Homework: One-night accommodation

Context

The aim of this exercise is to study accommodation prices in Vienna and identify the most under-priced hotel for a one-night stay in the city. The dataset 'hotels_one_night.xlsx' includes information about hotels and apartments in Vienna.[1] Specifically, the variables are defined as:

- 'price': Daily price for one night in USD.
- 'accommodation_type': Type of accommodation.
- 'stars': Number of stars.
- 'rating': average rating of the customers (out of 5).
- 'distance': Distance - from main city center.

Exploratory Data Analysis

1.a) We expect that the distance to the city center affects the price. Create the following scatter plots:

1. A scatter plot showing the distance compared to the price.
2. A scatter plot of the natural logarithm of the distance (i.e., ln(distance)) compared to the natural logarithm of the price (i.e., ln(price)).
3. Which transformation results in the strongest linear relationship between the two variables? Provide a numerical answer for supporting your claim.

1.b) Let us investigate the relationship between price and rating. Create the following scatter plots:

1. A scatter plot showing the rating compared to the natural logarithm of the price (i.e., ln(price)).
2. A scatter plot of the exponential of the rating (i.e., exp(rating)) compared to the natural logarithm of the price (i.e., ln(price)).

[1] Source: https://gabors-data-analysis.com/

3. Which transformation gives the strongest linear relationship between the two variables? Provide a numerical answer for supporting your claim.

Multiple linear regression

Consider the following linear regression:

$$\textbf{Model 1: } \ln(\text{price}_i) = \beta_0 + \beta_1 \text{Hotel}_i + \beta_2 \text{stars3}_i + \beta_3 \text{stars4}_i + \beta_4 \text{stars5}_i + \beta_5 \text{stars3\_Hotel}_i + \beta_6 \text{stars4\_Hotel}_i + \beta_7 \exp(\text{rating}_i) + \beta_8 \ln(\text{distance}_i) + \epsilon_i,$$

where

- Hotel_i is a dummy variable equal to 1 if the accommodation is a hotel,
- stars3_i is a dummy variable equal to 1 if the accommodation has a star of 3 or 3.5,
- stars4_i is a dummy variable equal to 1 if the accommodation has a star of 4 or 4.5,
- stars5_i is a dummy variable equal to 1 if the accommodation has a star of 5,
- stars3_Hotel is an interaction variable between stars3_i and Hotel_i,
- stars4_Hotel is an interaction variable between stars4_i and Hotel_i.

2.a) Estimate Model 1. What is the estimated value of β_1, which is associated with the hotel variable?

2.b) How does the variable stars4_Hotel impact the log price ? Provide a numerical interpretation.

2.c) How does the variable distance **(not the log distance)** impact the price **(not the log price)** ? Provide a numerical interpretation.

2.d) Predict the one-night price **(not the log price)** for a 4-star hotel with a rating of 3.5 at a distance of 1.

2.e) Which variables are not significant at a 95% confidence level? Hint: For each variable, perform a two-sided test with the Null hypothesis being the related parameter equal to 0.

Multiple linear regression: second model

We now consider a second model which is equivalent to Model 1 with the additional explanatory variable 'distance':

$$\begin{aligned}\textbf{Model 2: } \ln(\text{price}_i) \quad = \quad & \beta_0 + \beta_1 \text{Hotel}_i + \beta_2 \text{stars3}_i + \beta_3 \text{stars4}_i + \beta_4 \text{stars5}_i + \beta_5 \text{stars3\_Hotel}_i \\ & + \beta_6 \text{stars4\_Hotel}_i + \beta_7 \exp(\text{rating}_i) + \beta_8 \ln(\text{distance}_i) + \beta_9 \text{distance}_i + \epsilon_i,\end{aligned}$$

3.a) Discuss two statistical reasons for choosing Model 2 over Model 1.

3.b) Create a diagnostic plot to test the linearity assumption. What do you conclude about this assumption?

3.c) Create a diagnostic plot to test the homoskedasticity assumption. What do you conclude about this assumption?

3.d) Create a diagnostic plot to test the assumption of normality of the error term. What do you conclude about this assumption? Explain why this assumption is not critical in this context.

3.e) Using Model 2, identify the most under-priced accommodation with at least 4 stars.