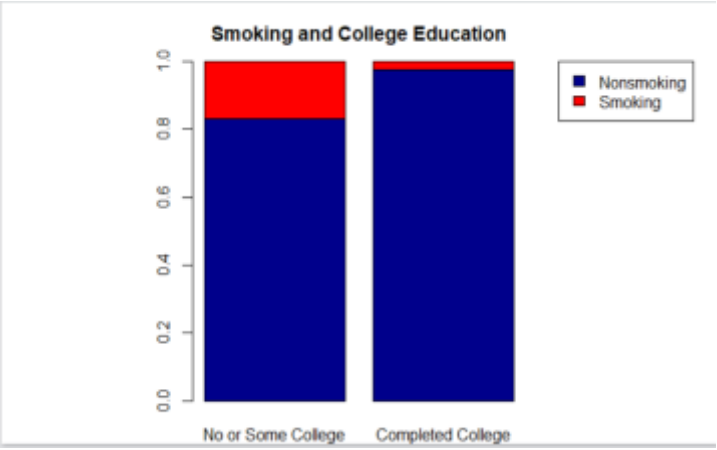
Team Case 1

Section B, Team 9: Vatsal Sanghavi, Annabelle Nguyen, Alper Sayiner, and Sophia Rowland

1. Before running any model, pick any two variables and attempt to show an (interesting) relation via visualization. This requires one to formulate a question, and to communicate clearly a conclusion based on data visualization. Ideally this would suggest ways to act on the issue. (For example, during pregnancy, does the amount of smoking seem to diminish with education level? This can provide guidance on where to place educational ads regarding smoking during pregnancies.)

During pregnancy, does the amount of smoking diminish when mothers have finished college? The stacked bar plot belows shows that there is a smaller proportion of smokers among mothers who have finished college than among mothers who have not finished college. This would provide guidance on placing educational ads regarding smoking during pregnancy in media and areas frequented by women who have not finished college.



2. Next consider the 10 dummy (binary) variables in Exhibit 1. Test independence among all 45 combinations. First using the traditional 0.05 rule. Second controlling the false discovery rate (conservative and FDR with q=0.001). Any discrepancies? If so, which one do you believe got it wrong this time?

Yes, there are discrepancies between the traditional method and the other methods for the boy and black combination and the boy and married combination as well as a discrepancy between the tradition method and the fdr method for the black and ed.smcol combination.

3. In your opinion, can any of the variables provided in Exhibit 1 help to predict birthweight? Since opinions do not provide strong arguments, provide a simple evidence based on data.

First, we find the correlation matrix for the data. Looking at the correlation between weight and every other variable, we concluded that there is a weak relationship between weight and every

other variable. Out of the 14 variables, mother's weight gain has the strongest correlation with baby's weight, of 0.21. We also checked if the correlations were significant, and the p-value of 2.2 x (10^-16) for each correlation proves that they are significant.

4. Run a multiple regression with 14 of the variables described in Exhibit 1 (all except for id, birmon, tri.none, and novisit). Which variables are statistically significant? Apply the 0.05 cut-off rule and also control for false discovery rate (conservative and FDR with q=0.001).

With the conservative approach, the statistically significant variables are black, married, boy, tri1, tri2, tri3, ed.hs, ed.smcol, ed.col, mom.age, smoke, cigsper, m.wtgain, mom.age2

With the FDR with q=0.001 approach the statistically significant variables are black, married, boy, tri1, tri2, tri3, ed.smcol, ed.col, mom.age, smoke, cigsper, m.wtgain, mom.age2

The following questions do not require additional data analysis calculations.
5. In rural health care, primary care clinics cannot rely on expensive screening tests. To increase screening capabilities of detecting risky pregnancies early (end of first trimester), consider using the model in Question 4 to forecast birthweight since it is inexpensive. Discuss issues with implementation and choice variables.

- Mother's weight gain variable captures her weight gain over the entire pregnancy. At the end of the first trimester, we don't know the mother's weight gain variable yet.
- At the end of the first trimester, we don't know if a mother is going to visit the doctor in the second and third trimesters. Therefore, we don't know the tri2 and tri3 variables yet.
- The correlations between baby's weight and other variables are very low. The highest correlation is between the baby's weight and mother's weight gain is the highest correlation, and we don't know the mother's weight gain as of the end of the first trimester.
- Even though the p-values of the coefficient estimates of the variables are significant, the low coefficient estimate of each variable makes the model unreliable.
- The R-squared value and adjusted R-squared value for the whole model is very low (0.11) so the model is not accurate in predicting a baby's weight at birth.

6. Consider more broadly the various departments in a hospital which forecasts demand for services (with potentially additional information) that is used to build simulation software to more precisely set the number of nurses and other employees. Briefly discuss potential issues regarding deployment. Propose a suggestion that addresses at least one of them (even if only partially).

- The model may predict that there are periods of high staff need and periods of low staff need, but you can't ethically lay off staff and rehire them again based on staff need.
- The model is based on past data and may not account for future events that haven't occurred yet, such as a protest, social event, natural disaster, etc.
    - Suggestion: if the hospital knows that such an event will occur, they can add additional staff to the model's prediction.
- The model could underpredict the health service demand and this could result in dramatic consequences.
    - Since being able to serve all emergency cases is more critical for an hospital than overhead costs, the hospitals should determine a safety level of minimum employees that would cover most of the standard deviation of the demand for health services. This is similar to the safety stock approach in inventory management.