MATH8009 – 2022-23 Project

Dataset: births.csv

Brian Higgins

1. Introduction

1.1 Data Introduction.

The dataset contains data on 250 mothers who gave birth to a single baby each. The variables in the data are given in the table below:

- Bweight – Numerical – Weight(g) of newborn babies
- Gestwks – Numerical – Number of gestational weeks
- Matage – Numerical – Age of month (Maternal Age)
- Lowbw – Binary - 0 = birth weight above 2,500g, 1 = birth weight below 2,500 g
- Preterm – Binary - 0 = baby not born premature, 1 = baby born premature
- Hyp – Binary - 0 = mother does not have hypertension, 1 = mother has hypertension
- Sex – Binary - 0 = male baby, 1 = female baby

1.2 Report Outline

We have been asked by a maternity hospital to review the dataset relating to births and then write a report to summarise our findings.

1.3 Dependent and Independent Variables

The dependent variable is the column “bweight” or the body weight of the baby at birth. We will use the other columns as independent predictor variables to see if there are any relationship between them and the dependent variable. For example, we will look to see if pregnancy length or the age of the mother had an effect on the body weight of the baby at birth or if hypertension had an effect. We will also create a Linear Regression Model to find which of the predicator columns is the best predicator of the body weight of a baby at birth.

1.4. Missing and Duplicate data.

A quick look at the Fig 1.4.1 and we can see that there is no missing data. In Fig 1.4.2 there are many duplicates but this is not surprising given the type of data, in this case, the most important zero result is that the Id column is not duplicated so we know each row is a unique entry.

```
Missing values in id : 0
Missing values in bweight : 0
Missing values in gestwks : 0
Missing values in matage : 0
Missing values in lowbw : 0
Missing values in preterm : 0
Missing values in hyp : 0
Missing values in sex : 0
```

Table 1.4.1 Missing columns

```
Duplicate values in id : 0
Duplicate values in bweight : 62
Duplicate values in gestwks : 159
Duplicate values in matage : 479
Duplicate values in lowbw : 498
Duplicate values in preterm : 498
Duplicate values in hyp : 498
Duplicate values in sex : 498
```

Table 1.4.2 Duplicate columns

1.5. Data

Fig 1.5 shows the head of the data and what it looks like, we can see the column names at the top, bweigth (Body Weight), gestwks (Gestational Weeks) and matage (Mothers age) are numeric and lowbw (Low Birth Weight), preterm (Pre-term delivery), hyp (Hypertension) and sex are categorial values of 0 and 1.

Head of the Data

id	bweight	gestwks	matage	lowbw	preterm	hyp	sex
1	2974	38.52	34	0	0	0	1
2	3270	41.22	30	0	0	0	0
3	2620	38.15	35	0	0	0	1
4	3751	39.80	31	0	0	0	0
5	3200	38.89	33	0	0	1	0
6	3673	40.97	33	0	0	0	1

Fig 1.5 Head of the Data

2. Descriptive Statistics

2.1 Explore the Numeric variables for Centrality and Spread

We have four continuous numerical variables, “id”, is not important as it is just an identifier, so we will not include it. In fig 2.1 you can see the Mean, Median, Standard Deviation and IQR for each of the columns. We will talk about each of the numeric variables separately in the next sections.

Summary of Numeric Columns

	Mean	Medain	Standard Dev	IQR
bweight	3136.88	3188.50	637.45	689.25
gestwks	38.71	39.11	2.32	2.20
matage	34.03	34.00	3.90	6.00

Table 2.1 Summary of Numeric Columns

2.1.1 bweight Variable

From Table 2.1 the mean for bweight is 3136.88 and the median is 3188.50 so on first glance as the mean is less than the median, although the amount is very small, which tells us this data is left-skewed. A skewness test gives us a value of -0.983, again showing a left skew as it has a minus value. It is below -.05 so this indicates a moderate skew.

Boxplot and Histogram

The Histogram Fig 2.1.1.1 and Boxplot Fig 2.1.1.2 to the right give us more details on what the data looks like. The histogram shows the data does have a left skew and a long tail to the left. We can see the boxplot gives us more details. The median is in the middle of the box but there are many outliers in the lower ranges which shows the left side in the histogram, which could be affecting the distribution.

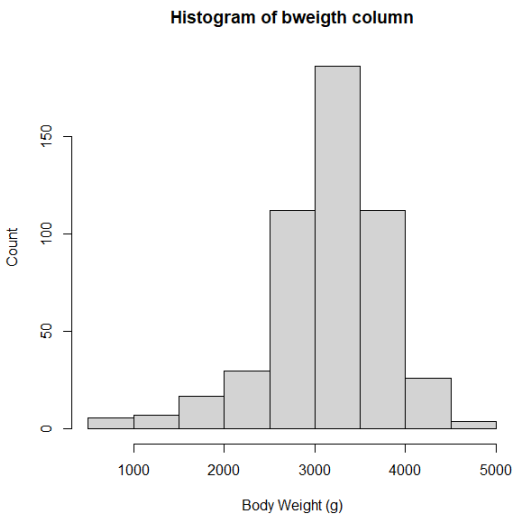


Fig2.1.1.1 Histogram of Bweight

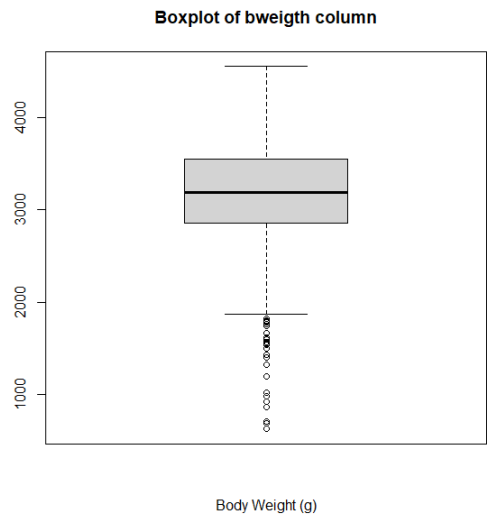


Fig2.1.1.2 Boxplot of Bweight

Sharpiro Wilks test:
Gives us a p-value of 1.256e-12 which is less than the level of significant of 0.05 and so we can reject the Null Hypothesis that the data is normally distributed. So it is skewed.

Since the data is skewed to measure centrality and spread we would use Median and IQR.

2.1.2 gestwks Variable

From Table 2.1 the mean is 38.17 and the median is 39.11 so as the mean is less than the median again, this is showing the data is left-skewed, and a skewness test confirms this with a big negative result of -2.136 which indicates that there is a big left skew.

Boxplot and histogram

The Histogram in Fig 2.1.2.1 shows clearly that there is more data on the right and indicates a left skew, the same as our findings above.

The Boxplot in Fig 2.1.2.2 also shows the median line is in the middle of the box but there are a large amount of outliers causing a skew.

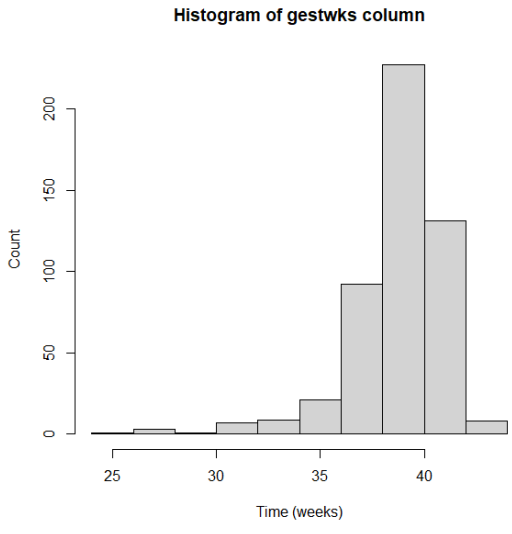


Fig 2.1.2.1 Histogram of gestwks

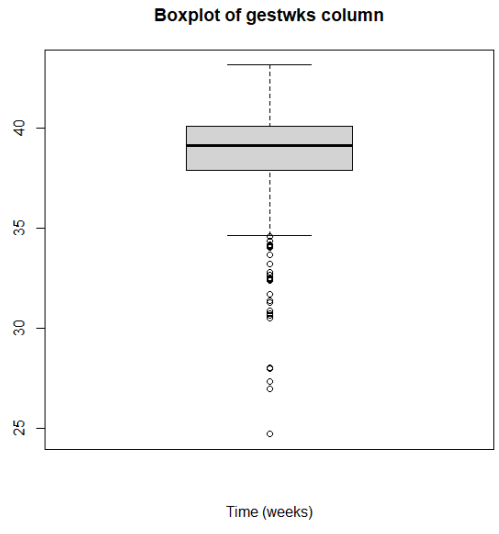


Fig2.1.2.2 Boxplot of gestwks

Shapiro Wilks test
Gives us a p-value 2.2e-16, again below 0.05 so we can reject the Null Hypothesis that the data is normally distributed.

Since the data is skewed to measure centrality and spread we would use Median and IQR.

2.1.3 matage

Table 2.1 shows the mean is 34.03 and the median is 34.00, as the mean is greater than the media this is showing a right skew, however the skewness test gives a -0.232 value which is showing a slight left skew. Below we will look into this in more detail to see which is right. Any value under -0.5 could be considered symmetric but we will look into this more.

Boxplot and histogram

The Histogram in Fig 2.1.3.1 to the right looks almost symmetric and this is backed up by the boxplot in Fig 2.1.3.2 which has the median in the middle of the data with whiskers that look almost the same size. There are no outliers this time to affect the distribution.

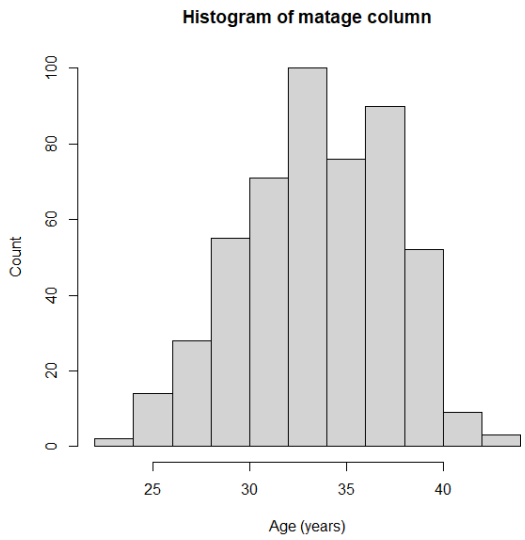


Fig 2.1.3.1 Histogram of matage

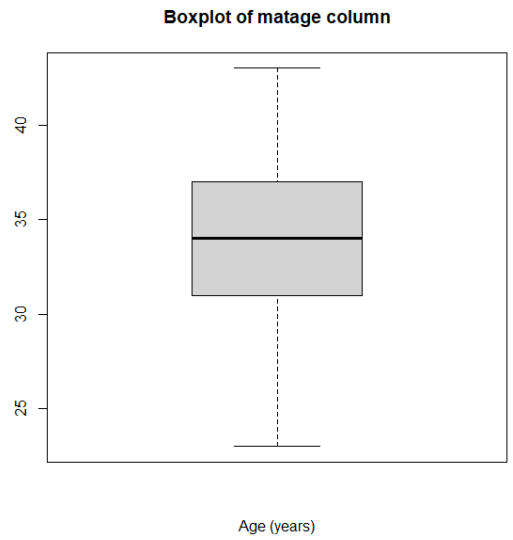


Fig2.1.3.2 Boxplot of matage

Shapiro Wilks test

This gives us a value of 1.517e-05 or 0.00001517 which is below 0.05 and so once again we reject the Null Hypothesis that the data is Normally Distributed. So even with the other results showing visually that the data looks symmetric and the values of the mean and median above it is not symmetric.

Since the data is skewed to measure centrality and spread we would use Median and IQR.

2.2 Questions

1) What proportion of babies are born prematurely?

   14% of the babies born are premature, this accounts for 69 of the 500 babies in the data.

2) What proportion of babies are born below 2,500 g?

   12% of the babies were below 2500, that's 60 babies out of the 500 babies.

3) What proportion of babies are born prematurely and below 2,500 g?

   7.6% of babies are premature and below 2500.

2.3.1 Boxplots for Birthweight by premature age

Fig 2.3.1 shows a boxplot of the birth weight of babies that are born at Normal Term and those that are born Premature. The first thing to note is that blue dots shows the amount of babies in the study so we can see that most babies are born in Normal Term, 86% or 431 out of 500 babies in the study. The median weight of Normal Term babies is higher than that for Premature babies we can see with the lines though the boxplots. The median weight for Normal Term babies also looks symmetric (outliers could affect his a small bit) with the median line in the middle, however for the Premature babies there is left skew of baby weights, with more premature babies born to higher weights for those data points.

50% of the Normal Term babies are born in a smaller range than the Premature babies. There are no outliers for the Premature babies but there are in the Normal Term, in fact some babies born to Normal Term weight less than the median babies born premature. So even babies born premature can weight more than babies born to Normal Term. However, babies born premature nearly all weigh less than babies born full term so it is healthier for babies to be born full term with a higher body weight.

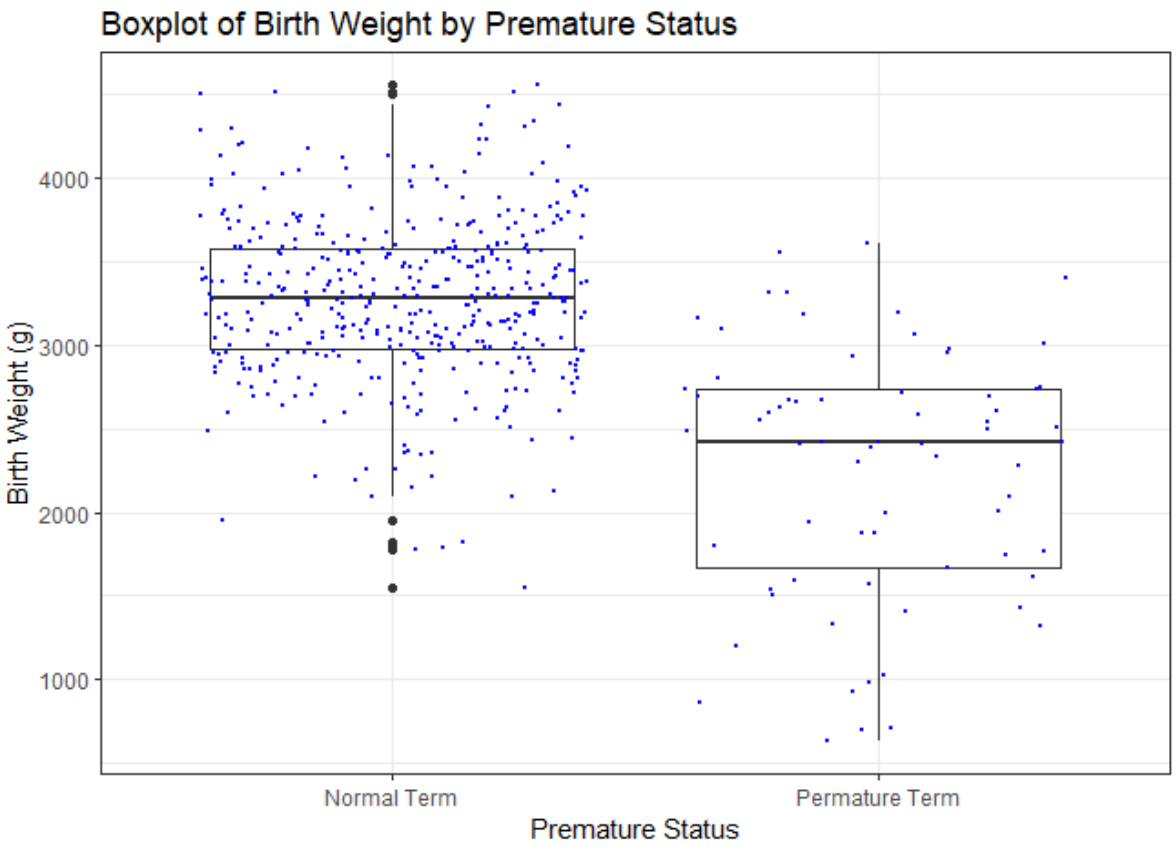


Fig 2.3.1 Boxplot of Body Weight by Premature Status

2.3.2 Boxplots for Birthweight by Hypertension

Fig 2.3.2 shows whether mothers had hypertension during their pregnancy or not using red dots to shows the amount of babies. We can see that most mothers who did not have Hypertension account for 86% of the data, or 428 out of 500 babies. The median for Mothers who did not have Hypertension is higher then that for Mothers who had Hypertension. The data for Mothers with hypertension looks symmetric but is likely affected by the outliers. For mothers with Hypertension the data appears to be left skewed. There are outliers in both groups with Mothers with no Hypertensions showing a great amount of outliers. There are a number of babies whose mothers had no hypertension that were born with low body weights.

Interestingly the data appears to show that for some mothers who did not have hypertension there were several babies that were born with low body weights indicating that there are other factors besides from just hypertension that had an effect on the body weight of new born babies. Mothers who did have Hypertension tended to have babies that had lower body weights so its important to monitor this.

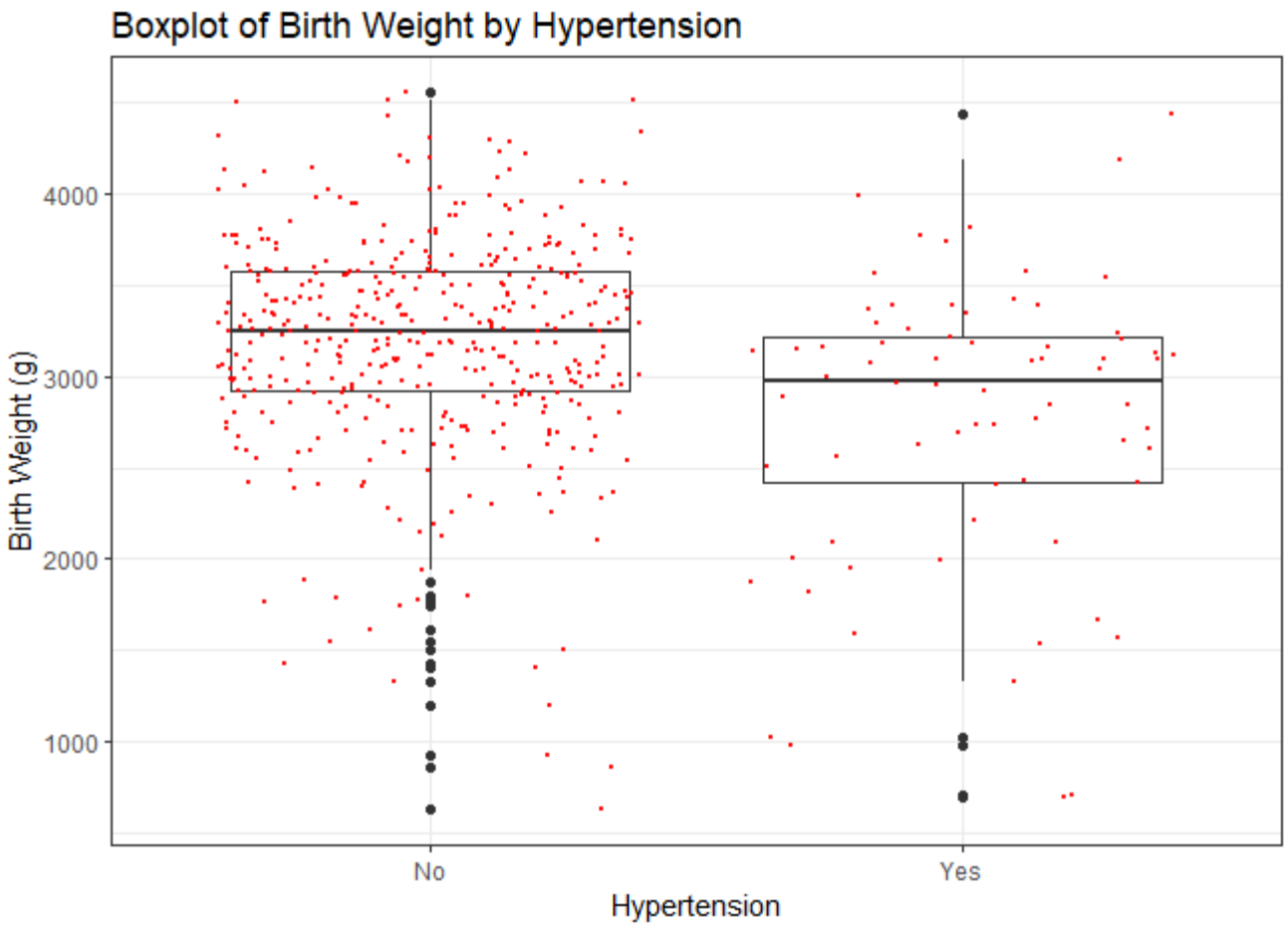


Fig 2.3.2 Boxplot of Body Weight by Hypertension in mothers.

3. Regression Analysis

3.1 Scatter plot of Birth Weight by Gestational Weeks

Fig 3.1 shows a scatter plot of the amount of weeks and the Birth Weight, we can see a strong positive relationship between the length of time and the birth weight of babies born. Showing that the babies who are go longer in term are born with larger weights in general.

The Correlation Coefficient of Birth Weight and Gestational weeks is 0.71 which is just as we have seen in the scatter plot in Fig 3.1 showing a strong positive correlation with indicates a relationship between the two.

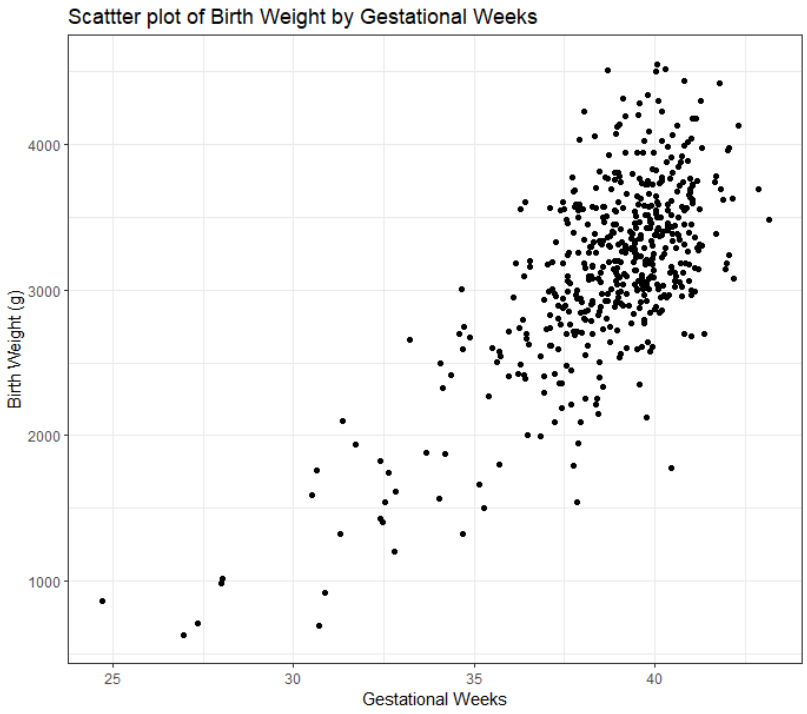


Fig 3.1 Scatter plot of Birth Weight by Gestational Weeks

3.2 Scatter plot of Birth Weight by Maternal Age

Fig 3.2 shows a scatter plot between the birth weight and the maternal age of the mother. There does not appear to be any relationship between these as the data looks like random noise.

The Correlation Coefficient of Birth Weight and Maternal Age is 0.014 which again, just like the scatter plot in Fig 3.2 is showing no correlation as the value is so low.

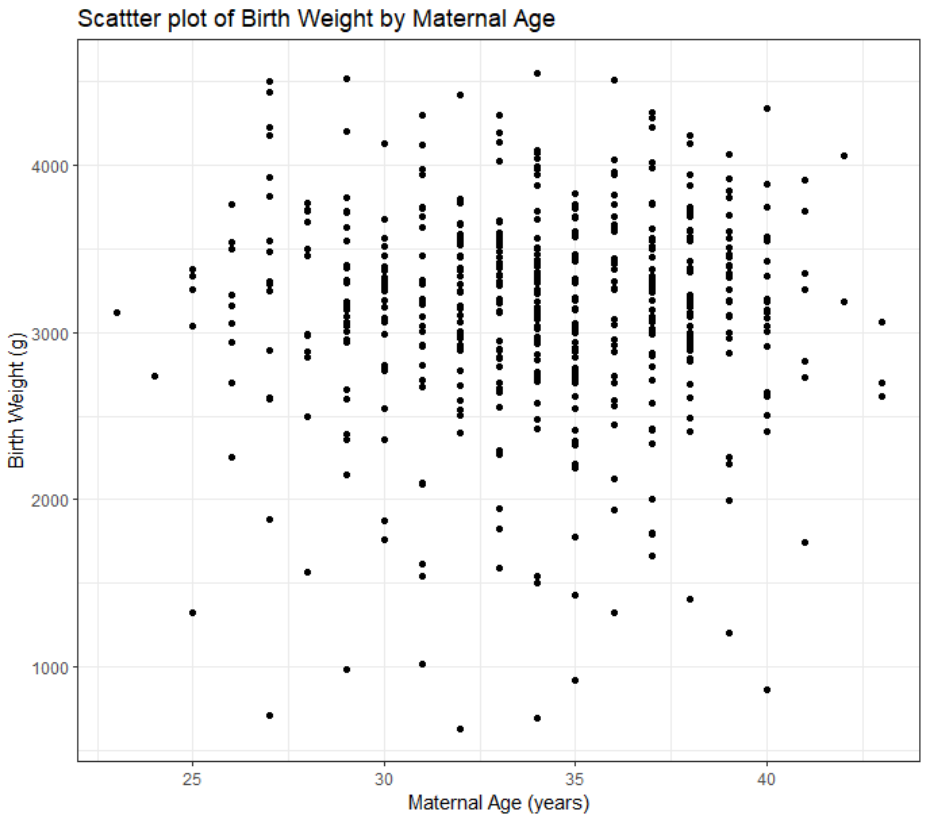


Fig 3.2 Scatter plot of Birth Weight by Maternal Age

3.3 Linear Regression models

1. **Model 1**: bweight = -4465.2 + 196.4 gestwks

   Birth Weight = -4465.2 + 196.4 Gestational Weeks

2. **Model 2**: bweight = 3053.87 + 2.44 matage

   Birth Weight = 3053.87 + 2.44 Maternal Age

3. **Model 3**: bweight = 3198.9 + -430.7 hyp

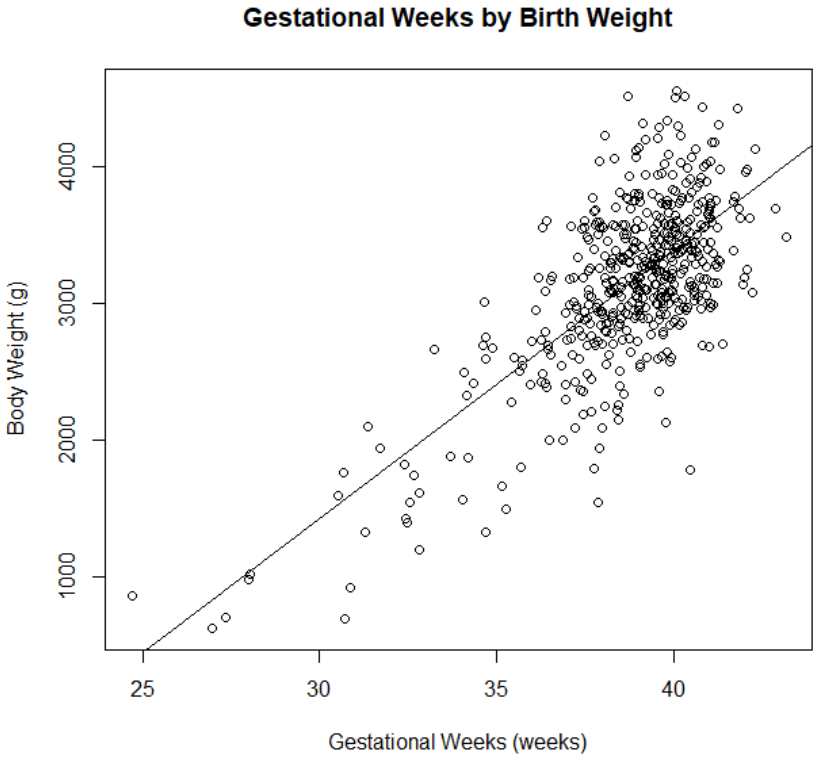
   Birth Weight = 3198.9 + -430.7 Hypertension

4. **Model 4**: bweight = 3279 + -1028 perterm

   Birth Weigth = 3279 + -1028 Premature

3.4 Scatter plots with Regression Lines

Shows the same plots with a regressions lines added. Now it's even clearer that on Fig 3.4.1 we can see a strong positive relationship whereas on Fig 3.4.2 the line is almost completely flat showing no relationship.



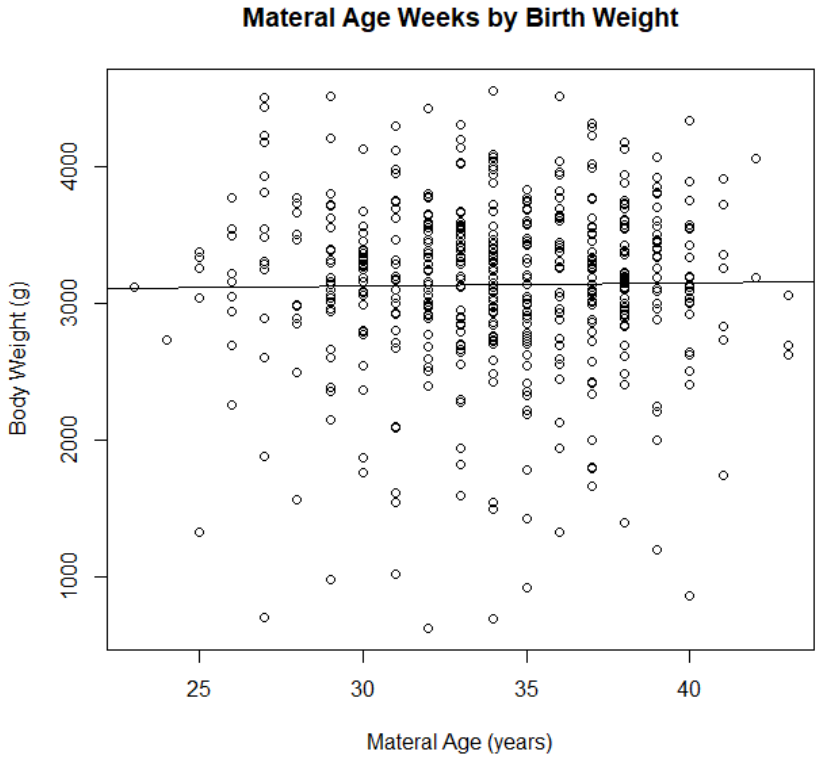


Fig 3.4.1 Gestational Weeks by Birth Weight

Fig 3.4.2 Maternal Age Birth Weight

3.5 Show the summary output from R for Model 1

Fig 3.5 shows the linear regression model results for the Dependent variable bweights and the independent predicator gestwks.

1. **Regression Coefficients**

   The regression coefficients shows the mathematical relationship between the independent predicator variable and the dependent variable. In this case the Independent predicator for the model is gestwks and the dependent variable is bweight.

2. **Coefficient P-Value**

   The p value indicates whether the gestwks predicator in this linear model thisis statistically significant.

   In the gestwks model the coefficient p-value is <2e-16 or 0.0000000000000002. A number well below our level of significance of 0.05 (standard level of significance used). The Null Hypothesis is that there is a relationship between the independent and dependent variables.

What this is saying is there is no relationship due to chance and so we can not reject the Null Hypothesis. There is a relationship between the dependent and independent variables, because the number is so small this indicates a very strong relationship. This is further backed up with the 3 *'s next to the P-Value.

3. Multiple R squared

Gives us a value for how well the model is fitting the data between 0 (not fitting well) and 1 (fitting perfectly) explained by the predicator variable. In the gestwks model this is giving a value of 0.51, so 51% of the data is being predicted. A high R squared value and a low p-value is the ideal outcome as it says our model is explaining the data well.

The R squared value will always increase when we have multiple independent variables, so when we have multiple predicators, we should use the Adjusted R squared value. Which we will see an example of in Section 3.7.

4. F-Statistic and P-value

If the F Statistic p-value is greater than 0.05, than no relationship that exists between **any** of the independent variables and the dependent variables. We should reject the Null hypothesis. If it is less than 0.05, then at least one independent variable is related to the dependent variable and we cannot reject the Null Hypothesis.

This is more important when there is more than one independent variable as the individual coefficient p-values may have an effect on each other. So we will see this in more used In section 3.7 when we check for multiple independent variables.

For the gestwks model we get a p-vlaue of <2e-016, the same as the p-value next to the coefficient, this makes sense as there is only one independent variable.

```
Call:
lm(formula = data$bweight ~ data$gestwks)

Residuals:
     Min       1Q   Median       3Q      Max
-1698.52  -276.73    -5.69   283.47  1381.08

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4465.207    334.243  -13.36   <2e-16 ***
data$gestwks   196.384      8.619   22.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 446.5 on 498 degrees of freedom
Multiple R-squared:  0.5104,    Adjusted R-squared:  0.5094
F-statistic: 519.2 on 1 and 498 DF,  p-value: < 2.2e-16
```

Fig 3.5 Linear Regression Model for gestwks

3.6 Which of the four models would you use to predict birthweight?

You will note in the table 3.6 that the Coefficient P-Value and the F-statistic P-Value are the same as there is only one independent variable. To help make sense of the table , what we are looking for is a p-value's below 0.05, our level of significance so that we can fail to reject the Null Hypothesis and hence show a relationship. This rules out matage as it has a p-value greater than 0.05. Not surprising given the scatter plot in Section 3.4 where the line was mostly flat.

The other three models have p-values less that 0.05 so we can not reject the Null Hypothesis and so there is a relationship between them and the dependent variables while they are the only independent variable.

Model 1: getwks and Model 4: preterm have the lower p-values compared to model 3:hyp. Where the models next differ is in there R -Squared Values and Adjusted R-Squared Values. Model 1: gestwks performs better with an R value of 0.50104. The adjusted square results are also lower but this will be used more when we have multiple coefficients.

So model 1: gestwks is the better model with a lower P-Value for the coefficients, F-statistic p-value and R squared value.

Model Results Table

	Coefficent:P-Value	R-Squared	Adjusted R-Square	F-Statistic P-Value
Model 1: gestwks	<2e-16	0.5104	0.5094	< 2.2e-16
Model 2: matage	0.739	0.0002226	-0.001785	0.7393
Model 3: hyp	7.73e-08	0.05638	0.05449	7.729e-08
Model 4: preterm	<2e-16	0.3098	0.3084	< 2.2e-16

Fig 3.6 Model Results

3.7 Multiple Linear Regression Model

Fig 3.7.1 shows the results of running a multiple linear model with all four independent variables added into the same Linear Regression Model. Previously when each independent predicator had in its own model, gestwks, hyp and preterm all had low p-values indicating they could be used to predict the dependent variable. When all four are run together this is not the case. Gestwks still has a very low P-Value of <2e-16, so no change and the F statistic P-value has stayed the same at <2.2e-16.

But the other two have increased. Now preterm is above 0.05 and is not significant anymore. Hyp still is below 0.05 but has also increased. The model its self has downgraded it from three*'s to one * as an indicator when it is used with other predicators, as gestwks is better. Before we look at the results and compare Model 1 with these results I will run the model again and this time only include gestwks and hyp, results showed in Fig 3.7.2.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3533.030    529.198  -6.676  6.6e-11 ***
data$gestwks   174.450     12.603  13.841  < 2e-16 ***
data$matage     -1.178      5.105  -0.231   0.8176
data$hyp1     -148.668     58.102  -2.559   0.0108 *
data$preterm1 -156.663     83.425  -1.878   0.0610 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 443.4 on 495 degrees of freedom
Multiple R-squared:   0.52,     Adjusted R-squared:  0.5161
F-statistic:   134 on 4 and 495 DF,  p-value: < 2.2e-16
```

Fig 3.7.1 Model Results

In fig 3.7.2 we can see the results for the three tests. I was interested in how the multiple predicator test of Model 5 and Model 6 would compare given that I had removed two of the predicators. You can see that the R squared and Adjusted R squared values dropped by small amounts, while the model did lose some accuracy, the drop is very small. If it was just between these two models I would take Model 6 because it is a simpler model with half the amount of predicators for a very small drop in accuracy.

However for this same reason it is model 1 with only the one predictor gestwks that is the best model over all. It has the same F statistic value as the other two and while the R squared and Adjusted R-squared values are lower than the other two models, it is a much simpler model with only one predicator. This tiny drop in accuracy is worth it to create a simpler model, especially when the F-statistic values have not changed.

Model Results Table

	R-Squared	Adjusted R-Square	F-Statistic P-Value
Model 1: gestwks	0.5104	0.5094	< 2.2e-16
Model 5: gestwks / matage / hyp / preterm	0.52	0.5161	< 2.2e-16
Model 6: geskwk / hyp	0.5165	0.5146	< 2.2e-16

Fig 3.7.2 Model Results

3.8 List potential issues with the data and the models.

1) Linear Regression models are greatly affected by outliers. We saw in the boxplots in section 2.3.2 that we had outliers. In future models we could drops these outliers and run the models again.

2) We saw in sections 2.3.1/2.3.2 that there are outliers that showed some babies born to normal term but weighing less that some premature babies. There may be other factors contributing to this, diet or other health considerations for the mother..

3) There is no indicator if babies born prematurely have any health issues. Is body weight a good indicator for health? On the lower scale perhaps not but there was no indicator given by the hospital what would be considered an unhealthy weight for a new born baby. It is possible with the increase in food density over the years that babies are just born heavier now compared to 20 years ago? Further conversation with the hospital would help. For now we will take it as a given that the median weight is an indicator of health.

4) Lastly, there is no indicator if labour was natural or had to be induced. This would affect the body weight of babies.

3 Conclusions

Overall we have seen that most babies are born to normal term, with 86% of babies in the study. We have also seen that that 88% of babies are born with a body weight above 2500g with the median weight of babies born to normal term being higher than premature babies.

More than any other factor we have seen that the Gestational length of the pregnancy is the biggest predicator in determining a healthy baby with a good body weight. Babies born premature are nearly always weight less than babies born full term so it is important to make sure mothers go the full length of term. There are some expectations to this, especially with mothers with Hypertension.

Hypertension was a factor in babies being born with lighter body weights so this is something the hospital should work with pregnant mothers to see if it can be reduced.

Maternal Age of the Mothers appears to have no affect on the body weight of babies which is a positive sign with families choosing to have babies later in life.

In Summary, the hospital should look into making sure that mothers go the full term length as it leads to healthier babies (this assumption was mentioned in section 3.8).