Multiple Regression Using OLS

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

March 24, 2022

Motivation

- Previously, we built the simple, bivariate linear regression model to help us estimate the effect some explanatory variable, x, had one some dependent variable, y.
- But our assumption that x is uncorrelated with $\hat{\epsilon}_i$ is probably unrealistic.
- We'd like to say that x affects y, *ceteris paribus*. So how do we get there using regression?
- Simple. We "control" for other factors simultaneously via "multiple regression" as we attempt to isolate x's affect on y.
- In this way, multiple regression, at long last, allows us to mimic experimental research designs.

Introduction to multiple linear regression

- Consider a model with two explanatory variables, "education" and "experience," with a dependent variable of "wage":

$$\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \epsilon_i.$$

- Had we omitted experience as a right-hand-side covariate, its effect would be in the error term.
- Thus, removing it from the error term allows us to estimate the effect of education on wage while *holding experience constant*.
- Similarly, we needn't (and shouldn't) be limited to just two explanatory variables.

Generalized multiple linear regression

- Multiple regression analysis allows us to examine the effect of x on y while controlling for potentially many other variables.
- To generalize to the k^{th} variable, we get the following:

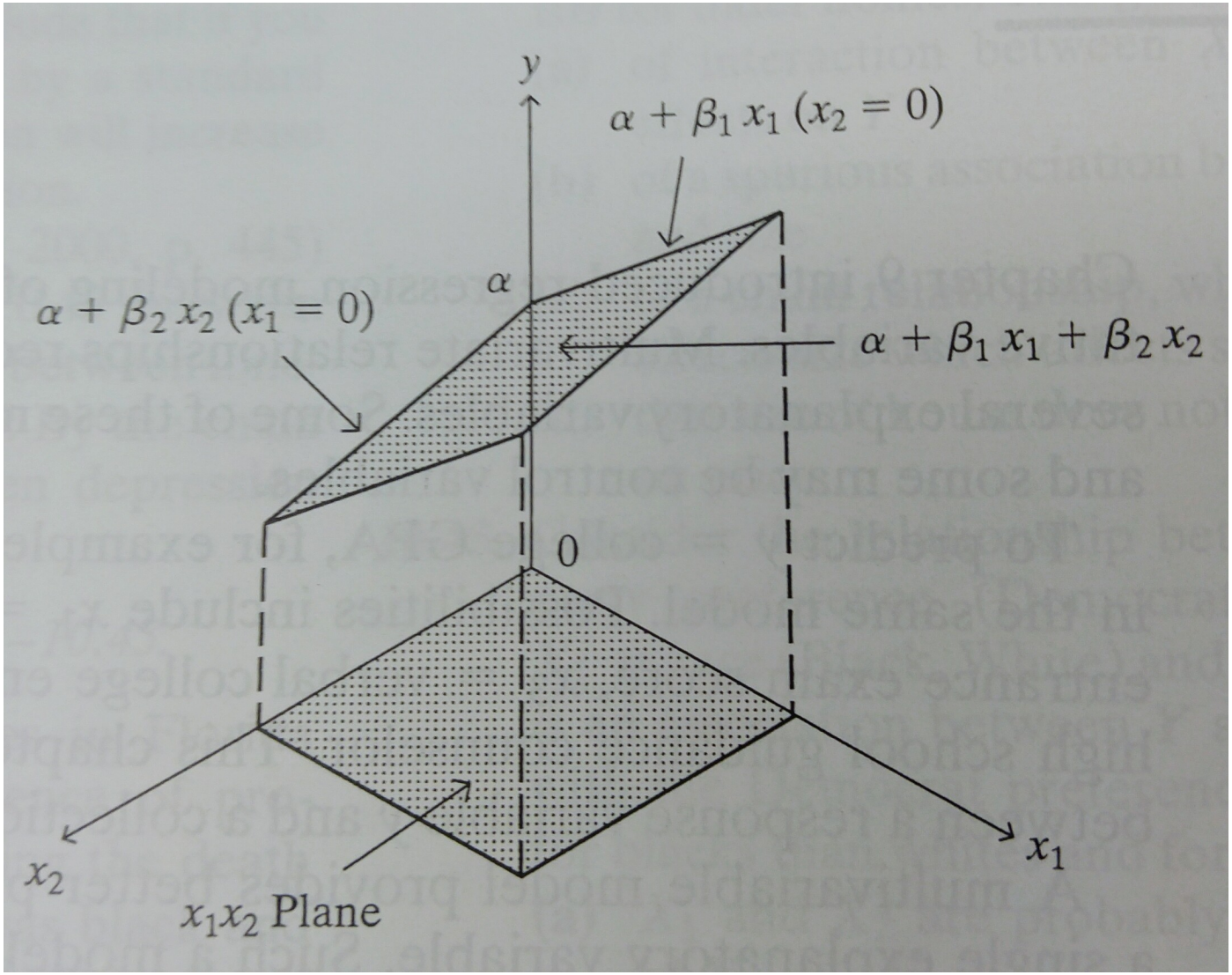
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki} + \hat{\epsilon}_i, \quad (1)$$

- Each x_k represents some independent variable, $\hat{\beta}_0$ is the intercept, the other $\hat{\beta}_k$s now represent the "partial slope coefficients" for each variable, and $\hat{\epsilon}_i$ represents the unaccounted variance (or error).
- As before, we continue to assume that what is contained in $\hat{\epsilon}$ is uncorrelated with the independent variables.

Estimating the multiple regression coefficients

- Previously, we used calculus and the scalar form of the linear model to derive via closed-form solution the values of $\hat{\beta}_1$ and $\hat{\beta}_0$ that minimize the sum of squared errors.
- We want to use the same logic in multiple regression, but the equations we derived for the bivariate model are no longer appropriate since they don't account for the linear effect of other independent variables.
- We could continue in the scalar form, but the addition of more variables makes the work rather tedious. We could proceed in the matrix form, but the math is beyond the scope of this class.
- Thus, from here on out, we rely on the machine to estimate our $\hat{\beta}$s.

A 2-d multiple regression plane



How to interpret multiple regression OLS output

- We refer to each $\hat{\beta}_k$ as the "partial slope coefficient."
- For each independent variable, we say, "A one-unit change in X_k has a predicted $\hat{\beta}_k$ effect on Y, *ceteris paribus*."
- By holding other variables constant, we can isolate the effect of a given X on Y.
- But remember, any leftover or unaccounted for variance is going into $\hat{\epsilon}$.

Why Multiple Regression?

- Wouldn't it be easier to estimate bivariate regressions piecemeal?
- Well, yes. But that doesn't mean we should. We use multiple regression to combat endogeneity in the error term.
- Remember, what explains the error? It's anything not included as an independent variable. By including more IVs, we are sucking error variance out of $\hat{\epsilon}_i$.

Assumption 1: Linearity

- Just as with the bivariate OLS model, we require linearity in our parameter.
- Again, we can allow non-linearity in our variables, just not in our partial slope coefficients.

Assumption 2: Random sampling

- As with the bivariate model, we assume that our data derive from a random sampling method.
- This doesn't guarantee us a representative sample, but it maximizes the likelihood we got one.
- As we mentioned above, though, absent time-series features in the data, we can assume random sampling for most cross-sectional research designs.

Assumption 3: No perfect collinearity

- This one's a little different from the bivariate context.
- If some variable, x is *perfectly collinear* with some other variable, y, then OLS can only estimate parameters for one of these variables.
- We're unlikely to run across many truly perfectly collinear relationships in the wild.
- Odds are when you see it is due to dummy variables. With categorical controls, we need to omit a "reference category."

Assumption 4: Zero conditional mean

- As with the bivariate model, we assume that the expected value of the error term is equal to zero, conditional upon the covariates, $E(\epsilon \mid \boldsymbol{X}) = 0$.
- The way violations of this assumption usually emerge is via omitted variable bias.
- If the data are correlated with the error, the simplest fix is to identify what variables are missing in the model and to put them there. Then they're out of the error, and your independent variables can't covary with it.

Assumption 5: Homoskedasticity

- As in the case of the bivariate model, we assume that the error term has the same variance given any value in the independent variables.
- Importantly, if assumption 5 fails and we have heteroskedasticity, we will not have efficient standard errors for the slope coefficients.

The Gauss-Markov Theorem

- Under assumptions 1 through 5, OLS produces $\hat{\boldsymbol{\beta}}$, are the best, linear, unbiased, estimates of $\boldsymbol{\beta}$. (BLUE)
- By "best," we refer to minimal variance—i.e., efficiency (follows from Assumptions 3 and 5).
- By "linear" we refer to the parameters (follows from Assumption 1).
- By "unbiased," we refer to the $\hat{\beta}$ estimates themselves (follows from Assumptions 2 and 4).

Inclusion of irrelevant variables

- We've touched on omitted variables, but what if we include some variable—call it a—that *doesn't* belong in the regression model? Will this bias our results?
- Suppose we have variables x_1 and x_2, which are theoretically linked to outcomes in y. Then we have:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 a_i + \hat{\epsilon}_i.$$

- If a is truly irrelevant to outcomes in y, then its partial slope coefficient is $\hat{\beta}_3 = 0$.
- Obviously, then, the effect of a will be canceled out, and we'll just be left with the relevant independent variables, the slope coefficients that matter, and the error term.

Omitted variable bias

- But now it's time to show that omitting a relevant variable is not without penalty to $\hat{\boldsymbol{\beta}}$.
- Suppose we specify an *under-specified* regression model: $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_{1i} + \tilde{\epsilon}_i$.
- And now suppose we should have included another variable, x_2. In the correct model, we would observe: $y = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\epsilon}_i$.
- Then we see that $\tilde{u}_i = \hat{\beta}_2 x_{i2} + \hat{\epsilon}_i.$, and so long as $\hat{\beta}_2 \neq 0$, $\sum \tilde{\epsilon}_i \neq \sum \hat{\epsilon}_i$, which ensures that $\tilde{\beta}_1 \neq \hat{\beta}_1$.
- Hence, omitted variables can bias coefficients for the variables that are included in an OLS model.

The normality assumption

- Building on the Gauss-Markov Theorem, we want to be able to hypothesis-test using our OLS results, so we need one additional assumption.
- We assume that the population error is normally distributed with mean zero.
- The addition of the normality assumption to the Gauss-Markov assumptions is sometimes termed the classical linear regression model (CLRM).

Hypothesis-testing in OLS

- The normality assumption allows us to assume that $\hat{\beta}$s are drawn from the t-distribution for the purposes of hypothesis-testing.
- Our null hypothesis is that an IV has no effect on the DV. That is, $\hat{\beta}_j = 0$. Our alternative hypothesis will be the value of $\hat{\beta}_j$ we derive from OLS.

Using a t-test with OLS

- To calculate t for a given coefficient, we simply measure the number of standard errors some $\hat{\beta}_k$ falls from its null hypothesis:

$$t = \frac{\hat{\beta}_{H_a} - \hat{\beta}_{H_0}}{\hat{\sigma}_{\hat{\beta}}}.$$

- Because the null hypothesis holds that $\hat{\beta} = 0$, we simplify the equation for t down to:

$$t = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}}}.$$

Confidence intervals for $\hat{\beta}$s

- As we did in other areas before, we can also express our uncertainty about $\hat{\beta}_k$ via confidence intervals.
- The calculation and interpretation are relatively identical as before:

$$\hat{\beta}_k \pm t^*(\hat{\sigma}_{\hat{\beta}}),$$

where t^* indicates the relevant critical threshold for rejecting the null hypothesis, depending upon degrees of freedom and whether one is conducting a one- or two-tailed test.

Overall significance of the regression

- Finally, we can report on the statistical significance of our OLS model as a whole.
- To do so, we turn to the F formula:

$$F = \frac{R^2/k}{(1-R^2)(n-k-1)},$$

where n indicates the number of observations in the regression, R^2 indicates overall model fit, and k indicates the number of $\hat{\beta}$s estimated.

- The null hypothesis here is that none of the IVs have any effect on the DV.

Why presentation of results matters

- If you've made it this far in the course and your reaction to OLS estimates is, "This makes complete and total sense to me," then congratulations.
- If you're like most of the world, however, and you don't just "get" statistical regression, that's why presentation of results matters.
- At the end of the day, after all your hard work, you want people to know just what the heck you're talking about.

Tabular presentation

- To be certain, tables remain the most popular way by which academics discuss the results of their statistical regressions.
- On the one hand, OLS beta coefficients are readily interpretable, but we're assuming a fair degree of competency on the part of our audience that may not be altogether realistic.
- That said, we can present tabular information showing the marginal effect X has on Y. Consider the following example.

Example of tabular results

Table: Democratic vote-shares in AL counties (2018)

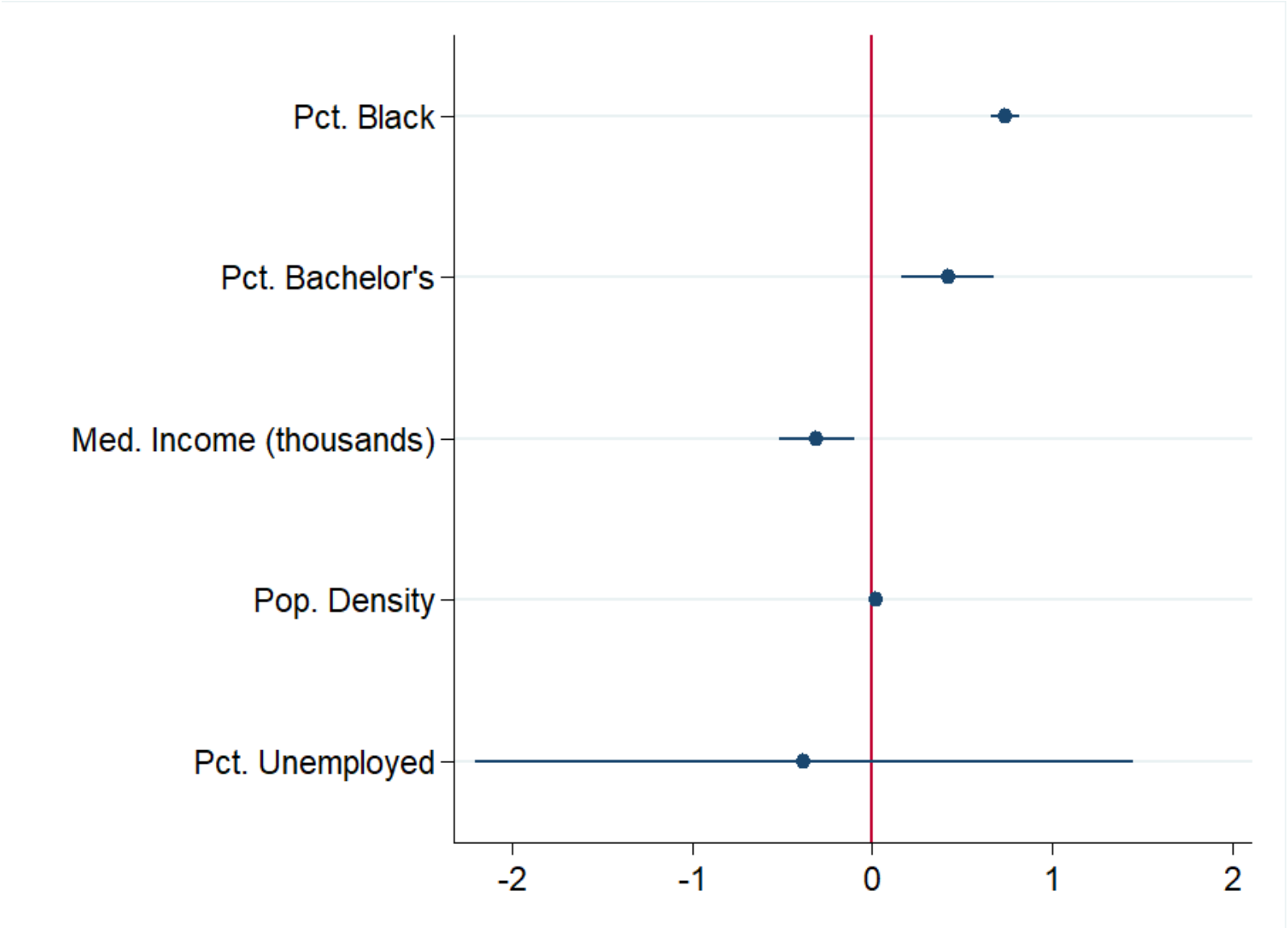
Variable	$\hat{\beta}_k$	$\hat{\sigma}_{\hat{\beta}_k}$	p
Pct. Black	0.74*	0.05	0.00
Pct. Bachelor's	0.42*	0.15	0.00
Med. Income (thousands)	-0.31*	0.13	0.01
Pop. Density	0.02*	0.01	0.02
Pct. Unemployed	-0.38	1.10	0.37
Intercept	21.37*	6.50	0.00
$N = 67$, $R^2 = 0.94$, $F = 203.46$			

Notes: The dependent variable is the percent of the county voting Democratic. Asterisks indicate statistical significance ($\alpha = 0.05$, one-tailed).

Graphical presentations of regression output

- Increasingly, particularly given advancements in computing software, scholars are turning to graphical means of presenting their regression results.
- Graphical presentations of regression estimates usually fall into one of two camps: Plots of all variables, their coefficients and confidence intervals and plots of individual variables and their effect on the dependent variable over its given values.

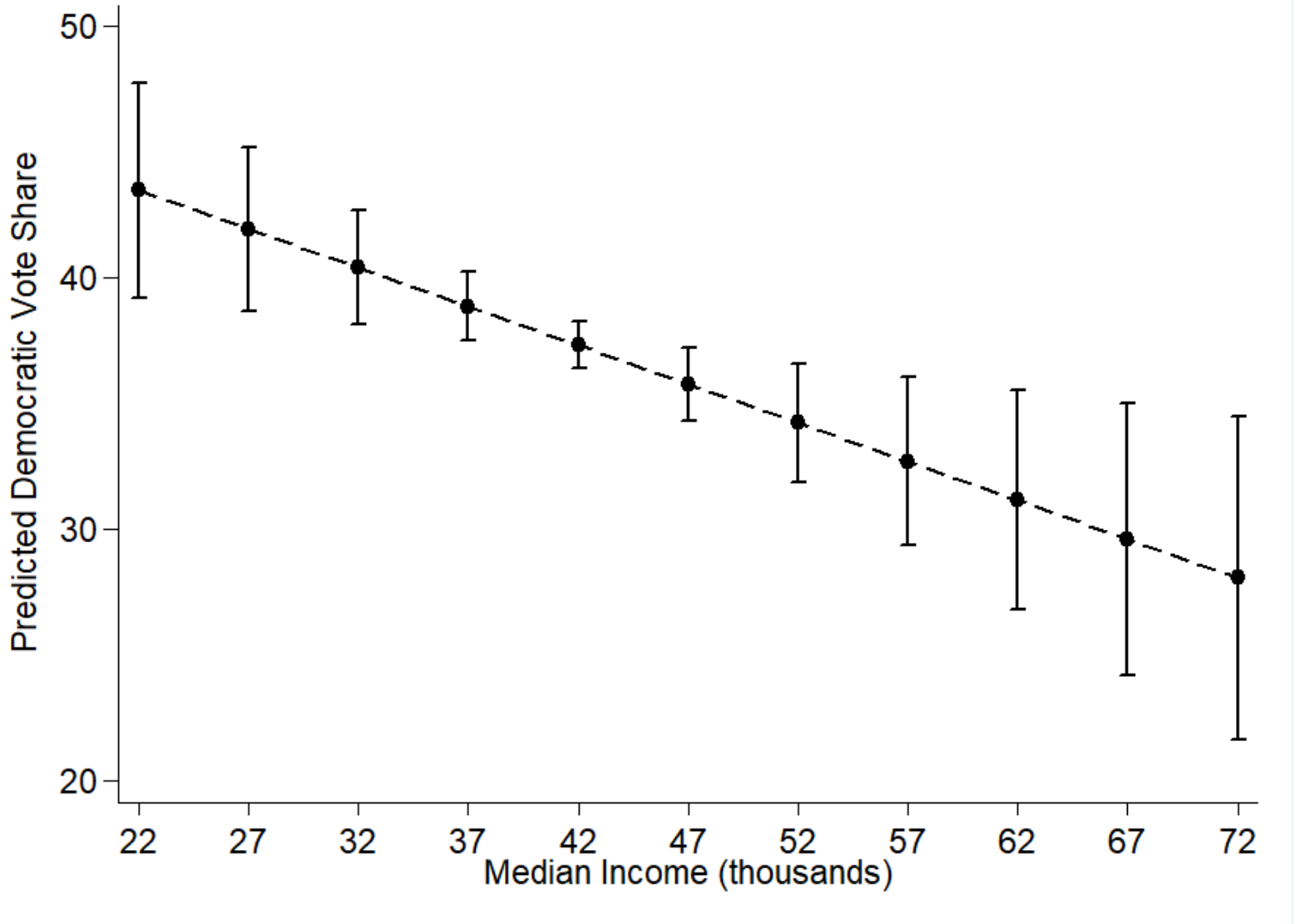
Rope and ladder plot



Assessing the previous graph

- The above graph looks a lot tidier than the table. If the confidence intervals don't overlap with 0 (red line), we have statistical significance.
- These type of figures make it easy to see the uncertainty in the estimates.
- One downside, though, is that with a big scale on the x-axis as we have here, it's pretty tough to get a bead on the values of the coefficients.

Plots on individual variables and their values



Assessing the previous graph

- I really can't over-emphasize how useful a graph like the one above is in helping to interpret regression estimates.
- These sorts of visualizations are so easily interpretable—not only for people who know what OLS is but also for the average layperson.
- My personal approach to presenting regression output is to display everything tabularly (possibly in an appendix).
- Then I find the most interesting results and plot them graphically for the reader (best of both worlds).

Discussion

- When our dependent variable is measured continuously, OLS is, given the Gauss-Markov assumptions are satisfied, the most appropriate statistical regression technique.
- Multiple regression allows us to estimate the effect of some x on y, even while controlling for other variables, thereby allowing us to mimic experimental research methods using observational data.
- As we proceed, we'll learn how to make inferences using OLS, how to test OLS assumptions, and how to revise our regression tehcniques when CLRM assumptions are not met.