Stats 506: PS4 Analysis with Stata

Diana Liang
11/15/2019

Stata as a statistical tool requires a different methodology compared to its contempories such as R and Python. The main difference is the use of a single data set at a time. While this feature may seem like a restriction, Stata can still perform many of the same analysis techniques tauted by other programs. Below are three various analyses that demonstrate Stata's utility.

Part 1: Mouse-Tracking Mixed Models

Previously R was used to analyze mouse-tracking data that would add numeric understanding to mouse movements for an experimental task. The table below was reproduced using Stata instead of R using the same data.

Mixed Model Coefficients for Effect of Condition on Log of Curvature Measures

Curvature Measure	Coefficient	Standard Error
total_dist	-0.16 (-0.23, -0.09)	0.0370304
max_dev	-0.51 (-0.73, -0.29)	0.1128828
avg_dev	-0.65 (-0.91, -0.40)	0.1299681
auc	-0.50 (-0.72, -0.27)	0.1147798

Part 2: 2015 RECS Homes with Internet

Once again the 2015 RECS data was used to analyze the proportion of houses in each census division and area type that had internet access. The data was also ordered to find the division with the greatest disparity of proportions between the urban and rural areas.

Disparity of Prop of Homes with Internet between Urban and Rural Areas

Division	Disparity in Prop	Urban Prop	Rural Prop
Mountain South	18.52 (7.19, 29.84)	85.27 (81.32, 89.21)	66.75 (58.26, 75.24)
East South Central	9.33 (-1.40, 20.06)	78.36 (70.54, 86.18)	69.03 (63.50, 74.55)
West North Central	7.68 (-2.45, 17.81)	88.00 (84.63, 91.38)	80.33 (71.49, 89.16)
Mountain North	5.50 (-6.19, 17.19)	87.42 (81.99, 92.85)	81.93 (73.82, 90.03)
West South Central	5.10 (-2.30, 12.50)	81.61 (76.41, 86.80)	76.50 (72.12, 80.88)
Pacific	3.43 (-4.51, 11.37)	88.71 (86.17, 91.25)	85.28 (77.44, 93.12)

South Atlantic	3.26 (-3.54, 10.05)	85.30 (82.63, 87.96)	82.04 (76.28, 87.80)
New England	1.78 (-2.48, 6.04)	87.57 (82.50, 92.64)	85.79 (82.36, 89.22)
East North Central	0.04 (-5.30, 5.39)	86.25 (83.76, 88.74)	86.21 (81.64, 90.78)
Middle Atlantic	-1.95 (-9.09, 5.19)	89.34 (83.85, 94.82)	91.29 (85.31, 97.26)

As shown above, Mountain South had the greatest disparity of internet proportion between urban and rural areas. Almost all of the division, except South Atlantic, seemed to have greater proportions of internet access in urban areas compared to rural areas, but many of the confidence intervals contain 0 or negative proportions so that may not hold true.

Part 3: Logistic Models for Drinking Water

The last analysis deals with CDC National Health and Nutrition Examination Survey (NHANES) for 2005-2006. The focus was on how the day of water intake effected the likelyhood to drink water, specifically whether that day was during the weekday or the weekend. Other variables such as season, age, and economic status were controlled for to make the final analysis more true to the actual effect. The data was cleaned and reorganized in Stata to achieve the necessary conditions for a logistic model, the summaries of two are shown below.

Logistic Model

Only one of the two days was used for analysis to control for the possibility of two data points for each respondent, so a mixed model was not necessary. The odds ratios are listed below but the coefficients are also listed as a more direct way to compare the two models.

Summary Table for Logistics Model

Indep. Variable	Odds Ratio	Log Coefficient	Marginal Effect
weekday	1.13 (1.02, 1.24)	0.12 (0.02, 0.21)	0.02 (0.00, 0.04)
winter	0.90 (0.82, 0.99)	-0.10 (-0.20, -0.01)	-0.02 (-0.04, -0.00)
age	1.08 (1.04, 1.13)	0.08 (0.04, 0.12)	0.02 (0.02, 0.03)
age^2	0.95 (0.94, 0.96)	-0.05 (-0.07, -0.04)	-0.01 (-0.01, -0.01)
gender	1.33 (1.21, 1.46)	0.28 (0.19, 0.38)	0.06 (0.04, 0.07)
pir	1.10 (1.07, 1.13)	0.09 (0.06, 0.13)	0.02 (0.01, 0.02)

Mixed Logistics Model

This mixed model included both possible days for each respondant and used the respondant variable as a random intercept. Unfortunately, this model does not automatically output odds ratios, so the coefficients will be used as a way to compare the two models.

Summary Table for Logistics Model

Indep. Variable	Log Coefficient	Marginal Effect
weekday	0.16 (0.04, 0.28)	0.02 (0.00, 0.03)
day	0.16 (0.06, 0.26)	0.02 (0.01, 0.03)
winter	-0.08 (-0.22, 0.06)	-0.01 (-0.02, 0.01)
age	0.23 (0.16, 0.29)	0.03 (0.03, 0.03)
age^2	-0.08 (-0.10, -0.06)	-0.01 (-0.01, -0.01)
gender	0.55 (0.41, 0.69)	0.06 (0.04, 0.07)
pir	0.17 (0.12, 0.22)	0.02 (0.01, 0.02)

Looking at the logistic coefficients and marginal effects for weekday, the original logistic model and the mixed logistic model do not differ enough to say that they provide different conclusions. The confidence intervals of the corresponding coefficients definitely include the coefficients of the other model and almost overlap between the models.