Regression Analysis

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

November 18, 2019

Introduction

- Lately, we've been examining bivariate relationships.
- But remember our three criteria for causality.
- How do we account for the influence of other, "lurking" variables?

Linear relationships

- We have an IV and a DV.
- H_a: IV has a positive/negative effect on DV.
- H_0: IV has *no* effect on DV.

Reviewing lines

- Recall what a line is:

$$Y_i = \beta_0 + \beta_1 X_i,$$

where Y_i is the DV, X_i is the IV, β_0 is the intercept, and β_1 is the slope coefficient.

Interpreting lines: $Y_i = \beta_0 + \beta_1 X_i$

- We interpret lines as follows: "For every one-unit increase in X, there is a corresponding β_1 change in Y.
- Let the sign on β_1 denote the directional relationship between IV and DV.
- Let the magnitude of β_1 denote the strength of this relationship.

Drawing and interpreting lines

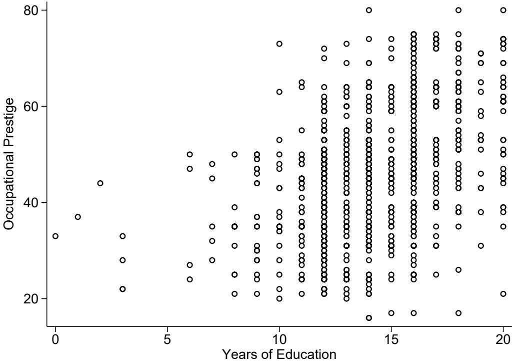
- Draw these lines:
  1. $Y_i = 2 + 3x_i$
  2. $Y_i = -1 - 2x_i$
  3. $Y_i = 5 + .5x_i$
- Draw and interpret the following:
  1. You hypothesize that increased flu vaccinations decrease the number of flu cases. You collect the number of flu shots administered and cases of flu here in Montgomery County and find: $\text{Cases}_i = 1050 - .25\text{Shots}_i$.

Lines and the scatterplot

- Imagine a scatterplot of data.
- We'd like to fit a line onto these data.
- This line would represent the direction and strength of association of our IV on the DV.

Race and partisanship in Alabama politics

- Clearly, there's a positive relationship.
- But how do we think about the *strength* of that relationship?
- Put differently, how do we fit a line across those data?



The regression line

- A linear regression is a technique by which you fit a line onto your scatterplot of data.
- We summarize the relationship between X and Y using the following:

$$Y_i \quad = \quad \beta_0 + \beta_1 x_i + u_i,$$

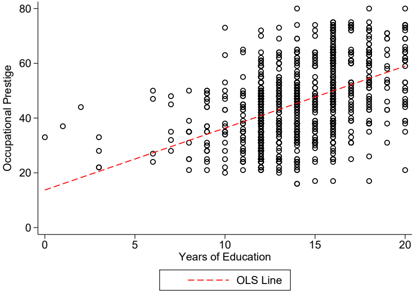
where now we have added i u_i, which denotes the "error," or how far a given observation was from the line itself.

Ordinary least squares (OLS)

- Ordinary least squares is the method by which we fit a line onto our scatterplot of data.
- This is termed the "line of best fits" because it minimizes the distance of data-points to a feasible regression line.

Interpreting OLS results

- Using OLS, we find that the effect of education on occupational prestige is: $Prestige_i = 13.7 + 2.3\text{Education}_i + u_i$.
- Interpret these results.



Hypothesis-testing with lines

- H_a will be a directional relationship.
- Therefore, H_a: $\beta_1 \lesseqgtr 0$.
- And H_0: $\beta_1 = 0$.

Uncertainty in OLS estimation

- As with our difference-of-means tests, we denote our uncertainty using standard errors.
- Every beta coefficient gets a standard error, and this is how we hypothesis-test.
- Therefore, we hypothesis test in OLS using a z-test:

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

Hypothesis-testing in OLS

- Assume that we use an α-level of 0.05, one-tailed.
- We find that $\hat{\sigma}_{\hat{\beta}_1} = 0.13$.
- What is z, and what is the critical threshold necessary to reject the null?
- Is education a statistically significant factor in one's occupational prestige?

Goodness of fit in OLS

- How good of a job does our IV do in explaining the variance in the DV?
- In OLS, we use R^2 to measure "goodness of fit."
- R^2 is on a scale of 0 to 1, where higher values denote stronger fit—more specifically, it is a measure of the proportion of variance in the dependent variable explained by the independent variable(s).
- In our running example, $R^2 = 0.22$.

Multiple Regression

- It is now time to address those lurking variables we discussed earlier.
- Multiple regression analysis allows us to account for the effect of some X on Y, *while simultaneously controlling for* the extraneous effect of some lurking variable, Z.

Multiple regression with OLS

- Theoretically, we could generalize our model such that:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i,$$

where k is the number of independent variables.

- We can estimate $n-1$ partial slope coefficients.

Interpreting multiple regression OLS output

- We refer to β_k as the *partial slope coefficient.*
- Thus, the effect of some X_k on Y is β_k, *holding all other variables constant*.
- That last bit's really important.

Practice Interpreting OLS Multiple Regression Results

Variable	Coefficient	Standard Error	z-value
Education	2.18*	0.13	17.07
Age	0.13*	0.03	5.58
Wealth Redistribution	-1.01*	0.45	-2.23
Female	0.11	0.72	0.15
Nonwhite	-2.22*	0.83	-2.68
Intercept	11.62*	2.42	4.81

Notes: $N = 1109$. $R^2 = 0.25$. Asterisks denote $p < 0.05$.

Multiple regression with categorical DVs

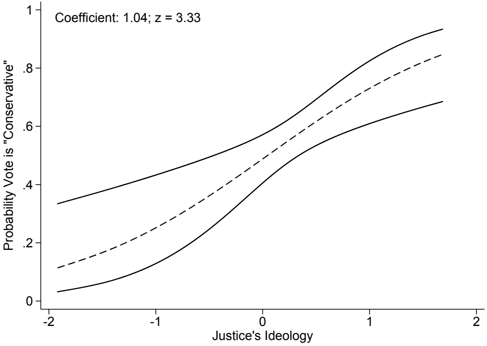
- When our DV is categorical, we could theoretically still use OLS in the multiple regression.
- This may be difficult to justify, however.
- When the DV is dichotomous, we use logit or probit (really the same thing).

Logit and probit models

- We're still interested in the marginal effect some IV has on a DV.
- Unlike with OLS, however, we allow this effect to be curvilinear.
- This means that we are unable directly to interpret our β_k coefficients from the logit/probit.
- We are still able to assess, however, "signs and significance."

Example of a Logistic Curve

- Suppose we're studying the likelihood that a supreme court justice casts a "conservative" vote in an abortion case ("1" if yes, "0" otherwise).
- Our DV is "conservative," and our IV of interest is ideology, measured liberal-to-conservative.



Practice Interpreting Probit/Logit Regression Results

Variable	Coefficient	Standard Error	z-value
Education	0.02	0.02	0.65
Age	-0.01*	0.005	-2.37
Ideology	-0.01	0.08	-0.15
Religiosity	-0.07	0.06	-1.25
Female	0.28*	0.13	2.14
Nonwhite	0.19	0.14	1.33
Intercept	-1.42	0.43	-3.27

Notes: DV="Gay/Bisexual." $N = 1109$. Asterisks denote $p < 0.05$.

Other types of statistical regressions

Level of Measurement

- Ordinal
- Nominal
- Event count
- Duration

Statistical Estimator

- Ordered logit/probit
- Categorical logit/probit
- Poisson/negative binomial
- Hazard/duration model