

Lab 05

This lab will be done in Python. For the lab to be complete, all questions must be answered.

Section 1: Multiple Regression

A healthcare non-profit is interested in understanding the impact of statewide demographic information and cigarette prices on cigarette sales. They feel that if any of these factors are significantly related to cigarette sales, it will help them figure out which areas should be targeted with anti-smoking messaging. They provided you with the `cigarette_sales.csv` dataset. Information about this data is in the table below.

Variable	Description
Age	Median age of persons living in a state
HS	% of people >25 years of age in a state who had completed high school
Income	Per capita personal income for a state in dollars
Black	% of black race living in a state
Female	% of females living in a state
Price	Weighted average price in cents of a pack of cigarettes in a state
Sales	Number of packs of cigarettes sold in a state per person

[This link](#) will be helpful for doing this analysis. However, as this is multiple regression, you will need to make sure the X array contains all the predictor variables.

- 1) Answer the following:
 - a) What is the outcome?
 - b) What are the predictors they want to understand the impact of?
 - c) What is the hypothesis?
- 2) Exploratory data analysis
 - a) Look at a few rows of the data to understand how it is structured.
 - b) Generate some summary statistics using `describe()`.
 - c) Look at the distributions and scatterplots of the data. A convenient function for doing this is `pairplot()` in Seaborn.
 - d) Do any of these variables look like they might violate the normality regression assumption or be correlated with other variables? Explain.
 - e) Generate a plot to check if there are outliers in the outcome. What do you see?
- 3) Multiple regression
 - a) Conduct a multiple regression analysis.
 - b) Are any of the variables significant?
 - c) Interpret the intercept and any statistically significant coefficients (i.e. what is their meaning in relation to sales?)

Section 2: Detecting Assumption Violations

Using the same data set and regression results from the prior section, do the following:

- 1) Collinearity
 - a) Compute the VIF for each covariate and explain what the results mean. Use [this link](#).
 - b) Compute all the pairwise correlations between the variables. [This link](#) shows 3 ways to do this. Just use `corr()`.
 - c) Remove the 3 variables with the highest p-values. Refit the model. How have the p-values for the other variables changed? Did R^2 change by much?
- 2) Model Fit
 - a) What does R^2 tell you about the fit of the second model?
 - b) As noted in the video on MLE, AIC is another measure of fit. Which model has the lowest AIC value (lowest is best)?
- 3) Outliers
 - a) Do a leverage plot. Are there influential outliers? Again, [this resource](#) is helpful.
- 4) Linearity & constant variance
 - a) Generate a predicted vs standardized residual plot. The `resid_studentized_internal` in `statsmodels` are the standardized residuals. The link above shows how to obtain them. Does the data meet the linearity assumption?
- 5) Normality
 - a) Do a Q-Q Plot of the residuals. Are the residuals normally distributed?