Multiple Regression with both Categorical and Quantitative Explanatory Variables

Recall that we are thinking about a data set with several variables recorded about 1753 movies. We are exploring building multiple regression models for a movie's international gross earnings in inflation-adjusted 2013 dollars (`intgross_2013`) based on the following 5 explanatory variables:

1. `budget_2013` (quantitative)
2. `run_time_min` (quantitative)
3. `imdb_rating` (quantitative)
4. `mpaa_rating` (categorical)
5. `bechdel_test_binary` (categorical)

First, let's load the data in, filter to include only MPAA ratings categories with a reasonable number of movies in them, set categorical variables to factors, and apply a log transformation to `intgross_2013`, `budget_2013`, and `run_time_min`.

```
library(readr)
library(dplyr)
library(ggplot2) # general plotting functionality
library(GGally) # includes the ggpairs function, pairs plots via ggplot2
library(gridExtra) # for grid.arrange, which arranges the plots next to each other

options(na.action = na.exclude, digits = 7)

movies <- read_csv("http://www.evanlray.com/data/bechdel/bechdel.csv") %>%
  filter(mpaa_rating %in% c("G", "PG", "PG-13", "R")) %>%
  mutate(
    bechdel_test = factor(bechdel_test, levels = c("nowomen", "notalk", "men", "dubious", "ok")),
    bechdel_test_binary = factor(bechdel_test_binary, levels = c("FAIL", "PASS")),
    mpaa_rating = factor(mpaa_rating, levels = c("G", "PG", "PG-13", "R"))
  ) %>%
  mutate(
    log_intgross_2013 = log(intgross_2013),
    log_budget_2013 = log(budget_2013),
    log_run_time_min = log(run_time_min)
  )
```

Our goals:

- Understand R's parameterization of linear models involving categorical variables (interpretation of fixed and interaction effects)
- See some examples of testing effects

1 categorical explanatory variable (one-way ANOVA)

Here are plots showing the relationship between the categorical variables and the response:

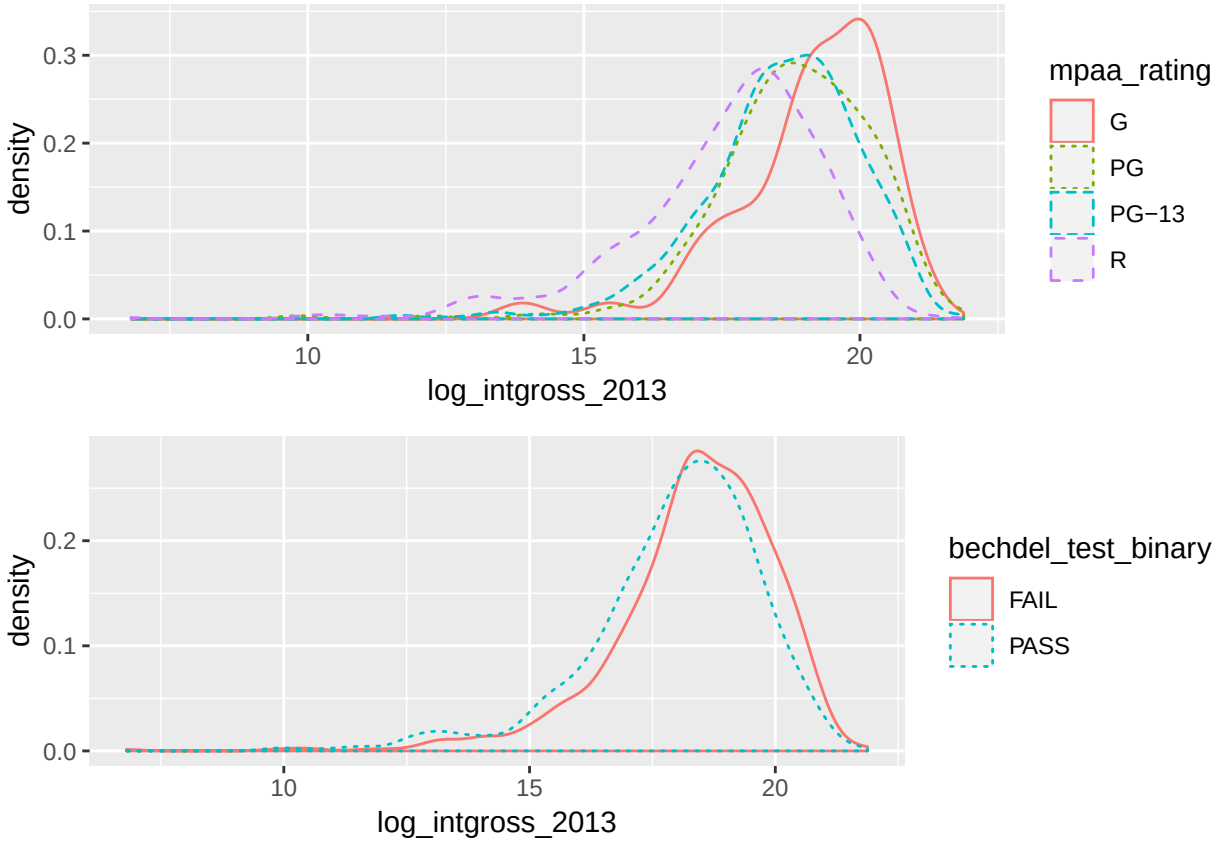
```
p_mpaa <- ggplot(data = movies, mapping = aes(x = log_intgross_2013, color = mpaa_rating, linetype = mpaa_rating)) +
  geom_density()

p_bechdel <- ggplot(data = movies,
  mapping = aes(x = log_intgross_2013, color = bechdel_test_binary, linetype = bechdel_test_binary)) +
  geom_density()

grid.arrange(p_mpaa, p_bechdel, ncol = 1)
```

```
## Warning: Removed 7 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 7 rows containing non-finite values (stat_density).
```



Explanatory variable has 2 categories

```
movies <- movies %>% filter(!is.na(bechdel_test_binary))
fit_bechdel <- lm(log_intgross_2013 ~ bechdel_test_binary, data = movies)
summary(fit_bechdel)
```

```
##
## Call:
## lm(formula = log_intgross_2013 ~ bechdel_test_binary, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5695  -0.8094   0.2046   1.1536   3.9065
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              18.37075    0.05418 339.091  < 2e-16 ***
## bechdel_test_binaryPASS -0.39968    0.08095  -4.937 8.68e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.682 on 1744 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.01378,    Adjusted R-squared:  0.01322
## F-statistic: 24.38 on 1 and 1744 DF,  p-value: 8.682e-07
```

What is the estimated equation for this fit? Define all variables involved.

How do the parameter estimates relate to the following R output?

```
group_means <- movies %>%
  group_by(bechdel_test_binary) %>%
  summarize(
    mean_log_earnings = mean(log_intgross_2013, na.rm = TRUE)
  ) %>%
  as.data.frame()
group_means
```

```
##   bechdel_test_binary mean_log_earnings
## 1                FAIL          18.37075
## 2                PASS          17.97107
```

```
levels(movies$bechdel_test_binary)
```

```
## [1] "FAIL" "PASS"
```

Explanatory variable has >2 categories

```
fit_mpaa <- lm(log_intgross_2013 ~ mpaa_rating, data = movies)
summary(fit_mpaa)
```

```
##
## Call:
## lm(formula = log_intgross_2013 ~ mpaa_rating, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7940  -0.7901   0.2335   1.0671   3.8756
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       19.1061     0.2266  84.319  < 2e-16 ***
## mpaa_ratingPG     -0.2831     0.2459  -1.151    0.250
## mpaa_ratingPG-13  -0.5210     0.2355  -2.213    0.027 *
## mpaa_ratingR      -1.5108     0.2337  -6.466 1.31e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.602 on 1742 degrees of freedom
##   (7 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.1062, Adjusted R-squared:  0.1047
## F-statistic:    69 on 3 and 1742 DF,  p-value: < 2.2e-16
```

What is the estimated equation for this fit? Define all variables involved.

How do the parameter estimates relate to the following R output?

```
group_means <- movies %>%
  group_by(mpaa_rating) %>%
  summarize(
    mean_log_earnings = mean(log_intgross_2013, na.rm = TRUE)
  ) %>%
  as.data.frame()
group_means
```

```
##   mpaa_rating mean_log_earnings
## 1           G          19.10614
## 2          PG          18.82303
## 3       PG-13          18.58513
## 4           R          17.59530
```

```
levels(movies$mpaa_rating)
```

```
## [1] "G"     "PG"    "PG-13" "R"
```

A Cautionary Tale - Regression with Ordered Factors

```
movies <- movies %>%
  mutate(
    bechdel_test_binary_ordered_factor = factor(bechdel_test_binary, levels = c("FAIL", "PASS"), ordered = TRUE)
  )

fit_bechdel_ordered_factor <- lm(log_intgross_2013 ~ bechdel_test_binary_ordered_factor, data = movies)
summary(fit_bechdel_ordered_factor)
```

```
##
## Call:
## lm(formula = log_intgross_2013 ~ bechdel_test_binary_ordered_factor,
##     data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5695  -0.8094   0.2046   1.1536   3.9065
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        18.17091    0.04048 448.929  < 2e-16
## bechdel_test_binary_ordered_factor.L -0.28262    0.05724  -4.937 8.68e-07
##
## (Intercept)                        ***
## bechdel_test_binary_ordered_factor.L ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.682 on 1744 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.01378,    Adjusted R-squared:  0.01322
## F-statistic: 24.38 on 1 and 1744 DF,  p-value: 8.682e-07
```

```
movies %>%
  group_by(bechdel_test_binary) %>%
  summarize(
    mean_log_earnings = mean(log_intgross_2013, na.rm = TRUE)
  ) %>%
  as.data.frame()
```

```
##   bechdel_test_binary mean_log_earnings
## 1                FAIL          18.37075
## 2                PASS          17.97107
```

Both Categorical and Quantitative Variables; no interactions

First, we subset to movies where all of our candidate explanatory variables are non-missing. This is necessary to ensure that all fits below are based on the same observations, which is needed for comparing models with `anova`.

```
movies <- movies %>%
  filter(!is.na(log_intgross_2013) & !is.na(mpaa_rating) & !is.na(bechdel_test_binary) &
           !is.na(log_budget_2013) & !is.na(log_run_time_min) & !is.na(imdb_rating))
```

Let's try a backwards selection type strategy: we'll use all the explanatory variables we're considering, then drop variables that don't seem to be contributing much to the fit.

```
fit_all_x <- lm(log_intgross_2013 ~ mpaa_rating + bechdel_test_binary + log_budget_2013 + log_run_time_min + imdb_rating,
  data = movies)
summary(fit_all_x)
```

```
##
## Call:
## lm(formula = log_intgross_2013 ~ mpaa_rating + bechdel_test_binary +
##     log_budget_2013 + log_run_time_min + imdb_rating, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8442 -0.5614  0.1202  0.6923  4.7576
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.62261    0.82268   3.188  0.00146 **
## mpaa_ratingPG            -0.07776    0.19030  -0.409  0.68286
## mpaa_ratingPG-13         -0.19903    0.18622  -1.069  0.28532
## mpaa_ratingR             -0.60253    0.18683  -3.225  0.00128 **
## bechdel_test_binaryPASS  -0.02416    0.06047  -0.400  0.68953
## log_budget_2013           0.80312    0.02686  29.904  < 2e-16 ***
## log_run_time_min         -0.10942    0.20959  -0.522  0.60170
## imdb_rating               0.38097    0.03589  10.614  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.221 on 1711 degrees of freedom
## Multiple R-squared:  0.4847, Adjusted R-squared:  0.4826
## F-statistic: 229.9 on 7 and 1711 DF,  p-value: < 2.2e-16
```

This initial fit suggests that after accounting for the associations between all other explanatory variables and log earnings, a movie's run time and whether or not it passes the Bechdel test do not account for a statistically significant amount of variation in earnings.

Let's conduct an F test to see whether we might drop both of these variables from the model. We fit a reduced model and compare with `anova`:

```r
fit_mpaa_budget_imdb <- lm(log_intgross_2013 ~ mpaa_rating + log_budget_2013 + imdb_rating,
  data = movies)
anova(fit_all_x, fit_mpaa_budget_imdb)
```

```
## Analysis of Variance Table
##
## Model 1: log_intgross_2013 ~ mpaa_rating + bechdel_test_binary + log_budget_2013 +
##     log_run_time_min + imdb_rating
## Model 2: log_intgross_2013 ~ mpaa_rating + log_budget_2013 + imdb_rating
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1   1711 2548.8
## 2   1713 2549.5 -2  -0.67022 0.225 0.7986
```

What is the result of the test?

In the reduced model fit above, what's the interpretation of the estimated coefficient for `mpaa_ratingR`?

Interactions between quantitative and categorical variables

Here's a model fit that includes an interaction between `log_budget_2013` and `mpaa_rating`, as well as a call to `anova` that compares this model with the previous model that did not include interactions.

```
fit_interaction <- lm(log_intgross_2013 ~ mpaa_rating + log_budget_2013 + imdb_rating + mpaa_rating:log_budget_2013,
  data = movies)
summary(fit_interaction)
```

```
##
## Call:
## lm(formula = log_intgross_2013 ~ mpaa_rating + log_budget_2013 +
##     imdb_rating + mpaa_rating:log_budget_2013, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8564 -0.5538  0.1145  0.6920  4.6334
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.48574    3.62192   0.962 0.335985
## mpaa_ratingPG                   -0.77310    3.85700  -0.200 0.841161
## mpaa_ratingPG-13                -3.05619    3.70830  -0.824 0.409969
## mpaa_ratingR                    -1.19013    3.66856  -0.324 0.745665
## log_budget_2013                  0.72899    0.20268   3.597 0.000331 ***
## imdb_rating                      0.37479    0.03232  11.596  < 2e-16 ***
## mpaa_ratingPG:log_budget_2013    0.03778    0.21556   0.175 0.860877
## mpaa_ratingPG-13:log_budget_2013 0.15941    0.20719   0.769 0.441763
## mpaa_ratingR:log_budget_2013     0.02891    0.20511   0.141 0.887913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.219 on 1710 degrees of freedom
## Multiple R-squared:  0.4865, Adjusted R-squared:  0.4841
## F-statistic: 202.5 on 8 and 1710 DF,  p-value: < 2.2e-16
```

```
anova(fit_interaction, fit_mpaa_budget_imdb)
```

```
## Analysis of Variance Table
##
## Model 1: log_intgross_2013 ~ mpaa_rating + log_budget_2013 + imdb_rating +
##     mpaa_rating:log_budget_2013
## Model 2: log_intgross_2013 ~ mpaa_rating + log_budget_2013 + imdb_rating
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   1710 2540.1
```

```
## 2   1713 2549.5 -3   -9.3923 2.1077 0.09733 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What's the estimated equation for this model?

In the model fit including interactions, what is the interpretation of the estimated coefficient for `mpaa_ratingPG-13:log_budget_2013`?

According to the hypothesis test, should we include the interaction in the model?