STAT 215B Assignment 1

Xiaowei Zeng

March 12, 2024

Abstract

The main job in this assignment is to investigate the connection between maternal smoking and infant health and is conducted **mainly** following the guideline in the instruction document. The first section introduces the dataset used for this analysis and some exploratory data analysis results. The second and the third section states the methods, results and findings in analyzing Claim 1 and Claim 2 stated in (Deborah Nolan, 2000). The significance level for the whole analysis is set as 0.05.

1 Dataset

The dataset forming the basis of this analysis is (a subset of) the Child Health and Development Studies (CHDS), a large survey on all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, California. It contains 1236 observations (live male births). The variables recorded for each birth are given in the following table:

Table 1: Variables description

Name	Description
bwt	Newborn weight (rounded to the nearest ounce)
gestation	Length of the pregnancy (days)
parity	Whether the baby is (1) or is not (0) the first-born
age	Age of the mother at conception (years)
height	Mother's height (inches)
weight	Mother's weight (pounds)
smoke	Whether the mother smokes (1) or not (0)

The variables gestation, age, height, weight, and smoke have some missing values. But the code for a missing value is not exactly the same across the variables. Using the `summary` command, I find that **the maximum of these variables are unusual**, e.g. 9 for smoke, 99 for age and height, 999 for gestation and weight. Then I replace all occurrences of the missingness code with R's missing value code, `NA`. The description statistics of the dataset after replacing missing values is shown in Table 2.

The correlation of covariates is of great importance in getting an initial feel for the data. As shown in Figure 1, only the infant born weight and gestation, height and weight have

Table 2: Summary of the dataset

Characteristic	**N = 1,236**	**Missing**
bwt	120 (109, 131)	0
gestation	280 (272, 288)	13
parity		0
not_first_born	921 (75%)	
first_born	315 (25%)	
age	26 (23, 31)	2
height	64 (62, 66)	22
weight	125 (115, 139)	36
smoke		10
not_smoke	742 (61%)	
smoke	484 (39%)	

a relatively strong positive relationship (> 0.40). Notice that only **numerical variables** are used here. To capture both linear and nonlinear relationship between the variables, I use **Spearman method** to calculate correlation coefficients rather than Pearson method.

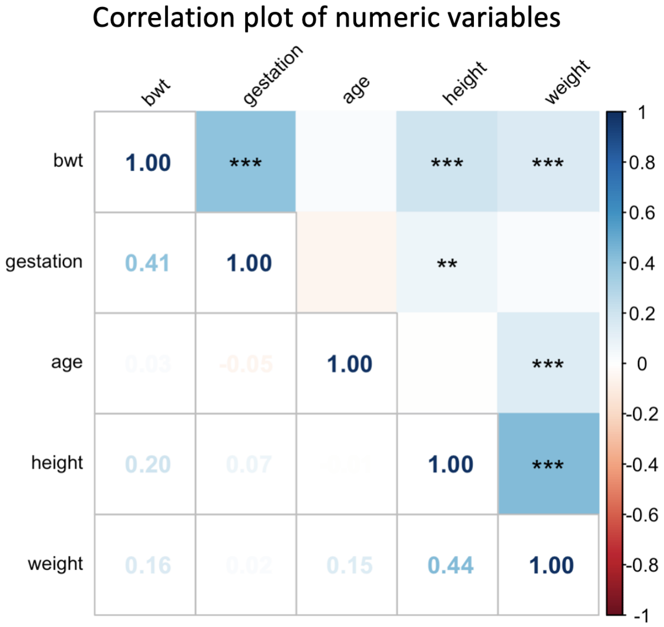

Figure 1: Spearman Correlation Heatmap

Since the two claims we intend to analyze are all related to the **smoke** variable, it is necessary to scratch the relationship between smoke and other covariates (bwt, age, height, gestation, height, weight). Since smoke is a categorical variable, we cannot calculate the correlation coefficients, but we can conduct some hypothesis tests.

Shapiro-Wilk normality tests are first employed on these covariates (categorized by smoke variable) to test normality of data, whose results can inform us of whether to use the parametric two sample t-test (which assumes normality) or the nonparametric Wilcoxon rank sum test (which don't have any assumption for distribution of data). The *p*-values show that only the infant born weight with mother who smokes pass the normality test. Therefore, **we should reject the normality assumption** and **Wilcoxon rank sum**

tests are then employed (Table 3).

Table 3: Results of Wilcoxon Rank Sum test

variable	p-value
bwt	<0.001
age	0.02
gestation	<0.001
height	0.56
weight	0.02

Except the height variable, the smoke variable has strong association with all other variables. It exactly makes sense because height is almost fixed for adults and determined by genes and growing environments, which has little to do with whether the mother smokes or not.

2 Analyzing Claim 1

Claim 1 states that mothers who smoke deliver premature babies more often than mothers who do not.

2.1 Difference in Gestation Distribution

Some graphical comparisons of the gestation distribution are made for smoking mothers to the gestation distribution of non-smoking mothers. Figure 2 shows an obvious left location shift for the gestation of babies whose mothers smoke. Figure 3 shows that the median (and the quantile range) of the gestation of babies whose mothers smoke is smaller. I don't use the whole data to draw the boxplot because there are many outliers that negatively influence its looking. **These two figures all support Claim 1.**

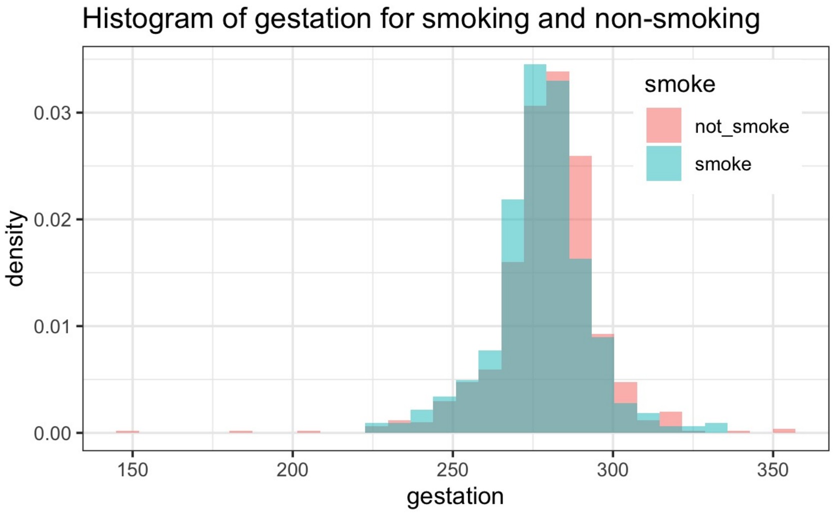

Figure 2: Histogram of Gestation

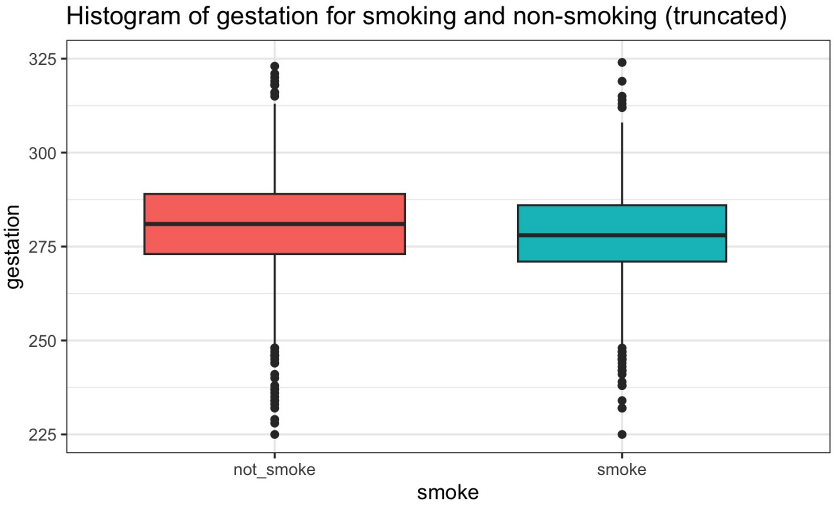

Figure 3: Boxplot of Gestation

2.2 Premature Indicator: Tabular Comparison

A full-term pregnancy is defined by the medical community as lasting 40 weeks. A premature birth is defined as occurring prior to the 37th week of gestation. I introduce a two-level factor variable, indicating whether or not each baby was born prematurely, to the babies data. The `cut` command is used to categorize the gestation variable by the cutpoint of $36 \times 7 = 252$.

Then I carry out a relevant tabular comparison of distributions using the factor premature and the factor smoke (Table 4). We can see that the proportion of the premature babies in the smoking group is 0.24% higher than that of the premature babies in the non-smoking group. It seems that there exist a difference, but we cannot know whether it's significant or not. **This tabular analysis partly supports Claim 1.**

Table 4: Tabular Comparison

	premature	mature
smoke	26 (5.42%)	454 (94.58%)
not_smoke	38 (5.18%)	695 (94.82%)

2.3 Premature Indicator: Graphical Comparison

I make a barplot to enable the comparison in the previous section to be carried out visually (Figure 4). **No obvious evidence in this barplot implies Claim 1.**

2.4 Premature Indicator: Contigency Table Analysis

In this section, I use the same table in Section 2.2 to carry out hypothesis tests of the null hypothesis that smoking and non-smoking mothers have the same rate of premature delivery. **Common association tests for contigency table are Fisher's Exact Test and Chi-square Test.**

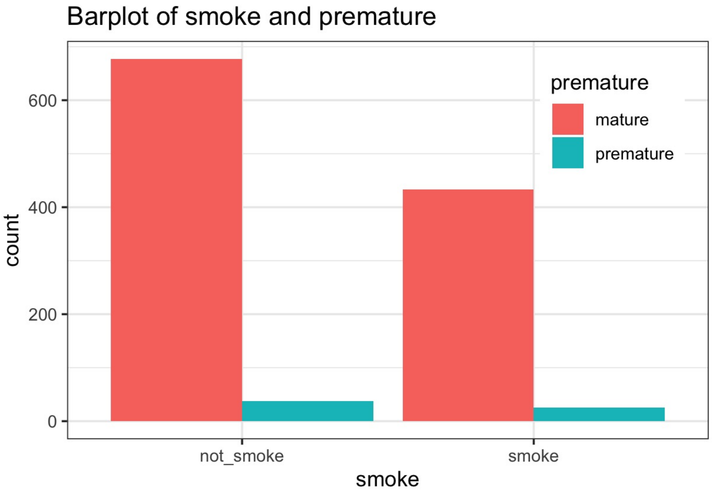

Figure 4: Barplot of Smoke and Premature

The results of both Fisher's Exact Test and Chi-square Test **show no evidence that smoke and premature are correlated** (Table 5).

Table 5: Results of Fisher's Exact Test and Chi-square Test

Test	p-value
Fisher's exact test	0.90
Chi-square test	0.96

From this view of point, **Claim 1 is protested.**

2.5 Gestation Variable: Two Sample Test

Gestation is a continuous variable, and thus a two sample test should be conducted. According to the guideline in the instruction document, the null hypothesis is that the overall average gestation time is the same for smoking mothers and non-smoking mothers; **the alternative hypothesis is that the overall average gestation time is shorter** for smoking mothers, compared to non-smoking mothers.

The same as the analysis in the end of Section 1, we cannot assume normality for the gestation data, and the **one-sided** Wilcoxon rank sum test is employed. The p-value of this test is $0.0004 < 0.05$, indicating that we should reject the null hypothesis, which is **a strong evidence for supporting Claim 1.**

2.6 Extra Analyses

2.6.1 Cluster Analysis

In the previous sections, we take only the smoke variable into consideration. However, grouping the mothers into smoking and non-smoking does not ensure that they are similar among other features. **The other features may also influence the comparison result for gestation because smoke variable is strongly correlated to other**

covariates as shown in the end of Section 1. Therefore, I try to use an unsupervised technique, cluster analysis, to **divide the mothers into many groups where within each group the mothers are as similar as possible.**

In the cluster analysis, I take into account a mother's age, height, weight and parity when placing her in a group. The number of cluster (k) should be chosen as 3 because it is evident that $k = 3$ identifies the clear inflection point in the scree plot (Figure 5).

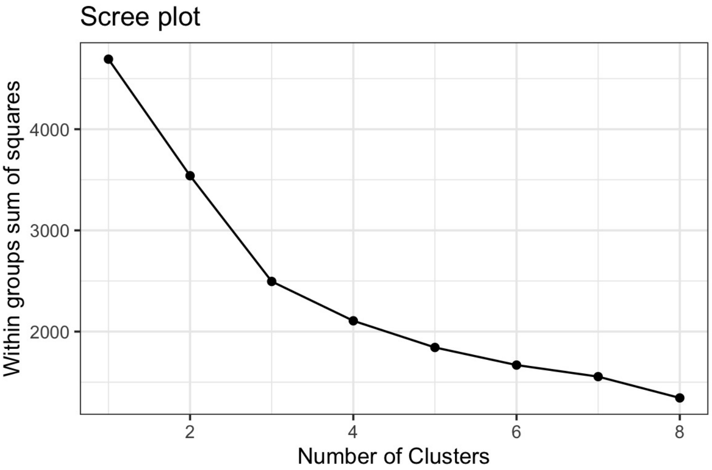

Figure 5: Scree Plot

The features of each cluster (mean) are shown in Table 6.

Table 6: Charateristics of Each Cluster

cluster	gestation	age	height	weight
1	279	29.8	65.9	146
2	278	27.3	62.4	116
3	281	23.6	64.2	124

Within each group, I compare the gestation to smokers and nonsmokers on both visual and quantitative scale. In each cluster, Figure 6 shows an obvious pattern that mothers who smoke have shorter gestation days. However, Wilcoxon rank sum test does not give significant results to all the clusters (Table 7), indicating that in the groups that have similar features, gestation time does not make much difference. This may be attributed to the correlation between the smoke variable and the other demographics of mothers.

Table 7: Results of Wilcoxon Rank Sum Test

Cluster	p-value	# of obs.
1	0.16	400
2	0.001	474
3	0.12	300

The conclusions led from the plot and the hypothesis test are inconsistent. **So Claim 1 is partly supported and partly protested.**

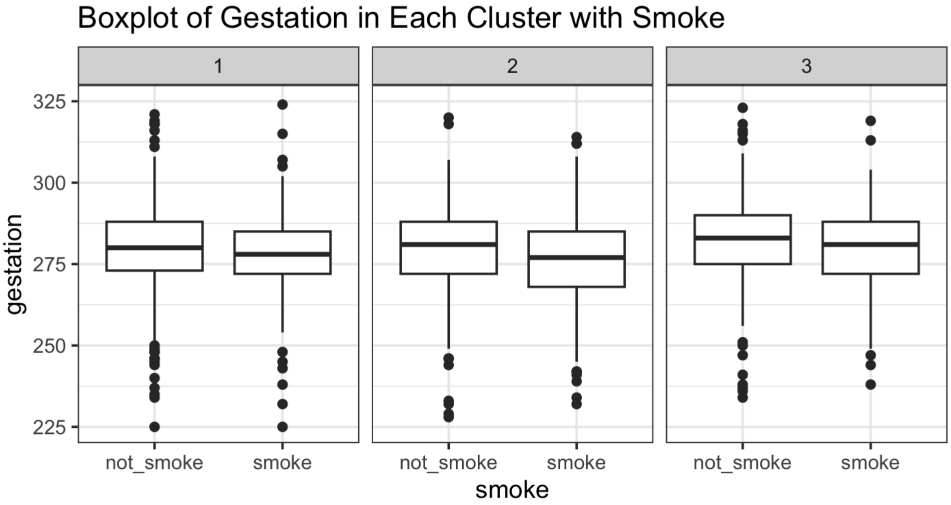

Figure 6: Boxplot of Gestation in Each Cluster with Smoke

2.6.2 Regression Analysis

Gestation may be affected by other covariates. I first conduct a regression analysis with gestation as the outcome, smoke, parity, age, height and weight as the independent variables. Then I use stepwise to remove variables (age, weight).

Table 8 shows that the smoke variable is significant (p-value$= 0.04 < 0.05$), which **supports Claim 1**.

Table 8: Results of Linear Regression Model

	Estimate	Std. Error	t value	Pr($\|t\|$)
(Intercept)	251.4709	11.7928	21.324	<0.001
smoke smoke	-1.9913	0.9519	-2.092	0.037
parity first_born	2.8144	1.0567	2.663	0.008
height	0.432	0.1841	2.346	0.019

The reason why we don't use born weight as the independent variable is that it is a collider, affected by both gestation and other covariates, which should not be added in the model.

2.7 Findings

Most of the results using **gestation** as the variable of main interest **support Claim 1**, while most of the results using **premature** as the variable of main interst **protest Claim 1**. This can be attributed to the information loss occurring in the binary transformation of variable gestation.

3 Analyzing Claim 2

Claim 2 states: Cigarette smoking has a stronger relationship to infant birth weight than several other relevant covariates. The only other covariates available in the data for us to check are parity, age, height and weight.

3.1 Univariate: Checking Parity, Age, Height, Weight

Notice that **the guideline in the instruction document does not include analyzing the age variable.** Since Claim 2 also includes the age variable, I will conduct the same analysis with age as parity, height and weight.

The analyzing streamline in this section is as follows:

1. If the variable of great interest (one of parity, age, height and weight) is continuous (a.k.a. numerical), first categorize the variable into a binary indicator.
   - Divide the mothers into "midage" ($>$ median age) and "young" ($\leq$ median age).
   - Divide the mothers into "tall" ($>$ median height) and "short" ($\leq$ median height).
   - Divide the mothers into "heavy" ($>$ median weight) and "light" ($\leq$ median weight)
2. Conduct the Shapiro Wilk normality test to test normality. If the data pass the test, then use t-test (which has greater power for normal data); if the data does not pass the test, use Wilcoxon rank sum test. **Compare the p-values, Hodges-Lehmann estimation and the 95% confidence intervals (CI) of the tests for smoke and the other variable (one of parity, age, height and weight) and make informal inference.** However, since the upper bound of 95% CI usually goes to infinity, I only compare the p-values and Hodges-Lehmann estimation in the following paragraph.
3. Conduct a quantitative test (t-test) to test the significance of Claim 2.

   Let $\theta_1 = |\beta_1| - |\beta_2|$, where β_1 is the coefficient for smoke variable, β_2 is the coefficient for the variable to compared. However, the true β_1 and β_2 are unobserved. It is hard to derive the distribution of $|\beta_1| - |\beta_2|$ if we do not know the signs of them. Therefore, I relevel the categorical factors to make sure that $\hat{\beta}_1 > 0, \hat{\beta}_2 > 0$, according to both prior knowledge and the OLS estimator $\hat{\beta}_1, \hat{\beta}_2$:

   - Prior knowledge: We expect that the first-born babies have lower birth weight; babies with mid-age mothers have higher birth weight; babies with tall mothers have higher birth weight; babies with heavy mothers have higher birth weight. Relevel X_2 to $\tilde{X}_2$ using the lower weight level as the reference level.
   - OLS estimator: Set smoking as the reference level of the `smoke` variable, and thus $\hat{\beta}_1 > 0$. Lead the regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. If $\hat{\beta}_2 > 0$, let $\tilde{X}_2 = X_2$; if $\hat{\beta}_2 < 0$, let $\tilde{X}_2 = 1 - X_2$.

Now we can use the one-sided t-test in the linear regression model to test

$$H_0 : \theta_1 = 0 \text{ v.s. } H_1 : \theta_1 > 0.$$

Then, the regression model turns to be

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 \tilde{X}_2 + \varepsilon \\ &= \beta_0 + (\theta_1 + \beta_2) X_1 + \beta_2 \tilde{X}_2 + \varepsilon \\ &= \beta_0 + \theta_1 X_1 + \beta_2 (X_1 + \tilde{X}_2) + \varepsilon \end{aligned}$$

The `lm` command will automatically give the p-value of the two-sided t-test. Since the sample size is very large ($n = 1174$), **we can use normal approximation to t distribution and thus can obtain the one-sided p-value by simply taking a half.**

To summarize, I will compare the difference in the average birth-weight between smoking and non-smoking mothers to the difference in the average birth-weight between first-borns and non-first-borns; between babies born to midage versus young women; between babies born to tall versus short women; between babies born to heavy versus light women.

At the significance level of 0.05, both Wilcoxon rank sum test and t-test show that the difference in average birth weight between smoking and non-smoking mothers is the most significant (Table 9). Parity and age are the least influenced on infant born weight, and then the weight, height, and finally smoke. Therefore, **Claim 2 is strongly supported.**

Table 9: Results of Wilcoxon rank sum test and t-test

	Hodges-Lehmann Est.	p-value for Wilcox test	$\hat{\theta}_1$	p-value for t-test
smoke	9.00	<0.001		
parity	2.00	0.015	5.87	<0.001
age	2.00	0.019	5.18	<0.001
height	6.00	<0.001	1.38	0.033
weight	5.00	<0.001	1.82	0.010

3.2 Visual Comparison: Multi-Panel Plots

Then I make a multi-panel histogram and a multi-panel boxplot which allow the comparisons to be carried out visually for whole distributions (Figure 7, Figure 8).

To ease visual comparison, I put the y-axes across the panels in exactly the same range. It seems that when we take a direct look at the difference between the distributions, we can lead to a similar conclusion as the aforementioned quantitative hypothesis test, that both smoke and height has the most significant impact on infant born weight, while smoke is more influenced than age, parity and weight. We can hardly tell that smoke is more influenced than height, which is consistent to the largest p-value in the t-test between smoke and height.

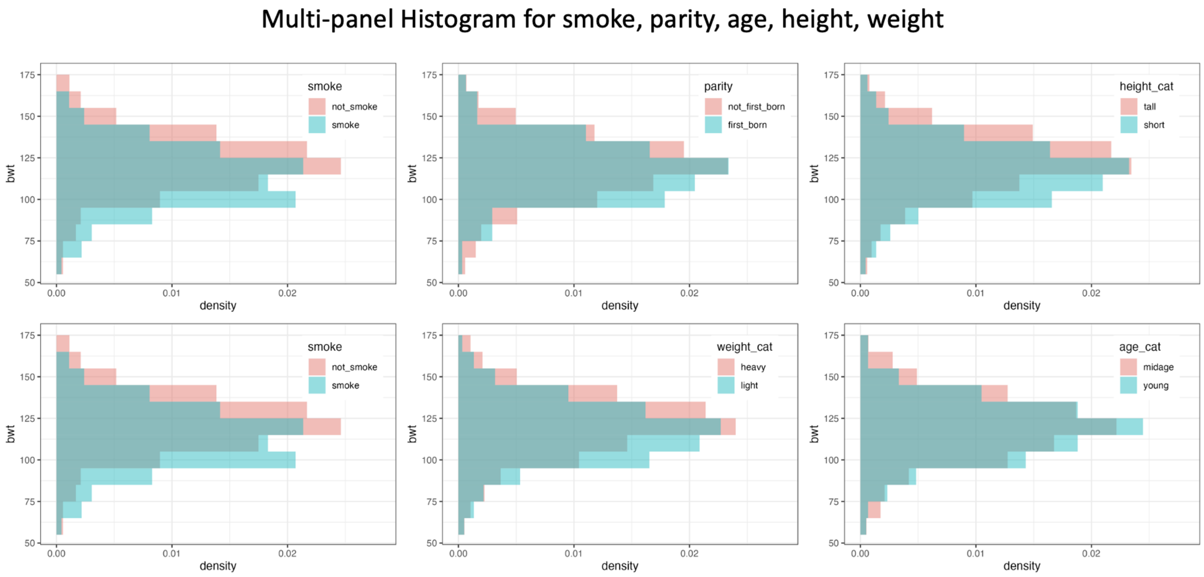

Figure 7: Multi-panel Histogram

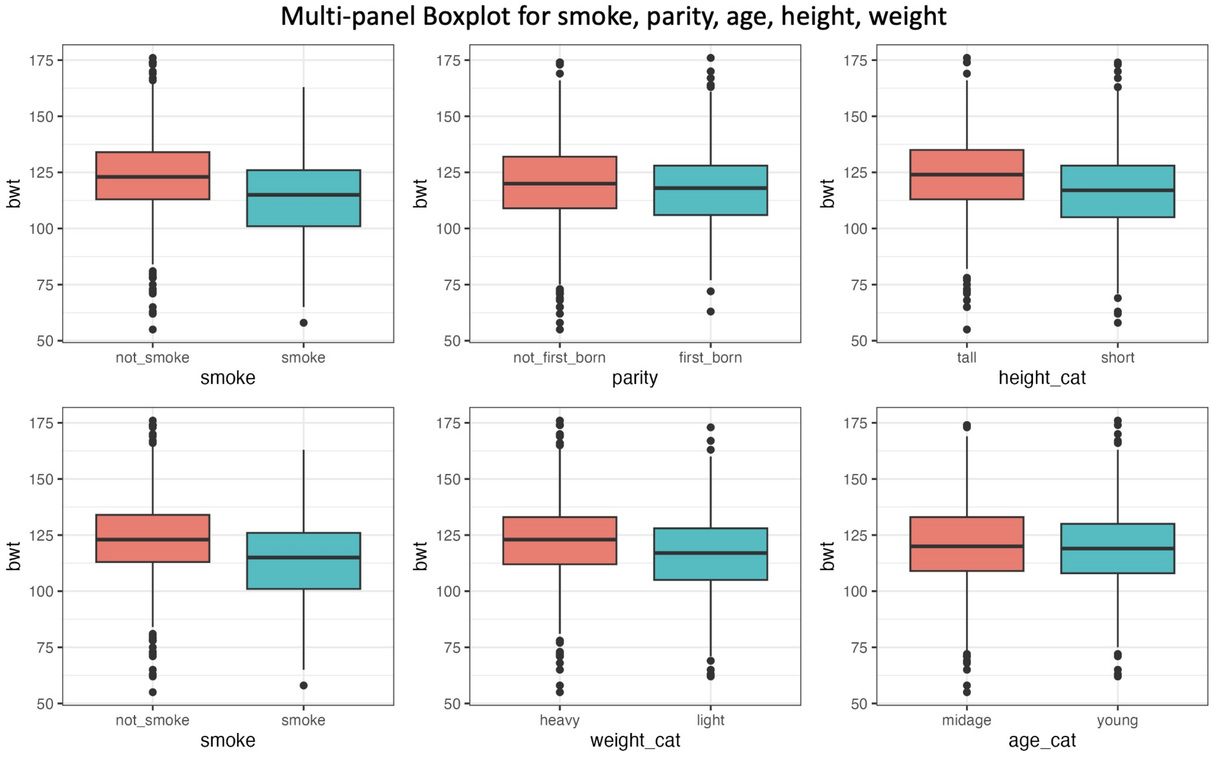

Figure 8: Multi-panel Boxplot

3.3 Multiple Regression

I first fit a multiple linear regression of birth-weight against age, height, weight, and parity (but not smoking status). Actually, we should still introduce age into the linear regression, but the p-value of age in this regression is 0.96, which is really close to 1. So it is very obvious that **the relationship between age and bwt is not as significant as that between smoke and bwt.**

Therefore, I will follow the guideline in the instruction document, which is, first fit a multiple linear regression of birth-weight against height, weight, and parity (but not smoking status); then fit a second regression with smoking status added to the model. The results are shown in Table 10.

Table 10: Results of Multiple Regression Model

		Model 1			Model 2	
	Estimate	Std. Error	Pr($\|t\|$)	Estimate	Std. Error	Pr($\|t\|$)
(Intercept)	31.19	13.57	0.022	30.86	13.13	0.019
height	1.25	0.23	<0.001	1.35	0.22	<0.001
weight	0.07	0.03	0.016	0.05	0.03	0.075
parity first_born	-1.83	1.20	0.126	-2.04	1.16	0.079
smoke smoke				-9.28	1.04	<0.001

The R^2 of Model 1 is 4.90%, and the R^2 of Model 2 is 10.97%. This means **the introduction of smoke variable explains even more variance than the variance that the other variables can explain, which strongly supports Claim 2.**

The coefficients of height and weight in both model 1 and model 2 do not change much after we introduce the smoke variable, which means smoke variable can only explain a little of the impact that other variables have on the infanct body weight. It really makes sense because the relationship between smoke and other variables is trivial as shown in the correlation part in Section 1.

Besides, the p-value of Analysis of Variance for the two models is far smaller than 0.001, indicating that the second model is much better fit, and thus the smoke variable has a strong relationship to infant birth weight. In conclusion, **Claim 2 is greatly supported.**

3.4 Pros and Cons of Multiple-Regression

- Advantage: It controls the effect of other variables.
- Disadvantage: With more variables included in the model, the effective sample size will become smaller, and then the statistical efficiency will decrease. Multiple regression assumes that the variances for different groups (categorized by the binary variables) are the same, while the univariate comparison/regression allows different variances, which is more flexible.

3.5 Reduction in Root Mean Square of Residuals

In this section, I conduct univariate regression analyses with only smoke, parity, height, weight and age, and then calculate the root mean square of the residuals (r.m.s.). The smoke variable has the largest reduction in the r.m.s. of variance, indicating that **it can explain the most variance in the infant born weight** (Table 11), which **can also support Claim 2.**

Table 11: Reduction in Root Mean Square of Residuals

variable	r.m.s. reduction (%)
smoke	0.57 (3.09%)
parity	0.02 (0.1%)
height	0.38 (2.1%)
height_cat	0.26 (1.4%)
weight	0.22 (1.22%)
weight_cat	0.19 (1.02%)
age	0.01 (0.04%)
age_cat	0.02 (0.12%)

3.6 Multi-Panel Figure

Finally, I produce a single multi-panel figure which does the following: for each bin created in a three-way classification by tall/short, heavy/light, parityi, visually compare the birth-weight distribution of smokers versus nonsmokers (Figure 9). This plot also supports Claim 2.

The advantage is that we can explicitly compare the distribution in eight bins with similar features in mother's height, weight and parity. The figure shows that in all the eight groups, those infants with nonsmoking mother has larger born weights.

The disadvantage is that it only gives us the intuition but not a quantitative result or the confidence of our conclusion.

3.7 Limitation

We can only compare the influence of two-level indicator on the infant born weight. However, the true relationship between the outcome and the variable (especially continuous variables) is usually non-binary and complicated. But if we try to compare the influence in infant born weight on a continuous scale, there's no consistent method for both binary (parity, smoke) and numerical ones. Therefore, it's a great challenge that needs further research and discussion.

References

Deborah Nolan, T. S. (2000). Stat labs: Mathematical statistics through applications. *Springer Texts in Statistics.*

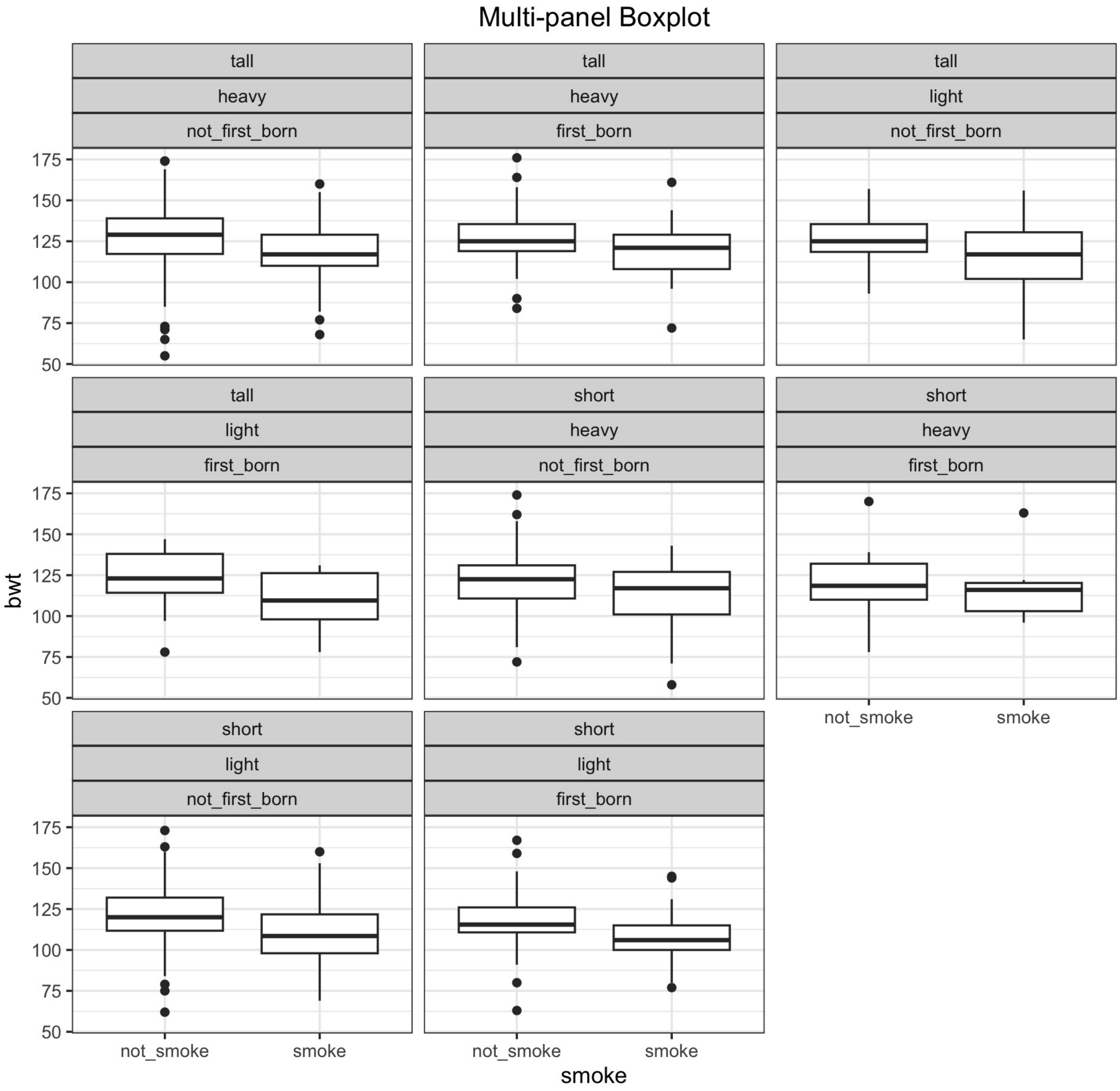

Figure 9: Multi-panel Boxplot