Math 189: Homework 6 Solution

Introduction

In this assignment we are studying whether it is possible to predict smoking status from various characteristics of a baby, such as birthweight and gestation. If the result of smoking is certain deformities or abnormalities, then we may be able to trace backwards this relationship, and on the basis of observed baby characteristics determine whether the mother was smoking. This could be tricky, because there may be other factors, other than smoking, which produce deformities.

But if the probability of being a smoker is high based on the baby's data, then we'd have grounds for further investigation, and potentially an intervention by government health services to offer counseling and addiction assistance programs. This could be useful in a cultural climate where there is now stigma associated with cigarette smoking, leading to under-reporting of the activity in self-response surveys.

Metadata for *babies.dat*

The following variables are measured in the dataset:

- **bwt**: Baby's weight at birth, to the nearest ounce
- **gestation**: Duration of the pregnancy in days, calculated from the first day of the last normal menstrual period.
- **parity**: Indicator for whether the baby is the first born (1) or not (0).
- **age**: Mother's age at the time of conception, in years
- **height**: Height of the mother, in inches
- **weight**: Mother's prepregnancy weight, in pounds
- **smoking Indicator**: for whether the mother smokes (1) or not (0); (9) denotes unknown.

```
baby <- read.table("babies.dat",header=TRUE)
head(baby)
```

```
##   bwt gestation parity age height weight smoke
## 1 120       284      0  27     62    100     0
## 2 113       282      0  33     64    135     0
## 3 128       279      0  28     64    115     1
## 4 123       999      0  36     69    190     0
## 5 108       282      0  23     67    125     1
## 6 136       286      0  25     62     93     0
```

Data Cleaning

It seems that some of the baby data has a value of 9 for the smoking indicator, which is a flag for where the status was unknown. We remove these records (we can later try to predict them too).

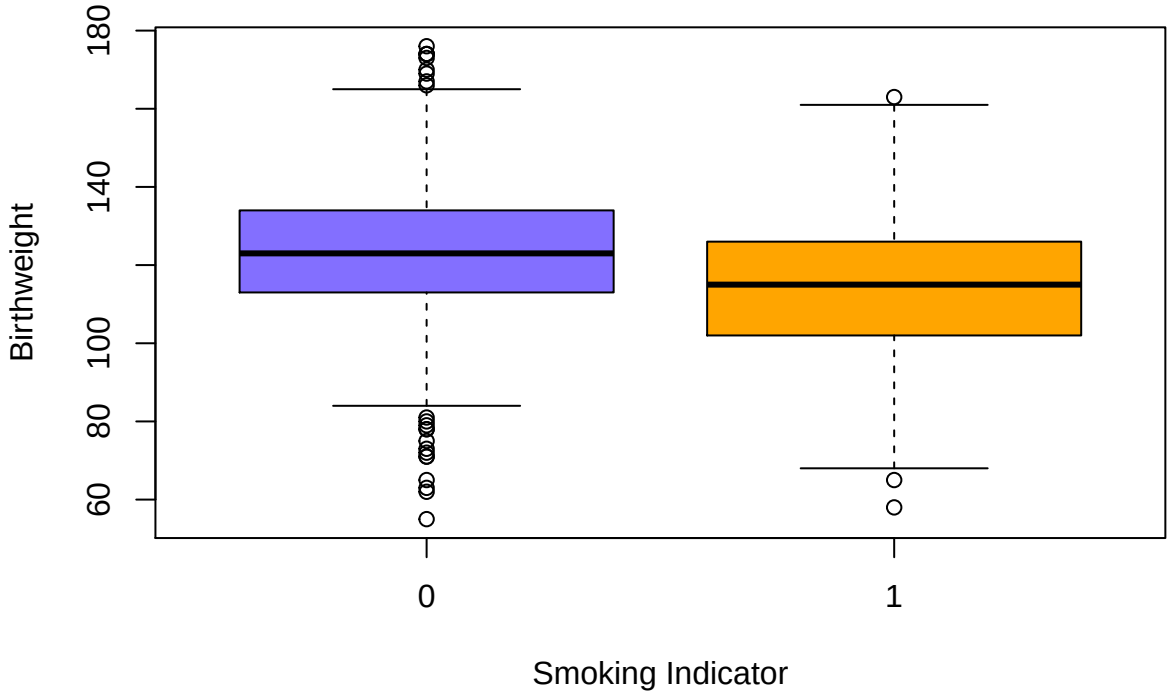
```
baby.clean <- baby[baby$smoke != 9,]
```

Exploring Associations

Here we utilize the boxplot techniques used in class lectures, exploring the relationship between smoking status and the other variables.

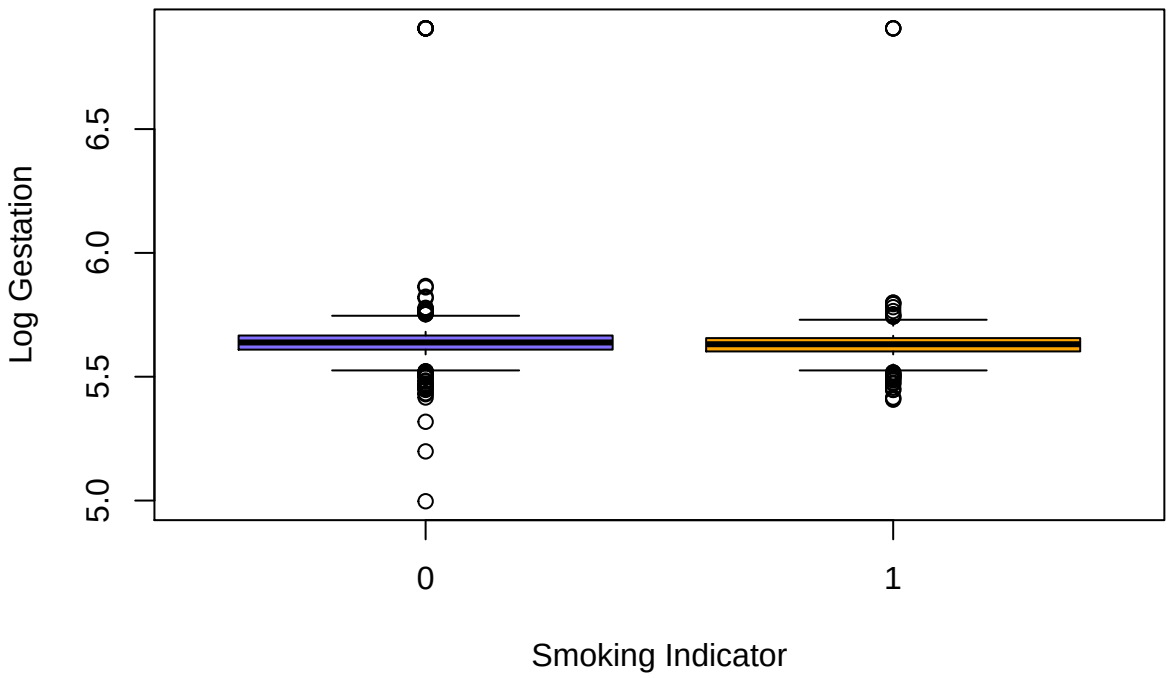
- First we examine birthweight by smoking status, and find there seems to be an overall drop due to smoking.

```
boxplot(baby.clean$bwt~baby.clean$smoke,xlab="Smoking Indicator",
        ylab="Birthweight",col=c("slateblue1","orange"))
```



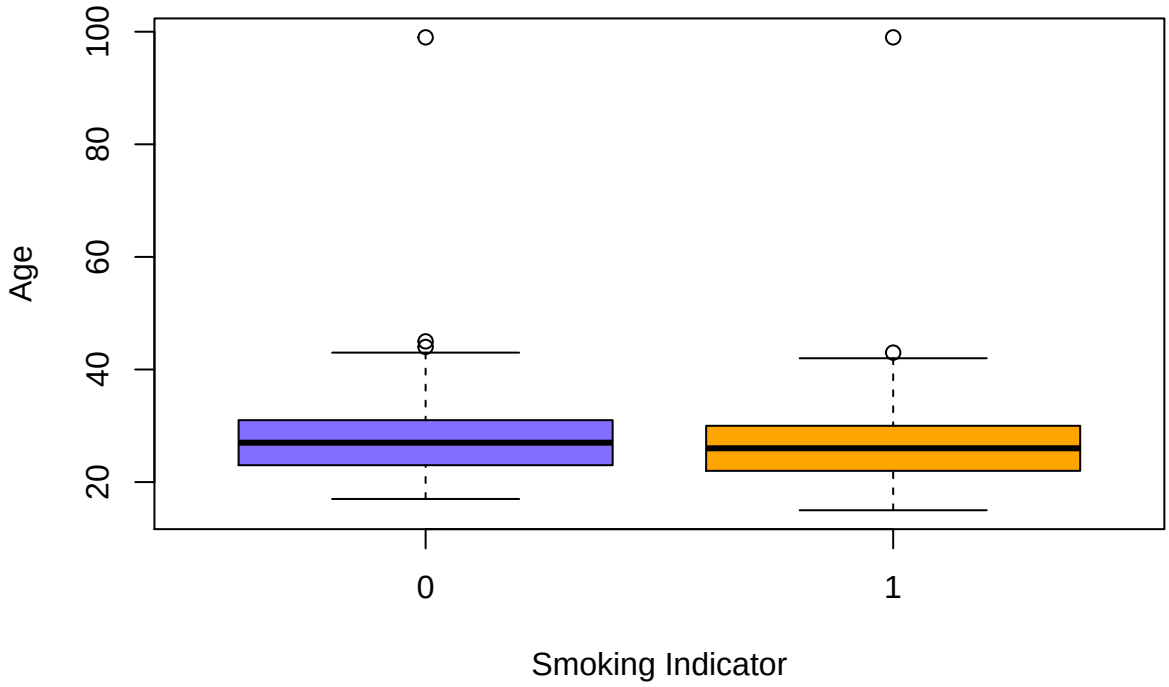
- Next we examine gestation. The relationship is impacted by the presence of a few outliers, which are probably data entry errors. We could clean these out, but instead examine gestation in log scale, which will automatically shrink the impact of the outliers. Overall, smoking seems to exert slight downard pressure on log gestation.

```
boxplot(log(baby.clean$gestation)~baby.clean$smoke,xlab="Smoking Indicator",
        ylab="Log Gestation",col=c("slateblue1","orange"))
```

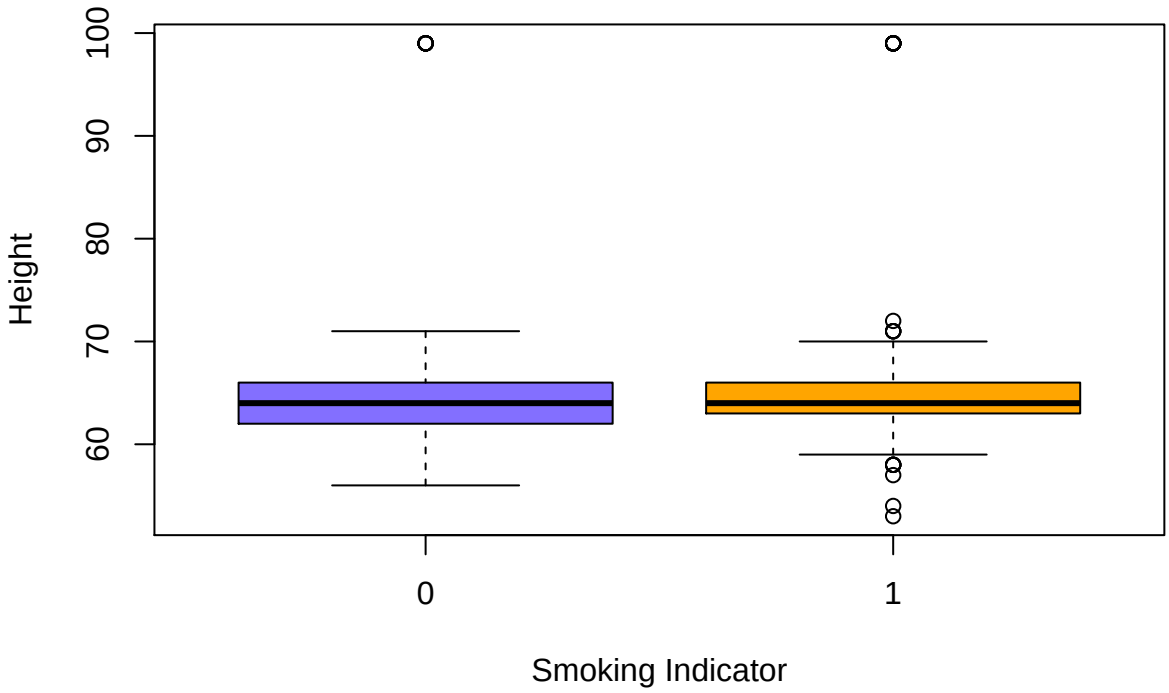


- The other variables seem to be less useful, especially as age, height, and weight are characteristics of the *mother*. To be thorough, we consider each of these. There are some more absurd outliers (99-year old mothers), which we could remove; however, there seems little point. The boxplots indicate these variables have little impact on smoking status.

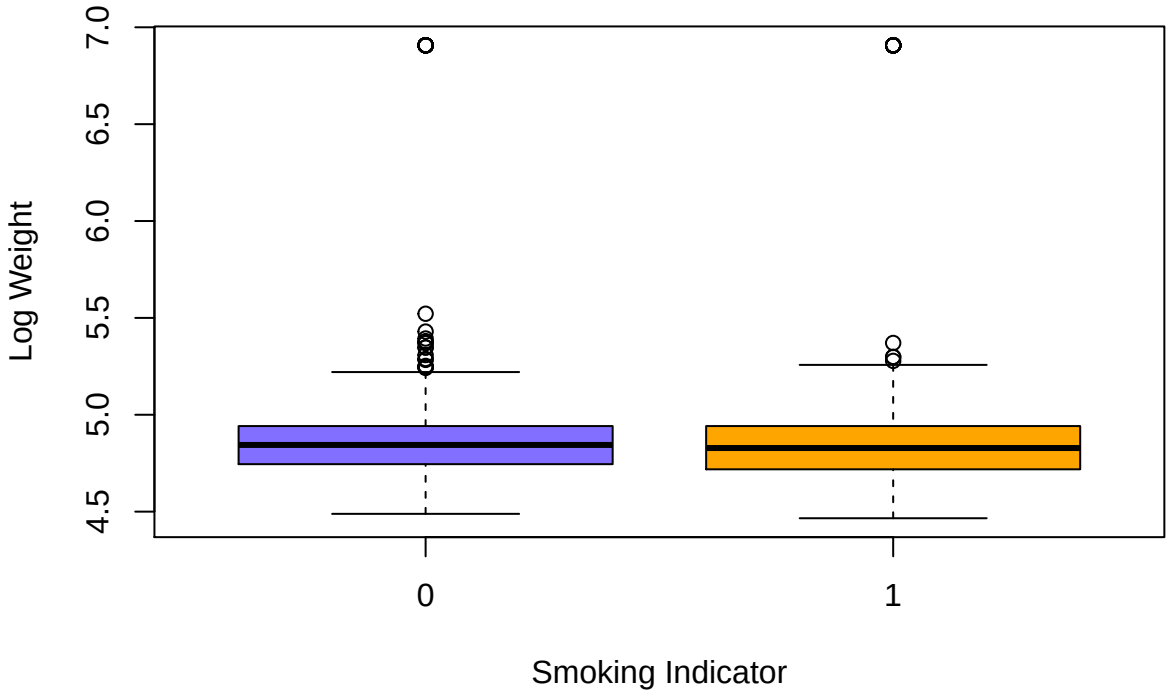
```
boxplot(baby.clean$age~baby.clean$smoke,xlab="Smoking Indicator",
        ylab="Age",col=c("slateblue1","orange"))
```




```
boxplot(baby.clean$height~baby.clean$smoke,xlab="Smoking Indicator",
        ylab="Height",col=c("slateblue1","orange"))
```




```
boxplot(log(baby.clean$weight)~baby.clean$smoke,xlab="Smoking Indicator",
        ylab="Log Weight",col=c("slateblue1","orange"))
```



- To consider parity, which is also binary, we instead use a classification table.

```
table <- c(dim(baby.clean[baby.clean$smoke == 0 & baby.clean$parity == 0,])[1],
           dim(baby.clean[baby.clean$smoke == 1 & baby.clean$parity == 0,])[1])
table <- rbind(table,
         c(dim(baby.clean[baby.clean$smoke == 0 & baby.clean$parity == 1,])[1],
           dim(baby.clean[baby.clean$smoke == 1 & baby.clean$parity == 1,])[1]))
rownames(table) <- c("Parity 0","Parity 1")
colnames(table) <- c("Smoker 0","Smoker 1")
table
```

```
##          Smoker 0 Smoker 1
## Parity 0      548      363
## Parity 1      194      121
```

- The probability of being a smoker, conditional on parity being zero, is estimated by 0.75. The unconditional probability of being a smoker is estimated as 0.3947798. Since these are fairly different, we are inclined to include parity in our model.

Splitting the Data

Using the cleaned data, we split into roughly an 80%, 20% split into training and test data. This gives a fairly large training set, but we still have over 300 observations in the test set. These ratios seem reasonable for our objectives.

```
baby.train <- baby.clean[1:980,c(1,2,3,7)]
baby.test <- baby.clean[981:1226,c(1,2,3,7)]
```

Logistic Regression Model

- We now fit the logistic model, using bwt, gestation, and parity as predictors.

```
library(ISLR)
all.fit <- glm(smoke~bwt+gestation+parity,data=baby.train,family=binomial)
summary(all.fit)
```

```
##
## Call:
## glm(formula = smoke ~ bwt + gestation + parity, family = binomial,
##     data = baby.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8105  -0.9928  -0.8145   1.2341   1.9546
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.0345436  0.5335114   5.688 1.29e-08 ***
## bwt         -0.0276700  0.0040217  -6.880 5.98e-12 ***
## gestation   -0.0004398  0.0008477  -0.519    0.604
## parity      -0.1471617  0.1562930  -0.942    0.346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1314.9  on 979  degrees of freedom
## Residual deviance: 1262.1  on 976  degrees of freedom
## AIC: 1270.1
##
## Number of Fisher Scoring iterations: 4
```

- We can dump gestation and parity, as their Z-statistics are not significant.

```
best.fit <- glm(smoke~bwt,data=baby.train,family=binomial)
summary(best.fit)
```

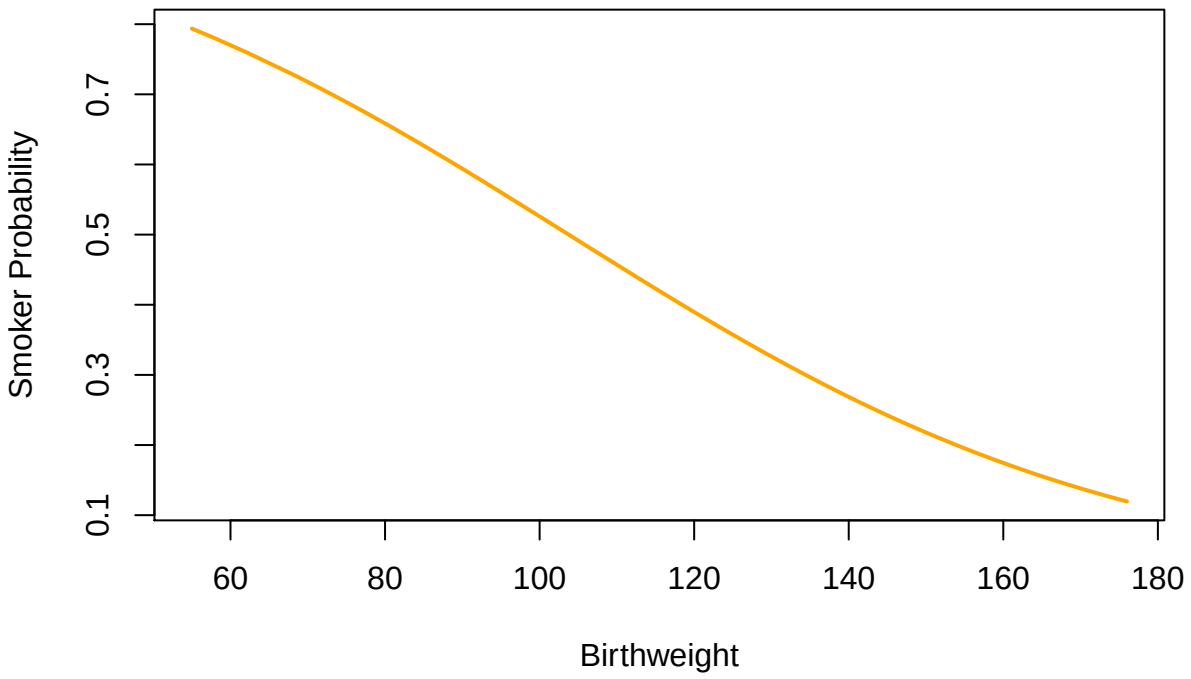
```
##
## Call:
## glm(formula = smoke ~ bwt, family = binomial, data = baby.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7769  -0.9937  -0.8099   1.2395   1.9052
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.867690   0.479839   5.976 2.28e-09 ***
## bwt         -0.027637   0.004009  -6.894 5.42e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1314.9  on 979  degrees of freedom
## Residual deviance: 1263.2  on 978  degrees of freedom
## AIC: 1267.2
##
## Number of Fisher Scoring iterations: 4
```

- We can examine the probability of smoker status based on the training data. Lower birthweight corresponds to a higher likelihood of being a smoker.

```
fitted.prob <- as.numeric(best.fit$fitted.values)
prob.smoke <- fitted.prob
bwt.smoke <- baby.train$bwt
M1 <- matrix(c(prob.smoke,bwt.smoke), ncol = 2)
M1 <- M1[order(M1[,1], decreasing = FALSE),]

plot(M1[,2], M1[,1],type='l',lwd=2,col="orange",
     xlab="Birthweight",ylab="Smoker Probability")
```



Classifying the Test Data

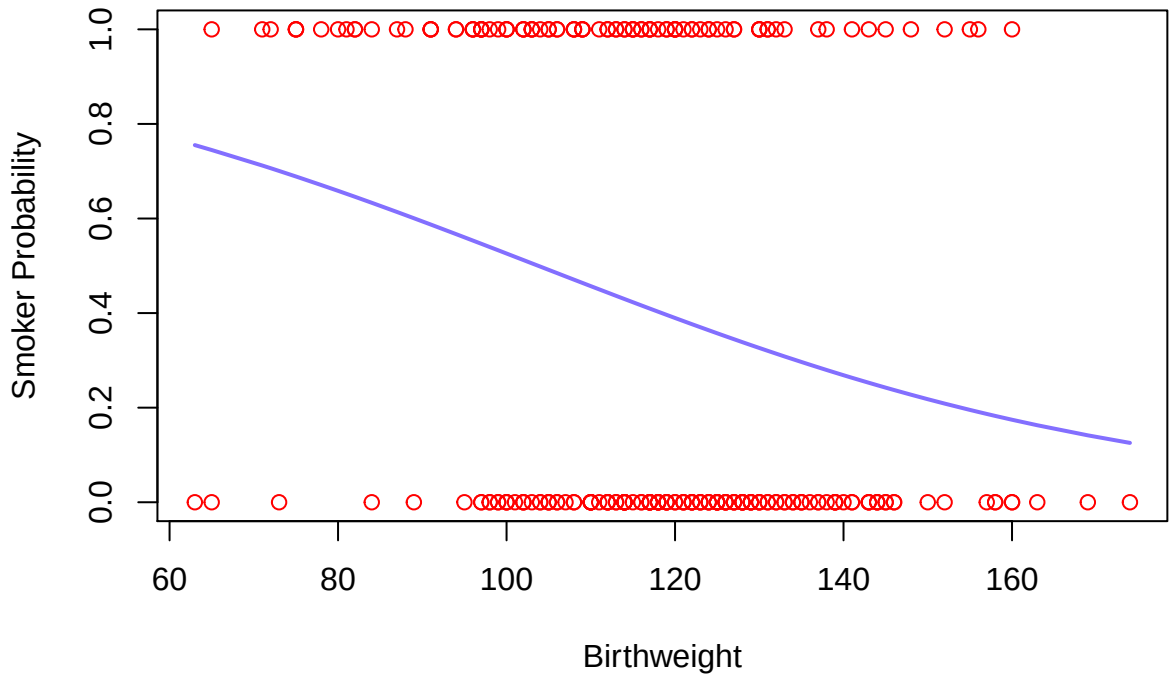
- Now for each record in the test data, we can compute the probability of smoker status as a function of Birthweight.

```
pred_all <- function(obs){
  x <- c(1,obs)
  pred <- as.numeric(as.numeric(x %*% best.fit$coefficients))
  pred <- 1/(1+exp(-pred))
  return(pred)
}
```

- We generate the probabilities, which are similar to the curve obtained on the training data. We also juxtapose the true categories.

```
prob.smoke <- c()
for(i in 1:dim(baby.test)[1])
{
  prob.smoke <- c(prob.smoke,pred_all(baby.test$bwt[i]))
}
M2 <- cbind(prob.smoke,baby.test$bwt,baby.test$smoke)
M2 <- M2[order(M2[,1], decreasing = FALSE),]

plot(M2[,2], M2[,1],type='l',lwd=2,col="slateblue1",
     xlab="Birthweight",ylab="Smoker Probability",ylim=c(0,1))
points(M2[,2],M2[,3],col=2)
```



- We need to assess performance. Let's consider any probability over .5 as corresponding to smoker, and any less than .5 as a non-smoker.

```
table <- c(dim(baby.test[baby.test$smoke == 0 & prob.smoke < .5,])[1],
           dim(baby.test[baby.test$smoke == 1 & prob.smoke < .5,])[1])
table <- rbind(table,
         c(dim(baby.test[baby.test$smoke == 0 & prob.smoke >= .5,])[1],
           dim(baby.test[baby.test$smoke == 1 & prob.smoke >= .5,])[1]))
rownames(table) <- c("Predict 0","Predict 1")
colnames(table) <- c("Smoker 0","Smoker 1")
table
```

```
##           Smoker 0 Smoker 1
## Predict 0      129       60
## Predict 1       20       37
```

- Success rate is 0.6747967. This is better than 50%, but there are enough errors to raise concern with this model.

Conclusion

We have built a predictive logistic regression model for smoking status on the basis of the Birthweight variable. Although some of the other variables, including gestation and parity, appeared to show promise initially, the turned out to be insignificant. After cleaning the date, we fit our model to the training data, and applied the model to our test data. Using a 50% threshold probability for classification, the overall success rate was 0.6747967. This indicated the model can be useful to do prediction, but makes enough Type I and II erors so as to be dubious for use in policy.

It is important to recall that correlation is not the same as causation. Similarly, we can build predictive models where the causal order is actually inverted; this may work statistically, even if it does not make sense scientifically. Here we have explored just such an exercise: although maternal smoking is causally linked to birth defects, here we have used baby Birthweight as a predictor for smoking status. We cannot, of course, conclude that low Birthweight causes a mother to be more likely to smoke, and yet it can be used as a predictor of smoking status in a statistical model.