Interpretting Transformed Models

Ex. 1: Log-Transformed Explanatory Variables

Consider the following population linear regression model for the mean, which focuses on the relationship between a country's per-capita gross domestic product (GDP_i) and its average life expectancy ($LifeExp_i$):

$$\mathbb{E}[LifeExp_i | \log(GDP_i)] = \beta_0 + \beta_1 \log(GDP_i),$$

where log() represents the natural logarithm.

1. According to this regression model, what is the mean life expectancy in a population with $GDP = x^*$? *Your answer will be in terms of the parameters β_0 and β_1.*

2. According to this regression model, what is the mean life expectancy in a population for which the GDP is doubled, $GDP = 2x^*$? *Your answer will be in terms of the parameters β_0 and β_1.*

3. What is the change in mean life expectancy between these two populations? *Your answer will be in terms of the parameters β_0 and β_1.*

4. Does your answer in part (3) depend on the initial value of x^*? (And what does it mean/why is it interesting if your answer doesn't?)

Exercise 2. Analyzing Movie Budgets and Gross Earnings From 2008

Let's revisit our fitted regression model for the total gross of a movie released in 2008 (measured in millions of dollars and denoted by y_i for the ith movie), modeled as a function of the movie's budget (x_i, also measured in millions of dollars). Based on our random sample of $n = 206$ movies, we found:

$$\widehat{\log(y_i)} = 0.604 + 0.846 \log(x_i)$$

1. Interpret the slope of this regression model *in context* and *on the original scale of the data.* Take $c = 2$.

2. The R output below shows the 95% confidence interval for the slope of the double-log-transformed model. How might we convert this confidence interval back to the original scale of the data? Calculate and interpret this converted 95% confidence interval *in context* and *on the original scale of the data.* Take $c = 2$.

```
ll_movies_lm <- lm(log(Total_Gross / 1000000) ~ log(Budget_2008 / 1000000), data = movies_08)
confint(ll_movies_lm, level = 0.95)
```

```
                             2.5 %    97.5 %
(Intercept)             0.06092836 1.1474486
log(Budget_2008/1e+06) 0.69416016 0.9970975
```

3. The R output below shows the point estimate and 95% confidence interval for the mean log gross of a movie with a $200,000,000 budget.

```
big_budget <- data.frame(Budget_2008 = 200000000)
ll_movies_lm |>
  broom::augment(newdata = big_budget, interval = "confidence", conf.level = 0.95)
```

```
# A tibble: 1 x 4
  Budget_2008 .fitted .lower .upper
        <dbl>   <dbl>  <dbl>  <dbl>
1   200000000    5.08   4.78   5.39
```

1. Convert both the point estimate and the lower and upper bounds of the 95% confidence interval back to the original data scale.

2. Do these numbers correspond to the estimated mean movie gross among all movies with a $200,000,000 budget? The median movie gross among movies in this subpopulation? Some other feature (like the standard deviation) of that movie gross distribution?

3. Interpret the 95% confidence interval *in context* and *on the original scale of the data.*