STAT-S 352 Problem Set 10

Upload your answers through the Assignments tab on Canvas by 11:59 pm,

Monday, November 10, 2025.

Important Note: Answer all questions and include R code when necessary. In general in this course, give explanations and/or working for all answers unless otherwise stated. Show your work for full credit.

Reminder: As a student at IU, you are expected to uphold and maintain professional and academic honesty and integrity. Academic integrity violations include: cheating, fabrication, plagiarism, interference, violation of course rules, and facilitating academic dishonesty. When you submit an assignment with your name on it, you are signifying that the work contained therein is yours, unless otherwise cited or referenced. Any ideas or materials taken from another source must be fully acknowledged.

1 Teaching Evaluation

In class we looked at the data set at `https://www.openintro.org/book/statdata/evals.csv` and tried to learn if "age", "average beauty score" and "rank" were good predictors in predicting professors' teaching evaluation scores (at least in Texas.) In this problem, we will look at another variable, "gender." The variable `gender` in `evals.csv` is coded as `female` (195 of the instructors) or `male` (268 of them.)

1. On average (and without considering any other variables), do female or male instructors get higher evaluations? How big is the difference, and is it statistically significant? Perform an appropriate hypothesis test and give a P-value.

   Hint: This is one review question for S350 (Chapter 11: Two sample location problem).

2. The relationship between perceived beauty and evaluation score might be different for female and male instructors. Separate the data into two subsets: `evalsFemale` for female instructors and `evalsMale` for male instructors. For each subset, fit a linear regression model to predict `score` from `bty_avg`. Plot the data with beauty on the x-axis and score on the y-axis, then add *both* regression lines to the plot, clearly distinguishing which line is which. Describe the difference between the two lines.
   Hints:

   (a) Review how to use the `subset()` function to separate a dataset.

   (b) The code `abline(3.9, 0.06, col = "orange")` adds an orange line with intercept 3.9 and slope 0.06 to an existing base R plot.

   (c) Another option is using `abline(lm(...), col)` to add a regression line to an existing R plot.

2 Rust

The file `rust.txt` contains county-level data on 2016 deaths from drug-related causes in the East North Central region, consisting of the states Ohio, Michigan, Indiana, Illinois, and Wisconsin, sometimes called the "Rust Belt." The variables are:

- `County`
- `Deaths`: number of drug-related deaths in 2016
- `Population` in 2016
- `Area`: land area of county in square miles
- `PctWhite` and `PctBlack`: percentage of the population who are white and black respectively
- `Income`: median household income
- `Trump`: proportion of 2016 Presidential election votes that went to Trump
- `Obama`: proportion of 2012 Presidential election votes that went to Obama

Our goal will be to fit a model that explains variation in drug-related deaths in these counties in 2016. The response variable will be the **rate of drug-related deaths per 100,000 population** (the standard way to measure this.) We'll take the counties as our units, so there's no need to weight the model by population.

1. Create a data frame and name it `rust2` with the following variables, and draw a pairs plot of the numeric variables:

   - `DeathRate`: drug-related deaths per 100,000 population
     (Formula: DeathRate=Deaths/Population*100,000)
   - `Density`: population divided by area
   - `PctWhite`
   - `logIncome`: natural log of income
   - `Swing`: Trump − (1− Obama). This measures (approximately) the change between the 2012 and 2016 elections: a value of 0.05 means a swing of five percentage points towards the Republicans.
   - `county`

   Notes:

   - Four out of six variables listed above do not come directly from the data `rust`. You need to create those variables following the given instructions.
   - Do not use `county` as a predictor.

2. Fit a model to predict `DeathRate` with `Swing`, `Density`, and the `Swing:Density` interaction as predictors; call this `q1model`. Find the AIC and compare it to the AIC of the model without the interaction. For the interaction model, write down the regression equations for predicting `DeathRate` from `Swing` for the following values of `Density`:

(a) `Density` = 55 (similar to Randolph County, the lowest density county in Indiana)

(b) `Density` = 369 (similar to Monroe County, IN)

(c) `Density` = 2375 (similar to Marion County, IN, i.e. Indianapolis)

Draw a graph of the data with these three regression lines on it, clearly distinguishing which line is which.

3. We wish to predict `DeathRate` using the other numeric variables in `rust2`, with no interactions. Write down the equations and AICs for the following models:

   (a) Forward selection starting from an empty model

   (b) Backward selection starting from a full model with all predictors

   (c) Both-ways stepwise selection starting from an empty model

   (d) Both-ways stepwise selection starting from a full model

   Do all these methods result in the same model? Write down the equations and find the AIC of the model(s) found by this process. (Note that the AIC given by the `step()` function differs from the AIC given by the `AIC()` function by a constant; either is fine but be consistent about which one you use.)