Regression: putting it all together

...

20 March 2019
PLSC 309

Review: OLS and Linear Regression

- Linear regression is a statistical model that helps us learn a function that uses explanatory variables to predict an outcome
- The function it learns is the best straight line fit
- The best fit is found by finding the straight line that minimizes the sum of squared errors

Review: assumptions

We make one large, overarching assumption: that our model is additive

We can see if this assumption holds by looking at our residuals. Are they...

- Homoscedastic? (have constant variance or spread)
- Normally distributed?
- Related to any of our explanatory variables?

Review: issues with interpreting our models

- Endogeneity
  - We have incorrectly flipped X and Y
- Omitted variable bias
  - There is some variable Z in our model that affects both our explanatory and outcome variables
- Interactions / multicollinearity
  - There is a combined effect of X_1 and X_2 that we do not account for in our model

Review: types of dependence

- Spatial dependence
  - Occurs when our observations are “clustered”
  - Effects standard errors
  - Tricks us into thinking we have more information than we really do
- Temporal dependence
  - Occurs when we measure our variables over time
  - Observation in time t is closely related to observation in time t-1
- These are sometimes called “autocorrelation” because they assign patterns that are really due to time and space to relationships between X and Y

Trucking in Aceh

Economists want to know when and how much people pay bribes in conducting business. So they rode around with truckers in Indonesia. The most basic question: does more police checkpoints mean that drivers pay a higher bribe?

- Outcome variable: $$$ of bribes
- Explanatory variable: number of police checkpoints

Trucking in Aceh

Economists want to know when and how much people pay bribes in conducting business. So they rode around with truckers in Indonesia. The most basic question: does more police checkpoints mean that drivers pay a higher bribe?

$$Bribe = \alpha + \beta_1(Num.\ checkpoint)$$

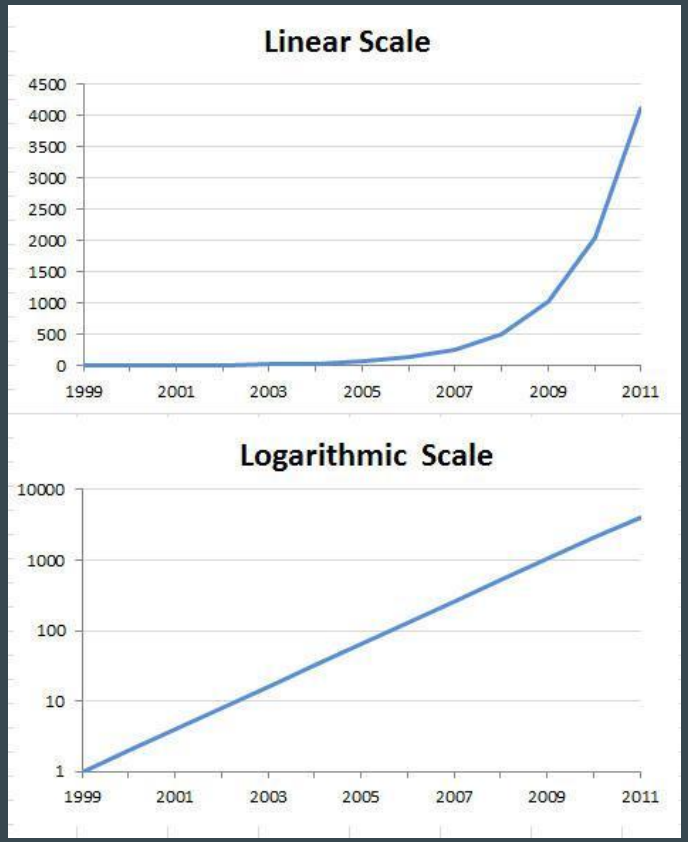
Trucking in Aceh

Economists want to know when and how much people pay bribes in conducting business. So they rode around with truckers in Indonesia. The most basic question: does more police checkpoints mean that drivers pay a higher bribe?

- Problem: neither X or Y is linear
    - This is common for any kind of count data

Simple fixes to non-linear relationships

- A linear models finds a *conditional probability*: E(Y | X)
- That means as long as you make constant changes to one of the variables, the *relative relationship* between those variables doesn't change



Trucking in Aceh

Economists want to know when and how much people pay bribes in conducting business. So they rode around with truckers in Indonesia. The most basic question: does more police checkpoints mean that drivers pay a higher bribe?

- Problem: neither X or Y is linear
- Solution: log both X and Y

Trucking in Aceh

Economists want to know when and how much people pay bribes in conducting business. So they rode around with truckers in Indonesia. The most basic question: does more police checkpoints mean that drivers pay a higher bribe?

$$log(Bribe) = \alpha + \beta_1 log(Num.\ checkpoint)$$

Trucking in Aceh

Economists want to know when and how much people pay bribes in conducting business. So they rode around with truckers in Indonesia. The most basic question: does more police checkpoints mean that drivers pay a higher bribe?

- Problem: bribes likely have spatial dependence
  - If more areas are prone to corrupt police...
  - ...this violates our assumptions of independence or random selection
- Solution: use clustered standard errors
  - This doesn't change our equation!

Trucking in Aceh

Economists want to know when and how much people pay bribes in conducting business. So they rode around with truckers in Indonesia. The most basic question: does more police checkpoints mean that drivers pay a higher bribe?

	Meulaboh OLS (1)
Log expected check-points in district	.663***
	(.081)
Observations	1,090

Trucking in Aceh, continued

They found a positive relationship between number of checkpoints and bribes. But they wanted to go further. Did all checkpoints affect the extent of bribes equally? Or where *greater number of guards*, and the *weapons those guards were carrying* have a relationship with the cost of the bribe?

- Explanatory variables:
  - Number of guards at checkpoint
  - Number of armed guards
- Outcome variable:
  - $$$ of bribe

Trucking in Aceh, continued

They found a positive relationship between number of checkpoints and bribes. But they wanted to go further. Did all checkpoints affect the extent of bribes equally? Or where *greater number of guards*, and the *weapons those guards were carrying* have a relationship with the cost of the bribe?

- Explanatory variables:
  - Number of guards at checkpoint
  - Number of armed guards
- Outcome variable:
  - $$$ of bribe

Trucking in Aceh, continued

They found a positive relationship between number of checkpoints and bribes. But they wanted to go further. Did all checkpoints affect the extent of bribes equally? Or where *greater number of guards*, and the *weapons those guards were carrying* have a relationship with the cost of the bribe?

$$log(bribes) = \alpha + \beta_1 Num.\ Guards + \beta_2 Guns$$

Trucking in Aceh, continued

They found a positive relationship between number of checkpoints and bribes. But they wanted to go further. Did all checkpoints affect the extent of bribes equally? Or where *greater number of guards*, and the *weapons those guards were carrying* have a relationship with the cost of the bribe?

	Log Payment	
	(1)	(2)
Gun visible	.166**	.154**
	(.066)	(.070)
Gun visible at subsequent checkpoint		.016
		(.024)
Number of officers at checkpoint	.047***	.050***
	(.010)	(.009)
Number of officers at subsequent checkpoint		−.003
		(.007)
Observations	5,260	4,968

Trucking in Aceh, wrap-up

1. Interactions:
    a. Could have included interaction b/w guns and number of officers
2. Spatial dependence:
    a. Accounted for through clustered standard errors
3. Temporal dependence
    a. Not accounted for!
4. Endogeneity
    a. Not really a concern
5. Omitted variable bias
    a. What about politically connected trucking firms?
6. Additivity violations
    a. Possible heteroscedasticity

Tropical economic development

Economists were interested in why certain countries have higher rates of GDP growth. According to traditional economic theory, this should be due to geography, resources, and other “structural” factors, instead of political institutions. One economist, Jeff Sachs, believed that tropical countries faced greater constraints to economic development.

- Explanatory variables: geographic factors
- Outcome variable: GDP

Tropical economic development

Sachs thought the following geographic factors would impact economic growth: presence of oil; tropical climate zone; whether a country was landlocked.

$$GDP = \alpha + \beta_1 Tropical + \beta_2 Oil + \beta_3 Landlocked$$

Tropical economic development

Sachs thought the following geographic factors would impact economic growth: presence of oil; tropical climate zone; whether a country was landlocked.

Table 3
Influence of Geography on the Level of Economic Development

	(1) lgdp65	(2) lgdp65	(3) lgdp65	(4) lgdp65	(5) lgdp74	(6) lgdp74	(7) lgdp74
Tropical Area	-0.83**	-0.76**			-0.46*		
(%)	(4.31)	(2.58)			(2.44)		
Pop 100 km		0.57					
(%)		(1.09)					
Oil	1.51**	1.30**	0.99**	1.18**	0.94**	0.72**	0.93**
	(11.30)	(5.71)	(8.10)	(12.52)	(11.07)	(5.79)	(9.24)
Malaria			-2.04**	-1.86**		-0.98	
Index 1966			(2.82)	(5.13)		(1.23)	
Andean				-0.31*			-0.32
				(2.46)			(1.88)
Landlock			-0.90**	-0.86**	-0.62**	-0.81**	-0.68**
			(6.12)	(4.51)	(3.69)	(5.13)	(3.09)
Constant	8.24	7.99	8.12	8.23	8.43	8.28	8.29
	(50.07)	(19.89)	(51.51)	(59.84)	(50.53)	(51.20)	(57.46)
Observations	10	10	10	10	10	10	10
Adj. R^2	0.71	0.74	0.81	0.89	0.82	0.69	0.73

Absolute values of t-statistics in parenthesis
* Significant at 10%level; ** Significant at 5% level
All regressions corrected for White heteroskedasticity-consistent standard errors and covariance

Problem one: interactions

The question we should ask ourselves when evaluating this problem is whether there is any *combined effect* of the explanatory variables:

- Do tropical oil states face unique challenges?
- What about landlocked tropical states?
- Or landlocked oil states?

Problem two: endogeneity

This is where we're probably good! It is almost impossible for a dynamic variable like GDP growth to affect something static, like geography.

Problem three: spatial dependence

Absolutely!

- Say that there is a “contagion” process, where if the neighboring country has a low GDP, it “infects” the countries next to it
- Then, if that “contagious” country is a tropical country, we will see a correlation, even if being a tropic country has no effect

Problem four: temporal dependence

Absolutely!

- Anything measured over time has temporal dependence
- If the tropical countries start in t=0 as poorer, then they will continue to be poorer, thereby producing a correlation when there is none
- Slightly less problematic because the geographic variables themselves don’t change with time

Problem five: omitted variable bias

Remember, an omitted variable is one that correlated with both X and Y variables

- What about colonialism?
- Colonial powers extracted resources, thereby making the colonized countries less rich (aka lowering GDP)
- Colonial powers imposed their own institutions, which might not be best suited to the colonized countries, (aka lowering GDP)
- Colonial powers were exclusively in temperate climates and most colonies were in tropical areas

Additivity violations

- Heteroscedasticity: there is non-constant variance
  - Yes!
  - Oil: not much variance at low levels of oil
  - Tropical countries: more of them; higher populations
- Relationships between X and errors: predicted values are less accurate at higher levels of an X variables
  - Probably?
  - Can't tell without looking at residuals

Achen's rule of three

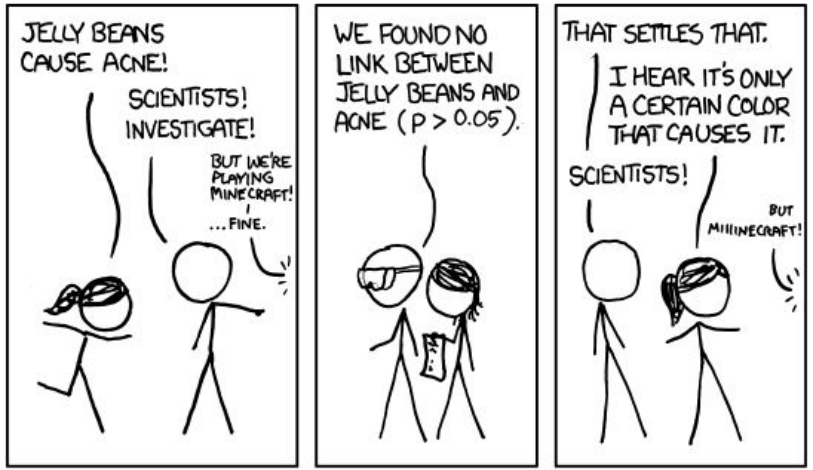
- If you have more than three explanatory variables it's almost impossible to interpret the regression
- This is because it's hard to think in more than three dimensions
- The more X variables you have the more likely that you will violate assumptions

Achen's rule of three

Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
β_1 Legislature	−0.519	−1.448*	−1.451*	−1.342*	−1.154+	−1.607*	−1.426*	−1.225+
	(0.39)	(0.68)	(0.66)	(0.66)	(0.66)	(0.68)	(0.68)	(0.69)
β_2 Military Legislature		1.012	0.834	0.766	0.938	1.052	1.033	0.569
		(0.68)	(0.70)	(0.68)	(0.70)	(0.67)	(0.68)	(0.66)
β_3 Military No Legislature		−1.614*	−1.623*	−1.499*	−1.194+	−1.687*	−1.581*	−1.284+
		(0.73)	(0.71)	(0.71)	(0.71)	(0.74)	(0.73)	(0.76)
β_4 Single Party Legislature		1.061*	1.079*	1.064*	1.019*	1.130*	1.000*	0.459
		(0.45)	(0.48)	(0.43)	(0.45)	(0.44)	(0.45)	(0.45)
β_5 Single Party No Legislature		−0.080	−0.433	0.005	0.301	−0.080	−0.041	−0.410
		(0.91)	(1.38)	(0.91)	(0.90)	(0.91)	(0.91)	(0.92)
β_6 Monarchy Legislature		1.402*	1.249*	0.885	0.835	1.624*	1.427*	0.251
		(0.64)	(0.60)	(0.64)	(0.58)	(0.66)	(0.64)	(0.79)
β_7 Monarchy No Legislature		2.194*	2.167+	1.772+	2.042+	2.425*	2.221*	1.167
		(1.11)	(1.12)	(1.08)	(1.05)	(1.14)	(1.11)	(1.25)
Log(GDPpc)	0.583	0.534	0.685+	0.302	0.132	0.570	0.563	0.818+
	(0.38)	(0.39)	(0.37)	(0.36)	(0.31)	(0.39)	(0.39)	(0.43)
Ethnic Frac.	−2.314**	−2.491**	−2.303**	−2.084**	−1.873**	−2.614**	−2.375**	−2.892**
	(0.64)	(0.67)	(0.72)	(0.65)	(0.63)	(0.68)	(0.68)	(0.69)
Sub-Saharan Africa	0.443	0.504	0.731	−0.027		0.625	0.541	1.034
	(0.59)	(0.61)	(0.62)	(0.57)		(0.62)	(0.61)	(0.71)
British Colony	1.224*	1.384**	1.162*	1.152*		1.319**	1.379**	1.454**
	(0.48)	(0.49)	(0.47)	(0.48)		(0.48)	(0.49)	(0.49)
Investment (% GDP)	0.168**	0.173**	0.186**	0.172**	0.179**	0.174**	0.169**	0.153**
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Govt Consumption	−0.204**	−0.210**	−0.193**	−0.177**	−0.182**	−0.216**	−0.207**	−0.210**
	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Inflation	−0.000*	−0.000*	−0.000*	−0.007**	−0.000*	−0.000*	−0.000*	−0.000*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
1960s	1.053*	0.960+	0.912+	0.821	0.832	0.963+	1.007+	1.270*
	(0.52)	(0.53)	(0.53)	(0.51)	(0.53)	(0.53)	(0.53)	(0.52)
1970s	1.410**	1.452**	0.887*	1.348**	1.346**	1.500**	1.485**	1.762**
	(0.36)	(0.37)	(0.38)	(0.36)	(0.37)	(0.37)	(0.37)	(0.37)
Polity						0.057+		
						(0.03)		
Communist							1.271	
							(1.30)	
Military	−0.422							
	(0.51)							
Single Party	0.763+							
	(0.40)							
Monarchy	1.502**							
	(0.52)							
Constant	−2.181	−1.119	−2.588	0.457	1.417	−0.974	−1.438	−3.190
	(2.65)	(2.83)	(2.61)	(2.67)	(2.19)	(2.87)	(2.84)	(3.08)
R^2	0.130	0.135	0.147	0.165	0.126	0.137	0.135	0.154
Observations	1576	1576	1279	1571	1576	1575	1576	1576
Countries	80	80	73	80	80	80	80	80

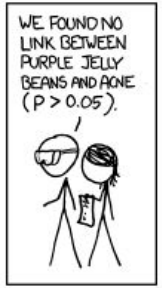
Be careful of running multiple models

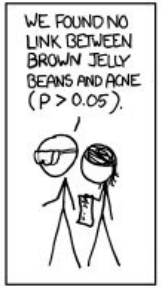
- You have noticed that in most of our examples, each regression has multiple columns, signifying different models with new variables included or old ones excluded
- Remember the probabilistic nature of p-values!
- There are an almost infinite number of variables you can include in most social science regressions. If you run an infinite number of models, your chance of finding a statistically significant relationship approaches 1!

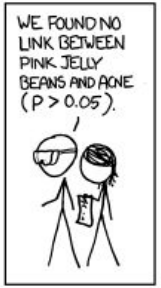

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.

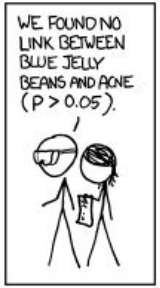
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).

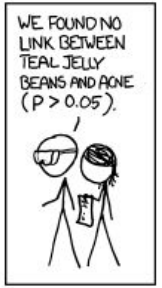
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

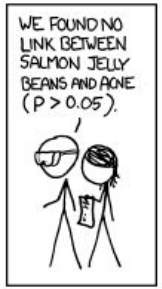

WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).

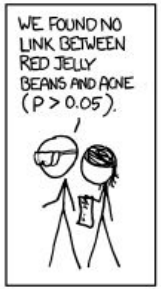

WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).

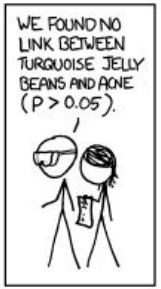

WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).

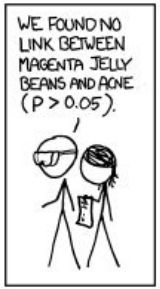

WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).

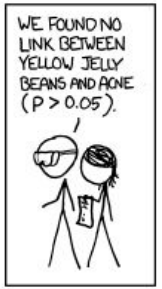

WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).

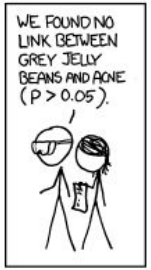

WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).

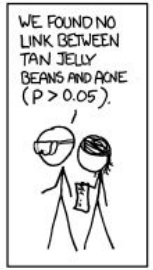

WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).

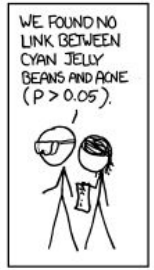

WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).

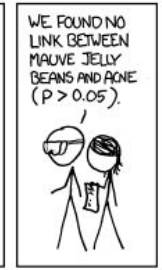

WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

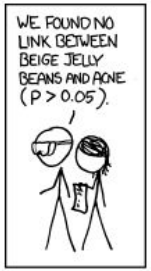

WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).

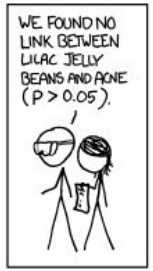

WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).

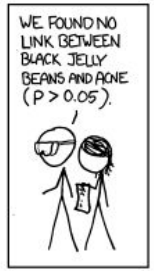

WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).

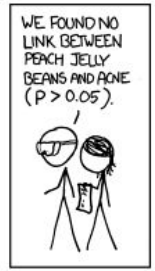

WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!

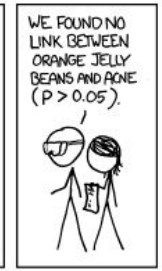

WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).

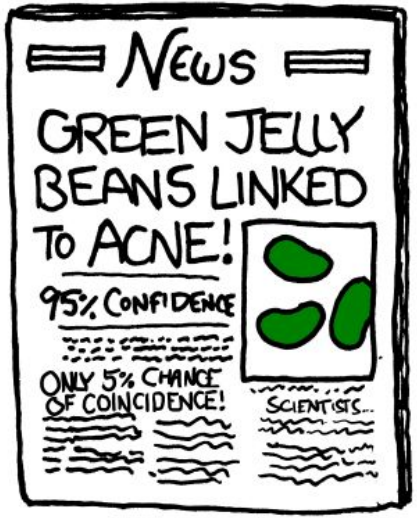

WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Review

We looked at a few different social scientific models, and evaluated them for a variety of assumptions:

1. Interactions
2. Endogeneity
3. Omitted variable bias
4. Temporal dependence
5. Spatial dependence
6. Additivity assumptions

Review

These categories are helpful, but they're not mutually exclusive!

- You can think of temporal dependence as omitted variable bias
  - The omitted variable is time
- Interactions are violations of additivity
  - It means that your errors will be correlated with your residuals since you fail to adjust for their combined effect
- It all comes back to linearity!