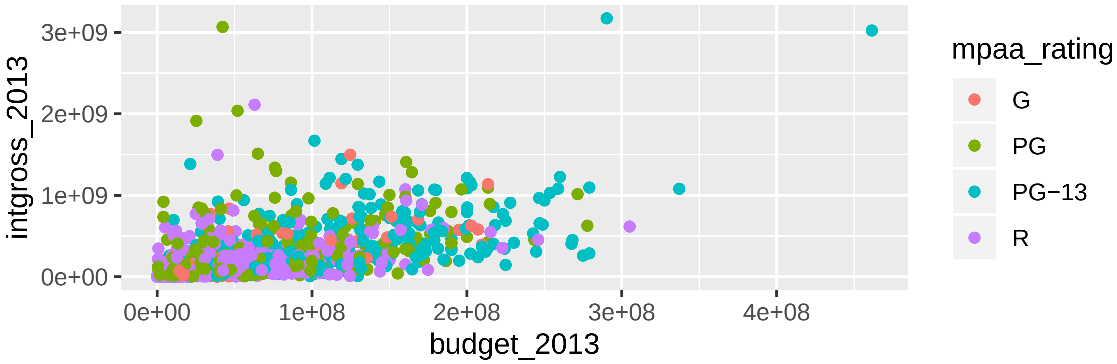
Regression with More Than 2 Levels in a Categorical Variable

Oct 30 2019 – Sleuth3 Chapters 9, 10

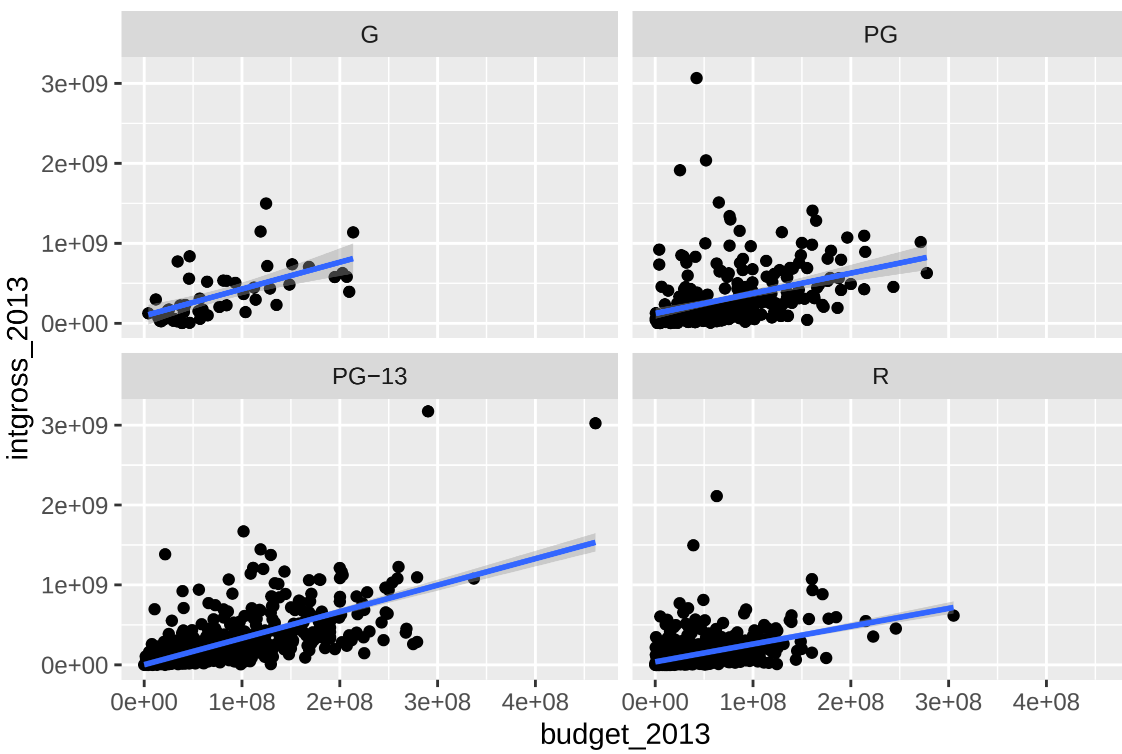
Example

Let's look at modeling a movie's international gross earnings in inflation-adjusted 2013 dollars (`intgross_2013`) as a function of its budget (`budget_2013`) and its MPAA ratings category (`mpaa_rating`, 4 levels: "G", "PG", "PG-13", and "R").

```r
ggplot(data = movies, mapping = aes(x = budget_2013, y = intgross_2013, color = mpaa_rating)) +
  geom_point()
```



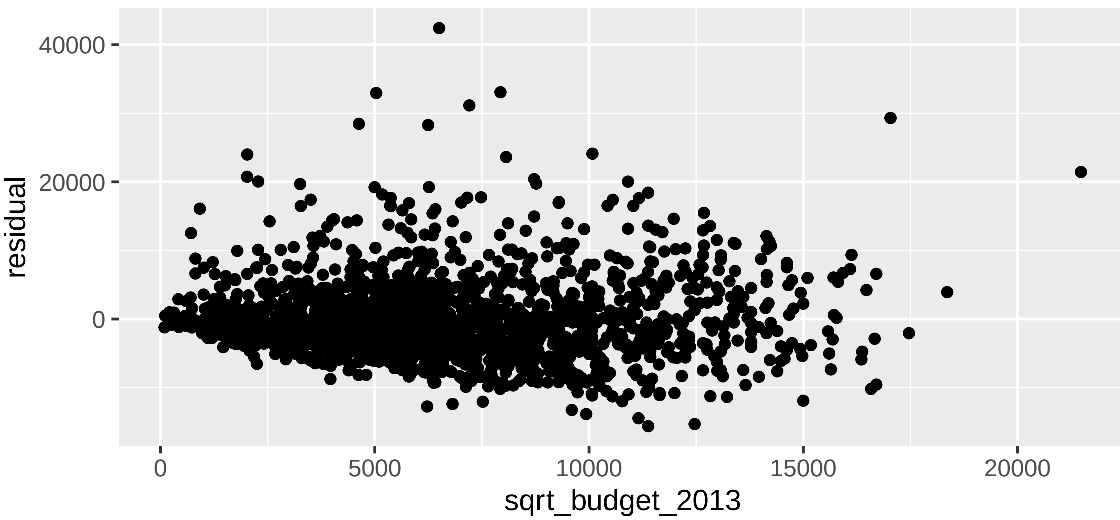

```r
ggplot(data = movies, mapping = aes(x = budget_2013, y = intgross_2013)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~ mpaa_rating)
```



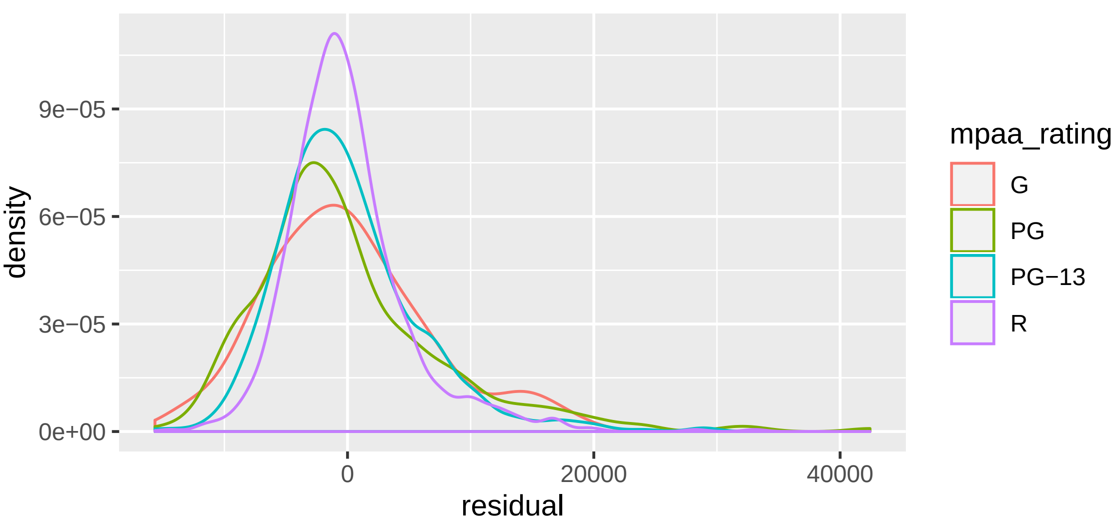
- Both variables are skewed right, with serious outliers.
- Let's try transforming both variables.
- This is essentially the same process we went through the other day, but now I am looking at 2 sets of diagnostic plots:
  - scatter plot of residuals vs budget (looking for: no trends in residuals, constant variance, no outliers)
  - residuals vs MPAA rating (looking for: equal variance across different groups, no outliers)

```
movies <- movies %>% mutate(
  sqrt_intgross_2013 = sqrt(intgross_2013),
  sqrt_budget_2013 = sqrt(budget_2013)
)
```

```
lm_fit <- lm(sqrt_intgross_2013 ~ mpaa_rating + sqrt_budget_2013, data = movies)
movies <- movies %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = movies, mapping = aes(x = sqrt_budget_2013, y = residual)) +
  geom_point()
```

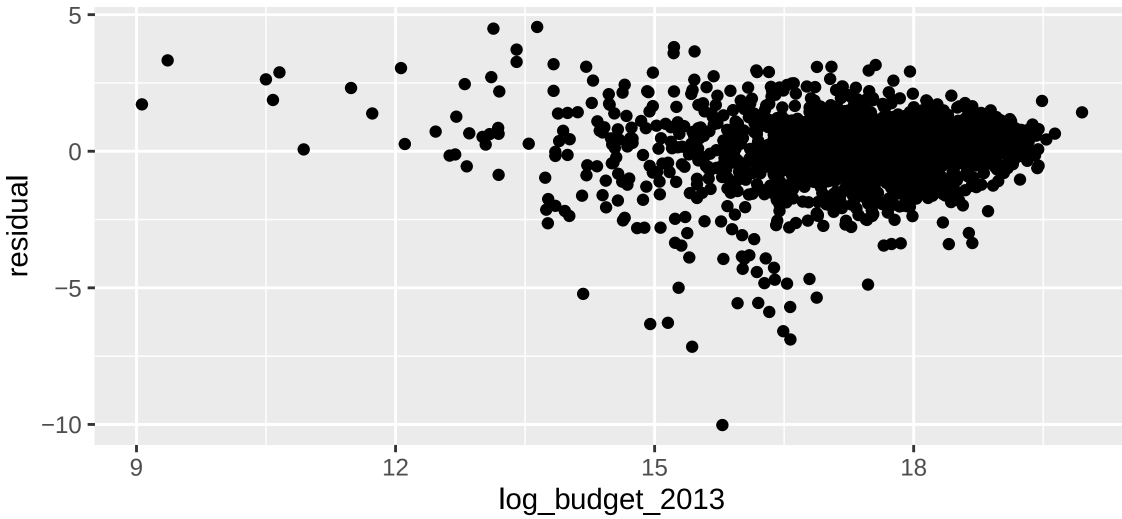



```
ggplot(data = movies, mapping = aes(x = residual, color = mpaa_rating)) +
  geom_density()
```

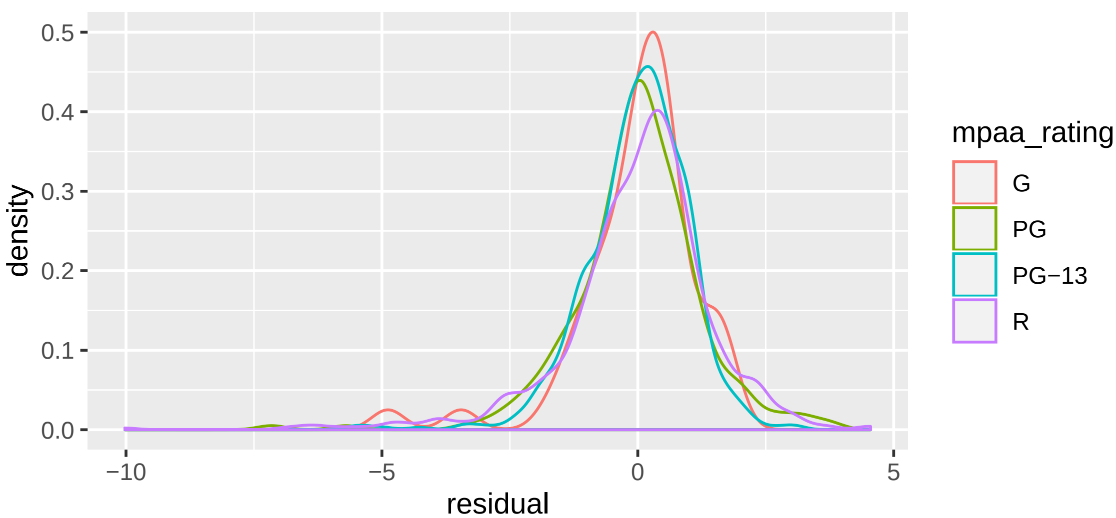


```r
movies <- movies %>% mutate(
  log_intgross_2013 = log(intgross_2013),
  log_budget_2013 = log(budget_2013)
)
```

```r
lm_fit <- lm(log_intgross_2013 ~ mpaa_rating + log_budget_2013, data = movies)
movies <- movies %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = movies, mapping = aes(x = log_budget_2013, y = residual)) +
  geom_point()
```

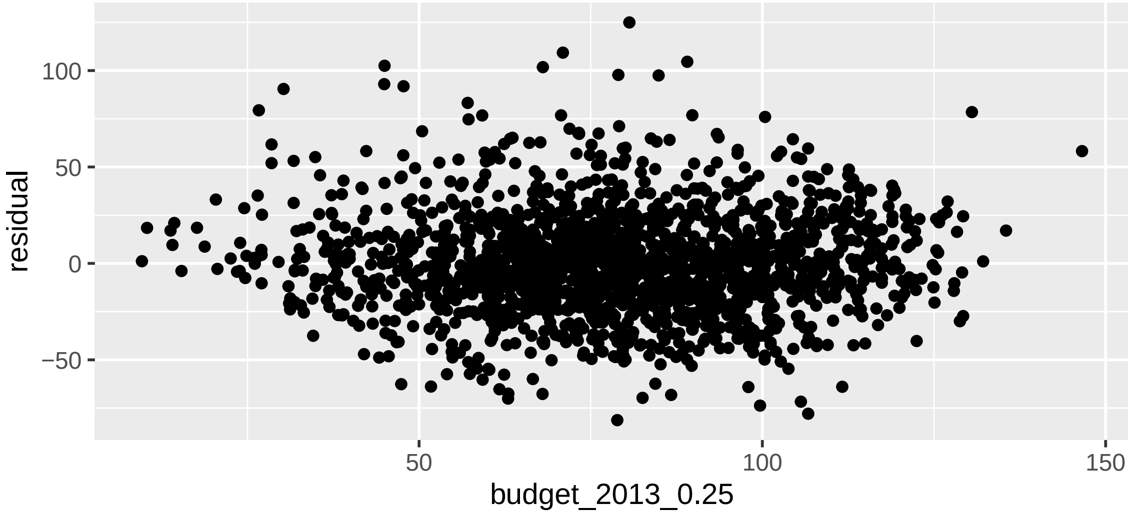



```r
ggplot(data = movies, mapping = aes(x = residual, color = mpaa_rating)) +
  geom_density()
```

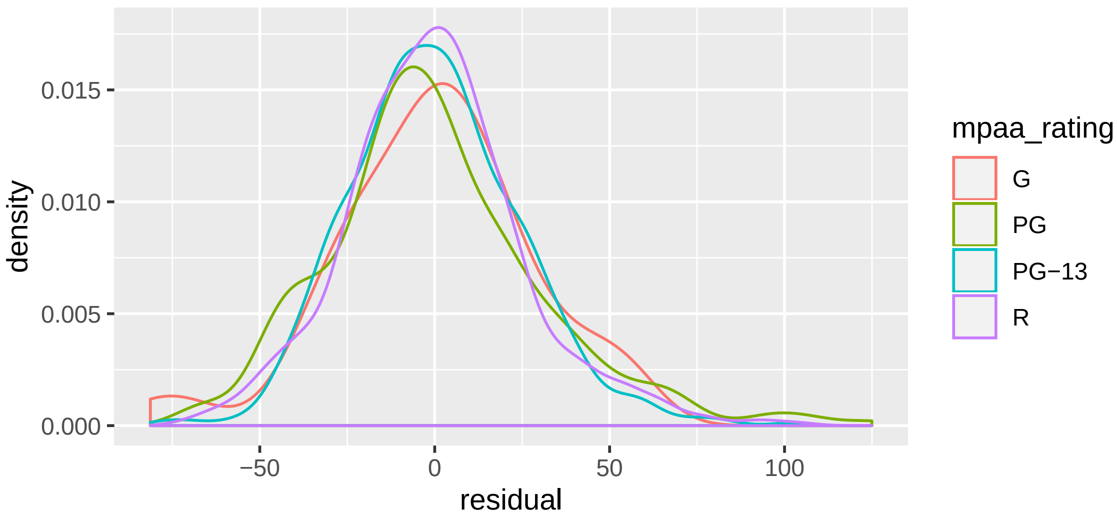


```r
movies <- movies %>% mutate(
  intgross_2013_0.25 = intgross_2013^0.25,
  budget_2013_0.25 = budget_2013^0.25
)
```

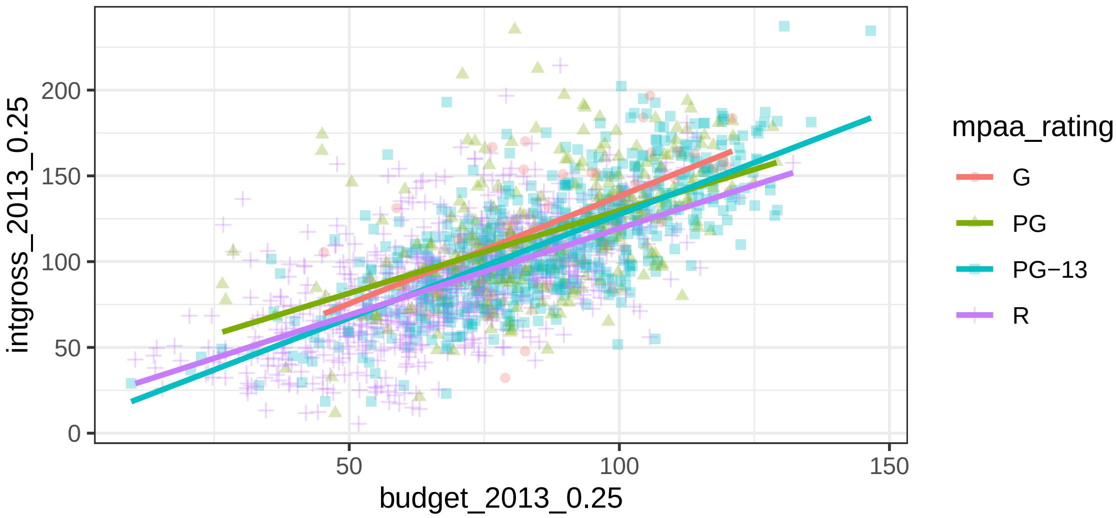
```r
lm_fit <- lm(intgross_2013_0.25 ~ mpaa_rating + budget_2013_0.25, data = movies)
movies <- movies %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = movies, mapping = aes(x = budget_2013_0.25, y = residual)) +
  geom_point()
```




```r
ggplot(data = movies, mapping = aes(x = residual, color = mpaa_rating)) +
  geom_density()
```



```r
ggplot(data = movies,
    mapping = aes(x = budget_2013_0.25, y = intgross_2013_0.25, color = mpaa_rating, shape = mpaa_rating)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```




```r
lm_fit <- lm(intgross_2013_0.25 ~ mpaa_rating + budget_2013_0.25, data = movies)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = intgross_2013_0.25 ~ mpaa_rating + budget_2013_0.25,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.275 -16.235  -1.364  14.516 124.960
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       27.98090    4.60233   6.080 1.48e-09 ***
## mpaa_ratingPG     -5.05739    4.01272  -1.260 0.207715
## mpaa_ratingPG-13 -10.66081    3.84390  -2.773 0.005606 **
## mpaa_ratingR     -14.75736    3.86195  -3.821 0.000137 ***
## budget_2013_0.25   1.08435    0.03062  35.416  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.14 on 1741 degrees of freedom
## Multiple R-squared:  0.4876, Adjusted R-squared:  0.4864
## F-statistic: 414.1 on 4 and 1741 DF,  p-value: < 2.2e-16
```

The `mpaa_rating` variable had four levels: "G", "PG", "PG-13", "R"

There are now 3 indicator variables for the PG, PG-13, and R catagories:

$$\texttt{mpaa\_ratingPG} = \begin{cases} 1 \text{ if mpaa\_rating} = \text{PG} \\ 0 \text{ otherwise (for all other categories)} \end{cases}$$

$$\texttt{mpaa\_ratingPG-13} = \begin{cases} 1 \text{ if mpaa\_rating} = \text{PG-13} \\ 0 \text{ otherwise (for all other categories)} \end{cases}$$

$$\texttt{mpaa\_ratingR} = \begin{cases} 1 \text{ if mpaa\_rating} = \text{R} \\ 0 \text{ otherwise (for all other categories)} \end{cases}$$

1. Write down a single equation for the estimated mean transformed international gross earnings as a function of the MPAA rating category and the transformed budget.

2. Write down separate equations for the estimated mean transformed international gross earnings as a function of the transformed budget for the G, PG, and PG-13 ratings categories

3. What is the interpretation of the coefficient labelled "`mpaa_ratingPG`" in the R output above?

4. Conduct a hypothesis test where the null hypothesis is the claim that in the population of all movies, the intercept of a line describing the relationship between transformed budget and transformed international gross earnings is the same for both G and PG movies.

Are the slopes the same?

```r
fit_different_slopes <- lm(intgross_2013_0.25 ~ budget_2013_0.25 * mpaa_rating, data = movies)
summary(fit_different_slopes)
```

```
##
## Call:
## lm(formula = intgross_2013_0.25 ~ budget_2013_0.25 * mpaa_rating,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.521 -16.310  -0.898  14.618 124.228
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     12.81828   18.59253   0.689    0.491
## budget_2013_0.25                 1.25375    0.20359   6.158 9.11e-10 ***
## mpaa_ratingPG                   20.64188   19.91849   1.036    0.300
## mpaa_ratingPG-13                -6.07255   19.09565  -0.318    0.751
## mpaa_ratingR                     5.59731   18.88251   0.296    0.767
## budget_2013_0.25:mpaa_ratingPG    -0.29099    0.21892  -1.329    0.184
## budget_2013_0.25:mpaa_ratingPG-13 -0.04609    0.20947  -0.220    0.826
## budget_2013_0.25:mpaa_ratingR     -0.24414    0.20862  -1.170    0.242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.08 on 1738 degrees of freedom
## Multiple R-squared:  0.4911, Adjusted R-squared:  0.489
## F-statistic: 239.6 on 7 and 1738 DF,  p-value: < 2.2e-16
```

5. Write down a single equation for the estimated mean transformed international gross earnings as a function of the MPAA rating category and the transformed budget.

6. Write down separate equations for the estimated mean transformed international gross earnings as a function of the transformed budget for the G and PG ratings categories.

7. How strong of evidence do the data provide against the null hypothesis that the slopes of lines describing the relationship between transformed budget and transformed international gross earnings are the same across all four MPAA ratings categories?

```r
anova(lm_fit, fit_different_slopes)
```

```
## Analysis of Variance Table
##
## Model 1: intgross_2013_0.25 ~ mpaa_rating + budget_2013_0.25
## Model 2: intgross_2013_0.25 ~ budget_2013_0.25 * mpaa_rating
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1   1741 1190013
## 2   1738 1181904  3      8109 3.9748 0.00778 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
0.5^7
```

```
## [1] 0.0078125
```

8. Do the data provide strong evidence of a difference in slopes between the G and PG categories? Between the G and PG-13 categories? Between the G and R categories?

Summary of ideas for today

- When considering transformations with multiple explanatory variables, look at plots of residuals vs. each explanatory variable
- If a categorical variable has I categories, there are $I - 1$ corresponding indicator variables in the model describing offsets from a baseline category.
- Although the same variable may appear in different models, the coefficient estimates, interpretations, and p-values depend on what other variables are included
- F tests about multiple coefficients and t tests about the individual coefficients can give seemingly contradictory results - trust the F test more