Homework 3

Due Thursday, May 28th at 11:59 PM

This homework is supposed to be more comprehensive in nature, covering Lectures 2-16 and Labs 2-8. Please write your solutions in an R Markdown file and submit your PDF output to Gradescope. All code for your solutions should be shown in your PDF file.

Introduction

This assignment asks you to carry out a complete statistical inference workflow from start to finish, using a single real-world dataset. The assignment is divided into two parts: **Question 1** focuses on a continuous (quantitative) outcome variable and will require you to use at least one interaction term in your fitted model, while **Question 2** focuses on a binary outcome variable. Both parts will follow the same four-phase workflow:

(i) *Pre-Specified Plan.* During this phase, you will provide your list of predictors of your choice. Before running any model, you will specify what your hypotheses are given those chosen variables.

(ii) *Descriptive statistics and data manipulation.* During this phase, you will look through the data and try to understand some of its nuances. This will allow you to see whether your chosen predictors are missing any values, whether they exhibit some skewness, or whether they are categorical, numerical, or other. Here, you will also apply any data manipulations that are appropriate for either the predictors you have chosen or for the outcome variable (e.g. removal of missing value, transformation of a predictor).

(iii) *Primary Analysis.* During this phase, you will carry out the model you specified in your *Pre-Specified Plan.*

(iv) *Secondary Analysis.* During this phase, you will re-examine and stress tests the findings of your *Primary Analysis*, hoping to get a better idea of how your results fit in within the broader context of possible results one might have gotten from this dataset.

It is important that you complete phase (i) before the subsequent phases in each part of the assignment; review Lecture 13 if you need a refresher on why defining our model before doing any inference is important for the validity and truthfulness of the findings.

Dataset

The dataset, provided in the `hotel_bookings.csv` file, contains records of roughly 120,000 individual hotel bookings across two hotels in Portugal.[1] Bookings span arrivals to the hotels between July 2015 and August 2017 and include both those that were seen through (honored) and that were not (canceled).

This is not a perfect dataset. There are some strings with value `NULL` (instead of a true missing value, `NA`) in some rows for which the `agent` or `company` are not there. As such, neither of these two columns

[1]This dataset was originally found from this Kaggle website: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand.

are all that helpful as predictors, so it makes sense to not include them. Moreover, you are not to use `reservation_status` or `reservation_status_date` as predictors since these two columns encode final information about the booking that were provided after the fact; using them as part of an inference model would be a clear case of data leakage. Lastly, the dataset contains a hefty amount of rows with identical values across all columns. This is a documented feature of the original data since each row represents a distinct booking event even if some covariate values might collide, meaning that these rows should not be removed.

For your reference, here is a brief idea for what all of the variables in the dataset are:

```
## Rows: 119,390
## Columns: 32
## $ hotel                           <chr> "Resort Hotel", "Resort Hotel", "Resort~
## $ is_canceled                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, ~
## $ lead_time                       <dbl> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, ~
## $ arrival_date_year               <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 201~
## $ arrival_date_month              <chr> "July", "July", "July", "July", "July",~
## $ arrival_date_week_number        <dbl> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,~
## $ arrival_date_day_of_month       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ stays_in_weekend_nights         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ stays_in_week_nights            <dbl> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, ~
## $ adults                          <dbl> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ children                        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ babies                          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ meal                            <chr> "BB", "BB", "BB", "BB", "BB", "BB", "BB~
## $ country                         <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "GBR~
## $ market_segment                  <chr> "Direct", "Direct", "Direct", "Corporat~
## $ distribution_channel            <chr> "Direct", "Direct", "Direct", "Corporat~
## $ is_repeated_guest               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_cancellations          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_bookings_not_canceled <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ reserved_room_type              <chr> "C", "C", "A", "A", "A", "A", "C", "C",~
## $ assigned_room_type              <chr> "C", "C", "C", "A", "A", "A", "C", "C",~
## $ booking_changes                 <dbl> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ deposit_type                    <chr> "No Deposit", "No Deposit", "No Deposit~
## $ agent                           <chr> "NULL", "NULL", "NULL", "304", "240", "~
## $ company                         <chr> "NULL", "NULL", "NULL", "NULL", "NULL",~
## $ days_in_waiting_list            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ customer_type                   <chr> "Transient", "Transient", "Transient", ~
## $ adr                             <dbl> 0.00, 0.00, 75.00, 75.00, 98.00, 98.00,~
## $ required_car_parking_spaces     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ total_of_special_requests       <dbl> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 3, ~
## $ reservation_status              <chr> "Check-Out", "Check-Out", "Check-Out", ~
## $ reservation_status_date         <date> 2015-07-01, 2015-07-01, 2015-07-02, 20~
```

Unless otherwise specified, assume a significance threshold of $\alpha = 0.05$ to be the default level for this assignment. You should use this value when discussing your findings.

Question 1 – Quantitative Outcome Variable

In this part of the assignment, your outcome variable is `adr`, the Average Daily Rates in Euros, which represents the average revenue per occupied room night across the length of the hotel stay. You may choose any of the following variables as your primary covariate: `lead_time`, `market_segment`, `hotel`, or `stays_in_week_nights`.

(a) State your primary covariate of interest and chose one hypothesized interaction term from the list above. Include a brief domain-based explanation for your choice. State the null and alternative hypothesis for this covariate using model parameters and words (do not forget the interaction), and then write out the full model using those parameters. [This will now be referred to as your primary model.]

(b) Provide some detailed descriptive statistics about all of the variables that feature in your primary model. This should include a summary table of those variables and some plots capturing the relationships between the model variables and `adr`. The kind of plots will depend on the type of predictor, but comment briefly with some observations on those plots, and be sure to plot the marginal distribution of `adr` since it is our outcome variable. Also state any transformations you would make to any of the variable (predictors or outcome), justifying your choice with a reasoned argument. (If you do not feel any transformations are necessary, just say so.)

(c) Fit the specified model to perform your primary analysis. When reporting your results, you should:

1. Explicitly restate your null hypothesis and primary model from (a). Present a summary of your model results, and state the decision for your null. Explain, in words, what this decision means.
2. Interpret the interaction term in words by including some discussion of the direction and magnitude of the interaction and its practical meaning within the context of the outcome variable.
3. For any statistically significant effect, report an appropriate effect size alongside its obtained p-value. For a regression coefficient, this would mean expressing the estimated effect in interpretable units. (You could do this by providing the percent change in `adr` based on the effect, the fraction of `adr`'s standard deviation explained by this effect, or by discussing some values contained in and not contained in a 95% confidence interval for this effect.)

(d) Perform some secondary analysis on your model. You should:

1. Assess whether the conditions of the primary model you tested are satisfied as described in Lecture 10. Produce relevant plots and describe, briefly, what they are intending to show and whether there exists evidence that the conditions are violated.
2. Fit at least one alternative model that differs meaningfully from your primary model and present a side-by-side comparison of the results (e.g. coefficient estimates, p-values). Describe the differences between the primary and alternative models, and comment on the robustness of your findings: does the alternative model yield markedly different conclusions, or do the conclusions from the primary model hold up?

For part (d), reasonable alternative models could be in the form of variable selection (including/excluding a predictor, adding/omitting/replacing a confounder) or, if you made use of a transformation in your primary model, applying a different one or removing it entirely. Another option might be rescaling a continuous predictor to be a categorical one or vice versa. Be explicit in what your alternative model is.

Question 2 – Binary Output Variable

In this part of the assignment, your outcome variable is `is_canceled`, a binary indicator variable that identifies whether the booking was ultimately canceled (encoded as a 1) or honored (encoded as a 0). You may choose any of the following as your primary covariate of interest: `lead_time`, `previous_cancellations`, or `deposit_type`.

(a) State your primary covariate of interest and one confounder from the list above. Include a brief domain-based explanation for your choice. State the null and alternative hypothesis for this covariate using model parameters and words, and then write out the full logistic regression model using those parameters. [This will now be referred to as your primary model.]

(b) For the descriptive statistics, you should:

1. State what you might expect the class distribution of `is_canceled` to be based solely on the variable name and description. Find the true class imbalance in the data, and comment on whether this imbalance might be an issue for your model.
2. Produce a summary table of your chosen predictor, stratified by `is_canceled`. Produce plots capturing the relationships between your chosen predictor and `is_canceled`. The kind of plot you will produce will depend on the type of predictor, but comment briefly with some observations.

(c) Fit the specified model to perform your primary analysis. When reporting your results, you should:

1. Explicitly restate your null hypothesis and primary model from (a). Present a summary of your model results, and state the decision for your null.
2. Interpret the coefficient of your covariate in terms of its odds ratio.
3. Regardless of your decision with respect to your null hypothesis, can you say that statistical significance alone can help you determine an optimal classification threshold for predicting cancellations? That is, does finding a statistically significant p-value automatically give a good classification threshold for predicting cancellations? Why or why not?

(d) Perform some secondary analysis on your model. You should:

1. Assess whether the conditions of the primary model you tested are satisfied as described in Lecture 16. Produce relevant plots and describe, briefly, what they are intending to show and whether there exists evidence that the conditions discussed in lecture are violated.
2. Fit at least one alternative model that differs meaningfully from your primary model and present a side-by-side comparison of the results (e.g. coefficient estimates, p-values). Describe the differences between the primary and alternative models, and comment on the robustness of your findings: does the alternative model yield markedly different conclusions, or do the conclusions from the primary model hold up?

For part (d), reasonable alternative models could be in the form of variable selection (including/excluding a predictor, adding/omitting/replacing a confounder) or, if you made use of a transformation in your primary model, applying a different one or removing it entirely. Another option might be rescaling a continuous predictor to be a categorical one or vice versa. Be explicit in what your alternative model is.