Lecture 9: Models With Data

36-350

24 September 2014

In Previous Episodes

- Until now: processing existing data into R
- String manipulation, scraping and collecting data

Today

- Using data frames for statistical purposes
- Manipulation of data into more convenient forms
- (Re-)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           race      
##  Min.   :0.000   Min.   :14.0   Min.   : 80   Min.   :1.00  
##  1st Qu.:0.000   1st Qu.:19.0   1st Qu.:110   1st Qu.:1.00  
##  Median :0.000   Median :23.0   Median :121   Median :1.00  
##  Mean   :0.312   Mean   :23.2   Mean   :130   Mean   :1.85  
##  3rd Qu.:1.000   3rd Qu.:26.0   3rd Qu.:140   3rd Qu.:3.00  
##  Max.   :1.000   Max.   :45.0   Max.   :250   Max.   :3.00  
##      smoke            ptl              ht                ui        
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.000  
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000  
##  Median :0.000   Median :0.000   Median :0.0000   Median :0.000  
##  Mean   :0.392   Mean   :0.196   Mean   :0.0635   Mean   :0.148  
##  3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:0.000
```

```
##  Max.   :1.000   Max.   :3.000   Max.   :1.0000   Max.   :1.000
##       ftv             bwt
##  Min.   :0.000   Min.   : 709
##  1st Qu.:0.000   1st Qu.:2414
##  Median :0.000   Median :2977
##  Mean   :0.794   Mean   :2945
##  3rd Qu.:1.000   3rd Qu.:3487
##  Max.   :6.000   Max.   :4990
```

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"    "ui"
##  [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")
```

Make it readable, again!

Let's make all 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)
```

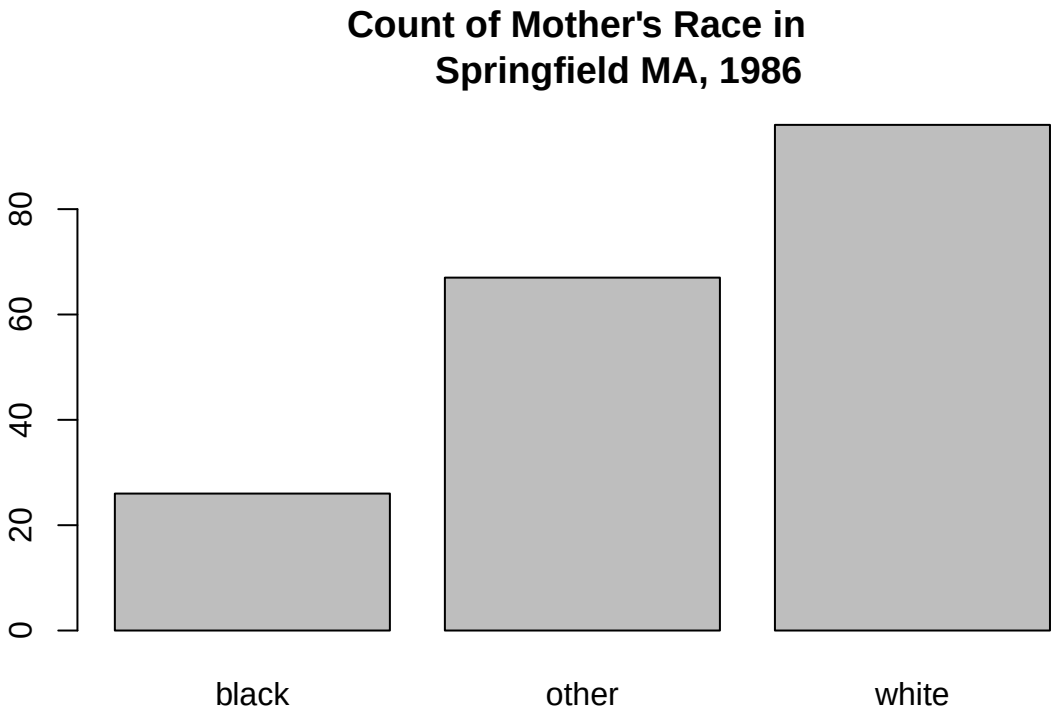
```
##  birthwt.below.2500   mother.age   mother.weight    race    mother.smokes
##  Min.   :0.000      Min.   :14.0   Min.   : 80   black:26   No :115
```

```
##  1st Qu.:0.000      1st Qu.:19.0   1st Qu.:110   other:67   Yes: 74
##  Median :0.000      Median :23.0   Median :121   white:96
##  Mean   :0.312      Mean   :23.2   Mean   :130
##  3rd Qu.:1.000      3rd Qu.:26.0   3rd Qu.:140
##  Max.   :1.000      Max.   :45.0   Max.   :250
##  previous.prem.labor hypertension uterine.irr physician.visits
##  Min.   :0.000       No :177      No :161     Min.   :0.000
##  1st Qu.:0.000       Yes: 12      Yes: 28     1st Qu.:0.000
##  Median :0.000                                Median :0.000
##  Mean   :0.196                                Mean   :0.794
##  3rd Qu.:0.000                                3rd Qu.:1.000
##  Max.   :3.000                                Max.   :6.000
##  birthwt.grams
##  Min.   : 709
##  1st Qu.:2414
##  Median :2977
##  Mean   :2945
##  3rd Qu.:3487
##  Max.   :4990
```

Explore it!

R's basic plotting functions go a long way.

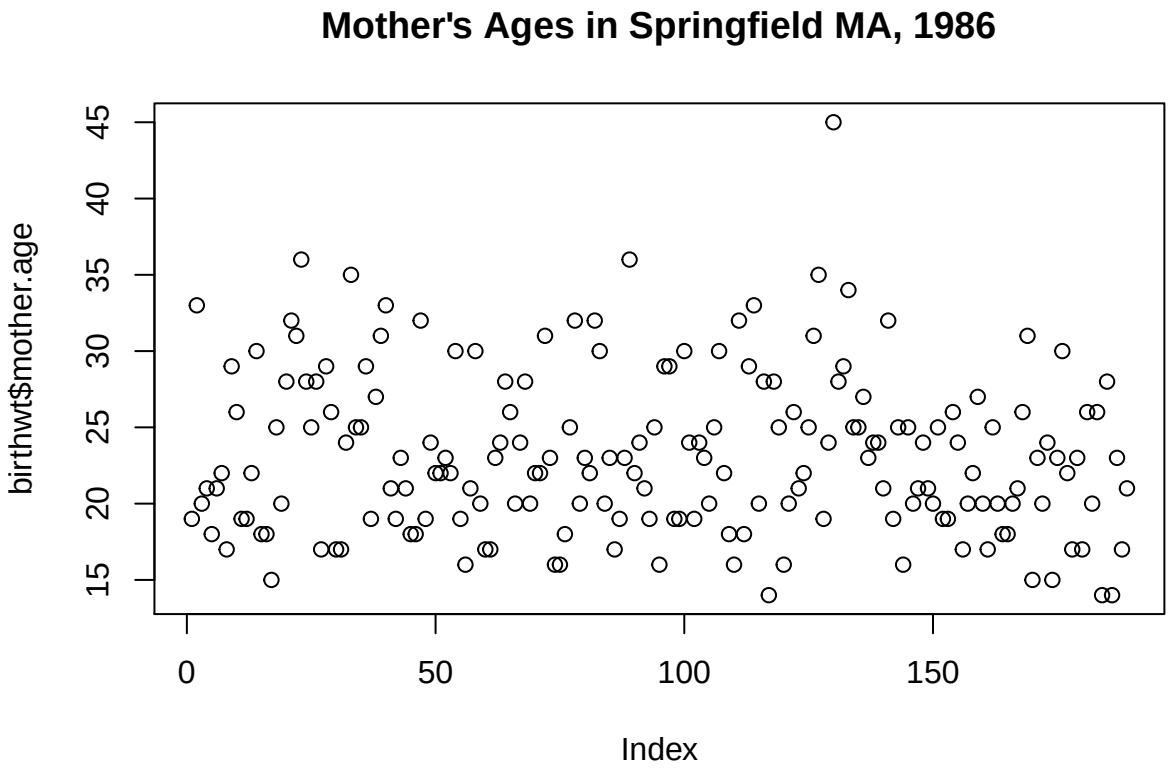
```
plot (birthwt$race)
title (main = "Count of Mother's Race in
       Springfield MA, 1986")
```



Explore it!

R's basic plotting functions go a long way.

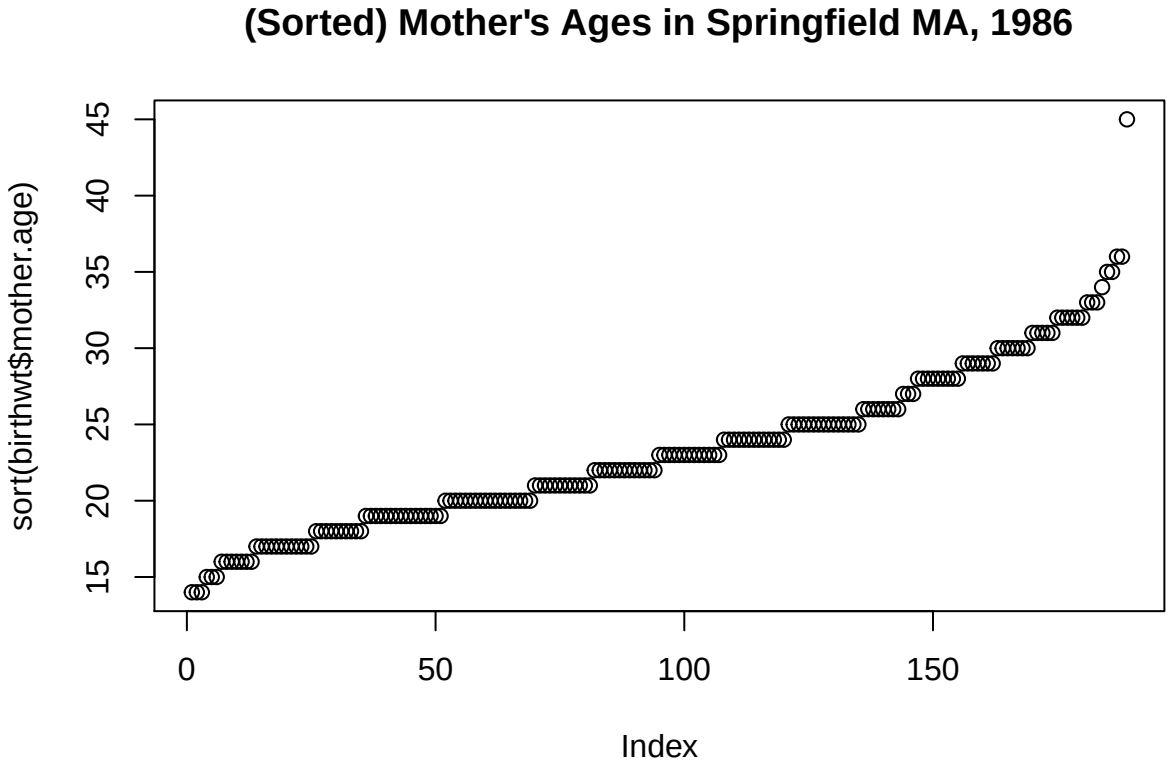
```
plot (birthwt$mother.age)
title (main = "Mother's Ages in Springfield MA, 1986")
```



Explore it!

R's basic plotting functions go a long way.

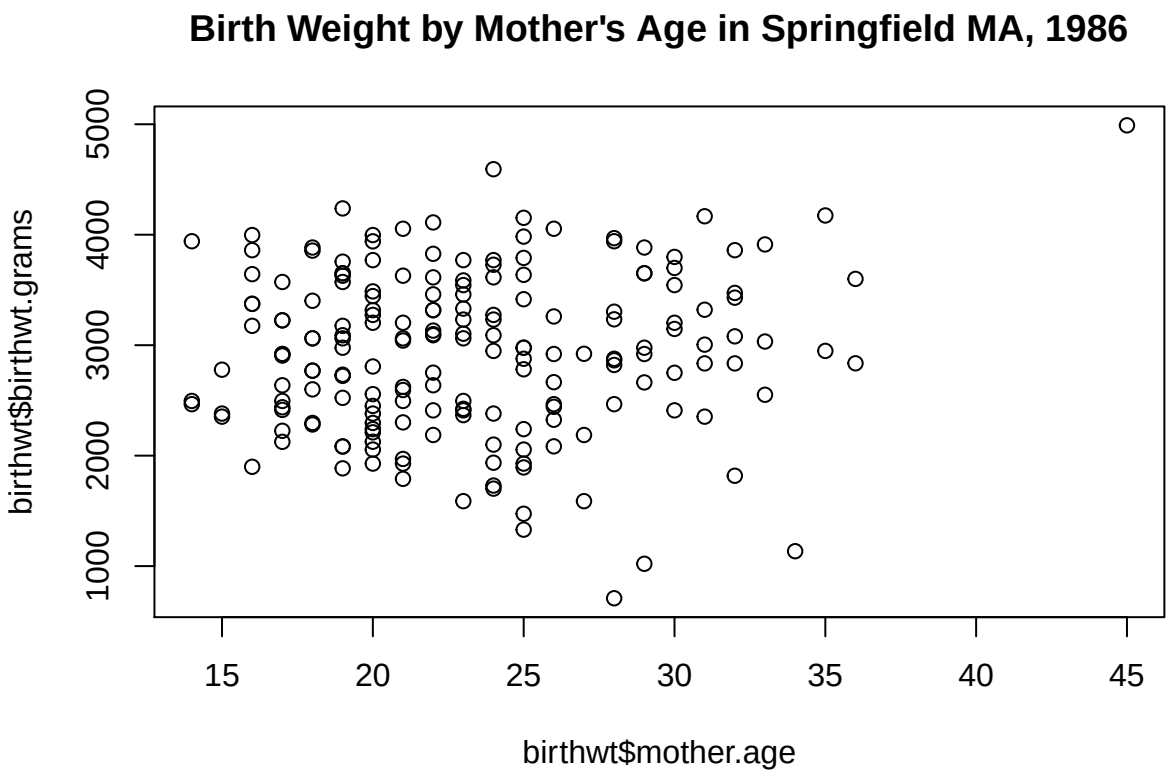
```
plot (sort(birthwt$mother.age))
title (main = "(Sorted) Mother's Ages in Springfield MA, 1986")
```



Explore it!

R's basic plotting functions go a long way.

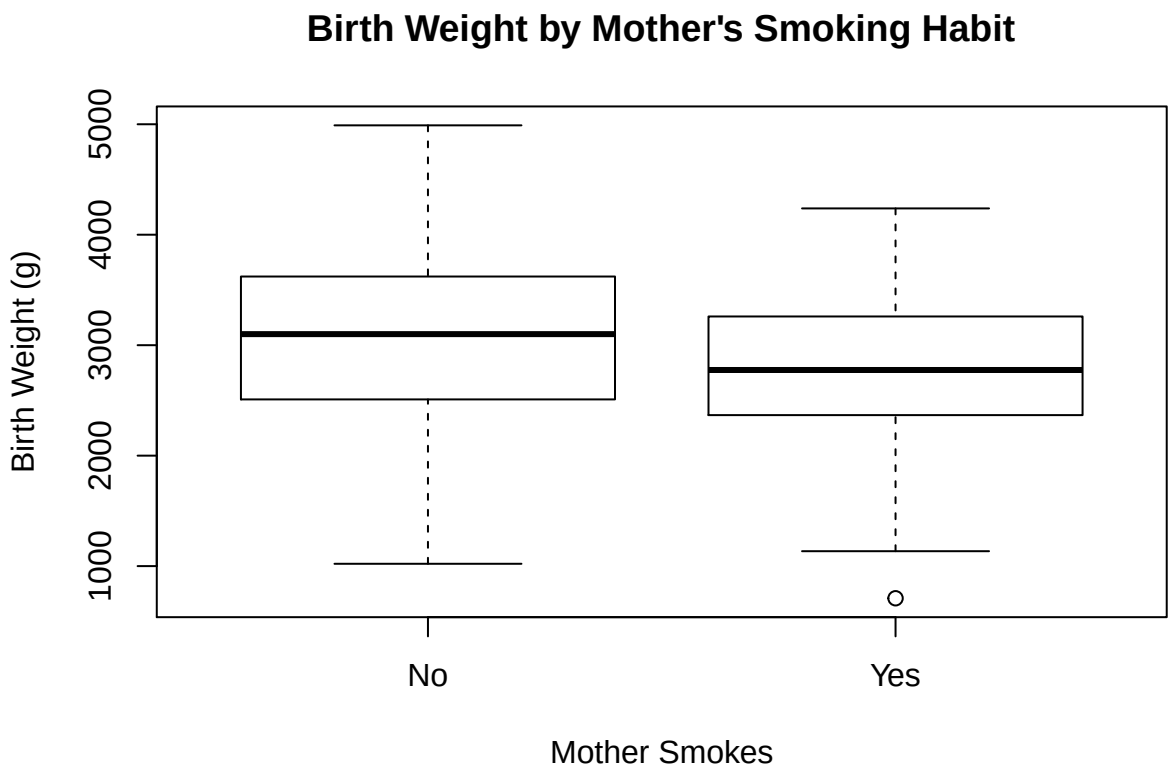
```
plot (birthwt$mother.age, birthwt$birthwt.grams)
title (main = "Birth Weight by Mother's Age in Springfield MA, 1986")
```



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="Birth Weight by Mother's Smoking Habit", ylab
```



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 == "Yes"] and birthwt$birthwt.grams[birthwt$mothe
## t = -2.73, df = 170.1, p-value = 0.007003
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -488.98  -78.57
## sample estimates:
## mean of x mean of y
##      2772      3056
```

Basic statistical testing

Does this difference match the linear model?

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

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.smokes, data = birthwt)
##
## Coefficients:
##      (Intercept)  mother.smokesYes
##             3056              -284
```

Basic statistical testing

Does this difference match the linear model?

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

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.smokes, data = birthwt)
##
## 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)        3055.7       66.9   45.65   <2e-16 ***
## mother.smokesYes   -283.8      107.0   -2.65   0.0087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 718 on 187 degrees of freedom
## Multiple R-squared:  0.0363, Adjusted R-squared:  0.0311
## F-statistic: 7.04 on 1 and 187 DF,  p-value: 0.00867
```

Basic statistical testing

Does this difference match the linear model?

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

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age, data = birthwt)
##
## Coefficients:
## (Intercept)   mother.age
##      2655.7         12.4
```

Basic statistical testing

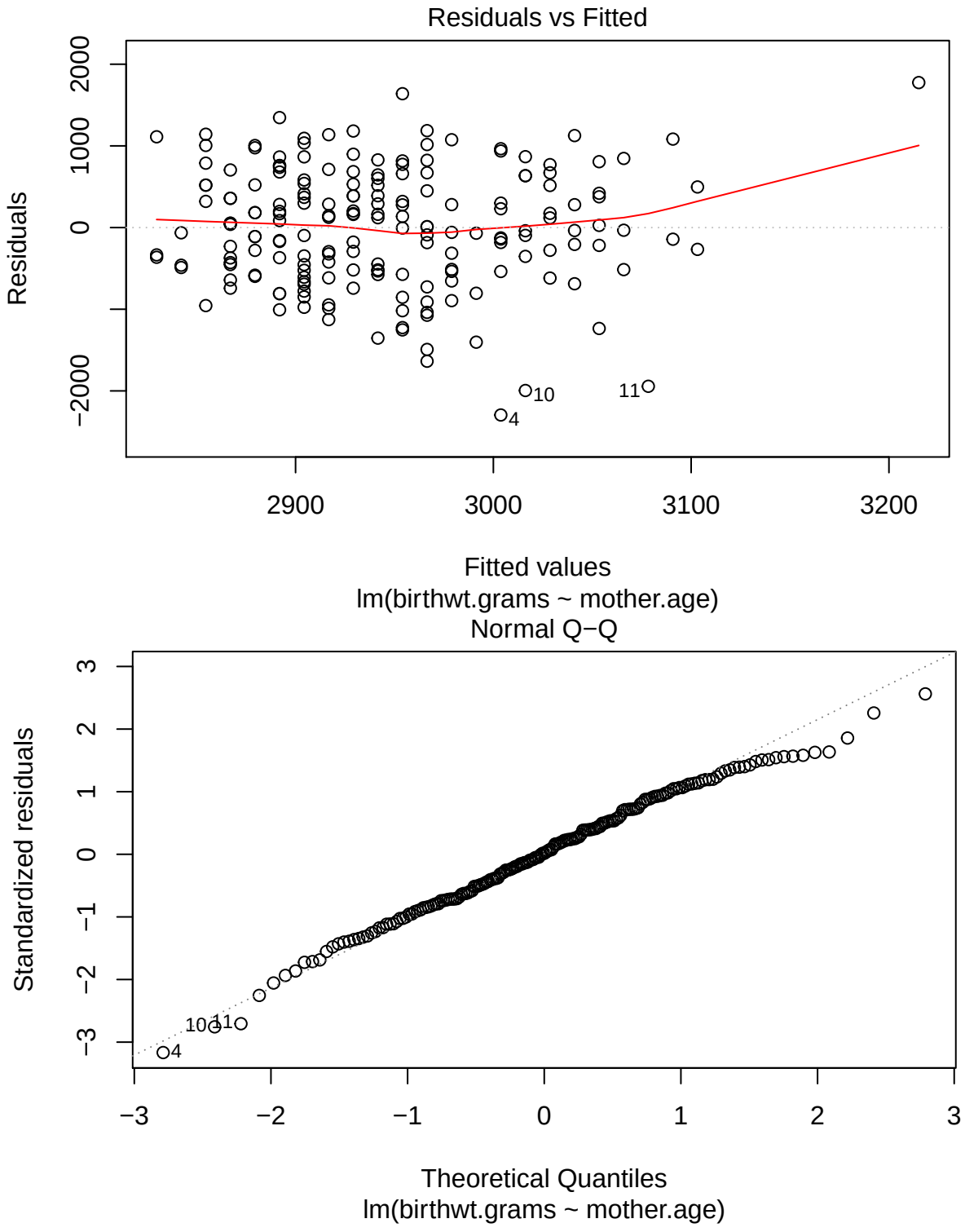
```
summary(linear.model.2)
```

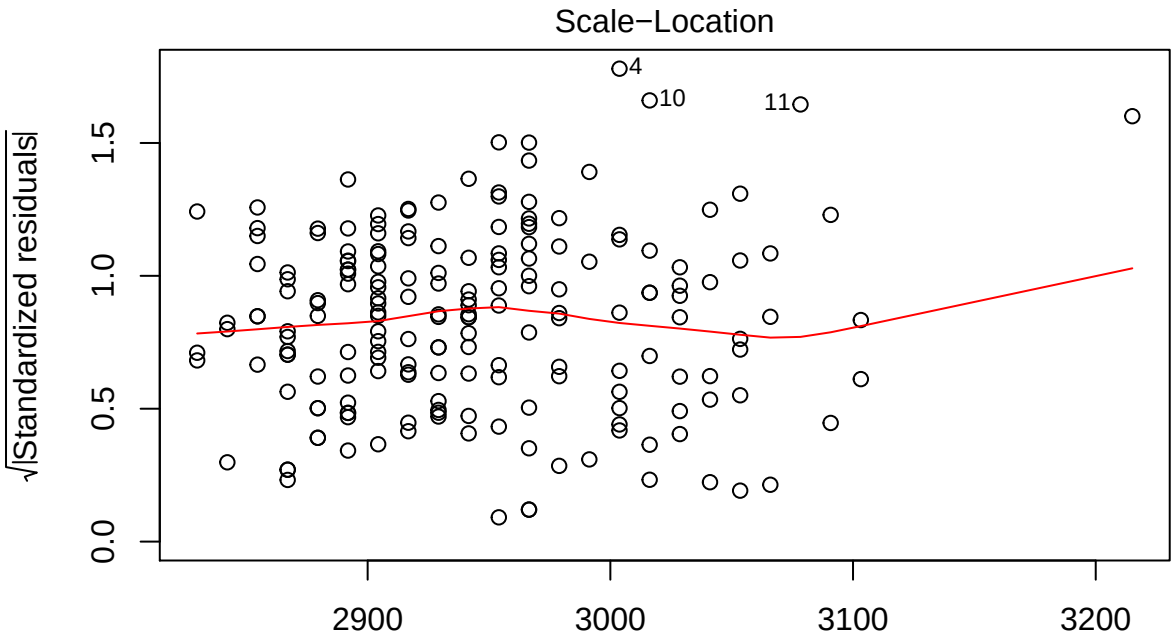
```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age, data = birthwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2294.8  -517.6    10.5   530.8  1774.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2655.7      238.9   11.12   <2e-16 ***
## mother.age      12.4       10.0    1.24     0.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 728 on 187 degrees of freedom
## Multiple R-squared:  0.00816,    Adjusted R-squared:  0.00285
## F-statistic: 1.54 on 1 and 187 DF,  p-value: 0.216
```

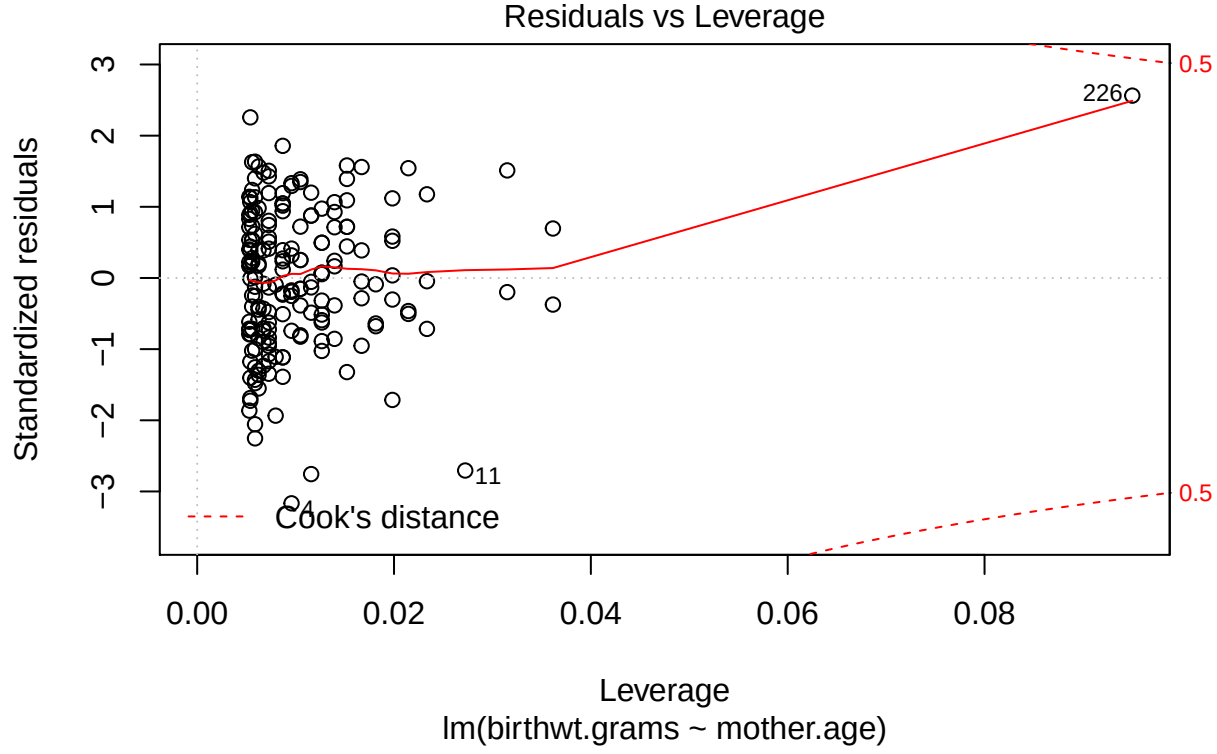
Basic statistical testing

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

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


Residuals vs Fitted
Residuals
2000
1000
0
−2000
2900
3000
3100
3200
10
11
4
Fitted values
lm(birthwt.grams ~ mother.age)
Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
10
11
4
Theoretical Quantiles
lm(birthwt.grams ~ mother.age)





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=birthwt.noout)
linear.model.3
```

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age, data = birthwt.noout)
##
## Coefficients:
## (Intercept)   mother.age
##     2833.27         4.34
```

Detecting Outliers

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

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age, data = birthwt.noout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2245.9  -511.2    26.4   540.1  1655.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2833.27     244.95   11.57   <2e-16 ***
## mother.age      4.34      10.35    0.42     0.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 717 on 186 degrees of freedom
## Multiple R-squared:  0.000946,   Adjusted R-squared:  -0.00443
## F-statistic: 0.176 on 1 and 186 DF,  p-value: 0.675
```

More complex models

Add in smoking behavior:

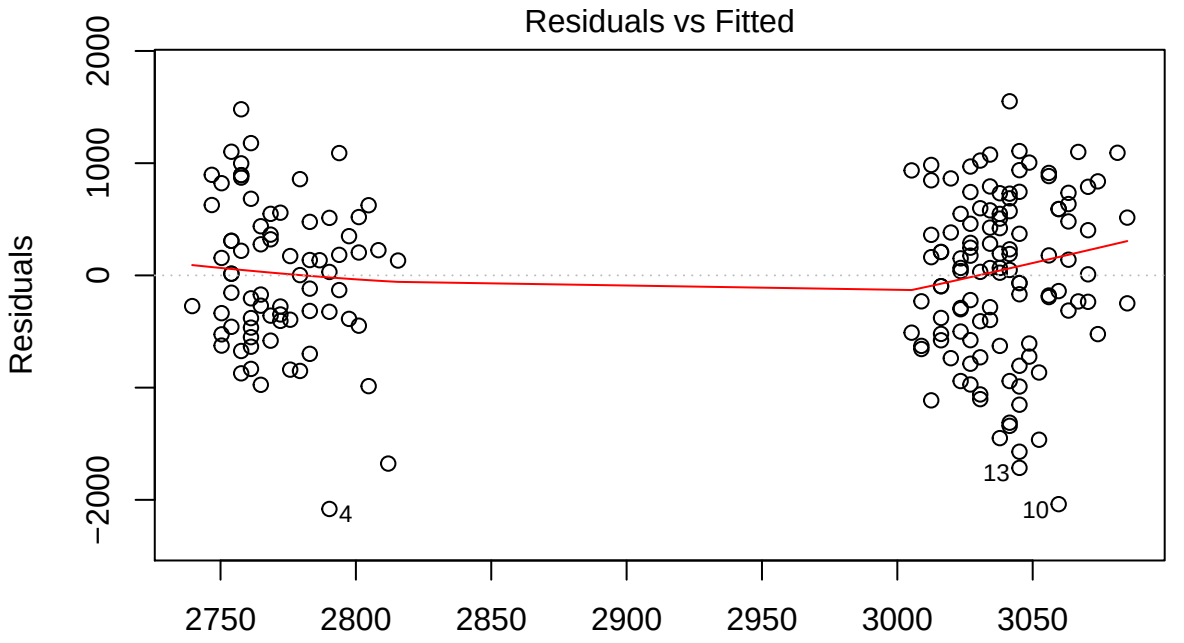
```
linear.model.3a <- lm (birthwt.grams ~ + mother.smokes + mother.age, data=birthwt.noout)
summary(linear.model.3a)
```

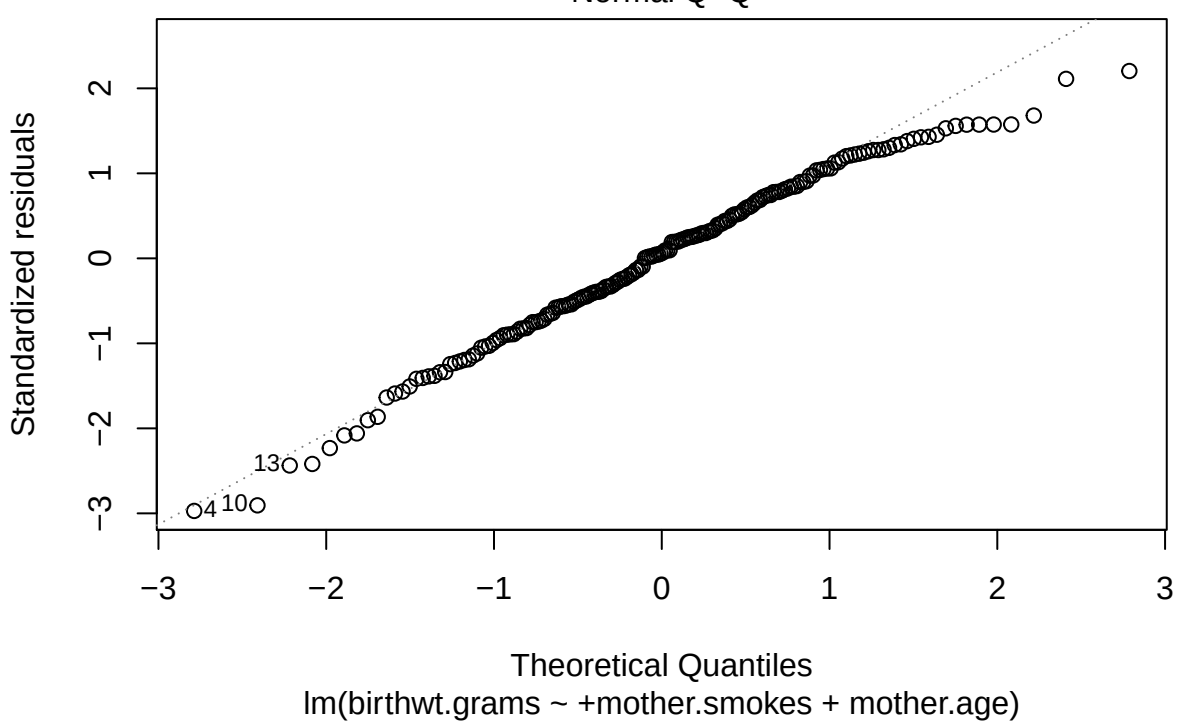
```
##
## Call:
## lm(formula = birthwt.grams ~ +mother.smokes + mother.age, data = birthwt.noout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

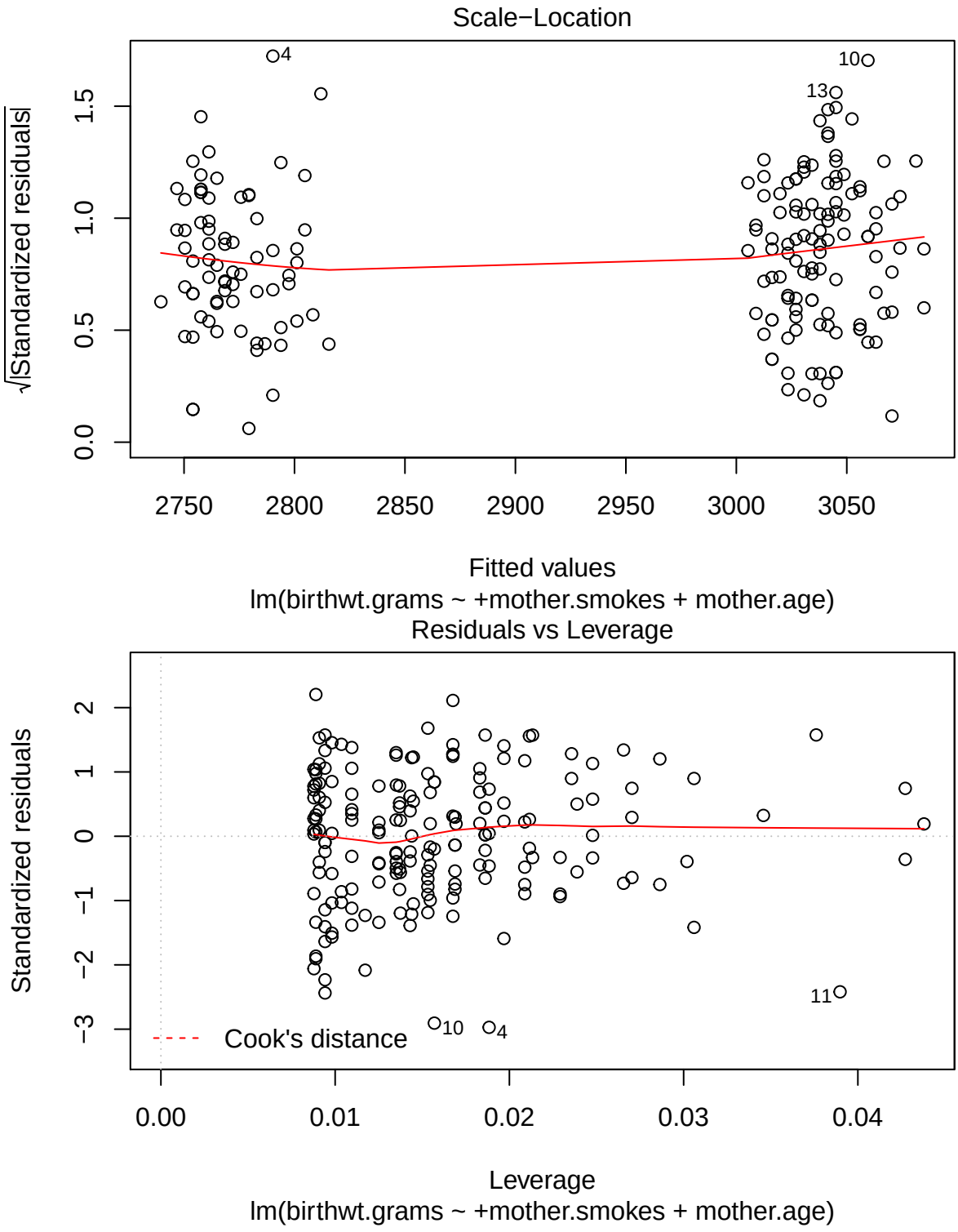
```
## -2081.2  -459.8    43.6   548.2  1551.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2954.58     246.28   12.00   <2e-16 ***
## mother.smokesYes  -265.76     105.60   -2.52    0.013 *
## mother.age           3.62      10.21    0.35    0.723
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 707 on 185 degrees of freedom
## Multiple R-squared:  0.034,  Adjusted R-squared:  0.0236
## F-statistic: 3.26 on 2 and 185 DF,  p-value: 0.0407
```

More complex models

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


Residuals vs Fitted
Residuals
2000
1000
0
−2000
2750
2800
2850
2900
2950
3000
3050
4
13
10
Fitted values
lm(birthwt.grams ~ +mother.smokes + mother.age)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
13
4
10
Theoretical Quantiles
lm(birthwt.grams ~ +mother.smokes + mother.age)



More complex models

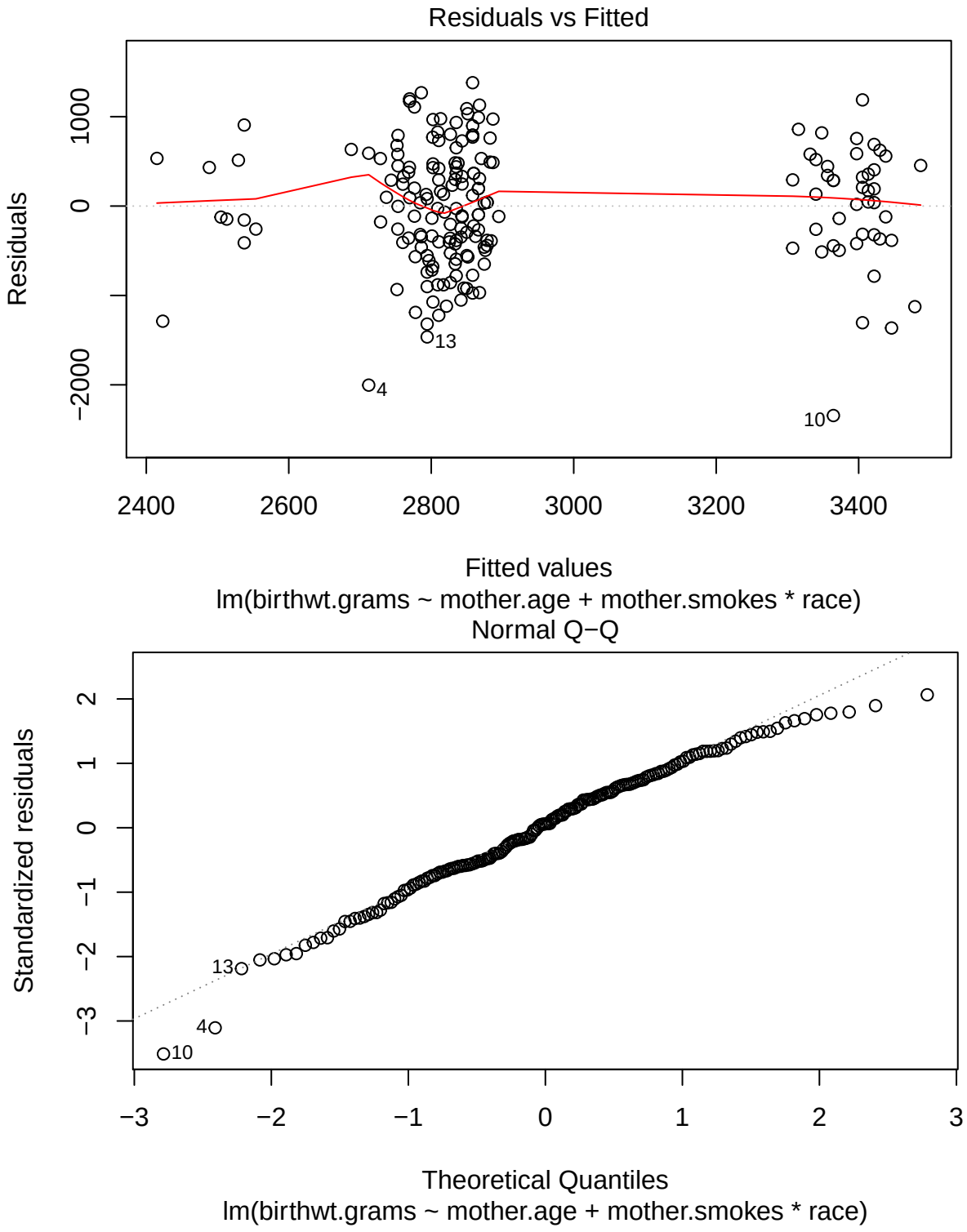
Add in smoking behavior:

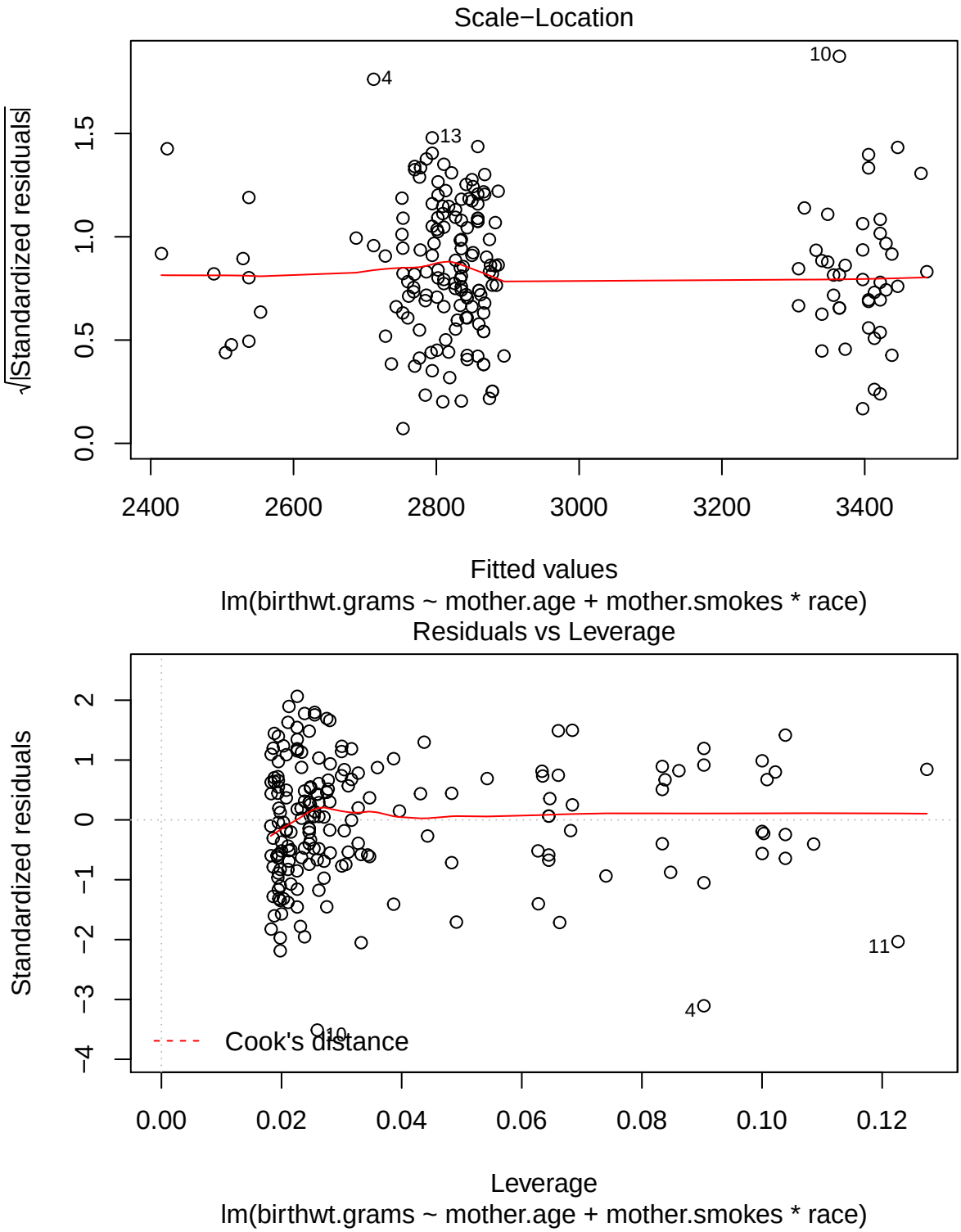
```r
linear.model.3b <- lm (birthwt.grams ~ mother.age + mother.smokes*race, data=birthwt.noout)
summary(linear.model.3b)
```

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age + mother.smokes * race,
##     data = birthwt.noout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2343.5  -413.7    39.9   480.4  1379.9
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3017.35     265.61   11.36   <2e-16 ***
## mother.age                    -8.17      10.28   -0.79   0.4277
## mother.smokesYes            -316.50     275.90   -1.15   0.2528
## raceother                    -18.90     193.67   -0.10   0.9224
## racewhite                    584.04     206.32    2.83   0.0052 **
## mother.smokesYes:raceother   259.00     349.87    0.74   0.4601
## mother.smokesYes:racewhite  -271.59     314.27   -0.86   0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 676 on 181 degrees of freedom
## Multiple R-squared:  0.136,  Adjusted R-squared:  0.107
## F-statistic: 4.75 on 6 and 181 DF,  p-value: 0.000163
```

More complex models

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


Residuals vs Fitted
Residuals
1000
0
−2000
13
4
10
2400
2600
2800
3000
3200
3400
Fitted values
lm(birthwt.grams ~ mother.age + mother.smokes * race)
Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
13
4
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(birthwt.grams ~ mother.age + mother.smokes * race)



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           mother.age
##            3360.516            -1116.393               -16.032
##       mother.weight            raceother            racewhite
##               1.932               68.814               247.024
##    mother.smokesYes  previous.prem.labor      hypertensionYes
##            -157.704               95.982              -185.278
##      uterine.irrYes     physician.visits
##            -340.092               -0.352
```

Everything Must Go (In), Except What Must Not

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

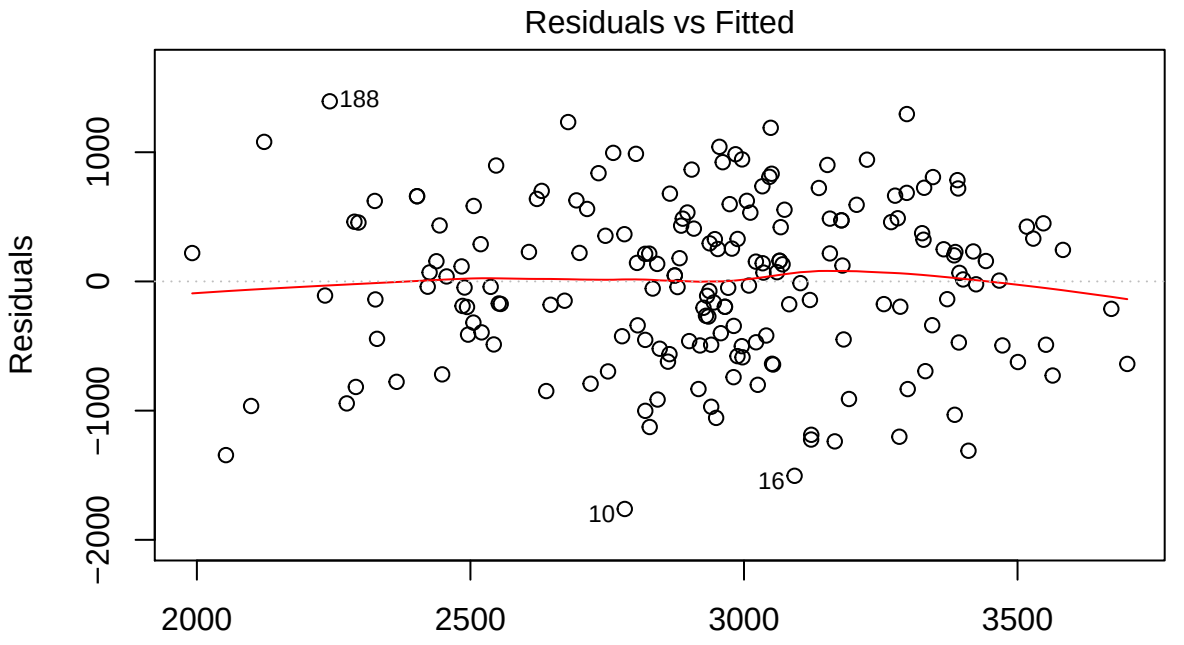
```
linear.model.4a <- lm (birthwt.grams ~ . - birthwt.below.2500, data=birthwt.noout)
summary(linear.model.4a)
```

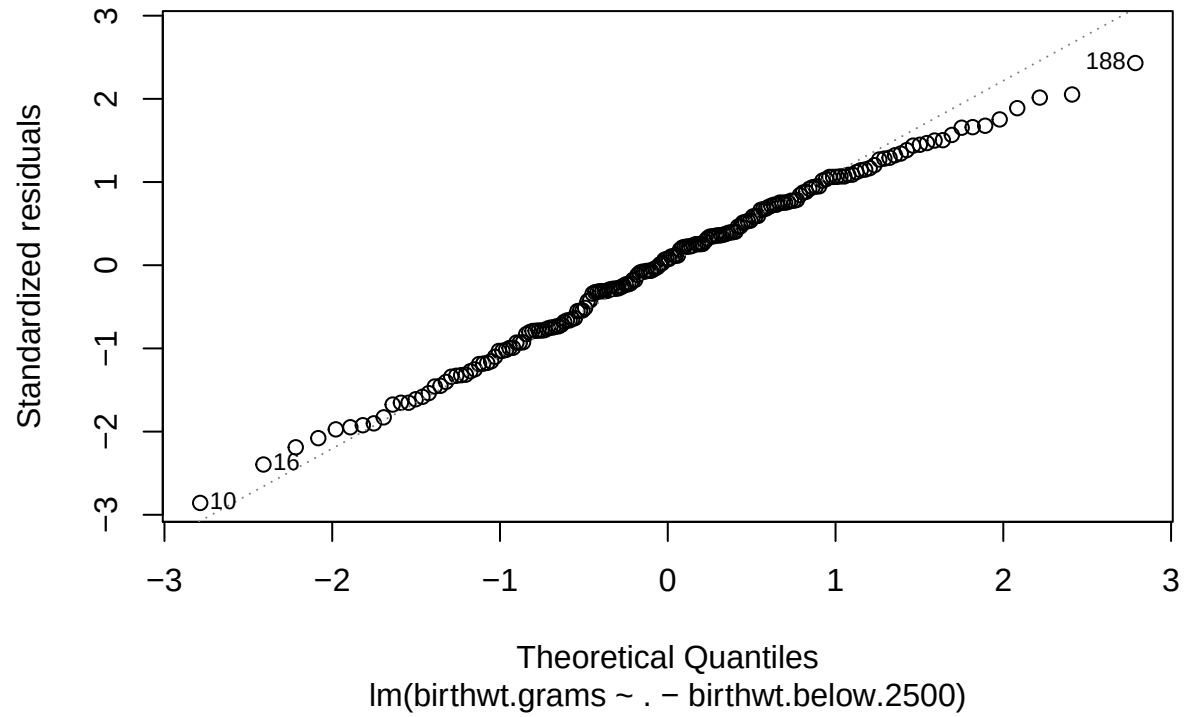
```
##
## Call:
## lm(formula = birthwt.grams ~ . - birthwt.below.2500, data = birthwt.noout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1761.1  -454.8    46.4   459.8  1394.1
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2545.58     323.20    7.88  3.2e-13 ***
## mother.age            -12.11       9.91   -1.22  0.22324
## mother.weight           4.79       1.71    2.80  0.00566 **
## raceother             155.61     156.56    0.99  0.32163
## racewhite             494.54     147.15    3.36  0.00095 ***
## mother.smokesYes     -335.79     104.61   -3.21  0.00158 **
## previous.prem.labor   -32.92     100.19   -0.33  0.74284
## hypertensionYes      -594.32     198.48   -2.99  0.00314 **
## uterine.irrYes       -514.84     136.25   -3.78  0.00021 ***
## physician.visits       -7.25      45.65   -0.16  0.87404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 638 on 178 degrees of freedom
## Multiple R-squared:  0.243,  Adjusted R-squared:  0.205
## F-statistic: 6.37 on 9 and 178 DF,  p-value: 8.26e-08
```

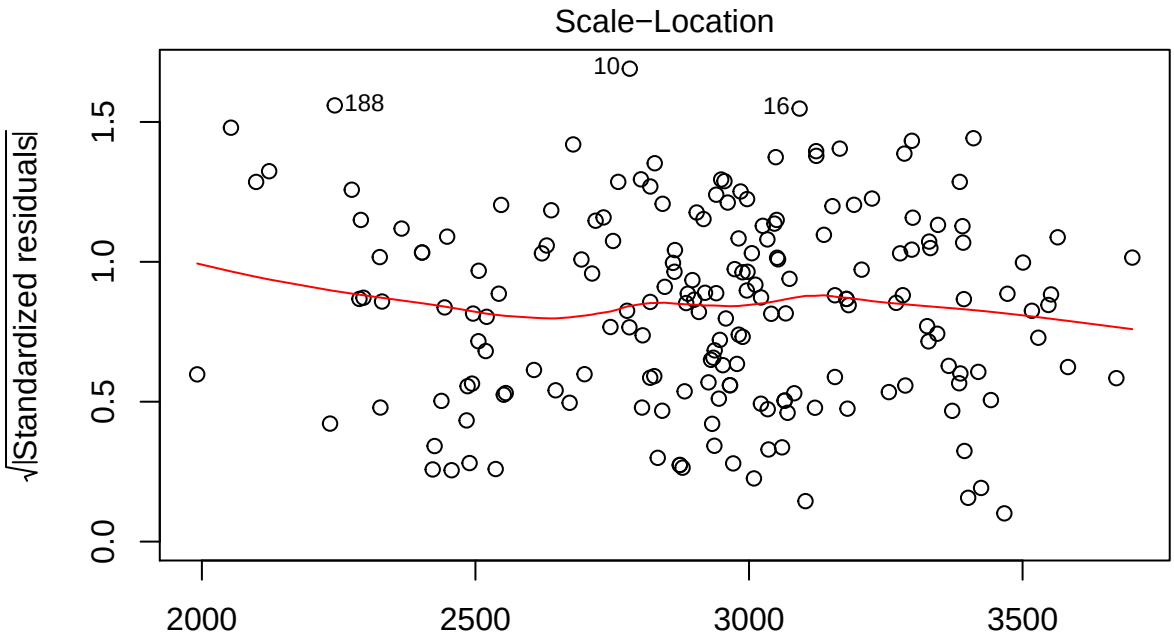
Everything Must Go (In), Except What Must Not

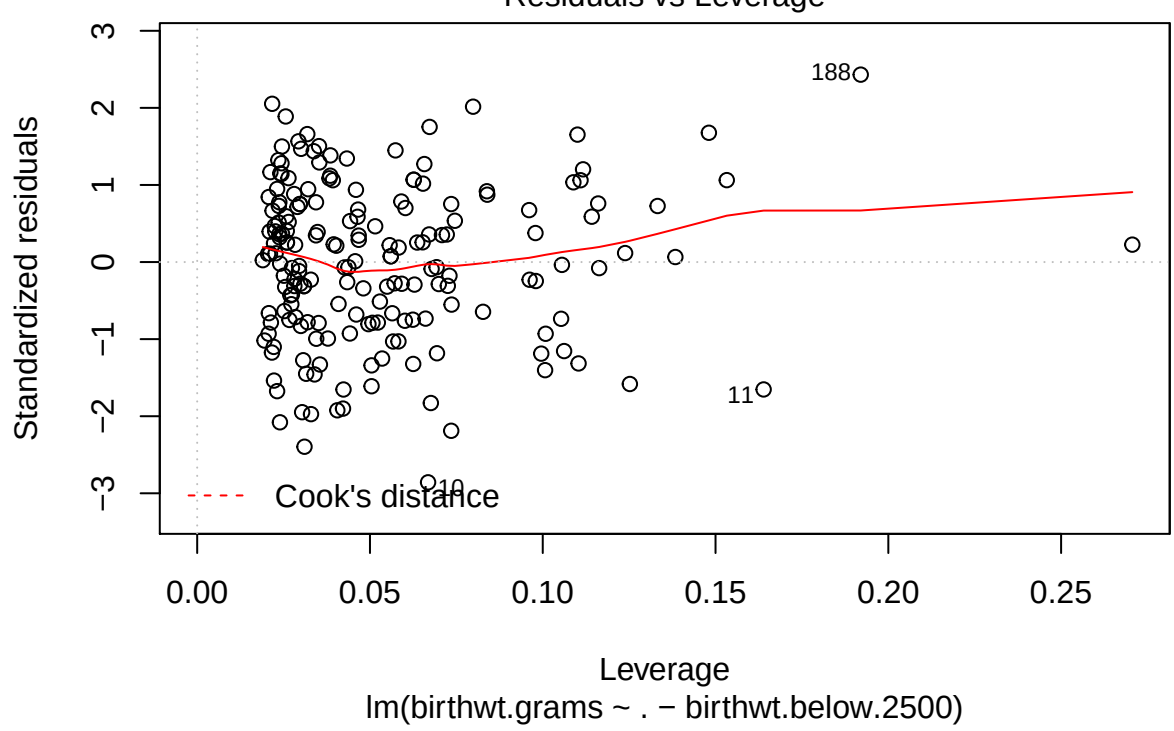
Whoops! One of those variables was `birthwt.below.2500` which is a function of the outcome.

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





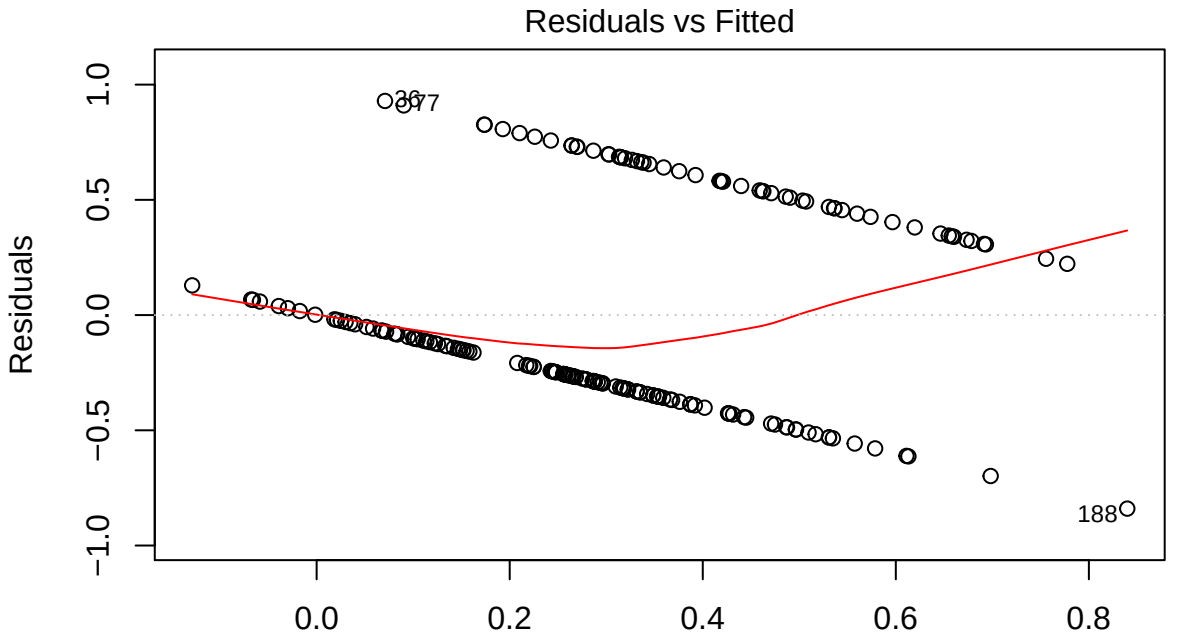


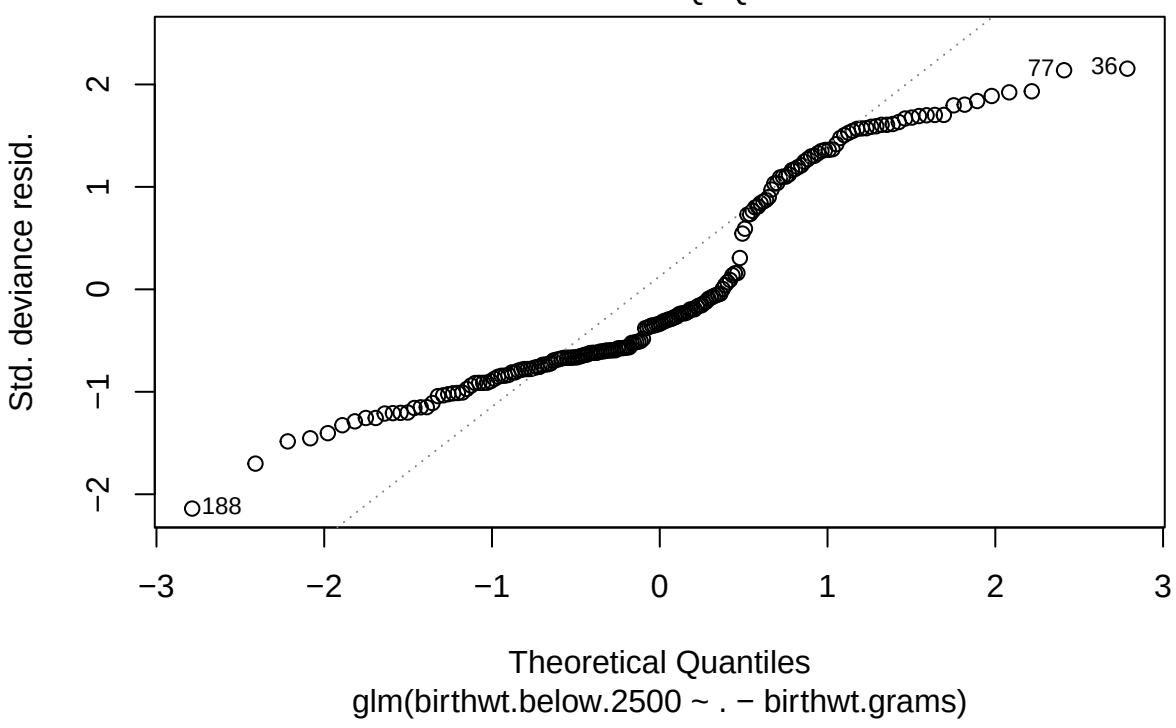


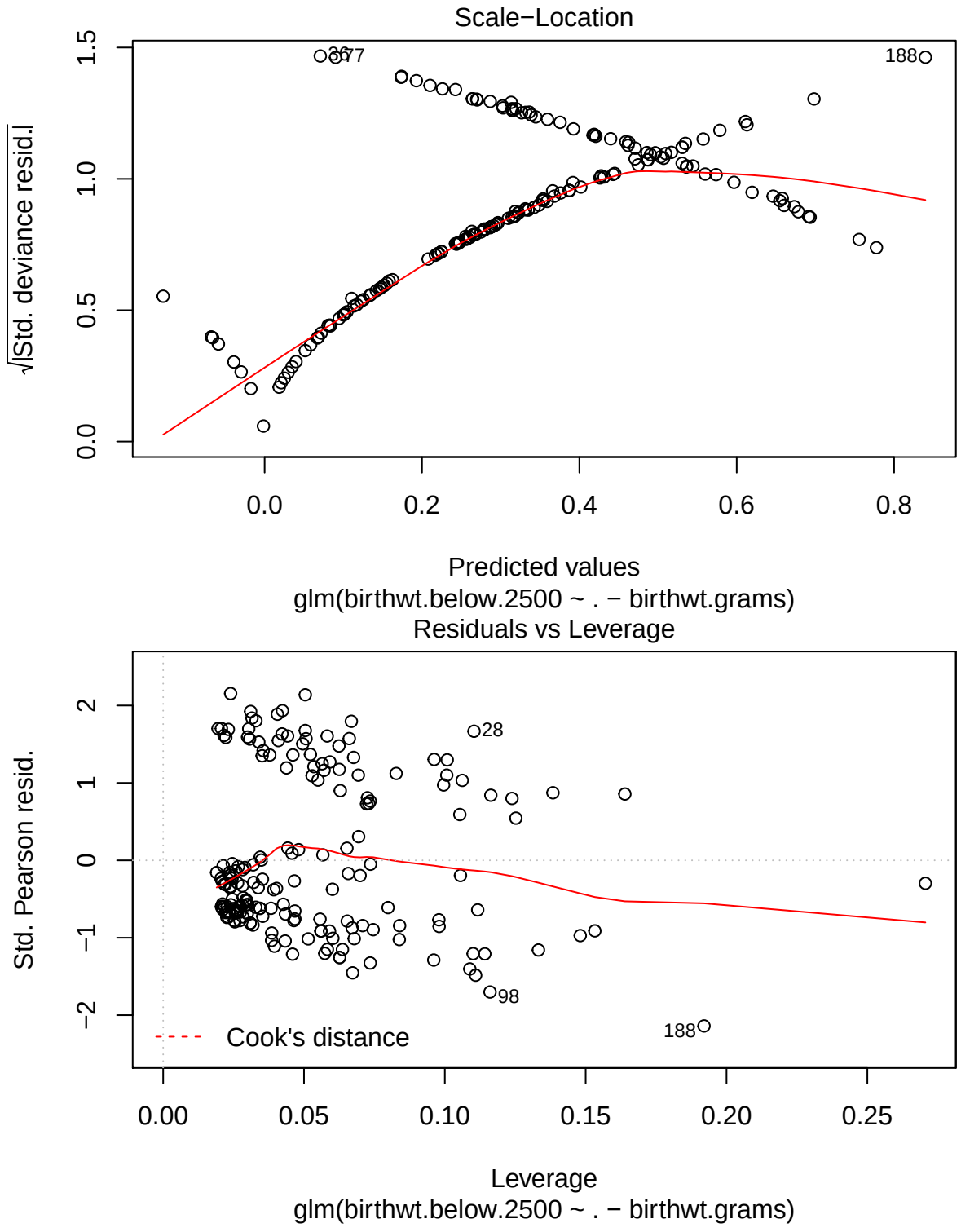
Generalized Linear Models

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=birthwt.noout)
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=birthwt.noout, family=binomial(link=logit))
```

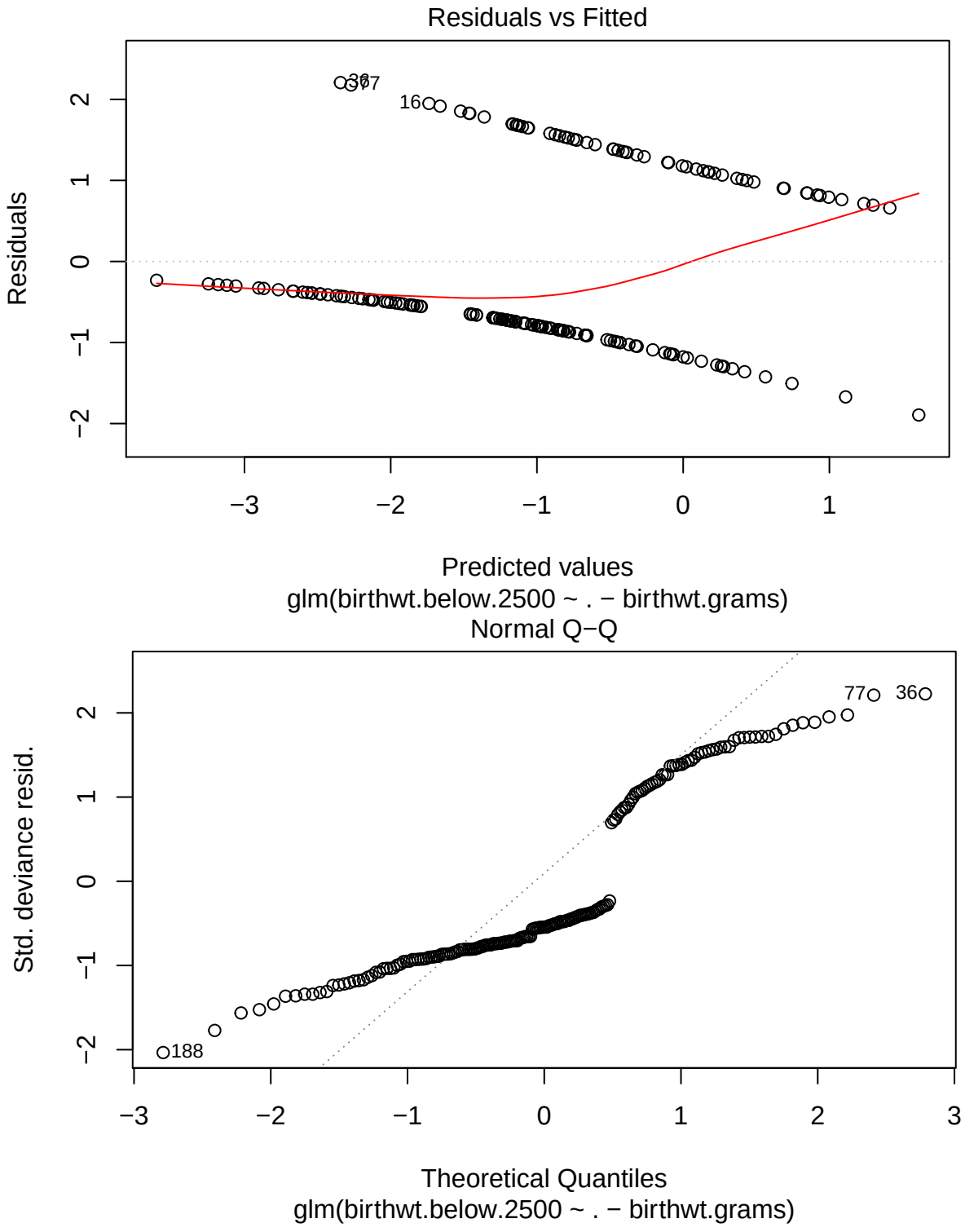
Generalized Linear Models

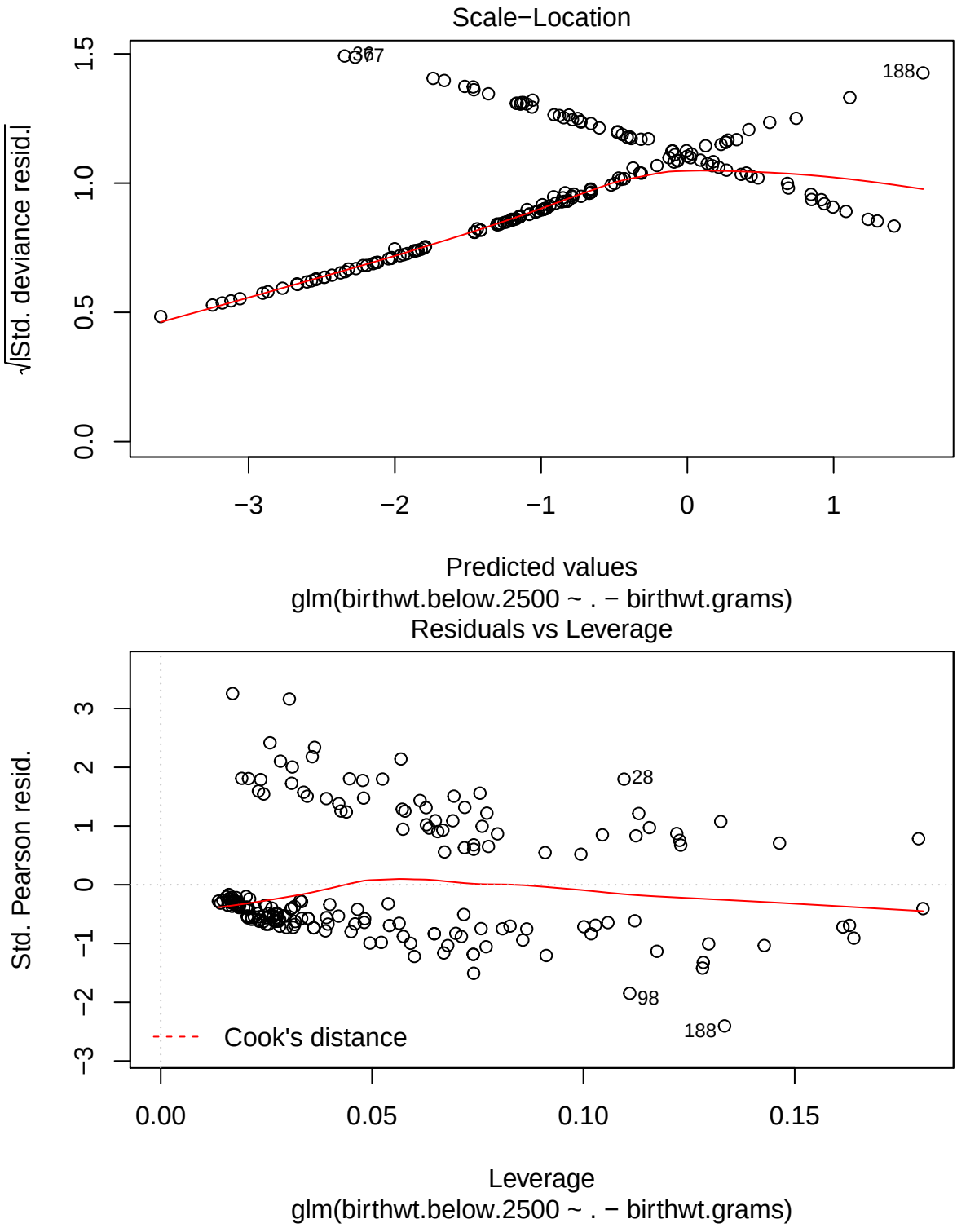
```
summary(glm.1)
```

```
##
## Call:
## glm(formula = birthwt.below.2500 ~ . - birthwt.grams, family = binomial(link = logit),
##     data = birthwt.noout)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.894  -0.822  -0.536   0.985   2.207
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.72183    1.25890    1.37   0.1714
## mother.age          -0.02754    0.03772   -0.73   0.4653
## mother.weight       -0.01547    0.00692   -2.24   0.0253 *
## raceother           -0.39550    0.53769   -0.74   0.4620
## racewhite           -1.26901    0.52718   -2.41   0.0161 *
## mother.smokesYes     0.93173    0.40236    2.32   0.0206 *
## previous.prem.labor  0.53955    0.34541    1.56   0.1183
## hypertensionYes      1.86052    0.69750    2.67   0.0076 **
## uterine.irrYes       0.76652    0.45895    1.67   0.0949 .
## physician.visits     0.06340    0.17243    0.37   0.7131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 233.92  on 187  degrees of freedom
## Residual deviance: 201.15  on 178  degrees of freedom
## AIC: 221.1
##
## Number of Fisher Scoring iterations: 4
```

Generalized Linear Models

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


Residuals vs Fitted
Residuals
Predicted values
glm(birthwt.below.2500 ~ . − birthwt.grams)
Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(birthwt.below.2500 ~ . − birthwt.grams)
188
77
36



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)
```

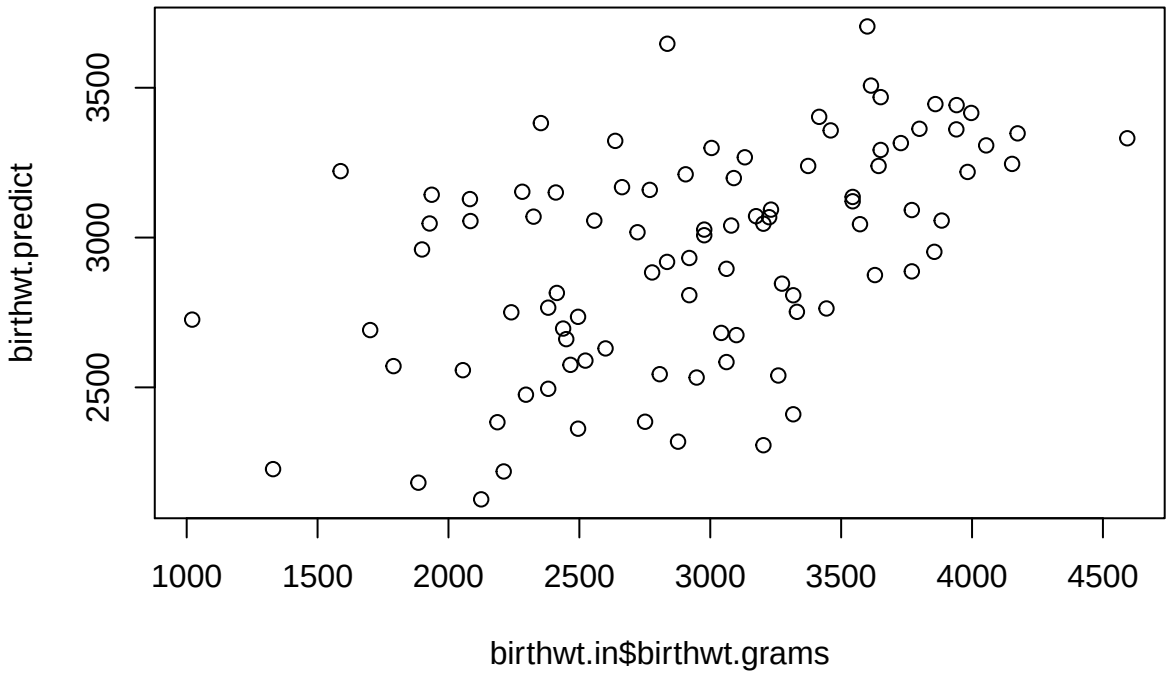
```
##
## Call:
## lm(formula = birthwt.grams ~ . - birthwt.below.2500, data = birthwt.in)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1705.2  -303.1    26.5   427.2  1261.6
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2514.89     450.24    5.59  2.8e-07 ***
## mother.age              7.05      14.93    0.47   0.6380
## mother.weight           2.68       2.89    0.93   0.3550
## raceother             113.95     224.52    0.51   0.6131
## racewhite             466.22     204.97    2.27   0.0255 *
## mother.smokesYes     -217.22     154.52   -1.41   0.1635
## previous.prem.labor  -206.09     143.73   -1.43   0.1553
## hypertensionYes      -653.59     281.79   -2.32   0.0228 *
## uterine.irrYes       -547.88     193.39   -2.83   0.0058 **
## physician.visits     -130.20      81.40   -1.60   0.1135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 644 on 84 degrees of freedom
## Multiple R-squared:  0.259,  Adjusted R-squared:  0.179
## F-statistic: 3.25 on 9 and 84 DF,  p-value: 0.00194
```

What Do We Do With This, Anyway?

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

```
## [1] 0.5084
```

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

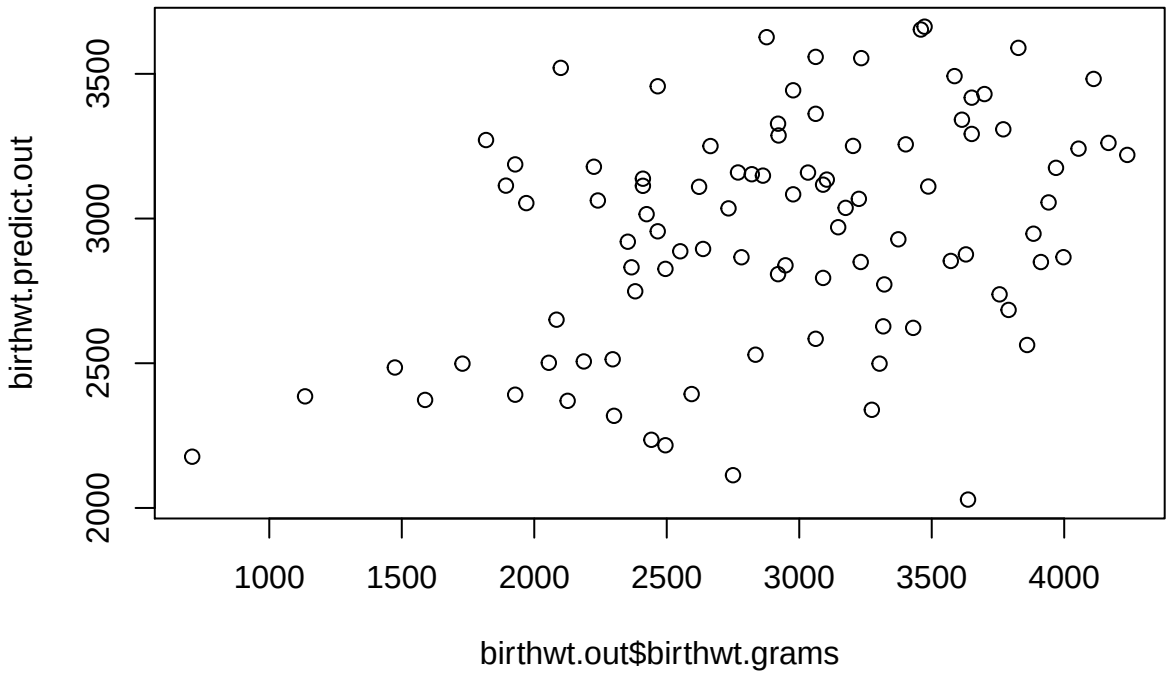


What Do We Do With This, Anyway?

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

```
## [1] 0.3749
```

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


birthwt.predict.out
2000
2500
3000
3500
1000
1500
2000
2500
3000
3500
4000
birthwt.out$birthwt.grams