

Variance Stabilizing Transformation: Log-Transforming the Outcome

2017-11-28

Preparation

In this set of notes, you will learn about transforming the outcome variable as a method for dealing with heteroscedasticity and non-normality. Specifically, we will look at transforming the outcome variable using a logarithmic 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.

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

# Read in data
movies = read_csv(file = "~/Dropbox/epsy-8251/data/movies.csv")
head(movies)
```

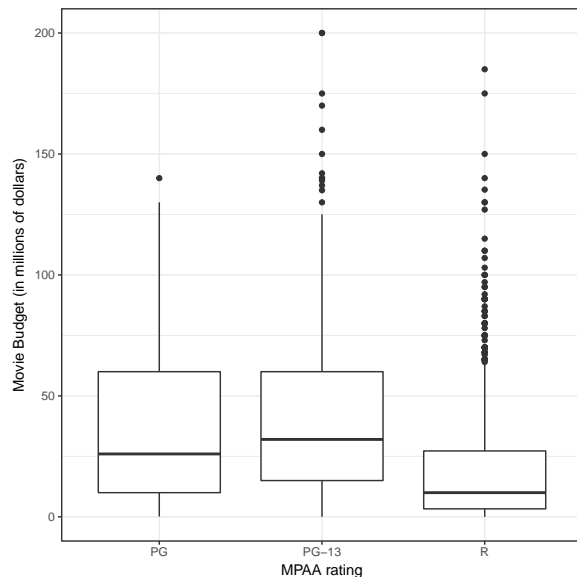
```
# A tibble: 6 x 4
```

	title	budget	age	mpaa
	<chr>	<dbl>	<int>	<chr>
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

Relationship between Budget and MPAA Rating

The primary research question we will address is whether there are mean differences in movie budget between movies with different MPAA Ratings. We will also examine the relationship between budget and MPAA rating.

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



```
# Compute summary statistics
movies %>%
  group_by(mpaa) %>%
  summarize(M = mean(budget), SD = sd(budget))
```

```
# A tibble: 3 x 3
  mpaa      M      SD
  <chr>  <dbl>  <dbl>
1 PG    36.45425 34.04058
2 PG-13 41.42601 34.83981
3 R     19.81063 24.60763
```

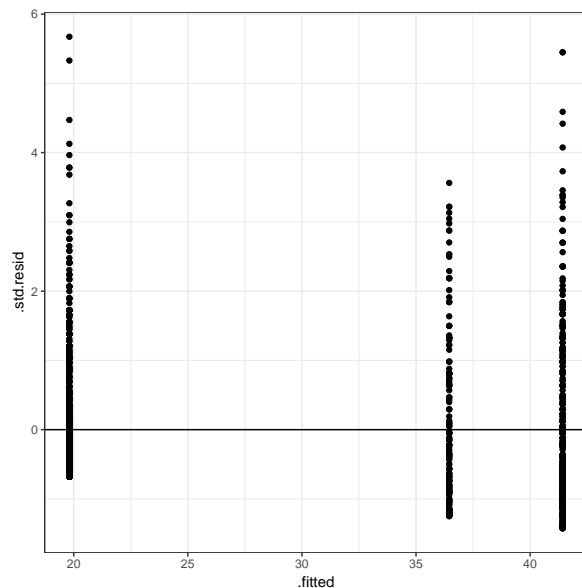
The side-by-side boxplots and summary statistics show there are differences in the mean budgets for the three MPAA ratings. All three conditional distributions of budget are positively skewed, and each seems to have movies with exceedingly large budgets (potential outliers). Lastly, looking at the box widths AND the overall range of each distribution, the plot suggests potential heterogeneity of variance.

To examine these issues further, we will fit a linear model including MPAA rating as a predictor of budget and scrutinize the residuals.

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

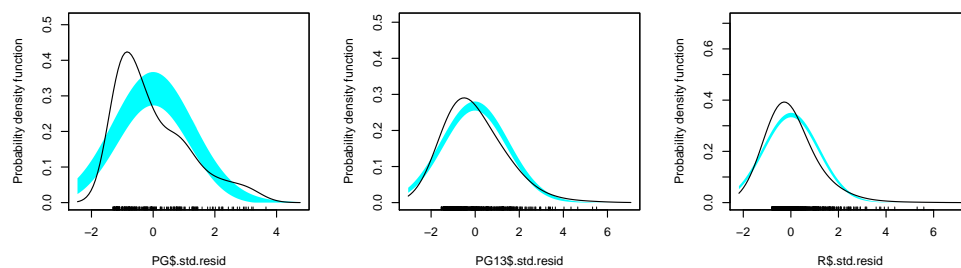
# Obtain residuals
out = augment(lm.1)
```

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



```
# Obtain each mpaa ratings residuals
PG = out %>% filter(mpaa == "PG")
PG13 = out %>% filter(mpaa == "PG-13")
R = out %>% filter(mpaa == "R")
```

```
# Density plots
par(mfrow = c(1, 3))
sm.density(PG$.std.resid, model = "Normal", main = "PG")
sm.density(PG13$.std.resid, model = "Normal", main = "PG")
sm.density(R$.std.resid, model = "Normal", main = "PG")
```



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

The residual plots confirm that non-normality is a problem. They also show some evidence of heteroskedasticity.

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

To alleviate problems of non-normality when the conditional distributions are right-skewed (or have high-end outliers) OR to alleviate heteroskedasticity, we can mathematically transform the outcome using a logarithm.

Any base can be used for the logarithm, but we will transform the outcome using the natural logarithm because of the interpretive value.

First, we will create the log-transformed variable as a new column in the data, and then we will use the log-transformed budget (rather than raw budget) in any analyses. To log-transform we use the `log()` function. This function has an argument, `base=`, to change the base used in the logarithm. By default this is set to base- e (the natural logarithm).

```
movies = movies %>% mutate(Lbudget = log(budget))
head(movies)
```

```
# A tibble: 6 x 5
```

	title	budget	age	mpaa	Lbudget
	<chr>	<dbl>	<int>	<chr>	<dbl>
1	'Til There Was You	23.0	20	PG-13	3.13549422
2	10 Things I Hate About You	16.0	18	PG-13	2.77258872
3	100 Mile Rule	1.1	15	R	0.09531018
4	13 Going On 30	37.0	13	PG-13	3.61091791
5	13th Warrior, The	85.0	18	R	4.44265126
6	15 Minutes	42.0	16	R	3.73766962

Some Quick Math

Recall that the logarithm is the inverse function of an exponent. As an example, consider the budget and log-transformed budget for *'Til There Was You*.

$$\begin{aligned}\ln(\text{Budget}) &= 3.135 \\ \ln(23.0) &= 3.135\end{aligned}$$

Or,

$$e^{3.135} = 23.0$$

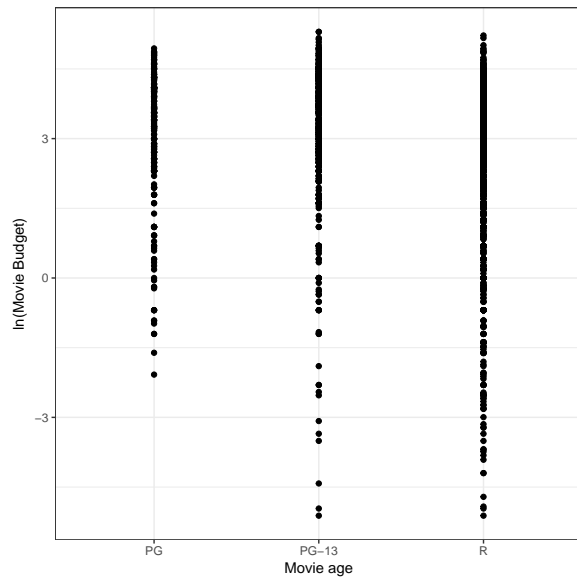
The logarithm answers the mathematical question:

e to what power is equal to 23.0?

Re-analyze using the Log-Transformed Budget

Now we will re-examine the scatterplot using the log-transformed outcome to see how this transformation affects the relationship.

```
# Plot the conditional distributions
ggplot(data = movies, aes(x = mpaa, y = Lbudget)) +
  geom_point() +
  theme_bw() +
  xlab("Movie age") +
  ylab("ln(Movie Budget)")
```



```
# Compute summary statistics
movies %>%
  group_by(mpa) %>%
  summarize(M = mean(Lbudget), SD = sd(Lbudget))
```

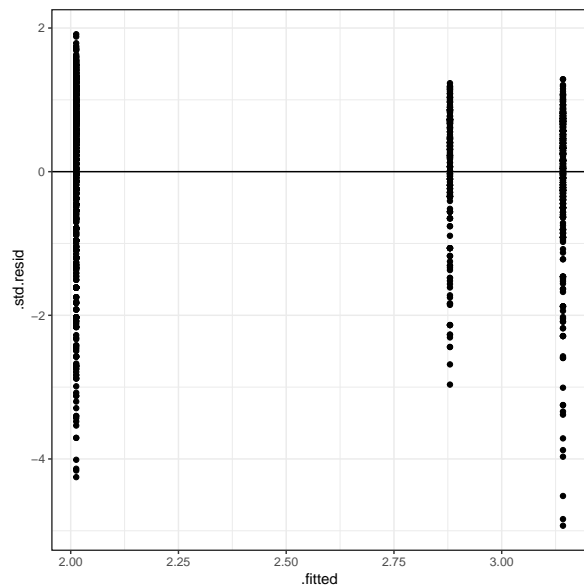
```
# A tibble: 3 x 3
  mpa      M      SD
  <chr>   <dbl> <dbl>
1 PG  2.879950 1.532475
2 PG-13 3.142182 1.509520
3 R    2.012803 1.780308
```

Log-transforming the outcome has drastically affected the scale for the outcome. Now the conditional distributions show movies that are potential outliers at the low-end of log-budget. The summary statistics also show that the conditional distributions have much more equal variances. We can probably also see this in the residuals.

```
# Fit linear model
lm.2 = lm(Lbudget ~ 1 + mpa, data = movies)

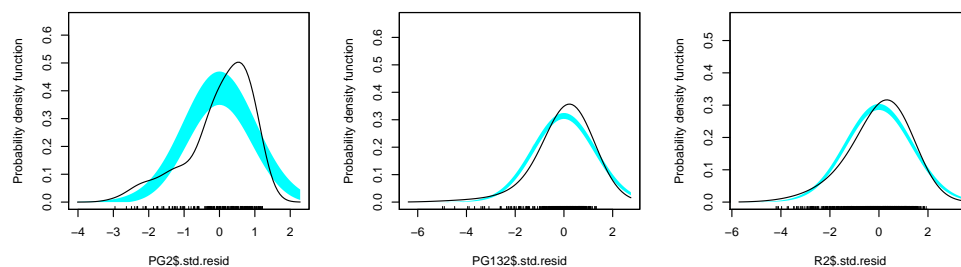
# Obtain residuals
out2 = augment(lm.2)

# Plot std. residuals vs. fitted values
ggplot(data = out2, aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  theme_bw()
```



```
# Obtain each mpaa ratings residuals
PG2 = out2 %>% filter(mpaa == "PG")
PG132 = out2 %>% filter(mpaa == "PG-13")
R2 = out2 %>% filter(mpaa == "R")

# Density plots
par(mfrow = c(1, 3))
sm.density(PG2$.std.resid, model = "Normal", main = "PG")
sm.density(PG132$.std.resid, model = "Normal", main = "PG")
sm.density(R2$.std.resid, model = "Normal", main = "PG")
```



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

The residuals show better fit to the normality assumption (although not perfect) and similarly to the homogeneity of variance assumption.

Interpreting the Regression Output

Let's examine the output from the model that regressed log-transformed budget on MPAA rating

```
summary(lm.2)
```

Call:

```
lm(formula = Lbudget ~ 1 + mpaa, 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 ***
mpaaPG-13	0.2622	0.1363	1.924	0.0545 .
mpaaR	-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

The model-level summary information suggests that differences in MPAA rating explains 8.9% 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, $F(2, 1803) = 88.28$, $p < 0.001$.

From the coefficient-level output we see that R chose PG (the first level in the `mpaa` variable alphabetically) to be the reference group. The fitted equation is

$$\ln(\text{Budget}_i) = 2.88 + 0.26(\text{PG-13}_i) - 0.87(\text{R}_i)$$

With log-transformations, there are two possible interpretations we can offer. The first is to interpret the coefficients using the log-transformed values. These we interpret in the exact same way we do any other regression coefficients (except we use log-outcome instead of outcome):

- The intercept, $\hat{\beta}_0 = 2.88$, is the average predicted log-budget for PG rated movies.
- The slope associated with PG-13, $\hat{\beta}_1 = 0.26$, indicates that PG-13 rated movies have a log-budget that is 0.26 higher than PG rated movies, on average.
- The slope associated with R, $\hat{\beta}_2 = -0.87$, indicates that R rated movies have a log-budget that is 0.87 lower than PG rated movies, on average.

Back-Transforming: A More Useful Interpretation

A second, probably more useful, interpretation is to back-transform log-budget to budget. To think about how to do this, we first consider a more general expression of the fitted linear model:

$$\ln(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1(X_1) + \hat{\beta}_2(X_2)$$

The left-hand side of the equation is in the log-transformed metric, which drives our interpretations. If we want to instead, interpret using the raw metric of Y , we need to back-transform from $\ln(Y)$ to Y . To back-transform, we use the inverse function, which is to exponentiate using the base of the logarithm, in our case, base- e .

$$e^{\ln(Y)} = Y$$

If we exponentiate the left-hand side of the equation, to maintain the equality, we also need to exponentiate the right-hand side of the equation.

$$e^{\ln(Y)} = e^{\hat{\beta}_0 + \hat{\beta}_1(X_1) + \hat{\beta}_2(X_2)}$$

Then we use rules of exponents to simplify this.

$$Y = e^{\hat{\beta}_0} \times e^{\hat{\beta}_1(X_1)} \times e^{\hat{\beta}_2(X_2)}$$

For our example, when we exponentiate both sides of the fitted equation:

$$\hat{\text{Budget}}_i = e^{2.88} \times e^{0.26(\text{PG-13}_i)} \times e^{-0.87(\text{R}_i)}$$

Substituting in the Dummy Variables to Interpret Effects

To interpret the effects (which are now interpreted using budget—not log-budget) we can substitute in the different dummy variable patterns and solve.

$$\begin{aligned} \textbf{PG Movie: } \hat{\text{Budget}}_i &= e^{2.88} \times e^{0.26(0)} \times e^{-0.87(0)} \\ &= 17.81 \times 1 \times 1 \\ &= 17.81 \end{aligned}$$

The model estimated budget for a PG rated movie is 17.81 million dollars.

$$\begin{aligned} \textbf{PG-13 Movie: } \hat{\text{Budget}}_i &= e^{2.88} \times e^{0.26(1)} \times e^{-0.87(0)} \\ &= 17.81 \times 1.30 \times 1 \\ &= 23.15 \end{aligned}$$

The model estimated budget for a PG-13 rated movie is 23.15 million dollars. This is 1.30 TIMES the estimated budget for a PG rated movie.

$$\begin{aligned} \textbf{R Movie: } \hat{\text{Budget}}_i &= e^{2.88} \times e^{0.26(0)} \times e^{-0.87(1)} \\ &= 17.81 \times 1 \times 0.42 \\ &= 7.48 \end{aligned}$$

The model estimated budget for an R rated movie is 7.48 million dollars. This is 0.42 TIMES the estimated budget for a PG rated movie.

Simply put, when we back-transform from interpretations of log- Y to Y the interpretations are multiplicatively related to the intercept rather than additive. We can obtain these multiplicative values (and the back-transformed intercept) by using the `exp()` function to exponentiate the coefficients from the fitted model, which we obtain using the `coef()` function.

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

(Intercept)	mpaaPG-13	mpaaR
17.8133826	1.2998280	0.4201486

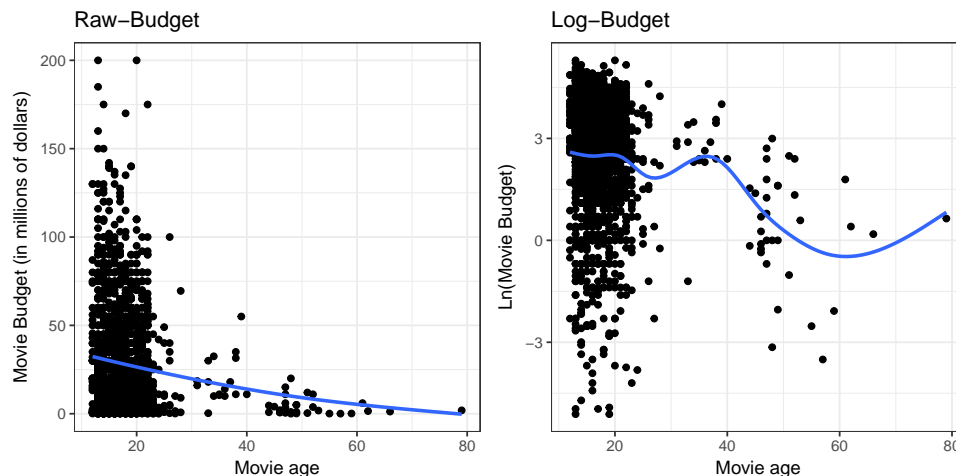
Relationship between Budget and Age

Suppose now that we wanted to include the age of a movie in the model as well as MPAA rating. Since we log-transformed the outcome, we need to see how age is related to log-budget. (We will also, for pedagogical purposes, examine the relationship with age and raw-budget as well.)

```
# Relationship with raw-budget
p1 = 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)") +
  ggtitle("Raw-Budget")

# Relationship with log-budget
p2 = ggplot(data = movies, aes(x = age, y = Lbudget)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  theme_bw() +
  xlab("Movie age") +
  ylab("Ln(Movie Budget)") +
  ggtitle("Log-Budget")

gridExtra::grid.arrange(p1, p2, nrow = 1)
```



Examining the relationship between age and raw-budget, we see a non-linear, negative effect of age on budget, in the sample. In other words, older movies tend to have a smaller budget, on average, but this decrease is not constant. This pattern of non-linear decline is referred to as *exponential decay*. The scatterplot also foreshadows issues with homogeneity of variance and also suggest the relationship may be nonlinear.

The relationship between age and log-budget is much more fortuitous for fitting a linear model. The exponential decay pattern we observed with raw-budget seems much more linear (most of the non-linearity observed in the loess smoother can be attributed to the sparsity of data on the right-hand side of the plot rather than to a non-linear relationship in the data). We also see that the log-transformed data conform much more to the assumption of homogeneity of variance.

Fitting the Linear Model and Interpreting the Coefficients

Although we are mostly interested in age as a control in our MPAA model, we will fit the simple regression using age to predict log-budget to get a feel for how to interpret coefficients when we have a continuous predictor.

```
lm.3 = lm(Lbudget ~ 1 + age, data = movies)
summary(lm.3)
```

Call:

```
lm(formula = Lbudget ~ 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 age of a movie 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, $F(1, 1804) = 39.27$, $p < 0.001$.

To interpret the coefficients, we will again exponentiate the fitted coefficients so we can interpret them using the raw-metric of budget.

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

(Intercept)	age
25.3821648	0.9568737

- The model estimated budget for a movie that was made in 2017 (age = 0) is 25.38 million dollars.
- Each one-year difference in age is associated with a 0.95 TIMES difference in budget, on average.

Rather than using the language of **TIMES difference** most social scientists would either use the language **fold difference**. In this case the slope coefficient would be interpreted as,

- Each one-year difference in age is associated with a 0.95-fold difference in budget, on average.

One could also use PERCENT difference.

- Each one-year difference in age is associated with a 5% budget decrease, on average.

This can be inferred directly from the slope coefficient in the `summary()` output; without having to exponentiate:

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 ***

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 -0.0392 (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.

If you use the language of percent decrease/increase, be very careful. Percent change and percentage change are sometimes interpreted differently! It is generally more clear to use the *X-fold difference* language.

Multiple Regression

Now we can fit a model that includes our focal predictor of MPAA rating, and our control predictor of age.

```
lm.4 = lm(Lbudget ~ 1 + age + mpaa, data = movies)
```

To determine if there is an effect of MPAA rating, after accounting for differences in age, we can use the nested *F*-test to compare the full model (lm.4) with a model that does not include MPAA rating, but does include age (lm.3).

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

Analysis of Variance Table

Model 1: Lbudget ~ 1 + age

Model 2: Lbudget ~ 1 + age + mpaa

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1804	5448.0				
2	1802	4957.2	2	490.8	89.206	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The test suggests that there is a statistically significant effect of MPAA rating even after accounting for differences in age; $F(2, 1802) = 89.21$, $p < .001$. To interpret these differences, we can examine the model output:

```
summary(lm.4)
```

Call:

```
lm(formula = Lbudget ~ 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

The model-level summary information suggests that differences in age and MPAA rating of a movie explains 11.0% of the variation in budget. (Remember, explaining variation in log-budget is the same as explaining variation in budget); $F(3, 1802) = 73.84$, $p < 0.001$.

To interpret the coefficients, we will again exponentiate the fitted coefficients so we can interpret them using the raw-metric of budget.

```
exp(coef(lm.4))
```

(Intercept)	age	mpaaPG-13	mpaaR
39.9523650	0.9578146	1.2317513	0.4050910

- The model estimated budget for a PG movie (reference group) that was made in 2017 (age = 0) is 39.95 million dollars.
- Each one-year difference in age is associated with a 4.3% decrease in budget, on average, controlling for differences in MPAA rating.
- PG-13 rated movies have a budget that is 1.23 times that for PG movies, on average, controlling for differences in age. This budget difference is not statistically significant; unadjusted $p = 0.123$.
- R rated movies have a budget that is 0.41 times that for PG movies, on average, controlling for differences in age. This budget difference is statistically significant; unadjusted $p < .001$.

To examine whether the age controlled budgets for PG-13 and R rated movies differ, we need to fit a model that uses one of these categories as the reference group.

```
# Create dummy variables
movies = movies %>%
  mutate(
    PG = if_else(mpaa == "PG", 1, 0),
    R = if_else(mpaa == "R", 1, 0)
  )

# Fit model (PG-13 is reference group)
lm.5 = lm(Lbudget ~ 1 + age + PG + R, data = movies)
summary(lm.5)
```

Call:

```
lm(formula = Lbudget ~ 1 + age + PG + 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)	3.896125	0.137975	28.238	< 2e-16 ***
age	-0.043101	0.006727	-6.407	1.89e-10 ***
PG	-0.208437	0.135045	-1.543	0.123
R	-1.112080	0.088222	-12.605	< 2e-16 ***

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

```
# Exponentiate the coefficients
exp(coef(lm.5))
```

```
(Intercept)      age      PG      R
49.2113784    0.9578146    0.8118522    0.3288740
```

- R rated movies have a budget that is 0.33 times that for PG-13 movies, on average, controlling for differences in age. This budget difference is statistically significant; unadjusted $p < .001$.

Adjusting the p-Values for the Multiple MPAA Rating Comparisons

By examining the size of the p -values, it is pretty clear that after adjusting them,

- The PG vs. PG-13 difference will remain non-significant.
- The PG vs R difference will remain significant.
- The PG-13 vs R difference will remain significant.

However, for completeness, we will obtain the Benjamini-Hochberg adjusted p -values.

```
# Vector of unadjusted p-values
p.values = c(0.123, 0.000000000000679, 0.0000000000000002)
```

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

```
[1] 1.2300e-01 1.0185e-12 6.0000e-16
```

Interaction Effect between MPAA Rating and Age

To study whether there is an interaction effect between MPAA rating and age, we will fit the interaction model and compare it to the main-effects model using the nested F -test.

```
lm.4 = lm(Lbudget ~ 1 + age + mpaa, data = movies)
lm.6 = lm(Lbudget ~ 1 + age + mpaa + age:mpaa, data = movies)

anova(lm.4, lm.6)
```

Analysis of Variance Table

```
Model 1: Lbudget ~ 1 + age + mpaa
Model 2: Lbudget ~ 1 + age + mpaa + age:mpaa
  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 test suggests that we should adopt the main-effects model. The interaction-effect was not statistically significant; $F(2, 1800) = 1.41$, $p = .244$.

Plotting the Fitted Main-Effects Model

Since we adopted the main-effects model, we can also produce a plot of the back-transformed fitted values (raw-budget) 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

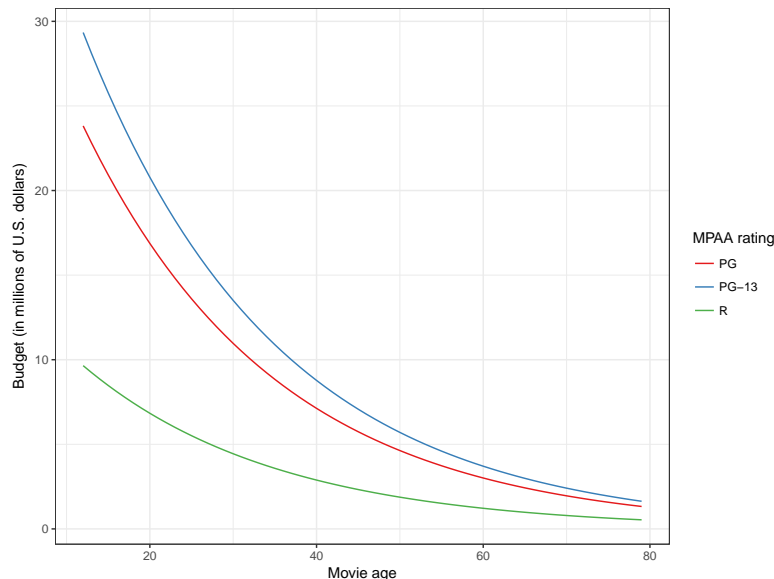
plot age vs. budget for each of the MPAA ratings. Prior to doing this, we will refit the main-effects model using the original `mpaa` variable instead of the two dummy variables. This will make creating this plot easier.

```
# 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)

# 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.