Does Mother's Smoking During Pregnancy Affect Baby's Birthweight?

David Hua; Gordon Wang

May 12, 2018

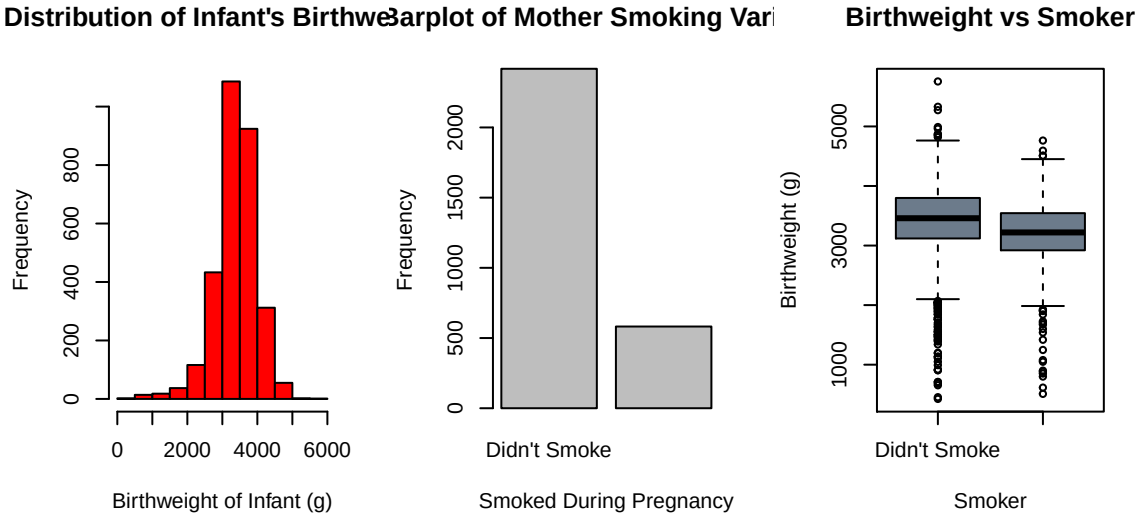
Introduction

Doctors have worried about low infant birthweight for many years because it is an important indicator of poorer overall health. It is generally believed that a mother's actions during pregnancy can greatly affect infant birthweight. We are especially interested in whether mother's smoking during pregancy affects a baby's birthweight after controlling for other factors such as mother's attributes and other behaviors during pregnancy.

Data Analysis

The dataset we analyzed contains 3000 observations and 12 different variables. The two main variables we are focusing on are birthweight, the birth weight of the infant in grams, and smoker, an indicator equal to 1 if the mother smoked during pregnancy and 0 otherwise.

We begin by doing simple univariate EDA on the data. We find that the distribution of birthweight is slightly left-skewed with a long left tail, indicating some low values are possible outliers. We also find that many more mothers, 2418, did not smoke during pregancy, versus 582 who did. In addition, we perform bivariate EDA, and speculate from the boxplot result that mothers' smoking has a negative relationship with birthweight.



First, we will regress birthweight on the main x variable of interest (smoker), without controlling for other factors.

```
##
## Call:
## lm(formula = birthweight ~ smoker, data = bw.data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3007.06  -313.06    26.94   366.94  2322.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3432.06      11.87 289.115   <2e-16 ***
## smoker       -253.23      26.95  -9.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 583.7 on 2998 degrees of freedom
## Multiple R-squared:  0.0286, Adjusted R-squared:  0.02828
## F-statistic: 88.28 on 1 and 2998 DF,  p-value: < 2.2e-16
```

Right away, we can see that smoker seems to have a large negative effect of 253.23 on birthweight on average, since its slope coefficient has a t-value of -9.396 with a p-value of less than 2e-16, indicating significance at 5% and 10% levels. However, we need to account for other factors, evident from the model's small multiple R-squared value of 0.0286.

We then run a regression controlling for other factors. Note that we will not include tripre0 to avoid the dummy variable trap caused by perfect collinearity between tripre0 and tripre1, tripre2, and tripre3. Also, because of redundancy between alcohol and drinks, we choose to not include alcohol since drinks gives strictly more information.

From the regression summary, we find that under a 5% signifiance level, based on the t-test statistics, the influential predictors of birthweight are nprevist, tripre2, tripre3, smoker, and unmarried.

Interestingly, tripre1 does not have a significant effect. However, we find high correlation among the tripre indicators, since tripre1, tripre2, and tripre3 have high variance inflation factors of 18.62, 14.92, and 4.23, respectively. To decide whether to drop or keep these three variables, we run a F-test for their joint significance. For our test, we have $H_0 : \beta_{tripre1} = 0$, $\beta_{tripre2} = 0$, $\beta_{tripre3} = 0$ and our H_A is that H_0 is false. The F-statistic's distribution under the null is the F random variable with (q = 3, n-k-1 = 290) degrees of freedom. Our result is a F-statistic of 5.2812 and p-value of 0.001244, so we reject H_0 at the 5% significance level, and conclude that tripre1, tripre2, and tripre3 are jointly significant in the model and shouldn't be dropped.

We then try to improve our model by dropping the individually insignificant variables of educ, age, and drinks. However, we need to justify their exclusion with another F-test for the joint significance of the variables. Our test has $H_0 : \beta_{educ} = 0$, $\beta_{age} = 0$, $\beta_{drinks} = 0$, and $H_A : H_0$ is false. The F-statistic's distribution under the null is the F random variable with (q = 3, n-k-1 = 290) degrees of freedom. With a result of a F-statistic of 0.3128 and p-value of 0.8162, we fail to reject H_0 at both 5% and 10% significance levels. Thus, the three variables aren't jointly significant in predicting birthweight and can be dropped.

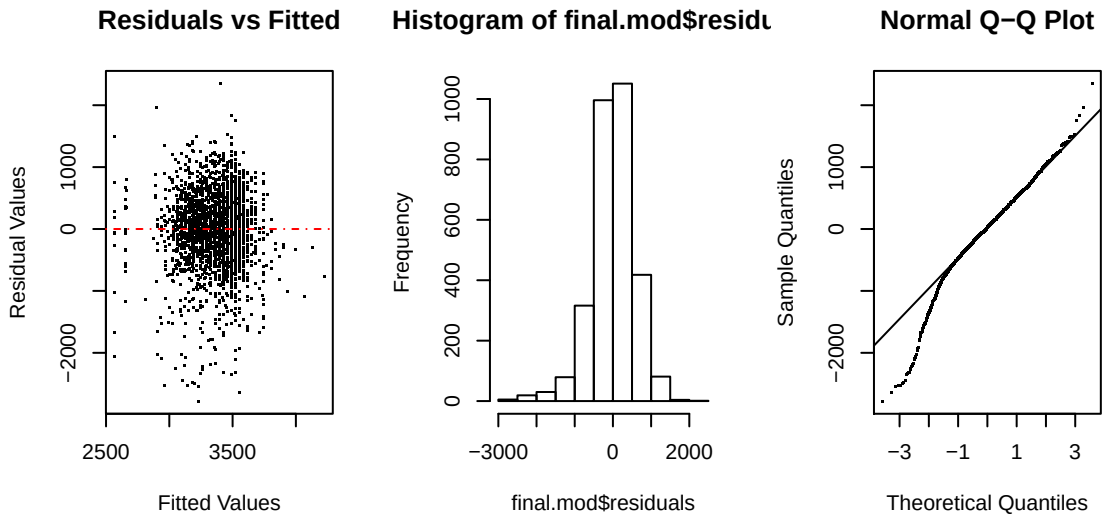
```
## Analysis of Variance Table
##
## Model 1: birthweight ~ smoker + unmarried + nprevist + tripre1 + tripre2 +
##     tripre3
## Model 2: birthweight ~ (nprevist + alcohol + tripre1 + tripre2 + tripre3 +
##     tripre0 + smoker + unmarried + educ + age + drinks) - tripre0 -
##     alcohol
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1   2993 953269695
## 2   2990 952970652  3    299043 0.3128 0.8162
```

In order to fine-tune our model's specification, we also check for potential interactions, specifically those between between smoker and the other variables, in order to better understand its effect on birthweight.

So we fit an unrestricted model with all the different interactions of smoker and find the only individually significant interaction is smoker*unmarried, with t-value of 2.652 and p-value of 0.00804. We want to check if we can exclude all the insignificant interactions from the unrestricted model using a F-test. For our test, H_0 : $\delta_{smoker*nprevist} = 0$, $\delta_{smoker*tripre1} = 0$, $\delta_{smoker*tripre2} = 0$, $\delta_{smoker*tripre3} = 0$, and H_A: H_0 is not true. The F-statistic's distribution under the null is the F random variable with (q = 4, n-k-1 = 288) degrees of freedom. We get a F-statistic of 0.9227 with p-value of 0.4496, so we fail to reject H_0 at the 5% and 10% significance levels, and conclude that other interactions are jointly insignificant and can be dropped.

Now, we have our final selected model. We include an interaction term between smoker and unmarried, which will affect our interpretation. This interaction means that the coefficients to smoker and unmarried variables individually no longer represent unique effects. The effect of one on smoking will depend on the value of the other.

```
##
## Call:
## lm(formula = birthweight ~ smoker + unmarried + nprevist + tripre1 +
##     tripre2 + tripre3 + smoker * unmarried, data = bw.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2779.42  -306.40    24.56   358.46  2348.83
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2891.86     105.16  27.499  < 2e-16 ***
## smoker            -230.93      33.50  -6.893 6.64e-12 ***
## unmarried         -241.37      31.20  -7.736 1.40e-14 ***
## nprevist            32.00       3.39   9.438  < 2e-16 ***
## tripre1            213.57     111.86   1.909  0.05632 .
## tripre2            275.72     110.28   2.500  0.01246 *
## tripre3            388.52     118.49   3.279  0.00105 **
## smoker:unmarried   145.98      55.95   2.609  0.00912 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 563.8 on 2992 degrees of freedom
## Multiple R-squared:  0.09558,    Adjusted R-squared:  0.09346
## F-statistic: 45.17 on 7 and 2992 DF,  p-value: < 2.2e-16
```



Looking at the residuals vs fitted plot for our final model, we see that regression assumptions of residuals having a mean of zero and being uncorrelated are met, but that the variance of the residuals might not be constant. Also, a histogram of the residuals and Normal Q-Q Plot shows the assumption of normality is violated, possibly due to influential outliers.

Lastly, we want to run a heteroskedasticity test, in order to see if the final model in fact does violate the homoskedasticity assumption.

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 46.12918, Df = 1, p = 1.1071e-11
```

Running the NCV test, we obtain a Chisquare value of 46.129 and p-value of 1.107e-11, indicating significance at the 5% level, so we reject H_0 of constant variance of residuals and accept H_A of heteroskedasticity. This confirms our suspicion from our earlier diagnostics.

Conclusions

Based on our final model, we find that smoking has a significant impact on infant birthweight. Specifically, the t-statistic for smoking in our final model is extremely negative with a coefficient of around -231. Note that since we also have an interaction term between smoker and unmarried, we interpret the coefficient to smoker as the expected decrease in birthweight (-231 grams) of a smoker vs. nonsmoker if unmarried = 0 (meaning the are unmarried).

The interaction term between smoker and unmarried is also worth mentioning. We find that there is a significant postive coefficient to this interaction which means that the effect of smoking on birthweight is different for unmarried/married people (or vice versa).

Lastly, we must be careful with our final conclusions. since we find that there is heteroskedasticity in our model. Because homoskedasticity is a requirement for our regression, we should take our result with a grain of salt.