R Notebook

Code

First and foremost, we clear all memory and load our library and dataset.

Hide

```
rm(list=ls(all=TRUE))

library(MASS)
data(birthwt)

birthwt
```

	low <int>	age <int>	lwt <int>	race <int>	smoke <int>	ptl <int>	ht <int>	ui <int>	ftv <int>
85	0	19	182	2	0	0	0	1	0
86	0	33	155	3	0	0	0	0	3
87	0	20	105	1	1	0	0	0	1
88	0	21	108	1	1	0	0	1	2
89	0	18	107	1	1	0	0	1	0
91	0	21	124	3	0	0	0	0	0
92	0	22	118	1	0	0	0	0	1
93	0	17	103	3	0	0	0	0	1
94	0	29	123	1	1	0	0	0	1
95	0	26	113	1	1	0	0	0	0

1-10 of 189 rows | 1-10 of 10 columns Previous **1** 2 3 4 5 6 … 19 Next

The dataset we will be using today is the birthwt dataset, which gives the risk factors assoicated with low infant birth weight. For our research question, we would like to know which factors in particular, if any, lead to low birth weight in babies.

To do so, let's first take a quick look at the overall data.

Hide

```
summary(birthwt)
```

```
      low               age              lwt              race           smoke       
 Min.   :0.0000   Min.   :14.00   Min.   : 80.0   Min.   :1.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   1st Qu.:1.000   1st Qu.:0.0000  
 Median :0.0000   Median :23.00   Median :121.0   Median :1.000   Median :0.0000  
 Mean   :0.3122   Mean   :23.24   Mean   :129.8   Mean   :1.847   Mean   :0.3915  
 3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :45.00   Max.   :250.0   Max.   :3.000   Max.   :1.0000  
      ptl               ht                ui              ftv              bwt        
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   : 709  
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2414  
 Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000   Median :2977  
 Mean   :0.1958   Mean   :0.06349   Mean   :0.1481   Mean   :0.7937   Mean   :2945  
 3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:3487  
 Max.   :3.0000   Max.   :1.00000   Max.   :1.0000   Max.   :6.0000   Max.   :4990  
```

We see here that there are 9 total variables. Race, smoke, ht, and ui are all factors. We convert them from integers to factor variables. Furthemore, we can eliminate the “bwt” column, since all it does it give us the weight of the baby. We will only be looking at whether the baby is less than 2.5 kg, or, the “low” column.

Hide

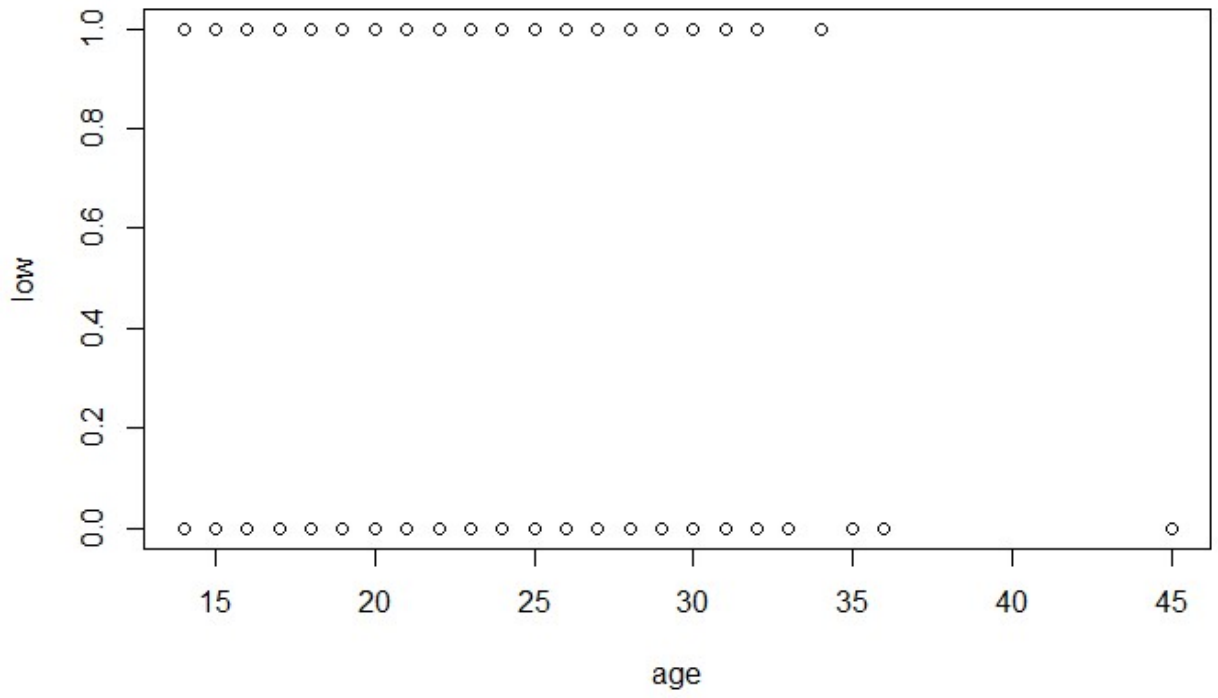
```
birthwt$race <- as.factor(birthwt$race)
birthwt$smoke <- as.factor(birthwt$smoke)
birthwt$ht <- as.factor(birthwt$ht)
birthwt$ui <- as.factor(birthwt$ui)
birthwt <- birthwt[,-10]
birthwt
```

	low <int>	age <int>	lwt <int>	race <fctr>	smoke <fctr>	ptl <int>	ht <fctr>	ui <fctr>	ftv <int>
85	0	19	182	2	0	0	0	1	0
86	0	33	155	3	0	0	0	0	3
87	0	20	105	1	1	0	0	0	1
88	0	21	108	1	1	0	0	1	2
89	0	18	107	1	1	0	0	1	0
91	0	21	124	3	0	0	0	0	0
92	0	22	118	1	0	0	0	0	1
93	0	17	103	3	0	0	0	0	1
94	0	29	123	1	1	0	0	0	1
95	0	26	113	1	1	0	0	0	0

To begin with, I plot each factor against low. However, because the response variable is binomial, we may not immediately see any particular trends.

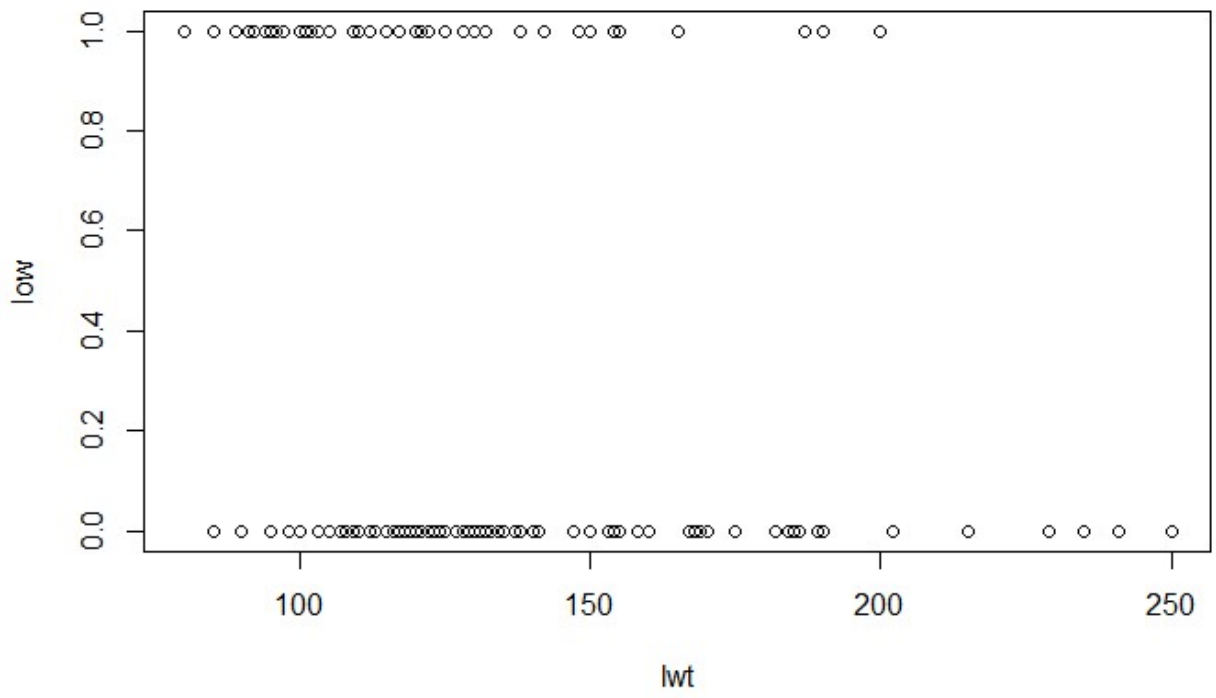
Hide

```
plot(low ~ age, data = birthwt)
```



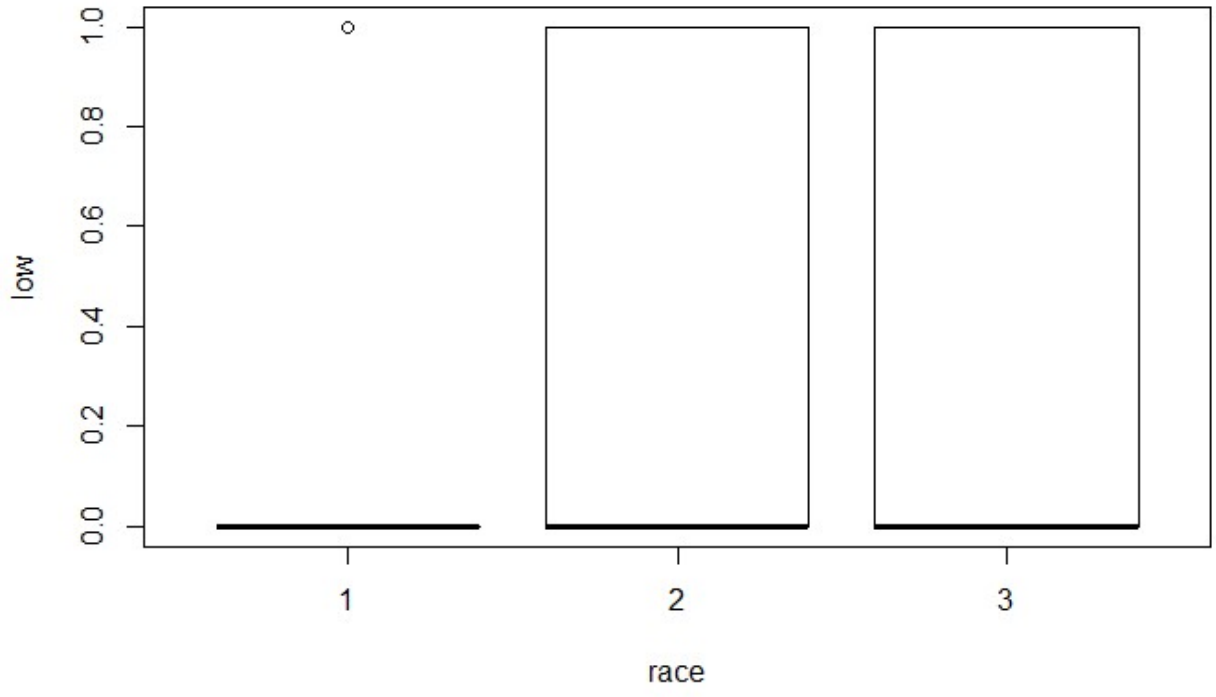
Hide

```
plot(low ~ lwt, data = birthwt)
```



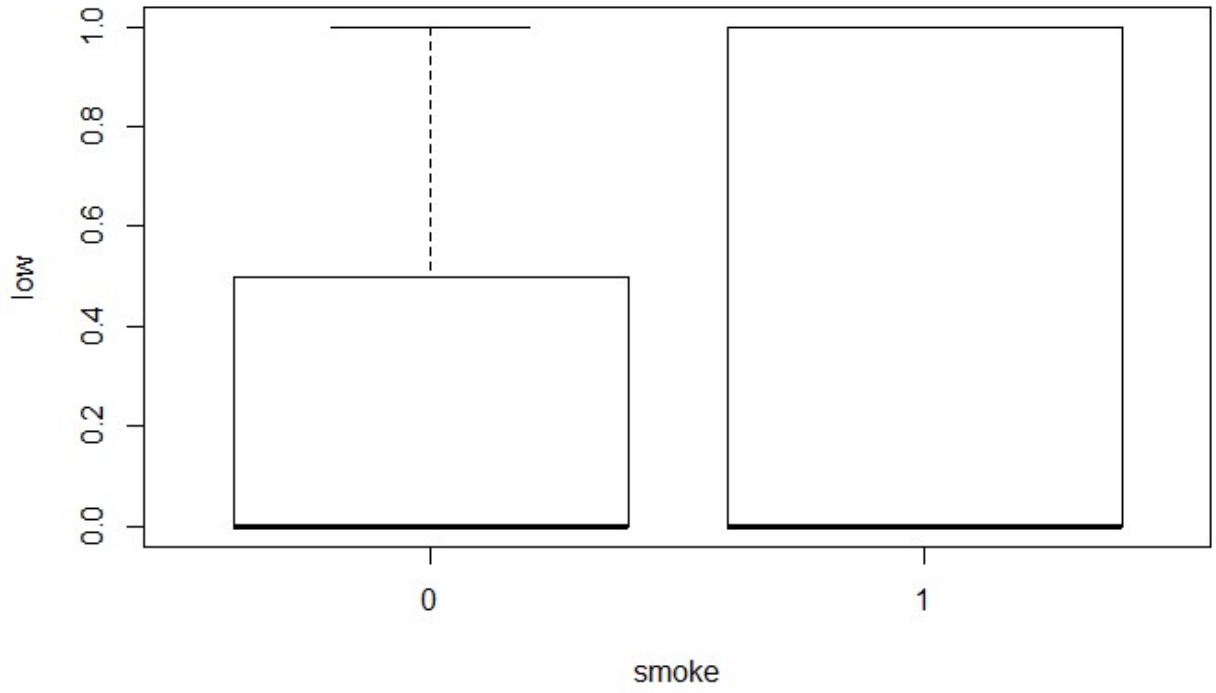
Hide

```
plot(low ~ race, data = birthwt)
```



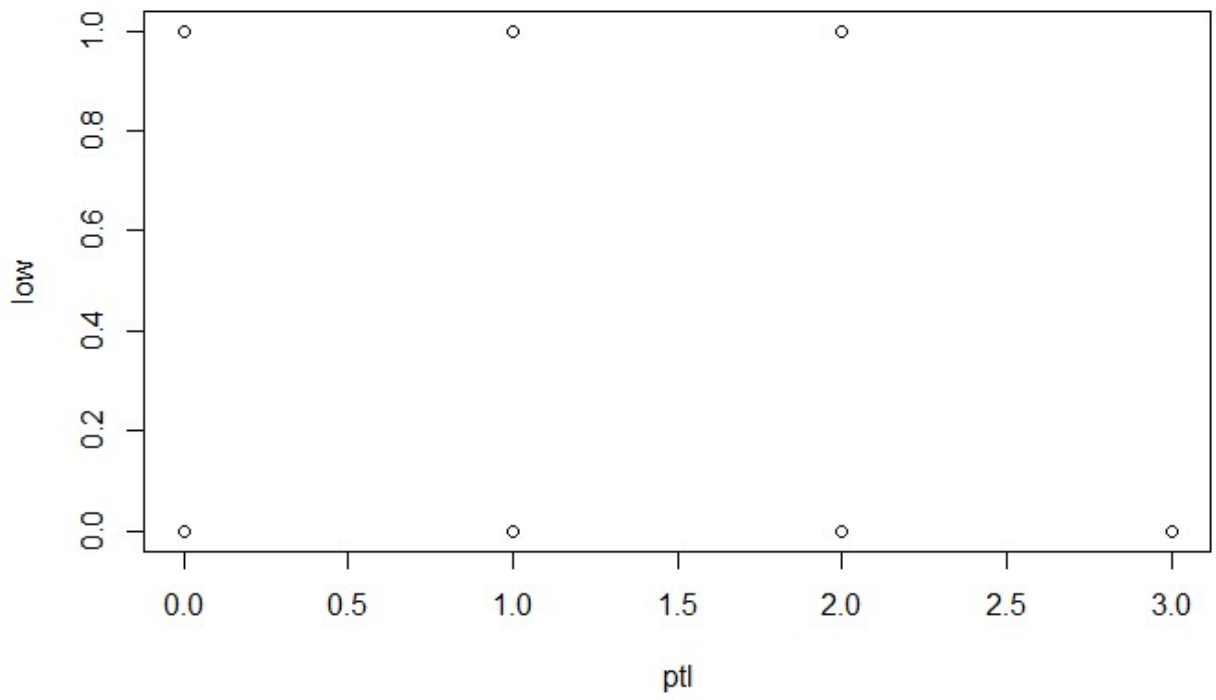
Hide

```
plot(low ~ smoke, data = birthwt)
```



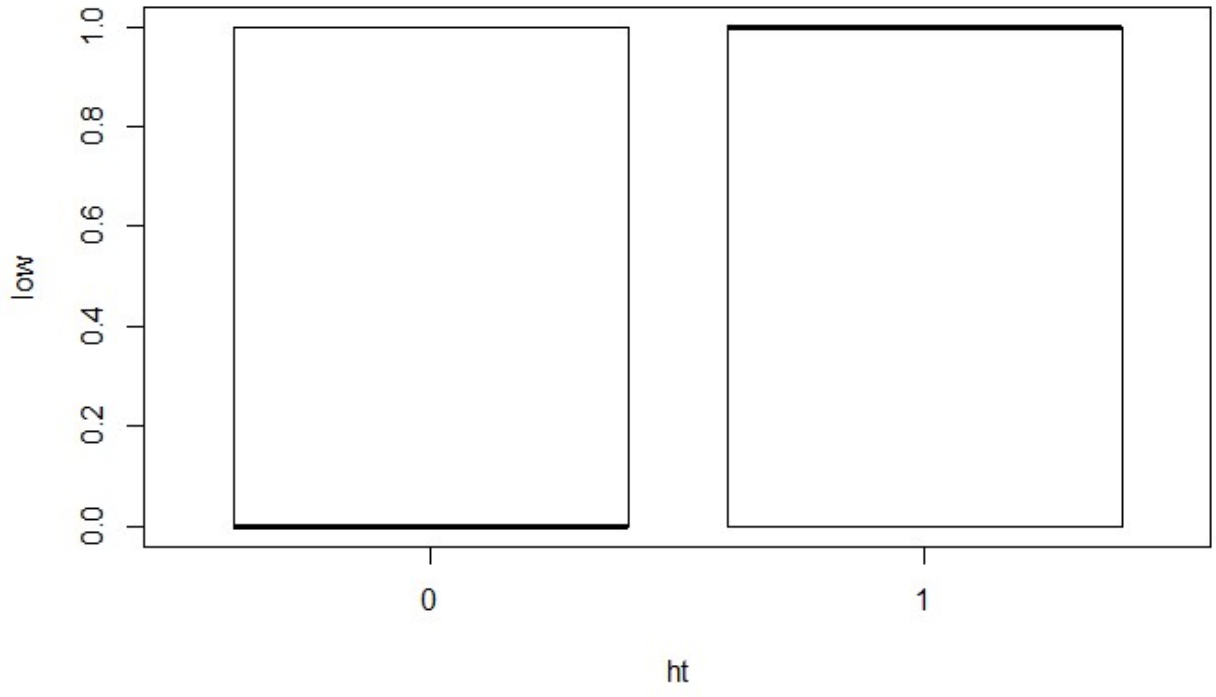
Hide

```
plot(low ~ ptl, data = birthwt)
```



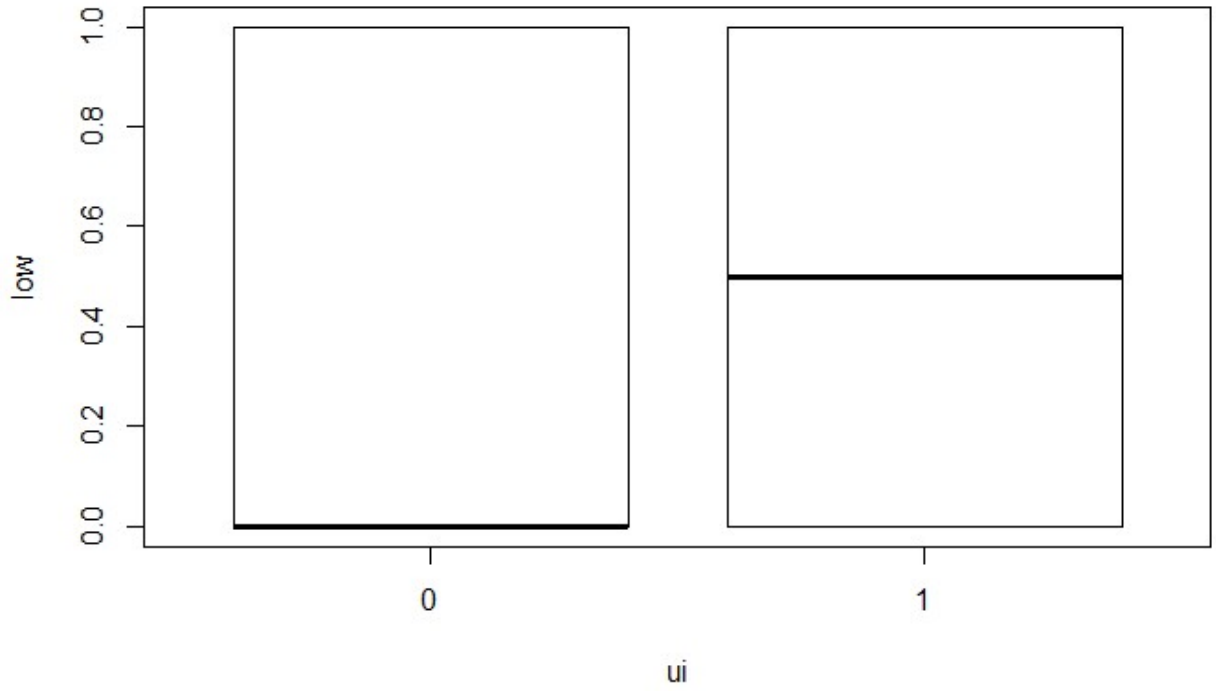
Hide

```
plot(low ~ ht, data = birthwt)
```



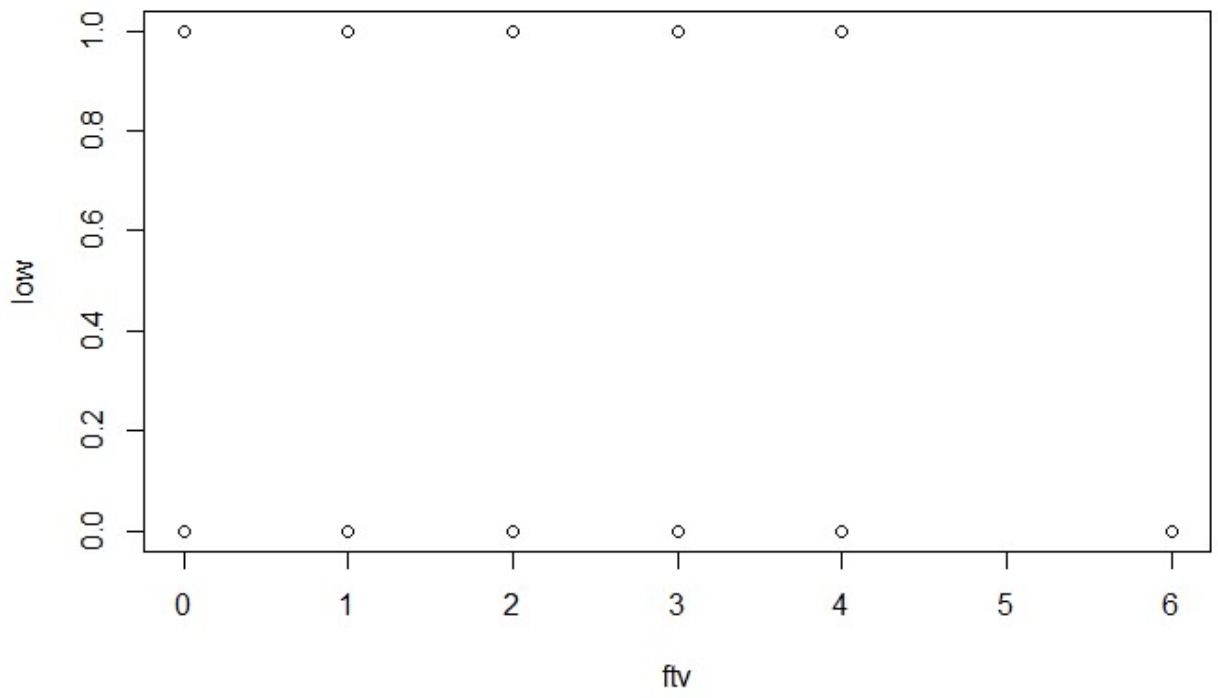
Hide

```
plot(low ~ ui, data = birthwt)
```



Hide

```
plot(low ~ ftv, data = birthwt)
```



Now, using the above plots as a guideline, we formulate our hypotheses. Our null hypothesis is that these 9 factors do not have an effect on low birthweight. This assumes the status quo, and is the hypothesis that we aim to reject.

My alternative hypothesis is that these 9 factors DO have an effect on low birthweight. This hypothesis covers everything other than the null hypothesis. Studie have shown that genetics as well as health conditions during pregnancy lead to health conditions of the offspring.

For our dataset, we will be using a Generalized Linear Model. To do so, we must first create the global model with all variables. Since we want to know which specific factors affect low birthweight, we do not take into consideration the pairwise interactions between two factors.

Hide

```
model.1 <- glm(low ~ age + lwt + race + smoke + ptl + ht + ui + ftv, data = birthwt, f
amily = "binomial")
summary(model.1)
```

```
Call:
glm(formula = low ~ age + lwt + race + smoke + ptl + ht + ui +
    ftv, family = "binomial", data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8946  -0.8212  -0.5316   0.9818   2.2125

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.480623   1.196888   0.402  0.68801
age         -0.029549   0.037031  -0.798  0.42489
lwt         -0.015424   0.006919  -2.229  0.02580 *
race2        1.272260   0.527357   2.413  0.01584 *
race3        0.880496   0.440778   1.998  0.04576 *
smoke1       0.938846   0.402147   2.335  0.01957 *
ptl          0.543337   0.345403   1.573  0.11571
ht1          1.863303   0.697533   2.671  0.00756 **
ui1          0.767648   0.459318   1.671  0.09467 .
ftv          0.065302   0.172394   0.379  0.70484
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 201.28  on 179  degrees of freedom
AIC: 221.28

Number of Fisher Scoring iterations: 4
```

We see that only some of our intercepts seem to show significant values. This is understandable since we have many parameters, and inclusion of all of our variables could lead to a complicated model.

Next, we reduce this global model by one factor each time, and compare the AIC values to one another. We opt for the model with the lowest AIC value without simplyfing the model too much.

Hide

```
options(na.action = "na.fail")
red.model <- dredge(model.1, rank = "AICc")
```

```
Fixed term is "(Intercept)"
```

Hide

```
red.model
```

```
Global model call: glm(formula = low ~ age + lwt + race + smoke + ptl + ht + ui +
    ftv, family = "binomial", data = birthwt)
---
Model selection table
     (Intrc)      age       ftv ht       lwt    ptl race smoke ui df   logLik  AICc de
lta
253 -0.08655                    + -0.015910 0.5032    +     +  +  8 -100.993 218.8
0.00
237  0.05628                    + -0.016730           +     +  +  7 -102.108 218.8
0.05
125  0.11790                    + -0.016580 0.6025    +     +     7 -102.449 219.5
0.73
254  0.46440 -0.02707           + -0.015180 0.5418    +     +  +  9 -100.713 220.4
1.65
109  0.35200                    + -0.017910           +     +     6 -104.124 220.7
1.92
238  0.43720 -0.01826           + -0.016290           +     +  +  8 -101.974 220.7
1.96
126  0.74110 -0.03194           + -0.015600 0.6462    +     +     8 -102.053 220.9
2.12
255 -0.10740           0.040160  + -0.016100 0.5023    +     +  +  9 -100.965 220.9
2.15
239  0.03615           0.040830  + -0.016940           +     +  +  8 -102.079 221.0
2.17
127  0.10890           0.020240  + -0.016680 0.6025    +     +     8 -102.442 221.7
2.90
189  0.68500                    + -0.017510 0.6408    +        +  7 -103.864 222.3
3.56
110  0.77730 -0.02117           + -0.017300           +     +     7 -103.939 222.5
3.71
256  0.48060 -0.02955  0.065300  + -0.015420 0.5433    +     +  + 10 -100.642 222.5
3.74
245 -2.14600                    +           0.5758    +     +  +  7 -104.056 222.7
3.95
221  0.56280                    + -0.015490 0.5339          +  +  6 -105.155 222.8
3.99
240  0.45480 -0.02051  0.059120  + -0.016520           +     +  +  9 -101.916 222.8
4.05
111  0.34490           0.017270  + -0.018000           +     +     7 -104.118 222.9
4.07
128  0.75500 -0.03374  0.048390  + -0.015780 0.6485    +     +     9 -102.014 223.0
4.25
205  0.72190                    + -0.016310                 +  +  5 -106.413 223.2
4.37
61   0.88620                    + -0.018120 0.7464    +           6 -105.425 223.3
4.53
157  0.83250                    + -0.015910 0.6274             +  5 -106.507 223.3
4.56
```

```
222  1.38200 -0.04223           + -0.014320 0.5932         +  +  7 -104.386 223.4
4.60
93   0.82390                    + -0.016690 0.6282         +     5 -106.574 223.5
4.69
246 -1.18400 -0.04033           +           0.6289    +    +  +  8 -103.404 223.6
4.82
229 -2.09200                    +                     +    +  +  6 -105.583 223.6
4.84
190  1.31100 -0.03163           + -0.016560 0.6793    +       +  8 -103.458 223.7
4.93
94   1.67400 -0.04561           + -0.015200 0.6896         +     6 -105.664 223.8
5.00
158  1.65800 -0.04372           + -0.014560 0.6842            +  6 -105.665 223.8
5.01
173  0.95220                    + -0.018640           +       +  6 -105.755 224.0
5.19
121 -0.35020                      -0.011940 0.6055    +    +     6 -105.775 224.0
5.23
249 -0.53980                      -0.011130 0.5202    +    +  +  7 -104.732 224.1
5.30
117 -2.02500                    +           0.6965    +    +     6 -105.812 224.1
5.30
29   1.09300                    + -0.017070 0.7256               4 -107.982 224.2
5.40
206  1.40000 -0.03407           + -0.015450                 +  +  6 -105.889 224.2
5.45
30   1.94500 -0.04663           + -0.015440 0.7828               5 -107.005 224.3
5.55
233 -0.38660                      -0.011980           +    +  +  6 -105.957 224.4
5.59
118 -0.92380 -0.04690           +           0.7501    +    +     7 -104.904 224.4
5.64
62   1.56800 -0.03531           + -0.016960 0.7864    +          7 -104.907 224.4
5.65
191  0.69170          -0.010570  + -0.017480 0.6405    +       +  8 -103.862 224.5
5.74
112  0.78910 -0.02250  0.036560  + -0.017450           +    +     8 -103.916 224.6
5.85
141  1.06800                    + -0.016920                  +  4 -108.306 224.8
6.04
247 -2.11700          -0.028730  +           0.5751    +    +  +  8 -104.039 224.9
6.09
223  0.57470          -0.023980  + -0.015420 0.5340         +  +  7 -105.145 224.9
6.12
230 -1.34400 -0.03102           +                     +    +  +  7 -105.176 225.0
6.18
77   1.08400                    + -0.018050                  +     4 -108.429 225.1
6.29
241 -2.01900                                 0.5738    +    +  +  6 -106.400 225.3
```

```
6.48
207  0.73390          -0.026190  + -0.016220                +  +  6 -106.401 225.3
6.48
122  0.27480 -0.03274              -0.010800 0.6491    +     +     7 -105.339 225.3
6.51
105 -0.10920                       -0.013260           +     +     5 -107.507 225.3
6.56
159  0.85420          -0.043420  + -0.015810 0.6267            +  6 -106.474 225.4
6.62
63   0.90260          -0.027090  + -0.018060 0.7451    +           7 -105.412 225.4
6.66
224  1.39100 -0.04325  0.023430  + -0.014370 0.5943          +  +  8 -104.376 225.6
6.77
95   0.84090          -0.038690  + -0.016560 0.6278          +     6 -106.547 225.6
6.77
250  0.02383 -0.02851              -0.010230 0.5605    +     +  +  8 -104.406 225.6
6.83
174  1.41300 -0.02269            + -0.018000           +        +  7 -105.534 225.7
6.90
231 -2.05700          -0.035330  +                     +     +  +  7 -105.558 225.7
6.95
113 -1.92900                                 0.6788    +     +     5 -107.719 225.8
6.98
142  1.74900 -0.03505            + -0.015930                    +  5 -107.740 225.8
7.02
248 -1.18300 -0.04075  0.009212  +           0.6296    +     +  +  9 -103.402 225.8
7.02
192  1.31300 -0.03223  0.016990  + -0.016580 0.6803    +        +  9 -103.453 225.9
7.13
96   1.67800 -0.04604  0.010350  + -0.015230 0.6902          +     7 -105.662 225.9
7.16
160  1.65900 -0.04390  0.004244  + -0.014570 0.6845            +  7 -105.665 225.9
7.16
78   1.76700 -0.03569            + -0.016960                 +     5 -107.842 226.0
7.23
175  0.96240          -0.016590  + -0.018590           +        +  7 -105.750 226.1
7.33
119 -1.97800          -0.050470  +           0.6938    +     +     7 -105.759 226.1
7.35
242 -1.08900 -0.03885                       0.6244    +     +  +  7 -105.763 226.1
7.36
123 -0.34500          -0.022590    -0.011800 0.6055    +     +     7 -105.765 226.1
7.36
31   1.11900          -0.057720  + -0.016920 0.7237              5 -107.923 226.2
7.39
114 -0.87750 -0.04455                       0.7289    +     +     6 -106.865 226.2
7.41
225 -1.96500                                          +     +  +  5 -107.934 226.2
7.41
```

```
234  0.01299 -0.01973              -0.011400           +     +  +  7 -105.791 226.2
7.42
    weight
253  0.101
237  0.098
125  0.070
254  0.044
109  0.039
238  0.038
126  0.035
255  0.034
239  0.034
127  0.024
189  0.017
110  0.016
256  0.016
245  0.014
221  0.014
240  0.013
111  0.013
128  0.012
205  0.011
61   0.010
157  0.010
222  0.010
93   0.010
246  0.009
229  0.009
190  0.009
94   0.008
158  0.008
173  0.008
121  0.007
249  0.007
117  0.007
29   0.007
206  0.007
30   0.006
233  0.006
118  0.006
62   0.006
191  0.006
112  0.005
141  0.005
247  0.005
223  0.005
230  0.005
77   0.004
241  0.004
```

```
207  0.004
122  0.004
105  0.004
159  0.004
63   0.004
224  0.003
95   0.003
250  0.003
174  0.003
231  0.003
113  0.003
142  0.003
248  0.003
192  0.003
96   0.003
160  0.003
78   0.003
175  0.003
119  0.003
242  0.003
123  0.003
31   0.003
114  0.002
225  0.002
234  0.002
 [ reached getOption("max.print") -- omitted 185 rows ]
Models ranked by AICc(x)
```

We see that the models with “ht, lwt, ptl, race, smoke, ui” and “ht, lwt, race, smoke, ui” have the same AICc of 218.8. We take the more simple of the two since they are considered equally parsimonious.

Hide

```
red.model.1 <- glm(low ~ lwt + race + smoke + ht + ui, data = birthwt, family = "binom
ial")
summary(red.model.1)
```

```
Call:
glm(formula = low ~ lwt + race + smoke + ht + ui, family = "binomial",
    data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7396  -0.8322  -0.5359   0.9873   2.1692

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.056276   0.937853   0.060  0.95215
lwt         -0.016732   0.006803  -2.459  0.01392 *
race2        1.324562   0.521464   2.540  0.01108 *
race3        0.926197   0.430386   2.152  0.03140 *
smoke1       1.035831   0.392558   2.639  0.00832 **
ht1          1.871416   0.690902   2.709  0.00676 **
ui1          0.904974   0.447553   2.022  0.04317 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 204.22  on 182  degrees of freedom
AIC: 218.22

Number of Fisher Scoring iterations: 4
```

We see now that the p-value of all of our intercepts fall within the 0.05 level, making them significant values.

Therfore, we can conclude that history of hypertension, mother’s weight, race, smoking status, and presence of uterine irritability all affect low birthweight in babies.