POLS201 Spring 2019

More About Dummy Variables and Jacobson Paper

April 3

Agenda

- Go into Moodle and let me know your preferred meet time
- Live import of a .csv into R (The 1998 Field CA Survey)
- Using dummy variables for categorical and ordinal IV's
- The Jacobson paper: summary and a breakout session
- Some risks of using regressions
    - Not typically relevant for your paper, but note for the final

Quick demo of importing R data

- Let's import 1998 data and look at these two variables
- V128: Respondent's Age and V131: Strength of liberal or conservative belief
- V131:
  - 1 'STRONG CONSERVATIVE'
  - 2 'NOT VERY STRONG CONSERVATIVE'
  - 3 'NOT VERY STRONG LIBERAL'
  - 4 'STRONG LIBERAL'
  - 8 'DON''T KNOW'
  - 9 'NOT APPLICABLE (NOT''CONSERVATIVE''OR''LIBERAL''ON Q103A)'
- Remember: select() chooses variables and filter() chooses observations

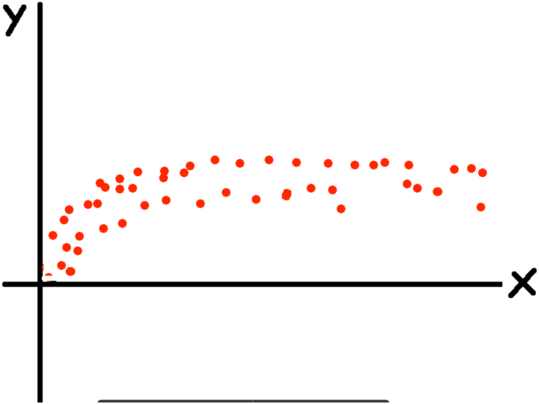
So How Good is your Regression?

- No one cares about predicting the past
  - An overly perfect model using past data is useless
- Worry if your predictions miss, a lot, and not randomly
- Have you thought of all the confounders?
- Linear regression is most appropriate for continuous variables. What if the DV isn't continuous?

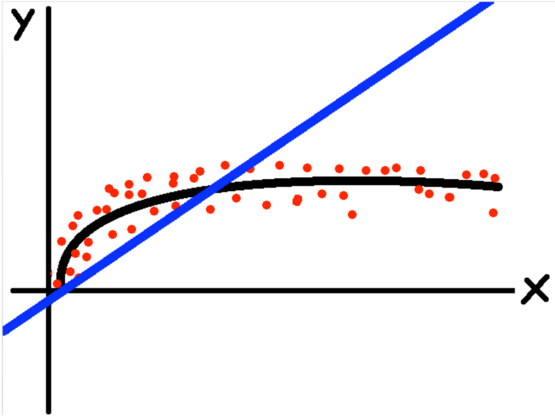
Linearity

- Pitfalls of fitting lines to non-linear relationships
  - Your estimates might be insignificant, even though there is indeed a relationship between your variables
- You fail to adequately control for what you want to control for.

Linearity



Linearity



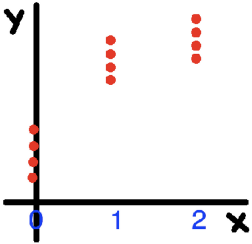
Linearity

- A function is linear if the coefficient is constant
  - Which means: it looks like a straight line
- Solution: Transform your variables.
- If you think a variable has an exponential effect on your DV, you can square it!
- In economics, variables are frequently converted to a logarithm to represent diminishing returns
- For categorical or ordinal variables? Dummy it out!

What is problematic here?

What's problematic?

DV= church attendance in days per year

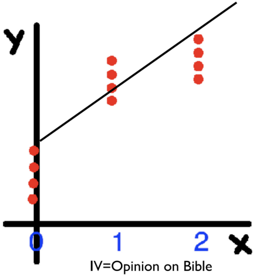


IV=Opinion on Bible

0=word of man **1= inspired** **2=word of god**

The problem? We assume linearity

- Recall our DV is church attendance

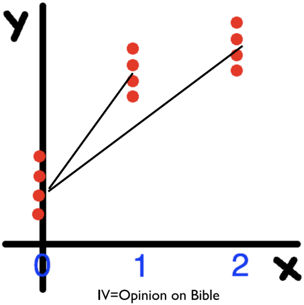


Linear Regression assumes the the marginal effect of moving from 0 to 1 is identical to the marginal effect of moving form 1 to 2. This is not necessarily the case in ordinal data where the distance between numbers is meaningless.

0=word of man **1= inspired** **2=word of god**

The solution? Use dummy variables!

- Sometimes called "indicator" variables



1. Choose a Baseline category. For example make "0" the baseline.
2. Find the effect of being "1" relative to "0"
3. Find the effect of being "2" relative to "0"

0=word of man **1= inspired** **2=word of god**

How to use dummy variables

- We can use dummy variables to capture marginal effects of variables with multiple categories

Attendance	Bible Code	Bible	D1	D2
4	1	Divine Inspired	1	0
3	1	Divine Inspired	1	0
4	2	Word of God	0	1
4	2	Word of God	0	1
4	1	Divine Inspired	1	0
0	0	Word of Man	0	0
2	1	Divine Inspired	1	0
1	0	Word of Man	0	0
4	2	Word of God	0	1
3	2	Word of God	0	1
4	1	Divine Inspired	1	0
1	0	Word of Man	0	0
0	0	Word of Man	0	0

Dummy Variables

- In a regression, both dummy variables MUST be interpreted relative to an **omitted** category.
- If there are no other variables in the regression, the **intercept** can be interpreted as the expected outcome for that omitted group.
- If there are other variables, compare predicted values!

Dummy Variables

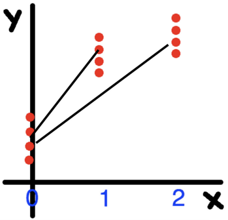
- We create dummy variables (or, if we're good at R, use "factor" variables)
- Each coefficient compares the mean for that group vs. mean of excluded category

```
Call:
lm(formula = Attendance ~ `Bible Code`, data = x1, x = TRUE)

Residuals:
   Min     1Q Median     3Q    Max 
 -1.40  -0.50   0.25   0.50   0.60 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     0.5000     0.3518   1.421 0.185645    
`Bible Code`1   2.9000     0.4720   6.145 0.000109 ***
`Bible Code`2   3.2500     0.4975   6.533 6.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7036 on 10 degrees of freedom
Multiple R-squared:  0.8407,	Adjusted R-squared:  0.8089 
F-statistic: 26.39 on 2 and 10 DF,  p-value: 0.0001025
```

Dummy Variables: Always Exclude One Variable

- You must ALWAYS exclude one category
- Example: If you had a dummy variable for male and a dummy variable for female, they are just mirrors of each other. One must drop!
- The choice about which to drop is arbitrary; customary to choose the most frequent or smallest value

Remember, you are determining the marginal effect relative to the baseline. This means that if your variable of interest takes on 3 values, you can only estimate 2 relative marginal effects



Jacobson Paper Breakout Session

- Theory?
- Hypothesis?
- Unit of Analysis?

TABLE 3

OLS Regression Estimates of the Effects of AFL-CIO Targeting on the Vote for Republican House Incumbents

Independent Variables	Freshmen Republicans	Senior Republicans
Intercept	25.27	49.77***
	(25.06)	(8.17)
Republican incumbent's vote in 1994 (two-party %)	.34*	.37***
	(.13)	(.06)
Bob Dole's district vote in 1996 (two-party %)	.33***	.21***
	(.09)	(.04)
Challenger has held elective public office	−.83	−1.85*
	(1.11)	(.81)
Natural log of spending by and on behalf of challenger	−2.16***	−2.12***
	(.61)	(.26)
Natural log of spending by and on behalf of incumbent	1.86	.16
	(1.64)	(.62)
AFL-CIO target	−4.12**	−.67
	(1.45)	(.93)
AFL-CIO target—video	−4.27**	.22
	(1.62)	(1.94)
Adjusted R^2	.72	.80
Number of cases	69	103

Note: The dependent variable is the percentage of the two-party vote won by the Republican incumbent; candidates are assumed to have spent at least $5,000 (spending below this total need not be reported); standard errors are in parentheses.

*$p < .05$ (two-tailed test)

**$p < .01$ (two-tailed test)

***$p < .001$ (two-tailed test)

Fake Data

Name	Status	Rep. 94 Vote Share	Dole 96 Vote Share	Challenger Quality	Challenger Spending	Incumbent Spending	AFL-CIO target	AFL-CIO Video
Smith	Freshman	53	60	1	$30	$40	0	1
Brown	Senior	64	58	0	$20	$80	0	0
Wilson	Senior	52	45	1	$60	$120	1	0

TABLE 1

The Fates of House Republicans Targeted by AFL-CIO Advertisements

	Freshmen	Nonfreshmen	Total
Not targeted by AFL-CIO	26	123	149
Losers	0	2	2
Percent losers	0.0%	1.6%	1.3%
Mean vote	62.4%	66.3%	65.6%
Target of at least one advertisement	23	17	40
Losers	5	2	7
Percent losers	21.7%	11.8%	17.5%
Mean vote	52.9%	61.6%	56.4%
Target of voter video guide	21	3	24
Losers	7	2	9
Percent losers	33.3%	66.7%	37.5%
Mean vote	50.9%	51.2%	51.0%

Note: The differences across categories of AFL-CIO targeting for both the percentage of losses and the mean share of the vote are significant at $p < .01$ or better in all three columns; the vote is measured as the two-party vote in the district; uncontested incumbents are excluded from this calculation.

Model Specification

- Why include previous vote share? Dole vote share? Challenger quality?
- Other things predict candidate vote - for example racial composition of district. Why not include those?
- Why does Jacobson run his model separately for first term Congress members?
- Why are there two separate variables for targeting?

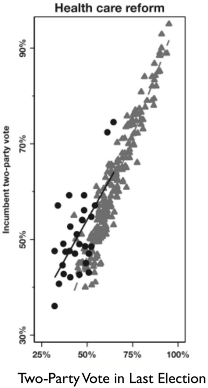
Refresh on Omitted Variable Bias

- An omitted variable is correlated with the IV and partly determines the DV
- It distorts the estimate of the IV coefficient
- But Jacobson thinks he has "fixed" that problem. Do you agree?
- In pairs, list potential omitted variables that might fit this definition.

	Targeted?	Observe DV	Treatment	Observe DV
R	Yes	O	X	O
	No	O	~X	O

The simplicity of Jacobson's model is a virtue

- But beware:



- Lingering issues
  - An "Identification" problem: Is there enough variation on the IVs? It would be best if there were un-targeted candidates with similar district/incumbent profiles.
  - A "linearity assumption" problem: Is accounting for a "linear" effect of our control sufficient?

Basic Problems:

- Regression won't run:
  - If your number of variables is greater than your number of observations
  - If your X variable perfectly predicts your Y variable
- Regression will drop variables if:
  - If your X variable doesn't vary
  - If your X variable is identical or nearly identical to another X variable

The List of Basic Problems (even if the regression does run)

- Too Little Variation
- Outliers
- Too Many or Too Few Observations
- Collinear Variables
- Non-Linear Effects
- Error-Term Issues: Residuals vary in range or systematic

Outliers

- Outliers are data points that take on extreme values (high or low) of either your IV or DV.
- The slope of your line (the marginal effect) can be heavily influenced by outliers.
    - Especially if you don't have a lot of observations

Multicollinearity

- Beware of putting multiple variables that are highly correlated into the same regression.
- Practical effect: Null results or even false results.
  - Increases risk of false negatives
- Don't lard your model with highly correlated IV's

Too many / Too few number of observations

- Too Many Observations (~10,000 plus):
  - Easy to achieve significance
  - But all you have done is explain the data, not the process
  - Machine learning uses large observations much more effectively
- Too Few Observations (~less than 30): -Too hard to achieve significance
- But also remember: you need to work with the data you have, not the data you wish you had :)