Tutorial 06, Hilary Term

Research Methods for Political Science (PO3600)

Stefan Müller

6 March 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]

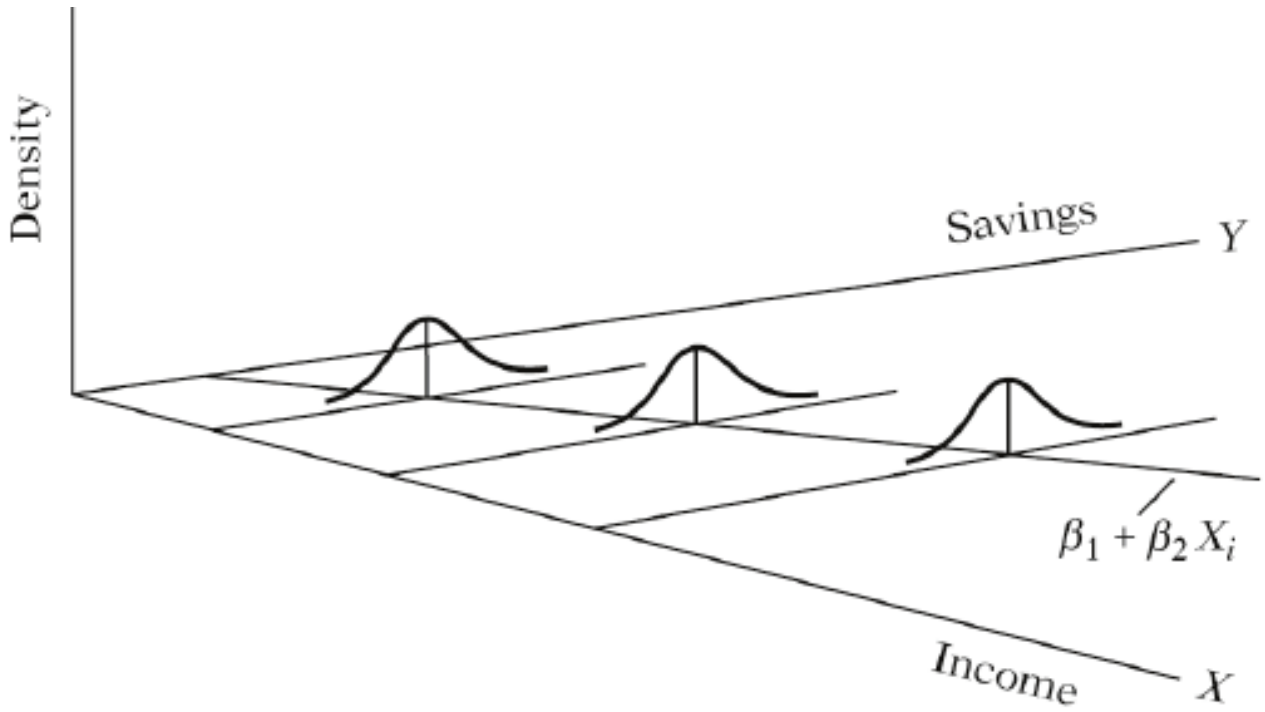
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-1615.23	6343.40	2927.33	2056.694	84
Residual	-5941.399	23353.250	.000	4386.110	84
Std. Predicted Value	-2.209	1.661	.000	1.000	84
Std. Residual	-1.330	5.227	.000	.982	84

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

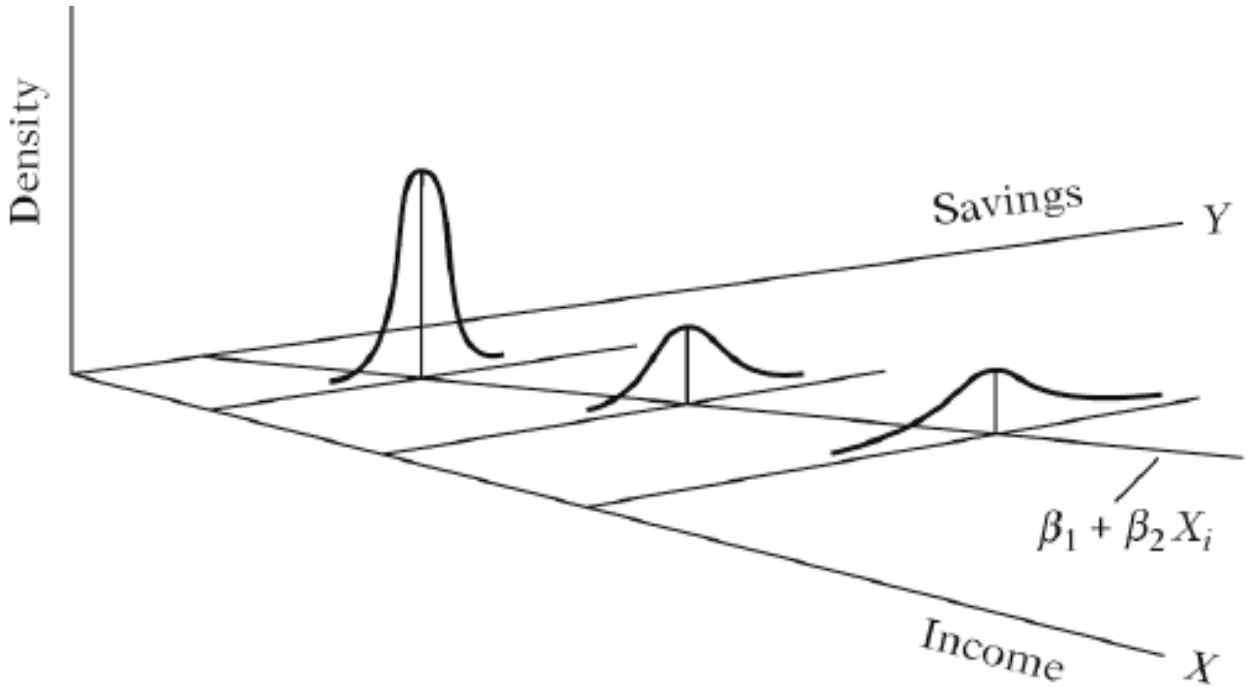
Regression Diagnostics

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)

Recap: Homoscedasticity



Recap: Heteroscedasticity

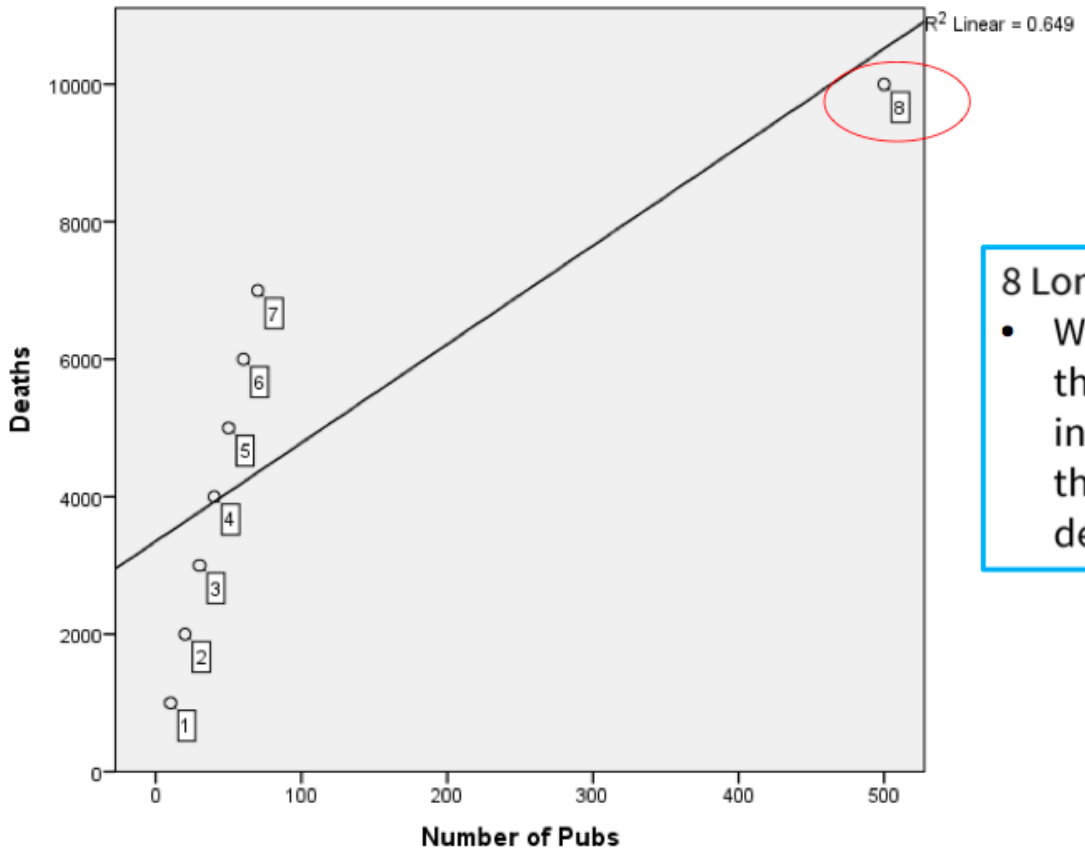


Recap: Homoscedasticity

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

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

Outliers vs. Influential Points

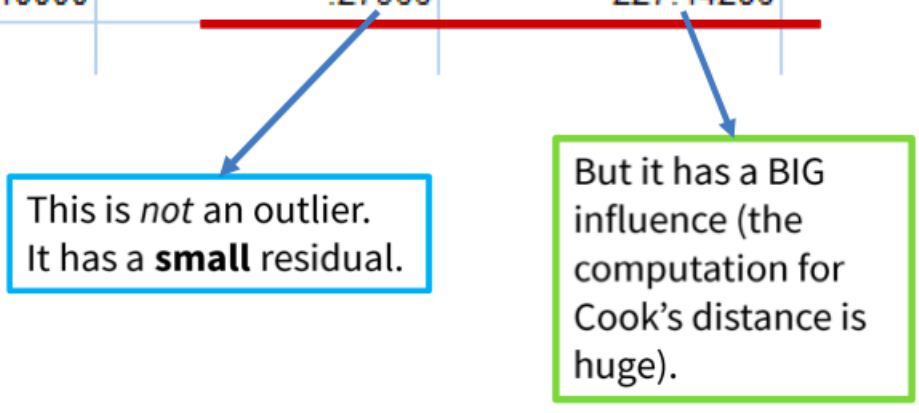


8 London boroughs.

- We want to know if the number of pubs in the area predicts the number of deaths.

Outliers vs. Influential Points

pubs	mortality	ZRE_1	COO_1
10	1000	-1.33839	.21328
20	2000	-.87895	.08530
30	3000	-.41950	.01814
40	4000	.03995	.00015
50	5000	.49940	.02294
60	6000	.95885	.08092
70	7000	1.41830	.17107
500	10000	-.27966	227.14286



Diagnostics: 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: Multicollinearity

- Perfect multicollinearity: SPSS will remove one predictor
- High multicollinearity when two of the independent variables are highly correlated
- Standard errors will be large: more uncertainty in the model
- Check multicollinearity by calculating Variance Inflation Factor (VIF): $VIF > 5$ is definitely cause for concern!

Exercise: Categorical Predictors and Interaction Terms

1. Use the ANES 2016 pilot dataset (available on Blackboard)
2. Download the codebook: https://tinyurl.com/anescodebook
3. Create summary statistics of the following variables: `pid1d`, `educ`, `gender`, `ftobama`
4. Create new variable called `piddem` which is based on `pidid` and coded binary: 1: Democrat; 0: Not Democrat. Pay attention to the original coding of missing values and recode them as "system missing"
5. Check `ftobama` and correct the mistake by creating a new variable `ftobamarecoded`

Check Codebook

```
ftobama                                                 Feeling thermometer - Barack Obama
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  ftobama, but 92 nonmissing values are not labeled

                 range:  [0,100]                      units:  1
         unique values:  101                  missing .:  0/1,200
    unique mv codes:  1                  missing .*:  2/1,200

              examples:  3
                         31
                         70      70 - Fairly warm or favorable feeling
                         90
```

Check Codebook

```
-------------------------------------------------------------------------------
pid1d                                                  Party ID - Republican first
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  pid1d

                 range:  [1,4]                         units:  1
         unique values:  4                     missing .:  0/1,200
    unique mv codes:  1                     missing .*:  612/1,200

        tabulation:  Freq.    Numeric  Label
                       229          1  Democrat
                       123          2  Republican
                       195          3  Independent
                        41          4  Something else
                       612         .b  [9]Not Asked
```

Check Codebook

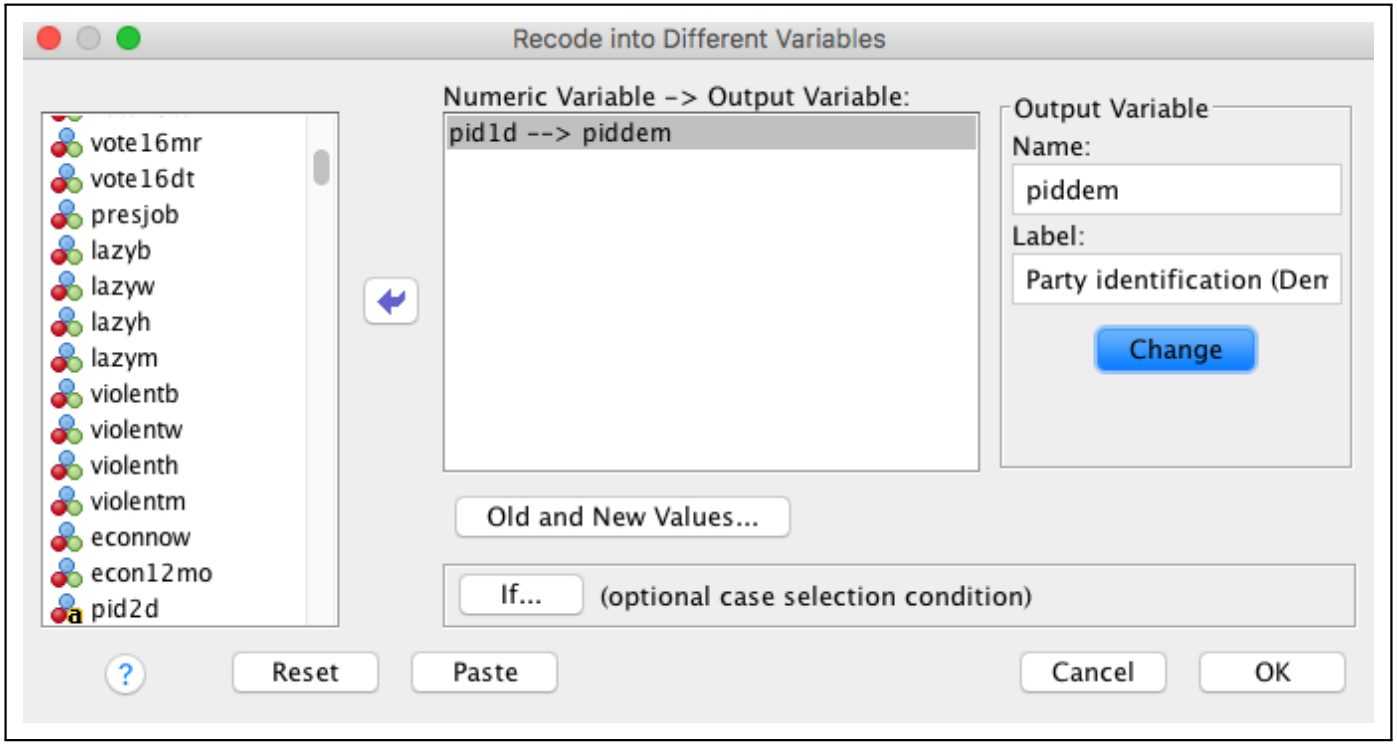
```
-------------------------------------------------------------------------------
educ                                                                Education
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  educ

                 range:  [1,6]                        units:  1
          unique values:  6                    missing .:  0/1,200

              tabulation:  Freq.   Numeric  Label
                             102         1  No HS
                             411         2  High school graduate
                             257         3  Some college
                             106         4  2-year
                             202         5  4-year
                             122         6  Post-grad
```

Check Codebook

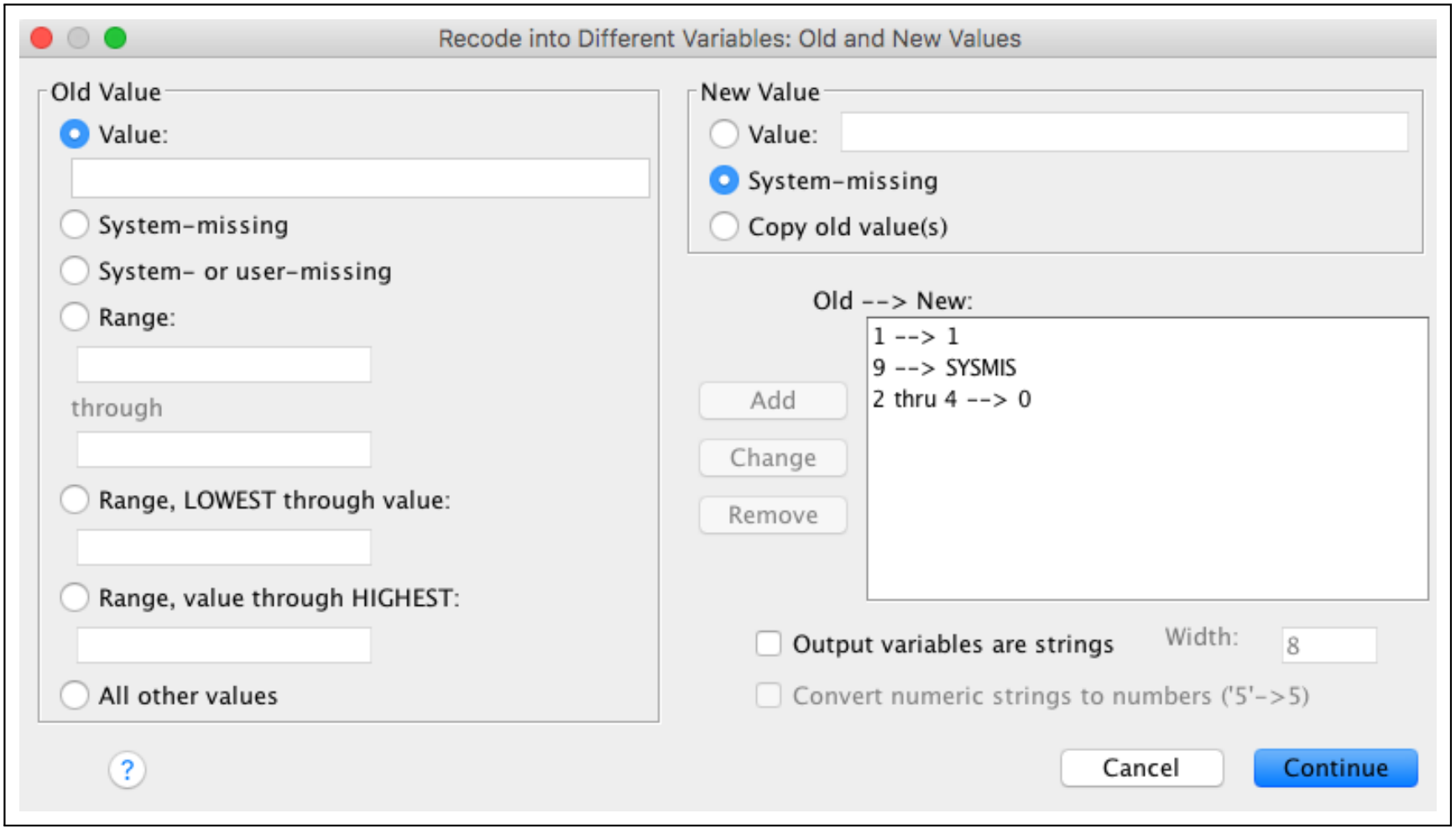
```
-------------------------------------------------------------------------------
gender                                                                 Gender
-------------------------------------------------------------------------------

                 type:  numeric (byte)
                label:  gender

                range:  [1,2]                        units:  1
        unique values:  2                    missing .:  0/1,200

            tabulation:  Freq.   Numeric  Label
                           570         1  Male
                           630         2  Female
```

Create a New Variable (I)

Transform $\longrightarrow$ Recode into Different Variable



Create a New Variable (II)



Exercise: Categorical Predictors and Interaction Terms

1. Dependent variable: `ftobamarecoded` (feeling thermometer for Obama)
2. Independent variables: `piddem`, `educ`, `gender`
3. Explain the coding of these variables
4. Interpret these coefficients (in substantive terms!)

Regression Results

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.626[a]	.392	.389	29.867

a. Predictors: (Constant), educ, Party identification (Democrat), gender

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	335341.201	3	111780.400	125.312	.000[b]
	Residual	520045.641	583	892.017		
	Total	855386.842	586			

a. Dependent Variable: ftobama

b. Predictors: (Constant), educ, Party identification (Democrat), gender

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	19.456	4.807		4.047	.000
	Party identification (Democrat)	47.872	2.551	.611	18.763	.000
	gender	5.400	2.490	.071	2.169	.031
	educ	.968	.818	.038	1.184	.237

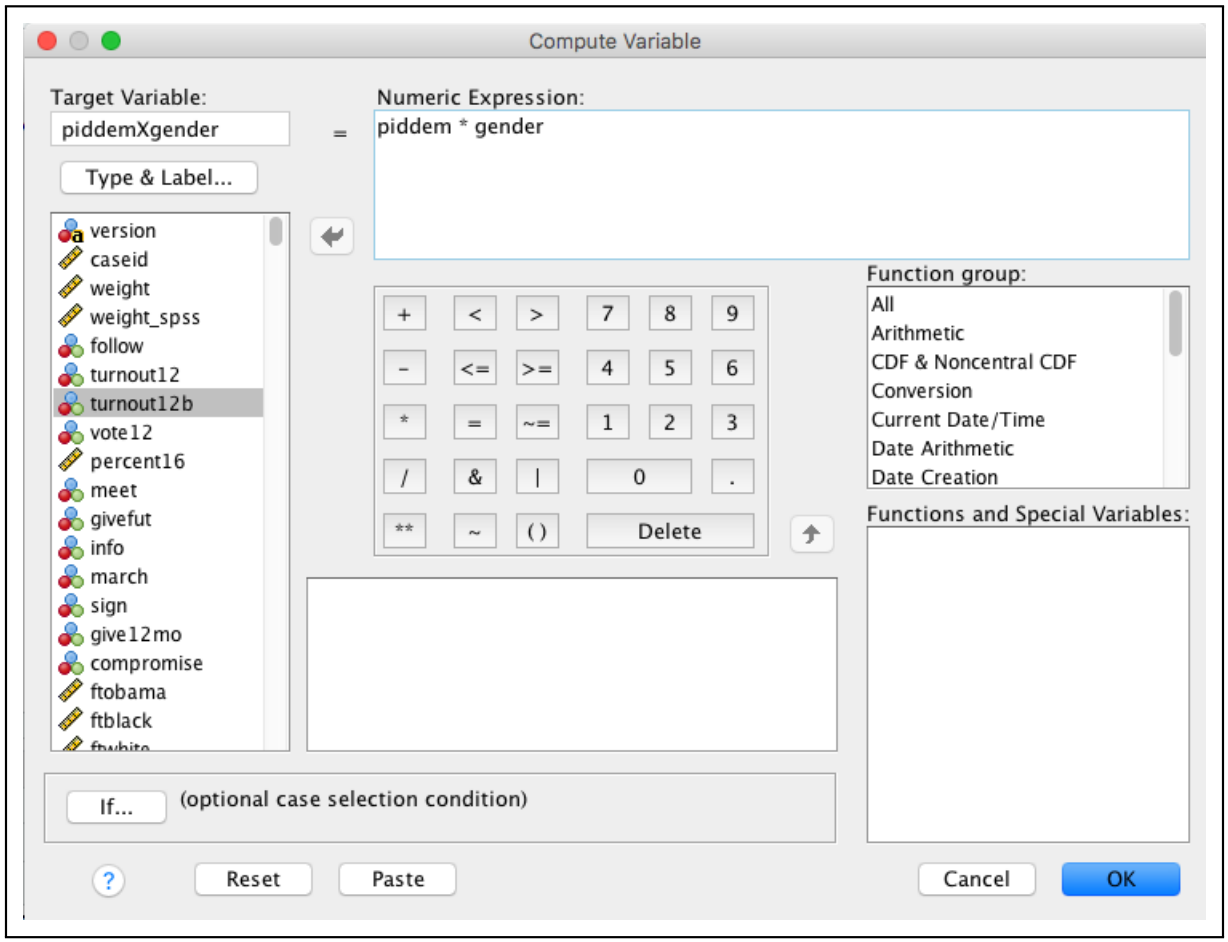
a. Dependent Variable: ftobama

Exercise: Categorical Predictors and Interaction Terms

1. Generate variable called `piddemXgender` which is an interaction of `gender` and `piddem`
2. Re-run the model and add `piddemXgender`
3. Interpret the coefficient of the interaction

Interaction Effect

Transform $\longrightarrow$ Compute Variable



Regression Results with Interaction

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.627[a]	.394	.390	29.851

a. Predictors: (Constant), piddemXgender, educ, gender, Party identification (Democrat)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	336759.705	4	84189.926	94.477	.000[b]
	Residual	518627.136	582	891.112		
	Total	855386.842	586			

a. Dependent Variable: ftobama

b. Predictors: (Constant), piddemXgender, educ, gender, Party identification (Democrat)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	15.712	5.647		2.782	.006
	Party identification (Democrat)	57.867	8.322	.739	6.953	.000
	gender	7.866	3.165	.103	2.486	.013
	educ	1.010	.818	.040	1.234	.218
	piddemXgender	-6.465	5.124	-.142	-1.262	.208

a. Dependent Variable: ftobama

Useful Resources

- Code to replicate analysis in SPSS: https://github.com/stefan-mueller/research-methods/blob/master/code/ht06/code_ht_06.sps
- A. F. Zuur et al. (2010). "A Protocol for Data Exploration to Avoid Common Statistical Problems". In: *Methods in Ecology and Evolution* 1.1, pp. 3–14