Global Primary School Completion

Brian & Hannah

“Education is the most powerful weapon which you can use to change the world.”

Goal: To Predict the primary school completion rate of each country

Target variable “Primary school completion rate”

Data Sources

- World Bank
  - A “financial institution that provides loans and grants to the governments of poorer countries for the purpose of pursuing capital projects”
  - Public data on all sorts of features of countries
- UNICEF
  - A division of the UN to “save children’s lives, to defend their rights, and to help them fulfil their potential”
  - Used their “State of the World’s Children” report

Independent Variables Investigated

- Child employment rates
- Proportion of GDP spent on education
- Population density
- **Urban population**
- **Proportion of agricultural land**
- **Adolescent birth rate**
- **Improved sanitation**
- **Average immunization**
- Average support in learning from fathers
- Region
- **Income relative to other countries**

Our First Model

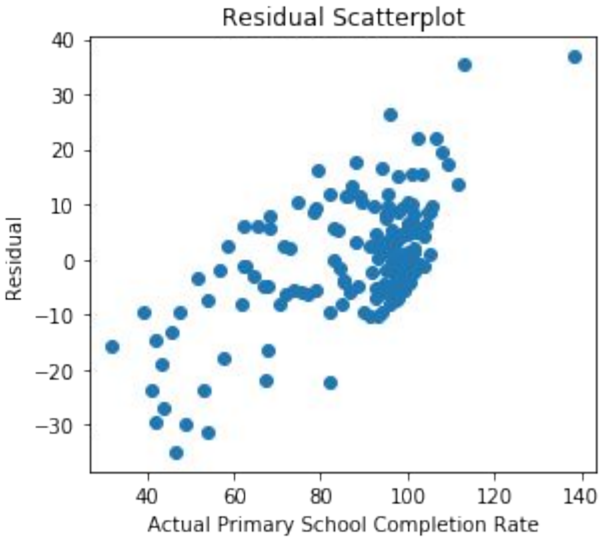
	coef	P>|t|
Intercept	69.6422	0.000
region[T.Europe & Central Asia]	-9.4332	0.004
region[T.Latin America & Caribbean]	-3.5407	0.330
region[T.Middle East & North Africa]	-13.3548	0.000
region[T.North America]	-13.8065	0.212
region[T.South Asia]	-0.9034	0.847
region[T.Sub-Saharan Africa]	-12.2644	0.001
avg_pop_density	-0.0010	0.528
avg_urban_pop	0.0462	0.420
agricultural_land	0.0155	0.719
adolescent_birth_rate	-0.1068	0.003
improved_sanitation_total	0.2412	0.000
immunization_avg	0.1127	0.108
relative_country_income	1.8889	0.240

R-squared:	0.693
Adj. R-squared:	0.666

- Wide variety of p values
- r^2 is pretty good

A Step in the Right Direction

R-squared:		0.632
Adj. R-squared:		0.618
	coef	**P>|t|**
Intercept	66.0661	0.000
avg_urban_pop	-0.0239	0.662
agricultural_land	-0.0285	0.506
adolescent_birth_rate	-0.1151	0.000
improved_sanitation_total	0.2877	0.000
immunization_avg	0.0883	0.198
relative_country_income	1.9333	0.233



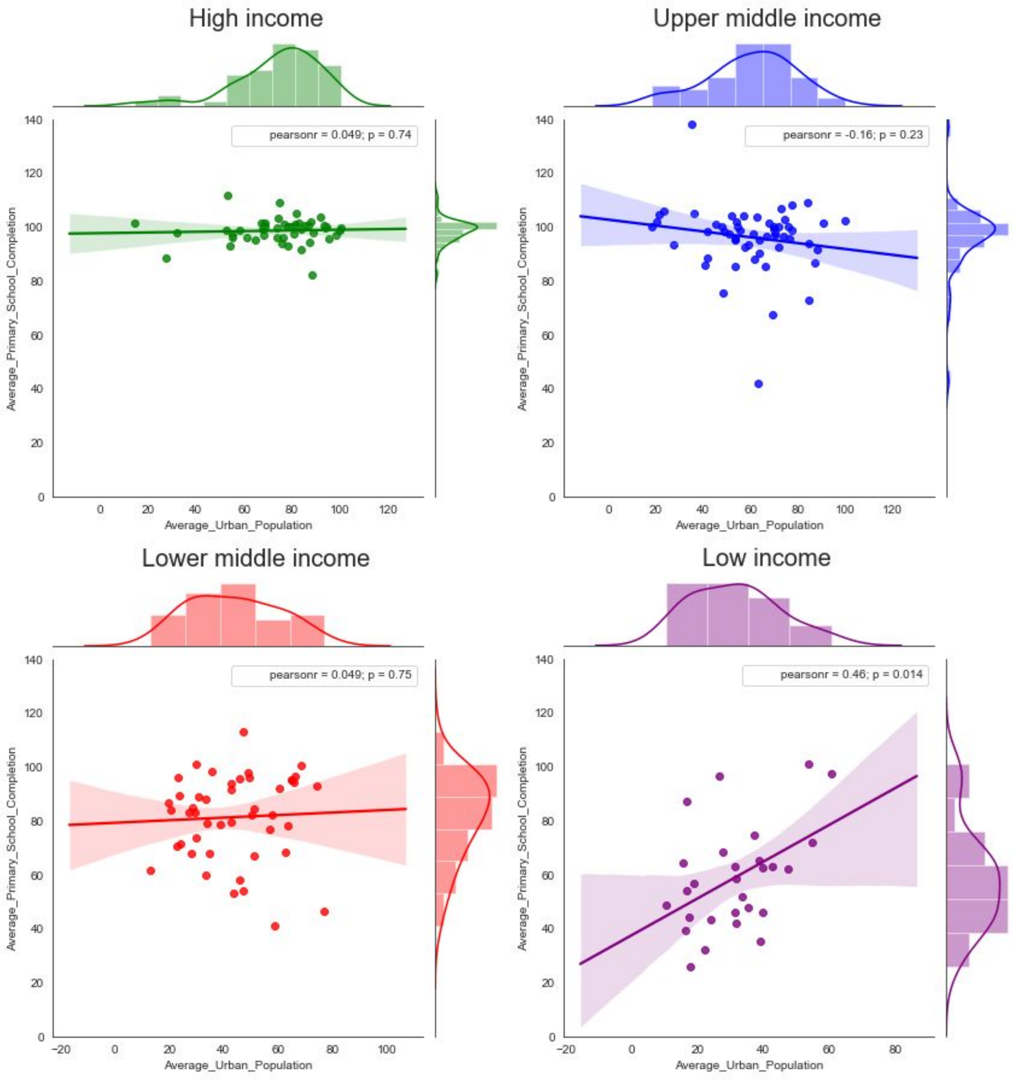
- Lower r^2, but this is not the end all, be all
- P values improved some
- No interaction terms
- Relative country income was not a good predictor - what's going on here?

Motivation for Final Model Adjustments

Our hypothesis with 95% confidence is:

H_0 : Income does not affect the Average Urban Population and Primary School Completion relationship.

H_a : Income does affect the Average Urban Population and Primary School Completion relationship.



Low income is the only income group with a significant p-value, so we can reject the null hypothesis for low income.

All the other income groups have insignificant p-values, so we fail to reject the null hypothesis for them.

An important mathematical discovery!

That took more than an hour of seven whole people's time to come to a conclusion on

Question: If you standardize your variables, and *then* compute your interaction terms, will you get the same p values in your new LR model as computing interaction terms and *then* standardizing them?

Answer: No!!!!!!!!!!!!!

- By standardizing, then computing, you are in essence standardizing twice:

$$f(v_1) \cdot f(v_2) = f^2(v_1 \cdot v_2)$$

Instead of

$$f(v_1 \cdot v_2)$$

(let f be the scaling function)

- MOTS: Always compute interaction terms and THEN scale.
  - Math makes sense

Another important mathematical discovery!

This one took less time to understand (yay)

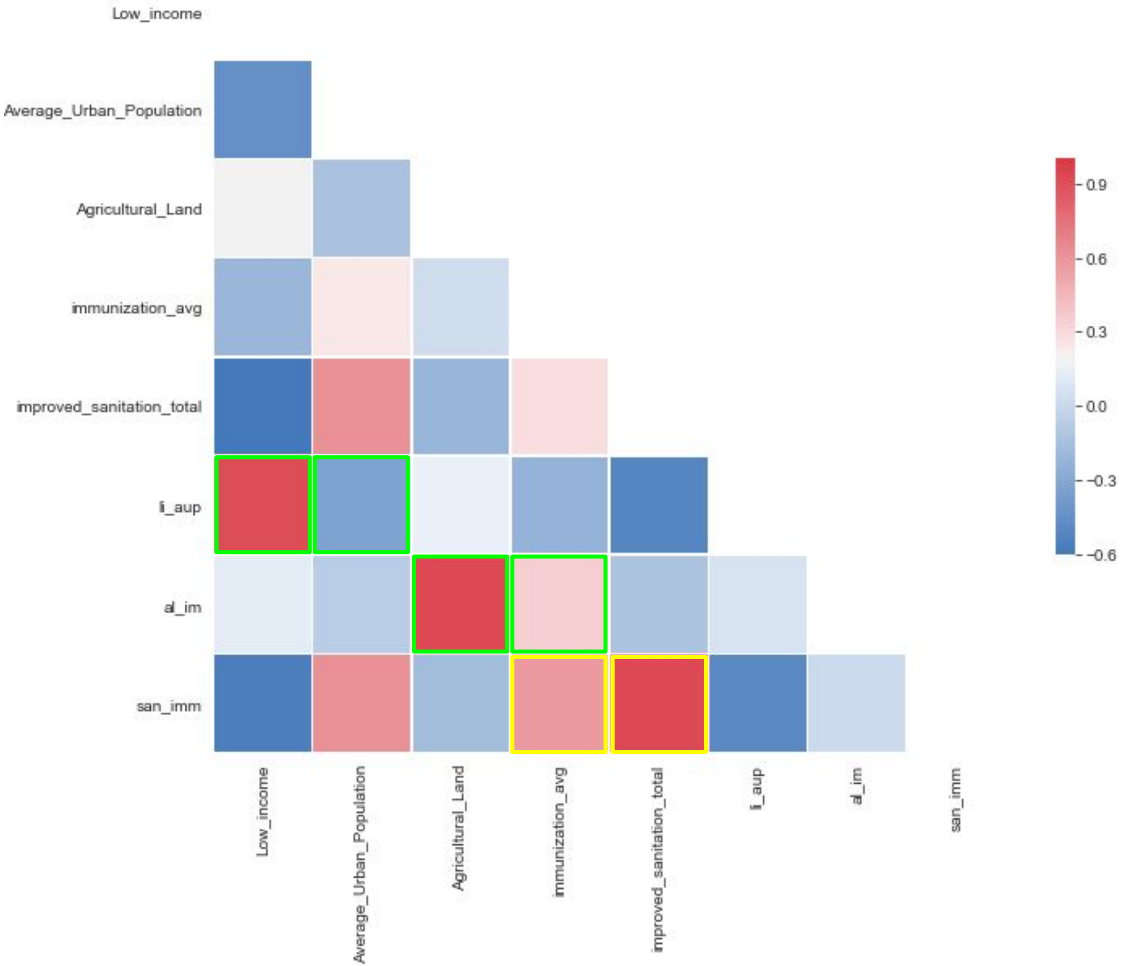
Question: Why is multicollinearity not a problem when including interaction terms? Don't variable X and variable Y perfectly predict the interaction XY?

Answer:

- With dummy variable columns C_1, C_2, and C_3, no new information is given by C_3, so the model struggles to assign a meaningful non-zero coefficient to C_3.
  - No new information because:
    - $C_1+C_2 = 0 \Rightarrow C_3 = 1$
    - Else, $C_3 = 0$
- With X and Y, additional information is given by XY that you couldn't get from just adding X and Y! So, a meaningful coefficient can be found for XY.

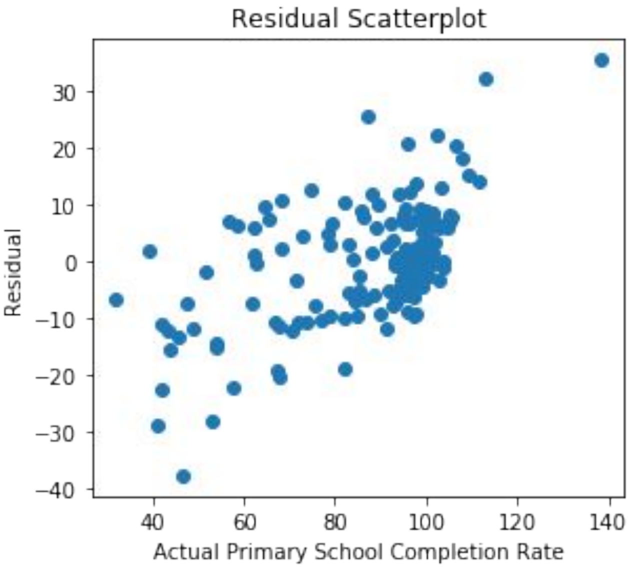
Another important mathematical discovery!

This one took less time to understand (yay)



Final Model

R-squared:		0.688
Adj. R-squared:		0.672
	coef	**P>|t|**
Intercept	92.0492	0.000
avg_urban_pop	-1.5156	0.168
agricultural_land	8.3424	0.050
adolescent_birth_rate	-4.9168	0.000
improved_sanitation_total	8.0647	0.000
immunization_avg	4.9752	0.003
low_income	-27.4197	0.000
li_aup	6.2117	0.004
al_ia	-9.7128	0.032



- Better r^2 and adjusted r^2
- Very low p values, for the most part
- Residual scatter plot improved

What can we conclude?

Biggest increasers of primary school completion:

- Having a lot of agricultural land
- Having improved sanitation
- Having higher rates of immunization

Biggest decreasers of primary school completion:

- Being a low income country
- The interaction of being a low income country with a high urban population
- Having a high adolescent birth rate