Investigation of a dataset: Cleaning, exploration and modeling

Using what you've learned all semester, perform data cleaning, EDA, and regression modeling on a dataset as detailed below in Parts 1 and 2. Please upload your R scripts (can be R Markdown along with its html if you'd like) with all code that accomplishes your cleaning, EDA, and modeling. Additionally, submit a final report of your important findings and conclusions as you would present them to your PI or hospital administrator. These will include your most informative plots, statistical tests and models that concisely tell the story of this dataset. Your final report can be R Markdown along with its html output, OR Word along with its pdf version.

Your summaries and interpretations of your findings are very important - your employer will count on you to concisely relate the story this data tells. However, your final report need not be verbose. You should use well-labeled plots to convey much of the information.

Your grade will be based on the process you go through to clean, explore, model, and analyze this dataset. You should demonstrate that the data science skills you've acquired to reason and communicate about health data.

Information on the variables is at the bottom of this assignment.

Part 1:

Perform cleaning of the dataset, correcting variable types or orders of levels as needed, and handling missing information appropriately. Perform EDA, investigating the story the data tells. In particular, investigate which features relate to birthweight of babies.

Be sure to include summary statistics, correlations, statistical tests, and visualizations of distributions of all variables. Also include bivariate or trivariate relationships between important variables, particularly as related to birth weight.

When performing EDA and statistical tests, be sure to follow Motulsky's advice on visualizing as well as calculating descriptive and inferential and statistics. Look beyond statistical significance to assess scientific significance (effect size matters). State accurately what you infer from p-value or confidence level, from power calculations and statistical tests that you perform.

Part 2:

Fit linear and logistic regression models to this daataset. Evaluate each model and use best practices to refine your linear and logistic regression models to arrive at a final interpretable model for each (linear and logistic). For each of your two final models, select one important categorical variable and one important numeric variable, stating what the fitted coefficient reveals about the impact of that predictor variable on the response variable in that model.

For your linear models, fit birth weight as a function of predictor variables. Evaluate your models according to best practices, and discuss your final interpretable model, including how you arrived at it. State what you think your final model reveals about factors that relate to birth weight, and what evidence you have for your claims.

For the logistic model, fit the dichotomous low birth weight variable as a function of predictor variables. As with the regression model, evaluate your models according to best practices, and discuss your final model, including how you arrived at it. State what you think your final model reveals about factors that relate to low birth weight, and what evidence you have for your claims.

Dataset variables (1000 births):

dad_age
    age of father of baby (years)

mom_age
    age of mother of baby (years)

maturity
    classify mother as of advanced maternal age or not (advanced / younger)

len_preg
    length of pregnancy (weeks)

is_premie
    classify baby as either premature or full-term (premie / fullterm)

num_visits
    number of visits to hospital during pregnancy

marital
    marital status of mother at time of birth (married / unmarried

mom_wt_gain
: mother's weight gain during pregnancy (pounds)

bwt
: birth weight of baby (pounds)

low_bwt
: classify baby as either low birthweight or not (low / notlow)

sex
: sex of baby (female / male)

smoke
: smoking status of mother (smoker / nonsmoker)

mom_white
: classify mother as either white or not (white / nonwhite)

mom_age_level
: age level of mother of baby (teens, early20s, late20s, early30s, 35+)