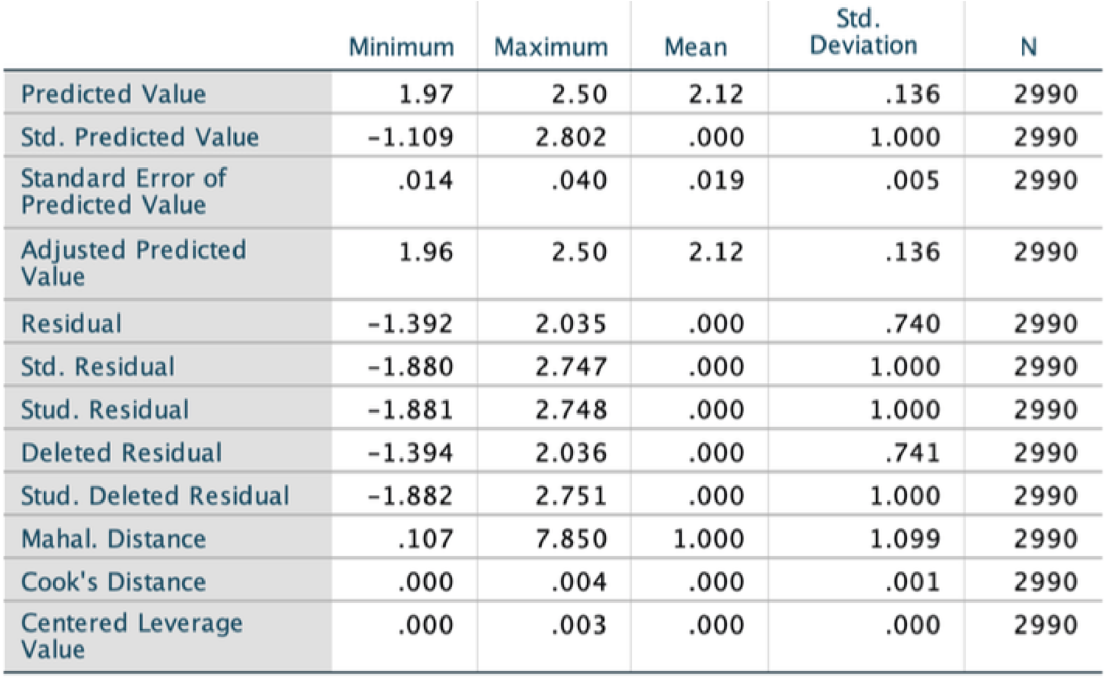
Homework Week 11 (Outliers, leverage, influence)

Simon Chen
Apr 12

Attention: *For the remainder of the assignments this semester, it is suggested that you work with the same dataset and group that you plan to use for your final project. This gives you an opportunity to get to know your data and to do preliminary analyses (with feedback).*

For this assignment, use the dataset and model you chose for Homework Week 9, and Week 10. (Or, alternatively, put another model together).

1) For your model, use the methods introduced in class to examine your data to determine if there are any **influence points**:
   (a) Calculate Cook's distance for each observation. Calculate the rule of thumb for large values. Which observations have large influence? Interpret this result.



	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1.97	2.50	2.12	.136	2990
Std. Predicted Value	-1.109	2.802	.000	1.000	2990
Standard Error of Predicted Value	.014	.040	.019	.005	2990
Adjusted Predicted Value	1.96	2.50	2.12	.136	2990
Residual	-1.392	2.035	.000	.740	2990
Std. Residual	-1.880	2.747	.000	1.000	2990
Stud. Residual	-1.881	2.748	.000	1.000	2990
Deleted Residual	-1.394	2.036	.000	.741	2990
Stud. Deleted Residual	-1.882	2.751	.000	1.000	2990
Mahal. Distance	.107	7.850	1.000	1.099	2990
Cook's Distance	.000	.004	.000	.001	2990
Centered Leverage Value	.000	.003	.000	.000	2990

Model: news clicks = $\alpha + \beta^*$ newsID+ εi

Cook's distance=4/2589=0.0015 The value of Di>0.0015 are considered possibly influential.

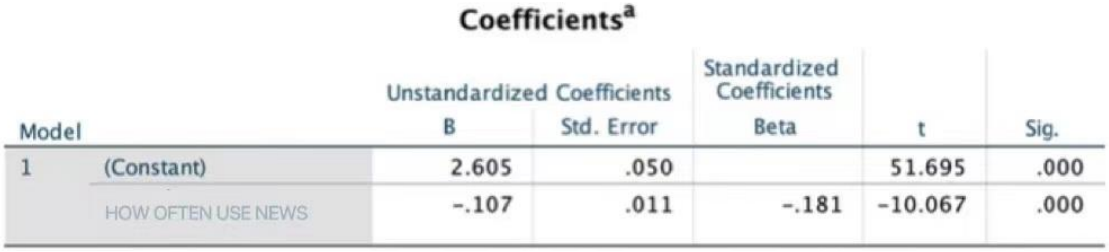
   (b) Pick a covariate of interest in your model (usually there are one or two you really are focused on). Calculate standardized DFBETAS for this covariate.

Calculate the rule of thumb. Which observations have large influence. Interpret this result.

DFBETAS=2/($n^{1/2}$)=0.037
|DFBETASij| ≥ 2/√(n) ,

An observation exerts influence on regression coefficient if |DFBETAS|>0.037

(c) Depending upon your question, choose *either* the observation with the largest Cook’s distance value or with the largest standardized DFBETAS value. Remove this from the dataset. Re-run the model you ran in #2. How do the results change?



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.605	.050		51.695	.000
	HOW OFTEN USE NEWS	-.107	.011	-.181	-10.067	.000

Removing it will have significant impact on the coefficient result of the model. B changed.

(d) For this observation, why do you think it is influential? (Relate this to leverage and residuals/outliers). What could you do to reduce it’s influence?

I think it is influential. The observation has both high leverage and high residual, so exerting influence on the regression analysis results. To reduce the influence, we can delete it.