PREDICTING BABY WEIGHT

[STAT W4702] STATISTICAL INFERENCE & MODELING PROJECT

Arushi Arora(aa3766), Alexander Makarov(adm2190), Eloi Morlaas(em3152), Gary Sztajnman(ggs2121)

December 15, 2015

INDEX

OVERVIEW

This project was conducted on the Low Birth Weight dataset collected in 1986 at Baystate Medical Center, Springfield, Massachusetts as a part of a bigger study on the factors influencing newborn infants' health and risk of serious health problems potentially leading to death. This dataset is distributed as a part of MASS library and contains **189 observations** and **10 variables**, among which bwt represents the newborn infant's weight in grams and is used as the variable of interest that we are trying to predict. The other 9 variables stand for different factors related to mother's physiological parameters, such as age, weight and race, her health-related habits and behavior during pregnancy (smoking habits, presence of uterine irritability and number of physician visits). Also there is a low birth weight indicator low, which is defined as a binary variable showing whether the weight of an infant is below 2500 grams or not. Brief description of each variable is provided in the table below.

The goal of our research is to identify relationship between these variables and infant weight and understand the influence of each of them on the explained variable. The project pursues both inferential and predictive goals as it is equally important to be able to infer about factors affecting newborn's health and to be able to react on the potential health risks in a timely manner when the model predicts the low birth weight outcome for a certain observation. In order to accomplish this goal we tried to fit multiple linear and non-linear models exploring the rationale that could provide the evidence for certain types of models and finding balance between interpretability and predictive power of the model.

CLEANING AND EXPLORING DATASET

For the purposes of the research the dataset was cleaned in the following way:

- birth weight variable bwt is converted from grams to kilograms to reduce the order of magnitude for estimated model coefficients and error values;
- factor variable race was assigned with proper labels white, black and other;
- physician visits were converted to a factor variable ftv with 3 labels 0, 1 and 2+;
- response is defined as an exact amount of infant's weight from bwt;
- all the columns are assigned with meaningful names.

Variable description table and summary statistics of the tidy dataset are provided below.

Variable	Description
baby.grams	weight of newborn infant in kg
mother.age	mother's age in years
mother.weight	mother's weight in pounds at last menstrual period
race	mother's race, factor variable with following labels: *white*, *black* or *other*
smoke	binary variable representing mother's smoking status during pregnancy
prem.labor	binary variable showing whether mother had a previous premature labor or not
hypertension	binary variable showing whether mother had hypertension or not
uterine	binary variable showing presence of uterine irritability
physician.visits	number of physician visits during the first trimester: *0*, *1* or *2+*

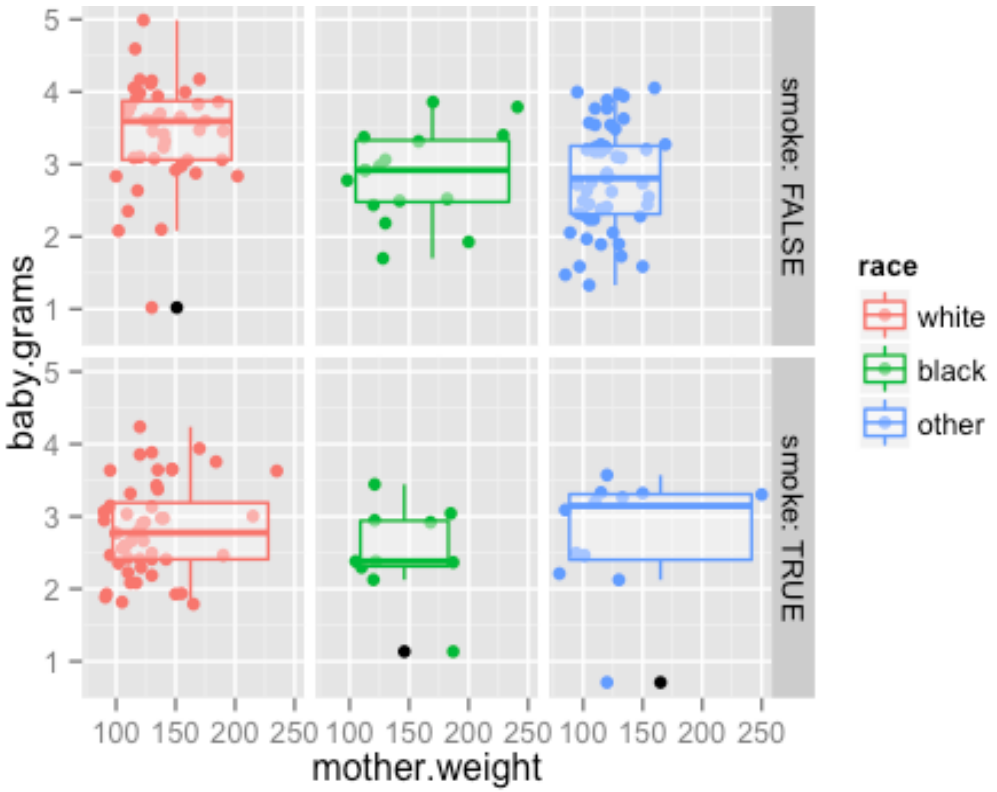
	BABY.GRAMS	MOTHER.AGE	MOTHER.WEIGHT
MIN	0.709	14.00	80.0
1ST QUARTER	2.414	19.00	110.0
MEDIAN	2.977	23.00	121.0
MEAN	2.945	23.24	129.8
3RD QUARTER	3.487	26.00	140.0
MAX.	4.990	45.00	250.0

RACE	SMOKE	PREM.LABOR	HYPERTENSION	UTERINE	PHYSICIAN.VISITS
white:96	TRUE :74	True: 30	True: 12	True: 28	0 visit: 100
black:26	FALSE: 115	False: 159	False: 177	False: 161	1: 47
other:67					2+: 42

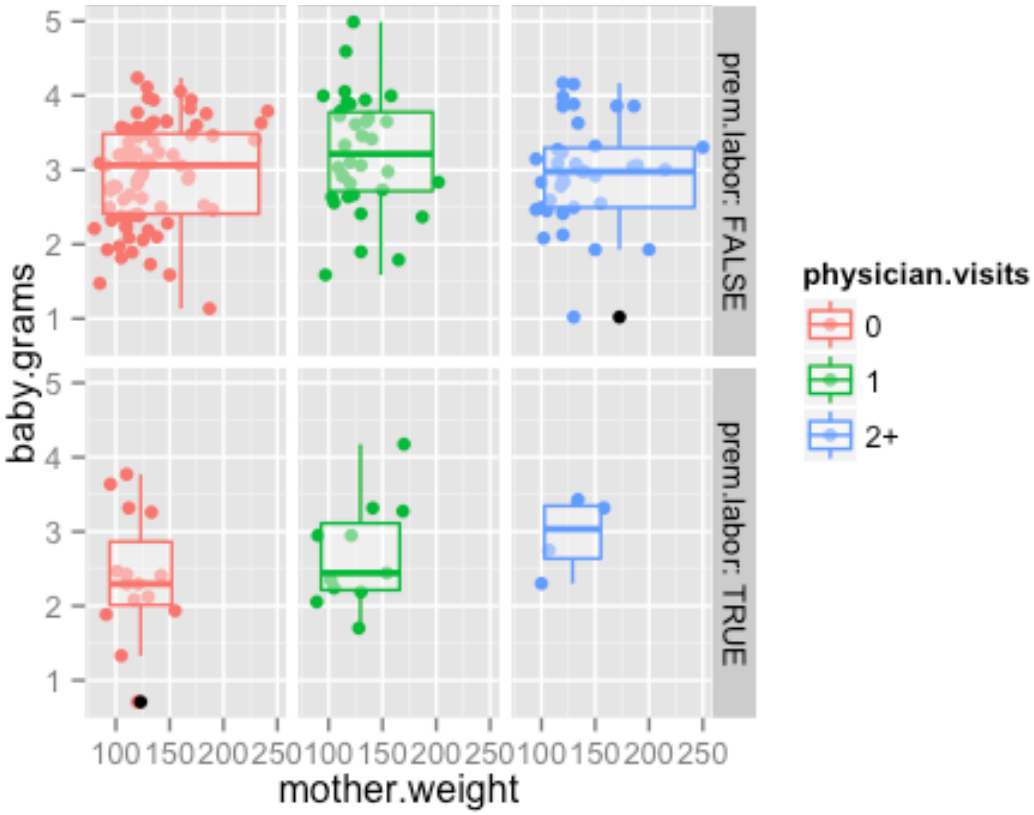
The datatset has only 2 quantitative variables apart from infant weights, however, as shown in the table below, they do not demonstrate strong correlation between each other, which suggests that these variables will not be sufficient themselves in explaining birth weight variation. Variable mother.age demonstrates the lowest correlation with baby.grams and will most probably be omitted in the prediction models further on.

```
##              baby.grams mother.age mother.weight
## baby.grams    1.00000000 0.09031781    0.1857333
## mother.age    0.09031781 1.00000000    0.1800732
## mother.weight 0.18573328 0.18007315    1.0000000
```

The following charts demonstrate boxplots and splits of the baby.grams data points vs mother.weight across various categorical and binary variables that make part of the working dataset.



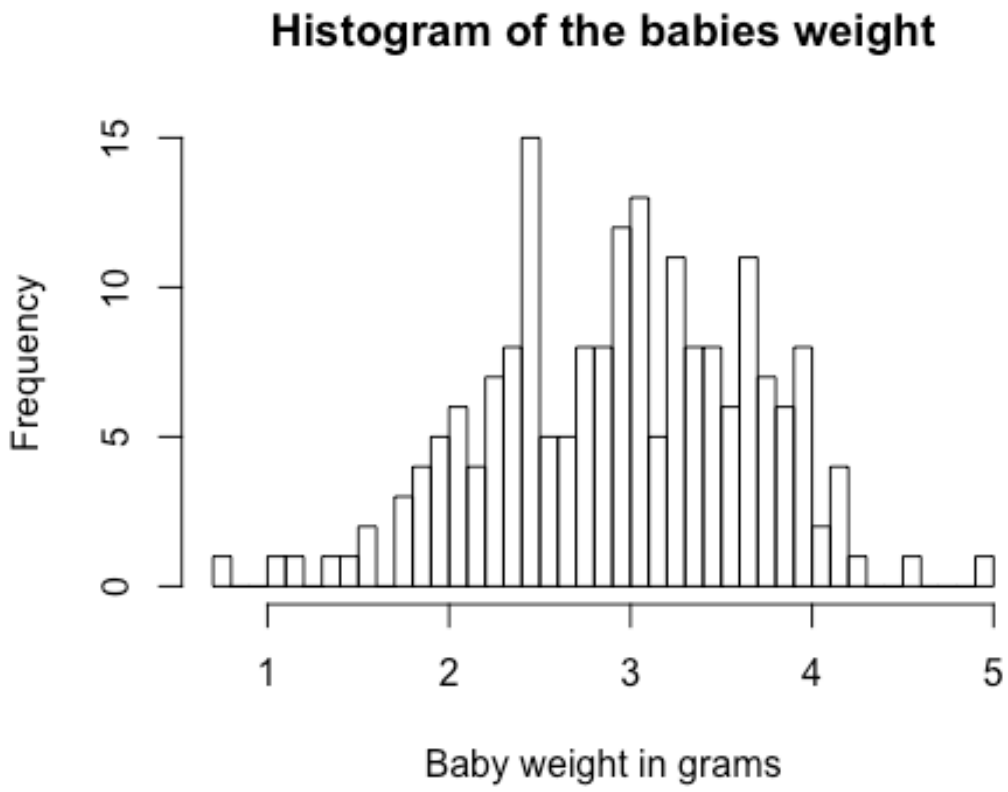
First chart shows some evidence in importance of race in predicting the risk of giving birth to low weight baby, as well as smoking habits during pregnancy. Facet scatterplots show that data point corresponding to each of these factors' combinations group around different median values, which can suggest their predictive power on the newborn infant's weight.



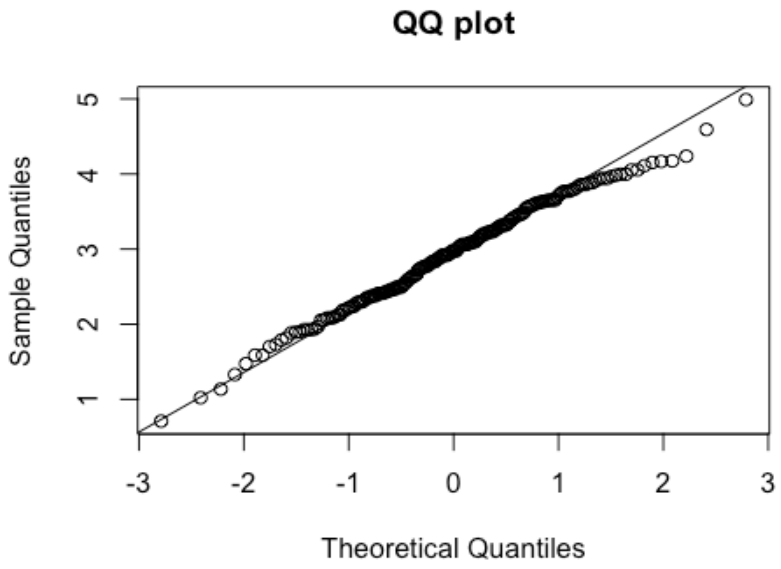
The second chart splits all the observations in sample into several groups by number of physician visits in the first trimester and occurrence of premature labor by each subject of the study. For mothers without previous premature births no significant difference is observed with respect to number of physician visits, whereas women who had premature labors before are exposed to the higher risk of giving birth to low weight baby if they do not pay enough visits to physician during the first trimester of their pregnancy term. However, we need to account for existing outliers in the sample dataset, as there are three observations of infants that were born with weight less than or equal to 1 kg, which significantly differs from the majority of observations in this dataset.

Non-Parametric Analysis

It is well known by doctors that low weight babies have a higher mortality rate than normal or over weight babies. Thus, understanding the factors that can influence the baby weight is an important question. So as to answer to such a broad question we will begin by looking at the shape of the distribution of the babies' weights. It will allow us to argue whether a parametric or a non parametric model is the best fit for this dataset.



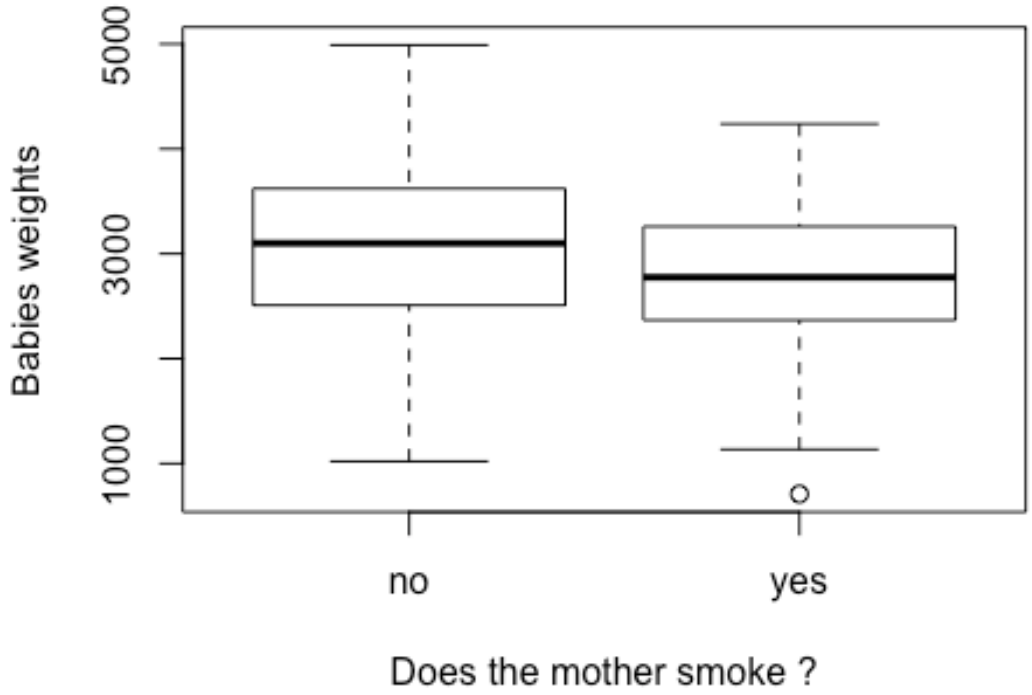
Given this histogram it is quite hard to estimate if the data is truly normally distributed, thus we can draw the corresponding QQ plot.



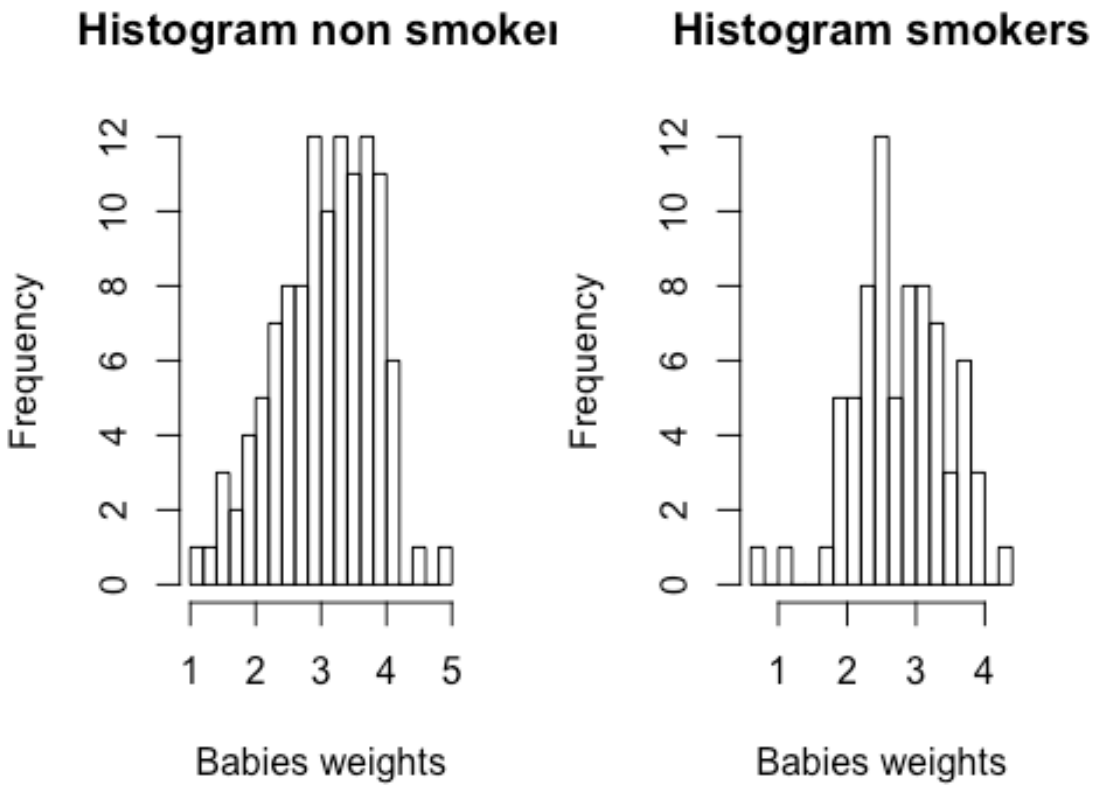
Here we can notice that the observations are quite well aligned on the line which means that the sample quantiles correspond to the quantiles of a theoretical normal distribution. So as to test this hypothesis we can run a Shapiro-Wilk test.

```
shapiro.test(bwt.grams$baby.grams)

##
##  Shapiro-Wilk normality test
##
## data:  bwt.grams$baby.grams
## W = 0.99244, p-value = 0.4353
```

We have a very high p-value that is significantly greater than 0.05 thus we can accept the null hypothesis of the test and conclude that our data is normally distributed.



We can see that the variable smoke seems to have a negative influence on the weight of the baby. We want to statistically verify our assumption. Even if we know that we can assume the data to be normally distributed we can nonetheless try a non parametric approach to answer the question by running a Mann-Whitney test. With this test we determine whether the median of the babies' weights differs between two groups: when the mother smokes or not. For this test to work we need the distribution of babies' weight to have the same shape in both groups. We can easily verify it in these histograms.



We can notice that the two histograms have approximately the same shape, then we can apply the Wilcoxon test to verify if the two populations have same central tendency or not without assuming them to follow the normal distribution.

```
wilcox.test(bwt.grams$baby.grams ~ bwt.grams$smoke)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bwt.grams$baby.grams by bwt.grams$smoke
## W = 5249.5, p-value = 0.006768
## alternative hypothesis: true location shift is not equal to 0
```

We can see that the *p-value* is below 0.05 thus we can reject the null hypothesis and conclude that the median of our two populations are not equal, thus the variable smoke appears to be a variable that can have a certain predictive power in predicting the baby's weight. Nevertheless, here we have seen that we can assume the data to be normally distributed (cf. the Shapiro-Wilk test) thus it is maybe more appropriate to do the alternative parametric test: a 2 sample t-test. Indeed, parametric tests are usually more powerful than their corresponding non parametric tests. Thus, in a non parametric test we are usually less likely to reject the null hypothesis when it is false. Then, if we run a t-test (we return into the parametric world) we have the following results:

```
t.test(bwt.grams$baby.grams[bwt.grams$smoke==FALSE], bwt.grams$baby.grams[bwt.grams$smoke==TRUE])

##
##  Welch Two Sample t-test
##
## data:  bwt.grams$baby.grams[bwt.grams$smoke == FALSE] and bwt.grams$baby.grams[bwt.grams$smoke == TRUE]
## t = 2.7299, df = 170.1, p-value = 0.007003
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.07857486 0.48897860
## sample estimates:
## mean of x mean of y
##  3.055696  2.771919
```

We can notice that this test is considering the mean and not the median as before. We have here a very low p-value that allows us to reject the null hypothesis and conclude that the difference in means of the two samples is not equal to 0. It confirms that the variable smoke is a discriminative variable for predicting the weight. We will confirm this intuition in the models we will build in the next parts.

Linear Models

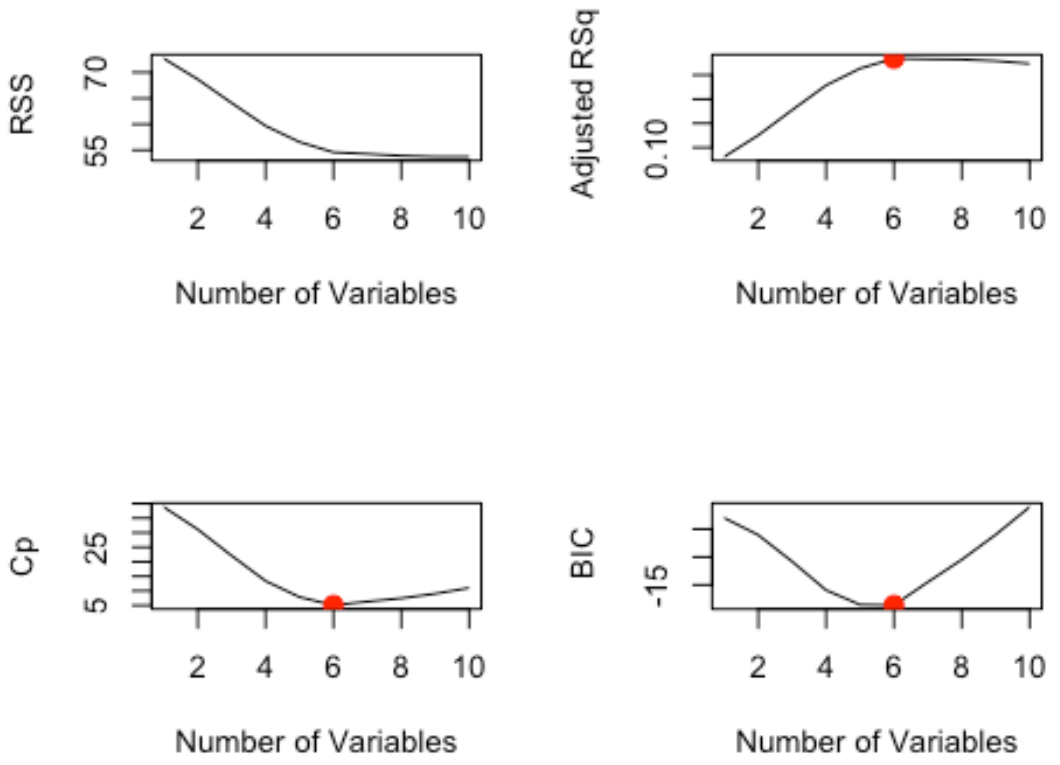
We would like to try different linear models in order to predict the weight of a baby. In order to decide which predictors to choose, we use multiple techniques:

- the best subset selection which fit a separate least squares regression for each possible combination of the p predictors.
- the ridge regression
- the lasso regression

Best Subset Selection

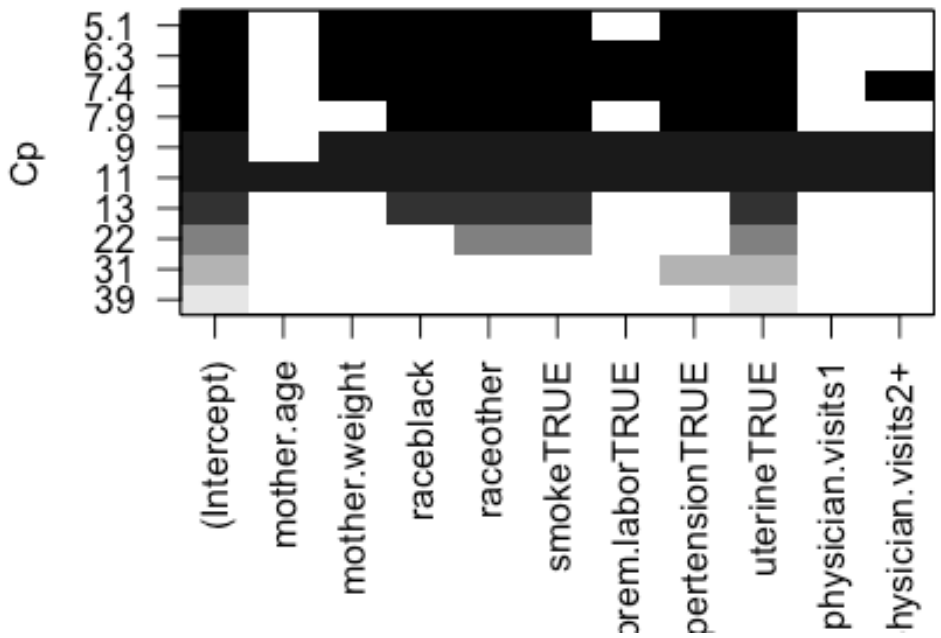
To begin we will first separate the data into a train and a test set so as to be able to compare our models on their test error rather than on their training error (we know that the training error is a poor estimate of the real error of a model). We choose a 75%/25% train/test split. We have to keep in mind here that the test set will only contain 48 observations, thus it might not be a perfect estimator of the overall performance of a model.

```
# Create train and test sets
set.seed(1)
train <- sample(1:nrow(bwt.grams), floor(0.75*nrow(bwt.grams)))
bwt.grams.train <- bwt.grams[train,]
bwt.grams.test <- bwt.grams[-train,]
```



We applied the best subset method on the training set and we compare 3 indicators: BIC, C_p and adjusted R^2. All 3 indicators converge in the idea that we should only consider 6 predictors in our analysis. The following plot is a summary of the results for C_p.

```
par(mfrow =c(1,1))
plot(regfit.full, scale ="Cp")
```



We select the predictors that minimize C_p: mother.weight, race, smoke, hypertension, uterine. Now we want to measure the quality of a linear regression using the predictors selected by the best subset method.

Call: lm(formula = baby.grams ~ mother.weight + race + smoke + hypertension + uterine, data = bwt.grams.train

## Coefficients:	*Estimate*	*Std. Error*	*p-value < 0.05*
Intercept	2.955	0.284	Yes
Mother.weight	0.004	0.002	Yes
Raceblack	-0.567	0.165	Yes
Raceother	-0.487	0.131	Yes
Smoketrue	-0.466	0.120	Yes
Hypertensiontrue	-0.675	0.215	Yes
Uterinetrue	-0.573	0.156	Yes

Multiple R-squared: 0.3147, Adjusted R-squared: 0.284
F-statistic: 10.25 on 6 and 134 DF, p-value: 2.508e-09

Based on the summary, we note that the linear regression model is decent as all predictors have a p-value lower than 0.05 however the R^2 is small with a value of 0.31 *ie* only 31% of the variance of the data is explained by our model. We will see that it is quite hard to do far much better.

We now apply our model on the test set in order to test its accuracy:

MSE best model = 0.467712

Here we note that this linear model has a low accuracy on our test set with a MSE of 0.46. Thus the corresponding mean squared error is: 0.678 Kg which is very high for predicting the weight of a baby as the weights range from 0.709 Kg to 4.99 Kg.

We will then try 2 shrinkage methods to see if we can improve our results.

SHRINKAGE METHODS

In both case (ridge and lasso), we will test 100 different tuning parameters - λs to find the one that minimize the error on 6 folds cross validation. As our database is quite small, we will change the default number of folds from k = 10 folds to k= 6.

RIDGE REGRESSION

Lambda min value = 0.1450829

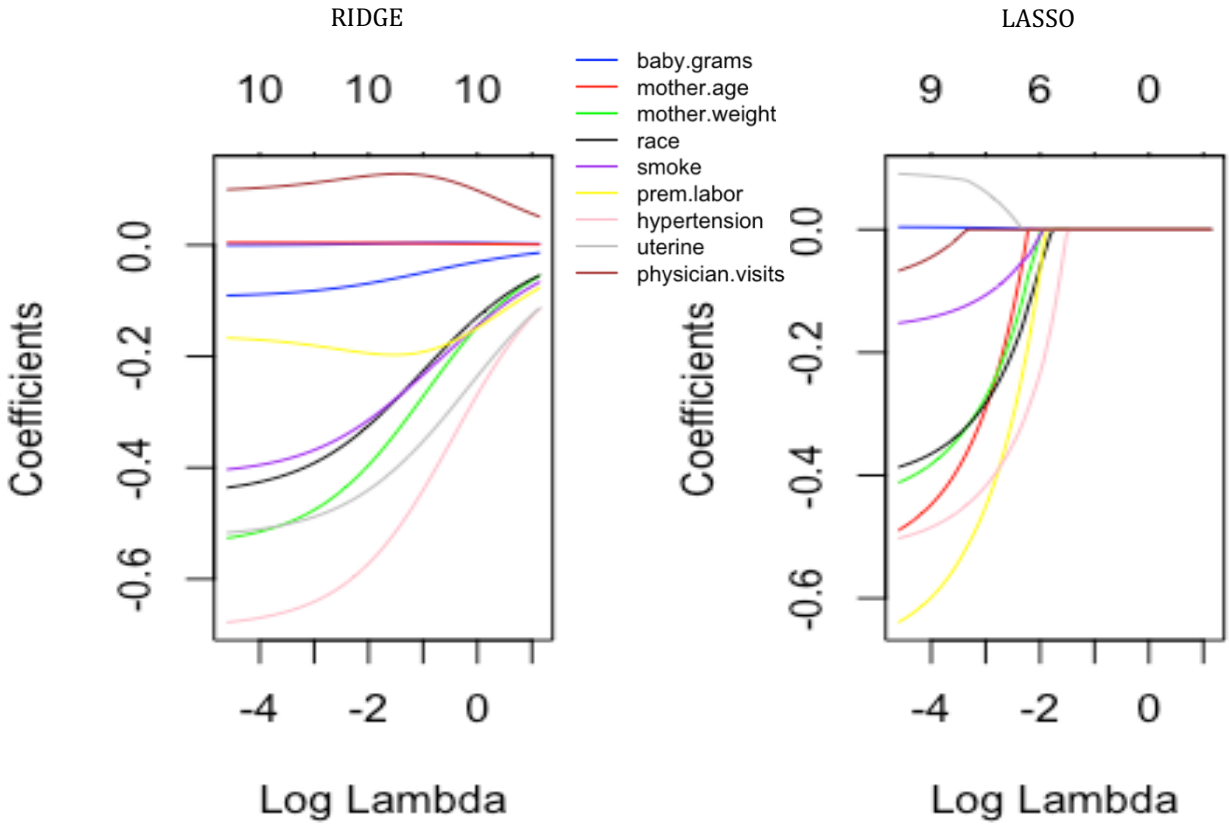
MSE ridge regression = 0.4169001

For the ridge regression we minimize our MSE for a tuning parameter of 0.145. We then perform the ridge regression on the full training set to compute the optimal coefficients. Finally, we test our model and obtain an MSE of 0.416.

LASSO REGRESSION

Lambda min value = 0.0178865

```
## MSE lasso =  0.4289326
```

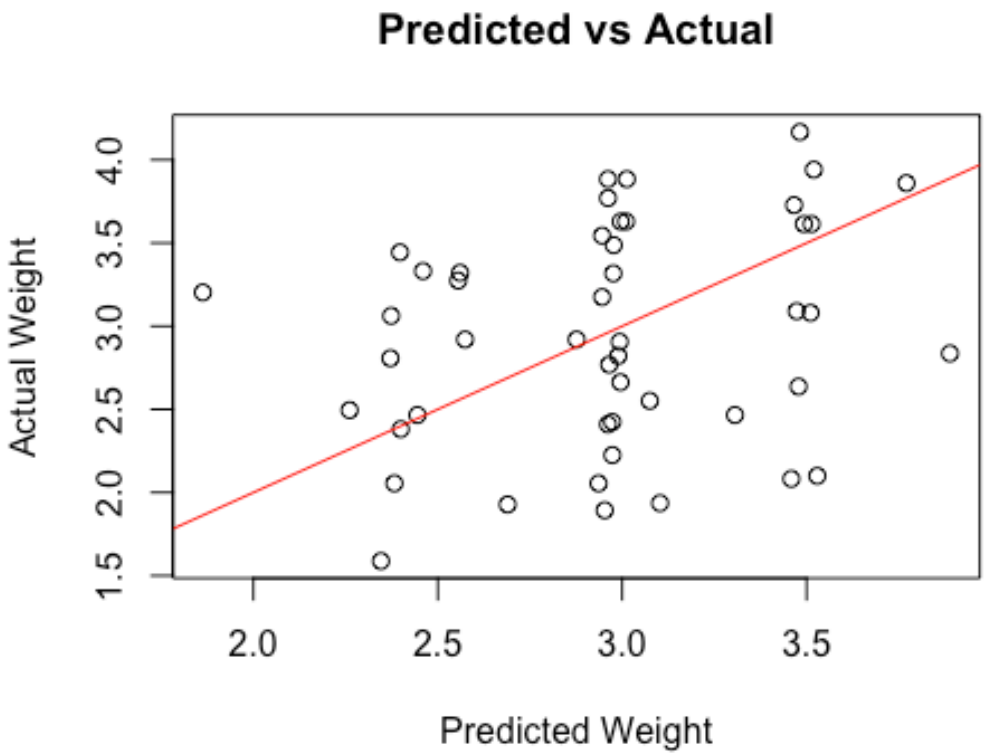


For the ridge regression we minimize our MSE for a tuning parameter of 0.017. We then perform the lasso regression on the full training set to compute the optimal coefficients. Finally, we test our model and obtain an MSE of 0.428. To sum up the results on subset selection we can see that shrinkage methods perform better than subset selection in term of test MSE. The best test MSE is achieved by the ridge regression. Now we can wonder if a polynomial model can bring even more predictive power to our existing model.

Polynomial Model

When we fit a polynomial model on the predictors obtained from best subset, we observe a Mean Squared Error of 0.4813745. The smaller the Mean Squared Error, the closer the fit is to the data. But, as the value of the MSE is high, it suggests that this model does not provide a good fit for the data. The plot also shows that there are irregularities in the prediction and that the polynomial model of degree 2 obtained by using predictors suggested by the best subset is not sufficient.

```
## MSE polynomial model =  0.4813745
```



Different models were tried by increasing the degree of the polynomial but still using the predictors suggested by the best subset and the following results were obtained:

```
poly.fit.2 = lm(baby.grams ~ hypertension + uterine + smoke + race
          + poly(mother.weight, 3), data = bwt.grams.train)
mean((predict.lm(poly.fit.2, bwt.grams.test) - bwt.grams.test[,1])^2)

## [1] 0.4640868

poly.fit.3 = lm(baby.grams ~ hypertension + uterine + smoke + race
          + poly(mother.weight, 4), data = bwt.grams.train)
mean((predict.lm(poly.fit.3, bwt.grams.test) - bwt.grams.test[,1])^2)

## [1] 0.4619314
```

We note that as the degree of the polynomial increases, the MSE decreases, but the drop is not significant, suggesting that these predictors are not sufficient enough to predict the correct baby weight.

When we remove the predictors with very low *p-values*, which were suggested by the best subset - namely `smoke`, `race` and add other predictors which were rejected by the best-subset, namely - `mother.age`, `prem.labor` and `physician.visits`, we see that the Mean Squared Error starts to decrease. A low MSE denotes a better fit. Thus, the predictors which were rejected by the best subset selection, were actually significant in predicting the correct birthweight.

```
poly.fit.4 = lm(baby.grams ~ hypertension + uterine + poly(mother.age,2)
          + poly(mother.weight,3), data = bwt.grams.train)
mean((predict.lm(poly.fit.4, bwt.grams.test) - bwt.grams.test[,1])^2)

## [1] 0.3890751
```

When we fit a polynomial model using mother.age as one of the predictors, we see a significant change in the Mean Squared Error value. Even though best-subset rejected mother.age, the lower Mean Squared Error denotes that the predictor does affect the baby weight - baby.grams. Thus, clearly the fact that mother.age affects the baby.grams cannot be rejected.

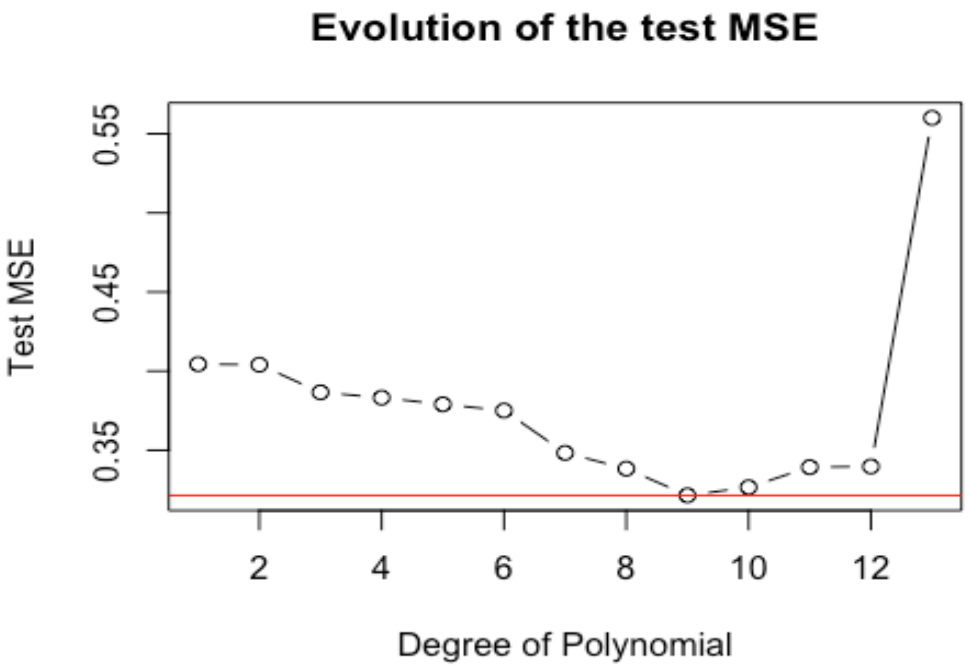
Now, we try another model, where we will choose some of the predictors suggested by the best-subset and some predictors from the previous model. We fit a polynomial of degree 2 on mother.age and a degree 3 polynomial on mother.weight. When we predict the baby.grams on the test set, we observe a mean squared error of 0.3865828, which is not a big significant change from 0.3890751 - the mean squared error from poly.fit.4.

```
poly.fit.5 = lm(baby.grams ~ hypertension + uterine + smoke + prem.labor
          + poly(mother.age,2) + poly(mother.weight,3), data = bwt.grams.train)
mean((predict.lm(poly.fit.5, bwt.grams.test) - bwt.grams.test[,1])^2)

## [1] 0.3865828
```

Now we increase the degrees of the polynomials to figure out if the mean squared error reduces as the degree of the polynomial increases.

```
poly.fit.6 = lm(baby.grams ~ hypertension + uterine + smoke + prem.labor
          + poly(mother.age,2) + poly(mother.weight,9), data = bwt.grams.train)
mean((predict.lm(poly.fit.6, bwt.grams.test) - bwt.grams.test[,1])^2)

## [1] 0.3214657
```

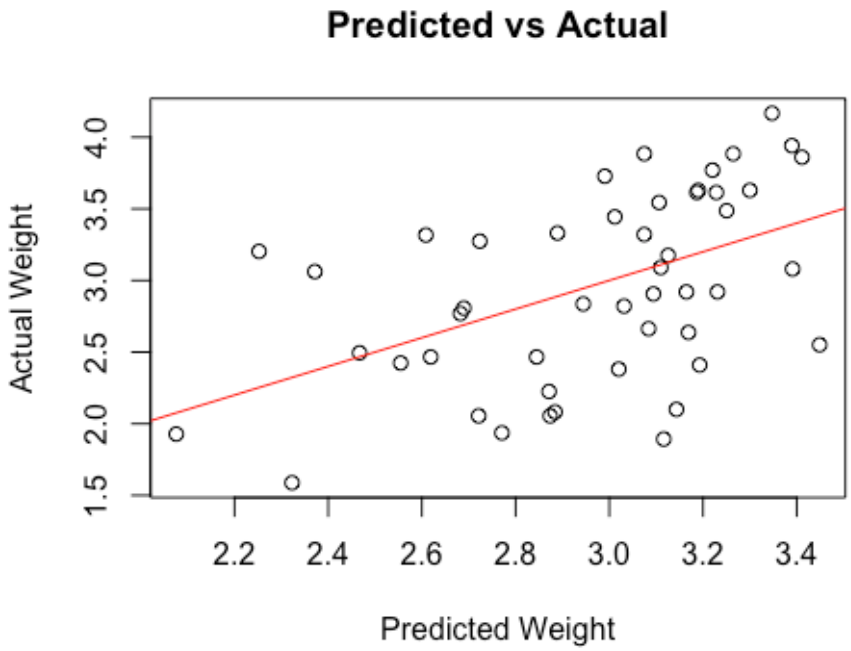
The lowest MSE obtained is 0.3214657 which is obtained when mother.weight is used as a predictor as a polynomial of degree 9. This can be confirmed from the following plot:



The plot above suggests that the lowest MSE is obtained when the degree is 9 and as the degree of the polynomial increases above 9, the MSE starts increasing. The MSE versus Degree of Polynomial plot is a U-shaped curve and clearly

shows that polynomials of degree more than 10 overfit to the train data and then produce poor predictions on the test data.

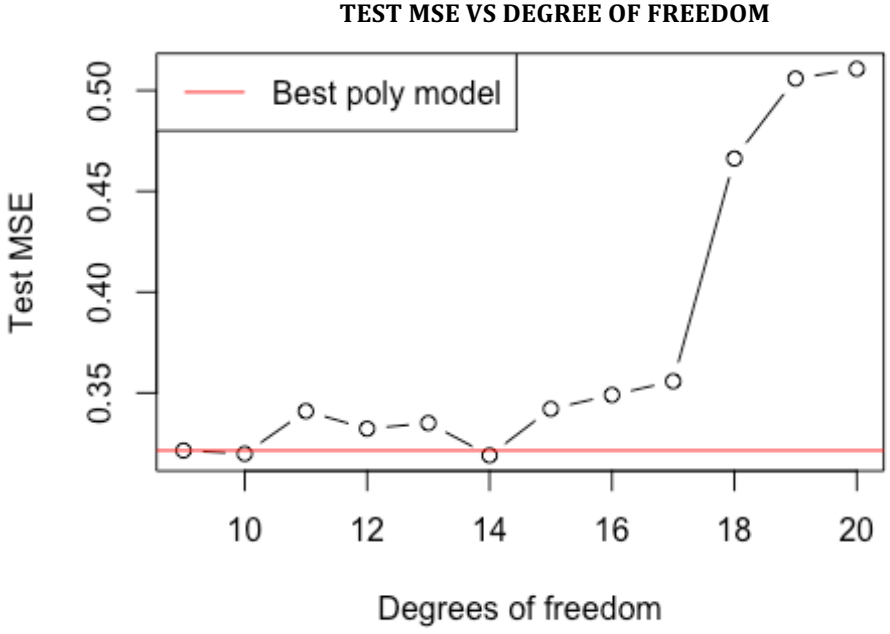
We now plot the predicted weight vs actual weight of the babies on the test data set for our best polynomial model:



If we compare this plot to the first plot of Predicted vs Actual baby weight, we can clearly see that this is a much better fitting model.

SPLINES ANALYSIS

Now that we have tried a lot of different polynomial regressions we can wonder if it is possible to improve our best polynomial model by introducing splines. Here we added in the regression formula several basis functions for the variable mother.weight. Between each knots we fit a $9 - degree - polynomial$. We tried different values for the number of degrees of freedom so as to find the best parameter. Here is the resulting plot:

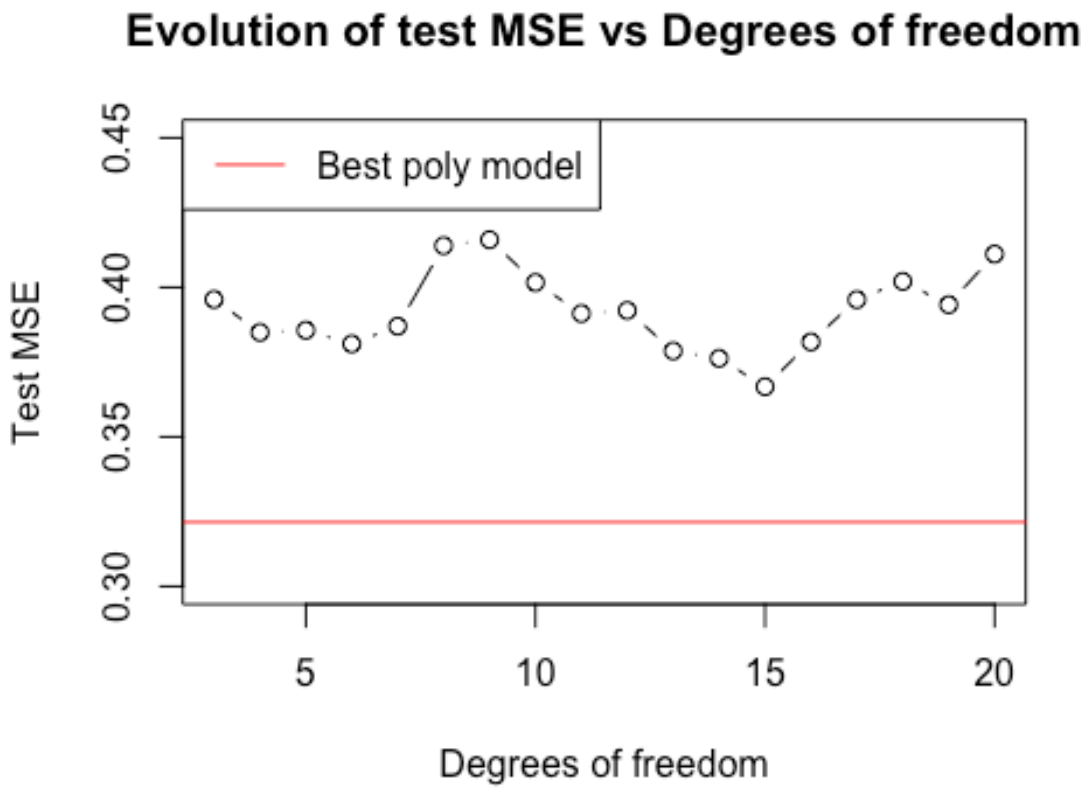


The minimum MSE is obtained when we have 14 degrees of freedom. With the R built-in function $bs()$, R automatically puts knots on the quantile values of the variable. Here for 14 degrees of freedom our knots are: $q_{16.7}$, $q_{33.3}$, q_{50}, $q_{66.7}$ and $q_{83.3}$. Thus between each quantile R fits a degree 9 polynomial on the mothers' weights. It also makes sure that the 1st, 2nd, ... and 8th derivatives are continuous at each knots. Thus the relation between the number of degrees of freedom d and the number of knots K is the following:

$$d = K + 9$$

We can see that this formula is verified in our case ($14 = 5 + 9$).

Natural Splines

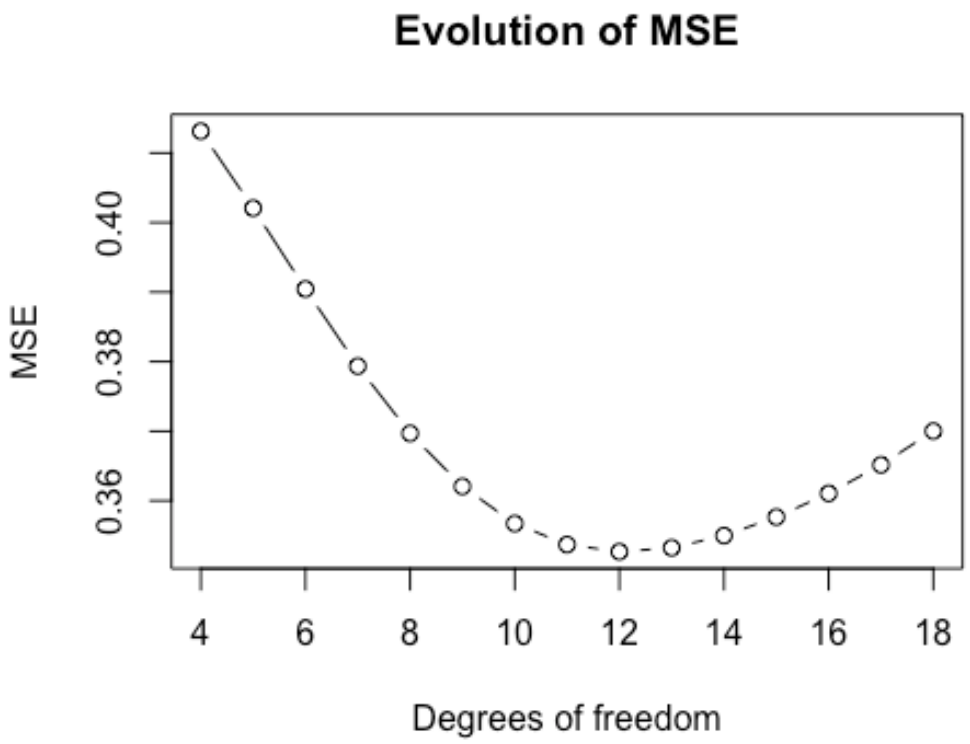


When we try with natural splines we have worse results than with the normal splines model. It is mainly due to the fact that R can only fit cubic natural splines, there is no degree argument in the R built-in function.

Now we can try to see if there is an improvement if we use smoothed splines. We have to use the General Additive Models R library to perform this analysis.

```
library(gam)

## Loaded gam 1.12
```

```
MSE = 4:18
for (k in 4:18){
  gam.fit = gam(baby.grams ~ prem.labor+uterine+ hypertension + smoke
          + s(mother.weight, k) + poly(mother.age, 2), data=bwt.grams.train)
  pred = predict(gam.fit, bwt.grams.test)
  mse = mean((pred-bwt.grams.test$baby.grams)^2)
  MSE[k-3] = mse
}
plot(4:18, MSE, type='b', main='Evolution of MSE', xlab='Degrees of freedom')
```



We can notice that the results are still not better than with our optimal model with degree 9 splines. The smoothing effect does not bring more predictive power to the final model. To conclude this part on splines we managed to find a model that outperforms slightly our best polynomial model. This was expected as splines models are more flexible than polynomial models. Nonetheless the improvement in test MSE is quite low and we can wonder if the splines model is really better than the polynomial model. Indeed, fitting a degree nine polynomial between each splines brings a lot of flexibility to the model but the increase of variance can be huge too. If we have had more observations we could have answered to this question by testing our models on a big test set. Nevertheless, we can run a ANOVA test to verify if the difference between our best polynomial model and our best splines model is really significant:

```
## Analysis of Variance Table
##
## Model 1: baby.grams ~ hypertension + uterine + smoke + prem.labor + poly(mother.age,
##     2) + poly(mother.weight, 9)
## Model 2: baby.grams ~ hypertension + uterine + smoke + prem.labor + bs(mother.weight,
##     df = 14, degree = 9) + poly(mother.age, 2)
##   Res.Df     RSS     Df   Sum of Sq      F         Pr(>F)
## 1    125  56.829
## 2    120  53.554     5     3.2754     1.4679     0.2054
```

We can see that the resulting p-value is roughly 0.21. Thus the p-value is really high so we may reject H_0: we can say that the difference of performance between those two models is not obvious. Thus we will maybe prefer to keep the less complex model *ie* the polynomial model.

Building Classification Model

The modeling approaches discussed above tried to use different combinations and transformations of the predictors available in the dataset to predict the exact weight of the newborn baby. None of the obtained models demonstrated solid quality results with respect to their MSE, that might suggest that these predictors are not enough to explain all the variance observed in the baby.grams response variable. However, the main goal of this research is to identify risk of giving birth to low-weight infant, which should be revealed during pregnancy period in order to be able to minimize this risk with appropriate medical involvement. For that we can reformulate our modeling problem as a classification problem, testing for threshold in dataset, which will split healthy infants from infants at risk, and fitting logistic regression on this binary outcome — *no* for healthy infants and *yes* for infants with low weight.

Testing for Classification Threshold

Conventional definition of low birth weight classifies a newborn infant of less than 2.5 kg as a low birth weight infant, and, as suggested by the recent studies the frequency of Low Birth Weight case occurrence is no more than 30%. Before we start modelling logistic regression on whether an infant will be born with normal or low weight, we need to test whether the dataset we are working on attributes the same frequency properties as the general population of such cases.

For this purposes we obtain bootstrapped estimate of the 30th percentile of baby.grams and compare it with 2.5 kg.

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = bwt.grams$baby.grams, statistic = boot.fn, R = 1000)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1*    2.495 0.0196722  0.07713608
```

The results of bootstrap test prove that 30th precentile estimate of baby.grams is equal to 2.495 and the threshold value of 2.5 kg that we are interested in falls into 95% confidence interval of this estimate [2.341; 2.649].

Fitting Logistic Regression

After we proved that the decision threshold for classification on this data can indeed be assumed to be equal to 2.5 kg, we now reshape our dataset to attribute this classification problem: response is now defined as a factor variable with level *no* if the weight is above 2.5 kg, and level *yes* if the weight is below this threshold. All the rest of the transformations remain the same.

```
##  below.2500   mother.age    mother.weight     race      smoke
##  yes: 59    Min.   :14.00   Min.   : 80.0   white:96   Mode :logical
##  no :130    1st Qu.:19.00   1st Qu.:110.0   black:26   FALSE:115
##             Median :23.00   Median :121.0   other:67   TRUE :74
##             Mean   :23.24   Mean   :129.8              NA's :0
##             3rd Qu.:26.00   3rd Qu.:140.0
##             Max.   :45.00   Max.   :250.0
##  prem.labor  hypertension    uterine       physician.visits
##  FALSE:159   Mode :logical   Mode :logical   0 :100
##  TRUE : 30   FALSE:177       FALSE:161       1 : 47
##              TRUE :12        TRUE :28        2+: 42
##              NA's :0         NA's :0
##
##
```

As the strength of relationships between different predictors and the weight of the infant was explored before, we will take only those predictors that were chosen by the best subset selection procedure while fitting linear models. Since the size of the dataset is relatively small, validating the model results is better be done with *k-fold* cross validation procedure. The optimal number of folds for this dataset was chosen before: *k = 6*.

```
## [1] 0.1971018 0.1947619
```

The logistic model produced quite good results with unbiased classification error of 0.191 after the cross-validation procedure.

The summary statistics of the model and analysis of deviance provided below, demonstrates that the choice of the predictors was appropriate for this model, since all of them, but uterine demonstrate p-values lower than 0.1 in *t-test* for individual significance and *chi-square test*, demonstrating that the model including this variable demonstrate statistically significant difference from *null* model.

```
## Call:  glm(formula = below.2500 ~ mother.weight + race + smoke + hypertension +
    prem.labor + uterine, family = binomial, data = bwt)
```

## Coefficients:	*Estimate*	*Std. Error*	*p-value < 0.05*
Intercept	0.125	0.968	No
Mother.weight	0.016	0.007	Yes
Raceblack	-1.301	0.528	Yes
Raceother	-0.854	0.441	Yes
SmokeTRUE	-0.867	0.404	Yes
HypertensionTRUE	-1.867	0.707	Yes
prem.laborTRUE	-1.129	0.450	yes
UterineTRUE	-0.751	0.459	No

```
## Residual deviance: 197.85  on 181  degrees of freedom
## AIC: 213.85
## Number of Fisher Scoring iterations: 4
```

```
## Analysis of Deviance Table

##
##              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                        188     234.67
## mother.weight  1   5.9813      187     228.69 0.014458 *
## race           2   5.4316      185     223.26 0.066153 .
## smoke          1   8.2444      184     215.01 0.004088 **
## hypertension   1   6.7672      183     208.25 0.009285 **
## prem.labor     1   7.7652      182     200.48 0.005326 **
## uterine        1   2.6307      181     197.85 0.104817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

However, the confusion matrix of this model reveals the following fact: it demonstrates solid prediction power classifying healthy infants *(10% classification error)*, however it fails to distinguish properly the low birth weight cases classifying *55.9%* of them inaccurately. This is the major drawback of this model, since it does not help identifying pregnancies with low birth weight risk, thus making timely medical intervention to support infant's and mother's health condition.

```
##              below 2.5 kg
## prediction yes  no
##       no       33  117
##       yes      26   13
```

Despite low predictive power for the cases of high low birth weight risk, this model gives an important inferential conclusion that the categorical factors that were picked for this model (smoking habits, hypertension, race, physician visits) include enough information to conclude that the infant will be born with a healthy weight, hence low risk of infant mortality. However, we should seek the relationships explaining low weight birth cases in other medical and demographical factors that were not collected for this research. In fact, what we are really interested in is to decrease the false negative rate (the number of children who are predicted healthy whereas they are in fact under 2.5 Kg) as most as possible without scarifying the overall accuracy. A good way to do this will be to change the value of the threshold used for the logistic function (equal to 0.5 by default). If we decrease it our model will do fewer false positive mistakes (note that the dummy model consisting in always predicting 'YES' will have a false negative rate equal to 0 but a very bad overall accuracy of roughly 30%...).

RESULTS AND CONCLUSION

The results of our models can be summarized as follows:

	Model	*Parameters*	*MSE*
Regression	Linear	Best subset (nb of predictors: 6)	0.47
		Cross validation + Ridge (Lambda = 0.15)	0.42
		Cross validation + Lasso (lambda = 0.02)	0.43
	Polynomial	Degree 3	0.39
		Degree 7	0.35
		Degree 9	0.321
	Splines	Splines (poly 9, 14 degrees of freedom)	0.319
		Natural splines (poly 3, 15 degrees of freedom)	0.37
		GAM (poly 3, 12 degrees of freedom)	0.35
Class.	Logistic regression	Bootstrap (decision threshold: 2.5 kg) Cross validation on best subset (nb of predictors: 7)	80.5% accuracy

The best regression model to predict the weight of a baby is the splines with 9 degrees of polynomial and 14 degrees of freedom. With this model, we obtain a Mean squared error 0.321 on the test set.

The best classification model to predict if a baby's weight will be below the threshold of 2.5 kilograms is a logistic regression using cross validation, best subset and bootstrap. With this model, we obtain an accuracy 80.5% on the test set.