Design of Experiments

Lab 4

Nick Levitt, Andre Duarte, Joshua Amunrud

June 28, 2017

In this assignment, we wish to investigate four different factors and how they affect a user's tendency to book a ride with Lyft. The table below shows the experimental factors and their levels.

Factor	Description	Low	High
x_1	Discount Amount (%)	10	50
x_2	Discount Duration (days)	1	5
x_3	Method of Dissemination	In-app	Email
x_4	Ride Type	Standard	Shared

Our goal here is to find the levels of each factor that simultaneously maximize the booking rate for Lyft using a factorial design.

Design Strategy

We opted for a full factorial design for this experiment. Our reason for this was that there are only four distinct factors, with two levels each, giving a total of $2^4 = 16$ conditions for experimentation. The usual constraint of data collection, which would normally hamper one's ability to run an experiment with 16 conditions, is not an issue given the design of this lab. Therefore, because we have the computational power to run the linear regression model on such a dataset, we determined that running the full factorial design will give the most insight into the effects of the different factors.

Analysis

For our analysis, we chose to fit a regression model using the main effects and two-way-interactions for each effect. We chose to not include any three-way interactions for our analysis due to the sparsity of effects principle (more information is given on this in the “Assumptions” section below).

Assumptions

As we are only given the percentage of booked rides in each case, we make some assumptions about the underlying data that generated these values. The first assumption is that the data was randomly/independently generated, and not, for example collected over a period of time for one set of factors and another period of time for another set. The second assumption is that the underlying data is normally distributed and not highly skewed.

We also assume that the sparsity of effects principle holds for our experiment, and that high-order interaction effects are not likely to be statistically significant. This allows us to omit the three-way and four-way interaction effects when fitting our linear regression model.

Results

The results of fitting a linear regression model to our design matrix and the associated results can be found in Table 1 below. It is important to remember that the model coefficient estimates for each effect (not the intercept) are to be interpreted as one half of the corresponding factor’s effect estimate (for example, the main effect for x_1 is $2 \times 0.31875 = 0.6375$. In other words, going from the low to the high level of x_1, holding all other variables constant, leads to an increase in 0.6375% in the booking rate on average).

Table 1: Regression Coefficients

	Estimate	**2.5%**	**97.5%**	**Std. Error**	**t value**	**Pr(> $\|t\|$)**
(Intercept)	7.03125	6.998324218	7.06417578	0.01281	548.944	3.81e-13
x_1	0.31875	0.285824218	0.35167578	0.01281	24.885	1.95e-06
x_2	0.18125	0.148324218	0.21417578	0.01281	14.151	3.17e-05
x_3	0.03125	-0.001675782	0.06417578	0.01281	2.440	0.058672
x_4	-0.21875	-0.251675782	-0.18582422	0.01281	-17.078	1.26e-05
x_1x_2	0.09375	0.060824218	0.12667578	0.01281	7.319	0.000746
x_1x_3	0.01875	-0.014175782	0.05167578	0.01281	1.464	0.203111
x_1x_4	0.21875	0.185824218	0.25167578	0.01281	17.078	1.26e-05
x_2x_3	-0.01875	-0.051675782	0.01417578	0.01281	-1.464	0.203111
x_2x_4	-0.01875	-0.051675782	0.01417578	0.01281	-1.464	0.203111
x_3x_4	-0.01875	-0.051675782	0.01417578	0.01281	-1.464	0.203111

Residuals

Figure 1 shows the diagnostic plots for the model residuals. In particular, we can see that they seem to be fairly normally distributed, as shown by the histogram and the QQ-Plot. The plot of Residuals vs. Order shows that there is no dependence of residuals over time. The plot of Residuals vs Fitted Valued shows that the residuals seem homoskedastic. All model assumptions are therefore validated by these figures.

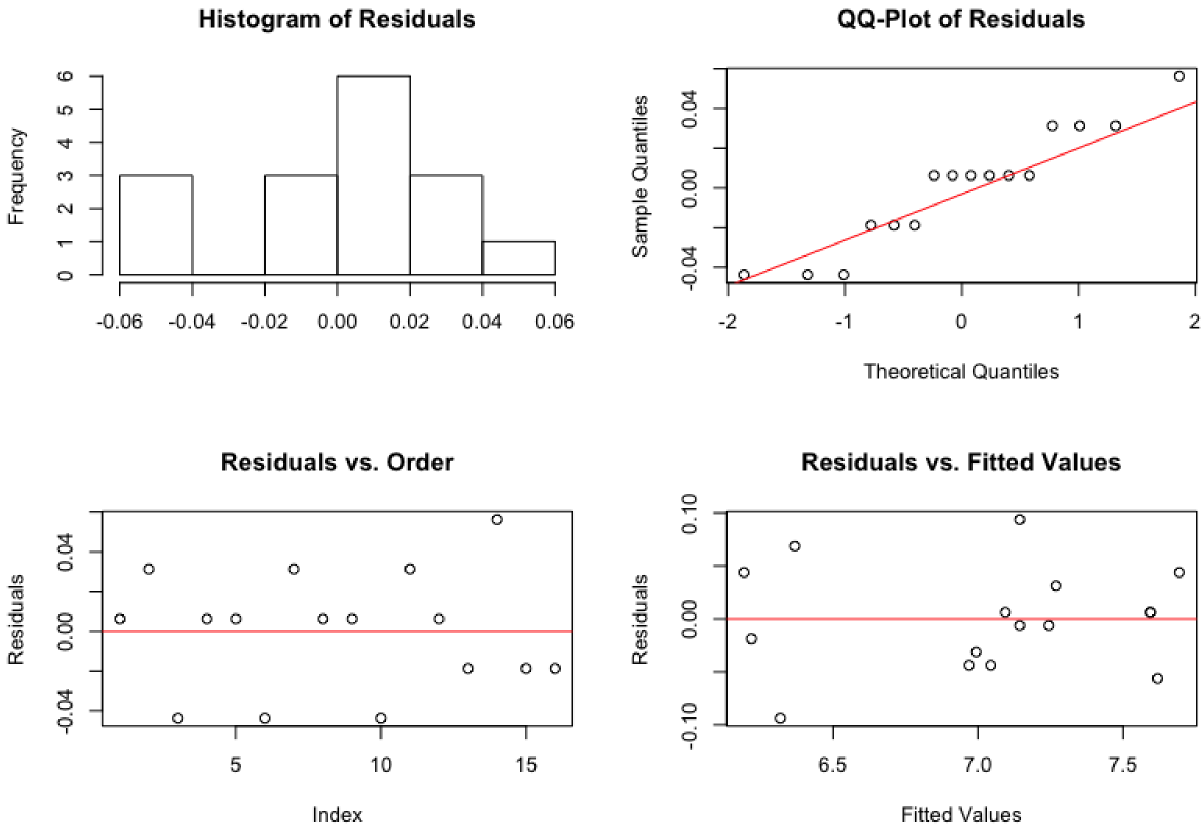

Figure 1: Diagnostic Plots

Main Effects

Three main effects were found to be statistically significant: the discount amount (x_1, A), the discount duration (x_2, B) and the ride type (x_4, D). The method of dissemination (x_3, C) was found to be statistically significant at the 10% significance level, but the 95% confidence interval for the fitted coefficient includes 0, so we feel confident in stating that the method of dissemination has negligible impact on booking rate. Plots of the predicted four main effects, as well as 95% confidence intervals can be seen in Figure 2.

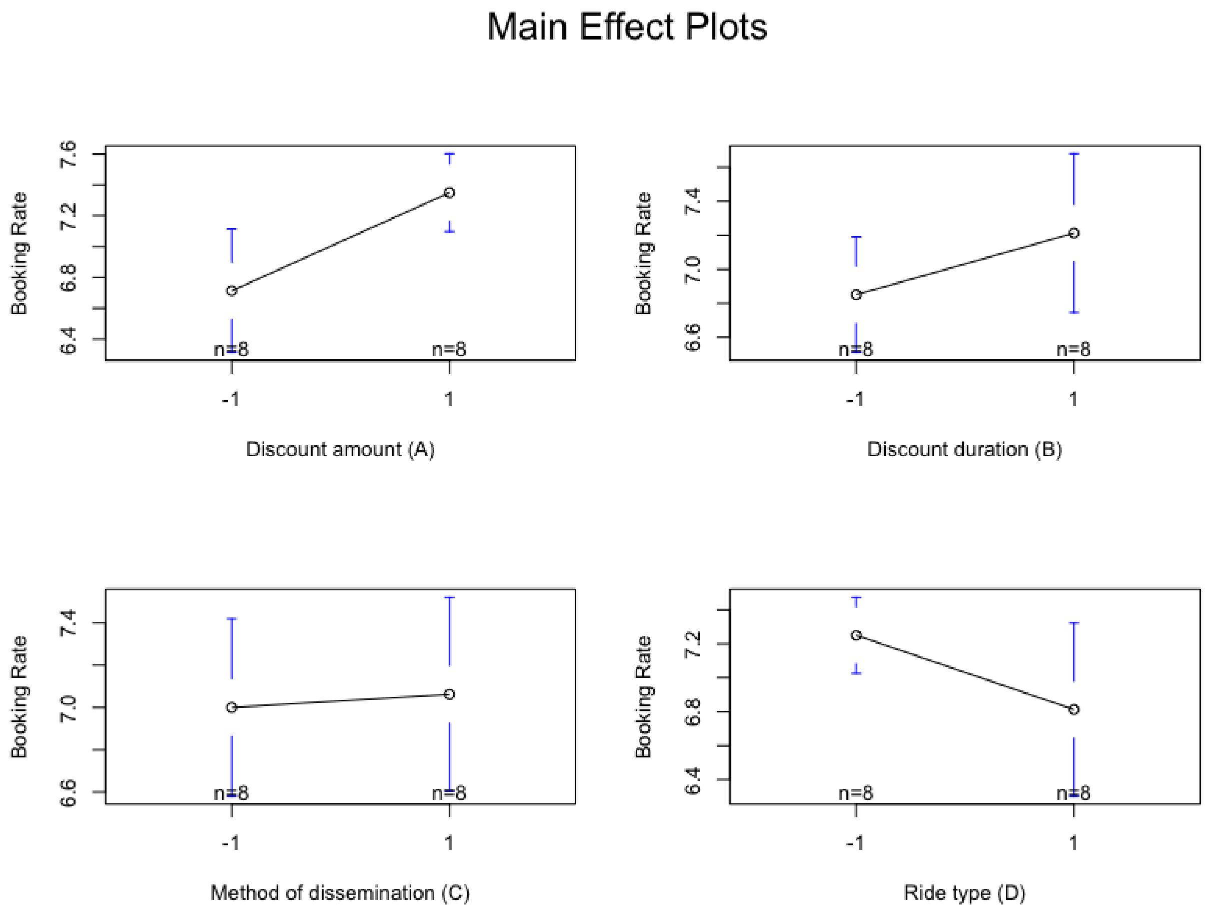

Figure 2: Main Effects

Interaction Effects

The interaction between discount amount (x_1, A) and discount duration (x_2, B) is statistically significant. It has a positive coefficient in the fitted linear regression model, indicating a compounding effect for these two factors. This means that offering a 50% discount has a more positive impact on booking rate when offered for 5 days compared to 1 day. Similarly, this indicates that offering a discount for 5 days has a more positive impact on booking rate when the discount offered is 50% compared to when the discount offered is 10%.

The other interaction effect found to be statistically significant was the interaction between discount amount (x_1, A) and ride type (x_4, D). As can be seen in Figure 3, when the discount amount is only 10%, the ride type offered has a large impact on the booking rate.

However, when the discount amount if 50%, the ride type offered has little impact on the booking rate.

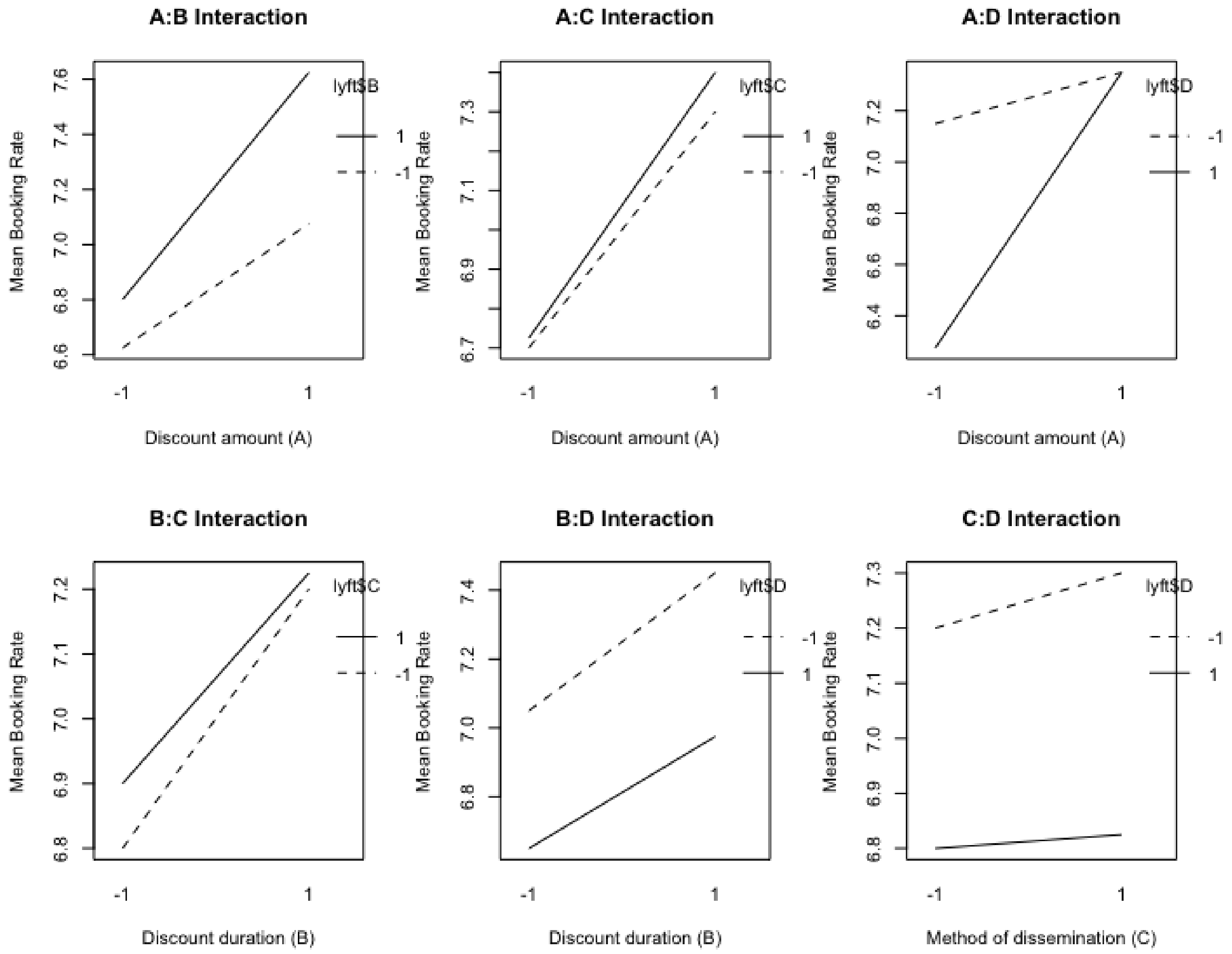

Figure 3: Interaction Effects

Recommendation

Based on our analysis, the best promotion would include a discount amount of 50% for shared rides which lasts 5 days and is disseminated via email. The predicted booking rate for this promotion is 7.69375. The true booking rate will be between 7.522663 and 7.864837 95% of the time.