W203 Lab 02 Report - Team 02

Ayoade Israel, Divya Menghani, Hamsini Sankaran, Sivakumar Thiyagarajan

```
## [1] 2998
```

0.1 Introduction

As today's housing stock ages, the choice to remodel a property is driven by many factors, some personal and some market-wide. For many owners, a major concern is the effect a remodel will have on future revenues, including those from renting and from selling the property. Some may turn to real estate professionals for guidance. In a 2019 survey performed by the National Association of Realtors, members estimated the financial benefit of different type of remodels, ranging from $2,500 for a closet renovation up to $20,000 for a full kitchen remodel.

While experts may provide broad guidelines, data-based approaches are needed to reduce uncertainty in the value of remodels. In the aggregate, remodeling accounts for $400 billion spent each year in the US.[1] Uncertainty in how much of this money can be recouped may contribute to the misallocation or underprovision of resources in this sector of the economy.

This study estimates the economic value for remodeling a home empirically, utilizing observations of house sales in Ames, Iowa. The data shows the timing of the last remodel before a house is sold, but does not distinguish between different types of remodels. Applying a set of regression models, I estimate the value that results immediately when a house is remodeled, and also the rate at which it decays over time.

0.2 Data and Methodology

We gathered the dataset from Kaggle. The dataset comprises metadata for more than 700,000 movies listed in the TMDB Dataset. It has 722986 rows and 20 columns. It is relevant and provides an opportunity to analyze the relationship between the budget and the revenue of the movies. The dataset contains unique information on films, with each movie being distinct from the others and identified by an id column in the dataset. Additionally, the distribution of the X and Y concept is identical, satisfying the assumptions of IID. Because each movie is not subject to multiple samplings, the dataset is regarded as cross-sectional.Each row in the data represents a movie. I performed all exploration and model building on a 30% subsample of the data. The remaining 70%, totaling 9992 rows, was used to generate the statistics in this report.

To determine the factors that affect movie revenue, we need to extensively analyze data and conduct research .To analyze the factors that affect movie revenue, we have defined inclusion criteria that could potentially impact the dependent variable "revenue" (Y), and selected "budget" as the independent variable. We are focusing on movies that have been officially released to the public, and to refine our analysis, we are excluding movies with budgets or revenues exceeding $999 and budget values of $5 billion and $800 million. After filtering the data, we are left with a sample size of r filtered_rows observations.

I am interested in analyzing the relationship between movie revenue and budget, and identifying the factors that impact this relationship. To determine the covariates that impact the dependent variable, we have chosen the following variables:

- Runtime: An important aspect of the movie-watching experience that has a correlation with revenue.
- Vote_count: A proxy for popularity and an indicator of the level of engagement a movie has received from the public.
- Popularity: A comprehensive metric that takes into account various factors such as user ratings, page views, and social media mentions.

[1] La Jeunesse, Elizabeth. "Healthy Home Remodeling: Consumer Trends and Contractor Preparedness." Harvard's Joint Center for Housing Studies (2019).

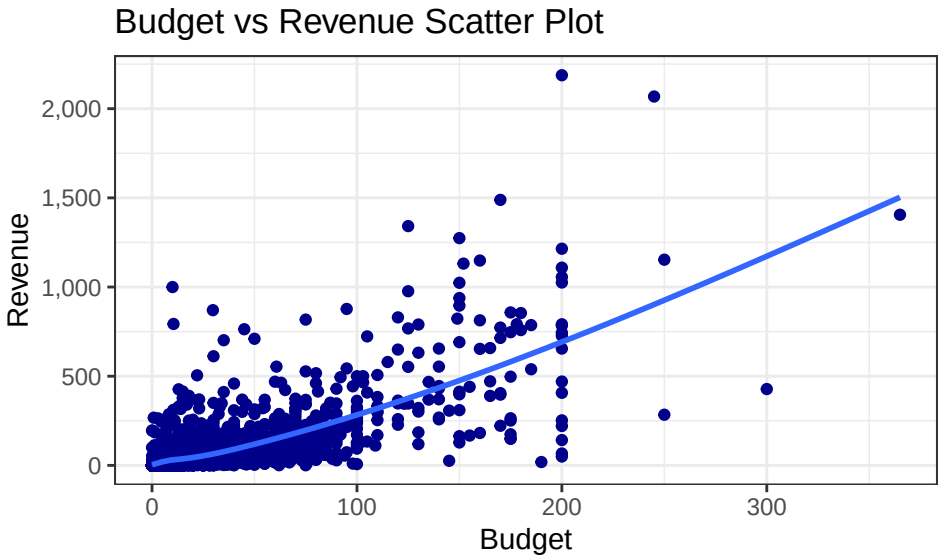

Figure 1: Revenue as a Function of Budget

- Title_length: A factor that can impact audience interest and expectations, with shorter titles being easier to remember and share.

We choose these covariates based on their potential impact on the dependent variable, and have been widely studied and recognized in the industry. By including these covariates in our analysis, we aim to gain insights into the factors that affect movie revenue and provide recommendations for filmmakers to optimize their marketing and production strategies. We operationalize the variables Budget(X) and Revenue(Y) in terms of amount. This form was chosen to best fit the relationship seen in exploratory plots. Figure 1 plots revenue as a function of budget. There is a general positive relationship between budget and revenue

- The age of the house when remodeled. It is possible that remodeling an older home results in a larger value increase than remodeling a newer home.
- The time between the remodel and the sale. It is possible that remodels become more or less valuable over time, and the rate of change may be different than that of houses in general.

Exploratory plots suggest that both effects exist and that both are roughly linear. I therefore create regression models in which the "boost" from remodeling increases by a fixed amount with the age of the house when remodeled, and also changes by another fixed amount with each year that passes after the remodel. In other words, I fit regressions of the form,

$$\widehat{revenue} = \beta_0 + \beta_1 \cdot Budget + \beta_2 \cdot votecount + \beta_3 \cdot titlelength + \beta_4 \cdot runtime + \mathbf{Z}\gamma$$

where R is an indicator for remodeling, β_1 represents the immediate increase in value per year the house existed before remodeling, β_2 represents the change in the value increase for each year that passes after the remodel, $\mathbf{Z}$ is a row vector of additional covariates, and γ is a column vector of coefficients.

I considered specifications that also include the modeling indicator R by itself (i.e. uninteracted). This type of model allows for the possibility that even a brand new house that is remodeled immediately increases in value. However, when fitting such models in the exploration set, the resulting coefficient was practically small (equivalent to reducing the age of a home by 1 to 2 years) and non-significant. To improve the precision of my estimates and the simplicity of the model, I removed this term.

0.3 Results

1 Stargazer Table for the linear model

	Dependent variable:
	log_revenue
log_budget	0.789*** (0.009)
vote_count	0.0002*** (0.00001)
log_title_length	0.123*** (0.034)
runtime	0.006*** (0.001)
Constant	-0.528*** (0.120)
Observations	6,992
R2	0.684
Adjusted R2	0.684
Residual Std. Error	1.622 (df = 6987)

Note: *p<0.05;* ***p<0.01;*** p<0.001