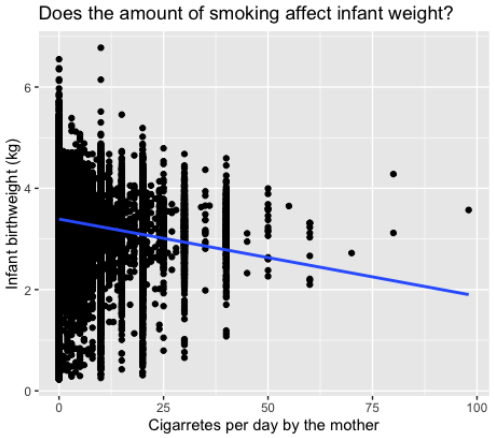
Data Science for Business Team Case 1
Team 18

Yinan Chen, Caryl Alexis Cohen, Tirth Pravin Gala,
Yuhe(Tiffany) Jin, Chenjie (Angelina) Sun

Question 1

The variables we chose are “Cigarettes per day by the mother” and “the infant birth weights”. There appears to be a discernible negative correlation between the number of cigarettes smoked by the mother per day and the birth weight of the infant. Specifically, as the maternal daily cigarette consumption increases, there is a corresponding decrease in the infant’s birth weight. This observed negative correlation aligns with well-established findings within the field of medical research and public health. Extensive literature has consistently reported adverse outcomes associated with maternal smoking during pregnancy, among which a reduction in birth weight is prominently documented. Infants born to mothers who engage in smoking tend to exhibit lower birth weights in comparison to those born to non-smoking mothers.



Question 2

Using the traditional 0.05 rule, we find that there are seven pairs that have p-values above 0.05. This means that we fail to reject the null hypothesis, suggesting that the variables in each pair are independent, in favor of the alternate hypothesis, which states the variables in these pairs are not independent. By using Bonferroni correction, we found that all nine pairs that have p-value greater than 0.05 will be known as independent as well. When we apply the Bonferroni method, which accounts for multiple tests by adjusting the significance level, all nine pairs appear to be independent. While the traditional 0.05 significance test indicates that two pairs are associated with “boy”, when applying the stricter Bonferroni correction, all nine pairs appear to be independent. This fits our biological understanding: a child’s gender is determined at conception. Specifically, when the sperm provides a Y chromosome, the result is a boy, and with an X chromosome, it's a girl. Thus, factors like maternal smoking during pregnancy are unlikely to influence a child‘s gender as it's essentially determined by the father's genetic contribution. Two pairs where discrepancies are noted. These are the “boy and black” and “boy and married” combinations. Bonferroni correction suggests independence for these values within each pair, while the 0.05 rule suggests dependence. The 0.05 rule likely got it wrong as the Bonferroni coefficient is much smaller and a more conservative estimate, with the value of 0.00111($\alpha = 0.05/45 = 0.00111$)

- traditional 0.05 rule for each pair:

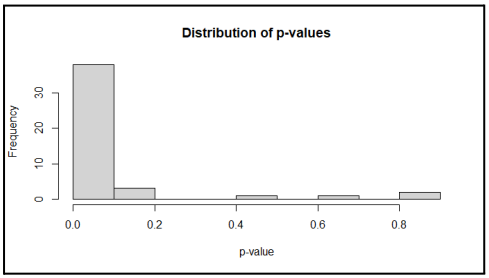
```
> for (i in 1:45){if(pvals[i] > 0.05){print(paste(pvals[i],ListLabels[i]))}}
[1] "0.605458247847503 boy and tri1"
[1] "0.124597786938161 boy and tri2"
[1] "0.121553586558693 boy and tri3"
[1] "0.476237095860618 boy and ed.hs"
[1] "0.184792787077862 boy and ed.smcol"
[1] "0.895930238864032 boy and ed.col"
[1] "0.812928337254288 boy and smoke"
```

The following 7 pairs have p-values larger than 0.05, which leads to the failure to reject the null hypothesis.

- Bonferroni correction:

```
> for (i in 1:45){if(pvals[i] > 0.00111){print(paste(pvals[i],ListLabels[i]))}}
[1] "0.605458247847503 boy and tri1"
[1] "0.124597786938161 boy and tri2"
[1] "0.121553586558693 boy and tri3"
[1] "0.00580415332696215 boy and black"
[1] "0.00564400710903288 boy and married"
[1] "0.476237095860618 boy and ed.hs"
[1] "0.184792787077862 boy and ed.smcol"
[1] "0.895930238864032 boy and ed.col"
[1] "0.812928337254288 boy and smoke"
```

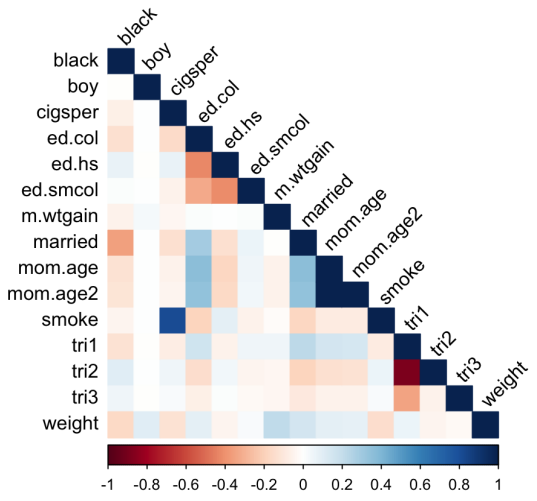
The following 9 pairs have p-values, larger than 0.00111, which leads to the failure to reject the null hypothesis.



Graph shows that the p-values are not distributed uniformly.

Question 3

To predict an infant's weight, we get an insight into the correlation of each variable with the infant's weight. With *weight* as the dependent variable, we conduct the correlation matrix with the correlation coefficients between *weight* and other variables and visualize the correlation into the heat map as below:



	weight
weight	1.00000000
black	-0.16271604
married	0.15219814
boy	0.10285159
tri1	0.06302941
tri2	-0.04442780
tri3	-0.02147732
ed.hs	-0.04666745
ed.smcol	0.02805578
ed.col	0.09586540
mom.age	0.09602286
smoke	-0.14285097
cigsper	-0.12594258
m.wtgain	0.21096743
mom.age2	0.08848520

From the heatmap and correlation matrix, we could find that:

1. Mother's weight gain during pregnancy is positively correlated with the infant's weight. This correlation with weight is significantly stronger compared to other variables. Following that, we have the infant being boy or not and the mother's marital status showing a strong and positive correlation.
2. Mother's being black or not, smoking or not and number of cigarettes smoked per day by mother showing a negative and strong correlation with infant's birth weight.

Therefore, we could conclude that the mother's weight gain during pregnancy is the best predictor of the infant birthweight from the dataset.

Question 4

By running the multiple linear regression model, we could find that all independent variables have its p-value smaller than 0.05 from the output. It indicates that all the variables are statistically significant in influencing an infant's birthweight.

The p-value for each independent variable also falls below the Bonferroni correction threshold of 0.003571429 (**$\alpha = 0.05/14 = 0.003571429$**). It leads to the consistent conclusion with the standard 0.05 cut-off rule about the variables' statistical significance. There is no discrepancy between the standard 0.05 cut-off rule and Bonferroni correction method in this analysis.

```
Call:
lm(formula = weight ~ black + married + boy + tri1 + tri2 + tri3 +
    ed.hs + ed.smcol + ed.col + mom.age + smoke + cigsper + m.wtgain +
    mom.age2, data = DATA)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5303 -0.2915  0.0213  0.3352  3.8260

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.281e+00  3.050e-02  74.769  < 2e-16 ***
black       -1.997e-01  3.573e-03 -55.892  < 2e-16 ***
married      6.064e-02  3.250e-03  18.657  < 2e-16 ***
boy          1.089e-01  2.418e-03  45.021  < 2e-16 ***
tri1         1.848e-01  1.358e-02  13.607  < 2e-16 ***
tri2         1.966e-01  1.388e-02  14.163  < 2e-16 ***
tri3         2.146e-01  1.570e-02  13.670  < 2e-16 ***
ed.hs        1.547e-02  3.756e-03   4.117 3.84e-05 ***
ed.smcol     3.180e-02  4.185e-03   7.599 2.99e-14 ***
ed.col       3.798e-02  4.498e-03   8.443  < 2e-16 ***
mom.age      3.638e-02  1.996e-03  18.226  < 2e-16 ***
smoke       -1.682e-01  6.288e-03 -26.748  < 2e-16 ***
cigsper     -3.711e-03  4.468e-04  -8.307  < 2e-16 ***
m.wtgain     8.900e-03  9.442e-05  94.260  < 2e-16 ***
mom.age2    -5.473e-04  3.458e-05 -15.828  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.538 on 198362 degrees of freedom
Multiple R-squared:  0.1101,    Adjusted R-squared:   0.11
F-statistic:  1753 on 14 and 198362 DF,  p-value: < 2.2e-16
```

Question 5

a) The model does not account for maternal nutrition and diet, which can be significant determinants of birth weight, especially in rural settings where malnutrition might be more prevalent.
b) Mothers in rural areas might not have regular prenatal check-ups due to the distance from clinics or other logistical challenges. Therefore, variables that require regular updates or monitoring might not be as reliable.
c) Mothers in rural areas might not have the same level of education or literacy as those in urban settings. This can lead to potential miscommunication or misunderstanding when they are asked about certain variables, such as their smoking habits or other relevant behaviors.
d) Mother's age squared does not seem necessary when already factoring in the mother's age.
e) The model is missing some key variables, such as the medical history of mother or father (not related to smoking), which can help determine the birthweight of the infants.