Nonlinearity: Log-Transforming the Outcome

2017-04-18

Preparation

In this set of notes, you will learn another method of dealing with nonlinearity. Specifically, we will look at transforming the outcome variable using a nonlinear transformation. The data we will use in this set of notes, *movies.csv*, contains data for $n = 1,806$ movies. We will use these data to explore two potential predictors of budget; namely age of a movie and the MPAA rating. The variables are:

- `title`: Movie's title
- `budget`: Movie's budget (in millions of U.S. dollars)
- `age`: Age of the movie; Computed by subtracting the movie's release date from 2017
- `mpaa`: MPAA rating (PG, PG-13, R)

These data are a subset of data from the `movies` data object included in the **ggplot2movies** package. The original data contains information on 24 variables collected from 28,819 movies.

```
movies = read.csv(file = "~/Google Drive/Documents/epsy-8251/data/movies.csv")
head(movies)
```

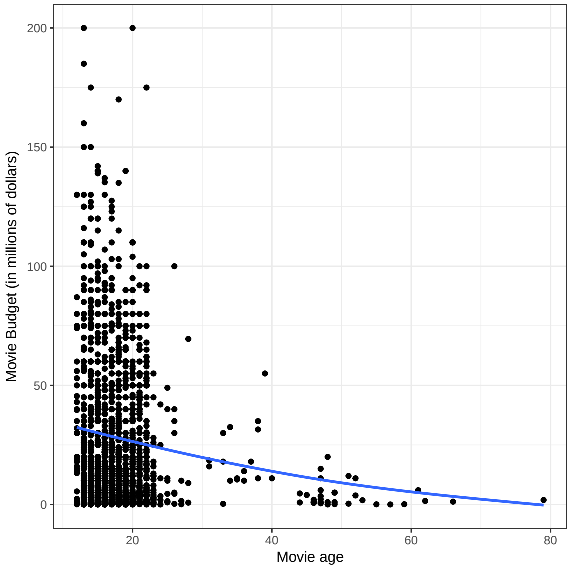
```
                     title budget age  mpaa
1         'Til There Was You   23.0  20 PG-13
2 10 Things I Hate About You   16.0  18 PG-13
3              100 Mile Rule    1.1  15     R
4             13 Going On 30   37.0  13 PG-13
5          13th Warrior, The   85.0  18     R
6                 15 Minutes   42.0  16     R
```

```
# Load libraries
library(dplyr)
library(ggplot2)
library(sm)
```

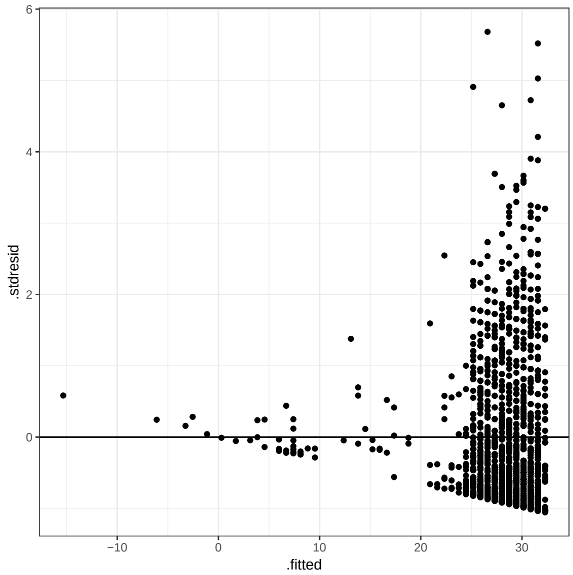
Relationship between Budget and Age

A quick examination of the data suggest that there is a negative effect of age on budget, in the sample. In other words, older movies tend to have a smaller budget, on average. The scatterplot also foreshadows issues with homogeneity of variance and also suggest the relationship may be nonlnear. To examine the nonlinearity issue more, we will fit a linear model and scrutinize the residuals.

```
ggplot(data = movies, aes(x = age, y = budget)) +
  geom_point() +
  geom_smooth(se = FALSE) +
    theme_bw() +
    xlab("Movie age") +
    ylab("Movie Budget (in millions of dollars)")
```

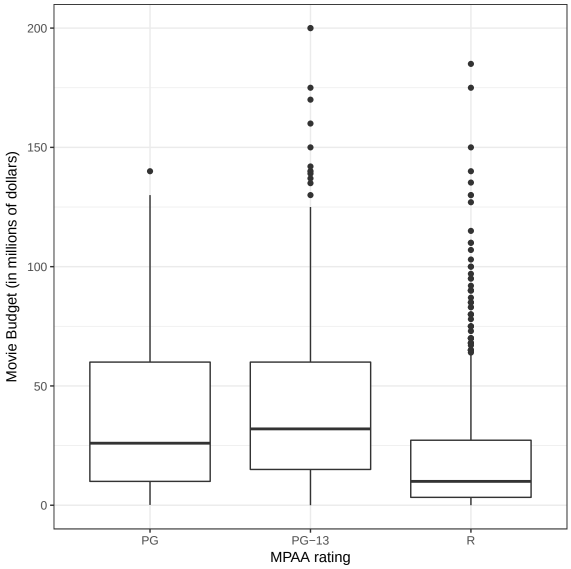


```
# Fit linear model
lm.1 = lm(budget ~ 1 + age, data = movies)

# Obtain residuals
out = fortify(lm.1)

# Plot std. residuals vs. fitted values
ggplot(data = out, aes(x = .fitted, y = .stdresid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  theme_bw()
```



The residual plot re-confirms the assumption of homoskedasticity is clearly violated. It also displays some minor issues with nonlinearity. Before we deal with this issue, we will also examine the relationship between budget and MPAA rating.

```r
ggplot(data = movies, aes(x = mpaa, y = budget)) +
  geom_boxplot() +
    theme_bw() +
    xlab("MPAA rating") +
    ylab("Movie Budget (in millions of dollars)")
```



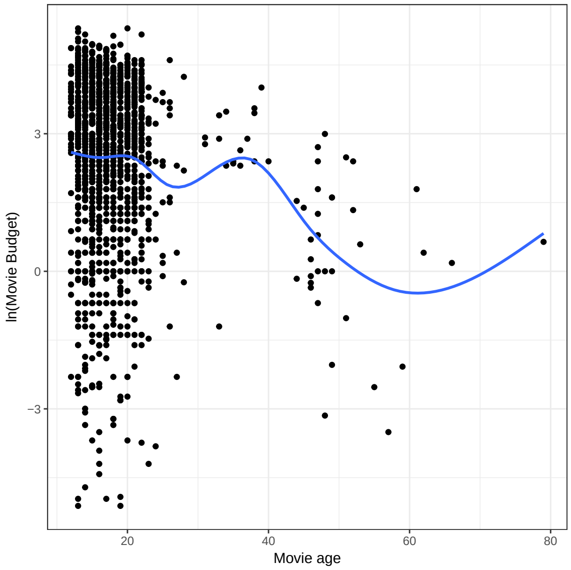
The side-by-side boxplots show there are differences in the median budgets for the three MPAA ratings. They also suggest that all three conditional distributions of budget are positively skewed with large outliers. Lastly, looking at the box widths AND the overall range of each distribution, the plot suggests potential heterogeneity of variance.

Transform the Outcome Using the Natural Logarithm (Base-e)

The non-linear relationship that we saw in the budget versus age scatterplot, is consistent with an *exponential decay* function. Exponential decay (and exponential growth) functions can be modeled by log-transforming the outcome variable and regressing this transformed outcome on the predictor(s). Any base can be used for the logarithm, but we will transform the outcome using the natural logarithm because of the interpretive value.

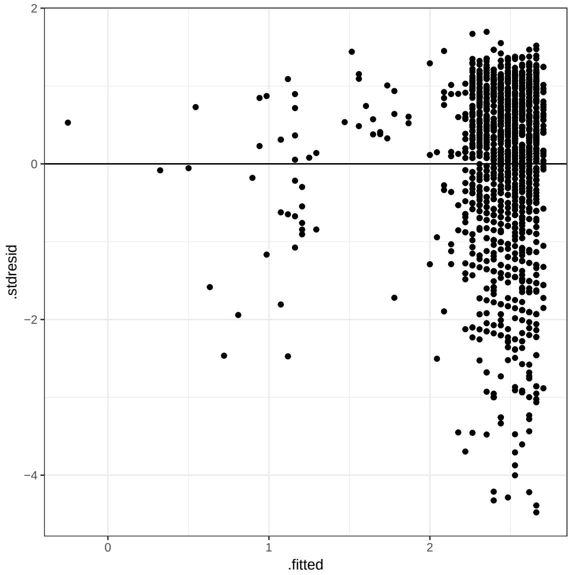
Rather than create the log-transformed variable as a new column in the data, we will just replace `budget` with `log(budget)` in all of the analysis we do (e.g., in the `lm()` and `ggplot()` calls). First, we will re-examine the scatterplot using the transformed outcome to see how this transformation affects the relationship.

```r
ggplot(data = movies, aes(x = age, y = log(budget))) +
  geom_point() +
  geom_smooth(se = FALSE) +
    theme_bw() +
    xlab("Movie age") +
    ylab("ln(Movie Budget)")
```



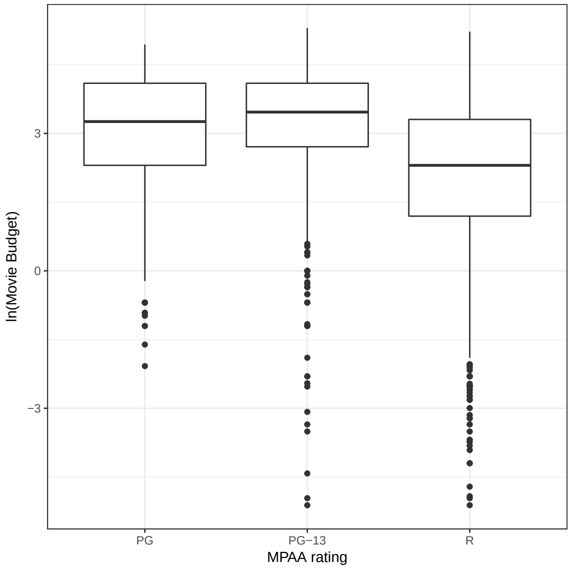
Log-transforming the outcome has affected the relationship in two ways. First, the exponential (curved) decay we saw previously is now much more linear. Secondly, the variance is much more homogenous (although still maybe problematic?). We can probably also see this in the residuals.

```r
# Fit linear model
lm.1 = lm(log(budget) ~ 1 + age, data = movies)

# Obtain residuals
out = fortify(lm.1)

# Plot std. residuals vs. fitted values
ggplot(data = out, aes(x = .fitted, y = .stdresid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  theme_bw()
```



The residuals show decidedly better fit to the linearity assumption. If we log-transform (or transform in any way) the outcome, this affects the relationship with all the predictors. Because of this, we should see how the relationship between other predictors are also affected.

```r
ggplot(data = movies, aes(x = mpaa, y = log(budget))) +
  geom_boxplot() +
    theme_bw() +
    xlab("MPAA rating") +
    ylab("ln(Movie Budget)")
```



There are still median differences in log-budget (differences in budget). The conditional distributions are now more symmetric, however the potential outlying observations are now at the low end of budget. The conditional distributions are also much more homogenous in their variation.

Interpreting the Regression Output

Before we include MPAA rating, let's examine the output from the model that regressed log-transformed budget on age.

```r
summary(lm.1)
```

```

Call:
lm(formula = log(budget) ~ 1 + age, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.7770 -0.8296  0.3348  1.2780  2.9459 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.234047   0.132228  24.458  < 2e-16 ***
age         -0.044084   0.007035  -6.267  4.6e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.738 on 1804 degrees of freedom
Multiple R-squared:  0.02131,	Adjusted R-squared:  0.02076 
F-statistic: 39.27 on 1 and 1804 DF,  p-value: 4.6e-10
```

The model-level summary information suggests that differences in movie age explains 2.1% of the variation in budget. (Remember, explaining variation in log-budget is the same as explaining variation in budget). Although this is a small amount of variation, it is statistically significant, $F(1,1804) = 39.27$, $p < 0.001$. The fitted equation is

$$\ln\left(\hat{\text{Budget}}_i\right) = 3.23 - 0.04(\text{Age}_i)$$

With log-transformations, there are two possible interpretations we can offer. The first is to interpret the coefficients using the log-transformed values:

- The intercept, $\hat{\beta}_0 = 3.23$, is the average predicted log-budget for movies that are zero years old (made in 2017).
- The slope, $\hat{\beta}_1 = -0.04$, indicates that each one-year difference in age is associated with a -0.04-unit difference in log-budget, on average.

A second, probably more useful, interpretation is to back-transform log-budget to budget. To do this, we exponentiate both sides of the fitted equation.

$$e^{\ln\left(\hat{\text{Budget}}_i\right)} = e^{3.23 - 0.04(\text{Age}_i)}$$

Now we use the rules of exponents to simplify this,

$$\hat{\text{Budget}}_i = e^{3.23} \times e^{-0.04(\text{Age}_i)}$$

Substitute in $\text{Age} = 0$ to estimate the average budget for movies made in 2017. The predicted average budget for movies made in 2017 is \$25.28 million.

$$\begin{aligned}\hat{\text{Budget}}_i &= e^{3.23} \times e^{-0.04(0)} \\ &= e^{3.23} \times 1 \\ &= 25.28\end{aligned}$$

What is the budget difference for movies that are made one year apart? To determine this substitute in two ages that differ by a year, say Age = 0 and Age = 1. Since we already know the predicted average budget for Age = 0, we only need to do this for Age = 1.

$$\begin{aligned}\hat{\text{Budget}}_i &= e^{3.23} \times e^{-0.04(1)} \\ &= 25.28 \times e^{-0.04} \\ &= 25.28 \times 0.96\end{aligned}$$

We could multiply out that last part, but leaving it like this tells us how the budgets differ. Namely that movies one year apart have, on average, budgets that are predicted to differ by a factor of 0.96 (a 0.96-fold difference). In general the back-transformed interpretations when we log-transform the outcome using the natural logarithm are:

- The average Y when $X = 0$ is predicted to be $e^{\hat{\beta}_0}$.
- Each one-unit difference in X is associated with a $e^{\hat{\beta}_1}$-fold (or factor) difference in Y, on average.

We can obtain these values by using the `exp()` function to exponentiate the coefficients from the fitted model, which we obtain using the `coef()` function.

```
exp(coef(lm.1))
```

```
(Intercept)         age 
 25.3821648   0.9568737 
```

Interpret the Slope Directly from the Output

Our fitted slope suggested a 0.96-fold difference in budgets for movies made one year apart, on average. This is equivalent to a 4% decrease in the average budget. This can be inferred directly from the slope coefficient in the `summary()` output.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.234047   0.132228  24.458  < 2e-16 ***
age         -0.044084   0.007035  -6.267  4.6e-10 ***
---
```

So, an alternative interpretation is to say that each one-year difference in movie age is associated with a 4% decrease in budget, on average. In general, multiplying the fitted slope coefficient by 100 will give a rough estimation of the percentage change. (Note: It is more accurate to use the formula $e^{\hat{\beta}_1} - 1$, which in our case gives $-.0392$ (a 3.92% decrease), but when you do not have a computer handy to compute this, the shortcut will suffice.) This shortcut only works when we have transformed the outcome using the natural logarithm; any other base would not give the percent change directly in the output.

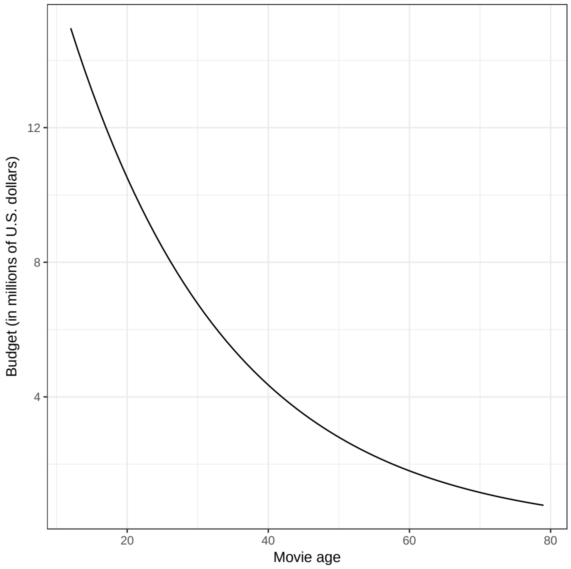
Plot of the Fitted Model

To plot the fitted model, we use a range of age values to predict the log-budgets. Then we back-transform the log-budget to budget and plot age vs. budget.

```r
# Set up data
plotData = expand.grid(
  age = 12:79
)

# Predict log-budget
plotData$Lbudget = predict(lm.1, newdata = plotData)

# Back-transform log-budget to budget
plotData$budget = exp(plotData$Lbudget)
head(plotData)
```

```
  age  Lbudget   budget
1  12 2.705040 14.95492
2  13 2.660956 14.30997
3  14 2.616872 13.69283
4  15 2.572788 13.10231
5  16 2.528705 12.53726
6  17 2.484621 11.99657
```

```r
# Plot
ggplot(data = plotData, aes(x = age, y = budget)) +
  geom_line() +
  theme_bw() +
  xlab("Movie age") +
  ylab("Budget (in millions of U.S. dollars)")
```



The plot helps us see the exponential relationship between movie age and budget. While overall, the effect of age is negative (older movies have, on average, a lower budget), the nonlinearity indicates that the effect of age varies by the age of the movie. For more recent movies, the negative effect of age is larger, while for older movies, this effect diminishes.

MPAA Rating as a Predictor of Log-Budget

Now we will fit a regression model using MPAA rating to predict budget, again using log-budget as our outcome. Since MPAA rating is a categorical predictor, we first create a dummy variable for each rating.

```r
# Create dummy variables
movies$pg = ifelse(movies$mpaa == "PG", 1, 0)
movies$pg13 = ifelse(movies$mpaa == "PG-13", 1, 0)
movies$r = ifelse(movies$mpaa == "R", 1, 0)
#head(movies)
```

Then we fit the model using any two of the dummy variables.

```r
lm.2 = lm(log(budget) ~ 1 + pg13 + r, data = movies)
summary(lm.2)
```

```

Call:
lm(formula = log(budget) ~ 1 + pg13 + r, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2582 -0.6265  0.3662  1.1753  3.2076 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.8800     0.1152  25.006  < 2e-16 ***
pg13          0.2622     0.1363   1.924   0.0545 .  
r            -0.8671     0.1261  -6.875 8.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.677 on 1803 degrees of freedom
Multiple R-squared:  0.08919,	Adjusted R-squared:  0.08818 
F-statistic: 88.28 on 2 and 1803 DF,  p-value: < 2.2e-16
```

Differences in MPAA rating explain 8.9% of the variation in movie budgets. This is statistically significant, $F(2,\ 1803) = 88.28$, $p < .001$. The fitted equation is

$$\ln\left(\widehat{\text{Budget}_i}\right) = 2.88 + 0.26(\text{PG13}_i) - 0.87(\text{R}_i)$$

Again, with log-transformations, there are two possible interpretations we can offer. The first is to interpret the coefficients using the log-transformed values:

- The intercept, $\hat{\beta}_0 = 2.88$, is the average predicted log-budget for PG rated movies (the reference group).
- The estimated PG-13 effect, $\hat{\beta}_1 = 0.26$, indicates that PG-13 rated movies are predicted to have a log-budget that is 0.26-units higher than PG rated movies, on average.
- The estimated R effect, $\hat{\beta}_2 = -0.87$, indicates that R rated movies are predicted to have a log-budget that is 0.87-units lower than PG rated movies, on average.

Better interpretation can be offered by back-transforming (exponentiating) the coefficients and interpreting the effects in the metric of budget.

```
exp(coef(lm.2))
```

```
(Intercept)        pg13           r 
 17.8133826   1.2998280   0.4201486 
```

- The average predicted budget for PG rated movies (the reference group) is $17.8 million.
- PG-13 rated movies are predicted to have a budget that is 130% of that for PG rated movies, on average.
- R rated movies are predicted to have a budget that is 42% of that for PG rated movies, on average.

Significant Differences in MPAA Rating?

If we are interested in whether these differences are statistically significant, we need to consider all threee pairwise comparisons. Currently we have the PG vs. PG-13 comparison and the PG vs. R comparison. We also need to determine the PG-13 vs. R comparison. (To obtain this we would need to fit another model using one of these ratings as the reference group.) Below are the comparisons and the unadjusted *p*-values obtained from fitting the `lm()`.

```
    Comparison        p
1 PG vs. PG-13 5.45e-02
2     PG vs. R 8.49e-12
3  PG-13 vs. R 2.00e-16
```

We adjust these *p*-values using the Benjamini-Hochberg adjustment (or some other method).

```
# Input the p-values into a vector
p = c(0.0545, 0.00000000000849, 0.0000000000000002)

# Adjust using the BH method
p.adjust(p, method = "BH")
```

```
[1] 5.4500e-02 1.2735e-11 6.0000e-16
```

```
    Comparison        p        BHp
1 PG vs. PG-13 5.45e-02 5.4500e-02
2     PG vs. R 8.49e-12 1.2735e-11
3  PG-13 vs. R 2.00e-16 6.0000e-16
```

The adjusted *p*-values suggest that we have statistically significant differences in the average budgets between PG and R rated movies and between PG-13 and R rated movies. There is not a statistically significant difference between the average budget for PG and PG-13 rated movies.

Multiple Predictors: Main Effects Model

Now we will fit a model that uses both age of a movie and MPAA rating to predict variation in budget, agin using the transformed log-budget as our outcome variable. In this model PG rated movies will again be our reference group.

```r
lm.3 = lm(log(budget) ~ 1 + age + pg13 + r, data = movies)
summary(lm.3)
```

```

Call:
lm(formula = log(budget) ~ 1 + age + pg13 + r, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4518 -0.6586  0.3298  1.1342  2.9966 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.687688   0.169910  21.704  < 2e-16 ***
age         -0.043101   0.006727  -6.407 1.89e-10 ***
pg13         0.208437   0.135045   1.543    0.123    
r           -0.903643   0.124876  -7.236 6.79e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.659 on 1802 degrees of freedom
Multiple R-squared:  0.1095,	Adjusted R-squared:  0.108 
F-statistic: 73.84 on 3 and 1802 DF,  p-value: < 2.2e-16
```

Differences in movie age and MPAA rating explain 11.0% of the variation in movie budgets. This is statistically significant, $F(3,\ 1802) = 73.84$, $p < .001$. The fitted equation is

$$\ln\left(\hat{\mathrm{Budget}}_i\right) = 3.69 - 0.04(\mathrm{Age}_i) + 0.21(\mathrm{PG13}_i) - 0.90(\mathrm{R}_i)$$

Here we will just provide the back-transformed interpretations of the fitted coefficients.

```r
exp(coef(lm.3))
```

```
(Intercept)         age        pg13           r 
 39.9523650   0.9578146   1.2317513   0.4050910 
```

- The average predicted budget for PG rated movies (the reference group) that was made in 2017 is $39.95 million.
- Each one-year difference in movie age is associated with a budget that is 4% lower, on average, cotrolling for differences in MPAA rating.
- PG-13 rated movies are predicted to have a budget that is 123% of that for PG rated movies, on average, controlling for differences in movie age.
- R rated movies are predicted to have a budget that is 41% of that for PG rated movies, on average, controlling for differences in movie age.

Multiple Predictors: Interaction Model

Let's also examine whether the effect of age varies by MPAA rating. To examine this, we fit the interaction model. Since we have multiple predictors that make up the effect of MPAA rating (`pg13` and `r`), we need to create multiple interaction terms to be included in the model, namely `pg13:age` and `r:age`.

```
lm.4 = lm(log(budget) ~ 1 + age + pg13 + r + age:pg13 + age:r, data = movies)
```

When there are multiple interaction effects, it is often better to use the nested F-test to examine whether or not there is an interaction effect. Recall that we will be comparing the main-effects model to the interaction model.

```
anova(lm.3, lm.4)
```

```
Analysis of Variance Table

Model 1: log(budget) ~ 1 + age + pg13 + r
Model 2: log(budget) ~ 1 + age + pg13 + r + age:pg13 + age:r
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1802 4957.2
2   1800 4949.4  2    7.7623 1.4115  0.244
```

The results of the nested F-test, $F(2,\ 1800) = 1.42$, $p = .244$, indicate that the interaction model is NOT explaining additional variation in budget. As such, we should stick with the main-effects model. (Note that if the interaction model was adopted, we would immediately create a plot to help interpret the effects.)

Plotting the Fitted Main-Effects Model

Since we adopted the main-effects model, we can also produce a plot ofthe back-transformed fitted values to aid interpretation of the effects. To create this plot, we use a range of age values and the different MPAA ratings to predict the log-budgets. Then we back-transform the log-budget to budget and plot age vs. budget for each of the MPAA ratings. Prior to doing this, we will refit the main-effects model using the original `mpaa` variable instead of the two dummy variables. This will make creating this plot easier.

```
# Re-fit main-effects model
lm.4 = lm(log(budget) ~ 1 + age + mpaa, data = movies)
summary(lm.4)
```

```
Call:
lm(formula = log(budget) ~ 1 + age + mpaa, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4518 -0.6586  0.3298  1.1342  2.9966

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.687688   0.169910  21.704  < 2e-16 ***
age         -0.043101   0.006727  -6.407 1.89e-10 ***
mpaaPG-13    0.208437   0.135045   1.543    0.123
mpaaR       -0.903643   0.124876  -7.236 6.79e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

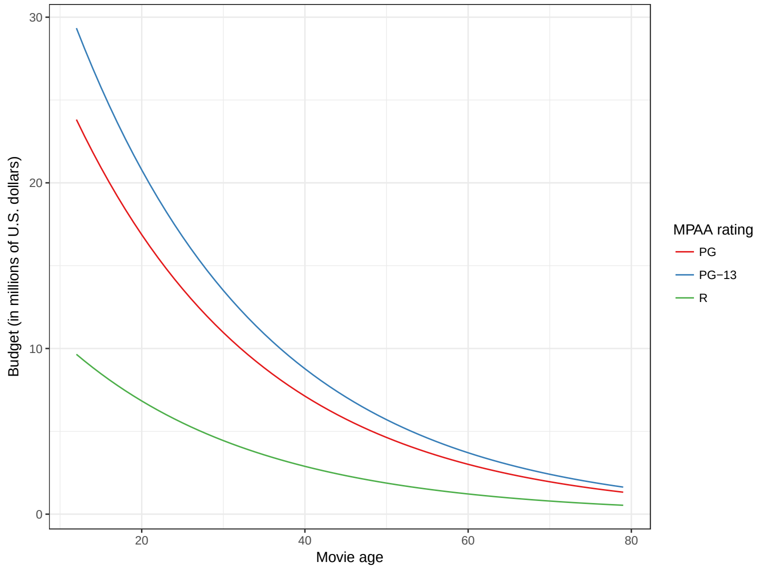
```
Residual standard error: 1.659 on 1802 degrees of freedom
Multiple R-squared:  0.1095,    Adjusted R-squared:  0.108
F-statistic: 73.84 on 3 and 1802 DF,  p-value: < 2.2e-16
```

Note that the regression output is the same as that for the main-effects model when we used the dummy variables, its just that R chose the reference category for us. Now when we set up our plotting data, we will use the original categories in the `mpaa` variable

```
# Set up data
plotData = expand.grid(
  age = 12:79,
  mpaa = c("PG", "PG-13", "R")
)

# Predict log-budget
plotData$Lbudget = predict(lm.4, newdata = plotData)

# Back-transform log-budget to budget
plotData$budget = exp(plotData$Lbudget)
head(plotData)
```

```
  age mpaa  Lbudget   budget
1  12   PG 3.170475 23.81881
2  13   PG 3.127374 22.81400
3  14   PG 3.084273 21.85158
4  15   PG 3.041172 20.92976
5  16   PG 2.998071 20.04683
6  17   PG 2.954970 19.20115
```

```
# Plot
ggplot(data = plotData, aes(x = age, y = budget, color = mpaa)) +
  geom_line() +
  theme_bw() +
  xlab("Movie age") +
  ylab("Budget (in millions of U.S. dollars)") +
  scale_color_brewer(name = "MPAA rating", palette = "Set1")
```



The plot helps us see (1) the exponential (decaying) relationship between movie age and budget for PG, PG-13, and R rated movies. It also helps us see the budget differences between PG, PG-13, and R rated movies are smaller for older movies. Even though we fitted a main-effects model, the fitted lines after we back-transform are not parallel.How non-parallel the lines are depends on the size of the coefficients associated with the MPAA effects (in this example). This is why, especially with transformed data, it is essential to plot the model to make sure you are understanding the interpretations from your coefficients.