Math 189: Homework 6

Math 189 Homework 6

Introduction

In this learning assignment, we used data surrounding child-birth and whether smoking impacts baby weight and gestation period. However, for this exercise we flipped the idea, and tried to see if we could predict whether or not the mother smoked based on factors from the baby and mother. After cleaning up the data and determining the suitable variables for prediction, we split the data into a training and test set and developed a logistic regression model to do the classification. We then determined the accuracy of the test, and graphed the results. The data can be found on GitHub.

Data

The data is taken from The Child Health and Development Studies (CHDS) data are presented in Stat Labs: Mathematical Statistics Through Applications by Deborah Nolan and Terry Speed (Springer). The variables that we will be looking at is as described below.

- **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.

```r
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
```

Methods & Analysis

1. Explore the data graphically in order to investigate the association between the **smoking indicator** and bwt, gestation, or *any of the other variables* that seem appropriate. Which of the other features seem most likely to be useful in predicting smoker status?

```r
any(is.na(baby))
```

```
## [1] FALSE
```

```r
summary(baby)
```

```
##       bwt          gestation         parity            age
##  Min.   : 55.0   Min.   :148.0   Min.   :0.0000   Min.   :15.00
##  1st Qu.:108.8   1st Qu.:272.0   1st Qu.:0.0000   1st Qu.:23.00
##  Median :120.0   Median :280.0   Median :0.0000   Median :26.00
##  Mean   :119.6   Mean   :286.9   Mean   :0.2549   Mean   :27.37
##  3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.:1.0000   3rd Qu.:31.00
##  Max.   :176.0   Max.   :999.0   Max.   :1.0000   Max.   :99.00
##      height          weight         smoke
##  Min.   :53.00   Min.   : 87   Min.   :0.0000
##  1st Qu.:62.00   1st Qu.:115   1st Qu.:0.0000
##  Median :64.00   Median :126   Median :0.0000
##  Mean   :64.67   Mean   :154   Mean   :0.4644
##  3rd Qu.:66.00   3rd Qu.:140   3rd Qu.:1.0000
##  Max.   :99.00   Max.   :999   Max.   :9.0000
```

It can be seen that the maximum for gestation is 999.0, age is 99.0, height is 99.0, weight is 999, and smoke is 9. These can definitely be seen as outliers. Therefore we need to clean up the data.

We are first going to sort the data such that the outlier values are the first values

```r
sort(baby$gestation, decreasing = TRUE)
```

```
##    [1] 999 999 999 999 999 999 999 999 999 999 999 999 999 353 351 338 336 330
##   [19] 330 329 328 324 323 323 321 320 319 319 318 318 318 318 316 316 315 315
##   [37] 315 314 313 313 313 312 312 311 310 309 308 308 308 308 307 307 307 306
##   [55] 306 306 306 306 306 305 305 305 305 304 304 304 303 303 303 303 303 302
##   [73] 302 302 302 302 302 302 302 302 302 301 301 301 301 301 300 300 300 300
##   [91] 300 300 300 300 300 299 299 299 299 299 299 299 299 299 298 298 298 298
##  [109] 298 298 298 298 298 297 297 297 297 297 297 297 297 297 297 297 297 296
##  [127] 296 296 296 296 296 296 296 296 296 296 295 295 295 295 295 295 295 295
##  [145] 295 295 295 295 295 294 294 294 294 294 294 294 294 294 294 294 294 294
##  [163] 294 294 294 294 293 293 293 293 293 293 293 293 293 293 293 293 293 293
##  [181] 293 293 293 293 293 293 293 293 293 293 293 292 292 292 292 292 292 292
##  [199] 292 292 292 292 292 292 292 292 292 292 292 292 292 292 292 292 292 292
##  [217] 292 292 292 292 291 291 291 291 291 291 291 291 291 291 291 291 291 291
##  [235] 291 291 291 291 291 291 291 291 291 290 290 290 290 290 290 290 290 290
##  [253] 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##  [271] 290 290 290 290 289 289 289 289 289 289 289 289 289 289 289 289 289 289
##  [289] 289 289 289 289 289 289 289 289 289 289 289 289 289 288 288 288 288 288
##  [307] 288 288 288 288 288 288 288 288 288 288 288 288 288 288 288 288 288 288
##  [325] 288 288 288 288 288 288 288 288 288 288 288 287 287 287 287 287 287 287
##  [343] 287 287 287 287 287 287 287 287 287 287 287 287 287 286 286 286 286 286
##  [361] 286 286 286 286 286 286 286 286 286 286 286 286 286 286 286 286 286 286
##  [379] 286 286 286 286 286 286 286 286 286 286 286 286 286 286 286 286 286 286
##  [397] 285 285 285 285 285 285 285 285 285 285 285 285 285 285 285 285 285 285
##  [415] 285 285 285 285 285 285 285 285 285 285 285 285 285 285 285 285 285 285
```

```
##  [433] 285 285 284 284 284 284 284 284 284 284 284 284 284 284 284 284 284 284
##  [451] 284 284 284 284 284 284 284 284 284 284 284 284 284 284 284 284 284 284
##  [469] 284 284 284 284 284 284 283 283 283 283 283 283 283 283 283 283 283 283
##  [487] 283 283 283 283 283 283 283 283 283 283 283 283 283 283 283 283 283 283
##  [505] 283 283 283 283 283 283 283 283 283 282 282 282 282 282 282 282 282 282
##  [523] 282 282 282 282 282 282 282 282 282 282 282 282 282 282 282 282 282 282
##  [541] 282 282 282 282 282 282 282 282 282 282 282 282 282 282 282 282 282 282
##  [559] 282 282 282 281 281 281 281 281 281 281 281 281 281 281 281 281 281 281
##  [577] 281 281 281 281 281 281 281 281 281 281 281 281 281 281 281 281 281 281
##  [595] 281 281 281 281 281 281 281 281 281 280 280 280 280 280 280 280 280 280
##  [613] 280 280 280 280 280 280 280 280 280 280 280 280 280 280 280 280 280 280
##  [631] 280 280 280 280 280 280 280 280 280 280 280 280 280 280 280 280 280 279
##  [649] 279 279 279 279 279 279 279 279 279 279 279 279 279 279 279 279 279 279
##  [667] 279 279 279 279 279 279 279 279 279 279 279 279 279 279 279 279 279 279
##  [685] 278 278 278 278 278 278 278 278 278 278 278 278 278 278 278 278 278 278
##  [703] 278 278 278 278 278 278 278 278 278 278 278 278 278 278 278 278 278 278
##  [721] 278 278 278 278 278 278 278 278 277 277 277 277 277 277 277 277 277 277
##  [739] 277 277 277 277 277 277 277 277 277 277 277 277 277 277 277 277 277 277
##  [757] 277 277 277 277 277 277 277 277 277 277 277 277 277 277 276 276 276 276
##  [775] 276 276 276 276 276 276 276 276 276 276 276 276 276 276 276 276 276 276
##  [793] 276 276 276 276 276 276 276 276 276 276 276 276 276 276 276 276 276 275
##  [811] 275 275 275 275 275 275 275 275 275 275 275 275 275 275 275 275 275 275
##  [829] 275 275 275 275 275 275 275 275 275 275 275 275 275 275 275 275 275 275
##  [847] 275 275 275 274 274 274 274 274 274 274 274 274 274 274 274 274 274 274
##  [865] 274 274 274 274 274 274 274 274 274 274 274 274 274 274 274 274 274 274
##  [883] 274 274 274 274 274 273 273 273 273 273 273 273 273 273 273 273 273 273
##  [901] 273 273 273 273 273 273 273 273 273 273 273 273 273 273 273 273 273 273
##  [919] 273 273 273 273 273 273 272 272 272 272 272 272 272 272 272 272 272 272
##  [937] 272 272 272 272 272 272 272 272 272 272 272 272 271 271 271 271 271 271
##  [955] 271 271 271 271 271 271 271 271 271 271 271 271 271 271 271 271 271 271
##  [973] 271 271 270 270 270 270 270 270 270 270 270 270 270 270 270 270 270 270
##  [991] 270 270 270 270 270 270 270 270 270 270 270 270 270 270 270 270 270 270
## [1009] 269 269 269 269 269 269 269 269 269 269 269 269 269 269 269 269 269 269
## [1027] 268 268 268 268 268 268 268 268 268 268 268 268 268 268 268 268 268 268
## [1045] 268 268 268 268 267 267 267 267 267 267 267 267 267 267 267 267 267 267
## [1063] 267 267 267 267 267 267 266 266 266 266 266 266 266 266 266 266 266 266
## [1081] 266 266 266 265 265 265 265 265 265 265 265 265 265 265 265 264 264 264
## [1099] 264 264 264 264 264 264 264 263 263 263 262 262 262 262 262 262 262 262
## [1117] 262 262 262 262 261 261 261 261 261 261 261 260 260 260 260 260 260 260
## [1135] 260 260 259 259 259 258 258 258 258 258 258 257 257 257 257 256 256 256
## [1153] 256 256 256 256 255 255 255 255 255 255 255 255 255 254 254 254 254 254
## [1171] 254 253 252 252 252 252 252 251 251 251 250 250 250 249 249 249 249 248
## [1189] 248 248 247 247 247 246 246 246 246 246 246 245 245 245 245 244 244 244
## [1207] 243 242 242 241 241 240 239 238 238 238 237 237 236 235 234 234 234 233
## [1225] 232 232 232 229 228 225 225 224 223 204 181 148
```

```r
sort(baby$age, decreasing = TRUE)
```

```
##    [1] 99 99 45 44 43 43 43 43 43 43 42 42 42 42 41 41 41 41 41 41 41 41 41 41
##   [25] 41 41 41 41 40 40 40 40 40 40 40 40 40 40 40 39 39 39 39 39 39 39 39 39
##   [49] 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 38 38 38 38 38 38 38 38 38
##   [73] 38 38 38 38 38 38 38 38 38 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37
##   [97] 37 37 37 37 37 37 37 37 37 37 37 37 37 37 36 36 36 36 36 36 36 36 36 36
##  [121] 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 35 35 35 35 35 35 35 35
```

```
##  [145] 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 34 34
##  [169] 34 34 34 34 34 34 34 34 34 34 34 34 34 34 34 34 34 34 34 34 34 34 34 34
##  [193] 34 34 34 34 34 34 34 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33
##  [217] 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33
##  [241] 33 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32
##  [265] 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 31 31 31 31 31 31 31 31
##  [289] 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31
##  [313] 31 31 31 31 31 31 31 31 31 31 31 31 31 31 30 30 30 30 30 30 30 30 30 30
##  [337] 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30
##  [361] 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30
##  [385] 30 30 30 30 30 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29
##  [409] 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29
##  [433] 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 28
##  [457] 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28
##  [481] 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28
##  [505] 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 27 27 27
##  [529] 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27
##  [553] 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27
##  [577] 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27
##  [601] 27 27 27 27 27 27 27 27 27 27 26 26 26 26 26 26 26 26 26 26 26 26 26 26
##  [625] 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26
##  [649] 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26
##  [673] 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26
##  [697] 26 26 26 26 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
##  [721] 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
##  [745] 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
##  [769] 25 25 25 25 25 25 25 25 25 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24
##  [793] 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24
##  [817] 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24
##  [841] 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 23
##  [865] 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23
##  [889] 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23
##  [913] 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23
##  [937] 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 22 22 22 22
##  [961] 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22
##  [985] 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22
## [1009] 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22
## [1033] 22 22 22 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21
## [1057] 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21
## [1081] 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 20 20
## [1105] 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
## [1129] 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
## [1153] 20 20 20 20 20 20 20 20 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19
## [1177] 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19
## [1201] 19 19 19 19 19 19 19 19 19 19 19 19 19 18 18 18 18 18 18 18 18 18 18 18
## [1225] 18 18 18 18 17 17 17 17 17 17 17 15
```

```r
sort(baby$weight, decreasing = TRUE)
```

```
##    [1] 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999
##   [19] 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999 999
##   [37] 250 228 220 217 215 215 215 210 210 202 200 200 198 197 196 192 191 190
##   [55] 190 190 190 189 185 185 185 185 182 182 181 180 180 180 180 180 180 180
##   [73] 178 178 177 177 176 175 175 175 175 175 175 175 175 174 171 170 170 170
```

```
##   [91] 170 170 170 170 170 170 170 169 169 168 165 165 165 165 165 165 165 165
##  [109] 164 164 163 162 162 162 162 160 160 160 160 160 160 160 160 160 160 160
##  [127] 160 160 160 160 160 159 159 159 159 158 157 157 157 156 156 156 156 156
##  [145] 155 155 155 155 155 155 155 155 155 155 155 155 155 155 155 155 155 155
##  [163] 154 154 154 154 154 153 153 152 152 151 150 150 150 150 150 150 150 150
##  [181] 150 150 150 150 150 150 150 150 150 150 150 150 150 150 150 150 150 150
##  [199] 150 150 150 150 150 149 149 149 149 148 148 148 148 148 148 148 148 148
##  [217] 147 147 147 147 147 147 147 147 147 146 146 146 146 145 145 145 145 145
##  [235] 145 145 145 145 145 145 145 145 145 145 145 145 145 145 145 145 145 145
##  [253] 145 145 145 145 145 145 145 145 145 145 145 145 145 145 145 145 145 145
##  [271] 145 144 144 144 144 143 143 143 143 143 143 142 142 142 142 142 142 142
##  [289] 142 142 142 141 141 140 140 140 140 140 140 140 140 140 140 140 140 140
##  [307] 140 140 140 140 140 140 140 140 140 140 140 140 140 140 140 140 140 140
##  [325] 140 140 140 140 140 140 140 140 140 140 139 139 139 139 138 138 138 138
##  [343] 138 138 138 138 137 137 137 137 137 137 137 137 137 137 137 137 137 137
##  [361] 137 136 136 136 136 136 136 136 136 136 136 136 135 135 135 135 135 135
##  [379] 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135
##  [397] 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135
##  [415] 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135 135
##  [433] 135 135 134 134 134 134 134 134 134 134 134 133 133 133 133 133 133 133
##  [451] 133 133 132 132 132 132 132 132 132 132 132 132 132 132 132 132 132 132
##  [469] 132 132 132 132 132 132 132 131 131 131 130 130 130 130 130 130 130 130
##  [487] 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130
##  [505] 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130
##  [523] 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130
##  [541] 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 130 129 129
##  [559] 129 129 129 129 129 129 129 129 128 128 128 128 128 128 128 128 128 128
##  [577] 128 128 128 128 128 128 128 128 127 127 127 127 127 127 127 127 127 127
##  [595] 127 127 127 127 127 127 127 127 127 127 127 127 127 127 127 126 126 126
##  [613] 126 126 126 126 126 126 126 126 126 126 126 126 126 125 125 125 125 125
##  [631] 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125
##  [649] 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125
##  [667] 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125 125
##  [685] 125 125 125 125 125 125 125 125 125 124 124 124 124 124 124 124 124 124
##  [703] 124 124 124 124 124 124 124 124 124 124 124 124 123 123 123 123 123 123
##  [721] 123 123 123 123 123 123 123 123 123 122 122 122 122 122 122 122 122 122
##  [739] 122 122 122 122 122 122 122 122 122 122 122 122 122 121 121 121 121 121
##  [757] 121 121 121 121 120 120 120 120 120 120 120 120 120 120 120 120 120 120
##  [775] 120 120 120 120 120 120 120 120 120 120 120 120 120 120 120 120 120 120
##  [793] 120 120 120 120 120 120 120 120 120 120 120 120 120 120 120 120 120 120
##  [811] 120 120 120 120 120 119 119 119 119 119 119 119 119 119 119 118 118 118
##  [829] 118 118 118 118 118 118 118 118 118 118 118 118 118 118 118 118 118 118
##  [847] 118 118 118 118 118 118 118 118 118 118 118 117 117 117 117 117 117 117
##  [865] 117 117 117 117 117 117 117 117 117 117 117 117 117 117 116 116 116 116
##  [883] 116 116 116 116 116 116 116 116 116 116 116 116 115 115 115 115 115 115
##  [901] 115 115 115 115 115 115 115 115 115 115 115 115 115 115 115 115 115 115
##  [919] 115 115 115 115 115 115 115 115 115 115 115 115 115 115 115 115 115 115
##  [937] 114 114 114 114 114 114 114 114 114 114 113 113 113 113 113 113 113 113
##  [955] 113 113 113 113 113 113 112 112 112 112 112 112 112 112 112 112 112 112
##  [973] 112 112 112 112 112 112 112 112 112 112 112 112 112 112 112 112 112 111
##  [991] 111 111 111 111 111 111 111 111 110 110 110 110 110 110 110 110 110 110
## [1009] 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110
## [1027] 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110
## [1045] 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110 110
```

```
## [1063] 110 110 109 109 109 109 109 109 109 109 109 109 108 108 108 108 108 108
## [1081] 108 108 108 108 108 108 108 108 108 108 107 107 107 107 107 107 107 107
## [1099] 107 107 107 107 107 107 107 106 106 106 106 106 106 106 105 105 105 105
## [1117] 105 105 105 105 105 105 105 105 105 105 105 105 105 105 105 105 105 105
## [1135] 105 105 105 105 105 104 104 104 104 104 104 104 104 104 104 104 104 104
## [1153] 103 103 103 103 103 103 103 103 103 103 103 103 103 103 103 102 102 102
## [1171] 102 102 102 102 102 102 102 101 101 101 100 100 100 100 100 100 100 100
## [1189] 100 100 100 100 100 100 100 100 100 100  99  99  99  99  99  99  98  98
## [1207]  98  98  98  98  97  97  97  96  96  96  96  95  95  95  95  95  94  94
## [1225]  93  93  93  93  92  91  90  90  90  89  89  87
```

```r
sort(baby$smoke, decreasing = TRUE)
```

```
##    [1] 9 9 9 9 9 9 9 9 9 9 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##   [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##   [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [186] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [223] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [260] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [297] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [334] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [371] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [408] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [445] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [482] 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [519] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [556] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [593] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [630] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [667] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [704] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [741] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [778] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [815] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [852] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [889] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [926] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [963] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1000] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1037] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1074] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1111] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1148] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1185] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1222] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Now we are going to clean up the data by taking out the outliers.

```r
baby.clean = baby[baby$smoke != 9,]
baby.clean = baby.clean[baby.clean$weight != 999,]
baby.clean = baby.clean[baby.clean$age != 99,]
```

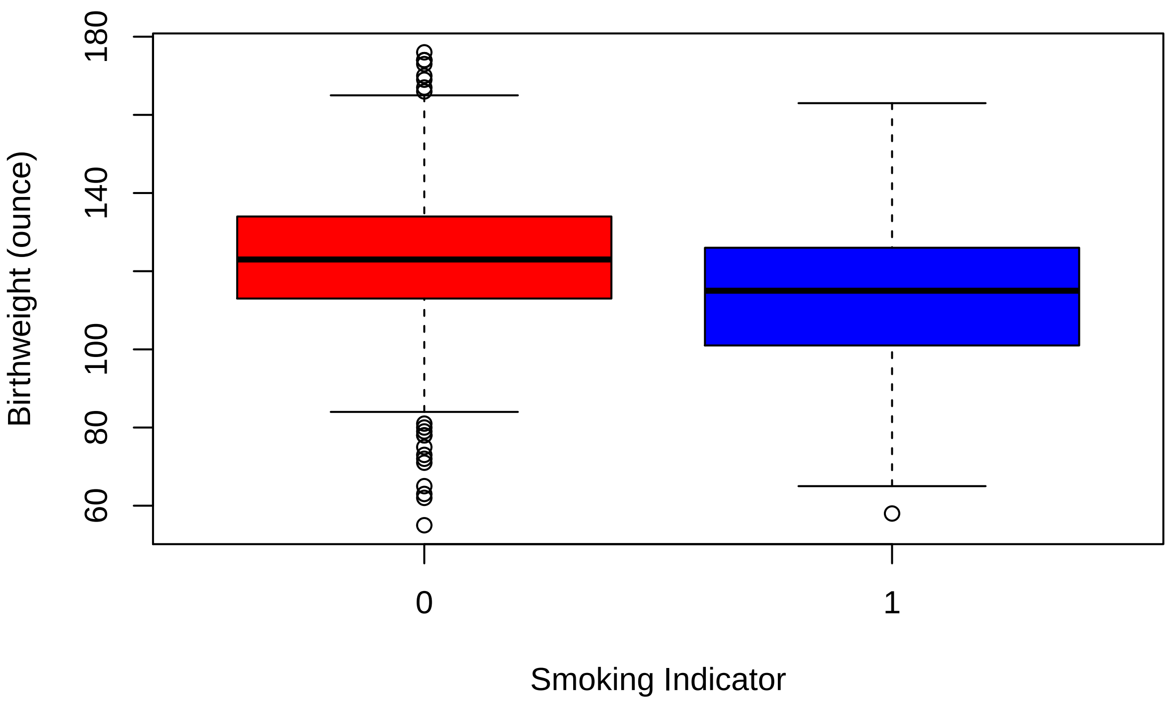
```
baby.clean = baby.clean[baby.clean$gestation != 999,]
baby.clean = baby.clean[baby.clean$height != 99,]
nrow(baby.clean)
```

```
## [1] 1174
```

As it can be seen the data has been cleaned up and the total number of rows have been reduced to 1174 from 1236.

Now we will explore the association between the smoking indicator and birth weight of the baby.

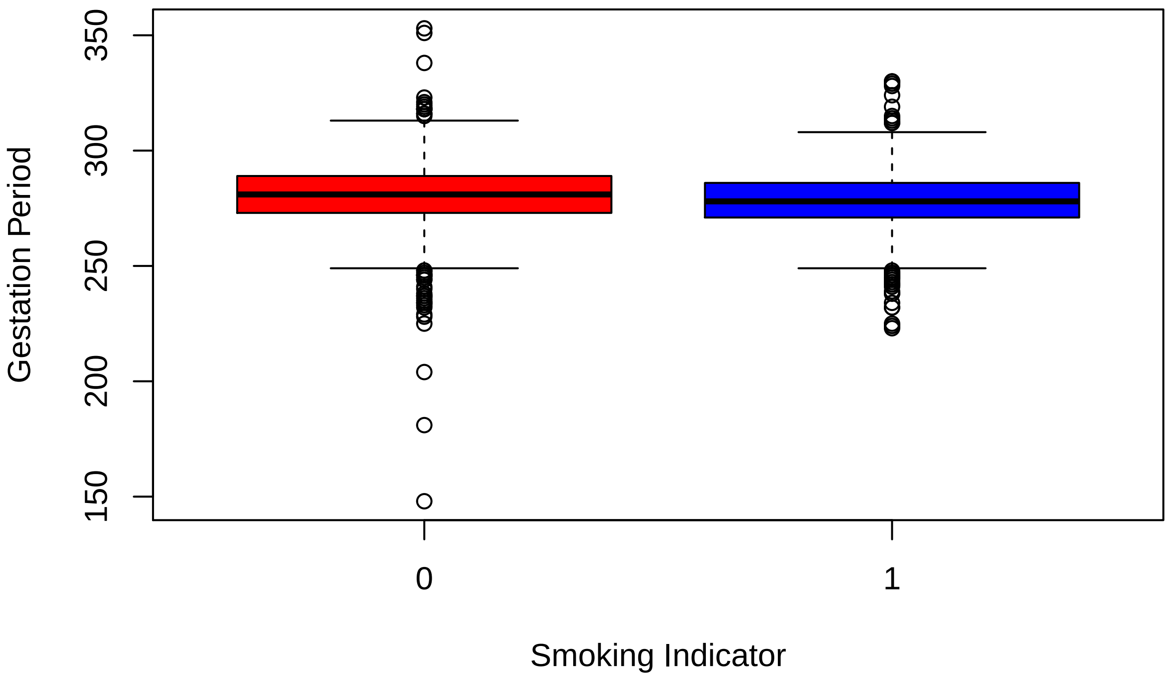
```
boxplot(baby.clean$bwt~baby.clean$smoke, xlab = "Smoking Indicator", ylab = "Birthweight (ounce)", col =
```



As it can be seen there is a difference between the bwt (birth weight of the baby) from mother's that do smoke and that do not smoke. Non-smoker babies weigh more than smoker babies on average.

Now we will see if there is an association between gestation period and smoking indicator.

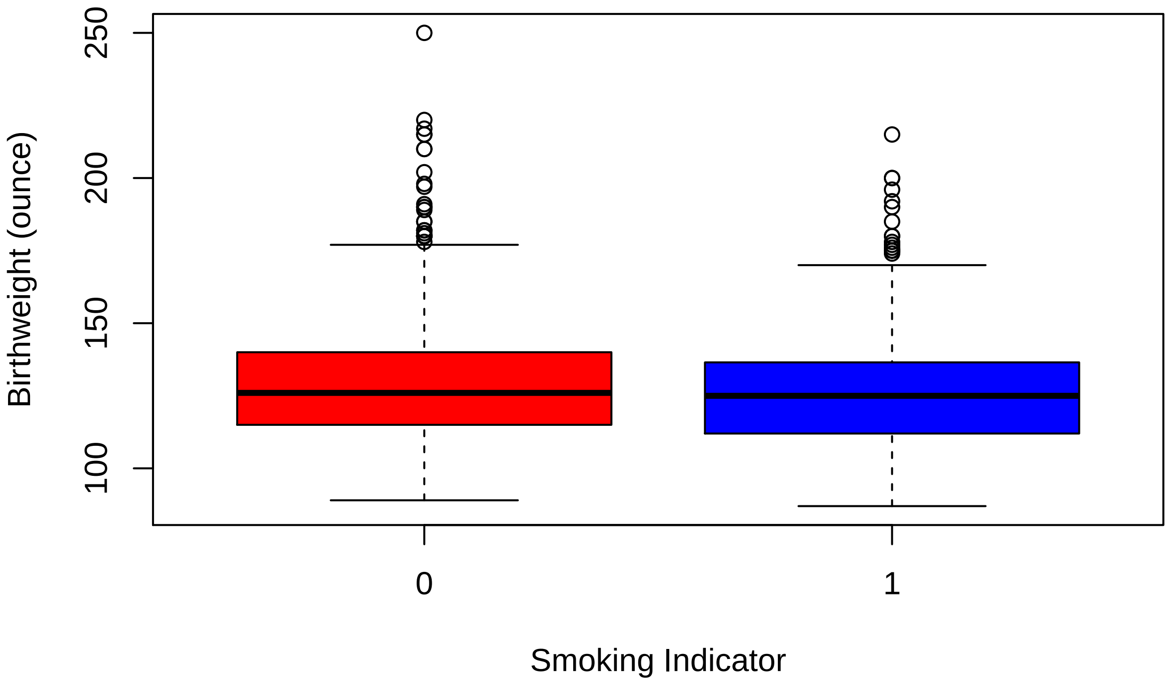
```
boxplot(baby.clean$gestation~baby.clean$smoke, xlab = "Smoking Indicator", ylab = "Gestation Period", co
```



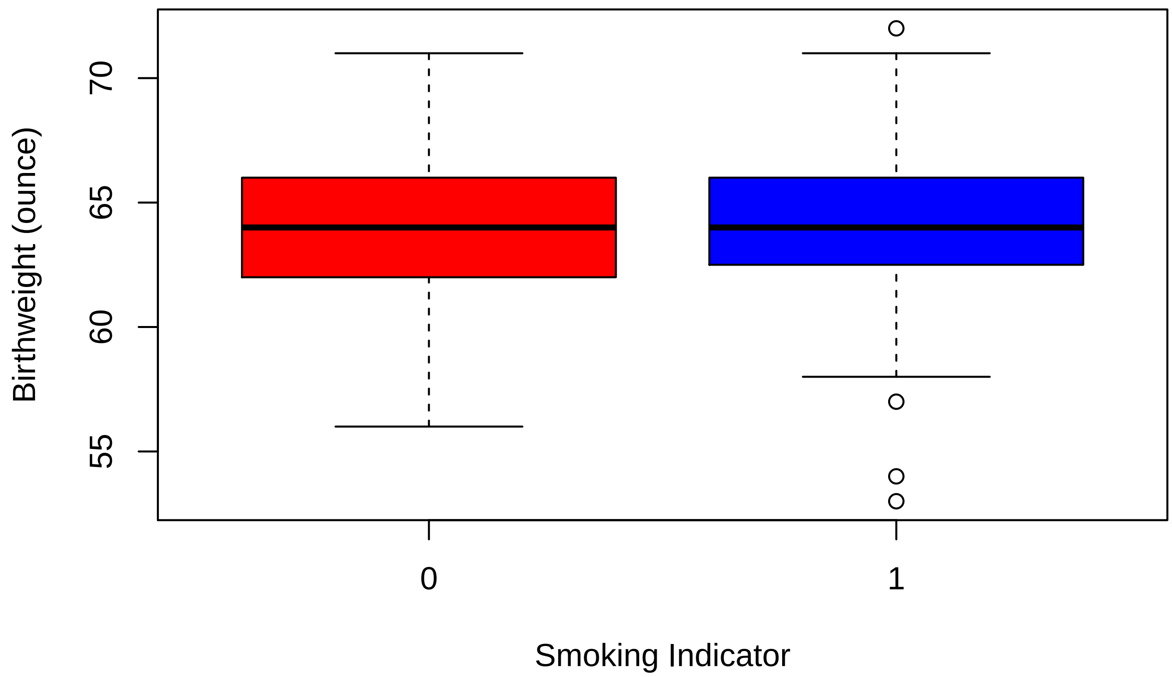
As it can be seen there seems to be no association. Therefore, the gestation period variable can be avoided.

Now we will look at whether there is an association between other variables and the smoking indicator.

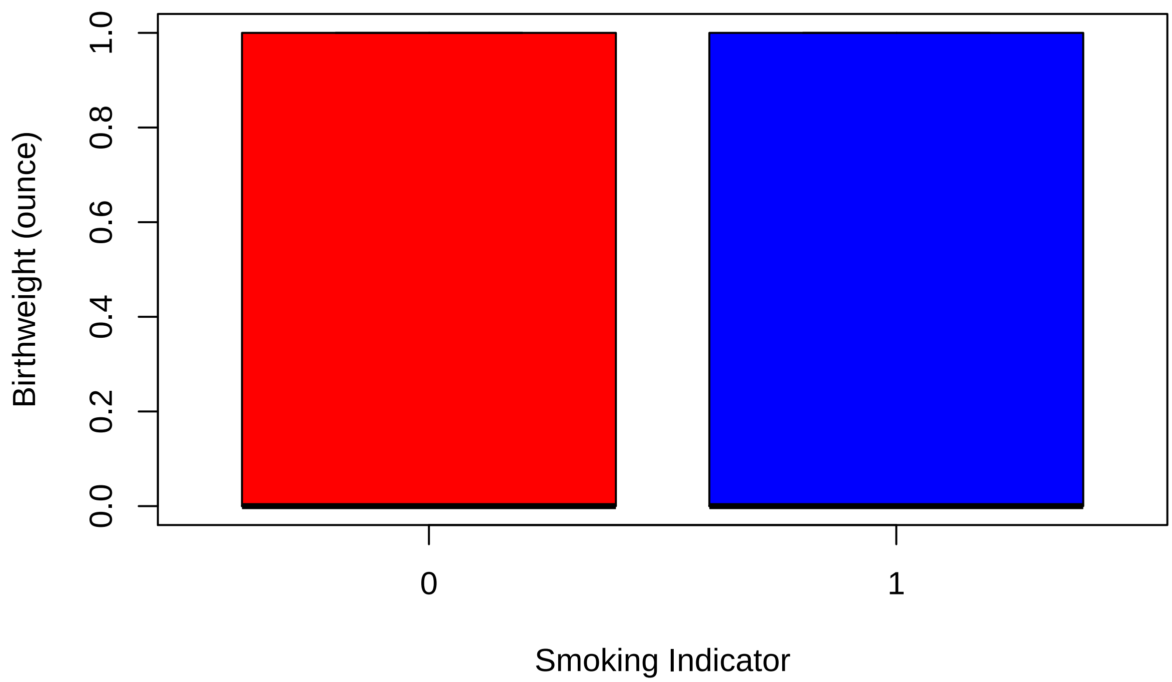
```
boxplot(baby.clean$weight~baby.clean$smoke, xlab = "Smoking Indicator", ylab = "Birthweight (ounce)", co
```



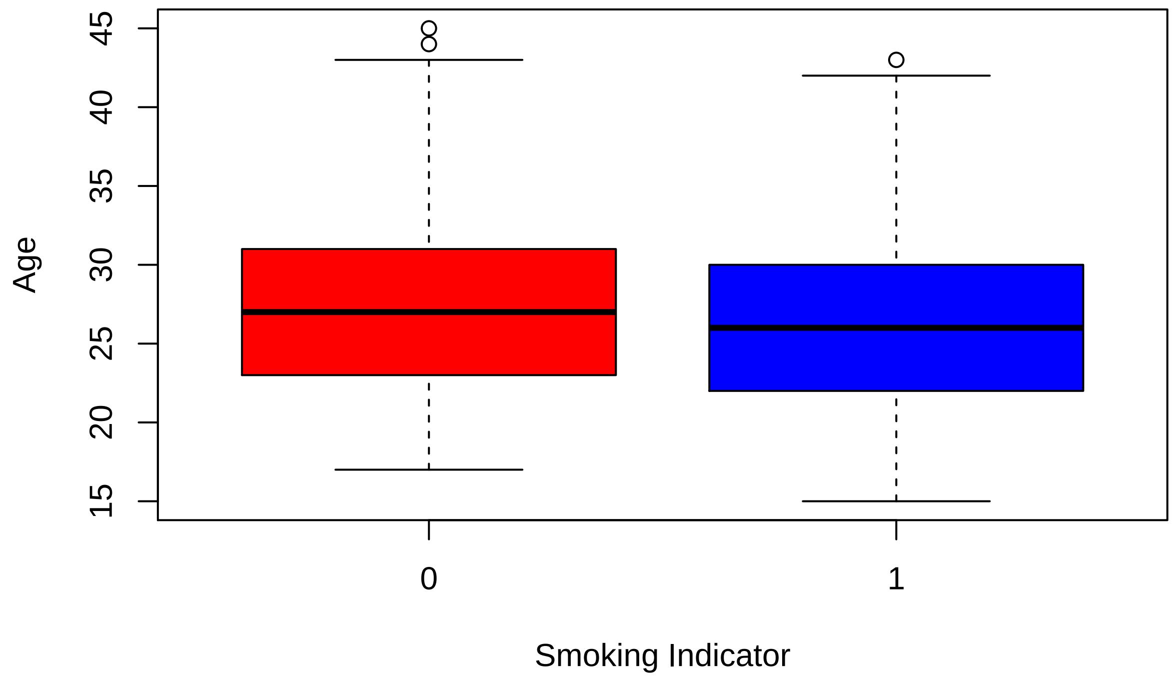

```
boxplot(baby.clean$height~baby.clean$smoke, xlab = "Smoking Indicator", ylab = "Birthweight (ounce)", co
```




```
boxplot(baby.clean$parity~baby.clean$smoke, xlab = "Smoking Indicator", ylab = "Birthweight (ounce)", c
```




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



There seems to be a little association between age and smoker indicator as well as weight and smoker indicator. However, there is no association between parity and smoker indicator as well as height and smoker indicator. Therefore, they can be avoided alongside gestation period.

2. There are 1236 observations. Split the data into a training set and a test set, of sizes that you select (and justify).

Using the clean data we will roughly split the data into an 80%, 20% split into training and test data which is standard.

```
n = nrow(baby.clean)
split = floor(0.8*n)
train_val = sample(1:n, split, replace = FALSE)
test_val = (1:n)[-train_val]
baby.train = baby.clean[train_val, c(1,4,6,7)]
baby.test = baby.clean[test_val, c(1,4,6,7)]
```

3. Perform logistic regression on the training data in order to predict smoking indicator using the variables that seemed most associated.

We will now fit the logistics model using bwt, age, and weight as predictors.

```
library(ISLR)
```

```r
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
## -1.7108  -0.9981  -0.8003   1.2197   1.9117
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.95087    0.47263   6.244 4.28e-10 ***
## bwt         -0.02824    0.00396  -7.131 9.94e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1264.2  on 938  degrees of freedom
## Residual deviance: 1208.5  on 937  degrees of freedom
## AIC: 1212.5
##
## Number of Fisher Scoring iterations: 4
```

```r
best.fit1 = glm(smoke~age, data = baby.train, family = binomial)
summary(best.fit1)
```

```
##
## Call:
## glm(formula = smoke ~ age, family = binomial, data = baby.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1259  -1.0245  -0.9385   1.3141   1.5486
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.34691    0.32008   1.084   0.2785
## age         -0.02761    0.01156  -2.388   0.0169 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1264.2  on 938  degrees of freedom
## Residual deviance: 1258.4  on 937  degrees of freedom
## AIC: 1262.4
##
## Number of Fisher Scoring iterations: 4
```

```r
best.fit2 = glm(smoke~weight, data = baby.train, family = binomial)
summary(best.fit2)
```

```
##
## Call:
## glm(formula = smoke ~ weight, family = binomial, data = baby.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1054  -1.0232  -0.9718   1.3241   1.5852
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.357662   0.422992   0.846    0.398
## weight      -0.005949   0.003272  -1.818    0.069 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1264.2  on 938  degrees of freedom
## Residual deviance: 1260.9  on 937  degrees of freedom
## AIC: 1264.9
##
## Number of Fisher Scoring iterations: 4
```

The p-values of age and weight is greater than 0.05. Therefore they can be dropped as predictors. This indicates that only bwt (babies weight at birth) is a good predictor for whether the mother is a smoker or non-smoker.

4. Generate the prediction probabilities for the test data, and discuss the results.

For each record in the test data, we can now compute the probability of smoker status as a function of Birthweight.

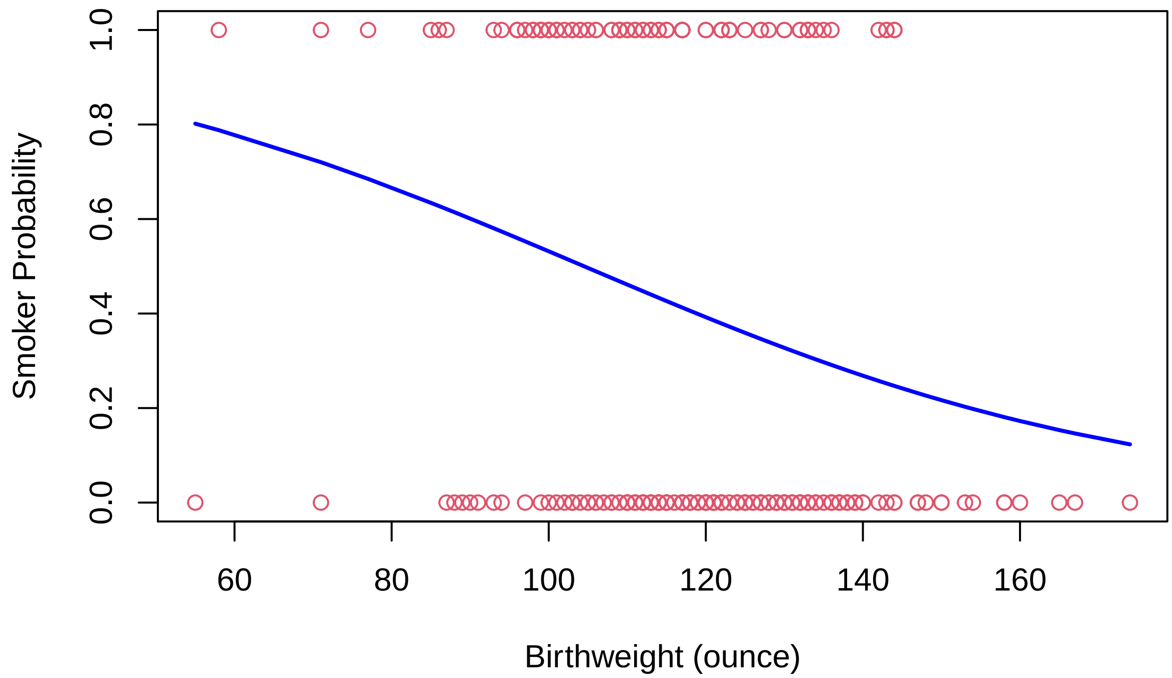
```r
pred_all = function(obs){
  x = c(1, obs)
  pred = as.numeric(as.numeric(x %*% best.fit$coefficients))
  pred = 1/(1+exp(-pred))
  return(pred)
}
```

Now we will generate the probabilities.

```r
prob_smoke = c()
for (x in 1:dim(baby.test)[1]) {
  prob_smoke = c(prob_smoke, pred_all(baby.test$bwt[x]))
}

A2 = cbind(prob_smoke,baby.test$bwt,baby.test$smoke)
A2 = A2[order(A2[,1], decreasing = FALSE),]
```

```r
plot(A2[,2], A2[,1], type = 'l', lwd = 2, col = "blue", xlab = "Birthweight (ounce)", ylab = "Smoker Pr
points(A2[,2], A2[,3], col = 2)
```



The graph above shows the probablity of the mother being a smoker depending on the weight of the baby such that if the baby weighs around 65oz then the probability of the baby being a smoker baby would be around 70%.

We will now assess the performance of the above probability. We will consider any probability over 0.5 as corresponding to smoker and any less than 0.5 as a non-smoker.

```r
pred_table = c(dim(baby.test[baby.test$smoke == 0 & prob_smoke < 0.5,])[1],
               dim(baby.test[baby.test$smoke == 1 & prob_smoke < 0.5,])[1])

pred_table = rbind(pred_table,
                   c(dim(baby.test[baby.test$smoke == 0 & prob_smoke >= 0.5,])[1],
                     dim(baby.test[baby.test$smoke == 1 & prob_smoke >= 0.5,])[1]))

rownames(pred_table) = c("Predict 0", "Predict 1")
colnames(pred_table) = c("Smoker 0", "Smoker 1")
pred_table
```

```
##             Smoker 0 Smoker 1
## Predict 0      134       54
## Predict 1       18       29
```

Now we will check the accuracy of this prediction table

```r
sum(diag(pred_table)) / nrow(baby.test)
```

```
## [1] 0.693617
```

Our prediction had about a 59.57% accuracy rate. Out of the 235 test data points, we correctly predicted 117 of the sample were nonsmokers and 23 of the samples were smokers. We incorrectly predicted 14 of the samples were smokers and 81 of the samples were non smokers.

Results & Conclusion

5. In your conclusion, discuss possible applications of such a predictive model. Comment on how it is possible to predict from variables that are not causing a phenomenon.

Since the model does not have a very high accuracy rate, it should not solely be used to form conclusions. Instead, it could be used as an indicator that further analysis is needed. It seems like a lower birth weight is correlated to mothers smoking. Birth weight can be a possible indicator for the baby's health, meaning that too low of a birth weight can suggest the baby was premature. Smoking is known to cause premature births. It is most likely that the birth weight does not have any effect on whether or not the mother smokes. However, we are still able to predict whether or not the mother smokes based on the birth weight because there is a relationship between the mother being a smoker/non smoker and the birth weight. For instance, it could be that smoking or not smoking causes birth weight to increase/decrease. While birth weight can be used to predict whether or not the mother smokes, birth weight does not have to cause the mother to smoke or to not smoke in order to be used as a predictor.