Data Preparation

Within *babies.data*, there are 1236 observations of 7 important variables related to the live births. The data cleaning procedure is mainly focused on the missing value removal and catagorical variable conversion. Among several variables, for example, *gestation* indicating the days of pregnancy, labels missing values with the number of 999. On the other hand, *parity* indicating whether the baby is the first-born, needs to be coded categorically with description of levels. The initial cleaning results on the data frame *babies* are shown below.

```
##    bwt gestation parity age height weight smoke
## 1  120       284 not.FB  27     62    100    No
## 2  113       282 not.FB  33     64    135    No
## 3  128       279 not.FB  28     64    115   Yes
## 4  123        NA not.FB  36     69    190    No
## 5  108       282 not.FB  23     67    125   Yes
## 6  136       286 not.FB  25     62     93    No
## 7  138       244 not.FB  33     62    178    No
## 8  132       245 not.FB  23     65    140    No
## 9  120       289 not.FB  25     62    125    No
## 10 143       299 not.FB  30     66    136   Yes
```

When working on the statistical claims, we need to add serveral categorical variables into the current data frame. *premature* is added as a two-level factor variable to indicate whether a baby was born prematurely; and a premature birth is defined as occuring prior to the 37th week ($37 \times 7 = 259$ days) of gestation. Similarly, *f.height* and *f.weight* are added to divide the mothers into groups based on median height and weight in the data. So the final clean data before any analysis is sketched below.

```
##    bwt gestation parity age height weight smoke premature f.height  f.weight
## 1  120       284 not.FB  27     62    100    No    not.PM    short     light
## 2  113       282 not.FB  33     64    135    No    not.PM    short     heavy
## 3  128       279 not.FB  28     64    115   Yes    not.PM    short     light
## 4  123        NA not.FB  36     69    190    No      <NA>     tall     heavy
## 5  108       282 not.FB  23     67    125   Yes    not.PM     tall     light
## 6  136       286 not.FB  25     62     93    No    not.PM    short     light
## 7  138       244 not.FB  33     62    178    No     is.PM    short     heavy
## 8  132       245 not.FB  23     65    140    No     is.PM     tall     heavy
## 9  120       289 not.FB  25     62    125    No    not.PM    short     light
## 10 143       299 not.FB  30     66    136   Yes    not.PM     tall     heavy
```

We can get an initial feel for the data by looking at the histogram of one of the important variable, *gestation*. The proportion of premature pregnancy (shorter than 259 days) is around 30 percent, which is suitable for analysis done on grouping the mothers based on this particular variable. Similar initial checks are also done but will not be shown here due to redundancy.

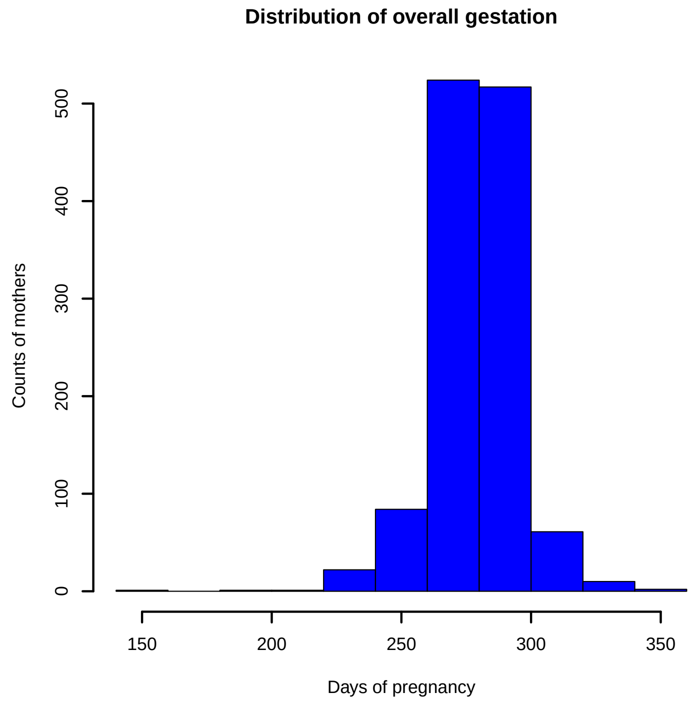

Figure 1: Initial check of the dataset

Claim 1

Mothers who smoke deliver premature babies more often than mothers who do not.

Long story short, this claim is NOT well supported with this dataset.

1. Graphical comparisons of the gestation distribution

I tried two different ways of graphical comparisons, histogram in parallel and group boxplot, both shown below. The gestation distributions of smoking mothers and of non-smoking mothers DO NOT show a significant difference in either mean or spread.

2. Tabular comparisons of the factor variables

With the added two-level factor variable, *premature*, indicating whether the baby was born prematurely, we use the two factors, *smoke* and *premature*, to carry out a relevant tabular comparison of distributions with results shown below.

```
tb.PS <- table(premature, smoke)
ftable(tb.PS)

##            smoke  No Yes
## premature
## not.PM           677 439
## is.PM             56  41
```

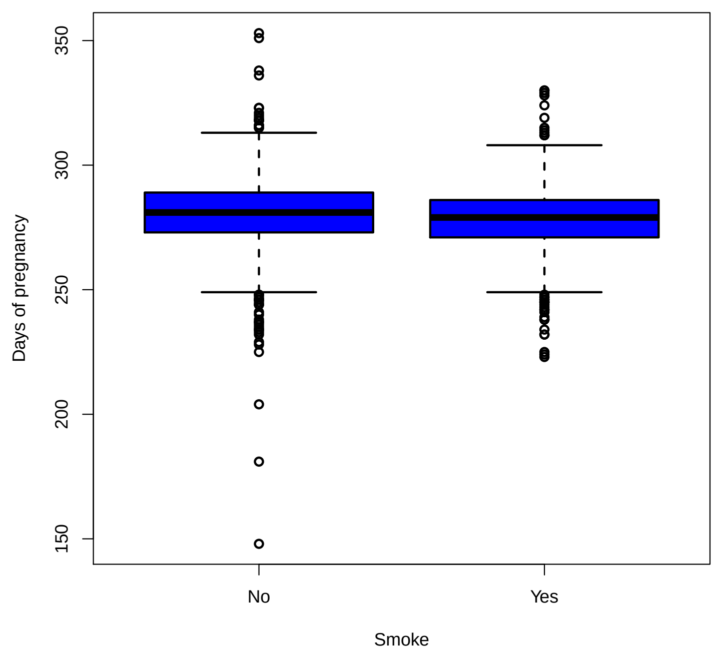

Figure 2: Graphical comparisons of the gestation distribution

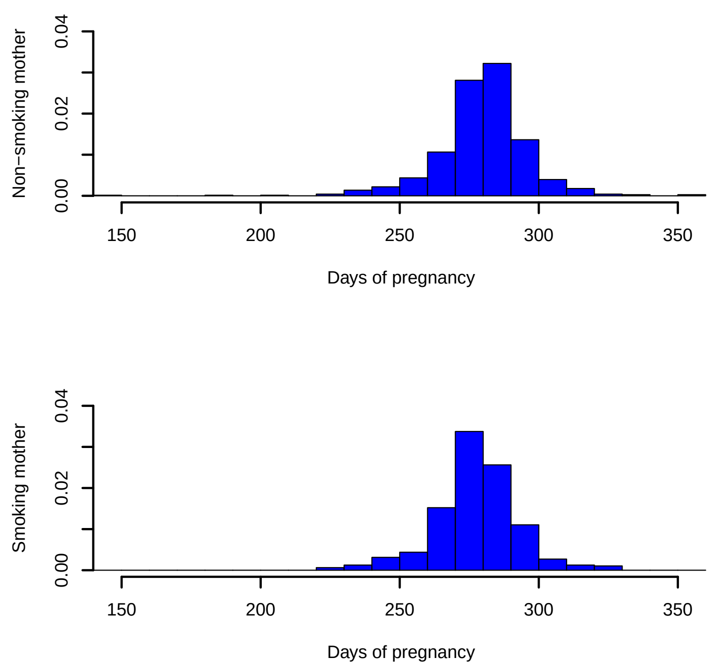

Figure 3: Graphical comparisons of the gestation distribution

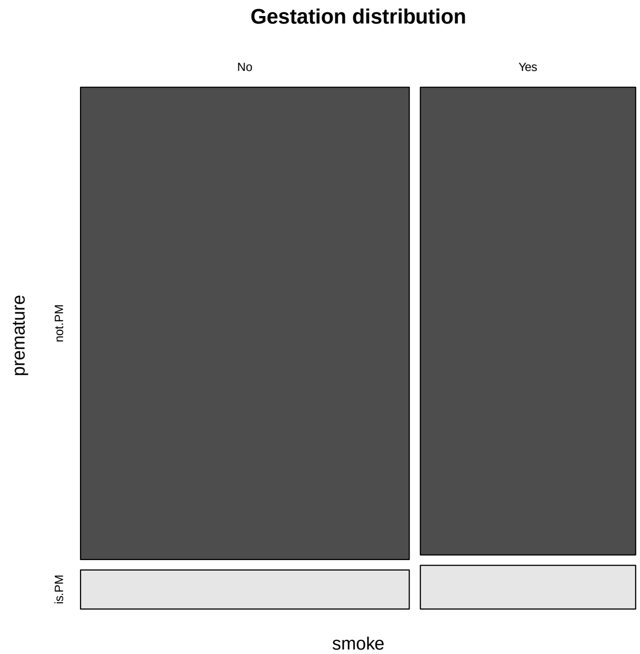

Figure 4: Visual tabular comparisons

3. Visual tabular comparisons

The figure shown below allows us to carry out visually the tabular comparisons of the factor variables, *smoke* and *premature*. We can NOT tell significant difference in the distribution of gestation between smoking and non-smoking groups from the figure.

4. Hypothesis test on the tabular comparison

With the given table shown above, we want to conduct hypothesis test formally for our claim.

H0: smoking and non-smoking mothers have the same rate of premature delivery
HA: smoking and non-smoking mothers DO NOT have the same rate of premature delivery

Two tests are conducted, $Chi - squre$ and $Fisher$, and both have shown a fairly large p-value, which indicates that we should not reject the null hypothesis.

```
chisq.test(tb.PS)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.PS
## X-squared = 0.2098, df = 1, p-value = 0.6469
```

```r
fisher.test(tb.PS)

##
##  Fisher's Exact Test for Count Data
##
## data:  tb.PS
## p-value = 0.5893
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7221 1.7530
## sample estimates:
## odds ratio
##      1.129
```

5. Hypothesis test on the overall average comparison

With the data, we want to conduct hypothesis test formally for a related question concerning the overall average gestation time for the *smoke* groups.

> *H0: smoking and non-smoking mothers have the same overall average gestation time*
> *HA: smoking mothers have shorter overall average gestation time*

Similarly, two one-sided tests are conducted, t and $Wilcox$, and both have shown a fairly large p-value, which indicates that we should not reject the null hypothesis.

```r
t.test(gestation ~ smoke, alternative = "less")

##
##  Welch Two Sample t-test
##
## data:  gestation by smoke
## t = 2.394, df = 1093, p-value = 0.9916
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##   -Inf 3.726
## sample estimates:
##  mean in group No mean in group Yes
##             280.2             278.0

wilcox.test(gestation ~ smoke, alternative = "less")

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  gestation by smoke
## W = 195867, p-value = 0.9996
## alternative hypothesis: true location shift is less than 0
```

Table 1: Comparison of univariate influence against smoke

Variable	Wald-test statistic	p-value (Normal)
parity	0.2851	0.3878
height	0.1185	0.4528
weight	0.1460	0.4419

Claim 2

Cigarette smoking has a stronger relationship to infant birth weight than serveral other relevant covariates.

Long story short, this dataset DOES support this claim.

1. First-borns comparisons

With the data, we want to conduct hypothesis test formally for comparing the influence from relevant variables on birth-weight *bwt*.

H0: the difference in the average bwt between smoking/non-smoking mothers is the same as that of firt-borns/non-first-borns

HA: the difference in the average bwt between smoking/non-smoking mothers is NOT the same as that of firt-borns/non-first-borns

With the assumption of $i.i.d$ samples, we conduct a Wald test on the groups, that is,

$w = \frac{|\delta A| - |\delta B|}{se(\delta A - \delta B)}$

$\delta A = \bar{X}_{smoker} - \bar{X}_{non-smoker}; \delta B = \bar{X}_{first-born} - \bar{X}_{non-first-born}$

$var(\delta A) = var(\bar{X}_{smoker}) + var(\bar{X}_{non-smoker}); var(\delta B) = var(\bar{X}_{first-born}) + var(\bar{X}_{non-first-born});$

$var(\delta A - \delta B) = var(\delta A) + var(\delta B) - 2cov(\delta A, \delta B)$

The details of the covariance calculations are given in the R code. So from this, we can get the test statistic and p-value for the test. The results are given below in the table for better comparison.

2. Mother height comparisons

Similar to 1., we conduct the test and show the results in the table below.

3. Mother weight comparisons

Similar to 1., we conduct the test and show the results in the table below.

4. Visual comparison on distributions

In addition to the given table shown above, we want to make multi-panel comparions of the whole distribution visually as the figure shown below.

5. Multiple linear regression without smoking status

With the data, we fit a linear regression model

$bwt_1 = \beta_{1,0} + \beta_{1,1}height + \beta_{1,2}weight + \beta_{1,3}parity.$

The summary is shown below and we check the fit by two plots. The one with $fitted \sim residual$ is for checking the linear model $Y \sim N(\beta^T x, \sigma)$; while the histogram of the residuals from the fitting is for checking the normality assumption of the residual distribution. From visual inspection, both assumptions are satisfied.

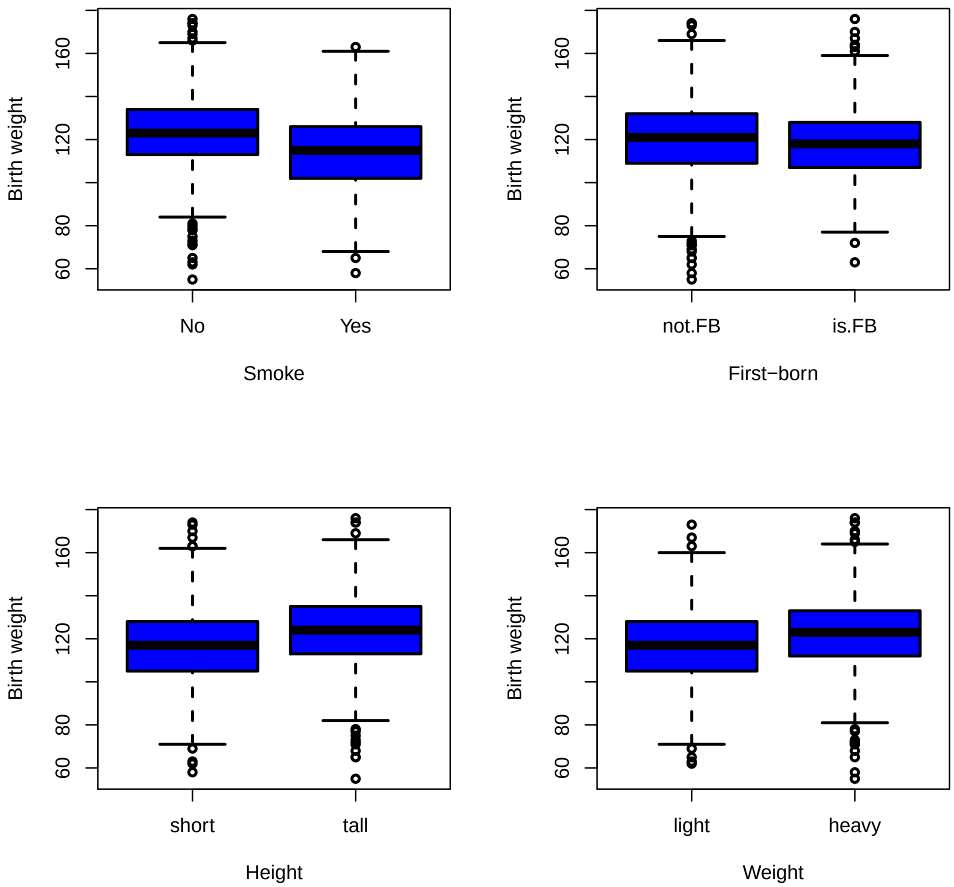

Figure 5: Difference in bwt comparison of relevant variables

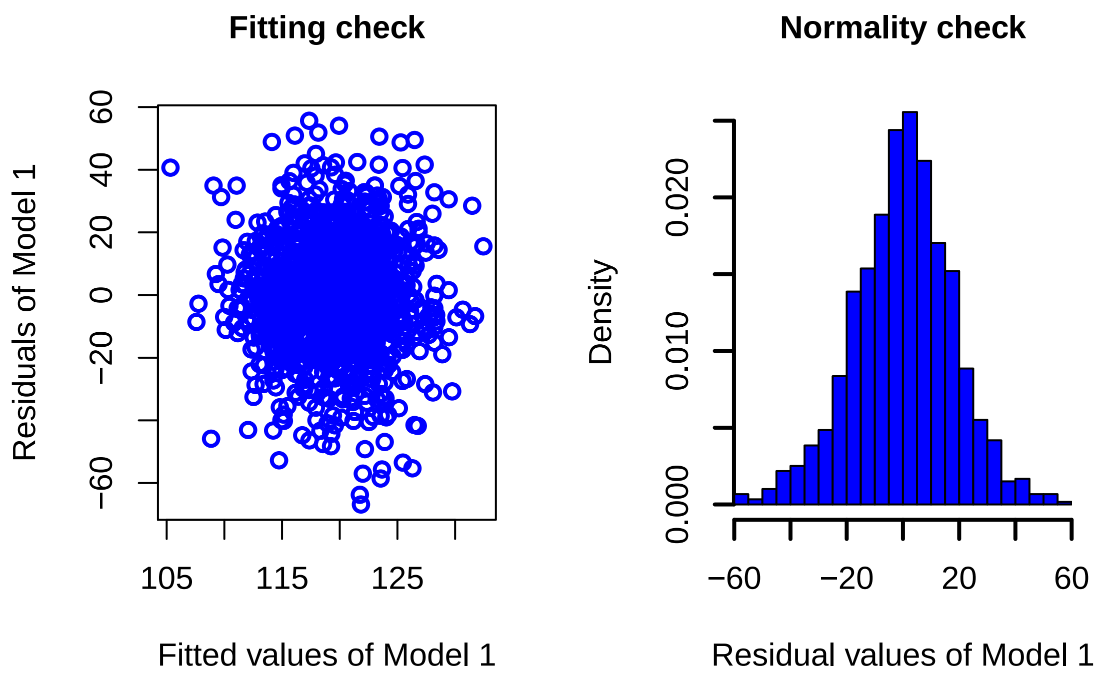

Figure 6: Model 1 fitting check

6. Multiple linear regression with smoking status

With the data, we fit another linear regression model

$bwt_1 = \beta_{1,0} + \beta_{1,1} height + \beta_{1,2} weight + \beta_{1,3} parity + \beta_{1,4} smoke.$

The model is also checked for normality assumptions and is well satisfied.

We summarize the two models in the table below for easy informal comparison. As shown in the table, the model 2 (with *smoke* factor) has lower residual standard error, higher R^2, lower p-value as compared to the model 1 (without *smoke* factor). We can conclude that M2 is better and that the *smoke* factor DOES have a strong relationship to infant birth weight *bwt*.

Meanwhile, ANOVA is carried out for formal comparison between the two linear models. From the ANOVA results, we can see that *smoke* explains the majority sum of squres (SS) compared to all the other relevant variables (*height,weight,parity*), with the smallest p-value. This also supports our claim 2.

Table 2: Comparison of linear regression models

Multiple Linear Regression	M1 : bwt height+weight+parity	M2 : bwt height+weight+parity+smoke
Residual standard error	17.94 (df=1193)	17.34 (df=1182)
Multiple R^2	0.04695	0.1068
Adjusted R^2	0.04456	0.1038
F-statistic	19.59	35.33
p-value	2.112e-12	<2.2e-16

```
anova(fit1)

## Analysis of Variance Table
##
## Response: bwt
##             Df Sum Sq Mean Sq F value  Pr(>F)
## height       1  15888   15888    49.4 3.5e-12 ***
## weight       1   2348    2348     7.3   0.007 **
## parity       1    675     675     2.1   0.148
## Residuals 1193 383827     322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(fit2)

## Analysis of Variance Table
##
## Response: bwt
##             Df Sum Sq Mean Sq F value  Pr(>F)
## height       1  16110   16110   53.57 4.6e-13 ***
## weight       1   2278    2278    7.57   0.006 **
## parity       1    628     628    2.09   0.149
## smoke        1  23478   23478   78.08 < 2e-16 ***
## Residuals 1182 355435     301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. Pros and Cons of multiple-regression approach

As compared to the univariate comparisons we carried out initially, the multiple-regression approach gives more accurate results of the relative influence on the birth-weight among serveral relevant variable. Null acceptance can be derived purely from the p-values of the Wald tests for univariate comparison, which seems to underestimate the relationship of *smoke* to *bwt*. The multiple regression models have shown us that the *smoke* is the single most influential variable for the *bwt* changes. However, the regression approach may be overly optimistic about the claim as visual inspection on the plots does not give that much confidence.