ENHANCING REGRESSION IN SPSS, PART II:

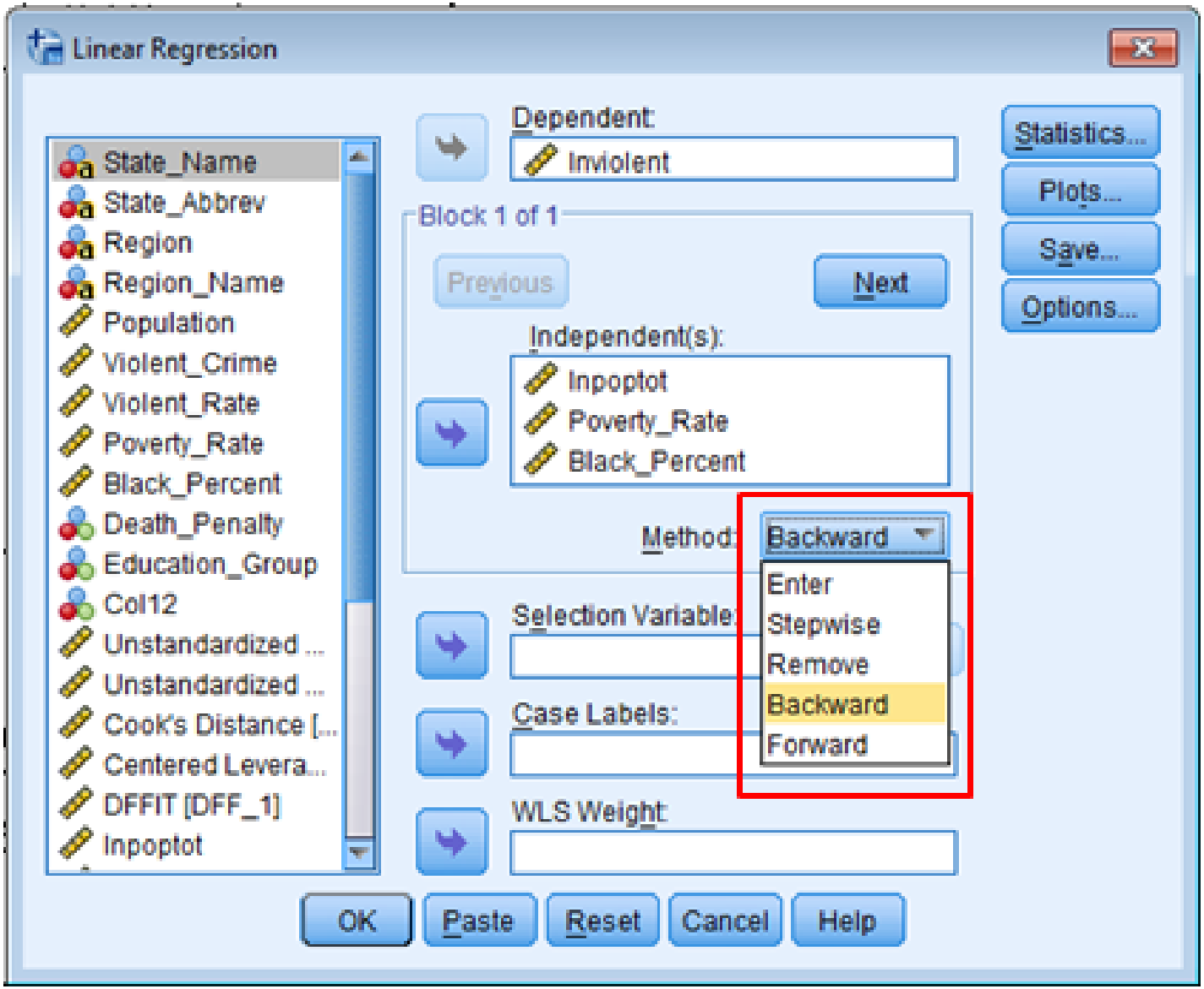
VARIATIONS IN REGRESSION ANALYSIS AND NON-RESPONSE ANALYSIS

Richard Lee Rogers

Last Update: April 2, 2016

STEPWISE REGRESSION

Backward Stepwise	Forward Stepwise
• Begin with all variables in model	• Start with no variables in the model
• Remove the least statistically significant variable	• Add the most significant variable from the pool of variables selected
• Continuing removing variables until all remaining variables are statistically significant	• Continue adding until all statistically significant variables are added to the model


Linear Regression
State_Name
State_Abbrev
Region
Region_Name
Population
Violent_Crime
Violent_Rate
Poverty_Rate
Black_Percent
Death_Penalty
Education_Group
Col12
Unstandardized ...
Unstandardized ...
Cook's Distance [...
Centered Levera...
DFFIT [DFF_1]
Inpoptot
Dependent:
Inviolent
Block 1 of 1
Previous
Next
Independent(s):
Inpoptot
Poverty_Rate
Black_Percent
Method:
Backward
Enter
Stepwise
Remove
Backward
Forward
Selection Variable:
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help

STEPWISE OUTPUT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.736	.703		6.739	.000
	lnpoptot	.035	.048	.086	.732	.468
	Poverty_Rate	.022	.016	.174	1.366	.178
	Black_Percent	.019	.005	.517	4.169	.000
2	(Constant)	5.222	.230		22.703	.000
	Poverty_Rate	.025	.016	.195	1.580	.121
	Black_Percent	.019	.004	.524	4.249	.000
3	(Constant)	5.570	.067		82.703	.000
	Black_Percent	.022	.004	.605	5.315	.000

a. Dependent Variable: lnviolent

MODEL PERFORMANCE

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.636[a]	.404	.366	.33810
2	.630[b]	.397	.372	.33646
3	.605[c]	.366	.353	.34157

a. Predictors: (Constant), Black_Percent, lnpoptot, Poverty_Rate

b. Predictors: (Constant), Black_Percent, Poverty_Rate

c. Predictors: (Constant), Black_Percent

d. Dependent Variable: lnviolent

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.640	3	1.213	10.615	.000[b]
	Residual	5.373	47	.114		
	Total	9.013	50			
2	Regression	3.579	2	1.790	15.807	.000[c]
	Residual	5.434	48	.113		
	Total	9.013	50			
3	Regression	3.296	1	3.296	28.254	.000[d]
	Residual	5.717	49	.117		
	Total	9.013	50			

a. Dependent Variable: lnviolent

b. Predictors: (Constant), Black_Percent, lnpoptot, Poverty_Rate

c. Predictors: (Constant), Black_Percent, Poverty_Rate

d. Predictors: (Constant), Black_Percent

Why Not Use Stepwise Regression All the Time?

- Analysis is always guided by a theoretical or conceptual framework

- Proper use
  - Exploratory analysis
  - Reduce the number of legitimate variables
  - Multicollinearity

REGRESSION BY REGION

Coefficients[a]

Region Name	Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
			B	Std. Error	Beta		
MW	1	(Constant)	3.422	1.067		3.207	.009
		lnpoptot	.151	.070	.564	2.161	.056
NE	1	(Constant)	.811	1.084		.748	.479
		lnpoptot	.314	.072	.855	4.364	.003
SO	1	(Constant)	8.734	1.550		5.634	.000
		lnpoptot	-.173	.101	-.393	-1.712	.106
WE	1	(Constant)	4.978	1.495		3.329	.008
		lnpoptot	.052	.101	.161	.516	.617

a. Dependent Variable: lnviolent

MODEL PERFORMANCE BY REGION

Model Summary[b]

Region_Name	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
MW	1	.564[a]	.318	.250	.22510
NE	1	.855[a]	.731	.693	.25063
SO	1	.393[a]	.155	.102	.38826
WE	1	.161[a]	.026	-.071	.39072

a. Predictors: (Constant), lnpoptot

b. Dependent Variable: lnviolent

ANOVA[a]

Region_Name	Model		Sum of Squares	df	Mean Square	F	Sig.
MW	1	Regression	.237	1	.237	4.669	.056[b]
		Residual	.507	10	.051		
		Total	.743	11			
NE	1	Regression	1.196	1	1.196	19.044	.003[b]
		Residual	.440	7	.063		
		Total	1.636	8			
SO	1	Regression	.442	1	.442	2.930	.106[b]
		Residual	2.412	16	.151		
		Total	2.854	17			
WE	1	Regression	.041	1	.041	.266	.617[b]
		Residual	1.527	10	.153		
		Total	1.567	11			

a. Dependent Variable: lnviolent

b. Predictors: (Constant), lnpoptot

APA TABLE

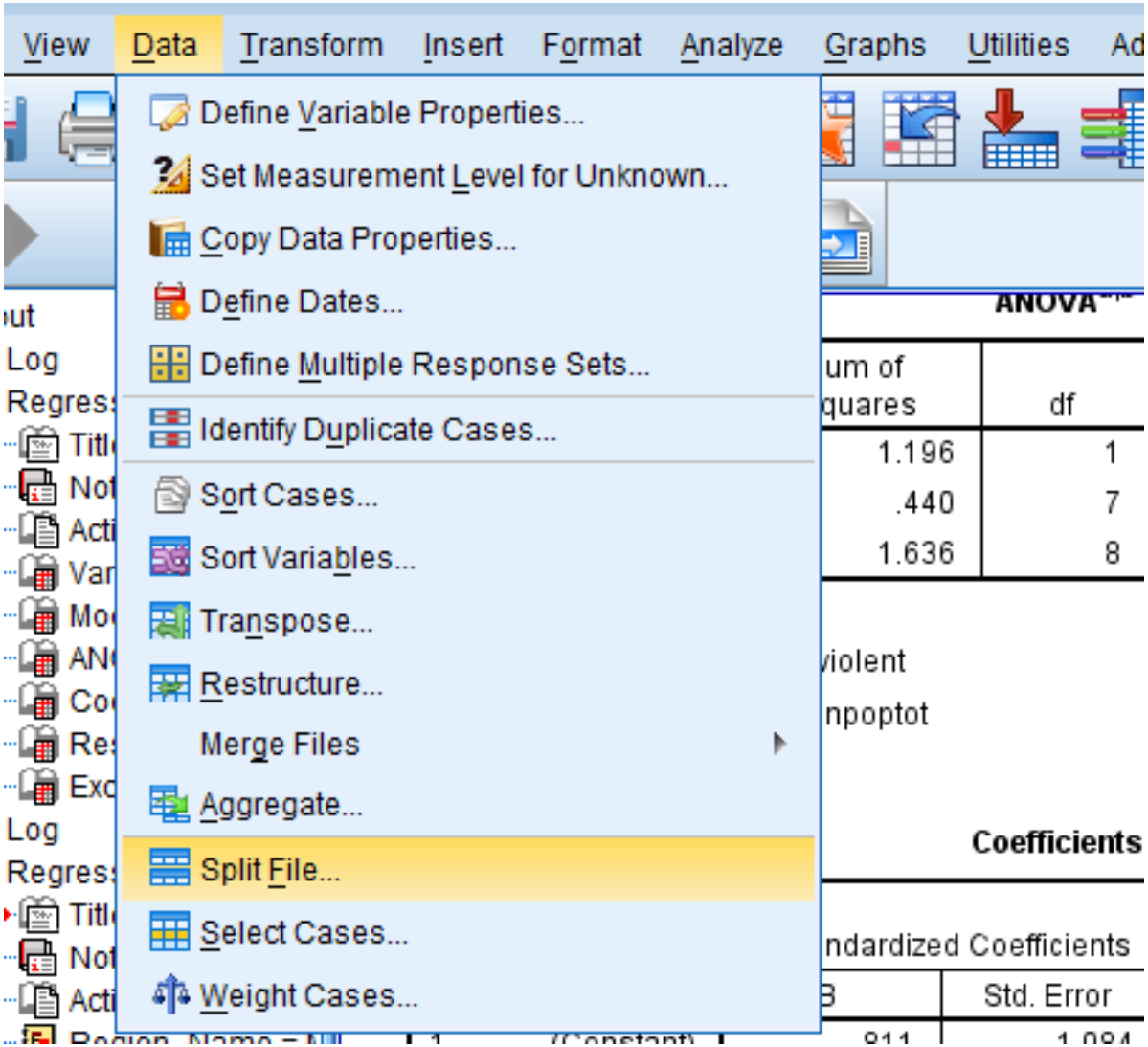
Table 3

Regression Analysis

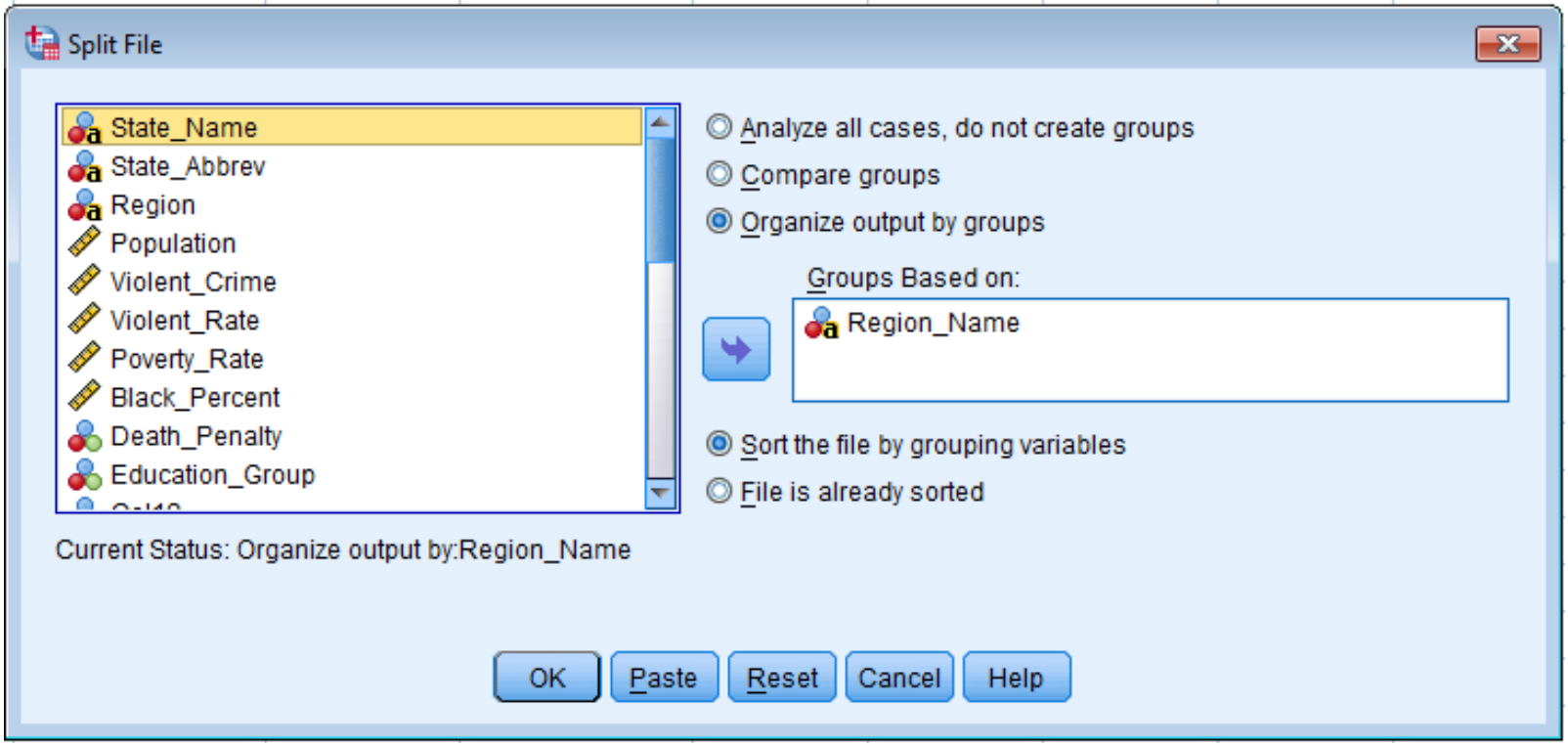
	All U.S.	Northeast	Midwest	South	West
Log of population	.14*	.31*	.15t	-.17	.05
Constant	3.64*	.81	3.42*	8.73*	4.98*
R2	.14*	.73*	.32	.16	.03

t p<.10 * p<.05

DATA > SPLIT FILE



SPLIT FILE COMMAND

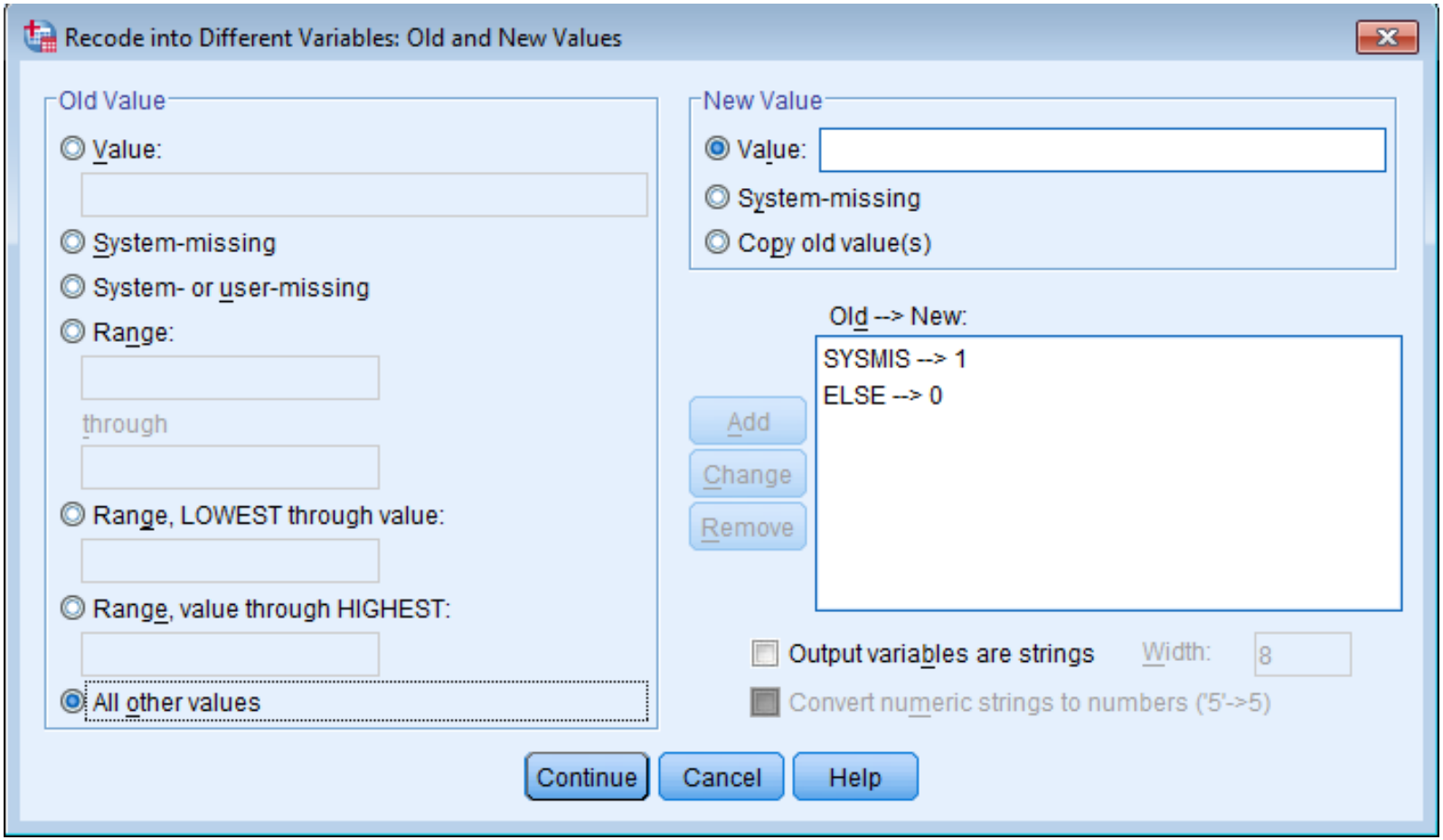


NON-RESPONSE ANALYSIS

U.S. COUNTY CRIME RATES

- Only 1735 of 3143 counties report a crime rate
- 55.2% response rate

IDENTIFYING MISSING DATA



REGRESSION FOR NON-RESPONDING COUNTIES

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	poptot	.000	.000	7.566	1	.006	1.000
	urbanp	-.007	.001	26.548	1	.000	.993
	popfemalep	.057	.018	10.664	1	.001	1.059
	whitep	.003	.002	2.162	1	.141	1.003
	Constant	-3.057	.873	12.277	1	.000	.047

a. Variable(s) entered on step 1: poptot, urbanp, popfemalep, whitep.

MULTIPLE IMPUTATION

1. Determine the relationship among using random samples of observations with no missing data (e.g., 10-20% random samples).
2. Do Step 1 several times, e.g., 5-10 times.
3. Estimate the value of missing data using each of the models. This gives multiple estimates per missing observation.
4. Insert the average of predicted estimates for missing data.
5. Do the regular analysis.

TO IMPUTE OR NOT IMPUTE

- Against:
  - You are making up data.
  - The results of the observations lacking missing data.

- For:
  - The bias caused by removing the observation may be worse than the harm caused by imputation. Sometimes imputation may actually be right.
  - The use of multiple imputations creates a range of values. If you believe in the theory of sampling, the average value of the estimates may be a truer estimate than any single estimate.

WHAT WE LEARNED

- Relying on regression output alone is not enough—we need to look beyond the fitted line and examine the residuals and influence statistics.
- Stepwise regression and regressions on subsamples can improve our analysis.
- Missing data can introduce biases into an analysis, some of which can be fixed by using a multiple imputation technique.