Specifying the Regression Model

Ciaran S. Phibbs

March 22, 2023

Background

- All too commonly an investigator focuses on using the correct regression model and doesn't adequately address the issues about how to specify the variables in the model, and if those variables are properly specified and meet the model assumptions.

Independent Variables

- Regression models make several assumptions about the independent variables
- The purpose of this talk is to examine some of the more common problems, and some methods of fixing them
- Focus on things that not be covered in standard MPH-level regression classes

Outline

- Heteroskedasticity
- Clustering of observations
- Data Aggregation
- Functional Form
- Testing for multicollinearity

Heteroskedasticity

$$Y_i = \beta_0 + \beta X + \varepsilon_i$$

- Assumes that the error terms are independent of x_i. Common pattern, as x gets bigger, e gets bigger.

Heteroskedasticity

Why does this matter?

- Biased standard errors
- Parameter estimates unbiased, but inefficient

Heteroskedasticity

- Simple solution, “robust” option in Stata uses Huber-White method to correct standard errors.
- May also consider transformation of variables, *e.g.*, log(X) instead of X as a RHS variable.

Clustering

$$Y_i = \beta_0 + \beta X + \varepsilon_i$$

- Assumes that the error terms are uncorrelated
- Clustering is a common problem in healthcare, for example, patients are clustered within hospitals

Clustering

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$$

- x_1 is a patient level variable, and x_2 is a hospital level variable
- Regression assumes there are as many hospitals as patients
- Result, the standard errors for β_2 are too small, no effect on parameter estimate.

Correcting for Clustering

- Generalized Estimating Equations (GEE) or other hierarchical methods can be used
- Alternatively, Stata “cluster” option uses a Huber-White correction of the standard errors.
- Both methods can yield essentially the same result, it depends on the data structure

Correcting for Clustering

- Hierarchical Linear Modeling. Method of formally incorporating hierarchical structures into the model.
  - Can use for non-linear models also.

- Need for HLM, vs. other methods will depend on structure of the data. Often very similar answers.

Example of Clustering

- I had a research project that looked at the effects of NICU patient volume and NICU level on mortality. NEJM 2007.
- I apologize for not using a VA example, but good example where I had the data.

Clustering

- Failure to make this correction happens all too often. It is easy to fix
- Extent of the correction varies with sample size, and with the number of clusters, relative to the number of observations.
- With big samples, the effects are fairly small. My example, N>48,000, >200 hospitals, 10 years of data, with repeat observations.

Example of Clustering

Level of Care/VLBW volume	OR	95% C.I.	unadjusted
Level 1 ≤10 VLBW infants	2.72**	(2.37, 3.13)	2.40, 3.07
Level 2 11-25 VLBW infants	1.88**	(1.56, 2.26)	1.64, 2.15
Level 2 >25 VLBW infants	1.22	(0.98, 1.52)	1.09, 1.36
Level 3B or 3C ≤25 VLBW	1.51**	(1.17, 1.95)	1.25, 1.78
Level 3B or 3C 26-50 VLBW	1.30**	(1.12, 1.50)	1.17, 1.42
Level 3B, 3C, or 3D 51-100	1.19*	(1.04, 1.37)	1.10, 1.29

Data Aggregation

- Many times, have a choice of how to organize data
- Data aggregation can matter:
  - In general, increased aggregation will reduce variance
  - Aggregation can also change the relationship between the variable of interest and the dependent variable

Example of Data Aggregation

- Data from Bartel, Bealieu, Phibbs, Stone., Am Econ J: Applied Econ 2014:6(2):231-259, and Winter et al., HSR 2021;56(6):1262-1270.
- Patient-level regressions, nurse staffing measured at different aggregations
  - Unit vs. hospital
  - Month vs. year

Effect of Data Aggregation, Unit vs. Hospital

	Hospital	Acute Care Units	ICUs
HPPD	-0.011***	-0.031***	-0.016***
% LPN	-0.194*	-0.041	0.215
% UAP	0.138*	0.088**	0.217
% Contract	0.180**	0.31***	0.333***

Why Data Aggregation Mattered in this Example

- ICUs and Acute Care units are very different units on several dimensions. In this case, especially the severity/nursing needs of the patients and the levels of nurse staffing. With much higher staffing levels, the effects smaller in ICUs. Combing them masks bigger effect on Acute Care units.

Why Data Aggregation Mattered in this Example

- Winter paper also showed that time aggregation mattered. Results different for year vs. month. Essentially, if you aggregate over a year the averaging is masking variation.

Why Data Aggregation Mattered in this Example

- In general, more aggregated data yield less precise estimates because the aggregation can mask the variance.
- General rule, try to avoid aggregation to the extent possible
  - Acknowledge aggregation as a limitation if you must aggregate

Functional Form

$$Y_i = \beta_0 + \beta X + \varepsilon_i$$

βX assumes that each variable in X has a linear relationship with Y

- This is not always the case, can result in a mis-specified model

You should check for the functional form for every non-binary variable in your model.

- There are formal tests for model specification, some of which you may have been exposed to in classes. But these tests don't really show you what you are looking at, just that model is mis-specified. Further, tend to be fairly week tests.

Using Dummy Variables to Examine Functional Form

1. Look carefully at the distribution of each variable
2. Create a set of dummy variables for reasonably small intervals, with no excluded category
3. Run model with no intercept

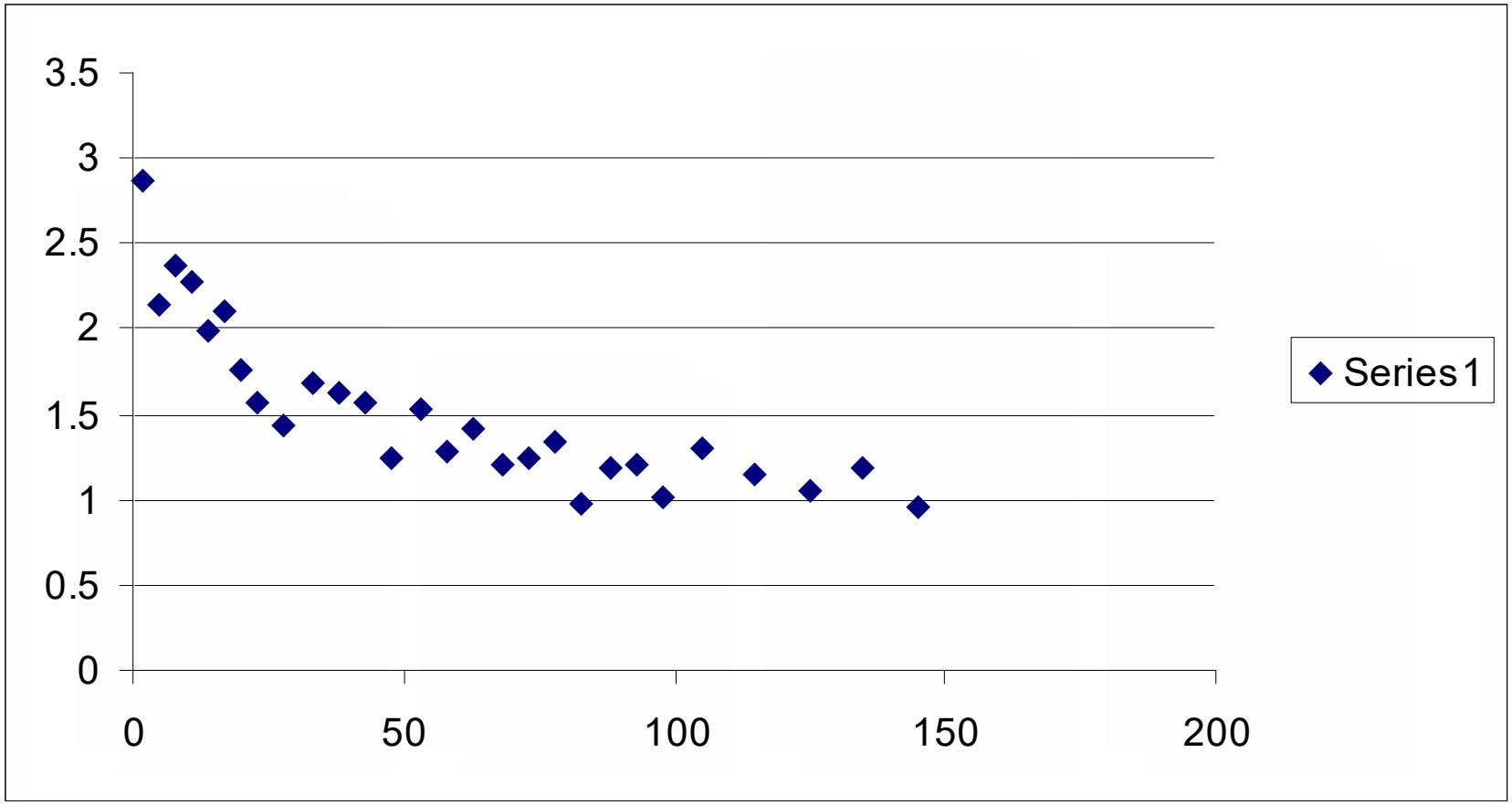
Example of Using Dummy Variables to Examine Functional Form

- Same NICU dataset as before.
- Effect of NICU patient volume on mortality
- Overall, and within level of care

Example of Using Dummy Variables to Examine Functional Form

- Graph out the parameter estimates for each dummy variable.
- Gives you a good idea of what the function looks like.
- Can use to determine which functional forms would be good starting points, or where to make the cuts for categorical variables.

Dummy Variable Look at Functional Form



Example of Using Dummy Variables to Examine Functional Form

- For some applications you may just want to use dummy variables, instead of a continuous functional form. This may be especially useful when there are complex relationships. It can be very difficult to get a continuous function to accurately predict across the entire range of values.
- Aside, categorical variables frequently easier to present to medical audiences.

Dummy Variables to Capture Complex Functional Forms

Level of Care/VLBW volume	OR	95% C.I.
Level 1 ≤10 VLBW infants	2.72**	(2.37, 3.13)
Level 2 11-25 VLBW infants	1.88**	(1.56, 2.26)
Level 2 >25 VLBW infants	1.22	(0.98, 1.52)
Level 3B or 3C ≤25 VLBW	1.51**	(1.17, 1.95)
Level 3B or 3C 26-50 VLBW	1.30**	(1.12, 1.50)
Level 3B, 3C, or 3D 51-100	1.19*	(1.04, 1.37)

Multicollinearity

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$$

- What if x_1 and x_2 are strongly correlated? Regression has trouble attributing effect to each variable.
  - Increases standard errors
  - Can affect parameter estimates, can even get offsetting effects with highly correlated variables

Testing for Multicollinearity

- First step, look at the simple correlations.
  - General rule of thumb, need to look of $r>0.5$
  - Note, can still have collinearity problems with $r<0.5$

Multicollinearity

- Strong simple correlation, you have a problem. But can be hidden problems not detected by simple correlations.
- Variance Inflation Factor ("/VIF" SAS, "vif" in Stata Regression Diagnostics) measures the inflation in the variances of each parameter estimate due to collinearities among the regressors
- Tolerance, which is 1/VIF
- VIF > 10 implies significant collinearity problem

Example of Correlation and VIF

- Study of nurse staffing and patient outcomes referred to above.
  - Problem variables. RN Tenure and RN Age
  - R=0.46
  - VIF range, 18-30, depending on subset

  Result, many fewer statistically significant results than we expected.

Fixing multicollinearity

- More observations. As long as there isn't perfect correlation, additional observations help.
- Revise data in ways that reduce the correlation
- In nurse staffing example, dropped age from model

Example of parameter effects of multicollinearity

- Average RN tenure on unit and average age of RNs on unit, corr = 0.46
  - Tenure only -0.013**
  - Age only -0.003 **

Example of parameter effects of multicollinearity

- Average RN tenure on unit and average age of RNs on unit, corr = 0.46
  - Tenure only -0.013**
  - Age only -0.003 **
  - Both tenure -0.003 ns
    Age -0.0051 ns

Multicollinearity

- Strong simple correlation, you have a problem. But can be hidden problems not detected by simple correlations.
- Regression, n-space, correlation on each of the regression planes can matter.
- Collin option in SAS, looks at how much of the variation in each eigen vector is explained by each variable. Intuitively, the correlation in the Nth dimension of the regression.

SAS Collin option

- SAS Model Y= var1 … varN/collin;
- Continue newborn example
- Birth weight and gestational age very correlated. R=0.56
- Simple model, only BW, GA, Black

Interpreting Collin output

- Condition index >10 indicates a collinearity problem
- Condition index >100 indicates an extreme problem
- There is strong correlation in the variance proportion if 2 or more variables have values >0.50.

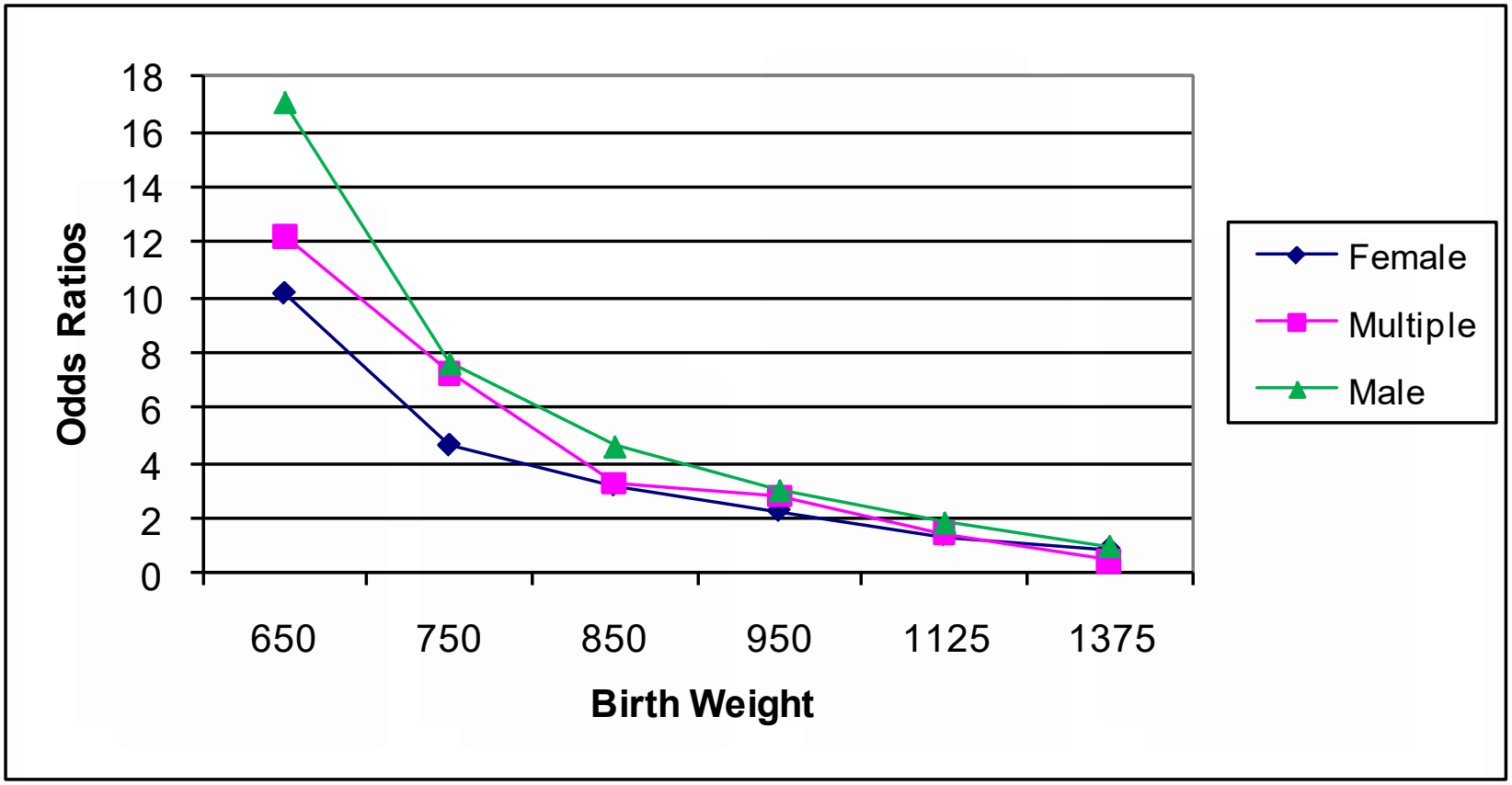
SAS Collin output

Eigen value	Condition index	Black	BW	GA
3.18	1.00	0.02	0.00	0.00
0.77	2.03	0.96	0.00	0.00
0.04	9.27	0.01	0.80	0.02
0.001	18.41	0.00	0.39	0.98

Fixing multicollinearity, NICU example

- Used dummy variables for BW in 100g intervals to 1000g, then 250g intervals.
- Separate BW dummies for singleton males, singleton females, and multiple births,
- Gestation in 1 week intervals.
- Max condition index < 8, i.e., no serious collinearity problem.
- Model predictions also improved.

Dummy Variables To Fix Collinearity



Bottom line recommendation

- Know your data
- Run lots of diagnostics to understand your data before you start running regressions
  - This will alert you to many potential problems so that you can address them

References

- Belsley, D.A., Kuh, E., and Welsch, R.E. (1980) Regression Diagnositics. New York, John Wiley & Sons.

Next lecture

Limited Dependent Variables

Ciaran Phibbs

March 29, 2023