Tutorial 05, Hilary Term

Research Methods for Political Science (PO3600)

Stefan Müller

20 February 2018

Trinity College Dublin
http://muellerstefan.net/research-methods

Multivariate Regression

What predicts wealth (measured as GDP per capita)?

Multivariate Regression

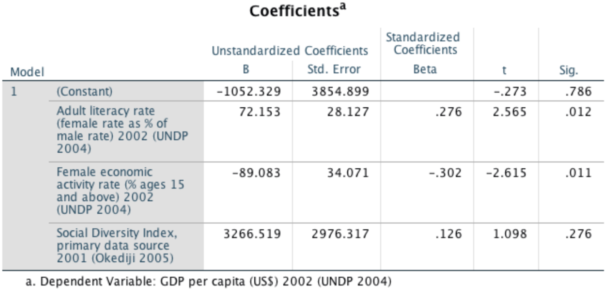
What predicts wealth (measured as GDP per capita)?

Dependent variable: GDP per capita (US$) 2002 (UNDP 2004)

Independent variables:

- `FM_Lit2002`: Adult illiteracy rate (% ages 15 and above) 2002 (UNDP 2004)
- `F_Work2002`: Female economic activity rate (% ages 15 and above) 2002 (UNDP 2004)
- `SDI`: Social Diversity Index, primary data source 2001 (Okediji 2005)

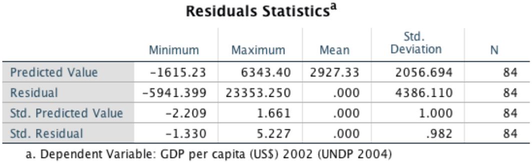
Regression Coefficients



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-1052.329	3854.899		-.273	.786
	Adult literacy rate (female rate as % of male rate) 2002 (UNDP 2004)	72.153	28.127	.276	2.565	.012
	Female economic activity rate (% ages 15 and above) 2002 (UNDP 2004)	-89.083	34.071	-.302	-2.615	.011
	Social Diversity Index, primary data source 2001 (Okediji 2005)	3266.519	2976.317	.126	1.098	.276

a. Dependent Variable: GDP per capita (US$) 2002 (UNDP 2004)



Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-1615.23	6343.40	2927.33	2056.694	84
Residual	-5941.399	23353.250	.000	4386.110	84
Std. Predicted Value	-2.209	1.661	.000	1.000	84
Std. Residual	-1.330	5.227	.000	.982	84

a. Dependent Variable: GDP per capita (US$) 2002 (UNDP 2004)

Regression Diagnostics

1. **Influential data points/outliers**
2. Independence/**autocorrelation** (errors associated with one observation not correlated with errors in any other observation)
3. **Linearity** (relationship should be linear)
4. **Homoscedasticity** (constant error variance)
5. Normality (errors should be normally distributed)
6. Model specification
7. Multicollinearity (predictors are highly correlated)
8. Leverage (extent to which predictor differs from mean of predictor)

Performing an F-test

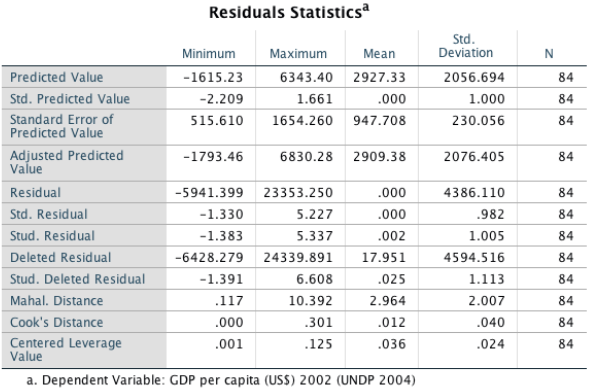
- F ratio tests H0 that *all* slopes in the model $= 0$
- $F = \frac{MS_M}{MS_R}$
- $MS_M = \frac{SS_M}{df}$ $(df = 1)$
- $MS_R = \frac{SS_R}{df}$ $(df = \text{n-p-1})$
- Important: even if the model fits quite poorly, F is mostly substantially larger than 1.

Diagnostics: Check for influential points

Cook's Distance

- How much the predicted scores for other observations would differ if the observation in question were not included?
- Cook's Distance: influence of an observation and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question.
- A common rule of thumb is that an observation with a value of Cook's D over 1.0 has too much influence.

Check for influential points



Residuals Statistics[a]

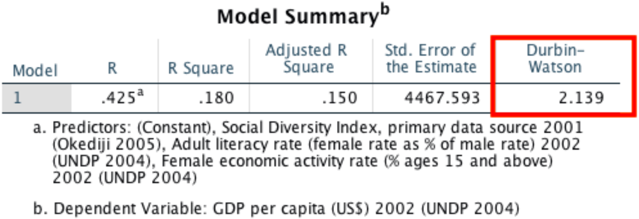
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-1615.23	6343.40	2927.33	2056.694	84
Std. Predicted Value	-2.209	1.661	.000	1.000	84
Standard Error of Predicted Value	515.610	1654.260	947.708	230.056	84
Adjusted Predicted Value	-1793.46	6830.28	2909.38	2076.405	84
Residual	-5941.399	23353.250	.000	4386.110	84
Std. Residual	-1.330	5.227	.000	.982	84
Stud. Residual	-1.383	5.337	.002	1.005	84
Deleted Residual	-6428.279	24339.891	17.951	4594.516	84
Stud. Deleted Residual	-1.391	6.608	.025	1.113	84
Mahal. Distance	.117	10.392	2.964	2.007	84
Cook's Distance	.000	.301	.012	.040	84
Centered Leverage Value	.001	.125	.036	.024	84

a. Dependent Variable: GDP per capita (US$) 2002 (UNDP 2004)

Diagnostics: Autocorrelation

- Assumption: observations are independent
- Durbin-Watson statistic should be between 1.5 and 2.5

Diagnostics: Autocorrelation



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.425[a]	.180	.150	4467.593	2.139

a. Predictors: (Constant), Social Diversity Index, primary data source 2001 (Okediji 2005), Adult literacy rate (female rate as % of male rate) 2002 (UNDP 2004), Female economic activity rate (% ages 15 and above) 2002 (UNDP 2004)

b. Dependent Variable: GDP per capita (US$) 2002 (UNDP 2004)

Diagnostics: Linearity and Homoscedasticity

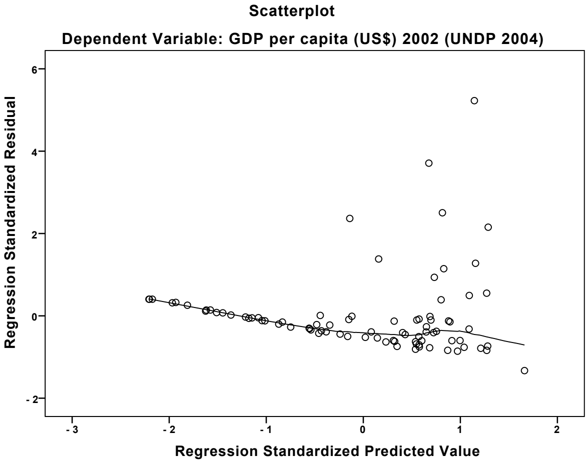
Linearity: We assume that the relationship between the response variable and the predictors is linear. If this assumption is violated, the linear regression will try to fit a straight line to data that do not follow a straight line.

Homoscedasticity: Variance of the residuals is homogeneous across levels of the predicted values.

If your residuals are normally distributed and homoscedastic, you do not have to worry about the linearity assumption!

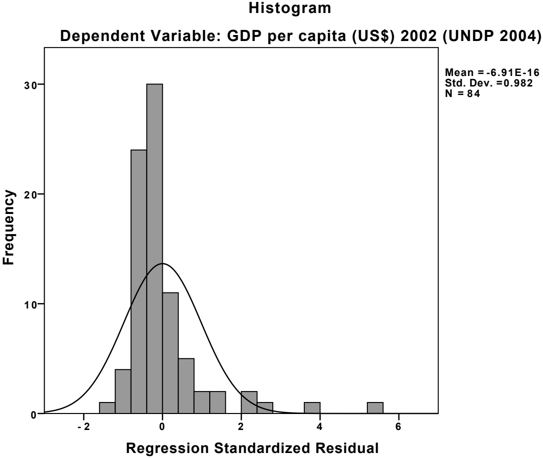
- Plot of `ZRESID` and `ZPRED`
- Interpret the plot

Homoscedasticity



Homoscedasticity

Residuals should be normally distributed.



Useful Explanation

SPSS Regression Diagnostics: https://stats.idre.ucla.edu/spss/seminars/introduction-to-regression-with-spss/introreg-lesson2/

Doing it all in R: https://github.com/stefan-mueller/research-methods/blob/master/code/ht05/ht_05_replicate_spss.R