes for the two groups) improves the model, but the effects are still not statistically significant
- Two outliers (Slovakia and Czech Republic) are clearly influential for Non-democracy model

11 / 43

Influential Cases



(Intercept)	0.94*
gini	0.00*
democrat	0.49*
gini:democrat	-0.01*
N	47
R ²	0.52
adj. R ²	0.48
Resid. sd	0.07

Standard errors in parentheses

* indicates significance at $p < 0.05$

- Czech Republic and Slovakia have unusually high levels of attitudes toward inequality
- When these cases are removed from the model there is a vast improvement in fit
- Both the Gini coefficient and Democracy have significant effects on attitudes and there is a strong interaction between them

12 / 43

Assumptions of Multiple Regression

The multiple regression model takes the following form:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$$

The assumptions of the model for ordinary least squares regression (OLS) concern the errors:

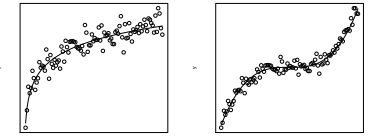
1. Linearity
2. Constant error variance
3. Normally distributed errors
4. Uncorrelated error terms
5. X's are independent of the errors

When these assumptions are met, the OLS estimators are unbiased and efficient estimates of the population parameters

13 / 43

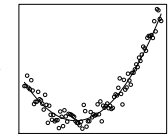
OLS and Nonlinearity(1): Transformable Nonlinearity

- Transformations of one or both variables can help straighten the relationship between two quantitative variables
- Possible only when the nonlinear relationship is simple and monotone
 - Simple implies that the curvature does not change – there is one curve
 - Monotone implies that the curve is always positive or always negative



(a)

(b)

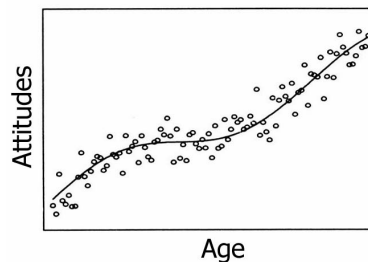


(c)

14 / 43

OLS and Nonlinearity (2): Polynomial Regression

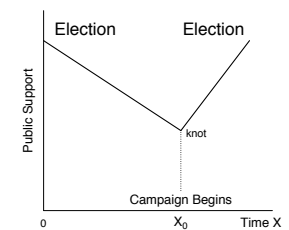
- When the relationship is not monotone or simple, we could try polynomial regression
- If there is only one bend in the curve, we fit a quadratic model – i.e., we could add an X^2 (age^2) term to the model
 - For every bend in the curve, we add another higher term to the model
- The two bends below suggest trying a cubic regression (i.e., include age , age^2 and age^3 as predictors)



15 / 43

OLS and Nonlinearity(3): Regression Splines

- Regression splines allow the regression line to change direction abruptly
- Piecewise polynomial functions that are constrained to join smoothly at points called knots
 - These are regression models with restricted dummy regressors
 - Separate regression lines are fit within regions (i.e., the range of X is partitioned) that join at knots



16 / 43

Handling Complex Nonlinearity: A more general way to think of regression

- Regression traces the conditional distribution of a dependent variable, Y , as a function of one or more explanatory variables, X 's

$$p(y|x_1, \dots, x_k) = f(x_1, \dots, x_k)$$

- Linear regression assumes a linear function but often this is not appropriate
- If the nonlinear relationship is complex, nonparametric regression and generalized additive models provide alternative ways of capturing the relationship
 - These models estimate the functional form from the data themselves (i.e., they do not assume linearity)

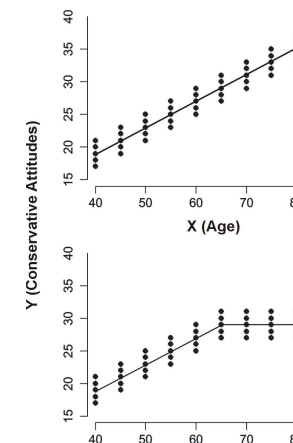
17 / 43

Modelling the Conditional Mean

- With large samples and when the values of X are discrete, it is possible to estimate the regression by directly examining the conditional distribution of Y
- Here we determine the mean of Y (could also use the median) at each value of X :

$$\mu = E(Y|x) = f(x)$$

- A naïve nonparametric regression line connects the conditional means
- Here, a linear regression would work well for the top graph, but poorly for the bottom graph.



18 / 43

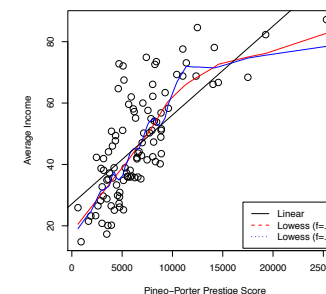
Modelling the Conditional Mean (2)

- With extremely large data sets or when the explanatory variable takes on discrete values, we can easily calculate conditional distributions
- In the 'real world' of social science data, however, we do not often have this luxury
 - If X is continuous, even when the sample size is large, we may not have enough cases at each value of X to calculate precise conditional means
- If we have a large sample we can dissect the range of X into narrow bins that contain many observations, obtaining fairly precise estimates of the conditional mean of Y within them
 - The smaller the sample size, the larger the bin sizes need to be, and thus the fewer the bins. As long as the relationship between X and Y is not too complicated this is fine; If there is complex nonlinearity it can become problematic

19 / 43

Locally Weighted Scatterplot Smoothing (Lowess)

- Lowess is a form of nonparametric regression that fits a separate weighted least squares regression line to each x_i value and then joins the fitted values together
- We choose a span for the proportion of the data to be included in each local regression that provides a smooth fit
- Especially useful when comparing to a linear regression fit
 - The blue line is $s=.2$; the red line is $s=.7$; the black line is the linear fit



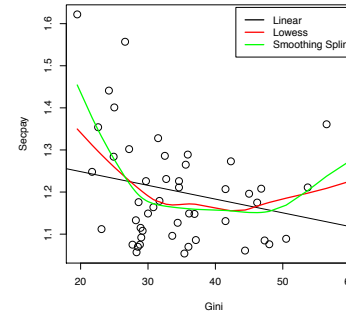
20 / 43

Smoothing Splines

- Smoothing splines offer a compromise between global polynomial regression and local polynomial regression
 - Different piecewise polynomial trends that are constrained to joined smoothly at the knots
 - Not as smooth as global polynomial regression, but generally behave much better at the peaks
- Rather than choose a span as for lowess curves, we usually choose the degrees of freedom – low degrees of freedom will fit a smooth curve; high degrees of freedom will give a rough curve
- Smoothing splines play an important role in Generalized Additive Models

21 / 43

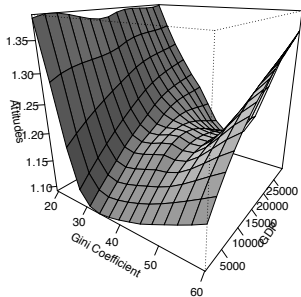
Linear and Nonparametric fits



- The black line is the linear fit, the red line is the lowess smooth from a local linear regression with a span of 0.6
- A clear departure from linearity in these data.
- An F-test comparing the RSS from the linear model with that from the more general trend of the lowess model allows us to assess whether the relationship is linear.

22 / 43

Multiple Nonparametric Regression



- The regression surface is clearly nonlinear for gini
- As with the simple model, we could test for nonlinearity using an F-test comparing the RSS of this model with the RSS of the linear model
- If we had only one more predictor, the lowess model would be impossible to interpret - we can't see in more than 3 dimension.

23 / 43

Generalized Additive Models

- Additive Regression Models overcome the curse of dimensionality by applying local regression to low dimensional projections of the data
- The nonparametric additive regression model is:

$$Y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \dots + f_k(x_{ik}) + \varepsilon_i$$

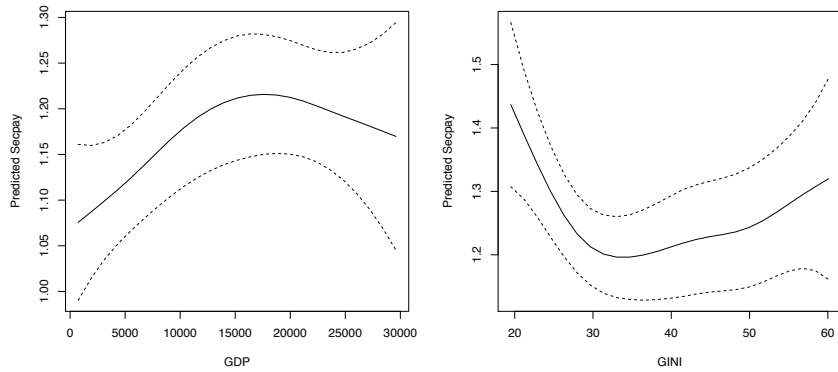
instead of

$$Y_i = \alpha + f_1(x_{i1}, x_{i2}, \dots, x_{ik}) + \varepsilon_i$$

- Additive models create an estimate of the regression surface by a combination of a collection of one-dimensional functions
 - In effect, then, they restrict the nonparametric model by excluding interactions between the predictors
 - An estimation procedure called "backfitting" is used to fit the models.

24 / 43

Generalized Additive Model (2): $secpay \sim s(gdp) + s(gini)$



Introduction

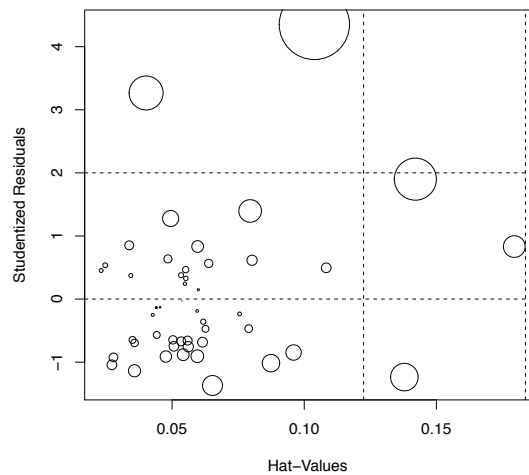
OLS Basics: Theory and Presentation of Results

Non-Linearity

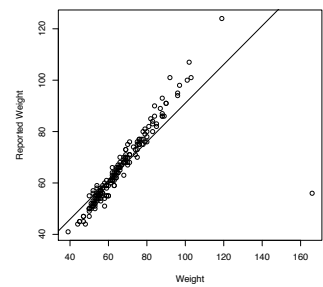
Outliers and Influential Data

Other Topics

Influential Cases: Influence Plots

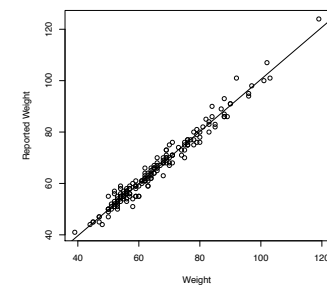


OLS and Influential Cases



(Intercept)	15.76*
	(2.50)
weight	0.75*
	(0.04)
N	183
R ²	0.70
adj. R ²	0.70
Resid. sd	7.58

Standard errors in parentheses
* indicates significance at $p < 0.05$



(Intercept)	-0.93
	(0.86)
weight	1.01*
	(0.01)
N	182
R ²	0.97
adj. R ²	0.97
Resid. sd	2.32

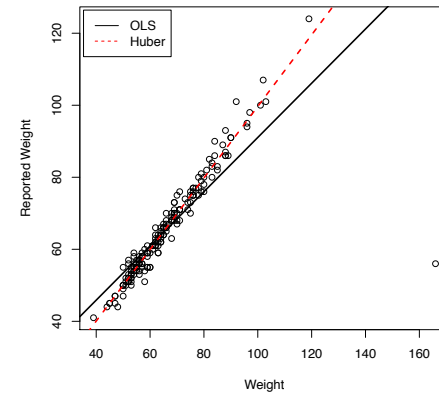
Standard errors in parentheses
* indicates significance at $p < 0.05$

Robust Regression (1)

- Robust regression (e.g., MM-Estimation with Huber weights) can give a significantly better fit to data with influential cases than does OLS.
 - Both slope coefficients and statistical significance can change drastically
 - Residual standard error can also be decreased significantly
 - Not the same as commonly used robust standard errors which are used to compensate for a unknown pattern of heteroskedasticity
- Robust regression assumes large sample sizes, however, so if we have a small one, we may wish to use bootstrapping in order to have greater confidence in our statistical inferences
- Robust regression can also be extended to generalized linear models

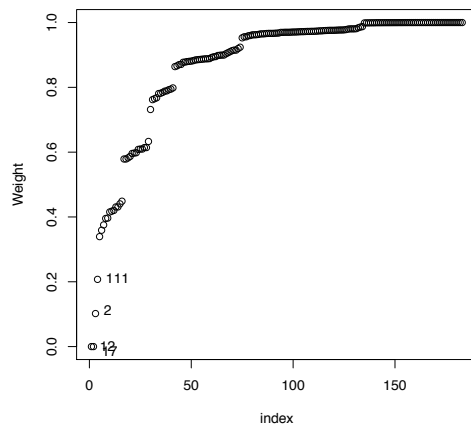
Robust Regression (2)

Figure: OLS and Huber Robust fits to Davis Data

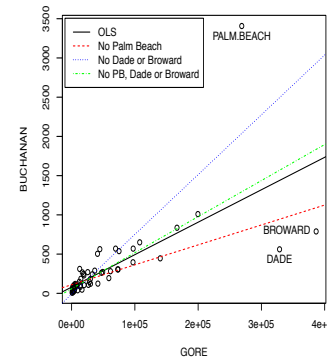


Robust Regression (3)

- We can even see how much weight was given to each case in the analysis:



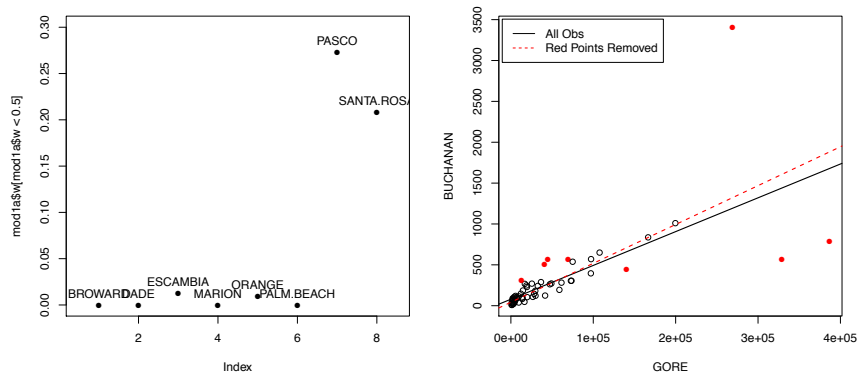
OLS and Joint Influence: Votes by County 2000 US Presidential Election



- If all three outliers are included, the regression line is similar as to when they are all deleted
- Jointly BROWARD and DADE have nearly equal influence to the influence of PALM BEACH
- If either PALM BEACH or BROWARD/DADE are deleted separately, however, the OLS fit changes dramatically

Robust Regression and Joint Influence

- Robust Regression can also potentially identify jointly influential points.



33 / 43

Introduction

OLS Basics: Theory and Presentation of Results

Non-Linearity

Outliers and Influential Data

Other Topics

34 / 43

Bootstrapping

Analytical sampling distribution of Robust Regression estimates are *not* valid for small or markedly non-normal samples, thus in those situations, we need something different.

- Bootstrap resampling provides a way of approximating a sampling distribution when a closed-form solution is unknown or unreliable.
- We all know that the sampling distribution of a mean is:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

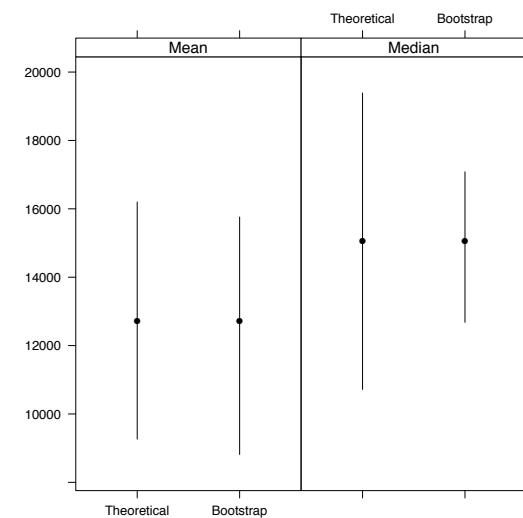
- For normal data with large-N, the sampling distribution:

$$\tilde{x} \sim N\left(\mu_{\frac{1}{2}}(x), 1.253 \frac{\sigma}{\sqrt{n}}\right)$$

35 / 43

Bootstrapping (2)

Consider finding the difference of medians or means across two groups.



36 / 43

Heterogeneity Through Mixtures

Sometimes we don't think that all observations are predicted by the same underlying equation.

- Can deal with this through interactions (in some circumstances).
- Can deal with this through multilevel models (in some other circumstances).
- Both of the ways above will use all X variables to predict all observations

Finite Mixture models allow us to estimate

1. Different models for different groups of observations, *and*
2. The probability of being a member of each of the k different groups

37 / 43

Finite Mixture of Duncan Data

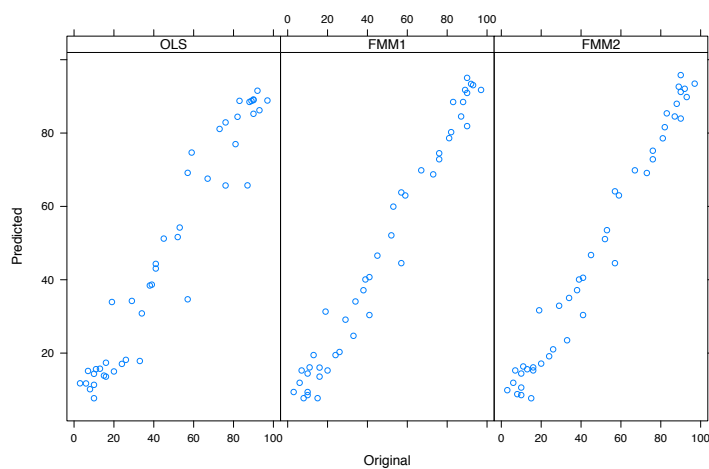
We could estimate the regression of prestige on income, education and type.

- However, assume prestige for some occupations (given type) is a function of income and for others, it is a function of education.
- Using both education and income to predict all observations will give us the wrong idea.
- However, we don't know which occupations are

	OLS	FMM: C1	FMM: C2
(Intercept)	17.89	-25.92	1.83
typeprof	-85.48	30.76	27.44
typewc	-86.68	60.44	-13.29
income	-0.29		0.84
education	-0.56	2.20	
typeprof:income	2.49		-0.02
typewc:income	1.59		0.11
typeprof:education	2.03	-1.25	
typewc:education	1.98	-2.12	
income:education	0.04		
typeprof:income:education	-0.06		
typewc:income:education	-0.05		

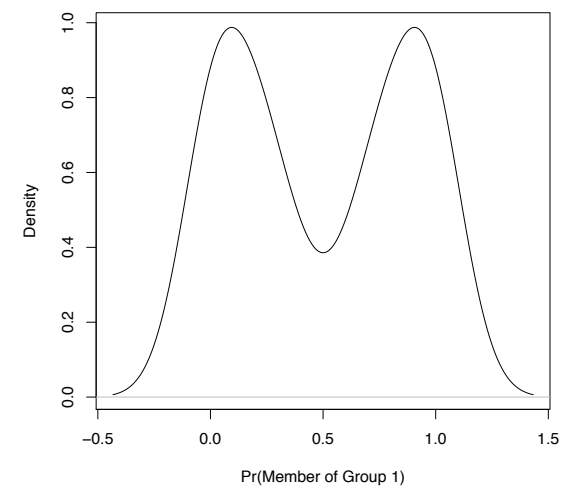
38 / 43

Predictions vs. Prestige



39 / 43

Probabilities of Group Membership



40 / 43

Missing Data and Multiple Imputation

If data are not missing completely at random (MCAR), then estimates we produce with those data may be biased. Here, we will discuss:

- What are the various types of missing data we can have?
- What tools do we have to bring missing observations back into our analysis?
- How can we account for our uncertainty about the imputed (i.e., made up) data?

41 / 43

Readings for Today

- * Fox (2008), Chapters 1 & 2
- * Fox (2002), Chapters 1 & 2
- Venables and Ripley (2002), Chapters 1-3

42 / 43

Tomorrow's class

- Everything you ever wanted to know about OLS

Readings:

- * Fox (2008) Chapters 5, 6, & 9
- * Fox (2002) Chapter 4

43 / 43