ASSIGNMENT #6

EPsy 8252

This assignment is intended to have you build your understanding of using information criteria for model selection. *Do not include any R syntax or output unless it is specifically required in the question.* Any graphics you include should be resized so that they do not take up more room than necessary and all should have an appropriate caption. Any equations should be appropriately typeset within the document.There are 11 points possible for the assignment (each question is worth one point unless otherwise noted).

The data which are presented in Chatterjee and Hadi (2012) and originally stem from McDonald and Schwing (1973) are from a study examining the effects of climate, socioeconomic condition, and pollution on mortality. There are 15 variables, which are:

Table 1
Description of variables (n = 60)

Variable	Description
X1	Mean annual precipitation (in inches)
X2	Mean January temperature (in degrees F)
X3	Mean July temperature (in degrees F)
X4	Percent of population over 65 years
X5	Population per household
X6	Median school years completed
X7	Percent of housing units that are sound
X8	Population per square mile
X9	Percent of nonwhite population
X10	Percent employment in white-collar jobs
X11	Percent of families with income under $3000
X12	Relative pollution potential of hydrocarbons
X13	Relative pollution potential of oxides of nitrogen
X14	Relative pollution potential of sulfur dioxide
X15	Percent relative humidity
Y	Total age-adjusted mortality from all causes

Model Fitting

Previous research and data fitting have suggested three potential models. Fit each of the models and use them to answer the questions on the assignment. There are no questions to answer in this section, you only need to fit the three models.

- **Model 1** includes a subset of eight predictors: X1, X2, X3, X4, X5, X6, X9, and X14.
- **Model 2** includes a subset of 14 predictors: all but X12.
- **Model 3** includes a subset of 14 predictors: all but X13.

Table of Model Summaries

Create a table that includes the following summaries for each of the three models. Remember, models are conventionally presented in columns and summaries in rows.

1. The estimate for error variance (where $\hat{\sigma}_\epsilon^2 = \frac{\sum \epsilon}{n}$)
2. The AIC value (where $\text{AIC} = n\ln(\hat{\sigma}_\epsilon^2) + 2K$)
3. The AICc value (where $\text{AICc} = \text{AIC} + \frac{2K(K+1}{n-K-1}$)
4. The Δ_i value
5. The w_i value

Model Selection

6. Which model will you select as the "best" model of the subset? Explain.
7. Interpret your selected model's weight of evidence (w_i).
8. Compute the evidence ratio ($\frac{w_i}{w_j}$, where w_i is the weight of evidence for your selected model) for the selected model versus each of the other two models in the subset. Interpret them **(2pts.)**

Multimodel Inference

9. Use model averaging to compute the coefficient estimate and the variance estimate for the predictors in your selected model. Present these estimates in a table. **(2pts.)**

References

- Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example* (5th ed.). New York: Wiley.
- McDonald, G. C., & Schwing, R. C. (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics, 15*, 463–481.