Example of matrix formulation of multiple regression

One of my favorite data sets contains a bunch of information about movies, including how the movie scores on the Bechdel test. A movie passes the Bechdel test if it satisfies 3 rules:

1. it has at least two women;
2. the women talk to each other; and
3. they talk to each other about something or someone other than a man.

The full data set contains the following variables:

- `year` is the year the movie was released
- `title` is the title of the movie
- `bechdel_test` is a version of the results of the Bechdel test with 5 categories, according to users of the website www.bechdeltest.com: "nowomen" means there are not at least two women in the movie; "notalk" means there are at least two women in the movie, but they don't talk to each other; "men" means there are at least two women in the movie, but they only talk to each other about men; "dubious" means there was some disagreement among users of bechdeltest.com about whether or not the movie passed the test; and "ok" means that the movie passes the Bechdel test.
- `bechdel_test_binary` is a the results of the Bechdel test with 2 categories, according to the users of the website www.bechdeltest.com: "PASS" means that the movie passed the test (i.e., its value for `bechdel_test` is "ok"); "FAIL" means it did not pass the test (i.e., its value for `bechdel_test` is something other than "ok")
- `budget` is the movie's approximate production budget, in the dollars of the year the movie was made.
- `domgross` is the movie's domestic gross earnings (i.e., total earnings from the U.S.), in the dollars of the year the movie was made.
- `intgross` is the movie's combined domestic and international gross earnings (i.e., total earnings both in the U.S. and internationally), in the dollars of the year the movie was made.
- `budget_2013` is the same as `budget` but in inflation-adjusted 2013 dollars.
- `domgross_2013` is the same as `domgross`, but in inflation-adjusted 2013 dollars.
- `intgross_2013` is the same as `intgross`, but in inflation-adjusted 2013 dollars.
- `imdb_rating` is the average rating for the movie by users of the website www.imdb.com, on a scale of 0 to 10 (higher ratings are better)
- `num_imdb_ratings` is the number of distinct users of www.imdb.com who have rated the movie.
- `mpaa_rating` is the MPAA rating for the movie, like PG or R.
- `run_time_min` is the length of the movie in minutes.

Suppose we want to model a movie's international gross earnings in inflation-adjusted 2013 dollars (`intgross_2013`) based on the following 5 explanatory variables: `budget_2013`, `run_time_min`, `imdb_rating`, `mpaa_rating`, `bechdel_test_binary`

To start with, let's just see what the design matrix for a first model looks like:

```
movies_fit <- lm(intgross_2013 ~ budget_2013 + run_time_min + imdb_rating + mpaa_rating + bechdel_test_binary, data = movies)

movies %>% select(
  budget_2013, run_time_min, imdb_rating, mpaa_rating, bechdel_test_binary
) %>%
  head()
```

```
## # A tibble: 6 x 5
##   budget_2013 run_time_min imdb_rating mpaa_rating bechdel_test_binary
##         <dbl>        <dbl>       <dbl> <fct>       <fct>
## 1    13000000           93         5.9 R           FAIL
## 2    45658735           95         7.1 R           PASS
## 3    20000000          134         8.1 R           FAIL
## 4    61000000          109         6.7 R           FAIL
## 5    40000000          128         7.5 PG-13       FAIL
## 6   225000000          128         6.3 PG-13       FAIL
```

```
head(model.matrix(movies_fit))
```

```
##   (Intercept) budget_2013 run_time_min imdb_rating mpaa_ratingPG mpaa_ratingPG-13 mpaa_ratingR bechdel_test_binaryPASS
## 1           1    13000000           93         5.9             0                0           1                        0
## 2           1    45658735           95         7.1             0                0           1                        1
## 3           1    20000000          134         8.1             0                0           1                        0
## 4           1    61000000          109         6.7             0                0           1                        0
## 5           1    40000000          128         7.5             0                1           0                        0
## 6           1   225000000          128         6.3             0                1           0                        0
```

In this design matrix the fifth column has the values

$$x_{i4} = \begin{cases} 1 \text{ if movie } i \text{ is rated PG} \\ 0 \text{ otherwise} \end{cases}$$