Introduction to Linear Regression

Rebecca C. Steorts, Duke University

STA 325, Chapter 3 ISL

Today

- Using data frames for statistical purposes
- Manipulation of data into more convenient forms
- Introduction to linear models and the model space

So You've Got A Data Frame

What can we do with it?

- Plot it: examine multiple variables and distributions
- Test it: compare groups of individuals to each other
- Check it: does it conform to what we'd like for our needs?

Test Case: Birth weight data

Included in R already:

```
library(MASS)
data(birthwt)
summary(birthwt)
```

```
##       low              age              lwt              r
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   Min.
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   1st Qu
##  Median :0.0000   Median :23.00   Median :121.0   Median
##  Mean   :0.3122   Mean   :23.24   Mean   :129.8   Mean
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0   Max.
##      smoke            ptl               ht
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st
##  Median :0.0000   Median :0.0000   Median :0.00000   Med
##  Mean   :0.3915   Mean   :0.1958   Mean   :0.06349   Mea
```

From R help

Go to R help for more info, because someone documented this (thanks, someone!)

```
help(birthwt)
```

Make it readable!

```
colnames(birthwt)
```

```
##  [1] "low"   "age"   "lwt"   "race"  "smoke" "ptl"   "ht"
##  [9] "ftv"   "bwt"
```

```
colnames(birthwt) <- c("birthwt.below.2500", "mother.age",
                       "mother.weight", "race",
                       "mother.smokes", "previous.prem.labor",
                       "hypertension", "uterine.irr",
                       "physician.visits", "birthwt.grams")
```

Making the factors more descriptive

```
birthwt$race <-
  factor(c("white", "black", "other")[birthwt$race])
birthwt$mother.smokes <-
  factor(c("No", "Yes")[birthwt$mother.smokes + 1])
birthwt$uterine.irr <-
  factor(c("No", "Yes")[birthwt$uterine.irr + 1])
birthwt$hypertension <-
  factor(c("No", "Yes")[birthwt$hypertension + 1])
```

Make it readable, again!

```
summary(birthwt)
```

```
##  birthwt.below.2500   mother.age    mother.weight      race
##  Min.   :0.0000     Min.   :14.00   Min.   : 80.0   black:26
##  1st Qu.:0.0000     1st Qu.:19.00   1st Qu.:110.0   other:67
##  Median :0.0000     Median :23.00   Median :121.0   white:96
##  Mean   :0.3122     Mean   :23.24   Mean   :129.8
##  3rd Qu.:1.0000     3rd Qu.:26.00   3rd Qu.:140.0
##  Max.   :1.0000     Max.   :45.00   Max.   :250.0
##  mother.smokes previous.prem.labor hypertension uterine.irr
##  No :115       Min.   :0.0000      No :177      No :161
##  Yes: 74       1st Qu.:0.0000      Yes: 12      Yes: 28
##                Median :0.0000
##                Mean   :0.1958
##                3rd Qu.:0.0000
##                Max.   :3.0000
##  physician.visits birthwt.grams
##  Min.   :0.0000   Min.   : 709
##  1st Qu.:0.0000   1st Qu.:2414
##  Median :0.0000   Median :2977
##  Mean   :0.7937   Mean   :2945
##  3rd Qu.:1.0000   3rd Qu.:3487
##  Max.   :6.0000   Max.   :4990
```

Explore it!

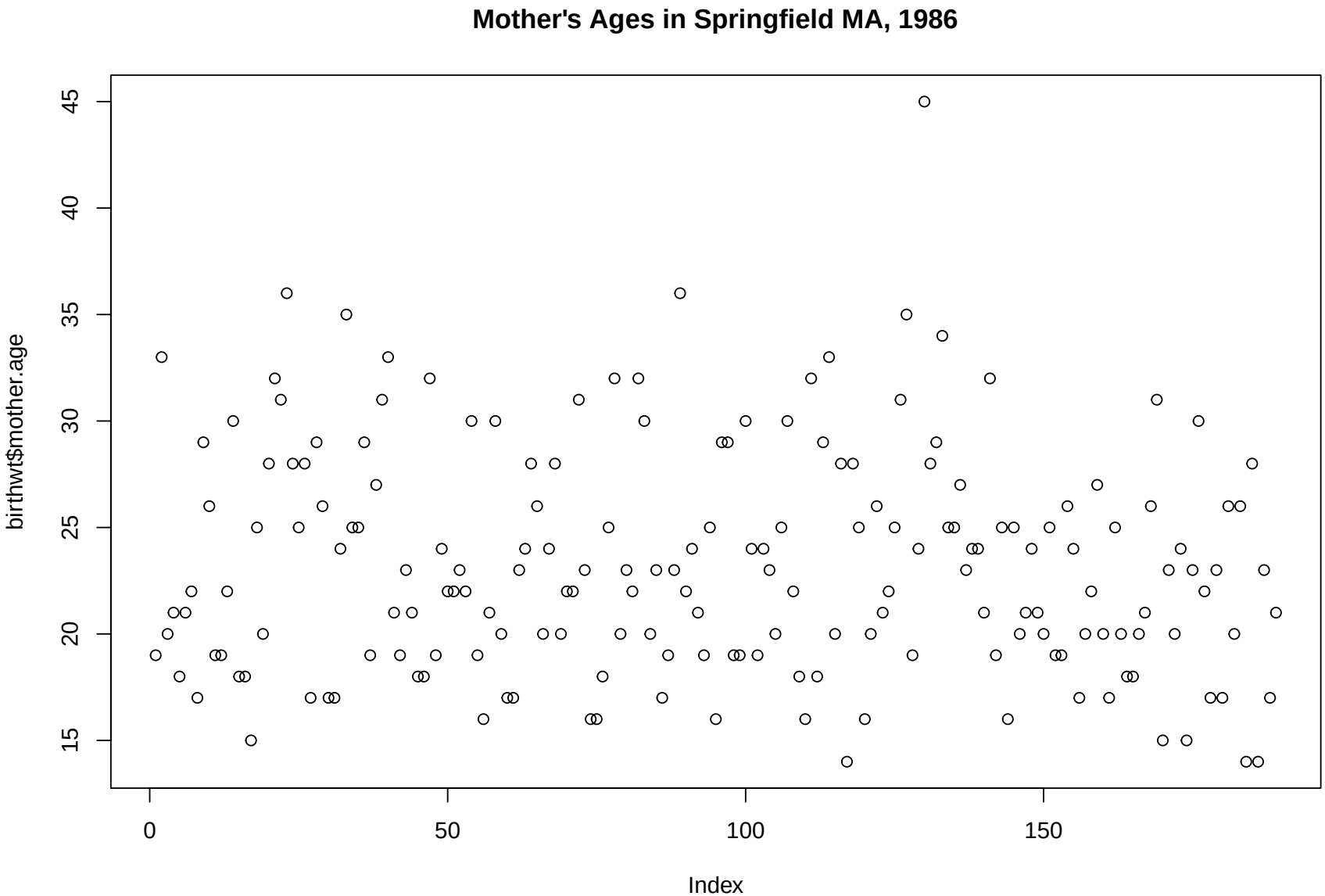
```
plot (birthwt$race, echo=FALSE)
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): "echo" is
## graphical parameter

## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = name
## axis.lty, : "echo" is not a graphical parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab =
## "echo" is not a graphical parameter

## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...
## not a graphical parameter
```

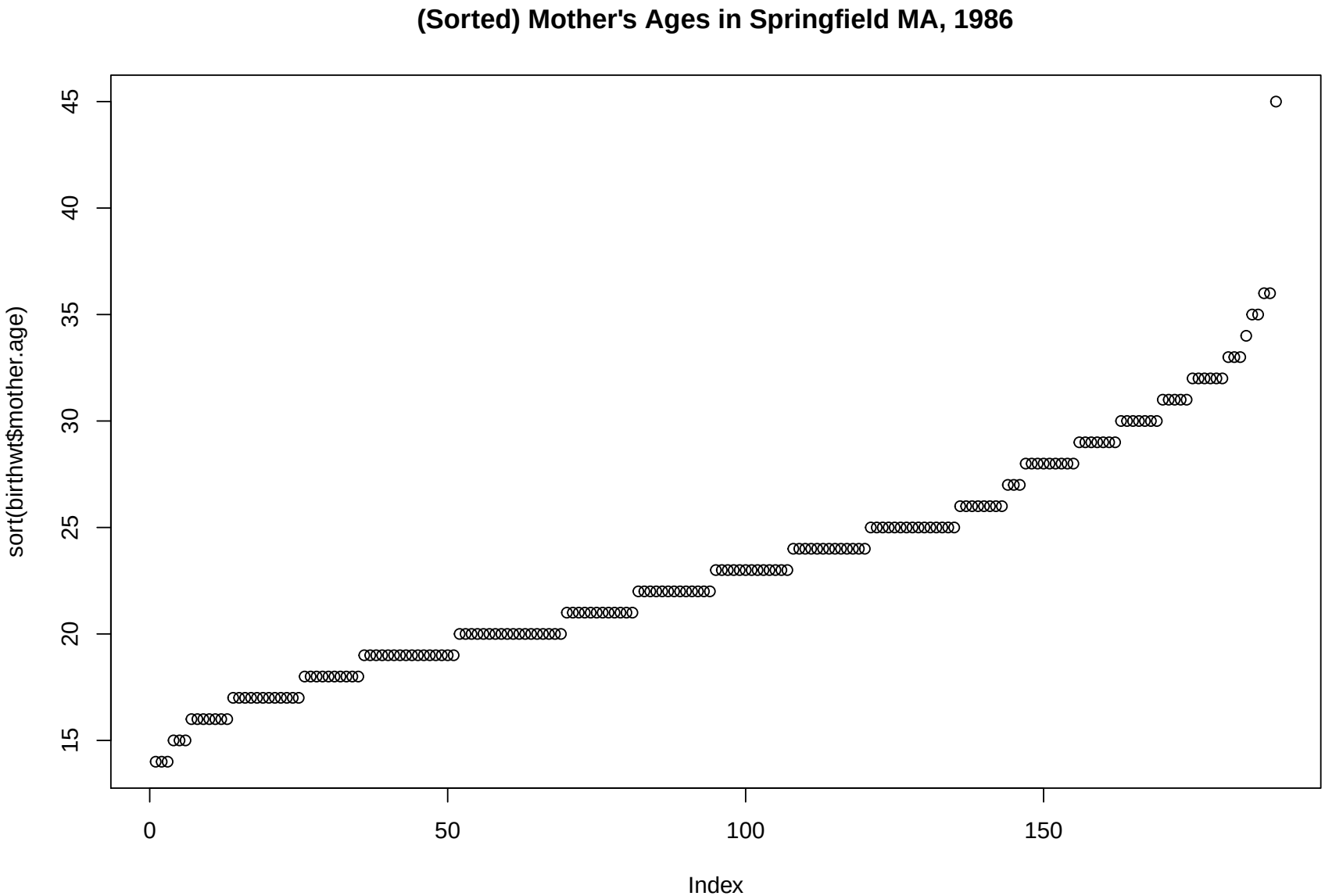
```
title (main = "Count of Mother's Race in
       Springfield MA, 1986")
```

Count of Mother's Race in Springfield MA, 1986

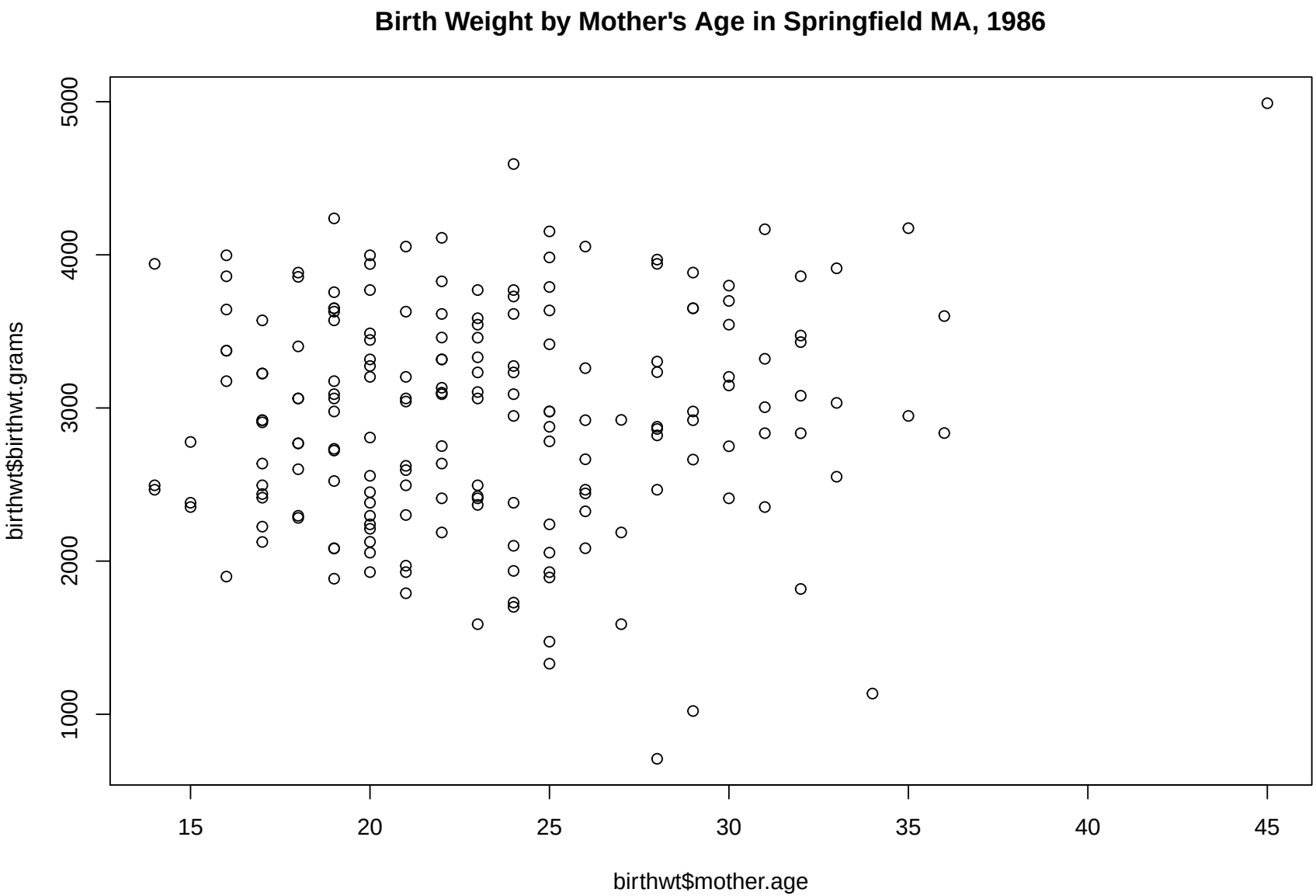
Explore it!


Mother's Ages in Springfield MA, 1986
birthwt$mother.age
15
20
25
30
35
40
45
0
50
100
150
Index

Explore it!


(Sorted) Mother's Ages in Springfield MA, 1986
sort(birthwt$mother.age)
15
20
25
30
35
40
45
0
50
100
150
Index

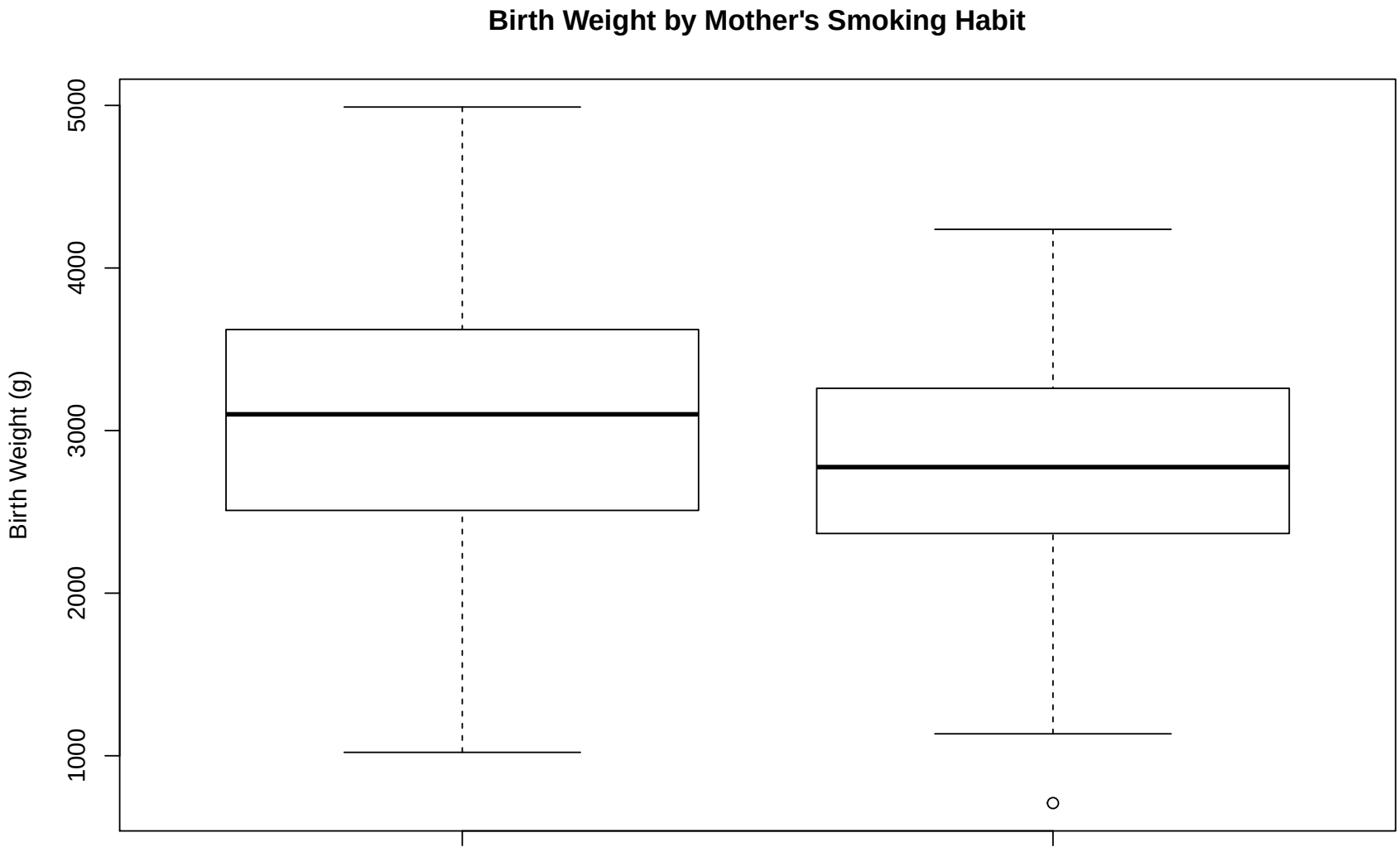
Explore it!


Birth Weight by Mother's Age in Springfield MA, 1986
birthwt$birthwt.grams
1000
2000
3000
4000
5000
15
20
25
30
35
40
45
birthwt$mother.age

Basic statistical testing

Let's fit some models to the data pertaining to our outcome(s) of interest.

```
plot (birthwt$mother.smokes, birthwt$birthwt.grams, main="B
```



Basic statistical testing

Tough to tell! Simple two-sample t-test:

```
t.test (birthwt$birthwt.grams[birthwt$mother.smokes == "Yes
        birthwt$birthwt.grams[birthwt$mother.smokes == "No"
```

```
##
##  Welch Two Sample t-test
##
## data:  birthwt$birthwt.grams[birthwt$mother.smokes == "Y
## t = -2.7299, df = 170.1, p-value = 0.007003
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -488.97860  -78.57486
## sample estimates:
## mean of x mean of y
##  2771.919  3055.696
```

Basic statistical testing

Does this difference match the linear model?

```
linear.model.1 <- lm (birthwt.grams ~ mother.smokes, data=b
linear.model.1
```

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.smokes, data = birth
##
## Coefficients:
##       (Intercept)  mother.smokesYes
##            3055.7            -283.8
```

Basic statistical testing

Does this difference match the linear model?

```
summary(linear.model.1)
```

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.smokes, data = birth
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2062.9  -475.9    34.3   545.1  1934.3
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3055.70      66.93  45.653  < 2e-16 **
## mother.smokesYes  -283.78     106.97  -2.653  0.00867 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Basic statistical testing

Does this difference match the linear model?

```
linear.model.2 <- lm (birthwt.grams ~ mother.age, data=birt
linear.model.2
```

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age, data = birthwt)
##
## Coefficients:
## (Intercept)   mother.age
##     2655.74        12.43
```

Basic statistical testing

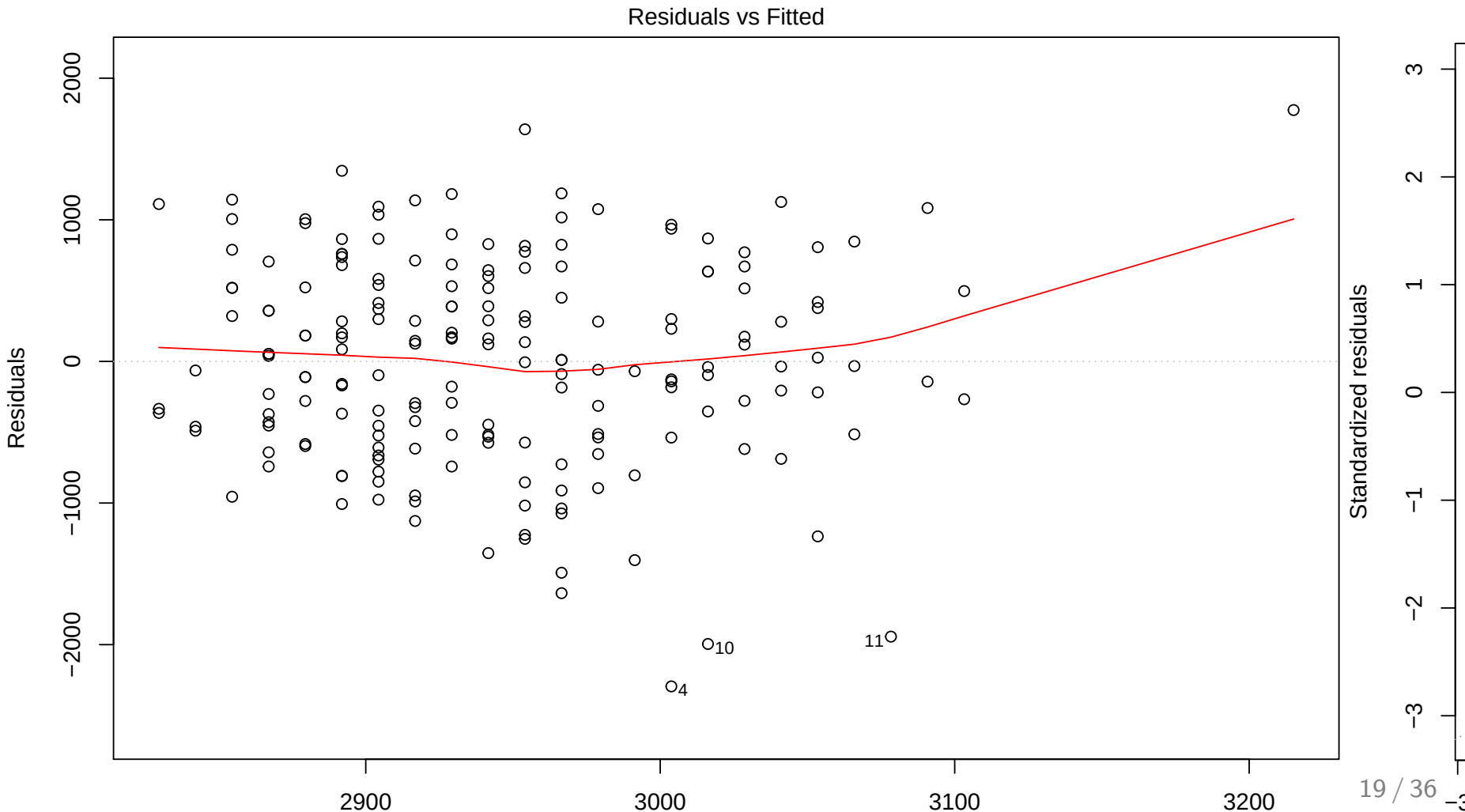
```
summary(linear.model.2)
```

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2294.78  -517.63    10.51   530.80  1774.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2655.74     238.86   11.12   <2e-16 ***
## mother.age     12.43      10.02    1.24    0.216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 728.2 on 187 degrees of freedom
```

Basic statistical testing

Diagnostics: R tries to make it as easy as possible (but no easier).
Try in R proper:

```
plot(linear.model.2)
```



Detecting Outliers

These are the default diagnostic plots for the analysis. Note that our oldest mother and her heaviest child are greatly skewing this analysis as we suspected.

```
birthwt.noout <- birthwt[birthwt$mother.age <= 40,]
linear.model.3 <- lm (birthwt.grams ~ mother.age, data=birt
linear.model.3
```

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age, data = birthwt.
##
## Coefficients:
## (Intercept)   mother.age
##    2833.273        4.344
```

Detecting Outliers

```
summary(linear.model.3)
```

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age, data = birthwt.
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2245.89  -511.24    26.45   540.09  1655.48
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2833.273    244.954   11.57   <2e-16 ***
## mother.age     4.344     10.349    0.42    0.675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 717.2 on 186 degrees of freedom
```

More complex models

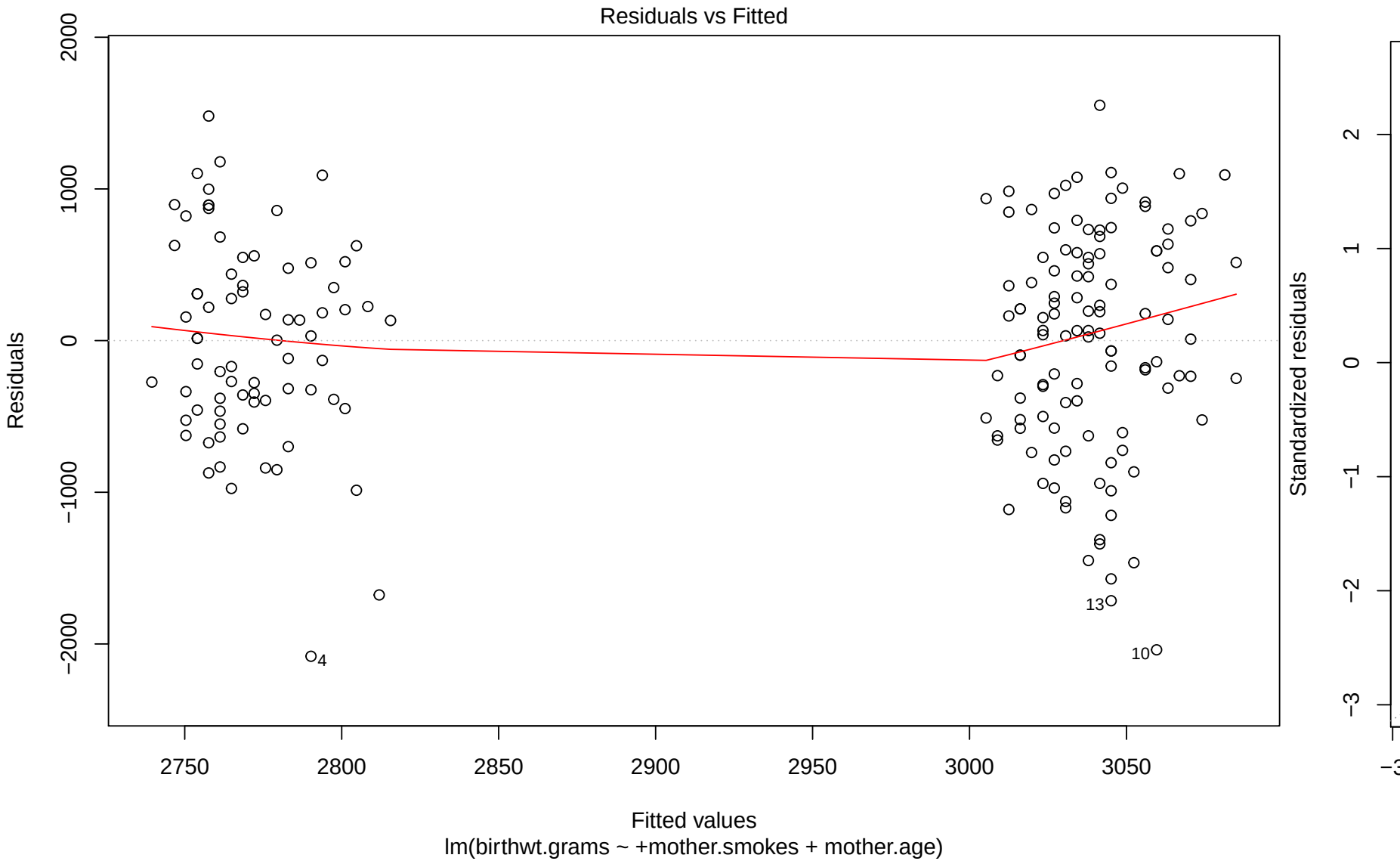
Add in smoking behavior:

```
linear.model.3a <- lm (birthwt.grams ~ + mother.smokes + mo
summary(linear.model.3a)
```

```
##
## Call:
## lm(formula = birthwt.grams ~ +mother.smokes + mother.age
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2081.22  -459.82    43.56   548.22  1551.51
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2954.582    246.280  11.997   <2e-16 **
## mother.smokesYes -265.756    105.605  -2.517   0.0127 *
## mother.age          3.621     10.208   0.355   0.7232
```

More complex models

```
plot(linear.model.3a)
```



More complex models

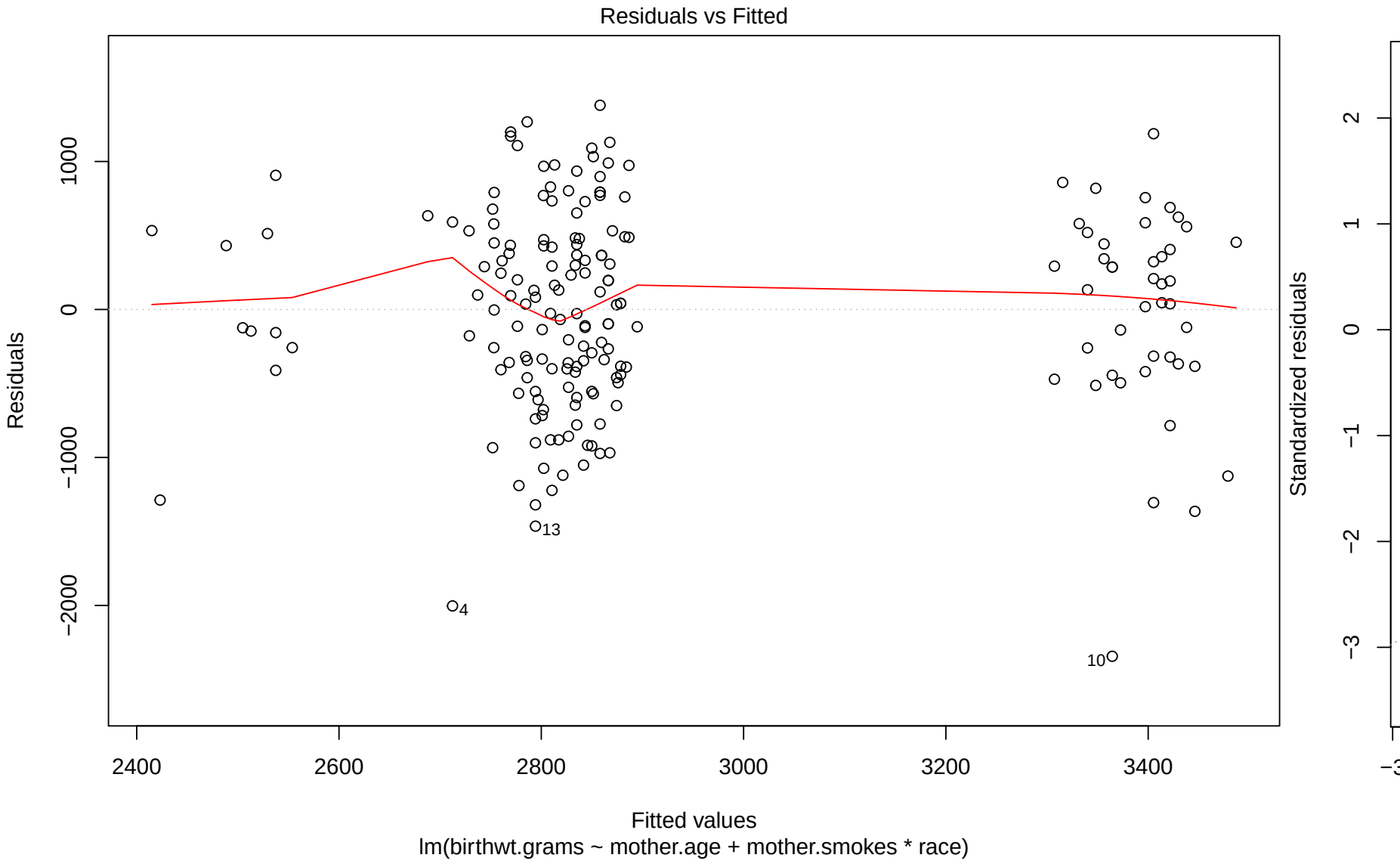
Add in smoking behavior:

```
linear.model.3b <- lm (birthwt.grams ~ mother.age + mother
summary(linear.model.3b)
```

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age + mother.smokes
##     data = birthwt.noout)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2343.52  -413.66    39.91   480.36  1379.90
##
## Coefficients:
##                            Estimate Std. Error t value P
## (Intercept)                3017.352    265.606  11.360
## mother.age                   -8.168     10.276  -0.795
```

More complex models

```
plot(linear.model.3b)
```



Everything Must Go (In)

Let's do a kitchen sink model on this new data set:

```
linear.model.4 <- lm (birthwt.grams ~ ., data=birthwt.noout
linear.model.4
```

```
##
## Call:
## lm(formula = birthwt.grams ~ ., data = birthwt.noout)
##
## Coefficients:
##          (Intercept)   birthwt.below.2500              mothe
##            3360.5163            -1116.3933                -16
##       mother.weight            raceother                race
##              1.9317              68.8145                 247
##    mother.smokesYes  previous.prem.labor      hypertensi
##           -157.7041              95.9825                -185
##      uterine.irrYes     physician.visits
##           -340.0918              -0.3519
```

Everything Must Go (In), Except What Must Not

Whoops! One of those variables was `birthwt.below.2500` which is a function of the outcome.

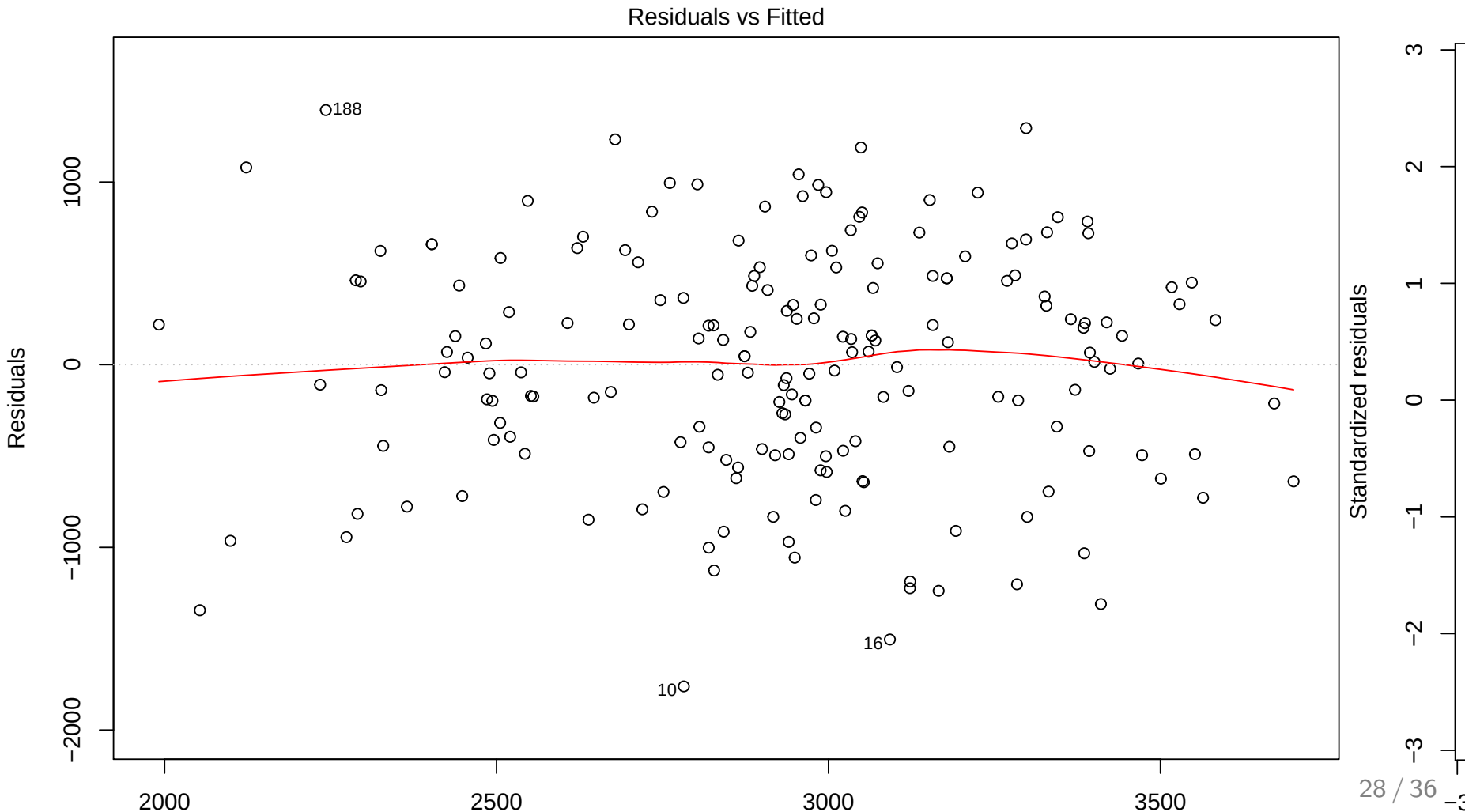
```
linear.model.4a <- lm (birthwt.grams ~ . - birthwt.below.25
summary(linear.model.4a)
```

```
##
## Call:
## lm(formula = birthwt.grams ~ . - birthwt.below.2500, dat
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1761.10  -454.81    46.43   459.78  1394.13
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2545.584    323.204   7.876 3.21e-13
## mother.age          -12.111      9.909  -1.222 0.223243
```

Everything Must Go (In), Except What Must Not

Whoops! One of those variables was `birthwt.below.2500` which is a function of the outcome.

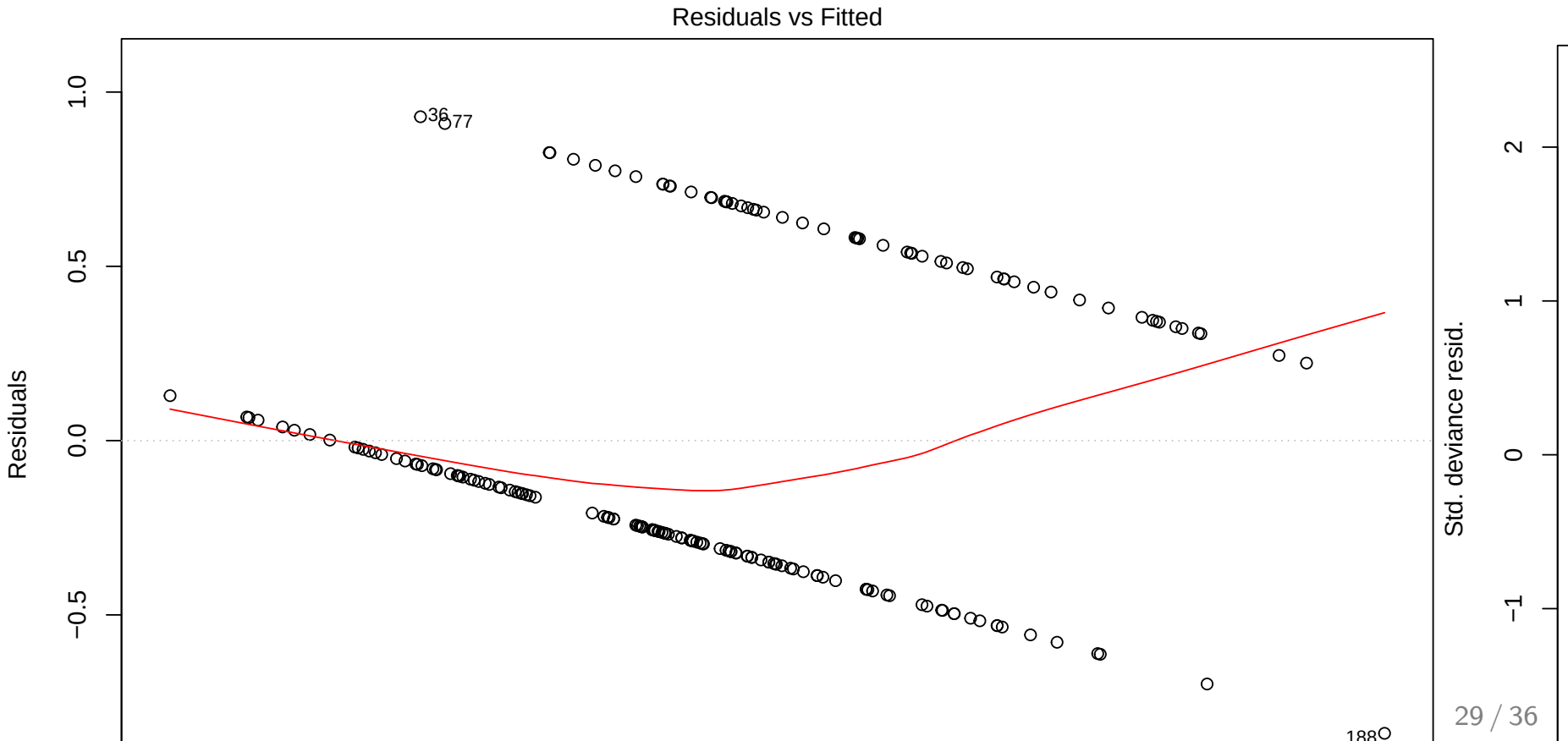
```
plot(linear.model.4a)
```



Generalized Linear Models

Maybe a linear increase in birth weight is less important than if it's below a threshold like 2500 grams (5.5 pounds). Let's fit a generalized linear model instead:

```
glm.0 <- glm (birthwt.below.2500 ~ . - birthwt.grams, data=
plot(glm.0)
```



Generalized Linear Models

The default value is a Gaussian model (a standard linear model).
Change this:

```
glm.1 <- glm (birthwt.below.2500 ~ . - birthwt.grams, data=
```

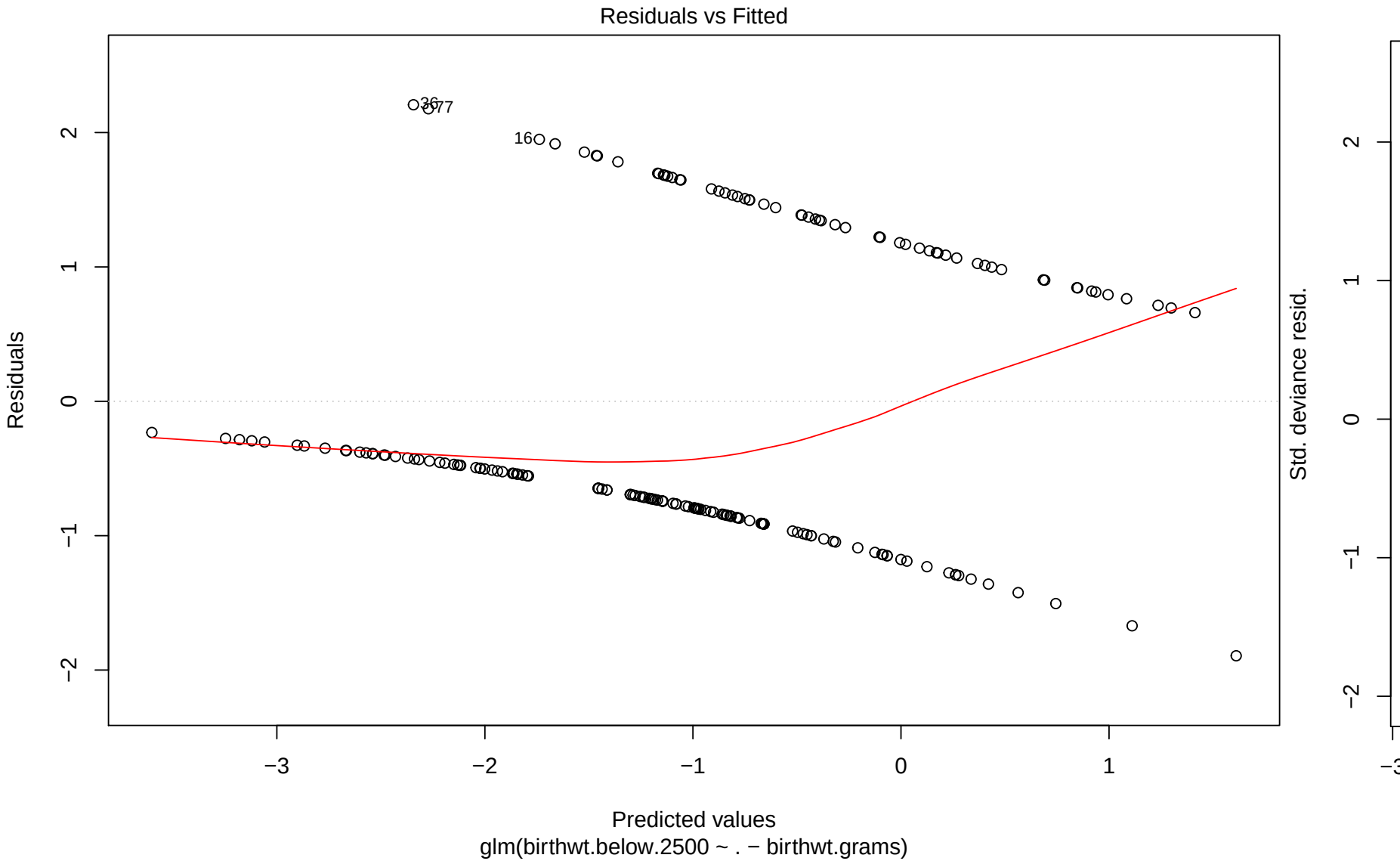
Generalized Linear Models

```
summary(glm.1)
```

```
##
## Call:
## glm(formula = birthwt.below.2500 ~ . - birthwt.grams, fa
##     data = birthwt.noout)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8938  -0.8222  -0.5363   0.9848   2.2069
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z
## (Intercept)        1.721830   1.258897   1.368  0.1714
## mother.age        -0.027537   0.037718  -0.730  0.4653
## mother.weight     -0.015474   0.006919  -2.237  0.0253
## raceother         -0.395505   0.537685  -0.736  0.4619
## racewhite         -1.269006   0.527180  -2.407  0.0160
```

Generalized Linear Models

```
plot(glm.1)
```



What Do We Do With This, Anyway?

Let's take a subset of this data to do predictions.

```
odds <- seq(1, nrow(birthwt.noout), by=2)
birthwt.in <- birthwt.noout[odds,]
birthwt.out <- birthwt.noout[-odds,]
linear.model.half <-
  lm (birthwt.grams ~
      . - birthwt.below.2500, data=birthwt.in)
```

What Do We Do With This, Anyway?

```
summary (linear.model.half)
```

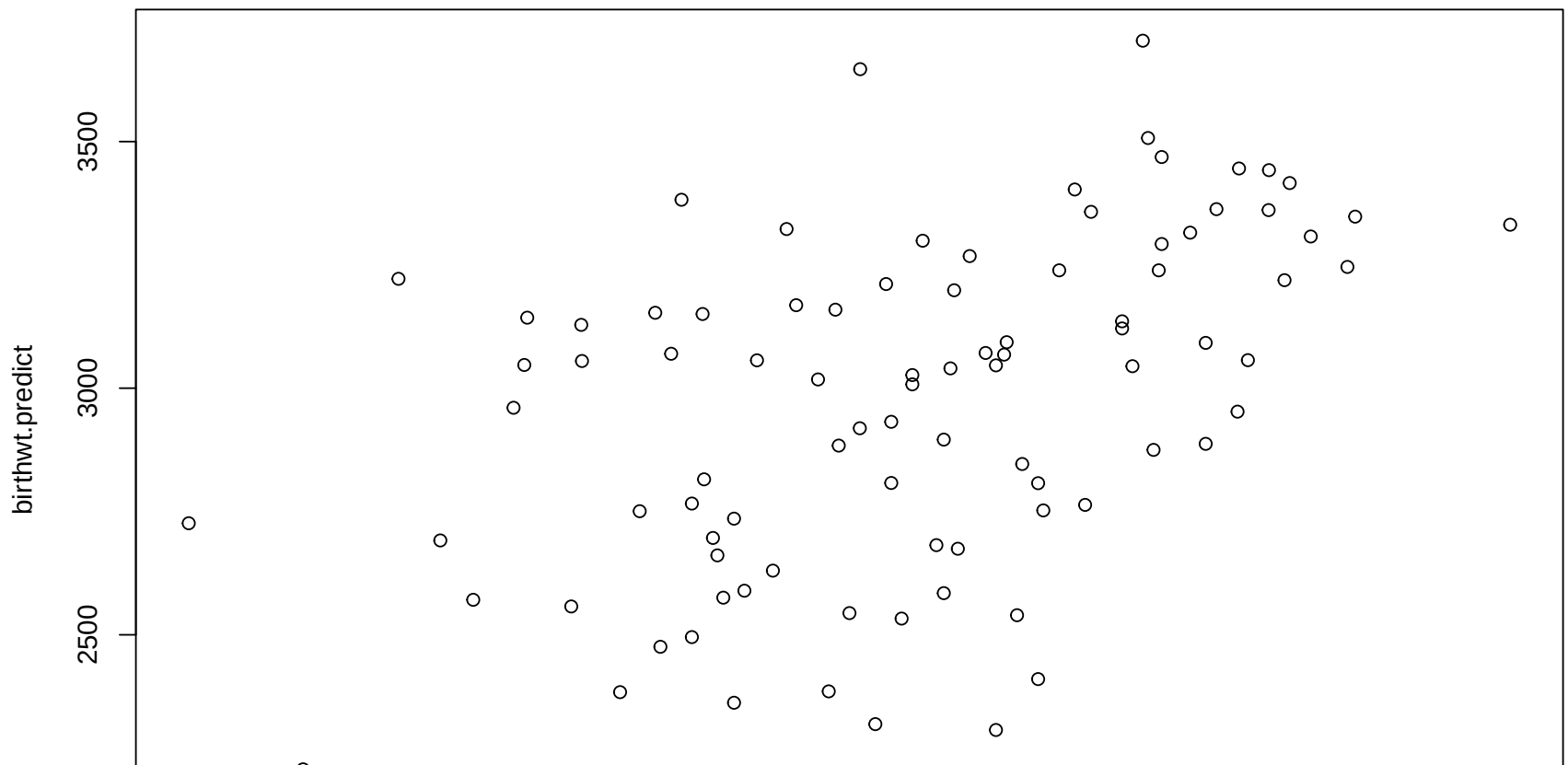
```
##
## Call:
## lm(formula = birthwt.grams ~ . - birthwt.below.2500, dat
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1705.17  -303.11    26.48   427.18  1261.57
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2514.891    450.245   5.586 2.81e-07
## mother.age           7.052     14.935   0.472  0.63801
## mother.weight        2.683      2.885   0.930  0.35501
## raceother          113.948    224.519   0.508  0.61312
## racewhite          466.219    204.967   2.275  0.02548
## mother.smokesYes  -217.218    154.521  -1.406  0.16349
```

What Do We Do With This, Anyway?

```
birthwt.predict <- predict (linear.model.half)
cor (birthwt.in$birthwt.grams, birthwt.predict)
```

```
## [1] 0.508442
```

```
plot (birthwt.in$birthwt.grams, birthwt.predict)
```



What Do We Do With This, Anyway?

```
birthwt.predict.out <- predict (linear.model.half, birthwt.
cor (birthwt.out$birthwt.grams, birthwt.predict.out)
```

```
## [1] 0.3749431
```

```
plot (birthwt.out$birthwt.grams, birthwt.predict.out)
```