Getting Data and Linear Models

James M. Flegal

Agenda

- Getting data into and out of R
- Using data frames for statistical purposes
- Introduction to linear models

Reading Data from R

- You can load and save R objects
  - R has its own format for this, which is shared across operating systems
  - It's an open, documented format if you really want to pry into it
- `save(thing, file="name")` saves `thing` in a file called `name` (conventional extension: `rda` or `Rda`)
- `load("name")` loads the object or objects stored in the file called `name`, *with their old names*

```
gmp <- read.table("http://faculty.ucr.edu/~jflegal/206/gmp.dat")
gmp$pop <- round(gmp$gmp/gmp$pcgmp)
save(gmp,file="gmp.Rda")
rm(gmp)
exists("gmp")
```

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

```
not_gmp <- load(file="gmp.Rda")
colnames(gmp)
```

```
## [1] "MSA"   "gmp"   "pcgmp" "pop"
```

```
not_gmp
```

```
## [1] "gmp"
```

- We can load or save more than one object at once; this is how RStudio will load your whole workspace when you're starting, and offer to save it when you're done
- Many packages come with saved data objects; there's the convenience function `data()` to load them

```
 data(cats,package="MASS")
summary(cats)
```

```
##  Sex         Bwt             Hwt
##  F:47   Min.   :2.000   Min.   : 6.30
##  M:97   1st Qu.:2.300   1st Qu.: 8.95
##         Median :2.700   Median :10.10
##         Mean   :2.724   Mean   :10.63
##         3rd Qu.:3.025   3rd Qu.:12.12
##         Max.   :3.900   Max.   :20.50
```

Non-R Data Tables

- Tables full of data, just not in the R file format
- Main function: `read.table()`
  - Presumes space-separated fields, one line per row
  - Main argument is the file name or URL
  - Returns a dataframe
  - Lots of options for things like field separator, column names, forcing or guessing column types, skipping lines at the start of the file...
- `read.csv()` is a short-cut to set the options for reading comma-separated value (CSV) files
  - Spreadsheets will usually read and write CSV

Writing Dataframes

- Counterpart functions `write.table()`, `write.csv()` write a dataframe into a file
- Drawback: takes a lot more disk space than what you get from `load` or `save`
- Advantage: can communicate with other programs, or even edit manually

Less Friendly Data Formats

- The `foreign` package on CRAN has tools for reading data files from lots of non-R statistical software
- Spreadsheets are special
  - Full of ugly irregularities
  - Values or formulas?
  - Headers, footers, side-comments, notes
  - Columns change meaning half-way down

Spreadsheets, If You Have To

- Save the spreadsheet as a CSV; `read.csv()`
- Save the spreadsheet as a CSV; edit in a text editor; `read.csv()`
- Use `read.xls()` from the `gdata` package
  - Tries very hard to work like `read.csv()`, can take a URL or filename
  - Can skip down to the first line that matches some pattern, select different sheets, etc.
  - You may still need to do a lot of tidying up after

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

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.2
```

```
data(birthwt)
summary(birthwt)
```

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

From R help

- Go to R help for more info, because someone documented this data

```
help(birthwt)
```

Make it Readable

```
colnames(birthwt)
```

```
##  [1] "low"   "age"   "lwt"   "race"  "smoke" "ptl"   "ht"    "ui"    "ftv"
## [10] "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

- Can 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])
```

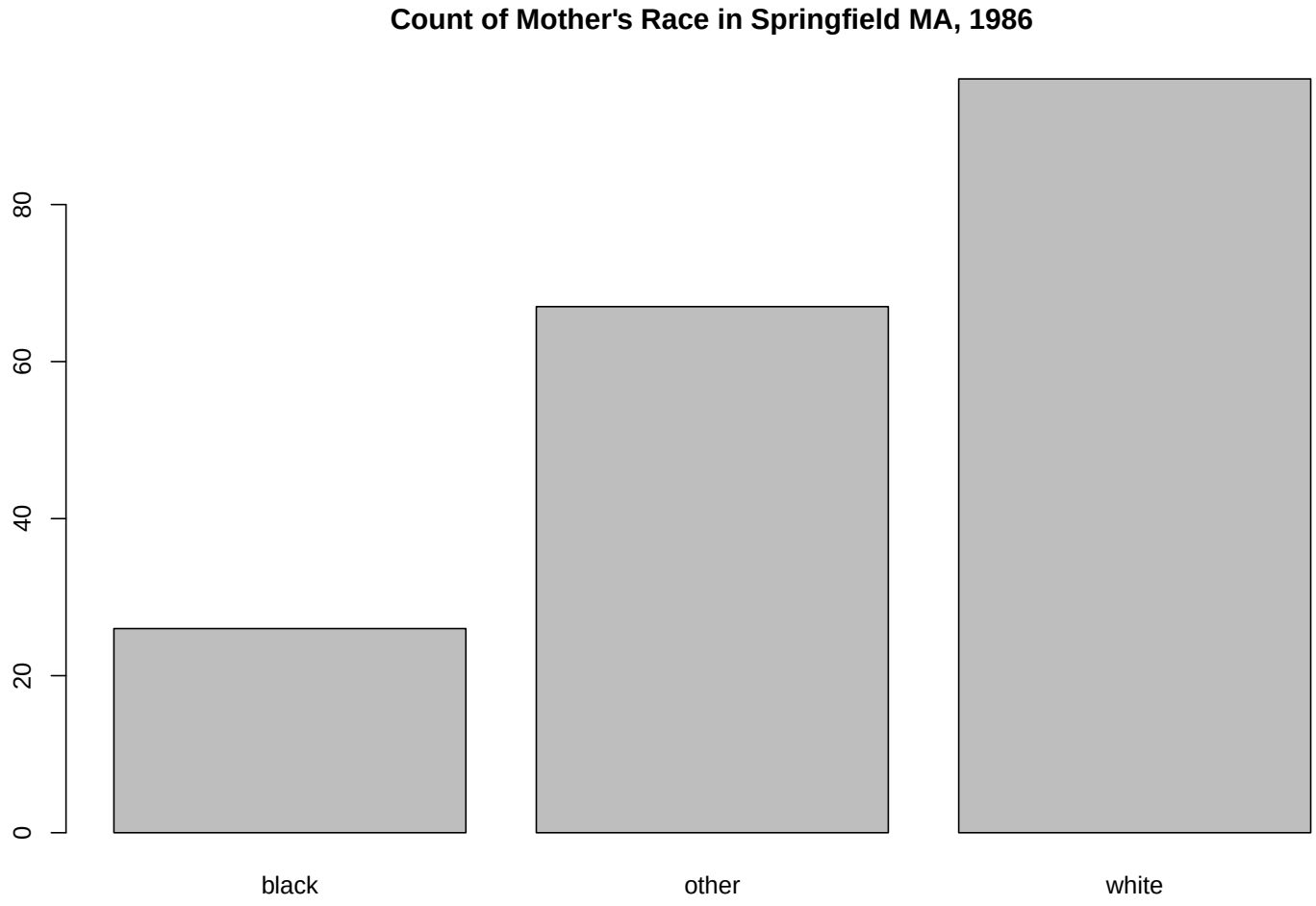
Make it Readable

```
summary(birthwt)
```

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

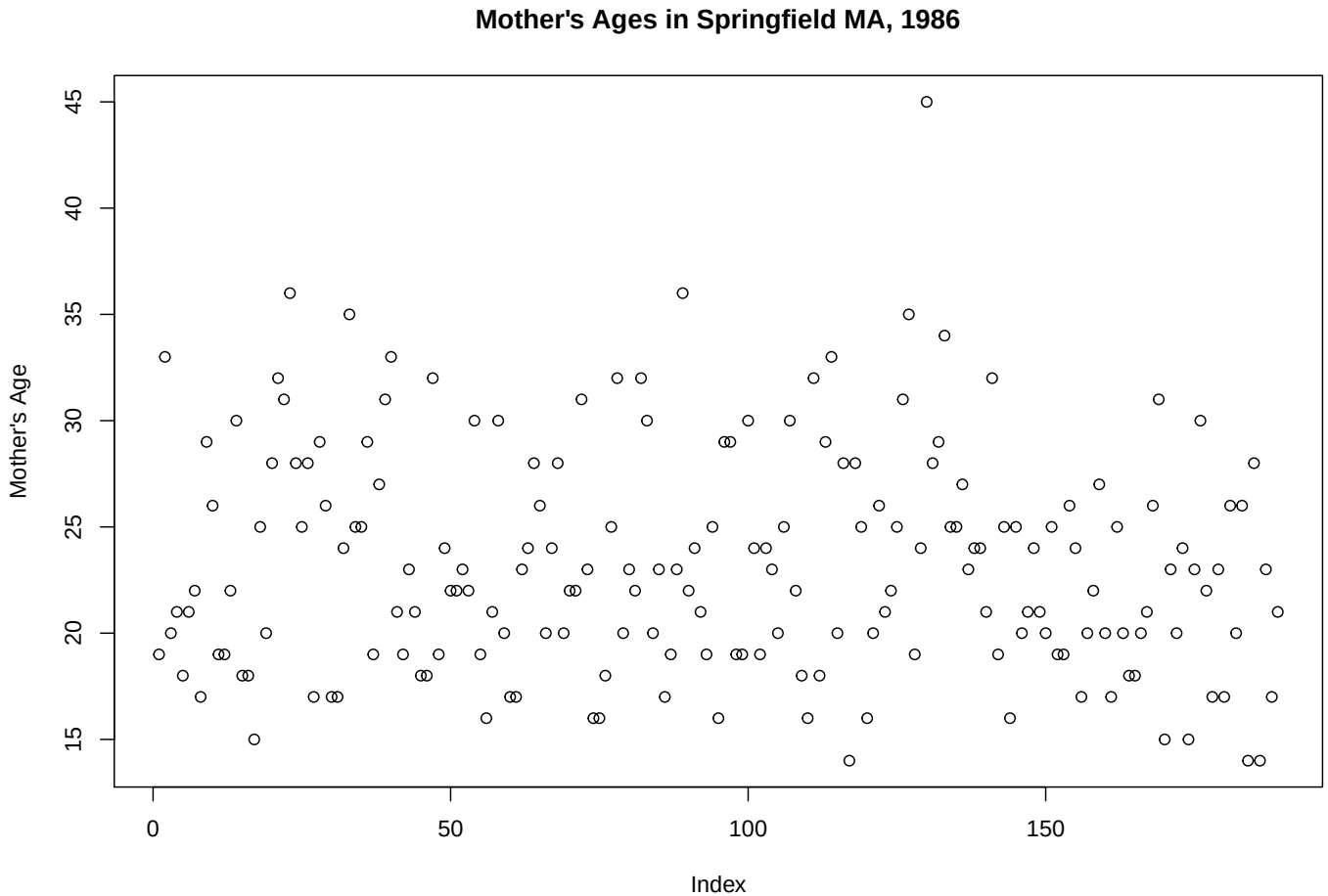
Explore It

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



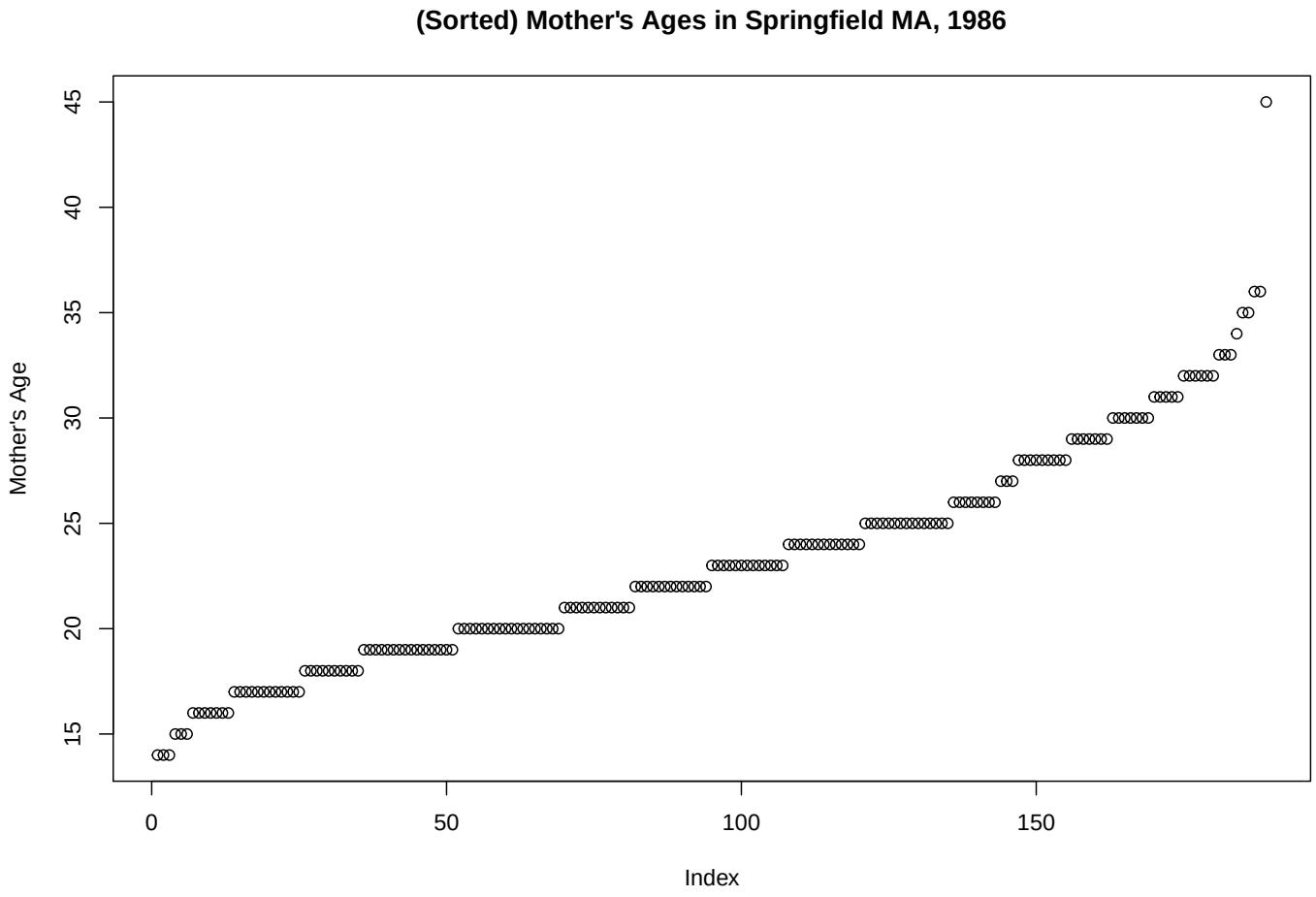
Explore It

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



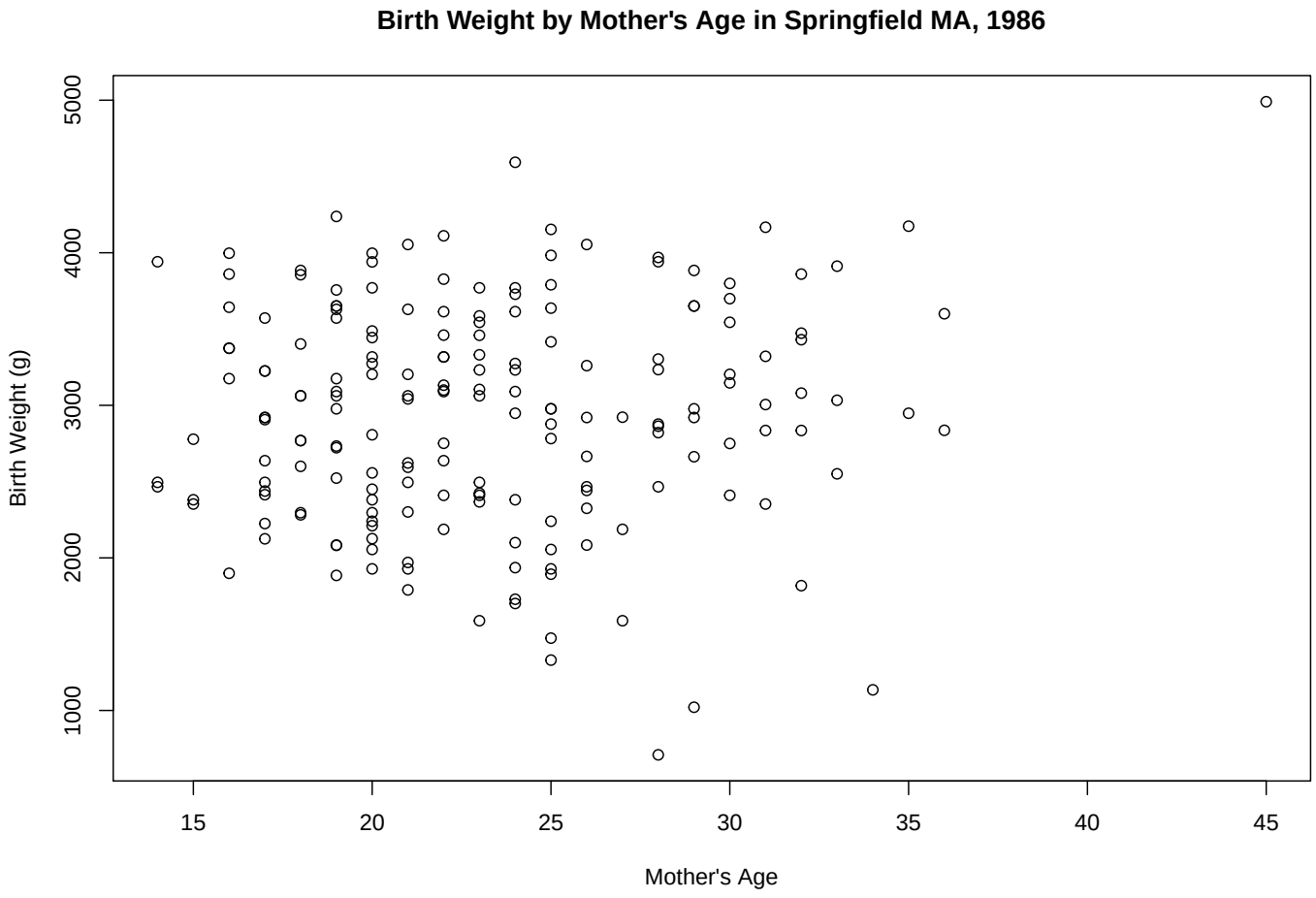
Explore It

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



Explore It

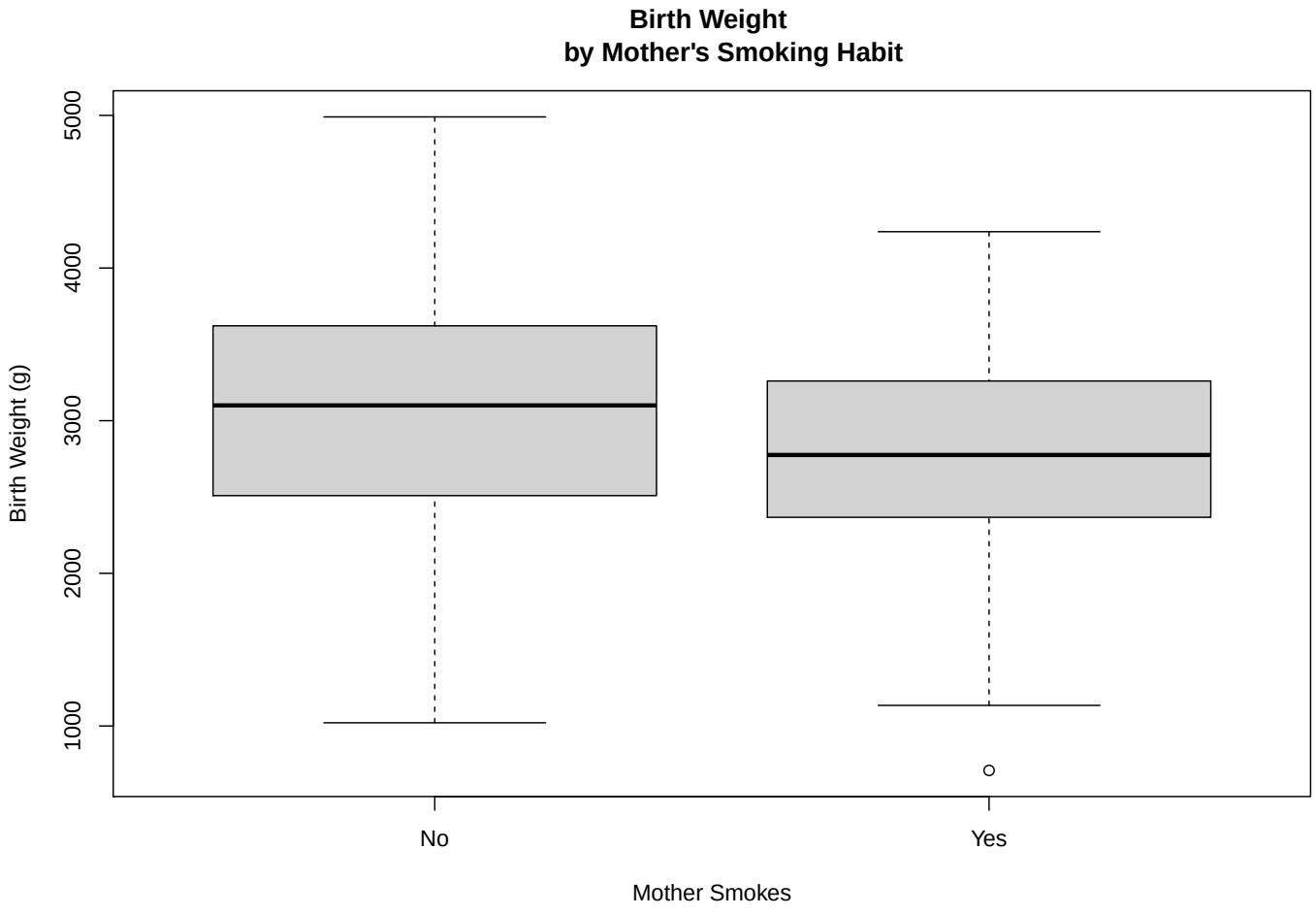
```
plot (birthwt$mother.age, birthwt$birthwt.grams, main =
      "Birth Weight by Mother's Age in Springfield MA, 1986",
      xlab="Mother's Age", ylab="Birth Weight (g)")
```



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 = "Birth Weight (g)", xlab="Mother Smokes")
```



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$birthwt.grams[birthwt$mother.s
## t = -2.7299, df = 170.1, p-value = 0.007003
## alternative hypothesis: true difference in means is not equal to 0
## 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=birthwt)
linear.model.1
```

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.smokes, data = birthwt)
##
## 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 = 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)       3055.70      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 ' ' 1
##
## Residual standard error: 717.8 on 187 degrees of freedom
## Multiple R-squared:  0.03627,    Adjusted R-squared:  0.03112
## F-statistic: 7.038 on 1 and 187 DF,  p-value: 0.008667
```

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
##     2655.74        12.43
```

Basic statistical testing

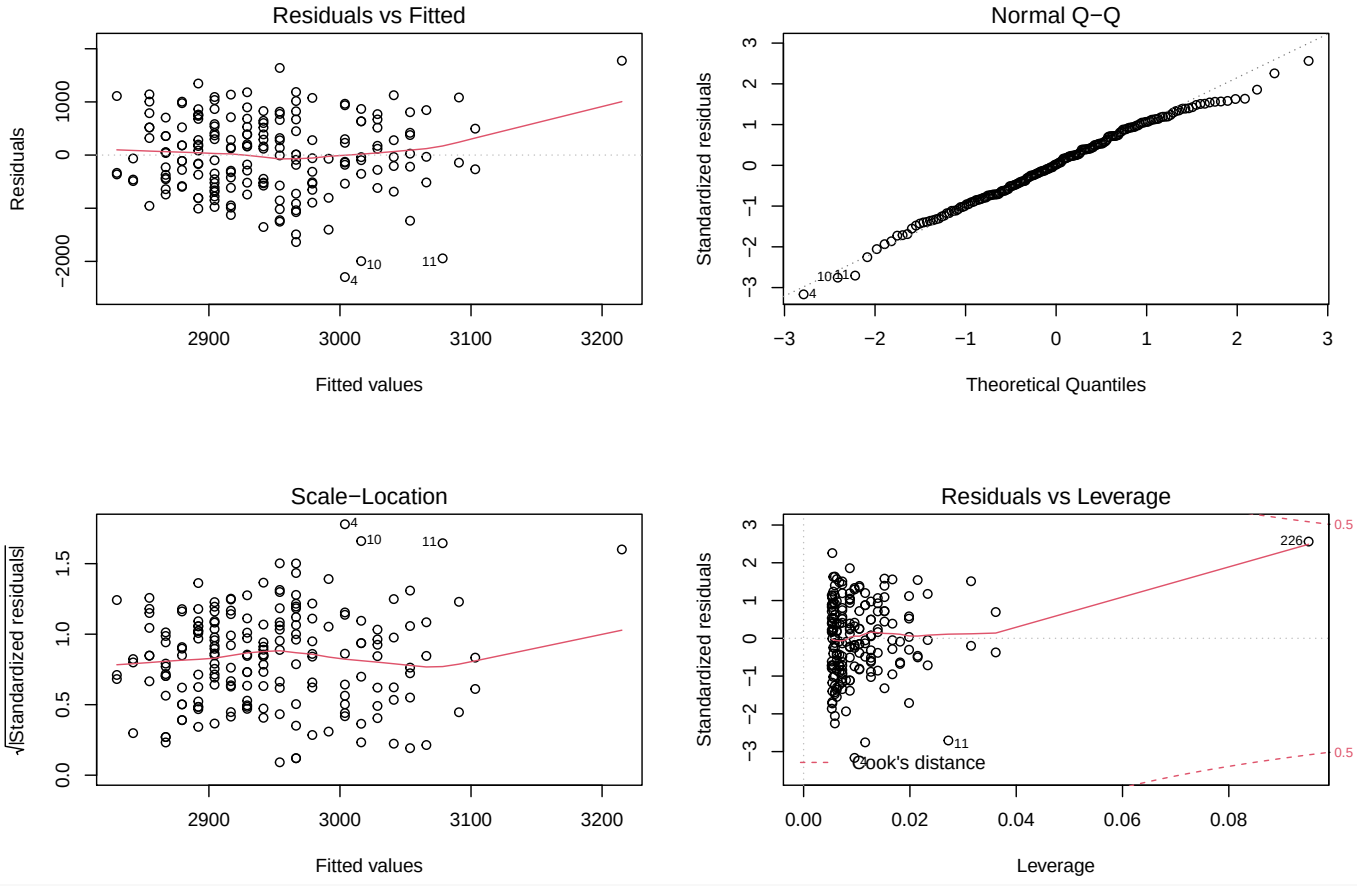
```
summary(linear.model.2)
```

```
##
## 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 ' ' 1
##
## Residual standard error: 728.2 on 187 degrees of freedom
## Multiple R-squared:  0.008157,   Adjusted R-squared:  0.002853
## F-statistic: 1.538 on 1 and 187 DF,  p-value: 0.2165
```

Basic statistical testing

- R tries to make diagnostics easy as possible. Try in R console.

```
par(mfrow = c(2, 2))
plot(linear.model.2)
```




```
par(mfrow = c(1, 1))
```

Detecting Outliers

- Note the oldest mother and her heaviest child are greatly skewing this analysis.

```
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
##    2833.273        4.344
```

Detecting Outliers

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

```
##
## Call:
## lm(formula = birthwt.grams ~ mother.age, data = birthwt.noout)
##
## 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 ' ' 1
##
## Residual standard error: 717.2 on 186 degrees of freedom
## Multiple R-squared:  0.0009461,  Adjusted R-squared:  -0.004425
## F-statistic: 0.1761 on 1 and 186 DF,  p-value: 0.6752
```

More complex models

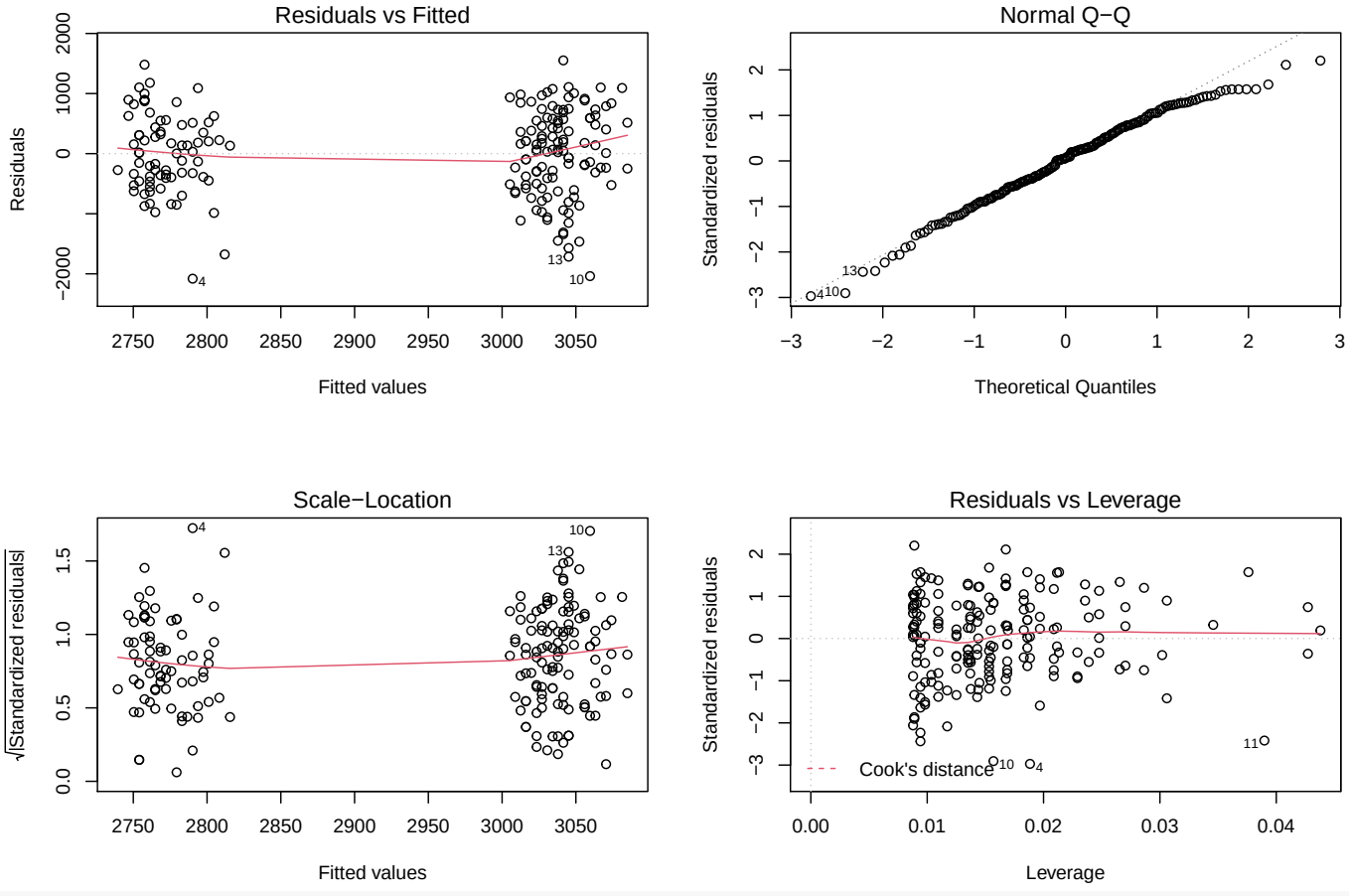
- Add in smoking behavior

```
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
## -2081.22  -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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 707.1 on 185 degrees of freedom
## Multiple R-squared:  0.03401,    Adjusted R-squared:  0.02357
## F-statistic: 3.257 on 2 and 185 DF,  p-value: 0.04072
```

More complex models

```
par(mfrow = c(2, 2))
plot(linear.model.3a)
```




```
par(mfrow = c(1, 1))
```

More complex models

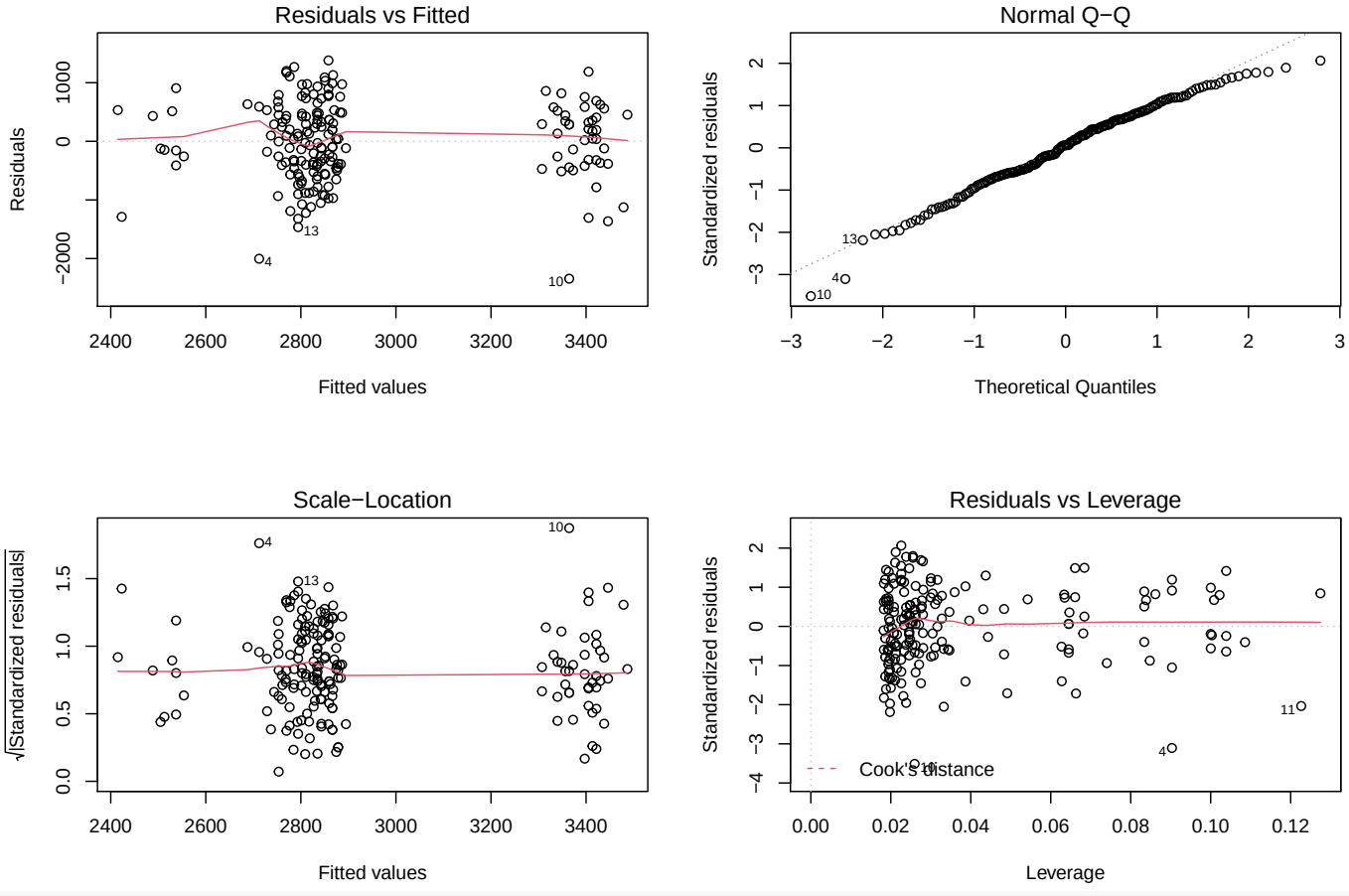
- Add in race

```
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
## -2343.52  -413.66    39.91   480.36  1379.90
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3017.352    265.606  11.360  < 2e-16 ***
## mother.age                     -8.168     10.276  -0.795  0.42772
## mother.smokesYes             -316.500    275.896  -1.147  0.25282
## raceother                     -18.901    193.665  -0.098  0.92236
## racewhite                     584.042    206.320   2.831  0.00517 **
## mother.smokesYes:raceother    258.999    349.871   0.740  0.46010
## mother.smokesYes:racewhite   -271.594    314.268  -0.864  0.38862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 676.1 on 181 degrees of freedom
## Multiple R-squared:  0.1359, Adjusted R-squared:  0.1073
## F-statistic: 4.746 on 6 and 181 DF,  p-value: 0.0001625
```

More complex models

```
par(mfrow = c(2, 2))
plot(linear.model.3b)
```




```
par(mfrow = c(1, 1))
```

Including everything

- Let's include everything 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
##           3360.5163           -1116.3933             -16.0321
##       mother.weight            raceother            racewhite
##              1.9317              68.8145             247.0241
##    mother.smokesYes  previous.prem.labor      hypertensionYes
##           -157.7041              95.9825             -185.2778
##      uterine.irrYes     physician.visits
##           -340.0918              -0.3519
```

Including everything

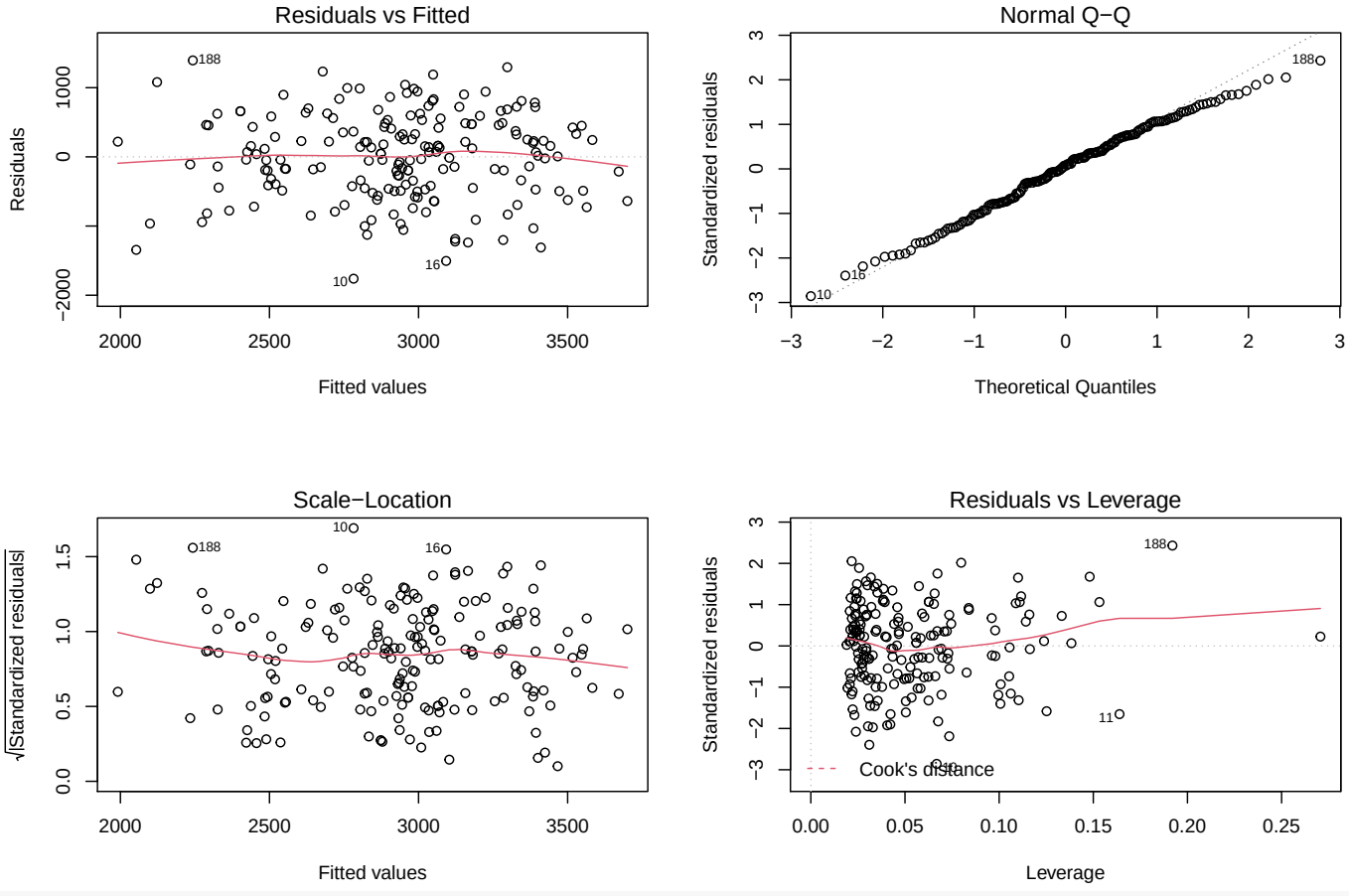
- Be careful! One of those variables `birthwt.below.2500` is a function of the outcome.

```
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
## -1761.10  -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
## mother.weight          4.789      1.710   2.801 0.005656 **
## raceother            155.605    156.564   0.994 0.321634
## racewhite            494.545    147.153   3.361 0.000951 ***
## mother.smokesYes    -335.793    104.613  -3.210 0.001576 **
## previous.prem.labor  -32.922    100.185  -0.329 0.742838
## hypertensionYes     -594.324    198.480  -2.994 0.003142 **
## uterine.irrYes      -514.842    136.249  -3.779 0.000215 ***
## physician.visits      -7.247     45.649  -0.159 0.874036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 638 on 178 degrees of freedom
## Multiple R-squared:  0.2435, Adjusted R-squared:  0.2052
## F-statistic: 6.365 on 9 and 178 DF,  p-value: 8.255e-08
```

Including everything

```
par(mfrow = c(2, 2))
plot(linear.model.4a)
```

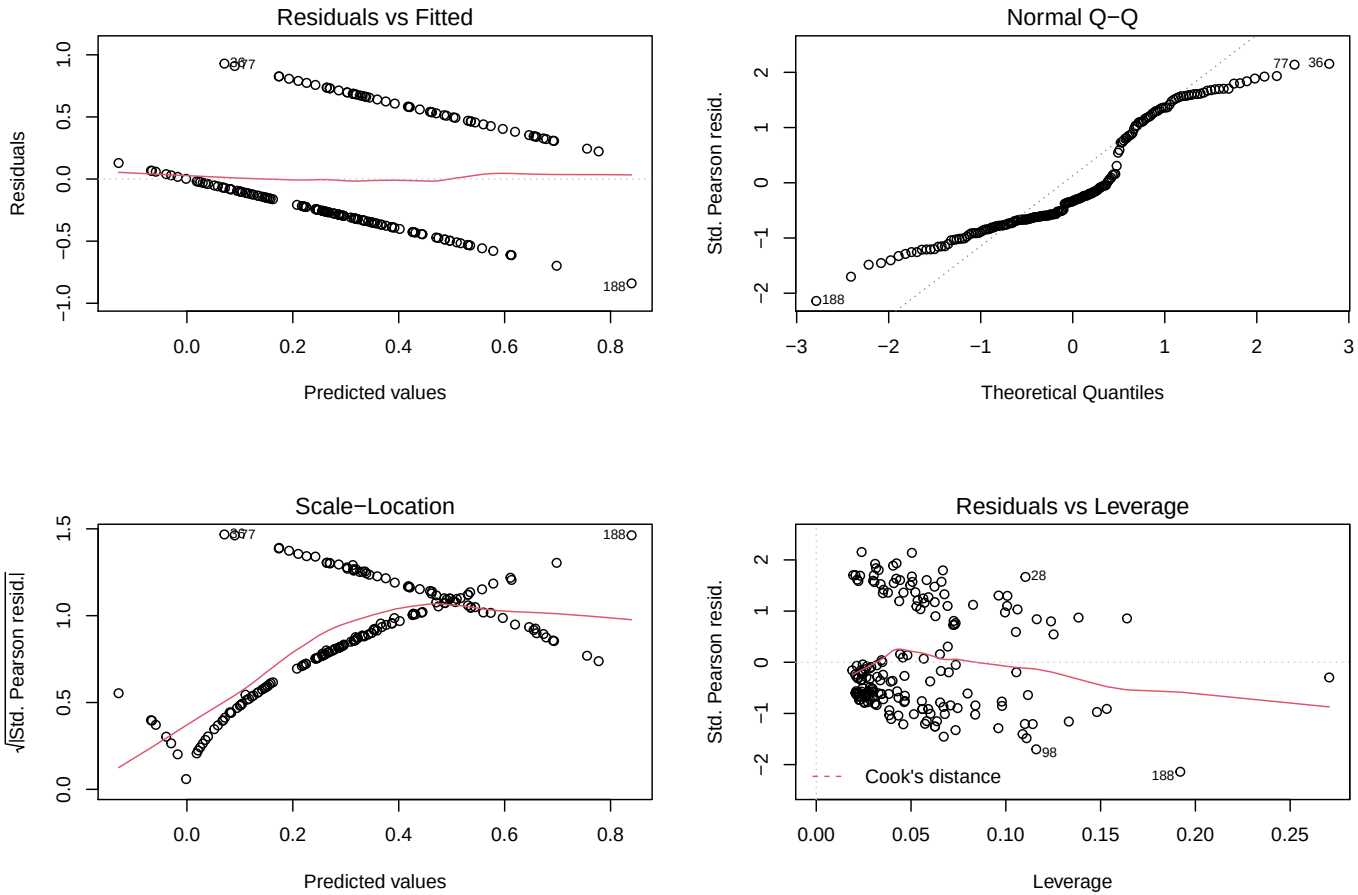



```
par(mfrow = c(1, 1))
```

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:

```
par(mfrow = c(2, 2))
glm.0 <- glm (birthwt.below.2500 ~ . - birthwt.grams, data=birthwt.noout)
plot(glm.0)
```




```
par(mfrow = c(1, 1))
```

Generalized Linear Models

- Default is a Gaussian model (a standard linear model)
- Let's change this!

```
glm.1 <- glm (birthwt.below.2500 ~ . - birthwt.grams, data=birthwt.noout, family=binomial(link=logit))
```

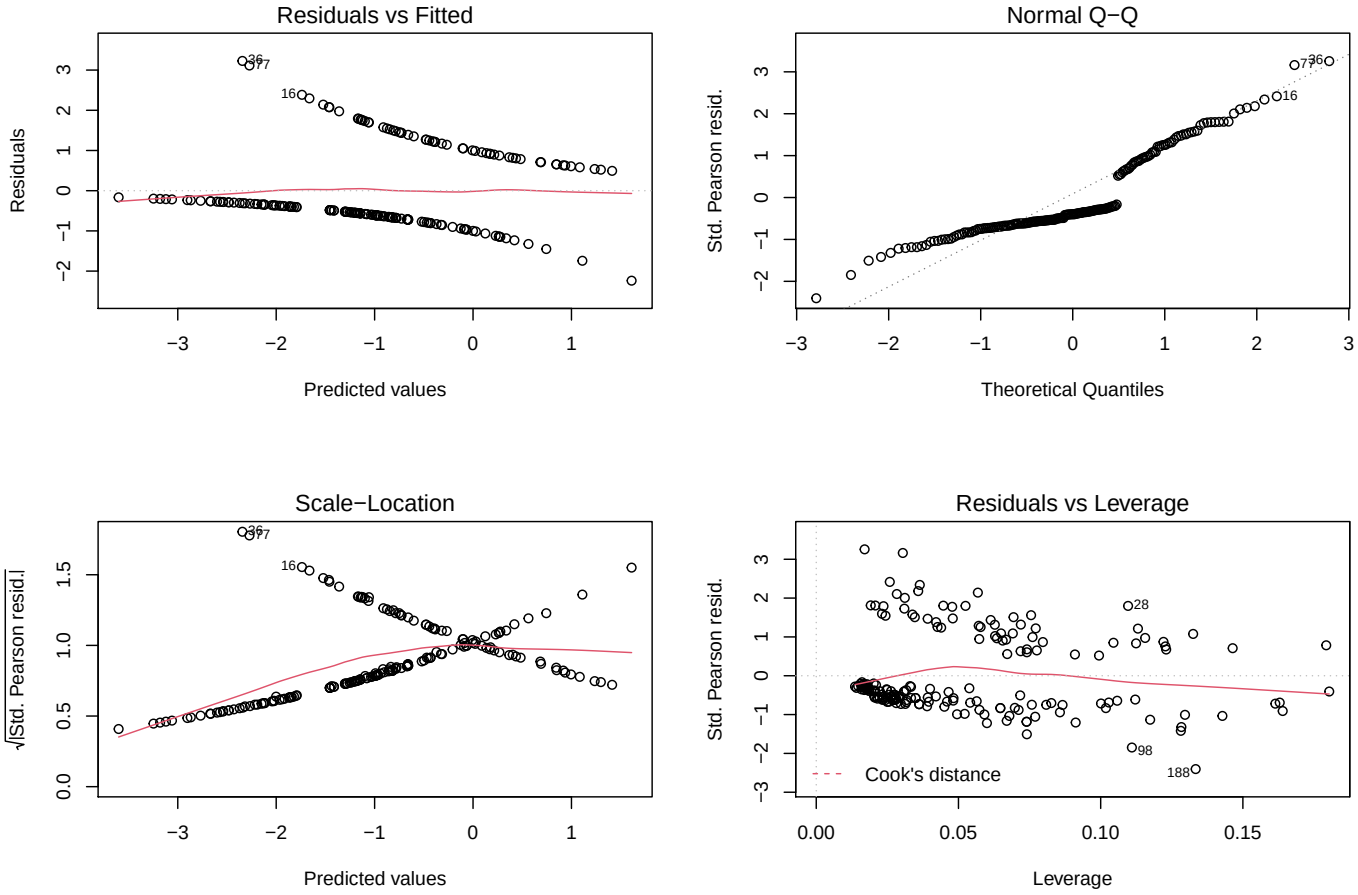
Generalized Linear Models

```
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
## -1.8938  -0.8222  -0.5363   0.9848   2.2069
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.721830   1.258897   1.368  0.17140
## mother.age          -0.027537   0.037718  -0.730  0.46534
## mother.weight       -0.015474   0.006919  -2.237  0.02532 *
## raceother           -0.395505   0.537685  -0.736  0.46199
## racewhite           -1.269006   0.527180  -2.407  0.01608 *
## mother.smokesYes     0.931733   0.402359   2.316  0.02058 *
## previous.prem.labor  0.539549   0.345413   1.562  0.11828
## hypertensionYes      1.860521   0.697502   2.667  0.00764 **
## uterine.irrYes       0.766517   0.458951   1.670  0.09489 .
## physician.visits     0.063402   0.172431   0.368  0.71310
## ---
## 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: 221.15
##
## Number of Fisher Scoring iterations: 4
```

Generalized Linear Models

```
par(mfrow = c(2, 2))
plot(glm.1)
```




```
par(mfrow = c(1, 1))
```

Why?

- 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)
```

Why?

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

```
##
## Call:
## lm(formula = birthwt.grams ~ . - birthwt.below.2500, data = birthwt.in)
##
## 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
## previous.prem.labor -206.093    143.726  -1.434  0.15530
## hypertensionYes     -653.594    281.795  -2.319  0.02280 *
## uterine.irrYes      -547.884    193.386  -2.833  0.00577 **
## physician.visits    -130.202     81.400  -1.600  0.11346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 643.7 on 84 degrees of freedom
## Multiple R-squared:  0.2585, Adjusted R-squared:  0.1791
## F-statistic: 3.254 on 9 and 84 DF,  p-value: 0.001942
```

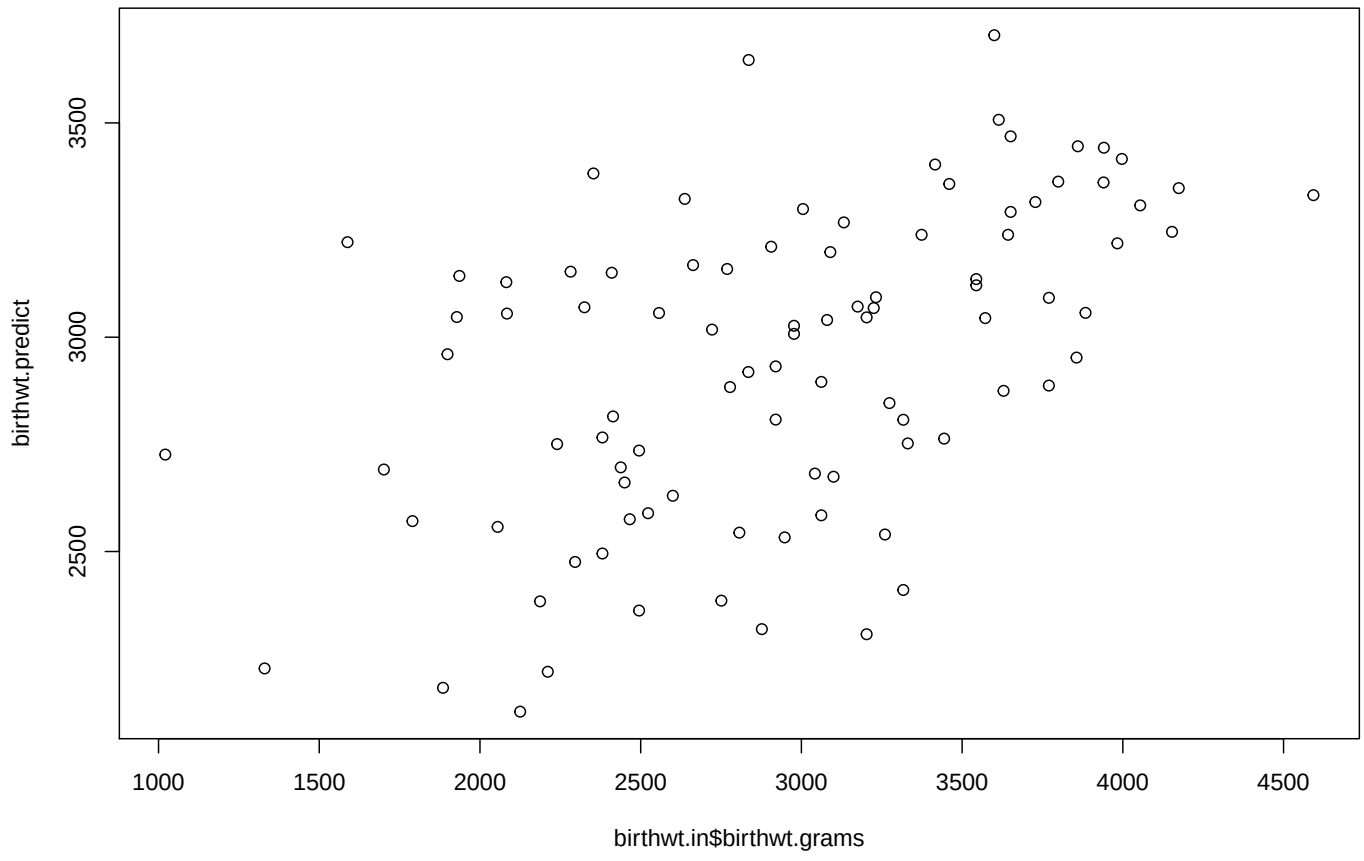
Prediction of Training Data

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

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

Prediction of Training Data

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



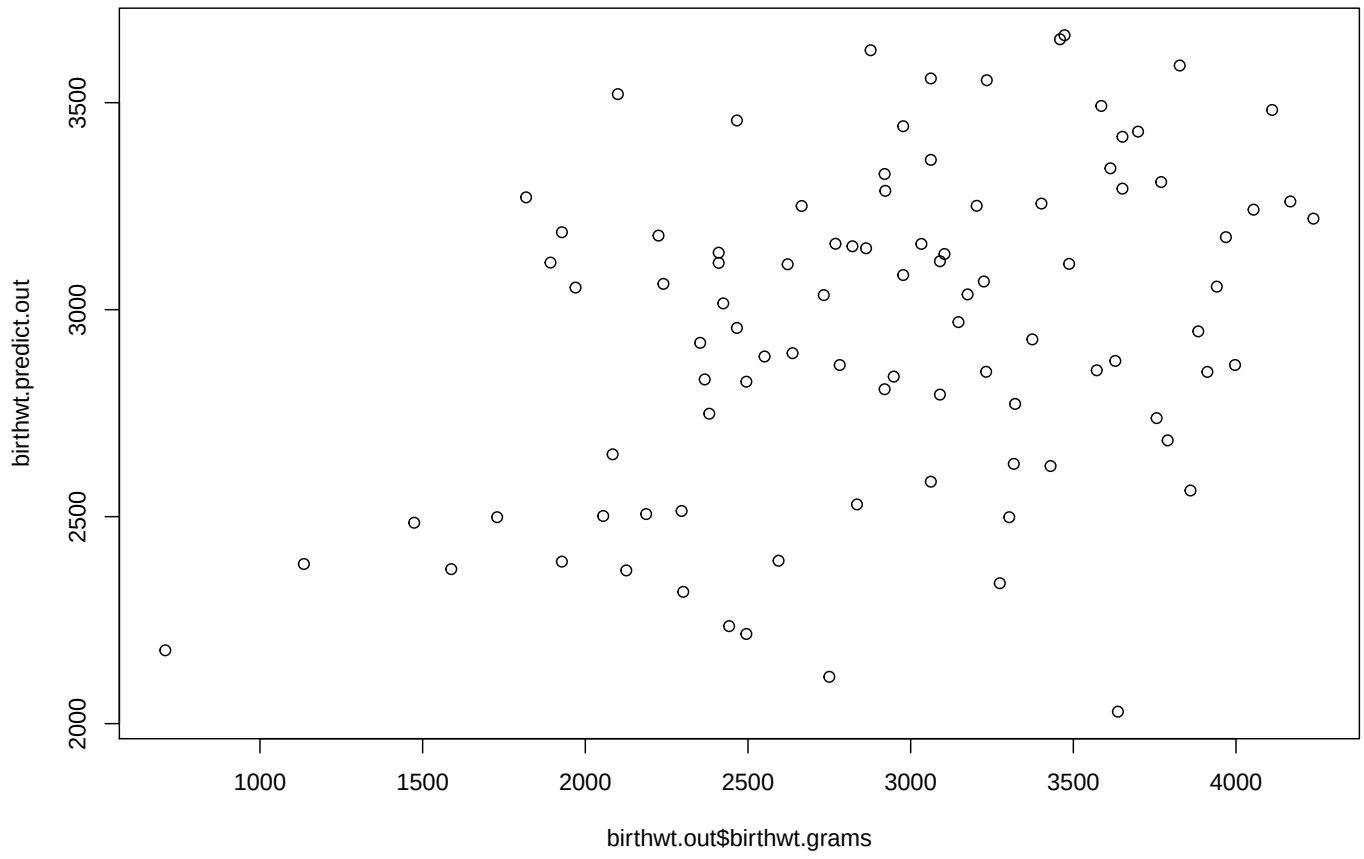
Prediction of Test Data

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

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

Prediction of Test Data

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



Summary

- Loading and saving R objects is very easy
- Reading and writing dataframes is pretty easy
- Linear models are very easy via `lm()`
- Generalized linear models are pretty easy via `glm()`
- Generalized linear mixed models via `lme4()` and `glmm()`