Stat 331 Final Project

People

2018-12-1

Discussion

Ultimately, it was determined that out of the models which we had considered, the model which had the greatest predictive power used the following covariates: mother's age, mother's body-mass index, smoke, mother's height, mother's ethnicity, interaction between mother's age and gestation, and interaction between parity and gestation. As determined in the previous section, the model does not seem to violate any of the regression assumptions. There was a point with great influence corresponding to an overweight 38 year old woman. However it would not be wise to exclude it as an "outlier" since there is no justification for why the model would not be able to predict the birth weight of 38-year old mothers.

As discussed in the model selection section, the covariate with the smallest p-value, and hence the most significant (in the presence of the other covariates) was the mother's age. It is worth noting that, as mentioned in the previous section, for model 1, the point with the highest influence was that with the highest age. It is also very interesting to note that both the gestation-age and the gestation-parity interactions were significant in this model. Both interaction covariates had positive coefficients. Nonetheless, further study would be required to determine the exact extent and nature of these interactions.

Consider two women of the same age, ethnicity, height, weight, parity, and gestation, but one of them smokes now while the other stopped smoking. Then, according to this model, the weight of the baby born to the woman who currently smokes is expected to be lower than the weight of the baby born to the woman who has stopped smoking (since all other factors have been held constant). However, most interestingly, this is true irrespective of when the mother stopped smoking (the coefficient + intercept is positive for both "used to smoke" and "smoked until pregnancy").

It appears as if the model also suggests that all other factors held constant, a Mexican mother is more likely to have a baby of greater weight than any other ethnicity. However, this is more likely due to the fact that out of the 1236 mothers in the dataset, only less than 50 were Mexican.

It should also be noted that the BMI is a significant covariate and the coefficient is positive. Thus, advice for prospective parents to reduce the probability of having an underweight baby would be to ensure that the mother has a healthy body-mass index, and to also stop smoking. The coefficients of the respective covariates suggest that it is better to quit smoking before pregnancy, however quitting at any point is still helpful.

Future Direction

As noted earlier, a significant portion of the father's data was missing. Due to this, the models did not use information from the father. However, this is a serious shortcoming since a child's genetic information is determined by both father and mother. Should this investigation be continued, data which is not missing the height/weight of fathers should be collected. Strategies used to reduce non-response bias could be employed as well (e.g. re-surveying a subset of the non-response group).

There were two socio-economic factors in the variable set: ethnicity, and income. It was very surprising that income did not have great predictive power (and hence discarded in the refined model 2). This could be due to the tendency that an individual has to lie about their income if they either make significantly less or significantly more than the average person. Changing the manner in which the question is phrased might reduce the measurement error. It would also be interesting to examine other biological and socio-economic interaction effects